universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Clustering Aviation Markets: Market Opportunity Identification through Machine Learning

verfasst von | submitted by

Gloria Widhalm

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 977
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Business Analytics
Betreut von | Supervisor:	Univ.-Prof. Dr. Jan Fabian Ehmke

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Univ.-Prof. Dr. Jan Fabian Ehmke, for his theoretical insights and guidance throughout this thesis, as well as for the opportunity to apply research concepts to a practical-oriented project. I am also deeply thankful to Omar Aitoulghazi, BSc MSc. for his support and expertise in advanced machine learning techniques and for patiently answering my numerous questions.

I would also like to thank Seabury Airline Strategy Group (ASG) for their cooperation, for the chance to work with interesting datasets, and for sharing relevant aviation domain knowledge that shaped many of my decisions.

Finally, I am grateful to my family, friends, and partner for their encouragement, patience, and understanding during the completion of this master thesis.

Abstract

This thesis presents a comprehensive global clustering analysis of non-directional aviation markets (airport-pairs) based on supply-side features derived from a 2024 flight schedule and enriched with demographic, tourism, and climate data. Multiple clustering algorithms (partition-, density- and distribution-based) were applied under varying parameter settings and data pre-processing techniques. Results were assessed using internal validation metrics (Silhouette Coefficient, Davies–Bouldin Index) and two-dimensional projections (UMAP) for qualitative inspection. Gaussian Mixture Modeling (GMM) delivered the highest internal scores overall. The interpretation of clustering results revealed both evenly-sized clusters, reflecting globally homogeneous markets, and imbalanced clusters, containing niche market segments. Applications for strategic network planning include identifying under-served market opportunities and tailoring route strategies. Overall, this research confirms that clustering aviation markets at a global scale is feasible and valuable, yet demands careful choice of features, algorithms, and interpretability techniques.

Kurzfassung

Diese Arbeit erläutert eine globale Clusteranalyse nicht-direktionaler Flugmärkte (Paare von Flughäfen) basierend auf Angebotsmerkmalen aus dem Flugplan 2024, ergänzt durch demografische, touristische und klimatische Daten. Es wurden verschiedene Clustering-Algorithmen (partitions-, dichte- und verteilungsbasiert) unter unterschiedlichen Parametern angewandt. Die Ergebnisse wurden mit internen Validierungsmetriken (Silhouette-Koeffizient, Davies–Bouldin Index) und zweidimensionalen Projektionen (UMAP) qualitativ verglichen. Gaussian Mixture Models erzielten insgesamt die höchsten Ergebnisse gemessen an den Metriken. Die Interpretation der Ergebnisse zeigte homogene Cluster sowie unausgeglichene Cluster mit Nischensegmenten. Strategische Anwendungsmöglichkeiten liegen in der Identifikation von Märkten mit Wachstumspotenzial und der Optimierung von Streckenstrategien. Insgesamt bestätigt die These die Machbarkeit und den Nutzen einer globalen Marktsegmentierung via Clustering, macht jedoch die Bedeutung einer sorgfältigen Auswahl von Merkmalen, Algorithmen und Interpretationsmethoden deutlich.

Contents

List of Tables

List of Figures

1. Introduction

In 2024, airlines operated more than 36 million direct flights around the world (based on the data in the flight schedule). The busiest route in the world in 2024, Jeju (CJU) to Seoul (GMP) in South Korea, had a capacity of 14.2 million seats alone, which is equivalent to roughly 39,000 daily seats (*Busiest Flight Routes in the World 2024* 2025).

The aviation industry is a complex and highly competitive field. Aviation has been shaped by the liberalization of markets, the emergence of low-cost carriers, and digital booking platforms. Identifying new growth opportunities and optimizing route networks are essential parts of airline planning processes. Strategic stakeholders often lack detailed methods to identify similar market structures within the global aviation network. This thesis addresses the gap by segmenting non-directional aviation markets at a global scale using unsupervised machine learning techniques. The overarching goal is to develop market groupings that can support network planning, resource allocation, and market opportunity identification for airlines, airports, and policy makers.

With over 30,000 global non-directional airport-pairs and a yearly flown capacity exceeding tens of millions of seats, manually segmenting markets for strategic insight is infeasible. Unsupervised clustering offers a data-driven approach to reveal latent market structures without the need for existing labels. Potential applications include the identification of under-served routes for network expansion, benchmarking performance between segments, and informing resource allocation decisions at airports. While prior studies have clustered airports or passenger behaviors, few have focused on airport-pair markets, and none have combined flight schedule data with demographic, tourism, and climate features at a global scale. By combining schedule-based measures (capacity, seasonality) with indicators of touristic interest (population, hotel density) and climate seasonality, both business-driven and leisure-driven market characteristics are included.

The thesis is structured as follows: First, chapter 2 introduces key terminologies relevant to this thesis. Chapters 3 and 4 give an overview of the theoretical foundations (CRISP-DM, clustering paradigms) and the literature on aviation clustering and market segmentation. The methodology chapter (5) defines the research questions and the methodological framework of this thesis. Data collection, exploration, and feature engineering are described in chapter 6. Next, chapter 7 outlines modeling experiments and parameter settings, while chapter 8 presents results and interpretation approaches of the clusters. The final chapter (9) concludes with addressing the research questions, discussion of the limitations, and future work.

2. Key Terminologies

Unsupervised Machine Learning
Machine learning techniques or algorithms that discover hidden patterns without human intervention or labeled data (ground truth) (R. Xu and Wunsch 2005).

Clustering
An unsupervised machine learning technique that groups similar objects or data points based on a measure of similarity (Jain, Murty and Flynn 1999).

Two Letter Codes for Airlines
Airlines are often represented in two letter codes such as "OS" for Austrian Airlines.

Three Letter Codes for Airports
Airports can be represented by three letter codes such as "VIE" for Vienna Airport, "JFK" for John F. Kennedy International Airport, or "FRA" for Frankfurt Airport.

Flight, Segment, Leg, Connection, Route and Itinerary
A flight refers to the connection between two airports via an aircraft as transportation mode. A segment and a leg can be used as synonyms for a flight, however, these terms refer to a part of a connection (there is more than one flight in the journey of the traveler), whereas a flight usually references a direct flight. When referencing a segment in this thesis, it refers to a market segment (non-directional airport pair). Route and itinerary are also used interchangeably in the scope of this thesis. They refer to the entire path that is flown from origin to destination airport.

Aviation Market, Direct Market, OD market
An aviation market can be defined as an airport pair (e.g., VIE-FRA) but can also refer to a regional or country pair (e.g., Europe-Asia). In the scope of this analysis, a market is

defined as an airport pair (if not stated otherwise). Furthermore, based on the available data (schedule of direct flights), only direct markets are included. Airport market, direct market and aviation market are treated as synonyms within this thesis. An important note is that these markets can be either directional or non-directional. Directional means that VIE-FRA is not the same market as FRA-VIE, whereas non-directional means that these two markets are consolidated into one market (e.g., FRA-VIE, often alphabetically ordered). An OD market refers to an origin-destination market. Again, these can be pairs of airports or regions. For OD markets, connections are usually also included (not only direct flights). The focus of this thesis lies on non-directional direct markets.

Local and Transfer Passengers

Local (or direct) passengers fly directly from their origin to their destination. They can also be referred to as single-leg passengers. Transfer or segment passengers need to transfer at one or multiple intermediate airports on their route (traveling via multiple legs) (Maertens, Grimme and Bingemer 2020).

Catchment Area

From a human geography perspective, a catchment area is the area that a city or service can attract visitors or customers (Lieshout 2012; Baltazar and Silva 2018). Other terms for catchment area are "Hinterland" or "Umland" and can be used interchangeably in most cases. From an aviation perspective, a catchment area refers to the geographical reach of the services of an airport to the surrounding area, population and economy. In other words, it can also be defined as the area most outbound travelers originate from and most inbound passengers travel to. There are different approaches to define catchment areas. One approach is the spatial definition within a fixed radius of for example 50km (Baltazar and Silva 2018). Another method is based on the assumption of a maximum travel time (e.g., 2 hours) to the airport. It is relevant to mention that these simplified approaches have certain disadvantages such as the assumption that the catchment areas and offered transport services are similar for all destinations from the airport, which is unrealistic (Lieshout 2012).

Embeddings

An embedding is a numerical representation of an object, text, image, or other data. In the context of this thesis, embeddings are a lower-dimensional representation of the higher-dimensional input features (tabular data).

3. Theoretical Background

The theoretical foundation for clustering aviation markets includes the principles of market segmentation, the data mining process based on the Cross Industry Standard Process for Data Mining (CRISP-DM, Chapman et al. 2000), and clustering techniques to identify groups of similar data points (markets) for strategic decision-making in the aviation industry.

3.1. Market Segmentation

Market segmentation is the process of creating smaller markets with similar characteristics based on a broad and heterogeneous market (Smith 1956; Leisch, Dolnicar and Grün 2018). In other words, market segmentation is a method of defining meaningful subgroups of individuals or objects by reducing the number of objects dealt with into a manageable number of slices (Teichert, Shehu and Wartburg 2008). These identified segments can be the basis for more informed strategic decisions and can help organizations design more targeted service offerings (Kotler and Keller 2016).

There are criteria that help evaluate the quality of the market segmentation approach and results:

- Measurable: The size and characteristics of the segments can be quantified and measured.
- Substantial: The subgroup is large enough to receive a targeted service.
- Accessible: The segments can be effectively targeted and served.

- Differentiable: The subgroups are conceptually different.

- Actionable: Well-designed programs can be created to attract and serve these specific segments (Kotler and Keller 2016).

One well-known framework in market segmentation is the STP model (Segmentation, Targeting, Positioning). The main purpose of this framework is to link the outcome of segmentation with strategic decision making (Kotler and Keller 2016; Leisch, Dolnicar and Grün 2018). The first step is segmentation, the process of identifying distinct groups (which is also the main focus of this thesis). The next step is targeting: The evaluation of the identified segments in terms of aspects such as their size or growth potential and the selection of target segments. In the positioning phase, an organization should implement strategies to ensure their services are distinctly perceived compared to competitors, while also aligning with the needs and characteristics of the target segment (Leisch, Dolnicar and Grün 2018).

In market segmentation theory, four classic segmentation methodologies are used to identify groups with similar characteristics, needs, or behaviors (Kotler and Keller 2016). The demographic segmentation perspective attempts to categorize individuals by age, gender, income, education, and other similar factors. Regional or spatial information such as region, country, city, or neighborhood is part of the concept of geographic segmentation (Leisch, Dolnicar and Grün 2018). In behavioral segmentation, individuals are grouped by, for example, purchase behavior, usage frequency, or brand loyalty. For example, airlines might distinguish between business travelers (frequent, often last-minute bookings) and leisure travelers (less frequent, advance purchases). The psychographic segmentation approach aims to classify by lifestyle, values, or personality traits (Kotler and Keller 2016). Although these approaches mainly focus on segmenting people (customers) into subgroups, some of them are applicable to other concepts of markets as well. For example, geographical segmentation can be applied to aviation markets (airport pairs) within different regions, such as domestic markets within the US (Cosmas et al. 2013). While the mentioned market segmentation approaches (demographic, geographic, behavioral, and psychographic) are often based on qualitatively defined rules (for example, age group between 18 and 24), there are different data-driven approaches for market segmentation as well, such as latent class analysis. Latent class analysis deals with discrete latent variables (categories) to identify subgroups in a dataset. It is often used to group people based on patterns of scores in survey questions (Weller, Bowen and Faubert 2020). Another

quantifiable market segmentation method is an unsupervised machine learning method: clustering. Cluster analysis is the task of grouping similar data points in the same group (called a cluster). More specifically, the input data are used to compute the similarity (through a distance metric) between data points (or in the case of this thesis, aviation markets) (R. Xu and Wunsch 2005; Dolnicar 2002).

Aviation Markets

The air transport field is determined by high fixed costs (aircraft acquisition, infrastructure) and relatively low marginal costs per additional passenger. These circumstances lead to natural monopolies on key routes and the development of hub-and-spoke networks. Hub-and-spoke is a system in which hubs (large, major airports) serve as central points for coordinating flights to and from spokes (smaller airports) (Doganis 2006). The deregulation and liberalization of airlines since the 1970s and the development of different revenue management techniques have shaped competition, capacity and fare structures in the current aviation industry (OECD 2010; Doganis 2006). Aviation has experienced multiple large changes in recent years, for example, the emergence of low-cost carriers or the COVID-19 pandemic. These have drastically shaped the aviation field (Zheng et al. 2025; Acar and Karabulak 2015). In such a dynamic industry, it is essential to understand the underlying market patterns and trends that are relevant for strategic decision making (Bhadra and Kee 2008; Zheng et al. 2025).

Aviation markets can be defined and segmented in multiple ways, from customer characteristics (Teichert, Shehu and Wartburg 2008) to airport catchment areas (Graham and Guyer 2000). A catchment area can be defined as the area most outbound travelers originate from and most inbound passengers travel to (Baltazar and Silva 2018). Typical market definitions in aviation are passengers (Harrison, Popovic and Kraal 2015), or airport pairs (Cosmas et al. 2013).

Passenger segmentation in aviation attempts to capture the demand side of air travel from a customer-centric perspective. It usually involves factors such as demographics, travel purpose, or price sensitivity (Teichert, Shehu and Wartburg 2008). Traditionally, airline customers were segmented based on two criteria: purpose of trip (business vs. leisure) and frequency of travel (frequent vs. non-frequent flyer). Other considered aspects are journey length (short haul vs. long haul), country or culture of origin, local vs. transfer passengers and time sensitivity (Harrison, Popovic and Kraal 2015). Fare product features such as refundable versus non-refundable tickets offer a customer-centric segmentation based on

flexibility and purchase behavior (Proussaloglou and Koppelman 1999). Markets can also be segmented by service model, distinguishing full-service carriers, and low-cost airlines (Efthymiou and Christidis 2023). Geographic region further separates intra-continental from inter-continental markets, capturing differences in regulation, demand drivers and network structures.

The segmentation of airport markets aims to identify similar markets with comparable route performance or demand and supply patterns (Cosmas et al. 2013; F. Yang and X. Yang 2017). Furthermore, airport market segmentation allows insight into the airport network for high-level strategic decisions, such as identifying market opportunities (Cosmas et al. 2013).

3.2. The Data Mining Process for Cluster Analysis

CRISP-DM is an iterative process that contains six relevant steps: project understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al. 2000). In the following section, these steps will be described with a focus on cluster analysis.

3.2.1. Project Understanding

The first phase of the CRISP-DM cycle is project understanding, also referenced as business understanding. The main idea is to understand the problem to be solved. Although this seems obvious, most data mining problems are usually not clear and straightforward (Provost and Fawcett 2013). The relevant parts of this phase include the definition of the objectives (from a business and technical perspective) and criteria for a successful project. An important aspect to mention is that poor descriptions of the project goals and success criteria can lead to large differences between the desired and actual results (Berthold et al. 2020; Provost and Fawcett 2013). Furthermore, it is essential to define and assess the necessary and available resources, especially the accessibility of data resources and the technical feasibility of the project with the available resources (Provost and Fawcett 2013). There are many relevant aspects that should be considered when

defining the project objectives from a technical perspective. Interpretability: Is it required to understand the delivered model and the detailed path to the solution? Interpreting and understanding decisions made by black-box models can be difficult (Doshi-Velez and Kim 2017). Reproducibility: If the analysis has multiple iterations, is it expected to not only obtain similar performance across the different runs but also similar models that can be directly compared? (Berthold et al. 2020) Furthermore, reproducibility is a vital scientific component since it involves the capability to run the identical code within the same environment and achieve consistent outcomes. If, for example, the results presented in research papers are not reproducible, it may undermine their scientific value (Kapoor and Narayanan 2022). Model flexibility: Should the model be able to adapt to more complicated situations? (risk of overfitting, Tan et al. 2019) Runtime: Computationally expensive methods might not be suitable for scenarios where building or applying a model has low runtime requirements (Berthold et al. 2020). Interestingness: Domain experts are often very knowledgeable about their field and it can be difficult to present them with new findings that are valuable and insightful for them (Berthold et al. 2020). An example scenario: the clustering analysis results in two clusters, where one cluster can be interpreted as business markets and the other cluster as leisure markets. The knowledge of which market is a business or a leisure market is presumably less valuable to an aviation expert than a more granular distinction between the markets.

For cluster analysis in the project understanding step, it is important to map the high-level objectives into specific clustering tasks (such as finding similar market groups), as well as specifying the evaluation and validation measures planned to be used (using internal validation indicators such as the Davies-Bouldin Index vs. comparing to ground truth labels). Interpreting and assessing the quality of clustering results can be very challenging as these aspects often highly depend on excessive domain knowledge and might not be straightforward (R. Xu and Wunsch 2005).

3.2.2. Data Understanding

The data understanding step contains the acquisition and exploration of the input data, as well as the evaluation of the data quality (Berthold et al. 2020). The most important aspect is to match the structure of the business problem with the available data. In many cases, a business problem consists of several different data mining tasks with different data

sources. The solutions of all subtasks are combined to answer the business question(s). One substantial issue in many data mining projects is that the available datasets are often collected for other purposes than the business objective or for no explicit purpose at all. Moreover, the availability of data are often related to different costs. While some datasets are publicly accessible (open source), other data sources require either a high effort for data acquisition or a high price to pay (Provost and Fawcett 2013). Other relevant questions in this phase evolve around the information contained in the data, incomplete data, and inconsistencies, as well as anomalies (outliers). Furthermore, a descriptive and visual analysis of the data distribution, statistical properties, and basic relationships between the attributes are part of the the data understanding phase. One potential outcome of this step is that the data does not suffice to solve the problem and it is necessary to review and adjust the project goals (Berthold et al. 2020).

In a clustering data mining project, outliers and anomalies identified during data understanding could influence distance calculations for the similarity measure. In addition, the distributions of the data can have an impact on feature transformations (for example, a skewed distribution could be log-transformed), and the choice of clustering algorithms to be applied (D. Xu and Tian 2015).

3.2.3. Data Preparation

After understanding the objectives of the project and the underlying data structures of the problem, the data preparation phase applies the findings from the previous steps to select and transform the raw data into features for the modeling phase (Provost and Fawcett 2013). Based on the identified data aspects from the data understanding phase, a subset of features is selected, missing values (and potentially also outliers) are removed or replaced by estimate values (such as imputing the mean) or actual (true) values from other sources, and new features are derived. The data analysis might reveal that there are redundant or irrelevant features that can be removed (for example, based on a correlation analysis). Typically, a significant portion of the project is dedicated to understanding and preparing the data (Berthold et al. 2020). In addition to feature selection and data cleaning, there are other relevant tasks involved in the data preparation process. Data must be formatted into the correct data types (for example, one hot encoding for categorical values) (Provost and Fawcett 2013). Scaling ensures that the influence on the

results of each feature is approximately the same by transforming the values to a common range. There are various scaling approaches, with min-max normalization and z-score standardization being the most popular ones (Berthold et al. 2020). Many data mining techniques such as Euclidean-based methods require scaled features because otherwise some attributes might dominate the distance calculations and lead to biased or skewed results. This is especially relevant for many clustering algorithms such as K-means. It is worth noting that selecting the scaling technique can significantly influence the results achieved (D. Xu and Tian 2015). Feature transformation methods such as logarithmic, power or reciprocal (inverse) transformations can help achieve a Gaussian-like distribution (normal distribution) for variables that have, for example, a skewed distribution (Tan et al. 2019). This is important due to the fact that numerous models operate under the assumption of a Gaussian distribution (Berthold et al. 2020).

Another key aspect is dimensionality reduction: many models run faster and are more accurate with fewer dimensions. A main issue with datasets with a large number of features is the curse of dimensionality: When adding more dimensions (features) to the dataset, the volume of the space increases exponentially and the data become sparse (Berthold et al. 2020). A simple explanation of this phenomenon is to imagine adding points first to a line (one-dimensional), then to a square (two-dimensional), and finally to a cube (three-dimensional). It becomes increasingly difficult to cover the area or space with points due to the large number of points required. The number of data points in a dataset generally does not increase when adding new attributes, which means that the data typically become more sparse with each new feature (Awan 2023). Other problems related to the curse of dimensionality are the increased computational time, the risk of overfitting (because the calculations become more complex, the model might fit to noise rather than underlying patterns) and the difference in distances between data points become less meaningful with increasing dimensions (Venkat 2018; Tan et al. 2019). Here is a concise overview of different dimensionality reduction methods:

- Principal Component Analysis (PCA): This technique creates a smaller number of features (principal components) based on linear combinations of the initial features by using a covariance matrix and the eigenvectors of the data. Thus, it preserves the variance of the original data (Pearson 1901; Hotelling 1933).

- Multi-Dimensional Scaling (MDS): MDS tries to preserve as much of the original pair-wise distances as possible by positioning the data points in the low-dimensional

space (Torgerson 1952; Kruskal 1964).

- t-Distributed Stochastic Neighbor Embedding (t-SNE): This non-linear dimensionality reduction method uses probabilities to convert high-dimensional distances into probabilities. The "t" stands for the t-distribution that is used in this technique. The similarity of the data points is measured based on the Kullback-Leibler divergence. The primary application of t-SNE is data visualization (Maaten and G. Hinton 2008). In contrast to PCA (which tries to maintain large pairwise distances to maximize variance), t-SNE tries to preserve small pairwise distances (therefore focusing on the pairwise similarities and relationships of the data points) (Nanga et al. 2021).

- Uniform Manifold Approximation and Projection (UMAP): UMAP encodes the high-dimensional data in a low-dimensional embedding by minimizing cross-entropy between fuzzy topological graph structures (McInnes, Healy and Melville 2018). UMAP is a more recent approach for non-linear dimensionality reduction and it has major advantages over t-SNE. UMAP is less computationally expensive, has higher memory efficiency, and it is better at maintaining both local and global structures in the data (Nanga et al. 2021).

- Autoencoders: Autoencoders are neural networks trained to reconstruct their input data, with common use cases in anomaly detection, denoising, and data compression. By forcing all information through a smaller “bottleneck” layer, they learn a nonlinear data representation. This latent layer then serves as a low-dimensional embedding that captures the most important features of the original data (G. E. Hinton and Salakhutdinov 2006; Svirsky and Lindenbaum 2024). The usage of autoencoder networks or similar deep learning techniques for cluster analysis is referred to as "deep clustering" (Wei et al. 2024).

Before proceeding to the modeling phase in a clustering project, it is recommended to test whether the data tend to form natural clusters. This can be achieved through the Hopkins statistic. Artificial random data points are introduced alongside the original data points in the data space, and then the Hopkins statistic value is calculated. A value below 0.5 suggests that the data are randomly distributed, which means the data are not suitable for cluster analysis. On the other hand, a value greater than 0.5 means that the data tend to form clusters (Banerjee and Dave 2004; Tan et al. 2019).

3.2.4. Modeling

Initially, the project understanding involves an exploration of potential machine learning techniques and models suitable for the project. In the modeling phase, a detailed comparison of various methodologies within a specific category suitable for the problem takes place (such as evaluating different clustering algorithms in terms of their strengths and weaknesses) (Provost and Fawcett 2013). It is critical to choose a technique that suits the data well in terms of the properties of the data. For example, scalability is important for large datasets. Another relevant aspect is that some methods can deal effectively with outliers while others are highly influenced by them (Berthold et al. 2020). Most of the time, algorithms are not implemented from scratch. Instead, optimized algorithms and tools are used, such as those available in the scikit-learn package (*scikit-learn 1.7.0 documentation* 2025). The result of this step is a model that captures patterns in the data (Provost and Fawcett 2013).

Formally, clustering groups data points based on a measure of similarity such that instances in the same cluster should be as similar as possible, while instances in different clusters should be as dissimilar as possible (D. Xu and Tian 2015). The most commonly used distance measure for clustering is the Euclidean distance. The generalization of the Euclidean distance is the Minkowski distance metric (with a parameter that can be adjusted, for Euclidean it is 2). Other distance measures are the Manhattan distance (which is known to be more robust than Euclidean, also a variant of the Minkowski distance with the parameter 1) and the Chebyshev distance (which is the maximum of all distances between two data points) (Tan et al. 2019; Provost and Fawcett 2013).

Main Types of Clustering Algorithms

One of the most well-known clustering techniques is K-means (and its other variants such as K-means++). Other prominent clustering methods are DBSCAN, hierarchical clustering (agglomerative and divisive) or distribution-based clustering techniques such as the Gaussian Mixture Model (GMM) (D. Xu and Tian 2015; H. Yang, Jiao and Pan 2021). In the following section, the main types of clustering algorithms are briefly described.

