A Report on

King County House Price Prediction

Github Link to Code: https://github.com/aman1608/Python/blob/master/King%20County%20House%20Price%20Prediction

Purpose of Study:

The Real Estate Sector is one of the most important contributing factors in a country's economic prosperity. Every year billions of dollars of Bank Credit are diverted to this sector. As was shown by the 2008 financial crisis, it is essential for the economy that the amount of a bank's real estate lending matches the actual value of the real estate. Getting a good estimate of the price of a house is hard even for the most seasoned real estate agents.

There is an obvious benefit to building a data-driven decision support tool for both banks and real estate agents, and such tools have been around for decades. They have typically used historical sales data to track prices in individual neighbourhoods and from that get average prices. With the advent of deep learning it is now possible to get a much more sophisticated valuation as we can now use other data types — such as images.

Objective:

To predict the sale prices for the houses in King County, which includes Seattle. The dataset includes homes sold between May 2014 and May 2015.

Data Source:

https://data.kingcounty.gov/

Data Structure:

The Dataset has 21 features and 21613 observations.

Methodology:

The dataset has been split into train, validation and test, with the Test data having 2217 observations, as against train and validation data having 9761 and 9635 observations respectively. We train the model on the Train data. Evaluate several models on the Validation data. The final model is then used for predicting the target variable (Price) for the Test data.

Tools Used :

Python

Libraries Used:

Pandas, Numpy, Seaborn, Matplotlib.Pyplot, Ppscore, Sklearn, Warnings.

Evaluation Metrics Used:

R-squared value (R^2) and Root Mean Square Error Score (RMSE).

EDA & Anomaly Detection:

Exploratory Data Analysis was conducted on the train dataset. We observed the following structure of the datasets provided:

- The train data has 9761 rows and 21 columns.
- The validate data has 9635 rows and 21 columns.
- The test data has 2217 rows and 21 columns.

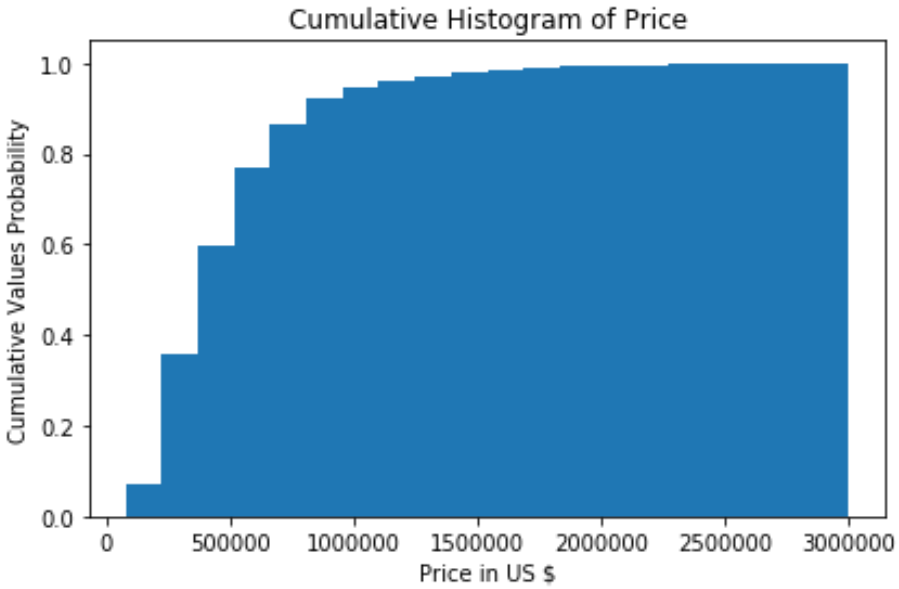
The summary statistics of the train dataset can be found in the github link mentioned above.

An Anomaly was found in the bedrooms variable, where a house had 33 bedrooms and 1.75 bathrooms. This record was changed to have 3 bedrooms and 1.75 bathrooms.

Univariate Analysis:

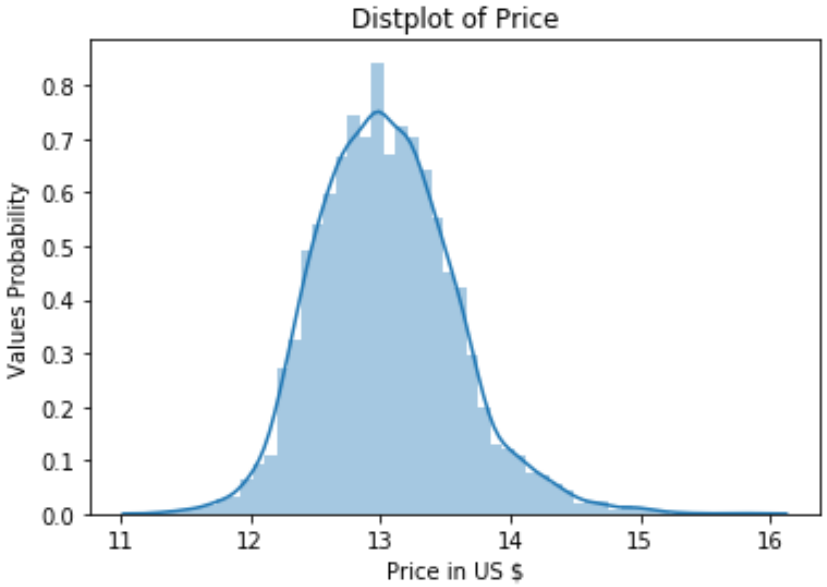
We analysed the following variables:

1. *Price*:



