**Hotel Booking Analysis**

**Alma Better Capstone Project**

**EXPLORATORY DATA ANALYSIS**

**Team member**

**Yashwant Raul**

**Mayur Marathe**

**Sanket Gawali**

**1. Abstract:**

Hotel industry is a very volatile industry and the bookings depend on variety of factors such  as type of hotels, location, seasonality, days of week and many more. This makes analyzing  the patterns available in the past data more important to help the hotels plan better. Using  the historical data, hotels can perform various campaigns to boost the business. We can use  the patterns to predict the future bookings using time series. we have done brainstorming to  find out insightful from the data by using various technique such as programming i.e. Python  (Pandas , Numpy, etc.), data wrangling, data visualization using libraries like seaborn, matplotlib etc.

**2. Problem Statement:**

Have you ever wondered when the best time of year to book a hotel room is? Or the optimal  length of stay in order to get the best daily rate? What if you wanted to predict whether or  not a hotel was likely to receive a disproportionately high number of special requests? This  hotel booking dataset can help you explore those questions!

This data set 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. All  personally identifying information has been removed from the data.

Explore and analyze the data to discover important factors that govern the bookings.

**3. Data Summary:**

  The main objective behind this project is to explore and analyze data to discover important factors that govern the bookings and give  insights to hotel management ,which can perform various campaigns to boost the business and performance

The main objective of this project is to find out which type of hotel is preferred by visitors with their special requirement. While analyzing we found that in total the data set contain below 1.2 lakh rows and 32 columns

The Dataset has 4 columns with float64 dtype, 16 columns with int64 dtype and 12 columns with object dtype

In the Dataset, we observed null values in the following columns:

4 null values in the children column

488 null values in the country column

16,340 null values in the agent column

112,593 null values in the company column

∙ **New columns added**

Total\_people i.e.: - adult +children’s +babies

Length\_of\_stay i.e.: - week\_night +weekend\_night

We have the following column names provided to us in the dataset. Let's understand the columns.

**hotel :** Defines the Type of the Hotels in the given data.

**Is\_canceled:** Defines the Status of the Booking ( Ex: Canceled )1 refers to Canceled and

0 suggests Not Canceled

**lead\_time:** Gives us the timing difference between the booking Time and the arrival  from the given data set 1.2 lakh rows and 32 columns

**arrival\_date\_year :** Represents the Year of Arrival of the Visitor (2015, 2016, 2017)

**arrival\_date\_month:** Represents the month of Guest ( Visitors ) Arrival From Jan To  Dec

**arrival\_date\_week\_number:** This represents the Week No. of The Visitors Arrival - 1  to 53

**arrival\_date\_day\_of\_month:** This gives the day number of month when the visitor  arrived - 1 to 31

**stays\_in\_weekend\_nights:** This gives the number of weekend nights, i.e.Saturday and  Sunday

**stays\_in\_week\_nights:** This gives the number of week nights, i.e. Monday to Friday **adults:** This gives the number of adults per booking

**children:** This gives the number of children per booking

**babies:** This gives the number of babies per booking

**meal:** This gives the type of meal preferred.

Undefined/SC means no meal package, BB means Bed & Breakfast, HB means Half  board (i.e., breakfast & one other meal – usually dinner), FB means Full board (i.e., breakfast, lunch & dinner).

**country:** This gives the country of origin of visitor .

**market\_segment:** This gives the group of people based on market .

Direct, Corporate, Online TA, Offline TA/TO, Complementary, Groups, Aviation Where,  TA: Travel Agents, TO: Tour Operators .

**distribution\_channel:** This mentions the type of distribution channel

Direct, Corporate, TA/TO, Undefined, GDS Features (cont.)

**is\_repeated\_guest:** This shows repeated customers 1 means repeated customer, 0  means not repeated.

**previous\_cancellations:** Represents the number of previous bookings that were  canceled by the customer before the current booking .

**previous\_bookings\_not\_canceled:** Represents the number of previous bookings not  canceled by the customer prior to the current booking.

