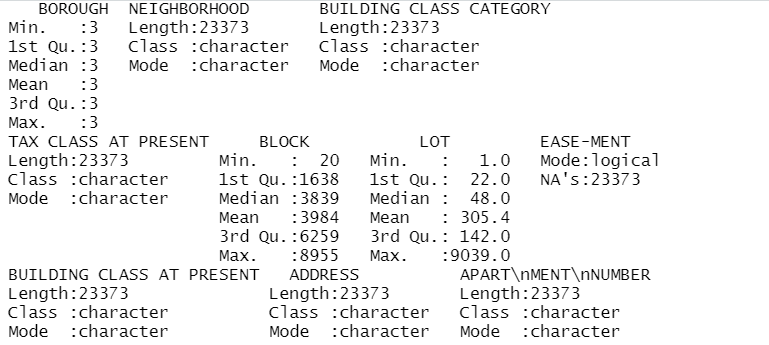
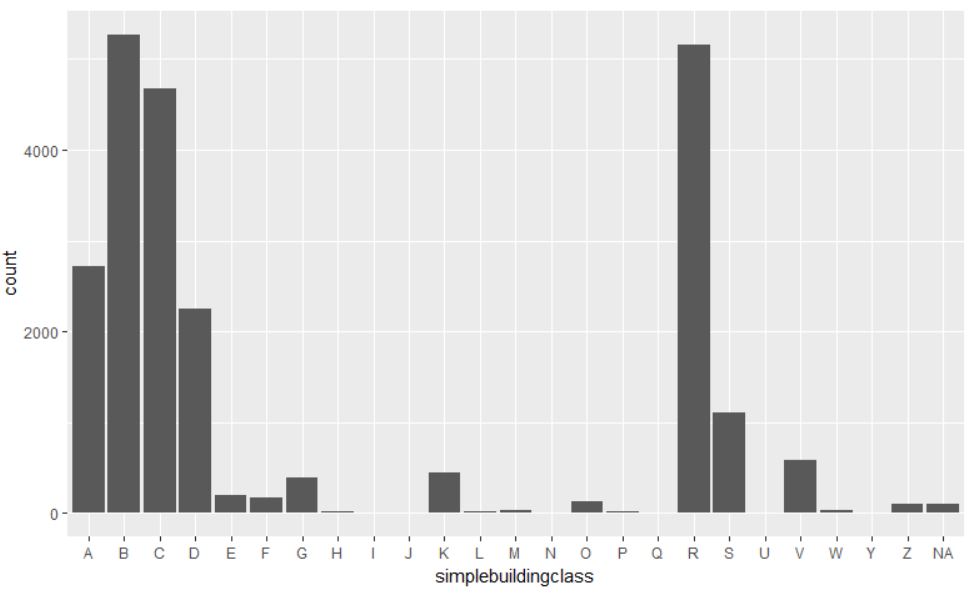
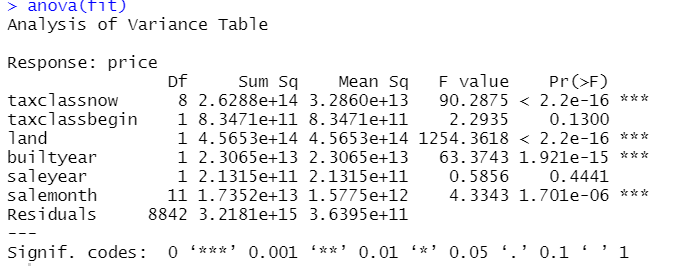
Question 1:

a. First of all, I use the summary function to see each variable infomations. Following image is part of results: 

Then I check the “NEIGHBORHOOD”, there are 63 unique values in neighborhood, such a big set of values. The “TAX CLASS AT PRESENT” has only 10 unique values and its name imply it may have relationship with final price. The “BUILDING CLASS” has data with letter and number, so I may subset the letter or number to create new variables. 

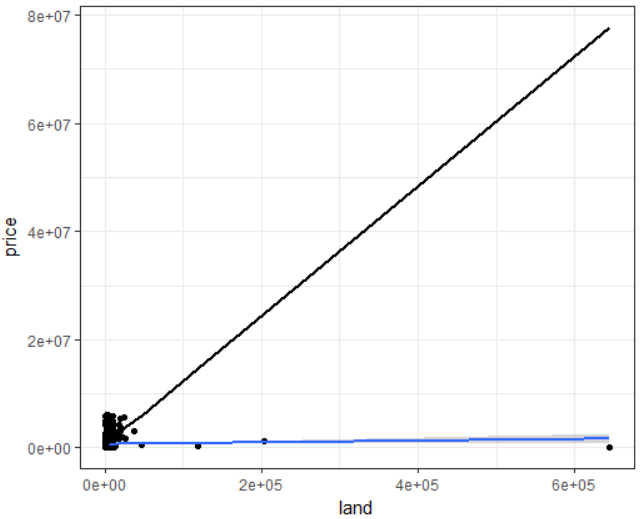
So in the plot of first letter of building class, there are some letters have more buildings than others.

b. I choose the multivariate regression. The independent features are tax classes, land square feet, built year, sale date from which I extract year and month. I exclude bottom 1% and top 1% price for outliers, the max value of origin price data is too big to mitigate its influence to linear regression. Also, I choose built year late than 1800 for there are many building have built year of 0, which is definitely wrong data.

