airbnb

STORYTELLING CASE STUDY: AIRBNB, NYC

Data insights of Airbnb in NYC

CONTENTS

- OBJECTIVE
- BACKGROUND
- AIRBNB DATA DESCRIPTION
- DATA ASSUMPTIONS - VARIABLES
- PROBLEM STATEMENT OF AIRBNB
- DATA HANDLING
- MISSING VALUE TREATEMENT
- ANALYSIS
- UNIVARIATE , BIVARIATE AND MULTI VARIATE ANALYSIS
- KEY FINDINGS - CHARTS
- APPENDIX - DATA METHODOLOGY
- CONCLUSION (INSIGHTS)

OBJECTIVE

- Airbnb is an online platform using which people can rent their unused accommodations.
- During the covid time, Airbnb incurred a huge loss in revenue.
- People have now started travelling again and Airbnb is aiming to bring up the business again and e ready to provide services to customers.

BACK GROUND

- For the past few months, Airbnb has seen a major decline in revenue.
- Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.
- So, analysis has been done on a dataset consisting of various Airbnb listings in New York.

AIRBNB DATA DESCRIPTION

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 ASSUMPTIONS - VARIABLES

```
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

PROBLEM STATEMENT OF AIRBNB

- For the past few months, Airbnb has seen a major decline in revenue. Now that the restrictions have started lifting and people have started to travel more, Airbnb wants to make sure that it is fully prepared for this change.
- The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset so as to increase the revenue. Our responsibility is to provide valuable insights to aid in decision making.

Importing libraries and reading the dataset

```
# Importing Necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

```
# Read and understand the dataset and check the first five rows
Airbnb_data = pd.read_csv('AB_NYC_2019.csv')
Airbnb_data.head()
```

	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_revie
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	

Creating features

1.1 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'
```

1.2 Categorizing the "minimum_nights" column into 5 categories

```
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'
```

1.3 Categorizing the "number_of_reviews" column into 5 categories

```
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'
```

Fixing columns and data types

```
# Check the datatypes of all the columns of the dataframe after categorizing the columns in data
Airbnb_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   id                              48895 non-null  int64  
 1   name                            48879 non-null  object 
 2   host_id                         48895 non-null  int64  
 3   host_name                       48874 non-null  object 
 4   neighbourhood_group             48895 non-null  object 
 5   neighbourhood                   48895 non-null  object 
 6   latitude                        48895 non-null  float64
 7   longitude                       48895 non-null  float64
 8   room_type                       48895 non-null  object 
 9   price                           48895 non-null  int64  
 10  minimum_nights                  48895 non-null  int64  
 11  number_of_reviews               48895 non-null  int64  
 12  last_review                     38843 non-null  object 
 13  reviews_per_month               38843 non-null  float64
 14  calculated_host_listings_count  48895 non-null  int64  
 15  availability_365                48895 non-null  int64  
 16  availability_365_categories     48895 non-null  object 
 17  minimum_night_categories        48895 non-null  object 
 18  number_of_reviews_categories    48895 non-null  object 
 19  price_categories                48895 non-null  object 
dtypes: float64(3), int64(7), object(10)
```

- reviews_per_month column is of object Dtype. datetime64 is a better Data type for this column.

There are total 48895 rows and 16 columns.

3.1 Categorical

```
# Categorical nominal
categorical_columns = Airbnb_data.columns[[0,1,3,4,5,8,16,17,18,19]]
categorical_columns
```

```
Index(['id', 'name', 'host_name', 'neighbourhood_group', 'neighbourhood',
       'room_type', 'availability_365_categories', 'minimum_night_categories',
       'number_of_reviews_categories', 'price_categories'],
      dtype='object')
```

3.2 Numerical

```
]: numerical_columns = Airbnb_data.columns[[9,10,11,13,14,15]]
   numerical_columns
```

```
]: Index(['price', 'minimum_nights', 'number_of_reviews', 'reviews_per_month',
          'calculated_host_listings_count', 'availability_365'],
         dtype='object')
```

3.3 Coordinates and date ¶

