Unsupervised Learning Assignment 1

TASK 1

Question 1:

Priciple Component Analysis(PCA) is used to reduce the correlation between variables by projecting them onto orthogonal planes which capture the most variation in the data. PCA's goal is to retain the most information from the variables whilst removing any repeated information/noise. Since the greatest variation of the data is projected onto the initial planes, the data can therefore be looked at in a lower dimensional space. This allows for better interoperability, visualization, and simpler models can be used reducing the need for larger data sets and also reducing the chances of over fitting. Since the correlation is removed between the variables the inference that can be taken from the models is more accurate. And with the removal of noise the models are less likely to start fitting to the noise.

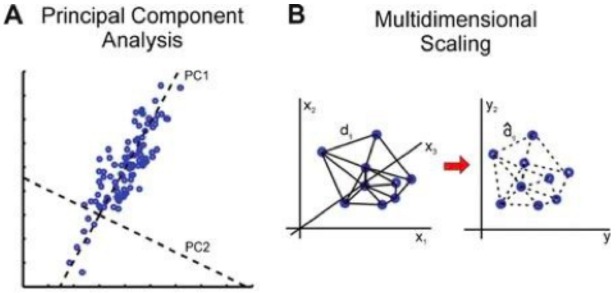
Multi-Dimensional Scaling(MDS) is used to visualize data in lower dimensions and help find clusters as they become more apparent in lower dimensions. MDS preseves the distance between the points found in higher dimension in lower dimension. This is done as it's hard to visualize distance between points and cluster in higher dimension. This method takes the calculated distances between each observation and finds co ordinates for the observation in a way that preserves the distances between them. These new co ordinates are generally kept at 3 or below as then when plotted it's easy to visualize the obsevations relative to each other.

Comparison

The biggest difference between the two methods is that PCA aims to preserve the variation in the data, whilst MDS aims to preserve the distances beween the data points and keep the spacial relationships.

Whilst PCA only works with continuous data, MDS is also able to work with ordinal data by rather preserving rank distances between the data.

The figure below highlights the differences between the two methods. PCA shown on the left has fitted the components along the decreasing lines of most variation, orthoganal to each other. Whilst MDS shown on the right preserves the spacial relationahips.

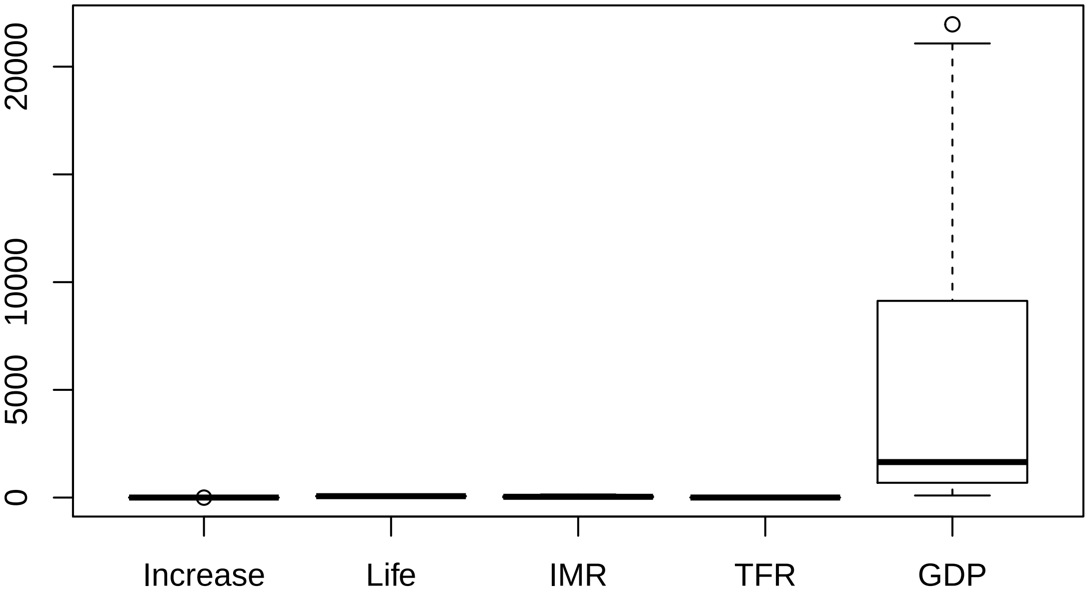


Question 2

Exploratory Data Analysis

A summary of the data is shown in the table below. The data contains 5 continuous values, all in different ranges. The different ranges can be viewed in the graph below. Since these variables are measured on different units the data will need to be scaled. If the variables are not scaled it will result in some variables contributing more and overpowering the other variables.

Country	Increase	Life	IMR	TFR	GDP
Albania : 1	Min. :-1.500	Min. :45.00	Min. : 7.0	Min. :1.30	Min. : 97.39
Argentina: 1	1st Qu.: 1.000	1st Qu.:60.60	1st Qu.: 9.0	1st Qu.:1.70	1st Qu.: 686.75
Australia: 1	Median : 1.500	Median :66.60	Median : 37.0	Median :2.90	Median : 1647.97
Austria : 1	Mean : 1.584	Mean :64.39	Mean : 44.8	Mean :3.46	Mean : 6349.31
Benin : 1	3rd Qu.: 2.400	3rd Qu.:72.50	3rd Qu.: 67.0	3rd Qu.:5.00	3rd Qu.: 9129.34
Boliva : 1	Max. : 4.400	Max. :75.00	Max. :143.0	Max. :7.20	Max. :21965.08
(Other) :19	NA	NA	NA	NA	NA



From the correlation table below, it can be seen the variables are all highly correlated, this goes to show they all contain similar information for ranking development.

