

DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

AI/ML Customer Segmentation: Exploratory Data Analysis

Group 23

Bruna Sousa, 20250526

Rui Sousa, 20250473

Alexandre Coelho, 20250475

Fall Semester 2025-2026

TABLE OF CONTENTS

1. Abstract.....	1
2. Introduction	1
2.1. Business context and segmentation challenge	1
2.2. Analytical objectives	1
3. Data analysis	1
3.1. Dataset description.....	1
3.2. Data quality assessment	2
3.2.1. Missing values.....	2
3.2.2. Duplicates	2
3.2.3. Coherence.....	2
3.3. Univariate and bivariate exploratory analysis	2
3.3.1. Categorical distributions	3
3.3.2. Numerical distributions.....	3
3.3.3. Behavioral trends	3
4. Results.....	5
5. Conclusion.....	5
Bibliographical References	6
Appendix A	7
Appendix B	8
Annexes.....	9
Annex: AI Usage Statement.....	9
Annex: Contribution Statement	9
Annex: Responsibility Statement	9

1. ABSTRACT

This report presents the development of a data-driven customer segmentation for Amazing International Airlines Inc. (AIAI), based on loyalty program data. It details the business challenge, data analysis process, and key findings from exploratory and behavioral analyses. After addressing data quality issues such as missing values and duplicates, the study identifies important patterns in customer behavior, including travel frequency, loyalty status, and lifetime value. The report concludes with a segmentation model that distinguishes active from inactive customers and defines three main behavioral profiles, providing insights into how different customer types interact with the airline.

2. INTRODUCTION

2.1. Business context and segmentation challenge

In a world increasingly driven by data, Amazing International Airlines Inc. (AIAI) seeks to enhance customer satisfaction, extend lifetime value, and maximize profitability through deeper insights into consumer behavior. This report addresses the challenge of identifying behavioral patterns and grouping customers based on travel habits, preferences, and loyalty characteristics to enable personalized marketing, targeted services, and optimized loyalty programs. Originating from a data-centric analytical approach, the project aims to translate findings into actionable strategies that improve customer experience and strengthen AIAI's position in the global airline industry.

2.2. Analytical objectives

In alignment with AIAI's strategic goal of leveraging data to enhance customer understanding, this study defines a set of analytical objectives that guide the segmentation process. These objectives translate the business challenge into measurable analytical tasks, ensuring that the resulting insights can be directly applied to marketing, loyalty, and service optimization decisions. Specifically, the analysis aims: to identify distinct customer segments within the loyalty program based on demographic, behavioral, and value-based variables; analyze travel behavior patterns such as flight frequency and distance, to understand how these behaviors relate to customer value; develop a data-driven segmentation model that supports targeted marketing and personalized service design; provide actionable business insights that help AIAI optimize loyalty rewards, pricing strategies, and communication channels for each segment.

3. DATA ANALYSIS

3.1. Dataset description

This analysis utilizes two core datasets provided by the airline, detailing customer profiles and their corresponding flight activities. The first, Customer Dataset, provides a static, demographic, and value-based snapshot of the Loyalty program members. It contains 16,921 records and 20 columns. These columns include one unique identifier (*Loyalty#*) and the others break down into 13 categorical features, 4 numerical features and 2 datetime features. The second dataset, Flights Dataset, is a transactional log of all flights taken by the members. It is a much larger dataset, containing 608,436

records and 10 variables. These variables are primarily quantitative, consisting of 8 numerical features and 1 datetime feature, detailing all flight and loyalty point activities from 2019 to 2021. Both datasets are linked by the *Loyalty#* column, which serves as the common customer identifier.

3.2. Data quality assessment

An in-depth data quality assessment was conducted to identify limitations, anomalies, and inconsistencies that could impact the reliability of the customer segmentation. The analysis focused on missing values, duplicates, and logical/temporal coherence.

3.2.1. Missing values

The analysis of missing values revealed two distinct scenarios. First, a negligible number of records (20) were missing identical values for both income and customer lifetime value. Given their minimal impact, the decision was made to drop these records to ensure the integrity of the modeling phase without the bias of imputation. More significantly, the *CancellationDate* column was missing for 86.45% of all customers. This widespread absence is not treated as an error but as a core characteristic of the dataset, indicating that these customers have not officially left the program.

