**Airbnb Booking Analysis**

**Sujit Musale**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Airbnb is a based in San Francisco, California, Its online Marketplace focused on short-term homestay.

In this dataset we have given the 49000 observation with 16 number of categorical variable.

Our experiment is for to know the How the various factor are contributing the company to check the minimum and maximum values of variable.to plot the graph and make visual for better understanding of the scenario.

We are learning about different host and areas, prediction about location, price and reviews, busiest host and why they are busiest, will check is there any noticeable difference of traffic among different areas and what could be reason for it. .

**1. Problem Statement-**

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data that can be analyzed and used for security, business decisions, understanding of customers' and providers' (hosts) behavior and performance on the platform, guiding marketing initiatives, implementation of innovative additional services and much more.

This dataset has around 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values.

Explore and analyze the data to discover key understandings (not limited to these) such as

1.What can we learn about different hosts and areas?

2. What can we learn from predictions? (ex: locations, prices, reviews, etc)

3. Which hosts are the busiest and why?

4. Is there any noticeable difference of traffic among different areas and what could be the reason for it?

**2. Introduction**

As we all know Airbnb act as Broker and charges commission one each booking.

You can find more details of Airbnb by clicking on this link - <https://en.wikipedia.org/wiki/Airbnb>

this project which is simply EDA( Exploratory Data Analysis) of given (Airbnb Booking Dataset) where I try to explore various factor which are doing impact on business by using python coding. I also eliminate some of the data rows which are having null values. Plotted various pie charts and bar charts to visualized the contribution of given variable.

### In the end we try to data distribution of variable to confirm the normal distribution of same. also plotted heat map to check the correlation of variable between each other.

## **3. Column’s Details**

* **Id –** unique id of each customer
* **Name –** apartment, room or home name.
* **Host\_id –** owner or host unique id
* **Host\_name -** name of owner or host.
* **Neighbourhood\_group –** especially the name of zone like manhantum, brooklyn etc.
* **Neighbourhood –** are near from booking location.
* **Latitude and longitude –** geographical coordinates of booking location.
* **Room type –** specify the type of room.
* **Price –** room booking price.
* **Minimum\_neights –** customer stay count in terms of nights.
* **Number\_of\_reviews –** total number of reviews given by the customer for particular property.
* **Last\_revies –** date of last review.
* **Review\_per\_month –** average review per month.
* **Availability\_365** – property available (days) in the year for customers**.**

**4. Steps involved:**

* **Import required library-** “pandas” is the very important library in python and use full for data mugging, manipulation and visualization.

“matplotlib” and “seaborn” this library or module is used to plot various type graph.

“numpy” this library is having specialty in numerical calculation.

* **Data cleaning-**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.

There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

* **Import csv file and null value treatment-**

We imported csv file for EDA and checked for null value.

Missing Data can occur when no information is provided for one or more items or for a whole unit. Missing Data is a very big problem in a real-life scenarios. Missing Data can also refer to as NA (Not Available) values in pandas. In Data Frame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed. For Example, suppose different users being surveyed may choose not to share their income, some users may choose not to share the address in this way many datasets went missing.

**Data is represented by two value:**

**None**: None is a Python singleton object that is often used for missing data in Python code.

**NaN** : NaN (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation

* **EDA-**

Performed EDA on the cleaned data to check various aspect.

