Comparative study of Airbnb and Property Sales Prices in New York

Laiba Rehman
MSc Data Analytics
National College Of Ireland
Dublin, Ireland
x20144032@student.ncirl.ie

***Abstract*— For this paper, three datasets are chosen based on one particular domain. On all of the three datasets the processes of data cleaning and data pre-processing is implemented. Laten on, in order to understand the nature of the data, the different analytical processes like descriptive, prescriptive, and exploratory was performed to get better insights about the data. These understandings were taken for further analysis and prediction of the variable "sale_price" of Sales property given in two datasets and Airbnb variable "price" given in one dataset. After the process of cleaning, there are a lot of unwanted columns which needed to be dropped and outliers which needed to be examined. Thus, the study was performed to see the correlation between the three datasets like does Airbnb or Property on sale Prices fluctuate according to the regions, reviews etc. All of my three datasets have been procured from the Kaggle website. Data Analysis and forecasting was performed using the integrated data warehouse. Prices were predicted using the available data. Significant correlation between Price and regions or Price and room type is identified by our analysis. For the purpose of prediction various machine learning models were used such as Multiple Linear Regression, Random Forest Regression, Decision Tree Regression, Gradient Boosting Regression and KNN. Sklearn libraries were used to perform machine learning activities in python. After implementing all of these models mentioned above the best model was chosen in each of the dataset based on the value of R^2 and adjusted R^2.**

Keywords—price, borough, New York, crime, room type, location

I. Introduction

In the whole world, New York real estate has always been the center of attention as millions of people live or go there for their livelihoods. The cost of living there is insane as New York is one of the most expensive and competitive cities in the world or all of America. New York city is a fascinating and interesting place, people love to visit there for the lights, food, fun, jobs and many more reasons. Around 8.5 Million people live there and 20 million in whole of America so it has several million people communicating every day in and out of New York for their jobs. New York is considered to be one of the most expensive cities to buy a real estate.

The traditional accommodation for the tourists is hotel but after the revolution of the new digital economy new concepts is emerging which is known as peer-to-peer accommodation which is gaining a lot of popularity. Since 2008, the hospitality services of the Airbnb have been changing around the world. Hence, the company is growing at a very faster rate now it has listings of 3 million in 181 countries and about 81000 cities. In the list of most visited cities in the world, New York surely comes around the top positions, as it has also received an award for having around 67 million of tourists in year 2019. The prices of motels, hotel and hostels are pretty expensive in New York making the Airbnb option quite good for the customers who are price sensitive. As everything has its pros and cons thus, the prices in Airbnb rentals can vary as well according to the property, location and reviews.

Fascinated by these reasons I decided to choose these datasets and this domain.

In this study the process of KDD has been followed in which the main target is finding the knowledge in data and emphasizing on the "high-level" application of particular data mining methods. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

Following are the questions answered in this project.

1) *What is the correlation between the room type, regions and reviews with respect to the Airbnb Prices and whether the Prices differ based on these three factors?*
2) *Which state is the most expensive to live in New York and which is affordable to live in?*
3) *Which States have the maximum number of rooms available?*
4) *In which county of New York maximum houses were sold?*
5) *How much the sales properties prices differ based on the regions?*

II. Related Work

1) Yu and Wu, previously attempted to predict the real-estate property prices using various machine learning algorithms like Random Forest regression, linear regression and SVR. Prices was classified into 7 classes by using the SVC, Random Forest, Naïve Bayes and Logistic Regression. For their project the best model predicted was SVR with RMSE value of 0.53 and 69% accuracy.[1]
2) In the study done by Ma et al., he used different models like Gradient Boosting Regression Trees, Linear Regression, Linear Regression for the analysis of the Beijing rental prices. The best model predicted was the

tree regression model with RMSE value of 1.05CNY/m^2-day.[2]

3) Another study which was more relevant to this work, it was done to investigate the sharing economy and hotels rental prices. It is a work done by Wang and Nicolau, they did a study on analyzing the listings of Airbnb by measuring the quantile regression and ordinary least squares.[3] A work very similar to this done by Masiero et al, he made use of quantile regression model in order to study about the relationship between the hotel prices and total traits.[4]
4) Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, Hoormazd Rezaei, in this study they tried to do something different by applying Neural Networks and getting analysis with the sentiments by calculating the Mean Absolute Error, Mean Squared Error and R^2 square for the analysis of Airbnb prices. In this project they stated that plethora of features can lead to weak performance and high variance of the model. The best model predicted on the training set was Lasso based it used advanced models like neural networks and SVR to achieve the higher value of R^2. In the test set, SVR performance was the best with MSE value of 0.147 and R^2 of 69%. [5] They didn't apply Random Forest Regression on this model. So, I thought to apply this in my project to check the results.
5) The host of the Airbnb should know what are the requirements of an Airbnb when compared with hotels. There are many features in the listings that affects the prices of Airbnb thus it becomes important to look at the relation between the features affecting the prices. Thus the host can take care of the attributes in the houses which also helps the community grow and by keeping the prices in check as well. The nature and behavior of the Airbnb host also play a major role in the renting.[6] Majority people do not know on what measures they should price their property. Therefore, Price attribute becomes one of the main factors which affects the system of accommodation, thus making it essential to find the determinant of price.[7]
6) One of the main important things is that the platform of Airbnb is regularly managed. It is often seen that some do take care of it while some do not, the decision about any financial activity should be taken by taking account of regulations so that all the visitors benefit by this platform.[8] While considering the attributes, one important factor is the ratings given by the visitors because ratings plays an important part in order to make the new visitors trust that the area is worth spending their money and time. It is considered a vital role to also look at the ratings of the competitors of business to know which option suits your requirement the best.[9] One of the impacts of increasing the Airbnb business was seen on the business of Hotels as there was a decrease in the number of people going there.[10]
7) Technology has played a vital role in the development of the Airbnb's because of which many accommodation problems have been solved and also helped in making customers act as entrepreneurs, making people aware of the new innovative ideas about using online resources giving people opportunities to book and view the properties as per their needs.[11]
8) Questions that could be raised which I found pretty interesting by reading these articles were that what difference does it made to the hotel prices before and after the Airbnb business started. Who are those people who used to go to the hotels but now they prefer Airbnb? or if they use both of the facilities equally now. It will be really helpful if we get to analyze that the popularity of Airbnb is just based on the prices or there are other factors involved as well.[12]
9) The previous studies done on the Airbnb data not sufficient when it comes of evaluation of performance and metrics. The author Tang and Sangani did the prediction of price on the listings of San Francisco Airbnb dataset. We can learn from this article is how he has turned the problem of regression into the problem of binary classification by splitting the dataset on the basis of median, which in turn effectively decreases the complexity of the task.
10) Kalehbasti et al., he studied on the regression model for the price prediction in New York Airbnb dataset. A couple of machine learning algorithms such as KMC, SVR, NN etc. were applied and based on the sentiment analysis data was integrated. The value of R^2 was calculated as 0.7246. For the evaluation of the metrics instead of original scale they used logarithmic scale.

