2023
2024

Foundations of Data Science Project Report

A COMPREHENSIVE ANALYSIS ON HOTEL BOOKINGS

BY SANSKRUTI JOSHI

TY BSC COMPUTER SCIENCE - A | Roll no.: 233331028

Progressive Education Society's

MODERN COLLEGE OF ARTS, SCIENCE AND COMMERCE

GANESHKHIND, PUNE – 411016

CERTIFICATE

This is to certify that **Sanskruti Joshi** of **TY BSC (Computer Science)** completed the project work titled **"Analysis on hotel bookings"** for the curriculum of Savitribai Phule Pune University during the academic year **2023-2024.**

INDEX

INTRODUCTION

Data Science

Data science is a field that involves using statistical and computational techniques to extract insights and knowledge from data. It encompasses a wide range of tasks, including data cleaning and preparation, data visualization, statistical modelling, machine learning, and more. Data scientists use these techniques to discover patterns and trends in data, make predictions, and support decision-making. They may work with a variety of data types, including structured data (such as numbers and dates in a spreadsheet) and unstructured data (such as text, images, or audio). Data science is used in a wide range of industries, including finance, healthcare, retail, and more.

Data science is a multidisciplinary field that uses statistical and computational methods to extract insights and knowledge from data. It involves a combination of skills and knowledge from various fields such as statistics, computer science, mathematics, and domain expertise.

The process of data science involves several steps, including data collection, cleaning, exploration, analysis, and interpretation. These steps are often iterative, and the process may be refined based on the results obtained.

Dataset

In tourism and travel related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry. However, the remaining tourism and travel industries like hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry's specific data. Hence, two hotel datasets with demand data are shared to help in overcoming this limitation.

The datasets were collected aiming at the development of prediction models to classify a hotel booking's likelihood to be cancelled.

This dataset contains information on records for client stays at hotels. More specifically, it contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

The dataset was obtained from IBM website, a free platform that stores datasets for data analysis. It was extracted in CSV format. The data format is mixed and not pre-processed. The specific subject area of this project is Revenue Management.

Each observation represents a hotel booking. Both datasets comprehend bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings that effectively arrived and bookings that were cancelled. Since this is hotel real data, all data elements pertaining hotel or costumer identification were deleted. Due to the scarcity of real business data for scientific and educational purposes, these datasets can have an important role for research and education in revenue management, machine learning, or data mining, as well as in other fields.

DATA ANALYSIS

Commonly, there are 4 steps in data science, namely, Data Collection, Data Preprocessing, Data Analysis (most commonly, EDA), and Data Interpretation.

Data collection

Data collection is the process of collecting and evaluating information or data from multiple sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. It is an essential phase in all types of research, analysis, and decision-making, including that done in the social sciences, business, and healthcare.

Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity.

For the project, the data was gathered from IBM. The data was already gathered in tabular form. Hence, after the dataset was extracted in CSV format, it is imported in the Python IDE, here, Google Colab.

After the dataset was safely imported, basic details about it were gathered, such as number of rows and columns, number of records, datatypes of each column, and so on.

Some statistical details were also included, like mean, median, standard deviation, and so on and so forth.

The important points to note from this step is that there are almost 12,000 records and more than 30 attributes. Such a vast dataset was selected for better accuracy and variability in graphs.

Data preprocessing

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.
Missing values were detected and removed from the dataset. Records having Nan values were replaced by the corresponding attribute's mean. Duplicate fields were also dropped.

Data Integration: This involves combining data from multiple sources to create a unified dataset.
In the dataset, there were four fields with similar information so they were combined into one and the resulting duplicate columns were removed. Here, 'date' was divided into four fields. Since the separation served no specific purpose, they were combined into one. In this process, Dimensionality reduction was also performed parallelly.

Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.
Depending on the field datatype, the appropriate data transformation technique was used to convert the respective attribute.
Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which

only take numerical data. Hence, Label Encoder was used on categorical variable, hotel.

One hot encoding is a technique that is also used to convert categorical variables into numeric datatype. In most scenarios, one hot encoding is the preferred way to convert a categorical variable into a numeric variable because label encoding makes it seem that there is a ranking between values. Hence, this is applied on Reservation Status attribute.

Data Reduction: This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

Data Discretization: This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering. In the dataset, Lead Time was categorised using equal frequency binning, into "High" and "Low".

