Technical document for EDA project

Airbnb

Prepared By

Member 1: **Kashif Kamran**
Email: kashifkamran639@gmail.com

Member 2: **Raja Chowdhury**
Email: rajachowdhury2468@gmail.com

Member 3: **Aman Jain**
Email: amanjn932@gmail.com

Member 4: **Sandipan Das**
Email: sandipan.das202@gmail.com

Notations:

- All the Codes are written in "`Courier New`" Fonts.
- Variable/Column Name are in *"Italic"*
- Visualization Tools used are Highlighted in **Bold**

1. Purpose of Document

- The purpose of this document is to understand the technical details of the analysis and explain each step followed to get the conclusion

2. Our Approach

We cleaned and preprocessed the data, analyze each column and then we performed the exploratory data analysis.

Steps performed in this EDA Projects

A. Data Exploration
B. Data Preparation
C. Exploratory Analysis
D. Key findings & Conclusions.

3. Data Exploration

In Exploration we dive into the data, we see top 5 rows, bottom 5 rows, shape, size and Geometry of our data using `bnb.head(), bnb.tail(), bnb.info() & bnb.describe()` Functions respectively.

Understanding different rows/variable in data set.

1. *id*- refers to the id of the hotel.
2. *name*- refers to the hotel name.
3. *host_id*- refers to the id of the hotel in charge.
4. *host_name*- refers to name of the hotel in charge.
5. *neighbourhood_group*- refers to the nearest place where the hotel has its other branches.
6. *neighbourhood*- refers to the nearest place of the hotel area location.
7. *latitude*-Here longitude refers to the geographic coordinates that specifies north and south position of the hotel.
8. *longitude*- Here longitude refers to the geographic coordinates that specifies east and west position of the hotel.
9. *room_type*- The type of the room is divided into three categories i.e private room, shared room and entire home (which means the entire hotel is booked).
10. *price*- refers to the cost of the room per night in dollars.
11. *minimum_nights*- refers to the minimum number of nights spent at that hotel.
12. *number_of_reviews*- refers to the reviews provided by the customer regarding the hotel.
13. *last_review*-refers to the date when the last review of a customer was recorded.
14. *reviews_per month*- refers to the reviews of the total number of customers in a month.
15. *calculated_host_listings_count*- It is total no listing done by particular host.
16. *availability_365*- refers to the availability of rooms out of the total 365 days of a year.

4. Data Preparation

In Data Preparation we prepare our data for Exploratory Analysis based on data exploration outcome, we came to know that we need to clean our data, handle outliner and stablish correlation in data set align with problem set if possible.

- Data Cleaning:

Using `df_nyc.isnull().sum()` We find some Null Values in host_name, name , last_review and reviews_per_month

host_name & name : Missing Values are less than 0.5% so will simply drop the rows using `df_nyc.drop()` corresponding to the missing values in those column.

last_review & reviews_per_month : for these column we found Null Values are for those properties which doesn't have any reviews. so replacing all the Null values of review_per_month with zero ('0') and drop the last_review as column has very high Null Values and negligible relevance with our problem set for EDA.

Also we found total of eleven entries need to be drop whose price is 0 so going forward excluding those entries having price = 0

Filtering the data

```
df = df_nyc[df_nyc['price'] !=0]
```

- Handling Outliers:

Two column *price & minimum_nights* are having Outliners confirmed using Boxplot chart.

Z-score method to handle out liner

We use z-score method to handle our outliners, in this method we add two more column for z-score "z_price, z_min_nights" then we filter our data set by putting conditions that only entries having z-score less than 3 will be considered.

- Correlations:

Using Heatmap we check the correlation between the numerical data however we did not find any potential correlation between the numerical variables except number of reviews and reviews per month.

