Assignment Cover Page

School / Department	School of Mathematical Sciences (SMS)
Course	B Sc (Hons) in Actuarial Studies (BAS) B Sc (Hons) in Industrial Statistics (BinDs)
Subject Code and Name	MST3024 Multivariate Analysis
Lecturer	Dr. Jane Teh Kimm Lii
Assignment Due Date	19th July 2022

STUDENT'S DECLARATION

1. I/ We hereby declare that this assignment is based on my/ our work except where acknowledgement of sources is made.

2. I/ We also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University or Sunway College or other institutions.

3. Submit "Turn-it-in" report: Yes ☑ No □

No	Name of Student	Student ID No.	E-mail Address	Signature
1	Timothy Chen Xian Yii	18098392	18098392@imail.sunway.edu.my	*chen*

Marker's Comments:

Marks Awarded: Date:

Executive Summary

The World Happiness Report 2019 depicts the happiness scores of all countries in the world based on several variables. The aim of this study is to determine the significance of the variables in contributing to happiness scores via multivariate techniques. The dataset (and additional “Continent” dataset) was obtained from Kaggle and data exploratory analysis is performed before in-depth analysis.

1. Visualizations such as summary statistics, boxplots, and violin plots are constructed to understand our dataset and below are some key points: -
   - Total 10 variables with continents included, 7 of which are continuous variables.
   - Score seems to be correlated with GDP, social support and life expectancy
   - Outliers are present but no deletions are made to preserve information.
   - Western continents such as Europe, North/South America are found to be happier in general.

2. To further analyze the variables, PCA and clustering are performed as a means to find out hidden structures in our variables. Score, GDP, social support, and life expectancy are largely loaded in PC1, which corresponds to the socio-economic status of a country while generosity and corruption ranks higher in PC2, indicating humbleness shown by people in the country.

3. A biplot was constructed and discovered that countries like Denmark and Singapore considerably positively loaded for both PC1 and PC2. Myanmar was seen as the humblest country in 2019 while Greece was the least humble. Countries like Afghanistan and Central African Republic have among the worst socio-economic status as they are hugely negatively loaded in PC1.

4. Clustering analysis shows 4 clusters in total, countries highlighted in blue such as New Zealand means that they are first world countries with developed infrastructures and economy. Moving negatively along PC1 shows a red cluster, which could indicate that these countries are on the lower side of socio-economic statuses.

5. Green cluster highlights that these countries show quite a significant level of humbleness while the light blue cluster shows the opposite.

6. In short, the happiness score is largely correlated with the GDP, social support, and life expectancy. Based on PCA, most countries located in quadrant 1 of the biplot (positively loaded for PC1 and PC2) belongs to Europe or North America continent, hence enforcing the fact that western continents are generally happier.

7. Apart from that, external factors such as the debt crisis in Greece, ongoing conflicts in Syria, and religious teachings in Myanmar highly affects their overall happiness scores.

Tables of Content

Introduction

The World Happiness Report is a survey conducted every year to reflect the state of happiness in each country. To date, the reports are being continuously used by important officials such as government bodies and large organizations as it acts as an indicator for better decision making. Experts from various fields have described that the measurements of the well-being of a country can be used beneficially when it comes to assessing the progress of different nations around the world.

The happiness scores and rankings use data from the Gallup World Poll (GWP) and are based on the answers to the life evaluation questions that were in the poll. The questions were answered based on the Cantril Ladder of Life Satisfaction, which let the respondents to imagine a ladder with the best possible score as 10 and the worst 0. From there, the respondents were asked to rate their own current lives based on that scale.

The columns following the happiness score estimate the extent to which six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to happiness scores in each country than they are in Dystopia, which is a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. (Sustainable Development Solutions Network, 2019).

Data Exploration & Analysis

The dataset was obtained from Kaggle, uploaded by Sustainable Development Solutions Network. The data originally consists of 9 variables, namely Country, Overall Rank, Score, GDP Per Capita, Social Support, Life Expectancy, Freedom, Generosity, and Corruption. For a more in-depth visualization, continents are added into the dataset by importing and merging another Excel spreadsheet which lists all countries and their respective continents. The merged columns are then converted from strings to factors before data exploration. Below shows the structure and summary statistics of the data after conversion: -

Table 1: Structure of All Variables in World Happiness Report 2019

Variable Names	Description
Country	List of all the name of the countries
Overall Rank	Ranking based on the happiness scores
Score	Total score based on the variables (including Dystopia index)
GDP Per Capita	Economic production score
Social Support	Score derived from support from friends and relatives
Life Expectancy	Expected life expectancy of the average citizen of the country
Freedom	Freedom given in letting people making their own life choices
Generosity	Level of generosity shown by citizens
Corruption	Perceived corruption by citizens (the higher the index, the lower the corruption)

Continent	List of continents according to country

From Table 1, there are a total of 156 observations and 10 variables, 2 of which are non-numeric while the overall rank is ordinal by default. To understand the numbers, a summary statistics was performed to show metrics of the data such as median, mean, quartile range (QR) etc.

Table 2: Summary Statistics of All Continuous Variables in World Happiness Report 2019

	Score	GDP_per_capita	Social_support	Life_expectancy	Freedom	Generosity	Corruption
Minimum	2.853	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1st QR	4.545	0.6028	1.056	0.5477	0.3080	0.1087	0.0470
Median	5.380	0.9600	1.272	0.7890	0.4170	0.1775	0.0855
Mean	5.407	0.9051	1.209	0.7252	0.3926	0.1848	0.1106
3rd QR	6.184	1.2325	1.452	0.8818	0.5072	0.2482	0.1412
Max	7.769	1.6840	1.624	1.1410	0.6310	0.5660	0.4530

The summary statistics shows that there are no missing values present in the dataset. It also shows that variables such as GDP_per_capita, Social_support, and Life_expectancy have higher mean values compared to other variables. This infers that these 3 variables contribute more in making a country happier. Based on the scores, generally speaking, most countries are well above average in the happiness scores as the maximum recorded is only 7.769 and more than 75% of it lies above 50% of the score of rank 1 according to the interquartile range.

Next, a correlation plot is constructed to visualize any correlation between the variables in the dataset.

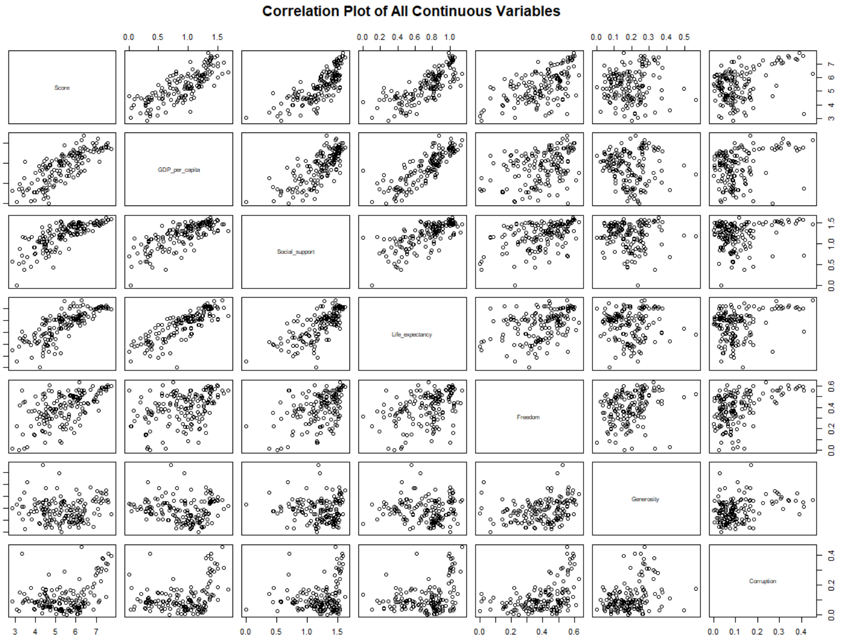


Figure 1: Correlation Plot of All Continuous Variables

The correlation plot suggests that the score is highly correlated with the GDP, social support, and life expectancy as the plots are shows a positive gradient, inferring a positive relationship while the rest of the variables does not show an obvious relation between each other.

A boxplot is used for the next step in visualization to further provide more insights into our dataset. Since the ‘Score’ variable is just the cumulative sum of all scores from the other 6 numeric variables, the Dystopia index (1.88), and also the residuals of the scores which varies from countries (Helliwell et al., 2019), the ‘Score’ variable is excluded from the plot as the values may skew the other boxplots.

