TASK

Capstone Project II

Visit our website

Introduction

WELCOME TO THE SECOND CAPSTONE PROJECT TASK!

For this Capstone project, we are going to perform clustering techniques on the dataset provided and analyse which method is the best. Secondly, we will perform PCA on our dataset to investigate if it helps the clustering of the observations.

Get in touch

Connect for support

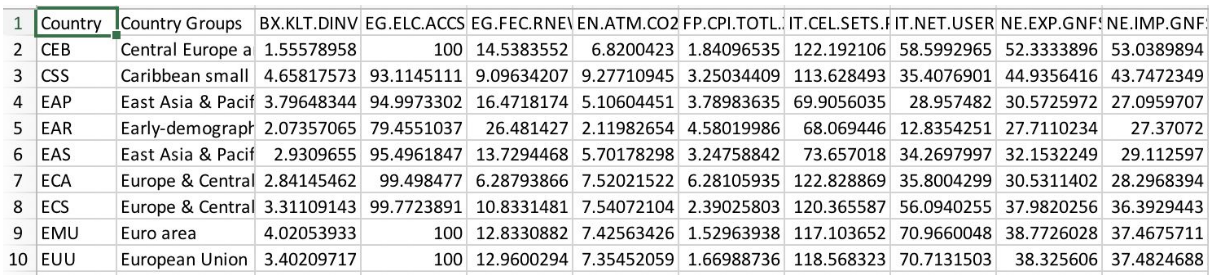
Remember that with our courses, you're not alone! You can contact your mentor to get support on any aspect of your course.

The best way to get help is to login to **www.hyperiondev.com/portal** to start a chat with your mentor. You can also schedule a call or get support via email.

Your mentor is happy to offer you support that is tailored to your individual career or education needs. Do not hesitate to ask a question or for additional support!

INTRODUCTION TO THE TASK

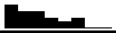
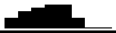
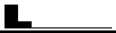
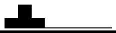
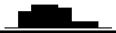
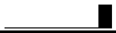
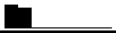
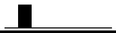
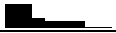
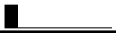
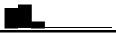
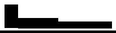
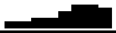
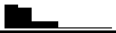
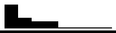
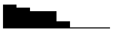
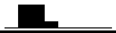
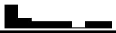
In this task, we explore the differences between various countries using unsupervised learning methods such as Principal Component Analysis (PCA) and various clustering techniques. The data we will be exploring contains 192 countries. There are 19 variables describing each country which contain population statistics, electricity and technology adoption as well as economic indicators such as inflation and trade data.



1	Country	Country Groups	BX.KLT.DINV	EG.ELC.ACCS	EG.FEC.RNE	EN.ATM.CO2	FP.CPI.TOTL.	IT.CEL.SETS.	IT.NET.USER	NE.EXP.GNF	NE.IMP.GNF
2	CEB	Central Europe a	1.55578958	100	14.5383552	6.8200423	1.84096535	122.192106	58.5992965	52.3333896	53.0389894
3	CSS	Caribbean small	4.65817573	93.1145111	9.09634207	9.27710945	3.25034409	113.628493	35.4076901	44.9356416	43.7472349
4	EAP	East Asia & Pacif	3.79648344	94.9973302	16.4718174	5.10604451	3.78983635	69.9056035	28.957482	30.5725972	27.0959707
5	EAR	Early-demograph	2.07357065	79.4551037	26.481427	2.11982654	4.58019986	68.069446	12.8354251	27.7110234	27.37072
6	EAS	East Asia & Pacif	2.9309655	95.4961847	13.7294468	5.70178298	3.24758842	73.657018	34.2697997	32.1532249	29.112597
7	ECA	Europe & Central	2.84145462	99.498477	6.28793866	7.52021522	6.28105935	122.828869	35.8004299	30.5311402	28.2968394
8	ECS	Europe & Central	3.31109143	99.7723891	10.8331481	7.54072104	2.39025803	120.365587	56.0940255	37.9820256	36.3929443
9	EMU	Euro area	4.02053933	100	12.8330882	7.42563426	1.52963938	117.103652	70.9660048	38.7726028	37.4675711
10	EUU	European Union	3.40209717	100	12.9600294	7.35452059	1.66988736	118.568323	70.7131503	38.325606	37.4824688

EXPLORING THE DATA

To improve the understanding of the data, the variables are renamed to have more intuitive names such as "**Birthrate**" instead of the original heading of "SP.DYN.CBRT.IN". The mean, standard deviation, range and distribution of each variable as well as the number of missing values per variable were observed. This is summarised in the table below.

