**Linear Regression Model**

To build the linear regression model, we remove the unnecessary variables including ids, dates, zipcodes, latitudes, etc. We also calculate the age of the houses by subtracting the year of built from the current year 2018. Afterwards, we regress against the price with the left variables.

During our first attempt to build the linear regression model, we find out some variables are linearly dependent on the others. For example, the square foot of living = the square foot of basement + the square foot of upper areas. We delete the items that can be interpreted as linear combinations of the others, and regress again.

During the second attempts, we found some variables are not significant based on t-test and partial F-test, such as the square foot of living area of the 15 nearest neighbors. We remove those insignificant variables, and regress on the remaining ones, including the number of bedrooms, the number of bathrooms, the square foot of lots, the floors, the conditions, the grades, the square foot of upper areas, the square foot of basements, and the ages of the houses.

As the summary of linear regression model demonstrates, all of those variables are significant by t-test. Then we run a F-test on the general adequacy of the model, which rejects the null hypothesis that the linear relationships does not exist. Afterwards, we run F-tests on the difference of sum of squares due to regression between the full model and the partial model with one variable deleted. The results also show that all of those variables significantly contribute to the final model.

The R square of the model is 61.82821%, not very optimal, so we tried several transformations on the data, such as logarithm and quadratic, but none of them generate better results.

Thus, we pick up the first-order linear regression model with coefficients such as

(Intercept): -1.088274e+06

Bedrooms: -4.903076e+04

Bathrooms: 4.968293e+04

sqft\_lot: -2.263292e-01

floors: 2.876962e+04

condition: 1.882481e+04

grade: 1.326334e+05

sqft\_above : 1.810695e+02

sqft\_basement: 2.029833e+02

age: 3.964099e+03

Next, we calculate the residuals, standardized residuals, studentized residuals, Press residuals, and the R-student residuals, and plot the normality probability, the R-student residuals against fitted values, and the R-student residuals against each regressors.

The normality probability plot is negatively skewed, but the all the other plots do not demonstrate severe normality problems.

We make conclusion that the linear regression model is adequate but not optimal compared to other non-linear model.