Hyperiondev

TASK

Capstone Project VII

Visit our website

Introduction

WELCOME TO YOUR FINAL CAPSTONE PROJECT TASK!

Well done on making it this far! For this Capstone project, we are first 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 an expert code reviewer to get support on any aspect of your course.

The best way to get help is to login to Discord at **https://discord.com/invite/hyperdev** where our specialist team is ready to support you.

Our team 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 dataset we will be exploring contains data on 44 countries. There are 20 variables for each country in total, with 19 describing each country through population statistics, electricity, and technology adoption, as well as economic indicators such as inflation and trade data.

	Country	Country Groups	BX.KLT.DINV	EG.ELC.ACCS	EG.FEC.RNE\	EN.ATM.CO2	FP.CPI.TOTL.	IT.CEL.SETS.F	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

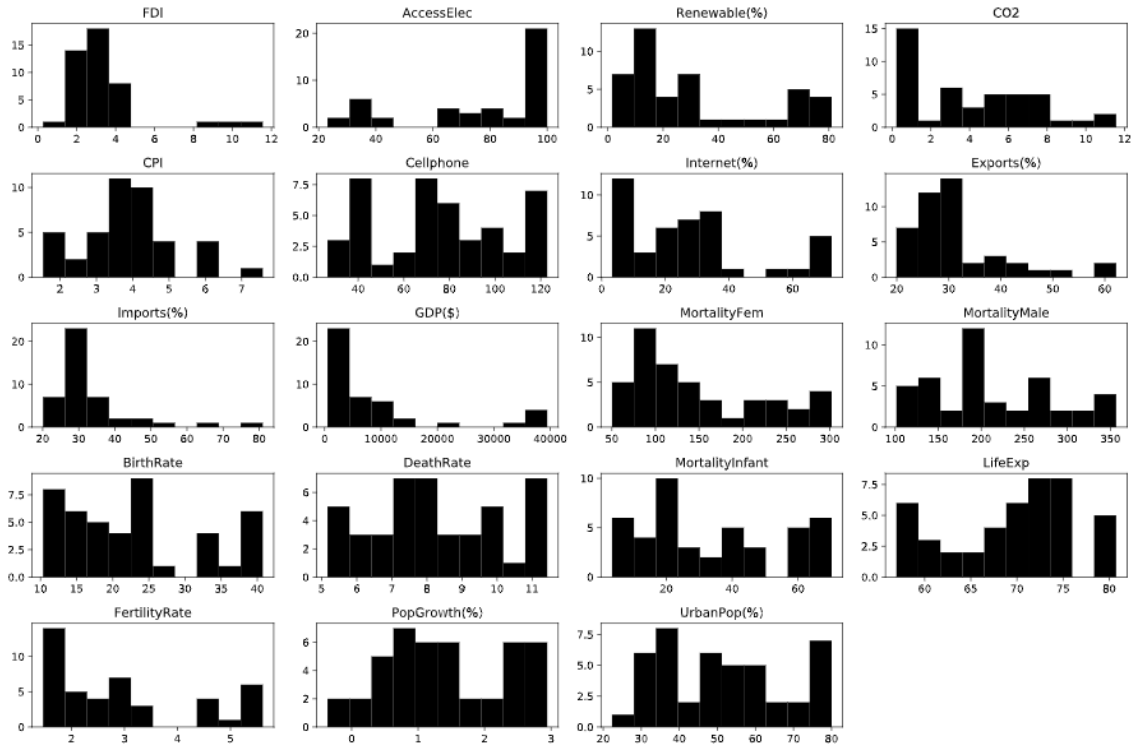
EXPLORING THE DATA

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, are observed. This is summarised in the table below.

