CS989 Big Data Fundamentals Coursework

Topic: Factors Affect Happiness Score in Different Countries in 2015

Student: WAN-TING TSENG
Registration Number: 201984399
Professor: Yashar Moshfeghi
Date: 11/11/2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The World Happiness Report is a milestone investigation about the happiness of the nations of worldwide that ranked around 158 countries using the feeling of their citizens recognized from themselves. The report was initially issued in 2012, and published once each year except 2014. It is indicated by The World Happiness Report that happiness and well-being are important signals of a country's overall developments, including economy and society, and the government should regard them as critical purposes of policy (John Helliwell, Richard Layard and Jeffrey Sachs, 2015).

The annual sample size of the survey is around 1,000 people. When measuring the current report, the Gallup would integrate the data that was collected in preceding years to enlarge the sample size to reduce the errors caused by random sampling.

In 2015 report, the data is based on surveys collected from individual person over 150 nations on a scale with the interval between 0 and 10 from 2012 to 2015 by Gallup separately for these reports. The data provides several information shown below:

1. Region
2. Happiness Rank
3. Happiness Score
4. Economy (GDP per Capita)
5. Family
6. Health (Life Expectancy)
7. Trust (Government Corruption)
8. Generosity
9. Dystopia Residual

Dystopia is an unreal country that people who live there is extremely unhappy in the world. To compare different factors in each country, the Dystopia norm is created. Variables which contains the lowest score indicate the attribute of Dystopia.

Through analysing these key variables, it might be clear to find the relationship between the Happiness Score and the factors that might affect the Happiness Rank.

In this report, we will first look at the overall happiness situation among the world. Then, we will look into more details by regions that are given in the dataset, and try to find out if there is something interesting.

Chapter 2

Key Challenges and Problems

Before analysing the data, the first important step is to check if there is any missing data. Although there is no missing data in these datasets, there are some values that are zero. By checking the information from the official website of The World Happiness Report, it is mentioned that not each nation is surveyed yearly. As a result, the value may be measured by records that are in former years. There are six different countries that each has one value is measured as zero in different variables in both years. Moreover, these nations were ranked at the bottom among all countries. As a consequence, their ranking might be affected slightly by the zero values.

In general, if there is outlier in the dataset, it is necessary to check if the outlier would have an impact on the outcome. In this case, although there are some outliers in Family, Trust (Government Corruption), Generosity and Dystopia Residual, it is not possible to change those values because the ranking will be affected. Consequently, it will be proper to maintain the original values.

The key aim of this report is to find out what factors might influence effectively to the ranking of happiness, and to figure out if the Happiness Score can become a reference index while evaluating development of a country. It is sometimes questioned if the source of data is objective and reliable. These reports only use seven variables to measure Happiness Score. However, there might be other factors that will affect the score and the ranking. As a result, these reports will first focus on analysing the relationships between seven factors and the Happiness Score. Then, it will use different analysis methods to analyse the data. Finally, there will be a reflection and conclusion to the analysis at the end of the report.

Chapter 3

General analysis and summary of the dataset

All data used in this report is accessed from Kaggle. The dataset consists of 158 rows and 11 columns in 2015, and 157 rows and 12 columns in 2016. Since there are many countries, it is quite difficult to visualise each country with their Happiness Score in a plot. As a result, it is clear to classify those nations by regions to compare their Happiness Score. From Figure 3.1, it shows the number of countries in each regions. Sub-Saharan Africa have the largest number of country compared to other regions.

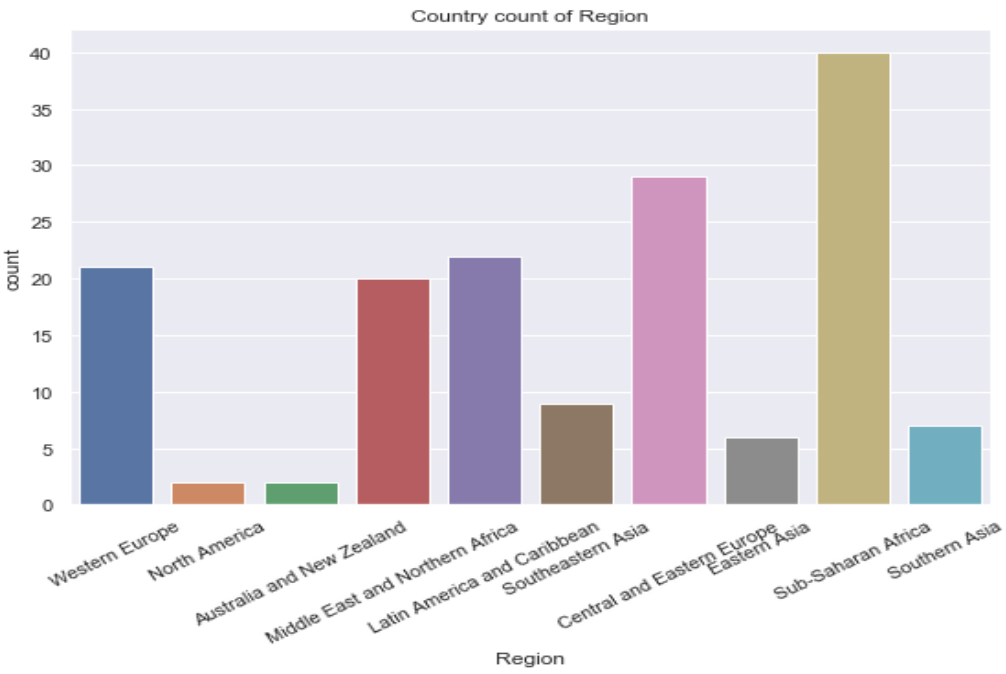

Figure 3.1: Country count of region

Figure 3.2 shows the distribution of Happiness Score in different regions. It is indicated that regions located at the top of the plot are developed countries, including Western Europe, North America, Australia and New Zealand. On the other side, regions like Sub-Saharan Africa and Southern Asia seem to have lower Score than the other regions.