```
coordinates = Airbnb_data.columns[[5,6,12]]
Airbnb_data[coordinates]
```

	neighbourhood	latitude	last_review
0	Kensington	40.64749	2018-10-19
1	Midtown	40.75362	2019-05-21
2	Harlem	40.80902	NaT
3	Clinton Hill	40.68514	2019-05-07
4	East Harlem	40.79851	2018-11-19
...	...	...	...
48890	Bedford-Stuyvesant	40.67853	NaT
48891	Bushwick	40.70184	NaT
48892	Harlem	40.81475	NaT
48893	Hell's Kitchen	40.75751	NaT
48894	Hell's Kitchen	40.76404	NaT

48895 rows × 3 columns

Missing values Treatment

```
# To see the sum of missing values for each column
Airbnb_data.isnull().mean()*100
```

```
id                                  0.000000
name                                0.032723
host_id                             0.000000
host_name                           0.042949
neighbourhood_group                 0.000000
neighbourhood                       0.000000
latitude                            0.000000
longitude                           0.000000
room_type                           0.000000
price                               0.000000
minimum_nights                      0.000000
number_of_reviews                   0.000000
last_review                        20.558339
reviews_per_month                  20.558339
calculated_host_listings_count      0.000000
availability_365                    0.000000
availability_365_categories         0.000000
minimum_night_categories            0.000000
number_of_reviews_categories        0.000000
price_categories                    0.000000
dtype: float64
```

Insights:

- last_review , reviews_per_month columns have around 20.56% missing values
- name and host_name have 0.03% and 0.04 % missing values respectively.

After treating, last_review, host_name, name ,most of the missing values is removed.