	Missing	Mean	StdDev	Min	Max
FDI	0	3.39	2.11	0.27	11.56
AccessElec	0	76.63	26.00	23.09	100.00
Renewable(%)	0	30.77	25.55	1.50	81.01
CO2	0	4.14	3.21	0.22	11.56
CPI	2	3.89	1.29	1.53	7.58
Cellphone	0	76.06	28.70	26.56	122.83
Internet(%)	0	27.39	20.54	3.25	72.29
Exports(%)	0	31.90	9.44	19.93	62.17
Imports(%)	0	32.98	11.35	20.10	80.99
GDP($)	0	8864.04	11163.53	522.03	39449.30

MortalityFem	0	149.35	76.51	50.16	302.16
MortalityMale	0	213.08	72.75	101.46	356.62
BirthRate	0	22.80	9.50	10.36	40.75
DeathRate	0	8.30	1.81	5.17	11.43
MortalityInfant	0	33.49	21.33	3.48	70.22
LifeExp	0	69.19	6.94	56.94	80.75
FertilityRate	0	2.96	1.29	1.47	5.61
PopGrowth(%)	0	1.41	0.92	-0.36	2.94
UrbanPop(%)	0	51.31	16.82	22.30	80.02

At first glance, the GDP per capita variable stands out as having a mean and standard deviation which are significantly 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. We can also get insight into the spread of the data by plotting histograms for each variable:



MISSING VALUES

It is common when working with datasets to have missing values. Below is a sample showing some missing data:

Country	Renewable(%)	CO2	CPI	Cellphone
CEB	14.54	6.82	1.84	122.19
CSS	9.10	9.28	3.25	113.63
EAP	16.47	5.11	3.79	69.91
EAR	26.48	2.12	4.58	68.07
EAS	13.73	5.70	3.25	73.66
ECA	6.29	7.52	6.28	122.83
ECS	10.83	7.54	2.39	120.37
EMU	12.83	7.43	1.53	117.10
EUU	12.96	7.35	1.67	118.57
FCS	51.77	0.83	3.62	38.27
FSM	1.50	1.10	nan	26.56
HIC	9.53	11.56	1.96	109.04

There are 2 missing values in total within the dataset - both in the CPI column. This is relatively few; however, if there were many missing values, excluding these cases from our dataset entirely would be detrimental to our 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 44 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 consumer price index (CPI) for the "FSM" country group:

```
Country Groups     Micronesia, Fed. Sts.
FDI                                 0.27
AccessElec                         64.53
Renewable(%)                        1.50
CO2                                 1.10
CPI                                  NaN
Cellphone                          26.56
Internet(%)                        20.00
Exports(%)                         23.51
Imports(%)                         80.99
GDP($)                           2861.77
MortalityFem                      156.79
MortalityMale                     185.22
BirthRate                          23.75
DeathRate                           6.27
MortalityInfant                    32.40
LifeExp                            68.58
FertilityRate                       3.46
PopGrowth(%)                       -0.33
UrbanPop(%)                        22.30
Name: FSM, dtype: object
```

