DATA MINING

AIRBNB NEW YORK 2019

http://insideairbnb.com/new-york-city/

Introduction

Airbnb is an online marketplace that links individuals who are searching for housing in any region to rent their properties. Presently, it encompasses more than 81 thousand cities worldwide and 191 countries.

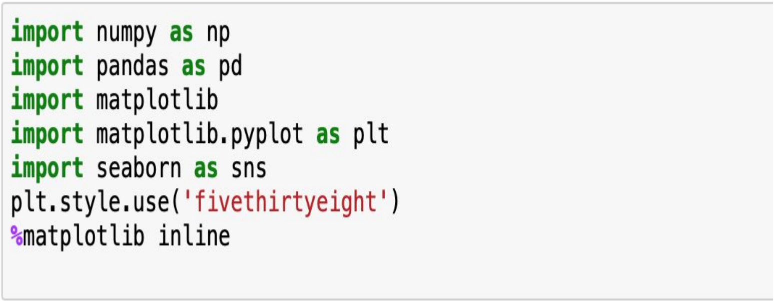
Since 2008 visitors and tourists utilize Airbnb to broaden possibilities for traveling and create an unusual personal experience. The numbers of listing and renting are listed in this data set for NYC, 2019. This free database is accessible on this page as part of Airbnb.

MOTIVATION

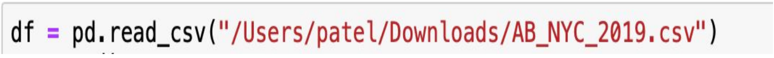
New York City (NYC) has a large Airbnb market of over 48,000 listings for the calendar year from August 2019 (correlating with a rents score of 48,000 per 468 square miles of rent equal to up to 102 leases per square miles). This research concentrates on trends and other pertinent information on Airbnb in NYC.

We will find out how rentals in the NYC neighborhoods are distributed. How do the costs differ for neighbourhoods, styles of assets and rental facilities? We'll discover how we can make predictions on prices depending upon the customers needs with machine learning models. Here, we'll focus on price prediction so we can immerse ourselves in future possible prices on the basis of actual price data from Airbnb. Using techniques of machine learning and powerful algorithms such as linear models of regression and gradient boosted regressor. Introducing new versions for ML Python, such as Scikit Teach, to work on.

IMPORTING LIBRARY AND DATA SET



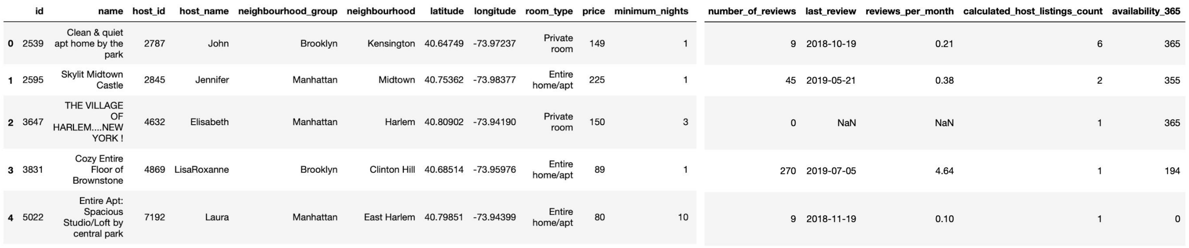
```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
%matplotlib inline
```



```
df = pd.read_csv("/Users/patel/Downloads/AB_NYC_2019.csv")
```

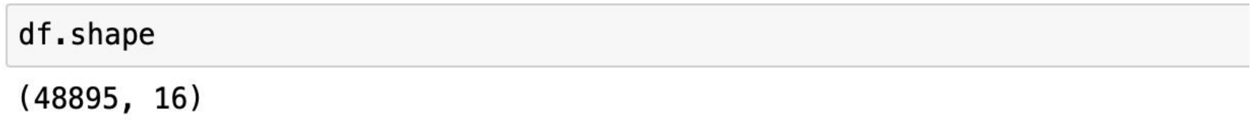
We will implement all the necessary files in this section, where we can find the facts to work with so let's extract our raw data. In order to make this Ipython Notebook work, we need to import the necessary libraries and frameworks before we begin.

Exploratory Data Set



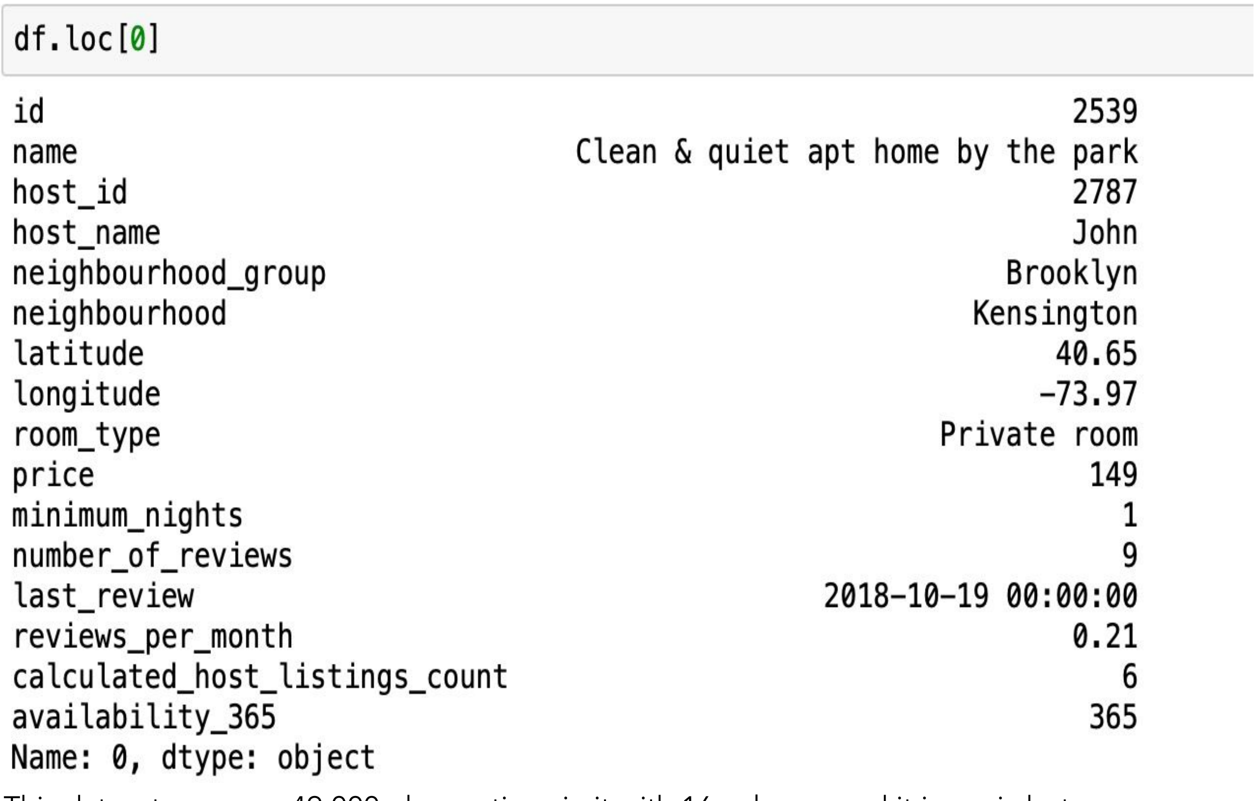
	id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
0	2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
1	2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
2	3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	365
3	3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
4	5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0

The database contains approx. 49,000 observations with 16 columns and a combination of categorical and numerical values.



```
df.shape
```

```
(48895, 16)
```