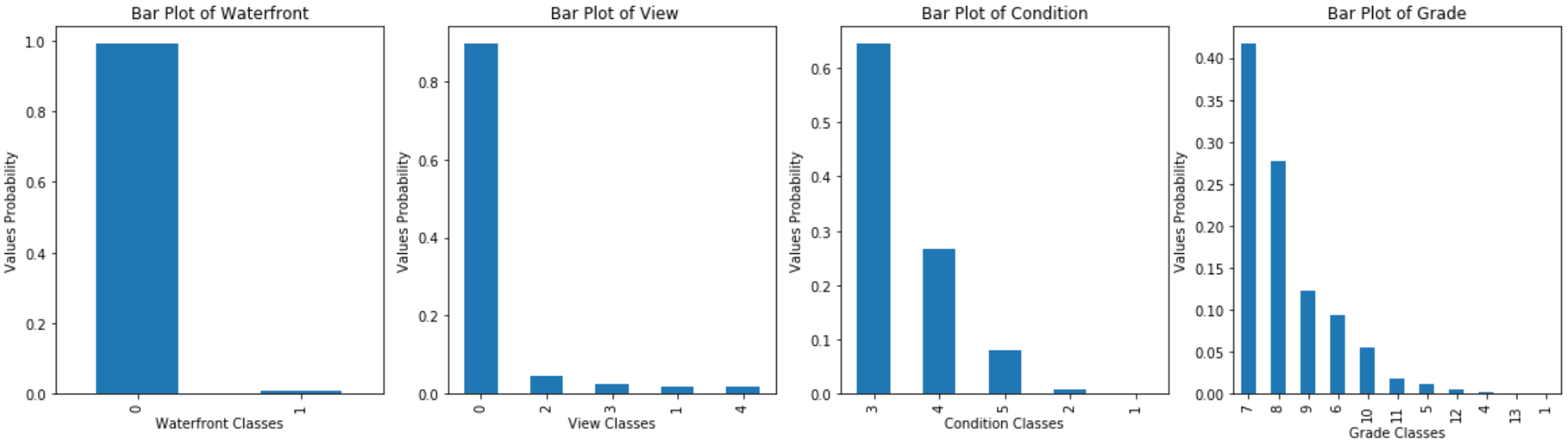
On plotting the Histograms, we observe that approximately 90% of the houses in the train data are priced less than or equal to US$1 million. Further, the price variable (target) is skewed to the right. We will need to log transform this variable so that it becomes normally distributed.

A normally distributed (or close to normal) target variable helps in better modelling the relationship between target and independent variables. In addition, linear algorithms assume constant variance in the error term. Constant Variance here means that when we plot the individual error against the predicted value, the variance of the predicted error value should be constant. Alternatively, we can also confirm this skewed behavior in the skewness metric.



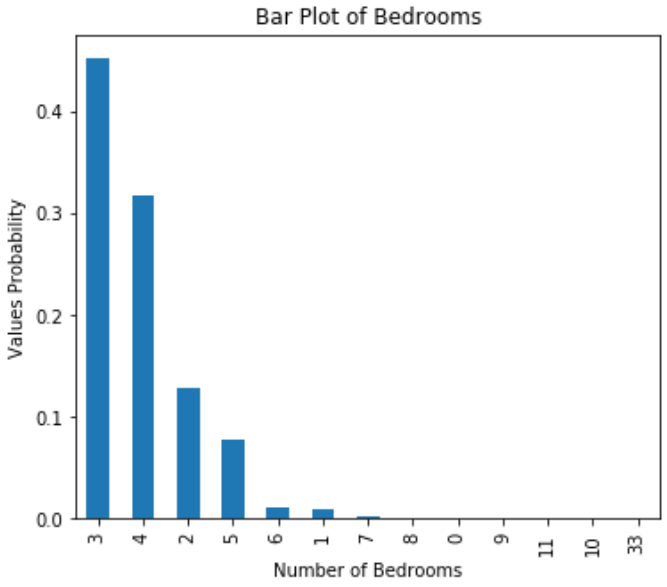
The Skewness of price was observed to be 0.4583. The Log transformation helped us to fix the skewness of the Target Variable, and it now looks closer to a normal distribution.

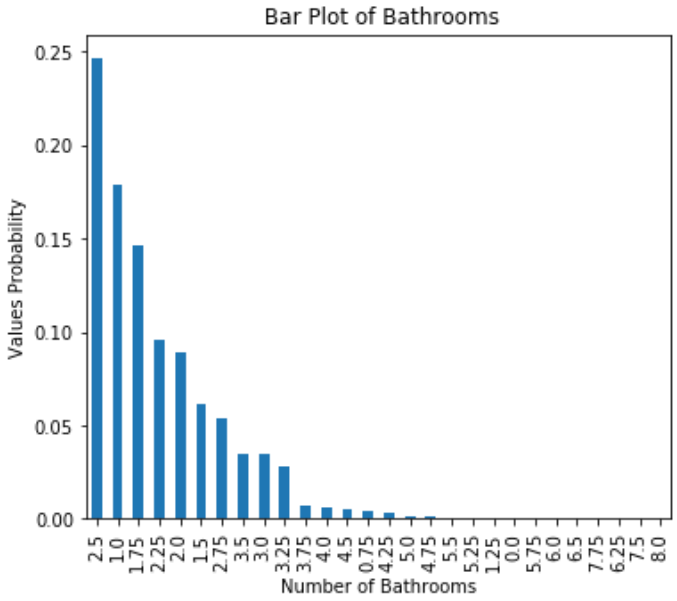
2. *Categorical Variables*

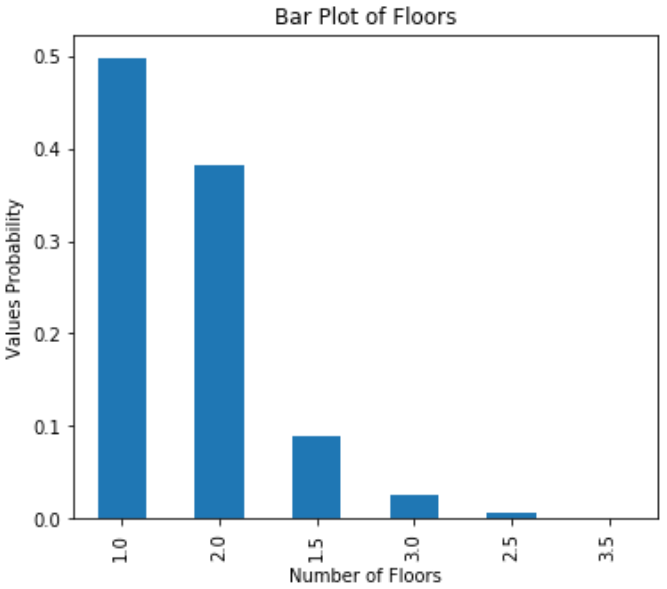


- The Bar plot of Waterfront shows that 99% of the houses in the train dataset do not have a waterfront.
- The Bar plot of View shows that approximately 95% of the houses in the train dataset have a view class of 0. Since this variable can have any meaning, we really cannot make any assumptions based on its values. Hence, we will ignore this variable.
- The Bar plot of Condition shows that approximately 65% of the houses in the train dataset have a Condition class of 3, followed by Condition Class 4 with 28% of the houses.
- The Bar plot of Grade shows that approximately 43% of the houses in the train dataset have a Grade class of 7, followed by Grade class 8 with approximately 28% of the houses.

