**Airbnb NYC 2019 Analysis and Visualization**

**MEGHA DETHE**

**Data science trainee Alma Better, Bangalore**

**Abstract:**

In the Airbnb booking process that Enable to book or schedule date and Length of the stay. How many room select and payment option in one place. We team members followed step by step process for the EDA capstone project like data collection,

Data wrangling, EDA, data visualization.

In this project we have taken *Airbnb* booking data for New York City of the year 2019 (henceforth will be represented as AIRBNB\_NYC\_2019). As stated in the problem statement this dataset contains around 49000 observations, distributed among 16 columns. These columns represent various categorical and Numerical data namely ***id, name, host id, host name, neighborhood group, neighborhood, latitude, longitude, room type, price, minimum nights, number of reviews, last review, reviews per month, calculated host listings count, availability 365.***

**Keywords: pandas, matplotlib, seaborn,**

**1.Problem Statement**

We are supposed to explore and analyze the data, and to assess some analysis. While some of them are already defined, we have framed certain query by ourselves. All total 6 queries are generated. Below we have listed them:

1. What can we learn from predictions? (ex: locations, prices, reviews, etc)
2. Which hosts are the busiest and why?
3. What can we learn about different host and area?
4. Which rooms has the lowest review?
5. Relationship between number of review and prices?

**2. Introduction**

Data mining of large datasets can give us valuable insight of various parameters. It can also help us to make certain decisions based upon those parameters. A lot of tools are in use for deep diving of the datasets, like Python, Pearl, R, MATLAB… to name a few. Among them, in this project we have used Python because not only it is most suited for Machine Learning, but we can further expand it within web applications. Online travel forums have become an extremely popular platform for sharing travel information, with a large number of reviews being posted daily. Airbnb websites such as www.airbnb.com and Booking.com have turned into very important resources for operators and travelers alike. While hotel operators use the forums to promote their hotels as well as to solicit feedback in an effort to improve their services, travelers use them to decide on which hotels to stay at during their vacation or business travels, as well as to share their experience with other travelers on the services of the hotels. A travel forum site, such as Booking.com, enables travelers to con- tribute descriptions, pictures, reviews and travelogs about the hotels they have stayed at and the tourist attractions they have visited. Individuals who visit a TripAdvisor site can gather numerous bits of information about a destination connected with other travelers around the world, to provide and read feedback/reviews on the satisfaction level for the hotels they stayed at.

## **3. Types of Airbnb Booking**

* **Online**

You need Internet access. You need to be ready for an influx of new customers. Not all online booking systems are created equal.

## **4. EDA**

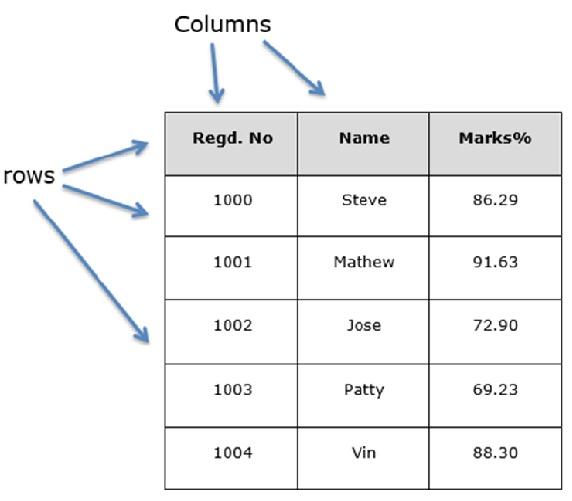
## EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Often, EDA involves data visualization, including creating graphs like histograms, scatter plots and box plots. Value: A data value is a piece of information, such as a number or a date. Variable: A data variable is a characteristic that you can measure, such as weight or income Import libraries and load dataset Check for missing values Visualizing the missing values replacing the missing values Asking Analytical Questions and Visualizations. Positive Correlation. Negative Correlation

# **5. Data Collection**

**Exploratory Data Analysis** is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. Exploratory data analysis has been promoted by John Tukey since 1970 to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed

1. **Pandas**

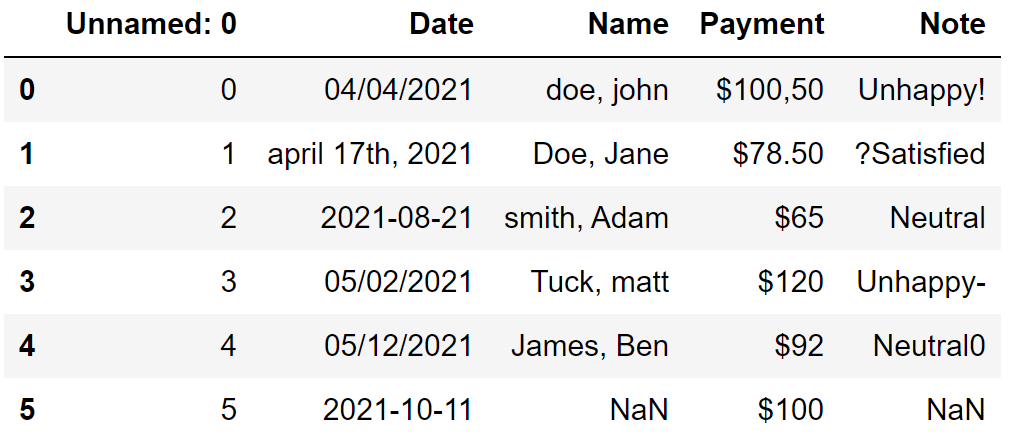
Pandas Data frame is two-dimensional size-mutable, potentially heterogeneous tabular data structure with 3abelled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and



Columns. Pandas Data Frame consists of three principal components, the data, rows, and columns. Dealing with Rows and Columns A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. We can perform basic operations on rows/columns like selecting, deleting, adding, and renaming. Column Selection: In Order to select a column in Pandas Data Frame, we can either access the columns by calling them by their columns name

1. **Data Cleaning**

It is an essential skill of Data Scientists to be able to work with messy data, missing values, inconsistent, noise, or nonsensical data. To work smoothly python provides a built-in module Pandas How do you remove NULL data in Python? Pandas Data Frame drop na() function is used to remove rows and columns with Null/Nan values. By default, this function returns a new Data Frame and the source Data Frame remains unchanged. We can create null values using none, pandas. Nat, and numpy. Pandas Data Frame drop na () function is used to remove rows and

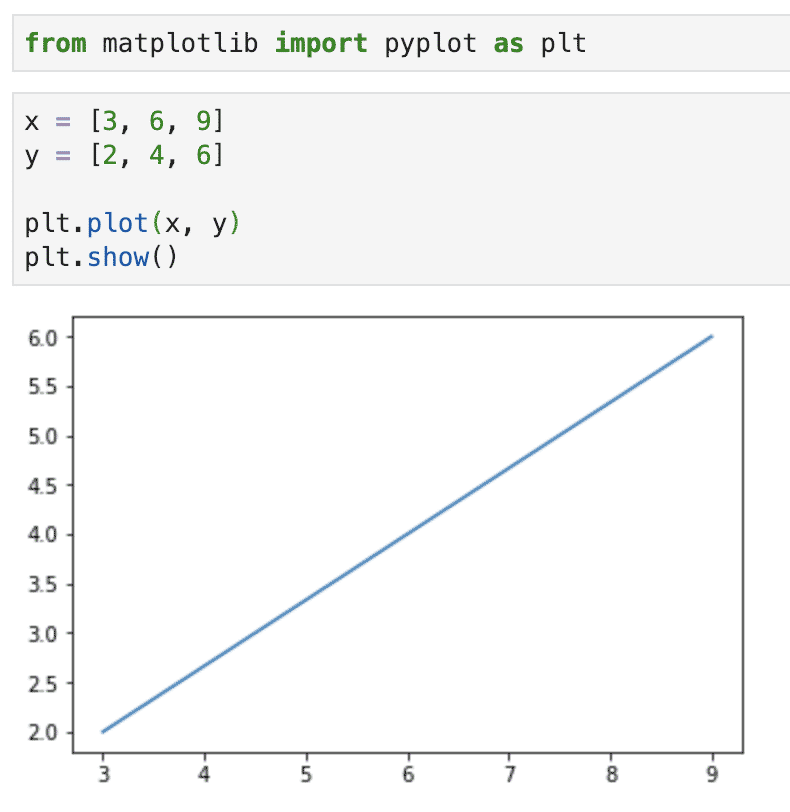


Columns with Null/NaN values. By default, this function returns a new Data Frame and the source Data Frame remains unchanged We can create null values using None, pandas NaT, and Numpy nan variables