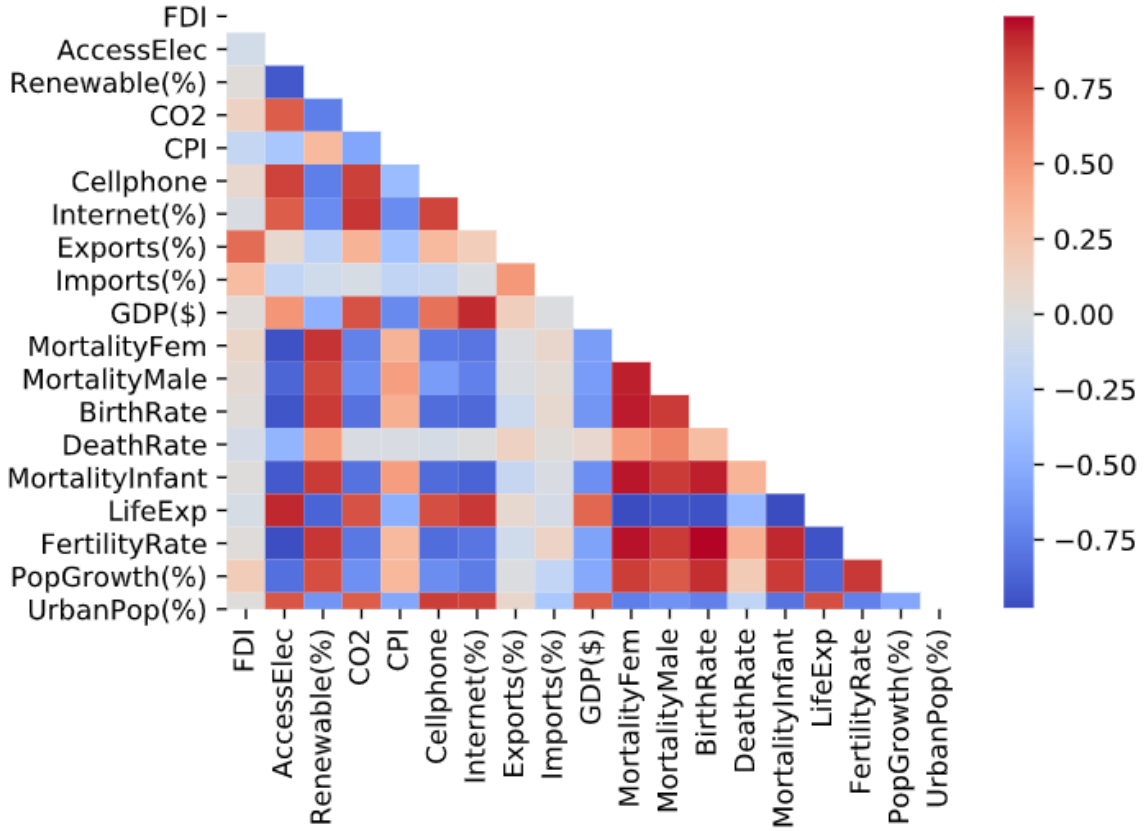
KNN will investigate other observations with similar values to all the other variables. After the identification, it will get the mean of the consumer prices of these observations and impute the value as the value of consumer price for Bermuda.

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.

Country	Original	Imputed
CEB	1.84	1.84
CSS	3.25	3.25
EAP	3.79	3.79
EAR	4.58	4.58
EAS	3.25	3.25
ECA	6.28	6.28
ECS	2.39	2.39
EMU	1.53	1.53
EUU	1.67	1.67
FCS	3.62	3.62
FSM	nan	3.89

CORRELATION ANALYSIS

From the plot below, most of the variables are highly positively or negatively correlated with each other. For example, access to electricity (% of the population) has a strong negative correlation with renewable energy, female mortality, and the fertility rate. Similarly, the birth rate is strongly positively correlated with infant mortality and fertility rate.



From the correlation plot, it is evident that Foreign Direct Investment (FDI) has a relatively 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 - as well as CO2 emissions!

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 fuel production or procurement 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

