## Introduction

The following project tries to understand trends when it comes to life expectancy and unemployment of people from various countries of the world. The data was retrieved from the World Bank, and contains details about health and population statistics from 2001 - 2015.

## Data Setup

Here is a Summary of the data that we shal use.

data<-read.csv("Health and Population Statistics\_Data.csv")  
dim(data)

## [1] 967 19

str(data)

## 'data.frame': 967 obs. of 19 variables:  
## $ ï..Series.Name: Factor w/ 29 levels "","Adolescent fertility rate (births per 1,000 women ages 15-19)",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Series.Code : Factor w/ 27 levels "","NY.GNP.PCAP.CD",..: 21 21 21 21 21 21 21 21 21 21 ...  
## $ Country.Name : Factor w/ 38 levels "","American Samoa",..: 2 3 4 5 6 7 8 9 10 11 ...  
## $ Country.Code : Factor w/ 38 levels "","ASM","AUS",..: 2 3 4 12 5 6 29 8 9 10 ...  
## $ X2001..YR2001.: Factor w/ 507 levels "","..","0.163782398652776",..: 2 112 175 281 462 271 279 315 241 299 ...  
## $ X2002..YR2002.: Factor w/ 518 levels "","..","0.194323704164731",..: 2 103 170 278 458 264 273 312 207 302 ...  
## $ X2003..YR2003.: Factor w/ 518 levels "","..","0","0.226214022293707",..: 2 106 166 280 457 270 278 307 206 301 ...  
## $ X2004..YR2004.: Factor w/ 511 levels "","..","0.182340379369523",..: 2 112 170 285 455 278 281 309 203 305 ...  
## $ X2005..YR2005.: Factor w/ 510 levels "","..","0.16497408488386",..: 2 100 160 276 395 267 265 298 197 300 ...  
## $ X2006..YR2006.: Factor w/ 516 levels "","..","0.120732265474933",..: 2 115 173 289 405 284 280 292 205 311 ...  
## $ X2007..YR2007.: Factor w/ 503 levels "","..","0.143209918485443",..: 2 104 172 278 391 275 269 280 195 299 ...  
## $ X2008..YR2008.: Factor w/ 505 levels "","..","0.117908308033479",..: 2 105 168 274 390 269 262 275 192 302 ...  
## $ X2009..YR2009.: Factor w/ 507 levels "","..","0","0.100000001490116",..: 2 105 169 274 386 268 264 276 196 297 ...  
## $ X2010..YR2010.: Factor w/ 501 levels "","..","0","0.105035188241282",..: 2 101 162 277 388 266 232 280 197 296 ...  
## $ X2011..YR2011.: Factor w/ 500 levels "","..","0","0.0807537363894052",..: 2 98 162 276 389 269 237 278 205 299 ...  
## $ X2012..YR2012.: Factor w/ 547 levels "","..","0.0921312064771722",..: 2 114 173 300 420 287 251 319 216 322 ...  
## $ X2013..YR2013.: Factor w/ 511 levels "","..","0.114837571175624",..: 2 110 179 297 403 287 246 296 216 316 ...  
## $ X2014..YR2014.: Factor w/ 479 levels "","..","0.131876861464529",..: 2 94 159 288 366 263 219 267 187 287 ...  
## $ X2015..YR2015.: Factor w/ 100 levels "","..","0.5222",..: 2 20 33 70 81 62 48 66 40 68 ...

tail(data)

## ï..Series.Name  
## 962 School enrollment, secondary (% gross)  
## 963   
## 964   
## 965   
## 966 Data from database: Health Nutrition and Population Statistics  
## 967 Last Updated: 12/16/2016  
## Series.Code Country.Name Country.Code X2001..YR2001. X2002..YR2002.  
## 962 SE.SEC.ENRR Vietnam VNM .. ..  
## 963   
## 964   
## 965   
## 966   
## 967   
## X2003..YR2003. X2004..YR2004. X2005..YR2005. X2006..YR2006.  
## 962 .. .. .. ..  
## 963   
## 964   
## 965   
## 966   
## 967   
## X2007..YR2007. X2008..YR2008. X2009..YR2009. X2010..YR2010.  
## 962 .. .. .. ..  
## 963   
## 964   
## 965   
## 966   
## 967   
## X2011..YR2011. X2012..YR2012. X2013..YR2013. X2014..YR2014.  
## 962 .. .. .. ..  
## 963   
## 964   
## 965   
## 966   
## 967   
## X2015..YR2015.  
## 962 ..  
## 963   
## 964   
## 965   
## 966   
## 967