It is easy to realise the western countries seem to earn higher Happiness Score than the developing countries. Take Western Europe for example, the Happiness Score mean is just under 7, which is much higher than the mean of Sub-Saharan Africa (around 4.3). It is straight forward to analyse the outcome that because of the high GDP, developed countries seem to have the high Happiness Score. However, the economy might not be the only one factor that affect the Score.

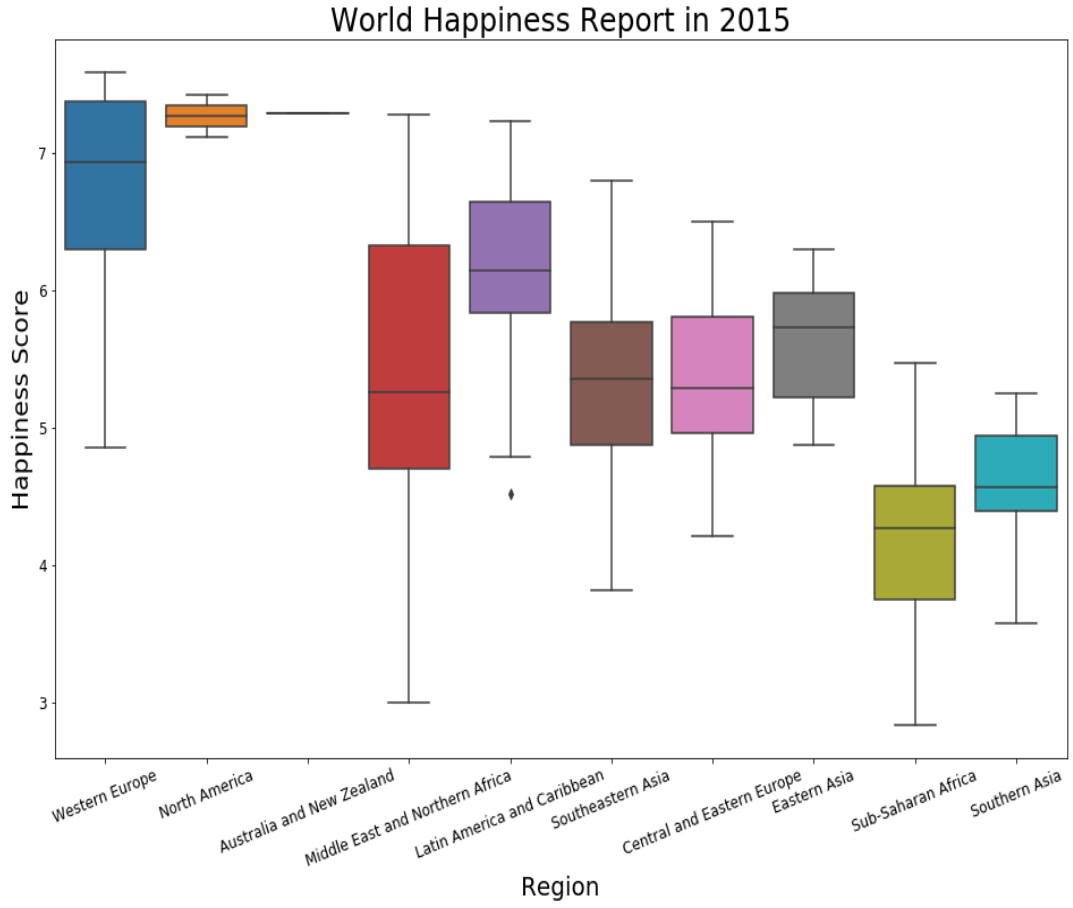

Figure 3.2: Box plot of different region

As a consequence, it is necessary to find out the correlation ship between Happiness Score and the other factors. From Figure 3.3, it is shown that Economy (GDP per capita), Family and Health (Life Expectancy) have a high correlation ships with the Happiness Score. On the other hand, Trust and Generosity are less correlated with the Score.

