Northeastern University

M6 FINAL PROJECT REPORT

ALY6010 and CRN:70487

Submitted by: Suprit Mestry

Instructor Name: Dr. Dee Chiluiza

Date Of Submission: 12/18/2021

PART 1: INTRODUCTION

a) Simple Linear Regression:

Linear regression models and their analysis allows us to estimate how the values or characteristics of a dependent variable changes with the changes in independent variable. Therefore simple linear regression mainly analyzes one independent and one dependent variable. It also helps us to determine how strong is the relationship of correlation between two quantitative variables.

The relationship is presented by equation: **Y = a +b(X)**, where ‘Y’ is dependent variable, ‘a’ is y-intercept, ‘b’ is slope and ‘X’ is independent variable.

For Example, a simple linear regression is performed for a sample period of 5 months between independent variable “Amount of Total Rainfall” and dependent variable “Number of Umbrellas sold”. This explains that the more the amount of rainfall in a month, more umbrellas will be sold in that particular month and vice versa. This model can be represented by a scatter plot and is concluded with having strong positive correlation between the variables.

b) Multiple Regression:

Multiple regression is the followup enhanced concept over simple regression analysis which is associated with one dependent variable and more than one independent variables. This regression technique is very effective as it allows multiple variable regression and also predicts complex correlations.

The relationship is presented by equation: **Y = a + b1(X1) + b2(X2) +…**, where ‘Y’ is dependent variable, ‘a’ is y-intercept, ‘b1, b2’ are consecutive slopes and ‘X1, X2’ are consecutive independent variables.

For Example, a multiple regression is performed for 5 samples between one dependent variable “House Pricing($)” and 2 independent variables “Size(feet^2)” and “Age of House(yrs)”. This model is also well represented with scatter plots. As “Age of House(yrs)” increases, the pricing will decrease with negative correlation whereas if “Size(feet^2)” increases, the pricing will increase with positive correlation.

c) History:

Regression analysis and correlation has developed and has made a huge impact in the world of statistics. Coming to the history of correlation and regression, Francis Galton made the invention of correlation in the late year of 1888 which made a huge impact in the field of statistics and analytics. With the subsequent efforts of Karl Pearson further established enhanced multiple regression and correlation techniques which helped to analyse modern regresion models and application in industries. In his four-volume biography of Galton, Pearson described the genesis of the discovery of the regression slope. In 1875, Galton had distributed packets of sweet pea seeds to seven friends; each friend received seeds of uniform weight, but there was substantial variation across different packets. Main works of Galton included early considerations of regression and correlation, generality of regression slope, development of correlation and regression mathematically by Pearson and rise to modern multiple regression analysis.

d) Industrial Example and Importance:

For field of interest, concept of regression model and correlation can be used in Business Forecasting. The field of Business forecasting is a essential thing for any business related industry and company to survive in the market. Therefore defining correlation between two independent and dependent variable affecting the company takes place. Further, Regression models and techniques uses the past historical data of the entities of the company and compare the correlation with the current data to predict the future outcomes and values. For example, we consider company’s sales values as dependent variable and the fixed interest rates as independent variable. Hence Regression analysis is used to measure the strength of correlation between the variables and models are represented in a sophisticated graphical or statistical manner. This further helps to forecast the future outcomes like the potential loss or profit that a company can make from their workflow of sales in the market. This will help companies excel in the marketing as well as in the businesses.

REFERENCES

1) Simple linear regression. (n.d.). Wayne W. LaMorte, MD, PhD, MPH Retrieved May 31, 2016., from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/ (https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/)

2) Bevans, R. (2020, October 26). An introduction to multiple linear regression. Scribbr. Retrieved December 18, 2021, from https://www.scribbr.com/statistics/multiple-linear-regression/ (https://www.scribbr.com/statistics/multiple-linear-regression/)

3) Jeffrey M. Stanton. 01 Dec 2017, from https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537 (https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537)

4) Saint-Leger, R. (2016, October 26). Business forecasting using historical data and regression anlaysis. Small Business - Chron.com. Retrieved December 12, 2021, from https://smallbusiness.chron.com/business-forecasting-using-historical-data-regression-anlaysis-32844.html (https://smallbusiness.chron.com/business-forecasting-using-historical-data-regression-anlaysis-32844.html)

PART 2: SECTION OF ANALYSIS

Task 1: Exploring “Faithful” from R directory

Task 1.1

After running code “?faithful” in R console, an ‘Old Faithful Geyser Data’ which is pre-existing data set gets displayed in RStudio. This data set is the description of 2 variables that are ‘Waiting Time’ & ‘Eruption Time’ for eruptions. This dataset is collected from the location at Yellowstone National Park, Wyoming, USA.