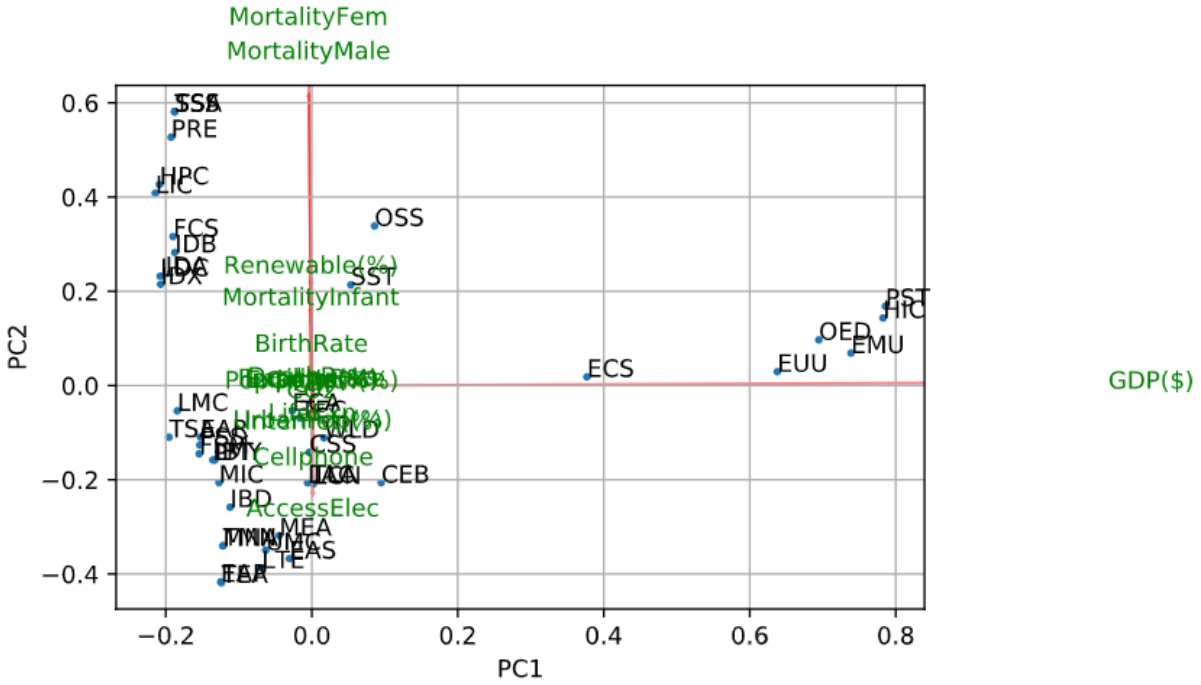
Principal Components Analysis (PCA) is a method for finding the underlying variables (i.e. principal components) that best differentiate the observations by determining the directions along which your data points are most spread out. Since the determination of the principal components is based on finding 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

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.

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6 ...**
Standard deviation	1.12e+04	9.02e+01	2.74e+01	1.52e+01	1.12e+01	8.52e+00
Proportion of Variance Explained	1.00e+00	6.53e-05	6.01e-06	1.85e-06	1.00e-06	5.82e-07
Cumulative Proportion	1.25e+08	1.25e+08	1.25e+08	1.25e+08	1.25e+08	1.25e+08