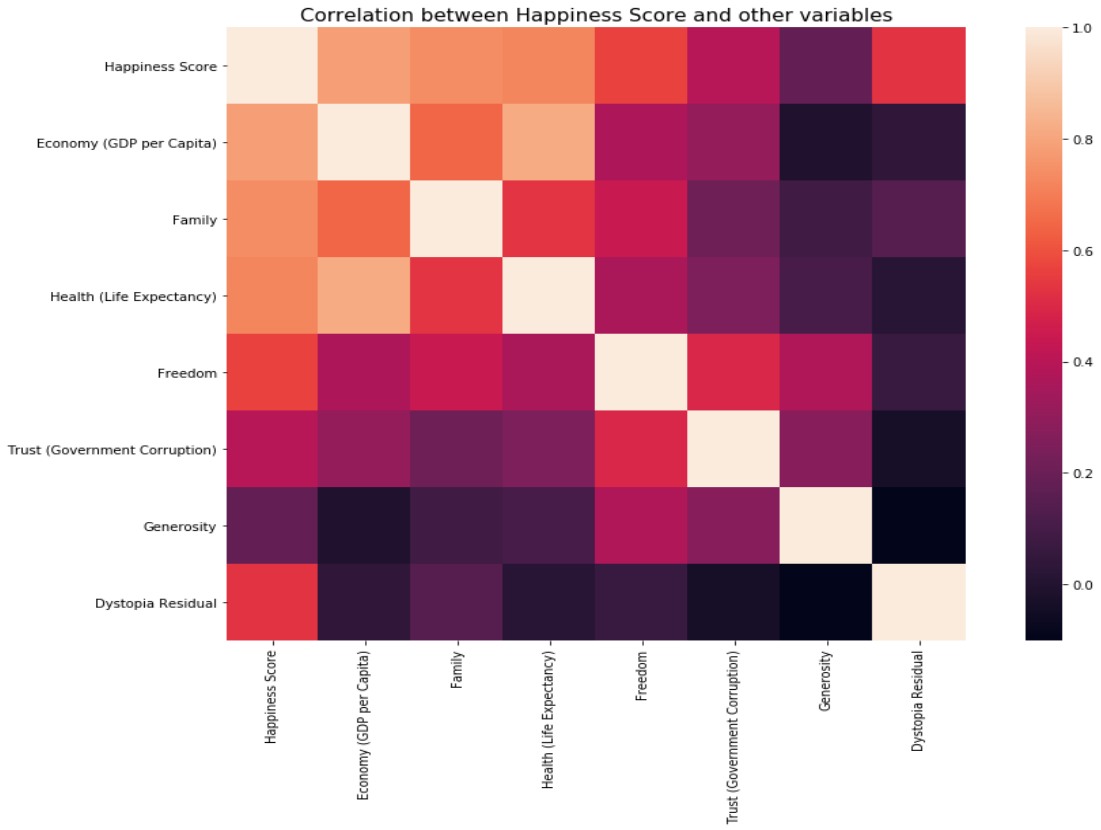

Figure 3.3: Heat map for Happiness Score and all variables

Look into more details, the Figure 3.4 shows the numbers of correlation ship between Happiness Score and the other variables. From Figure 3.3, it is observed clearly that the correlation between Happiness Score and Economy is around 0.8, which means the higher the GDP per capita is, the higher the Happiness Score is. However, the correlation among Happiness Score and Generosity is under 0.2, which means the score of Generosity will not play an important role on the Happiness Score.

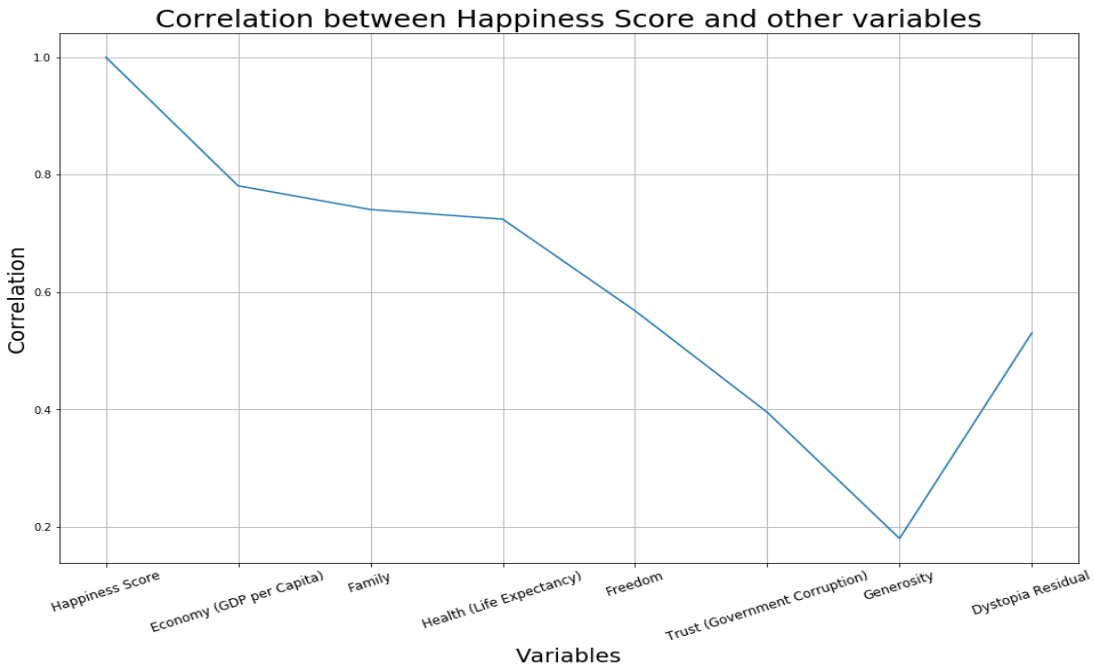

Figure 3.4: Correlation between Happiness Score and other factors

As mentioned above, there is a strong related connection between Happiness Score and Economy. Consequently, from Figure 3.5, it is obvious to state that regions with high GDP seem to have higher Happiness Score. For example, Western Europe in blue colour sits on the right top of the plot. On the other side, regions such as Sub-Saharan Africa in yellow colour has a low GDP and also get a low score in happiness. The countries in Sub-Saharan Africa almost locate in the left bottom of the scatter plot.

Figure 3.5: Distribution of Happiness Score and Economy

Now let's draw attention to each region to see what factors play a significant role to affect Happiness Score.

3.1 Western Europe, North America, Australia and New Zealand

First, let's look at the regions that earns high Happiness Score than the other regions. From Figure 3.6, it is shown that Family, Freedom, Trust and Dystopia Residual seem to have a high correlation ship with the Happiness Score. However, Health is less related with the Happiness Score. The reason for this is that those countries are developed nations, so they already have good improvement on health field. As a result, only few people need specialist help (John Helliwell, Richard Layard and Jeffrey Sachs, 2015).

