HW4

**For this HW, I convert the excel version of the dataset to CSV file. In addition, during the data analytics, I find the read.csv will read number with 1000 separator as a factor/character. So I disable the use of 1000 separator in excel for all the numbers in the dataset.**

**CODE:**

ny <- read.csv('C:/Users/dingj/Desktop/data\_analytics/hw4/rollingsales\_manhattan.csv',, stringsAsFactors = FALSE,skip=4)

attach(ny)

summary(`SALE

PRICE`)

boxplot(`SALE

PRICE`)

plot(`TAX CLASS AT TIME OF SALE`,`SALE

PRICE`)

plot(`YEAR BUILT`,`SALE

PRICE`)

plot(`YEAR BUILT`,`SALE

PRICE`)

plot(`GROSS SQUARE FEET`,`SALE

PRICE`)

plot(`SALE DATE`,`SALE

PRICE`)

new = apply(ny[c(15,16,17)],1,function(z) all(z!=0))

ny2 <- ny[new,]

new2 = apply(ny2[c(20)],1,function(z) all(z!=0))

ny2 <- ny2[new2,]

train\_ny2 <- ny2[1:2000,]

attach(train\_ny2)

mm <- lm(SALE.PRICE~GROSS.SQUARE.FEET+YEAR.BUILT+ZIP.CODE+TAX.CLASS.AT.TIME.OF.SALE+LAND.SQUARE.FEET)

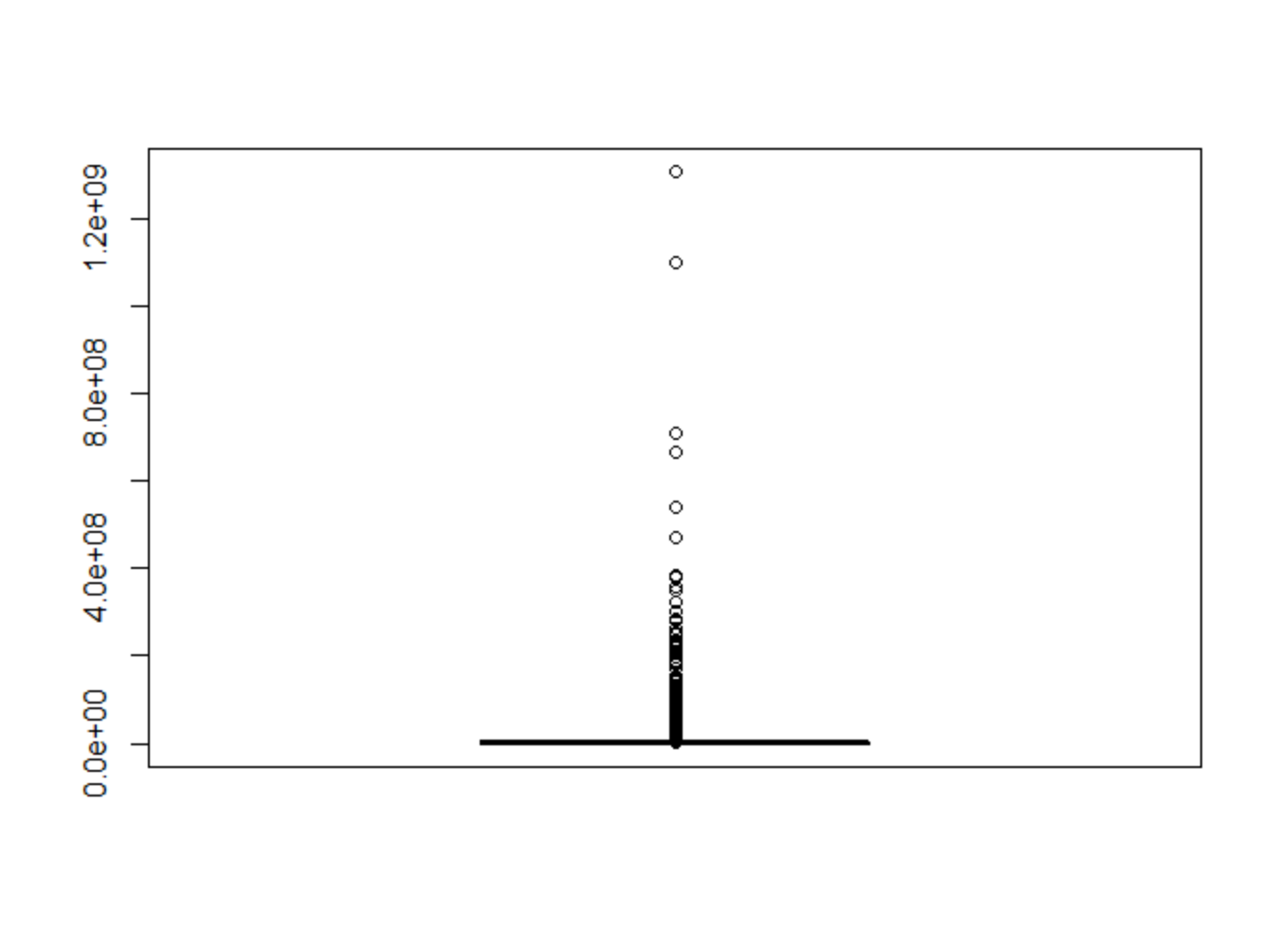
newdata <- data.frame(GROSS.SQUARE.FEET=2465,YEAR.BUILT=46200,ZIP.CODE=10034,TAX.CLASS.AT.TIME.OF.SALE=4,LAND.SQUARE.FEET=14980)