III. Methodology

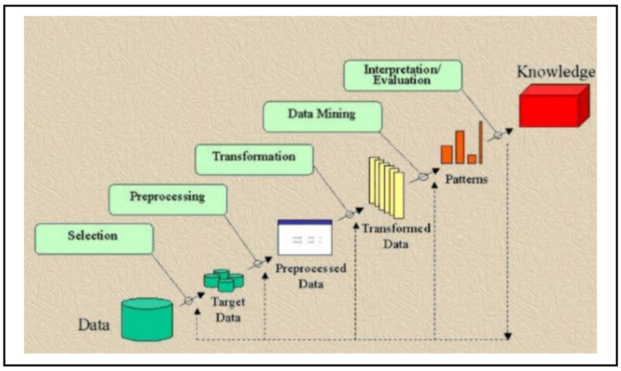
Many handful techniques are available to extract unstructured data which is meaningful. Here, we have two unique methodologies i.e., KDD and CRISP-DM which can be implemented on our projects based on the structure and requirements of our dataset. These methodologies are applied because they help us to make our implementation successful and also help us to manage, organize and plan our project in order to make it in a proper flow.

KDD i.e. (Knowledge Discovery in Databases) is a process of finding knowledge or information in data and then using that data in a broad way in data mining technique. With the help of this process, we prepare the data, do the selection process, prior knowledge is applied based on some facts, data is cleaned if it has any outliers or missing values and then analysis is performed on the calculated results to gain some very rich useful insights.

KDD process consist of following steps which are:

- **Data Cleaning:** In this process the outliers and noises are removed by the collecting the necessary information to make a model which is more efficient and accurate for data analysis.

- **Dataset Integration:** After the process of cleaning is completed. The data is combined from the various source in the Datawarehouse. The ETL process is applied for integration.
- **Data Selection:** Once the process of setting our goal is done, the process of Data Selection, where the relevant data for our analysis is selected and the data from the collection is retrieved.
- **Data Transformation:** The process of transformation is caried out to convert the data into appropriate from and to prepare it for further analysis of Data mining. In this with the help of transformation methods or dimensionality reduction, an effective number of variables are reduced in order to find the data for invariant representations.
- **Data mining:** The main objective is to decide whether the KDD process is going to be applied on Regression, Classification or Clustering algorithms for data implementation. Here a pattern of interest is searched and observed in the relevant data represented in a particular form.
- **Pattern Evaluation:** When the patterns represented are finalized than the process of evaluation of pattern starts. The final patterns are observed for more insights and based on that further visualizations can be made.
- **Knowledge Representation:** A knowledge Representation is a kind of a technique that uses visualization tools in order to represent the results or insights of the project and even the reports are made out of it.



Graphical Representation of steps of KDD process.

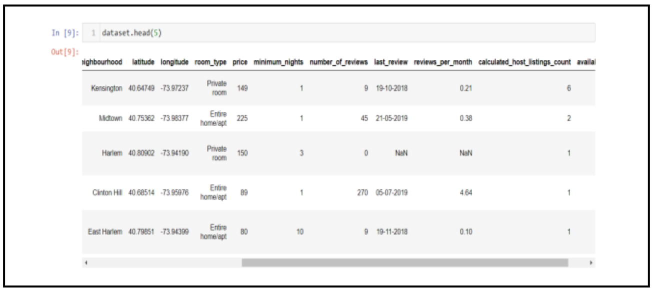
The intent of this project is to investigate and gain insights by using different techniques of data mining and to apply the different Machine Learning algorithms. The main objective of KDD process is evaluation whereas the main focus of DRISP-DM is business, therefore I have chosen KDD process for the evaluation of my project.

A. Dataset Description

Dataset-1 New York City Airbnb

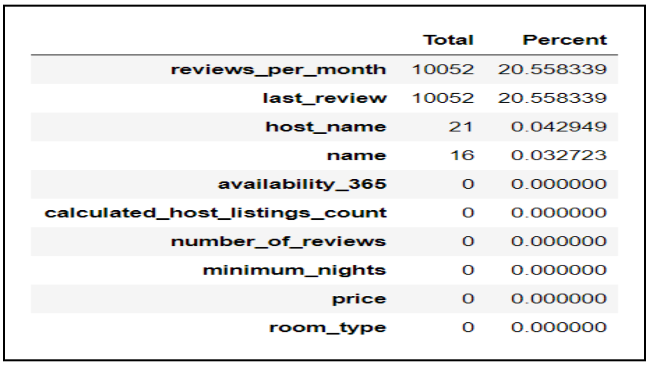
1) *Data Preprocessing:*

The Airbnb dataset on New York consist of 49000 observation and 16 columns in it. It comprises of a mixture of both numeric and categorical values. First step was to load the CSV file in the Jupyter notebook using Pandas library and read_csv function. After that using the head of Dataset, I could see the couple of things that the presented 16 columns provided a rich amount of information in which depth of data exploration could be performed. Some missing values could also be seen represented as NAN which also needs to be taken care of later in order to increase the efficiency of the model.



```
In [9]:  dataset.head(5)
Out[9]:
```

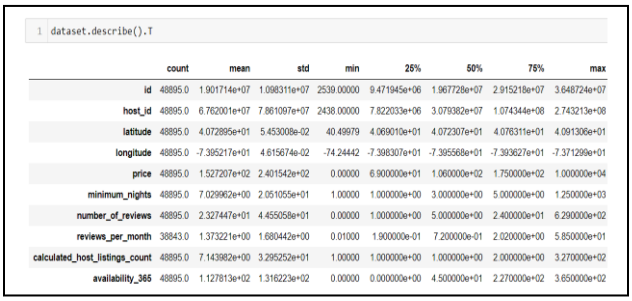
ighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availal
Kensington	40.64749	-73.97237	Private room	149	1	9	19-10-2018	0.21	6	
Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	21-05-2019	0.38	2	
Harlem	40.80902	-73.94190	Private room	150	3	0	NaN	NaN	1	
Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	05-07-2019	4.64	1	
East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	19-11-2018	0.10	1	



	Total	Percent
reviews_per_month	10052	20.558339
last_review	10052	20.558339
host_name	21	0.042949
name	16	0.032723
availability_365	0	0.000000
calculated_host_listings_count	0	0.000000
number_of_reviews	0	0.000000
minimum_nights	0	0.000000
price	0	0.000000
room_type	0	0.000000

