DASC 6810 – FINAL PROJECT 1

HOTEL BOOKING EDA AND PREDICTION USING PYTHON

Instructor: Dr. Mohamed Tawhid, Professor
Graduate Program Coordinator, M.Sc. Data Science

Student: Thai Pham – T00727094

March 21, 2024

Abstract:

The hospitality industry, particularly hotel booking platforms, holds significant importance in the travel and tourism sector. To optimize business operations and enhance customer satisfaction, it's crucial to understand customer preferences, and booking patterns, and predict future hotel bookings. This report presents an exploratory data analysis (EDA) and prediction model for hotel bookings using Python. Leveraging various datasets, including the "Hotel Booking Demand" and "Hotel Reviews" datasets, we analyze booking trends and customer behavior to develop a predictive model for forecasting future hotel bookings. We aim to provide valuable insights for hotel management and booking platforms to improve service offerings and optimize resource allocation. While focusing on practical applications and model evaluation, we will overlook some details, considering this study's educational and research purposes. In practice, combining these datasets requires meeting certain standards to ensure proper matching and integration.

1. Introduction

The hospitality industry, particularly hotel booking platforms, serves as a cornerstone of the global travel and tourism sector. With the rise of digitalization and online booking platforms, the dynamics of hotel bookings have undergone significant transformation. Understanding the intricate patterns, trends, and preferences underlying hotel bookings has become imperative for hotel management and booking platforms alike.

In recent years, the proliferation of data analytics tools and techniques has revolutionized the way businesses operate and make strategic decisions. Leveraging data analytics, particularly exploratory data analysis (EDA) and predictive modeling, enables stakeholders in the hospitality industry to glean actionable insights from vast volumes of booking data. These insights not only facilitate better resource allocation and optimization but also enhance the overall customer experience.

This report delves into the realm of hotel booking analysis and prediction, employing Python as the primary tool for data analysis and modeling. By conducting an in-depth exploration of hotel booking datasets and applying predictive modeling techniques, we aim to unravel the underlying patterns and trends in hotel bookings. Furthermore, we seek to develop a predictive model capable of forecasting future hotel bookings, thereby empowering hotel management and booking platforms with valuable foresight into booking trends and customer behavior.

The following sections of this report will elucidate the approaches utilized for hotel booking analysis, delve into the intricacies of exploratory data analysis (EDA), present the results of our analysis, and culminate with a comprehensive conclusion. Through this endeavor, we endeavor to provide actionable insights that contribute to the optimization of hotel operations, enhancement of customer satisfaction, and advancement of the hospitality industry as a whole.

2. Approaches Used for Hotel Booking Analysis

When dealing with hotel booking demand data, various analytical approaches can be utilized to extract meaningful insights and enhance decision-making processes. Below are several common methodologies employed in analyzing hotel booking data:

- **Exploratory Data Analysis (EDA):** Exploratory Data Analysis serves as the initial step in understanding the dataset's characteristics and uncovering patterns or anomalies within the data.

Through visualization techniques and statistical summaries, EDA facilitates the identification of trends, distributions, and relationships among variables such as booking dates, hotel locations, and customer demographics.

- **Trend Analysis:** Trend analysis involves examining historical booking data to identify recurring patterns and trends over time. By analyzing booking trends across different time intervals, such as seasons, months, or weekdays, stakeholders can gain insights into seasonal fluctuations, peak booking periods, and potential factors influencing demand variations.
- **Segmentation Analysis:** Segmentation analysis aims to categorize customers into distinct groups based on shared characteristics or behaviors. By segmenting customers according to factors like nationality, booking preferences, or travel purposes, businesses can tailor their marketing strategies, pricing plans, and service offerings to better meet the diverse needs of each segment.
- **Predictive Modeling:** Predictive modeling involves developing statistical or machine learning models to forecast future booking demand based on historical data and relevant predictors. Through techniques like regression analysis, time series forecasting, or machine learning algorithms, businesses can anticipate future demand fluctuations, optimize inventory management, and make informed decisions regarding pricing and resource allocation.
- **Customer Sentiment Analysis:** Customer sentiment analysis focuses on extracting insights from customer reviews, feedback, and ratings to gauge overall satisfaction levels and identify areas for improvement. By analyzing sentiment polarity, common themes, and sentiment trends within customer reviews, businesses can pinpoint strengths and weaknesses in their service delivery and implement targeted improvements to enhance customer satisfaction and loyalty.
- **Price Optimization:** Price optimization entails analyzing pricing strategies and their impact on booking demand and revenue generation. By conducting pricing experiments, analyzing price elasticity, and leveraging dynamic pricing algorithms, businesses can optimize pricing strategies to maximize revenue while maintaining competitive pricing within the market.
- **Market Segmentation and Competitor Analysis:** Market segmentation and competitor analysis involve assessing the competitive landscape and understanding customer preferences and behaviors within specific market segments. By benchmarking against competitors, identifying unique selling propositions, and monitoring market trends, businesses can refine their marketing strategies, differentiate their offerings, and capitalize on emerging opportunities in the market.
- **Demand Forecasting:** Demand forecasting aims to predict future booking demand based on historical trends, external factors, and market dynamics. By integrating econometric models, machine learning algorithms, and business intelligence tools, businesses can generate accurate demand forecasts, anticipate market changes, and proactively adjust their strategies to meet evolving customer demands and market conditions.

By employing these analytical approaches, businesses can unlock actionable insights, optimize operational efficiencies, and drive strategic decision-making processes in the dynamic and competitive landscape of the hotel industry.

