World Happiness Prediction

Report II

Andrea Gabriela Pop

January 14, 2021

The aim of this report was to build a model to predict how happiness score (or ladder score) is distributed among countries, this was done using clustering, discriminant analysis, and classification. Indeed, for a better understanding of the distribution of happiness throughout the world, factor analysis was employed. The analysis focuses on data from the World Happiness 2020 and data related to alcohol consumption from The World Bank. The analysis will be based on 145 countries around the world.

1 Introduction

Subjective well-being is no longer studied only from a psychological perspective. In the last five decades, the effect of happiness in other disciplines is being object of study in the economic, scientific, and political fields (Clark and Oswald, 1994; Krantz and Manuck, 1984; Radcliff, 2001). Specifically, Easterlin (1974) awakened the real interest in the analysis of the impact of happiness on the economy. Among the contributions to the literature of the respective article, it can be highlighted the demonstration of the correlation between life satisfaction and GDP per capita. Furthermore, positive attitudes indirectly influence the qualification of the health status of individuals (Taylor, Buunk, and Aspinwall, 1990).

During the last decades, the interest in happiness analysis has increased. One of the important data sources related to happiness is the World Happiness Report, an annual report published by the Sustainable Development Solutions Network of the United Nations. This report ranks world happiness by country based survey data measuring different topics such as happiness score of each participant, health status, freedom, trust in government, generosity, life expectancy, GDP per capita, family. Therefore, it allows for identifying potential factors that influence happiness. This information could be helpful since it can be merged with government records to determine the individual's trust in government and the preferred form of government: autocracy, democracy, or oligarchy. Furthermore, it can be established socio-legal factors such as preferences for limits on government: constitutional totalitarian or parliamentary limits of power. In sum, this information is helpful for multinational organizations that are looking to determine potential investment countries with a stable economy, stable government, and a healthy supply of labour.

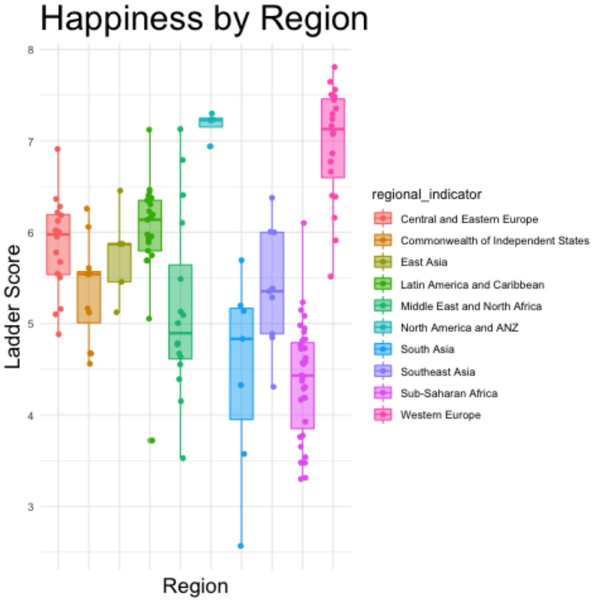
2 Data Cleaning

The concepts of missing values and outliers are important to be taken into account in order to successfully manage data. If missing values are not handled properly, inference and prediction about the data would be probably inaccurate. Therefore, before starting this analysis, missing values and outliers are going to be analysed and commented. In general, if less than 5% of the observations are missing, the missing observations can simply be deleted without any significant ramifications, Harrell (2001). Nevertheless, if more than 5% of the data is missing, deleting the missing data will result in a reduced sample size. For the purposes of this report, missing values have been replaced with the mean of the column. Consequently, after replacing missing value with the mean, the new dataset does not have any missing values and it can proceed with the analysis. On the other hand, it is not efficient to proceed on dropping outliers because omitting these observations would lead to loss of relevant data for the analysis. For example, countries with really low happiness score which will be relevant for the analysis.

3 Data Visualization

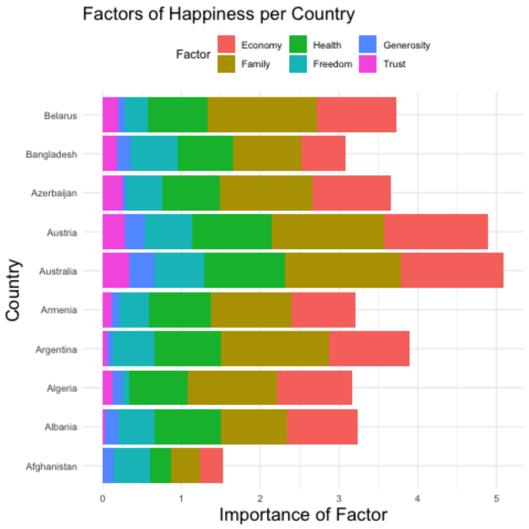
Before starting with clustering, classification, or any other analysis, it is relevant to visualize how happiness score and other factors are distributed. Figure 1 shows the happiness scores by world regions. In this case, Western Europe and North America presented higher mean of happiness scores, whereas Sub-Saharan Africa, South Asia, the middle east, and North Africa showed inferior mean scores in happiness with respect to the other regions. Moreover, this could be relevant when referring to the cluster analysis, which will be shown in Section 5.

