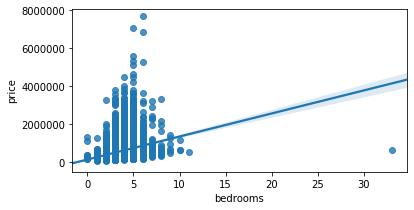
Capstone project two milestone report

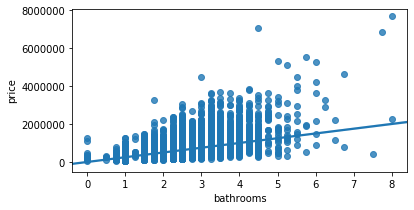
When comes to housing purchasing and selling, buyers and sellers might not have all the information they need to make the decision. By going through all the existing sold housing data, it allows clients: buyers and sellers of houses and brokers to have clearer ideas on how much their houses are worth in the market. This way, it will helps them to reduce the number of hours needed to spend on doing research on houses when comes to the decision how much to bid for a house or how much to sell for a house.

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 perform any data cleansing. The dataset includes many essential features of a house like 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. The dataset is perfect for doing regression modeling.

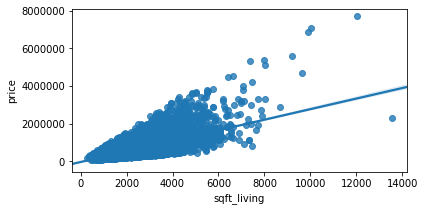
For the data analysis part, I applied seaborn library’s regplot to identify the relationship between each of the housing features like the total number of bathroom, total number of bedrooms etc with the housing price. Many features show strong positive correlation with housing price; they are the total 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*