We see that the data after row 962 is not useful, hence we shall remove it. Also, we shall clean the given data to change the column names into identifiable years.

data<-data[1:962,]  
names(data)<-(c("Series","sCode","Country","cCode",2001:2015)) ## Simplifying the names

## Exploratory Data Analysis

### 1. One Variable Analysis

In this section we shall analyze the avearge life expectancy of males, and females, along with the avearage life expectancy irrespective of gender.

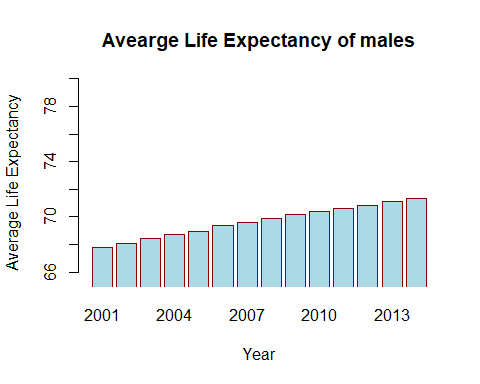
require(ggplot2)

## Loading required package: ggplot2

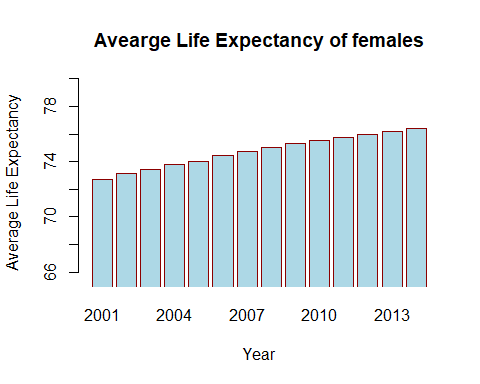
lm <- data[data$sCode == "SP.DYN.LE00.MA.IN",-c(1:4)] ## Life Expectancy of males  
lf <- data[data$sCode == "SP.DYN.LE00.FE.IN",-c(1:4)] ## Life Expectancy of females  
lt <- data[data$sCode == "SP.DYN.LE00.IN",-c(1:3,5:17,19)] ## Life Expectancy, total for 2014  
  
## Finding the Columnwise average:  
avglm <- apply(apply(lm,2,as.numeric),2,mean,na.rm=TRUE)  
avglf <- apply(apply(lf,2,as.numeric),2,mean,na.rm=TRUE)  
lt$`2014`<-as.numeric(as.character(lt$`2014`))  
avglt<-lt$`2014`  
names(avglt)<-lt$cCode  
avglt<-avglt[!is.na(avglt)]

Now we shall plot each of these values:

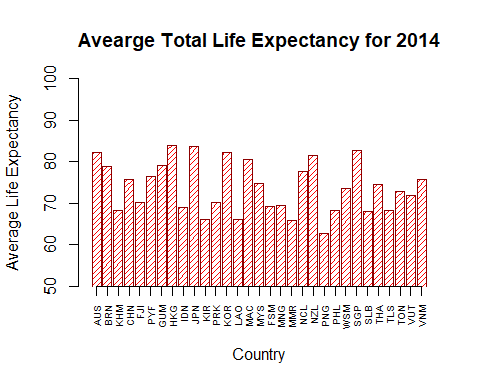
## Graph for life expectancy of males.  
barplot((avglm[1:14]),main = "Avearge Life Expectancy of males",xlab = "Year",ylab = "Average Life Expectancy", col = "lightblue", border = "darkred",ylim = c(65,80),xpd = FALSE)



## Graph for life expectancy of females.  
barplot((avglf[1:14]),main = "Avearge Life Expectancy of females",xlab = "Year",ylab = "Average Life Expectancy", col = "lightblue", border = "darkred",ylim = c(65,80),xpd = FALSE)



## Graph for total life expectancy by country  
b <- barplot(avglt,main = "Avearge Total Life Expectancy for 2014",xlab = "Country",ylab = "Average Life Expectancy", col = "red", border = "darkred",ylim = c(50,100),xpd = FALSE,xaxt="n",density = 25)  
axis(1,b,names(avglt),las=3,cex.axis = 0.6)



The graphs show a similar increase in life expectancy from 2001 to 2014. The life expectancy of males has always been lesser than that of females. Also we see that in 2014, Honk Kong had the highest average life expectancy.

