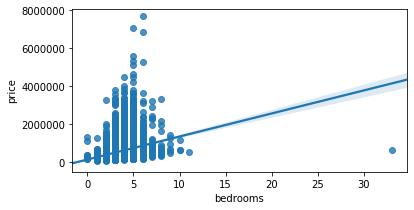
Capstone project two milestone report

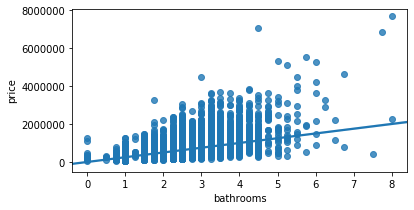
When comes to the decision of buying or selling a house, the buyers and the sellers of the house most likely do not have enough information they need to make the sound decision. By going through all the existing sold housing data and using them to create housing price prediction models, it allows clients, the buyers and sellers of houses also the brokers, to be on the same page. They get to have a clearer idea on how much their houses are worth in the market. This way, it would be easier for buyers to make the decisions on how much to bid for a house and for sellers to know what are the reasonable prices to sell their houses. It would saves countless hours for all the parties that are involved in the housing transaction deal.

The housing dataset that I obtained from Kaggle has a total of 21613 rows and 21 columns in total. These are the house sale prices for King County that were sold between May 2014 and May 2015. The dataset is clean. I did not have to perform any data cleansing tasks on the dataset. The dataset includes many essential features of a house such as the total number of bedrooms it has, the total number of bathrooms it has, and the total square feet of the house and more. All these columns are numerical values. The dataset is perfect for doing regression modeling.

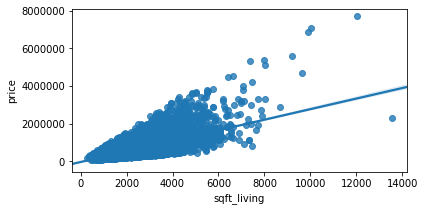
After I loaded the CSV file into Python Pandas, I applied Pandas methods to find out the shape of the dataframe which is the total number of rows and columns it has. I also checked out the datatype for each of these columns. Later, I applied seaborn library’s regplot to identify the relationship between each of the housing features like the total number of bathroom, total number of bedrooms etc with its depending value, the housing price. Many features show strong positive correlation with housing price; these fields are the total number of bedrooms, the total number of bathrooms and the total number of square feet of living.



*It shows positive correlation between number of bedrooms with housing sales price*

