Module Assignment
Module 4
QMB-6304 Foundations of Business Statistics

Write a simple R script to execute the following:

Preprocessing

1. Load the file "6304 Module 4 Assignment Data.xlsx" into R. This file contains information on 99 used automobiles for sale.
2. Using the numerical portion of your U number as a random number seed, take a random sample of 30 cars using the method presented in class. This will be the your primary data for your assignment.

```
#Preprocessing
#Carolina Aldana Yabur
#U25124553

rm(list=ls())
install.packages("rio")
install.packages("moments")
library(rio)
library(moments)
automobiles_data=import("6304 Module 4 Assignment Data.xlsx")
colnames(automobiles_data)=tolower(make.names(colnames(automobil
es_data)))
set.seed(25124553)
automobiles.sample=automobiles_data[sample(1:nrow(automobiles_da
ta),30),]
attach(automobiles.sample)
```

Analysis

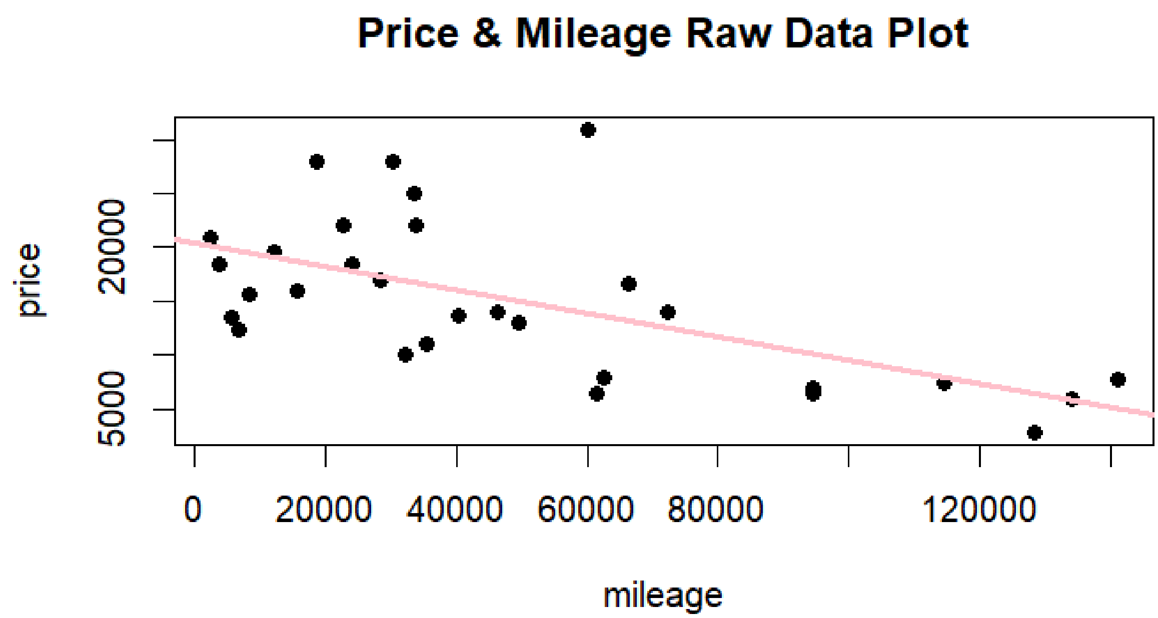
Using your primary data frame:

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

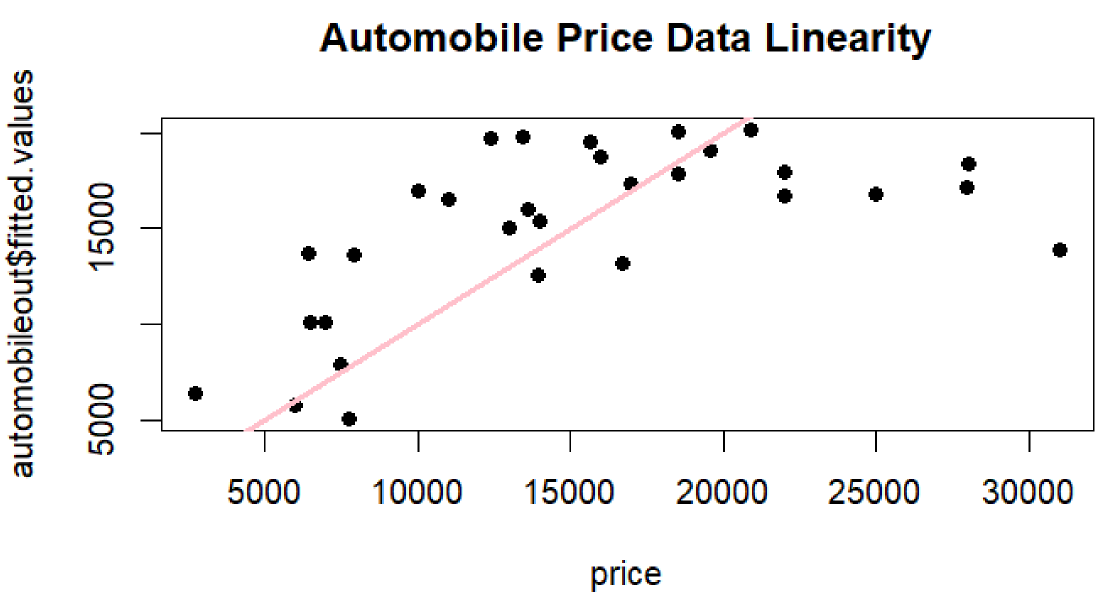
```
> #Analysis
> #Part 1
> str(automobiles.sample)
'data.frame':	30 obs. of  4 variables:
 $ price  : num  27990 16699 24999 6470 2750 ...
 $ mileage: num  30218 66296 33677 94396 128330 ...
 $ age    : num  3 6 3 12 22 13 2.5 3 8 3 ...
 $ make   : chr  "Lexus" "Lexus" "Lexus" "Honda" ...
```

