DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Amazing International Airlines Inc. Clustering Analysis

Group 23

Bruna Sousa, 20250526

Rui Ferreira, 20250473

Alexandre Coelho, 20250475

Fall Semester 2025-2026

TABLE OF CONTENTS

1. Executive Summary

This report presents a data-driven segmentation strategy developed to transition the AIAI's airline from a generic marketing model to a precision engagement framework. After evaluating multiple clustering techniques, K-Means was selected as the most effective method for analyzing the value, behavioral, and seasonality dimensions. This methodology identified five distinct customer profiles, revealing critical opportunities to maximize return on investment through targeted resource allocation rather than mass-market spending.

The analysis shows a clear split between customers who deliver profit and those who drive growth. **Cluster 2 ("core flyers")** was identified as the airline's operational backbone, characterized by high stability and the highest average customer lifetime value; the strategy for this group shifts from acquisition to retention, utilizing "cancel flexibility" and premium service upgrades to maximize revenue capture. **Cluster 1 ("emerging momentum customers")** represents the growth engine. To maintain this group, we recommend reinvesting profits to fund a "silver status trial". Even though this results in a projected ROI of -52% for this specific segment, it makes sure we keep their full spending potential during this critical growth phase.

Still regarding the high-value segments, the analysis showed specific opportunities in niche and volume-driven groups. **Cluster 0 ("once-a-year traditionalists")** presents a clear seasonal pattern, traveling almost exclusively during winter holidays. The strategy maximizes margin by suppressing marketing spend for most of the year and concentrating budget on festive campaigns to achieve a moderate 12% ROI. In contrast, **cluster 4 ("steady loyalists")** offers a good flight frequency but has the lowest income profile among clusters. Recognizing their price sensitivity, the proposed strategy avoids luxury upselling in favor of a budget-savvy explorer content stream and a 5th flight bonus, using points rewards to keep their consistent activity. Finally, the report shows operational efficiency regarding **cluster 3 ("low-engagement drifters")**, cutting all advertising spend to save costs, moving them to a free automated email list.

The financial impact of this segmented portfolio strategy is visible. The simulation projects a **global ROI of 108%**, with the initial marketing investment fully recovered in the **first six months**. This shows that the model pays for itself quickly and generates real value.

2. Introduction

In phase 1, an initial segmentation strategy distinguished between active and inactive customers, with clustering focused solely on active users to identify behavioral profiles. However, in phase 2, this assumption was revised to include the full customer base. Clustering both active and inactive customers jointly allows inactivity to emerge naturally as a behavioral pattern, enabling the identification of churned or at-risk segments and supporting more data-driven insights into customer lifecycle dynamics.

Data preparation and feature engineering were refined accordingly, prioritizing behavioral relevance and interpretability. Core aggregated activity variables, such as total number of flights, total distance flown, and accumulated points, were retained due to their high variability and strong differentiation

between engagement levels (see figures 1–5). In addition, exploratory analysis revealed pronounced temporal variation in travel behavior across months and years (see figure 6), motivating the creation of higher-level indicators capturing recurring and seasonal flight patterns (see figures 7–8). Complementary lifecycle features, including recency and tenure, were incorporated to distinguish recently active customers from those with declining engagement. Furthermore, the original Customer Lifetime Value (CLV) variable was replaced by a revised metric, as the initial version showed limited dispersion (See figure 9). The engineered CLV exhibits greater variability and stronger alignment with observed customer behavior, providing a more meaningful measure of customer value (see figure 10).

Overall, Phase 2 reflects an evidence-driven refinement of the initial analytical strategy, resulting in a richer and more informative behavioral representation that serves as a robust foundation for the clustering methodologies described in the following section.

3. Methodology

3.1. Feature Selection

To enable the exploration of clustering solutions using the newly engineered features, a feature selection process was first conducted. Building on the conclusions of Phase 1, variables that were deemed unsuitable for clustering were removed at an early stage. These included customer identifiers, personal information, geographic attributes, and raw date variables. Such features do not contribute to behavioral similarity and would artificially inflate distances between customers, as these values are naturally unique.

Although categorical variables were encoded using one-hot encoding (see figure 11), this step was performed primarily to support later cluster profiling and visualization. These variables were not retained in the clustering input, as the segmentation strategy focused on numerical features that directly reflect customer behavior and value.

Feature selection was further refined using correlation analysis of flight-related variables (see figure 12). On the matrix, we can see some redundancy features, such as distance with points accumulated, or more flights made, bigger total distance. As discussed previously, several engineered features were found to better represent customer behavior than their original counterparts. Consequently, redundant variables including total distance flown, accumulated and redeemed points, and the absolute number of flights with companions, were removed in favor of ratio-based and normalized indicators. This resulted in a more compact and informative feature set, as confirmed by the updated correlation matrix after feature selection (see figure 13).

3.2. Feature Scaling

Following feature selection, all retained variables were scaled to ensure balanced contribution to distance-based clustering algorithms. A comparative analysis between Min–Max scaling and Standard scaling was performed using income as a reference variable (see figure 14). Standard scaling was

selected, as it better preserved relative distances between observations and resulted in more stable and interpretable clustering behavior.

3.3. Outlier Detection

Given the sensitivity of centroid-based clustering methods to extreme values, an outlier detection step was introduced prior to clustering. The DBSCAN algorithm was applied exclusively as a preprocessing tool to identify atypical customers whose behavior could disproportionately influence cluster centroids.

To determine an appropriate value for the ε (epsilon) parameter, a k-distance plot was analyzed (see figure 15). The point of maximum curvature in the plot suggested an optimal ε value of approximately ***ε = 1.8***, which was subsequently adopted. Using this configuration, DBSCAN identified a small proportion of customers as outliers, corresponding to approximately ***9.4%*** of the dataset, while the remaining ***90.6%*** formed the core customer population.

This step allowed the clustering analysis to focus on the most representative customer behaviors, while preserving outliers for separate inspection. The detected outliers are subsequently analyzed in relation to the final cluster profiles to evaluate their potential alignment with the identified segments.

3.4. Segmentation Methodology Overview

To ensure analytical clarity while respecting the report length constraints, clustering results are presented by **method** rather than by **perspective**. Each clustering technique is evaluated consistently across the three segmentation perspectives, **Value**, **Behavioral**, and **Seasonality**, following a unified decision logic: the method is applied, its performance is assessed using appropriate metrics and visual diagnostics, and a clear decision is taken regarding its suitability. This structure preserves methodological transparency, ensures interpretability, and avoids unnecessary repetition across perspectives. Table 3, 4 and 5 summarize the performance metrics for the clustering techniques applied to each perspective.

3.5. K-Means and Hierarchical Clustering

In the **value-based segmentation**, K-Means systematically achieves the highest explained variance when compared with hierarchical and density-based alternatives (see figure 16), where the R^2 curve remains consistently above competing methods across all tested values of k. Although the Elbow Method indicates an optimal range between k = 2 and k = 4, a k = 5 solution was selected due to its higher explanatory power and superior interpretability. Silhouette analysis confirms acceptable cluster separation (see figure 17), with average silhouette scores remaining positive and stable across the selected k values. The additional cluster enables a meaningful split of new customers into **High-Potential** and **Low-Value** profiles, increasing business relevance without compromising quality.

For **behavioral features**, K-Means again outperforms competing methods, achieving the highest R^2 values across all tested cluster sizes (see figure 26). While Silhouette analysis suggests a mathematical

optimum at k = 2 (see figure 27), this solution was rejected as overly reductive, merely separating high- and low-activity users. A k = 5 solution was therefore retained, as it provides a more balanced and informative segmentation, capturing both dominant behavioral patterns and niche customer profiles.

In the **seasonality** analysis, K-Means effectively captures recurring temporal travel patterns. Although k = 2 and k = 3 show slightly higher Silhouette scores (see figure 37), these solutions oversimplify seasonal behavior. A k = 4 solution was therefore retained, aligning with the four dominant seasonal profiles observed in monthly flight distributions (see figure 36).

3.6. DBSCAN & HDBSCAN

DBSCAN was applied for **value based segmentation**, using $\varepsilon \approx 0.2$, as suggested by the k-distance plot (see figure 18). This configuration results in a highly fragmented solution with 15 clusters and low explained variance ($R^2 = 0.365$), dominated by numerous small clusters and weak separation. HDBSCAN reduces fragmentation but remains dominated by two large clusters with similar explanatory power ($R^2 = 0.362$), failing to provide meaningful value-based segmentation.

For **behavioral features**, DBSCAN with an $\varepsilon = 0.35$, (see figure 28) produces four clusters with a highly uneven distribution, including one dominant cluster containing the majority of observations and several negligible ones, despite a moderate R^2 of 0.4839. HDBSCAN further exacerbates this imbalance, identifying only two clusters with extreme skewness, indicating poor intra-cluster homogeneity.

From a seasonality perspective, the results are consistent with the previous analyses. DBSCAN k-distance plot shows an $\varepsilon = 0.4$, (see figure 38) making DBSCAN generate nine clusters with over 5% of observations classified as noise and a dominant cluster containing most customers. HDBSCAN reduces noise but remains strongly imbalanced and achieves lower explanatory power ($R^2 = 0.4403$).

3.7. Mean-Shift Clustering

Mean-Shift produces for **value perspective** a binary clustering solution, merging heterogeneous profiles, including Mass Market and VIP customers, into a single dominant cluster, while isolating only a small low-income group (see figure 19). This lack of granularity limits its interpretability and strategic usefulness.

In the **behavioral segmentation**, Mean-Shift again yields a two-cluster solution, distinguishing a small low-activity group from a large cluster characterized by near-average behavior across metrics (see figure 29).

For **seasonality features**, Mean-Shift identifies four clusters using a quantile of 0.25, successfully capturing multiple density peaks corresponding to distinct seasonal preferences, such as Winter- and Summer-oriented travelers (see figure 39).

3.8. Gaussian Mixture Models (GMM)

In the **value perspective**, we first performed model selection using AIC and BIC, which favor a Full Covariance structure over the Diagonal alternative (see figures 20 and 21), indicating strong feature correlations. A 4-component Full Covariance GMM was selected, achieving $R^2 = 0.5127$ with balanced cluster sizes, enabling a clear differentiation between customer value profiles.

For **behavioral data**, the Full Covariance GMM again outperforms the Diagonal model (see figures 30 and 31). A 4-component solution provides a good balance between statistical fit and interpretability, achieving $R^2 = 0.5427$ and capturing meaningful behavioral heterogeneity.

In the **seasonality analysis**, the Full Covariance GMM achieves exceptionally low AIC/BIC values, with a clear elbow at k = 4 (see figures 40 and 41). The resulting model captures the dominant seasonal structures with an R^2 of 0.4617.

3.9. Self-Organizing Maps (SOM)

In the **value perspective**, SOM component planes reveal a distinct diagonal stratification. High Income and CLV values concentrate in the upper-left region, clearly distinguishing premium customers from the mass market, which the hits map confirms is densely clustered in the bottom-left area (see figures 22 and 23). The tenure isolates a distinct strip of new customers along the bottom edge, while points utilization peaks in the upper-right, inverse to income. The U-Matrix highlights a clear distance barrier between these regions, confirming that differences between value segments are abrupt rather than gradual. While K-Means partitions the customer map into distinct regional blocks, Hierarchical clustering shows a boundary zone separating the premium tier from the mass market (see figures 24 and 25).

In the **behavioral** SOM, the top-right corner of the map identifies a structurally isolated inactive segment, characterized by high recency and negative engagement trends (see figures 32 and 33). In contrast, the bottom region highlights customers with positive behavioral trends, corresponding to Growing Loyalists. SOMS with K-Means and Hierarchical clustering validate this separation. K-Means partitions core behavioral blocks, while Hierarchical clustering confirms the topological boundaries that distinctively isolate inactive users from active segments. (see figures 34 and 35).

For **seasonality**, SOMs display clear geometric separation across the map, all season flyers dominate the bottom-left region, characterized by low seasonality scores, while summer travelers are distinctly concentrated in the bottom-right (see figures 42 and 43). SOMS with K-Means reinforces this structure by partitioning the map into distinct seasonal blocks, while Hierarchical clustering validates the topology, effectively isolating the seasonal summer niche while aggregating the off-peak profiles into a broader upper segment (see figures 44 and 45).

4. Results & Validation

4.1. Clustering Performance Across Perspectives

K-means clustering was applied independently to the three analytical perspectives: Value-based, Behavioral, and Seasonality segmentation. Table 2 reports the clustering quality metrics obtained for each perspective.

In the Value perspective, K-means clearly outperforms alternative methods, achieving the highest R^2 score, the strongest Silhouette value among viable solutions, and a high *Calinski–Harabasz* index, indicating well separated and interpretable clusters. Density-based methods, while competitive on isolated metrics, either produced an excessive number of clusters or collapsed the data into too few groups.

In the Behavioral perspective, density-based clustering achieved very strong separation metrics but resulted in only two clusters, effectively collapsing customer behavior into a binary structure. K-means provides a more informative segmentation by distinguishing multiple behavioral patterns rather than collapsing customers into a small number of broad groups, while maintaining competitive clustering quality.

For the Seasonality perspective, density-based methods again achieved higher Silhouette and lower Davies-Bouldin scores. However, these methods proved more sensitive to parameter choices and tended to fragment seasonal patterns into uneven or unstable clusters. K-means offered a more balanced and interpretable representation of seasonal travel behavior with 4 clusters, which was critical for cross-perspective integration.

Overall, K-means consistently delivered a robust trade-off between statistical quality, interpretability, and methodological consistency across perspectives. (See the tables 3,4,5)

4.2. Final Cluster Construction and Size Distribution

While K-means produced stable results across perspectives, the initial clustering was overly detailed for direct interpretation. To address this, hierarchical clustering using *Ward* linkage was applied to the K-means centroids as a post-processing step, consolidating structurally similar clusters without reassigning individual customers (see figure 46). The final solution consists of five clusters, selected to balance statistical structure with managerial interpretability. This consolidation enables coherent customer narratives while preserving the underlying behavioral distinctions identified by K-means.

The resulting cluster size distribution is imbalanced, with one dominant segment and several smaller but clearly differentiated groups. This imbalance reflects the natural heterogeneity of the customer base. No cluster is small or excessively dominant to compromise analytical usefulness (see figure 47).

4.3. Cluster Profiling and Behavioral Interpretation

Cluster profiling was conducted using standardized feature means to highlight relative deviations from the overall customer population. Feature importance was quantified using a supervised model (decision tree), confirming *Income* and *Spring_Ratio* as primary drivers of cluster differentiation (see figure 48). This hierarchy confirms that the obtained segmentation is primarily driven by financial capacity and secondarily by seasonal travel preferences. Parallel coordinate plots and heatmaps provide complementary representations of these patterns (see the figures 49 and 50).

4.3.1. Cluster 0, the “once-a-year traditionalists”

This small segment is defined by extreme seasonality, with strong Fall and Winter concentration and a very high Seasonality Index. This temporal clustering suggests a strong correlation with the Christmas holidays and Autumn getaway. Consequently, value-related indicators such as *CLV* and *Tenure* remain well below average, suggesting episodic, event-driven travel rather than sustained engagement.

4.3.2. Cluster 1, the “emerging momentum” customers

Customers in this cluster exhibit a strong positive flight trend despite below-average current value and tenure. This pattern indicates increasing engagement over time, distinguishing this segment from structurally low-value customers. Visual inspection using *PCA* and *UMAP* projections shows that this cluster lies closer to the cluster 2 segment than to low-engagement groups, suggesting a transitional position in the customer lifecycle and positioning it as a high-potential growth segment (see figures 56 and 57).

4.3.3. Cluster 2, the “core flyers”

Representing the largest segment, this group displays near average behavior across most metrics, with slightly positive *Income*, *CLV*, and flight frequency. Low seasonality and stable usage patterns suggest necessity-driven and predictable travel behavior, forming the operational backbone of the airline’s customer base.

4.3.4. Cluster 3, the “low-engagement drifters”

This cluster exhibits consistently low engagement across almost all dimensions, including flight frequency, *CLV*, and travel distance. The absence of strong seasonality or growth signals suggests limited responsiveness to traditional loyalty mechanisms.

4.3.5. Cluster 4, “steady loyalists”

Customers in this segment demonstrate stable, long-term behaviour, with slightly above-average tenure and consistent engagement across features. This group is characterised by the lowest income levels among all clusters, indicating higher price sensitivity despite frequent usage. Compared to Cluster 2, “steady loyalists” show lower behavioural volatility, suggesting a more established and predictable relationship with the airline.

4.4. Demographic segmentation

Demographic analysis reveals distinct profiles that challenge traditional assumptions. Geographically, customers are well-distributed across regions. However, education acts as a sharp differentiator: cluster 4 is composed almost exclusively of customers with “college” education, contrasting with the prevalence of Bachelor’s degrees in the high volume cluster 2 and growth-oriented cluster 1.
Regarding marital status, while the general population is predominantly married, cluster 4 shows a notable overrepresentation of single users. Finally, the analysis of enrollment dates validates the classification of cluster 1 (“emerging momentum”): as expected, the vast majority of customers acquired during the 2021 promotional campaign are concentrated in this segment, explaining their lower tenure but positive growth trajectory. (See figures 51-54)

4.5. Visual and statistical assessment of clusters

Dimensionality reduction techniques were applied to assess whether the identified clusters exhibit coherent structures when projected into lower-dimensional spaces. Principal Component Analysis (PCA) was used as a linear baseline to visualize global variance patterns, revealing partial overlap between segments. This overlap is expected, as PCA preserves directions of maximum variance rather than cluster separation, and the clustering was performed in the full standardized feature space rather than in the reduced dimensions. (See figures 55 and 56)

To complement this linear view, *UMAP* was employed to explore non-linear neighborhood structure. In contrast to *PCA*, the *UMAP* projection reveals more clearly differentiated manifolds for several clusters, particularly those characterized by distinct behavioral and seasonal patterns, while still showing proximity between structurally similar groups. This suggests that the identified segments reflect underlying structure in the data rather than random partitioning (see figure 57). These visual inspections are used as qualitative diagnostics to support the internal clustering metrics and profiling results. To validate model robustness, outliers that were removed previously were projected onto the final clusters, this confirmed that extreme behavioral values aligned mostly with high activity clusters (clusters 1 and 2), supporting the robustness of the segmentation with respect to extreme behavioural values (see table 6).