Fig 1

Missing Values: The sum of all the null values and percentage is taken to know how many missing values are present and in which variables. After evaluating, it was seen that 20% of the values are missing in the columns: reviews_per-month and last_review.



```
dataset.describe().T
```

	count	mean	std	min	25%	50%	75%	max
id	48895.0	1.901714e+07	1.098311e+07	2539.00000	9.471945e+06	1.967728e+07	2.915218e+07	3.648724e+07
host_id	48895.0	6.762001e+07	7.861097e+07	2438.00000	7.822033e+06	3.079382e+07	1.074344e+08	2.743213e+08
latitude	48895.0	4.072895e+01	5.453008e-02	40.49979	4.069010e+01	4.072307e+01	4.076311e+01	4.091306e+01
longitude	48895.0	-7.395217e+01	4.615674e-02	-74.24442	-7.398307e+01	-7.395568e+01	-7.393627e+01	-7.371299e+01
price	48895.0	1.527207e+02	2.401542e+02	0.00000	6.900000e+01	1.060000e+02	1.750000e+02	1.000000e+04
minimum_nights	48895.0	7.029962e+00	2.051055e+01	1.00000	1.000000e+00	3.000000e+00	5.000000e+00	1.250000e+03
number_of_reviews	48895.0	2.327447e+01	4.455058e+01	0.00000	1.000000e+00	5.000000e+00	2.400000e+01	6.290000e+02
reviews_per_month	38843.0	1.373221e+00	1.680442e+00	0.01000	1.900000e-01	7.200000e-01	2.020000e+00	5.850000e+01
calculated_host_listings_count	48895.0	7.143982e+00	3.295252e+01	1.00000	1.000000e+00	1.000000e+00	2.000000e+00	3.270000e+02
availability_365	48895.0	1.127813e+02	1.316223e+02	0.00000	0.000000e+00	4.500000e+01	2.270000e+02	3.650000e+02

Fig 2

Exploring Numerical Variables: To get more insights I used the describe function on Dataset which gave me some very useful unprecedented insights that the average price of rooms is 157 dollars and they reached up to the maximum of 10000 dollars. People, on an average prefer to spend 7 days in rooms which indicated that people mostly prefer a week of holiday. One interesting insight was that someone stayed in Airbnb for almost 4 years i.e., 1250 days.

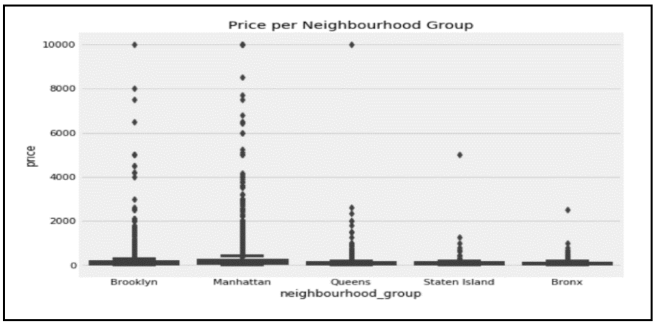


Fig 3

Removing Outliers: Fig 3 represents the graph made with the help of a boxplot between price and neighbourhood_group I could see that there are many outliers present which could affect the efficiency of my model so I decided to collect necessary information and remove the outliers on the basis of quantile function by setting the range of min and max threshold to 0.05 and 0.95 respectively and then removing the points which do not lie in this range. Later on, the graph looked like as shown in below figure 4.

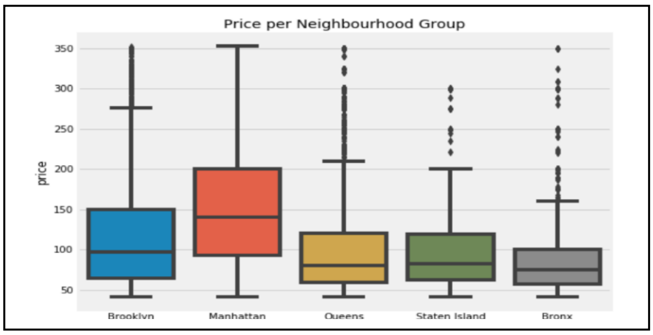


Fig 4

Encoding Categorical Variables: I had categorical data which needed to be converted to numeric. Thus, by importing the preprocessing module from sklearn library I imported the LabelEncoder() class. I chose three of my categorical variables "neighbourhood_group", "neighbourhood" and "room_type" to be converted to numerical variables.

2) Data Transformation:

By observing the nature of our dataset further things that could be stated was that few columns: "name", "host_name" and "id" are insignificant and irrelevant to our analysis of data. Moreover, the column "last_review" consist of date but I do not have a variable of reviews so keeping the variable "last_review" is of no use. Thus, I decided to drop these variables.

3) *Data Visualization:*

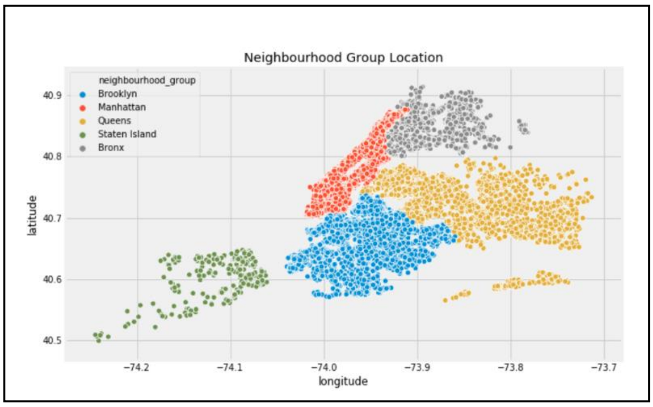


Fig 5

Manhattan is the most expensive area to live in and Staten island was listed as having rooms which are low priced.

Manhattan and Brooklyn have the maximum number of hotels which were shown by the map graphs.

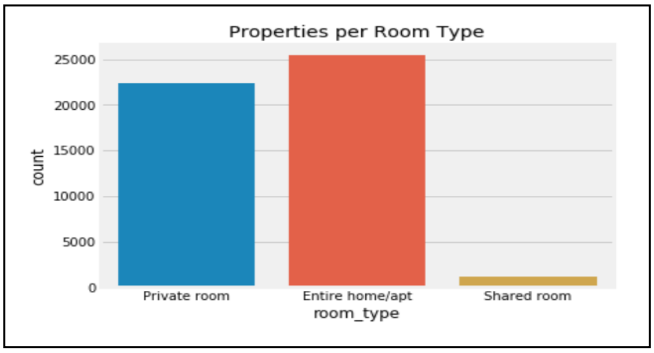


Fig 6

When it comes to booking a room of a particular type it was seen that majority of the people rent out entire apartment on Airbnb followed by the private rooms. The reason behind this could be that people go on holidays with their families and they require privacy which they could not get in shared rooms probably the reason why only few people opt for shared rooms.