Variable	Missing	Mean	Standard Deviation	Min	Max	Histogram
BirthRate	1	22.34	10.83	8.3	50.03	
Cellphone	2	89.25	42.27	1.18	209	
CO2	4	4.88	6.45	0.024	40.74	
CPI(%)	10	4.53	4.02	-2.43	28.19	
DeathRate	1	8.09	2.99	1.47	16.57	
Electricity (%)	0	78.24	30.94	1.5	100	
Exports (%)	7	43.54	34.73	0.11	298.34	
FDI (%)	6	5.88	10.99	-15.99	105.79	
Fertility	3	2.93	1.49	1.06	7.49	
GDP($)	1	13364.51	18779.18	231.19	1e+05	
Imports (%)	7	49.25	31.83	0.066	284.97	
Internet (%)	2	33.1	27.2	0.25	93.39	
LifeExp	3	70.24	8.62	47.56	82.98	
MortalityFem	4	144.7	102.84	28.71	508.16	
MortalityInfant	7	27.86	24.73	2	106.7	
MortalityMale	4	211.74	108.24	68.68	588.84	
PopGrowth (%)	0	1.52	1.55	-2.1	11.22	
Renewable (%)	1	31.66	29.84	0	96.83	

UrbanPop (%)	0	56.41	23.91	9.09	100	

At first glance, the GDP per capita variable stands out as having a mean and standard deviation which is dramatically higher than the other variables. This makes sense as most of the other variables are percentages or ratios per 1000 people while GDP per Capita is in US$. This indicates that scaling the data will be useful to keep the GDP per capita from impacting the analysis disproportionately.

MISSING VALUES

Most observations in the data have at least 1 missing value. Below is a sample showing some missing data.

```
##CODE  COUNTRY    ForeignInv AccessElec RenewEnerg CO2Emmis ConsumPric
##  ABW      Aruba        7.57          93.4     5.46         24.7         2.08
##  BMU      Bermuda      3.88          100.     2.39         9.35         NA
##  CUB      Cuba         NA            100.0    13.2         3.39         NA
##  DJI      Djibouti     3.23          53.3     34.4         0.607         3.95
##  DMA      Dominica     4.93          94.8     8.91         1.95         3.21
##  ERI      Eritrea      4.30          39.7     81.2         0.117        NA
```

There are 63 missing values in total. These missing values are distributed among 23 countries, therefore approximately 12% of our observations contain missing data. This is a significant amount of observations to exclude from the analysis!

There exist a variety of techniques for substituting missing values with statistical prediction. This process is generally referred to as ‘missing data imputation’.

A very powerful imputation method is to use bagging. For each variable, a bagged tree is created via all the other variables in the dataset using 10 bootstrap replications. For every missing value, the appropriate bagged tree is used to predict the value. By using imputation, all 192 countries could be used in the rest of the analysis.

Another option is K-Nearest Neighbour Imputation, which is based on a variation of the Gower Distance. Consider the first missing variable: the variable consumer price of Bermuda.

CODE	COUNTRY	ForeignInv	AccessElec	RenewEnerg	CO2Emmis	ConsumPric
BMU	**Bermuda**	**3.88**	**100.**	**2.39**	**9.35**	**NA**

