363 Final Project Report

Kelsey Evans, Megan Ahern, Armin Thomas

May 4, 2019

Introduction

Background and Motivation The motivation for this project was to examine relationships between different indicator categories, like socio-economic, energy, and environmental indicators. We each chose several indicators that were personally interesting. For example, Megan was interested in the environment, so she made sure to include factors like forest area and CO2 emissions, Kelsey was interested in factors relating to women's rights, like Gender in the Constitution and Fertility rates, and Armin was interested in factors relating to development, such as the legal rights index. We hoped to see these different indicators relate to each other.

Design and Primary Questions We've chosen to use PCA, Factor Analysis, Cluster Analysis, and MANOVA for this project.

Some questions:

- How do the indicators relate to one another?
- How do categorical variables predict continuous variables?
- Can we find interesting insight by grouping countries together?
- Might there be latent factors that explain these relationships?

Data We scraped our data off of the World Bank site, using information from the year 2016. The cleaning process included putting the data into an excel spreadsheet which was converted to a CSV so it could be used in R. We took the log of about half of these indicators after examining normal quantile plots. We chose the following indicators:

Categorical:

- **Region**: (NAm = North America, SLA = South/Latin America, AS = Arab States, EUR = Europe, AF = Africa, AP = Asia/Pacific)
- **IncomeGrp**: (1.L = Low, 2.LM = Lower Middle, 3.UM = Upper Middle, 4.H = High)

Continuous:

- **Gender**: Mention of Gender in the Constitution, units: 1=yes; 0=no (this is technically a binary variable, but we used it as continuous)
- **logLegalRights**: Strength of Legal Rights, units: 0=weak to 12=strong (while this is a 0-12 scale, we are also confident in using it as a continuous variable)
- **logCO2**: CO2 output, units: metric tons per capita
- **LogEnergy**: Energy Consumption, units: kg of oil equivalent per capita
- **CleanCook**: Access to Clean Cooking Fuel, units: % of population
- **logFertility**: Fertility Rate, units: births per woman
- **logPopBelow**: Percent of Population below 5 meters of elevation, units: % of population
- **ForestArea**: Forest Area, units: % of land area
- **logRefugee**: Refugee Population by country of asylum, units: # of persons

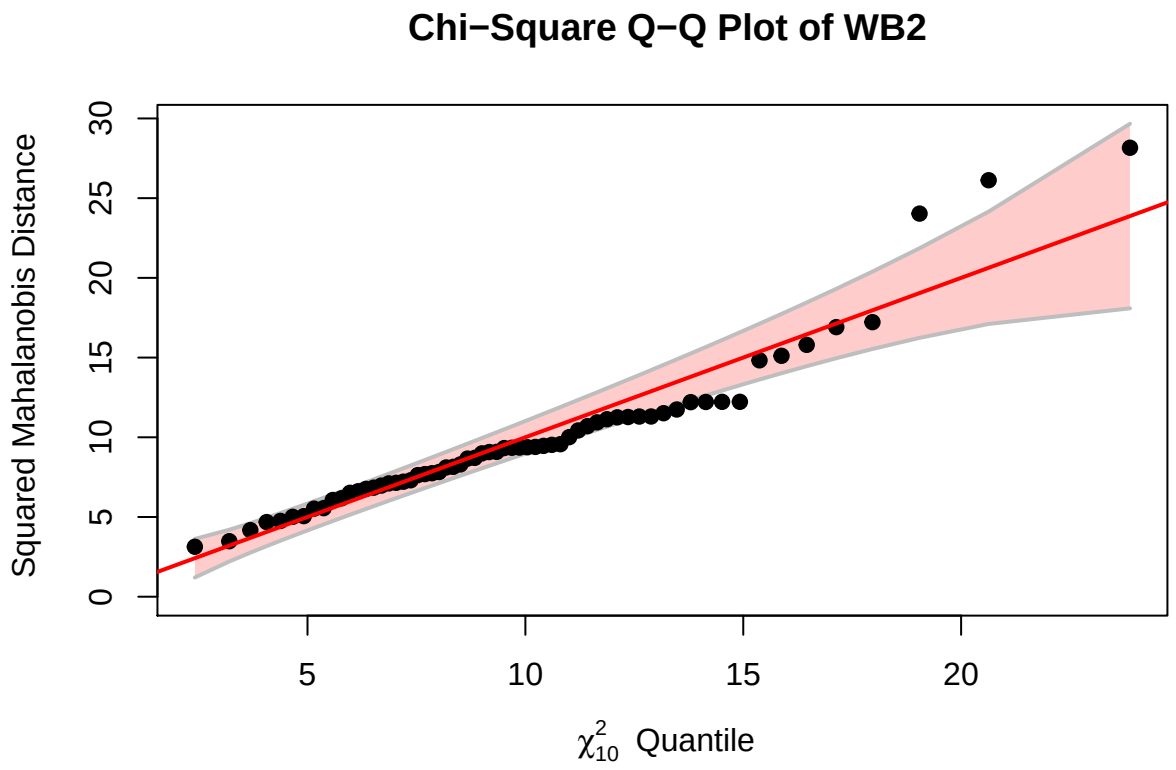
- **logOutput**: Renewable Energy Output, units: % of total electricity output

Here is an example of the data:

```
##            Region IncomeGrp Gender CleanCook ForestArea logRefugee
## Albania       EUR      3.UM      1     77.42 28.1218970   4.927254
## Algeria        AS      3.UM      1     92.62  0.8244393  11.453515
## Angola         AF      3.UM      1     48.05 46.3072104   9.652137
## Bahrain        AS       4.H      1    100.00  0.7840617   5.602119
## Bangladesh     AP       1.L      1     17.72 10.9579782  12.528906
## Belarus       EUR      3.UM      0     98.18 42.6301065   7.408531
##                logCO2 logFertility logEnergy  logOutput logPopBelow
## Albania     0.6824721    0.5382462  6.695126  4.6051702   1.9559964
## Algeria     1.3130272    1.0210110  7.186220 -1.1310824  -0.2777310
## Angola      0.2556714    1.7394130  6.300860  3.9735870   0.3619272
## Bahrain     3.1548600    0.7095125  9.268043  0.0000000   3.5346103
## Bangladesh -0.7783958    0.7438403  5.403672  0.2044824   2.1878635
## Belarus     1.9023997    0.5498540  7.982253 -0.2037334   0.0000000
##            logLegalRights
## Albania         1.7917595
## Algeria         0.6931472
## Angola          0.0000000
## Bahrain         0.0000000
## Bangladesh      1.6094379
## Belarus         0.6931472
```

