BIG DATA ANALYSIS

September 2019

Exploratory Data Analysis - Analysis of Social Economy Indicators of G-20 Countries

Kadek Ardya Novi Diani
09211850096004

A. Data Explanation

In this paper utilizing data social economy of G-20 countries in 2015, that is shared in World Bank official website. Here is the link for the data:
https://databank.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG/1ff4a498/Popular-Indicators#
There are 10 variables of social economy that being used in this analysis, they are:

- **Exports of goods & services (% of GDP)**
  Represent the value of all goods and other market services provided to the rest of the world.
- **Fertility rate (birth per woman)**
  The number of children that would be born to a woman if she were to live to the end of her childbearing years.
- **GDP growth (annual %)**
  Annual percentage growth rate of GDP at market prices based on constant local currency.
- **Immunization (% of children ages 12-23 months)**
  Measures the percentage of children ages 12-23 months who received the measles vaccination before 12 months.
- **Imports of goods & services (% of GDP)**
  The value of all goods and other market services received from the rest of the world.
- **Inflation (annual %)**
  The rate of price change in the economy as a whole.
- **Life expectancy at birth (years)**
  The number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
- **Mortality rate, under 5 years (per 1000 live births)**
  The probability per 1,000 that a newborn baby will die before reaching age five
- **Population growth (annual %)**
  Annual population growth rate. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
- **Revenue (% of GDP)**
  Cash receipts from taxes, social contributions, and other revenues (except grants).

Here is member of G-20 organization:

- Argentina
- Australia
- Brazil
- Canada
- China
- European Union
- France
- Germany
- India
- Indonesia
- Italy
- Japan
- Korea
- Mexico
- Russia
- Saudi Arabia
- South Africa
- Turkey
- United Kingdom
- United States

Considering European Union is an organization instead of a country, then in this paper will not including analysis for European Union, and only focusing on condition of countries member.

B. Purpose of Analysis

G-20 (Group of 20) is an organization that consist of finance ministers and central bank governors from 19 developed and developing countries, and European Union, which has a mandate to promote global economy growth, international trading, and regulation of financial markets.

Therefore, this paper will analyze character of those 19 countries that listed as the member of G-20 based on social economy point of view.

Here is the list of the purpose that would like to expose in this paper:

- To know overall condition of social economy indicator of all country member G-20.
- To clarify position of Indonesia among G-20 countries based on social economy indicator.
- To analyze the dependency among 10 social economy indicator that is mentioned in previous section for G-20 countries.
- To understand social economy condition of each countries in G-20 through classification method.

C. Results & Analysis

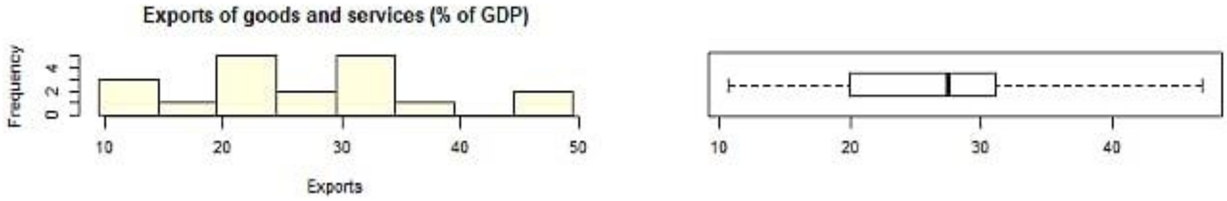
1. Statistics Descriptive

To figuring out the condition of each social economy indicator, it is useful to use statistics descriptive method. Here is the simple statistics descriptive results for those 10 variables of social economy:

var	exports	fertility	GDP Growth	immunization	imports	inflation	life expect	mortality	pop growth	revenue
rata	26,21	1,90	2,50	92,53	25,98	2,91	76,92	12,11	0,91	24,67
varians	100,89	0,15	7,93	38,15	77,81	60,07	32,25	139,57	0,41	78,07
maks	46,86	2,51	8,00	99,00	38,85	26,58	83,79	43,60	2,56	44,29
min	10,71	1,24	-3,55	75,00	11,78	-16,91	62,65	3,00	-0,11	12,42

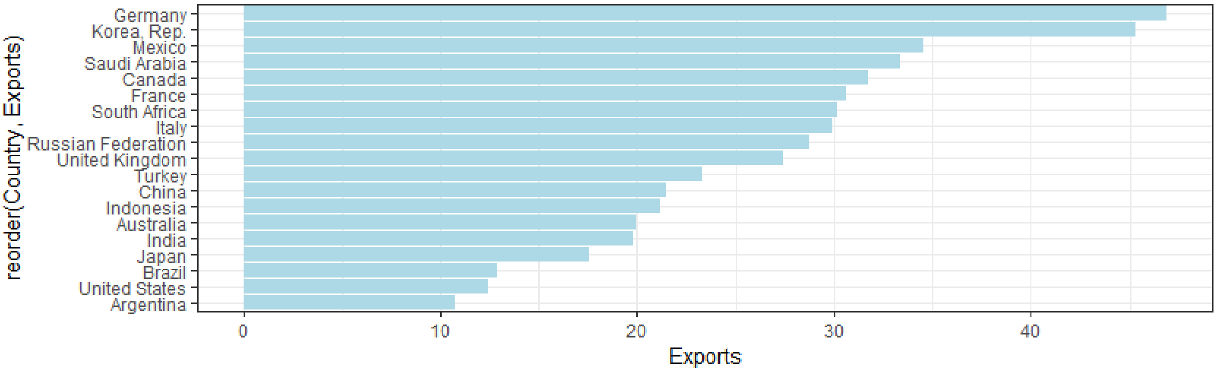
The easier way to understand the condition of each variables is through data visualization. Below is the histogram and boxplot of each variables, which shows data distribution of the G-20 countries.