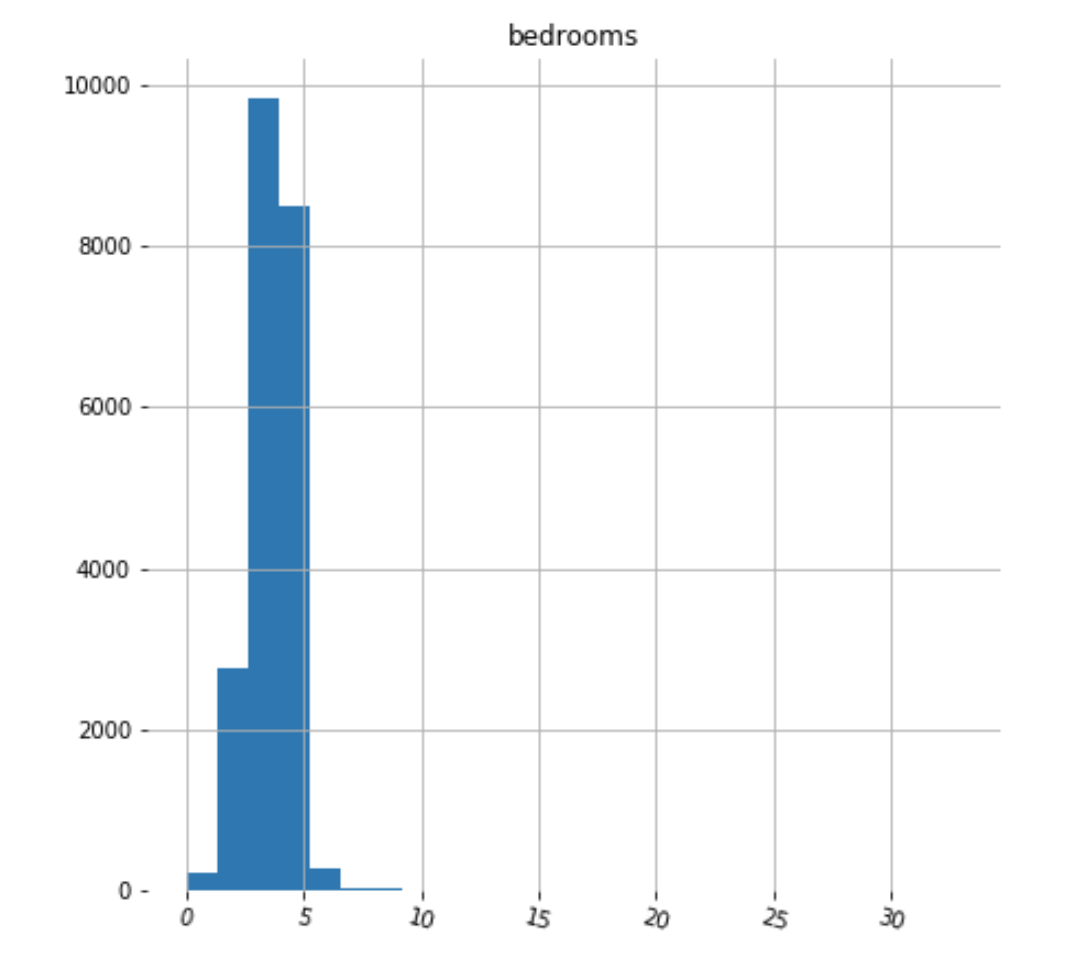


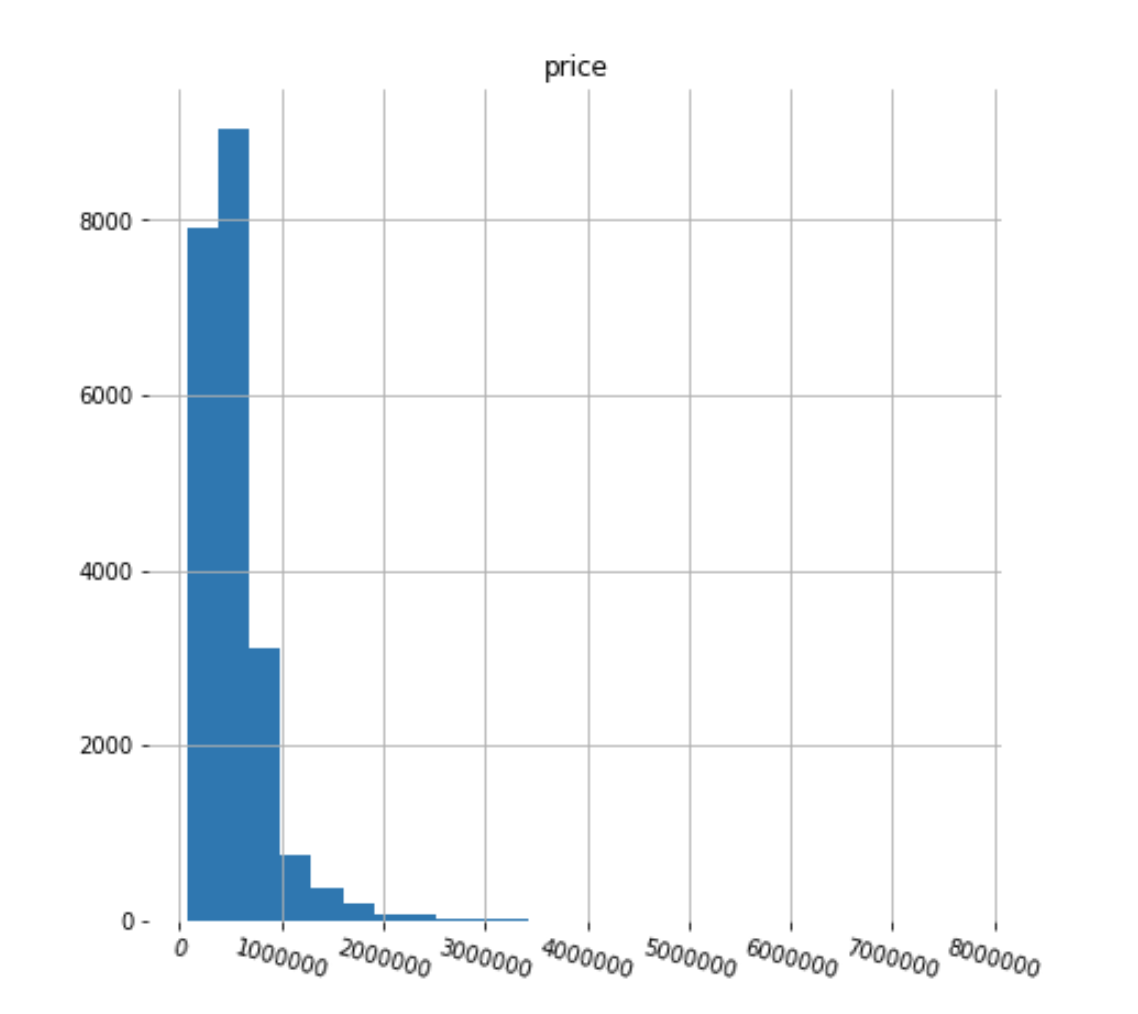
*It shows positive correlation between number of bathrooms with housing sales price*

