Reservation Cancellation Prediction

Team Members - Roles:

Siva Mani Venkat Korlakunta (11599987) – Dataset collection, Feature Extraction

Akhila Bommareddy (11614902) – Built Code, Implementation

Gopi Krishna Merugumala (11554008) – Built Code, Studied References/related work

Srikanth Reddy Gangavarapu (11609968) – Testing, Documentation

Aditya Kapilavai (11594174) – Built code, Studied References/related work

Abstract:

- The online hotel reservation channels have significantly changed the behavior of the customer and it gives lot of booking options to customers without making them to wait at the hotels for more time for a booking and it do saves the energy of the customers.

- In the same time, some amount of reservations are getting cancelled by the customers due to no shows or because of some unforeseen conditions such as issues in their schedule or might be because of changes in their plans lead to cancellations of the reservation. This cancellation becomes free or sometimes they will cost low amount which is an advantage for the customers but not for the hotel owners which will decrease the revenue of the owners.

- The main aim of this project is to help the owners of the hotel to increase their revenue by better understanding of the customers whether they are going to stay or not in the hotel. For this, we are going to use Decision tree which gives more accuracy rate.

Dataset:

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	hotel	is_cancele	lead_time	arrival_da	arrival_da	arrival_da	arrival_da	stays_in_v	stays_in_v	adults	meal	country	market_se	distributio	previous_	reserved_	assigned_	booking_c	deposit_t	agent	custor
2	0	0	9	2015	7	27	1	0	2	2	FB	PRT	Direct	Direct	0	C	C	0	No Depos	303	Transi
3	0	1	75	2015	7	27	1	0	3	2	HB	PRT	Offline TA	TA/TO	0	D	D	0	No Depos	15	Transi
4	0	1	23	2015	7	27	1	0	4	2	BB	PRT	Online TA	TA/TO	0	E	E	0	No Depos	240	Transi
5	0	0	35	2015	7	27	1	0	4	2	HB	PRT	Online TA	TA/TO	0	D	D	0	No Depos	240	Transi
6	0	0	18	2015	7	27	1	0	4	2	HB	ESP	Online TA	TA/TO	0	G	G	1	No Depos	241	Transi
7	0	0	7	2015	7	27	1	0	4	2	BB	GBR	Direct	Direct	0	G	G	0	No Depos	250	Transi
8	0	1	60	2015	7	27	1	2	5	2	BB	PRT	Online TA	TA/TO	0	E	E	0	No Depos	240	Transi
9	0	1	45	2015	7	27	2	1	3	3	BB	PRT	Online TA	TA/TO	0	D	D	0	No Depos	241	Transi
10	0	1	40	2015	7	27	2	1	3	3	BB	PRT	Online TA	TA/TO	0	D	D	0	No Depos	241	Transi
11	0	0	36	2015	7	27	2	1	3	3	BB	PRT	Online TA	TA/TO	0	D	D	0	No Depos	241	Transi
12	0	1	43	2015	7	27	2	1	3	3	BB	PRT	Online TA	TA/TO	0	D	D	0	No Depos	241	Transi
13	0	0	70	2015	7	27	2	2	3	2	HB	ROU	Direct	Direct	0	E	E	0	No Depos	250	Transi
14	0	1	45	2015	7	27	2	2	3	2	BB	PRT	Online TA	TA/TO	0	G	G	0	No Depos	241	Transi
15	0	0	107	2015	7	27	2	2	5	2	BB	PRT	Online TA	TA/TO	0	A	A	0	No Depos	240	Transi
16	0	1	47	2015	7	27	2	2	5	2	BB	PRT	Online TA	TA/TO	0	G	G	0	No Depos	240	Transi
17	0	0	50	2015	7	27	2	2	5	2	HB	IRL	Online TA	TA/TO	0	E	F	1	No Depos	241	Transi
18	0	0	76	2015	7	27	2	4	10	2	BB	OMN	Offline TA	TA/TO	0	D	D	0	No Depos	243	Contr
19	0	0	1	2015	7	27	2	0	1	2	BB	ARG	Online TA	TA/TO	0	H	H	0	No Depos	240	Transi
20	0	0	0	2015	7	27	2	0	1	2	BB	PRT	Online TA	TA/TO	0	A	D	0	No Depos	240	Transi
21	0	0	10	2015	7	27	2	0	2	2	BB	PRT	Online TA	TA/TO	0	G	G	0	No Depos	241	Transi
22	0	0	5	2015	7	27	2	0	2	2	BB	IRL	Online TA	TA/TO	0	E	E	0	No Depos	240	Transi
23	0	0	17	2015	7	27	2	0	3	2	BB	ESP	Direct	Direct	0	F	F	0	No Depos	250	Transi

T-test:

A t-test is a statistical hypothesis test that analyzes two groups' means to see if there is a statistically significant difference between them. It uses the t-distribution, which is comparable to the normal distribution but has significantly fatter tails, when the sample size is small to moderate and/or the population standard deviation is unknown.

