Appendix

#Code for Desctriptive Statistics Analysis

```
LE1<-read.csv("D:/col.name_changed_raw data.csv",header=TRUE)

#numerical representation

#descriptive_statistics=measure of central tendency & variance

#m.of.central tendency=mean,mode&median

#m.0f.variability=range,variance&standard deviation

nrow<-nrow(LE1)

ncol<-ncol(LE1)

dim(LE1)

print(LE1,limit=5)

attributes(LE1)

names(LE1)

summary(LE1)

install.packages('modeest')

library('modeest')

med<- median(LE1$life_expectency,na.rm = TRUE)

mean<- mean(LE1$life_expectency,na.rm = TRUE)

mode<-mfv(LE1$life_expectency,na_rm = TRUE)

max<-max(LE1$life_expectency,na.rm=TRUE)
```

```
min<-min(LE1$life_expectency,na.rm=TRUE)

range<-max-min

variance<-var(LE1$life_expectency,na.rm = TRUE)

standard_deviation<-sd(LE1$life_expectency,na.rm = TRUE)

quartiles<-quantile(LE1$life_expectency,na.rm = TRUE)

IQR<-IQR(LE1$life_expectency,na.rm = TRUE)

# Kurtosis and skewness:

install.packages('e1071')

library('e1071')

skewness<-skewness(LE1$life_expectency,na.rm = TRUE)

kurtosis<-kurtosis(LE1$life_expectency,na.rm = TRUE)

#Hence the data is negatively skewed with the majority of data values greater than mean

# Histogram with mean,median and density curve.

# Histogram with density instead of count on y-axis

install.packages('ggplot')

library('ggplot2')

install.packages('moments')

library('moments')

graphical_data<- ggplot(LE1, aes(x=life_expectency)) +

geom_histogram(aes(y=..density..),binwidth=.5,
```

```
                colour="black", fill="white")+
geom_density(alpha=.2, fill="green") + geom_vline(aes(xintercept=med),
color="blue", lwd=1)
graphical_data
graphical_data+geom_vline(aes(xintercept=mean),
color="red", linetype="dashed", size=1)
```

#Code for Cleaning data by imputation concept

```
#checking nan's
library('dplyr')
install.packages('naniar')
library('naniar')
LE1%>%summary(LE1)
# Removing the columns having more than 75% missing data. as they having more than 75% missing values they won't contribute to the model.
threshold<-0.75 #for a 75% cut-off
LE1 <- LE1 %>% select(where(~mean(is.na(.))< threshold))
head(LE1)
install.packages('car')
library('car')
```

```
qqplot<-qqPlot(LE1$life_expectency)

#range of life expectancy

range<-range(LE1$life_expectency,na.rm=TRUE)

dim(LE1)

colnames(LE1)

head(LE1)

install.packages('mice')

library('mice')

install.packages('tidyverse')

library('tidyverse')

#contains null for all the columns except one

#LE1%>% slice(218:266)

LE2<-LE1%>% slice(1:217)

md.pattern(LE1)

sapply(LE1, function(x) sum(is.na(x)))

names(LE1)

cols <- c("LE2$LE_t_birth","LE2$acc_elect","LE2$adj_NNI",

        "LE2$NNI_capit","LE2$HIV(0-14)","LE2$not_prim",

        "LE2$prim_25+","LE2$Bch_25+","LE2$inf_mort",

        "LE2$prim=age","LE2$lit_rate","LE2$real_int",
```

```
        "LE2$pop_grow","LE2$pop_dense","LE2$pop_total",
        "LE2$hlth_capit","LE2$hlth_GDP","LE2$unemp",
        "LE2$GDP_grth","LE2$GDP_capit","LE2$crude_brth",
        "LE2$renew_eng","LE2$HIV(15-49)","LE2$safe_wtr","LE2$pov_pop")
cols
 LE3 <- LE2[1:2]
#LE1 = data.frame()
LE3.names <- c("C_Name", "C_code","comp_edu")
data.frame(LE3, stringsAsFactors = TRUE)
LE3$C_Name<-LE2$C_Name
LE3$comp_edu<-LE2$comp_edu
LE3
LE2$C_Name
drop <- c("C_Name", "Continent","C_Code","comp_edu")
LE2 <- LE2[,!(names(LE2) %in% drop)]
names(LE2)
LE2 <- log(LE2)
imputations <- mice(LE2,method = 'pmm')
print(imputations,limit=5)
complete(imputations)
```

```
newdataframe<-complete(imputations)

newdataframe

stripplot(imputations, pch = 20, cex = 1.2)

sapply(newdataframe, function(x) sum(is.na(x)))

drop_column<- c("renew_eng", "pov_pop","not_prim","GDP_grth")

newdataframe= newdataframe[,!(names(newdataframe) %in% drop_column)]

sapply(newdataframe, function(x) sum(is.na(x)))

install.packages("xlsx", dependencies=TRUE)

write.csv(newdataframe, "D:/Analysed_data.csv", row.names=FALSE)
```