```
Airbnb_data1.isnull().mean()*100
```

```
id                                  0.0
name                                0.0
host_id                             0.0
host_name                           0.0
neighbourhood_group                 0.0
neighbourhood                       0.0
latitude                            0.0
longitude                           0.0
room_type                           0.0
price                               0.0
minimum_nights                      0.0
number_of_reviews                   0.0
last_review                         0.0
reviews_per_month                   0.0
calculated_host_listings_count      0.0
availability_365                    0.0
availability_365_categories         0.0
minimum_night_categories            0.0
number_of_reviews_categories        0.0
price_categories                    0.0
dtype: float64
```

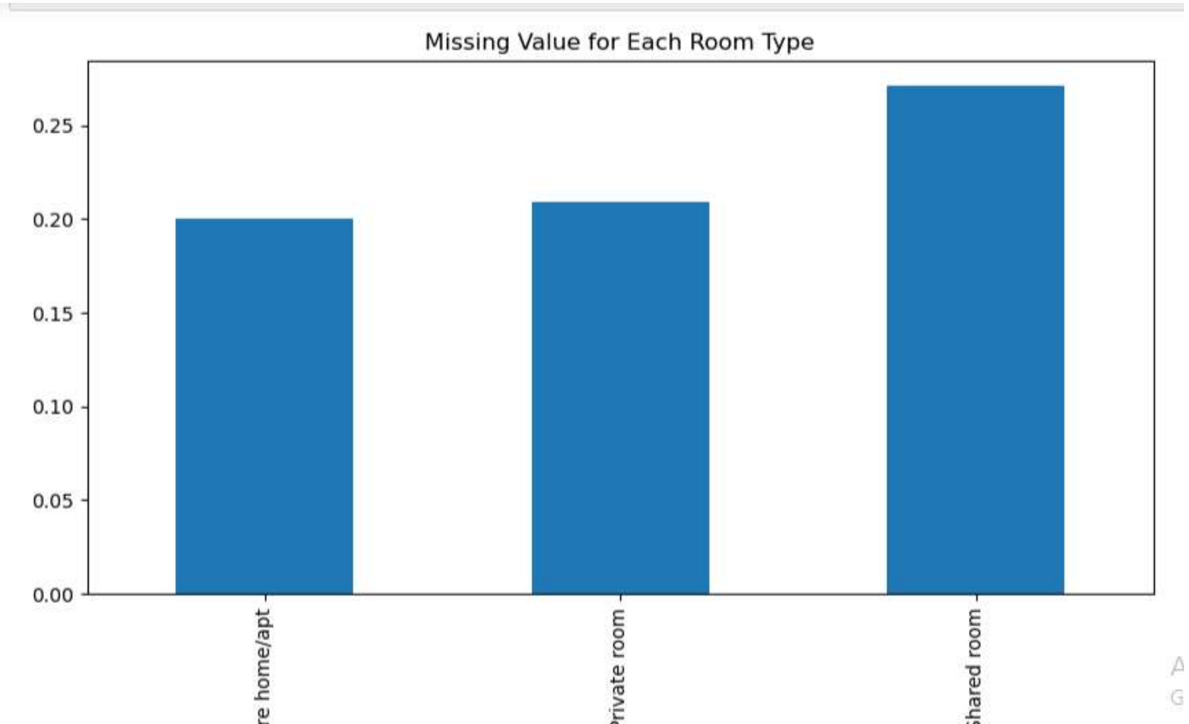
ANALYSING MISSING VALUE

Insights:

The Each neighbourhood_group has 19% missing values in last _review feature.

Insights:

The Each neighbourhood_group has about 22 % missing values in 'last_review' feature.



Insights:

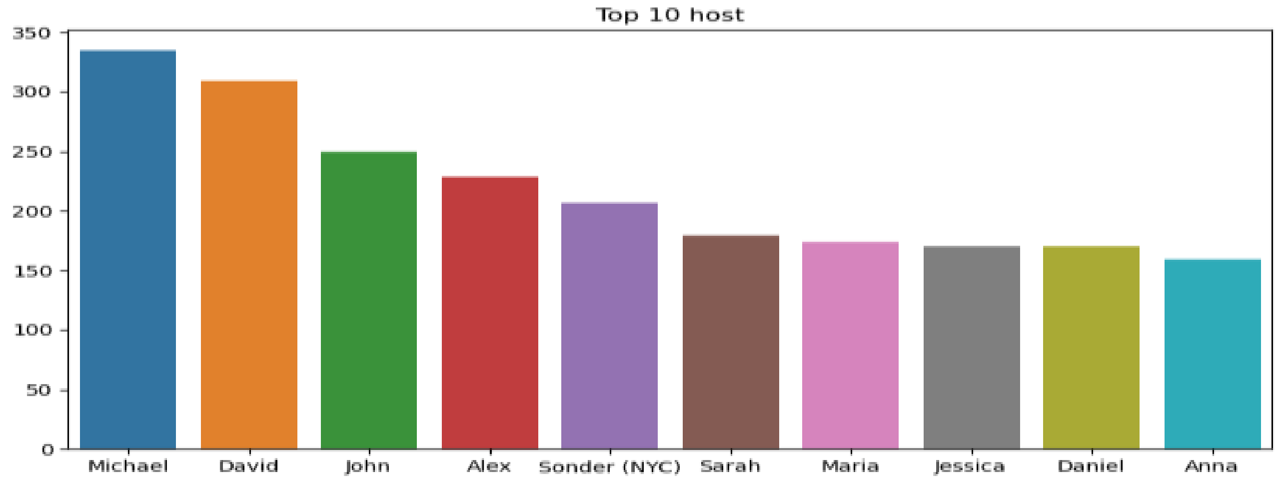
- The pricing is higher when 'last_review' feature is missing .
- reviews are less likely to be given for shared rooms
- When the prices are high reviews are less likely to be given
- The above analysis seems to show that the missing values here are not MCAR (missing completely at random)

UNIVARIATE ANALYSIS

5.3 host_name

```
: Airbnb_data1.host_name.value_counts()

