Data Mining Project

Master in Data Science and Advanced Analytics

NOVA Information Management School
Universidade Nova de Lisboa

Exploratory Data Analysis for Customer Segmentation in AIAI Loyalty Program

Group 24

Catarina Mendinhas, 20250422
Bárbara Franco, 20250388
Khadija Ennaifer, 20250439

Fall/Spring Semester 2025-2026

Contents

Abstract

This project supports Amazing International Airlines Inc. (AIAI) in developing a data-driven customer segmentation strategy. Using loyalty program (CustomerDB) and flight activity (FlightsDB) data from 2019 to 2021, we conduct Exploratory Data Analysis (EDA) to assess data quality, understand behavioral patterns, and prepare a modeling-ready dataset. We diagnose missingness, duplicates, and logical inconsistencies; engineer interpretable features (e.g., activity, redemption behavior, companionship, seasonality); aggregate monthly flight records to one row per customer; and merge the flight summary with the customer table.

Key discoveries include the identification of a substantial zero-income cohort representing 25% of customers, and weak correlations between demographic attributes and Customer Lifetime Value. The hypothesis of segmenting the customers into four groups was created after analyzing topics such as seasonality, proportion of flights, proportion of customer lifetime value, redemption of points, and the mean distance per flight. These insights motivate specific preprocessing choices and define a feature set that will underpin Phase 2 clustering. The outcomes lay a solid foundation for personalised marketing, improved retention, and more efficient points economics.

Keywords: Customer Segmentation, Exploratory Data Analysis, Loyalty Program, Flight Activity, Feature Engineering, Personalised Marketing.

1 Introduction

Amazing International Airlines Inc. (AIAI) operates in an intensely competitive market where customers face low switching costs, fares are highly transparent, and loyalty is influenced by alliance networks, co-branded cards, and flexible earning/redemption ecosystems. In this context, data-driven customer segmentation is essential to personalise offers, improve retention, optimise redemption economics (and breakage liability), and allocate marketing investment efficiently.

This project analyses AIAI's customer (CustomerDB) and flight activity (FlightsDB) data from 2019–2021 to build a solid foundation for segmentation. We explore behavioural and value patterns across members, consolidate flight history to a customer-level view, and combine it with demographics and loyalty metadata to enable actionable insights.

Specifically, we pursue the following analytical objectives: **Establish a unified Dataset** by aggregating monthly flight records to one row per loyalty member and merging the flight summary with the customer table; **Assess data quality**, **Choose interpretable features**; **Quantify distributions and relationships** to identify a behaviour-first feature set suitable for clustering, while anticipating preprocessing needs; **Define hypotheses and KPIs** that will guide Phase 2 clustering and subsequent marketing activation. The project follows the CRISP–DM methodology (Chapman et al., 2000), this Introduction corresponds to Business Understanding, while the subsequent analysis aligns with Data Understanding and prepares a clear path for Data Preparation and Modeling in the second part.

2 Data Analysis

2.1 Description of the Datasets

This project uses two primary datasets provided by Amazing International Airlines Inc. (AIAI).

CustomerDB contains detailed information about loyalty program members. The raw extract has 16,921 rows and 21 columns, including identifiers (*Loyalty#*), demographics (*Gender, Education, Marital Status, Province, City, PostalCode, Country*), program metadata (*EnrollmentDateOpening, CancellationDate, LoyaltyStatus, EnrollmentType*), and numeric measures such as *Income* and *Customer Lifetime Value (CLV).* Most fields are categorical/text; Income and CLV are numeric.

FlightsDB records monthly flight activity for loyalty members over 2019–2021, with 608,436 rows and 10 columns: *Loyalty#, Year, Month, YearMonthDate, NumFlights, NumFlightsWithCompanions, DistanceKM, PointsAccumulated, PointsRedeemed, DollarCostPointsRedeemed.* Variables are primarily numeric and capture frequency, distance, and points behaviour at the month level. Because there are multiple months per customer, FlightsDB must be consolidated (aggregated) to one row per customer before merging with CustomerDB.

2.2 Data Quality

We performed a structured, systematic set of checks and transformations to ensure analytical quality and to give solutions for constructing a clean dataset .

2.2.1 Data Type Verification

Customers Dataset
The columns *EnrollmentDateOpening* and *CancellationDate* were stored as objects, so they were converted to datetime type. They need to be converted to datetime to enable proper temporal operations, such as calculating durations, filtering by date or making temporal visualizations. The remaining features are in the correct datatype.

Flights Dataset
The feature *YearMonthDate* is stored as object and it was converted to datetime type for the same reasons as EnrollmentDateOpening and CancellationDate. 15.4% of the values in *NumFlights* and *NumFlightsWithCompanions* have decimal values(e.g., 9.9 flights appearing 7,537 times in the dataset), requiring rounding and casting to integer to properly reflect count data. *PointsAccumulated, PointsRedeemed, DollarCostPointsRedeemed* and *DistanceKM* are floats as well, however, they only have two decimal places, which is reasonable given that they represent points and kilometers. The other variables are in the expected type.

2.2.2 Missing Values

Customers Dataset
We identified missing values in three fields: *Income* (20), *CLV* (20), and *CancellationDate* (14,611). The 20 missing Income cases exactly overlap the 20 missing CLV cases, suggesting a shared origin or dependency. Further investigation revealed that these 20 customers have identical EnrollmentDateOpening and *CancellationDate* values—meaning they enrolled and cancelled on the same day. Cross-referencing with FlightsDB confirmed these *Loyalty#* values have zero flight activity records. Since these customers never actively participated in the program and provide no behavioural data for segmentation, we dropped these 20 rows entirely.

Since *CancellationDate* is present in 2,286 rows (13.5%), we infer 2,286 customers left the program (churned) and the remainder (14,635 or 86.5%) are active members. The missing values here represent active customers, which is the expected pattern and requires no imputation.

Flights Dataset
No missing values were observed in the original flight activity columns, indicating high data completeness.

2.2.3 Duplicates

Customers Dataset
No exact row duplicates were found when considering all columns. However, upon deeper inspection, the 'Unnamed: 0' column (row index) showed 20 duplicates—these correspond exactly to the rows with missing Income/CLV that enrolled and cancelled the same day. Additionally, 327 instances of duplicated Loyalty# existed with differing attributes (different Last Names, cities, and other demographics), indicating data entry errors rather than legitimate related accounts. All observations with duplicated loyalty numbers were dropped, since they consisted of only 1.9% of the data.

