Final Exam

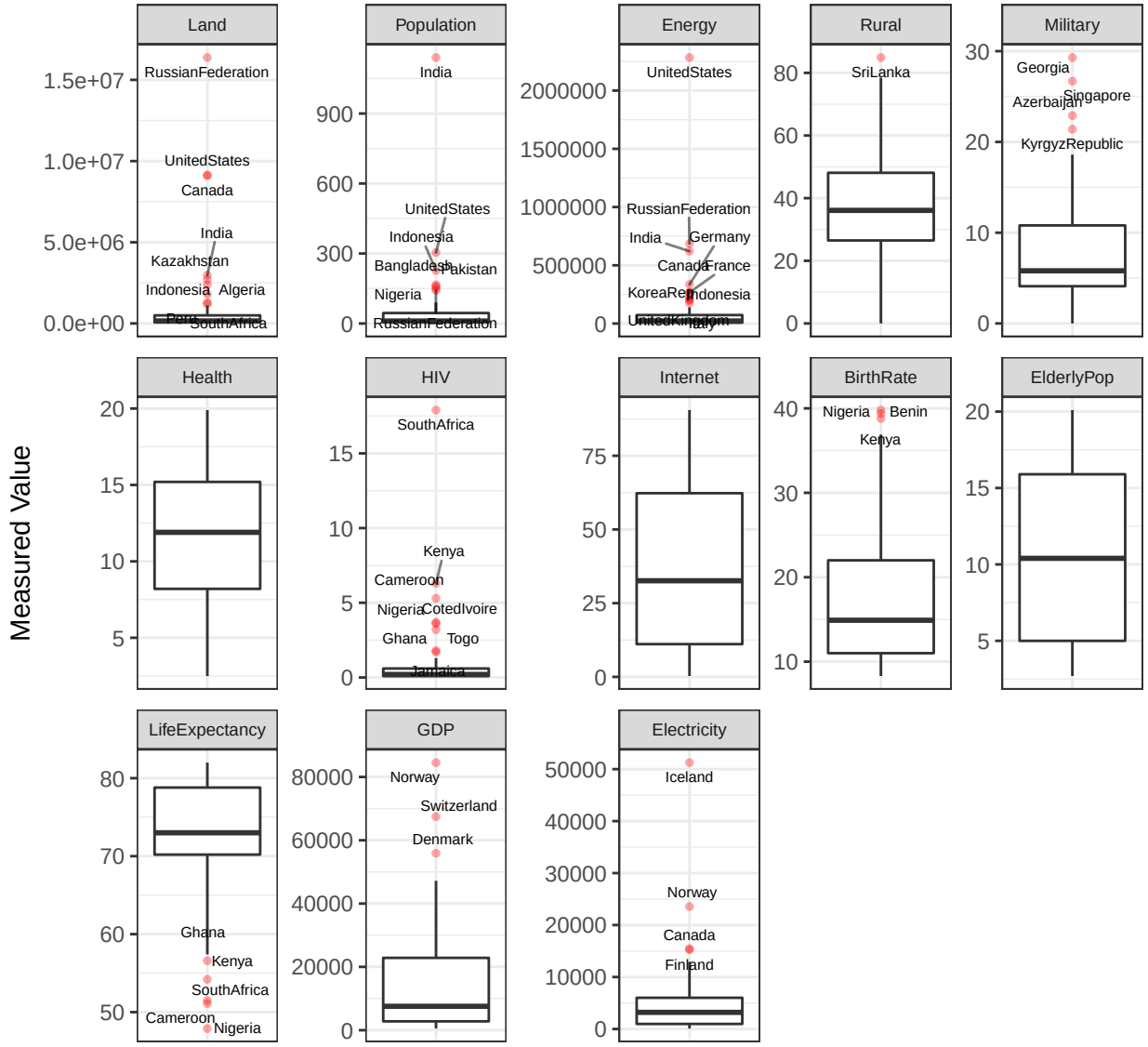
Cody Frisby

5/02/2017

77 Countries

Part A

A first look at the data, using the univariate box plot, we can see that quite a few of the measured characteristics contain countries that could be considered outliers, or very unusual.



As can be seen from the above plot, there are many countries among the measured characteristics that are unusual. Beginning with the top left box plot, *Land*, **Russia** is far beyond the other countries with **Canada** and the **United States** also beyond the bulk of the other countries, among others.

India is highly unusual for the variable *Population* with other outliers **United States** and **Indonesia**, among others.

For the variable *Energy* the **United States** is far beyond the rest of the countries with **Russia** and **India**, among others, being unusual as well.

Sri Lanka appears to be the most unusual country in terms of the percent of the population living in rural areas.

Georgia, **Singapore**, **Azerbajan**, and the **Kyrgyz Republic** are all unusual for *Military* spending per GDP.

The percent of the population with *HIV*, African countries appear to be the only one's that are highly unusual with **South Africa** far away from the rest.

African countries are those that are unusual with *BirthRate* and *LifeExpectancy* variables: **Nigeria**, **Benin**, and **Kenya** (*BirthRate*) and **Nigeria**, **Cameroon**, and **South Africa** being the furthest away from the bulk of the countries in terms of *LifeExpectancy*.

Norway appears to be unusual for both *GDP* (per capita) and *Electricity* (also per capita). **Iceland** appears to be way beyond the others for **Electricity** per capita.

