DATA MINING PROJECT

Master in Data Science and Advanced Analytics

NOVA Information Management School

Universidade Nova de Lisboa

Loyalty Segmentation

Group 80

Felix Diederichs, 20250489

Paul Harnos, 20250487

Tjark Bunjes, 20250505

Fall Semester 2025-2026

TABLE OF CONTENTS

ABSTRACT

Provide a concise summary of your objectives, methodology, and key findings. This should give readers an immediate understanding of your study's scope and contributions.

1. INTRODUCTION

"Loyalty is about more than a program, a department, or a tangible redemption offer."

The international strategy consultancy McKinsey emphasized with this statement that the airline industry is undergoing a structural shift in which traditional loyalty programs no longer secure genuine customer allegiance. Once powerful drivers of repeat purchases through miles, status tiers, and redemption perks, these programs have evolved into profit centers built on B2B mile sales and co-branded credit cards. Yet post-pandemic devaluations, benefit cuts, and complex rules have eroded trust and perceived value. Travelers increasingly diversify their loyalty, joining multiple programs and distributing spending across carriers, showing attachment to the brand experience rather than to program mechanics (McKinsey, 2023). Loyalty economics, once central to airline strategy, are losing behavioral influence as travelers now expect seamless, personalized experiences instead of transactional rewards. Emotional factors such as recognition and care drive loyalty more than point accumulation, while points inflation and overcrowded tiers reduce perceived fairness and satisfaction (McKinsey, 2023). In response, airlines rely on data-driven segmentation to tailor engagement. Combining demographic, psychographic, and behavioral perspectives, particularly the latter two, provides deeper explanations of loyalty (Dolnicar, Grün & Leisch, 2018) and enables personalized strategies that restore the emotional foundation of loyalty (Forbes Advisor, 2025).

Against this backdrop, AIAI Airlines has launched a segmentation initiative to regain customer loyalty through greater relevance, recognition, and personalized experience design. This segmentation initiative originates from a pedagogical impetus embedded in data-centric inquiry. Guided by the CRISP-DM framework, the project translates business objectives into a structured data mining plan that links Business Understanding with Data Understanding and iterative refinement (Wirth & Hipp, 2000). Our team's task is to derive an interpretable segmentation of AIAI's loyalty customers to support differentiated strategies at both cluster and individual levels. Defined as an unsupervised grouping problem, the analysis seeks to move beyond generic point-based incentives toward personalized offers and communications that reflect customers' diverse motivations and travel behaviors.

Our analytical objectives encompass three complementary perspectives: value-based segmentation, distinguishing customers by their economic contribution and cost-to-serve; behavioral segmentation, capturing travel patterns such as flight frequency, distance, and redemption behavior; and demographic segmentation, applied for accessibility while acknowledging its limited explanatory power without behavioral context (Dolnicar, Grün & Leisch, 2018). These perspectives are integrated into a unified, actionable framework that guides AIAI's strategy across proposition, pricing, and service design.

Following the CRISP-DM process, we define business success criteria such as higher share of wallet and loyalty satisfaction, focusing on Data Understanding to guide the next phase. Using AIAI's three-year loyalty and flight data, we assess data quality, variable relationships, and behavioral patterns that

suggest preliminary customer groups. These exploratory insights support a targeted Data Preparation plan covering selective cleaning, feature construction, and scaling where needed for Phase 2, when clustering will formally identify and validate these emerging segments.

2. DATA ANALYSIS

This section presents a systematic exploration of AIAI's loyalty dataset to surface patterns and constraints that will guide subsequent preparation and clustering.

2.1. Dataset Overview

The provided Data consists of two core tables covering a three-year horizon (2019-2021). CustomerDB contains 16,921 loyalty members with 21 attributes, including important features like Loyalty#, Age, Income, LoyaltyStatus, Customer_Lifetime_Value (CLV), and EnrollmentDate. FlightsDB contains 608,436 monthly activity records with 10 attributes, including Loyalty#, NumFlights, DistanceKM, PointsAccumulated, and PointsRedeemed. The observation units are customers in CustomerDB and monthly customer-flight transactions in FlightsDB; the relationship is one-to-many through the key Loyalty#.

