Module Assignment
Module 6
QMB-6304 Foundations of Business Statistics

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

Preprocessing

1. Load the file "6304 Module 6 Assignment Data.xlsx" into R. This file contains information on 1338 instances of an adult being hospitalized somewhere in the United States. The variables included the patient's age, body mass index (bmi), whether or not they were a smoker, and the total final charges for hospital care submitted to the patient or a third-party payer. This is your master data set.
2. Using the numerical portion of your U number as a random number seed, take a random sample of 150 cases from the full data set using the method presented in class. Convert smoking status to a factor variable. This will be your primary data set for analysis.

```
#Carolina Aldana Yabur
#U25124553

#Preprocessing
rm(list=ls())
setwd("C:/Users/calda/Desktop")
library(rio)
master_dataset=import("6304 Module 6 Assignment Data.xlsx")
colnames(master_dataset)=tolower(make.names(colnames(master_data
set)))
set.seed(25124553)
primary_dataset=master_dataset[sample(1:nrow(master_dataset),150
),]
as.factor=(primary_dataset$smoker)
attach(primary_dataset)
```

Analysis
Using your primary data set:

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

```
> #Analysis
> #Part 1
> str(primary_dataset)
'data.frame':	150 obs. of  4 variables:
 $ age     : num  20 53 54 40 48 26 57 19 63 19 ...
 $ bmi     : num  22 38.1 31.9 28.1 30.8 ...
 $ smoker  : chr  "no" "no" "no" "yes" ...
 $ charges: num  1965 20463 10929 22332 10141 ...
```

2. Conduct a full regression analysis including all variables and using the “charges” variable as the dependent variable.

```
> #Part 2
> hospital.out=lm(charges~age+bmi+smoker, data=primary_dataset)
> summary(hospital.out)

Call:
lm(formula = charges ~ age + bmi + smoker, data =
primary_dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-8794.4 -3532.6 -1493.2   811.3 29848.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -10856.49    3033.88  -3.578  0.00047 ***
age            236.45      37.33   6.333  2.8e-09 ***
bmi            330.50      91.02   3.631  0.00039 ***
smokeryes    19659.14    1317.21  14.925  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6386 on 146 degrees of freedom
Multiple R-squared:  0.6536,	Adjusted R-squared:  0.6465 
F-statistic: 91.82 on 3 and 146 DF,  p-value: < 2.2e-16
```

The three independent variables are statistically significant, meaning that all of them have an impact on the total **charges** for hospital care. The independent variables have a positive relationship with the dependent variable **charges**, so for every one unit increase in age or bmi, or if a person is a smoker, then the expected value of total **charges** for hospital care will increase accordingly.

- age → expected total charges for hospital care increase by $236.45 for every one unit
- bmi → expected total charges for hospital care increase by $330.50 for every one unit
- smoker → expected total charges for hospital care increase by $19659.14 if the person is a smoker.

On the other hand, the R^2 is 0.6536, which is high and means that the model explains 65% of the variation in y (**charges**). Overall, the model is a good fit.

3. Show your model output. Interpret the beta coefficients in your output in terms of the variable's estimated impact on the y. Include an appropriate discussion of the beta coefficient p values.

```
> #Part 3
> summary(hospital.out)

Call:
lm(formula = charges ~ age + bmi + smoker, data =
primary_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-8794.4 -3532.6 -1493.2   811.3 29848.1

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10856.49    3033.88  -3.578  0.00047 ***
age            236.45      37.33   6.333  2.8e-09 ***
bmi            330.50      91.02   3.631  0.00039 ***
smokeryes    19659.14    1317.21  14.925  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6386 on 146 degrees of freedom
Multiple R-squared:  0.6536,  Adjusted R-squared:  0.6465
F-statistic: 91.82 on 3 and 146 DF,  p-value: < 2.2e-16
```

From the summary, the intercept beta coefficient is **-10856.49**. This would mean that when all independent variables are 0, then the total charges are -$10856.49. However, this does not make sense in this context because the hospital is not going to charge a negative amount, so it is not meaningful as part of the conclusions.

The p-values of **age, bmi**, and **smoker** are less than the standard significance of 0.05:

- **age:** 2.8e-09 < 0.05
- **bmi:** 0.00039 < 0.05
- **smokeryes:** < 2e-16 < 0.05

This means that we reject the null hypothesis that these variables have no impact on the **charges** and conclude that these coefficients are statistically significant in this model.