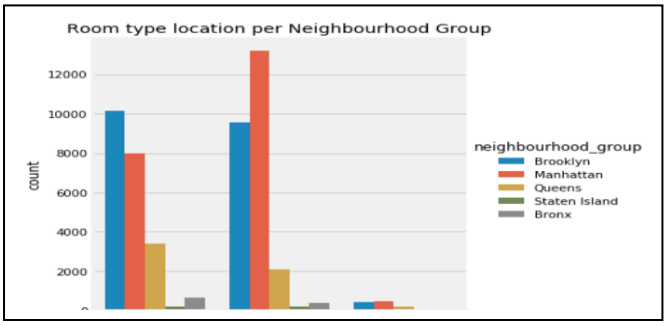


Fig 7

The next graph gave me the insight that the Manhattan has the max number of apartments of Entire home or Shared room category whereas the category Private room availablility were found to be maximum in Brooklyn.

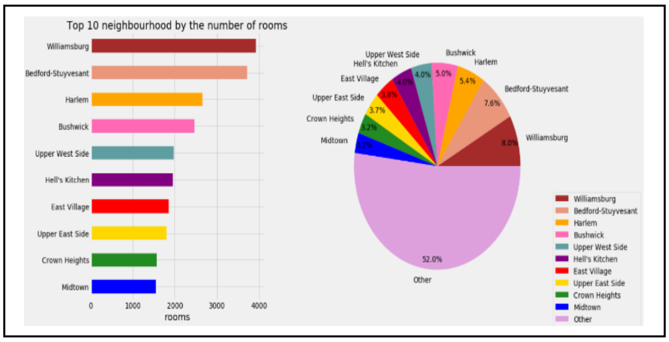


Fig 8

Fig 8 tells us in which neighborhoods was the maximum number of room present which we could see with the help of this graph. It gave me an insight that Hariem Williamburg, and Bedford-Stuyvesant have the highest number of rooms.

4) *Data Mining:*

Multiple Linear Regression: A Multiple Linear Regression is a supervised regression model. We still use the word "linear" here as in Linear Regression because here we assume that the dependent variable is directly proportional to the linear combination of independent variables. It's a statistical measure or a predictive analysis used to determine the change in response variable on the basis of multiple independent variables. More independent variables are added in the dataset to get better explanation of the outcome variable.

In New York City Airbnb dataset, I have used Multiple Regression Model for the prediction of outcome variable i.e. Price.

Gradient Boosting Regression: Gradient Boosting is an ensemble learning boosting which form a strong learner by combining the weak learners. It is based on an intuition that when the best possible next model is combined with the previous models, the error in the predictions get minimize. The main objective here is to minimize the error by training the estimators using the iterative approach. For the dataset New York City Airbnb, I have used this model because while evaluating the models the accuracy this model had was better than the others. It will be explained in detail in the Evaluation section.

Dataset-2 NYC Property Sales

This dataset is about every New York city building, building unit or apartment sold over a period of one year in the New York city property market. This dataset has been taken from Kaggle website. It contains 84549 rows and 20 columns. Some of the attributes in this dataset are sale price, location, sale date, address of the building units that were sold. Some of the attributes that were majorly focused on were:

Borough: It holds the names of the counties in which the property was located. It is in the numerical format but contains categorical data like (1) for Manhattan, (2) for Bronx, (3) for Brooklyn, (4) for Queens and (5) for Staten Island.

Building class at present and Building class at time of sale: It contains the information about the type of building at various interval of time.

1) *Data Preprocessing:*

First, I loaded the CSV file in the Jupyter notebook using Pandas library and read_csv function. After that using the head of Dataset, I could see the number of things that there were duplicated values in my dataset and some missing values were also there represented as a blank space which were need to be taken care of.

I started with removing the duplicates as it is an essential skill to get the counts of data which is accurate because counting the same thing multiple times can severely affect the efficiency of the model. To do this, first I took the sum of all the duplicated then by using drop_duplicates function I removed the duplicated values.

After this I decided to see the description of every column by using .info() function. There were a lot of variables which were not in appropriate datatype so to covert these pandas library function to_numeric was used.

Then I decided to check the columns which had null values by using. isnull() function. The insight provided by this function was that the dataset had three columns "Sale Price", "Land Square Feet","Gross Square Feet" which contained null values. I decided to made a graph after counting the sum of all missing values.

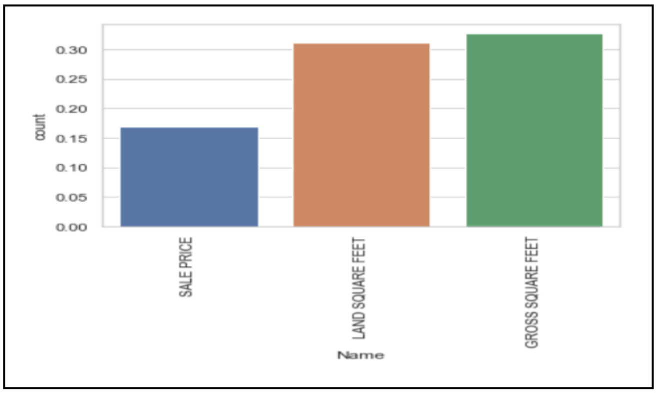


Fig 9

The missing values in columns "Land Square Feet" and "Gross Square Feet" were filled by taking the mean value of the column at individual level. There were a large number of missing values so I also thought to drop these missing values but when I check the MSE score of my model it was better when the mean values were taken.

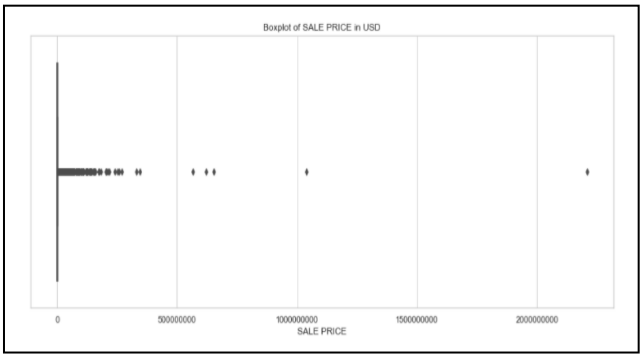


Fig 10

When I check for the target variable which is "Sale Price" by using the boxplot.

Here I could see an incredible lot of values close to $0 and the outliers could also be noticed which were mainly greater than 500000. I got one insight from the dataset description that the $0 property value was actually the deeds transfer between parties. Such as a property given by the parents to their child after retirement. Thus, there was no use of having this data to predict the sale value of the property. After removing these values, the graph looked like this.

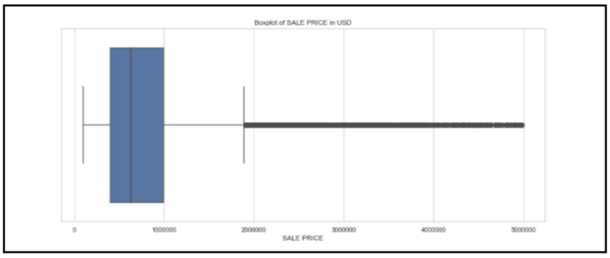


Fig 11

After plotting the distplot of the sale price, It was easily predicted that it was highly rightly skewed so to get better results log transformation was used. It was necessary to do this because the "Gross Square Feet" is decently correlated with the target variable so when the regression model will be applied it will make a significant difference.

