Project Report

Airbnb Data Analysis

BY Ajit kumar

1. General Description

1.1 Product Perspective & Problem Statement:

Since 2008, guests and hosts have used Airbnb to expand on travelling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in Amsterdam, Netherland for 2019. Content This data file includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions**.**

The objective of the project is to perform data visualization techniques to understand the insight of the data. This project aims to apply **Exploratory Data Analysis (EDA**) and **Business Intelligence tools** such as **Power BI** to get a visual understanding of the data.

Objectives:

Research Questions-

Regarding the Host

- Who are top earners
- Is there any relationship between monthly earning and prices

Regarding the Neighbourhood

- Any particular location getting maximum number of bookings
- Price relation with respect to location

Regarding the reviews

- Relationship between Quality and Price

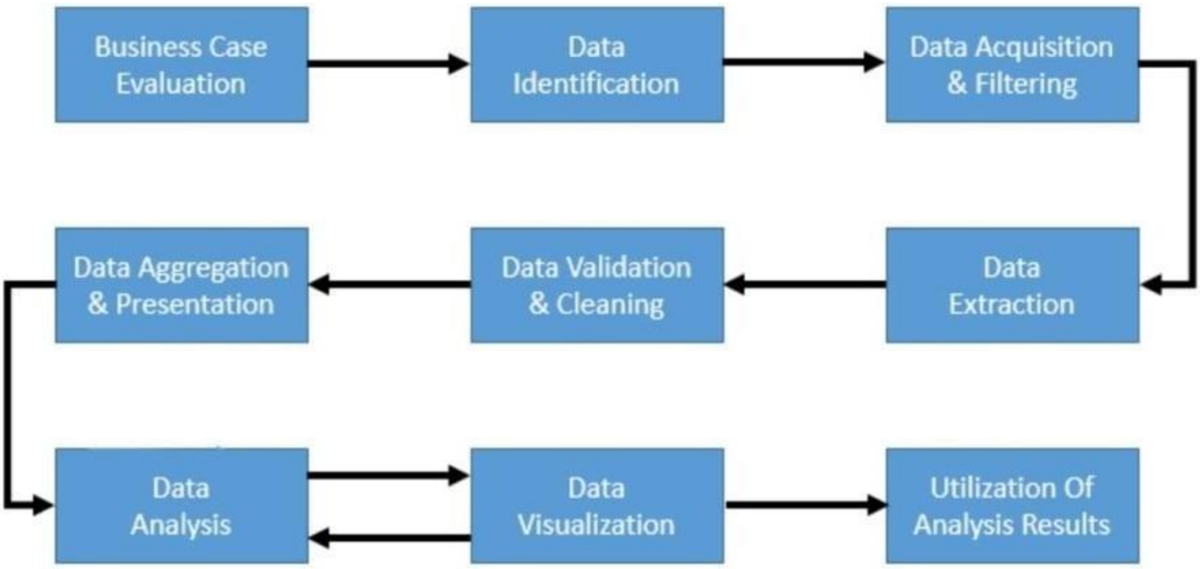
Regarding Price

- Price vs amenities
- Price vs location

1.2 Tools used :

Business Intelligence tools and libraries works such as **Python**-Numpy, Pandas, Seaborn, Matplotlib, Excel, Power BI are **used** to build the whole framework.

2. Design Details


Business Case Evaluation
Data Identification
Data Acquisition & Filtering
Data Extraction
Data Validation & Cleaning
Data Aggregation & Presentation
Data Analysis
Data Visualization
Utilization Of Analysis Results

2.1 Dataset/ Data Acquisition:

The dataset had information regarding the reviews with respect to listing id. This data had all the information regarding the listings. It had Host name, location, neighbourhood, price, review score and number of review, latitude, longitude ,room type.

You can find the dataset on the given link: https://drive.google.com/drive/folders/1ANkgtAT0Pdp2r86IxFKv9vKYmnsYjJDO?usp=sharing

2.2 Data Description:

The dataset had information regarding the reviews with respect to listing id. This data had all the information regarding the listings. It had Host name, location, neighbourhood, price, review score and number of review, latitude, longitude, room type. etc..

The features in the dataset can be described as follows:

1. room id - This is the identity number of the property listed by a particular host.

2.survey id - This is the identity number of survey.

3. name - It stands for the name of the property listed by the host.

4. Host id - It is the identity number of the hosts who have registered on Airbnb website.

5. room type - This represent the various types of room listed by host.

6. Country – name of the country where the survey has conducted.

7. City - name of the city where the survey has conducted.

8. neighbourhood- These are the names of the neighbourhood or locations present in the city.

9. latitude - These represent the coordinates of latitude of the property listed.

10. longitude - These represent the coordinates of longitude of the property listed.

11. price - This is the rent of the property listed in euro.

12. ministay - This represent the minimum number of nights customer rented the property.

13. reviews - This represent the number of customers reviewed the property.

14. overall_satisfaction – customars has given a rating to a places in 0 to 5.