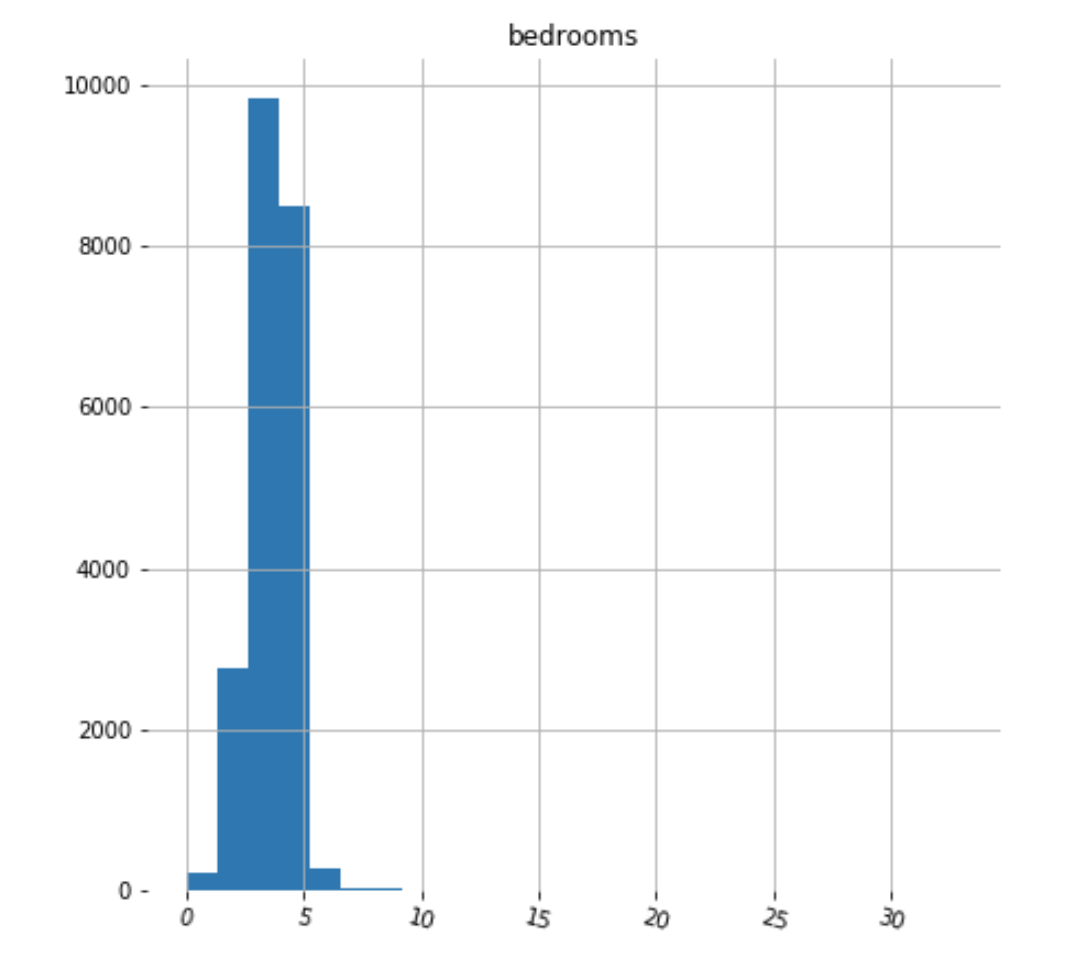


*It shows positive correlation between number of bathrooms with housing sales price*

