Project Documentation: Airline Customer Value Analysis Case

1. Introduction

This document provides documentation for the end-to-end data analysis and clustering project based on the "Airline Customer Value Analysis Case." The project involves tasks such as Exploratory Data Analysis (EDA), Feature Engineering, Clustering, and Interpretation of the clusters. The analysis is conducted using Python in a Jupyter notebook environment.

1.1. Project Goals

The primary goals of this project are as follows:

1. **Exploratory Data Analysis (EDA):** Understand the structure and characteristics of the provided dataset.
2. **Feature Engineering:** Select relevant features for clustering and perform preprocessing as needed.
3. **Clustering:** Apply the K-means clustering algorithm to segment customers into distinct groups.
4. **Interpretation:** Analyze and interpret the characteristics of each cluster to derive actionable insights.
5. **Recommendations:** Provide business recommendations based on the identified customer segments.

2. Technical Details

2.1. Libraries Used

The analysis uses the following Python libraries:

- Pandas
- NumPy
- Matplotlib
- Seaborn
- Scikit-learn

2.2. Data Source

The dataset, named "flight.csv," contains customer information for a airline company. The data includes various attributes such as customer ID, flight details, and demographic information. The data is loaded into a Pandas DataFrame for analysis.

2.3. Data Preprocessing

2.3.1. Date Column Handling

The date columns (`FFP_DATE`, `FIRST_FLIGHT_DATE`, `LOAD_TIME`, `LAST_FLIGHT_DATE`) are parsed into `datetime` format using the `parse_dates` parameter in `pd.read_csv`. Special attention is given to the `last_flight_date` column, where invalid dates are identified and corrected.

2.3.2. Data Cleaning

Duplicate rows are removed, and missing values are handled appropriately. The `member_no` column, if unique for each row, is dropped as it does not contribute to the analysis.

3. Exploratory Data Analysis (EDA)

3.1. Descriptive Statistics

Descriptive statistics are generated for both numerical and categorical columns. The distribution of numerical features is visualized through histograms and box plots. Correlation between variables is explored using a pairplot.

3.2. Missing Values and Duplicates

The presence of missing values and duplicated rows is checked. Missing values are handled, and duplicated rows are removed.

4. Feature Engineering

4.1. Outlier Detection and Handling

Outliers in numerical columns are identified using z-score, and the affected rows are either corrected or removed.

4.2. Feature Selection

Based on the analysis, 3-6 relevant features (`avg_discount`, `points_sum`, `seg_km_sum`, `flight_count`) are selected for clustering.

4.3. Feature Scaling and Transformation

Standard scaling is applied to the selected features. Power transformation is performed to address skewed data.

5. Clustering

5.1. Optimal Number of Clusters

The elbow method is utilized to determine the optimal number of clusters for K-means clustering.

5.2. K-Means Clustering

K-means clustering is applied to the scaled and transformed features. The resulting clusters are visualized using PCA.

6. Interpretation

6.1. Cluster Characteristics

Clusters are interpreted based on the statistical summary of features within each cluster.

6.2. Customer Segmentation

Customers are segmented into four classes:

- **Class 0:** High-value, high-frequency customers.
- **Class 1:** High-value, lower-frequency customers.
- **Class 2:** Low-value, high-frequency customers.
- **Class 3:** Low-value, low-frequency customers.

7. Recommendations

7.1. Business Insights

Based on the cluster analysis, the following recommendations are proposed:

- **Class 0:** Focus on loyalty programs and exclusive perks to maintain customer engagement.
- **Class 1:** Improve marketing strategies or enhance service quality to encourage more frequent transactions.
- **Class 2:** Provide personalized offers to increase transaction value, such as referral programs.
- **Class 3:** Conduct targeted marketing and gather feedback to understand reasons for low engagement.

8. Conclusion

This document summarizes the end-to-end process of analyzing and clustering airline customer data. The insights gained from this analysis can guide business strategies to better cater to the diverse needs of different customer segments.