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, 20001111
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 2019–2021, we conduct Exploratory Data Analysis (EDA) to assess data quality, understand behavioural patterns, and prepare a modelling-ready dataset. We diagnose missingness, duplicates, and logical inconsistencies; engineer interpretable features (e.g., activity, redemption behaviour, companionship, seasonality); aggregate monthly flight records to one row per customer; and merge the flight summary with the customer table.

The analysis reveals heavy-tailed engagement where a minority of members drive a disproportionate share of value. Key discoveries include the identification of a precise points accrual formula (PointsAccumulated = 10% × DistanceKM), a substantial zero-income cohort representing 25% of customers, and weak correlations between demographic attributes and Customer Lifetime Value. These insights motivate specific preprocessing choices (transformations, scaling, outlier handling, encoding) 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** by aggregating monthly flight records to one row per loyalty member and merging the flight summary with the customer table
- **Assess data quality** (missingness, duplicates, logical inconsistencies, type normalisation) and document mitigations that preserve analytical integrity
- **Choose interpretable features** that capture engagement volume (flights, distance, points), redemption behaviour, seasonality, companionship (social travel), and membership context (tenure/churn proxies)
- **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 first stage of the work applies Exploratory Data Analysis (EDA) to understand the datasets, assess their quality, and surface patterns that inform feature design and preparation choices for clustering. By following 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 Data Understanding

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 Preparation

We performed a structured 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

EnrollmentDateOpening and CancellationDate were converted from text to datetime. The Country column was identified as containing only "Canada" across all records, making it irrelevant for segmentation and warranting removal during feature engineering.

Flights Dataset

YearMonthDate was converted from text to datetime. NumFlights and NumFlightsWithCompanions were stored as floats with many 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.

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 completenes.

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, 164 instances of duplicated Loyalty# existed with differing attributes (different Last Names, cities, and other demographics), indicating data entry errors rather than legitimate related accounts. To preserve merge integrity with FlightsDB, we retained the first occurrence per Loyalty# and dropped the remainder, reducing the customer table to 16,757 unique members.

Flights Dataset

We removed 2,903 exact duplicate rows. These duplicates represent complete redundancy where all field values were identical, requiring straightforward deletion to prevent overcounting activity metrics.

2.2.4 Logical Inconsistencies

Customers Dataset

The 164 duplicated Loyalty# entries with conflicting customer information (different names, cities) represent data quality issues that were resolved by keeping the first occurrence. Additionally, we noted that at least 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.

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.

Important Discovery: points Accrual Mechanism: Analysis revealed a systematic relationship between distance and points: **PointsAccumulated = 10% of DistanceKM**. Specifically, we verified that all rows with DistanceKM > 0 have corresponding PointsAccumulated values following this precise formula (rounded to integer points). This discovery confirms the airline's consistent accrual mechanism across all flights, validates data consistency and internal logic, will inform feature engineering decisions (avoiding redundant features), and enables accurate forecasting of points liability for the loyalty program.

2.3 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.

2.3.1 Flight-level (monthly) features

Table 1: Engineered Features at Monthly Flight Level

Feature	**Description**
ActiveMonth	Binary flag (1 if PointsAccumulated > 0; else 0)
RedeemRate	PointsRedeemed / PointsAccumulated (0 when denominator is 0)
AvgDistancePerFlight	DistanceKM / NumFlights (0 when NumFlights = 0)
CompanionshipRatio	NumFlightsWithCompanions / NumFlights (0 when NumFlights = 0)
MonthSeason	Winter, Spring, Summer, Autumn (from Month)

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

Monthly records were consolidated to a single customer-grain table (flights_agg; 16,737 rows) using groupwise totals, means, and behaviour flags:

- **Volume and distance:** TotalFlights, AvgFlightsPerMonth, TotalDistanceKM, AvgDistanceKM
- **Points:** TotalPointsAccumulated, AvgPointsAccumulated, TotalPointsRedeemed, AvgPointsRedeemed, TotalDollarCostRedeemed, AvgDollarCostRedeemed, AvgRedeemRate
- **Travel style:** AvgDistancePerFlight, AvgCompanionshipRatio
- **Temporal activity:** monthsActive_2019, monthsActive_2020, monthsActive_2021
- **Seasonality:** Flights_Spring, Flights_Summer, Flights_Autumn, Flights_Winter; PeakSeason; PeakMonth; MonthMatchesSeason
- **Yearly participation:** Flew_2019, Flew_2020, Flew_2021; YearsActive (0–3)

Quality consideration: Given the discovered Points = 10% × DistanceKM relationship, we will avoid including both PointsAccumulated and DistanceKM as separate features in clustering to prevent multicollinearity.