**reserved\_room\_type:** Represents the type of room reserved 'C', 'A', 'D', 'E', 'G',  'F', 'H', 'L', 'P', 'B'

**assigned\_room\_type:** Represents type of room whose possession is given at the time of arrival. 'C', 'A', 'D', 'E', 'G', 'F', 'H', 'L', 'P', 'B'

**booking\_changes:** Represents the number of bookings changed

**deposit\_type:** Represents the types of deposit No Deposit, Non Refund, Refundable 24. **agent:** Represents the Agent Id.

**company:** Represents the Company Id.

**day\_in\_waiting\_list:** Represents the Number of days the booking was in the waiting  list before confirmation.

**customer\_type:** This Gives us Type of customer Contract, Group, Transient,  Transient-party.

**adr:** means average daily rate.

**required\_car\_parking\_spaces:** Number of car parking spaces required by the customer.

**total\_of\_special\_requests:** Number of special requests made by the customer.

**reservation\_status:** Status of reservation Canceled, Check-Out, No-Show. **reservation\_status\_date:** Date at which the last status was updated .

**4. Steps involved in the Data Analysis:**

* **Data Wrangling:** After loading the dataset, we performed this method by cleaning,  organizing, and transforming raw data into the desired format which makes us to understand  the data clearly. This process helped us to tackle the unwanted data, to produce accurate  results, to make better decision.
* **Null Value Treatment:** Our data set contains a small number of null values. wwe  used various method to. Which might tends to disturb our accuracy hence we dropped sand somes and fill null values by feature engineering at beginning of our project in order to get better result

Null values are present in agent, company children, & Country column. So we dropped columns agent, company which has 94.3 % Null values . We will replace null values in other column i.e.  Children, Country with feature Engineering.

Country  Column has total 195 contraries in the world. we will replace null values of country column with mode value

* **Standerdisation of Features:** Our main motive through this step is scale out data into uniform format that would allow to us to utilize the data in better way while performing different operations
* **EDA analysis:** Exploratory Data Analysis is one of the most efficient methods used  to Analyze the given data sets. After loading the dataset, we performed many methods by  comparing our target variable that is booking analysis with other independent variables. Using  the exploratory data analysis, we can summarize the characteristics of the data sets  which are important, and in this EDA Capstone Project we have made use of the  statistical graphs and the other visualization methods such as. . Bar Plot, Line Plot , Count Plot, Pie Chart, Heatmap etc.
* **Visualization:** Visualization is one of the most efficient methods of data  representation in readable plots and very easy to interpreted.

**5. Exploratory Analysis:**

The preliminary analysis of data to discover relationships between measures in the data and to gain an insight on the trends, patterns, and relationships among various entities present in the data set with the help of statistics and visualization tools is called Exploratory Data Analysis (EDA).

Exploratory data analysis is cross-classified in two different ways where each method is either graphical or non-graphical. And then, each method is either univariate, bivariate or multivariate.

**Univariate Analysis**

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

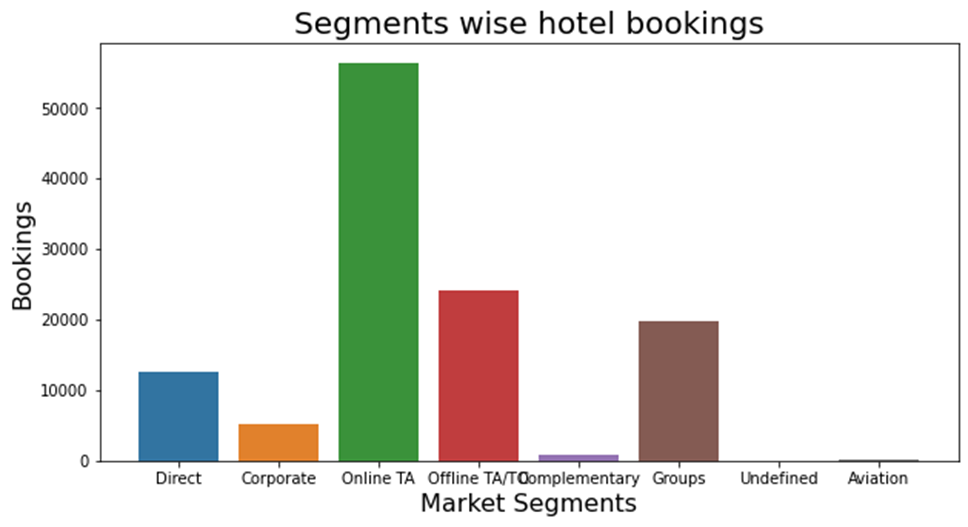
 Univariate data can be described through:

* **Frequency Distribution Tables**

The frequency distribution table reflects how often an occurrence has taken place in the data. It gives a brief idea of the data and makes it easier to find patterns.

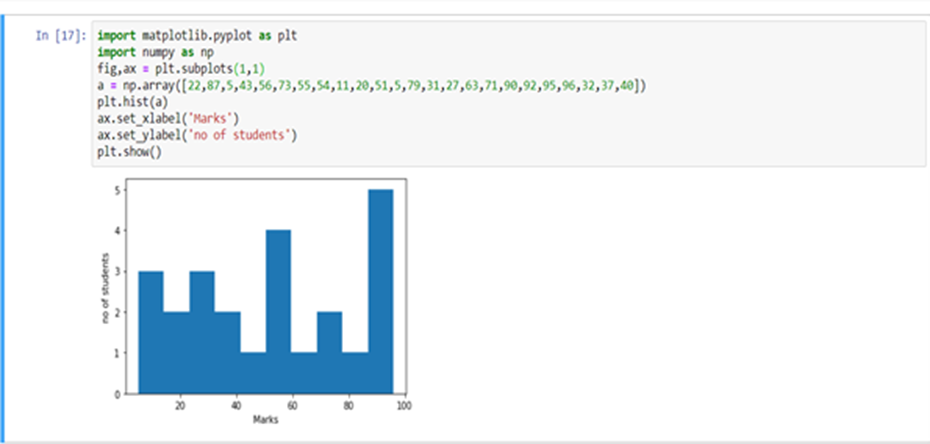
* **Bar Charts**

The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.



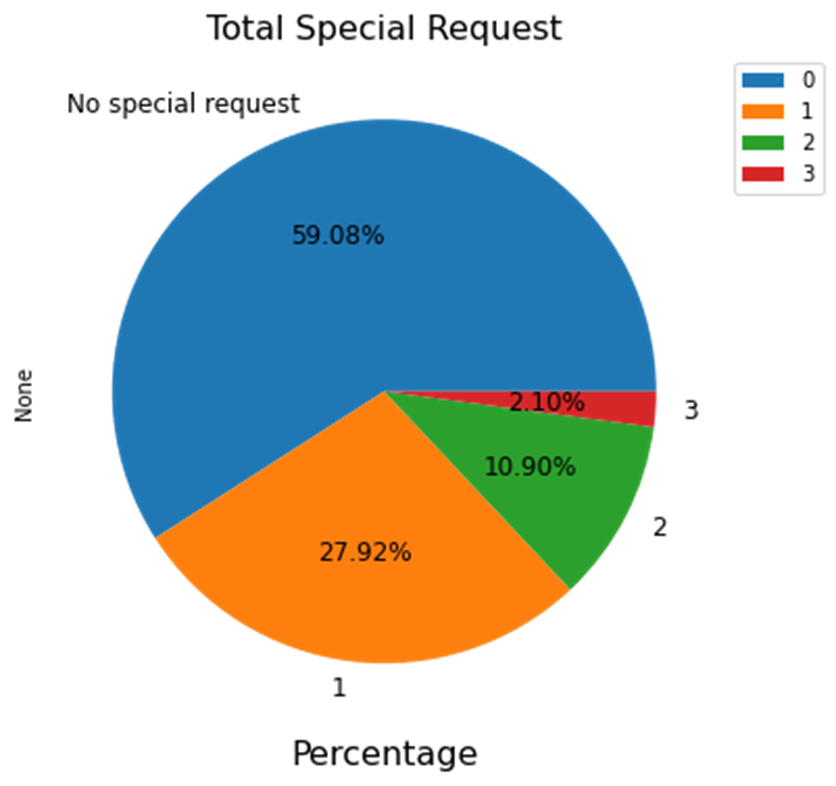
* **Histograms**

Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.



