# DECODING NYC AIRBNB: UNVEILING HIDDEN INSIGHTS

BY
DEEPAK PALSAVDIYA
BHAVINI BHAVESH HENIYA
DIPALI PAWAR

# AGENDA

*Objective*

*Data life cycle*

*Analysis methods*

*Recommendations*

*appendix:*

- *Data sources*
- *Data methodology*
- *Data model assumptions*

# OBJECTIVE

*To Conduct a thorough analysis of New York Airbnb Dataset.*

*Ask effective questions that can lead to data insights*

*process,analyze and share findings by data visualization and*

*statistical techniques*

## *DATA LIFE CYCLE*

*In the first phase the data captured and loaded into various environment.*

*Once data is cleaned, EDA is done and new features are created.*

*Then Meaningful insights are derived using various analytical methods.*

# 1. Importing libraries and reading the data

```
# import required libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
[ ] # import data set

data=pd.read_csv('AB_NYC_2019.csv')
data.head()
```

| | id | name | host_id | host_name | neighbourhood_group | neighbourhood | latitude | longitude | room_type | price | minimum_nights | n |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | |

# 2. Creating features

**2.1 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'
```

**2.2 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'
```

**2.3 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'
```

**Note:** By categorizing, we are able to better understand relationships and connections between things and better communicate our findings.

# 3. Data types

```
[ ] data.columns
```

```
Index(['id', 'name', 'host_id', 'host_name', 'neighbourhood_group',
       'neighbourhood', 'latitude', 'longitude', 'room_type', 'price',
       'minimum_nights', 'number_of_reviews', 'last_review',
       'reviews_per_month', 'calculated_host_listings_count',
       'availability_365', 'availability_categories',
       'minimum_nights_categories', 'number_of_reviews_categories',
       'price_categories'],
      dtype='object')
```

```
[ ] categoorical_columns= data.columns[[0,1,3,4,5,8,16,17,18,19]]
    categoorical_columns
```

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

```
[ ] # 2- Numeric Columns
```

```
numerical_columns=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 Coordinate and Date
```

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

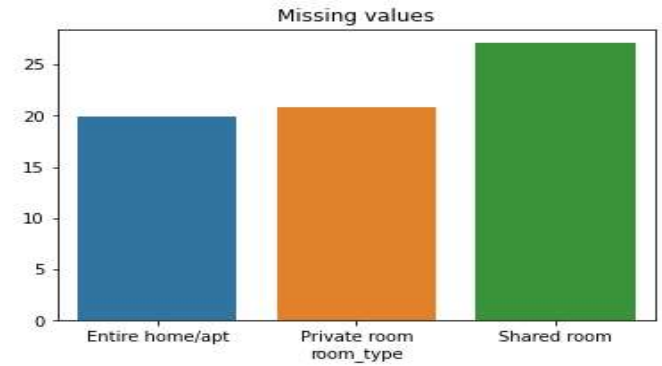
```
Index(['neighbourhood', 'latitude', 'last_review'], dtype='object')
```

# 5. Missing values

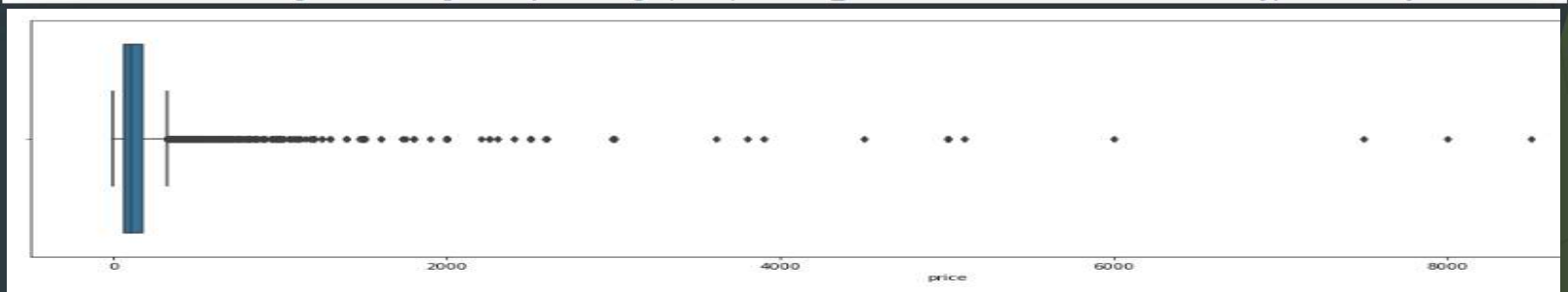
```
[ ]  #Checking null values once again

