

DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Amazing International Airlines Inc.

Exploratory Data Analysis

Group 4

Pedro Santos, 20250399

Miguel Correia, 20250381

Pedro Fernandes, 20250418

Tiago Duarte, 20250360

Fall Semester 2025-2026

ABSTRACT

This project aims to build a customer segmentation strategy for Amazing International Airlines Inc. (AIAI) using three years of loyalty program and flight activity data. The analysis followed the CRISP-DM methodology, focusing on Business Understanding, Data Understanding, and Data Preparation.

During the exploratory phase, several data issues were identified and corrected, including duplicated customer identifiers, invalid date sequences, and inconsistent point balances. A left merge between CustomerDB and FlightsDB allowed the creation of new aggregated variables such as TotalFlights, TotalPointsAccumulated, TotalFlightsWithCompanions, and TotalPointsRedeemed, as well as derived features like RedeemRatio, Tenure, and IsActive.

The results show strong right-skewed distributions in Income and Customer Lifetime Value, meaning that a small group of customers contributes most of the airline's revenue. The analysis also revealed a steady increase in cancellations from 2016 to 2021 and confirmed that higher loyalty tiers include fewer but more valuable customers. These findings provide a clean and structured dataset ready for clustering, supporting future work on customer segmentation and loyalty optimization.

TABLE OF CONTENTS

1. Introduction	1
2. DATA ANALYSIS	1
2.1. CustomerDB:.....	1
2.2. Flights:	2
3. RESULTS.....	3
4. CONCLUSION.....	5
Bibliographical References	6
Appendix.....	7
Annexes	15
Annex 1 (AI USAGE STATEMENT)	15
Annex 2 (CONTRIBUTION STATEMENT)	16
Annex 3 (responsibility statement)	17

LIST OF FIGURES

Figure 1 - Descriptive Statistics of customers with "0" Income	7
Figure 2 - Income Distributions (Original, Group Median and KNN).....	7
Figure 3 - Points Redeemed vs Accumulated - Inconsistencies	8
Figure 4 - Relative Share of Flights per Enrollment Type	8
Figure 5 - Program tenure for costumers with CLV = 0	9
Figure 6 - Income by Education Level.....	9
Figure 7 - DistanceKM > NumFlights	10
Figure 8 -FlightsDB correlation matrix	10
Figure 9 - MergedDB correlation matrix	11
Figure 10 - Income vs Customer Lifetime Value hexbin.....	11
Figure 11 - Points Redeemed vs Numflights hexbin.....	12
Figure 12 - Points Accumulated vs NumFlights hexbin	12
Figure 13 - Total flights by Enrollment Tyoe	13
Figure 14 - Smoothed Cancellation Trend.....	13
Figure 15 - New Features Summary	14

1. INTRODUCTION

Amazing International Airlines Inc. (AIAI) operates a loyalty program serving a diverse customer base with varied travel frequencies, spending habits, and service preferences. However, its current uniform management approach leads to inefficient resource allocation and missed revenue opportunities. A data-driven segmentation strategy is thus essential to identify meaningful customer groups, enabling more targeted marketing, personalized engagement, and improved retention.

The core business issue lies in AIAI's limited ability to distinguish high-value frequent travelers from occasional flyers, resulting in undifferentiated rewards, suboptimal pricing, and weak marketing precision.

The analytical objective is to uncover behavioral, value-based, and demographic patterns explaining variations in loyalty and profitability. By examining variables such as lifetime value, flight frequency, redemption behavior, and socioeconomic attributes, the study aims to define actionable customer profiles that inform marketing and service strategies.

Following the CRISP-DM framework, this phase focuses on Business Understanding and Data Understanding, establishing the analytical foundation for clustering and segmentation modeling in the subsequent phase.

2. DATA ANALYSIS

This section evaluates the main datasets: CustomerDB and FlightsDB, assessing their structure, quality, and suitability for segmentation. The objective is to identify patterns, redundancies, and anomalies that inform feature engineering decisions for the clustering phase

2.1. CustomerDB:

This dataset gives us information about all the clients of the company. We will extract mainly value and geographical segmentation. The dataset consists of 16 921 records across 19 features.

- **Loyalty#:** This is supposed to be the unique identifier of the customer and key of the database, but it has a lot of repeated values. We must keep only one, as an ID cannot be duplicated.
- **First Name, Last Name and Customer Name:** The name of the customer is not a useful attribute, all 3 features should be dropped.
- **Latitude and Longitude:** There are 49 unique results in each feature across the dataset, meaning we expect these points to be reference locations and not the real locations of customers. Each one of these 49 values relates to a pair Latitude-Longitude.
- **Income:** Income values are very right skewed due the high number of zeros, reflecting heterogeneity. It has 20 missing values and these need to be treated.
- **Customer Lifetime Value:** Also right skewed due to the high number of values near 0. This will be interesting to value based segmentation since we are interested in the higher value customers. It has 20 missing values that need to be evaluated.
- **Country:** Since all customers are from Canada, this variable is constant and should be excluded.
- **Province or State:** Customers are distributed across 11 provinces or states. This variable may be interesting to geographic insights.

- **City:** There are 29 cities across 11 provinces, providing an additional layer of geographic granularity for the analysis. The City variable is closely related to Latitude and Longitude, as larger metropolitan areas contain multiple coordinate points (typically three to four), while smaller cities tend to have only one. We concluded that there is some redundancy here.
- **Postal Code:** There are 75 unique postal codes, offering greater geographic granularity than the coordinate variables. Given the high level of granularity, this variable may contribute limited value to segmentation and could introduce unnecessary noise into the analysis.
- **Gender:** There is an almost equal distribution of male and female in the dataset.
- **Education:** There is a predominance of customers holding a Bachelor's degree, followed by those with a College education, while Doctorate, Master's, and High School levels are less represented. This may be a good variable to help understand how customers with different education levels behave.
- **Location Code:** There is also an almost equal distribution of Suburban, Rural and Urban clients, which may be useful for geographical segmentation.
- **Marital Status:** More than half of the customers are married, followed by single and divorced individuals. This variable may influence travel companionship patterns, as married customers are more likely to travel with partners or family members
- **LoyaltyStatus:** Star is the most common loyalty status followed by Nova and Aurora. By analyzing the univariate LoyaltyStatus plot and the bivariate LoyaltyStatus x Customer Lifetime Value plot, the assumption that exists is that the hierarchy follows the order: Star -> Nova -> Aurora, being Aurora the most premium tier of the loyalty program. This variable is useful to understand how the different tiers in the loyalty program behave.
- **EnrollmentDateOpening:** This variable spans from 2015 to 2021, allowing the calculation of customer tenure within the loyalty program.
- **CancellationDate:** Most of the values are NA's (86,35%) and we will assume these are still active customers. This variable has valid entries (true cancelled memberships) in around 13,65% of the dataset, and those may be used to predict churn. This variable is also useful, as stated before, to calculate the tenure of customers in the program. When combining with the above, it was found that there are some customers who have a CancellationDate before an EnrollmentDateOpening, which, is impossible to happen – this consists of an inconsistency.
- **EnrollmentType:** Almost all our customers did not enroll through the 2021 promotion. However, it may be interesting to find if these customers have different behaviors and whether the promotion was successful or not.

