SCS2211

Laboratory II – Group Assignment

Group 21

18000061 - J. H.S. Abethunge
18000088 - U. J. Achinthya
18001181 - E. B. P. Perera
18001521 - C. D. Satharasinghe

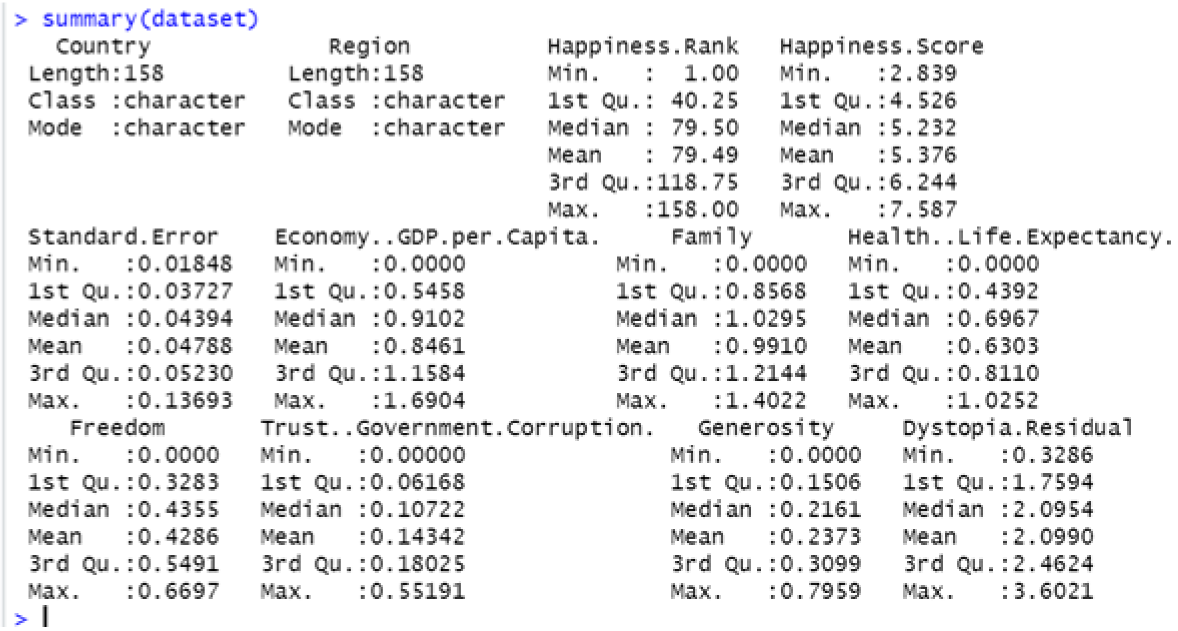
2021/03/27

Content

1. Plots and Observations

We considered the happiness score (of year 2015) depending on the economy, GDP per capita, family health , life expectancy and freedom.

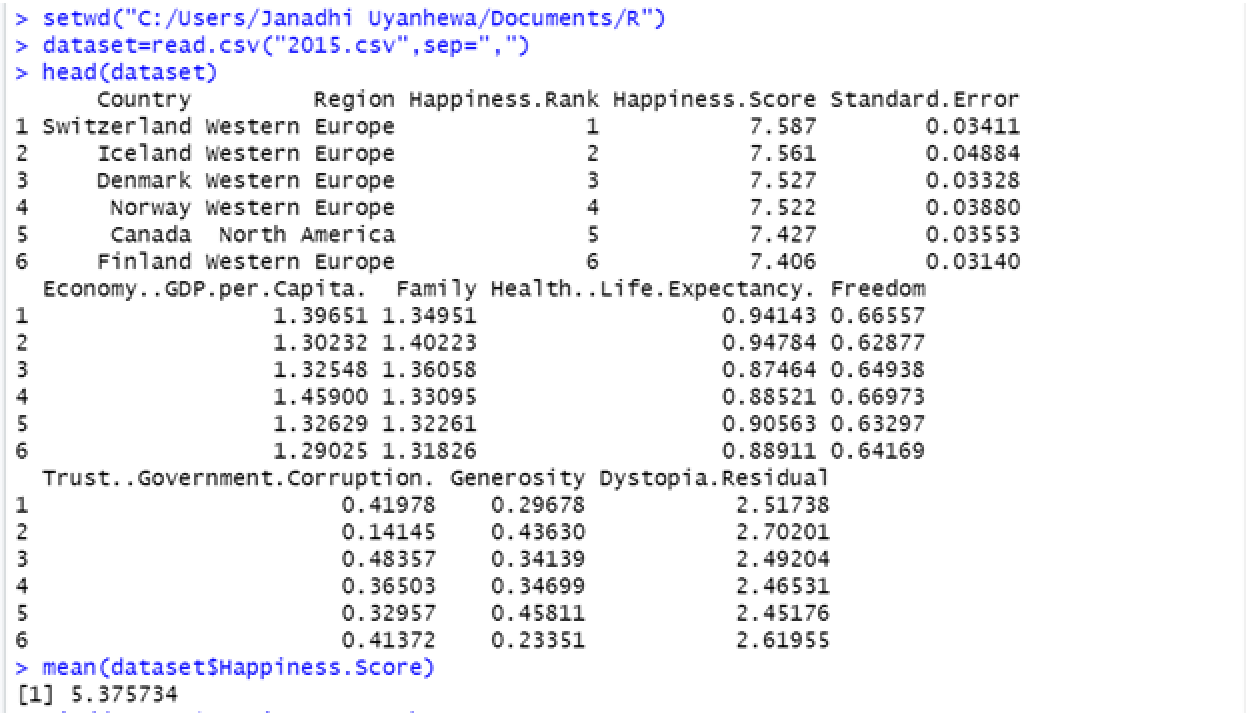
The summary of the variables are as follows,



```
> summary(dataset)
   Country              Region          Happiness.Rank   Happiness.Score
 Length:158         Length:158         Min.   :  1.00   Min.   :2.839  
 Class :character   Class :character   1st Qu.: 40.25   1st Qu.:4.526  
 Mode  :character   Mode  :character   Median : 79.50   Median :5.232  
                                       Mean   : 79.49   Mean   :5.376  
                                       3rd Qu.:118.75   3rd Qu.:6.244  
                                       Max.   :158.00   Max.   :7.587  
 Standard.Error    Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.
 Min.   :0.01848   Min.   :0.0000           Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.03727   1st Qu.:0.5458           1st Qu.:0.8568   1st Qu.:0.4392  
 Median :0.04394   Median :0.9102           Median :1.0295   Median :0.6967  
 Mean   :0.04788   Mean   :0.8461           Mean   :0.9910   Mean   :0.6303  
 3rd Qu.:0.05230   3rd Qu.:1.1584           3rd Qu.:1.2144   3rd Qu.:0.8110  
 Max.   :0.13693   Max.   :1.6904           Max.   :1.4022   Max.   :1.0252  
    Freedom       Trust..Government.Corruption.   Generosity     Dystopia.Residual
 Min.   :0.0000   Min.   :0.00000               Min.   :0.0000   Min.   :0.3286  
 1st Qu.:0.3283   1st Qu.:0.06168               1st Qu.:0.1506   1st Qu.:1.7594  
 Median :0.4355   Median :0.10722               Median :0.2161   Median :2.0954  
 Mean   :0.4286   Mean   :0.14342               Mean   :0.2373   Mean   :2.0990  
 3rd Qu.:0.5491   3rd Qu.:0.18025               3rd Qu.:0.3099   3rd Qu.:2.4624  
 Max.   :0.6697   Max.   :0.55191               Max.   :0.7959   Max.   :3.6021  
> |
```

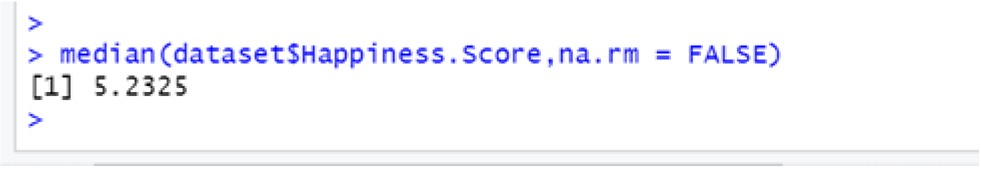
The happiness score

Analyzing the Happiness.score column, it can be seen that the statistical mean of it is 5.375734.



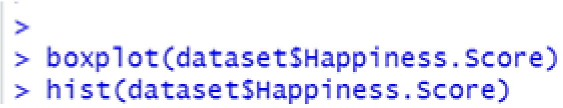
```
> setwd("C:/Users/Janadhi Uyanhewa/Documents/R")
> dataset=read.csv("2015.csv",sep=",")
> head(dataset)
      Country         Region Happiness.Rank Happiness.Score Standard.Error
1 Switzerland Western Europe              1           7.587        0.03411
2     Iceland Western Europe              2           7.561        0.04884
3     Denmark Western Europe              3           7.527        0.03328
4      Norway Western Europe              4           7.522        0.03880
5      Canada  North America              5           7.427        0.03553
6     Finland Western Europe              6           7.406        0.03140
  Economy..GDP.per.Capita.  Family Health..Life.Expectancy. Freedom
1                  1.39651 1.34951                  0.94143 0.66557
2                  1.30232 1.40223                  0.94784 0.62877
3                  1.32548 1.36058                  0.87464 0.64938
4                  1.45900 1.33095                  0.88521 0.66973
5                  1.32629 1.32261                  0.90563 0.63297
6                  1.29025 1.31826                  0.88911 0.64169
  Trust..Government.Corruption. Generosity Dystopia.Residual
1                       0.41978    0.29678           2.51738
2                       0.14145    0.43630           2.70201
3                       0.48357    0.34139           2.49204
4                       0.36503    0.34699           2.46531
5                       0.32957    0.45811           2.45176
6                       0.41372    0.23351           2.61955
> mean(dataset$Happiness.Score)
[1] 5.375734
```

The median of Happiness.score is 5.2325,

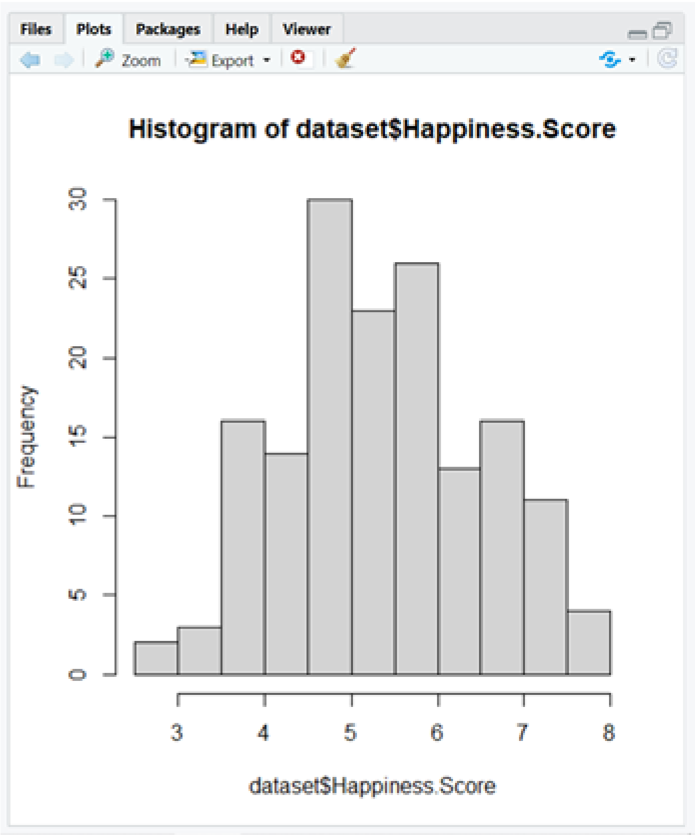


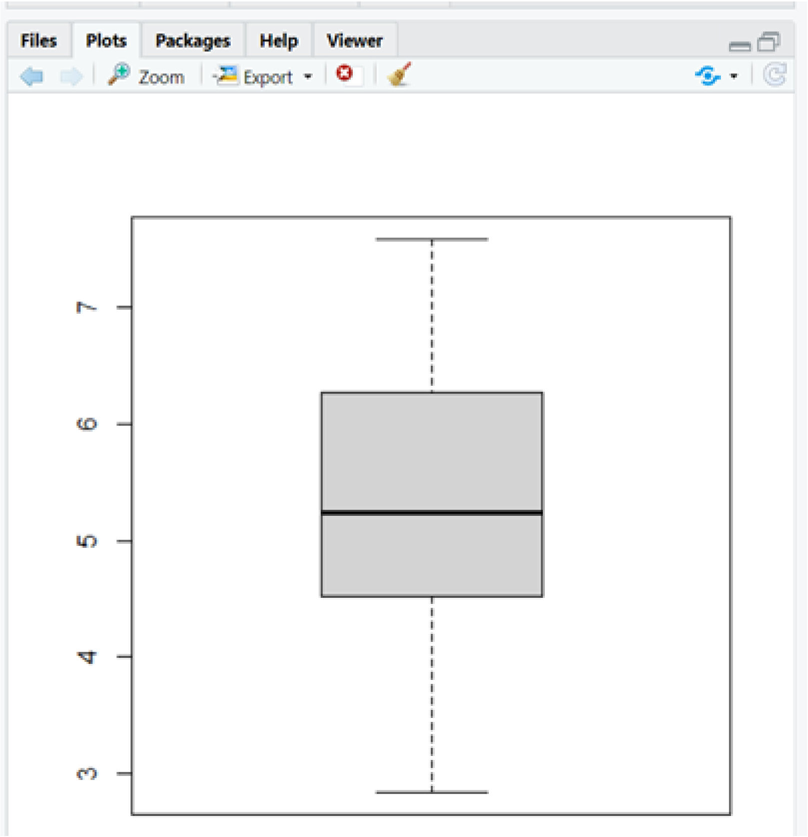
```
>
> median(dataset$Happiness.Score,na.rm = FALSE)
[1] 5.2325
>
```

Box plot and histogram are used to get an idea of how the data has been distributed.



```
>
> boxplot(dataset$Happiness.Score)
> hist(dataset$Happiness.Score)
```

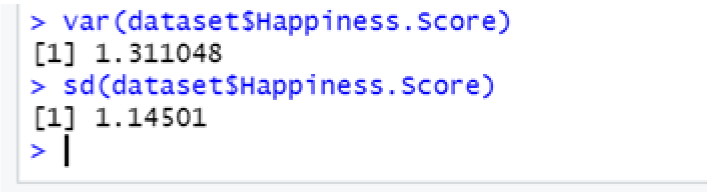




The histogram shows that the data set of Happiness.Score is positively skewed (this is also clear because the mean is greater than the median).

When we consider the box plot, it’s clear that the interquartile range lies between 4.5 and 6.1 (approximately) and the median is approximately 5.2.

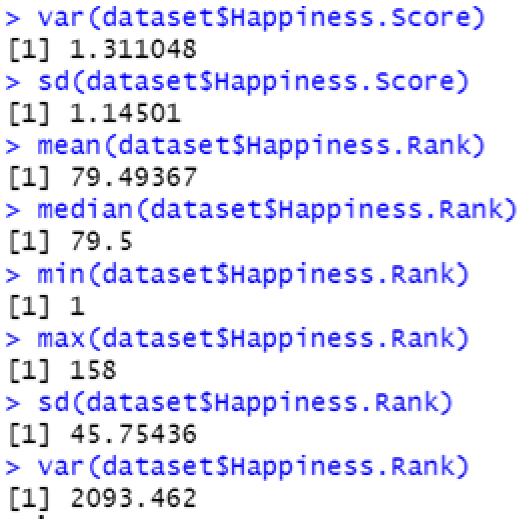
The variance and standard deviation are,



```
> var(dataset$Happiness.Score)
[1] 1.311048
> sd(dataset$Happiness.Score)
[1] 1.14501
> 
```

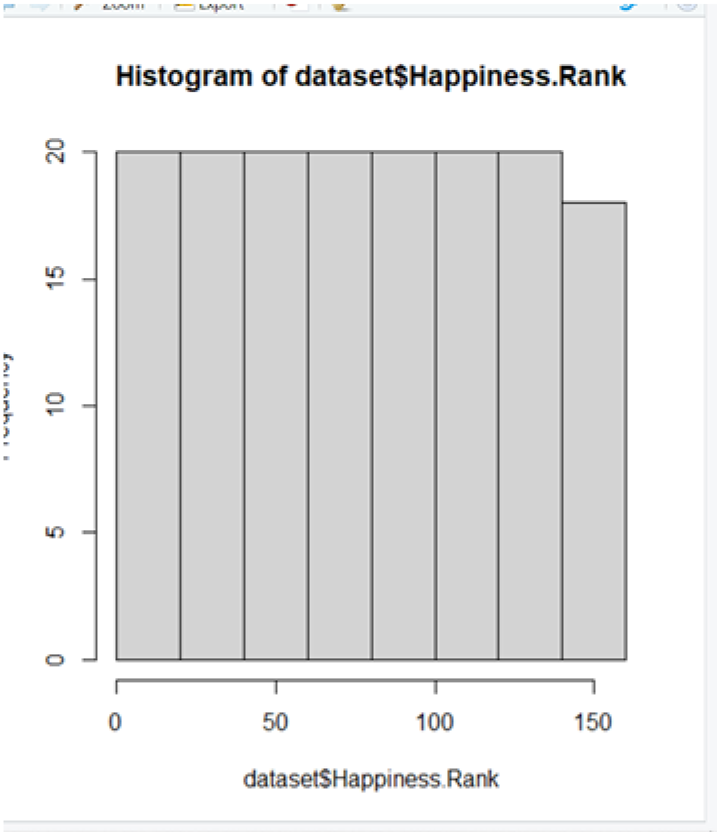
The happiness rank

Analyzing the Happiness.score column, it can be seen that the mean, median, mode, min, max, variance and the standard deviation of it are,



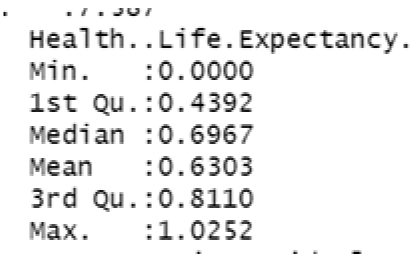
```
> var(dataset$Happiness.Score)
[1] 1.311048
> sd(dataset$Happiness.Score)
[1] 1.14501
> mean(dataset$Happiness.Rank)
[1] 79.49367
> median(dataset$Happiness.Rank)
[1] 79.5
> min(dataset$Happiness.Rank)
[1] 1
> max(dataset$Happiness.Rank)
[1] 158
> sd(dataset$Happiness.Rank)
[1] 45.75436
> var(dataset$Happiness.Rank)
[1] 2093.462
```

Using Histogram we can see that the data are symmetrically distributes (also it's clear because mean and median are almost equal) and the boxplot shows that the IQR lies between 55 and 100(approximately).



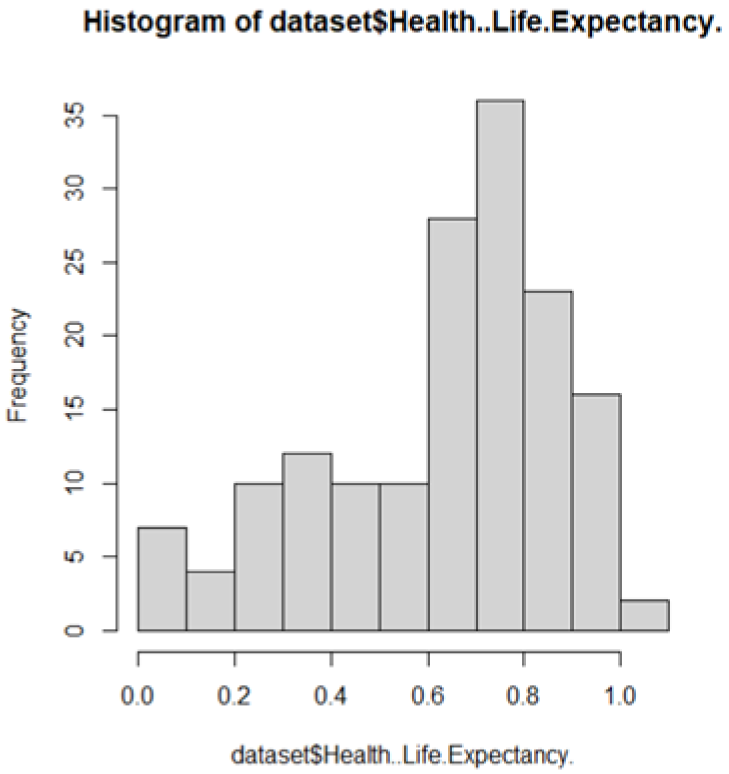
The Life expectancy

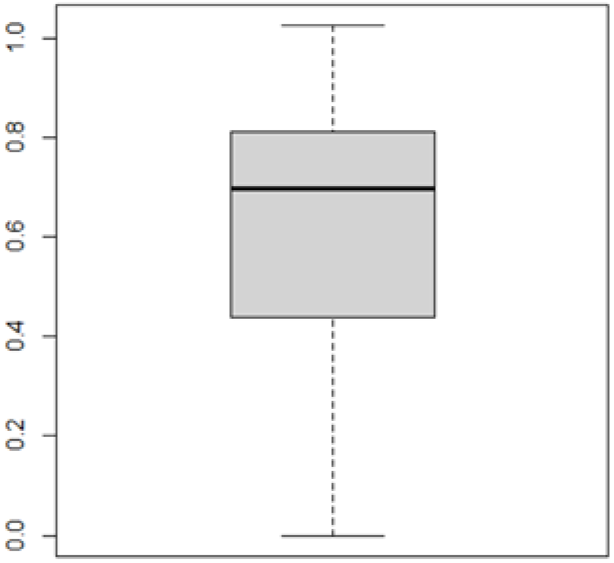
Analyzing the Life Expectancy column the following values are taken for mean, median, mode, min, max, variance and the standard deviation,



```
 Health..Life.Expectancy.
 Min.   :0.0000
 1st Qu.:0.4392
 Median :0.6967
 Mean   :0.6303
 3rd Qu.:0.8110
 Max.   :1.0252
```

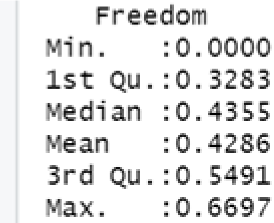
By observing the histogram and the boxplot, we can see that the dataset is negatively skewed and the IQR is between 0.6 and 0.8.





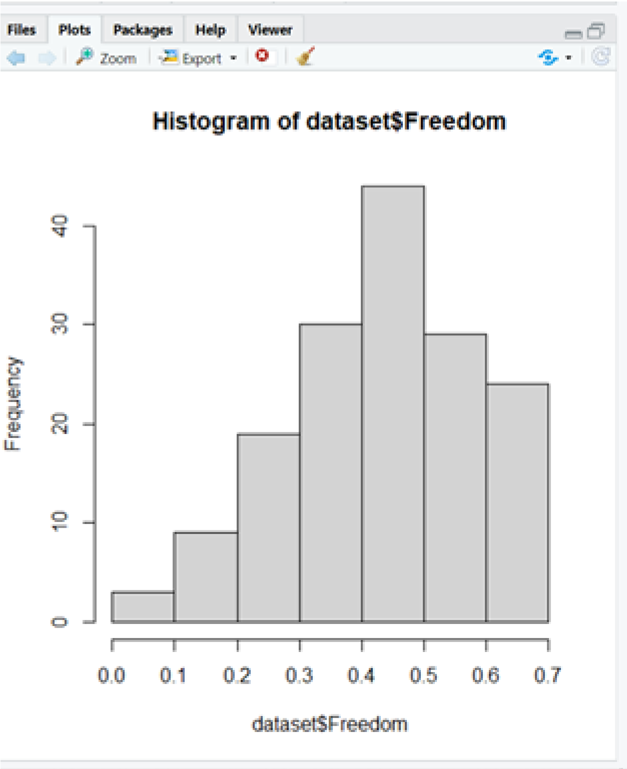
Freedom

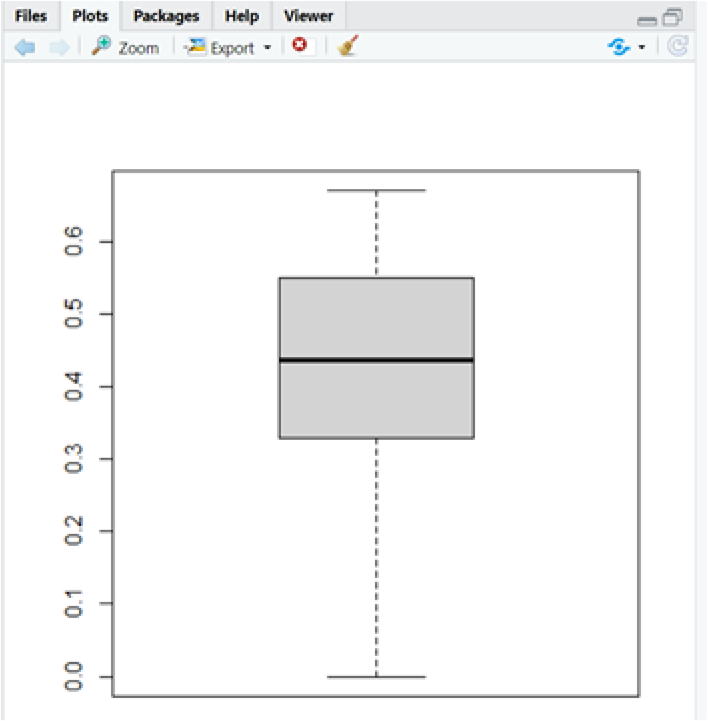
The mean, median, mode, min, max, variance and the standard deviation values associated with the variable Freedom are,



```
    Freedom
 Min.   :0.0000
 1st Qu.:0.3283
 Median :0.4355
 Mean   :0.4286
 3rd Qu.:0.5491
 Max.   :0.6697
```

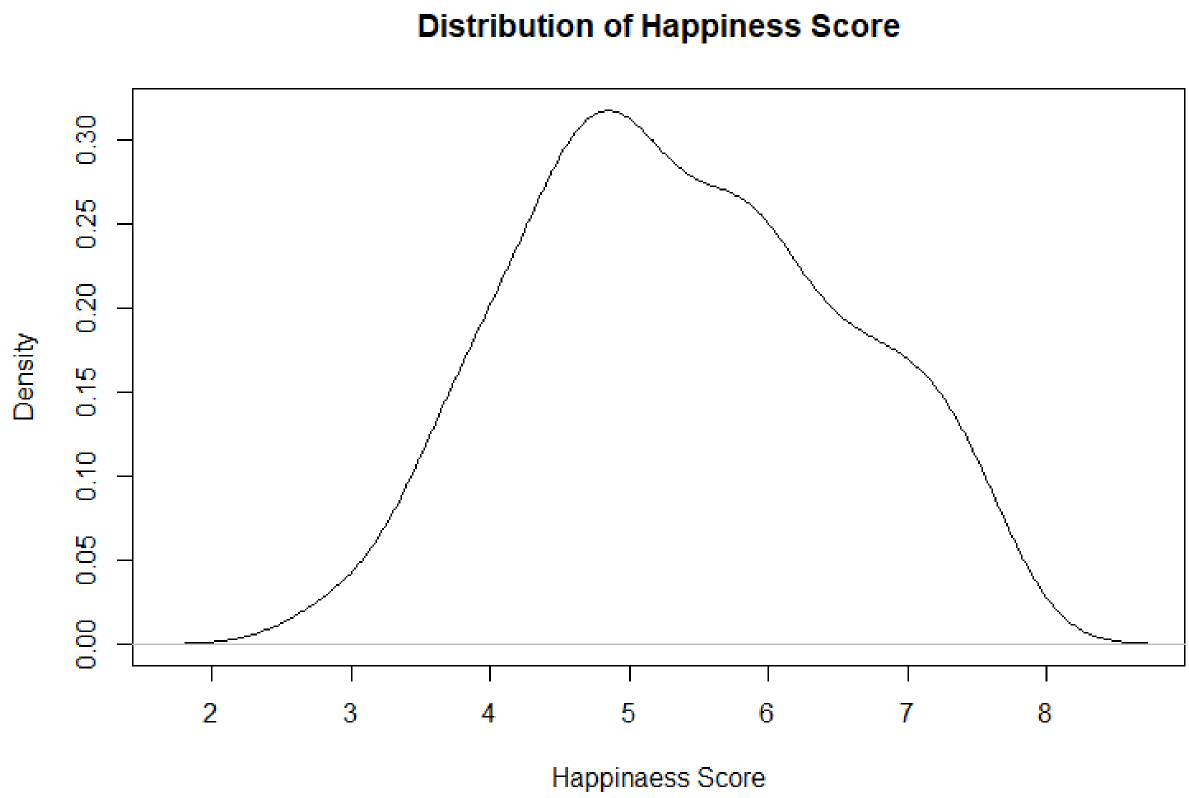
By observing the histogram and the boxplot, we can see that the dataset is negatively skewed and the IQR is between 0.4 and 0.5.



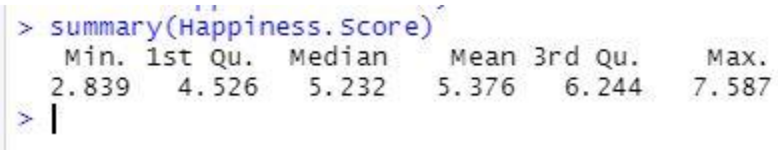


2. Distributions of Data

- ***Distribution of Happiness Score among Countries:***



This shows how the happiness score is distributed among countries.
Summary statistics:



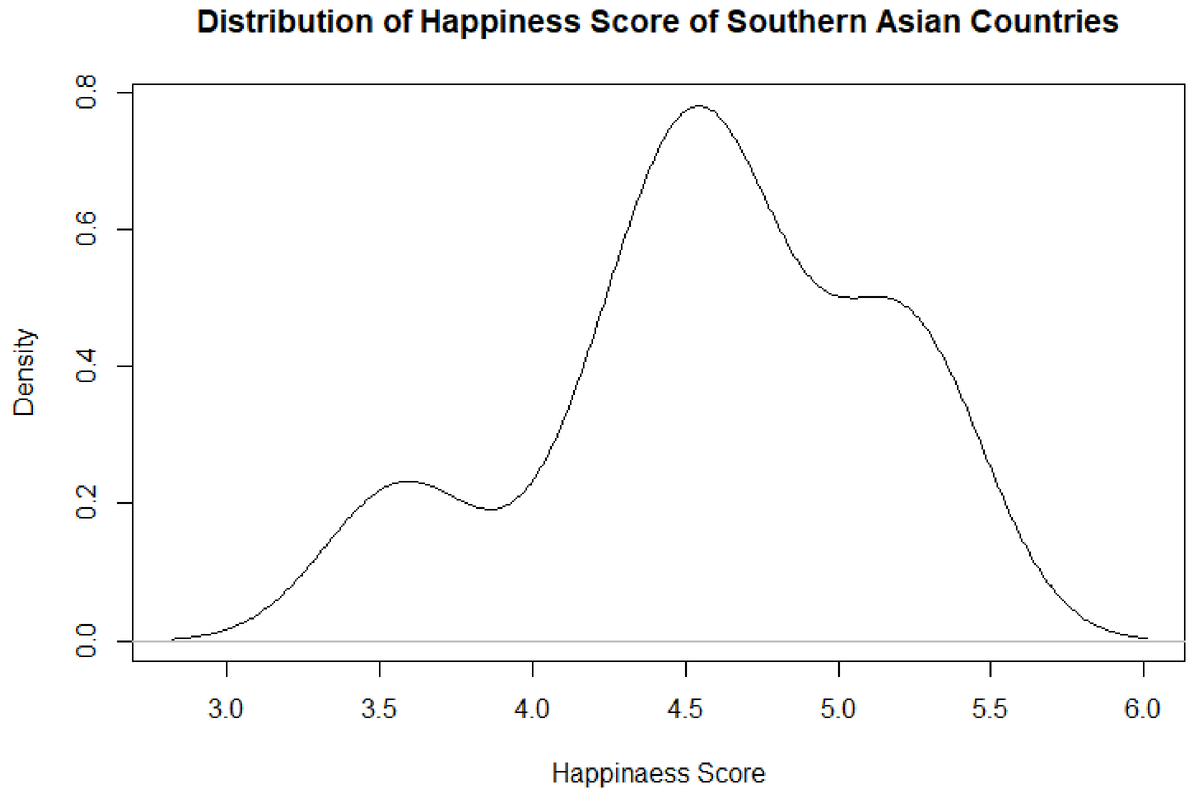
```
> summary(Happiness.Score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.839   4.526   5.232   5.376   6.244   7.587 
> |
```

Conclusion:

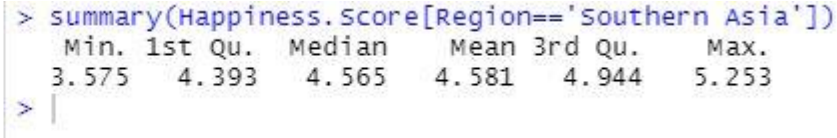
Happiness scores of countries have an average of 5.23.

Now, lets see how this happiness score varies for some regions.

1. Southern Asia Region



Summary Statistics:

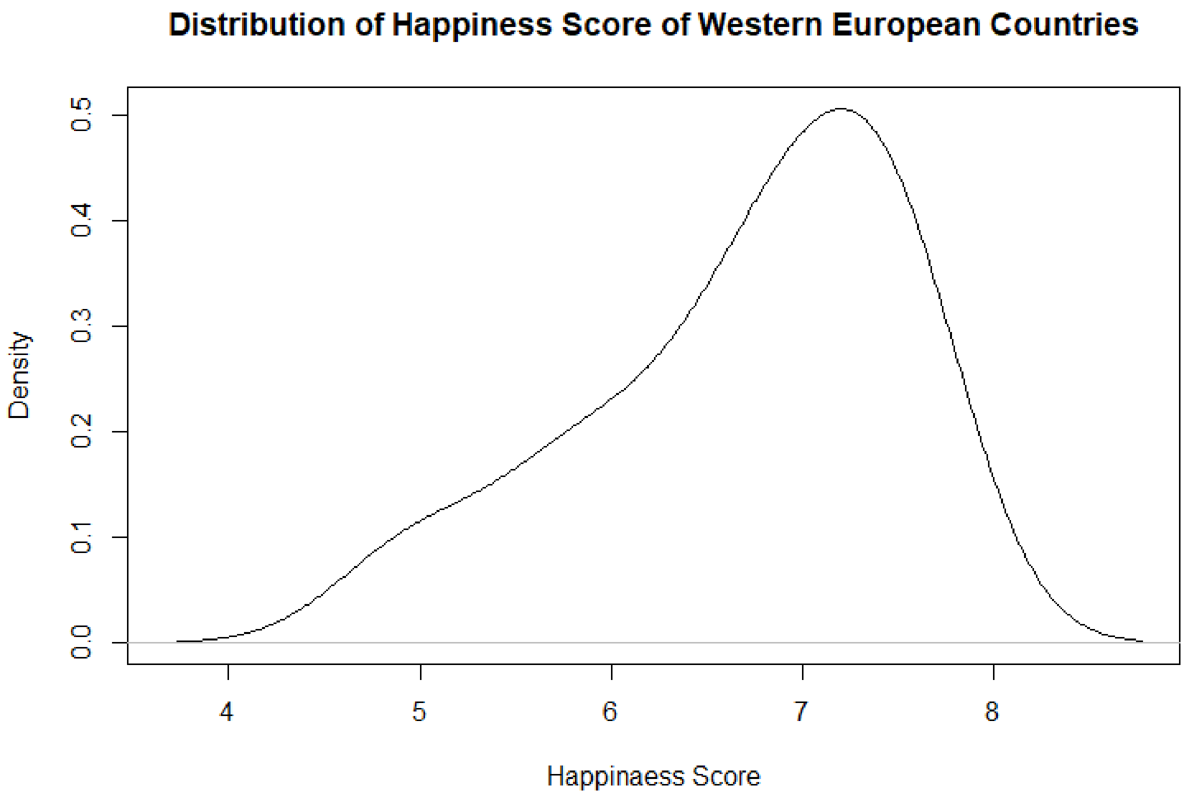


```
> summary(Happiness.Score[Region=='Southern Asia'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.575   4.393   4.565   4.581   4.944   5.253 
> 
```

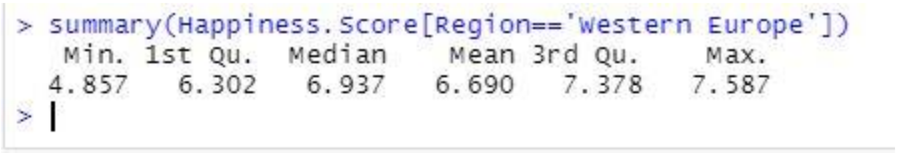
Conclusion:

Southern Asian countries have an average of 4.58 happiness score. (Comparatively low score.)

2. Western Europe Region



Summary Statistics:



```
> summary(Happiness.Score[Region=='Western Europe'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  4.857   6.302   6.937   6.690   7.378   7.587 
> |
```

Conclusion:

Western European countries have an average of 6.69 happiness score which is comparatively a high score.

3. Testing Hypothesis

Researchers believed that the average world happiness score is less than 5.4 from the surveys they have done from 2015-2018. But, from the collected sample dataset in 2019, they have claimed that the average happiness score is 5.407096. So, now they want to check whether we can believe the new hypothesis, that the average world happiness score is greater than or equal to 5.4.

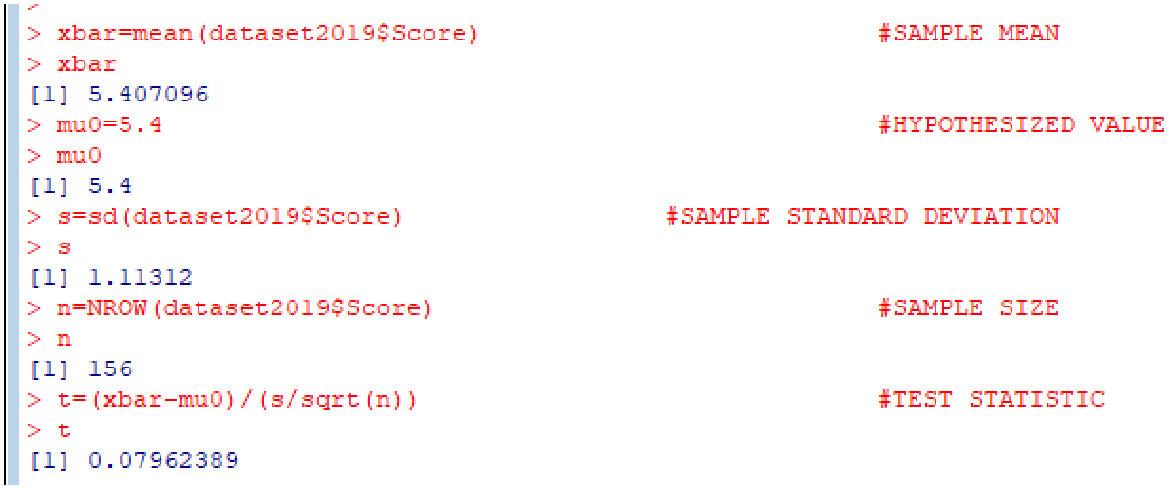
Null Hypothesis

- Average World Happiness score is less than 5.4.

Alternative Hypothesis

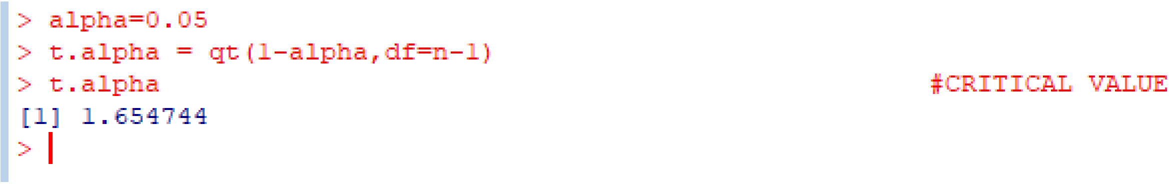
- Average World Happiness score is greater than or equal to 5.4.

As we are not given the population standard deviation, we decided to do the t-test for the sample dataset collected in 2019.

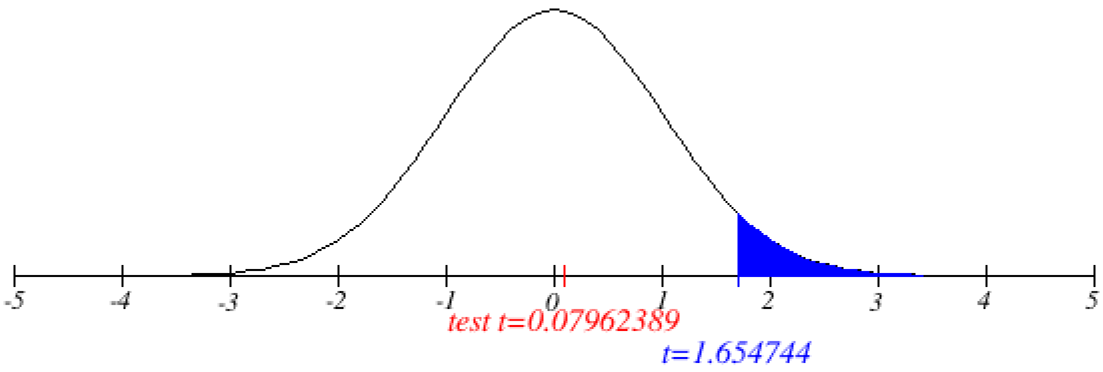


```
> xbar=mean(dataset2019$Score)                              #SAMPLE MEAN
> xbar
[1] 5.407096
> mu0=5.4                                                   #HYPOTHESIZED VALUE
> mu0
[1] 5.4
> s=sd(dataset2019$Score)                      #SAMPLE STANDARD DEVIATION
> s
[1] 1.11312
> n=NROW(dataset2019$Score)                                 #SAMPLE SIZE
> n
[1] 156
> t=(xbar-mu0)/(s/sqrt(n))                                  #TEST STATISTIC
> t
[1] 0.07962389
```

We compute the critical value at .05 significance level.



```
> alpha=0.05
> t.alpha = qt(1-alpha,df=n-1)
> t.alpha                                                      #CRITICAL VALUE
[1] 1.654744
> 
```

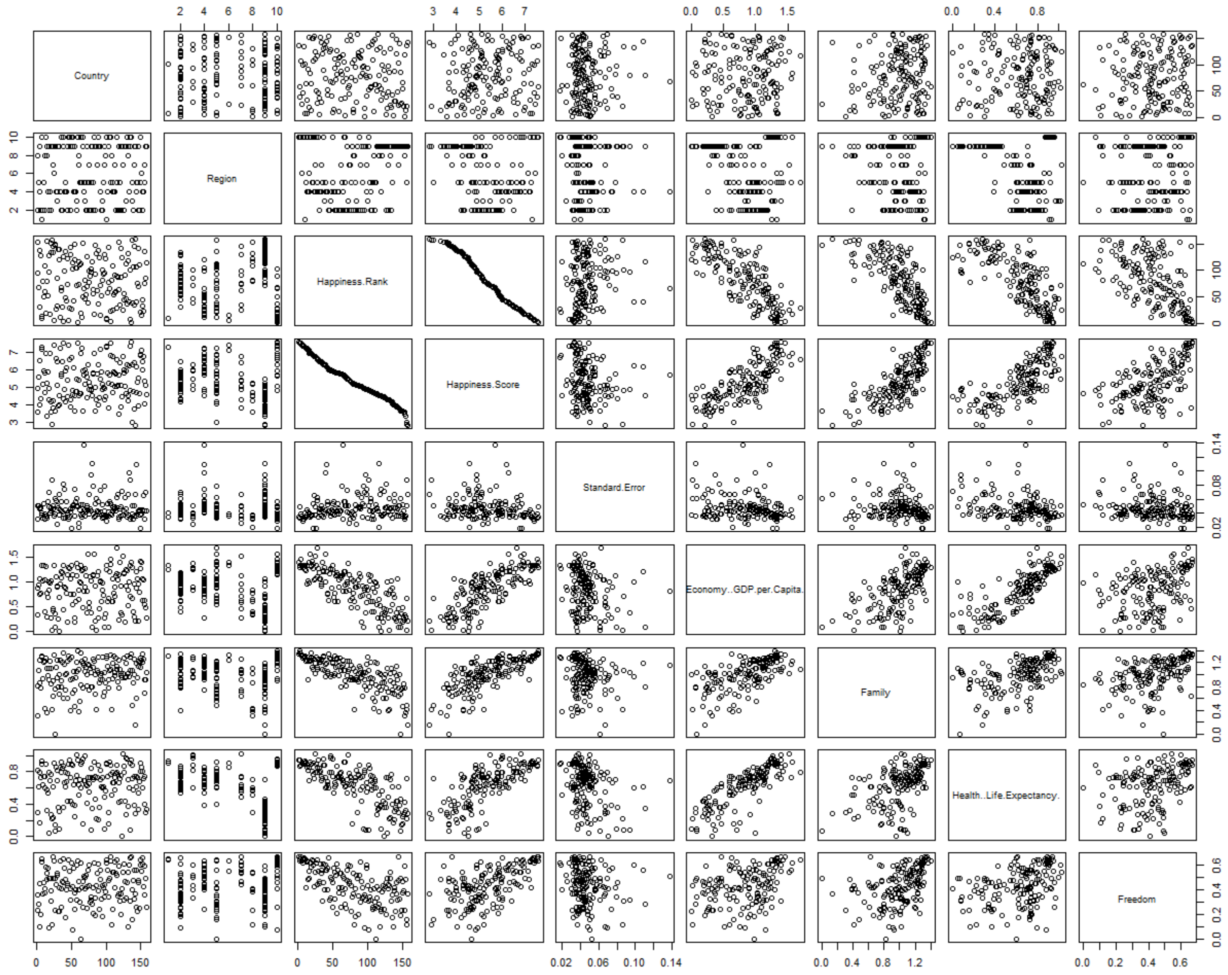


From the t-test the test statistic does not exceed the critical value. Hence we cannot reject the null hypothesis.

4. Multivariate Data

Plotting multivariate data elements:

- Country, Region, Happiness Rank, Happiness Score, Standard Error, Economy GDP, Family, Health Life Expectancy, and Freedom.



Code:

```
> dataset <- read.csv("/2015.csv")
> plot(dataset[1:9])
```

(Data Sample: Word Happiness Report – 2015 data sample.)

5. Strong Relationship Analysis

Description

By observing the multivariate plot obtained in part 4, we can see there are two responsive variables and four explanatory variables.

Responsive Variables:

- Happiness Rank
- Happiness Score

Explanatory Variables:

- Economy GDP per Capita
- Family Success Rate
- Health Life Expectancy
- Freedom

From those explanatory variables two explanatory variables are significantly show a strong, positive and linear relationship against 'Happiness Score' response variable.

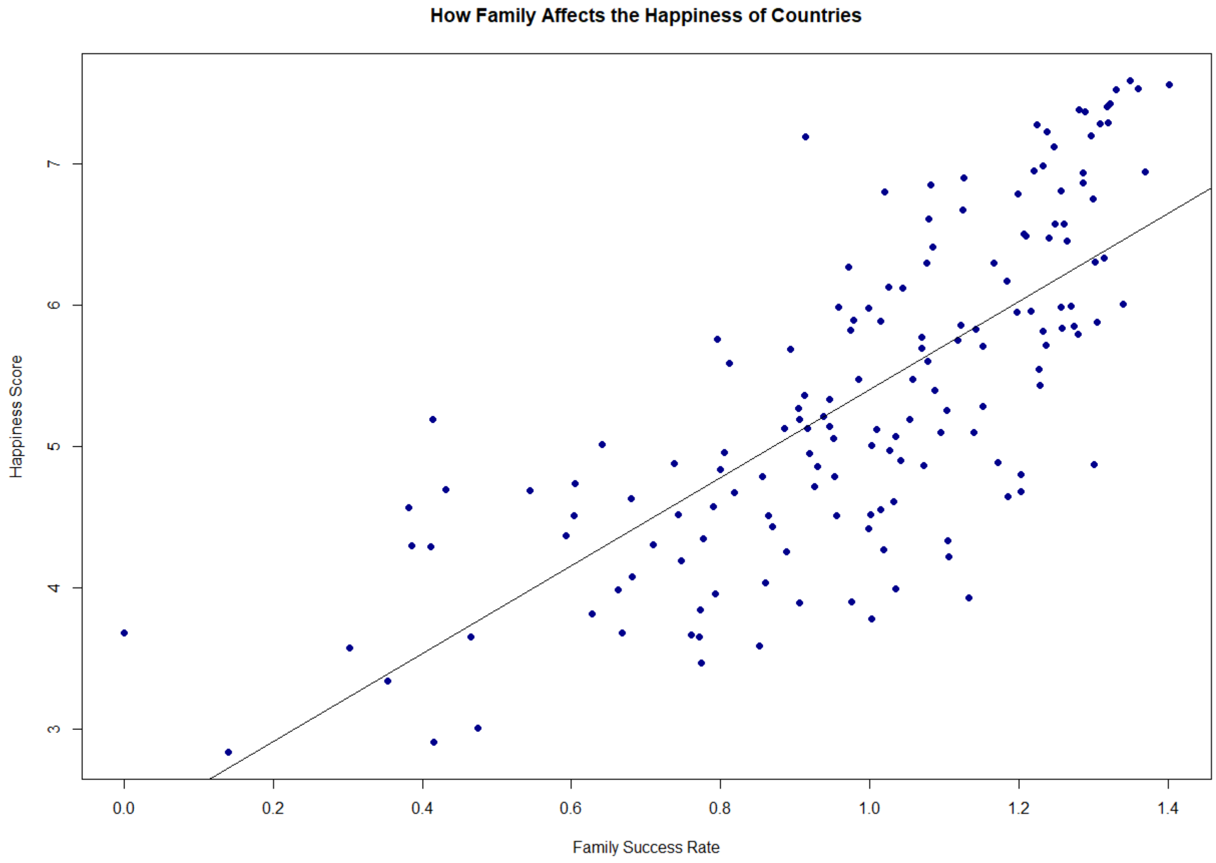
Two variables which depicts strongest relationship in multivariate data plot,

- **'Happiness Score' and 'Family Success Rate'.**
- **'Happiness Score' and 'Economy - GDP per Capita'.**

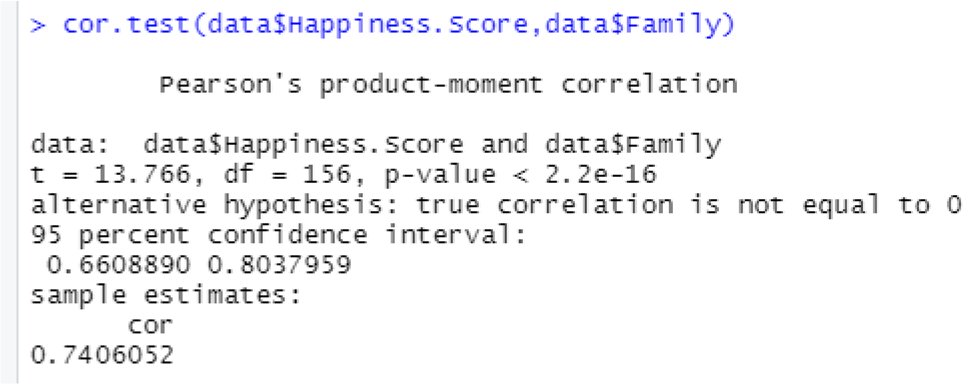
The analysis of above mentioned responsive and explanatory variables are stated below.

Happiness Score and Family Success Rate

Least Square Regression Line



Correlation



```
> cor.test(data$Happiness.Score,data$Family)

        Pearson's product-moment correlation

data:  data$Happiness.Score and data$Family
t = 13.766, df = 156, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6608890 0.8037959
sample estimates:
      cor
0.7406052
```

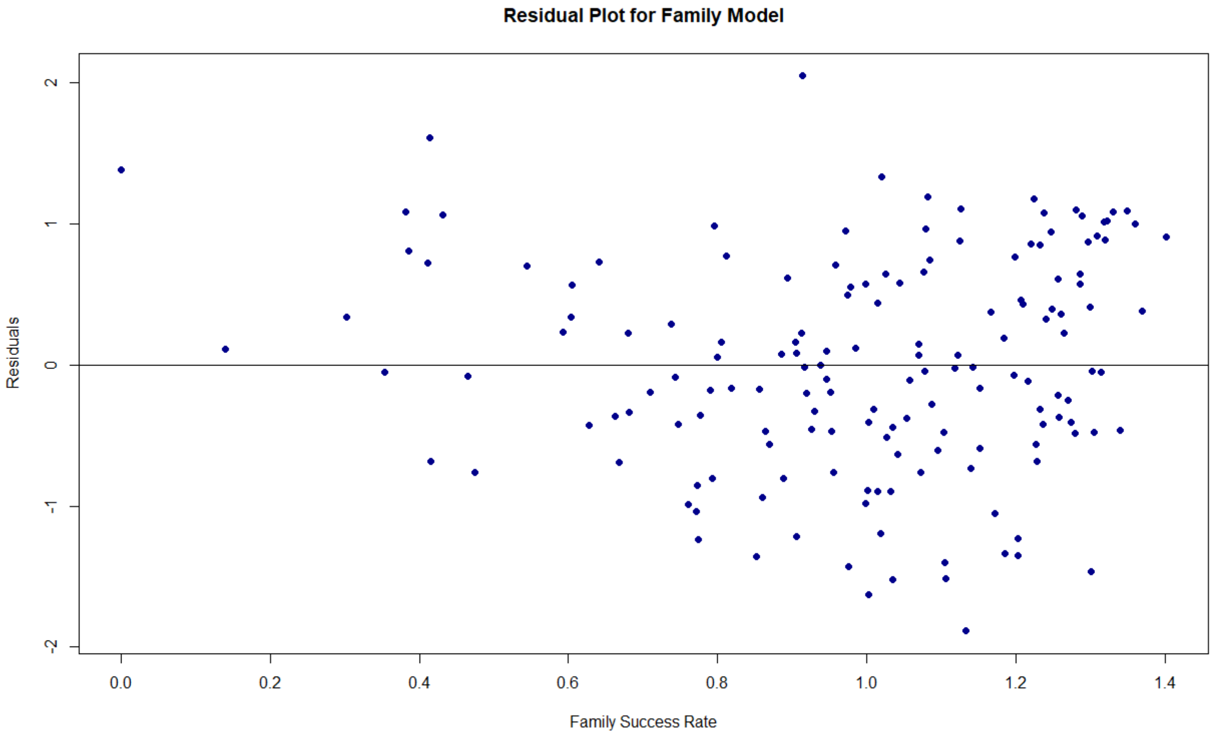
Code:

```
          > plot(Family, Happiness.Score, main="How Family Affects the Happiness of Countries"
, xlab="Family Success Rate" , ylab="Happiness Score",col="darkblue" ,pch=16)
          > familymodel <- lm(Happiness.Score~Family)
          > abline(familymodel)
          > plot(Family, family.res, main="Residual Plot for Family Model" , xlab="Family Success
Rate" , ylab="Residuals",col="darkblue" ,pch=16)
          > abline(0,0)
```

Conclution:

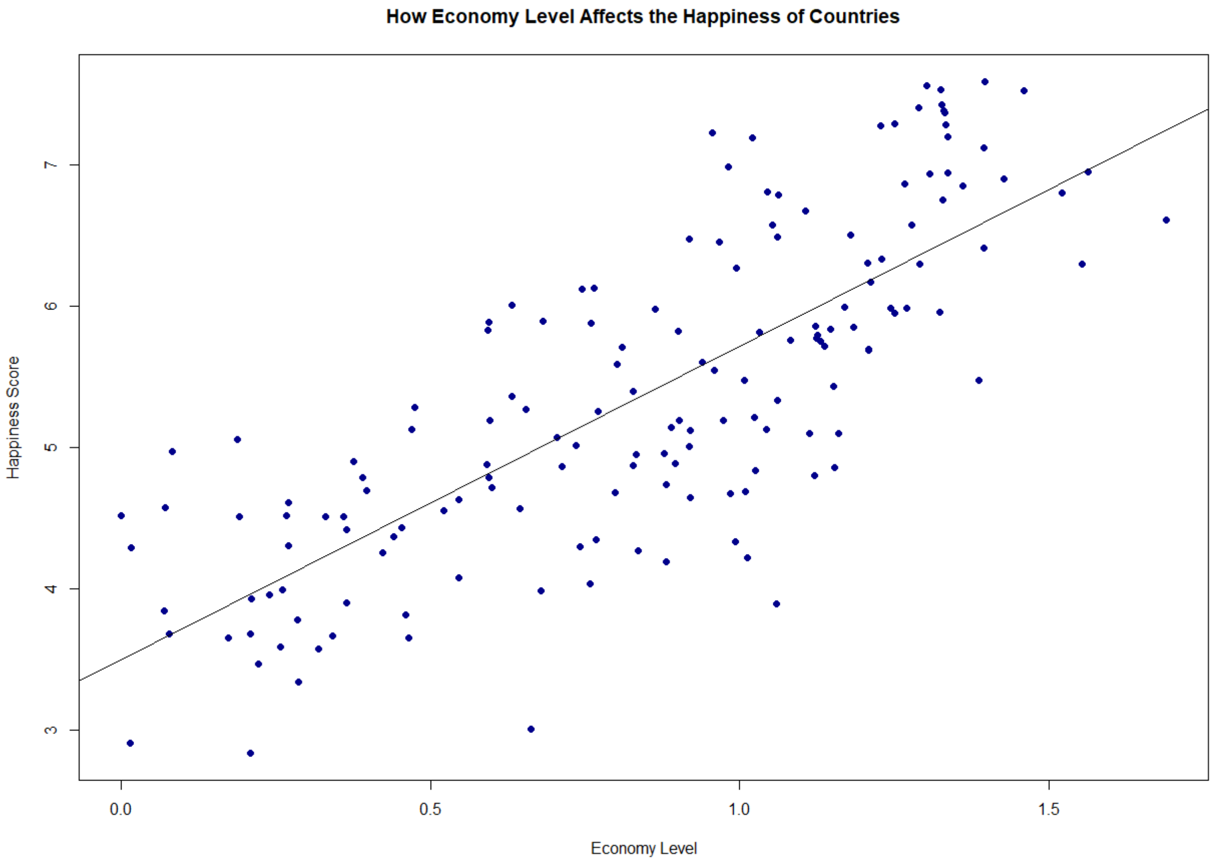
As correlation value is greater than 0.7 the relationship is stronger than moderate level. Family Success Rate has highly moderate relationship against Happiness Score.

Residual Plot

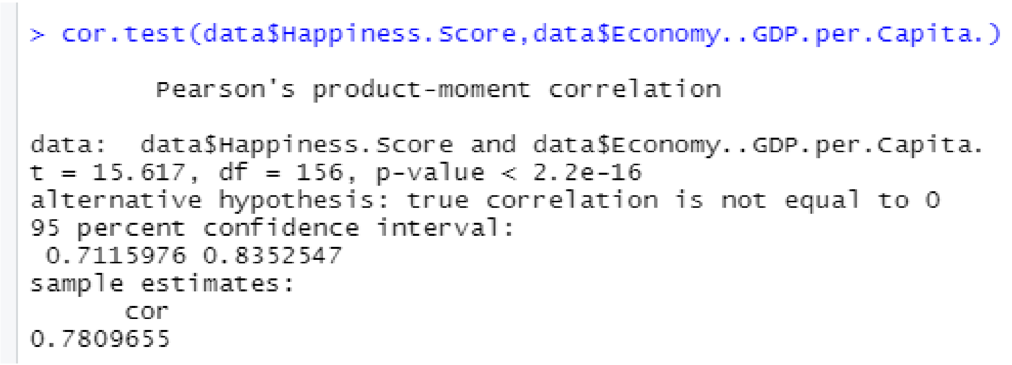


Happiness Score and Economy - GDP per Capita.

Least Square Regression Line



Correlation



```
> cor.test(data$Happiness.Score,data$Economy..GDP.per.Capita.)

        Pearson's product-moment correlation

data:  data$Happiness.Score and data$Economy..GDP.per.Capita.
t = 15.617, df = 156, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7115976 0.8352547
sample estimates:
      cor
0.7809655
```

Code:

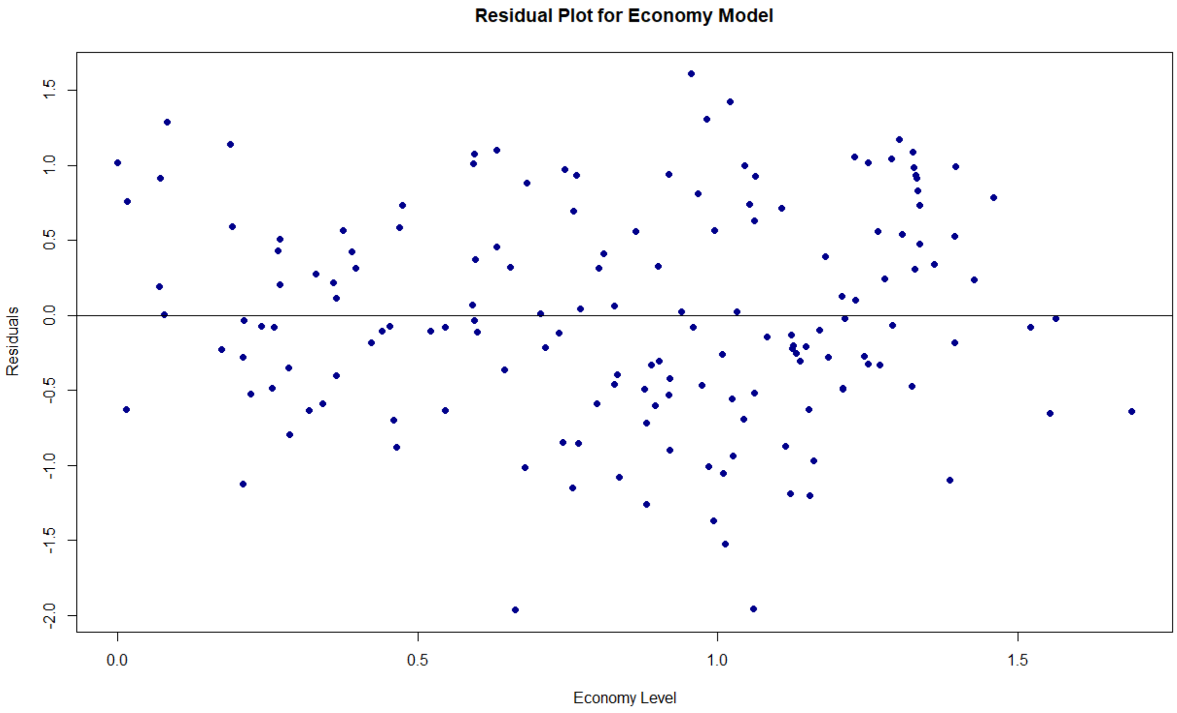
```
> plot(Economy..GDP.per.Capita., Happiness.Score, main="How Economy Level Affects the Happiness of Countries" , xlab="Economy Level" , ylab="Happiness Score",col="darkblue" ,pch=16)
> economymodel <- lm(Happiness.Score~Family)
> abline(economymodel)
>  plot(Economy..GDP.per.Capita., economy.res, main="Residual Plot for Economy Model" , xlab="Economy Level" , ylab="Residuals",col="darkblue" ,pch=16)
```

Conclution:

As correlation value is greater than 0.7 the relationship is stronger than moderate level. Economy has highly moderate relationship against Happiness Score.

```
> abline(0,0)
```

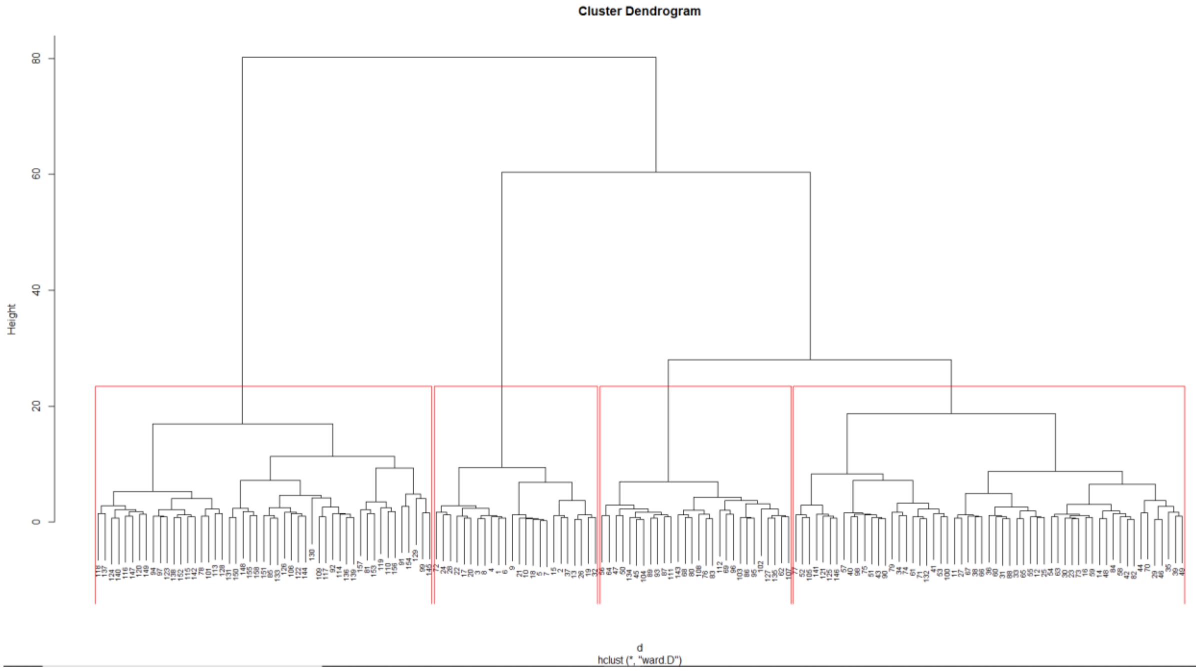
Residual Plot



6. Clustering Dataset

This clustering is executed to explanatory variables of 2015 data sample.

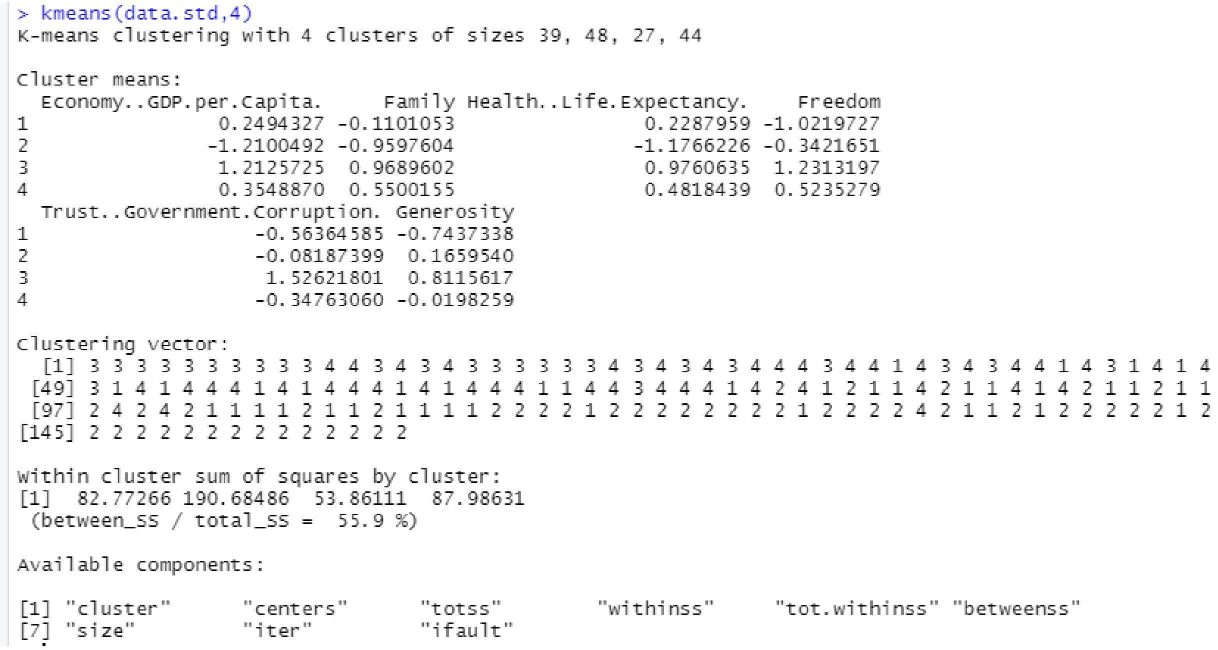
- Hierarchical clustering



Code:

```
> data.std=scale(data[6:11])
> d<-dist(data.std,method="euclidean")
> rect.hclust(hclust(d, method = "ward.D"),k=4,border = "red")
```

- Non - Hierarchical clustering



```
> kmeans(data.std,4)
K-means clustering with 4 clusters of sizes 39, 48, 27, 44

Cluster means:
  Economy..GDP.per.Capita.     Family Health..Life.Expectancy.    Freedom
1                0.2494327 -0.1101053                0.2287959 -1.0219727
2               -1.2100492 -0.9597604               -1.1766226 -0.3421651
3                1.2125725  0.9689602                0.9760635  1.2313197
4                0.3548870  0.5500155                0.4818439  0.5235279
  Trust..Government.Corruption. Generosity
1                   -0.56364585 -0.7437338
2                   -0.08187399  0.1659540
3                    1.52621801  0.8115617
4                   -0.34763060 -0.0198259

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 4 4 3 4 3 4 3 3 3 3 3 3 4 3 4 3 4 3 4 4 4 3 4 4 1 4 3 4 3 4 4 1 4 3 1 4 1 4
 [49] 3 1 4 1 4 4 4 1 4 1 4 4 4 1 4 1 4 4 4 1 1 4 4 3 4 4 4 1 4 2 4 1 2 1 1 4 2 1 1 4 1 4 2 1 1 2 1 1
 [97] 2 4 2 4 2 1 1 1 1 2 1 1 2 1 1 1 1 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2 2 4 2 1 1 2 1 2 2 2 2 2 1 2
[145] 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]  82.77266 190.68486  53.86111  87.98631
 (between_SS / total_SS =  55.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Code:

> kmeans(data.std,4)

Methods used:

- The explanatory variables Economy-GDP per capita, Family relationship, Health-Life expectancy, Freedom, Trust-Government Corruption and Generosity are used to cluster the dataset and obtain natural groups.
- The above data set is clustered under hierarchical and non-hierarchical clustering.
- We use 'hclust' to cluster under hierarchical clustering which is agglomerative hierarchical clustering.
- 'K-means' method is used to cluster under non-hierarchical clustering.

Observation:

- We obtained 4 groups separated under hierarchical clustering (Separated by red rectangles as shown in figure ****).
- In k-means clustering, we separated the dataset into 4 clusters and obtained 4 groups 39, 48, 27, and 44 in size.

Conclusion:

- Above dataset has a natural grouping as all the explanatory variables can be divided into clusters.

7. Team Details

Group : 21

Member	Contribution
18000061 - J. H.S. Abethunge	• Distribution of data • Multivariate Plot • Strong Relationship Analysis – regression line and residual plot
18000088 – U. J. Achinthya	• Observation and Plots
18001181 - E. B. P. Perera	• Testing Hypothesis
18001521 - C. D. Satharasinghe	• Clustering Dataset • Strong Relationship Analysis – Description and Correlation

Data Set : World Happiness Report
Link: https://www.kaggle.com/unsdsn/world-happiness