Flights Dataset
The duplicated loyalty numbers were also present in this dataset, however, as it was not clear which flight activity belong to which customer, we were not able to associate the flight activity with the corresponding client. For this reason, the duplicated *Loyalty#* were dropped both in the customers dataset and in the flights one. This dataset has 2,903 exact duplicate rows. However, after deleting those Loyalty numbers stops having duplicated information.

2.2.4 Logical Inconsistencies

Customers Dataset
We haven't found impossible values or contradictory data in this data set.

Flights Dataset
We flagged 5,901 rows (0.97%) where *NumFlights* = 0 but *DistanceKM* > 0. This logical inconsistency—recording distance without recording flights—is impossible and requires correction. These rows were flagged for correction by setting all activity metrics to zero or for exclusion prior to modelling. (see Figure 1)

2.8% of customers have redeemed more points than they earned, which is logically inconsistent. (see Figure 2)

1.2% of the customers have a cancellation date prior to the enrollment rate, which is not plausible. It was assumed that these dates were in the wrong order; thus, they were switched. (see Figure 3) Additionally, some customers have flight activity before they enroll. This would greatly affect the engineered features and the visualization, so a new column *CorrectEnrollmentDate* was created, which corresponds to the minimum between *EnrollmentDateOpening* and the date of the first flight of each customer.

2.2.5 Outliers

Customers Dataset
The variable *Income* does not show outliers. *Customer Lifetime Value*'s outliers consist of 8.8% of the data. These values are much higher so they could indicate customers with exceptional behaviour compared to the majority, and could be a group very important to highlight when considering personalized marketing. (see Figure 4)

Flights Dataset
The features *NumFlightsWithCompanions*, *DollarCostPointsRedeemed, PointsRedeemed, NumFlights, DistanceKM, PointsAccumulated* have the following outlier percentages: 17.5%, 5.8%, 5.8%, 0.6%, 0.3% and 0.3%, respectively. Most of this information isn't very relevant since the dataset is based on the flights and so each customer has 36 rows because there are 36 month in 3 years. It is expected that in most months of the year, there is no flight activity for the customers, so these variables as very right skewed, which the highest concentration on zero. However, the big percentage of high outliers in NumFlightsWithCompanions may suggest a grouping of customers that in a single month travel a lot with companions, since the percentage of Numflights doesnt come close to the 17%, these customers probably travel with company the majority of the time. (see Figure 5)

2.3 Variables' Distributions

CustomersDB
For demographics, both gender have a similar representation in the data, the majority of customers is married, followed by single and only then divorced and a striking amount has a Bachelor as the highest level of education. Most customers are from Ontario, British Columbia, and Quebec. An important insight is that all customers are from Canada, which makes this feature irrelevant from client segmentation, thus it will be dropped in the next deliverable. Most customers are in the Star tier, followed by Nova and only then Aurora. Furthermore, almost all customers enrolled in the Standard way, with some entering via a 2021 Promotion. For monetary, income and Customer Lifetime value have more concentrated in between 0 and 10 000 but contrary to CLV, income is well distributed until 100 000. 25% of customers have Income = 0, potentially representing unemployed customers, students, or individuals who chose not to disclose income. This substantial cohort requires special consideration in segmentation. (see Figure 6)

FlightsDB
The analysis was also done for this data set but the results are not very relevant.

2.4 Feature Engineering

We engineered features at two levels: (1) monthly flight records, and (2) aggregated customer-level flight summaries. We also created interpretable variables on the customer table to support segmentation.

2.4.1 Flight-level (monthly) features

'YearMonth' was created since the day is always 1. 'Season' was created in order to compute new features in the customers dataset and to contribute to customer segmentation without algorithms. As shown in

Table 1

2.4.2 Customer-level flight aggregation (one row per member)

Several new features were added to the already existing customer information and put into a new dataset *customers_merged*. New categorical features were mainly cumputed for profiling and for trying to segment the customers only with analysing and visualizing the data. Numerical features were added with their possible contribution to the clustering algorithms in mind. As shown in Table 2
Following the addition of this information some new insights were found: The preferred season to travel is Autumn and the majority of customers travel all year round, although the there are more flights registered in Summer (refer to the dashboard) . Most customers have travelled very recently with the Recency_Month 60% quantile still being 0 (0 months since the last travel). (see Figure 7)

2.5 Relationships between variables and Insights

Highly correlated features include PropNumFlights, PropNrFlightsWithCompanions, PropPoints and PropPointsRedeemed all with each other. Months_In_Program with Diversity_Season. NumFlightsWithCompanions_Max with Diversity_Season and TotalPoints. NumFlights_Max with Diversity_Season. (see Figure 8) Points_Redeemed and DollarCostPointsRedeemed have a linear relationship (see Figure 9) and so do 'DistanceKM' with 'PointsAccumulated' (Specifically, the Accumulated Points are 10% of the Distance in KM) (see Figure 10). These make them redundant.

All the customers with 0 income, which amount for at least 25%, have College as the highest educational level. In Canada, college is a preparatory school for university, this means that these customers are all probably still in university. From 30000 income and up we only have customers with only a bachelor degree, between 10000 and 30000 there is a mixture between Master, Doctor and High School. The income and the education does not seem to affect the number of flights. It does not seem that income affects the Proportion of flights per month. (see Figure 11)The plot of Customer Lifetime Value vs Proportion of flights is very similar. (see Figure 12)

Although Income and Customer Lifetime Value does not show correlation, the category in Income and Customer Lifetime value always match. The proportion of flights seems to decrease slightly with Customer Lifetime Value and Income. (see Figure 12)

The customers with higher recency have, on average, a bigger proportion of flights per month. (see Figure 13)

Customers who usually travel only in one season have higher ratio of redeemed points, have a lower proportion of flights per month but have much longer distanced trips, and have a bigger ratio of flights with companions and proportion of customer lifetime value per month

Customers who travel all year long have a much higher maximum number of flights and flights with companions in a month, but have a much smaller proportion of CLV per month and ratio of redeemed points. They are the ones that have travelled more recently as well. (see Figure 14)

As expected customers that have a smaller proportion of flights, travel to more distanced locations, have a bigger ratio of trips with companions and they also have a much higher recency. (see Figure 15)

There are no customers with low income or CLV in the Aurora Tier. And the loyalty tiers seems to be related with the Customer Lifetime Value, with Aurora having the biggest average of CLV and Star the lowest. (see Figure 16) (see Figure 17) (see Figure 18)