: Michael          335
  David            309
  John             250
  Alex             229
  Sonder (NYC)     207
                  ...
  Krisztián          1
  Kila               1
  Maisha             1
  Martin & Hande     1
  Rusaa              1
```

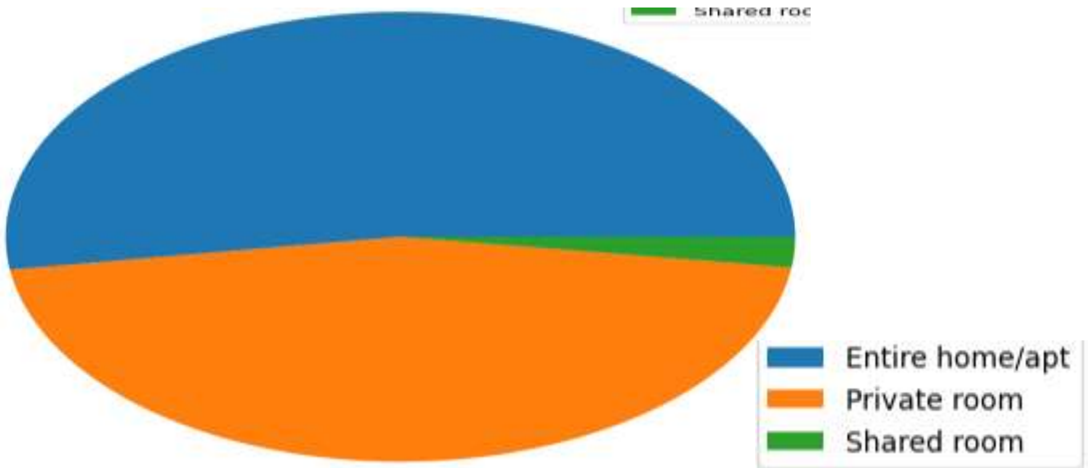


5.7 room_type

```
]: Airbnb_data1.room_type.value_counts(normalize=True)

]: Entire home/apt    0.523454
   Private room       0.454754
   Shared room        0.021792
   Name: room_type, dtype: float64

]: plt.figure(figsize=(8,8))
   plt.title("Room Type")
   plt.pie(x = Airbnb_data1.room_type.value_counts(normalize=True),
   labels = Airbnb_data.room_type.value_counts(normalize= True).index)
   plt.legend()
   plt.show()
```



MOST CONTRIBUTING NEIGHBOURS

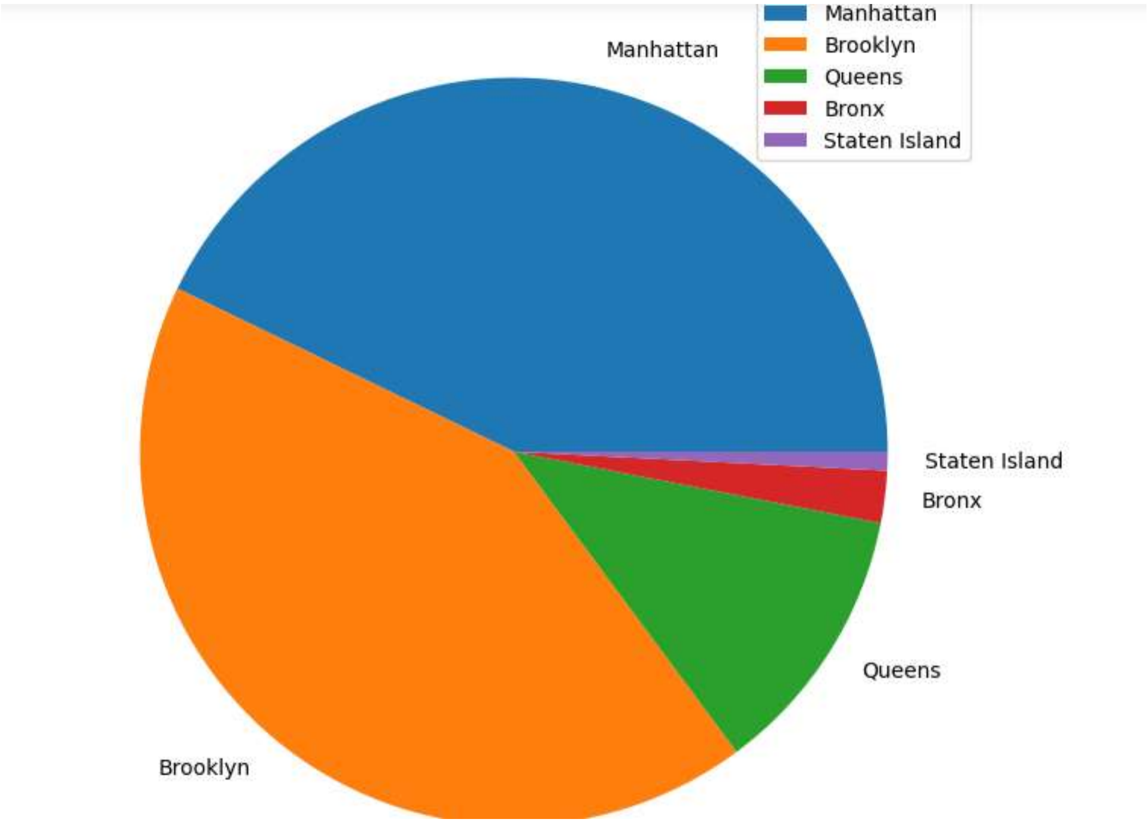
5.4 neighbourhood_group