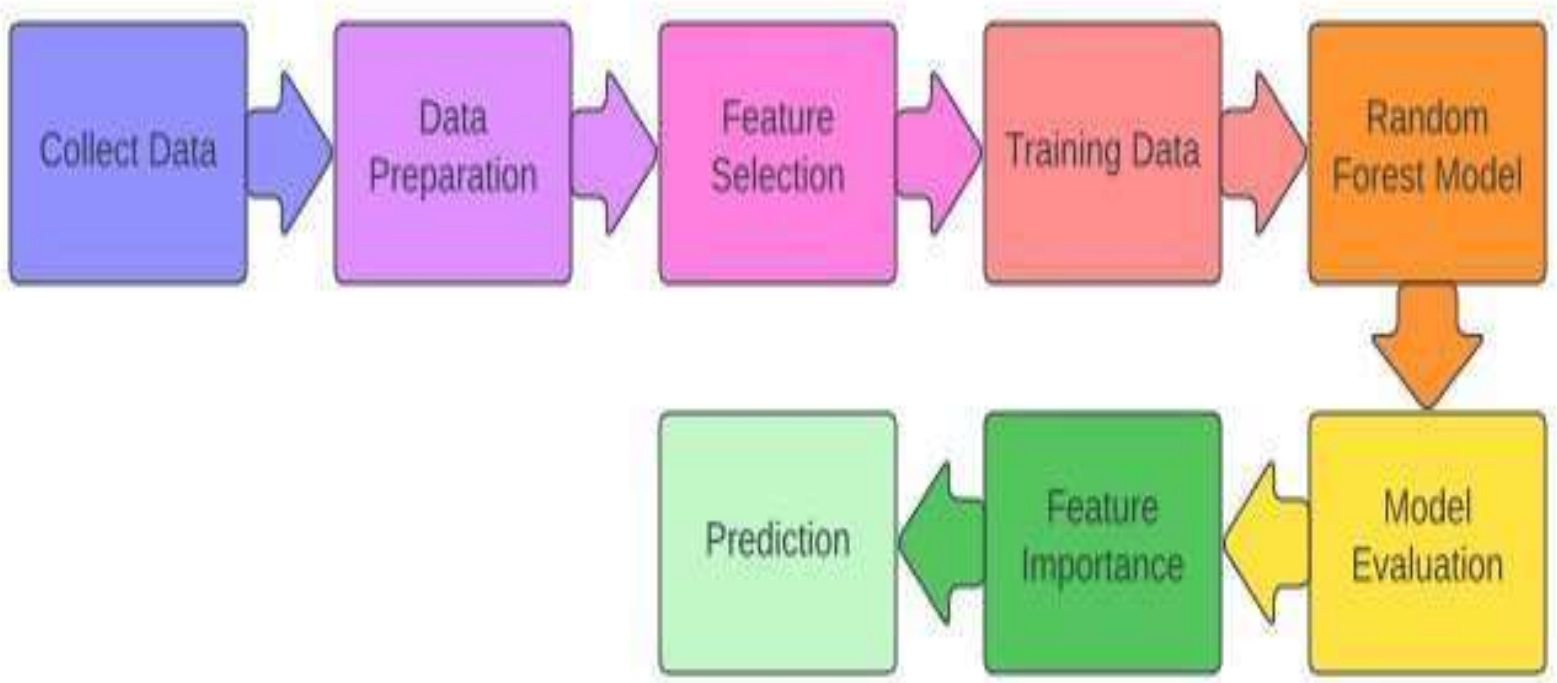
Decision Tree Algorithm:

One advantage of using a decision tree for this task is that it can easily handle both categorical and continuous input features. This is important because reservations may have a variety of different features that could be relevant to the cancellation decision, such as the date of the reservation, the type of room reserved, the number of guests, and the payment method.