Customers who already left the program share some characteristics. Their proportion of flights is almost half of the one from active customers. On the contrary, they have almost double the Proportion of customer lifetime value. Adicionally, The maximum number of flights or flights with companions in a single month is a lot smaller but their ratio of points redeemed is slightly higher. They also do shorter distanced trips. (see Figure 19)

3 Results

After the exploratory data analysis and considering the problem at hand, there are certain numerical features that are more relevant for segmentation. 'TotalPoints' and 'TotalFlights' are highly correlated, so we would be introducing similar information with both, we choose 'TotalFlights' since it has higher correlations with the remaining features. Features that showed key findings and that were different across different categorical variables possess very important information for segmentation, thus 'Recency_Months', 'PropNrFlights', 'Ratio_Redeemed_Points', 'MeanDistancePerFlight', 'PropCLV', 'PropNrFlightsWithCompanions' and 'NumFlights_Max' will be introduced to the clustering algorithms. 'Customer Lifetime Value' and 'Income' will also be added since they do not have correlations with the other variables and are very important considering the business needs.

All the categorical and discrete variables will be later used for profiling, but they will not be used for the segmentation itself.

It will be very helpful in the profiling part, to see if the algorithms are able to clearly identify the group of customers who have cancelled the program, since, as we have seen in the data analysis, they have very similar features. This way, we could understand the patterns and identify customers at risk of churn and prepare a personalized marketing strategy. This is a pressing matter because the cancellation rate of the AIAI's program has been rapidly increasing. (see Figure 20)

Bonus questions 2 and 5 were accomplished and also revealed interesting insights about the data. The dashboards presents in a clear way KPI's and relationships between variables. The geospacial analysis aligns well with the predicted clusters.

4 Conclusion

This project cleaned and combined AIAI's customer and flight data into one clear record per loyalty member. We fixed errors, removed duplicates, and built useful features to show how people fly and use points.

With the added features, we can now see peak travel seasons, proportion of flights and points for each customer and who redeems points (savers vs spenders)

For the next phase we will drop irrelevant columns such as 'Unnamed: 0' and 'Country' and put 'Loyalty#' as the index of the customers data set. 'NumFlights' and 'NumFlightsWithCompanions' require rounding. This will be followed by outlier treatment. For features with a very high percentage of outliers, no treatment will be applied, since they could represent a group of customers with very extreme characteristics and behaviours, for example for 'MeanDistancePerFlight' and 'PropCLV'. Outliers in 'Recency_months could refer to customers who have cancelled a long time ago. For variables with smaller percentages, multiple methods of outlier treatment will be applied and the one that fits the data the best will be chosen. Then, scalling of the select features for clustering will be scalled, probably with StandardScaler() since there are mutlitple outliers in the data set. Several different clustering techniques will be employed, followed by profiling. They will be compared and the best one will be elected. Lastly, real actions will be sugestion for the marketing strategy, according to the behaviour and the characteristics of each group.

We hypothesied that at least 4 groups will be indentified. 1: **Seasonal travellers** that travel only in one season, as they have higher ratios of redeemed points and flights with companions, big proportion of customer lifetime value per month and longer distanced trips and small proportion of flights. **All year round travellers**, they are costumers that have travelled more recently. Contrary to the seasonal travellers, they have a small proportion of CLV per month and ration of redeemed points. They also have the higher maximum number of flights and flights with companions in a month. **At risk of churn**. These are clients with very small proportion of flights, maximum number of flights or flights with companions in a single month. Contrary to all year round travellers, they have a striking proportion of CLV and ratio of redeemed points. They also do shorter distanced trips are and the ones with bigger recency, as expected. There will probaby be another group, **Average customer** that shows more moderate behaviours and characteristics.

5 References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc. Accessed: 2025-10-11.

6 Appendices

Appendix A: Engineered Features

6.0.1 Flight-level (monthly) features

Table 1: Engineered Features at Monthly Flight Level

Feature	Description
YearMonth	Just the year and month from ‘YearMonthDate’
Season	corresponding Season of the month

6.0.2 Customer-level flight aggregation (one row per member)

Table 2: Engineered Features for Customer Dataset

Feature	Description
TotalFlights	Total number of flights per customer
TotalFlightsWithCompanions	Total number of flights with companions per customer
TotalPoints	Total Points per customer
TotalPointsRedeemed	Total Points Redeemed per customer
NumFlights_Max	Maximum number of flights in a month
NumFlightsWithCompanions_Max	Maximum number of flights with companions in a month
Racio_Flights_Companions	TotalFlightsWithCompanions / TotalFlights if TotalFlights is not 0
Racio_Points_Redeemed	TotalPointsRedeemed / TotalPoints if TotalPoints is not 0
MeanDistancePerFlight	The mean of the distance of the flights per customer
MostFrequentSeason	The season with the most amount of flights per customer
Diversity_Season	Number of seasons with flights per customer
Churn	1 if CancellationDate present (churned), 0 if missing (active)
Recency_Months	Number of months since the last flight; -1 if there are no flights
Months_In_Program	Number of months in the program within the flights dataset interval
Months_Since_Enrollment	CancellationDate - EnrollmentDateOpening in months
PropNrFlights	TotalFlights / Months_In_Program if the latter is not 0
PropNrFlightsWithCompanions	TotalFlightsWithCompanions / Months_In_Program if the latter is not 0
PropPoints	TotalPoints / Months_In_Program if the latter is not 0
PropPointsRedem	TotalPointsRedeemed / Months_In_Program if the latter is not 0
CLV_Category	CustomerLifetimeValue in 4 bins: Low, Medium, High, Very High
Income_Category	Income in 4 bins: Low, Medium, High, Very High
Recency_Category	Recency_Months in 5 bins: No Flights, Very High, High, Medium, Low
PropNumFlight_Category	PropNumFlights in 4 bins: Low, Medium, High, Very High