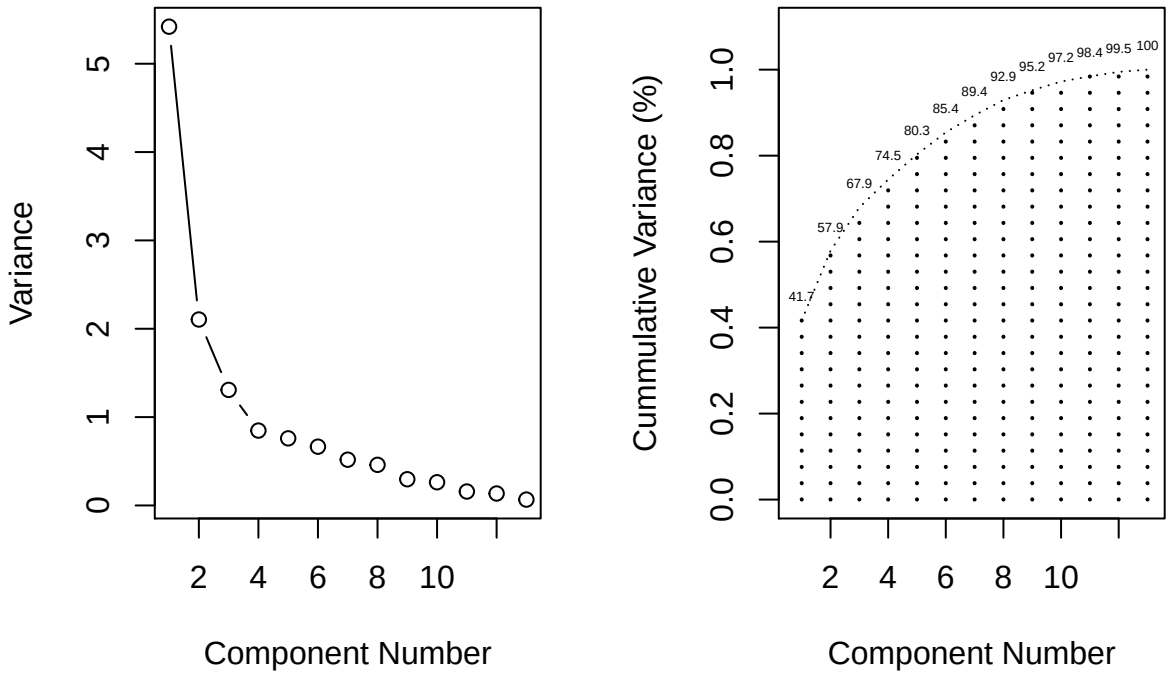
Regarding relationships among the variables that are notable the following is simply a summary. I forgo displaying the correlation matrix as it is rather large.

- *Energy* consumption and *Land* are correlated (0.65).
- *Rural* is negatively correlated with *Internet*, *LifeExpectancy*, and *GDP* with values of -0.65, -0.61, and -0.58.
- *LifeExpectancy* has a correlation coefficient of 0.6 with *Health* and a negative correlation of -0.61 with *HIV*.
- *Internet* is correlated with *BirthRate*, *ElderlyPop*, *LifeExpectancy*, *GDP*, and *Electricity* with values of -0.7, 0.74, 0.73, 0.82, and 0.63 respectively.
- *BirthRate* and *ElderlyPop* are negatively correlated, -0.85, which intuitively makes sense, as well as -0.77 with *LifeExpectancy* and *BirthRate*.

With 13 variables, it is difficult to graphically represent them all and draw conclusions. Usually the scatter plot matrix is the best in showing the bivariate relationships among all variables but with 13 that plot is difficult to interpret and show in a small space. I like the above univariate box plot and then the subsequent correlations of all the variables to begin exploring the data when there are this many variables.

Part B & C

Using the principal component analysis procedure we may be able to describe a majority of the variation in the data using indices far less in number than 13. Using the correlations of the variables, since the variances of the variables are definitely NOT different, I plot the variances of the principal components and also the cumulative sum of the PC variances to guide us in the reduction of dimensions.