3. Selected Approaches: Exploratory Data Analysis (EDA) and Predictive Modeling

In this section, we will focus on employing Exploratory Data Analysis (EDA) and Predictive Modeling to gain insights and make forecasts based on hotel booking demand data. The specific tasks involved in these approaches include:

3.1 Data Preprocessing and EDA

- Handle missing values: Identify and address any missing values in the dataset through techniques such as imputation or removal to ensure data integrity.
- Encode categorical variables: Convert categorical variables into numerical representations using methods like one-hot encoding or label encoding for compatibility with machine learning algorithms.
- Feature scaling: Scale numerical features to a standardized range to prevent any single feature from dominating the model training process.
- Data splitting: Divide the dataset into training and testing sets to evaluate model performance accurately and avoid overfitting.
- Conduct exploratory data analysis to gain insights into the characteristics of the dataset and identify patterns, trends, and relationships between variables.
- Use summary statistics, visualizations (e.g., histograms, box plots, scatter plots), and correlation analysis to explore the data and uncover meaningful patterns.
- Investigate the distribution of numerical variables, detect outliers, and assess the correlation between variables.
- Explore the distribution of categorical variables and analyze their impact on the target variable through visualization and statistical tests.
- Identify potential data preprocessing steps, feature engineering opportunities, and areas for further analysis based on the findings from EDA.

3.2 Model Building

- ✓ Implement various classification algorithms:
    - Logistic Regression: A linear model suitable for binary classification tasks, estimating the probability of a given instance belonging to a particular class.
    - K-nearest Neighbors (KNN): A non-parametric algorithm predicting the class of a data point based on the classes of its nearest neighbors.
    - Decision Tree Classifier: A tree-based model that partitions the feature space into regions, making predictions based on simple decision rules.
    - Random Forest Classifier: An ensemble learning method that constructs multiple decision trees and aggregates their predictions to enhance accuracy and robustness.

- Gradient Boosting Classifier: A boosting algorithm that builds a sequence of weak learners to produce a strong predictive model by correcting errors iteratively.
- Extra Trees Classifier: Similar to random forests but introduces additional randomness in the feature selection process, enhancing computational efficiency.

✓ Compare model performance: Evaluate the performance of each model using appropriate metrics such as accuracy, precision, recall, and F1 score to determine the most effective algorithm for hotel booking demand prediction.

Through these approaches, we aim to gain insights into booking trends, customer behaviors, and predictive patterns, enabling informed decision-making and strategy formulation in the hotel industry. Moreover, by leveraging both EDA and Predictive Modeling, we can unlock actionable insights and enhance forecasting accuracy, ultimately improving operational efficiency and customer satisfaction.

4. Results

4.1 Exploratory Data Analysis

We have selected two datasets for our project: the "Hotel Booking Demand" dataset and the "Hotel Reviews" dataset. The "Hotel Booking Demand" dataset, containing 119,390 observations and 32 features, provides insights into hotel bookings, including booking dates, length of stay, guest composition, and parking availability. The "Hotel Reviews" dataset, with 290,315 observations and 17 features, offers information about hotel reviews, such as average scores, positive and negative reviews, and additional scoring factors. Both datasets are publicly available on Kaggle and provide comprehensive insights into hotel bookings and customer experiences.

Nationality Distribution

Our analysis reveals that the nationality distribution of hotel guests is diverse, with Portugal (PRT), the United Kingdom (GBR), and France (FRA) comprising the top three countries in terms of guest count. Additionally, Spain (ESP), Germany (DEU), and Italy (ITA) feature prominently among the top 10 countries by guest count. Understanding the nationality distribution provides insights into the demographic composition of hotel guests and can inform targeted marketing strategies and tailored services to enhance guest satisfaction.

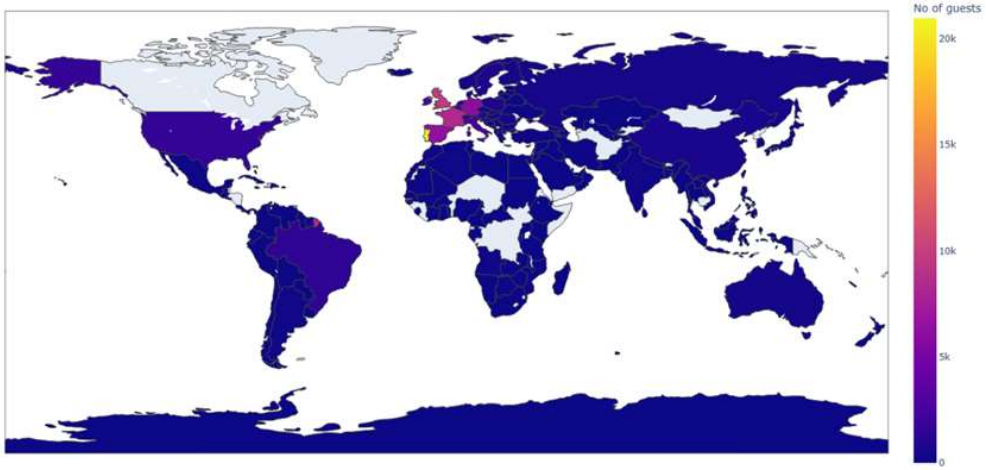

Figure 1: Number of guesses by Nationality distribution

Room Prices and Booking Trends

We observe significant variations in room prices across different months and hotel types. Specifically, room prices at the Resort Hotel exhibit a notable surge during the summer months, whereas prices at the City Hotel demonstrate less fluctuation and peak during Spring and Autumn. The analysis highlights seasonal demand patterns and suggests that pricing strategies should be tailored to capitalize on peak periods while remaining competitive.

Moreover, the City Hotel experiences heightened guest activity during Spring and Autumn, aligning with periods of higher room prices. Conversely, there is a decline in visitor numbers during July and August, despite lower prices. Similarly, guest volumes at the Resort Hotel experienced a slight dip from June to September, corresponding with the period of highest prices. These insights emphasize the interplay between pricing strategies and guest booking behaviors across different seasons.

