BUS212A World Happiness Case

Yuzhou Liu, Yutian Lai, Leiyuxiang Wu

11/19/2018

Excutive Summmary

The main goal of this world happiness data research is to identified whether there are similarities among “happy countries”. Using 2015-2017 world happiness report data of 156 cocuntries, we cluster the coutries into 8 groups using 4 models with 2 clustering algorithm, K-means and Hierarchical. The final result concludes that the relative happier country group do have more similarities, thus support the words, that happy countries do seems alike.

World Happiness Dataset Intro

The Dataset comes from the World Happiness Report website, recording about observations from 156 countries and regions. Most of the variables are the country citizen responses to certain questions related to life happiness level, like *LifeLadder* and *SocSupp*. The other variables are simply meassures of country basic economics and demographic status, like *LnGDPpc* and *LifeExp*. As mentioned before, the purpose is to find whether there are similarities between “happy countries”, so the data is suitable and adequate for our goal.

```
## Observations: 156
## Variables: 20
## $ country      <chr> "Afghanistan", "Albania", "Algeria", "Angola", "A...
## $ region       <chr> "South Asia", "Central and Eastern Europe", "Midd...
## $ LifeLadder   <dbl> 3.631519, 4.586040, 5.294638, 3.794838, 6.387958,...
## $ change7      <dbl> -0.68833780, -0.79144096, -0.16892910, NA, 0.1124...
## $ forn_ladder  <dbl> 4.068487, NA, NA, NA, 5.843226, 4.100641, 7.24934...
## $ local_ladder <dbl> 3.853136, NA, NA, NA, 6.439717, 4.417696, 7.32390...
## $ SE_life      <dbl> 0.04235984, 0.05581279, 0.05708700, 0.07991932, 0...
## $ LnGDPpc      <dbl> 7.462610, 9.338126, 9.540703, 8.741957, 9.842001,...
## $ GDPpc        <dbl> 1741.6875, 11363.0957, 13914.7236, 6260.1328, 188...
## $ LifeExp      <dbl> 52.01333, 68.87155, 65.60486, 52.46071, 67.39848,...
## $ SocSupp      <dbl> 0.5250745, 0.6395764, 0.7769767, 0.7652755, 0.905...
## $ SEsoc        <dbl> 0.011118961, 0.010078066, 0.011170883, 0.01574368...
## $ LifeChoice   <dbl> 0.4452942, 0.7263402, 0.4391773, 0.3741727, 0.853...
## $ SEChoice     <dbl> 0.011448964, 0.009421178, 0.019161487, 0.01834626...
## $ Generosity   <dbl> 0.1790535, 0.2599754, 0.1289877, 0.1068286, 0.163...
## $ SEGen        <dbl> 0.008330869, 0.009048301, 0.012327445, 0.01149865...
## $ Corruption   <dbl> 0.8797045, 0.8867784, 0.6983430, 0.8335404, 0.847...
## $ SECorr       <dbl> 0.006009077, 0.005774963, 0.019292729, 0.01601409...
## $ OECD         <int> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ Power        <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

Data Prapration

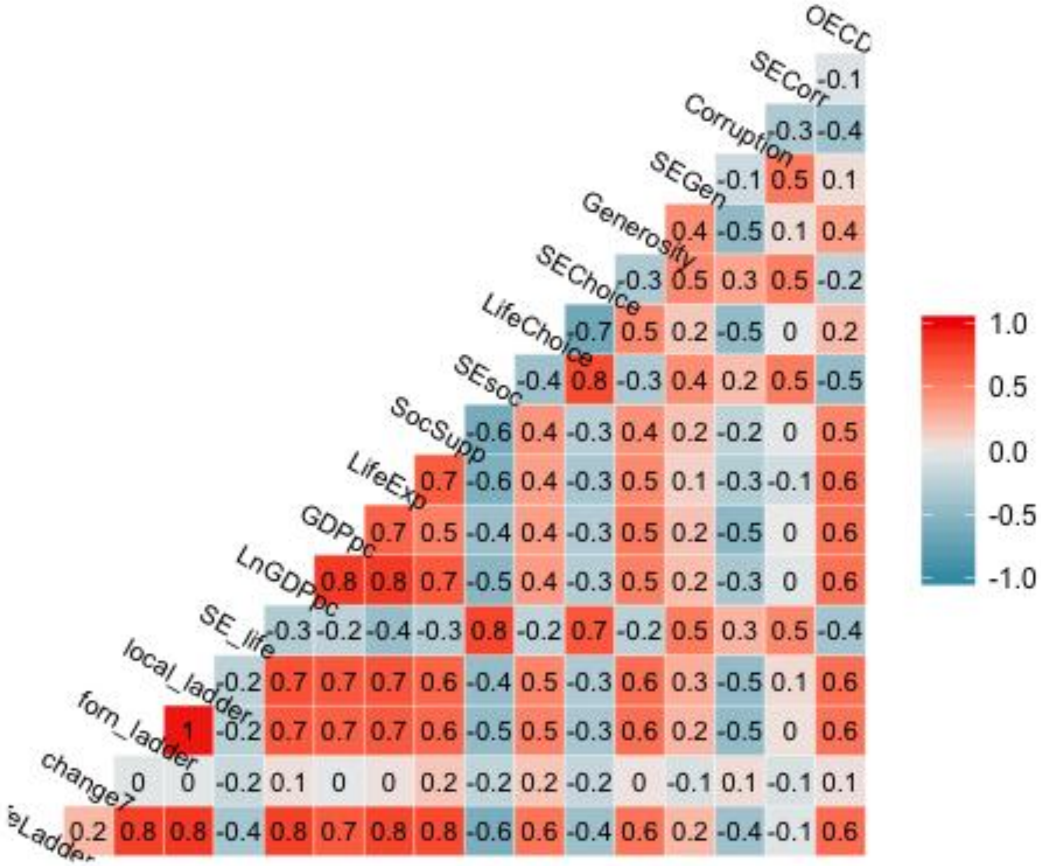
All the variables in the happiness dataset are in numberic from and suitable for the clustering alfrithem used, except for *region* as categorical.

First, we recode the *region* from 10 categories down to 6 categories including Asia, Europe, Africa, North American, South American (did not merge with North American because coutries has significant differences) and Commonwealth of Independent state. Then we break down the *region* to 5 dummies using Aisa as basic case.

