Travel Tide Segmentation Analysis

Goal:

We wanted to build meaningful customer segments from TravelTide's data so the business can improve loyalty and retention through targeted rewards. We built a pipeline in Colab that prepares, cleans, analyzes, and clusters the data, then exports it for use in BI tools like Tableau.

1) Setup & Project Paths

- First, we connected Google Colab to Google Drive, so all work and files are stored in Drive.
- We created a standard folder structure:
    - data/raw → input files (CSV exports or DB extracts).
    - data/processed → outputs after cleaning and feature engineering.
    - output/models, reports, scripts → for models, reports, and code.
- Installed all the required Python libraries (pandas, SQLAlchemy, scikit-learn, etc.).

2) Data Connection

- The pipeline can take input from two places:
    1. **CSV files** (already exported into data/raw/).
    2. **Postgres database** using a secure connection string.
- A single switch (SOURCE = "csv" or "db") decides which source to use.
- In both cases, the four key tables are loaded: users, sessions, flights, hotels.

3) Data Cleaning

- Converted date/time fields to proper datetime objects (e.g., session start/end, hotel check-in/out).
- Recalculated hotel stay length (nights) when it was missing or invalid, using check-in and check-out dates.
- Converted numeric fields (e.g., base_fare_usd, page_clicks) from text to numbers.
- Added derived fields, such as session_minutes (duration of each session).
- Removed duplicate rows from key tables (users, sessions, flights, hotels).

This step ensured the data is consistent and ready for analysis.

4) Data Analysis (EDA)

- Ran some quick checks to understand the data distribution.
    - Example: Histogram of session page clicks.
    - Example: Histogram of flight fares.
- This step helps spot outliers, errors, or strange patterns before modeling.

5) Feature Engineering (Building User-Level Table)

We transformed the raw data into user-level metrics, combining behavior across sessions, flights, and hotels.

- **Session features (per user):**
    - Total sessions, bookings, cancellations, average clicks, average time spent, last session date.
- **Trip/monetary features (from flights & hotels):**
    - Total spend on flights and hotels, total nights and rooms booked.
- **Behavioral / RFM metrics:**

    - **Recency**: Days since last booking/session.
    - **Frequency**: Number of trips booked.
    - **Monetary**: Total spend.
    - Cancel rate and booking rate.
    - Months since sign-up.
- **Quality filters:**
    - Kept only users with at least 2 sessions and at least 6 months on the platform.
    - Winsorized (trimmed) extreme values at the 99% level to reduce outlier impact.

Final feature set for clustering:
recency_days, frequency_trips, monetary_total, cancel_rate, book_rate, avg_page_clicks, avg_session_minutes, total_nights

6) Segmentation (K-Means)

- Standardized all features (scaled to mean=0, std=1).
- Tested different numbers of clusters (k = 3 to 6).
- Chose the best k using **silhouette score**, a measure of how well clusters are separated.
- Trained final **K-Means** model and assigned each user a cluster label.
- Built quick cluster profiles by averaging features in each cluster (e.g., high spenders, frequent travelers, bargain seekers).

7) Outputs

- Exported results into Drive:
    - customer_segments.csv → the full dataset with clusters.

    - user_features.csv → simplified table with features + cluster (ready for Tableau).

8) Optional (BI / Git)

- Use the processed files in **Tableau** to visualize:
    - Segment sizes, retention rates, churn risk, and potential reward targeting.
- If saving to GitHub, large CSVs should be tracked with **Git LFS** (to handle >100MB files).