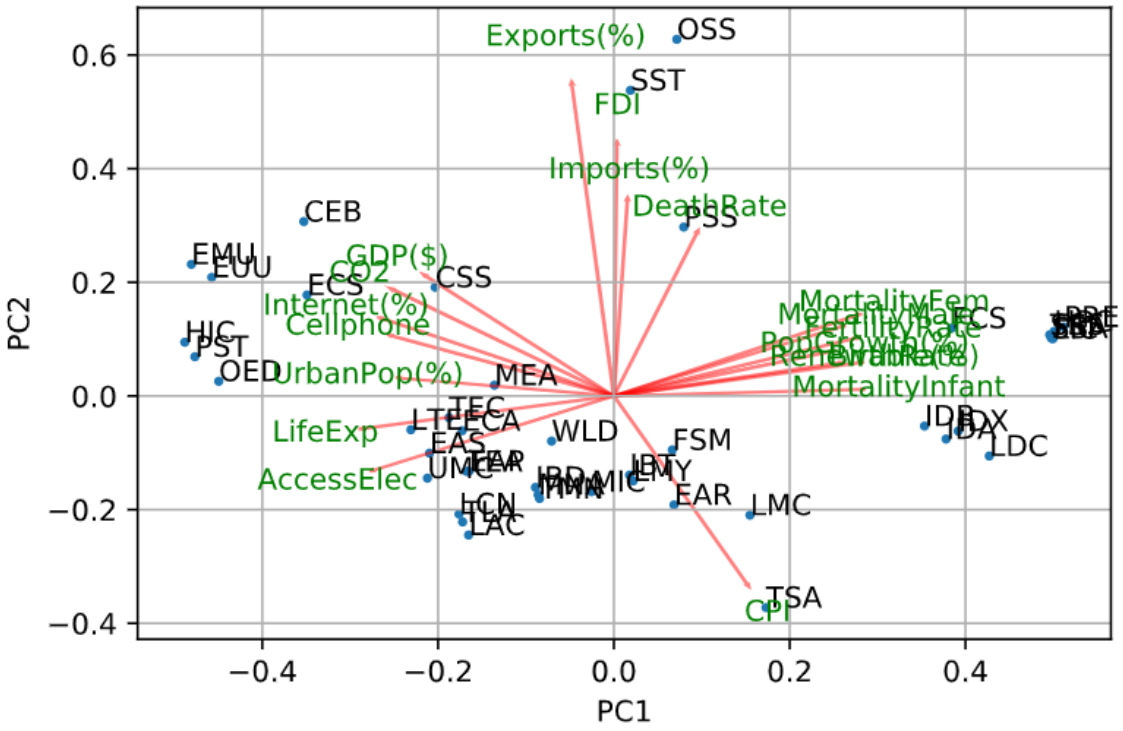
If we consider the biplot for these components, as expected, the first principal component is dominated by GDP which is on a much larger scale than the other variables (as seen during data exploration). This makes it difficult to see how countries vary with respect to the other variables or read the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



It can, however, be seen that the country groups HI and PST (High Income and Post-demographic dividend) have the highest 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 100% and 0.0000653% 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. This measure makes it possible to read the biplot (below) 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.

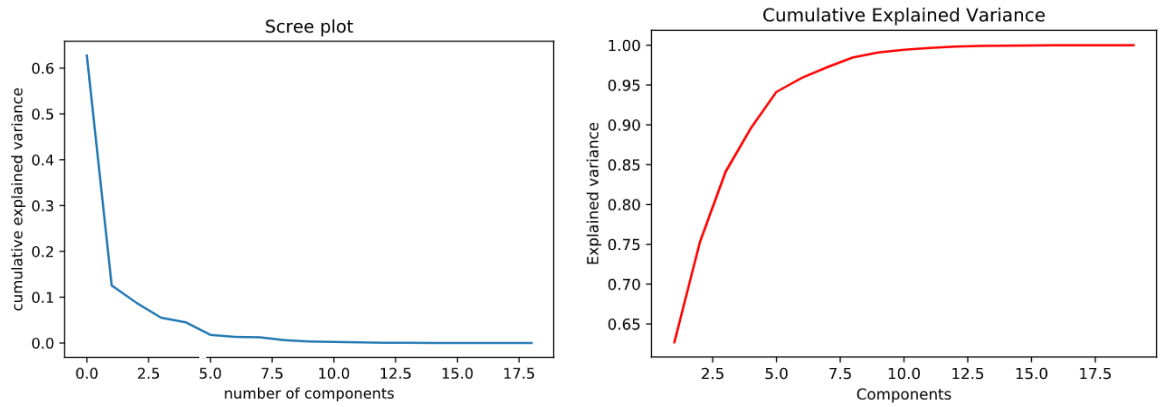
Countries with a lower standard of living are grouped to the right-hand side in the first principle component, such as the groups: FCS (Fragile and conflict-affected situations), SSA (Sub-Saharan Africa (excluding high income)), etc. These countries tend to have high mortality rates and high population growth rates.

In the centre, we have countries such as LAC (Latin America & Caribbean (excluding high income)) and MIC (Micronesia, Fed. Sts.). These are middle-income countries which are still developing but have a higher standard of living than those just discussed.