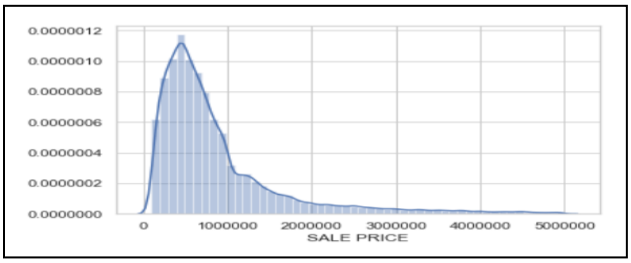


Fig 12 before log transformation

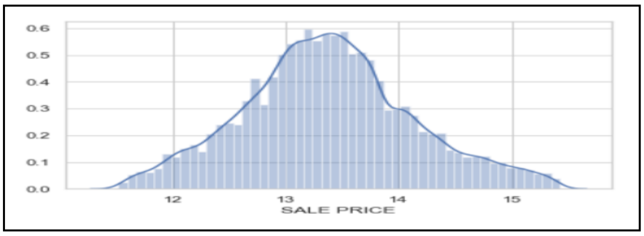


Fig 13 after log transformation

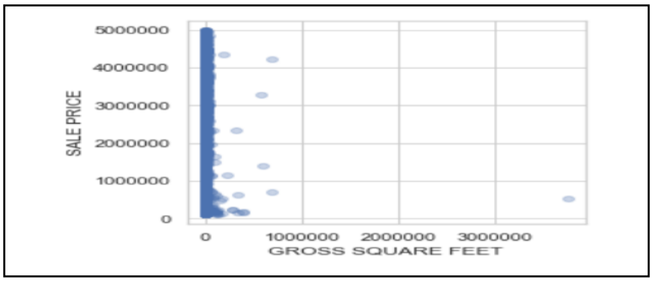


Fig 14 before removing values greater than 10000

After plotting the scatterplot for the independent variable i.e., "Gross Square Feet" and "Land Square Feet" it was clearly visible that they also have similar issue as "Sale Price". There was a number of values which could be removed in order to normalize the data. Thus, most of the values greater than 10000 in both cases were removed. Later on, the scatter plot looked like this.

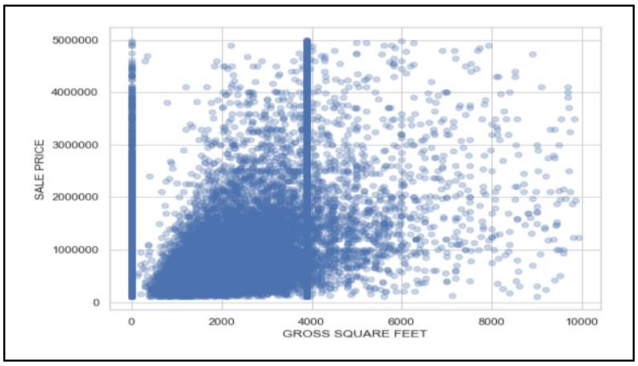


Fig 15

When we check the independent variable "Total Units". It is the sum of residential units and commercial units. It tells us for how many people the particular buildings were constructed. Total Units are grouped with "Sale Price" using groupby function and count function to check the number of particular unit counts. For 0 value of Total Units there were 15554 houses that were listed in Sale price and there was one outlier with 2261 value Total Units there was only one house. Thus, I decided to remove these values in order to improve the efficiency of my model.

One hot Encoding was done on "Borough", "Building Class Category", "Tax Class At Present", "Tax Class At Time Of Sale" in which categorical variables were converted into dummy variables and replaced with categorical variables.

2) *Data Transformation:*

The column "EASE-MENT" has no data so I decided to remove that from the dataframe. There was a column which was not named therefore no data could be gathered from that column so that column was removed as well.

Next, I check the "Apartment Number" column and found that it is around 50% empty, thus decided to remove that column. Then for the "Address" column at first, I thought I could use this column for geographical data analysis but to do so I also needed that the "ZIP CODE" column should have contained the same information which wasn't true in this case so I decided to remove the "Address" column.

3) *Data Visualization:*

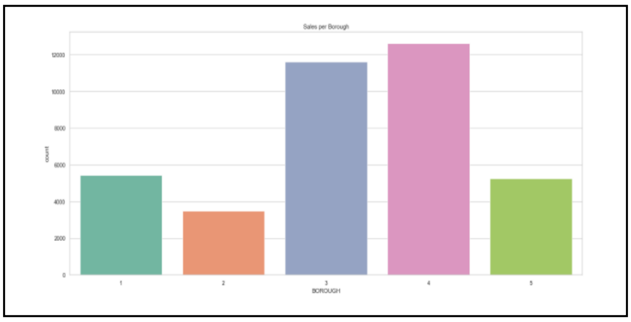


Fig 16

This graph is a countplot gave me an insight that in 3 i.e. Brooklyn maximum houses were sold.

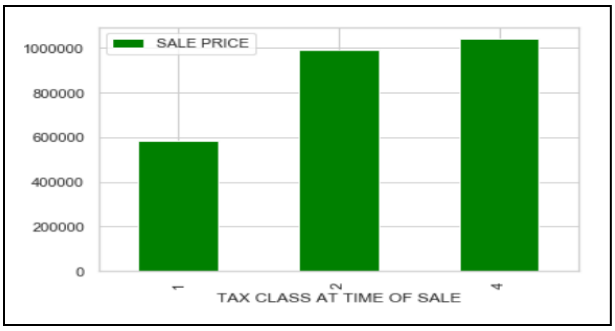


Fig 17

In the bar graph “Tax Class At Time Of Sale” vs Sale Price the “Tax Class At Time Of Sale” represent 4 tax classes (1,2,3 and 4). In class 1 it includes most of the residential property of upto three residential units (for example: offices, small number of family homes) Class 2 includes primarily residential property such as condominiums and cooperatives and Class 4 includes other properties like garage buildings, factories, warehouse, etc. Thus, this graph gave the visualization that for class 4 the sale prices were the highest.

BOROUGH: Manhattan: 1, Bronx: 2, Brooklyn: 3, Queens: 4, Staten Island: 5

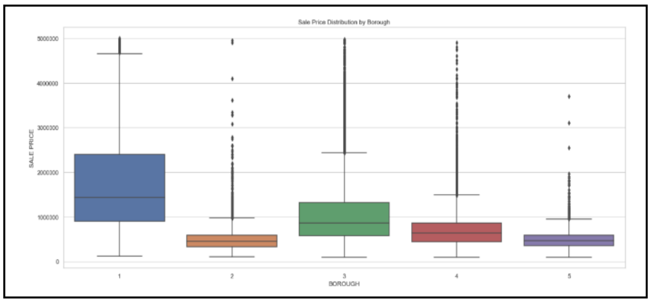


Fig 18

The above boxplot Sale Price vs Borough graph gave the visualization that the sales price distribution of borough 1 which is Manhattan was much wider than the other 4 boroughs.

