DATA MINING PROJECT

Master in Data Science and Advanced Analytics

**NOVA Information Management School**

Universidade Nova de Lisboa

# Amazing International Airlines Inc.

## Clustering

**Group 4**

Pedro Santos, 20250399

Miguel Correia, 20250381

Pedro Fernandes, 20250418

Tiago Duarte, 20250360

Fall Semester 2025-2025

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1. EXECUTIVE SUMMARY

This project delivers a data-driven customer segmentation for AIAI that combines behavioral and value-based perspectives to support actionable business decisions. The analysis shows that customer behavioral and engagement patterns, rather than demographics, are the primary drivers of differentiation, while value metrics refine strategic prioritization.

Four segments were identified. The Established Active Core forms the revenue backbone and should be protected through retention and value expansion. The Emerging High-Growth segment, though smaller, shows rapid engagement acceleration and represents the highest future ROI, justifying proactive investment. Seasonal Solo Travelers display peak-period activity with limited off-season engagement and should be addressed through targeted activation with controlled spend. The At-Risk Low-Engagement segment presents elevated churn risk and requires selective retention of high-value customers while minimizing investment in low-value dormant accounts.

Overall, the segmentation provides a clear prioritization framework that enables AIAI to allocate marketing resources more effectively, strengthen loyalty, and drive sustainable long-term value.

# 2. INTRODUCTION

This section connects the findings from Phase 1 (Exploratory Data Analysis) with the objectives of the clustering phase, while also noting that some data preparation steps were performed during Phase 2 as part of the modeling process. It documents the treatment of inconsistencies (strange values), duplicated values, missing values and outliers, as well as the feature engineering and scaling decisions, ensuring the dataset is appropriate for effective and reliable clustering. Correcting data types is not going to be included in the report (these were already explored in deliverable 1 report).

## 2.1. Strange Values

### 2.1.1. Income

Data visualization (Figure 3) showed that about 25% of records reported zero income, initially assumed to reflect customers without income. Further analysis indicated these were predominantly adult customers with college education, suggesting zero values were anomalous rather than genuine. These values were replaced with NaN and two imputation methods were tested. As Group Median distorted the distribution, KNN was selected for providing more realistic results (Figure 4). The 20 originally missing income values were imputed using the same approach for consistency.

### 2.1.2. CancellationDate before EnrollmentDateOpening

A consistency check identified records where CancellationDate precedes EnrollmentDateOpening, which is logically invalid. Among records with both dates available, 199 cases (1.1%) were found and their CancellationDate was set to NaN rather than removing the entire record, preserving customer information while preventing invalid dates from affecting the analysis.

### 2.1.3. Points Redeemed > Points Accumulated

A total of 462 customers (2.76%) presented TotalPointsRedeemed greater than TotalPointsAccumulated (Figure 5), which is logically impossible at the aggregate level. Although this may occur in individual months, it should not happen overall in totals. To correct this inconsistency, TotalPointsRedeemed was capped at TotalPointsAccumulated, ensuring that RedeemRatio (Feature created in Feature Engineering) ≤ 1 while preserving all other customer information.

## 2.2. Duplicated Values

### 2.2.1. Duplicated Loyalty#

We identified numerous duplicated Loyalty# entries (164 - equivalent to 0.97%) in the CustomerDB dataset. Our deduplication strategy prioritized active customers (those with null CancellationDate values). Among active customers, we selected records with complete information, and in case of ties, the entry with the most recent opening date was chosen.

### 2.2.2. Duplicated entries in FlightsBD

Duplicate analysis identified 2,903 records (0.48%) with repeated combinations of Loyalty#, Year, and Month. These duplicates were removed using this composite key, ensuring each customer–time combination appears only once and preserving the integrity of the temporal analysis.

## 2.3. Missing Values

### 2.3.1. Customer Lifetime Value

Analysis of the FlightsDB dataset showed that 20 customers (0.12%) have no associated flight records, corresponding exactly to the 20 missing CLV values. As CLV is derived from flight activity, these values were imputed with 0, representing a valid absence of value rather than missing data.

### 2.3.2. CancellationDate

The CancellationDate variable contains 14,490 missing values (86.35%), corresponding to customers who have not cancelled their membership. Instead of imputing these values, the variable was used to engineer IsActive and Tenure features. Afterward, CancellationDate was removed, as its information was fully captured by the derived variables.

### 2.3.3. Point-related Features

Missing value analysis showed that 20 records (0.12%) had missing values in TotalPointsAccumulated and TotalPointsRedeemed, corresponding to customers with no flight activity. As no behavioral information was available for these cases (and we don’t want customers without activity), the records were removed, resulting in a negligible reduction of the dataset from 16,757 to 16,737 records.

The RedeemRatio variable contained 1,518 missing values (~9%), corresponding to customers who accumulated points but never redeemed any. Since a value of zero represents a valid non-redemption behavior, these cases were imputed with 0, preserving interpretability.

## 2.4. Univariate Outliers

The Interquartile Range (IQR) method was used to identify outliers, as it is robust to extreme values and suitable for non-normal distributions. Following (Han et al., 2023), observations beyond Q1 − 1.5 × IQR and Q3 + 1.5 × IQR were classified as suspected outliers, a criterion analogous to the 3σ rule for normal distributions.

Outliers were mainly concentrated on Customer Lifetime Value (8.9%), TotalPointsRedeemed (1.0%), and TotalFlightsWithCompanions (0.1%), while temporal features showed negligible extremes. In total, 9.9% of records contained at least one outlier. Rather than removing approximately 10% of the data (this would be a lot), winsorization was applied to cap extreme values at IQR boundaries, preserving all 16,737 records for clustering analysis. Final boxplots can be seen in Figure 6,

## 2.5. Feature Engineering

Table 1 - Feature Engineering

| Feature Name | Circumstance | Why? | Formula / Logic | Reason for Clustering |
|---|---|---|---|---|
| IsActive | New feature (CustomerDB) | Identify active memberships | True if CancellationDate is NaN; False otherwise | Separates active vs churned customers |
| Tenure | New feature (CustomerDB) | Measure loyalty duration | (CancellationDate or 31-12-2021 − EnrollmentDateOpening) / 365 | Captures customer maturity |
| TotalFlights | Merge Preparation | Total flight activity | Sum of NumFlights per customer | Measures engagement |
| TotalFlightsWithCompanions | Merge Preparation | Group travel behavior | Sum of NumFlightsWithCompanions | Identifies social travelers |
| TotalPointsAccumulated | Merge Preparation | Points earned | Sum of PointsAccumulated | Measures activity level |
| TotalPointsRedeemed | Merge Preparation | Points used | Sum of PointsRedeemed | Measures reward usage |
| RedeemRatio | Merge Preparation (Derived) | Redemption efficiency | TotalPointsRedeemed / TotalPointsAccumulated (0 if undefined) | Distinguishes redeemers |
| Flights_Period1 | Temporal aggregation | Reduce monthly dimensionality | Sum of flights Jan–Jul (Figure 7 and Figure 8) | Preserves monthly activity |
| Flights_Period2 | Temporal aggregation | Reduce monthly dimensionality | Sum of flights Aug–Sep (Figure 7 and Figure 8) | Preserves monthly activity |
| Flights_Period3 | Temporal aggregation | Reduce monthly dimensionality | Sum of flights Oct–Dec (Figure 7 and Figure 8) | Preserves monthly activity |
| Avg_Monthly_Flights | Derived | Average activity | Total flights / 12 | Stable activity metric |
| Seasonality_Score | Derived | Monthly variability | Std(monthly flights) / Mean(monthly flights) | Identifies seasonal travelers |
| Trend% | Derived | Growth trend | (Flights_2021 − Flights_2019) / Flights_2019 × 100 | Separates growing vs declining |

### 2.6. Scaling

Clustering algorithms such as K-means and Hierarchical Clustering rely on distance-based measures to group observations. Consequently, variables with larger numerical ranges can dominate distance calculations and bias the clustering results. In this dataset, features such as Customer Lifetime Value exhibit substantially larger scales than ratio-based variables (e.g., RedeemRatio). Without scaling, high-magnitude features would disproportionately influence the clustering process, reducing the relevance of other informative variables.

To address this issue, the StandardScaler methodology was applied. This approach standardizes each feature by centering the data around zero and scaling it to unit variance. As a result, all variables are placed on a comparable scale while preserving the relative differences between observations. Unlike range-based scaling methods, StandardScaler maintains the original distribution shape and retains information about variability, making it well suited for datasets containing natural extreme values. This preprocessing step ensures that the clustering algorithms operate on balanced and comparable feature representations.

## 3. METHODOLOGY

### 3.1. Feature Selection and Redundant Features

Feature selection aimed to reduce dimensionality and remove redundant or non-informative variables before clustering. Identifying attributes and low-informative geographic variables were discarded, retaining Province/State due to its lower cardinality. Redundant temporal and highly correlated features were removed, including YearMonthDate and variables with near-perfect correlations (Figure 9 and Figure 10). Absolute date variables were excluded after their information was captured through engineered features. This process resulted in a compact, non-redundant feature set optimized for clustering.

### 3.2. Perspective Definition

Now that the relevant features have been chosen, they were grouped into perspectives based on the type of information they convey and their relevance for clustering. The Behavioral Perspective includes variables that describe how customers interact with the airline and how their engagement evolves over time, capturing travel intensity, temporal dynamics, and loyalty behavior. This perspective comprises Tenure, avg_monthly_flights, seasonality_score, trend%, TotalFlightsWithCompanions, and the aggregated temporal variables flights_period2 and flights_period3, as these features directly reflect usage patterns and are well suited for distance-based clustering.

The Value Perspective focuses on the economic contribution of customers, incorporating Income, Customer Lifetime Value, TotalPointsRedeemed, and RedeemRatio. These variables quantify purchasing power, realized and potential value, and loyalty program engagement, making them essential for distinguishing high and low-value customer segments.

Categorical variables were intentionally excluded from the clustering stage due to their incompatibility with distance-based algorithms. Instead, they were used exclusively for cluster profiling to enhance interpretability. These variables include Gender, Education, Marital Status, Province/State, and

Location Code (demographic perspective), as well as LoyaltyStatus, EnrollmentType, and IsActive (program-related perspective). This approach ensures that cluster formation is driven by quantitative behavioral and value metrics, while categorical features provide descriptive context for the final segments.

## 3.3. Multivariate Outliers

To identify multivariate outliers, DBSCAN was used on the scaled feature space, allowing detection of observations with unusual combinations of features based on local density. DBSCAN was chosen for its ability to label low-density points as noise without assuming any data distribution. Using eps = 2.5 (Figure 11 - elbow method) and min_samples = 22 (rule of thumb -> 2*number of features), the method identified 50 outliers (0.30%), retaining 99.7% of customers. This low rate confirms that the data was already well-behaved after univariate treatment, and the cleaned dataset was used for clustering.

## 3.4. Algorithm Selection, Parameter Tuning, and Evaluation Metrics

The clustering methodology followed a comparative and data-driven process in which multiple algorithms were evaluated before selecting the final solution. The goal was to identify a segmentation that balances cluster quality, stability, and interpretability for both behavioral and value perspectives.

