

# Citations

*The $20 Billion Question for Guyana*, Clifford Krauss, July 20, 2018

<https://www.nytimes.com/2018/07/20/business/energy-environment/the-20-billion-question-for-guyana.html>

*2012 Census Data*, Guyana Bureau of Statistics

<https://statisticsguyana.gov.gy/>

GitHub ISQS-6350-Group3-Project

<https://github.com/H4L3ST0RM/ISQS-6350-Group3-Project/>

Cover template from Inter-American Development Bank - IADB.org

# 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

## Loading required package: HSAUR2

## Loading required package: tools

##   
## 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

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

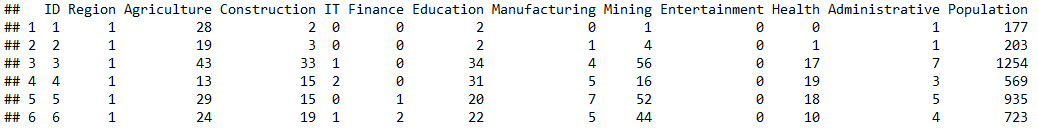
##   
## Attaching package: 'mclust'

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

## Registered S3 methods overwritten by 'huge':  
## method from   
## plot.sim BDgraph  
## print.sim BDgraph

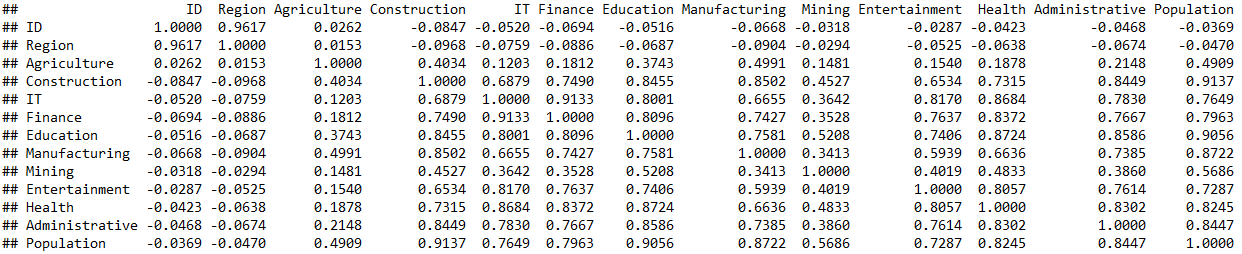
### Read in Data

census <- read.csv("https://raw.githubusercontent.com/H4L3ST0RM/ISQS-6350-Group3-Project/master/Guyana/data/cleaned\_census\_data2.csv") # retrieve from GitHub  
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]  
options(digits = 3)  
# head(mydata) screenshot

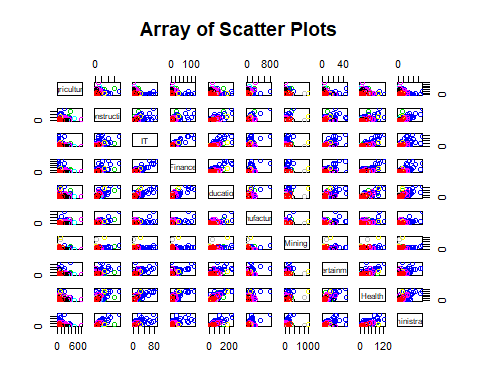


### Correlation Matrix

cor <- cor(mydata)  
# cor screenshot

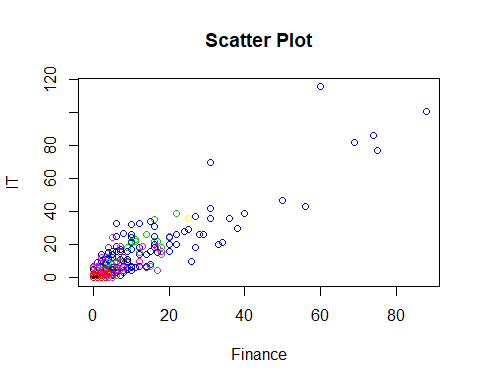


plot(myindustries, col=mydata$Region, main = "Array of Scatter Plots")



Notice that, with the exception of agriculture, all of the industries appear to be positively correlated with population. Mining has a weaker correlation than the other industries.