To assess internal consistency, feature distributions were compared across clusters using non-parametric *Kruskal–Wallis* tests. Statistically significant differences were observed for the main value, behavioral, and seasonality variables ($p < 0.05$). Given the large sample size, statistical significance was interpreted as confirmation of structural coherence rather than practical importance, which was evaluated through effect magnitude and business interpretation.

4.6. RFM consistency check

A closer inspection of dominant RFM (Recency, Frequency, Monetary) profiles within each cluster further clarifies their behavioural meaning. Cluster 0 is primarily dominated by the RFM score 211, representing approximately 45% of the segment. This profile corresponds to customers with low recency and frequency but moderate monetary value, consistent with infrequent, event-driven travel concentrated around specific periods of the year. The top three RFM profiles account for nearly 70% of the cluster, indicating a relatively homogeneous behavioural pattern. Cluster 3 shows the strongest alignment with classical RFM segmentation. Almost 60% of customers fall into the 111 profile, with the top three profiles explaining over 99% of the segment. These scores reflect very low engagement across all RFM dimensions, confirming this cluster as structurally disengaged.

Cluster 1 is mainly associated with intermediate RFM profiles such as 322, 321, and 311, which together account for approximately 49% of the cluster. This concentration supports the interpretation of this segment as transitional, composed of customers increasing engagement but not yet consolidated into high-value behaviour. In contrast, clusters 2 and 4 exhibit a more diffuse distribution of RFM profiles, with the top three profiles explaining only 17% and 22% of customers, respectively. This dispersion indicates that traditional RFM metrics alone are insufficient to capture their structure, which is instead driven by longer-term behavioural stability and tenure (see Tables 7 and 8).

5. Strategic Recommendations

Drawing on the behavioral insights and financial valuation detailed in Appendix C, it is recommended that AIAI shifts from a generic approach to a **precision-based segmentation strategy**. The following table shows specific objectives and tactics for each cluster:

Table 1 - Strategic marketing recommendations for each customer segment

Cluster	Strategic objective and marketing strategy	Service customization	Loyalty program optimization
Cluster 0: the once-a-year traditionalists	Suppress all marketing investment from February to August and concentrate 100% of the budget in a festive calendar campaign (Sep-Dec), targeting high-yield bookings for Christmas and New Year's Eve.	Since they travel in high-risk weather months, sell disruption insurance as an additional add-on. This generates pure margin to improve this cluster's profitability.	Enable temporary "points pooling" during the holidays. This encourages them to book their entire family group with AIAI (increasing volume) rather than splitting bookings across cheaper competitors.
Cluster 1: the emerging momentum	Instead of standard ads, use the budget to fund a subsidized "silver status trial" for 6 months. This creates high switching costs, preventing them from switching to competitors during their growth phase.	Tag these IDs as future high-value customers in the system. Even on economy tickets, offer them priority boarding or fast track security.	Implement a 2x points multiplier for their first 6 months. This increases the liability on the balance sheet (costing part of the budget) but ensures they consolidate all their travel with AIAI.
Cluster 2: the core flyers	Use the budget to lower the price of premium features (such as empty middle seat), encouraging customers to experience and eventually depend on higher-margin services.	Allow members to cancel one non-refundable itinerary per year in exchange for full flight credit, removing the purchase risk for their most expensive trips.	Offer an annual 50% off companion voucher' after reaching a threshold of a certain number of flights.
Cluster 3: the low-engagement drifters	Move communication to a fully automated welcome sequence (email/app push) that educates the user on how to use the service, ensuring they remain engaged without manual intervention or ad spend.	Maintain a standard economy experience to establish a baseline expectation.	Maintain the account for data tracking purposes only, but exclude them from all bonus campaigns, status offers, or promotions.
Cluster 4: the steady loyalists	Since this group is highly price-sensitive but frequent, the strategy moves away from upselling and focuses on locking in their travel habits.	Replace generic luxury content with a 'Budget-Savvy Explorer' guide, offering cost-saving tips and deals to align with their price-sensitive, high-frequency profile.	Introduce a "5th flight bonus", unlocking extra perks after every five flights, keeping them loyal to our airline despite aggressive pricing from competitors.

6. Conclusion

This project successfully evolved from an initial analysis of active users to a comprehensive segmentation of the entire customer base, demonstrating that effective strategy requires looking beyond simple transaction volume. By integrating inactive users and adopting a multi-perspective K-Means methodology across value, behavioral, and seasonality dimensions, the analysis overcame the limitations of other clustering techniques. This technical approach revealed five distinct profiles, ranging from the high-growth "emerging momentum" group driven by the 2021 promotion to the "steady loyalists" frequent but lower income customers. The integration of demographic data was important, as it allowed a transition from generic marketing to precision strategies, such as silver status trials and companion vouchers for families, linking technical insights to revenue protection and growth objectives. To validate the economic viability of these strategies, a financial simulation model was developed. This model projects that these targeted marketing strategies yield a positive return on investment, transforming technical insights into business value.

However, the analysis is subject to certain limitations that must be acknowledged. The ambiguity in the data dictionary regarding educational levels suggests potential inconsistencies in legacy coding. Furthermore, the financial projections rely on standardized business assumptions for ticket prices and margins rather than granular operational cost data. Consequently, while the simulated ROI is promising, future research directions must prioritize the integration of real-time profitability metrics and the development of dynamic migration models to predict segment shifts. Additionally, qualitative research (such as surveys) is recommended to verify the motivations assumed for each cluster.

Finally, the proposed marketing strategy, particularly the status trials, must undergo controlled market validation to confirm these simulated financial returns before a full-scale implementation. A complementary video-based cluster demonstration is presented in Appendix D, providing a visual and practical illustration of the proposed segmentation-driven strategies.

7. ANNEXES

7.1. AI Usage Statement

AI tools were used as support instruments throughout the project, both for language refinement and for learning and idea development. Tools such as ChatGPT and Gemini were used to improve sentence clarity, adjust academic tone, translate or rephrase Portuguese text into English, and to support understanding of theoretical concepts and implementation approaches, particularly in the bonus components.

All analytical reasoning, methodological choices, business interpretations, and final decisions were developed by the group. AI tools were used in an advisory capacity only and did not generate original insights or conclusions.

A significant portion of the code implementation was based on and adapted from notebooks provided during practical classes. These were extended and modified by the group to meet the specific project requirements, with AI tools used occasionally to clarify syntax, explore alternative approaches, and refine existing ideas.

7.2. Contribution Statement

Student Alexandre Coelho (Student ID: 20250475)

- Bonus: financial impact modelling
- Report: results and validation, appendix A, appendix B, appendix C
- Notebook: feature engineering, data clustering perspectives, Gaussian Mixture Model, merge clusters, test significance test

Student Bruna Sousa (Student ID: 20250526)

- Video
- Report: executive summary, strategic recommendations, conclusion, annexes
- Notebook: feature scaling, outliers removal (DBSCAN), RFM analysis, K-means and hierarchical clustering, DBSCAN, HDBSCAN, mean-shift, SOMs

Student Rui Ferreira (Student ID: 20250473)

- Bonus: implementation and demonstration of interactive visualization

- Report: introduction, methodology, appendix A, appendix B, appendix D
- Notebook: feature engineering, feature encoding, feature selection, data clustering perspectives, profiling, clusters visualization

7.3. Responsibility Statement

We certify that this submission represents our original analytical effort. Although we used class notebooks for code development and AI tools as specified, the resulting insights and strategic recommendations are the core of our independent critical thinking. We take full responsibility for the accuracy and ethics of this report.

Bibliographical References

Software and AI Tools

OpenAI. (2025). *ChatGPT* [Large language model]. OpenAI. https://chat.openai.com/

Google DeepMind. (2025). *Gemini* [Large language model]. Google. https://gemini.google.com/

Canva. (2025). *Canva* [Computer software]. https://www.canva.com/

Power BI. (2025). *Power BI* [Data visualization software]. Microsoft. https://powerbi.microsoft.com/

Envato. (2025). *Envato Elements* [Design resources platform]. https://elements.envato.com/

ByteDance. (2025). *CapCut* [Video editing software]. https://www.capcut.com/

Freepik Company, S.L. (2025). *Freepik: Graphic resources for designers* [Online resource]. https://www.freepik.com/

Libraries and Documentation

Python Software Foundation. (2025). *Python: Official documentation* [Computer software]. https://docs.python.org/3/

Plotly Technologies Inc. (2025). *Dash: Web application framework for Python* [Computer software]. https://dash.plotly.com/

Plotly Technologies Inc. (2025). *Plotly Express: Interactive data visualization library for Python* [Computer software]. https://plotly.com/python/

Pandas Development Team. (2025). *pandas: Data analysis and manipulation library* [Computer software]. https://pandas.pydata.org/

NumPy Developers. (2025). *NumPy: Fundamental package for scientific computing in Python* [Computer software]. https://numpy.org/

Matplotlib Developers. (2025). *Matplotlib: Visualization with Python* [Computer software]. https://matplotlib.org/

GeoPandas Developers. (2025). *GeoPandas: Python tools for geographic data* [Computer software]. https://geopandas.org/

Scikit-learn Developers. (2025). scikit-learn: Machine learning in Python [Computer software]. https://scikit-learn.org/

Appendix A. Figures

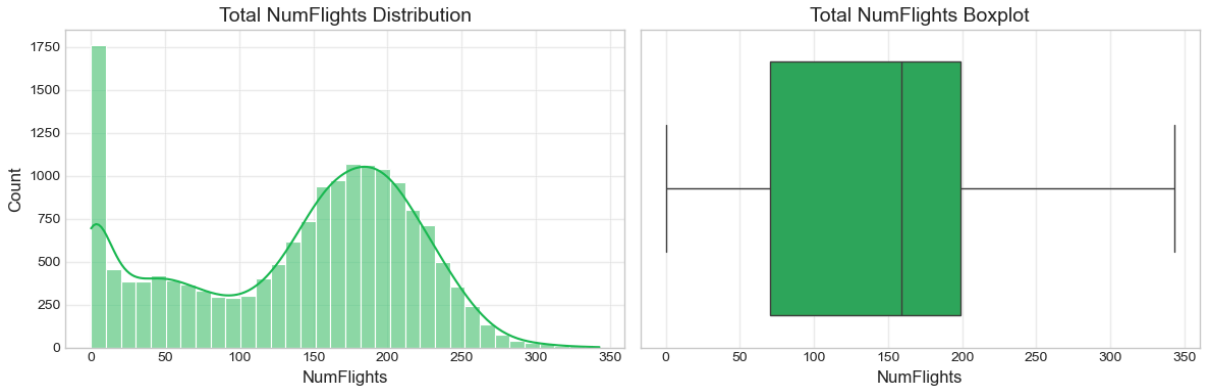

Figure 1 - Feature Engineering: Cumulative Customer Totals (Total_Flights)

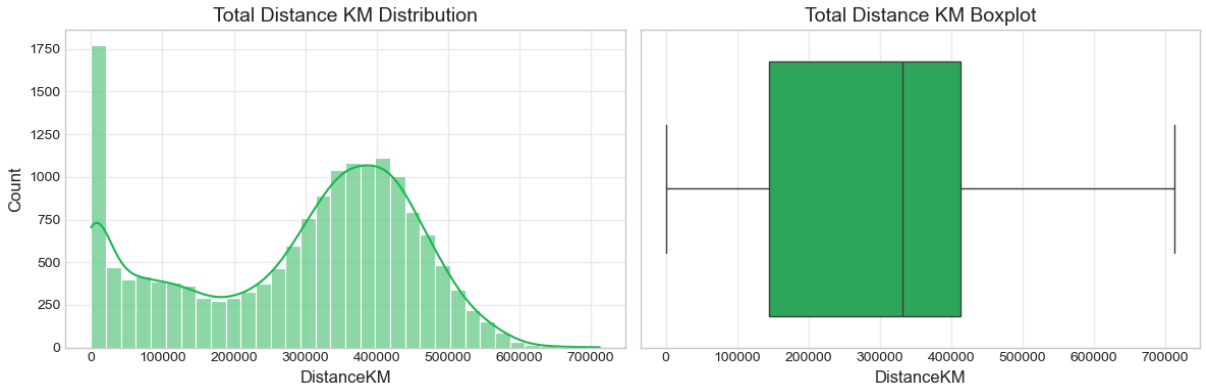

Figure 2 - Feature Engineering: Cumulative Customer Totals (Total_DistanceKM)

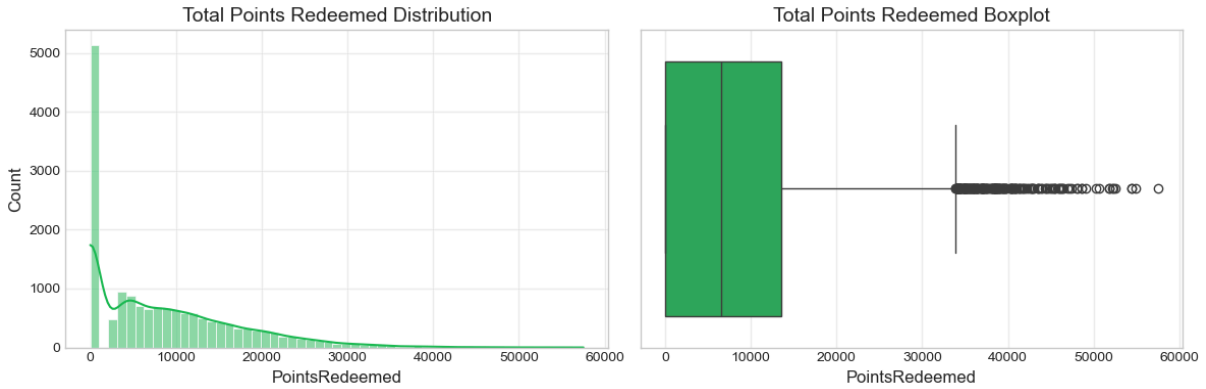

Figure 3 - Feature Engineering: Cumulative Customer Totals (Total_Points_Redeemed)

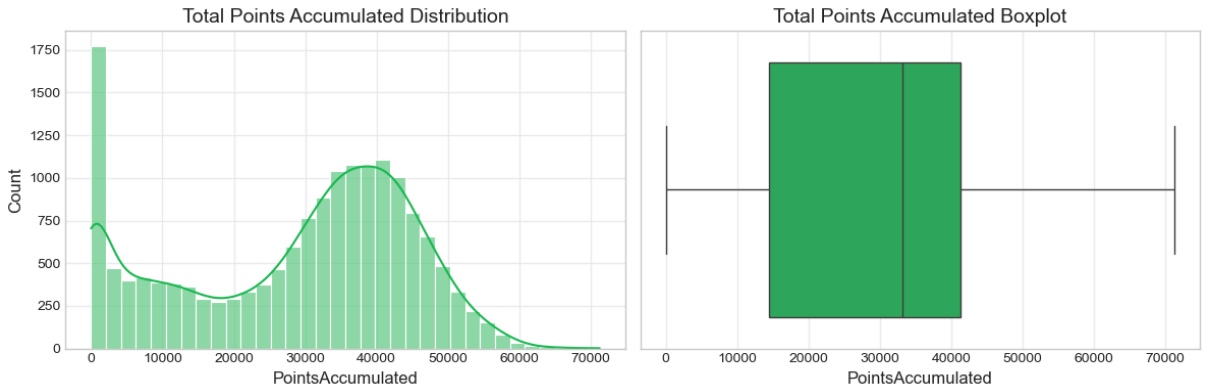

Figure 4 - Feature Engineering: Cumulative Customer Totals (Total_Points_Accumulated)

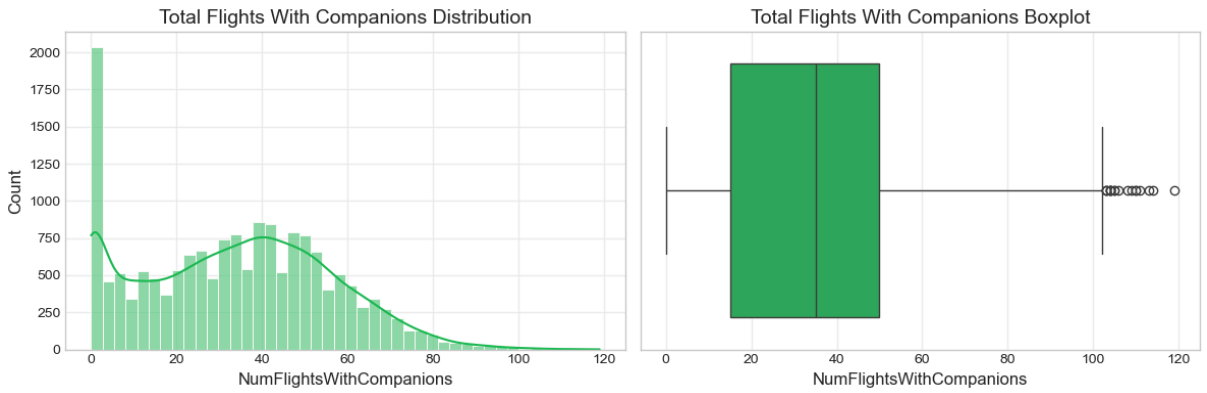

Figure 5 - Feature Engineering: Cumulative Customer Totals (Total_Flights_With_Companions)