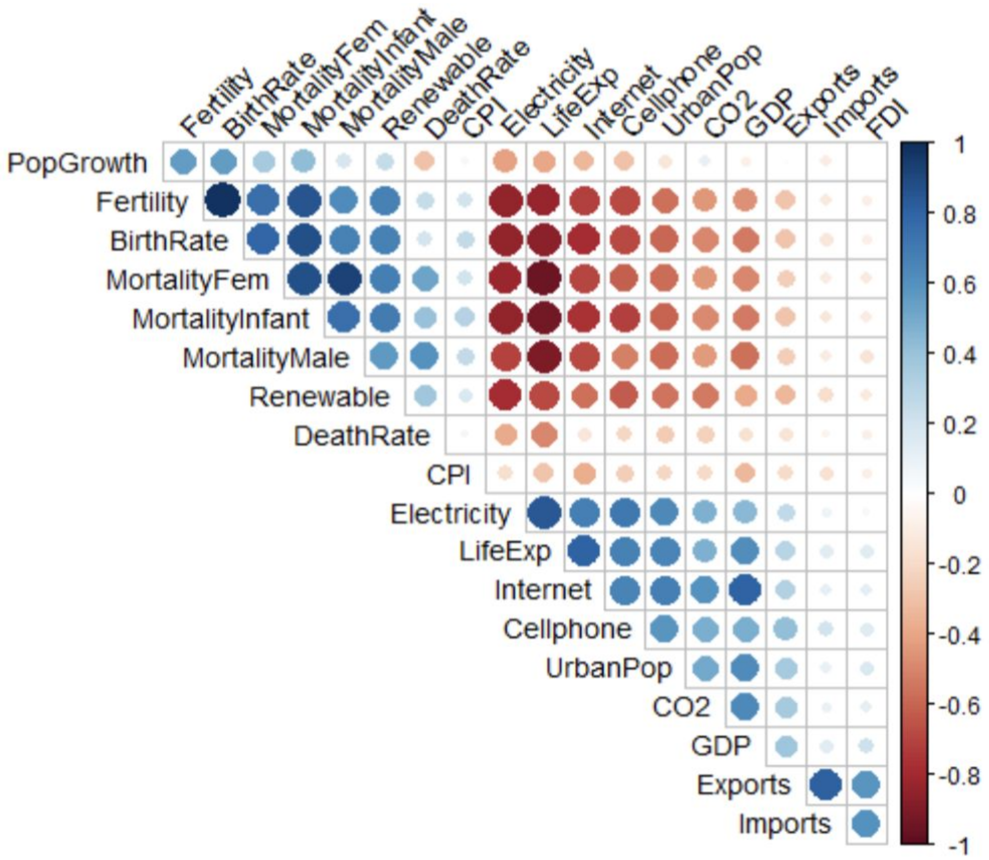
KNN will investigate other observations with a similar number of cellphone subscribers, internet users and all the other variables. After the identification, it will get the mean of the consumer prices of these observations and impute the value as the value of consumer price for Bermuda.

Here is the same sample as before, now with the imputed values. It shows us that all the missing variables have been imputed and the data is now ready for PCA.

```
## CODE   COUNTRY    ForeignInv   AccessElec    RenewEnerg
## ABW     Aruba     7.5681598    93.35629     5.46471591
## BMU     Bermuda   3.8769921    100.00000    2.39176852
## CUB     Cuba      5.0441724    99.96390     13.15763195
## DJI     Djibouti  3.2341533    53.30377     34.43334712
```

CORRELATION ANALYSIS

From the plot below, most of the variables are highly positively or negatively correlated with each other. Access to electricity (% of population) and Female Mortality are strongly negatively correlated. Internet users (% of population) and GDP per capita are highly positively correlated.



From the correlation plot, it is evident that Foreign Direct Investment (FDI) has a strong positive correlation to imports and exports. Access to electricity is positively correlated to cellphone subscriptions, internet usage, life expectancy and percentage of people who live in urban areas.

These correlations are intuitive as people who have electricity can use electronics such as phones, urban areas are more likely to have electricity than rural areas and, generally, countries with electricity access are more likely to have better healthcare, thereby increasing life expectancy. The predictors that have a strong negative correlation to electricity are the various mortality rates, the fertility rate and the percentage of renewable energy consumption.

The last correlation is interesting as it seems to suggest that countries which have high access to electricity are less likely to use renewable energy. This may point to the fact that countries with high access to electricity historically haven't needed to invest as heavily in renewable energy infrastructure as they can already provide for their countries' electricity needs with their existing fossil fuels production techniques.

There are other intuitive correlations such as population growth to fertility rates and birth rates. Overall, there are many variables that have strong negative and positive correlations with each other. This makes the data a good candidate for PCA. PCA will be able to reduce variables which encode similar types of differences between countries in a way that requires fewer dimensions.

PCA: UNSTANDARDISED DATA

Principal Components Analysis (PCA) is a method for finding the underlying variables (i.e. principal components) that best differentiate the observations by determining the directions along which your data points are most spread out. Since the determination of the principal components is based on finding the direction that maximises the variance, variables with variance that are much higher than the other variables tend to dominate the analysis purely due to their scale.

