NOVA IMS
INFORMATION MANAGEMENT SCHOOL

AMAZING INTERNATIONAL AIRLINES INC.

DATA MINING – DELIVERABLE 1

Master in **Data Science and Advanced Analytics**
Instructors: Fernando Bação, Farina Pontejos, Gaspar Pereira, Ana Caleiro
2025/2026
GROUP 27

GONÇALO RILHAS, 20250490
GUILHERME SANTOS, 20250510
LUANA PINTO, 20250481
MARTA PEDRO, 20250500

TABLE OF CONTENTS

1. ABSTRACT

This report presents an exploratory data analysis conducted for Amazing International Airlines Inc. (AIAI), aiming to support the development of a data-driven customer segmentation strategy. Following the CRISP-DM methodology, the project's main goal is to find and understand different customer groups within AIAI's loyalty program. A precise segmentation will allow AIAI to make well informed, customer-focused choices, leading to personalized service, better rewards, and focused marketing. Descriptive statistics and visualizations were used to find distributions, anomalies, and links across demographic, behavioral, and value-based characteristics. Several features were created to capture more complete dimensions of customer behavior, preparing the data for clustering. The analysis shows main points about customer diversity, loyalty status, and travel activity, which are expected to guide segmentation design in later project parts. This report creates a firm base for making understandable and usable customer clusters in the next stage of AIAI's marketing strategy work.

2. INTRODUCTION

Amazing International Airlines Inc. (AIAI) works in a very competitive and fast-changing airline market, where customer loyalty and personalized service are key. As customer desires keep growing, AIAI needs to create marketing strategies that meet different traveler preferences, spending habits, and service priorities. Acknowledging these differences through data-driven segmentation is important for improving satisfaction, retention, and income.

This project attempts to find customer patterns and routines within the company's loyalty program. The analysis focuses on which characteristics, demographic, behavioral, or value-based, are most helpful in splitting customer groups. Following the CRISP-DM methodology, this report covers the Business Understanding and Data Understanding phases, setting a basis for later modeling and segmentation.

This study is meant to give AIAI helpful information about its customer database, allowing it to create targeted strategies that match service offerings with customer needs and long-term loyalty goals.

3. DATA ANALYSIS

This section presents the exploratory analysis carried out in the Data Understanding phase of the CRISP-DM methodology. The objective was to evaluate the quality and structure of AIAI's datasets, understand the main data distributions, and identify relationships that could later inform customer segmentation modeling.

3.1. Data set overview

The analysis was based on two core datasets. The first, *DM_AIAI_CustomerDB.csv*, includes 16,921 customers described by 20 variables encompassing demographic, loyalty, and value-based attributes. The second, *DM_AIAI_FlightsDB.csv*, contains 608,436 monthly records documenting flight activity between 2022 and 2024. Both datasets were linked via the unique customer identifier *Loyalty#*, enabling the integration of customer information with flight data.

3.2. Data quality and structural validation

A structural audit was performed to secure consistency and reliability across both datasets.

In the customer database, the variable *CancellationDate* displayed 86.3% missing values. This was confirmed to represent active customers, not data errors, leading to the creation of a new binary variable, *Is_Active*. The resulting distribution, shown in the Annex 4 (Figure 1), illustrates this 86.3% (Active) vs 13.7% (Inactive) split. The variables *Income* and *Customer Lifetime Value* (CLV) exhibited minimal missingness (0.12%) and were removed before modeling.

Regarding duplicates, 164 repeated *Loyalty#* identifiers (1.93% of customers) were found and removed to make sure each record represented a single customer. In the flights' dataset, 2,903 duplicated rows were identified but kept, as they represented inactivity and had behavioral meaning. It is also important to check that variables are in the correct data type, which led to changes in some variables.

3.3. Univariate and categorical exploration

Exploration of numerical variables revealed meaningful anomalies and asymmetries. Most notably, 25.37% of customers reported an income of zero (Figure 2, Annex 4). This was interpreted as a valid value, possibly showing students or customers who didn't want to share their income or don´t work yet. Moreover, the variables *Income*, *CLV*, *Total_Flights*, and *DistanceKM* all presented strong right-skewed distributions, suggesting that a small number of "super-travelers", people who travel a lot, or high-value customers that contribute disproportionately to total business value (Figure 2, Annex 4). This skewness in CLV is also reflected in the "Frequency of Customer Value Tiers" graph (Figure 3 on Annex 4), which shows the "Medium" tier as the most common.

An analysis of the raw flights database also revealed strong seasonal patterns, with travel peaking, as expected, in the summer months (July) and again in December, as seen in the "Total Flights by Month and Season" chart (see Figure 4, Annex 4).

Categorical distributions provided more context on customer composition. The "Star" is the largest, followed by Nova and Aurora, reflecting a pyramid structure (Figure 5, Annex 4). The mean fidelity age at cancellation is similar across all groups, with most cancellations 1 year after joining the fidelity program ("Cancelation mean time per LoyaltyStatus", see Figure 6, Annex 4). To mitigate these early cancellations, some measures can be implemented according to the loyalty status of the customer:

Star - Offer exclusive experiences and early access to promotions to maintain engagement and reward long-term loyalty

Nova - Introduce tier-up incentives and personalized offers to encourage progression to the Star level

Aurora - Provide welcome bonuses, tailored discounts, and improved communication to build stronger brand attachment and reduce early cancellations.