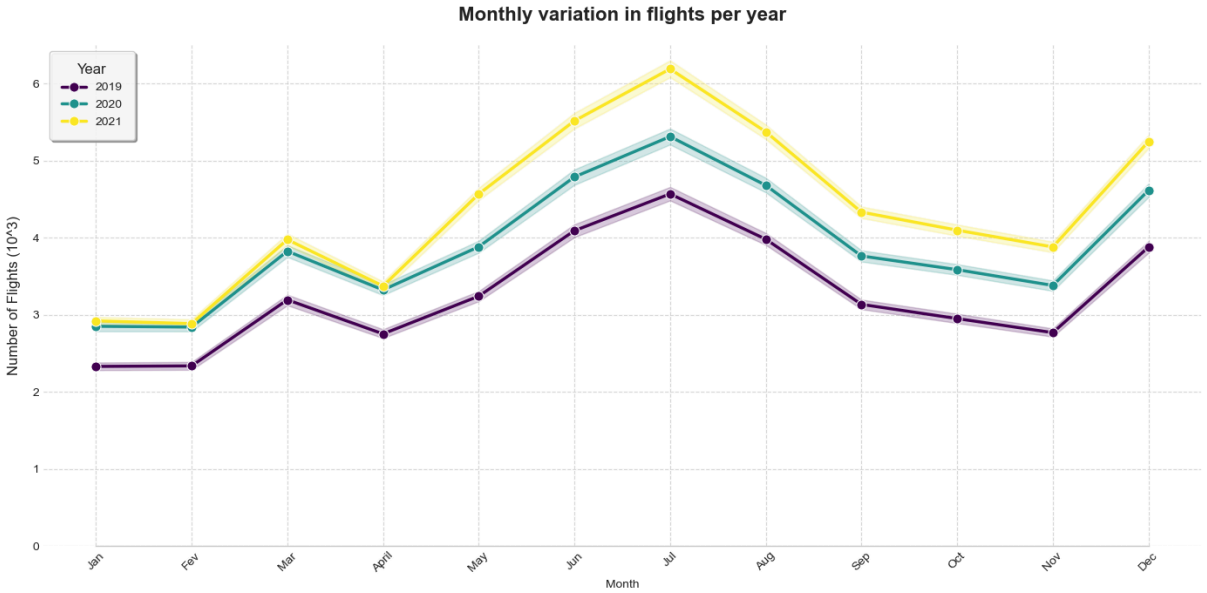

Figure 6 - Monthly variation in flights per year

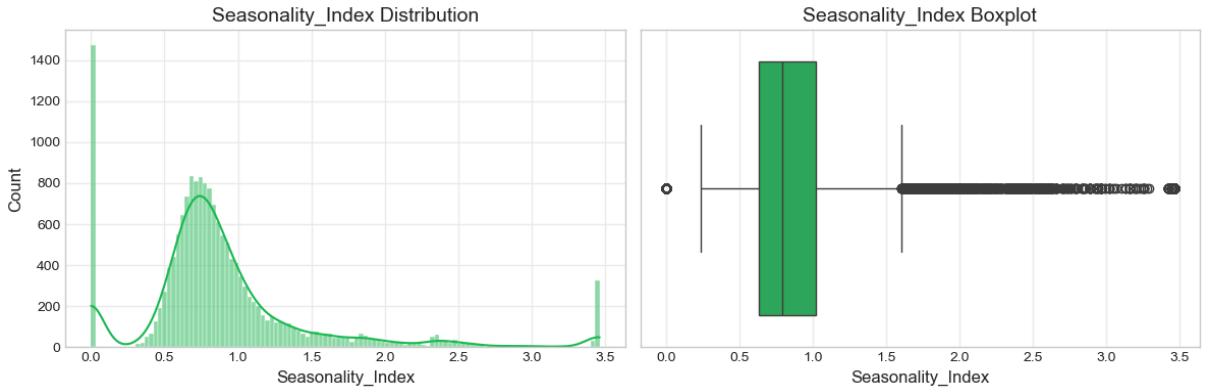

Figure 7 - Feature Engineering: Seasonality Index (Seasonality_Index)

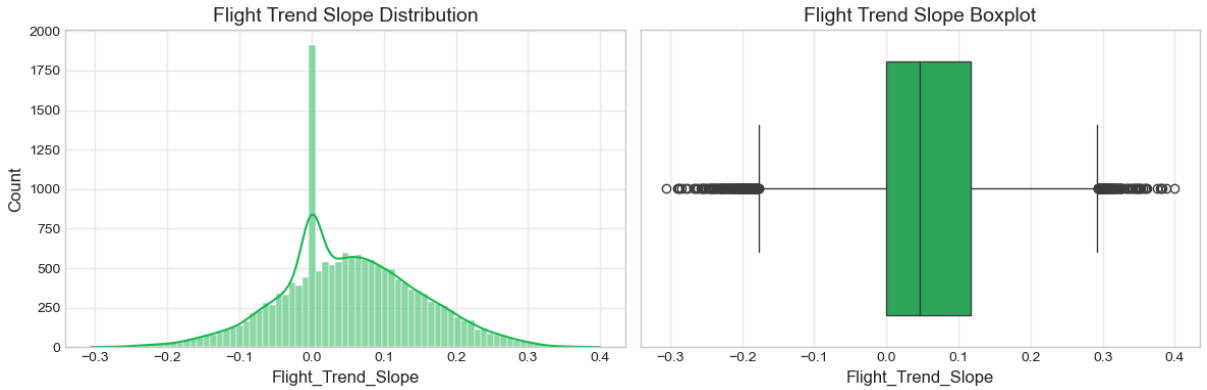

Figure 8 - Feature Engineering: Flight Trend Slope (Flight_Trend_Slope)

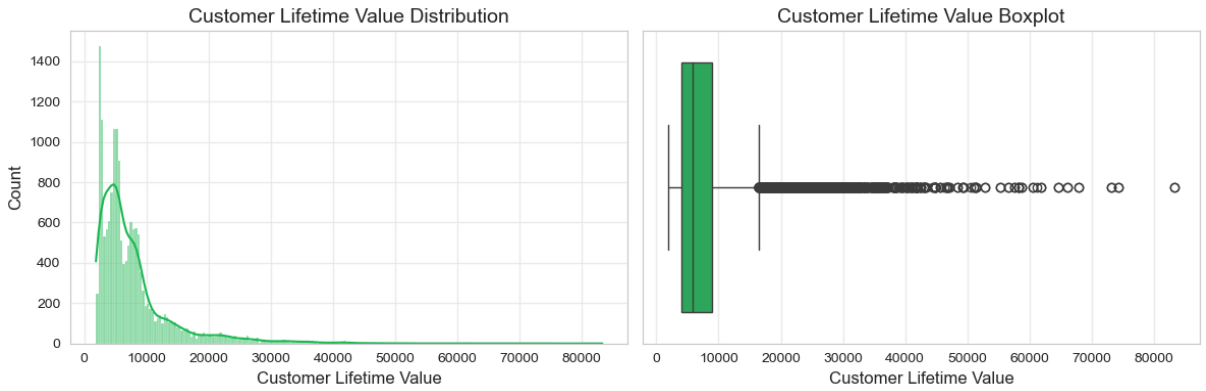

Figure 9 - Original Customer Lifetime Value

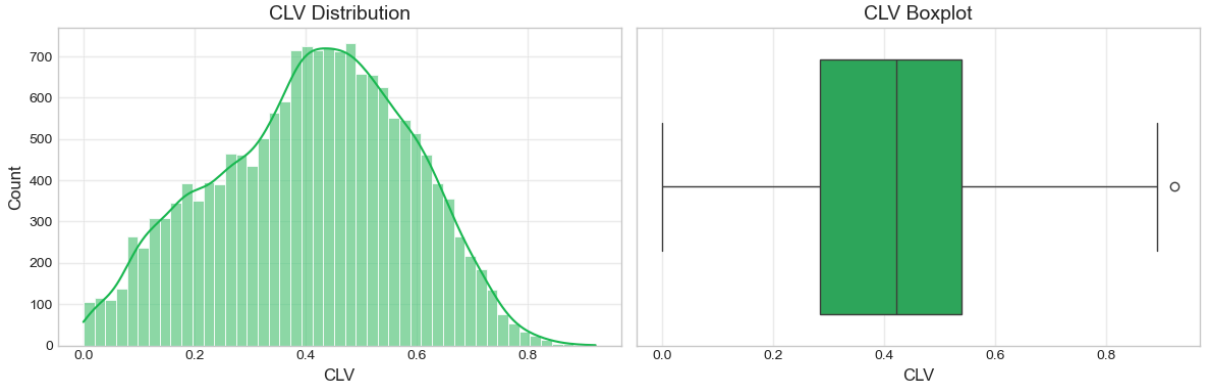

Figure 10 - Feature Engineering: Customer Lifetime Value (CLV)

```
Features Métricas (18): ['Income',
Features One-Hot (52): ['oh_female
Features Unused/Descartadas (13):
```

