Team Report – NYC Real Estate Analysis

Team Members:

Arun Sivakumar
Shalvika Mishra
Shreyas Putturaju

Executive Summary – NYC Real Estate Analysis

Project Goal:

The objective behind this project is to build a platform which can help users find recommendations of property based on the specifications provided by them. The project also helps in depicting the houses on a geographical map of the USA. It shows the houses labeled as clusters categories which in turn helps us examine the geographical difference between the clusters found.

Data Source and Schema:

This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period. This dataset contains the location, address, type, sale price, and sale date of building units sold. The dataset has 84549 instances, we will be using 80% as training and 20% as testing data. The dataset consists of 22 fields. The dataset depicts records of every building, apartments etc. sold in New York City.

Project Modules:

We broke down our project into below modules which are described in detail in the team project report.

1. Importing Libraries
2. Data Preprocessing
3. Data Visualization
4. Clustering
5. Cluster Analysis
6. Classification -Random Forest
7. Visualizations and Geographical Plots
8. Recommendation of the best match

Methods Used:

1) Unsupervised Learning - Clustering
2) Supervised Learning -Random Forest Classification
3) Visualization and Geographical Plotting – GeoPandas

Conclusion:

After clustering we could conclude that cluster 0 (red) has the highest values from approximately 0.9-1.0 (has 50,000 values), and other clusters are not well defined by sales price, but we hypothesize that there is a geographical significance in separating cluster 1 and cluster 2. We can also observe that cluster 0 has most values and we can consider them as the high-priced houses and cluster 1 has low prices and cluster 2 has medium price range. We conclude that k means gives us clusters with high prices as a separate cluster namely cluster 0 and the other clusters are separated by geographical significance and residential units. Residential units' significance is known from the important feature bar graph of Random Forest classifier.
The Recommend functionality helped us give the closest significant match based on Gross Square Feet and Cluster value. It gave us significant result i.e. SALES PRICE,ADDRESS and NEIGHBORHOOD corresponding to the given input i.e., Gross Square Feet and Cluster value{SALES PRICE,ADDRESS and NEIGHBORHOOD } were given as recommendations to the users.

Team Report

Data Source:

This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period. This dataset contains the location, address, type, sale price, and sale date of building units sold.

Data URL: https://www.kaggle.com/datasets/new-york-city/nyc-property-sales

Video Link : https://youtu.be/eLxSgbnx8ag

Data Schema and Size:

The dataset has 84549 instances, we will be using 80% as training and 20% as testing data. The dataset consists of 22 fields. The dataset depicts records of every building, apartments etc. sold in New York City.

Project Modules:

1. Importing Libraries
2. Data Preprocessing
3. Data Visualization
4. Clustering
5. Cluster Analysis
6. Classification -Random Forest
7. Visualizations and Geographical Plots
8. Recommendation of the best match

Objective:

The objective behind this project is to build a platform which can help users find recommendations of property based on the specifications provided by them. The project also helps in depicting the houses on a geographical map of the USA. It shows the houses labeled as clusters categories which in turn helps us examine the geographical difference between the clusters found.

Data Pre-processing:

Several data cleansing steps were required before moving into data exploration and machine learning tasks:

1. Deleting Unnecessary Columns and Cleansing

   We dropped the columns which were empty and like an iterator for example EASE-MENT, Unnamed: 0, SALE DATE and were not relevant to our analysis.
   We deleted the duplicates and checked that it worked by finding the shape of the dataset and description of every column.

2. Calculating New feature "Building Age"

   We did some analysis and thought it would be better to convert YEAR BUILT to the BUILDING AGE i.e.,2017 (as data is collected for 2017) - YEAR BUILT. This precisely given how old the building is, we further dropped YEAR BUILT.

3. Mean Imputation

For Land Square Feet and Gross Square Feet we replaced the null values with mean values.

4. Specifying categorical and numerical variables

We also needed to change the format of the columns to reflect the proper data types (all datatypes initially were of type 'object').

5. Converting into appropriate data type for Tax class

We sliced and converted the Tax class to numerical variables i.e. we changed 1A, 1B, 1C, 1D into 1 . We made this conversion based on an article mentioned in {https://www1.nyc.gov/assets/finance/downloads/pdf/brochures/class_1_guide.pdf}

6. Taking Significant Values

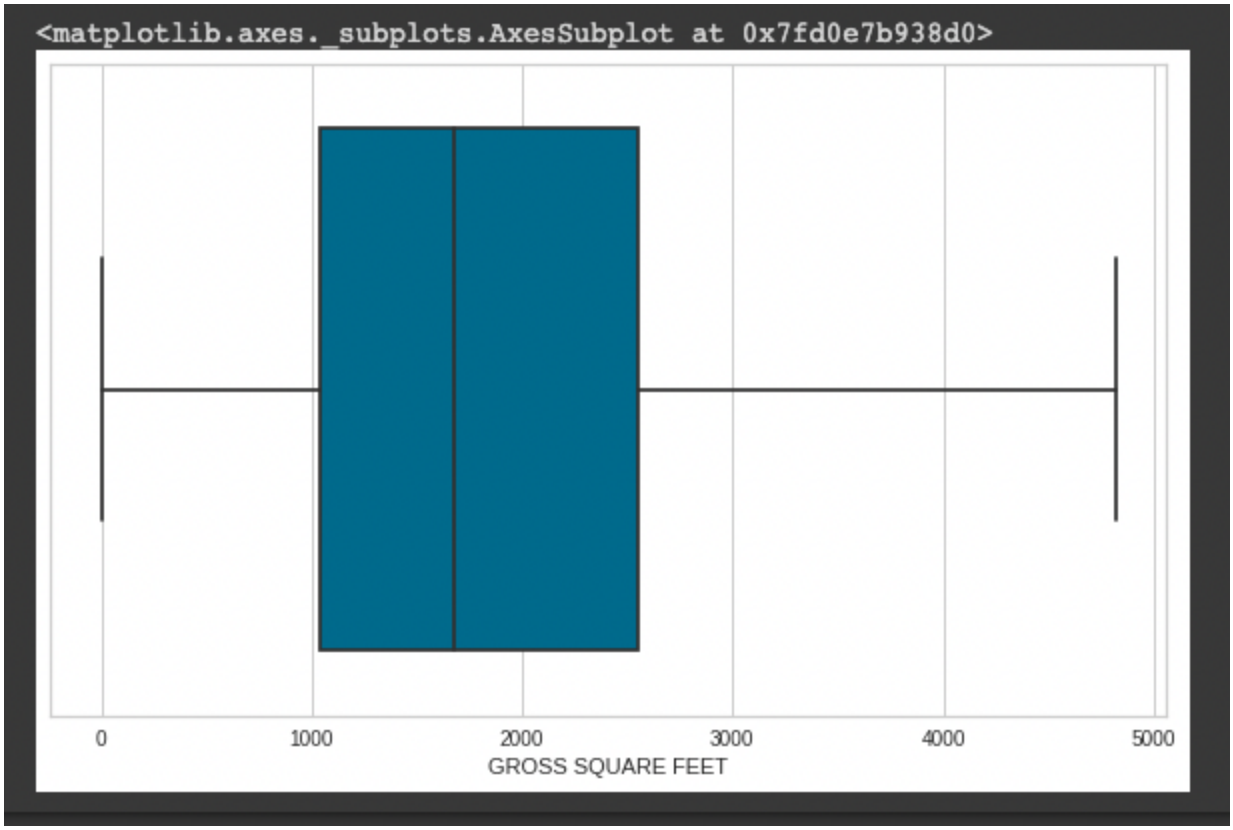
We observed that there are a lot of Sales Price values which were insignificant hence in the training data we only selected values of Sales Price which were greater than 50000,because the minimum house causes approximately 50K in the New York state.

7. L2 Normalization

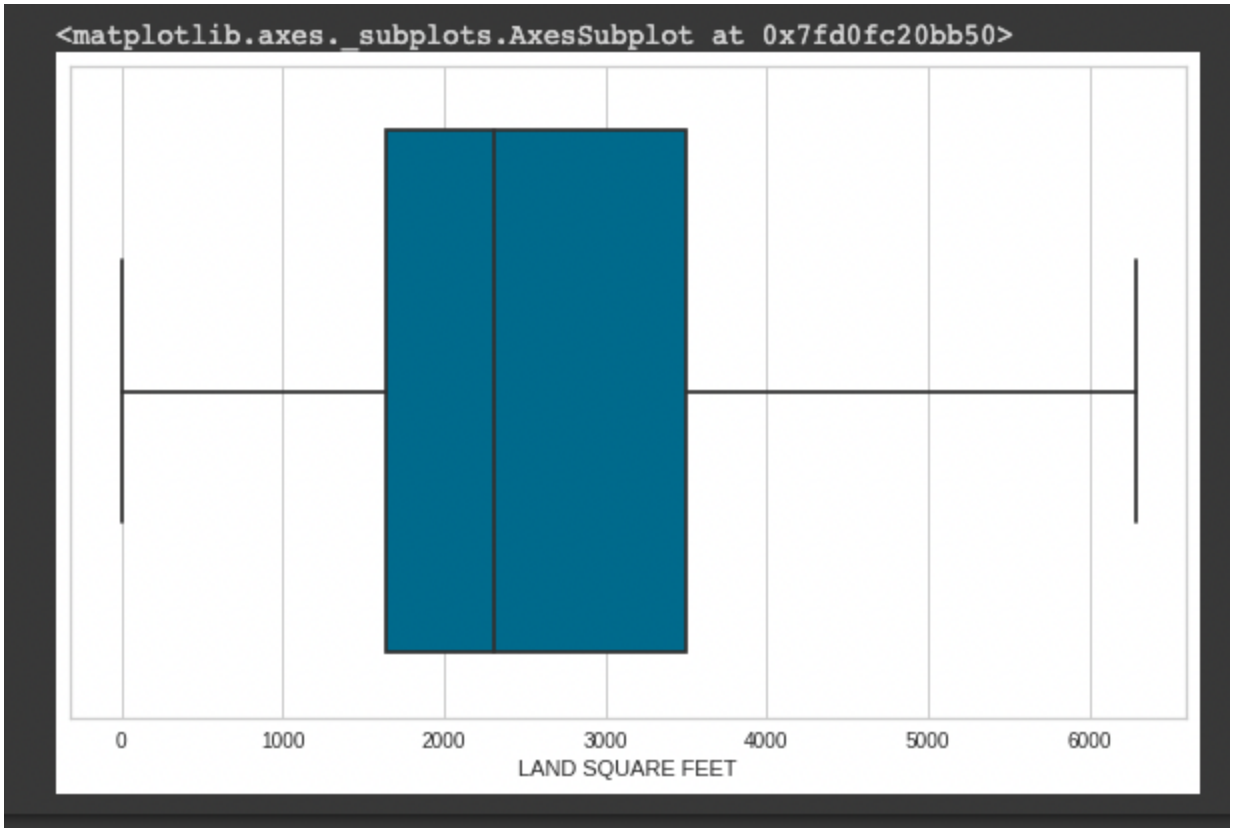
We also normalized the data using L2 normalization using sklearn

Data Visualization before Normalization:

We plotted box plots for all the numerical columns i.e., GROSS SQUARE FEET, LAND SQUARE FEET, TOTAL UNITS, RESIDENTIAL UNITS, and we further analyzed GROSS SQUARE FEET vs SALE PRICE, LAND SQUARE FEET vs SALE PRICE. Further we also did mean imputation for GROSS SQUARE FEET, LAND SQUARE FEET.