Second, the numberic inputs has no extrem values, despite with some NA values. Since the clustering methods used are all based on distance meassure, we try to minimize the influence of these missing value by simply fill column means in, which has the smallest distance measure to other data points.

Third, we rules out 4 varibales that has relative higher correlations, which will be discussed in detail in the following session. Finally, the all numberical dataset get scaled to same level to equal theirs weight in following clustering process.

Inputs correlations



From the chart above, we can see that the inputs quality overall is good, and there are only a few variable pairs has high correlation. We make some adjustments to the inputs (rule out 4 variables), where the correlations excess 0.7.

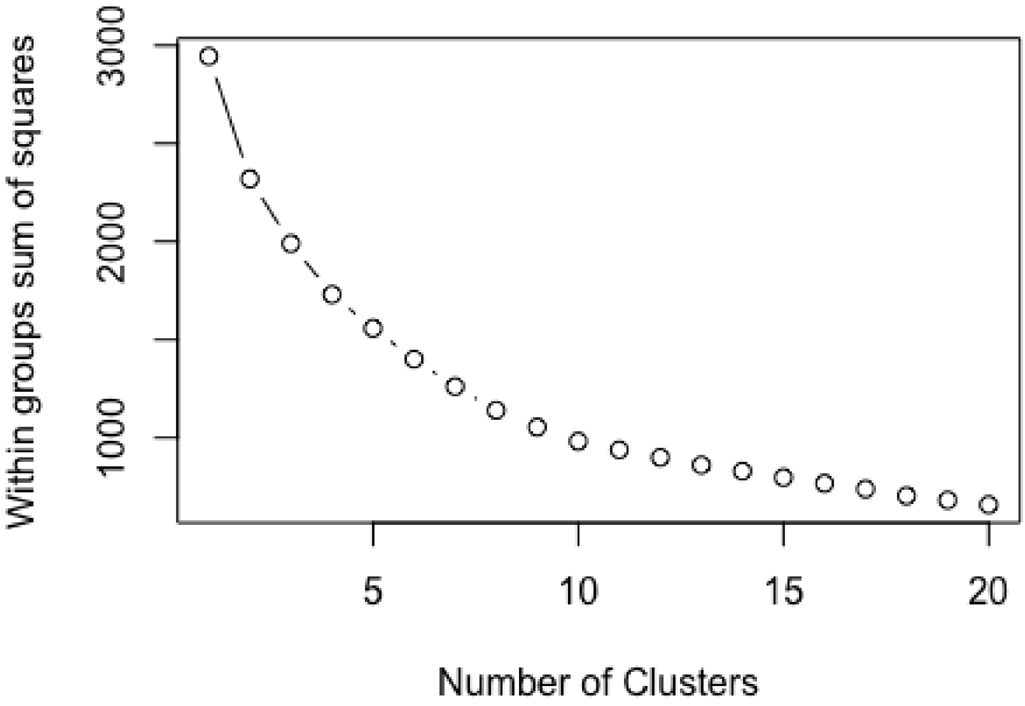
First, the foreign born residents lifeladder response and local born residents responses are almost perfectly related and both have relative high correlation with overall lifeladder response. We believe it shows that the happiness level of citizens are simliar whether they are foreign born or natives.Therefore, we decide to rule out both and only keep the overall lifeladder response.

Second, the GDP per capital and its log transfer reasonablly shows high correlations. The GDP data distribution, as we checked, roughly follows a lognormal form, so we decide to keep the log transfer of GDP per capital for its roughly normal distribution.

Third, the standard error of social support question response surprisingly has relative high correlation with standard error of lifeladder response and standard error of free-choice response. Since lifeladder response is the main happiness aspect we interested in, we rules out the standard error of social support question response.

Optimal K value selection

When choosing the best number of clusters, we decide to select k=8 based on Maslow's 'hierarchy of needs' theory in psychology proposed by Abraham Maslow in 1943. At the beginning, he created a classification system which reflected the universal needs of society as its base and then proceeding to more acquired emotions. Maslow used the terms "physiological," "safety," "belonging and love," "esteem," and "self-actualization" to describe the pattern through which human motivations generally move. Later on people improved this theory and expanded it into eight stages developmental model, including "Physiological needs,""Safety needs,""Belonging needs,""Self-esteem needs,""Cognitive needs,""Aesthetic needs,""Self-actualization needs," and "Self-transcendence needs". This eight levels of demand can greatly explain how happy people feel in their countries.



The above plot of Average Within-Cluster Squared Distance in terms of different number of clusters also indicates that clusters of 8 is a wise choice.

Cluster without Life-Ladder

K mean Model

After determing the best K value for clustering, we apply K means clustering approach to forming good clusters. By repeatedly assigning each record to one of the K clusters we pre-determine , we intend to find out an optimal clustering that can minimize the dispersion within clusters. In this way, the clusters we get are the most homogeneous.

As the best K value is 8, we set k to 8. Due to randomness of k means clustering, we try the same assigning process for 100 times by setting "nstart" to 100, so we can get a stable result.

