TALLINN UNIVERSITY OF TECHNOLOGY
School of Business and Governance

[redacted]

# HAPPINESS SCORE ANALYSIS

Multivariate Statistical Analysis Project

Tallinn 2022

# TABLE OF CONTENTS

# 1. RESEARCH PROBLEM

According to Jeremy Bentham, an 18th century philosopher, the best society is where its citizens are the happiest (Layard & Layard, 2011). It should, therefore, also be a goal of each and every country to maximize the happiness of its citizens. However, happiness is complex and is influenced by many different factors. For over two decades, scientists have studied different variables that could affect happiness. For example, economic wealth has historically often been associated with a person's happiness, however, what can actually be observed is that even though some countries have gotten richer, the citizens of those countries have not gotten happier (Layard & Layard, 2011). This is especially true in situations where GDP per capita has increased but there is much inequality in the country and because of that the well-being and therefore the happiness of most people has not gotten any better (Stiglitz *et al.*, 2008). However, economic factors cannot be fully excluded either as previous research does indicate a relationship between them and happiness does exist. For example, Peggy Schyns in her article "Cross national differences in happiness: Economic and Cultural factors explored" does prove a correlation between happiness and GDP per capita, especially when additional, in this case cultural, factors are added (Schyns, 1998). Even though no concrete list has been compromised, it is often agreed that multiple facets play into what factors influence happiness, everything from governmental policies to the individual's household (Sujarwoto *et al.*, 2017).

Due to the importance of people's happiness on a society and considering the fact that there is no agreed upon list of factors to determine the happiness of a country's citizens, it is also the aim of this project to take a closer look into some of the factors that have previously been looked at. Using decision-tree and clustering methods, we aim to answer the two research questions set forth in this project.

(1) What are the main factors that influence whether a country has a low or a high happiness score? How influential is GDP per capita?
(2) How can countries be grouped based on the happiness score factors?

Research question (1) is inferential and therefore the tree-based method is used to determine the main factors that influence whether a country is considered to have a low or a high happiness ranking. Research question (2) is also inferential but with the main goal of grouping the observations, which is why cluster analysis is the method used to better understand the data and answer the question.

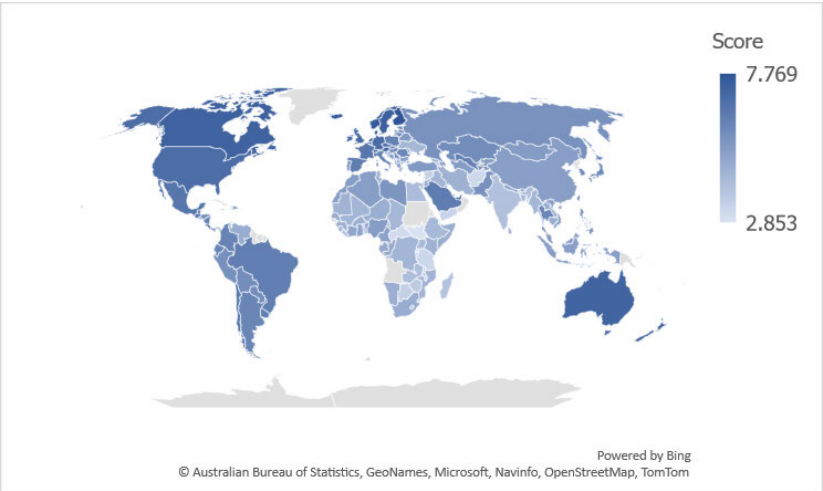
# 2. DATA

The data for the project was extracted from the dataset "World Happiness Report" that was published in community platform Kaggle. The present analysis is based on the 2019 World Happiness Report that ranks 156 countries according to happiness levels. Although there are newer reports available from the official website, the 2019 report was chosen due to its accessibility and format. As some of the data was still unavailable for a few countries, it was substituted with forecasts, interpolation and extrapolation to complete the dataset.

The countries' ranking was set up according to the overall score of happiness from six factors. The score was set up with 0 - 10 scale with 10 being the best possible life and 0 being the worst possible life. The six underlying factors that resulted the happiness scores were: GDP per capita in purchasing power parity (PPP), social support as national average of binary response, healthy life expectancy extracted from WHO, freedom to make life choices (satisfaction with freedom to choose), generosity (response to donation) and corruption perception (the regularity of corruption).

