# Data Cleaning & Visualization

Much of the clenaing was done in Python. It consisted of dropping race, religion, sex, village names, & village number data We decided to focus only on the industry categories.

### Install the R Required Packages

#install.packages("MVA")  
#install.packages("psych")  
library(psych)

## Warning: package 'psych' was built under R version 3.6.1

library(MVA)

## Warning: package 'MVA' was built under R version 3.6.1

## Loading required package: HSAUR2

## Warning: package 'HSAUR2' was built under R version 3.6.1

## Loading required package: tools

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(mclust)

## Warning: package 'mclust' was built under R version 3.6.1

## Package 'mclust' version 5.4.5  
## Type 'citation("mclust")' for citing this R package in publications.

##   
## Attaching package: 'mclust'

## The following object is masked from 'package:psych':  
##   
## sim

### Read in Data

options( warn = -1, width = 160 )  
census <- read.csv("https://raw.githubusercontent.com/H4L3ST0RM/ISQS-6350-Group3-Project/master/Guyana/data/cleaned\_census\_data2.csv")  
#census\_perc <- read.csv("../data/cleaned\_census\_perc\_data2.csv")  
focus <- c('ID','Region','Agriculture', 'Construction', 'IT', 'Finance', 'Education', 'Manufacturing','Mining', 'Entertainment', 'Health','Administrative','Population')  
census$ID <- seq.int(nrow(census))  
mydata <- census[,focus]  
myindustries <- mydata[,3:12]  
head(mydata)

## ID Region Agriculture Construction IT Finance Education Manufacturing Mining Entertainment Health Administrative Population  
## 1 1 1 28 2 0 0 2 0 1 0 0 1 177  
## 2 2 1 19 3 0 0 2 1 4 0 1 1 203  
## 3 3 1 43 33 1 0 34 4 56 0 17 7 1254  
## 4 4 1 13 15 2 0 31 5 16 0 19 3 569  
## 5 5 1 29 15 0 1 20 7 52 0 18 5 935  
## 6 6 1 24 19 1 2 22 5 44 0 10 4 723

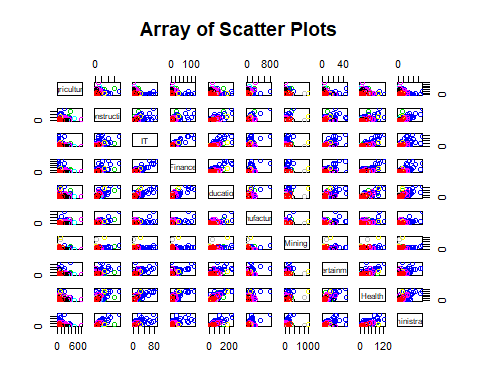
options(digits = 3)