15. Location – it has given a code of the locations.

16. Bedrooms – no. of bedrooms present In the property.

17. Bathrooms – no. of bathrooms present in the property.

18. Last modification - This represent the date when the property was last reviewed.

2.3 Data Transformation:

In the Transformation Process, we will Transform our original datasets excel fil into jupyter notebook for data Exploration and performing Exploratory Data Analysis using python programming language

2.4 Exploratory Data Analysis (EDA)

- EDA Using Python

Exploratory Data Analysis is an approach to analyse the datasets to summarize their main characteristics in form of visual methods.
EDA is nothing but a data exploration technique to understand various aspects of the data. The main aim of EDA is to obtain confidence in a data to an extent where we are ready to engage a machine learning model.
EDA is important to analyse the data it's a **first steps in data analysis process**.
Exploratory data analysis help us to finding the errors, discovering data, mapping out data structure, finding out anomalies. Exploratory data analysis is important for business process because we are preparing dataset for deep through analysis that will detect you business problem.

Steps Involved in EDA :-

- Data Sourcing
- Data Cleaning
- Univariate analysis with visualisation
- Bivariate analysis with visualisation
- Derived metrics

Data Sourcing: Data Sourcing is the process of gathering data from multiple sources as external or internal data collection.
There are two major kind of data which can be classified according to the source: 1. Public data 2. Private data

Data Cleaning : After collecting the data , the next step is data cleaning. Data cleaning means that you get rid of any information that doesn’t need to be there and clean up by mistake.
Data Cleaning is the process of clean the data to improve the quality of the data for further data analysis and building a machine learning model. The benefit of data cleaning is that all the incorrect and irrelevant data is gone and we get the good quality of data which will help in improving the accuracy.

analysis with visualisation: Visualisation is the presentation of the data in the graphical or visual form to understand the data more clearly. Visualisation is easy to understand the data.
Easily analyse the data and summarize it. Easily understand the features of the data. Help to find the trend or pattern of the data. Help to get meaningful insights from the data.

- Important Charts for Visualisation:

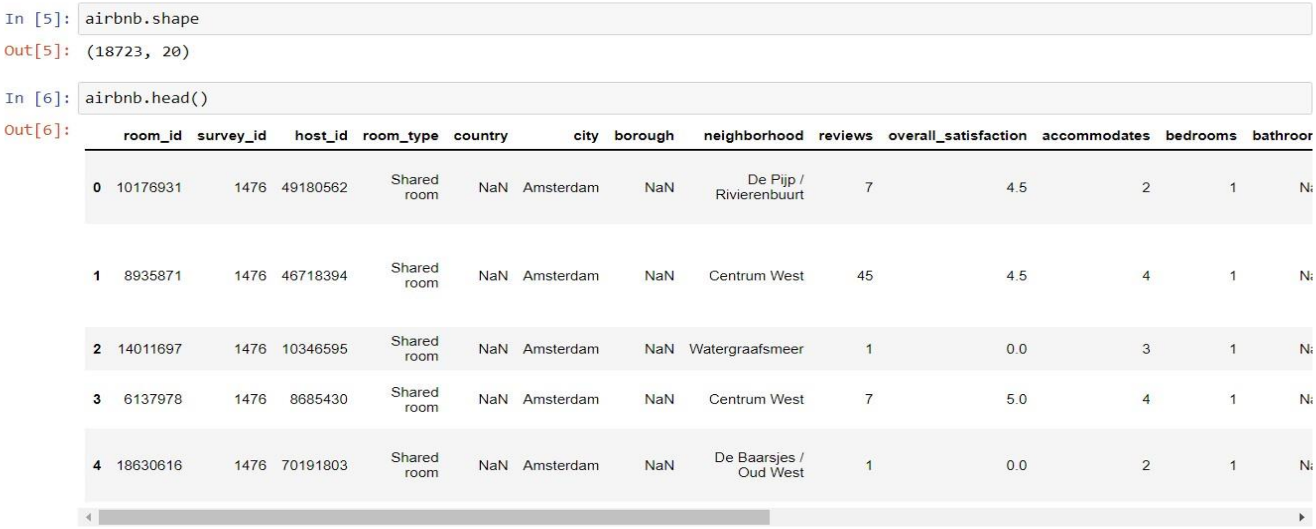
  1.Histogram
  2.Bar Chart
  3.Box plot
  4.pie chart
  5.Heatmap
  6.Scatter plot
  7.Line chart etc..

For Detailed Understanding of EDA process, prefer a Low Level Design Document :

https://drive.google.com/file/d/1lSIUPXfGMSA1E6T0lgH2_shByD5f1/view?usp=share_link

Data Exploration:

Checking the first 5 rows of the dataset and the dataset consist of 18723 observations (rows) and 20 features (columns).



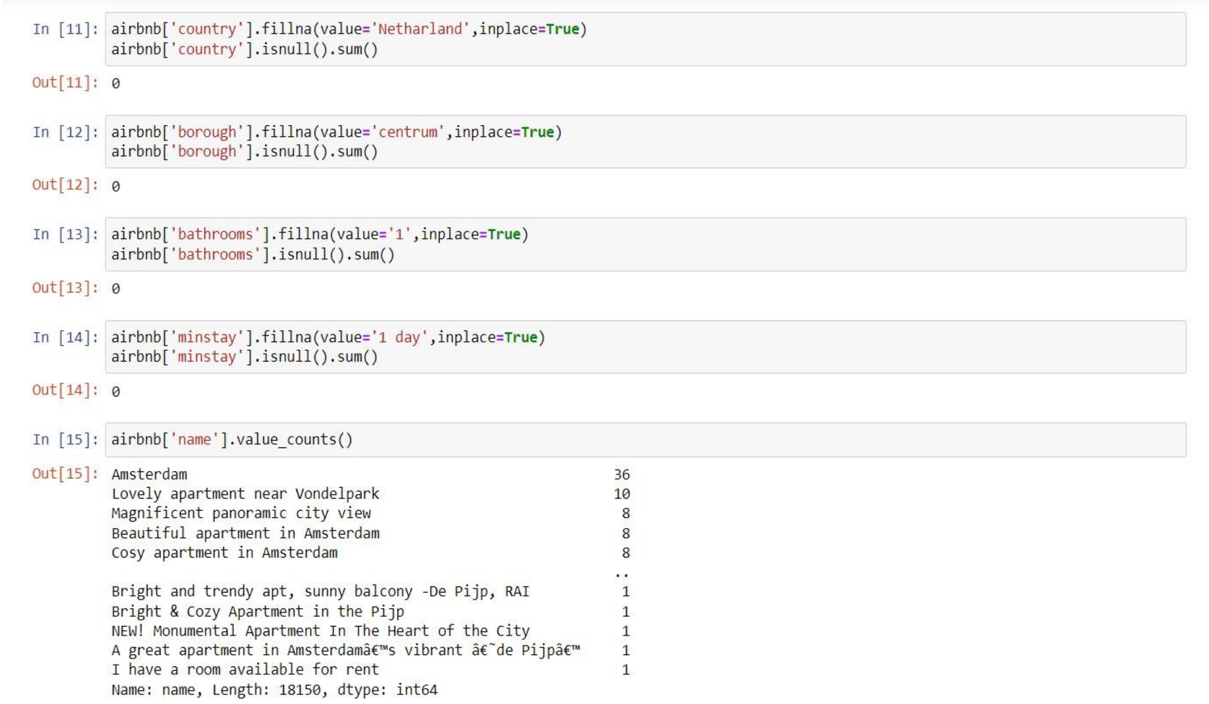
```
In [5]: airbnb.shape
Out[5]: (18723, 20)

In [6]: airbnb.head()
Out[6]:
```

	room_id	survey_id	host_id	room_type	country	city	borough	neighborhood	reviews	overall_satisfaction	accommodates	bedrooms	bathroor
0	10176931	1476	49180562	Shared room	NaN	Amsterdam	NaN	De Pijp / Rivierenbuurt	7	4.5	2	1	Na
1	8935871	1476	46718394	Shared room	NaN	Amsterdam	NaN	Centrum West	45	4.5	4	1	Na
2	14011697	1476	10346595	Shared room	NaN	Amsterdam	NaN	Watergraafsmeer	1	0.0	3	1	Na
3	6137978	1476	8685430	Shared room	NaN	Amsterdam	NaN	Centrum West	7	5.0	4	1	Na
4	18630616	1476	70191803	Shared room	NaN	Amsterdam	NaN	De Baarsjes / Oud West	1	0.0	2	1	Na

Data Cleaning:

Wrong data can skew and distort analysis results. Data input into Big Data analyses can be disorganized without any evidence of validity. Its complexity can more make it difficult to come at a set of proper validation constraints. This phase sets often complex validation rules and eliminates any known wrong data.

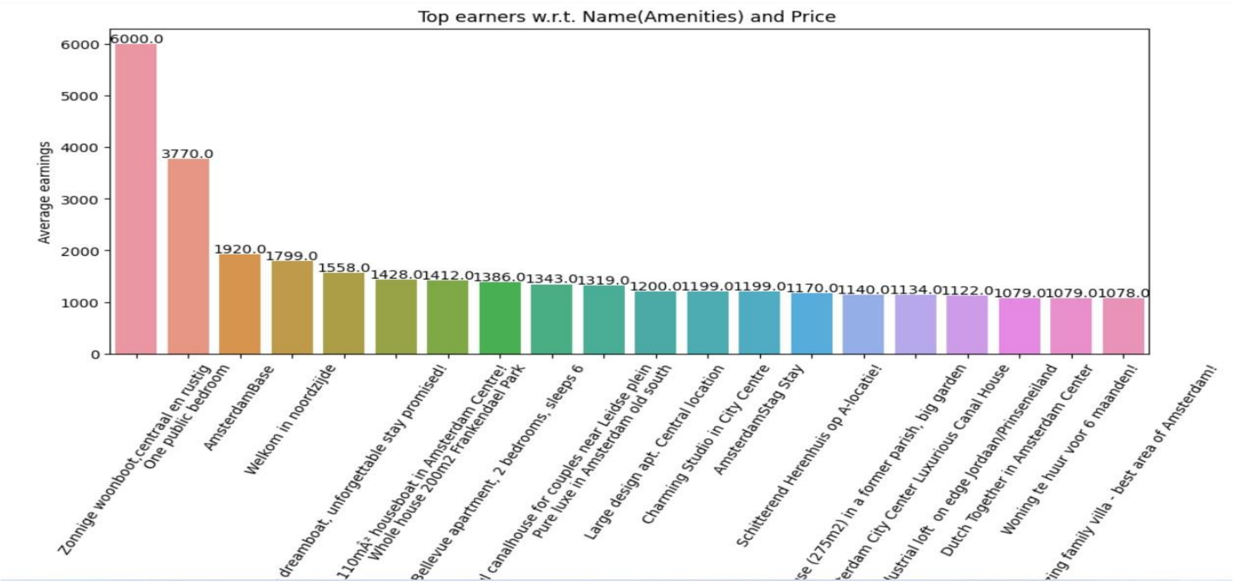
Fixing the null values: We have filled the null values i.e for country-Netharland, Borough-centrum, bathroom-1, mainstay-1day, name-apartment/shared/private room.

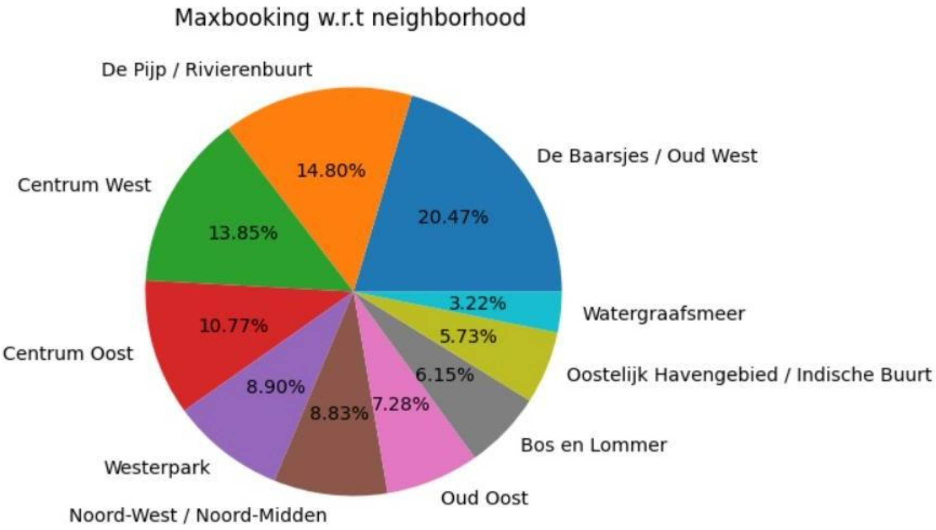


```
In [11]: airbnb['country'].fillna(value='Netharland',inplace=True)
         airbnb['country'].isnull().sum()

Out[11]: 0

In [12]: airbnb['borough'].fillna(value='centrum',inplace=True)
         airbnb['borough'].isnull().sum()

Out[12]: 0

In [13]: airbnb['bathrooms'].fillna(value='1',inplace=True)
         airbnb['bathrooms'].isnull().sum()

Out[13]: 0

In [14]: airbnb['minstay'].fillna(value='1 day',inplace=True)
         airbnb['minstay'].isnull().sum()

Out[14]: 0

In [15]: airbnb['name'].value_counts()

Out[15]: Amsterdam                                                 36
         Lovely apartment near Vondelpark                          10
         Magnificent panoramic city view                            8
         Beautiful apartment in Amsterdam                           8
         Cosy apartment in Amsterdam                                8
                                                                   ..
         Bright and trendy apt, sunny balcony -De Pijp, RAI         1
         Bright & Cozy Apartment in the Pijp                        1
         NEW! Monumental Apartment In The Heart of the City         1
         A great apartment in Amsterdamâ€™s vibrant â€˜de Pijpâ€™    1
         I have a room available for rent                           1
         Name: name, Length: 18150, dtype: int64
```

Data Analysis and Data Visualization:

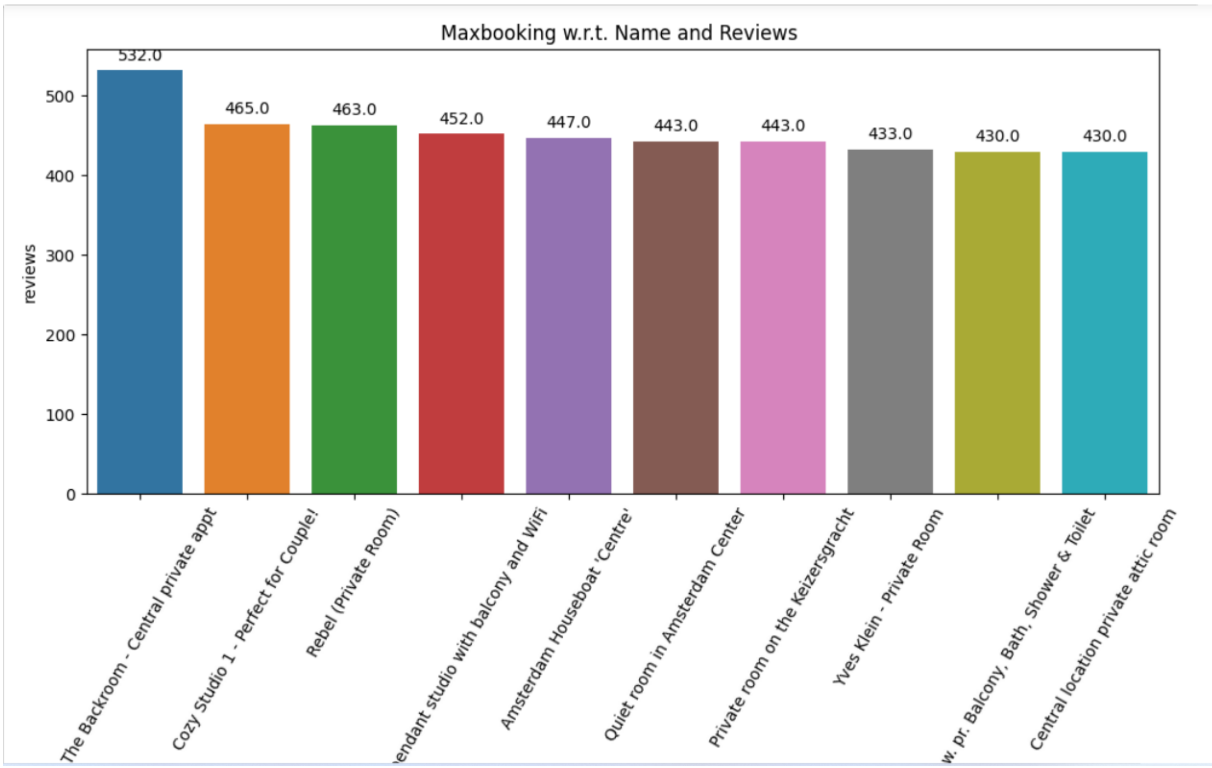
Used to graphically demonstrate the analysis results for effective analysis by company users. Present users with the ability to make visual analysis, allotting for the discovery of answers to questions that users have not yet even formed. The same results may be performed in a number of various ways, which can change the presentation of the results. Use the most proper visualization technique by keeping the business domain in context.

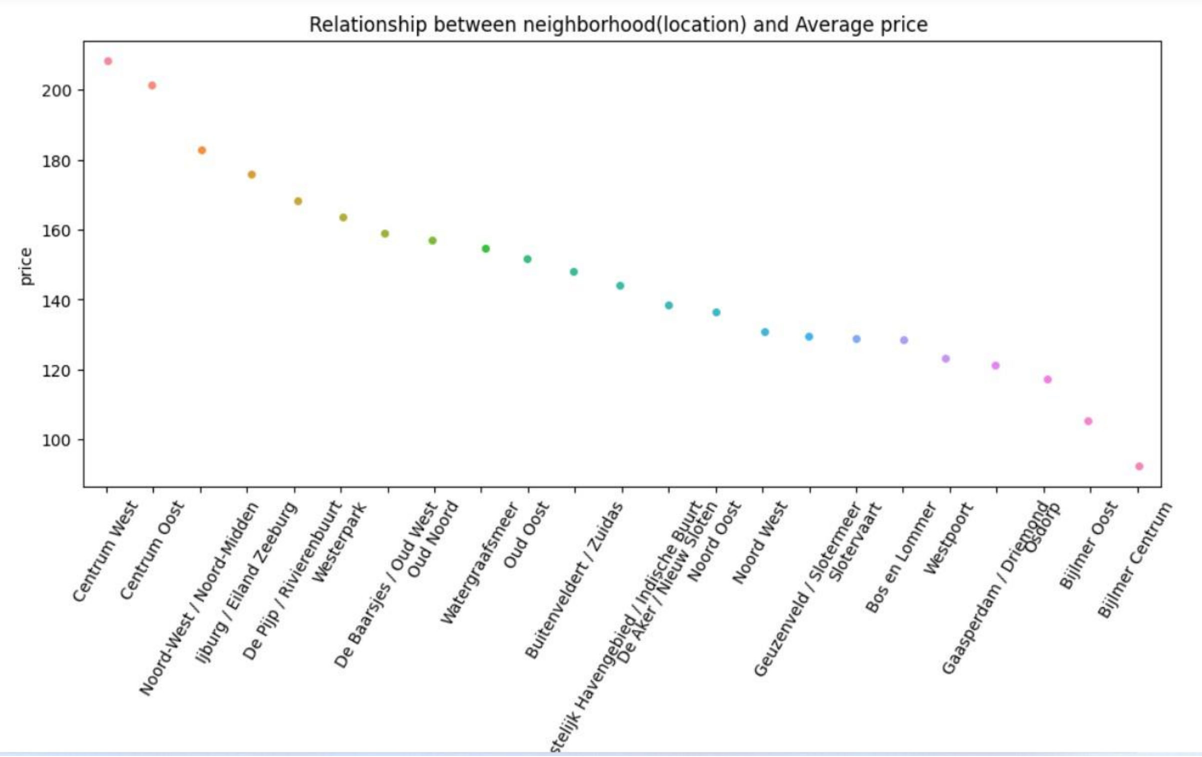


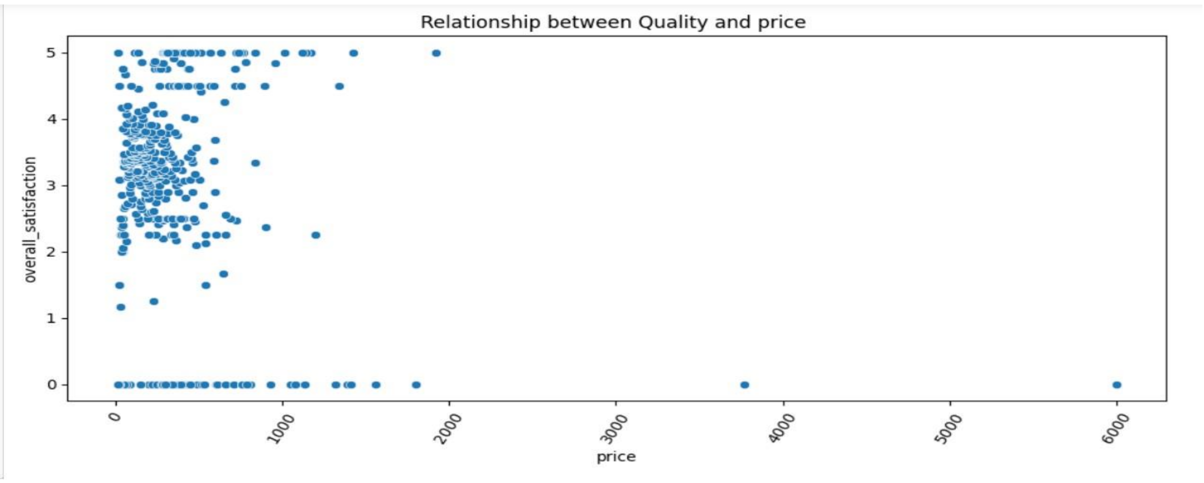


Conclusion:

'De Baarsjes / Oud West' place is having a maximum no. of bookings of 20.47% and having a max. count of 3289

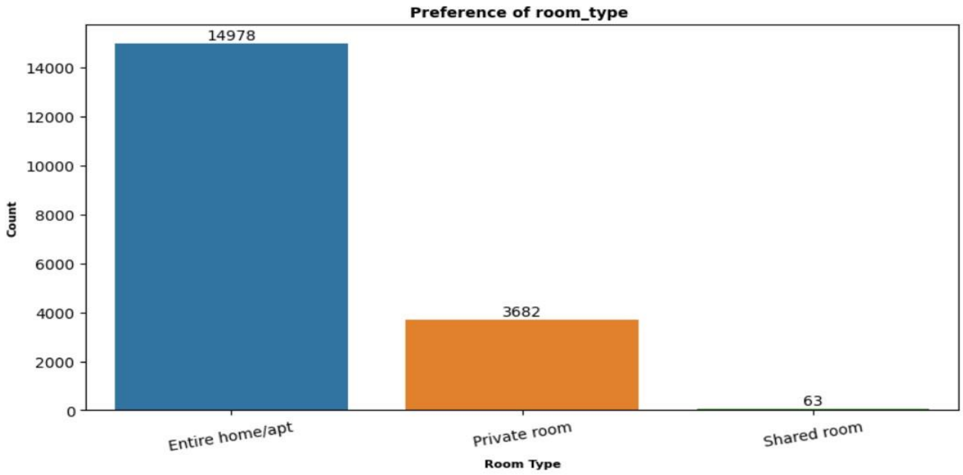

Maxbooking w.r.t. Name and Reviews
reviews
0
100
200
300
400
500
532.0
465.0
463.0
452.0
447.0
443.0
443.0
433.0
430.0
430.0
The Backroom - Central private appt
Cozy Studio 1 - Perfect for Couple!
Rebel (Private Room)
endant studio with balcony and WiFi
Amsterdam Houseboat 'Centre'
Quiet room in Amsterdam Center
Private room on the Keizersgracht
Yves Klein - Private Room
w. pr. Balcony, Bath, Shower & Toilet
Central location private attic room


Relationship between neighborhood(location) and Average price
price
100
120
140
160
180
200
Centrum West
Centrum Oost
Noord-West / Noord-Midden
IJburg / Eiland Zeeburg
De Pijp / Rivierenbuurt
Westerpark
De Baarsjes / Oud West
Oud Noord
Watergraafsmeer
Oud Oost
Buitenveldert / Zuidas
stelijk Havengebied / Indische Buurt
De Aker / Nieuw Sloten
Noord Oost
Noord West
Geuzenveld / Slotermeer
Slotervaart
Bos en Lommer
Westpoort
Gaasperdam / Driemond
Osdorp
Bijlmer Oost
Bijlmer Centrum



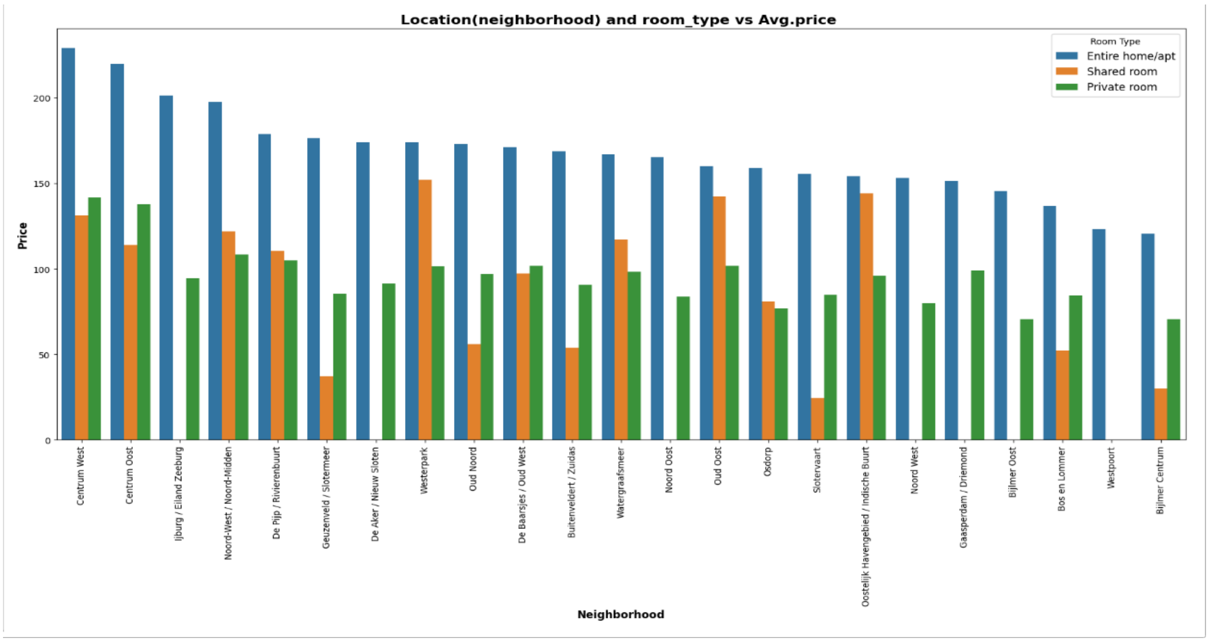
Conclusion:

In the above scatterplot plot we can see that if the 'price' is higher than the 'overall_satisfaction(quality)' is less and where the 'price' is less than the 'overall_satisfaction(quality)' is high. For example: price=313 then the overaoverall_satisfaction(quality) is 5.0 and in other side price=6000 then the overall_satisfaction(quality) is 0.0



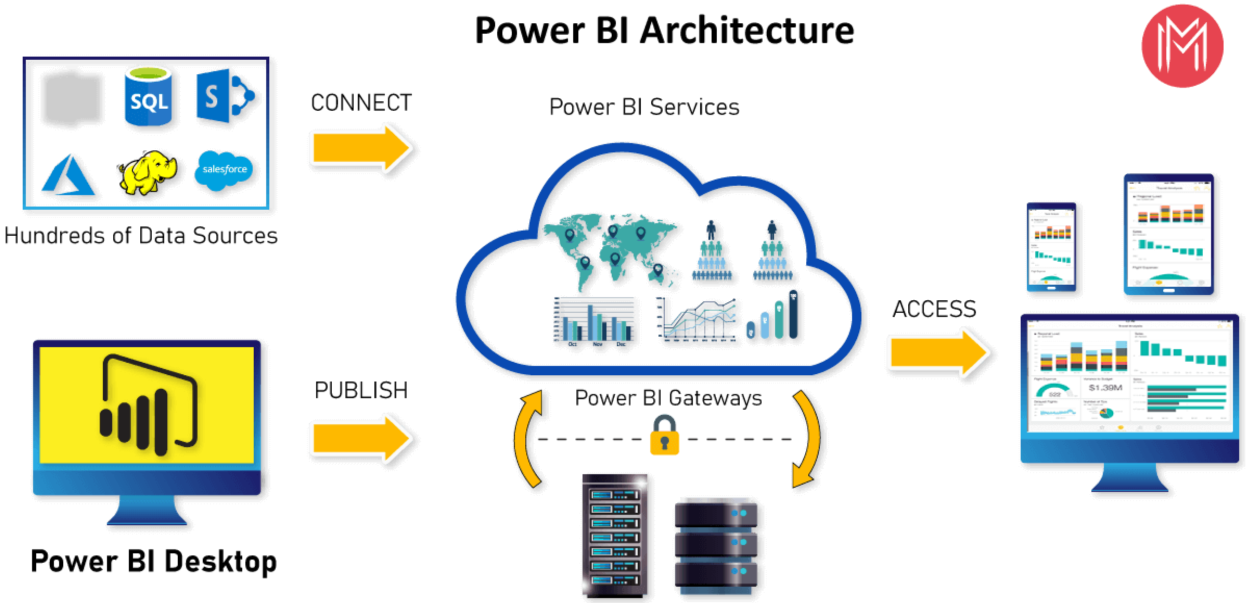
Conclusion:

From the above visualization we can clearly see that the most prefered room type by the guests is Entire home/apt and the less prefered room type is shared room and private room.

