Predicting Housing Price In a Well Defined Neighborhood

Judy Dae

May 5th, 2020

1. Introduction

1.1 Background

In the Real Estate market, one of the most interesting questions is "How much is that house?". That question can be asked in various situations. The answers to that question can help greatly when a realtor needs to put a house on the Market. Homeowners often want to know what their house's worth. In most cases, it is the biggest investment after all. A buyer wants to know they paid a fair price for the house they are interested in. The loan officer wants to make sure the money they are lending out is a good amount. The investors want to know if they are getting a good deal. In short, the question of "How much is that house?" is a very common question that is being asked, and the answer to that question is critical to many perspect.

1.2 Problem

The current tool many Real Estate Professionals use is called CMA. Realtor will login to the professional association account. Use search queries to pull the data according to a particular address they are interested in. After they have some data, they will need to manually choose the houses that they believe are comparable to their subject property. Including by distance to their subject property, the number of bedrooms, the year of the house was built. Then they will need to estimate the price using their best judgement. Most of the time, realtors just pick a number between the highest and the lowest. This method can create inaccurate data and many times cause unpleasant conversation among realtors and their clients.

1.3 Interest

As a realtor myself, I often want to have some more reliable ways to help me to get my client the best answer they deserve. More data based, less human interference result.

2. Data acquisition and cleaning

2.1 Data sources

As a realtor, I have the access to the real estate market data in our area. I login to har.com as professionals. From there I run a query to get all the information for all the houses that has been sold in the last year in the neighborhood that I am interested in. For this project, I used the neighborhood of one of my clients. His house is at 22 Hessenford St, Surgarland, TX 77479. I then download all the data into a csv file. I then uploaded the file to my python environment.

2.2 Data cleaning

The data downloaded from the website contains a lot of information. After I get the csv file into a data frame, I noticed there are some problems with the dataset.

First, there are many missing values in the dataset. I decided to replace the value with the most reasonable approach I can think of. For example, when the number of fireplaces is missing, I replaced the value with 0 since that is what is most likely the case.

Second, there are data that is not in the neighborhood. The reason I know that is because the neighborhood I am interested in does not have houses that are zoned to a certain high school. I decided to drop those data from my dataset.

Third, we know schools are an important factor for the housing price. In order for me to use the school information in my model, I will need to change the value from *string* to *int.*

Fourth, some columns that I need for my model are the wrong datatype, I changed the datatype to to *int* as well.

After I fixed these problems I checked for outliers in the data. I found there were some extreme outliers, mostly because we have some huge mansions sitting on a nearby golf course. In order to create a more reasonable model, I decided to eliminate those data from my dataset.

I feel more comfortable with the dataset I have right now to start working with my ML model.

2.3 Feature Selection

After data cleaning, there were 54 features, after examination, there are many features that are irrelevant. For example, the listing agent information, the MLS number, and etc.

There were also redundancies, such as all the houses in the neighborhood have the same zip code, in the city, and in the same school district.

After I discarded all the irrelevant and redundant data, I inspected the correlation of independent variables. And this helped me to decide the features I should use.

```
[12]: df_ori[['Building SqFt','Year Built','Lot Size','Baths Total','Stories','Close Price']].corr()
```

[12]:

	Building SqFt	Year Built	Lot Size	Baths Total	Stories	Close Price
Building SqFt	1.000000	0.574365	0.623881	0.776606	0.663900	0.910774
Year Built	0.574365	1.000000	0.283882	0.600529	0.394940	0.601003
Lot Size	0.623881	0.283882	1.000000	0.471454	0.357623	0.583859
Baths Total	0.776606	0.600529	0.471454	1.000000	0.588283	0.780316
Stories	0.663900	0.394940	0.357623	0.588283	1.000000	0.549614
Close Price	0.910774	0.601003	0.583859	0.780316	0.549614	1.000000