Data Normalization: This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

INTERPRETATION

EDA

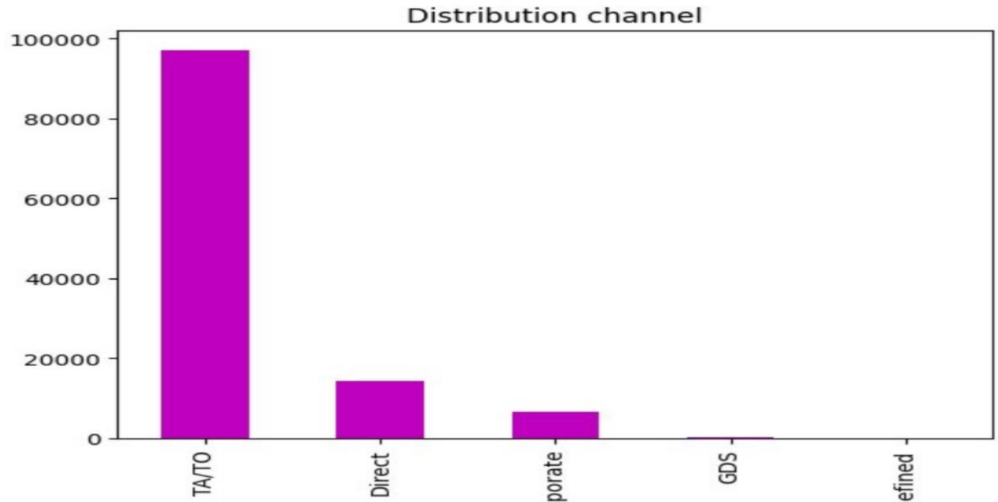
EDA, or Exploratory Data Analysis, refers back to the method of analysing and analysing information units to uncover styles, pick out relationships, and gain insights. There are various sorts of EDA strategies that can be hired relying on the nature of the records and the desires of the evaluation.

Data Visualization

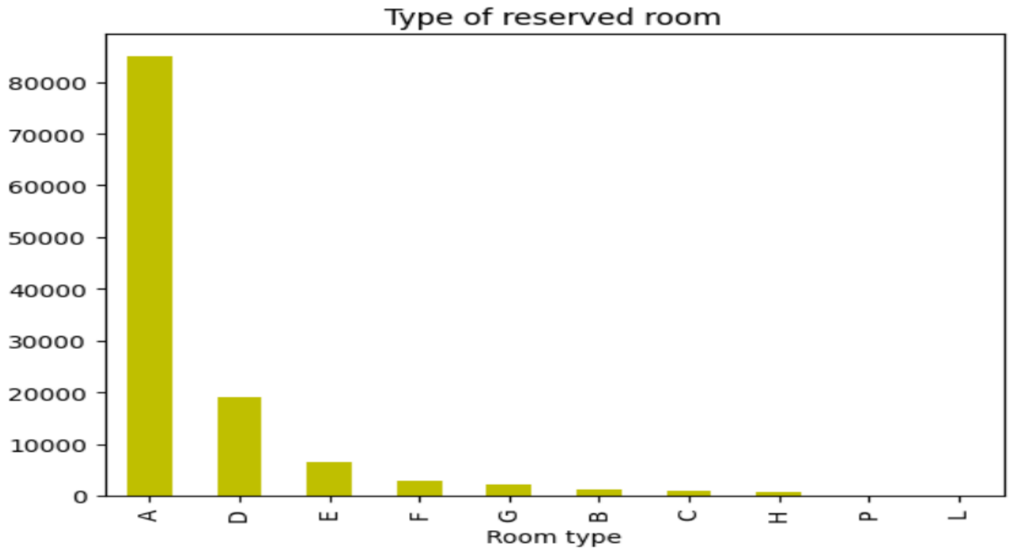
Data Visualization is the process of analysing data in the form of graphs or maps, making it a lot easier to understand the trends or patterns in the data.

Univariate Analysis: This sort of evaluation makes a speciality of analysing character variables inside the records set. It involves summarizing and visualizing a unmarried variable at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.

1. Bar graph: frequency chart for qualitative variables. Used to assess the most-occurring and least-occurring categories within a dataset.