Educationally, most members held at least a bachelor's degree, while "Standard" remained the main enrollment channel (65%), with the "2021 promotion" also contributing substantially (Figure 5, Annex 4).

Geographically, the customer base is highly concentrated in Canada's major urban centers. An analysis of the customer distribution map (see Figure 9, Annex 4) provides a clear visual for the Province or State and Location Code variables. The map illustrates that most customers are clustered in a few important areas, notably a large cluster in Ontario (mostly Toronto and Quebec) and smaller, significant clusters in British Columbia (Vancouver) and near the New Scotland area. This concentration directly explains the multi-peaked distributions observed in the Latitude and Longitude histograms (Figure 2, Annex 4), confirming that the customer footprint mirrors Canada's main population hubs. An interactive version of this map is available at

[https://martap18.github.io/Data-Mining-Project/canada_customers_map_enhanced.html] for detailed study.

In terms of marital status, an equilibrium between single and married customers was observed. This attribute showed unexpected results when compared to the *Companion_Rate*, since most customers across all groups travel predominantly solo (Companion Rate < 0.3) and married customers showed just a slightly higher median companion rate compared to single or divorced customers (Figure 7, Annex 4).

3.4. Feature engineering and data aggregation

To enable customer-level segmentation, the flight activity data were aggregated by *Loyalty#,* creating a behavioral profile for each customer. In addition to *Is_Active* and *Total_Flights*, two main engineered features were introduced:

- *Fidelity_Age_Years*, measuring the customer's loyalty length in years.
- Behavioral ratios**,** including *Redemption_Behaviour* (redeemed vs. accumulated points) and *Companion_Rate* (flights with companions as a share of total flights).

3.5. Relationship analysis

The multivariate analysis, visualized in the "General pair plot with client and flight metrics" (Figure 8, Annex 4), revealed several patterns of interest. Strong positive correlations were observed among travel volume metrics, *Total_Flights, DistanceKM*, and *PointsAccumulated,* confirming that they capture the same behavioral dimension. CLV showed a moderate positive correlation with these indicators, suggesting that travel frequency contributes to lifetime value but does not fully determine it. On the other hand, Income displayed a weak correlation with behavioral variables, influenced by the "zero income" group, indicating that income may not be a reliable predictor of travel activity.

4. RESULTS

The exploratory findings from section 3 provide the justification for specific feature engineering, selection, and preprocessing decisions required for the clustering phase. This section connects the data insights to the analytical steps necessary for the Data Preparation phase of CRISP-DM.

4.1. Feature Engineering Choices

The analysis of data structure and quality directly guides the creation of key variables for clustering:

- The discovery that *CancellationDate* contained 86.3% missing values (Sec 3.2) was interpreted as representing active customers. This justified the engineering of the binary feature *Is_Active*, which will be critical for segmenting customers based on their engagement status (active vs. inactive).

- The raw *FlightsDB* data (Sec 3.1) required aggregation to the customer level. This justified the creation of new features (Sec 3.4) like *Fidelity_Age_Years, Redemption_Behaviour*, and *Companion_Rate*. These engineered features are essential for capturing customer fidelity, program engagement, and social travel habits, which are richer indicators than simple demographics.

4.2. Transformation and Scaling

The distributions in the univariate analysis (Sec 3.3) require transformations for clustering algorithms to perform well:

- The strong right-skewed distributions of *Income, CLV*, and *Total_Flights* (Sec 3.3) confirm the need for solid normalization or logarithmic transformation. This decision is essential to prevent distance-based algorithms (like K-Means) from being dominated by a small number of high-value customers.

- The "Income = 0" group (Sec 3.3), was identified as a valid behavioral subgroup, not an error. Therefore, this group will be preserved rather than excluded. This justifies treating 'Income' as a mixed variable or binning it, rather than simply scaling it numerically.

4.3. Feature Selection and Reduction

The relationship analysis (Section 3.5) offers a solid basis for selecting features that enhance both model efficiency and interpretability:

- The strong positive correlations found between *Total_Flights, DistanceKM,* and *PointsAccumulated* (Sec 3.5) indicate high redundancy. This finding justifies the elimination of redundant travel metrics or the use of dimensionality reduction algorithms, like PCA, to reduce collinearity.

- The weak correlation between Income and travel routine (Sec 3.5), combined with the finding that Marital Status doesn't predict companion travel (Sec 3.3), supports prioritizing behavioral

variables (*Companion_Rate, Redemption_Behaviour*) over demographics (*Income, Marital Status*) as drivers for segmentation

5. CONCLUSIONS

This exploratory analysis provided a strong support for the next Data Preparation and Clustering phases of the CRISP-DM framework. The findings clarified data structure, quality, and behavior, providing a foundation for effective segmentation.

Based on the results, the clustering process will focus on identifying customer groups that differ in behavioral intensity, loyalty engagement, and value contribution. Engineered variables such as *Is_Active*, *Fidelity_Age_Years, Redemption_Behaviour*, and *Companion_Rate* will play a central role in capturing engagement and travel dynamics, ensuring the resulting clusters are compartmentally meaningful.