The countries to the left-hand side in the first principle component are those that have a good standard of living, such as the groups: HIC (High income), CEB (Central Europe and the Baltics), EMU (Euro area), EEU (European Union), etc. We see that these countries are correlated with GDP, cellphone and Internet usage, life expectancy etc.

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 country groups that are extremely above average in these variables are PSS (Pacific island small states), OSS (Other small states) and SST (Small states). These 3 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. As such, 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 Explained Variance plot, below.



The first 5 principal components together explain around 90% 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.

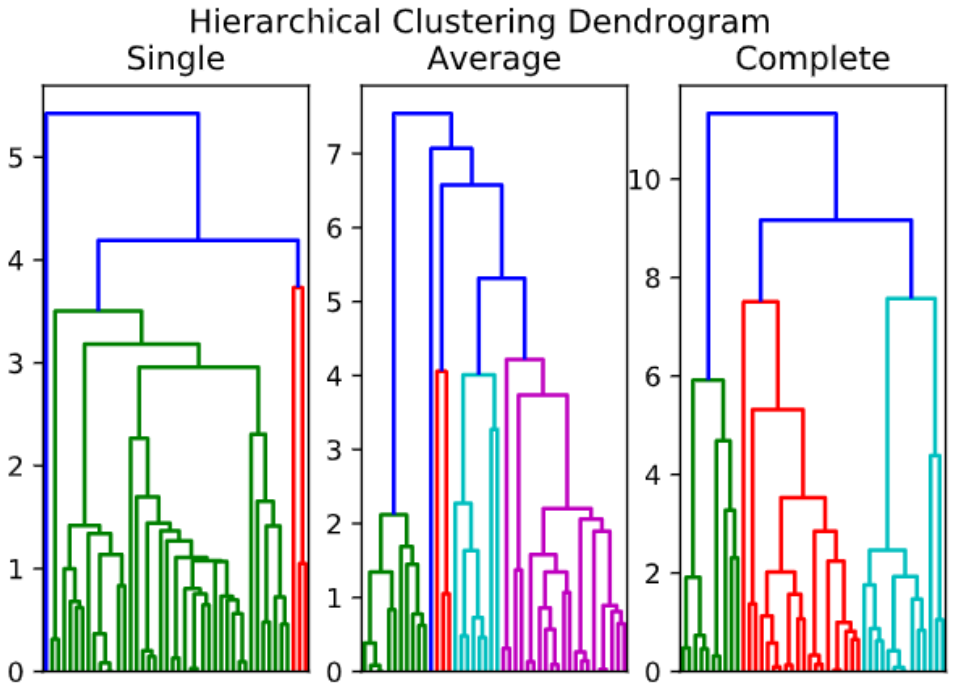
CLUSTER ANALYSIS

We will perform both Hierarchical Clustering and K-means with these data and compare the results.