2.2. Flights:

This dataset gives us information about all the flights and points of a customer in a given month of a given year. We will extract mainly behavior, economic value and geographical segmentation. The dataset consists of 608436 records across 9 features.

- **Loyalty#:** This value is the connection to the customerDB. Each line of this dataset consists of a key (Loyalty#-Month-Year) of each customer. This is supposed to be the unique identifier of the customer, but it has a lot of duplicated values (these will be dropped).
- **Year:** There are essentially 3 years: 2019, 2020 and 2021. It may be important to understand the evolution of the number of flights.

- **Month:** All months are present (1 to 12). Same as the above, it may be important to understand seasonality and in what months customers fly.
- **YearMonthDate:** This variable presents the same redundant information from year and month. Should be removed.
- **NumFlights:** There is a high frequency of zero values, indicating that in most months, most customers did not take any flights. This pattern suggests potential seasonality in travel behavior.
- **NumFlightsWithCompanions:** This variable follows a similar distribution from NumFlights. This is important to understand if customers flight alone or with companions. May be useful to create clusters of leisure travelers, family travelers or work travels.
- **DistanceKM:** The distribution is also like that of the NumFlights variable. However, some records showed non-zero distances in months with zero flights, indicating data inconsistencies. This variable has a perfect correlation with PointsAccumulated, so the decision was to drop it.
- **PointsAccumulated:** This variable gives us the points earned in a month and has a perfect correlation with the kilometers made. To maintain consistency of “Points”, the above one will be dropped.
- **PointsRedeemed:** There is a high number of 0 points and then a nearly normal distribution from 2000 to 7000 with mean in 4000. This feature is important because it reflects customers’ engagement with the loyalty program, indicating how actively they redeem earned rewards and thus differentiating loyalty-driven from value-hoarding behaviors.
- **DollarCostPointsRedeemed:** It reflects the distribution of PointsRedeemed, with a correlation of 1 and so it should be dropped.

3. RESULTS

Strange values:

- **Income Variable treatment:** Twenty customers had missing income values and 25% of all entries report “0” income. At first, we thought that these may be children or students without any income. However, as can be seen in [Figure 1](#), most of these customers are from Ontario and Toronto, mainly single (although only 57%, which means 43% are either married or divorced), male, and with College education (above 18 years) - so these should be treated as strange values rather than real income. We compared two imputation methods: K-Nearest Neighbors (KNN) and Group Median. The Group Median approach significantly distorted the original distribution, while KNN provided smoother, more realistic results. We therefore selected KNN for income imputation. More details in [Figure 2](#).
- **Date inconsistencies:** A brief validation identified customers whose cancellation date preceded their enrollment date. Initially, these cases were treated as data errors. However, as can be seen in [Figure 3](#) (customers enrolled in 2021 had flights pre-2021), they may reflect customers who canceled their membership and reactivated it under the same loyalty number. To address this, whenever such inconsistencies occurred, the invalid cancellation dates were replaced with NaN values, aligning them with cases of customers who had not canceled.
- **Points Accumulated and Points Redeemed:** Some customers show more points redeemed than accumulated, which is logically impossible. At first, when the monthly records were checked, it was assumed that it was normal for some months to show customers spending more points than they earned in that month. However, it should not be possible for this to

occur in overall totals. ([Figure 4](#)). We fixed them by capping TotalPointsRedeemed to match TotalPointsAccumulated, ensuring RedeemRatio ≤ 1 , while keeping all other information intact.

- **Education-Income Anomaly:** Bachelor's degree holders earned twice the income of Master's, High School or Below, and Doctorate degree holders, who had similar income levels. This unexpected finding warrants further investigation, as we anticipated higher earnings for advanced degree holders. More details in [Figure 5](#).

Duplicated values:

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

Missing values:

- **Missing Value Treatment in Customer Lifetime Value:** Twenty customers lacked Customer Lifetime Value data. Analysis revealed these were customers who had never taken a flight, so we appropriately imputed zeros for these cases. We also found; these zero-value customers had been enrolled in the program for an average of 44 months - contrary to our first thought that these may be new joiners to the program ([Figure 6](#)).

Removing Redundant Values:

- **Redundancy:** Customer Name, First Name, and Last Name are redundant and don't give any meaningful information, so were dropped.
- **Geographic Variables:** We evaluated geographic granularity across Country, Province/State, City, Postal Code and Latitude/Longitude levels. Since Country had only one unique value, it provided no information. We observed minimal differences in clustering quality between Province/State and City levels, so we retained Province/State due to its lower cardinality and dropped the remaining geographic variables including Latitude, Longitude, and Postal Code.
- **Temporal Variables:** In the flights dataset, we removed YearMonthDate since Year and Month columns already provided this information with easier accessibility.

Data Types:

- Flight counts and flights with companions contained float values that were truncated to integers for consistency. We identified months where customers recorded zero flights but non-zero distance kilometers, indicating data quality issues, as can be seen in [Figure 7](#).

Correlations:

- **Perfect Correlations:** We discovered perfect correlations between two variable pairs: DistanceKM-PointsAccumulated and PointsRedeemed-DollarCostPointsRedeemed ([Fig. 8](#)). To maintain consistency in points logic, we dropped DistanceKM and DollarCostPointsRedeemed.
- **Correlations:** After the merge, we found 91% correlation between TotalPointsAccumulated and TotalFlights after merging datasets, leading us to drop TotalPointsAccumulated and retain TotalFlights. More details in [Figure 9](#).

