DATA INSIGHTS OF AIRBNB IN NYC

BACKGROUND

Airbnb is an online platform using which people can rent their unused accommodations.

During the covid time, Airbnb incurred a huge loss in revenue.

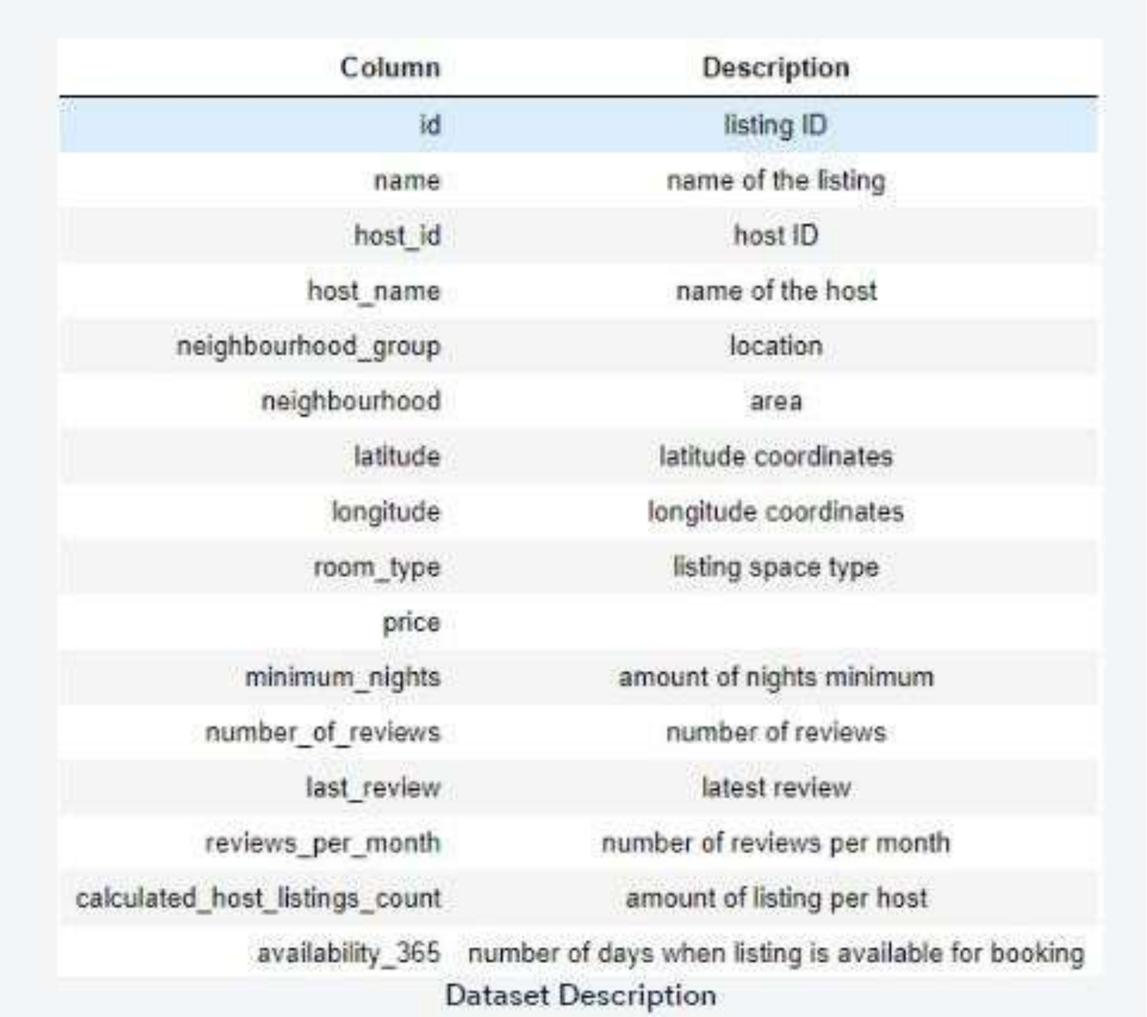
People have now started travelling again and Airbnb is aiming to bring up the business again and ready to provide services to customers.

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.

