



Data Mining

Customers Segmentation in Airlines Industry:

Data Mining Approach

Filipe Esteves 20250470

Mariana Freitas 20250369

Miguel Caramelo 20250378

Miguel Silva 20250379

Abstract

This project is set to build a customer segmentation, through data analysis and exploration, conducted to uncover patterns, trends and relationships within the given data. Results demonstrate space for improvement for the company's structure and approach, as well as future research refinement.

Introduction

In today's business environment, analysing data is crucial in order to achieve success and improved outcomes, specially when dealing with a very competitive industry.

This project focuses on the company Amazing International Airlines Inc. (AIAI) seeking personalized services and marketing strategies for their customers. Our objective is to group customers according to their economic contribution through a Value-Based Segmentation, analysing purchasing habits and travel behaviours through a Behavioural Segmentation, and finally, categorizing customers by age, occupation, or any other relevant attributes and patterns.

Importations

To execute this project, we needed to integrate some libraries. The *sqlite3* and *os* libraries are meant to serve as a lightweight data storage, helping us import the file where our dataset relies. We included *pandas* for efficient data manipulation and analysis, *numpy* that provides us numerical operations to handle arrays, the *random* library that allows us to introduce randomness to enhance the diversity of our data analysis. *Seaborn*, *matplotlib* and *math* are crucial when it comes to visualization of the data at hand, as well as the trends within our datasets. The *datetime* library was used to convert the variable 'year' into a object with the datetime format. *Sklearn* and *scipy* will co-operate with *numpy* and *pandas* for data analysis and modeling.

Data Exploration

- **Duplicates and Missing Data**

First step was to visualize the data at hand in order to have a general idea of: what information we had to analyze, initial takeaways from the data as well as what could be wrong within the data itself. Through the use of the features *head()*, *tail()*, *shape*, *info()* and *describe*, we were able to know what would be the next steps.

Initial conclusion is the inexistence of duplicated values and missing values on the flights dataset, however, within the customer dataset there's also no duplicated values, 20 missing values from the variable 'income', 20 missing

values from the variable 'Customer Life Value', in addition to 14 611 missing values from the variable 'CancellationDate'. This indicates a positive information about the desire of the customer to continue their relationship with the airline.

- **Histograms**

Through the histograms we are able to visualize the data presented and analyze it. Throughout the dataset, there are a lot of values as zero, therefore, to better visualize the data to come to conclusions, we also present each numerical variable with all values above zero.

From the Flights dataframe, we can recognize some behaviours, AIAI customers seem to take less flights as the variable 'NumFlights' is predominant throughout 1 to 11 flights per customer, moreover, these trips are typically made alone as the variable 'NumFlightsWithCompanions' shows very few flights where our customer didn't fly alone, this might suggest these trips are made with a specific purpose, p.e. work purposes.

In terms of how far they travel, our customers don't go far, mostly travelling less than 30000 kilometers accumulated through all flights made by each, along with somewhat lower points, in 'PointsAccumulated' very much correlated with 'DistanceKM', as the former adds up depending on the latter.

'PointRedeemed' and 'DollarCostPointsRedeemed' also show a very high correlation, presenting corresponding behaviours, these variables tell us that only customers with more points accumulated eventually use them, pointing out how the newly arrived customers end up not meeting the perks of being a active AIAI customer.

From the Customers dataframe, the built histograms convey irrelevancy in the variables 'Location Code' and 'Gender', most of our customers have a bachelor degree, are married, are within the *Star* loyalty status and have a standard enrollment type. The latter suggests the 2021 Promotion hasn't reached many customers.

Then, we proceeded with a bivariate Categorical Distribution, in order to get a visual of how two variables intercept with each other, but it mainly supports the suggestions given by the initial histograms.

- **Box Plots**

The Box Plots support the analysis made through the information given by the histograms about the customers' behaviours. The values of all variables are skewed to the left, again showing the minimal relationship the customer has with the company as well as lack of retention proven by the lower points acquired by each client.

There are some outliers in the variables 'NumFlightsWithCompanions', 'PointsRedeemed' and 'DollarsCostPointsRedeemed'. These likely represent AIAI's most profitable customers, the one's who travel more, alone or with companions, in addition to having more points accumulated to which they take advantage of by redeeming them. Thus demonstrating the company's capability of retention and customer satisfaction.