The data frame of this data set consists of 272 observations(rows) and 2 variables(columns). This data set was sourced and collected from the reference “Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39”

Task 1.2

```
my_dataa <- faithful
library(tibble)
my_dataa <- as.data.frame(my_dataa)
my_dataa <- subset(my_dataa,select=c(2,1))       #Rearranged Columns
aa<-head(my_dataa,3)                   #First 3 records
bb<-tail(my_dataa,3)                   #Last 3 records
cc<-data.frame(rbind(aa,bb))
knitr::kable(cc)
```

	waiting	eruptions
1	79	3.600
2	54	1.800
3	74	3.333
270	90	4.417
271	46	1.817
272	74	4.467

Task 1.3

Average value of waiting time is: **70.897 minutes.**
Average value of eruption time is: **3.488 minutes.**

```
#Calculating quartiles for variable Waiting
qw_25= quantile(faithful$waiting,0.25)
qw_50= quantile(faithful$waiting,0.50)
qw_75= quantile(faithful$waiting,0.75)
matr1<-matrix(c(qw_25, qw_50, qw_25),nrow = 3,ncol=1,byrow=F)
rownames(matr1)<-c("25th quantile waiting value is","50th quantile waiting value is","75th quantile waiting value
is")
colnames(matr1)<-"VALUE"
knitr::kable(matr1)
```

	VALUE
25th quantile waiting value is	58
50th quantile waiting value is	76
75th quantile waiting value is	58

```
#Calculating quartiles for variable Eruptions
qe_25= round(quantile(faithful$eruptions,0.25),2)
qe_50= round(quantile(faithful$eruptions,0.50),2)
qe_75= round(quantile(faithful$eruptions,0.75),2)
matr2<-matrix(c(qe_25, qe_50, qe_25),nrow = 3,ncol=1,byrow=F)
rownames(matr2)<-c("25th quantile eruption value is","50th quantile eruption value is","75th quantile eruption va
lue is")
colnames(matr2)<-"VALUE"
knitr::kable(matr2)
```

	VALUE
25th quantile eruption value is	2.16
50th quantile eruption value is	4.00
75th quantile eruption value is	2.16

The correlation coefficient for two variables is: 0.901
The coefficient 0.901 indicates correlation between the two variables which is strong in nature.
The determination coefficient for two variables is: 0.811
The determination coefficient 0.811 states that the dependent variable is changed or affected by 81.1% with changes in independent variable. Other factors only affect the changes by 18.9%