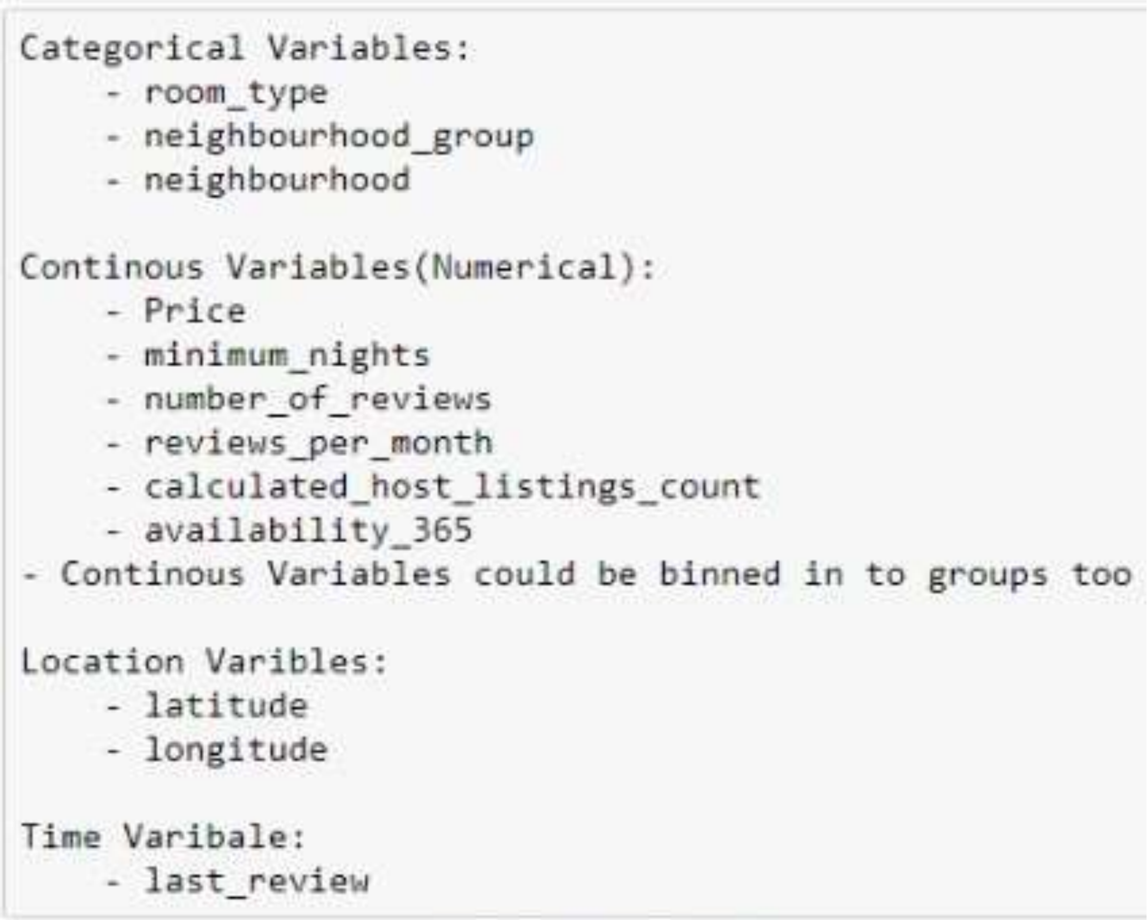
AIRBNB DATA DESCRIPTION



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

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.

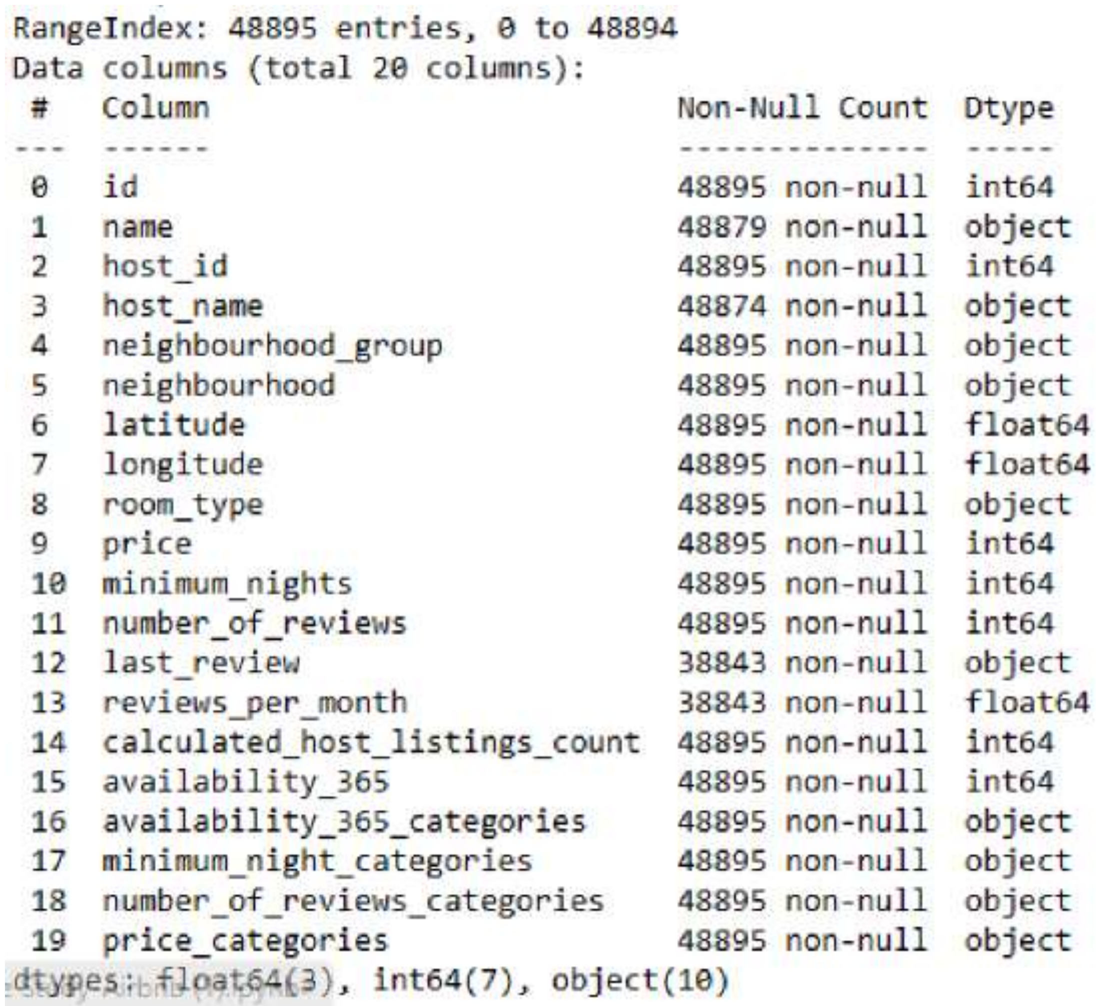
The different leaders at Airbnb want to understand some important insights based on various attributes in the dataset so as to increase the revenue. Our responsibility is to provide valuable insights to aid in decision making.

PROBLEM STATEMENT

DATASET

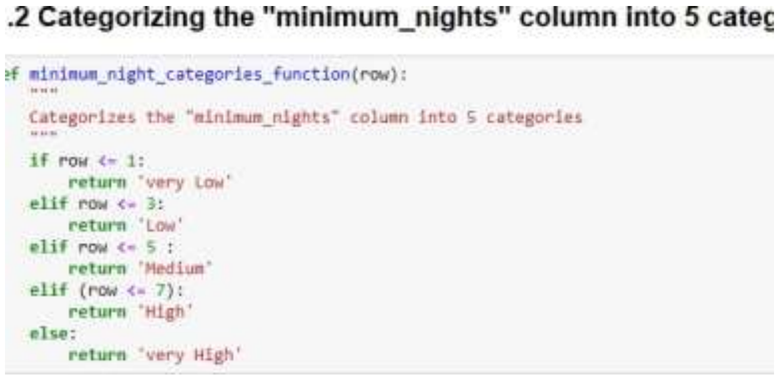
There are total 48895 rows and 16 columns.