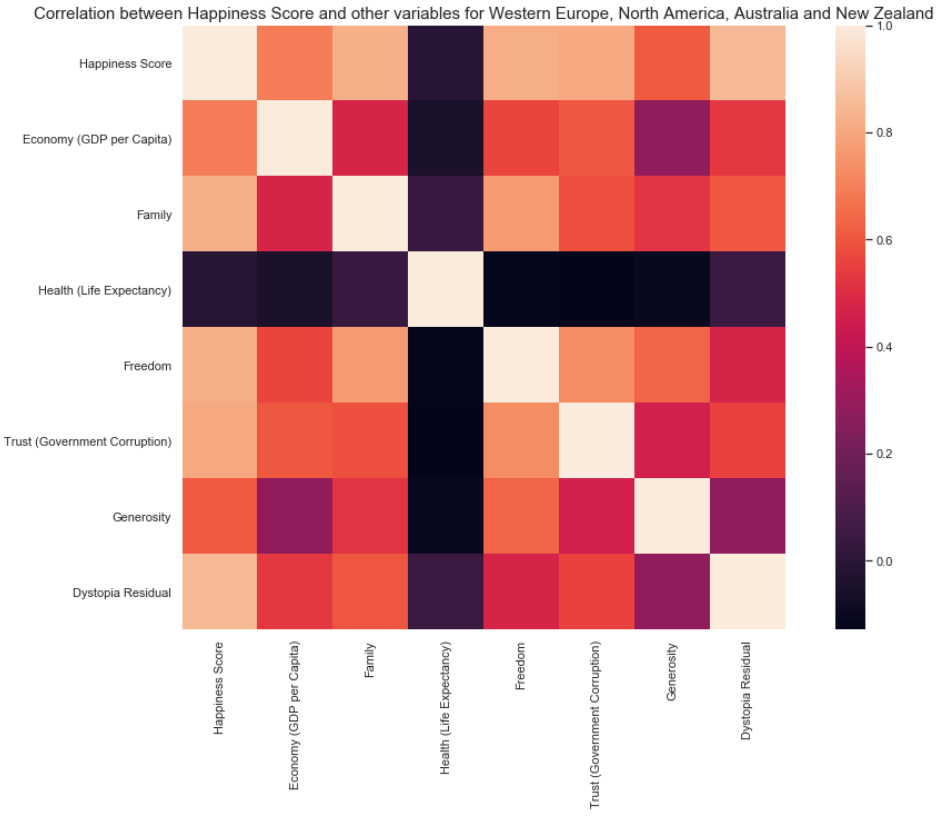

Figure 3.6: Happiness Score and other variables for Western Europe, North America, Australia and New Zealand

3.2 Eastern Asia, South Eastern Asia, Central and Eastern Europe, Middle East and Northern Africa, Latin America and Caribbean

Second, from the Figure 3.7, the correlation between Happiness Score and variables seems to be lower than it is in the high score regions shown in Figure 3.6. Nevertheless, the correlation in Health in the Figure 3.7 is higher than the number in the Figure 3.6. As mentioned before, many people in these regions still struggle with the health problem. As a result, they might not pay lots of attention on other factors such as Economy.

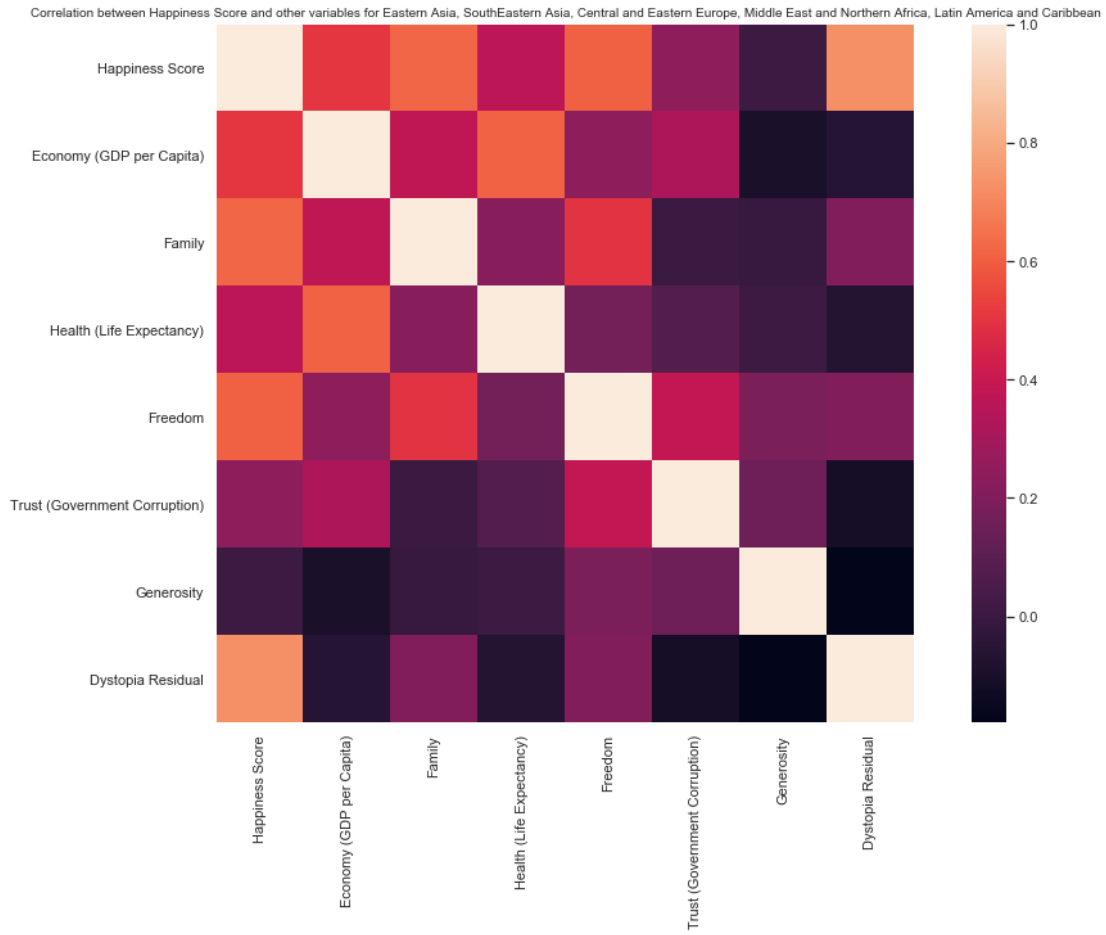

Figure 3.7: Heat map for Happiness Score and other variables for Eastern Asia, South Eastern Asia, Central and Eastern Europe, Middle East and Northern Africa, Latin America and Caribbean

3.3 Sub-Saharan Africa and Southern Asia

From Figure 3.8, the correlation between the Happiness Score and all variables seem to present a similar phenomenon that is in the Figure 3.8. For instance, both of them show decrease in correlation ship in all factors. The reason for this might be the increase in the population in poor citizen in those regions.