$$NullHypothesis: H_0: \rho = 0$$

$$AlternativeHypothesis: H_1: \rho \neq 0$$

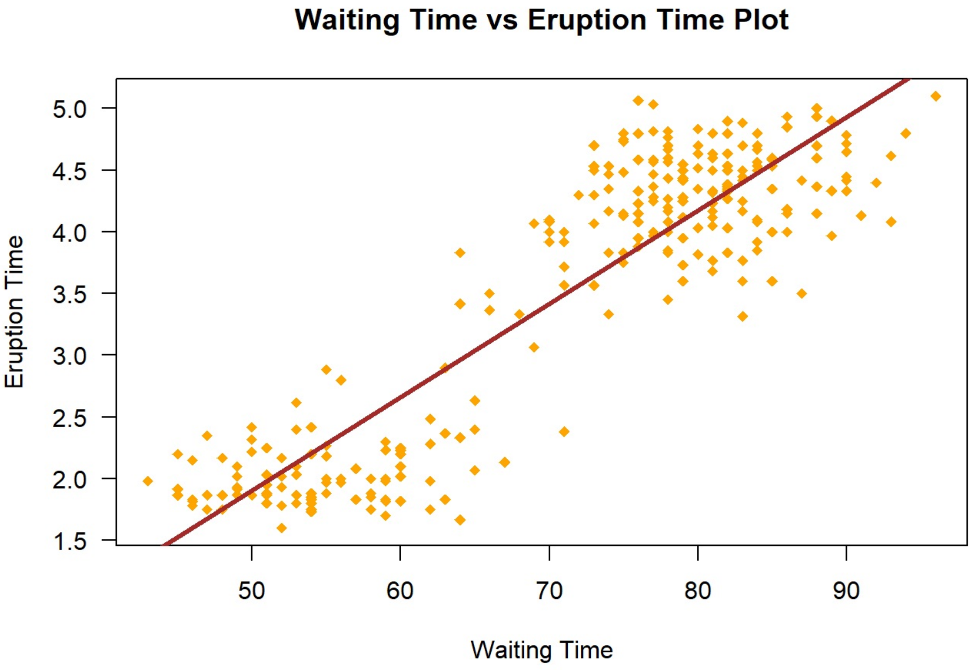
```
#Calculating Test Values and CriticaL Value
r1=cor(faithful$waiting,faithful$eruptions)
d1=(cor(faithful$waiting,faithful$eruptions)^2)
```

The variables **‘waiting’ and ‘eruptions’** has a test value of: **34.089**
The variables when alpha=0.01 have critical value of: **2.594**

From the above results, we used here calculated right critical value as the test value is a positive value. Hence by calculating the parameters we have enough evidence including the test value, critical value, correlation ad determination coefficients, we can reject the null hypothesis.

Task 1.4

```
plot(faithful$eruptions~faithful$waiting,
               xlab = "Waiting Time",
               ylab = "Eruption Time",
               main = "Waiting Time vs Eruption Time Plot",
               pch=18,col="orange",
               frame=T,
               las=1)
abline(lm(faithful$eruptions~faithful$waiting),
       col="brown",
       lty=1,
       lwd=3)
```



Task 1.5

```
SmyData = lm(formula = faithful$eruptions~faithful$waiting, data = faithful)
summary(SmyData)
```

```
##
## Call:
## lm(formula = faithful$eruptions ~ faithful$waiting, data = faithful)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.874016   0.160143  -11.70   <2e-16 ***
## faithful$waiting  0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Task 1.6

```
ww<-summary(lm(faithful$eruptions~faithful$waiting))
dr<-attributes(ww)
knitr::kable(dr)
```

x
call
terms
residuals
coefficients
aliased
sigma
df
r.squared
adj.r.squared
fstatistic
cov.unscaled

x
summary.lm

Task 1.7

```
uu <- lm(faithful$eruptions~faithful$waiting)

faithful$new_expect_y <- uu$fitted.values
gg <- faithful$eruptions-faithful$new_expect_y
faithful$R_val <- gg
qty <-rbind(faithful[1:3,],faithful[270:272,])
rownames(qty) <- c(1,2,3,4,5,6)
knitr::kable(qty)
```

eruptions	waiting	new_expect_y	R_val
3.600	79	4.100592	-0.5005919
1.800	54	2.209893	-0.4098932
3.333	74	3.722452	-0.3894522
4.417	90	4.932499	-0.5154993
1.817	46	1.604870	0.2121304
4.467	74	3.722452	0.7445478

Task 1.8

Equation is represented as:

$$\hat{y} = a + bx$$

Task 1.9

20 minutes waiting period have expected eruption time of: `-0.361456` minutes.
100 minutes waiting period have expected eruption time of: `5.688784` minutes.

Task 2:

Task 2.1

a) Linear regression model has a equation: **Y'= a + bX**. Similarly Multiple regression is: **Y'= a + b1X1 + b2X2 +…**

```
SystolicBP = c(132,143,153,162,154,168,137,149,159,128,166)
age = c(52,59,67,73,64,74,54,61,65,46,72)
weight = c(173,184,194,211,196,220,188,188,207,167,217)

TData = data.frame(SystolicBP, age, weight)
t_table <- lm(SystolicBP ~ age + weight,  data=TData)
summary(t_table)
```