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



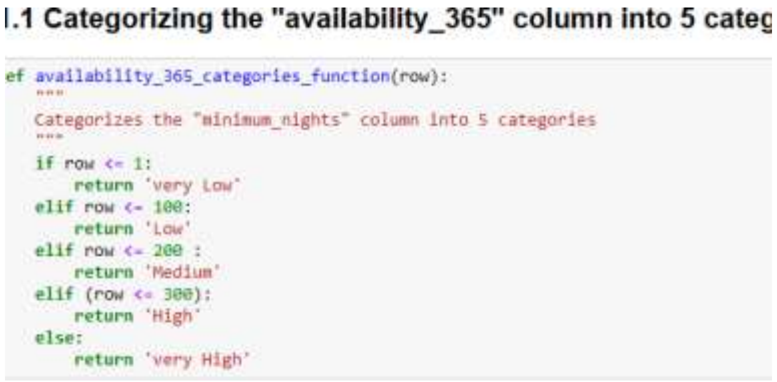
```
RangeIndex: 48895 entries, 0 to 48894
Data columns (total 20 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              48895 non-null  int64
 1   name                            48879 non-null  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)
```

CREATING FEATURES



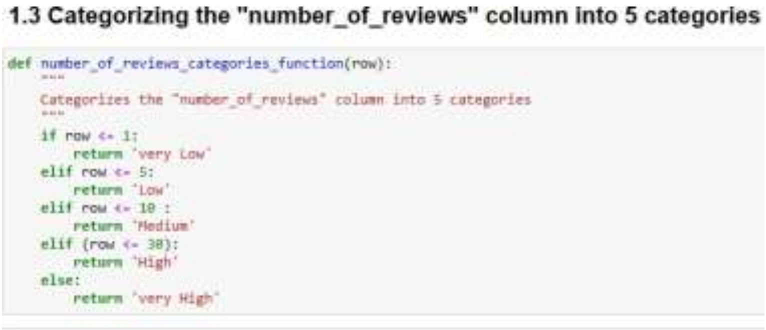
.2 Categorizing the "minimum_nights" column into 5 categ

```
ef 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'
```



l.1 Categorizing the "availability_365" column into 5 categ

```
ef 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'
```

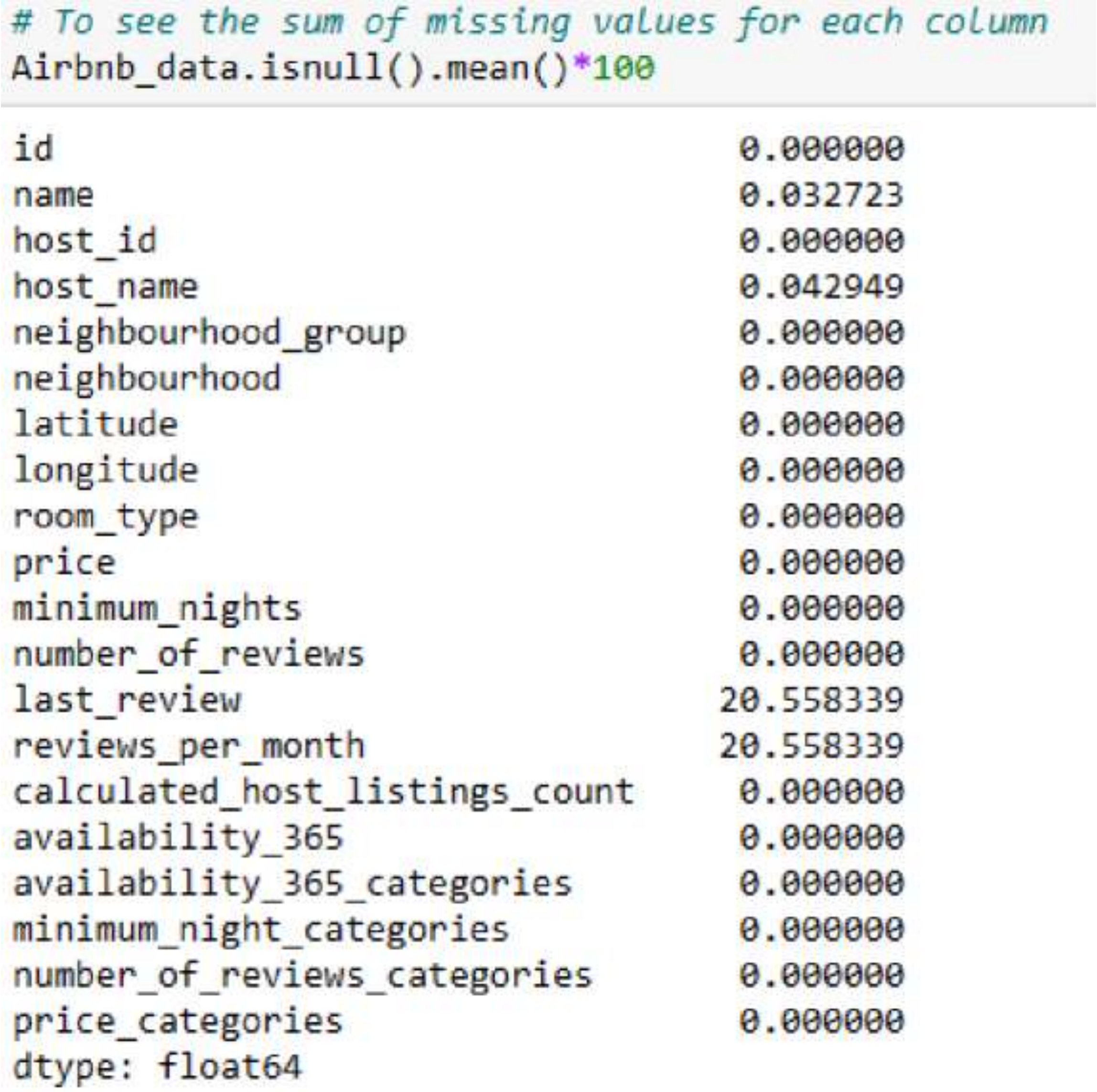


1.3 Categorizing the "number_of_reviews" column into 5 categories

```
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'
```

