TravelTide Data Retrieval and Customer Behavior Analysis Project

1. Executive Summary

This project analyzes how customers interact with the TravelTide travel booking platform and how those interactions translate into bookings, cancellations, and revenue. The analysis combines data from four relational tables, namely sessions, users, flights, and hotels, to create a clean analytical dataset. Behavioral features are engineered, and exploratory analysis is performed at both the session level and the user level. The analysis shows that customer value is highly concentrated, meaning that a relatively small group of users contributes a large share of total spend. Engagement signals such as session duration and click intensity are closely linked to booking behavior and can be used to estimate booking intent. Cancellations are not random and tend to align with certain booking patterns, which makes them important for risk control and reward strategy design. Bundled behavior, where users book both flights and hotels, is associated with higher engagement and presents strong cross-sell opportunities. Overall, the analysis supports data-driven recommendations for reward targeting, discount governance, and future segmentation and predictive modeling.

2. Business Problem and Objectives

Travel platforms aim to increase conversion and revenue while controlling cancellation rates and discount-related costs. To design an effective reward strategy, it is necessary to understand which users are most valuable, which behaviors indicate booking intent, and which behaviors signal a higher risk of cancellation. The objectives of this project are to retrieve and integrate TravelTide relational data into a single analysis-ready dataset, clean and validate the data to avoid misleading results, engineer interpretable behavioral and trip-related features, perform exploratory data analysis to uncover actionable behavioral patterns, and provide recommendations that can guide rewards and personalization strategies.

3. Data Sources and Structure

The project uses four relational tables. The users table contains demographic and account information, such as birthdate and signup date. The sessions table captures browsing behavior, including page clicks, timestamps, and booking or cancellation signals. The flights table contains flight-related trip details and costs. The hotel's table includes hotel stay details and pricing information. A key design choice in this project is to use the sessions table as the base table. Additional user and trip attributes are joined to sessions. This approach preserves the behavioral timeline while allowing deeper analysis through enrichment.

4. Data Retrieval and Integration

Data is extracted through structured SQL joins between sessions, users, flights, and hotels. Join keys such as user_id and trip_id are validated to ensure consistency. Session-level granularity is preserved before aggregating data to the user level. Relational integration is a common source of silent analytical errors, such as duplicated rows and inflated spend. In this project, integration is treated as a core analytical step, and join logic is checked and validated before any analysis is performed.

5. Data Cleaning and Quality Checks

The cleaning process focuses on making the dataset reliable and suitable for analysis. Datetime columns are standardized into proper datetime formats. Missing values are assessed carefully, distinguishing between cases that indicate no booking and those that represent unknown information. Duplicates created during joins are removed or reconciled. Numeric columns such as spend, clicks, and duration are validated and corrected where necessary. Extreme values, such as unusually high page clicks, are inspected and treated cautiously. From a statistical perspective, many travel-related metrics are heavily skewed, particularly spend and click-based variables. Outliers can dominate averages and distort correlations, so distributions are examined explicitly, and conclusions are not based solely on mean values.

6. Feature Engineering

Feature engineering is a central component of this project. The goal is to convert raw system logs into interpretable behavioral indicators. User and account features include user age derived from birthdate and account age in months derived from the signup date. These features help differentiate behavior across demographic and lifecycle groups. Session behavior features include session duration in minutes as a proxy for engagement, clicks per minute as a proxy for interaction intensity, booking type, and booking status indicators to classify sessions by user actions, and cancellation flags to identify trips that were canceled. Trip-related features include trip duration in days, trip category such as business or leisure patterns depending on classification logic, and trip distance in kilometers derived from origin and destination coordinates. Value-related features include flight spend, hotel spend, and total spend per trip or per user. Total spend serves as a core indicator for value-based segmentation. These features map directly to business questions related to conversion, cancellation risk, product preferences, and customer value.

7. Exploratory Data Analysis

Exploratory data analysis is used to identify patterns and generate hypotheses for modeling and reward strategy design. Distribution analysis shows that spend and engagement variables follow right-skewed distributions, where most users spend relatively little, and a small number of users spend a lot. Click-based metrics also show heavy tails, meaning that extreme users can distort averages. This highlights the importance of using medians, percentiles, and grouped analysis rather than relying only on means. Analysis of conversion and engagement shows that higher engagement often signals stronger booking intent. Longer sessions and higher click intensity align with higher booking likelihood. These engagement variables can therefore serve as early indicators of booking intent and support real-time personalization. User-level aggregation reveals a group of high-value users with substantially higher total spend. This pattern is typical in customer value analysis and supports targeted loyalty and retention strategies. Cancellation analysis shows that cancellations cluster around specific behavioral and booking patterns rather than occurring randomly. This insight supports more controlled discounting and reward confirmation policies. Product mix analysis shows that flights contribute the majority of total spend, while hotel-related activity tends to increase session duration and overall engagement. This indicates strong potential for flight and hotel bundling strategies.

8. Key Insights

First, customer value is highly concentrated, with a small set of users driving a large share of revenue. Second, engagement metrics such as session duration and click intensity are useful behavioral signals for booking intent. Third, cancellation behavior follows identifiable patterns and can be actively managed. Fourth, bundling flights and hotels is an effective growth lever that increases engagement and total value. Fifth, aggregating data at the user level is essential, as session-level logs are noisy and less stable for decision making.

9. Recommendations

For rewards strategy, value-based tiers should be used instead of flat rewards, focusing on users with high spend and consistent trip completion. High intent behavior should be rewarded carefully, with incentives triggered only when engagement signals indicate real intent. For personalization and conversion, early intent triggers based on session duration and click intensity should be implemented. Support or offers should be triggered at the right moment, such as after a threshold level of engagement. For cancellation risk control, discount governance should be applied, with large discounts limited for historically high cancellation patterns. Rewards should be confirmed after trip completion to reduce exploitation and wasted incentives. Cancellation prediction models should be explored as the next step. For cross-sell and bundling, flight and hotel bundles should be promoted with targeted messaging, particularly for users showing hotel browsing signals. Bundling incentives should be optimized rather than applied broadly to protect margins.

10. Limitations and Future Work

This project is exploratory and based on observational historical data. It identifies associations rather than causal relationships. Some behaviors, such as high engagement, may correlate with booking without directly causing it. Session logs may contain noise, such as idle time inflating session duration, and spend measures can be affected by join duplication if not carefully validated. Future work should include formal customer segmentation using clustering techniques based on user-level features such as value, engagement, cancellation rate, and product mix. Predictive models for booking propensity and cancellation risk should be developed and evaluated using proper train-test splits and performance metrics. Reward strategies should be validated through controlled experiments such as A and B testing. Time-aware features capturing recent activity, trends, and seasonality should be added to improve model robustness and real-world applicability.

11. Conclusion

This project demonstrates a complete data workflow that includes relational data retrieval, data cleaning, feature engineering, and exploratory analysis. The analysis produces clear and actionable insights while maintaining statistical caution and business relevance. It provides a strong foundation for reward optimization and personalization and can be extended into formal segmentation and predictive modeling frameworks.