3. Discrete Numerical Variables

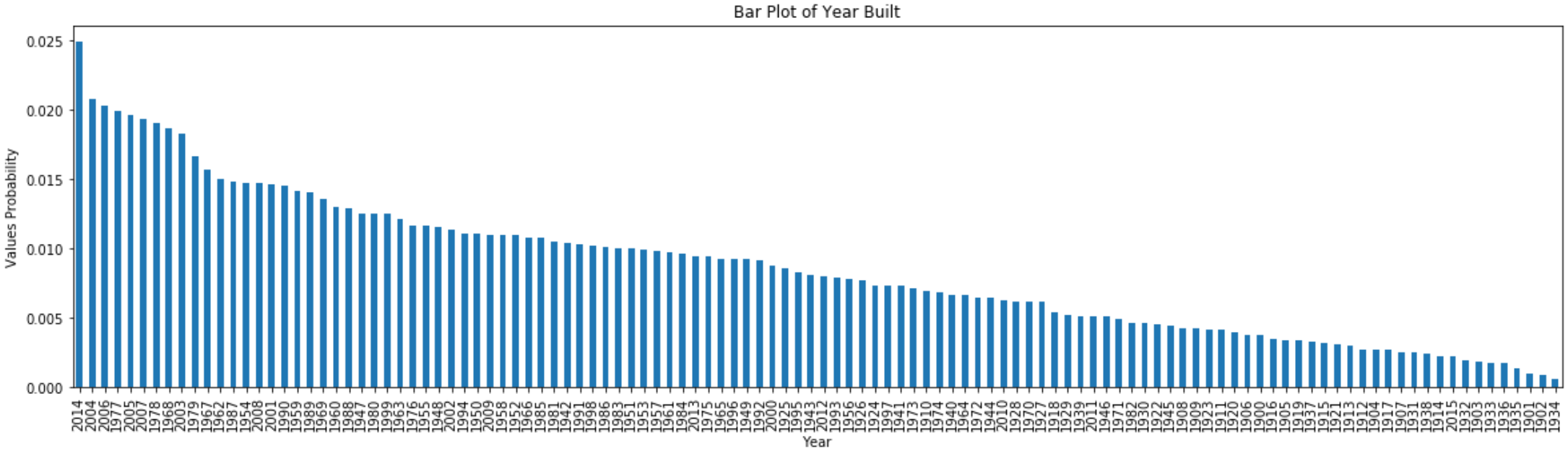






- More than 45% of the houses have 3 bedrooms, followed by approximately 32% of the houses having 4 bedrooms.
- 24% of the houses have 2.5 bathrooms, followed by 14% of the houses having 1 bathroom.
- There is 50% probability of a house being single storied.
- Let us check the skewness of these variables, and see if we need any transformations.

4. *Year Built*

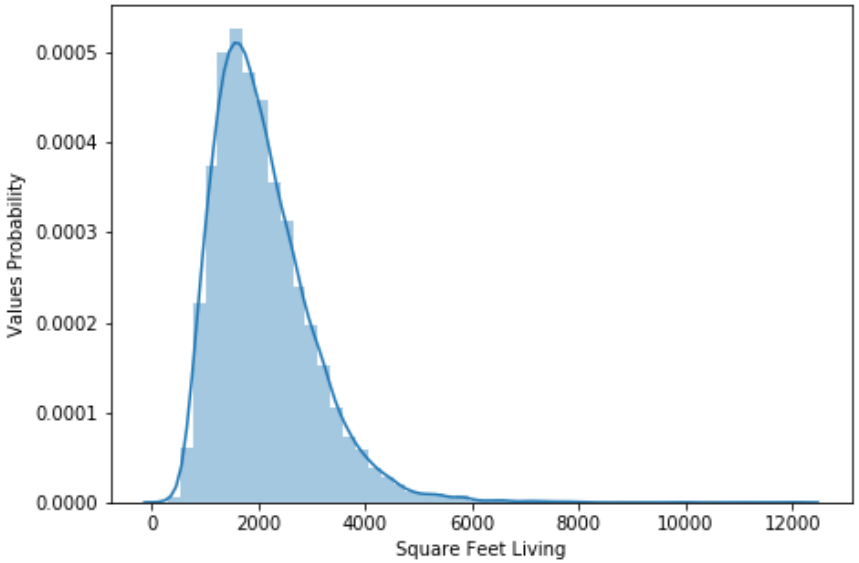


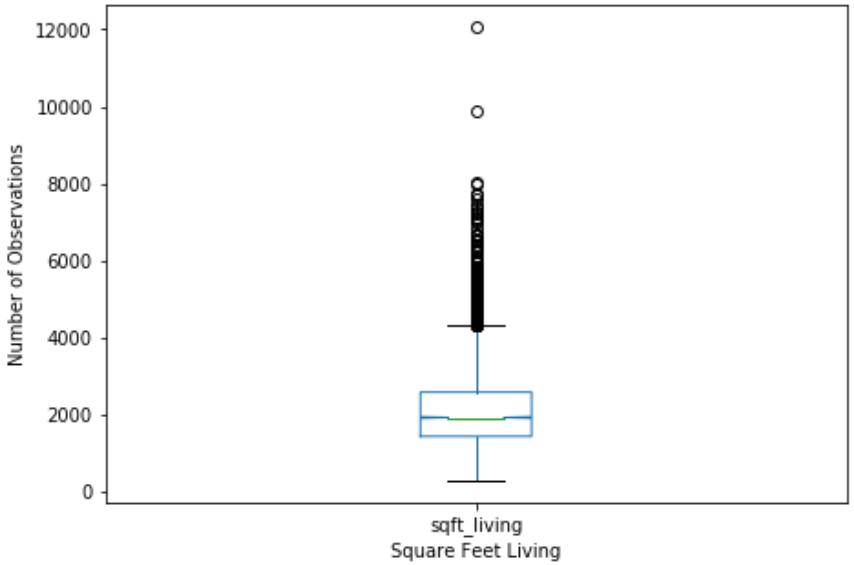
As we see above, the Year Built Variable is skewed to the right. 25% of the houses sold are built in 2014. The least number of houses sold are those that were built in 1934. Intuitively, a newly built house should have a higher price. We will see that during the Bivariate Analysis.