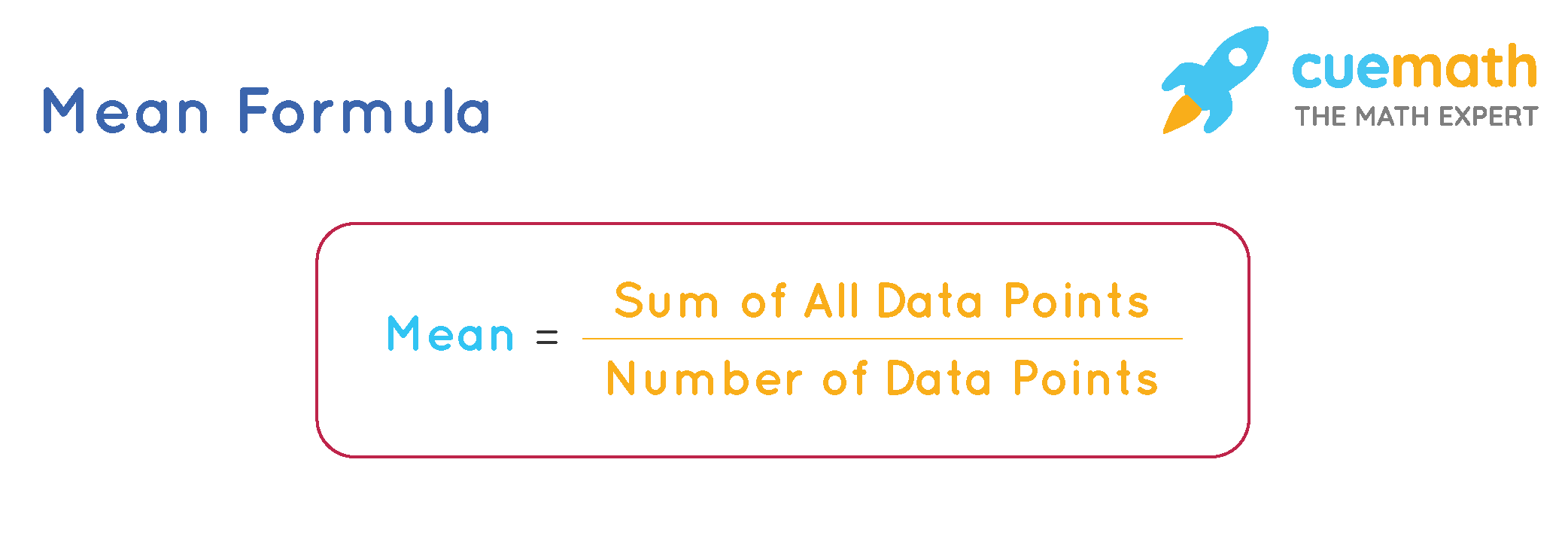
* **Conclusion-** Made conclusion from performed EDA.

**5. Mathematics concept’s:**

* **mean:** Mean is one of the important and most commonly used measures of central tendency. There are several types of means in mathematics. In statistics, the mean for a given set of observations is equal to the sum of all the values of a collection of data divided by the total number of values in the data.

**Example**- average room prices of various room type.

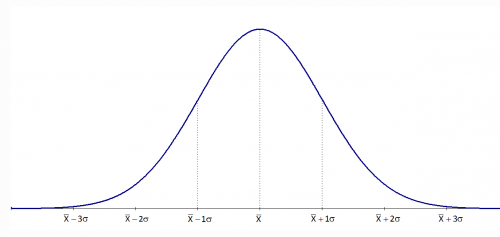
Calculated by below formula:

****

.

* **Normal Distribution:**

Normal distribution, also known as the Gaussian distribution,

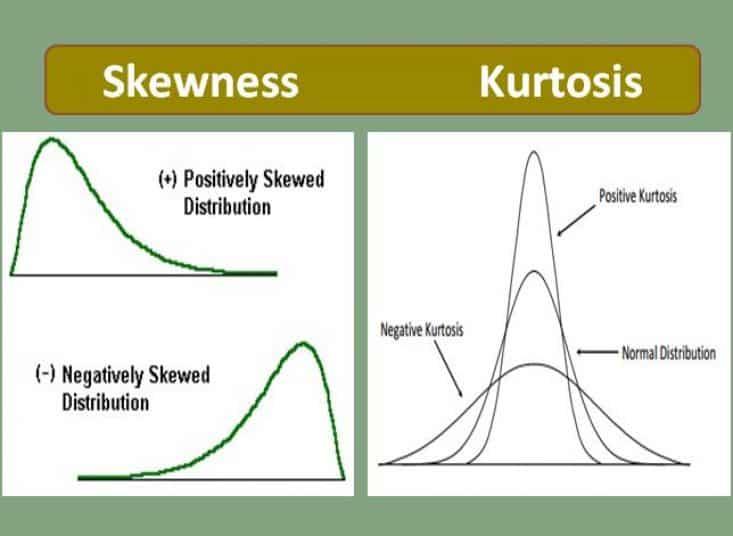


The standard normal distribution has two parameters: the mean and the standard deviation.