The most highly ranked country was Finland (7.77) followed by other European countries like Denmark, Norway and Iceland. The lowest ranked countries were South Sudan (2.85), followed by the Central African Republic, Afghanistan and Tanzania. Estonia was in the rank 55 with the happiness score 5.89.

The countries with the highest values of the factors will also be pointed out. GDP per capita was the highest for Qatar, whereas Iceland had the most social support. Healthy life expectancy and perceptions of corruption were the highest in Singapore. Interestingly Uzbekistan had the highest freedom to make life choices and Myanmar had the highest generosity factor.



Graph 1. The world map coloured by the happiness score.

The above graph 1 shows that developed countries tend to have higher happiness score, whereas developing countries have lower happiness score. However, as happiness score also takes into account factors like freedom, generosity and corruption, then being a forefront country might not always determine the highest happiness level for the citizens. One of the examples is China, which is an advanced and industrial country, but in the overall happiness ranking is in the 93$^{th}$ position.

Table 1. Descriptive statistics of factors influencing the happiness score

| | Min | Max | Median | Mean | Std. dev |
|---|---|---|---|---|---|
| Score | 2.85 | 7.77 | 5.38 | 5.41 | 1.11 |
| GDP per capita | 0.00 | 1.68 | 0.96 | 0.91 | 0.4 |
| Social support | 0.00 | 1.62 | 1.27 | 1.21 | 0.3 |
| Healthy life expectancy | 0.00 | 1.14 | 0.79 | 0.73 | 0.24 |
| Freedom to make life choices | 0.00 | 0.63 | 0.42 | 0.39 | 0.14 |
| Generosity | 0.00 | 0.57 | 0.18 | 0.18 | 0.1 |
| Perception of corruption | 0.00 | 0.45 | 0.09 | 0.11 | 0.09 |

The table 1 shows the descriptive statistics of the factors. As it can be seen, all the factors had a minimum of 0 for some country in the list. The mean of happiness score was 5.41 and the median was similarly 5.38. The score logically had the highest standard deviation as other factors were similar amount-wise. At this point, it would be suitable to point out that as the numbers do not have as much variety, there are some restrictions for the range of this project.

# 3. METHODOLOGY

In this project there has been used two multivariate methods, which the authors considered most appropriate in answering the research questions. The first research question was to find out the main factors that influence whether a country has a “low” or a “high” happiness score and also how influential GDP per capita is in it. For answering this question, the tree-based method was used. The second research question was to investigate how countries can be grouped based on the happiness score factors, and for this question the cluster analysis was examined. The tree-based method can be applied for either regression or classification problems. It uses a series of conditional statements to partition training data into subsets. Each successive split adds some complexity to the model, which can be used to make predictions. The cluster analysis is a type of unsupervised learning technique, used to find commonalities between data elements and grouping similar observations into respective clusters.

Both of the methods were performed in R. With the tree-based method classification trees were used to analyse the data set. First, the continuous variable was recorded as a binary variable, which takes on a value “yes” or “no”, depending on if the variable “score” is higher or lower than 5. Then the “tree ()” function was used to fit the tree in order to classify the “high” using all the variables except “score”.

With cluster analysis a non-hierarchical clustering procedure was used, where the k-means partitioning was applied. First the K-means technique was used to determine the number of clusters by interpreting the scree plot. Based on the scree plot, it was wise to use three clusters and therefore the next step was to distribute the observations to the said number of clusters with the K-means technique.

