363 Final Project Code

Kelsey Evans, Megan Ahern, Armin Thomas

May 4, 2019

```
#Making the data sets

library(tcltk)
library(rpanel)
```

```
## Package `rpanel', version 1.1-4: type help(rpanel) for summary information
```

```
library(sp)
library(tkrplot)
library(lattice)
library(SpatialEpi)
library(MASS)
library(biotools)
```

```
## ---
## biotools version 3.1
```

```
##
```

```
library(DiscriMiner)
library(klaR)
library(readr)
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.5.3
```

```
library(rela)
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 3.5.3
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following object is masked from 'package:psych':
##
##     AUC
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
library(GPArotation)
library(aplpack)
library(fpc)
library(cluster)
library(ape)
library(rms)
```

```
## Warning: package 'rms' was built under R version 3.5.3
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 3.5.3
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:ape':
##
##     zoom
```

```
## The following object is masked from 'package:psych':
##
##     describe
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

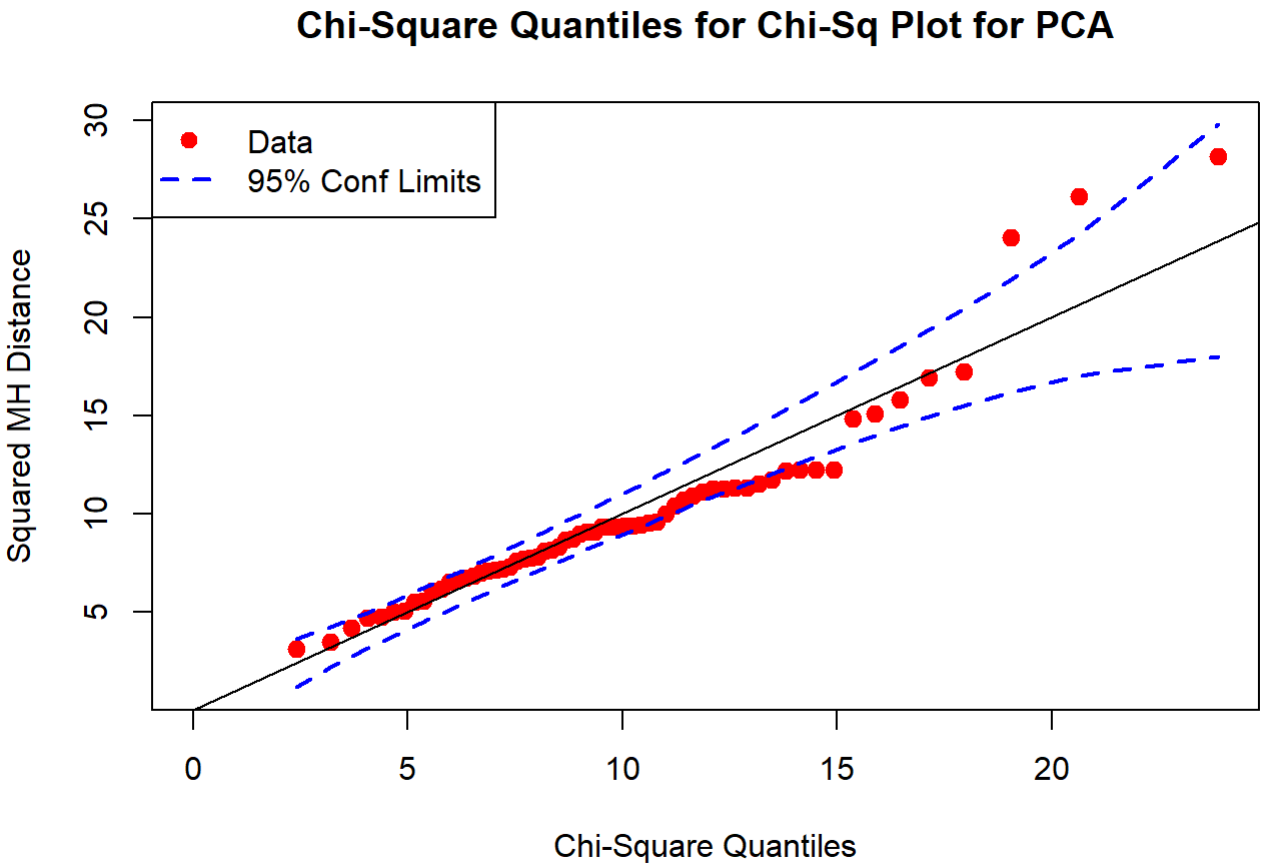
```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
library(Hmisc)
library(survival)
library(Formula)
library(ggplot2)
library(contrast)
```

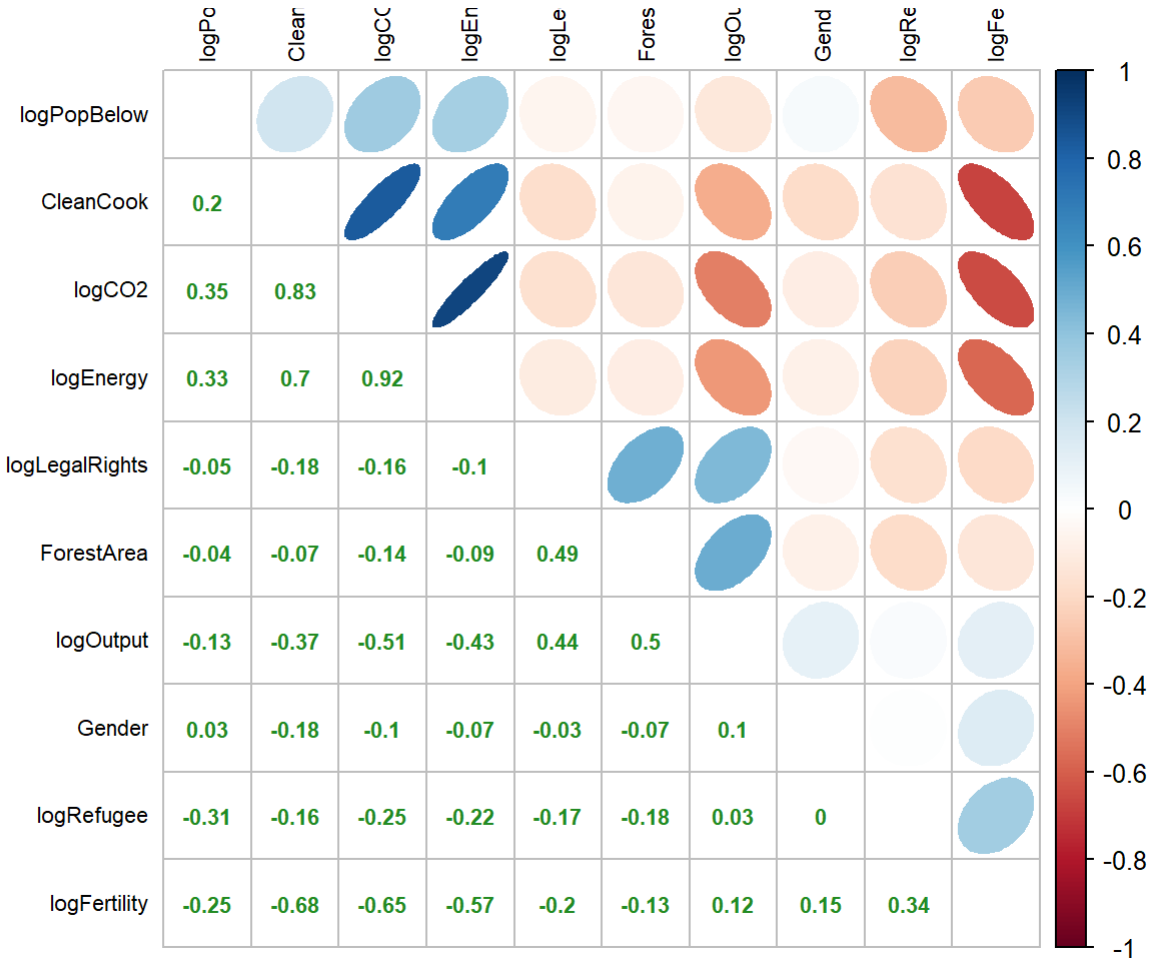
```
## Warning: package 'contrast' was built under R version 3.5.3
```

```
library(sandwich)

WB5 <- read.csv("C:/Users/HP/Documents/3. S&DS 363/WB5.csv", header = TRUE)
rownames(WB5) <- WB5[, 1]
WB5 <- WB5[, -1]
WB5$logRefugee <- log(WB5$Refugee)
WB5$logCO2 <- log(WB5$CO2)
WB5$logFertility <- log(WB5$Fertility)
WB5$logEnergy <- log(WB5$Energy)
WB5$logOutput <- log(WB5$Output)
WB5$logPopBelow <- log(WB5$PopBelow)
WB5$logLegalRights <- log(WB5$LegalRights)
WB5 <- WB5[, c(1:4, 8, 13:19)]
WB5 <- WB5[complete.cases(WB5[, 1:12]), 1:12]
is.na(WB5) <- sapply(WB5, is.infinite)
WB5[is.na(WB5)] <- 0
WB5 <- droplevels(WB5, exclude = "")
WB5 <- WB5[c(1:55, 57:64), ]


WB1 <- WB5[ ,c(1, 2, 4, 7, 8, 9)]   #MANOVA
WB2 <- WB5[ , c(3:12)]     #all else
```

```
source("http://www.reuningscherer.net/STAT660/R/CSQPlot.r.txt")
CSQPlot(WB2, label = "Chi-Sq Plot for PCA")
```




```
corrplot.mixed(cor(WB2), lower.col = "forest green", upper = "ellipse", tl.col = "black", numbe
r.cex = 0.7, order = "hclust",tl.pos = "lt", tl.cex = 0.7)
```




```
#PCA
#Principal components (cor = true means that it is scaled):
pc1 <- princomp(WB2, cor=TRUE)
#Objects created:
names (pc1)
```

```
## [1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"
## [7] "call"
```

```
#Summary:
print(summary(pc1),digits=2,loadings=pc1$loadings,cutoff=0)
```

```
## 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
```

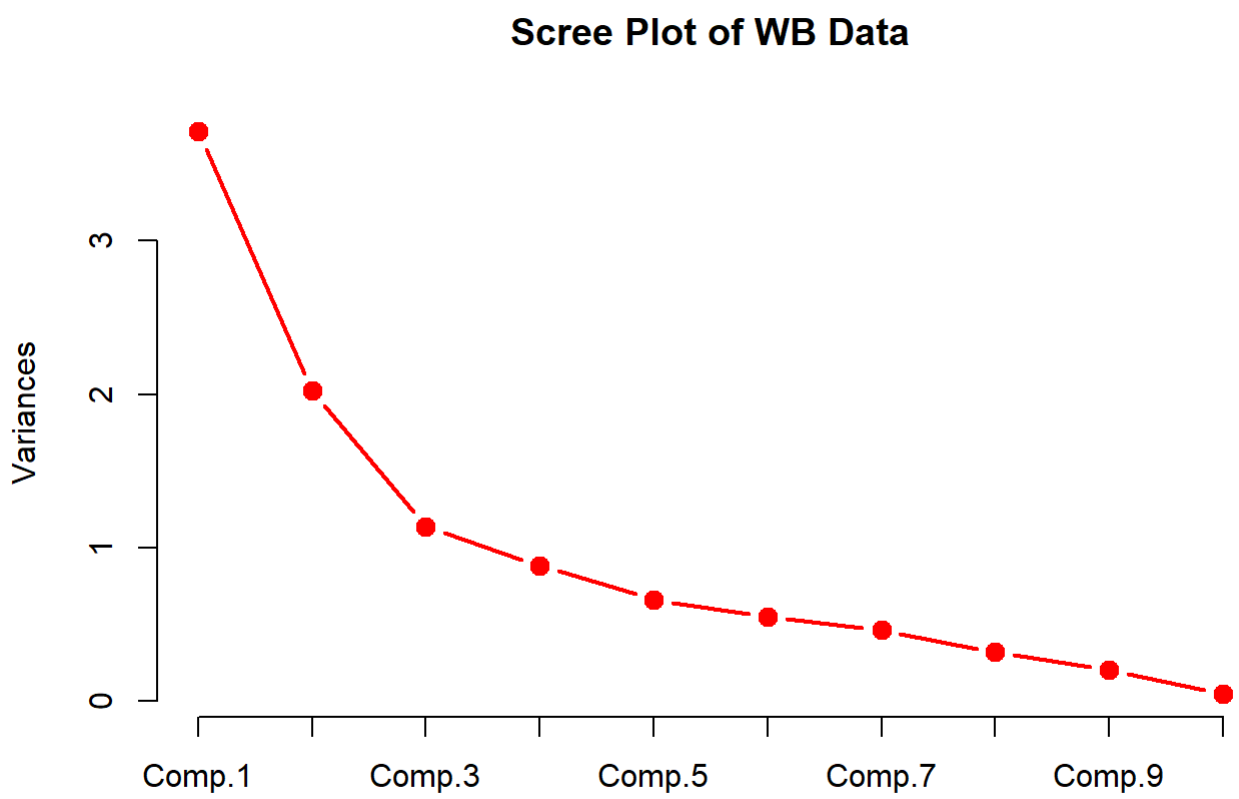
```
## Warning in if (loadings) {: the condition has length > 1 and only the first
## element will be used
```