```r
#Run K-mean clustering
wh.km.df1 <- whappy.df[,-1]
set.seed(1)
wh.km1 <- kmeans(wh.km.df1, centers = 8, nstart = 100)

#Include cluster result
wh.km.df1$Cluster <- wh.km1$cluster
wh.km.df1$country <- rownames(whappy.df)

#The number of countries in each cluster
pander(data.frame(clustersize = wh.km1$size,
           row.names = c("Cluster1", "Cluster2","Cluster3",
                         "Cluster4", "Cluster5", "Cluster6",
                         "Cluster7", "Cluster8")))
```

	clustersize
Cluster1	41
Cluster2	23
Cluster3	18
Cluster4	4
Cluster5	12
Cluster6	16
Cluster7	28
Cluster8	14

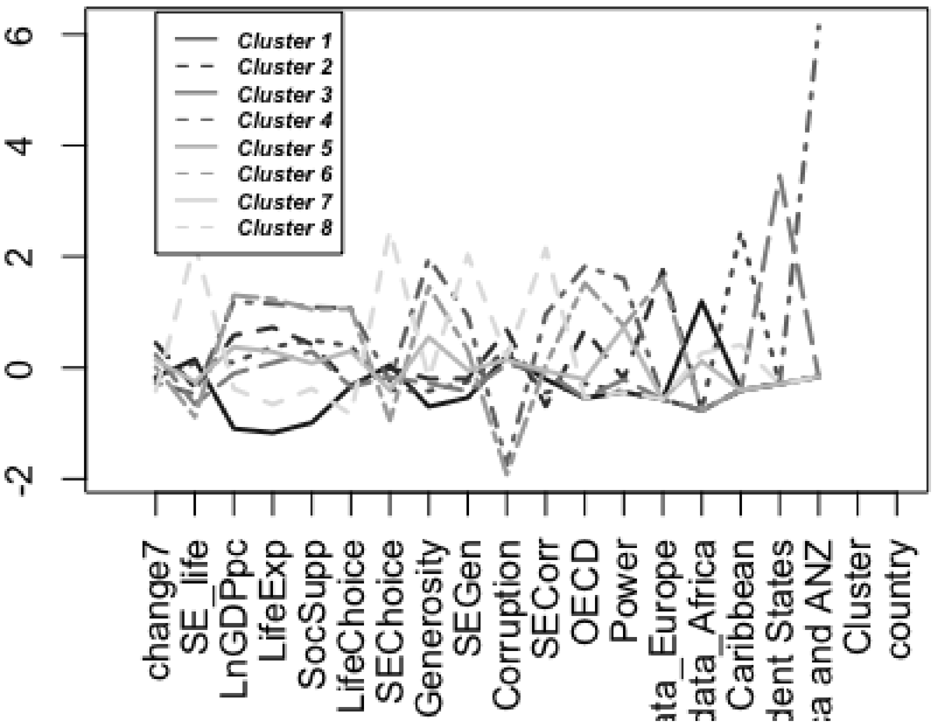
```r
#Cluster result display
out.km1 <- aggregate(country~ Cluster, wh.km.df1, I)
panderOptions("table.split.table", 90)
pander(data.frame(Cluster1=out.km1$country[[1]][1:10],
           Cluster2= out.km1$country[[2]][1:10],
           Cluster3= out.km1$country[[3]][1:10],
           Cluster4= out.km1$country[[4]][1:10],
           Cluster5= out.km1$country[[5]][1:10],
           Cluster6= out.km1$country[[6]][1:10],
           Cluster7= out.km1$country[[7]][1:10],
           Cluster8= out.km1$country[[8]][1:10]))
```

Table continues below

Cluster1	Cluster2	Cluster3	Cluster4
Afghanistan	Albania	Argentina	Australia
Bangladesh	Bosnia and Herzegovina	Bolivia	Canada
Benin	Bulgaria	Brazil	New Zealand
Botswana	Croatia	Chile	United States
Burkina Faso	Cyprus	Colombia	NA
Cameroon	Czech Republic	Costa Rica	NA
Central African Republic	Estonia	Dominican Republic	NA
Chad	Greece	Ecuador	NA
Congo (Brazzaville)	Hungary	El Salvador	NA
Congo (Kinshasa)	Italy	Guatemala	NA

Cluster5	Cluster6	Cluster7	Cluster8
Armenia	Austria	Bahrain	Algeria
Azerbaijan	Belgium	Cambodia	Angola
Belarus	Denmark	China	Belize
Georgia	Finland	Egypt	Bhutan
Kazakhstan	France	Hong Kong SAR, China	Burundi
Kyrgyzstan	Germany	India	Haiti
Moldova	Iceland	Indonesia	Jamaica
Russia	Ireland	Iran	Laos
Tajikistan	Luxembourg	Israel	Lesotho
Turkmenistan	Malta	Japan	Malaysia

Cluster Atrributes Summary



We also plot centrod plot combined with classification of countries and find some patterns:

1.countries in Cluster1 are generally developed countries with high Log GDP per capita, high life expectancy and less corruption.

2.countries in Cluster2 normally have weak postion in international relationship because they aren't power countries. But a clear misclassification is Russia which has national power.

3.countries in Cluster3 have relatively large increase in LifeLadder score but their corruption is relatively severe.

4.countries in Cluster4 are all wealthy and powerful countries that have long life expectancy, well social wealfare, strong social support, multiple life choices, high Generosity and huge national power.

5.countries in Cluster5 have relatively more corruption and weak national power.

6.countries in Cluster6 are mainly from Africa whose Log GDP per capita is much lower than overall mean. They also have low life expectancy, less social support, less Generosity and weak national power.

7.countries in Cluster7 have largest decrease in LifeLadder score and have fewest life choices and weak national power.

8.countries in Cluster8 are generally pretty normal and don’t have key features.

Hierarchical clustering

For hierarchical clustering model, we exclude the LifeLadder column and build clusters using some or all the remaining columns. We start with each cluster comprising exactly one record and then progressively agglomerating (combining) the two nearest clusters until there is just one cluster left at the end, which consists of all the records.

When calculating the distance of records, we use Euclidean distance. And then we apply this matrix derived from Euclidean distance into hierarchical clustering model, setting number of clusters to 8 for the purposes of comparing models

According to hierarchical clustering model, number of countries classified into 8 groups is 36,41,9,42,4,17,6,1 respectively. Then we list part of the countries classified into these 8 groups.