Key clustering hypotheses include:

- high-frequency, high-value travelers;
- low-income, low-activity members (emerging customers);
- socially oriented or family travelers;
- inactive or churn-prone customers.

The preprocessing strategy will involve handling of missing and skewed data, normalization of continuous variables, encoding of categorical attributes, and feature reduction to minimize redundancy. Clustering algorithms such as K-Means or Hierarchical Clustering will be tested for performance and interpretability.

Overall, this analysis defines a clear roadmap for segmentation, enabling AIAI to derive interpretable, data-driven customer clusters that can guide targeted marketing and fidelity strategies in subsequent phases.

6. ANNEXES

Annex 1:

AI tools were used for coding syntax assistance (ChatGPT), literature review brainstorming and general formality corrections (Gemini), and report proofreading (Grammarly for grammar checks and Gemini). All analytical insights, business interpretations, and strategic recommendations represent original group analysis and thinking.

Annex 2:

Guilherme Santos (Student ID: 20250510)
- Fitness industry research for Business Understanding
- Introduction section and Descriptive Statistics
- Notebook organization

Gonçalo Rilhas (Student ID: 20250490)
- Poster Organization and Design

Luana Pinto (Student ID: 20250481)
- Data quality assessment (strange values, outliers)
- Data Visualization
- Notebook organization

Marta Pedro (Student ID: 20250500)
- Feature engineering
- Abstract, results interpretation sections on the notebook, report coordination

All members contributed to collaborative discussions and ideation.

Annex 3:

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.

Annex 4:

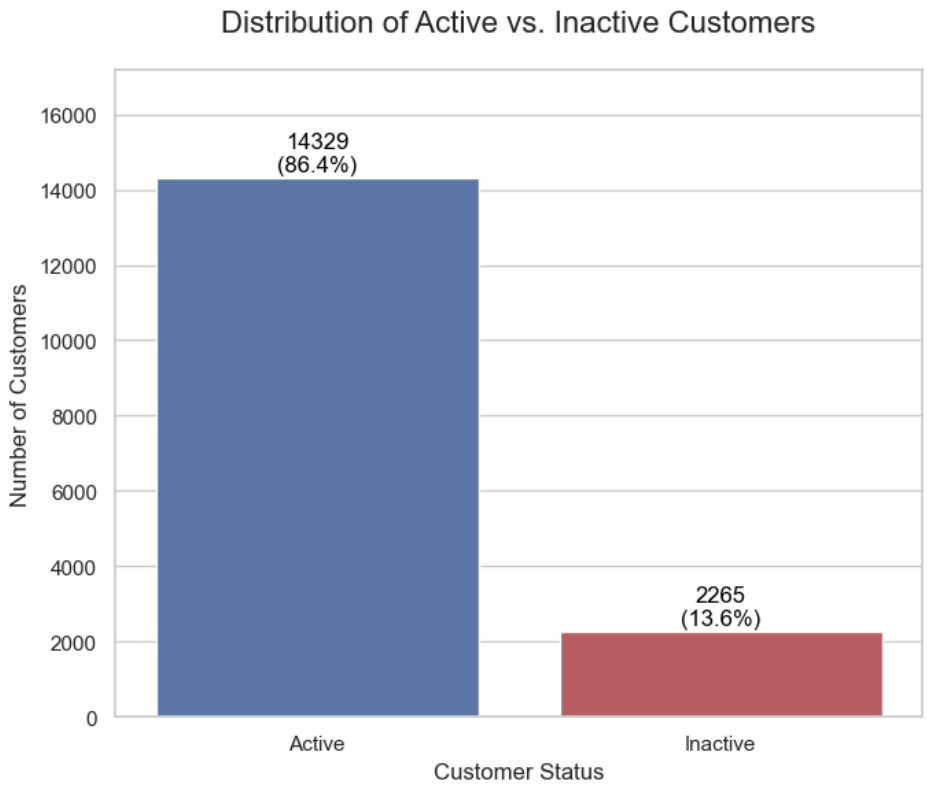

Figure 1 – Activity status distribution

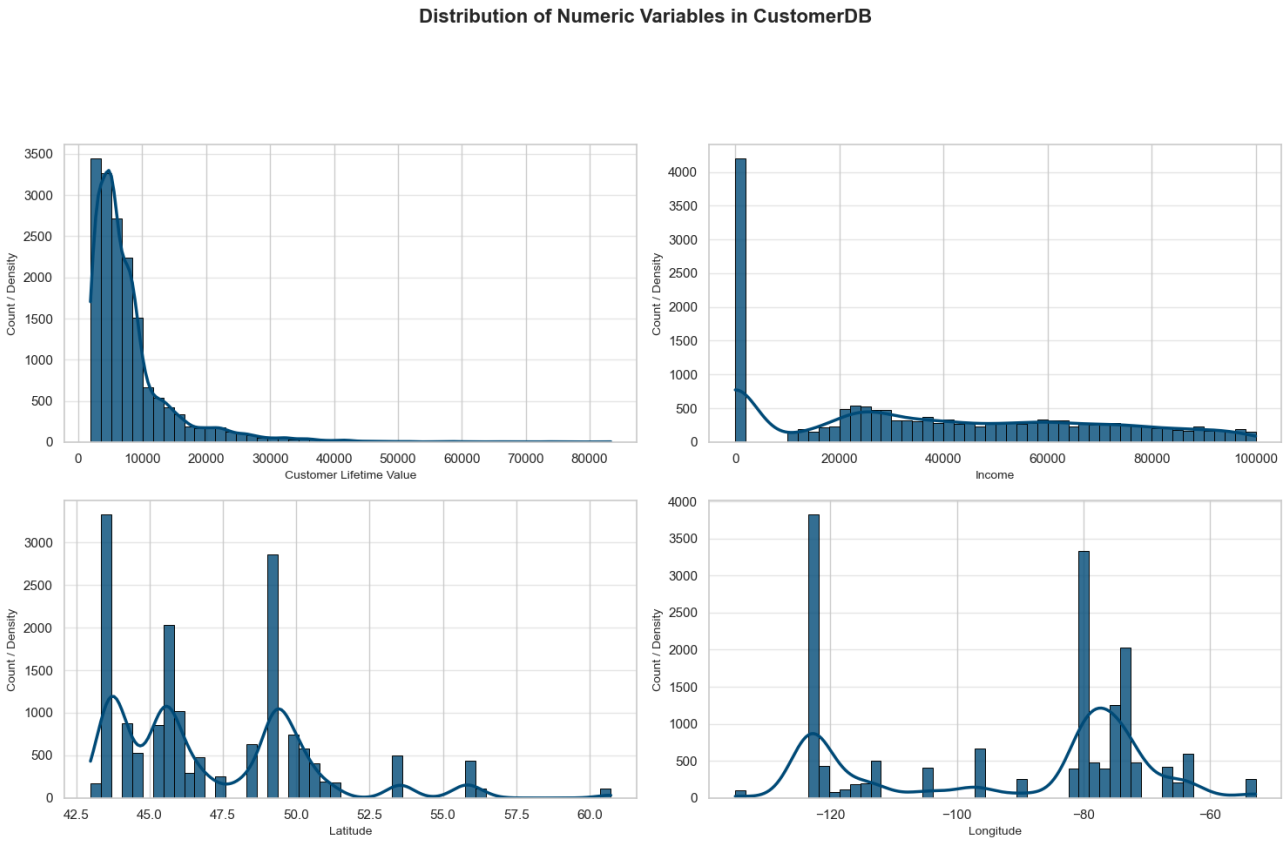

Figure 2 – Income, CLV, latitude and longitude distributions

