Real-Time Personalized Marketing for Retention - Using Spark & Kafka

Team Members:
Roy Efroni, 206483216
Tom Saacs, 318455573
Adam Bublil, 207271610
Shelly Levy, 318901550

1. Introduction
The hospitality industry serves a diverse customer base with varying needs and preferences. Traditional marketing strategies often fail to cater to individual behaviors, leading to lower retention and satisfaction rates. Additionally, processing large datasets efficiently is a challenge. This project aims to address these challenges by leveraging Apache Spark and Kafka to develop a data-driven solution for personalized marketing and cancellation prediction.

---

2. Problem Description
Hospitality Industry Focus
The hospitality industry must accommodate diverse customer needs while ensuring an optimized operational workflow.
Customer Diversity
Hotels cater to a wide range of customers with unique preferences, requiring personalized marketing strategies.
Need for Targeted Marketing
Generic marketing approaches reduce retention and satisfaction. A data-driven approach is needed to address individual customer behaviors effectively.
Data Processing Challenge
Large booking datasets require efficient processing with tools like Spark and Kafka to enable real-time segmentation and targeted marketing efforts.

---

3. Project Overview
Segmentation with Clustering
Using Apache Spark, customers are clustered based on behavioral data, enabling personalized marketing campaigns.
Cancellation Prediction
By incorporating clustering data, a cancellation classification model predicts the likelihood of cancellations, allowing for proactive marketing interventions (e.g., offering SPA vouchers to high-risk customers). It can do it in real-time using Kafka.

---

4. Relevance to Big Data

Volume

The project handles hundreds of thousands of records, including bookings, demographics, and revenue.

Scalability with Apache Spark

Spark enables distributed computing, ensuring efficient processing of large datasets.

Velocity (Real-Time Processing) with Kafka

Kafka facilitates real-time data processing from booking systems, loyalty programs, and feedback platforms.

---

5. Literature Review

Recent research (Thakur et al., 2024; Singh & Verma, 2025; Caetano, 2019) highlights best practices for customer segmentation in hospitality using K-Means clustering. Key insights focus on feature selection, evaluation methods, explainability, and scalability to enhance decision-making in hotels and Airbnb.

Feature Selection: Enhancing Cluster Accuracy

To improve clustering precision, entropy-based feature selection is used to retain high-impact variables, while removing correlated features prevents redundancy (*Singh & Verma, 2025*). This ensures distinct customer segments with meaningful differentiation.

Best Evaluation Methods for Clustering

Explainability with SHAP & LIME

For transparent decision-making, SHAP highlights feature importance, while LIME explains individual customer assignments (*Thakur et al., 2024*). This makes clustering models interpretable for marketing teams and revenue managers.

Optimizing Customer Segmentation in Hospitality with Scalable K-Means and Real-Time Analytics

Recent research (*Singh & Verma, 2025; Caetano, 2019*) highlights scalable K-Means clustering for customer segmentation in hotels and Airbnb. Apache Spark MLlib accelerates clustering for large-scale datasets, while Kafka enables real-time customer updates, ensuring dynamic segmentation (*Singh & Verma, 2025*). Leveraging these insights, businesses can identify high-value vs. low-value guests, predict booking cancellations, and offer personalized incentives (e.g., loyalty perks, discounts) to at-risk customers, enhancing retention and revenue optimization (*Caetano, 2019*).

To determine optimal clusters, robust metrics ensure segmentation quality (Thakur et al., 2024):

- Silhouette Score → Measures compactness.
- Davies-Bouldin Index → Evaluates separation.
- Calinski-Harabasz Index → Assesses density.
- Elbow Method → Identifies the ideal number of clusters.

---

6. Dataset

Content

- Hotel booking records from 2022 to 2024.

- Reservations and customer details, including nationality, booking source, room type, price, booking date and more.

Structure

- Data Type: Structured data.
- Key Features: plan (meal), arrival_month, nights, room_type, price, child_number, source.

Source

The dataset is extracted from the company's internal hotel management and booking systems.

