Abhilash Kumar Dikshit   
R Practice: ALY 6010

Module 1 Final Milestone

Week 2

# Exploratory Data Analysis (EDA) allows us to:

* **Better understand the data:** Getting domain knowledge by reading some articles about the topic
* **Build intuition about the data:** Check if the data agree with our domain knowledge.
* **Generate hypotheses:** Understand how the data was generated, find insights, and try to predict the output.
* **Exploring anonymized data:** Explore individual features, check if the values match with our domain knowledge, explore features relations.

Initially, I began by downloading the raw dataset *“AB\_NYC\_2019.csv”* from [Kaggle](https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?resource=download) and using the libraries below for further Exploratory Data Analysis (EDA) report.

|  |  |
| --- | --- |
| Fig 1: In order to clean up the data, it was sorted by descending *price*; column 1 to 4 was dropped as it was insignificant; column headers were converted to capital letters and removed the top 40 % and bottom 25% of the uncleaned data. |  |

*Fig1: Cleaned Dataset Info and Top and Bottom 3 Of Cleaned Dataset*

|  |  |
| --- | --- |
| *Fig2: Data set’s variable and respective data types for Uncleaned and Cleaned Data frame, Summary of data frame* | Fig 2: As per the analysis, we have 7336 observations and 12 attributes available for our data analysis report and the class types of the following attributes have 4 characters, 5 integers and 8 numeric data types. We mainly focussed on the Airbnb rental prices, reviews per month, availability based on room type for the year 2019 for specific neighbourhood groups. |

Abhilash Kumar Dikshit

R Practice: ALY 6010

|  |  |
| --- | --- |
|  |  |
| *Fig3: Average Price for Room Types (Uncleaned & Cleaned Data frame)*  Fig 3: Entire home/apartment average price was 211.80 USD in uncleaned data frame and 127.30 USD in cleaned data frame whereas for Private rooms average price was 89.80 USD in uncleaned data frame and 122.60 USD in cleaned data frame. Moreover, shared rooms average price was 70.10 USD in uncleaned data frame and 120.30 in cleaned data frame. | |

|  |  |
| --- | --- |
| Fig 4: For price comparison across different neighborhood and room types, Bronx, Brooklyn, Manhattan, and Queens had the highest median price out of all of them i.e., 125. It also has the least number of outliers. Staten island has the lowest median price i.e., 120. | *Fig4: Price Across Different Neighbourhood Groups* |

|  |  |
| --- | --- |
| *Fig5:* *Prices for Different Room Types* | Fig 5: For price comparison across different neighborhood and room types, entire home or apartment has the highest median price out of all of them i.e., 125. It also has the least number of outliers whereas Private room has a median value of 120. Shared room has the lowest median value i.e., 115. |

Abhilash Kumar Dikshit

R Practice: ALY 6010

|  |  |
| --- | --- |
| *Fig6: Distribution of Room Types in New York City* | Fig 6: For distribution of room type in New York city, customers prefer Entire home or apartment type for stay and it was the most popular as compared to private and shared room types in New York city, USA as depicted in the bar plot for the year 2019. |

|  |  |
| --- | --- |
| Fig 7: Below we have differentiated the neighbourhood in 5 groups as shows in the given table.  Out of Bronx, Brooklyn, Manhattan, Queens, and Staten Island, the percentage of Airbnb’s located in Brooklyn is 45% which is the highest among all whereas Manhattan is at 39% followed by Queens which is at 13%, Bronx at 2% and Staten Island is below 1% | *Fig7:* *Airbnb’s Count for Neighbourhood in NY City* |

|  |  |
| --- | --- |
| *Fig8:* *No. of Reviews Density Plot Using Latitude and Longitude for Specific Room Types* | Fig 8: In Uncleaned plot, density of number of reviews in New York city is much higher and the value goes up to 600 for Entire home/apt room type. When compared the same to Shared room type, we see the density is much less.  For Cleaned Dataset, we can clearly analyse the density of number of reviews across all 3 room types and Entire home/apt room type has the highest denser plot based on the reviews received for year 2019 for NY city. |

Abhilash Kumar Dikshit

R Practice: ALY 6010

|  |  |
| --- | --- |
| Fig 9: We can clearly depict there is an exponential growth of the reviews received from year 2011-2019 from the bar plot. Moreover, we see a huge jump of reviews received in 2019 as compared to the previous years. For a decade, entire home/apt room types were always in demand as compared to other room types. Also, there was a positive trend of reviews received from the customers after 2015 due to the advanced of mobile application and introduction of 4G bandwidth. | Chart, bar chart  Description automatically generated  *Fig9: Reviews of Room Types From 2011-2019* |

|  |  |
| --- | --- |
| A picture containing text  Description automatically generated  *Fig10: Frequency tables* | Fig 10: Looking at the frequency table data for the neighbourhood group, Manhattan has the highest frequency of 3500 as compared to the other ones.  Then we converted the frequency table to a proportion table. The output shows the relative proportions of each value in our example vector.  Afterwards, we multiplied the proportion table that we have initialized with 100. The output shows the percentages of each character element in our example vector.  To find the Cumulative frequency, we applied the “cumsum” function. The output shows a cumulative frequency table of our input data. |

**References:**

1. *GitHub: Datasets Used in Plotly Examples and Documentation* ([November 5, 2022](http://127.0.0.1:64200/rmd_output/4/#ref-R-Kaggle));
2. *Descriptive Statistics and Graphics- Easy Guides* ([October 28, 2022](http://127.0.0.1:64200/rmd_output/4/#ref-R-Material)); Kervizic ([Feb 9, 2020](http://127.0.0.1:64200/rmd_output/4/#ref-R-Career));
3. *Data Visualization Tips for More Effective and Engaging Design*. n.d. Tableu. <https://www.tableau.com/learn/articles/data-visualization-tips>.
4. *Descriptive Statistics and Graphics- Easy Guides*. October 28, 2022. <http://www.sthda.com/english/wiki/descriptive-statistics-and-graphics>.
5. Kabacoff, Robert I. May 2015. *R in Action, Second Edition*. <https://www.manning.com/books/r-in-action-second-edition>.
6. Kervizic, Julien. Feb 9, 2020. *Data Science Career Path & Progression*. <https://medium.com/analytics-and-data/data-science-career-path-progression-ab5140cfbc84>.
7. *Mapping in r: Latitude and Longitude Coordinates*. October 29, 2019. <https://map-rfun.library.duke.edu/01_georeference.html>.