New Features and Key Findings:

- **Feature Engineering and Variable Selection (Figure 15):** We created an "IsActive" boolean feature to replace the CancellationDate column, which contained predominantly null values. Customers without a cancellation date were classified as active. We also created the feature "Tenure", is crucial as it quantifies the duration of the customer's relationship with AIAI, enabling the assessment of loyalty, retention, and temporal behavior patterns.
- **Merge Strategy:** We performed a left merge to preserve the 20 customers not present in the FlightsDB. Before the merge, we created aggregate features: TotalFlights, TotalPointsAccumulated, TotalFlightsWithCompanions, and TotalPointsRedeemed, separate from monthly and yearly flight metrics.
- **Income- Customer Lifetime Value Relationship:** Although overall correlation between Income and Customer Lifetime Value was very low (0.024), we observed a strong concentration of low-income customers among those with low lifetime value, suggesting a meaningful relationship at the lower end of the distribution, as can be seen in [Figure 10](#).
- **Points and Flight Behavior:** While PointsRedeemed and NumFlights showed modest overall correlation, we identified strong concentrations in specific ranges: 2.5-4 flights with 3,000-4,500 points redeemed ([Figure 11](#)). Similarly, NumFlights and PointsAccumulated showed clustering at 2.5 flights and between 7.5-11 flights with 50-250 PointsAccumulated ([Figure 12](#)).
- **Enrollment Type Differences:** Standard enrollment customers exhibited twice the number of flights and flights with companions compared to other enrollment types, despite having equivalent income levels and customer lifetime values ([Figure 13](#)). This suggests enrollment type significantly influences engagement patterns independent of economic factors.
- **Cancellation Patterns:** The smoothed trend reveals a steady and almost uninterrupted rise in cancellations from 2016 through 2021 ([Figure 14](#)). This sustained upward movement suggests increasing customer attrition over time, pointing to unresolved loyalty or satisfaction issues that have intensified in recent years.

4. CONCLUSION

This exploratory phase established a solid foundation for clustering by resolving data inconsistencies, removing redundancy, and engineering features that reflect customer value, behavior, and profile diversity. Based on these insights, we propose a three-perspective clustering approach:

- Value-based, distinguishing customers by profitability and spending potential.
- Behavioral, capturing travel frequency, loyalty engagement, and redemption activity.
- Demographic-geographic, identifying lifestyle and location-driven patterns.

We will begin with K-Means for its interpretability and complement it with Hierarchical Clustering to validate stability. After scaling and dimensionality checks, we expect to identify 3–5 meaningful customer segments, ranging from high-value loyal travelers to low-engagement members. These clusters will provide a data-driven basis for targeted marketing, loyalty optimization, and strategic decision-making in the next project phase.

BIBLIOGRAPHICAL REFERENCES

Baçaõ, F. L. (2025). Data Preparation - S5 class

Baçaõ, F. L. (2025). Information Visualization - S4 class

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

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

APPENDIX

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

Figure 1 - Descriptive Statistics of customers with "0" Income

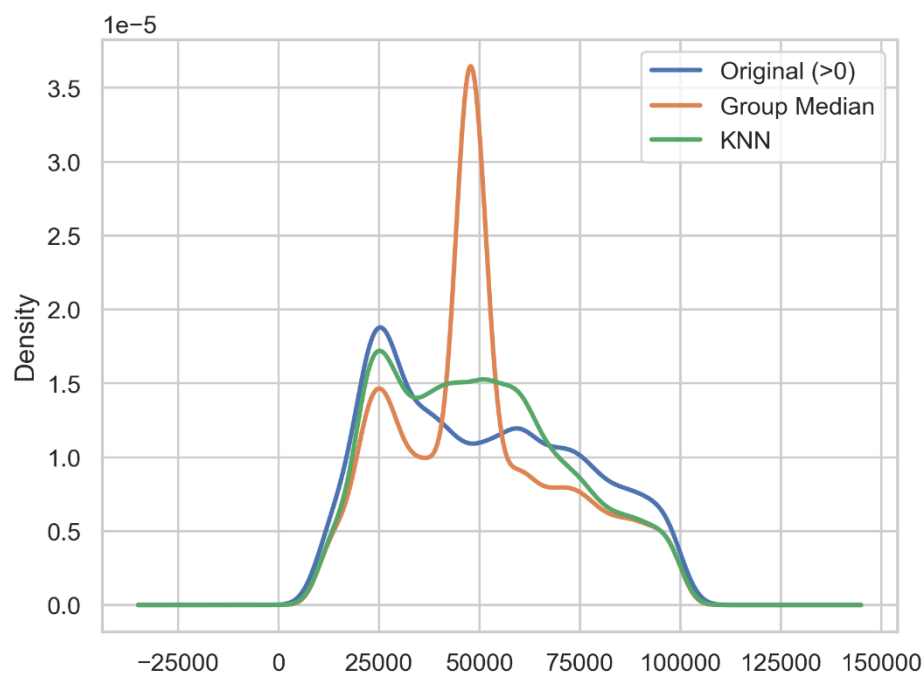


Figure 2 - Income Distributions (Original, Group Median and KNN)

Relative Share of Flights by Enrollment Type (2019-2021)