2.3.3 Customer-level features (CustomerDB)

Note: Country column will be removed as it contains only ”Canada” and provides no segmentation value.

Table 2: Engineered Features for Customer Dataset

Feature	Description
Churn	1 if CancellationDate present (churned), 0 if missing (active)
IncomeGroup	Income tertiles (Low/Medium/High) or separate Zero group
MaritalGroup	Married vs. Not Married (Single/Divorced)
EnrollmentYear	Year extracted from EnrollmentDateOpening
EducationLevelCode	Ordinal encoding of education level
ProvinceActivity	Customer density in province
PostalArea	First three characters of postal code
PostalAreaActivity	Customer density in postal area
LoyaltyScore	Ordinal mapping of LoyaltyStatus
ValueToIncomeRatio	CLV / (Income + 1) to handle zero-income cases
MarriedHighIncome	1 if Married and in High income group; else 0

2.3.4 Merged dataset construction

We left-joined the aggregated flight summary (flights_agg; 16,737 unique Loyalty#) to the deduplicated customer table (16,757 unique Loyalty#), producing a modelling-ready, customer-grain dataset (16,757 rows). Approximately 20 customers have no flight summary (NA across flight-derived columns), which is expected when no valid flight history exists in 2019–2021 or rows were filtered by quality rules. Uniqueness of Loyalty# is preserved in the merged table.

2.4 Exploratory Data Analysis

We used visual and summary analyses to characterise distributions, relationships, and seasonality across both the customer and aggregated flight datasets.

- **Distribution Analysis:** Histograms and boxplots for Income, CLV, TotalFlights, DistanceKM, and points metrics reveal heavy right tails—a minority of members drive disproportionate value and activity.

- **Relationship and Correlation Analysis:** Scatter plots, count plots, and heatmaps examine dependencies among demographic, behavioural, and points variables. Activity measures (flights, distance, points accrued) are strongly related; redemption measures align with dollar-cost redemption; travel style (companionship) and redeem rate add complementary, less-correlated signals. Seasonal patterns are evident (e.g., Summer peaks).

Key findings from EDA are summarised in the Results section and directly inform feature selection, transformations (e.g., log1p for heavy-tailed variables), scaling, outlier treatment, and encoding strategies for Phase 2 clustering.

3 Results

3.1 Distribution Analysis

3.1.1 Customers Dataset

The customer dataset contains demographic and loyalty information such as Income, Customer Lifetime Value (CLV), EnrollmentDateOpening, and Marital Status. Analysis shows that Income and CLV are heavily right-skewed: most customers have modest values while a small minority exhibit very high values. Specifically, at least 25% of customers have Income = 0, suggesting they may be unemployed, students, or chose not to disclose income information. Income shows a mean of approximately 70,000 with a maximum reaching 158,932. CLV shows a median around 8,000 with outliers above 40,000. This indicates that a limited group of high-value members contributes a disproportionate share of overall value.

For categorical variables such as Gender, Education, and Marital Status, distributions are uneven but interpretable. Gender shows a well-balanced distribution between Female and Male customers (approximately 50-50). A large percentage hold Bachelor degrees, suggesting an educated customer base. More than half of customers are married, which may correlate with companion travel patterns. LoyaltyStatus reveals almost half are classified as ”Star” status, with fewer in Nova/Aurora tiers. This concentration suggests the Star tier may be too broad and could benefit from sub-segmentation. Geographic analysis confirms all customers are in Canada, making the Country column irrelevant for segmentation.

3.1.2 Flights Dataset

The flights dataset captures operational activity (NumFlights, DistanceKM, PointsAccumulated, DollarCostPointsRedeemed) across 2019–2021. Most numeric variables are positively skewed: the majority of members flew infrequently and over shorter distances, while a smaller subset is highly active and long-haul oriented. NumFlights typically ranges from 0-5 per month, with some frequent flyers reaching 20+ flights. Points metrics show a similar pattern, with many months of zero redemption and a smaller group with high redemption values, indicating distinct ”saver” vs. ”spender” behaviours. Outliers are present, especially in DistanceKM and redemption-related measures, requiring careful treatment in preprocessing.

3.2 Relationships Between Variables

3.2.1 Customers Dataset

We examined pairwise relationships and grouped comparisons to understand how customer characteristics relate to value and loyalty behaviour. Consistent with the scatter and correlation analysis, the linear association between Income and CLV is weak (correlation coefficient approximately 0.3), indicating that higher income does not strongly predict higher lifetime value. The 25% zero-income cohort requires special consideration in segmentation and may represent a distinct behavioural group.