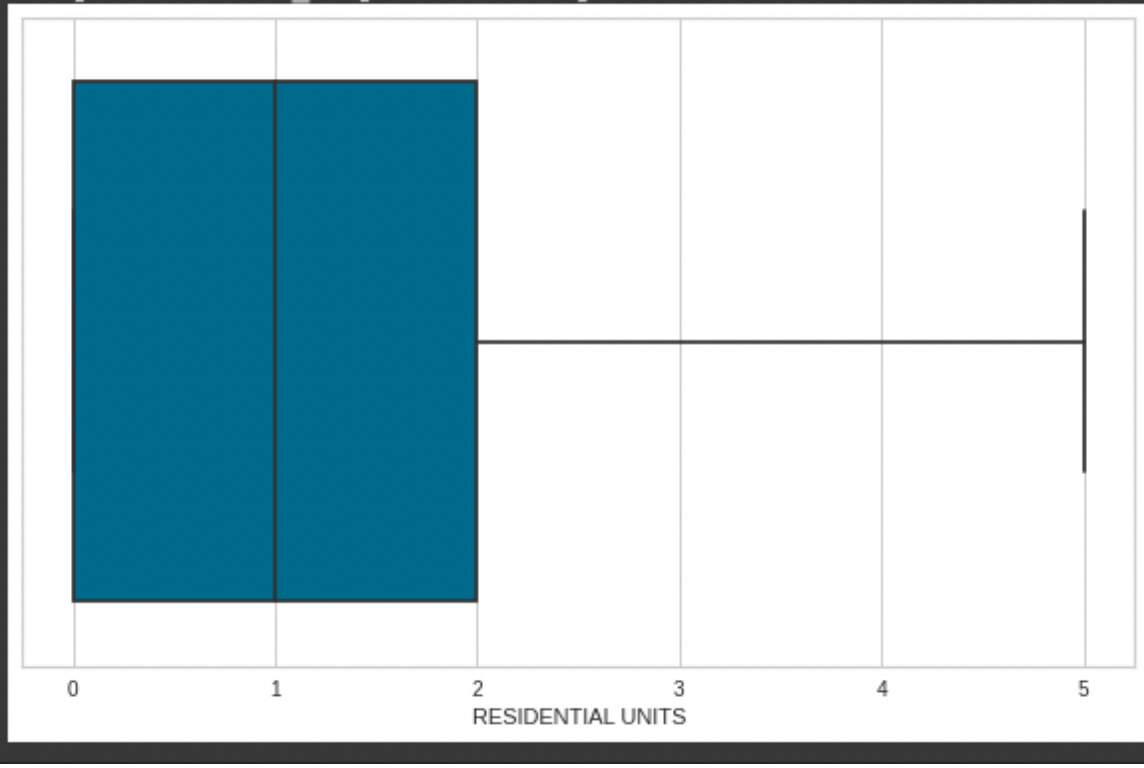


From the above graph we can see the median of GROSS SQUARE FEET between 1000 -2000.

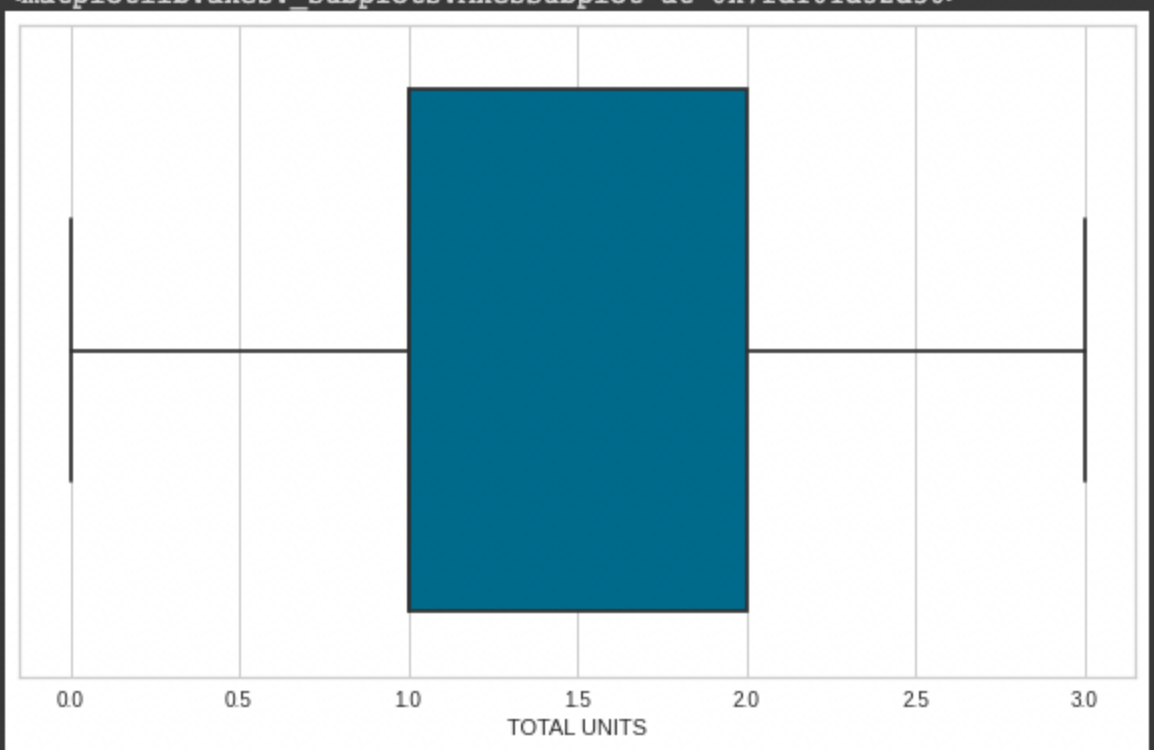


From the above graph we can see the median of LAND SQUARE FEET between 2000 -3000.

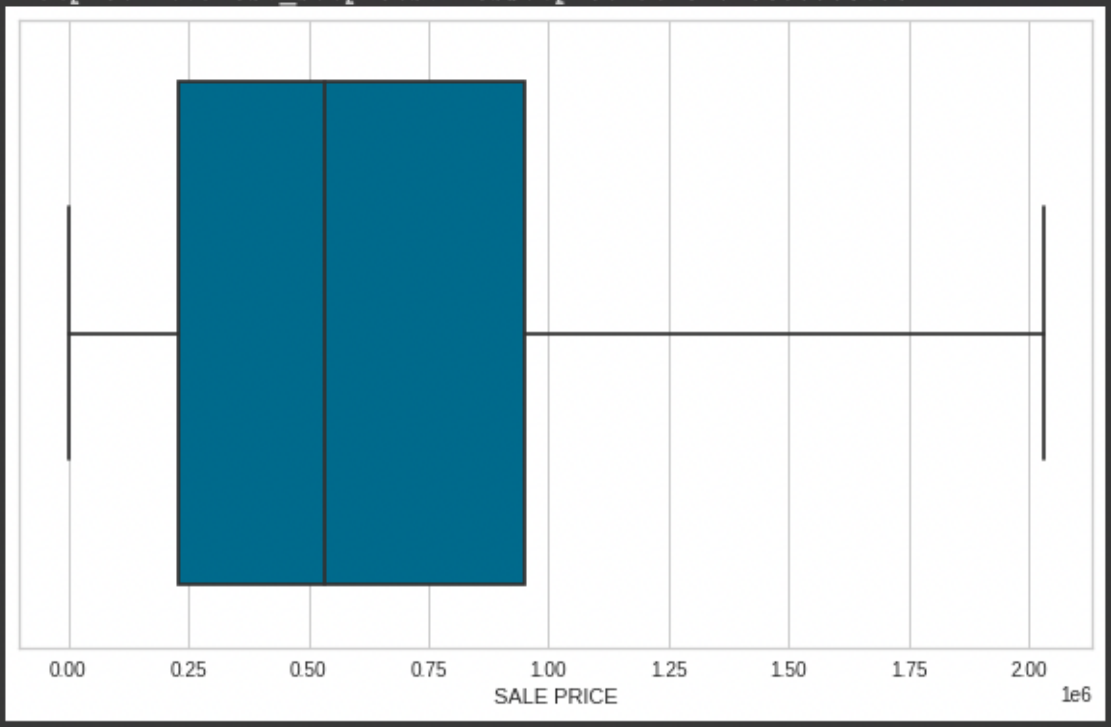
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd0fc21ebd0>
```



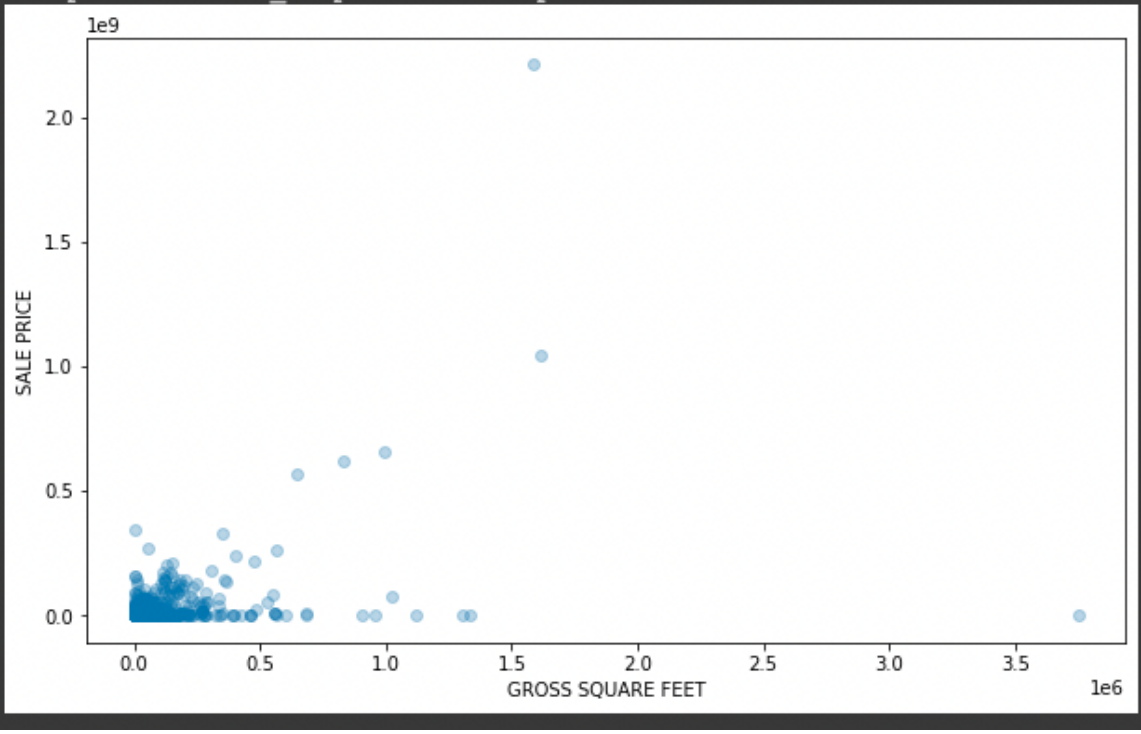

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd101d32d50>
```

