In figure 1, we can notice that about 98% house prices are below 2,000,000 dollars, 90% of which are below 1,000,000 dollars, which also means that people prefer to buy the house that prices are lower 1,000,000 dollars. Also, because there are numeral longitude and latitude values, our team are interested in geographical distribution of house prices in King County. Based on figure 2, we can find that the latitude of the prices beyond 39,9950 is almost higher 47.5. On the contrary, the latitude of the prices below 39,9950 is nearly lower 47.5. there is one more thing is that majority houses are locate the longitude over –121.5. Thus, we guess this area could be more suitable for living.

Moreover, we also observe the data set includes one column called “years of build”. Therefore, we want to know the tendency of the number of houses built in 1900 to 2015. From figure 3, we know that the tendency of the number of houses is upward based on linear regression. Meanwhile, we also observe that some years are decline trend such as in the period from 1926 to 1934, 1997 to 1982 and 2006 to 2011 and so on.

Normally, the house living area, condition, location will straightly affect the house prices. So, we also wonder the distribution of the house living area, lot area, sale price based on the number of bedrooms, bathrooms, conditions and so forth. In figure 4 and 5, we know that house prices rise with the increase in the number of bedrooms and bathrooms. This is reasonable. The more bedrooms and bathrooms, the larger the house size and the higher the price. Furthermore, we observe that when the number of bedrooms is greater or equal to 8, there is not any abnormal high price house. By contrast, as the number of bedrooms is lower 8, the houses show abnormal prices. Especially, some houses with 5 or 6 bedrooms are over 6,000,000. Same as the bedrooms, the more the bathrooms, the higher the prices. The average price of the houses with 8 bathrooms is highest. The house with 2 or 2.5 bathrooms is more common. From image 6, we find that the better the condition, the higher the grade and the price. Correspondingly, the worse the condition, the lower the grade and the price. When the live area is over 5000 square footages, the condition score is greater or equal to 3 and the grade is over 10. In picture 7, we notice that most houses have 2 floors, at least 4 bedrooms and 4 bathrooms when the live area is over 5,000 square footages. Also, majority houses have at least two floors, at least three bedrooms and 2 bathrooms.

Feature selection

After we explored house price attribute and its attributes other variable attributes. We want to know more what factors or variables have a greater impact on the house price. So, we are obliged to implement the feature selection to remove redundant from the dataset as so to select “perfect” subset and rank the variables in the dataset by their importance. We find that the house price distribution is heavily positively skewed, so we decide to perform a square root transformation on price so that the new “sqrt\_price” variable will be close to normal distribution that won’t cause too much data discrepancy and used as the main outcome variable in analyses. Firstly, we set a random seed to make the result can be repeated. Then, we use the repeated cross validation method to repeat 10 times resampling iterations in trainControl function for training our model. In the model, we want to learn the rank of the impact of each feature on the house price. According to the K- nearest neighbors' algorithm, we start to train our model and estimate each feature has a great influence on the price. From the figure 8, we can know that live space has a greatest impact on the price followed by grade and sqft\_living15. The lot\_zise and sqft\_15 is almost no effect for the house price. Thus, for our future analysis, we can rely on this information to research the top 10 variables how to affect the house price. The company also know how to increase the value of the house. For instance, the company can increase the house grade by select a good view and expand the living area.

Radom Forest

In this module, we divided the data into two subsets for the random forest model. One subset is training dataset to train the model. The other is testing subset to test the accuracy of model. Initially, we use the train dataset to build our random forest model. Meanwhile, we also would like to find the importance of each variable to house price. Based on the result of the “%IncMSE” and “IncNodePurity” values in random forest model, we will rank the variables by importance. Eventually, we are going to leverage testing dataset to determine how well the model performs on the test set. In picture 9, we notice that the “latitude” is most important to house price in Mean Decrease Accuracy (%IncMSE) chart. However, the “grade” in Mean Decrease Gini (IncNodePurity) chart is most important. The importance of each variable is distinct. Another interesting thing we noticed is that the number of bedroom and bathroom have a little impact, which is conflict with our expectation. We believe that the more the bedrooms and bathrooms, the higher the price or the more influence. Furthermore, from image 10, we can know that the MSE value of the model is 4370.372, which means that this model has a good performance in the testing dataset. The company can use it to predict the house price or assist company making decision.