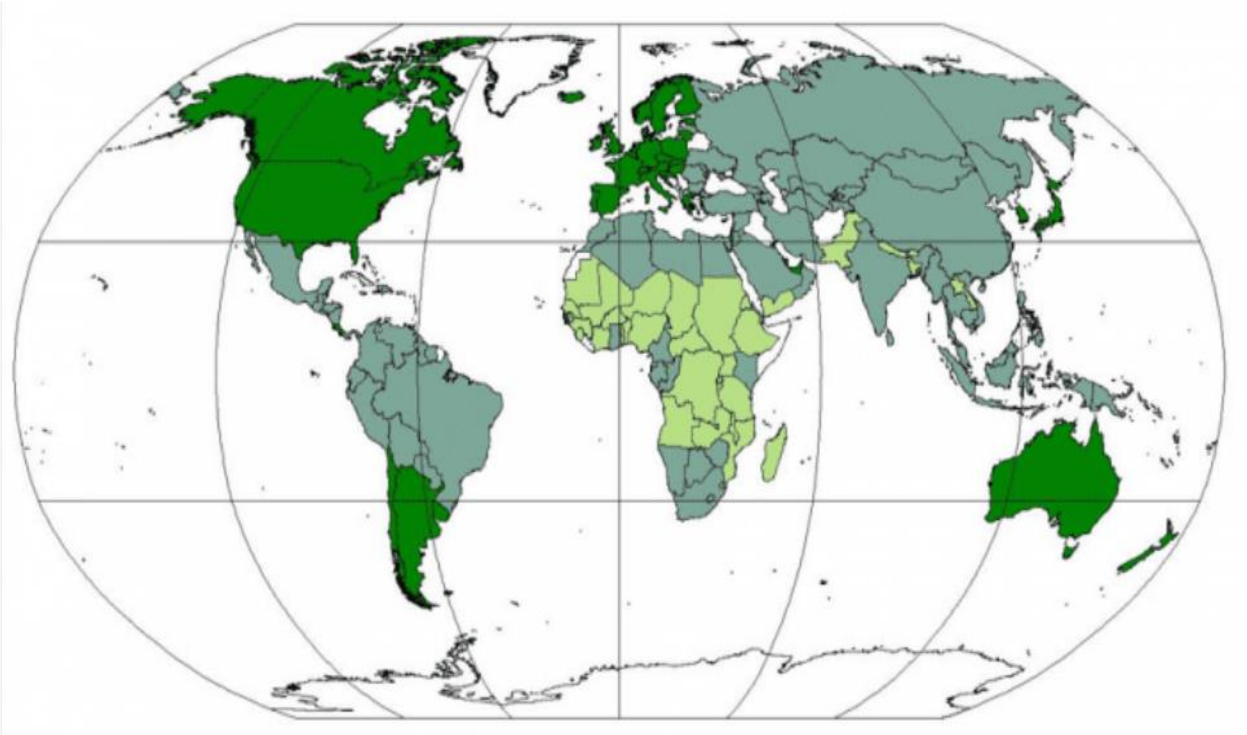
STATE OF THE WORLD 2010

DANA 4840 - Classification II



Team Members

Kiran Anga (100368125)
Priya Yadav (100366834)
Radhika Maini (100340257)
Sai Prasanna (100365191)

Submitted To

Edward Chiu

Table of Contents

Introduction

The data is provided by the World Health Statistics department which is a compilation of 214 world countries for the year 2010.

Data capture various indicators that help in determining the status of country's health in different areas such as – nation health and education expenditure, birth, mortality and life expectancy, national GDP per capita, GDP annual growth, and demographic and socioeconomic statistics.

With the use of clustering, we are trying to group countries basis similar characteristics and label them basis the cluster groups formed.

Aim of the Analysis

The main objective behind the analysis was to understand the nature behind the developed, underdeveloped and developing countries and what measures can be taken by the respective countries and supporting organizations to help them live a better life.

Target Audience

Target audience is the economists and researchers from various parts of the world and WHO because we wanted them to be aware of the factors which are responsible for determining the Human Development Index (HDI) of the country so that areas of improvement can be identified.

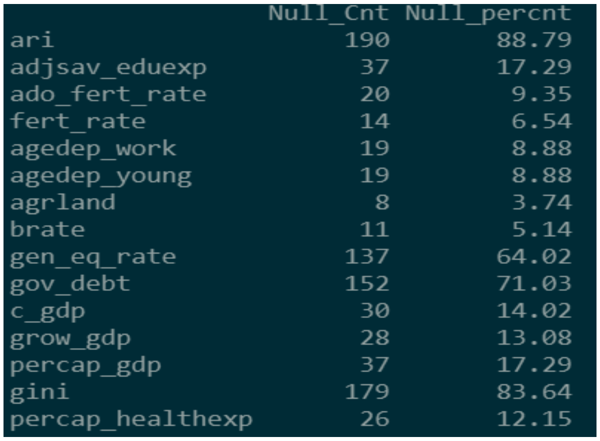
Data Preprocessing

Data has 214 observations and 33 columns. There are 4 categorical and 29 continuous variables. All 4 categorical variables are unique for each of the observations; hence we remove them. The variables are- Year, Year Code, Country, and Country Code.

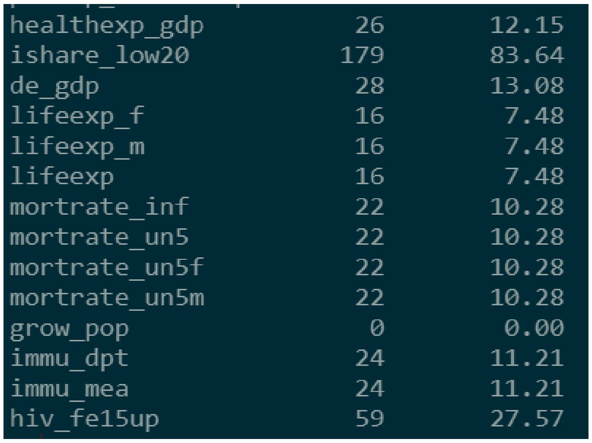
From the data distribution plot, we see that it is not normal for most of the variables, hence we need to scale the data before performing clustering.

Data distribution plot for all the variables is attached in the appendix.