Several clustering approaches were tested, including K-means, hierarchical clustering, density-based methods (DBSCAN, Mean Shift Clustering and Gaussian Mixture Model), and Self-Organizing Maps. Each algorithm was chosen to explore different assumptions about cluster shape and data structure. K-means was implemented using k-means++ initialization with multiple random restarts to improve convergence stability. The number of clusters was evaluated for values between two and eight using the $R^2$ metric (Figure 12), and inertia and silhouette score (Figure 13), which jointly assess variance explained, compactness, and separation.

Hierarchical clustering was evaluated using Ward, Complete, Average, and Single linkage methods (Figure 12). Ward's linkage produced more compact and balanced clusters and showed consistent behavior across both perspectives. $R^2$ trends were used to compare hierarchical results with K-means solutions.

Density-based methods, including DBSCAN and Mean Shift, were tested to assess whether clusters were driven by local density patterns. DBSCAN parameters were selected using a k-distance graph for the neighborhood radius and a rule-of-thumb for the minimum number of points (as explained in 3.3). Mean Shift was evaluated under different bandwidth settings and assessed using the silhouette score and Davies-Bouldin index. While these methods were effective at identifying density structures and outliers, they produced less stable and less interpretable segmentations in this context.

Gaussian Mixture Models were evaluated using different numbers of components and covariance structures, with model selection guided by the AIC and BIC criteria. Although probabilistic clustering captured uncertainty in cluster membership, it did not provide clear advantages over simpler methods.

Finally, a two-stage methodology combining Self-Organizing Maps and K-means was adopted. SOMs were trained with hyperparameters selected through a structured grid search evaluated using quantization error and topographic error. K-means clustering was then applied to the SOM neurons, and the number of clusters was selected based on inertia and silhouette scores (Figure 14), and interpretability.

This implementation approach ensured that the final clustering solution was selected through objective evaluation, careful parameter tuning, and methodological comparison, resulting in a robust and interpretable customer segmentation

# 4. RESULTS & VALIDATION

## 4.1. Metrics

Clustering performance was evaluated using quantitative metrics focused on cluster quality and comparability. The primary metric was $R^2$ (variance explained), used to consistently compare algorithms and numbers of clusters. This analysis was complemented with Inertia and Silhouette Score for partitioning methods, and the Davies-Bouldin Index for density-based approaches, ensuring that solutions with poor separation were not selected.

Model comparison was supported by a custom evaluation function that computed $R^2$ and the number of clusters across all approaches using the same feature space. Based on this evaluation, the SOM followed by K-means with four clusters achieved the highest $R^2$ for the behavioral perspective (Figure 15), indicating superior representation of engagement patterns. For the value perspective, K-means with three clusters provided the best performance (Figure 16). These results confirm that the final models were selected through objective evaluation rather than heuristic choice.

## 4.2. Cluster Profiling and Business Interpretation

In this subtopic, first it will be profiled each perspective, and then these are going to be merged and profile both on a metric and a non-metric approach.

The behavioral clustering (Figure 17) revealed four clearly differentiated customer profiles based on engagement patterns. The largest segment corresponds to an Established Active Core (63,7%), characterized by consistent year-round travel, high flight frequency, frequent companion bookings, and stable engagement. This group represents the operational backbone of the business, with predictable behavior and strong loyalty. A smaller Seasonal segment (7.5%) displays extremely concentrated travel in specific periods, reflecting vacation-driven usage with minimal engagement outside peak seasons. The Emerging High-Growth segment (3.8%), although limited in size, shows rapid increases in activity despite short tenure, indicating strong future potential and successful early engagement. Finally, the At-Risk Low-Engagement segment (25%) exhibits persistently low and declining activity, suggesting failed activation or early churn risk.

From the value perspective (Figure 18), three economically distinct segments emerged. The Low-Value Non-Engagers (50.8%) form the largest group, contributing limited revenue and showing weak interaction with the loyalty program. In contrast, the Premium High-Value (17.3%) segment generates a disproportionate share of total Customer Lifetime Value despite representing a smaller fraction of customers, making it the most critical segment for retention. The Active Loyalty Enthusiasts (31.9%) display high redemption intensity and strong program engagement, even though their overall monetary contribution is moderate, highlighting the loyalty program's effectiveness for this group.

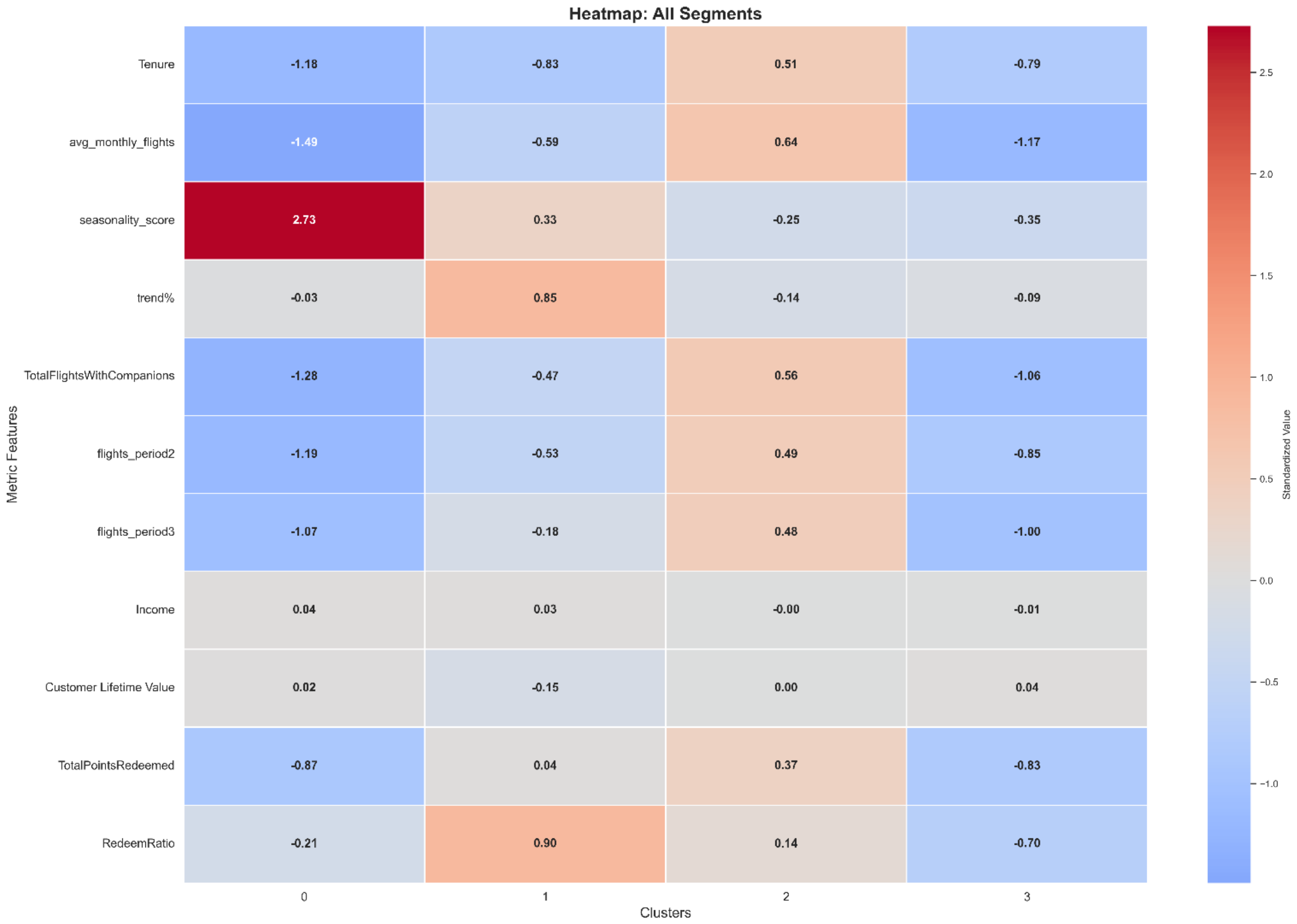


Figure 1  - Final cluster's heatmap

By creating centroids and using hierarchical clustering to merge them (Figure 19), merged behavioral and value segmentation resulted in four final customer profiles (Figure 1) that jointly capture engagement patterns and economic contribution. The Established Active Core (Cluster 2 (63.68%)), the largest segment, combines consistent year-round travel, frequent companion bookings, and active loyalty participation, providing stable and predictable revenue. The Emerging High-Growth segment (Cluster 1 (7.84%)) shows rapidly increasing engagement despite low tenure and moderate current value, indicating strong future potential and high return on targeted growth initiatives. The Seasonal Low-Value segment (Cluster 0 (7.48%)) is characterized by highly concentrated travel during peak periods, low loyalty engagement, and limited economic contribution, suggesting vacation-driven usage rather than long-term loyalty. Finally, the At-Risk Low-Engagement segment (Cluster 3 (20.99%)) exhibits minimal and declining activity with weak program interaction; although some customers retain positive value, this group represents the highest churn risk and requires targeted intervention.

Overall, the integrated segmentation demonstrates that engagement and value do not always align, reinforcing the importance of a combined perspective for actionable customer strategy.

Post-clustering analysis using categorical variables further enriched interpretation. Demographic characteristics (Figure 20) showed minimal variation across clusters, indicating that customer segmentation is driven primarily by how customers behave and the value they generate, rather than by who they are. In contrast, program-related variables (Figure 21) such as enrollment type and active status varied meaningfully across clusters, validating the behavioral profiles and providing insights into acquisition quality, engagement success, and churn risk. For example, high growth and at-risk segments predominantly consist on 2021 promotion enrollment – which can mean that the promotion was either good and precipitated in some cases. Also, 40% of at-risk customers are inactive, which means that it doesn't make sense to try and reactivate them.

## 4.3. Feature Importance

Now that the profiling has been made, a search has been conducted to analyze the feature importance (Figure 22). Analysis shows that behavioral variables are the main drivers of the final segmentation. The most discriminative features are avg_monthly_flights ($R^2$ = 0.74), seasonality_score (0.63), and TotalFlightsWithCompanions (0.58), indicating that flight frequency, temporal consistency, and social travel behavior form the core dimensions separating customers. Tenure and activity in specific periods provide additional differentiation between established and newer customers, while program engagement variables contribute moderately. In contrast, Customer Lifetime Value and Income show negligible explanatory power, confirming that behavior, rather than demographics or value, drives the clustering and supports the separation of behavioral and value perspectives.

## 4.4. Multi-Method Visualization, Cluster Validation, and Marketing Insights

To validate the final merged clusters, three dimensionality reduction techniques were applied: PCA, t-SNE, and UMAP, each providing a complementary view of the same high-dimensional structure and supporting robust interpretation.

PCA was used as a linear baseline with two components selected to balance interpretability and information retention, capturing 53% (Figure 23) of the total variance. Although this choice introduces some overlap between clusters, it provides a stable global view of engagement patterns, with overlap interpreted as a natural consequence of dimensionality compression rather than weak segmentation. Two non-linear methods were then applied to enhance visual separation. t-SNE was tuned through a grid search over the perplexity parameter, selecting the configuration that maximized silhouette score and visual stability. UMAP was tuned using grid searches over the number of neighbors and minimum distance parameters, producing the most compact and clearly separated clusters while preserving both local and global structure.