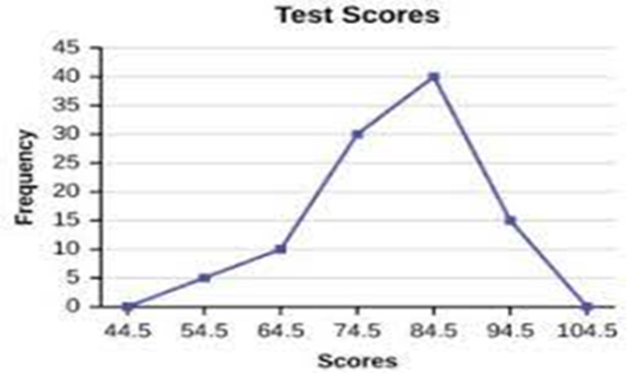
* **Pie Charts**

Pie charts are mainly used to comprehend how a group is broken down into smaller pieces. The whole pie represents 100 percent, and the slices denote the relative size of that particular category.



* **Frequency Polygons**

Similar to histograms, a frequency polygon is used for comparing datasets or displaying the cumulative frequency distribution.



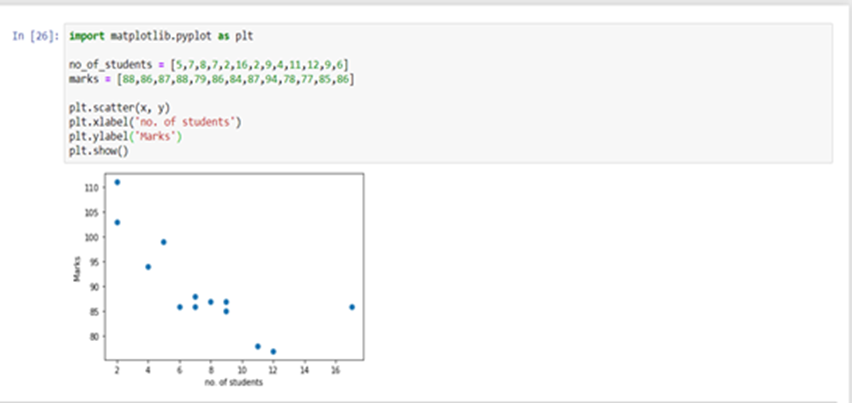
**Bivariate Analysis**

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables. There are three types of bivariate analysis.