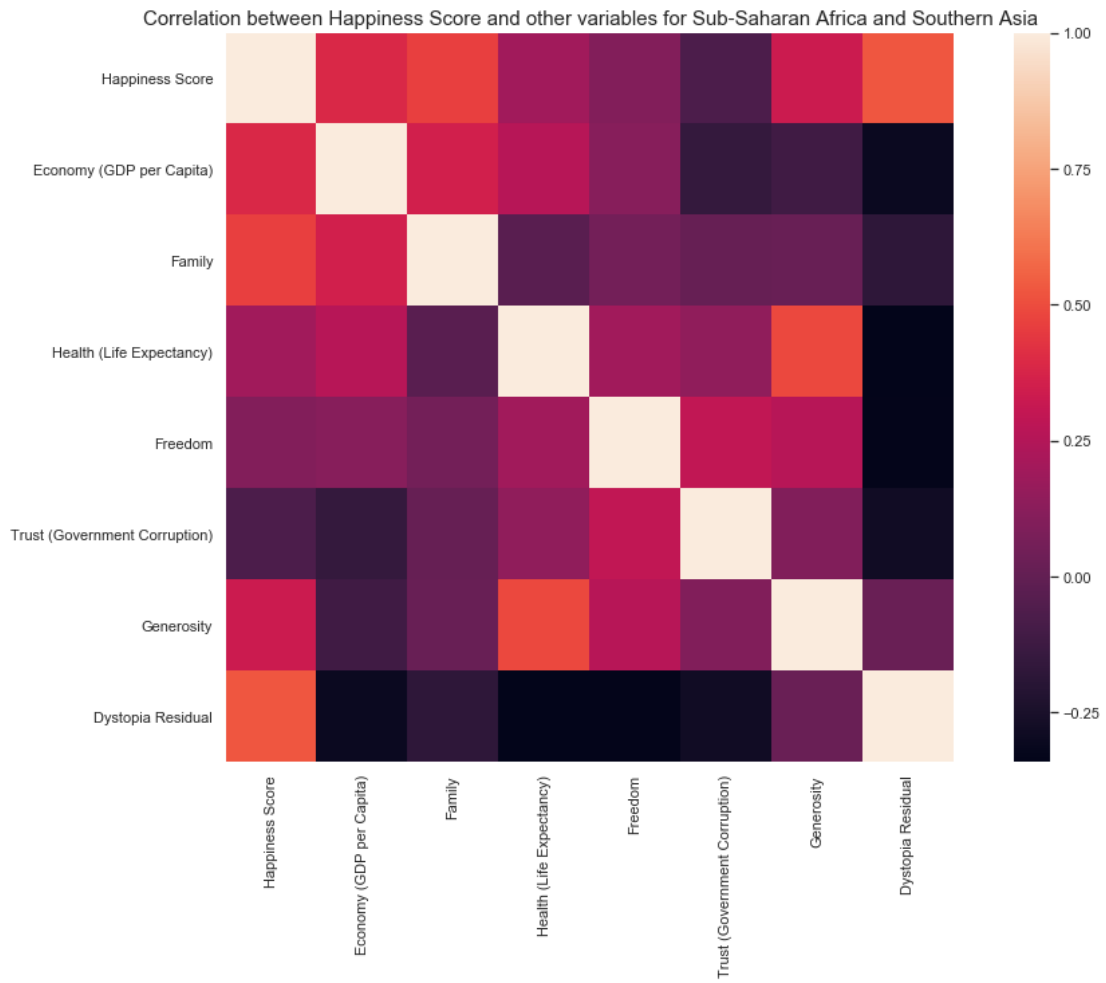

Figure 3.8: Heat map for Happiness Score and other variables in Sub-Saharan Africa and Southern Asia

3.4 Countries that Rank Head and Tail in all nations

Figure 3.9 shows the count of countries that rank in the first 50 position. It is shown that in the top 50 countries, most of them are in Western Europe, followed by Latin America and Africa. Figure 3.10 shows the count of countries

that rank in the bottom 50 position. It is clear that most countries in Sub-Saharan are ranked at the bottom of 158 countries.