---

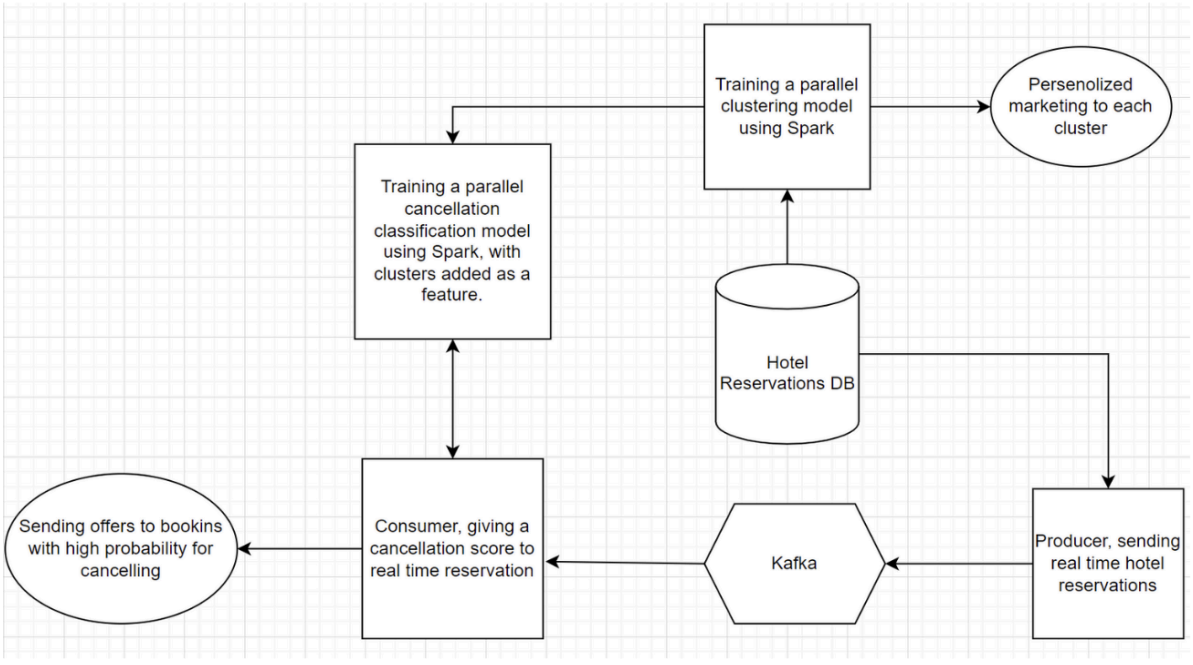
7. Solution Approach

1. **Load and Merge Data** - Integrating multiple data sources.
2. **Exploratory Data Analysis (EDA)** - Identifying patterns and trends.
3. **Data Cleaning** - Handling missing values and inconsistencies.
4. **Feature Engineering** - Creating meaningful features for modeling and deleting unnecessary ones.
5. **Build Clustering Model (Spark)** - Segmenting customers based on booking behavior.
6. **Evaluate & Fine-Tune Clustering** - Optimizing segment definitions.
7. **Business Decision - Personalized Marketing** - Utilizing clusters for targeted marketing.
8. **Add Cluster as a Feature** - Enhancing predictive capabilities.
9. **Build Cancellation Classification Model (Spark)** - Predicting cancellations.
10. **Evaluate & Fine-Tune Cancellation Model** - Improving model performance.
11. **Real-Time Evaluation with Kafka** - Processing "live" bookings.
12. **Business Decision - Targeted Interventions** - Implementing strategic marketing interventions.

---

8. Project Architecture

1. **Data Ingestion** - Real-time reservations are sent via Kafka.
2. **Storage** - Data is stored in the hotel reservations database.
3. **Model Training** - Spark-based clustering and classification models are trained in parallel.
4. **Marketing Strategies** - Personalized marketing offers are tailored for each customer cluster.
5. **Cancellation Prediction** - Predictions inform targeted interventions for high-risk bookings.
6. **Real-Time Evaluation** - Kafka-based consumers process real-time reservations.