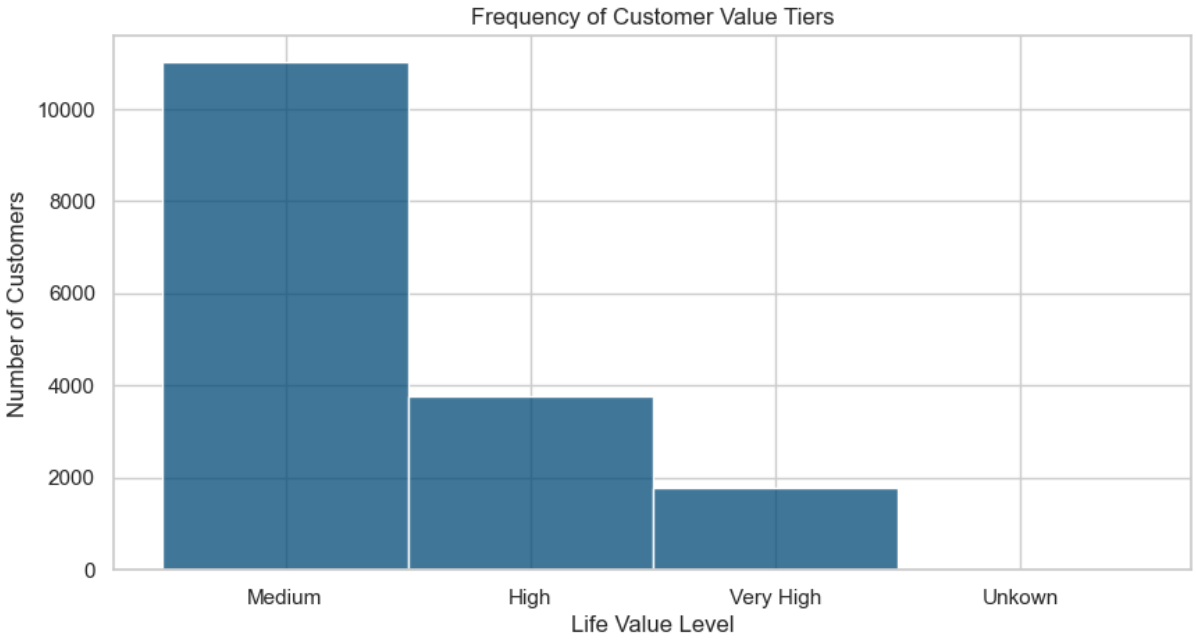

Figure 3 – Customer Value Tiers distributions

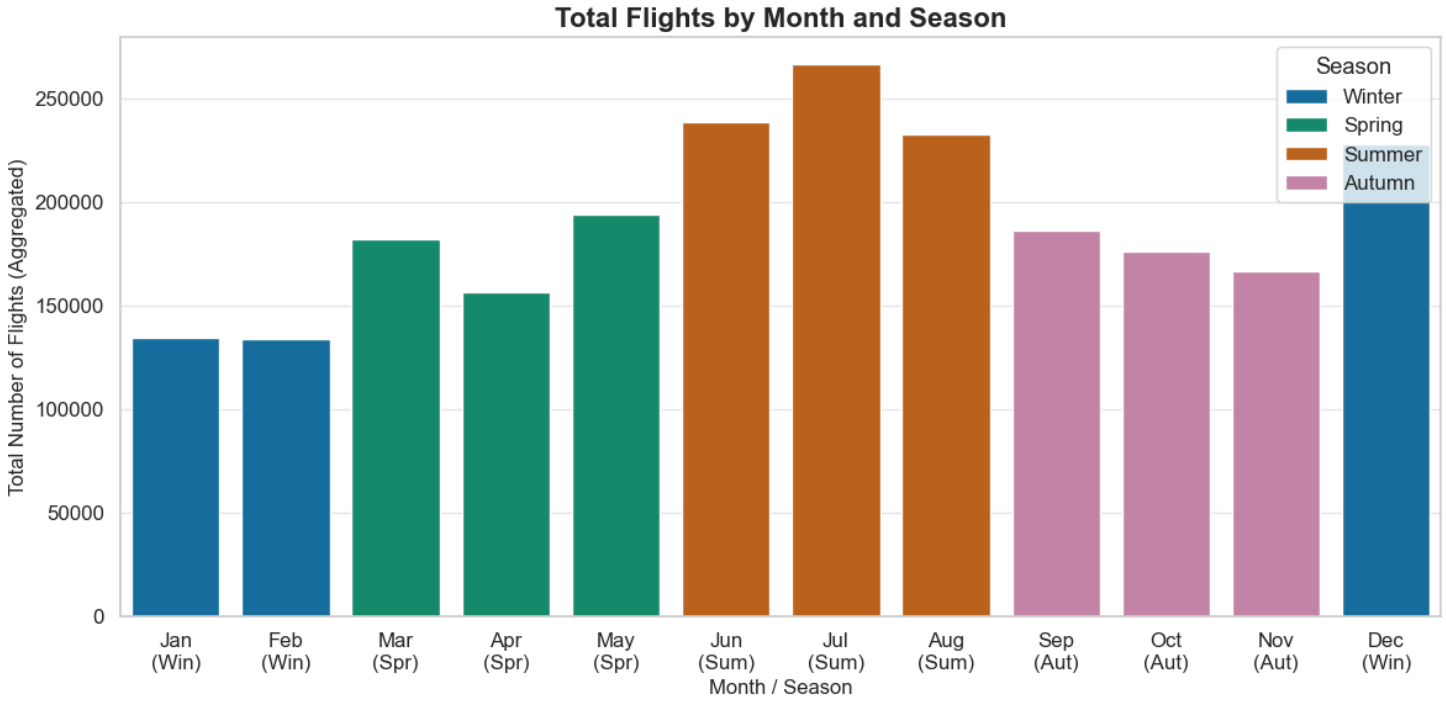

Figure 4 – Flights per month/season