4) *Data Mining:*

Random Forest Regression: Random forest is based on Ensemble Learning. This method takes various machine learning algorithms and combine them together to generate one significant machine learning algorithm which results in more robust predictions. Random Forest can be done in following steps:

a) First step is to choose any K random data points from the dataset.

b) Based on these K data points build a Decision tree.

c) Select the number of trees which you want to construct and repeat STEPS (a) and (b).

d) Let each one of your tests choose the category to which your new data point belongs to and the category which wins with the majority vote assign it the new data.

In NYC Property Sales dataset, I have used Random Forest Regression Model for the prediction of outcome variable i.e., Sale Price.

Decision Tree Regression: Besides Classification, Decision Tree can also be applied in regression problems. In Classification, most of the time decision tree is given more preference and it act as a classifier. A Decision Tree is used for visualization to constitute decision and decision making. It consists of two nodes named as decision node and leaf node. The one with the branches is decision node, they are kind off like a test which predicts the result in a ‘Yes or No’ and this test is performed on the attributes of the dataset.

In NYC Property Sales dataset, I have used Decision Tree Regression Model for the prediction of outcome variable i.e., Sale Price. These models will be explained in detail in the Evaluation section.

Dataset-3 Brooklyn Home Sales, 2003 to 2017

This dataset was found on the Kaggle website. It is a dataset which is concatenated from two datasets taken from two different sources. Some part of this is taken from the NYC Department of Finance (Housing Sales Data) from which my 2nd dataset is also extracted and the other part from the NYC Department of City Planning (MapPLUTO and PLUTO). This dataset is about the available sales properties in Brooklyn. This dataset contains 390883 rows and 111 columns. By looking at the dataset it can be clearly seen with the naked eyes that this dataset has a lot of columns which are duplicated and some which have no impact on the predictions or questions that I’m going to answer. The response variable is Sale_Price.

1) *Data Processing:*

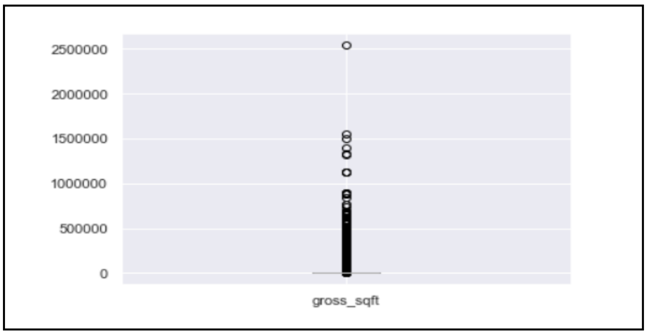
First, I loaded the CSV file in the Jupyter notebook using Pandas library and read_csv function. After that using the head of Dataset, I could see the number of things that a lot of columns which had missing values represented as NAN. My objective here would be to clean this data as much as possible. The first step is to check in how many variable what numbers of null values are present.

In the sale price column, the values with $0 were associated with a property transfer between parties such as parents have given their property to the children after retirement. Thus, there was no use of including that data in the prediction of sale price.

In many variables zeros and NANs are replaced with mode and median. Some of the categorical variables missing values were filled with 0. For majority of the continuous variables the missing values are replaced with mean or median.

The address variable is split into two columns: “street name” and “number”. The “street name” is converted to categorical variable and I decided to drop the “address” and “number” columns.

There are some missing values in “tax Class” which are replaced by the values of the column “tax class at sale”. Then this variable is converted to categorical. Same is done with the column “building class” by replacing its missing values with values of “building class at sale”. Some variables are categorized.



Fig

Outliers were removed using the z score value.

2) *Data Transformation:*

As discussed before this dataset has a lot of columns which are either do not have any impact on our predictions or are duplicated. Thus, 59 columns are removed based on various ground such as some columns were unnamed, had majority of values missing, duplicated data, not important for our prediction etc.

Next step is to check the variable that whether there are some variables highly correlated to each other and this was done by evaluating Variable Inflation Factor. The multicollinearity amount in the regression variables is measured by the VIF. If VIF is above 10 it is a cause of concern because that indicates that there is a high correlation between the variables. For this dataset all the variables with VIF greater than 6 are removed one by one in order to reduce the correlation to the minimum.

3) *Data Mining:*

Gradient Boosting Regression: Gradient Boosting is an ensemble learning boosting which form a strong learner by combining the weak learners. It is based on an intuition that when the best possible next model is combined with the previous models, the error in the predictions get minimize. The main objective here is to minimize the error by training the estimators using the iterative approach. For the dataset Brooklyn Home Sales, 2003 to 2017, I have used this model because while evaluating the models the accuracy this model had was better than the others. It will be explained in detail in the Evaluation section.

K-Nearest Neighbors(KNN): KNN is a supervised learning technique. It is a non-parametric method stating the fact that it does not make assumptions about the prescribed data set. It is a very intuitive algorithm which can be performed in following steps:

a) Its first step is to choose a number (K) of neighbors that you're going to have in your algorithm.
b) As per Euclidean distance (or any other distances as per your choice), for a new data point select the K nearest neighbors.
c) After selecting them observe each category by counting the number of data points.
d) The category where you counted the most number of neighbors assign your new data point to that category.
e) Your model is ready.

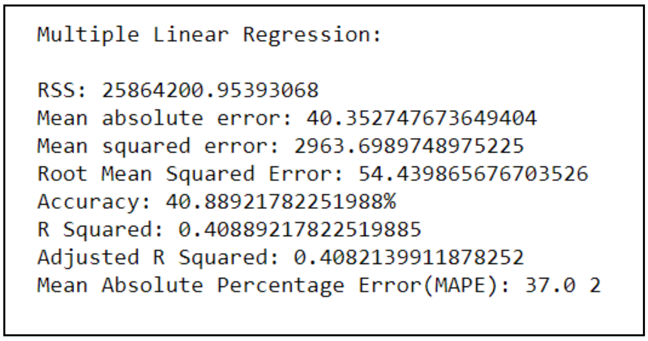
IV. EVALUATION

In order to do the evaluation first the dataset is needed to be split into two datasets: first one training dataset and the second one test dataset. The model is trained on the training dataset and then the trained model is applied on the fresh dataset i.e., test data in order to investigate how much accurate was our training model able to predict the outcome variable. Thus, 70% of dataset is assigned to training dataset and 30% to the test dataset.

Dataset 1:

1. Multiple Linear Regression:

The Multiple Linear Regression is executed and then the result obtained are visualized in the below figure.



