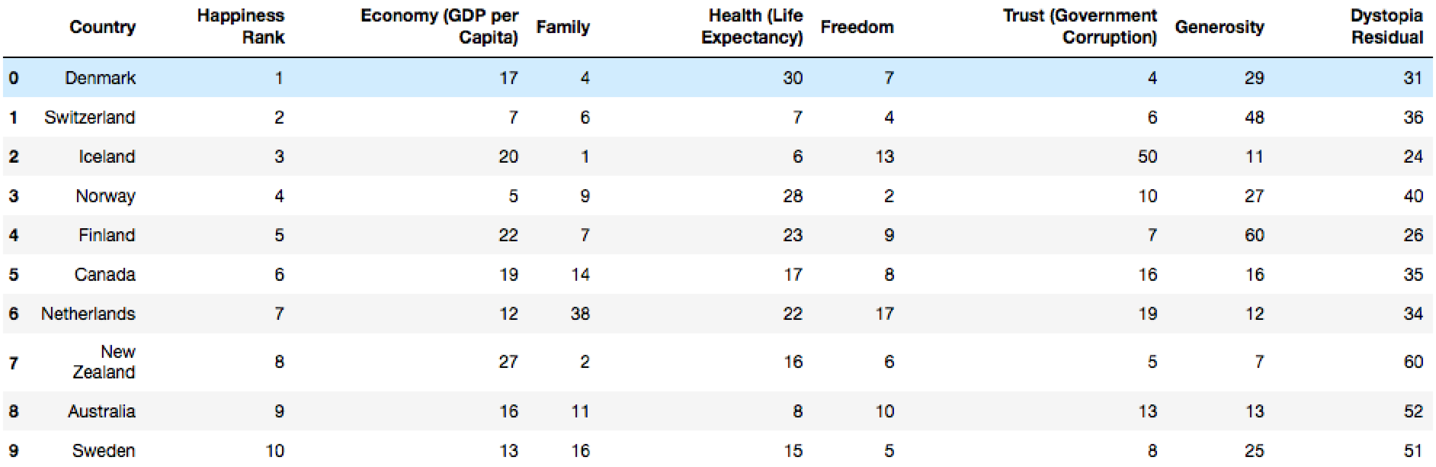
RMBI3110 Assignment 2 --- Clustering

Li Chung Yat
Student ID: 20422979
Lo Hau Yi
Student ID:20421470

Perform Explanatory Data Analysis

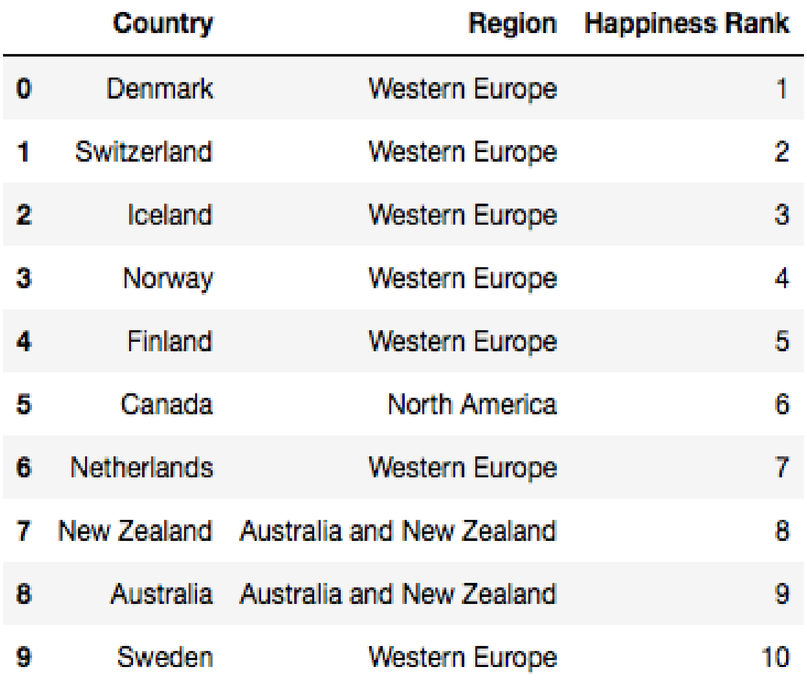
Data Exploration:
This dataset mainly focused on happiness rank of different countries and the seven measures to calculate happiness scores of these countries. There is a total of 157 countries included. The 7 happiness score estimators will be target variables to carry out clustering analysis. It is because attribute ‘Country’ and ‘Region’ will be our dependent variable and ‘Happiness score’ and its confidence intervals are derived by these predictors which will cause high correlation if they are deemed as target variables. There are no missing value or unreasonable value and they are all numerical data, ranging from 0 to 1.2 except ‘Dystopia Residual’.

Data Analysis:
Considering the top 10 happiest countries and the rankings of their estimators,



	Country	Happiness Rank	Economy (GDP per Capita)	Family	Health (Life Expectancy)	Freedom	Trust (Government Corruption)	Generosity	Dystopia Residual
0	Denmark	1	17	4	30	7	4	29	31
1	Switzerland	2	7	6	7	4	6	48	36
2	Iceland	3	20	1	6	13	50	11	24
3	Norway	4	5	9	28	2	10	27	40
4	Finland	5	22	7	23	9	7	60	26
5	Canada	6	19	14	17	8	16	16	35
6	Netherlands	7	12	38	22	17	19	12	34
7	New Zealand	8	27	2	16	6	5	7	60
8	Australia	9	16	11	8	10	13	13	52
9	Sweden	10	13	16	15	5	8	25	51

Most of the top 10 countries that ranks high in ‘Family’, ‘Freedom’ and ‘Trust (Government Corruption)’ Also tends to have good performance in ‘Economy (GDP per Capita)’ and ‘Health (Life Expectancy)’ Relevancy of ‘Generosity’ and ‘Dystopia Residual’ with Happiness Rank are not conclusive using this table, using clustering could probably give more insight on these 2 attributes.