9. Dataset Overview

- **66 Features** (categorical, numerical, date, string).
- **~170,000 Records**.

Missing Values

- **20.03% missing data, many empty features**.
- Features with more than 60% missing data are removed, 10 total.
- Addressed some by business understanding, for example In vip_code we filled the NAs with 0 since they were reservations that were not VIP.

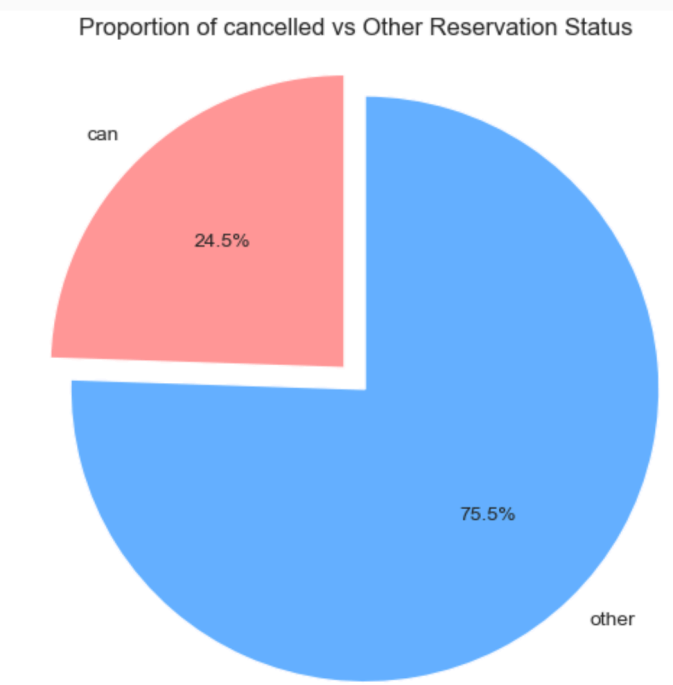
Outlier Detection

- Identified in key features: guests, price, nights.
- Addressed based on business logic, for example, removed the reservations of the room “presidential” since the prices there were very high and after a check with the company we decided that this room is not relevant to us.