MISSING VALUES ANALYSIS

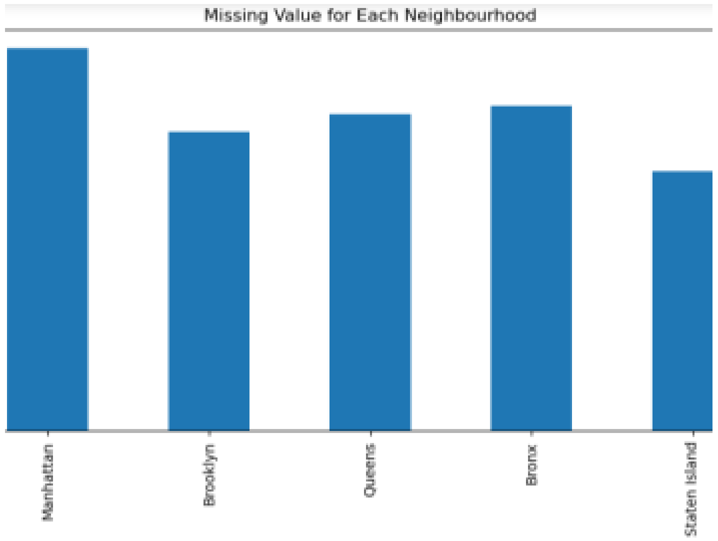
last_review , reviews_per_month columns have around 20.56% missing values
name and host_name have 0.03% and 0.04 % missing values respectively.

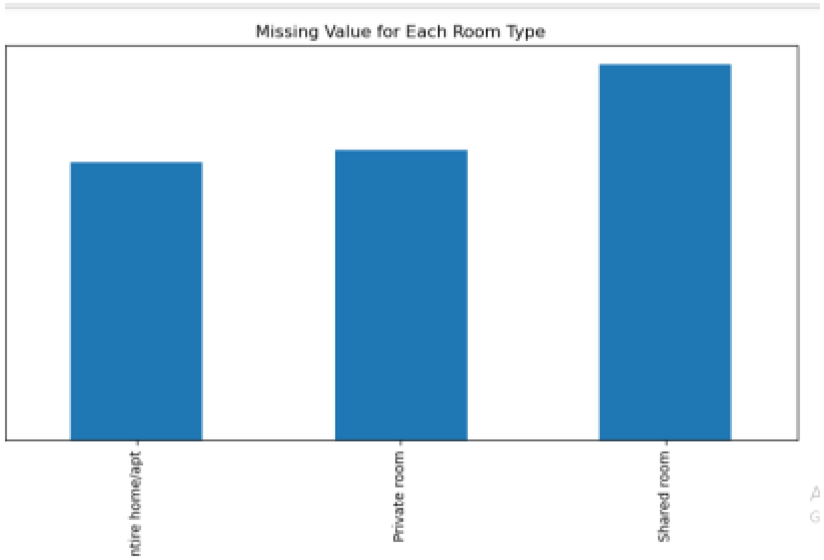


```python
# To see the sum of missing values for each column
Airbnb_data.isnull().mean()*100
```

```
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
```