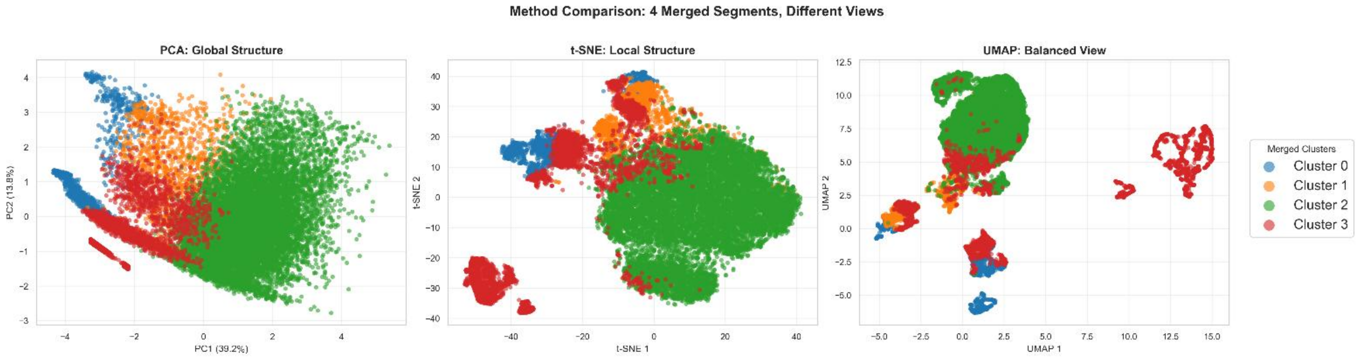


Figure 2 - PCA, t-SNE and UMAP

As can be seen in Figure 2, across all parameter settings and methods, the same four clusters consistently emerged, confirming robustness of segmentation. The Established Active Core (green) remains the largest and most cohesive group, the At-Risk segment (red) shows internal heterogeneity, and the Seasonal (blue) and Emerging High-Growth (orange) segments remain clearly distinguishable. These consistent results confirm that parameter choices support a stable and actionable segmentation.

From a marketing perspective, visual separation supports clear actionability. The Established Active Core should be prioritized for retention and value expansion, the Emerging High-Growth segment for acceleration and early loyalty lock-in, the Seasonal segment for targeted peak-period activation, and the At-Risk segment for churn prevention rather than upselling. Overall, the multi-method validation reinforces both the robustness of the clustering and its relevance for strategic marketing decisions.

### 4.5. Classifying Multivariate Outliers

The multivariate outlier analysis (Figure 24) identified a small group of customers with extreme and contradictory behavioral patterns, particularly an unusually high trend% value far above the mean (14.17 standard deviations above). This signal conflicts with their low flight frequency, short tenure, declining recent activity, and weak engagement in other metrics, indicating statistical anomalies rather than meaningful profiles. Although these customers are not invalid, their highly volatile behavior makes them unsuitable for segmentation-based targeting, and they were therefore excluded to preserve the stability and interpretability of the final clusters.

## 5. STRATEGIC RECOMMENDATIONS

Table 2 - Strategic Recommendations summary

| Tier | Strategic Goal | Target Segments | Share | Primary Focus | Upselling & Cross-Selling Targets | Actions to Avoid |
|---|---|---|---|---|---|---|
| Tier 1 Protect & Grow | Protect high-value customers and accelerate high-potential growth | Established Active Core; Emerging High-Growth | 60% | Retention, loyalty deepening, controlled value expansion | Upselling: premium cabins, flexible fares, priority services<br>Cross-selling: family/group travel, companion add-ons, co-branded cards, partners | Aggressive discounting, mass promotions, generic campaigns |

| | | | | | | |
|---|---|---|---|---|---|---|
| Tier 2 Activate & Convert | Increase engagement and usage frequency | Seasonal Low-Value | 20% | Peak activation, off-season conversion | Upselling: seat upgrades, ancillaries<br>Cross-selling: insurance, lounges, bundled vacations | High-cost loyalty programs, continuous personalization |
| Tier 3 Rescue & Recover | Prevent churn and selectively recover value | At-Risk Low-Engagement | 15% | Churn prevention, segmented reactivation | Selective win-back offers for higher-value customers | Upselling before re-engagement, broad campaigns |
| Maintenance | Maintain brand presence at minimal cost | Residual Low-Engagement | 5% | Low-cost communication | Newsletters, reminders | Active targeting, personalization |

Tier 1 is the highest strategic priority, combining stable revenue with future growth potential. The Established Active Core should be protected through retention and value expansion, while the Emerging High-Growth segment justifies proactive investment due to its strong acceleration and high future ROI. Tier 2 includes Seasonal Solo Travelers with limited current value but activation potential. This segment should be addressed through targeted peak-period and off-season initiatives that increase frequency while maintaining cost efficiency. Tier 3 represents the highest churn risk. At-Risk Low-Engagement customers require selective retention of high-value cases, while low-value dormant customers should be managed with minimal investment.

**Executive Summary:** Customer engagement drives segmentation more than demographics. Investment should prioritize Tier 1, selectively activate Tier 2, and carefully manage Tier 3 to maximize ROI and control costs.

## 6. CONCLUSION

This project developed a robust and actionable customer segmentation for AIAI by combining advanced analytics with clear business objectives. A rigorous data preparation and feature engineering process ensured high data quality, while the separation into behavioral and value perspectives captured both customer engagement patterns and economic contribution, overcoming the limitations of single-perspective segmentation.

Methodologically, multiple clustering algorithms were systematically compared using objective metrics, leading to the selection of a SOM-based approach for behavioral segmentation and K-means for value segmentation. Merging these perspectives produced four statistically robust and business-relevant segments. Validation using PCA, t-SNE, and UMAP confirmed cluster stability, and feature importance analysis showed that behavioral variables are the main drivers of differentiation, directly supporting targeted strategies for retention, growth, activation, and churn prevention.

Despite its strengths, the approach relies on historical behavior, feature selection, and parameter choices, and uses static snapshots that may not capture short-term dynamics. Future work could incorporate temporal models, probabilistic clustering, and additional data sources such as pricing sensitivity or campaign responses. Overall, the framework provides AIAI with a scalable and interpretable foundation for data-driven marketing and loyalty management.

# BIBLIOGRAPHICAL REFERENCES

Bação, F. L. (2025). *Data Mining: Dimensionality Reduction*.

Bação, F. L. (2025). *Data Mining: Partition-based Algorithms (K-means)*.

Bação, F. L. (2025). *Data Mining S12: Association Rules*.

Bação, F. L. (2025). *Data Preparation – Aula S5*.

Bação, F. L. (2025). *Information Visualization – Aula S4*.

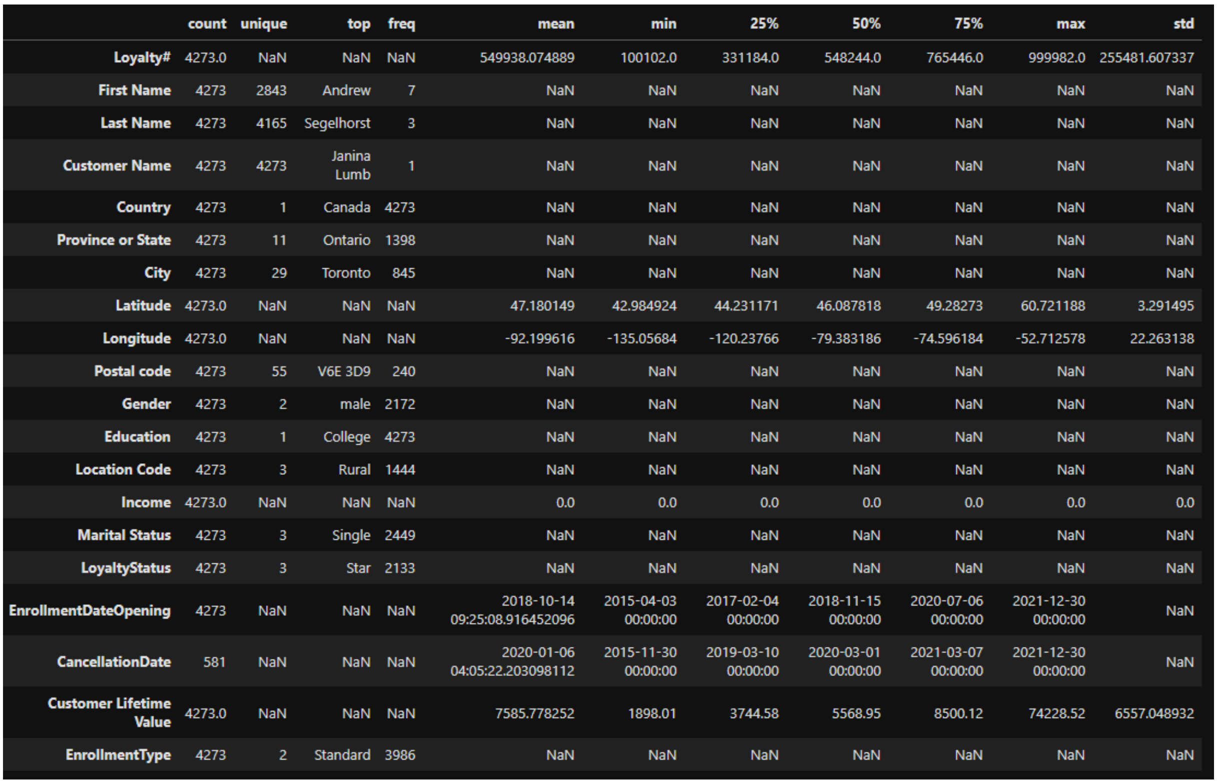
Han, J., Pei, J., & Tong, H. (2023). *Data Mining: Concepts and Techniques*.