### Correlation Matrix

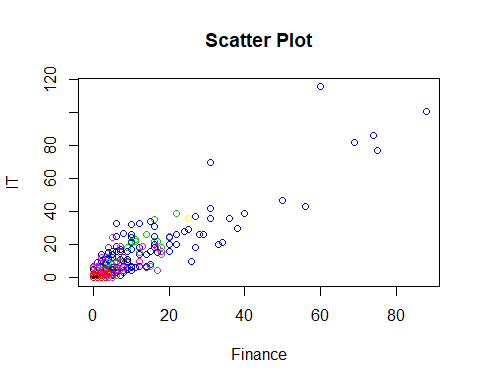
cor <- cor(mydata)  
cor

## ID Region Agriculture Construction IT Finance Education Manufacturing Mining Entertainment Health Administrative Population  
## ID 1.0000 0.9617 0.0262 -0.0847 -0.0520 -0.0694 -0.0516 -0.0668 -0.0318 -0.0287 -0.0423 -0.0468 -0.0369  
## Region 0.9617 1.0000 0.0153 -0.0968 -0.0759 -0.0886 -0.0687 -0.0904 -0.0294 -0.0525 -0.0638 -0.0674 -0.0470  
## Agriculture 0.0262 0.0153 1.0000 0.4034 0.1203 0.1812 0.3743 0.4991 0.1481 0.1540 0.1878 0.2148 0.4909  
## Construction -0.0847 -0.0968 0.4034 1.0000 0.6879 0.7490 0.8455 0.8502 0.4527 0.6534 0.7315 0.8449 0.9137  
## IT -0.0520 -0.0759 0.1203 0.6879 1.0000 0.9133 0.8001 0.6655 0.3642 0.8170 0.8684 0.7830 0.7649  
## Finance -0.0694 -0.0886 0.1812 0.7490 0.9133 1.0000 0.8096 0.7427 0.3528 0.7637 0.8372 0.7667 0.7963  
## Education -0.0516 -0.0687 0.3743 0.8455 0.8001 0.8096 1.0000 0.7581 0.5208 0.7406 0.8724 0.8586 0.9056  
## Manufacturing -0.0668 -0.0904 0.4991 0.8502 0.6655 0.7427 0.7581 1.0000 0.3413 0.5939 0.6636 0.7385 0.8722  
## Mining -0.0318 -0.0294 0.1481 0.4527 0.3642 0.3528 0.5208 0.3413 1.0000 0.4019 0.4833 0.3860 0.5686  
## Entertainment -0.0287 -0.0525 0.1540 0.6534 0.8170 0.7637 0.7406 0.5939 0.4019 1.0000 0.8057 0.7614 0.7287  
## Health -0.0423 -0.0638 0.1878 0.7315 0.8684 0.8372 0.8724 0.6636 0.4833 0.8057 1.0000 0.8302 0.8245  
## Administrative -0.0468 -0.0674 0.2148 0.8449 0.7830 0.7667 0.8586 0.7385 0.3860 0.7614 0.8302 1.0000 0.8447  
## Population -0.0369 -0.0470 0.4909 0.9137 0.7649 0.7963 0.9056 0.8722 0.5686 0.7287 0.8245 0.8447 1.0000

plot(myindustries, col=mydata$Region, main = "Array of Scatter Plots")

 Notice that, with the exception of agriculture, all of the industries appear to be positively correlated with population.

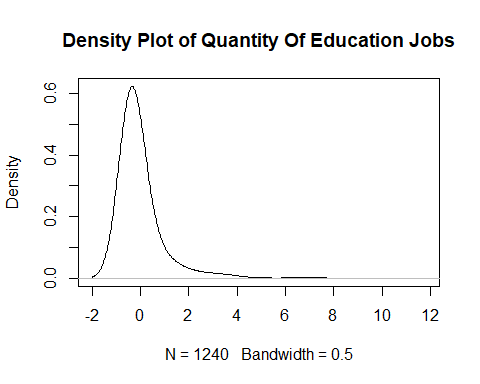
bvbox(myindustries[,c('IT','Finance')],type = "p",  
 xlab = "Finance",  
 ylab = "IT",  
 pch = 1,  
 col = mydata$Region,  
 main = "Scatter Plot")

 outliers = 718, 587, 775, 654, 523, 582, 738, 219, 953 460, 219, 738, 582, 1018, 967, 1028, 963, 951

### Variable Distributions

Most all of the percentage variables are normally distributed.

plot(density(scale(myindustries$Education), bw = .5, kernel ="gaussian"), main="Density Plot of Quantity Of Education Jobs")



All of the industries, when scaled have an almost normal distribution. They are all right skewed. This is likely due to some outliers.

