**Airbnb Bookings Analysis**

**Lakdawala Ali Asgar, Tembhare Kunal**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

The project “Airbnb Bookings Analysis” is based on discovering key understandings about hosts, areas and their traffic.

In this analysis, we’re provided with a dataset with some records. We did basic inspection, data cleaning to remove suspicious data and to avoid errors.  We’ve used different plots to visualize our analysis in the easiest way. This can be used to improve no of booking

***Keywords: EDA, Airbnb, Data visualization, and exploration.***

**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 entire 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 of categorical and numeric values. Explore and analyze the data to discover key understandings.

**2. Introduction**

We live in an era where data is produced and circulated in an enormous amount. Those data can be collected and allow us to infer meaningful results and make well-informed decisions. However, as the number of data increases, we need to visualize the data to help us in conducting data analysis. By using visualization tools, we can deliver a message to our audience and inform them about our findings.

This project explores a dataset from a technology company, maps the result clearly through visualization tools, and gives new insight to the public and other relevant parties.

We will explore and visualize the dataset from Airbnb in New York using basic exploratory data analysis techniques. We will find out the distribution of every Airbnb listing based on their location, including their price range, room type, listing name, and other related factors.

This San-Francisco-based startup offers you someone’s home as a place to stay instead of a hotel. Airbnb allows you to be a host for anyone anywhere with rooms/beds available in your personal space.

Our goal here is to explore the data and find useful insights from the data and find out different relations between the columns.

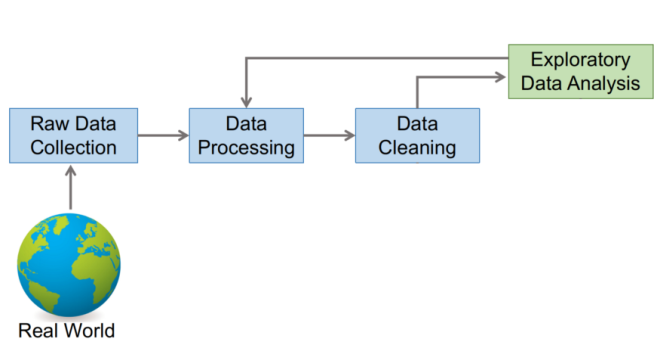
Basic descriptions of columns after univariate analysis are

* name = Description about the listings.
* host\_id = unique id for each listed host.
* host\_name = Hostname for the listings.
* neighbourhood\_group = Location
* neighborhood = Area
* latitude = Latitude coordinates
* longitude = Longitude coordinates
* room\_type = Listing space types
* price = Price in dollars
* minimum\_nights = minimum nights required to stay
* number\_of\_reviews = No. of reviews written for the listing
* last\_review = Last reviewed date for the listing
* reviews\_per\_month = Total review per month for the listings
* calculated\_host\_listings\_count = Total no of listing against the host id
* availability\_365 = Available days of a listing in a year.

# 3. **Steps involved:**

1. Acquire and loading data

For this project, we are using jupyter notebook IDE with a python programming language to write our script. Airbnb data is available at<http://insideairbnb.com/.> This data file includes all needed information to find out more about hosts, geographical availability, and necessary metrics to make predictions and draw conclusions.



1. Exploratory Data Analysis

After loading the dataset, we performed this method by comparing all the columns. This process helped us to figure out patterns, spot anomalies, detect outliers, and find fascinating relations among the variables. By conducting EDA, we can turn an almost usable or unusable dataset into a usable dataset. The major components of exploratory data analysis are Cleaning datasets, exploring and visualizing data.

* Cleaning dataset

The next step is cleaning up the data. Often the data we load has various faults, such as typos, missing values, incomplete data. By cleaning up, the data quality will be better to use for further analysis.

* Null & unfitting values Treatment

Our dataset contains many null values and a tiny number of rows, so we tried to preserve as many rows as possible by replacing null values with suitable values.

* Non-Graphical and Graphical Univariate Analysis

We analyze just one variable to describe the data and find patterns that exist within it. For which we used different graphs like a bar, histogram, box, etc. By which we came to know about some outliers.