```
##
## 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
```

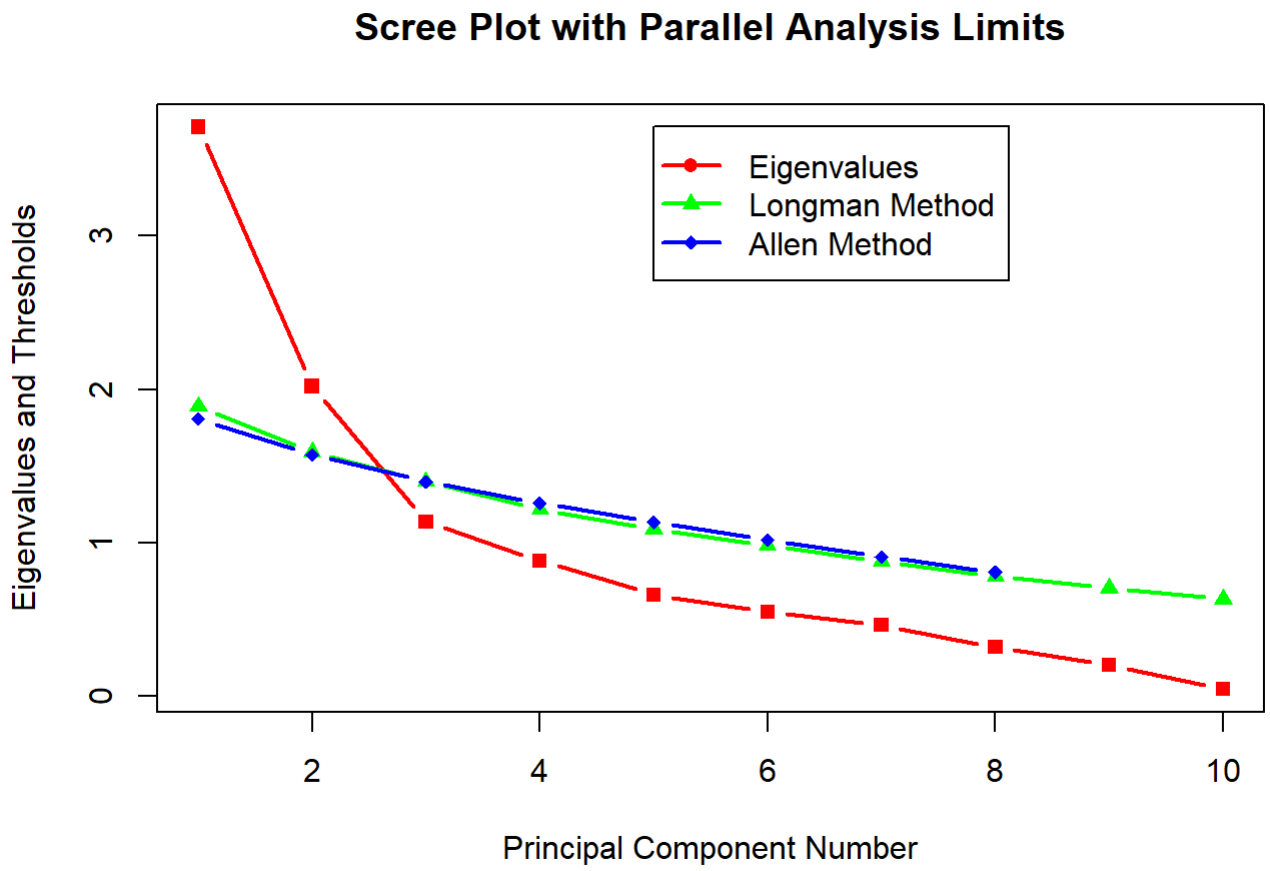
```
#Eigenvalues:
round(pc1$sdev^2,2)
```

```
##  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
```

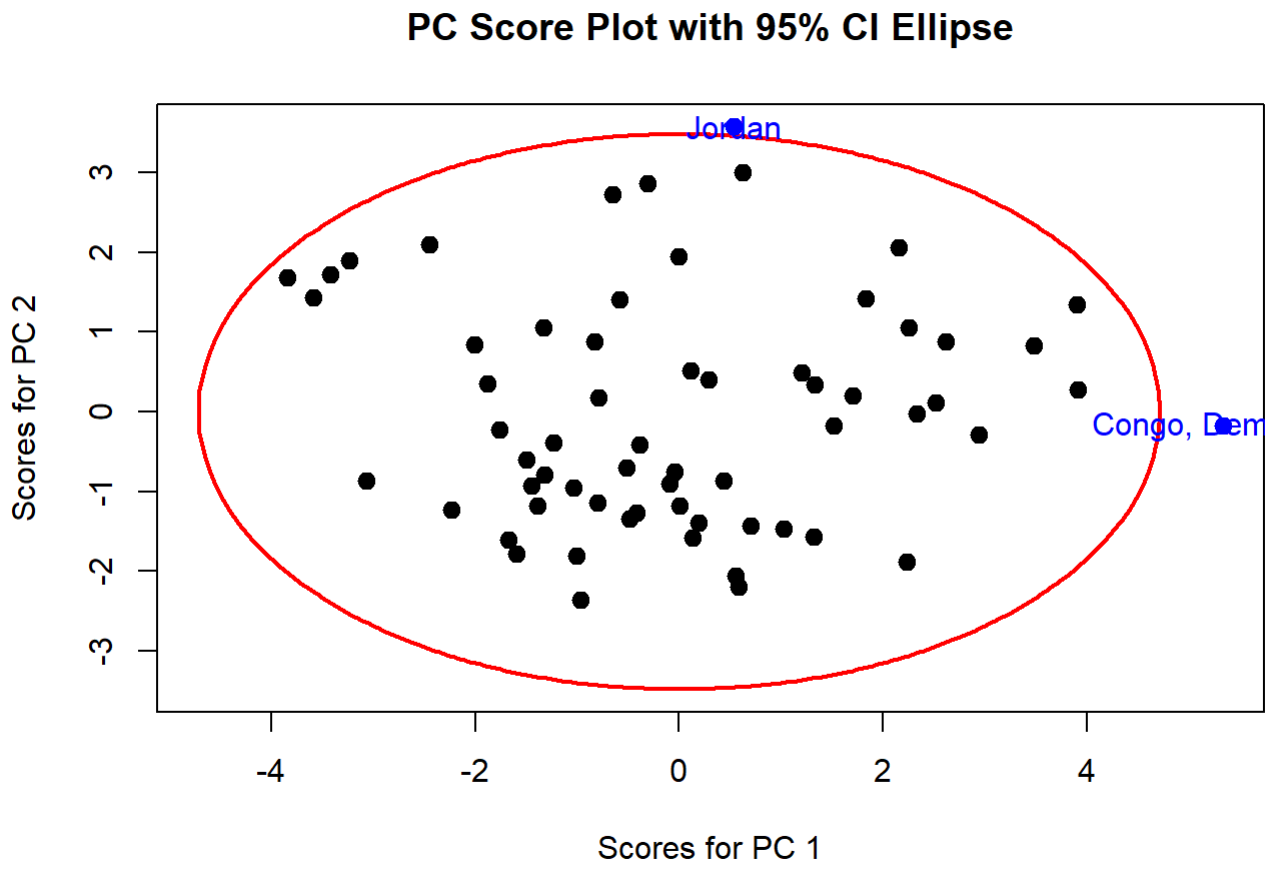
```
#Screeplot:
screeplot(pc1,type="lines",col="red",lwd=2,pch=19,cex=1.2,main="Scree Plot of WB Data")
```



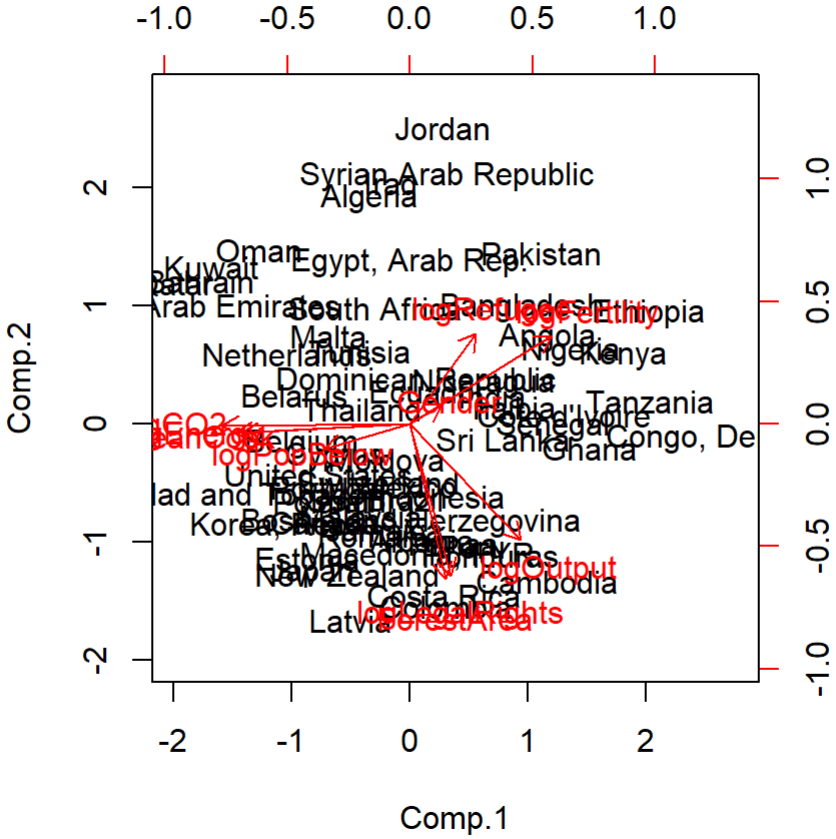

```
#Parallel analysis:
source("http://www.reuningscherer.net/STAT660/R/parallel.r.txt")
parallelplot(pc1)
```




```
source("http://reuningscherer.net/stat660/r/ciscoreplot.R.txt")

#Scoreplot with confidence intervals (used JDRS code):
WB2$names <- rownames(WB2)
ciscoreplot(pc1, c(1,2), WB2[, 11])
```




```
WB2 <- WB2[, -11]

#Biplot:
biplot(pc1, choices = c(1,2), pc.biplot = T)
```



We will use three components:

- The eigenvalues for the first 3 components are all greater than 1.
- According to the scree plot, the elbow also seems to be around 4, meaning we would keep the first 3 components
- The parallel analysis plot suggests using 3 components.
- However, the amount of variance explained by three components is only around 70%, while usually the cutoff would be 80%. Despite this, we would still like to keep only three components.

In a situation where you really needed to describe 80 percent of variance….

Story?

We chose to look at variables with a correlation 0.4 or larger in each component. There is a possible interpretation for two out of the five components.

- For component 1: Contains access to cooking fuel, CO2 per capita, and energy per capita. These all relate to individual energy use within a country.
- For component 2: Contains forest area, energy ouput and legal rights - unclear interpretation.
- For component 3: Contains the binary variable gender in constitution and population below sea level - unclear interpretation.

In our case, only the first component appears to tell a story about the variables it contains.

No trends in the score plot, which is good

PCA was appropriate for our data, once certain variables were logged, because we observed significant correlation between some variables (e.g. CO2 output and Clean Fuel Cooking). Also the chi-squared quantile plot, because it was quite linear, indicated that PCA was appropriate for our data set. There were no outliers on our Chi-

Squared quantile plot (i.e. only a couple of points outside of the 95% confidence bands). However, we did have two outliers on the score plot, but in a data set of 64 that's entirely possible.

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. The KMO measure was 0.71 which is an acceptable value for factor analysis.

```
#Factor analysis
#chisq, corr
#(see above)

fact1 <- paf(as.matrix(WB2))
fact1$KMO
```

```
## [1] 0.72049
```

```
(fact1 <- fa(WB2, nfactors = 3, fm = "pa"))
```

```
## Factor Analysis using method =  pa
## Call: fa(r = WB2, nfactors = 3, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##                   PA1   PA2   PA3    h2    u2 com
## Gender          -0.25 -0.08  0.18 0.056 0.944 2.0
## CleanCook        0.95 -0.02 -0.17 0.803 0.197 1.1
## ForestArea       0.06  0.73 -0.01 0.519 0.481 1.0
## logRefugee      -0.09 -0.22 -0.49 0.323 0.677 1.5
## logCO2           0.89 -0.14  0.14 0.988 0.012 1.1
## logFertility    -0.75 -0.33 -0.08 0.656 0.344 1.4
## logEnergy        0.75 -0.10  0.17 0.741 0.259 1.1
## logOutput       -0.31  0.62 -0.04 0.558 0.442 1.5
## logPopBelow      0.04 -0.07  0.58 0.362 0.638 1.0
## logLegalRights   0.00  0.69  0.05 0.472 0.528 1.0
##
##                        PA1  PA2  PA3
## SS loadings           3.11 1.60 0.77
## Proportion Var        0.31 0.16 0.08
## Cumulative Var        0.31 0.47 0.55
## Proportion Explained  0.57 0.29 0.14
## Cumulative Proportion 0.57 0.86 1.00
##
##  With factor correlations of
##       PA1   PA2  PA3
## PA1  1.00 -0.17 0.43
## PA2 -0.17  1.00 0.01
## PA3  0.43  0.01 1.00
##
## Mean item complexity =  1.3
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are  45  and the objective function was  5.57 with
Chi Square of  322.36
## The degrees of freedom for the model are 18  and the objective function was  0.58
##
## The root mean square of the residuals (RMSR) is  0.03
## The df corrected root mean square of the residuals is  0.05
##
## The harmonic number of observations is  63 with the empirical chi square  5.46  with prob <
1
## The total number of observations was  63  with Likelihood Chi Square =  32.39  with prob <
0.02
##
## Tucker Lewis Index of factoring reliability =  0.865
## RMSEA index =  0.127  and the 90 % confidence intervals are  0.045 0.175
## BIC =  -42.18
## Fit based upon off diagonal values = 0.99
## Measures of factor score adequacy
##                                                  PA1  PA2  PA3
## Correlation of (regression) scores with factors   1.00 0.89 0.81
## Multiple R square of scores with factors          1.00 0.78 0.65
## Minimum correlation of possible factor scores     0.99 0.57 0.30
```

```
repro1 <- fact1$loadings%*%t(fact1$loadings)
resid1 <- cor(WB2)-repro1
round(resid1,2)
```

```
##                Gender CleanCook ForestArea logRefugee logCO2 logFertility
## Gender           0.90      0.09       0.01       0.05   0.09        -0.05
## CleanCook        0.09      0.07      -0.12      -0.16   0.00         0.02
## ForestArea       0.01     -0.12       0.47      -0.02  -0.09         0.15
## logRefugee       0.05     -0.16      -0.02       0.71  -0.13         0.17
## logCO2           0.09      0.00      -0.09      -0.13   0.16        -0.01
## logFertility    -0.05      0.02       0.15       0.17  -0.01         0.31
## logEnergy        0.08      0.01      -0.07      -0.09   0.21        -0.02
## logOutput        0.08     -0.07       0.07       0.12  -0.13         0.08
## logPopBelow     -0.06      0.26       0.01      -0.04   0.23        -0.20
## logLegalRights   0.01     -0.16      -0.01       0.01  -0.07         0.04
##                logEnergy logOutput logPopBelow logLegalRights
## Gender              0.08      0.08       -0.06           0.01
## CleanCook           0.01     -0.07        0.26          -0.16
## ForestArea         -0.07      0.07        0.01          -0.01
## logRefugee         -0.09      0.12       -0.04           0.01
## logCO2              0.21     -0.13        0.23          -0.07
## logFertility       -0.02      0.08       -0.20           0.04
## logEnergy           0.39     -0.13        0.20          -0.04
## logOutput          -0.13      0.52       -0.05           0.02
## logPopBelow         0.20     -0.05        0.66          -0.03
## logLegalRights     -0.04      0.02       -0.03           0.53
```

