Project Report: Customer Segmentation of Hotel Bookings

1. Project Overview

The objective of this project is to segment hotel customers into distinct groups based on their booking behavior, stay patterns, and spending levels.
By using K-Means clustering, we can better understand customer profiles, allowing the hotel to design personalized marketing strategies, improve retention, and optimize pricing.

2. Dataset

1. Source: Hotel Booking Dataset (119,390 records, 36 features)
2. Selected Features for Clustering (numeric only):
    a. Lead time (days before arrival booking is made)
    b. Length of stay (weekend & weekday nights)
    c. Number of adults, children, babies
    d. Is repeated guest (loyalty indicator)
    e. Previous cancellations & successful bookings
    f. Booking changes
    g. Days on waiting list
    h. ADR (Average Daily Rate – proxy for spending)
    i. Car parking requirements
    j. Total special requests

3. Methodology

1. **Data Cleaning:** Removed personal info (name, email, phone, credit card), handled missing values with median imputation.
2. **Feature Scaling:** Standardized all numeric features using StandardScaler.
3. **Model Selection:**
    a. Tested **K = 2 to 10** clusters using **Elbow Method** and **Silhouette Score**.
    b. Selected **K = 4** as optimal.
4. **Clustering:** Applied **K-Means** (4 clusters).

5. **Validation & Visualization:** Used PCA (2D projection) to visualize clusters.
6. **Profiling:** Calculated average feature values per cluster to interpret behaviors.

4. Cluster Profiles & Insights

1. **Cluster 0 – Budget Short-Term Guests**
    a. Lead time: 124 days (plan well in advance)
    b. Short stays: 2–3 nights
    c. Very low ADR (~89)
    d. Few/no cancellations, low special requests
    e. Interpretation: Cost-conscious planners, low-maintenance customers.
2. **Cluster 1 – Standard Mid-Spenders**
    a. Lead time: 88 days
    b. Short stays: 2–3 nights
    c. ADR slightly higher (~97)
    d. Very few cancellations or changes
    e. Interpretation: Average guests — not too demanding, steady revenue generators.
3. **Cluster 2 – High-Spending Families**
    a. Longer stays: 6+ nights (week + weekend)
    b. Often include children (0.5 avg children per booking)
    c. Highest ADR (~144) and more special requests (~1.0)
    d. Rare cancellations
    e. Interpretation: Valuable family vacationers — premium customers who need attention.
4. **Cluster 3 – Loyal Repeat Guests**
    a. Very short lead time (~31 days) → book last minute
    b. Short stays: ~2 nights
    c. Almost all are repeat guests (0.98 loyalty flag)

d. Higher history of cancellations (0.49)

e. Lower ADR (~64)

f. Interpretation: Loyal but price-sensitive guests, may cancel often. Need loyalty perks.

5. Recommendations

1. **Cluster 0 (Budget Planners):** Offer **early bird discounts** and packages.
2. **Cluster 1 (Standard Guests):** Maintain **steady pricing & upsell add-ons** (meals, services).
3. **Cluster 2 (High-Spending Families):** Provide **premium family packages**, kids' services, and targeted promotions.
4. **Cluster 3 (Loyal Guests):** Launch a **loyalty rewards program**, flexible cancellation policies, and exclusive perks.

6. Deliverables

1. Cleaned dataset with cluster labels → hotel_booking_with_clusters.csv
2. Cluster profile summary → cluster_profiles.csv
3. Visualizations → Elbow & Silhouette plots, PCA cluster scatter plot
4. Final report (this document)

7. Conclusion

The clustering analysis revealed **four distinct customer segments** with clear differences in booking behavior, spending, and loyalty.
By targeting each cluster with customized strategies, the hotel can:

- **Boost revenue** (focus on Cluster 2 families),
- **Improve retention** (Cluster 3 loyalists),
- **Fill occupancy efficiently** (Clusters 0 & 1 planners).