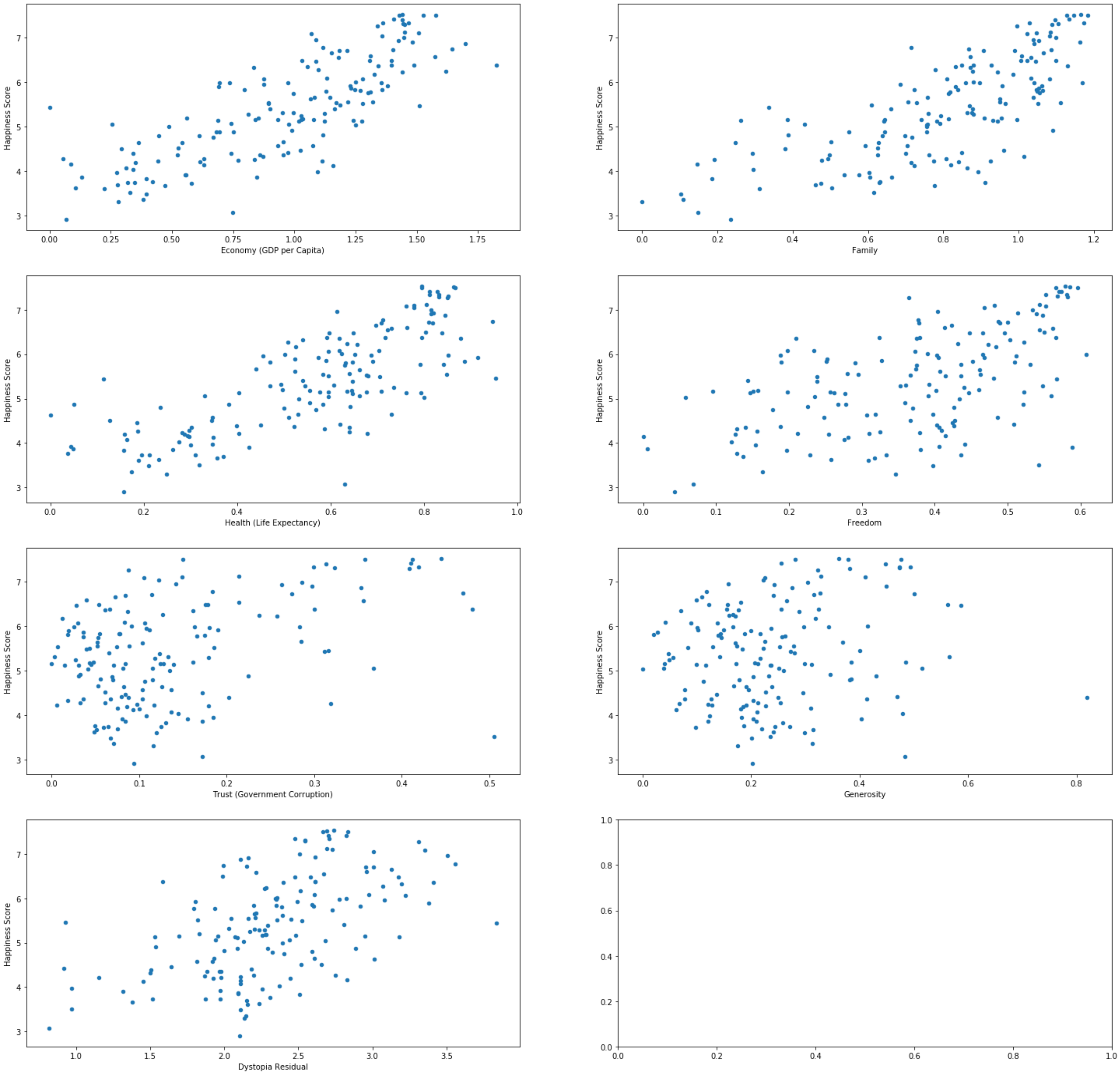
	Country	Region	Happiness Rank
0	Denmark	Western Europe	1
1	Switzerland	Western Europe	2
2	Iceland	Western Europe	3
3	Norway	Western Europe	4
4	Finland	Western Europe	5
5	Canada	North America	6
6	Netherlands	Western Europe	7
7	New Zealand	Australia and New Zealand	8
8	Australia	Australia and New Zealand	9
9	Sweden	Western Europe	10

The table on the left is to study the Region where top 10 happy countries located.

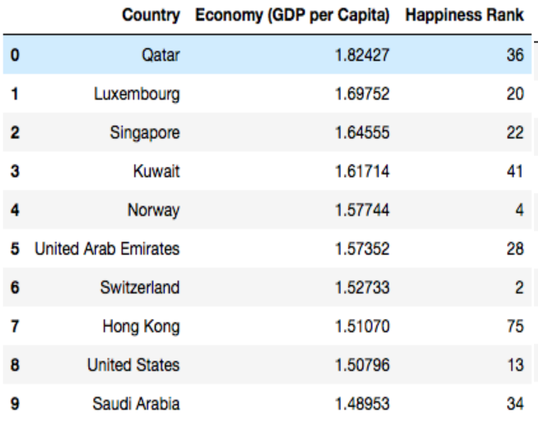
The finding is interesting because 7 out of 10 happiest countries are from Western Europe. Consider the estimators ranking table from the above, we can observe that they all rank top 20 in ‘Family’, ‘Freedom’ and ‘Trust’ predictors. Except Iceland in which their citizens have low trust towards the government and it makes sense as Iceland government is facing bankruptcy.

As both Australia and New Zealand are the happiest countries, we could interpret that region Australia and New Zealand is one of the happiest regions among the World.

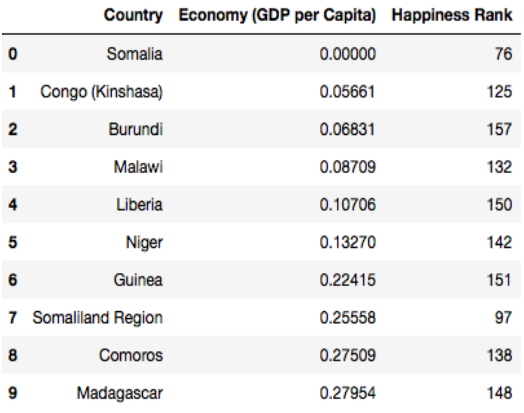
By conducting scatter plot, we can study the relationship between Happiness Score and the 7 estimators,



We can observe that ‘Economy’, ‘Family’, ‘Health’, ‘Freedom’ and ‘Dystopia Residual’ have some linearity with Happiness Score. Whereas, ‘Trust’ and ‘Generosity’ seems to have only a little relationship. However, this observation is contradictory to results we obtained from the top 10 happiest countries table. Some of the top 10 happiest countries ranking high in ‘Trust’ is actually special cases and it is not comprehensive as it cannot explain some countries ranking low in ‘Trust’ but still place high in happiness ranking.



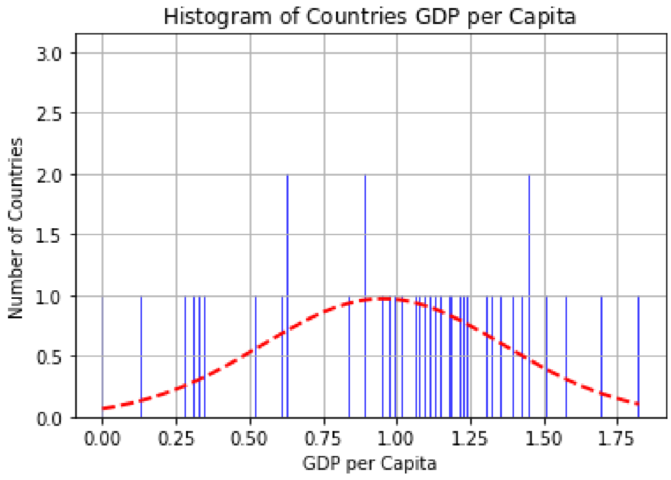
	Country	Economy (GDP per Capita)	Happiness Rank
0	Qatar	1.82427	36
1	Luxembourg	1.69752	20
2	Singapore	1.64555	22
3	Kuwait	1.61714	41
4	Norway	1.57744	4
5	United Arab Emirates	1.57352	28
6	Switzerland	1.52733	2
7	Hong Kong	1.51070	75
8	United States	1.50796	13
9	Saudi Arabia	1.48953	34

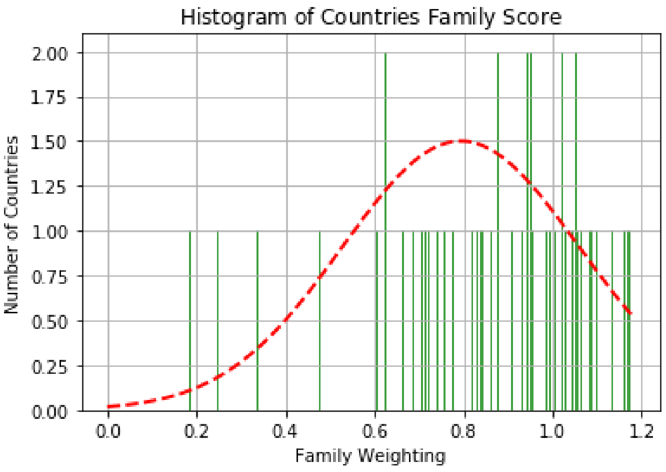


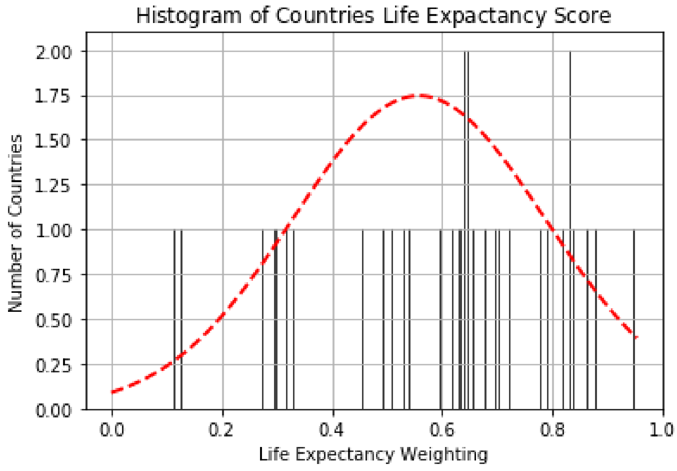
	Country	Economy (GDP per Capita)	Happiness Rank
0	Somalia	0.00000	76
1	Congo (Kinshasa)	0.05661	125
2	Burundi	0.06831	157
3	Malawi	0.08709	132
4	Liberia	0.10706	150
5	Niger	0.13270	142
6	Guinea	0.22415	151
7	Somaliland Region	0.25558	97
8	Comoros	0.27509	138
9	Madagascar	0.27954	148