We then check the missing data proportion in the overall dataset.



```
                   Null_Cnt Null_percnt
ari                     190       88.79
adjsav_eduexp            37       17.29
ado_fert_rate            20        9.35
fert_rate                14        6.54
agedep_work              19        8.88
agedep_young             19        8.88
agrland                   8        3.74
brate                    11        5.14
gen_eq_rate             137       64.02
gov_debt                152       71.03
c_gdp                    30       14.02
grow_gdp                 28       13.08
percap_gdp               37       17.29
gini                    179       83.64
percap_healthexp         26       12.15
```



```
healthexp_gdp            26       12.15
ishare_low20            179       83.64
de_gdp                   28       13.08
lifeexp_f                16        7.48
lifeexp_m                16        7.48
lifeexp                  16        7.48
mortrate_inf             22       10.28
mortrate_un5             22       10.28
mortrate_un5f            22       10.28
mortrate_un5m            22       10.28
grow_pop                  0        0.00
immu_dpt                 24       11.21
immu_mea                 24       11.21
hiv_fe15up               59       27.57
```

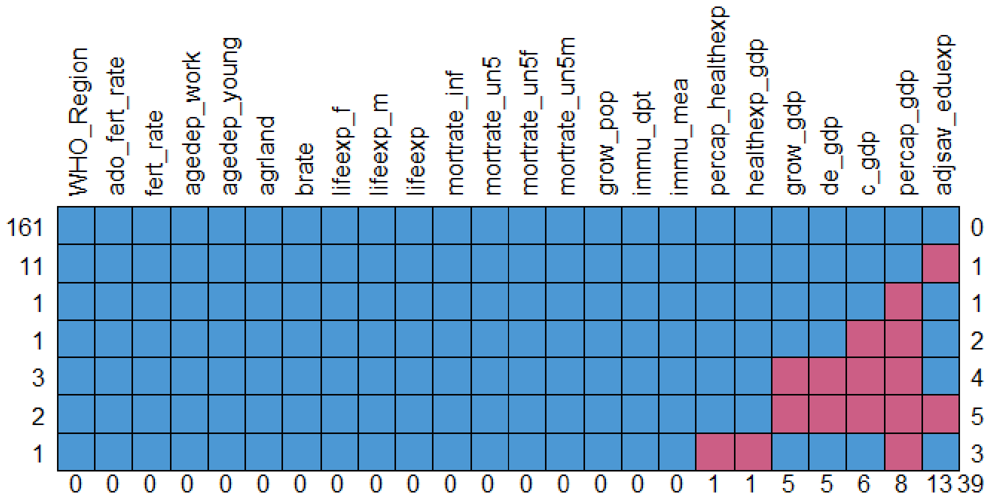
As part of data cleaning, we first dropped columns and rows which had more than 25% null values. As imputation for highly sparse variables and observations would lead to bias in the data and would not be appropriate.

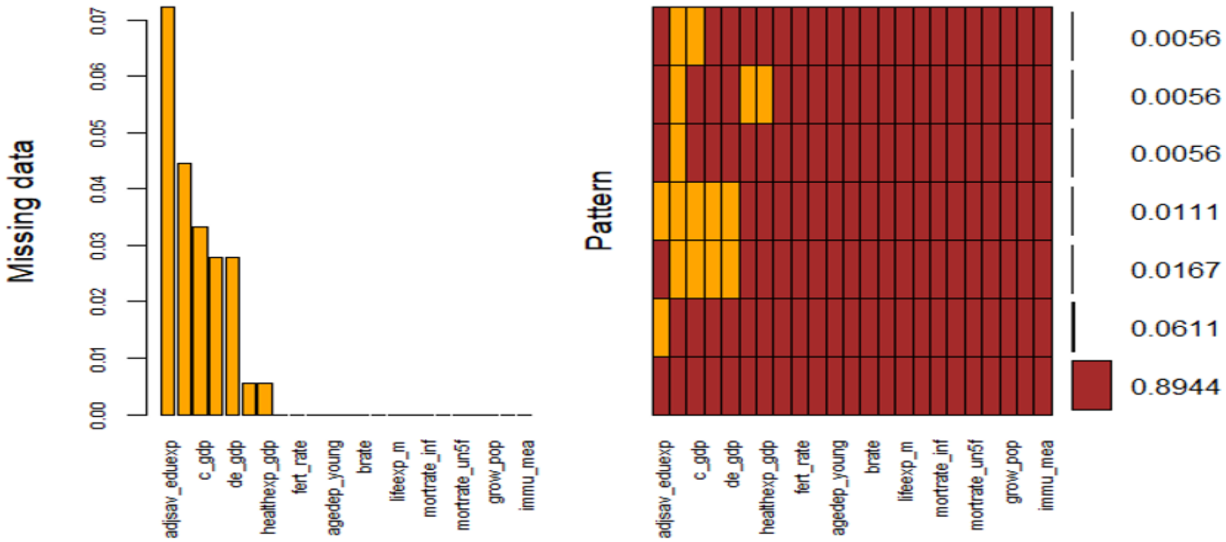
Dropped columns list with null proportion:

ari	(89%)
gen_eq_rate	(64%)
gov_debt	(71%)
gini	(84%)
ishare_low20	(84%)
hiv_fe15up	(28%)

After dropping columns and rows, we are left with 180 observations and 24 columns. The data still had a few null values which were Missing at Random (MAR), hence we used the Mice package to impute the missing values.

Checking the missing value distribution:





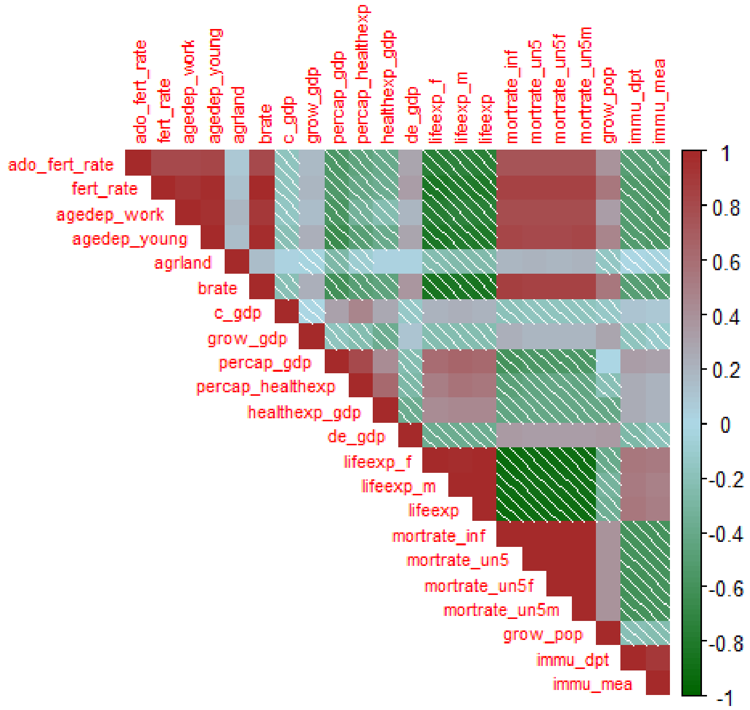
From the above plots, we see 161 observations (89.4%) of the data have no null values. “adjsav_eduexp” has the maximum number of null values contributing to 6.11% followed by c_gdp with 1.6% null values and so on.