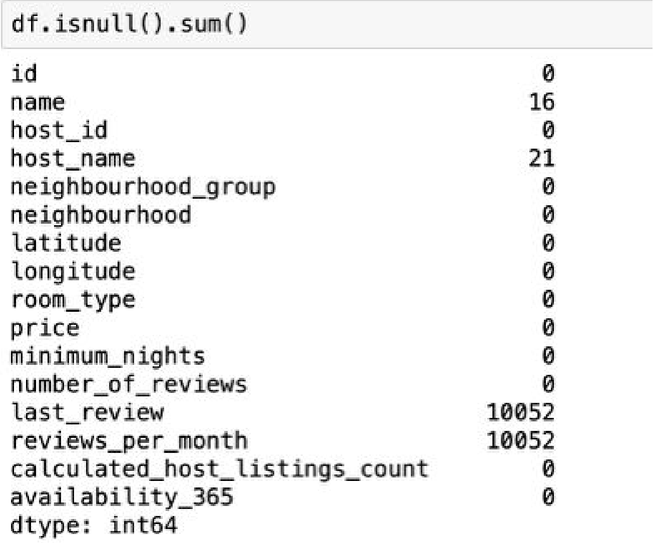
```
df.loc[0]
```

```
id                                                          2539
name                               Clean & quiet apt home by the park
host_id                                                     2787
host_name                                                   John
neighbourhood_group                                     Brooklyn
neighbourhood                                         Kensington
latitude                                                   40.65
longitude                                                 -73.97
room_type                                           Private room
price                                                        149
minimum_nights                                                 1
number_of_reviews                                              9
last_review                                  2018-10-19 00:00:00
reviews_per_month                                           0.21
calculated_host_listings_count                                 6
availability_365                                             365
Name: 0, dtype: object
```

This dataset conveys 49,000 observations in it with 16 columns and it is a mix between categorical and numeric values. After loading the dataset in and from the head of Airbnb dataset we can see a number of things. These 16 columns provide a very rich amount of information for deep data exploration we can do on this dataset. We do already see some missing values, which will require cleaning and handling of NaN values.

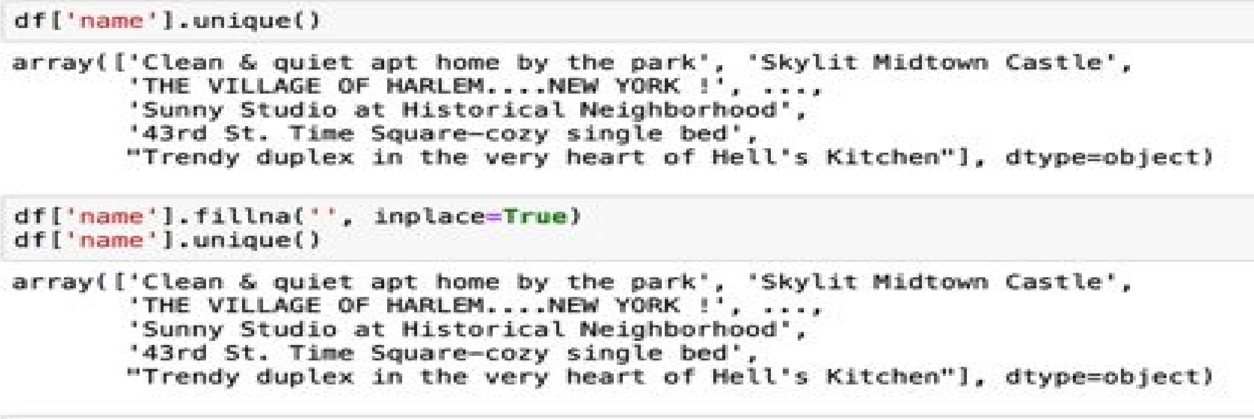
PREPROCESSING

Data preprocessing is a technique in data mining that requires raw data conversion into a functional format. Real-world data are frequently incomplete, inconsistent, or incomplete in certain conduct or trends, and may have many bugs. The preprocessing of data is a proven way to resolve these problems. The preprocessing of data prepares raw data for further processing. Information can have many unrelated and missing parts. Data cleaning is practiced to tackle this portion. It does the handling of missing data, noisy data, etc.



```
df.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
dtype: int64
```

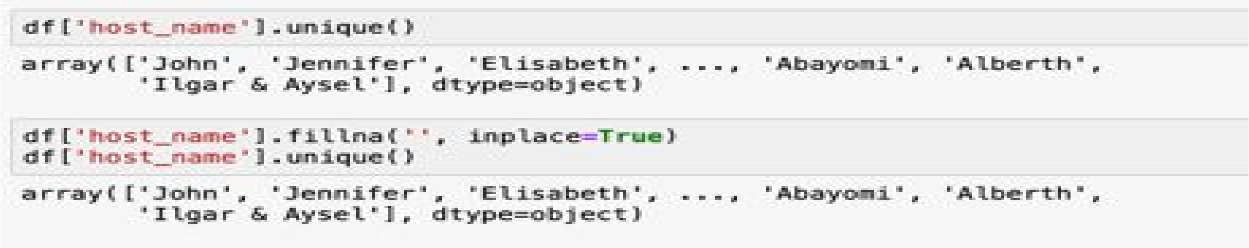


```
df['name'].unique()
```

```
array(['Clean & quiet apt home by the park', 'Skylit Midtown Castle',
       'THE VILLAGE OF HARLEM....NEW YORK !', ...,
       'Sunny Studio at Historical Neighborhood',
       '43rd St. Time Square-cozy single bed',
       "Trendy duplex in the very heart of Hell's Kitchen"], dtype=object)
```

```
df['name'].fillna('', inplace=True)
df['name'].unique()
```

```
array(['Clean & quiet apt home by the park', 'Skylit Midtown Castle',
       'THE VILLAGE OF HARLEM....NEW YORK !', ...,
       'Sunny Studio at Historical Neighborhood',
       '43rd St. Time Square-cozy single bed',
       "Trendy duplex in the very heart of Hell's Kitchen"], dtype=object)
```

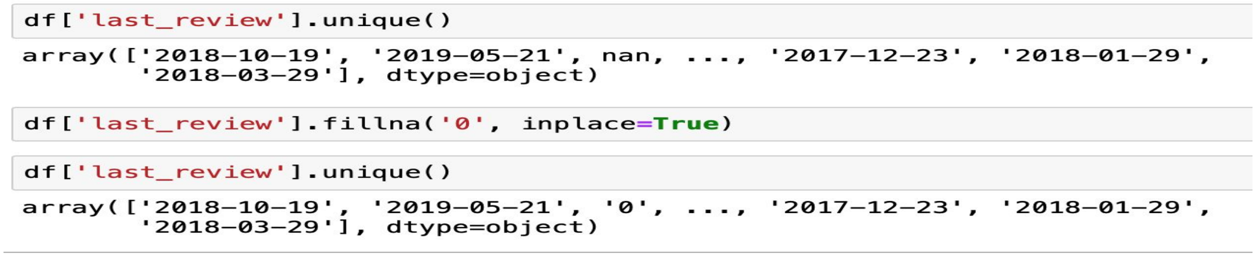


```
df['host_name'].unique()
```

```
array(['John', 'Jennifer', 'Elisabeth', ..., 'Abayomi', 'Alberth',
       'Ilgar & Aysel'], dtype=object)
```

```
df['host_name'].fillna('', inplace=True)
df['host_name'].unique()
```

```
array(['John', 'Jennifer', 'Elisabeth', ..., 'Abayomi', 'Alberth',
       'Ilgar & Aysel'], dtype=object)
```



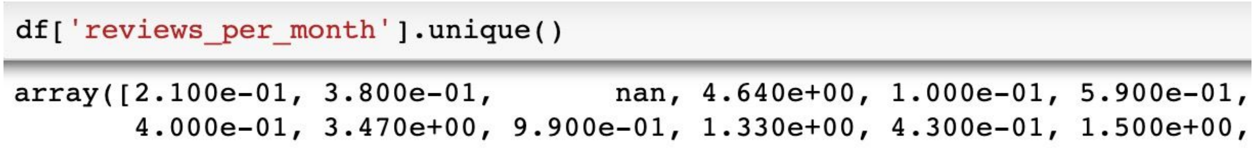
```
df['last_review'].unique()
```

```
array(['2018-10-19', '2019-05-21', nan, ..., '2017-12-23', '2018-01-29',
       '2018-03-29'], dtype=object)
```

```
df['last_review'].fillna('0', inplace=True)
```