# Dimensionality Reduction Analysis

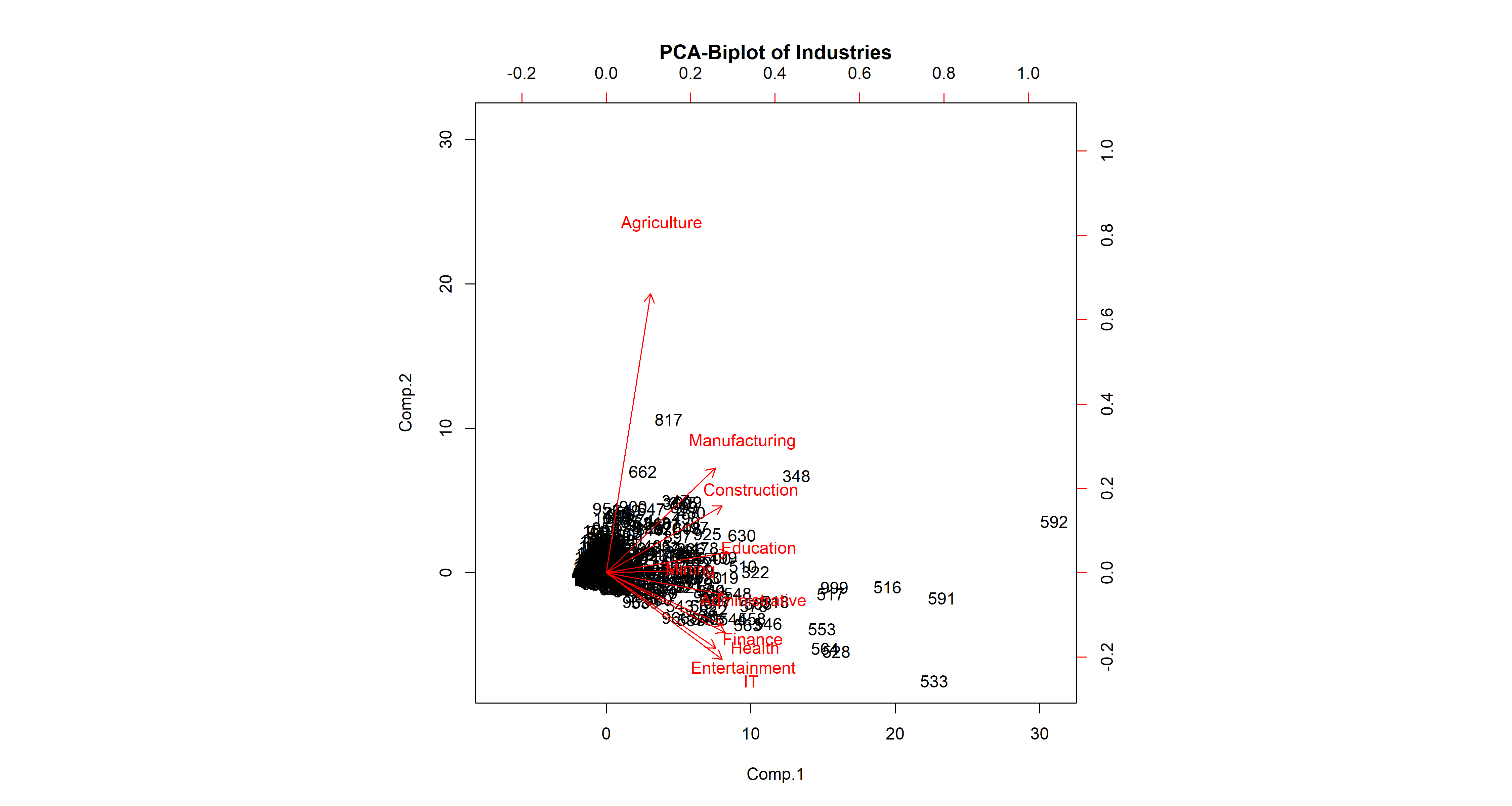
Apply dimension reduction analysis correctly and discuss the motivation behind that. Also provide interesting insights into the results.

### Principal Component Analysis

census\_pca <- princomp(myindustries, cor = T)  
  
summary(census\_pca, cor=T, loadings = T)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Standard deviation 2.605 1.069 0.895 0.6287 0.5348 0.4740 0.3450 0.31106 0.28539 0.25585  
## Proportion of Variance 0.679 0.114 0.080 0.0395 0.0286 0.0225 0.0119 0.00968 0.00814 0.00655  
## Cumulative Proportion 0.679 0.793 0.873 0.9127 0.9413 0.9637 0.9756 0.98531 0.99345 1.00000  
##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10  
## Agriculture 0.131 0.827 0.481 0.162 0.122   
## Construction 0.343 0.198 -0.488 0.109 -0.114 0.482 -0.308 0.430 -0.262   
## IT 0.344 -0.257 0.155 0.256 -0.310 0.125 0.449 -0.636   
## Finance 0.347 -0.157 0.188 -0.538 0.332 0.638   
## Education 0.361 0.159 0.440 0.124 -0.338 -0.706 -0.115   
## Manufacturing 0.323 0.311 0.163 -0.345 -0.355 -0.382 -0.574 -0.216   
## Mining 0.200 -0.946 -0.163 -0.104 0.157   
## Entertainment 0.325 -0.224 0.445 0.387 -0.664 -0.201   
## Health 0.353 -0.177 0.196 0.407 -0.532 -0.319 0.494   
## Administrative 0.349 -0.318 0.511 0.109 0.627 0.305

biplot(census\_pca$scores[,1:2], census\_pca$loadings[,1:2],col=c("black", "red"), cex = 1, main="PCA-Biplot of Industries")