Partition-Based Clustering

The most widely used clustering algorithm K-means (and its improved version K-means ++) are partition-based clustering methods (D. Xu and Tian 2015). The main idea of

K-means is to have a pre-defined number of center points, also called centroids, that define which data points belong to a certain cluster by the distance to the centroids of all clusters. A data point belongs to the cluster with the smallest distance to its centroid. These centroids can be artificial points (for K-means, they become the mean of all assigned data points) or actual data points (as in the K-Medoids clustering algorithm) (Jain, Murty and Flynn 1999). The disadvantages of clustering methods based on partition is that they require in most cases a pre-defined number of clusters, they often get stuck in local optima and they are not suitable for non-convex data (D. Xu and Tian 2015). The correct number of clusters k can be determined using the elbow method (Thorndike 1953; Tan et al. 2019) or the gap statistic (Tibshirani, Walther and Hastie 2001). The elbow method approximates the number of clusters by running the clustering algorithm multiple times with an increasing number for k while computing a validation measure such as within cluster sum of squares (WCSS) for each iteration. The results are visualized in a plot, where the x-axis represents the number of clusters, while the y-axis corresponds to the validation measure. A good choice for k is suggested by the "elbow" point in the graph (Tan et al. 2019). The gap statistic tries to determine the right number of clusters by statistically evaluating how well-separated clusters are in contrast to random distributions (null reference distribution of the data) (Tibshirani, Walther and Hastie 2001). In contrast to the elbow method, the gap statistic provides a more objective perspective due to the comparison against random distributions (Jacob, Daniel and Aneke 2024).

Density-Based Clustering

Density-based algorithms identify clusters based on high-density areas within the data. With this approach, they can work with arbitrary shapes of clusters and are more flexible (D. Xu and Tian 2015). These clustering algorithms are more robust against outliers as data points can also be left as noise (not part of any cluster). However, if there are many areas of varying densities and high dimensions in the data, density-based clustering methods might struggle (Berthold et al. 2020). The most popular density-based algorithm is DBSCAN (Density-Based Spatial Clustering). It has two important parameters: the minimum number of neighbors (data points) and a specified radius (epsilon). If a selected data point has enough neighbors in the specified region (based on the radius), it gets assigned to the current cluster. DBSCAN does the same check for all neighbors of the selected data point until there are no more points within the radius. Then it starts a new cluster and repeats the mentioned steps until there is no data point with enough neighbors left. The remaining data points are marked as outliers (noise) (Jain, Murty and Flynn 1999; Tan et al. 2019).

Hierarchical Clustering
Hierarchical clustering techniques group objects step-by-step into a cluster hierarchy which can be visualized using a tree-like diagram called a dendrogram (D. Xu and Tian 2015). In the agglomerative hierarchical clustering approach, each data point has its own cluster in the beginning, and the two data points that are closest to each other (based on a distance metric) will be merged into a cluster. This process is repeated until all data points are in one big cluster. Another main type of hierarchical clustering is divisive or top-down which is the opposite of agglomerative hierarchical clustering: The algorithm starts with assigning all data points to a single cluster and splits the clusters step-by-step until each cluster only contains one data point. Divisive hierarchical clustering is less common and computationally more expensive than agglomerative hierarchical clustering (Berthold et al. 2020; R. Xu and Wunsch 2005). There are two important aspects of hierarchical clustering: the computation of a distance matrix (distance between each pair of clusters) and the linkage type (such as single linkage which is the shortest distance between any two data points in the clusters or the average linkage which is the average distance between all pairs of data points in the clusters) (D. Xu and Tian 2015).

Distribution-Based Clustering
The assumption that the data are made up of probabilistic distributions, such as Gaussian distributions, is the basis of distribution-based clustering techniques. Data points that are closer to the center of a certain cluster have a higher probability or confidence of belonging to that cluster whereas data points that are further away from the center have a lower confidence of belonging to that cluster (Jain, Murty and Flynn 1999; D. Xu and Tian 2015). One popular distribution-based clustering method is Gaussian Mixture Model (GMM), which uses a two-step expectation-maximization approach. Unlike hard clustering methods such as K-means, GMM performs soft clustering (one data point can belong to multiple clusters with certain probabilities). It is relevant to mention that distribution-based clustering is not recommended if the data do not have a particular underlying distribution (Tan et al. 2019).

3.2.5. Evaluation

To ensure the effectiveness and reliability of the results, it is essential to verify their validity. It is also crucial to determine whether the model meets the original project goals.

Additionally, it is relevant to consider the model's applicability in real-world situations, as it may contain simplifications or assumptions that might not (always) hold outside controlled environments. For example, achieving over 99% accuracy in fraud detection lab-tests could still result in excessive false alarms in a business setting (Provost and Fawcett 2013). There are different types of evaluation approaches: quantitative and qualitative, as well as internal and external validations (Berthold et al. 2020; Tan et al. 2019). In the case of clustering, a qualitative assessment could be the visual inspection of the clusters in a two- or three-dimensional plot. An example of a quantitative and internal validation approach is the use of scores to gain insight about the quality of the clustering, such as the calculation of the Silhouette Coefficient (D. Xu and Tian 2015). External validation can be the evaluation by a domain expert or the comparison with the labels of the ground truth classes, if available (Tan et al. 2019). Furthermore, in many cases it is important that stakeholders understand the model or at least its high-level behavior, ensuring clarity about its functionality. If improvements to the model are needed, revisiting the data preparation phase to derive new features or try different feature transformations is a reasonable approach (Provost and Fawcett 2013).

The choice of validation metrics for evaluation depends on the objective of the project and available data. Different validation metrics assess different aspects of the quality of the cluster analysis. There is a distinction between internal validation metrics (no ground truth knowledge) and external validation metrics (available ground truth knowledge) (Tan et al. 2019; D. Xu and Tian 2015). Examples for internal validation indicators are the Silhouette Coefficient, Davies-Bouldin Index, Calinski-Harabasz Index and the Dunn Index. The Silhouette Coefficient measures how similar a data point is to its own cluster compared to other clusters, with a range of -1 to 1. A score of 1 indicates that the object is very close to its assigned cluster and far away from neighboring clusters. A score close to 0 signals that the sample lies between two neighboring clusters (close to the decision boundary), while -1 indicates an assignment to the wrong cluster (Rousseeuw 1987; Berthold et al. 2020). Another clustering validation metric is the Davies-Bouldin Index: The average similarity ratio for each cluster relative to its most similar cluster is determined by considering both the inter-cluster and intra-cluster differences. Lower scores for the Davies-Bouldin Index are desirable, as they indicate better separation and compactness (Davies and Bouldin 1979; D. Xu and Tian 2015). The Calinski-Harabasz Index compares the intra-cluster variance (degree of dispersion) with the inter-cluster variance. For this score, larger values reflect better clustering results (Wang and Y. Xu 2019). The Dunn Index evaluates the clustering quality by dividing the minimum

inter-cluster distance by the maximum intra-cluster distance, higher values indicate better clustering (Berthold et al. 2020). External validation indicators are the Rand Index (RI), the Adjusted Rand Index (ARI, extension of the Rand Index), or the Fowlkes-Mallows Index. These indexes are based on ground-truth data (D. Xu and Tian 2015). This thesis focuses on clustering without ground-truth knowledge.

Interpreting clusters means translating the cluster labels into groups and explanations that make sense for decision makers. One method is post hoc explanation, where, for example, a simple classifier (e.g., a decision tree or a random forest) is trained on the cluster labels after the clustering. The feature importance values reveal which features have a strong influence on the segmentation of the data points (H. Yang, Jiao and Pan 2021; Hu et al. 2024). Another approach is to inspect representative data points, such as centroids or medoids, to read the typical values for each cluster and relate them to known groups in the domain (Park and Jun 2009). Analyzing the balance and stability of cluster sizes can also reveal relevant insights (D. Xu and Tian 2015).

3.2.6. Deployment

Once technical benchmarks or results can no longer be improved, the deployment phase begins. However, if the outcomes are not sufficient to solve the problem, the project may be concluded without deploying the system in a practical setting (Provost and Fawcett 2013; Berthold et al. 2020). The process often goes back to the problem understanding phase, potentially to refine the problem definition, modify or expand data sources, or to start a new iteration for a better solution. Typically, the initial version serves as a working prototype. During the deployment phase, this prototype is integrated into an existing system and enhanced for performance (Provost and Fawcett 2013). For models put into regular application, it is important to periodically review the assumptions they rely on and verify whether they remain valid or if the underlying data or business problem has changed (Berthold et al. 2020). Therefore, it is also relevant to monitor model results over time using validation metrics (for example, cluster stability indices or other validation metrics) (Gama et al. 2014). Version control, reproducible machine learning pipelines, documentation, and stakeholder training are other aspects that are relevant in the deployment phase of a data mining project (Provost and Fawcett 2013).

4. Literature Review

Several studies have explored clustering techniques and network analysis in the aviation industry (Güvercin, Ferhatosmanoglu and Gedik 2021; Prabhakar and Anbarasi 2021; F. Yang and X. Yang 2017). Some papers concentrate on clustering airports (Gao 2021; Postorino and Versaci 2014) or segmenting passengers rather than airport markets (Üstebey, Yelmen and Zontul 2020; Robles and Sarathy 1986; Maliha and Hougen 2024).

Although clustering and segmentation studies help to understand high-level groupings of markets and passengers, they often treat airports as origins and destinations in the journey of travelers. The distinction between "Airport-to-Airport" and "Door-to-Door" modeling was introduced by J. Mandel and B. Mandel 2021. “Airport-to-Airport” is a simplification and describes the passenger flow from and to airports whereas “Door-to-Door” aims to integrate more information about the mobility needs of passengers (e.g. where passengers actually come from and where they want to go, not just the airports themselves). Their research suggests that traditional models with a focus on direct airport connectivity might miss passenger demand patterns and that the integration of additional information other than airline and airport-related data leads to better results in identifying market opportunities.

4.1. Clustering Airport Markets

There are two papers that are relatively similar to this thesis in their approach to cluster airport markets. One paper focuses on global aviation market clustering based on network performance indicators (F. Yang and X. Yang 2017) and the other paper combines clustering with financial portfolio theory to support strategic planning by evaluating airport markets as financial investments (Cosmas et al. 2013).

F. Yang and X. Yang 2017 apply hierarchical clustering to create global aviation market clusters based on airline route network performance. The objective of their study is to evaluate the competition across different markets and also to run different fare strategies. They define 11 network performance indicators (one example of such an indicator is RPK which means revenue per passenger kilometer). The indicators are reduced to four main components through Principal Component Analysis: market comprehensive factor (including most of the variance), market network traffic flow coverage factor (combining passenger flow, capacity and other factors), market potential factor (including average connect time) and market network passenger load factor (including available seat kilometers, ASK). The authors then combine the principal components into one composite score that serves as the basis for the clustering. Markets are grouped at different levels: intercontinental (for example, North America to Europe), national (country level), and on airline level. One interesting finding is that most of the national markets are in the same cluster as their return markets except for the national pair "Italy" and "China" ("CNIT" and "ITCN" are not in the same cluster). The objective of this paper is clustering markets based on performance indicators rather than market segmentation, however, the approach could be extended to include other aspects relevant for market segmentation as well.

In the paper of Cosmas et al. 2013, clustering techniques and financial portfolio management are combined and applied to the U.S. airline industry. They define a market as an origin-destination airport pair, also referred to as a city-pair route. Their focus is to segment these OD markets into statistically similar peer groups using cluster analysis for strategic network planning purposes. For example, to identify which markets have high growth and low volatility. The three key features used are passenger volume, average annual yield (revenue per passenger mile), and market concentration, measured using the Herfindahl-Hirschman Index (HHI). For the clustering step, a two-step algorithm is used. First, pre-clusters are generated through minimizing the average log-likelihood distance between clusters. In the second step, a hierarchical clustering algorithm is applied to the pre-clusters. One major aspect is the use of financial portfolio theory to include risk factors in strategic decisions. As an example, the authors use the Sharpe Ratio to compare the expected returns of the OD markets considering the risk associated with such investments. Seven distinct market clusters are identified based on 10,000 analyzed US domestic OD markets between 2000 and 2010. In their analysis, one finding is that low-yield markets have, on average, performed better than the industry in passenger and revenue growth, whereas high-yield markets have underperformed on average. It is highlighted that current forecasting and market share models are often too generalized

and insensitive to distinct characteristics of different markets. By applying OD market clustering, it becomes possible to balance the need for generalizable models and more realistic market representations. This paper focuses on strategic network planning and financial aspects, but does not take into account demographic factors related to the broader context of the markets, such as characteristics of the surrounding regions or countries.

Although the focus lies on origin-destination flows in urban transport rather than on airport markets, the study by Fang et al. 2021 can also be applied to aviation flows. The paper demonstrates how clusters of OD flows can be detected at multiple spatial scales (different levels of aggregation, meaning different within neighborhood distances). The authors use the OPTICS clustering algorithm in their analysis. Point of origin information (where passengers initially come from) is a valuable aspect for analyzing aviation markets as it can enable precise market-level analysis (Berry and Jia 2010; Cristea, Hummels and Roberson 2017).

4.2. Airport Clustering and Airport Network Analysis

The study from Gao 2021 explores airport clustering (of US hub airports) based on passenger movements. Their results imply that the location of an airport is relevant in determining its flight schedules (considering the arrivals of passengers at security checkpoints). Postorino and Versaci 2014 focus on airport clustering as well, using a fuzzy-based approach. One of the chosen criteria for the clustering are also passenger movements. Their proposed method successfully identifies airport clusters with similar functional roles (applied to the Italian airport network).

Prabhakar and Anbarasi 2021 analyze air transport networks using social network analysis (SNA). In their paper, they highlight the importance of connectivity and centrality characteristics in air transport. The objective is to create a better understanding of the global aviation network and to identify airports and countries that play a major role in the network. For example, Hartsfield-Jackson International airport in Atlanta has been ranked as the busiest airport worldwide because of its high degree centrality, with Paris Charles de Gaulle airport coming closely after Atlanta. The paper by Güvercin, Ferhatosmanoglu and Gedik 2021 also uses network-based features such as betweenness

centrality and flight connectivity for clustering airports in order to predict flight delays more accurately. They combine clustering techniques with time-series forecasting and were able to achieve a higher accuracy for delay prediction than traditional individual airport-based models.

4.3. Passenger Segmentation

Teichert, Shehu and Wartburg 2008 claim that the traditional approach to differentiate travelers between business and economy passengers is not suitable to reflect more complex customer preferences. The authors propose an approach to use latent class modeling combined with behavioral and socio-demographic variables to find more granular and refined passenger segments. Data on stated preferences by more than 5,800 airline passengers were used for the analysis. Based on the profiles derived, three potential areas for service specialization for airlines are revealed: comfort, efficiency, and price-oriented offers. In the study of Wu and Berger 2018, IBM Watson survey data (more than 127.000 passenger records) are used to cluster airline customers based on demographics, flight as well as airport experience using K-means. The scope of the paper is consumer-centric and provides insights into passenger behavior (business vs. non-business traveler clusters). Zhou et al. 2020 also use survey data in their analysis. The authors aim to identify existing and potential passenger segments across airport and non-airport samples from there preference survey in Western Australia about transportation modal choices. A mixture-model clustering approach (Expectation Maximization) is used for the segmentation process.

Yelmen, Üstebey and Zontul 2020 apply a self-organizing map (SOM) to airline ticket sales data. Self-organizing maps are used for dimensionality reduction by producing a low-dimensional representation of the data. SOMs are a specific type of neural networks. The authors cluster passengers by purchase and loyalty behavior as well as customer return. The study illustrates how SOMs can help uncover nuanced passenger profiles (for example, high return in contrast to low return passengers).

Han et al. 2022 introduce a variation of the K-means algorithm (Han-Kmeans), which extends the traditional K-means by selecting initial centroids based on local density. The goal of their analysis is to increase customer satisfaction and customer loyalty by

improving passenger segmentation. When applied to Chinese airline passenger records with a time period of two-months, Han-Kmeans achieved better results than the standard K-means and yielded four actionable passenger segments for targeted marketing strategies.

Time-series information plays a vital role in the analysis of aviation data, as it reflects important temporal patterns such as seasonality in tourism. The very recent study by Zheng et al. 2025 builds time series features based on local share. Passengers can fly directly from their origin to their destination (local passengers) or need to transfer to intermediate airports (transfer passengers) (Maertens, Grimme and Bingemer 2020). Local share is defined as the proportion of local passengers relative to total passengers (Zheng et al. 2025). The analysis aims to reveal how local share has changed over time through economic expansion, the rise of low-cost carriers, and other shifts in the industry by analyzing local share patterns of thousands of origin and destination pairs in the US air transport system. From a technical perspective, the authors apply K-means to extract trend and seasonality features and use hierarchical clustering on the normalized raw time series data.

Clustering methods and other market (or passenger) segmentation techniques have been used for a variety of application areas in aviation, ranging from airline network performance assessment (F. Yang and X. Yang 2017), clustering of airports (Gao 2021; Postorino and Versaci 2014), or segmentation of aviation customers (Teichert, Shehu and Wartburg 2008; Yelmen, Üstebey and Zontul 2020). This thesis builds on existing literature and aims to expand the research by the application of clustering techniques to airport markets while incorporating a comprehensive dataset that includes economic, demographic, and seasonal factors. Moreover, the analysis has a global perspective, including various regions across the world.

5. Methodology

In this thesis, cluster analysis is used to create groups of similar aviation markets (airport pairs) with the potential application of market opportunity identification. The implementation from a technical standpoint uses Python alongside various libraries for data processing, machine learning, and visualization, such as numpy, pandas, scikit-learn, matplotlib, and seaborn. The thesis follows the CRISP-DM process including the project understanding, data understanding, data preparation, modeling, and evaluation phases (Chapman et al. 2000). The last phase (deployment) is not part of this thesis.

5.1. Research Questions

This thesis aims to answer the following key questions:

- Primary objective: **Which clustering technique(s) are effective for segmenting aviation markets?**
  Evaluation of the clustering can be achieved with validation metrics such as Silhouette Coefficient.

- Secondary objective: **Can markets with growth potential be identified through clustering results?**

The secondary objective, which involves a comprehensive validation and analysis of potential growth markets, falls outside the scope of this work. This is due to the necessity of actual demand data to compare against the supply-side features (capacity). Unfortunately, demand data were not available for this thesis. Therefore, the focus shifted entirely to creating meaningful clustering results based on internal validation metrics. In

practice, airlines could combine their demand information with the dataset and clustering results of this thesis to detect opportunities for new routes or increased flight frequencies.

Literature Review
A targeted literature review on papers related to clustering and market segmentation in aviation was conducted. References were traced to identify studies that apply clustering or other segmentation techniques for airports, passengers, and airport markets. This process revealed two key studies that also focus on clustering airport pairs, several studies that apply cluster analysis to airports, and different approaches to segment passengers in aviation. This review established the state-of-the-art regarding clustering and passenger segmentation in aviation, highlighted typical data sources and features, and informed potential choices for data transformation methods and clustering algorithms.

Data Sources and Preparation
Multiple data sources (flight schedule, tourism-related information, climate data) are integrated into a single dataset (6.1). These heterogeneous datasets are merged through standardized airport identifiers (three letter airport codes) and geographic joins (for example, population through two-letter country codes), producing a feature matrix in which each row represents an airport-pair market.

The handling of missing values and outliers is approached with care, as the global analysis requires the inclusion of as many airports as possible. Fortunately, the majority of data sources were fairly complete for the airports listed in the flight schedule. Consequently, only a minimal number of airports is excluded. Plausibility checks are performed for extreme values and numerical ranges. Following an extensive data exploration phase, the appropriate features are chosen and transformed.

Modeling
Various experiments are executed using different clustering algorithms to determine effective methods for creating meaningful groups of airport markets. The evaluation of the clustering outcomes includes a comparison of internal validation metrics along with a visual inspection in reduced two-dimensional representations. In addition, different interpretation approaches are described.

Potential Outcomes and Value of the Results

While a detailed interpretation of cluster characteristics is left to domain experts, this is an outline how different aviation stakeholders can benefit from the clustering results.

Airlines can group their existing and potential airport pairs by clusters and compare performance metrics such as load factors between clusters. Market opportunities can be identified through a comparison of the supply and demand patterns of the different clusters and markets.

Airports can make operational decisions based on cluster assignments, such as resource allocation for seasonal and tourist-heavy markets. Infrastructure investments can be derived by benchmarking clusters or markets with similar characteristics, for example, prioritizing runway expansion. These use cases could also be improved through demand data or actual passenger movements.

Tourism boards and governments can target promotional campaigns by identifying under-served regions with high clustering similarity to established tourist hubs. Focusing on regional pairs instead of airport pairs presents an interesting direction for this use case.

By conducting a comprehensive analysis by domain experts, the existence of the identified clusters can be potentially validated. Consequently, these clusters could be used as probabilistic classes that enable supervised learning applications such as market classification.

6. Data Collection, Exploration and Preparation

According to the CRISP-DM framework (Chapman et al. 2000), the stages of data preparation, modeling, and evaluation often involve a significant amount of iteration. This iterative nature is also a key aspect of this thesis. Therefore, the first part of this chapter describes the multiple data sources, followed by an initial feature list. Next, various data exploration steps are described. Finally, the refined features obtained after extensive iterative feature engineering and experimentation are outlined alongside different applied data transformation techniques.

6.1. Data Sources

Various characteristics of airports and markets were selected to represent airport markets. The data gathering process took into account aspects of data availability and quality.

Flight Schedule
This flight schedule contains flights for the entire year 2024 on a daily basis as well as the operating airline, origin, destination, departure, and arrival times and the available seats in the aircraft (provided by ASG). The total number of records (direct flights) in the schedule is approximately 36 million.

Many relevant features can be derived from the flight schedule such as the capacity (supply, or sum of seats) between two airports, or the number of flights in the morning hours. Average fares in a market would be valuable for this analysis, however, the data are unfortunately not available (on a global scale and in an appropriate level of detail).

Airport Information

Airport-related characteristics are, for example, the number of airlines operating at the airport, the number of runways (reflecting the size of the airport), or the elevation of the airport. There are two data sources with airport information used in this thesis. An airport list containing data about coordinates and geographical information (country, region) is provided by ASG, the other list is obtained from "OurAirports" as well as the runway information about airports. OurAirports is an initiative in which members maintain global aviation data records, completely in the Public Domain (*OurAirports* 2025). The airports list from ASG has roughly 12,000 entries, whereas the list retrieved from OurAirports has around 80,000 airport entries (including military and helicopter airports as well). The runway dataset from OurAirports has roughly 47,000 entries. These lists are filtered by the actual airports present in the flight schedule (number of unique airports in the flight schedule 2024: 3,991). This analysis focuses only on existing markets based on the flight schedule, resulting in approximately 60,000 directional markets and 30,000 non-directional markets (this means that, for example, the airports VIE-FRA and FRA-VIE are consolidated into one row FRA-VIE).

Business

Passengers and markets are often segmented in business and leisure (Leisch, Dolnicar and Grün 2018). The business interest in the area around an airport can be represented by the number of companies within a certain radius around the airport (Baltazar and Silva 2018). Although there are a few datasets available with company information such as industry and company size, these datasets can not be properly matched with airport data as geographical information is provided through region codes and countries (not coordinates or addresses).

Tourism

To reflect tourist interest in the area around an airport, data such as the number of hotels or hotel rooms within a certain radius and the proximity to points of interest (for example, beaches or monuments) is explored. The OpenStreetMap API (Application Programming Interface) provides information on both hotels and points of interest within a certain area (*OpenStreetMap* 2025). Based on different query settings and sample runs, only the number of hotels are included in the analysis. The inclusion of the number of hotel rooms would be suitable for the analysis (more granular), however, this information is not available for many hotels (for example, only 10% of the hotels around the Vienna Airport within a 10 km radius have room information in the OpenStreetMap data, for

John F. Kennedy International Airport it is approximately 4%). The imputation would skew the actual hotel room counts too much. Obtaining points of interest data for all global regions would be out of scope because of the querying and data processing duration. Even acquiring hotel information alone requires several hours for areas with a radius of 50 km. Gathering details on points of interest would involve numerous searches across various categories (not only hotels) and also the need to expand the radius, consequently extending the query and data processing duration even further.

The number of hotels within a radius of 10 km: Lee and Jang 2011 describe that mainly business travelers stay at airport hotels or hotels close to the airport. Furthermore, the data availability of hotels for many airports with a lower radius than 10 km was scarce, therefore a low radius with sufficient data (10 km) is selected.
The number of hotels within a radius of 50 km: The area of 50 km around the airport can be interpreted as its catchment area (Baltazar and Silva 2018). Therefore, the number of hotels within a radius of 50 km is suitable to represent the tourist interest in the area. A large number of hotels in the region around the airport can suggest that there are touristic destinations close to the airport.

Demographic
Information such as the population of the country of the airport or the closest city to the airport, the gross domestic product (GDP) of the country are taken into account. Given that certain countries have more than hundreds of large and publicly accessible airports (*National Transportation Statistics (NTS)* 2019), it would be better to refer to the population of the city closest to the airport. Unfortunately, the definition of 'city close to the airport' is unclear, and obtaining such a data set is beyond the scope of this analysis. In the feature list, the GDP per capita is included at the country level. This feature has the same disadvantages as the population (large countries tend to have many airports). Unfortunately, a more detailed GDP dataset (per region or also per industry) is not available on a global scale. Population and GDP per capita information is collected from *World Bank Group - International Development, Poverty and Sustainability* 2025.

