Travel-Tide Customer Segmentation & Perks assignment Project

Author: Jawad Mofleh

05.02.2026

EXECUTIVE SUMMARY

This project presents a comprehensive customer segmentation analysis for the Travel-Tide travel booking platform using behavioral, transactional, and demographic data. The goal is to enable personalized perk assignment that improves booking conversion, retention, and customer lifetime value.

Data from sessions, users, flights, and hotels were integrated under strict criteria: only sessions after January 4, 2023, and users with more than seven sessions were included. The final dataset contains 49,211 sessions, 16,099 trips, and 5,442 users.

Thirty-seven user-level features were engineered to describe engagement, travel frequency, spending, seasonality, discount usage, and cancellation behavior. Dimensionality was reduced using Principal Component Analysis (PCA), retaining 95% of variance in 20 components. K-Means clustering was applied to the PCA output, resulting in five meaningful customer segments. Cluster validation tests (statistical validation of clusters) applied to ensure validity of clusters.

Clusters exhibit clear differences in spending, booking frequency, and travel preferences. Based on cluster profiles, five tailored perks were recommended: Exclusive Discounts, Free Checked Bag, Free Cancellation, Free Hotel Night with Flight, and Free Meal. These perks are expected to improve personalization, increase conversions, and optimize incentive spending.

The project demonstrates how behavioral data can be transformed into actionable business insights using a robust analytical pipeline.

DATA & PREPROCESSING

Primary database was a PostgreSQL database with four tables namely, users, sessions, flights, and hotels. These tables were joined using user_id and trip_id to reconstruct each customer's journey from browsing to booking.

Key preprocessing steps:
- Conversion of timestamp fields to datetime
- Creation of session duration from session_start and session_end
- Removal of canceled trips from modeling dataset
- Handling missing values through default imputation
- Correction of invalid hotel night values
- Removal and clipping of extreme outliers using the IQR method
- Deduplication prior to aggregation

The cleaned pipeline produced three base tables:
- Session Base Table (session-level)
- Trip Table (one row per valid trip)
- User Feature Table (one row per user)

These steps ensured high data quality and modeling reliability.

FEATURE ENGINEERING & PCA

Thirty-seven features created. Feature categories are:
- Engagement: num_sessions, avg_clicks_per_session, avg_session_duration
- Frequency: num_trips, num_flights, num_hotels
- Monetary: total_spend, avg_money_spent_per_flight, avg_money_spent_per_hotel
- Behavior: cancellation_rate, avg_time_after_booking
- Seasonality: winter, spring, summer, fall trip counts
- Preference indices:
  - hotel_hunter_index
  - flight_fanatic_index
  - bundle_index

Features were standardized and fed into PCA. Twenty principal components retained 95% of total variance. Early components primarily captured spending and engagement, while later components captured seasonality and discount behavior.

PCA reduced noise and multicollinearity, improving clustering performance.

CLUSTERING & SEGMENT PROFILES

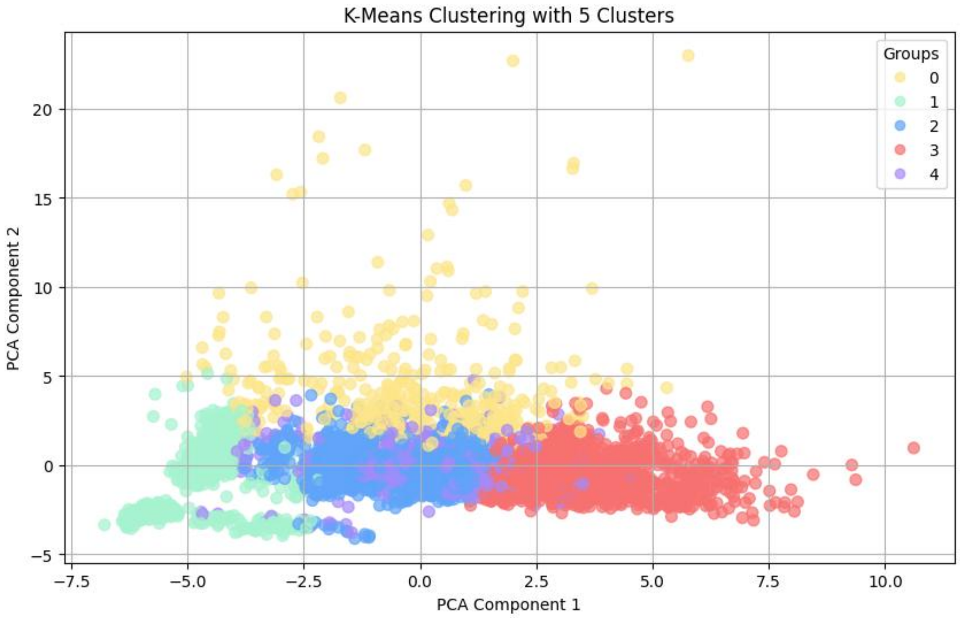
K-Means was tested across multiple k values. Elbow analysis suggested 5–6 clusters, while silhouette scores favored 3. Final selection of k=5 was chosen for business interpretability and evidence of five behavioral peaks.