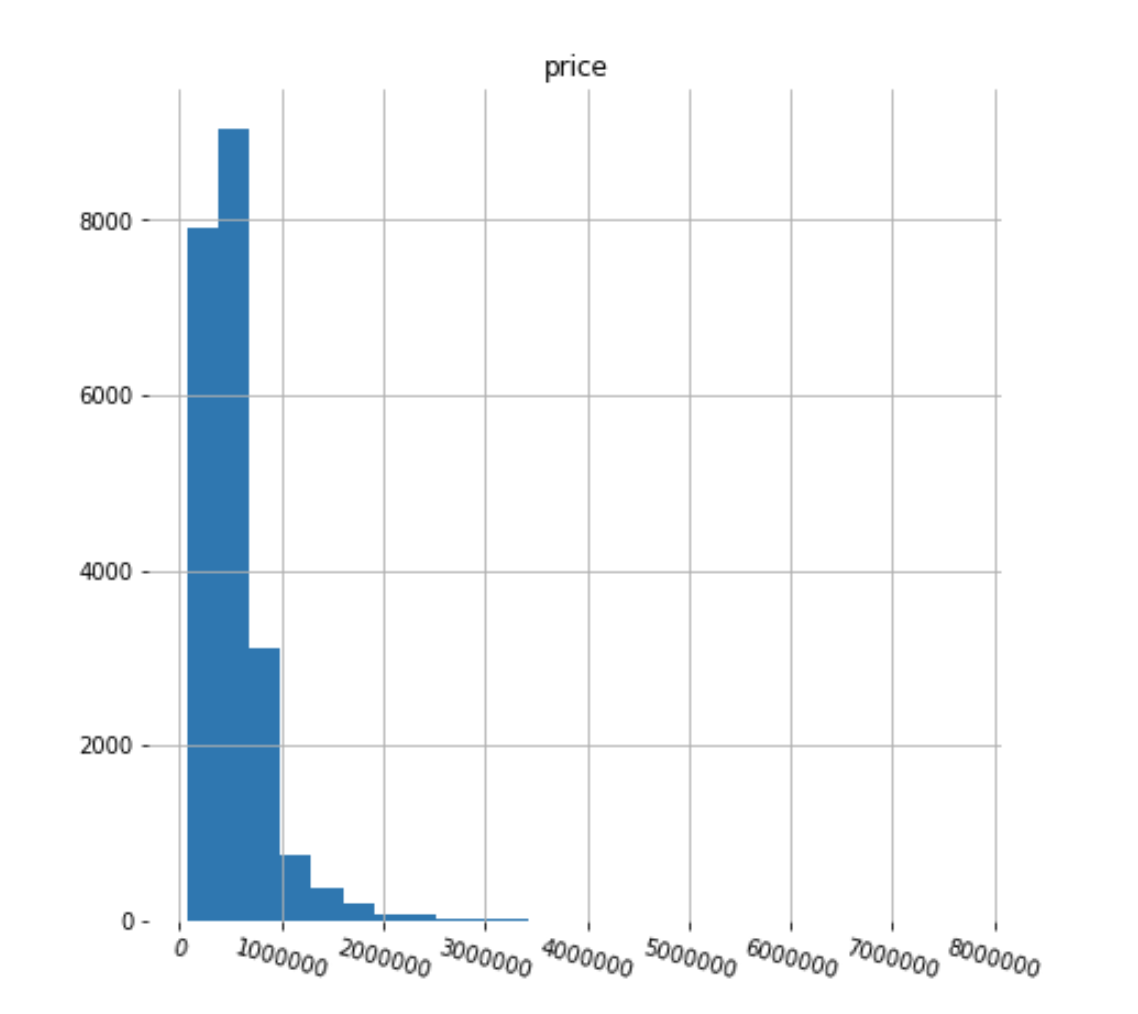


*It shows positive correlation between total square feet of the house with housing sales price*

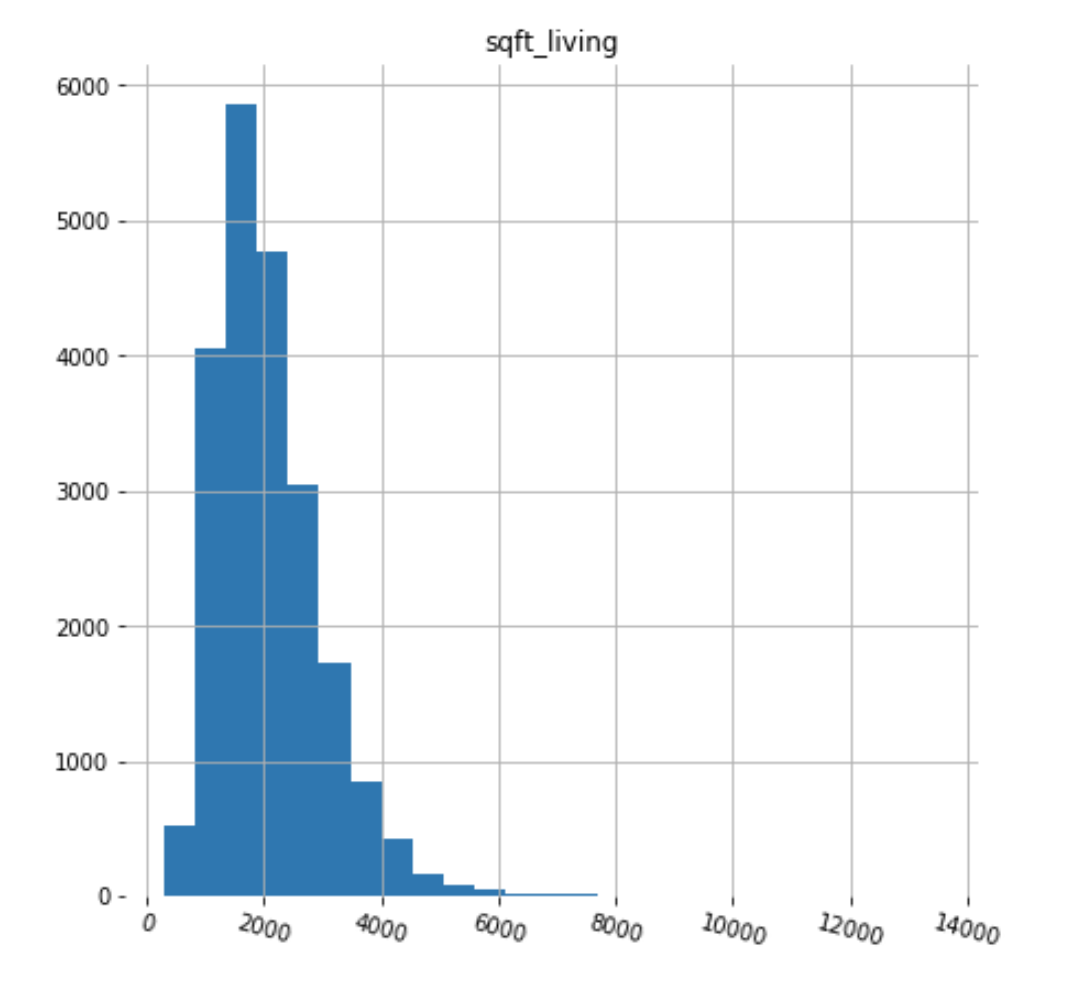
I plotted some of these features into histograms so I could understand these data better. A lot of the houses in King County has two bathrooms and many are built around 2,000.