     (data.isnull().sum()/len(data)*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
dtype: float64
```

# 5.1 Missing value analysis




**'Shared room' has the highest missing value percentage (27 %) for 'last_review' feature while to other room types has only about 20 %.**




- 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)

# 6. Analysis

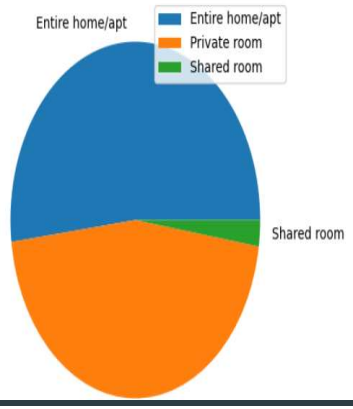
```
# room type

data.room_type.value_counts()
```

```
Entire home/apt    25409
Private room       22326
Shared room         1160
Name: room_type, dtype: int64
```

```
[ ] data.room_type.value_counts(normalize =True)*100
```

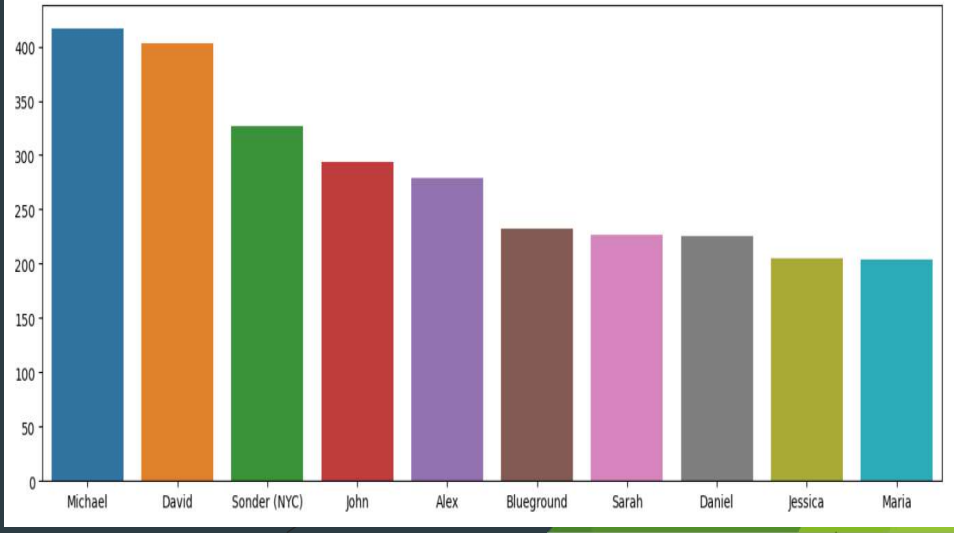
```
Entire home/apt    51.966459
Private room       45.661111
Shared room         2.372431
Name: room_type, dtype: float64
```

```
[ ] plt.figure(figsize=(5,5))
    plt.pie(x=data.room_type.value_counts(normalize =True)*100, labels =data.room_type.value_counts(normalize =True).index)
    plt.legend()
    plt.show()
```