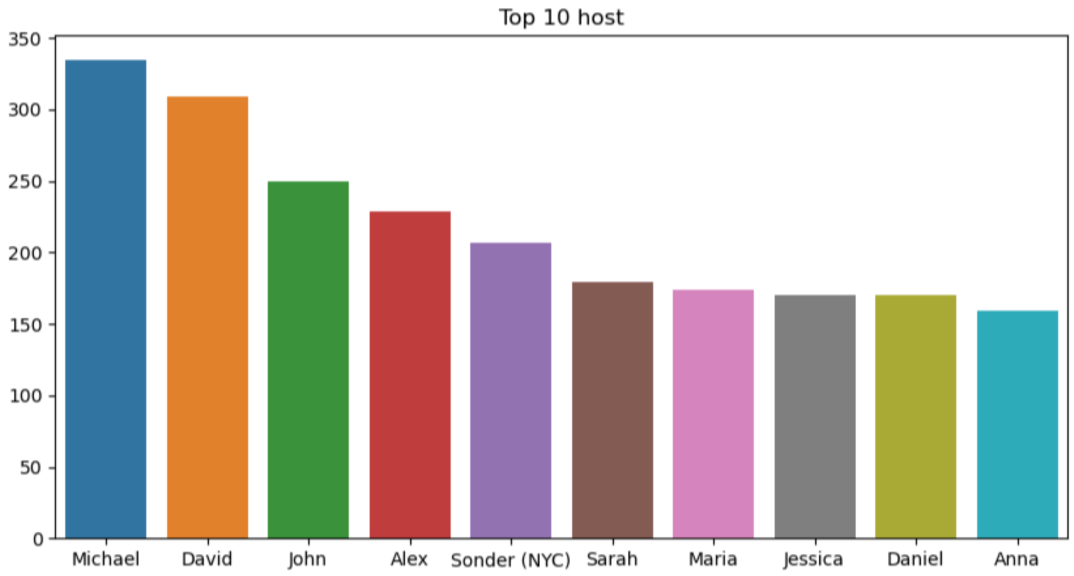
ANALYZING MISSING VALUE

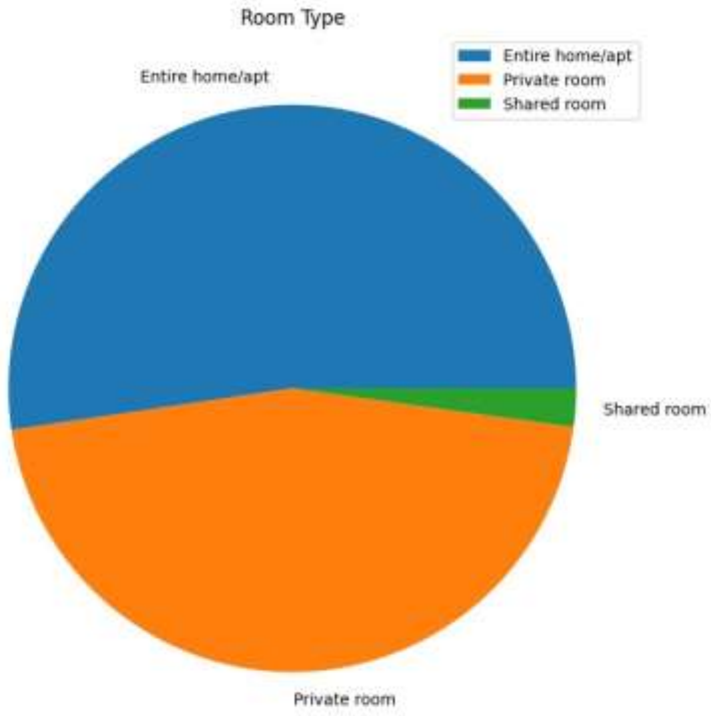




The Each neighborhood group has 19% missing values in 'last _review' feature.

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





UNIVARIATE ANALYSIS

LAST REVIEW FEATURE

The pricing is higher when 'last_review' feature is missing .

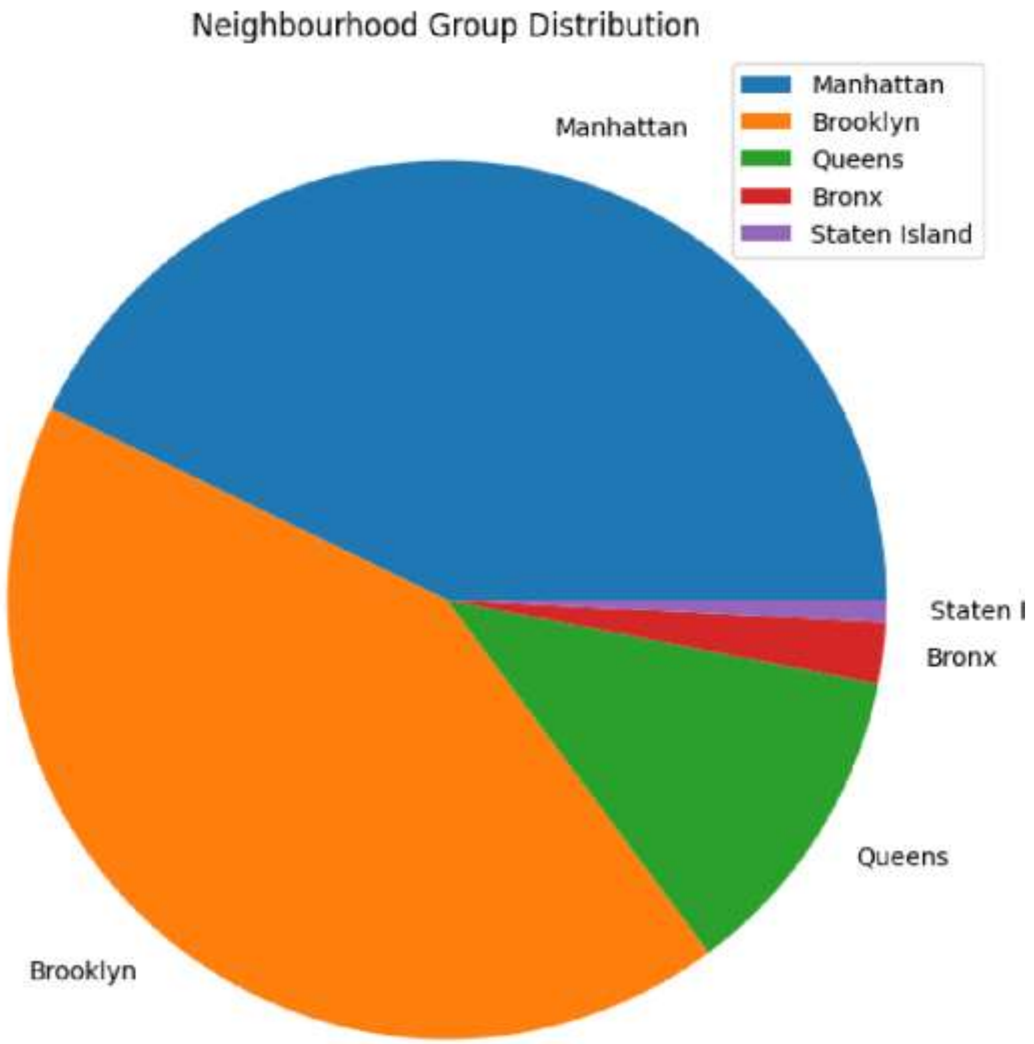
reviews are less likely to be given for shared rooms

