SRI RAMACHANDRA
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Category - I Deemed to be University) Porur, Chennai
SRI RAMACHANDRA ENGINEERING AND TECHNOLOGY

CSE 240 – DATA SCIENCE WITH R

NAME	JOSELYN DIANA CINDRELLA M
UNIQUE ID	E0120017
YEAR	II
QUARTER	Q1
DEPARTMENT	AI & ML
FACULTY NAME	Prof.RAMYA M
ACADEMIC YEAR	2021-2022

PROJECT REPORT

CONTENTS

WORLD HAPPINESS REPORT -2020

PROBLEM STATEMENT:

Social well-being of citizens in a country is not only defined by the Macro-economic factors such as Gross Domestic Product (GDP) Indicators but also by social indicators such as Freedom to make life Choices and Healthy Life Expectancy.

The dataset contains world happiness report-2020, it analyzes Happiness score based on key factors like health life **expectancy, Freedom, generosity, perceptions of corruption, social support.**

OBJECTIVE:

- To analyze the happiness score across all countries With the given dataset
- To select appropriate prediction model algorithm that combines several features to estimate the happiness score.
- To determine what factors affect the decrease in happiness score
- To visualize the average health across the region ,GDP,social support based on happiness

DATA LOADING:

Loading all packages:

```
####Loading packages
install.packages("dplyr")
library(dplyr)
install.packages("tidyverse")
library(tidyverse)
install.packages("caTools")
library(caTools)
install.packages("ggplot2")
library(ggplot2)
install.packages("reshape2")
library(reshape2)
install.packages("tidyr")
library(tidyr)
install.packages("ggthemes")
library(ggthemes)
install.packages("corrgram")
library(corrgram)
install.packages("corrplot")
library(corrplot)
install.packages("rpart")
library(rpart)
install.packages("plotrix")
library(plotrix)
```

Code:

```
#Loading Data set(happiness_2020)
happiness_2020 <- read.csv("2020.csv")
print(getwd())
head(happiness_2020,10)
```

Output:

```
> happiness_2020 <- read.csv("2020.csv")
> print(getwd())
[1] "/Users/apple"
> head(happiness_2020,2)
  Country.name Regional.indicator Ladder.score Standard.error.of.ladder.score upperwhisker lowerwhisker
1      Finland     Western Europe       7.8087                     0.03115630     7.869766     7.747634
2      Denmark     Western Europe       7.6456                     0.03349229     7.711245     7.579955
  Logged.GDP.per.capita Social.support Healthy.life.expectancy Freedom.to.make.life.choices  Generosity
1              10.63927      0.9543297               71.90083                    0.9491722 -0.05948202
2              10.77400      0.9559908               72.40250                    0.9514443  0.06620178
  Perceptions.of.corruption Ladder.score.in.Dystopia Explained.by..Log.GDP.per.capita Explained.by..Social.support
1                 0.1954446                 1.972317                         1.285190                    1.499526
2                 0.1684895                 1.972317                         1.326949                    1.503449
  Explained.by..Healthy.life.expectancy Explained.by..Freedom.to.make.life.choices Explained.by..Generosity
1                             0.9612714                                  0.6623167               0.1596704
2                             0.9793326                                  0.6650399               0.2427934
  Explained.by..Perceptions.of.corruption Dystopia...residual
1                               0.4778573            2.762835
2                               0.4952603            2.432741
```

DATA EXPLORATION:

Code:

```
#structure of the data
str(happiness_2020)
```

```
> str(happiness_2020)
'data.frame':   153 obs. of  20 variables:
 $ Country.name                            : chr  "Finland" "Denmark" "Switzerland" "Iceland" ...
 $ Regional.indicator                      : chr  "Western Europe" "Western Europe" "Western Europe" "Western Europe" ...
 $ Ladder.score                            : num  7.81 7.65 7.56 7.5 7.49 ...
 $ Standard.error.of.ladder.score          : num  0.0312 0.0335 0.035 0.0596 0.0348 ...
 $ upperwhisker                            : num  7.87 7.71 7.63 7.62 7.56 ...
 $ lowerwhisker                            : num  7.75 7.58 7.49 7.39 7.42 ...
 $ Logged.GDP.per.capita                   : num  10.6 10.8 11 10.8 11.1 ...
 $ Social.support                          : num  0.954 0.956 0.943 0.975 0.952 ...
 $ Healthy.life.expectancy                 : num  71.9 72.4 74.1 73 73.2 ...
 $ Freedom.to.make.life.choices            : num  0.949 0.951 0.921 0.949 0.956 ...
 $ Generosity                              : num  -0.0595 0.0662 0.1059 0.2469 0.1345 ...
 $ Perceptions.of.corruption               : num  0.195 0.168 0.304 0.712 0.263 ...
 $ Ladder.score.in.Dystopia                : num  1.97 1.97 1.97 1.97 1.97 ...
 $ Explained.by..Log.GDP.per.capita        : num  1.29 1.33 1.39 1.33 1.42 ...
 $ Explained.by..Social.support            : num  1.5 1.5 1.47 1.55 1.5 ...
 $ Explained.by..Healthy.life.expectancy   : num  0.961 0.979 1.041 1.001 1.008 ...
 $ Explained.by..Freedom.to.make.life.choices: num  0.662 0.665 0.629 0.662 0.67 ...
 $ Explained.by..Generosity                : num  0.16 0.243 0.269 0.362 0.288 ...
 $ Explained.by..Perceptions.of.corruption : num  0.478 0.495 0.408 0.145 0.434 ...
 $ Dystopia...residual                     : num  2.76 2.43 2.35 2.46 2.17 ...
```