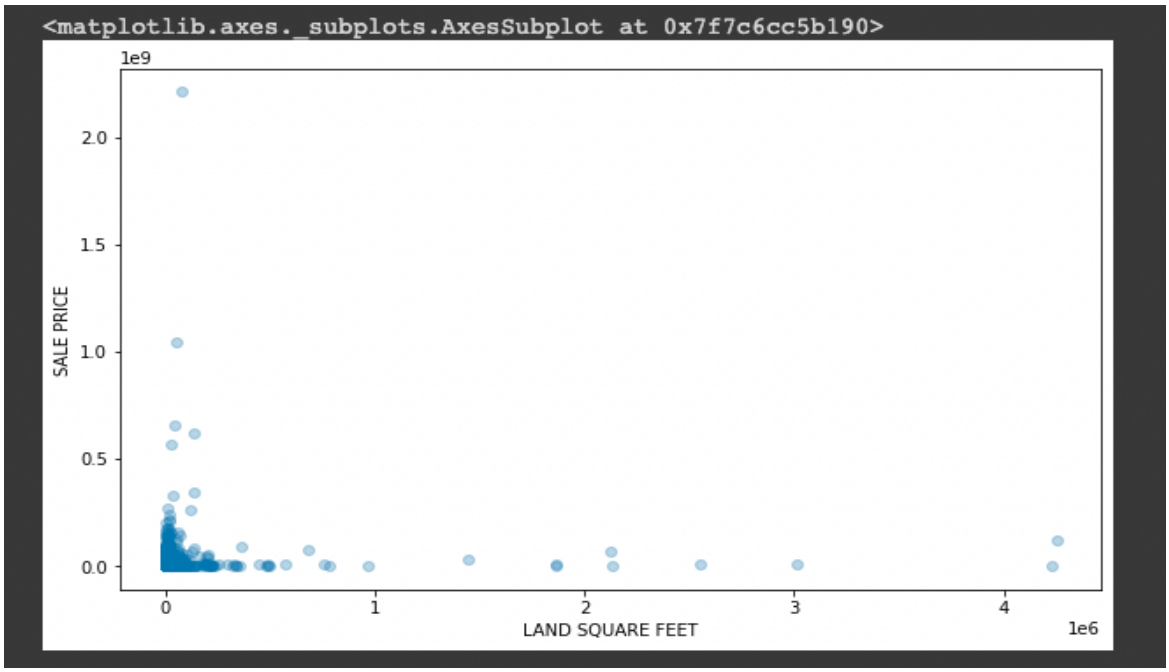


```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd0ddd86bd0>
```