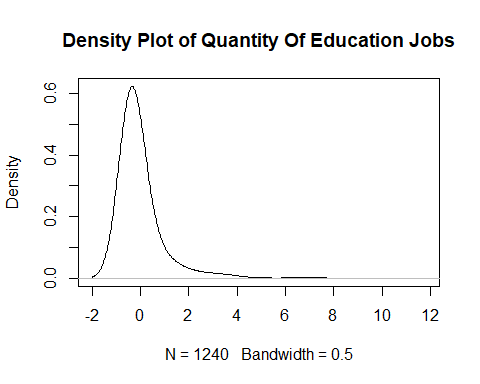
bvbox(myindustries[,c('IT','Finance')],type = "p",  
 xlab = "Finance",  
 ylab = "IT",  
 pch = 1,  
 col = mydata$Region,  
 main = "Scatter Plot")

 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

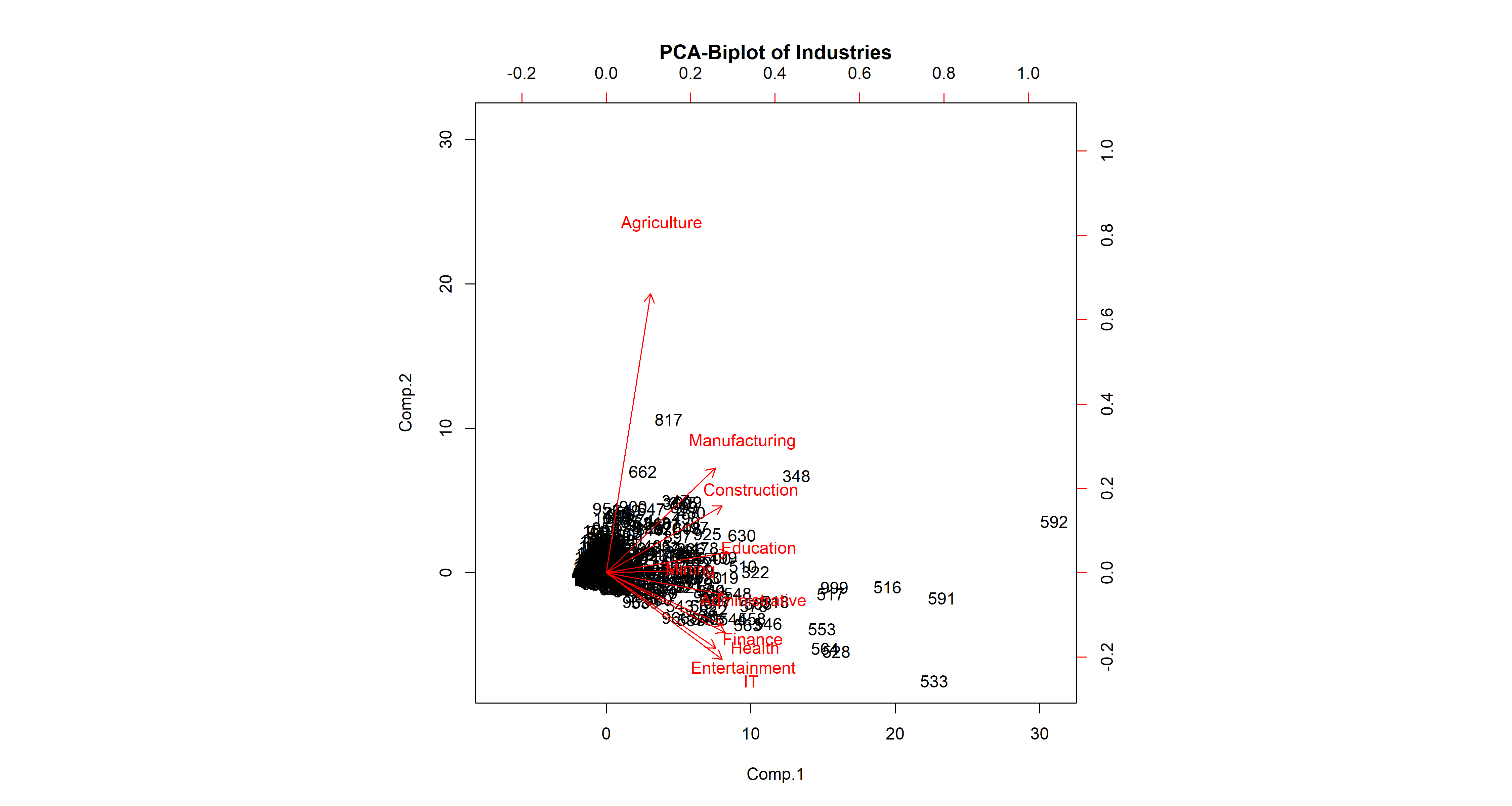
Apply dimension reduction analysis correctly and discuss the motivation behind that. Also provide interesting insights into the results.