```
len1 <- length(resid1[upper.tri(resid1)])
(RMSR1 <- sqrt(sum(resid1[upper.tri(resid1)]^2)/len1))
```

```
## [1] 0.10671
```

```
sum(rep(1,len1)[abs(resid1[upper.tri(resid1)])>0.05])/len1
```

```
## [1] 0.62222
```

```
fact2 <- fa(WB2, nfactors = 3, fm = "pa", rotate = "varimax")
repro2 <- fact2$loadings%*%t(fact2$loadings)
resid2 <- cor(WB2)-repro2
round(resid2,2)
```

```
##                Gender CleanCook ForestArea logRefugee logCO2 logFertility
## Gender           0.94     -0.01      -0.03       0.01   0.03         0.02
## CleanCook       -0.01      0.20       0.01       0.00  -0.02        -0.05
## ForestArea      -0.03      0.01       0.48      -0.02   0.02         0.06
## logRefugee       0.01      0.00      -0.02       0.68   0.03         0.04
## logCO2           0.03     -0.02       0.02       0.03   0.01         0.03
## logFertility     0.02     -0.05       0.06       0.04   0.03         0.34
## logEnergy        0.04     -0.03       0.02       0.04   0.06         0.03
## logOutput        0.08      0.03       0.04       0.05   0.00         0.00
## logPopBelow      0.00      0.01       0.00      -0.01  -0.01         0.00
## logLegalRights  -0.01     -0.07      -0.01       0.00   0.01        -0.04
##                logEnergy logOutput logPopBelow logLegalRights
## Gender              0.04      0.08        0.00          -0.01
## CleanCook          -0.03      0.03        0.01          -0.07
## ForestArea          0.02      0.04        0.00          -0.01
## logRefugee          0.04      0.05       -0.01           0.00
## logCO2              0.06      0.00       -0.01           0.01
## logFertility        0.03      0.00        0.00          -0.04
## logEnergy           0.26     -0.01        0.00           0.03
## logOutput          -0.01      0.44        0.03          -0.01
## logPopBelow         0.00      0.03        0.64          -0.03
## logLegalRights      0.03     -0.01       -0.03           0.53
```

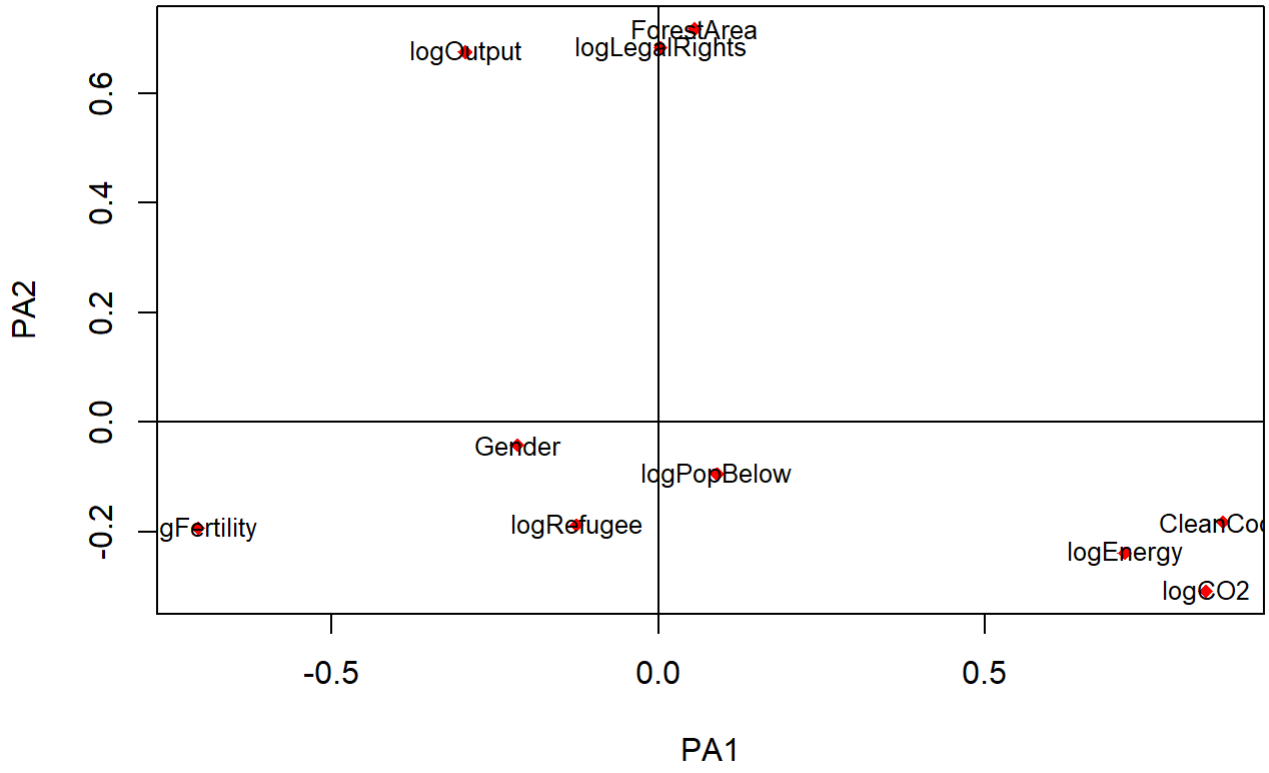
```
len2 <- length(resid2[upper.tri(resid2)])
(RMSR2 <- sqrt(sum(resid2[upper.tri(resid2)]^2)/len2))
```

```
## [1] 0.031034
```

```
sum(rep(1,len2)[abs(resid2[upper.tri(resid2)])>0.05])/len2
```

```
## [1] 0.13333
```

```
#Loading plot
plot(fact2$loadings, pch=18, col='red')
abline(h=0)
abline(v=0)
text(fact2$loadings, labels=names(WB2),cex=0.8)
```