* Outlier treatment:

Different rows were removed from the dataset based on the conclusion of the univariate analysis, i.e. data that listed price and availability\_365 is 0 and a value that had a z value > 3 (used for some graphs).

| https://lh5.googleusercontent.com/Z3QYczkFvUK2XfvfJrtcrMmmU1v6P55rtly77Yl6RqnvW9BIPXOGgIhjfiCMge4yvJjwNdRP9vM3Qh3Q9j5diiOiHTtgaAEOFjW5V451P0hUqOMHKzLZpMyLsUsULzQLRpKznslYTO8  **Fig 1 Distplot of availibility\_365 before** | https://lh4.googleusercontent.com/7o6-OnmfwMf1V3q4Itsbmu6iSVeZOD-YMO_pDb-27du_1GwM6Bgb-w8bS2msGNmu4G9hTOCSiq49Ybsj6oql0iUytqEwAttGxlyG20EETiQ4cN6grrgFAsRgfxEzoGKubSEJ2THWFkg  **Fig 2 Distplot of availibility\_365 after** |
| --- | --- |

* Correlation Analysis:

The relationship between two or more different variables of data is shown employing different graphs, created using different python libraries such as matplotlib and snsborn. The most commonly used graph is a grouped bar plot or bar chart, with each group representing the levels of one variable and each bar representing the levels of another variable within the group.

* Difference of traffic among different areas

| https://lh4.googleusercontent.com/GCK98f5OUvNFV27ytbWV8ejY5p2U_6zjWG5SC9qeE6Ztx7ZK0BJ4ULsiy5is7S9YhhOcypWdRN4OQR6h8sngY2uk-kX2hJXR2s_bmrBVRlVVQ9n677cER1bh8dmS0msh7bMvHKLMpJ4  **Fig 3 map of New York** | **Fig 4 map of New York with density of apartments** |
| --- | --- |
| Density of hotels is large in manhattan and places in close contact of manhattan | |

| * Words used to attract people | * Does reviews affect the price |
| --- | --- |
| * https://lh3.googleusercontent.com/64gahGE44DBHwysw6CubxTwrq0TPXBWHTLqLK4ue5ASbfUbm7X2YCnfEQvVZJGu3BY-tGjOzRUsRAfPYY3uadypNUPIBaqOZFHW6Q_ugA-alcMTF5IWURuUe1Mem1g   **Fig 5 most common words used for apartment’s name** | https://lh5.googleusercontent.com/qOOe-_2jSB-0BUJ1c1Ych7GpyYUix83idU1wAWo0uas8iHr5mowJKgsGoMsuitH-Y01KyBPG9PZNQ1o8D3UUQJEUBiLcmWtS8JqNYkPH4Dwo22FHqFzjCafGOF4rVw  **Fig 6 Distplot of no of reviews vs. price** |
| Most of words are the location, view, and type of rooms. | lower the price higher is the no of reviews |

* Which hosts are the busiest and why?

| C:\Users\asus\Pictures\Screenshots\Screenshot (73).png  **Fig 7 bar plot of top 10 busy hosts by total review per month** | C:\Users\asus\Pictures\Screenshots\Screenshot (73).png  **Fig 8 bar plot of top 10 busy hosts by total review per month** |
| --- | --- |
| Most of the hosts have their price below average(median) i.e 112 | |

| https://lh5.googleusercontent.com/yKPZEb9mmbg60NsfU1ItYD-xUmuLQgWVU8fwVOfs2C4jDWsESz818IAsE_ydne-MHz1Xp9BBsqzixsD5FG9icCWGqgNlECqMAjPIdEI-x4AwxTNGh6Obm1_mq1Bo3_3ZQkdyvyWRmBI  **Fig 9 Pie chart of busy hosts neighborhood group** | https://lh3.googleusercontent.com/XspNv9-mZiwyRBpy_zw9_Yc1fBwAPmPRRn7-clj--MNAW6L8ISqHlfIK14BiEDL92w5Pai6tkXa8vUjMda2G9Okz4IsulGAWUUHet2RgtC-sek8t-mALqrd7xXyDFv1YUy0L39s_BmI  **Fig 10 Pie chart of busy hosts neighborhood** | |
| --- | --- | --- |
| Most of the listing are in financial district of Manhattan | |