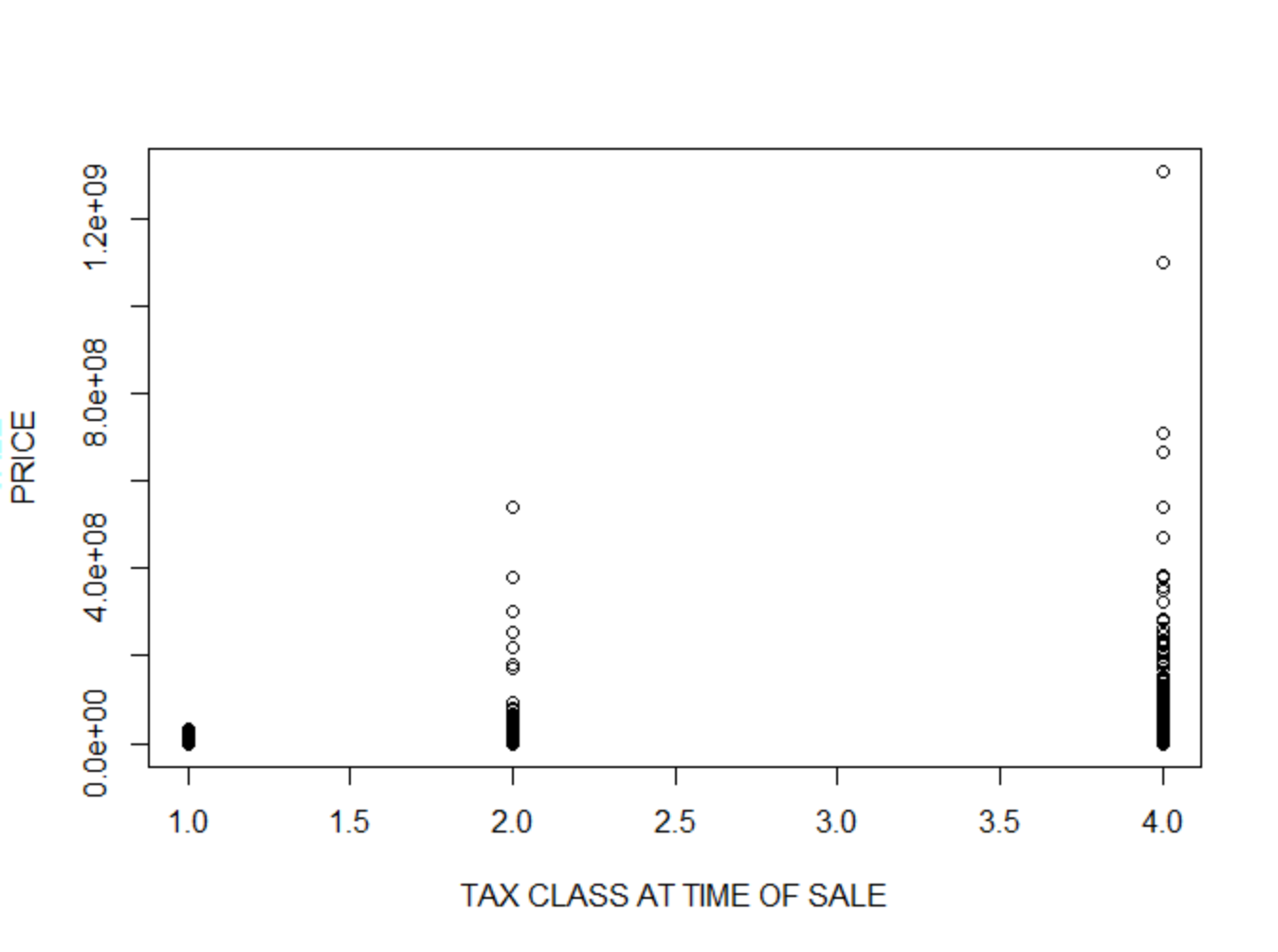
predict.lm(mm,newdata)