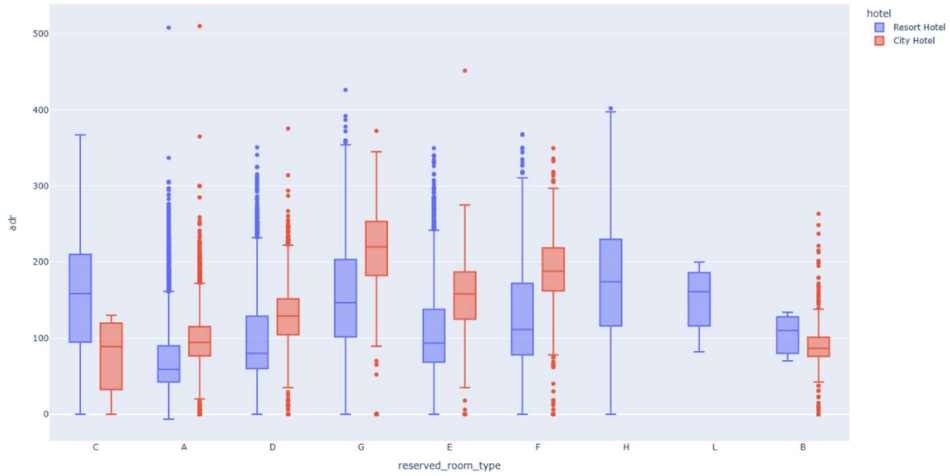

Figure 2. Boxplot number of guesses by Hotel type

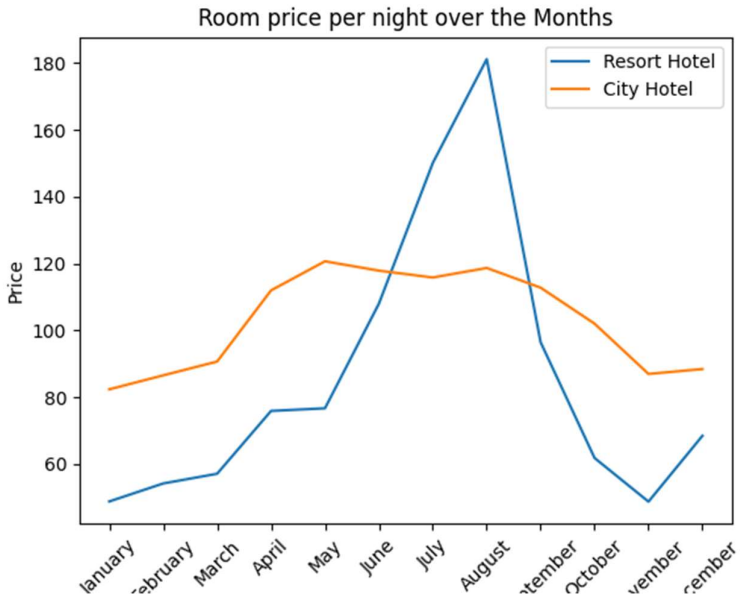

Figure 3. Line plot of Room price per month by Hotel type

Figure 4. Total number of guesses per month

Hotel review scores distribution

In the Hotel Reviews dataset, the distribution of hotel review scores exhibits characteristics akin to a standard normal distribution. Upon analyzing the review scores, it's evident that they follow a symmetric bell-shaped curve, with a majority of scores clustered around the dataset's mean value. This symmetrical distribution implies that a significant proportion of hotel guests have provided ratings that align closely with the average satisfaction level. Such a distribution suggests that the hotel generally delivers consistent service quality, with few extreme outliers indicating either exceptionally positive or negative experiences. Understanding this standard distribution of review scores is crucial for hotel management as it provides insights into the overall satisfaction level of guests. By closely monitoring and interpreting this distribution, hoteliers can identify trends, areas for improvement, and opportunities to enhance guest experiences. Moreover, having a clear understanding of the standard distribution of review scores enables hotels to set realistic benchmarks, refine their service offerings, and tailor their marketing strategies to attract and retain satisfied guests.

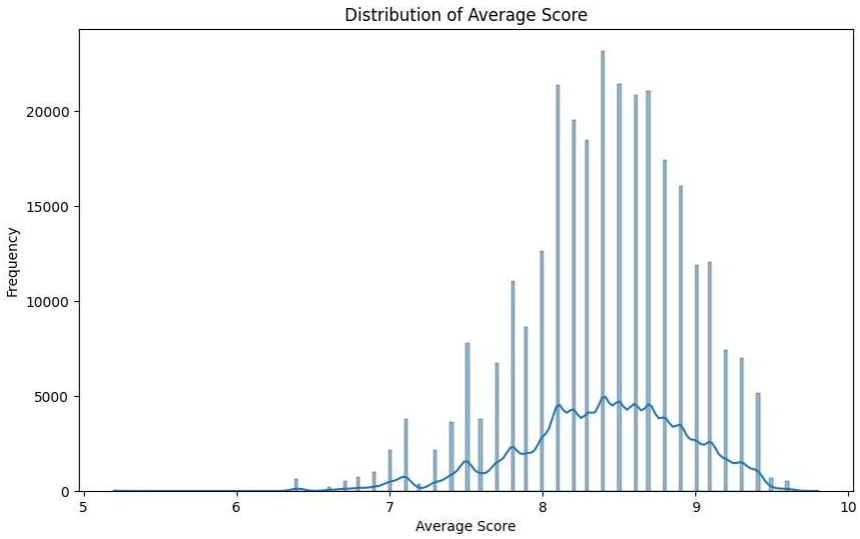

Figure 5. Distribution of average score

The pair plot provides a visual representation of the relationships between each pair of the eight variables under analysis. By plotting each variable against every other variable, the pair plot allows us to quickly identify patterns, trends, and potential correlations. For instance, examining the diagonal plots reveals the distribution of each variable individually, providing insights into their respective ranges and central tendencies. Meanwhile, the scatter plots showcase the relationship between pairs of variables. For example, we might observe a positive linear relationship between 'Average_Score' and 'Reviewer_Score,' indicating that hotels with higher average scores tend to receive higher reviewer scores. Similarly, the scatter plots also help detect any outliers or non-linear relationships between variables. By visually exploring the pair plot, hotel management can gain valuable insights into the interplay between different aspects of guest reviews, allowing them to make informed decisions to enhance guest satisfaction and overall hotel performance