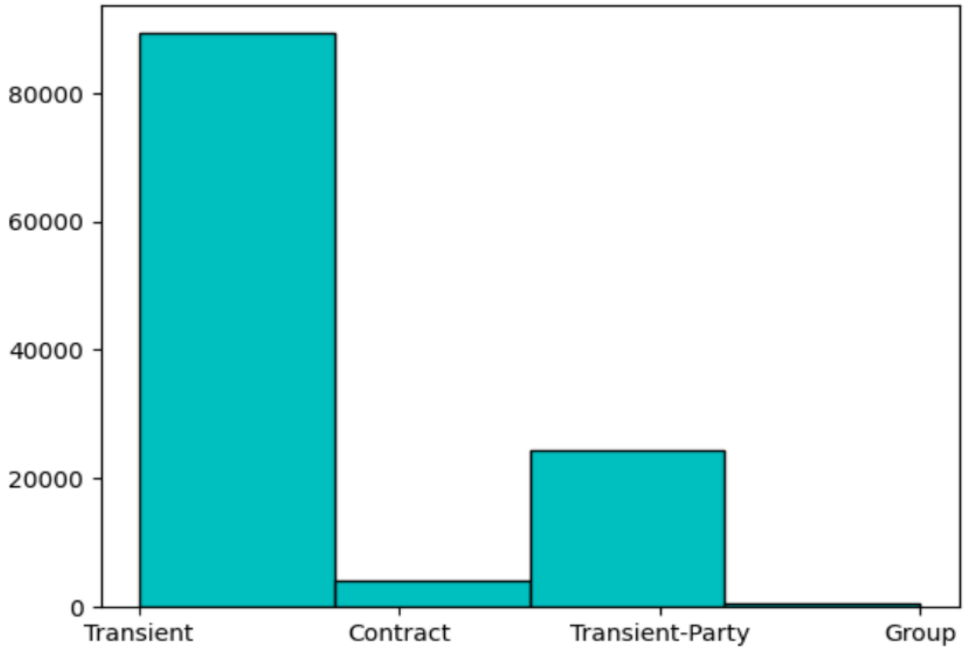


Interpretation of Distribution Channel: The TA/TO, that is, Travel Agents are the most frequently used medium for booking hotel appointments. Almost none of the guests prefer to go by themselves.

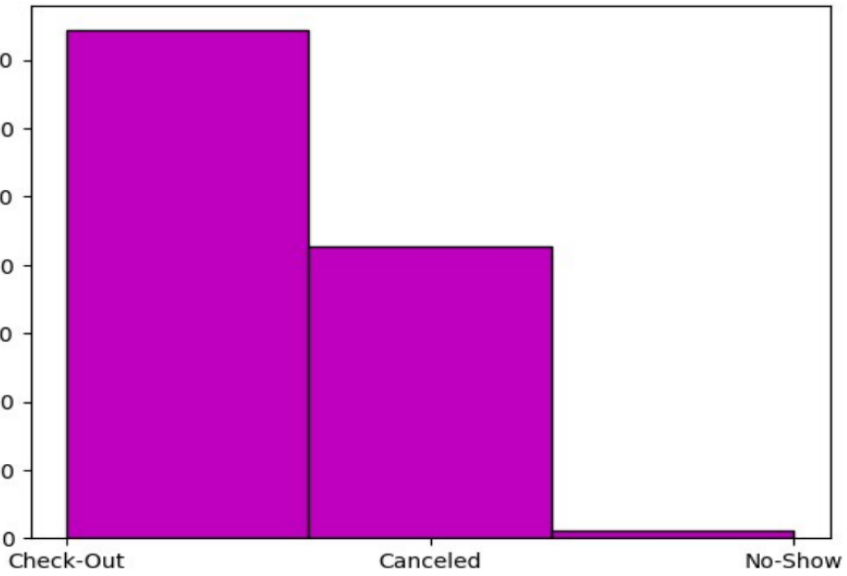


Interpretation of Reserved Room: The rooms are labelled alphabetically. It can be seen that room A is the most popular. Rooms P and L are never reserved and hence can be assumed to kept for last minute bookings.

2. **Histogram:** variation of a bar chart in which data values are grouped together and put into different classes.

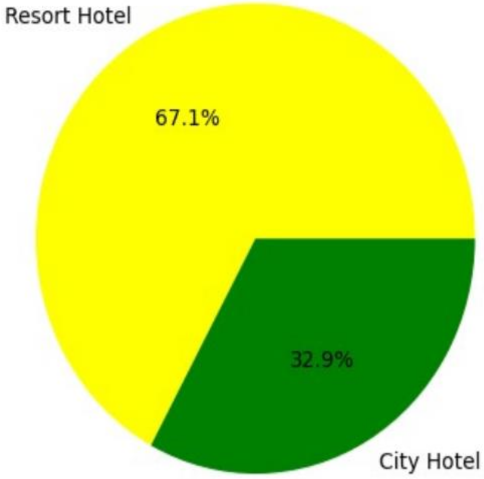


Interpretation of Customer Type: Skewed-left histogram. Transient customers are the most common, while contract and group customers are rare enough to be considered outliers.