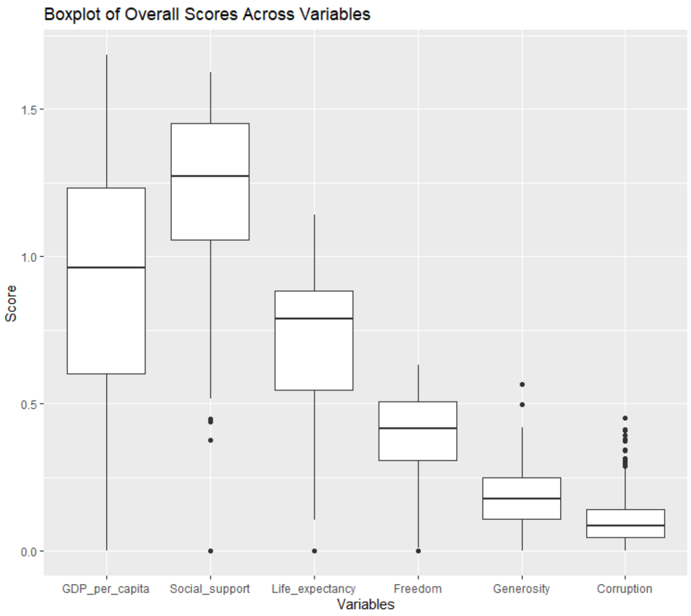


Figure 2: Boxplots of All Continuous Variables Except Score

Based on the boxplots, it is observed that on average, the social support variable is higher than GDP and life expectancy. This narrows our conclusion in determining which variables contributes the most in the happiness scores. The general consensus is still the same as in the summary statistics where the first 3 variables are placed higher in mean score then the rest. However, it is also observed that there are quite a few number of outliers (denoted in black dots), especially for corruption variables. For the sake of accuracy, the outliers are not removed as the dataset is small and every observation is crucial.

The boxplots are further modified by grouping them with their respective continents and each variables. The results for each boxplots were then arranged in a grid for aesthetic and readability purposes. The output of the results is shown below: -

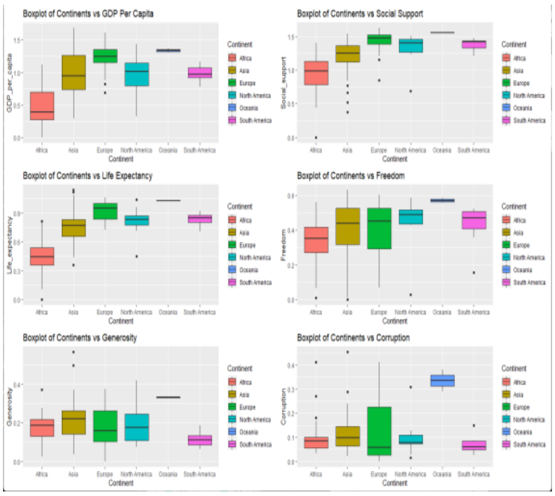


Figure 3: Boxplots of Continents Against All Variables

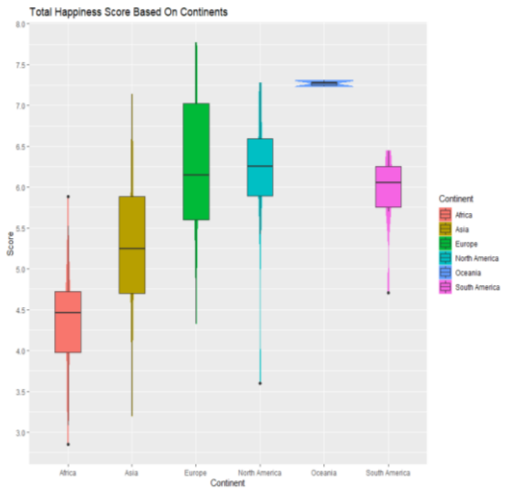


Figure 4: Violin Plot of the Overall Score

Based on Figure 3, the shape of the boxplots for GDP, social support, and life expectancy seems to be similar, indicating a correlation between the 3 variables. For Oceania, the boxplot is mostly thin for all plots as the only countries in that group is Australia and New Zealand. It is also found that the shape of the boxplots for all continents except Oceania & South America in generosity variable are all similar in size. The corruption variable varies a lot for Europe as the boxplot is longer than the rest.

Lastly, a violin plot of the overall scores is constructed and shown in Figure 4. From there, it is found that for continents such as Asia and Africa, the happiness scores tend to be lower compared to other western dominated continents such as Europe, North & South America, and Oceania. It is also suggested that part of the scores in South America are ranked higher than the 3rd interquartile range of the boxplot since the density is thicker compared to the rest of the line.

Further Data Analysis

Bivariate techniques are usually not sufficient in giving us insights into the dataset. Hence, additional multivariate analysis techniques such as principal component analysis (PCA) are adopted into the research so that complex datasets can be broken down to be more interpretable and gives us further hidden patterns and insights by reducing the dimensionality of the data while retaining the most important trends and patterns at the same time (Lever, 2017).

In this section, two techniques will be explored, which is PCA, followed by clustering analysis. The explanations from PCA are used for further explaination in clustering analysis.

1. Principal Component Analysis (PCA)

PCA is a multivariate technique in which its main purpose is to reduce the dimensionality of the dataset so that it is easier to interpret the data. Only numeric variables are used in this analysis as PCA works most optimally when the input data is all numeric in nature. Hence, any variables that is not continuous are excluded from this analysis. The PCA is carried out and the loadings are shown below: -

Table 3: Loading Vectors for the first two Principal Components (PC)

Variable	PC1	PC2
Score	0.4759	-0.0284
GDP per capita	0.4548	-0.2134
Social Support	0.4366	-0.2071
Life Expectancy	0.4502	-0.1779
Freedom	0.3322	0.3621
Generosity	0.0482	0.6938
Corruption	0.2465	0.5163

In PC1, score, GDP, social support, and life expectancy have a rather large weight placed. This suggests that the first component of the loading vectors corresponds to the overall social-economic status and well-being of the country. As the GDP of a country increases, the rate of poverty decreases, so there will be more financially capable person on average. Hence, there will be more support given in the family on average and their lifestyle can be very well-managed and healthy, which leads to higher life expectancy and better satisfaction in life.

In PC2, generosity and corruption seem to have the largest weight placed. This suggests that the second component of the vectors corresponds to the willingness to help those in need in response to high levels of corruption in the country. As corruption rises, people see the need in helping those affected by it.

Next, a biplot was then plotted to visualize the loading vectors and how each observation interacts with the components. Below is a diagram showing the biplot: -