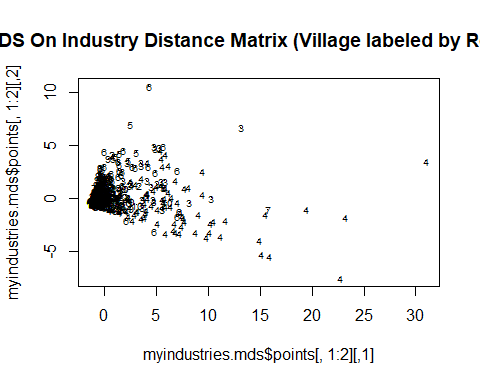
 Agriculture dominates component 2, while the rest of the industries are pretty evenlty spread through component 1. Also of note, finance, IT, Health, Entertainment, and Health are negative on component 2, all of which would theoretically be indicative of a more prosperous, urban society. ### a) Create a scaled distance matrix for observation.

census.mds = cmdscale(dist(scale(myindustries)), eig = T)   
(cumsum(census.mds$eig)/sum(census.mds$eig))[1:10]

## [1] 0.679 0.793 0.873 0.913 0.941 0.964 0.976 0.985 0.993 1.000

### Performing Graphical MDS on Distance Matrix

myindustries.mds = cmdscale(dist(scale(myindustries)), eig = T)   
   
eignv <- abs(myindustries.mds$eig)   
   
plot(myindustries.mds$points[,1:2], pch = ".", main="MDS On Industry Distance Matrix (Village labeled by Region)")  
text(myindustries.mds$points[,1:2], labels = mydata$Region, cex = 0.6)



Notice that region 4 contains nearly all of the outliers. Region 4 contains the largest city and capital Georgetown which would contain the greatest diversity of industry as compared to other regions as other regions are mainly rainforest or rural farming.

### Convert Correlation Matrix to Distance

Below we are getting the distances between each variable.

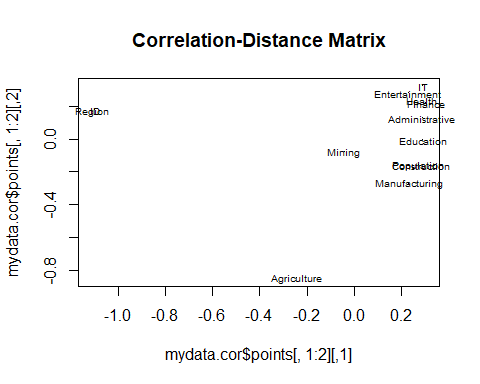
mydata.cor = cmdscale(cor2dist(cor(mydata)),eig=T)  
mydata.cor

## $points  
## [,1] [,2]  
## ID -1.0965 0.17706  
## Region -1.1107 0.17105  
## Agriculture -0.2422 -0.84986  
## Construction 0.2834 -0.16280  
## IT 0.2923 0.32125  
## Finance 0.3050 0.21810  
## Education 0.2931 -0.00738  
## Manufacturing 0.2324 -0.27140  
## Mining -0.0409 -0.07919  
## Entertainment 0.2308 0.28021  
## Health 0.2897 0.23507  
## Administrative 0.2899 0.12625  
## Population 0.2738 -0.15836  
##   
## $eig  
## [1] 3.19e+00 1.21e+00 8.27e-01 4.21e-01 2.87e-01 2.27e-01 1.21e-01 1.09e-01 8.31e-02 6.84e-02 5.07e-02 3.63e-02 1.30e-16  
##   
## $x  
## NULL  
##   
## $ac  
## [1] 0  
##   
## $GOF  
## [1] 0.664 0.664

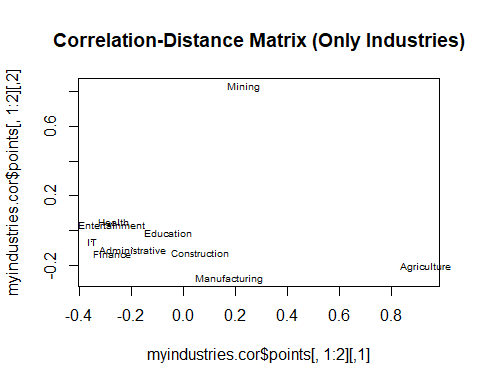
myindustries.cor = cmdscale(cor2dist(cor(myindustries)),eig=T)  
myindustries.cor