| https://lh4.googleusercontent.com/MC1APF6yVRFD1ThehQTE1CuyzO3GbxvUQrfVuvAyVDt8iybO81wzOcqPe_1tNWs6aKgb_chFVvH1FU03EDEovvDotaLihuPNmmM62_G8pJV6CZlaqm0UnX9XCgSIQBWxV9lIBZYEOVE  **Fig 11 Map representing apartments location** | https://lh5.googleusercontent.com/KrxCn-Yi6TO8vVrRLjDb4zTKqPO20oX5ANRHKefedOY8CnkVYaBP_y6WyfGb9k3EVMgl42HpXduMGKI6Sw4AXfEoi0K7htDW7XY6w0TF57PMH01Gcu07zS5FBJr2pTGA5NKb-KkLP9s  **Fig 12 bar plot showing their minimum required nights** |
| --- | --- |
| Most of their listing in lower manhattan or are near airport and probably focus on tourist rentals | |

* Which hosts are most popular and why?

| C:\Users\asus\Pictures\Screenshots\Screenshot (181).png  **Fig 13 bar plot of top 10 popular hosts by total review per month** | C:\Users\asus\Pictures\Screenshots\Screenshot (182).png  **Fig 14 bar plot of most popular hosts median price** |
| --- | --- |
| Most of the hosts have their price above average(median) i.e 112 | |

| **https://lh6.googleusercontent.com/NXlfdVmWxM0O-XtPCXDo6jiwsu86Zm0COaPwyO51S2i4eTdB-Z5r3uEnTV2KGqrj05BG1MNhj6XXRn3YBt4PF3TNqmL3UUIdHr3bOUbSQvaklIHtAkOm4UipSO0w1S_CTcjRGkQj8AM**  **Fig 15 Pie chart of popular hosts neighborhood group** | https://lh6.googleusercontent.com/X1wt9U7ILju6_OUTmvt972rtEKCrw1rdFeRXvhP35hsqcfyDs6eQDlRAz1NA4qzhi-9W9Ocp4IOD_7NJdgBoRVr_zJteLN3S6xAJRY-ROC6PYMDWpOwXuwV1uwE71g  **Fig 16 Pie chart of most popular hosts neighborhood** |
| --- | --- |
| Most of the listing of Manhattan | |

| https://lh3.googleusercontent.com/rzqFeaIjB6U7oGKGZbKp8FWKHWDHqRKEEJ4JLbroKO1s5z97g91LOP230MTicN2iLAuGhvc42_qY_NQRn8dX9tIwp1yFDy1qYSA6rQqFr4X5P8-GWBhgY-8i-KsnUmlnNs5d5ez9i-A  **Fig 17 Map representing apartments location** | https://lh4.googleusercontent.com/QGJizi3Q9LnJPGKlglaMOxyspLpMgeNiLFm2gZ469u9OSrYj0e1eSVzF0NGN_OrcLb7l3C381vThmEGBXuBZmng1siQrCQlau_bAj6A_q9g7ZBakUyw3fv02wAx2VOaca5jIY19q1Dg  **Fig 18 bar plot showing their minimum required nights** |
| --- | --- |
| Most of their listing in lower Manhattan or are near Manhattan and probably focus on business rentals | |

