**Module Assignment**

**Module 6**

**QMB-6304 Foundations of Business Statistics**

**#pre-processing**

**#Debjani Sarma**

**rm(list=ls())**

**library("rio")**

**library("moments")**

**getwd()**

Write a simple R script to execute the following:

**Preprocessing**

1. Load into R the data included in “6304 Module 6 Assignment Data.xlsx”. This data set includes information on 19,763 single family homes sold in Davidson County, Tennessee (the Nashville metro area) in 2013. The variables included are property.address, property.city, sale.price, land.value, building.value, total.value, and finished.area. This is your master data set.

**master.data=import("6304 Module 5 Assignment Data.xlsx")**

1. Using the method presented in class and applying the numerical portion of your U number as a random number seed take a random sample of 4000 of the homes in the master data set. Carefully consider whether any of the numeric variables should be converted to factors, and convert those variables to factors. Additionally, make certain your sample includes only the city of NASHVILLE. This will be your primary data set.

**nashville.data = master.data[master.data$property.city == "NASHVILLE", ]**

**nashville.data$bedrooms=as.factor(nashville.data$bedrooms)**

**nashville.data$full.bath=as.factor(nashville.data$full.bath)**

**set.seed(24173877)**

**my.nashdata=nashville.data[sample(1:nrow(nashville.data),4000),]**

**Analysis**

Using your primary data set:

1. Show the results of applying the str() command.

**> str(my.nashdata)**

**'data.frame': 4000 obs. of 9 variables:**

**$ property.address: chr "8017 STALLION CT" "1213 SHELBY AVE" "718 26TH AVE N" "2116 SUNSET PL" ...**

**$ property.city : chr "NASHVILLE" "NASHVILLE" "NASHVILLE" "NASHVILLE" ...**

**$ sale.price : num 146500 454638 50000 794560 438000 ...**

**$ land.value : num 36000 85000 11000 214500 100000 ...**

**$ building.value : num 95600 228500 46200 419600 326500 ...**

**$ total.value : num 131700 326100 57200 666400 426500 ...**

**$ finished.area : num 1504 1702 1010 3306 2672 ...**

**$ bedrooms : Factor w/ 3 levels "2","3","4": 2 2 1 3 3 2 1 2 2 2 ...**

**$ full.bath : Factor w/ 4 levels "0","1","2","3": 2 3 2 4 4 3 2 3 3 3 ...**

**> attach(my.nashdata)**

1. Conduct a multiple linear regression on the data with sale.price as the dependent variable and all other variables as independents, excluding property.address.

**> multi.reg = lm(sale.price ~ land.value + building.value +**

**+ total.value + finished.area + bedrooms + full.bath,**

**+ data = my.nashdata)**

As a part of this:

* 1. Show the R summary of your model’s output.

**> summary(multi.reg)**

**Call:**

**lm(formula = sale.price ~ land.value + building.value + total.value +**

**finished.area + bedrooms + full.bath, data = my.nashdata)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-785181 -40945 -3938 37781 9650075**

**Coefficients:**

**Estimate Std. Error t value Pr(>|t|)**

**(Intercept) 9.334e+04 1.986e+05 0.470 0.638**

**land.value -2.532e+00 5.703e-01 -4.440 9.25e-06 \*\*\***

**building.value -2.832e+00 5.743e-01 -4.932 8.47e-07 \*\*\***

**total.value 3.825e+00 5.664e-01 6.753 1.66e-11 \*\*\***

**finished.area -1.765e-01 8.696e+00 -0.020 0.984**

**bedrooms3 -7.498e+02 8.393e+03 -0.089 0.929**

**bedrooms4 7.614e+03 1.197e+04 0.636 0.525**

**full.bath1 -6.630e+04 1.984e+05 -0.334 0.738**

**full.bath2 -7.582e+04 1.984e+05 -0.382 0.702**

**full.bath3 -8.941e+04 1.987e+05 -0.450 0.653**

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 198200 on 3990 degrees of freedom**

**Multiple R-squared: 0.4363, Adjusted R-squared: 0.435**

**F-statistic: 343.1 on 9 and 3990 DF, p-value: < 2.2e-16**

* 1. Give written interpretations of the beta coefficients in terms of the actual case at hand. Interpret beta coefficients with p <= .10.

Coefficients:

Intercept: It indicates that if all independent variables are zero, the predicted sales price would be $93340. This estimate is not very meaningful in real life, since a house with 0sqft of land or finished area does not exist. Also, with a p-value of 0.638, the intercept is not statistically significant which means that it doesn't contribute much to the overall prediction of sale price.

Land Value: It indicates that for every unit increase in land value, the sale price decreases by $2.532. With extremely low p-value (9.25e-06), the land value's effect is highly significant, which mean that it contributes to the overall prediction of sale price.