Regarding the beta coefficients, all of them have a positive relationship with the dependent variable **charges**.

- The beta coefficient of the independent variable **age** is **$236.45**, which means that for every unit increase (in this case, another year of life) in **age**, the expected total charges for hospital care go up by **$236.45**, holding all the other independent variables constant.

- The beta coefficient of the independent variable **bmi** is **$330.50**, which means that for every unit increase in **bmi**, the expected total charges for hospital care go up by **$330.50**, holding all the other independent variables constant.

- The beta coefficient of the independent variable smoker is **$19659.14**. In this case, this is a binary variable, so being a smoker increases the expected charges for hospital care by **$19659.14**, holding all the other independent variables constant.

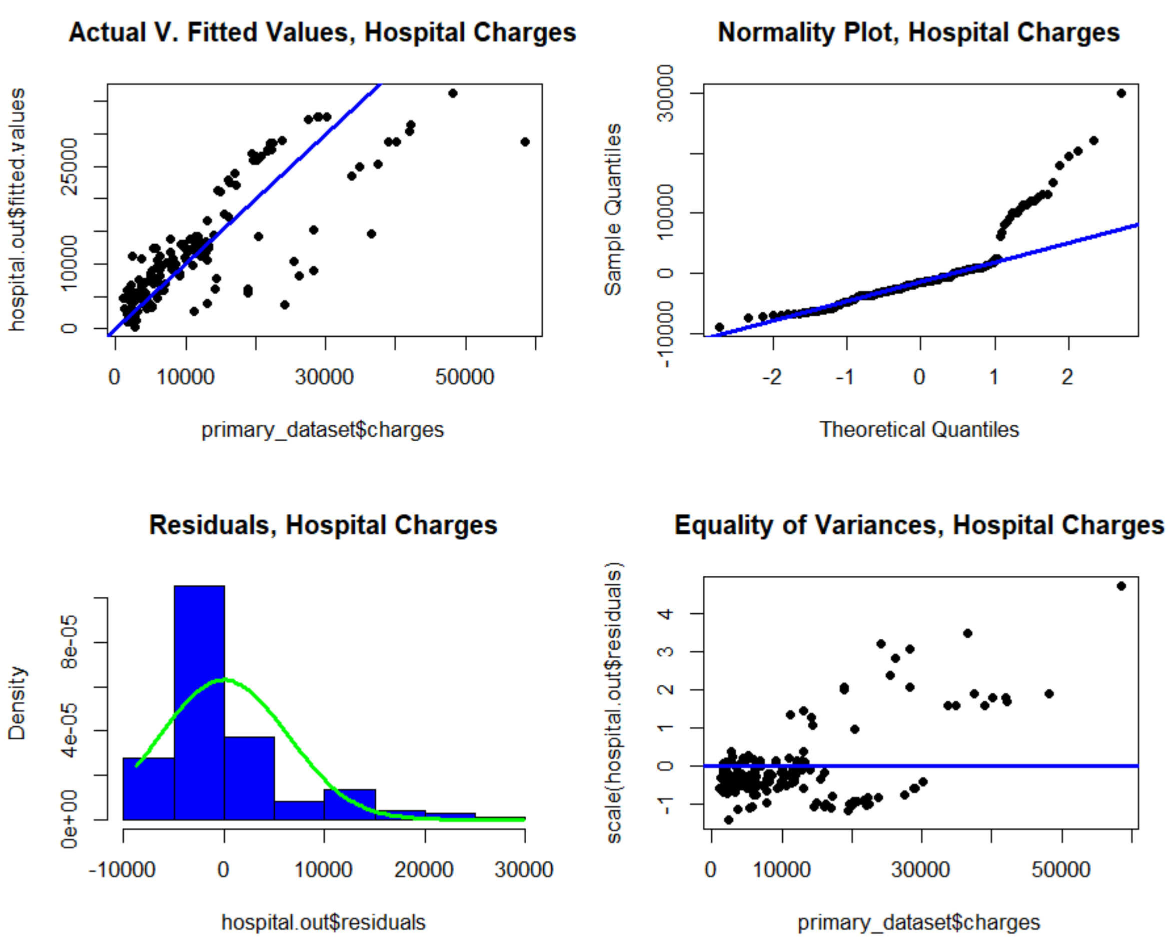
4. Report the confidence interval for each beta coefficient in your model.

```
> #Part 4
> confint(hospital.out)
                  2.5 %     97.5 %
(Intercept) -16852.4891 -4860.4869
age            162.6603   310.2324
bmi            150.6125   510.3908
smokeryes    17055.8769 22262.4098
```