Outlier Treatment

On the final imputed dataset, we checked for Outliers using Mahalanobis Distance with 0.95% probability. We found 28 observations were classified as outliers, we then checked the data and found that though the values were falling out of the probability range, but those were legitimate values. As different countries can have a varied range of values for a specific indicator. Hence, we chose not remove the observations identified as outlier using Mahalanobis distance.

Correlation Plot

We now check if the variables have correlation amongst them. We draw the correlation plot to check the same:



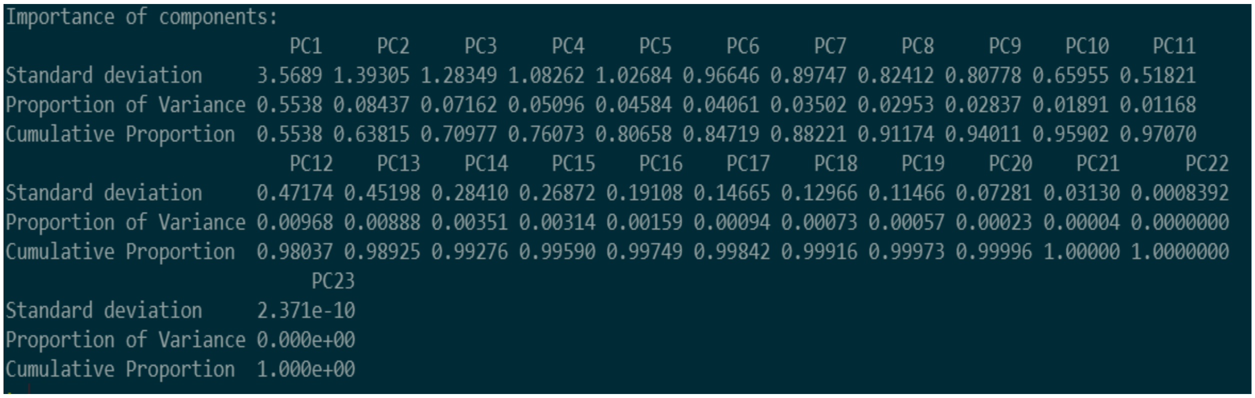
From the above figure, we see that correlation exists between the variables. For instance:

- life expectancy is highly negatively correlated with mortality rates and have positive correlation with fertility rate and % of working age population
- Mortality rate is positively correlated with fertility rate, birth rate and % of working population

Principal Component Analysis (PCA)

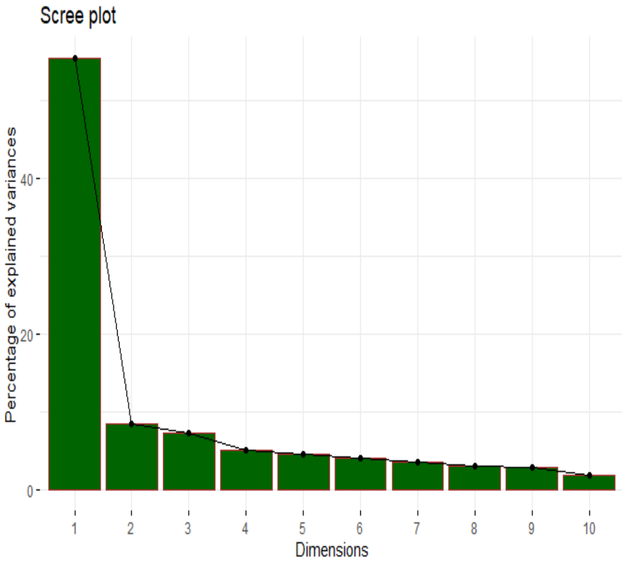
Since we observed correlation amongst variables, to reduce dimensionality of the dataset, we conducted PCA on the dataset to confine the number of variables used for analysis.