2. Conduct a simple regression using the price variable as the dependent variable and mileage as the independent variable.

```
> #Part 2
> automobileout=lm(price~mileage, data=automobiles.sample)
> summary(automobileout)

Call:
lm(formula = price ~ mileage, data = automobiles.sample)

Residuals:
    Min      1Q  Median      3Q     Max 
-7319.9 -3679.3  -925.2  2364.1 17095.5 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.045e+04  1.708e+03  11.970 1.58e-12 ***
mileage     -1.092e-01  2.692e-02  -4.057 0.000361 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5885 on 28 degrees of freedom
Multiple R-squared:  0.3702,  Adjusted R-squared:  0.3477 
F-statistic: 16.46 on 1 and 28 DF,  p-value: 0.0003609

plot(mileage, price,pch=19, main="Price & Mileage Raw Data
Plot")
abline(automobileout, lwd=3,col="pink")
```




```
plot(price, automobileout$fitted.values,pch=19, main="Automobile
Price Data Linearity")
abline(0,1,lwd=3,col="pink")
```



3. Building on your results in Part 2 give clear written interpretations of your model's beta coefficients and associated p values. Make certain your beta coefficient interpretations are in terms of the case provided.

```
> #Part 3
> automobileout_summary=summary(automobileout)
> beta_coefficients= automobileout_summary$coefficients
> beta_coefficients
                 Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 20449.2039795 1.708352e+03 11.970137 1.579890e-12
mileage        -0.1092225 2.692439e-02 -4.056638 3.609364e-04
```

From the summary, the value of the intercept beta coefficient is **20449.2039795**, and the value of the slope beta coefficient is **-0.1092225**. The regression equation is therefore:

Price of Automobile=20449.2039795-0.1092225 X Mileage

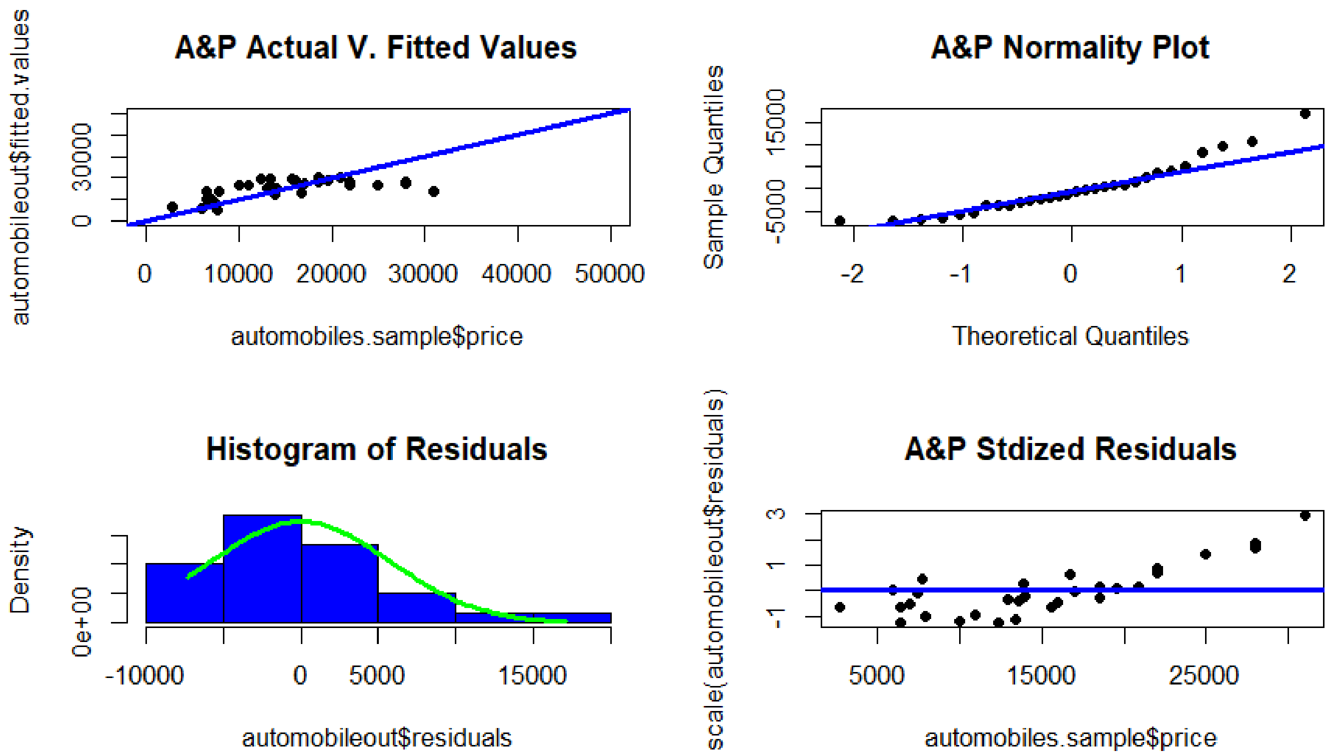
The equation means that when the mileage of the automobile is **0**, the price of the automobile is approximately **$20449**. In addition, since the value of the slope beta coefficient is **-0.1092225**, then for every one unit increase in the mileage of the automobile, the price of the automobile goes down by **$0.1092225**.

