Katholieke
Universiteit
Leuven

**Faculty of
Science**

# REPORT OF STATISTICAL ANALYSIS

Happiness Data set of 2016

**Eugene Ambe Ndamukong (r0417710)**
**Master of Statistics**
**Course of Statistical Consulting**

Academic year 2016–2017

# Contents

# 1 Introduction

This report gives a summary of the statistical methods which were used to investigate 2 research questions of the study on *Happiness* in several countries. The research questions included:

- Using the variables *Economy, Health, Family, Freedom,Trust, Generosity*, how can the countries be grouped into homogeneous groups?
- Using the same variables, make prediction of the *Happiness score*?

It was hypothesized that there are no homogeneous groups in the data. In regards to the two research questions, the analysis was done in 2 separate sections.

# 2 Happiness data set

The data set used for this analysis consisted of information on the wellbeing(*Happiness*) of residents in various countries. This information was collected on a survey which was executed in the year 2016 in 157 countries. The characteristics or variables which were recorded and were of interest to this analysis included: *Economy, Health, Family, Freedom,Trust* and *Generosity*. But prior to grouping the countries into homogeneous groups, the data was checked for already existing structures. The existence of potential grouping structures in the data using these 6 variables, gives an idea of how many optimal groups can possibly result from a grouping algorithm. The visualization of the data was done through a scatter plot(Appendix A) and biplot (Appendix B) using the six stated variables. Each data point in both plots represented a country. There didn't seem to be any distinct or clear grouping structures in the data especially using the biplot. However, the scatter plot seems to show very poor grouping of the data points.

# 3 Methods

Base on the first research question of the analysis, the data was clustered using 4 agglomerative clustering analysis and further with a non-hierarchical clustering algorithm. The cluster variable means of the most optimal algorithm were used as initial centers for the non-hierarchical clustering algorithm. The grouping of the countries was based on smallest euclidean distance possible among them as measure of similarity. The idea of using a subsequent clustering analysis after the previous was to optimize the grouping of the countries and ensure the homogeneity of the final clusters. Also, in regards to the second research question, a regression model was built to make a prediction of the *Happiness score*.

## 3.1 Clustering Analysis

Considering that there were no predefined groups, clustering algorithms were preferred to group the countries based on how close these countries were in 6 dimensional space. The algorithm was stopped based on a criterion which indicated how homogeneous the final clusters were. SAS (version 9.4) was used for this analysis.

### 3.1.1 Hierarchical or Agglomerative Clustering

This type of clustering involved merging clusters based on a similarity criterion i.e smallest euclidean distance. Four Hierarchical clustering methods were used in this section i.e single linkage, complete linkage, Centroid and Ward method. Among the four algorithms, the best was chosen based on which had the highest $R^2$ through the clustering history. The optimal number of hierarchical clusters were obtained from the dendrogram of the best algorithm. The optimal number clusters from the dendrogram were identified by considering the very long arms or gaps between the clusters.

### 3.1.2 Non-hierarchical Clustering

This clustering analysis involved assigning countries to consistently changing clustering centers until no further improvement of the algorithm was needed. The type of non-hierarchical clustering used here was K-means clustering. The cluster means from the Ward's method(best hierarchical clustering algorithm) were used as initial centers. The final clusters were characterized or labeled using the of the 6 variables. These labels were obtained by calculating the average means of the variables across the obtained clusters. The means of the variables which were higher than the respective average means were used as labels for the clusters.

### 3.2 Prediction Model: Linear Model

In regards to the second research question of this study, the predictions of the *happiness score* using the 6 mentioned characteristics or variables, was determined. The recorded *happiness score* ranged from 2.905 to 7.526. This was achieved using a linear model. First, 10-fold cross validation approach was used to build the model using 80% of the data to calibrate the model while 20% to test the model. Out of the 80% of the data, 60% was used to train and 20% to validate obtained by randomly splitting data in every iteration. However, the representation of the validation set changed every time the model was built and was used to select the best model for prediction. Therefore, the final model was one which was robust to changes in the data but still make accurate predictions. The finally selected model was tested on the test set and used to predict the *happiness score* on a simulated data set.