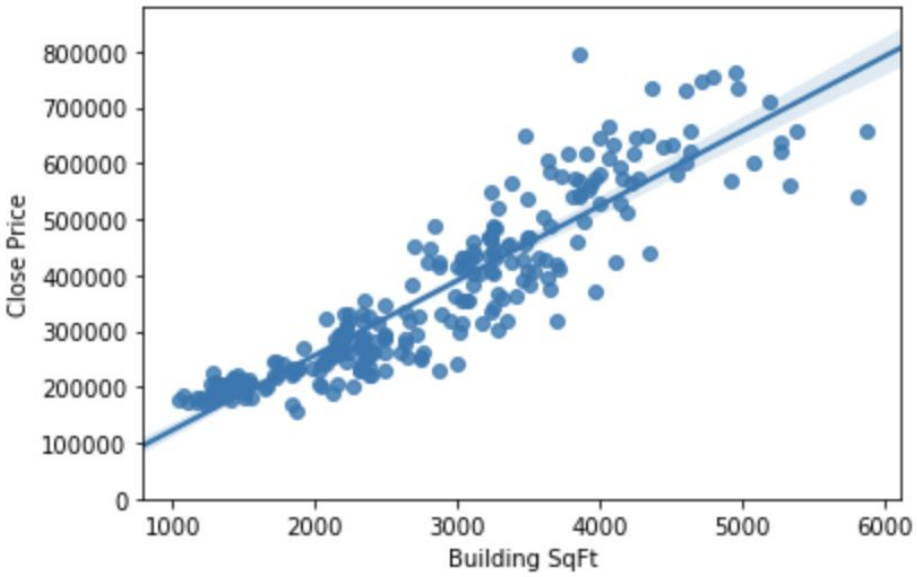
3. Exploratory Data Analysis

3.1 Data Visualization

Building SqFt Vs. Sale Price and Year Built Vs Sale Price

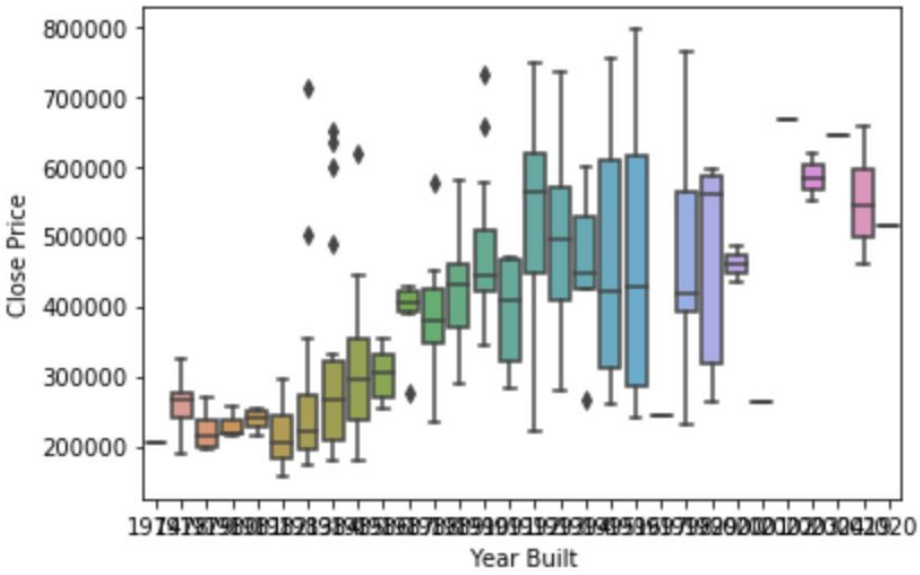
[13]:
```
sns.regplot(x="Building SqFt", y="Close Price", data=df_ori)
plt.ylim(0,)
```