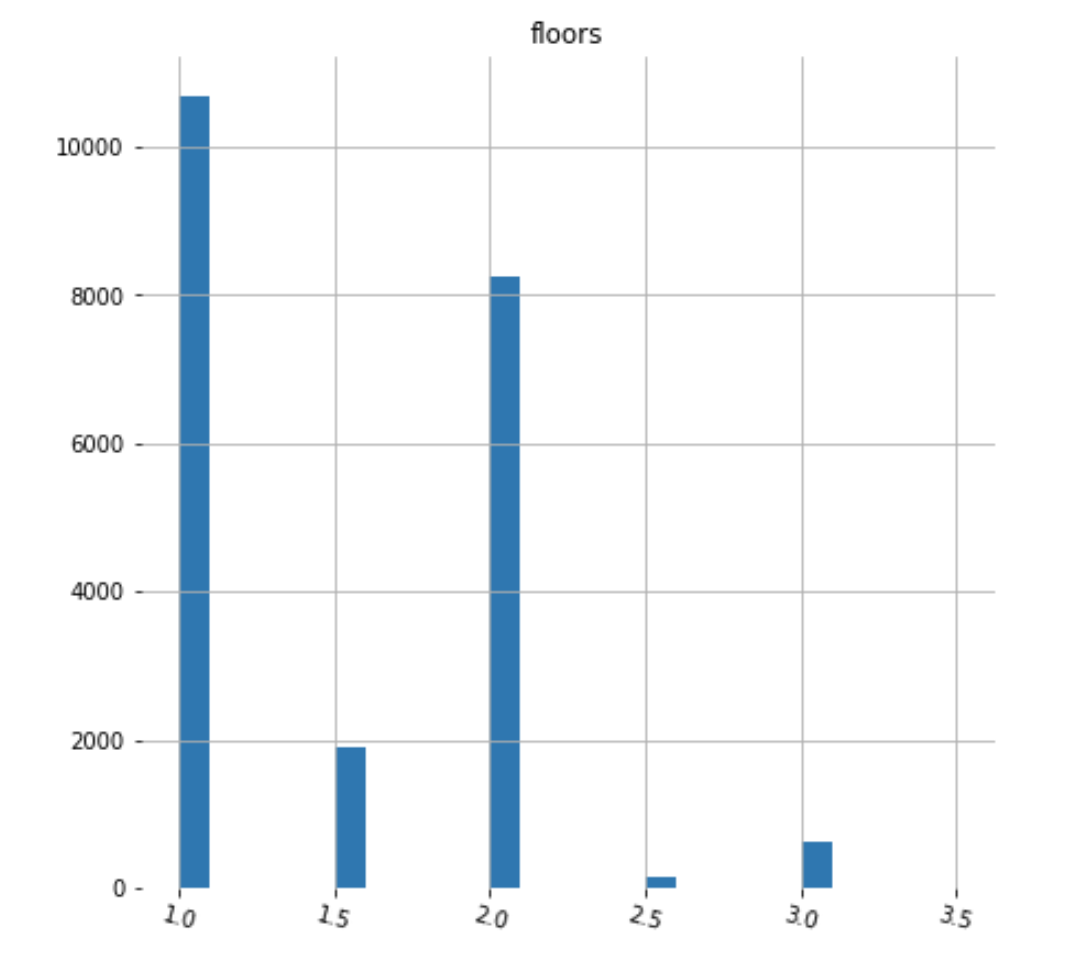
*It shows that majority of the houses in Kingston has 3-4 bedrooms*