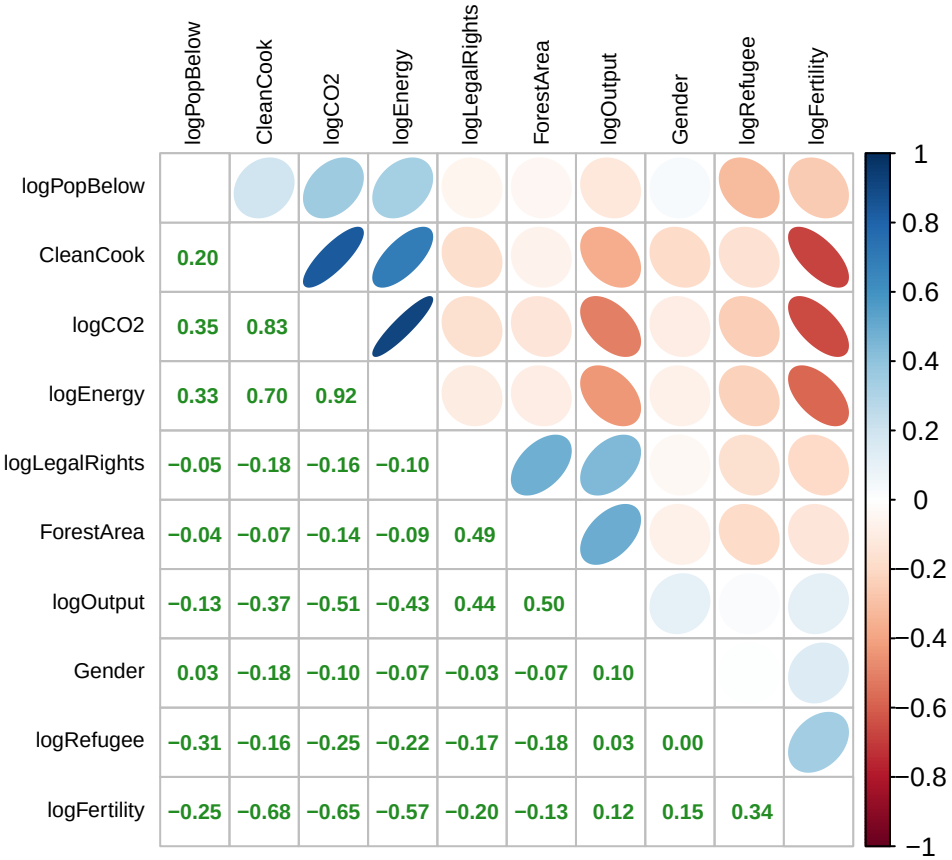
Principal Components Analysis

First, we will do principal components analysis to try to explain some of the variability in our data. In order to do this, our data would optimally be multivariate normal. We can examine this by creating a chi-square quantile plot:



The chi-square quantile plot looks approximately normal. We have only two points outside of the 95% confidence bands, so we feel confident that we can perform PCA on the data. Additionally, the multivariate normality will be useful for the other methods.

We also would like to look at a correlation matrix just to get an idea of which variables are strongly correlated:



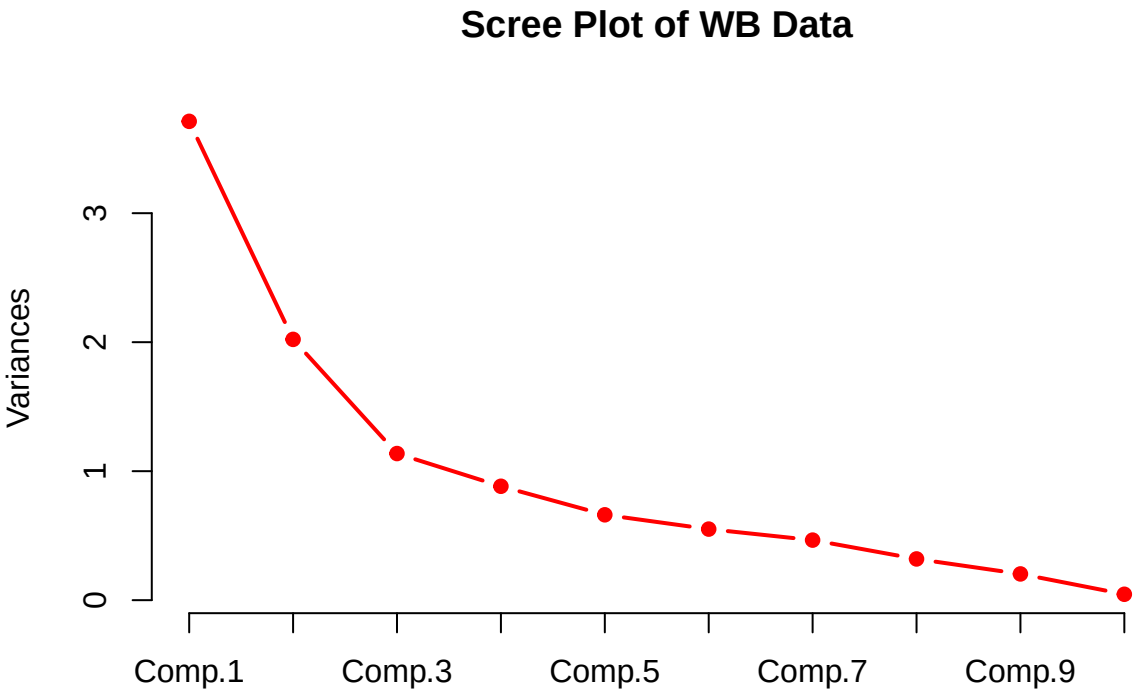
Now that we've examined the data, we can continue with Principal Components Analysis. In order to figure out how many components we'd like to keep, we will use three methods: the eigenvalue method, a scree plot, and parallel analysis.

The eigenvalues are as follows:

```
##  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
##    3.71    2.02    1.14    0.88    0.66    0.55    0.47    0.32    0.20
## Comp.10
##    0.05
```

The first three components are all greater than 1, so this test indicates that we should keep three components.

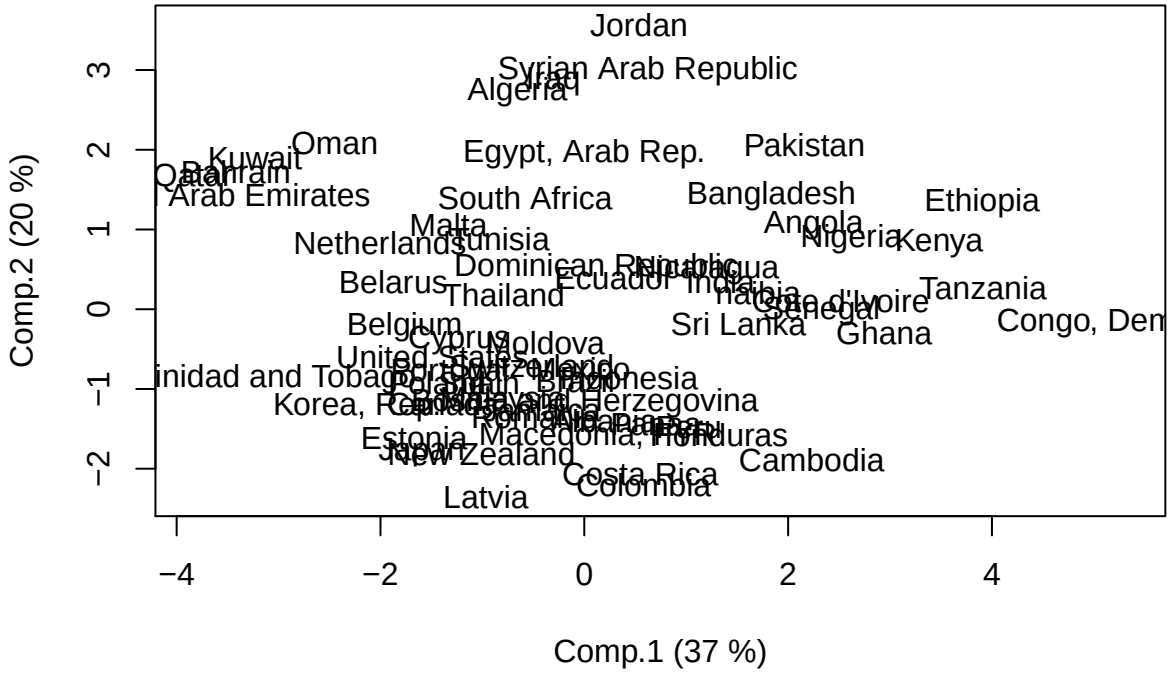
Next we will look at the scree plot:



This plot has an "elbow" at 4, which also indicates that we should keep three components.