## 4 Results

### 4.1 Agglomerative Clustering

Merging of close clusters increases the within-group sum of squares error which in turn decreases the between-group sum of squares error. The between-group sum of squares is proportional to the $R^2$ and was expected to be large for a good clustering algorithm. In figure 1,the clustering history starts from 157 clusters till 1 cluster along with a decrease in the $R^2$ as the clusters become heterogeneous.

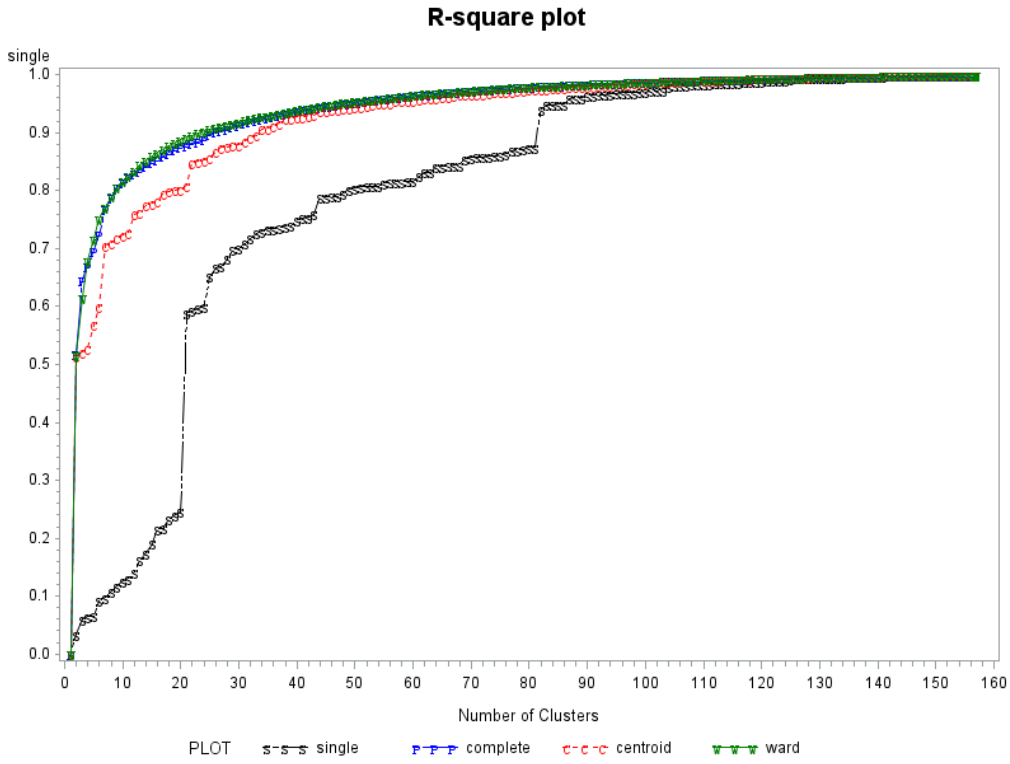


Figure 1: Change in $R^2$ after merging clusters

Based on figure 1, the Ward's method seemed to be the most appropriate for determining optimal number of clusters. This is because the $R^2$ values of this algorithm appear to decrease the least compared to the other algorithms i.e topping the $R^2$ plot. However, the complete linkage algorithm seem to perform more or less the same as the ward's method in most of the clustering history.
The dendrogram (figure 2)below, of the Ward's method, indicates that there are possibly 6 clusters using the indicated red line. This can also be estimated and confirmed from figure 1 between 1-10 clusters where a significant drop in $R^2$ is seen.