5. *Year Renovated*

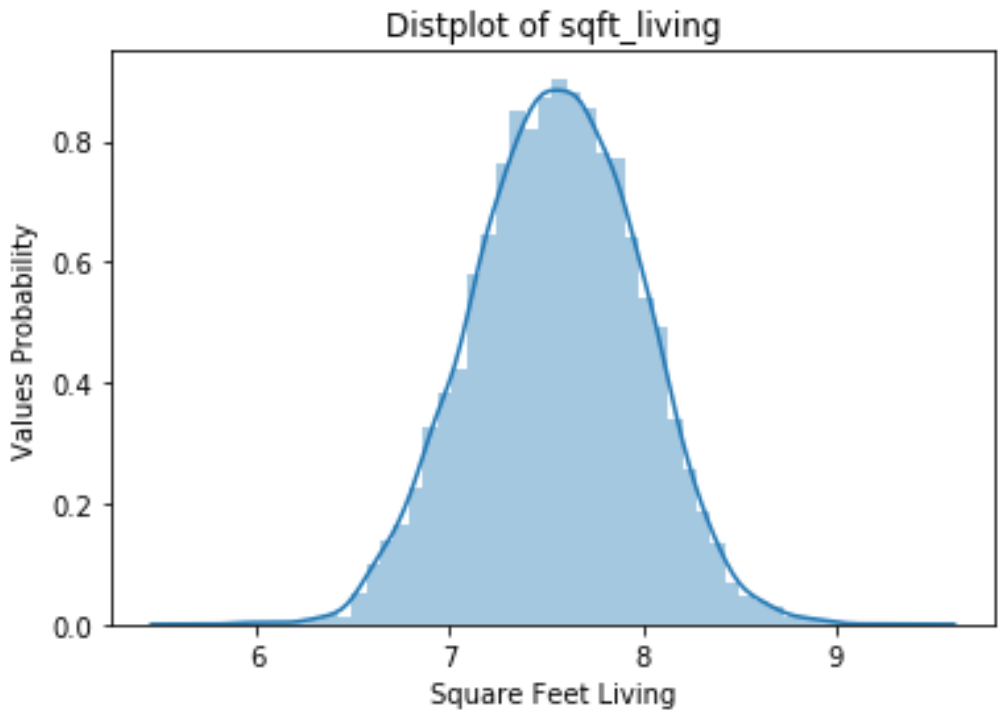
We see that approximately 96% of the houses were not renovated. Hence, we don't see much significance of this variable.

6. Square Feet Living



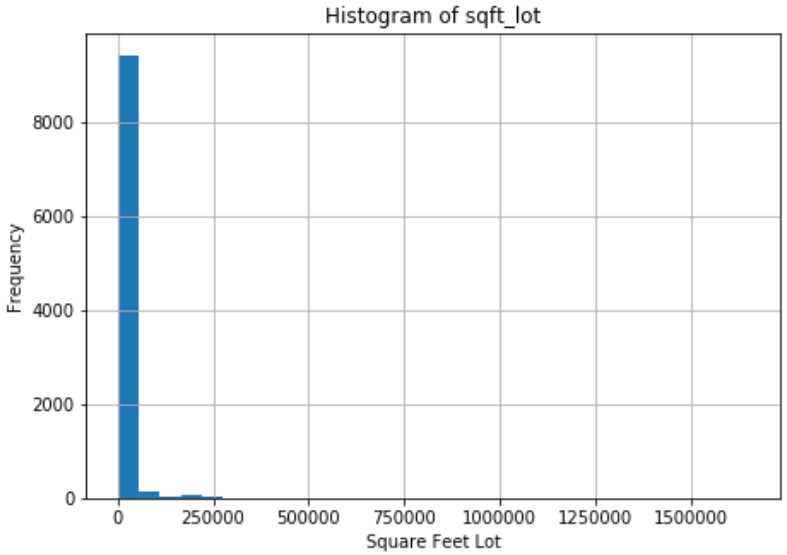


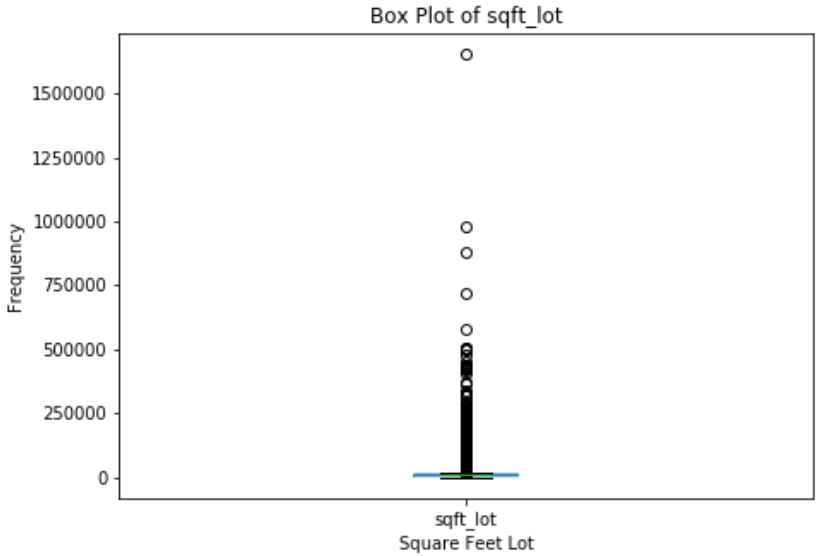
Most of the houses in our train dataset have a square feet living area less than 4000 square feet. The box plot shows some outliers. The median square feet living area is around 2000 square feet. As computed earlier, the mean square feet living area is 2084 square feet approximately. We observed a Skewness of 1.4258 in the variable. So, we log transformed the variable.