```
#copy all code but with rotation, show that it's smaller therefore better with the rotation.
```

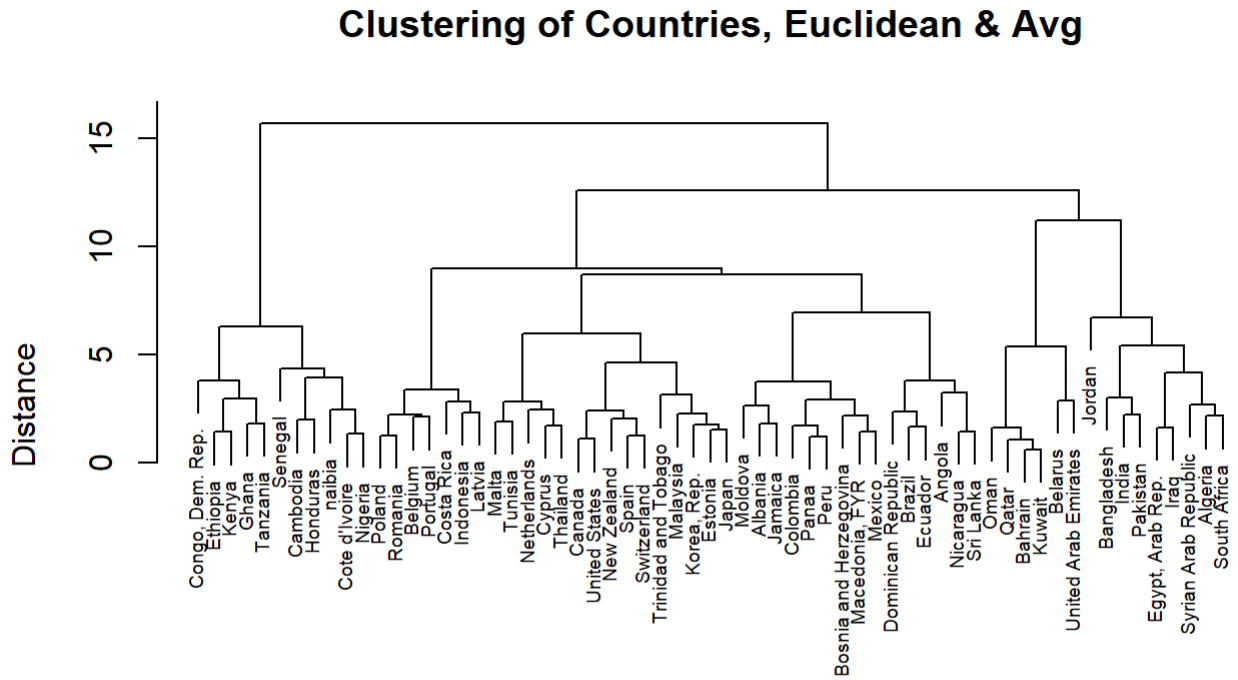
Referring to the previous question, we will use four factors…

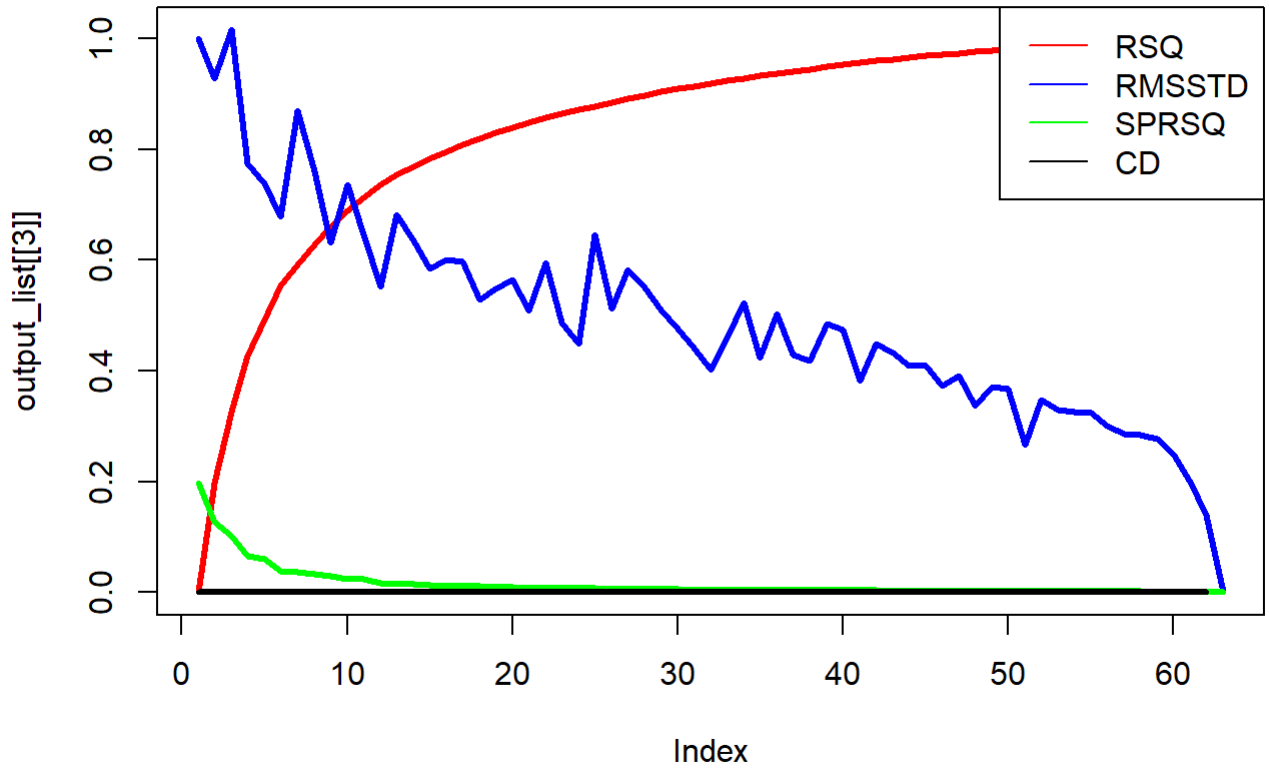
rmsr is 0.1, portion > 0.05 is 0.58

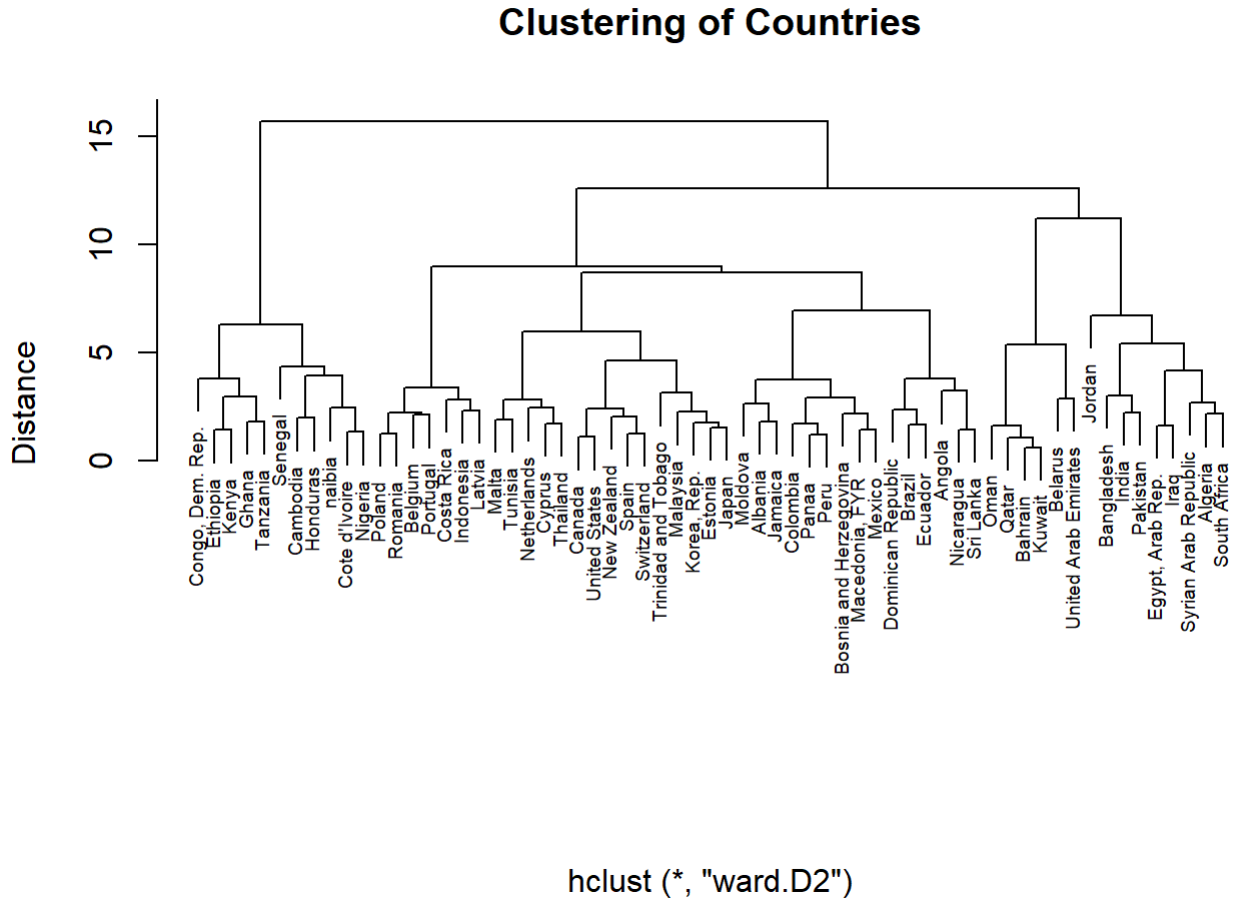
two clusters load heavily on factor 2, the third doesnt

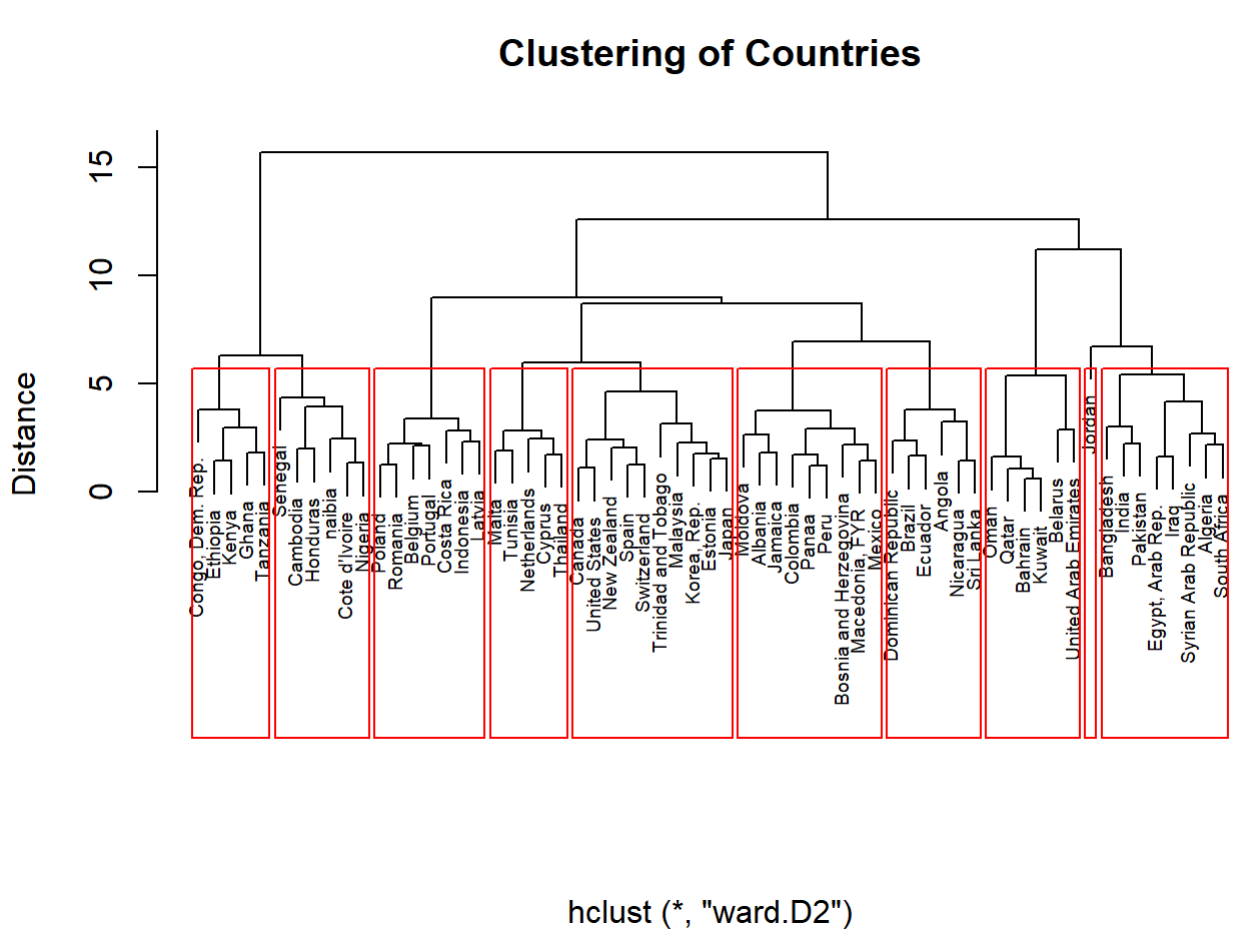
there's also not a great story. explain why

Tried two methods, this creates best graph

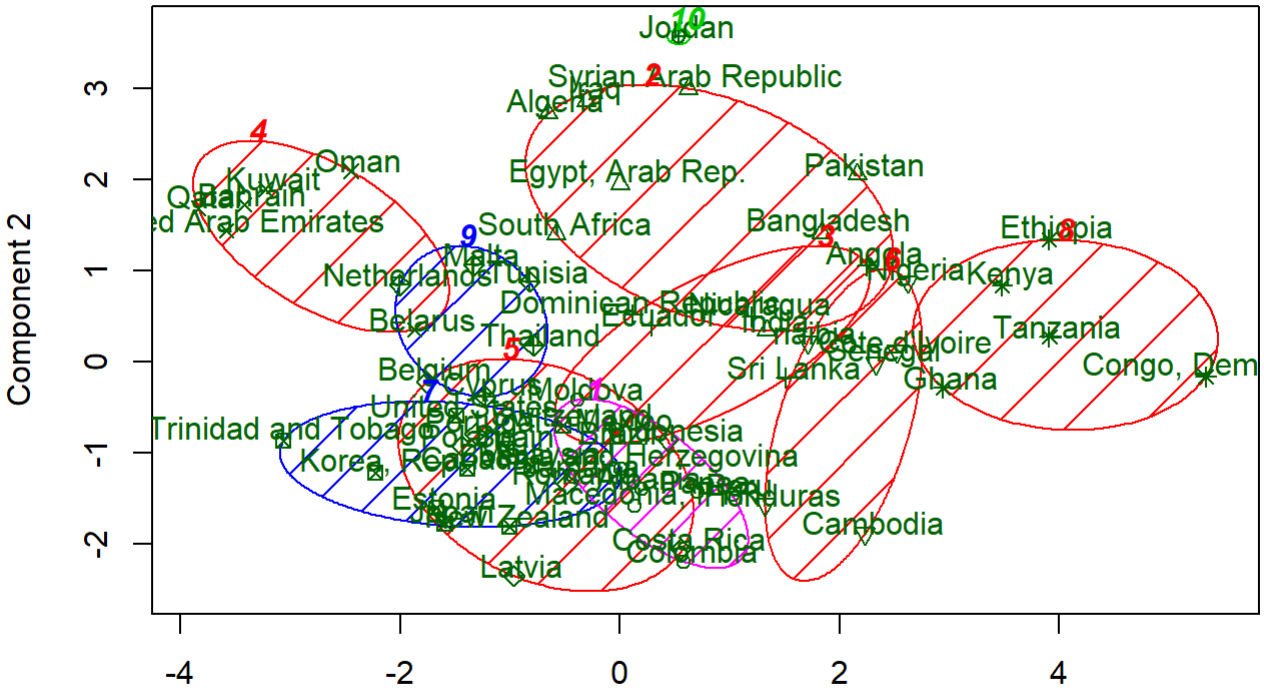





Clustering of Countries
Distance
0
5
10
15
Congo, Dem. Rep.
Ethiopia
Kenya
Ghana
Tanzania
Senegal
Cambodia
Honduras
naibia
Cote d'Ivoire
Nigeria
Poland
Romania
Belgium
Portugal
Costa Rica
Indonesia
Latvia
Malta
Tunisia
Netherlands
Cyprus
Thailand
Canada
United States
New Zealand
Spain
Switzerland
Trinidad and Tobago
Malaysia
Korea, Rep.
Estonia
Japan
Moldova
Albania
Jamaica
Colombia
Panaa
Peru
Bosnia and Herzegovina
Macedonia, FYR
Mexico
Dominican Republic
Brazil
Ecuador
Angola
Nicaragua
Sri Lanka
Oman
Qatar
Bahrain
Kuwait
Belarus
United Arab Emirates
Jordan
Bangladesh
India
Pakistan
Egypt, Arab Rep.
Iraq
Syrian Arab Republic
Algeria