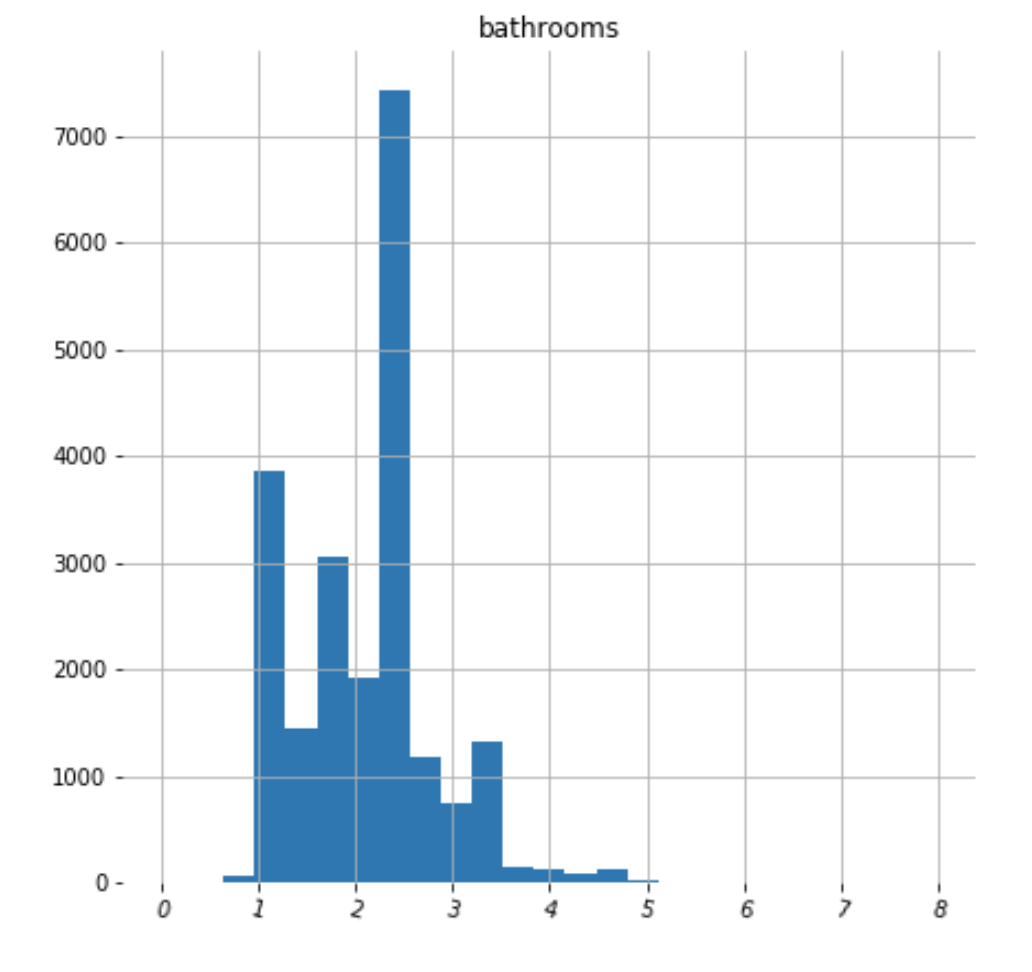
*It shows that majority of the housing prices in Kingston falls below 1 millions dollar*