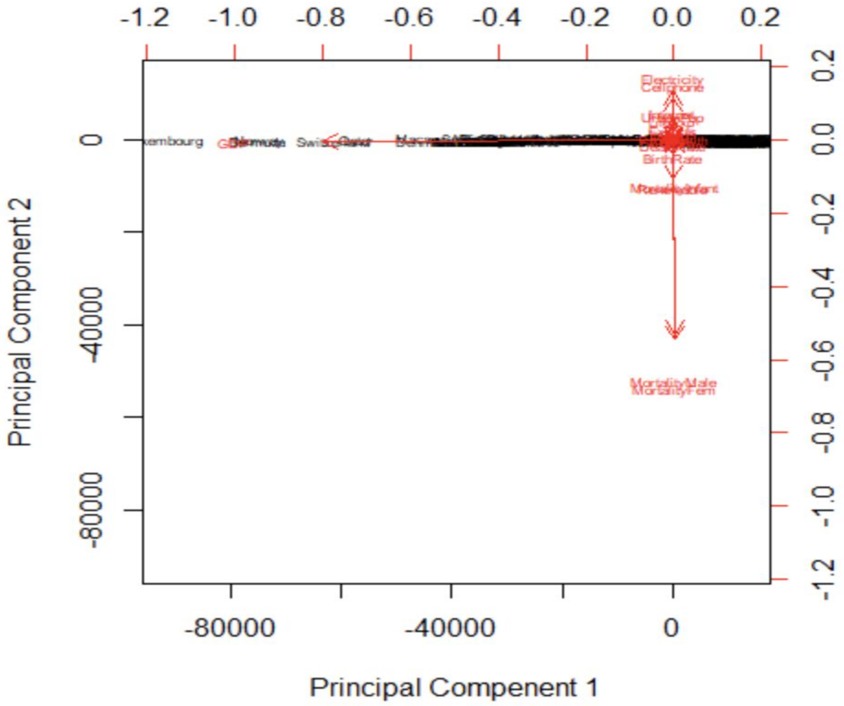
```
## Importance of components:

##                          PC1      PC2     PC3      PC4     PC5      PC6
## Standard deviation       1.8e+04  128.20  46.27    38.49   23.76    19.98
##Proportion of Variance    9.9e-01  0.0     0.00001  0.00    0.00     0.00
```

```
##Cumulative Proportion  9.9e-01      0.99  0.99999    1.00       1.00        1.00
```

The procedure shows the standard deviation associated with each of the 19 components. It also shows the amount of variance that the principal component comprises in comparison to the total variance.

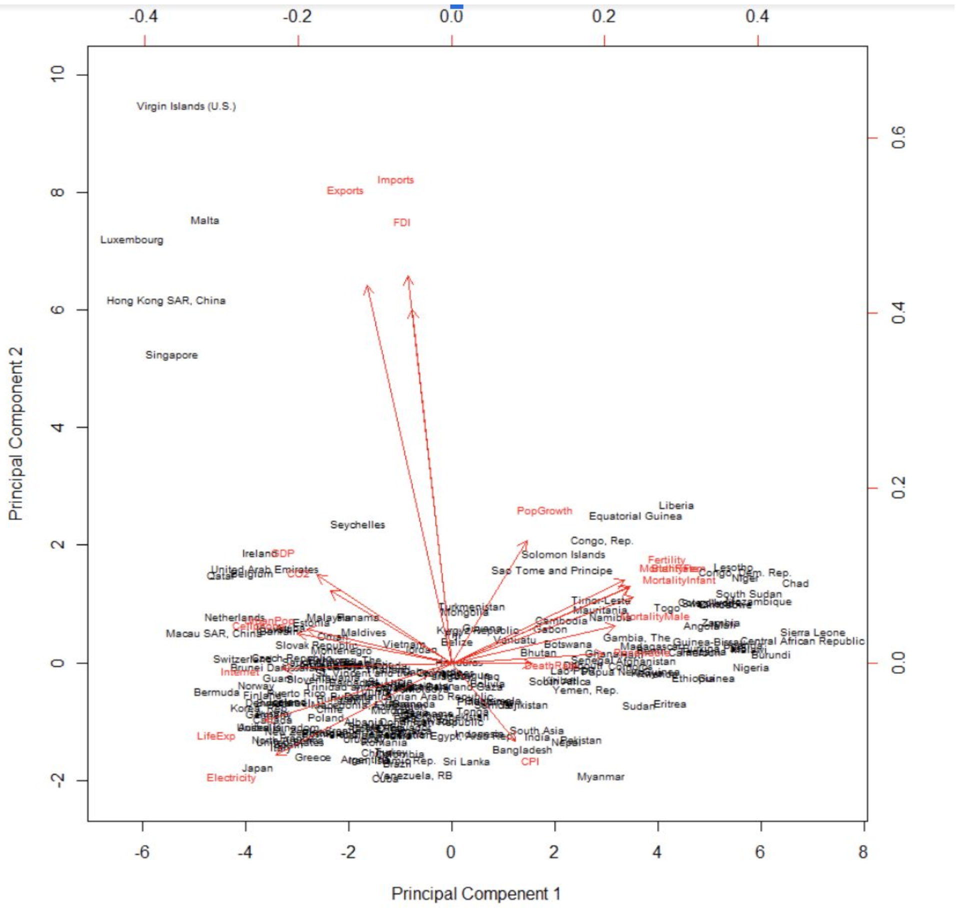
As expected, the first principal component is dominated by GDP which is on a much larger scale than the other variables (as seen during data exploration). This makes it difficult to see how countries vary with respect to the other variables or read the biplot as most countries are overlapping.