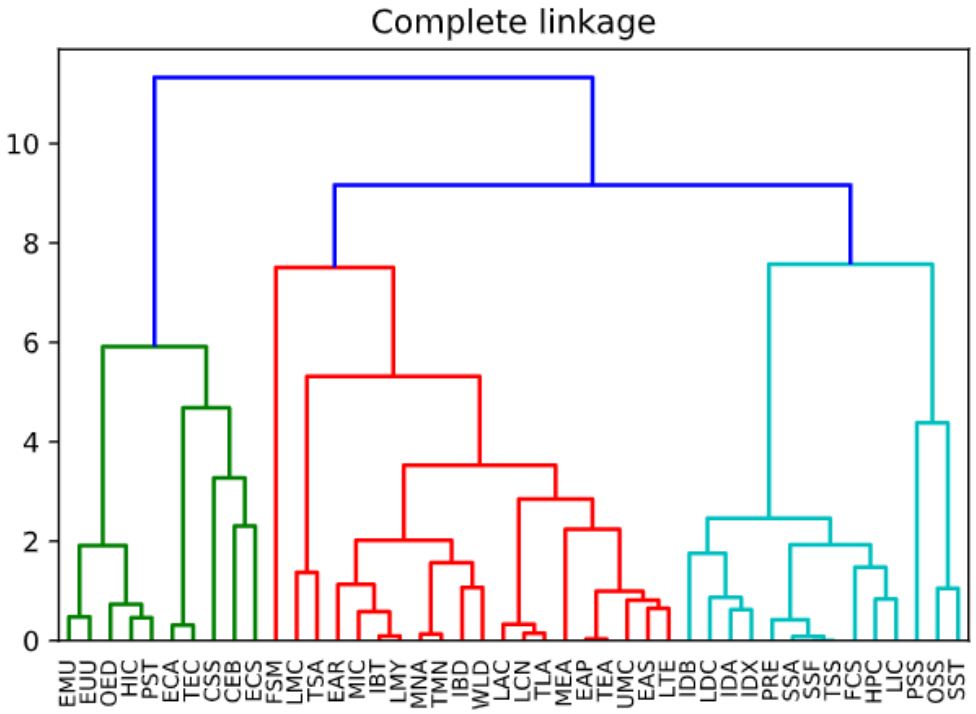
Hierarchical clustering

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.