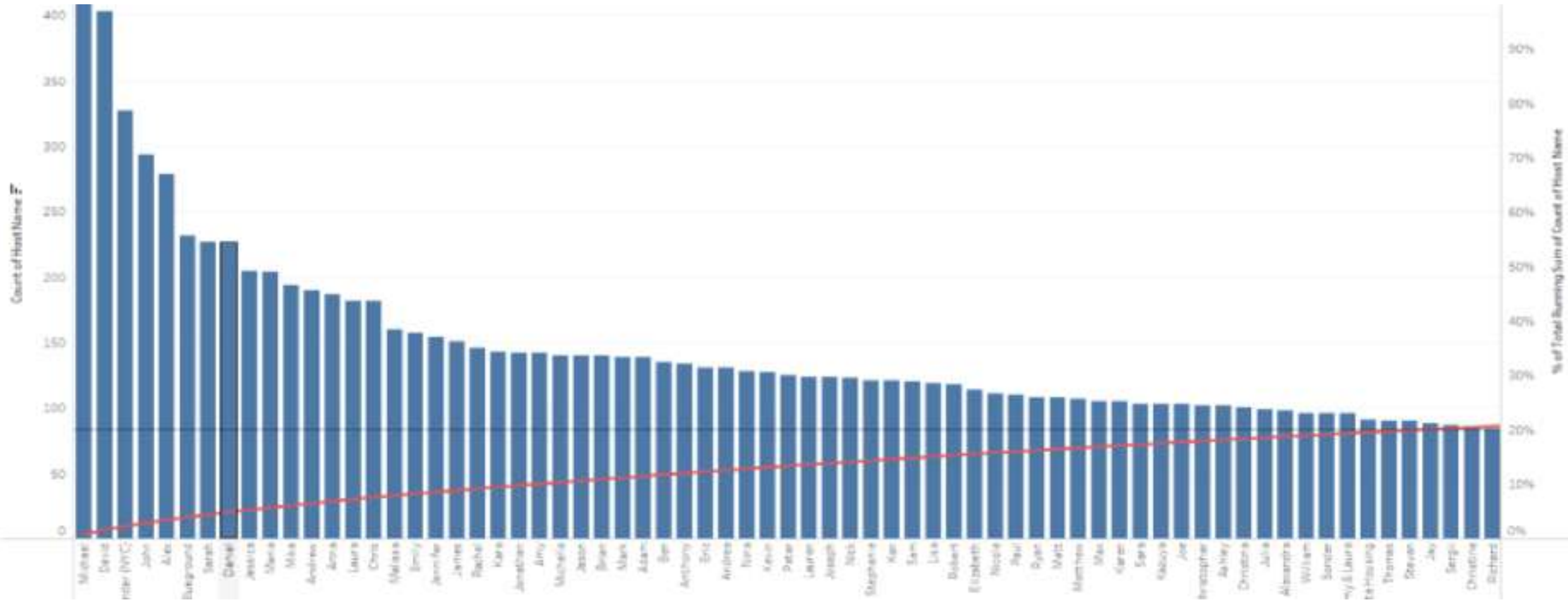
When the prices are high reviews are less likely to be given

The above analysis seems to show that the missing values here are not MCAR (missing completely at random)

MOST CONTRIBUTING NEIGHBOURS

What are the neighborhoods needed to target?
81% of the listing are Manhattan and Brooklyn neighborhood group