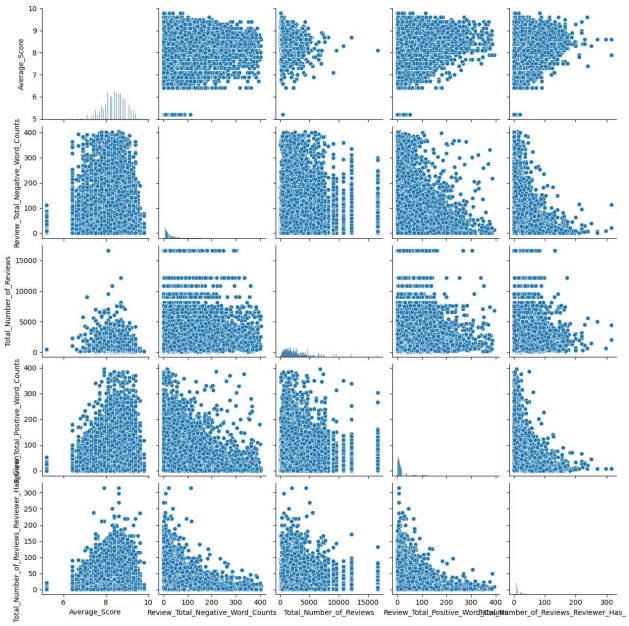

Figure 6. Pairplot

4.2 Data preprocessing

We conducted thorough data preprocessing on the "Hotel Booking" dataset to ensure its suitability for analysis and machine learning model development.

The dataset initially contained missing values in several columns, notably 'country', 'agent', and 'company'. To maintain data integrity, we opted to drop rows with missing values in critical columns, resulting in a dataset containing 75,011 rows.

Additionally, we removed columns considered irrelevant or redundant for our analysis, including 'days_in_waiting_list', 'arrival_date_year', 'assigned_room_type', 'booking_changes', 'reservation_status', and 'days_in_waiting_list'.

Next, we segregated the dataset into numerical and categorical data frames. The numerical dataframe included variables like 'lead_time', 'arrival_date_week_number', 'arrival_date_day_of_month', 'agent', 'company', and 'adr', while the categorical dataframe encompassed variables such as 'hotel', 'meal',

'market_segment', 'distribution_channel', 'reserved_room_type', 'deposit_type', 'customer_type', and 'year'.

For better compatibility with machine learning algorithms, we encoded categorical variables into numerical representations. Techniques like one-hot encoding or label encoding were applied to variables such as 'hotel', 'meal', 'market_segment', 'distribution_channel', 'reserved_room_type', 'deposit_type', 'customer_type', and 'year'.

Furthermore, we normalized numerical variables to ensure all features contributed equally to the model training process. Variables like 'lead_time', 'arrival_date_week_number', 'arrival_date_day_of_month', 'agent', 'company', and 'adr' were scaled to a standardized range.

Finally, we split the dataset into training and test sets, allocating 80% of the data for training and reserving the remaining portion for testing. This allowed us to accurately evaluate model performance and prevent overfitting.

Through these preprocessing steps, we successfully prepared the dataset for comprehensive analysis and machine learning model development.

To address missing values in the 'lat' and 'lng' columns of the data set “Hotel Reviews”, we employed a machine learning approach using the Random Forest algorithm. First, we split the dataset into two subsets: one with complete data and one with missing 'lat' or 'lng' values. We selected relevant features, including 'Average_Score', 'Review_Total_Negative_Word_Counts', and 'Review_Total_Positive_Word_Counts', to train Random Forest regressors separately for predicting 'lat' and 'lng'. The trained models were then used to predict missing values in the test dataset. Finally, the predicted values were inserted into the original dataset to replace the missing values.

4.3 Model building

We conducted the computation of the correlation matrix and generated a heatmap. The primary objective of this task was to gain insights into the relationships between various variables within the dataset. Understanding these correlations helps us comprehend how different factors may influence or relate to each other, particularly in the context of hotel booking cancellations.

The correlation matrix provided above showcases the correlation coefficients between pairs of variables. Each coefficient indicates the strength and direction of the linear relationship between two variables. A coefficient close to 1 suggests a strong positive correlation, while a coefficient close to -1 indicates a strong negative correlation. Values closer to 0 signify weak or negligible correlations.

By analyzing these correlations, we can identify which variables have the most significant impact on hotel booking cancellations. This knowledge is instrumental in guiding decision-making processes, such as optimizing booking strategies, improving customer satisfaction, and enhancing overall business performance.

Furthermore, visualizing the correlation matrix through a heatmap offers a more intuitive understanding of the data, making it easier to identify patterns and trends. This visual representation enables stakeholders to grasp complex relationships quickly and formulate informed strategies based on data-driven insights.