Following image is my anova analysis of fit model.

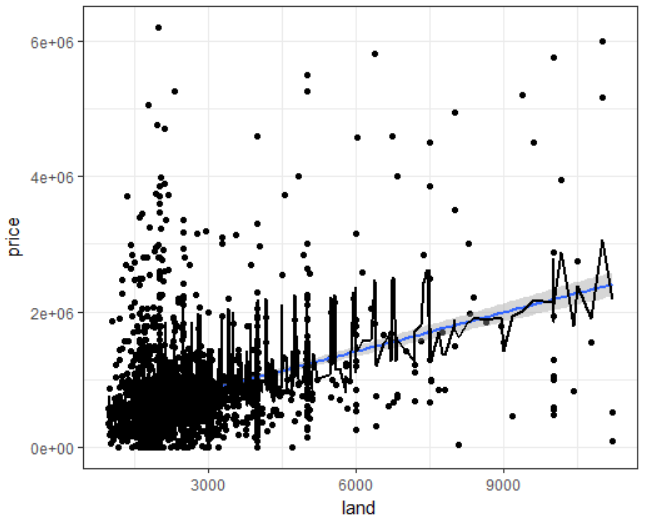
The tax class now, land area, built year and sale month have strong indicate of sale price.

Question 2:

a. 

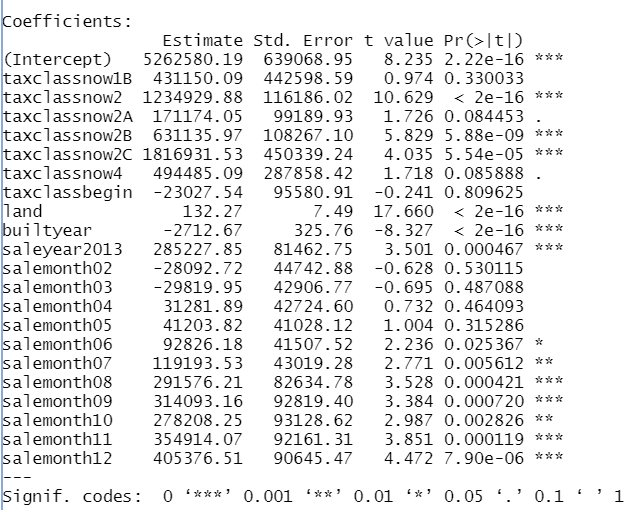
due to the outliers in land (data have land over 10000 the actual line and predict line have a huge gap. Then I choose data which have land in 1% to 99% and exclude the 0 data before.

The update image is following.



Now the prediction have a far less RMSE than model before (622k to 1927k).

B.The significance can calculate easily via summary function.



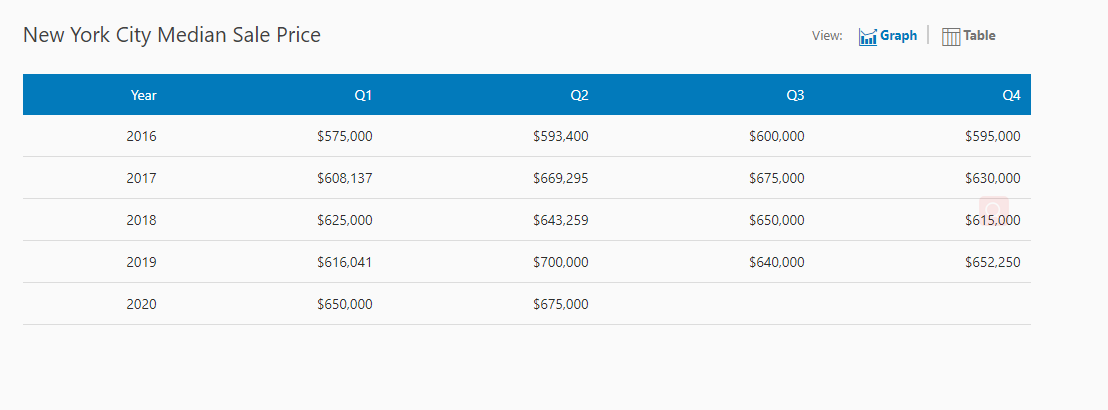
The tax class have relative strong prediction power, also the sale month have influence to the final price. The built year have strong negetive relation to the final price, which is easily understand.

C. From my perspective, the zipcode and block number can help predict the price. The different zipcode and block have different environment that are important to house price, like surronding hospitals and schools.

Question 3:

From the significance table, we can find the quarter 3 and 4 have strong positive influence on price. I find this is interesting so I search for history NYC real estate market data. So I find some median data of each quarter in past few years. However the data don’t suggest the quarter 3 & 4 will show a high price. This different of model and real data may caused by dataset itself (e.g. we only use the 2013 and 2014 brooklyn data) or the outlier in the data.

New York City Median Sale Price[[1]](#endnote-1)



1. propertyshark https://www.propertyshark.com/mason/market-trends/residential/nyc-all [↑](#endnote-ref-1)