### 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")

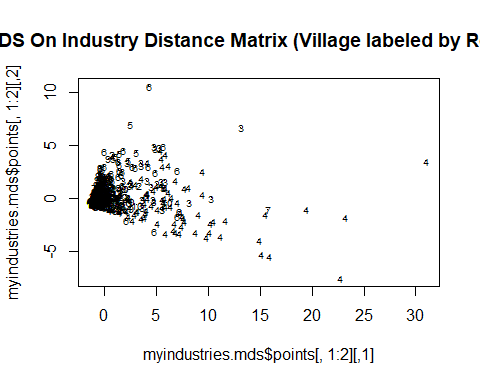
 Agriculture dominates component 2, while the rest of the industries are pretty evenlty spread through component 1. Also of note, finance, IT, Health, Entertainment, and Health are negative on component 2, all of which would theoretically be indicative of a more prosperous, urban society. ### Creating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

### Converting Correlation Matrix to Distance

Below we are getting the distances between each variable.

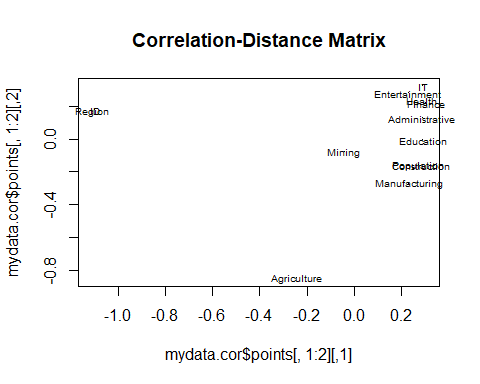
mydata.cor = cmdscale(cor2dist(cor(mydata)),eig=T)  
mydata.cor

## $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 1.96e-16  
##   
## $x  
## NULL  
##   
## $ac  
## [1] 0  
##   
## $GOF  
## [1] 0.664 0.664

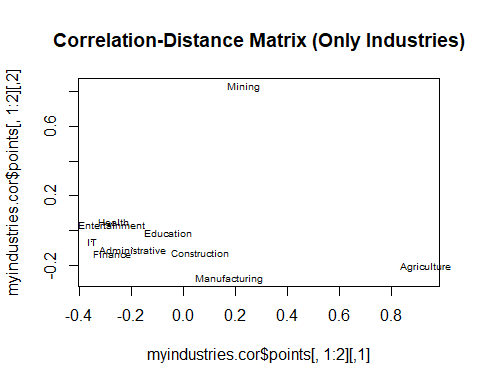
myindustries.cor = cmdscale(cor2dist(cor(myindustries)),eig=T)  
myindustries.cor

## $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 -1.91e-16  
##   
## $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.