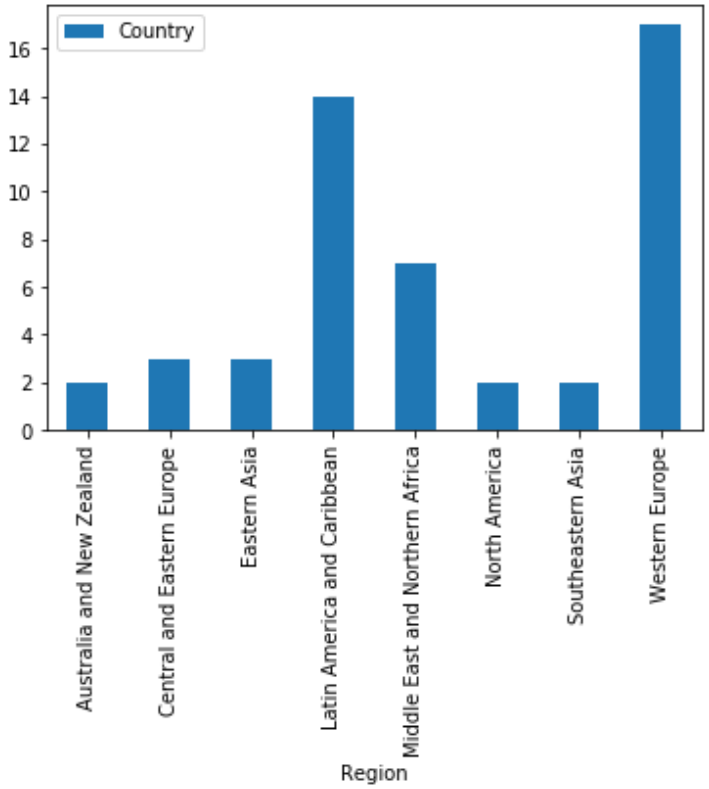

Figure 3.9: Count of top 50 countries in different regions

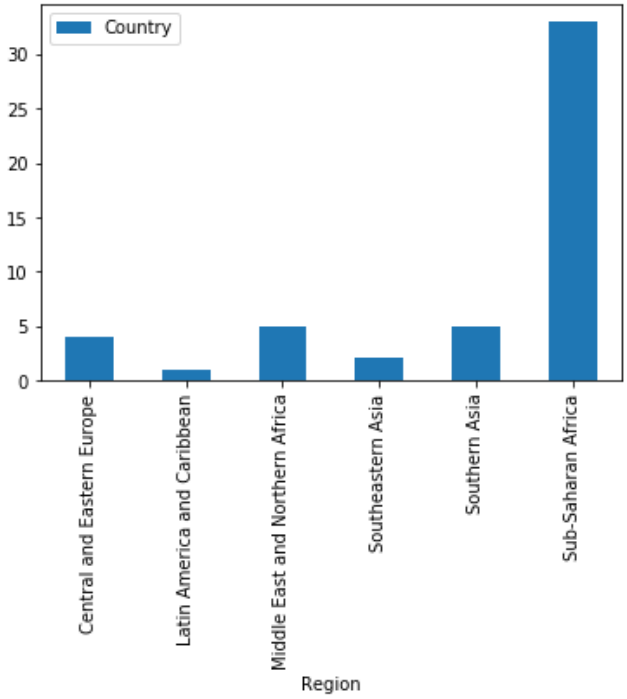

Figure 3.10: Count of bottom 50 countries in different regions

3.5 Summary

As mentioned before, different regions might have distinct factors on their Happiness Score. It is quite interesting that none of their Happiness Score has the high correlation ship in Economy. It is indicated that people regard other variables as more important factors than the GDP.

It is straight forward to find that the Western Europe, North America seem to have higher Happiness Score than the countries in Africa. However, the strong factors that have an important effect on each region are different.

Table 3.1: Factors affect Happiness Score in Different Region

Region	High Related Factors	Less Related Factors
Western Europe, North America, Australia and New Zealand	(1) Family (2) Freedom (3) Dystopia Residual (4) Trust	(1) Health
Eastern Asia, South Eastern Asia, Central and Eastern Europe, Middle East and Northern Africa, Latin America and Caribbean	(1) Family (2) Freedom (3) Dystopia Residual	(1) Trust (2) Generosity
Sub-Saharan Africa and Southern Asia	(1) Family (2) Dystopia Residual	(1) Trust (2) Freedom

Chapter 4

Unsupervised Analysis - Clustering

The aim of using unsupervised Analysis is to find out the patterns in the data. It means that the data given to the algorithms of unsupervised approaches is not labelled. By using the unsupervised methods, it might be possible to figure out some interesting structures that are not easily to be seen (Kurama, 2019).

In this report, two methods are examined, including K-Means clustering and Agglomerative clustering.

4.1 K-Means clustering

K-Means is a clustering method that categorise data into n groups, which the n is a given specific number. Then, each point will be clustered to the cluster which has the nearest mean. Finally, there will be a new point in the centre of each cluster. The process will be repeated for each point. Moreover, the feature of data needs to be numerical. It is a top-down approach. The goal of K-means algorithm is to select the central points that minimise the inertia, the criteria is below:

$$\sum_{i=0}^{n} \min_{\mu_j \epsilon C} (\|x_i - \mu_j\|^2)$$

There are two ways to find the appropriate number of clusters. First, given a number to the cluster. From the Figure 3.9, the number of clustering is given as 5, and outcome of each number can be seen. It is shown that when the group of clusters is three, the performance is better than the other. Second, the appropriate group of clusters can be selected by giving a range of numbers. From Figure 4.2, the performance is the best where the number of group is 3.

Figure 4.1: Results of Given Clusters – K-Means

Figure 4.2: Results of Different Clusters – K-Means

There might be some disadvantage of this method. From the outcome in Figure 4.2, although 3 groups seem to be the appropriate number for clusters, it is not

as best as shown in the Figure 4.3. The distance between each cluster is not clear enough to cluster recognise each point belongs to which clusters.

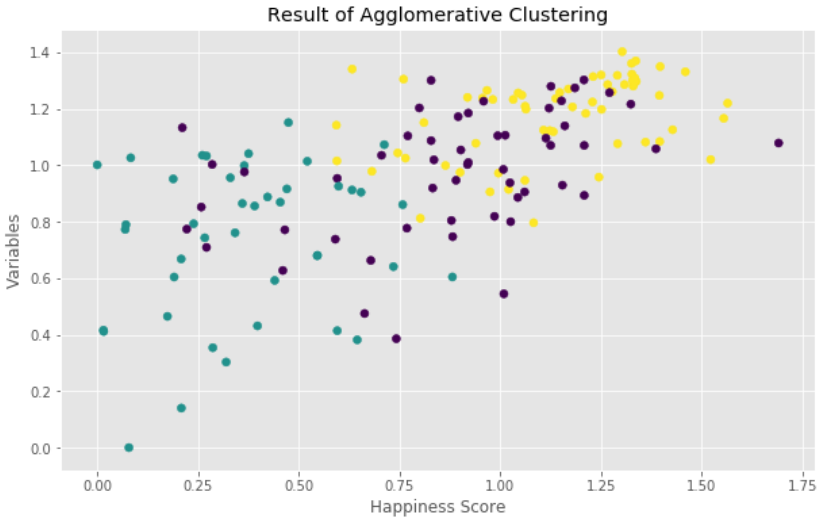

Figure 4.3: Result of Agglomerative Clustering

Looking into more details, the Figure 4.4 shows clustering by different variables and Happiness Score. It is shown that each of them is clear classified.