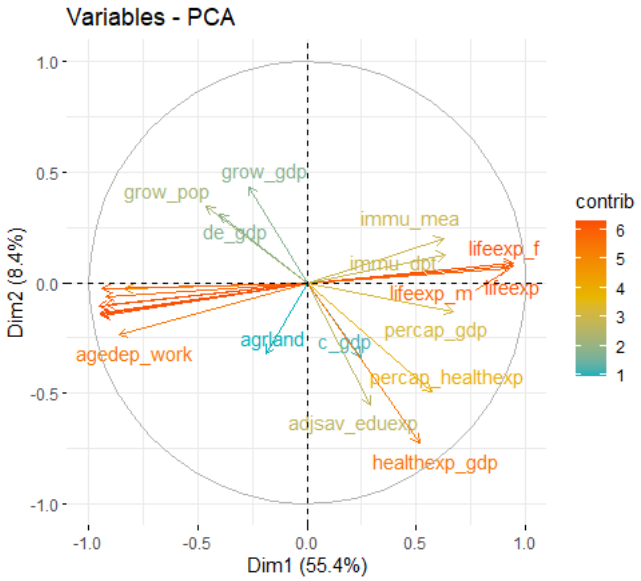
We chose first 6 principal components as they constituted 85% of the total variation in the data.



```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     3.5689 1.39305 1.28349 1.08262 1.02684 0.96646 0.89747 0.82412 0.80778 0.65955 0.51821
Proportion of Variance 0.5538 0.08437 0.07162 0.05096 0.04584 0.04061 0.03502 0.02953 0.02837 0.01891 0.01168
Cumulative Proportion  0.5538 0.63815 0.70977 0.76073 0.80658 0.84719 0.88221 0.91174 0.94011 0.95902 0.97070
                          PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21      PC22
Standard deviation     0.47174 0.45198 0.28410 0.26872 0.19108 0.14665 0.12966 0.11466 0.07281 0.03130 0.0008392
Proportion of Variance 0.00968 0.00888 0.00351 0.00314 0.00159 0.00094 0.00073 0.00057 0.00023 0.00004 0.0000000
Cumulative Proportion  0.98037 0.98925 0.99276 0.99590 0.99749 0.99842 0.99916 0.99973 0.99996 1.00000 1.0000000
                            PC23
Standard deviation     2.371e-10
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

Scree Plot and Variable PCA of explaining the variable variance for the dimensions and correlated variables:



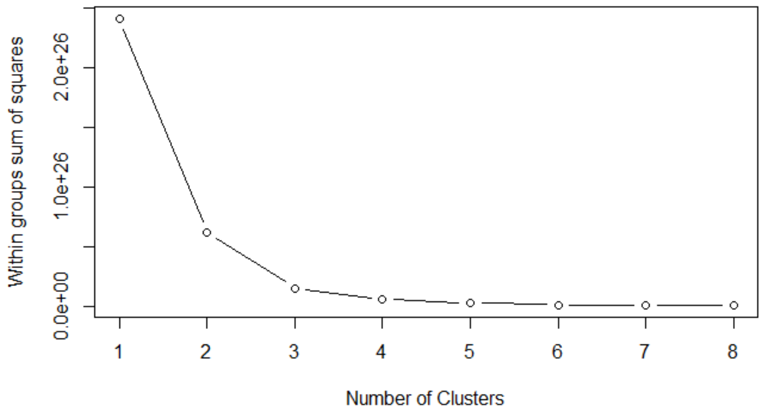


Clustering

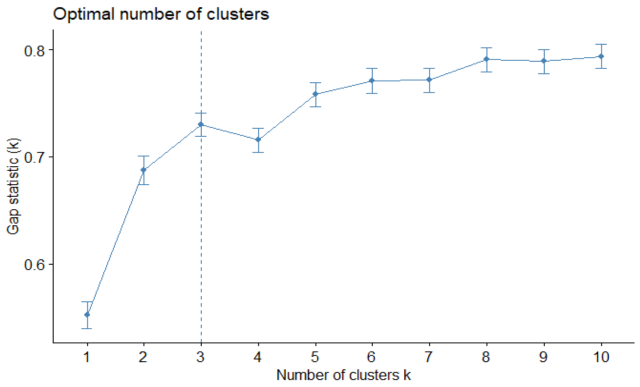
After we reduced the dimensions of the data and chose 6 principal components, then we had to perform the clustering analysis. First, we had to decide the optimal no of k. We performed Within sum of square method and the Gap Statistics Method to obtain the optimal k value as 3.

We decided to choose k-means over k-medoids because the clustering result with k-means was giving better results. Also, this was one of the stated advantages of the k-means approach that if performed in continuous iterations it gives a better result.

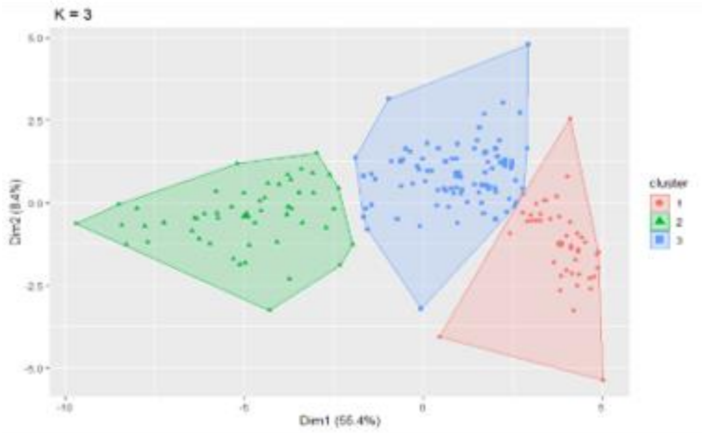
WSS Method

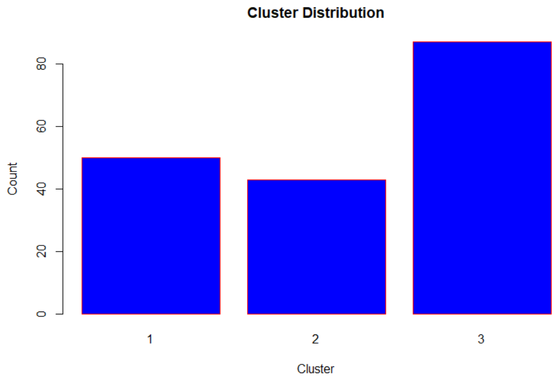


Gap Statistics Method



Clustering distribution:





From the above diagram we see that for the first two dimensions alone, we got non overlapping clusters.

Cluster Labelling

In the cluster distribution that we got which evidently represent the three distinct values where the clusters can be labelled as developing, developed and the underdeveloped countries by looking at the cluster features. Countries falling under Developed group had high GDP, high life expectancy, higher immunization whereas countries that were categorized as Under Developed had high mortality and fertility rate, low GDP, more agriculture land, less spend on education and health, etc. The three groups had clear distinct characteristics.

* *Box plot for different clusters attached in the appendix*

Cluster Validation

To validate the cluster results we performed Lavene Test and One-way Anova Test to check if three cluster groups are statistically different. So, we divided the data into 5 segments viz. Socio Economic, Mortality Rate/100 births, GDP, Life Expectancy & Immunization and Birth Rate. We then compared the group means for one random indicator from each segment.

We conducted the hypothesis test where,