Figure 1: Distribution of Happines by Region



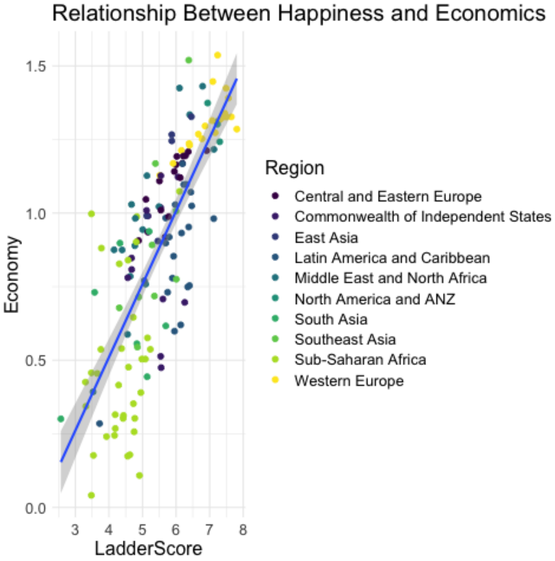
When referring to the importance of each factor in the happiness score of a country, Figure 2 revealed that Economy (GDP per capita) and Family are the most important factors in evaluating a country's happiness. This will be taken into account for the cluster analysis, establishing the relationship between GDP and happiness score. On the other hand, generosity and trust seemed to be the variables less related to a higher happiness score in a country.

Figure 2: Importance of each Factor



Unsurprisingly, Figure 3 shows that the happiest countries and world regions generally tended to be ones with strong and stable economies. Therefore, the happiness score is positively correlated with GDP per capita. This is expected since more economic stability and higher GDP per capita generally encourage stable and comfortable family life as well as increases the availability of proper medical resources and healthcare.

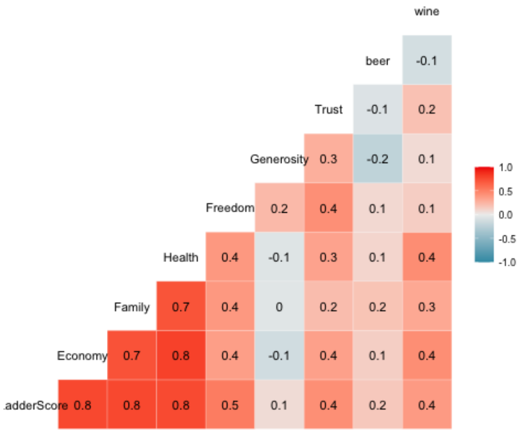
Figure 3: Happiness and GDP relationship



4 Factor Analysis

Factor analysis is an analytical tool that creates linear combinations of factors to abstract the variable's underlying commonality. Indeed, as the variables have an underlying commonality, fewer factors capture most of the variance in the data set. In general, in economics, the number of factors used are three. Nevertheless, this project also presents factor analysis with two factors with the main aim of understating the differences when using only two or more factors. In fact, the predictability of factor analysis is better than PCA. It must be taken into account that factor analysis is also sensitive to outlying and missing values. In this report, as it was stated in Section 2, the missing values were replaced with the mean of the factor since omitting these values could lead to a loss of relevant data for the analysis.

Figure 4: Correlation between factors



	Economy	Family	Health	Freedom	Generosity	Trust	beer	wine
adderScore	0.8	0.8	0.8	0.5	0.1	0.4	0.2	0.4
Economy		0.7	0.8	0.4	-0.1	0.4	0.1	0.4
Family			0.7	0.4	0	0.2	0.2	0.3
Health				0.4	-0.1	0.3	0.1	0.4
Freedom					0.2	0.4	0.1	0.1
Generosity						0.3	-0.2	0.1
Trust							-0.1	0.2
beer								-0.1

Figure 4 presents the correlation between different factors that are supposed to affect the ladder score or also known as happiness score. As it was expected, economy, family, and health are highly correlated with the happiness score.