Appendix B: Visualizations

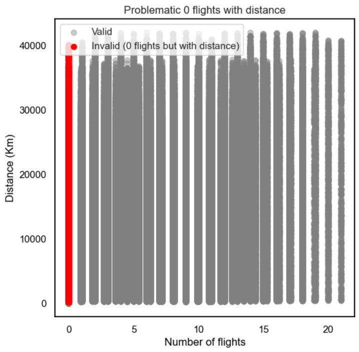

Figure 1: Observations with 0 flights but with 'DistanceKM' higher than 0

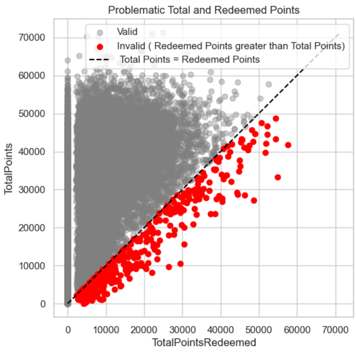

Figure 2: Total Points Redeemed vs Total Points

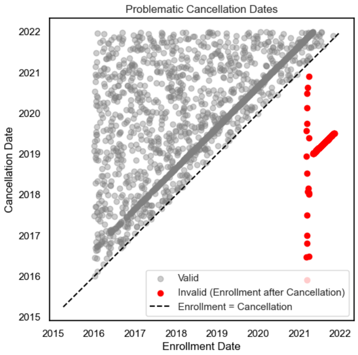

Figure 3: Enrollment Date vs Cancellation Date

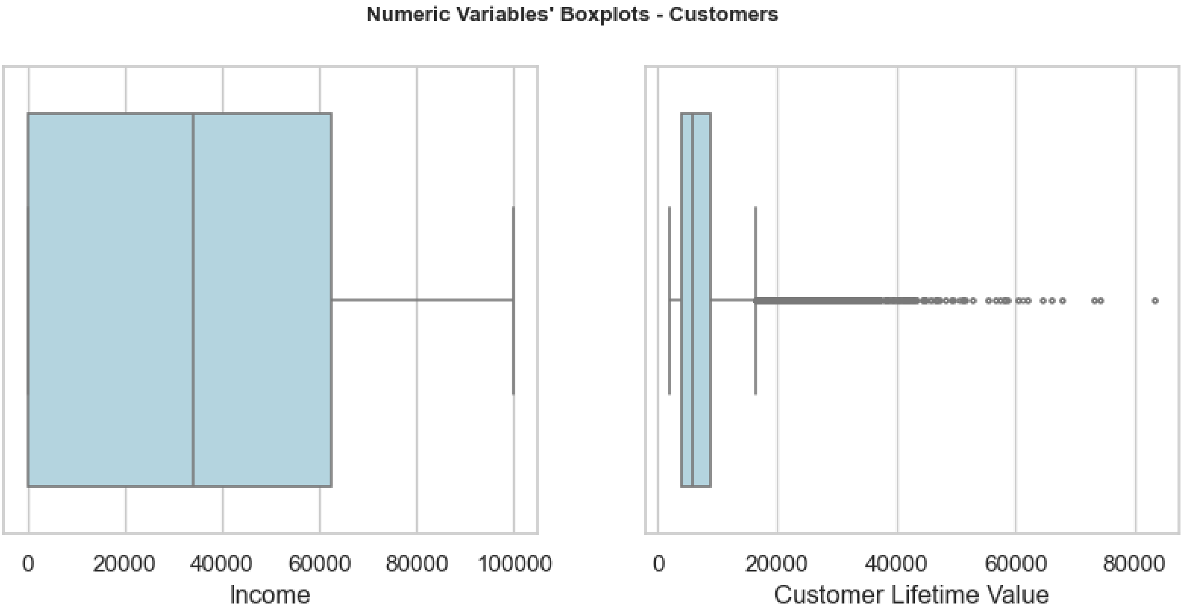

Figure 4: CustomerDB outliers