	Increase	Life	IMR	TFR	GDP
Increase	1.0000000	-0.7309788	0.7529396	0.8550333	-0.4764443
Life	-0.7309788	1.0000000	-0.9220168	-0.8870387	0.6871920
IMR	0.7529396	-0.9220168	1.0000000	0.8761029	-0.6773342
TFR	0.8550333	-0.8870387	0.8761029	1.0000000	-0.6013634
GDP	-0.4764443	0.6871920	-0.6773342	-0.6013634	1.0000000

Data preprocessing

Before MDS can be applied there are a few steps that need taken.

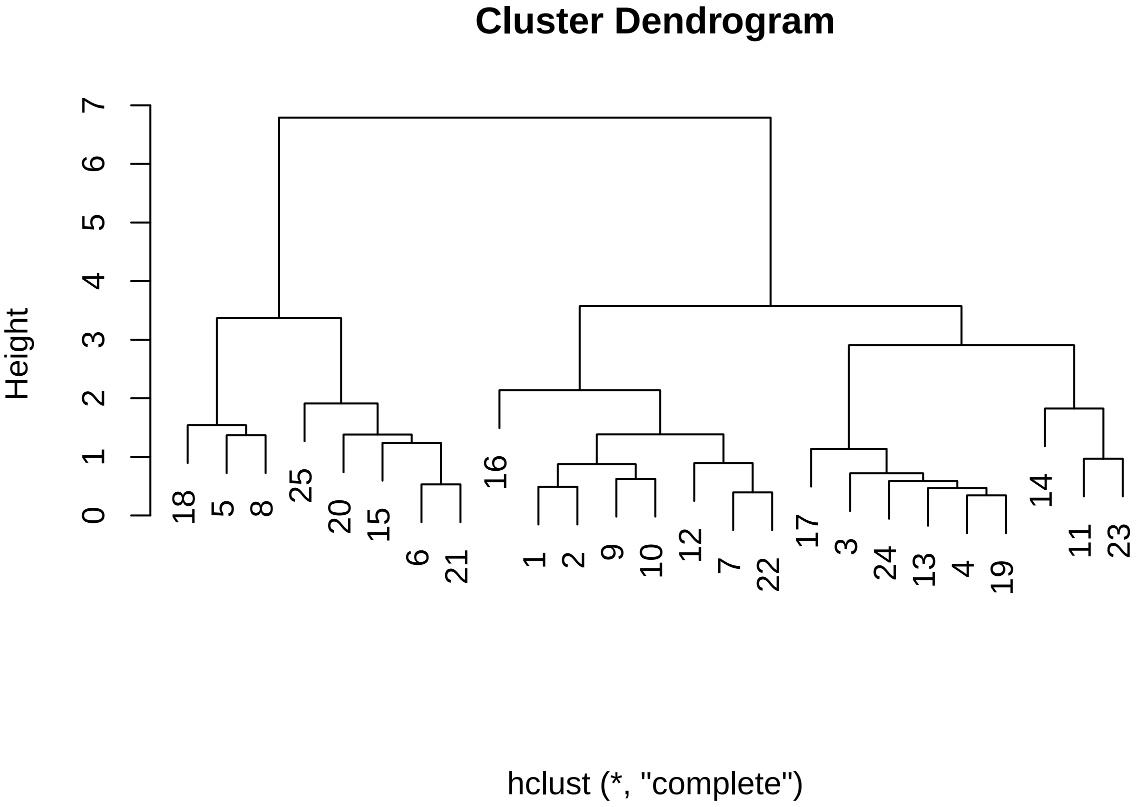
1. Explore the data to make sure there are no missing values and find out what type of data you are working with -Is it ordinal? -Is the data measured on the same scale?

2.If the data is not measured on the same scale it will need to be scaled so that it is all in the same range and one variable won't over power another

3. Once the variables are on the same scale, the distances between the observations need to be calculated and put into a distance matrix before MDS can be performed. The distance metric will depend on the data itself. These are the distances the MDS will try preserve.

Clustering

Clustering of the data was performed for the purpose of visualizing how the methods preform. Hierachal clustering with complete distances was used. From the dendrogram 6 clusters were chosen.

