Module Assignment
Module 6
QMB-6304 Foundations of Business Statistics

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.

2. 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.

```
#Carolina Aldana Yabur
#U25124553

#Preprocessing
rm(list=ls())
library(rio)
masterdf=import("6304 Module 5 Assignment Data.xlsx")
colnames(masterdf)=tolower(make.names(colnames(masterdf)))
masterdf$bedrooms= as.factor(masterdf$bedrooms)
masterdf$full.bath= as.factor(masterdf$full.bath)
intermediatedf= subset(masterdf, masterdf$property.city ==
"NASHVILLE")
set.seed(25124553)
primarydf= intermediatedf[sample(1:nrow(intermediatedf),4000),]
attach(primarydf)
```

Analysis

Using your primary data set:

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

```
> #Analysis
> #Part 1
> str(primarydf)
'data.frame':  4000 obs. of  9 variables:
 $ property.address: chr  "5510  TENNESSEE AVE" "1016  GARFIELD
ST" "1809 OTTER CREEK  RD" "110 DELLWAY  DR" ...
 $ property.city   : chr  "NASHVILLE" "NASHVILLE" "NASHVILLE"
"NASHVILLE" ...
 $ sale.price      : num  138500 58600 1330000 161250 306700 ...
 $ land.value      : num  45000 40000 378800 15000 60300 ...
 $ building.value  : num  213800 54100 21400 79100 190100 ...
 $ total.value     : num  258800 94100 405300 94100 258200 ...
 $ finished.area   : num  2027 1056 2028 910 1715 ...
 $ bedrooms        : Factor w/ 3 levels "2","3","4": 2 2 3 1 3 2
1 2 2 3 ...
 $ full.bath       : Factor w/ 4 levels "0","1","2","3": 4 2 4 3
3 3 2 3 2 3 ...
```

2. Conduct a multiple linear regression on the data with sale.price as the dependent variable and all other variables as independents, excluding property.address. As a part of this:
    a. Show the R summary of your model's output.

```
> #Part 2
> homes.out=lm(sale.price~land.value+building.value+total.value
+              +finished.area+bedrooms+full.bath,
+              data=primarydf)
> summary(homes.out)

Call:
lm(formula = sale.price ~ land.value + building.value +
total.value +
    finished.area + bedrooms + full.bath, data = primarydf)

Residuals:
    Min      1Q  Median      3Q     Max
-621336  -38357   -4589   40689 4175819

Coefficients:
```