(We have also recognized that these three components only explain around 70% of the data, while 80% would be ideal, though it would require five components. However, all of the other tests indicated that three components was best, so we would only expand to 5 if 80% of variability was explicitly required.)

To check the validity of our tests, we made a score plot to see if there were any unusual trends in the data:



There are no other observable trends, so we are confident in our tests.

We looked at the loadings and the individual components to see if there was any story told by the data. We looked at components with an absolute value of greater than 0.4, and only the first component seems to tell a

consistent story. The variable exceeding 0.4 are CleanCook, logCO2, and logEnergy, which are all energy use factors. The other components, while significant, don't tell a discernable story.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     1.9266143 1.4218750 1.0661516 0.93960587 0.8135158
## Proportion of Variance 0.3711843 0.2021729 0.1136679 0.08828592 0.0661808
## Cumulative Proportion  0.3711843 0.5733571 0.6870251 0.77531098 0.8414918
##                           Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.7422594 0.68230705 0.56530123 0.45091710
## Proportion of Variance 0.0550949 0.04655429 0.03195655 0.02033262
## Cumulative Proportion  0.8965867 0.94314096 0.97509751 0.99543013
##                           Comp.10
## Standard deviation     0.213772498
## Proportion of Variance 0.004569868
## Cumulative Proportion  1.000000000
##
## Loadings:
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## Gender           0.09   0.07   0.66   0.72   0.12   0.02   0.03   0.15
## CleanCook       -0.45  -0.03  -0.22   0.19  -0.04   0.24  -0.23   0.12
## ForestArea       0.10  -0.56  -0.12   0.07  -0.13   0.51   0.54   0.31
## logRefugee       0.17   0.33  -0.42   0.41  -0.67  -0.19   0.07   0.17
## logCO2          -0.50  -0.01  -0.03   0.15  -0.04  -0.01   0.16  -0.21
## logFertility     0.38   0.32   0.11  -0.14  -0.05   0.10   0.51  -0.43
## logEnergy       -0.46  -0.03   0.00   0.18  -0.08  -0.07   0.34  -0.52
## logOutput        0.29  -0.42   0.02   0.17  -0.29   0.25  -0.48  -0.57
## logPopBelow     -0.23  -0.09   0.56  -0.40  -0.65  -0.10  -0.02   0.15
## logLegalRights   0.11  -0.55  -0.08   0.10   0.06  -0.75   0.15  -0.01
##                 Comp.9 Comp.10
## Gender           0.04   0.02
## CleanCook        0.72   0.28
## ForestArea      -0.07  -0.01
## logRefugee      -0.08  -0.02
## logCO2           0.04  -0.81
## logFertility     0.53  -0.03
## logEnergy       -0.32   0.50
## logOutput        0.00  -0.09
## logPopBelow      0.09   0.04
## logLegalRights   0.29   0.01
```

By using PCA, we saw a component of environmental/energy related variables as explaining a lot of the variance in the world bank data. This is what we set out to see for this project - whether certain groups of variables could help explain trends in the data. However, for PCA there isn't always a coherent story - it's more about explaining the most variance. It is nice when there is also a story, but this isn't always the case.

Factor Analysis

We chose to do factor analysis next because it contains aspects of PCA, at least in choosing how many factors to look for. We referred to the correlation matrix from the previous section to look at which variables were most strongly correlated.

We observed that the largest correlations were positive correlations between log Energy and log CO2, log CO2 and clean cooking fuel, and log Energy and clean cooking fuel. The best negative correlation was between fertility and clean cooking fuel. We saw poor correlations between gender and all other indicators. Most interestingly, Gender and log Legal Rights have no notable correlation.

Observing our correlation matrices, we would say that factor analysis may be appropriate because we observed some correlations between indicators that may share a latent factor. For instance, a latent factor could simply be Energy or Development.

First we computed the KMO measure to see if factor analysis was appropriate:

```
## [1] 0.7204946
```

The KMO measure is 0.72, so factor analysis is a satisfactory option.

Next, we referred to the previous section and chose to use 3 factors by using the PCA method of choosing factors.

We chose to do the Principal Axis Factoring (PAF) method of factor analysis, both with and without the varimax rotation. For each, we computed the RMSR and the percent of residuals greater than 0.05 to determine which method was better.

For PAF:

RMSR:

```
## [1] 0.1067124
```

Percent of residuals:

```
## [1] 0.6222222
```

For PAF with varimax:

RMSR:

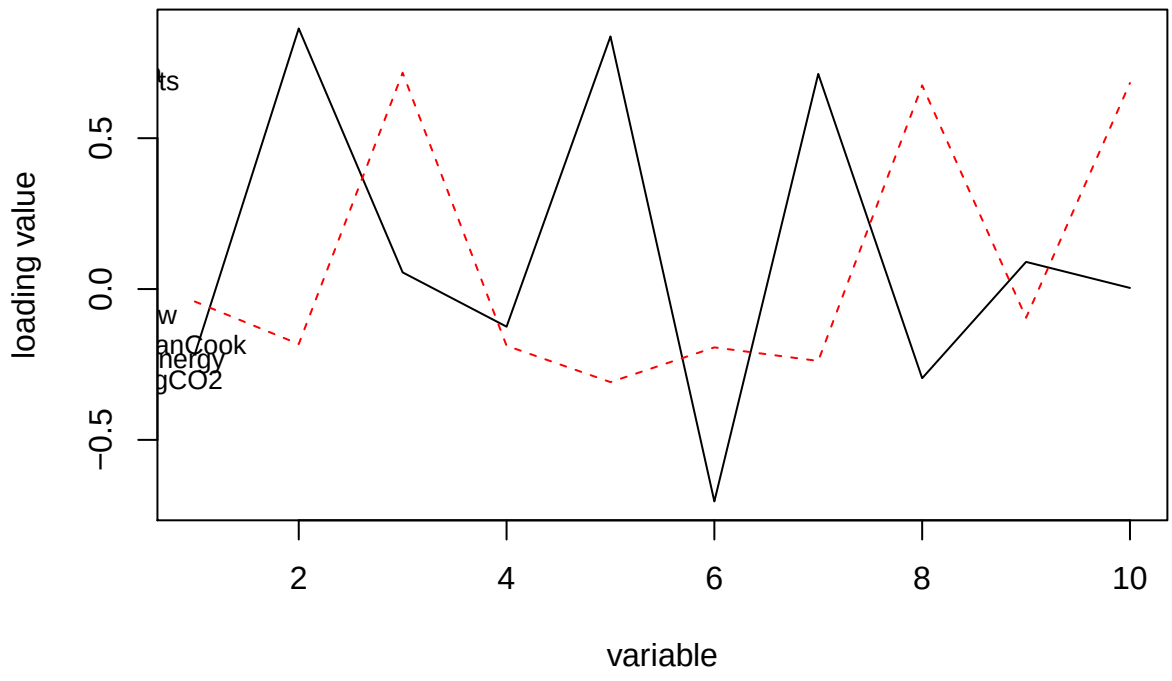
```
## [1] 0.03103372
```

Percent of residuals:

```
## [1] 0.1333333
```

Clearly, the results are better with the varimax rotation. We will use this to create a loading plot:

```
plot(fact2$loadings)
text(fact2$loadings, labels=names(WB2),cex=0.8)
```

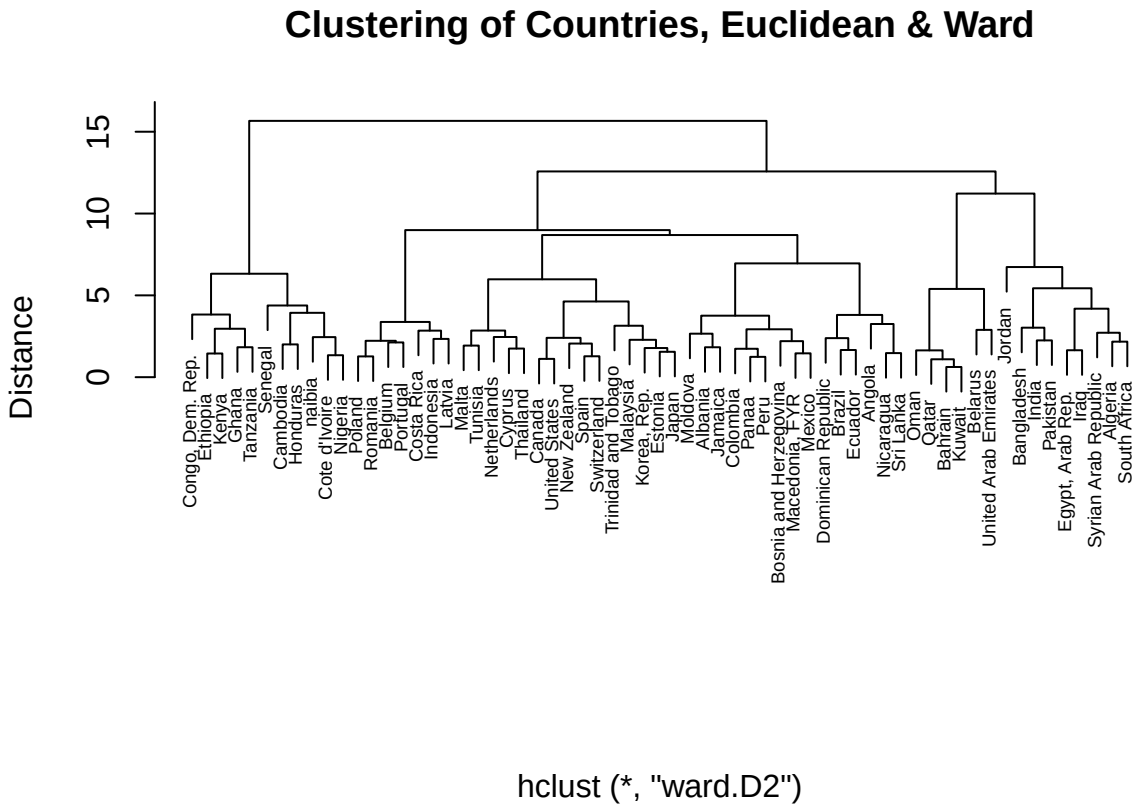


Though the plot seems to contain three groups of variables, it's difficult to ascertain exactly what the factors would be. It seems that two of the groups of variables load more heavily on the first pricipal axis, and two as well on the second. Interestingly, similarly to PCA, it seems that the energy-related variables remain grouped together to form a factor. This could be because we used the PCA method to choose the number of factors, or because energy-related variables are generally very strong predictors in general.

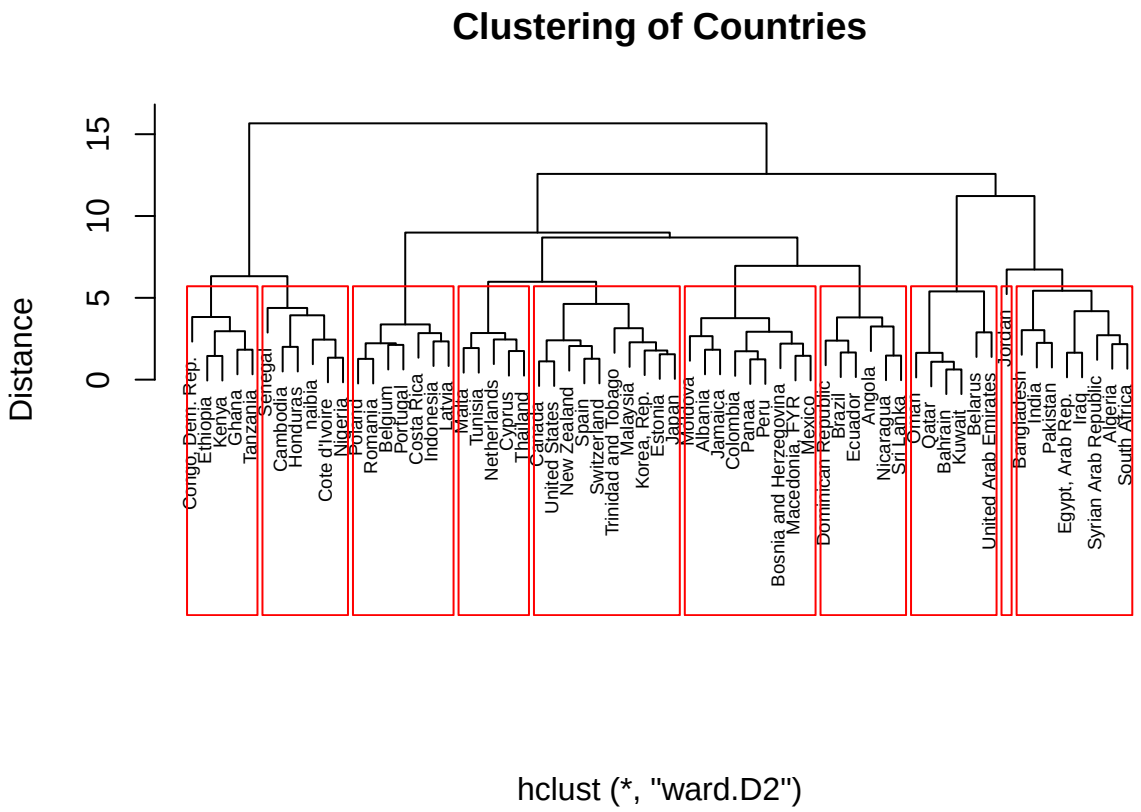
Cluster Analysis

For cluster analysis, we first created a data set that was normally scaled. This is important for cluster analysis because variables should be on the same scale in order to see how close they are to each other.

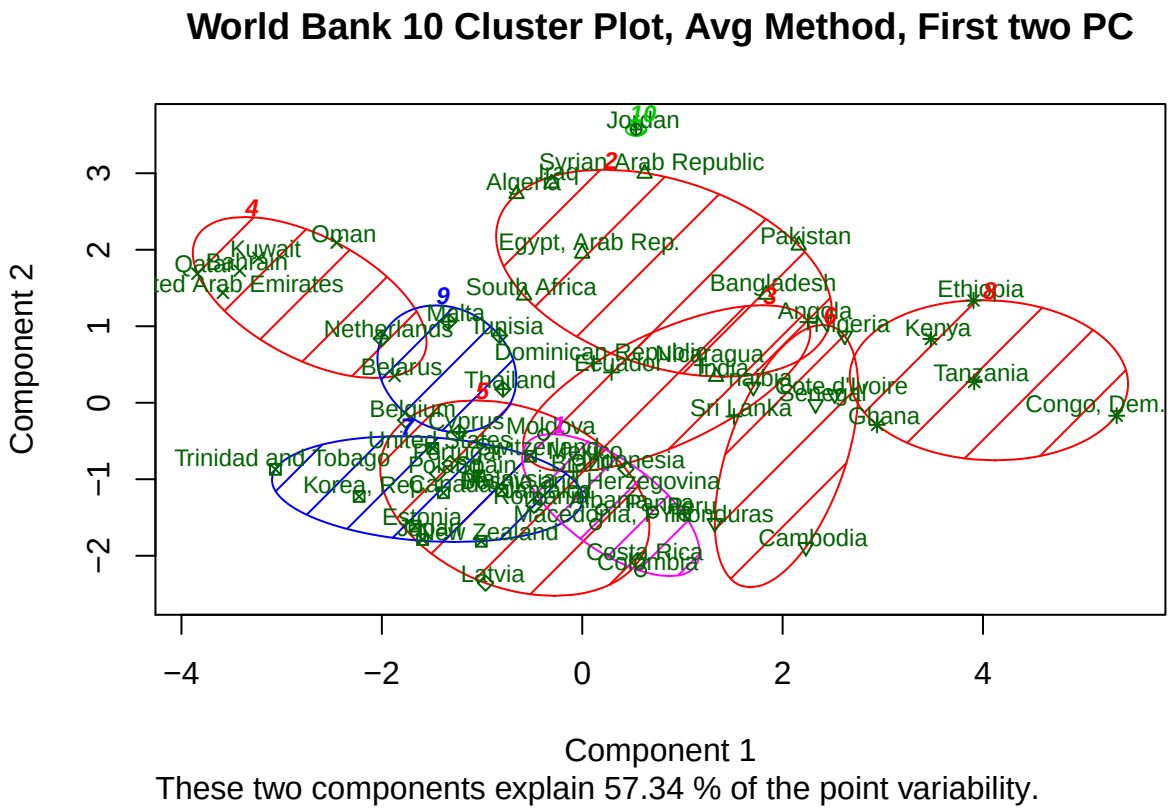
We tried several methods of clustering and found that the dendrogram we liked best was Euclidean Distance with Ward Method Linking:



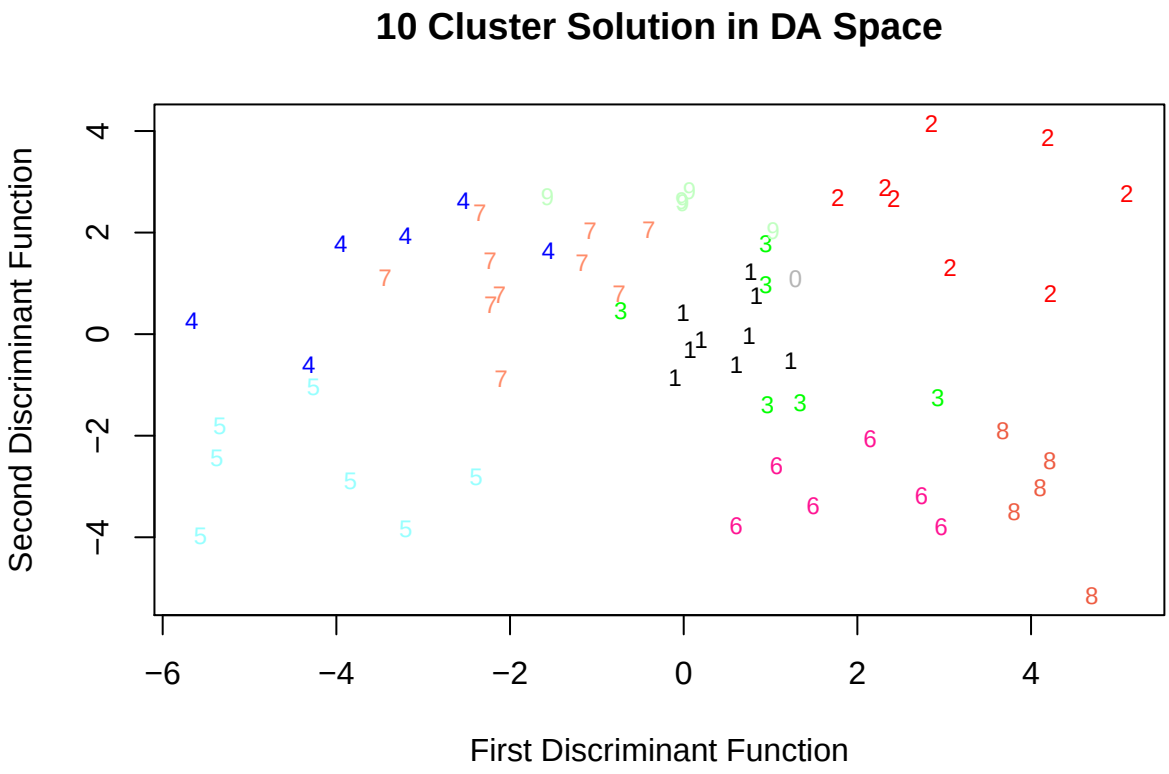
After trying various values, we settled on 10 clusters:



Next, we looked at Principal Components and Discriminant Analysis graphs of the clusters.