South Africa
hclust (*, "ward.D2")

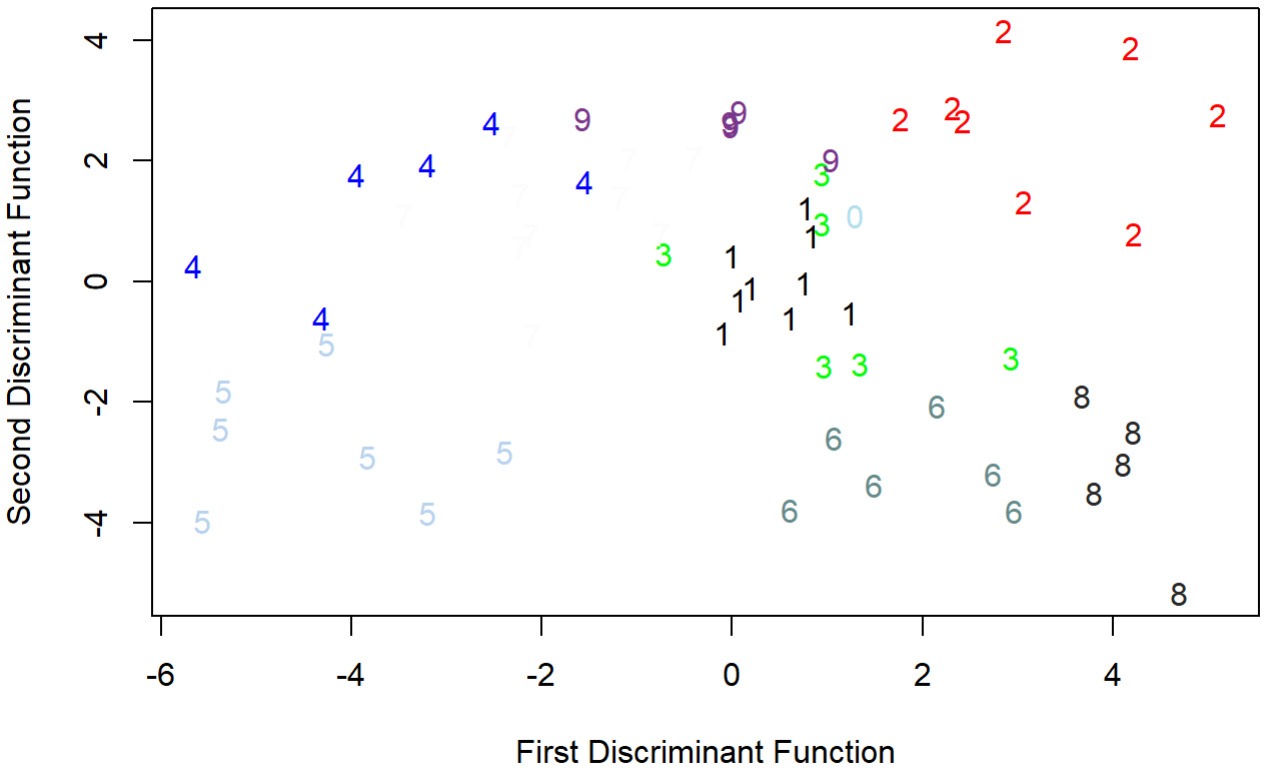

Clustering of Countries
Distance
0
5
10
15
Congo, Dem. Rep.
Ethiopia
Kenya
Ghana
Tanzania
Senegal
Cambodia
Honduras
naibia
Cote d'Ivoire
Nigeria
Poland
Romania
Belgium
Portugal
Costa Rica
Indonesia
Latvia
Malta
Tunisia
Netherlands
Cyprus
Thailand
Canada
United States
New Zealand
Spain
Switzerland
Trinidad and Tobago
Malaysia
Korea, Rep.
Estonia
Japan
Moldova
Albania
Jamaica
Colombia
Panaa
Peru
Bosnia and Herzegovina
Macedonia, FYR
Mexico
Dominican Republic
Brazil
Ecuador
Angola
Nicaragua
Sri Lanka
Oman
Qatar
Bahrain
Kuwait
Belarus
United Arab Emirates
Jordan
Bangladesh
India
Pakistan
Egypt, Arab Rep.
Iraq
Syrian Arab Republic
Algeria
South Africa
hclust (*, "ward.D2")

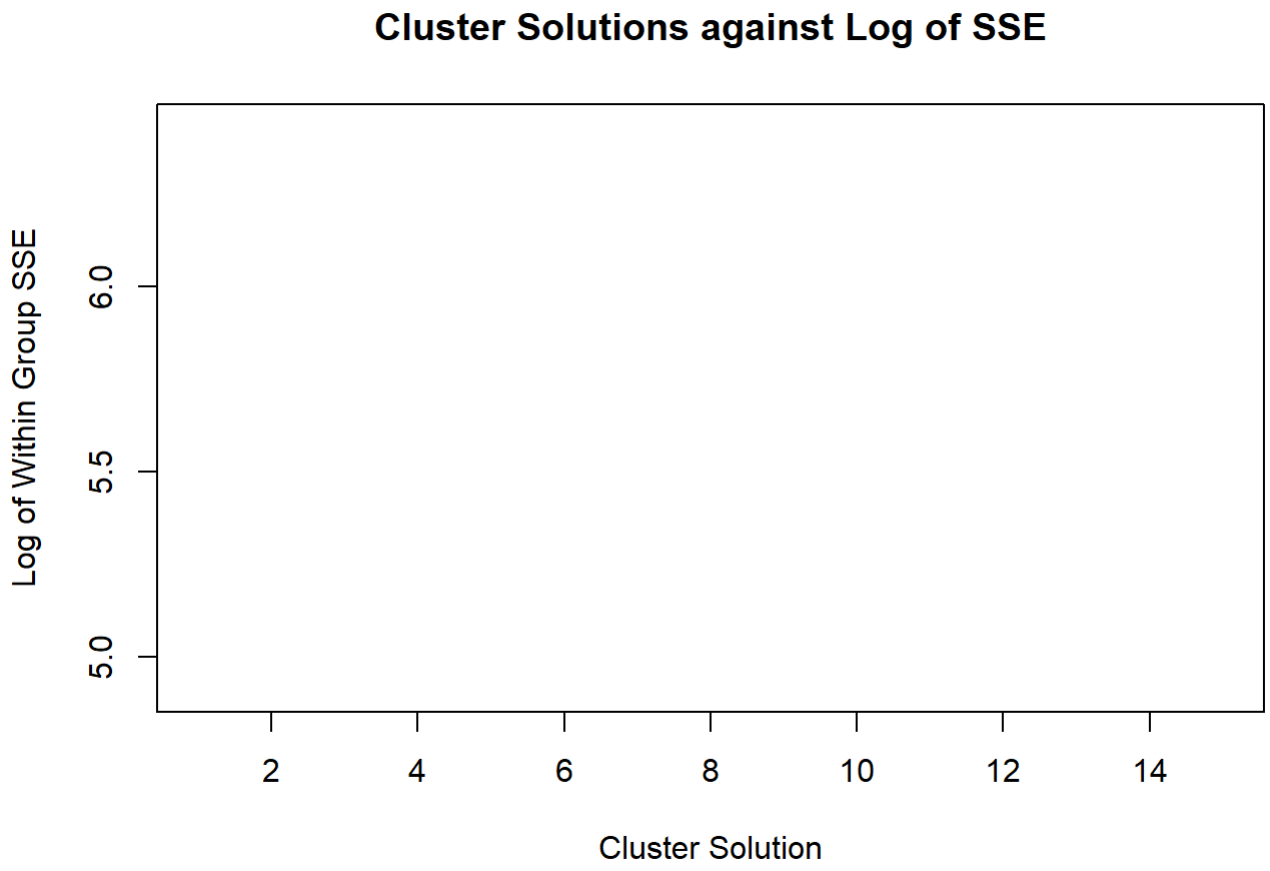
World Bank 10 Cluster Plot, Avg Method, First two PC

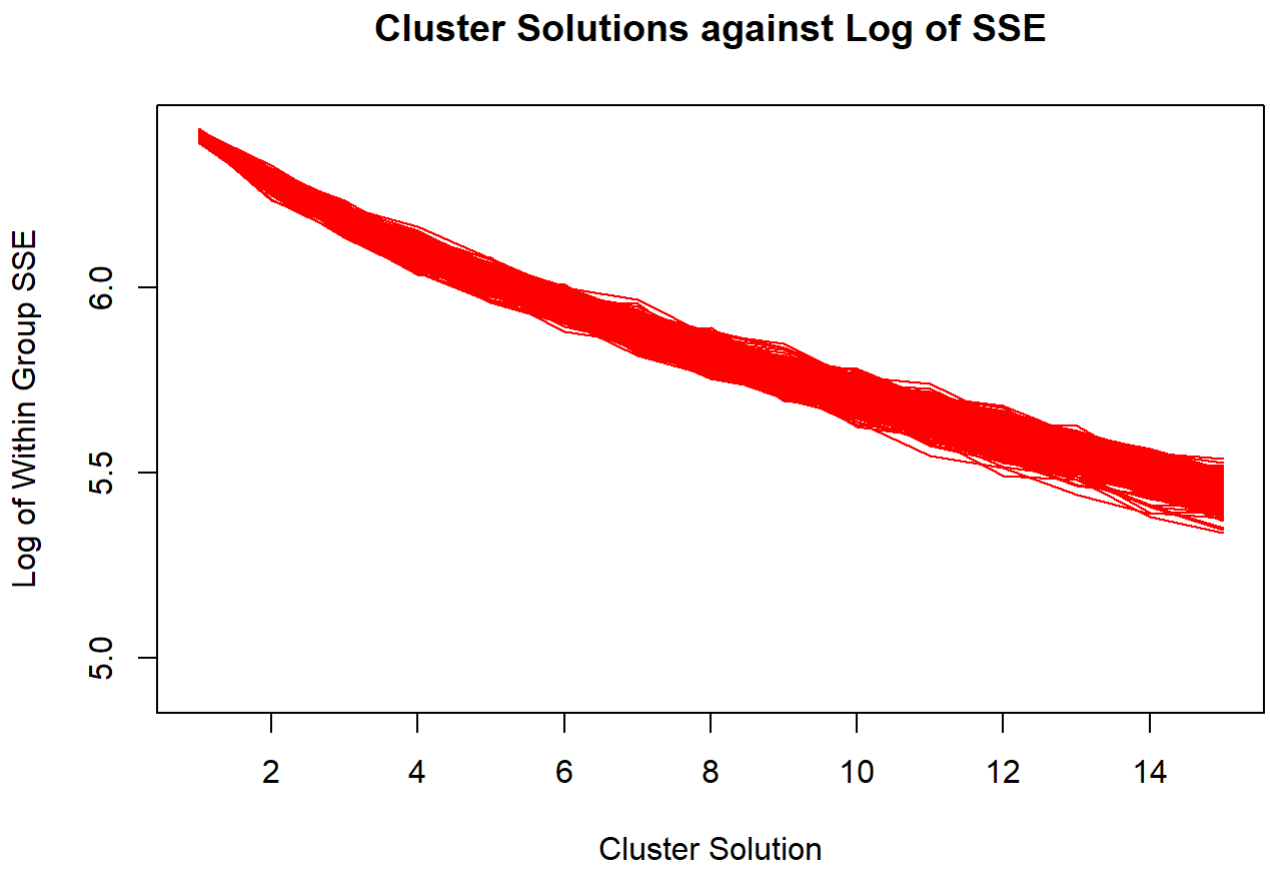


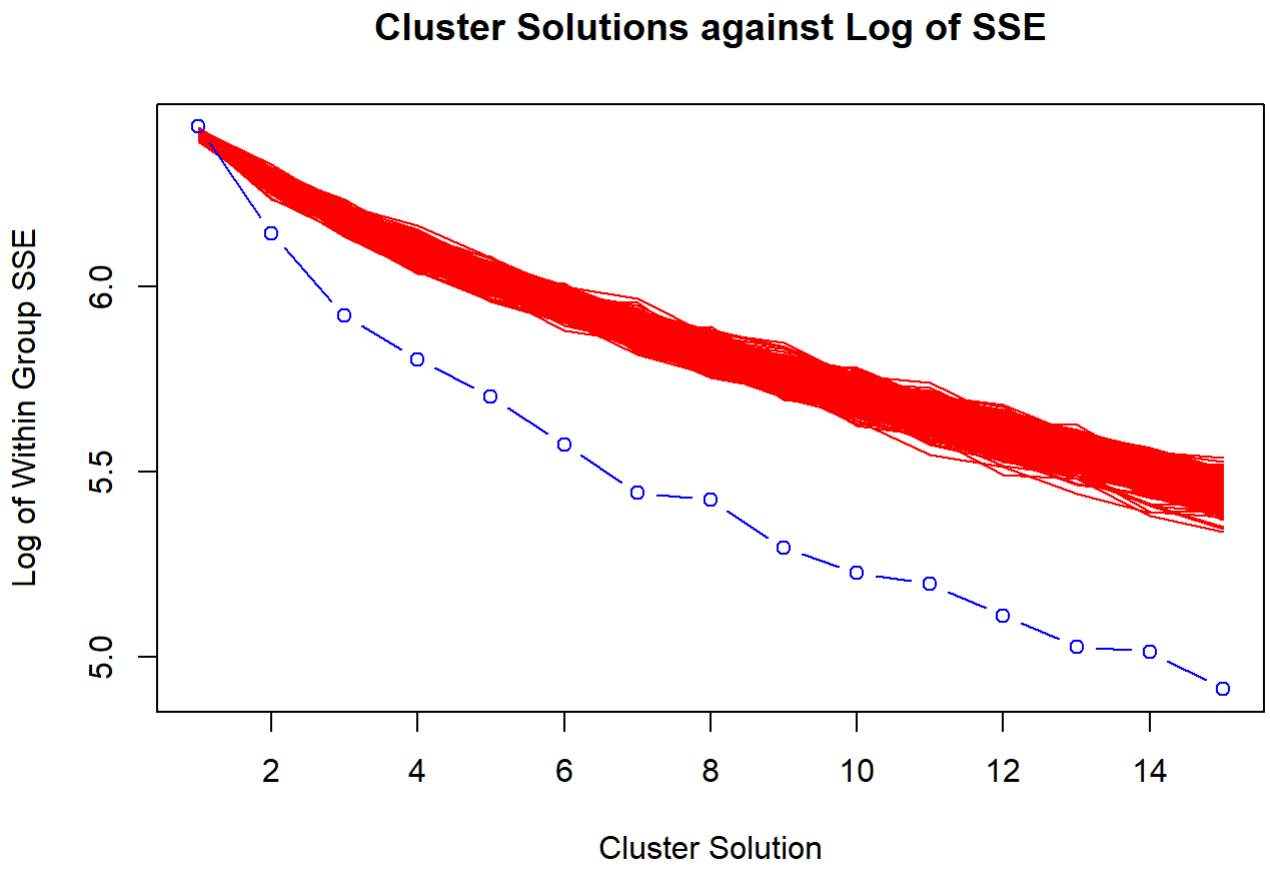
These two components explain 57.34 % of the point variability.

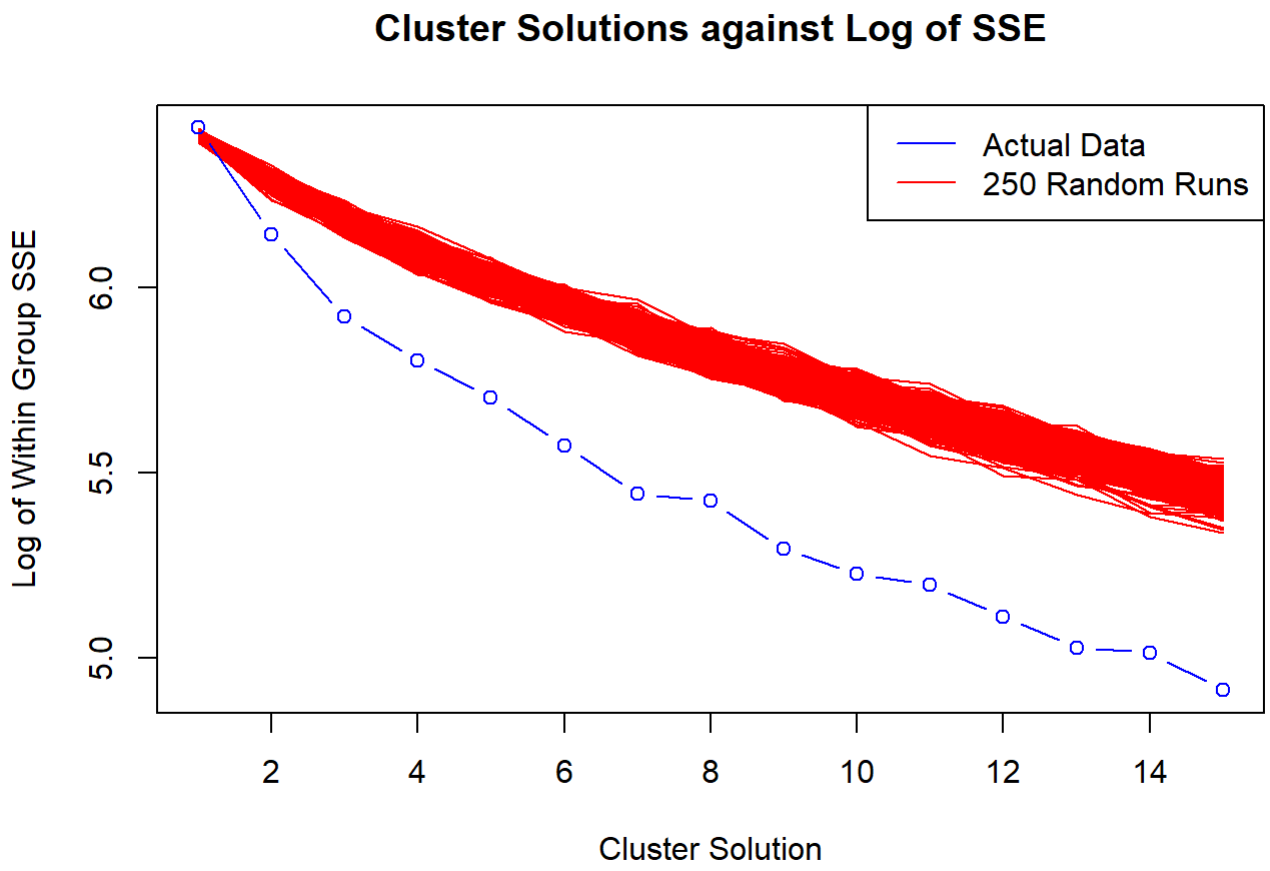
10 Cluster Solution in DA Space

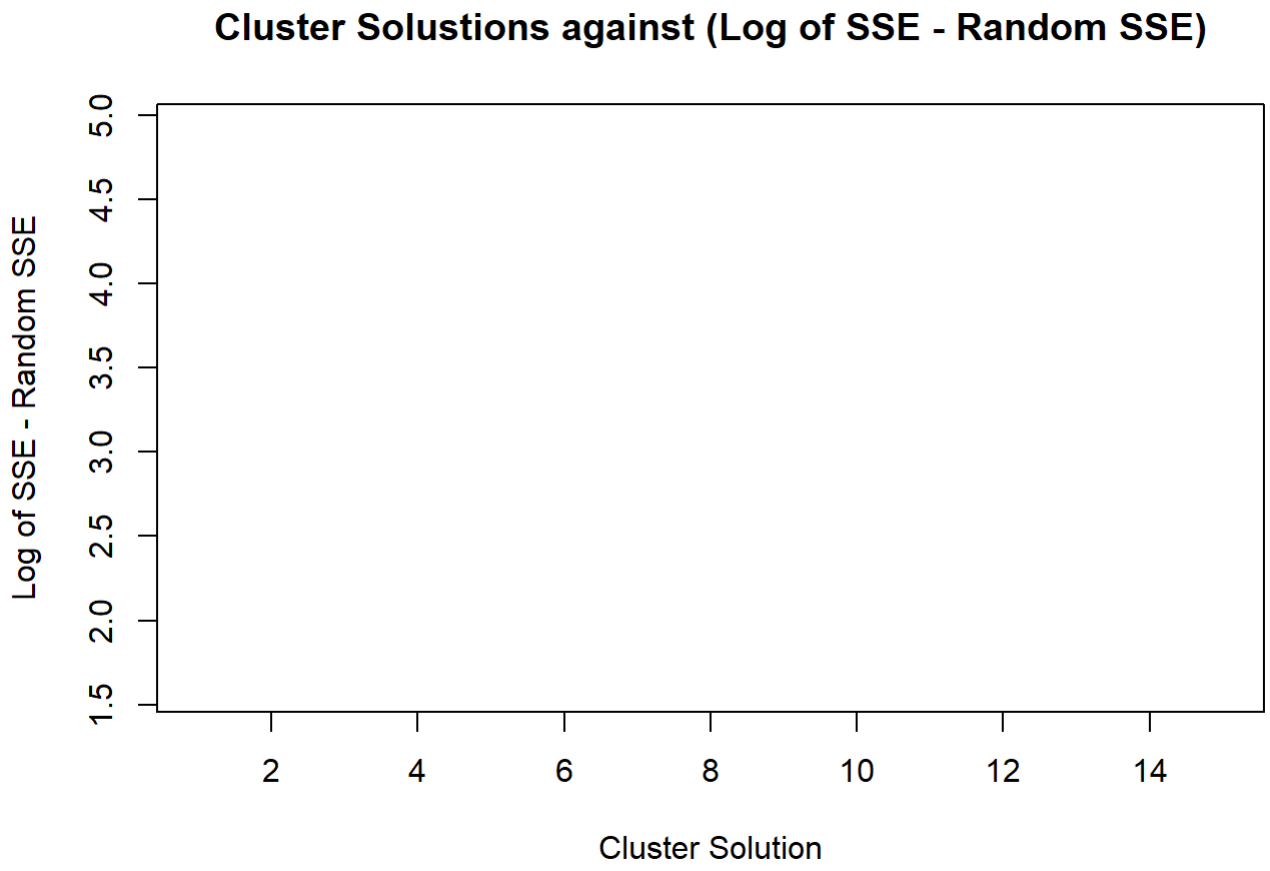


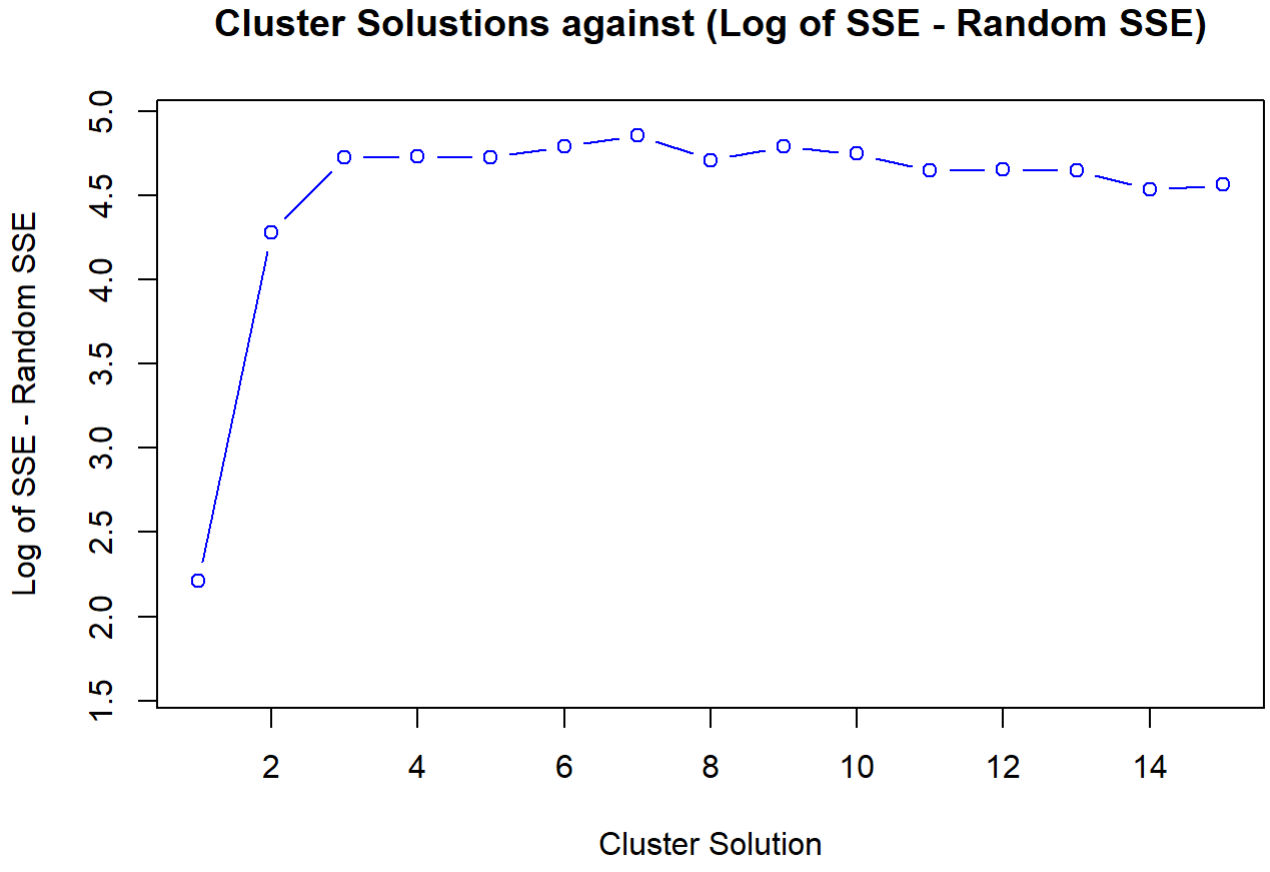

Cluster Solutions against Log of SSE