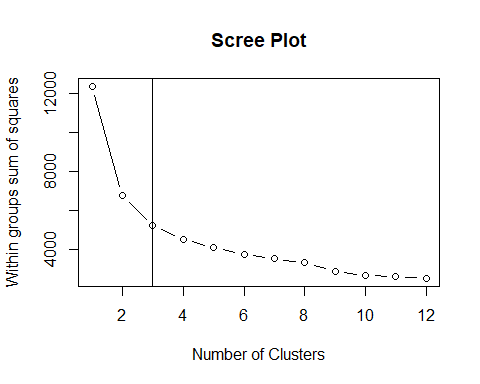
Health & Entertainment is an interesting grouping. This could possibly relate to economic prosperity. It could be suspected that more economically proserpous villages will have more healthcare and entertainment jobs.

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

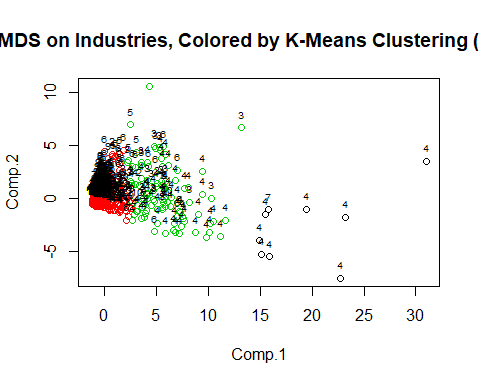
myindustries.s <- scale(myindustries)  
  
  
plot.wgss = function(mydata, maxc) {  
 wss = numeric(maxc)  
 for (i in 1:maxc)  
 wss[i] = kmeans(mydata, centers=i, nstart = 10)$tot.withinss  
 plot(1:maxc, wss, type="b", xlab="Number of Clusters",  
 ylab="Within groups sum of squares", main="Scree Plot")  
}   
plot.wgss(myindustries.s, 12) # Elbow test.  
abline(v=3)

 ### K-Means Clustering

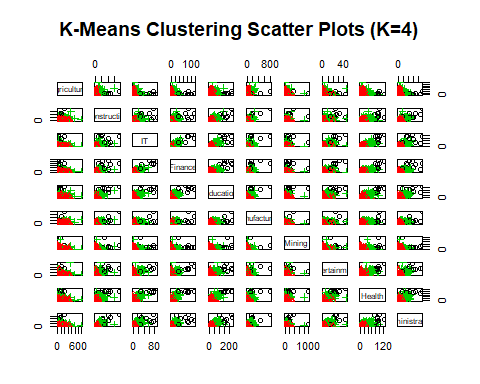
km <- kmeans(myindustries.s, centers = 3, nstart = 10)  
table(km$cluster)

##   
## 1 2 3   
## 9 1129 102

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 7.712 7.998 6.89 5.173 4.974 6.271 6.711 6.314  
## 2 -0.110 -0.231 -0.212 -0.216 -0.24 -0.224 -0.115 -0.216 -0.233 -0.241  
## 3 1.185 2.032 1.670 1.683 2.05 2.025 0.837 1.834 1.992 2.108

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 67 150 201 126 170 176 46 30 61 102  
## 3 0 1 20 64 3 11 0 1 0 2