```
##
## Call:
## lm(formula = SystolicBP ~ age + weight, data = TData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4640 -1.1949 -0.4078  1.8511  2.6981
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.9941    11.9438   2.595  0.03186 *
## age           0.8614     0.2482   3.470  0.00844 **
## weight        0.3349     0.1307   2.563  0.03351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.318 on 8 degrees of freedom
## Multiple R-squared:  0.9768, Adjusted R-squared:  0.9711
## F-statistic: 168.8 on 2 and 8 DF,  p-value: 2.874e-07
```

Therefore from above analysis summary, the multiple regression formula is represented as: **Y'= 30.994 + 0.861(X1) + 0.334(X2)**

b)

```
#Coefficients
cor_age=cor(SystolicBP, age)
print(paste("Correlation coeff for age vs systolic blood pressure is",cor_age))
```

```
## [1] "Correlation coeff for age vs systolic blood pressure is 0.978693374257093"
```

```
det_age=cor_age^2
print(paste("Determination coeff for age vs systolic blood pressure is",det_age))
```

```
## [1] "Determination coeff for age vs systolic blood pressure is 0.957840720814735"
```

```
cor_weight=cor(SystolicBP, weight)
print(paste("Coefficient of correlation for weight vs systolic blood pressure is",cor_weight))
```

```
## [1] "Coefficient of correlation for weight vs systolic blood pressure is 0.970564376293148"
```

```
det_weight=cor_weight^2
print(paste("Coefficient of determination for weight vs systolic blood pressure is",det_weight))
```

```
## [1] "Coefficient of determination for weight vs systolic blood pressure is 0.941995208529306"
```

```
#Table
matx1 <- matrix(c(cor_age,det_age,cor_weight,det_weight),nrow=4,ncol=1,byrow=F)
rownames(matx1)<-c("Correlation coeff for Age vs Systolic blood pressure","Detemination coeff for Age vs Systolic
blood pressure","Correlation coeff for Weight vs Systolic blood pressure","Detemination coeff for Weight vs Systo
lic blood pressure")
colnames(matx1)<-"VALUE"
knitr::kable(matx1)
```

	VALUE

Correlation coeff for Age vs Systolic blood pressure	0.9786934
Detemination coeff for Age vs Systolic blood pressure	0.9578407
Correlation coeff for Weight vs Systolic blood pressure	0.9705644
Detemination coeff for Weight vs Systolic blood pressure	0.9419952

c)

```
corf_aw=cor(age, weight)
Multiple_R_coef= sqrt((cor_age^2+cor_weight^2 - 2*cor_age*cor_weight*corf_aw) / (1-corf_aw^2))
Mul_R_square=Multiple_R_coef^2
matx2<-matrix(c(Multiple_R_coef,Mul_R_square),nrow = 2,ncol=1,byrow=F)
rownames(matx2)<-c("Multiple R value is","Multiple R square value is")
colnames(matx2)<-"VALUE"
knitr::kable(matx2)
```

	VALUE
Multiple R value is	0.9883558
Multiple R square value is	0.9768471

d)

```
F_test <- (Mul_R_square/2)/((1-Mul_R_square)/8)
Tst_Value <- Multiple_R_coef*(sqrt(9/1-Mul_R_square))
Crit_Val <- qt(1-0.01/2, 9)
matx3<-matrix(c(F_test, Tst_Value, Crit_Val),nrow = 3,ncol=1,byrow=F)
rownames(matx3)<-c("F Test value is","Hypothesis Test Result is","Critical Value is")
colnames(matx3)<-"VALUE"
knitr::kable(matx3)
```

	VALUE
F Test value is	168.764566
Hypothesis Test Result is	2.799535
Critical Value is	3.249836

e)

Based on the above results obtained along with the scatter plots, we can say that using with two independent variables **'age' & 'weight'** can be used to predict dependent variable **'Systolic Blood Pressure'** as both the variable are **strongly correlated** with Systolic Blood Pressure with respective correlation coefficients of **0.978 and 0.97** respectively.

f)

```
#We can use the equation Y = a1X1 + a2X2 + C where a1,a2=slopes & X1,X2=variables
Y1 = (0.861*45)+(0.334*135)+30.994
matx5<-matrix(c(Y1),nrow = 1,ncol=1,byrow=F)
rownames(matx5)<-c("The Systolic BP at weight 135 vs age 45 is:")
colnames(matx5)<-"VALUE"
knitr::kable(matx5)
```

	VALUE
The Systolic BP at weight 135 vs age 45 is:	114.829

g)