Climate
Data sources with various climate-related information, such as humidity, rainfall, temperature, or wind speed, are explored. For simplicity, the focus within this analysis lies on temperature information only. Monthly average temperatures from weather stations around the world are collected from TerraClimate 2025 (Public Domain). The temperature

datasets are divided into yearly minimum and maximum temperatures. Each dataset contains around 450 million entries (daily temperature records for the weather stations). Based on the minimum and maximum temperatures, the mean temperature is computed. These monthly temperatures should represent climate information and, to some extent, seasonality.

6.2. Data Exploration of Initial Features

The initial list of variables included in the analysis is derived from the different data sources explored (6.1). The number of features (for the initial feature list before feature engineering) is 92: 53 capacity features (annual and weekly), 24 temperature features, and all features related to the airport twice for origin and destination.

To get an overview of the data properties such as distribution, minimum and maximum values, a YData profiling report is created (*YData data quality for Data Science* 2025). This report includes summary statistics, missing values, the number of zero values, extreme values, and helpful visualizations such as histograms. Some exemplary features are analyzed in more detail in the following paragraphs. More information on the other features can be found in the YData profiling report (A.1.1). A plausibility analysis of extreme values is performed to find erroneous outliers. A general rule is to keep outliers if they are actual values as they could be relevant for further analysis.

Initial Features

Feature	Data Type	Description
Annual capacity	Integer	The sum of all seats of all direct flights (for the year 2024) based on actual flown data (flight schedule provided by ASG).
Seasonal capacity	Integer	The sum of all seats of all direct flights in a certain time period (weekly, for the year 2024) based on actual flown data, represented by 52 features (weekly sum of seats of direct flights between two airports).
Number of airlines	Integer	The number of airlines operating (departing and arriving) at an airport based on actual flown data.
Distance	Float	The distance (in meters) between two airports in a market based on a list of airports with latitude and longitude.
Elevation	Float	The elevation (in meters) of the airport based on OurAirports data (missing values filled via Google Elevation API).
Number of runways	Integer	The number of runways of the airport based on OurAirports data.
Population	Integer	The population of the airport country (year 2023) based on World Bank data.
GDP per capita	Float	The gross domestic product per capita of the airport country (year 2023) based on World Bank data.
Number of hotels within 10 km	Integer	The count of hotels within a 10 km radius around the airport (from OpenStreetMap API).
Number of hotels within 50 km	Integer	The count of hotels within a 50 km radius around the airport (from OpenStreetMap API).
Monthly average temperature	Float	The average temperature per month at the weather station closest to each airport according to TerraClimate data, 24 features (12 months per airport in a market).

Table 6.1.: Summary of initial features, their data types, and descriptions.

Origin and Destination Country

Although the country information of the origin and destination airports are not part of the clustering process, they are necessary to match, for example, population data.

Furthermore, it is interesting to analyze how often certain countries appear in the flight schedule. The figure 6.1 represents how many direct flights have a certain country as their origin. Based on the word cloud visualizations, the countries that occur most frequent as origin country are the United States (15.7%, meaning almost 10,000 markets originate from a US-based airport), China (13.7%), Russia (4%), and Great Britain (3.7%). These numbers are an indication for the relevance of certain countries in the global aviation network. It is worth mentioning that five of the ten most frequent countries are in Europe. In total, 16.7% of the markets in the flight schedule originate in Europe, compared to 15.7% in the United States. In addition, more than half of the markets originate from the top ten countries in this list. The values for the destination countries are very similar to those of the origin countries.

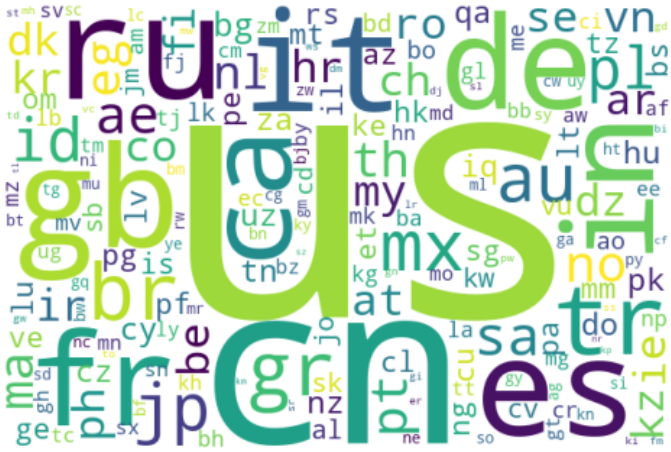

Figure 6.1.: Word cloud visualization of the countries of the origin airport

Capacity Data

The annual capacity feature represents the size of an airport market, by summing up all available seats for all flights between the two airports. The market with the highest number of seats on a yearly basis is Jeju International to Seoul Gimpo (CJU-GMP) with approximately 7.5 million seats (in one direction). This is plausible since OAG (Official Aviation Guide, one of the largest aviation data providers) published a capacity of 14 million seats (non-directional) (*Busiest Flight Routes in the World 2024* 2025).

The market with the smallest number of seats is between two French airports (Toulouse–Blagnac, Le Bourget). This might be a data error as there is a capacity of 8 seats in week 27 but no capacity in all other weeks. Another possibility is that there is only a single flight operated with a small aircraft in 2024. Based on Grant 2024, the average number of seats

per flight was 160 in 2024. If there is on average only one flight per month with 160 seats, the annual capacity in this scenario would be 1,920. A further analysis of annual capacity values revealed that approximately 6,000 directional markets have a capacity of less than 1,920 seats (which is approximately 10% of the directional markets). In conducting a global analysis of airport markets, it is determined that airport pairs with very limited capacity are insignificant and are therefore excluded (removing airport pairs below 1,920 seats per year). The histogram of the distribution of the annual capacity values (6.2) shows that most of the markets have an annual capacity below 200,000 seats. 3,000 markets (approximately 5%) have a capacity of more than 400,000 seats per year, and about 500 markets (less than 1%) have an annual capacity of more than 1 million seats.

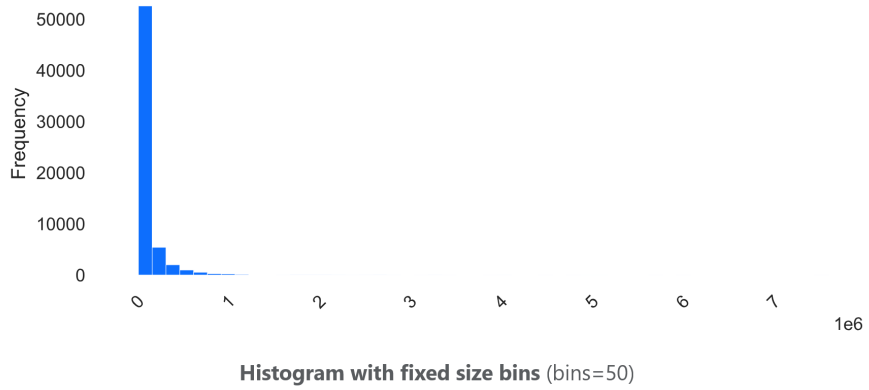

Figure 6.2.: Histogram of annual capacity, data distribution of values based on YData profiling report

The weekly capacity features all follow a similar data distribution as the annual capacity (right-skewed with many values in the lower capacity range). The capacities for each week are derived from the same data source as the annual capacity for the entire year (the flight schedule provided). The minimum of 0 is plausible, since some markets are seasonal and have no flights in some weeks of the year. The sum of the capacity of all weeks is the annual capacity for the entire year 2024 (plausibility check). Among the 52 weekly capacity attributes, the 95th percentile predominantly falls between 7,000 and 10,000 seats per week. This indicates that 95% of the markets have a weekly capacity of up to 10,000 seats. The market with the largest weekly capacity (185,000) in the first week of the year is the same one that has the highest annual capacity, namely CJU-GMP.

Elevation, Number of Airlines, Distance, and Number of Runways

Elevation may be negative for some airports, as the lowest point on the is the Dead Sea, which is approximately 420 meters below sea level (Geographic 2025). The highest airport is around 4,400 meters above sea level (*Daocheng Yading Airport* 2025). Elevation may also be inaccurate for some airports (since a small error in GPS coordinates can lead to a large error in elevation). However, this should not be a big issue for the analysis, since a plausibility check is integrated in the data cleaning step. For simplicity, all locations below 420 meters or above 5,000 meters sea level are removed.

The number of airlines operating at an airport is between 1 and 140 which seems plausible at first glance, although 140 airlines is potentially a bit high. There are two airports with 140 airlines (Rome Fiumicino "Leonardo da Vinci" Airport - FCO, and Paris Charles de Gaulle Airport - CDG), the official FCO airport websites state that 75 airlines operate at the airport (*Airlines - AEROPORTI DI ROMA* 2025). However, the 140 airlines in the flight schedule for FCO are unique. Potentially, there are some airlines that do not fly frequently to the airport and are therefore not counted in the official numbers. For CDG, the Wikipedia site states that there are 105 airlines operating at that airport (*Charles de Gaulle Airport* 2025). Therefore, the estimated range for the number of airlines at each airport appears somewhat reasonable, even though it might not be completely accurate, which is acceptable for this analysis.

The distance is calculated using the airport coordinates and the great-circle distance function from OSMnx (version 2.0.4, Boeing 2025). The maximum distance is 17 million meters (17,000 km) between two airports: Salvador International Airport in Brazil (SSA) and Mitchell River Airport in Australia (MXQ) which is plausible. The minimum value is 300 meters between two airports: Suavanao Airport (VAO), and Kaghau Airport (KGE) (which are on two different small islands in the Pacific Ocean). This distance is too small because according to *Great Circle Mapper* 2025 these airports are 130 km apart (which is still very close). The assumption is that the coordinates are inaccurate and therefore the resulting distance is too small, which is acceptable for the purposes of the analysis). Airports in Alaska come next with slightly larger distances between them, measuring a few thousand kilometers, which seems reasonable.

The range of number of runways is plausible (between 1 and 8). The airport with the highest number of runways is Chicago O'Hare International Airport (ORD) with 8 runways. More than 40% of the airports have one runway, and more than 30% have two.

Population and GDP per Capita
The country with the smallest population is Tuvalu (an island state in the Pacific) with approximately 10,000 inhabitants in 2023. The country with the highest population is India, with approximately 1.4 million people in 2023. GDP per capita in the year 2023 is plausible, the values lie between 193 USD (Burundi) and 128,000 USD (Luxembourg).

Number of Hotels
The range for the number of hotels around the airports (radius 10 km and 50 km) seems to be plausible. It lies between 0 and 2150 (10 km), and 0 and 3500 (50 km). The airport with the highest number of hotels within a 10 km radius is Denpasar Airport in Bali (DPS) with more than 2,000 hotels. Within 50 km radius, the highest number of hotels is approximately 3,500 and these are also located around DPS airport based on OpenStreetMap data. The next airport is Paris Charles de Gaulle Airport (CDG) with almost 3,000 airports within a 50 km radius. Although these numbers seem reasonable, the data quality might not be accurate (difficult to verify). Furthermore, the assumption with many hotels within a small radius might indicate a business focus and many hotels within a larger radius might indicate touristic focus, might not hold.

Temperature
The temperature related features (monthly average temperatures) seem to be within a reasonable range (the lowest average temperature is -49°C (near Aeroport Pos. Ust'-Nera in Russia - USR, the minimum and maximum temperature are between -42°C and -48°C based on data from NOAA, *National Centers for Environmental Information (NCEI)* 2025). This range seems plausible, potentially slightly too low. The highest average temperature is 41°C (near Iranshahr Airport in Iran - IHR). The minimum and maximum temperature 31°C and 44°C based on NOAA data. In this case, the range also seems plausible, potentially slightly too high.

6.3. Missing Values

The basis for identifying missing values is the number of unique airports in the flight schedule, which is 3,991 (as this is also the filtering criterion for airports). There are two airports missing from the airports list provided by ASG that are present in the flight schedule (these two are manually added). 120 airports do not have elevation information,

the missing values are retrieved using the airport coordinates and the Google Elevation API (Google 2025). Some airport coordinates are not completely accurate because the resulting elevation values are far below sea level due to the coordinates pointing to a location in an ocean for some airports). These are manually corrected. Joining the airports with the runway dataset requires an "id" (airport identifier column of the OurAirports dataset), 11 airports do not have the necessary identifier and are filtered out. Furthermore, 228 of the remaining airports have missing runway information and are also removed from the airport list for simplification purposes. It is worth mentioning that missing runway information can be imputed with the mean or also using K-Nearest Neighbor (KNN). 3,752 airports are remaining after merging with the runway dataset. 6 of them have missing country codes that are manually added. All countries have population data for the year 2023, however, not all countries have GDP per capita information for the year 2023. Therefore, the years 2019-2022 are used to fill in the missing values (starting with the most recent year). The merging process with the airports list reveals that 43 of the airports have no population information because they can not be matched with a country and are mostly islands such as the Christmas Island. These entries are removed. After joining population and GDP per capita data, 3,709 airports remain. No airports have to be removed in the merging process for hotel and temperature data, which means out of the initial 3,991 airports, 3,709 airports can be used for the analysis (removing approximately 7% of the records).

6.4. Feature Selection

The first approach is to use all initial features and only apply scaling as transformation step (z-score standardization) and no other transformations. This type of feature set is referenced as “temp capacity wide” because there are 24 temperature columns (mean temperature for both origin and destination airport on a monthly basis) and 52 capacity columns (weekly sum of seats of all flights between the origin and destination). An overview of the initial experimental clustering results can be found in chapter 8.

The second version of the feature set uses statistically derived features for temperature and capacity. For example, instead of using one capacity column for each week of the year, the mean capacity, standard deviation of the capacity, minimum and maximum capacity and the week of the minimum and maximum capacity are created. The adjusted

features for the monthly mean temperature are also the mean, standard deviation, summer mean temperature (June-August), winter mean temperature (December-February), and the annual amplitude (the difference between the warmest and coldest month). One main reason for this decision is that the number of features representing capacity and temperature in the initial feature list is too high. With so many features for temperature and capacity, they are given far more importance than the other characteristics of markets. Additionally, the correlation analysis reveals strong correlations for temperature and capacity features. These features can result in less effective clustering outcomes, since redundant or highly correlated features do not contribute meaningful information to the clustering process but are included in the calculation of distances (Sambandam 2003).

The next analysis step includes applying block-wise PCA on temperature and capacity features. The idea behind it is to extract the underlying patterns and seasonality and create principal components that represent most of the information (keeping 90% of the variance of the initial 24 temperature and 52 capacity features). For the capacity features, this results in only one principal component. This can indicate that most markets have similar seasonal profiles (with potentially large differences in the magnitude of the capacity but not in the pattern) and that there are no complex seasonality patterns, as it would probably require more components to capture large differences. It is also important to mention that only the year 2024 is considered for the analysis (because of data availability) and that other years can have different seasonal profiles. For the temperature features, a similar picture emerges as 90% of the variance in the 24 temperature columns can be represented in two principal components. From a practical point of view, this observation mirrors seasonal temperature cycles in most places of the world (higher temperatures in summer and lower temperatures in winter) (Masson-Delmotte et al. 2021).

At a later stage, markets are combined and use characteristics related to the market itself, rather than those associated with airports, since the objective of this thesis is to cluster the airport pairs and not the airports. This means that the directionality is removed and markets become non-directional. The approach from airport-related to market-related features is to compute the absolute difference between all airport-related features, for example, the difference in average temperature or the difference in elevation of the two airports. The following example illustrates why the absolute difference is used instead of maintaining positive and negative values: the population of the United States is roughly 340 million and for Germany 80 million. If the "normal" difference for FRA-JFK is computed, the result is 80-340=-260. The same calculation for JFK-MUC

would be 340-80=260, the absolute difference in both cases is 260. Putting this into, e.g., the Manhattan distance formula (7.1), the calculation for the "normal" difference for the part in the formula with the population feature would be: |-260-260| = 520, for absolute difference the result is: |260-260| = 0. Since the markets are non-directional (meaning JFK-MUC and MUC-JFK are consolidated because there is not enough data to properly represent directionality, the absolute difference is the better approach for non-directional markets). In the example case, both markets are between Germany and the United States, so the difference in population should be the same for FRA-JFK and JFK-MUC regardless of the ordering of the airports in the market.

Following the analysis of different initial clustering results and validation metrics such as the Silhouette Coefficient, other flight schedule-based features (in addition to the existing capacity features) are created. The weekly capacity features and also the statistically derived capacity features (such as mean and standard deviation) are able to represent seasonality patterns to some extent, however, they are not able to represent all important characteristics of markets. The time of the day and the days within a week on which flights depart are not covered. Furthermore, some markets do not have flights during certain periods of time (high seasonality). Therefore, the morning and evening flights ratio (morning or evening flights divided by total flights), difference of the mean capacity on a weekday and on a weekend day, as well as the number of weeks with zero capacity are extracted from the schedule.

After the data exploration and feature engineering phase and some clustering iterations (refer to chapter 8 for further information), a final set of features is determined for further analysis and experimentation.

6.5. Data Exploration of Refined Features

The refined features are based on non-directional markets and the list involves a set of 18 features (6.2, 6.3). The focus lies on attributes extracted from the flight schedule (8 features). Compared to the initial feature list (6.1), which includes 52 capacity and 24 temperature features, this collection offers a more balanced distribution of features per category.

Refined Features (Part 1)

Feature	Data Type	Description
Difference mean weekday vs. weekend capacity	Float	Difference between the mean sum of seats on a weekday and the mean sum of seats on a weekend day.
Number of weeks with zero capacity	Integer	Number of calendar weeks with no flights between the two airports.
Mean capacity weekly	Float	Average weekly sum of seats between the two airports across the year.
Standard deviation capacity weekly	Float	Standard deviation of the weekly sum of seats between the two airports across the year.
Week of max capacity	Integer	Calendar week number in which the highest weekly sum of seats occurs.
Morning flights ratio	Float	Ratio of flights departing between 06:00 and 10:00 to total flights (peak morning hours).
Evening flights ratio	Float	Ratio of flights departing between 17:00 and 21:00 to total flights (peak evening hours).
Number of airlines in the market	Integer	Count of unique airlines operating flights between the two airports.
Distance	Float	Distance in meters between the two airports based on great-circle distance.
Difference of summer mean temperature	Float	Absolute difference between mean June–August temperatures at the closest weather stations.
Difference of winter mean temperature	Float	Absolute difference between mean December–February temperatures at the closest weather stations.
Difference of annual amplitude	Float	Absolute difference between the annual temperature amplitudes (warmest minus coldest month) of the two airports.

Table 6.2.: Summary of refined features, their data types, and descriptions (first part, continues on next page).

Refined Features (Part 2)

Feature	Data Type	Description
Difference of elevation	Float	Absolute difference in elevation (meters) between the two airports.
Difference of number of runways	Integer	Absolute difference in runway counts between the two airports.
Difference of number of hotels within 10 km	Integer	Absolute difference in counts of hotels within a 10 km radius of each airport.
Difference of number of hotels within 50 km	Integer	Absolute difference in counts of hotels within a 50 km radius of each airport.
Difference of population	Integer	Absolute difference between the 2023 population values of the airport countries in the market.
Difference of GDP per capita	Float	Absolute difference between 2023 GDP per capita of the airport countries in the market.

Table 6.3.: Summary of refined features, their data types, and descriptions (second part).

The analysis of the data distributions of the features revealed that most of the features are right-skewed. As examples, kernel density estimate (KDE) plots of the mean capacity (6.3) and the distance (6.4) demonstrate the skewed distributions. As a result, numerous markets typically have a relatively small capacity and are located at shorter distances as opposed to longer ones. Generally, lower values for various characteristics tend to be more common than higher ones.

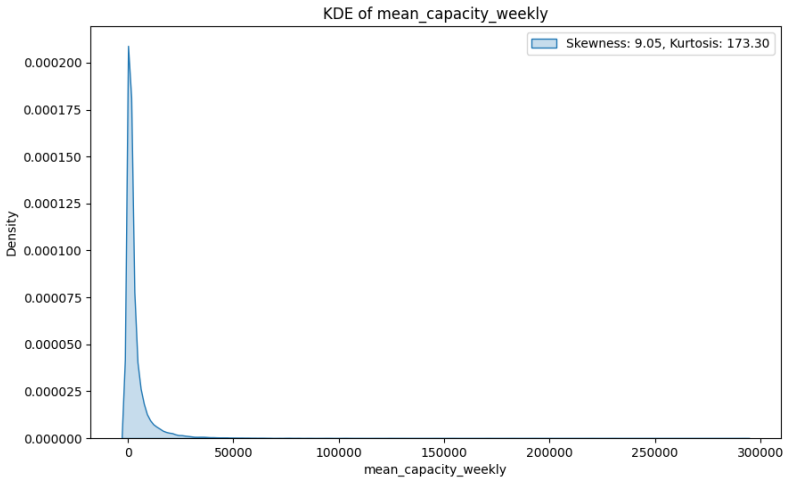

Figure 6.3.: Kernel Density Estimate (KDE) plot of mean capacity for non-directional markets

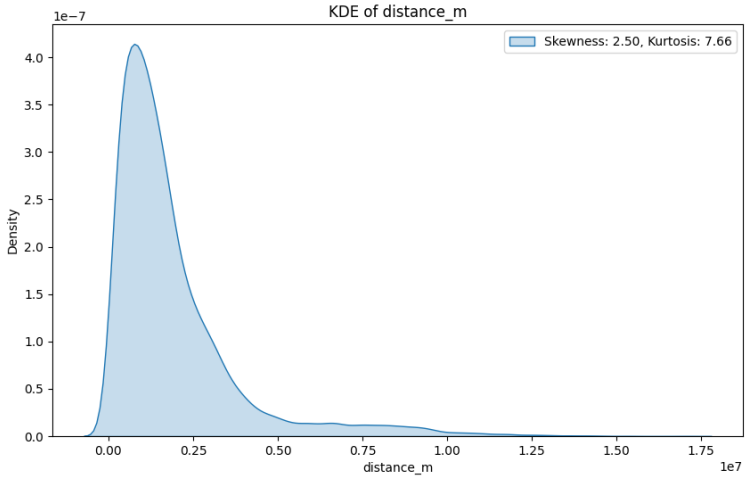

Figure 6.4.: KDE plot of distance between airports for non-directional markets

Two characteristics are not distributed with a right skew: the difference between the mean weekday capacity and mean weekend capacity, as well as the ratio of morning flights. Almost all values for the capacity difference (weekday vs. weekend) lie between -1,000 and 1,000 (6.5). Approximately 50 markets have an absolute difference above 1,000. The highest value is around 6,500 capacity difference between weekday and weekend for the

non-directional market MEL-SYD (Melbourne Airport, Sydney Airport - Kingsford Smith International Airport). This means that on an average weekday, there are 6,500 more seats available than on an average day of the weekend in the MEL-SYD non-directional market. The KDE plot of the morning flights ratio (6.6) shows that a substantial number of non-directional markets do not have morning flights, numerous markets have between 30% and 50% of their flights scheduled in the morning, and certain markets have exclusively morning flights.

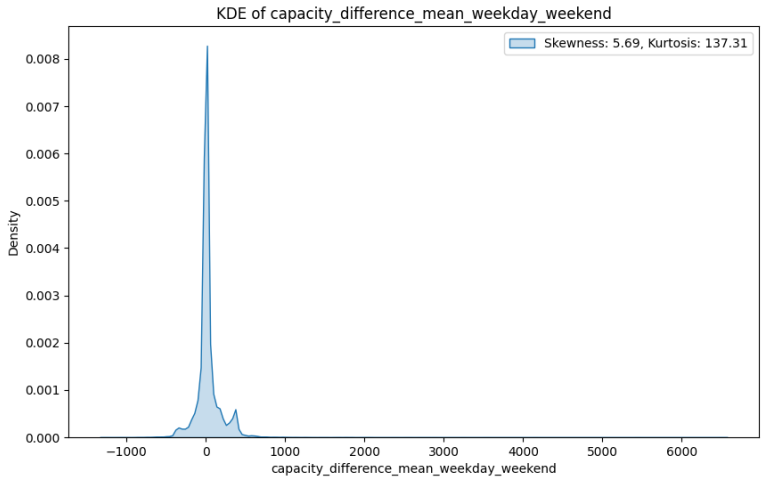

Figure 6.5.: KDE plot of difference between the mean weekday capacity and mean day of the weekend capacity for non-directional markets

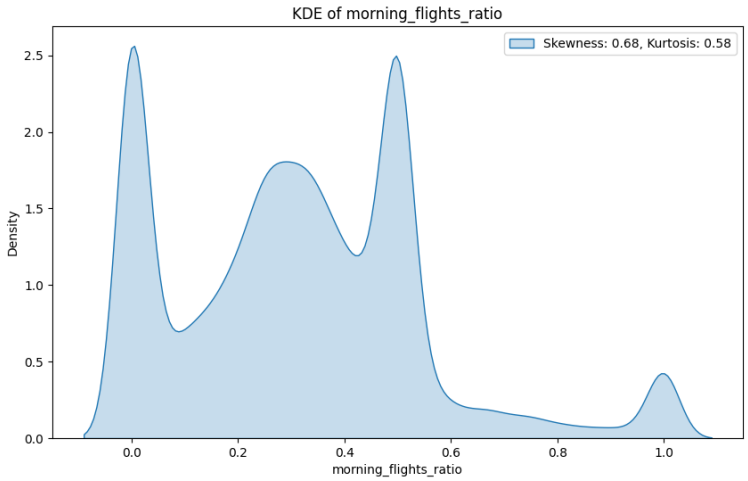

Figure 6.6.: KDE plot of morning flights ratio for non-directional markets

The relationships between the features (bi-variate) are analyzed using a correlation matrix (using Pearson correlation). The correlation analysis reveals that there are strong correlations between capacity and temperature attributes. Furthermore, some of these features are redundant to a certain extend as the provide similar information. One example for this is the annual capacity and the mean capacity. Both features reflect the size of the market and have a very high correlation. Therefore, the annual capacity is removed. The week of minimum capacity and the number of weeks with zero capacity also represent similar information. As many weeks have a capacity of zero, the week of minimum capacity is not an ideal feature because any week with zero capacity could be selected and this does not hold much information. The week of minimum capacity is also removed. The mean temperature features (mean temperature, summer and winter mean temperature) have high correlation values as well, the mean temperature is therefore removed. In addition, the standard deviation of the temperature and the annual amplitude have similar meanings as they demonstrate the variation in temperature. The standard deviation is removed, as the annual temperature amplitude is a feature derived from the temperature domain. Although the mean winter temperature and the annual amplitude have a rather high correlation as well, both features stay in the feature list because of their relevance. The first correlation matrix (6.7) displays the features before the removal of highly correlated and redundant attributes.