It can, however, be seen that Luxembourg has the largest GDP per Capita and the wealth dispersion among countries is very heavily skewed to the right with there being a large wealth gap between the poor and wealthy countries. The second principal component deals with the other variables, with the correlated variables going in the same direction. Due to the scale issue, there seems to be little variation based on the 2nd principal component. This is confirmed by looking at the proportion of variance explained which is 99.99% and 0.005% for the 1st and 2nd principal components respectively.

In order to learn more about the data through PCA, the data was scaled prior to performing PCA. After this, the Biplot below can be read more easily, while gaining more insight into possible clusters in the data.

PCA - Standardised Data



The first principal component seems to separate the data into 2 directions, which shows the strength of the negative correlations mentioned above. The variables with the largest positive loading values are the various mortality rates, the fertility rate and renewable energy. While the variables with significant negative loading values are the technology and electricity access, urbanisation level, GDP per capita and life expectancy. Therefore, the 1st principal component seems to summarise a general standard of living.

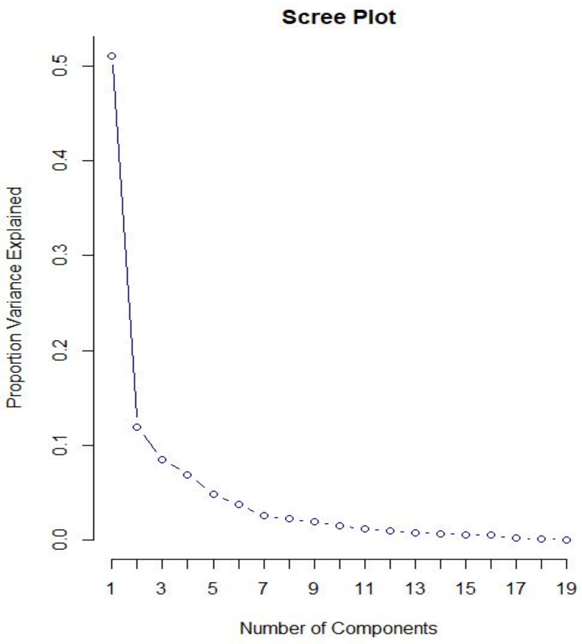
Most of the countries that have a generally lower standard of living are African countries such as Sierra Leone, Nigeria, Central African Republic and Burundi. This seems to indicate that many African countries have mortality rates that are above average, have larger families than more Western countries, but have been more proactive in setting up renewable energy in their respective countries. Internet and cellphone access is much lower than in the wealthier countries and a much lower percentage of Africans tend to live in urban areas.

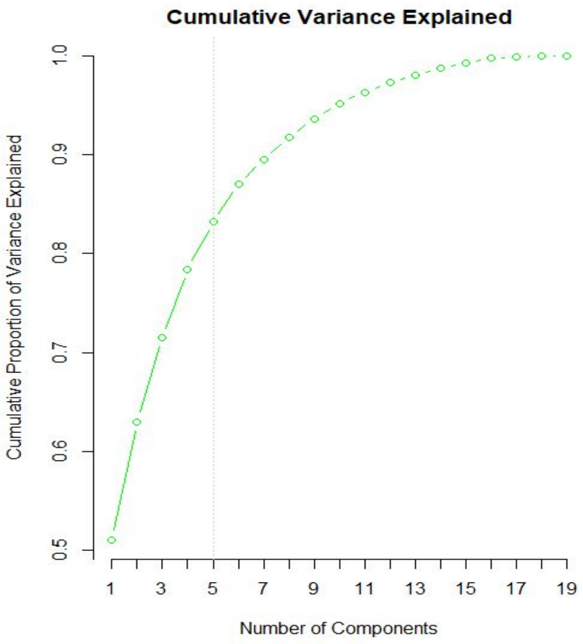
The countries in the centre of the 1st principal component are the countries which are still developing but have a higher standard of living than many African countries. This includes Sri Lanka, India, Jordan and Belize.