3.2.2. Duplicates

The duplicate analysis began with the Customer dataset, where a critical data integrity failure was discovered: the *Loyalty#* was not unique. Further inspection revealed that these duplicates were errors where distinct customers (e.g.: individuals with different names) had been assigned the same *Loyalty#*. To resolve this data conflict, the only methodologically correct choice was to eliminate all data contaminated by these compromised *Loyalty#* IDs.

3.2.3. Coherence

The coherence analysis tested business logic and temporal integrity between the datasets. It was confirmed that 100% of customers who left the program (and did not re-enroll) still maintain a *Loyalty* status value. This showed that *LoyaltyStatus* reflects historical tier achievement rather than current membership activity, making it unsuitable to indicate active or inactive status. The number of customers that had his *CancellationDate* before the *EnrollmentDateOpening* corresponds to 8.86% of the customers that left the program and strongly suggest a business scenario where members re-enrolled in the loyalty program, highlighting a valuable segment of returned customers for further attention in the segmentation analysis. Another inconsistency found was the customers flagged with the 2021 promotion who had enrolled in previous years. This suggests the *EnrollmentType* variable cannot be strictly relied upon for calculating customer longevity. Additional tests confirmed high internal data integrity, showing no inconsistencies between flight counts and companions, distance travel and points accumulated, or the dollar value of redeem points. Checks also verified that customer names were correctly recorded.

3.3. Univariate and bivariate exploratory analysis

The exploratory analysis aimed to understand the main characteristics of the dataset, identify potential patterns, and detect anomalies or irregularities that could influence the modelling phase. Both univariate and bivariate analyses were performed, covering categorical, numerical, and behavioral

variables. Various visualization techniques were applied, such as count plots, histograms, boxplots, scatter plots, line plots and a heat map to better interpret the data distributions, relationships and extract relevant insights.

3.3.1. Categorical distributions

The categorical analysis was conducted exclusively on the customer dataset. From the visualizations, it was observed that all customers are from Canada, indicating that *Country* offers no variability. Similarly, there is little variation in *Gender* or *Location Type*, suggesting that these features may have limited discriminative power for segmentation purposes.

Regarding education, most loyalty members hold a Bachelor's degree, followed by those with College level education. In terms of *marital status*, most customers are married, which may indicate a stable and mature demographic segment. The *Loyalty Status* distribution reveals that most clients belong to the *Star* tier, followed by *Nova* and *Aurora*.

3.3.2. Numerical distributions

The numerical analysis began with the variables from the customer dataset. There are more enrolments on the loyalty program than cancellations, which indicates good customer retention. However, as shown in figure 1, cancellations have increased steadily since 2019, while enrolments remained relatively stable until 2021, when a sharp rise occurred, likely due to a major promotional campaign that attracted new members.

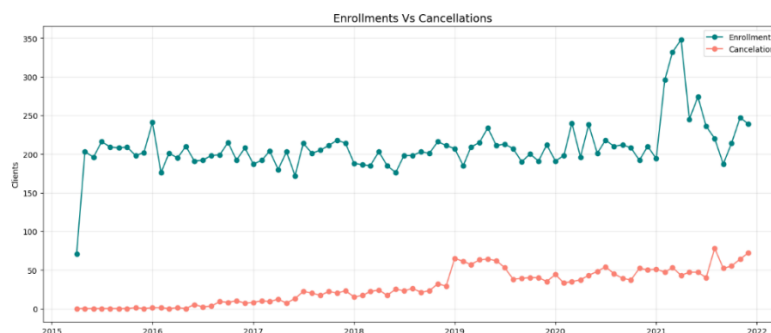


Figure 1 – Customer enrollments on loyalty program and cancellations throughout the years

Regarding *Income* and *Customer Lifetime Value*, both variables display a wide range of values. The presence of zero-income entries may correspond to unemployed individuals or students. Since the flight dataset contains monthly records, many values for *NumFlights* are zero, as customers do not fly every month.