# 4. RESULTS AND CONCLUSION

## 4.1. Decision tree

Research question (1) focuses on the main factors that determine whether a country has a low or a high happiness score. For the purpose of this research and based on the data, we chose the breaking point between “High” and “Low” happiness scores to be five out of ten. In order to classify the observed countries as having a low or a high happiness score, we used the tree-based method. According to figure 1, the main factor that determines the classification of a country’s happiness score, is “healthy life expectancy”. When a country has a “healthy life expectancy” score that is lower than 0.6505, then the next determining factor is “GDP per capita”. When “GDP per capita” is lower than 5.305, then these countries are classified as having “low” happiness scores. However, if “GDP per capita” is greater than 0.5305, then the next determining factor is “generosity”. If the score is lower than 0.1705, then the country is assigned a “low” happiness score, whereas if the “generosity” score is higher than 0.1705, then “freedom to make life choices” is the deciding factor as a score lower than 0.4285 results in a “high” happiness score and a “freedom to make life choices” score greater than 0.4285 means that a country has a happiness score greater than 5. (Figure 1)

When moving right from the healthy life expectancy internal node, then the countries are not yet classified as having a “high” happiness score. When a country has a “healthy life expectancy” score of at least 0.6505 but a “social support” score lower than 1.0765, then the country is still classified as having a “low” happiness score. However, if the “social support” score is greater than 1.0765, then the next deciding variable is “GDP per capita”. From there the classification process follows that if the “GDP per capita” is lower than 0.9715, then the next decision will be made based on the value of the “freedom to make life choices” score. If it is lower than 0.4795, then the country’s final classification decision will be made based on the value of “perceptions of corruption”. If the variable’s value is lower than 0.0425 the country is classified as having a “high” happiness score and if greater than 0.0425 then the country has a “low” happiness score. Country is also considered to have a “high” happiness score when “GDP per capita” is greater than 0.9715

or when the “freedom to make life choices” score is greater than 0.4795. (Figure 1)

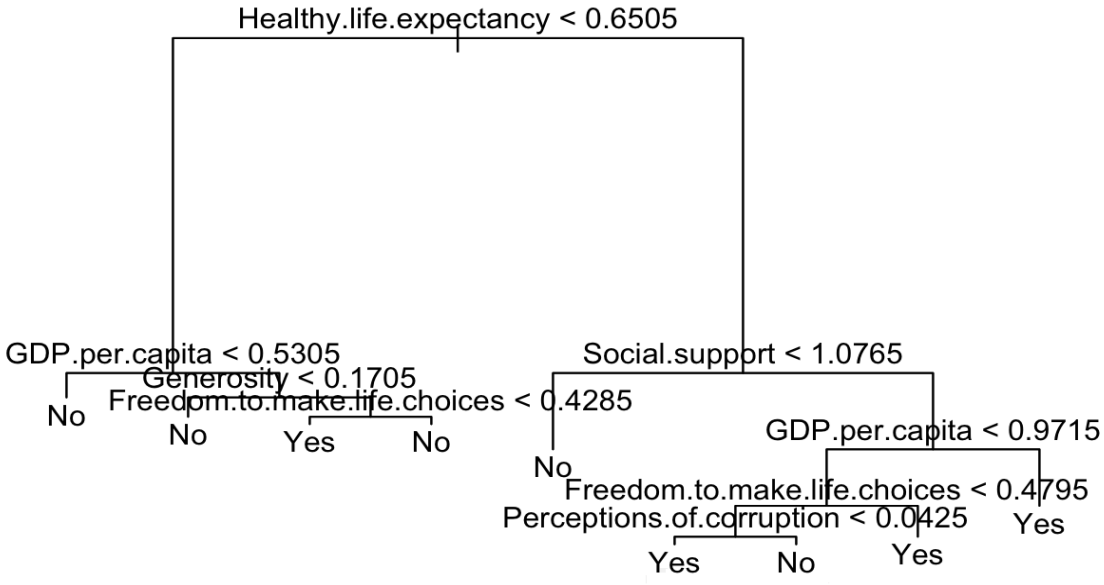



Figure 1. Tree-based method

With the tree-based method the first research question was answered. Based on the tree-based method it came out that the most important factor that influences whether a country has a “low” or “high” happiness score is “healthy life expectancy”. But what also came out is that “GDP per capita” and “social support” are also quite important. “Generosity”, “freedom to make life choices”, and “perception of corruption” are also statistically important, however to a lower extent than the three previously mentioned.

The second part of research question (1) relates to the importance of “GDP per capita”. As can be seen from the decision tree, “GDP per capita” is not the most influential factor when it comes to the happiness score being classified as “high” or “low”. However, as mentioned it does share the second/third place and is therefore still quite important. This confirms previous research published that even though “GDP per capita” is not the only deciding factor, it does still play an important role, especially when the other factors are more social than economic.