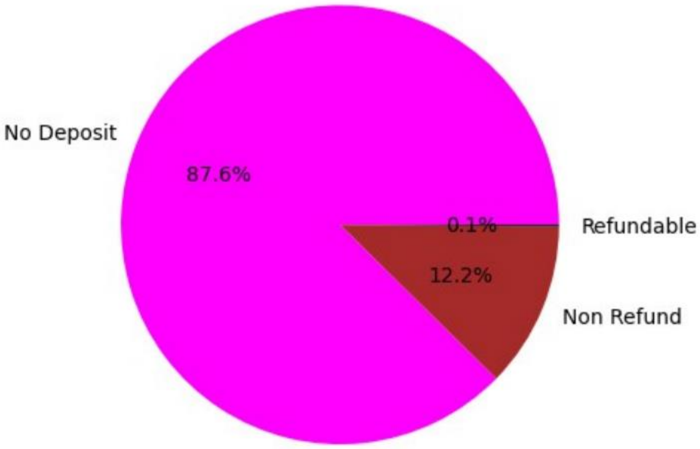


Interpretation of Reservation Status: Skewed-left histogram. Most of the customers transitioned through the normal process, namely, checking in and then checking out. A noticeable number of them cancelled but very few people were a no-show. So, most people tried to let the hotel know of their absence.

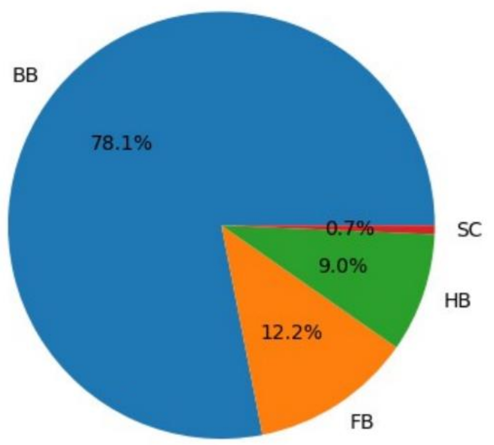
3. Pie chart: assess the relative sizes of categories to the entire dataset. At a minimum, pie charts require one categorical variable.



Interpretation of Hotel type: From the chart, nearly three-fourth of the total customers booked Resort Hotel. Hence, it is more popular and famous. This increases the chances of bookings in Resort Hotel.



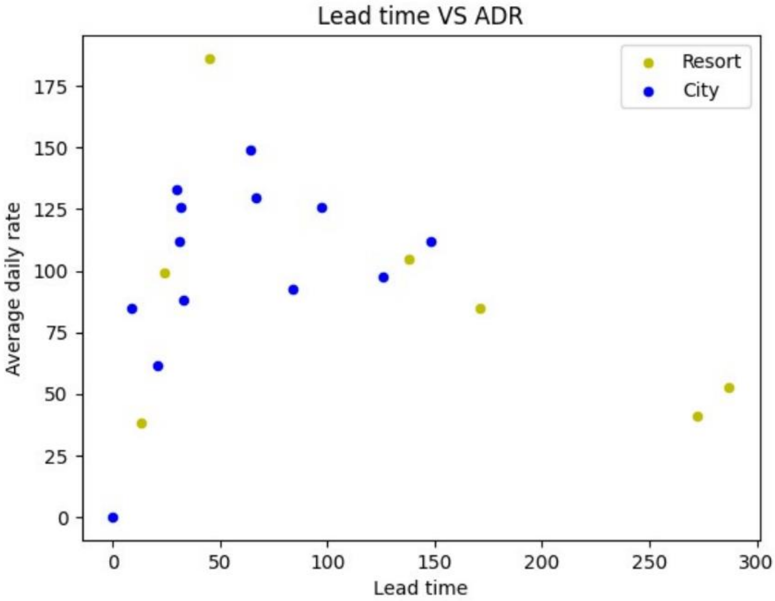
Interpretation of Deposit Type: Almost 90% of customers chose not to give a deposit. From the ones that did, not even 1% chose to give refundable deposits. Hence, not giving a deposit is clearly preferred, but the ones that do trust the hotel.