Building value: It indicates that for every unit increase in building value, the sale price decreases by $2.832. With extremely low p-value (8.47e-07), the building value's effect is highly significant, which mean that it contributes to the overall prediction of sale price.

Total value: It indicates that for every unit increase in total value, the sale price increases by $3.825. With extremely low p-value (1.66e-11), the total value's effect is highly significant, which mean that it contributes to the overall prediction of sale price.

The coefficient for finished area, bedrooms 3, bedrooms 4, full bath1, full bath 2, full bath 3, are not statistically significant in predicting sale prices, as the p-values are greater than 0.10. This indicates that these independent variables do not meaningfully impact in predicting sale price.

Multiple R-squared as 0.4363 suggests that 43.63% of the variance in sale process is explained by the model.

* 1. Assess your model’s conformance with the LINE assumptions of regression. State whether you believe your model to be in conformance with these assumptions of regression.

**par(mfrow=c(2,2))**

**#Linearity**

**plot(my.nashdata$sale.price, multi.reg$fitted.values,**

**xlab = "Sale Price", ylab="Fitted values",**

**pch=19,main ="Actual vs Fitted Values, Linearity")**

**abline(0,1,col="red",lwd=3)**

**#Independence**

**plot(scale(multi.reg$residuals), pch=19, main="Independence plot",**

**ylab="Scaled residuals")**

**abline(0,0, col="red", lwd=3)**

**#Normality**

**qqnorm(multi.reg$residuals, pch=19, main = "Normality Plot")**

**qqline(multi.reg$residuals, col="red", lwd=3)**

**residuals = multi.reg$residuals**

**#histogram**

**hist(multi.reg$residuals, col="red", main = "Histogram of Residuals",**

**probability = TRUE, ylim = c(0, 250e-08),**

**xlim = c(-2e+06, 2e+06), xlab= "Residuals")**

**max.res=max(multi.reg$residuals)**

**curve(dnorm(x,mean(multi.reg$residuals),sd(multi.reg$residuals)),**

**from=c(-2e+06, 2e+06),**

**to=max.res,lwd=3,col="blue", add=TRUE)**

**skewness(multi.reg$residuals)**

**kurtosis(multi.reg$residuals)**

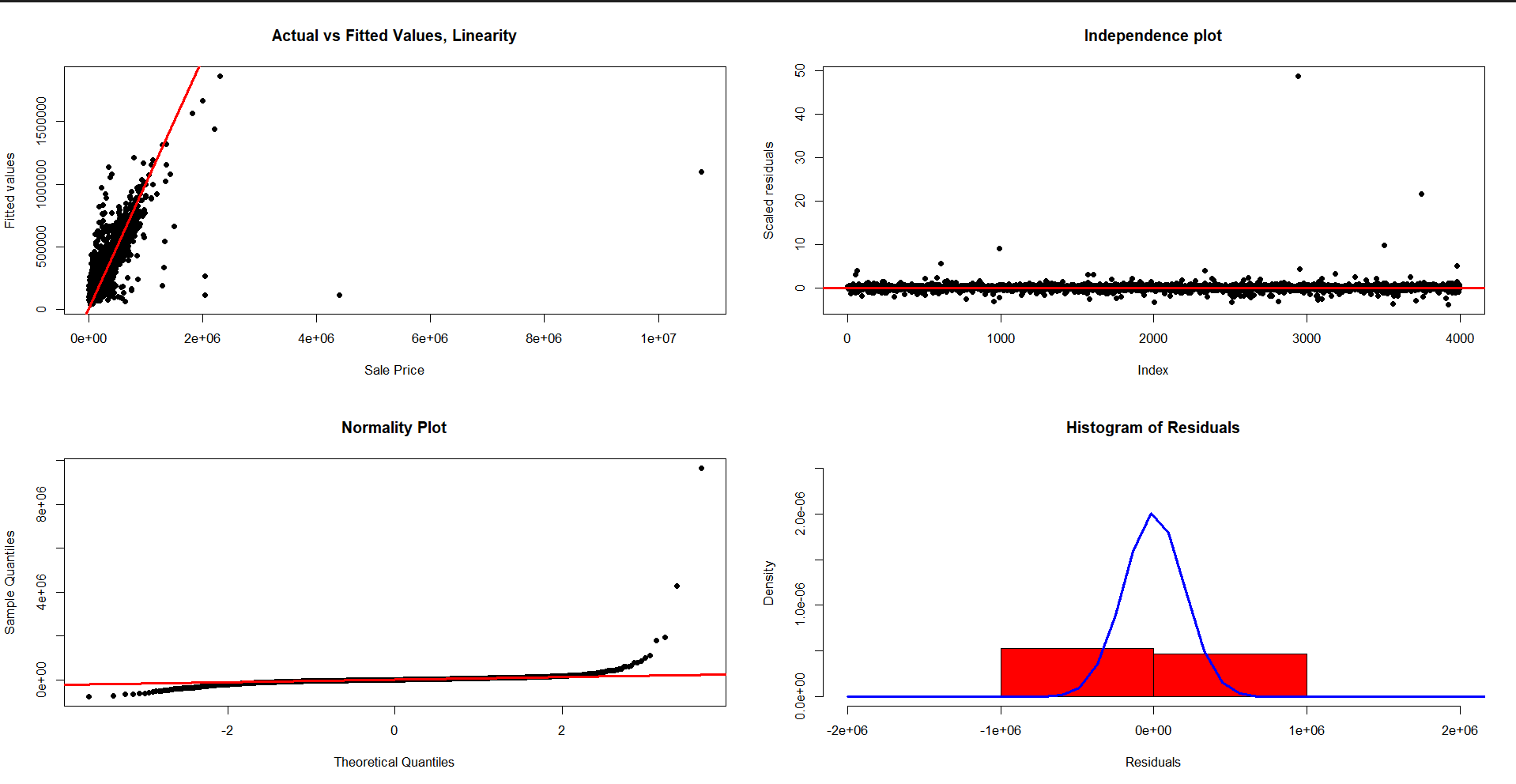
**#Equality of variances**

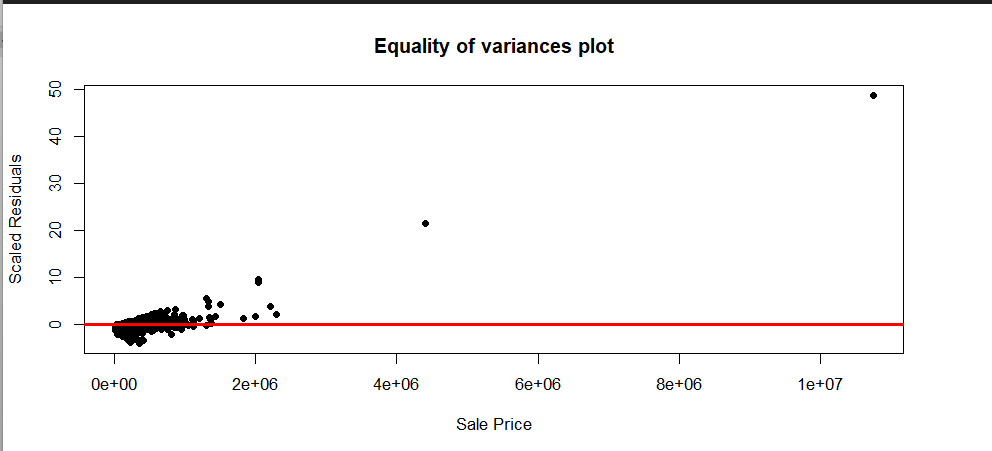
**plot(sale.price,scale(multi.reg$residuals),**

**pch=19, main="Equality of variances plot",**

**xlab = "Sale Price", ylab="Scaled Residuals")**

**abline(0,0,col="red",lwd=3)**





Linearity: Although there are some extreme outliers and scatter points around the line, majority of the data points follow a linear trend. This suggest that we are in conformity with the assumption of linearity, with minor deviations.