2.2. Univariate Analysis

The demographic profile of the customer base is evenly split by gender and geographically concentrated in the States Ontario and British Columbia, Canada (over 50%), with an almost equal distribution of urban, suburban, and rural areas. Marital status skews toward married members (58%), followed by singles (27%) and divorced individuals (15%). Income is heavily zero-inflated, with a right-skewed distribution: most nonzero values cluster near 25,000 USD and taper toward 100,000 USD. Education levels are dominated by bachelor's degrees (63 %), followed by college (25%), with minor shares for high school and master's/doctorate. Loyalty tiers exhibit a clear hierarchy - Star (46%), Nova (34%), Aurora (20%) while the CLV is highly right-skewed, showing many low-value members and few high-value outliers. Enrollment surges in early 2021 align with a 7 % promotional cohort, while cancellations show upward spikes in early 2019, mid-2020, and late 2021. Flight activity has slightly increased over the years, showing clear intra-year seasonality with peaks primarily during the summer months. Excluding zero-flight months, most members complete one to 17 flights per month, with strongly right-skewed distributions for NumFlights and DistanceKM (650 - 42,000 km per month). Companion travel is rare. Points variables (Accumulated, Redeemed, DollarCostRedeemed) share similar skewness, with 50 - 60% of all monthly records containing zeros, indicating inactive periods rather than missing data. Early segmentation cues emerge: multimodal latitude/longitude clusters around Toronto, Vancouver, and Montreal; income peaks around 25k suggest distinct socioeconomic groups; CLV outliers mark a high-value niche; and flight-frequency separation points to frequent versus occasional travelers.

2.3. Data Quality Assessment

A structured data-quality evaluation was conducted following the CRISP-DM framework to assess completeness, consistency, and plausibility before progressing to feature engineering. This step aimed to detect missingness patterns, duplicate records, and anomalies that could affect clustering

performance. In the CustomerDB, missing values occurred only in Income, CLV, and CancellationDate. A perfect correlation (r = 1) was found between missingness in Income and CLV, suggesting both originate from the same underlying process. Since CancellationDate is missing only for active customers, this is considered intentional rather than a quality issue. FlightsDB showed no missing values across its ten variables, indicating full transactional coverage from 2019 to 2021. Duplicate records were present in both databases. Because Loyalty# serves as the unique customer ID and the relational key between the two datasets, duplicates compromised the one-to-many structure required for linking customers to their flights. All duplicate Loyalty# entries were therefore removed to maintain relational integrity. Additionally, the redundant "Unnamed: 0" column in CustomerDB was dropped as it provided no analytical value. Another inconsistency was found in the EnrollmentType column: several customers labeled as part of the "2021 Promotion" cohort had enrollment dates before 2021, representing clear outliers.

Several anomalies were also identified that could bias clustering or distort similarity metrics. Customers with zero income but positive CLV, all holding a college education, likely reflect non-income funding such as parental support or measurement inconsistencies. These cases were retained but flagged for monitoring. Some records reported positive DistanceKM and PointsAccumulated despite zero flights, which were corrected by setting both values to zero. A few customers showed average flight distances exceeding 14.000 kilometers, an unrealistic upper limit; these entries were set to missing to prevent distortion. Moreover, flight metrics from 2019, including NumFlights and related KPIs, contained decimal values, likely due to aggregation across reporting periods. As confirmation from the data provider was unavailable, these records were retained but noted for sensitivity checks in later analyses. Finally, 1.234 customers had positive CLV despite no recorded flights between 2019 and 2021. Their enrollment dates ranged from 2015 to 2021, suggesting that CLTV may capture non-flight revenue, pre-2019 transactions, or reporting delays.

3. RESULTS

Interpret your exploratory findings to justify specific feature engineering choices for clustering. Explain which patterns, correlations, and data characteristics lead to preprocessing decisions. Connect data insights, demonstrating how analysis findings inform feature selection, transformation, and scaling approaches.