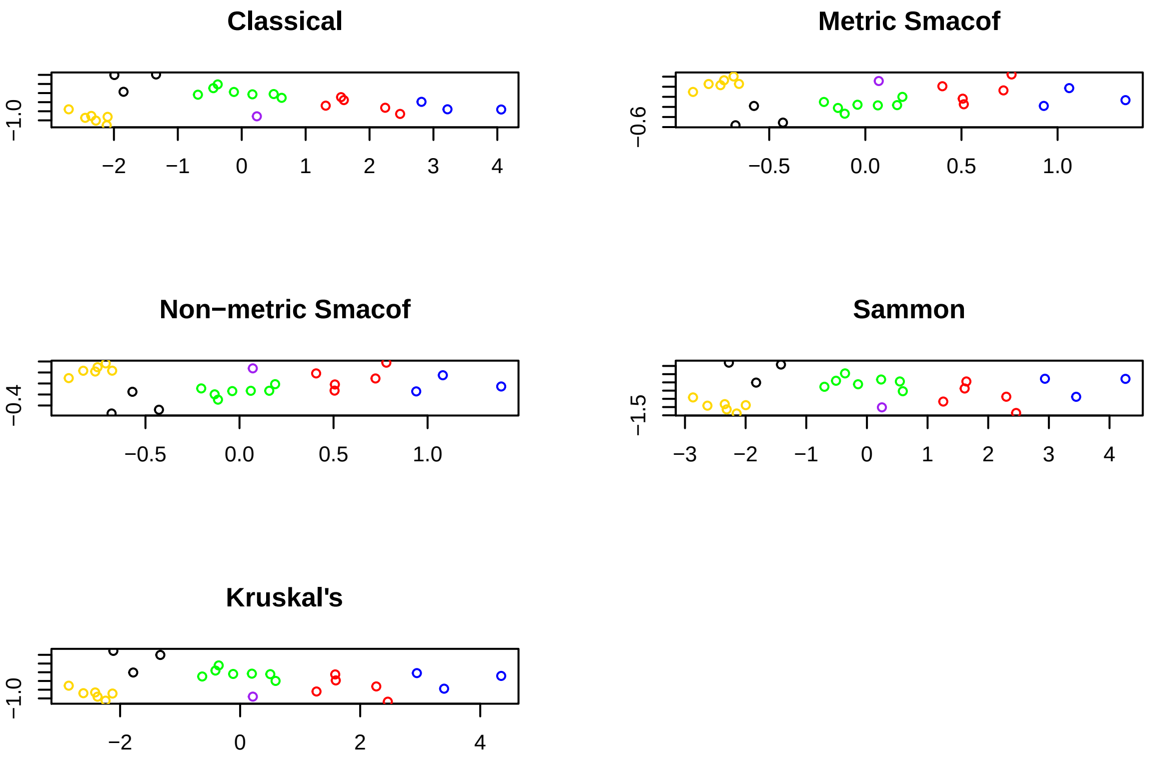


Choosing between methods

Different MDS functions were explored to find the method that works best with this data.

Since euclidean distance was used R^2 and stress were used for comparison between the methods.

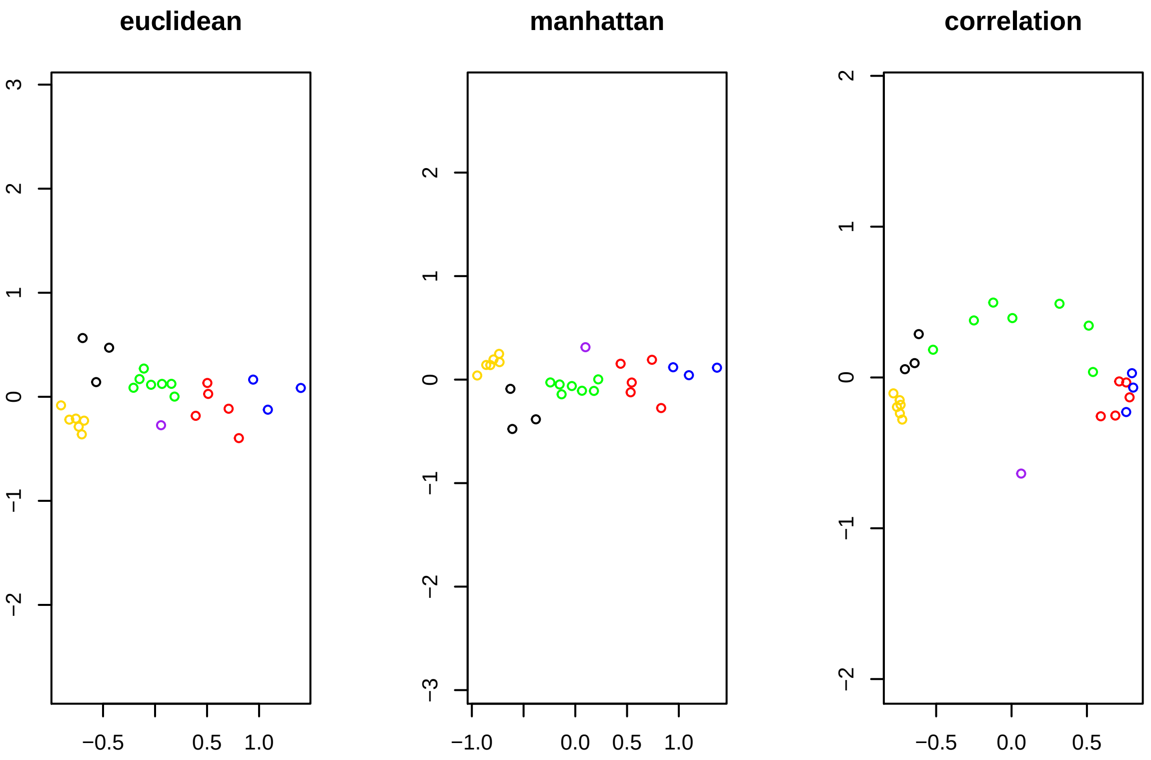
The non-metric ordinal smacof method worked best, and so was used for the rest of the assignment. This gives indication that the data was non-linear.



Method	R^2	p value	Stress
Classical	0.9858	0	0.0838
Metric Smacof	0.9924	0	0.0507
Non-Metric Smacof	0.9943	0	0.0251
Sammon	0.9907	0	0.0538
Kruskals	0.9943	0	0.0579