The wealthier countries with higher standards of living are on the left and include more European countries or countries that have large oil reserves such as the UAE and Qatar.

The 2nd principal component is dominated by exports, imports and Foreign Direct Investment (FDI), which we saw earlier were positively correlated. This can be summarised as a principal component indicating trade and investment levels. It makes sense that countries which have high investments would be investing in manufacturing products that can be exported. Raw materials for the production may need to be imported leading to the correlation between the variables. The countries that are extremely above average in these variables are Malta, the Virgin Islands (U.S.), Luxembourg, Hong Kong and Singapore. These 5 likely represent a cluster in the cluster analysis which will be performed below.

In PCA, the first few principal components are the variables that explain most of the variation in the data. And therefore, when using PCA for dimensionality reduction, we need to choose an appropriate number of principal components that explain a significant portion of the variation in our data. This decision will be aided by the Scree plot and Cumulative Variance Explained Plot below.





The first 5 principal components together explain over 80% of the variance. We can therefore use them to perform cluster analysis. This is what we refer to as dimensionality reduction. We began with 19 variables and now we have 5 variables explaining most of the variability.

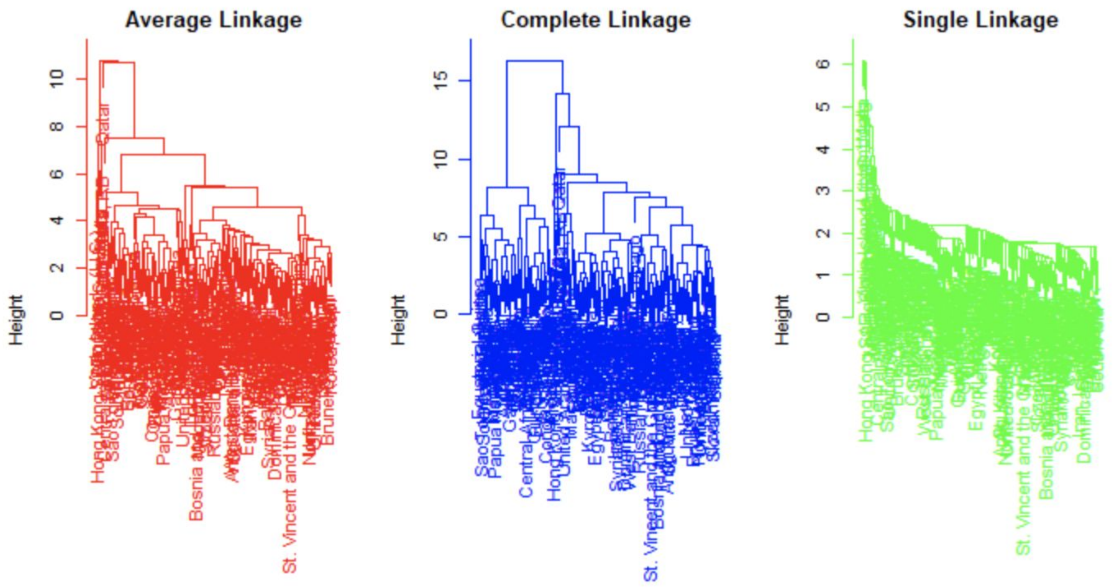
CLUSTER ANALYSIS

We will perform both Hierarchical Clustering and K-means to this data and compare the results.