3.1. Feature Engineering

This section highlights only the most important engineered features; a complete overview is provided in the detailed report. The feature set builds on RFM logic and behavioral segmentation to capture customer activity, loyalty engagement, and economic value.

Key variables include DaysSinceLastFlight, is_active_12m, and avg_flights_per_year, which measure recency and frequency of engagement, distinguishing active from inactive members. Loyalty lifecycle features such as is_current_loyalty_member and is_previous_member classify customers along retention stages. Behavioral variables like companion_flight_ratio, avg_distance, and season, differentiate business versus leisure travels and reveal temporal preferences.

Because the original Customer Lifetime Value proved unreliable, the newly engineered Monetary_Value_2021 serves as the central monetary proxy. It combines annual flight distance (60 %), income (20 %), and historical CLV (20 %) using robust MinMax scaling to approximate the customer's economic contribution in 2021. This composite feature is essential for evaluating the value of active customers and forms the analytical basis for subsequent clustering.

3.2. Bivariate Analysis

This chapter also summarizes only the central findings; all detailed statistical outputs are documented in the full report. The bivariate analysis examined pairwise correlations and groupwise comparisons to identify relationships that inform clustering and feature selection.

Within the FlightsDB, DistanceKM and NumFlights show a strong positive correlation, while PointsAccumulated is almost perfectly aligned with distance, indicating functional redundancy. PointsRedeemed correlates only moderately with accrual, separating active redeemers from passive collectors. A clear negative correlation between avg_distance and NumFlights (≈ -0.6) reveals that frequent flyers tend to travel shorter distances, consistent with commuter or business traveler patterns. NumFlightsWithCompanions correlates moderately with NumFlights (≈ 0.6), suggesting that leisure or family travelers typically take more total flights. Across customer variables, Income and CLV are moderately correlated overall but weaken once zero-income cases are excluded, suggesting heterogeneous value generation beyond earnings. Education and marital status are related: customers with college or bachelor's degrees are more often single, while married individuals have the highest mean income (≈ 43.8 k) and CLV (≈ 8.1 k), followed by divorced and then single customers. Loyalty tiers display consistent value differences - Aurora members show the highest income and CLV, followed by Nova and Star tiers - confirming that higher loyalty status aligns with greater economic contribution and travel engagement. No material correlations were observed between unrelated categorical variables, confirming structural independence across most demographic attributes. However, enrolled members fly significantly more often than non-members, underlining the behavioral impact of program participation. The 2021 promotion further reinforces this: enrollment volumes and flight activity increased markedly during the campaign period, validating its effectiveness and highlighting the strategic potential of targeted acquisition initiatives.

Together, these relationships reveal four clear customer tendencies that guide the upcoming clustering phase: (1) high-frequency short-haul commuters, (2) low-frequency long-haul travelers, (3) leisure and family travelers with companion patterns, and (4) inactive or pre-promotion members.

4. CONCLUSION

4.1. Business Strategy

Retaining active loyalty members is considerably easier and more profitable than acquiring new customers or reactivating those who have left the program. Consequently, AIAI's strategic focus should lie in strengthening engagement among existing active members while progressively re-engaging inactive and ex-loyalty customers. Customers are managed along the dimensions of Loyalty Status x Activity x Value x Behavior to systematically increase retention, reactivation, and conversion.

The four meta- and sub-segments provide the structural foundation for this approach. Active loyalty members should be kept engaged through personalized offers, recognition, and exclusive benefits designed to maintain satisfaction and reduce churn risk. Inactive loyalty members can be reactivated by targeted campaigns emphasizing lost tier privileges, bonus point opportunities, or tailored incentives that restore engagement. Active ex-loyalty members, who continue to travel without program participation, should be encouraged to rejoin by highlighting clear incremental value such as fast-track status recovery or reward matching. Finally, inactive ex-loyalty members require a combined reactivation and reacquisition strategy based on personalized communication, behavioral profiling, and incentive alignment.

To implement these strategies effectively, AIAI must identify behavioral subgroups-such as frequent short-haul versus long-haul travelers or business versus leisure flyers, as well as demographic segments defined by income, education, or geography. These distinctions will help uncover the drivers of customer value and enable differentiated engagement strategies tailored to the needs and potential of each group. By linking loyalty behavior with underlying motivations and socioeconomic characteristics, AIAI can direct resources toward initiatives that maximize long-term value creation and strengthen overall program performance.