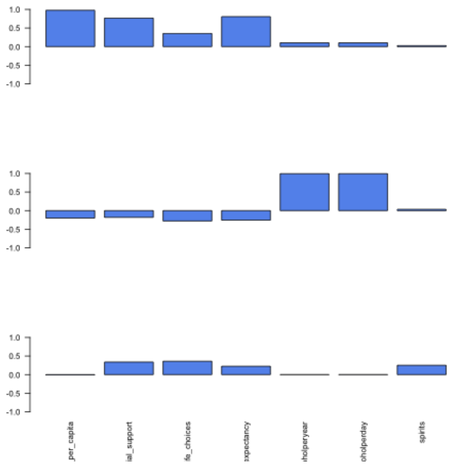
The variables employed for this analysis are related to some of the factors that could affect the happiness score and some alcohol drinks variables. Table 1 presents the results of the factor analysis with no rotation. The first three columns present the 3 factors whereas the last column represents the uniqueness. On the one hand, the uniqueness is referred to as noise and corresponds to the proportion of variability which cannot be explained by a linear combination of the factors. Therefore, a high uniqueness for a variable indicates that the factors do not account well for its variance. In this case, the logged GDP per capita shows small uniqueness, as well as the alcohol per day and alcohol per year drunk. However, the rest of the variables present a considerable value of uniqueness, for example, spirits' uniqueness is almost 1, which means that it cannot be explained by a linear combination of the factors. On the other hand, the loadings are the contribution of each original variable to the factor. Thus, variables with high loading are well explained by the factor. In this case, the logged GDP per capita shows a high loading and a small uniqueness. For social support and healthy life, the loadings are high but the error is quite high which means that the correlation could be noisy. Finally, although alcohol per year and alcohol per day showed a small uniqueness, the loadings are not really well explained by the factor.

Table 1: Factor analysis with no rotation

	Factor 1	**Factor 2**	**Factor 3**	**Uniqueness**
logged gdp per capita	0.977 215 14	−0.199 875 1	−0.010 117 885 2	0.005 000 0
social support	0.766 212 11	−0.180 770 3	0.338 918 446 6	0.265 381 8
freedom to make life choices	0.353 478 85	−0.278 350 8	0.359 753 604 1	0.668 171 7
healthy ife expectancy	0.806 712 99	−0.255 959 8	0.225 196 260 4	0.232 990 1
alcoloholperyear	0.102 163 68	0.993 507 7	0.000 406 369 8	0.005 000 0
alcoholperday	0.102 257 92	0.993 497 6	0.000 672 495 1	0.005 000 0
spirits	0.025 421 96	0.034 542 4	0.251 546 605 4	0.934 936 1

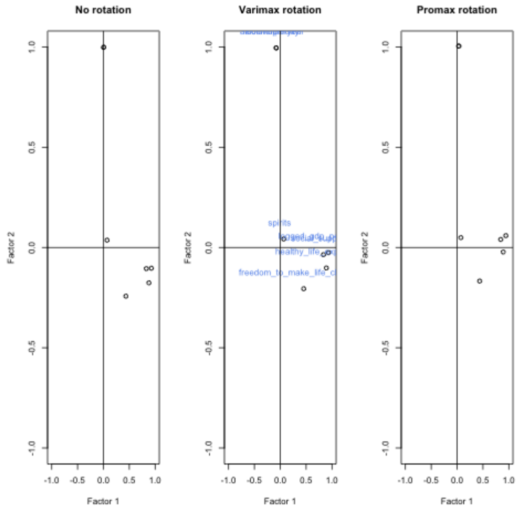
In factor analysis, the solution is going to be unique except the rotation. The purpose of a rotation is to produce factors with a mix of high and low loadings and few moderate-sized loadings. The idea behind it is to give meaning to the factors, which helps interpret them. From a mathematical viewpoint, there is no difference between a rotated and unrotated matrix, meaning that the fitted model is the same, the uniquenesses are the same, and the proportion of variance explained is the same. Figure 5 shows these results in a bar plot. The first bar shows a weighted average, the second the slope and in the last component presents the curvature.

Figure 5: Bar Plot 3 Factors and no rotation



For a better interpretation of these results, Figure 6 shows a scatter plot of the first and second loadings of three-factor models, the first one with no rotation, the second with varimax rotation, and the last one with promax rotation. In general, if two variables have large loadings for the same factor, they will probably have something in common. Therefore, it appears that factor 1 accounts for social support, freedom, and health.

Figure 6: Scatter Plot 3 Factor Analysis



As it was stated at the beginning of this section, this report presents also factor analysis with two factors, which is represented in Table 2. However, in this case, the null hypothesis, H0, is that the number of factors in the model (2 factors) is sufficient to capture the full dimensionality of the data set is rejected. However, the p-value is less than 0.05, which indicates that the number of factors is too small.