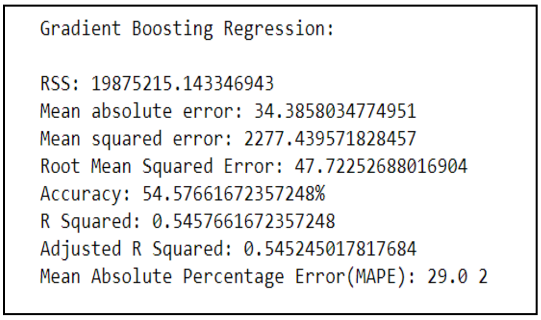
```
Multiple Linear Regression:

RSS: 25864200.95393068
Mean absolute error: 40.352747673649404
Mean squared error: 2963.6989748975225
Root Mean Squared Error: 54.439865676703526
Accuracy: 40.88921782251988%
R Squared: 0.40889217822519885
Adjusted R Squared: 0.4082139911878252
Mean Absolute Percentage Error(MAPE): 37.0 2
```

Fig 20

The value of R^2 is .40889 and adjusted R^2 is .40821. Then the values of RSS is 25864200.96, the value of MSE is 2963.69, the value of RMSE is 54.43, the value of MAE is 40.352 and at last I value of MAPE came out to be 37.02.

2. Gradient Boosting Regression:

In the second time this model was applied to the dataset and the following are the insight that I get after implementation of the model.



```
Gradient Boosting Regression:

RSS: 19875215.143346943
Mean absolute error: 34.3858034774951
Mean squared error: 2277.439571828457
Root Mean Squared Error: 47.72252688016904
Accuracy: 54.57661672357248%
R Squared: 0.5457661672357248
Adjusted R Squared: 0.545245017817684
Mean Absolute Percentage Error(MAPE): 29.0 2
```

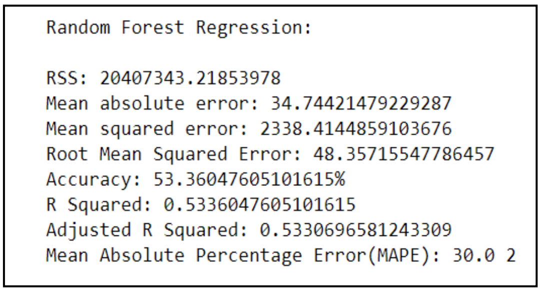
Fig 21

The value of R^2 is .5457 and adjusted R^2 is .5452. Then the values of RSS is 19875215.14, the value of MSE is 2277.43, the value of RMSE is 47.72, the value of MAE is 34.385 and at last I value of MAPE came out to be 29.02.

As it can be seen that this a better model between the above two but just to confirm it another model was applied.

3. Random Forest Regression:

This is the third extra model which was applied to check if Gradient Boosting algorithm is the better option for this dataset.



```
Random Forest Regression:

RSS: 20407343.21853978
Mean absolute error: 34.74421479229287
Mean squared error: 2338.4144859103676
Root Mean Squared Error: 48.35715547786457
Accuracy: 53.36047605101615%
R Squared: 0.5336047605101615
Adjusted R Squared: 0.5330696581243309
Mean Absolute Percentage Error(MAPE): 30.0 2
```

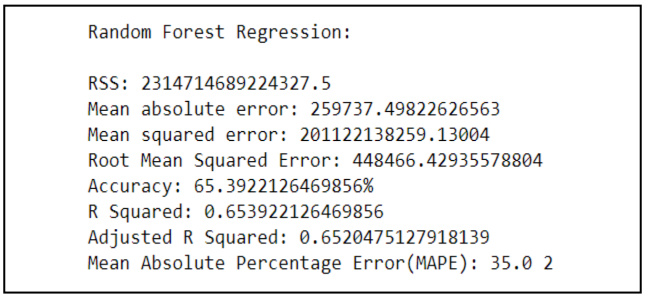
Fig 22

The R2 value for this model is .5336 which is less than the R2 of Gradient Boosting Model. Thus, it is proved that Gradient Boosting Regression is the better option for this dataset.

Dataset 2:

1. Random Forest Regression:

The first model that is applied on this dataset is Random Forest Regression. The results it provided are shown in the below figure.



```
Random Forest Regression:

RSS: 2314714689224327.5
Mean absolute error: 259737.49822626563
Mean squared error: 201122138259.13004
Root Mean Squared Error: 448466.42935578804
Accuracy: 65.3922126469856%
R Squared: 0.653922126469856
Adjusted R Squared: 0.6520475127918139
Mean Absolute Percentage Error(MAPE): 35.0 2
```

Fig 23

The value of R^2 is .6539 and adjusted R^2 is .6520. Then the value of RSS is 2314714689224327, the value of MSE is 201122138259, the value of RMSE is 448466.42, the value of MAE is 259737.49 and at last value of MAPE came out to be 35.02

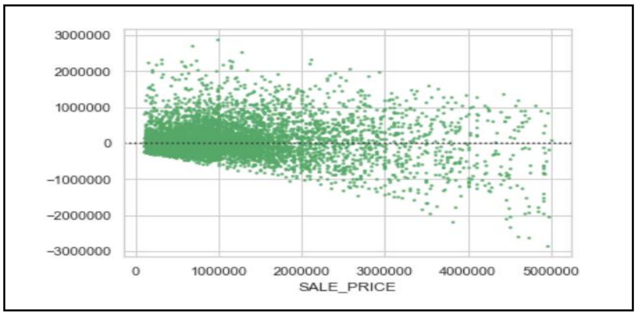
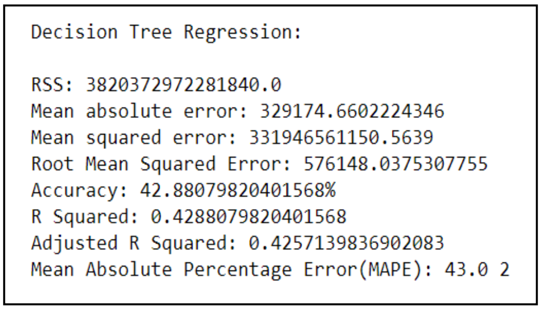


Fig 24 Random Forest Regression

This is a residual plot that do not show any pattern which means this model provided an optimum fit to data. This could not be said for the residual plot based on the Linear Regression as it showed a slight downward going pattern shown below.

2. Decision Tree Regression:

In the second time this model was applied to the dataset and the following are the insight that I get after implementation of the model.



