Technical Report – Travel Tide Customer Segmentation Jupyter Notebook

Introduction

This technical report provides a detailed, chapter-by-chapter explanation of the Jupyter Notebook developed for the Travel Tide Customer Segmentation project. The purpose of this document is to transparently describe the analytical logic, data processing steps, and modeling decisions implemented throughout the notebook. The report is written for a technical audience and serves as formal documentation of the final project.

Notebook preparation: Data Collection

- Connection to TravelTide Server with beekeeper Studio.
- SQL coding: Filtering cohorts after January 4, 2023 for user sessions count greater than 7.
- Export CSV file for Jupyter Notebook.

---

1. EDA & Preprocessing

1.1 Converting Data Formats

The notebook begins with the standardization of data formats, primarily focusing on date and time fields across all datasets. Proper conversion ensures consistency in downstream calculations, enables time-based aggregations, and prevents type-related errors during analysis.

1.2 First Analysis & Data Cleaning

An initial exploratory analysis is conducted to understand distributions, detect anomalies, and assess overall data quality. Summary statistics and visual inspections reveal implausible values and extreme outliers that could distort analytical results.

Findings of the Analysis

Key issues identified include skewed distributions, extreme values in numerical fields, and inconsistencies in trip-related metrics.

1.2.1 Outlier Removal

Extreme observations that clearly represent data errors or non-representative behavior are removed to improve data reliability. This project uses only outlier clipping to preserve information.

1.2.2 Outlier Clipping

For variables where outliers may still carry behavioral meaning, values are clipped to reasonable bounds rather than removed, preserving information while limiting distortion.

1.2.3 Cleaning: Negative Hotel Nights

Trips with negative hotel nights are identified as invalid records and corrected or removed, ensuring logical consistency in travel-related features.

1.3 CSV Export for the Cleaned Data

Cleaned datasets are exported to CSV files, creating clear checkpoints in the workflow and supporting reproducibility.

---

2. Detection of Canceled Trips

This section focuses on identifying and isolating canceled trips. Correctly labeling cancelations prevents canceled activity from being misinterpreted as completed travel behavior. The dataset is split accordingly, and non-canceled trips are exported for subsequent feature engineering.

2.1 CSV Export for Non-Canceled Trips

Only completed trips are retained for behavioral aggregation, ensuring accurate representation of customer travel activity.

---

3. Feature Engineering

Feature engineering transforms raw transactional data into user-level behavioral indicators suitable for segmentation.

3.1 Feature Engineering of df_session

Session data is aggregated to derive engagement metrics such as session frequency and activity intensity.

3.2 Feature Engineering of df_trips

Trip-level data is transformed into indicators reflecting booking behavior, travel frequency, and trip characteristics.

3.3 Feature Engineering of df_user

User-level attributes are cleaned and enriched with derived metrics.

3.3.1 Feature Engineering of df_user_1

Additional transformations consolidate user features into a single analytical table.

---

4. PCA Analysis

4.1 Cleanup of the Working DataFrame df_users

4.1.1 Handling Non-Numerical Values

Non-numeric columns are removed or encoded to prepare the dataset for numerical modeling techniques.

4.1.2 Handling NaN Values

Missing values are handled through imputation or exclusion to ensure model stability.

4.2 Scaling the Data in df_users

All numerical features are scaled to ensure equal contribution in distance-based algorithms.

4.3 Principal Component Analysis (PCA)

PCA is applied to reduce dimensionality and address multicollinearity while retaining most of the variance.

4.4 PCA Results

The resulting components reveal underlying behavioral structure and enable visual inspection of customer distribution.

---

5. Silhouette Score

The Silhouette Score is used to quantitatively evaluate clustering performance across different numbers of clusters. This analysis supports the selection of an optimal cluster count by balancing cohesion and separation.

---

6. K-Means Clustering

Using the selected number of clusters, K-Means clustering is applied to the PCA-transformed data. Each user is assigned to a cluster based on similarity in behavioral patterns.

---

7. Cluster Analysis

7.1 Joining the Final DataFrame df_customer_seg

Cluster labels are merged back with the engineered feature set to enable interpretability.

7.2 Cleanup of the Final DataFrame

Final data cleaning ensures consistency and usability of the segmentation output.

7.3 CSV Export of the Final DataFrame

The complete customer segmentation dataset is exported for downstream use.

7.4 EDA of the Final DataFrame

Exploratory analysis is conducted to compare clusters and identify defining characteristics.

7.5 Manual Evaluation

Clusters are manually reviewed to validate plausibility and business relevance.

---

8. Conclusion

This technical report documents a structured, reproducible analytics pipeline for customer segmentation. By combining rigorous preprocessing, feature engineering, dimensionality reduction, and clustering, the notebook provides a robust foundation for data-driven customer analysis.

Prepared by Frank Pirkl – Data Analytics / Data Science