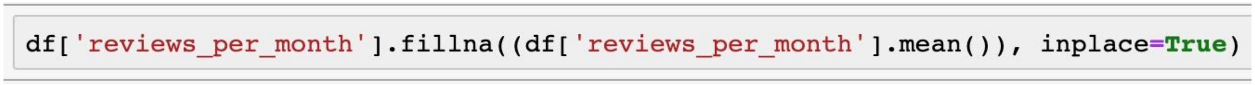
```
df['last_review'].unique()
```

```
array(['2018-10-19', '2019-05-21', '0', ..., '2017-12-23', '2018-01-29',
       '2018-03-29'], dtype=object)
```

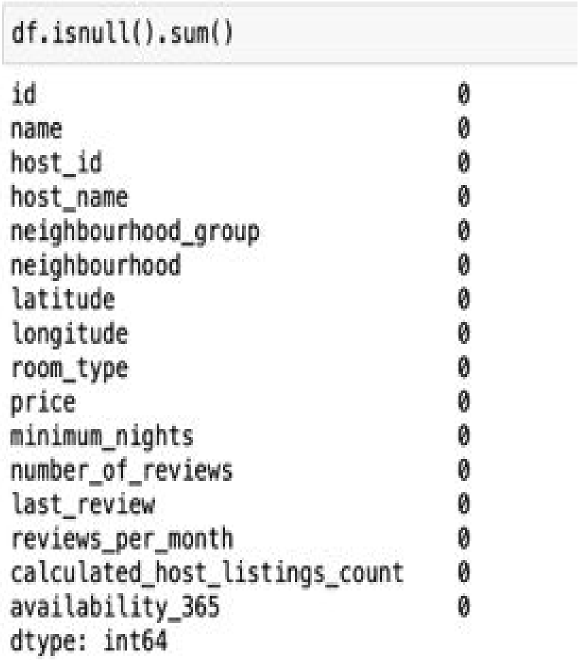


```
df['reviews_per_month'].unique()
```

```
array([2.100e-01, 3.800e-01,       nan, 4.640e+00, 1.000e-01, 5.900e-01,
       4.000e-01, 3.470e+00, 9.900e-01, 1.330e+00, 4.300e-01, 1.500e+00,
```



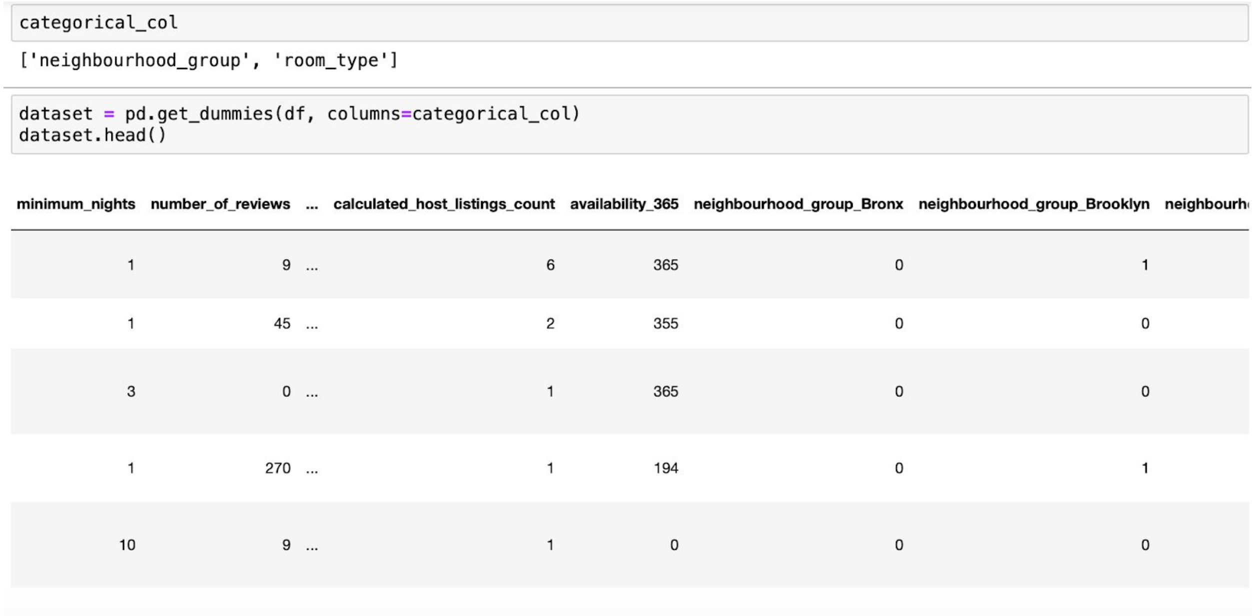
```
df['reviews_per_month'].fillna((df['reviews_per_month'].mean()), inplace=True)
```



```
df.isnull().sum()
```

```
id                                0
name                              0
host_id                           0
host_name                         0
neighbourhood_group               0
neighbourhood                     0
latitude                          0
longitude                         0
room_type                         0
price                             0
minimum_nights                    0
number_of_reviews                 0
last_review                       0
reviews_per_month                 0
calculated_host_listings_count    0
availability_365                  0
dtype: int64
```

In our case, the lack of data observed is not so crucial to our dataset. In the context of our database, we can suggest other things: the "name" and "hostname" columns are meaningless and unrelated to our data analysis, so we just replaced their null values to ' ' , and the "last review" and "review per month" columns need to be handled very efficiently. "last review" is date in order; if no reviews have been done for the listing-the date is simply not present, so we just replace the null values of last review with 0. For "review per month" we can add the mean for missing values.



```
categorical_col
```

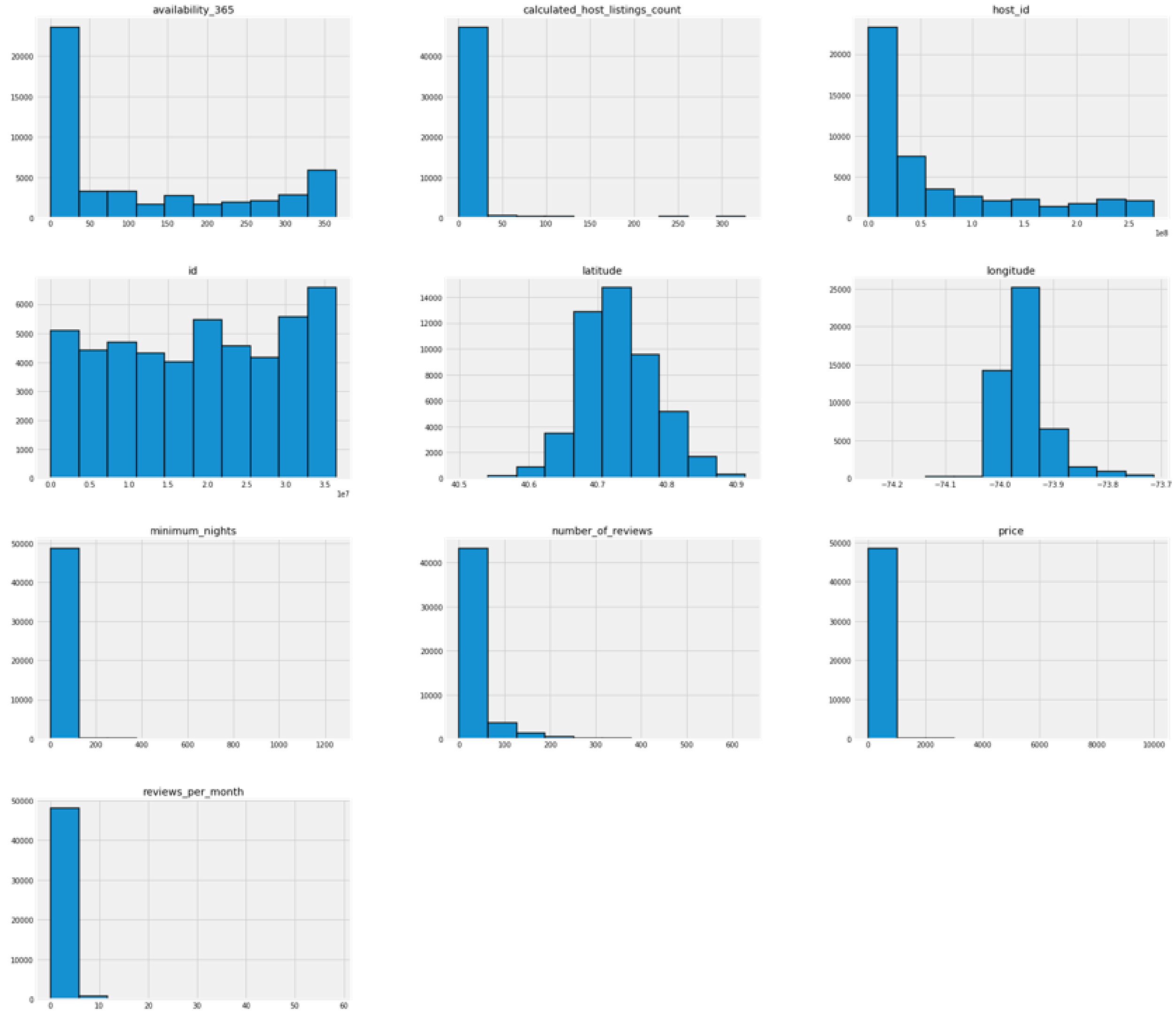
```
['neighbourhood_group', 'room_type']
```

```
dataset = pd.get_dummies(df, columns=categorical_col)
dataset.head()
```