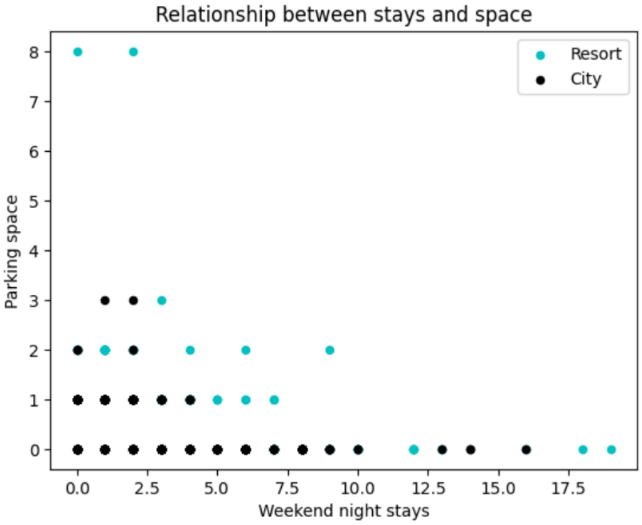
Interpretation of Meal type: Three quarters of customer want BB, that is, Bed and Breakfast. Almost 10% prefer Half Board and barely 1% likes to be Self Catered. This showcases people's preference for staying a night, having breakfast, and moving on.

Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.

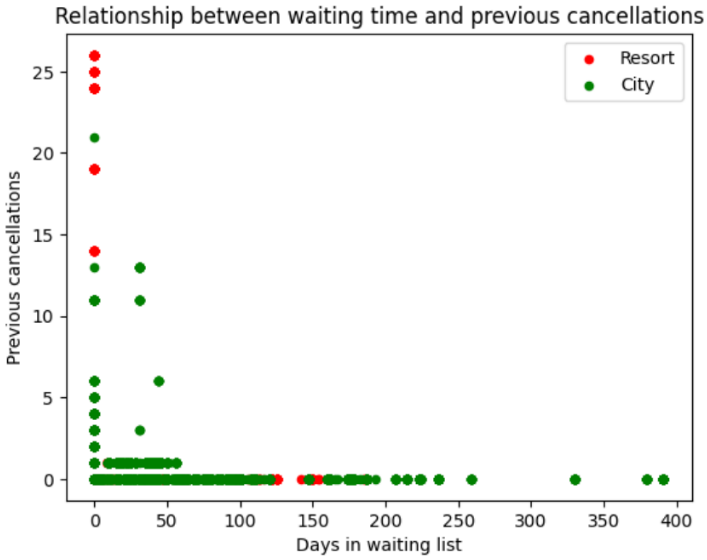
1. Scatter plot: show relationships between pairs of continuous variables. The pattern of dots on a scatterplot allows you to determine whether a relationship or correlation exists between two continuous variables. If a relationship exists, the scatterplot indicates its direction and whether it is a linear or curved relationship.



Interpretation: For City Hotel, there appears to be no correlation between ADR and lead time. The datapoints are clustered and averages about 100 for lead time and ADR. There is one outlier which has zero lead time and ADR. This could point towards a day when the hotel was closed. For Resort Hotel, there appears to be a weak negative correlation, with a few outliers. Hence, the variables have an inverse relationship for Resort Hotel.

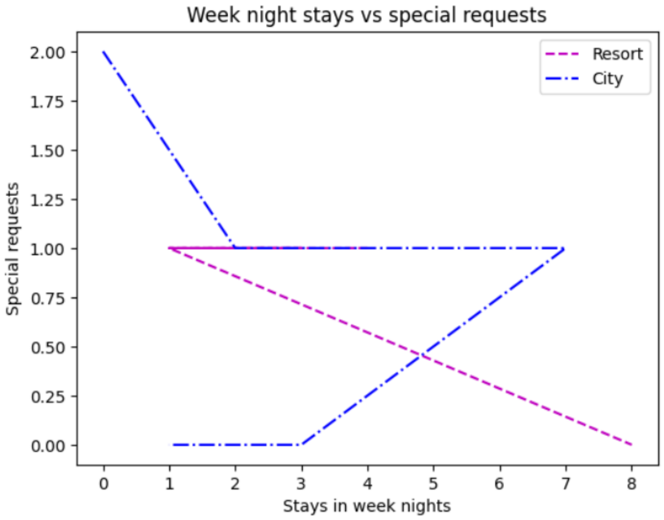


Interpretation: For City Hotel, most customers did not want car parking space, regardless of number of nights to stay. For Resort Hotel, there is a very weak negative correlation. Hence, people staying longer require less parking space than those whose duration is shorter.