* Ideal room conditions to gain max out of it

| https://lh3.googleusercontent.com/NTO4X5WETdEezHQdLGC4iyOYFwwH4twwF6rF6rsnNkUcF_fQbRymUHqYp3qNeLvakmh2yBZL9nqvP0C0r4oZvAbqwW_neR3g3Vqfc_WrGHI4TRTiEpIViTQGT495NlloLHnwFxeamWs | https://lh6.googleusercontent.com/79Kxs_0Qem4-5GmGfM0WdLurxxwJRmoP_Q7jGqps06-L-OnNK96d07yB4RUYRyKjQ1L7owahlFmXMdEI9SY0YGGRh_VGc53HB7b4OZT5kGJvbViesZthqUi5JYvScw-lBM7VifAJ5A4 |
| --- | --- |
| https://lh3.googleusercontent.com/_R-0_hFUmzmZBlmVpkbyFRiqkgabqojo1DBcZmKiUeMoBUl9ci2jRRRA7uI16zL5dt6yVSvWHLJC2KJTD2jPkSPhsWW4KyHxyPW_FPCLzwG3uEvDv7rt_bB-hG4MwcAy6apeFH7wTl8 | https://lh6.googleusercontent.com/jJk0pprE1VlViMAYTLjQBOIt-VYCSSORE9PJSKLfjCJc7lW18KsNuDKmcp5AxiA-NAs-zDAIVSPt0QByHVNN6B9ugUoje6iirOwUk3ygnd-pMj3-ZYqlBVKVb5jwFW0cT057n3HyIVg |
| People prefer rooms in the range of 50-200 in or near Manhattan with  min nights as (1,2,3 or 30) | |

* Price distribution neighborhood wise

| **https://lh6.googleusercontent.com/oW-6orpNrShIWwtQV51lk4Td8tO8KuZsDL6yOknBOd8kodJwpMJfutiFKd7oizty8YJYRdKiHUHP1O6S769nzVYQ3aGmT0wFL4u2TNIK11ezc-6RYHUjAB9-72l4PMPblDvfcnGM=s0** | **https://lh3.googleusercontent.com/egkwZojrx_qyAHrp6sGDT-JisMFRiAXD5GUIPs0bCZoocS2Ta_oqs-HKdTsSknaWdJTy_onZ7RmWzTPTiEYmPH0pyShVvn4z1Iyn2MYUUV_RDBNmDFlIIHgoBOX9DWVHwezgLRyy=s0** |
| --- | --- |
| **https://lh5.googleusercontent.com/bnZhqlKeHViBydWX9kIBp_kz94WPx6fDUFcz3_5EOK_eX9cyKF4_yopyCxJ7_dD8GQaxZLP_Tq_Q_cJHCrxP3ulAO18M_KkpGt6ACERleyuKAJ7BKR7FTRez6sxQMTFiL8rkU_4n=s0** | **https://lh6.googleusercontent.com/AUWxy2kNT7RyxzQYmwcpXl9SzskGSRsLYpba6NlSRRDU5TrmZQTxJkogRj7RNKbJVuqCq9Trb2Cl8PTcJAmXFN2LTjWlQWcy_TxOwywKfQPfsGx9E-aYh4mmgQYs7g1EZHgrlLBX=s0** |
| **https://lh4.googleusercontent.com/jCLvir5q3rk8wdaaKm81cTQKsS9ZKSBdVmEDz1A_u3CsxN2Daj8LwPNWt0GTb7gKF-b9Amj-vXNRJx6mhS2gEe7P1NJeFbm21sVa9skFwXYq9mlzgoFQ0Do6cBMlp7AGCbtMtZUD=s0** | Flatiron District and Tribeca are the most expensive neighborhoods in Manhattan. Dumbo is the most expensive neighborhood in Brooklyn. Riverdale in the Bronx is the most expensive. Fort Wadsworth is the most expensive neighborhood on Staten Island. |

**4. Conclusion:**

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

Starting with loading the data, So far we have done identification of variables and data types, Missing value treatment, Non-Graphical and Graphical Univariate Analysis, outlier treatment, and Correlation Analysis and learned that

* Availability of the room in a year does not affect the reviews of the data set.
* There are 36 percent of inactive hosts. Which are directly proportional to the density of listings in the neighborhood.
* Density of hotels is large in Manhattan and places in close contact of Manhattan
* Typical words that are used by the host to make their listing easily searchable are the location, view, and type of rooms.
* Lower the price higher is the no of reviews
* Most of the hosts have their price below average (median) i.e. 112. Most of their listings are in lower Manhattan or are near airports and probably focus on tourist rentals.
* Most of the hosts have their price above average (median) i.e. 112, Most of the listings are in lower Manhattan or near Manhattan and probably focus on business rentals.
* People prefer rooms in the range of 50-200 in or near Manhattan with min nights as (1,2,3,30)

**References-**

1. GeeksforGeeks
2. Stack Overflow
3. medium