minimum_nights	number_of_reviews	...	calculated_host_listings_count	availability_365	neighbourhood_group_Bronx	neighbourhood_group_Brooklyn	neighbourh
1	9	...	6	365	0	1	
1	45	...	2	355	0	0	
3	0	...	1	365	0	0	
1	270	...	1	194	0	1	
10	9	...	1	0	0	0	

We had two columns room_type and neighborhood_groups which contains categorical data so we had to use dummy variables. We used it to convert the categorical data into numerical values so it can help us to get better performing results for our algorithm and to price prediction.

Histogram:

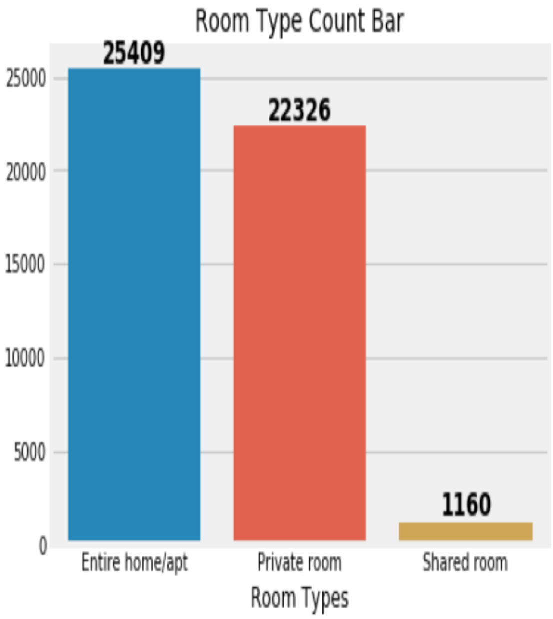

availability_365
0
5000
10000
15000
20000
0
50
100
150
200
250
300
350
calculated_host_listings_count
0
10000
20000
30000
40000
0
50
100
150
200
250
300
host_id
0
5000
10000
15000
20000
0.0
0.5
1.0
1.5
2.0
2.5
1e8
id
0
1000
2000
3000
4000
5000
6000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1e7
latitude
0
2000
4000
6000
8000
10000
12000
14000
40.5
40.6
40.7
40.8
40.9
longitude
0
5000
10000
15000
20000
25000
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
minimum_nights
0
10000
20000
30000
40000
50000
0
200
400
600
800
1000
1200
number_of_reviews
0
10000
20000
30000
40000
0
100
200
300
400
500
600
price
0
10000
20000
30000
40000
50000
0
2000
4000
6000
8000
10000
reviews_per_month
0
10000
20000
30000
40000
50000
0
10
20
30
40
50
60

ANALYSIS

Data analysis is a process of data inspection, cleaning, transformation and modeling with the aim of finding valuable information, reporting findings and assisting to make better predictions. Data analysis plays a role in making more scientific decisions in today's business world and helps businesses to function more efficiently. The three key aspects of data analytics are, Descriptive Analysis, Diagnostic Analysis & Predictive Analysis.

Descriptive Analysis

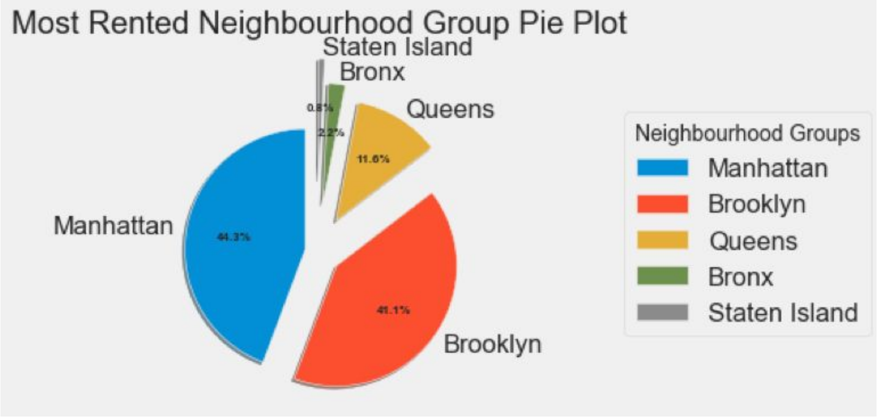
Descriptive analytics are the basis of analysis it provides BI tools and dashboards, which helps us to understand certain findings better. It answers our basic questions like "HOW MANY," "WHEN," "WHERE," and "WHAT."

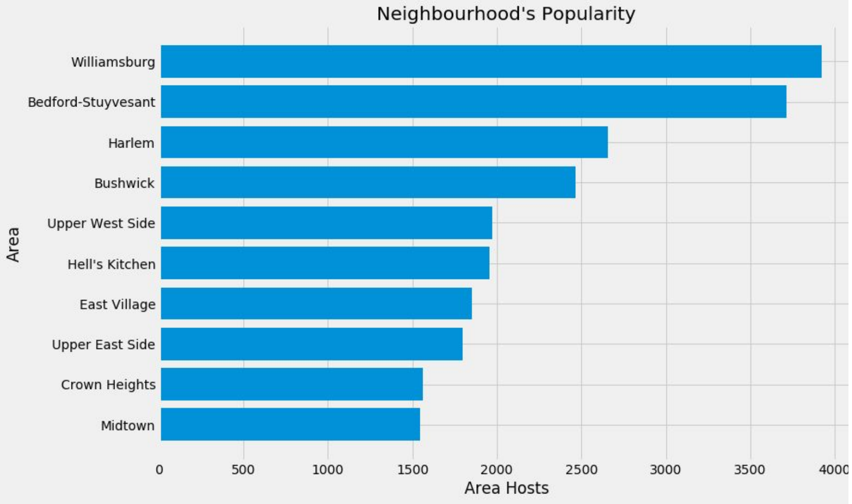


Number of Room Types Available

According to the graphic above, you can observe that the most rented type of room is the "Entire home / apartment" with a total count of 25409 followed by the "Private Room" and the "Shared Room" with 22326 and 1160. Evidently, all the people renting an Airbnb prefer a whole home.