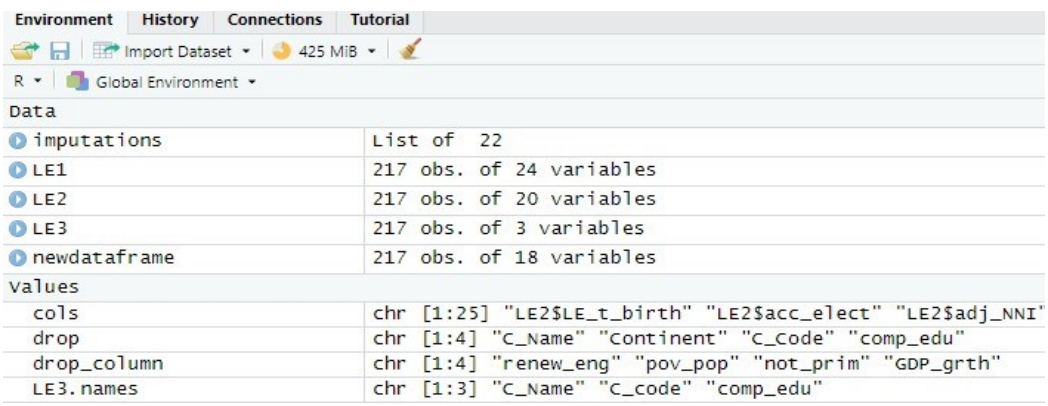

Figure 1: Dataset after imputation

Source: Created in RStudio

#Code for fixing Collinearity

```
install.packages('faraway')

library("faraway")
```

```
newdataframe<-read.csv("D:/Analysed_data.csv",header=TRUE)

head(newdataframe)

names(newdataframe)

model1<-lm(newdataframe$life_expectency~.,data=newdataframe)

summary(model1)

corr<-cor(newdataframe)

corrround(corr,digits=2)

library(corrplot)

install.packages('mctest')

library('mctest')

corrplot.mixed(corr, lower.col = "black", number.cex=.5)

vif_score<-vif(newdataframe)

vif_score

model2<-lm(newdataframe$life_expectency~newdataframe$crude_brth+
newdataframe$Inf_mor+newdataframe$hlth_capit)

summary(model1)

model3<-lm(newdataframe$life_expectency~newdataframe$crude_brth+

newdataframe$Inf_mor+newdataframe$GDP_capit)

summary(model3)

anova(model2,model1)

anova(model3,model1)
```

```
drop_column<-c("crude_brth")

newdata<-newdataframe[,!names(newdataframe)%in%drop_column]

View(newdata)
```

Filter

	life_expectency	Inf_mort	acc_elect	adj_NNI	NNI_capit	HIV.0.14.	HIV.15.49.	prim.age	unempl	real_int	pop_grow	pop_dens	pop_total
1	4.171815	3.8372995	4.581902	2.377936155	1.939102890	5.298317	7.170120	4.434745	3.2580965	2.5449492	0.838577094	4.0650770	17.4541
2	4.364028	2.1517622	4.605170	-1.920888088	-0.555094139	4.605170	4.605170	4.637858	2.4397350	1.6082069	-0.784480961	4.6460007	14.8642
3	4.342246	2.9957323	4.600158	1.077779874	-0.034105009	5.298317	7.244228	4.618724	2.7166795	2.1409885	0.659581785	2.8946014	17.5779
4	4.310463	2.3025851	4.605170	-0.363537804	-0.841220619	4.605170	4.605170	4.761128	2.8249440	1.9623181	0.593018779	5.6224278	10.9207
5	4.381133	0.9162907	4.605170	1.394135471	2.783276031	4.605170	4.605170	4.656109	1.6620304	1.4037770	-1.720136065	5.1007223	11.2534
6	4.113281	3.9100210	3.821449	0.115621130	-6.401888731	8.732305	9.680344	4.002388	2.5273274	2.0675820	1.176472317	3.2397615	17.2757
7	4.344013	1.7227666	4.605170	1.627288395	1.891085266	4.605170	4.605170	4.650232	2.4414771	2.1265402	-0.149143157	5.3968764	11.4836
8	4.339471	2.0918641	4.605170	2.124964358	1.334433208	4.605170	8.575462	4.590536	2.2864557	2.3815776	-0.006624400	2.7985504	17.6208