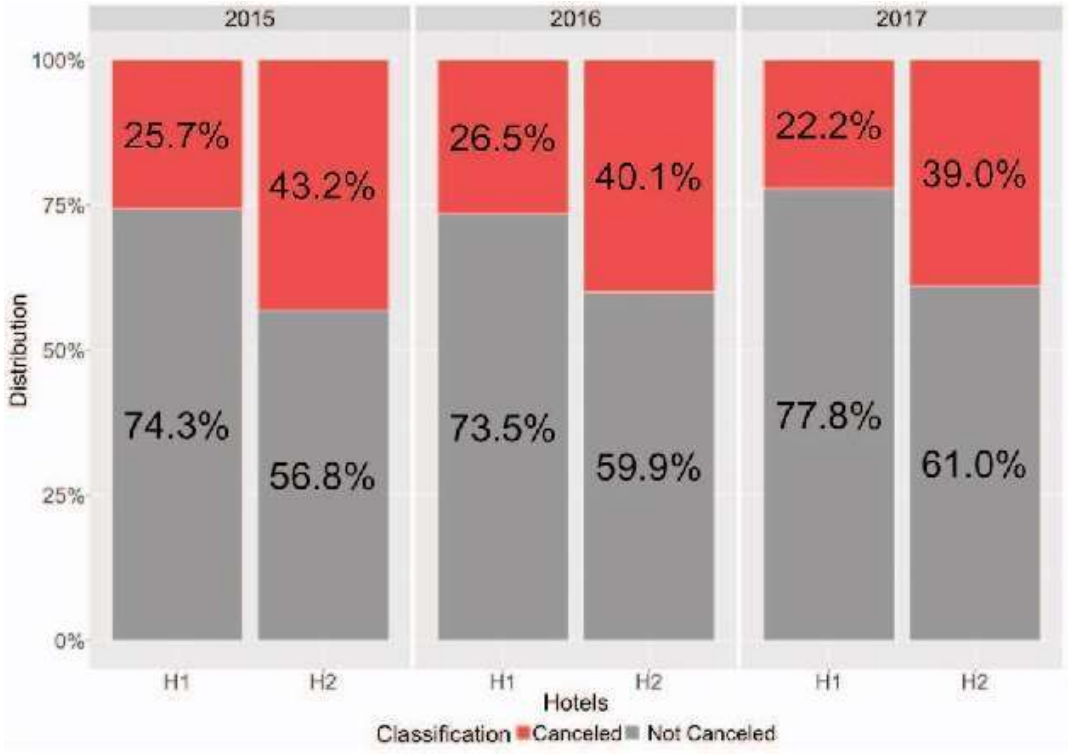
Another advantage of using a decision tree is that it is easy to interpret and visualize. This can be helpful for understanding which features are the most important for predicting cancellations and for communicating the results of the model to stakeholders. Decision trees can also be easily modified and updated as new data becomes available, making them a flexible choice for this type of project.

A decision tree can be a useful tool for predicting reservation cancellations because of its ability to handle diverse input features, its interpretability, and its flexibility.

Flowdiagram/Model:

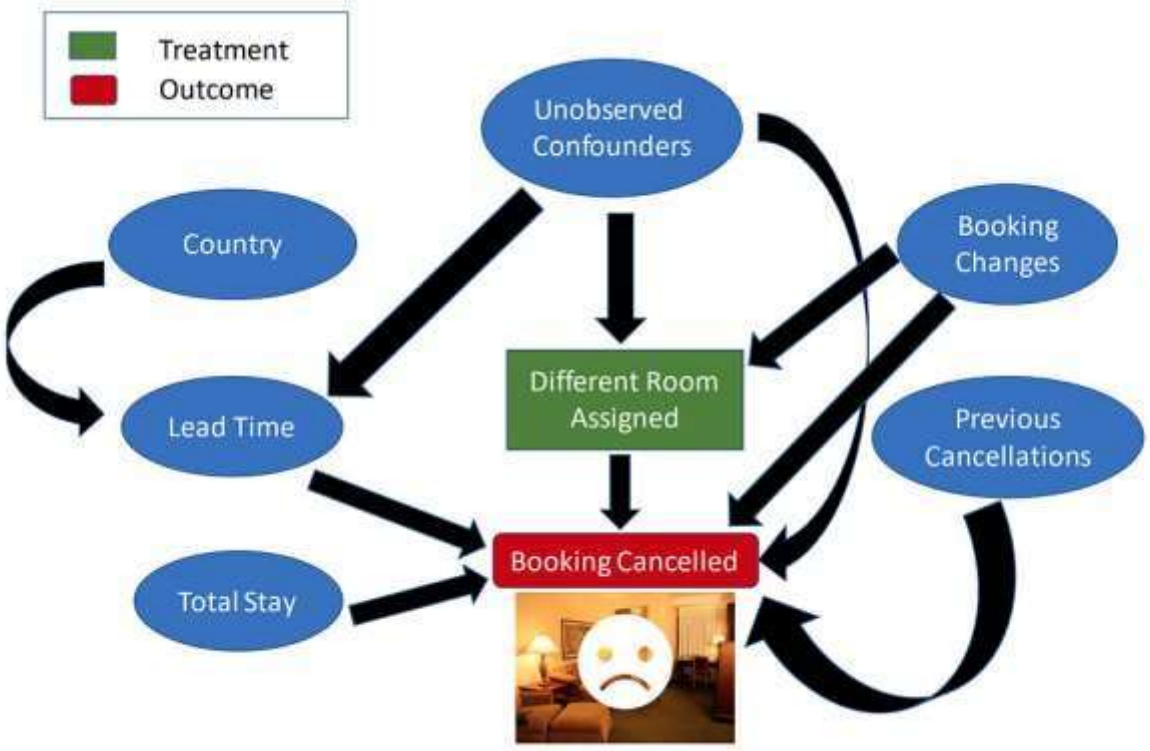


Visualization



This depicts the distribution of cancellation of the hotels per year.