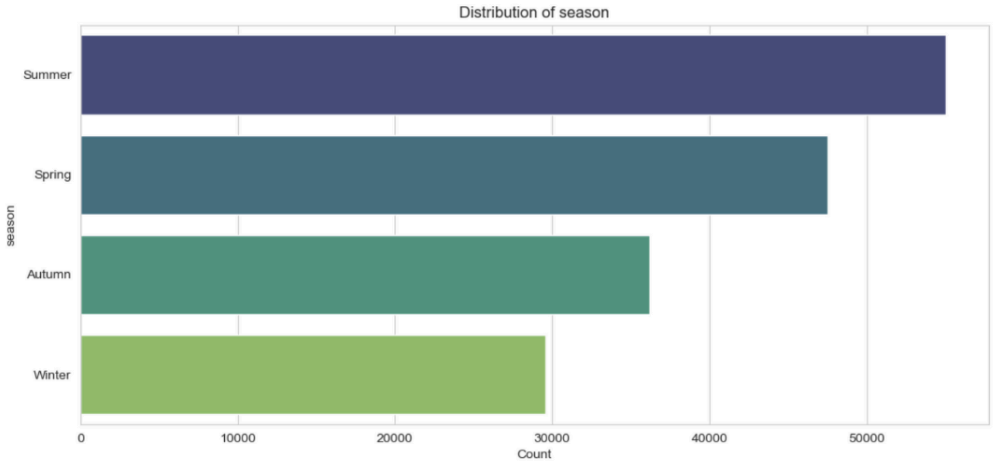
---

10. Exploratory Data Analysis (EDA)

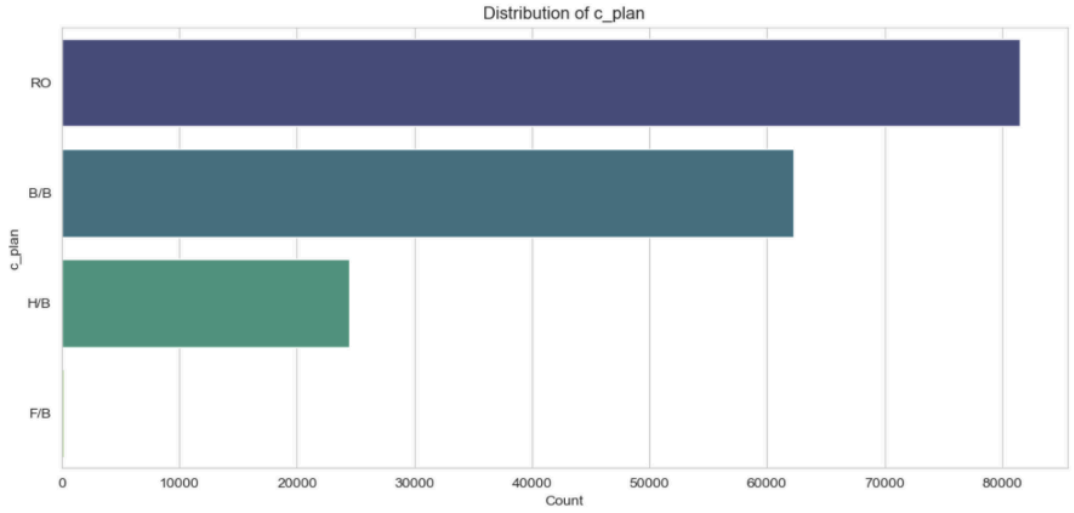
Reservation Status - 24.5% of bookings are canceled, requiring targeted retention strategies.



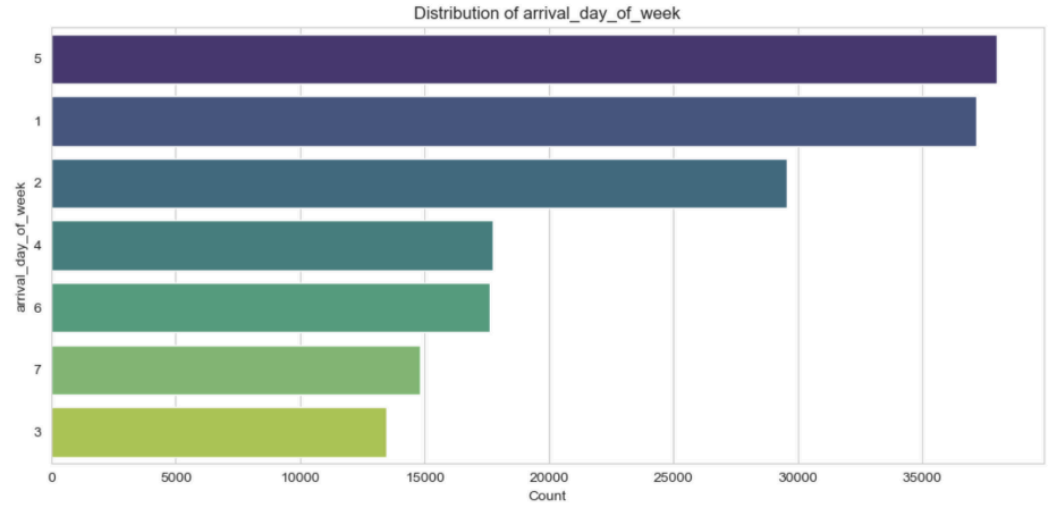
Seasonal Trends - Higher booking rates in summer, indicating peak seasonality.