```r
#Run Hierarchical clustering
wh.h.df1 <- whappy.df[,-1]
matrix1 <- dist(wh.h.df1)
wh.hc1 <- hclust(matrix1,method="complete")
cut.hc1<-cutree(wh.hc1,k=8)

#Include cluster result
wh.h.df1$Cluster <- cut.hc1
wh.h.df1$country <- rownames(whappy.df)

#The number of countries in each cluster
pander(data.frame(clustersize = as.vector(table(cut.hc1)),
           row.names = c("Cluster1", "Cluster2","Cluster3",
                         "Cluster4", "Cluster5", "Cluster6",
                         "Cluster7", "Cluster8")))
```

	clustersize
Cluster1	36
Cluster2	41
Cluster3	9
Cluster4	42
Cluster5	4
Cluster6	17
Cluster7	6
Cluster8	1

```r
#Cluster result display (Partial)
out.h1 <- aggregate(country~ Cluster, wh.h.df1, I)
```

```
pander(data.frame(Cluster1=out.h1$country[[1]][1:10],
           Cluster2= out.h1$country[[2]][1:10],
           Cluster3= out.h1$country[[3]][1:10],
           Cluster4= out.h1$country[[4]][1:10],
           Cluster5= out.h1$country[[5]][1:10],
           Cluster6= out.h1$country[[6]][1:10],
           Cluster7= out.h1$country[[7]][1:10],
           Cluster8= out.h1$country[[8]][1:10]))
```

Table continues below

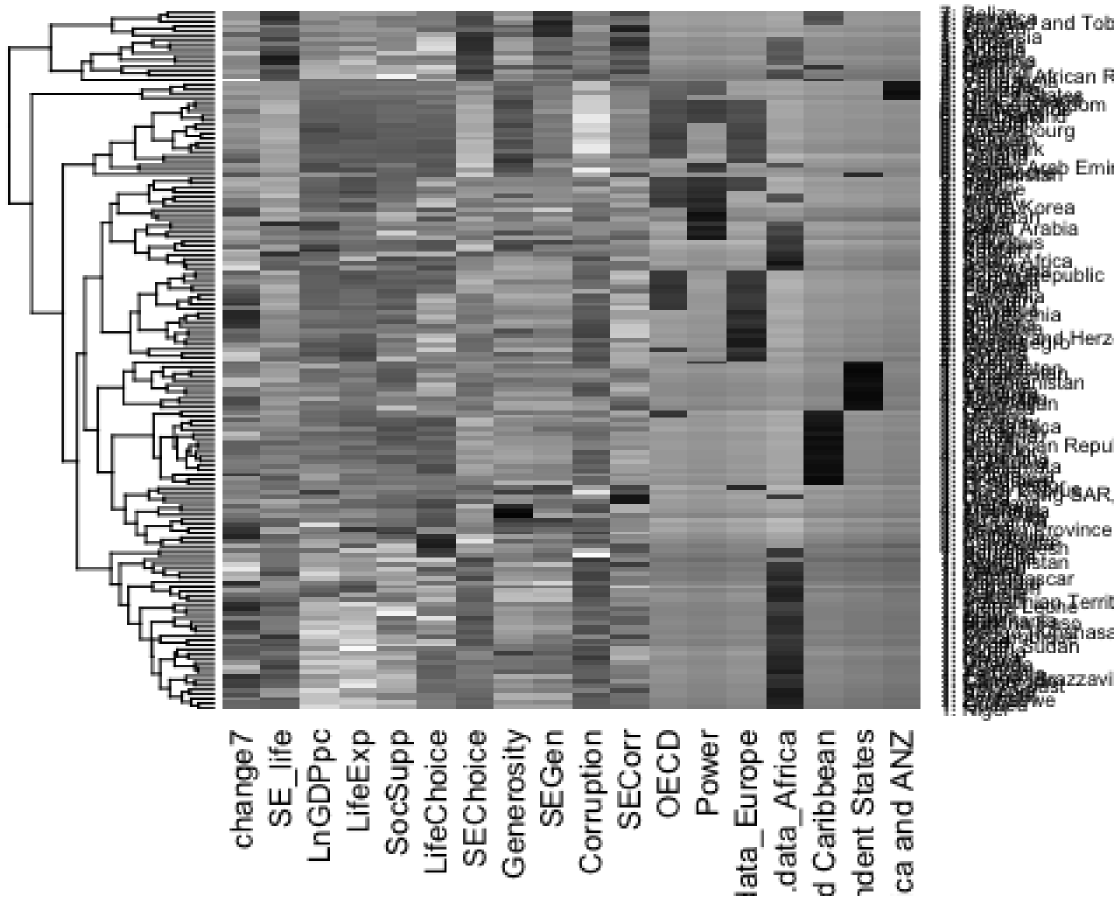
Cluster1	Cluster2	Cluster3	Cluster4
Afghanistan	Albania	Algeria	Argentina
Benin	Bahrain	Angola	Armenia
Burkina Faso	Bosnia and Herzegovina	Burundi	Azerbaijan
Cameroon	Botswana	Central African Republic	Bangladesh
Chad	Bulgaria	Haiti	Belarus
Congo (Brazzaville)	China	Lesotho	Bolivia
Congo (Kinshasa)	Croatia	Namibia	Brazil
Ethiopia	Cyprus	Sudan	Cambodia
Gabon	Czech Republic	Syria	Chile
Ghana	Egypt	NA	Colombia

Cluster5	Cluster6	Cluster7	Cluster8
Australia	Austria	Belize	Venezuela
Canada	Belgium	Bhutan	NA
New Zealand	Denmark	Jamaica	NA