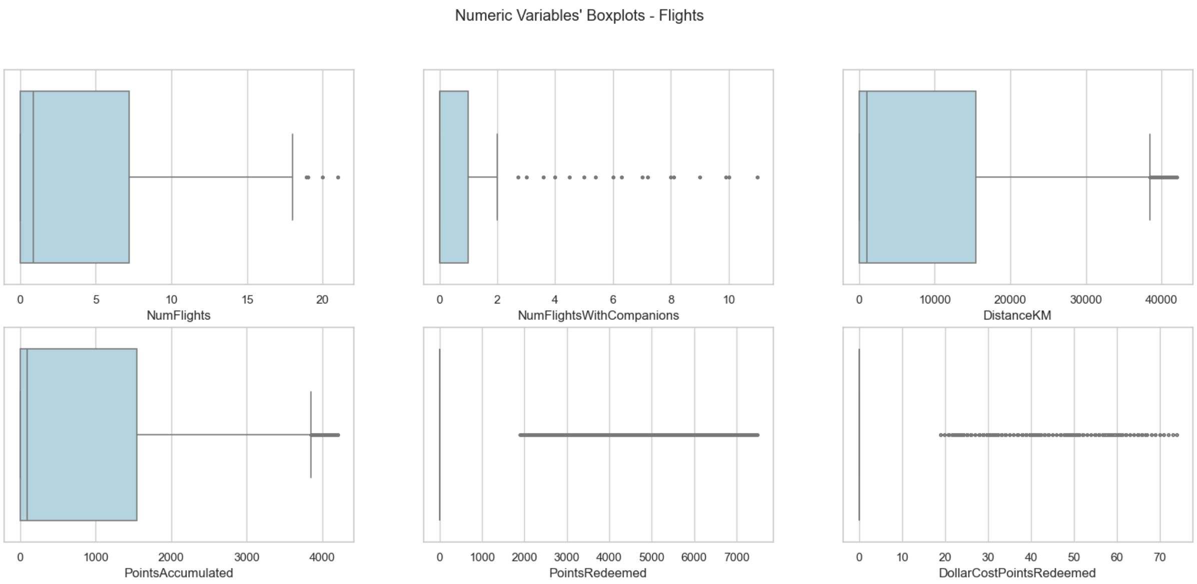

Figure 5: FlightsDB outliers

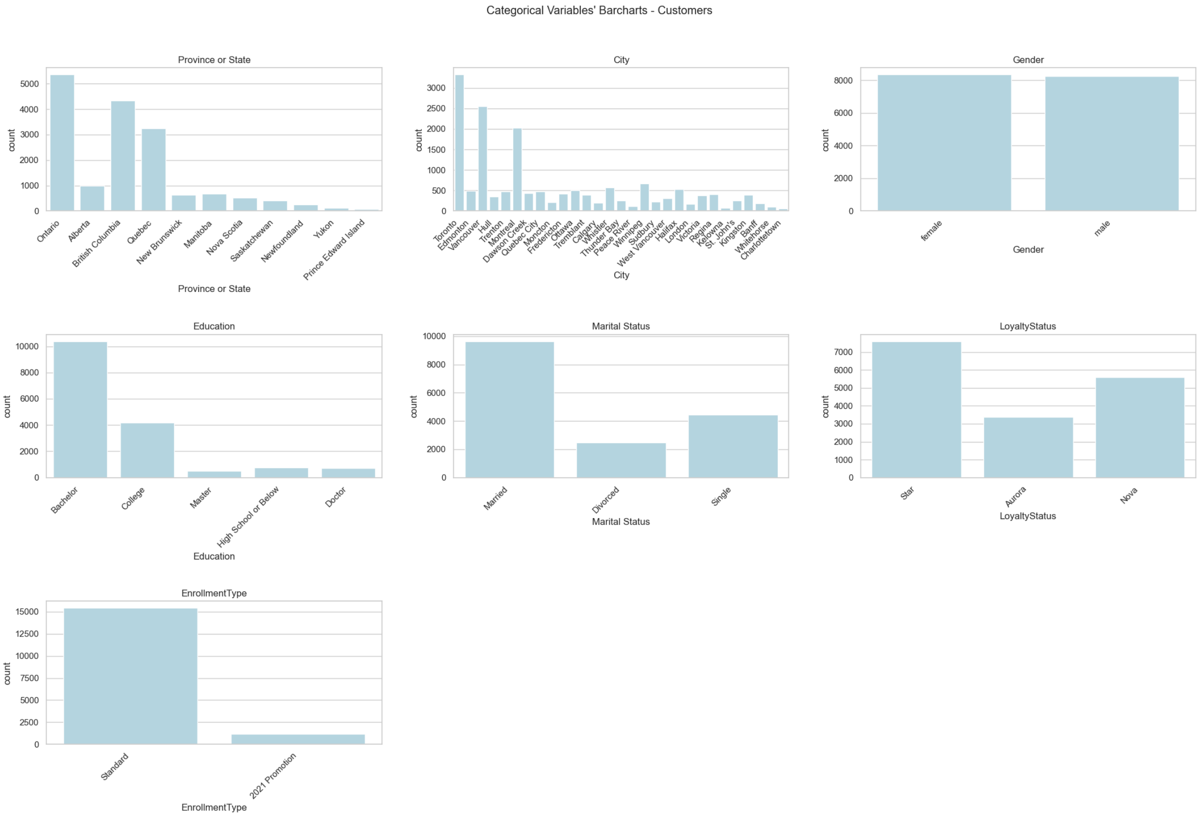

Figure 6: Barcharts of customers' variables

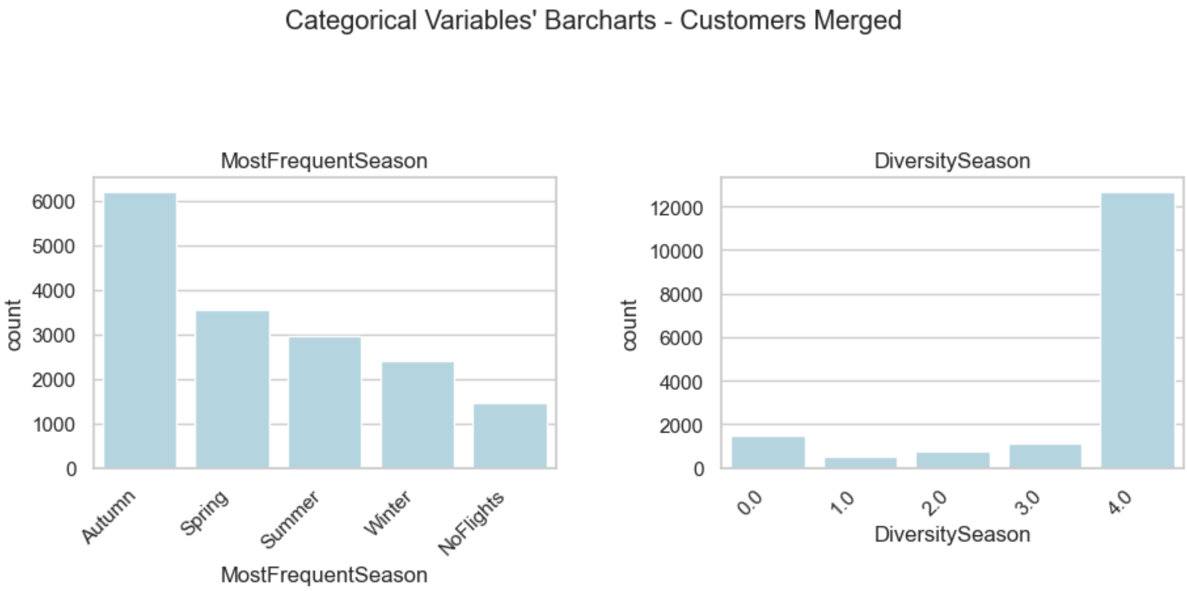

Figure 7: Bar chart of 2 new features

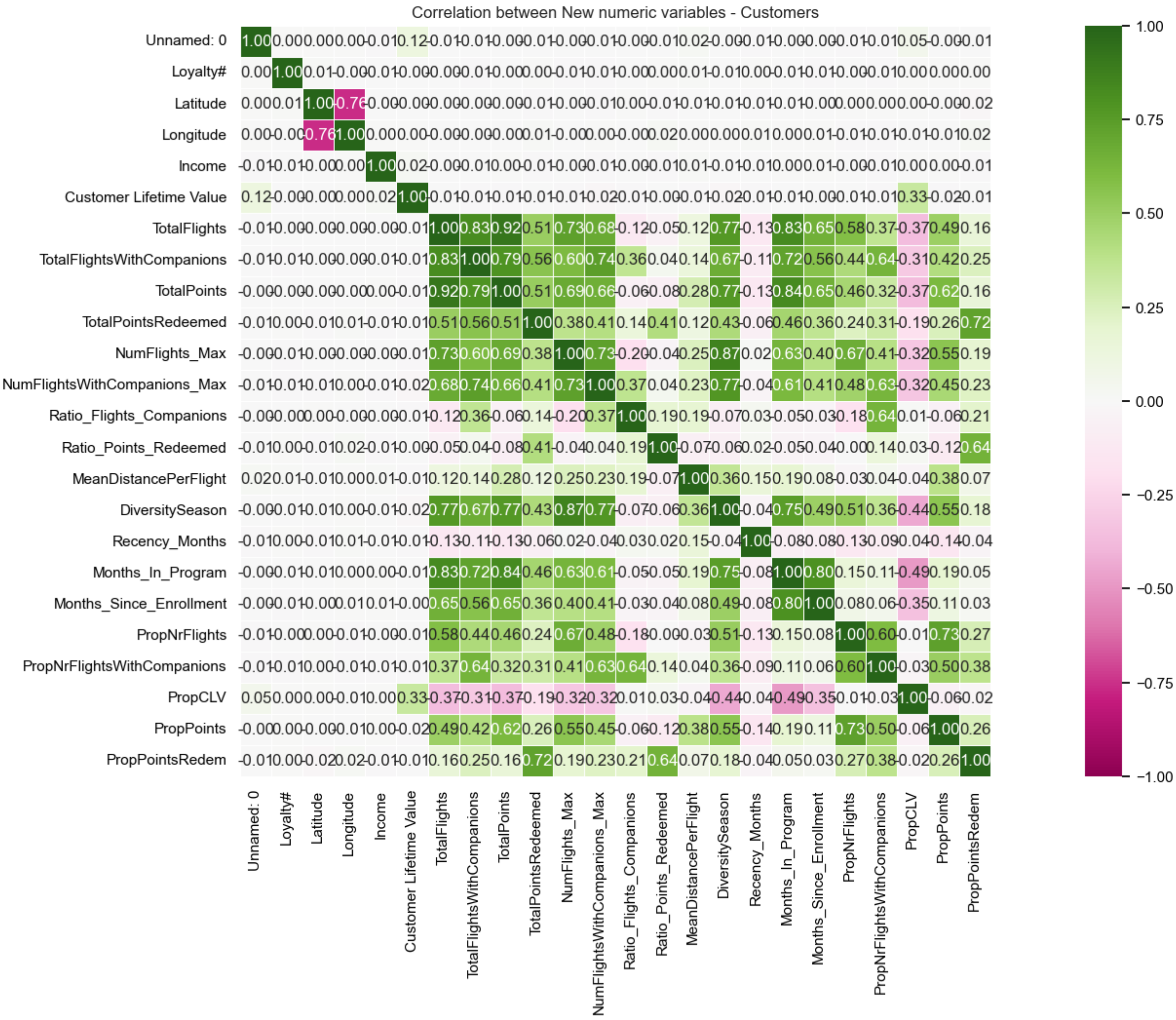

Figure 8: Correlation matrix with all features in customers_merged