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.747e+05  1.139e+05   1.534   0.1251
land.value      -2.211e-01  3.331e-01  -0.664   0.5069
building.value  -6.613e-01  3.354e-01  -1.972   0.0487 *
total.value      1.472e+00  3.303e-01   4.456 8.59e-06 ***
finished.area    1.231e+01  5.001e+00   2.461   0.0139 *
bedrooms3       -7.566e+03  4.771e+03  -1.586   0.1129
bedrooms4       -3.315e+03  6.879e+03  -0.482   0.6300
full.bath1      -1.393e+05  1.139e+05  -1.223   0.2214
full.bath2      -1.341e+05  1.140e+05  -1.176   0.2395
full.bath3      -1.460e+05  1.142e+05  -1.278   0.2012
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 113800 on 3990 degrees of freedom
Multiple R-squared:  0.6595,  Adjusted R-squared:  0.6588
F-statistic: 858.8 on 9 and 3990 DF,  p-value: < 2.2e-16
```

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

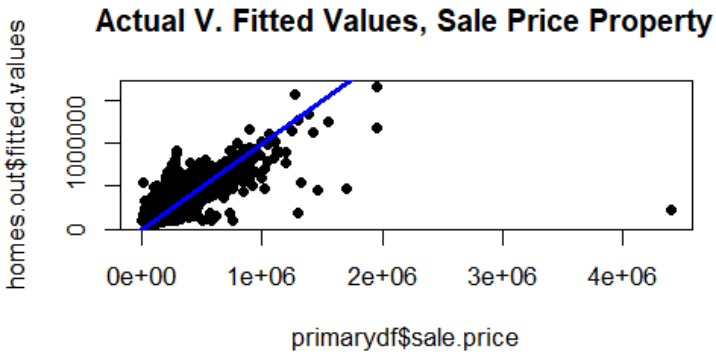
From the summary, as we can see with the intercept beta coefficient, we can say that the expected value of a home property in Nashville is **$174,700** when all independent variables are 0.

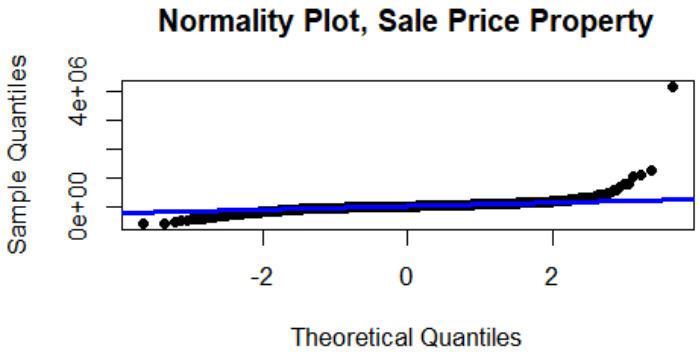
The beta coefficients with p <= .10 are **building.value, total.value**, and **finished.area**, which means that we reject the null hypothesis of these variables as having no impact on the sale price of the home properties in Nashville, and have enough evidence to conclude that these coefficients are statistically significant in this model.

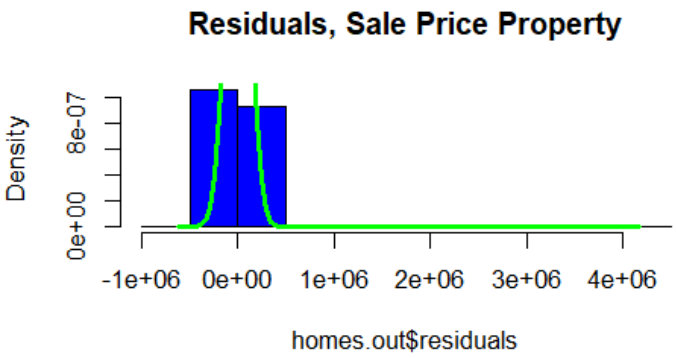
- The beta coefficient of the independent variable **building.value** is **-0.6613**, which means there is a negative relationship between the sale price and the building value of a home property in Nashville. For every one unit increase in the building value of a property, the expected sale price of the home property goes down by **$0.6613**, holding all other independent variables constant.
- The beta coefficient of the independent variable **total.value** is **1.472**, which means there is a positive relationship between the sale price and the total value of a home property in Nashville. For every one unit increase in the total value of a property, the expected sale price of the home property goes up by $1.472, holding all other independent variables constant.
- The beta coefficient of the independent variable **finished.area** is **12.31**, which means there is a positive relationship between the sale price and the finished area of a home property in Nashville. For every one unit increase in the finished area of a property, the expected sale price of the home property goes up by $12.31, holding all other independent variables constant.

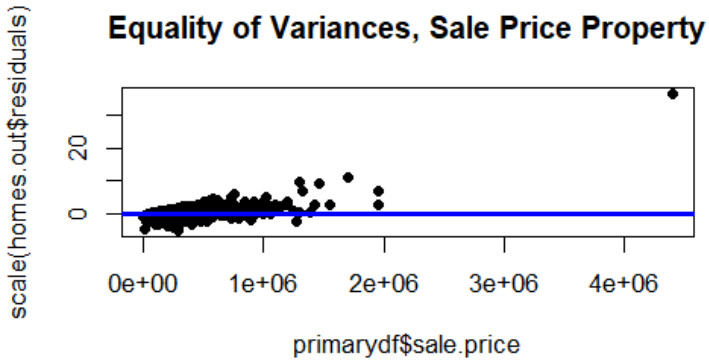
c. Assess your model's conformance with the LINE assumptions of regression. State whether you believe your model to be in conformance with these assumptions of regression.

```
#c
par(mfrow=c(2,2))
#Linearity
plot(primarydf$sale.price, homes.out$fitted.values,
     pch=19, main="Actual V. Fitted Values, Sale Price
Property")
abline(0,1,col="blue", lwd=3)
#Normality
#QQ Plot for Residuals
qqnorm(homes.out$residuals,pch=19,main="Normality Plot, Sale
Price Property")
qqline(homes.out$residuals,col="blue",lwd=3)
#Histogram of Residuals
hist(homes.out$residuals,col="blue",main="Residuals, Sale Price
Property",
     probability="TRUE")
curve(dnorm(x,
mean(homes.out$residuals),sd(homes.out$residuals)),
      from= min(homes.out$residuals),to=
max(homes.out$residuals),
      col="green",lwd=3,add=TRUE)
#Equality of Variances
plot(primarydf$sale.price,scale(homes.out$residuals),
     pch=19, main="Equality of Variances, Sale Price Property")
abline(0,0,col="blue",lwd=3)
```









Linearity: Apart from a few data points, most data points seem to follow a strong linear pattern, indicating normal distribution. We can conclude it is in conformity with the linearity assumption.

Normality: In the QQ Plot, most of the data points follow a normal distribution, although in the tails the points deviate slightly from the line. Since most of the data follows a normal distribution, the QQ Plot indicates that the model is in conformity with the normality assumption. In addition, the histogram shows that residuals are slightly skewed to the right, but the histogram seems bell-shaped. Since most residuals seem to follow normal distribution, it appears that the model is in conformity with the normality assumption.

Equality of Variances: There is no clear evidence of a specific pattern in the residuals that suggests that it does not follow normal distribution, so it appears that the model is in conformity with the equality of variances assumption.

Therefore, it appears that the model is in conformity with the LINE assumptions of regression.

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

Since the model is in conformity with the LINE assumptions of regression and the R^2 is high (0.6595), I think the model is a good fit to my primary data set. However, the residual plot shows evidence of under forecasting, particularly in one of the tails. I believe the model would become a better fit if some of the independent variables with less statistical significance were excluded from the multiple regression model.

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.**