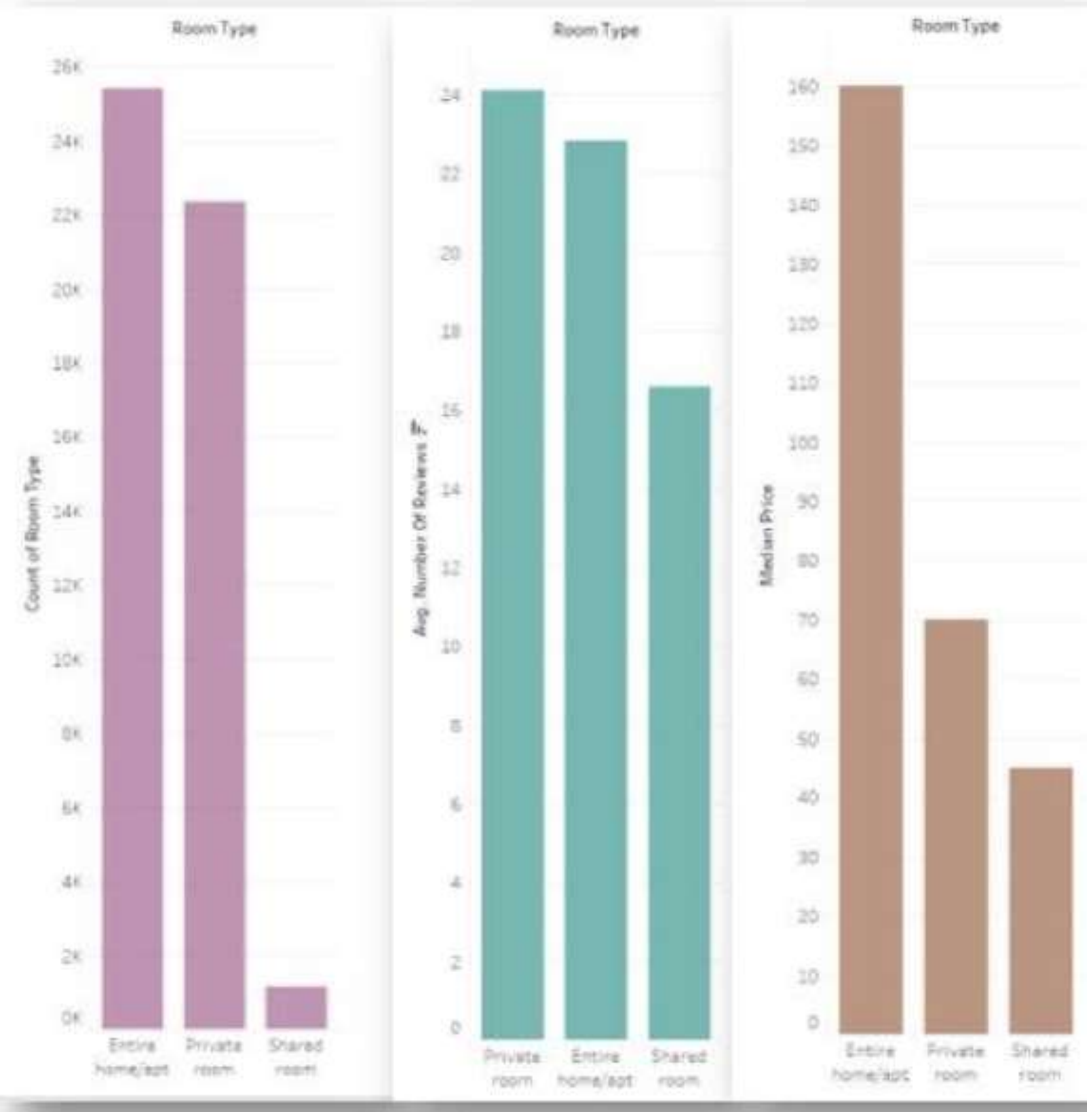


EVERY HOST MATTER

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

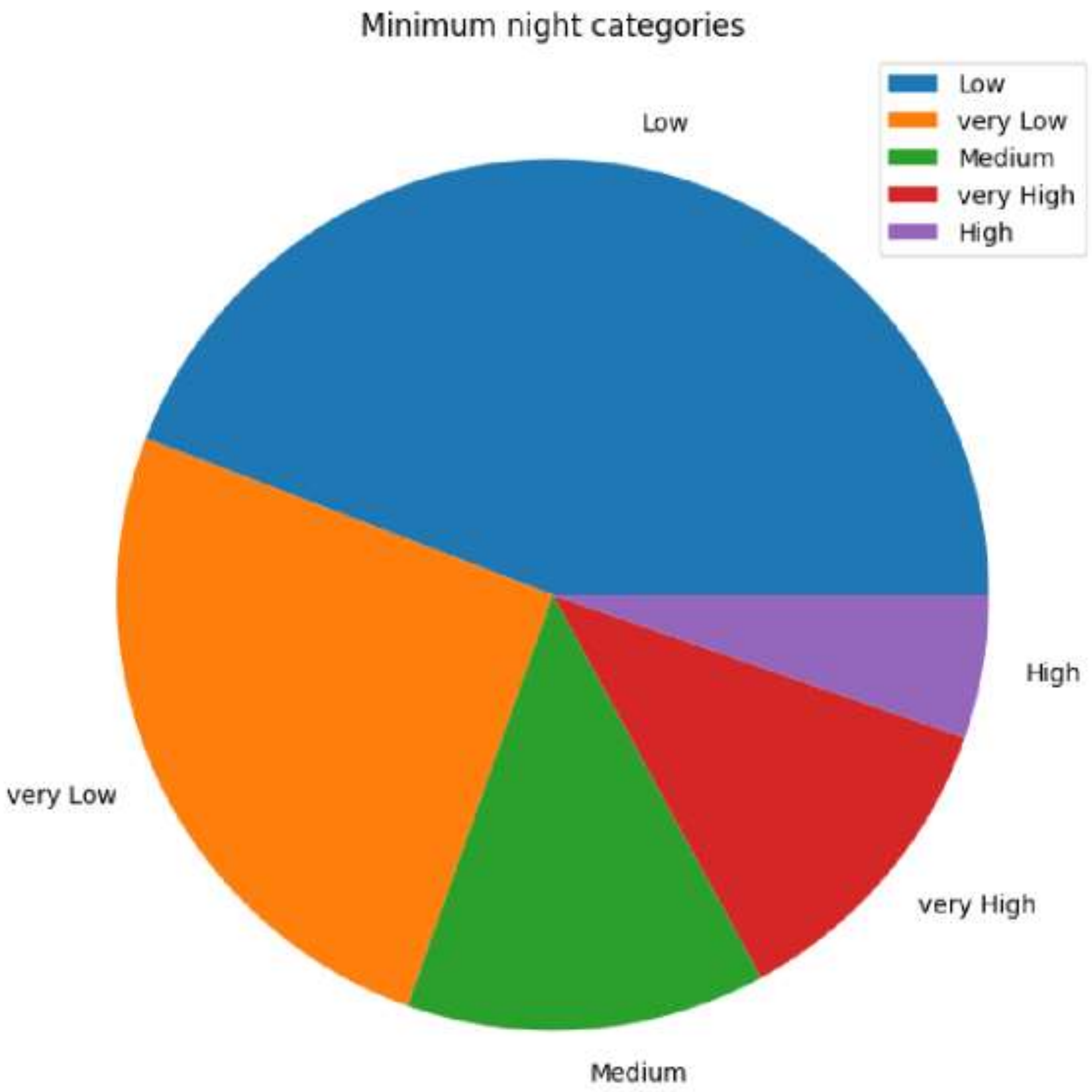
THE PROBLEMS OF SHARED ROOMS

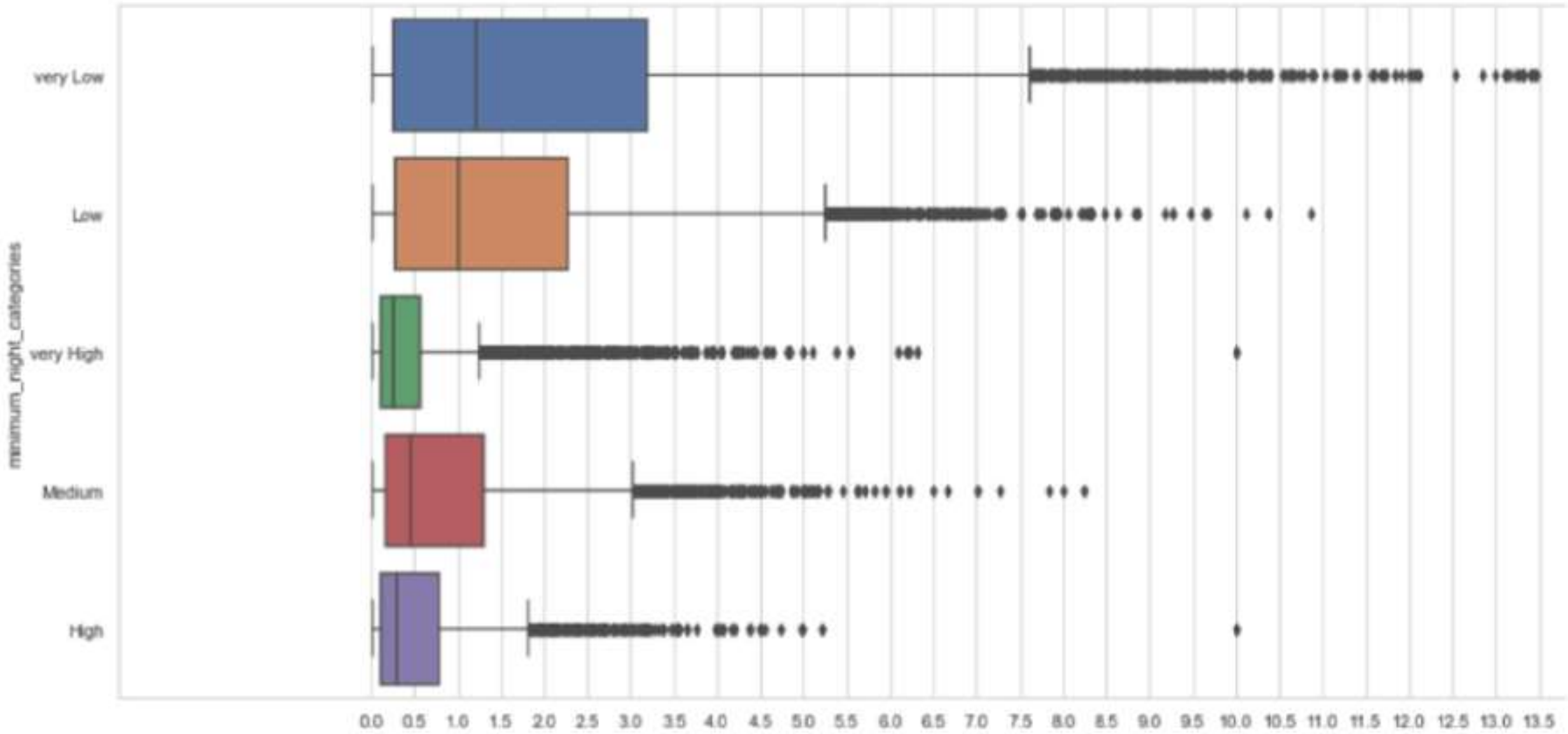
Median rates for shared rooms are significantly lower.
They are less likely to be reviewed.
Shared rooms only accounts for 2% of the total types of rooms.