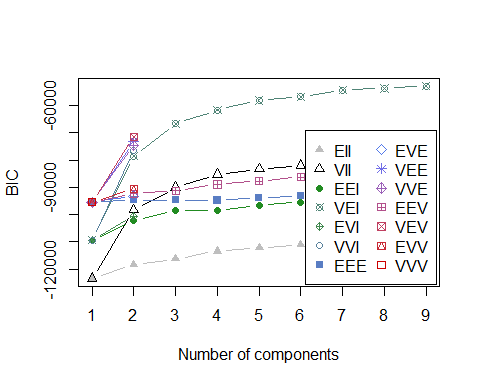
### Model Clustering

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

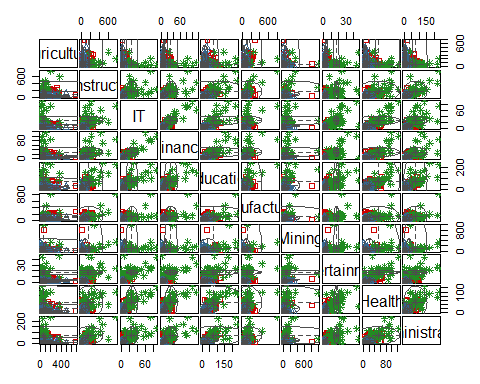
table(mc$classification, mydata$Region)

##   
## 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

Lets first perform exploratory factor analysis on the following industries: \* “IT” \* “Health” \* “Construction” \* “Manufacturing” \* “Entertainment” \* “Finance”

ind <- c("IT", "Health", "Construction", "Manufacturing", "Entertainment", "Finance")  
efa <- factanal(myindustries[,ind], 2)  
print(efa$loadings, cut =0.5)

##   
## Loadings:  
## Factor1 Factor2  
## IT 0.915   
## Health 0.759   
## Construction 0.869   
## Manufacturing 0.787   
## Entertainment 0.733   
## Finance 0.796   
##   
## Factor1 Factor2  
## SS loadings 2.916 2.153  
## Proportion Var 0.486 0.359  
## Cumulative Var 0.486 0.845

Based on the results of the EFA performed above, lets assume their is a latent variable representing white collar jobs, and a latent variable representing blue collar.

Let us model the blue collar jobs as: \* Construction \* Manufacturing

Let us model the white collar jobs as: \* IT \* Health \* Entertainment \* Finance

Below is the model we are using for our CFA Analysis.

White -> IT, lambda1, NA White -> Entertainment, lambda2, NA White -> Health, lambda3, NA Blue -> Construction, lambda4, NA Blue -> Manufacturing, lambda5, NA White -> Finance, lambda6, NA White <-> Blue, rho, NA IT <-> IT, theta1, NA Entertainment <-> Entertainment, theta2, NA Health <-> Health, theta3, NA Construction <-> Construction, theta4, NA Manufacturing <-> Manufacturing, theta5, NA Finance <-> Finance, theta6, NA White <-> White, NA, 1 Blue <-> Blue, NA, 1

industry\_model <- specifyModel(file = "https://raw.githubusercontent.com/H4L3ST0RM/ISQS-6350-Group3-Project/master/Guyana/src/model2.txt")

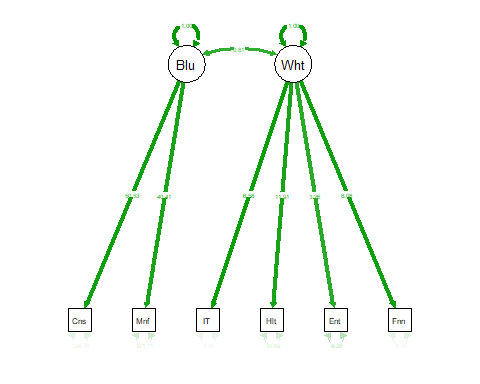
## NOTE: it is generally simpler to use specifyEquations() or cfa()  
## see ?specifyEquations

industry\_sem <- sem(industry\_model, cov(myindustries[,ind]), nrow(myindustries[,ind]))   
  
summary(industry\_sem)