Visualization



The above figure represents the predictions for the hotel rooms cancellation.

Data Description and Potential Statistical Tests

- There are several methods that can be used for predicting hotel booking cancellations in a machine learning project. Some of these methods include:

1. Logistic Regression: This is a statistical method used to analyze a dataset and identify the relationships between the input features and the output variable (i.e., cancellation). It can be used to predict the probability of a booking being canceled.
2. Decision Trees: This is a method that uses a tree-like model to make predictions. Each node in the tree represents a decision based on an input feature, and the leaf nodes represent the output (i.e., cancellation).
3. Random Forest: This is an ensemble method that combines multiple decision trees to improve the accuracy of the predictions.
4. Neural Networks: This is a type of deep learning algorithm that can be used for prediction tasks. It involves creating a network of interconnected nodes (i.e., neurons) that can learn to make predictions based on the input data.
5. Support Vector Machines (SVM): This is a method that tries to find the best boundary between two classes of data (i.e., canceled and not canceled bookings). It can be used to predict whether a booking will be canceled or not based on the input features.

Project Design and Milestones:

For this analysis we are going to use Python, because it contains many libraries and built-in algorithms to perform testing. It best suits for the exploratory data analysis and data analysis projects. Software we use is Google colab/Jupyter/Spyder.

- We are going to use Python 3 to its following packages:

- **Pandas:** Pandas is an open-source Python library used for data manipulation and analysis. It provides a wide range of functions and tools for handling and manipulating structured data, particularly in the form of tables or "data frames".
- **Matplotlib:** Matplotlib is a data visualization library in Python used to create high-quality static, animated, and interactive visualizations in Python. It provides a wide range of functions for creating various types of charts and plots, such as line plots, scatter plots, bar plots, histogram, heatmaps, and more.

Libraries:

- **Seaborn:** Seaborn is a data visualization library in Python that is built on top of Matplotlib. It provides a higher-level interface for creating attractive and informative statistical graphics. Seaborn makes it easy to create a variety of visualizations such as scatter plots, line plots, bar plots, heatmaps, and more.
- **Sklearn:** Scikit-learn (sklearn) is a popular open-source Python library for machine learning. It provides a variety of algorithms and tools for data preprocessing, feature selection, dimensionality reduction, model selection, and performance evaluation.
- **Pycountry**: PyCountry is a Python library that provides a way to access ISO databases containing country-related information. It can be used to retrieve information about countries such as their name, official name, languages spoken, currencies used, and more.

Importing the dataset

```
## Importing Data
from google.colab import files
uploaded = files.upload()
data = pd.read_csv('hotel_bookings.csv')
```

Choose Files hotel_bookings.csv

- **hotel_bookings.csv**(text/csv) - 16855599 bytes, last modified: 2/13/2020 - 100% done

Saving hotel_bookings.csv to hotel_bookings.csv

```
[3] ## Showing the first 5 rows of Data
data.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

Data Preprocessing

```
[4]  ## Data Preprocessing
     ## Copy the dataset
     df = data.copy()
```

```
[5]  ## Find the missing value
     df.isnull().sum().sort_values(ascending=False)[:10]
```

```
company                 112593
agent                    16340
country                    488
children                     4
reserved_room_type           0
assigned_room_type           0
booking_changes              0
deposit_type                 0
hotel                        0
previous_cancellations       0
dtype: int64
```