Choosing a distance metric

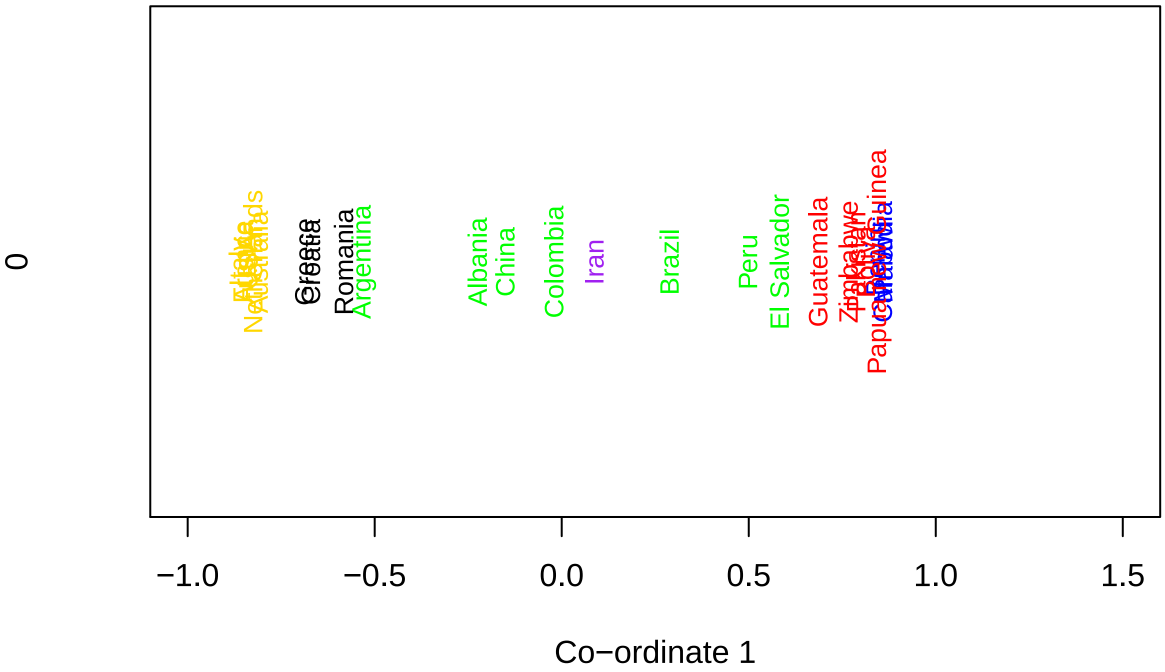
Ordinal MDS chosen above was used to explore the best distance metric for this method. The euclidean distance worked the best, this can be seen in the table as it has the highest R^2 and lowest stress value meaning the MDS was able to capture the initial distance matrix the best. By using euclidean distance it will help us determine what the co-ordinates are comprised of, which is not usually possible with MDS.



Distance	R^2	p value	stress	f norm
euclidean	0.9944	0	0.0249	71
manhattan	0.9935	0	0.0303	170
correlation	0.9632	0	0.0736	6

1D Scaling

The distances projected into 1D corordinates are mapped on the graph below. It places the different clusters closer together relatively well with the exception of two instances. The first being Iran which was placed in the area of the green cluster which makes sense based on it's placement in the dendrogram (having only a small height between Iran an the green clsuter). The second being Pakistan, Zimbabwe and Cambodia being placed closer together compared rather than closer to the clusters.

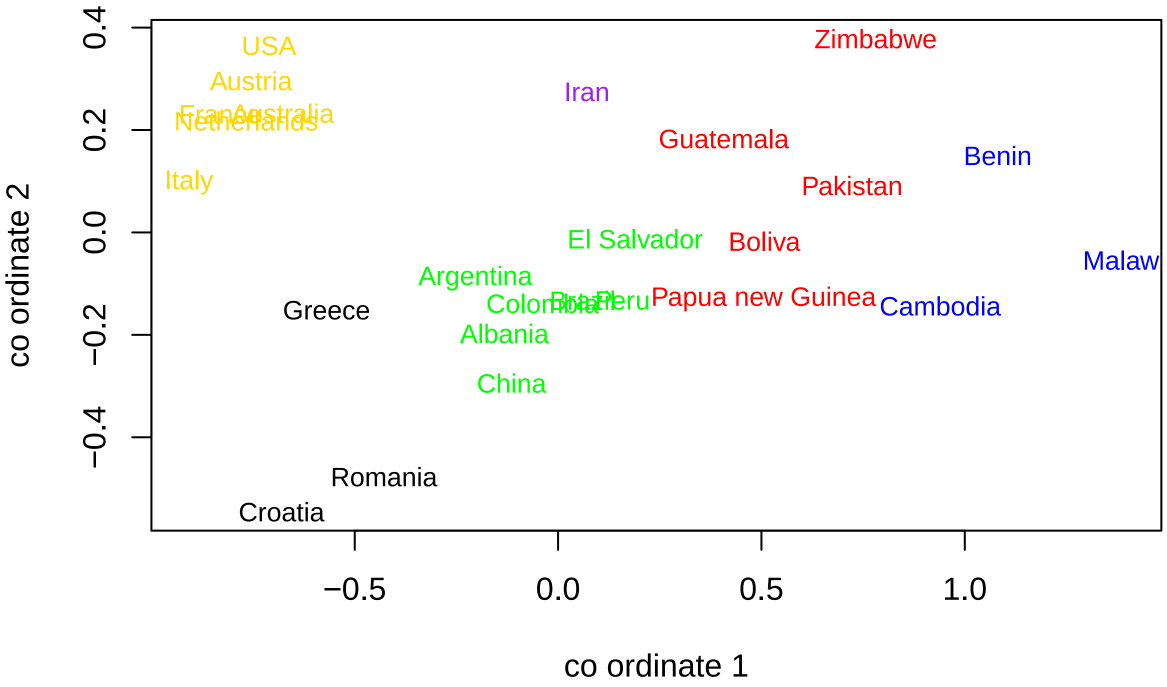


Increase	0.8259304120042
Life	-0.881163900564394
IMR	0.890973090914814
TFR	0.902931824148641
GDP	-0.774440294489298

All the variables are highly correlated as they are all ranked measurements of development. This can be seen as all the variables are highly correlated with the first co-ordinate. Because of this the first component being a combination of all the factors it goes to show that it will be a good measure to rank development on. Visually this can be seen in the graph, first world countries are on the left, whilst third world countries are on the right. Besides Iran the clusters are still kept together based on rank.