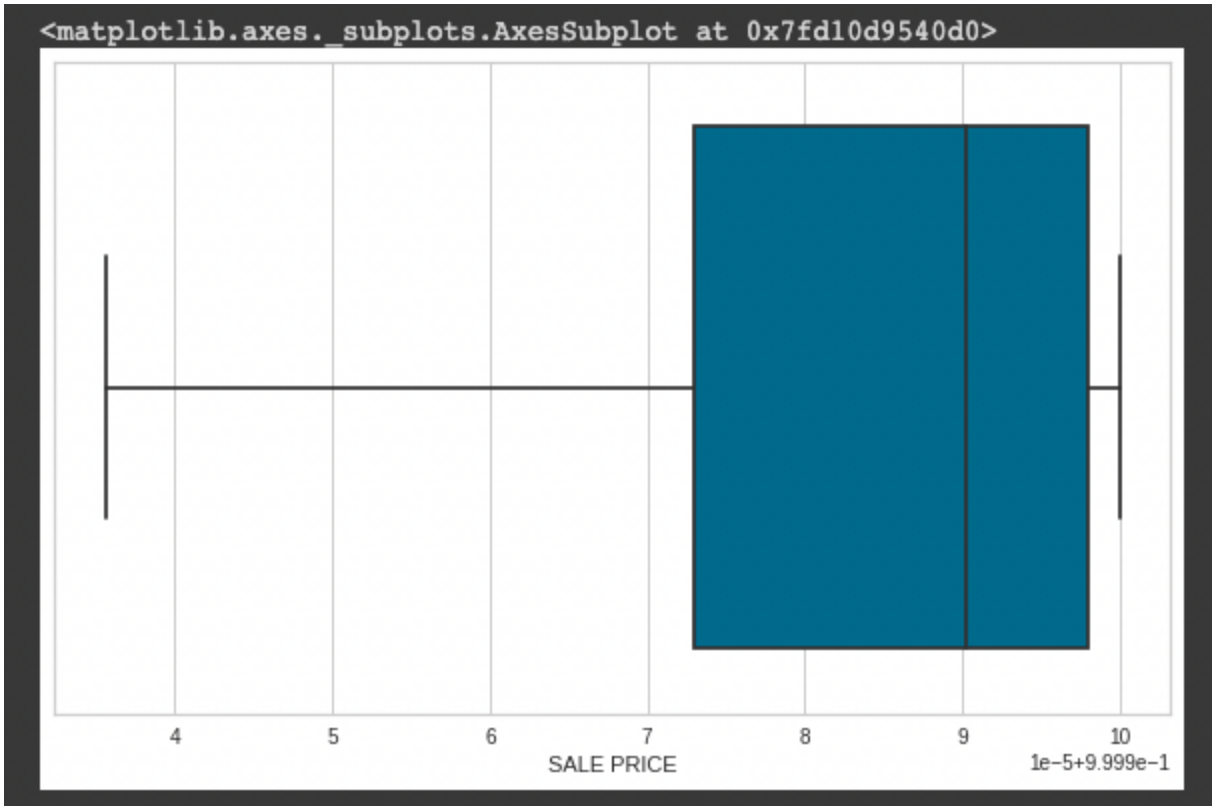



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7c6cc7b590>
```

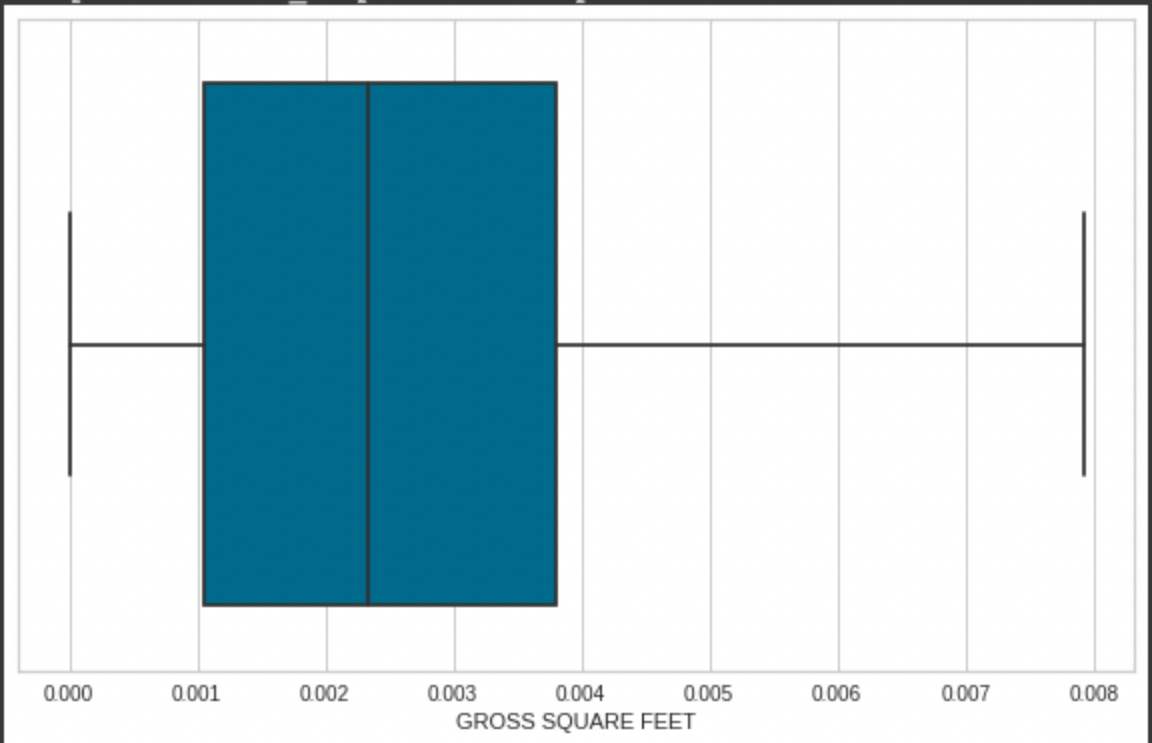




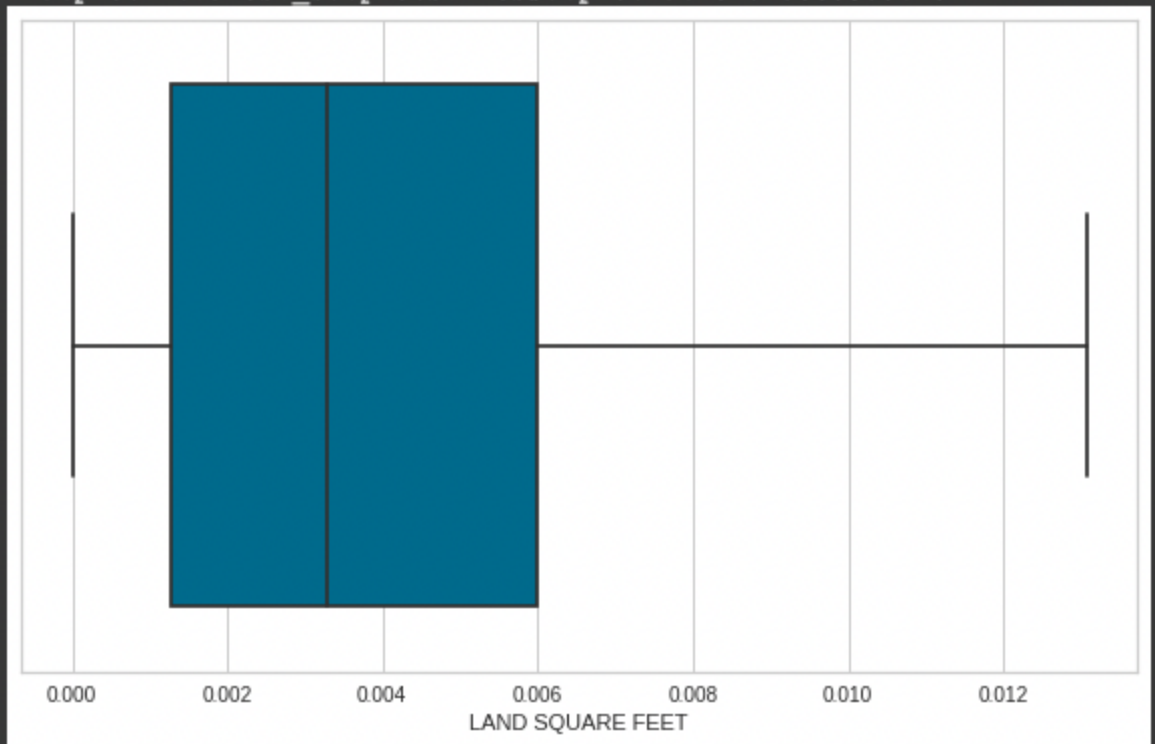
Data Visualization after Normalization:



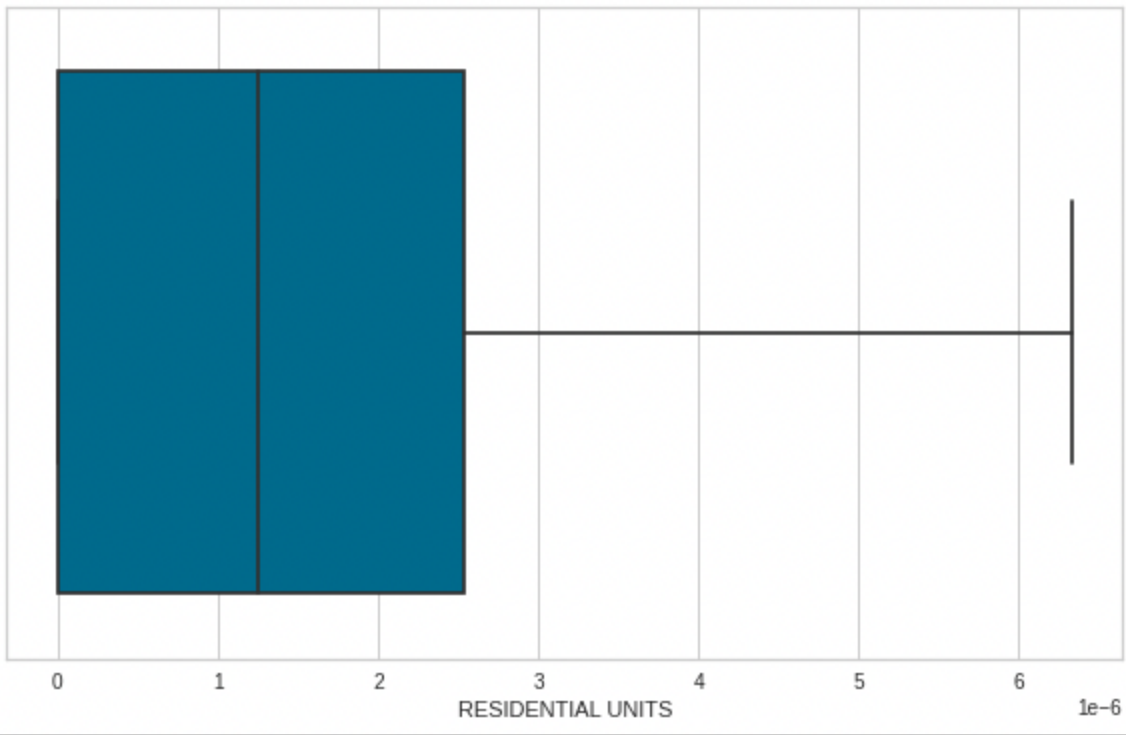
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd0d9dc4410>
```

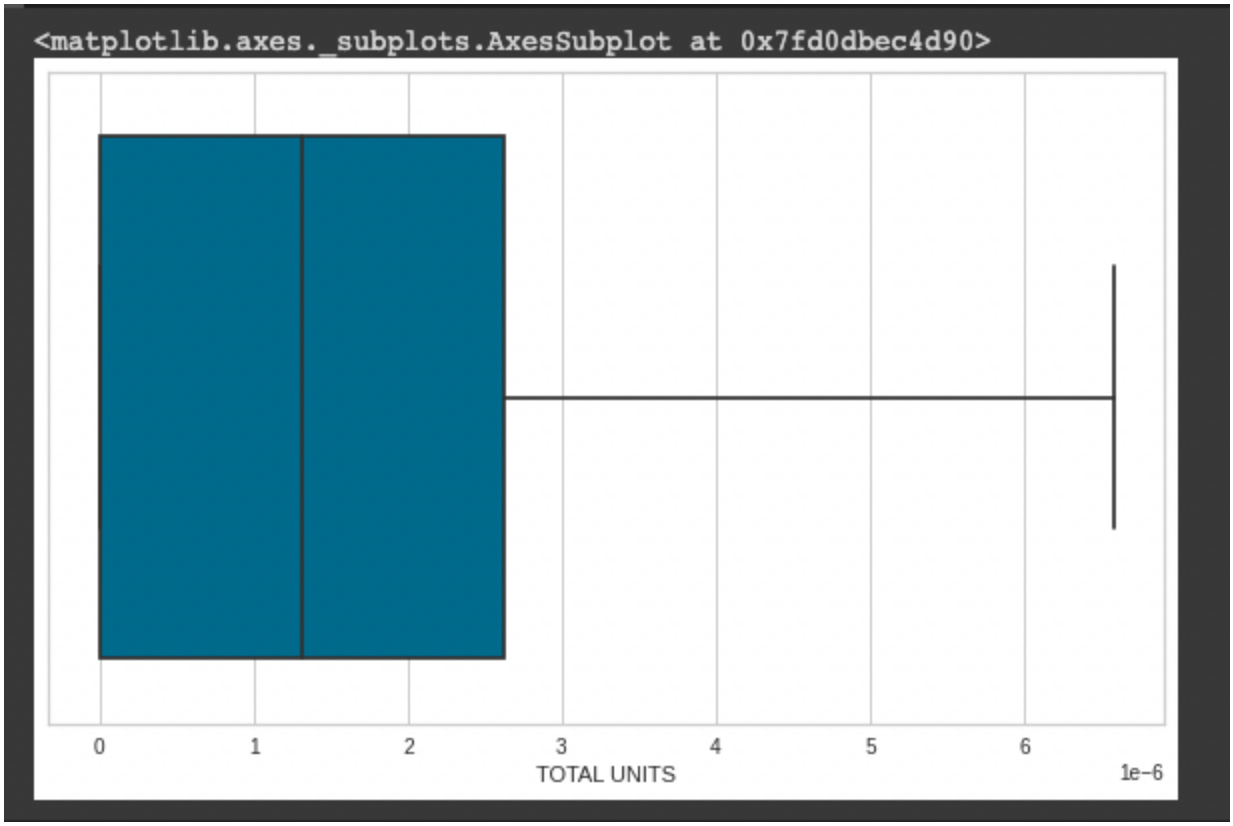


```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd0e3573b10>
```