**Bivariate Analysis of two Numerical Variables**(Numerical-Numerical)

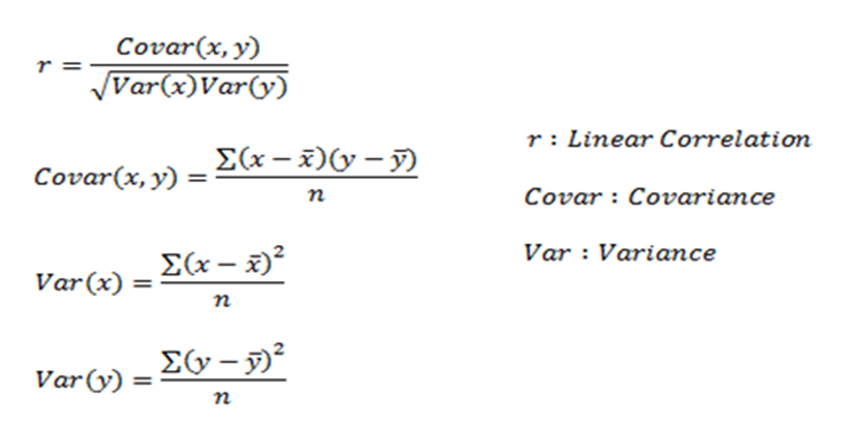
* **Scatter Plot**

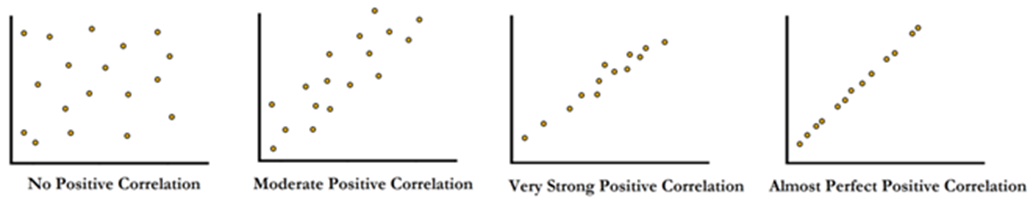
A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or nonlinear) and strength of the relationship between two variables.



* **Linear Correlation**

Linear **C**orrelation represents the strength of a linear relationship between two numerical variables. If there is no correlation between the two variables, there is no tendency to change along with the values of the second quantity.



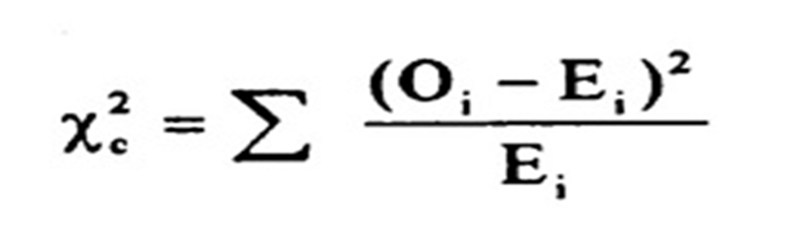
Here, r measures the strength of a linear relationship and is always between -1 and 1 where -1 denotes perfect negative linear correlation and +1 denotes perfect positive linear correlation and zero denotes no linear correlation.

**Bivariate Analysis of two categorical Variables** (Categorical-Categorical)

* **Chi-square Test**

The chi-square test is used for determining the association between categorical variables. It is calculated based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table. A probability of zero indicates a complete dependency between two categorical variables and a probability of one indicates that two categorical variables are completely independent.

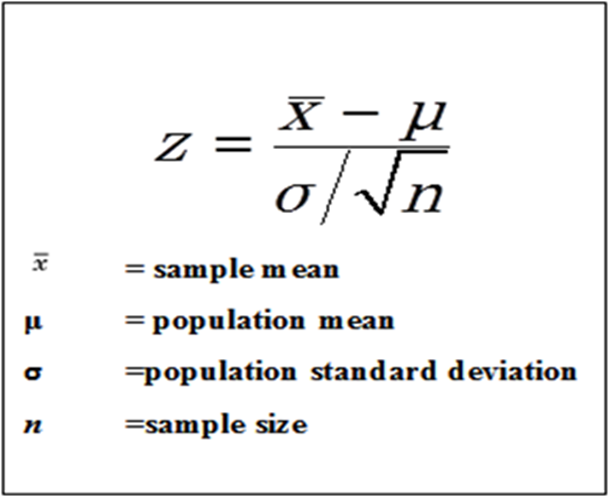
Here, subscript c indicates the degrees of freedom, O indicates observed value, and E indicates expected value.

****

**Bivariate Analysis of one numerical and one categorical variable** (Numerical-Categorical)

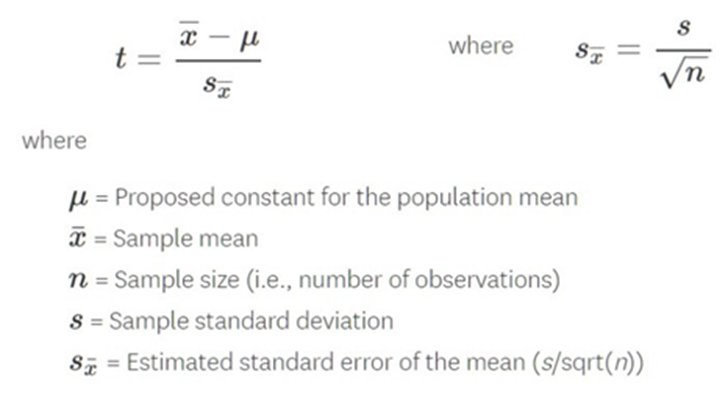
* **Z-test and t-test**

  Z and T-tests are important to calculate if the difference between a sample and population is substantial.