3.3.3. Behavioral trends

The behavioral analysis compared categorical and numerical variables from the customer dataset to uncover meaningful relationships and trends. A geovisualization analysis was also conducted to explore the distribution across Canada's states and provinces (Appendix A) and an interactive EDA Dashboard (Appendix B). The median Income is slightly lower for the *Star* loyalty group compared to *Aurora* and *Nova*, but the overall distributions are consistent across all three tiers (Figures 2 and 3). This suggests that *Loyalty Status* is not a strong differentiator of customer income levels. Similarly, the

distribution of *Customer Lifetime Value* shows that high-value customers (outliers) appear across all loyalty statuses. Although *Aurora* members have a slightly higher median *CLV*, the pattern indicates that loyalty tier is a weak predictor of lifetime value. This suggests that the current segmentation strategy may not effectively differentiate customers.

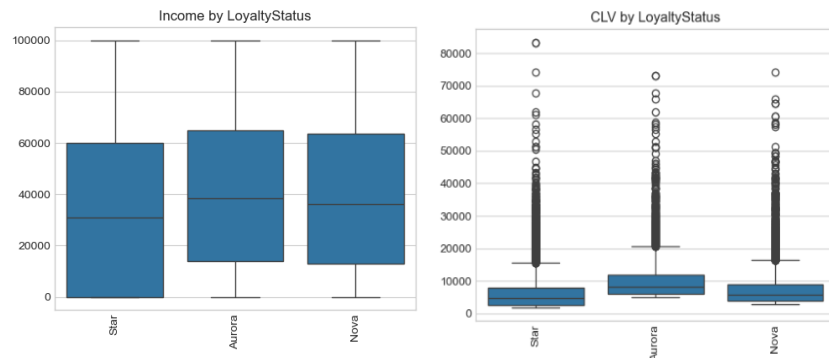


Figure 2 – Customer Income per Loyalty Status Figure 3 – Customer Lifetime Value per Loyalty Status

Additionally, flight frequency exhibits clear seasonal patterns. As expected, customers travel more frequently during holiday periods such as Christmas, Easter, and summer, with a major peak observed during the summer months.



Figure 4 – Monthly variation in flights per year

Finally, a positive relationship between number of flights, distance and points accumulated (0.61), suggests that frequent flyers tend to cover longer routes and earn more points. However, the correlation between points accumulated and points redeemed is very low (0.19). This suggests that customers are accumulating points but not using them. This behavior may indicate that the reward structure is unattractive or that point expiry policies limit customer engagement.

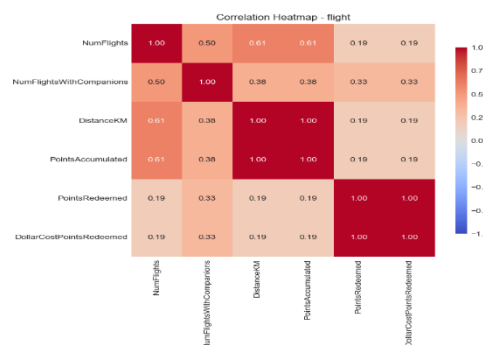


Figure 5 – Correlation between flights database variables

4. RESULTS

Based on the exploratory data analysis, several actions are proposed to prepare the dataset for clustering and ensure meaningful customer segmentation. The *LoyaltyStatus* variable will be retained, but it should not be interpreted as an indicator of current activity, as it reflects a customer's historical achievements rather than their active or inactive status. To address this limitation, a new Boolean variable will be created to indicate each customer's current membership status, allowing the segmentation analysis to clearly distinguish between active and inactive clients.

Similarly, the *EnrollmentType* variable will be treated cautiously and used only as an exploratory attribute, since inconsistencies were identified in the historical enrollment data that prevent it from serving as a reliable proxy for customer longevity. Continuous variables such as total spend, miles traveled, and points accumulated will undergo appropriate transformations and normalization to handle asymmetrical distributions and differences in scale, ensuring comparability and more stable clustering results.

The exploratory analysis also revealed that demographic attributes such as gender, country, and location type show little variability and contribute limited explanatory value. In contrast, behavioral variables like *NumFlights*, *DistanceKM*, *PointsAccumulated*, and *Customer Lifetime Value* displayed greater variation and stronger associations, indicating their relevance for capturing different levels of engagement and travel intensity.