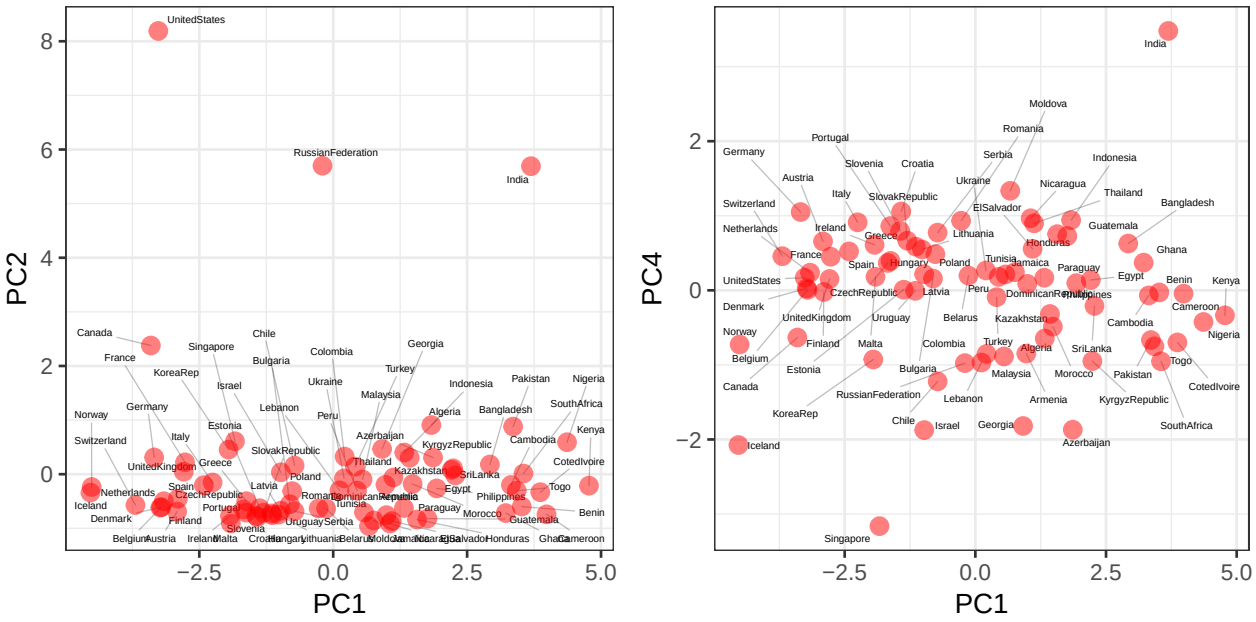


Components 1-4 explain around 75% of the variation in the data. We would gain another 10% if we include components 5 and 6. Looking at the plot on the left it begins to flatten out after the 4th component. Considering the amount of variation explained by the first 4 components AND the scree diagram showing a leveling out after the 4th one (and for the sake of brevity), I consider an interpretation of the first 4 components.

Table 1: First 4 principal components from 77 countries dataset

	PC1	PC2	PC3	PC4
Land	-0.0284141	0.5549205	0.1255714	-0.0846777
Population	0.0943356	0.4645047	-0.0718885	0.4659794
Energy	-0.0746941	0.6022667	0.1160382	0.1260395
Rural	0.3160566	-0.0117572	-0.0971178	0.3521915
Military	0.1082910	0.3143528	-0.4526782	-0.6176207
Health	-0.3031178	-0.0442405	0.1580708	0.2690996
HIV	0.1733835	-0.0081951	0.6601101	-0.1489378
Internet	-0.3914187	0.0110050	0.0888806	-0.0758374
BirthRate	0.3510608	-0.0071725	0.2379301	-0.1161037
ElderlyPop	-0.3573879	-0.0398960	-0.0805174	0.2048087
LifeExpectancy	-0.3786649	-0.0295520	-0.2925731	0.0025903
GDP	-0.3589288	0.0648030	0.2029433	-0.0784681
Electricity	-0.2734190	0.0758698	0.3045315	-0.3039304

Table 1 shows the coefficients for the first 4 principal components. The interpretation of them will be illustrated with accompanying plots.

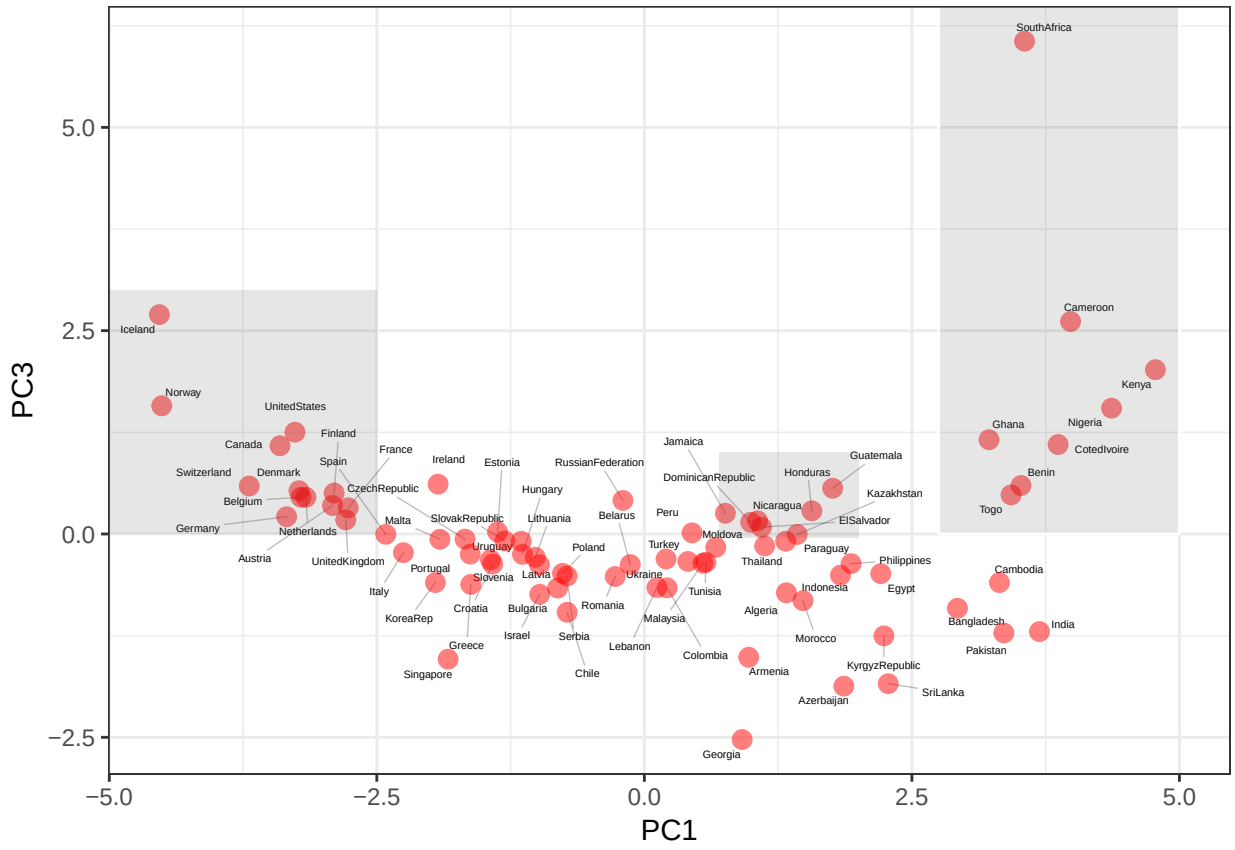


PC1 (as can be seen on the plot above) has a lot of spread from around -3 to 5 on the *PC1* axis. This is due to loadings on the the original variables (negative values) *Health*, *Internet*, *ElderlyPop*, *LifeExpectancy*, *GDP*, *Electricity*, and positive values *Rural* and *BirthRate*. This component might be thought of as the overall "quality of life" index. With extreme negative scores being a high "quality of life" and large positive scores indicating a lower overall "quality of life". We can see some grouping among countries on the African continent, Central America, and Western Europe/North America, among others (see below *PC1* by *PC3* plot).