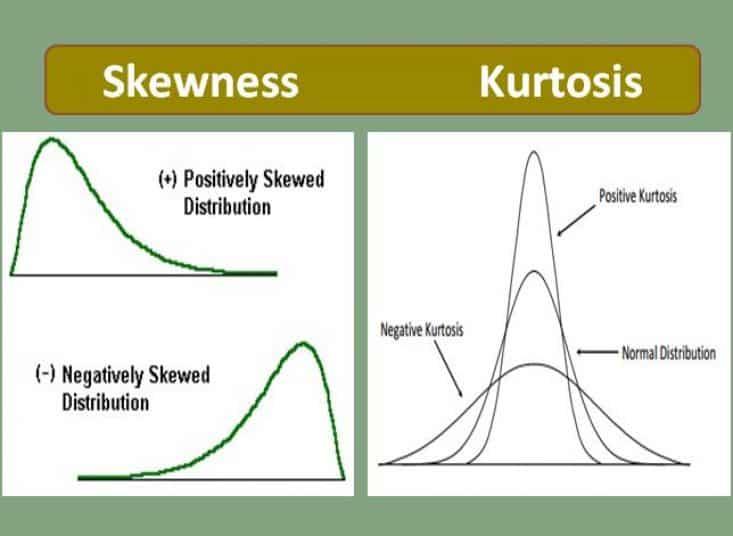
The normal distribution model is important in statistics and is key to the [Central Limit Theorem](https://www.investopedia.com/terms/c/central_limit_theorem.asp) (CLT). This theory states that averages calculated from independent, identically distributed random variables have approximately normal distributions, regardless of the type of distribution from which the variables are sample

is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.

“[Skewness](https://www.investopedia.com/terms/s/skewness.asp)” measures the degree of symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero.



“[Kurtosis](https://www.investopedia.com/terms/k/kurtosis.asp)” measures the thickness of the tail ends of a distribution in relation to the tails of a distribution. The normal distribution has a kurtosis equal to 3.0.



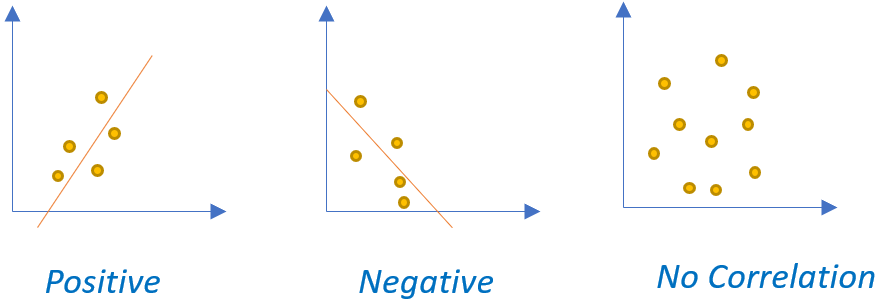
* **correlation:**

**Correlation** refers to a process for establishing the relationships between two variables. You learned a way to get a general idea about whether or not two variables are related, is to plot them on a “[scatter plot](https://byjus.com/maths/scatter-plot/)”. While there are many

Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. The correlation coefficient is usually represented using the symbol r, and it ranges from -1 to +1.

There are two type of correlation.

* Positive correlation
* Negative correlation



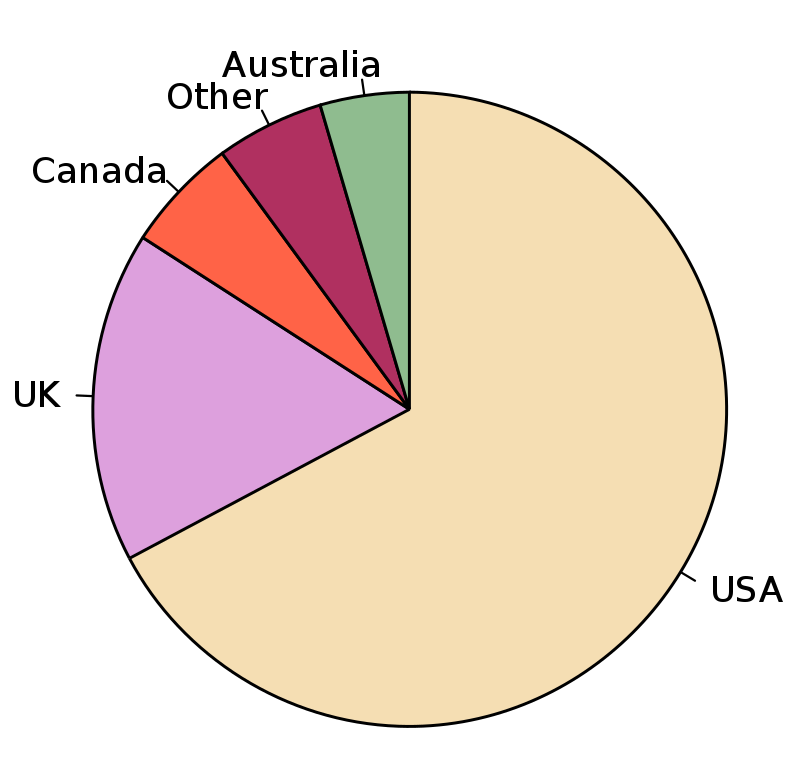
Above figure shows the positive, negative and no correlational data.