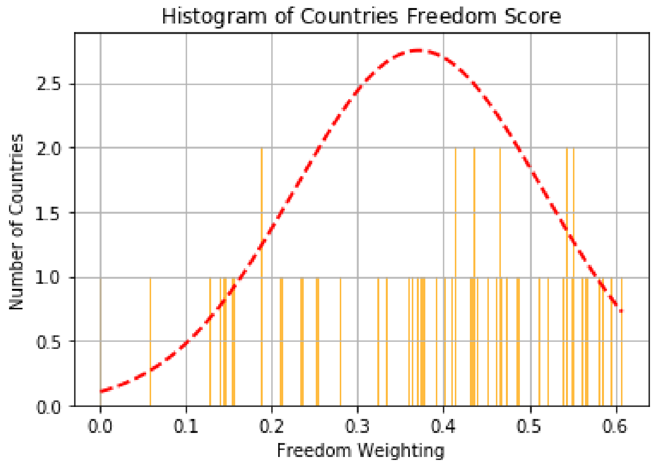
These two tables shown the top 10 richest and poorest countries. The topic on whether money could buy happiness has always been widely discussed, based on this Happiness Rank research’s result. It is fair to conclude that to some extent money is related to our happiness.

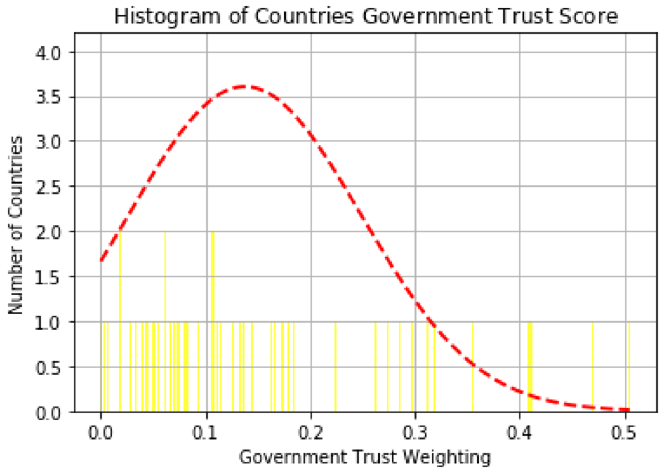
Skewness check:
A normally distributed attribute in dataset will be ideal for performing clustering analysis. By checking the distribution of data in each estimator using histogram, we can find out the attributes that are skewed.

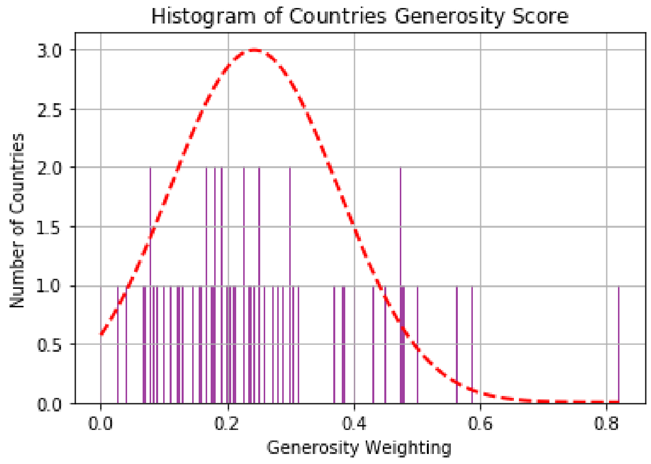

Histogram of Countries GDP per Capita

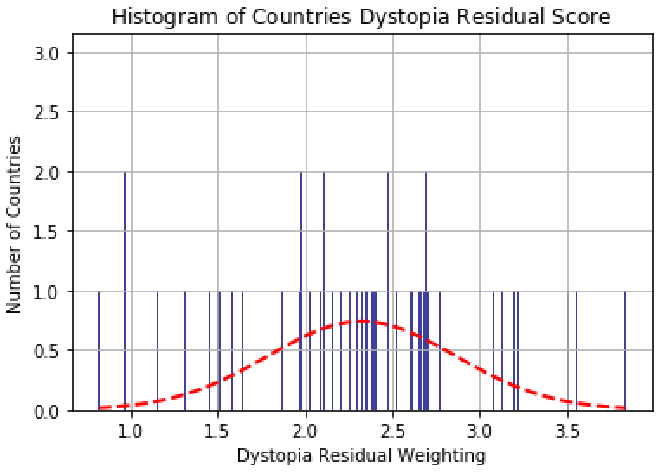

Histogram of Countries Family Score


Histogram of Countries Life Expactancy Score


Histogram of Countries Freedom Score


Histogram of Countries Government Trust Score


Histogram of Countries Generosity Score

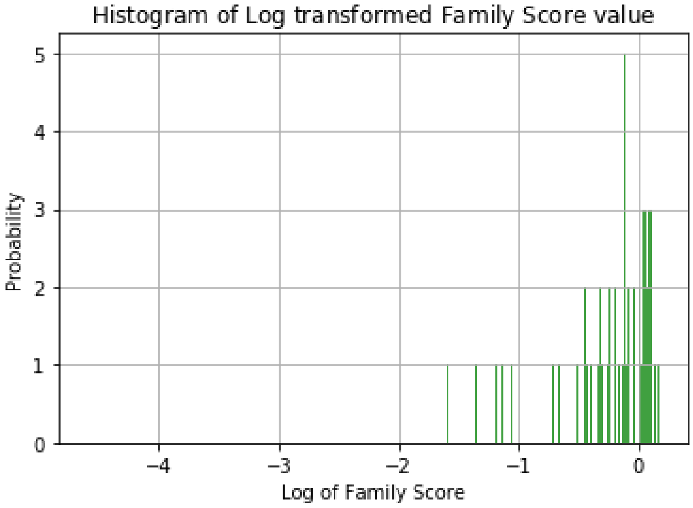

Histogram of Countries Dystopia Residual Score

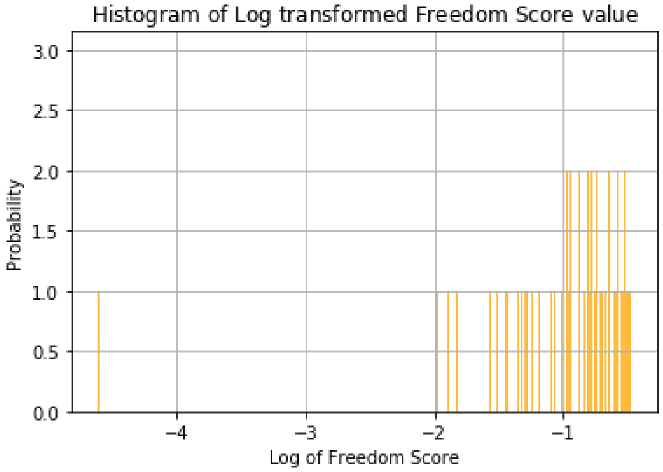
The 7 normality plots show that ‘Family’, ‘Life Expectancy’, ‘Freedom’, ‘Government Trust’ and ‘Generosity’ attributes are skewed.