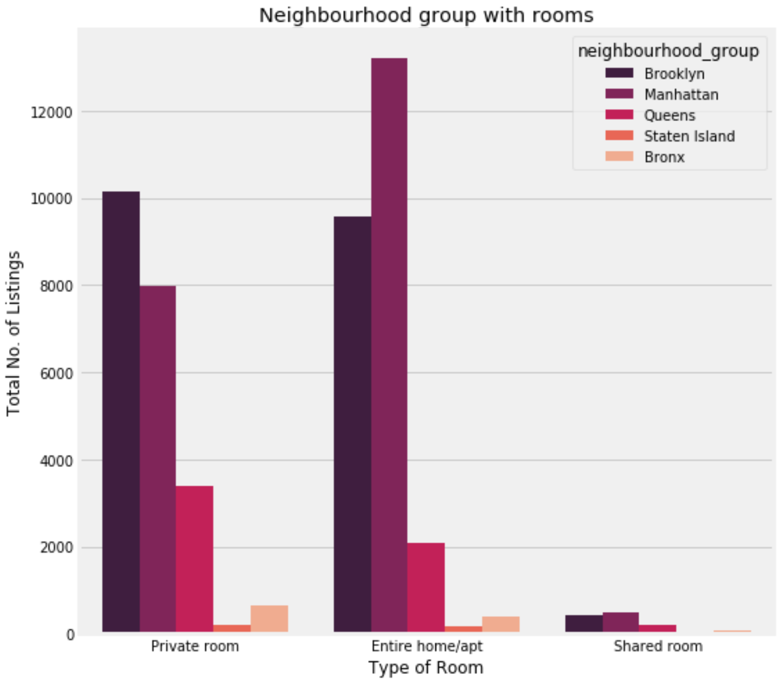
Percentage Representation of Neighbourhood Group in Pie





Neighbourhood Popularity

With the graph on the side, we found that Williamsburg in Brooklyn was eventually the most rented neighborhood.



Quantity of each type of room in each region by count plot

Manhattan and Brooklyn accumulate more than 75% for its properties in New York. They have brought together most of the best selling entire houses and private space. Therefore, the cost of renting in these regions will be higher.

Distribution Price by Neighbourhood Groups

As shown, in these graph our data is distributed by cost and divided by the regions so we can evaluate them according to the figure.

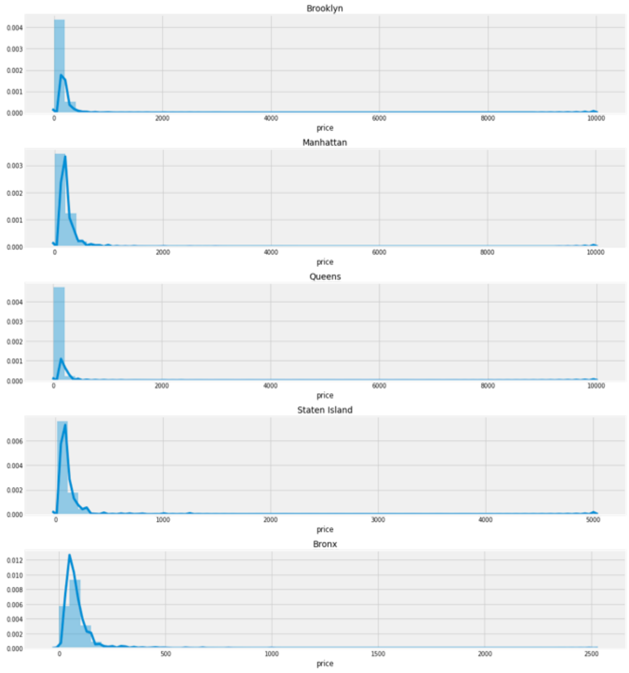
Brooklyn averages around 70-500$, per night depending upon the neighbourhood.

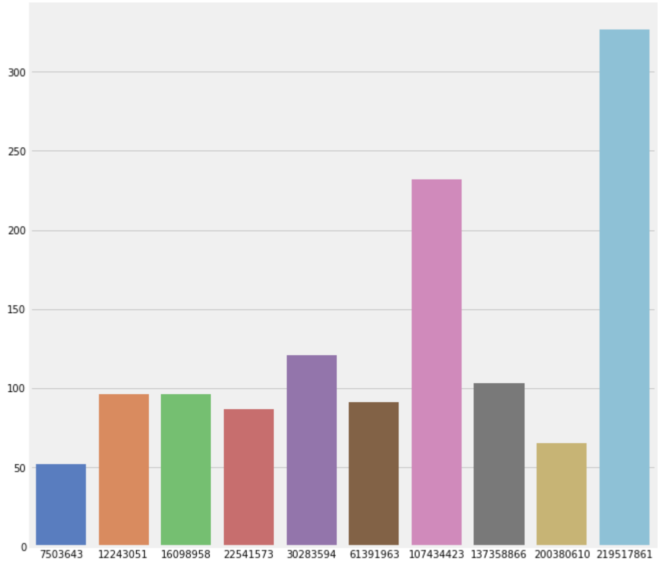
Manhattan averages around 80-490$

Queens averages around 60-280$

Staten Island averages around 50-800$

Bronx averages around 50-450$



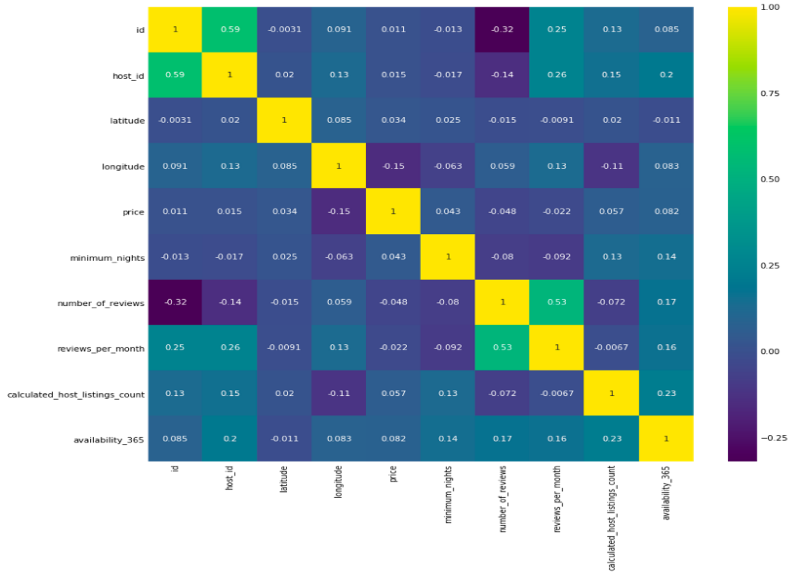


Top 10 Host IDs

This is the list of top 10 users who take the most advantage of airbnb by hosting at least more than 50 listings, while the top most has more than 300 listings. This graph portrays that distribution.

Diagnostic Analysis

Diagnostic data analytics is the method by which information is analyzed to understand the cause and the occurrence or why. This answers questions like "What has been done" and "Why has been made?" Diagnostic data analysis, in particular, helps to explain the justification for the occurrence of something. Its method of analytics is used by companies as they build more links between information and detect behavioral patterns.



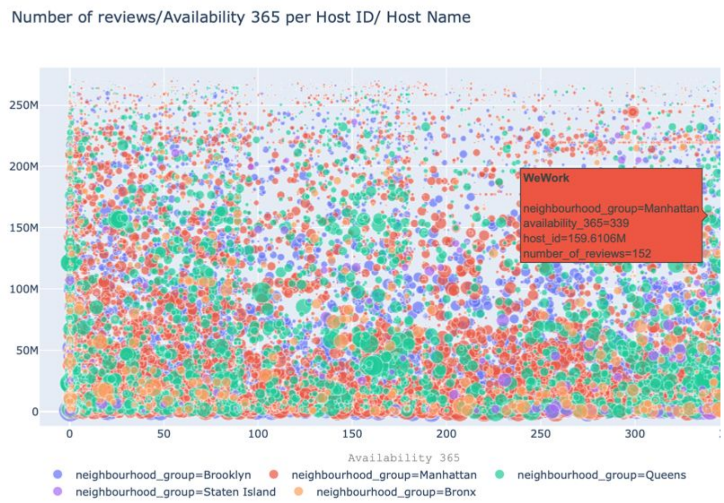
	id	host_id	latitude	longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
id	1	0.59	-0.0031	0.091	0.011	-0.013	-0.32	0.25	0.13	0.085
host_id	0.59	1	0.02	0.13	0.015	-0.017	-0.14	0.26	0.15	0.2
latitude	-0.0031	0.02	1	0.085	0.034	0.025	-0.015	-0.0091	0.02	-0.011
longitude	0.091	0.13	0.085	1	-0.15	-0.063	0.059	0.13	-0.11	0.083
price	0.011	0.015	0.034	-0.15	1	0.043	-0.048	-0.022	0.057	0.082
minimum_nights	-0.013	-0.017	0.025	-0.063	0.043	1	-0.08	-0.092	0.13	0.14
number_of_reviews	-0.32	-0.14	-0.015	0.059	-0.048	-0.08	1	0.53	-0.072	0.17
reviews_per_month	0.25	0.26	-0.0091	0.13	-0.022	-0.092	0.53	1	-0.0067	0.16
calculated_host_listings_count	0.13	0.15	0.02	-0.11	0.057	0.13	-0.072	-0.0067	1	0.23
availability_365	0.085	0.2	-0.011	0.083	0.082	0.14	0.17	0.16	0.23	1