```
[6]  ## Drop Rows where there is no adult, baby and child
     df = df.drop(df[(df.adults+df.babies+df.children)==0].index)
```

Visualization of Data

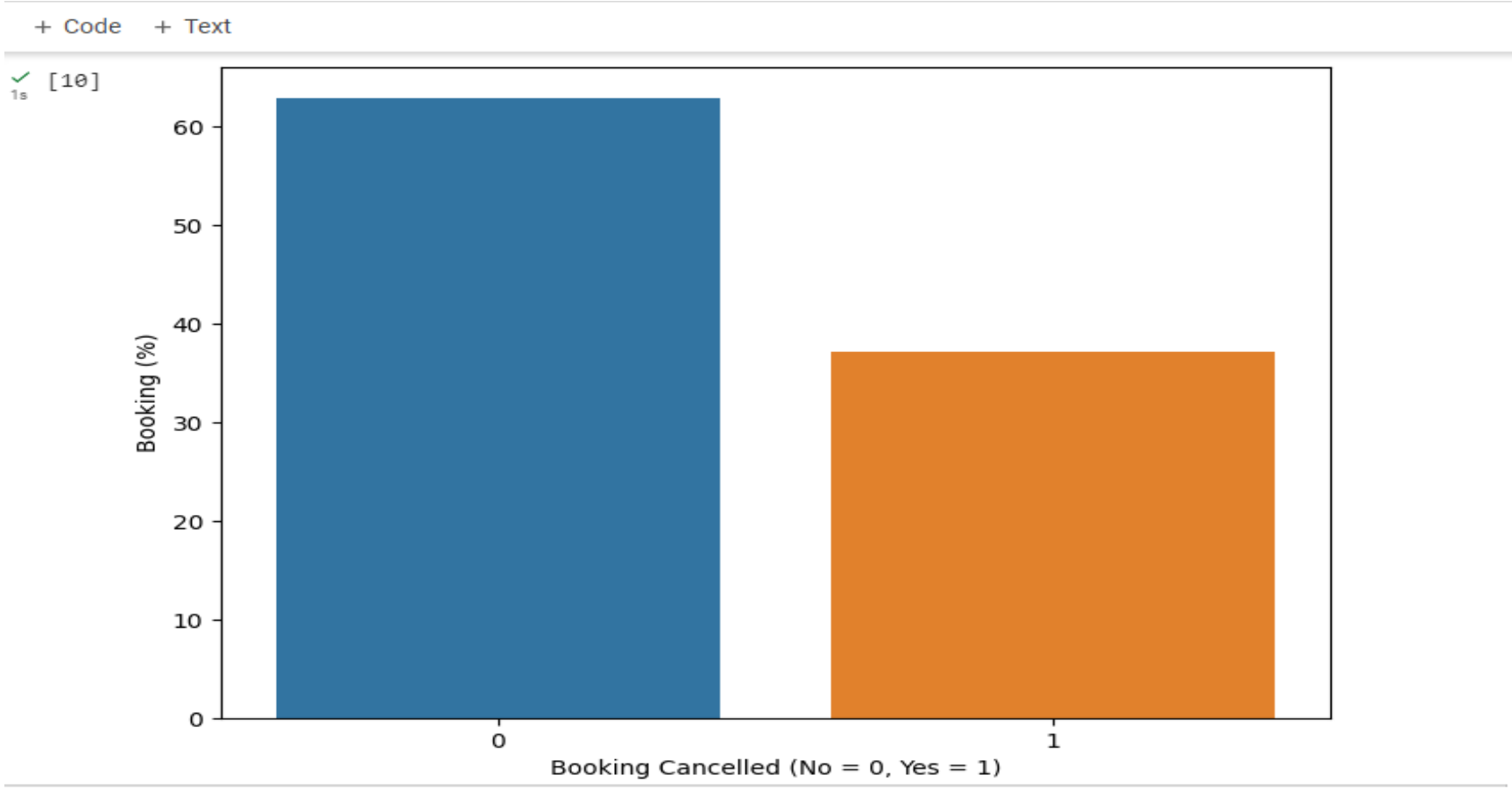