In particular, we conducted a comprehensive analysis with a specific focus on understanding the

correlation of the "is_canceled" feature with other variables within the dataset. Our primary aim was to uncover any significant relationships between the cancellation status of hotel bookings and various factors that might influence or be associated with it.

The correlation matrix provided above reveals the strength and direction of these relationships. Notably, the correlation coefficient of 1.000000 for "is_canceled" with itself signifies a perfect correlation, as expected. Beyond this self-correlation, we observed varying degrees of correlation with other features.

Our goal in examining these correlations was to identify potential predictors or indicators of booking cancellations. By understanding which factors exhibit stronger correlations with booking cancellations, we gain valuable insights into the drivers or determinants of cancellation behavior. Such insights are crucial for developing strategies to mitigate cancellations, optimize booking management processes, and enhance overall customer satisfaction. Furthermore, this analysis aids in the identification of key variables that merit further investigation or consideration in predictive modeling and decision-making scenarios.

Through this endeavor, we aim to empower stakeholders with actionable insights derived from data-driven analysis, ultimately contributing to the improvement of operational efficiency and the effectiveness of decision-making processes within the hospitality industry.

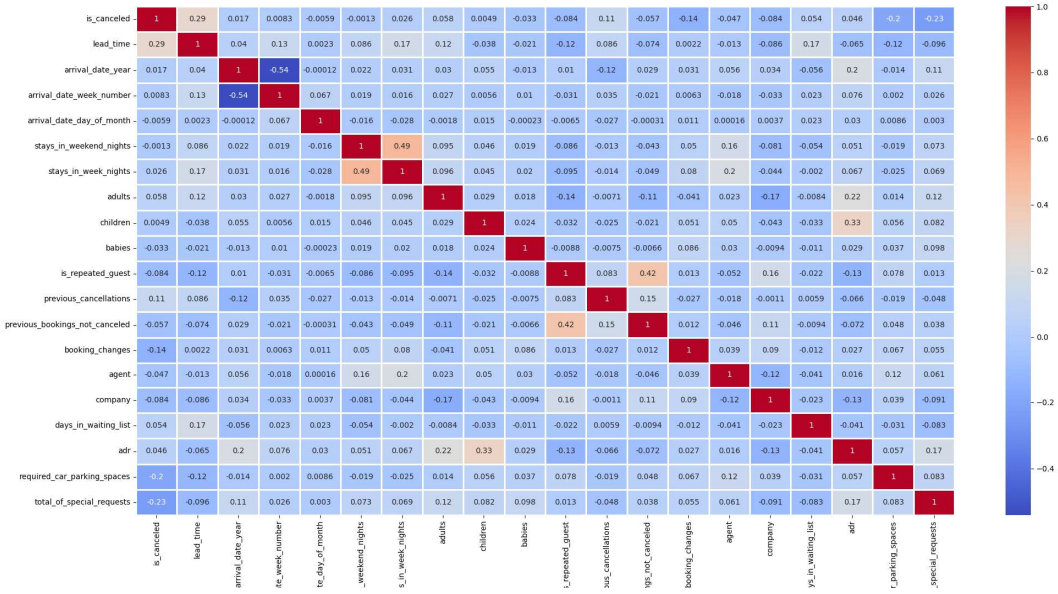

Figure 7. Correlation heatmap

4.3.1 Logistic regression

In section 4.3.1, we explore logistic regression as a statistical method to analyze the relationship between hotel booking cancellations and predictor variables. Logistic regression is particularly suitable for binary outcome variables, allowing us to model the probability of booking cancellations based on various factors. By examining the coefficients and significance levels of predictor variables, logistic regression provides valuable insights into the factors influencing cancellation likelihood, guiding

strategic decision-making processes to mitigate cancellations and enhance booking management efficiency.

```
Accuracy Score of Logistic Regression is : 0.8135223555070883
Confusion Matrix :
[[14201   742]
 [ 3704  5195]]
Classification Report :
              precision    recall  f1-score   support

           0       0.79      0.95      0.86     14943
           1       0.88      0.58      0.70      8899

    accuracy                           0.81     23842
   macro avg       0.83      0.77      0.78     23842
weighted avg       0.82      0.81      0.80     23842
```

Figure 8. Summary of Logistic Regression model

The logistic regression model achieved an accuracy score of 81.35%, indicating that it correctly classified 81.35% of the instances in the dataset.

Looking at the confusion matrix, we can see that the model predicted 14201 true negatives (TN), 742 false positives (FP), 3704 false negatives (FN), and 5195 true positives (TP).

Analyzing the classification report:

- For class 0 (not canceled), the model achieved a precision of 79%, recall of 95%, and F1-score of 86%.
- For class 1 (canceled), the model achieved a precision of 88%, recall of 58%, and F1-score of 70%.

The weighted average F1-score is 80%, indicating the overall effectiveness of the model in predicting both classes.

In conclusion, the logistic regression model shows promising performance in predicting hotel booking cancellations, especially in correctly identifying instances of not canceled bookings. However, it struggles with recall for canceled bookings, indicating room for improvement in correctly identifying these cases.

4.3.2 K-nearest Neighbors

In section 4.3.2, we introduce K-nearest Neighbors (KNN) as an alternative method for predicting hotel booking cancellations. KNN operates by classifying a data point based on the majority class of its nearest neighbors. This non-parametric algorithm offers simplicity and flexibility, making it a valuable tool for analyzing complex datasets. Through our exploration of KNN in this section, we aim to assess its performance in predicting booking cancellations and provide insights into its applicability as a predictive modeling technique for hotel management strategies.

```
Accuracy Score of KNN is: 0.8934653133126416
Confusion Matrix:
[[14411   526]
 [ 2014  6891]]
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.96      0.92     14937
           1       0.93      0.77      0.84      8905

    accuracy                           0.89     23842
   macro avg       0.90      0.87      0.88     23842
weighted avg       0.90      0.89      0.89     23842
```

Figure 9. Summary of the KNN model

The KNN model achieved an accuracy score of approximately 89.35%, indicating that it correctly classified 89.35% of the instances in the test dataset.