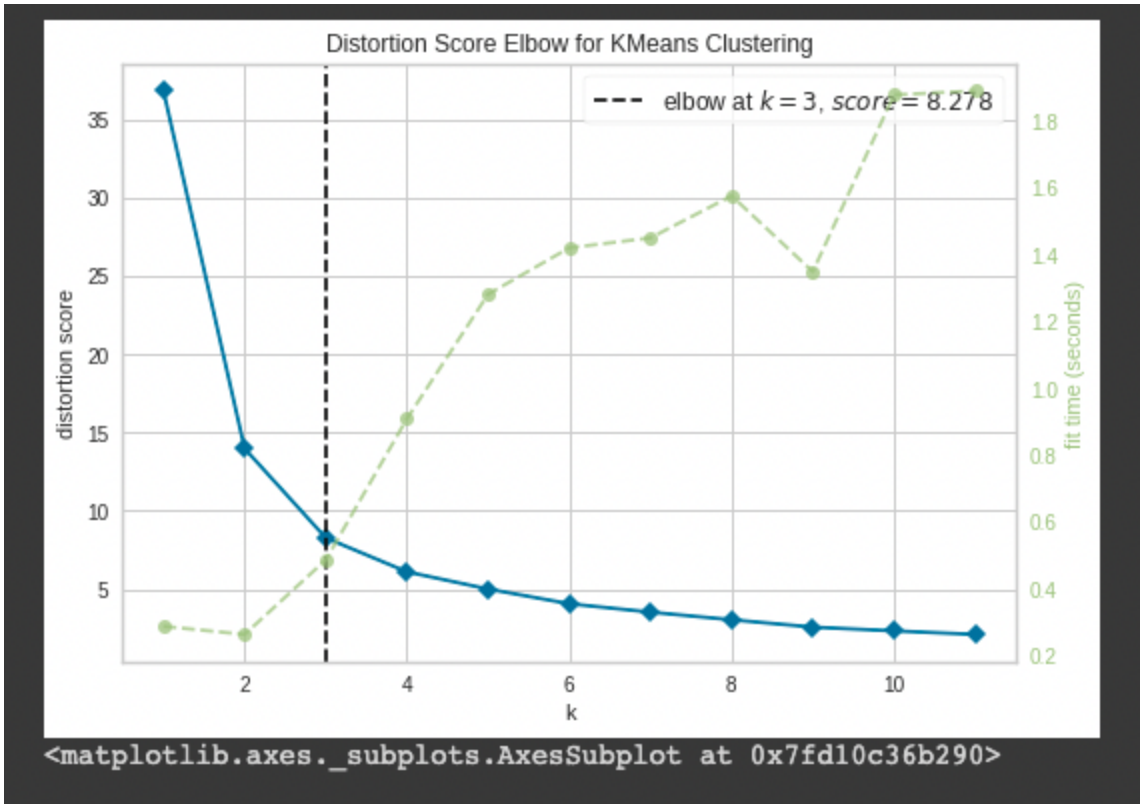

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fd0de47c950>
```





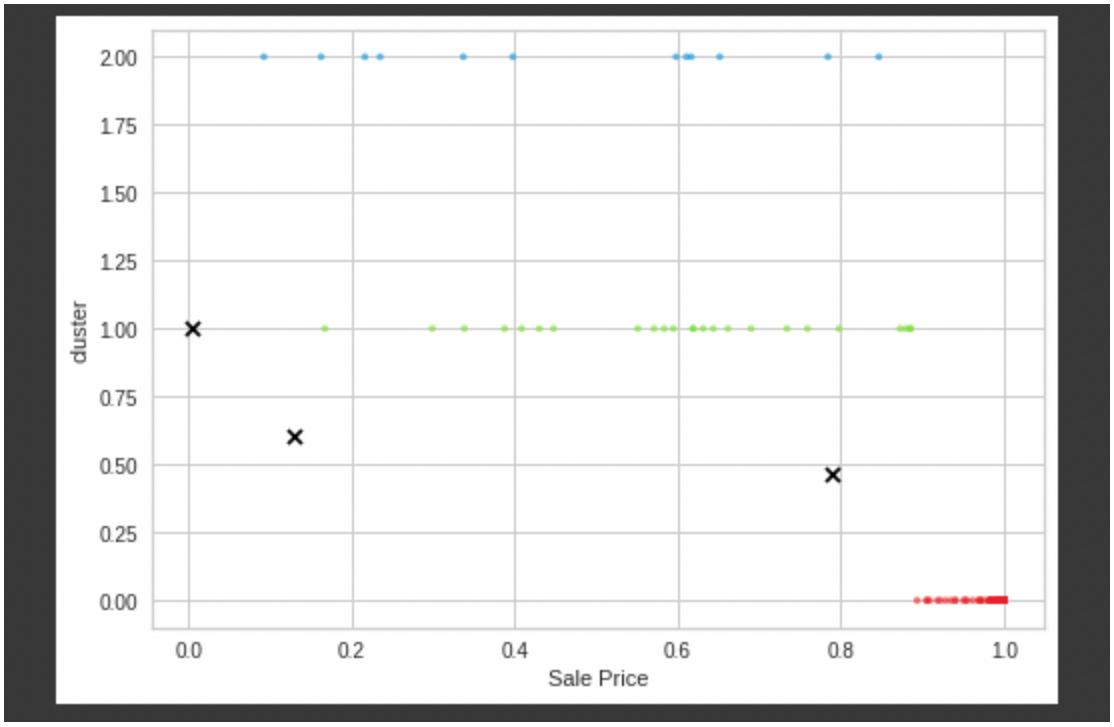
Clustering:

After Data preprocessing we performed K-means clustering, the k-means algorithm tries to minimize distortion, which is defined as the sum of the squared distances between each observation vector and its dominating centroid. We further used the elbow method to determine the optimal K value.



From the above diagram the elbow is at K= 3 hence we did clustering taking K value as 3.
We got 3 distinct clusters and as depicted below i.e., SALE PRICE vs CLUSTER:

We can see from the below figure that we got 1 cluster separate considering sale price aspect which is cluster 0 (red) and it is discussed further below.



labels= 0-red , 1-green ,2-blue

In the above diagram we can clearly see that the cluster with red color has the highest SALE PRICE .We further went on to finding in depth cluster analysis :

Cluster Information: {No of value in each cluster}

```
Counter({0: 57054, 1: 24, 2: 12})
```

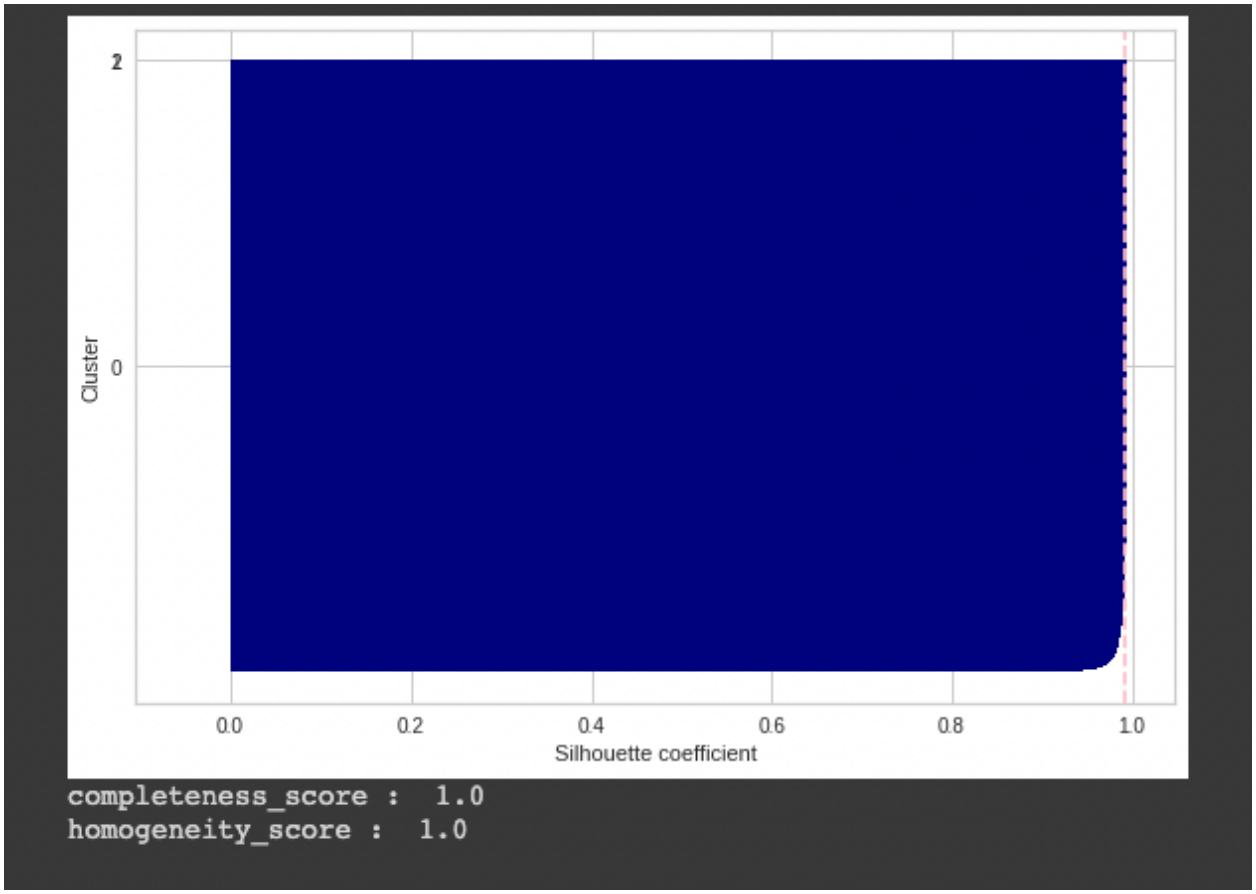
```
[257]  Centroids = scaled_train.groupby(["cluster"]).mean()[["GROSS SQUARE FEET","SALE PRICE"]]
```