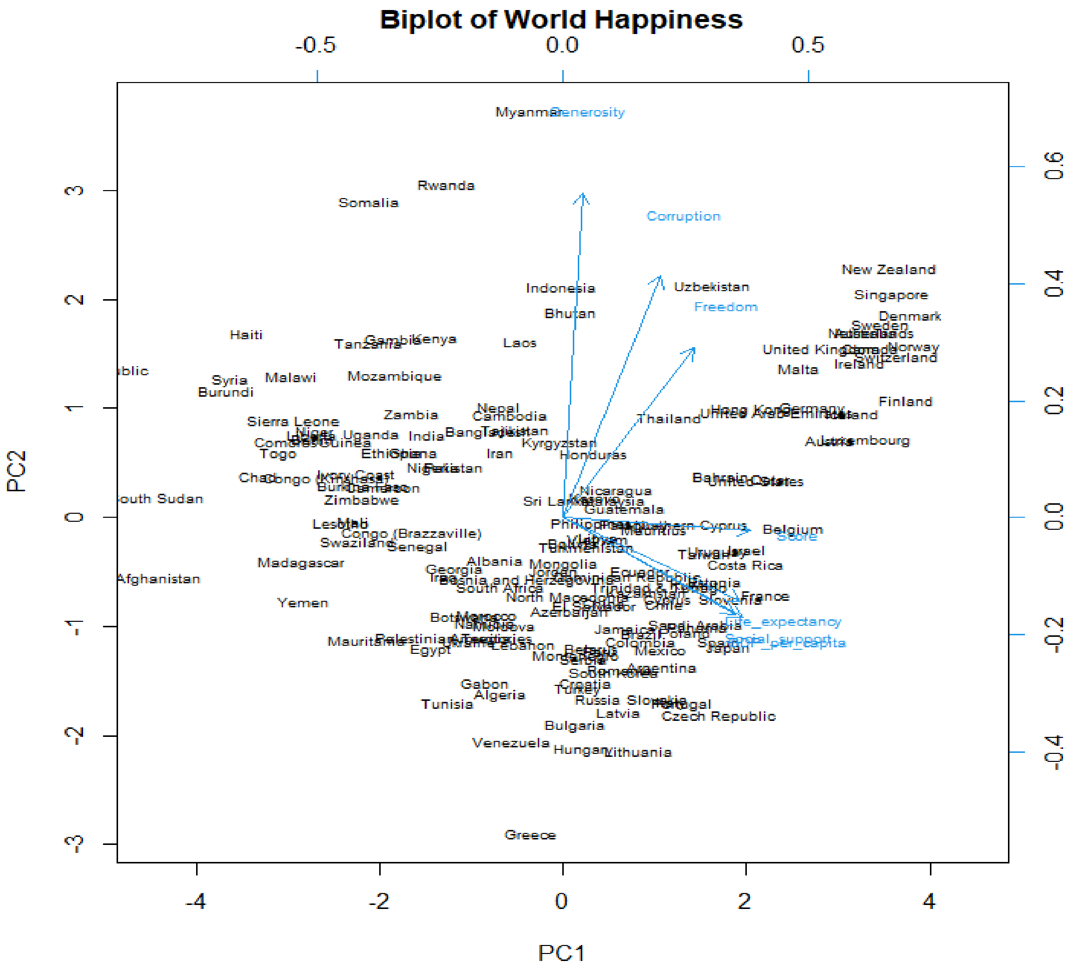


Figure 4: Biplot of World Happiness Report 2019

Most of the variables except for generosity is positively loaded in the first PC, with "Score" variable being the most parallel to the PC1 axis. Meanwhile, generosity is largely positively loaded on the second PC instead, close to being parallel to PC2 as well.

Countries such as Singapore, New Zealand, and Denmark are better in terms of social-economic status. This is evident from the fact that these countries are developed and have economic presence worldwide, hence families have more opportunities and necessary funds in providing support to their loved ones to ease their financial burden coming from various shortcomings. This also leads to a more extended life expectancy as they can have a better and healthy lifestyle on average, which leads to a more happier country in general. Conversely, countries such as Afghanistan and South Sudan is negatively loaded in PC1, which indicates that the support was minimal or non-existent due to the fact that the countries are not well-developed and have a lower GDP on average than other countries.

On the other hand, countries such as Myanmar are largely positively loaded on PC2, which infers that Myanmar is the most humble country in the world since PC2 corresponds to the humility of the general populace while Greece is deemed to be the least humble country based on the 2019 report. The high levels of generosity in Myanmar could evidently caused by the teachings of religion and the act of giving practiced by said religion (Cole, 2015). Meanwhile, the low levels of generosity found in Greece could be caused by the debt crisis ever since 2010 (Amadeo, 2020), hence it likely caused people in Greece to developed a more selfish attitude.

Lastly, a scree plot is constructed as it helps in determining how many principal components should be included during the analysis and also the eigenvalues that is recommended to be used for clustering analysis later on.

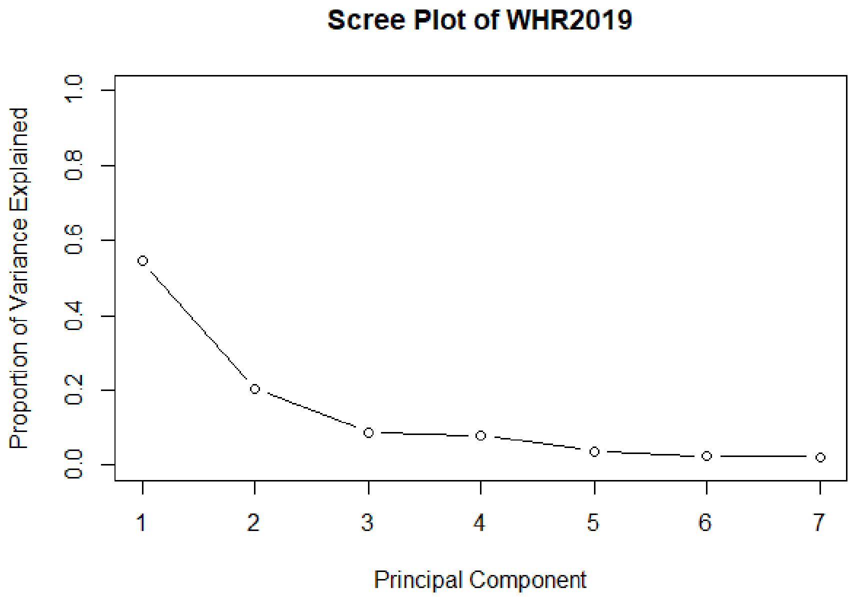


Figure 5: Scree Plot of World Happiness Report 2019

According to the plot, the elbow is located at PC2, which captures around 70% variance in the data and this is sufficient enough for the analysis. This means that a plot of PC1 and PC2 is enough to approximately explains the variance in the dataset. However, an additional PC3 can also be included as the variance captured will be more than 80% when PC3 is included in the analysis. PC3 mainly explains on the willingness to help others in combating corruption but more plots have to be constructed and it is not as easily interpretable compared to only two PCs.

2. Clustering Analysis

Clustering analysis is a technique which groups similar observations into a number of clusters based off of several variables for each individual observation. With the help of PCA, clustering analysis allows the further interpretation of similar observations based on the clusters according to the principal components explained. Most common clustering method is the k-means clustering, whereby the number of k (centroid) is pre-determined based on the eigenvalues of the scree plot. The scree plot produced an eigenvalue of 2 since the elbow is located at PC2, hence k=2 is sufficient. However, it is more logical to have a cluster for each ends the PCs, i.e. high & low scores for both PC1 and PC2, hence the value of k is chosen to be 4. The percentage of variance explained will be higher as well since we are taking 4 eigenvalues, i.e. 4 PCs.

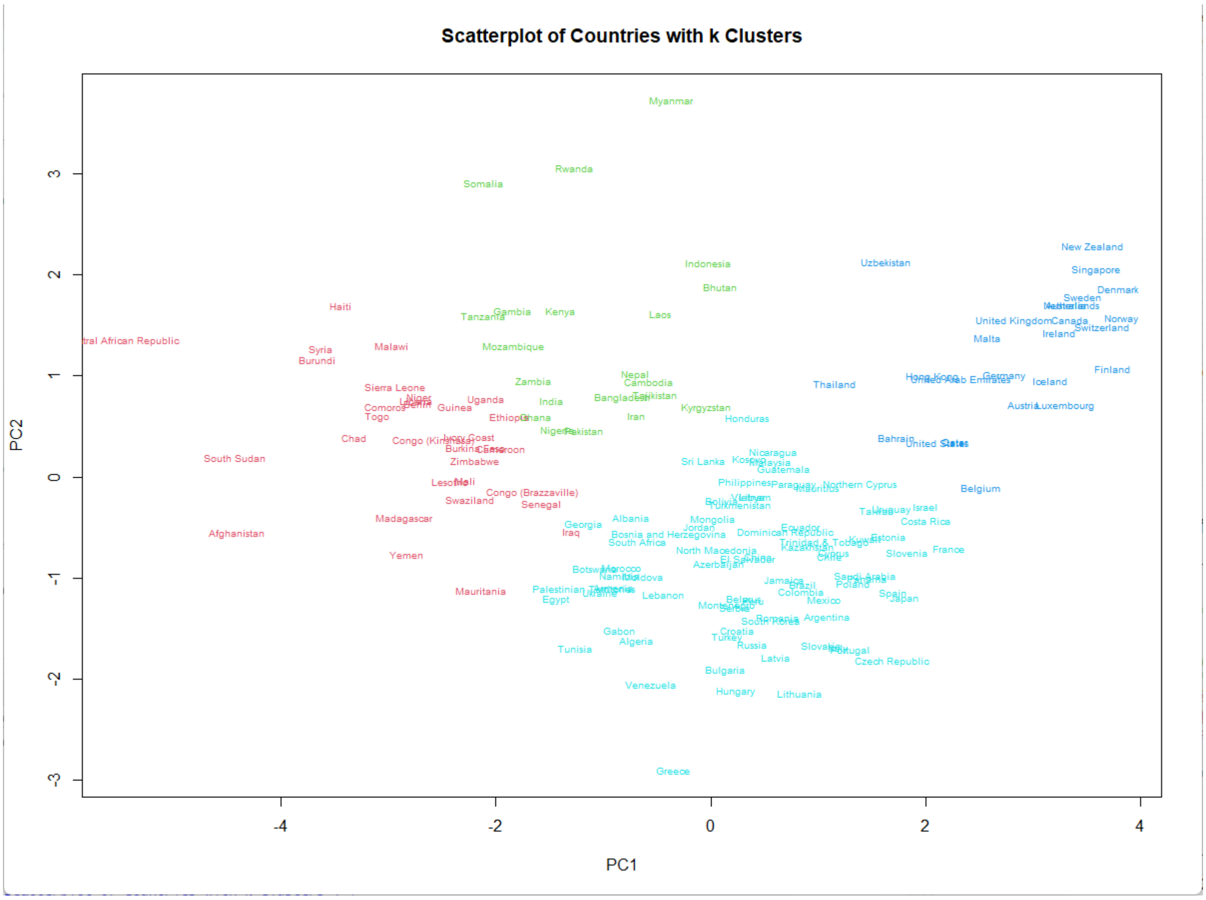


Figure 6: Scatterplot of All Countries based on k Clusters

With k=4, there will be a total of 4 clusters present in the scatterplot. Based on the scatterplot, it is observed that clusters highlighted in blue corresponds to countries which has a high socio-economic status since they are positively loaded at PC1 and also have high levels of humbleness. The green clusters represent countries such as Myanmar have shown a high humility status among their citizens while the light blue clusters represent the opposite of the green clusters since they tend to be more negatively loaded on PC2. This cluster mainly shows countries which are have lower humbleness shown such as Malaysia and Phillipines since they ranked lower based on PC2. It is also inferred that although Japan and France are very developed countries, they belong to the light blue clusters due to their low humility shown. Last but not least, the red clusters on the left side of the plot represents countries which has low socio-economic status and are underdeveloped because they are largely negatively loaded in PC1.

Discussion

In summary, it is safe to assume that one country's happiness ratings would always be dependent on the GDP per capita, overall life expectancy of its citizens, and also the aspect of social support from friends or family members based on the 1st component. First world countries with established GDP would automatically rank higher in the happiness leaderboard since GDP are correlated with the life expectancy and social support in general.

With PCA, it was observed that Myanmar was crowned as the humblest country based on its score and Greece as the least humble country in 2019. To further explain, clustering analysis suggested that apart from Myanmar, Rwanda and Somalia have similar observations as Myanmar, making them being grouped into the same cluster together. Similarly, countries like Afghanistan and Syria ranked lower in terms of PC1. This is expected as there is still the ongoing conflict in 2019, which further causes the nation to fall lower in socio-economic status.

Based on the clusters observed, some countries which look like they are fully developed such as Japan and France, ends up not belonging to the red clusters since the people tend to not be that humble. This could be caused by the fact that some developed countries tend to overworked people and offer little to no resources in coping with mental health, which led to lowering the positive mental attitude of its citizens.

To further improve the analysis, it is recommended that a more thorough preprocessing procedure should be carried out such as removing outliers since it could skew the factor scores, making it more confusing to interpret. A dummy variable can be considered in the analysis to include parts of the categorical data that is in the dataset such as continent, rank etc. For PCA, a third component can be introduced in PCA as it captures more variance of data but at the cost of easier interpretation of variables as it might be hard to group them together.

Conclusion

In short, based on PCA, a country's happiness largely correlates with the GDP of the country, along with life expectancy and social support. The level of generosity also correlates with lower corruption perceived in a country. With clustering analysis, countries like Singapore and New Zealand demonstrates a high socio-economic presence worldwide alongside humbler citizens. Conversely, countries such as Afghanistan and Syria have relatively low socio-economic status likely due to ongoing conflicts. Lastly, although having moderate socio-economic status, Myanmar was crowned as the humblest country in 2019 due to it being the most loaded in PC2 whilst Greece the least humble country as most likely caused by the debt crisis. Overall, developed western continents such as Europe and North America tends to be happier due to their culture and how they were raised.

References

Amadeo, K. (2020, May 18). *Understand the Greek Debt Crisis in 5 Minutes*. The Balance. Retrieved June 15, 2022, from https://www.thebalance.com/what-is-the-greece-debt-crisis-3305525#:%7E:text=Since%20the%20debt%20crisis%20began,scheduled%20debt%20payments%20beyond%202060

Cole, D. (2015, November 28). *You'll Never Guess The Most Charitable Nation In The World*. NPR. Retrieved June 15, 2022, from https://www.npr.org/sections/goatsandsoda/2015/11/28/457101304/youll-never-guess-the-most-charitable-nation-in-the-world

Helliwell, J. F., Huang, H., & Wang, S. (2019, March). *Statistical Appendix 1 for Chapter 2 of World Happiness Report 2019*,. https://s3.amazonaws.com/happiness-report/2019/WHR19_Ch2A_Appendix1.pdf

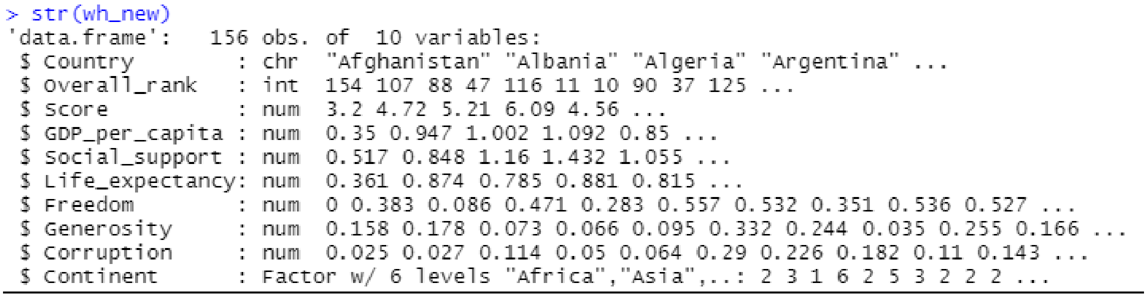
Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

List of Countries by Continent 2022. (2022). [Dataset]. https://worldpopulationreview.com/country-rankings/list-of-countries-by-continent

Sustainable Development Solutions Network. (2019, November 27). *World Happiness Report* (Version 2) [Dataset]. https://www.kaggle.com/datasets/unsdsn/world-happiness

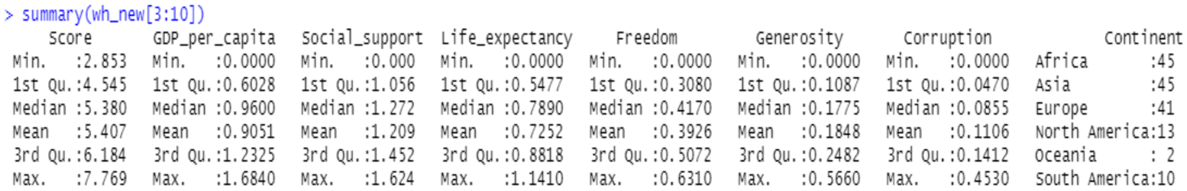
Appendix A – Tabular values

Structure of Variables



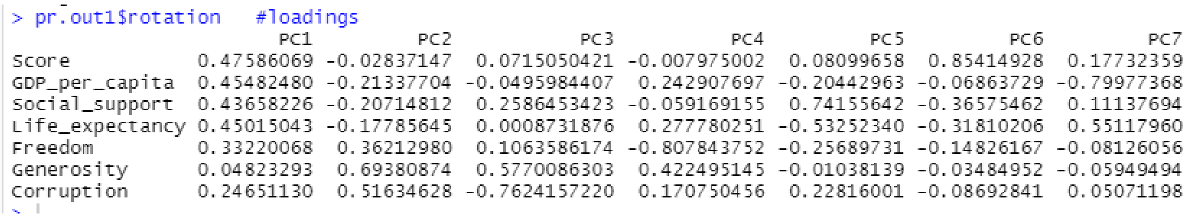
```
> str(wh_new)
'data.frame':   156 obs. of  10 variables:
 $ Country        : chr  "Afghanistan" "Albania" "Algeria" "Argentina" ...
 $ Overall_rank   : int  154 107 88 47 116 11 10 90 37 125 ...
 $ Score          : num  3.2 4.72 5.21 6.09 4.56 ...
 $ GDP_per_capita : num  0.35 0.947 1.002 1.092 0.85 ...
 $ Social_support : num  0.517 0.848 1.16 1.432 1.055 ...
 $ Life_expectancy: num  0.361 0.874 0.785 0.881 0.815 ...
 $ Freedom        : num  0 0.383 0.086 0.471 0.283 0.557 0.532 0.351 0.536 0.527 ...
 $ Generosity     : num  0.158 0.178 0.073 0.066 0.095 0.332 0.244 0.035 0.255 0.166 ...
 $ Corruption     : num  0.025 0.027 0.114 0.05 0.064 0.29 0.226 0.182 0.11 0.143 ...
 $ Continent      : Factor w/ 6 levels "Africa","Asia",..: 2 3 1 6 2 5 3 2 2 2 ...
```

Summary Statistics



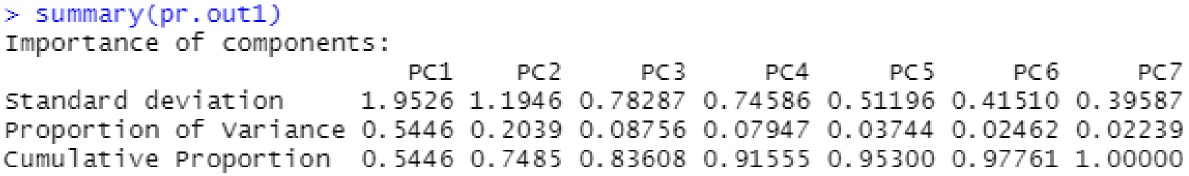
```
> summary(wh_new[3:10])
     Score       GDP_per_capita   Social_support  Life_expectancy     Freedom         Generosity       Corruption              Continent
 Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Africa       :45
 1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477   1st Qu.:0.3080   1st Qu.:0.1087   1st Qu.:0.0470   Asia         :45
 Median :5.380   Median :0.9600   Median :1.272   Median :0.7890   Median :0.4170   Median :0.1775   Median :0.0855   Europe       :41
 Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252   Mean   :0.3926   Mean   :0.1848   Mean   :0.1106   North America:13
 3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818   3rd Qu.:0.5072   3rd Qu.:0.2482   3rd Qu.:0.1412   Oceania      : 2
 Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410   Max.   :0.6310   Max.   :0.5660   Max.   :0.4530   South America:10
```

PCA Loadings



```
> pr.out1$rotation   #loadings
                       PC1         PC2           PC3          PC4         PC5         PC6         PC7
Score           0.47586069 -0.02837147  0.0715050421 -0.007975002  0.08099658  0.85414928  0.17732359
GDP_per_capita  0.45482480 -0.21337704 -0.0495984407  0.242907697 -0.20442963 -0.06863729 -0.79977368
Social_support  0.43658226 -0.20714812  0.2586453423 -0.059169155  0.74155642 -0.36575462  0.11137694
Life_expectancy 0.45015043 -0.17785645  0.0008731876  0.277780251 -0.53252340 -0.31810206  0.55117960
Freedom         0.33220068  0.36212980  0.1063586174 -0.807843752 -0.25689731 -0.14826167 -0.08126056
Generosity      0.04823293  0.69380874  0.5770086303  0.422495145 -0.01038139 -0.03484952 -0.05949494
Corruption      0.24651130  0.51634628 -0.7624157220  0.170750456  0.22816001 -0.08692841  0.05071198
```

Summary importance of PCA



```
> summary(pr.out1)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.9526 1.1946 0.78287 0.74586 0.51196 0.41510 0.39587
Proportion of Variance 0.5446 0.2039 0.08756 0.07947 0.03744 0.02462 0.02239
Cumulative Proportion  0.5446 0.7485 0.83608 0.91555 0.95300 0.97761 1.00000
```

Appendix B – R code

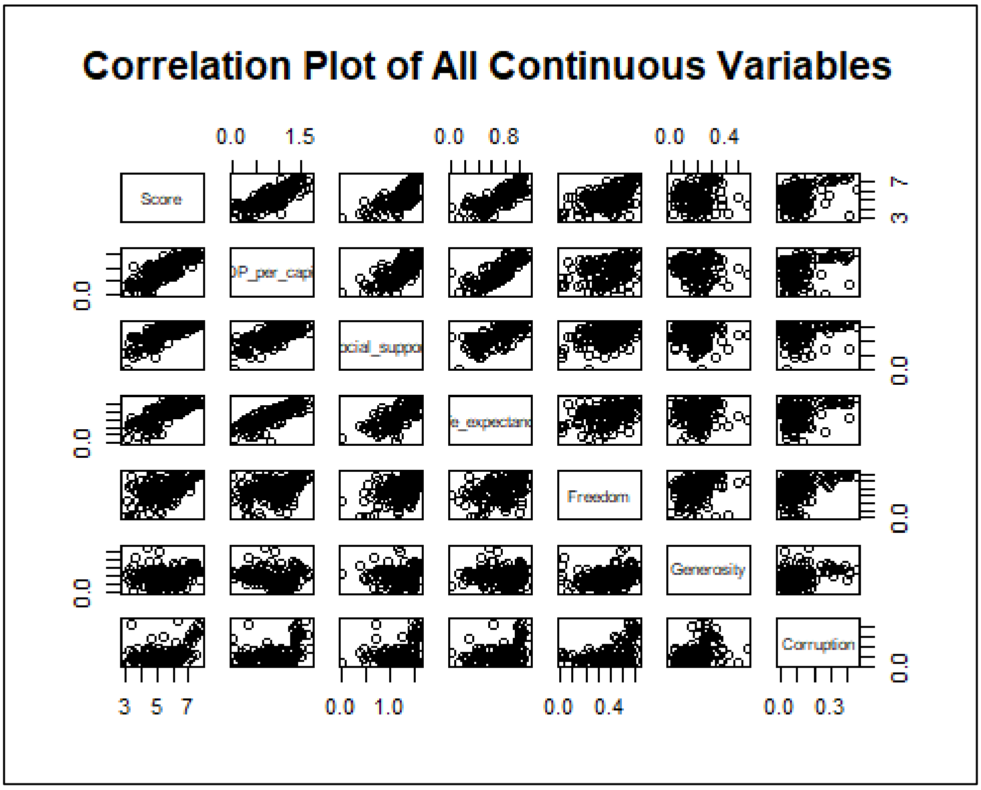
#Import library and dataset

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.1.3

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.1.3

library(psy)

## Warning: package 'psy' was built under R version 4.1.3

library(MASS)
library(psych)

## Warning: package 'psych' was built under R version 4.1.3

##
## Attaching package: 'psych'

## The following object is masked from 'package:psy':
##
##     wkappa

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

whdata <- read.csv("WHR2019.csv", header = TRUE)
whdata_raw <- read_xls("WHR2019 Raw Data.xls", sheet="Figure2.6")  # to ad
```

```
d dystopia index and residual
continents <- read.csv("Continents.csv", header = TRUE, fileEncoding="UTF-
8-BOM")  # to add continent to each country

wh_new <- merge(whdata,whdata_raw[, c("Country","Dystopia (1.88) + residua
l")], by="Country")
wh_new <- merge(whdata,continents[, c("Country","Continent")], by="Country
")
View(wh_new)
```

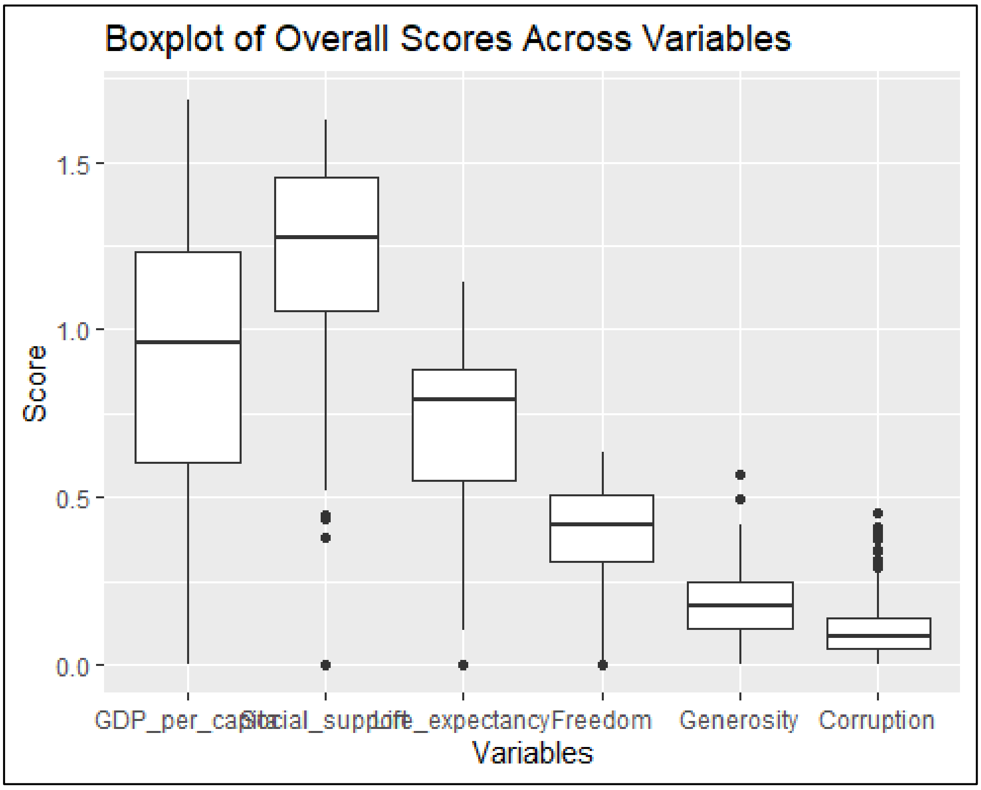
Exploratory Analysis

```
# Structure of each data
wh_new$Continent <- as.factor(wh_new$Continent)
str(wh_new)

## 'data.frame':    156 obs. of  10 variables:
##  $ Country        : chr  "Afghanistan" "Albania" "Algeria" "Argentina"
...
##  $ Overall_rank   : int  154 107 88 47 116 11 10 90 37 125 ...
##  $ Score          : num  3.2 4.72 5.21 6.09 4.56 ...
##  $ GDP_per_capita : num  0.35 0.947 1.002 1.092 0.85 ...
##  $ Social_support : num  0.517 0.848 1.16 1.432 1.055 ...
##  $ Life_expectancy: num  0.361 0.874 0.785 0.881 0.815 ...
##  $ Freedom        : num  0 0.383 0.086 0.471 0.283 0.557 0.532 0.351 0.
536 0.527 ...
##  $ Generosity     : num  0.158 0.178 0.073 0.066 0.095 0.332 0.244 0.03
5 0.255 0.166 ...
##  $ Corruption     : num  0.025 0.027 0.114 0.05 0.064 0.29 0.226 0.182
0.11 0.143 ...
##  $ Continent      : Factor w/ 6 levels "Africa","Asia",..: 2 3 1 6 2 5
3 2 2 2 ...

# Summary Statistics
summary(wh_new[3:10])

##      Score       GDP_per_capita   Social_support  Life_expectancy
##  Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477
##  Median :5.380   Median :0.9600   Median :1.272   Median :0.7890
##  Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252
##  3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818
##  Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410
##     Freedom          Generosity       Corruption             Continent
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Africa       :45
##  1st Qu.:0.3080   1st Qu.:0.1087   1st Qu.:0.0470   Asia         :45
##  Median :0.4170   Median :0.1775   Median :0.0855   Europe       :41
##  Mean   :0.3926   Mean   :0.1848   Mean   :0.1106   North America:13
##  3rd Qu.:0.5072   3rd Qu.:0.2482   3rd Qu.:0.1412   Oceania      : 2
##  Max.   :0.6310   Max.   :0.5660   Max.   :0.4530   South America:10

#correlation plots
pairs(wh_new[3:9], main="Correlation Plot of All Continuous Variables")
```



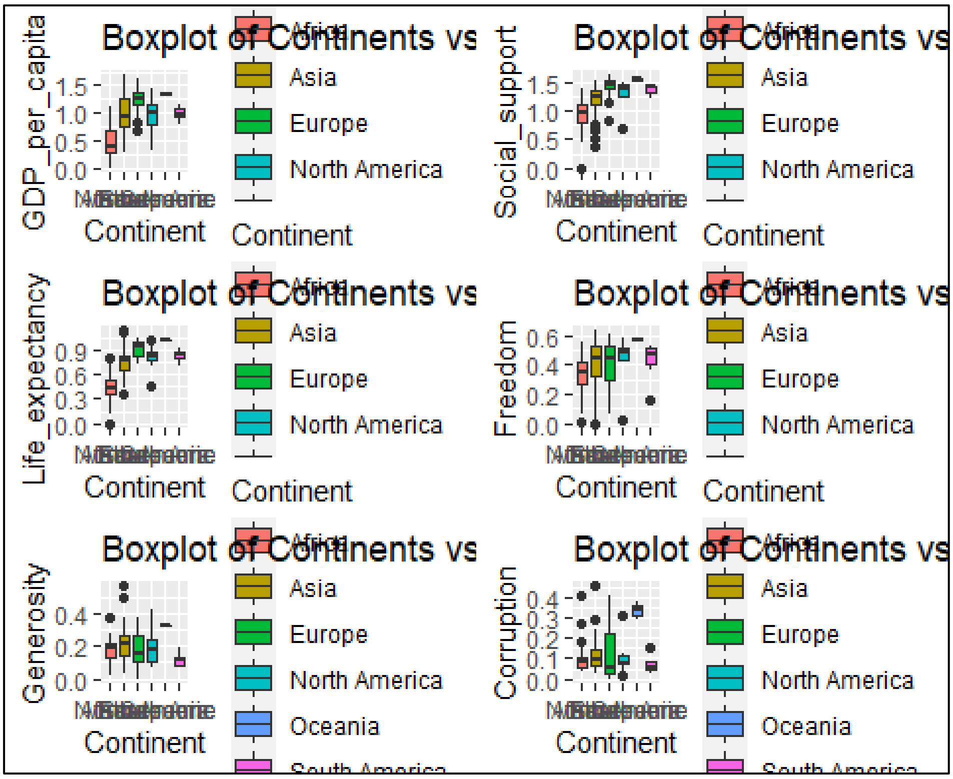
Box plots

```
bp1 <- ggplot(stack(wh_new[4:9]), aes(x=ind,y=values)) +
  geom_boxplot() +
  labs(y="Score", x="Variables", title="Boxplot of Overall Scores Across V
ariables")

bp1 # Box plot for each variable regardless of countries or continents
```

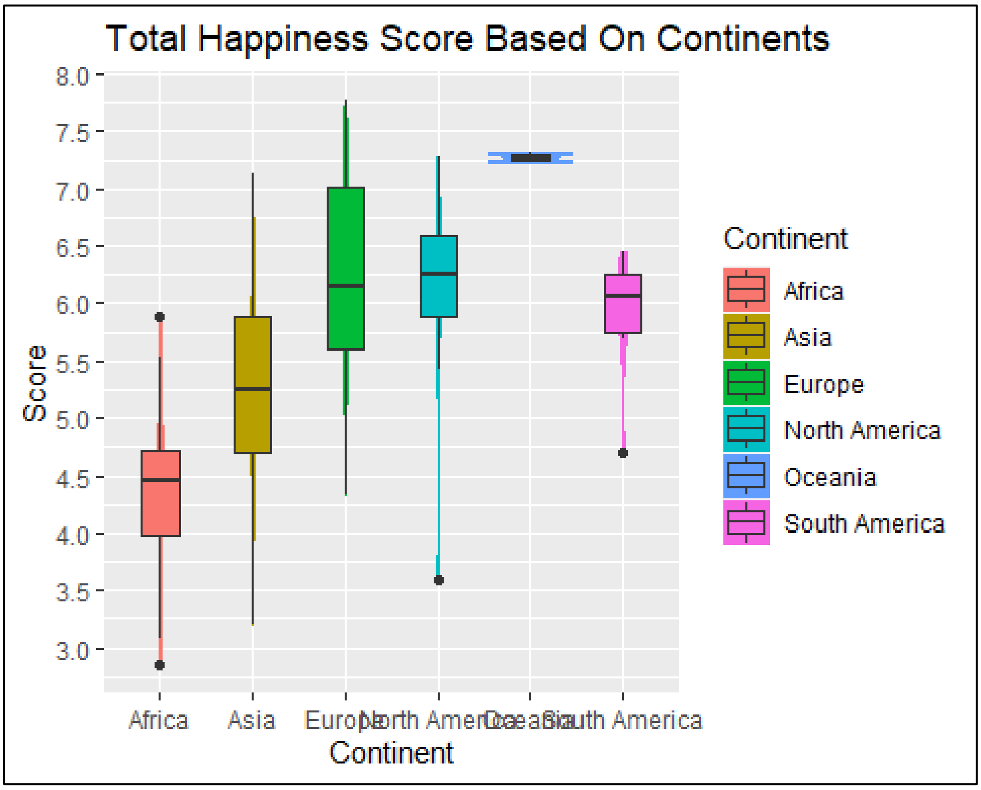


```
# Box plots for each variables based on continents
gdp <- ggplot(data = wh_new, aes(x = Continent, y = GDP_per_capita, fill=C
ontinent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs GDP Per Capita")

soc_supp <- ggplot(data = wh_new, aes(x = Continent, y = Social_support, f
ill=Continent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs Social Support")

life_exp <- ggplot(data = wh_new, aes(x = Continent, y = Life_expectancy,
fill=Continent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs Life Expectancy")

free <- ggplot(data = wh_new, aes(x = Continent, y = Freedom, fill=Contine
nt)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs Freedom")

genero <- ggplot(data = wh_new, aes(x = Continent, y = Generosity, fill=Co
ntinent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs Generosity")

corrupt <- ggplot(data = wh_new, aes(x = Continent, y = Corruption, fill=C
ontinent)) +
  geom_boxplot() +
  ggtitle("Boxplot of Continents vs Corruption")
```

```
grid.arrange(gdp, soc_supp, life_exp, free, genero, corrupt, ncol=2, nrow=
3)
```



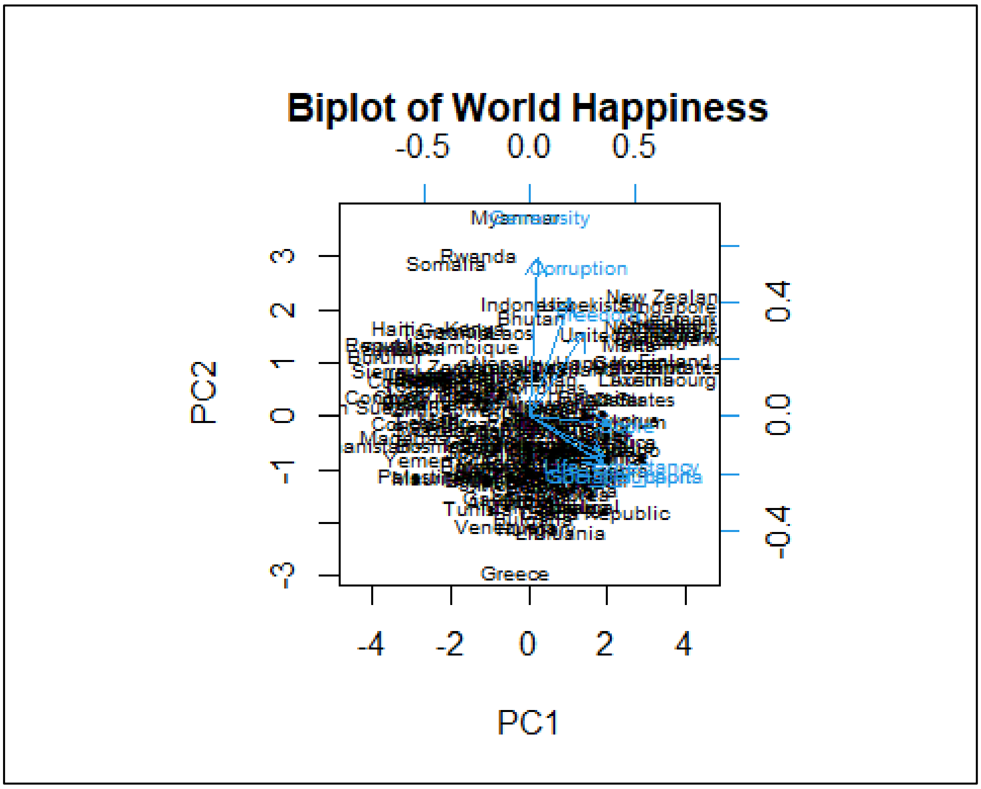
```
# Violin Plot of Score based on continent
ggplot(wh_new, aes(Continent, Score, fill=Continent)) +
  geom_violin(aes(color = Continent), trim = T)+
  scale_y_continuous("Score", breaks= seq(0,30, by=.5))+
  geom_boxplot(width=0.4)+
  theme(legend.position="right") +
  ggtitle("Total Happiness Score Based On Continents")
```



Principal Component Analysis

```
wh_new2 <- wh_new[,-1]
rownames(wh_new2) <- wh_new[,1] #change row names

whdat <- wh_new2[2:8]
pr.out1 = prcomp(whdat, scale = TRUE)
pr.out1$rotation   #loadings

##                        PC1         PC2           PC3          PC4
   PC5
## Score          0.47586069 -0.02837147  0.0715050421 -0.007975002  0.08
099658
## GDP_per_capita  0.45482480 -0.21337704 -0.0495984407  0.242907697 -0.20
442963
## Social_support  0.43658226 -0.20714812  0.2586453423 -0.059169155  0.74
155642
## Life_expectancy 0.45015043 -0.17785645  0.0008731876  0.277780251 -0.53
252340
## Freedom         0.33220068  0.36212980  0.1063586174 -0.807843752 -0.25
689731
## Generosity      0.04823293  0.69380874  0.5770086303  0.422495145 -0.01
038139
## Corruption      0.24651130  0.51634628 -0.7624157220  0.170750456  0.22
816001
##                         PC6         PC7
## Score           0.85414928  0.17732359
## GDP_per_capita -0.06863729 -0.79977368
## Social_support -0.36575462  0.11137694
## Life_expectancy -0.31810206  0.55117960
## Freedom          -0.14826167 -0.08126056
## Generosity       -0.03484952 -0.05949494
## Corruption       -0.08692841  0.05071198
```

```
#biplot
biplot(pr.out1 , scale =0, col=c(9,4), cex = 0.6,xlim=c(-4.5,4.5), main="B
iplot of World Happiness")
```



```
#proportion of variance
summary(pr.out1)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
  PC7
## Standard deviation     1.9526 1.1946 0.78287 0.74586 0.51196 0.41510 0.
39587
## Proportion of Variance 0.5446 0.2039 0.08756 0.07947 0.03744 0.02462 0.
02239
## Cumulative Proportion  0.5446 0.7485 0.83608 0.91555 0.95300 0.97761 1.
00000
```

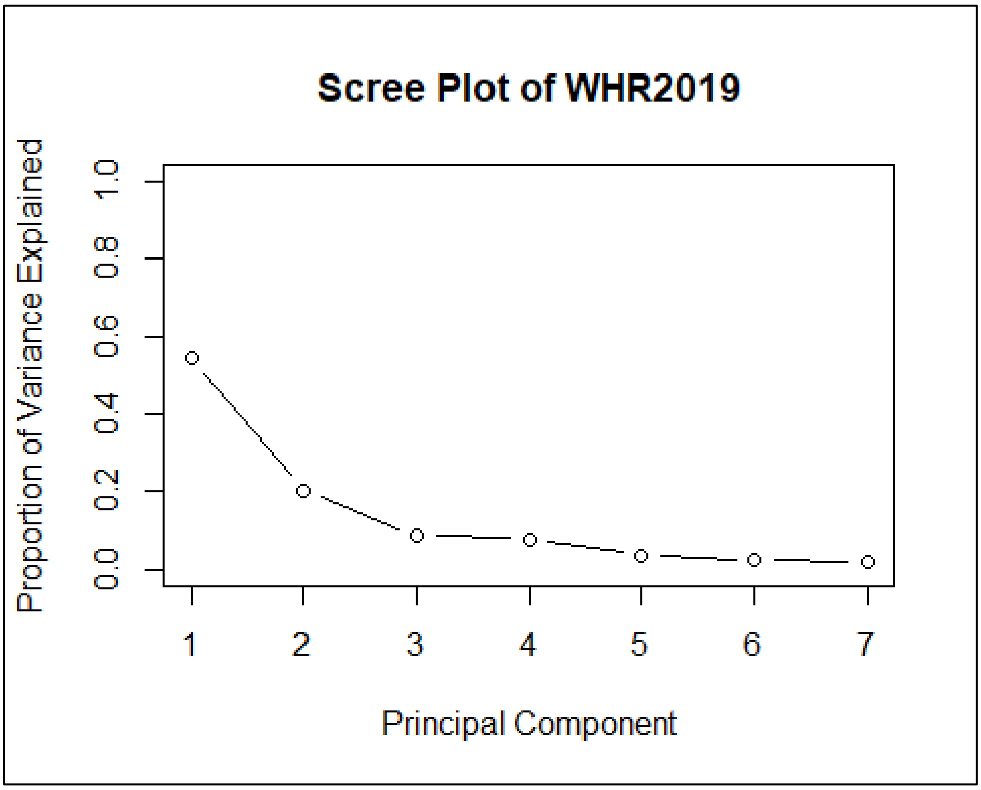
```
pr.var1 =pr.out1$sdev^2
pr.var1
```

```
## [1] 3.8125442 1.4271391 0.6128853 0.5563073 0.2621029 0.1723061 0.15671
51
```

```
pve1=pr.var1/sum(pr.var1)
pve1
```

```
## [1] 0.54464917 0.20387702 0.08755504 0.07947247 0.03744327 0.02461516 0
.02238787
```

```
#scree plot
plot(pve1 , main="Scree Plot of WHR2019",xlab=" Principal Component ", yla
b=" Proportion of Variance Explained ", ylim=c(0,1) ,type="b")
```



Clustering Analysis

```
set.seed(18098392)
km1 = kmeans(pr.out1$x[,c(1:2)],centers=4,nstart=20)
km1

## K-means clustering with 4 clusters of sizes 31, 21, 25, 79
##
## Cluster means:
##          PC1        PC2
## 1 -2.7514707  0.3624552
## 2 -1.0505362  1.4081597
## 3  2.8860592  1.2331790
## 4  0.4456377 -0.9067967
##
## Clustering vector:
##             Afghanistan                  Albania                  Alge
ria
##                       1                        4                      
  4
##               Argentina                  Armenia                Austra
lia
##                       4                        4                      
  3
##                 Austria               Azerbaijan                  Bahr
ain
##                       3                        4
```

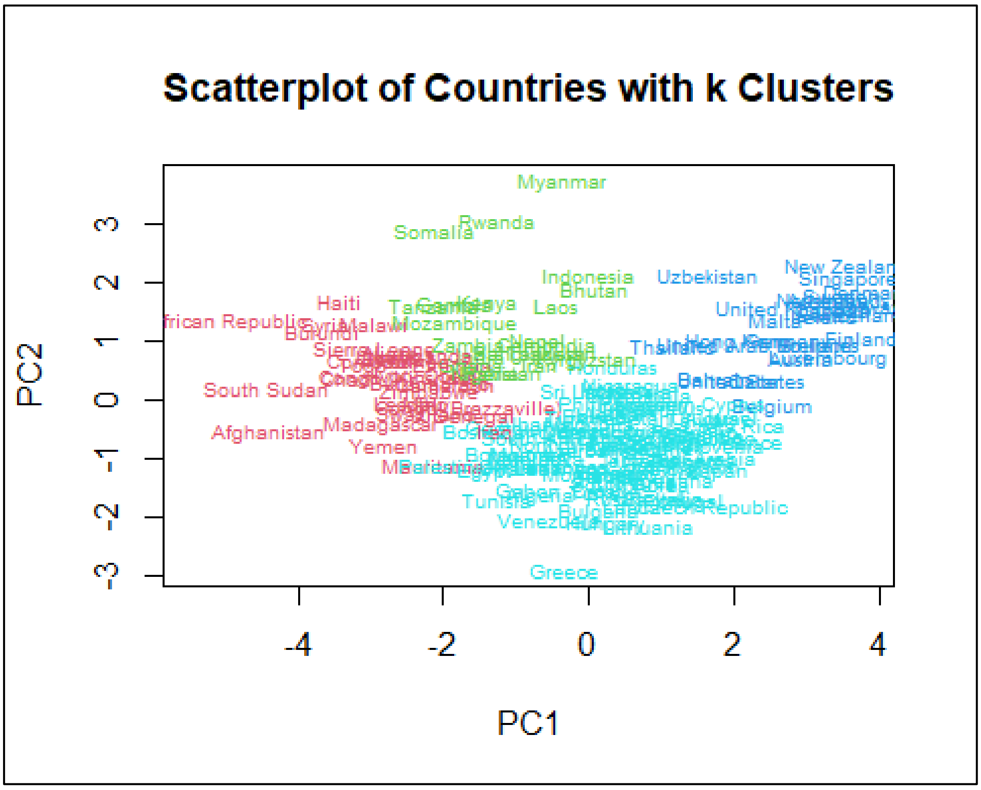
```
  3
##               Bangladesh                  Belarus                  Belg
ium
##                        2                        4
  3
##                    Benin                   Bhutan                  Boli
via
##                        1                        2
  4
##   Bosnia and Herzegovina                 Botswana                   Bra
zil
##                        4                        4
  4
##                 Bulgaria             Burkina Faso                  Buru
ndi
##                        4                        1
  1
##                 Cambodia                 Cameroon                   Can
ada
##                        2                        1
  3
## Central African Republic                     Chad                    Ch
ile
##                        1                        1
  4
##                    China                 Colombia                  Como
ros
##                        4                        4
  1
##      Congo (Brazzaville)         Congo (Kinshasa)               Costa R
ica
##                        1                        1
  4
##                  Croatia                   Cyprus           Czech Repub
lic
##                        4                        4
  4
##                  Denmark       Dominican Republic                  Ecua
dor
##                        3                        4
  4
##                    Egypt              El Salvador                  Esto
nia
##                        4                        4
  4
##                 Ethiopia                  Finland                   Fra
nce
##                        1                        3
  4
##                    Gabon                   Gambia                  Geor
gia
##                        4                        2
  4
##                  Germany                    Ghana                   Gre
```

```
ece
##                        3                        2
  4
##                Guatemala                   Guinea                    Ha
iti
##                        4                        1
  1
##                 Honduras                Hong Kong                  Hung
ary
##                        4                        3
  4
##                  Iceland                    India                Indone
sia
##                        3                        2
  2
##                     Iran                     Iraq                  Irel
and
##                        2                        1
  3
##                   Israel                    Italy              Ivory Co
ast
##                        4                        4
  1
##                  Jamaica                    Japan                   Jor
dan
##                        4                        4
  4
##               Kazakhstan                    Kenya                   Kos
ovo
##                        4                        2
  4
##                   Kuwait               Kyrgyzstan                     L
aos
##                        4                        2
  2
##                   Latvia                  Lebanon                  Leso
tho
##                        4                        4
  1
##                  Liberia                    Libya                Lithua
nia
##                        1                        4
  4
##               Luxembourg               Madagascar                   Mal
awi
##                        3                        1
  1
##                 Malaysia                     Mali                    Ma
lta
##                        4                        1
  3
##               Mauritania                Mauritius                   Mex
ico
##                        1                        4
```

```
  4
##                  Moldova                 Mongolia               Montene
gro
##                        4                        4
  4
##                  Morocco               Mozambique                  Myan
mar
##                        4                        2
  2
##                  Namibia                    Nepal              Netherla
nds
##                        4                        2
  3
##              New Zealand                Nicaragua                    Ni
ger
##                        3                        4
  1
##                  Nigeria          North Macedonia          Northern Cyp
rus
##                        2                        4
  4
##                   Norway                 Pakistan  Palestinian Territor
ies
##                        3                        2
  4
##                   Panama                 Paraguay                     P
eru
##                        4                        4
  4
##              Philippines                   Poland                 Portu
gal
##                        4                        4
  4
##                    Qatar                  Romania                   Rus
sia
##                        3                        4
  4
##                   Rwanda             Saudi Arabia                  Sene
gal
##                        2                        4
  1
##                   Serbia             Sierra Leone                Singap
ore
##                        4                        1
  3
##                 Slovakia                 Slovenia                  Soma
lia
##                        4                        4
  2
##             South Africa              South Korea              South Su
dan
##                        4                        4
  1
##                    Spain                Sri Lanka                Swazil
```

```
and
##                        4                        4
  1
##                   Sweden              Switzerland                    Sy
ria
##                        3                        3
  1
##                   Taiwan               Tajikistan                 Tanza
nia
##                        4                        2
  2
##                 Thailand                     Togo        Trinidad & Tob
ago
##                        3                        1
  4
##                  Tunisia                   Turkey             Turkmenis
tan
##                        4                        4
  4
##                   Uganda                  Ukraine     United Arab Emira
tes
##                        1                        4
  3
##           United Kingdom            United States                  Urug
uay
##                        3                        3
  4
##               Uzbekistan                Venezuela                  Viet
nam
##                        3                        4
  4
##                    Yemen                   Zambia                 Zimba
bwe
##                        1                        2
  1
##
## Within cluster sum of squares by cluster:
## [1] 37.72601 26.33020 22.57840 95.03737
##  (between_SS / total_SS =  77.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.wi
thinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
plot(pr.out1$x[, 1:2], type="n", main = "Scatterplot of Countries with k C
lusters") +
text(pr.out1$x[, 1:2], rownames(whdat), col=(km1$cluster+1), cex = 0.6)
```



1. One Drive

2. Google Drive