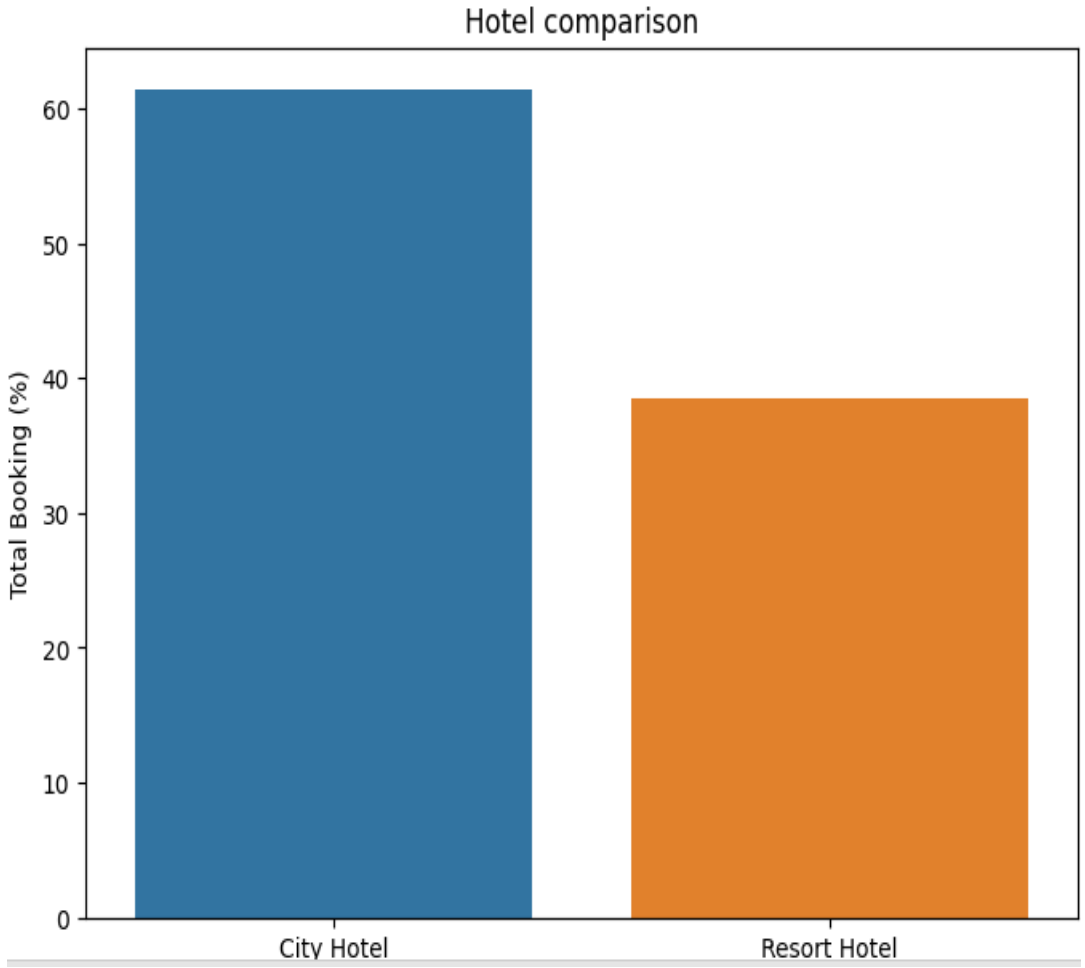
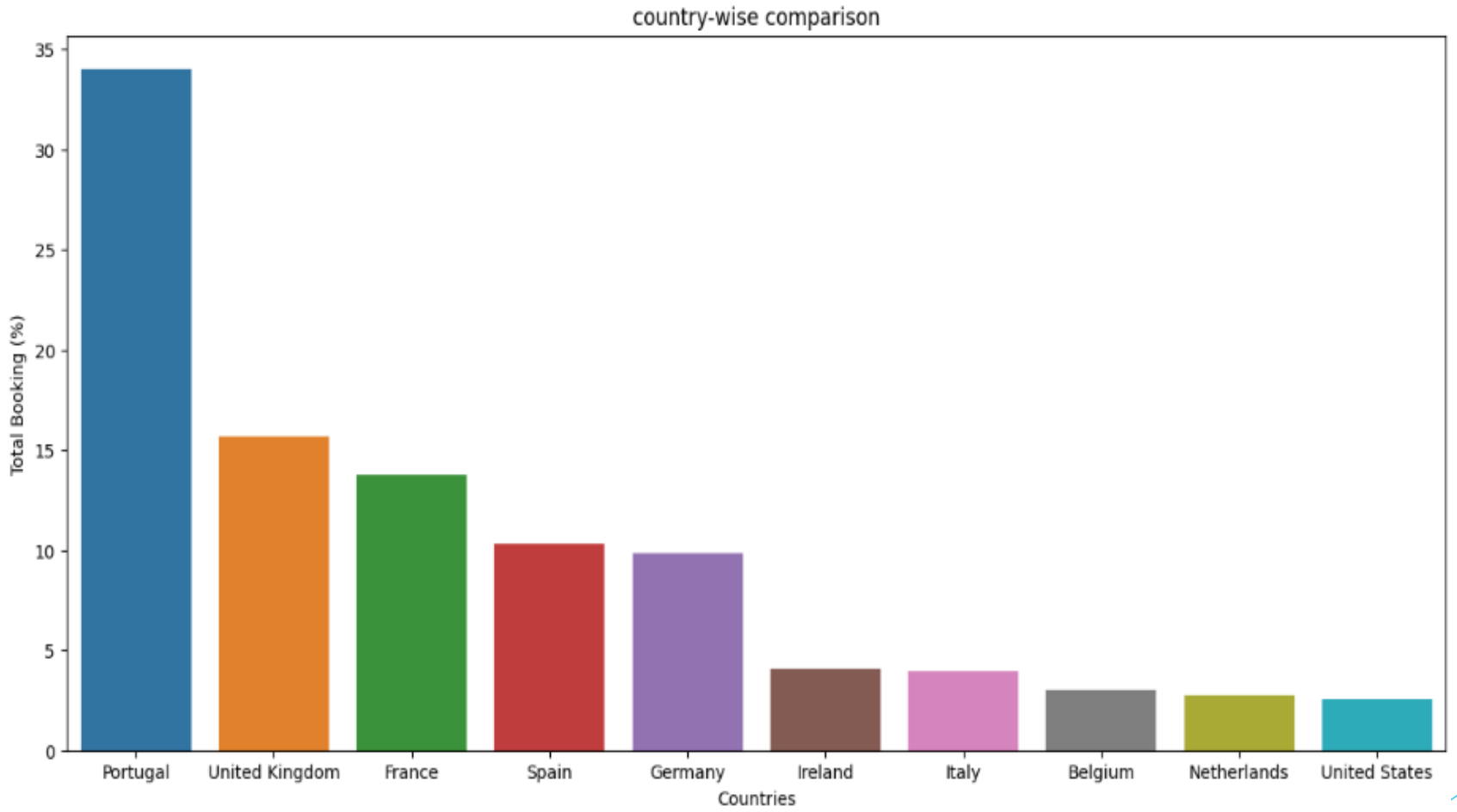


Fig. Graph to check how many bookings cancelled

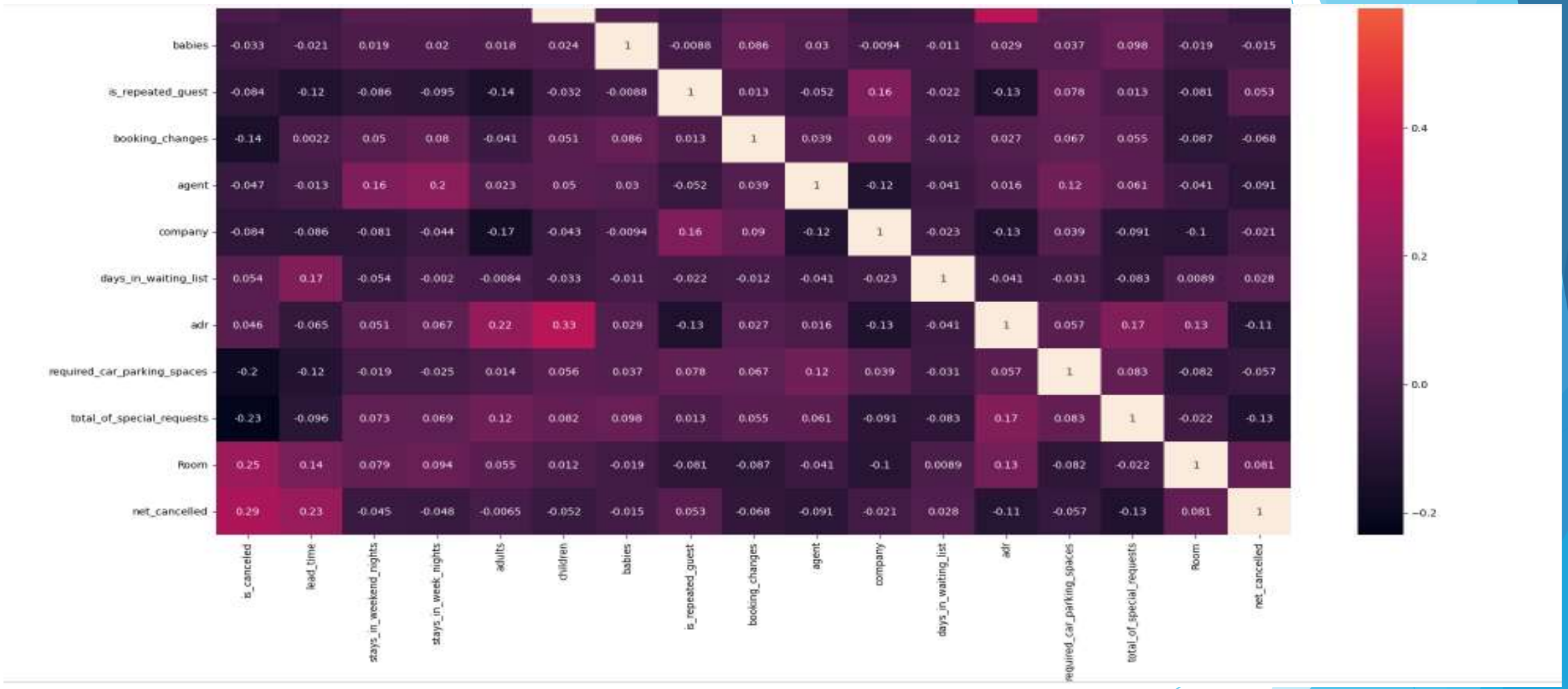
Comparison between type of hotels



Calculating total number of bookings country-wise



Plotting the heat map for the columns



T-test Results

Reservation_Cancellation (1).ipynb

File Edit View Insert Runtime Tools Help Last saved at 14:50

+ Code + Text