```
Airbnb_data1.neighbourhood_group.value_counts(normalize=True)*100
```

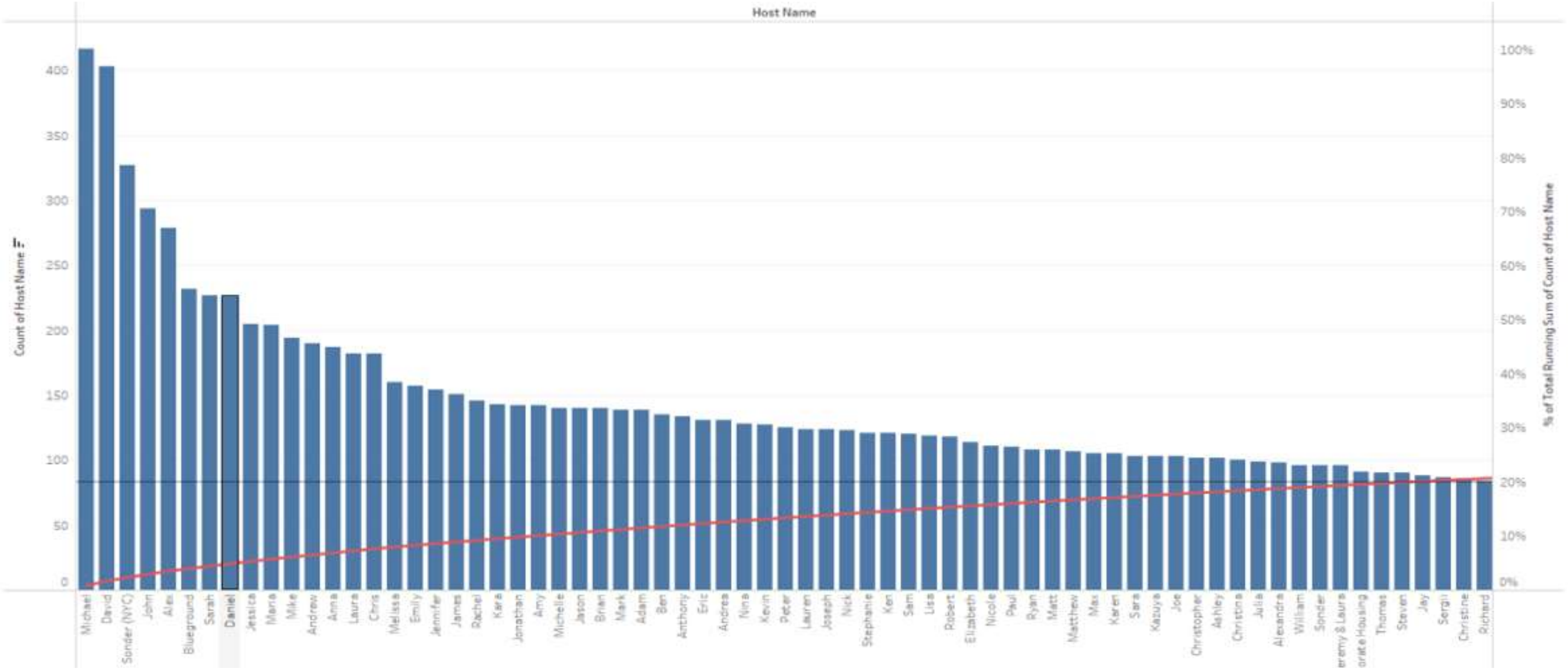
```
Manhattan        42.814456
Brooklyn         42.345638
Queens           11.777131
Bronx             2.253935
Staten Island     0.808841
Name: neighbourhood_group, dtype: float64
```

Insights:

- What are the neighbourhoods they need to target?
- 81 % of the listing are Manhattan and Brooklyn neighbourhood_group



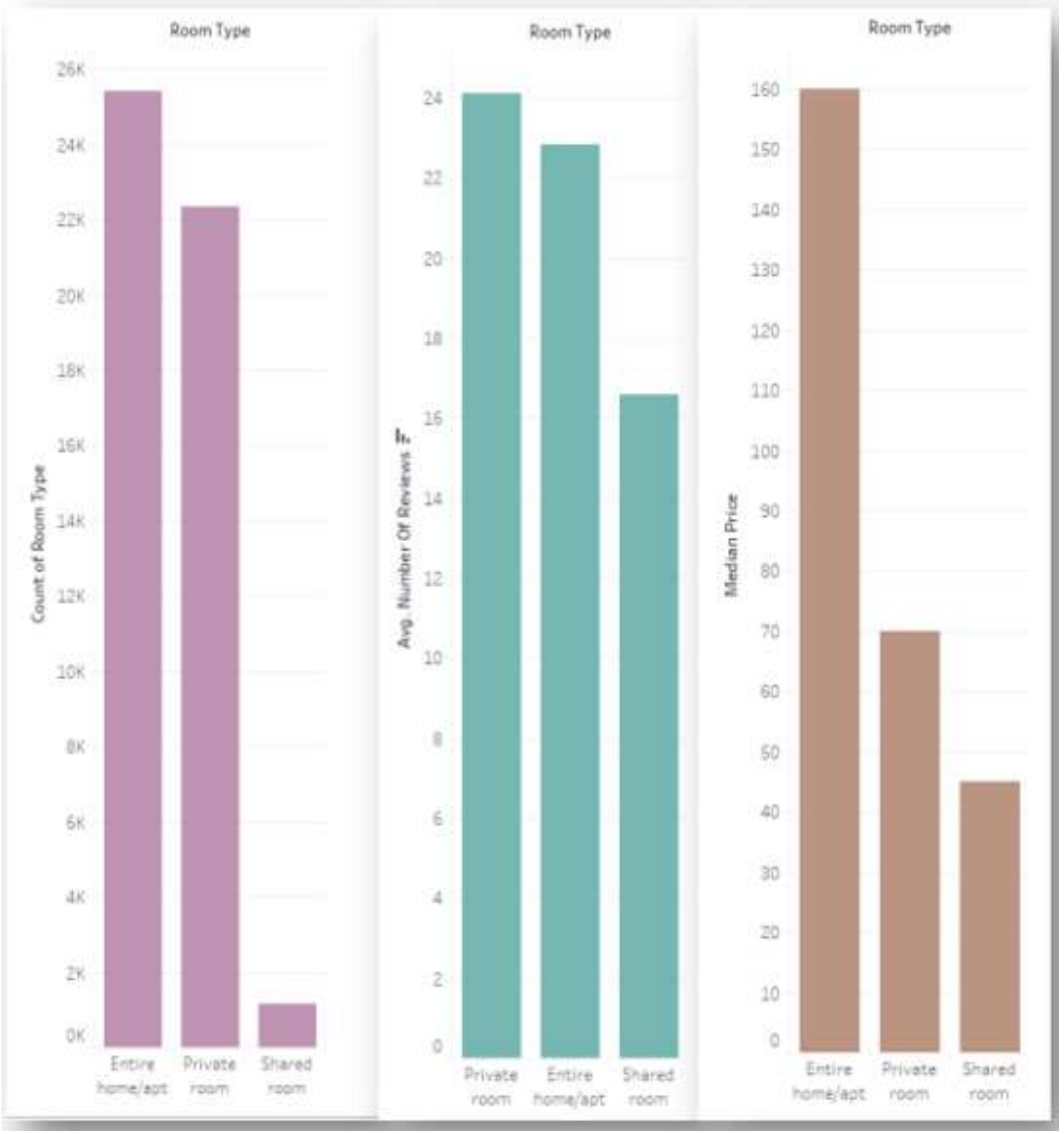
EVERY HOST MATTER