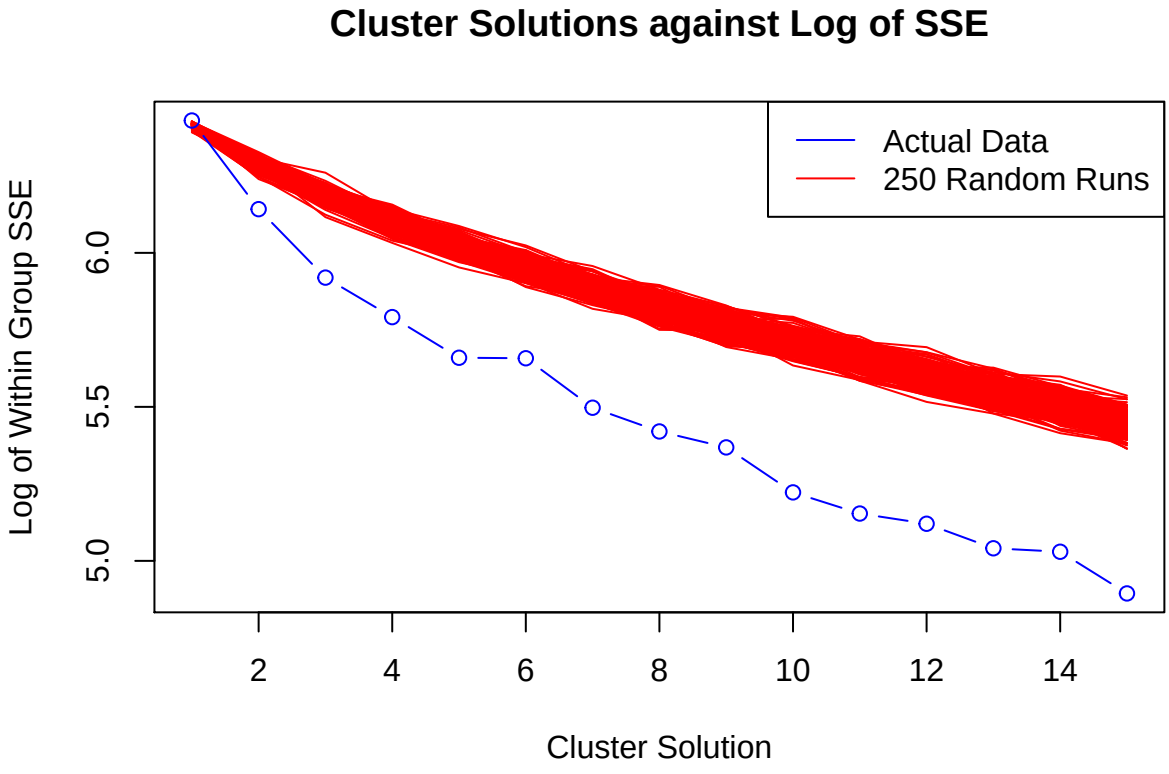


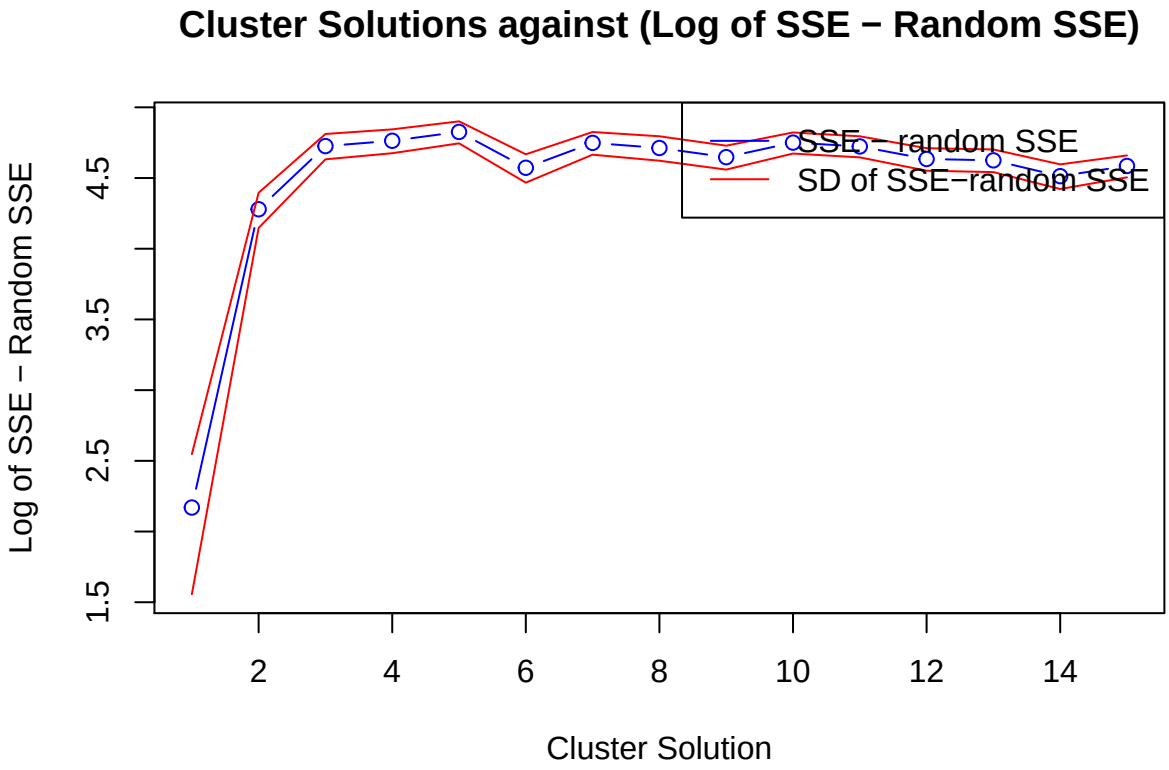
These two components explain 57.34 % of the point variability.



In both graphs, the ten clusters are pretty easily discernible.

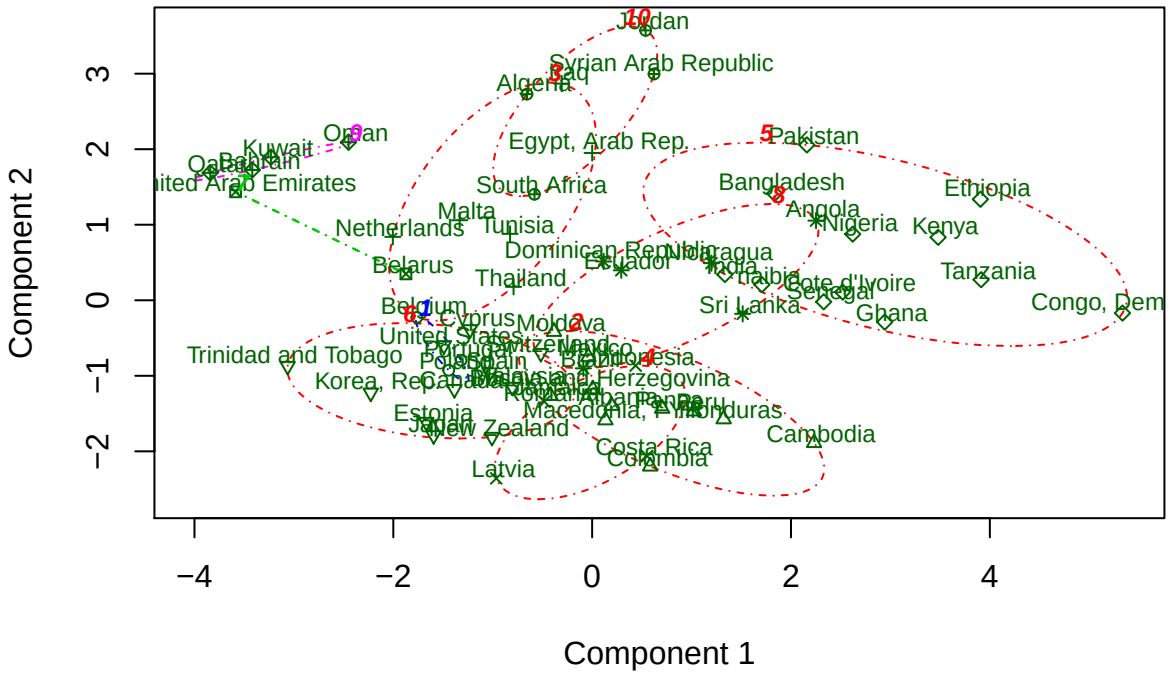
After this, we applied k-means to the data. First, we created the Cluster Solutions Against Log of SSE and Cluster Solutions against (Log of SSE - Random SSE) graphs to check on the number of clusters we were using.





Interestingly, these graphs seem to suggest using five clusters instead of ten. Howver, in looking at the original hclus_eval graph, an argument could be made for five or ten clusters. The same can be said for just looking at the graph and estimating the number of clusters - we could either choose five bigger clusters of countries or ten smaller ones.

Finally, we looked at the PCA graph with the new k-means data. Just like before, the 10 clusters are visible.



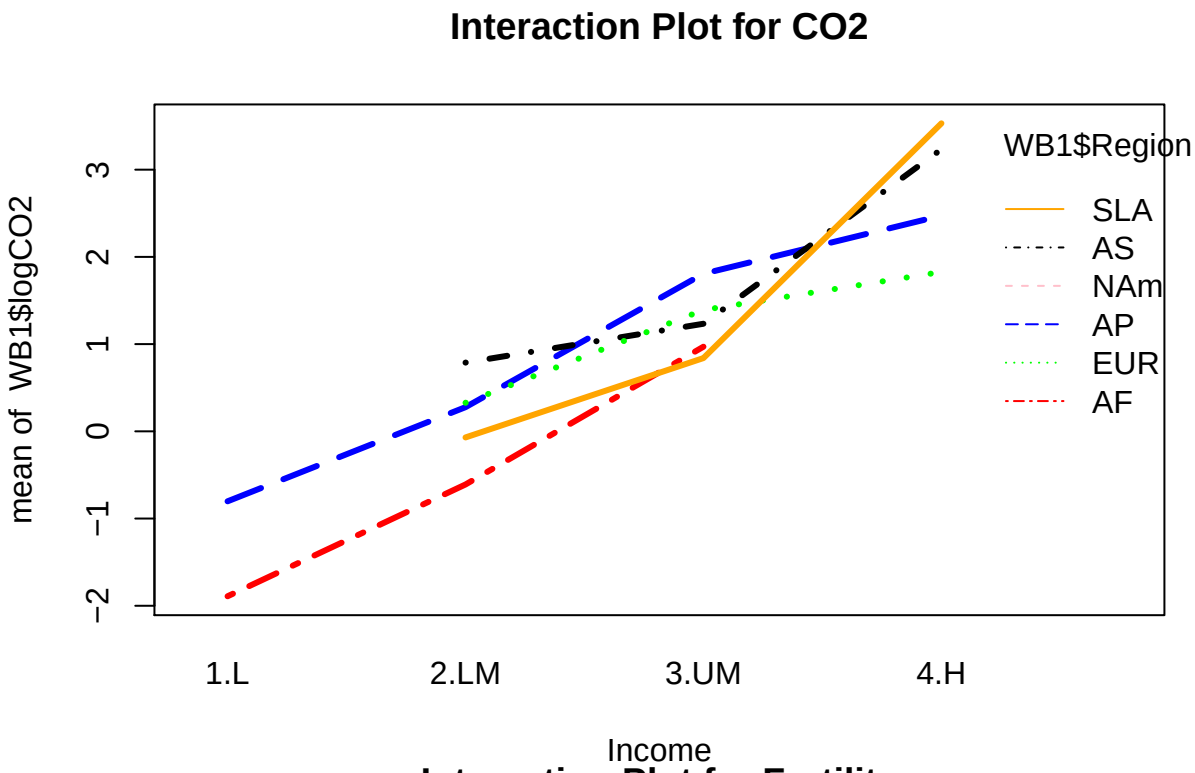
These two components explain 57.34 % of the point variability.

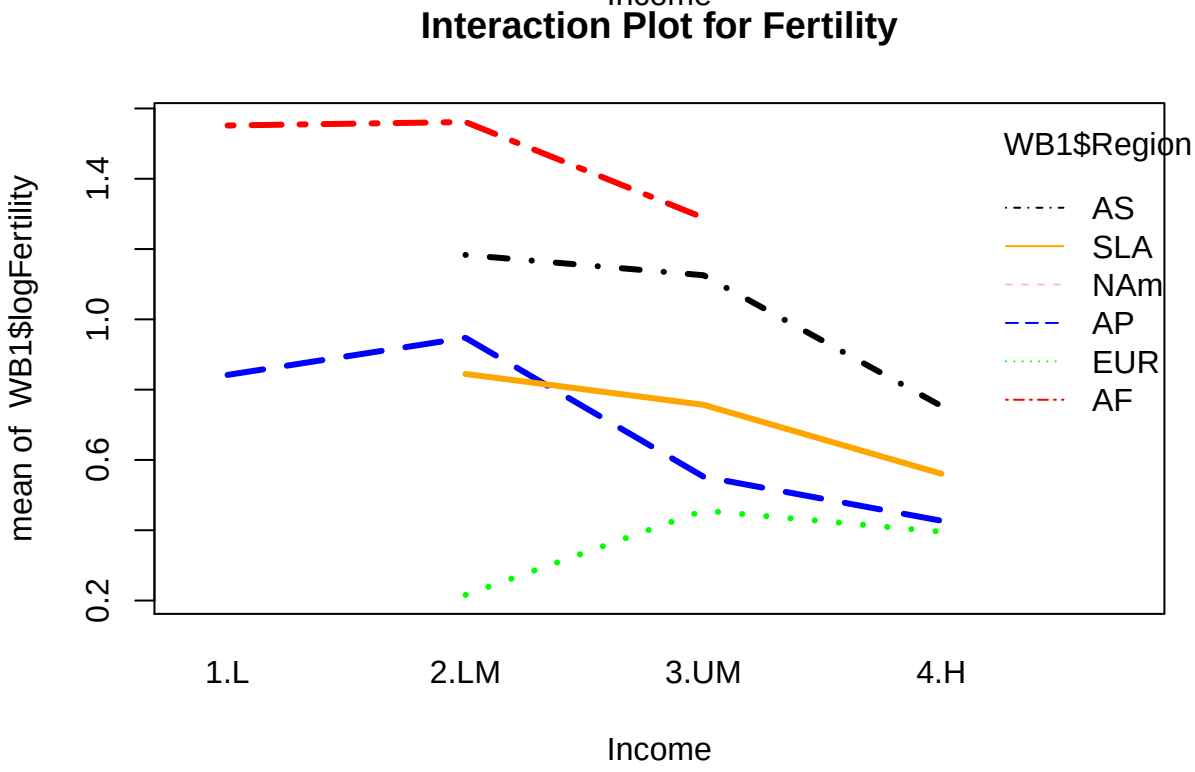
It's difficult to compare the results of cluster analysis to the other methods, as we clustered on countries instead of on variables. A topic for future analysis could be looking at the ten clusters of