### 2. Two Variable analysis

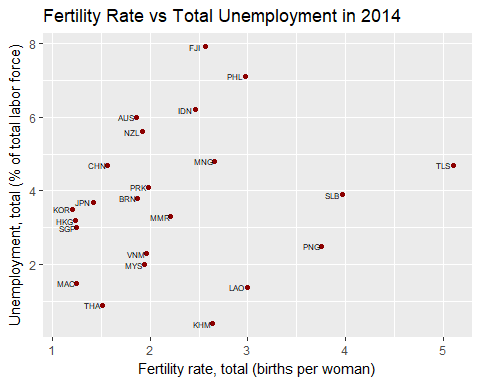
#### a) Total Fertility Rate(ie Births per woman), vs Unemployment rate(ie % of total labor force).

We shall do this for the year 2014, and try to establish a relationship between the two variables. The motive here, is to see whether a high number of Births per woman(a very rough estimate of population), results in more Unemployment.

samp <- data[(data$Series == "Unemployment, total (% of total labor force)" | data$Series == "Fertility rate, total (births per woman)"),-c(2,5:17,19)] ## This gives us the required data for 2014.  
samp[,4]<-as.numeric(as.character(samp[,4]))  
  
samp <- na.omit(samp) ## Remove all countires, the data of which is not available.  
u <- samp[samp$Series=="Unemployment, total (% of total labor force)",2:4] ## Contains Data about unemployment  
names(u)<-c("Country","cCode","Unemployment Value")  
f <- samp[samp$Series == "Fertility rate, total (births per woman)",2:4] ## Contains Data about fertility  
names(f)<-c("Country","cCode","Fertility Value")  
  
## Cleaning the Data in the desired manner  
samp <- merge(f,u,by=c("Country","cCode"))   
samp <- samp[order(samp$`Fertility Value`),]  
rm(f,u)  
print(samp)

## Country cCode Fertility Value Unemployment Value  
## 10 Korea, Rep. KOR 1.205 3.5  
## 6 Hong Kong SAR, China HKG 1.234 3.2  
## 12 Macao SAR, China MAC 1.243 1.5  
## 19 Singapore SGP 1.250 3.0  
## 8 Japan JPN 1.420 3.7  
## 21 Thailand THA 1.512 0.9  
## 4 China CHN 1.562 4.7  
## 1 Australia AUS 1.859 6.0  
## 2 Brunei Darussalam BRN 1.874 3.8  
## 16 New Zealand NZL 1.920 5.6  
## 13 Malaysia MYS 1.944 2.0  
## 23 Vietnam VNM 1.961 2.3  
## 9 Korea, Dem. Peopleâ<U+0080><U+0099>s Rep. PRK 1.979 4.1  
## 15 Myanmar MMR 2.204 3.3  
## 7 Indonesia IDN 2.463 6.2  
## 5 Fiji FJI 2.564 7.9  
## 3 Cambodia KHM 2.635 0.4  
## 14 Mongolia MNG 2.655 4.8  
## 18 Philippines PHL 2.977 7.1  
## 11 Lao PDR LAO 2.991 1.4  
## 17 Papua New Guinea PNG 3.757 2.5  
## 20 Solomon Islands SLB 3.966 3.9  
## 22 Timor-Leste TLS 5.100 4.7

gfu <- ggplot(samp,aes(`Fertility Value`,`Unemployment Value`))+  
 geom\_point(color="darkred")+  
 labs(title = "Fertility Rate vs Total Unemployment in 2014",y="Unemployment, total (% of total labor force)",x="Fertility rate, total (births per woman)")+  
 geom\_text(aes(samp$`Fertility Value`,samp$`Unemployment Value`),label = samp$cCode,size=2.2,nudge\_x = -.1)  
   
print(gfu)



The graph above, reflects the Unemployment in some countries against the fertility rate for the year 2014. This graph will be further analyzed in the upcoming sections.