MINIMUM NIGHT CATEGORIES

Low Category in minimum nights feature contribute 40%.

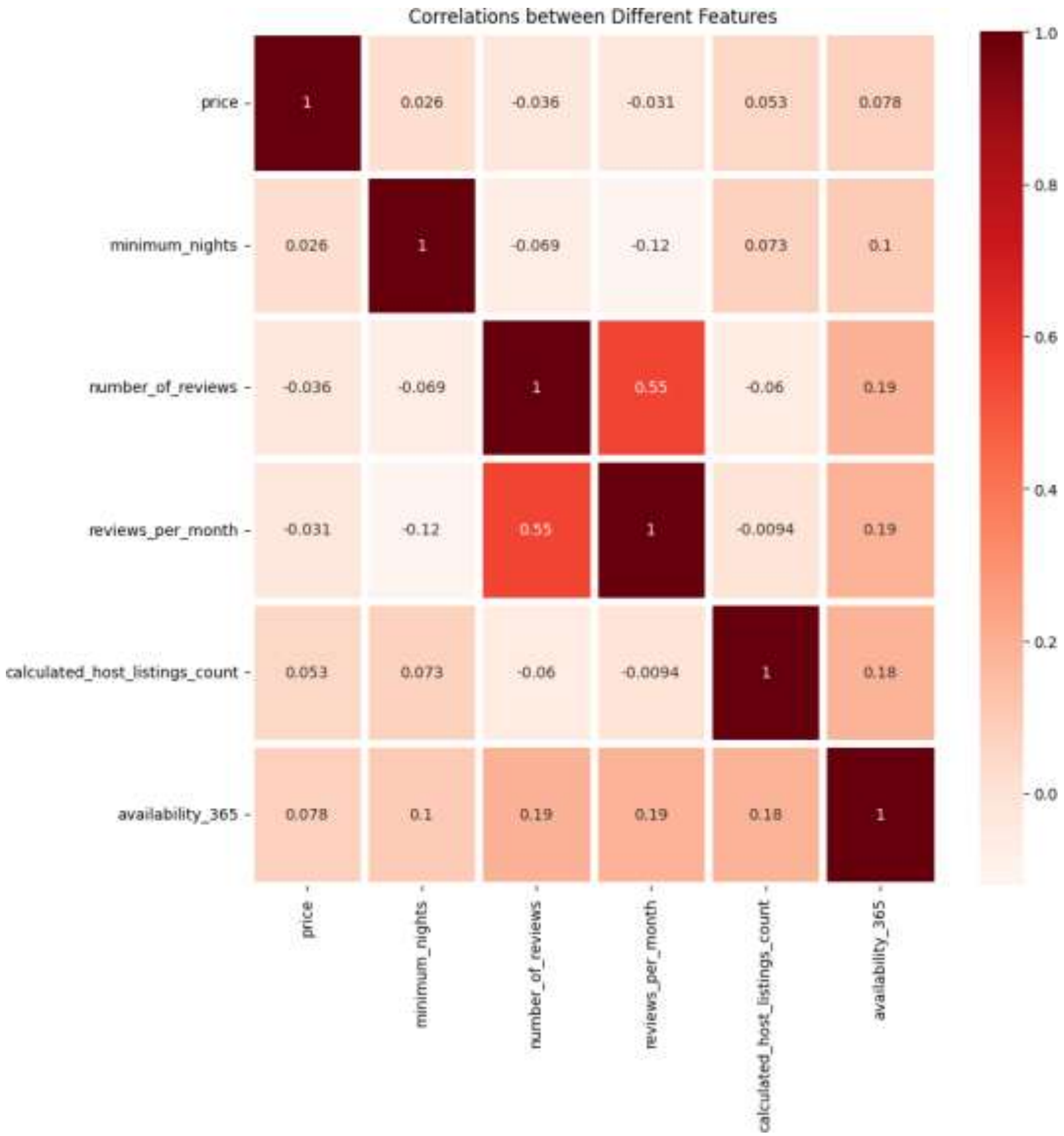




EFFECT OF MINIMUM NIGHT CATEGORIES

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

BIVARIATE AND MULTIVARIATE ANALYSIS



DATA METHODOLOGY

Conducted a thorough analysis of New York Airbnb's Dataset.

Cleaned the data set using python.

Derived the necessary features.

Used group aggregation , pivot table and other statistical methods.

Created charts and visualization using Tableau.

CONCLUSION

Strong significant insights are delivered 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.

THANK YOU