Matplotlib Development Team. (2025). *Matplotlib Tutorials*. https://matplotlib.org/stable/tutorials/index.html

scikit-fuzzy Developers. (2025). *scikit-fuzzy: Fuzzy Logic Toolbox for Python*. https://pypi.org/project/scikit-fuzzy/

Seaborn Project. (2025). *Seaborn Tutorial*. https://seaborn.pydata.org/tutorial.html

# APPENDIX



| | count | unique | top | freq | mean | min | 25% | 50% | 75% | max | std |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Loyalty#** | 4273.0 | NaN | NaN | NaN | 549938.074889 | 100102.0 | 331184.0 | 548244.0 | 765446.0 | 999982.0 | 255481.607337 |
| **First Name** | 4273 | 2843 | Andrew | 7 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Last Name** | 4273 | 4165 | Segelhorst | 3 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Customer Name** | 4273 | 4273 | Janina Lumb | 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Country** | 4273 | 1 | Canada | 4273 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Province or State** | 4273 | 11 | Ontario | 1398 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **City** | 4273 | 29 | Toronto | 845 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Latitude** | 4273.0 | NaN | NaN | NaN | 47.180149 | 42.984924 | 44.231171 | 46.087818 | 49.28273 | 60.721188 | 3.291495 |
| **Longitude** | 4273.0 | NaN | NaN | NaN | -92.199616 | -135.05684 | -120.23766 | -79.383186 | -74.596184 | -52.712578 | 22.263138 |
| **Postal code** | 4273 | 55 | V6E 3D9 | 240 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Gender** | 4273 | 2 | male | 2172 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Education** | 4273 | 1 | College | 4273 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Location Code** | 4273 | 3 | Rural | 1444 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Income** | 4273.0 | NaN | NaN | NaN | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **Marital Status** | 4273 | 3 | Single | 2449 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **LoyaltyStatus** | 4273 | 3 | Star | 2133 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **EnrollmentDateOpening** | 4273 | NaN | NaN | NaN | 2018-10-14 09:25:08.916452096 | 2015-04-03 00:00:00 | 2017-02-04 00:00:00 | 2018-11-15 00:00:00 | 2020-07-06 00:00:00 | 2021-12-30 00:00:00 | NaN |
| **CancellationDate** | 581 | NaN | NaN | NaN | 2020-01-06 04:05:22.203098112 | 2015-11-30 00:00:00 | 2019-03-10 00:00:00 | 2020-03-01 00:00:00 | 2021-03-07 00:00:00 | 2021-12-30 00:00:00 | NaN |
| **Customer Lifetime Value** | 4273.0 | NaN | NaN | NaN | 7585.778252 | 1898.01 | 3744.58 | 5568.95 | 8500.12 | 74228.52 | 6557.048932 |
| **EnrollmentType** | 4273 | 2 | Standard | 3986 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

Figure 3 - Descriptive Statistics of customers with "0" Income

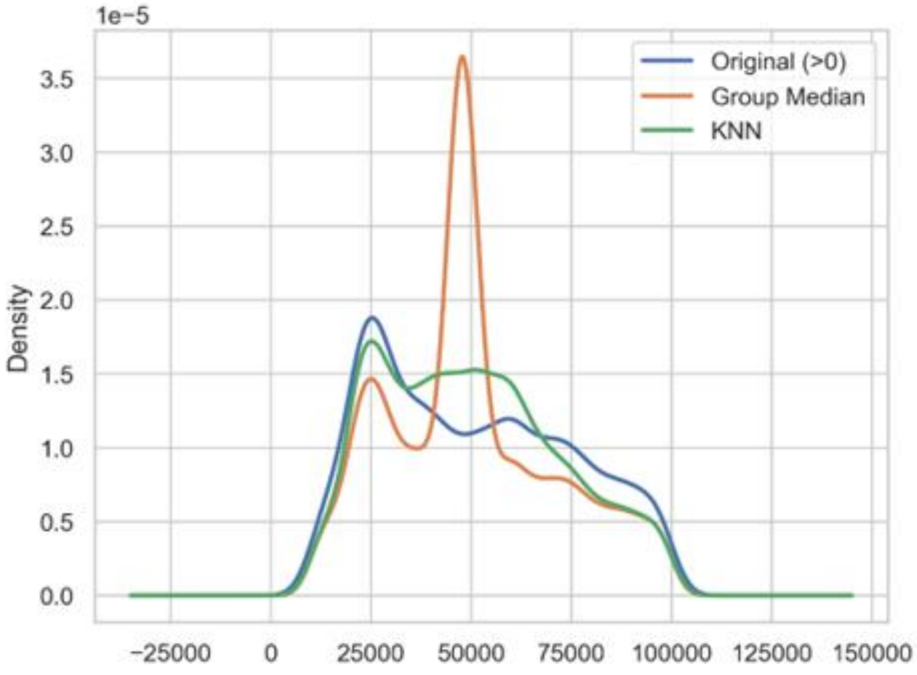


Figure 4 - Income Distributions (Original, Group Median and KNN)