Figure2: New dataset after reducing collinearity

Source: Created in RStudio

#Code for best fit of Multiple Linear Regression model

```
library(olsrr)

library(leaps)

newdataframe<-read.csv("D:/Analysed_data.csv",header=TRUE)

head(newdataframe)

#build full model using all columns

full.model<-lm(life_expectency ~ .,data=newdataframe)

summary(full.model)

#built reduced model from the features obtained
```

```
reduced.model<-lm(life_expectency ~ crude_brth+adj_NNI+NNI_capit +
acc_elect + HIV.0.14.+pop_dens+Inf_mort +
real_int +pop_total +pop_grow +prim.age + HIV.15.49.,data=newdataframe)
summary(reduced.model)


#analysis of variance table,which calculates the sum of squares for each variable
anova(full.model)
#comparing the full and the reduced models
anova(reduced.model,full.model)
#plots the standardized residuals against fitted values for FULL model
stdres_fullmodel<-rstandard(full.model)
par(mfrow=c(2,2))
plot(full.model$fitted.values,stdres_fullmodel,pch=16,
ylab="Standardized Residuals",xlab="fitted y",
ylim=c(-3,3), main="Full model",col=("red"))
abline(h=0,lwd=2,col="blue")
abline(h=2,lty=2,lwd=2,col="green")
abline(h=-2,lty=2,lwd=2,col="green")
#plots the QQ-plot for full model
qqnorm(stdres_fullmodel, ylab="Standardized Residuals",
```

```
xlab="Normal Scores", main="QQ Plot for Full model",col="red")

qqline(stdres_fullmodel,lwd=2,col="blue")

#plots the standardized residuals against fitted values for FULL model

stdres_reducedmodel<-rstandard(reduced.model)

plot(reduced.model$fitted.values,stdres_reducedmodel,pch=16,

ylab="Standardized Residuals",xlab="fitted y",

ylim=c(-3,3), main="Reduced model",col="red")

abline(h=0,lwd=2,col="blue")

abline(h=2,lty=2,lwd=2,col="green")

abline(h=-2,lty=2,lwd=2,col="green")

#plots the QQ-plot for full model

qqnorm(stdres_reducedmodel, ylab="Standardized Residuals",

xlab="Normal Scores", main="QQ Plot for Reduced model",col="red")

qqline(stdres_reducedmodel,lwd=2,col="blue")

#An alternative way to get the above plots

par(mfrow=c(2,2))

plot(full.model,main="FULL Model",col=c("purple"),lwd=2)

par(mfrow=c(2,2))

plot(reduced.model,main="REDUCED Model",col=c("purple"),lwd=2)

#AIC values for full and reduced models
```

```
AIC(full.model)

AIC(reduced.model)

#Calculating the Mallow's Cp

library(olsrr)

ols_mallows_cp(full.model,reduced.model)

#Calculate min cp from all models

full.model.cp<-lm(life_expectency~.,data=newdataframe,x=TRUE) #note the additional x=TRUE term

#define our y variablk\es & design matrix X:

X <- full.model.cp$x

y <-newdataframe$life_expectency

library('leaps')

all.models<- leaps(X, y, int = FALSE, strictly.compatible = FALSE, method="Cp")

#Plot all cp

plot(all.models$size, all.models$Cp, log="y", xlab="|M|", ylab=expression(C[p]),ylim=c(1,200),col="blue")

lines(all.models$size, all.models$size,col="red",lwd=2)

#evaluate min cp and consider columns rated to it

min.cp<- all.models$Cp == min(all.models$Cp)

min.cp #this finds the smallest C_p value

min(all.models$Cp) #gives the min C_p value
```

```
min.cp<- all.models$which[min.cp, ] #this finds the corresponding model with the smallest C_p

min.cp #this lists the parameters included in the model

#we can save this as a multiple linear model

#best.model.cp<-(lm())

#summary(best.model.cp)

#AIC(best.model.cp)

#AIC(full.model.cp)

#stepwise selection

install.packages('faraway')

library("faraway")

newdataframe<-read.csv("D:/Analysed_data.csv",header=TRUE)

head(newdataframe)

#Forward feature selection

model_1 <- lm(life_expectency ~ 1 , data = newdataframe)

model_1

names(model_1)

step_1 <- step(model_1 ,   scope =
~crude_brth+Inf_mort+acc_elect+adj_NNI+NNI_capit+HIV.0.14.+HIV.15.49.+prim.age+un
empl+real_int+pop_grow+pop_dens+pop_total+hlth_capit+hlth_GDP+GDP_capit+safe_wtr,
data = newdataframe,method = "forward")

summary(step_1)
```

```
#Backward feature selection

model_2 <- lm(life_expectency ~ . , data = newdataframe)

step_2 <- step(model_2,method = "backward")

summary(step_2)
```