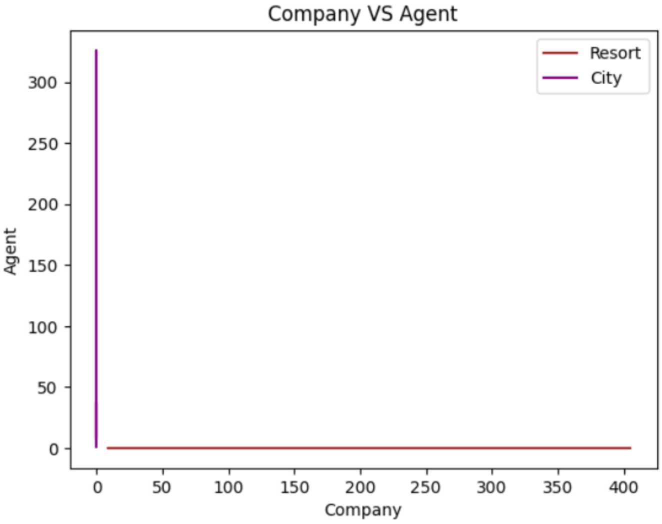


Interpretation: For City Hotel, with a few exceptions, that is, outliers, customers who did not previously cancel had to wait around 50 days. Hence, their algorithm for waiting did not take previous cancellations into account. For Resort Hotel, the ones who cancelled previously seem to wait the least, so perhaps they wish to give them a better experience.

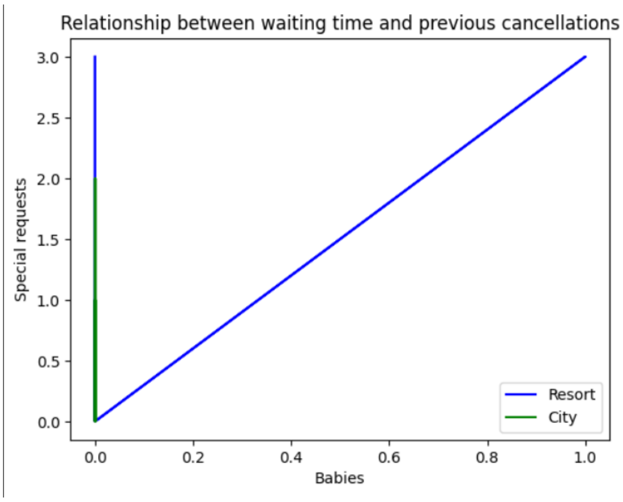
2. Line plot: track many types of measurements by time, values of another variable, and categories. Emphasize trends and patterns.



Interpretation: Resort Hotel has a strong negative correlation between the number of special requests and number of week nights of stay. City Hotel has different patterns in different sections. One can infer that on particular days, City's customers have more requests when the duration is longer, and on other days, its vice-versa.