Log of Within Group SSE
5.0
5.5
6.0
2
4
6
8
10
12
14
Cluster Solution

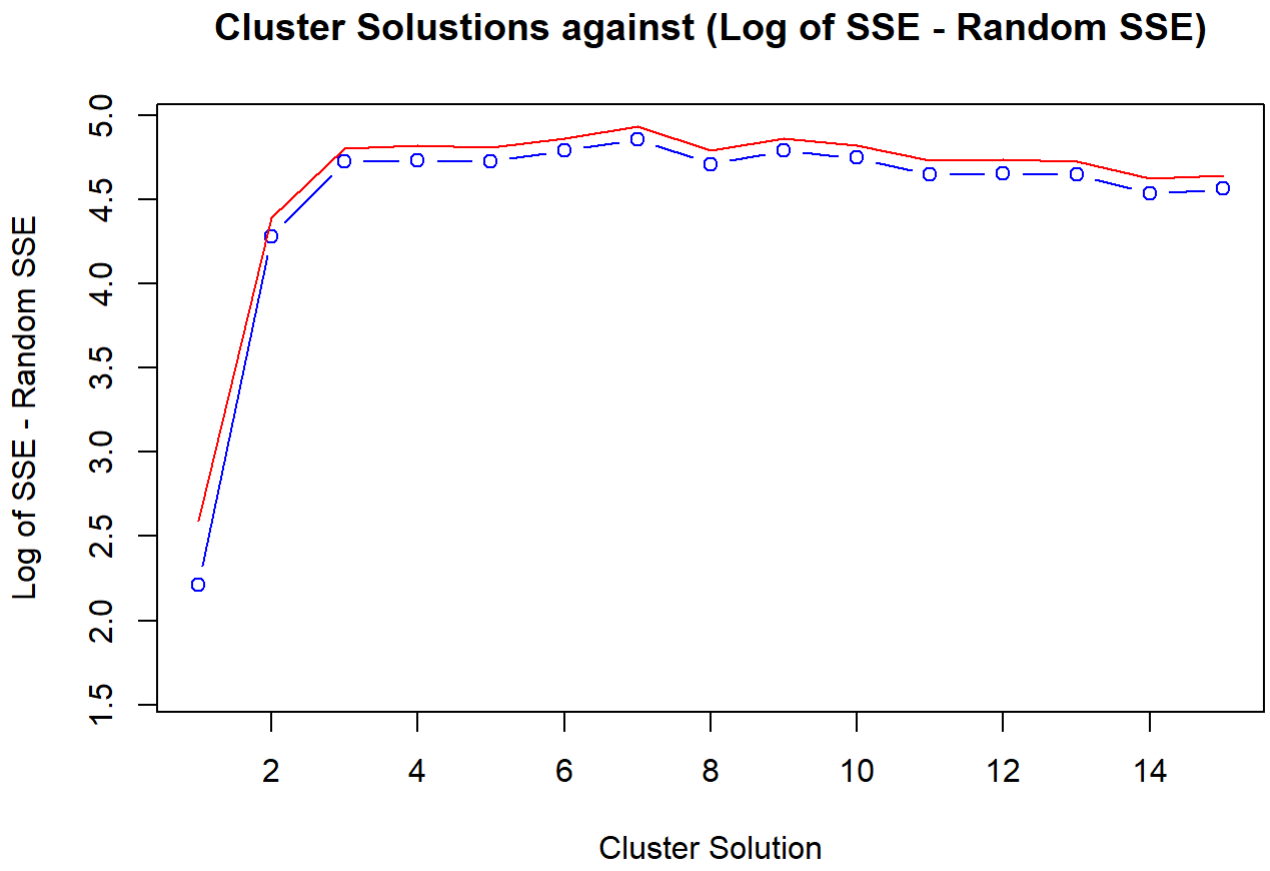

Cluster Solutions against Log of SSE
Log of Within Group SSE
5.0
5.5
6.0
2
4
6
8
10
12
14
Cluster Solution

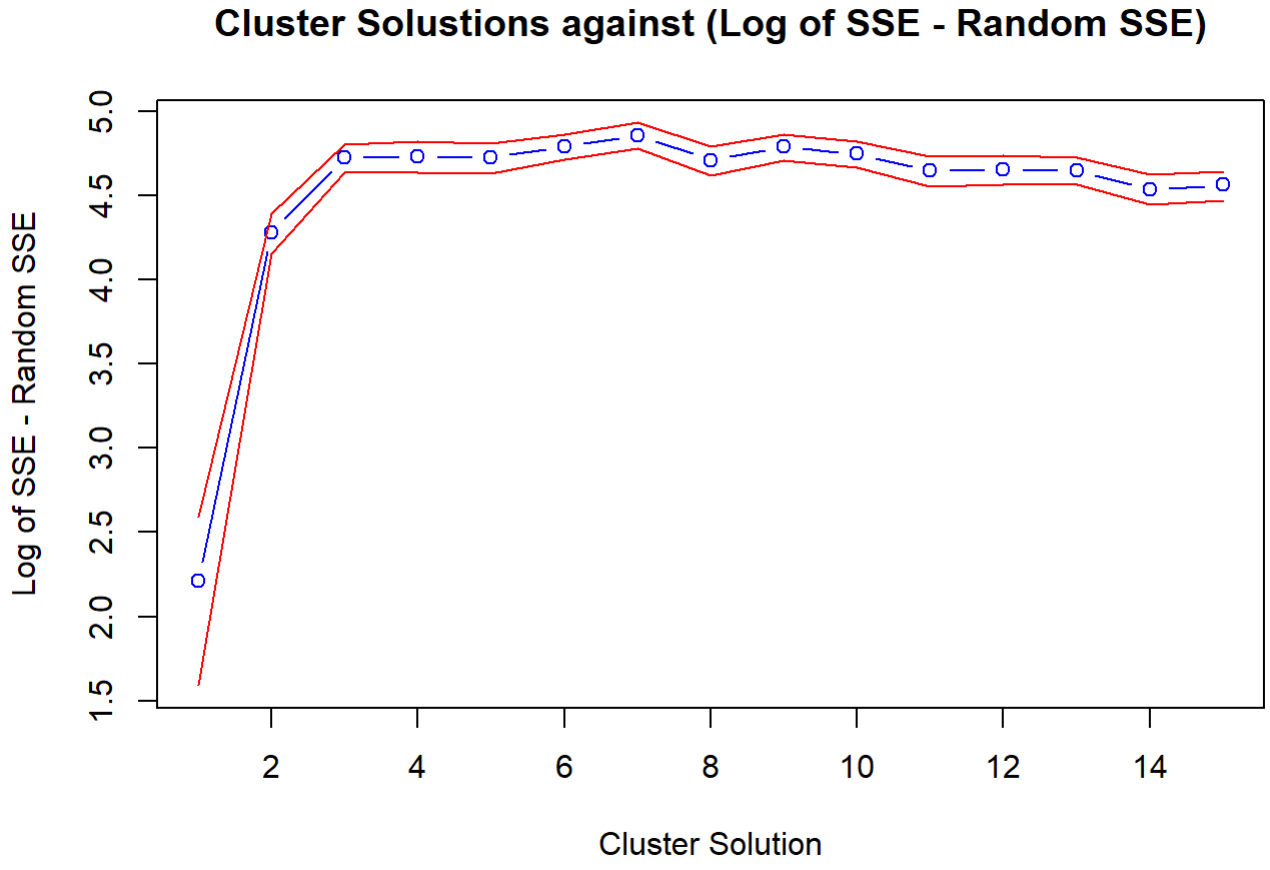

Cluster Solutions against Log of SSE
Log of Within Group SSE
5.0
5.5
6.0
2
4
6
8
10
12
14
Cluster Solution

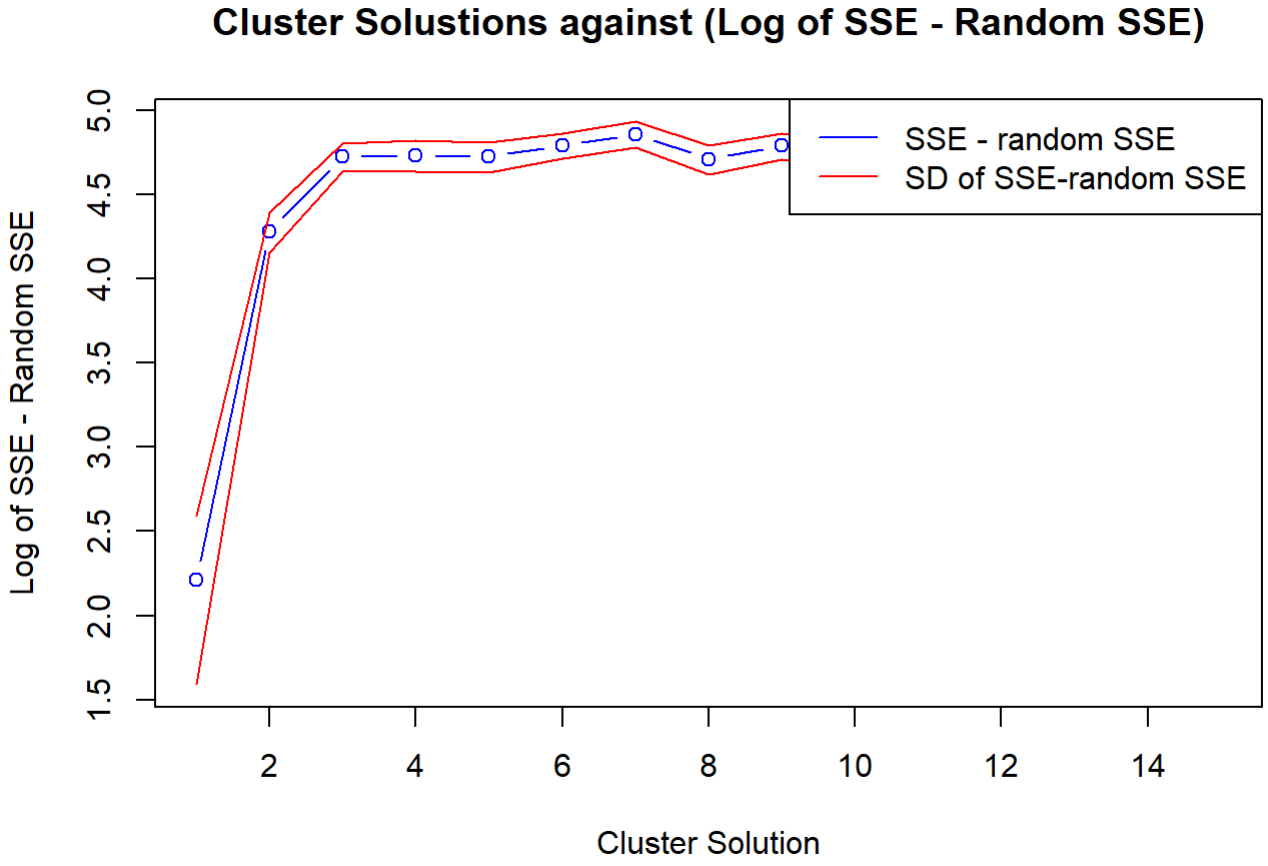

Cluster Solutions against Log of SSE
Actual Data
250 Random Runs
Log of Within Group SSE
5.0
5.5
6.0
2
4
6
8
10
12
14
Cluster Solution

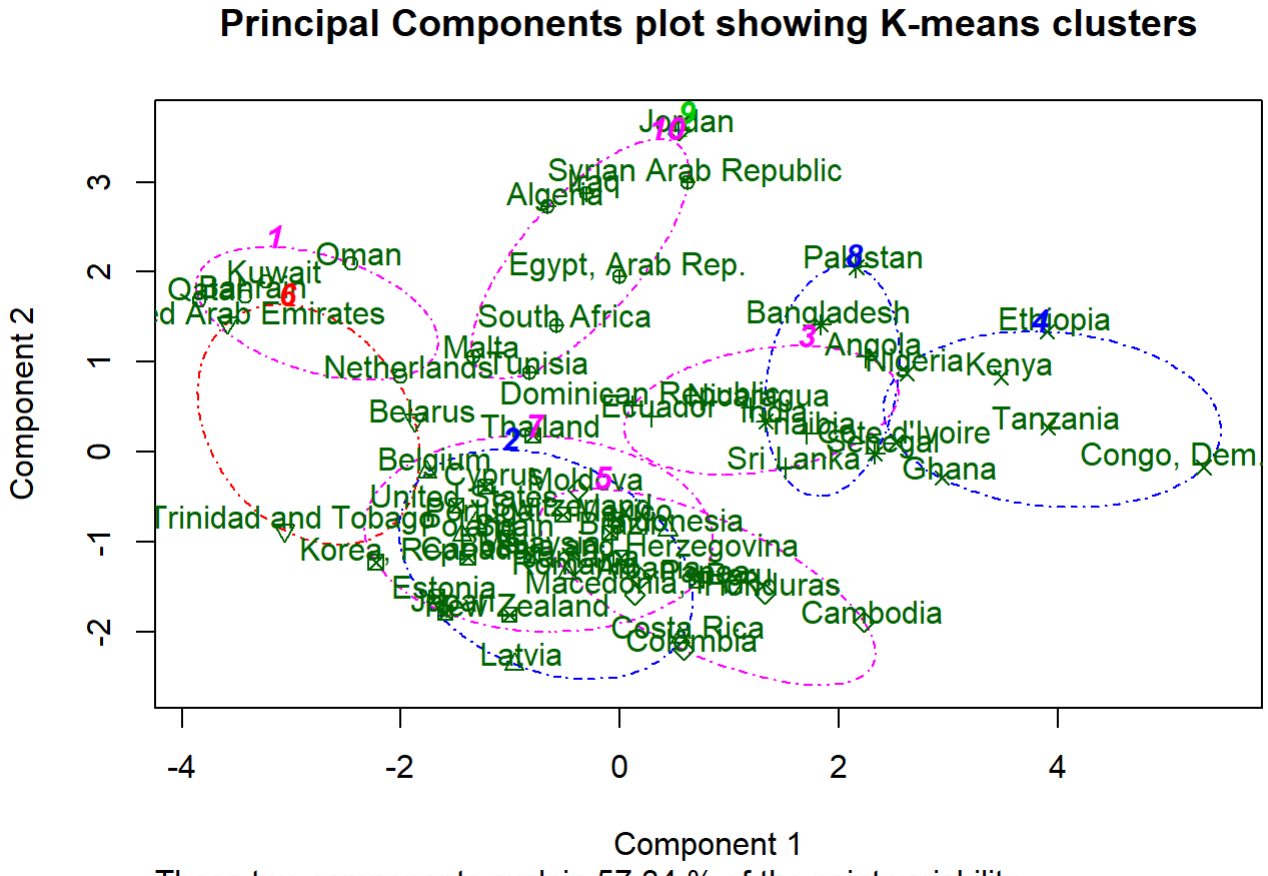

Cluster Solustions against (Log of SSE - Random SSE)
Log of SSE - Random SSE
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2
4
6
8
10
12
14
Cluster Solution


Cluster Solustions against (Log of SSE - Random SSE)
Log of SSE - Random SSE
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2
4
6
8
10
12
14
Cluster Solution

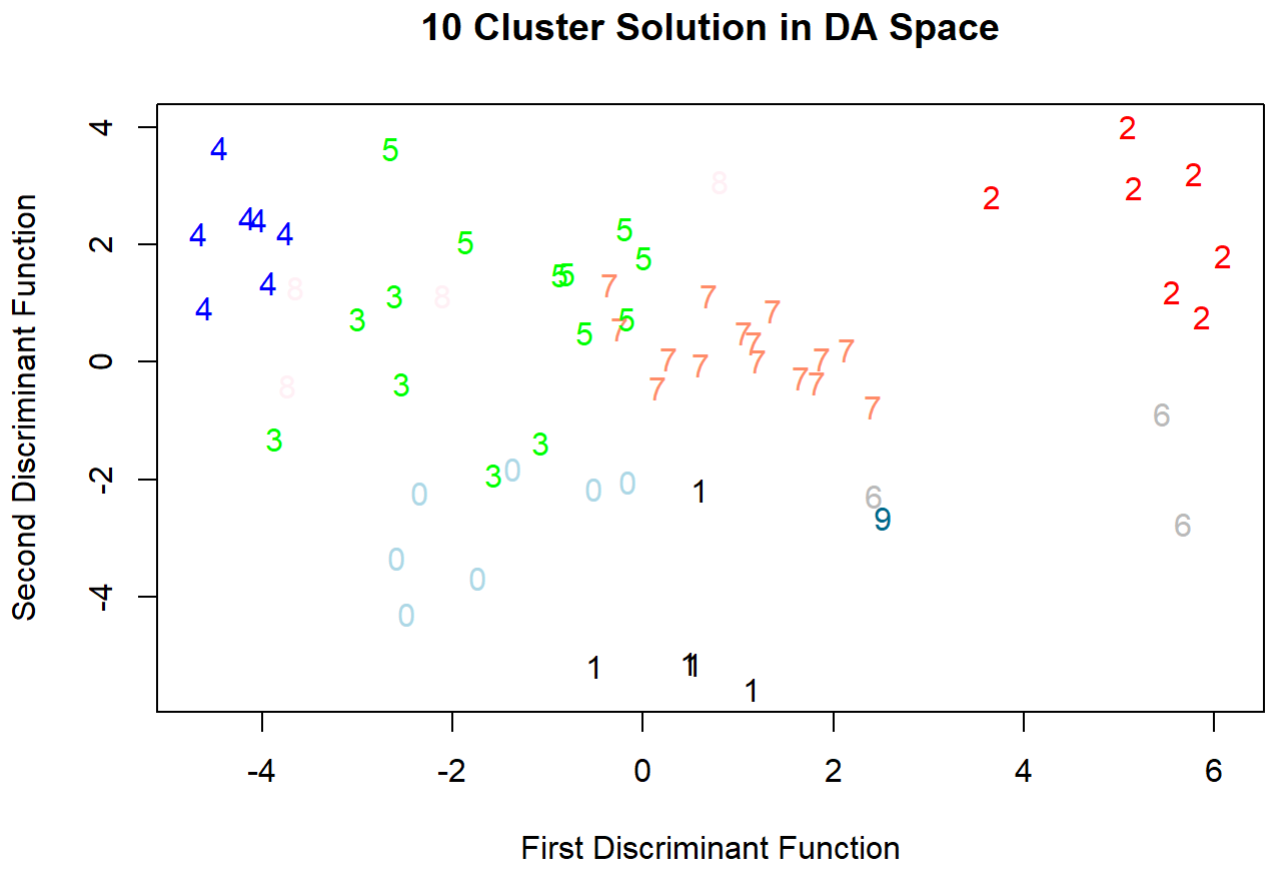

Cluster Solustions against (Log of SSE - Random SSE)
Log of SSE - Random SSE
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2
4
6
8
10
12
14
Cluster Solution


Cluster Solustions against (Log of SSE - Random SSE)
Log of SSE - Random SSE