Table 2: Factor analysis with no rotation

	Factor 1	**Factor 2**	**Uniqueness**
logged gdp per capita	0.835 579 59	−0.035 535 47	0.300 542 6
social support	0.835 579 59	−0.035 535 47	0.300 542 6
freedom to make life choices	0.453 433 93	−0.205 266 46	0.752 265 2
healthy ife expectancy	0.891 142 98	−0.102 025 11	0.195 455 9
alcoloholperyear	−0.078 699 31	0.996 645 04	0.001 000 0
alcoholperday	−0.078 507 80	0.996 660 10	0.001 000 0
spirits	0.067 042 45	0.042 452 66	0.993 710 7

5 Cluster Analysis

The cluster analysis is an exploratory method to identify similarity structures in data. The objects to be examined in a cluster analysis can be persons, products, or different units such as countries or companies. By using cluster analysis, these objects can be grouped into clusters based on their characteristics and similarities. Each cluster should be as similar as possible (homogeneous) and at the same time be as different as possible from the other clusters (heterogeneous). When deciding the number of clusters, certain rules are used to determine how the objects are grouped into clusters. The outcome of this process depends not only on the choice of clustering algorithm but also on how the distance or similarity between the objects is determined. At the beginning of the cluster analysis, it is important to scale the data and then decide the number of clusters. For estimating the number of clusters, three methods were used for this project: the silhouette, wss, and Partitioning Around Medoids (see Figure 7, Figure 8 and Figure 9). The last algorithm is known to be is more robust than the other ones. Consequently, according to these 3 methods and the majority rule, the best number of clusters selected for the analysis will be two.

Figure 7: Partitioning Around Medoids (PAM)

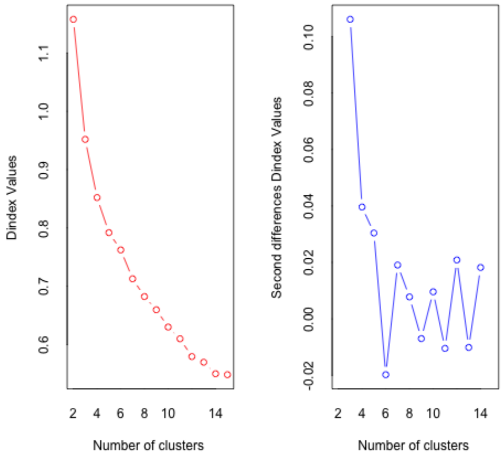


Figure 8: WSS Method

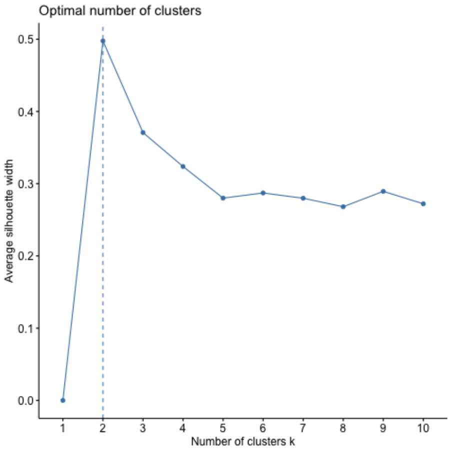
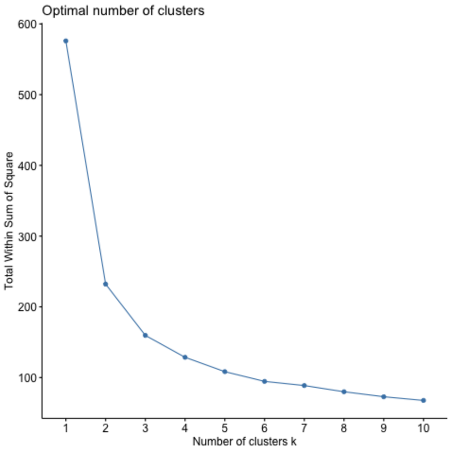
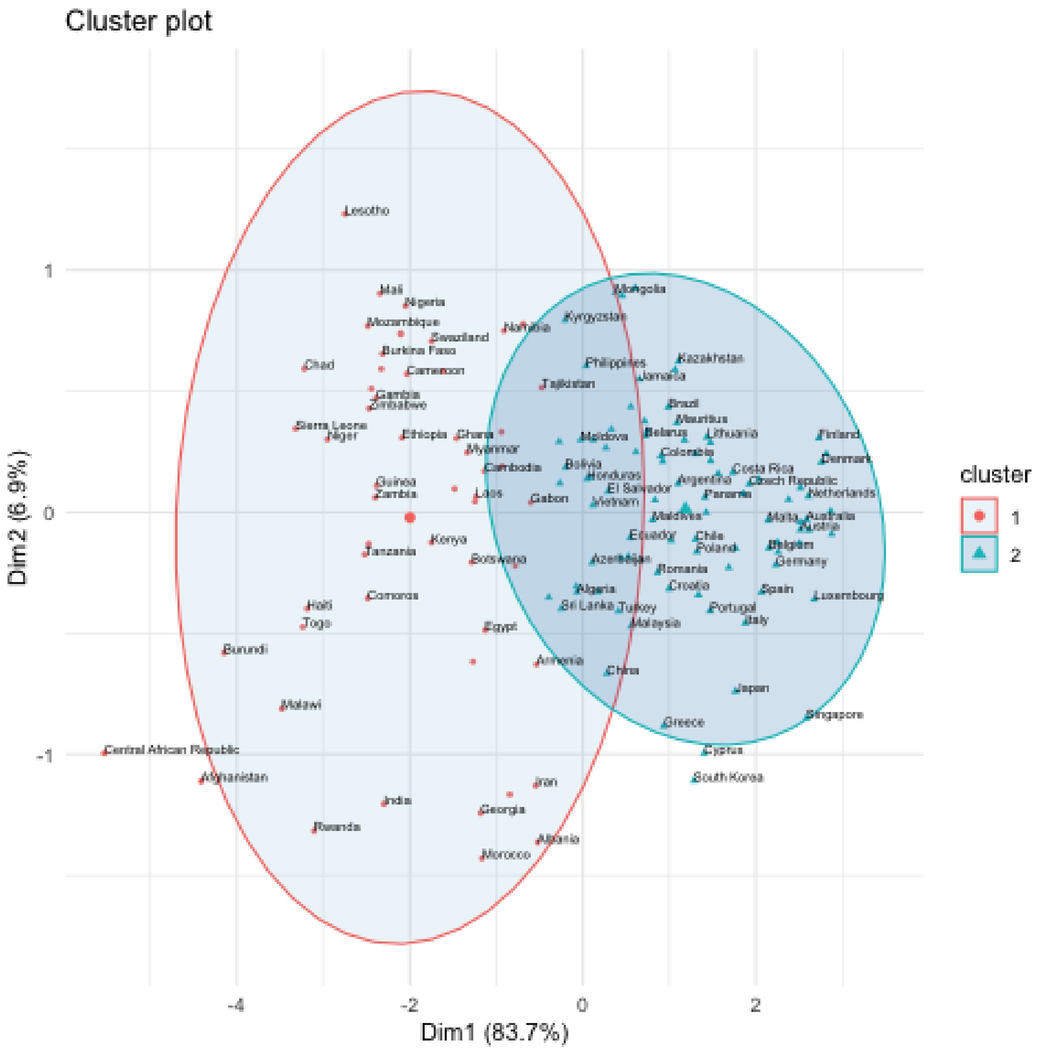