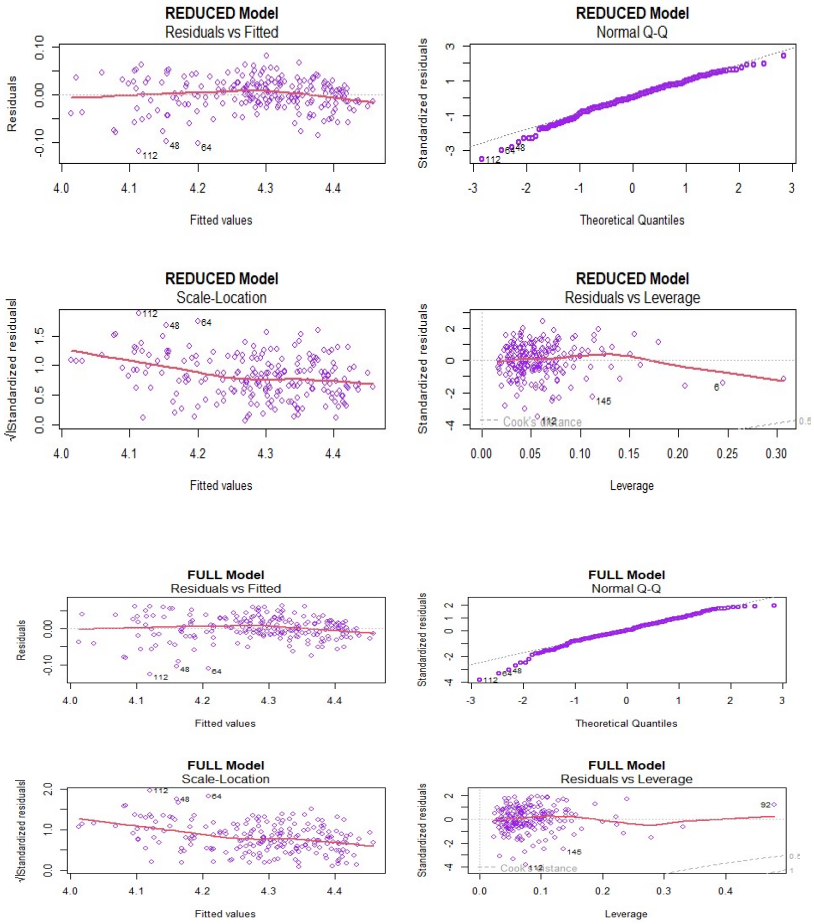

Figure 3: Alternate view of Full Model vs Reduced Model

Source: Created in RStudio

```
#Stepwise feature selection/regression

#stepwise selection

initial_model<- glm(life_expectency ~ . , data = newdataframe)

selection <- step(initial_model ,

                  scope = list(lower = ~ 1 ) ,

                  data = newdataframe ,

                  direction = "both")

selection

#eval <- selection$coefficients

#eval

#summary(selection)

#reduced model based on the stepwise feature selection

stepwise_reduced.model<- lm( life_expectency ~ crude_brth + Inf_mort + acc_elect +
adj_NNI + NNI_capit + HIV.0.14. +HIV.15.49. + prim.age + real_int + pop_grow +
pop_dens + pop_total + hlth_capit + safe_wtr, data=newdataframe)

best.model.cp<- stepwise_reduced.model

summary(best.model.cp)

#Standardized full model and reduced model

stdres_best.model.cp<-rstandard(best.model.cp)

AIC(best.model.cp)

ols_mallows_cp(best.model.cp,full.model)
```

