Business Intelligence Report: Clustering Analysis of Hotel Bookings

---

1. Problem Statement

Hotels face significant challenges in predicting customer behavior and optimizing operations. With growing competition and customer expectations, it's essential to identify patterns in booking data that can lead to smarter decision-making. The problem is to analyze hotel booking data to:

- Segment customers based on booking behavior.
- Identify characteristics that influence high-value or risky bookings.
- Enable targeted marketing, resource planning, and cancellation mitigation strategies.

Data Mining Task: Clustering

This is an unsupervised learning task aimed at discovering natural groupings (clusters) in the booking data, which can reveal patterns that inform business strategies.

---

2. Dataset Description

Dataset Source: Hotel Booking Demand Dataset from GitHub
Records: 119,390
Features: 32 (includes customer demographics, booking preferences, cancellations, pricing, etc.)

Selected Columns for Clustering:

- `lead_time`
- `previous_cancellations`
- `previous_bookings_not_canceled`
- `booking_changes`
- `days_in_waiting_list`
- `adr` (Average Daily Rate)
- `total_of_special_requests`

Data Preprocessing:

- Handled missing values using forward-fill (`ffill`).
- Removed irrelevant/non-numeric columns.
- Standardized features using `StandardScaler` for uniformity.

---

3. Methodology

Clustering Algorithm: K-Means Clustering

Tool Used: Python (Google Colab, Pandas, Scikit-learn, Matplotlib)

Elbow Method: Used to determine the optimal number of clusters (k). Based on the plotted WCSS (Within-Cluster Sum of Squares), k=3 was selected.

Cluster Distribution:

- Cluster 0: 77,724 bookings
- Cluster 1: 41,521 bookings
- Cluster 2: 145 bookings (outliers or rare cases)