United States	Finland	Laos	NA
NA	Germany	Malaysia	NA
NA	Iceland	Trinidad and Tobago	NA
NA	Ireland	NA	NA
NA	Luxembourg	NA	NA
NA	Malta	NA	NA
NA	Netherlands	NA	NA

Cluster Atrributes Summary

The heat map below from hierarchical clustering is crowed with 156 observations, so it is kind of hard to see patterns and atrributes. The relative clear pattern is that the cluster members share similar value on regional dummies, power and OECD (indicates international status), which means these variables contributes more in this clustering method.

```
## [1] "Heat Map"
```



We also plot heatmap combined with classification of countries and find some patterns:

1.countries in Cluster1 are mainly from Africa whose Log GDP per capita is much lower than overall mean.

2.countries in Cluster2 are kind of normal in every aspect except their corruption is relatively severe.

3.countries in Cluster3 have main features of short healthy life expectancy at birth and poor social support.

4.countries in Cluster4 have the largest enchancement in LifeLadder score.

5.countries in Cluster5 are all wealthy and powerful countries that have long life expectancy, well social wealfare.

6.countries in Cluster6 are generally developed countries with high Log GDP per capita, multiple life choice and less corruption.

7.countries in Cluster7 normally have weak postion in international relationship because they aren’t power countries.

8.countries in Cluster8 only include Venezuela which has huge decrease in LifeLadder score, few life choices, low Generosity and high rate of corruption all because of its hyperinflation.

Cluster with Life-Ladder

K mean Model

We include life ladder variable with everything else equal to the K-means model before.