2D scaling

Using 2 cordinates allows the clusters to be captured better as we can now see that Iran is a cluster on it's own. It also seperates Zimbabwe and Pakistan further from Combodia and the blue cluster. None of the clusters overlap in this 2D space indicating to a better fitting soltion compared to the 1D solution.



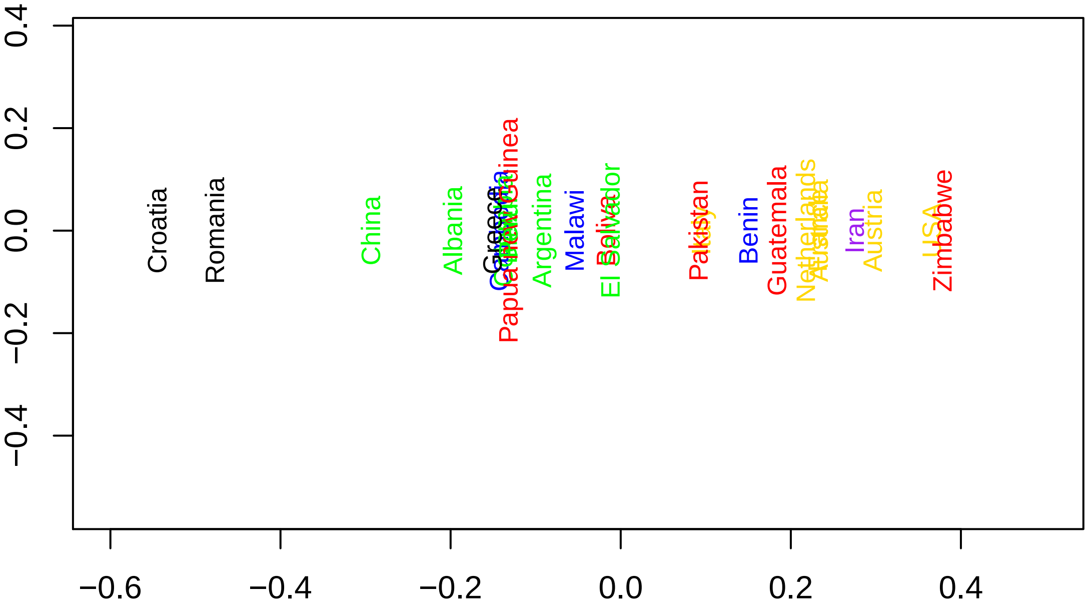
Cambodia and Zimbabwe moved further apart due to the second component. The biggest differences between these two countries are in Increase IMR and GDP. This gives indication that these are measures in the second dimension.

	Country	Increase	Life	IMR	TFR	GDP
8	Cambodia	2.8	50.1	116	5.3	97.39
25	Zimbabwe	4.4	52.4	67	5.0	686.75

When adding in the second co ordinate Iran moved out of the area of the green cluster into it's own space. When looking at how Iran differs from the other coutries in the green cluster we can see it's GDP value is higher than the rest of the values, this gives more indication to the second dimension measuring GDP. Besides EL Salvidor it also has outliying values in the TFR column and Increase column suggesting these are also measurements of the second co ordinate. Since El Salvidor had similar values in TFR and Increase but still remained in range of the green cluster goes to show that GDP is the biggest contributer to the co ordinate. El Salvidor is near the top right of the cluster closer to Iran also proving to this point.

	Country	Increase	Life	IMR	TFR	GDP
1	Albania	1.2	69.2	30	2.9	659.91
2	Argentina	1.2	68.6	24	2.8	4343.04
7	Brazil	1.5	64.0	58	2.9	3219.22
9	China	1.1	66.7	44	2.0	341.31
10	Colombia	1.7	66.4	37	2.7	1246.87
12	El Salvador	2.2	63.9	46	4.0	988.58
16	Iran	2.3	67.0	36	5.0	9129.34
22	Peru	1.7	64.1	64	3.4	1674.15