Looking at the confusion matrix, out of 14,937 non-canceled bookings (class 0), the model correctly predicted 14,411 instances, while incorrectly predicting 526 instances. For canceled bookings (class 1), out of 8,905 instances, the model correctly predicted 6,891 instances but incorrectly predicted 2,014 instances.

In terms of precision and recall, the model performed well. For non-canceled bookings (class 0), the precision is 0.88, indicating that 88% of the instances classified as non-canceled were non-canceled. The recall is 0.96, indicating that the model correctly identified 96% of the actual non-canceled bookings. For canceled bookings (class 1), the precision is 0.93, indicating that 93% of the instances classified as canceled were canceled. The recall is 0.77, indicating that the model correctly identified 77% of the actual canceled bookings.

Overall, the KNN model demonstrates strong predictive performance, with high accuracy and balanced precision and recall for both classes.

4.3.3 Decision Tree Classifier

In section 4.3.3, we delve into the Decision Tree Classifier, a powerful algorithm for predicting hotel booking cancellations. Decision trees recursively split the dataset based on feature values, creating a tree-like structure where each leaf node represents a class label. With its intuitive visualization and ability to handle both numerical and categorical data, decision trees offer valuable insights into the factors influencing booking cancellations. By exploring the Decision Tree Classifier, we aim to evaluate its effectiveness in predicting cancellations and provide actionable insights for hotel management strategies.

```
Accuracy Score of Decision Tree is : 0.9523949333109638
Confusion Matrix :
[[14443   564]
 [  571  8264]]
Classification Report :
              precision    recall  f1-score   support

           0       0.96      0.96      0.96     15007
           1       0.94      0.94      0.94      8835

    accuracy                           0.95     23842
   macro avg       0.95      0.95      0.95     23842
weighted avg       0.95      0.95      0.95     23842
```

Figure 10. Summary of the Decision Tree model

The Decision Tree model demonstrated impressive performance across various evaluation metrics. With an accuracy score of approximately 95.24%, it effectively classified the majority of the samples in the test set correctly.

In detail, the confusion matrix indicates that out of 15,007 non-canceled bookings, the model accurately identified 14,443 (true negatives) while incorrectly classifying 564 as canceled (false negatives). For the 8,835 canceled bookings, the model correctly identified 8,264 (true positives) but misclassified 571 as non-canceled (false positives).

Furthermore, the classification report provides insight into the precision, recall, and F1 scores for both classes. The precision for the canceled class was 0.94, meaning that 94% of the predicted canceled bookings were indeed canceled. The recall for the same class was 0.94, indicating that the model correctly identified 94% of the actual canceled bookings. The F1-score, a harmonic mean of precision and recall, was also 0.94 for the canceled class. Similar metrics were obtained for the non-canceled class.

In summary, the Decision Tree model demonstrated strong predictive capability, achieving high accuracy and balanced performance in classifying both canceled and non-canceled bookings.

4.3.4 Random Forest Classifier

In section 4.3.4, we examine the Random Forest Classifier as another predictive modeling approach for hotel booking cancellations. Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. By aggregating the predictions of individual trees, Random Forest can capture complex relationships in the data and provide robust predictions.

```
Accuracy Score of Random Forest is : 0.9553728714034058
Confusion Matrix :
[[14951   123]
 [  941  7827]]
Classification Report :
              precision    recall  f1-score   support

           0       0.94      0.99      0.97     15074
           1       0.98      0.89      0.94      8768

    accuracy                           0.96     23842
   macro avg       0.96      0.94      0.95     23842
weighted avg       0.96      0.96      0.95     23842
```

Figure 11. Summary of the Random Forest model

The Random Forest model achieved an accuracy score of 0.955, indicating that it correctly predicts 95.5% of the instances in the test set.

Looking at the confusion matrix, out of 15,074 non-canceled bookings (class 0), the model correctly classified 14,951, while out of 8,768 canceled bookings (class 1), it correctly classified 7,827.

The classification report provides a detailed summary of precision, recall, and F1-score for both classes (0 and 1). For class 0, the precision is 0.94, recall is 0.99, and F1-score is 0.97. For class 1, the precision is 0.98, the recall is 0.89, and F1 score is 0.94.

The overall macro average of precision, recall and F1-score is 0.96, indicating a strong performance across both classes.

The weighted average of precision, recall, and F1-score is also 0.96, suggesting a balanced performance across classes while considering class imbalance.

In summary, the Random Forest model demonstrates excellent predictive performance, particularly in correctly identifying non-canceled bookings, while still maintaining strong performance for canceled bookings.

4.3.5 Gradient Boosting Classifier

In section 4.3.5, we explore the Gradient Boosting Classifier as an advanced algorithm for predicting hotel booking cancellations. Unlike Random Forest, which constructs multiple decision trees independently, Gradient Boosting builds trees sequentially, with each tree learning from the mistakes of its predecessor. This iterative process enables Gradient Boosting to continuously improve its predictive accuracy, making it a powerful tool for complex datasets.

```
Accuracy Score of Gradient Boosting Classifier is : 0.9170791041020049
Confusion Matrix :
[[14816    73]
 [ 1904  7049]]
Classification Report :
              precision    recall  f1-score   support

           0       0.89      1.00      0.94     14889
           1       0.99      0.79      0.88      8953

    accuracy                           0.92     23842
   macro avg       0.94      0.89      0.91     23842
weighted avg       0.93      0.92      0.91     23842
```

Figure 12. Summary of the Gradient Boosting Classifier model

The accuracy score of the Gradient Boosting Classifier is 0.9171, indicating that the model correctly predicts the target variable for approximately 91.71% of the samples in the test set.