Differences in CLV across Gender or Marital Status are limited based on grouped analysis. Education shows some relationship with Income but not directly with CLV. Urban dwellers (based on Province and PostalCode analysis) show slightly higher activity levels and potentially higher CLV. Overall, demographic attributes alone are not strong determinants of member value, reinforcing the need to incorporate behavioural signals from flight activity for effective segmentation.

3.2.2 Flights Dataset

Operational variables show expected dependencies. NumFlights and DistanceKM are strongly and positively related (as expected—more flights typically mean more total distance). PointsAccumulated follows the precise formula discovered: Points = 10% of DistanceKM. This systematic relationship was validated across all non-zero distance records, confirming the airline's consistent accrual mechanism and creating very high correlation (r $\approx$ 1.0) between these variables. PointsRedeemed is strongly associated with DollarCostPointsRedeemed (r $\approx$ 0.95), confirming internal consistency of redemption values and suggesting we can use one as a proxy for the other.

Data quality checks revealed 5,901 rows (0.97%) where NumFlights = 0 but DistanceKM > 0; these were flagged for correction or exclusion. Beyond core activity measures, features such as NumFlightsWithCompanions, AvgCompanionshipRatio, and AvgRedeemRate exhibit weaker correlations with primary volume metrics, suggesting they capture complementary behavioural dimensions useful for segmentation beyond just flight frequency.

3.3 Correlation Analysis

We used correlation matrices to identify redundancy and multicollinearity prior to clustering and to validate earlier relationship findings.

3.3.1 Customers Dataset

Correlations among customer features are generally modest, indicating most demographic variables provide independent information. Income and CLV exhibit weak linear correlation (r $\sim$ 0.3), supporting separate inclusion of both variables. Demographic variables (Education, Marital Status) show low association with CLV. Latitude and Longitude show geographic patterns but may be redundant with Province/City categoricals. No pairs appear highly redundant (except Latitude/Longitude vs. Province), so most customer-side variables can be retained with appropriate scaling/encoding for integration with behavioural features.

3.3.2 Flights Dataset

The strongest correlations appear among activity and accrual measures. NumFlights, DistanceKM, and PointsAccumulated form a highly correlated triad. DistanceKM and PointsAccumulated show near-perfect correlation (r $\approx$ 1.0) due to the 10% formula discovered during data preparation. This suggests we should include only one of DistanceKM or PointsAccumulated in clustering models to avoid multicollinearity. Redemption value in points aligns closely with its dollar cost (r $\approx$ 0.95), suggesting one can serve as a proxy for the other.

By contrast, AvgCompanionshipRatio, seasonality-derived indicators, and AvgRedeemRate correlate weakly with the core volume metrics, indicating that they add orthogonal information on travel style and redemption behaviour. At the aggregated customer level, totals and their corresponding averages (e.g., TotalPointsAccumulated vs. AvgPointsAccumulated; TotalDistanceKM vs. AvgDistanceKM) are highly collinear, as expected from construction. This supports selecting one representative from each highly coupled pair for clustering to reduce redundancy.

Overall, the exploratory analysis revealed clear behavioural patterns and heavy-tailed engagement among AIAI's loyalty members. The results validate the cleanliness of the consolidated, customer-grain dataset, highlight meaningful yet partially independent behavioural signals (volume, redemption, travel style, seasonality), and directly inform the feature selection and preprocessing strategy (transformations, scaling, outlier handling, and logical constraints) for the forthcoming clustering phase.

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.

Key Findings:

- A small group of frequent flyers takes most trips and earns most points.
- **Points = 10% of distance flown** — a simple, perfect rule.
- Income doesn't predict value well. Behavior matters more.
- **1 in 4 customers reports $0 income** — likely students or non-disclosers.

We fixed:

- 20 fake sign-ups (same-day cancel, no flights) → removed
- 164 mixed-up customer IDs → kept one
- 2,903 repeated flight records → deleted
- 5,901 impossible entries (distance but zero flights) → flagged
- Rounded weird numbers like "9.9 flights" to whole numbers

New features show:

- How often and far people fly
- Who redeems points (savers vs. spenders)
- Who travels with others
- Peak travel seasons
- Who's active each year

Next Step:

1. Clean skewed numbers and encode categories
2. Group customers with clustering (aim for 4–6 segments)
3. Name each group (e.g., elite flyers, point savers, family travelers, students) and suggest real actions

This gives AIAI a **solid, trustworthy dataset** and a clear plan to better understand and reward customers — using real behavior, not just guesses.

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: Data Overview

CustomerDB Structure (After Cleaning)

FlightsDB Structure (After Cleaning)

Appendix B: Key Discoveries

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