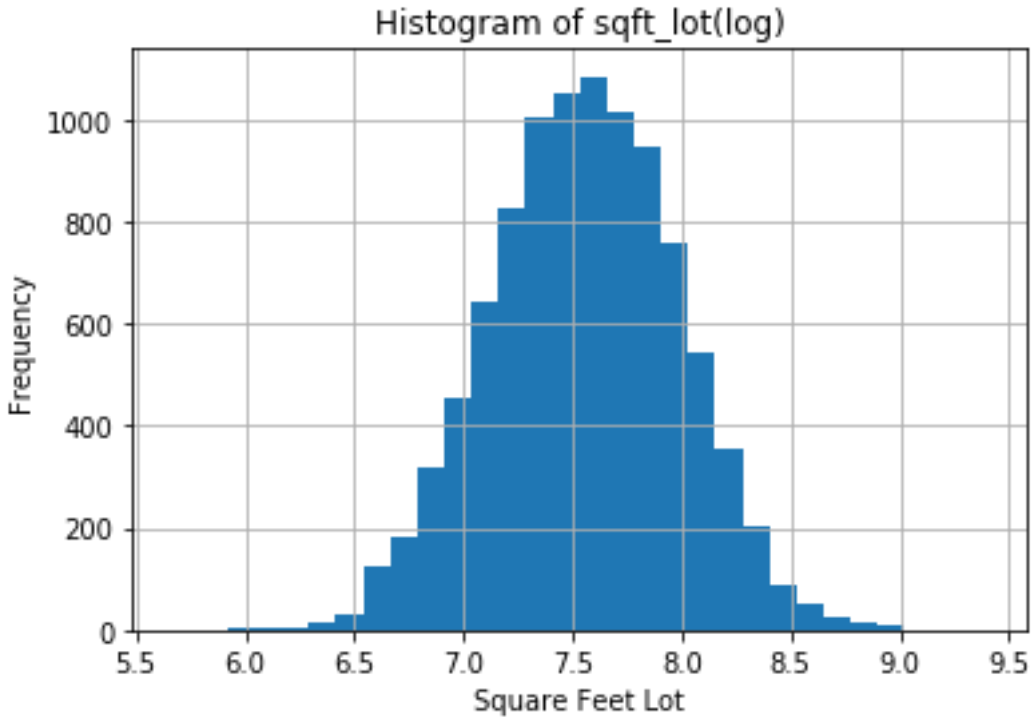
Now the distribution looks much closer to normal and the effect of extreme values has been significantly subsided.

7. Square Feet Lot



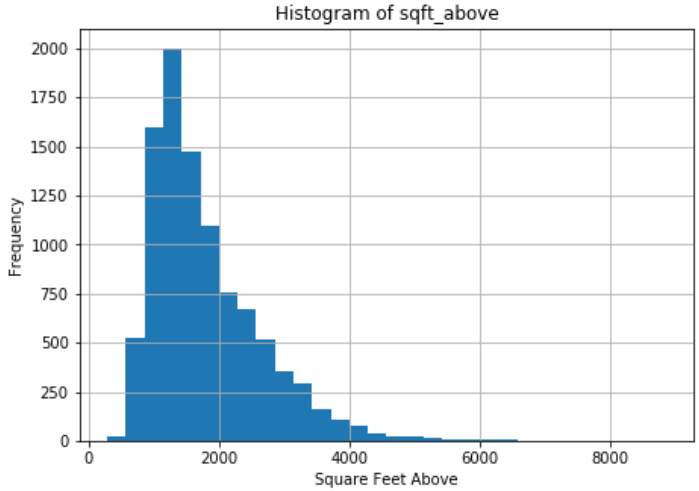


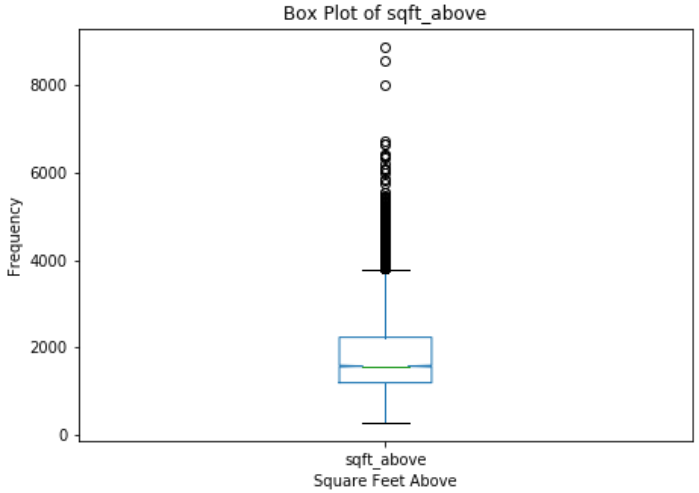
As we see above, most of the houses in our dataset has Lot size of less than 62,500 square feet. There do exist some outliers in the variable. We observed a Skewness of 13.6039 in the variable. So, we log transformed this variable.



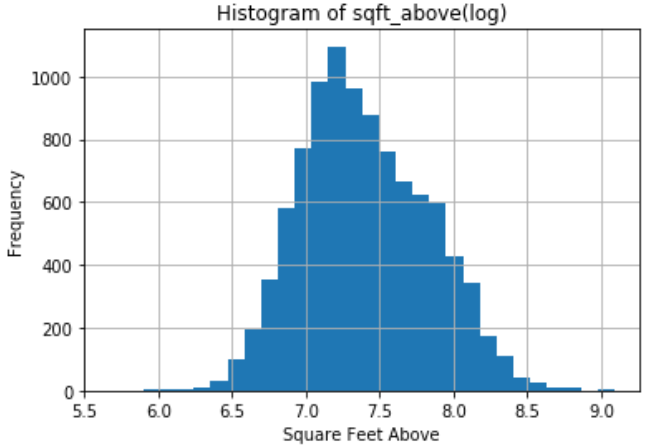
Now the distribution looks much closer to normal and the effect of extreme values has been significantly subsided.