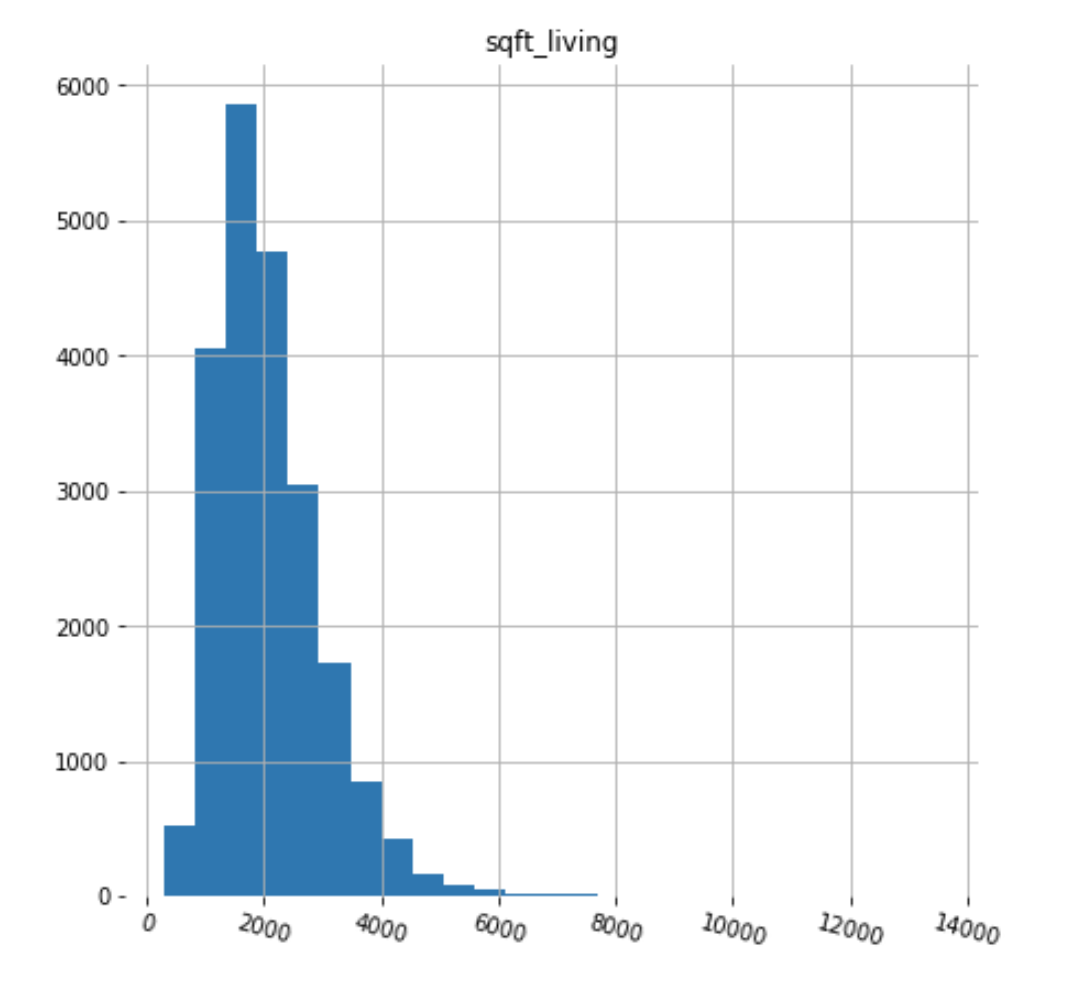


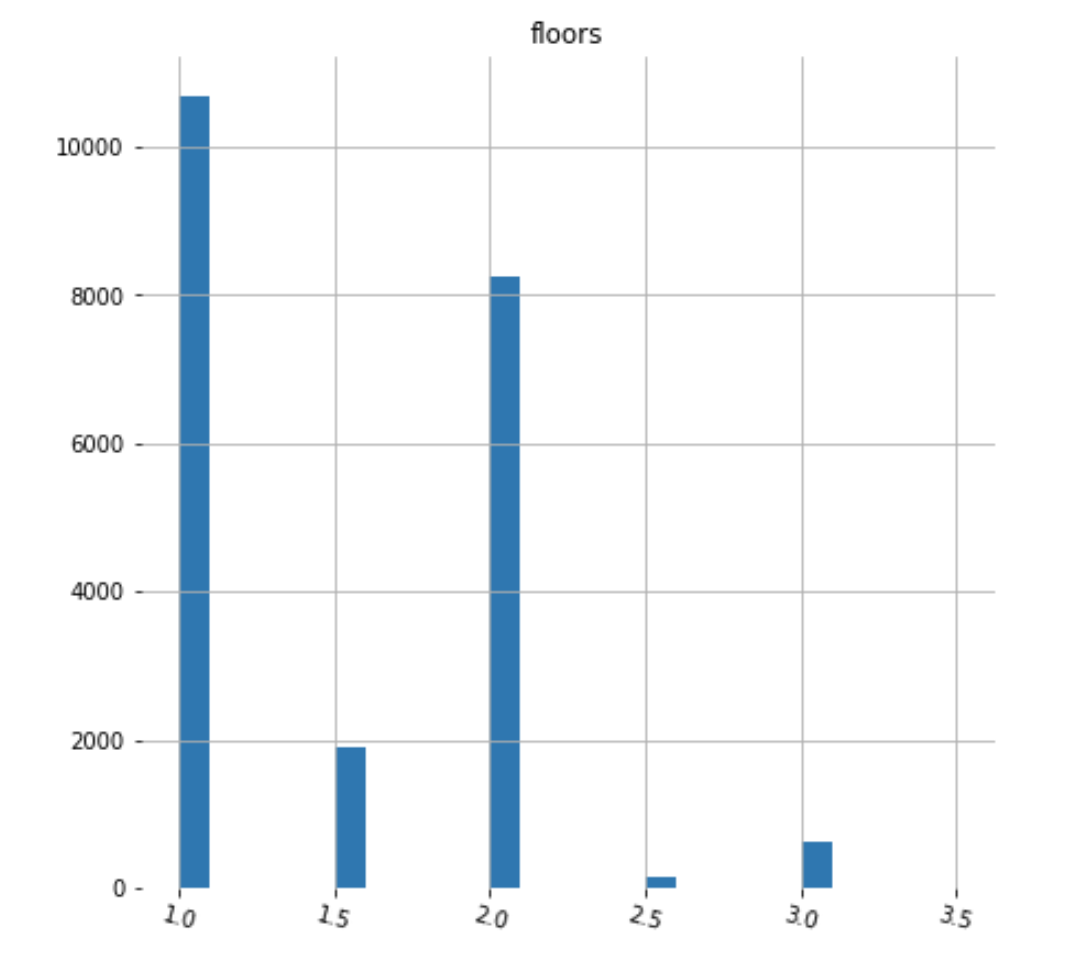
*It shows positive correlation between total square feet of the house with housing sales price*