Cluster Summary and Perk Assignments to Clusters

This K-means clustering plot (below) shows that most of the separation between the five clusters occurs along PCA Component 1, indicating that the primary dimension driving segmentation is captured on the horizontal axis. One cluster (red) is clearly separated on the right with high PC1 values, suggesting a distinct group with strong underlying behavioral characteristics relative to others. On the left side, another cluster (green) occupies low PC1 values, forming a contrasting group. The remaining clusters (blue, purple, yellow) are more concentrated around the center and partially overlap, implying they share some similarities but still differ in subtler ways captured by the combination of PCA Component 1 and 2. The key takeaway is that the data contains one or two strongly distinct traveler segments, plus several moderately differentiated groups clustered around average behavior. PCA Component 2 mainly adds vertical spread but does not create strong horizontal separations, meaning it refines distinctions rather than defining primary segments. Overall, the clustering appears reasonable: there is meaningful structure, but not extreme separation, which is typical for real behavioral data. These clusters can now be interpreted by examining feature averages per cluster to assign practical labels such as high-activity travelers, low-engagement users, deal-seekers, or group-oriented travelers

The main strategies to assign perks for customers were:

- Flight perks for flight-heavy users
- Hotel upgrades for hotel-focused users
- Discounts for price-sensitive users
- Re-engagement offers for at-risk users



Cluster (0):

Younger, hotel-focused, high hotel spenders with low overall travel engagement, taking fewer but more luxurious or longer hotel-centric trips. Low flight activity.
Assigned Perk: Free hotel night with flight

Justification: While they are hotel-focused, a "free hotel night with flight" could encourage them to combine their hotel stays with flights, potentially increasing their overall engagement and flight activity. This is a compromise from the ideal "Premium Hotel Upgrades" due to the limited options provided.

Cluster (1):

Highly engaged, frequent travelers who actively seek and use discounts, often bundling their travel. Likely to be older and travel year-round, with high session counts, clicks, and usage of flight/hotel discounts.
Assigned Perk: Exclusive discount

Justification: This cluster already shows high engagement and actively seeks discounts. Exclusive discounts would directly reward their behavior and encourage continued high activity and loyalty. This aligns perfectly with their discount-seeking nature.

Cluster (2):

Efficient and active travelers, like Cluster one but with less emphasis on discounts and lower hotel spending. Value bundled services, with moderate engagement, flights, and

hotels.
Assigned Perk: Free meal

Justification: This cluster is efficient and active. A free meal can add perceived value and convenience to their trips without directly focusing on discounts or large financial incentives. It's a simple, tangible benefit that enhances their travel experience, especially if they are looking for efficiency.

Cluster (3):

High-spending, often family or group travelers, taking long-haul flights, booking significantly in advance, and less interested in hotels as a primary focus. Highest `avg_money_spent_per_flight`, `avg_flight_km`, `avg_bags`, `avg_rooms`, `avg_seats`, `num_fam_and_fri_trips`, and `avg_time_after_booking`.

Assigned Perk: Free checked bag

Justification: This cluster takes long-haul flights and often travels with family/friends (`num_fam_and_fri_trips`, `avg_bags`, `avg_seats` are high). A free checked bag directly addresses a practical need and cost associated with their travel style, providing significant value for group or long-distance trips.

Cluster (4)

Highly engaged users who spend a lot of time browsing and exploring options, possibly looking for shorter stays, and are often interested in bundled deals. Highest `avg_clicks_per_session` and `avg_session_duration`. High `bundle_index`, `hotel_hunter_index`, and `flight_fanatic_index`, but lower `avg_night`.

Assigned Perk: Free cancellation fee

Justifications: This cluster is highly engaged in browsing and exploration, suggesting they might be indecisive or frequently changing plans. A "free cancellation fee" perk offers flexibility and reduces booking friction, encouraging them to book more frequently knowing they have an option to change without penalty, fitting their exploratory nature.

Limitations:

1. The dataset is a static snapshot of all customers and trips (one row per user, probably the same user different trips with different objectives, holiday or business)
2. Limited psychographic data (no customer satisfaction data and information is available in the dataset.)
3- Assumptions are based on K-Means and PCA these are just mathematical assumption and our decision K-means=5.

Business Impact

This segmentation and personalization framework enables targeted marketing, improved user experience, and more efficient incentive spending. Expected outcomes include higher conversion rates, increased retention, and greater customer lifetime value.

Suggested Validation Strategies

A future A/B testing framework is proposed to validate perk effectiveness. Users within each cluster would be randomly assigned to control and treatment groups, where the treatment group receives a cluster-specific perk. Booking conversion rate and revenue per user would be compared using statistical tests such as a two-proportion z-test and Welch's t-test. This experimental design would allow causal evaluation of whether personalized perks lead to improved business outcomes.

Measure lift in conversion and revenue per cluster: Quantify how much a perk increases bookings and spending for each customer segment compared to users who did not receive the perk.

Future Work:

- Dynamic segmentation
- Lifetime value modeling
- Real-time personalization
- Expanded experimentation

Conclusion:

The project delivers a scalable, interpretable segmentation framework that connects customer behavior to targeted business actions. It is based on behavioral data of the customers, best mathematical available calculations and validations that could have some variability from reality.

-End-