Overall, these steps will refine the dataset by emphasizing the most informative features, mitigating noise from inconsistent or low-variance attributes, and ultimately supporting the identification of robust and actionable customer segments for strategic decision-making.

5. CONCLUSION

The proposed approach adopts a two-phase segmentation strategy to ensure both clarity and analytical relevance. The macro-segmentation divides customers into "Active" and "Inactive" groups using the newly engineered *IsActive* variable, allowing the model to focus on customers who maintain a current relationship with the brand. The subsequent micro-segmentation, applied only to active customers, employs the K-Means algorithm on key behavioral variables such as *NumFlights*, *DistanceKM*, *PointsAccumulated*, and *CLV* selected for their strong ability to capture customer value and engagement.

The preprocessing pipeline includes skewness correction (via logarithmic or Box-Cox transformations) and normalization (using *StandardScaler*), ensuring stability and fairness in distance-based computations. Demographic and inconsistent variables, such as *EnrollmentType*, were excluded to minimize noise and bias.

We expect to identify three main active customer segments: the high-value frequent flyers, occasional and seasonal travelers and the inactive or disengaged Customers, where we are looking for re-engagement strategies. Overall, this segmentation will enable more optimized resource allocation and targeted marketing actions with data-driven understanding of customers, to improve the current behavior and value.

BIBLIOGRAPHICAL REFERENCES

Software and AI Tools

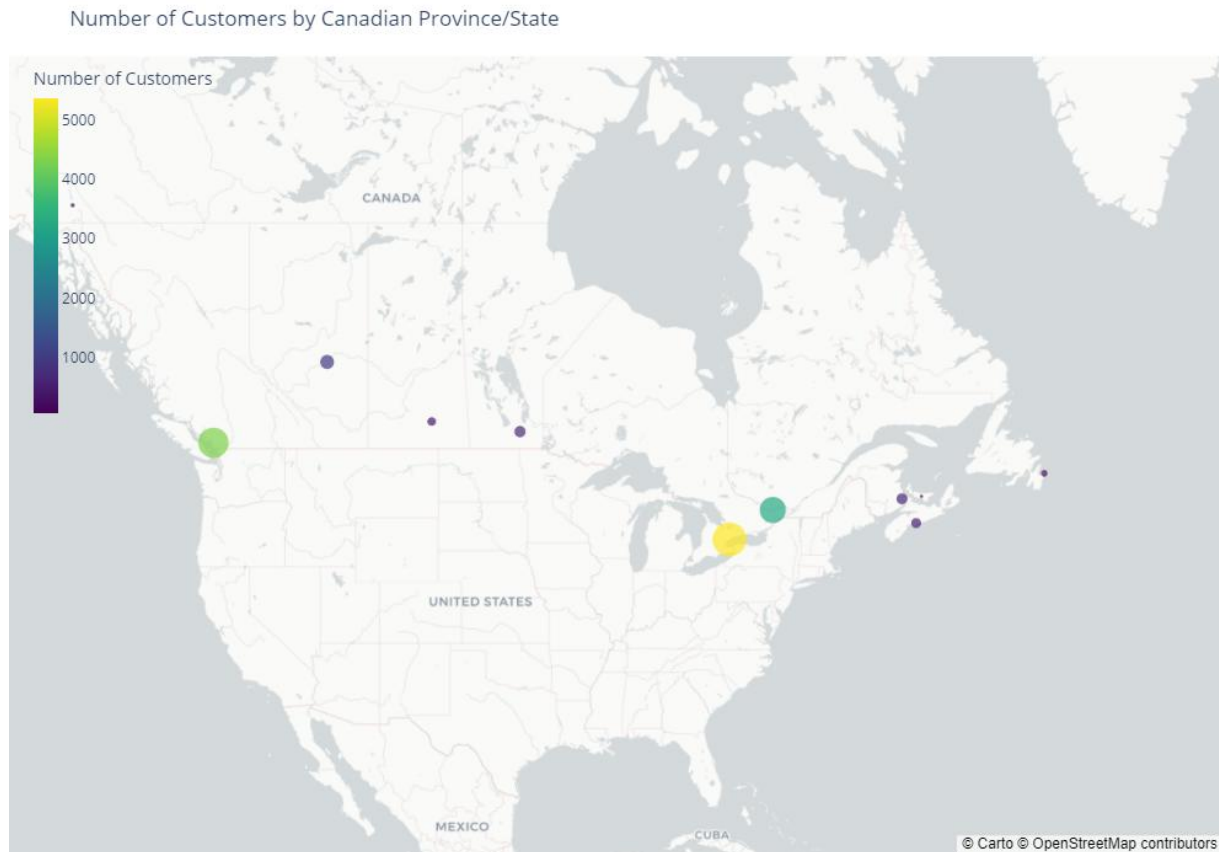
- OpenAI. (2025). *ChatGPT* [Large language model]. OpenAI. <https://chat.openai.com/>
- Google DeepMind. (2025). *Gemini* [Large language model]. Google. <https://gemini.google.com/>
- Canva. (2025). *Canva* [Computer software]. <https://www.canva.com/>
- Power BI. (2025). *Power BI* [Data visualization software]. Microsoft. <https://powerbi.microsoft.com/>
- Envato. (2025). *Envato Elements* [Design resources platform]. <https://elements.envato.com/>