The second co ordinate can be seen in the graph below

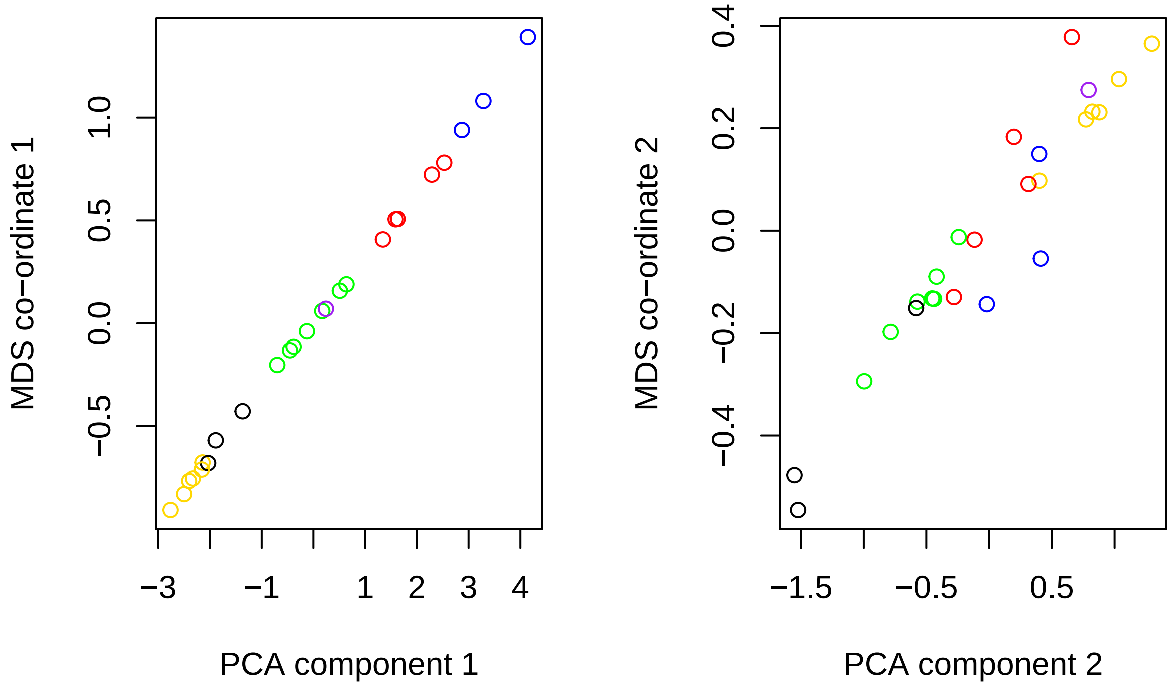


The correlation between each variable and the second co ordinate are shown below. From this it can be seen that the largest part of the second component is made up of GDP, then Increase then a bit of TFR. This confirms what we were seeing earlier.

Increase	0.449114236695348
Life	0.0727411166934483
IMR	-0.0857452809622315
TFR	0.177724448398359
GDP	0.540090147672412

Using PCA to interpret

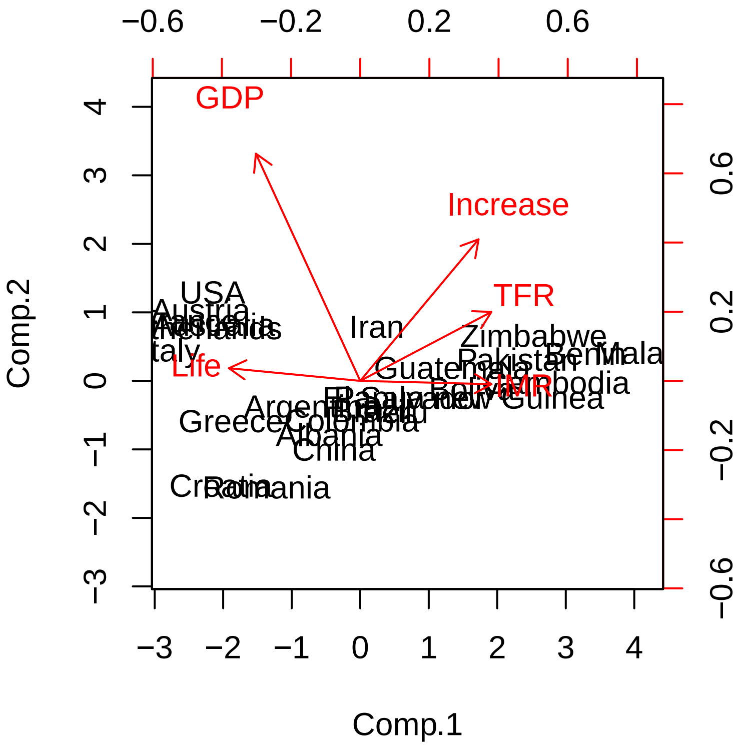
When using euclidian distance PCA and classical MDS result in the same thing. From comparing methods earlier, it was seen that the methods produced similar results. The differences between the non-metric MDS we performed and PCA can be viewed in the graphs below. There is some variation between the methods but they seem to have the same general trend. Since, they had the same general trends the loading scores of PCA were used to help confirm what each axis consisted of.



From the loading scores and the biplot it can be seen that the second dimension consists of mainly GDP(Gross Domestic Product) followed by Increase(Population Growth Increase) and then TFR(total fertility rate). This confirms our initial deductions.

```
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## Increase  0.428  0.511  0.645  0.284  0.241
## Life     -0.474         0.476  0.162 -0.721
## IMR       0.475        -0.417  0.634 -0.446
## TFR       0.474  0.249        -0.700 -0.472
## GDP      -0.377  0.821 -0.426
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## SS loadings       1.0    1.0    1.0    1.0    1.0
## Proportion Var    0.2    0.2    0.2    0.2    0.2
## Cumulative Var    0.2    0.4    0.6    0.8    1.0
```

It is interesting to note in the biplot that Zimbabwe is not near the end of the GDP axis, which could mean that this is newer data, as the GDP of Zimbabwe has started to increase since around 2007.



TASK 2

Exploratory Data Analysis

A summary of the data is shown in the table below. The data contains 7 continuous values, all in different ranges due to the distances of the races. Since these variables are measured on different ranges the data will need to be scaled. If the variables are not scaled it will result in some variables contributing more and overpowering the other variables.