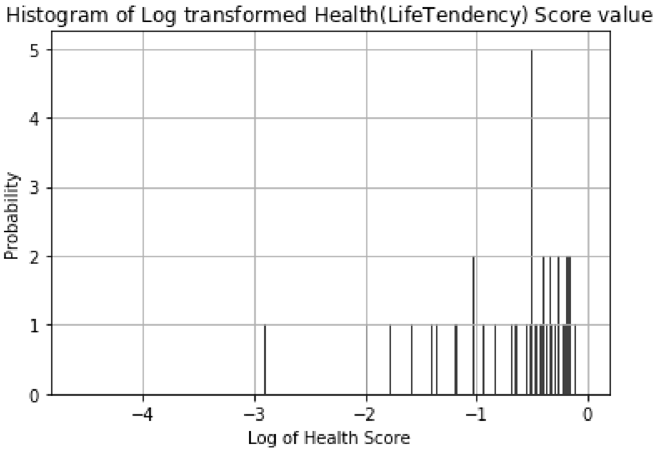
There exists extreme value in ‘Government Trust’ and ‘Generosity’ which is a good indicator that we should take log on these predictors’ value.

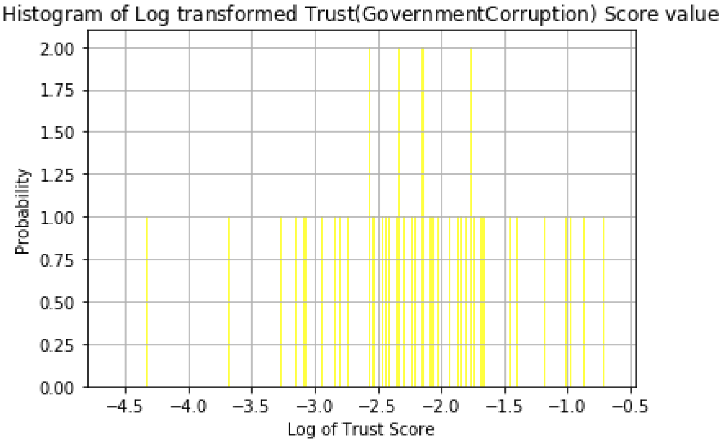
‘Economy (GDP per Capita)’ and ‘Dystopia Residual’ are normally distributed.

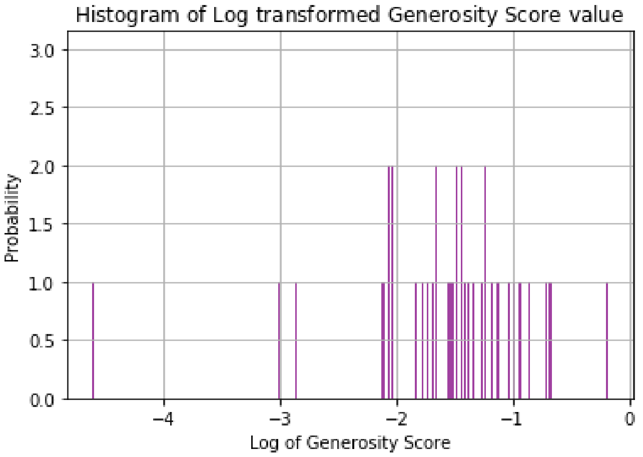
We decide to perform logarithmic transformation on all of these skewed attributes and plot their probability distribution graph again to see whether data transformation brings us improvement or drawbacks on the data set. Before taking log on these attributes, I have print out the summary statistics of these attributes and observed that all these numerical estimators contain a 0 which will be a deterrent for our logarithm. Therefore, I add 0.01 to each country in each estimators. Using the log value of skewed dataset, the plots are as follow:











‘Family’, ‘Freedom’ and ‘Health’ clearly show that it is much more skewed after taking log, it is because raw data of these attributes are high dense in high score, i.e. left skewed, and Python package’s math log seems to have strong log power in large extreme value but not for small one. The 0.01 value in each attribute becomes a large negative value (relatively speaking).

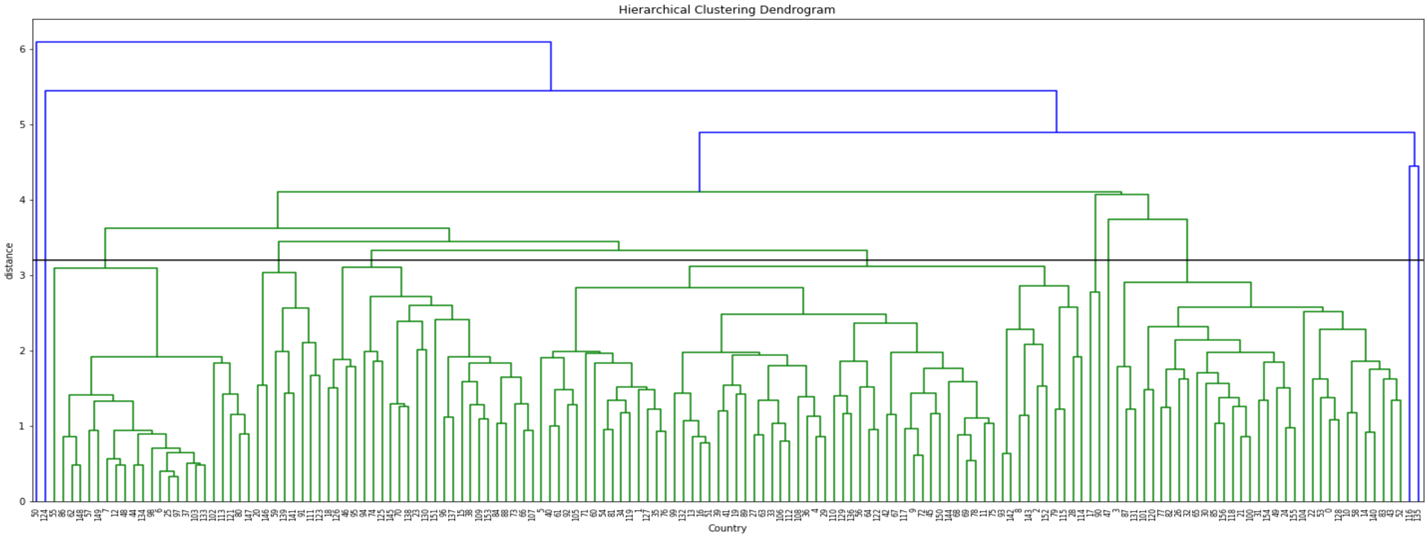
‘Trust’ and ‘Generosity’ are normalized. For these 2 estimators, our finalized dataset to perform clustering will replace the raw value by value after log.

Standardization of data:
We did not perform standardization on this dataset as the data range do not exhibit extreme distances. When calculating distance or similarity measures, no attribute will be dominating others. Furthermore, it is unnecessary to reduce the effect of outliers bring to the dataset as clustering can filter out outliers easily.

No dropping or combination of estimators as well. Each of the measures are uniquely meaningful. Like grouping ‘Trust’, ‘Economy’ and ‘Freedom’ as ‘Government performance’ is reasonable but this will lose some meanings from these 3 underlying attributes. Although clustering does not have good performance in high dimension dataset, the business insight carried along in these estimators are more valuable.

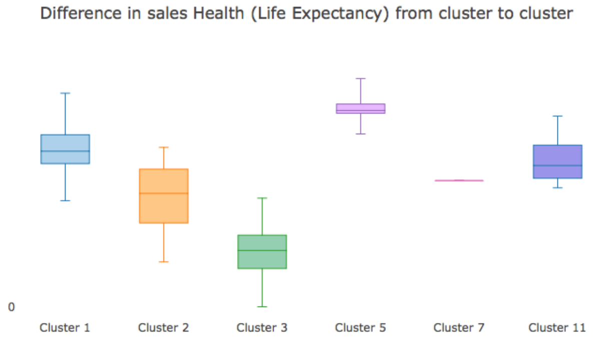
Perform Hierarchical clustering. Using Dendrogram and elbow plot to select the optimal number of clusters. Compare the results.