On the other hand, the **associate p-value of the intercept beta coefficient** is **1.579890e-12,** and the **associated p-value of the slope beta coefficient** is **3.609364e-04**. The p-value of the intercept beta coefficient (**1.579890e-12)** is less than the standard significance level 0.05. Therefore, we reject the null hypothesis of the intercept being equal to 0. Furthermore, the p-value of the slope beta coefficient (**3.609364e-04)** is also less than the standard significance level 0.05, therefore, we reject the null hypothesis that mileage has no impact on the price of an automobile. There is statistically significant evidence to conclude that mileage has a significant and negative impact on the price of an automobile.

4. Building on your results in Part 2 determine and state whether your model is in conformity with the LINE assumptions of regression.

```
#Part 4
par(mfrow=c(2,2))
#Linearity
plot(automobiles.sample$price, automobileout$fitted.values,
     pch=19, xlim=c(0,50000), ylim=c(0,50000), main="A&P Actual
V. Fitted Values")
abline(0,1,col="blue",lwd=3)
#Normality
qqnorm(automobileout$residuals,pch=19,main="A&P Normality Plot")
qqline(automobileout$residuals,col="blue",lwd=3)
hist(automobileout$residuals,col="blue",
     main="Histogram of Residuals", probability = "TRUE")
```

```
curve(dnorm(x,mean(automobileout$residuals),sd(automobileout$res
iduals)),
      from= min(automobileout$residuals),to =
max(automobileout$residuals),col="green",lwd=3,add=TRUE)
#Equality of Variances
plot(automobiles.sample$price,scale(automobileout$residuals),
     pch=19, main="A&P Stdized Residuals")
abline(0,0,col="blue",lwd=3)
```



The model is in conformity with the LINE assumptions of regression.

Linearity: After applying standard limitations to both X and Y, most of the points follow the diagonal line. There is some concern with a few points deviating at the higher end, as it appears the model does not forecast well for more expensive cars. However, since the majority of points follow normal distribution, we can conclude it is in conformity with the linearity assumption.

Normality: In the QQ Plot, most of the data follows a normal distribution, although there are some exceptions at the higher end. In addition, the histogram shows that residuals are slightly skewed to the right. However, since most residuals follow a 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 suggest that it does not follow normal distribution, so it appears that the model is in conformity with the equality of variances assumption.

5. A two-year-old Honda CR-V with approximately 25,000 miles is offered for sale. What does your model predict the price of this car should be?

```
> honda_crv=data.frame(mileage=25000)
> predict(automobileout,honda_crv,interval = "predict")
       fit      lwr      upr
1 17718.64 5390.853 30046.43
```

Based on the model, the predicted price of the two-year-old Honda CR-V with approximately 25,000 miles is approximately $**17719.**

6. The vehicle shown below is a 1934 Rolls Royce Phantom II Continental Owens Drophead Sedanca Coupe, a custom-built vehicle which is likely one of a kind and highly collectable. Its odometer shows 1275 miles. The last time the vehicle sold was for a price in excess of $600,000. Explain why it would be inappropriate to use your model to predict the price of this car.

This model is not appropriate because it only considers the independent variable mileage. This particular vehicle is a custom-built vehicle, likely one of a kind and highly collectable. Due to its unique characteristics that other vehicles do not have, the price of it would be higher. If we were to predict the price of this car using this simple regression model, we would see that the estimated price $20310 is much lower than the price of it last time it was sold ($600,000).

```
> #Part 6
> rollsroycephantom=data.frame(mileage=1275)
> predict(automobileout,rollsroycephantom,interval =
"predict")
       fit      lwr     upr
1 20309.95 7772.089 32847.8
```

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions and 2) the results of those 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 before you leave the classroom. No collaboration of any sort is allowed on this assignment.**