8. Square Feet Above



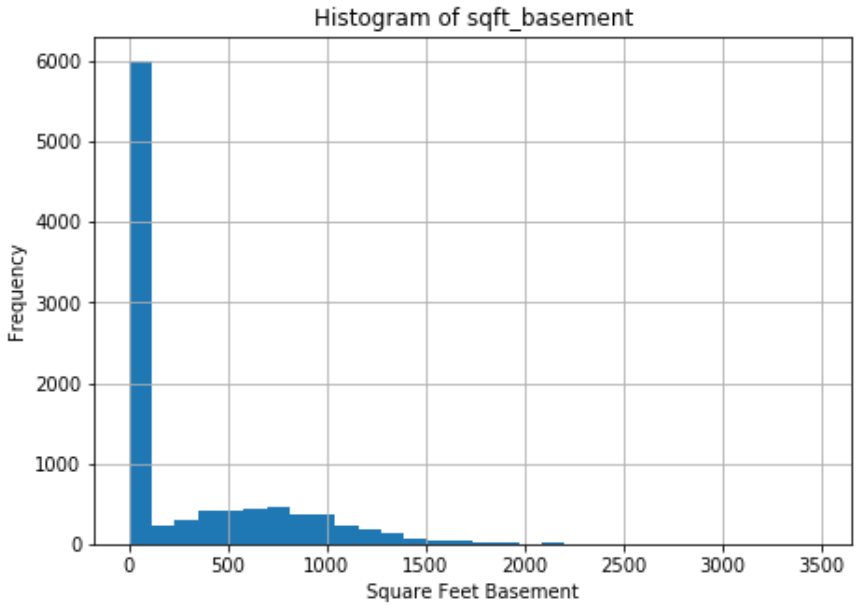


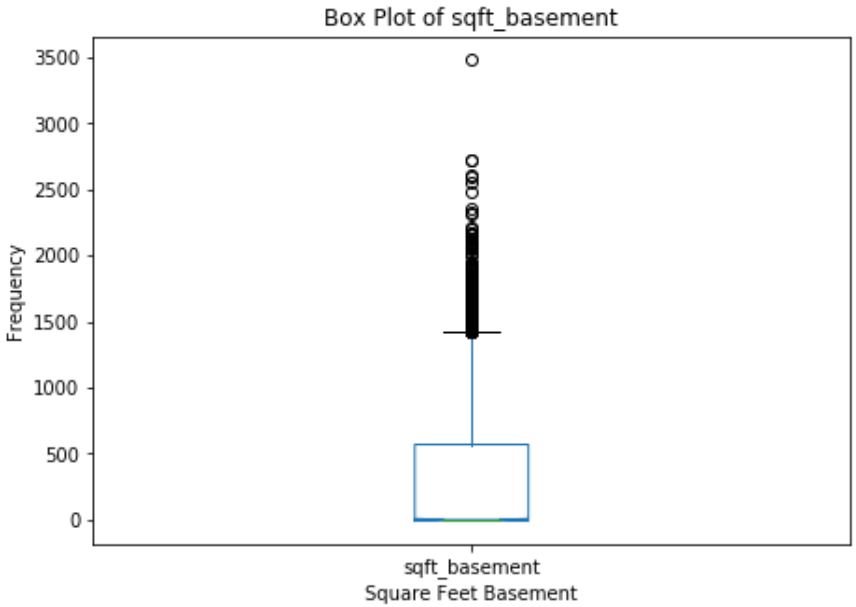
We observed a Skewness of 1.4527 in the variable. So, we log transformed the variable.



Now the distribution looks much closer to normal and the effect of extreme values has been significantly subsided.