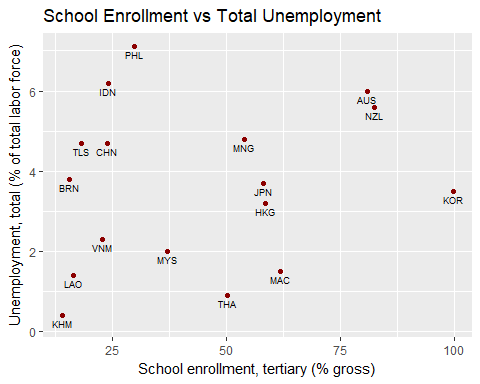
#### b) School enrollment, tertiary (% gross), vs Unemployment rate(ie % of total labor force).

We shall consider Unemployment of the year 2014, while Primary school Enrollment of the year 2010. As the people enrolled in Tertiary school in 2010, would form a part of the working population in 2014. Then, we will try to establish a relationship between the two variables. The motive here, is to see whether primary education, affects Unemployment.

samp2 <- data[data$Series == "Unemployment, total (% of total labor force)",c(1,3,4,18)]  
##We select the column numbers 1,3,4,18 for the required data  
  
names(samp2)<-names(data[,c(1,3,4,14)])  
## The change in name is to facilitate rbind  
  
samp2<-rbind(samp2,data[data$Series == "School enrollment, tertiary (% gross)",c(1,3,4,14)])  
##rbind combines the two sampled data frames into 1  
  
## This gives us the required data .  
  
samp2[,4]<-as.numeric(as.character(samp2[,4]))  
samp2 <- na.omit(samp2) ## Remove all countires, the data of which is not available.  
u <- samp2[samp2$Series=="Unemployment, total (% of total labor force)",2:4] ## Contains Data about unemployment  
names(u)<-c("Country","cCode","Unemployment Value")  
f <- samp2[samp2$Series == "School enrollment, tertiary (% gross)",2:4]   
names(f)<-c("Country","cCode","School Enrollment")  
  
## Cleaning the Data in the desired manner  
samp2 <- merge(f,u,by=c("Country","cCode"))   
samp2 <- samp2[order(samp2$`Unemployment Value`),]  
rm(f,u)  
print(samp2)

## Country cCode School Enrollment Unemployment Value  
## 3 Cambodia KHM 14.06062 0.4  
## 15 Thailand THA 50.20262 0.9  
## 9 Lao PDR LAO 16.35676 1.4  
## 10 Macao SAR, China MAC 61.78807 1.5  
## 11 Malaysia MYS 37.13811 2.0  
## 17 Vietnam VNM 22.68768 2.3  
## 5 Hong Kong SAR, China HKG 58.48681 3.2  
## 8 Korea, Rep. KOR 99.66034 3.5  
## 7 Japan JPN 58.07524 3.7  
## 2 Brunei Darussalam BRN 15.65481 3.8  
## 4 China CHN 23.94792 4.7  
## 16 Timor-Leste TLS 18.14971 4.7  
## 12 Mongolia MNG 53.82303 4.8  
## 13 New Zealand NZL 82.51750 5.6  
## 1 Australia AUS 80.91708 6.0  
## 6 Indonesia IDN 24.19967 6.2  
## 14 Philippines PHL 29.75401 7.1

gsu <- ggplot(samp2,aes(`School Enrollment`,`Unemployment Value`))+  
 geom\_point(color="darkred")+  
 labs(title = "School Enrollment vs Total Unemployment",y="Unemployment, total (% of total labor force)",x="School enrollment, tertiary (% gross)")+  
 geom\_text(aes(samp2$`School Enrollment`,samp2$`Unemployment Value`),label = samp2$cCode,size=2.5,nudge\_y = -.2)  
   
print(gsu)



The graph above, reflects the Unemployment in some countries(in 2014) against tertiary school enrollment(in 2010). This graph will be further analyzed in the upcoming sections.