Libraries and Documentation

- GeoPandas Developers. (2025). *GeoPandas: Python tools for geographic data* [Computer software]. <https://geopandas.org/>
- Plotly Technologies Inc. (2025). *Plotly Express: Interactive data visualization library for Python* [Computer software]. <https://plotly.com/python/tile-scatter-maps/>
- Python Software Foundation. (2025). *Python: Official documentation* [Computer software]. <https://docs.python.org/3/>
- Matplotlib Developers. (2025). *Matplotlib: Visualization with Python* [Computer software]. <https://matplotlib.org/>

APPENDIX A

GEOSPATIAL ANALYSIS FINDINGS

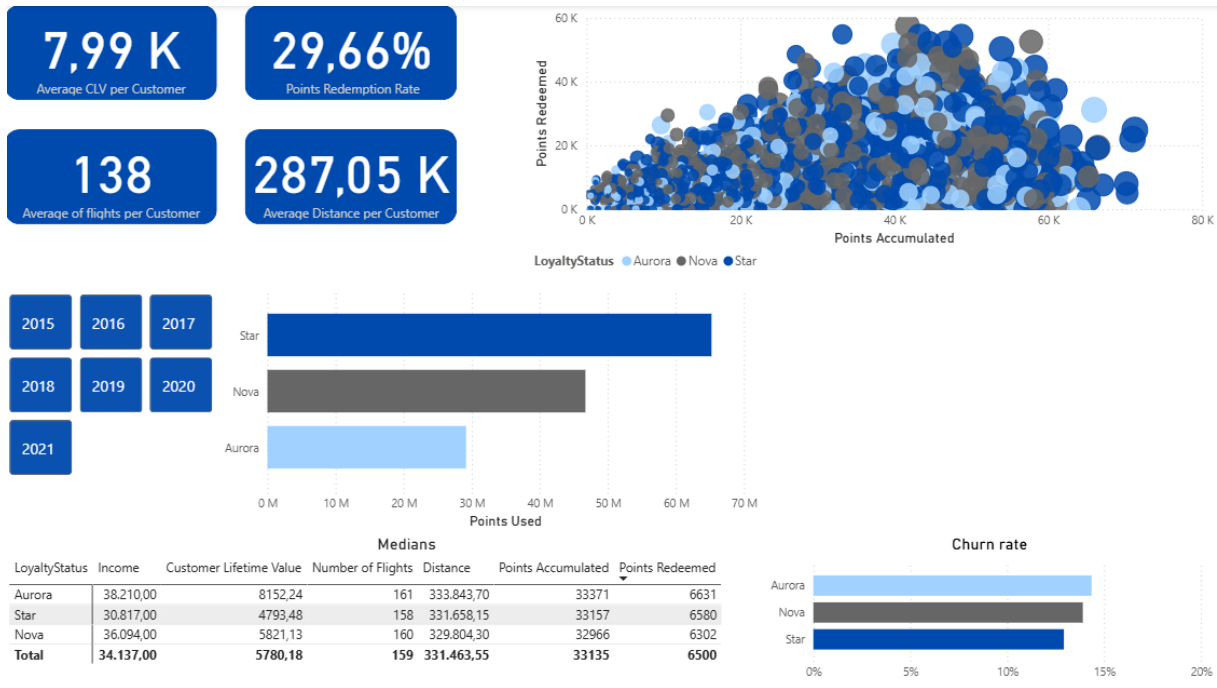


The geospatial analysis reveals that the airline's customer base is heavily concentrated in Canada's most populous provinces: Ontario (ON), Quebec (QC), and British Columbia (BC). This visualization confirms where customers reside, but a critical data limitation prevents further route analysis. Specifically, the flights database lacks origin and destination airport data for individual journeys, making it impossible to map flight routes or identify specific operational hubs.

An unexpected insight emerged when comparing customer behavior across regions. The hypothesis was that provinces with major international airports (like Ontario) would exhibit higher average travel frequency. However, the data shows that the mean *NumFlights* and *DistanceKM* per customer are surprisingly consistent across all provinces, suggesting that the average traveler profile is homogeneous nationwide. Furthermore, the *LoyaltyStatus* mode is "Star" in every province, reinforcing the finding that this attribute has low variability and limited value for creating distinct customer segments.

APPENDIX B

POWER BI DASHBOARD



The Power BI dashboard provides a critical layer of dynamic validation and interactive exploration that this static report cannot. It translates the key findings from the exploratory data analysis, such as the trends in member enrollments versus cancellations and the seasonal flight patterns into a live, filterable environment. This allows stakeholders to move beyond static charts and directly engage with the data, verifying the conclusions drawn from the code analysis and providing a tangible tool for business users.