## 4.2. Cluster analysis

The cluster analysis is being used to find answers to the second research question. Research question (2) concentrates on grouping the countries based on happiness score factors. Firstly, an optimal number of clusters was found by the scree plot (figure 2). The optimal number of clusters would be 3 meaning that the countries would be divided into 3 groups.

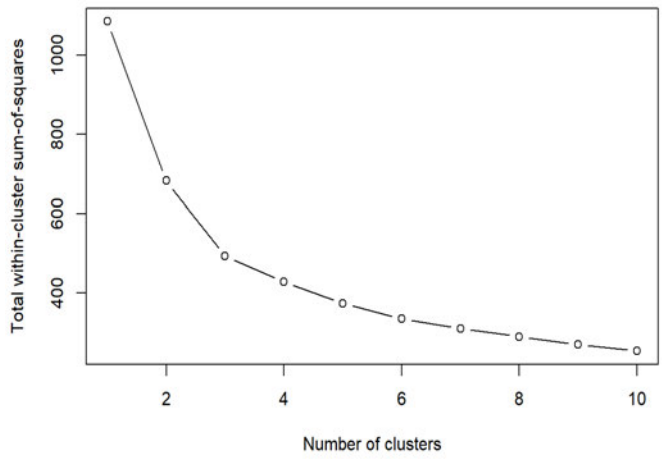



Figure 2. Optimal number of clusters.

There were two options to consider when conducting a cluster analysis. There is an option to include the happiness score or leave it out. In this case both versions were implemented to see the differences.

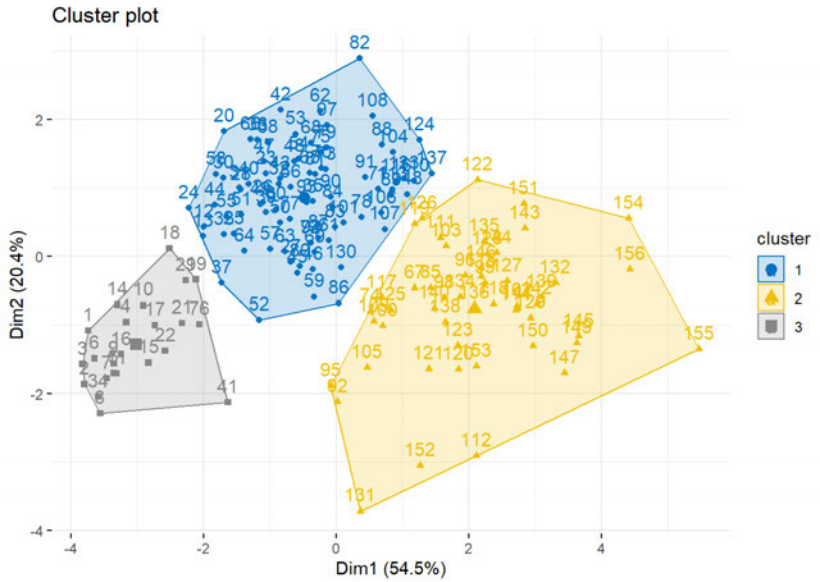



Figure 3. Cluster plot with happiness score.

According to figure 3 with the happiness score included, the smallest cluster is 3 that includes mostly higher happiness score ranking countries. Cluster 1 includes countries that are somewhere in the middle and cluster 2 includes countries that are more in the bottom of the ranking list. The results make sense as the countries with similar score are put together in a group.