Figure 11 - Feature Encoding Code Section

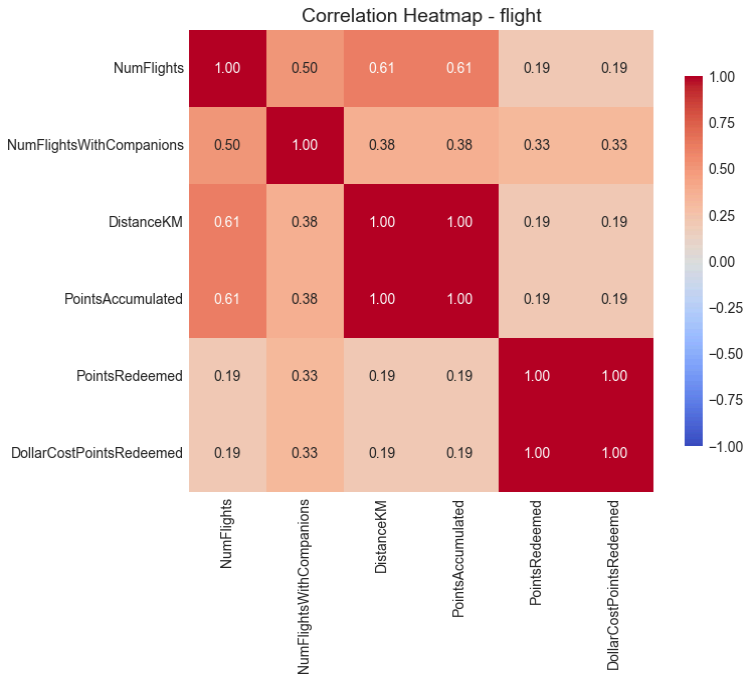

Figure 12 - Correlation Heatmap Numeric Flight Features

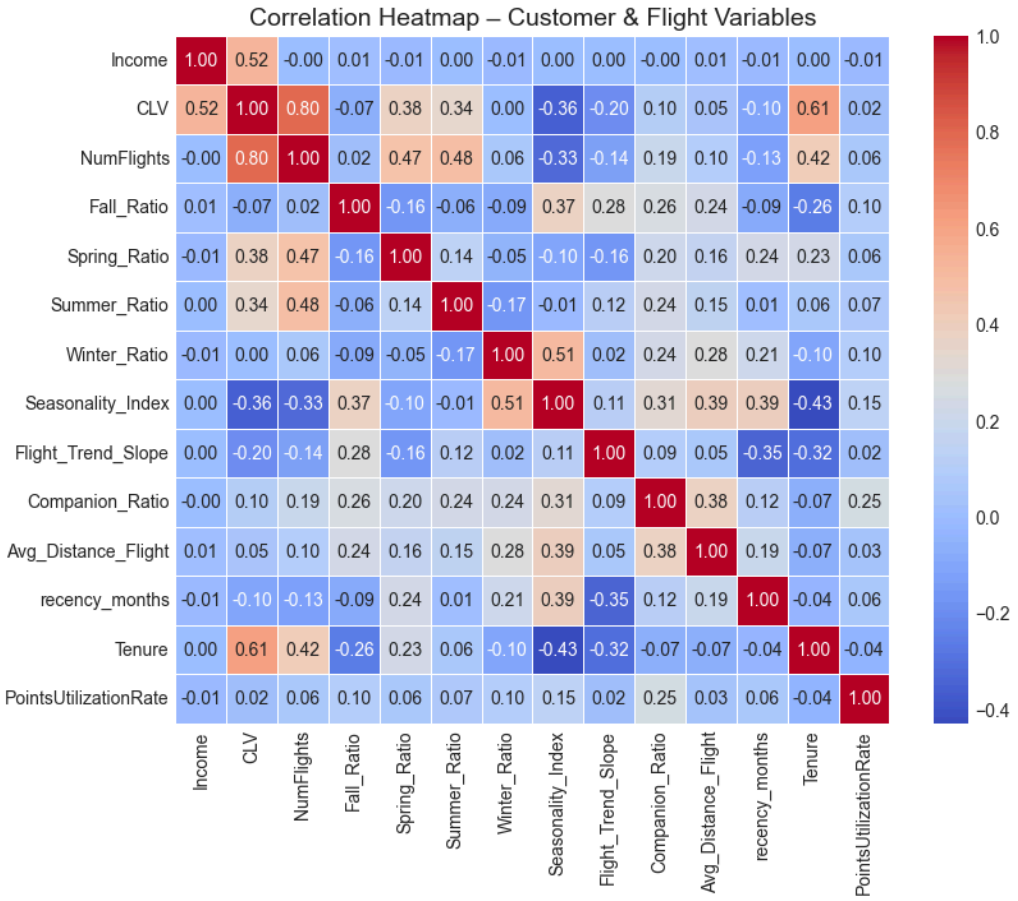

Figure 13 - Correlation Heatmap Feature Selection

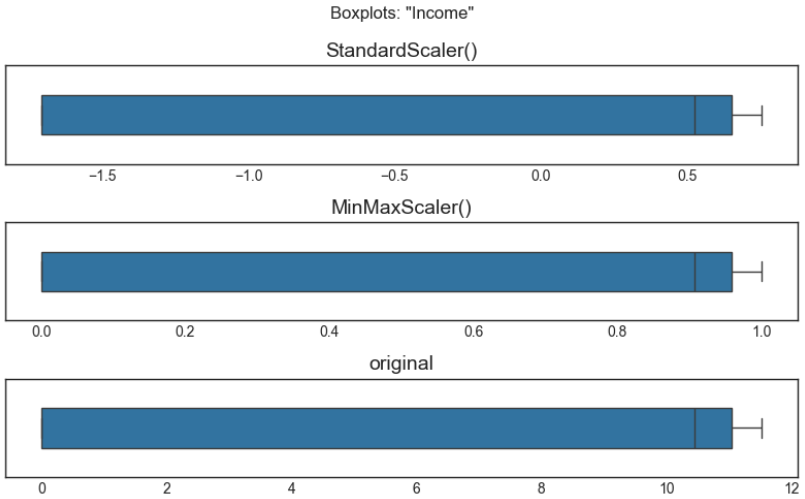

Figure 14 - Scaling Comparison Using Income

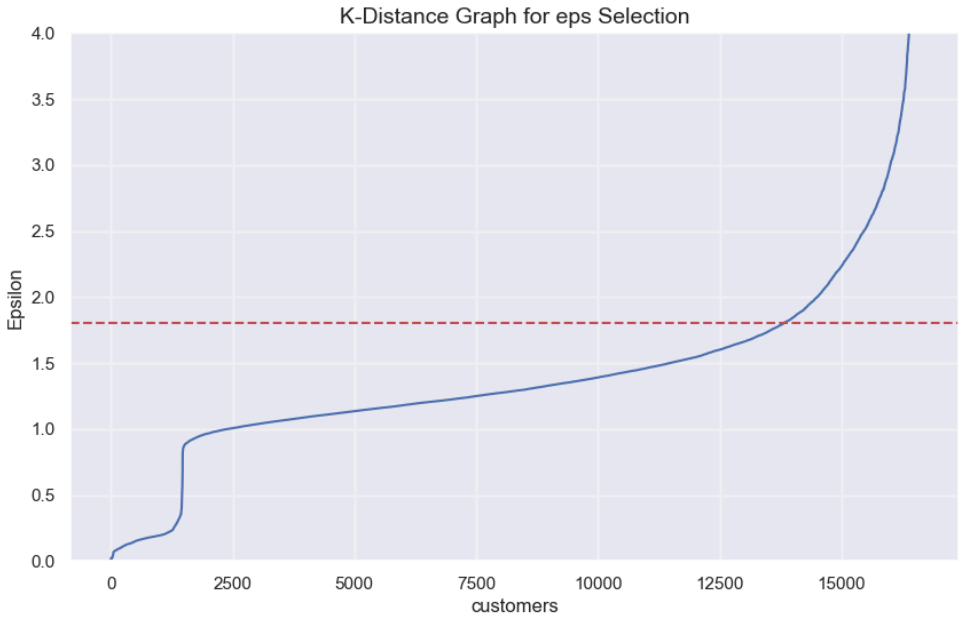

Figure 15 - K-Distance Graph for eps Selection

Figure 16 - Value Based - R2 plot for various clustering methods

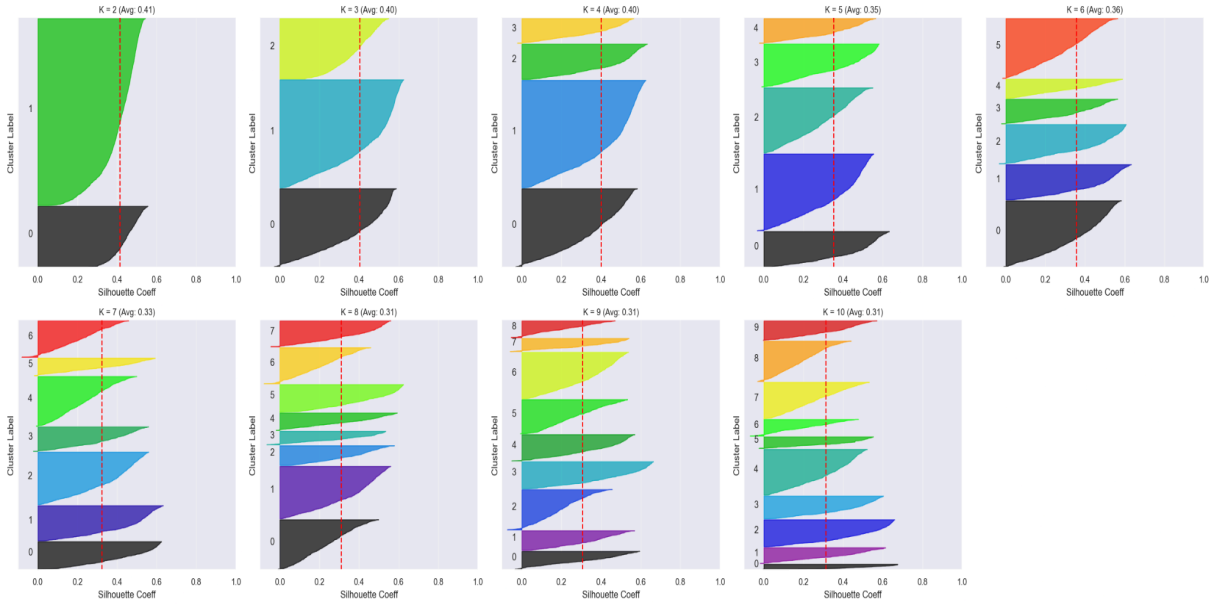

Figure 17 - Value Based - Silhouette Plots

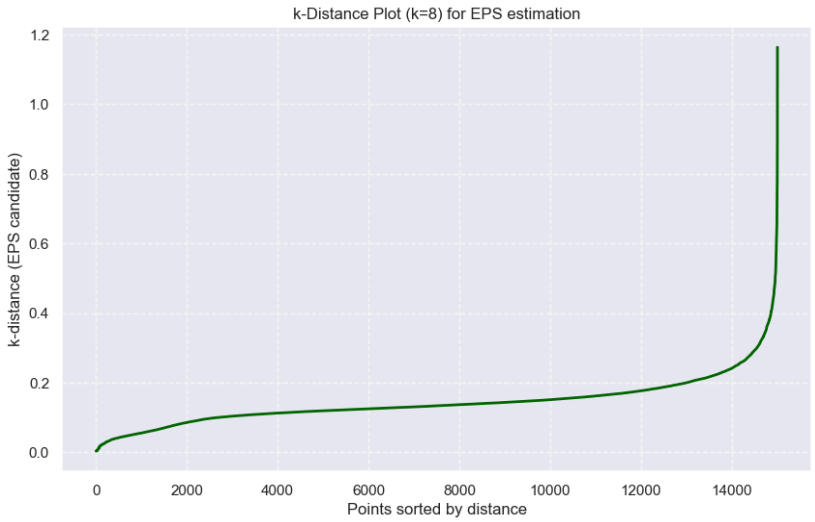

Figure 18 - Seasonality - K-Distance Plot for EPS Estimation

```
                    CLV    Income  PointsUtilizationRate    Tenure
Cluster_MS_val
0              0.304810  0.579511              -0.061984  0.067785
1             -0.474672 -1.706172              -0.030252  0.112212

Cluster Sizes:
Cluster_MS_val
0    11331
1     3677
dtype: int64
```

Figure 19 - Value Based - Mean Shift Output

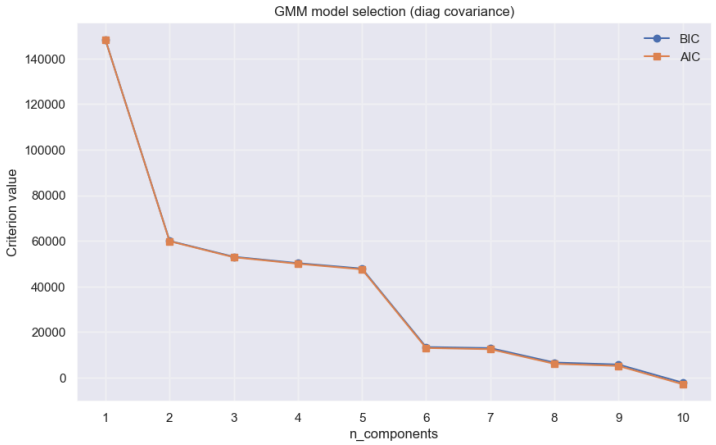

Figure 20 - Value Based - Gaussian Mixture Model (Diagonal Covariance)

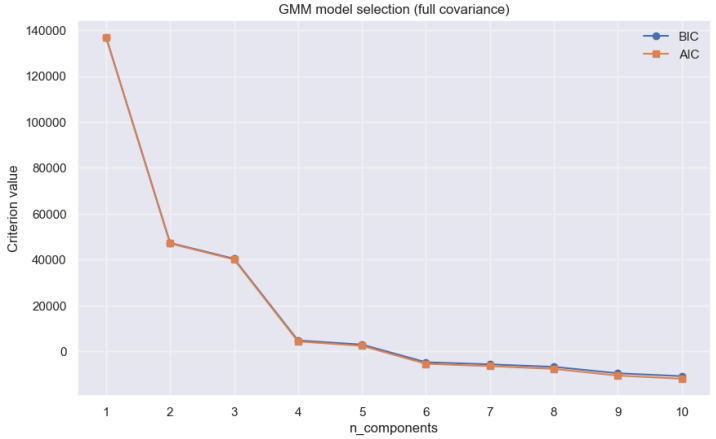

Figure 21 - Value Based - Gaussian Mixture Model (Full Covariance)

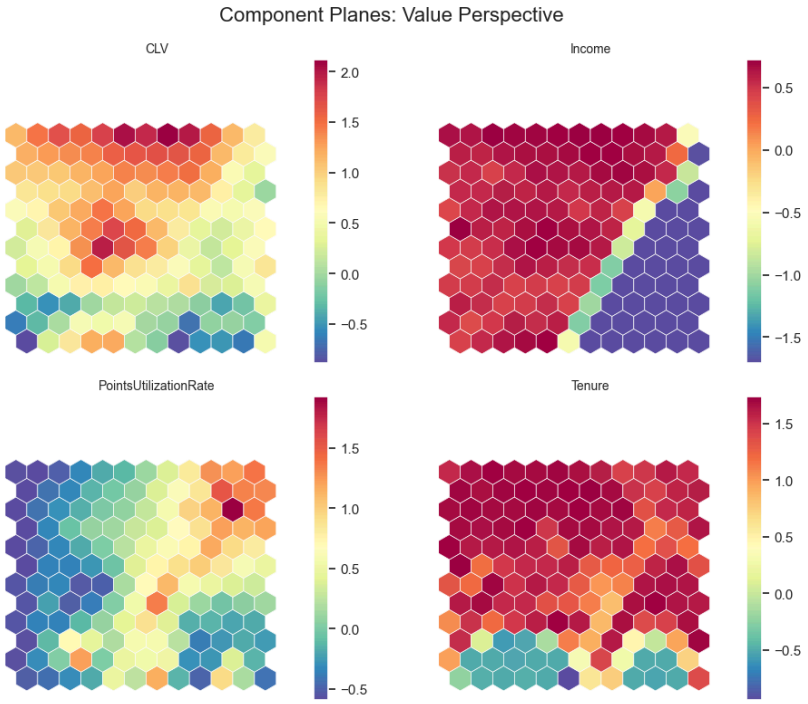

Figure 22 - Value Based - Component Planes

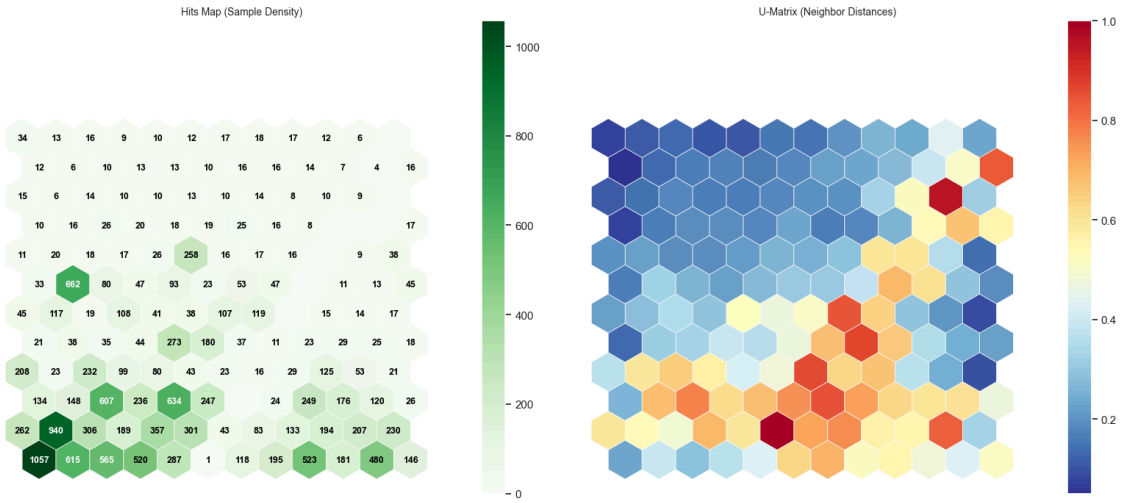

Figure 23 - Value Based - U-Matrix

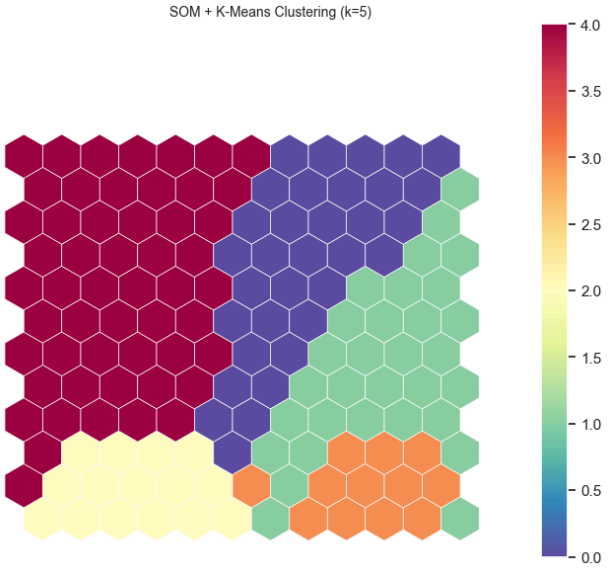

Figure 24 - Value Based - SOMS with K-means

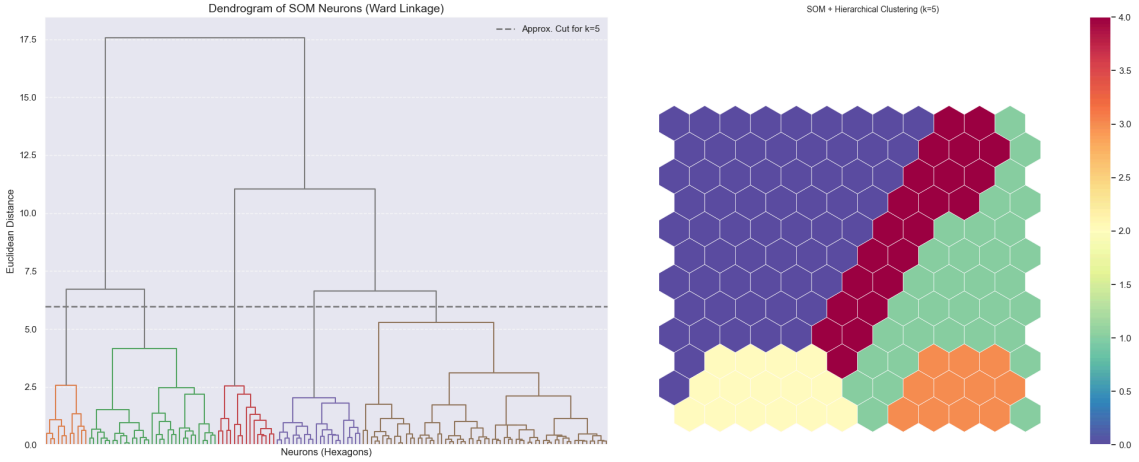

Figure 25 - Value Based - SOMS with Hierarchical clustering

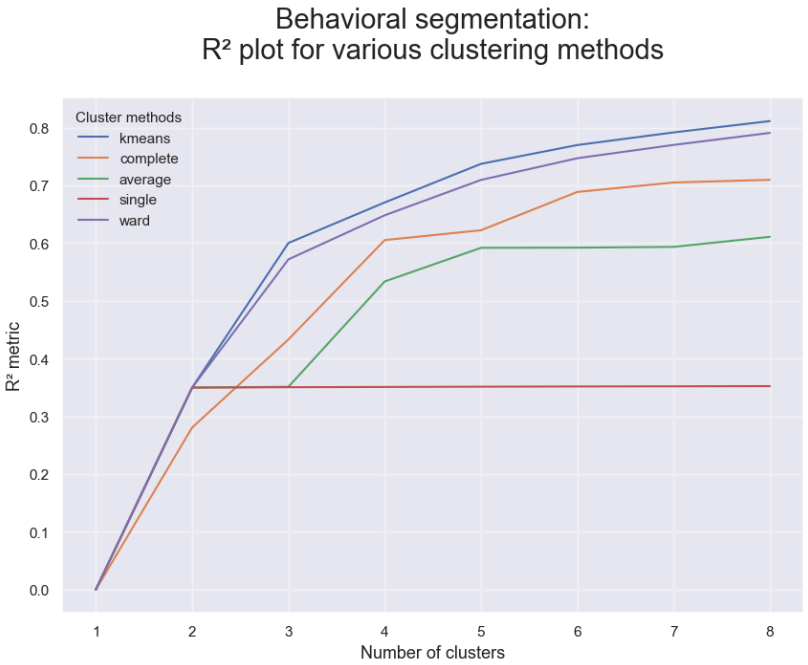

Figure 26 - Behavioral - R2 plot for various clustering methods

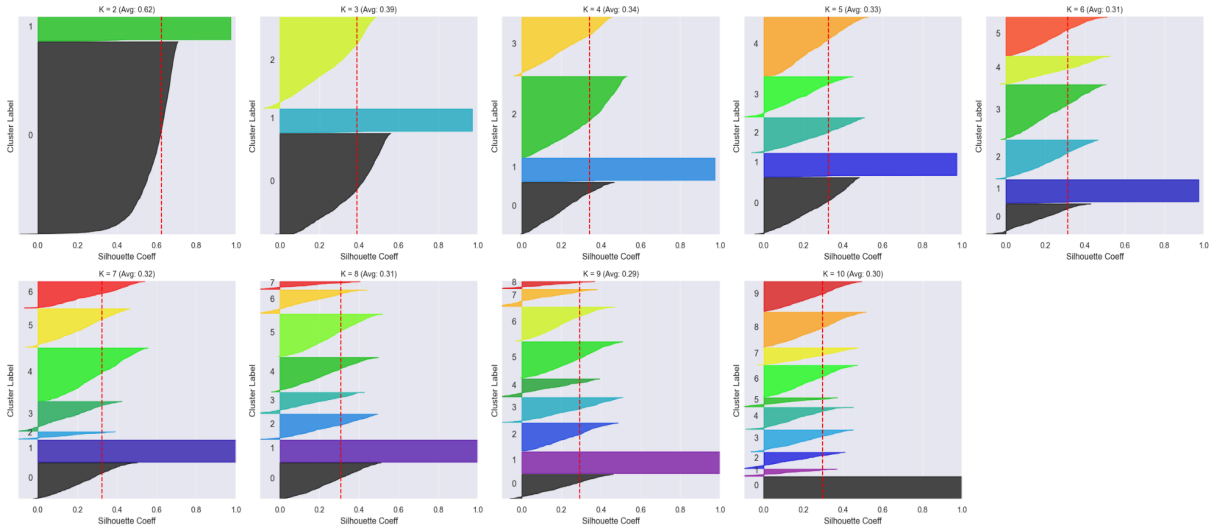

Figure 27 - Behavioral - Silhouette Plots

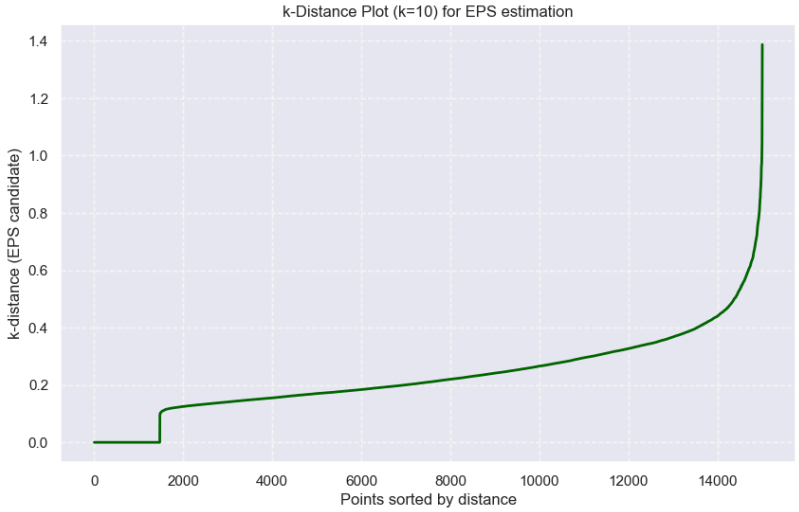

Figure 28 - Behavioral - K-Distance Plot for EPS Estimation

```
                    NumFlights  Avg_Distance_Flight  recency_months  \
Cluster_MS_behave
0                     0.380839             0.088449       -0.167991
1                    -2.735457            -1.439752       -0.585659

                    Companion_Ratio  Flight_Trend_Slope
Cluster_MS_behave
0                          0.118713            0.133270
1                         -1.695019           -0.520084

Cluster Sizes:
Cluster_MS_behave
0    13456
1     1552
dtype: int64
```