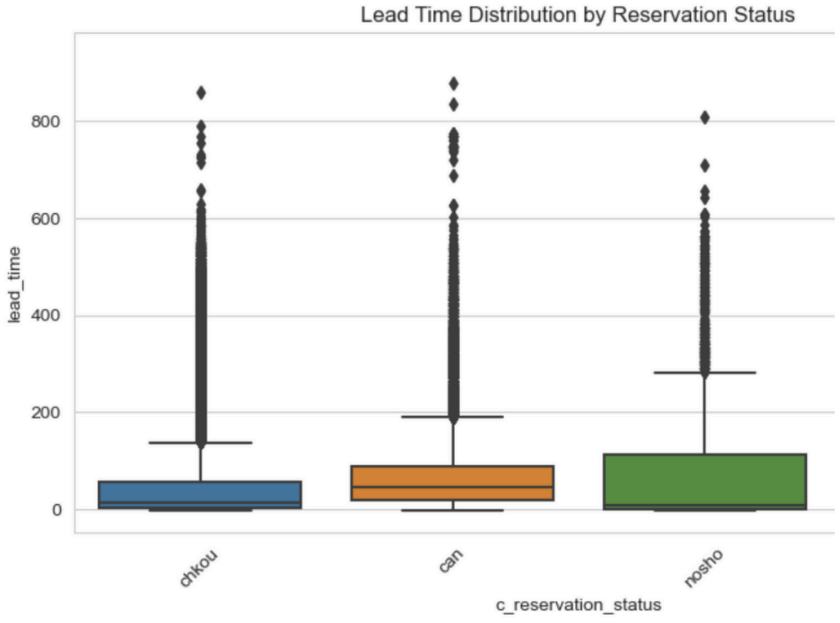
Booking Plans - Most popular: RO (Room Only) and B/B (Bed & Breakfast).



Arrival Day Trends - Most arrivals occur on weekends (Friday & Saturday).



Lead Time Analysis - Cancellations tend to have longer lead times, suggesting early-booking discounts may reduce cancellation rates.

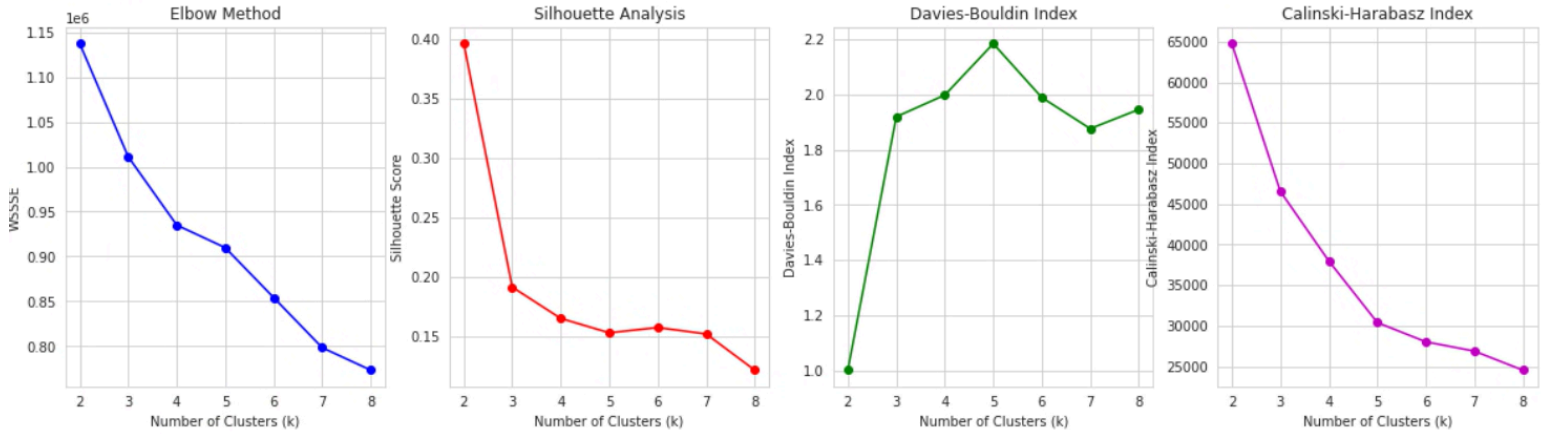


11. Methodology and Technical Implementation

We began by preparing the data for modeling through careful feature selection—removing columns with high missing values, low entropy, or those deemed irrelevant based on business logic. We also engineered new features, such as `lead_time` (the duration between reservation and arrival), and created dummy variables where appropriate. Additionally, we excluded non-relevant rows, such as group or company reservations.