**6.Analysis**

**• Matplotlib**

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open source alternative to MATLAB. Developers can also use matplotlib’s APIs (Application Programming Interfaces) to embed plots in GUI applications. A Python matplotlib script is structured so that a few lines of code are all that is required in most instances to generate a visual data plot. The matplotlib scripting layer overlays two APIs: The pyplot API is a hierarchy of Python code objects topped by matplotlib pyplot An OO (Object-Oriented) API collection of objects that can be assembled with greater flexibility than pyplot. This API provides direct access to Matplotlib’s backend layers. Numpy is a package for scientific computing. Numpy is a required dependency for matplotlib, which uses numpy functions for numerical data and multi-dimensional arrays as shown in the following code snippet:

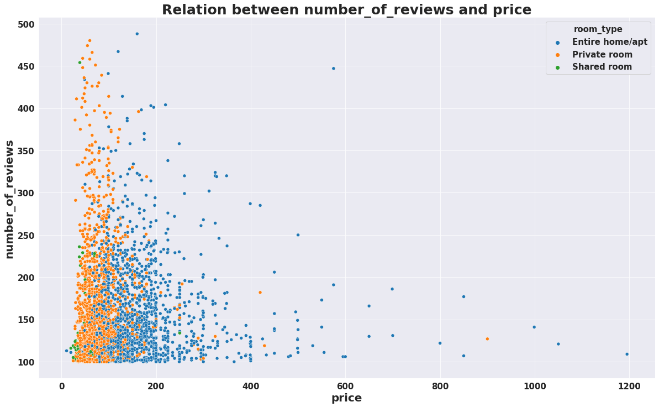


1. **Visualization**

Data Visualization represents the text or numerical data in a visual format, which makes it easy to grasp the information the data express. We, humans, remember the pictures more easily than readable text, so Python provides us various libraries for data visualization like matplotlib, seaborn, plotly, etc. In this tutorial, we will use Matplotlib and seaborn for performing various techniques to explore data using various plots. Analysis is the simplest form of analysis where we explore a single variable. Univariate analysis is performed to describe the data in a better way. We perform Univariate analysis of Numerical and categorical variables differently because plotting uses different plots.

**• Seaborn**

Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily Data Visualization is the art of representing data in the form of graphs. It is a useful tool for professionals who work with data, i.e., financial analysts, business analysts, data analysts, data scientists this tutorial can be divided into three main parts.

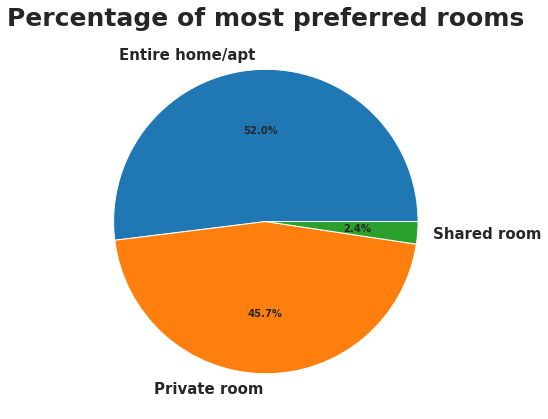


The first part will talk about installing seaborn and loading our dataset. In the second part, we will discuss some common graphs in Seaborn. Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions. Seaborn is a Python data visualization library based on the Matplotlib library. It provides a high-level interface for drawing attractive and informative statistical graphs