Code:

```
#dropping unnecessary columns
happiness_2020 <- happiness_2020[, -c(2,4,5,6,13,14,15,16,17,18,19)]
happiness_2020
head(happiness_2020)
```

‘

Output:

```
> head(happiness_2020)
  Country.name Ladder.score Logged.GDP.per.capita Social.support Healthy.life.expectancy Freedom.to.make.life.choices
1      Finland       7.8087              10.63927      0.9543297                71.90083                      0.9491722
2      Denmark       7.6456              10.77400      0.9559908                72.40250                      0.9514443
3  Switzerland       7.5599              10.97993      0.9428466                74.10245                      0.9213367
4      Iceland       7.5045              10.77256      0.9746696                73.00000                      0.9488919
5       Norway       7.4880              11.08780      0.9524866                73.20078                      0.9557503
6  Netherlands       7.4489              10.81271      0.9391388                72.30092                      0.9085478
   Generosity Perceptions.of.corruption Dystopia...residual
1 -0.05948202                0.1954446            2.762835
2  0.06620178                0.1684895            2.432741
3  0.10591104                0.3037284            2.350267
4  0.24694422                0.7117097            2.460688
5  0.13453263                0.2632182            2.168266
6  0.20761244                0.3647171            2.352117
```

Renaming column –

Code:

```
#Renaming columns for convenience
happiness_2020 <- happiness_2020 %>% rename(c("Country" = "Country.name" ,
                                              "Score" = "Ladder.score",
                                              "GDP" = "Logged.GDP.per.capita",
                                              "Family" = "Social.support",
                                              "Health" = "Healthy.life.expectancy",
                                              "Freedom" = "Freedom.to.make.life.choices",
                                              "Corruption" = "Perceptions.of.corruption",
                                              "Dystopia.residual" = "Dystopia...residual"))

names(happiness_2020)
```

Output:

```
> names(happiness_2020)
[1] "Country"          "Score"            "GDP"              "Family"           "Health"           "Freedom"
[7] "Generosity"       "Corruption"       "Dystopia.residual"
```

Creating a new column “continent”

```
# Creating a new column for continents
happiness_2020$Continent <- NA
happiness_2020$Continent[which(happiness_2020$Country %in% c("Israel", "United Arab Emirates", "Singapore", "Thailand", "Taiwan Province of China",
                                                             "Qatar", "Saudi Arabia", "Kuwait", "Bahrain", "Malaysia", "Uzbekistan", "Japan",
                                                             "South Korea", "Turkmenistan", "Kazakhstan", "Turkey", "Hong Kong S.A.R., China", "Philippines",
                                                             "Jordan", "China", "Pakistan", "Indonesia", "Azerbaijan", "Lebanon", "Vietnam",
                                                             "Tajikistan", "Bhutan", "Kyrgyzstan", "Nepal", "Mongolia", "Palestinian Territories",
                                                             "Iran", "Bangladesh", "Myanmar", "Iraq", "Sri Lanka", "Armenia", "India", "Georgia",
                                                             "Cambodia", "Afghanistan", "Yemen", "Syria"))] <- "Asia"
happiness_2020$Continent[which(happiness_2020$Country %in% c("Norway", "Denmark", "Iceland", "Switzerland", "Finland",
                                                             "Netherlands", "Sweden", "Austria", "Ireland", "Germany",
                                                             "Belgium", "Luxembourg", "United Kingdom", "Czech Republic",
                                                             "Malta", "France", "Spain", "Slovakia", "Poland", "Italy",
                                                             "Russia", "Lithuania", "Latvia", "Moldova", "Romania",
                                                             "Slovenia", "North Cyprus", "Cyprus", "Estonia", "Belarus",
                                                             "Serbia", "Hungary", "Croatia", "Kosovo", "Montenegro",
                                                             "Greece", "Portugal", "Bosnia and Herzegovina", "Macedonia",
                                                             "Bulgaria", "Albania", "Ukraine"))] <- "Europe"
happiness_2020$Continent[which(happiness_2020$Country %in% c("Canada", "Costa Rica", "United States", "Mexico",
                                                             "Panama","Trinidad and Tobago", "El Salvador", "Belize", "Guatemala",
                                                             "Jamaica", "Nicaragua", "Dominican Republic", "Honduras",
                                                             "Haiti"))] <- "North America"
happiness_2020$Continent[which(happiness_2020$Country %in% c("Chile", "Brazil", "Argentina", "Uruguay",
                                                             "Colombia", "Ecuador", "Bolivia", "Peru",
                                                             "Paraguay", "Venezuela"))] <- "South America"
happiness_2020$Continent[which(is.na(happiness_2020$Continent))] <- "Africa"
```

Code:

```
#Adding country ranks in the data set according to their scores
happiness_2020 <- happiness_2020 %>% mutate(Rank = row_number())
options(max.print=100000)
happiness_2020
```