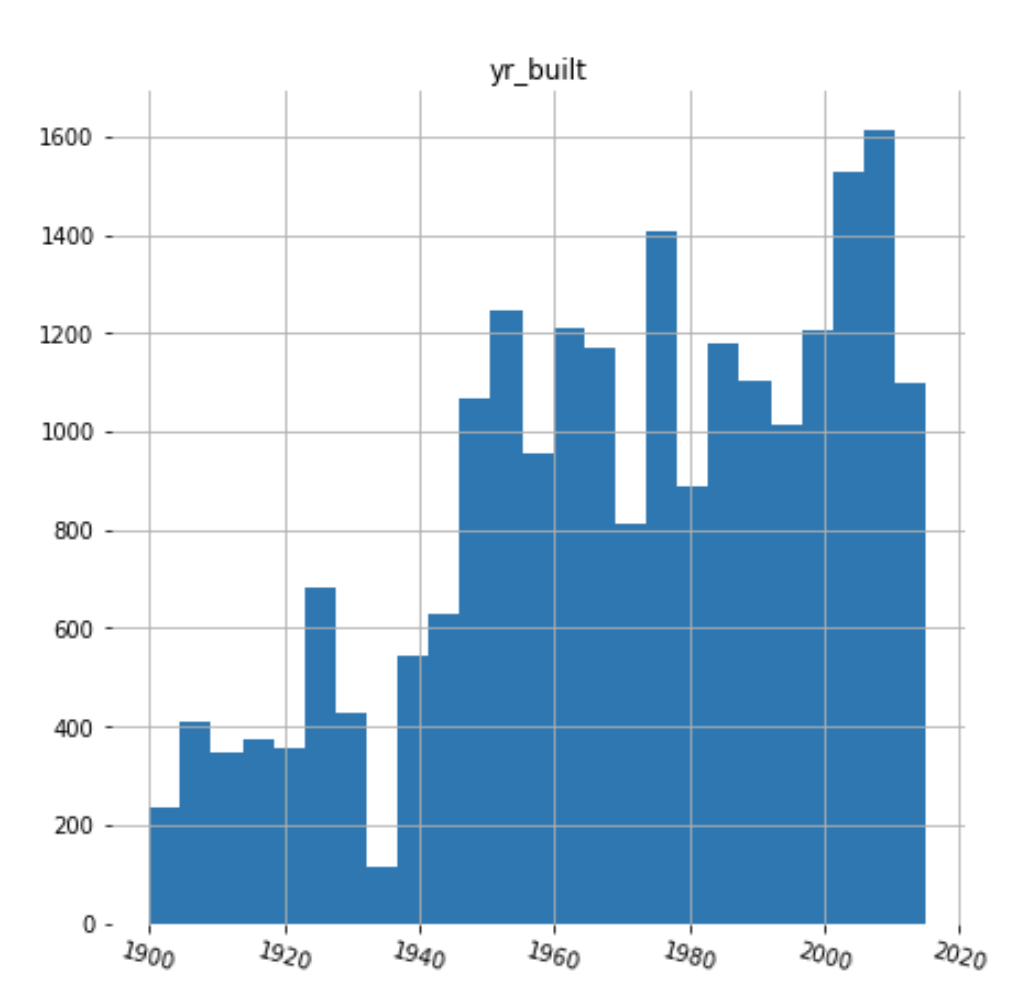
*It shows that majority of the houses in Kingston has around 2000 square feet*



*It shows that majority of the houses in Kingston are either 1 floor house or 3 floors house*

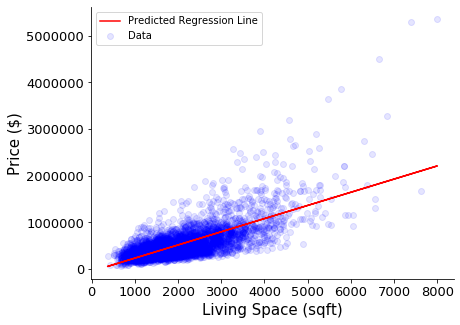


*It shows that majority of the houses in Kingston has 2 bathrooms*



*It shows that a lot of the houses in Kingston fairly new, built around 2000*

After I have a good idea on the dataset that I am working with, I started applying machine learning techniques on the dataset. I first created a simple linear regression model using a single feature, total living space, to predict its housing price. I split the dataset into two, eighty percent of them for the training purpose and remaining twenty percent for the testing purpose. After I trained the linear regression model with eighty percent of the dataset, the trained linear regression model shows a y intercept of negative 47236 and coefficient of 282, which means for each increment of living square feet, the housing price goes up by 282 dollars. Later, I used the mean squared error and R2 score to determine the accuracy of this linear regression model. The results are 25489 for the mean squared error and 0.496 for the R2 score.