countries to see what each has in common. For instance, the United States, Canada, Switzerland, and Japan are all in the same cluster, and are all wealthy and more developed countries. These countries are likely similar in terms of socioeconomic, environmental, and other indicators.

###MANOVA

For MANOVA, we will look at the categorical variables of income and region, and the continuous variables of CleanCook, logCO2, logFertility, and logEnergy.

To start off, we made a couple of interaction plots to see how the data looked by region and income:





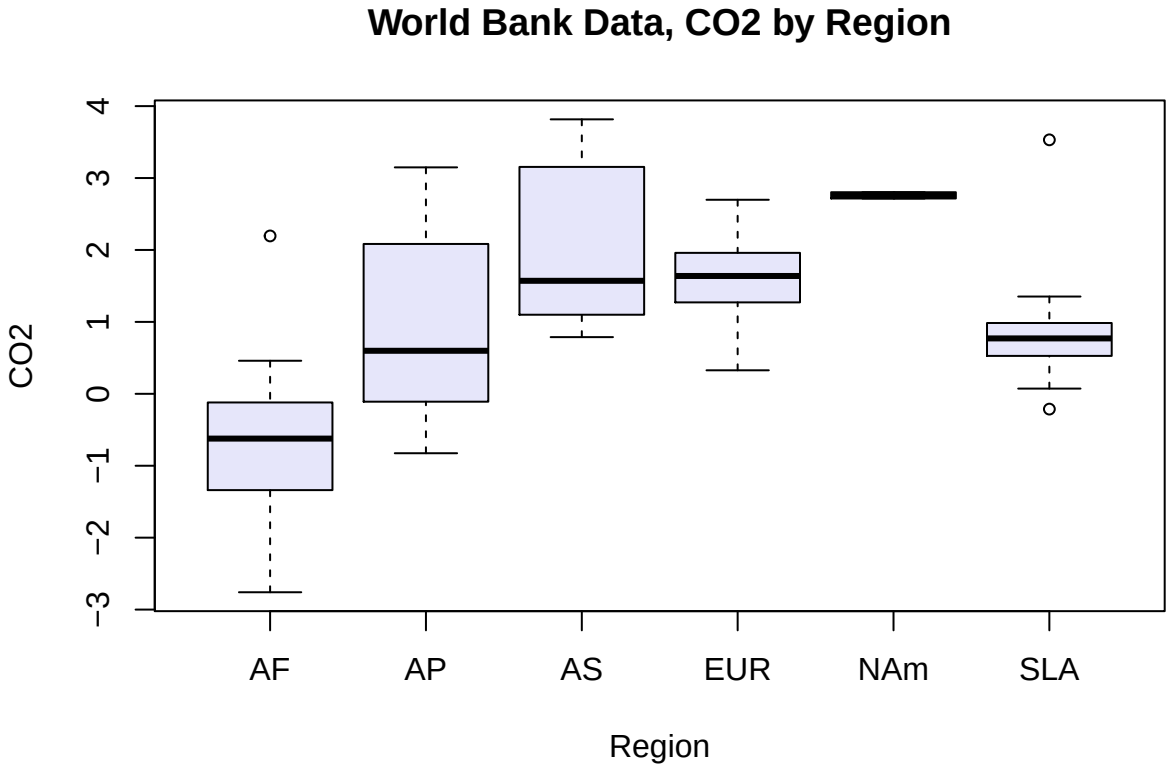
Looking at the two graphs, we see that there's a possible interaction between income and region, though we will have to run a regression to make sure. But with fertility, there's a significant difference by region. Fertility still decreases as income increases, but region seems to be more significant. As a note, North America does not appear in either interaction plot. This is because during data cleaning, the United States was the only North American country with complete data for this subset of data we chose for MANOVA.

Next, we ran several regressions and found that the best one predicted the continuous variables using income, region, and the interaction between income and region.