##   
## Model Chisquare = 326 Df = 8 Pr(>Chisq) = 1.23e-65  
## AIC = 352  
## BIC = 269  
##   
## Normalized Residuals  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.201 -0.471 0.000 0.076 0.322 1.610   
##   
## R-square for Endogenous Variables  
## IT Entertainment Health Construction Manufacturing Finance   
## 0.921 0.714 0.822 0.887 0.815 0.889   
##   
## Parameter Estimates  
## Estimate Std Error z value Pr(>|z|)   
## lambda1 6.579 0.1440 45.70 0.00e+00 IT <--- White   
## lambda2 3.263 0.0888 36.73 2.26e-295 Entertainment <--- White   
## lambda3 11.911 0.2885 41.28 0.00e+00 Health <--- White   
## lambda4 50.531 1.1896 42.48 0.00e+00 Construction <--- Blue   
## lambda5 40.409 1.0183 39.68 0.00e+00 Manufacturing <--- Blue   
## lambda6 8.081 0.1828 44.21 0.00e+00 Finance <--- White   
## rho 0.808 0.0118 68.25 0.00e+00 Blue <--> White   
## theta1 3.693 0.2659 13.89 7.40e-44 IT <--> IT   
## theta2 4.276 0.1888 22.64 1.60e-113 Entertainment <--> Entertainment  
## theta3 30.693 1.4890 20.61 2.06e-94 Health <--> Health   
## theta4 324.758 37.7309 8.61 7.49e-18 Construction <--> Construction   
## theta5 371.112 27.1090 13.69 1.17e-42 Manufacturing <--> Manufacturing  
## theta6 8.180 0.4766 17.16 5.15e-66 Finance <--> Finance   
##   
## Iterations = 40

#summary(industry\_sem)

semPaths(industry\_sem, 'std','est')



options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices   
criteria = summary(industry\_sem)   
print('SRMR')

## [1] "SRMR"

print(criteria$SRMR)

## [1] 0.0244

print(criteria$SRMR < 0.05)

## [1] TRUE

print('GFI')

## [1] "GFI"

print(criteria$GFI)

## [1] 0.921

print(criteria$GFI > 0.9)

## [1] TRUE

print('AGFI')

## [1] "AGFI"

print(criteria$AGFI)

## [1] 0.794

print(criteria$AGFI > 0.9)

## [1] FALSE

The standard root means square difference (SRMR) is less than 0.05, and our Goodness-of-fit index (GFI) is greater than 0.9. Both of these measures indicate that we have a decent model. Our adjusted goodness of fit index on the other hand is at 0.79, well below 0.9, indicating this isn’t such a great model.

Given that 2 of the 3 measures indicate this is a good model, I think we can assume the model is adequate for our purposes.

parameters = summary(industry\_sem)  
est = parameters$coef[7,]  
conf.L = est$Estimate - 1.96 \* est$`Std Error`  
conf.U = est$Estimate + 1.96 \* est$`Std Error`  
'Estiate'

## [1] "Estiate"

est$Estimate

## [1] 0.808

'Upper'

## [1] "Upper"

conf.U

## [1] 0.831

'Lower'

## [1] "Lower"

conf.L

## [1] 0.784

Above is the estimated correlation between our two factors. There appears to be a noticeable correlation between our “blue” and “white” factors.

# Conclusion

The purpose of the project was to propose the likely impact of the incoming oil industry by evaluating existing industry to population correlations by region as provided by the national census. From the results, we have identified two major observations:

Agriculture & Mining - Little Correlation  
All Other Industry - High Correlation

Ultimately, there are two groups of industry: production industries (Agriculture, Mining, Oil) and support service industries (Power, Hotels, Finance, etc.). The data proves that as production industries change, those industries impact support service industries, but do not have much impact on other existing production industries. We conclude that the introduction of the oil industry will have the effect of any change in a production industry and will impact almost exclusively support service industries.

These support service industries include Manufacturing, Power, Sewer, Construction, Vehicles, Transportation, Hotels, IT, Finance, Real Estate Professional, Administrative, Pubic Sector, Education, Health, Entertainment, Membership Orgs, Other Goods and Overseas. We suggest that investment in Guyana as a result of the incloming oil industry be directed to one of these industries.