```
[ ] #Host_name
    data.host_name.value_counts()
```

```
Michael                417
David                  403
Sonder (NYC)           327
John                   294
Alex                   279
                      ...
Rhonycs                  1
Brandy-Courtney          1
Shanthony                1
Aurore And Jamila        1
Ilgar & Aysel            1
Name: host_name, Length: 11452, dtype: int64
```

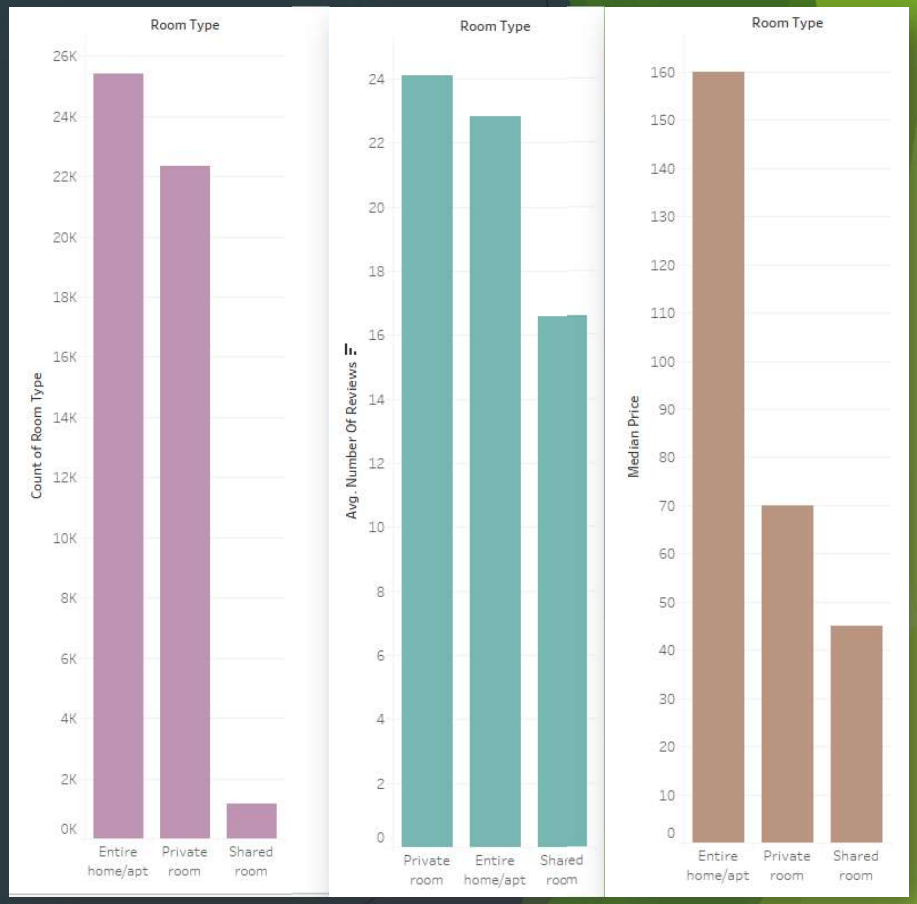


# THE PROBLEMS WITH SHARED ROOMS

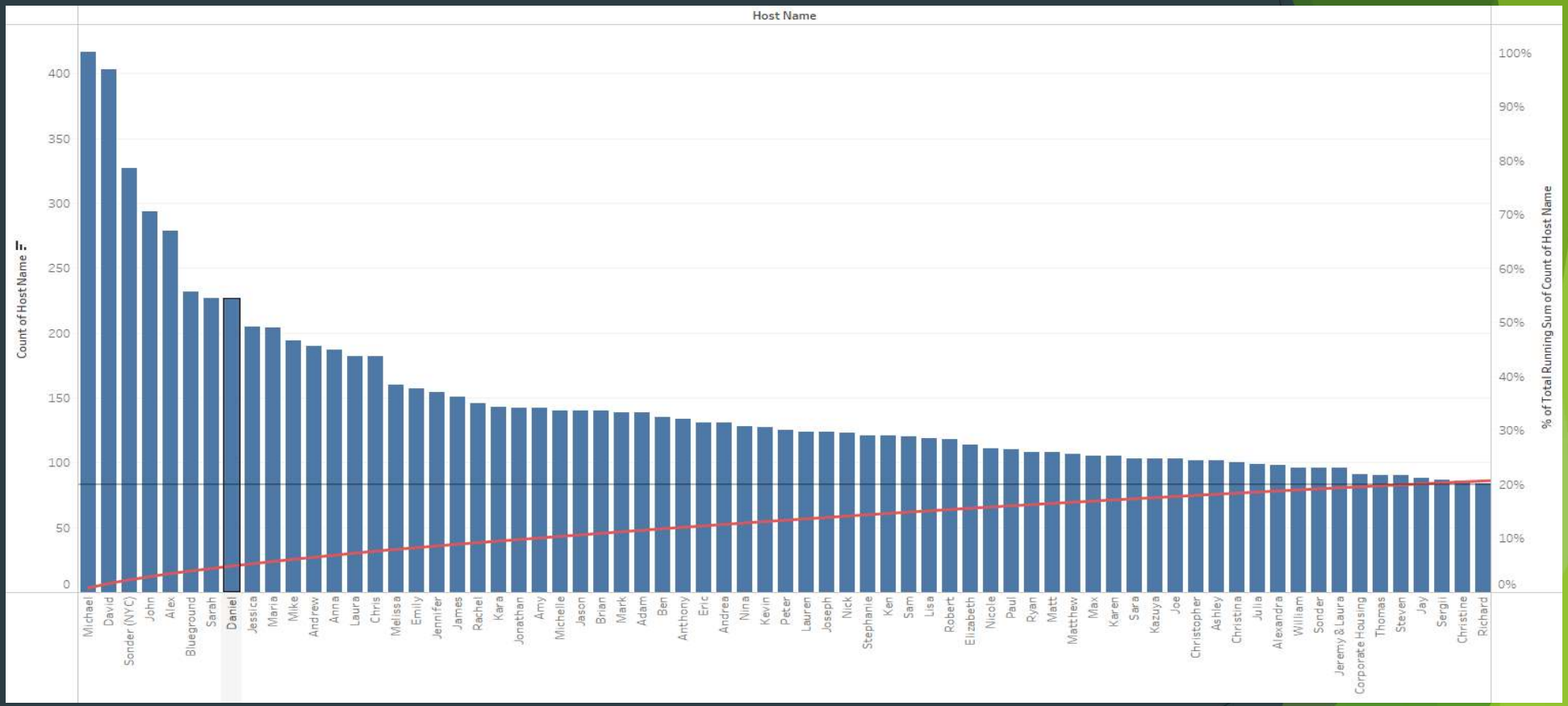
*Shared rooms only account for 2 % of the total types of rooms.*

*They are less likely to be reviewed.*

*Median rates for shared rooms are significantly lower.*

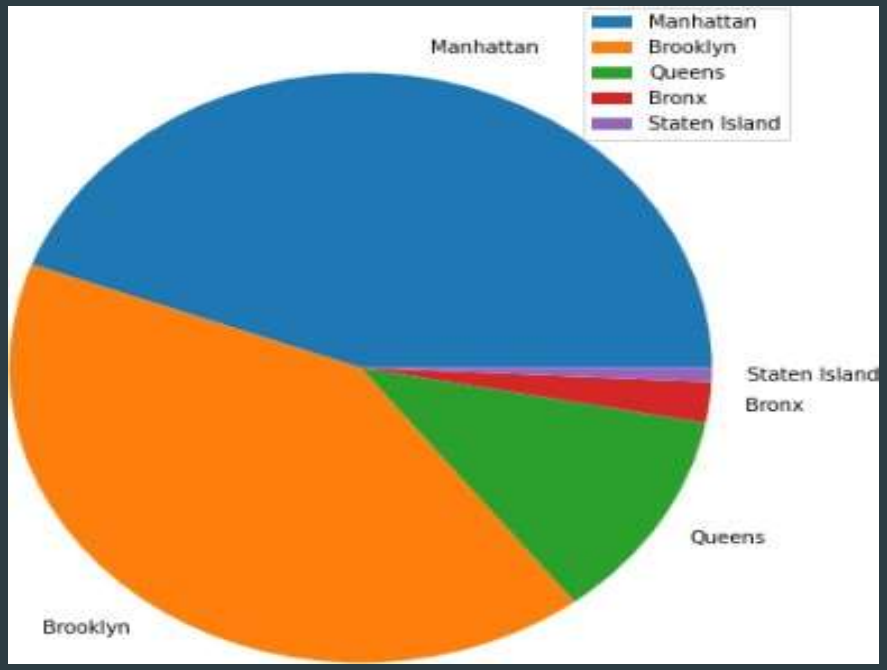