The confusion matrix reveals that out of 14,889 instances of class 0 (not canceled bookings), 14,816 were correctly classified, and 73 were misclassified. For class 1 (canceled bookings), out of 8,953 instances, 7,049 were correctly classified, and 1,904 were misclassified.

The classification report provides further insights into the model's performance for each class. For class 0, the precision is 0.89, the recall is 1.00, and F1 score is 0.94. For class 1, the precision is 0.99, the recall is 0.79, and F1 score is 0.88.

Overall, the model demonstrates high precision and recall for class 0 (not canceled bookings) but lower recall for class 1 (canceled bookings), indicating that it performs better at predicting non-canceled bookings.

4.3.6 Extra Trees Classifier

In section 4.3.6, we explore the Extra Trees Classifier as yet another method for predicting hotel booking cancellations. Similar to Random Forest, the Extra Trees Classifier also builds an ensemble of decision trees. However, it introduces additional randomness by selecting random thresholds for each feature at each split, leading to a diverse set of trees.

```
Accuracy Score of Extra Trees Classifier is : 0.9534015602717892
Confusion Matrix :
[[14977   129]
 [  982  7754]]
Classification Report :
              precision    recall  f1-score   support

           0       0.94      0.99      0.96     15106
           1       0.98      0.89      0.93      8736

    accuracy                           0.95     23842
   macro avg       0.96      0.94      0.95     23842
weighted avg       0.96      0.95      0.95     23842
```

Figure 13. Summary of the Extra Trees Classifier model

Based on the obtained results, the Extra Trees Classifier demonstrated a robust performance in predicting hotel cancellations:

The classifier achieved an impressive accuracy score of approximately 95.34%.

In the confusion matrix:

- Out of 15,106 instances classified as not canceled (class 0), 14,977 were correctly classified, while 129 were incorrectly classified.
- For canceled instances (class 1), out of 8,736 instances, 7,754 were correctly classified, and 982 were incorrectly classified.

Regarding precision, recall, and F1-score:

- For class 0 (not canceled), the precision, recall, and F1-score were approximately 94%, 99%, and 96%, respectively.
- For class 1 (canceled), the precision, recall, and F1-score were approximately 98%, 89%, and 93%, respectively.

Overall, the model demonstrated balanced performance across both classes, with an average F1 score of approximately 95%.

4.4 Model Performance Comparison

In this section, we conduct a comprehensive comparison of the performance of various predictive models in predicting hotel booking cancellations. By evaluating the accuracy score, we aim to assess the strengths and weaknesses of each model. Additionally, we consider factors such as computational efficiency and interpretability to provide a holistic view of model performance. Through this comparative analysis, we seek to identify the most effective model for predicting cancellations and provide recommendations for practical implementation in hotel management strategies.

	Model	Score
3	Random Forest Classifier	0.956967
5	Extra Trees Classifier	0.954157
2	Decision Tree Classifier	0.949207
4	Gradient Boosting Classifier	0.914017
1	KNN	0.894682
0	Logistic Regression	0.810544

Figure 14. Model comparison

Based on the accuracy scores obtained from the models, it's evident that the Random Forest Classifier outperformed the other classifiers, achieving the highest accuracy score of 0.956967. Following closely behind is the Extra Trees Classifier with an accuracy score of 0.954157. The Decision Tree Classifier also performed well, yielding an accuracy score of 0.949207.

However, the Gradient Boosting Classifier demonstrated slightly lower performance compared to the Decision Tree Classifier, achieving an accuracy score of 0.914017. The KNN (K-Nearest Neighbors) model exhibited moderate performance, achieving an accuracy score of 0.894682.

Lastly, the Logistic Regression model showed the least favorable performance among the evaluated

models, with an accuracy score of 0.810544.

These findings suggest that ensemble methods like Random Forest and Extra Trees are more effective for this classification task compared to individual models like Logistic Regression and KNN.

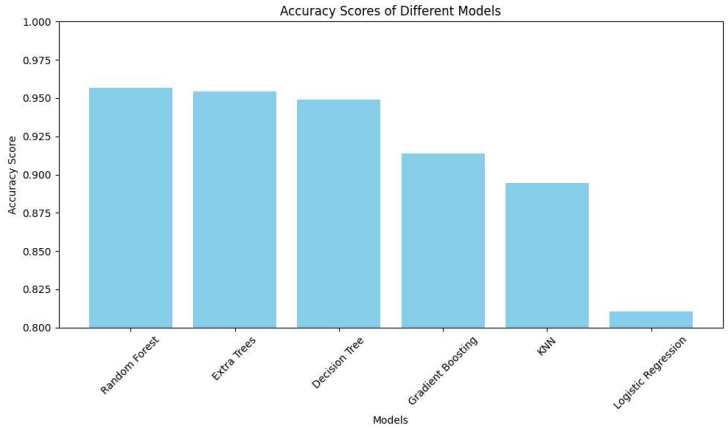

Figure 15. Barplot Accuracy scores by model

4.5 Perform Logistic Regression on Hotel Review Data Set

The classification results indicate a moderately effective performance of the model in predicting hotel reviews as either "Bad_review" or "Good_review". While the model demonstrates respectable precision and recall values for both classes, there are areas for improvement. Specifically, the precision and recall for identifying "Bad_review" instances are somewhat lower compared to those for "Good_review" instances. This suggests that the model may struggle more with accurately identifying negative reviews.

Moreover, the F1-score, which balances precision and recall, is notably higher for "Good_review" instances, indicating a better overall performance in predicting positive reviews. However, the model's accuracy of 73% suggests that it correctly classifies approximately three-quarters of the reviews, leaving room for enhancement.