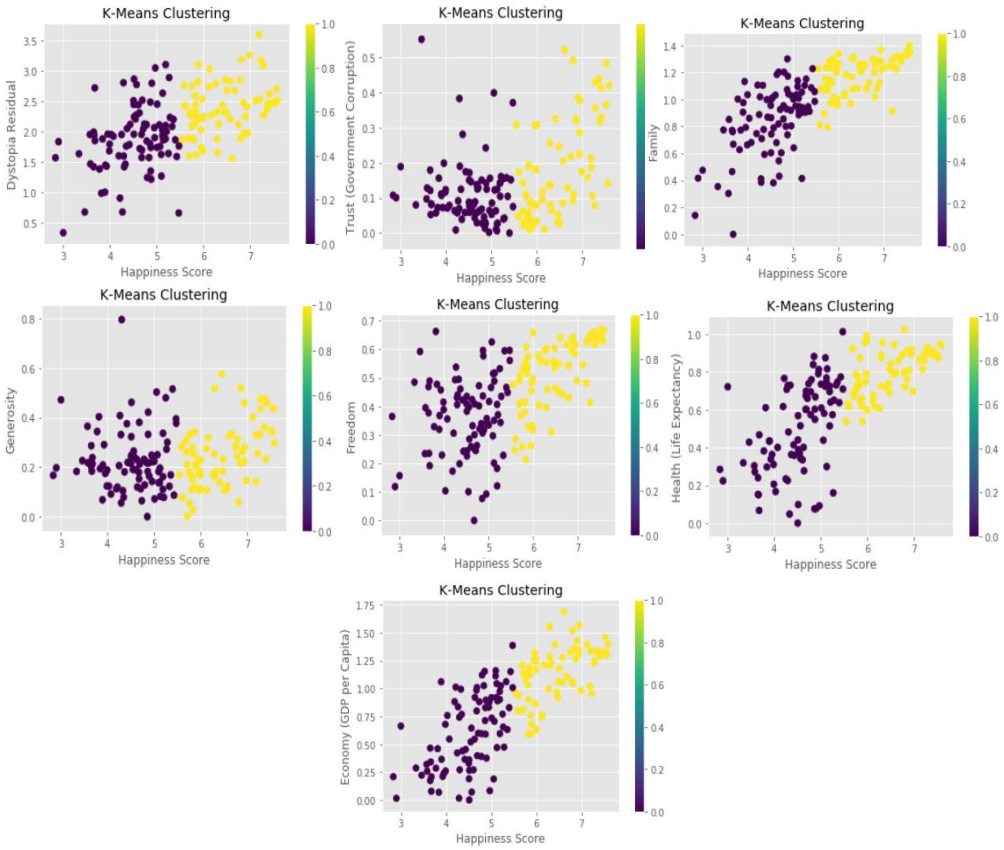

Figure 4.4: Scatter plot for clusters

4.2 Agglomerative Clustering

Agglomerative Clustering is another method used in unsupervised analysis. In this approach, each data is originally considered as an independent group (Reddy, 2018). After the process repeats several times, the clusters that are similar to each other integrate and stop when one cluster is formed (Reddy, 2018). It is a bottom-up approach.

There are seven different kinds of linkage algorithms (Xu, 2009):

i. **The single linkage algorithm**. It is also called the nearest neighbour method. The distance between different groups is minimised by two nearby clusters.
ii. **The complete linkage algorithm.** It measures the farthest distance among numbers of clusters to determine the distance. It is quite useful in small groups.
iii. **The group average linkage algorithm.** It is also named as the unweighted pair group method average (UPGMA). It use the mean of the distance among all data, which comes from different cluster individually.
iv. **The weighted average linkage algorithm.** It is known as the weighted pair group method average (WPGMA). The difference between this and the previous one is that its distance is measured by the amount of data points in individual group.
v. **The centroid linkage algorithm**. It is called the unweighted pair group method centroid (UPGMC). Two different clusters will integrate into one cluster by the distance of the average.
vi. **The median linkage algorithm.** It is also known as the weighted pair group method centroid (WPGMC). It is similar to the average linkage algorithm.
vii. **Ward' s method.** It also called the minimum variance method. This method is to minimise the total of the square errors.

Graph methods regard every points in number of clusters while measuring the distance, including single linkage, complete linkage and average linkage (Xu, 2009). The rest of them are named geometric methods using the geometric middles to calculate the distances (Xu, 2009).

There are three scores that will be calculated after using Agglomerative Clustering in Table 4.1.

1. Silhouette Score: It measures how each point is close to other similar one that is in the same cluster compared to another cluster. The range of this score is between -1 and 1. The best score is 1 and the worst is -1. If the value is near 0, it means overlapping groups.

2. Completeness Score: Determine the percentage of how the points are assigned into the correct clusters. The range of this score is from 0 to 1. 1 means that the all data is in the right groups.

3. Homogeneity Score: The percentage of data points that are in the same cluster with same class. It is bounded between 0 and 1, where 1 performance there is single type in a cluster.

From the Table 4.1, it is indicated that the Completeness Score is almost 1, which means all data is assigned into the correct clusters. The other two score is just about 20%.

Table 4.1: Comparison of Agglomerative Methods

Score	**Outcome**
Silhouette Score	0.201798019674489
Completeness Score	1.0000000000000004
Homogeneity Score	0.18942238841663253

Chapter 5

Supervised Analysis

Supervised approaches use algorithms which examples are already labelled among a training dataset to predict the potential inputs (Igual, Laura. ; Seguí, Santi., 2017). There are several general methods in supervised analysis, including linear regression, Logistic regression, K nearest neighbours, Decision tree, Naïve Bayes, etc.

In this case, the aim of analysing this dataset is to determine what factor plays an important role in the Happiness Score. There are four supervised methods used in this report, linear regression, logistic regression, decision tree and K-Nearest Neighbours. The percentage of testing data is 30% of the total.

5.1 Linear Regression

Linear Regression use given variables (X) to predict the value of Y. In this case, the Y is the Happiness Score, and the X is all variables, such as Economy, Health and so on. Here, the KFold is set as 5 to valid the data. The result is shown in the Table 5.1. The Mean Absolute Error means the proportion of how the prediction is not correct by the linear regression. Mean Squared Error has better performance when the squared error is low. Higher R^2 which is close to 1 is the best. As the result shown below (Table5.1), the R^2 is too small, only around 20%, which means the model might not be able to predict the value.

Table 5.1: Result of Linear Regression

	Linear Regression
Mean Absolute Error	-0.2738918041236298
Mean Squared Error	-0.1044601285727718
R^2	0.19073672283119275

5.2 Logistic Regression

Logistic Regression is an algorithm that categorise data points into two different groups. In this case, there is no obvious classification, however we can still run the logistic regression.