[13]:
```
(0, 879100.9799725484)
```



[14]:
```
sns.boxplot(x="Year Built", y="Close Price", data=df_ori)
```

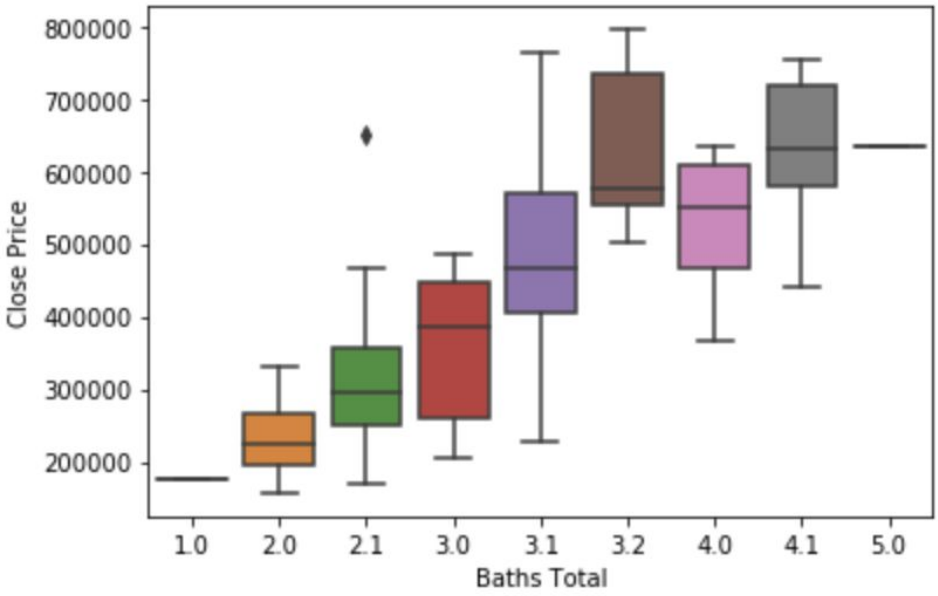
[14]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa53441afd0>
```



Bathroom Total Vs Sale Price and Stories Vs Sale Price

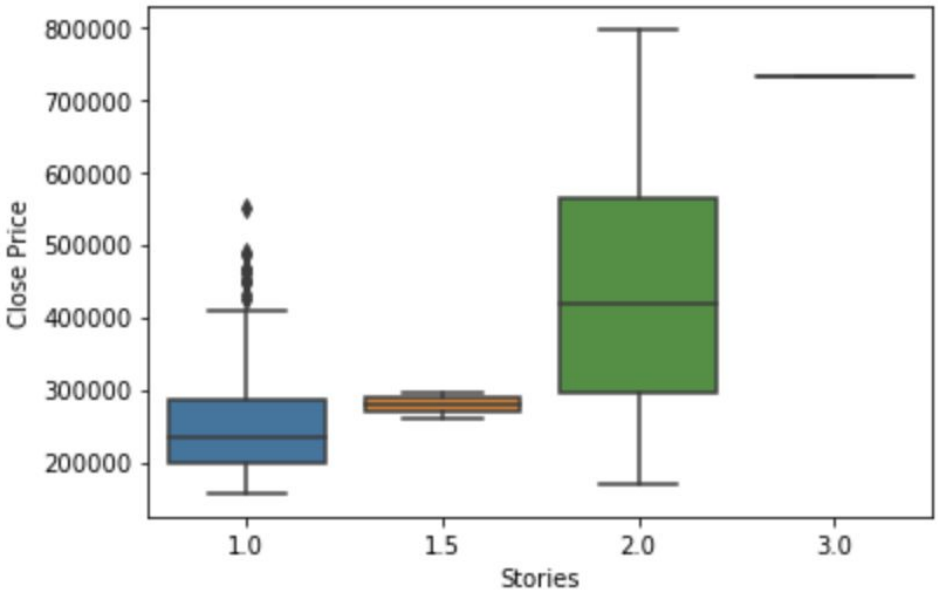
```
15]: sns.boxplot(x="Baths Total", y="Close Price", data=df_ori)
```

```
15]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa534142748>
```




```
16]: sns.boxplot(x="Stories", y="Close Price", data=df_ori)
```

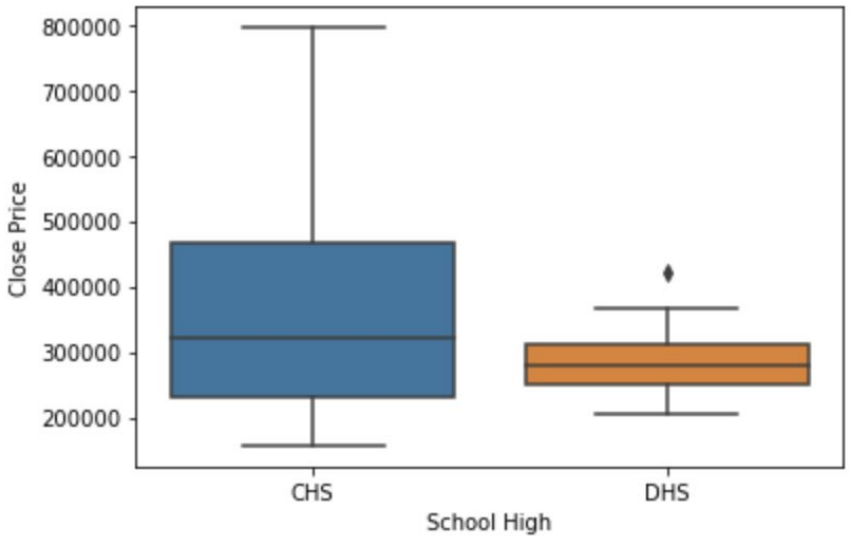
```
16]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa53403f438>
```



High School Vs Sale Price