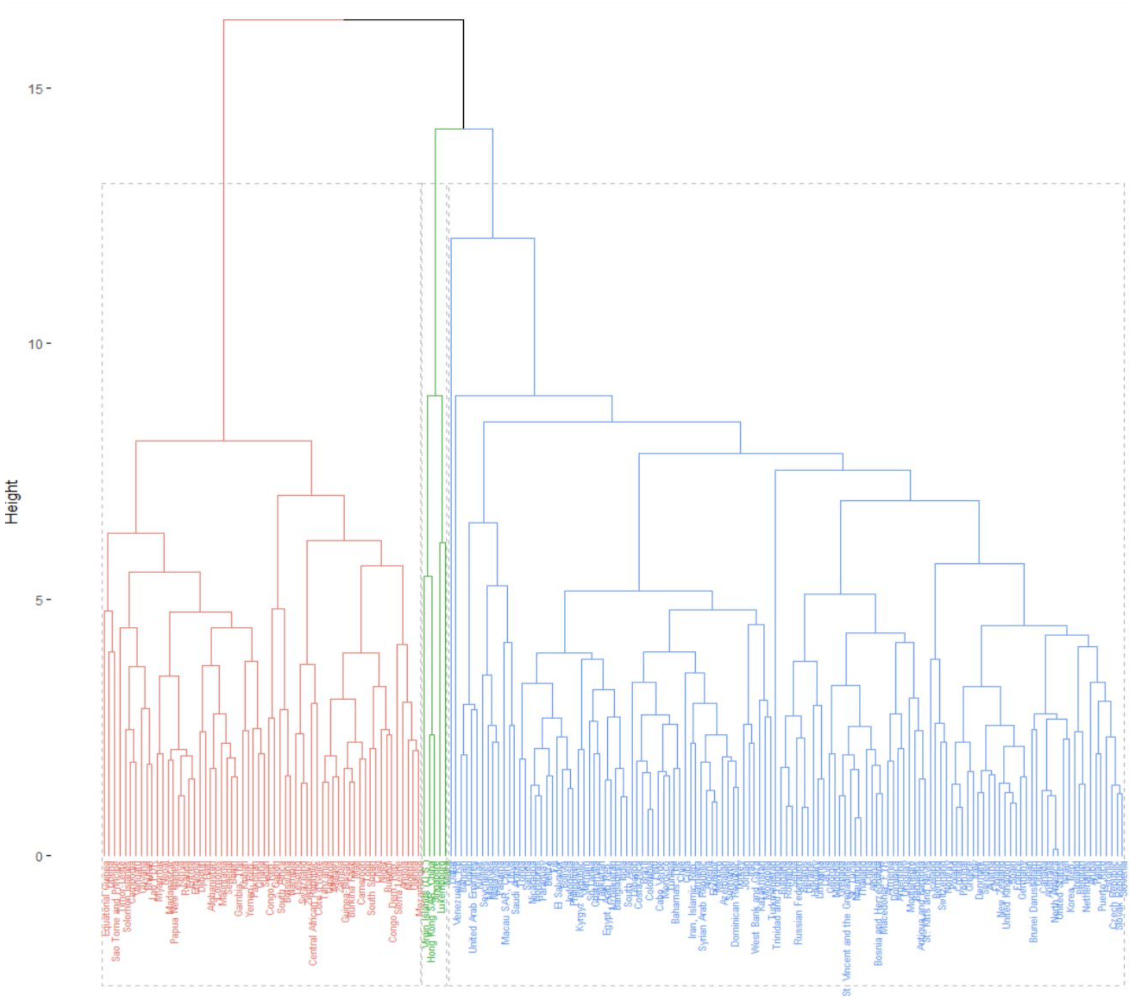
Hierarchical clustering

Hierarchical clustering has the advantage that we can see the clusters visually in a dendrogram and don't have to specify the number of clusters before running the algorithm. However, we will have to decide the number of clusters after the algorithm runs.

For the distance metric between observations, Euclidean distance was used which is the most common way to measure distance. In order to determine the method used to measure the distance between clusters, we plotted the various dendrograms for the single, complete and average linkage methods.



From the dendrograms above, the complete linkage method creates the most balanced dispersion of clusters and will therefore be the method of choice for the rest of this analysis. A clearer dendrogram for the complete linkage method is shown below with the colouring of the 3 clusters.