```r
#Run K-mean clustering
wh.km.df2 <- whappy.df
set.seed(1)
wh.km2 <- kmeans(wh.km.df2, centers = 8, nstart = 100)

#Include cluster result
wh.km.df2$Cluster <- wh.km2$cluster
wh.km.df2$country <- rownames(whappy.df)

#The number of countries in each cluster
pander(data.frame(clustersize = wh.km2$size,
          row.names = c("Cluster1", "Cluster2","Cluster3",
                        "Cluster4", "Cluster5", "Cluster6",
                        "Cluster7", "Cluster8")))
```

	clustersize
Cluster1	18
Cluster2	43
Cluster3	4
Cluster4	13
Cluster5	23
Cluster6	12
Cluster7	27
Cluster8	16

```r
#Cluster result display#Cluster result display (Partial)
out.km2 <- aggregate(country~ Cluster, wh.km.df2, I)
pander(data.frame(Cluster1=out.km2$country[[1]][1:10],
          Cluster2= out.km2$country[[2]][1:10],
          Cluster3= out.km2$country[[3]][1:10],
          Cluster4= out.km2$country[[4]][1:10],
          Cluster5= out.km2$country[[5]][1:10],
          Cluster6= out.km2$country[[6]][1:10],
          Cluster7= out.km2$country[[7]][1:10],
          Cluster8= out.km2$country[[8]][1:10]))
```

Table continues below

Cluster1	Cluster2	Cluster3	Cluster4
Argentina	Afghanistan	Australia	Algeria
Bolivia	Bangladesh	Canada	Angola
Brazil	Benin	New Zealand	Belize
Chile	Botswana	United States	Bhutan
Colombia	Burkina Faso	NA	Haiti
Costa Rica	Burundi	NA	Jamaica
Dominican Republic	Cameroon	NA	Laos
Ecuador	Central African Republic	NA	Lesotho
El Salvador	Chad	NA	Malaysia
Guatemala	Congo (Brazzaville)	NA	Namibia
Cluster5	Cluster6	Cluster7	Cluster8
Albania	Armenia	Bahrain	Austria
Bosnia and Herzegovina	Azerbaijan	Cambodia	Belgium
Bulgaria	Belarus	China	Denmark
Croatia	Georgia	Hong Kong SAR, China	Finland
Cyprus	Kazakhstan	India	France
Czech Republic	Kyrgyzstan	Indonesia	Germany
Estonia	Moldova	Iran	Iceland
Greece	Russia	Israel	Ireland
Hungary	Tajikistan	Japan	Luxembourg
Italy	Turkmenistan	Kuwait	Malta

Cluster Atrributes Summary

After the inclusion of life ladder variable, the K-means cluster set bearly changes. Therefore, the cluster attributes do not change a lot too.

Hierarchical clustering

We include life ladder variable with everything else equal to the hierarchical model before.

```
#Run Hierarchical clustering
wh.h.df2 <- whappy.df
matrix2 <- dist(wh.h.df2)
wh.hc2 <- hclust(matrix2,method="complete")
cut.hc2<-cutree(wh.hc2,k=8)
```

```
#Include cluster result
wh.h.df2$Cluster <- cut.hc2
wh.h.df2$country <- rownames(whappy.df)

#The number of countries in each cluster
print("The number of countries in each cluster")
```

```
## [1] "The number of countries in each cluster"
```

```
pander(data.frame(clustersize = as.vector(table(cut.hc2)),
           row.names = c("Cluster1", "Cluster2","Cluster3",
                         "Cluster4", "Cluster5", "Cluster6",
                         "Cluster7", "Cluster8")))
```

	clustersize
Cluster1	40
Cluster2	76
Cluster3	9
Cluster4	4
Cluster5	15
Cluster6	4
Cluster7	7
Cluster8	1