Component 2 (*PC2*) heavily loads on *Population*, *Land*, and *Energy*. This is illustrated with a bi-variate plot of *PC1* by *PC2* (shown above left) with extreme positive *PC2* values among countries with large *Land*, *Energey*, and/or *Population*.

PC3 loads heavily on *HIV* (positive scores correspond to high occurrences of HIV) and *Military* (larger negative scores correspond to higher spending per GDP on military). There seems to be some clustering of countries that are in western Europe or Scandinavia as can be seen on the *PC1* by *PC3* scatter-plot above. We can even begin to see some clustering of the countries geographically! I've drawn some rectangles over those countries that are geographically close as well as close on the plot.

PC4 shows much spread from negative to positive scores much like *PC1* (see above *PC1* by *PC4* plot). *Population* is a dominant variable of this component with large positive values corresponding to large populations (e.g. **India**) and *Military* with large negative scores corresponding to greater spending on the military per GDP.

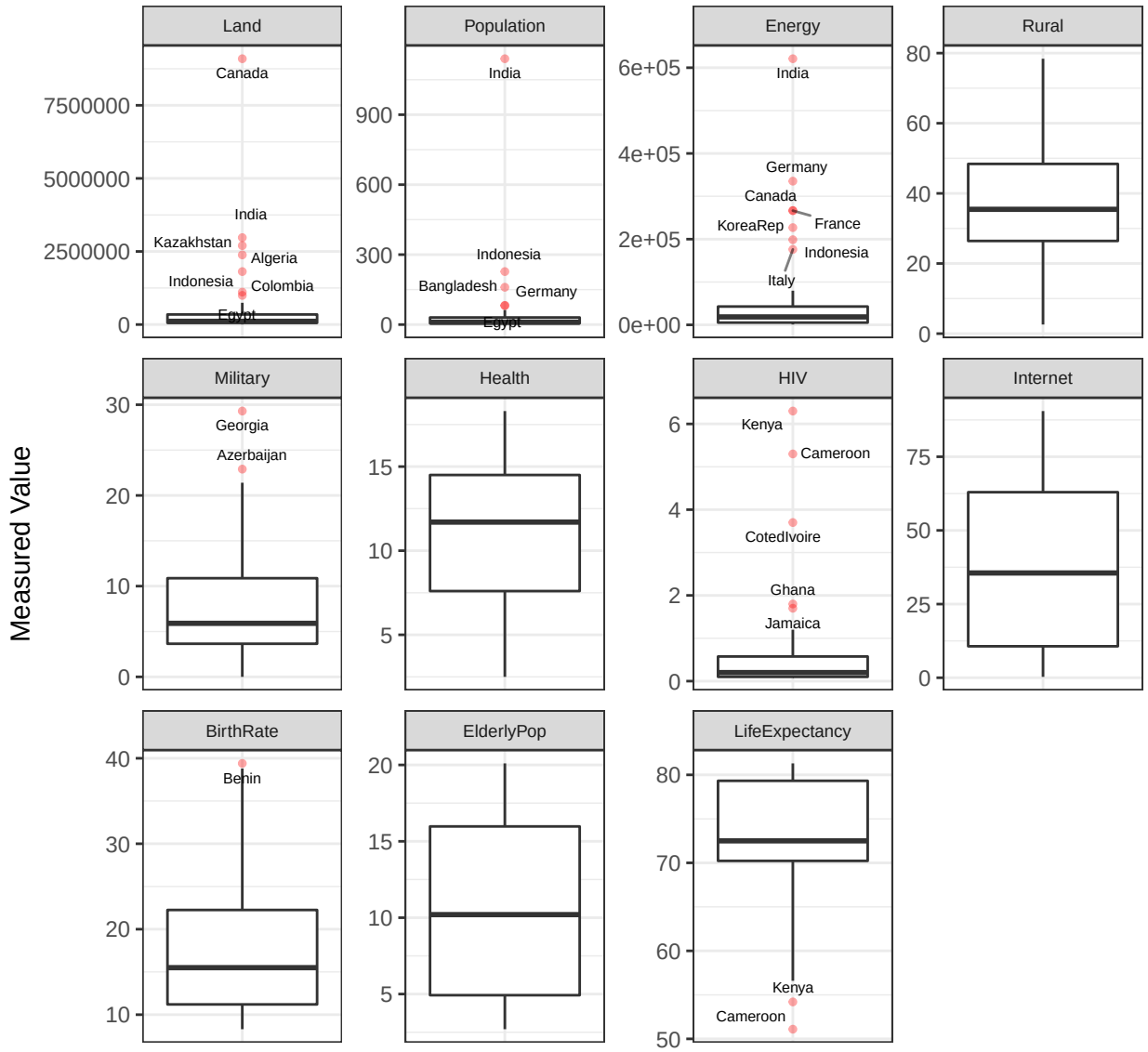


Part D

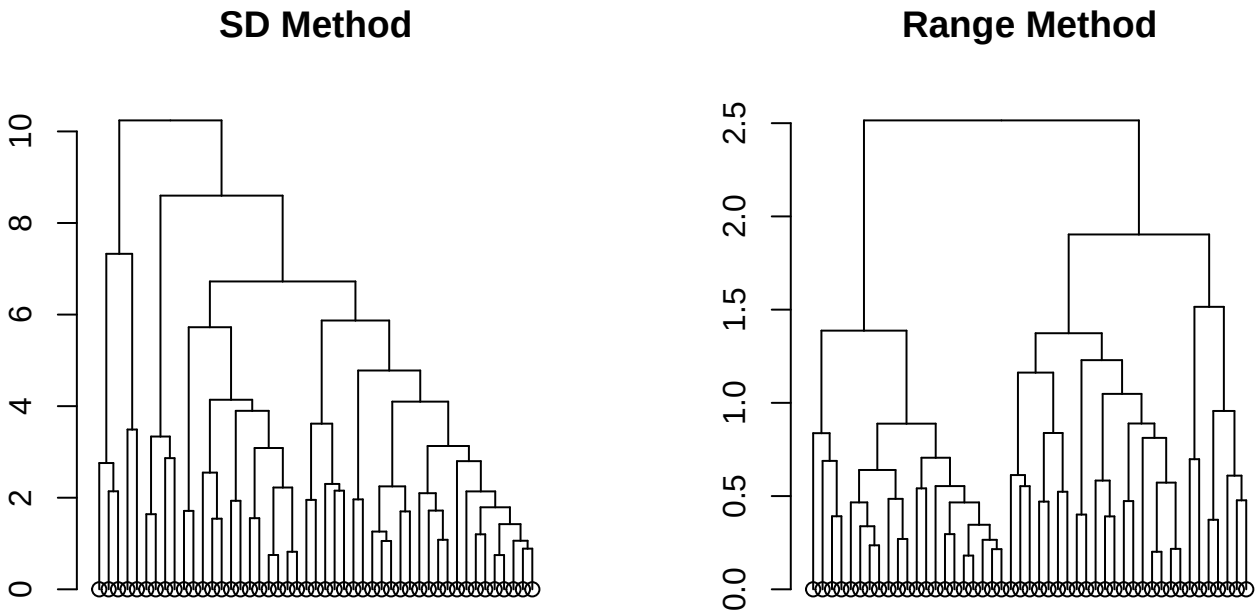