Figure 29 - Behavioral - Mean Shift Output

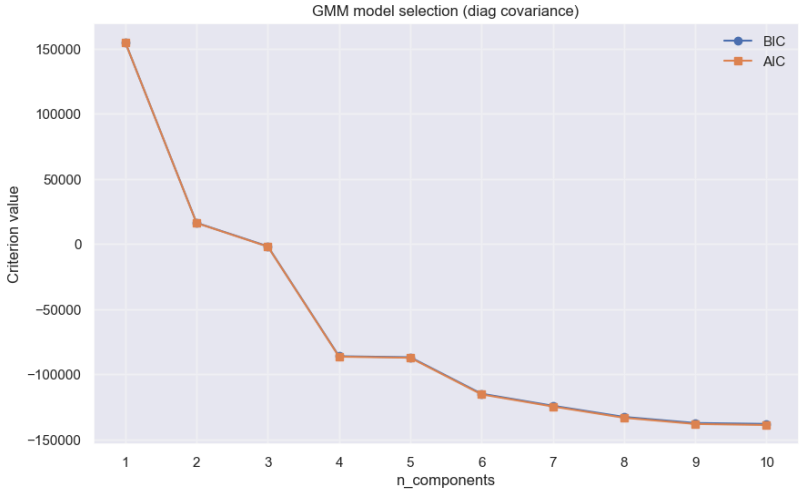

Figure 30 - Behavioral - Gaussian Mixture Model (Diagonal Covariance)

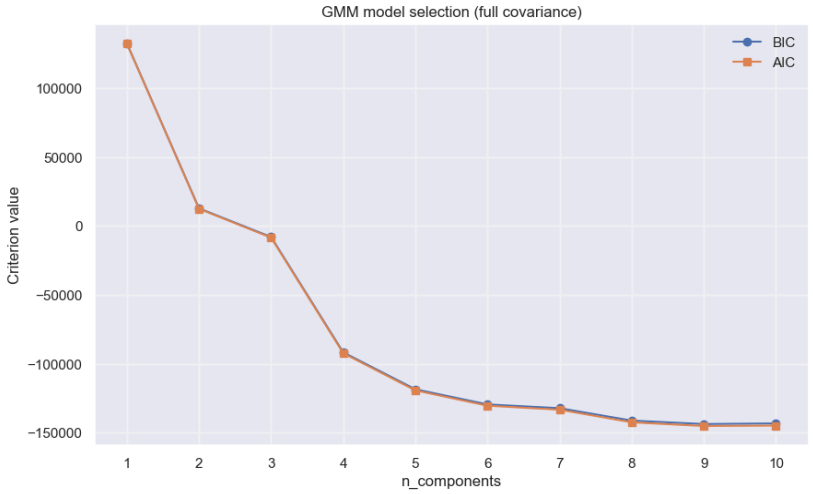

Figure 31 - Behavioral - Gaussian Mixture Model (Full Covariance)

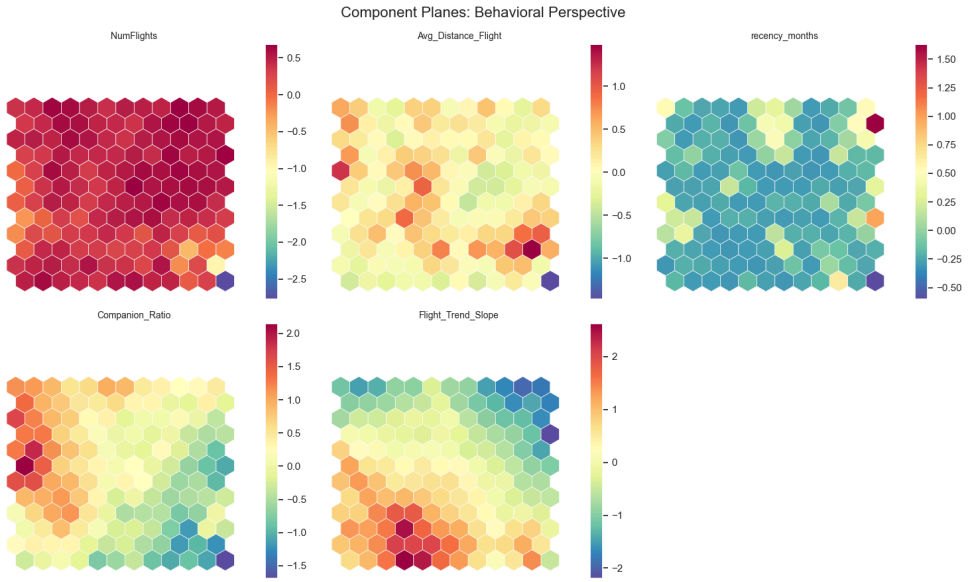

Figure 32 - Behavioral - Component Planes

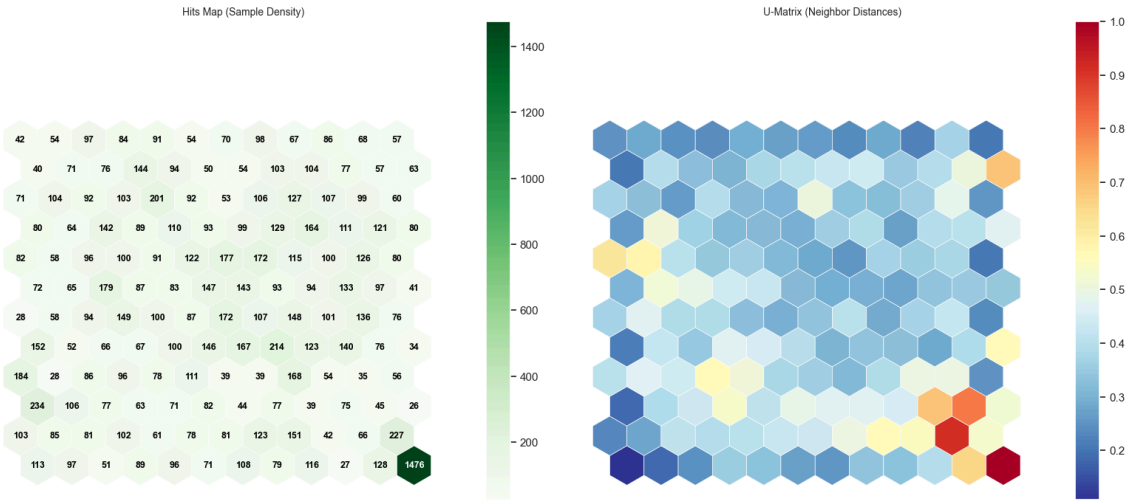

Figure 33 - Behavioral - U-Matrix

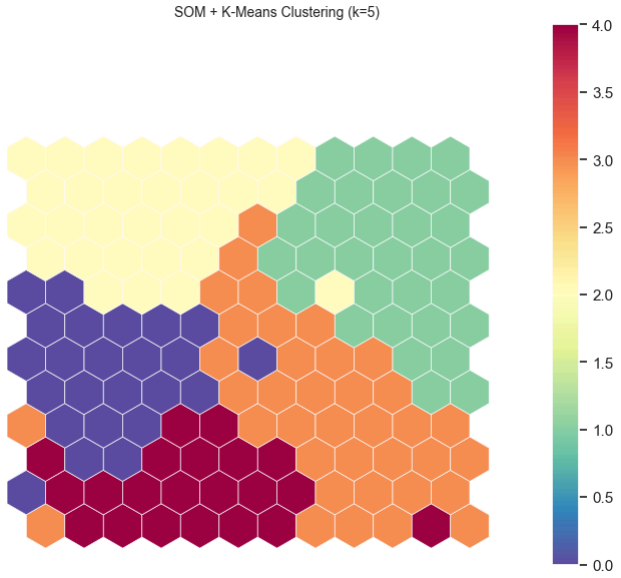

Figure 34 - Behavioral - SOMS with K-means

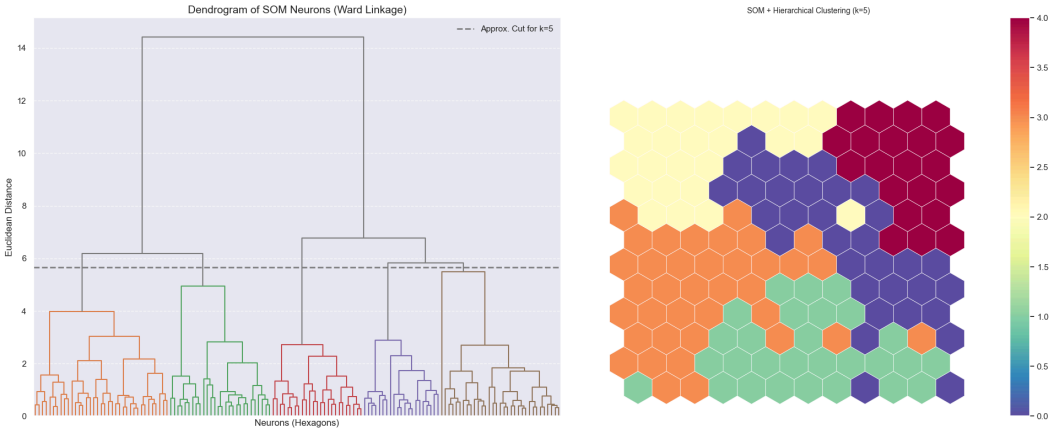

Figure 35 - Behavioral - SOMS with Hierarchical Clustering

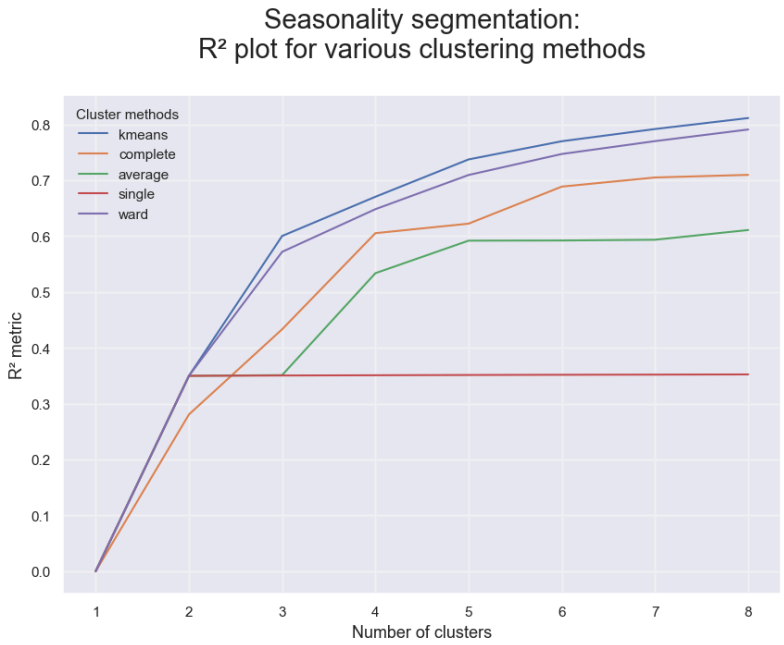

Figure 36 - Seasonality -R2 plot for various clustering methods

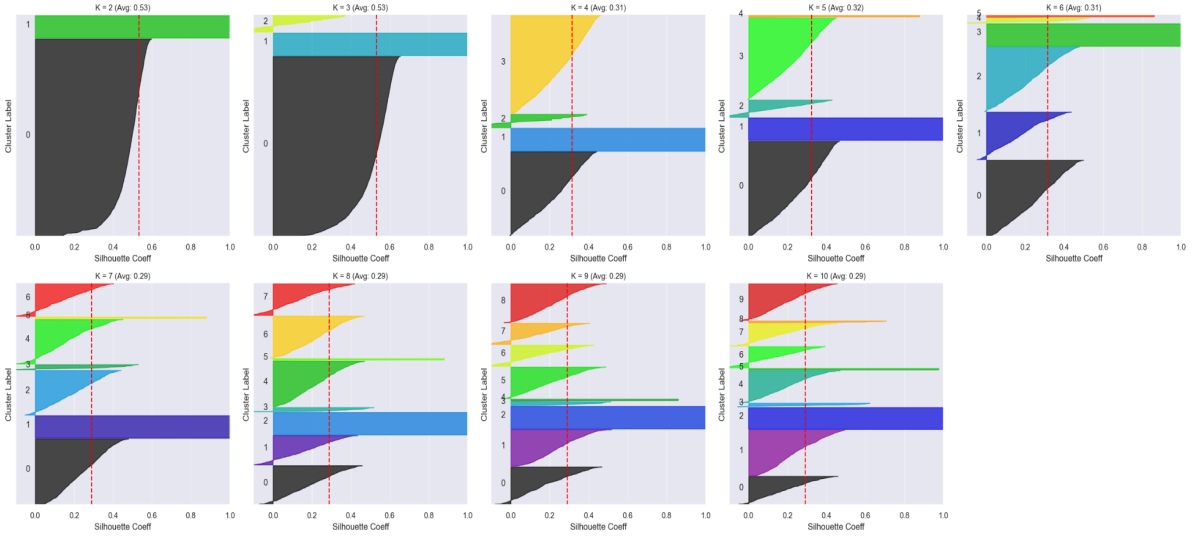

Figure 37 - Seasonality - Silhouette Plots

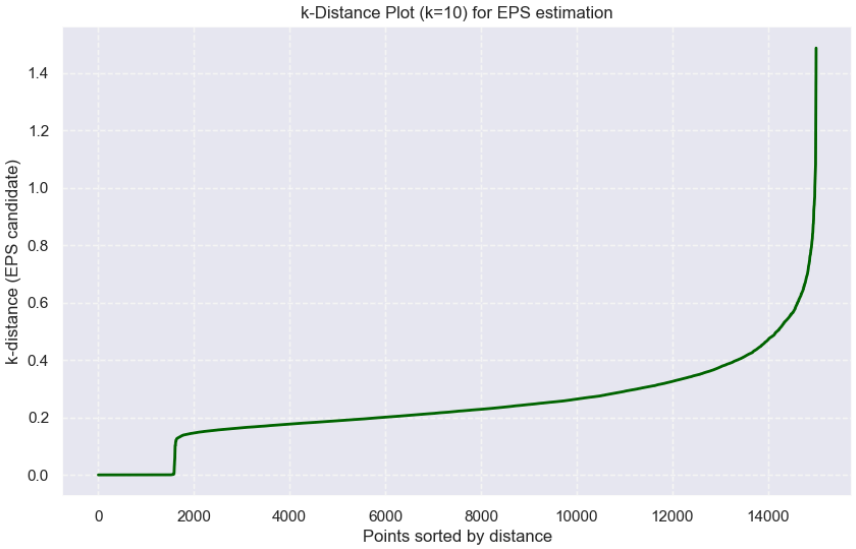

Figure 38 - Seasonality - K-Distance Plot for EPS Estimation

Cluster_MS_season	Fall_Ratio	Spring_Ratio	Summer_Ratio	Winter_Ratio
0	0.059061	0.210136	0.283665	0.017095
1	-1.360274	-1.418826	-1.660585	-1.278877
2	3.225712	-1.407216	-1.448851	-0.111739
3	-0.576804	-1.414078	-1.652724	3.975444

Cluster_MS_season	Seasonality_Index
0	-0.129779
1	-1.469139
2	2.113138
3	3.477844

Cluster Sizes:

Cluster_MS_season	
0	13094
1	1473
2	293
3	148

dtype: int64

Figure 39 - Seasonality - Mean Shift Output

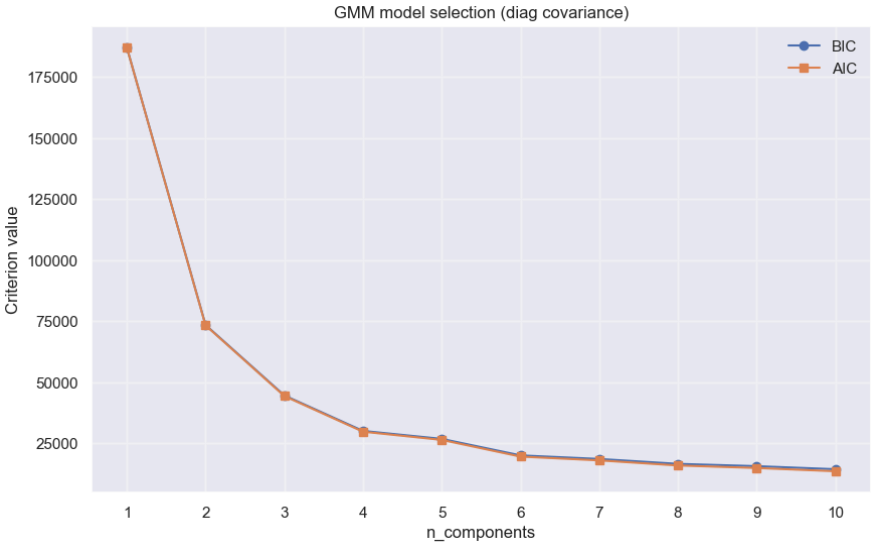

Figure 40 - Seasonality - Gaussian Mixture Model (Diagonal Covariance)