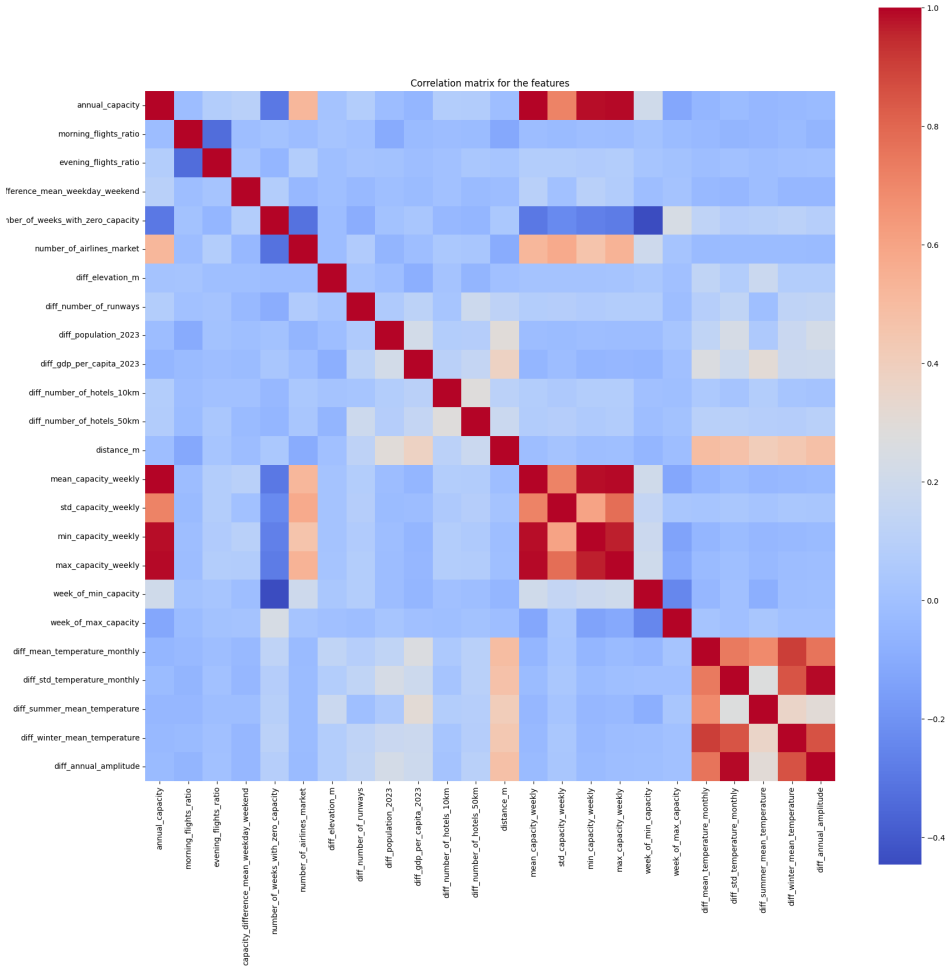

Figure 6.7.: Correlation matrix of features for non-directional markets before removing redundant and highly correlated features

In contrast, the second correlation matrix (6.8) illustrates a significant reduction in the number of highly correlated features. In the appendix, there are the correlation matrices for the initial feature list and the intermediate feature version (using statistically derived variables for temperature and capacity but with airport-related features instead of the differences, see A.1 and A.2).

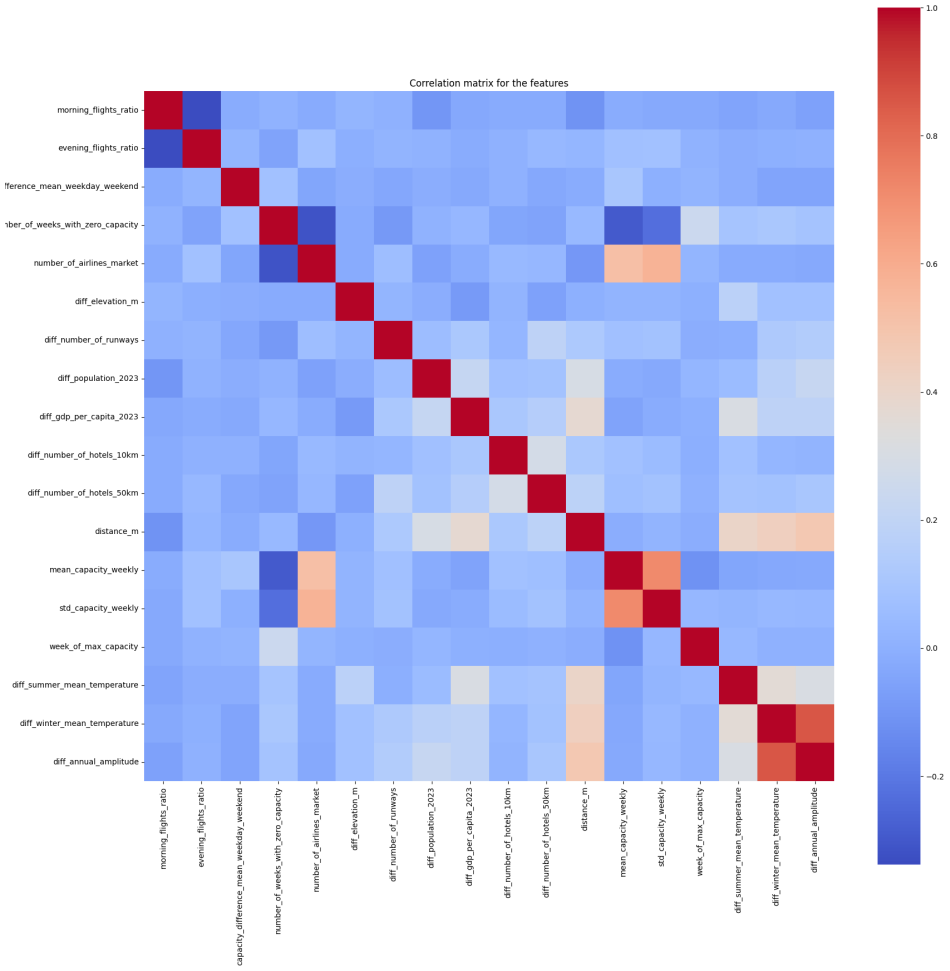

Figure 6.8.: Correlation matrix of features for non-directional markets after removing redundant and highly correlated features

6.6. Feature Transformations

Scaling and Standardization

Feature scaling is an important step in many data mining projects (Provost and Fawcett 2013). z-score standardization is used instead of min-max normalization, due to the

assumption of normal distribution in many algorithms and techniques. For initial experimental clustering runs, the z-score standardization is used. Since numerous features have large outliers, robust z-score standardization is applied to most clustering runs to achieve more balanced feature scales.

Let $\{x_i\}_{i=1}^{N}$ be the sample. X is a numerical feature and N is the number of records.

$$\sigma = \text{standard deviation}(X), \qquad \mu = \text{mean}(X)$$

The formula for z-score standardization is as follows:

$$z_i \;=\; \frac{x_i - \mu}{\sigma} \tag{6.1}$$

The interquartile range (IQR) is the difference between the third and first quartile.

$$\tilde{x} = \text{median}(X), \qquad \text{IQR} = Q_3(X) - Q_1(X)$$

The formula for robust z-score standardization is as follows:

$$z_i robust \;=\; \frac{x_i - \tilde{x}}{\text{IQR}} \tag{6.2}$$

The robust z-score standardization centers each feature on its median and scales by the IQR, a measure more resistant to outliers. By replacing the mean with the median and the standard deviation with IQR, the resulting scores prioritize the central part of the data, making the clustering less sensitive to extreme values (Berthold et al. 2020).

Transformation of Data Distribution

Due to the highly skewed data distributions of many features, most experiments are performed using power transformations. Many models assume a Gaussian distribution or perform poorly on highly skewed data (Tan et al. 2019). The Yeo-Johnson transformation (Yeo and Johnson 2000) is the choice for the transformation because it can reduce skewness and can handle negative and zero values.

The Yeo–Johnson Transform y_i is defined for each data point x_i and parameter λ by:

$$y_i = \begin{cases} \dfrac{(x_i+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0,\ x_i \geq 0, \\ \ln(x_i+1), & \lambda = 0,\ x_i \geq 0, \\ -\dfrac{(-x_i+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2,\ x_i < 0, \\ -\ln(-x_i+1), & \lambda = 2,\ x_i < 0. \end{cases} \tag{6.3}$$

This type of power transformation is continuous in λ and handles both positive and negative data values. Yeo-Johnson transformation seeks to minimize skewness and stabilize variance while preserving the order of the data.

KDE plots of the transformed mean weekly capacity (6.9) and the transformed distance (6.10) show the distributions of the features after the power transformation. The feature distributions are less skewed compared to their previous state (mean weekly capacity: 6.3, distance: 6.4).

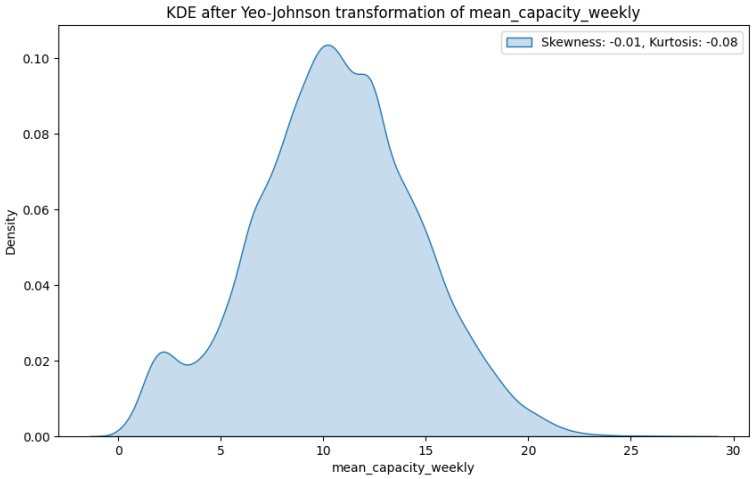

Figure 6.9.: KDE plot of transformed mean weekly capacity for non-directional markets

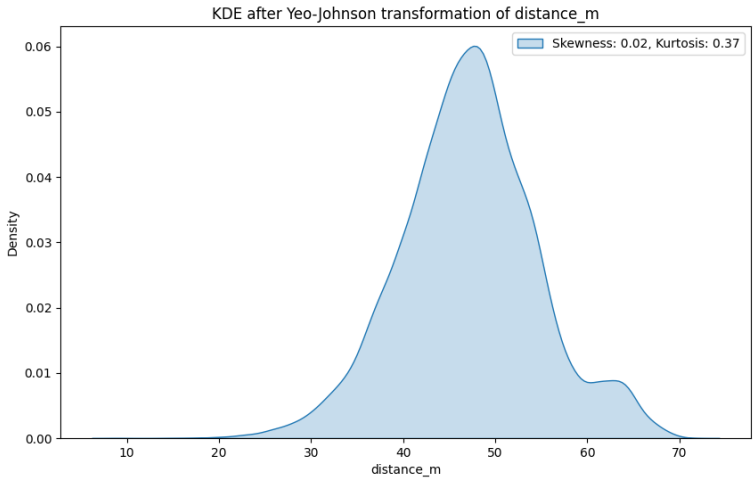

Figure 6.10.: KDE plot of transformed distance between airports for non-directional markets

Dimensionality Reduction

Integrating dimensionality reduction before clustering simplifies complex data into fewer dimensions. Using all features without reducing the dimensions might lead to the curse of dimensionality. By focusing on the most important patterns, techniques like UMAP and PCA remove noise and redundant information, resulting in a potentially clearer separation of markets. Reduced dimensions also speed up clustering algorithms, making them more efficient. Four widely adopted techniques are considered: Principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and autoencoders. Each technique provides different advantages and trade-offs in relation to the data characteristics and the goal of the analysis.

When feature relationships are mainly linear or approximately lie on a linear subspace, PCA is a reasonable option. The first clustering runs with dimensionality reduction are performed using PCA. It uses relatively fast matrix operations, making it suitable for large datasets and experimental runs. By projecting data onto orthogonal components that capture as much variance as possible, PCA transforms high-dimensional inputs into a lower-dimensional space (Pearson 1901; Hotelling 1933). There are different approaches to define how many components the data should be reduced to. For example, one can

use a predefined number of components (e.g., two components for a 2D visualization) or define the percentage of variance that should be captured. Most of the PCA experiments use the setting to keep 90% of the variance in the data. The main disadvantage of PCA is that real-world datasets are usually more complex (not linear). The dataset in this thesis has predominantly non-linear relationships (based on a visual analysis of the scatterplot matrix of all features, see A.3). Since linear projections cannot fully capture the feature interactions, non-linear techniques such as autoencoders, t-SNE and UMAP are further analyzed.

Autoencoders, a type of neural network models, learn to reconstruct inputs through a compressed bottleneck layer. By selecting suitable parameters such as the number of hidden layers or the dimension of the latent space, autoencoders can uncover nuanced, non-linear embeddings (Svirsky and Lindenbaum 2024). However, the process of parameter selection involves numerous experiments and can be very time consuming. Furthermore, the vague nature of cluster analysis makes it even more challenging to find a suitable representation of the data. An Autoencoder is trained using PyTorch (*PyTorch* 2025) with 3 hidden layers (128, 64 and 5 as bottle neck layer for the number of dimensions in the latent space), Rectified Linear Unit (ReLU) as activation function, and a learning rate of 0.01 as base settings. After a few iterations of training, hyperparameter adjustments, and comparisons of feature representations without significant improvements of the clustering results, alternative dimensionality reduction methods are further explored. The repetitive and time-consuming process of trying different data transformations and dimensionality reduction settings seems inefficient when using an autoencoder (potentially also due to the relatively complex dataset).

t-SNE and UMAP are non-linear dimensionality reduction techniques that require fewer parameters than autoencoders. For visualization and exploratory analysis in two or three dimensions, t-SNE is a popular choice. It models pairwise relationships via probability distributions to preserve local neighborhood structure. Although t-SNE is suitable for revealing subgroups in the data, it suffers from a relatively high computational cost and it is non-deterministic (running t-SNE multiple times on the same dataset can yield different results). The main adjustable parameter in t-SNE is perplexity. Low values emphasize local relationships, whereas high values capture more of the global structure in the data (Maaten and G. Hinton 2008). Based on the objective of the analysis and several experiments, a perplexity value of 40 is determined for the reduction of the dimensionality.

UMAP is a more recent approach and quite similar to t-SNE. It outperforms t-SNE in most cases, as it is faster, better at preserving the global structure of the data (not only the local structure) and deterministic (stable results for multiple runs with the same random seed) (Nanga et al. 2021). UMAP constructs a graph structure with weighted edges in both a high-dimensional and low-dimensional space, optimizing the representation of the low-dimensional graph to capture as much information as possible. In the aviation market data experiments in this thesis, UMAP generates the most meaningful embeddings, leading to its selection for subsequent clustering and analysis. In the official UMAP website, McInnes 2018 provides recommended settings for using UMAP for clustering. The most relevant setting for UMAP is the number of data points considered as neighbors. This parameter works similar to the perplexity in t-SNE, as lower values emphasize the local structure, and higher values focus on the global structure in the data. A value of 30 for the number of neighbors tries to capture both local and global structures, focusing more on the global one, as too low values can result in patterns of noise in the data rather than actual subgroups. The minimum distance parameter adjusts how densely points are packed together, for clustering a very low value (in this case, 0) is beneficial to create clearer cluster separations. Setting the number of dimensions to 10 for the low-dimensional space proved to be a reasonable choice in the experiments.

Comparing the two-dimensional representations of the dataset (refined features, with robust standardization and Yeo-Johnson transformation) using PCA, t-SNE and UMAP further emphasizes that UMAP is better suited for this analysis, as clearer subgroups can be identified (see figures 6.11, 6.12 and 6.13). It is relevant to mention that the reduction of features to a two-dimensional space is for the purpose of finding a suitable dimensionality reduction technique through a visual analysis, and it might not ideally represent the actual distribution of data points in the multidimensional space. In the actual cluster analysis, more than two dimensions are used.

The PCA plot of the scaled and transformed features (6.11) reveals that the majority of data points are concentrated in a single large cluster, while some points are farther away from this dense area.

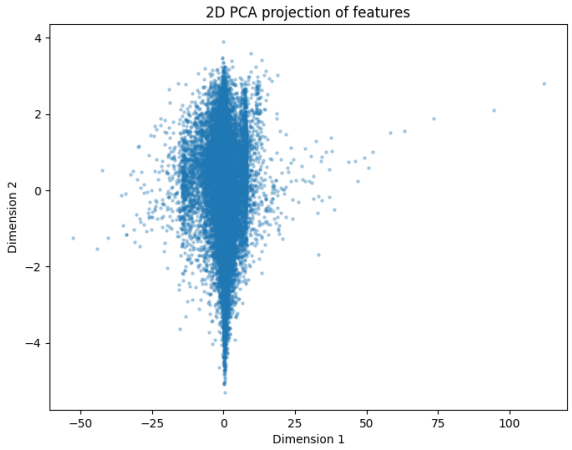

Figure 6.11.: PCA plot of scaled and transformed features

The t-SNE plot of the scaled and transformed features (using Manhattan distance, 6.12) shows one large cluster and a few smaller ones with slightly varying densities.

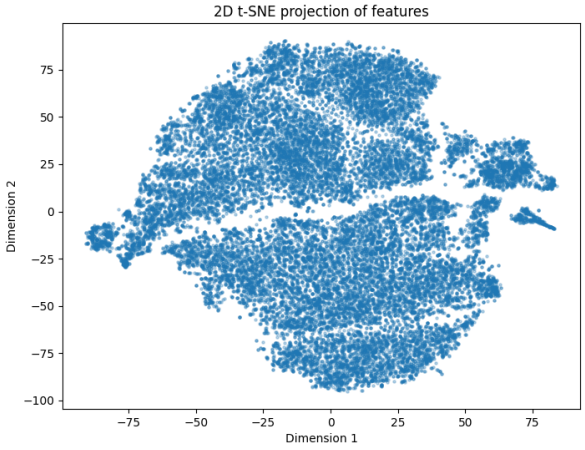

Figure 6.12.: t-SNE plot of scaled and transformed features, using Manhattan distance

The UMAP plot of the scaled and transformed features shows multiple clusters that can be visually distinguished in the two-dimensional space (6.13). Additionally, the clusters appear to have arbitrary forms rather than spherical ones. With respect to the density of the different regions, it is difficult to evaluate in these plots because of the overlap of points. An alternative perspective merges when comparing the two-dimensional UMAP visualization of all features (6.13) against the visualization with reduced features (6.14). In the latter case, UMAP is utilized in two steps: initially reducing the dimensions to 10 components for clustering, followed by further reduction to two dimensions for visualization. This results in an embedding that is even more suitable to form clusters. UMAP is applied twice because the visualization should be based on the same feature representation used in the clustering step.

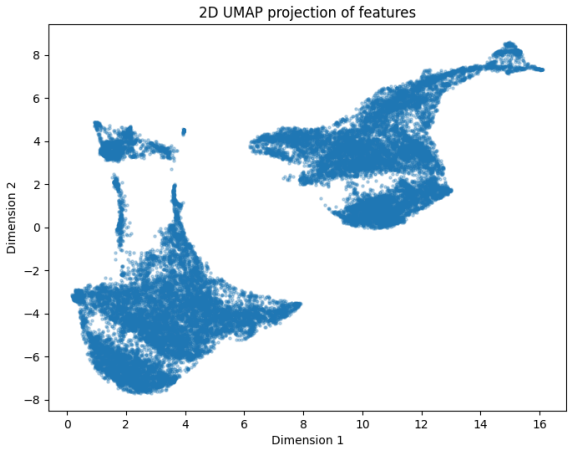

Figure 6.13.: UMAP plot of scaled and transformed features, using Manhattan distance

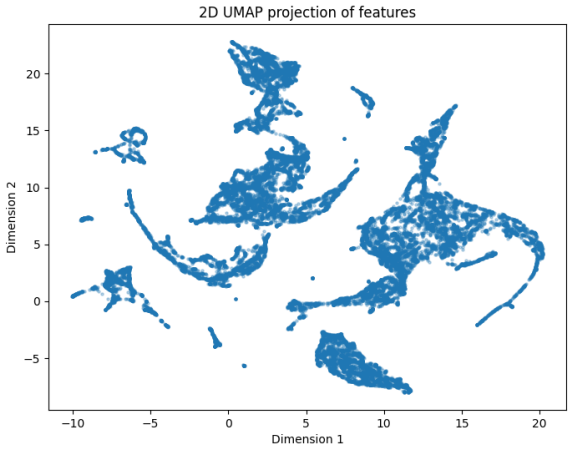

Figure 6.14.: UMAP plot of scaled, transformed and reduced features (using UMAP, reduced to 10 components before the visualization), using Manhattan distance

The reduced number of components (dimensions) using UMAP is determined by running HDBSCAN multiple times (in the range between 5 and 15 dimensions). The decision to use HDBSCAN as the algorithm is made because it does not require much parameter tuning with changing feature representations. The highest Silhouette Coefficient of 0.68 is achieved with 10 UMAP components. Additionally, the authors of UMAP (McInnes 2018) point out that typically UMAP is used to reduce data to 10-20 dimensions. Considering that there are 18 refined features, 10 dimensions appear to be a reasonable balance between preserving information and mitigating the effects of the curse of dimensionality.

7. Modeling

After determining that clustering is a suitable approach for this analysis, the modeling step involves selecting appropriate clustering algorithms for the dataset and objectives. Before applying different clustering techniques to the data, the Hopkins' statistic (Banerjee and Dave 2004) is used to evaluate the cluster tendency of the data. The cluster tendency for both the initial features (92 features, roughly 60,000 directional markets) and the refined features (18 features, roughly 30,000 non-directional markets) is very close to 1 (more than 0.99) which demonstrates that the data have a high potential to form clusters.

7.1. Distance Metric

Choosing an appropriate distance metric is essential because it directly shapes similarities and differences between data points, influencing the formation of clusters (Berthold et al. 2020). In the initial experiments on the aviation features, the Manhattan distance yields a more stable and interpretable segmentation than the Euclidean distance. Regarding outliers, Manhattan penalizes large deviations linearly rather than quadratically. Therefore, the Manhattan distance is used for further analysis.

Minkowski Distance

The Minkowski distance of order p between two points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ is defined as

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{7.1}$$

By varying $p \geq 1$, this family of metrics has different norms: $p = 1$ yields the Manhattan distance, $p = 2$ the Euclidean distance, and larger p place greater emphasis on large differences in attributes (Berthold et al. 2020).

Manhattan Distance

The Manhattan (or city-block) distance is the special case of the Minkowski distance with $p = 1$:

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{7.2}$$

This metric measures the sum of absolute coordinate differences, analogous to navigating a grid of city streets.

7.2. Clustering Algorithms

To improve understanding of the data, the initial algorithm is the widely used K-means clustering technique (note: whenever K-means is referenced, the specific algorithm used is K-means++). K-means partitions the data into k clusters by iteratively assigning each point to its nearest centroid and then updating centroids to minimize the within-cluster sum of squares. A significant advantage of K-means is its relatively short runtime, making it a good choice for quick evaluations (D. Xu and Tian 2015). Experiments are conducted with different parameter settings and feature combinations. To address one disadvantage of K-means (the sensitivity to outliers for the positioning of the centroids), K-medoids is also tested. Unlike K-means, K-medoids uses actual data points as centroids (medoids) rather than artificially created ones (Jain, Murty and Flynn 1999). Unfortunately, K-medoids has a runtime considerably higher than K-means, making this algorithm unsuitable for test runs with numerous parameter configurations, and it does not yield better results than K-means on this dataset. The elbow method is used to determine an appropriate number of clusters, using the within-cluster sum of squares as the evaluation measure. The elbow plot with up to 50 clusters (7.1) suggests that an optimal number of clusters falls between 5 and 10. To compare different iterations, clustering is performed with 5, 7 and 10 clusters.