There does appear to be countries that are similar and dissimilar using the first 4 principal components. As illustrated above with the *PC1* by *PC3* plot, we appear to have some groupings of similar countries that at times even appears to correlate with their geographical locations. Those countries that have large negative scores for *PC1* would be the "best" to live in since they would tend to correspond to better overall health, longer life expectancy, higher GDP, and perhaps even less crowded. I'm surprised that the US doesn't appear to have higher scores regarding military spending.

50 Countries

It is readily apparent that the variances of the variables are not the same. Prior to any clustering, we first need to standardize the variances of the variables. I first chose divide the elements of each variable by their respective standard deviations. I then tried using the `diff` of each variables' range. This standardization technique produces a dendogram that has a much clearer structure. Additionally, **Canada**, **India**, and **Indonesia** are outliers (among others) and will heavily influence any clustering methods so they will be removed (see box-plot below).

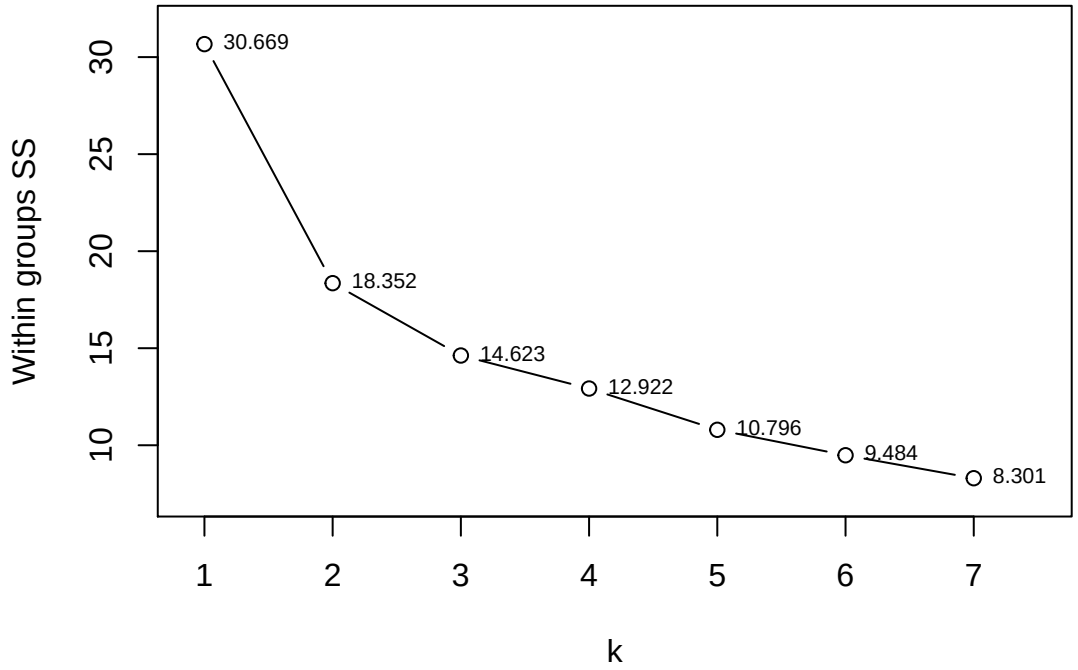


After removing the above mentioned 3 countries I use **complete linkage** and draw the dendogram below using both standardization methods mentioned above. (The labels are not shown due to over-crowding).