Output:

```
> head(happiness_2020,5)
      Country  Score      GDP    Family   Health   Freedom  Generosity Corruption Dystopia.residual Rank
1     Finland 7.8087 10.63927 0.9543297 71.90083 0.9491722 -0.05948202  0.1954446          2.762835    1
2     Denmark 7.6456 10.77400 0.9559908 72.40250 0.9514443  0.06620178  0.1684895          2.432741    2
3 Switzerland 7.5599 10.97993 0.9428466 74.10245 0.9213367  0.10591104  0.3037284          2.350267    3
4     Iceland 7.5045 10.77256 0.9746696 73.00000 0.9488919  0.24694422  0.7117097          2.460688    4
5      Norway 7.4880 11.08780 0.9524866 73.20078 0.9557503  0.13453263  0.2632182          2.168266    5
>
```

Code:

```
#Correlation matrix
#To find out which factors correlate the most with the happiness quotient
library(corrgram)
library(corrplot)
library(rpart)
library(ggplot2)
library(ggthemes)

str(happiness_2020)
num.cols <- sapply(happiness_2020,is.numeric)
cor.data <- cor(happiness_2020[,num.cols])
print(cor.data)
```

Output:

```
                    Score        GDP       Family      Health     Freedom  Generosity  Corruption Dystopia.residual
Score             1.00000000  0.77537440  0.765000757  0.77031629  0.59059678  0.06904313 -0.41830509       0.480278943
GDP               0.77537440  1.00000000  0.781813583  0.84846862  0.41901865 -0.11839937 -0.33472908      -0.062063061
Family            0.76500076  0.78181358  1.000000000  0.74274409  0.47886318 -0.05678035 -0.21052960      -0.002800699
Health            0.77031629  0.84846862  0.742744090  1.00000000  0.44884619 -0.07185211 -0.35384121      -0.039947769
Freedom           0.59059678  0.41901865  0.478863180  0.44884619  1.00000000  0.25372112 -0.42014450       0.062571264
Generosity        0.06904313 -0.11839937 -0.056780354 -0.07185211  0.25372112  1.00000000 -0.27848023      -0.021784952
Corruption       -0.41830509 -0.33472908 -0.210529601 -0.35384121 -0.42014450 -0.27848023  1.00000000       0.017850913
Dystopia.residual  0.48027894 -0.06206306 -0.002800699 -0.03994777  0.06257126 -0.02178495  0.01785091       1.000000000
Rank              -0.98526254 -0.77730185 -0.749488309 -0.76974621 -0.58344125 -0.04492203  0.39435477      -0.468968922
                          Rank
Score             -0.98526254
GDP               -0.77730185
Family            -0.74948831
Health            -0.76974621
Freedom           -0.58344125
Generosity        -0.04492203
Corruption         0.39435477
Dystopia.residual -0.46896892
Rank               1.00000000
```

DATA CLEANING:

Since there is no null value present, there is no need of cleaning.

Code:

```
colSums(is.na(happiness_2020))
```

Output:

```
> colSums(is.na(happiness_2020))
         Country            Score              GDP           Family           Health          Freedom
               0                0                0                0                0                0
      Generosity       Corruption Dystopia.residual             Rank
               0                0                0                0
```

DATA VISUALIZATION:

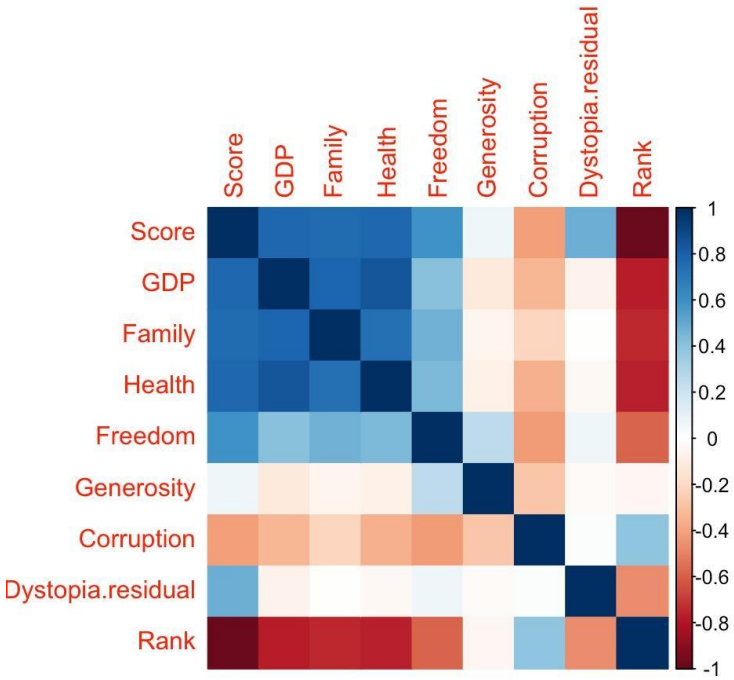
Heat map –

Code:

```
#graph 1
#correlation matrix
cor <- cor(happiness_2020)
str(happiness_2020)
colnames(happiness_2020)
head(happiness_2020)
summary(happiness_2020)
print(corrplot(cor.data, method = 'color'))

#Hence, happiness most strongly correlates with the attributes of GDP, Health, Social Support(Family) and Freedom.
```

Output:

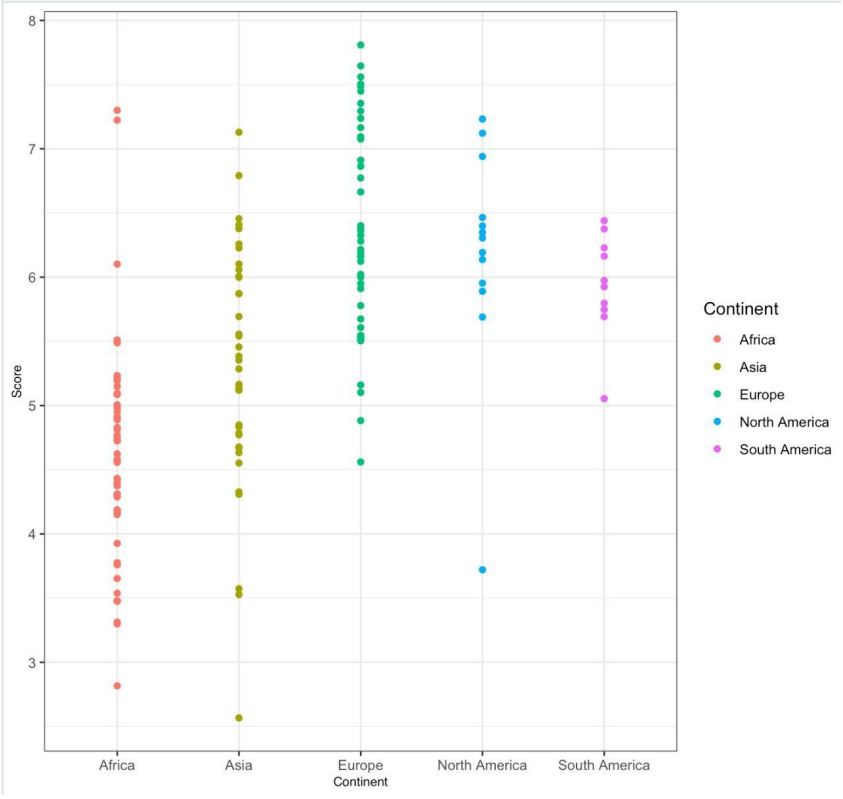


Scatterplot for Happiness scores Vs. continent-

Code:

```
#graph 2
#Happiness score distribution across different continents using a scatter plot
options(repr.plot.width=10, repr.plot.height=8)
gg1 <- ggplot(happiness_2020,
              aes(x=Continent,
                  y=Score,
                  color=Continent))+
  geom_point() + theme_bw() +
  theme(axis.title = element_text(family = "Helvetica", size = (8)))
gg1
```

Output:

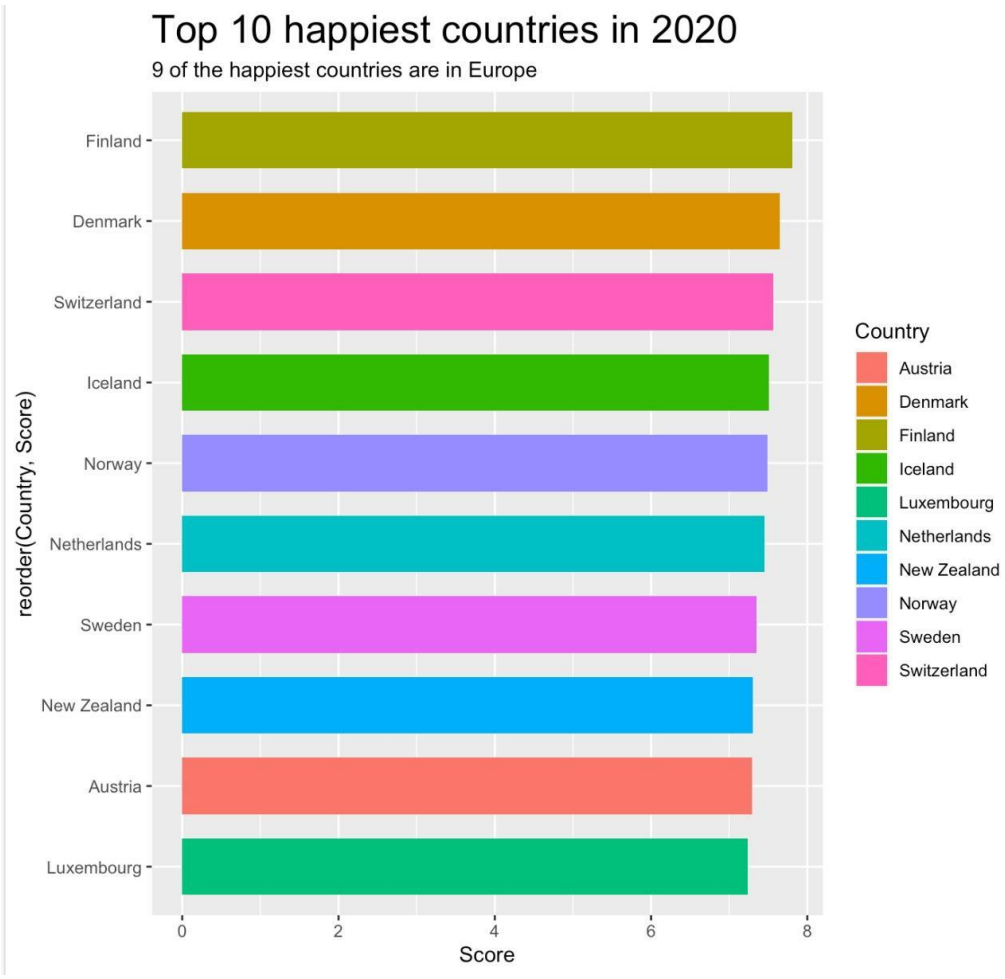


Bar graph for happiest countries-

Code:

```
#graph 3
#Top 10 countries (countries with highest happiness score)
gg2 <- ggplot(happiness_2020[1:10,], aes(x = reorder(Country, Score), y=Score, fill = Country)) +
  ggtitle("Top 10 happiest countries in 2020", subtitle = "9 of the happiest countries are in Europe") +
  geom_bar(stat="identity", width=0.7) + theme(plot.title = element_text(size=20)) + coord_flip()
gg2
```

Output:

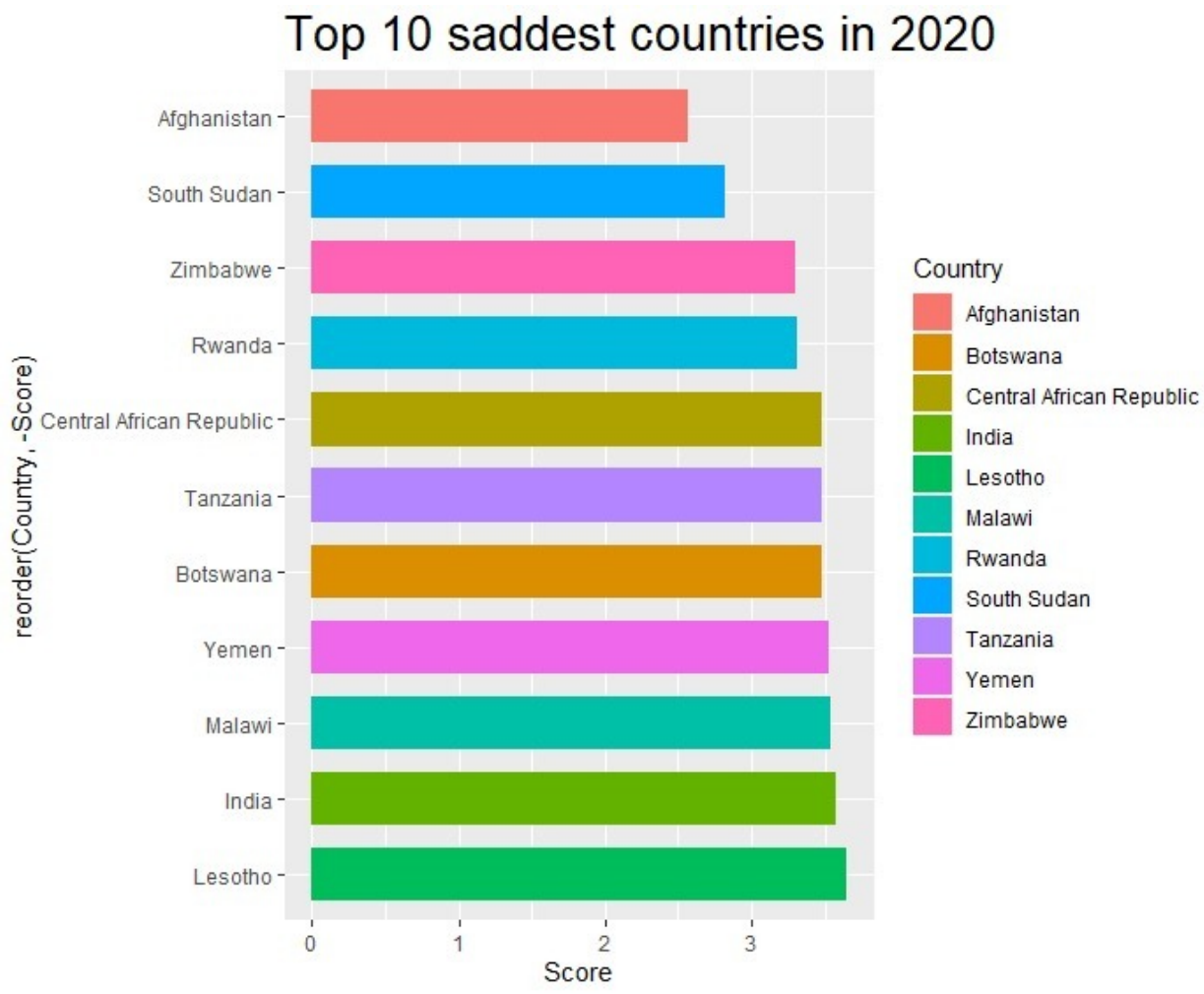


Bar graph for saddest countries-

Code:

```
#graph 4
#Bottom 10 countries (countries with least happiness score)
gg3 <- ggplot(happiness_2020[143:153,], aes(x = reorder(Country, -Score), y=Score, fill = Country)) +
  ggtitle("Top 10 saddest countries in 2020") + geom_bar(stat="identity", width=0.7) +
  theme(plot.title = element_text(size=20)) + coord_flip()
gg3
```

Output:

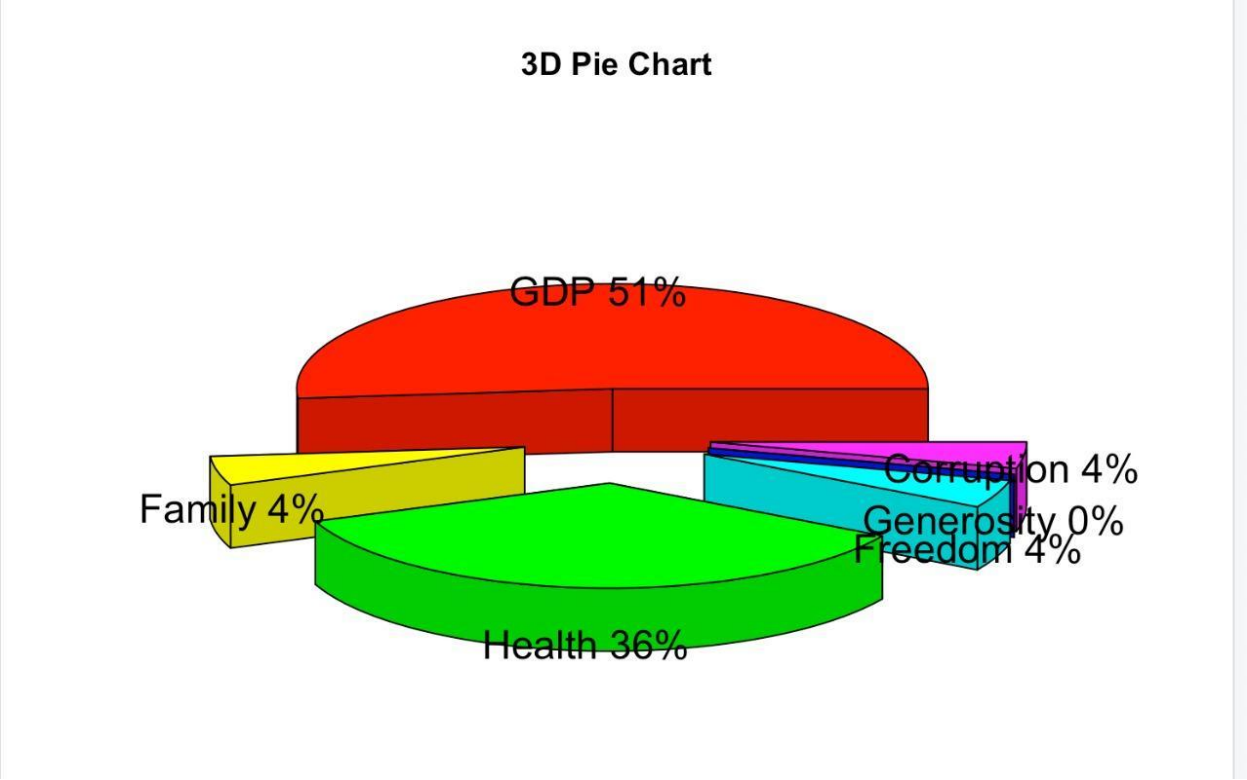


Pie chart-

Code:

```
library(plotrix)
slices<-c(9.295706,0.8087211,6.444553,0.7833602,0.00001,0.7331202)
pct<-round(slices/sum(slices)*100)
lbls<-paste(c("GDP","Family","Health","Freedom","Generosity","Corruption")," ",pct,"%",sep="")
pie3D(slices,labels=lbls,explode=0.3,
      main="3D Pie Chart")
```

Output:

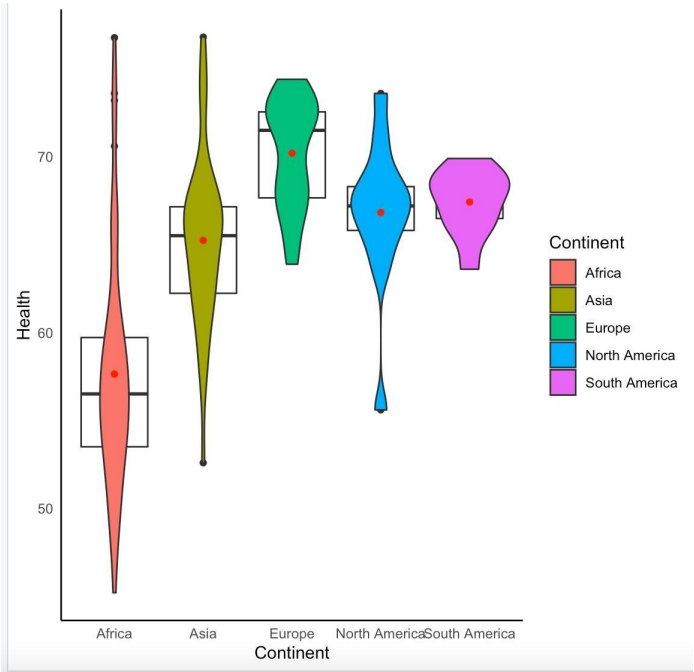


Violin plot for average health across region

Code:

```
#What is the average health across the different regions
gg4 <- ggplot(happiness_2020, aes(x=Continent, y = Health))+
  geom_boxplot()+
  geom_violin(aes(fill=Continent))+
  theme_minimal()+
  stat_summary(geom = 'point', fun = 'mean', color='red')+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))
```

Output:

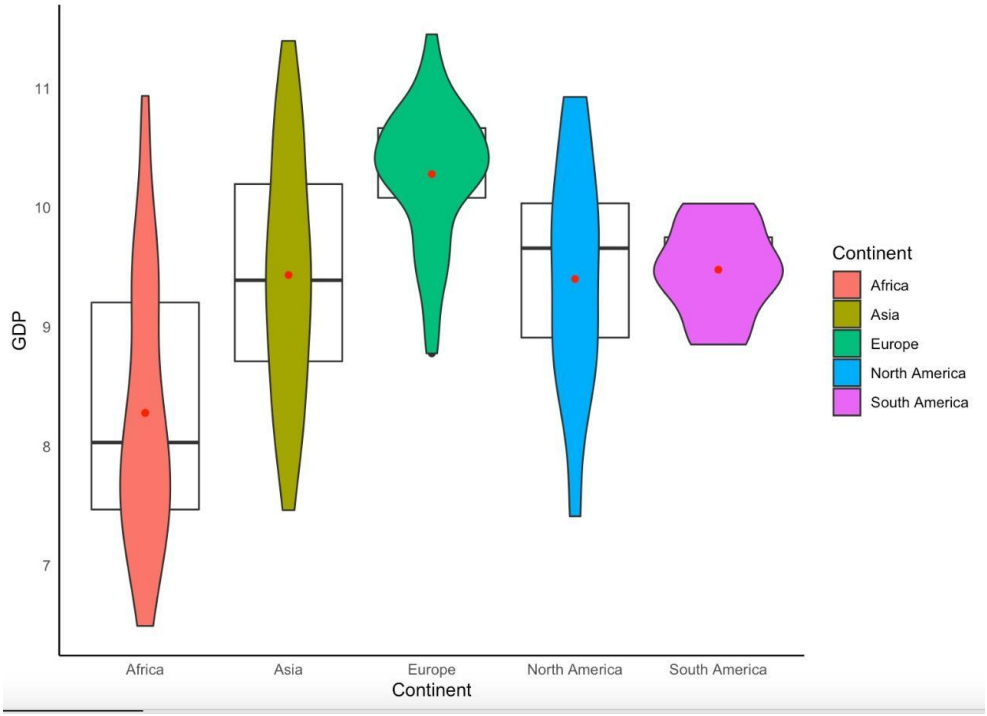


Violin plot for average GDP Vs. continent

Code:

```
#What is the average GDP across the different regions
gg5 <- ggplot(happiness_2020, aes(x=Continent, y = GDP))+
  geom_boxplot()+
  geom_violin(aes(fill=Continent))+
  theme_minimal()+
  stat_summary(geom = 'point', fun = 'mean', color='red')+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))
gg5
```

Output:

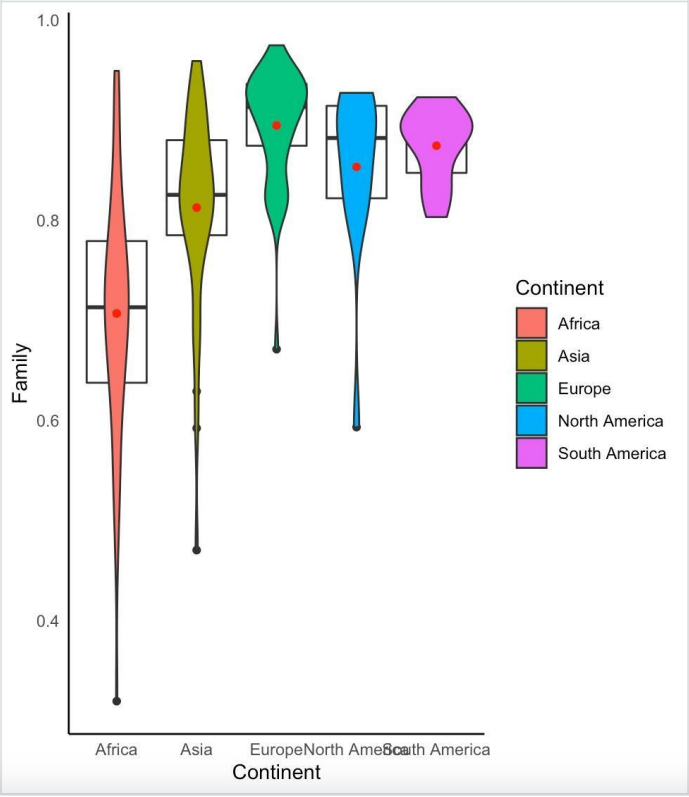


Violin plot for average social support Vs. continent

Code:

```
#What is the average social support(family) across the different regions
gg6 <- ggplot(happiness_2020, aes(x=Continent, y = Family))+
  geom_boxplot()+
  geom_violin(aes(fill=Continent))+
  theme_minimal()+
  stat_summary(geom = 'point', fun = 'mean', color='red')+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_line(colour = "black"))
gg6
```

Output:



ML ALGORITHM MODEL:

Code:

```
#linear regression
s=sample.split(happiness_2020,SplitRatio=0.7)
train=subset(happiness_2020,split=T)
test=subset(happiness_2020,split=F)

#predicting the value
MODEL1=lm(Score~GDP,data=happiness_2020) #creating a regression model
p=predict(MODEL,test)
a<-data.frame(GDP=10.99993)
result=predict(MODEL,a)
print(result)
RSE=sigma(MODEL)/mean(test$Score)
print(RSE)
acc<-sqrt(mean((test$Score-p)^2))
acc
summary(MODEL1)

MODEL2=lm(Score~Health,data=happiness_2020)
a1<-data.frame(Health=74.40250)
result2=predict(MODEL2,a1)
RSE2=sigma(MODEL2)/mean(test$Score)
print(RSE2)
p2=predict(MODEL2,test)
acc2<-sqrt(mean((test$Score-p2)^2))
acc2
print(result2)
summary(MODEL2)

MODEL3=lm(Score~Family,data=happiness_2020)
a2<-data.frame(Family=0.9559908)
result3=predict(MODEL3,a2)
RSE=sigma(MODEL3)/mean(test$Score)
print(RSE)
p3=predict(MODEL3,test)
acc3<-sqrt(mean((test$Score-p3)^2))
acc3
print(result3)
summary(MODEL3)
```

Output:

```
> RSE=sigma(MODEL)/mean(test$Score)
> print(RSE)
[1] 0.1287579
> RSE2=sigma(MODEL2)/mean(test$Score)
> print(RSE2)
[1] 0.1300139
> RSE=sigma(MODEL3)/mean(test$Score)
> print(RSE)
[1] 0.1313121
> range(happiness_2020$Score)
[1] 2.5669 7.8087
> result=predict(MODEL,a)
> result
       1
6.696427
> s=sample.split(happiness_2020,SplitRatio=0.7)
> train=subset(happiness_2020,split=T)
> test=subset(happiness_2020,split=F)
>
> #predicting the value
> MODEL1=lm(Score~GDP,data=happiness_2020) #creating a regression model
> p=predict(MODEL,test)
> a<-data.frame(GDP=10.99993)
> result=predict(MODEL,a)
> print(result)
       1
6.696427
> RSE=sigma(MODEL)/mean(test$Score)
> print(RSE)
[1] 0.1287579
> acc<-sqrt(mean((test$Score-p)^2))
> acc
[1] 0.7001014
```

```
> summary(MODEL1)

Call:
lm(formula = Score ~ GDP, data = happiness_2020)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29256 -0.52524  0.02843  0.57109  1.38802

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.19865    0.44586  -2.688  0.00799 **
GDP          0.71774    0.04757  15.088  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7047 on 151 degrees of freedom
Multiple R-squared:  0.6012,    Adjusted R-squared:  0.5986
F-statistic: 227.6 on 1 and 151 DF,  p-value: < 2.2e-16

>
> MODEL2=lm(Score~Health,data=happiness_2020)
> a1<-data.frame(Health=74.40250)
> result2=predict(MODEL2,a1)
> RSE2=sigma(MODEL2)/mean(test$Score)
> print(RSE2)
[1] 0.1300139
> p2=predict(MODEL2,test)
> acc2<-sqrt(mean((test$Score-p2)^2))
> acc2
[1] 0.7069308
> print(result2)
       1
6.681984
```

```
> summary(MODEL2)

Call:
lm(formula = Score ~ Health, data = happiness_2020)

Residuals:
     Min       1Q   Median       3Q      Max
-1.75466 -0.58222  0.09589  0.56470  1.57394

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.350239   0.530158  -4.433 1.78e-05 ***
Health       0.121397   0.008178  14.845  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7116 on 151 degrees of freedom
Multiple R-squared:  0.5934,    Adjusted R-squared:  0.5907
F-statistic: 220.4 on 1 and 151 DF,  p-value: < 2.2e-16

>
> MODEL3=lm(Score~Family,data=happiness_2020)
> a2<-data.frame(Family=0.9559908)
> result3=predict(MODEL3,a2)
> RSE=sigma(MODEL3)/mean(test$Score)
> print(RSE)
[1] 0.1313121
> p3=predict(MODEL3,test)
> acc3<-sqrt(mean((test$Score-p3)^2))
> acc3
[1] 0.7139899
> print(result3)
       1
6.504995
```

```
> summary(MODEL3)

Call:
lm(formula = Score ~ Family, data = happiness_2020)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01071 -0.38261 -0.04146  0.46455  2.12511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1926     0.3925  -0.491    0.624
Family        7.0059     0.4800  14.596   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7187 on 151 degrees of freedom
Multiple R-squared:  0.5852,    Adjusted R-squared:  0.5825
F-statistic: 213.1 on 1 and 151 DF,  p-value: < 2.2e-16

>
> range(happiness_2020$Score)
[1] 2.5669 7.8087
>
```

8. Result

Based on the analysis made, the countries in the European South American continent have a better happiness score.

The factors GDP, Health and Family plays a major role in contributing to the country happiness score.

Based on the linear regression model, the accuracy of our model is 70% hence our regression is good.