Figure 9: Silhouette Method



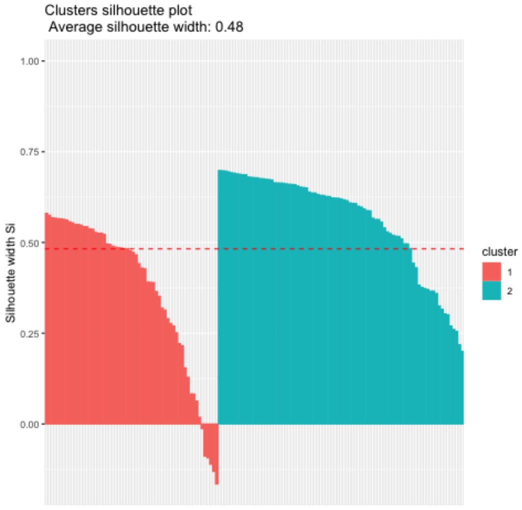
With the aim of the Euclidean distance, the distance between two points can be calculated as a linear distance. Once the proximity measure has been calculated, the actual grouping of the data is carried out using a clustering algorithm. The variables that were input into the model were happiness, GDP per capita, health, and social support. The cluster plot is represented by Figure 10 and it allowed to group countries with similar happiness and socioeconomic characteristics. The first cluster represents in general developing countries and the second one represents more developed countries. For example, the Central African Republic, Togo or Haiti are included in the first cluster and they are situated at the left and bottom of the graph. On the other hand, at the right and top of the graph, countries like Austria, Finland, Denmark, or the Netherlands can be found. Moreover, countries with similar socioeconomic situation and happiness score, for example, Spain, Portugal or Italy, form part of the second cluster and they are situated close to each other. However, it is important to mention that there is overlapping and some countries do not fit correctly. Indeed, the elipson of the first cluster is larger because there is more variance for developing countries, thus developing countries have more variability.