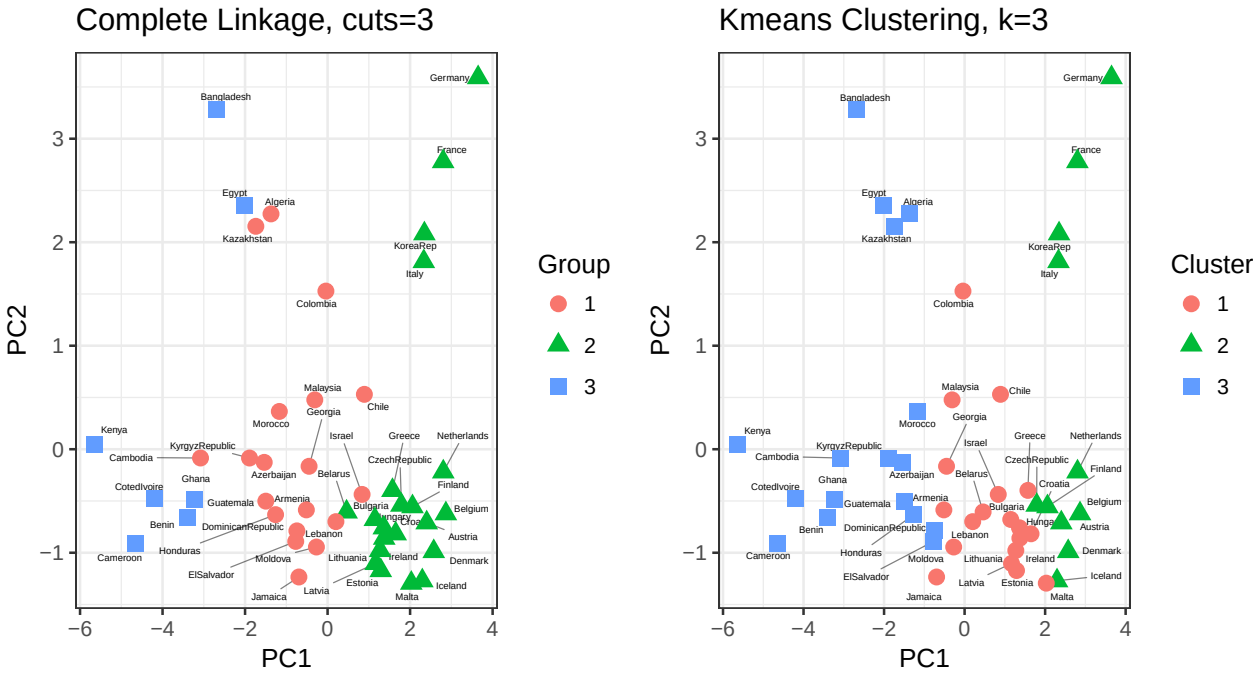


The dendogram on the left doesn't strongly suggest an ideal number of groups to cut the data into but the one on the right suggests the presence of two distinct groups. I use the range method hereafter.

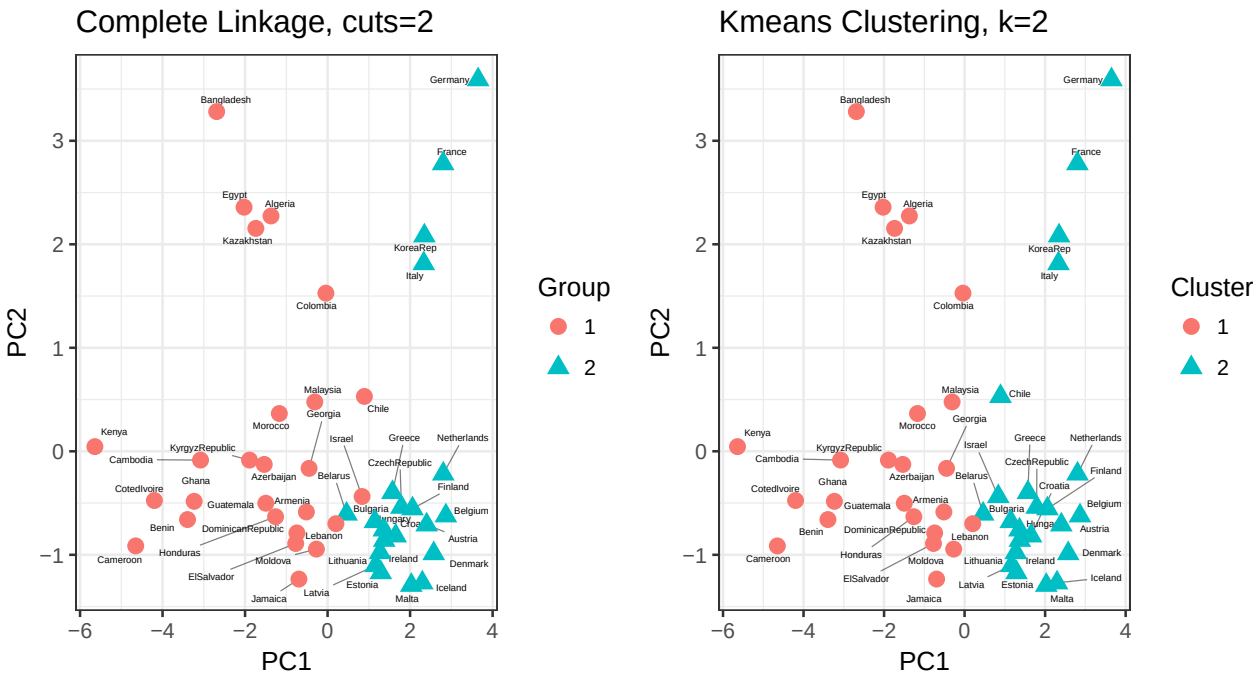
The below plot, k vs. within groups sum of squares shows the within group sum of squares for 1 to 7 values for k to try to get an indication for the value for k.