## Advanced Analysis

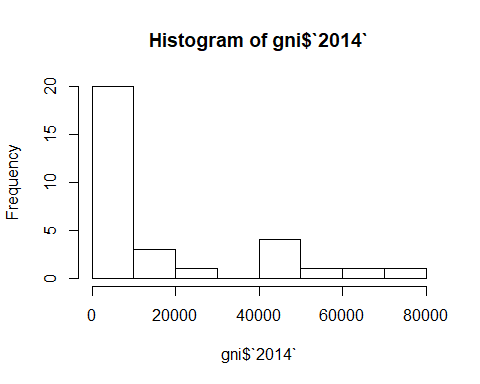
### 1. Clustering

k-means clustering is another simple way of examining and organizing multi-dimensional data. As with hierarchical clustering, this technique is most useful in the early stages of analysis when you're trying to get an understanding of the data, e.g., finding some pattern or relationship between different factors or variables. **We shall try to cluster countries with respect to GNI per capita in the year 2014**

gni<-data[data$Series=="GNI per capita, Atlas method (current US$)",c(3,4,18)]  
## Storing the GNI for 2014, along with Country name and Code  
  
gni[,3]<-as.numeric(as.character(gni[,3]))  
gni<-gni[!is.na(gni[,3]),]  
## Removing Data which is not available

First, we shall try to estimate the number of centers. We shall use histograms as we are analyzing univariable data:

hist(gni$`2014`)



We see that broadly, there are 2 clusters, hence we shall perform kmeans clustering with 2 centers:

gniClust <- kmeans(gni[,3],2,100)  
gniClust$size ## Gives us the size of each cluster

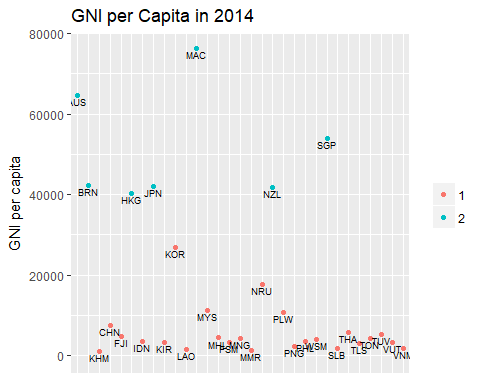
## [1] 24 7

gniClust$centers ## Gives us the 2 centers

## [,1]  
## 1 5688.333  
## 2 51567.143

Now we shall plot the GNI per capita of each country using ggplot:

gniClust$cluster <- as.factor(gniClust$cluster)  
cp <- ggplot(data = gni,mapping = aes(x=gni$Country,y=gni$`2014`))+  
 geom\_point(aes(color = gniClust$cluster))+  
 theme(axis.title.x=element\_blank(),  
 axis.text.x=element\_blank(),  
 axis.ticks.x=element\_blank(),  
 legend.title = element\_blank())+  
 ylab("GNI per capita")+  
 labs(title = "GNI per Capita in 2014")+  
 geom\_text(aes(x=gni$Country,y=gni$`2014`),label = as.character(gni$cCode),size=2.5,size=2.5,nudge\_y = -1500)  
  
cp



Thus we see how the countries are clustered according to the average GNI per capita for 2014.

### 2. Linear regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X. We shall try to perform linear regression on the two bi-variable graphs that we have made so far.

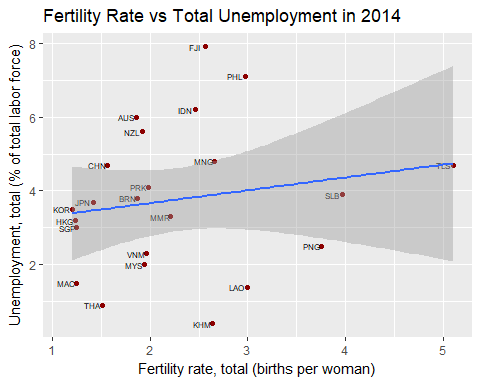
#### a) Total Fertility Rate(ie Births per woman), vs Unemployment rate(ie % of total labor force).