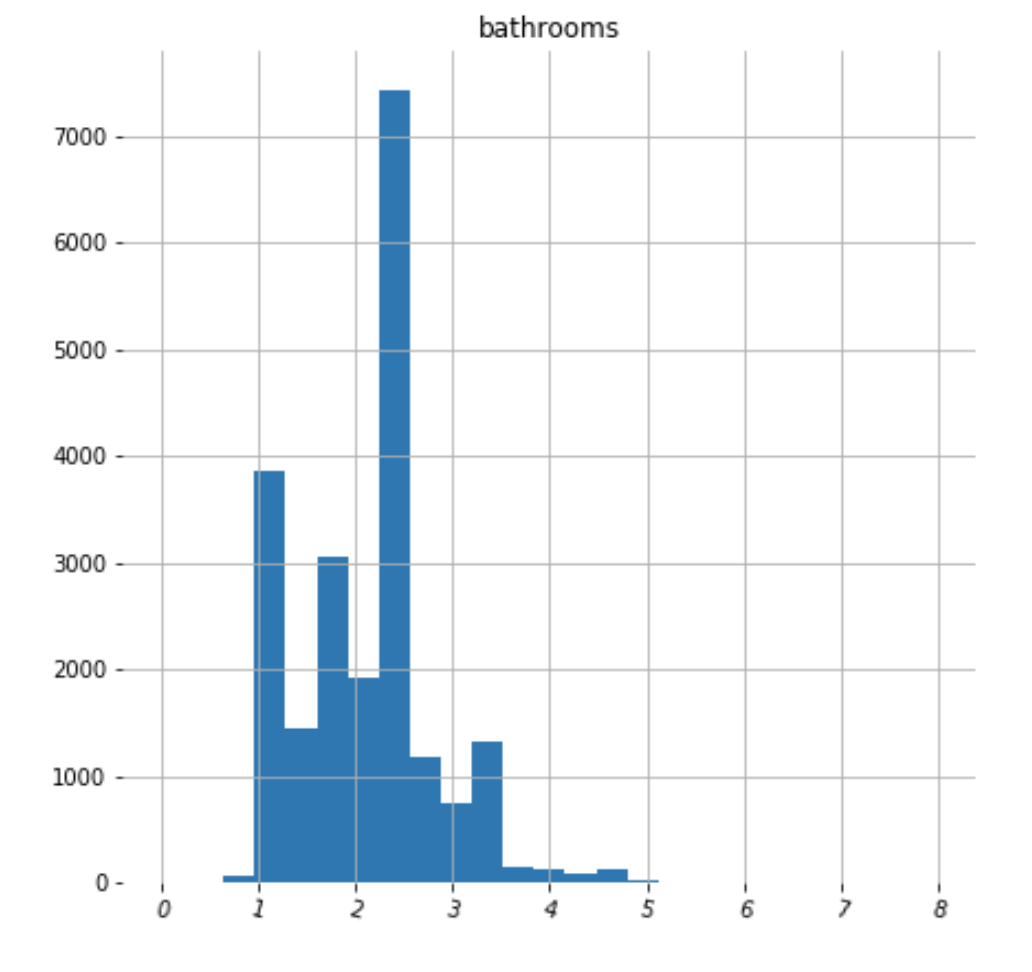
Later, I plotted all these features into histograms so I could understand these data better. A lot of the houses in King County has two bathrooms and many are built around 2,000.