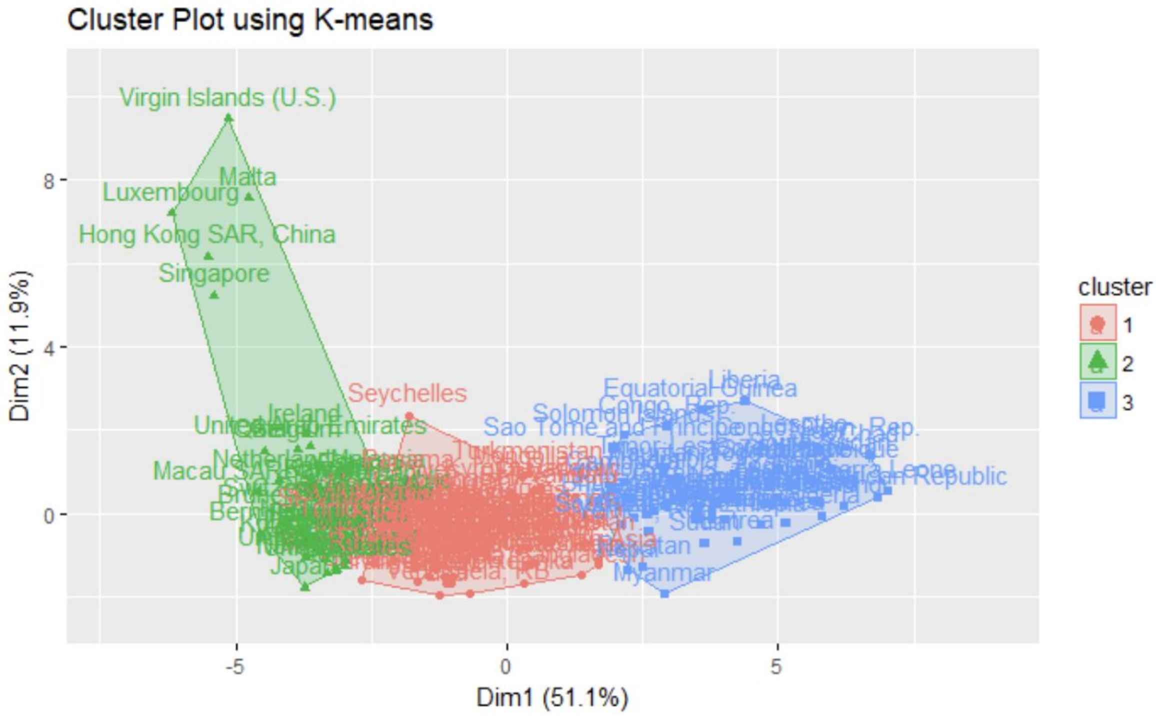


With *k=3*, the clusters are of size 127, 60 and 5 respectively. The small middle cluster of 5 contains Malta, the Virgin Islands(U.S.), Luxembourg, Hong Kong and Singapore. These are the same countries mentioned earlier which we expected to form a cluster because of their high values for trade and FDI. The medium size cluster contains developing countries with a high concentration of this cluster coming from the African continent. These countries are clustered together because of having a lower standard of living based on high mortality rates, lower incomes and limited access to electricity and technology. The final cluster contains all the European countries as well as countries from the Middle East and wealthy countries such as the US, Australia and Canada. This cluster has the middle income to upper-income countries who have access to electricity and technology and have high life expectancies. This divide closely mimics what we saw in PCA.

K–means

K-means is a very popular clustering partitioning algorithm that is fast and efficient and scales well for large datasets. It is an iterative process, so observations can switch between clusters while the algorithm runs until it converges at a local optimum. This method is not robust when it comes to noise data and outliers and is not suitable for clusters with non-convex shapes. Another drawback with K-means is the necessity of specifying K in advance.

For our analysis, it seems that the shape of clusters is likely to be regular based on the PCA biplot. K will be set to 3. To ensure that an optimal solution is found, the algorithm was initiated 30 times with random cluster centres. A visualisation of the clusters is shown in the figure below.



Based on the clustering, it seems that K-means has clustered the countries based on general living conditions, i.e. developing countries, semi-developed and developed countries. With the developing countries having higher poverty, higher mortality rates and less access to electricity and technology and vice versa for the developed countries. The semi-developed countries seem to have an average standard of living by containing characteristics of both the developed and developing countries. The 5 high trade countries are grouped with the developed countries as k-means doesn't find those exports, imports and FDI to be enough of a differentiator as they share the same general living standard with the developed countries.

Both hierarchical clustering and K-means separated the same countries as being developing, mainly African countries. They mainly differ in the separating of the other countries with hierarchical separating based on trade levels and k-means separating based on the general living standards.

Compulsory Task 1

This dataset is from the US Arrests Kaggle challenge (**link**). A description of the data is given as: "This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas."

Follow these steps:

- Use the dataset **UsArrests.csv** included in this folder to generate a similar in-depth PCA report of the data, explore as much as you can, motivate the pre-processing steps you take, and interpret the outcomes of any analyses.
- You are also required to do an application of two clustering techniques and an analysis of the clusters they generate.
- Push all the work that you have generated for this project to GitHub.

Completed the task(s)?

Ask your mentor to review your work!

Review work

Things to look out for:

1. Make sure that you have installed and set up all programs correctly. You have set up **Dropbox** correctly if you are reading this, but **Python or Notepad++** may not be installed correctly.
2. If you are not using Windows, please ask your mentor for alternative instructions.

Rate us

Share your thoughts

Hyperion strives to provide internationally-excellent course content that helps you achieve your learning outcomes.

Think that the content of this task, or this course as a whole, can be improved or think we've done a good job?

Click here to share your thoughts anonymously.