We had built a graph for the assesment of these 2 variables earlier. The data corresponds to 2014. We shall try to fit a Linear Regresion model on this graph, and do some more analysis pertaining to regression.

fit1 <- lm(samp$`Unemployment Value`~samp$`Fertility Value`)  
summary(fit1)

##   
## Call:  
## lm(formula = samp$`Unemployment Value` ~ samp$`Fertility Value`)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.487 -1.500 -0.044 1.046 4.038   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.9705 1.0503 2.828 0.0101 \*  
## samp$`Fertility Value` 0.3478 0.4254 0.817 0.4229   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.966 on 21 degrees of freedom  
## Multiple R-squared: 0.03084, Adjusted R-squared: -0.01532   
## F-statistic: 0.6681 on 1 and 21 DF, p-value: 0.4229

gfu<-gfu+geom\_smooth(method = "lm")  
gfu



The results above can be summarized in the following manner:

* The slope of the regression line is 0.34, that is if all other parameters are constant, for a unit increase in fertility rate, Unemployment increases by 0.34% of the total population, ie, **For the given data, There is an increase in unemployment if there is an increase in population.**
* The p value for the slope is 0.42, which is very high than the standard, hence the above result cannot be generalized, and is insufficient for any prediction.
* The R-squared value is 0.03, that is, Fertility Rate accounts for only 3% of the variability in Unemployment.
* There is an increase in variability of unemployment rate in countries with high fertility rate. This is primarily because the little data that we have for countries with very high fertility rate.

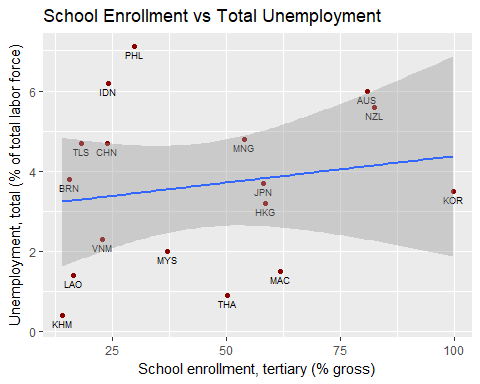
#### b) School enrollment, tertiary (% gross), vs Unemployment rate(ie % of total labor force)

We had built a graph for the assesment of these 2 variables earlier. The unemployment rate corresponds to 2014, while tertiary school enrollment corresponds to 2010. We shall try to fit a Linear Regresion model on this graph, and do some more analysis pertaining to regression.

fit2 <- lm(samp2$`Unemployment Value`~samp2$`School Enrollment`)  
summary(fit2)

##   
## Call:  
## lm(formula = samp2$`Unemployment Value` ~ samp2$`School Enrollment`)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8399 -1.5450 -0.1219 1.4061 3.6526   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.05396 0.96941 3.150 0.0066 \*\*  
## samp2$`School Enrollment` 0.01322 0.01899 0.696 0.4969   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.031 on 15 degrees of freedom  
## Multiple R-squared: 0.03131, Adjusted R-squared: -0.03327   
## F-statistic: 0.4848 on 1 and 15 DF, p-value: 0.4969

gsu<-gsu+geom\_smooth(method = "lm")  
gsu



The results above can be summarized in the following manner:

* The slope of the regression line is 0.01, that is if all other parameters are constant, for a unit increase in tertiary school enrollment, Unemployment increases by 0.01% of the total population, ie, **There is an increase in unemployment if there is an increase in tertiary school enrollment**
* The p value for the slope is 0.50, which is very high than the standard, hence the above result cannot be generalized, and is insufficient for any prediction.
* The R-squared value is 0.03, that is, Fertility Rate accounts for only 3% of the variability in Unemployment.

## Coclusion

* There has been an increase in average life expectancy from 2001 to 2014.
* The life expectancy of males has always been lesser than that of females.
* In 2014, Honk Kong had the highest average life expectancy.
* Counries can broadly be divided into 2 clusters with respect to their GNI per capita.
* The data is inadequate to make any prediction when it comes to the causes of unemployment in countries.

## Reflections

The analysis of the given data was a very enriching experience. Though there were cetain difficulties which were encountered: The biggest concern was the high percentage of missing data. This lead to less accurate analysis in certain sections, as certain factors with high percentage of missing data had to be ignored. But overall, this project provided a great opportunity to learn new techniques of data analysis, and was ideal for someone who is interested in this domain.