```
[17]: sns.boxplot(x="School High", y="Close Price", data=df_ori)
```

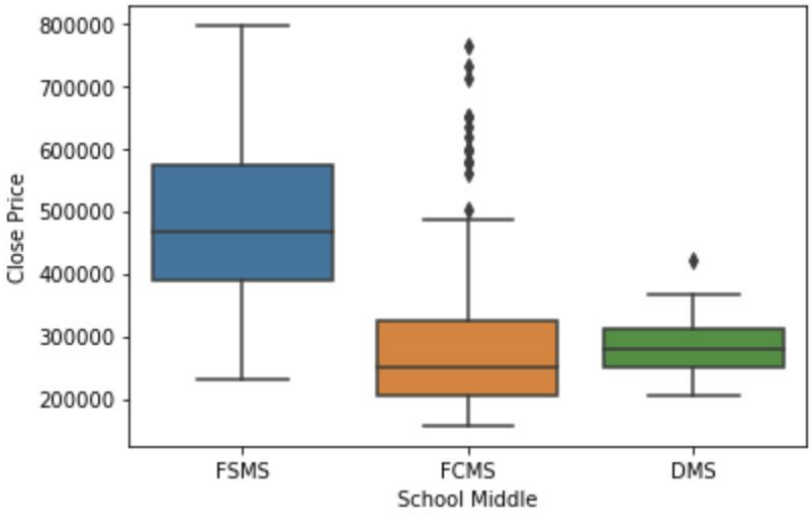
```
[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa51c7aff28>
```



Middle School Vs Sale Price

```
[18]: sns.boxplot(x="School Middle", y="Close Price", data=df_ori)
```

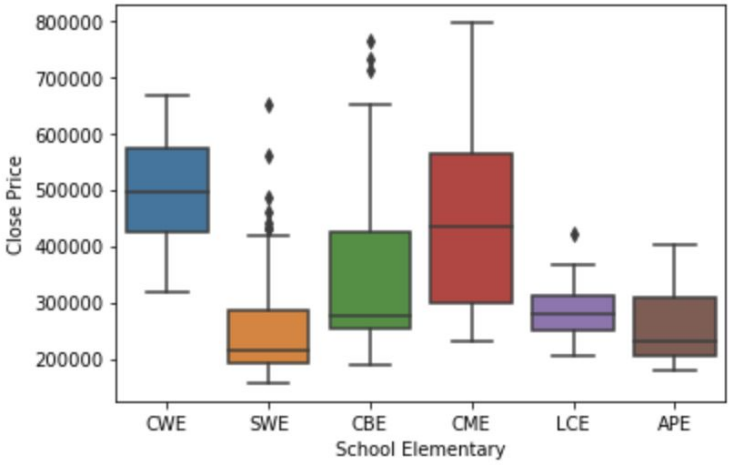
```
[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa51c6b7550>
```



Elementary Vs Sale Price