Figure 10: Cluster Plot



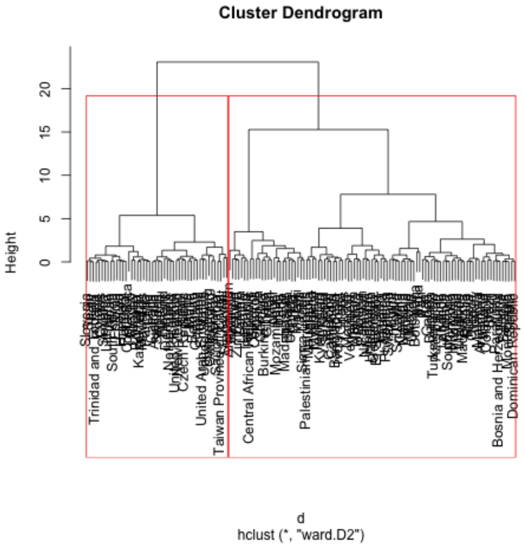
The Silhouette plot represented by Figure 11, measures how well the observations are clustered and it estimates the average distance between clusters. Therefore, observations with negative silhouette are probably placed in the wrong cluster. In this case, the observations of cluster two are correctly clustered, however, for the first cluster, some countries were probably placed in the wrong cluster. This is reasonable as it was commented before and since the first cluster presents more variability. In the case of the second cluster, there are not observations with negative silhouette, which means that the countries are placed in the right cluster.

Figure 11: Silhouette Plot



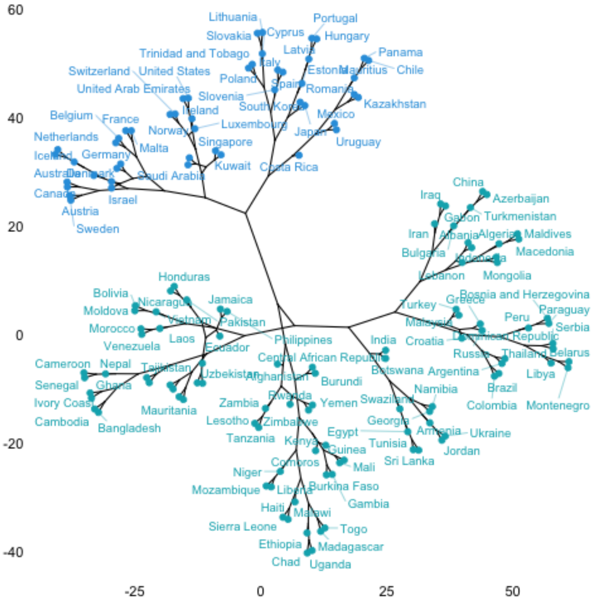
For a clear overview of the clusters, now the variables selected were happiness score and GDP per capita since it was one of the factors that contributes more to the happiness score of a country. The result of the clustering algorithm can be displayed with the aid of the dendrogram, which is represented by Figure 12. All countries are listed individually on the lower side of the dendrogram. Therfore, first of all, each country corresponds to a cluster, which can be seen from the fact that each case has its own horizontal line. Moreover, the dendrogram shows two different selections, which are more visually represented by Figure 13.

Figure 12: Dendrogram Plot



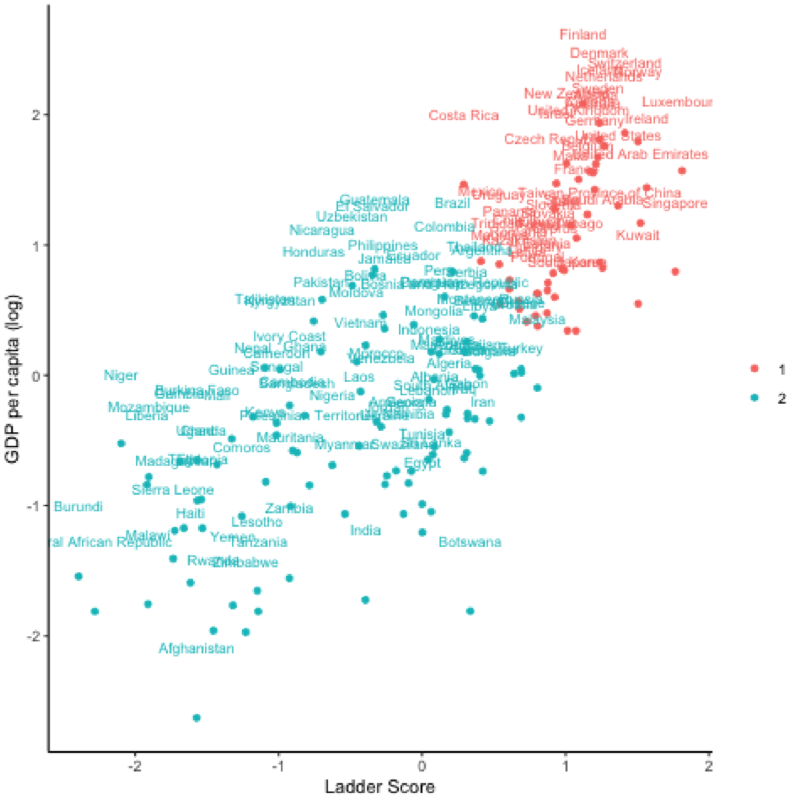
Similarly, Figure 13 represents a more visual Dendrogram. As it is expected, the plot shows European and other developed countries at the top, while other countries like Uganda or Ethiopia are situated at the bottom of the graph. It is also interesting to mention that Figure 13, in comparison with Figure 12, shows clearly countries with similar socioeconomic characteristics and happiness scores. For instance, within the same group, countries like Sweden, Netherlands, or Austria are closer and situated more at the left, while Portugal, Hungary, or Cyprus are situated at the right but still belonging to the same group as Sweden.