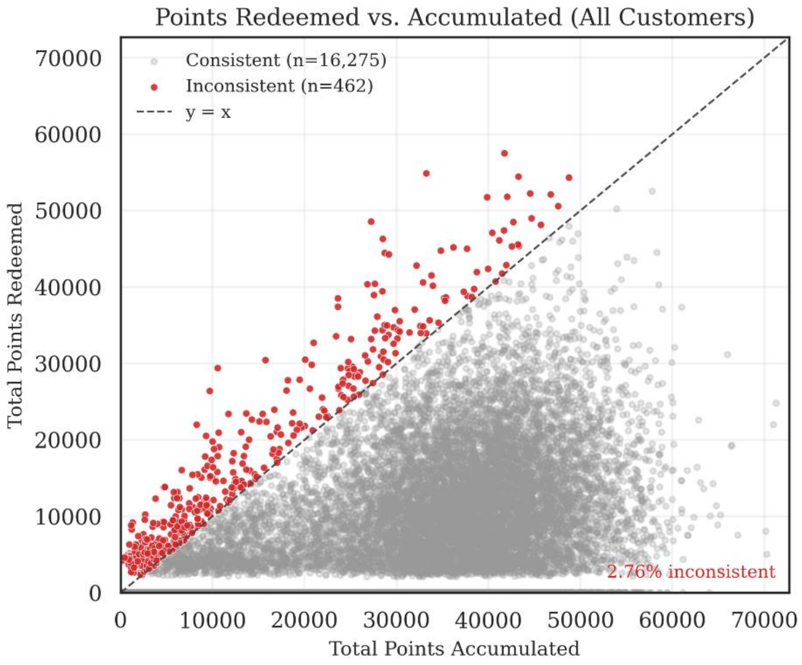


Figure 5 - Points Redeemed higher than Points Accumulated

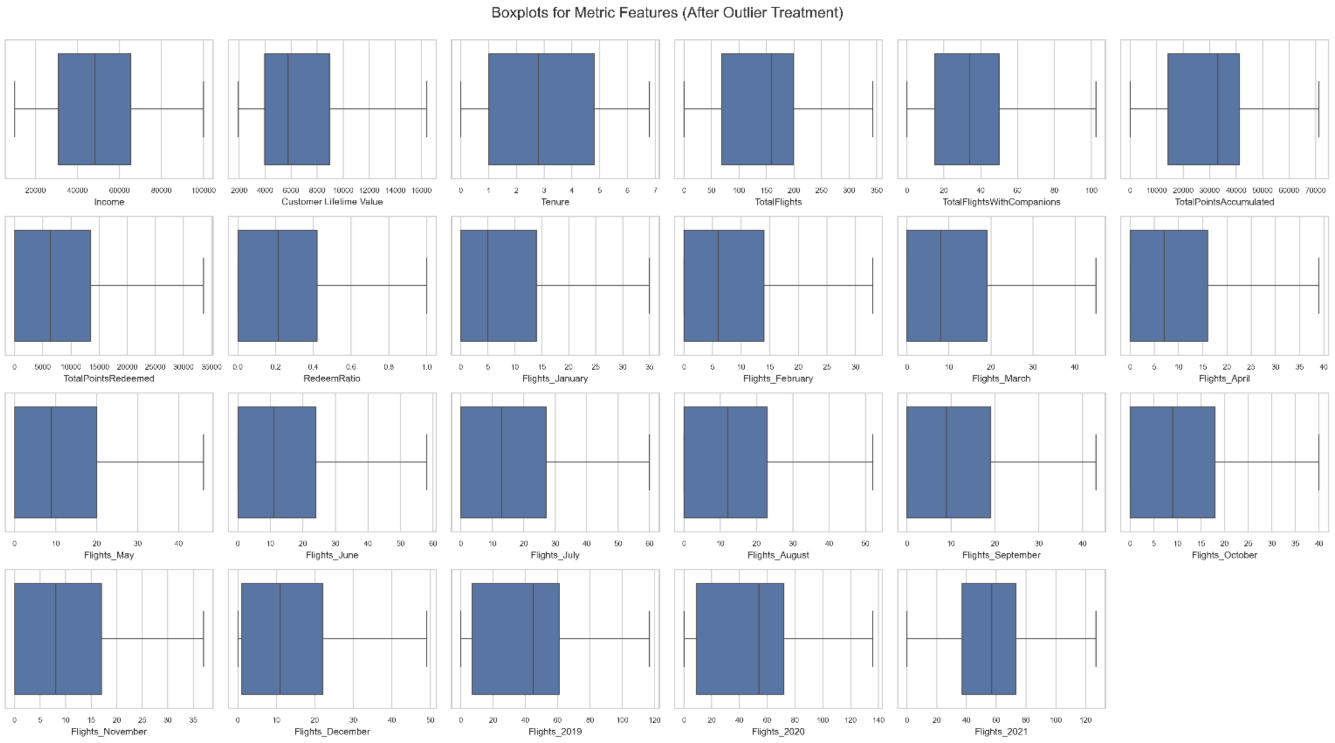


Figure 6 - Boxplots after Outlier Treatment

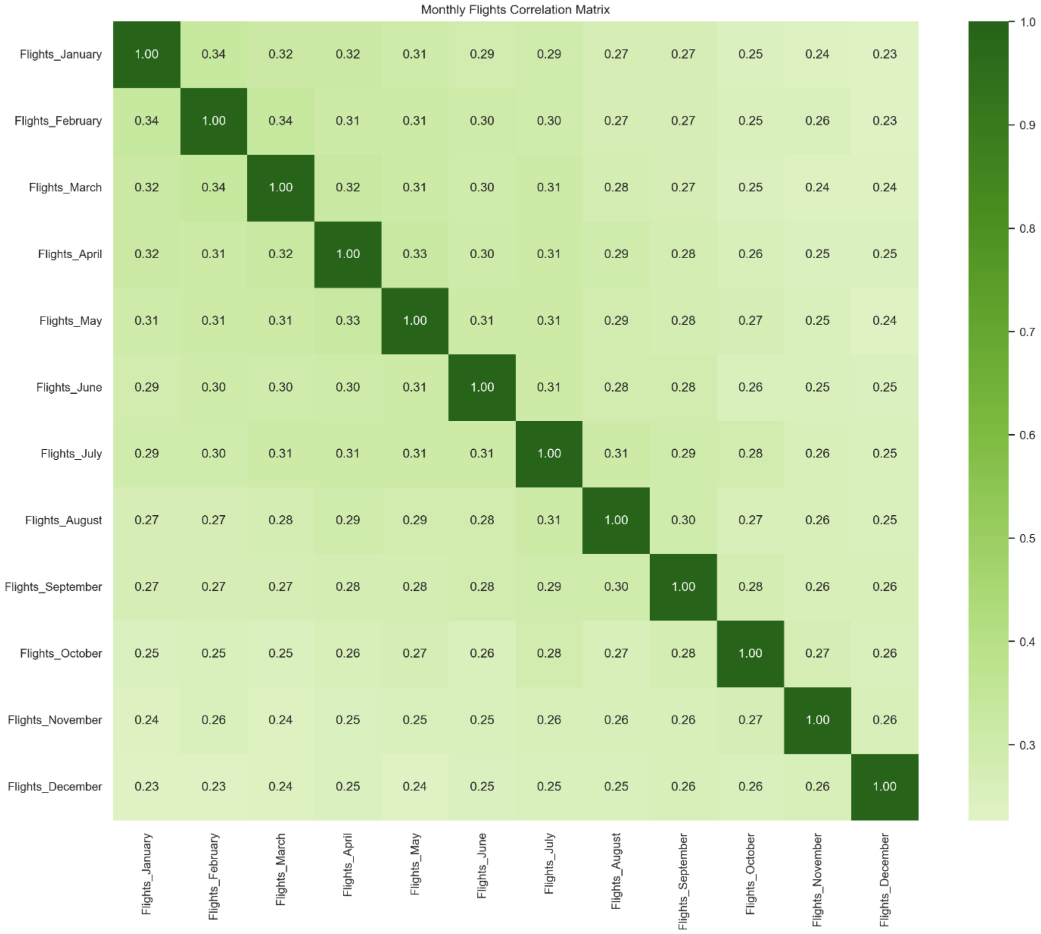


Figure 7 - Monthly Flights Correlation Matrix

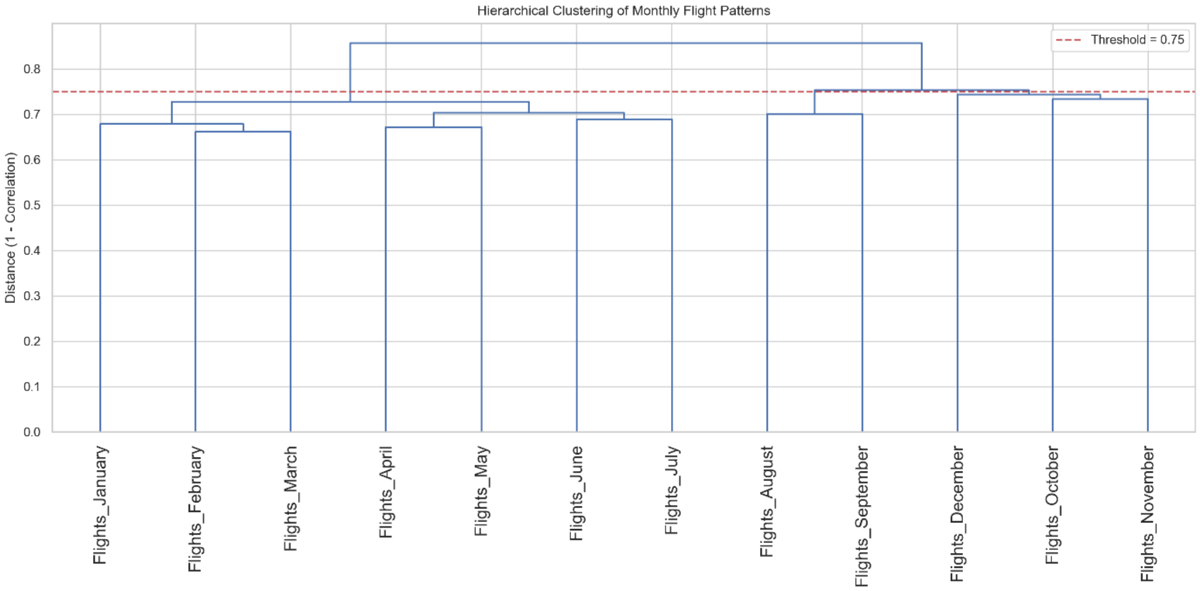


Figure 8 - Dendogram of Monthly Flights Corr Matrix

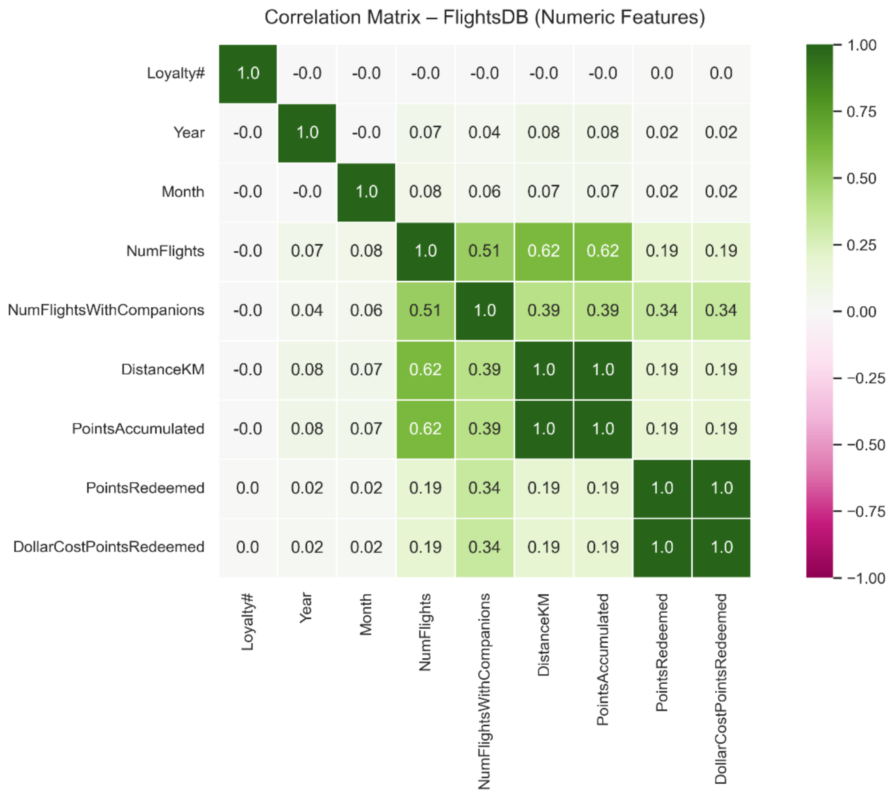


Figure 9 - FlightsDB Correlation Matrix

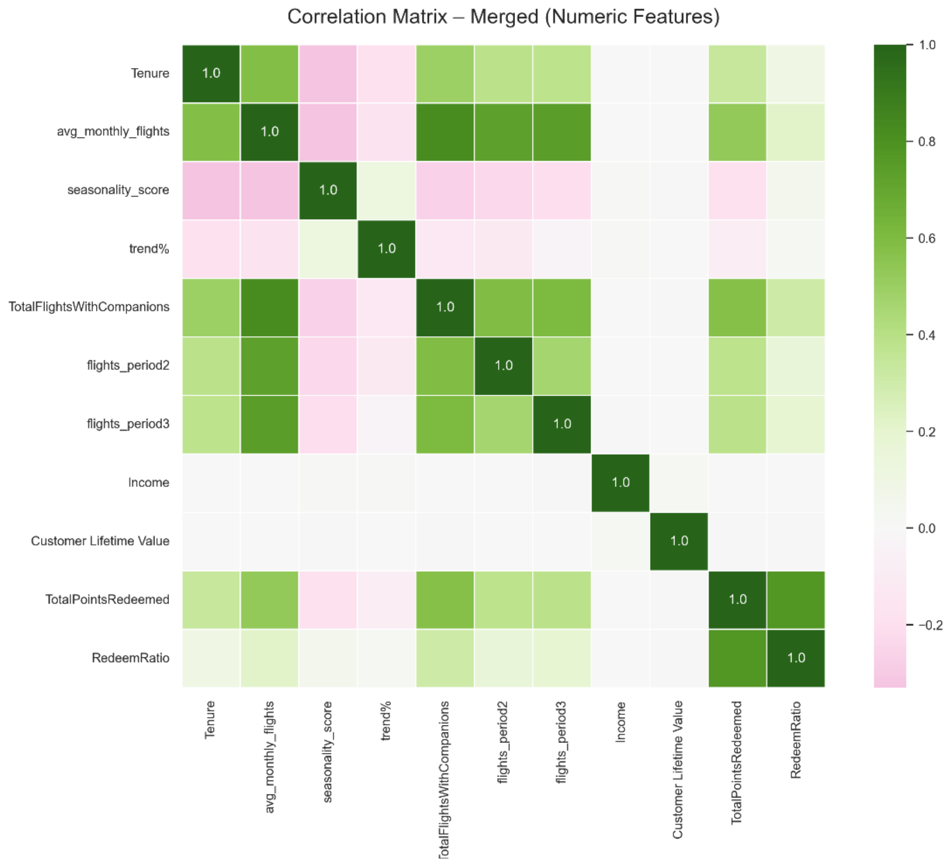


Figure 10 - Merged Correlation Matrix

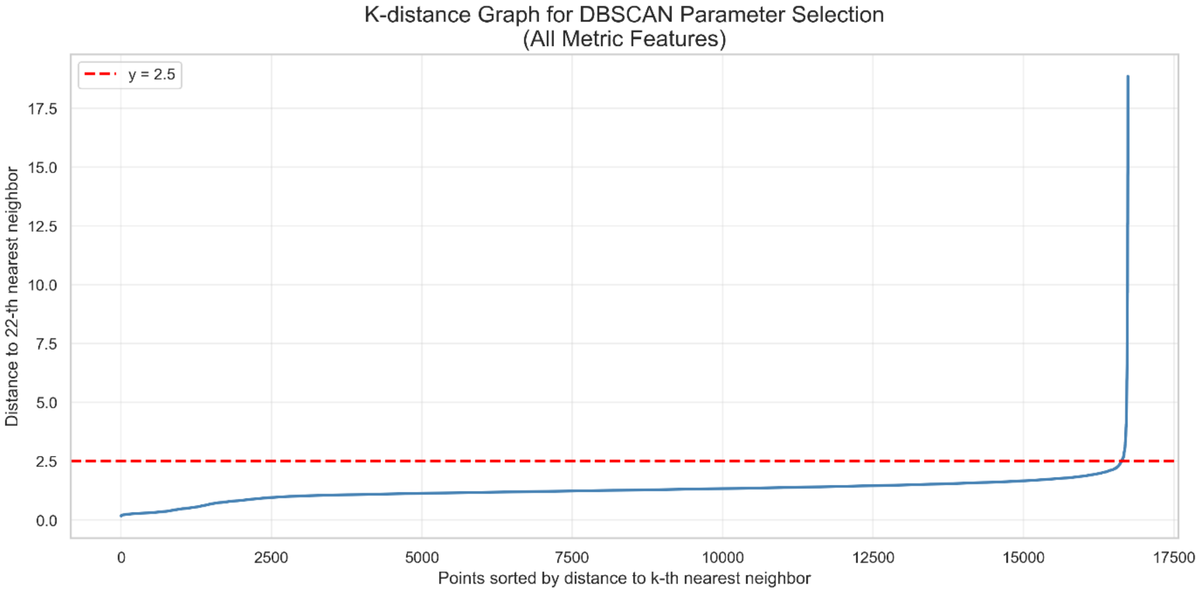