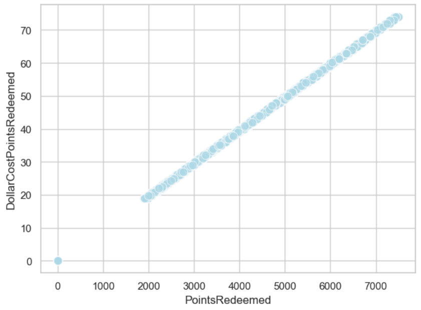

Figure 9: Points Redeemed vs DollarCostPointsRedeemed

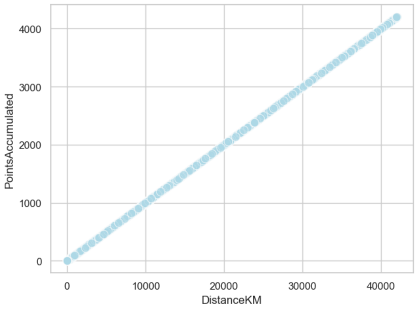

Figure 10: Distance in KM vs Points Accumulated

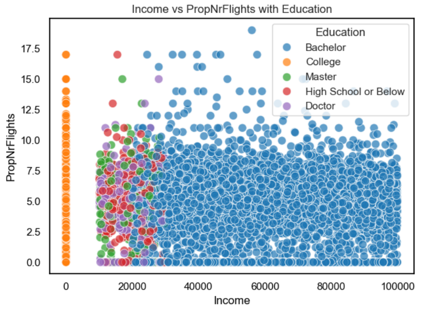

Figure 11: Income vs Proportion of flights per month

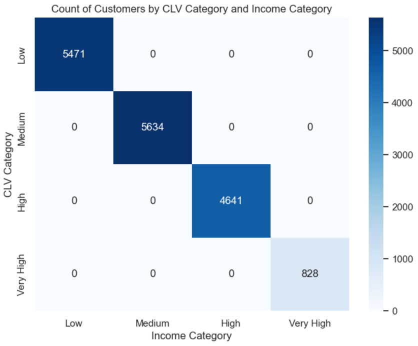

Figure 12: Income vs Customer Lifetime Value

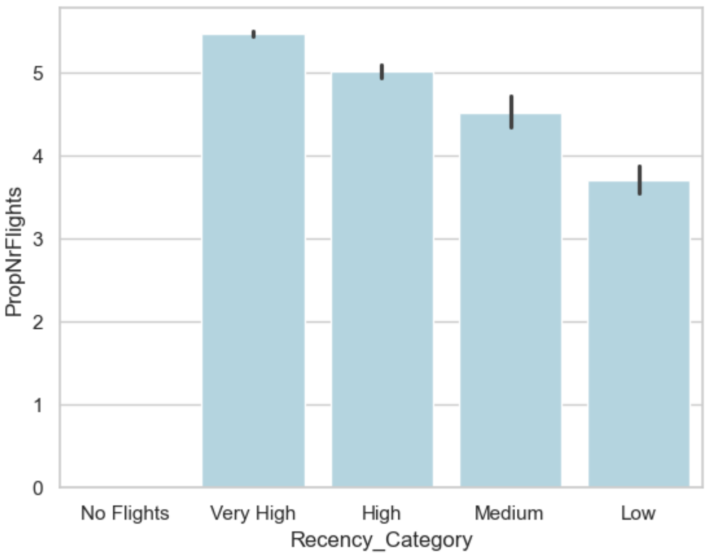