9. *Square Feet Basement*

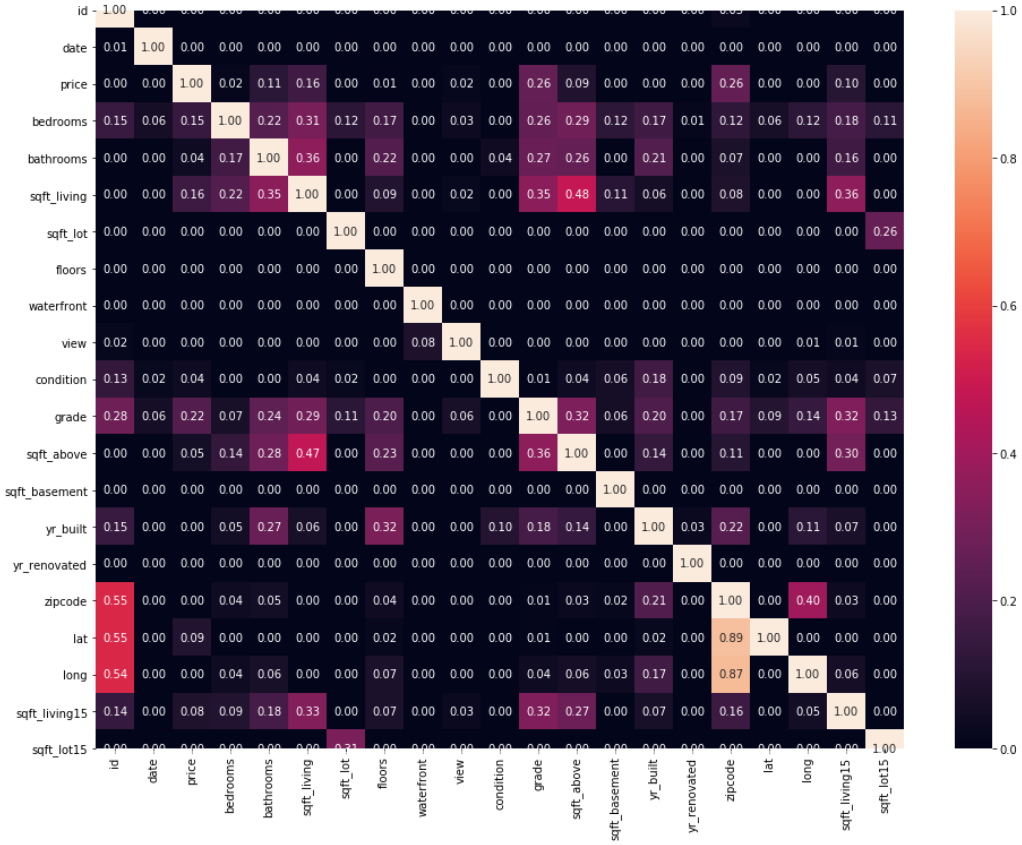




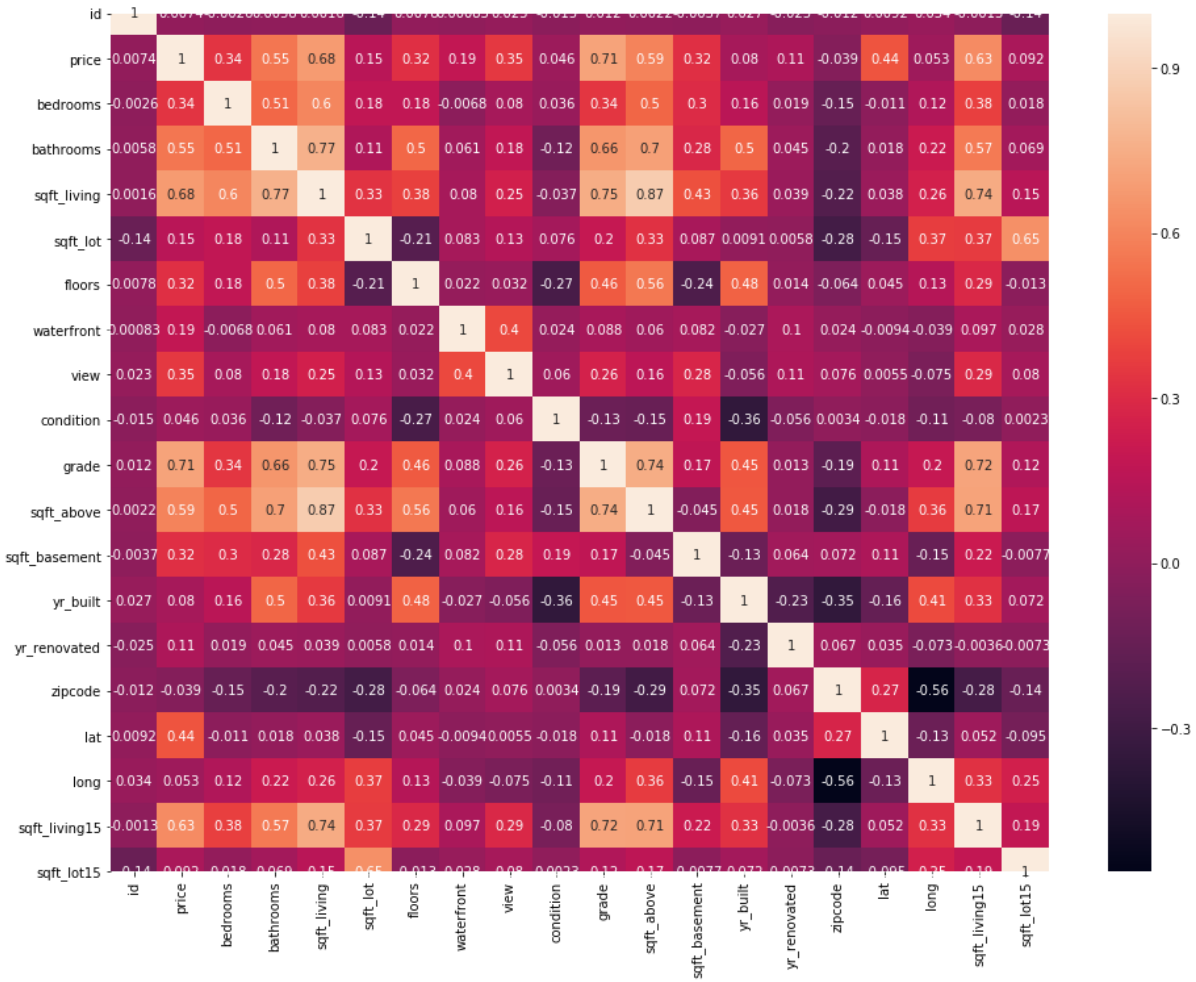
The Square Feet Basement variable looks skewed to the right. But log transformation is not possible, as most of the houses do not have a basement, and thus have a 0 value for the variable.

Bivariate Analysis:

Before we proceed with the Bivariate Analysis, we checked the Predictive Power Score for the Variables. Let's say, we have two columns and want to calculate the predictive power score of A predicting B. In this case, we treat B as our target variable and A as our (only) feature. We can now calculate a cross-validated Decision Tree and calculate a suitable evaluation metric. When the target is Numeric, we can use a Decision Tree Regressor and calculate the Mean Absolute Error (MAE). When the target is Categoric, we can use a Decision Tree Classifier and calculate the weighted F1. It calculates the score on randomly selected 5000 rows at maximum. The PPS Matrix helps us to find non-linear relationships in the data.



We also plotted a Correlation Matrix to further get an understanding of the relationships.



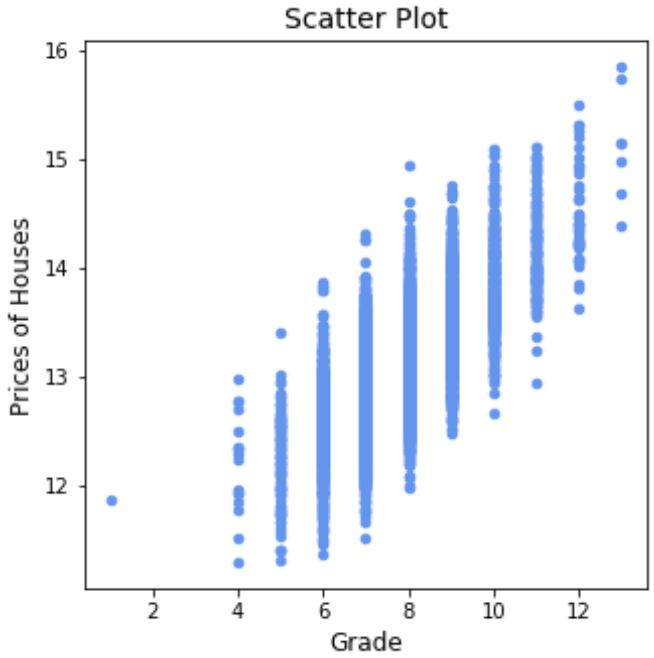
A score of 0.47 has been assigned for sqft_living & sqft_above. This is an obvious relationship. Let us explore the relationships as defined by the Corelation Matrix. A point to notice here is that bedrooms and price have a Pearson score of only 0.34. The PPS Matrix and the Correlation Matrix

shows that sqft_living and sqft_above are corelated with a Pearson score of 0.71 and a PPS score of sqft_living to sqft_above of 0.47. Hence, we will choose one of these variables in our model.

We cannot see any significant linear relationship between price and floors, condition, waterfront, year renovated and zipcode respectively. However, some linearity in relationship can be observed between price and latitude, longitude, sqft_lot, year built and bedrooms respectively.