- **Income and Customer Lifetime Value**

Surprisingly, our customers are a great target financially. The 'Income' variable suggest a purchasing power that AIAI can take advantage of, when marketing their products, however, the relationship with each customer doesn't last. The scatter plot that joins these two variables shows exactly AIAI biggest challenge, no matter the annual income, customer lifetime value gets stuck at early stages. Outliers, in this case, are the example to follow.

- **Correlation**

The correlation heatmap was done with the spearman correlation between all numerical variables in flights dataset. We observed some high correlations between 'DistanceKm' and 'NumFlights', 'PointsAccumulated' and 'NumFlights', the same applies to 'NumFlightsWithCompanion' that has a slightly lower but still high correlation with 'DistanceKm' and 'PointsAccumulated'. Looking at filtered dataset (rows only containing positive values in 'NumFlights' column) we observed much less significant values between variables.

Then, we decided to create three new features, 'PointsRedeemRatio' that shows the 'PointsRedeemed' by 'Pointsaccumulated' ratio, then we added the average distance per flight feature and finally a ratio between 'NumFlightsWithCompanions' and 'NumFlights'.

The last heatmap shows the correlation between numeric features including the ones created by us. About the new features added correlations with other variables we can emphasize a slight correlation of 'FlightsWithCompanionRatio' with 'PointsRedeemed' and 'DollarCostPointsRedemmed'.

Optional Bonus Components

1. Geo-Spatial Insights

This visualization was chosen because it facilitates the observation of customer distribution patterns across different regions and highlights variations in Customer Lifetime Value (CLV) through a color-based representation.

It provides an intuitive way to identify geographical clusters of high-value customers, which will be valuable for further segmentation in Phase 2.

The corresponding code for this visualization can be found at the end of the notebook `eda_flights.ipynb`.

2. Interactive EDA Dashboard

The interactive dashboard was developed as an innovative approach to data exploration.

It allows users to dynamically analyze key variables, filter insights, and visualize trends without modifying the underlying code.

The implementation is contained in InteractiveEDADashboard.py, and it runs according to the dependencies listed in requirements.txt.

The link to access Interactive EDA Dashboard is as follow:

<https://interactive-eda-dashboardh3vq5prmkuytzn8vgcg8cp.streamlit.app/>

Future Steps on Data Preprocessing

Our future step for this project is to focus on relevant variables, in order to be able to view the data clearly. For this we plan on dropping the missing values from the variables 'Income' and 'Customer Life Value' as it represents a minimal amount of data in comparison to the whole dataset. We won't be working with the variables 'Location Code', 'Gender', 'First Name' and 'Last Name' as these categorical variables don't seem to add any insights to our analysis.

We are also set to merge both datasets, regarding flights and customers, for future research refinement to proceed with clustering.

Conclusion

Thus, the exploratory analysis establishes a solid foundation for the next step, Data Preparation and Clustering. The insights derived from the EDA highlight that customer retention and engagement are the most relevant factors, therefore customer behaviour and loyalty must be prioritized for the clustered model. Also, the preprocessing strategy will address outliers, missing values, and identification of distinct customer groups that differ in value, loyalty, and regional distribution.

We expect to deliver a functional strategic design for retention and marketing per customer segment, enabling AIAI to perform at its best.

Annex

Artificial Intelligence was used to support and refine our notebooks, specially assisting in the making of the geo spacial and interactive dashboard. AI was also used for debugging code.

- Abstract/Conclusions
- Feature Engineering on Flights dataset
- Geo Spacial/ Data Exploration on Customers dataset

Mariana Freitas 20250369

- Report Analysis/Results Interpretation
- Data Exploration/Correlations on Flights dataset/ Interactive DashBoard Code
- Poster Design

Miguel Caramelo 20250378

- Poster design
- Interactive DashBoard Code, Data Exploration/Correlations on Customers dataset
- Report Analysis/Results Interpretation

Miguel Silva 20250379

- Geo Spacial
- Data Exploration on Customers dataset
- Data Visualization/Feature Engineering

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