This Bring us to the end of Data Preparation and our final data has 48149 entries, 0 to 48894, total 15 columns:

#	Column	Non-Null Count	Data Type
0	id	48149 non-nulls	int64
1	name	48149 non-nulls	object
2	host_id	48149 non-nulls	int64
3	host_name	48149 non-nulls	object
4	neighbourhood_group	48149 non-nulls	object
5	neighbourhood	48149 non-nulls	object
6	latitude	48149 non-nulls	float64
7	longitude	48149 non-nulls	float64
8	room_type	48149 non-nulls	object
9	price	48149 non-nulls	int64
10	minimum_nights	48149 non-nulls	int64
11	number_of_reviews	48149 non-nulls	int64
12	reviews_per_month	48149 non-nulls	float64
13	calculated_host_listings_count	48149 non-nulls	int64
14	availability_365	48149 non-nulls	int64

5. Exploratory analysis

We Explore and analyze the data to discover key understandings such as :

a) What can we learn about different hosts and areas?
b) What can we learn from predictions? (ex: locations, prices, reviews, etc)
c) Which hosts are the busiest and why?
d) Is there any noticeable difference of traffic among different areas and what could be the reason for it?

a) What can we learn about different hosts and areas?

- Using **Scatter plot** to visualize the different areas
- Using `df.groupby()` function we aggregate the *host_name* for *calculated_host_listing_count* and using **bar chart** to visualize the top hosts.
- Using `df.value_conts()` *for neighburhood_group* and **Pie Chart** analyze the booking percentage in different neighborhood group.
- Using `df.value_conts()` *for neighburhood* and **Horizontal Bar Graph** to visualize the Top 10 Neighbor by Demand.

b) <u>What can we learn from predictions? (ex: locations, prices, reviews, etc)</u>

- **Scatter Plot** with price column as Color bar parameter is used to visualize the price variation at different areas.
- **Classic Bar plot** to visualize the *price* of different room type in different *neighbourhood_group*.
- To Understand the Price Distribution of different *neighbourhood_group* we use a very beautiful **Violin Plot**. The use case is more the symmetry better the distribution.
- Using `df.groupby()` function we aggregate *neighbourhood_group* by sum of *number_of_reviews* then visualize using the P**ie Chart.**

c) Which hosts are the busiest and why?

Busiest hosts would be those who have maximum number of reviews as people are booking frequently at those hosts or One having highest total no minimum night stay.

- To understand the busiest host based on *number_of_reviews* we group the dataset using `df.groupby()` function and sort the table using `df.sort()` by *number_of_reviews* in descending order, then visualize using **Vertical Bar Graph.**
- To understand the busiest host based on *minimum_nights* we group the dataset using `df.groupby()` function and aggregate at *minimum_nights* to get the total night stay at different host, and finally visualize using **Vertical Bar Graph.**
- To confirm the observation "Busiest Host mostly are those who hosting Either Private rooms or Entire Home/Apt" we Horizontal Bar Graph for different room type using `df.value_conts()` for *room_type* column.

d) <u>What can we learn about different hosts and areas?</u>

- Using Pie Chart to visualize the percentage of room preferred
- Using **countplot bar chart** with hue = *room_type* analyzing the different total no of different *room_type* in different *neighbourhood_group*
- Using Bar Plot with `df.groupby()` and aggregate by median minimum_nights to analyze the minimum night stay at different room types.
- To analyze the availability of different room type we use Box Plot for visualization.

6. Conclusion:

- We can conclude from the analysis that Manhattan is the Top neighbourhood group by number of listings and highest rental prices 7 out of 10 top Host are from Manhattan followed by Brooklyn. One of the Probable reasons for most preferred Neighbour Group is that Manhattan is a world-famous for its museums, stores, parks and theatres - and its substantial number of tourists thus attract Entire Home/Apt as favourite stay options and stayed longer, as demand is high prices are much higher in this borough.
- Brooklyn also has significant number of bookings because in Brooklyn some famous bridges, parks, museums, islands and other tourist places but with more affordable prices as compared to Manhattan. It also received the maximum number of reviews.
- Other Three neighbourhood groups namely Queens, Bronx and Staten island are observing very less listing options available, especially on Staten Island. Considering that those are residential areas, it is possible that many guests choose these locations to save up money or perhaps to visit family and friends who live in this area.