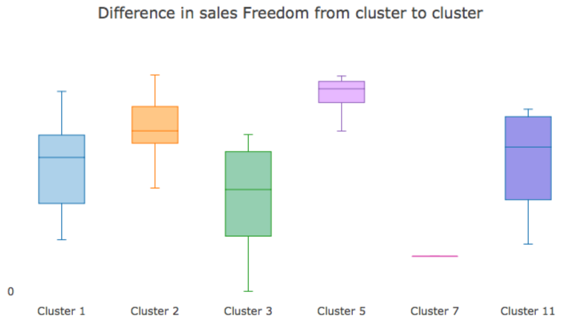
Dendrogram:

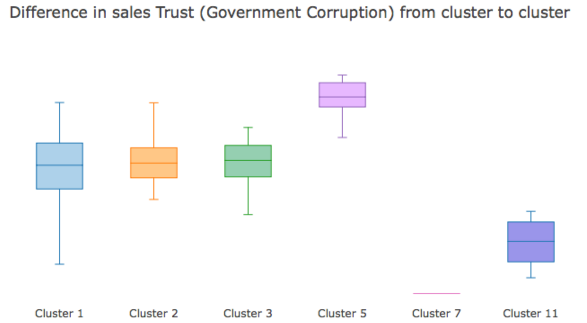


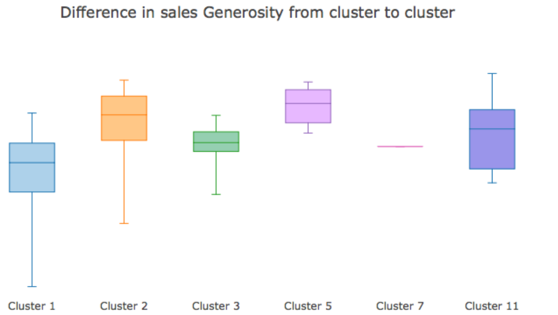
11 clusters can be observed when maximum distance is set at 3.2 which seems to minimise the sum of square error within clusters. 5 outliers will be resulted.

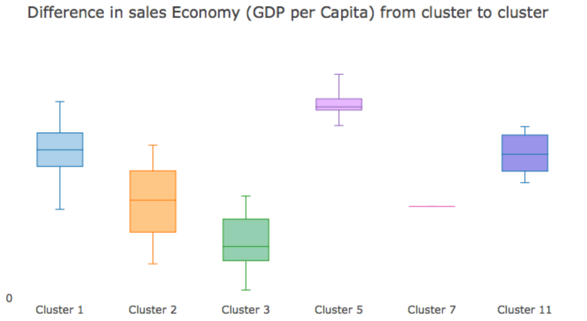
Clustering Result:

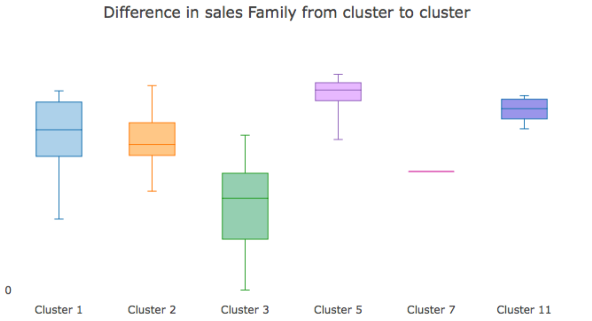


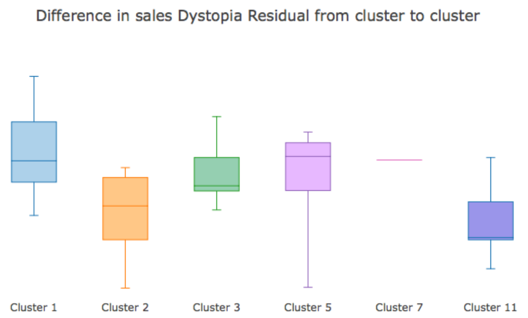












As cluster 7 only contains 2 samples, I will deem it as outlier as well in this case. We have 5 clusters at the end.
Counties belong to Cluster 1 have significant features which are long life expectancy, good economy and high dystopia. And all predictors of cluster 2's countries are ranged around average or a bit above average. Cluster 2 should be countries ranking in the middle in happiness rank.
Cluster 3 are countries with low freedom, health, economy and family score but generous and trustful to their government. I suspect these are the Middle East or Africa countries.

Cluster 5 and 11 tends to give more significant features. Cluster 5 has high score in all attributes except in 'Dystopia Residual'. This cluster has high similarity with the Western European countries group mentioned at the very beginning. Whereas, cluster 11 has decent health, family, GDP and generosity but low in government trust and dystopia while ranging large in freedom score. I suspect this cluster is a hybrid of some European countries and East Asia countries. The similarities and differences of some countries in these two regions suit the cluster description well. Like Hong Kong and Spain, strong economic power, good medical supports, low trust in government, family-based culture and people disfavor of dystopia but Spain enjoy more freedom and Hong Kong does not. This may explain the high variety in the freedom score.

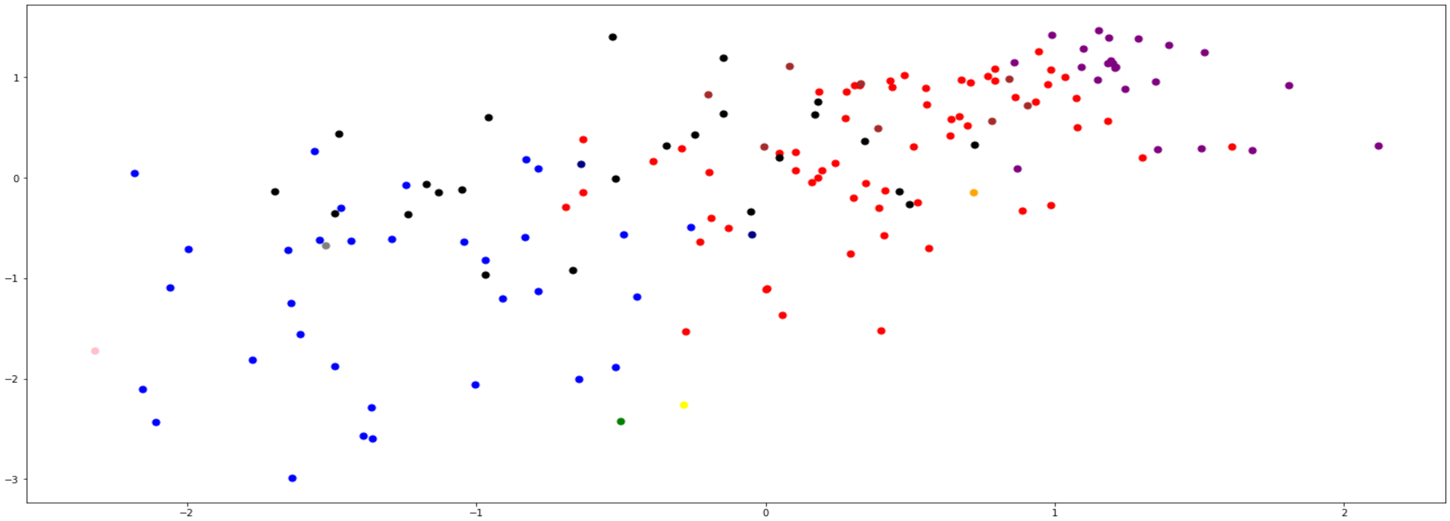
Cluster 1 is a large cluster including a total of 62 countries. Mean Happiness Score is 5.82. By comparing the mean Happiness Score in this cluster to the whole population, Cluster 1 has z-score = 0.376 and implies that people living in Cluster 1's countries are happier than average population. This is a big group of 'happy' countries, it is hard for a machine to cluster this group and is understandable that the result is not satisfactory. Happiness variate between different people, some people will deem money as an important criterion, but some will not. We can clearly observe that this big group has high variety in all attributes and we would describe this group as unexplainable cluster using algorithm.