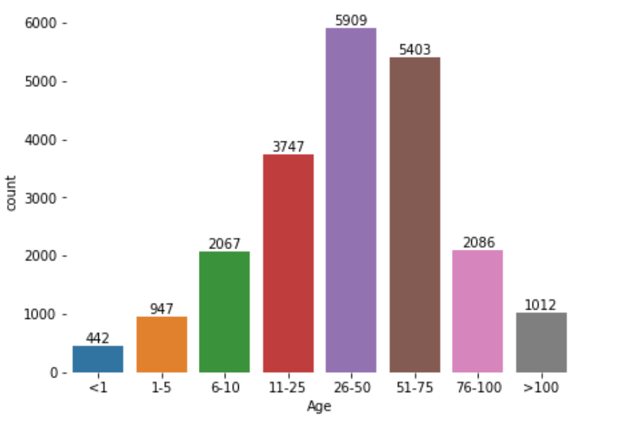


*Simple linear regression model*

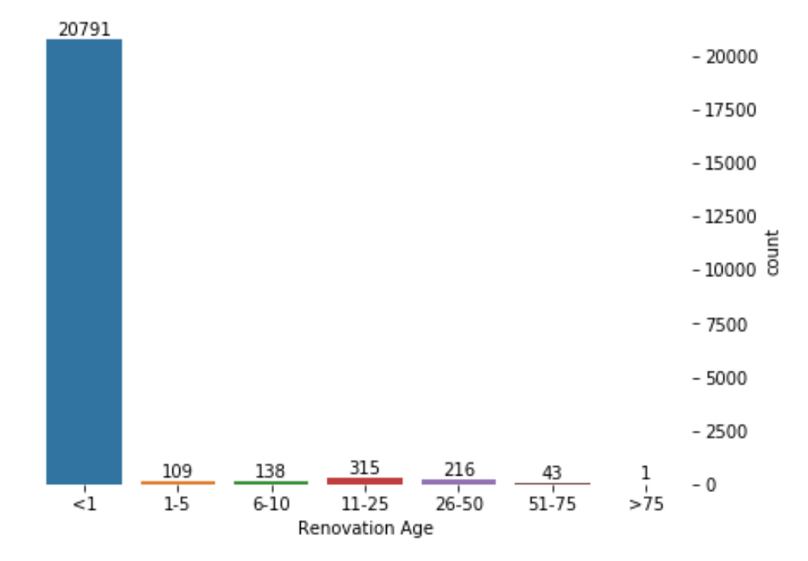
Next, I added more features to the regression models, using multiple features to predict the house pricing, which makes it multi-regression model. The features I used are: the total number of bedrooms, the total number of bathrooms, the total square feet of living, the total square feet of lot, the total number of floor and zip code. The trained the multi-linear regression model with the same data, and it shows a y intercept of negative 57221293 and several coefficients. I used this trained model to predict the remaining twenty percent of the datasets that excluded house price column, and then compared the predicted house price with the actual house price, it has mean squared error of 248514 and R2 score of 0.5188. Based on the mean squared error and R2 score results, this multi-linear regression model performs slightly better than the linear regression model mentioned earlier.

Again, I created another multi-regression model. This time, including all the features in the dataset: the built year, square feet of the basement, condition, year renovated and more. This multi-regression model shows a mean squared error of 193693 and R2 score of 0.7077. It performs better than previous two models.