If the probability of Z is small, the difference between the two averages is more significant.

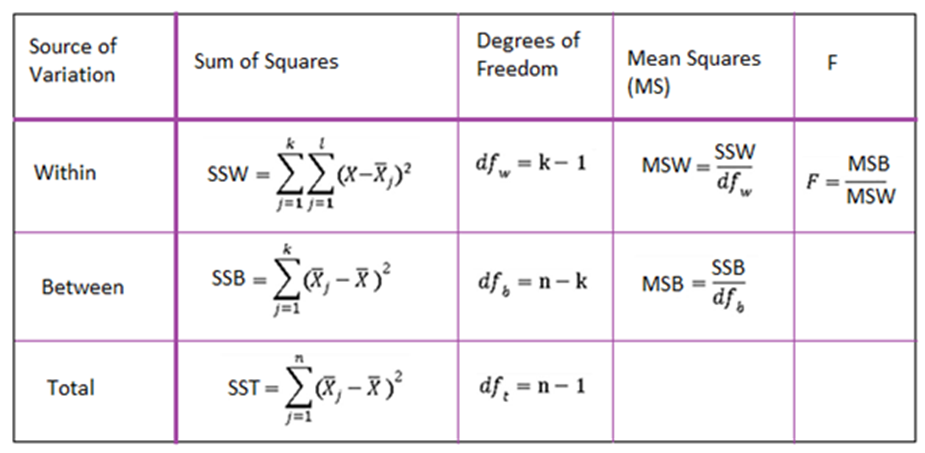
**T-Test**



If the sample size is large enough, then we use a Z-test, and for a small sample size, we use a T-test.

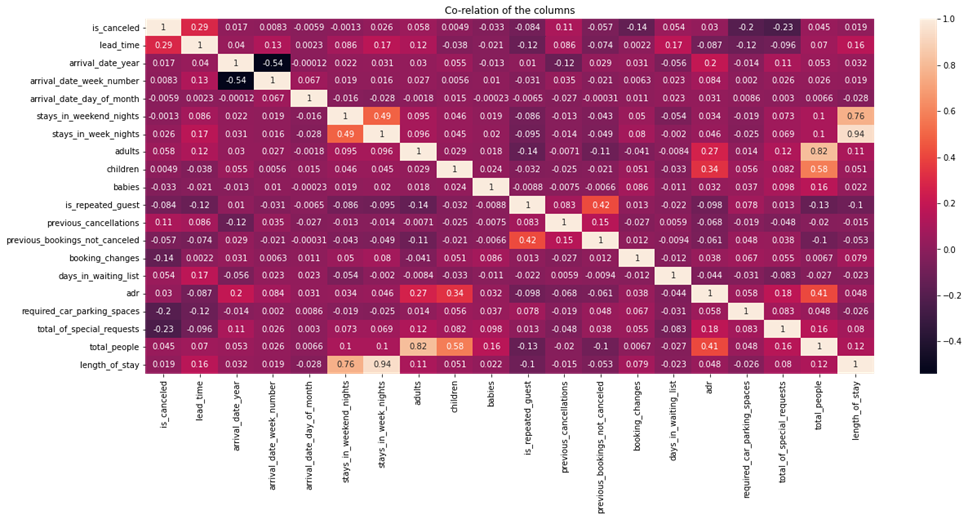
* **ANALYSIS OF VARIANCE (ANOVA)**

The ANOVA test is used to determine whether there is a significant difference among the averages of more than two groups that are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable for more than two categories of a categorical variable.