Figure 13: Tree Dendrogram Plot



Nevertheless, in this case, K-Means clustering provides more satisfactory than hierarchal clustering results. Figure 14 represents more clearly the clusters when taking into account only the happiness score and logged GDP per capita. For example, developed countries as Sweden, Finland, and Denmark are represented at the top and right (they are included in the cluster one), whereas African countries, Afghanistan or Niger are represented at the left and bottom of the graph (they are included in the second cluster).

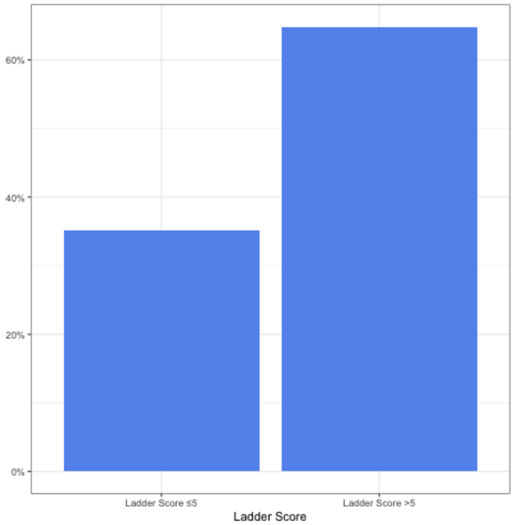
Figure 14: Cluster Plot (Happiness and GDP)



6 Discriminant Analysis and Classification

Logistic regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Naive Bayes were used to predict if the happiness score (or Ladder Score) is higher than 5 or not. The reason for selecting 5 as a breaking point is because the World Gallup considers that areas with a ladder scale superior to 5 are coping areas whereas equal and inferior to 5 are not coping areas, so, therefore, more unfavorable areas considering the happiness score and other socioeconomic variables. Figure 15 shows the distribution in percentage of this categorial variable. Before modeling, a preprocessing stage was needed which included splitting the data into training and test sets (75/25 split).

Figure 15: Distribution of Happiness)



The confusion matrix for these models on test data are shown in Table 3, whereas Table 4 presents Naive Bayes, and some of the most relevant metrics of these models are shown in Table 5.

Table 3: Combined Confusion Matrix and Classification

	Predicted Class 0 (LScore $<$ or $=$ 5)			Predicted Class 1 (LScore$>$5)		
	LR	LDA	QDA	LR	LDA	QDA
$TrueClass0$	10	35	41	0	16	10
$TrueClass1$	2	6	11	25	88	83

Note: LR (Logistic Regression)

The evaluation of these models is done using the confusion matrix and comparison of the metrics, for example, accuracy, sensitivity, specificity, false positives, and false negatives. The confusion matrix shows 4 different values for each model. The true positives refer to the cases in which it was predicted that a country had a ladder score higher than 5 and they really have a ladder higher than that value. In the case of true negatives, it means that it was predicted that the country did not reach a ladder score higher than 5 (thus, it is not a coping or unfavorable area) and they do not reach the ladder score higher than 5. False positives denote that it was predicted yes but the countries do not actually reach the ladder score, (Type I error). Finally, the false negatives denote that it was predicted no, but the countries actually have a ladder score higher than 5, (Type II error).

Table 4: Naive Bayes

	Predicted Class 0 (LScore $<$ or $=$ 5)	Predicted Class 1 (LScore$>$5)
True Class 0	12	2
True Class 1	1	22

By looking at Logistic Regression (LR) confusion matrix, it could be denoted that 25 countries actually have a ladder score higher than 5, while 12 countries do not reach that ladder score. The model predicted, however, that 27 countries have a ladder higher than 5 whereas 10 countries did not. On the other hand, Naive Bayes shows that 24 countries actually have a ladder score higher than 5 while 13 countries do not reach that ladder score. Nevertheless, this model predicted that 23 countries have a ladder higher than 5 whereas 14 countries did not. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) follow the same methodology. Consequently, for deciding which model would be the best in this case, it is important to take a look at Table 5 which shows the comparison metrics of all these models.