Once the data was cleaned and normalized, we converted it into a Spark DataFrame composed of RDDs and applied the parallelized K-Means algorithm. To determine the optimal number of clusters, we ran the model multiple times with different values of K. Spark's distributed processing made this step efficient, although computing the clustering evaluation metrics for each model required more time.

The clustering evaluation metrics suggest that **K = 4** is the optimal choice for segmentation. The **Elbow Method** indicates a natural break at **K = 4**, where the reduction in WSSSE slows down, making additional clusters less effective. The **Silhouette Score**, which measures cluster separation, is highest at **K = 2** but drops significantly afterward, with **K = 4** still maintaining a reasonable balance. The **Davies-Bouldin Index**, which favors lower values for better clustering, increases beyond **K = 4**, suggesting that adding more clusters reduces overall separation quality. The **Calinski-Harabasz Index**, which rewards well-separated clusters, is highest at **K = 2** but declines sharply, with **K = 4** still maintaining relatively good separation. Considering all four metrics, **K = 4 provides the best trade-off between compactness, separation, and interpretability**, making it the ideal choice for this clustering task (and we know that we have more than 2 groups in our data).

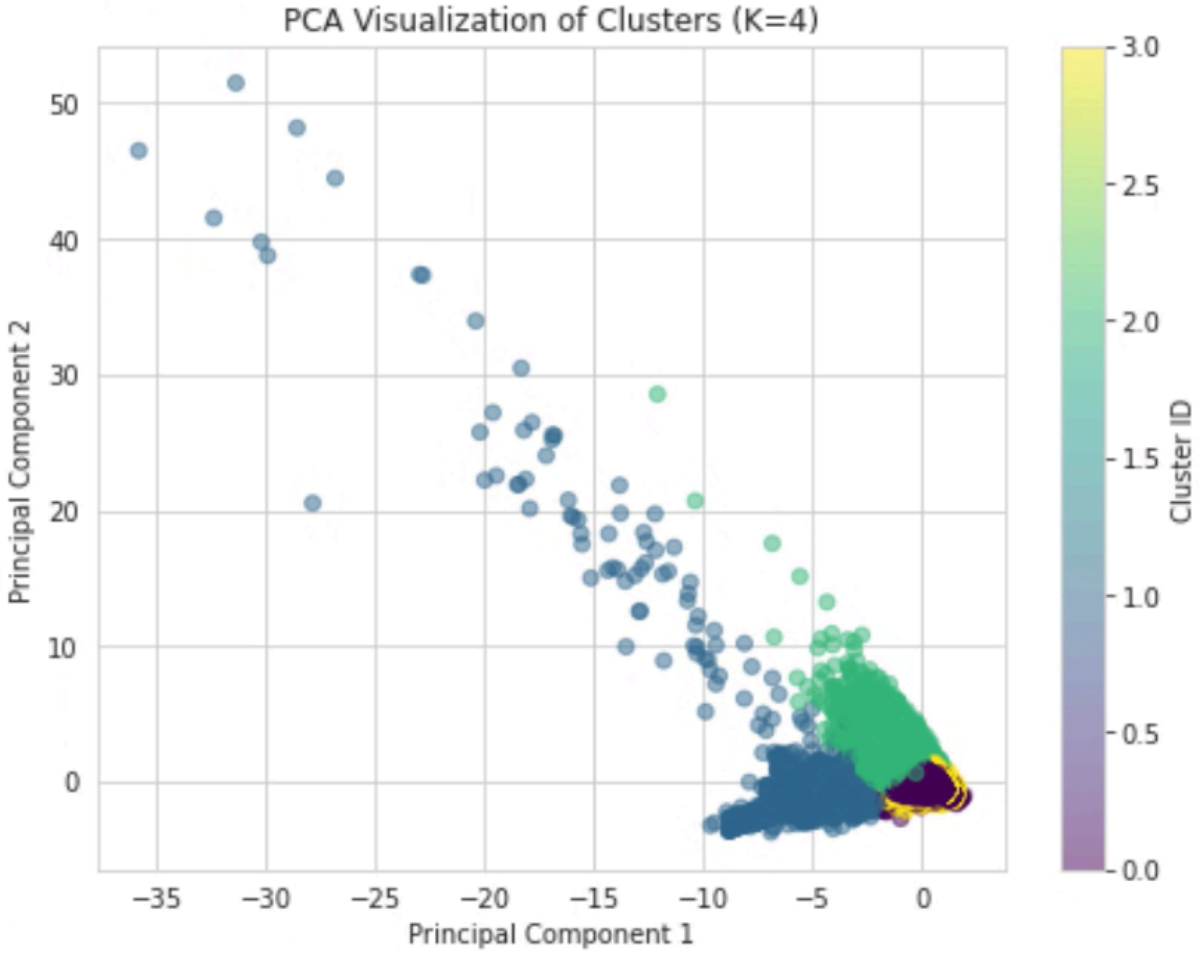


e then applied PCA to reduce the feature space to two dimensions, allowing us to visualize the optimal clustering model with 4 clusters. Afterward, we re-ran the clustering without including the target variable (`c_reservation_status`) to ensure the clusters remained unbiased and could be used as features in the classification model. Since reservation status had minimal influence on the clustering compared to features like price, lead time, and length of stay, the cluster structure remained largely unchanged. Thanks to Spark's distributed processing, this step was executed quickly and efficiently.

---

12. Results

PCA for 2 features on the 4 clusters, can see a clear separation, cluster 0 and 3 are similar.