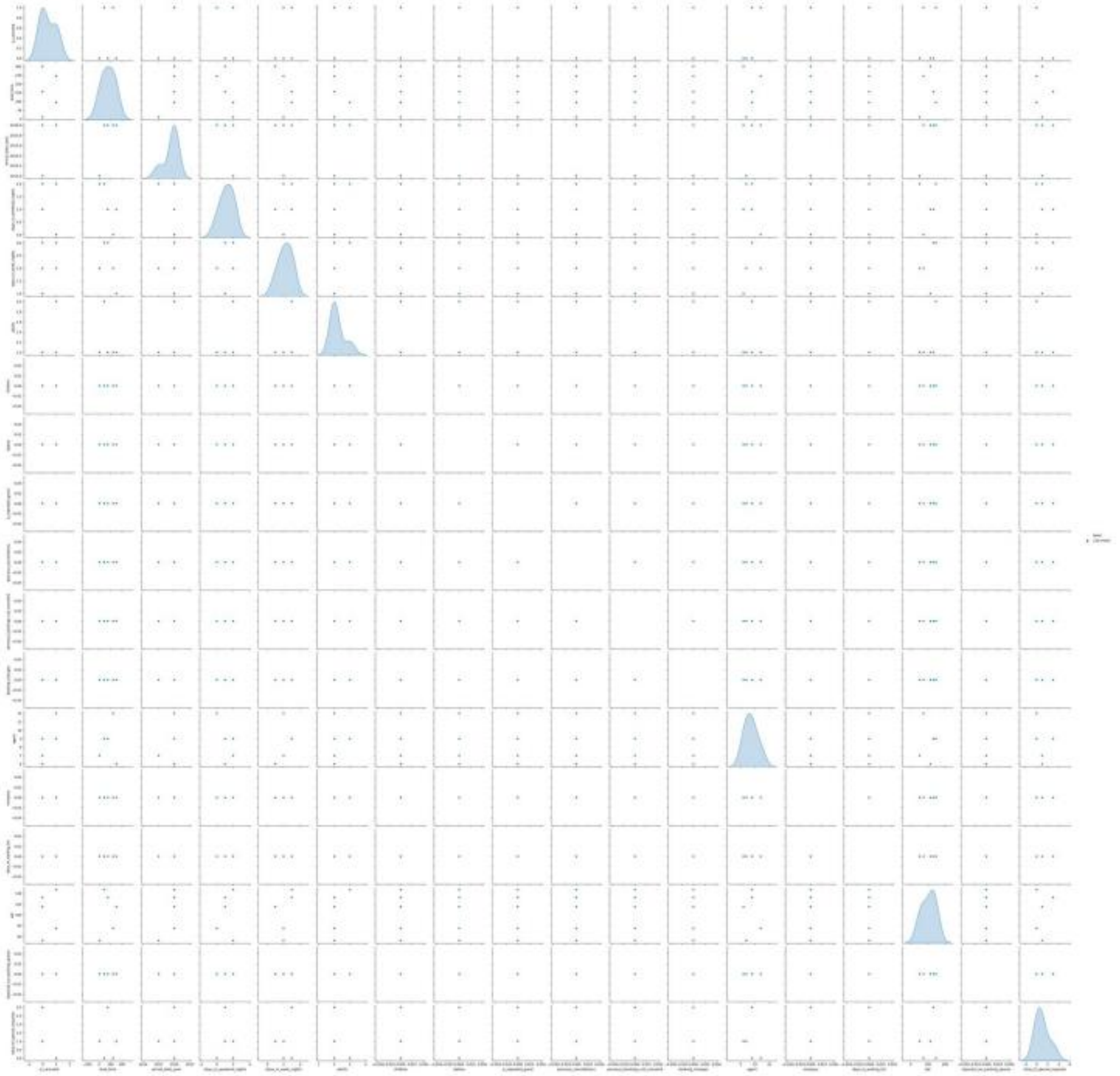


Interpretation: Customers of Resort Hotel very clearly prefer to make bookings through a company, while those of City Hotel favour agents.



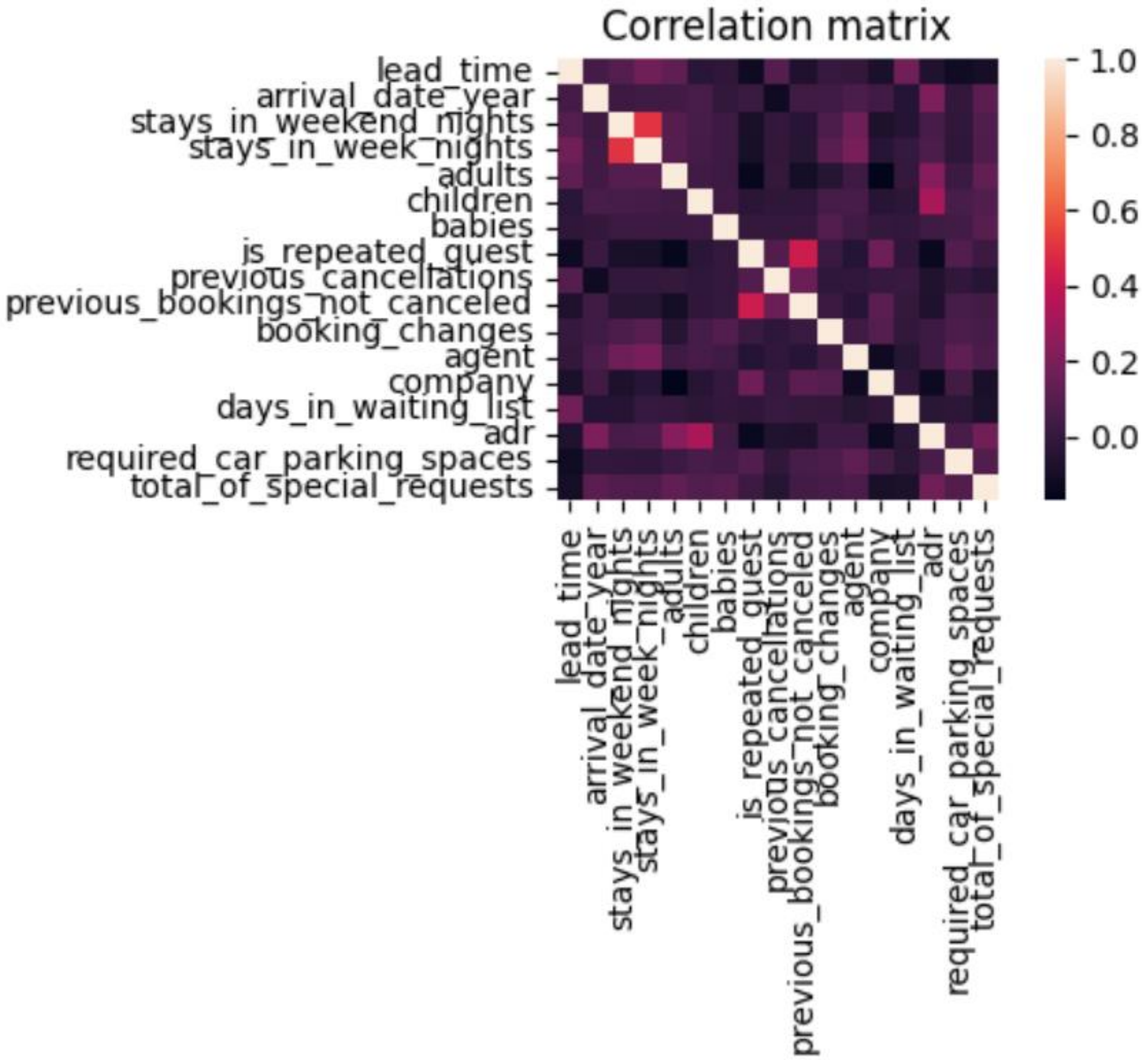
Interpretation: Resort Hotel, excluding some outliers, shows strong positive correlation between the number of special requests and babies. This shows that this hotel does not provide facilities needed to care for infants. City Hotel, on the other hand, shows that it is the people with no toddlers who ask for special requests the most, Multivariate Analysis: Multivariate analysis extends bivariate evaluation to encompass greater than variables. It ambitions to apprehend the complex interactions and dependencies among more than one variable in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.