It looks like $k = 3$ is adequate. When $k = 3$ the $\frac{between_{SS}}{total_{SS}} = 47\%$. Below I plot the two methods in a two-dimensional PCA space cutting the members into 3 groups.



For $k = 3$ the two methods are fairly similar when visualizing the clusters in a two-dimensional space with k-means doing, perhaps, a better job. When $k = 2$ the two methods' clusters are nearly identical, with disagreement on **Israel** and **Chile** but the $\frac{between_{SS}}{total_{SS}} = 40.2\%$. Partitioning, for the most part, appears to be on the line $PC1 = 0$.



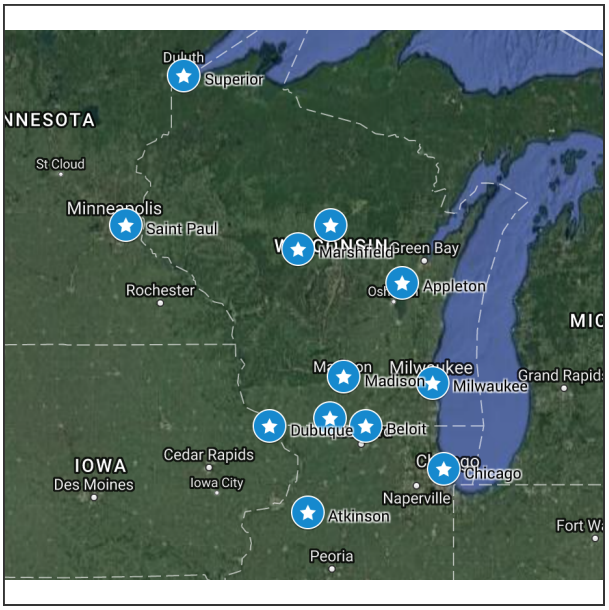
Road Distances

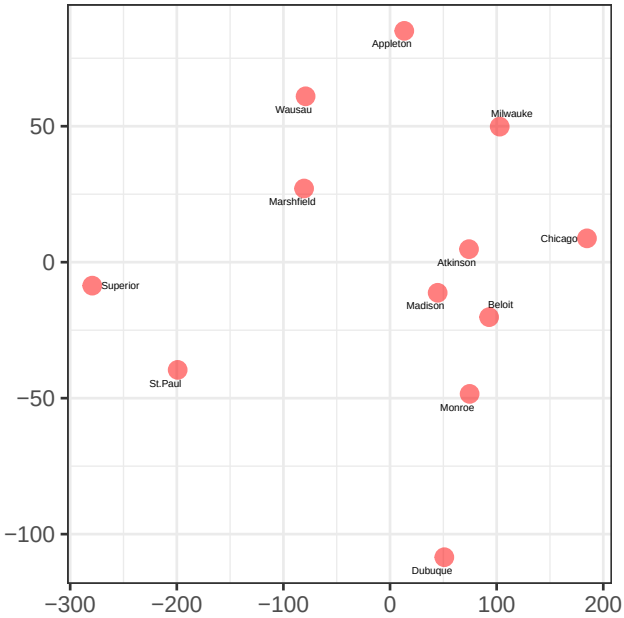
The 1, 2, and 3 dimension coordinates are located in table 2. Dimension 1 is comparable for all three. Dimension 2 is similar for the 2 and 3 dimension solutions.

Table 2: MDS solution for k = 1, 2, 3 dimensions

	d1	d21	d22	d31	d32	d33
Appleton	-13.37076	-13.37076	85.067148	-13.37076	85.067148	3.6080223
Beloit	-92.94157	-92.94157	-20.205916	-92.94157	-20.205916	-6.4943345
Atkinson	-74.07473	-74.07473	4.804039	-74.07473	4.804039	-0.7958586
Madison	-44.68148	-44.68148	-11.252521	-44.68148	-11.252521	-2.1590158
Marshfield	80.61250	80.61250	27.097883	80.61250	27.097883	6.0497082
Milwauke	-102.87582	-102.87582	49.849553	-102.87582	49.849553	10.8397572
Monroe	-74.66603	-74.66603	-48.422639	-74.66603	-48.422639	-9.8816566
Superior	279.27573	279.27573	-8.621892	279.27573	-8.621892	-62.7300520
Wausau	79.19504	79.19504	60.997371	79.19504	60.997371	1.3731325
Dubuque	-50.92029	-50.92029	-108.488036	-50.92029	-108.488036	-6.9417434
St.Paul	199.16640	199.16640	-39.595481	199.16640	-39.595481	74.5956825
Chicago	-184.71900	-184.71900	8.770492	-184.71900	8.770492	-7.4636415

Using Google maps I located all the cities and dropped a marker on them. The two-dimensional solution is shown to the right of it. Additionally, to get Chicago on the right-hand side of the plot I multiply dimension 1, of the two-dimension solution, by -1.





The two-dimensional solution appears to match fairly well with the actual map. It appears to be rotated slightly when compared to the actual map, e.g. **Dubuque** is west of **Chicago** but on the plot it is below it.

Track Records

Note: *The longer distance races appear to be in minutes while the shorter races appear to be in seconds. We can convert the minutes to seconds or leave them, it will not change our solution since we will be operating on the correlation matrix.*

First I take a look at the eigen values from the correlation matrix of the variables (of which we have 6) and the cumulative variance that they explain in the data.