Cluster 2 has mean Happiness Score of 4.83 and z-score = -0.485. This cluster group is expected to be the countries having Happiness ranking above 80. Countries are mainly from Southern or South-East Asia and Sub-Saharan Africa. Most of these countries are developing countries, like Vietnam, Philippines and South Africa. These countries' economy is still growing but not as poor as countries in undeveloped wild jungle, enjoying democracy and people are more generous, all these features well-fit the clustering results.

Cluster 3 has mean Happiness Score of 4.06 and z-score = -1.16 Cluster 3 are referring to group of countries that has the unhappiest citizens as 4.06 can only rank 135 in Happiness Rank in the research. If we take a look on the region of these countries belong to, almost all of them are from South Asia and Sub-Saharan Africa. The result of this cluster is satisfactory as the real-life implication is reasonable and logical that the region is still undeveloped and have low life expectancy. Some of these countries are democratic but some are rule under absolute power, thus, there exist high variety in 'Freedom' score.

Cluster 4 is the happiest countries that we have discussed on, i.e. the Western European and Australia region group. The result of this cluster is obvious as we can see the result from the significant features of this cluster is highly similar to the top 10 happiest countries we discussed on at the beginning.

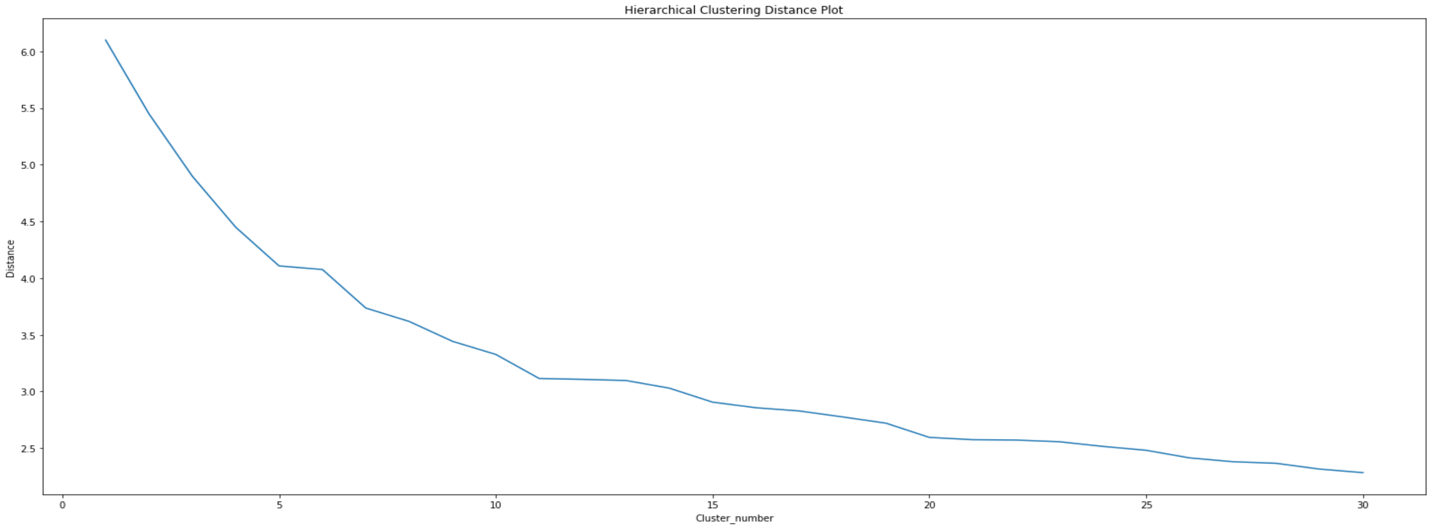
Cluster 11 is a small group of countries that are scatter and no obvious observation can be conclude except the low trust in government. These countries have the lowest score in 'Trust (Government corruption) attribute, I believe it implies this cluster are countries that have low happiness due to corruption Mean value of happiness is below average. Thailand, Indonesia, Ukraine are all countries that suffer political power rivalry among countries and parties.



The graph above show how cluster is distributed in dataset. It can only provide little observational conclusion as it is plotted based on only 2 out of the 7 dimensions.

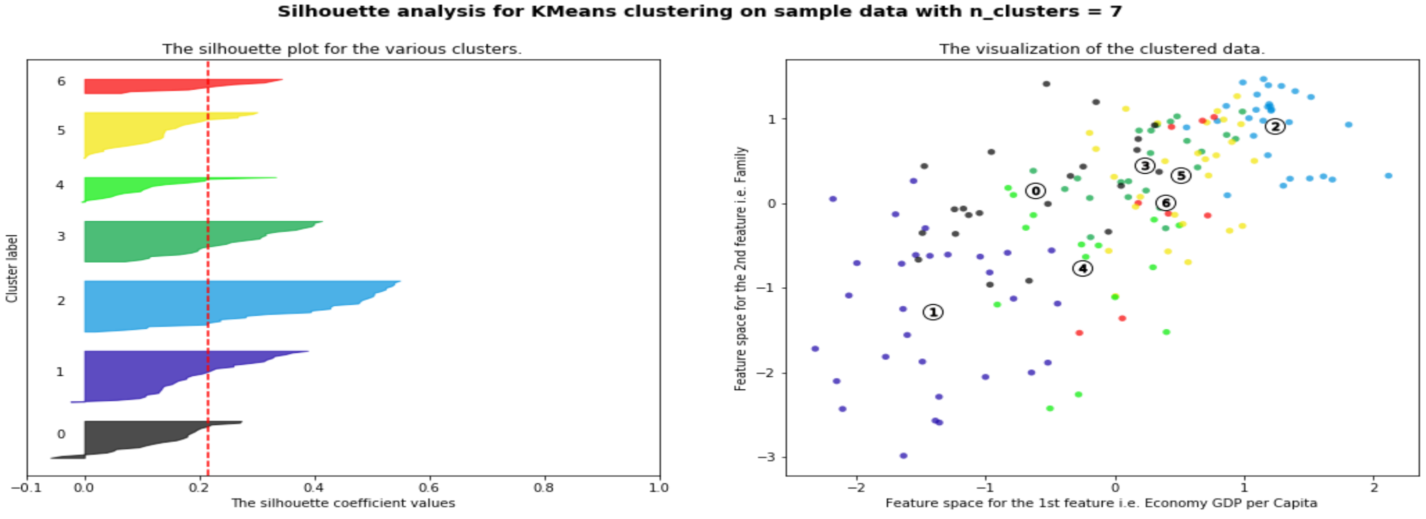
Perform K-means. Using Elbow plot and silhouette analysis to determine the optimal number of clusters. Compare the results.

Elbow Plot:



As we can see in elbow plot, the twisting point is around 5 to 10.

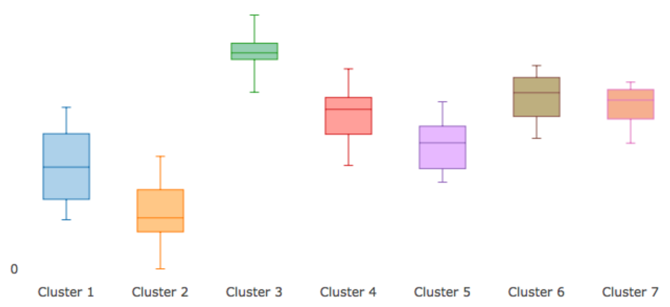
Silhouette:



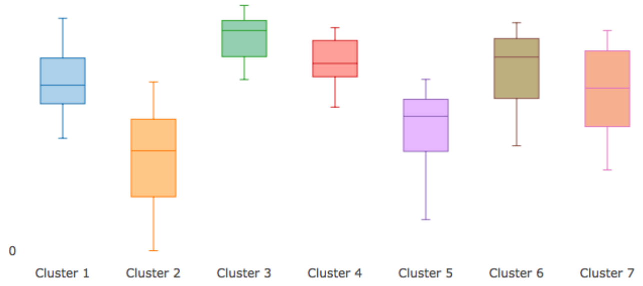
When number of cluster equals to 7, silhouette will be maximum. But K = 6 also gives similar result, I decided to plot both of them out and see which one give better interpretation of real-life data. We both think K = 7 provides better business insight and meaning to the dataset.