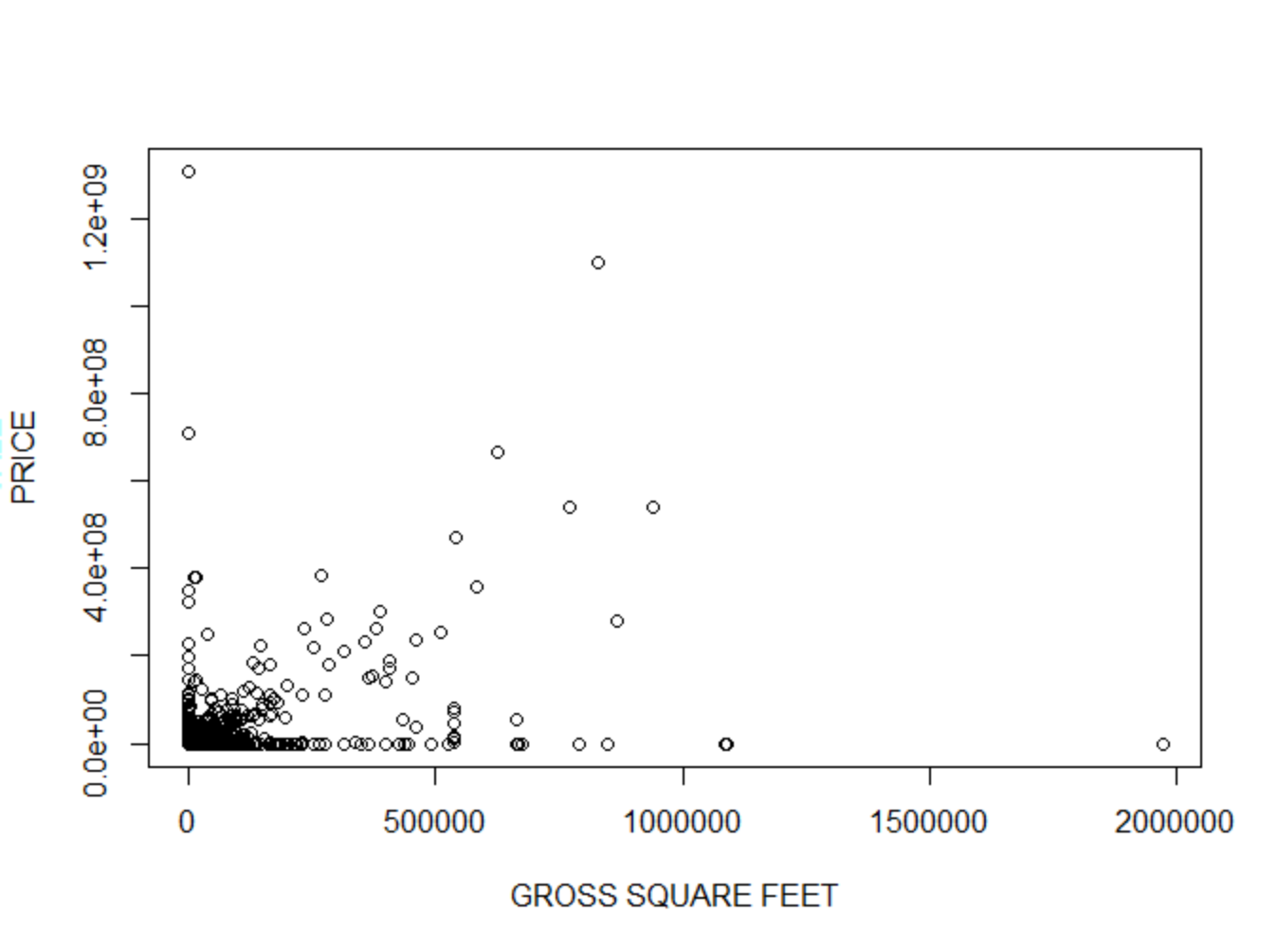