There are three important measurements in logistic regression (Joshi, 2016):

(1) Precision: The proportion of how positive the predicted observations is compared to the total one. High precision means that the false positive rate is low.
(2) Recall: The ratio of how positive the predicted observations is compared to the total one in real class.
(3) F1-score: The aggravated mean of precision and recall.

Figure 5.1 shows the result of logistic regression. It is shown that the score of 3 and 7 is not correctly categorised through this regression. The score of 5 presents the highest proportion of the dataset.

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.45	0.75	0.56	12
5	0.81	0.62	0.70	21
6	0.25	0.75	0.38	4
7	0.00	0.00	0.00	7
accuracy			0.52	48
macro avg	0.30	0.42	0.33	48
weighted avg	0.49	0.52	0.48	48

Figure 5.1: Result of Logistic Regression

5.3 Decision Tree

Decision Tree method is to spilt the dataset into small pieces and try to build a model that fits to those pieces. There are two kinds of decision tree, Classification trees and Regression trees. Classification trees use classified variables and see how they fall into different classes. Regression trees use continuous variables and predict the values. In this case, we use Classification trees to see if the dataset can be clear classified. The performance of Classification trees is shown in the Figure 5.2. The data is categorised into six classes, from 2 to 7. It is indicated that only data which score is 6 has low precision and recall in this method.

	precision	recall	f1-score	support
2	1.00	1.00	1.00	1
3	0.67	0.67	0.67	6
4	0.60	0.64	0.62	14
5	0.64	0.41	0.50	17
6	0.30	0.50	0.37	6
7	0.60	0.75	0.67	4
accuracy			0.56	48
macro avg	0.63	0.66	0.64	48
weighted avg	0.59	0.56	0.56	48

Figure 5.2: Result of Classification Trees

5.4 K-Nearest Neighbours

K-Nearest Neighbours method is to data points that are near each other and calculate the mean of them as their scores. The algorithm is based on Euclidean Distance. In this case, the result is shown in the Figure 5.3. It is indicated that only the group of score 2 does not performance well in this method. The other groups present much higher proportions.

	precision	recall	f1-score	support
2	0.00	0.00	0.00	1
3	0.71	0.83	0.77	6
4	0.87	0.87	0.87	15
5	0.88	0.88	0.88	16
6	0.71	0.83	0.77	6
7	1.00	0.75	0.86	4
accuracy			0.83	48
macro avg	0.70	0.69	0.69	48
weighted avg	0.82	0.83	0.83	48

Figure 5.3: Result of K-Nearest Neighbours

5.5 Discussion

Among these supervised methods, it is shown that Classification trees and FK-Nearest Neighbours seem to be appropriate to analyse this dataset. In linear regression, if the KFold is set to less than 10 times, the R^2 score will increase, and the predicted model will be more effective. Moreover, different approaches will classify the data into different numbers of classes. For example, some is categorised into 6 groups and some is 5 groups.

Chapter 6

Reflections

The dataset used in this report does not contain lots of information. It might be possible if I could compare the changes between years. However, the variables used in each year is also not the same, it might be difficult to compare them in the same level.

The dataset does not have a clear classification for each variable. As a result, it might be difficult to cluster the data. For example, when clustering the Happiness Score (target value) and all variables, it is quite difficult to classify each cluster (see Figure 4.3). However, if you cluster the Happiness Score (target value) with each variable separately, it is shown as clear classification.

Chapter 7

Conclusion

Although the happiness investigation results are regard as an important issue in the world, the calculation of Happiness Score still not reach an agreement. Some experts believe that the survey from people is not objective enough to represent their countries' score.

Moreover, there is still numbers of factors needed to be considered into calculating the Happiness Score. For example, unemployment rate, which can show a labour market of a country. The dataset might also provide the proportion of investigation in gender and age. As a consequence, we can also see the difference between generation.

In unsupervised analysis, it is better to cluster the data into two or three groups. If we just look at the Happiness Score with one variable each time, it is shown as a clear classification.

In terms of supervised analysis, it is appropriate to use K-Nearest Neighbours or Decision Trees to analyse to data.

Overall, the import factors that play a significant role in Happiness Score is easily to observe. However, the finding that each region has different factors that affect their Happiness Score is an issue that can be discussed furthermore by their government.

Appendix A

Environment

Python version: Python 3.6.5

Dataset derived from: https://www.kaggle.com/unsdsn/world-happiness

Packages used:

- pandas
- numpy
- matplotlib.pyplot
- seaborn
- sklearn
- cluster from sklearn
- metrics from sklearn
- mixture from sklearn
- datasets from sklearn
- scale from sklearn.preprocessing
- LabelEncoder from sklearn.preprocessing
- KMeans from sklearn.cluster
- LogisticRegression from sklearn.linear_model
- LinearRegression from sklearn.linear_model
- train_test_split from sklearn.model_selection
- KFold from sklearn.model_selection
- DecisionTreeClassifier from sklearn.tree
- KNeighborsClassifier from sklearn.neighbors
- GaussianNB from sklearn.naive_bayes

Reference

Igual, Laura. ; Seguí, Santi. (2017). *Introduction to data science : a Python approach to concepts, techniques and applications.* Cham, Switzerland: Springer.

John Helliwell, Richard Layard and Jeffrey Sachs. (2015, 4 23). *World Happiness Report 2015*. Retrieved from World Happiness Report: https://worldhappiness.report/ed/2015/

Joshi, R. (2016, 9 9). *Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures*. Retrieved from EXSILIO Solutions: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Kurama, V. (2019, 6 7). *HOW TO USE UNSUPERVISED LEARNING WITH PYTHON TO FIND PATTERNS IN DATA*. Retrieved from builtin: https://builtin.com/data-science/unsupervised-learning-python

Reddy, C. (2018, 12 20). *Understanding the concept of Hierarchical clustering Technique*. Retrieved from Medium: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Xu, R. (2009). *Clustering.* New Jersey: John Wiley & Sons.