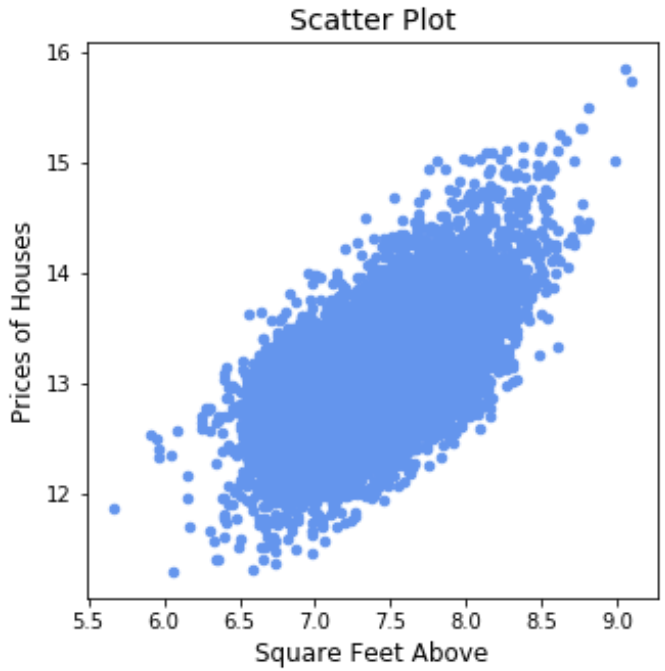
We will not use the latitude and longitude variable for modelling, as it will make the model complex. These variables should ideally be used in conjunction. Moreover, they have been covered under the zipcode variable, which does not have a significant linear relationship with price.

1. ***Price & Grade***



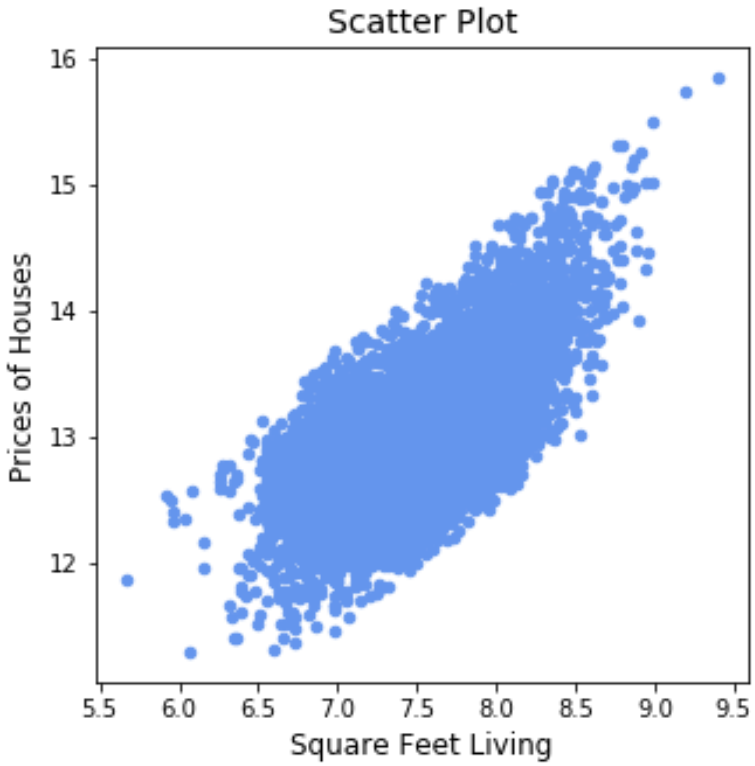
The scatter plot shows a positive correlation and linear relationship between the variables, with a Pearson score of 0.71.

2. *Price & Square Feet Above*



We observe that there is a positive correlation between the price of a house and its sqft_above. This also is intuitive, as higher the sqft area of the house, higher should be the price. A Pearson Score of 0.59 further strengthens the intuition.

3. Price & Square Feet Living



As in the case of sqft_above, sqft_living is also correlated to price, having a Pearson score of 0.68.

Final Model:

We created a Multiple Linear Regression Model with Price as the Dependent Variable and Grade, Square Feet Living, Square Feet Lot & Square Feet Above.

We got an RMSE of $241154.99 and a R^2 score of 53.89%.

The model summary is enumerated below:

Intercept	8.6888
Coefficient for Grade	0.2189
Coefficient for Square Feet Living	-0.2118
Coefficient for Square Feet Lot	0.5982
Coefficient for Square Feet Above	-0.2118

Interpretation of the Model:

The Intercept being 8.69, for every 1% increase in the log transformed sqft_living, while keeping the grade, log transformed sqft_lot and log transformed sqft_above constant, the log transformed price of the house increases by:

({(1 + 0.01)^0.59823638} - 1)*100 = (1.00597040213665811788838467504558 - 1)*100 = 0.5970 = 0.60%

The Intercept being 8.69, for every 1% increase in the log transformed sqft_lot, while keeping the grade, log transformed sqft_living and log transformed sqft_above constant, the log transformed price of the house increases by:

({(1 + 0.01)^(-0.02985227)} - 1)*100 = (0.99970300414878398640319631987213 - 1)*100 = -0.0297 = -0.03%

The Intercept being 8.69, for every 1% increase in the log transformed sqft_above, while keeping the grade, log transformed sqft_lot and log transformed sqft_living constant, the log transformed price of the house increases by:

({(1 + 0.01)^(-0.21181249)} - 1)*100 = (0.99789461508421724663744417942 44 - 1)*100 = - 0.2105 = -0.21%

The Intercept being 8.69, for every 1 unit increase in the grade, while keeping the log transformed sqft_living,log transformed sqft_lot and log transformed sqft_above constant, the log transformed price of the house increases by:

(e^0.21890008 - 1)*100 = (1.24470689936035931616778694960 95 - 1)*100 = (0.24471*100) = 24.47%

Conclusion:

The actual prices have a positive skew.

The R^2 of the final model wasn't very high: 53.89%.

It is very difficult to have a high R^2 value for a model with the given variables. Maybe adding in more variables to the data can help in predicting the House prices better. This goes with the intuition that the price of a house is not just dependent on the variables available in the dataset. The price of a house also depends on various external factors such as the prevailing economic situations in the region, personal bias, sentiments of the buyers and so on.