**• Charts**

1. **Pie Chart**

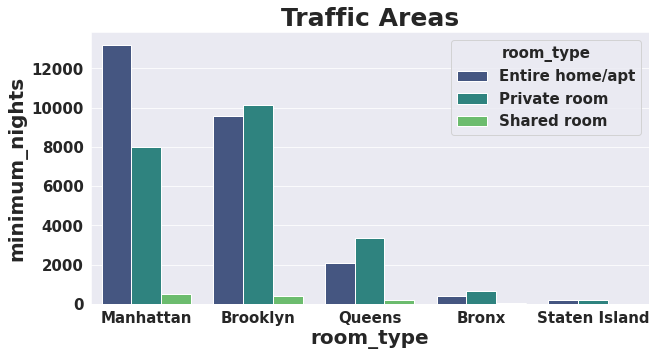
A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents. While it is named for its resemblance to a pie which has been sliced, there are variations on the way it can be presented. The earliest known pie chart is generally credited to William Play fair’s Statistical Breviary Pie charts are very widely used in the business world and the mass media. However, they have been criticized, and many experts recommend avoiding them, as research has shown it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts. Pie charts can be replaced in most cases by other plots such as the bar chart, box plot, dot plot, etc For ex.

****

**2. Bar Plot**

Bar graphs/charts provide a visual presentation of categorical data.[4] Categorical data is a grouping of data into discrete groups, such as months of the year, age group, shoe sizes, and animals. These categories are usually qualitative. In a column (vertical) bar chart, categories appear along the horizontal axis and the height of the bar corresponds to the value of each category. Bar charts have a discrete domain of categories, and are usually scaled so that all the data can fit on the chart. When there is no natural ordering of the categories being compared, bars on the chart may be arranged in any order. Bar charts arranged from highest to lowest incidence are called Pareto charts. A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths 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.

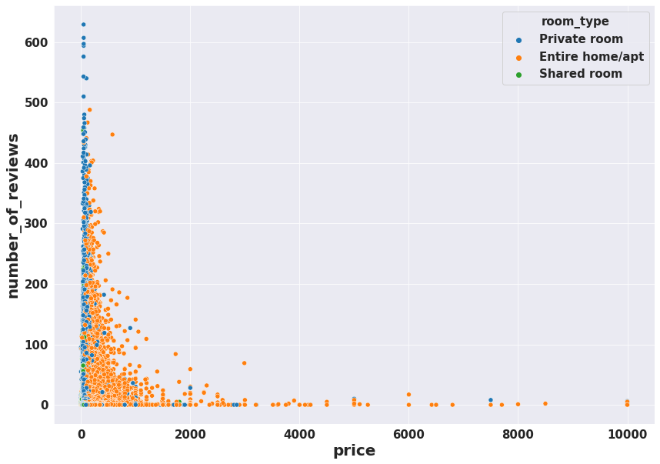
For ex.

****

**3. Scatter Plot**

A scatter plot can be used either when one continuous variable is under the control of the experimenter or the other depends on it or when both continuous variables are independent. If a parameter exists that is systematically incremented and/or decremented by the other, it is called the control parameter or independent variable and is customarily plotted along the horizontal axis. The measured or dependent variable is customarily plotted along the vertical axis. If no dependent variable exists, either type of variable can be plotted on either axis and a scatter plot will illustrate only the degree of correlation (not causation) between two variables. A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval. For example, weight and height would be on the y-axis, and height would be on the x-axis. Correlations may be positive (rising), negative (falling), or null (uncorrelated). If the dots' pattern slopes from lower left to upper right, it indicates a positive correlation among the 3 variables being studied. If the pattern of dots slopes from upper left to lower right, it indicates a negative correlation.

For an example

****

1. **Conclusion**

We conclude: Manhattan is most preferred by guests, followed by Queens, Staten Island Brooklyn and Bronx.Average review however shows that guests are satisfied with other neighbourhood too. Maybe this is due to the fact that Manhattan has most hosts, thus fetching a high value of total reviewes.the guests generally choose cheaper hosts, which is quite obvious.People are preferring entire home/apt or private room which are present in Manhattan , Brooklyn, and Queens.As the price is increasing then number of reviews decreasing so we can say that more price is better dealing with customers.

**• Reference**

1. Wikipedia

2. https://www.activestate.com/resou rces/quick-reads/what-is matplotlib-in-python-how-to-use it-for plotting/#:~:text=Matplotlib%20is %20a%20cross%2Dplatform,emb

3. “Think Python: How to think like a computer scientist”, Allen B Douney, *O’REILLY*, 2002