Table 5: Comparation metrics

	LR	LDA	QDA	Naive Bayes
Accuracy	0.9459	0.8483	0.8552	0.9189
Sensitivity	0.83333	0.8537	0.7885	0.9231
Specificity	1.000	0.8462	0.8925	0.9167
Pos Pred Value	1.000	0.6863	0.8039	0.8571
Neg Pred Value	0.9259	0.9362	0.8830	0.9565

Table 5 shows the metrics that are going to be taken into account for deciding the method that would be the best for this dataset. The decision will be based on the accuracy which is calculated as the number of all correct predictions divided by the total number of the dataset. In general, accuracy indicates how right the predictions are. Nevertheless, an issue related to the accuracy is that it assumes positive and negative errors are equal. However, this could not be true in all cases. Therefore, sensitivity which is calculated as the number of correct positive predictions divided by the total number of positives, and specificity (the number of correct negative predictions divided by the total number of negatives), are also taken into account for the comparison of the models. Finally, it will be referred to as the false positive rate which is calculated as the number of incorrect positive predictions divided by the total number of negatives.

It is clear that Logistic Regression shows higher values for almost all the parameters, excluding the sensitivity value which is lower than Naive Bayes model. In general, the higher values of Logistic Regression parameters are due to the model maximizes the conditional data likelihood function. The second reason is that the feature values in Logistic Regression are dependent and there is more correlation between these features which contribute to the prediction of the new data points. As it can be noticed in Table 5, Naive Bayes shows lower accuracy than Logistic Regression and this might be due to the reason that the features are independent of each other and each feature for the Naıve Bayes classifier contributes individually to the prediction of the new data point. Secondly, the features are not correlated and this assumption decreases the number of parameters that must be estimated to learn the classifier and this might be the reasoning the algorithm shows sometimes lower performance values for sensitivity, specificity or accuracy. Nevertheless, it must be stated that Naive Bayes's performance is better for every parameter than LDA and QDA, which models show lower values in the parameters.

On the other hand, sensitivity performance is higher in the Naive Bayes model, this denotes how often the model chooses the positive class when the observation is in fact in the positive class. Finally, in the case of specificity performance, Logistic Regression presents a value of 100%. Thus, if the best classifier is chosen only by looking at the value of specificity is because we do not want any false alarms or false positives.

Therefore, it could be confirmed that Logistic Regression is a decent classifier for the World Happiness dataset considering the relatively larger number of true positive and true negative values, as well as the other metrics. This model showed a high accuracy of the overall accuracy of the model is 94.59%. However, it must be stated that Naive Bayes's performance is also great in terms of sensitivity, as well as it presents a high accuracy of 91.89% which is significantly close to Logistic Regressio's accuracy value.

7 Conclusion

The present study aims to shed light on the relationship between the happiness or ladder score with other socioeconomic factors such as GDP, health status or social support, and the distribution of these indicators within different countries around the world. Initially, it was stated that factors, for example, GDP, family, and health were related to a higher happiness score.

After carrying out the analysis and plotting the data, the following conclusions were derived. On the one hand, factor analysis indicated that the logged GDP per capita shows a high loading and a small uniqueness, which contributed to explaining the factor. Moreover, GDP was the variable which showed one of the higher correlation with ladder score (or happiness score).

On the other hand, the analysis of this report showed that happiness across the world is highly dependent on the region. As it was shown, countries in Europe and North America are happier than countries of Sub-Saharan region of Africa which also lag behind in development. After carrying out clustering on factors, the different plots and dendrograms showed a similar pattern as plotting happiness score by regions.In general, developed countries, which in fact presented a higher ladder score and better socioeconomic indicators, were placed on the same cluster and closer to each other. All the clustering analysis was based on k=2 according to the three methods used: the silhouette plot, wss, and Partitioning Around Medoids (see Figure 7) algorithm which is known to be is more robust.

Finally, discriminant analysis and classification showed that Logistic Regression and Naive Bayes are decent classifiers for predicting a ladder score higher than 5.

As a recommendation, further research could better explain the data by employing more variables in order to better explain individual happiness while controlling for education or environmental factors that could impact people's happiness. This is due to it is complicated to employ aggregate data with countries as the unit of analysis to explain individual happiness.

8 References

Easterlin, R.A. (1974), Does Economic Growth Improve the Human Lot? Some Empirical Evidence. Nations and Households in Economic Growth. Academic Press, pp. 89-125.

Ferrer-i-Carbonell, A., Frijters, P. (2004), How Important is Methodology for the estimates of the determinants of Happiness?. The Economic Journal, 114, pp. 641-659.

Graham, C. (2009), Happiness Around the World: The Paradox of Happy Peasants and Miserable Millionaires. Oxford University Press.

Harrell, Frank E. 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer.

Krantz, D.S., Manuck S.B. (1984), Acute psychophysiologic reactivity and risk of cardiovascular disease: A review and methodologic critique. Psychological Bulletin, 96(3), pp. 435-464.

Radcliff, B. (2001), Politics, markets and life satisfaction: the political economy of human happiness. American Political Science Review, No95.

Taylor, S., Buunk, B.P., Aspinwall, L.G. (1990), Social comparison, stress, and coping. Personality and Social Psychology Bulletin, 16(1), pp. 74-89.