```
[258]  print(Centroids)
```

```
         GROSS SQUARE FEET  SALE PRICE
cluster
0                 0.003058    0.999940
1                 0.738822    0.602389
2                 0.211300    0.462427
```

we can see the cluster centroids in the above fig and we can see that cluster 0 has high prices houses and 1 medium and 2 even lower

We can see that cluster 0 has most values and we can consider them as the high priced houses and other clusters as low prices

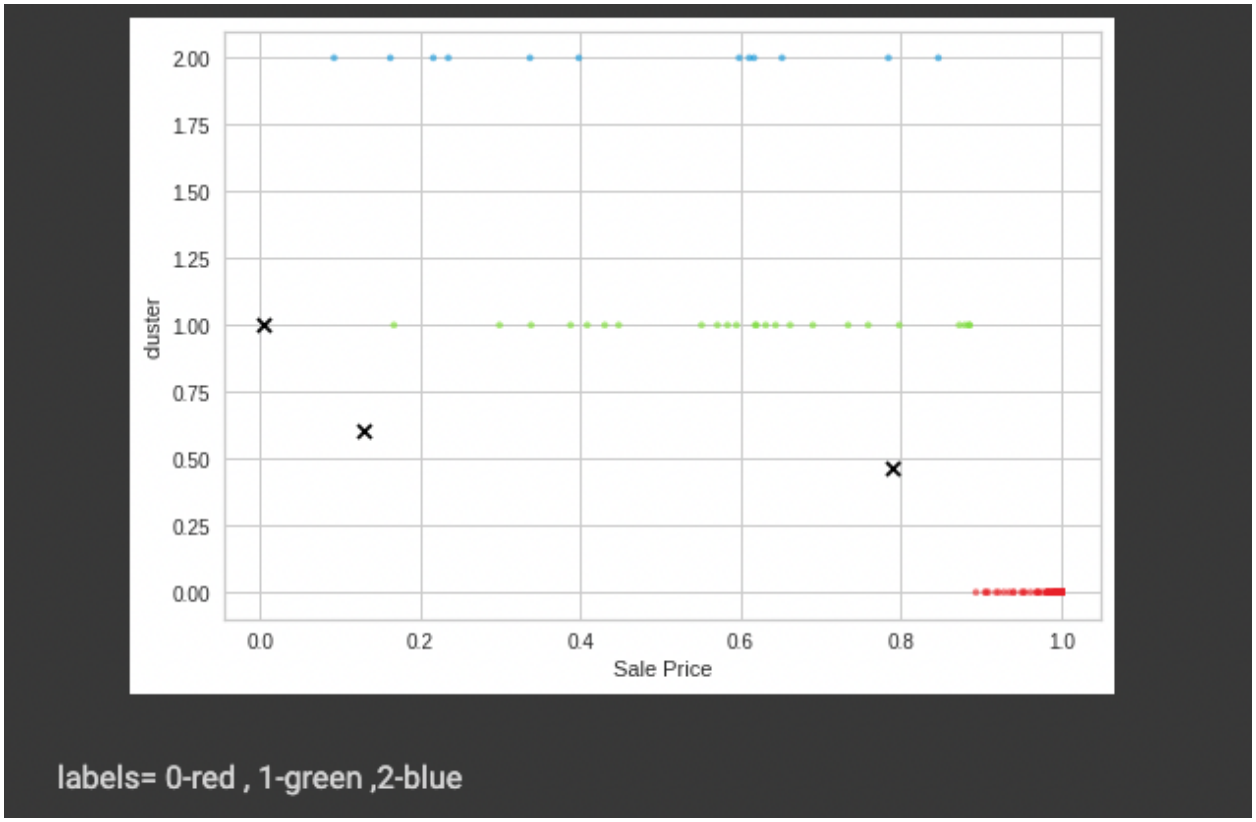


Hence because of clustering we could find out which set of records would fit in which cluster.

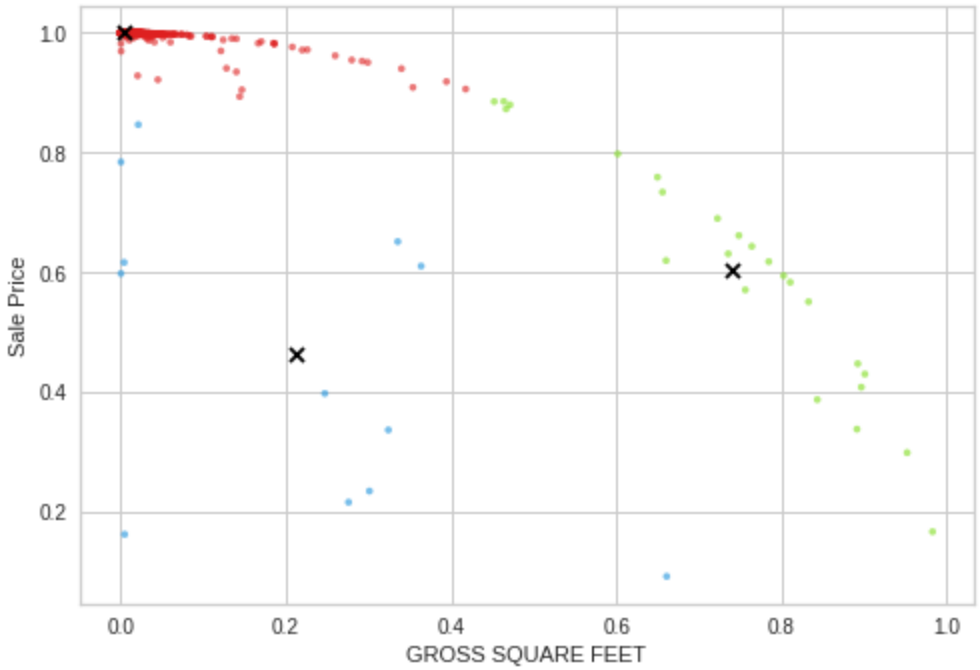
Cluster Analysis:

We further went on to find the cluster centroids. And we could clearly observe that for cluster 2 and cluster 3 the centroids were very close.

We further observed that cluster 0 (red) has the highest values from approximately 0.9-1.0 (almost 50,000 values), and other clusters are not well defined by sales price, but we hypothesize that there is a geographical significance in separating cluster 1 and cluster 2.We can also observe that cluster 0 has most values and we can consider them as the high-priced houses and cluster 1 has low prices and cluster 2 has medium price range



We evaluated clustering results by calculating the silhouette score and it was approximately 95%.



We visualized the cluster and their respective centroids with gross square feet and sales price as we find that gross square feet is an important feature for classifying the clusters using random forest classifier in later part of the notebook