Clustering Result:

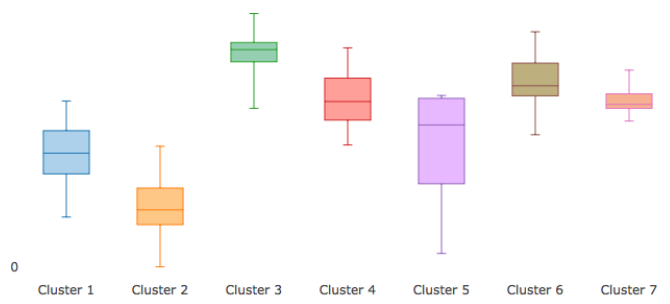
Difference in sales Economy (GDP per Capita) from cluster to cluster



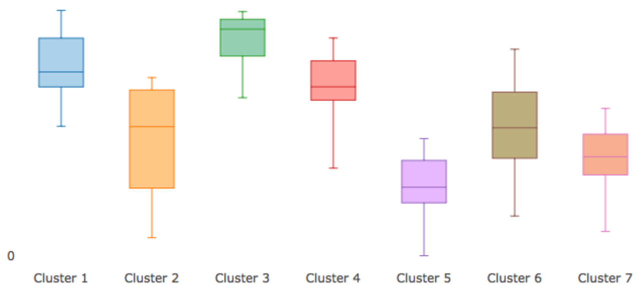
Difference in sales Family from cluster to cluster



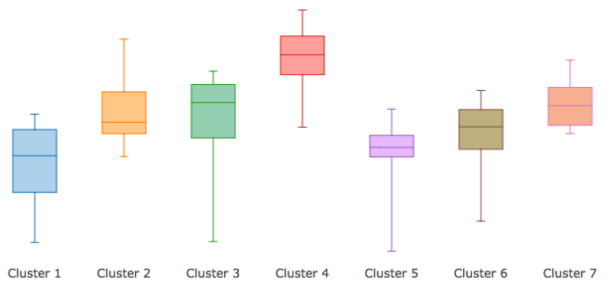
Difference in sales Health (Life Expectancy) from cluster to cluster



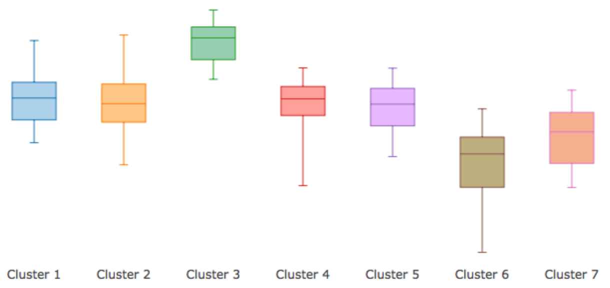
Difference in sales Freedom from cluster to cluster



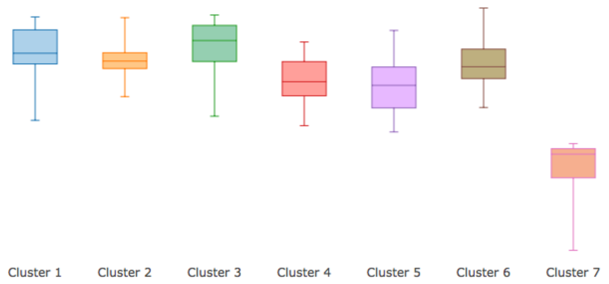
Difference in sales Dystopia Residual from cluster to cluster



Difference in sales Trust (Government Corruption) from cluster to cluster



Difference in sales Generosity from cluster to cluster



As we have already performed hierarchical clustering, we suspect the clustering results will not provide much difference, but we hope to see if K-Mean can provide a reasonable result on the unexplained cluster in our hierarchical clustering.

Cluster 1 in K-Mean has high similarity to the Cluster 2 in hierarchical clustering which has happiness score below average and majority of these countries located in Southern or South-East Asia and Sub-Saharan Africa. Mean value drop by a small value, but in conclusion they are referring to the same group of countries in the dataset.

Similarly, we can observe that Cluster 2 is the unhappiest group we had referred to in the dataset, most of them are from Sub-Saharan Africa. Mean value also increases by a bit but not significantly. Whereas, Cluster 3 is the happiest Western European Countries. There are some new data observed and worth noting about in which they are all from the unexplained cluster. All of these new data points do not belong to Western Europe but in Middle East, like Saudi Arabia and Kuwait. It is believed that they should have high similarity with Qatar where is being developed in high speed and rich in crude oil.

Cluster 4 are all belong to unexplained cluster before performing K-Mean. These countries have decent score in all areas of estimation and have the best score in Dystopia Residual. After further studying the data tabs, we found a surprising fact that most of these split data are from Latin America and Caribbean region. They score a relatively high average Happiness Score within the cluster, around 6.30. It is surprising that countries in this region score relatively high in Economy sector. But reasonable because there are lots of gangsters' group or privates in these areas which results in high dystopia score.

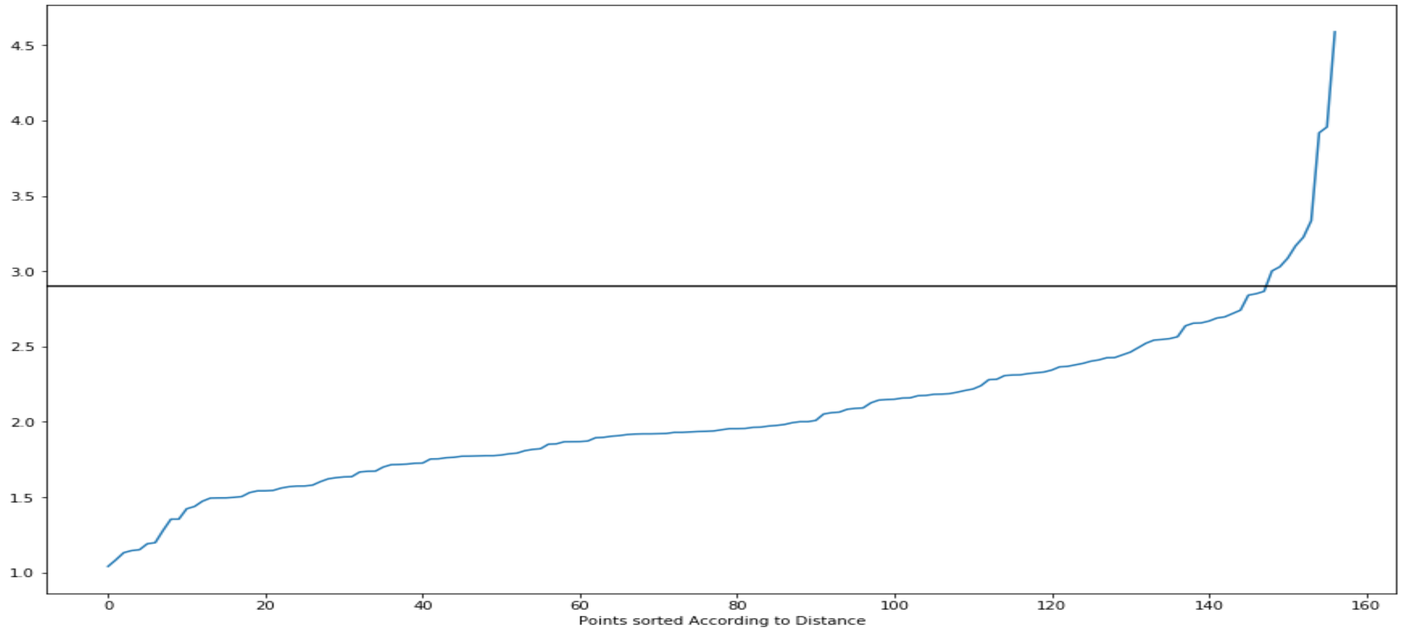
This cluster is the most interesting result I found which is a group of countries coming from different parts of Africa as well as Central and Eastern Europe. Cluster 5 has significant features on having low freedom and having most of the time as the group scoring below average. It is strange that there exist countries that score a little bit lower than average on all attributes but not being assigned to Cluster 3. After carrying out extra observation on these countries, we discovered that these are countries suffering from wars or terrorists such as Iraq, Syria and Georgia. All those Eastern European countries are actually very close to Middle-East or Africa and news reports reflected that Russia deem these 'Boarder land' as strategic focus area. People living in these countries are not happy for sure and average happiness index is about 4.38.