Figure 11 - K-distance Graph for DBSCAN Outlier Detection

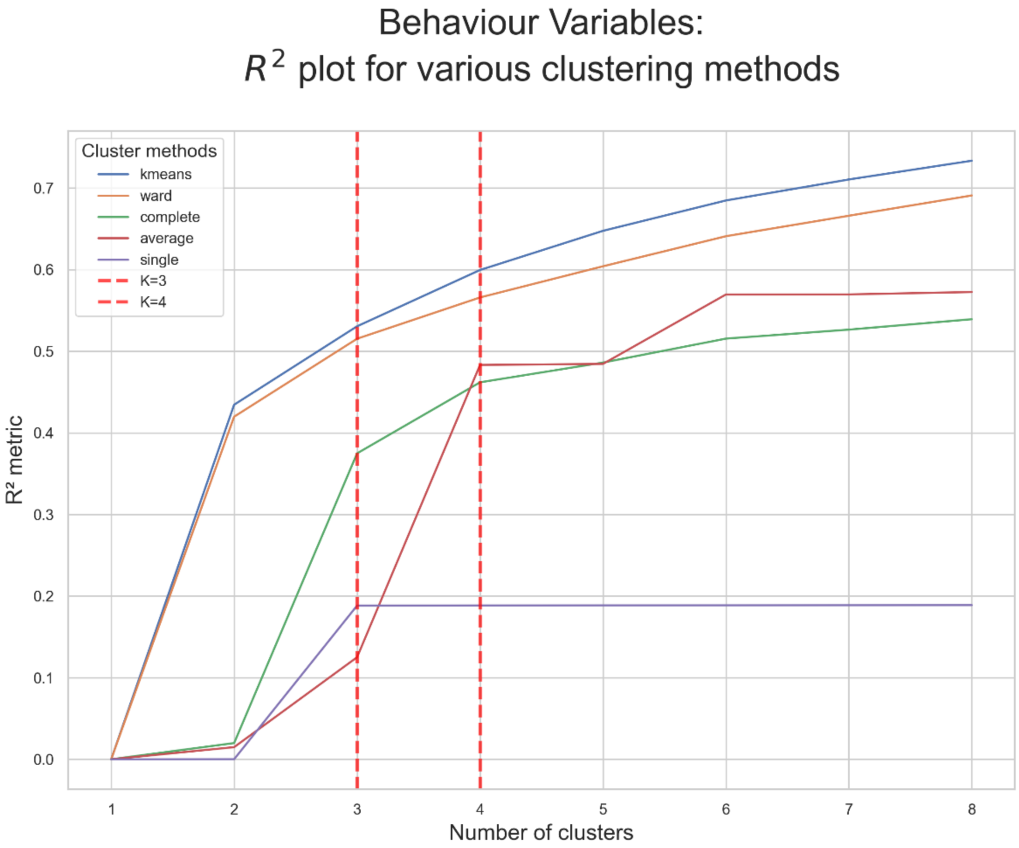


Figure 12 - R2 plot for Kmeans and Hierarchical Clustering (behavioural perspective)

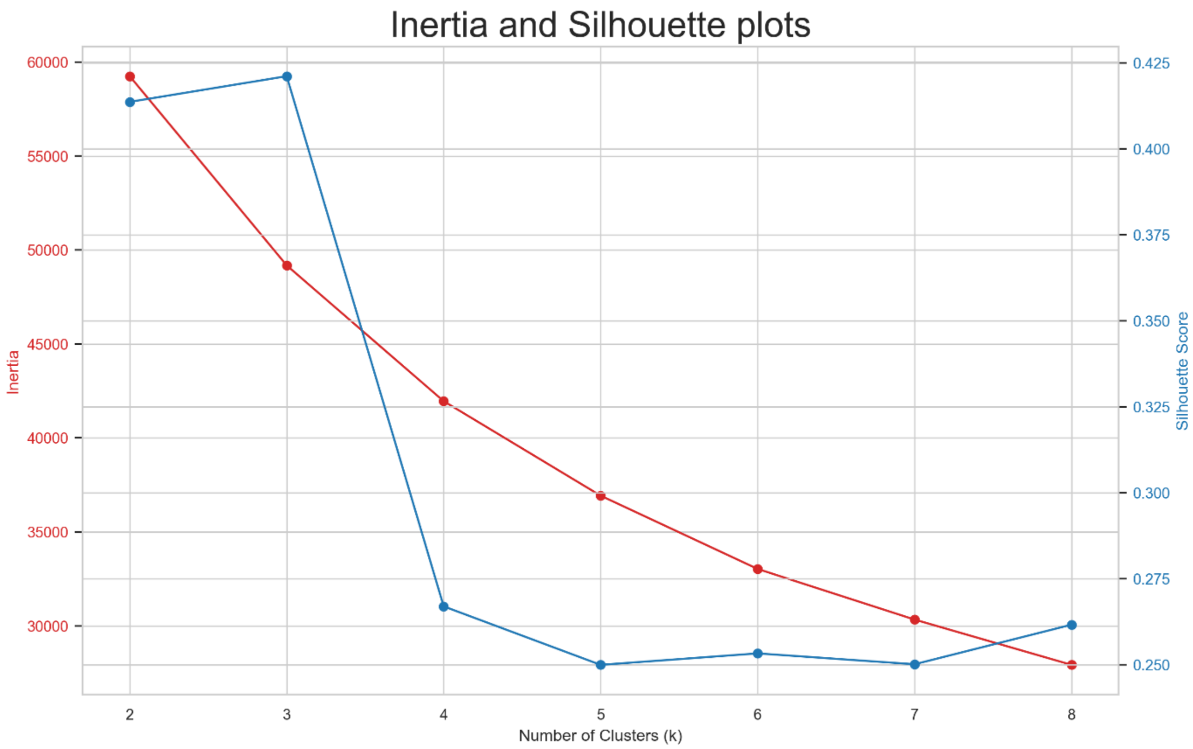


Figure 13 - Inertia and Silhouette Plots for Kmeans (behavioural perspective)

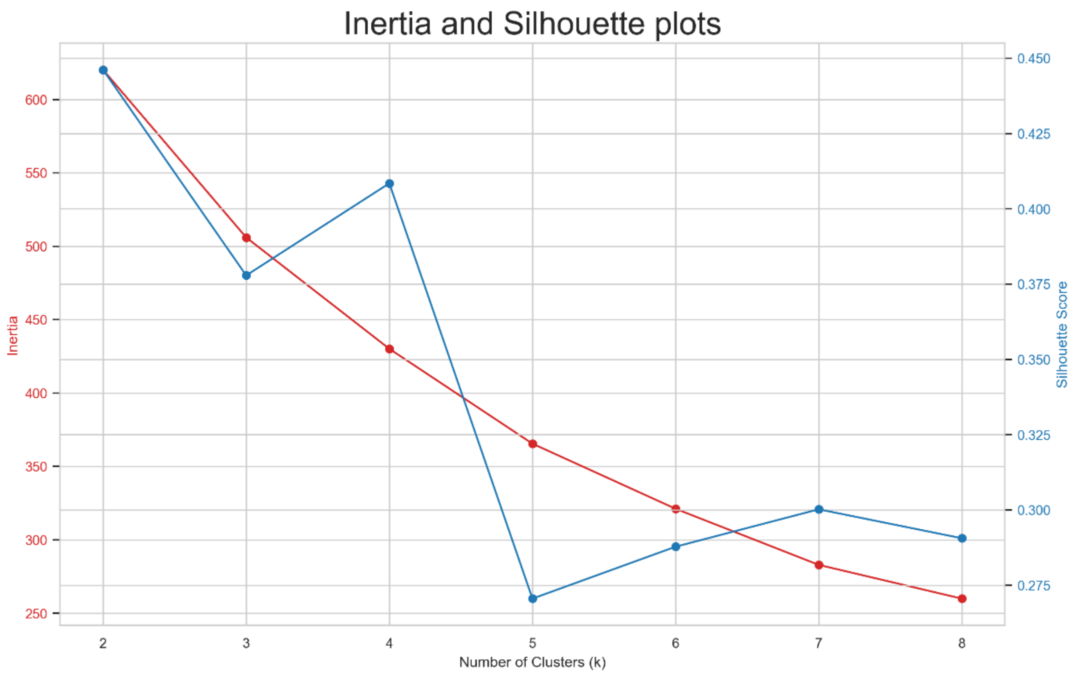


Figure 14 - Inertia and Silhouette Plots for SOM (behavioural perspective)

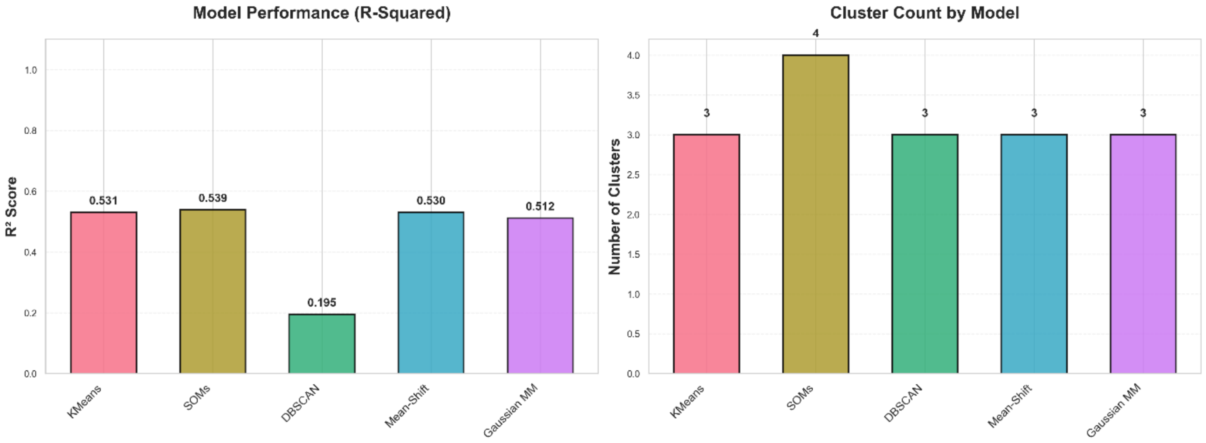


Figure 15 - Model Performance for behavioural perspective

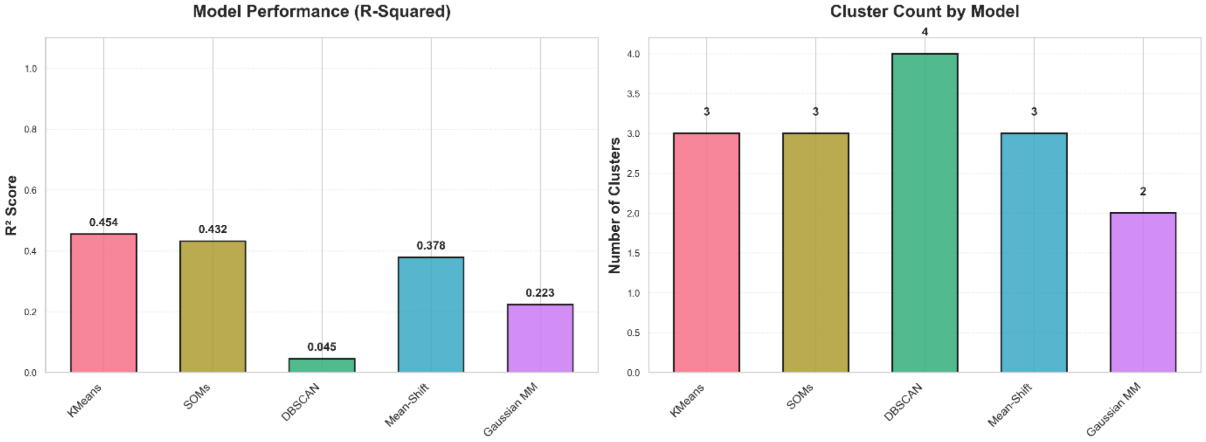
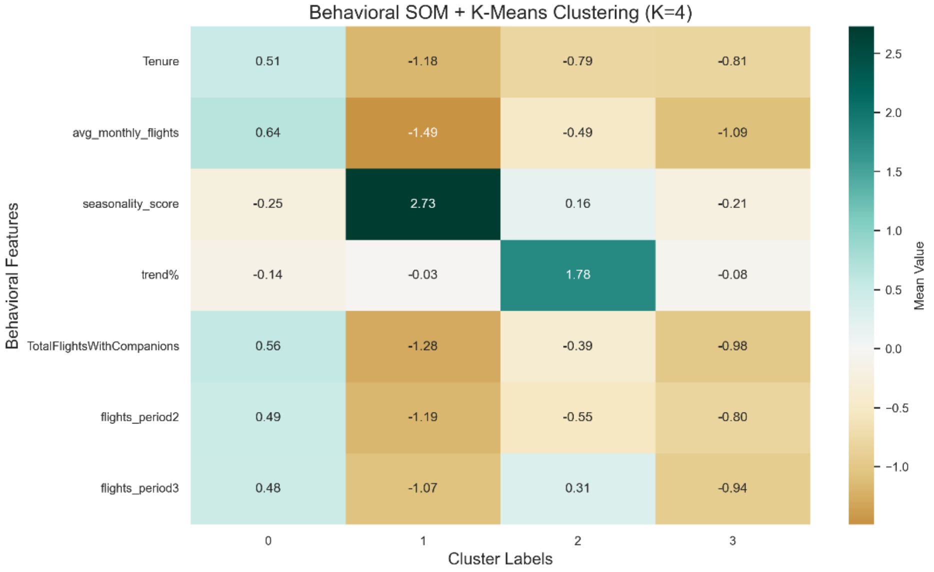