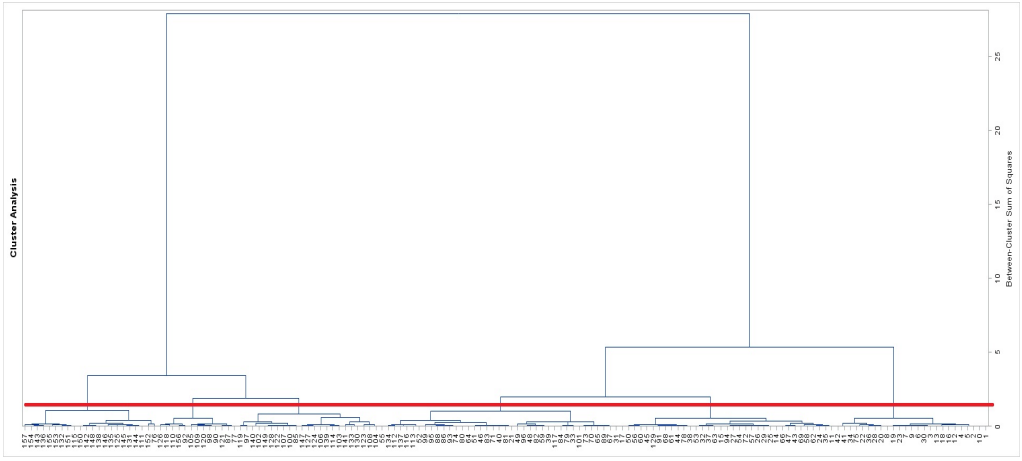


Figure 2: Dendrogram from the Ward's method

## 4.2 Non-Hierarchical Clustering

Cluster means of the 6 clusters obtained from Ward's method were used as initial centers for K-means clustering. This algorithm had a high Pseudo F-statistic of 100.61 indicating that the algorithm resulted in clusters having a very low heterogeneity. The algorithm also had a cubic clustering criterion of 10.93 which was greater than 3 indicating that the clustering was good. The cluster means and mean averages of the variables can be seen in the table below

Table 1: Cluster means of variables per clusters from K-means clustering

| Cluster | **Economy** | **Family** | **Health** | **Freedom** | **Trust** | **Generosity** |
|---|---|---|---|---|---|---|
| 1 | 0.230 | 0.283 | 0.233 | 0.285 | 0.123 | 0.268 |
| 2 | 0.475 | 0.695 | 0.267 | 0.324 | 0.124 | 0.279 |
| 3 | 0.851 | 0.442 | 0.562 | 0.234 | 0.104 | 0.197 |
| 4 | 0.955 | 0.882 | 0.551 | 0.404 | 0.101 | 0.224 |
| 5 | 1.497 | 1.056 | 0.816 | 0.543 | 0.322 | 0.373 |
| 6 | 1.260 | 0.939 | 0.727 | 0.356 | 0.093 | 0.168 |
| **Averages** | | | | | | |
| | 0.879 | 0.716 | 0.526 | 0.358 | 0.145 | 0.252 |

The green values were the means which were higher than their respective average means. Based on these averages, the clusters were labeled. This algorithm ended up with 13 countries in cluster 1; 28 countries in cluster 2; 16 countries in cluster 3; 38 countries in cluster 4; 23 countries in cluster 5 and 39 countries in cluster 6.

## 4.3 Predictive Model

The model was trained using 10-fold cross validation whereby 10 models were built out of which the best(lowest mean square error) on the validation set was selected. The best model had a mean square error of 0.199(lowest) i.e eighth model in figure 3. In order to use a linear model, its assumptions had to be fulfilled. The model diagnostics to investigate the model assumptions are shown in figure 4.