HeatMap Correlation Visualization

In order to measure the linear correlation between two variables X and Y, the Pearson Correlation coefficient is used, which determines the correlation of columns pair-by-pair except for Null.There is no strong correlation except for number of reviews and reviews per month

Predictive Analysis

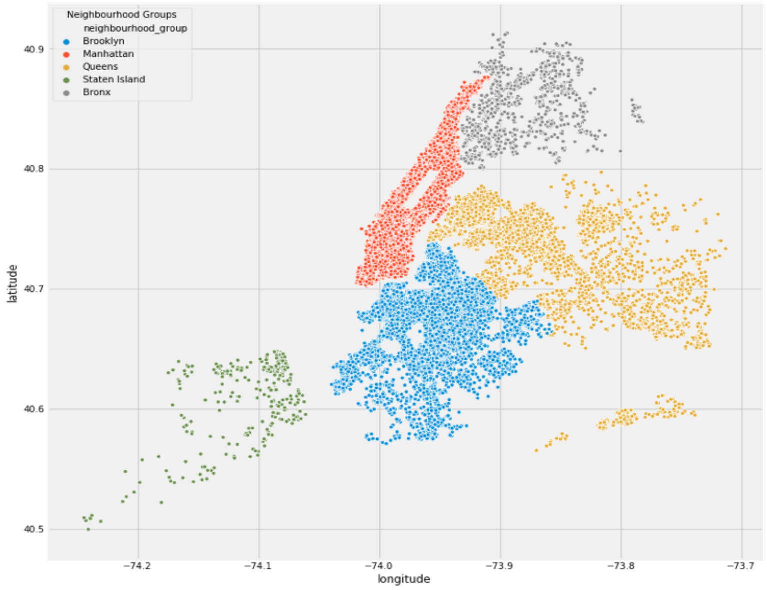
Predictive analysis is used to describe patterns, correlations, and triggers, and to try to answer' WHAT IS LIKE TO HAPPEN' questions. This type of analysis uses past data to predict future results. This analysis is a further step forward from the concise and diagnostic analysis. Predictive analysis uses the data presented to forecast event outcomes objectively.



Scatter Graph showing Number of reviews, Availability 365 per Host ID and Host Name

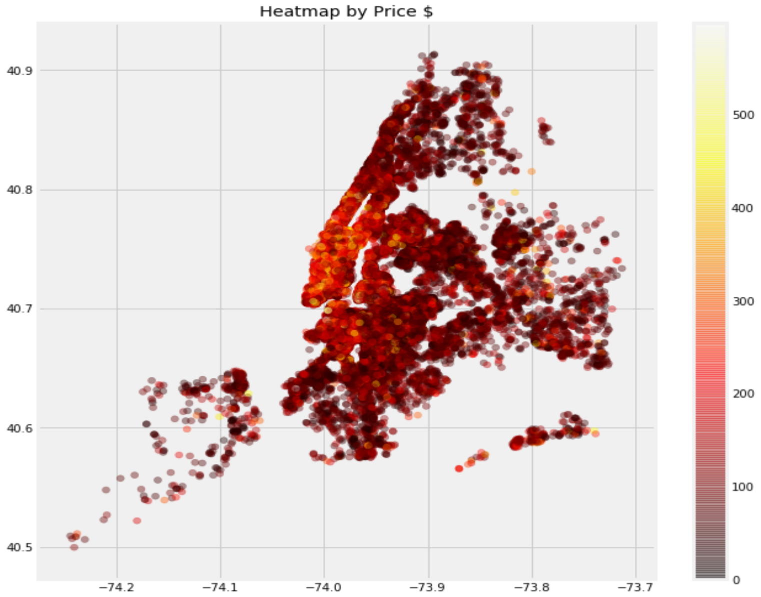
Data Visualizations

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected can be exposed.



Spatial Analysis by Neighbourhood Groups

This is a spatial analytical map, meaning it shows us the properties offered in the state of New York in different colors representing different neighbourhood and their location by longitude and latitude. For example we can understand that Bronx & Staten Island seem to have the lowest listings.



Price Overview in a HeatMap Normalization

This is a price-set heatmap, it helps us to identify that the most expensive offerings are in Manhattan and it is also the most rented. While generating this plot we had to lower the range to go up to $600 because, the highest offered property on airbnb prices at $10,000 and that makes the graph look unclear and harder to visualize. The rest of the regions seem to have equally distributed demands and supplies.

Mean Squared Error: Measuring the average of the error squares (average squared difference between the estimated values and actual value), or the average squared error regression of the estimator, of the procedure for determining the unexpected quantity.

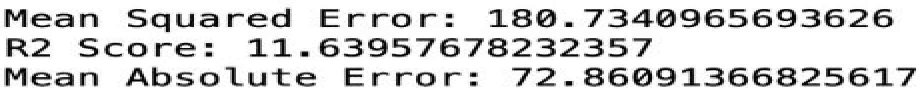
R^2 : It is a decision coefficient and is the score function of regression. The best score is 1.0 and can be negative, as the model could be worse arbitrary.

Mean absolute error regression loss: It is a difference measurement between two continuous variables. Suppose X and Y are parallel observation variables that express the same phenomenon.

Linear Regression Model

Linear regression is intended to be used for modeling the relation between the two variables by a linear equation. The methodology uses statistics to trace a trend line in a sequence of data points. The trend line in our report explains the relation between actual and forecast values. There are a variety of ways for estimating linear regression between an independent variable and a dependent variable under analysis. One of the most common methods for calculating unknown variables is the normal least square method, which visually converts the sum of the vertical distances between the datasets and the trend line.

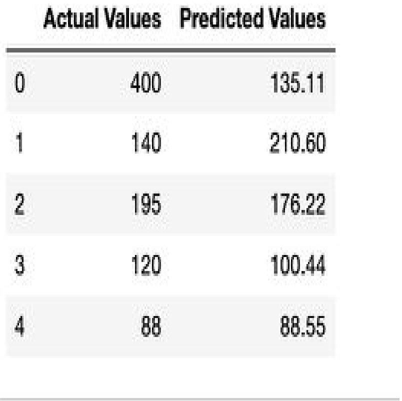
Predictions & Metrics:



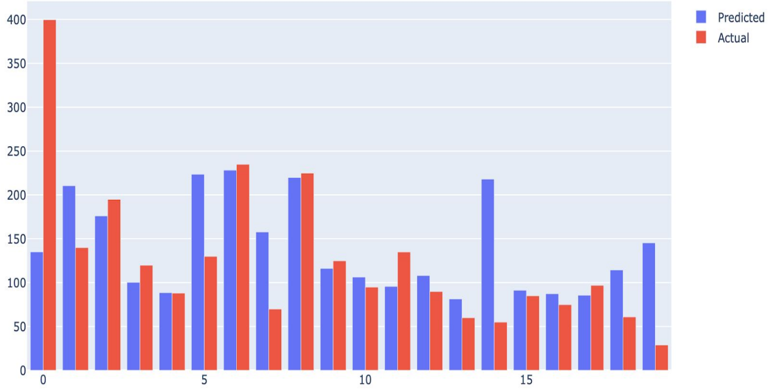
```
Mean Squared Error: 180.7340965693626
R2 Score: 11.63957678232357
Mean Absolute Error: 72.86091366825617
```