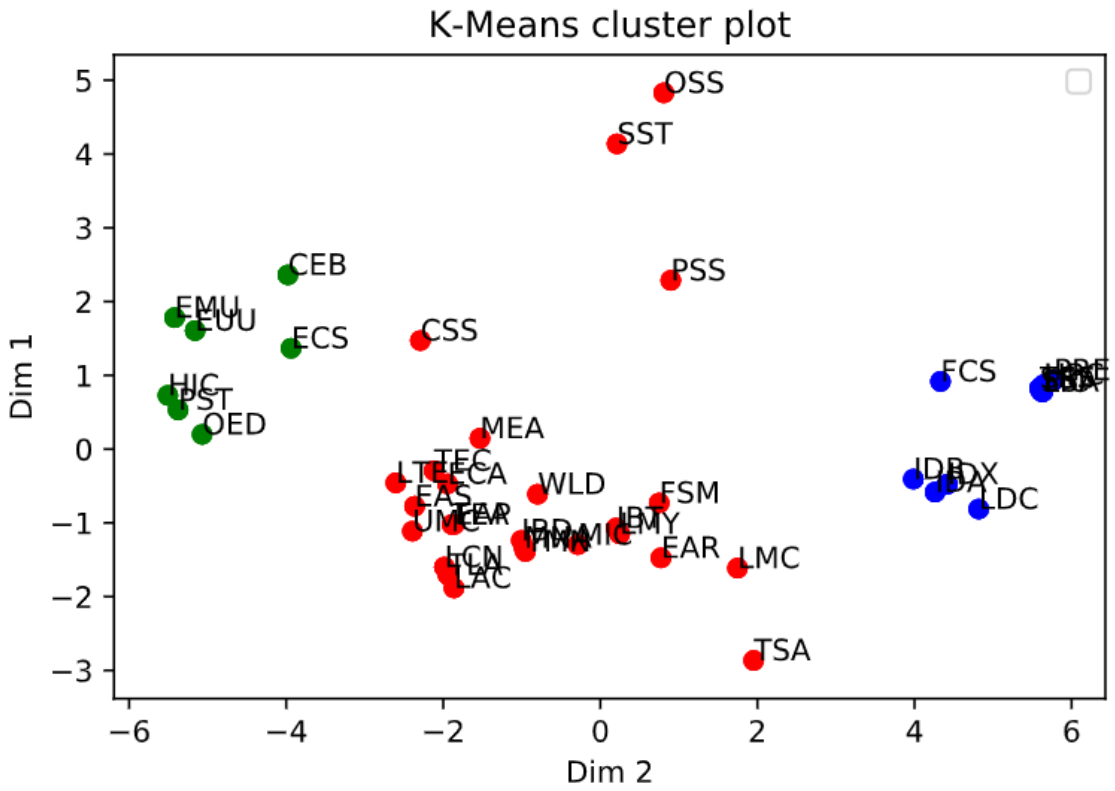


With *k=3*, the clusters are of size 10, 20, and 14 respectively. Within the pale blue cluster, the right-most branch at height 6 contains the 'Small states' country groups we noted earlier: PSS, OSS, and SST. These are the same countries mentioned earlier which we expected to form a cluster because of their high values for trade and FDI. The pale blue cluster at large contains country groups that are the least developed. The red cluster contains developing countries. 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 green cluster contains all the wealthy, most developed countries. This cluster has the middle- 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.

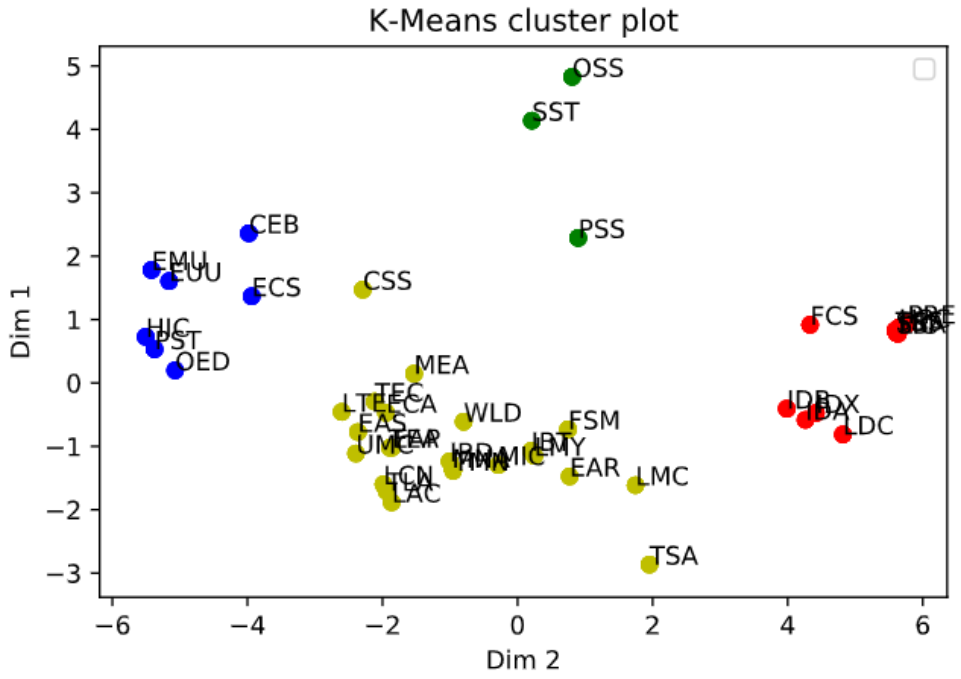
For our analysis, it seems that the shape of clusters is likely to be regular based on the PCA biplot. K will be set to 3. A visualisation of the clusters is shown in the figure below.



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. The developing countries have 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 3 high-trade countries are also in the semi-developed cluster, 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 grouped the countries together similarly. 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. This results in a few differences, such as into which group the 3 high-trade countries fall. Other small differences occur at the boundaries between the different clusters, for example, which group CSS (Caribbean small states) falls into.

If we were to set K=4, then the three high-trade countries are separated into their own cluster:



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. Try and see if you can find anything common within each cluster that has been found.
- Push all the work that you have generated for this project to GitHub.

If you are having any difficulties, please feel free to contact our specialist team **on Discord** for support.

Completed the task(s)?

Ask an expert to review your work!

Review work

Rate us

Share your thoughts

HyperionDev 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.