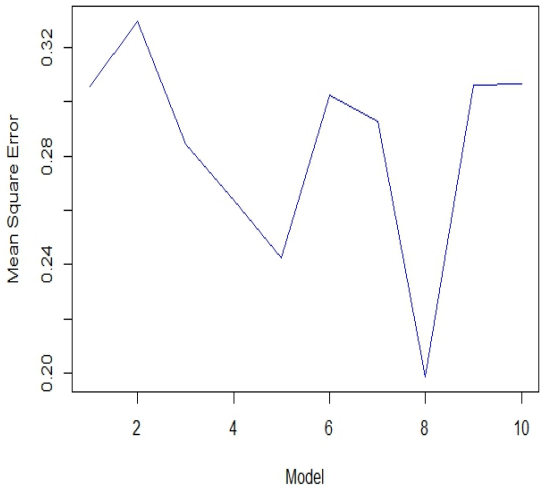


Figure 3: Mean square errors versus models of 10-fold CV

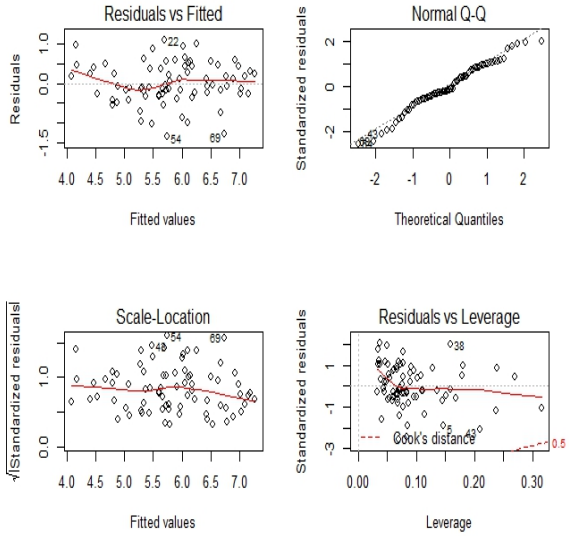


Figure 4: Model diagnostic of best model

The model was summarized in appendix 2 below having a coefficient of determination of 0.698(69.8% of explained variance). Only the *Health, Freedom* and *Generosity* were seen to be insignificantly related to *happiness score*. In addition, there was no multicollinearity among the variables as they all had variance inflation factors below 10 i.e Health=3.024503; Freedom=2.023522 ;Trust=1.622991; Generosity=1.283430; Family=1.823949; Economy=3.647138.
Furthermore, the selected model was tested on the test set and was shown to have a mean square error of 0.537. Predictions were made on simulated data. These predictions along with the prediction and 95% confidence intervals are shown in figure 5-6.

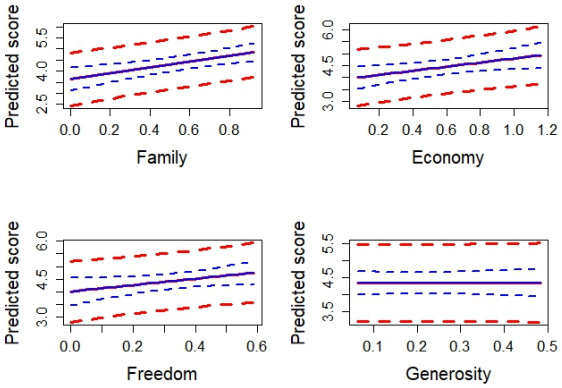


Figure 5: Predictions of happiness score along with prediction interval

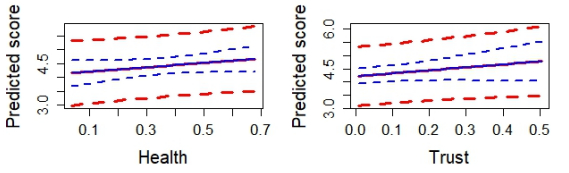


Figure 6: Predictions of happiness score along with prediction interval

The violet line is the predicted *happiness score*, the dashed blue and red lines are the respective confidence and predictive intervals. In addition, the prediction intervals were larger than the confidence intervals. The prediction intervals seem constant unlike the confidence intervals for all 6 variables. Also, the predicted *happiness scores* were increasing along with the respective variables except in the case of *generosity*. In each case of the 6 figures, predictions were made by varying one of the variables while keeping the other 5 variables constant. For example:*Economy* was left to vary while keeping the other 5 variables constant at their means.

