TravelTide Rewards Segmentation – Detailed Project Report

---

1 Project Goal

TravelTide asked me to boost its new rewards program by matching every active customer with one of five possible perks. Using last year's click-stream and booking data I built a quick segmentation model so marketing can send the right perk to the right user.

---

2 Data Sources and Cleaning

I worked with four Snowflake tables covering **01 Jan – 31 Dec 2023**:

Table	Core fields I used	Rows after cleaning
sessions	session_id, visit_date, device, referrer	12.8 M
flights	booking_id, segments, pax, bags, spend_usd	742 K
hotels	booking_id, nights, ADR, city_tier, cancelled_flag	195 K
demographics	user_id, country, age_band, loyalty_status	3.1 M

Cleaning rules

- Kept only sessions **after 04 Jan 2023** so the window aligns with campaign start-up.
- **Active user** = ≥ 7 sessions in the period → **≈ 702 K users**.
- Removed rows with corrupt timestamps or non-positive spend.
- Filled missing numeric values with the median and categoricals with the mode.
- Dropped columns with > 97 % nulls.

---

3 Feature Engineering & Scaling

For a simple, explainable model I created **six numeric features** per user:

Feature	Description
total_sessions	Total site visits in 2023
cancellation_rate	Cancellations ÷ bookings
discount_usage_rate	Share of sessions that used a promo code

Feature	Description
total_nights	Hotel nights booked
total_checked_bags	Bags booked on flights
total_base_fare	Sum of base fares in USD

I scaled these columns to **[0, 1]** with *MinMaxScaler* so no single metric would dominate the distance calculation.

---

4 Dimensionality Reduction (for visual checks only)

I applied **PCA** and plotted the first two components. The scatter revealed four dense clouds, which suggested that a small-k K-Means model would work (see notebook cell 86).

---

5 Clustering Method

- **Elbow method.** I plotted K-Means inertia for k = 1…10; the elbow landed at **k = 4** (see notebook cell 84 → 86).
- **Model.** K-Means with random_state = 42, default parameters.
- **Result.** Every of the 702 K active users now has a cluster label 0-3.

Note: The notebook does **not** compute silhouette scores or bootstrap stability. Those extras can be added later if needed.

---

6 Cluster Profiles and Perk Matches

Cluster	Rough share	Behaviour highlights	Recommended perk
0 – Discount Seekers	~34 %	High promo-code use, low cancels	🎁 *Exclusive Discounts*
1 – Light Packers	~22 %	Few trips, almost no bags	🧳 *Free Checked Bag*
2 – Risk-Averse Planners	~19 %	Highest cancel-rate (≈46 %)	❌ *No Cancellation Fees*
3 – Premium Vacationers	~25 %	Long hotel + flight combos, top-quartile spend	🏨 *1 Free Night with Flight*

The mapping simply fixes the biggest pain-point visible in each cluster's stats.

7 Validation Plan

I haven't run a full offline uplift simulation in the notebook yet. Instead, I propose a live **50 / 50 A/B test** inside each cluster:

- **Metric:** Confirmed booking within 21 days of email.
- **Sample size:** ≥ 10 K users per arm gives ±0.4 pp precision on a 5 % base-rate.
- **Duration:** ~4 weeks.

---

8 Implementation Road-Map

Phase	Task	Owner	ETA
Data push	Add cluster column to marketing schema	Data Eng	+1 wk
Email build	Create 4 creatives, 1 per perk	CRM	+3 wk
A/B send	50 % treatment per cluster	Growth Ops	+7 wk
Model refresh	Re-train monthly via Airflow	DS Team	+9 wk

---

9 Risks & Mitigations

- **Cold-start users:** Until they hit 7 sessions they default to *Exclusive Discount*.
- **Seasonal drift:** KS test alerts if any feature shifts >4 σ.
- **Margin impact:** Finance dashboard will track revenue per booking by cluster.

---

10 Next Steps

1. Wrap the A/B test and measure real lift.
2. Add credit-card spend and NPS as new features for richer signals.
3. Consider uplift-based clustering to optimise directly for revenue.

Key Takeaways

- Four K-Means clusters built from six simple behaviour metrics give TravelTide an actionable way to personalise perks.
- The method matches what is coded in the project notebook (cells 61 → 86).
- A live test is queued so we can turn modelling insight into measured revenue.