**Multivariate Analysis**

Multivariate analysis is required when more than two variables have to be analyzed simultaneously. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph and thus multivariate analysis is used to study more complex sets of data.

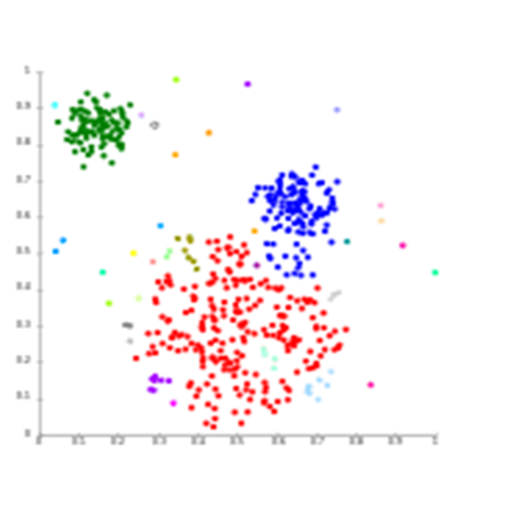
****

Types of Multivariate Analysis include Cluster Analysis, Factor Analysis, Multiple Regression Analysis, Principal Component Analysis, etc. More than 20 different ways to perform multivariate analysis exist and which one to choose depends upon the type of data and the end goal to achieve.

The most common ways are:

* ***Cluster Analysis***

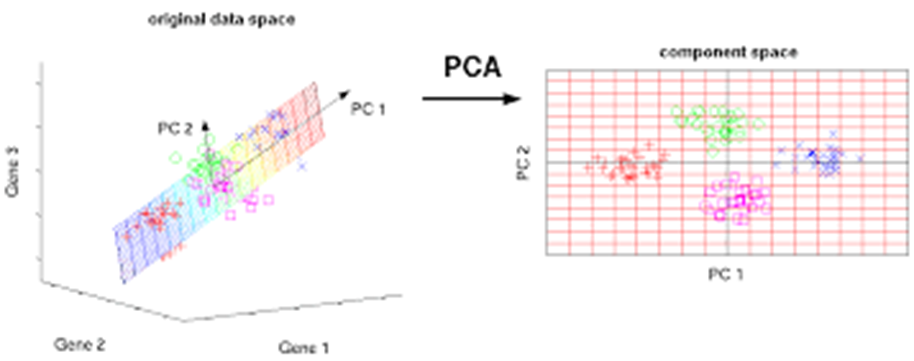
Cluster Analysis classifies different objects into clusters in a way that the similarity between two objects from the same group is maximum and minimal otherwise. It is used when rows and columns of the data table represent the same units and the measure represents distance or similarity.