Figure 13: Recency vs Proportion of flights per month

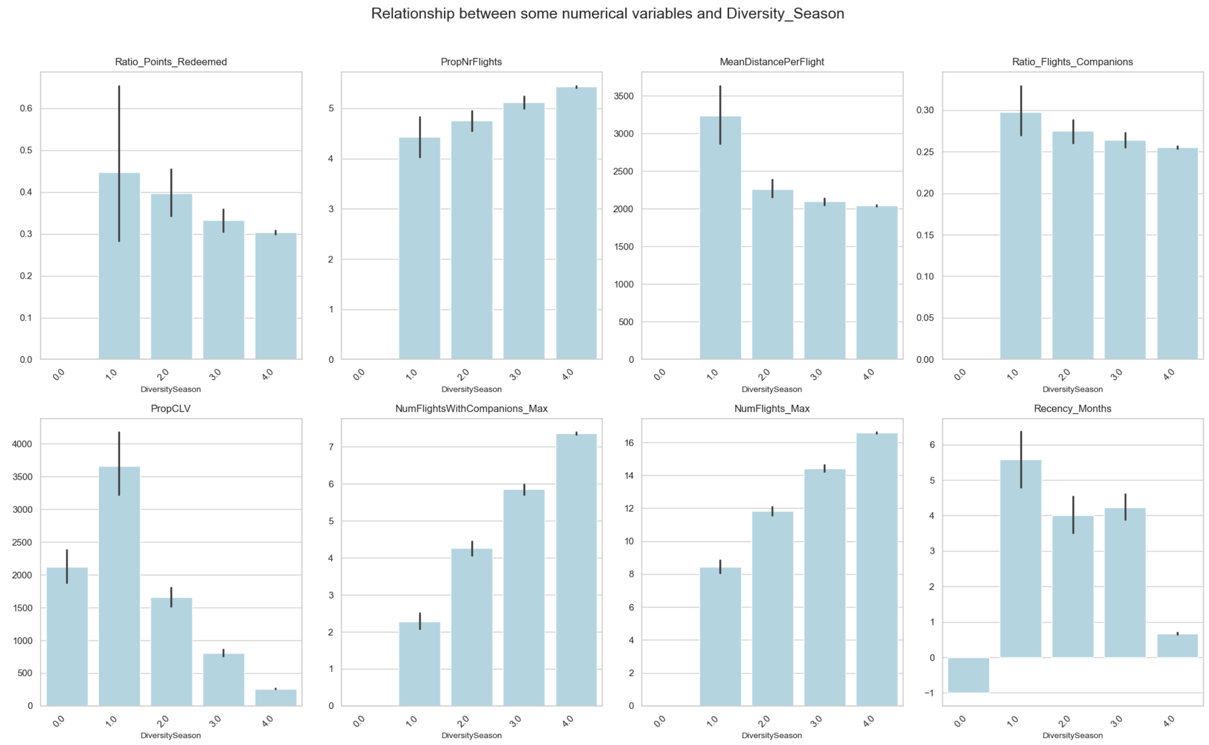

Figure 14: Relationship between Diversity Season (the amount of seasons where each customers travels) with some numerical features

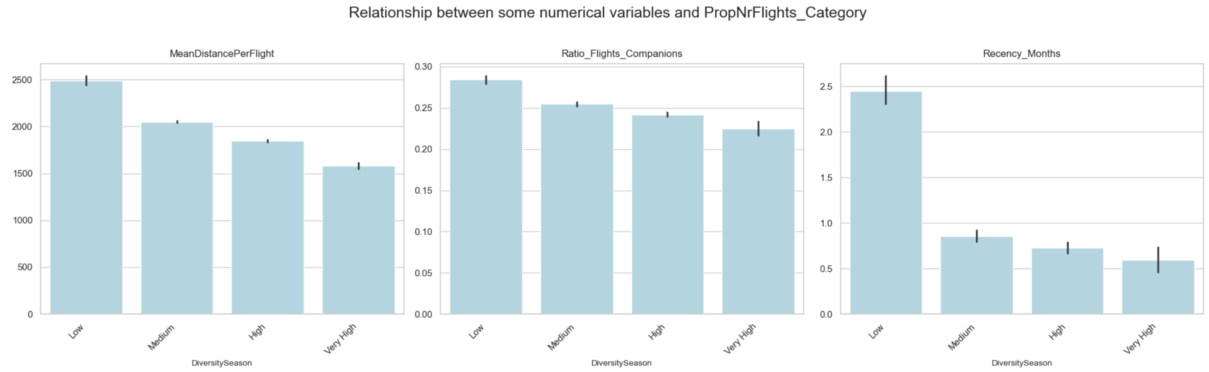

Figure 15: Relationship between the Proportion of flights per month and some other numerical features