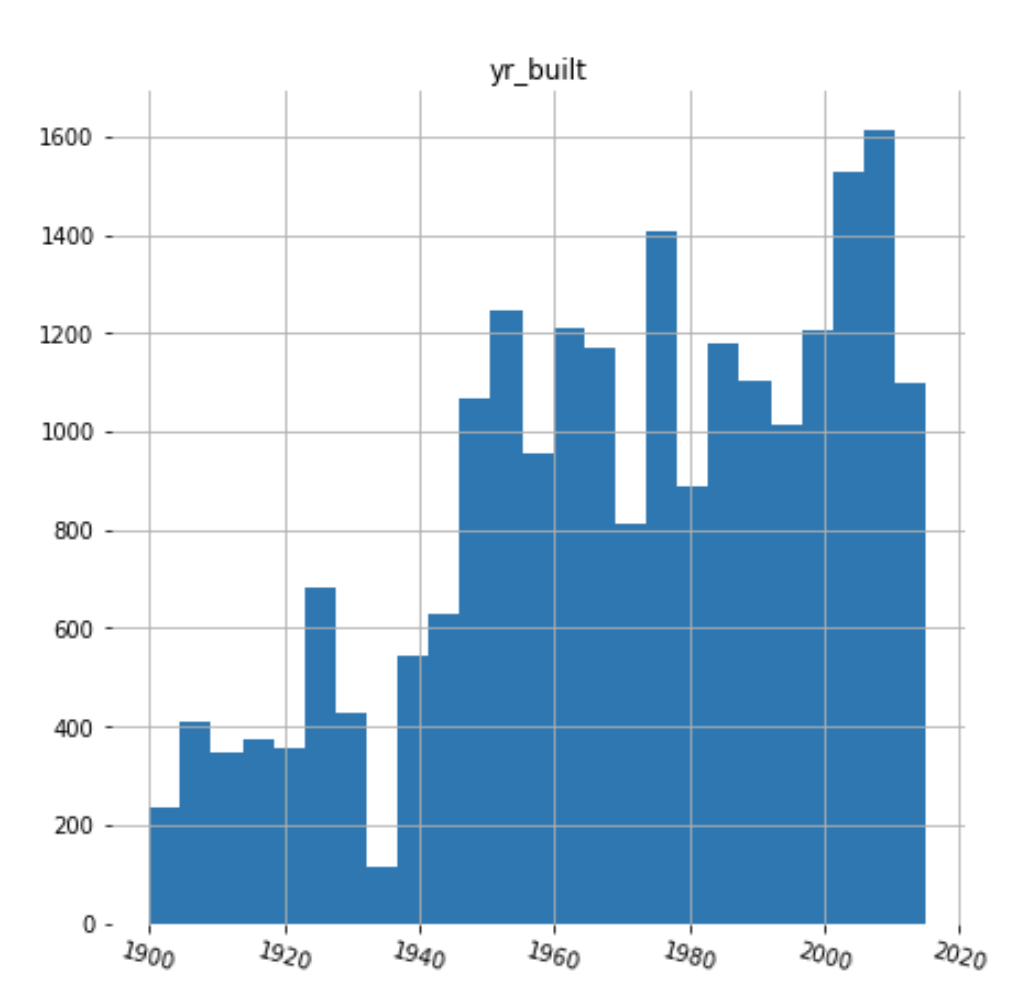




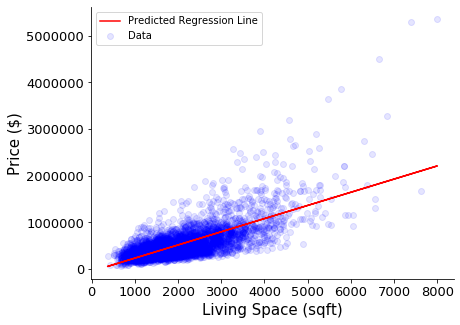








First, I tried to create a simple linear regression model using the single feature, total living space, to predict its housing price. I splitted the dataset into two, eighty percent of them into training data and remaining twenty percent for the testing purpose. The trained linear regression models show 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. 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*

And then I tried to use 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. This multi-linear regression model shows a y intercept of negative 57221293 and several coefficients. It has a 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 perform slightly better than the linear regression model mentioned earlier.

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

At the end, I created another multi-regression model but this time, I included additional columns. I applied binning to built year and renovated year, this way, the built year column is 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. This model has a mean squared error of 191879 and R2 score of 0.7131. It performs better than all the previous models.