H (0) = There is no difference in the variable means of the 3 cluster groups
H (a) = There is a significant difference between the means of the 3 cluster groups

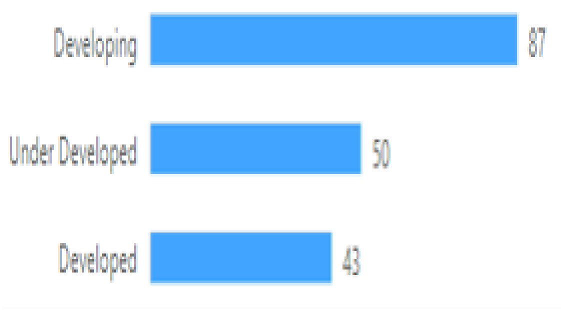
As a result, we deduced from the p value which was less than alpha=0.05 we rejected the null hypothesis and accepted the alternate hypothesis i.e the mean of the cluster groups is significantly different from each other.

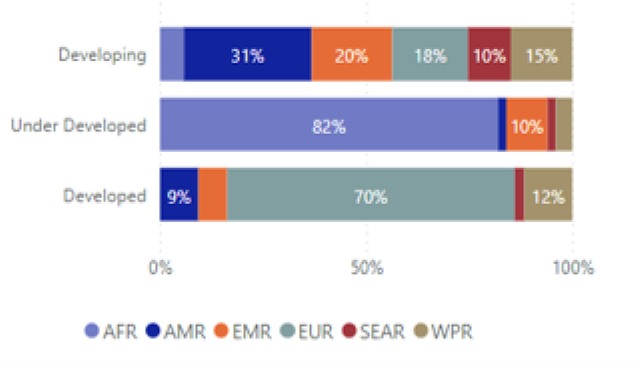
Deductions/results from cluster analysis

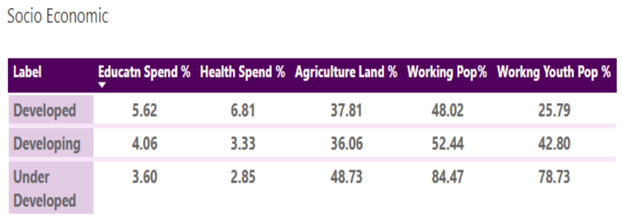
After performing the clustering analysis, we got some interesting insights from our analysis. 48.3% of the total countries were classified as developing, 27.7% countries were named as underdeveloped whereas 23.8% were categorized as developed countries. We noticed that 82% of the underdeveloped countries were from Africa and 70% of the developed countries were from Europe. The main reason for this was the good rate of gross domestic product (GDP) and Europe is one of the initial nations in the world to have an industrial revolution was also one of the reasons that these nations were developed. Also, the expenditure on education and health awareness was much higher than in other nations for the developed countries.

Another interesting insight that we got from our results is that China being a major economy in the world is still regarded as a developing nation. Countries like Africa were considered underdeveloped as they had high mortality rate, low GDP, and poor education systems. We also found that most of the underdeveloped and developing countries had more agricultural land and they had an agrarian economy.

Given Below are some of the key findings from our analysis

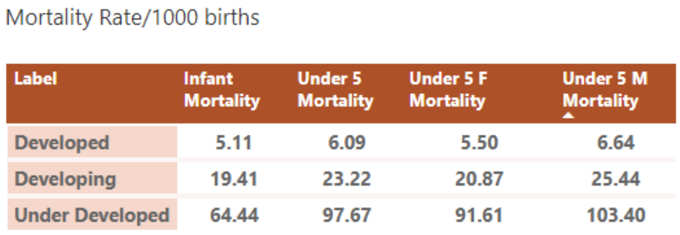






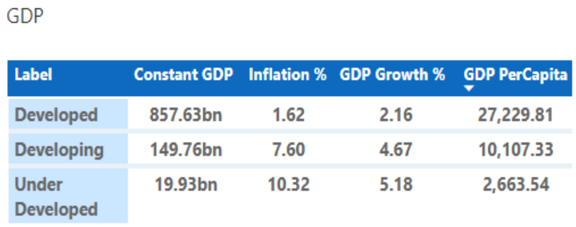
Socio Economic

Label	Educatn Spend %	Health Spend %	Agriculture Land %	Working Pop%	Workng Youth Pop %
Developed	5.62	6.81	37.81	48.02	25.79
Developing	4.06	3.33	36.06	52.44	42.80
Under Developed	3.60	2.85	48.73	84.47	78.73



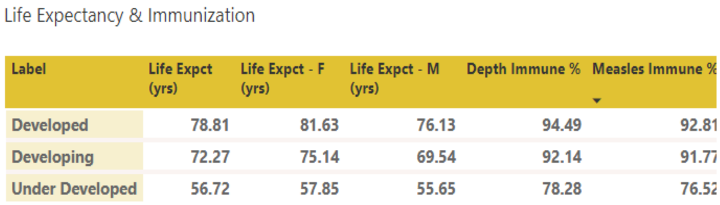
Mortality Rate/1000 births

Label	Infant Mortality	Under 5 Mortality	Under 5 F Mortality	Under 5 M Mortality
Developed	5.11	6.09	5.50	6.64
Developing	19.41	23.22	20.87	25.44
Under Developed	64.44	97.67	91.61	103.40



GDP

Label	Constant GDP	Inflation %	GDP Growth %	GDP PerCapita
Developed	857.63bn	1.62	2.16	27,229.81
Developing	149.76bn	7.60	4.67	10,107.33
Under Developed	19.93bn	10.32	5.18	2,663.54



Life Expectancy & Immunization

Label	Life Expct (yrs)	Life Expct - F (yrs)	Life Expct - M (yrs)	Depth Immune %	Measles Immune %
Developed	78.81	81.63	76.13	94.49	92.81
Developing	72.27	75.14	69.54	92.14	91.77
Under Developed	56.72	57.85	55.65	78.28	76.52

Future Analysis Improvements

- We had tagged region against every country, that introduced hierarchy in the data. We can run hierarchical clustering and compare the result with K-Means partitioning clustering to see if there’s any difference in the result as a next step