## $points  
## [,1] [,2]  
## Agriculture 0.9337 -0.20669  
## Construction 0.0654 -0.12622  
## IT -0.3502 -0.06365  
## Finance -0.2729 -0.13231  
## Education -0.0581 -0.00999  
## Manufacturing 0.1764 -0.27280  
## Mining 0.2343 0.82797  
## Entertainment -0.2723 0.03679  
## Health -0.2644 0.05196  
## Administrative -0.1919 -0.10506  
##   
## $eig  
## [1] 1.34e+00 8.55e-01 4.39e-01 2.87e-01 2.28e-01 1.20e-01 1.11e-01 8.33e-02 6.62e-02 -5.66e-17  
##   
## $x  
## NULL  
##   
## $ac  
## [1] 0  
##   
## $GOF  
## [1] 0.623 0.623

plot(mydata.cor$points[,1:2], pch = ".",main="Correlation-Distance Matrix")  
par(mar=c(1,1,1,1))  
text(mydata.cor$points[,1:2], labels = colnames(mydata), cex = 0.6)



plot(myindustries.cor$points[,1:2], pch = ".", main="Correlation-Distance Matrix (Only Industries)")  
par(mar=c(1,1,1,1))  
text(myindustries.cor$points[,1:2], labels = colnames(myindustries), cex = 0.6)



Above you can see how the various industires relate to one another. The main outlier, Agriculture, was the first industry to the country and region location is largely unrelated to subsequent industry. Family farms were established under indentured servitude where land was available under British rule and those farms continue to this day under private ownership.

Agriculture and Mining, the two largest outliers, are naturally going to be less dependent on the other industries, and more dependent on geographic location/resource availability.

In the chart above that includes population, it can be seen that construction & population are right on top of one another. This makes sense, since the amount of construction projects should correlate with the populatoin size.

Also notice how closely construction and manufacturing are. This is an interesting grouping, seeing how they are both classic blue collar jobs.

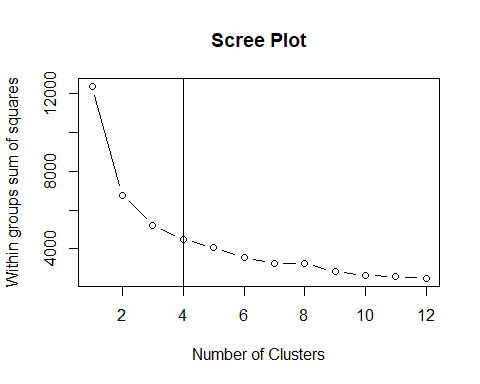
Health & Entertainment is an interesting grouping. This could possibly relate to economic prosperity. It could be suspected that more economically proserpous villages will have more healthcare and entertainment jobs.

Finance & Administrative grouping is also interesting.

Education is also government supported. Schools are established everywhere, including indigenous rainforest regions where many industries do not exist. Which indicates why it is .

# Cluster Analysis

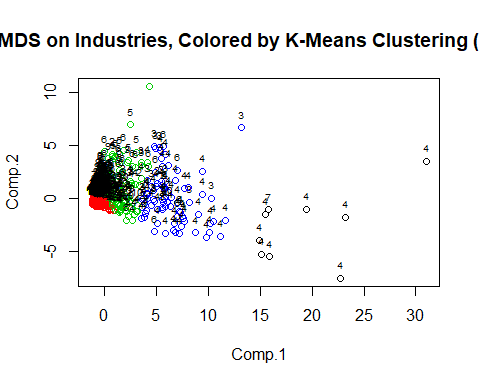
myindustries.s <- scale(myindustries)  
  
  
plot.wgss = function(mydata, maxc) {  
 wss = numeric(maxc)  
 for (i in 1:maxc)  
 wss[i] = kmeans(mydata, centers=i, nstart = 10)$tot.withinss  
 plot(1:maxc, wss, type="b", xlab="Number of Clusters",  
 ylab="Within groups sum of squares", main="Scree Plot")  
}   
plot.wgss(myindustries.s, 12) # Elbow test.  
abline(v=4)