EigenVals	Cummulative	Proportion
5.1141876	0.8523646	0.8523646
0.5626992	0.9461478	0.0937832
0.1529570	0.9716406	0.0254928
0.0993158	0.9881933	0.0165526
0.0549247	0.9973474	0.0091541
0.0159157	1.0000000	0.0026526

We can see that the vast majority of the variance is captured in the first factor. If we use $k = 2$ we capture approximately 94.6%. This is an acceptably large amount and k is still much less than the number of original variables.

Table 4: Principal component factor analysis solution for k = 2

	factor1	factor2	communality
m100	-0.9265990	-0.2662346	0.9294667
m200	-0.9397054	-0.2730455	0.9576002
m400	-0.8985491	-0.3451597	0.9265256
m800	-0.9403466	0.1455774	0.9054445
m1500	-0.9354910	0.3292861	0.9835727
m3000	-0.8976470	0.4104961	0.9742772

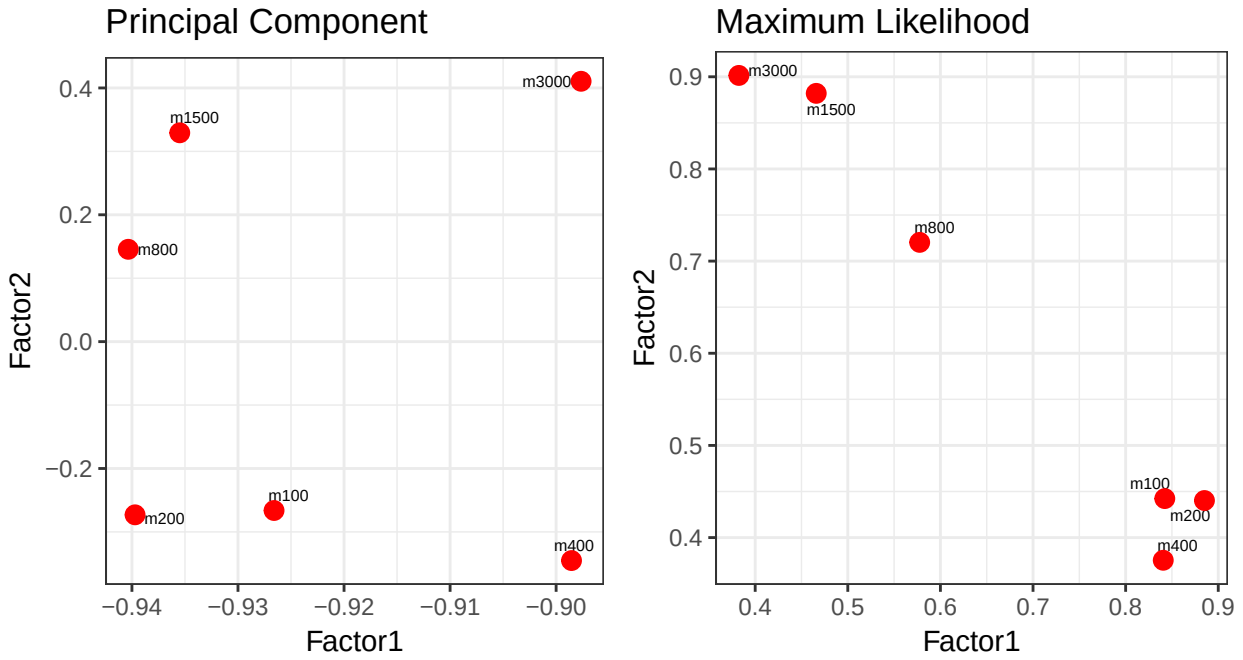
Table 4 shows the $k = 2$ principal component solution. *Factor1* has large negative values for all the loadings. *Factor2* appears to be partitioning the variables by sign based on shorter and longer distances.
Using $k = 2$ and the maximum likelihood method we get a p-value slightly bigger than 0.01, which with $\alpha = 0.01$ we would not reject the null hypothesis that 2 factors is sufficient. Considering this, and that most of the factor loadings are essentially 0 if we allow $k = 3$, I also fit the model using maximum likelihood and $k = 2$. Table 5 contains the rotated solutions.

Table 5: Maximum likelihood solution for k = 2

	factor1	factor2	communality
m100	0.8422872	0.4423248	0.0948991
m200	0.8851007	0.4402205	0.0228027
m400	0.8406144	0.3753932	0.1524492
m800	0.5776266	0.7203202	0.1474868
m1500	0.4659371	0.8820054	0.0050000
m3000	0.3824041	0.9014735	0.0411125

I tend to like the loadings produced using the maximum likelihood method as they appear a little easier to interpret. Most premier runners don't run all distance races. The sprinters tend to just run the 100-400 meter races while the medium distance runners tend to run 800 meter and above races. The two-factor solution using maximum likelihood loads heavily on the first three distances (*factor1*) and then on the next three distances (*factor2*).

As the plot below shows, both methods appear to group the 6 variables into two groups by shorter and longer distances. The maximum likelihood includes a rotation while the principal component does not.



By way of illustration, below is a plot of the two-factor solution, using maximum likelihood, with the points labeled by their respective country. Most of the countries are grouped around (0, 0) with a few outliers, **USA** with a large negative score for *factor1* (the USA usually has fast female sprinters), **COK** and **KOR_N** with larger positive *factor1* scores. **SAM** has a large *factor2* score with a *factor1* score around 0. Perhaps Samoa is exceptionally slow in the longer distance races but about average for the shorter ones.