- We can add more relevant variables in the data that help in determining the status of the country such as literacy rate, doctors per 1000 persons, gross national income of the country, etc. Adding more variables can help in better clustering results
- A country’s Human Development Index(HDI) is calculated using the indicators life expectancy, education and Gross National Income (GNI). Using the data available, we can predict the HDI and then label the countries as Under Developed, Developing or Developed basis the HDI range and validate the cluster results obtained using the cluster analysis

Recommendation From Analysis

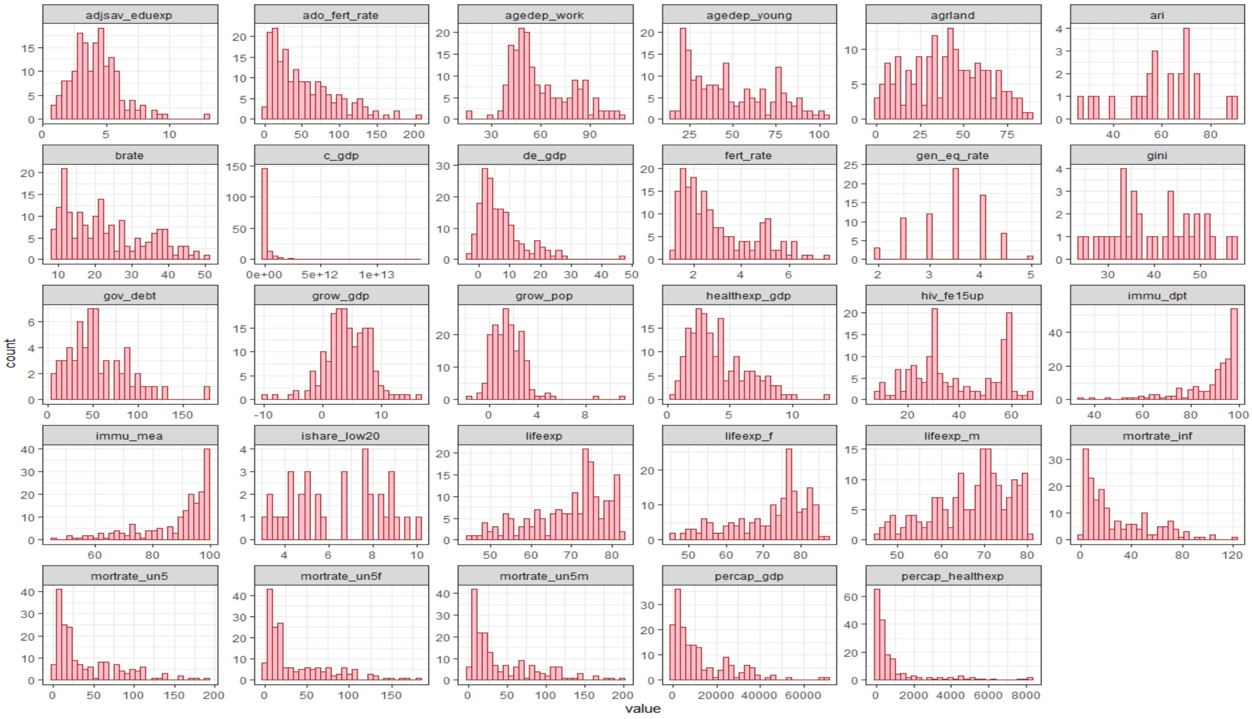
- Increasing literacy levels and spending on education is a critical aspect in enhancing the economic development of an underdeveloped and developing country
- Underdeveloped countries should diversify away from agriculture and aim toward manufacturing as a means of stimulating economic growth
- Government should emphasize building good health infrastructure and provide immunization so that there is a high life expectancy and less mortality rate leading to a happy life

APPENDIX

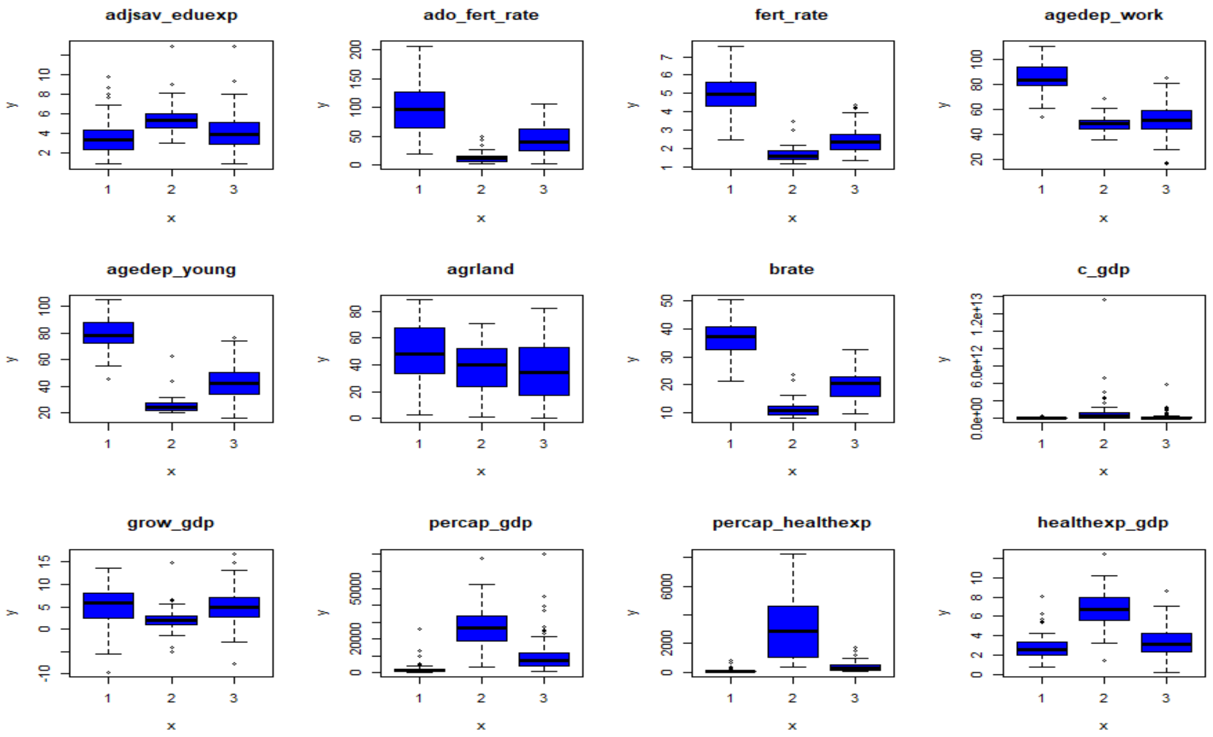
Data Variable Description

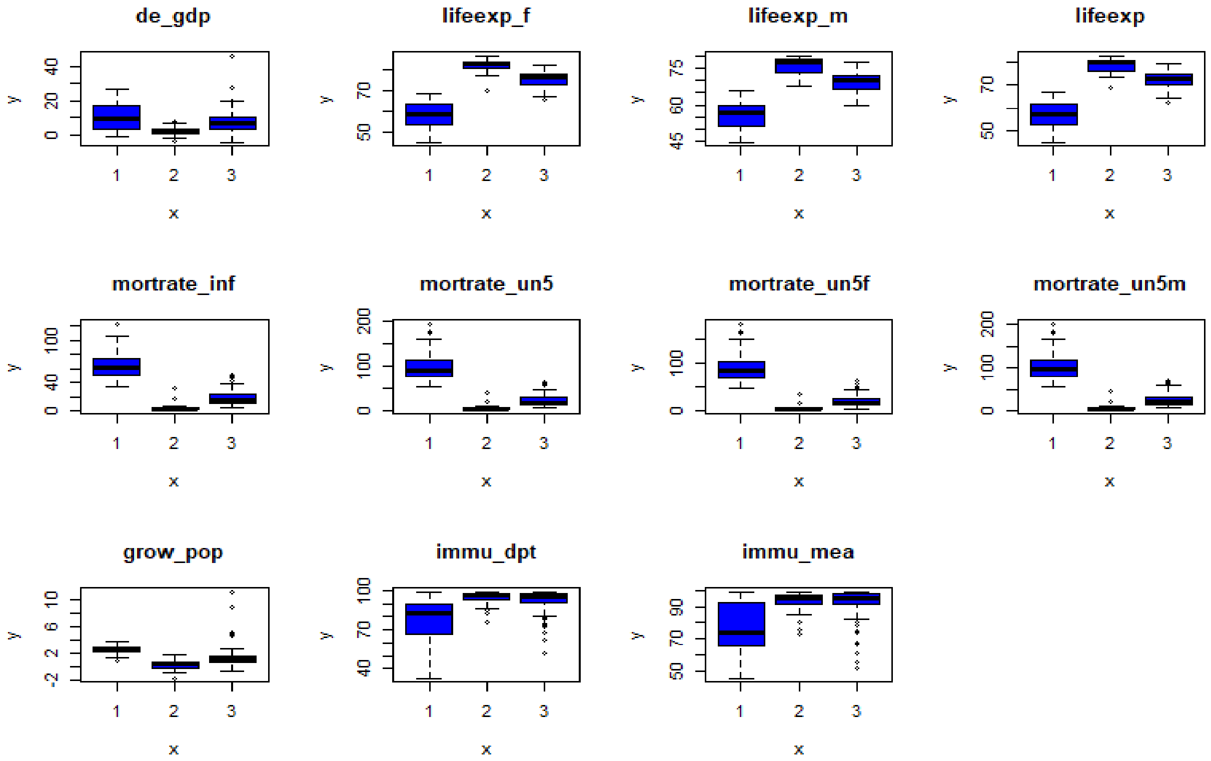
Variables	Type	Full Name
Year	Char	Year
YearCode	Char	YearCode
Country me	Char	Country Name
Country Code	Char	Country Code
ari	Num	ARI treatment (% of children under 5 taken to a health provider)
adjsav_eduexp	Num	Adjusted savings: education expenditure (% of GNI)
ado_fert_rate	Num	Adolescent fertility rate (births per 1,000 women ages 15-19)
agedep_work	Num	Age dependency ratio (% of working-age population)
agedep_young	Num	Age dependency ratio, young (% of working-age population)
agrland	Num	Agricultural land (% of land area)
brate	Num	Birth rate, crude (per 1,000 people)
gen_eq_rate	Num	CPIA gender equality rating (1=low to 6=high)
gov_debt	Num	Central government debt, total (% of GDP)
fert_rate	Num	Fertility rate, total (births per woman)
c_gdp	Num	GDP (constant 2005 US$)
grow_gdp	Num	GDP growth (annual %)
percap_gdp	Num	GDP per capita, PPP (constant 2005 international $)
gini	Num	GINI index
percap_healthexp	Num	Health expenditure per capita (current US$)
healthexp_gdp	Num	Health expenditure, public (% of GDP)
ishare_low20	Num	Income share held by lowest 20%
de_gdp	Num	Inflation, GDP deflator (annual %)
lifeexp_f	Num	Life expectancy at birth, female (years)
lifeexp_m	Num	Life expectancy at birth, male (years)
lifeexp	Num	Life expectancy at birth, total (years)
mortrate_inf	Num	Mortality rate, infant (per 1,000 live births)
mortrate_un5	Num	Mortality rate, under-5 (per 1,000 live births)
mortrate_un5f	Num	Mortality rate, under-5, female (per 1,000)
mortrate_un5m	Num	Mortality rate, under-5, male (per 1,000)
grow_pop	Num	Population growth (annual %)
immu_dpt	Num	Immunization, DPT (% of children ages 12-23 months)
immu_mea	Num	Immunization, measles (% of children ages 12-23 months)
hiv_fe15up	Num	Women's share of population ages 15+ living with HIV (%)

Data distribution for all the variables:



Cluster Distribution





R Code