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2
4
6
8
10
12
14
Cluster Solution





These two components explain 57.34 % of the point variability.



-10 clusters (argument for 5, both in original and in the randomized sum of squared errors (second plot)) - argument for 10 in the hclust eval graph, also looks nice with 10

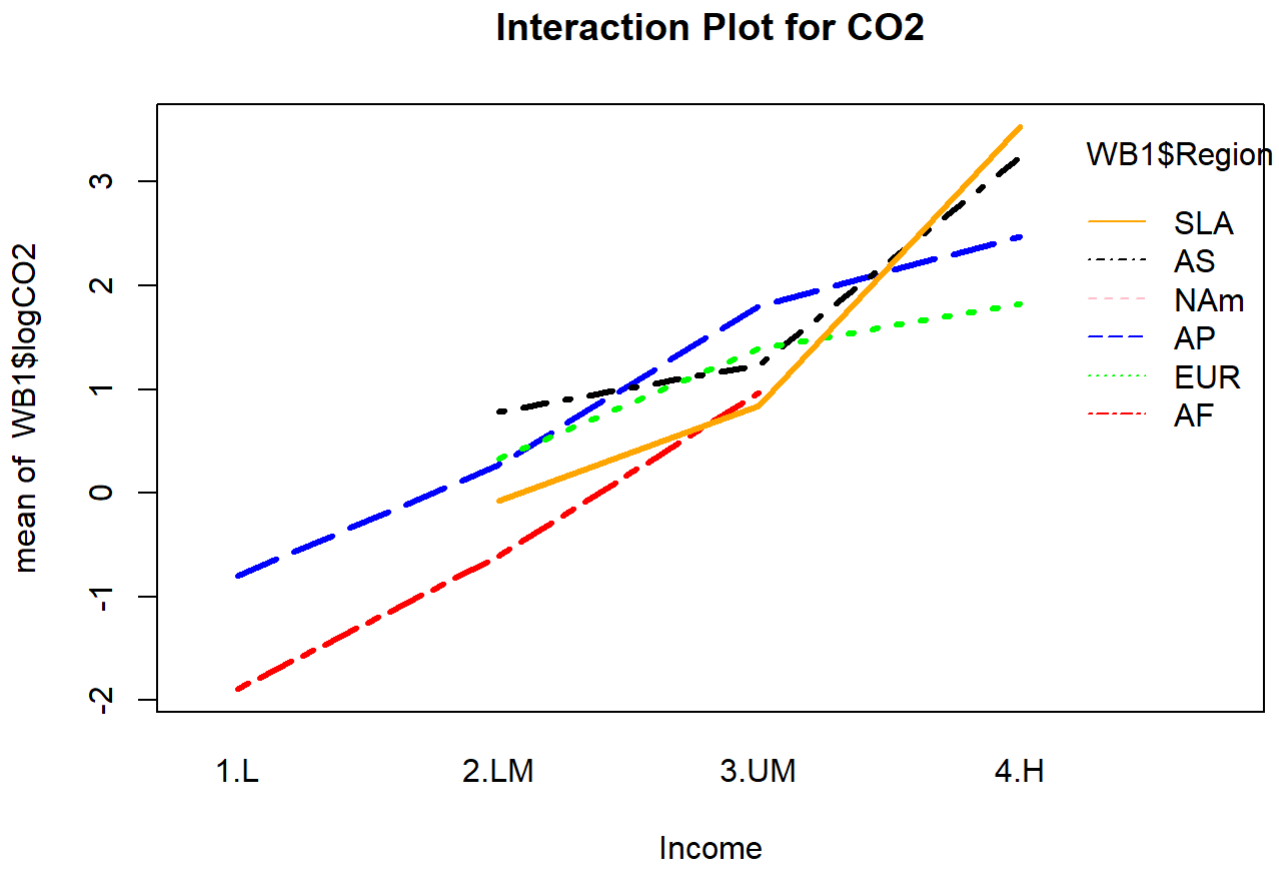
-The PCA plot seems more “visibly” divided than the discriminant analysis plot, this makes sense because we showed earlier that our data was good for PCA.

-Also we clusetered on countreis, not on variables, which can sometimes make it hard to tell why certain things are grouped together. but since PCA and factor is more focused on the variables we looked at grouping by country for this instead.

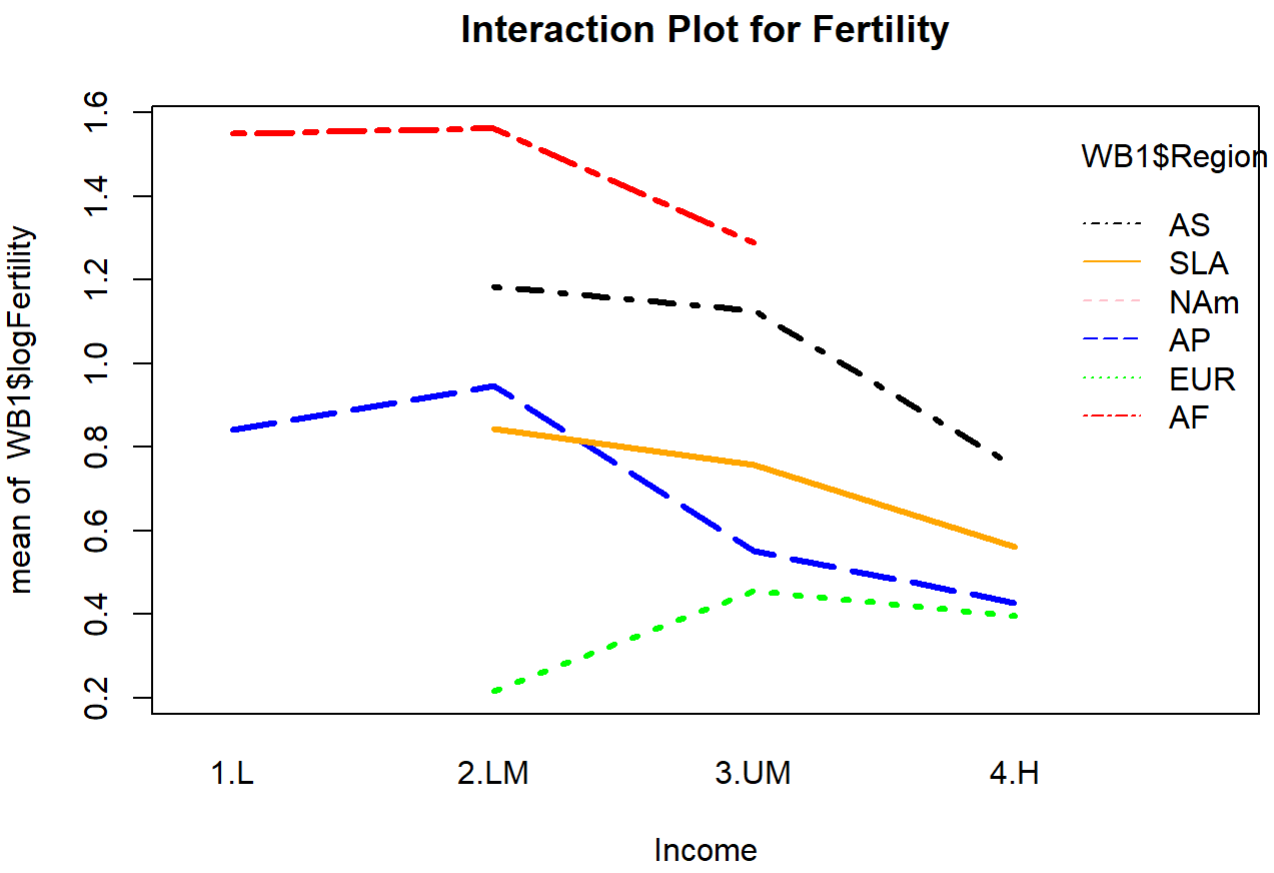
for manova the variables are:

CleanCook (percent access to clean cooking fuel) logCO2 (CO2 output) logFertility logEnergy (energy output) And the categorical predictors are IncomeGrp (1.L = Low, 2.LM = Lower Middle, 3.UM = Upper Middle, 4.H = High) and Region (NAm = North America, SLA = South/Latin America, AS = Arab States, EUR = Europe, AF = Africa, AP = Asia/Pacific)