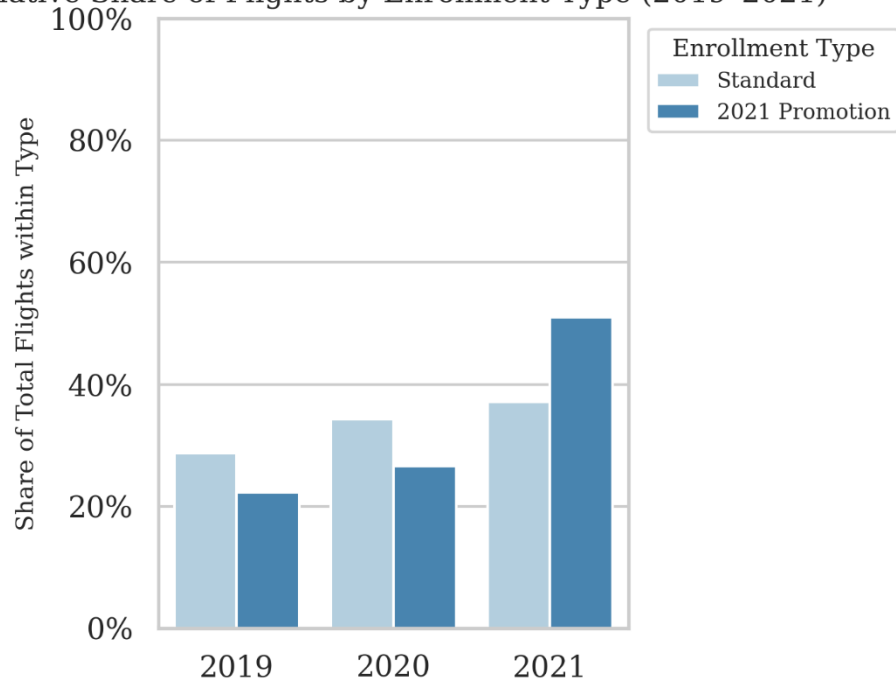


Figure 4 - Relative Share of Flights per Enrollment Type

2

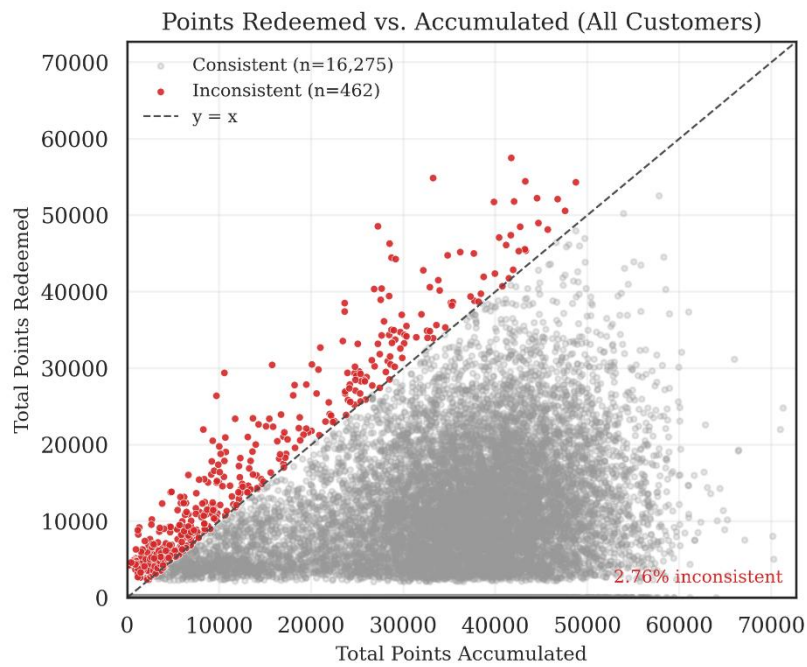


Figure 3 - Points Redeemed vs Accumulated - Inconsistencies

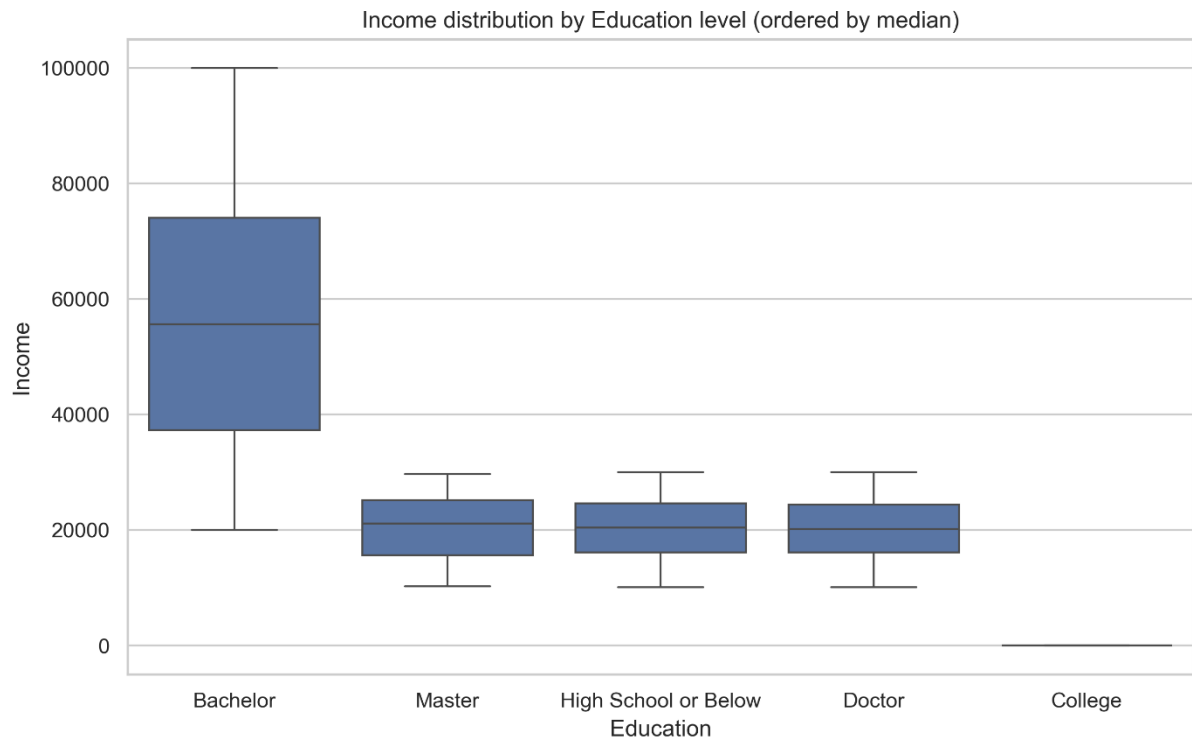


Figure 6 - Income by Education Level

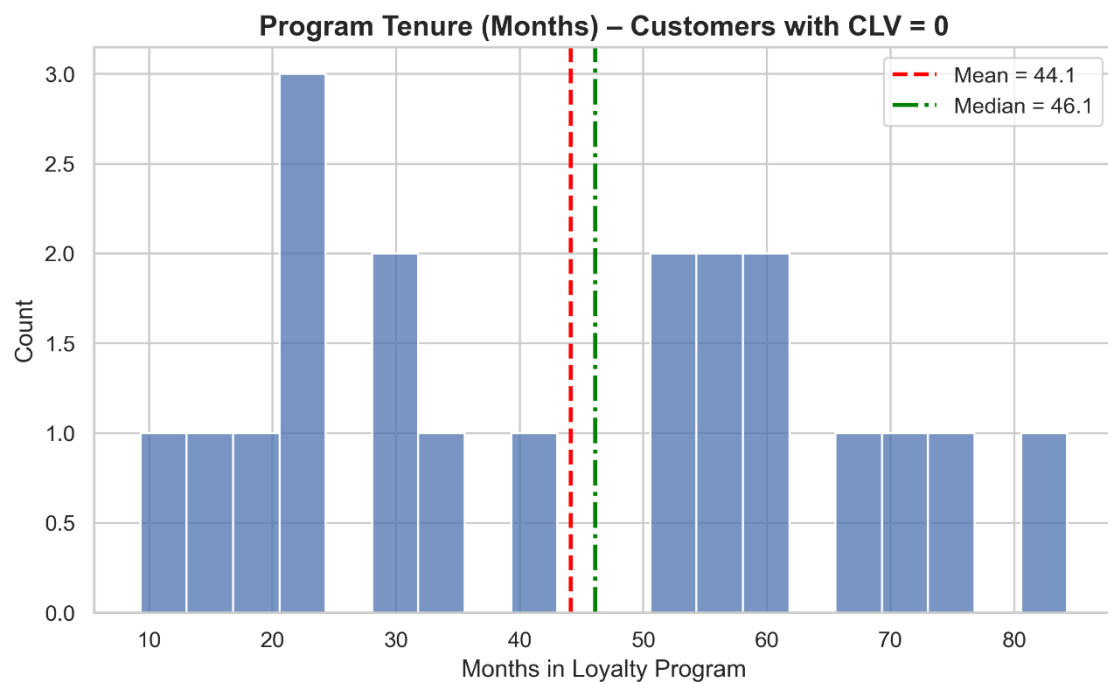


Figure 5 - Program tenure for customers with CLV = 0

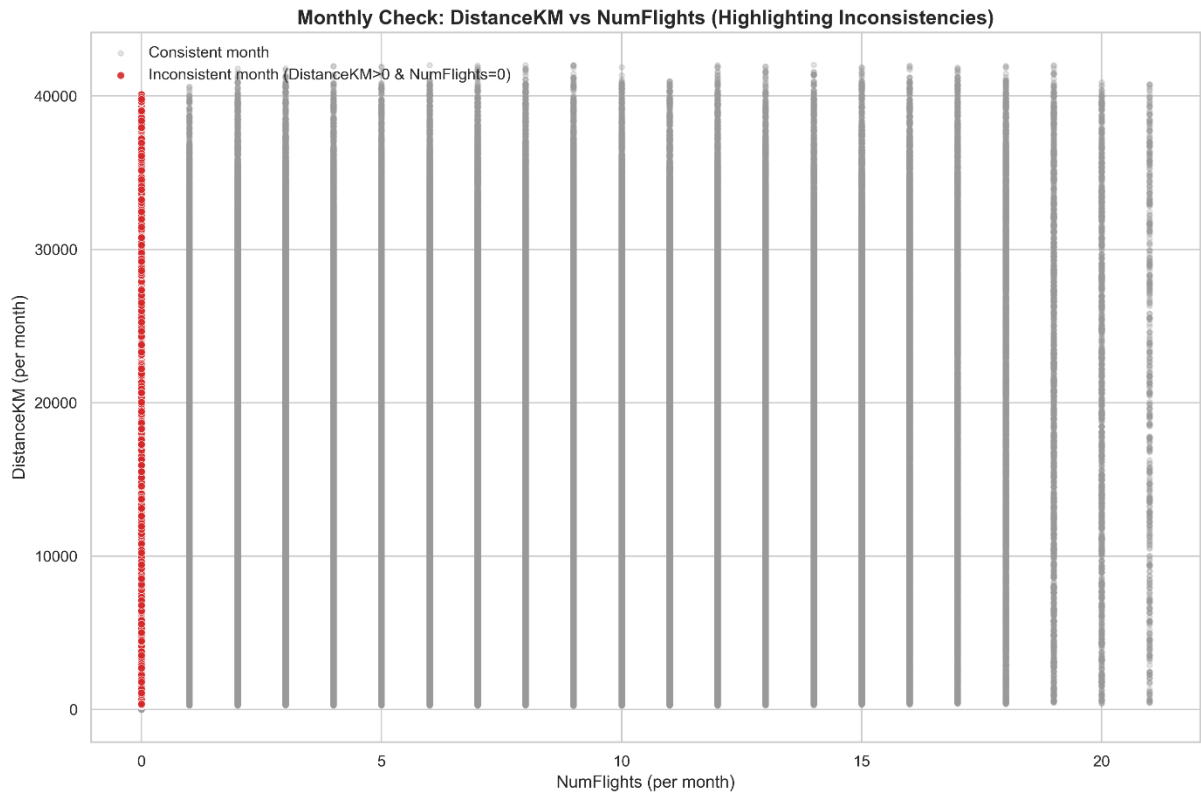


Figure 7 - DistanceKM > NumFlights

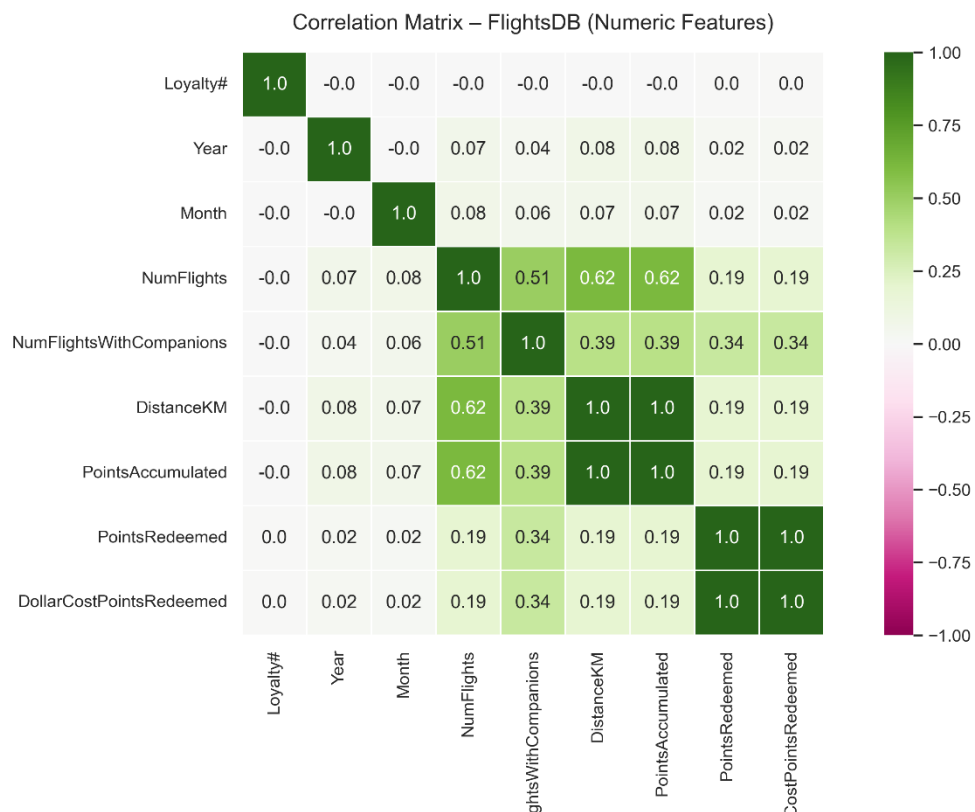


Figure 8 -FlightsDB correlation matrix

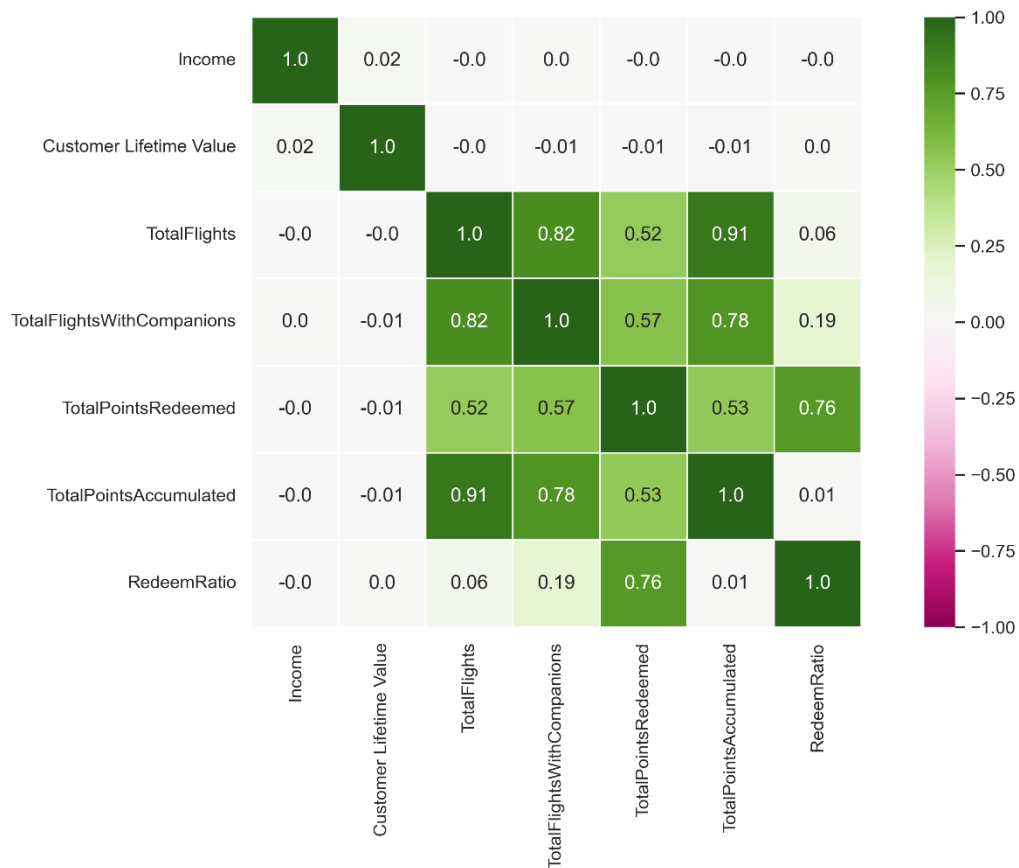


Figure 9 - MergedDB correlation matrix

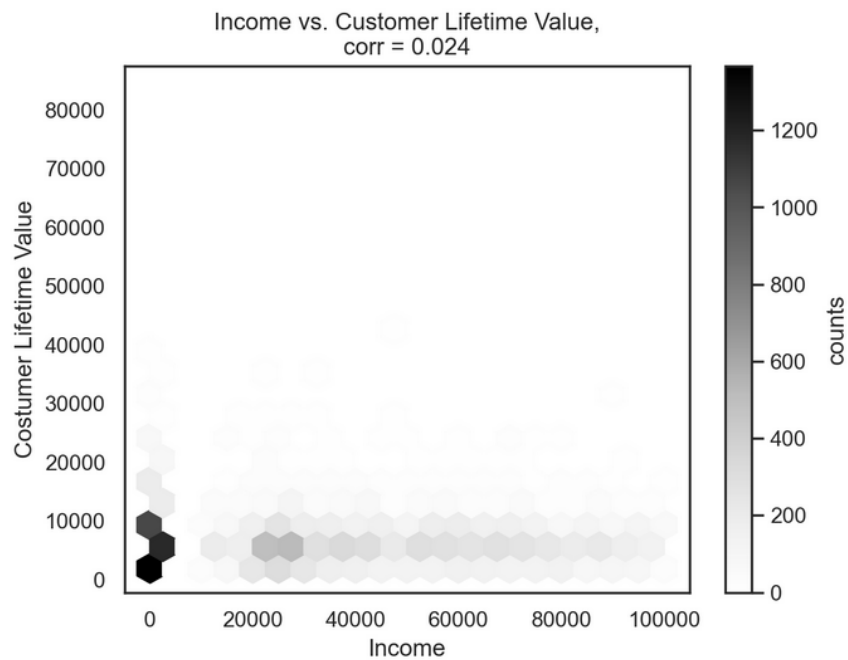


Figure 10 - Income vs Customer Lifetime Value hexbin

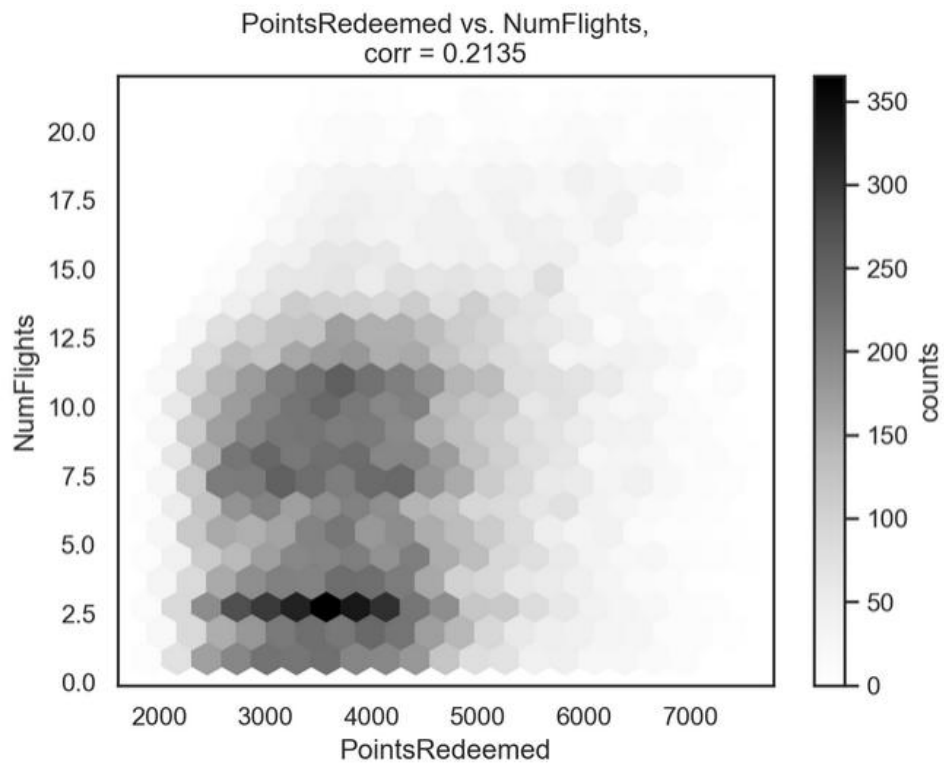


Figure 11 - Points Redeemed vs Numflights hexbin

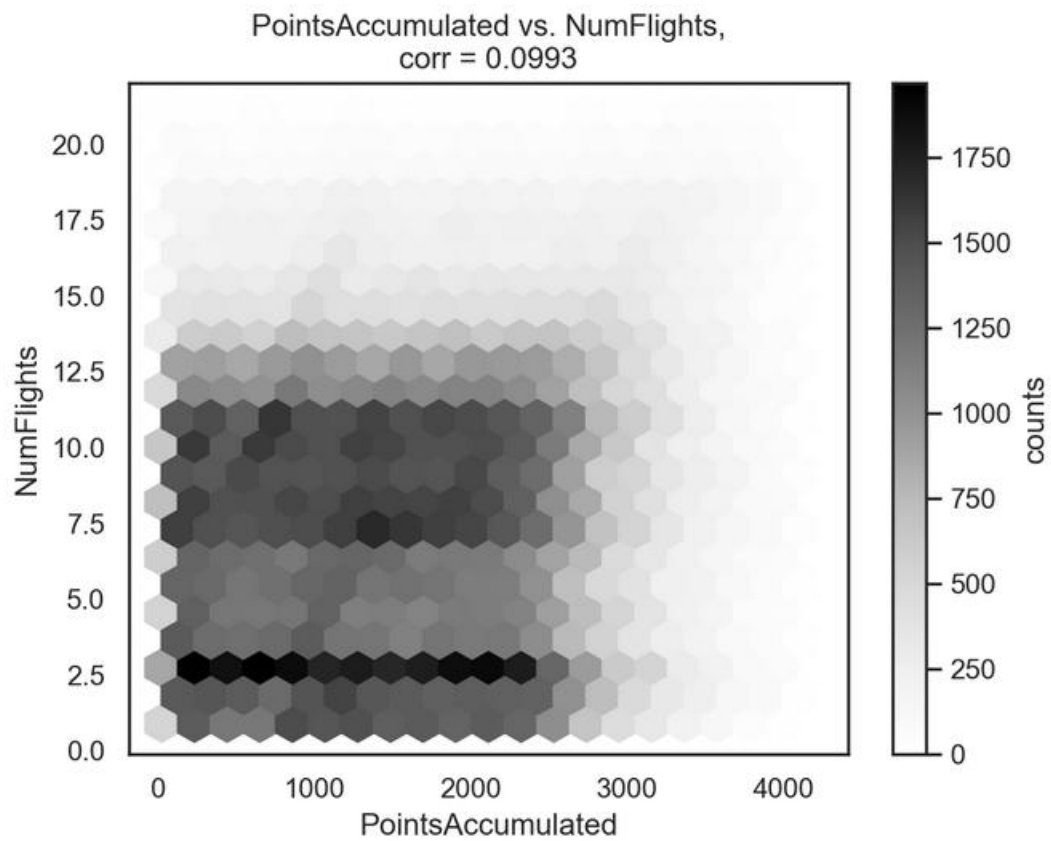


Figure 12 - Points Accumulated vs NumFlights hexbin