**6. plots:**

* **Pie chart-**

A pie chart (or a circle chart) is a circular [statistical graphic](https://en.wikipedia.org/wiki/Statistical_graphics), which is divided into slices to illustrate numerical proportion. In a pie chart, the [arc length](https://en.wikipedia.org/wiki/Arc_length) of each slice (and consequently its [central angle](https://en.wikipedia.org/wiki/Central_angle) and [area](https://en.wikipedia.org/wiki/Area)) is [proportional](https://en.wikipedia.org/wiki/Proportionality_(mathematics)) to the quantity it represents. While it is named for its resemblance to a [pie](https://en.wikipedia.org/wiki/Pie) which has been sliced, there are variations on the way it can be presented.

Let’s have a look on population of English language native speakers across the word.

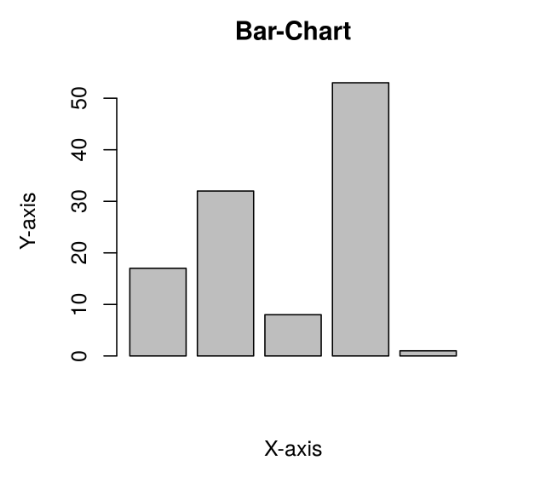


* **Bar chart**

A bar chart or bar graph is a chart or graph that presents [categorical data](https://en.wikipedia.org/wiki/Categorical_variable) with rectangular bars with [heights](https://en.wikipedia.org/wiki/Height) or [lengths](https://en.wikipedia.org/wiki/Length) proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

A bar graph shows comparisons among [discrete](https://en.wikipedia.org/wiki/Discrete_variable) [categories](https://en.wikipedia.org/wiki/Categorical_variable). One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Some bar graphs present bars clustered in groups of more than one, showing the values of more than one measured variable.

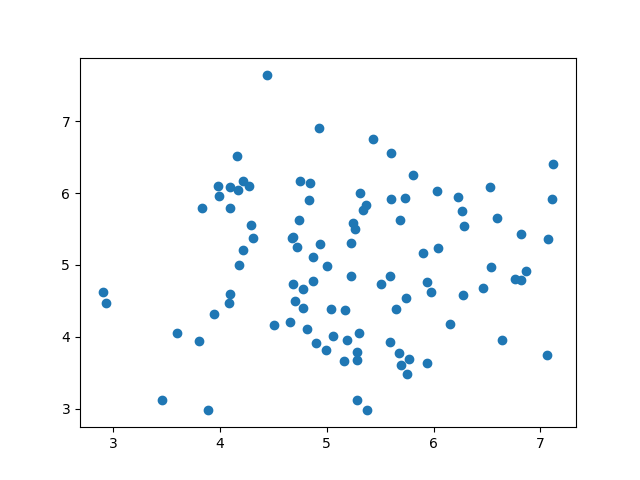
Let’s have look on below simple bar graph.



* **Scatter plot-**

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scatter diagram, or scatter diagram)  is a type of [plot](https://en.wikipedia.org/wiki/Plot_(graphics)) or [mathematical diagram](https://en.wikipedia.org/wiki/Mathematical_diagram) using [Cartesian coordinates](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) to display values for typically two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the [vertical axis](https://en.wikipedia.org/wiki/Vertical_axis).

Let’s have the look on below mention scatter plot.



**7. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya