STORYTELLING CASE STUDY – AIRBNB NYC

CONTRIBUTORS:

SARAVANAN RAJU

ABHIRAM N

GAYATRI CHAUHAN

AGENDA

- Objective
- Background
- Key Findings
- Recommendations
- Appendix
  - Data Sources
  - Data Methodology
  - Data Model Assumptions

OBJECTIVE

- Conduct thorough analysis of Airbnb NYC dataset.
- Ask effective justifications that can lead to data insights.
- Process, Analyse and share findings by Data Visualization and Statistical Techniques.

BACKGROUND

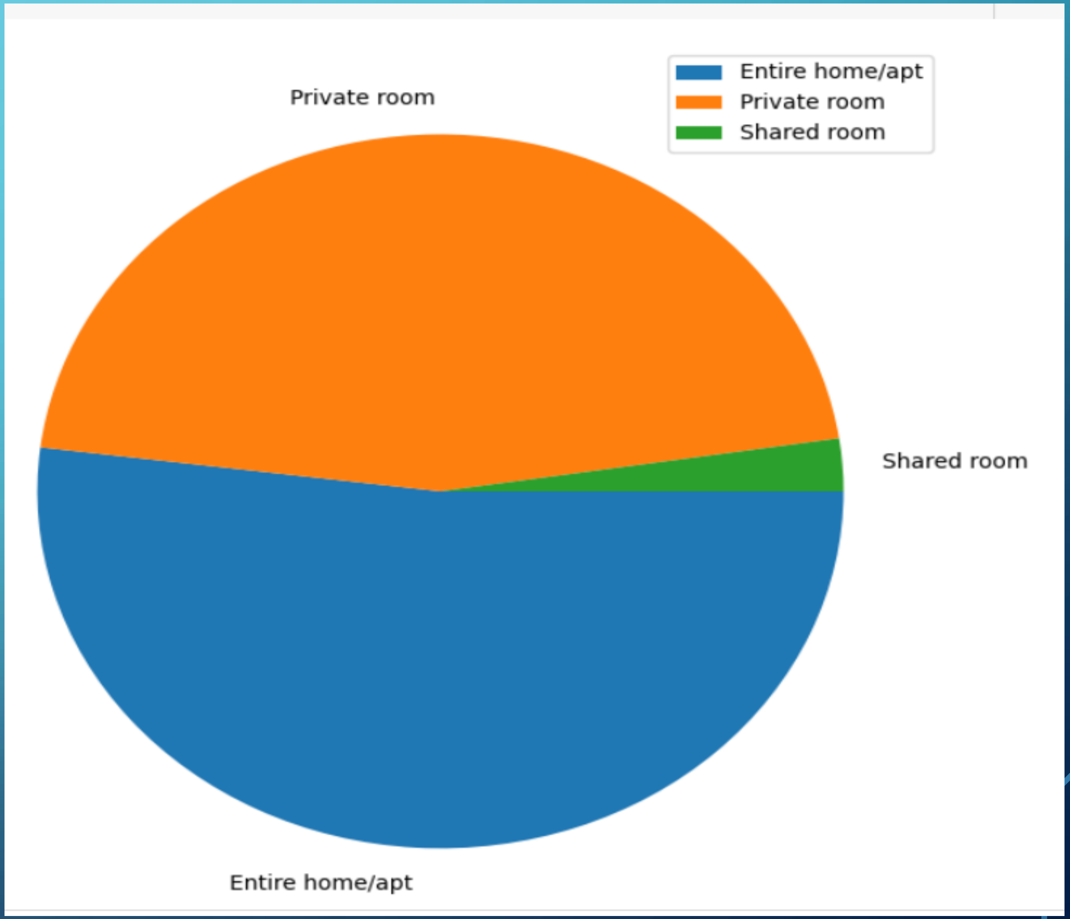
- Airbnb, an online marketplace for short – and long-term homestays and experiences.
- For the past few months, Airbnb has seen a major decline in revenue.
- Now that the restrictions have been lifted and people have started to travel more.
- Airbnb wants to make sure that it is fully prepared for the change.

DATA ANALYSIS STEPS

- In the first phase the data is captured and loaded for cleansing & data preparation.
- Once data is cleaned, Exploratory data analysis(EDA) is done and new features are created.
- Meaningful insights are created using various analytical methods.

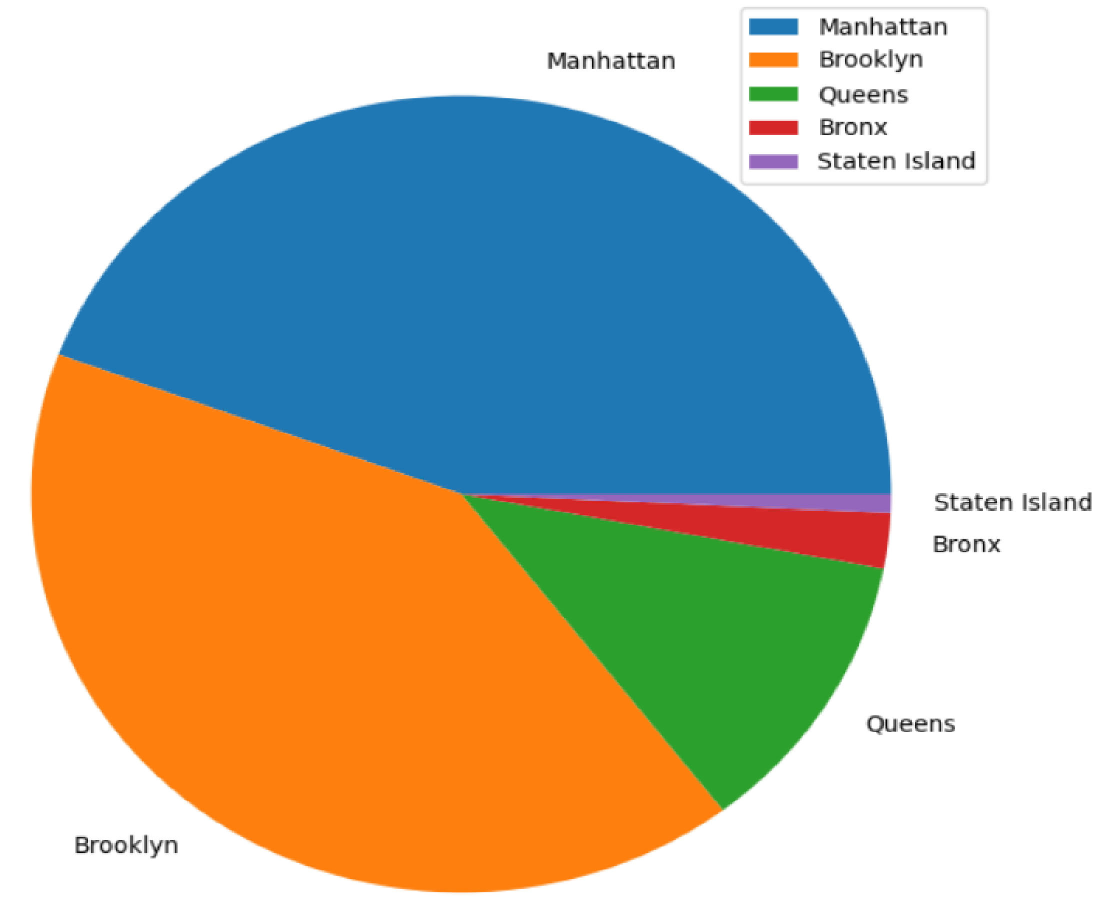
THE PROBLEM WITH SHARED ROOMS

- Shared rooms only account for 2% of the total type of rooms.
- They are less likely to be reviewed.



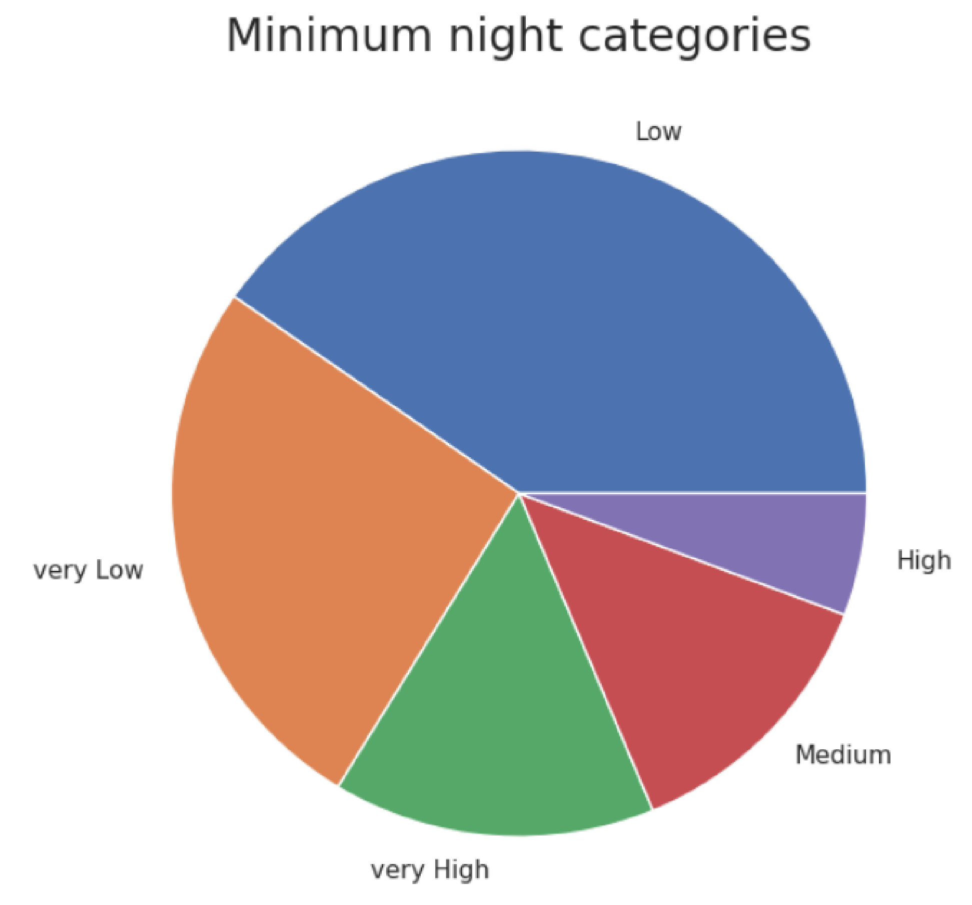
MOST CONTRIBUTING NEIGHBOURHOODS

- 81% of the listings are Manhattan and Brooklyn neighbourhood group.
- Staten Island has the least contribution.



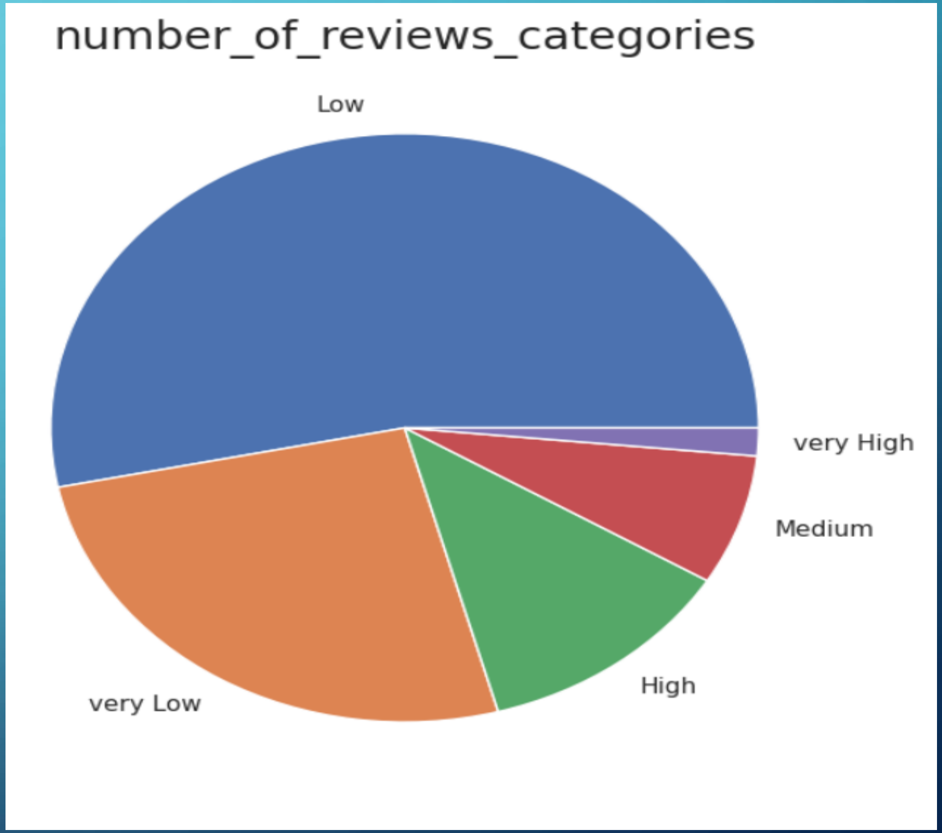
MINIMUM NIGHT CATEGORIES

- Low category in minimum night feature contributes 40%.



EFFECT OF MINIMUM NIGHTS ON REVIEWS

- Customers are more likely to provide reviews for lower number of minimum nights.



KEY FINDINGS AND RECOMMENDATIONS

- Data collection team can collect data about review scores so that it can strengthen the review analysis.
- A Clustering machine learning model to identify groups of similar objects in datasets with two or more variable quantities can be made.
- Shared accommodations has the least preferences. These need to be inspected and customized to private rooms to meet customer demand.
- More than 80% of the listings are in Manhattan and Brooklyn neighbourhood.
- Threshold of minimum nights should be less than 10 nights to make property more customer-oriented.

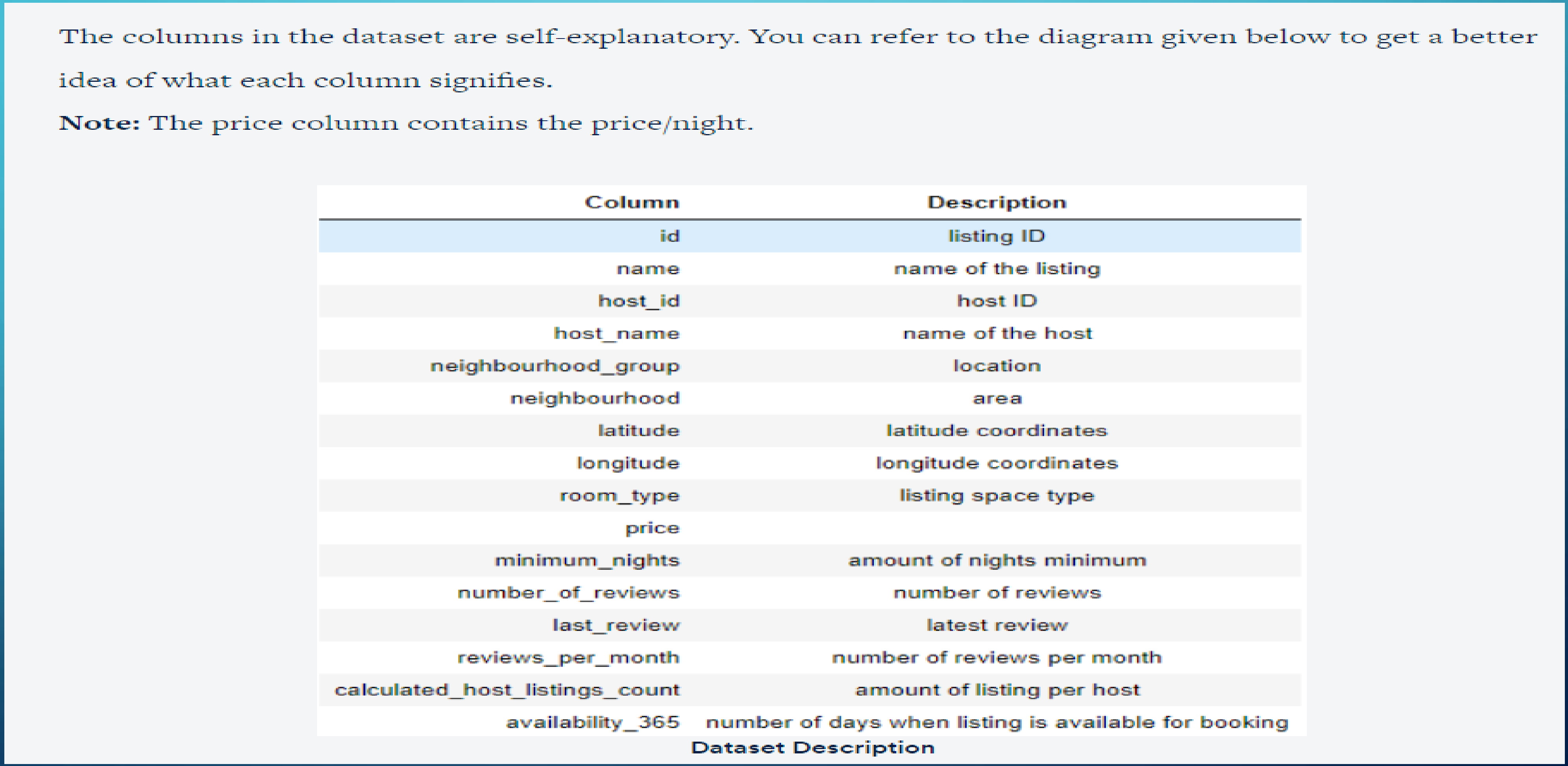
APPENDIX

DATA SOURCES

DATA METHODOLOGY

DATA ASSUMPTIONS

DATA SOURCES



The columns in the dataset are self-explanatory. You can refer to the diagram given below to get a better idea of what each column signifies.

Note: The price column contains the price/night.

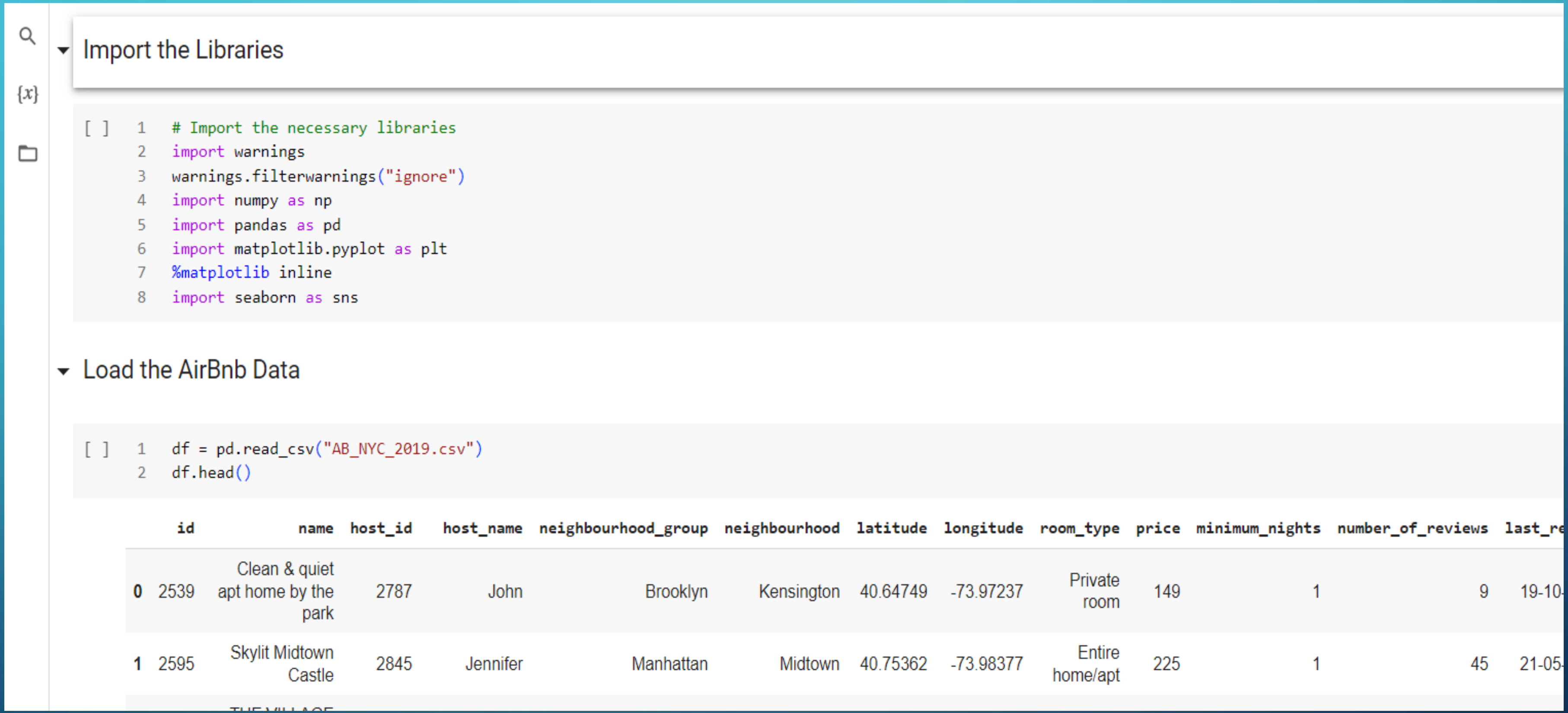
Column	Description
id	listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	location
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type
price	
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

Dataset Description

DATA METHODOLOGY

- Conducted Data Analysis on Airbnb, NYC dataset.
- Data-Cleaning, Preparation & adding features were done through Phyton.
- Used group aggregation, pivot table and other statistical methods.
- Created charts and Visualization through Python & Power-BI

DATA METHODOLOGY – DATA CLEANING/PREPARATION



Import the Libraries

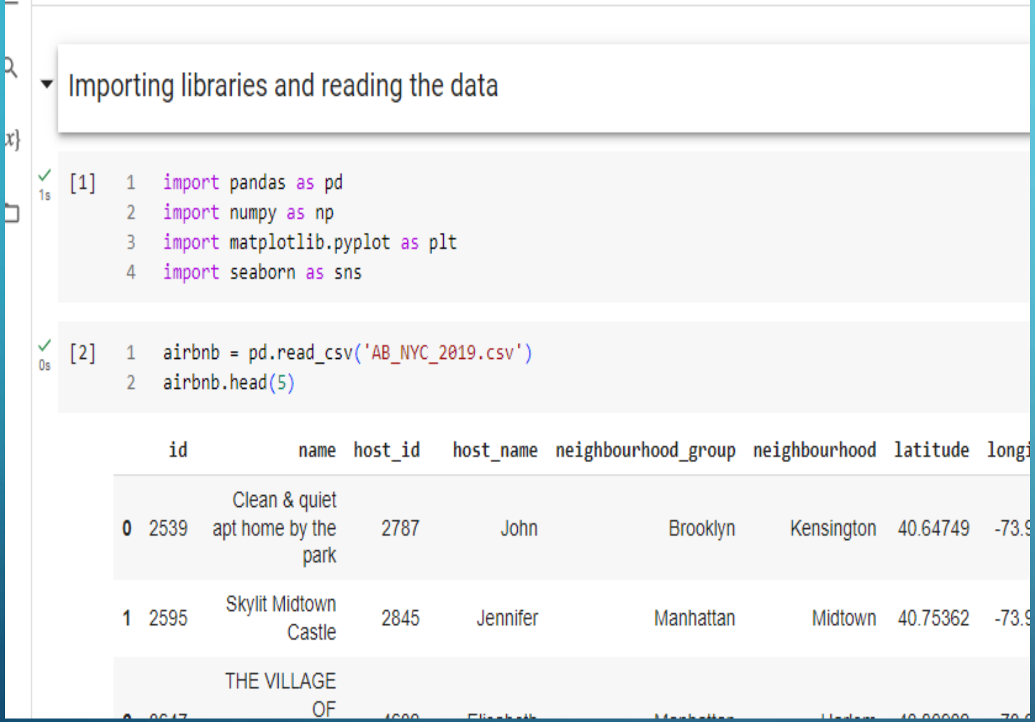
```
# Import the necessary libraries
import warnings
warnings.filterwarnings("ignore")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Load the AirBnb Data

```
df = pd.read_csv("AB_NYC_2019.csv")
df.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_re
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	19-10
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-05
		THE VILLAGE											

EXPLORATORY DATA ANALYSIS

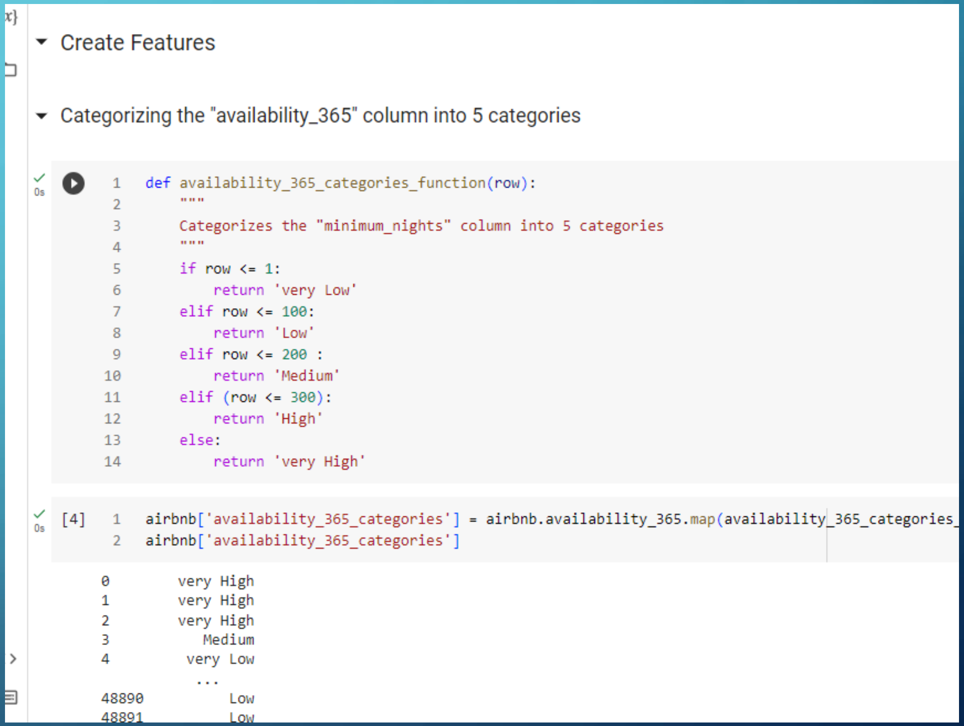


Importing libraries and reading the data

```
[1]
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[2]
airbnb = pd.read_csv('AB_NYC_2019.csv')
airbnb.head(5)
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longi
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.9
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.9
[illegible]	[illegible]	THE VILLAGE OF	[illegible]	Elisabeth	Manhattan	Harlem	[illegible]	[illegible]



Create Features

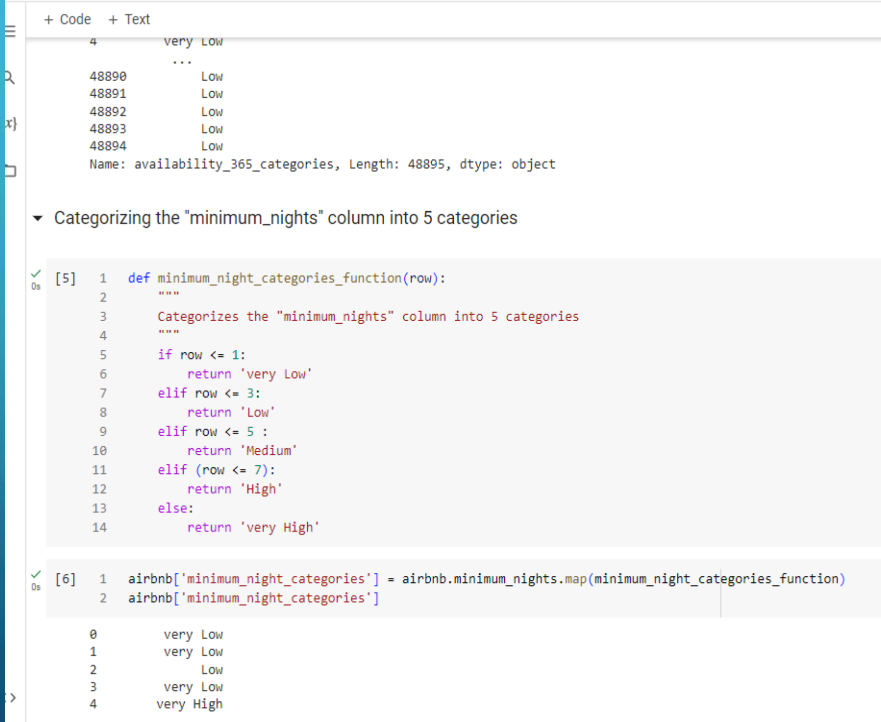
Categorizing the "availability_365" column into 5 categories

```
def availability_365_categories_function(row):
    """
    Categorizes the "minimum_nights" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 100:
        return 'Low'
    elif row <= 200 :
        return 'Medium'
    elif (row <= 300):
        return 'High'
    else:
        return 'very High'
```

```
[4]
airbnb['availability_365_categories'] = airbnb.availability_365.map(availability_365_categories_
airbnb['availability_365_categories']
```

```
0        very High
1        very High
2        very High
3           Medium
4         very Low
           ...
48890          Low
48891          Low
```

EXPLORATORY DATA ANALYSIS – CONTINUED…



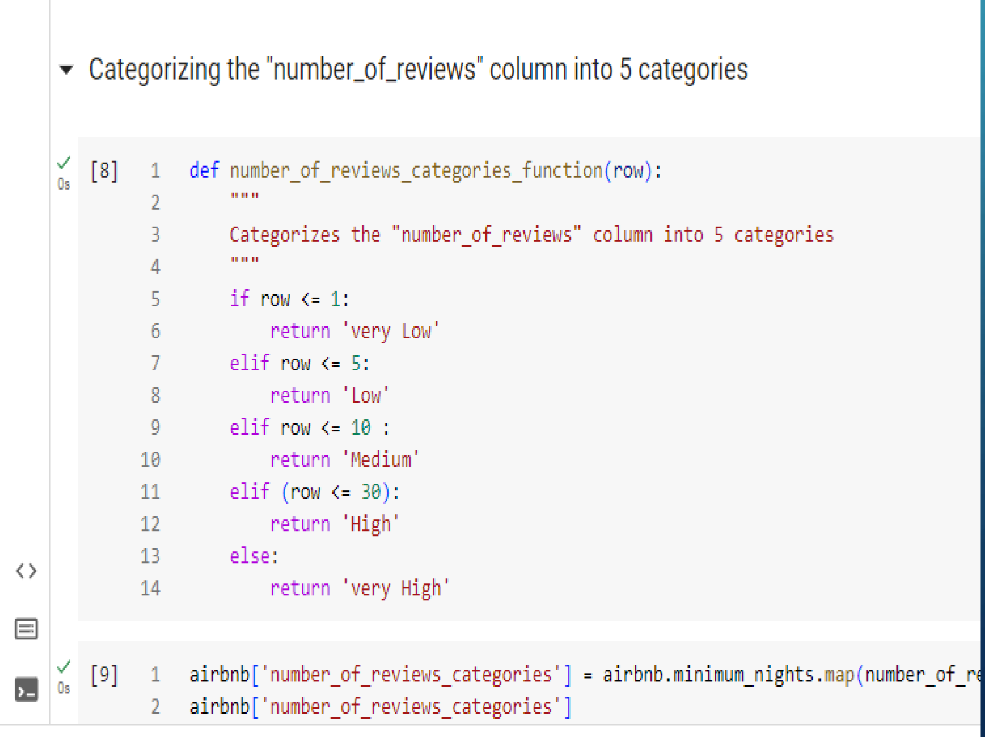
```
4         very Low
            ...
48890         Low
48891         Low
48892         Low
48893         Low
48894         Low
Name: availability_365_categories, Length: 48895, dtype: object
```

Categorizing the "minimum_nights" column into 5 categories

```
[5]
def minimum_night_categories_function(row):
    """
    Categorizes the "minimum_nights" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 3:
        return 'Low'
    elif row <= 5 :
        return 'Medium'
    elif (row <= 7):
        return 'High'
    else:
        return 'very High'
```

```
[6]
airbnb['minimum_night_categories'] = airbnb.minimum_nights.map(minimum_night_categories_function)
airbnb['minimum_night_categories']
```

```
0         very Low
1         very Low
2              Low
3         very Low
4        very High
```

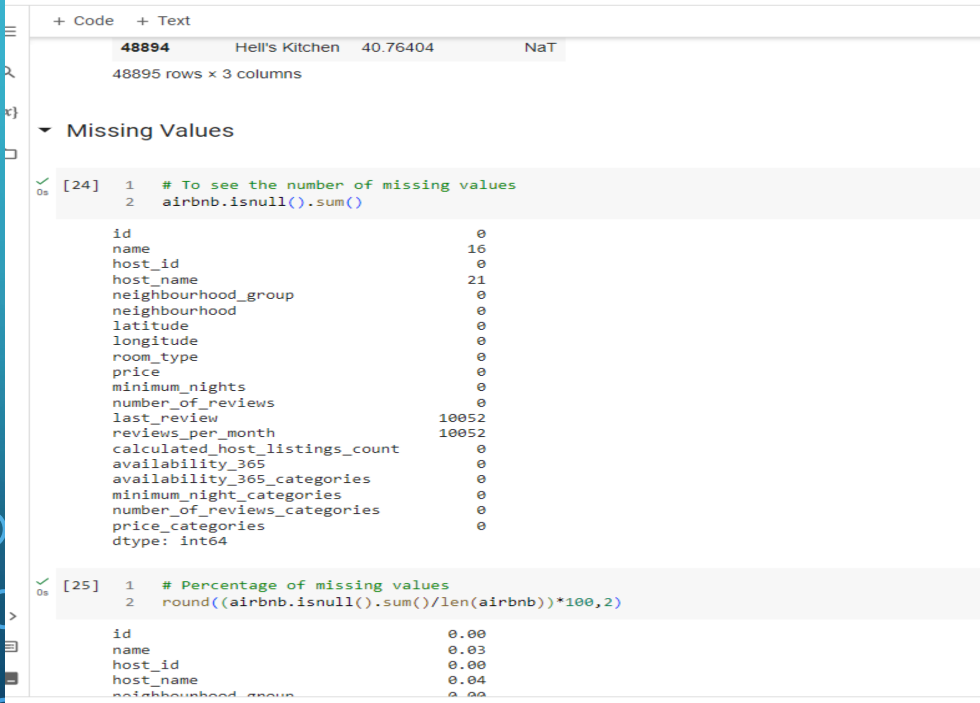


Categorizing the "number_of_reviews" column into 5 categories

```
[8]
def number_of_reviews_categories_function(row):
    """
    Categorizes the "number_of_reviews" column into 5 categories
    """
    if row <= 1:
        return 'very Low'
    elif row <= 5:
        return 'Low'
    elif row <= 10 :
        return 'Medium'
    elif (row <= 30):
        return 'High'
    else:
        return 'very High'
```

```
[9]
airbnb['number_of_reviews_categories'] = airbnb.minimum_nights.map(number_of_re
airbnb['number_of_reviews_categories']
```

EXPLORATORY DATA ANALYSIS – CONTINUED…



```
48894        Hell's Kitchen    40.76404             NaT
48895 rows × 3 columns
```

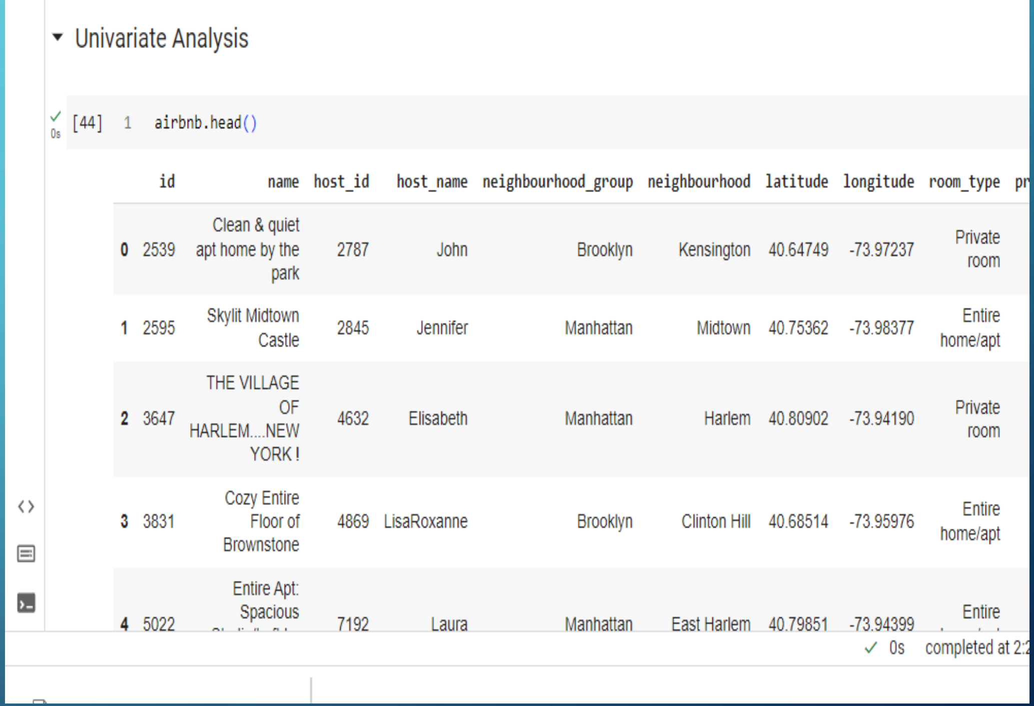
Missing Values

```
[24]
# To see the number of missing values
airbnb.isnull().sum()
```

```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
availability_365_categories           0
minimum_night_categories              0
number_of_reviews_categories          0
price_categories                      0
dtype: int64
```

```
[25]
# Percentage of missing values
round((airbnb.isnull().sum()/len(airbnb))*100,2)
```

```
id                                 0.00
name                               0.03
host_id                            0.00
host_name                          0.04
neighbourhood_group                0.00
```

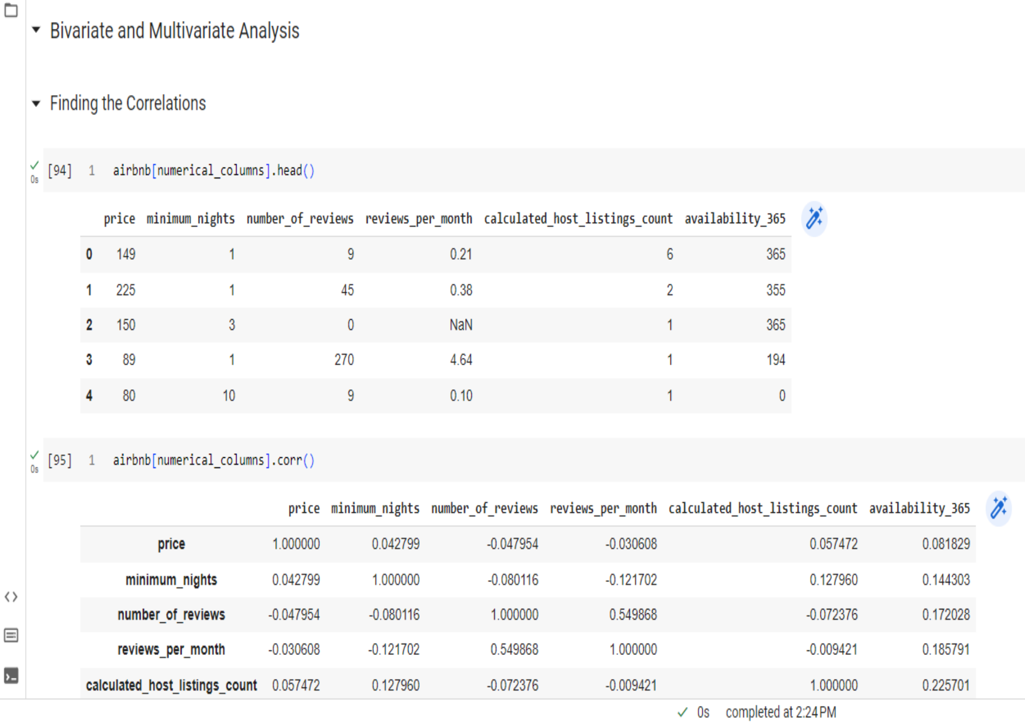


Univariate Analysis

```
[44]
airbnb.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	pr
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	
4	5022	Entire Apt: Spacious	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire	

EXPLORATORY DATA ANALYSIS – CONTINUED…



Bivariate and Multivariate Analysis

Finding the Correlations

[94]

```
airbnb[numerical_columns].head()
```

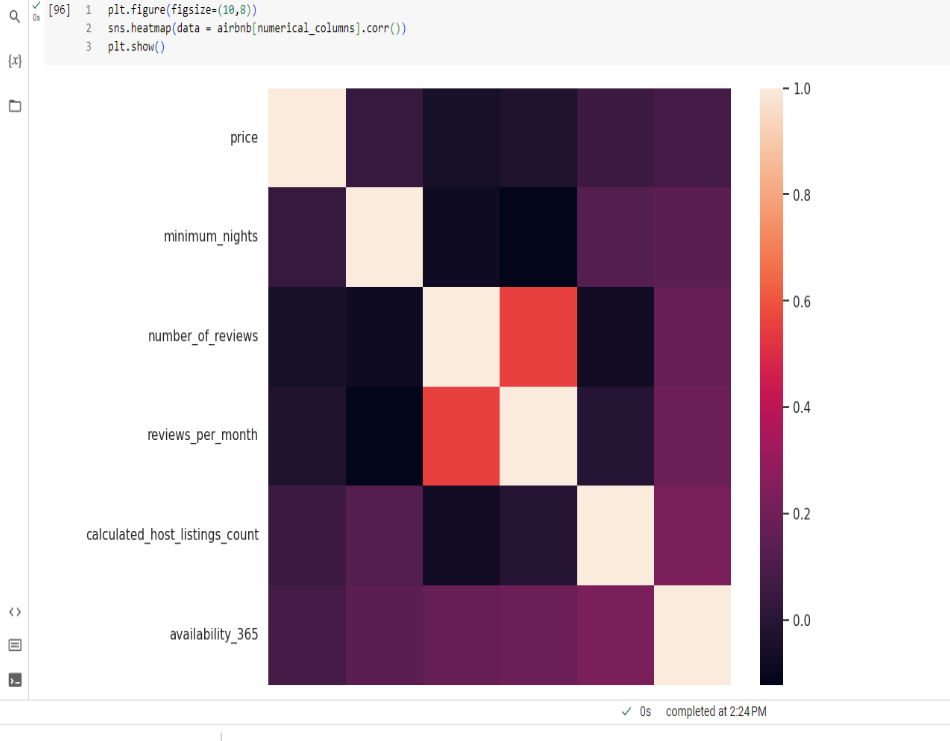
	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0	149	1	9	0.21	6	365
1	225	1	45	0.38	2	355
2	150	3	0	NaN	1	365
3	89	1	270	4.64	1	194
4	80	10	9	0.10	1	0

[95]

```
airbnb[numerical_columns].corr()
```

	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
price	1.000000	0.042799	-0.047954	-0.030608	0.057472	0.081829
minimum_nights	0.042799	1.000000	-0.080116	-0.121702	0.127960	0.144303
number_of_reviews	-0.047954	-0.080116	1.000000	0.549868	-0.072376	0.172028
reviews_per_month	-0.030608	-0.121702	0.549868	1.000000	-0.009421	0.185791
calculated_host_listings_count	0.057472	0.127960	-0.072376	-0.009421	1.000000	0.225701

✓ 0s completed at 2:24 PM

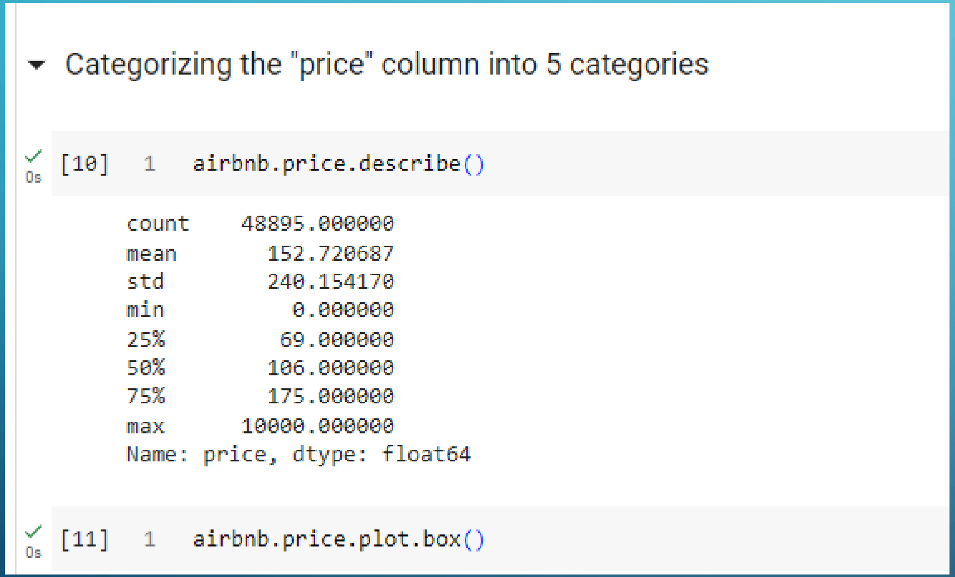


[96]

```
plt.figure(figsize=(10,8))
sns.heatmap(data = airbnb[numerical_columns].corr())
plt.show()
```

✓ 0s completed at 2:24 PM

EXPLORATORY DATA ANALYSIS – CONTINUED…

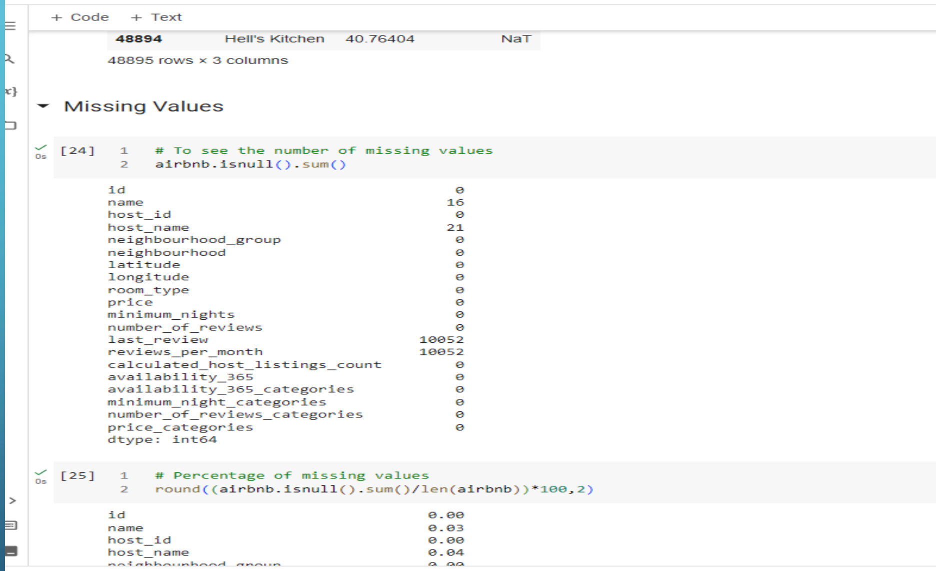


Categorizing the "price" column into 5 categories

```
[10]  airbnb.price.describe()
```

```
count    48895.000000
mean       152.720687
std        240.154170
min          0.000000
25%         69.000000
50%        106.000000
75%        175.000000
max      10000.000000
Name: price, dtype: float64
```

```
[11]  airbnb.price.plot.box()
```



```
48894        Hell's Kitchen    40.76404         NaT
48895 rows × 3 columns
```

Missing Values

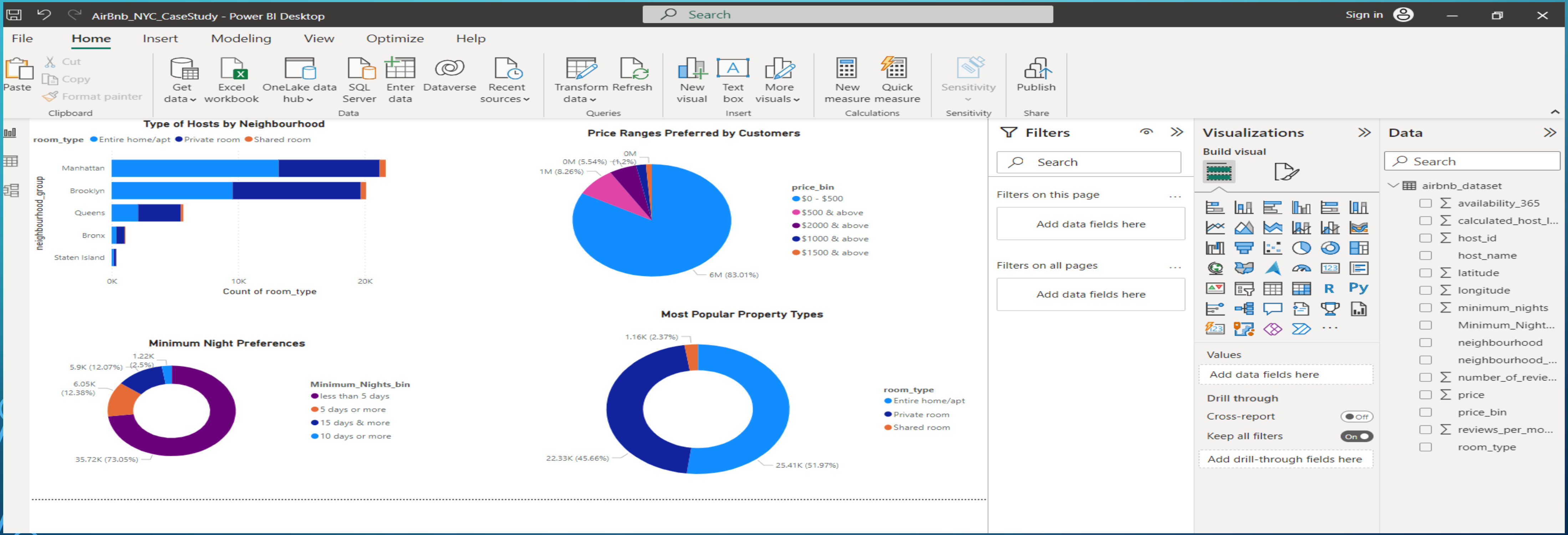
```
[24]  # To see the number of missing values
      airbnb.isnull().sum()
```

```
id                                    0
name                                 16
host_id                               0
host_name                            21
neighbourhood_group                   0
neighbourhood                         0
latitude                              0
longitude                             0
room_type                             0
price                                 0
minimum_nights                        0
number_of_reviews                     0
last_review                       10052
reviews_per_month                 10052
calculated_host_listings_count        0
availability_365                      0
availability_365_categories           0
minimum_night_categories              0
number_of_reviews_categories          0
price_categories                      0
dtype: int64
```

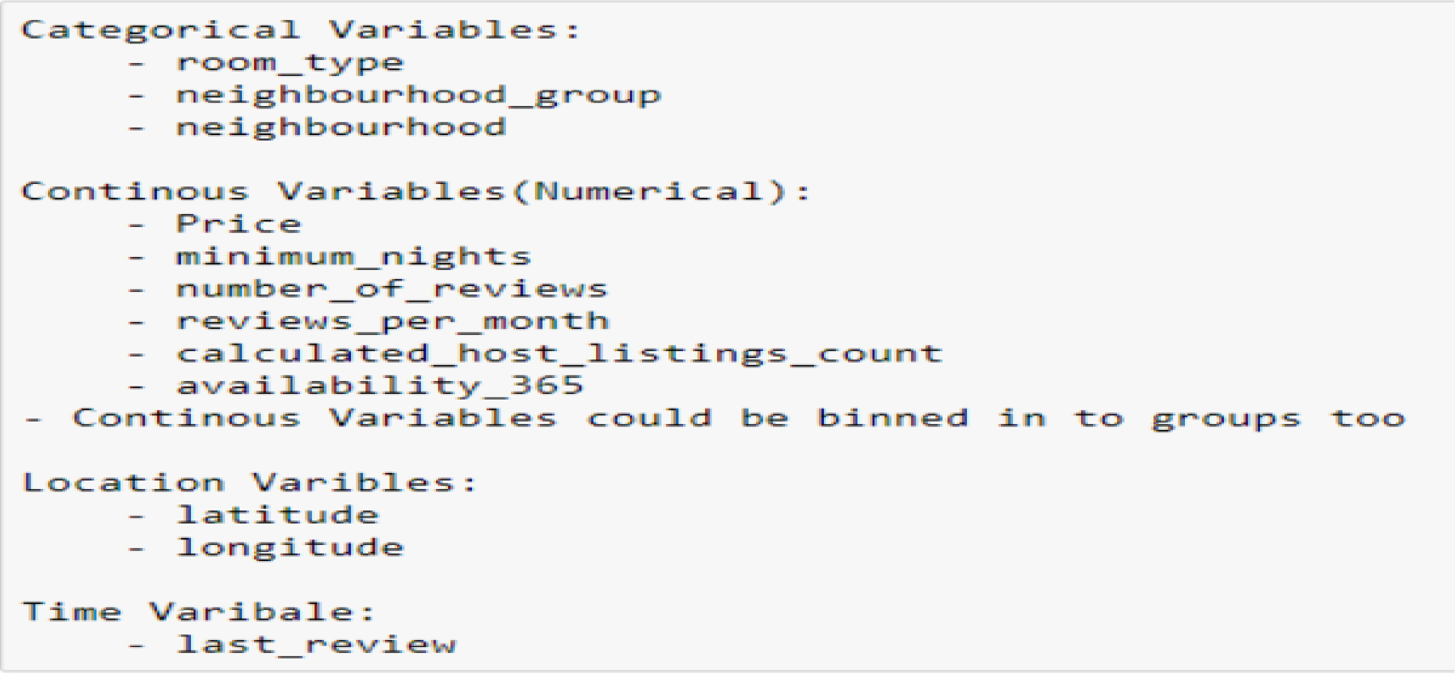
```
[25]  # Percentage of missing values
      round((airbnb.isnull().sum()/len(airbnb))*100,2)
```

```
id                                 0.00
name                               0.03
host_id                            0.00
host_name                          0.04
neighbourhood_group                0.00
```

DATA METHODOLOGY – VISUALIZATION THRU POWER BI



DATA ASSUMPTIONS



```
Categorical Variables:
      - room_type
      - neighbourhood_group
      - neighbourhood

Continous Variables(Numerical):
      - Price
      - minimum_nights
      - number_of_reviews
      - reviews_per_month
      - calculated_host_listings_count
      - availability_365
- Continous Variables could be binned in to groups too

Location Varibles:
      - latitude
      - longitude

Time Varibale:
      - last_review
```

Variable Categories