```
[ ]  ## t-test
     from scipy.stats import ttest_ind

     # Perform t-test on the extracted groups of data
     t_statistic, p_value = ttest_ind(group1_data, group2_data, equal_var=False)

     # Print results
     print("T-statistic:", t_statistic)
     print("P-value:", p_value)

     # Define significance level
     alpha = 0.05

     # Check for statistical significance
     if p_value < alpha:
         print("Reject null hypothesis. There is a statistically significant difference between the two groups based on the attribute.")
     else:
         print("Fail to reject null hypothesis. There is no statistically significant difference between the two groups based on the attribute.")
```

```
T-statistic: -3.4018430874330403
P-value: 0.000669595825894014
Reject null hypothesis. There is a statistically significant difference between the two groups based on the attribute.
```

```
[ ]  ## Modelling
     ## Converting Categorical variables to numerical
     def transform(dataframe):


         ## Import LabelEncoder from sklearn
         from sklearn.preprocessing import LabelEncoder

         le = LabelEncoder()
```

Evaluation of the Model

```
[30] def Score(clf,x_train,y_train,x_test,y_test):
         train_score = clf.score(x_train,y_train)
         test_score = clf.score(x_test,y_test)

         print("========================================")
         print(f'Training Accuracy of our model is: {train_score}')
         print(f'Test Accuracy of our model is: {test_score}')
         print("========================================")


     Score(clf,x_train,y_train,x_train,y_train)
```

```
========================================
Training Accuracy of our model is: 0.9956043710224032
Test Accuracy of our model is: 0.9956043710224032
========================================
```

Conclusion

- To Conclude our project that we are finding the accuracy values to help the owners of the hotel by the better understanding of the customers whether they are leaving the hotel or staying in the hotel for this we used the decision tree algorithm and we predict the training and testing accuracy values for our algorithm.

- Overall, the development of reservation cancellation prediction models has been a valuable tool for businesses in the hospitality industry. By reducing the number of cancellations, these models can help hotels and other businesses improve their profitability and customer satisfaction.

Resources and Related Projects:

- Hotel Booking Cancellation Prediction using ML algorithms

  Published By: M. Venkata Rakesh; S. Prasanna Kumar; Yogitha; R Aiswarya

  https://ieeexplore.ieee.org/document/9742843

Short description: Cancellation of hotel bookings can have a negative effect on the hospitality industry and can be a challenging issue for management decisions. To avoid the negative consequences of cancellations, hotels often implement policies and overbooking techniques, but these can ultimately harm the hotel's reputation and income. To address this issue, machine learning models have been developed that use past data to predict whether a booking is likely to be canceled or not. In this project, two hotels, the Resort hotel and the City hotel, were examined to see how certain management actions impacted revenue and cancellations. The ML models were able to help management anticipate the number of cancellations that might occur, leading to a more thoughtful approach to policies and decision-making.

Resources and Related Projects:

- Prediction of Hotel Booking Cancellation using CRISP-DM

  Published By: Zharfan Akbar Andriawan; Satriawan Rasyid Purnama; Adam Sukma Darmawan; Ricko

  https://ieeexplore.ieee.org/document/9299011

Short description: The online travel industry has experienced significant growth in recent years, with digital transactions related to online travel reaching USD 755.4 billion in 2019. The tourism and hospitality industry is an important sector within this market, with online reservation systems being a popular choice for hotel bookings. However, cancellations of online hotel reservations can create challenges for hotel management and result in lost revenue. To address this issue, data science can be used to predict whether a booking will be canceled or not, thereby allowing hotels to minimize their financial losses. In this study, the CRISP-DM framework was used to analyze datasets related to hotel booking requests and a tree-based algorithm was employed to make predictions. The results showed that the Random Forest model was the most accurate, with a value of 0.8725, and that the time difference between the booking and arrival date was the most important feature in predicting the likelihood of cancellation.

Resources and Related Projects:

- Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue

  Published by: Nuno Antonio, Ana Maria De Almeida, Luis Nunes.

  https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue

  Short description: Cancellation of hotel bookings can have a significant impact on demand management decisions in the hospitality industry. Cancellations can make it difficult to accurately forecast demand, which is a critical aspect of revenue management. To address this issue, hotels often implement strict cancellation policies and overbooking strategies, which can also have a negative impact on revenue and reputation. This study used data from four resort hotels to demonstrate that it is possible to build accurate prediction models for booking cancellations using data science. By treating booking cancellation as a classification problem, the authors were able to develop models that predicted cancellations with over 90% accuracy, showing that it is indeed possible to predict cancellations with a high degree of accuracy. This allows hotel managers to more accurately predict net demand, improve cancellation policies, and implement better overbooking tactics, ultimately leading to more effective pricing and inventory allocation strategies.