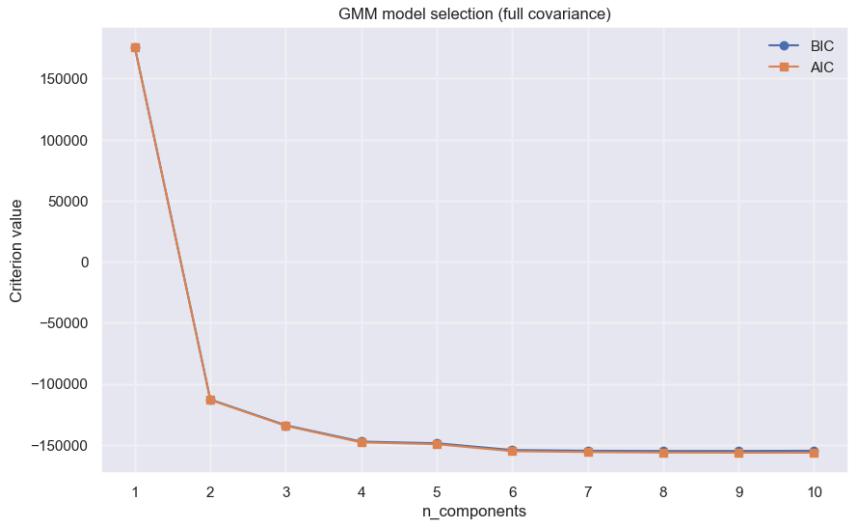

Figure 41 - Seasonality - Gaussian Mixture Model (Full Covariance)