5. Determine and state whether your model appears to be in conformity with the LINE assumptions of regression. Show appropriate graphics where needed to justify your conclusions.

```
#Part 5
par(mfrow=c(2,2))
#Linearity
plot(primary_dataset$charges, hospital.out$fitted.values,
     pch=19, main="Actual V. Fitted Values, Hospital Charges")
abline(0,1,col="blue", lwd=3)
```

```
#Normality
#QQ Plot for Residuals
qqnorm(hospital.out$residuals,pch=19,
       main="Normality Plot, Hospital Charges")
qqline(hospital.out$residuals,col="blue",lwd=3)
#Histogram of Residuals
hist(hospital.out$residuals,col="blue",
     main="Residuals, Hospital Charges",
     probability = TRUE)
#Equality of Variances
plot(primary_dataset$charges,scale(hospital.out$residuals),
     pch=19, main="Equality of Variances, Hospital Charges")
abline(0,0,col="blue",lwd=3)
curve(dnorm(x, mean(hospital.out$residuals),
sd(hospital.out$residuals)),
      from= min(hospital.out$residuals),to=
max(hospital.out$residuals),
      col="green",lwd=3,add=TRUE)
```



Linearity: Most data points follow a strong linear pattern, indicating normal distribution. Therefore, we can conclude that the model meets the linearity assumption.

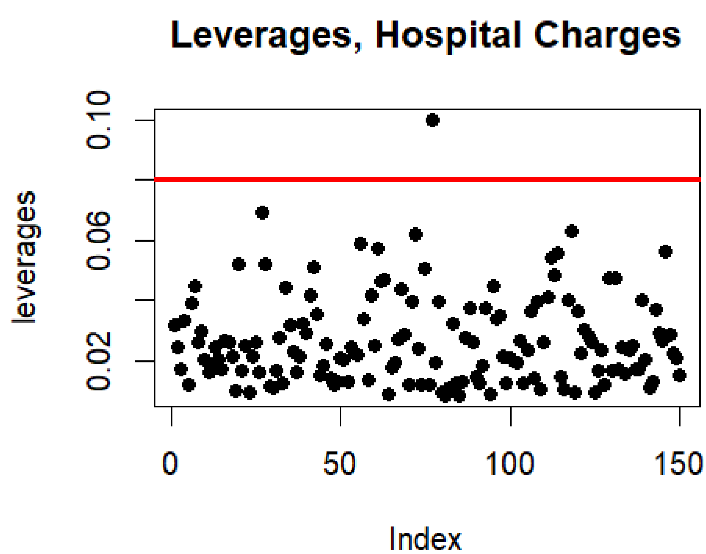
Normality: In the QQ Plot, most of the data points follow the line, which indicates normal distribution, expect for some deviation at one of the ends. Despite this, since the majority of points follow a normal pattern, the model seems to satisfy the normality assumption. In addition, the histogram shows slight right skewness in the residuals, but due to its bell-shape curve, it appears that the residuals mostly follow a normal distribution. So we can conclude that the model meets the normality assumption.

Equality of Variances: There is no distinct pattern in the residuals, so it appears the model is in conformity with the equality of variances assumption.

Therefore, the model seems to be in conformity with the LINE assumptions of regression.

6. Determine whether any of the data points in your reduced data set have a high leverage in influencing the plot of the regression. Show appropriate analytics to support your conclusion. Also, report the observations from your reduced data set (if any) which have such high leverage.

```
#Part 6
leverages=hat(model.matrix(hospital.out))
plot(leverages,pch=19,main="Leverages, Hospital Charges")
abline(3*mean(leverages),0,col="red",lwd=3)
```



It appears that there is one data point that has high leverage in influencing the plot of the regression. This data point is above the red line, which means is above 3 times the mean of the leverages.

```
> primary_dataset[leverages>(3*mean(leverages)),]
    age   bmi smoker  charges
848  23 50.38     no 2438.055
```

The data point that is above 3 times the mean of the leverages is from this row. It refers to a person aged 23, with a bmi of 50.38, who is not a smoker and whose total charges for hospital care are $2438.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is "Assignment 5" of our course and your name as it appears in Canvas. Your code chunks and analysis 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.**