4.2. Multivariate Analysis and Next Steps

The exploratory analysis has already revealed clear clustering tendencies even before applying formal algorithms. Active-loyal members demonstrate the highest flight frequency and monetary value, confirming their role as AIAI's core high-value segment. Active non-loyal customers show moderate flight activity but much lower monetary contribution, suggesting strong potential for conversion into loyalty members. Geographic patterns reinforce this insight, as customers from Peace River and Calgary exhibit above-average travel frequency and spending, marking these cities as important regional growth hubs. In addition, bachelor-educated travelers emerge as the most valuable educational group, consistently outperforming others in both flight activity and value contribution. Customers who joined during the 2021 promotion also show higher monetary performance, highlighting the campaign's success in attracting quality members.

In the next phase, AIAI will implement comprehensive preprocessing to ensure analytical rigor and model readiness. This includes outlier treatment, missing-value handling, and variable standardization-essential steps for distance-based clustering algorithms such as k-means or Gaussian mixtures. Once data integrity and comparability are established, the analysis will shift from feature-level groupings to customer-level multivariate clustering. This transition will enable a deeper understanding of customer diversity within and across segments, helping to uncover hidden behavioral or value-driven patterns.

The overarching goal remains to identify and characterize high-value customers, understand the behavioral and demographic mechanisms that differentiate them, and apply these insights to design personalized engagement, retention, and loyalty strategies that sustain long-term business growth.

BIBLIOGRAPHICAL REFERENCES

Forbes Advisor. (2025). Customer segmentation: the ultimate guide. June 2024. URL: https://www.forbes.com/advisor/business/customer-segmentation/.

McKinsey & Company. (2023). Travel invented loyalty as we know it. Now it's time for reinvention. https://www.mckinsey.com/industries/travel/our-insights/travel-invented-loyalty-as-we-know-it-now-its-time-for-reinvention#/

Rüdiger Wirth and Jochen Hipp. (2020). "CRISP-DM: Towards a standard process model for data mining". In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Vol. 1. Manchester. 2000, pp. 29–39. URL: http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.

Sara Dolnicar, Bettina Grün, and Friedrich Leisch. (2018). Market Segmentation Analysis. Jan. 2018. DOI: 10.1007/978-981-10-8818-6. URL: https://doi.org/10.1007/978-981-10-8818-6.

ANNEXES

AI Usage Statement

AI tools were used for coding syntax assistance (ChatGPT and Github Copilot for for pandas operations and very complex visualizations, report proofreading and Debugging (ChatGPT for grammar checks). All analytical insights, business interpretations, and strategic recommendations represent original group analysis and thinking.

Contribution Statement

Tjark Bunjes (Student ID: 20250505)
- Introduction section, poster design

Paul Harnos (Student ID: 20250487)
- Feature engineering ideation, RFM calculations

Felix Diederichs (Student ID: 20250489)
- Multivariate Exploration

All members contributed to collaborative discussions and ideation and to everything else what is not listed above

Annex Responsibility Statement

We, the group members listed above, certify that this report represents our original analytical work and interpretations. While AI tools were used as specified 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.

Example of an unnumbered list:

- Item 1
- Item 2
- Item 3

Example of a numbered list:

1. Item 1
2. Item 2
3. Item 3

Table 0.1 – Illustrative table

Title	Title
Text	Number
Text	Number
Text	Number

Infos about Structure and Formatting etc

The report should be written in English.

In the case references point to the “source code” make sure the code is correctly annotated.

All bibliographic references should be presented in the APA standard; this standard also applies the formatting of references and respective forms of referencing throughout the text[1].

Text highlighted in yellow like this are for your information, and not part of the formatting. Make sure to remove them from your report.

The following instructions should be followed for the writing of the report:

This word file provides a "standard" report structure. The format of the report (headers, spacings, fonts, and other formations) should be employed as defined in this template. However, students may change the structure and the titles according to their n