# EVERY HOST MATTER




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

# MOST CONTRIBUTING NEIGHBORHOODS

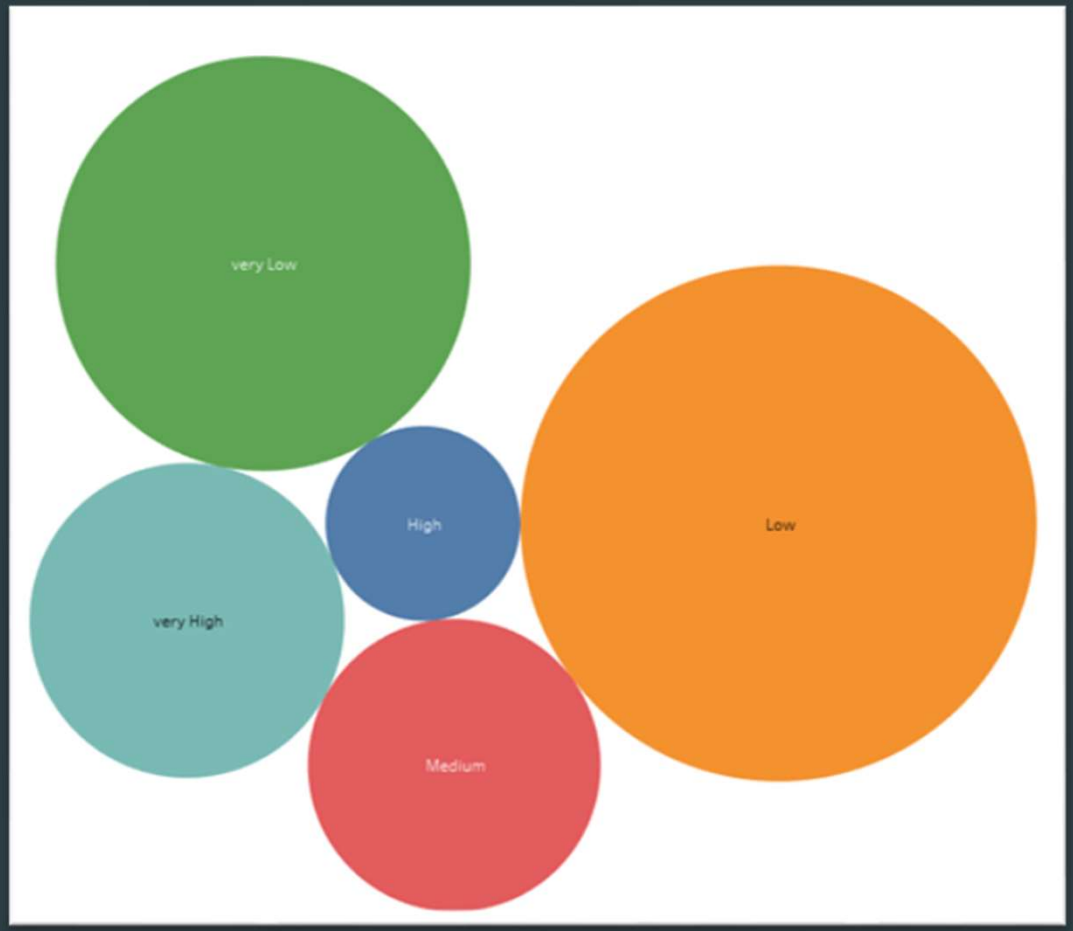



| | |
|---|---|
| Manhattan | 44.301053 |
| Brooklyn | 41.116679 |
| Queens | 11.588097 |
| Bronx | 2.231312 |
| Staten Island | 0.762859 |

*81 % of the listing are Manhattan and Brooklyn neighborhood group*

*Staten Island has the lowest contribution.*