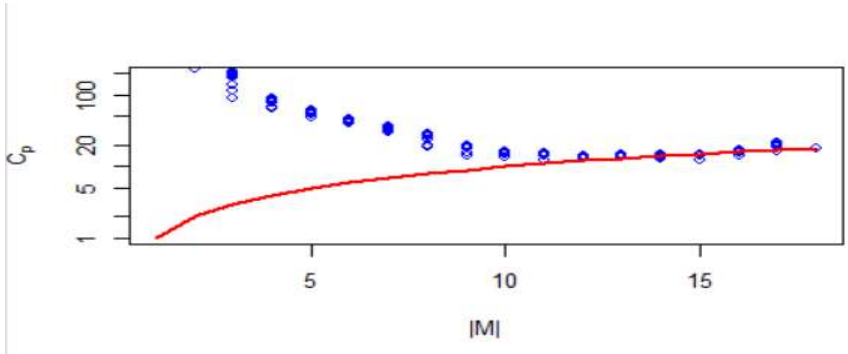

Figure 4: Cp score using leaps

#Code for finding differences of average life expectancies across the continents using One wayAnowa Method

```
#Reading the imputed data from csv file after the working directory is changed

#Reading the data from csv file after the working directory is changed

df2<-read.csv("C://Users//44776//Downloads//Life_Expectancy_Data1.csv",header=TRUE)

dim(df2)

install.packages('mice')

library(mice)

#Perform imputation for NA values using MICE

imputations<-mice(df2,method = "cart")

df2<-complete(imputations,1)

#calculating the group means of continents

group.means<-tapply(df2$SP.DYN.LE00.IN,df2$Continent,mean)

group.means
```

```
#boxplot of Life Expectancy vs Continent

boxplot(df2$SP.DYN.LE00.IN~df2$Continent,main='Life Expectancies versus continents', xlab='Continent', col="sky blue", ylab = "Life Expectancy ",)

#Shapiro-Wilk normality test

anova1way<-aov(df2$SP.DYN.LE00.IN~as.factor(df2$Continent),data=df2)

summary(anova1way)

df2$residuals<-anova1way$residuals

shapiro.test(df2$residuals)

#Levene's Test for Homogeneity of Variance

install.packages('car')

library(car)

leveneTest(df2$SP.DYN.LE00.IN~factor(df2$Continent))

#One-way analysis of means (not assuming equal variances) using Welch test

data.welchtest<-oneway.test(df2$SP.DYN.LE00.IN~factor(df2$Continent),data=df2)

data.welchtest

#Pairwise comparsions

#Bonferroni post-hoc test

cat("Bonferroni post-hoc test","\n")

pairwise.t.test(df2$SP.DYN.LE00.IN,df2$Continent,p.adj="bonferroni")

#Tukey post-hoc test

cat("/n","Tukey post-hoc test","\n")
```

tukey.data<-TukeyHSD(anova1way)