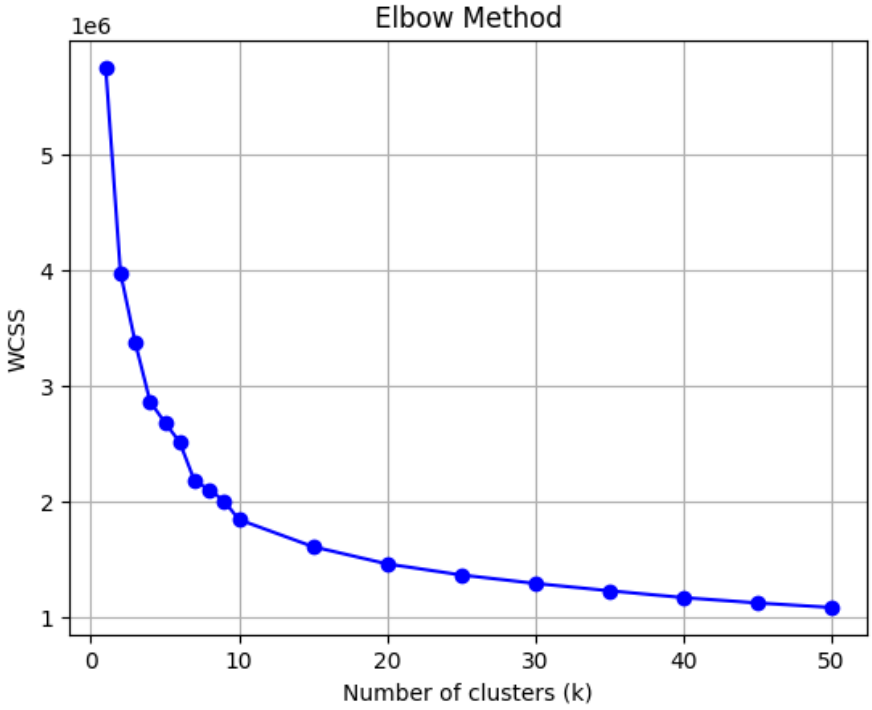

Figure 7.1.: Elbow plot with the number of clusters k on the x-axis and the within-cluster sum of squares on the y-axis

DBSCAN appears to be a suitable option for further clustering analyses based on UMAP visualizations of features within the two-dimensional space (6.13, 6.14). DBSCAN groups points into clusters of arbitrary shape by requiring each core point to have at least epsilon-radius density of neighbors (minimum samples) while labeling sparser points as noise. This density-based approach automatically discovers clusters without specifying the number of clusters beforehand, and is robust to outliers. The data has arbitrary shapes, and the densities of data points appear to vary across different areas. The k-distance plot (7.2) can help determine a suitable starting point for epsilon (Yin et al. 2023). This approach is similar to the elbow plot for K-means, the position of the elbow is relevant. As a simplification, the number of minimum points parameter is set to two times the number of features (in this case 20, as there are 10 UMAP reduced features). Based on this graph, a reasonable epsilon value is 0.75.

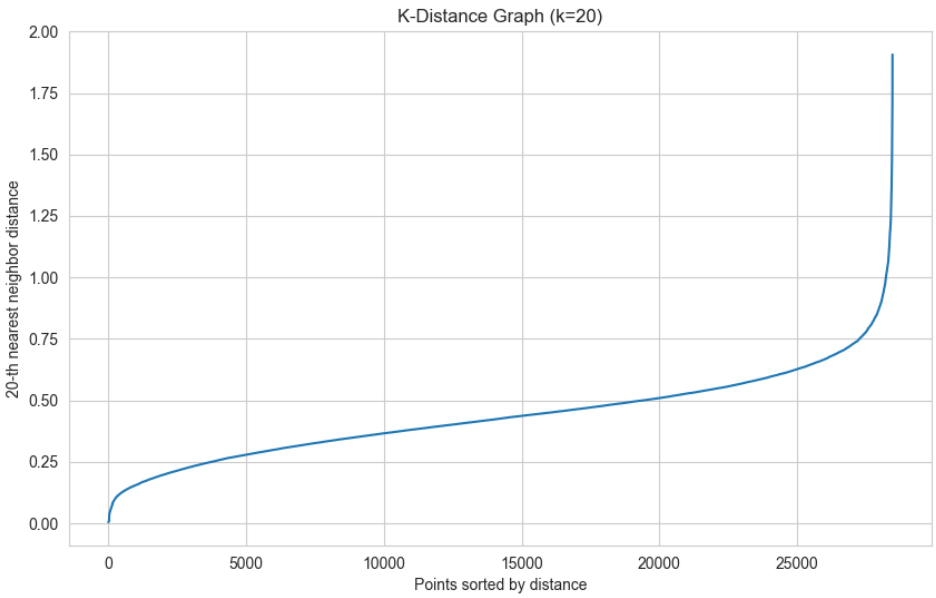

Figure 7.2.: K-distance plot for determining a starting point for the epsilon value for DBSCAN with the data points sorted by distance on the x-axis and the 20-th nearest neighbor distance on the y-axis

In addition to DBSCAN, the OPTICS algorithm is worth exploring as it is a density-based algorithm that dynamically adjusts the density parameter settings (epsilon and minimum number of points). OPTICS stands for Ordering Points To Identify the Clustering Structure and it computes a reachability plot for the data points that enables the extraction of clusters at different density levels (Ankerst et al. 1999). The authors of UMAP (McInnes 2018) point out that UMAP does not necessarily create well-defined spherical clusters. Therefore, using K-means may not be ideal. Instead, they advise using HDBSCAN (Hierarchical DBSCAN), since UMAP assumes uniform density and does not preserve density well. This indicates that HDBSCAN might perform better than OPTICS with the UMAP-reduced features. HDBSCAN builds a hierarchy of density-based clusters. This algorithm varies the density threshold, then selects the most "persistent" clusters in the built hierarchy (Campello, Moulavi and Sander 2013). With only a minimum cluster size (and optional minimum samples) to tune, HDBSCAN robustly handles noise and variable-density data points, making it a strong choice for a complex aviation dataset.

The feature embeddings indicate that certain markets might be situated near the decision boundaries that separate clusters. Consequently, it is worth examining soft clustering methods, such as the Gaussian Mixture Model (GMM), which is a distribution-based algorithm. GMM represents the data as a weighted sum of multivariate normal distributions, where each component corresponds to a cluster. Fitting a GMM relies on the Expectation–Maximization algorithm, an iterative two-step procedure. In the expectation step, the algorithm computes the posterior probability that each point belongs to each Gaussian component (cluster), based on the current estimates of means, covariances, and mixture weights. In the maximization step, it updates those parameters to maximize the expected log-likelihood. These steps repeat until convergence, yielding soft cluster assignments and estimates that best explain the data (Fraley and Raftery 2002). Due to the relatively large dataset, hierarchical clustering algorithms such as agglomerative hierarchical clustering are not considered.

7.3. Internal Validation Metrics

For internal validation of the clustering results, the Silhouette Coefficient (values close to 1 preferable) and the Davies-Bouldin Index (small values preferable) are used. The implementation from scikit-learn is used for these metrics (*scikit-learn 1.7.0 documentation* 2025). In addition to the validation metrics, a visual inspection of a two-dimensional representation of the clusters is conducted.

Silhouette Coefficient

For each point x_i, let

$$a_i = \frac{1}{|C_i| - 1} \sum_{x_j \in C_i,\, j \neq i} d(x_i, x_j), \qquad b_i = \min_{C \neq C_i} \frac{1}{|C|} \sum_{x_j \in C} d(x_i, x_j)$$

The Silhouette of x_i is

$$s_i \;=\; \frac{b_i - a_i}{\max(a_i,\, b_i)} \tag{7.3}$$

The overall score is the average $S = \frac{1}{N} \sum_{i=1}^{N} s_i$

The Silhouette Coefficient measures how well each point lies within its assigned cluster versus the nearest other cluster. Values near +1 indicate well-separated clusters, while values near −1 suggest misclassifications (Rousseeuw 1987).

Davies–Bouldin Index

Let each cluster C_i have centroid μ_i and scatter

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} d(x, \mu_i).$$

$$d(x, \mu_i) \text{ ... cohesion between point } x \text{ and its centroid } \mu_i$$

$$S_i \text{ ... within cluster scatter for cluster } i$$

The Davies–Bouldin index is defined as

$$\text{DB} \;=\; \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{S_i + S_j}{d(\mu_i,\, \mu_j)}. \tag{7.4}$$

$$d(\mu_i, \mu_j) \text{ ... separation between cluster } i \text{ and } j$$

$$K \text{ ... the set of } k \text{ clusters}$$

This index computes the average "worst-case" ratio of within-cluster scatter to between-cluster separation. Lower values indicate more compact, well-separated clusters (Davies and Bouldin 1979).

7.4. Experimental Settings

This is a summary of the different techniques used alongside the parameter settings (determined by multiple experiments). Figure 6.14 illustrates the feature embeddings, providing a visualization of how these settings are represented in a two-dimensional space.

Type	Techniques	Parameters for Main Results
Metric for distance calculations	Manhattan, Euclidean	Manhattan distance (as it is known to be more robust)
Scaling	Standard and robust z-score standardization	RobustScaler from scikit-learn (with centering, with scaling, with the interquartile range using the 25th and 75th quantiles)
Transformation of data distribution	Yeo-Johnsohn transformation	PowerTransformer from scikit-learn (using Yeo-Johnsohn transformation to reduce skew)
Dimensionality reduction	PCA, Autoencoder, t-SNE, UMAP	UMAP reducing to 10 components
Clustering algorithm	Partition-, density-, and distribution-based	K-means++, DBSCAN, OPTICS, HDBSCAN, Gaussian Mixture Model

Table 7.1.: Summary of experimental settings and used techniques

8. Results

The refined list of features (6.5), as well as the parameter settings and techniques described in the modeling chapter (scaling, transformation, and dimensionality reduction, 7.1) are the basis for the clustering results using various clustering algorithms (if not stated otherwise). The scikit-learn implementations of the clustering algorithms are used for these experiments, and the default settings of the version 1.7.0 are used.

8.1. Overview of Main Clustering Results

The default settings and techniques for the clustering results are the refined features, robust z-score standardization, Yeo-Johnson transformation, and UMAP dimensionality reduction with 10 components. The main clustering results are listed in the table 8.1.

Main Clustering Results

Algorithm	Clusters	Noise Points	Silhouette | DB Index	Settings
K-means	10	0	0.66 | 0.49	k=10
DBSCAN	4	250	0.63 | 1.17	epsilon=0.75, minimum points=20
DBSCAN	21	2,500	0.48 | 0.58	epsilon=0.5, minimum points=15
OPTICS	266	23,000	0.04 | 0.16	minimum samples=10, xi=0.05
HDBSCAN	4	0	0.68 | 0.32	minimum cluster size=20, selection method=eom
HDBSCAN	116	21,000	0.31 | 0.28	minimum cluster size=20, selection method=leaf
HDBSCAN	19	400	0.59 | 0.60	minimum cluster size=20, selection method=eom, no power transformation
HDBSCAN	773	15,000	0.75 | 0.12	minimum cluster size=5, selection method=eom, no power transformation
GMM	4	0	0.63 | 0.53	k=4
GMM	10	0	0.63 | 0.54	k=10
GMM	500	0	0.87 | 0.20	k=500

DB Index ... Davies-Bouldin Index; k ... number of components

Table 8.1.: Summary of main clustering results

8.2. K-means

The result for K-means++ clustering includes 10 clusters (8.1). The number of clusters is determined using the elbow plot (7.1), the highest Silhouette Coefficient (0.66) and the lowest Davies-Bouldin Index (0.49) for the runs with 5, 7 and 10 clusters. The visualization

shows generally well-separated clusters and also highlights one of the limitations of K-means when applying it to non-spherical data shapes (for example, the dark blue cluster on the right side). Some clusters overlap, such as the red and purple clusters in the central region, which indicates that some data points lie near the decision boundary and might be reassigned under different settings. Based on the shape of the clusters and the seemingly varying densities, the next step is to apply density-based algorithms.

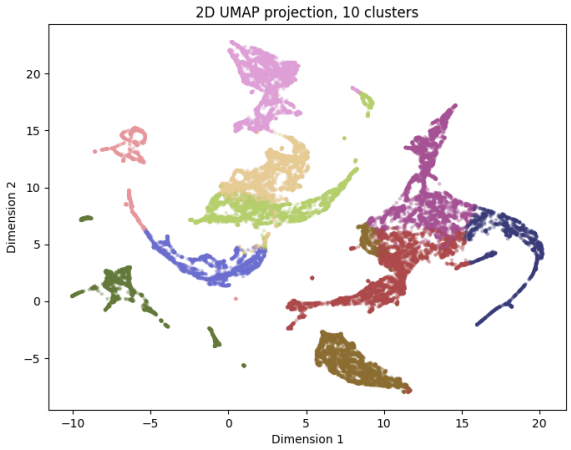

Figure 8.1.: UMAP projection of K-means++ clustering result with 10 clusters

8.3. DBSCAN

Two major challenges arise when applying DBSCAN to the dataset. Primarily, the highest validation metric scores often occur with a minimal number of clusters, sometimes as low as a single cluster (for instance, achieving a Silhouette Coefficient of 0.89 with just one cluster). Another challenge is the substantial portion of the data being classified as noise in many cluster runs. The initial DBSCAN run (8.2), with the epsilon value set to 0.75 and the number of minimum points set to 20 has a Silhouette Coefficient of 0.63 and a Davies-Bouldin Index of 1.17. The 4 clusters seem well separated. However, in the two-dimensional visualization it seems that a more granular separation could be better (for example, in the bottom right the large green cluster). Furthermore, there is one barely visible cluster in the bottom center and there are roughly 250 noise points, which is very low compared to other density-based cluster runs.

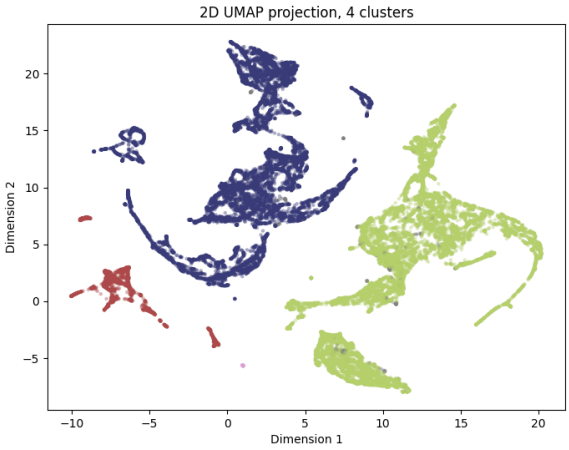

Figure 8.2.: UMAP projection of DBSCAN clustering result with 4 clusters, epsilon: 0.75, minimum points: 20

Dividing around 30,000 airport markets into just four distinct groups may offer limited insight. Consequently, the parameters are altered to produce a greater number of clusters. Using slightly smaller values for the epsilon parameter (0.5) and the minimum number of points (15), the Silhouette Coefficient decreases to 0.48 (worse) and the Davies-Bouldin Index also decreases to 0.58 (better). The resulting number of clusters increases to 21 and the number of noise points rises to approximately 2,500 (8.3). Assessing whether this clustering result is better than the previous one using internal validation metrics is difficult because one metric improves while the other deteriorates. In addition, these metrics serve as indicators of the quality of the clusters. The cluster results can be compared taking into account the objectives of the project and performing further analysis. The DBSCAN result with 21 clusters includes a few large clusters (for example, the dark purple and green ones in the central region of the plot) and many very small clusters. This could indicate that many markets are similar (large clusters), whereas some airport markets belong to niche groups.

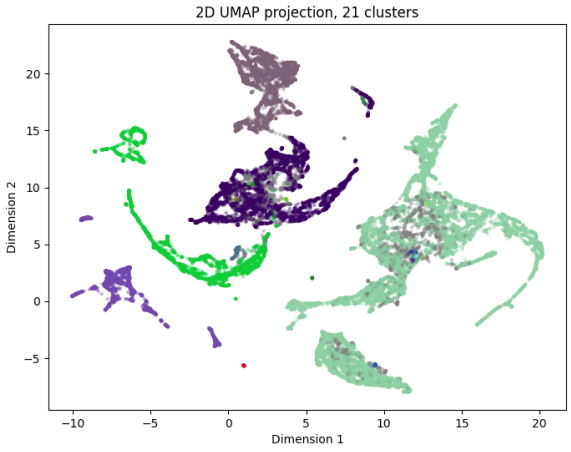

Figure 8.3.: UMAP projection of DBSCAN clustering result with 21 clusters, epsilon: 0.5, minimum points: 15

8.4. OPTICS

OPTICS, a density-based clustering algorithm, is particularly well-suited for datasets characterized by significantly varying densities. One big advantage is that the epsilon value does not have to be set in advance and is dynamically adjusted by the algorithm based on the density of the regions (Ankerst et al. 1999). In this clustering experiment, the initial parameter for the minimum number of samples required to define a dense region is set to 10 (the same as the number of features). The xi parameter, which represents the steepness of the reachability plot constructed by the OPTICS algorithm, is configured to 0.05 (the default value, larger values produce a greater number of clusters). The main finding in this cluster result (8.4) is the extremely high number of noise points (more than 23,000). Other parameter settings for OPTICS result in similarly high number of noise points (minimum samples in the range of 10 to 30, xi value between 0.05 and 0.15). The Silhouette Coefficient for these cluster runs has values around 0 and the Davies-Bouldin Index lies between 0.15 and 0.5). Due to the large number of noise points, the OPTICS algorithm is not considered further.

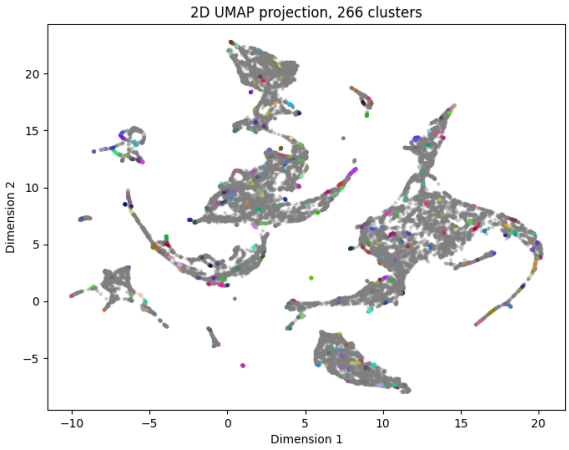

Figure 8.4.: UMAP projection of OPTICS clustering result with 266 clusters, minimum samples: 10, xi value: 0.05

8.5. HDBSCAN

Another density-based clustering technique is HDBSCAN, an extension of DBSCAN. It extracts a hierarchy of density clusters and selects the most stable ones. One important parameter is the minimum cluster size, it defines how small the groupings can be (Campello, Moulavi and Sander 2013). This parameter is set to 20 (based on two times the number of features, similar to the minimum samples parameter in OPTICS). The clustering result with these settings (8.5) yields 4 clusters and is very similar to the DBSCAN clustering result with fewer clusters (8.2). There are two large clusters, one smaller one, one tiny, barely visible cluster, and no noise points. The Silhouette Coefficient is 0.68 and the Davies-Bouldin Index is 0.32 (both are better than the internal metrics for the 4 DBSCAN clusters). Although a small number of clusters offers to some extent only limited information on the numerous airport markets, these internal validation metrics are technically strong and robust.

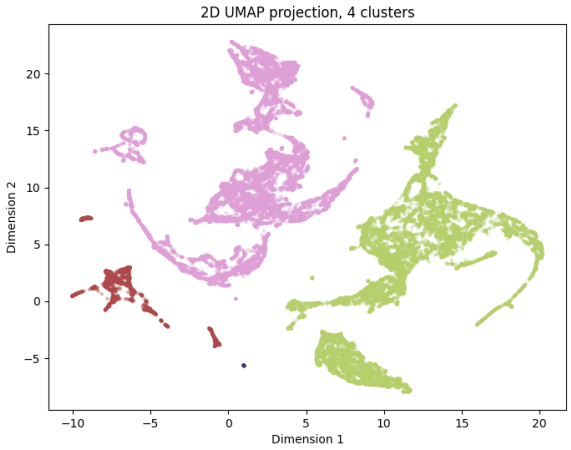

Figure 8.5.: UMAP projection of HDBSCAN clustering result with 4 clusters, minimum cluster size: 20, cluster selection method: eom

With "leaf" as the cluster selection method instead of the default "eom", more fine-grained and homogeneous clusters can be found. eom refers to "Excess of Mass", this method finds the most persistent clusters according to *scikit-learn 1.7.0 documentation* 2025. These settings result in 116 clusters, more than 21,000 noise points, a Silhouette Coefficient of 0.31 and a Davies-Bouldin Index of 0.28 (8.6). An interesting observation is that the Davies-Bouldin Index technically has very good values for clustering results with many clusters (and also many noise points). One potential explanation is the existence of extremely compact clusters that reduce the value of the Davies-Bouldin Index. Additionally, these tiny, well-separated clusters might indicate high-quality market segments, with the disadvantage of many unclassified markets.

A relevant factor is that the Yeo-Johnson transformation is applied on the features to mitigate skewness and transition towards a more Gaussian-like distribution (Yeo and Johnson 2000). Although this method can improve clustering results, it can also complicate interpretation due to the transformation involved. Therefore, it is valuable to analyze the clustering results for features without the Yeo-Johnson transformation as well (with robust z-score standardization and dimensionality reduction applied). This form of feature embedding tends to be more fragmented compared to the embedding when the power transformation is applied.

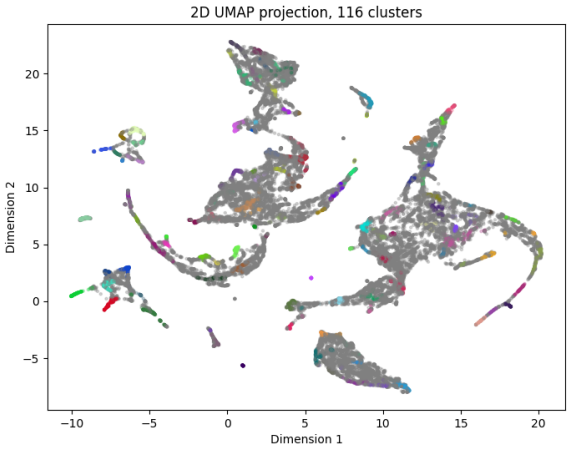

Figure 8.6.: UMAP projection of HDBSCAN clustering result with 116 clusters, minimum cluster size: 20, cluster selection method: leaf

Using a minimum cluster size of 20 (matching roughly two times the number of features) with the "eom" method yields 19 clusters and approximately 400 noise points (8.7). One of the resulting clusters encompasses over 25,000 markets, indicating highly imbalanced clusters. Here, the Silhouette Coefficient results in a value of 0.59 and a Davies–Bouldin Index of 0.60. Despite the lower internal validation scores, this configuration offers a balance between cluster granularity and interpretability, producing a manageable number of broad market segments rather than hundreds of very small groups. Depending on the goal of the analysis, this can be advantageous.

Setting the minimum cluster size to 5 (relatively low value to identify many clusters) and using the 'eom' selection method produce 773 clusters (plus roughly 15,000 noise points, see figure 8.8). The Silhouette Coefficient for this cluster result is 0.75, and the Davies–Bouldin Index falls to 0.12, reflecting extremely tight and well-separated small clusters. Most clusters contain only a handful of markets, and the large volume of noise points suggests that HDBSCAN is treating lower density regions as not assignable. In each scenario, not applying a power transform alters the density landscape created by UMAP. This suggests that HDBSCAN frequently either over-fragments the data or consolidates various markets into just a few large clusters. A similar observation can be made when analyzing the cluster results of the default feature representation (with power transformation).