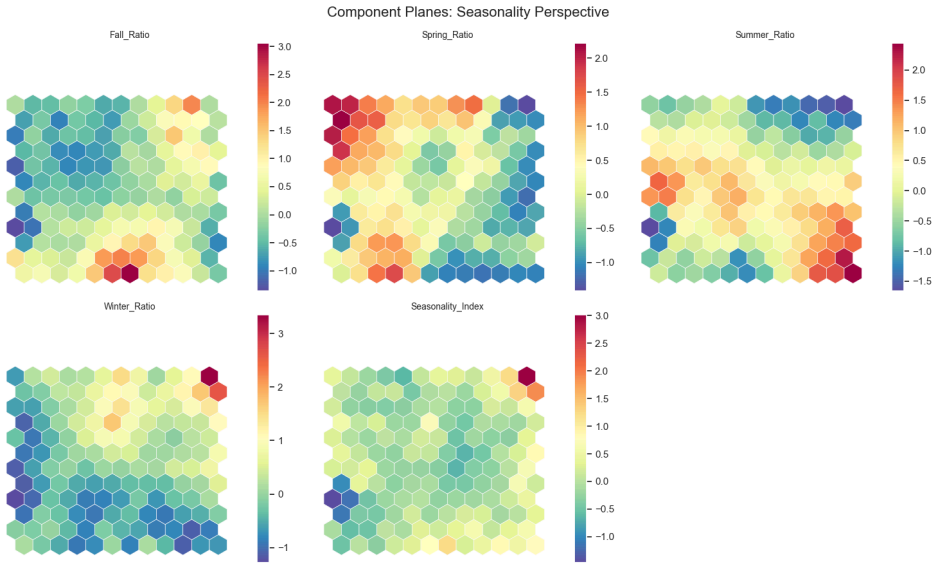

Figure 42 - Seasonality - Component Planes

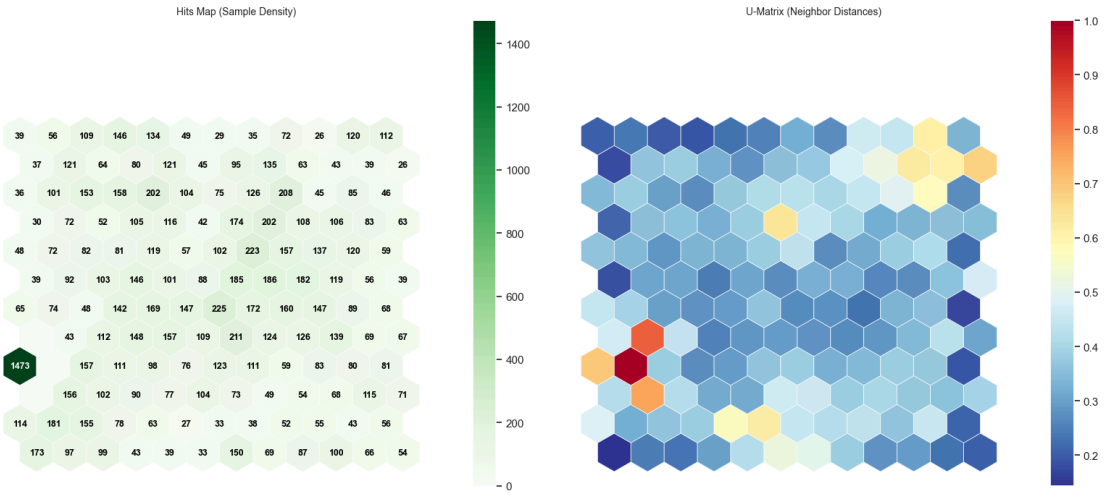

Figure 43 - Seasonality - U-Matrix

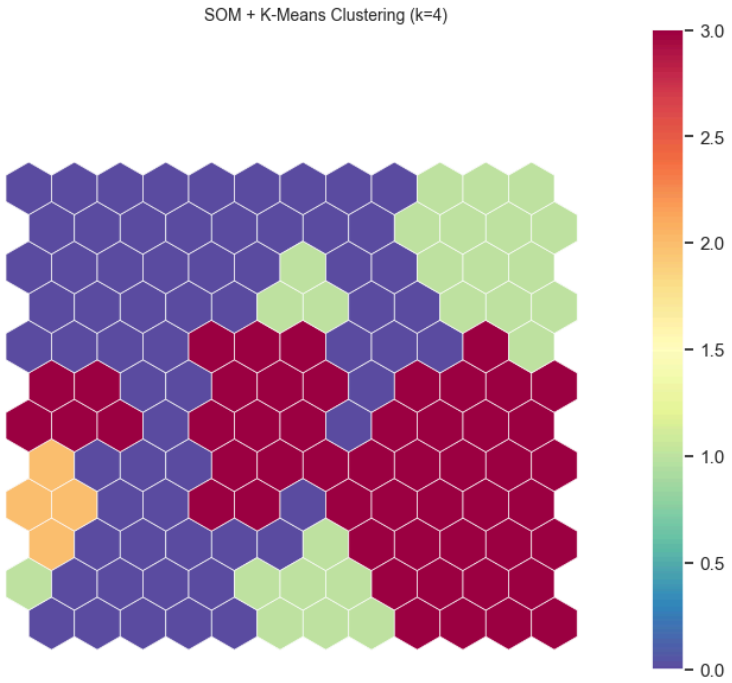

Figure 44 - Seasonality- SOMS with K-means

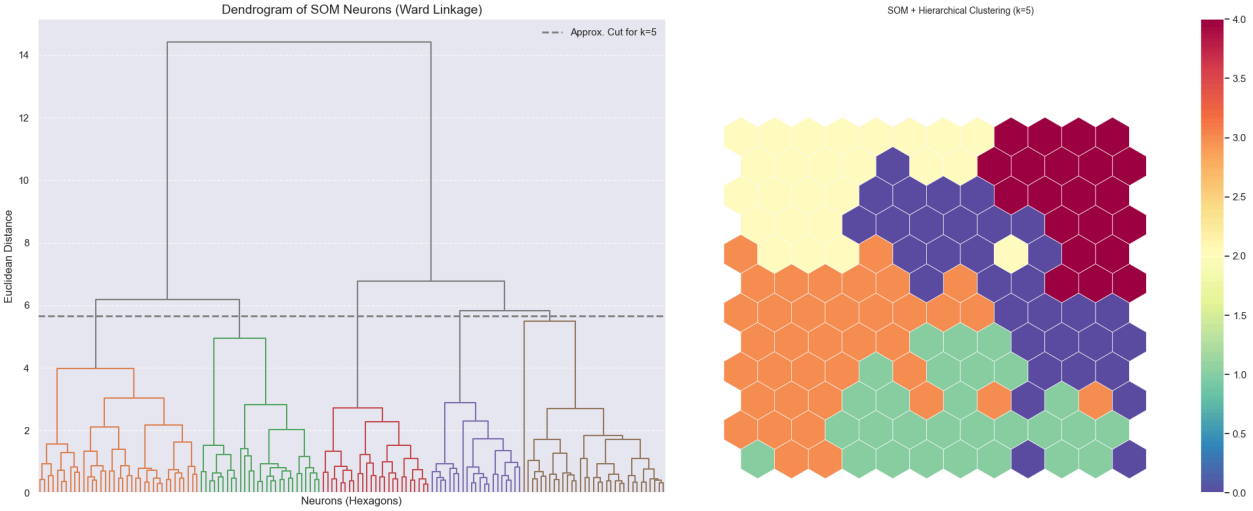

Figure 45 - Seasonality - SOMS with Hierarchical Clustering

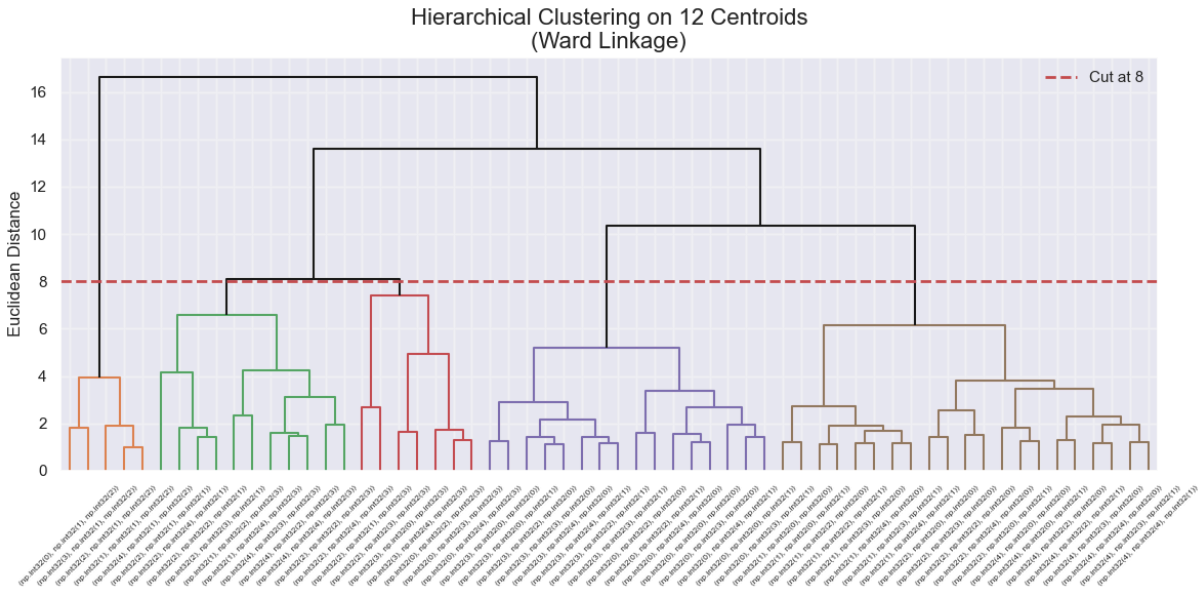

Figure 46 - Hierarchical Clustering used to merge our final clusters

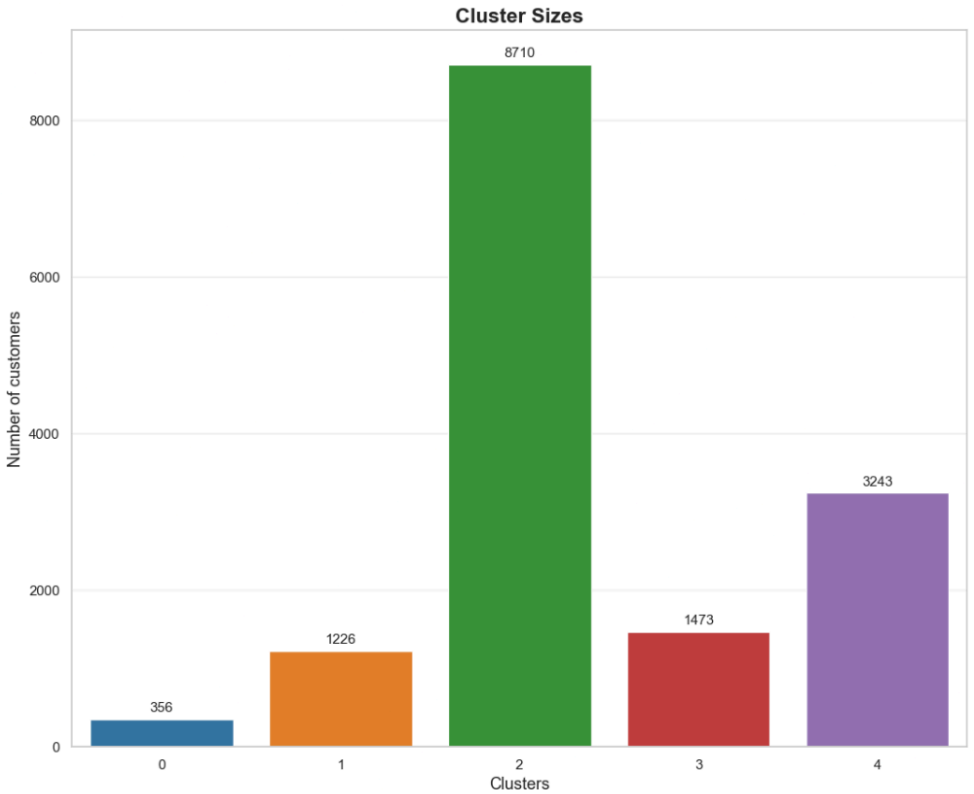

Figure 47 - Dimension of each cluster

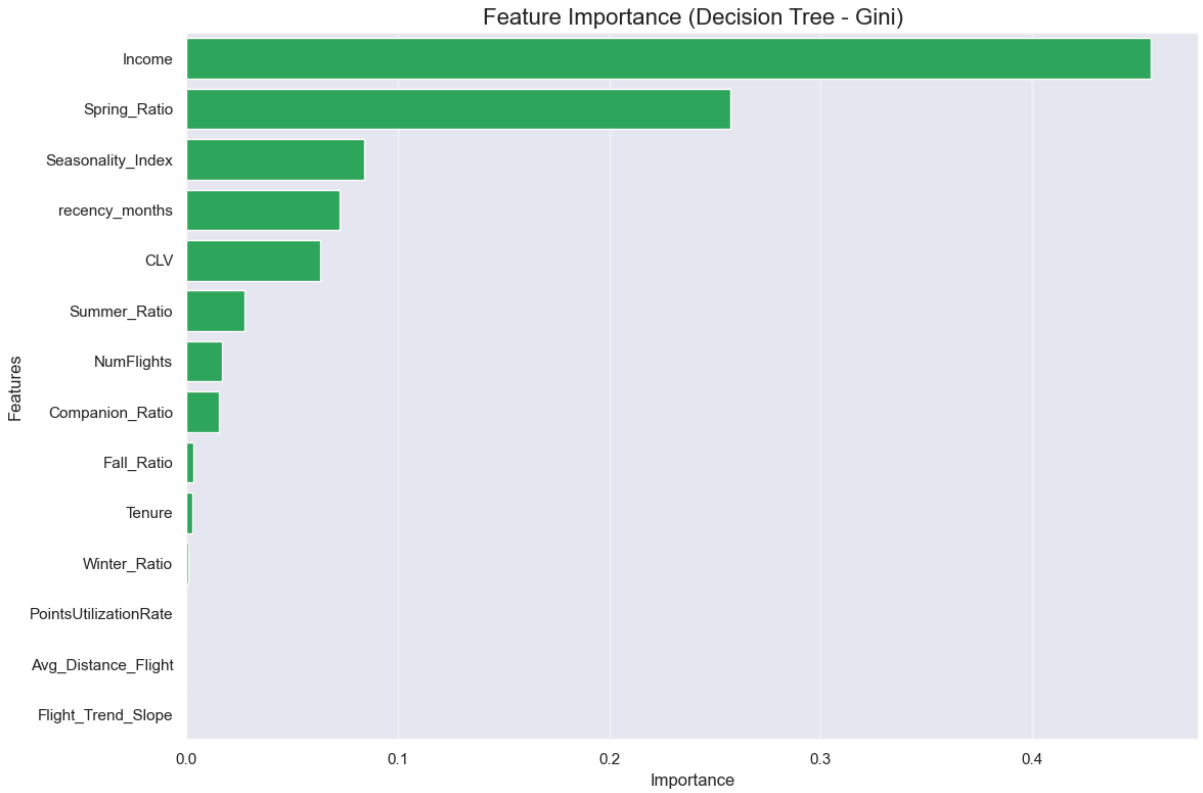

Figure 48 - Feature Importance using Decision Tree

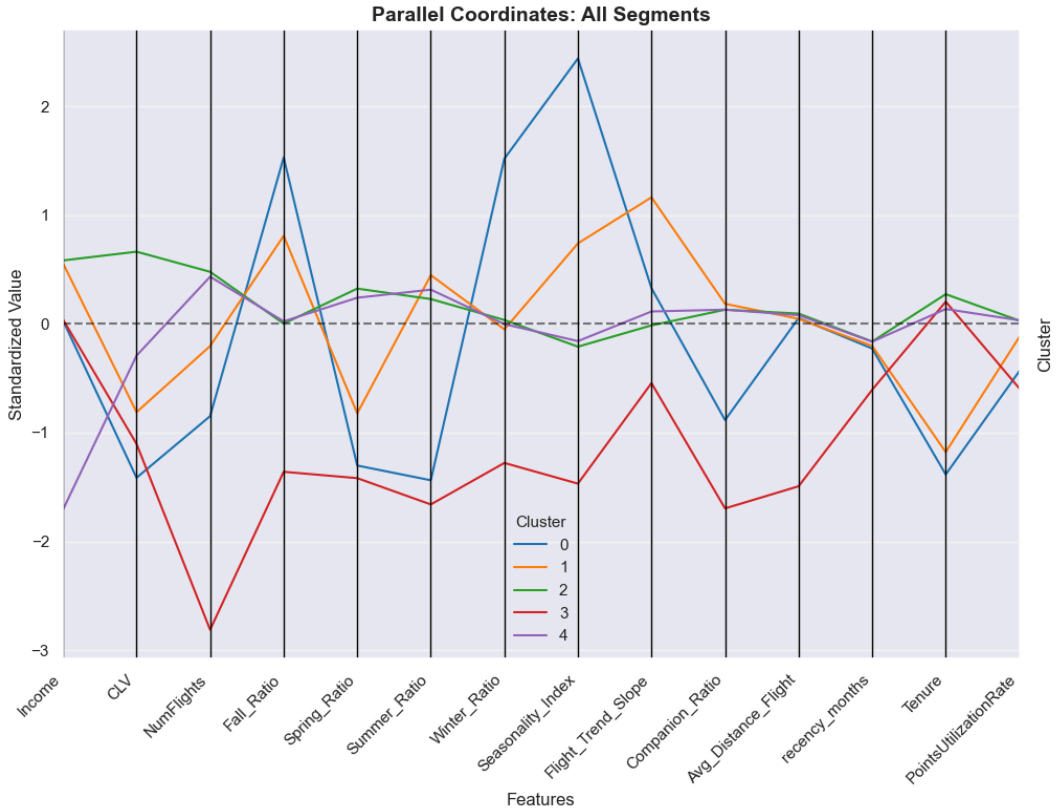

Figure 49 - Parallel Coordinates

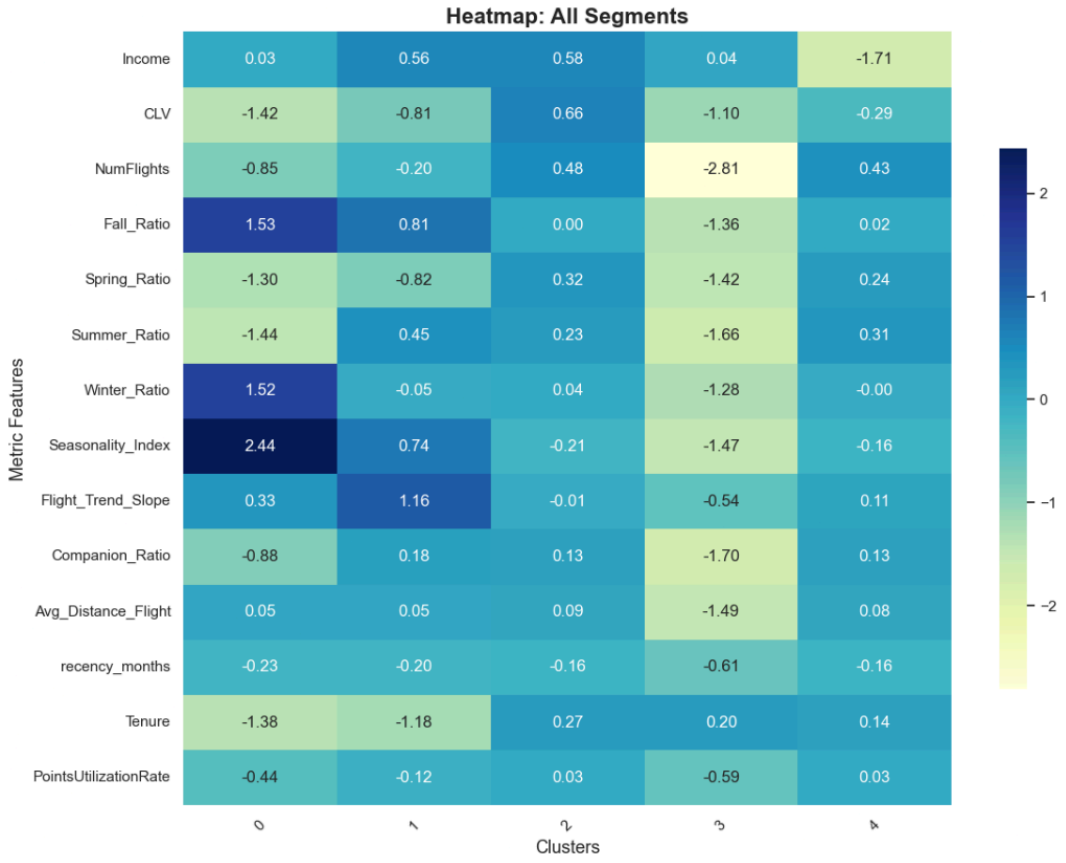

Figure 50 - Heatmap of the variables in each cluster

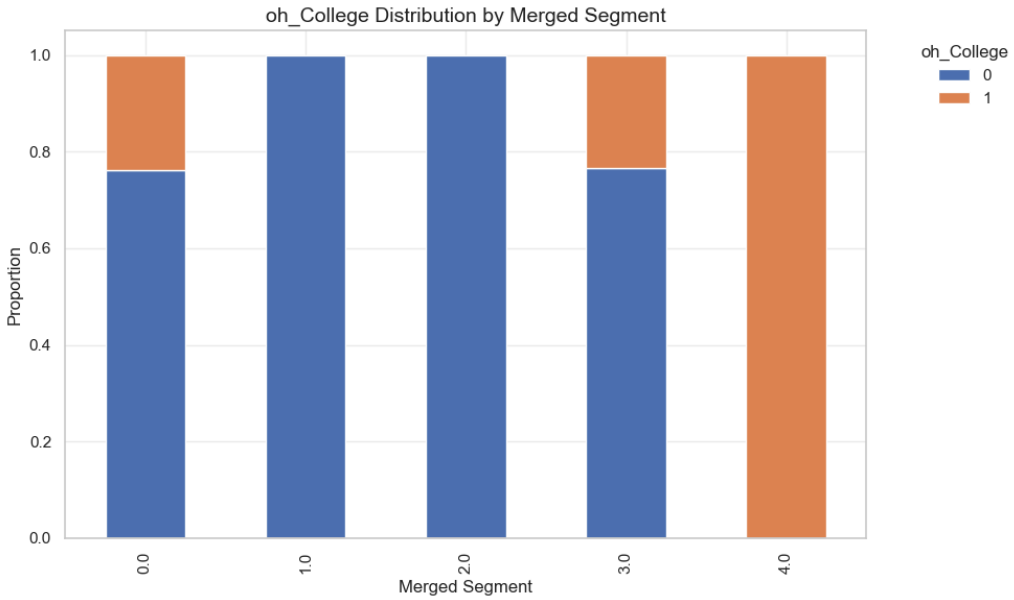

Figure 51 - Distribution of College degree per Cluster

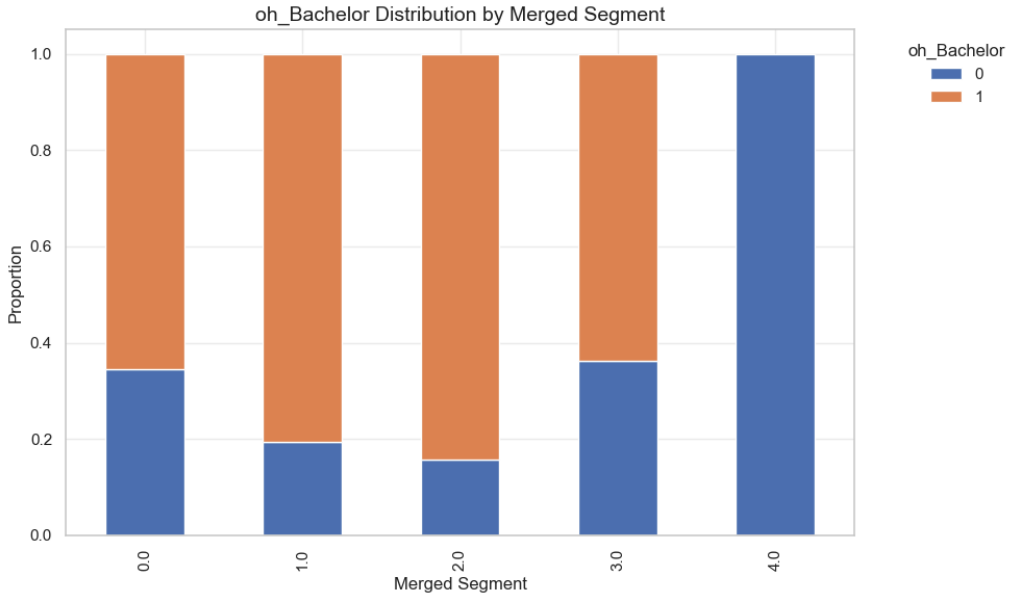

Figure 52 - Distribution of Bachelor degree per Cluster

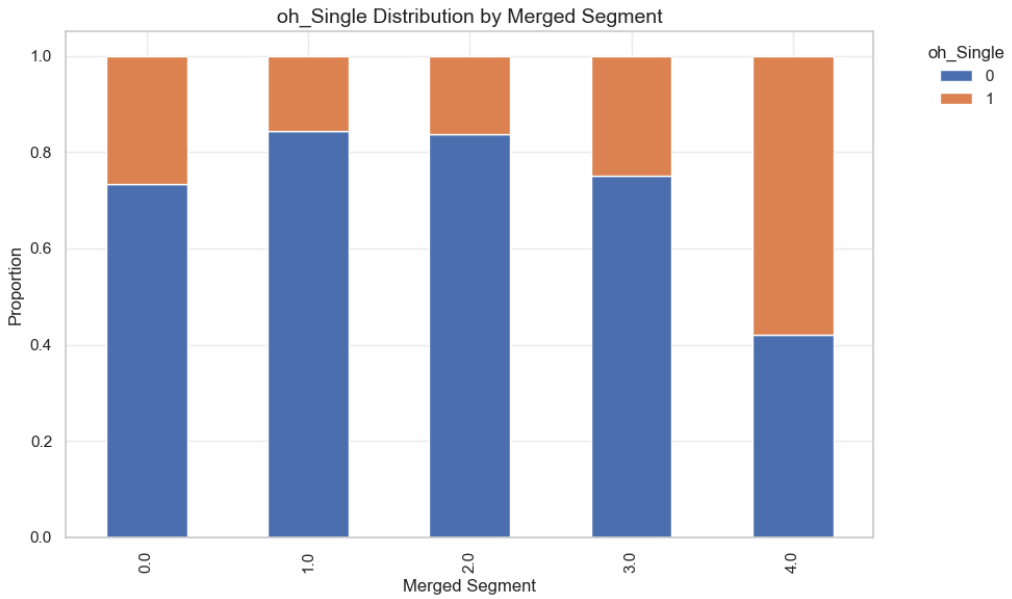

Figure 53 - Distribution of Single customers per Cluster

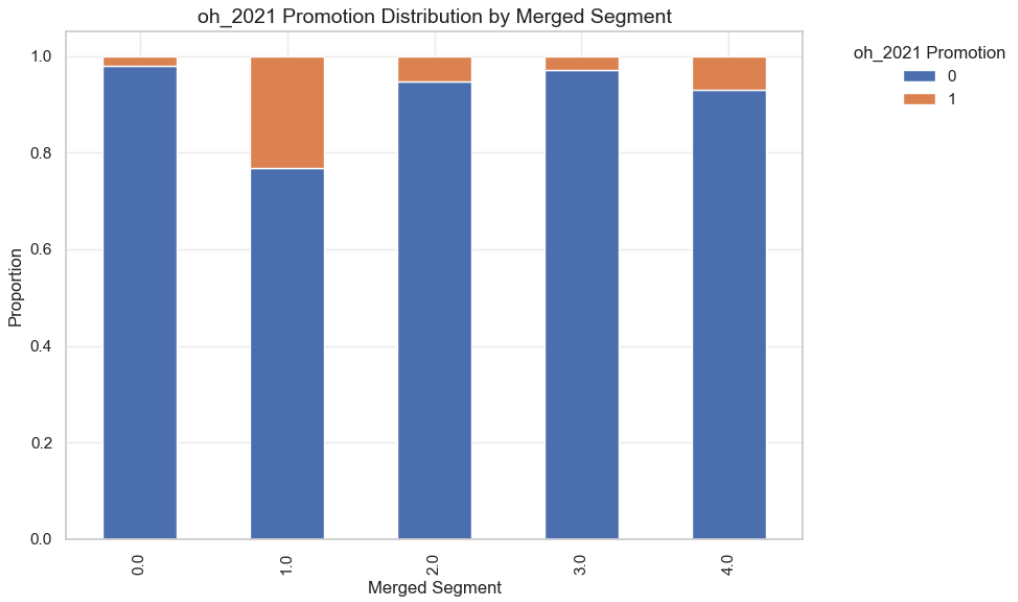

Figure 54 - Distribution of promotion 2021(enrollment type) per Cluster

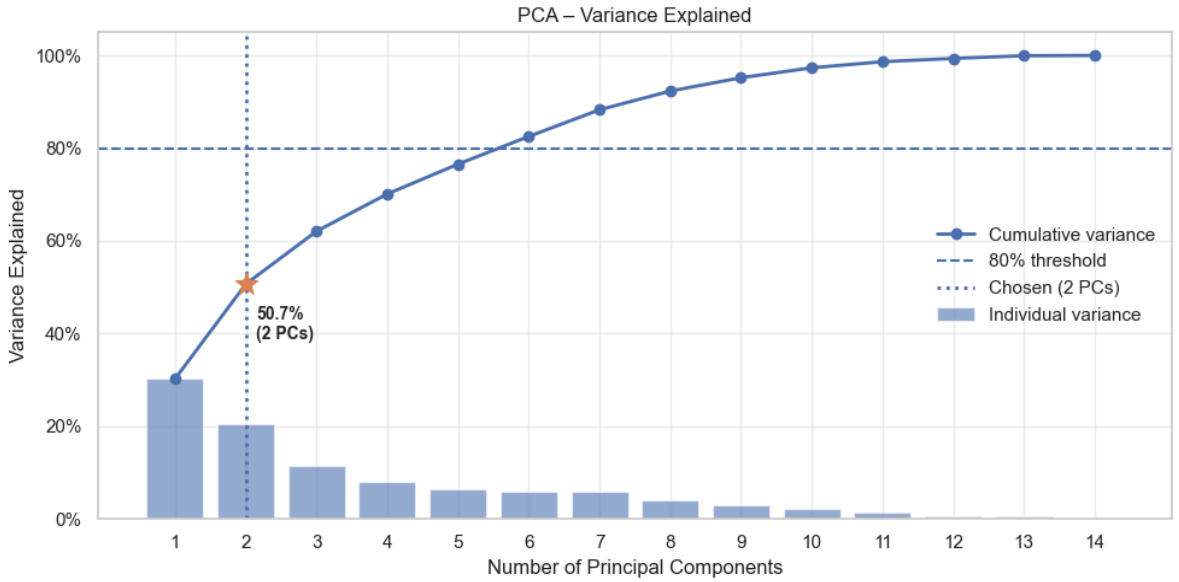

Figure 55 - PCA variance

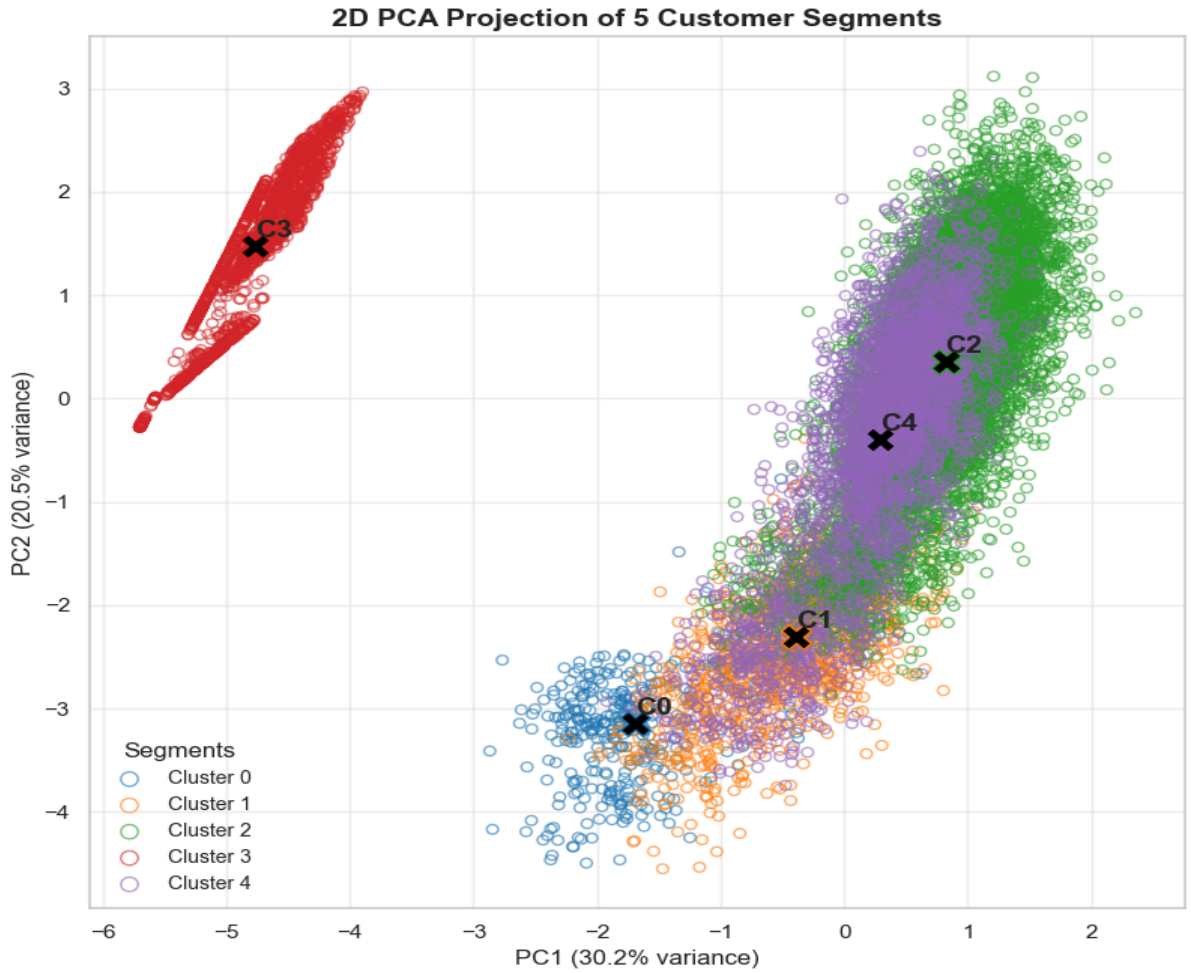

Figure 56 - 2D PCA visualization of the 5 clusters

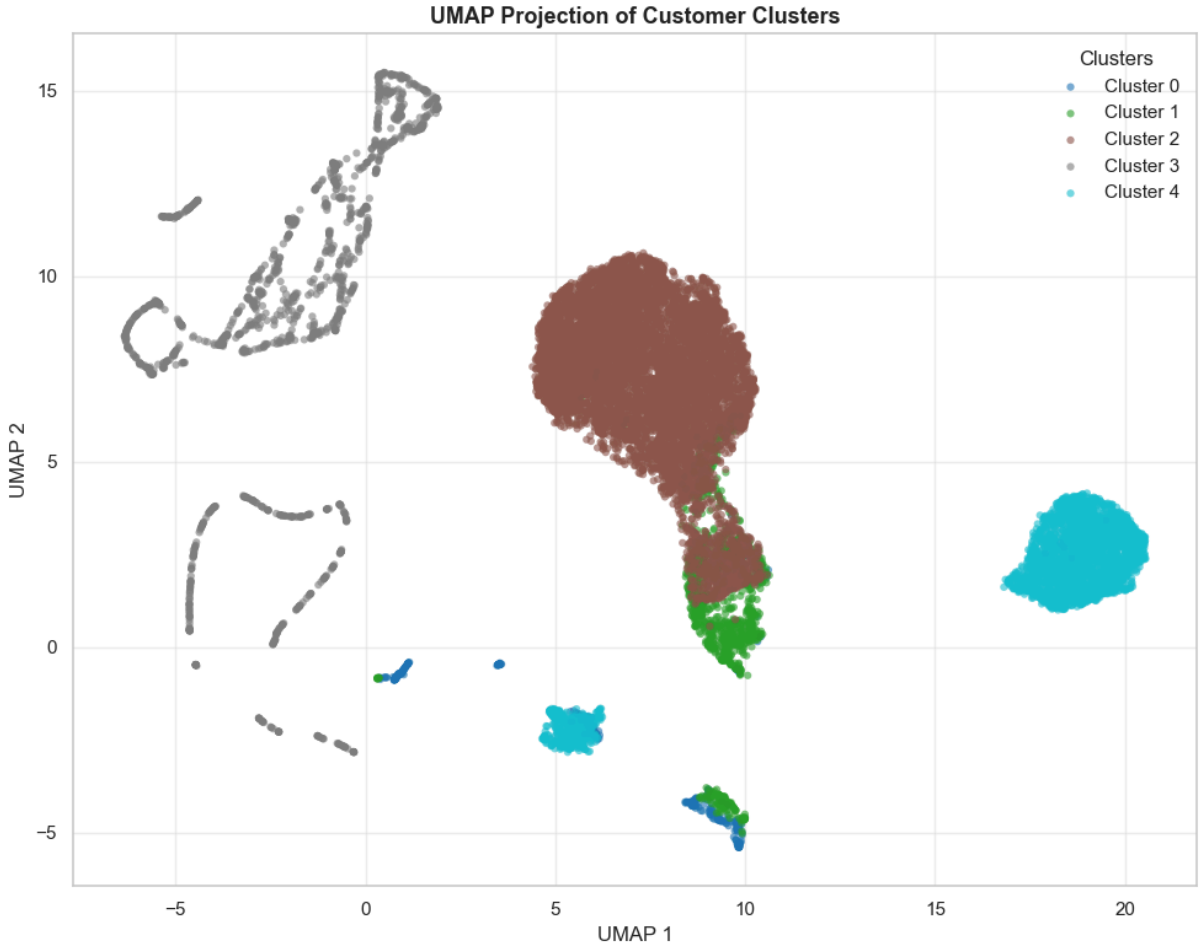

Figure 57 - UMAP visualization of the 5 clusters

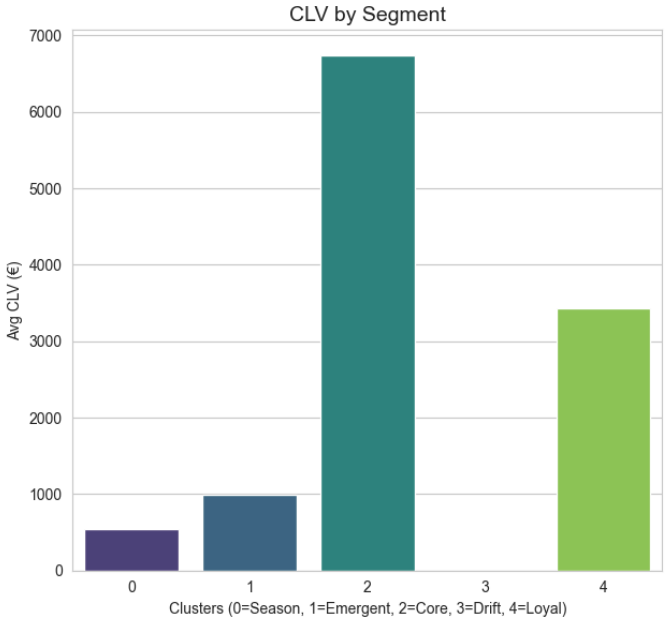

Figure 58 - Customer lifetime value by cluster

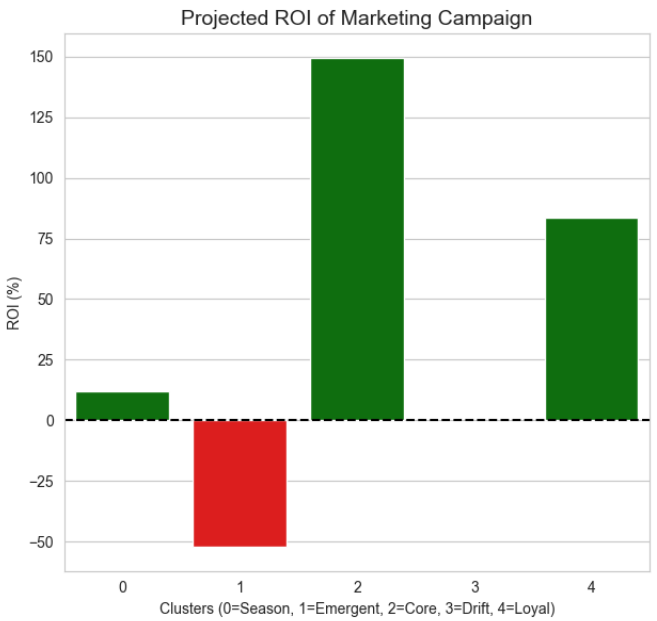

Figure 59 - Projected ROI in percentage per Cluster

Appendix B. Tables

Table 2 - Metrics for each perspective of the K-Means

Perspective	Number of Clusters	R^2 Score	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
Value	5	0.7369	0.353	0.999	10507.8
Behavioral	5	0.7676	0.317	1.012	12387.1
Seasonality	4	0.5703	0.309	1.134	6636.5

Table 3 - Value perspective metrics

Perspective	Model Name	Number of Clusters	R^2 Score	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
Value	K-Means	5	0.7369	0.353	0.999	10507.8
Value	SOM + K-Means	5	0.6262	0.234	1.244	6284.5
Value	SOM + Hierarchical	5	0.6248	0.226	1.243	6245.8
Value	Gaussian Mixture	4	0.5127	0.217	1.507	5261.7
Value	DBSCAN	15	0.3792	-0.216	0.913	616.3
Value	HDBSCAN	3	0.3606	0.352	1.136	4224.6

Table 4 - Behavioral perspective metrics

Perspective	Model Name	Number of Clusters	R^2 Score	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
Behavioral	K-Means	5	0.7676	0.317	1.012	12387.0
Behavioral	Gaussian Mixture	4	0.5427	0.134	1.996	5935.5
Behavioral	DBSCAN	4	0.5288	0.244	0.642	5198.7
Behavioral	HDBSCAN	2	0.4729	0.629	0.3	13446.8
Behavioral	SOM + K-Means	5	0.4416	0.115	1.41	2965.9
Behavioral	SOM + Hierarchical	5	0.4007	0.087	1.699	2507.4

Table 5 - Seasonality *perspective metrics*

Perspective	Model Name	Number of Clusters	R^2 Score	Silhouette Score	Davies-Bouldin	Calinski-Harabasz
Seasonality	K-Means	4	0.5703	0.309	1.134	6636.5
Seasonality	DBSCAN	9	0.5661	0.431	0.45	2322.5
Seasonality	SOM + K-Means	4	0.5462	0.291	1.175	6020.3
Seasonality	Gaussian Mixture	4	0.4617	0.257	2.474	4289.1
Seasonality	HDBSCAN	3	0.4515	0.567	0.748	6090.0
Seasonality	SOM + Hierarchical	4	0.2517	0.121	1.656	1682.4

Table 6 - Outliers classification

Predicted Cluster	Number of Outliers
0	316
1	452
2	500
3	0
4	298

Table 7 - Top 3 RFM Profiles per Cluster

Cluster	RFM_Score	Count	% Cluster
0	211	161	45.22
0	111	48	13.48
0	311	38	10.67
1	322	295	24.06
1	321	155	12.64
1	311	145	11.83
2	555	648	7.44
2	455	438	5.03
2	545	401	4.6
3	111	883	59.95
3	112	451	30.62
3	113	128	8.69
4	321	305	9.4
4	554	203	6.26
4	543	202	6.23

Table 8 - Cumulative Coverage of Top 3

Cluster	% Cluster
0	69.37
1	48.53
2	17.07
3	99.26
4	21.89

Table 9 - Average Financial Estimates by Cluster

Cluster	Annual Revenue	Annual Profit	CLV
0	1495.51	269.19	538.38
1	2077.34	249.28	997.12
2	7481.83	1122.27	6733.64
3	0	0	0
4	4287.94	428.79	3430.35

Table 10 - Campaign ROI Projection

Cluster	Annual Revenue	Marketing Cost	Number of Customers	Projected Revenue	Profit Gain	ROI(%)
0	532400.00	7120.0	356	612260.00	7986.00	12.16
1	2546820.00	79690.0	1226	2928843.00	38202.30	-52.06
2	65166708.33	391950.0	8710	74941714.58	977500.62	149.39
3	0.00	0.0	1473	0.00	0.00	NA
4	13905775.00	113505.0	3243	15991641.25	208586.62	83.77