* ***Principal Component Analysis (PCA)***

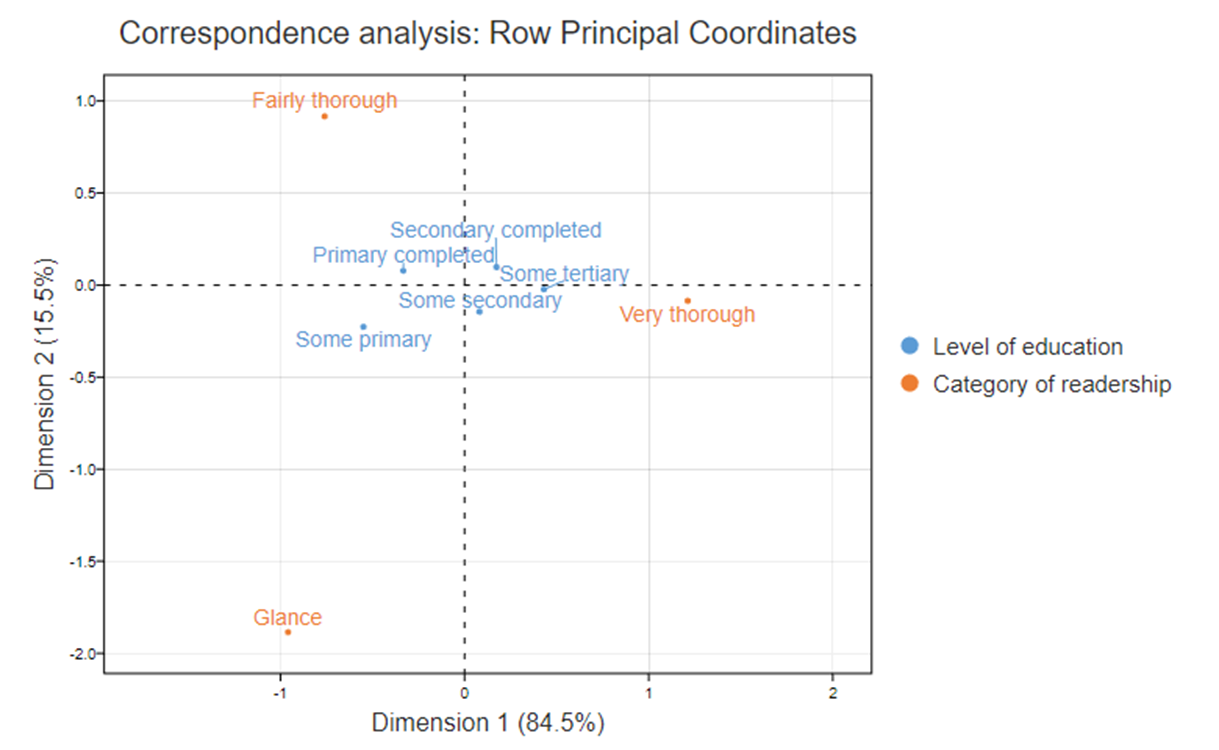
Principal Components Analysis (or PCA) is used for reducing the dimensionality of a data table with a large number of interrelated measures. Here, the original variables are converted into a new set of variables, which are known as the “Principal Components” of Principal Component Analysis.

PCA is used for the dataset that shows multicollinearity. Although least squares estimates are biased, the distance between variances and their actual value can be really large. So, PCA adds some bias and reduces standard error for the regression model.



* ***Correspondence Analysis***

Correspondence Analysis using the data from a contingency table shows relative relationships between and among two different groups of variables. A contingency table is a 2D table with rows and columns as groups of variables.



**6. The Conclusive Findings of the Study**

City hotels having maximum booking, which are the most preferred hotel type by the guests. We can say City hotel is the busiest hotel.

     Considering the years, Bookings for 2016 were highest with 48%. Bookings were increasing on a yearly basis. For 2016 increment was too high so there was a steep fall in bookings for 2017.

     The month of August saw the highest number of hotel bookings for both types of hotels. The lowest hotel bookings were in January. Daily rates are also high in the month of August Where low in January so January is the best time to book a hotel with cheaper rate & confirmed booking.

    Guests preferred no deposit type bookings because of the flexibility to   cancel or change bookings without losing any money. About 87.63% bookings are without any deposit. 12.24% bookings deposit are non refundable.

   Out of total booking 37% booking were canceled. With high no. of booking, cancellation is also higher in city hotels and out of that canceled booking, hotels with no deposit have higher cancellation rate.

     Repeated guests are not more likely to cancel their bookings.

Considering the rates City hotel has the highest ADR. That means city hotels are generating more revenues than the resort hotels.

    Most people prefer a 2 night stay. Optimal length of stay is 7   days. As an increase in length of stay ADR decreases. To get the best adr you need to stay more than a month.

     For long stay guests prefer resort hotels.

     More involvement in bookings with Online TA and Offline TA/TO that means maximum hotels are booked with agents only.

   Special requests are more for an adult. As an increase in special requests and increasing in people ADR will be more.

     Most of the guests are coming from Portugal(PRT),Great Britain(GRB), France(FRA),Spain(ESP) and Germany(DEU).

     Finally checkout the reservation status, 36% of bookings got canceled.1% of bookings never arrived at hotel without canceling bookings.

**7. References :**

* Google
* Alma better capstone project sample project section