Cluster Summary

From the centroids averages, we can infer about each cluster-

🔹 **Cluster 0: Short-Stay Local Families**

- 2.3 adults, 0.9 children, 0.18 babies → Likely small families or couples.
- Low VIP code (1.7) → Not frequent VIP guests.
- Short stays (2.16 nights) with moderate price (1961 local currency).
- Very short lead time (22 days) → Likely last-minute bookings.
- Mostly Israeli guests (99.7%).
- High reservation rate (82% confirmed, low cancellations).
- 88% from the hotel site.
- High meals plan (B/B 50%, H/B 13%).
- Season distribution is balanced, no strong seasonal preference.

👉 Summary: Local families or couples making short, last-minute trips. Prefer budget-friendly stays with breakfast included. 🚗🏡

🔹 Cluster 1: Long-Stay Time-Share Guests

- 3.87 adults, 0.36 children, 0.05 babies → Larger groups, fewer kids.
- Very low VIP code (0.1) → Rarely repeat or VIP guests.
- Long stays (6.55 nights) with high price (3133 local currency).
- Very long lead time (317 days) → Bookings planned nearly a year in advance.
- Mostly Israeli guests (99.9%).
- Lowest cancellation rate (8.5%) compared to other clusters.
- Mostly R/O plan (96%) → No meals included.
- Usually don't come on holidays, have the same week each year.

Summary: Long-stay time-share guests who book well in advance and typically visit during the same week each year. They prefer room-only (R/O) plans without meals and have low cancellation rates. Mostly Israeli guests traveling in larger groups, but with fewer children. Not holiday travelers, suggesting they have fixed vacation schedules. 🏡📅

🔹 Cluster 2: Large Families Spending More Per Stay

- 2.46 adults, 1.56 children, 0.24 babies → Larger families with kids.
- Moderate VIP code (2.99) → Some frequent visitors.
- Medium stays (3.74 nights) but very high price (6271 local currency).
- Moderate lead time (67 days).
- 98% Israeli guests.
- Higher mix of reservation types, including cancellations (40%!) and no-shows.
- 61% of them are in the Summer, 40% on Sundays.
- Mostly from the hotel site (65%) and from external (27%) sites.
- B/B and H/B (breakfast & half-board) are preferred.