```
##  Response CleanCook :
##                           Df Sum Sq Mean Sq  F value    Pr(>F)
## WB5$IncomeGrp              3  48000 16000.1 120.4408 < 2.2e-16 ***
## WB1$Region                 5   7724  1544.7  11.6279 2.712e-07 ***
## WB1$Region:WB1$IncomeGrp  8   3529   441.2   3.3208  0.004453 **
## Residuals                 46   6111   132.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logCO2 :
##                           Df Sum Sq Mean Sq  F value    Pr(>F)
## WB5$IncomeGrp              3 92.111 30.7037 151.5742 < 2.2e-16 ***
## WB1$Region                 5  7.536  1.5072   7.4405 3.487e-05 ***
## WB1$Region:WB1$IncomeGrp  8  6.406  0.8008   3.9533  0.001256 **
## Residuals                 46  9.318  0.2026
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logFertility :
##                           Df Sum Sq Mean Sq F value    Pr(>F)
## WB5$IncomeGrp              3 4.6406 1.54686 44.9352 1.017e-13 ***
## WB1$Region                 5 4.3388 0.86776 25.2079 3.842e-12 ***
## WB1$Region:WB1$IncomeGrp  8 0.4762 0.05953  1.7292     0.117
## Residuals                 46 1.5835 0.03442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logEnergy :
##                           Df Sum Sq Mean Sq  F value    Pr(>F)
## WB5$IncomeGrp              3 50.191 16.7304 113.4349 < 2.2e-16 ***
## WB1$Region                 5  2.564  0.5129   3.4775  0.009498 **
## WB1$Region:WB1$IncomeGrp  8  7.092  0.8864   6.0102 2.792e-05 ***
## Residuals                 46  6.784  0.1475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As is evidenced in the significance charts, both income and region, as well as their interaction, is significant for all of the continous variables. There is one exception: The interaction term is not significant for fertility. On a cursory glance, this makes sense, as our initial interaction plot indicated that so much variability in fertility was explained by region that it might not even need the interaction.