```
#MANOVA

#example of an interaction plot, we will use fertility for the example.

interaction.plot(WB1$IncomeGrp, WB1$Region, WB1$logCO2, lwd=3,col=c("red","blue","black", "gree
n", "pink", "orange"),xlab="Income",main="Interaction Plot for CO2")
```




```
interaction.plot(WB1$IncomeGrp, WB1$Region, WB1$logFertility, lwd=3,col=c("red","blue","black",
"green", "pink", "orange"),xlab="Income",main="Interaction Plot for Fertility")
```



Fertility Fertility decreases as income increases, but it does differ by region. Africa and Arab States have higher fertility rates, while Asia and South America have moderate fertility rates, and Europe has lower fertility rates.

after manova, don't forget to comment - income, region, interaction are all significatn except for fertility - makes sense bc probably the region already served to explain significatn variability.

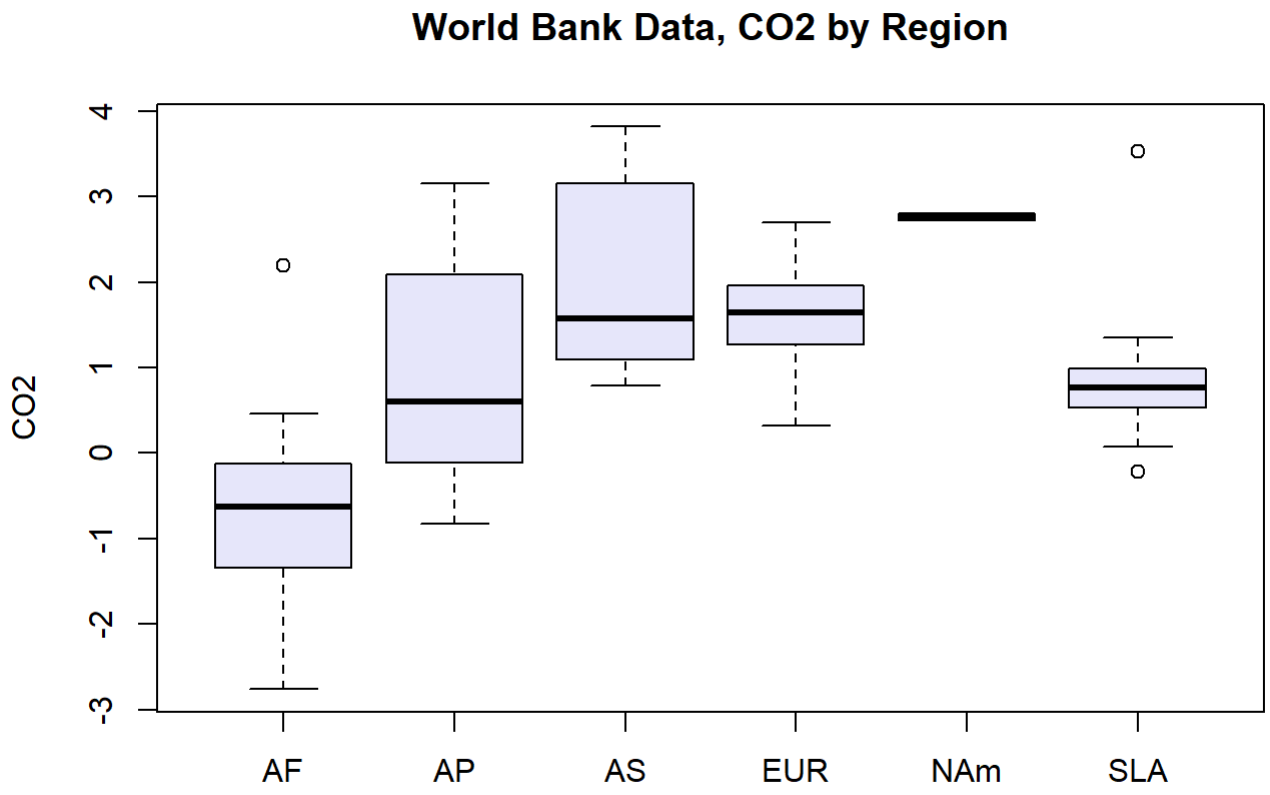
```
mod1 <- manova(as.matrix(WB1[ , c(3:6)]) ~ WB5$IncomeGrp + WB1$Region + WB1$IncomeGrp*WB1$Regio
n)
summary.aov(mod1)
```

```
##  Response CleanCook :
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## WB5$IncomeGrp             3  48000   16000  120.44 < 2e-16 ***
## WB1$Region                5   7724    1545   11.63 2.7e-07 ***
## WB1$Region:WB1$IncomeGrp  8   3529     441    3.32  0.0045 **
## Residuals                46   6111     133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logCO2 :
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## WB5$IncomeGrp             3   92.1   30.70  151.57 < 2e-16 ***
## WB1$Region                5    7.5    1.51    7.44 3.5e-05 ***
## WB1$Region:WB1$IncomeGrp  8    6.4    0.80    3.95  0.0013 **
## Residuals                46    9.3    0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logFertility :
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## WB5$IncomeGrp             3   4.64   1.547   44.94 1.0e-13 ***
## WB1$Region                5   4.34   0.868   25.21 3.8e-12 ***
## WB1$Region:WB1$IncomeGrp  8   0.48   0.060    1.73    0.12
## Residuals                46   1.58   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response logEnergy :
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## WB5$IncomeGrp             3   50.2   16.73  113.43 < 2e-16 ***
## WB1$Region                5    2.6    0.51    3.48  0.0095 **
## WB1$Region:WB1$IncomeGrp  8    7.1    0.89    6.01 2.8e-05 ***
## Residuals                46    6.8    0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.manova(mod1, test = "Wilks")
```

```
##                          Df  Wilks approx F num Df den Df  Pr(>F)
## WB5$IncomeGrp             3 0.0217    30.96     12    114 < 2e-16 ***
## WB1$Region                5 0.0790     8.25     20    144 1.4e-15 ***
## WB1$Region:WB1$IncomeGrp  8 0.1642     3.16     32    160 9.4e-07 ***
## Residuals                46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

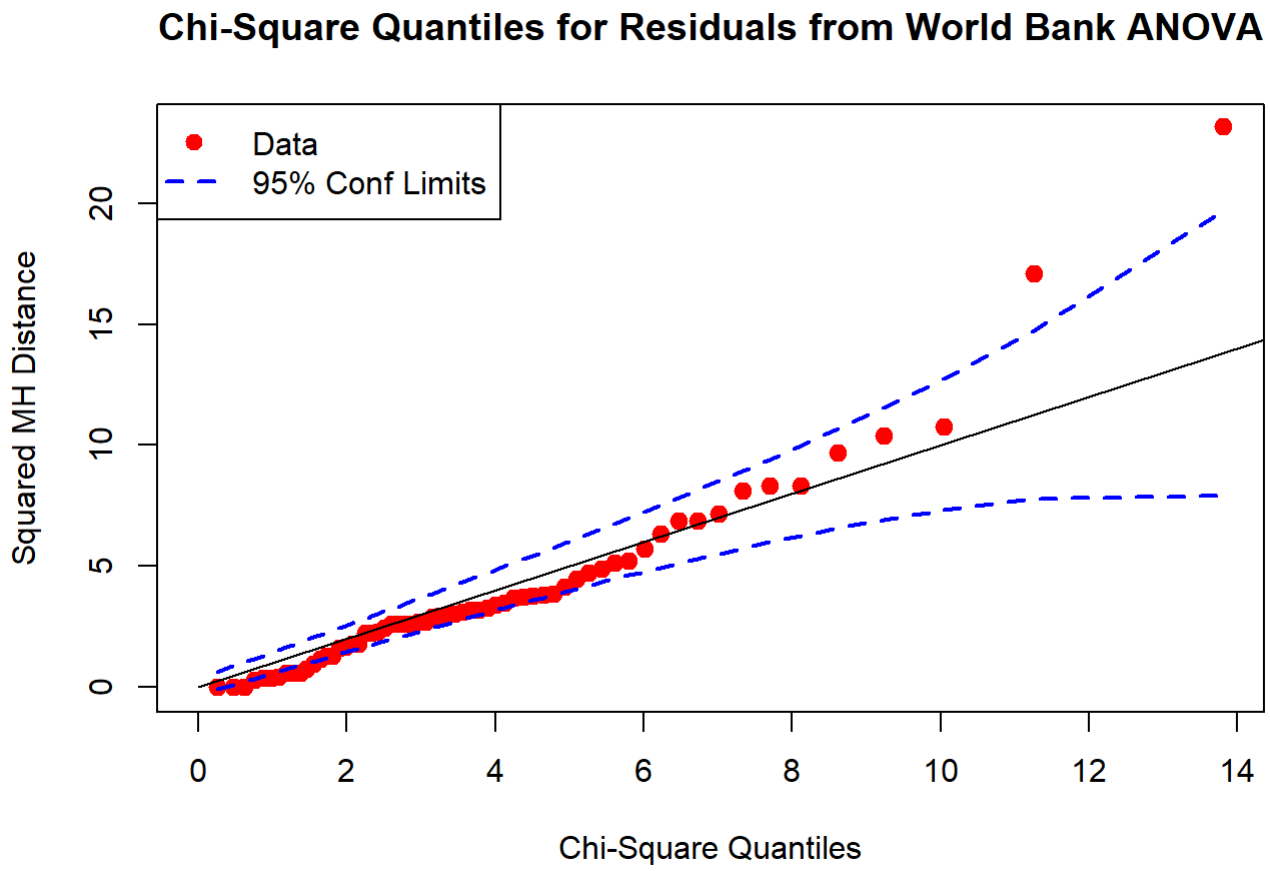
```
boxplot(logCO2 ~ Region, data = WB1, col = "lavender", ylab = "CO2", main = "World Bank Data, CO
2 by Region")
```




```
WB1aov <- lm(logCO2 ~ Region, data = WB1)
contrast1 <- contrast(WB1aov, list(Region = c("SLA", "AF", "AS", "NAm", "EUR")),list(Region = "A
P"),type='average')
print(contrast1,X=TRUE)
```

```
## lm model parameter contrast
##
##   Contrast    S.E.    Lower  Upper   t df Pr(>|t|)
## 1  0.31746 0.35138 -0.38617 1.0211 0.9 57   0.3701
##
## Contrast coefficients:
##   (Intercept) RegionAP RegionAS RegionEUR RegionNAm RegionSLA
## 1           0       -1      0.2       0.2       0.2       0.2
```

```
source("http://www.reuningscherer.net/STAT660/R/CSQPlot.r.txt")
CSQPlot(mod1$residuals, label = "Residuals from World Bank ANOVA")
```



Explain north america in boxplot. residual chi sq look good.

overall: region and income are significant predictors of the other continuous variables, as is the interaction. except for fertility. here's why.

when we did contrasts, we contrasted Asia as the constant against the other regions. the contrast was not statistically significant. we think it's because asia mean is close to group mean and also it has the largest spread (excluding outliers).