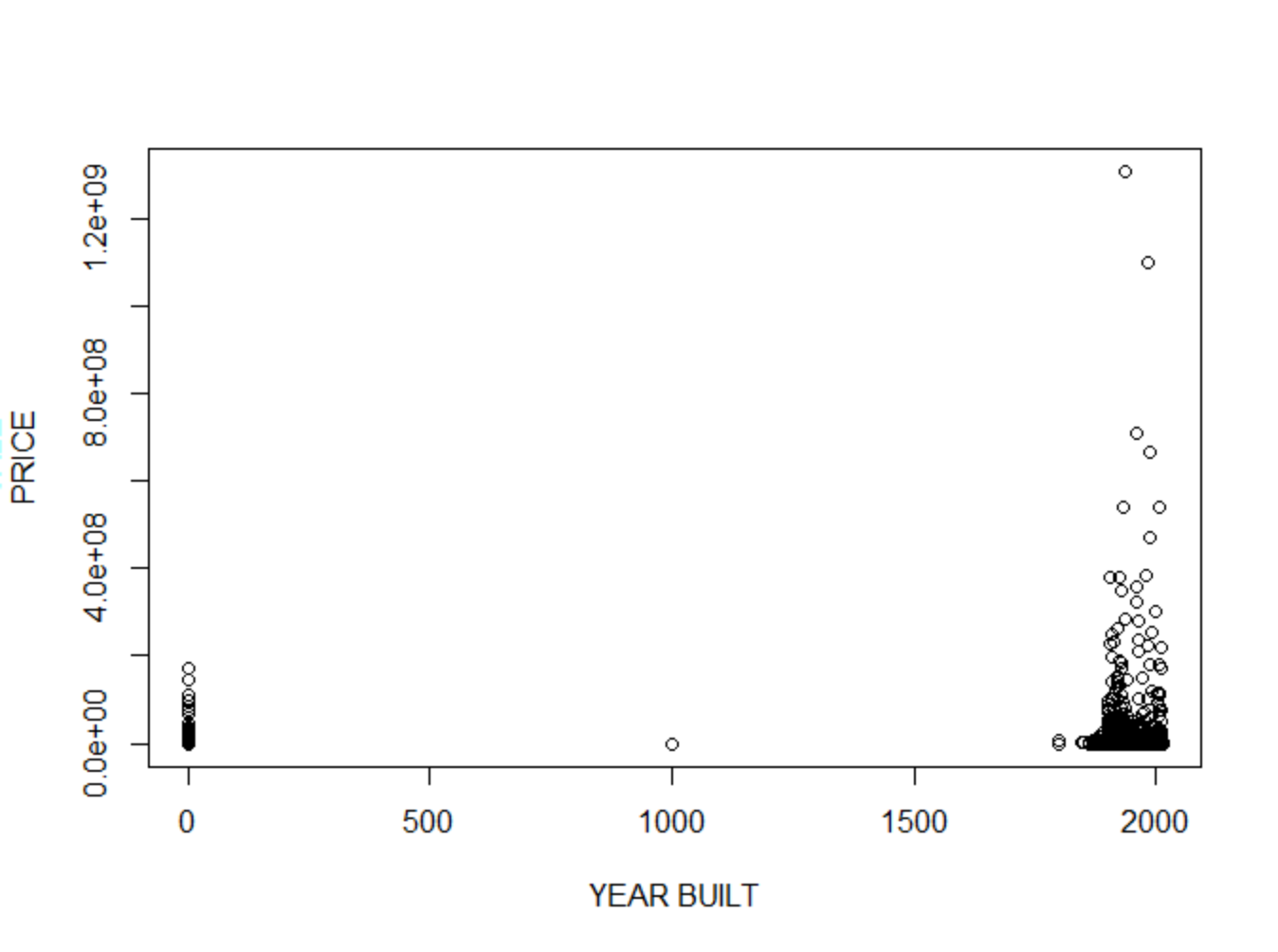
summary(mm)