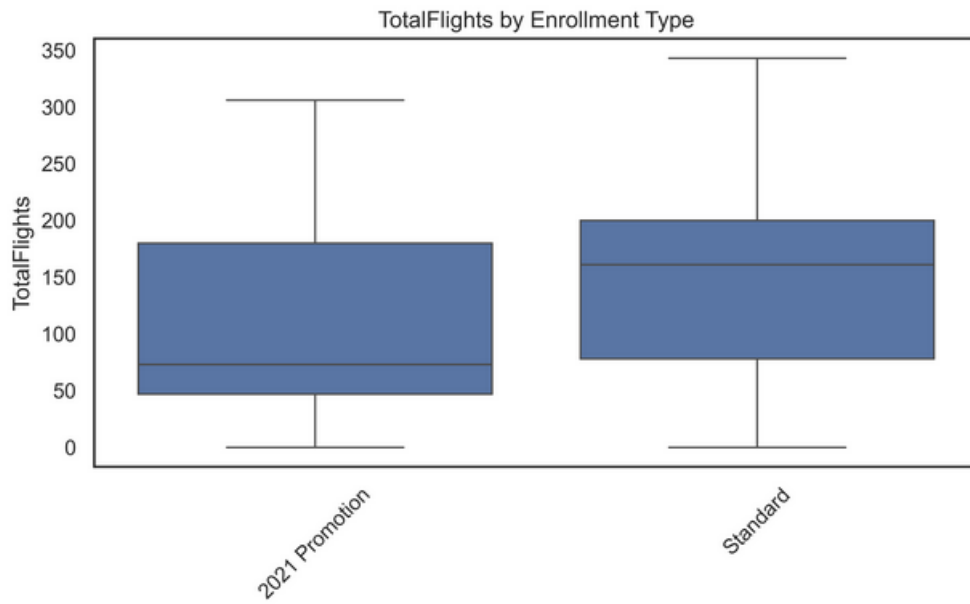


Figure 13 - Total flights by Enrollment Tyoe

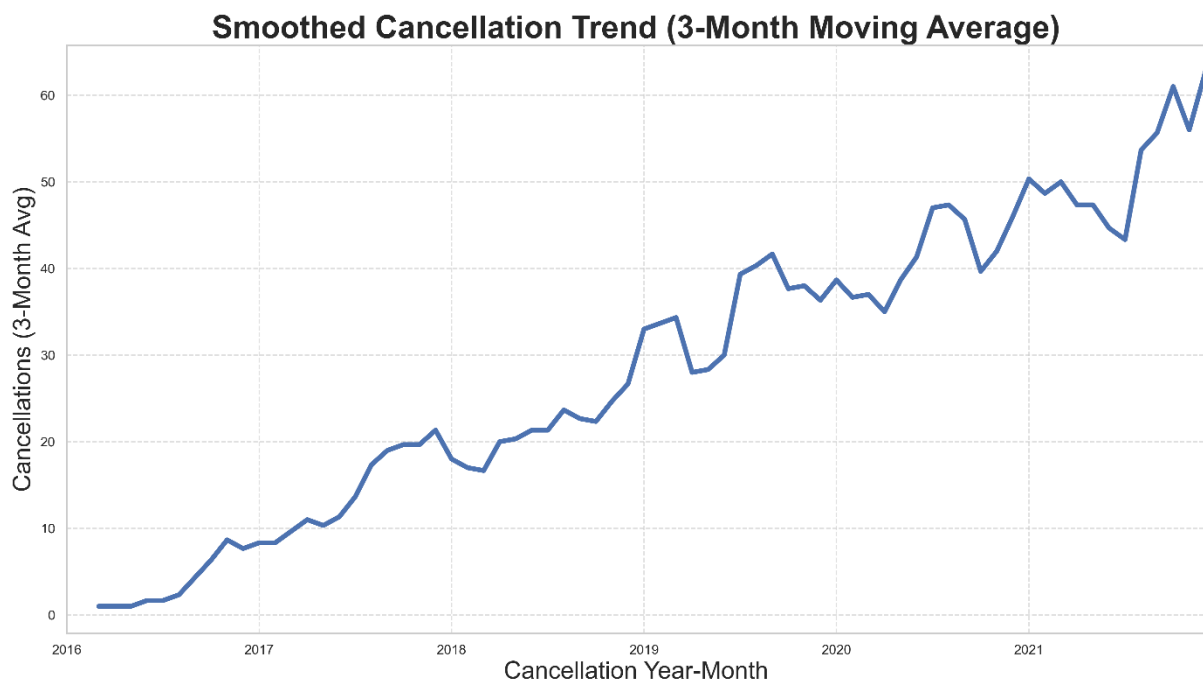


Figure 14 - Smoothed Cancellation Trend

Feature Name	Source	Description	Formula / Logic	Reason
IsActive	CustomerDB	Indicates whether the customer's membership is still active.	<code>True</code> if 'CancellationDate' is <code>NaN</code> ; <code>False</code> otherwise.	To identify currently active members vs churned ones.
Tenure	CustomerDB	Duration (in years) since the customer joined the loyalty program until the cancellation date, or 31/12/2021 if still active.	<code>(CancellationDate.fillna(2021-12-31)-EnrollmentDateOpening).days/365</code>	To estimate customer loyalty length and maturity.
TotalFlights	FlightsDB (aggregated)	Total number of flights taken by the customer over the entire period.	<code>sum(NumFlights)</code>	To measure engagement and travel frequency.