Figure 16 - Model Performance for value perspective



| Behavioral Features | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Tenure | 0.51 | -1.18 | -0.79 | -0.81 |
| avg_monthly_flights | 0.64 | -1.49 | -0.49 | -1.09 |
| seasonality_score | -0.25 | 2.73 | 0.16 | -0.21 |
| trend% | -0.14 | -0.03 | 1.78 | -0.08 |
| TotalFlightsWithCompanions | 0.56 | -1.28 | -0.39 | -0.98 |
| flights_period2 | 0.49 | -1.19 | -0.55 | -0.80 |
| flights_period3 | 0.48 | -1.07 | 0.31 | -0.94 |

Figure 17 - Behavioral Som + Kmeans Clustering

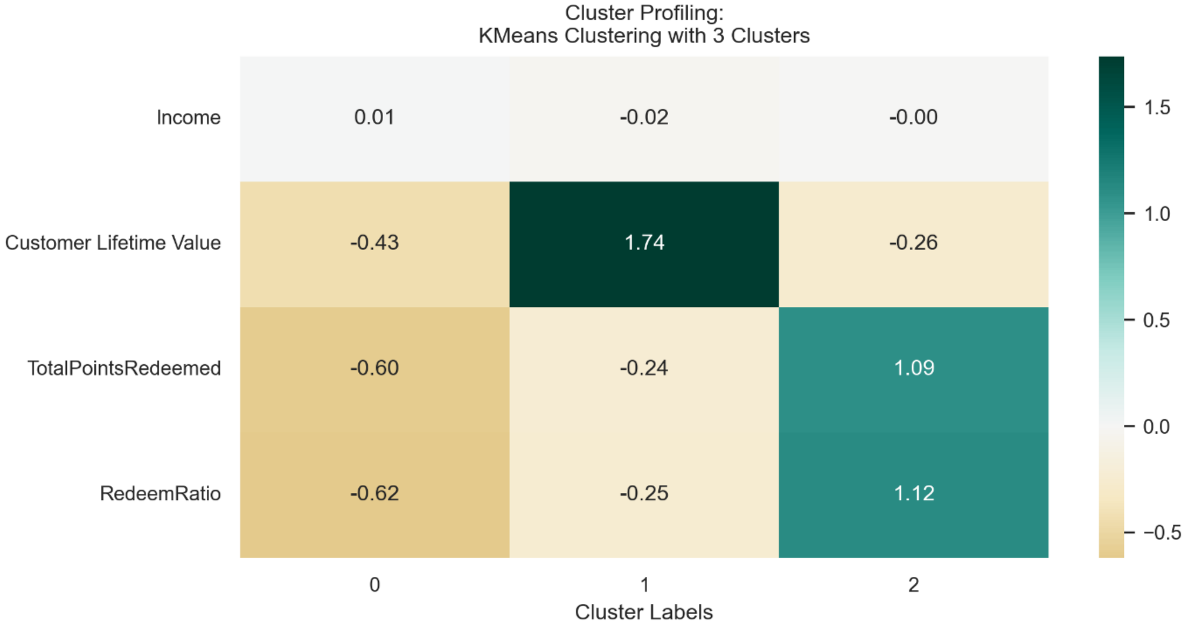


Figure 18 - Kmeans Value

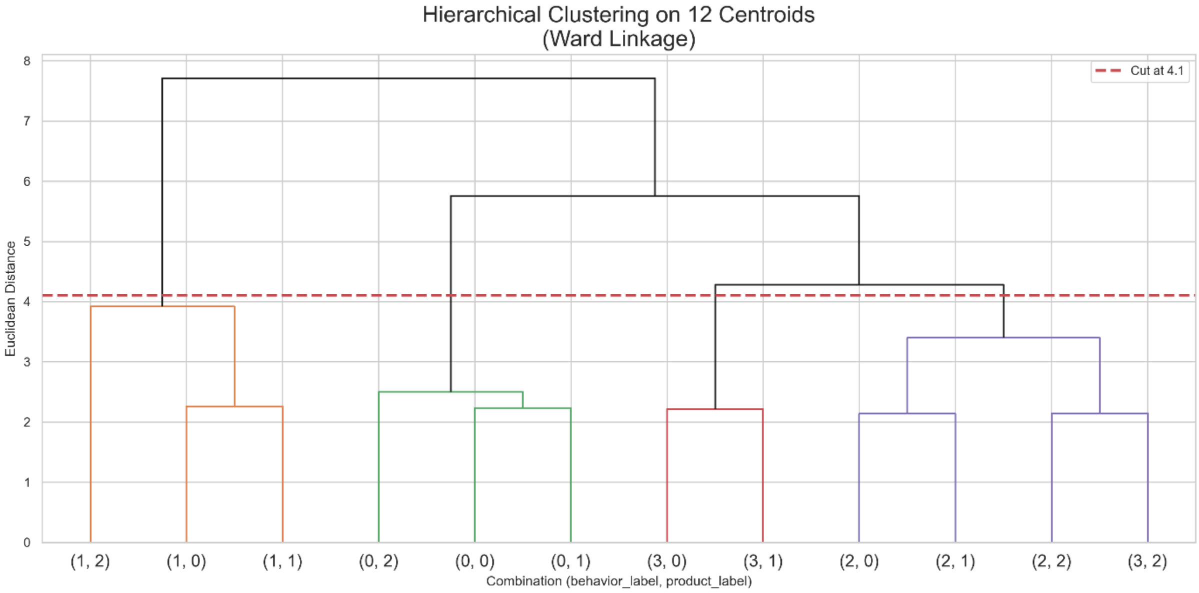


Figure 19 - Hierarchical Clustering Merge

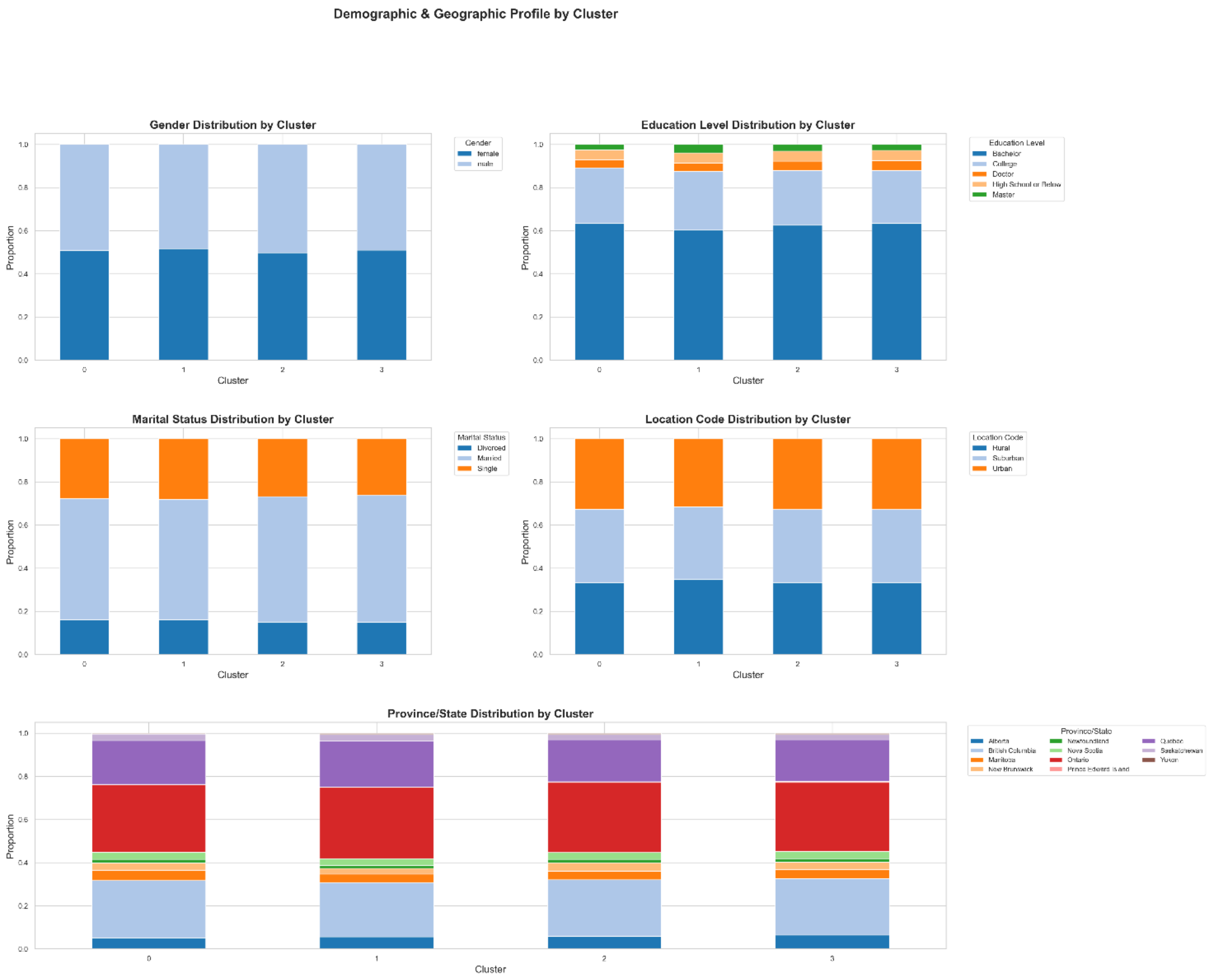


Figure 20 - Demographic Profiling per cluster

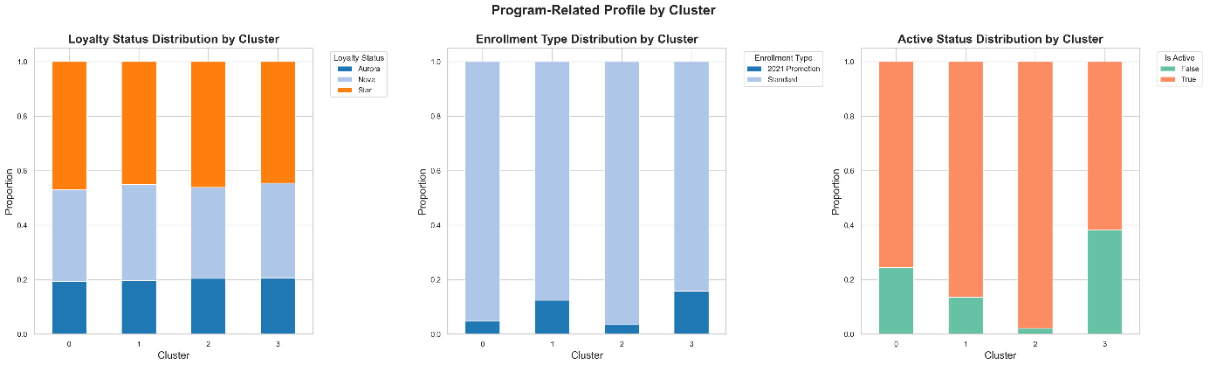


Figure 21 - Program related Profiling per cluster

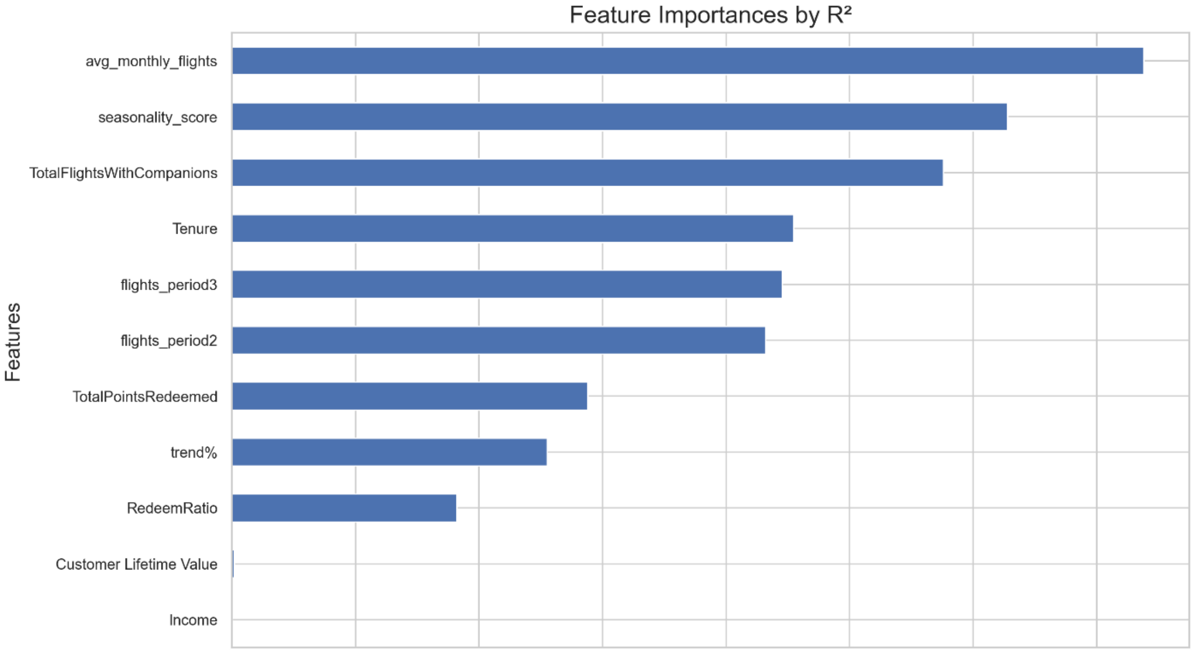


Figure 22 - Feature Importance

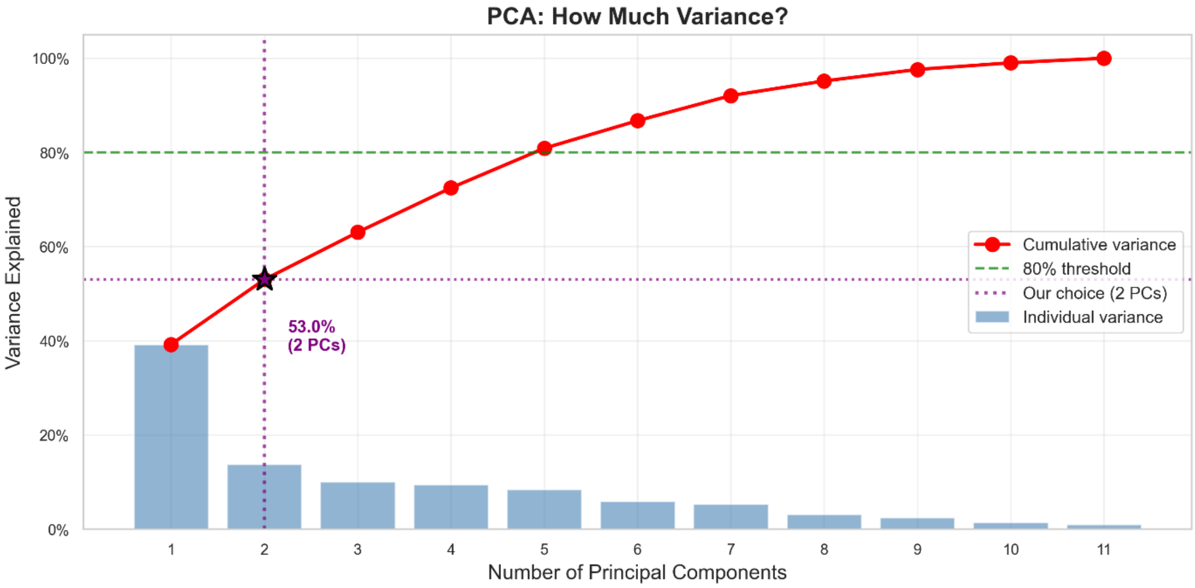


Figure 23 - PCA variance

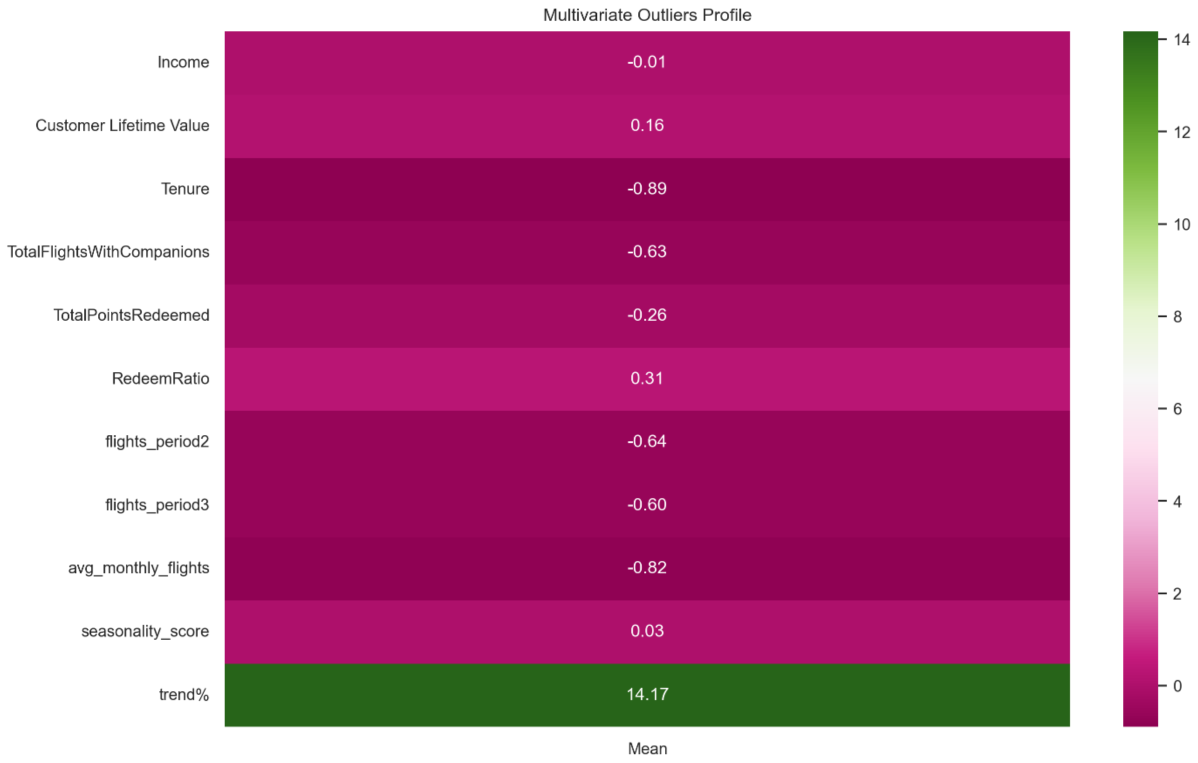


Figure 24 - Multivariate Outliers Profile

# ANNEXES

## ANNEX 1 (AI USAGE STATEMENT)

AI tools were used exclusively through ChatGPT and Claude to assist in technically demanding and conceptually complex components of the Clustering Phase.

Specifically, AI tools contributed to:

- Claude for assisting with the implementation and tuning logic of advanced clustering techniques, such as the Fuzzy C-Means algorithm used in Bonus Option 2, particularly regarding parameter interpretation and methodological consistency;
- Claude for supporting the design and validation of multi-method dimensionality reduction approaches (PCA, t-SNE, and UMAP), in particular the UMAP 3D visualization and interpretation of projection limitations;