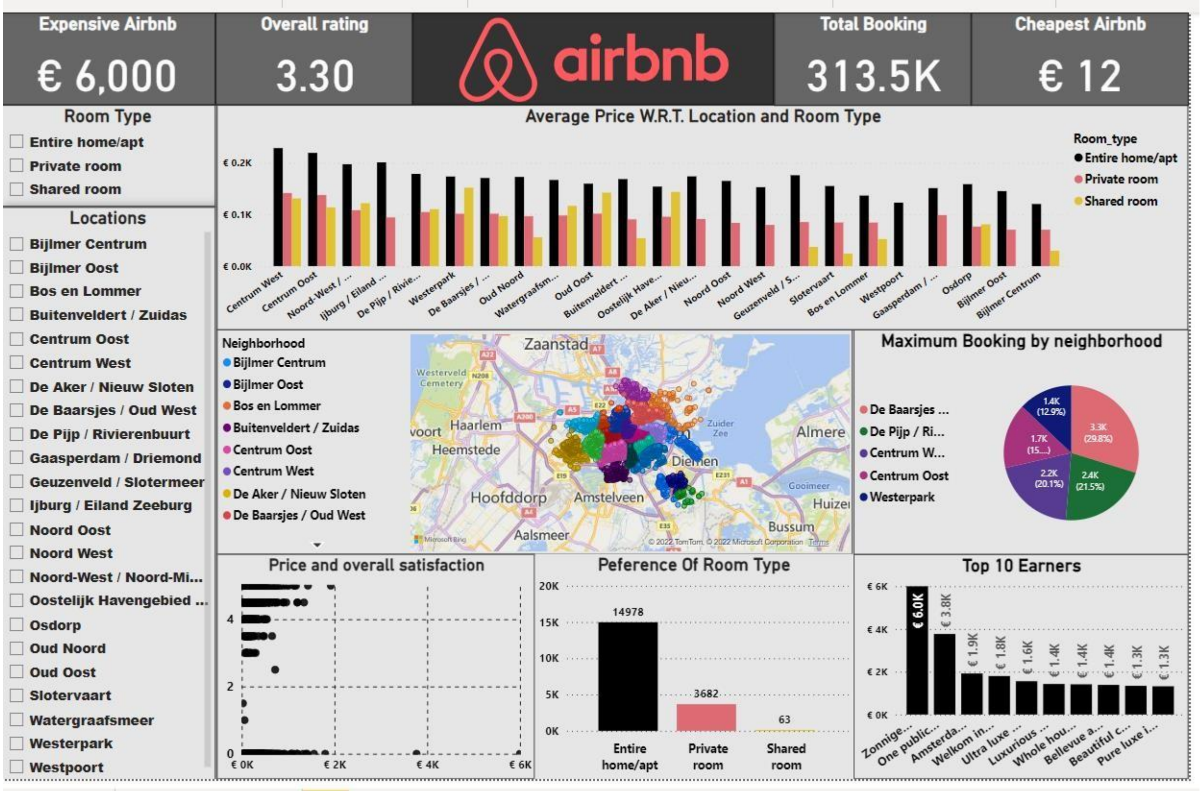


Business Intelligence Tool

Power BI Architecture:



Creating power bi report for better understanding of dataset and for the stakeholders to understand the data in a better way and solving the business problems and taking a right decisions for increasing the profitability.



Q & A:

Q1) What’s the source of data?
The dataset had information regarding the reviews with respect to listing id.
This data had all the information regarding the listings. It had Host name, location, neighbourhood, price, review score and number of review, latitude, longitude ,room type.

You can find the dataset on the given link:
https://drive.google.com/drive/folders/1ANkgtAT0Pdp2r86IxFKv9vKYmnsYjJDO?usp=sharing

Q 2) What was the type of data?
The data was the combination of numerical and Categorical values.

Q 3) What’s the complete flow you followed in this Project?
Refer slide 4th for better Understanding

Q 6) What techniques were you using for data pre-processing?

- Describing the data features of dataset.
- Removing unwanted attributes.
- Visualizing relation of independent variables with each other and output variables.
- Checking and changing Distribution of continuous values.
- Cleaning data and imputing if null values are present.

Q 7) How training was done or what models were used?

- Before diving the data in visualization set we performed Analysis of problem statements.
- Using various visualization library and different coding in python we try to analyze the data .
- Creating Dashboard Visualization using Business Intelligence Tool power bi.