TotalFlightsWithCompanions	FlightsDB (aggregated)	Total number of flights taken with companions.	<code>sum(NumFlightsWithCompanions)</code>	To identify social or family-oriented travelers.
TotalPointsAccumulated	FlightsDB (aggregated)	Total loyalty points earned by the customer.	<code>sum(PointsAccumulated)</code>	To measure customer earning behavior and activity level.
TotalPointsRedeemed	FlightsDB (aggregated)	Total loyalty points redeemed or spent by the customer.	<code>sum(PointsRedeemed)</code>	To evaluate how customers use their earned benefits.
RedeemRatio	FlightsDB (derived)	Efficiency of point redemption relative to points accumulated.	<code>TotalPointsRedeemed / TotalPointsAccumulated</code> (replace inf with NaN)	To measure loyalty engagement and reward utilization.
Flights_January ... Flights_December	FlightsDB (pivot)	Total number of flights per month; each column represents a month (1-12).	<code>groupby(['Loyalty#', 'Month'])['NumFlights'].sum().unstack(fill_value=0)</code>	To detect seasonal or temporal travel patterns.

Figure 15 - New Features Summary

ANNEXES

ANNEX 1 (AI USAGE STATEMENT)

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

Specifically, ChatGPT contributed to:

- Building the interactive map that plots the relationships of Latitude - Longitude vs City vs Province or State, as this involved advanced programming;
- Refining data transformation logic involving date operations (e.g., Tenure computation with conditional CancellationDate handling) and aggregations across multi-period flight activity;
- Supporting the interpretation of distributional asymmetries, especially logarithmic and right-skewed variables (e.g., Income and Customer Lifetime Value), and their implications for segmentation readiness;
- Assisting in structuring the CRISP-DM narrative, connecting exploratory insights to business relevance for subsequent clustering phases.

All statistical reasoning, interpretive insights, and business conclusions remain the group's original analytical work.

Any computational assistance received was auxiliary to substantive analytical judgment and is acknowledged herein.

ANNEX 2 (CONTRIBUTION STATEMENT)

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

Pedro Santos (Student ID: 20250399)

- Merge Datasets
- Business Report (Extra)
- Report
- Visualizations

Miguel Correia (Student ID: 20250381)

- Visualizations
- Data Description

Pedro Fernandes (Student ID: 20250418)

- Preprocessing
- Poster
- Visualizations

Tiago Duarte (Student ID: 20250360)

- Data Quality
- Visualizations
- Geo-Spatial Insights (Extra)

All members contributed to collaborative discussions and ideation.

ANNEX 3 (RESPONSIBILITY STATEMENT)

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