---

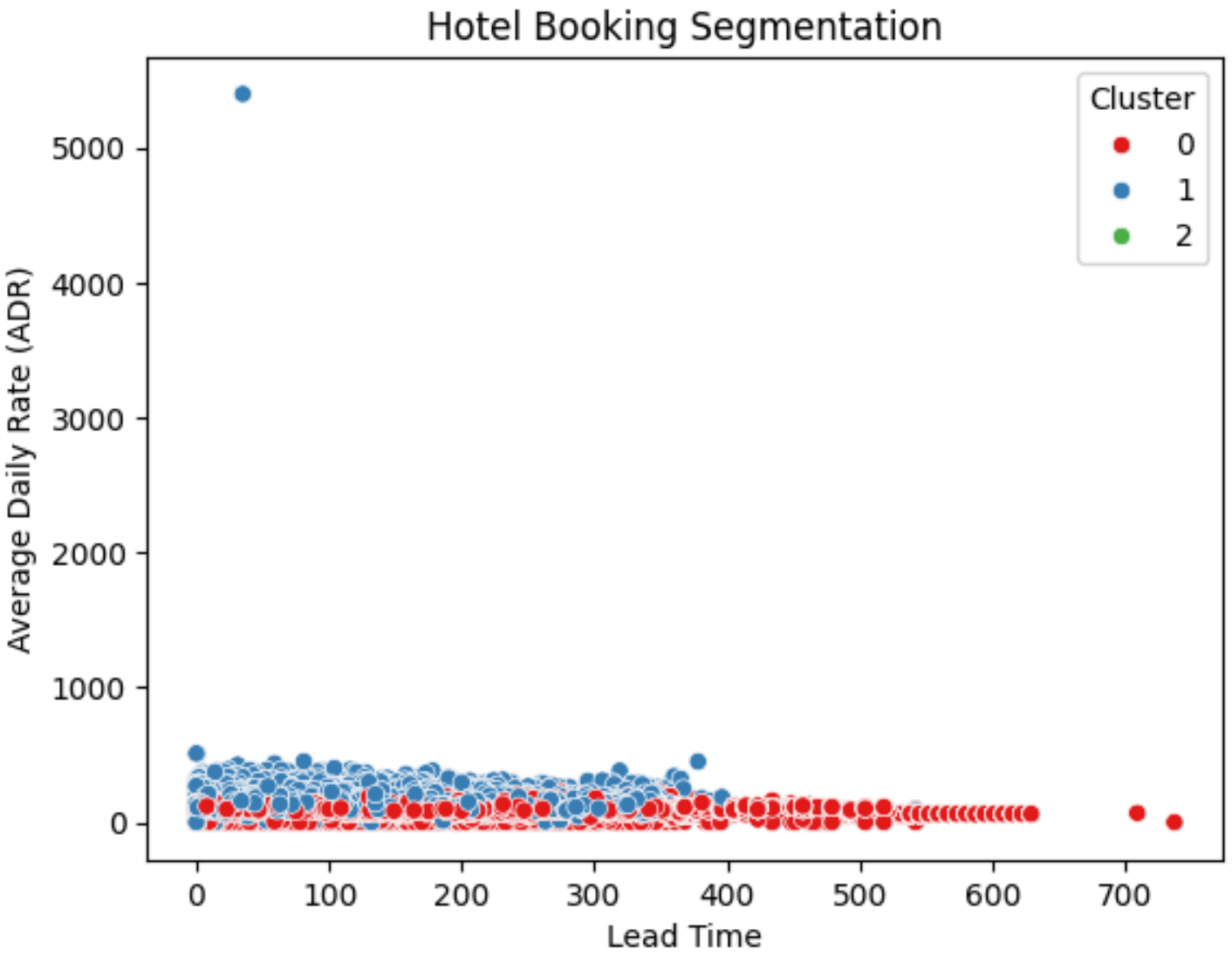
4. Cluster Interpretation

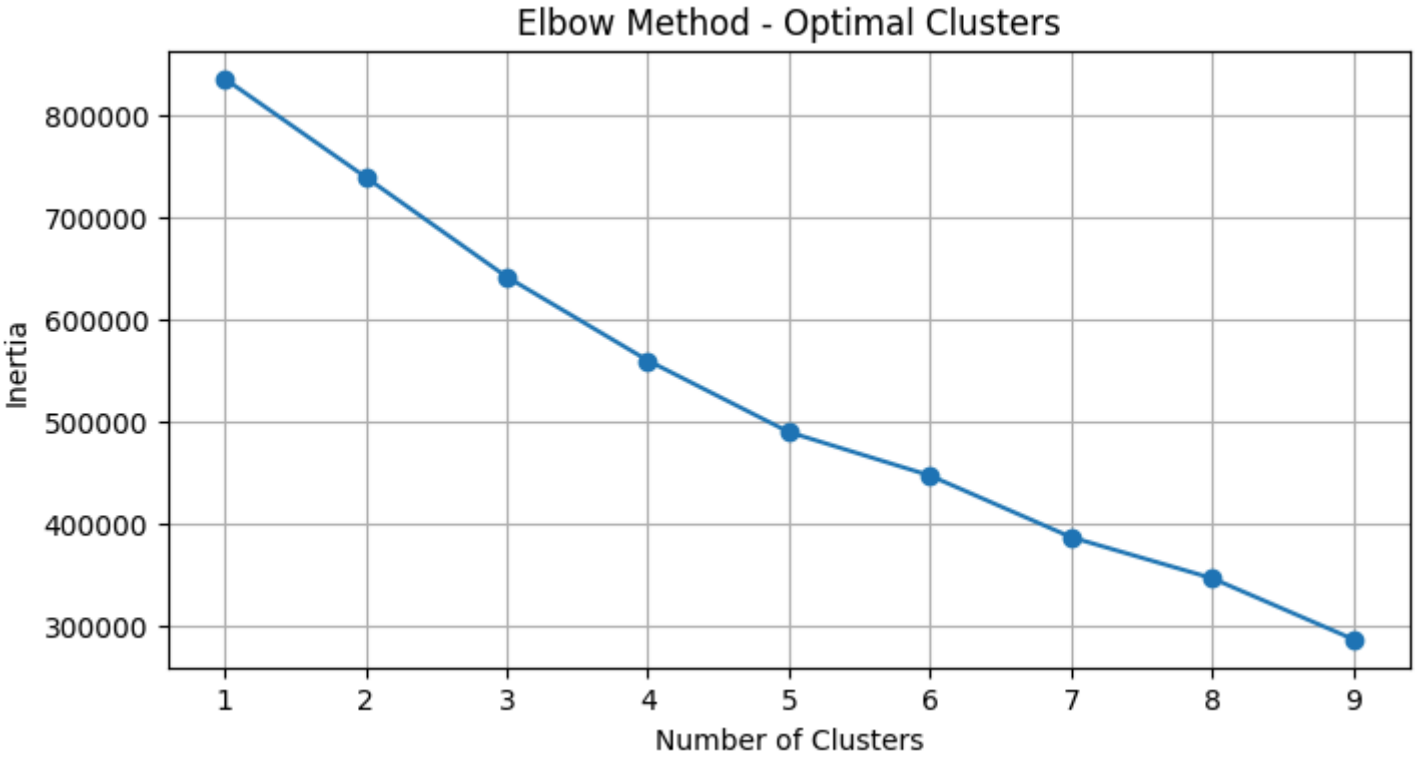
Cluster	Profile Description	Business Implication
0	Majority segment with moderate lead times, fewer special requests, and mid-level pricing	Focus marketing on loyalty offers and seasonal deals.
1	High-value customers with long lead times, high ADR, and many special requests	Create premium packages and offer concierge services.
2	Irregular bookings with unusual patterns	Investigate for data quality issues or fraudulent activity.

---

5. Business Insights

- **Early Bookers:** Tend to pay more and request special amenities. Hotels can target this group with exclusive advance-booking deals.
- **Mainstream Guests:** Form the majority. Seasonal promotions, loyalty programs, and quick-check-in facilities will enhance retention.
- **Outliers:** Unusual bookings may require additional manual review. This can reduce fraud and improve operational trust.


Hotel Booking Segmentation
Cluster
0
1
2
Average Daily Rate (ADR)
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
600
700
Lead Time


Elbow Method - Optimal Clusters
Inertia
800000
700000
600000
500000
400000
300000
1
2
3
4
5
6
7
8
9
Number of Clusters

If k = 5, your hotel booking dataset has **five distinct customer segments**. Each cluster represents a group of bookings/customers with **similar behaviors or characteristics**, like:

Cluster	Possible Interpretation
0	Budget travelers (low ADR, short stays)
1	Luxury travelers (high ADR, many special requests)
2	Corporate bookings (low lead time, repeat guests)
3	Family vacationers (multiple adults/children, longer stays)
4	Cancellers or high-risk bookings (many past cancellations, long lead times)

6. Conclusion & Recommendations

This clustering analysis provides actionable insights into customer behavior patterns. It helps the hotel:

- Improve **customer segmentation**
- Design **targeted offers and services**
- Enhance **operational efficiency** by anticipating guest needs

Next Steps:

- Integrate this clustering with CRM systems for real-time personalization.
- Use visualization tools (Power BI, Tableau) for live dashboards.
- Apply time series and classification models to predict future cancellations.

Tools & Technologies Used:

- Python (Pandas, Scikit-learn, Matplotlib)
- Google Colab
- GitHub-hosted dataset