```
[19]: sns.boxplot(x="School Elementary", y="Close Price", data=df_ori)
```

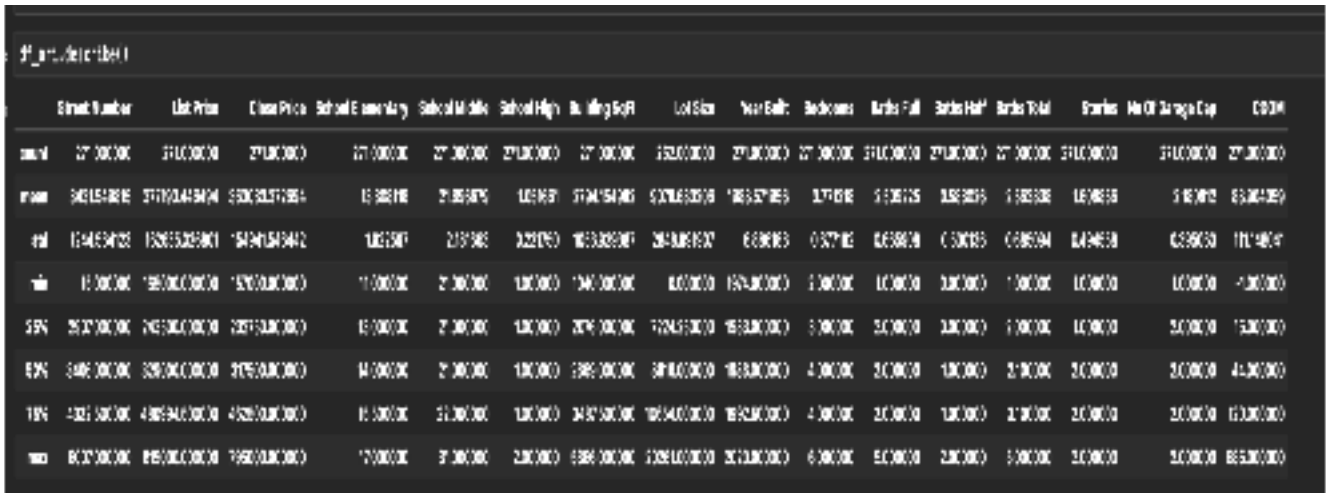
```
[19]: <matplotlib.axes._subplots.AxesSubplot at 0x7fa51c641668>
```



3.2 Data Statistics to demonstrate the data attribute of the dataset.

It would be interesting to look at the dataset for some statistics value. I used the descrip() function to display the most common statistics factor for the dataset.

It is really interesting to paint a picture of the neighborhood from the statistics point of view. This chart offered a vivid description of the neighborhood.

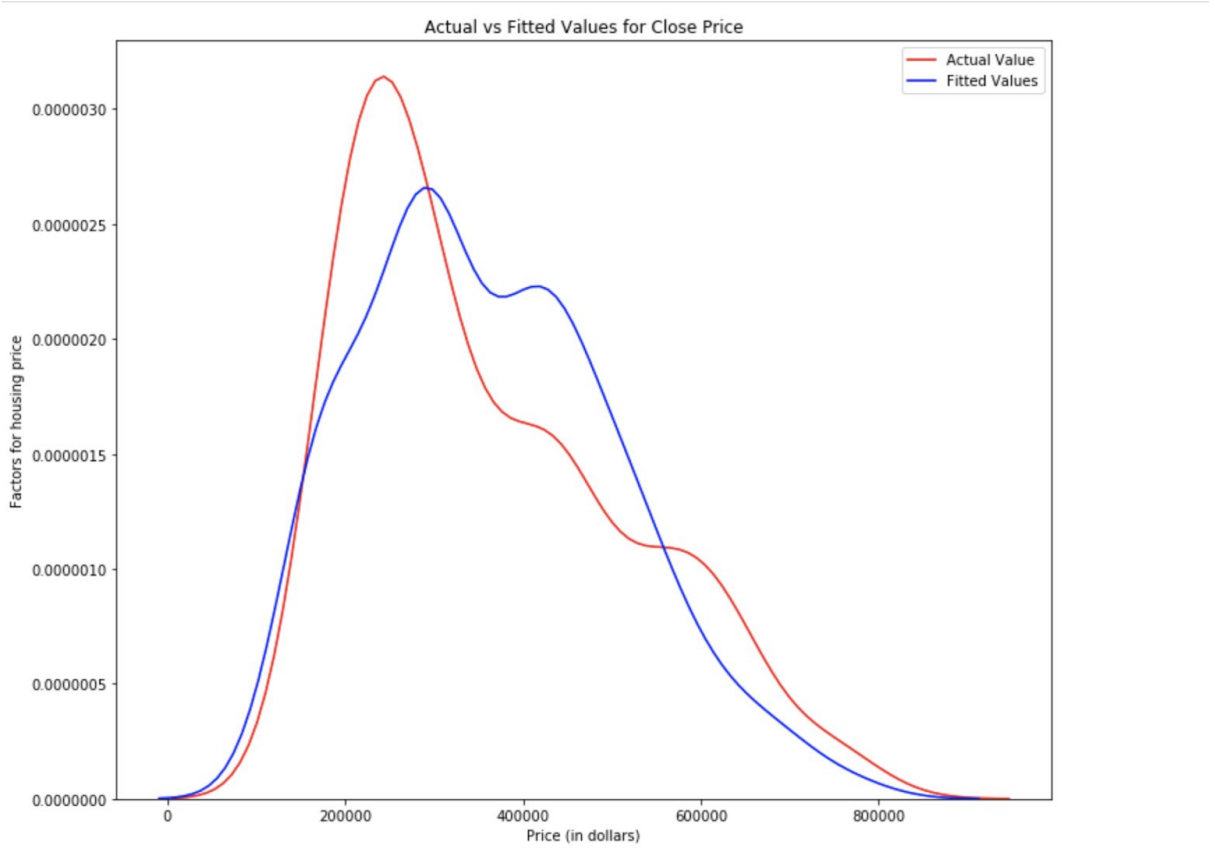


4. Predicting Model

In the project, we are trying to create a model to predict housing price. Given data with different features, the output should be a continuous result, so this is a regression problem.

4.1 Linear Regression with one single most correlated feature

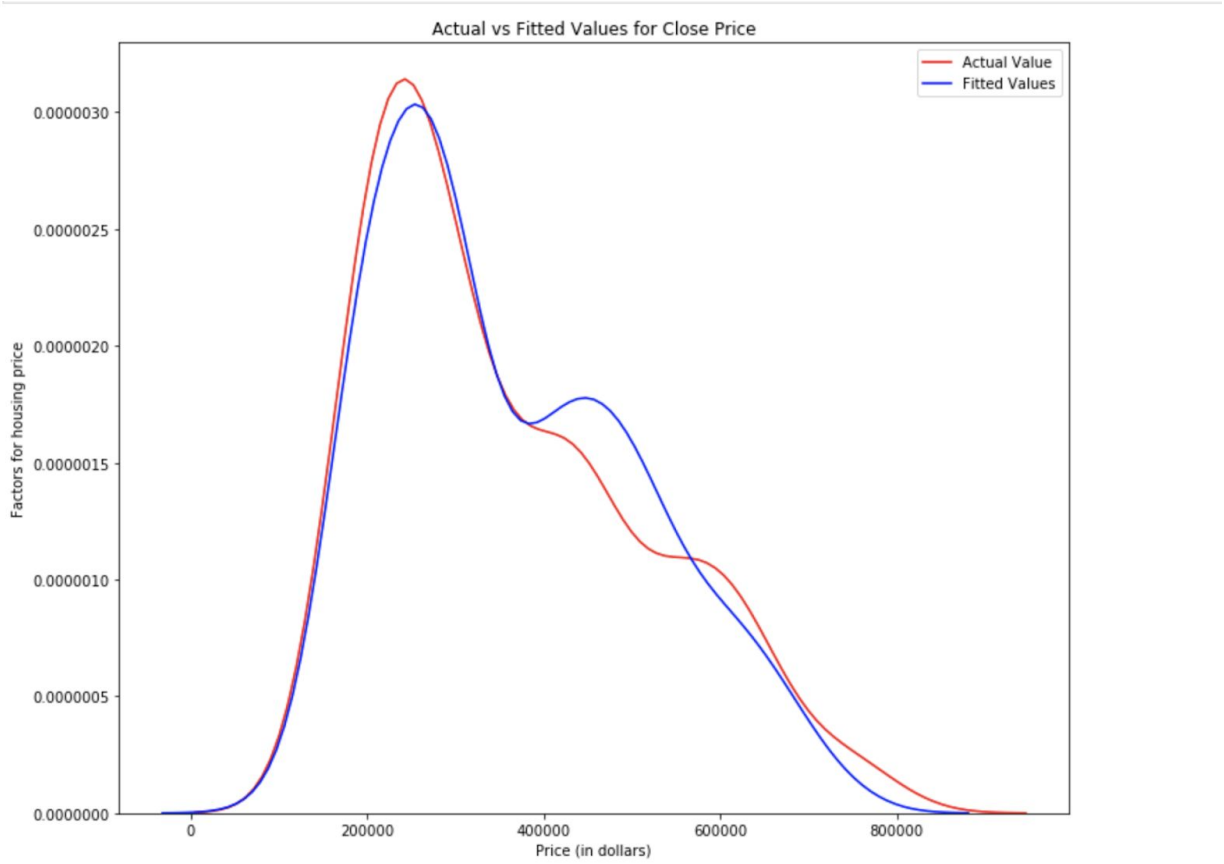
Using LinearRegression function from sklearn package. I am able to create this :