Results

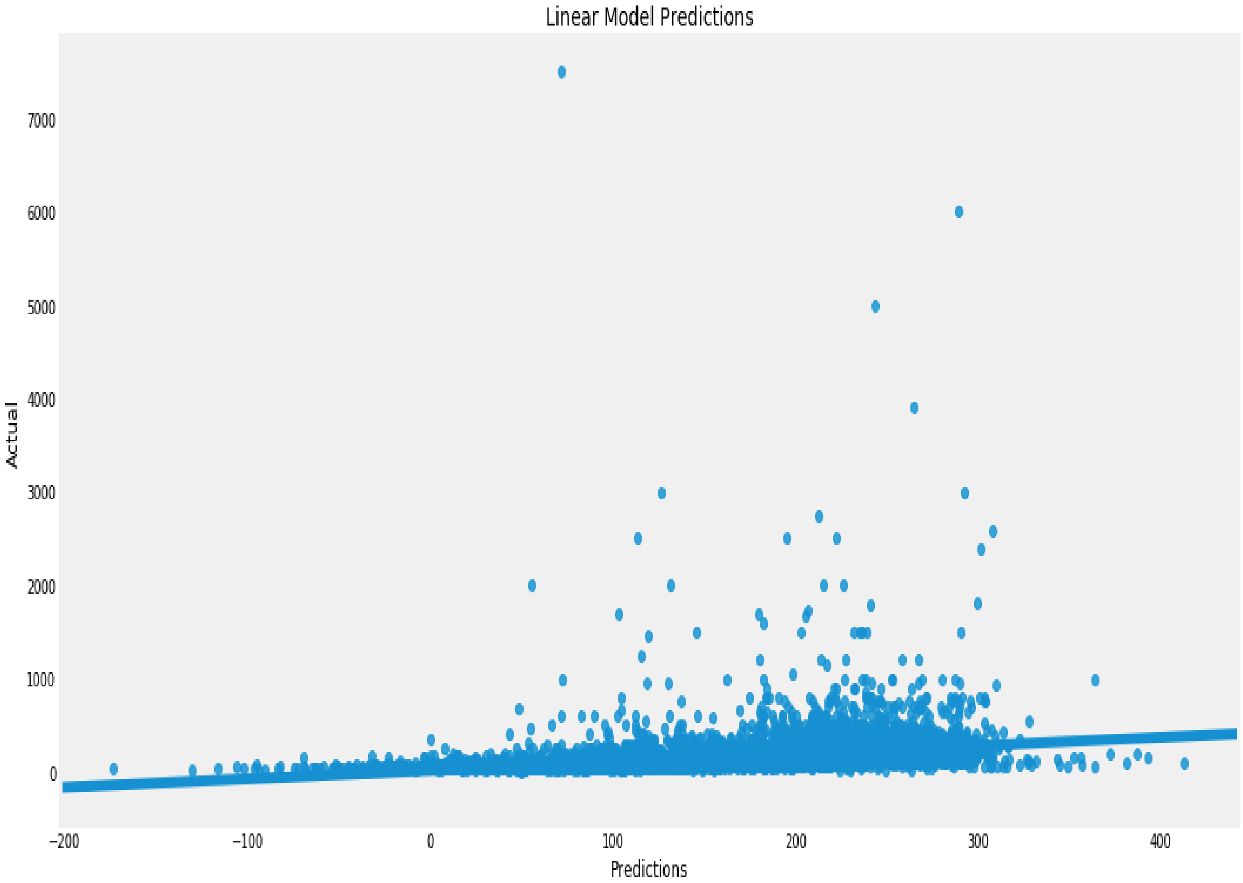
Actual Values VS Predicted Values



	Actual Values	Predicted Values
0	400	135.11
1	140	210.60
2	195	176.22
3	120	100.44
4	88	88.55



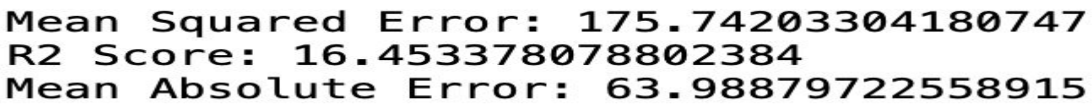
Linear Model Predictions



Gradient Boosted Regressor Model

Gradient boosted is a machine learning technique that creates a prediction model as a whole of weak models for prediction. It constructs the model stage-specifically, as do other boosting methods.Gradient Boosted helps us build an additive model in a more modern way, it makes it easy to optimize the arbitrary differentiable loss function. Using GB means in each stage we will encounter a regression tree is fit on a gradient of any given loss. Every supervised learning algorithm has the purpose of defining and reducing a loss function.Applying the gradient boosted algorithm involves continuous exploitation of residual patterns and refining and improving a model with weak predictions. Once we are at a point where there is no residual model, we should avoid modeling residues, otherwise it could lead to over-fitting. Algorithmically, we minimize our loss function.

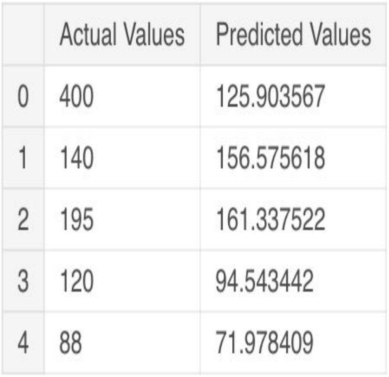
Predictions & Metrics:



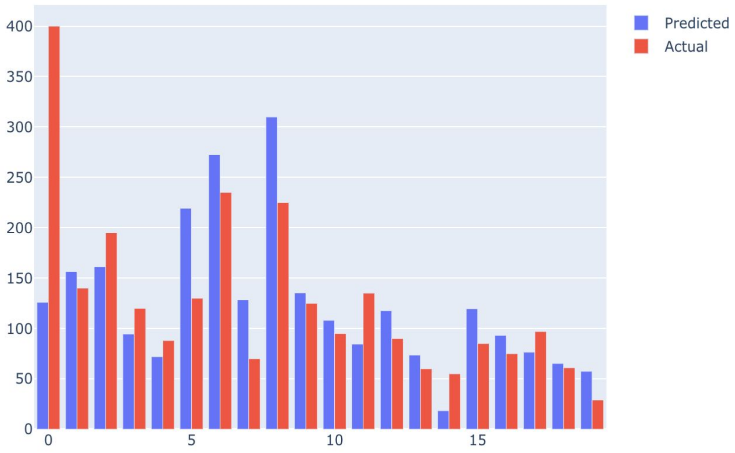
```
Mean Squared Error: 175.74203304180747
R2 Score: 16.453378078802384
Mean Absolute Error: 63.98879722558915
```

Results

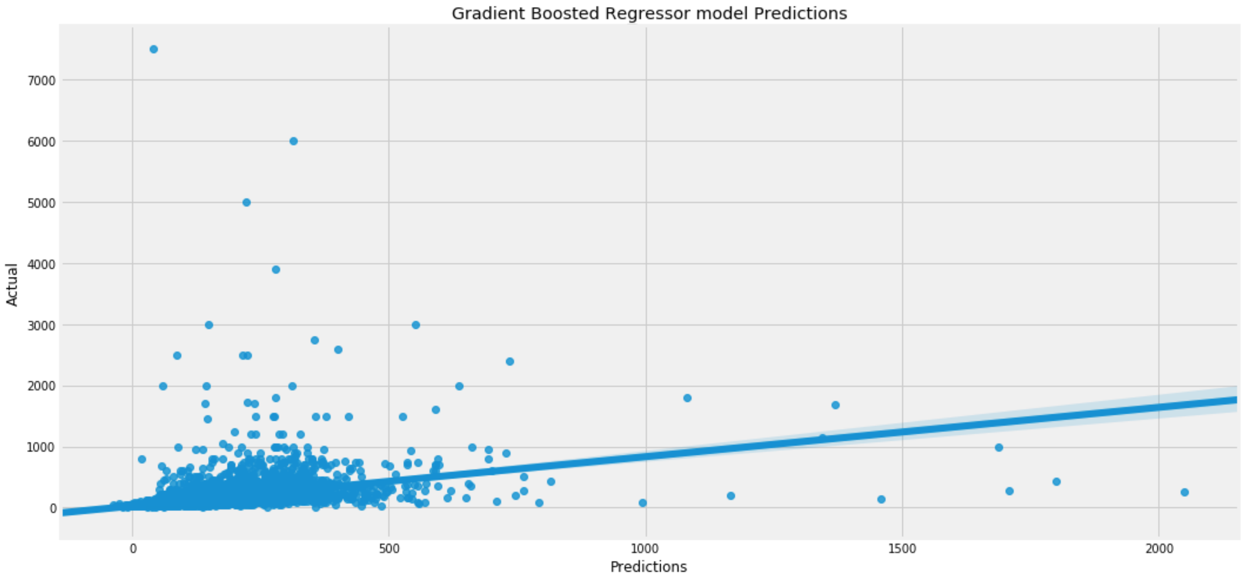
Actual Values VS Predicted Values



	Actual Values	Predicted Values
0	400	125.903567
1	140	156.575618
2	195	161.337522
3	120	94.543442
4	88	71.978409