Independence: The residuals generally appear randomly scattered around zero, which suggests that there is no strong pattern. This indicates we are in conformity with the assumption of independence.

Normality: The residuals deviate from the normal distribution, especially in the tails. The skewness is 31.91127 and kurtosis is 1470.976, both are very high. These values suggest that the residuals are not symmetrically distributed, indicating a violation of the normality assumption. Histogram of Residuals also shows non-normal distribution which adds to the conclusion.

Equality of Variance : The residual plot shows some extreme outliers and the residuals forms a fan-shaped pattern. This suggests a violation of the assumption of equal variance (homoscedasticity)

1. Given this analysis, do you consider your model to be a good fit to your primary data set?

Based on the above analysis, the R-squared value of 0.4363 indicates that the model explains approximately 43.63% of the variance in sale prices, which is considered moderate. However, violations of the LINE assumptions—specifically, the non-normality of residuals and heteroscedasticity—suggest that the model may not be the best fit for the data.

**Mean Sale Price:**

**mean(my.nashdata$sale.price) = 266666.3**

The Residual Standard Error (RSE) of 198200 is relatively large compared to the mean sale price, indicating that the model may not provide highly accurate predictions.

There are 3 independent variables that contribute to explaining the variability in predicted sale prices.

Also, the model seems to be in conformity with certain LINE assumptions, such as linearity and independence of residuals.

Considering all these factors, we can conclude that the model is a moderate fit but not a good fit overall.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions, 2) the results of those instructions, and 3) any interpretations asked for in the assignment instructions. The first line of your script file should be a “#” comment line showing your name as it appears in Canvas. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed and submitted by the time stated on Canvas. No collaboration of any sort is allowed on this assignment. Please remember the prohibition on using screen shots in your deliverable.**