The single featured Linear Regression model is showing something promising. Based on this result, I further the model by using Linear Regression with multiple features.

Based on the previous steps in the data analysis. I decided to use the following features: *'Building SqFt', 'Baths Total', 'Year Built','Bedrooms','Stories','No Of Garage Cap','School High', 'School Middle', 'School Elementary', the target value is the* ***Sale Price.***

4.1 Linear Regression model with more features



From this model, we can confirm the features we used for our model are pretty accurate. To further study this model, I decided to use the training set and test set approach.

4.3. Training set and test set to study the model preformance

The concept of the training set and test set is to create two subsets of data from our original dataset: one subset is used as a training set, the other subset is used as a test set. . We use the training dataset to train our model till we have a pretty good result from the model. We then use our test set to test the model to evaluate the model performance.

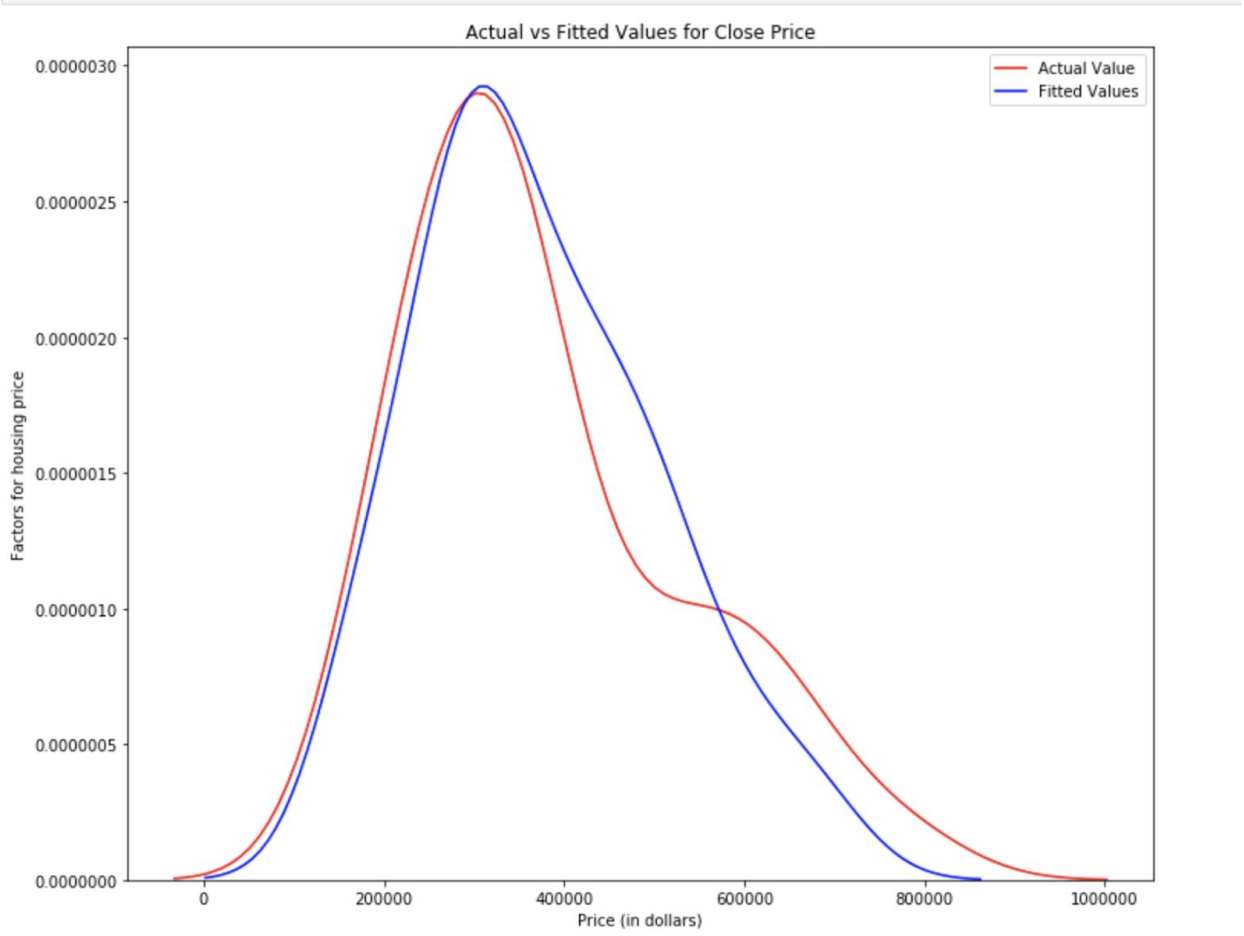