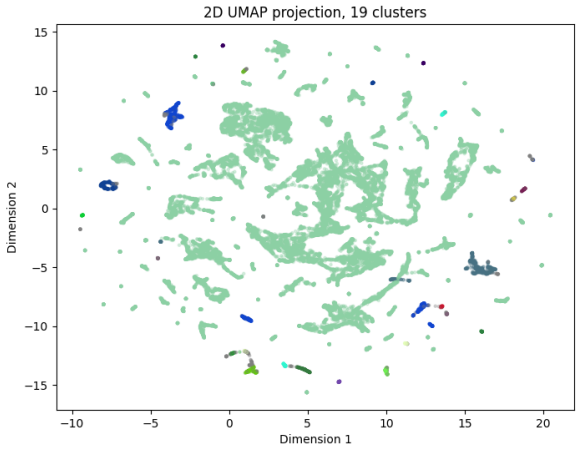

Figure 8.7.: UMAP projection of HDBSCAN clustering result with 19 clusters, minimum cluster size: 20, cluster selection method: eom, no Yeo-Johnson transformation

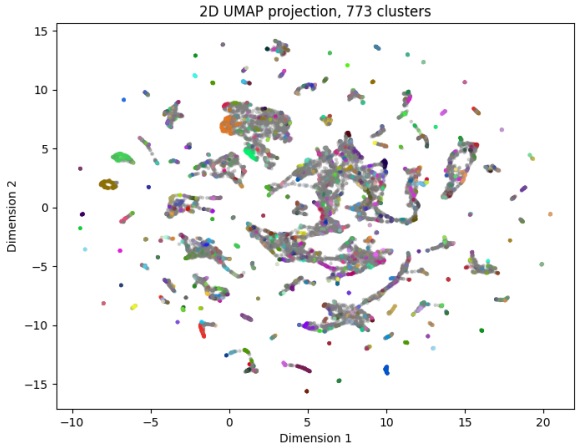

Figure 8.8.: UMAP projection of HDBSCAN clustering result with 773 clusters, minimum cluster size: 5, cluster selection method: eom, no Yeo-Johnson transformation

8.6. Gaussian Mixture Model

To this point, only hard clustering algorithms are used. Given that there is some overlap among clusters in the different clustering results, it is worthwhile to investigate a soft clustering approach such as the Gaussian Mixture Model as well. Mixture models generalize K-means by taking into account both the covariance structure of the dataset

and the means of the underlying Gaussian distributions (Fraley and Raftery 2002). The main parameter for the GMM is the number of components (clusters). Based on the number of clusters of previous cluster results, it was decided to use 4 (from the DBSCAN and HDBSCAN runs) and 10 (from the best K-Means run) as the number of components. Other parameters are the covariance type which is set to "full" and the initialization method which is set to "k-means++".

The cluster result with 4 clusters (8.9) has a Silhouette Coefficient of 0.63 and Davies-Bouldin Index of 0.53 which is slightly worse than the HDBSCAN result for 4 clusters from the perspective of internal validation metrics. In the two-dimensional representation, the Gaussian-contours (ellipses at different standard deviation radii) visualize the probability of the cluster membership and also the overlap of clusters.

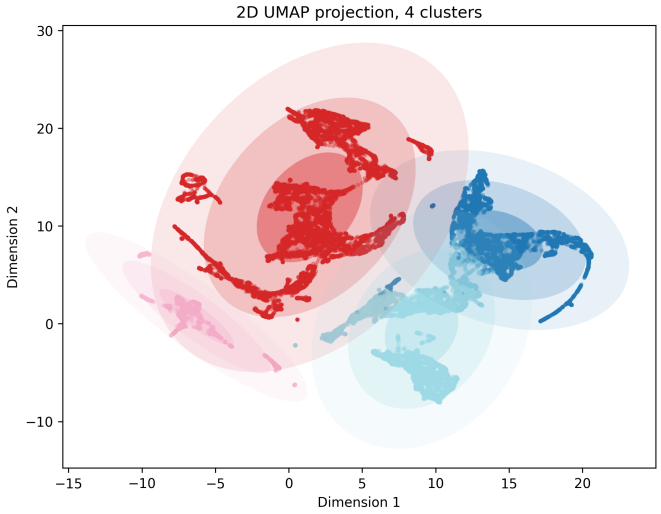

Figure 8.9.: UMAP projection of GMM clustering result with 4 clusters

The GMM cluster run with 10 clusters (8.10) resulted in a Silhouette Coefficient of 0.63 and Davies-Bouldin Index of 0.54 (almost the same as for 4 clusters). An interesting observation is that the clusters appear to be more evenly balanced compared to the cluster results achieved with density-based algorithms. For example, in the GMM result with 10 clusters, the largest five clusters contain data points in the range between 2,700 and 5,000, whereas in the HDBSCAN run with 4 clusters, the two largest clusters have roughly 13,000 markets, and the HDBSCAN run with 19 clusters, the largest cluster has even roughly 25,000 markets.

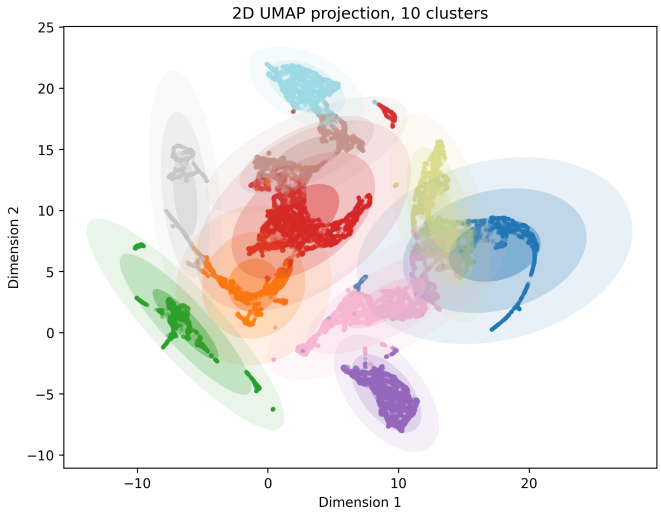

Figure 8.10.: UMAP projection of GMM clustering result with 10 clusters

When running GMM with various different numbers of components (range between 4 and 500), the cluster run with 500 clusters (8.11) has the highest Silhouette Coefficient (0.87) and one of the lowest Davies-Bouldin Index values (0.2).

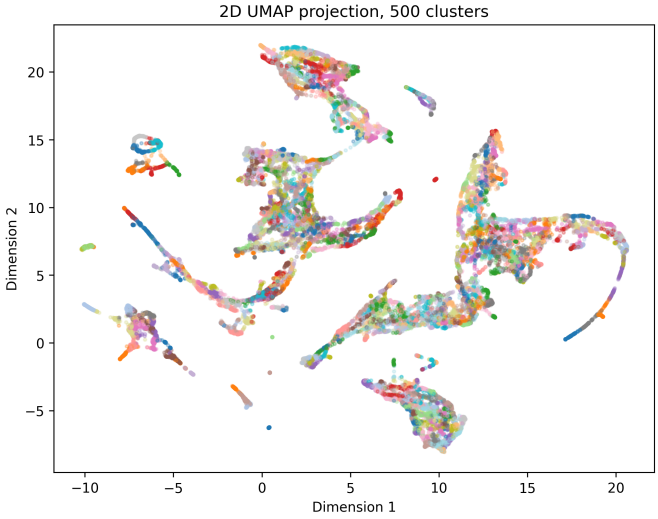

Figure 8.11.: UMAP projection of GMM clustering result with 500 clusters

The plot with scores from different GMM runs (8.12) demonstrates that, in general, the Silhouette Coefficient increases with the number of clusters and Davies-Bouldin decreases, indicating technically better clustering results. An advantage of using GMM is the absence of noise points in the results. However, this can also be perceived as a drawback, as in certain markets, it may be appropriate to exclude markets from clusters if they significantly differ from others. On the other hand, excluding a large part of the markets (in some cases even more than half of the markets) could make the results unusable for various applications.

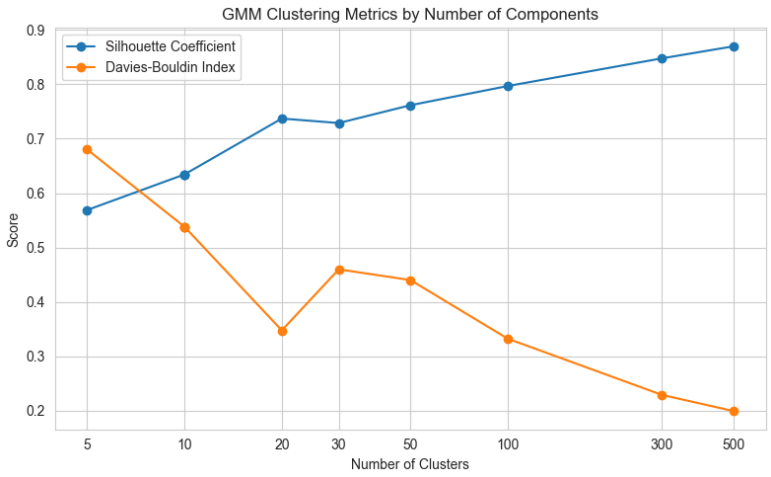

Figure 8.12.: Plot of various GMM runs with the number of clusters on the x-axis and the internal validation metrics on the y-axis

8.7. Interpretation

Given that the number of global direct, non-directional airport markets is approximately 30,000, conducting an in-depth analysis on each cluster or even market is not possible. There are various different techniques that can be used to make sense of the clustering results.

Interpretable clustering is a more recent method aimed at better understanding cluster formations. An option is to train a classification model on the cluster labels such as a decision tree or a random forest classifier to obtain the feature importance values and analyze them. Another approach is to use an explainable cluster algorithm such as explainable K-means (this iteratively runs K-means and builds an explainable clustering tree to better understand the cluster assignments) (H. Yang, Jiao and Pan 2021; Hu et al. 2024). For this analysis, the first approach is used, training a random forest classification model with 200 estimators on the refined features (without scaling, transformation or reduction of the dimensions). The feature importance values for the GMM clustering results for 10 and 500 clusters are created, see figures 8.13 and 8.14. The highest feature importance value in both results is the mean capacity difference between a weekday, and a day of the weekend. This suggests that schedule variability over the week is a major factor. Interestingly, the feature with the second highest feature importance is the

number of weeks with no capacity for the 10 clusters, reflecting seasonal services, while for the 500 clusters it is the difference in annual temperature amplitude, representing different climates. Although lower in the ranking, the physical distance still contributes meaningfully, reflecting operational profiles (short-haul vs. long-haul markets). In terms of additional characteristics, the two results have some common features, in a different order, such as the weekly mean capacity, the standard deviation of the weekly capacity, and the difference in the mean temperature in winter. However, the significance and combination of these features differ substantially in the results of the clusters. This outcome is to some extent expected, as markets grouped into fewer, larger clusters share different similarities than markets grouped into many, smaller clusters. The interpretable clustering approach aids domain experts in validating the clustering results (Hu et al. 2024).

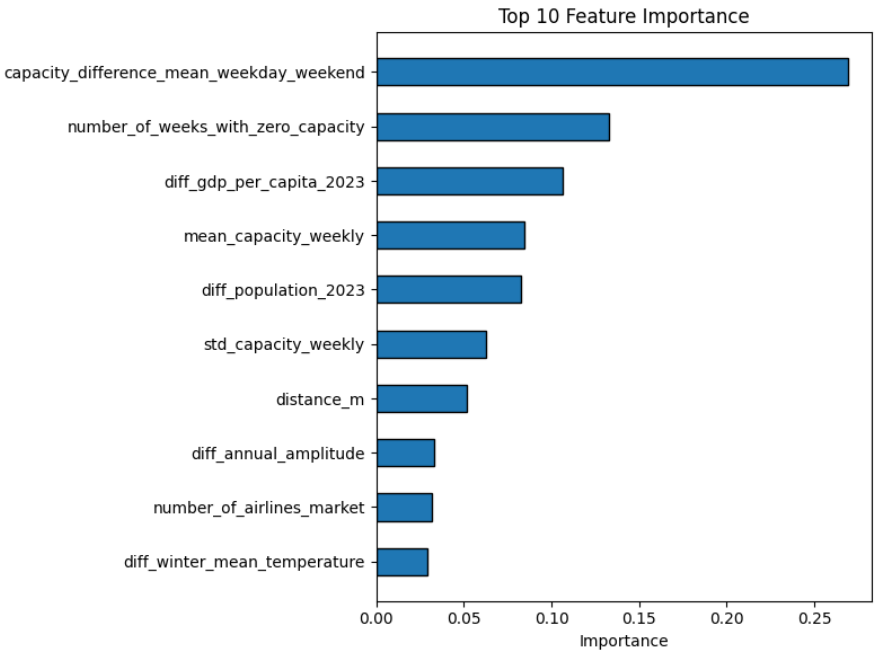

Figure 8.13.: Feature importance values for GMM clustering result with 10 clusters

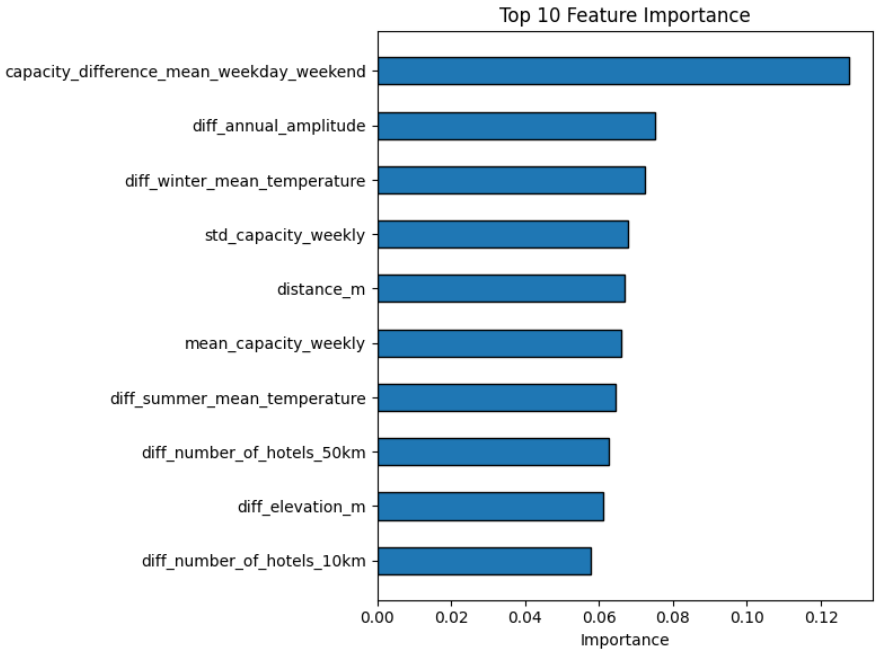

Figure 8.14.: Feature importance values for GMM clustering result with 500 clusters

An alternative interpretation strategy involves analyzing cluster representatives directly (centroids or medoids) (Park and Jun 2009). Given that the clustering results are based on reduced components, this analysis is not ideal, since the reduced feature values lack interpretability. Consequently, one might also calculate the medoids using the scaled and transformed features (without reduction) for a more comprehensible analysis. However, these medoids should be handled with caution, as they are not based on the same features as clustering. An example table showcasing the cluster representatives for the Gaussian Mixture Model result with 10 clusters is presented in table 8.2. There are only a small number of features due to limited space. These are selected based on the importance of the features in the interpretable cluster analysis. The mean difference in mean capacity between a weekday and a day if the weekend may serve as an indicator to distinguish business-oriented routes from leisure-oriented ones. Positive values indicate greater availability of services during weekdays, whereas negative values suggest increased weekend service, typically associated with leisure travel. The year-round and seasonal service can be distinguished by the number of weeks without capacity.

Example of Representative Markets

Market	Mean weekday-weekend capacity difference	Weeks with zero capacity	Mean weekly capacity	GDP per capita difference (2023)
Marignane (MRS) → Nantes Atlantique Airport (NTE)	110	0	9,459	0
Bergamo/Orio Al Serio (BGY) → Helsinki-Vantaa (HEL)	-70	0	2,116	13,922
Gdansk Lech Walesa (GDN) → Dionisios Solomos (ZTH)	372	27	314	1,344
Julius Nyerere International (DAR) → Lusaka Intl (LUN)	13	0	1,147	106
Bradley Intl (BDL) → Northern Kentucky International (CVG)	0	33	284	0
King Abdulaziz Intl (JED) → Osmani International (ZYL)	87	0	1,359	29,542
Nizhnevartovsk (NJC) → Tolmachevo (OVB)	-3	0	1,611	0
Goteborg-Landvetter Airport (GOT) → Palma de Mallorca (PMI)	-233	16	1,901	22,007
Biju Patnaik (BBI) → Chennai International Airport (MAA)	-11	0	5,848	0
Alicante Airport (ALC) → Klagenfurt (KLU)	-2	21	437	22,524

Table 8.2.: Representative markets (medoids) of the GMM clustering result with 10 clusters

From a technical perspective, the cluster balance can also be an indication of the cluster quality (D. Xu and Tian 2015). Depending on the objective of the analysis, the desired outcome might be evenly sized clusters. Many GMM cluster results have rather balanced clusters, an example is the GMM result with 10 clusters (cluster counts: 8.15, cluster visualization: 8.10). The range for cluster sizes is between 1,000 and 5,000 markets. This clustering result might be well-suited for further analysis towards market classification. It is worth mentioning that highly imbalanced clusters can also be a reasonable result if the underlying data actually contain these patterns, and very small clusters might reflect niche market segments that contain valuable insights. An example of highly imbalanced clusters is the HDBSCAN result with 4 clusters (cluster counts: 8.16, cluster visualization: 8.5). Here, there are two large clusters with roughly 13,000 markets, one smaller one with almost 2,000 markets, and a tiny group with only 33.

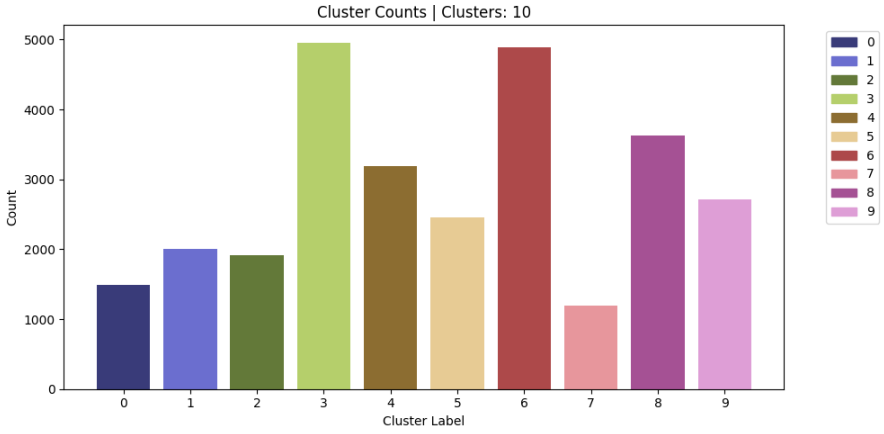

Figure 8.15.: GMM cluster counts with 10 clusters

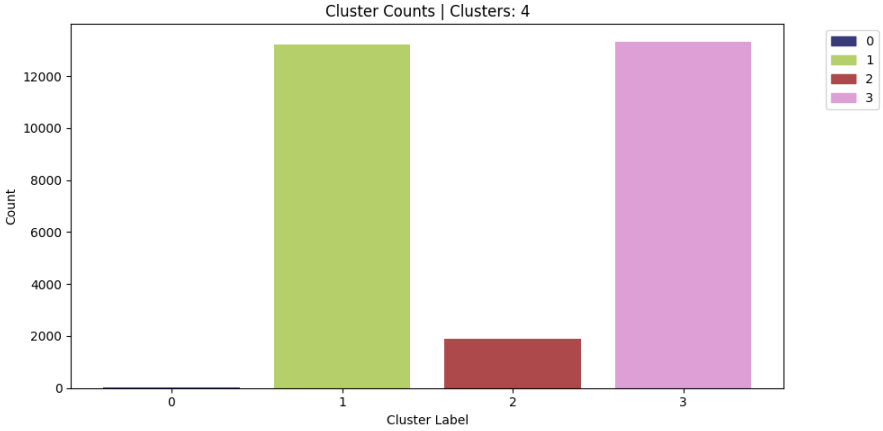

Figure 8.16.: HDBSCAN cluster counts with 4 clusters

Another idea is to select a few airport markets (or considering all markets linked to one airport at a time). Visualization on a world map might be a helpful tool in this context, as airport markets have a geographical focus, making them easier to interpret. Figure 8.17 provides an example of such a visualization.

Depending on the size of the cluster, it is also possible to use the cluster label as a filter on the world map. In figure 8.18, the smallest cluster (7) of the GMM result with 10 clusters demonstrates this. It is clearly visible that around 1,200 markets are too many to analyze in detail (especially because of the strong overlap). However, some patterns, such as the strong focus on Europe and the east cost of North America, can be depicted.

Figure 8.17.: Visualization of clustering result for all airport markets in a world map that either have their origin or destination in Vienna Airport (VIE); markets are represented as lines (using great-circle distance); based on the GMM clustering result with 10 clusters; visualized with Leaflet (OpenStreetMap contributors)

Building on the idea of selecting a few airport markets, one potential approach is to manually choose some markets of interest, and then plot the feature values in multiple bar charts and to analyze why markets are in the same cluster (or why they are not). As an example, 15 non-directional airport markets are selected that connect Vienna to different places around the world (short-haul and long-haul markets, see list A.3.1). Two of these selected markets are in the cluster 0: TFS-VIE (Tenerife Airport to Vienna Airport), and NRT-VIE (Narita Airport to Vienna Airport). As mentioned above, when analyzing the values of the importance of features, the difference in the mean capacity between a weekday and a day of the weekend is the main characteristic driving the segmentation of the markets. For this feature, both TFS-VIE and NRT-VIE have very similar values (8.19). Other features with similar values for TFS-VIE and NRT-VIE are the difference in the number of hotels within 10 km and 50 km around the airport, the standard deviation of the weekly capacity, the difference in GDP per capita (2023) and the difference in elevation. This might indicate that both markets have a similar touristic interest, more capacity on weekends, suggesting rather leisure-related travel. Looking at the feature

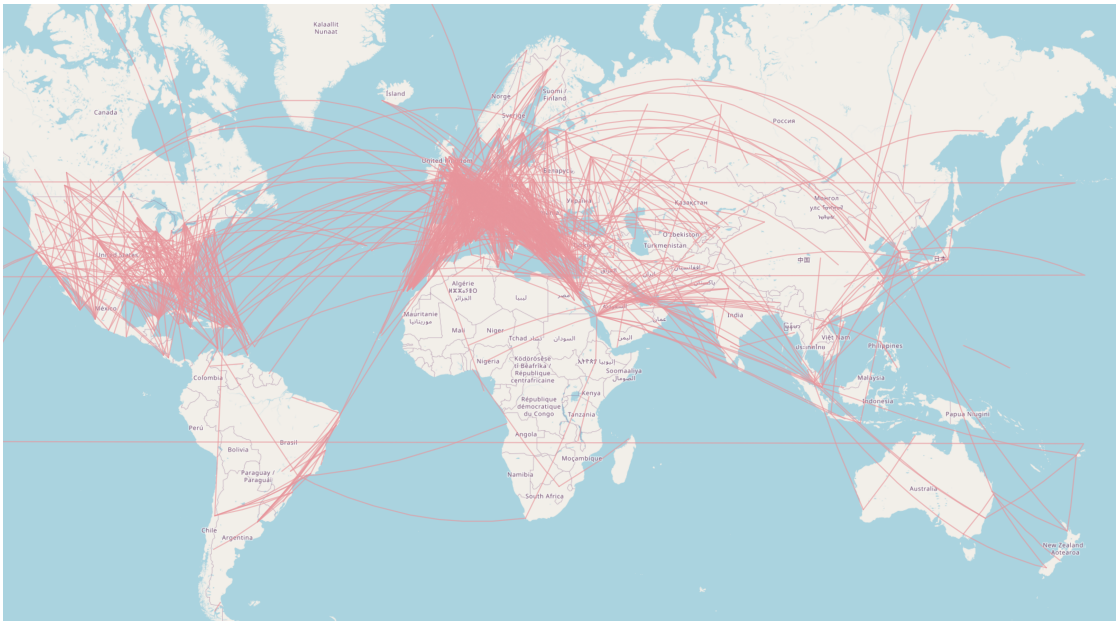

Figure 8.18.: Visualization of clustering result for airport markets in a world map that are part of the cluster 7; markets are represented as lines (using great-circle distance); based on the GMM clustering result with 10 clusters; visualized with Leaflet (OpenStreetMap contributors); Note: some markets are not correctly visualized as this is a 2D visualization (for example, some lines go across the continents instead of the Pacific Ocean)

values for the UMAP reduced components, all of them have similar values for TFS-VIE and NRT-VIE but they cannot be interpreted. For more details, see the full feature matrix for the selected markets in the appendix (A.4).

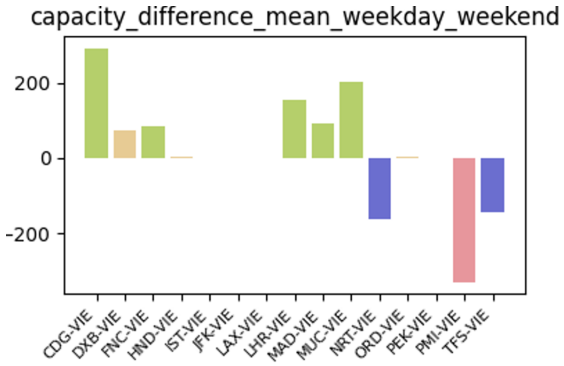

Figure 8.19.: Visualization of feature values for the mean capacity between a weekday and a day of the weekend for manually selected markets related to Vienna Airport; based on the GMM clustering result with 10 clusters

The described interpretative methods offer a possible framework for conducting a more detailed analysis of the clustering results. The primary aim of this thesis is to establish technically significant clusters, with the comprehensive analysis intended for aviation specialists.