- ChatGPT in assisting in structuring feature engineering logic related to temporal aggregation, seasonality, growth trends, and behavioral metrics used in clustering;
- Claude for assisting with the implementation, logic understanding and implementation of Bonus Option 1;
- Claude for assisting with implementing and designing Sankey diagram and bubble chart.

All statistical reasoning, interpretive insights, and business conclusions remain the group's original analytical work.
Any computational assistance received was auxiliary to substantive analytical judgment and is acknowledged herein.

# ANNEX 2 (CONTRIBUTION STATEMENT)

The presented points were discussed by all the members but were mainly done by:

**Pedro Santos (Student ID: 20250399)**

- Preprocessing

- Fuzzy Clustering Implementation (Bonus Option 2)

- Report

- Clustering Implementation and interpretation

- Clusters Visualization

**Miguel Correia (Student ID: 20250381)**

- Preprocessing

- Profiling

**Pedro Fernandes (Student ID: 20250418)**

- Preprocessing

- Fuzzy Clustering Implementation (Bonus Option 2)

- Report

**Tiago Duarte (Student ID: 20250360)**

- Financial Impact Modeling (Bonus Option 1)

- Video Presentation slide support

- Clustering Implementation and interpretation

- Combining segments

- Clusters Visualization

All members contributed to collaborative discussions and ideation.

# ANNEX 3 (RESPONSIBILITY STATEMENT)

We, the group members listed above, certify that this report represents our original analytical work and interpretations. While AI tools were used as specified above, all insights, conclusions, and recommendations are the result of our independent analysis and critical thinking. We take full responsibility for the accuracy and quality of this submission.