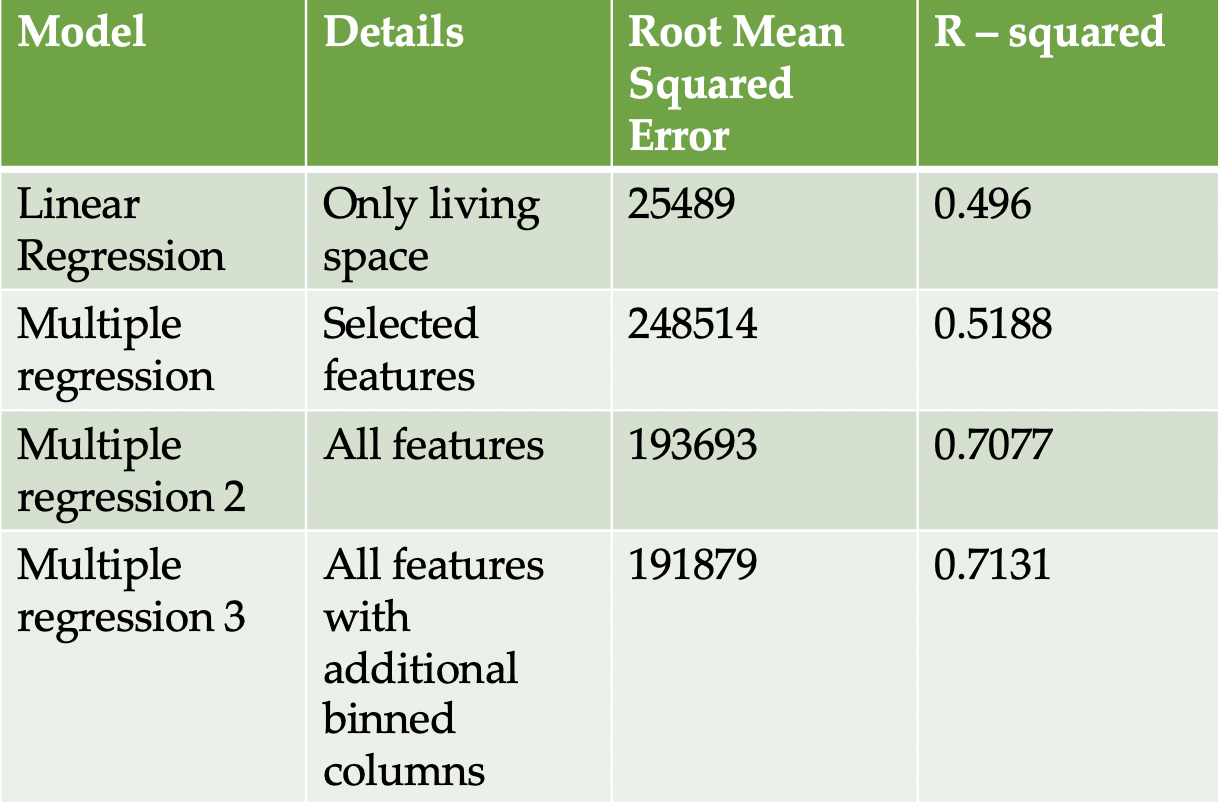
At the end, not fully satisfy with the accuracy of these models, I tried to see if there is anything else I could do to improve the accuracy of the result. I applied binning to the built year column and renovated year column; so instead of showing the year the house is built or renovated, the dataframe would have multiple columns that shows the age of the house or how many years has past since it was last renovated. The built year column and the renovated year column are divided into multiple columns, partitioned into intervals of less than 1 year of built time, between 1 to 5 years of built time and more. Same for the renovation age, it is divided into multiple columns, partitioned into intervals of less than 1 year, between 1 to 5 years renovation time and more. Adding these additional columns into the multi-linear regression model, the testing result show slight improvement. It now has a mean squared error of 191879 and R2 score of 0.7131. It performs better than all the previous models.



*It shows most of the houses are between 11 to 75 years old*



*Many of these houses just got renovated before it was sold*



*Comparing the prediction accuracy of these models*

For this project, I only applied linear regression model and multi-linear regression models for the prediction. There are many other machine learning models that are also useful in predicting the house prices. For example, using KNN regression model, given a house, its features, the KNN regression model would be able to use the given features to find the most similar house in the dataset in term of these features, then return its sold price.

We could also use polynomial regression model. However, it could often lead to overfitting, in which, the model would show low mean square of error but not great at predicting the house price.