8.8. Computational Time

The execution time is not a critical factor for this cluster analysis because clustering tasks are usually not performed on a frequent basis, and computational duration is only a secondary consideration. Nevertheless, an overview of the time required to achieve clustering results can be useful. Completing the full clustering pipeline from feature extraction and transformations to clustering and visualizing results, typically takes a few minutes, depending on the chosen techniques and the cached intermediate results such as the reduced features. The most computationally demanding steps are dimensionality reduction and visualization, both typically handled using UMAP (PCA is a bit faster). Generally, the clustering process itself is quite short, in the cases of a large number of clusters (such as in some GMM scenarios), the process might take slightly longer.

9. Conclusion

This thesis demonstrates how clustering can group over 30,000 airport-pair markets worldwide using supply-side data (flight capacity, seasonality) in addition to demographic, tourism, and climate features. Five different algorithms (K-means++, DBSCAN, HDBSCAN, OPTICS and Gaussian Mixture Model) are compared with internal validation scores and UMAP plots, revealing both balanced global segments and niche, imbalanced ones. Different interpretation approaches are presented, via cluster representatives and feature comparisons, world map visualizations, and a random forest classifier that highlights the features that drive the market segmentation.

9.1. Discussion

The systematic evaluation of clustering techniques confirms that multiple algorithms can yield technically sound market segmentation, but differ in cluster granularity and quality. Regarding the primary research question "*Which clustering technique(s) are effective for segmenting aviation markets?*", various aspects can be observed.

The best clustering result is obtained with 500 clusters and the Gaussian Mixture Model (Silhouette Coefficient of 0.87, Davies-Bouldin Index of 0.2). Overall, the cluster results using GMM yield relatively balanced clusters and achieved good internal validation scores (for Silhouette Coefficient in the range between 0.55 and 0.87 and for Davies-Bouldin Index between 0.7 and 0.2). GMM seems to be a strong choice for this dataset. Moreover, soft cluster assignments have the potential to represent real market segments, particularly in cases where markets are positioned near the decision boundaries separating clusters. Based on previous work, expectation maximization approaches have been used in the analysis of different aviation datasets (Zhou et al. 2020; Suh and Ryerson 2019).

For density-based algorithms, clustering results usually fall into one of two distinct categories: either a small number of clusters with minimal noise points, or numerous clusters accompanied by a high number of noise points. HDBSCAN achieves the technically best cluster results from the density-based algorithms. One result yields 4 clusters with no noise points, a Silhouette Coefficient of 0.68 and a Davies-Bouldin Index of 0.32. In contrast, another configuration results in 773 clusters, accompanied by 15,000 noise points, a Silhouette Coefficient of 0.75, and a Davies-Bouldin Index of 0.12. However, the presence of these 15,000 unassigned markets is probably an issue for further analysis (as roughly half of the markets are noise points).

Although the features are not spherically shaped, K-means achieves decent results in terms of validation metrics (for 10 clusters, a Silhouette Coefficient of 0.66 and a Davies-Bouldin Index of 0.49). One significant benefit of K-means is that it is widely known and easier to understand compared to other algorithms. Moreover, centroids offer a straightforward way to interpret the clusters, though it is important to note that a single data point may not be a perfect representation of an entire cluster, particularly when clusters are large and have irregular shapes. The results of the K-means clustering are to some extent similar (more balanced than density-based clusters) to the GMM results, which is expected because GMM generalizes K-means (Fraley and Raftery 2002).

Addressing the second research question "*Can markets with growth potential be identified through clustering results?*", it can be concluded that supply-side clusters highlight structural market differences, which can serve as indicators of growth opportunities. However, without direct demand or fare data, definitive identification of growth markets remains speculative. Consequently, these clustering results can help find market opportunities when combined with granular demand data.

9.2. Limitations

An essential aspect of cluster analysis is its objective. Based on this, different data sources, features, data transformations, and clustering techniques are more or less suitable. For example, in the application area of market opportunity identification, potentially a higher number of clusters is desirable, whereas in the case of market classification, a smaller, manageable number of clusters is more useful. In particular, the choice of data

sources and features has a significant impact on the clustering results and, therefore, requires careful consideration. When dealing with a relatively complex aviation dataset, selecting an appropriate method for dimensionality reduction and a suitable visualization technique to display the results are crucial components to achieve effective clustering results. One of the primary difficulties encountered in this thesis arises from the fact that minor adjustments in the features or parameter settings of the techniques used can result in substantially different outcomes and feature representations. Furthermore, the ambiguous nature of clustering, coupled with the challenges of interpreting and validating clusters, makes the analysis even more complex.

Data Completeness and Granularity
Global features (such as population, GDP per capita) are at the country level. Especially for large countries, demographic aspects such as population can vary considerably across different airports and areas. Hotel counts from OpenStreetMap vary in completeness across regions. In addition, point of interest and business-related attributes could improve the clustering results.

Absence of Demand Data
The lack of fare and passenger data limits direct assessment of the growth potential of markets.

Catchment Area Simplification
Markets are defined strictly as airport-pairs, ignoring traveler origination within broader regions (door-to-door perspective as described by J. Mandel and B. Mandel 2021).

Algorithm Sensitivity
Small hyperparameter changes can yield substantially different clusters and feature representations.

Clustering Constraints
Applying and analyzing hierarchical clustering on 30,000 markets is infeasible.

Non-directional Markets
As a simplification, directional markets are consolidated to non-directional markets (for example, VIE-FRA and FRA-VIE to FRA-VIE) because of the lack of features that can distinguish the markets properly.

9.3. Future Directions

Building on the findings of this thesis, the following areas for future research directions are recommended:

Clustering on a More Detailed Subset of the Data
Running the clustering on an airline-specific subset of the data, desirably extended with demand data in combination with hierarchical clustering is an interesting direction for the application of clustering on aviation data.

Refined Catchment Modeling, OD markets
Focusing exclusively on airport pairs overlooks potential details. It might be more insightful to concentrate on cities and regions by considering catchment areas, such as identifying all airports within London as components of London's catchment. This approach emphasizes the actual origins and destinations of passenger journeys. Moreover, at present, the focus is solely on direct markets, derived from the provided flight schedule. By incorporating single and double connections, the analysis can more effectively capture origin-destination markets and the network structure.

Focus on Time Series Features
Instead of only focusing on the data of one year such as the flight schedule of 2024 or the GDP per capita of 2023, deriving features that capture the change over time such as capacity growth can yield interesting insights.

Network Characteristics and Graph-based Methods
Centrality measures such as degree and betweenness provide more information on the structure of the aviation network, such as highly interconnected airport hubs and regions. By integrating these metrics with existing market attributes and graph-based techniques such as spectral clustering, this strategy can improve clustering results.

Interpretable Clustering Methods
Evaluation of other interpretable clustering techniques such as explainable K-Means variants (Hu et al. 2024) can provide more information for interpretation.

Deep Clustering
Implementation and exploration of more recent deep clustering approaches, such as autoencoder-based embedding models, can offer enhanced non-linear representations by also optimizing feature extraction and cluster assignment objectives (Wei et al. 2024).

Expert Interpretation and Validation
Combining internal metrics with qualitative assessments from aviation experts can provide more information on the quality of the cluster results.

9.4. Practical Implications and Final Remarks

The practical value of clustering aviation markets is based on clear objectives, high-quality data, domain expertise, and iterative experimentation. While clustering of 30,000 global non-directional markets is feasible, interpretation poses the greatest challenge. Ultimately, clustering provides a structural framework for market segmentation. These market groups can help airlines and airports prioritize route development, benchmark markets, and allocate resources. Adding demand data and more interpretable methods would make the approach even more powerful. Successful implementation depends on selecting suitable features, adjusting methodological decisions, and ensuring that the results are in harmony with the intended objectives.

Bibliography

Acar, A. Zafer and Selçuk Karabulak (20th Oct. 2015). "Competition between Full Service Network Carriers and Low Cost Carriers in Turkish Airline Market". In: *Procedia - Social and Behavioral Sciences*. 11th International Strategic Management Conference 207, pp. 642–651. ISSN: 1877-0428. DOI: 10.1016/j.sbspro.2015.10.134. URL: https://www.sciencedirect.com/science/article/pii/S1877042815052672 (visited on 12/06/2025).

Airlines - AEROPORTI DI ROMA (2025). AEROPORTI DI ROMA. URL: https://www.adr.it/web/aeroporti-di-roma-en/pax-fco-airlines (visited on 19/06/2025).

Ankerst, Mihael et al. (1st June 1999). "OPTICS: ordering points to identify the clustering structure". In: *SIGMOD Rec.* 28.2, pp. 49–60. ISSN: 0163-5808. DOI: 10.1145/304181.304187. URL: https://dl.acm.org/doi/10.1145/304181.304187 (visited on 13/07/2025).

Awan, Abid (13th Sept. 2023). *The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions*. URL: https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning (visited on 07/06/2025).

Baltazar, Maria Emilia and Jorge Silva (4th July 2018). "Airports Catchment Area Size Definition: a Portuguese Case Study". In: Publisher: @ Air Transport Research Society. URL: http://hdl.handle.net/10400.6/9639 (visited on 28/05/2025).

Banerjee, A. and R.N. Dave (July 2004). "Validating clusters using the Hopkins statistic". In: *2004 IEEE International Conference on Fuzzy Systems*. 2004 IEEE International Conference on Fuzzy Systems. Vol. 1. ISSN: 1098-7584, 149–153 vol.1. DOI: 10.1109/FUZZY.2004.1375706. URL: https://ieeexplore.ieee.org/document/1375706 (visited on 07/06/2025).

Berry, Steven and Panle Jia (2010). "Tracing the Woes: An Empirical Analysis of the Airline Industry". In: *American Economic Journal: Microeconomics* 2.3. Publisher: American Economic Association, pp. 1–43. ISSN: 1945-7669. URL: https://www.jstor.org/stable/25760397 (visited on 12/06/2025).

Berthold, Michael R. et al. (2020). *Guide to Intelligent Data Science: How to Intelligently Make Use of Real Data*. 2nd ed. 2020. Texts in Computer Science. Cham: Springer International Publishing Imprint: Springer. ISBN: 978-3-030-45574-3. URL: https://doi.org/10.1007/978-3-030-45574-3 (visited on 05/04/2025).

Bhadra, Dipasis and Jacqueline Kee (1st Jan. 2008). "Structure and dynamics of the core US air travel markets: A basic empirical analysis of domestic passenger demand". In: *Journal of Air Transport Management* 14.1, pp. 27–39. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2007.11.001. URL: https://www.sciencedirect.com/science/article/pii/S0969699707000993 (visited on 23/04/2025).

Boeing, Geoff (3rd May 2025). "Modeling and Analyzing Urban Networks and Amenities With OSMnx". In: *Geographical Analysis*, gean.70009. ISSN: 0016-7363, 1538-4632. DOI: 10.1111/gean.70009. URL: https://onlinelibrary.wiley.com/doi/10.1111/gean.70009 (visited on 19/06/2025).

Busiest Flight Routes in the World 2024 (2025). URL: https://www.oag.com/busiest-routes-world-2024 (visited on 19/06/2025).

Campello, Ricardo J. G. B., Davoud Moulavi and Joerg Sander (2013). "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer, pp. 160–172. ISBN: 978-3-642-37456-2. DOI: 10.1007/978-3-642-37456-2_14.

Chapman, Pete et al. (2000). "CRISP-DM 1.0: Step-by-step data mining guide". In: *SPSS inc* 9.13, pp. 1–73.

Charles de Gaulle Airport (17th June 2025). In: *Wikipedia*. Page Version ID: 1296117976. URL: https://en.wikipedia.org/w/index.php?title=Charles_de_Gaulle_Airport&oldid=1296117976 (visited on 19/06/2025).

Cosmas, Alex et al. (May 2013). "Market clustering and performance of U.S. OD markets". In: *Journal of Air Transport Management* 28, pp. 20–25. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2012.12.006. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7185465/ (visited on 08/04/2025).

Cristea, Anca D, David Hummels and Brian Roberson (2017). "Estimating the Gains from Liberalizing Services Trade: The Case of Passenger Aviation*". In.

Daocheng Yading Airport (14th Apr. 2025). In: *Wikipedia*. Page Version ID: 1285560472. URL: https://en.wikipedia.org/w/index.php?title=Daocheng_Yading_Airport&oldid=1285560472 (visited on 19/06/2025).

Davies, David L. and Donald W. Bouldin (1979). "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 224–227. DOI: 10.1109/TPAMI.1979.4766909.

Doganis, Rigas (2006). *The Airline Business*. Google-Books-ID: J2jOHCFK0YkC. Routledge. 330 pp. ISBN: 978-0-415-34614-6.

Dolnicar, Sara (12th July 2002). "A Review of Data-Driven Market Segmentation in Tourism". In: *Journal of Travel & Tourism Marketing* 12.1. Publisher: Routledge _eprint: https://doi.org/10.1300/J073v12n01_01, pp. 1–22. ISSN: 1054-8408. DOI: 10.1300/J073v12n01_01. URL: https://doi.org/10.1300/J073v12n01_01 (visited on 10/06/2025).

Doshi-Velez, Finale and Been Kim (2nd Mar. 2017). *Towards A Rigorous Science of Interpretable Machine Learning*. DOI: 10.48550/arXiv.1702.08608. arXiv: 1702.08608[stat]. URL: http://arxiv.org/abs/1702.08608 (visited on 07/06/2025).

Efthymiou, Marina and Panayotis Christidis (1st July 2023). "Low-Cost Carriers route network development". In: *Annals of Tourism Research* 101, p. 103608. ISSN: 0160-7383. DOI: 10.1016/j.annals.2023.103608. URL: https://www.sciencedirect.com/science/article/pii/S0160738323000816 (visited on 14/06/2025).

Fang, Mengyuan et al. (10th June 2021). *An adaptive Origin-Destination flows cluster-detecting method to identify urban mobility trends*. DOI: 10.48550/arXiv.2106.05436. arXiv: 2106.05436[cs]. URL: http://arxiv.org/abs/2106.05436 (visited on 11/06/2025).

Fraley, Chris and Adrian E Raftery (1st June 2002). "Model-Based Clustering, Discriminant Analysis, and Density Estimation". In: *Journal of the American Statistical Association* 97.458. Publisher: ASA Website _eprint: https://doi.org/10.1198/016214502760047131, pp. 611–631. ISSN: 0162-1459. DOI: 10.1198/016214502760047131. URL: https://doi.org/10.1198/016214502760047131 (visited on 13/07/2025).

Gama, Joao et al. (2014). "A survey on concept drift adaptation". In: *ACM Computing Surveys* 46.4. ISSN: 0360-0300. DOI: 10.1145/2523813.

Gao, Yi (1st Jan. 2021). "What is the busiest time at an airport? Clustering U.S. hub airports based on passenger movements". In: *Journal of Transport Geography* 90, p. 102931. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2020.102931. URL: https://www.sciencedirect.com/science/article/pii/S0966692320310085 (visited on 20/02/2025).

Geographic, National (2025). *Elevation*. URL: https://education.nationalgeographic.org/resource/elevation (visited on 19/06/2025).

Google (2025). *Elevation API*. Google for Developers. URL: https://developers.google.com/maps/documentation/elevation/overview (visited on 18/06/2025).

Graham, Brian and Claire Guyer (1st Dec. 2000). "The role of regional airports and air services in the United Kingdom". In: *Journal of Transport Geography* 8.4, pp. 249–262.

ISSN: 0966-6923. DOI: 10.1016/S0966-6923(00)00021-1. URL: https://www.sciencedirect.com/science/article/pii/S0966692300000211 (visited on 10/06/2025).

Grant, John (Mar. 2024). *Why Is Average Flight Capacity Increasing? | Aviation Market Analysis*. URL: https://www.oag.com/blog/average-flight-capacity-increasing-at-fastest-rate-ever (visited on 19/06/2025).

Great Circle Mapper (2025). Flight Distance and Duration Calculator - Airport and Aviation Database - Great Circle Mapper. URL: https://www.greatcirclemapper.net/ (visited on 19/06/2025).

Güvercin, Mehmet, Nilgun Ferhatosmanoglu and Bugra Gedik (May 2021). "Forecasting Flight Delays Using Clustered Models Based on Airport Networks". In: *IEEE Transactions on Intelligent Transportation Systems* 22.5. Conference Name: IEEE Transactions on Intelligent Transportation Systems, pp. 3179–3189. ISSN: 1558-0016. DOI: 10.1109/TITS.2020.2990960. URL: https://ieeexplore.ieee.org/document/9091035 (visited on 12/02/2025).

Han, Congshi et al. (Oct. 2022). "Optimization of K-means algorithm and its application in airline industry customer segmentation". In: *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*. 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), pp. 1370–1373. DOI: 10.1109/ICDSCA56264.2022.9988380. URL: https://ieeexplore.ieee.org/document/9988380 (visited on 28/03/2025).

Harrison, Anna, Vesna Popovic and Ben Kraal (12th June 2015). "A new model for airport passenger segmentation". In: *Journal of Vacation Marketing* 21. DOI: 10.1177/1356766715571390.

Hinton, G. E. and R. R. Salakhutdinov (28th July 2006). "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786. Publisher: American Association for the Advancement of Science, pp. 504–507. DOI: 10.1126/science.1127647. URL: https://www.science.org/doi/abs/10.1126/science.1127647 (visited on 07/06/2025).

Hotelling, Harold (1933). "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6, p. 417.

Hu, Lianyu et al. (1st Sept. 2024). *Interpretable Clustering: A Survey*. arXiv.org. URL: https://arxiv.org/abs/2409.00743v1 (visited on 09/06/2025).

Jacob, Afolayan Jimoh, Joseph Daniel and Chikezie Samuel Aneke (2024). "Determination Of Optimal Number Of Clusters Using Gap Statistics And Elbow Methods". In: 9.3.

Jain, A. K., M. N. Murty and P. J. Flynn (1st Sept. 1999). "Data clustering: a review". In: *ACM Comput. Surv.* 31.3, pp. 264–323. ISSN: 0360-0300. DOI: 10.1145/331499.331504. URL: https://dl.acm.org/doi/10.1145/331499.331504 (visited on 07/06/2025).

Kapoor, Sayash and Arvind Narayanan (14th July 2022). *Leakage and the Reproducibility Crisis in ML-based Science*. DOI: 10.48550/arXiv.2207.07048. arXiv: 2207.07048[cs]. URL: http://arxiv.org/abs/2207.07048 (visited on 07/06/2025).

Kotler, Philip and Kevin Lane Keller (2016). *Marketing management*. 15 [edition]. Boston: Pearson. 692 pp. ISBN: 978-0-13-385646-0.

Kruskal, Joseph (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1.

Lee, Seul Ki and SooCheong (Shawn) Jang (1st Mar. 2011). "Room Rates of U.S. Airport Hotels: Examining the Dual Effects of Proximities". In: *Journal of Travel Research* 50.2. Publisher: SAGE Publications Inc, pp. 186–197. ISSN: 0047-2875. DOI: 10.1177/0047287510362778. URL: https://doi.org/10.1177/0047287510362778 (visited on 28/05/2025).

Leisch, Friedrich, Sara Dolnicar and Bettina Grün (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Accepted: 2021-11-05T04:03:34Z. ISBN: 978-981-10-8817-9. URL: https://library.oapen.org/handle/20.500.12657/51281 (visited on 10/06/2025).

Lieshout, Rogier (1st Nov. 2012). "Measuring the size of an airport's catchment area". In: *Journal of Transport Geography*. Special Section on Accessibility and Socio-Economic Activities: Methodological and Empirical Aspects 25, pp. 27–34. ISSN: 0966-6923. DOI: 10.1016/j.jtrangeo.2012.07.004. URL: https://www.sciencedirect.com/science/article/pii/S0966692312001706 (visited on 28/05/2025).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605. ISSN: 1533-7928. URL: http://jmlr.org/papers/v9/vandermaaten08a.html (visited on 07/06/2025).

Maertens, Sven, Wolfgang Grimme and Stephan Bingemer (1st Jan. 2020). "The development of transfer passenger volumes and shares at airport and world region levels". In: *Transportation Research Procedia*. INAIR 2020 - CHALLENGES OF AVIATION DEVELOPMENT 51, pp. 171–178. ISSN: 2352-1465. DOI: 10.1016/j.trpro.2020.11.019. URL: https://www.sciencedirect.com/science/article/pii/S2352146520308723 (visited on 12/06/2025).

Maliha, Maisha and Dean Hougen (July 2024). "Aviation Passenger Segmentation through GANs Integrating K-Means Clustering: Innovating Airline Optimization". In: *NAECON 2024 - IEEE National Aerospace and Electronics Conference*. NAECON 2024 - IEEE National Aerospace and Electronics Conference. ISSN: 2379-2027, pp. 194–199. DOI: 10.1109/NAECON61878.2024.10670657. URL: https://ieeexplore.ieee.org/document/10670657 (visited on 09/06/2025).

Mandel, Julian and Benedikt Mandel (9th June 2021). "The Hammer & Dance process based on Door-to-Door modelling". In: *Airlines in a Post-Pandemic World: Preparing for Constant Turbulence Ahead*. Vol. 1. ISBN: 978-0-367-71582-3.

Masson-Delmotte, Valérie et al., eds. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157896.

McInnes, Leland (2018). *Using UMAP for Clustering — umap 0.5.8 documentation*. URL: https://umap-learn.readthedocs.io/en/latest/clustering.html (visited on 08/07/2025).

McInnes, Leland, John Healy and James Melville (9th Feb. 2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. DOI: 10.48550/arXiv.1802.03426. arXiv: 1802.03426[stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 07/06/2025).

Nanga, Salifu et al. (8th July 2021). "Review of Dimension Reduction Methods". In: *Journal of Data Analysis and Information Processing* 9.3. Number: 3 Publisher: Scientific Research Publishing, pp. 189–231. DOI: 10.4236/jdaip.2021.93013. URL: https://www.scirp.org/journal/paperinformation?paperid=111638 (visited on 13/06/2025).

National Centers for Environmental Information (NCEI) (2025). URL: https://www.ncei.noaa.gov/ (visited on 19/06/2025).

National Transportation Statistics (NTS) (2019). DOI: 10.21949/1503663. URL: https://tinyurl.com/rosapntlbtsNTS (visited on 18/06/2025).

OECD (11th Jan. 2010). *Globalisation, Transport and the Environment*. OECD. URL: https://www.oecd.org/en/publications/globalisation-transport-and-the-environment_9789264072916-en.html (visited on 12/06/2025).

OpenStreetMap (2025). URL: https://openstreetmapjl.readthedocs.io/en/stable/overview.html (visited on 18/06/2025).

OurAirports (2025). URL: https://ourairports.com/about.html (visited on 17/06/2025).

Park, Hae-Sang and Chi-Hyuck Jun (1st Mar. 2009). "A simple and fast algorithm for K-medoids clustering". In: *Expert Systems with Applications* 36.2, pp. 3336–3341. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2008.01.039. URL: https://www.sciencedirect.com/science/article/pii/S095741740800081X (visited on 14/07/2025).

Pearson, Karl (1901). "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.

Postorino, Maria Nadia and Mario Versaci (2014). "A Geometric Fuzzy-Based Approach for Airport Clustering". In: *Advances in Fuzzy Systems* 2014, pp. 1–12. ISSN: 1687-7101, 1687-711X. DOI: 10.1155/2014/201243. URL: http://www.hindawi.com/journals/afs/2014/201243/ (visited on 20/02/2025).

Prabhakar, Nikhilesh and L. Jani Anbarasi (8th Mar. 2021). "Exploration of the global air transport network using social network analysis". In: *Social Network Analysis and Mining* 11.1, p. 26. ISSN: 1869-5469. DOI: 10.1007/s13278-021-00735-1. URL: https://doi.org/10.1007/s13278-021-00735-1 (visited on 20/02/2025).

Proussaloglou, Kimon and Frank S. Koppelman (1st Oct. 1999). "The choice of air carrier, flight, and fare class". In: *Journal of Air Transport Management* 5.4, pp. 193–201. ISSN: 0969-6997. DOI: 10.1016/S0969-6997(99)00013-7. URL: https://www.sciencedirect.com/science/article/pii/S0969699799000137 (visited on 14/06/2025).

Provost, Foster and Tom Fawcett (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, UNITED STATES: O'Reilly Media, Incorporated. ISBN: 978-1-4493-7429-7. URL: http://ebookcentral.proquest.com/lib/univie/detail.action?docID=1323973 (visited on 06/06/2025).

PyTorch (2025). PyTorch. URL: https://pytorch.org/ (visited on 08/07/2025).

Robles, Fernando and Ravi Sarathy (1st Feb. 1986). "Segmenting the commuter aircraft market with cluster analysis". In: *Industrial Marketing Management* 15.1, pp. 1–12. ISSN: 0019-8501. DOI: 10.1016/0019-8501(86)90039-8. URL: https://www.sciencedirect.com/science/article/pii/0019850186900398 (visited on 09/06/2025).

Rousseeuw, Peter J. (1st Nov. 1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257 (visited on 15/07/2025).

Sambandam, R. (Mar. 2003). "Cluster Analysis Gets Complicated". In: *Marketing Research* 15, pp. 16–21.

scikit-learn 1.7.0 documentation (2025). URL: https://scikit-learn.org/stable/ (visited on 07/06/2025).