```
#We can use the equation Y = a1X1 + a2X2 + C where a1,a2=slopes & X1,X2=variables
Y2 = (0.861*60)+(0.334*182)+30.994
matx6<-matrix(c(Y2),nrow = 1,ncol=1,byrow=F)
rownames(matx6)<-c("The Systolic BP at weight 182 vs age 60 is:")
colnames(matx6)<-"VALUE"
knitr::kable(matx6)
```

	VALUE

The Systolic BP at weight 182 vs age 60 is: 143.442

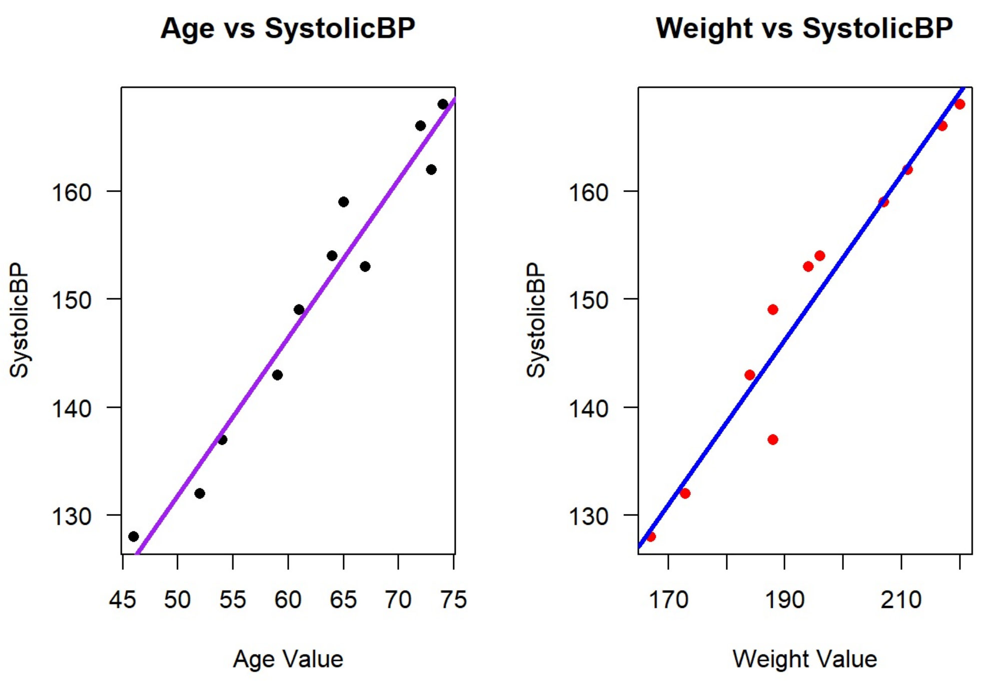
OBSERVATIONS:
From the above task analysis, we can observe that both independent variables age and weight are strongly correlated to dependent variable systolic blood pressure. Thus they have positive correlation between them. Here we also performed and calculated F Test value to check whether the regression model is a better fir to the data provided. At the end with the help of linear regression equation derived, we calculated expected Systolic BP for respective values of variables.

Task 2.2

```
#Scatter Plots
par(mfrow=c(1,2))

aplot <- plot(SystolicBP~age,
               xlab = "Age Value",
               ylab = "SystolicBP",
               main = "Age vs SystolicBP",
               pch=16,col="black",
               frame=T,
               las=1)
abline(lm(SystolicBP~age),
       col="purple",
       lty=1,
       lwd=3)