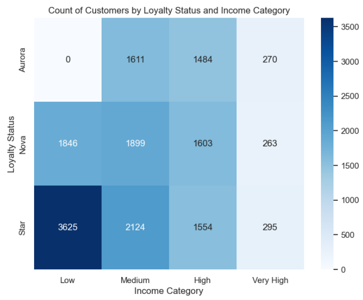

Figure 16: Income vs Loyalty Status

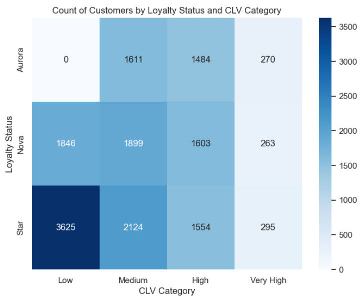

Figure 17: CLV vs Loyalty Status

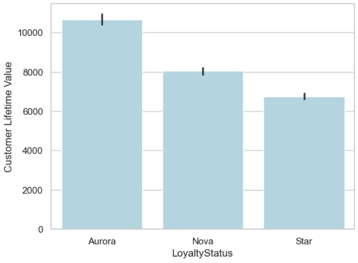

Figure 18: Loyalty Status vs CLV

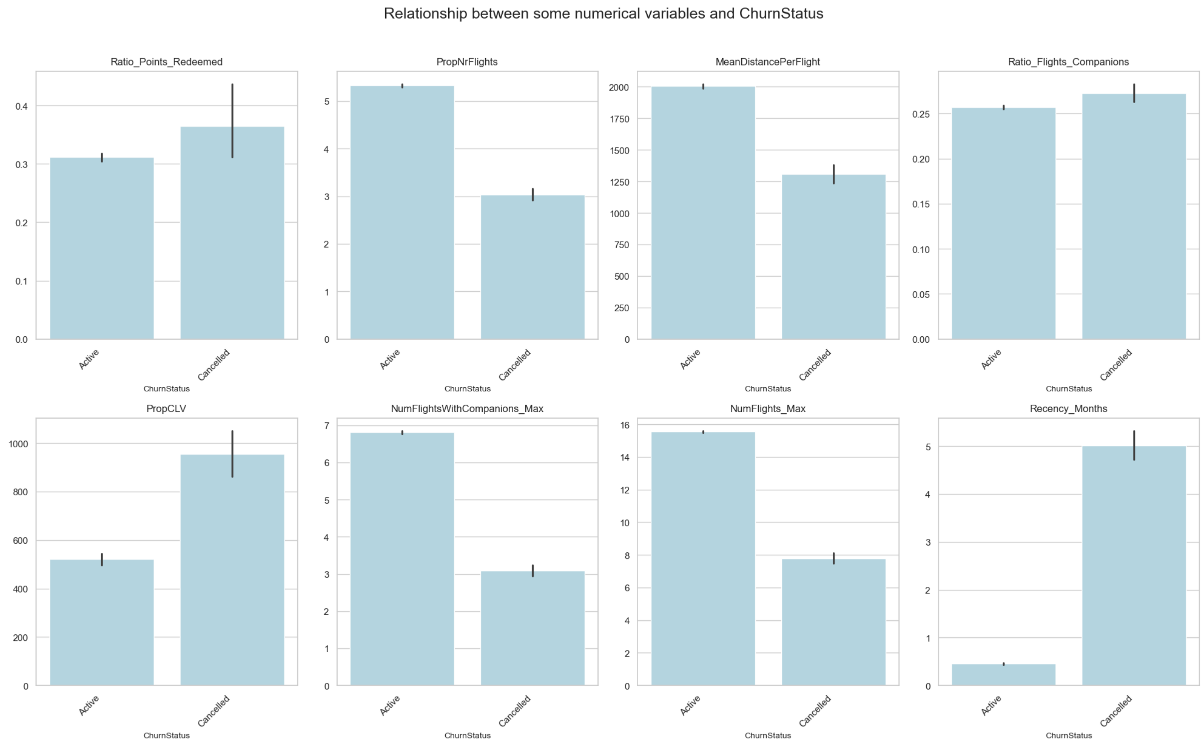

Figure 19: Relationship between Churn Status and other numerical features

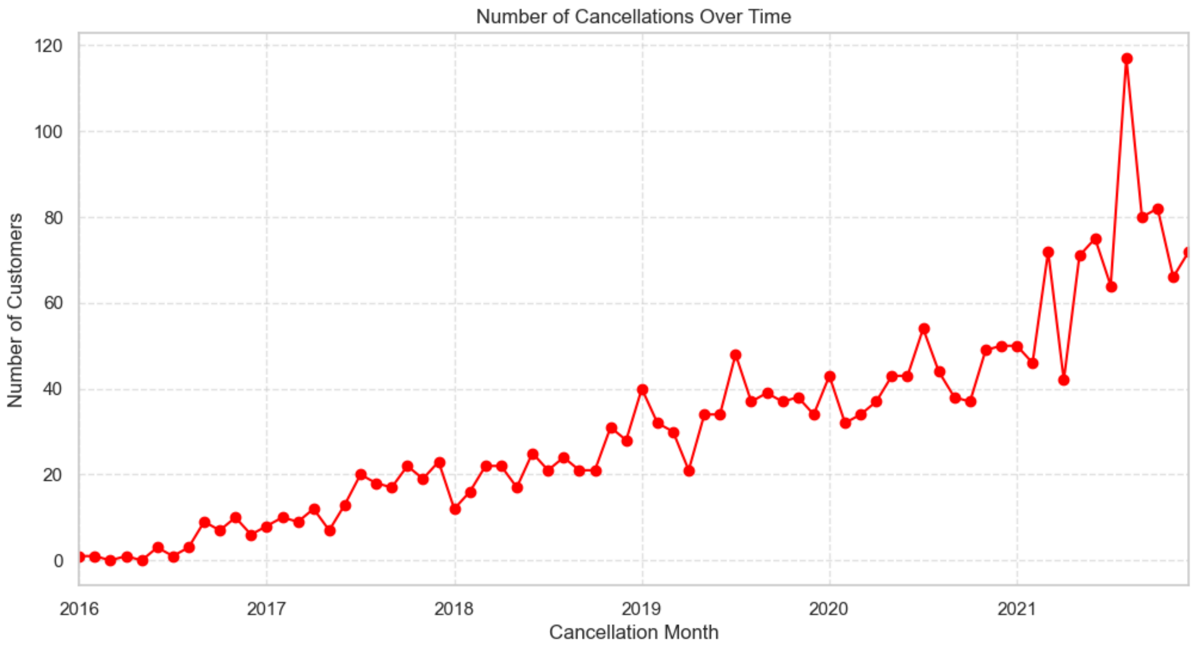

Figure 20: Churn Status over time

Appendix C: Data Quality Actions Summary

Issue	Count/Percentage	Action Taken
Same-day churners	20 (0.12%)	Dropped entirely
Duplicate Loyalty#	164	Kept first occurrence
Exact duplicate flights	2,903	Removed
Unnamed: 0 duplicates	20	Part of same-day churners
NumFlights = 0, Distance > 0	5,901 (0.97%)	Flagged for correction
Float NumFlights	7,537 occurrences of 9.9	Rounded to integer
Income = 0	25% of customers	Flagged for special treatment
Country = Canada only	100%	Column to be removed

Table 3: Summary of Data Quality Issues and Resolutions

Annexes

Some of the visualizations, Streamlit code segments, and the geospatial analysis notebook were developed with the assistance of AI-based tools to support implementation and visualization tasks. However, the majority of the analysis, interpretation, and development work was carried out manually by the authors.

Responsibility Statement

The authors declare that they are responsible for the content of this report. Certain visualizations, Streamlit code segments, and the geospatial analysis notebook were partially developed with the support of AI-based tools to facilitate implementation and visualization since the libraries were not still covered in class. Nonetheless, all analysis, interpretation, and validation of results were performed manually by the authors, who assume full responsibility for the final work.

Contribution

All team members contributed actively to the development of this project. The work was organized and shared through a common GitHub repository that was updated throughout the project time. The main analysis notebook was developed jointly by all team members, while specific components were led by different contributors: the written report was primarily prepared by Barbara and Khadija; the Streamlit dashboards were mainly developed by Caterina; the geospatial analysis notebook was produced by Khadija; the visualizations in the analysis notebook were primarily created by Barbara . As for the Canva materials, they were created collectively by all the members. Throughout the project, tasks such as data understanding, preparation, feature engineering, and model development were carried out collaboratively.