 ### K-Means Clustering

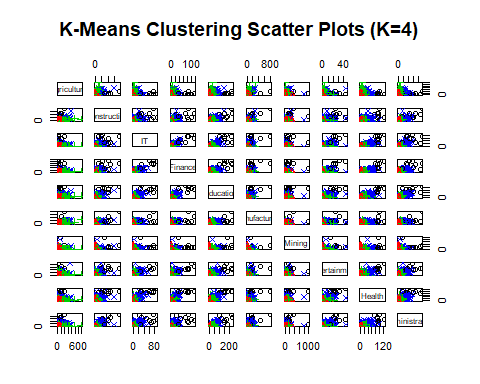
km <- kmeans(myindustries.s, centers = 4, nstart = 10)  
table(km$cluster)

##   
## 1 2 3 4   
## 9 971 181 79

pca <- princomp(myindustries.s)  
plot(pca$scores[,1:2], col=km$cluster, main="MDS on Industries, Colored by K-Means Clustering (K=4)")  
text(pca$scores[,1:2], labels = mydata$Region, cex = 0.6, pos=3)



plot(myindustries, col=km$cluster, pch = km$cluster,main="K-Means Clustering Scatter Plots (K=4)")



km$centers

## Agriculture Construction IT Finance Education Manufacturing Mining Entertainment Health Administrative  
## 1 0.419 5.978 7.7118 7.998 6.894 5.173 4.974 6.271 6.711 6.314  
## 2 -0.303 -0.311 -0.2537 -0.267 -0.346 -0.310 -0.166 -0.266 -0.314 -0.304  
## 3 1.219 0.380 0.0996 0.189 0.506 0.528 0.191 0.171 0.339 0.224  
## 4 0.877 2.268 2.0117 1.933 2.309 2.008 1.035 2.165 2.320 2.500

table(km$cluster,mydata$Region)

##   
## 1 2 3 4 5 6 7 8 9 10  
## 1 0 0 0 8 0 0 1 0 0 0  
## 2 63 132 186 84 157 131 44 29 48 97  
## 3 4 19 19 52 16 49 2 1 13 6  
## 4 0 0 16 54 0 7 0 1 0 1

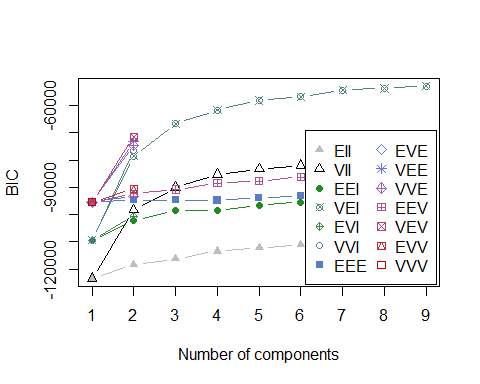
### Model Clustering

mc <- Mclust(myindustries)  
#table(mc$classification)

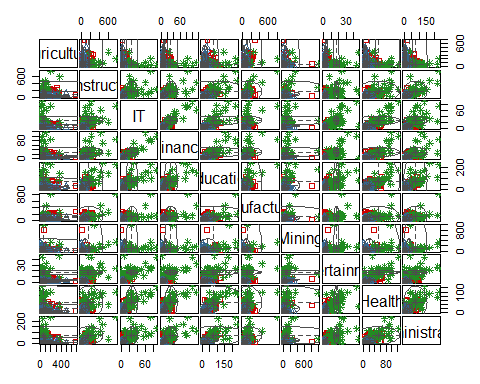
table(mc$classification, mydata$Region)

##   
## 1 2 3 4 5 6 7 8 9 10  
## 1 9 32 36 12 22 22 11 1 3 21  
## 2 0 3 22 60 3 13 0 1 1 5  
## 3 24 24 49 10 43 19 13 16 18 17  
## 4 3 16 23 7 17 14 3 1 2 37  
## 5 0 7 9 5 8 4 0 2 8 3  
## 6 11 18 13 26 23 36 1 0 4 4  
## 7 17 40 45 8 47 57 14 8 19 9  
## 8 0 0 5 40 0 3 1 0 1 0  
## 9 3 11 19 30 10 19 4 2 5 8

plot(mc, what = "BIC")



plot(mc, what = "classification")



The cluster analyses above came out as being quite volatile. Industry appears to be a poor indicator of Region. It would be interesting to see how the cluster analyses mapped out on a geographic map.

# Confirmatory Factor Analysis

Define appropriate latent variables and apply CFA correctly and discuss the motivation. Provide insights into the results

# Conclusion

Present meaningful findings and discuss the pros and cons of the study. Suggest some future works for post analysis.