tukey.data

plot(tukey.data)

```
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
PD.CHEX.PC.CD  SH.XPD.CHEX.GD.ZS  SL.UEM.TOTL.NE.ZS  NY.GDP.MKTP.KD.ZG  NY.GDP.PCAP.CD  SP.DYN.CBRT.IN  SH.HIV.INCD  SH.H2O.S
MDW.ZS  SI.POV.LMIC  SE.COM.DURS
  5   3  SP.DYN.LE00.IN  EG.ELC.ACCS.ZS  NY.ADJ.NNTY.KD.ZG  NY.ADJ.NNTY.PC.KD.ZG  SH.HIV.INCD.14  SE.PRM.UNER  SE.PRM.CUAT.ZS
  SE.TER.CUAT.BA.ZS  SP.DYN.IMRT.IN  SE.PRM.CMPT.ZS  SE.ADT.LITR.ZS  FR.INR.RINR  SP.POP.GROW  EN.POP.DNST  SP.POP.TOTL  SH.X
PD.CHEX.PC.CD  SH.XPD.CHEX.GD.ZS  SL.UEM.TOTL.NE.ZS  NY.GDP.MKTP.KD.ZG  NY.GDP.PCAP.CD  SP.DYN.CBRT.IN  SH.HIV.INCD  SH.H2O.S
MDW.ZS  SI.POV.LMIC  SE.COM.DURS
  5   4  SP.DYN.LE00.IN  EG.ELC.ACCS.ZS  NY.ADJ.NNTY.KD.ZG  NY.ADJ.NNTY.PC.KD.ZG  SH.HIV.INCD.14  SE.PRM.UNER  SE.PRM.CUAT.ZS
  SE.TER.CUAT.BA.ZS  SP.DYN.IMRT.IN  SE.PRM.CMPT.ZS  SE.ADT.LITR.ZS  FR.INR.RINR  SP.POP.GROW  EN.POP.DNST  SP.POP.TOTL  SH.X
PD.CHEX.PC.CD  SH.XPD.CHEX.GD.ZS  SL.UEM.TOTL.NE.ZS  NY.GDP.MKTP.KD.ZG  NY.GDP.PCAP.CD  SP.DYN.CBRT.IN  SH.HIV.INCD  SH.H2O.S
MDW.ZS  SI.POV.LMIC  SE.COM.DURS
  5   5  SP.DYN.LE00.IN  EG.ELC.ACCS.ZS  NY.ADJ.NNTY.KD.ZG  NY.ADJ.NNTY.PC.KD.ZG  SH.HIV.INCD.14  SE.PRM.UNER  SE.PRM.CUAT.ZS
  SE.TER.CUAT.BA.ZS  SP.DYN.IMRT.IN  SE.PRM.CMPT.ZS  SE.ADT.LITR.ZS  FR.INR.RINR  SP.POP.GROW  EN.POP.DNST  SP.POP.TOTL  SH.X
PD.CHEX.PC.CD  SH.XPD.CHEX.GD.ZS  SL.UEM.TOTL.NE.ZS  NY.GDP.MKTP.KD.ZG  NY.GDP.PCAP.CD  SP.DYN.CBRT.IN  SH.HIV.INCD  SH.H2O.S
MDW.ZS  SI.POV.LMIC  SE.COM.DURS
Warning message:
Number of logged events: 78
> df2<-complete(imputations,1)
> group.means<-tapply(df2$SP.DYN.LE00.IN,df2$Continent,mean)
> group.means
          Africa              Asia Australia/Oceania            Europe     North America     South America
        64.11014          74.61739          73.87860          79.46093          75.99767          75.09100
>
```

Figure 5: Group Means of different Continents

```
46:1   (Top Level) ÷                                                    R Script ÷
Console   Terminal ×   Background Jobs ×
R 4.2.2 · ~/
Fit: aov(formula = df2$SP.DYN.LE00.IN ~ as.factor(df2$Continent), data = df2)

$`as.factor(df2$Continent)`
                                   diff       lwr        upr     p adj
Asia-Africa                   10.5072495  7.778881 13.2356180 0.0000000
Australia/Oceania-Africa       9.7684610  6.060322 13.4765998 0.0000000
Europe-Africa                 15.3507913 12.593067 18.1085155 0.0000000
North America-Africa          11.8875221  8.844023 14.9310216 0.0000000
South America-Africa          10.9808564  6.544232 15.4174812 0.0000000
Australia/Oceania-Asia        -0.7387885 -4.485334  3.0077567 0.9930198
Europe-Asia                    4.8435418  2.034387  7.6526965 0.0000214
North America-Asia             1.3802725 -1.709905  4.4704500 0.7931804
South America-Asia             0.4736068 -3.995168  4.9423815 0.9996442
Europe-Australia/Oceania       5.5823303  1.814354  9.3503071 0.0004358
North America-Australia/Oceania  2.1190611 -1.862835  6.1009571 0.6449399
South America-Australia/Oceania  1.2123954 -3.913641  6.3384315 0.9839700
North America-Europe          -3.4632692 -6.579396 -0.3471427 0.0197115
South America-Europe          -4.3699350 -8.856693  0.1168228 0.0611547
South America-North America   -0.9066657 -5.574518  3.7611862 0.9934930
```

Figure 6: Output of Tukey post-hoc test on Continents