After this, we looked at a few univariate contrasts, for which we decided to use region. Befre that, we looked at some boxplots:

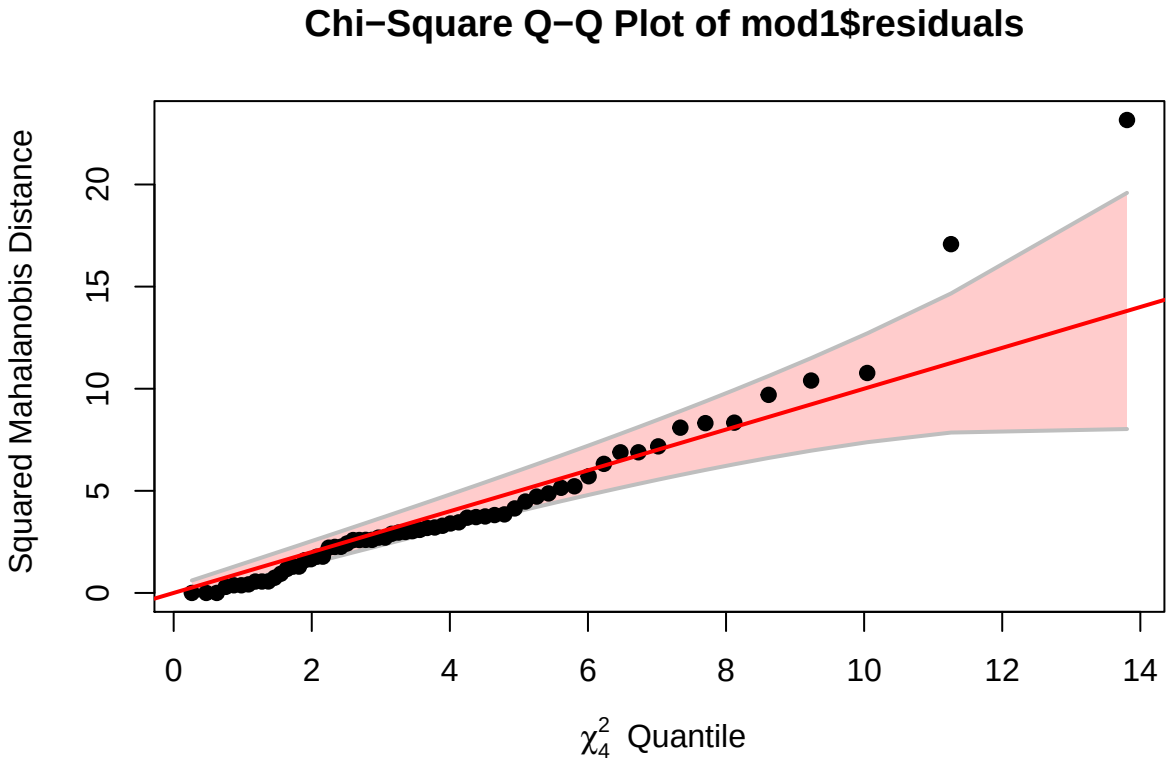


Just like earlier, North America contains only one data point and this is evident in the boxplot. To look at the contrasts, we set Asia Pacific as the constant.

```
## lm model parameter contrast
##
##    Contrast      S.E.      Lower    Upper   t df Pr(>|t|)
## 1 0.3174577 0.3513805 -0.3861697 1.021085 0.9 57   0.3701
##
## Contrast coefficients:
##   (Intercept) RegionAP RegionAS RegionEUR RegionNAm RegionSLA
## 1           0       -1      0.2       0.2       0.2       0.2
```

The contrast was not statistically significant. We think it's because the Asia Pacific mean is close to group mean, and also it has the largest spread (excluding outliers).

Finally, we looked at a chi-square residual plot to check the accuracy of our MANOVA:



The residuals appear approximately multivariate normal, which means we feel confident in our analysis.

MANOVA was the first time we introduced categorical variables. Here, it's easier to see which variables were the explanatory and response variables, as opposed to the other methods which have more to do with creating groups of variables than creating an equation to predict a response variable.

Conclusion

We looked back on the questions that we stated at the beginning of our project, and noted a few insights:

- How do the indicators relate to one another?
  - The biggest trend we noticed was the tendency of energy variables often move together.
- How do categorical variables predict continuous variables?
  - They were very significant predictors. In the example of fertility, region and income were so significant in predicting fertility rates that the interaction between them was not significant.
- Can we find interesting insight by grouping countries together?
  - Our findings from Cluster Analysis were ultimately very different than those of PCA and Factor Analysis because we based it on countries. We were able to see some interesting groupings, namely a group of highly developed countries (e.g. US, Canada, Switzerland, and Japan). The reason behind these groupings might be a topic for future research.
- Might there be latent factors that explain these relationships?
  - Certainly. Many of our indicators might be the results of other hidden indicators, such as a "development" indicator.

—