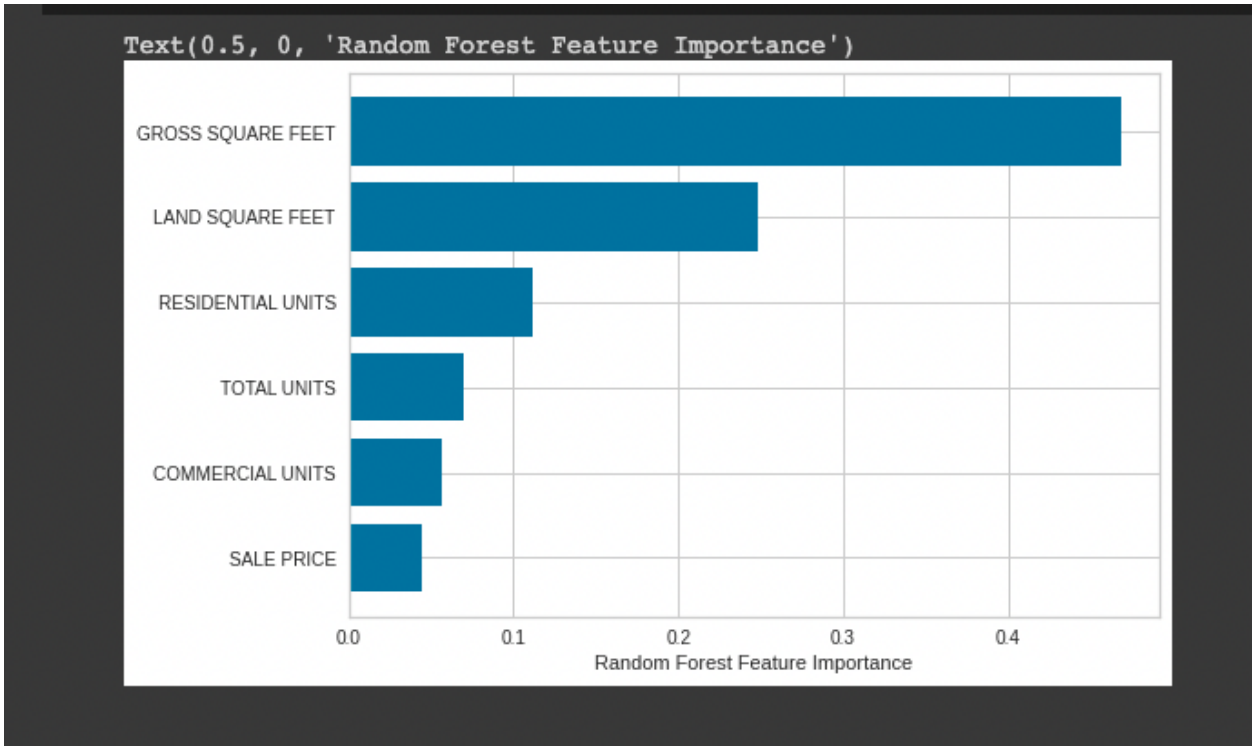
Classification:

Random Forest Classification

Next step was classification. We took cluster value as the target for Random forest classifier. We deleted Sales Price from the dataset.

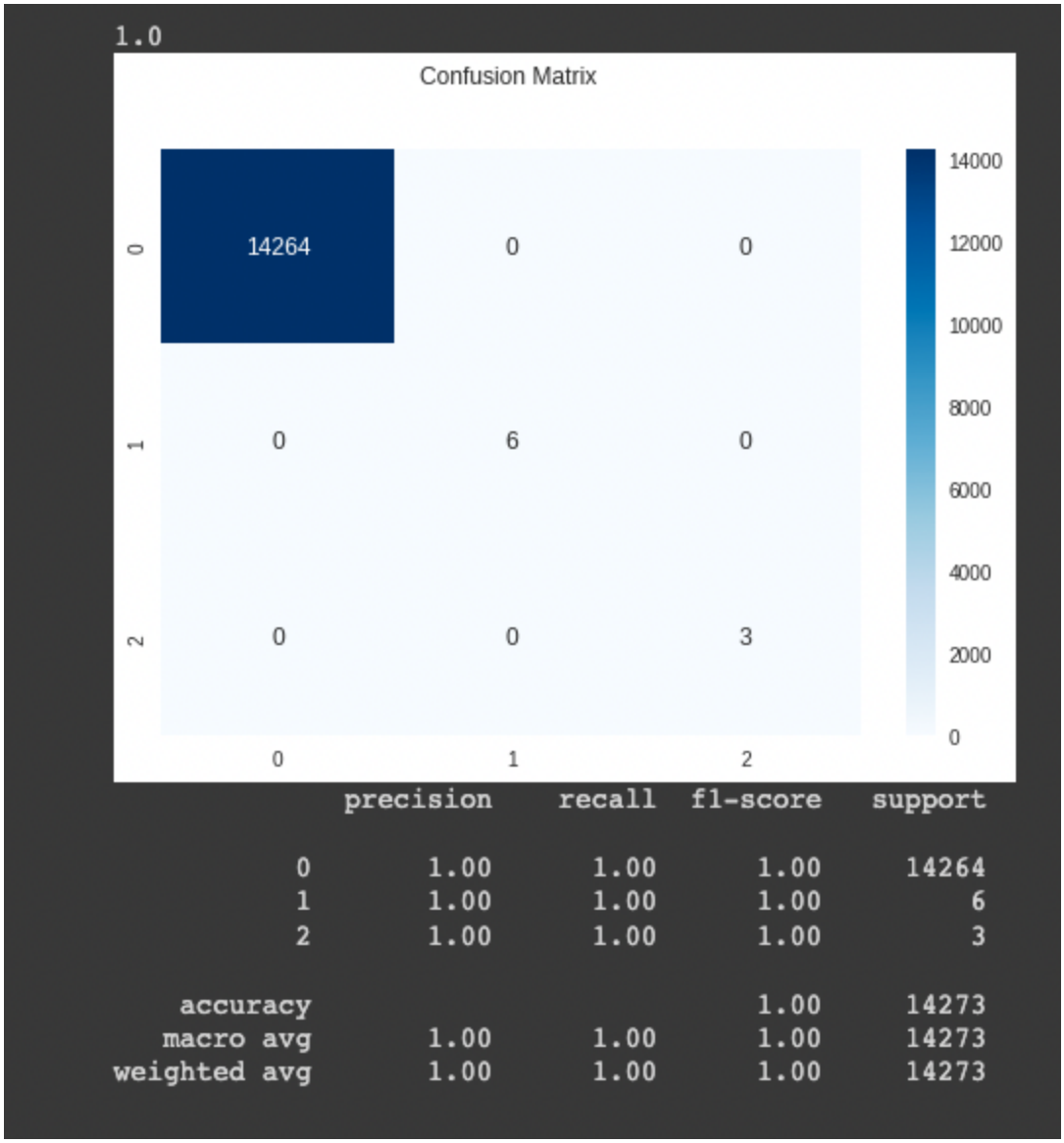
The deletion of sales price is done so that we can predict values that don't have sales price. we later consider them as user queries

We removed the target(cluster) to create a test set .We went on to splitting data into training and test sets with 80:20 proportions . And instantiated the model with 100 decision trees.We also performed feature importance using random forest and the below image shows the important features:



As shown in the figure above we can see that GROSS SQUARE FEET and LAND SQUARE FEET had great significance in classifying the values thus we use gross square feet for recommendation and we also visualized sales price against gross square feet for this reason.

We further predicted test data and plotted the confusion matrix and then we plotted the ROC-AUC curve and calculated the total area under the curve :

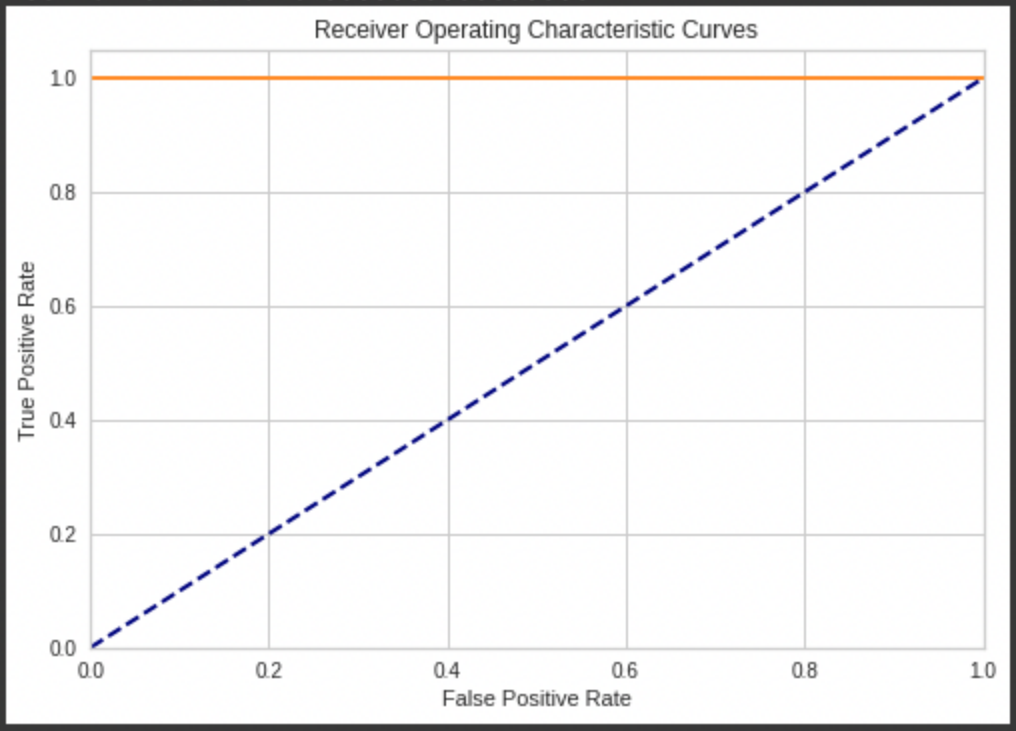


	precision	recall	f1-score	support
0	1.00	1.00	1.00	14264
1	1.00	1.00	1.00	6
2	1.00	1.00	1.00	3
accuracy			1.00	14273
macro avg	1.00	1.00	1.00	14273
weighted avg	1.00	1.00	1.00	14273

Most of the class values were predicted correctly, as none of them were misclassified as other classes.

The classification report shows us F1 Score of 1 shows us that there are no misclassifications.