bplot <- plot(SystolicBP~weight,
               xlab = "Weight Value",
               ylab = "SystolicBP",
               main = "Weight vs SystolicBP",
               pch=16,col="red",
               frame=T,
               las=1)
abline(lm(SystolicBP~weight),
       col="blue",
       lty=1,
       lwd=3)
```



PART 3: CONCLUSIONS

Looking at all the task analysis performed in this project, it helped me to gain overall experience regarding the linear and multiple regression analysis. The main aim of performing the mentioned tasks in this project to to understand why the concept of correlation was invented along with the enhancement in the applications of regressions techniques used in modern statistics.
Firstly, the overall Introduction section gave a brief ideology and the importance of using the concept of linear and multiple regression. As mentioned Frascis Galton along with mathematician Person contributed their ideas of regression analysis for modern statistics. We alsp learned and understood how this concept in modern industry like in Businesses for sales strategies in finance and other industries. This made me much clear how the concept is well formulated and applied in modern statistics.

First section of analysis consisted of simple linear regression analysis. These sub tasks taught how to import data set named ‘faithful’ which is pre-existing eruption data set loaded in R Studio. The main focus in these tasks was to understand the usage of “r in-line” codes. According to professors guidance and steps, I was able to perform and complete tasks using “r in-line” which was a completely new thing for me to learn other than using R-chunk codes. Along with it quantile function were studied as this was refreshing analysis from previous projects. Finally, hypothesis testing was carried out based on the faithful data set. Here I learned to write in-line hypothesis statements with the help provided by the professor. Further we gathered evidence like critical value and test value which helped to reject the null hypothesis. Plotting scatter plots with regression lines with modifications and presenting the parameter values using knitr::kable() function was another new skill concept learned by me in this project.

In the second part of analysis section, we performed multiple regression analysis. By the concept of multiple regression this data set had 2 independent variables and 1 dependent variable. It is a medical data set of variables systolic BP, age and weight. Overall this analysis helped to calculate F test, critical values and multiple regression equation. We also used matrix functions with knitr::kable() function to organize and present the values of regression analysis in a tabular manner.
New skills gained my me in this project was to use matrix function with knitr::kable() function, summary() was quite challenging as I learned from my previous mistakes of how precisely a code has to be written. Other skills learned were punctuality and regularity is need to complete these task analysis. This project helped me gained many statistical skills like regression, correlation and multiple descriptive statistics techniques with will overall hone my approach to statistics and insights generated. This will help me in future while working in a practical industry of interest.
Only recommendation I would make is to making these tasks analysis and projects as a group project as these projects makes it lengthy for one individual to complete and understand all of the concepts at full potential. Overall, it is innovative project.

BIBLIOGRAPHY

1) Simple linear regression. (n.d.). Wayne W. LaMorte, MD, PhD, MPH Retrieved May 31, 2016., from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/ (https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/)
2) Bevans, R. (2020, October 26). An introduction to multiple linear regression. Scribbr. Retrieved December 18, 2021, from https://www.scribbr.com/statistics/multiple-linear-regression/ (https://www.scribbr.com/statistics/multiple-linear-regression/)
3) Jeffrey M. Stanton. 01 Dec 2017, from https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537 (https://www.tandfonline.com/doi/full/10.1080/10691898.2001.11910537)
4) Saint-Leger, R. (2016, October 26). Business forecasting using historical data and regression analysis. Small Business - Chron.com. Retrieved December 12, 2021, from https://smallbusiness.chron.com/business-forecasting-using-historical-data-regression-anlaysis-32844.html (https://smallbusiness.chron.com/business-forecasting-using-historical-data-regression-anlaysis-32844.html)
5) Zach. (2021, February 12). Multiple R vs. R-squared: What’s the difference? Statology. Retrieved December 19, 2021, from https://www.statology.org/multiple-r-vs-r-squared/ (https://www.statology.org/multiple-r-vs-r-squared/)

PART 4: APPENDIX

A R markdown file names “M6_Project_Mestry.Rmd” and a HTML file named “M6_Project_Mestry.html” is been attached and submitted in form of report.

ACKNOWLEDGMENTS

I would like to present as well as express my gratitude and accountability to the to all the colleagues and mainly Northeastern University to give me this opportunity to complete the Final Project report. I want to thank specially to my **Professor Dr.Dee Chiluiza** whose help, stimulating suggestions on doubts and giving clarity on all the task completion which eventually encouraged me for precise fabrication of the final report.
I would also like to thank the **TA - Yameng Guo** who regulary conducted doubt section if every week which overall helped me to understand and clear my doubt along with the professor in the class as well in zoom meeting conducted by the TA.