MS-2112 Multivariate Statistical Analysis

Project Report on World Happiness Analysis

Academic Year: 2018 – 2019

Student: Quoc Tuan Vinh, Ngo

ID: 704526

Contents

1. Motivation

2. Research Questions

3. Theoretical Background

4. Data Description

5. Data Analysis

5.1. Univariate Analysis

5.2. Bivariate Analysis

5.3. Multivariate Analysis

5.3.1. Principle Component Analysis

5.3.2. k-Mean Clustering

5.3.3. Multivariate Linear Regression

6. Conclusion

7. Evaluation

References

Appendix

Note: *The total number of pages slightly exceeds 10 pages (11 pages) because some space-consuming Figures are kept in-text to increase the report's readability. The 10-page requirement can still be met if those Figures are moved to the Appendix.*

1. Motivation

Finland is well-known for its performance in many global ranking leaderboards (Statistic Finland, 2018). In a recent publish conducted by the United Nations (UN) (YLE, 2019), Finland is again ranked as the happiest country in the world for the second consecutive year. As a foreigner coming from a tropical land, the author is truly amazed, because to him, happiness is about beaches, sunshine and parties. Finland, in contrast, has a cold weather and dark winter, but the people are still the happiest in the world! Thus, it makes the author wonder what are the characteristics of happiness defined by citizens in a country. To utilize the content learnt from this course and to satisfy a part of his curiosity, the author therefore would like to dig into this ranking – the *World Happiness Report (WHR).*

WHR is an international annual report published by the UN, comparing amongst countries according to the perception of their citizens on various life factors. Methodology-wise, the report ranks countries based on the *average happiness score*, by asking the citizens to evaluate the quality of their current lives, averaged over the last 03 years. In addition, it tries to explain this score by modeling it with other factors, such as GDP per capita, social support – to name a few. Even though there are various controversies about its methodologies, metrics and data (Smith, 2017), WHR still gives an idea of how quality of life varies across countries globally.

In this project, the author only uses the data provided by WHR to conduct his own analysis. That means, he does not repeat what was conducted in WHR before because his ultimate purpose is different from WHR. Primarily, this project aims at *finding latent relationships among factors used and trying to explain the characteristics of happiness defined by the citizens over the world,* instead of ranking countries as what WHR did. Secondly, the focus of WHR also changes year by year - such as, in 2018 WHR focused on international migration, in 2019 it is the effect of technologies on lives; thus, their variables are more complex. The author only picks a subset of data from WHR to use in this project. Therefore, the findings of this analysis might also differ from WHR's.

2. Research Questions

Concretely, this analysis aims at answering the following questions:

(1) To which extent are the independent life metrics correlated to each other?
(2) Is it possible to explain the variances in the metrics with fewer components? If yes, what could those components be and how much could they explain the variances?
(3) Do countries perform differently across continents? If yes, what are the differences?
(4) Can some countries be grouped (clustered) together?
(5) Which variables can be best used to explain *Happiness Score*?
(6) What are the characteristics of the happiest and least happy countries?

3. Theoretical Background

In this section, the author will instead discuss briefly some statistical methods used in this project that are not mentioned in the course.

Sample correlation: There are several methods to measure correlation between two variables, such as Pearson's, Spearman's or Kendell's (Bonett & Wright, 2000). While the Pearson correlation evaluates the linear relationship between two continuous variables, the

Spearman correlation evaluates the monotonic relationship between two continuous or ordinal variables. Thus, Spearman is more useful if at least one variable is not normally distributed.

Multivariate linear regression: Brooks (2008) explains a regression model in a multivariate setting. There are five assumptions for a linear regression model to meet in order to confirm its validity:

(1) *Zero error mean:* This assumption is always true if the estimation uses Ordinary Least Square method.

(2) *Homoscedasticity:* The variance of model errors is constant across values of the dependent variables. Breusch-Pagan test can be used to test this assumption with the null hypothesis supporting the homoscedasticity of model residuals. Otherwise, the residuals are deemed heteroscedasticity.

(3) *No autocorrelation:* There should be no correlation among U. Durbin-Watson test can be used for testing first-order autocorrelation, and Breusch-Godfrey can be used for higher order autocorrelation. This symptom normally exists in time series data, which is not our case. Yet, it is also good to check if this issue exists. The null hypothesis of these two tests is that there is no autocorrelation among the residuals.

(4) *No multicollinearity:* There should be no correlation between residuals and independent variables. For a given predictor, multicollinearity can be assessed by computing a score called the Variance Inflation Factor (VIF), measuring how much the variance of a regression coefficient is inflated due to multicollinearity in the model. As a rule of thumb, a VIF value that exceeds 10 indicates a problematic amount of collinearity (Gareth, et al., 2014).

(5) *Normality:* The residuals are assumed to follow a normal distribution. Normality can be checked with Jarque-Bera or Shapiro-Wilk test, whose null hypothesis is that the variable is normally distributed.

The above-mentioned null hypotheses are accepted if their statistical *p-value* is higher than the *alpha level of 0.05*.

4. Data Description

The dataset is primarily obtained from *WHR 2018 Report* (United Nations, 2018). In its original report, WHR has a total of 14 variables. In this project report, the author only considers 07 variables of his interest. Details of those chosen metrics are as following:

(1) ***GDP per capita (USD)****:* This metric compares GDP on a purchasing power parity basis divided by population. Data is extracted from the World Development Indicators (WDI) released by World Bank in 2017.

(2) ***Social support***: It shows the national average of the binary responses (either 0 or 1) to the question "*If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them or not*?".

(3) ***Life expectancy***: This metric measures the average time an organism is expected to live, based on the year of its birth. Data is based on World Health Organization and World Development Indicators.

(4) ***Freedom to choices***: It shows the average of the binary responses to the question "*Are you satisfied or dissatisfied with your freedom to choose what you do with your life*?"

(5) ***Generosity*****:** This is the residual of regressing the answers from the questions "*Have you donated money to a charity in the past month?"* on GDP per capita.

(6) ***Perception of corruption***: It shows the national average of the binary responses to the questions "*Is corruption widespread throughout the government or not?*" and "*Is corruption widespread within business or not?*"

(7) ***Happiness score***: This metric measures the national average answer to the Cantril ladder question, asking people to evaluate the quality of their current lives on a scale of 0 to 10.

These 07 variables are all numeric and continuous. Among them, the first 06 variables are independent factors that evaluate quality of life. Meanwhile, the last one is the dependent metric that is used for final ranking in WHR. Thus, it is interesting to explore the relationship between those 06 independent factors and their impact on the happiness score. In addition, the author thinks it might be interesting to include the ***(8) Continent*** in order to inspect the relationship amongst these factors within different cluster groups.

Due to data unavailability of some countries, there are only 131 countries in this analysis in total. ***Table 1*** shows the sample view of the final dataset.

Table 1: Sample of the final dataset (5 first countries, in alphabet order)

Country	Continent	GDP per Capita (USD)	Social Support	Life Expectancy	Freedom to Choices	Generosity	Perception of Corruption	Happiness Score
Afghanistan	Asia	1737.397	0.491	52.340	0.427	-0.106	0.954	2.662
Albania	Europe	11774.816	0.638	69.052	0.750	-0.035	0.876	4.640
Algeria	Africa	13908.342	0.807	65.699	0.437	-0.195	0.700	5.249
Argentina	South America	18835.887	0.907	67.539	0.832	-0.186	0.841	6.039
Armenia	Asia	8389.288	0.698	65.126	0.614	-0.132	0.865	4.288

5. Data Analysis

5.1. Univariate Analysis

In this part, each variable is analyzed separately. To begin with, ***Figure 1*** shows the histogram of 07 numerical variables and 01 categorical variables, and ***Table 2*** represents descriptive statistical figures of each numerical attribute.

The histograms below show one fact that except from *Happiness Score*, all independent variables are non-normally distributed, because the shapes do not resemble a normal distribution curve. This conclusion can be confirmed by Jarque-Bera Test applied on each variable. Apart from *Happiness Score*, all of others have their values of much less than 0.05, preventing us from accepting the null hypothesis. This result will affect which correlation method to use in bivariate analyses consequently. Especially, the data is highly skewed in the case of *GDP per Capita* and *Perception of Corruption*. There are various methods to normalize data, such as taking square root or logarithms, etc.; however, in the scope of this project, the author does not perform data transformation, as it is relatively effortful.

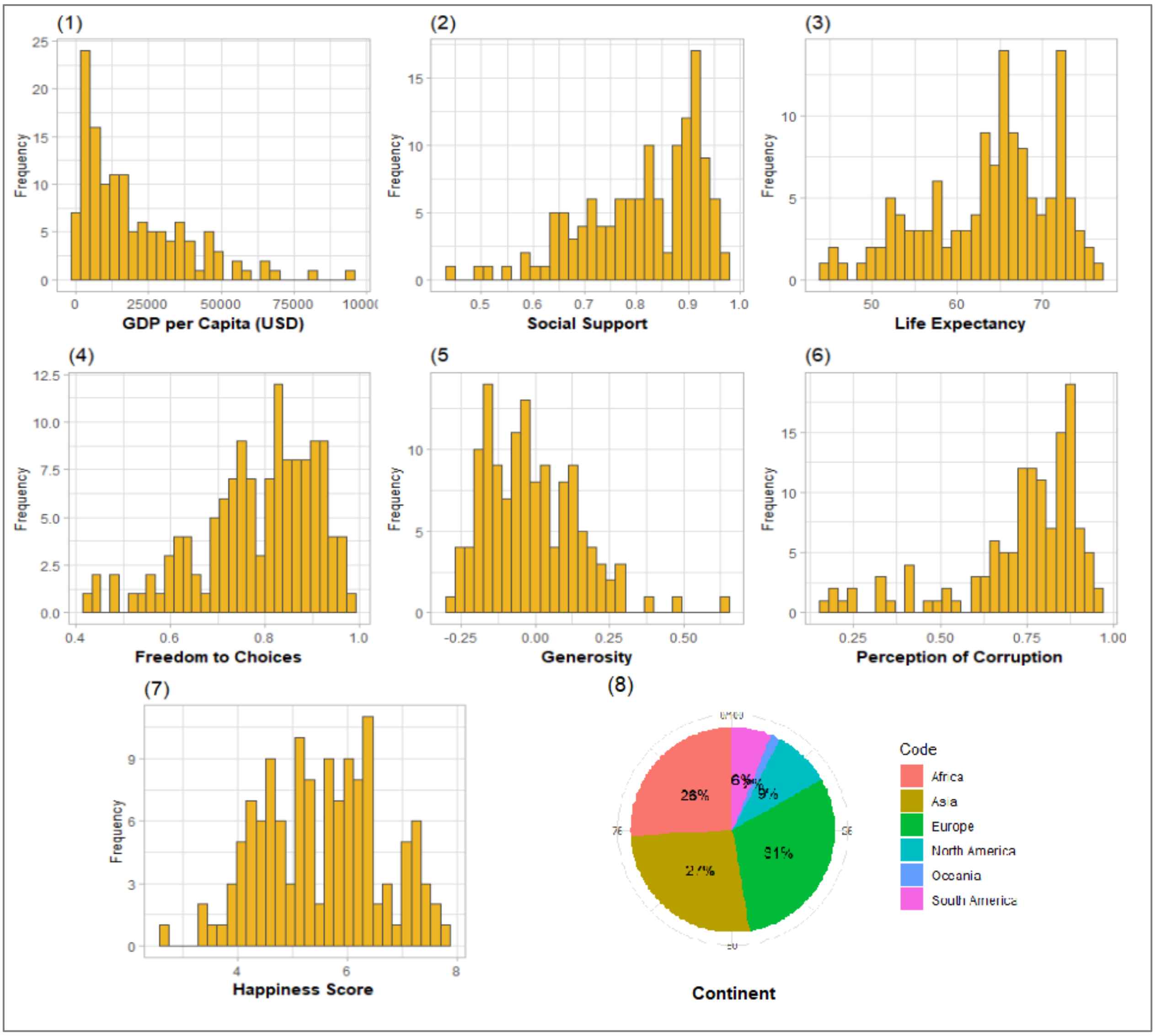


Figure 1: Univariate histograms of (1) *GDP per Capita*, (2) *Social Support*, (3) *Life Expectancy*, (4) *Freedom to Choices*, (5) *Generosity*, (6) *Perception of Corruption*, (7) *Happiness Score*, and Pie chart for *Continent* (8)

From the pie chart, one can say that most countries in this analysis are from Europe, Asia and Africa. In contrast, Oceania accounts for approximately 2% of the dataset (*n = 2 countries*).

Table 2: Descriptive Statistical Figures of 07 Numerical Variables

Metric	GDP per Capita (USD)	Social Support	Life Expectancy	Freedom to Choices	Generosity	Perception of Corruption	Happiness Score
Min	757.038	0.436	44.387	0.427	-0.297	0.162	2.662
Max	95342.50	0.967	76.536	0.985	0.629	0.954	7.788
Median	13908.34	0.831	65.379	0.814	-0.035	0.781	5.594
Mean	19517.62	0.811	63.795	0.783	-0.015	0.732	5.532
Standard deviation	18483.79	0.114	7.453	0.125	0.160	0.181	1.088
Skewness	1.422	-0.854	-0.599	-0.832	0.901	-1.526	-0.004
p-value for Jarque Bera	3.997×10^{-15}	3.896×10^{-4}	0.0128	5.633×10^{-4}	2.378×10^{-6}	5.995×10^{-15}	0.3426

In addition, their data scales also differ significantly. For instance, *Social Support, Perception of Corruption* and *Freedom to Choices* have a scale of 0.0 to 1.0, *Life Expectancy* has a scale of 44 to 77, while that of *GDP per Capita* is up to approximately 95,342. This acknowledges that data should be standardized if the author wishes to perform any specific exploratory analysis.

5.2. Bivariate Analysis

In this part, the author would like to present the relationship among 06 independent numeric variables. The relationship between the dependent variable *Happiness Score* and other independent variables will be discussed further in section 5.3.3 Multivariate Linear Regression. ***Figure 2*** below provides a comprehensive outlook on the relationship of each pair, grouping by *Continent* variable through colors.

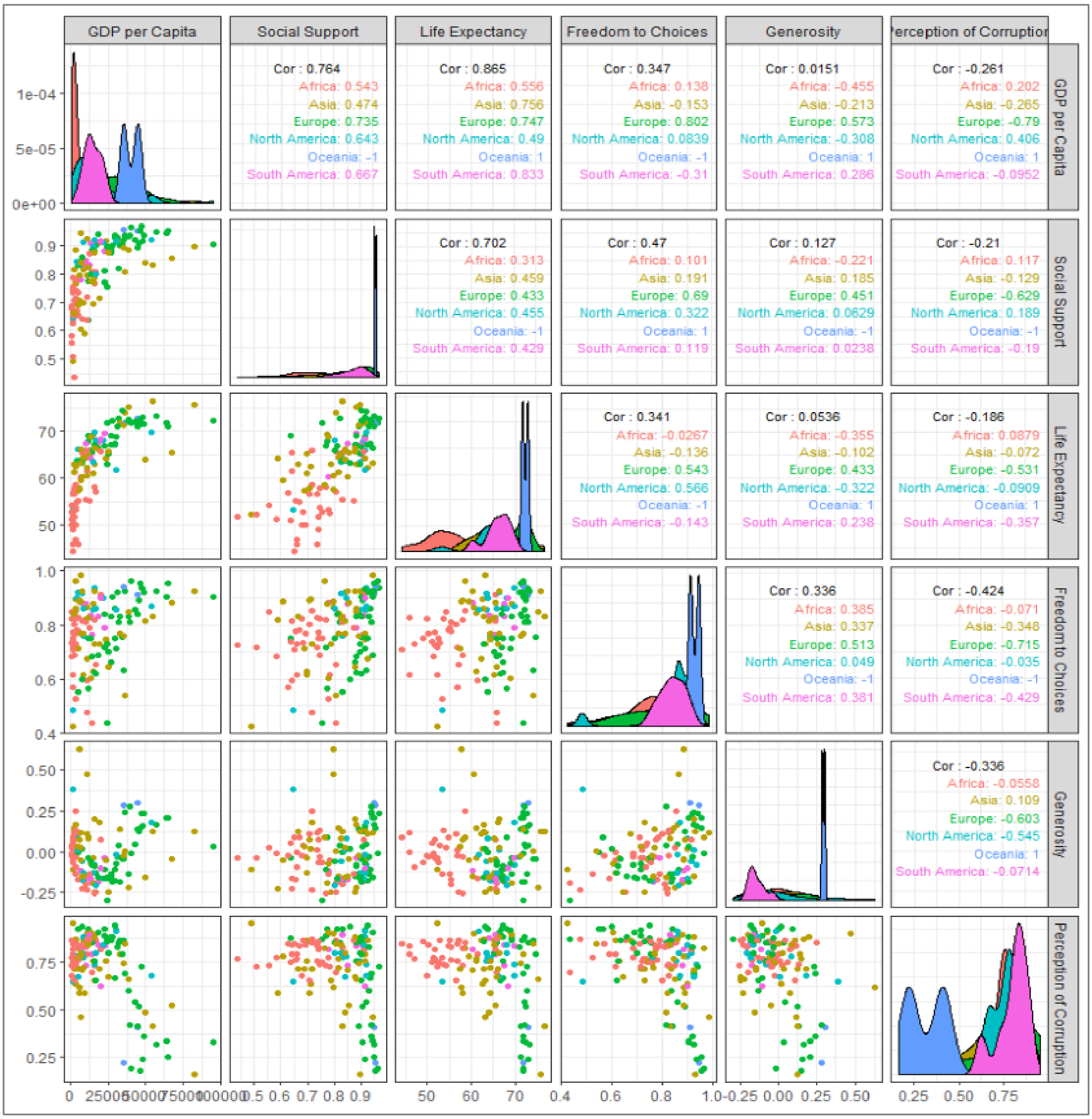


Figure 2: Pair plot for 06 independent factors

Histogram plots in ***Figure 2***'s diagonal elaborate those presented in ***Figure 1*** by separating into distinct continents. The upper right boxes provide Spearman correlation coefficients, and the lower left boxes shows pair scatter plots. In Section 5.1, the author learnt that the dataset is non-normally distributed and highly skewed. Therefore, *Spearman correlation* should be used to describe the relationship between each pair of variables using a monotonic function.

It can be seen that there are high correlations between *(1) GDP per Capita - Life Expectancy*, *(2) GDP per Capita - Social Support* and *(3) Social Support - Life Expectancy*. The coefficients of these 03 pairs are all positive, suggesting that the covariances among them can be explained with fewer variables. Particularly, in the case of *(1)*, we can see the a near-linear relationship between *GDP per Capita* and *Life Expectancy*. Altogether, the data shows that the wealthier a country is, the more likely the people live longer and receive more support from the society, which makes sense to the author completely. In addition, *Perception of Corruption* always has negative coefficients, presuming that the higher perception of corruption in a country is, the less likely that country does well in other life factors.

Continent-wise, apart from the other 03 mentioned pairs, it is worth pointing out that in Europe, most correlation coefficients have significant values, indicating latent relationship among those factors. Oceania has only $n = 2$ countries, and according to its absolute correlation coefficients, it seems that these 2 countries always variates in the same manner.

5.3. Multivariate Analysis

There are several methods to analyze a multivariate continuous-numeric dataset. In this project, the author answers the research questions mainly by three methods: *Principle Component Analysis (PCA)*, *k-mean Clustering* and *Multivariate Linear Regression*.

5.3.1. Principle Component Analysis

As explained above, there are strong correlation among some factors, thus the author believes that 06 independent variables could be explained in a better way via PCA. Furthermore, the author also recalls the finding from previous step, which is that the dataset does not share the same scale among variables. Thus, this activity is conducted by function *princomp* in R, using correlation matrix instead of covariance matrix. ***Figure 3*** shows the scree plot for the variances explained by each principle component (PC).

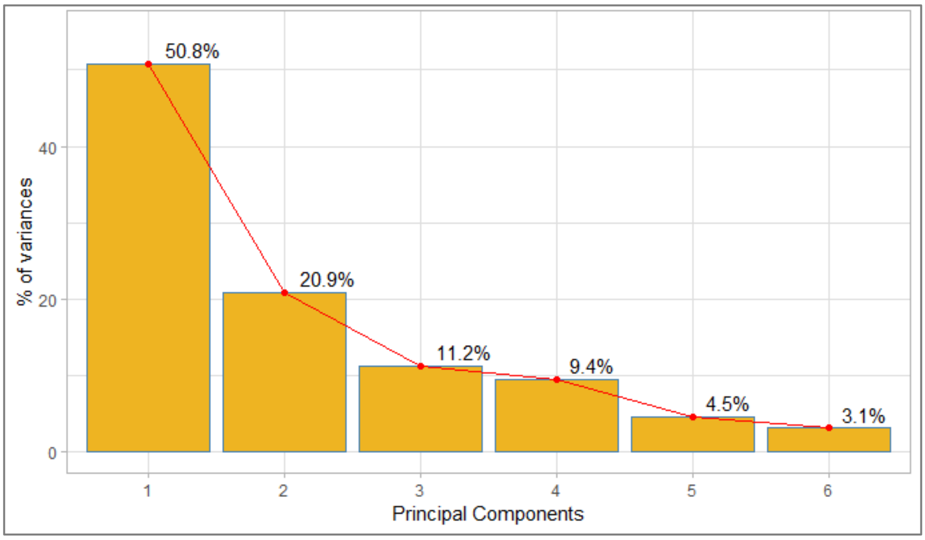


Figure 3: Variances explained by each Principle Component

As a rule of thumb, we should choose as many components as needed to explain at least 90% of total variance. In this case, the author chooses the first four components as they can explain up to **92.4%** of the total variances contained in the dataset. ***Figure 4*** displays score plots two pairs of PC: the *first two PCs* and *the third and forth PCs.*

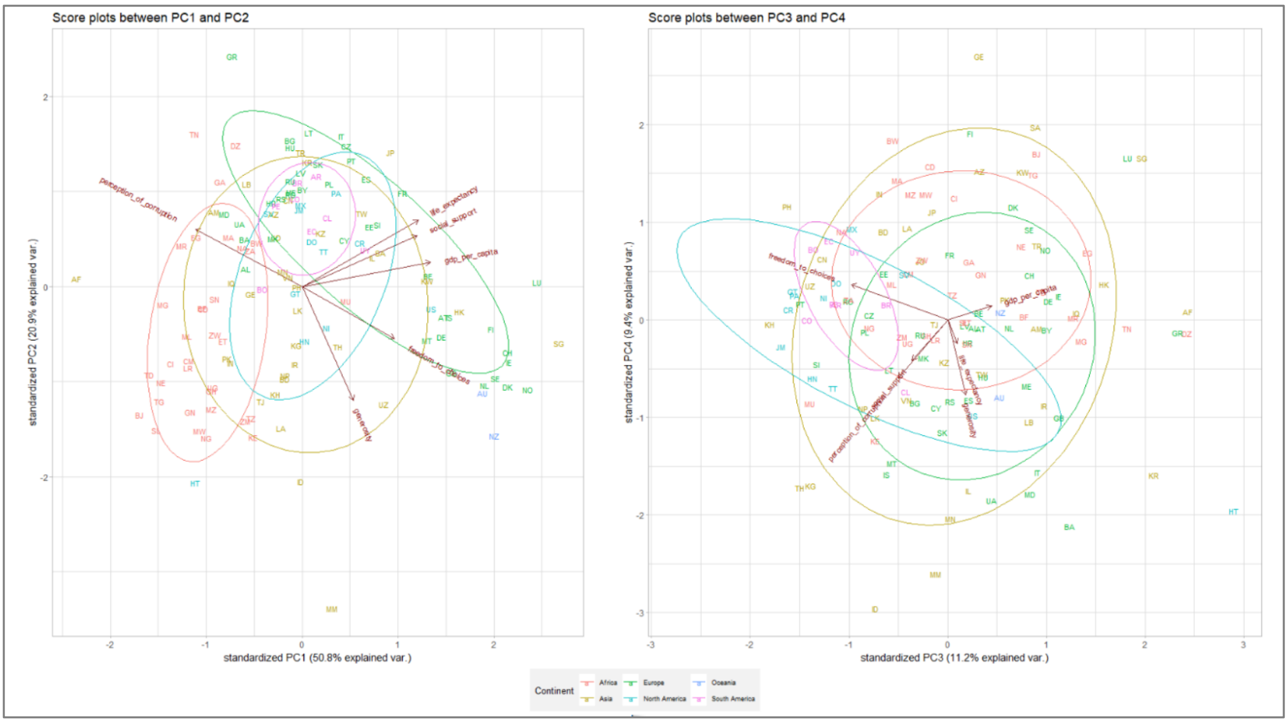


Figure 4: PC Score and Loading Plots. *See Appendix 1 for better resolution.*

In this figure, each country is plotted by its country code with regard to its score in PCs. The arrows show the loadings of each PC, unique up to sign. Color grouping is again used to inspect the differences among continents and the circle is used to emphasize this grouping effectively. The first PC has relatively high values in *GDP per Capita*, *Social Support*, *Life Expectancy*, *Freedom to Choices*, *Generosity* (highest in *GDP per Capita*) and negative value in *Perception of Corruption.* As a citizen, one will feel enjoyable if he is wealthy and generous, able to live long, has freedom to make decisions, receives support from others and has no corruption in government. This finding also matches with what was concluded earlier through pair plots. Thus, to the author's opinion, this component seemingly explains the ***Well-Being of Citizens***, and this is the most important PC.

The second component is somewhat different: It has high positive loading values in *Perception of Corruption, Life Expectancy,* and high negative values *Freedom to Choices*, *Generosity* (highest values in *Life Expectancy* and *Generosity* respectively). Thus, this PC can be seen as an explanation for the ***Inequality in Society***. That is, when there are distinct human classes in a society – meaning in that society, there are gaps among rich and poor people, they are less caring about each other, and there is less freedom in low-class group – then there might be a chance that society is perceived as highly corrupted, the citizens are somewhat neutral in generosity and the overall income is the average of the rich and poor.

The third component is fairly tricky, because it has high negative loading values in contradictory factors between *Freedom to choices, Social Support* and *Perception of Corruption,* and high positive loading value in *GDP per Capita.* Thus, the author assumes this component represents the ***Belief of Citizens in the Society***. In other words, even though the citizens have no freedom at all and low social support, they still have positive believes in the governing party, and thus have relative high *GDP per Capita*.

The last PC is also hard to interpret. It has high negative loading values in *Generosity*, *Perception of Corruption* and *Social Support*, and positive loading value in *Freedom to Choices*. That is, this component seemingly gives high values to countries whose citizens are introverted, independent and somewhat do not care very much about what others are doing but only themselves. Due to this assumption, their perception of corruption is consequently low. Thus, the author identifies this PC as the ***Autonomy of Citizens***.

Based on the interpretation of those four PCs, one can further highlight some *continent-wise* conclusions in 03 continents with highest shares in this analysis:

- ***Africa***: One can see that all African countries have low values in PC1 - meaning that the well-being condition of their citizens is not good. In addition, the majority of African countries have negative values in PC2 – meaning that the inequality does not exist greatly in the society. Plus, the majority of them have high values in PC4 – meaning that the people there are relatively independent and autonomous.
- ***Asia*:** This continent is quite mixed. The majority of them has around-center values in PC1 – meaning that they have an average well-being condition, except some noticeable outliers such as Singapore, Hong Kong or Kuwait. In PC2, there are some equal countries such as Myanmar or Indonesia, and some very unequal are Japan or Turkey.
- ***Europe*:** It is quite progressive here in Europe when many countries stay at an above-average level for the Well-being PC. Yet, there seems to be two distinct groups here: One including Western and Northern countries has extremely high PC1, and the other one including Southern and Eastern countries has relatively lower values. The first group also have lower PC2 values – meaning that they have less inequality in society.

5.3.2. k-Mean Clustering

k-Mean Clustering by function *clusplot* in R is an advancement of PCA, because this method clusters countries into different groups based on *distances from the "centers"*, plotted with regard to the first two PCs that we have discussed in 5.3.1. Since the author does not have any particular preference for *k*, he eventually chose *k = 2, 3, 4* as it is sufficient to visualize in this paperwork report. **Figure 5** below shows the clustering results.

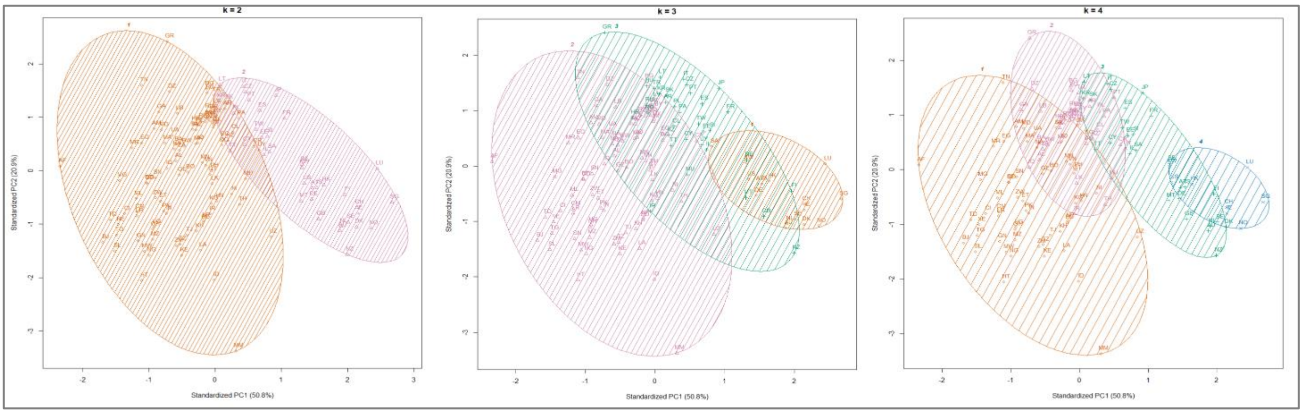


Figure 5: k-Mean Clustering *(k = 2,3,4). See Appendix 2 for better resolution.*

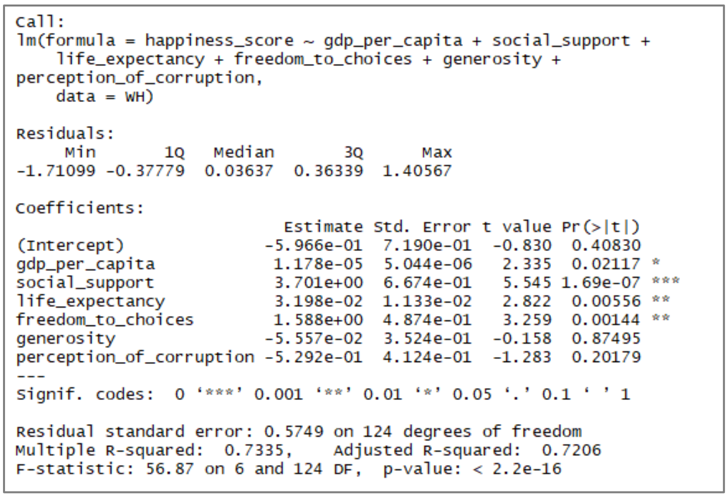
Let us recall that x-axis (PC1) represents ***Well-being of the Citizens*** and y-axis (PC2) shows ***Inequality in Society***. The results can be interpreted as below:

- ***k = 2***: It seems to cluster based on PC1 only. The results are classified into two groups (with some overlapping): Group 1 consists of countries with low well-being, while Group 2 contains all countries with positive well-being scores.
- ***k = 3:*** This clustering seems to be an extension of *k = 2*. While Group 1 in *k = 2* is almost the same as in *k = 3*, Group 2 in *k = 2* is now divided into 2 smaller group in *k = 3*: One group contains high inequality but average well-being, and one small group has very high well-being and relatively low values for PC2.
- ***k = 4:*** This clustering is somewhat different from *k = 2, 3*. Firstly, on the far right, there is one group of "*elite countries*", such as Singapore, Luxemburg, Norway, Switzerland, etc. with very high well-being and relatively low values for inequality. The second group consists of developed countries with high well-being of citizens, but the inequality gap in societies is still high, such as Japan, France or United Kingdom. The third group is comprised of countries with average well-being but high inequality in society, such as Poland, Belarus or Argentina. The last group consists of countries having low well-being and low inequality in society, such as Laos, Myanmar or Nigeria.

From the interpretation above, the author suggests using ***k = 4*** in this analysis, because it seems to capture all variances of both PC1 and PC2.

5.3.3. Multivariate Linear Regression

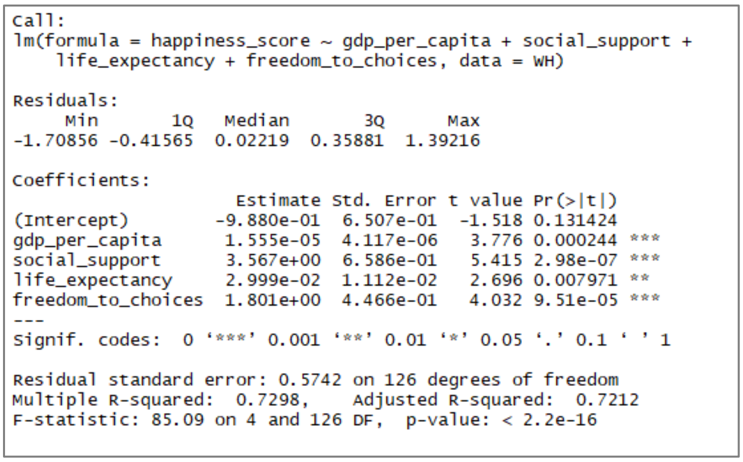
In previous section, the author explores the relationship among independent variables though expressing them in 04 different PCs. In this section, the author tries to understand which factors and to which extend they affect *Happiness Score*. To do so, the author firstly performs multivariate linear regression between *Happiness Score* and the other 06 independent variables as described in ***Figure 6***.



```
Call:
lm(formula = happiness_score ~ gdp_per_capita + social_support + 
    life_expectancy + freedom_to_choices + generosity + 
perception_of_corruption, 
    data = WH)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.71099 -0.37779  0.03637  0.36339  1.40567 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -5.966e-01  7.190e-01  -0.830  0.40830    
gdp_per_capita           1.178e-05  5.044e-06   2.335  0.02117 *  
social_support           3.701e+00  6.674e-01   5.545 1.69e-07 ***
life_expectancy          3.198e-02  1.133e-02   2.822  0.00556 ** 
freedom_to_choices       1.588e+00  4.874e-01   3.259  0.00144 ** 
generosity              -5.557e-02  3.524e-01  -0.158  0.87495    
perception_of_corruption -5.292e-01  4.124e-01  -1.283  0.20179    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5749 on 124 degrees of freedom
Multiple R-squared:  0.7335,	Adjusted R-squared:  0.7206 
F-statistic: 56.87 on 6 and 124 DF,  p-value: < 2.2e-16
```

Figure 6: Regression Modeling of Happiness Score with all Factors

Overall *p-value* is approximately zero, letting the model be accepted. Among 06 variables, *Generosity* and *Perception of Corruption* have *p-value* > 0.05, suggesting removing them out of the model. The author re-runs the regression without two insignificant variables in ***Figure 7***.



```
Call:
lm(formula = happiness_score ~ gdp_per_capita + social_support +
    life_expectancy + freedom_to_choices, data = WH)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70856 -0.41565  0.02219  0.35881  1.39216

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -9.880e-01  6.507e-01  -1.518 0.131424
gdp_per_capita     1.555e-05  4.117e-06   3.776 0.000244 ***
social_support     3.567e+00  6.586e-01   5.415 2.98e-07 ***
life_expectancy    2.999e-02  1.112e-02   2.696 0.007971 **
freedom_to_choices 1.801e+00  4.466e-01   4.032 9.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5742 on 126 degrees of freedom
Multiple R-squared:  0.7298,	Adjusted R-squared:  0.7212
F-statistic: 85.09 on 4 and 126 DF,  p-value: < 2.2e-16
```

Figure 7: Adjusted Modeling of Happiness Score with significant Factors

The adjusted model confirms that all variables are now significant. *p-value* remains approximately zero and Adjusted R-squared increases. In order to confirm the validity of the model, diagnostic tests need to be performed. ***Table 3*** shows the results for the diagnostic test discussed earlier in Section 2 for the Adjusted Model.

Table 3: Results of Diagnostic Tests

Diagnostic Test	Results
Residual Check	Sum and mean of residuals are 4.92 x 10^{-16} and 3.76 x 10^{-18}.
*Breusch-Pagan Test for **Homoscedasticity***	*p-value* of probability higher than Chi-Squared is 0.058.
*Durbin-Watson Test for **Autocorrelation***	Test *p-value* = 0.169.
*VIF for **Multicollinearity***	*GDP per Capita*: 2.28; *Social Support*: 2.23; *Life Expectancy*: 2.71; *Freedom to Choices*: 1.23.
*Shapiro-Wilk test for **Normality***	Test *p-value* = 0.695

Based on the results above, the adjusted model is accepted with the following explanations:

- *Zero error mean*: We can see that sum and mean of residuals are 4.92 x 10^{-16} and 3.76 x 10^{-18} respectively, which confirms this assumption is true.
- *Homoscedasticity*: Since *p-value* > 0.05, the null hypothesis is accepted. That is, variance of residuals from our model is constant across the values of happiness score. Requirement for homoscedasticity is met.
- *No autocorrelation: p-value* of Durbin-Watson test is higher than 0.05, we also accept the null hypothesis. That is, there is no first-order autocorrelation among residuals.
- *No multicollinearity:* VIFs from our model is relatively low (much less than 10), meaning that there is no severe multicollinearity among presented factors.
- *Normality*: H_o is accepted because *p-value* is much higher than 0.05.

Interested audience can also review diagnostic plots in **Appendix 3** for further proof. In short, *Normal Q-Q plot* is a near-straight line, implying normally-distributed residuals. *Residuals vs Fitted* shows a non-linear pattern between residuals and fitted values, implying all non-linear relationships are explained by the model.

In conclusion, *Happiness Score* can be explained by four factors: *GDP per Capita, Social Support, Life Expectancy* and *Freedom to Choices*. All of these have positive coefficients, showing positive support to *Happiness Score*. The author does not explain the weight of coefficients since factors have different scales, solving it would exceed his available resources.

6. Conclusion

In summary, the analysis attempted to answer all related questions from Section 2.

(1) *To which extent are these factors correlated to each other?*

In multivariate and bivariate analyses, we explored the distribution of individual variables and the relationship among each pair. We have concluded that our variables are skewed, and that *GDP per Capita*, *Social Support* and *Life Expectancy* are monotonically correlated to each other. Among continents, Europe has a strong behavior in correlation among factors.

(2) *Is it possible to explain the variances in those metrics with fewer components? If yes, what could those components be and how much could they explain the variances?*

One focus of this analysis is to express factor variances through fewer components. PCA proves that this objective can totally be done with standardized data. With four PCs, 92.4% of total variances in the original dataset can be explained. The author also tried to interpret those PCs; in a decreasing order of importance, they are ***Well-Being of Citizens, Inequality in Society, Belief of Citizens in the Society*** and ***Autonomy of Citizens***.

(3) *Do countries in different continents perform differently? If yes, what are the differences?*

Based on those PCs, the author spotted some distinct behavior across continents. Africa has low performance in well-being condition, but the people seem to be independent and equal in society. In contrast, Europe, where is traditionally famous for its liberty and wealth, performs well; yet, there are still gaps within the continent itself. Clouding relatively in the centers, Asian countries have average performance in most PCs.

(4) *Can some countries be grouped (clustered) together?*

One certain way to group countries is based on *Continent* as explained above. Additionally, k-mean clustering offers another method to group countries based on distance calculation of PC1 and PC2 across countries. With experiments on different *k* values, the author suggests a grouping option with *k = 4*. Since Well-being PC still accounts for the highest proportion (50.8%), it holds the most relevance in the clustering result.

(5) *Which variables can be best used to explain Happiness Score?*

The author also experimented regression models between *Happiness Score* and six independent variables. According to the Adjusted Model from Section 5.3.2, four variables *GDP per Capita, Social Support, Life Expectancy* and *Freedom to Choices* are the most significant ones, and they all have positive effects on Happiness Score.

(6) *What are the characteristics of happiest and least happy countries?*

Finally, let us relate the whole analysis to the initial motivation. Finland, as mentioned in Section 1, is the happiest country in the world (highest *Happiness Score*), followed by Denmark and Norway. On the other extreme, the least happy countries (lowest *Happines Score)* are Afghanistan, Tanzania and Malawi. **Table 4** presents the score of these countries on four principle components.

Table 4: Scores of Happiest and Least Happy Countries

	Country	PC1	PC2	PC3	PC4
Happiest Countries	Finland	3.425	-0.504	0.187	1.434
	Denmark	3.742	-1.183	0.541	0.870
	Norway	4.117	-1.213	0.816	0.541
Least Happy Countries	Afghanistan	-4.101	0.095	1.993	0.065
	Tanzania	-0.929	-1.551	0.045	0.193
	Malawi	-1.867	-1.701	-0.183	0.966

It can be seen that the strongest distinction between these two groups lies upon PC1 (***Well-being of the Citizens***). The happiest countries have very high level of well-being condition, while the least happy countries are on the opposite. Unfortunately, other PCs do not show any clear distinction between these two groups, so no further conclusion can be made.

Therefore, for **self-reflecting** purposes, this analysis awakens one truth in the author's mind: *Well-being - the state of being comfortable and healthy - is the foundation of happiness.*

7. Evaluation

There are some concerns regarding to this analysis as below:

- Firstly, our univariate data is highly skewed and not normally distributed. PCA does not strictly require normality if the purpose is only to do exploratory analysis, which is our case. However, ideally, one should consider *Independent Component Analysis* method or try to normalize the dataset first in order to improve the robustness and readability of results. Additionally, data transformation is highly recommended, such as taking square root or logarithms, to make univariate less skewed. Univariate data should also be standardized before the linear regression section, so that the standardized coefficients could be more explainable.
- The interpretation of PCs is rigorous and subjective. The author has tried his best to intuitively explain the PC in the most reasonable way. However, as the author does not have sufficient knowledge in social science-related topics to completely understand the meaning behind those PCs, the interpretation might not be fully correct.
- The multivariate linear regression can also be implemented on four PCs instead of all independent variables. The correlation amongst independent variables might make the regression model unreliable.

At the end, these concerns do not invalidate the results of this analysis. However, it is important to keep these in mind when reading the results.

Lastly, this report still has room for improvement in the following aspects:

- Data standardization and normalization as explained above.
- Adding more life factors that can show more angles in life, such as weather, demographic, entertainment offers, etc.
- Extending to more multivariate methods, such as Discriminant Analysis or Independent Component Analysis, etc.

References

Bonett, D. & Wright, T., 2000. Sample size requirements for Pearson, Kendall, and Spearman correlations. *Psychometrika,* Issue 65, pp. 23 - 28.

Brooks, C., 2008. *Introductory Econometrics for Finance.* ISBN-13: 9780521694681 ed. Cambridge: Cambridge University Press.

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R., 2014. *An Introduction to Statistical Learning: With Applications in R.* s.l.:Springer Publishing Company, Incorporated..

Smith, K., 2017. *That world happiness survey is complete crap.* [Online]
Available at: https://flowingdata.com/2012/04/25/world-happiness-report-makes-statisticians-unhappy/
[Accessed 21 3 2019].

Statistic Finland, 2018. *Finland among the best in the world.* [Online]
Available at: http://www.stat.fi/tup/satavuotias-suomi/suomi-maailman-karjessa_en.html
[Accessed 21 3 2019].

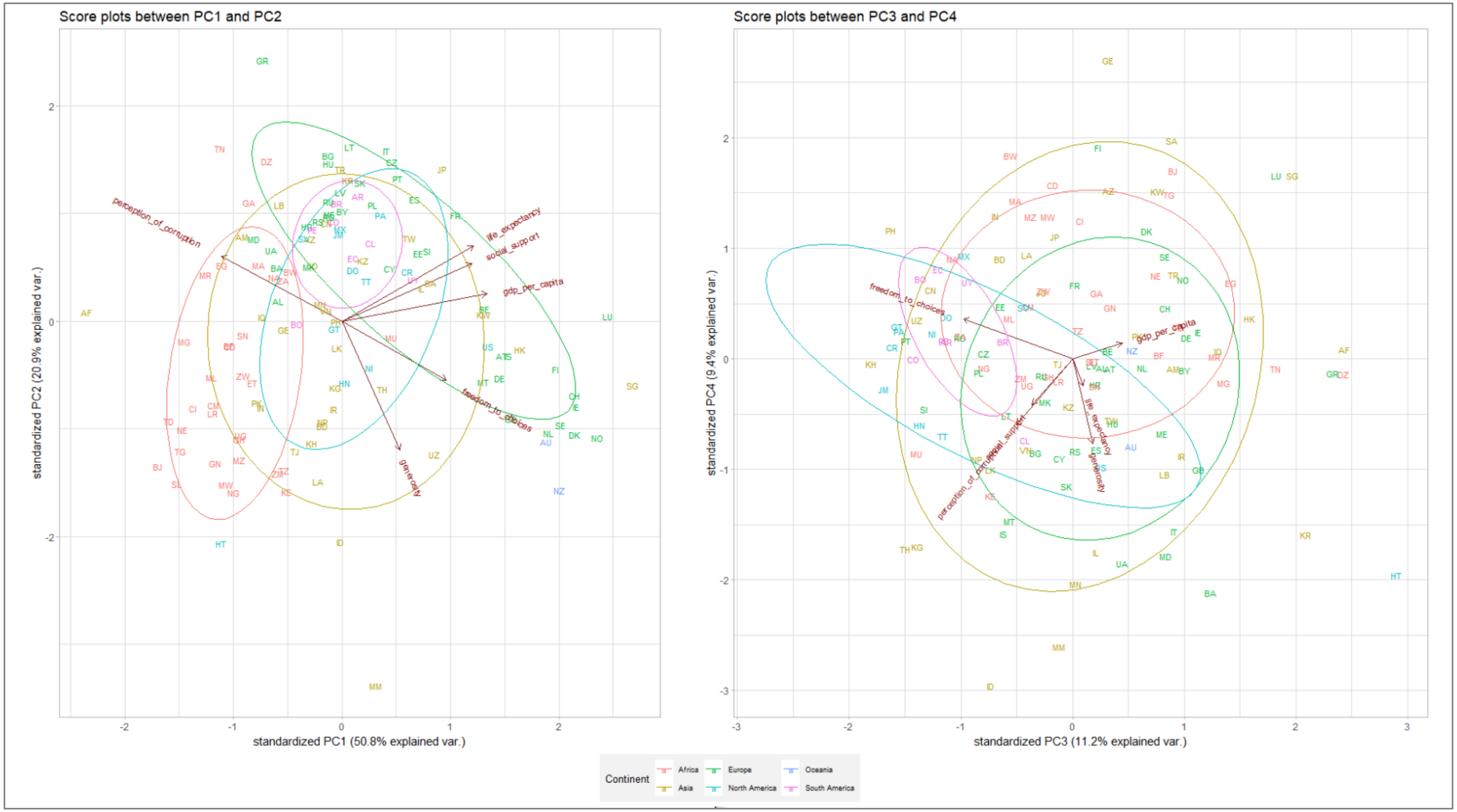
United Nations, 2018. *World Happiness Report,* s.l.: United Nations Sustainable Development Solutions Network.

YLE, 2019. *Finland: Still the happiest country in the world (says UN report).* [Online]
Available at:
https://yle.fi/uutiset/osasto/news/finland_still_the_happiest_country_in_the_world_says_un_report/10698146
[Accessed 21 3 2019].

Appendix

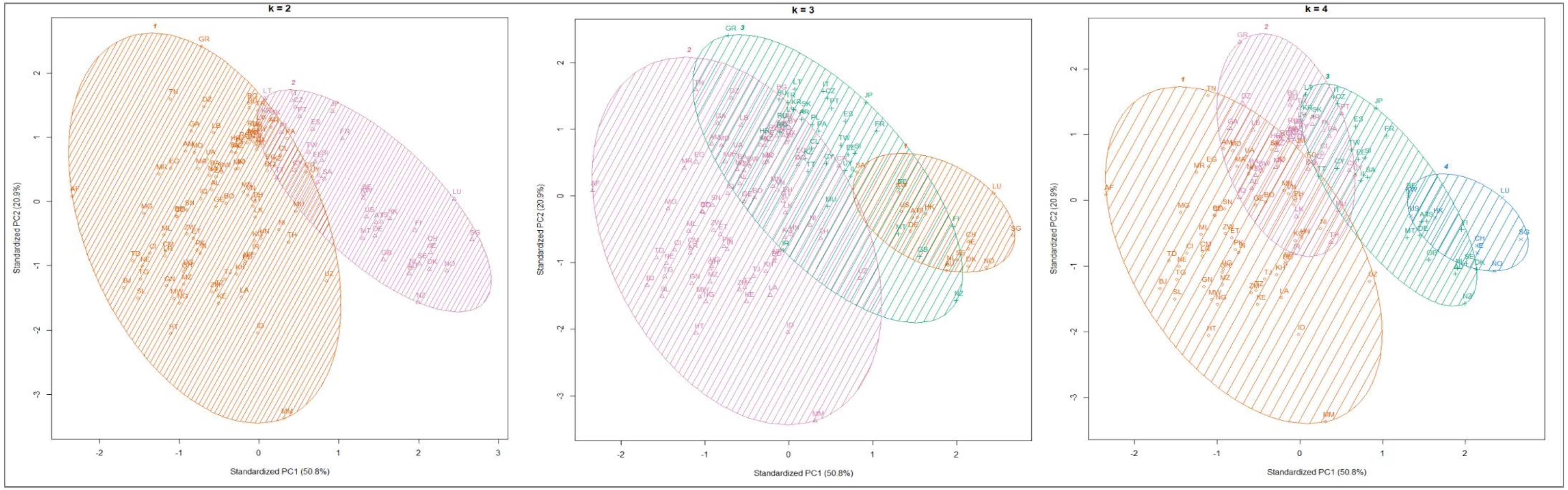
Appendix 1

PC Score and Loading Plots (Full Page)



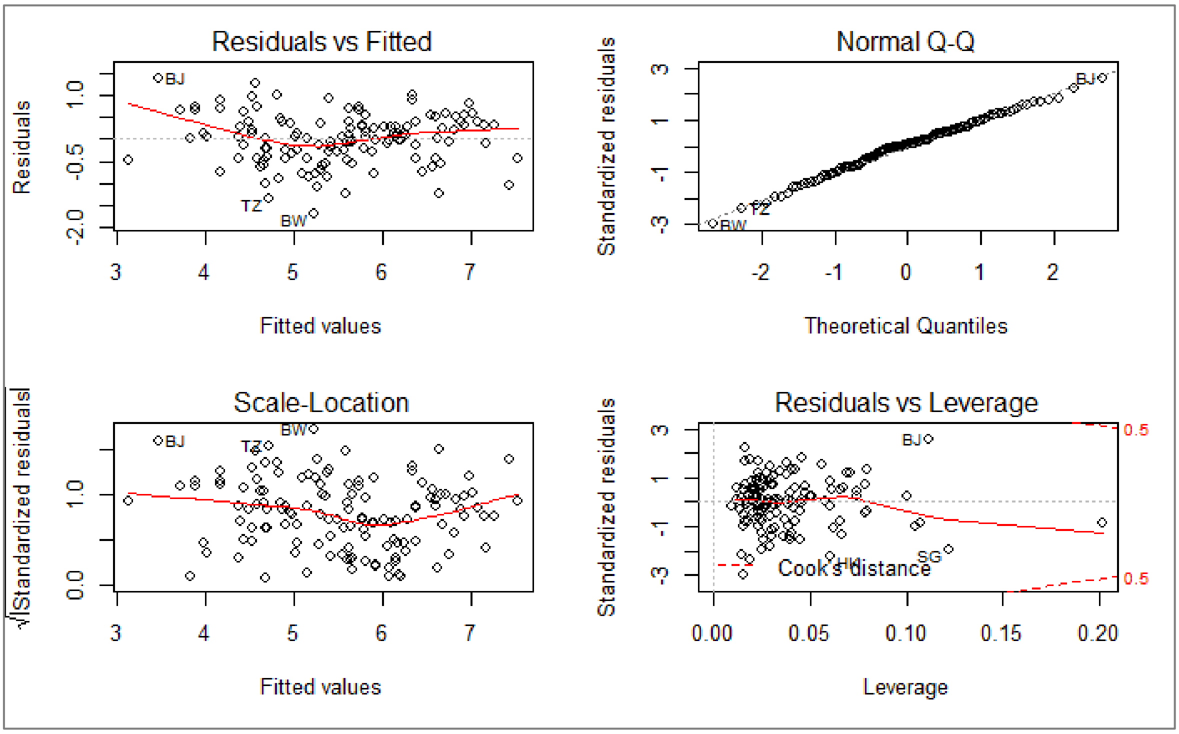
Appendix 2

k-Mean Clustering *(k = 2,3,4)* (Full Page)



Appendix 3

Diagnostic Plots for the Adjusted Regression Model


Residuals vs Fitted
Residuals
1.0
-0.5
-2.0
BJ
TZ
BW
3
4
5
6
7
Fitted values
Normal Q-Q
Standardized residuals
3
1
-1
-3
BJ
TZ
BW
-2
-1
0
1
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
1.0
0.0
BJ
TZ
BW
3
4
5
6
7
Fitted values
Residuals vs Leverage
Standardized residuals
3
1
-1
-3
BJ
HK
SG
Cook's distance
0.5
0.5
0.00
0.05
0.10
0.15
0.20
Leverage