👉 Summary: High-spending families, possibly for special occasions or holidays. Stay a bit longer and pay significantly more per stay. 🎉👨‍👩‍👧‍👦

🔹 Cluster 3: Discount & Promotional Guests

- 2.3 adults, 0.68 children, 0.11 babies → Mostly small families or couples.
- High VIP code (5.97) → Frequent returning guests.
- Short stays (2.19 nights) with lower spending (2049 local currency).
- Short booking lead time (25 days).
- Mostly Israeli guests (93%), the lowest percentage among clusters.
- High cancellation rate (32%) and the highest no-show rate (2%).
- 98% book through external internet platforms (not directly via the hotel).
- Highest percentage of holiday travelers.
- Prefer Studio rooms (7.5%), the most budget-friendly option.

- Bookings spread across the year but peak on Thursdays (28.5%), likely for weekend stays.

👉Summary: Frequent, deal-seeking guests who take advantage of online discounts and promotions. They often book short stays close to their travel date, prefer budget-friendly Studio rooms, and travel during holidays and weekends. Many book through third-party websites, and they have the highest no-show and second highest cancellation rates, indicating less commitment to their bookings. 🏨💰

The hotel can leverage this insight to deliver more personalized marketing and services tailored to each customer cluster.

Classification

After adding the cluster ID of each reservation as a feature to the classification model, we split the data into training, validation, and test sets based on the reservation date to ensure the model does not learn from future information. We used Spark to run a Random Forest model and efficiently perform hyperparameter tuning using a grid search on the validation set. Finally, we trained the model on the combined training and validation sets using the best hyperparameters. Below are the results of the best model evaluated with a standard classification threshold of 0.5:

```
Confusion Matrix:
 prediction    0.0   1.0
y_label
0           11004   844
1            2237  1393
```

AUC-ROC: 0.8024871512993786,
Accuracy: 0.8021708231037602,
F1-score: 0.7823005235328373.

We can see that the model outperforms the baseline accuracy of 75.5% (a model which just predicts the bias), indicating that it provides meaningful predictive power and can indeed offer valuable support to the company's decision-making.

Kafka-

To simulate real-time hotel bookings, we implemented a Kafka producer that streams 100 reservations from the test data every 5 seconds. A Kafka consumer in Spark listens to the test_data topic, where it parses, transforms, and feeds incoming reservations into the trained Random Forest model to generate live cancellation predictions. These predictions, including the predicted class and cancellation probability, are written to an in-memory Spark table and monitored until all test reservations are processed. Once complete, we extract the predictions, link them back to the original reservation IDs, and rank them by cancellation probability, enabling the hotel to identify high-risk reservations and take proactive action. We applied the ranking and added IDs to the entire test dataset as an example, but it can easily be adapted to work on each incoming batch individually, depending on the hotel's operational needs.
Further explanation is provided in the code comments and the accompanying README file.

---

13. Conclusion

This project demonstrates the power and flexibility of Apache Spark and Kafka in building a scalable, real-time machine learning pipeline tailored for the hospitality industry. Spark's distributed computing capabilities enabled efficient processing of large-scale reservation data, from clustering for customer segmentation to training and tuning a high-performing classification model. Each cluster revealed distinct customer behaviors—ranging from high-spending families to deal-seeking short-stay guests—allowing for more personalized and effective marketing strategies. Our classification model, enhanced by the cluster information, significantly outperformed the baseline model, providing more accurate cancellation predictions. Kafka allowed us to simulate and handle real-time reservation streams, enabling the deployment of predictive analytics at scale. Together, Spark and Kafka provided a robust infrastructure for both batch and streaming workflows, empowering the hotel to make data-driven, real-time marketing and retention decisions. This highlights the practical value of modern Big Data tools in solving complex business challenges with speed, scale, and intelligence.

---

14. References

- Fonseca, J. P. M. R. da. (2019). Harnessing big data to inform tourism destination management organizations (Master's thesis). Universidade Nova de Lisboa.
- Gupta, A. (2024). The convergence of big data analytics and CRM practices: A review. International Journal of Computer Trends and Technology, 72(7), 74-82.
- Machine learning applied to tourism: A systematic review. (n.d.). Journal Name, Volume(Issue), Page numbers. DOI or URL