I used package *sklearn.* First I use the function *random to create* my training set and test set. And I used the training set and *LinearRegression* function to get the multi-linear function. Using the fit function, I was able to find the coefficients.

Next, I use the test dataset to check the performance. I calculate the Variance score, Mean absolute error, MSE, and R2-Sore.

4.4 Result

Coefficients: *[[-3.09434057e+05 1.15535596e+02 3.95142319e+04 1.49031630e+03 -1.61795618e+04 -2.80165649e+04 2.35726653e+03 2.29120663e+04 1.10463663e+03]]*
Residual sum of squares: *3868490604.29*
Variance score: *0.83*

Mean absolute error*: 44055.88*
Residual sum of squares (MSE)*: 3868490604.29*
R2-score: *0.77*



5. Conclusion.

In this study, I analyzed the relationship between some features of a house and its sale price. I identified SqFt, total number of bathrooms, Schools, Year Built and Stories are among the most important features that affect price of a house in a specific neighborhood. I built regression models to predict the price for a house given those important features. The model performed fine for the goal of this project.

6. Future direction of the model

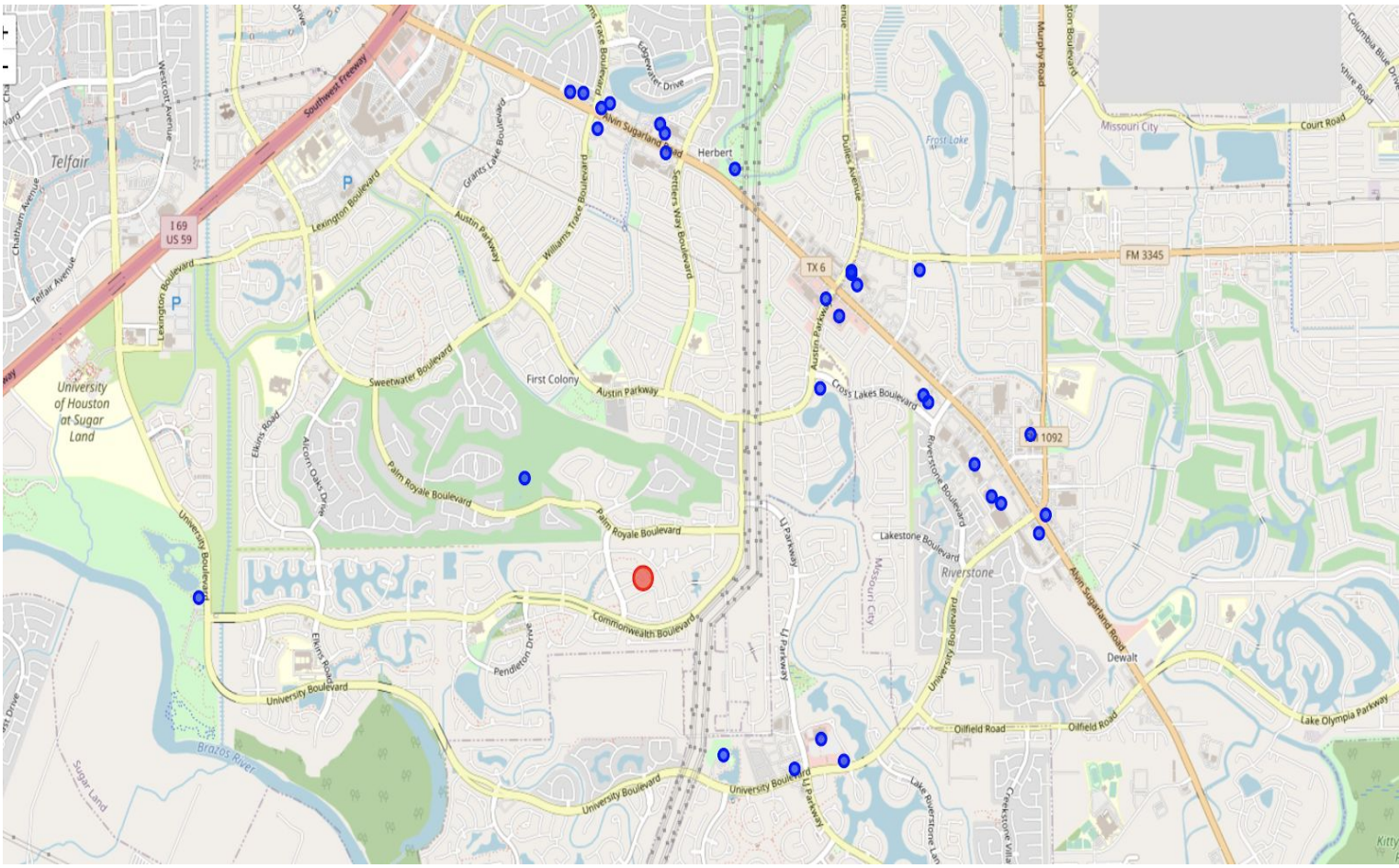
I would like to study different neighborhoods to confirm if this model can be generalized in different areas. I also would like to further study if the smaller subdivision or certain street can be extra features to improve the model's performance.

7. Neighborhood Information

Another perspective for this project is to present potential buyers with the neighborhood information. This information is compiled to show the potential buyer the basic geographic information, and popular venues in the nearby area.

I used foursquare API to generate the information and create a map. In order to do that, first we need to find out the longitude and Altitude, then make an API request call to foursquare and get the *venue* data in a Jason file.

Read Jason file into data frame. Then use *Folium* to create the map.



This concludes my project. The project is designed to help user to pinpoint the house price in a specific neighborhood, and provide local venue information around the neighborhood.