# MINIMUM NIGHT CATEGORIES

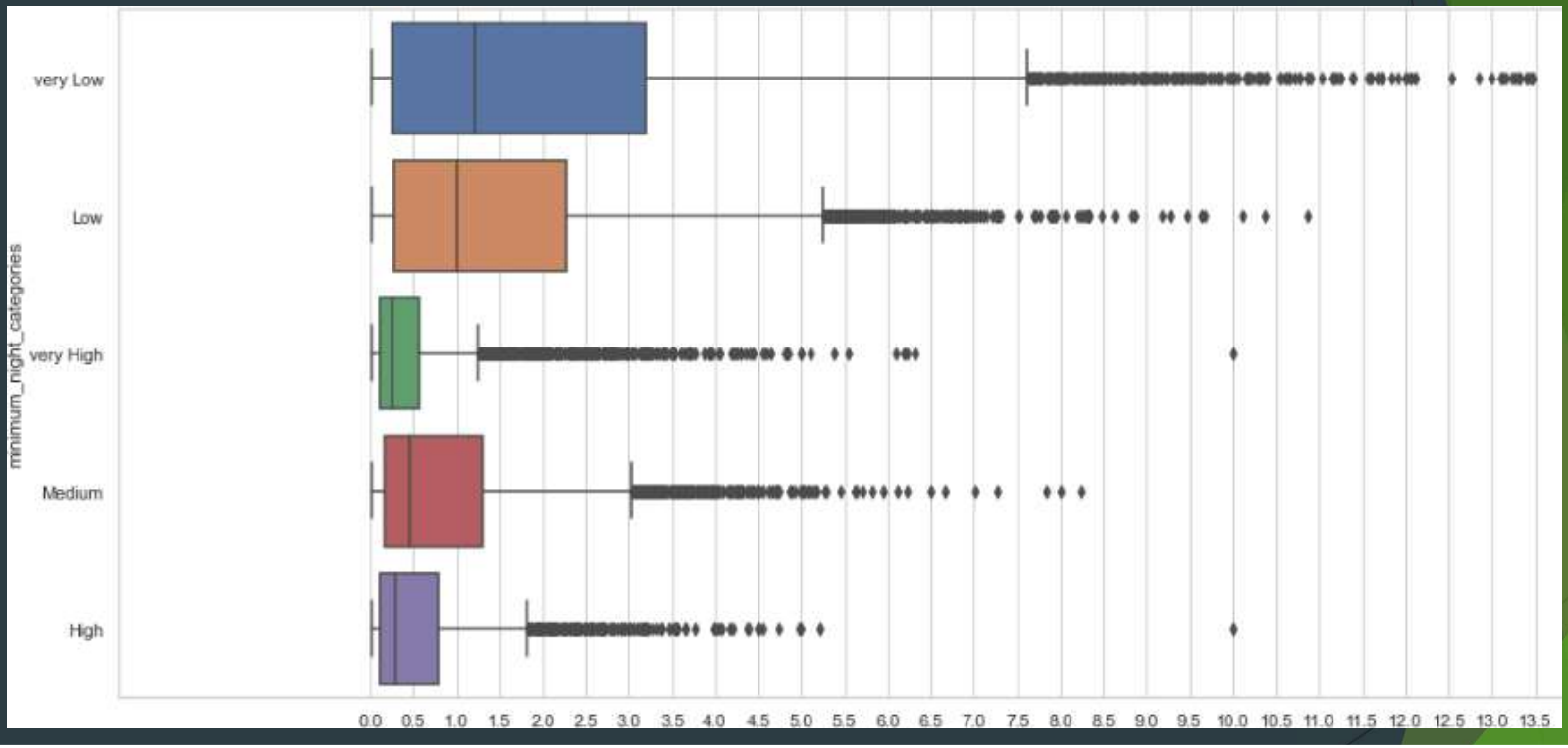



*Minimum night category percentages*

| | |
|---|---|
| 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 ON REVIEWS



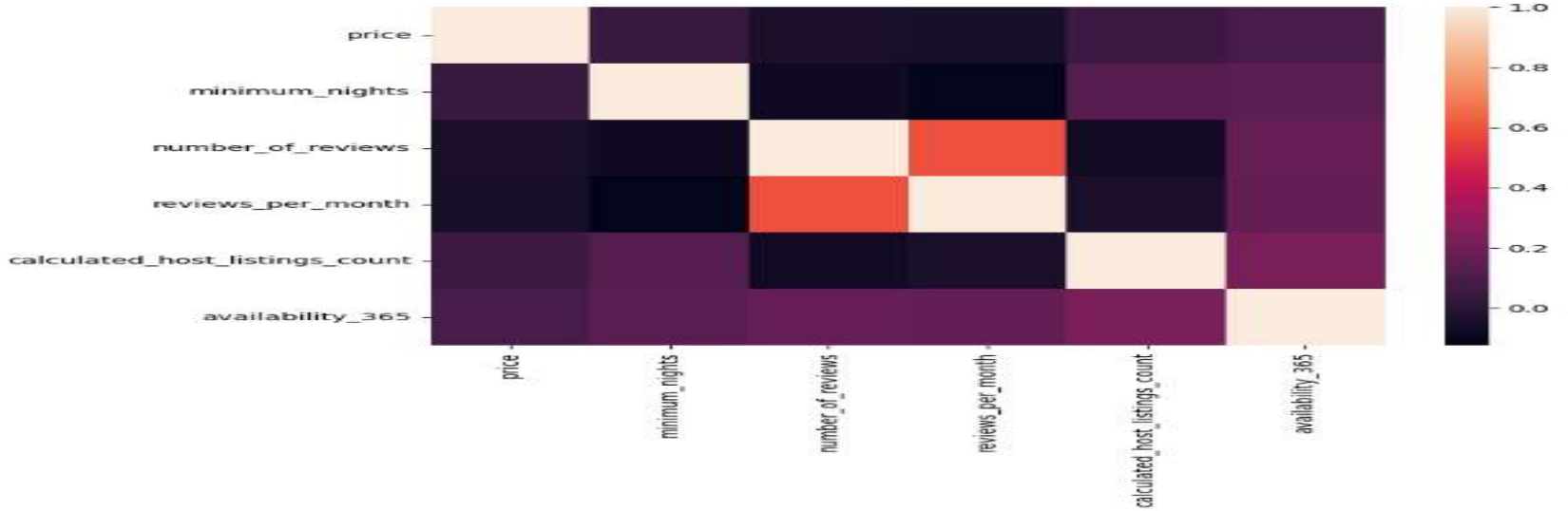

*Customers are more likely to leave reviews for lower number of minimum nights.*

# 7. Bivariate and Multivariate Analysis