# 5 Interpretation

## 5.1 Clustering

$R^2$ refers to the ratio of between-group sum of squares to the total sum of squares of the clusters. The $R^2$ gives a measure of the difference between clusters which is expected to be higher if the cluster are highly different and otherwise. Increase in $R^2$ implies a decrease in semi partial $R^2$(ratio of within sum of squares to total sum of squares) which is a measure of the homogeneity of the clusters. $R^2$ decreases as more distant observations are added to a cluster thereby making the clusters heterogeneous. Heterogeneous clusters lead to a reduction in the between-group sum of squares and an increase in the within-group sum of squares(increase in semi partial $R^2$). The high pseudo F-statistic and cubic clustering algorithm(>3) of the K-means clustering ensured that the homogeneity of the 6 clusters was well optimized.
The very long arms or gaps in the dendrogram from the ward's method(figure 2), are indications of distinct groups as a result of clustering. Therefore, the clusters were chosen these arms and such that other long arms were not included within the chosen clusters.
The mean averages of the 6 variables across the 6 clusters of K-means(Table 1) clustering algorithm, were used to label the clusters. Cluster 1 and 2 had mainly countries which had only high *Generosity* and for this reason they were labeled *Generous and unique* and *Very Generous and unique* countries, respectively. Cluster 3, however, had only countries that had high life expectancy at birth(*Health*) hence were labeled as *Baby survival and unique* countries. Cluster 4 consist mainly of countries which only have a good *Economy*, *Family* well being, high life expectancy at birth(*Health*) and a lot of *Freedom*. Therefore, cluster 4 was labeled as *conserving-citizen* countries because every other characteristic of these countries are higher than average except that they are not generous and think their governments are corrupt. Cluster 5 consisted of countries which were good in every surveyed variable hence were labeled as *Perfect wellbeing* countries. Cluster 6 consisted of countries which had a good *Economy*, *Family* well being, high life expectancy at birth(*Health*) and hence were labeled as *High-GDP-but-family-first* countries.

## 5.2 Prediction

A linear model was chosen to make predictions because they are easier to interpret. In figure 4, the top left plot shows that the residuals show no strong trends hence the independence assumption of linear model was not violated. The top right plot shows that the residuals were normally distributed as the plot does not show any heavy tails and the residuals are more or less aligned to the line. The bottom left plot shows that the homoscedasticity assumption was not violated as the residuals show a constant tube trend or constant variance. The bottom right plot shows that none of the residuals are leverage points as none of them is beyond the confidence limits (red dashed line). Therefore, it seems all assumptions of the linear model were not violated implying that the using a linear model was appropriate.
The plots of figure 5-6 were interpreted as the change in the predicted *happiness score* with change in of the variables while keeping the others constant at their means. The prediction intervals were larger than the confidence intervals because they account for both the uncertainty of the population mean and distribution of the data. In most of the case, the confidence intervals were wider at the ends than in the middle. The confidence intervals are interpreted as the intervals within which there is a 95% confidence of containing the true value of the *happiness score*. Also from the plots, the predicted scores linearly related to the variables can also be confirmed from the model summary(appendix C) as the most related matched the highly significant variables. The plots that were more or less related to the predicted scores were those that were insignificant in the model summary(appendix C).