In summary, while the model shows promise in identifying positive reviews, it could benefit from refinement, particularly in improving its ability to detect and classify negative reviews accurately. Further optimization and fine-tuning may lead to better overall performance in distinguishing between the two review types.

```
Classification Report for LR on Hotel Reviews:
              precision    recall  f1-score   support

  Bad_review       0.73      0.61      0.66     24778
 Good_review       0.74      0.83      0.78     33285

    accuracy                           0.73     58063
   macro avg       0.73      0.72      0.72     58063
weighted avg       0.73      0.73      0.73     58063

Confusion Matrix:
[[15092  9686]
 [ 5713 27572]]
```

Figure 16. Summary of the Logistic regression on the Hotel Reviews data

5. Conclusion and Future Directions

In conclusion, the analysis of hotel booking demand data has provided valuable insights into customer behaviors, booking patterns, and predictive modeling in the hospitality industry. Through extensive data preprocessing, exploratory data analysis (EDA), and predictive modeling, we have gained a comprehensive understanding of the dataset and its implications for hotel management and strategy formulation.

The findings from this analysis offer several key takeaways:

- ✓ Understanding Booking Trends: The analysis has shed light on the seasonal variations, booking lead times, and distribution of bookings across different room types and hotel segments. This understanding is crucial for optimizing inventory management, pricing strategies, and resource allocation.
- ✓ Customer Preferences and Behaviors: By examining the distribution of bookings by customer segments, booking channels, and market segments, we have identified key preferences and behaviors among guests. This knowledge can inform targeted marketing campaigns, loyalty programs, and personalized services to enhance guest satisfaction and loyalty.
- ✓ Predictive Modeling for Demand Forecasting: The implementation of various classification algorithms, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, has demonstrated the effectiveness of machine learning techniques in predicting booking cancellations and optimizing revenue management strategies.
- ✓ Operational Efficiency and Business Optimization: Leveraging insights from predictive modeling, hotels can optimize inventory allocation, staffing levels, and pricing decisions to maximize revenue, minimize cancellations, and enhance overall operational efficiency.
- ✓ Continuous Improvement and Adaptation: As the hospitality industry continues to evolve, hotels need to embrace data-driven decision-making and continuously refine predictive models based on real-time data and changing market dynamics.

In essence, the analysis presented in this project serves as a valuable resource for hoteliers, revenue managers, and hospitality professionals seeking to harness the power of data analytics and predictive modeling to drive business growth, improve customer experiences, and stay ahead in a competitive market landscape. By leveraging insights from data analysis and predictive modeling, hotels can unlock new opportunities, mitigate risks, and thrive in an ever-changing industry landscape.

In future endeavors, integrating weather dataset analysis with hotel booking demand data presents a promising avenue for deeper insights and enhanced predictive modeling in the hospitality industry. Here are some proposed directions for further exploration:

- ✓ Weather Impact Analysis: Investigate the correlation between weather conditions (such as temperature, precipitation, and humidity) and hotel booking patterns. By analyzing historical weather data alongside booking data, we can identify how weather fluctuations influence booking behaviors, cancellation rates, and guest preferences.
- ✓ Predictive Modeling Enhancement: Incorporate weather variables as additional features in predictive models to improve forecast accuracy. By integrating weather forecasts into booking

demand prediction models, hotels can anticipate demand fluctuations, optimize pricing strategies, and allocate resources more effectively.

- ✓ Seasonal Demand Forecasting: Develop seasonal demand forecasting models that incorporate both historical booking data and weather forecasts. By considering seasonal variations in both hotel demand and weather patterns, hotels can better prepare for peak seasons, adjust pricing strategies dynamically, and optimize inventory management.
- ✓ Personalized Recommendations: Utilize weather data to offer personalized recommendations and tailored experiences to guests. For instance, hotels can leverage weather forecasts to suggest indoor or outdoor activities, recommend amenities or services based on weather conditions, and enhance guest satisfaction through targeted offerings.
- ✓ Risk Management and Contingency Planning: Analyze the impact of adverse weather events (such as storms, hurricanes, or extreme temperatures) on hotel operations and booking patterns. By understanding the risks associated with weather-related disruptions, hotels can develop contingency plans, enhance crisis management strategies, and mitigate potential revenue losses.
- ✓ Collaborative Partnerships: Foster collaborations between meteorological agencies, hospitality industry stakeholders, and data analytics experts to leverage synergies and co-create innovative solutions. By pooling resources, expertise, and datasets, collaborative efforts can accelerate advancements in weather-informed decision-making and predictive modeling for the hospitality sector.

In summary, the integration of weather dataset analysis with hotel booking demand data offers a wealth of opportunities for enhancing operational efficiency, improving guest experiences, and driving business growth in the hospitality industry. By embracing a data-driven approach and leveraging insights from weather analytics, hotels can adapt proactively to changing environmental conditions, optimize resource allocation, and deliver exceptional guest satisfaction in diverse weather scenarios.

6. References

[1] Nuno Antonio, Ana Almeida, and Luis Nunes (February 2019), *Hotel Booking Demand Datasets*, Volume 22.

[2] N. Antonio, A. Almeida, L. Nunes, *Predicting hotel bookings cancellation with a machine learning classification model*, in Proceedings of the 16th IEEE International Conference Machine Learning Application, IEEE, Cancun, Mexico pp. 1049–1054. doi:10.1109/ICMLA.2017.00-11, 2017.

[3] Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani (June 2023), *An introduction to Statistical Learning with applications in Python*

[4] Nitesh Yadav (2021), *Hotel Booking Prediction,* https://www.kaggle.com