```
[ ] data[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 | -0.050564 | 0.057472 | 0.081829 |
| **minimum_nights** | 0.042799 | 1.000000 | -0.080116 | -0.124905 | 0.127960 | 0.144303 |
| **number_of_reviews** | -0.047954 | -0.080116 | 1.000000 | 0.589407 | -0.072376 | 0.172028 |
| **reviews_per_month** | -0.050564 | -0.124905 | 0.589407 | 1.000000 | -0.047312 | 0.163732 |
| **calculated_host_listings_count** | 0.057472 | 0.127960 | -0.072376 | -0.047312 | 1.000000 | 0.225701 |
| **availability_365** | 0.081829 | 0.144303 | 0.172028 | 0.163732 | 0.225701 | 1.000000 |

```
[ ] plt.figure(figsize=(8,6))
    sns.heatmap(data=data[numerical_columns].corr())
    plt.show()
```



# Conclusion

Strong significant insights are derived based on various attributes in the dataset.

Ample amount and variety 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.

A clustering machine learning model toidentify groups of similar objects in datasets with two or more variable quantities can be made.

# *APPENDIX - 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.***

| 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 |

# APPENDIX –DATA METHODOLOGY

- *Conducted a thorough analysis of NewYork Airbnbs Dataset.*
- *Cleaned the data set using python.*
- *Derived the necessary features.*
- *Used group aggregation, pivot table and other statistical methods.*
- *Created charts and visualizations using Tableau.*

# APPENDIX - 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
```


Thank you