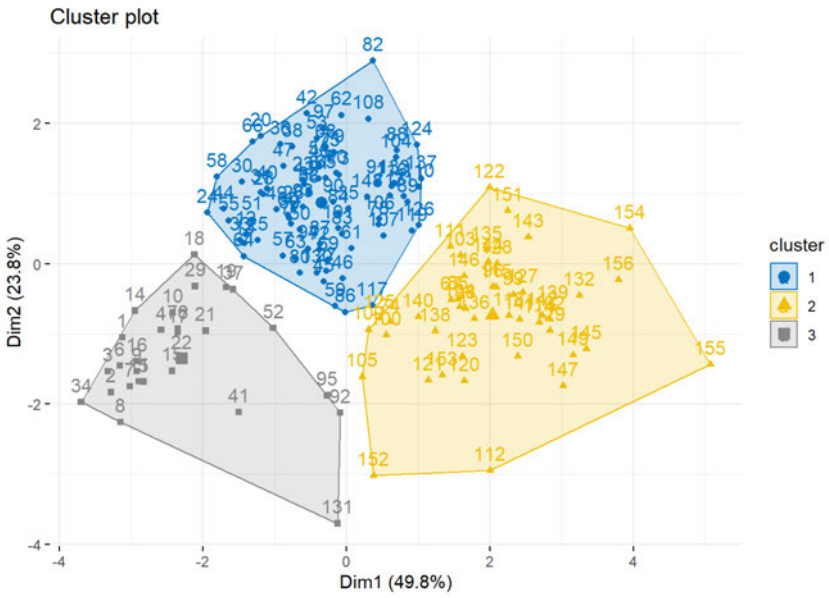



Figure 4. Cluster plot without happiness score.

In figure 4, the happiness score factor was excluded to check for the grouping of countries then. Although mostly it was similar, there were changes in the clusters 2 and 3. In cluster 3 the number of countries went from 23 to 28 (appendix 1) including some of the lower happiness score ranking countries like Myanmar (131), Bhutan (95) and Indonesia (92). The cluster 2 got smaller going from 52 countries to 46 countries (appendix 1) excluding some of the countries to cluster 3. This means that when the happiness score was removed then other factors had a bigger impact in grouping some of the countries differently. Cluster 1 mostly still included about the same number of countries with a medium happiness ranking.

Therefore, it seems like figure 3 gives more accurate grouping of countries as the six factors alone might not be efficient enough to group the countries by happiness as efficiently. Figure 4 seems to give more randomness by including countries like Bhutan to cluster 3 of otherwise similar developed countries. Therefore, the answer to research question (2) would be that countries are grouped into three groups according to similar happiness scores and conditions.

The next potential research that could be conducted regarding the happiness score could be aimed at predicting future happiness scores based on the factors. Due to the nature of the data used in this project, we were unable to create such predictions ourselves, however, with different data, it could be an excellent new direction in which to take this topic. One limitation we encountered was the relatively small number of variables. With a larger number of factors, the results might have been quite different. For future research on the topic, we recommend gathering a larger number and variety of factors to be able to come to a more complete list of factors that influence the happiness score. The other limitation that occurred was that, as the factors had very similar numerical values for all countries, then the differences were not as distinctive, leaving out some interesting findings to be discovered.

# LIST OF REFERENCES

Layard, P. R. G., & Layard, R. (2011) Happiness: Lessons from a new science. Retrieved December 22, 2022, from https://books.google.ee/books?hl=en&lr=&id=KUmbTWcTDKcC&oi=fnd&pg=PT8&ots=nyB5bApcZp&sig=4u7KKP7mDrs-LBAEhRY-s9vaccg&redir_esc=y#v=onepage&q&f=true

Schyns, P. (February 1998). Crossnational Differences in Happiness: Economic and Cultural Factors Explored. Retrieved December 27, 2022, from https://link.springer.com/content/pdf/10.1023/A:1006814424293.pdf?pdf=inline%20link

Stiglitz, J. E., Sen, A., Fitoussi, J.-P. (2008). Measurement of Economic performance and social progress. Retrieved December 27, 2022, from https://ec.europa.eu/eurostat/documents/8131721/8131772/Stiglitz-Sen-Fitoussi-Commission-report.pdf

Sujarwoto, S., Tampubolon, G., (2017, October 6). Individual and Contextual Factors of Happiness and Life Satisfaction in a Low Middle Income Country. Applied Research in Quality of Life. Retrieved December 22, 2022, from https://link.springer.com/article/10.1007/s11482-017-9567-y#ref-CR45

# APPENDICES

**Appendix 1.** Number of countries in the clusters.

| | With happiness score | Without happiness score |
|---|---|---|
| Cluster 1 | 81 | 82 |
| Cluster 2 | 52 | 46 |
| Cluster 3 | 23 | 28 |

Table 1. Number of countries in the clusters.