Appendix C . Financial Impact Modeling

This section translates the customer segmentation into financial outcomes over a one-year marketing campaign horizon. The objective is to quantify expected returns, assess investment trade-offs, and support data-driven prioritization of customer segments.

Customer Lifetime Value Estimation

The Customer Lifetime Value (CLV) variable created by us in the feature engineering represented more like a "score" used for segmentation rather than a monetary measure. To enable financial evaluation, a new profit-based *CLV* was constructed.

Annual revenue was estimated by the number of flights per year and multiplying it by an assumed average ticket price per segment. Annual profit was then computed using segment-specific operating margins. Finally, *CLV* was calculated as annual profit multiplied by an estimated retention horizon, reflecting expected customer longevity.

This approach yields a financially interpretable *CLV* that allows direct comparison across segments. The results indicate a strong concentration of value in cluster 2, with an average *CLV* of approximately €6,734. Cluster 4 follows with €3,430, driven by long retention despite lower annual spend. Cluster 1 and cluster 0 show substantially lower CLV, while cluster 3 generates negligible lifetime value because all of the customers have zero flights. (See figure 58 and table 9)

ROI Projections by Segment

ROI projections were simulated for a one-year targeted marketing campaign, assuming a 15% lift in annual revenue for contacted customers. Marketing costs were assigned per customer and aggregated at cluster level. Incremental profits were estimated using a conservative margin on the additional revenue.

The results show substantial heterogeneity across segments. Cluster 2 delivers a very high ROI of approximately 149%, confirming its role as the primary financial engine of the customer base. Cluster 4 also generates a strong positive ROI of around 84%, supporting investment in retention-oriented initiatives. Cluster 0 achieves a modest positive ROI of roughly 12%, indicating that narrowly timed seasonal campaigns can be marginally profitable.

Cluster 1 exhibits a negative one-year ROI of −52%. This outcome is intentional rather than problematic: investment in this segment is designed to accelerate customer progression toward the Core Flyers segment, trading short-term profitability for long-term value creation. Cluster 3 was excluded from paid marketing activities and therefore shows no meaningful return. (See figure 59 and table 10)

Cost–Benefit Analysis

At the portfolio level, the total marketing investment of approximately €592,000 generates a projected incremental profit of €1.23 million, resulting in a net benefit of €640,000 and an overall ROI of 108%. These results demonstrate that a segmented investment strategy substantially outperforms

a uniform allocation of marketing resources, with most value creation driven by high-engagement segments.

Table 11 - Final Cost-benefit analysis

FINAL COST-BENEFIT ANALYSIS	
Total Investment Required	592.265,00€
Total Projected Profit Gain	1.232.275,55€
Net Benefit (Profit - Cost)	640.010,55€
Global Project ROI	108.06%

Investment Timeline and Expected Returns

To assess financial risk and payback dynamics, a cumulative investment timeline was constructed. As illustrated in figure 60. The analysis shows that the initial marketing investment is recovered within the first half of the year, after which cumulative profits increase steadily.

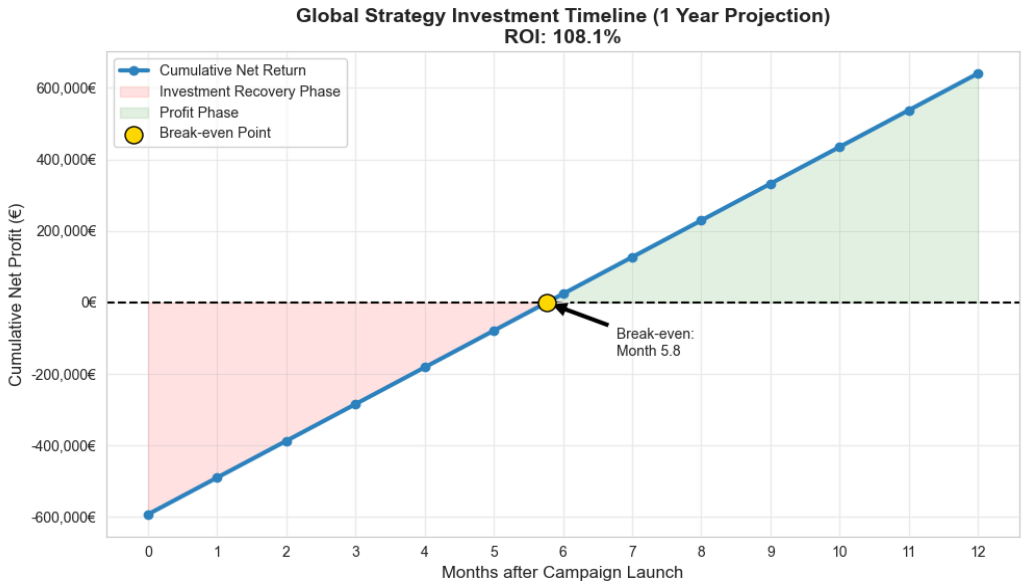

Figure 60 - Global Strategy Investment Timeline

Assumptions and Scope

This financial model is based on explicit assumptions regarding average ticket prices, operating margins, marketing costs, expected revenue lift, and retention horizons. The results should therefore be interpreted as scenario-based estimates rather than precise financial forecasts. Nevertheless, the framework provides a consistent and decision-oriented basis for comparing segments and prioritizing marketing investments.

Appendix D. Interactive Cluster Visualization Dashboard

An interactive cluster visualization dashboard was developed; the source code is available in the corresponding folder of our GitHub repository, and a video demonstration can be accessed at the following link: https://www.youtube.com/watch?v=FDur9wqpiYw.