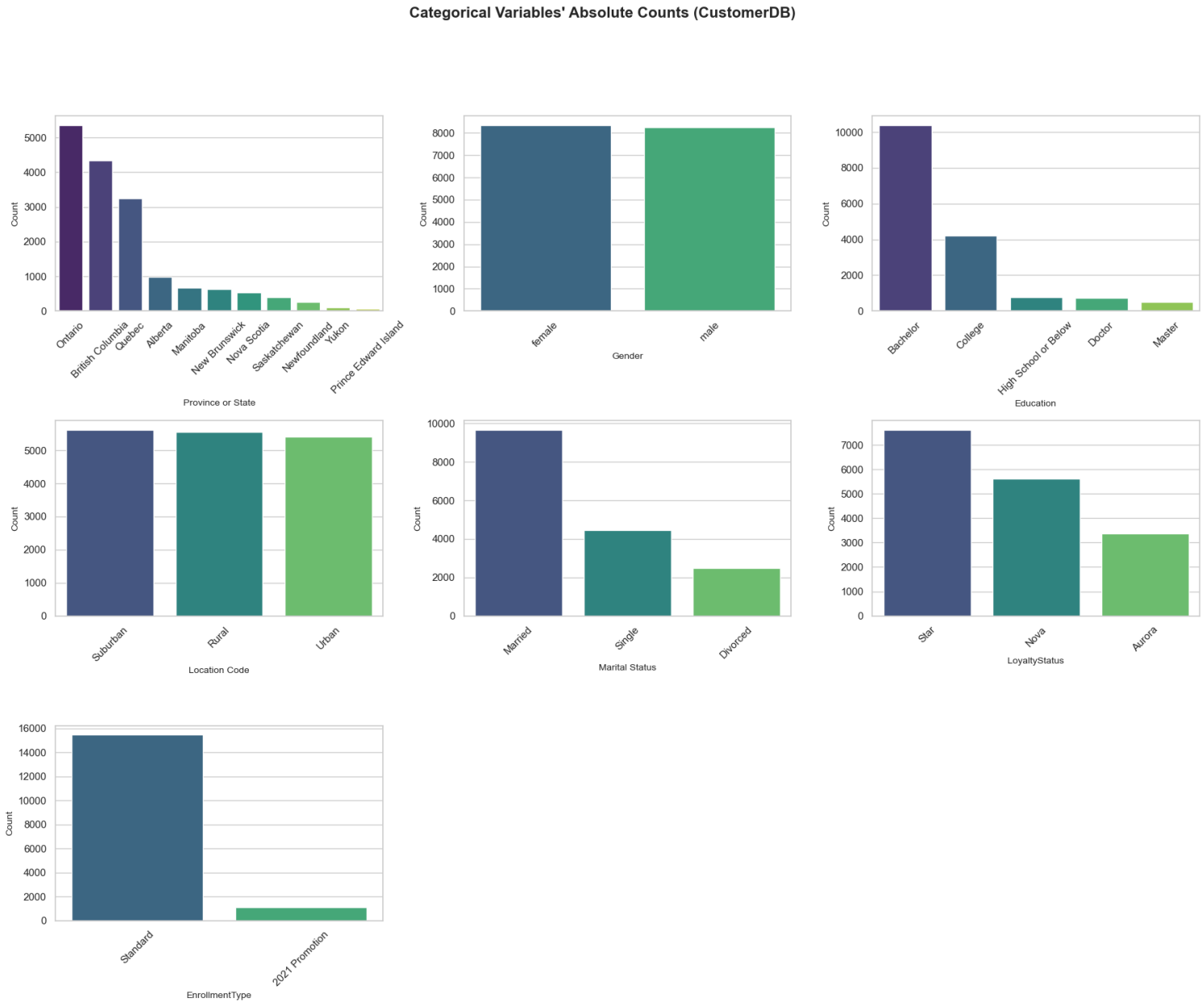

Figure 5 – Categorical variables distributions

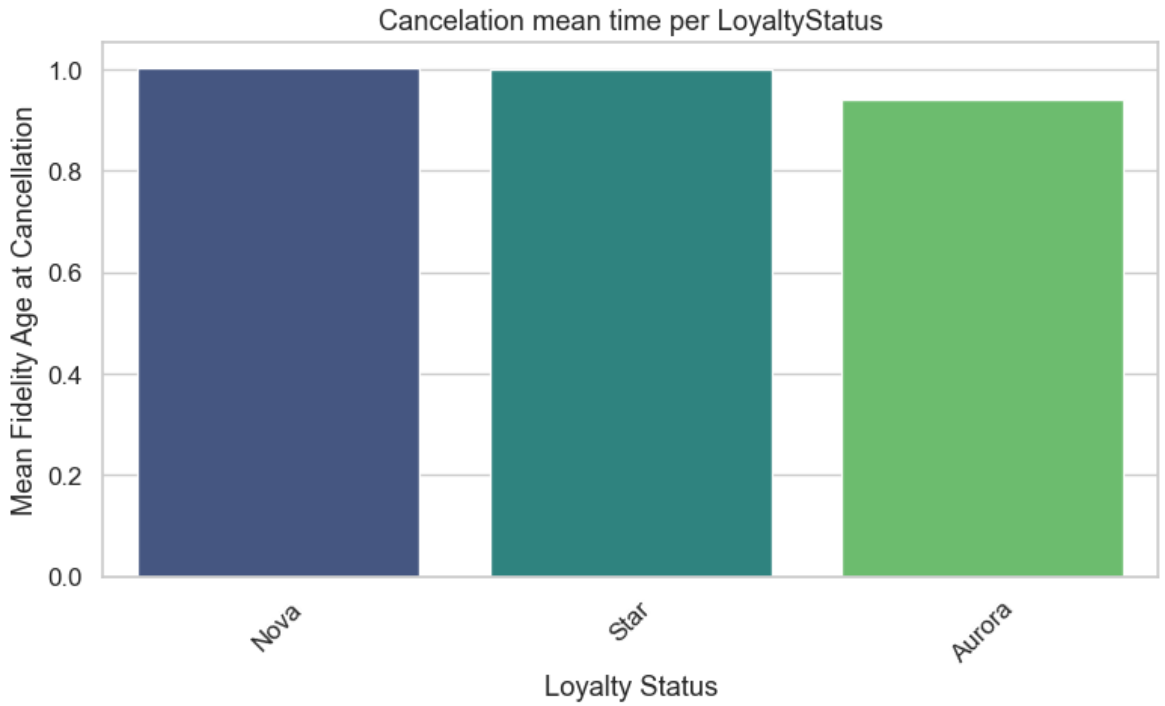

Figure 6 – Cancelation mean time per Loyalty status