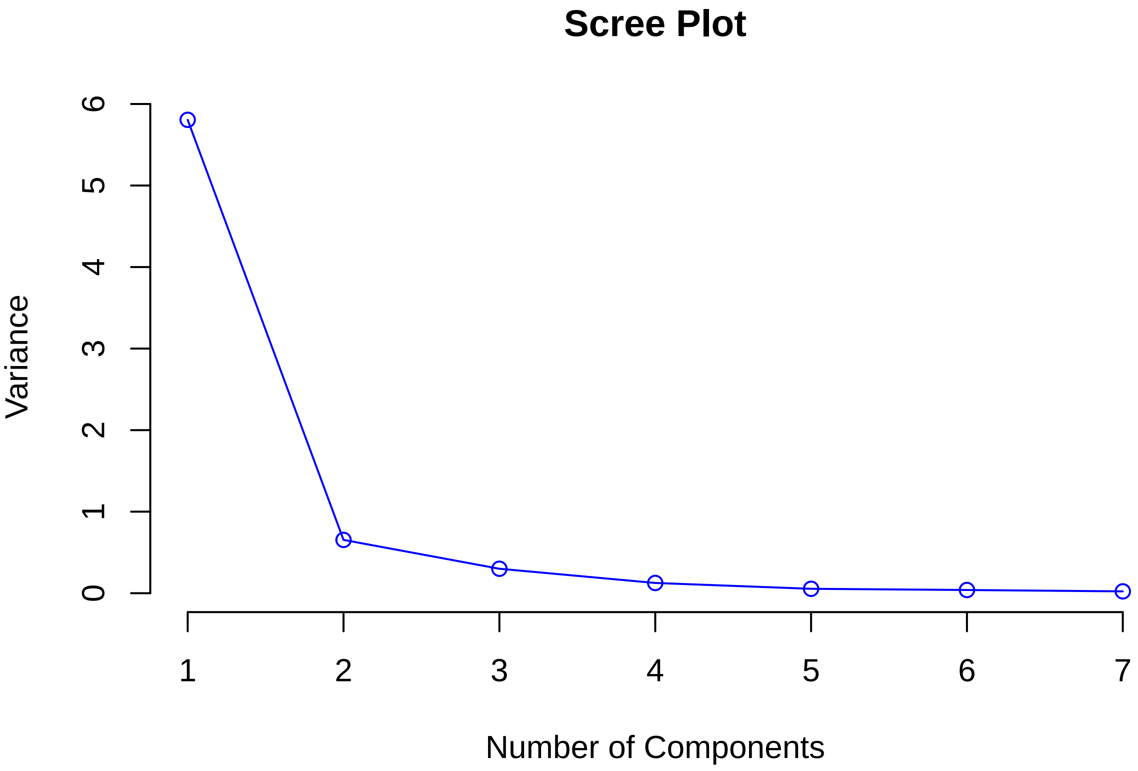
COUNTRY	X100m	X200m	X400m	X800m	X1500m	X3000m	Ma
argentin: 1	Min. :10.79	Min. :21.71	Min. :47.99	Min. :1.890	Min. :3.870	Min. : 8.450	Mir
australi: 1	1st Qu.:11.27	1st Qu.:22.88	1st Qu.:51.55	1st Qu.:2.000	1st Qu.:4.115	1st Qu.: 8.860	1st (
austria : 1	Median :11.60	Median :23.54	Median :53.30	Median :2.050	Median :4.250	Median : 9.340	Medi
belgium : 1	Mean :11.62	Mean :23.64	Mean :53.41	Mean :2.076	Mean :4.325	Mean : 9.448	Mea
bermuda : 1	3rd Qu.:11.92	3rd Qu.:24.43	3rd Qu.:55.04	3rd Qu.:2.150	3rd Qu.:4.470	3rd Qu.: 9.840	3rd (
brazil : 1	Max. :12.90	Max. :27.10	Max. :60.40	Max. :2.330	Max. :5.810	Max. :13.040	Max
(Other) :49	NA	NA	NA	NA	NA	NA	

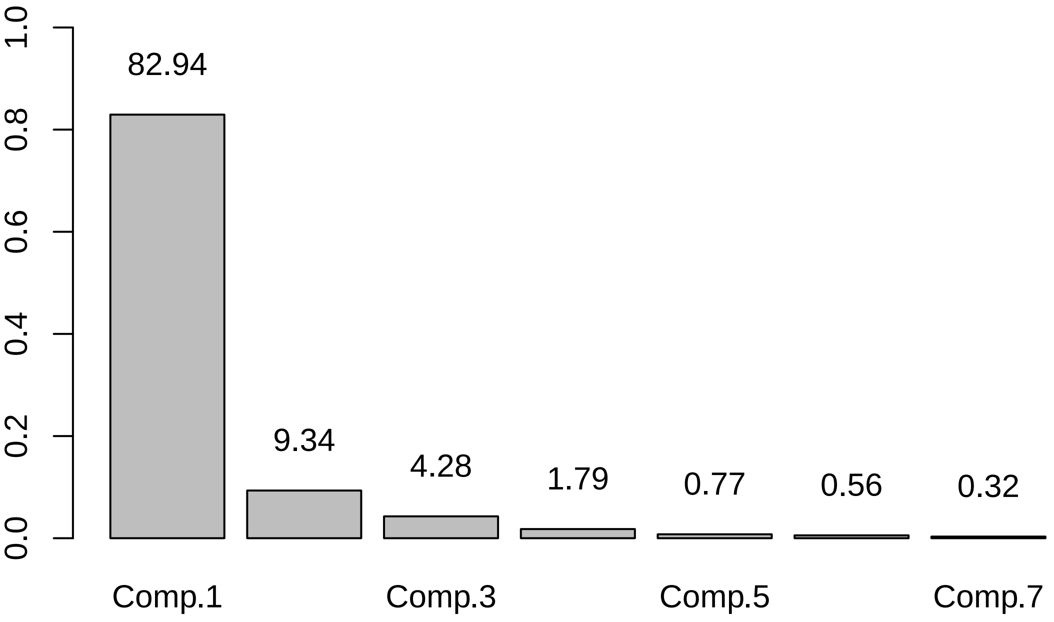
	X100m	X200m	X400m	X800m	X1500m	X3000m	Marathon
X100m	1.0000000	0.9527911	0.8346918	0.7276888	0.7283709	0.7416988	0.6863358
X200m	0.9527911	1.0000000	0.8569621	0.7240597	0.6983643	0.7098710	0.6855745
X400m	0.8346918	0.8569621	1.0000000	0.8984052	0.7878417	0.7776369	0.7054241
X800m	0.7276888	0.7240597	0.8984052	1.0000000	0.9016138	0.8635652	0.7792922
X1500m	0.7283709	0.6983643	0.7878417	0.9016138	1.0000000	0.9691690	0.8779334
X3000m	0.7416988	0.7098710	0.7776369	0.8635652	0.9691690	1.0000000	0.8998374
Marathon	0.6863358	0.6855745	0.7054241	0.7792922	0.8779334	0.8998374	1.0000000

The variables are highly correlated with each other, this gives agood indication that the information can be projected into a smaller sub-space, and that PCA will be a useful method.