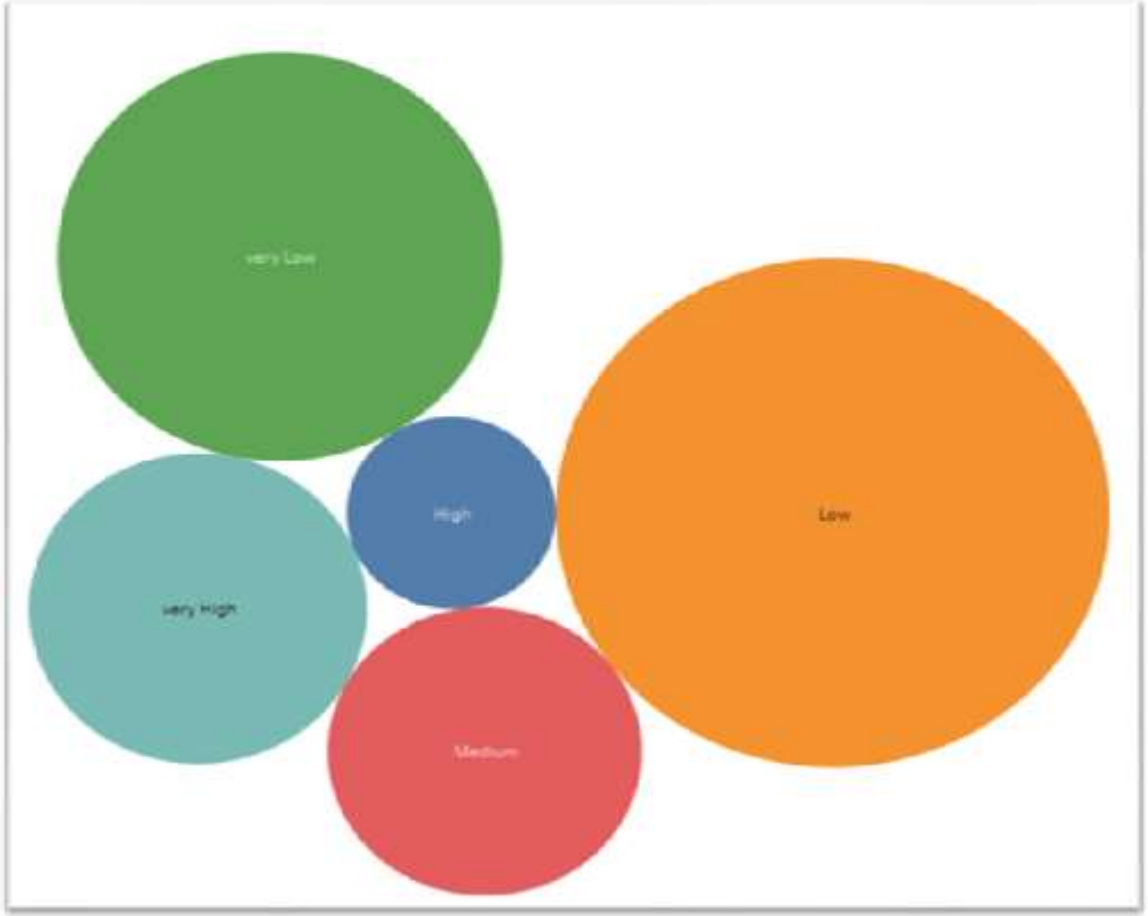
- The top 60 hosts only make up 20% of the total host count!

THE PROBLEMS OF SHARED ROOMS

- Median rates for shared rooms are significantly lower.
- They are less likely to be reviewed.
- Shared rooms only accounts for 2% of the total types of rooms.



MINIMUM NIGHT CATEGORIES

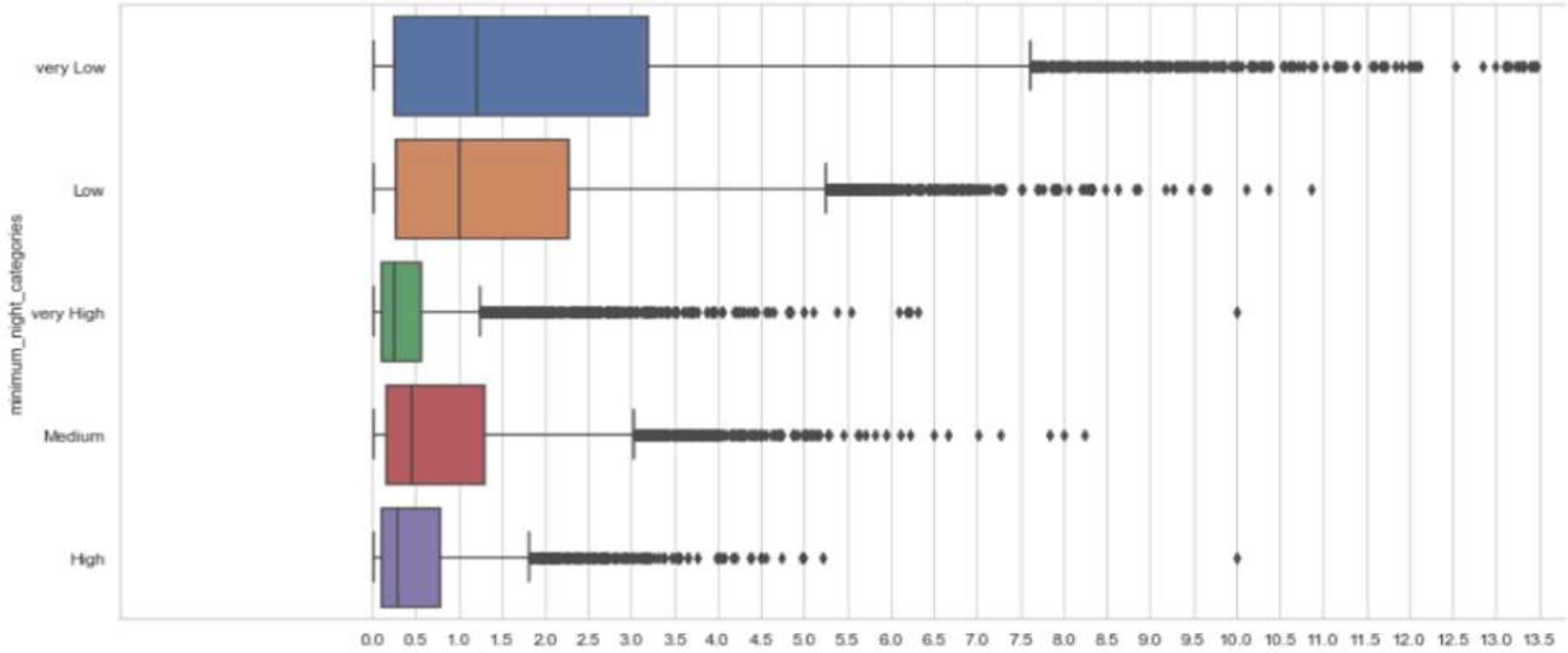


Minimum night category percentages

Low	40.280192
very Low	26.014930
very High	14.997444
Medium	12.960425
High	5.747009

- Low category in minimum night feature contributes 40 %

EFFECT OF MINIMUM NIGHT CATEGORIES



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

BIVARIATE AND MULTIVARIATE ANALYSIS

6.1 Finding the correlations

```
numerical_columns = Airbnb_data1.columns[[9,10,11,13,14,15]]
Airbnb_data1[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
3	89	1	270	4.64	1	194
4	80	10	9	0.10	1	0
5	200	3	74	0.59	1	129

DATA METHODOLOGY

- Conducted a through analysis of New York Airbnbs Dataset.
- Cleaned the data set using python.
- Derived the necessary features.
- Used group aggregration , pivot table and other statistical methods.
- Created charts and visualization using Tableau.

CONCLUSION

- Strong significant insights are deliverd based on various attributes in the dataset.
- Ample amount and variate of visuals have can used in the presentations for the stake-holders.
- Data collection team should collect data about review scores so that it can strengthen the later analysis.


airbnb