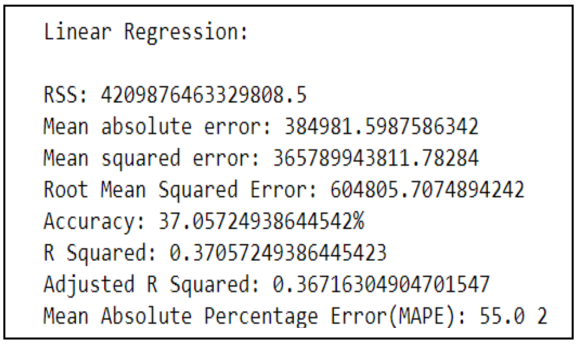
```
Decision Tree Regression:

RSS: 3820372972281840.0
Mean absolute error: 329174.6602224346
Mean squared error: 331946561150.5639
Root Mean Squared Error: 576148.0375307755
Accuracy: 42.88079820401568%
R Squared: 0.4288079820401568
Adjusted R Squared: 0.4257139836902083
Mean Absolute Percentage Error(MAPE): 43.0 2
```

Fig 25

The value of R^2 is .4288 and adjusted R^2 is .4257. Then the values of RSS is3820372972281840, the value of MSE is 331946561150, the value of RMSE is 576148.03, the value of MAE is 329174.66 and at last value of MAPE came out to be 43.02

3. Linear Regression:

This is the third model applied on the Dataset 2 to check if any other model exist which could be better than Random Forest for this dataset but because its R^2 value was less than Random Forest algorithm so we decided to remove this model.



```
Linear Regression:

RSS: 4209876463329808.5
Mean absolute error: 384981.5987586342
Mean squared error: 365789943811.78284
Root Mean Squared Error: 604805.7074894242
Accuracy: 37.05724938644542%
R Squared: 0.37057249386445423
Adjusted R Squared: 0.36716304904701547
Mean Absolute Percentage Error(MAPE): 55.0 2
```

Fig 26

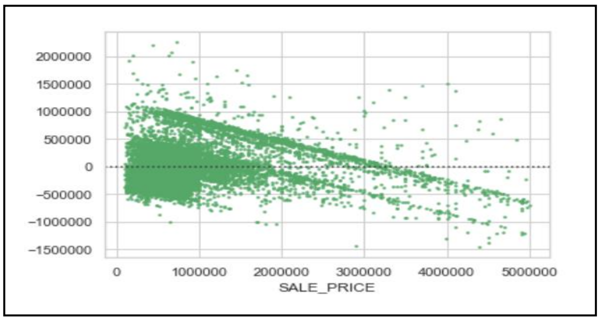


Fig 27 Linear Regression.

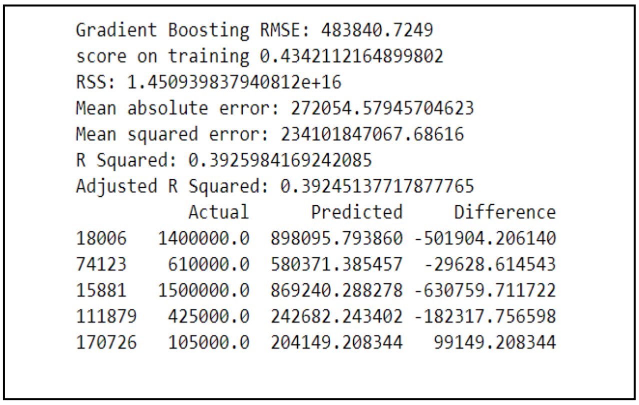
Dataset 3:

1. KNeighbors Regression:

After the implementation of this model on the third dataset. The R^2 value was 0.1014 which was really bad on its own. So, I decided to go for the another model. I even tried taking different values of K but it was still performing very poor.

2. Gradient Boosting Regression:

After applying this algorithm on the model, the R^2 value that I got was .14766. After looking at the result of both the model I wasn't satisfied with the results so I decided to do the Grid search on the better model i.e., Gradient Boosting Regressor to find out the best parameters. Feature-selection class from sklearn library was also used to improve the dataset.



```
Gradient Boosting RMSE: 483840.7249
score on training 0.4342112164899802
RSS: 1.450939837940812e+16
Mean absolute error: 272054.57945704623
Mean squared error: 234101847067.68616
R Squared: 0.3925984169242085
Adjusted R Squared: 0.39245137717877765
           Actual      Predicted     Difference
18006   1400000.0  898095.793860 -501904.206140
74123    610000.0  580371.385457  -29628.614543
15881   1500000.0  869240.288278 -630759.711722
111879   425000.0  242682.243402 -182317.756598
170726   105000.0  204149.208344   99149.208344
```

Fig 28

V. Conclusion

To Conclude in a general way, I have applied Linear Regression and Gradient Boosting Regression Model on the first dataset of Airbnb New York, in which Gradient Boosting came out to be the better one. In the second dataset of Sales Property in New York I used Random Forest regression and Decision Tree regression model, in which Random Forest came out to be the better one. Lastly, for the third dataset of Sales properties in Brooklyn, I used KNeighnour Regressor and Gradient Boosting regression model in which Gradient Boosting Regression model came out to be the better one. I have reached to a conclusion that no model is perfect and the performance of model vary on the basis of attributes. It is unprecedented what may change your model performance without calculations.

VI. Future work

To do things in future I would like to try out the method of two-step modeling while selecting the features more carefully, in which the training set is divided into K groups on the basis of price range and for individual groups a separate model is build. In order to predict labels will be classified at first and then the price regression will run. A particular thing which I couldn't do in this project was the prediction of sale property that whether it will be sold or not which could be done here by using classification. At last, It will be great if the model has transferability if applies on various cities.

References

[1] H. Yu and J. Wu, "Real estate price prediction with regression and classification," CS229 (Machine Learning) Final Project Reports, 2016

[2] Y. Ma, Z. Zhang, A. Ihler, and B. Pan, "Estimating warehouse rental price using machine learning techniques.," International Journal of Computers, Communications & Control, vol. 13, no. 2, 2018..

[3] D. Wang and J. L. Nicolau, "Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb. com," International Journal of Hospitality Management, vol. 62, pp. 120–131, 2017.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] https://arxiv.org/abs/1907.12665.

[6] Tarik Dogru, Makarand Mody and Courtney Suess, "Adding evidence to the debate: Quantifying Airbnb's disruptive impact on ten key hotel markets", *Tourism Management*, vol. 72, pp. 27-38, 2019.

[7] Dan Wang and Juan L. Nicolau, "Price determinants of sharing economy-based accommodation rental: A study of listings from 33 cities on Airbnb.com", *International Journal of Hospitality Management*, vol. 62, pp. 120-131, 2017.

[8] Jeroen Oskam and Albert Boswijk, "Airbnb: the future of networked hospitality businesses", *Journal of Tourism Futures*, 2016.

[9] Giovanni Quattrone et al., "Who benefits from the "Sharing." economy of Airbnb?", *Proceedings of the 25th international conference on the world wide web*, 2016.

[10] Arup Varma et al., "Airbnb: Exciting innovation or passing fad?", *Tourism Management Perspectives*, vol. 20, pp. 228-237, 2016.

[11] Katherine Goree, Battle of the beds: the economic impact of Airbnb on the hotel industry in Chicago and San Francisco.

[12] Makhmoor Bashir and Rajesh Verma, "Airbnb disruptive business model innovation: Assessing the impact on the hotel industry", *International Journal of Applied Business and Economic Research*, vol. 14.4, pp. 2595-2604, 2016.