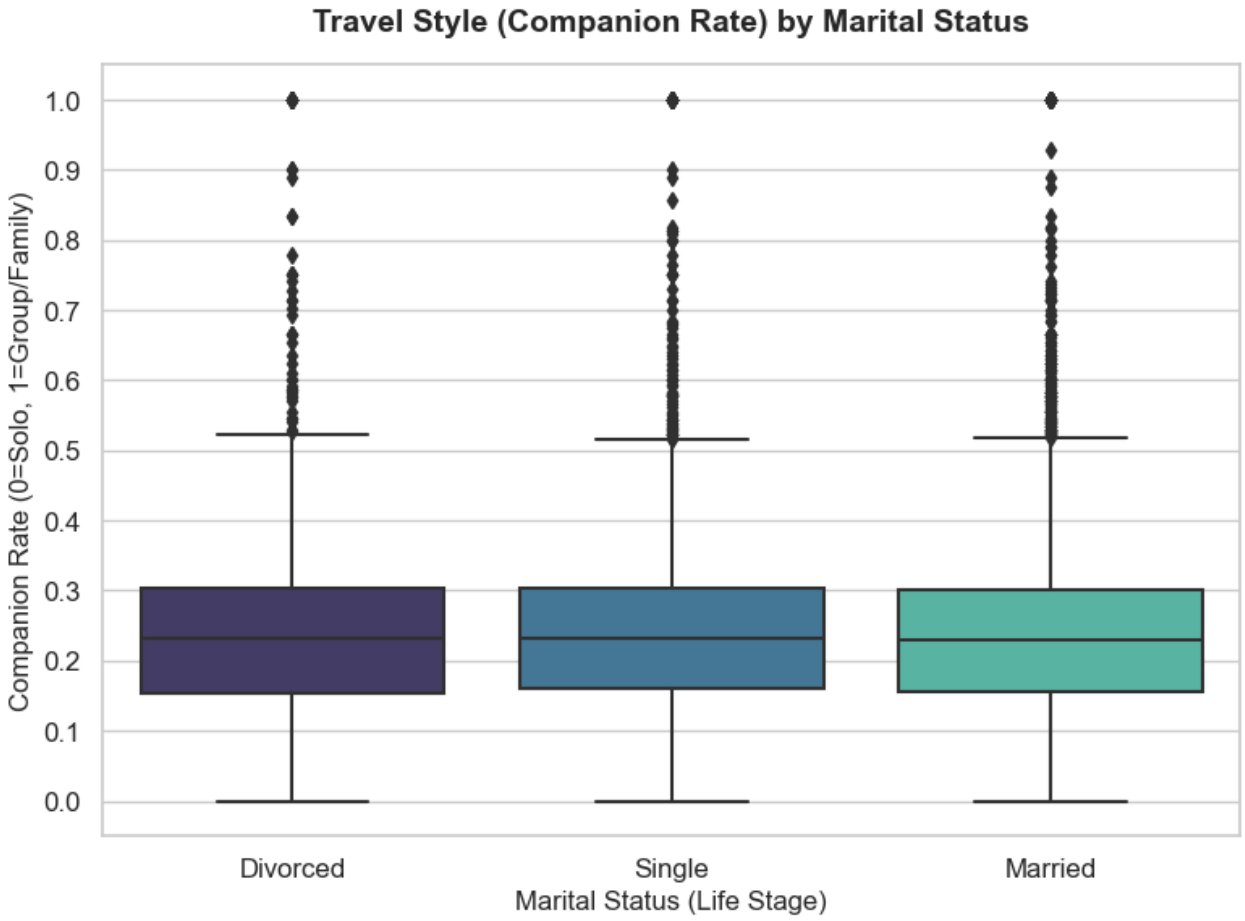

Figure 7 – Companion rate per marital status

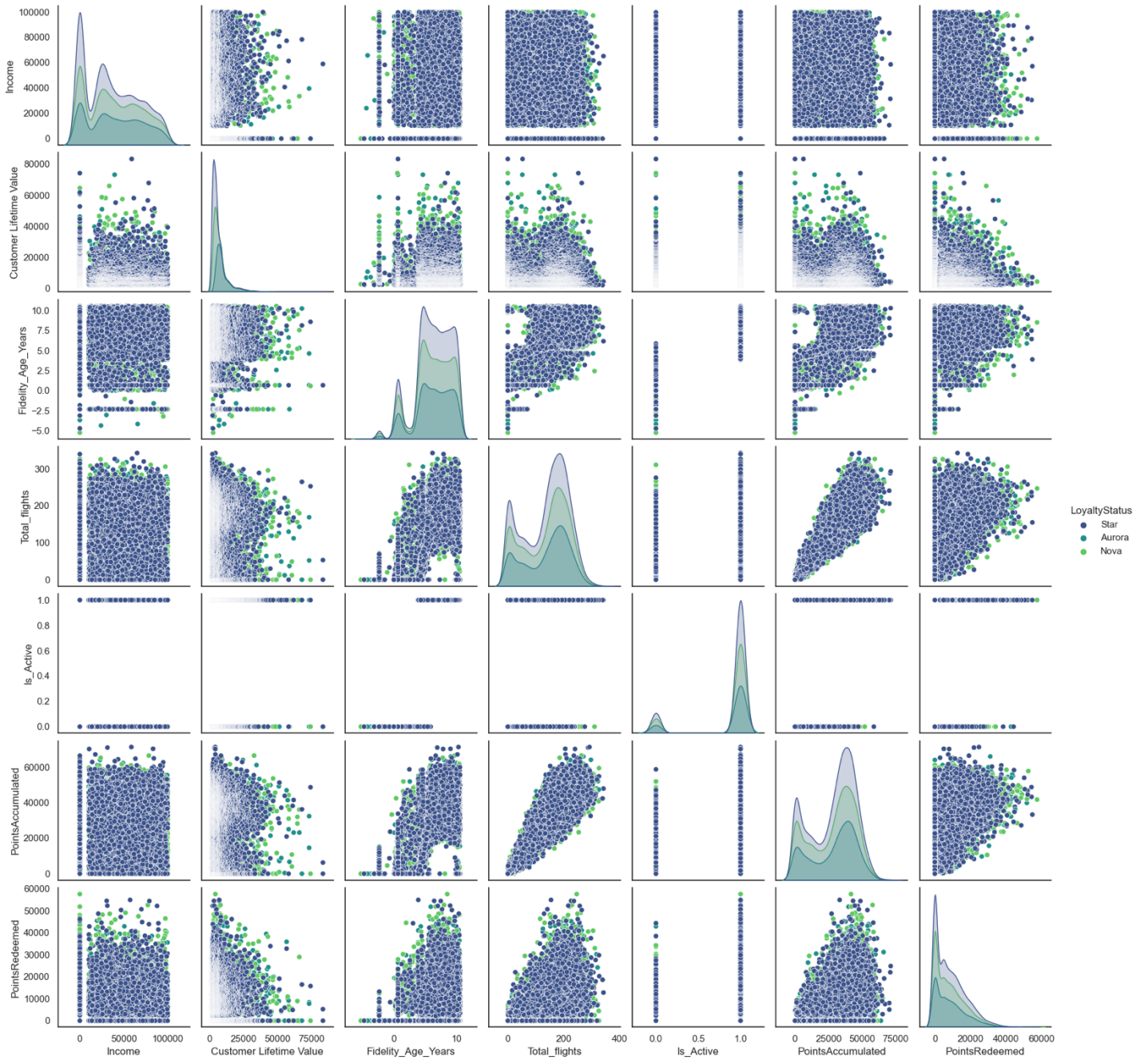

Figure 8 – General pair plot

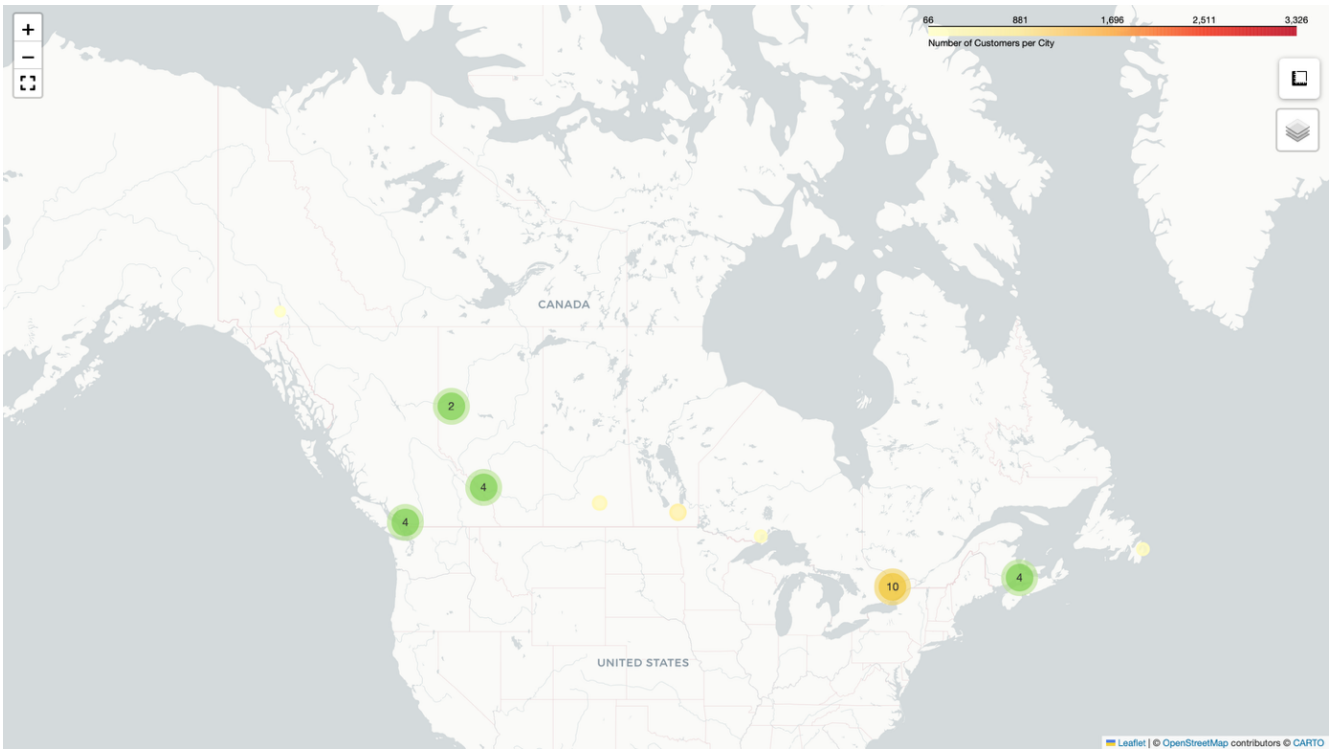

Figure 9 – Customer Geographic Distribution (**Note:** an interactive version of this map is available at https://martap18.github.io/Data-Mining-Project/canada_customers_map_enhanced.html)