```
AUC for Class 1: 1.0
AUC for Class 2: 1.0
AUC for Class 3: 0.9999999999999999
```



AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0.0; one whose predictions are 100% correct has an AUC of 1.0.As seen in above image we obtained 100% accurate predictions for Class 1 , 100 % accurate prediction for Class 2 and 99% accurate predictions for Class 3.

Visualization and Geographical Plots:

We further went ahead and plotted our cluster and sales price values on the geographical map of the USA. Based on the BOROUGH field. The BOROUGH field has 5 values which are Manhattan, Bronx, Brooklyn, Queens, Staten Island. We leveraged the geopy, folium library fully for these plots. And the plotting helped us understand why few clusters had significant values as many houses were located nearby or in the same area.

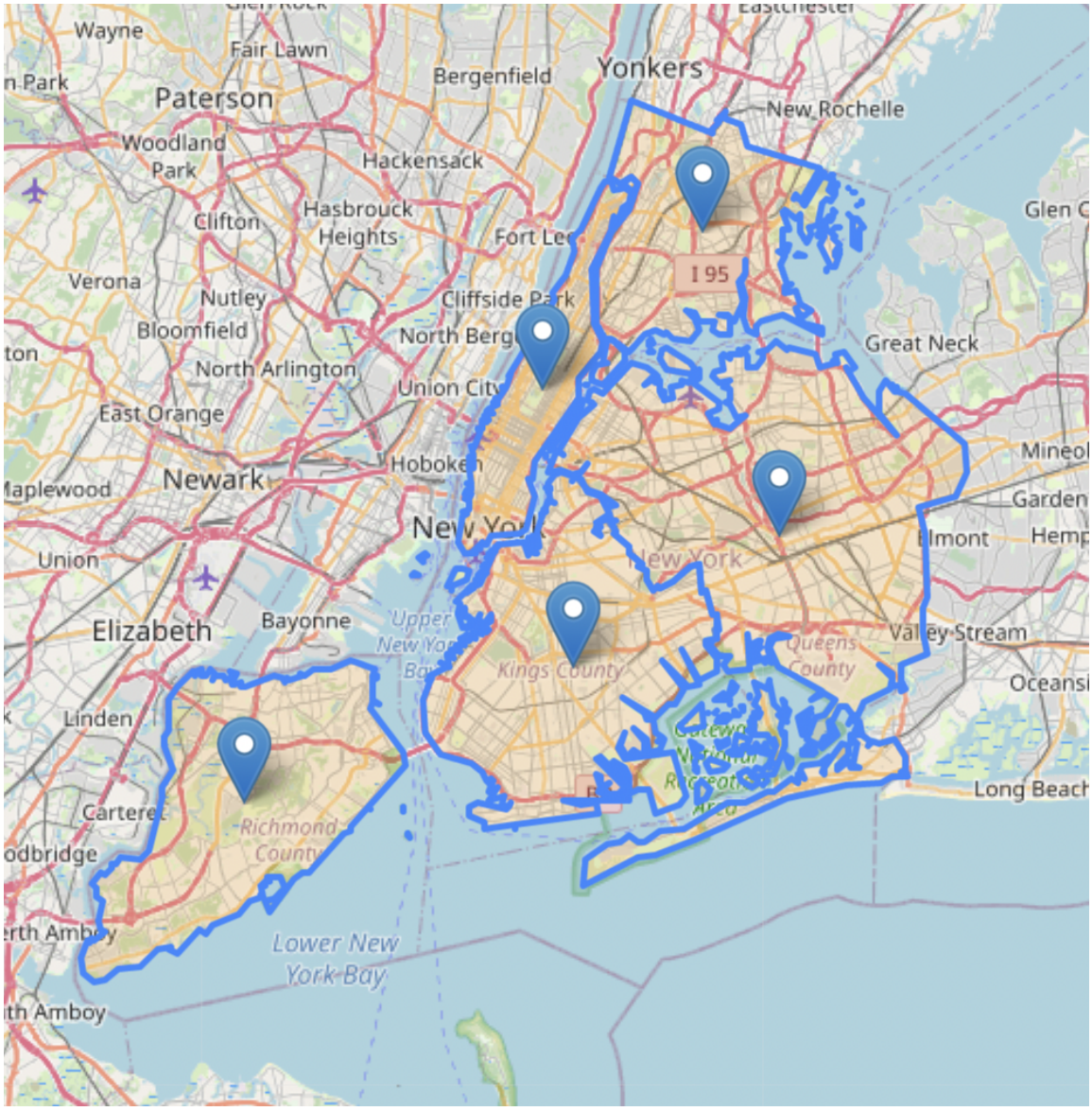


Fig : NEW YORK City BOROUGHS

We plotted New York city boroughs to find if there is geographical significance on valuing the houses.we suspect that some neighborhoods have the houses valued higher and we could compare the sales price of training values with the newly plotted predictions which don't have sales price (we consider the newly plotted houses as user queries)

Clusters as seen on the map:

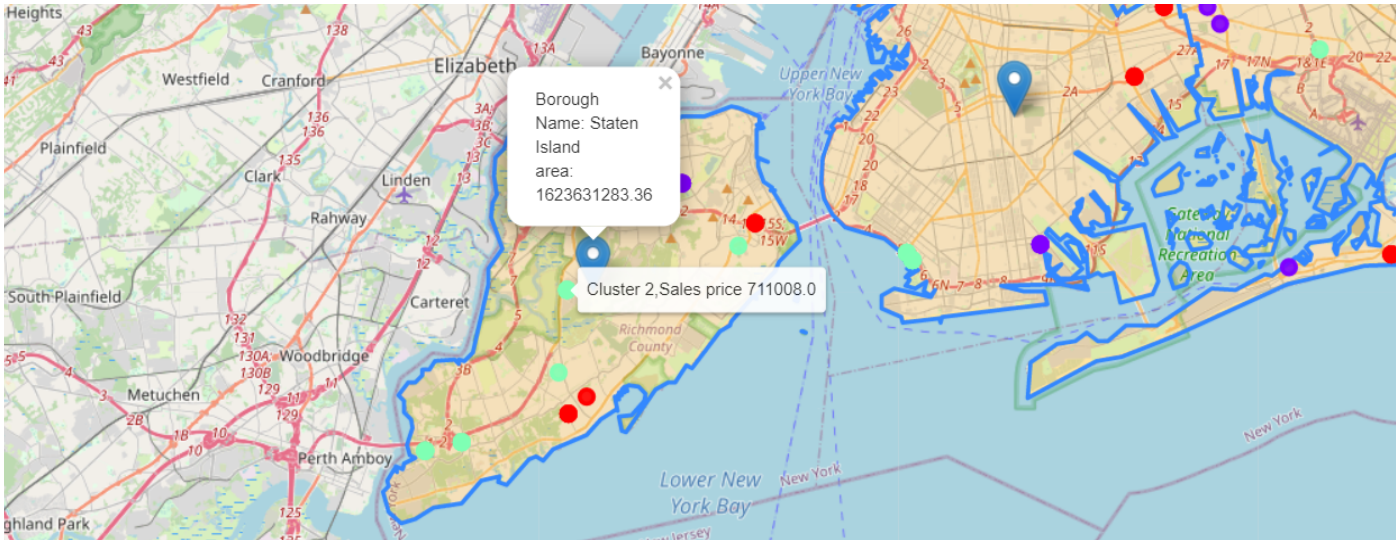
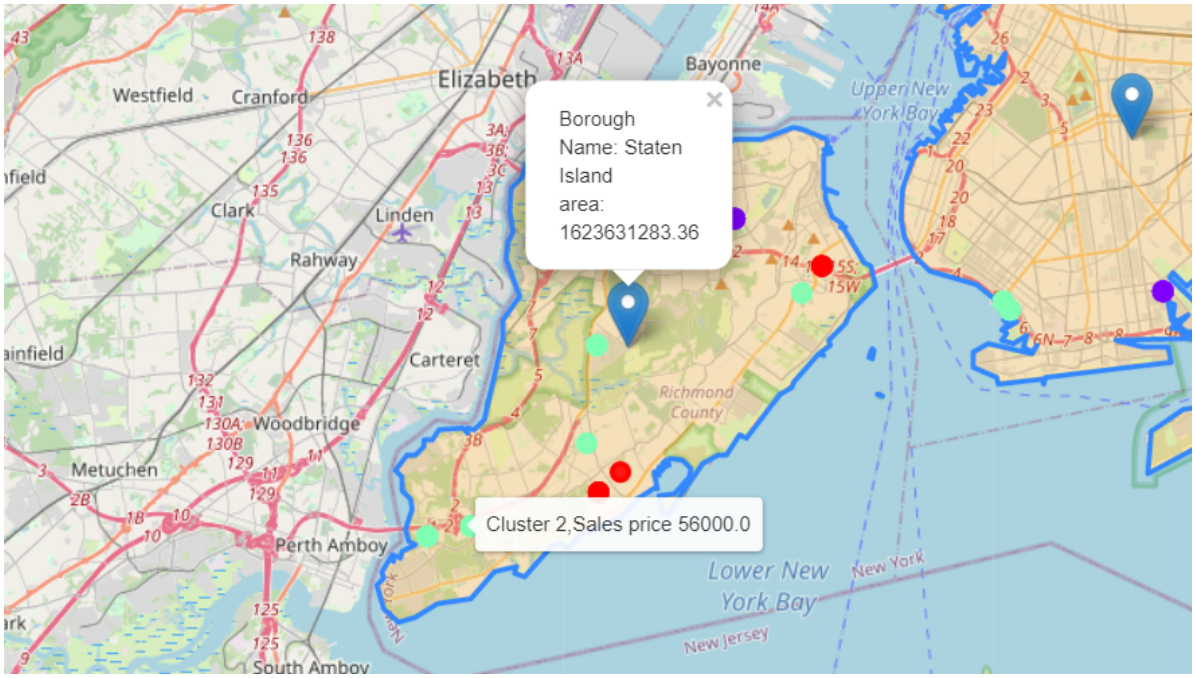
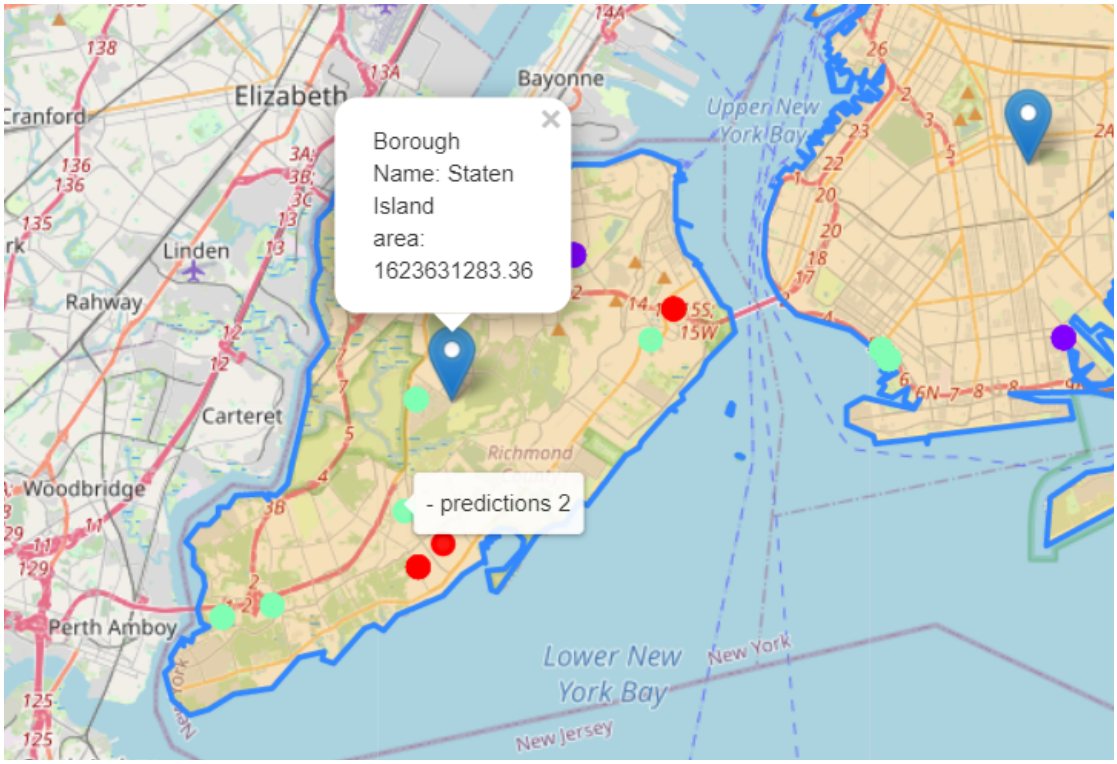


Fig: Cluster which has sales price

We can see that clusters which has sales price has a sales price of 71K



we can see that there is another cluster in the same category (2) with a sales price of 56K



We can see there is a user query plotted as prediction 2 (we consider data that dont have sales price as a user query) and we can make out that this property can have a valuation between 56K-71K dollars.

We can say that all cluster 2 category in staten island can have valuations between 56K-71K $ considering all the nearby property values

Recommendation of the best match:

We added this feature to recommend the closest match as per the value fed by the user and recommended the closest. we use gross square feet given by the user and get the closest match within the cluster which almost has the same gross square feet value.

We conclude that k means gives us clusters with high prices as a separate cluster namely cluster 0 and the other clusters are separated by geographical significance and residential units. Residential units' significance is known from the important feature bar graph of Random Forest classifier.

Geographical significance is evident from the above fig where staten island has high priced houses which is cluster 0 (red) and some medium priced houses , we can only find cluster 0 and cluster 1 in staten island showing that high priced neighborhoods have only

	GROSS SQUARE FEET	Cluster	point	closest_GSF	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQUARE FEET	BUILDING AGE	NEIGHBORHOOD	ADDRESS	SALE PRICE
0	18522	0	(18522, 0)	(18523.0, 0)	24	0	24	4,489.0000	97	ALPHABET CITY	629 EAST 5TH STREET	16,232,000.0000
1	5992	0	(5992, 0)	(5994.0, 0)	1	1	2	1,983.0000	116	UPPER EAST SIDE (59-79)	160 EAST 70TH STREET	9,350,000.0000
2	67854	0	(67854, 0)	(67652.0, 0)	78	0	78	15,000.0000	92	FLUSHING-NORTH	132-70 SANFORD AVENUE	18,522,529.0000
3	900652	0	(900652, 0)	(829024.0, 0)	894	8	902	141,836.0000	42	KIPS BAY	460-520 2ND AVENUE	620,000,000.0000
4	63423	0	(63423, 0)	(63000.0, 0)	66	0	66	13,125.0000	77	HIGHBRIDGE/MORRIS HEIGHTS	11 WEST 172 STREET	10,500,000.0000

From the above screenshot we can see that Recommend functionality helped us give the closest significant match based on Gross Square Feet and Cluster value. It gave us significant output such as SALES PRICE , NEIGHBORHOOD corresponding to the give input i.e. Gross Square Feet and Cluster value.

CONTRIBUTION:

Arun: Data Preprocessing, Clustering, Classification, Visualization and Geographical Plots, Cluster Analysis,Project Integration and Unit Testing

Shalvika: Data Preprocessing, Visualization, Feature Importance, Cluster Analysis, Report,Project Integration and Unit Testing

Shreyas: Data Preprocessing, Classification, Recommendation for Users, Cluster Analysis ,Project Integration and Unit Testing