The dashboard's primary contribution is its ability to visually and interactively prove the report's most critical hypotheses. While the report states that *Loyalty Status* is a weak predictor of customer value, the dashboard demonstrates this by allowing a direct, visual comparison of median *CLV* and *Income* across tiers. Furthermore, it provides the tools for deeper, cross-filtered analysis that a static document cannot, reinforcing the conclusions about customer behavior and ultimately validating the entire segmentation approach.

To fully interact with the dashboard and view all analyses, please open the Power BI file (Group23_EDA_Bonus5_Interactive_EDA_Dashboard.pbix) located within the submitted OPTIONAL Bonus on moodle.

ANNEXES

ANNEX: AI USAGE STATEMENT

AI tools were used exclusively for writing refinement and language accuracy purposes. Specifically, ChatGPT and Gemini were employed to improve sentence clarity, adjust tone to an academic/university standard, and translate or rephrase text into English when necessary. All content ideas, analytical insights, and business interpretations were fully developed by the group members, independently of AI assistance.

Most of the code implementation was based on and adapted from the notebooks provided during class sessions, with necessary modifications and extensions made by the group to meet the project requirements. AI usage was strictly limited to linguistic and stylistic enhancement only.

ANNEX: CONTRIBUTION STATEMENT

Student Alexandre Coelho (Student ID: 20250475)

- Bonus: interactive EDA dashboard.
- Report: data analysis, results.
- Notebook: business understanding, data understanding customer database, data preparation, data exploration.

Student Bruna Sousa (Student ID: 20250526)

- Poster.
- Report: data analysis, results, report structure.
- Notebook: business understanding, coherence checks, notebook structure, data preparation.

Student Rui Ferreira (Student ID: 20250473)

- Bonus: geo-spatial insights.
- Report: abstract, introduction, conclusion, bibliographical references, appendix, annexes.
- Notebook: business understanding, data understanding flights database, data preparation, data exploration (flights), feature engineering.

All members contributed equally to discussions, ideation, and validation of results throughout the project.

ANNEX: 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 in the AI Usage Statement, and the project code was primarily developed based on the notebooks and examples provided in class, all insights, conclusions, and recommendations are the result of our independent analysis and critical thinking. We take full responsibility for the accuracy, integrity, and quality of this submission.