Gradient Boosted Regressor Model Predictions:



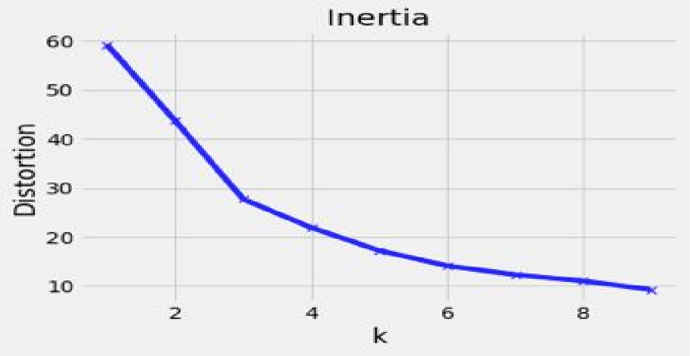
Clustering algorithm

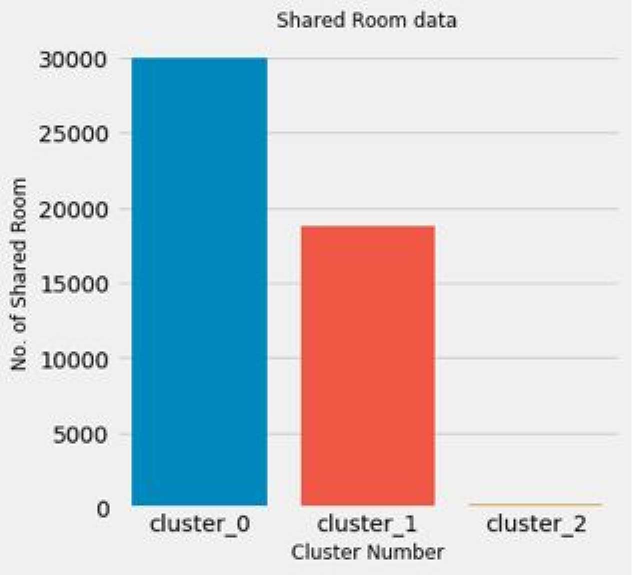
In machine learning clustering is an important tool. It involves grouping data points using a clustering algorithm which assigns the data points to a certain category as required.

For example, if someone is evaluating to do business on a shared room in New York, how are we able to set a fair price? For this matter we can use the cluster model.

K-means

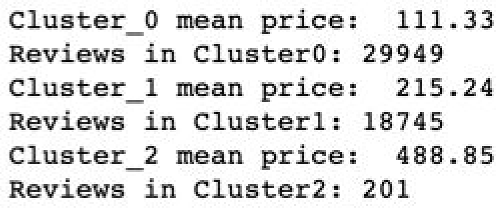
The cluster number is described by this elbow matter





The cluster 0 domain has the highest proportion of shared rooms. The least shared rooms are in group 2. Group 0 is the most appropriate cluster.

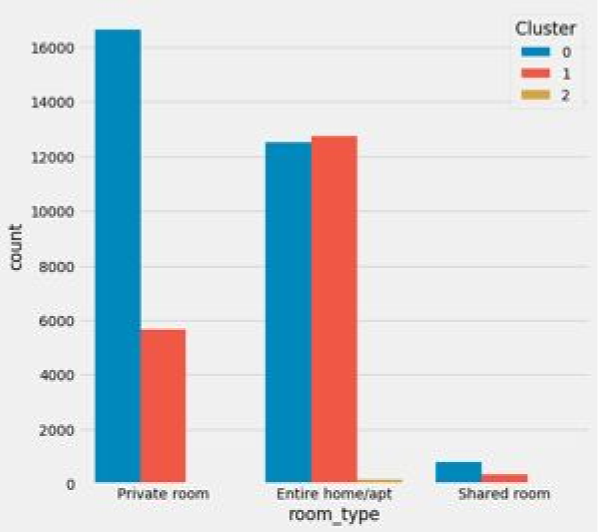
Check the cluster 0, cluster 1, and cluster 2 for their average price.



```
Cluster_0 mean price:  111.33
Reviews in Cluster0: 29949
Cluster_1 mean price:  215.24
Reviews in Cluster1: 18745
Cluster_2 mean price:  488.85
Reviews in Cluster2: 201
```

According to clusters 0, 1, and 2, average price, and we can set our shared room average rate at NY for $271.81.

We can then see which rooms in each cluster are more popular.

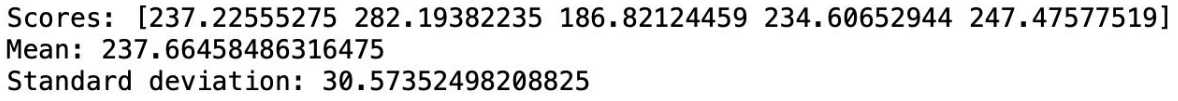


To conclude, We found that if we have a private sharing room, we should refer to pricing in cluster 1. We will relate to cluster 0 of the price when we have an entire apt. Manhattan and Brooklyn can be selected for the city.

K-fold cross validation

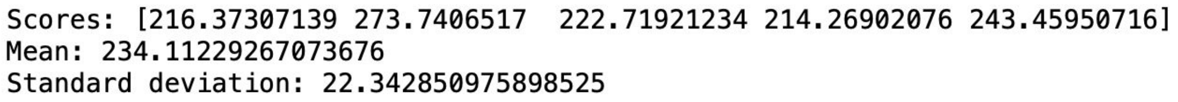
Cross-validation is one of the techniques used to test machine learning models ' effectiveness and is also a resampling tool of validating a model if the data is small. K-Fold is a standard model that is easily understood and usually produces less distraction from other models. It ensures that every observation from the original data set is allowed to appear in the course and test set.

Linear Model:



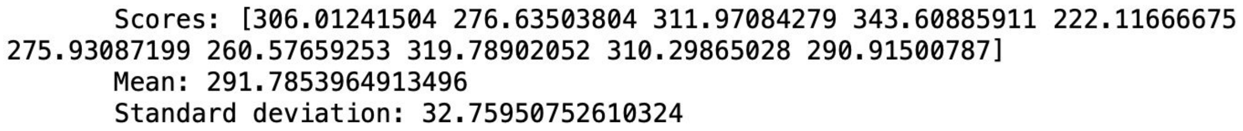
```
Scores: [237.22555275 282.19382235 186.82124459 234.60652944 247.47577519]
Mean: 237.66458486316475
Standard deviation: 30.57352498208825
```

Gradient Boosted Model:



```
Scores: [216.37307139 273.7406517  222.71921234 214.26902076 243.45950716]
Mean: 234.11229267073676
Standard deviation: 22.342850975898525
```

K mean:



```
        Scores: [306.01241504 276.63503804 311.97084279 343.60885911 222.11666675
 275.93087199 260.57659253 319.78902052 310.29865028 290.91500787]
        Mean: 291.7853964913496
        Standard deviation: 32.75950752610324
```

Conclusion

For the 2019 year, this Airbnb dataset seemed to be a very rich dataset with a wide range of columns, which enabled us to explore that critical column in detail.

We concluded, that manhattan has most expensive rooms available and it has most expensive availability. The price of single room is less compared to whole apartment. Shared rooms are cheaper but are less available in manhattan and brooklyn.

To predict prices over the years, we have to use predictive analysis using the latest stack technology. In the context of Artificial Intelligence (AI), we have used machine learning and have applied the latest and most optimized algorithms, such as the linear regression model, and the gradient boosted regression models.

Ultimately, we found several interesting relationships among features and clarified each step of the process.