Variation Proportions

The first component explains 82.94% of the variation in the data, whilst the second component explains 9.34% of the data. From the scree plot we can see that 2 components should be efficient to explain most of the information contained in the data.





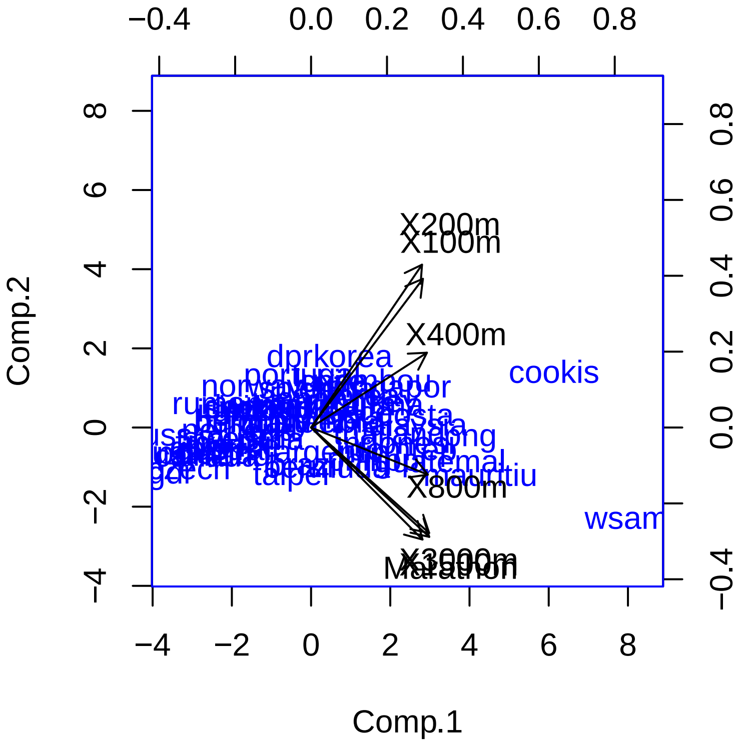
Interpretation of components

Since the input variables are scaled, and the loading scores for component one are pretty similar, the first component takes into account all the types of races and so accounts for the countries overall track ability.

The second component shows the rank of countries over the range of difference distance races. Higher values in the second component indicate the country is bad at short distance races, whilst a negative value would indicate the country is worse at long distance races. And so the where a country lies on the second component gives indication to what types of races they are better at.

```
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## X100m     0.368  0.490  0.286  0.319  0.231  0.620       
## X200m     0.365  0.537  0.230               -0.711 -0.109
## X400m     0.382  0.247 -0.515 -0.347 -0.572  0.191  0.208
## X800m     0.385 -0.155 -0.585         0.620        -0.315
## X1500m    0.389 -0.360         0.430        -0.231  0.693
## X3000m    0.389 -0.348  0.153  0.363 -0.463        -0.598
## Marathon  0.367 -0.369  0.484 -0.672  0.131  0.142       
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

The biplot reiterates this, with all the variables pointing positively along the first component, since the axis measures the overall abilty the better athletic countries are on the left whilst the worst are on the right. The second component has shorter races positively on the second component and longer races on the negative. Countries which are better at long distance races would be at the bottom of the plot, whilst visa versa countries which are better at short distance would be at the top of the plot.



Ranking on the first component should give a good indication of a country's athletic excellence, as its made up of the overall performace over all the races. The negative of the component values were shown below, this is because the lower the variables (time) the better the result. The top 3 countries with the biggest negative values were: Great Dominican Republic, Russia and the USA, these countries are known for having good athletic abilities. The bottom three countries with the biggest positive values are: Wsamoa, Cookis and Mauritius, these countries aren't known for their athletic ability. And so this co ordinate as rank of the countries atheleic excellence makes sense to me from my understanding of countries athletic abilities.

A country that stood out was Kenya as I always understood them to be ranked high in terms of athletic ability, and so it being in the middle of the rank seemed odd. This could be due to them being better long distance races which would be captured better by the second component.

Rank	Country
1	gdr
2	ussr
3	usa
4	czech
5	frg
6	gbni
7	poland
8	canada
9	finland
10	italy
11	australi
12	rumania
13	france
14	sweden
15	netherla
16	nz
17	belgium
18	norway
19	hungary
20	austria
21	switzerl
22	ireland
23	denmark
24	taipei
25	kenya
26	spain
27	portugal
28	israel
29	brazil
30	mexico
31	japan
32	columbia
33	bermuda
34	dprkorea
35	argentin
36	chile
37	china
38	greece
39	india
40	korea
41	luxembou
42	turkey
43	philippi
44	burma

Rank	Country
45	thailand
46	singapor
47	indonesi
48	domrep
49	malaysia
50	costa
51	guatemal
52	png
53	mauritiu
54	cookis
55	wsamoa