# 6 Conclusion

The hypothesis of no existence of clusters was rejected as the result of the cubic clustering criterion was above 3. The number of centers of the K-means algorithm can be determined through the agglomerative algorithm. The combination of both clustering algorithms also helped maximize the purity of the finally obtained 6 clusters.
The prediction of the *happiness score* increases with each variable except the case of *generosity*. No matter how generous the country is, the people's happiness is not affected but they seem to be more concerned about the other 5 characteristics. The observed *happiness score* is related to the variables more or less the same way as the predicted *happiness score* is related to the variables.

# 7 Appendix

## A Scattered plots

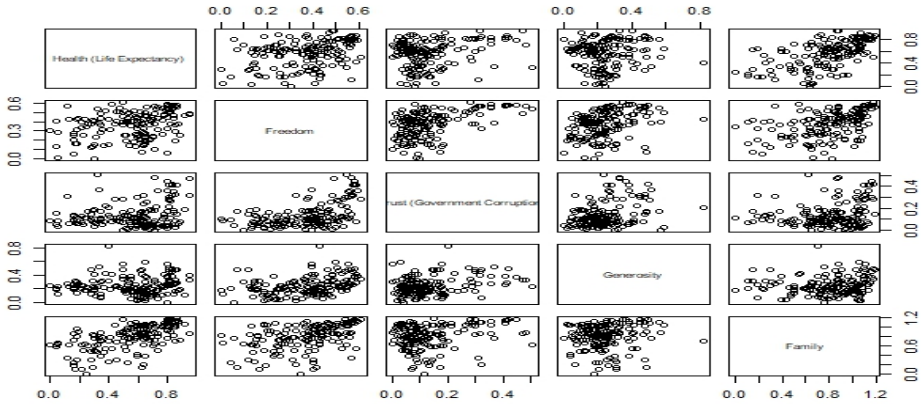


Figure 7: scattered plots

## B Biplots

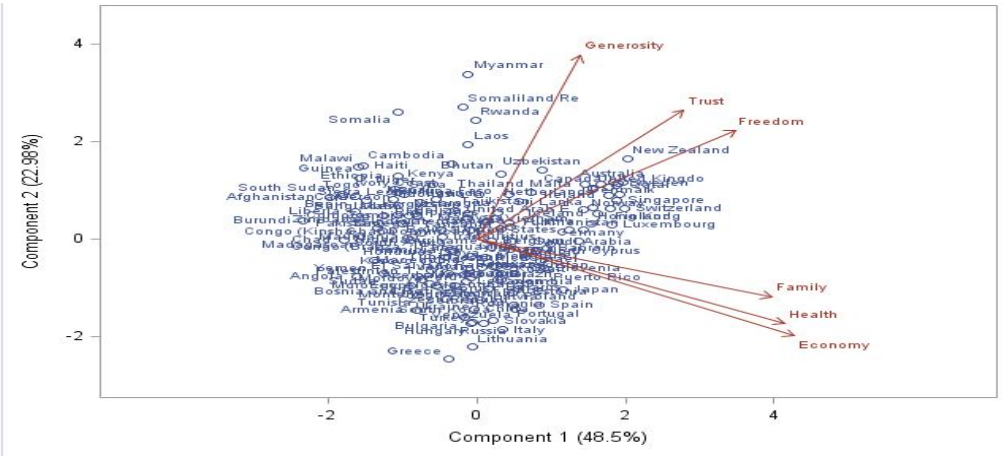


Figure 8: Biplot

## C Model summary

Table 2: Model summary of Predictive model

| Term | Coefficient | Standard error | P-value |
|---|---|---|---|
| **Intercept** | 2.542 | 0.289 | $7.32*e^{-13}$ |
| **Economy** | 0.855 | 0.366 | 0.02 |
| **Family** | 1.319 | 0.361 | $4.99*10^{-4}$ |
| **Health** | 0.808 | 1.448 | 0.152 |
| **Freedom** | 1.267 | 0.683 | 0.0677 |
| **Trust** | 1.663 | 0.637 | 0.011 |
| **Generosity** | 0.033 | 0.473 | 0.945 |