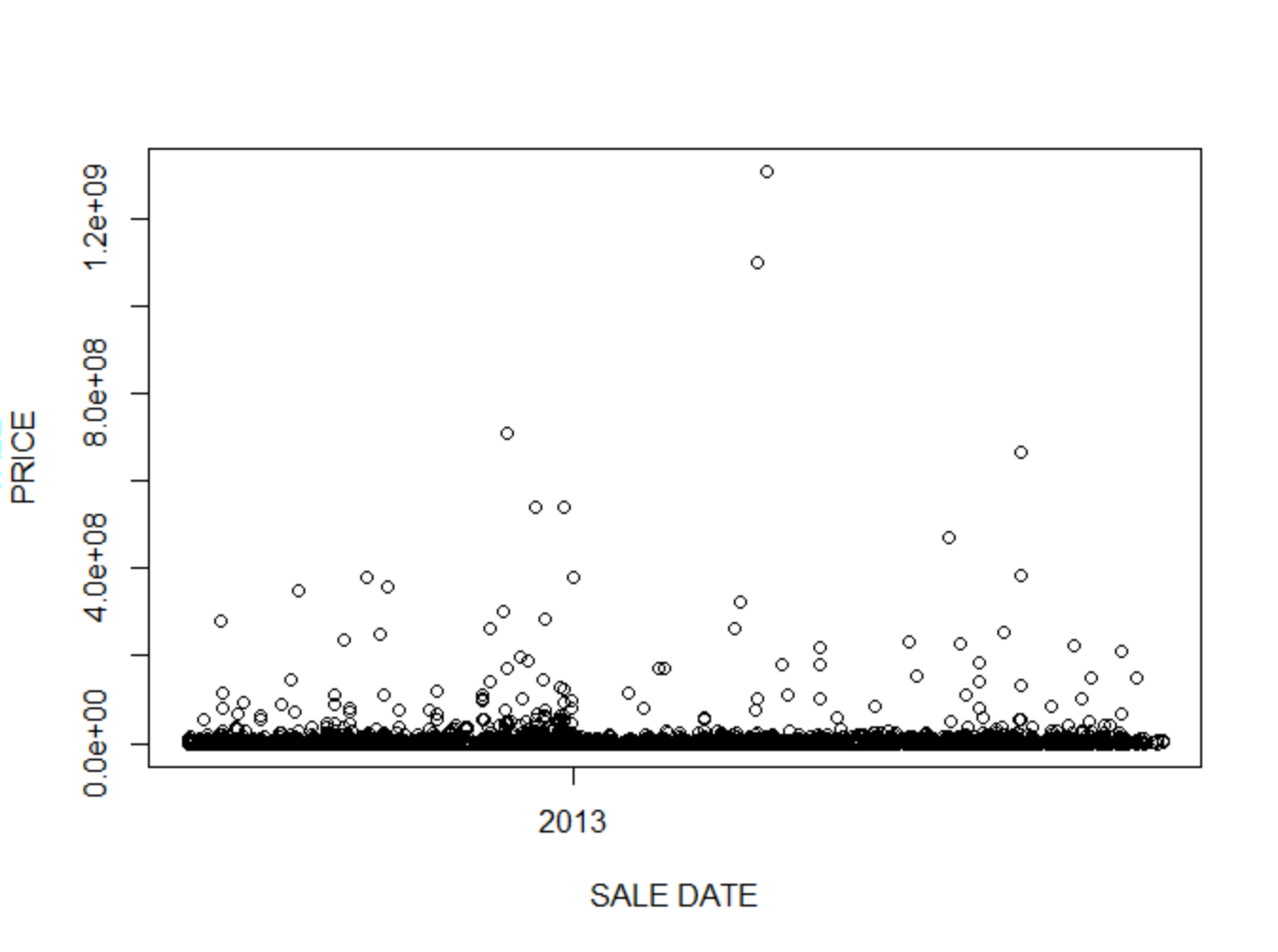
PART1:

1.





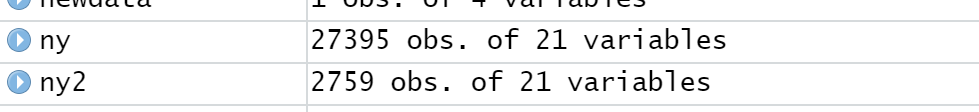




According to my analysis, the sale price for the house are mostly under 1 million dollars and the median is 450,000 dollars. Based on my multiple plots of the sales price against other variables in the dataset, there is not a very clear pattern indicating that the sales price heavily depends on a specific variable. However, it is obvious from the plots that sales prices have some relations with sale price, years build, gross squre foot and year of the sales and Zip code and Tax class at the time of sale and land square feet.

b).

I am going to use the regression model. There are a lot 0 values in both independent variables and dependent variables that I am interested in. Those 0 value will influence the correctness of any model that I am trying to use. Therefore I decide to discard the rows that has a 0 value in any variables that I am interested in. After the data clearing, I still have 2759 rows.

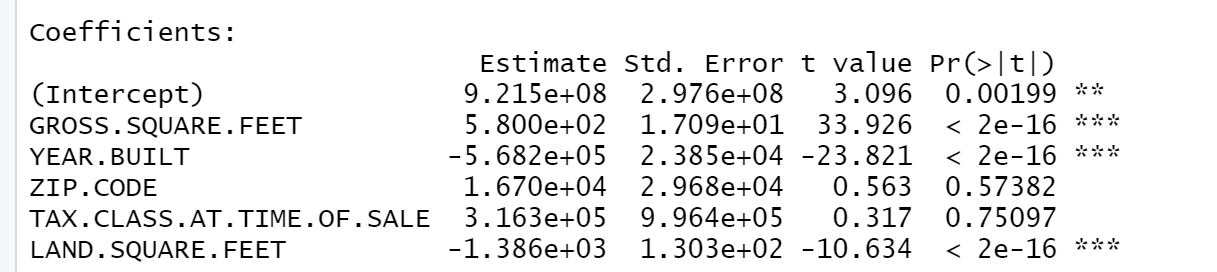


Part2:

a)

I want to predict the sales price based on the years build, gross square foot and year of the sales and Zip code and Tax class at the time of sale. I build the regression model for my prediction. After the data clearing, I have a 2759 rows for my dataset. So I choose to use the first 2700 rows as my training data for regression model. I use the last 59 rows as test cases and I test them manually one by one. Most of the sale price differ a lot from the actual sale price.

b)



I performed a significance test on my regression model. The p value of the gross square feet,year built and land square feet are less than 0.05 meaning the null hypothesis is invalid. However, the p value for zip code and tax class at time of sale is larger than 0.05. Therefore , these two variables have little correlation with the sale price.

c) The size of the house and years of the build of the house are correlated to the sale price, this observation is as expected in the real world. However, I believed the zip code will also have huge influence on the sale price of the house, but it is not according to my model. I think the reason is that there are not sufficient datapoints.

Part3:

1. For the regression model, I found that my prediction is not very accurate. One reason is that I do not have sufficient datapoints. After clearing, I only have 2765 datapoints for train and test. In addition, I think some variables that I picked don’t have clear linear relationship with the sales price which makes the model inaccurate. Finally, the regression model is easily affected by outliers, in my dataset, there is a property that the gross square feet is 2465, yet the sale price is 12,000,000 which is significantly higher than the price of the other properties.