```
#Cluster result display (Partial)
out.h2 <- aggregate(country~ Cluster, wh.h.df2, I)
pander(data.frame(Cluster1=out.h2$country[[1]][1:10],
           Cluster2= out.h2$country[[2]][1:10],
           Cluster3= out.h2$country[[3]][1:10],
           Cluster4= out.h2$country[[4]][1:10],
           Cluster5= out.h2$country[[5]][1:10],
           Cluster6= out.h2$country[[6]][1:10],
           Cluster7= out.h2$country[[7]][1:10],
           Cluster8= out.h2$country[[8]][1:10]))
```

Table continues below

Cluster1	Cluster2	Cluster3	Cluster4
Afghanistan	Albania	Algeria	Australia
Benin	Argentina	Angola	Canada
Botswana	Armenia	Burundi	New Zealand
Burkina Faso	Azerbaijan	Central African Republic	United States
Cameroon	Bahrain	Haiti	NA
Chad	Bangladesh	Lesotho	NA

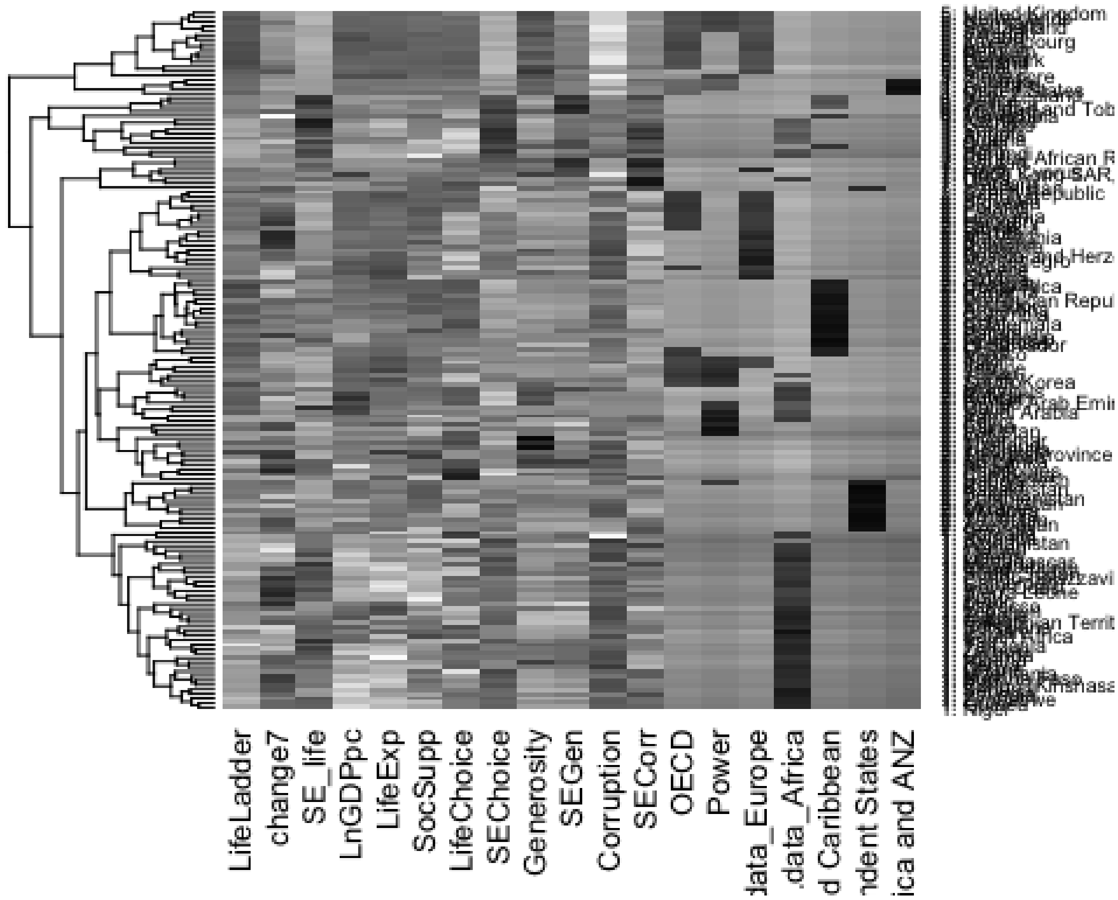
Congo (Brazzaville)	Belarus	Namibia	NA
Congo (Kinshasa)	Bolivia	Sudan	NA
Ethiopia	Bosnia and Herzegovina	Syria	NA
Gabon	Brazil	NA	NA

Cluster5	Cluster6	Cluster7	Cluster8
Austria	Belize	Bhutan	Venezuela
Belgium	Jamaica	Hong Kong SAR, China	NA
Denmark	Malaysia	Laos	NA
Finland	Trinidad and Tobago	Libya	NA
Germany	NA	North Cyprus	NA
Iceland	NA	Uzbekistan	NA
Ireland	NA	Vietnam	NA
Luxembourg	NA	NA	NA
Malta	NA	NA	NA
Netherlands	NA	NA	NA

Cluster Atrributes Summary

The heat map below is generally the same as hierarchical clustering with out life ladder . The relative clear pattern is that the cluster members share similar value on regional dummies, power and OECD (indicates international status).