```
ibrary(dplyr)
library(tidyverse)
#install.packages("mice")
library(mice)
#install.packages("VIM")
library(VIM)
library(cluster)
#install.packages("psych")
library(psych)
#install.packages("corrplot")
library(corrplot)
#install.packages("factoextra")
library(factoextra)
#install.packages("sjmisc")
library(sjmisc)
library(readr)
library(rgl)
library(cluster)
library(car) #- lavene test
```

```
data_org <- read_csv("D:/Semester 4/DANA/Project/Final Project/data.csv")
data <- data_org

#Removing the columns Year and YearCode since they are constant.
#Removin the column Country Code, since its like primary key and unique for each row.
data <- data[,-c(1,2,3)]

#Dealing with Missing values
Null_Cnt <- sapply(data, function(x){ sum(is.na(x))})
Null_percnt <- sapply(data, function(x){ round((sum(is.na(x))/length(x))*100,2) })
Null_Smry <- cbind(Null_Cnt,Null_percnt)
Null_Smry <- as.data.frame(Null_Smry)

col_list <- row.names(Null_Smry)[Null_percnt > 25] #"ari" "gen_eq_rate"  "gov_debt" "gini"
"ishare_low20" "hiv_fe15up

data <- data[,!(names(data) %in% col_list)]

# removing rows with more than 25% null value
data<- data[which(round(rowSums(is.na(data))/dim(data)[2]*100)<22),] #34 observations
removed - only 1 NO-OBS in data
data


summary(data)

###########################################################
#Data Imputation using MICE
############################################################
md.pattern(data, color = c("orange","dark green"),rotate.names = TRUE) #161 observations
without NA's

mice_plot <- aggr(data, col=c('brown','orange'),
                  numbers=TRUE, sortVars=TRUE,
                  labels=names(data), cex.axis=.7,
                  gap=3, ylab=c("Missing data","Pattern"))

imputed_Data <- mice(data[,-c(1,2)] , m=5, maxit = 50, method = 'cart', seed = 500)
summary(imputed_Data)

#complete_data <- merge_imputations(data,imputed_Data,summary="hist")

complete_data <- complete(imputed_Data,5)
nrow(complete_data[complete.cases(complete_data),]) #180
```

```
#corPlot(complete_data, numbers=FALSE, zlim = NULL, n.legend=5,
scale=TRUE,stars=TRUE,  MAR=TRUE, cex.axis=0.6)
corrplot(cor(complete_data), type = 'upper', method = 'shade',col=colorRampPalette(c("dark
green","lightblue","brown"))(100), tl.cex = 0.7)

#From the correlation matrix we see that, the Fertility rate and Mortality rate is highly correlated.