1. **Pair plot:** Matrix of scatterplots that lets you understand the pairwise relationship between different variables in a dataset. Gives us an idea of the relationship between each pair of variables in our dataset.



Interpretation: Most attributes seem to be correlated to each other normally, or in a left-skewed relationship, that is, the mean is lesser than the median.

2. **Heat map:** represents the magnitude of individual values within a dataset as a colour. The variation in colour may be by hue or intensity.



Interpretation: There is a moderate correlation between number of weekend nights and week nights, repeated guests and the ones who did not previously cancel, children and ADR (Average Daily Rate). The rest of the attributes do not corelate to each other.

SCOPE OF PROJECT

- Descriptive analytics can be employed to further understand patterns, trends, and anomalies in data.

- Used to perform research in different problems like: bookings cancellation prediction, customer segmentation, customer satiation, seasonality, among others.

- Researchers can use the datasets to benchmark bookings' prediction cancellation models against results already known.

- Machine learning researchers can use the datasets for benchmarking the performance of different algorithms for solving the same type of problem (classification, segmentation, or other.

- Educators can use the datasets for machine learning classification or segmentation problems.

- Educators can use the datasets to obtain either statistics or data mining training.

CONCLUSION

The success factoring a profitable hotel industry has been changing over time, driven by global competition and increasingly high customer expectations. Hotels focus on customer satisfaction and to exceed customer expectations. There are factors in the study which affect the business of the hotels. Factors such as location, ADR, Deposits charged, wait time, etc. We also have channels like distribution channel, Market segment to focus on to get more revenue.

- Majority of people prefer A-room type so hotels should increase their numbers to get more revenue.
- Chances of cancellation is high when there are no deposits taken by hotels, so hotels should take minimum deposits to minimise the rate of cancellation.
- Transient customers cancels more often but when people book in groups it leads to lesser cancellations, hence hotels should provide some offers focusing transient customers to decrease cancellations.
- Maximum number of bookings are in the month May to August, so hotels should provide exciting deal to customers to increase their booking in off season. As hotels are getting less repeated customers so management should take customer's feedback and improve the hotel facilities to increase the count of their repeated guests.
- Every year there is 25-30% cancellation for resort hotels and 40-45% cancellation for city hotels.

BIBLIOGRAPHY

- Geeksforgeeks
- Wikipedia
- Medium
- Science direct
- Towards Data Science