Smith, Wendell R. (1st July 1956). "Product Differentiation and Market Segmentation as Alternative Marketing Strategies". In: *Journal of Marketing* 21.1. Publisher: SAGE Publications Inc, pp. 3–8. ISSN: 0022-2429. DOI: 10.1177/002224295602100102. URL: https://doi.org/10.1177/002224295602100102 (visited on 10/06/2025).

Suh, Daniel Y. and Megan S. Ryerson (1st Aug. 2019). "Forecast to grow: Aviation demand forecasting in an era of demand uncertainty and optimism bias". In: *Transportation*

Research Part E: Logistics and Transportation Review 128, pp. 400–416. ISSN: 1366-5545. DOI: 10.1016/j.tre.2019.06.016. URL: https://www.sciencedirect.com/science/article/pii/S1366554518314947 (visited on 20/02/2025).

Svirsky, Jonathan and Ofir Lindenbaum (8th July 2024). "Interpretable Deep Clustering for Tabular Data". In: *Proceedings of the 41st International Conference on Machine Learning*. International Conference on Machine Learning. ISSN: 2640-3498. PMLR, pp. 47314–47330. URL: https://proceedings.mlr.press/v235/svirsky24a.html (visited on 26/05/2025).

Tan, Pang-Ning et al. (2019). *Introduction to data mining*. Second edition, global edition. NY NY, [Ann Arbor]: Pearson, ProQuest Ebook Central. ISBN: 978-0-273-77532-4.

Teichert, Thorsten, Edlira Shehu and Iwan von Wartburg (1st Jan. 2008). "Customer segmentation revisited: The case of the airline industry". In: *Transportation Research Part A: Policy and Practice* 42.1, pp. 227–242. ISSN: 0965-8564. DOI: 10.1016/j.tra.2007.08.003. URL: https://www.sciencedirect.com/science/article/pii/S0965856407000729 (visited on 09/06/2025).

TerraClimate (2025). *TerraClimate*. Climatology Lab. URL: https://www.climatologylab.org/terraclimate.html (visited on 18/06/2025).

Thorndike, Robert L. (Dec. 1953). "Who Belongs in the Family?" In: *Psychometrika* 18.4, pp. 267–276. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02289263. URL: https://www.cambridge.org/core/product/identifier/S0033312300048730/type/journal_article (visited on 13/06/2025).

Tibshirani, Robert, Guenther Walther and Trevor Hastie (2001). "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63.2. Publisher: [Royal Statistical Society, Oxford University Press], pp. 411–423. ISSN: 1369-7412. URL: https://www.jstor.org/stable/2680607 (visited on 13/06/2025).

Torgerson, Warren (1952). "Multidimensional scaling: I. Theory and method". In: *Psychometrika* 17.4.

Üstebey, Serpil, İlkay Yelmen and Metin Zontul (2020). "Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers". In: Accepted: 2023-04-06T05:43:57Z Publisher: Dr. Öğr. Üyesi Fatma Kutlu Gündoğdu. ISSN: 1304-0448. URL: http://arelarsiv.arel.edu.tr/xmlui/handle/20.500.12294/3695 (visited on 09/06/2025).

Venkat, Naveen (7th Sept. 2018). *The Curse of Dimensionality: Inside Out*. DOI: 10.13140/RG.2.2.29631.36006.

Wang, Xu and Yusheng Xu (July 2019). "An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index". In: *IOP Conference Series: Materials Science and Engineering* 569.5. Publisher: IOP Publishing, p. 052024. ISSN: 1757-899X. DOI: 10.1088/1757-899X/569/5/052024. URL: https://dx.doi.org/10.1088/1757-899X/569/5/052024 (visited on 13/06/2025).

Wei, Xiuxi et al. (14th July 2024). "An overview on deep clustering". In: *Neurocomputing* 590, p. 127761. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2024.127761. URL: https://www.sciencedirect.com/science/article/pii/S0925231224005320 (visited on 07/06/2025).

Weller, Bridget E., Natasha K. Bowen and Sarah J. Faubert (1st May 2020). "Latent Class Analysis: A Guide to Best Practice". In: *Journal of Black Psychology* 46.4. Publisher: SAGE Publications Inc, pp. 287–311. ISSN: 0095-7984. DOI: 10.1177/0095798420930932. URL: https://doi.org/10.1177/0095798420930932 (visited on 10/06/2025).

World Bank Group - International Development, Poverty and Sustainability (2025). URL: https://www.worldbank.org/ext/en/home (visited on 18/06/2025).

Wu, Yujuan and Paul D Berger (Mar. 2018). "A CLUSTER ANALYSIS APPROACH TO MARKET SEGMENTATION IN THE AIRLINES INDUSTRY". In: *International Journal of Management Information Technology and Engineering* 6.3, pp. 17–26.

Xu, Dongkuan and Yingjie Tian (1st June 2015). "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2, pp. 165–193. ISSN: 2198-5812. DOI: 10.1007/s40745-015-0040-1. URL: https://doi.org/10.1007/s40745-015-0040-1 (visited on 07/04/2025).

Xu, Rui and D. Wunsch (May 2005). "Survey of clustering algorithms". In: *IEEE Transactions on Neural Networks* 16.3, pp. 645–678. ISSN: 1941-0093. DOI: 10.1109/TNN.2005.845141. URL: https://ieeexplore.ieee.org/document/1427769 (visited on 07/06/2025).

Yang, Fan and Xiujuan Yang (22nd Oct. 2017). "An Aviation Market Hierarchical Cluster Model Considering Dimension Reduction of Airline Route Network Performance". In: *Proceedings of the 1st International Conference on Big Data Research*. ICBDR '17. New York, NY, USA: Association for Computing Machinery, pp. 98–103. ISBN: 978-1-4503-5356-4. DOI: 10.1145/3152723.3152737. URL: https://doi.org/10.1145/3152723.3152737 (visited on 23/02/2025).

Yang, Haoyu, Lianmeng Jiao and Quan Pan (July 2021). "A Survey on Interpretable Clustering". In: *2021 40th Chinese Control Conference (CCC)*. 2021 40th Chinese Control Conference (CCC). ISSN: 1934-1768, pp. 7384–7388. DOI: 10.23919/CCC52363.

2021.9549986. URL: https://ieeexplore.ieee.org/document/9549986 (visited on 09/06/2025).

YData data quality for Data Science (2025). URL: https://ydata.ai (visited on 19/06/2025).

Yelmen, İlkay, Serpil Üstebey and Metin Zontul (28th July 2020). "Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers". In: *Journal of Aeronautics and Space Technologies* 13.2. Section: Articles, pp. 227–233. URL: https://jast.hho.msu.edu.tr/index.php/JAST/article/view/430 (visited on 11/06/2025).

Yeo, In-Kwon and Richard A. Johnson (1st Dec. 2000). "A new family of power transformations to improve normality or symmetry". In: *Biometrika* 87.4, pp. 954–959. ISSN: 0006-3444. DOI: 10.1093/biomet/87.4.954. URL: https://doi.org/10.1093/biomet/87.4.954 (visited on 17/06/2025).

Yin, Lifeng et al. (Jan. 2023). "Improvement of DBSCAN Algorithm Based on K-Dist Graph for Adaptive Determining Parameters". In: *Electronics* 12.15. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, p. 3213. ISSN: 2079-9292. DOI: 10.3390/electronics12153213. URL: https://www.mdpi.com/2079-9292/12/15/3213 (visited on 10/07/2025).

Zheng, Xufang et al. (22nd Feb. 2025). *Examining the Dynamics of Local and Transfer Passenger Share Patterns in Air Transportation*. arXiv.org. URL: https://arxiv.org/abs/2503.05754v1 (visited on 12/06/2025).

Zhou, Heng et al. (2020). "Market segmentation approach to investigate existing and potential aviation markets". In: *Transport Policy* 99 (C). Publisher: Elsevier, pp. 120–135. URL: https://ideas.repec.org//a/eee/trapol/v99y2020icp120-135.html (visited on 09/06/2025).

A. Appendix

The appendix includes additional visualizations of the data exploration phase, more details on the experimental clustering runs, additional information on the interpretation of the clustering results, and also a reference to the code.

A.1. Data Exploration

A.1.1. YData Profiling Report

The YData profiling report is in the directory that contains the code.

A.1.2. Correlation Matrix Initial Features

The following correlation matrices are barely visible in terms of the feature correlations. However, they illustrate the number of highly correlated features. This feature version is referred to as "Temp/Cap wide".

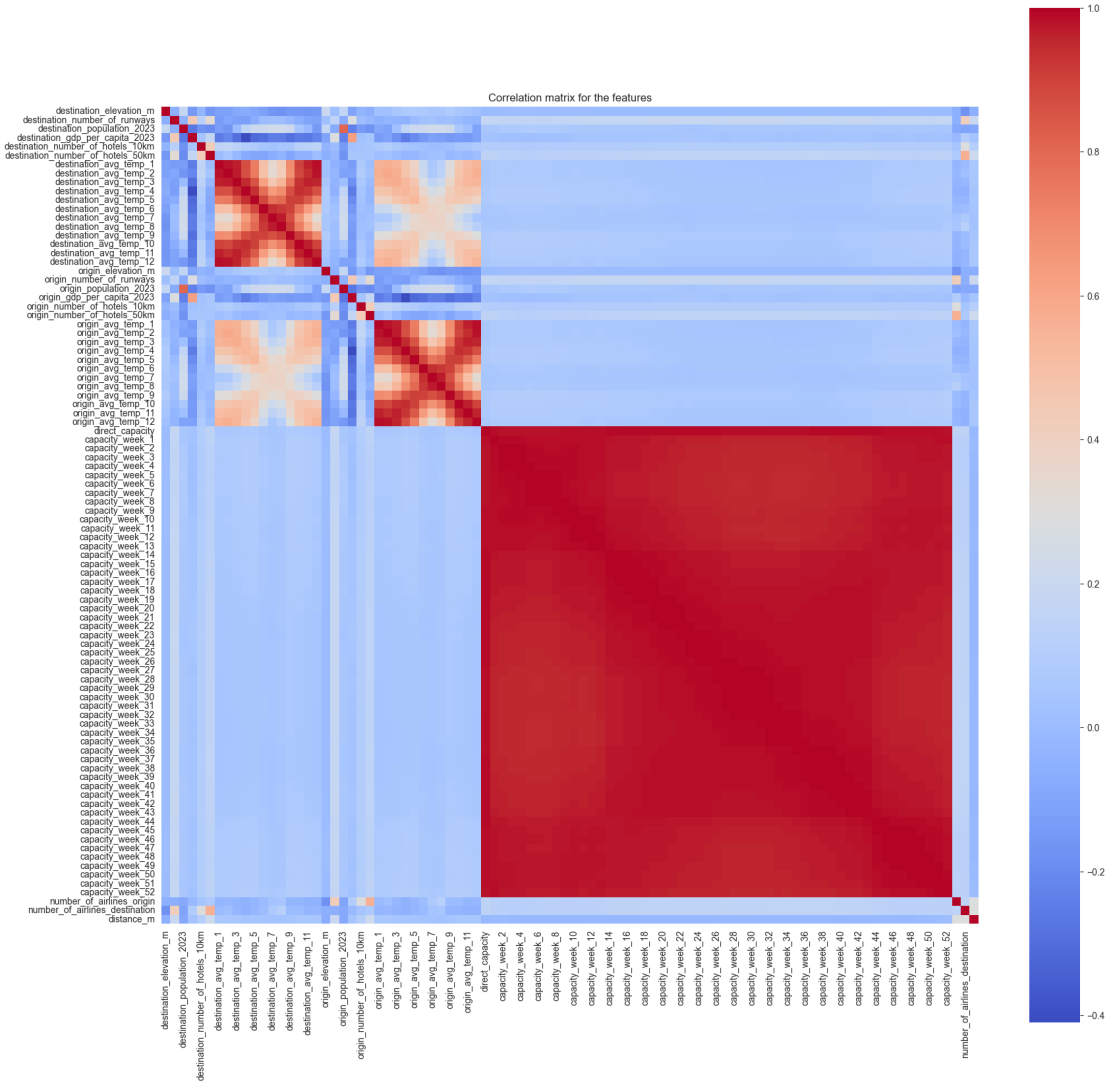

Figure A.1.: Correlation matrix of initial features

A.1.3. Correlation Matrix Intermediate Features

This version has statistically derived variables for temperature and capacity, but uses airport-related features instead of the differences (as in the refined feature list). This feature version is referred to as "Temp/Cap v2". Another intermediate feature version ("Temp/Cap PCA") applies PCA block-wise (meaning on the capacity and temperature features individually).

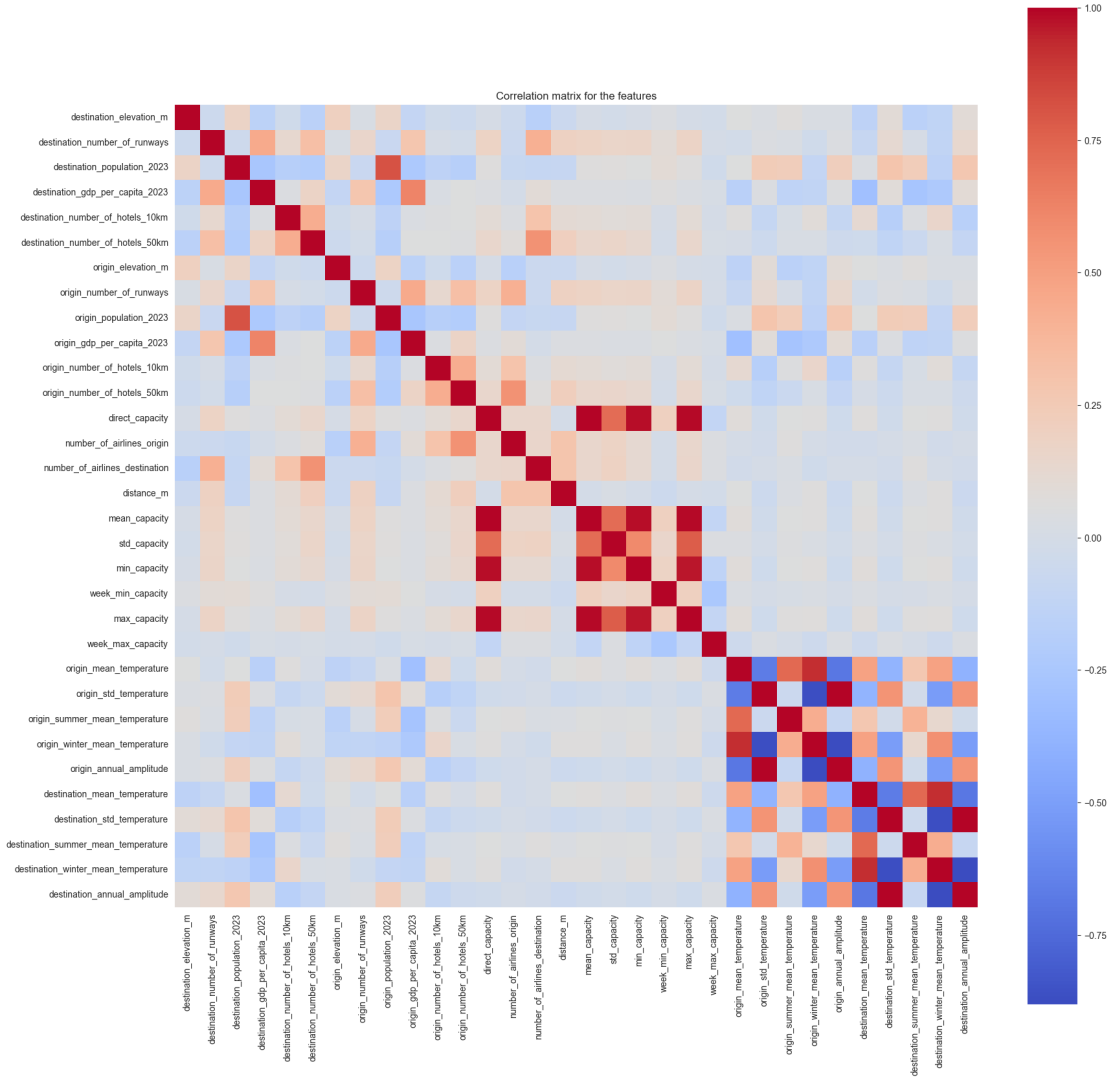

Figure A.2.: Correlation matrix of intermediate features (statistically derived features for temperature and capacity)

A.1.4. Scatterplot Matrix of Refined Features

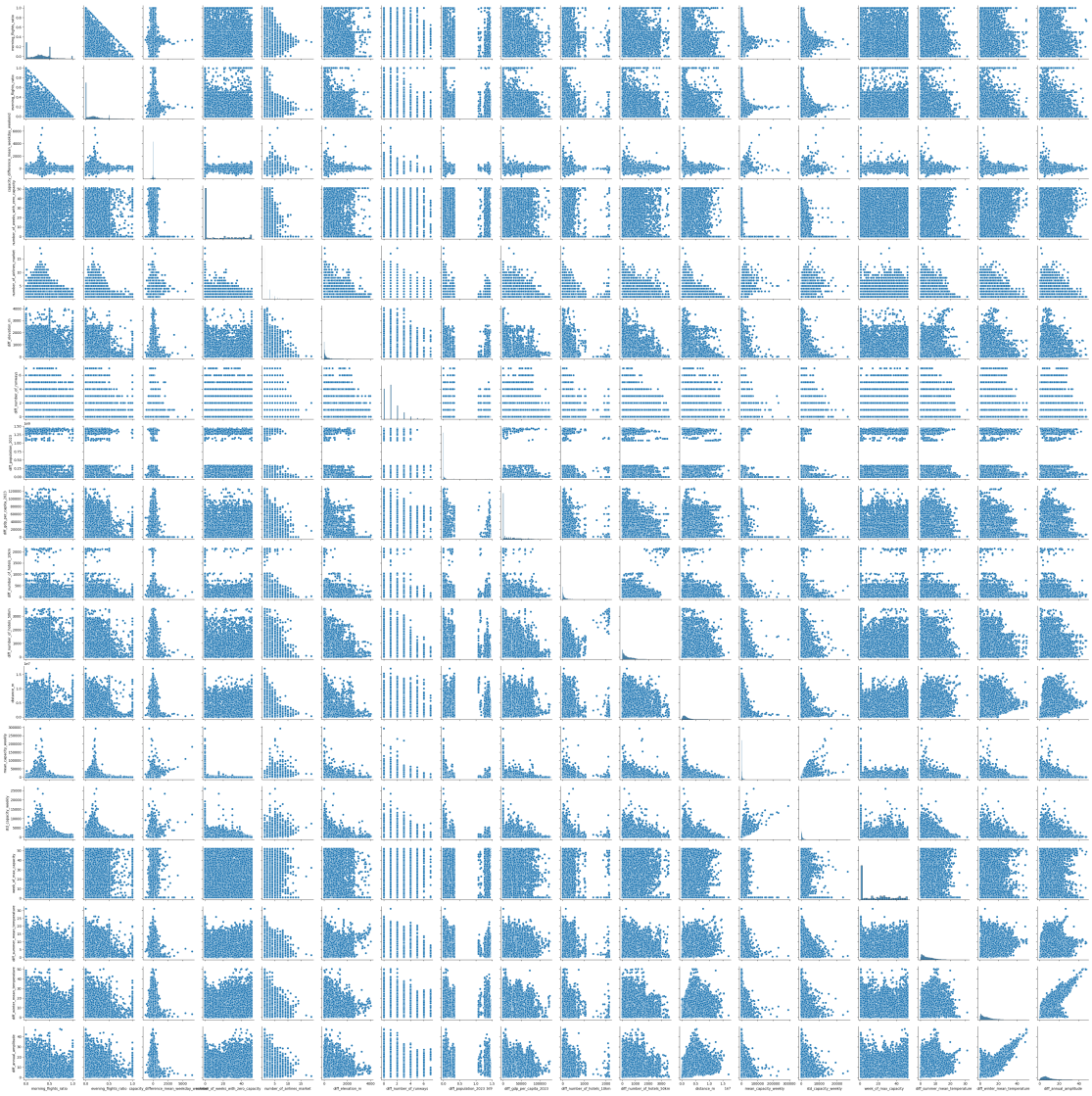

Figure A.3.: Scatterplot matrix of refined features, demonstrating bi-variate feature relationships, mostly non-linear

A.2. Clustering

A.2.1. Experimental Clustering Runs

Experimental Clustering Runs (Part 1)

Algorithm	Clusters	Noise Points	Silhouette | DB Index	Settings
K-Means				
K-Means	7	0	0.10 | 2.46	Temp/Cap v2, Autoencoder, standard scaling, $k = 7$
K-Means	10	0	0.17 | 1.33	Temp/Cap wide, no DR, standard scaling, $k = 10$
K-Means	15	0	0.20 | 1.38	Temp/Cap wide, PCA, standard scaling, $k = 15$
K-Means	15	0	0.72 | 0.39	UMAP (10 dims), Yeo–Johnson, robust scaling, $k = 15$
DBSCAN				
DBSCAN	1	10	0.89 | 0.30	Temp/Cap PCA (90% var.), standard scaling, $\varepsilon = 25$, minPts=34
DBSCAN	2	349	0.61 | 1.30	Temp/Cap v2, PCA (90% var.), standard scaling, $\varepsilon = 8$, minPts=25
DBSCAN	4	247	0.63 | 1.17	UMAP (10 dims), Yeo–Johnson, robust scaling, $\varepsilon = 0.75$, minPts=20
DBSCAN	21	2,461	0.48 | 0.58	UMAP (10 dims), Yeo–Johnson, robust scaling, $\varepsilon = 0.50$, minPts=15

DB Index ... Davies–Bouldin index; k denotes number of clusters/components; ε and minPts (for DBSCAN) are shown under "Settings"

Table A.1.: Summary of detailed clustering runs across algorithms (first part).

Experimental Clustering Runs (Part 2)

Algorithm	Clusters	Noise Points	Silhouette | DB Index	Settings
OPTICS				
OPTICS	266	23,637	0.05 | 0.16	UMAP (10 dims), minimum samples=10, $\xi = 0.05$
HDBSCAN				
HDBSCAN	773	14,958	0.75 | 0.12	no power transform, UMAP (10 dims), minimum cluster size=5, method=eom
HDBSCAN	4	0	0.68 | 0.31	Yeo–Johnson, robust scaling, UMAP (15 dims), minimum cluster size=20, method=eom
HDBSCAN	13	622	0.65 | 0.24	Yeo–Johnson, robust scaling, UMAP (8 dims), minimum cluster size=20, method=eom
HDBSCAN	12	678	0.64 | 0.25	Yeo–Johnson, robust scaling, UMAP (5 dims), minimum cluster size=20, method=eom
HDBSCAN	152	18,625	0.63 | 0.19	no power transform, UMAP (10 dims), minimum cluster size=20, method=leaf
Gaussian Mixture Model				
GMM	7	0	0.02 | 2.51	PCA (10 dims), Yeo–Johnson, robust scaling, $k = 7$
GMM	25	0	0.74 | 0.41	UMAP (10 dims), Yeo–Johnson, robust scaling, $k = 25$
GMM	50	0	0.76 | 0.44	UMAP (10 dims), Yeo–Johnson, robust scaling, $k = 50$
GMM	300	0	0.85 | 0.23	UMAP (10 dims), Yeo–Johnson, robust scaling, $k = 300$
GMM	500	0	0.87 | 0.20	UMAP (10 dims), Yeo–Johnson, robust scaling, $k = 500$

min_samples/ξ (for OPTICS) and minimum cluster size/method (for HDBSCAN) are shown under "Settings"

Table A.2.: Summary of detailed clustering runs across algorithms (second part).

A.3. Interpretation of Clustering

A.3.1. Manually Selected Airports Related to Vienna Airport

This list of manually selected airports is used for cluster interpretation purposes.

- Frankfurt am Main Airport (FRA) ↔ Vienna International Airport (VIE)
- Munich Airport (MUC) ↔ VIE
- London Heathrow Airport (LHR) ↔ VIE
- Paris–Charles de Gaulle Airport (CDG) ↔ VIE
- Cristiano Ronaldo Madeira International Airport (FNC) ↔ VIE
- Tenerife South–Reina Sofía Airport (TFS) ↔ VIE
- Adolfo Suárez Madrid–Barajas Airport (MAD) ↔ VIE
- Palma de Mallorca Airport (PMI) ↔ VIE
- John F. Kennedy International Airport (JFK) ↔ VIE
- Los Angeles International Airport (LAX) ↔ VIE
- Chicago O'Hare International Airport (ORD) ↔ VIE
- Tokyo Haneda Airport (HND) ↔ VIE
- Narita International Airport (NRT) ↔ VIE
- Beijing Capital International Airport (PEK) ↔ VIE
- Dubai International Airport (DXB) ↔ VIE
- Istanbul Airport (IST) ↔ VIE

A.3.2. Feature Matrix of Manually Selected Airports Related to Vienna Airport

Figure A.4.: Visualization of feature bar charts of refined features and UMAP reduced components for manually selected markets related to Vienna Airport; based on the GMM clustering result with 10 clusters

A.4. Code

A detailed description of the code structure can be found in the README.md file (within the code directory). Additionally, there are even more clustering results in the results folder in the code directory than mentioned in the thesis. Most of them are used to find suitable parameters.