#Finding outliers using mahalanobis distances
# Finding the center point
complete_data.center  = colMeans(complete_data)

# Finding the covariance matrix
complete_data.cov = cov(complete_data)

# Finding distances
distances <- mahalanobis(x = complete_data , center = complete_data.center , cov =
complete_data.cov, tol=1e-40)


# Cutoff value for ditances from Chi-Sqaure Dist.
# with p = 0.95 df = 6 which in ncol(df_num)
cutoff <- qchisq(p = 0.95 , df = ncol(complete_data))
complete_data$distances <- as.factor(ifelse(distances > cutoff, 0, 1))

complete_data_num <- complete_data[,-c(24)]

#complete_data_num = cbind(data[,c(1,2)],complete_data_num)
#complete_data_num<- complete_data_num[,-c(1,2)]
##################################################################################
#####
#We will confine the number of variables used for the analysis using PCA.
##################################################################################
#####
data_PCA <- prcomp(complete_data_num, scale = TRUE)
summary(data_PCA)
print(data_PCA)
#data_PCA$x

#dev.off()

# Selecting top 6 Principal components
comp <- data.frame(data_PCA$x[,1:6])


#Scree Plot
fviz_eig(data_PCA, barfill = "dark green", barcolor ="brown", linecolor  = "black" )
```

```
#Graph of individuals
fviz_pca_ind(data_PCA,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
###############################################################################
#############
#Variable-PCA
fviz_pca_var(data_PCA,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)

#Biplot
fviz_pca_biplot(data_PCA, repel = TRUE,
                col.var = "#2E9FDF", # Variables color
                col.ind = "#696969"  # Individuals color
)

# Eigenvalues
eig.val <- get_eigenvalue(data_PCA)
eig.val
###############################################################################
#############
# Determine number of clusters using WSS and Gap Stat
wss <- (nrow(complete_data_num)-1)*sum(apply(complete_data_num,2,var))
for (i in 2:8) wss[i] <- sum(kmeans(complete_data_num,
                                     centers=i)$withinss)
plot(1:8, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")

fviz_nbclust(x = scale(complete_data_num),FUNcluster = kmeans, method = 'gap_stat' )

#Euclidean distance
ed = dist(complete_data_num)


####### k=3
km2 <- kmeans(comp, 3, nstart=25, iter.max=1000)
km2$size # 50 43 87

# Adding the region and country to the data
complete_data = cbind(data[,c(1,2)],complete_data)
```

```
complete_data$cluster <- km2$cluster
table(complete_data$WHO_Region, complete_data$cluster)
##          1  2  3
##AFR     41  0  5
##AMR      1  4 26
##EMR      5  3 17
##EUR      0 30 16
##NA- AMR  0  0  1
##SEAR    1  1  9
##WPR      2  5 13
```

```
fviz_cluster(km2, geom = "point", data =complete_data_num) + ggtitle(" K = 3")
```

```
names(complete_data)[1] <- "Country"

complete_data[,c(1,2,27)]<-lapply(complete_data[,c(1,2,27)], as.factor)

par(mfrow=c(3,4)) # define 2x5 multiframe graphic
for (i in 3:(ncol(complete_data)-2)) # make box plots for all columns except the cluster label
{
  if (is.numeric(complete_data[,i])) # if numeric -> boxplot
  {
    plot(complete_data$cluster, complete_data[,i], main= colnames(complete_data)[i], col=
"blue")
  }
  else # if factor -> barplot
  {
    count<-table(complete_data[,i],complete_data$cluster)
    barplot(count, legend = rownames(count), main= colnames(complete_data)[i],
col=rainbow(6))
  }
}

par(mfrow=c(1,1)) # define 2x5 multiframe graphic

count<-table(complete_data[,"cluster"],complete_data$cluster)
barplot(count, legend = rownames(count), main= colnames(complete_data)[27], col=rainbow(6))

barplot(count,
        main="Cluster Distribution",
        xlab="Cluster",
        ylab="Count",
        border="red",
        col="blue"
)

write.csv(complete_data, "D:/Semester 4/DANA/Project/Final
Project/Complete_data_with_cluster.csv",row.names = FALSE)


################################################
###Cluster Validation
################################################
# Bartlett's test when data is normally distributed
# H(0) = There is no difference between the variances of 3 cluster groups
# H(a) = The 3 cluster groups have variance

bartlett.test(c_gdp ~ cluster, data = complete_data)

# p-value =  2.2e-16, means variance in c_gdp is significantlly different for the 3 cluster groups

# Lavene's test when data is not normally distributed

leveneTest(c_gdp ~ cluster, data = complete_data)

# p-value =  0.001226 , means variance in c_gdp is significantlly different for the 3 cluster
groups

##########################Anova Test ############################

# H(0) = There is no difference in the gdp means of 3 cluster groups
# H(a) = There is difference between gdp means of 3 cluster groups

oneway.test(c_gdp ~ cluster, data = complete_data, var.equal = TRUE)
# p-value =  0.000645 , means variance in c_gdp is significantlly different for the 3 cluster
groups




#2.------------------------------------------
# Health Spend

bartlett.test(healthexp_gdp ~ cluster, data = complete_data)   # pvalue - 0.04 , reject null

oneway.test(healthexp_gdp ~ cluster, data = complete_data, var.equal = TRUE)
# p-value =  2.2e-16 , means variance in c_gdp is significantlly different for the 3 cluster groups
```

```
#3.------------------------------------------------
# Mortality rate

bartlett.test(mortrate_inf ~ cluster, data = complete_data)   # pvalue - 1.07e-14 , reject null

oneway.test(mortrate_inf ~ cluster, data = complete_data,  var.equal = TRUE)
# p-value =  2.2e-16 , means variance in c_gdp is significantlly different for the 3 cluster groups

#4.------------------------------------------------
# Life Expectancy

bartlett.test(lifeexp ~ cluster, data = complete_data)   # pvalue - 3.927e-05 , reject null

oneway.test(lifeexp ~ cluster, data = complete_data,  var.equal = TRUE)
# p-value =  2.2e-16 , means variance in c_gdp is significantlly different for the 3 cluster groups

#5.------------------------------------------------
# Birth Rate


bartlett.test(brate ~ cluster, data = complete_data)   # pvalue - 4.006e-05 , reject null

oneway.test(brate ~ cluster, data = complete_data,  var.equal = TRUE)
# p-value =  2.2e-16 , means variance in c_gdp is significantlly different for the 3 cluster groups
```