```
## [1] "Heat Map"
```



In the final clusters, we find some patterns:

1.countries in Cluster1 are mainly from Africa whose Log GDP per capita is much lower than overall mean. They also have low life expectancy at birt and poor social support.

2.countries in Cluster2 have the largest enchancement in LifeLadder score.

3.countries in Cluster3 have main features of short healthy life expectancy at birth and poor social support. They are similar to countries in Cluster1 but their life choices are much lower.

4.countries in Cluster4 are all wealthy and powerful countries that have long life expectancy, well social wealfare.

5.countries in Cluster5 are generally developed countries with high Log GDP per capita, multiple life choice and less corruption.

6.countries in Cluster6 are kind of normal in every aspect except their corruption is relatively severe.

7.countries in Cluster7 normally have weak postion in international relationship because they aren't power countries.

8.countries in Cluster8 only include Venezuela which has huge decrease in LifeLadder score, few life choices, low Generosity and high rate of corruption all because of its hyperinflation.

Impact of including the happiness Life Ladder

For K-mean clustering, the inclusion of happiness did not significantly impact the clustering result. In factor, only about 2 countries changed clusters. Therefore, it shows that the country happiness is largely depended on other variables, and did not capture something different.

For Hierarchical clustering, the inclusion of happiness did significantly impact the clustering result. The cluster set seems to be more unbalanced with single jumbo group. The major change is the merge of cluster 2 and 4, whose major differences lies in OECD (whether the country is a member of OECD) and power. The country happiness capture the similarity among these countries dispite their difference in international status. Therefore, it basically shows that a country happiness level has little relation with international status.

The best Clustering set

We would choose the hierarchical clustering model with life ladder. First, the reason why we choose not to drop life ladder is that this variable is actually the most relative varaible that reflects the country's happiness . If we drop life ladder, the clustering process is no longer about whether the country's people is having a happy life or not, while more like grouping different countries based only on their different features. Given that we are figuring out the answer to the question about happiness in different countries, life ladder is certainly one that we need to include.

For the selection of model, we compared these 4 models by looking into 4 different clustering mechanics behind these models. Extreme values, deviation from mean value, and some similar features between different clusters all appear in k means clustering model. For example, one certain cluster contains countries that all have low international power,but Russia is assigned to this cluster while Russia has a very high value of power.

In addition, k means clustering model has trouble distinguishing countries that have similar and close values, some countries that have similar values and in the same region are assigned to different clusters.

In contrary, clusters that hierarchical clustering model have are more reasonable and distinguishable.

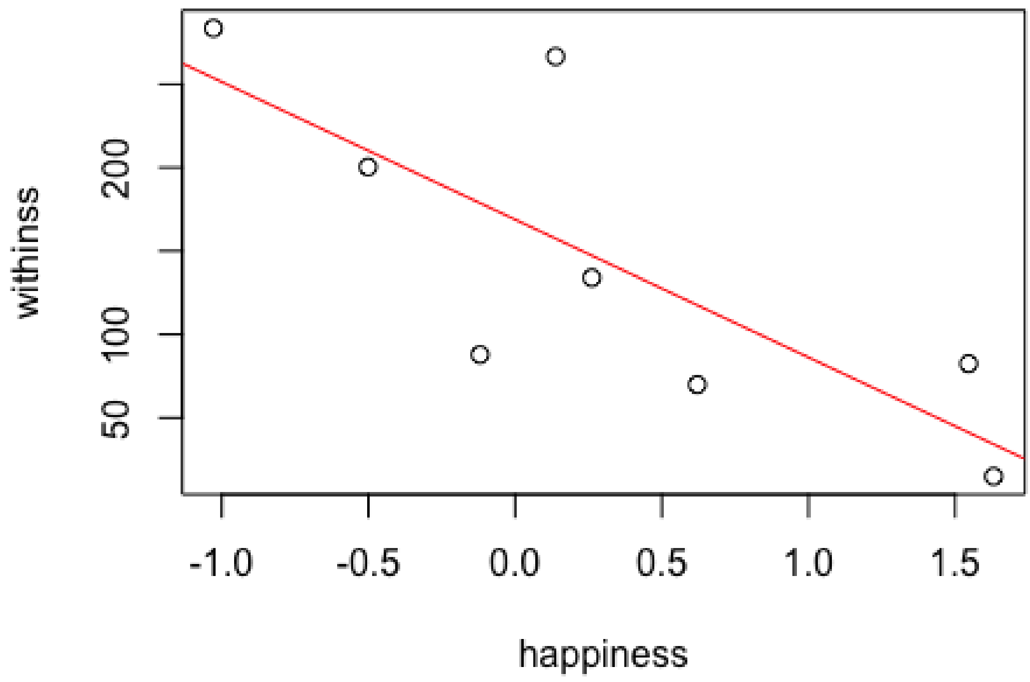
Cluster set with defined lable

	Lable
Cluster1	LowGDP/Africa
Cluster2	HappinessEnhancement

Cluster3	LowlifeExp/LowSocialsupport
Cluster4	Wealthy/Powerful/LonglifeExp
Cluster5	Developed/MoreLifechoice/LessCorruption
Cluster6	Corruption
Cluster7	Normal/LessInternationalStatus
Cluster8	Venezuela

Tolstoy’s assertion to countries (Conclusion)

From the following result display, we can safely conclude that the country cluster with higher happiness level do have more similarities shown by smaller within cluster sum square. In other words, Tolstoy’s assertion, “All happy families are alike”, can be generalized to countries.



Appendix Pros/Cons of K-Means and Hier

K-Means Clustering

Strengths

1. Easy to implement
2. If k is a small number, k means clustering is faster in computation.

Weaknesses

1. k value must be predetermined.
2. Sensitive to scale.

Hierarchical Clustering

Strengths

1. Dendrogram outputs are informative structures if the tree is not too complex.

Weaknesses

1. For large datasets, hierarchical clustering requires complex and slow computation.
2. low stability when reordering data.
3. sensitive to outliers.