Cluster 6 citizens have low trust towards their government but rating quite high in all other aspect. This is quite similar to the Cluster 11 we have before. However, this new cluster also consist of half data points from undefined cluster we had. Compare this cluster to the Cluster 11 we had in hierarchical cluster, Cluster 11 can provide more information or insights to us, but this merged group cannot. It is like grouping countries with low trust to government with some unexplained countries which have high similarity. Despite the fact that new data like Spain and Taiwan do have some news or crisis showing that their governments have serious corruption, most of others are not capable to imply any real-life explanation. They are merged just because they have high similarity in some of the 7 attributes.

It is fair to describe Cluster 7 as not a generous group of countries. It is the most significant observation we can make in the graphs. However, the results reflect that there is little relationship between generosity and happiness as the mean happiness score value of this cluster is 5.43 which is above average. But to some extent that, it is essential to be generous so that you can be the happiest countries. Take a look into the cluster label, Russia and China are 2 of the big countries in the World that introduce and kind of adopting the communism. Based on the clustering result, a hypothesis is raised. 'Countries that introduces and adopts communism are in fact the most selfish countries in the World?'. It is ironic that communism is to promote the idea of sharing of wealth and helping out each other but in reality, these countries score the lowest in generosity.

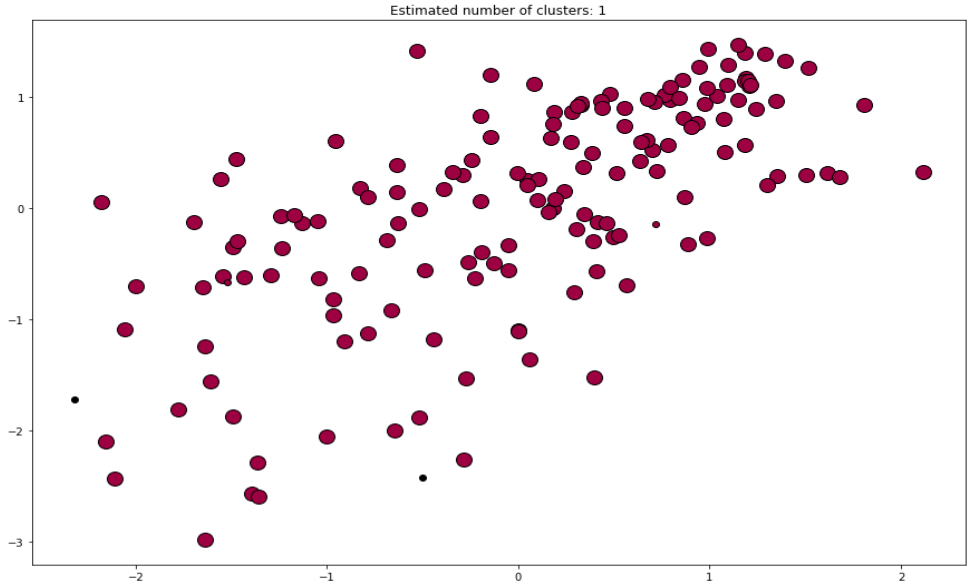
Use DBSCAN to cluster the data.

Eps and Minimum Point:
At the very beginning, we have to choose the number of minimum points within the Eps radius.
In common practice, number of minimum points chosen depends on the number of dimension we have in the dataset. In this case, we have 7 estimators in total. The number of minimum points will be given by 2*7 = 14.



In the above plot, we measure the distance of every point to its kth-nearest neighbour. The elbow point is ranging between 2.8 to 2.9. We picked 2.9 as the Eps radius for further processes.

Clustering Result:



The result is not satisfactory, all data points are grouped into 1 single cluster. The reason behind is that DBSCAN is not a suitable tools to handle high dimensional data. There are no interpretation we can made using a one-big cluster so they are not going to be shown or analysed as analysing a single cluster has no different in analysing our raw data during the data exploratory. There are 155 countries deemed as either core points or border points but only 2 are screened out as residuals.

The 2 noise points are Somalia and Syria. Both of them have unique characteristics in which Somalia is the only country infamous for privates and Syria are suffering in civil wars. These could be good explanations to high dissimilarity with others and not being included in the only cluster generated using DBSCAN.

Business implication:

In this assignment, we used 7 predictors to explain happiness scoring, which are life expectancy, freedom, trust, generosity, economy, family and Dystopia Residual respectively. All of these predictors have certain effect and have correlation to our predictor: happiness (and unhappiness). K-means clustering provides more and better business insights to our results, since it shows more clusters, as points that were in the same cluster in the hierarchical clustering was further divided in K-means, thus it provides more meaningful results for further analysis use. Also the observations found in hierarchical could also be seen in K-means clustering.

Based on the hierarchical clustering of clusters 3, it is observed that unhappy countries mostly have certain features, such as low life expectancy, low economy and low family. Then with results from cluster 11, it is shown that a low trust in government also contributes to unhappiness of a country.

Then, with reference to Hierarchical clustering and explanatory data analysis, it is seen that the happiest countries have a certain common features, including high life expectancy, great freedom, high trust in Government, high GDP per Capita and high family.

K-means clustering also provides new insight to our dataset. In order to be happiest, in fact having a higher generosity is necessary. In order to score an average score in happiness, the factor of having a high generosity do not correlate strongly though. Also, countries that adopt communism approach have an ironically low score on generosity. For countries suffering from wars or terrorists, the happiness scoring is lower than average. Trust in government also do not correlate strongly to a high average happiness scoring.

Aside from the predictors, it is observed that Western Europe and Australia and New Zealand are the happiest regions among the world, topping the top 10 countries.

The implication in our data explorations are as follow: in order to boost happiness to a top ranking within the globe from an average scoring, if I were the government of an average happiness scoring country, I would try to boost all of the indicators inside the study, such as boosting generosity among citizens, through improving quality of institutions and promoting a lawful society[1]. Freedom and Government trust could be improved through a more effective government and a more democratic society. A better family planning policy and incentive for giving birth could also increase the family indicator. Through policies as such, it could also help to promote freedom, family and government trust. These elements are essential to be enhanced to be above average on happiness scoring or reaching the top.

To achieve an average happiness score, if I were the government of a country with low happiness scoring, I would focus on bettering basic infrastructure: like improving health care policies to achieve higher life expectancy and having a good economy so that economy is better. Better family planning policies and improvement in government corruption situation is also necessary. Also, I would minimise negative events such as terrorist attacks and wars, that directly threatens citizen’s safety and life.

Aside from these, I would take reference from mainly the countries from Western Europe, Australia and New Zealand that topped the happiness scoring. Indeed, there are some similarities on their society and government policies, such as strong social support (good welfare policies), good governance (trustworthy and low corruption rate) and strong sense of community[2]. They are also developed countries with good basic infrastructure, often accompanied with free healthcare and education, and with the high sense of social security[3].

[1] https://greatergood.berkeley.edu/article/item/does_good_government_make_people_more_generous

[2] https://www.nationalgeographic.com/travel/top-10/2016-worlds-happiest-countries/

[3] https://www.inverse.com/article/42364-happiest-countries-in-world-nordic-region