- **Exports of goods & services (% of GDP)**



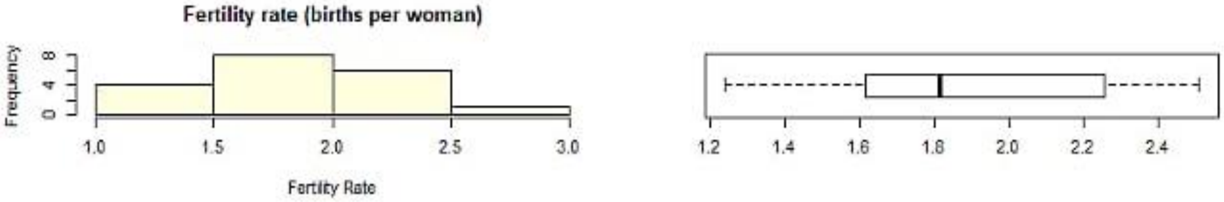
Average amount of exports in G-20 countries is 26.2% of GDP, with varies in 10% up to 50%, which shows the variance is quite high (variance result is 100.89).

Below is the bar chart of G-20 countries for exports variable:

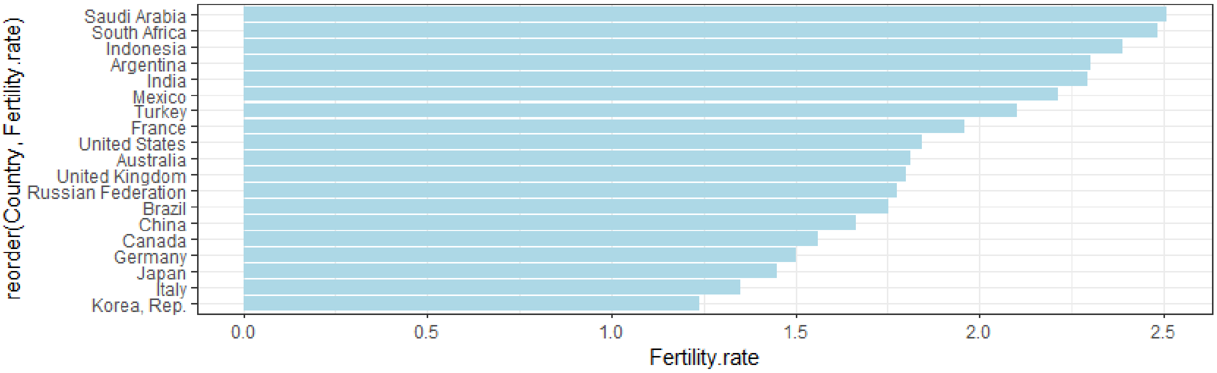


Country which has the highest exports portion are Germany & Korea, while the lowest is Argentina. Indonesia position is still below the average, which is around 20%.

- **Fertility rate (birth per woman)**

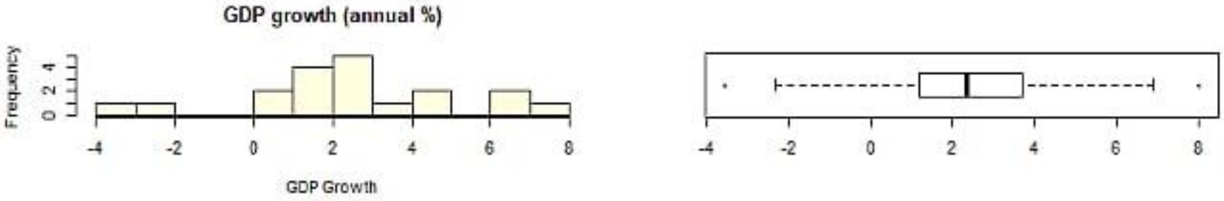


Fertility rate average of G-20 countries are around 1.9, or easier to say that average child to born from 1 woman is 2 children. The variance is quite small, data is varying only from 1 to 3 children, but the density data mostly in 2 children.

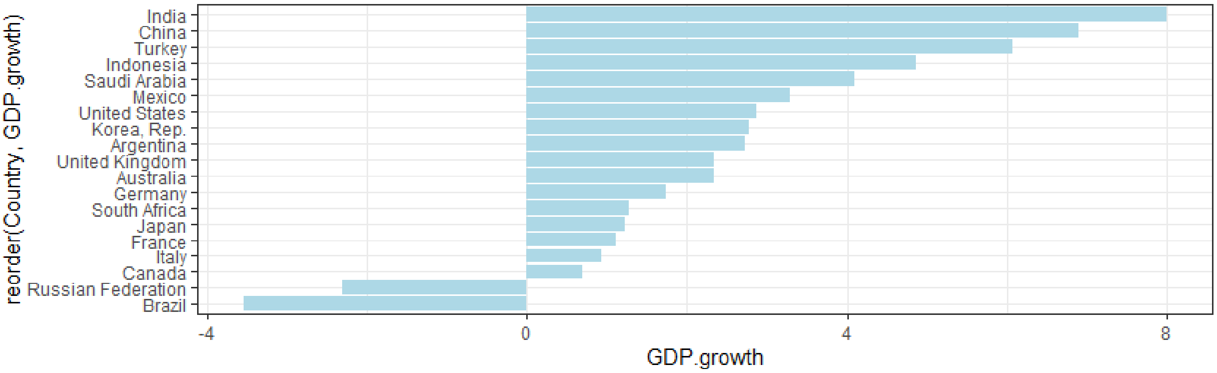


Indonesia's fertility rate is above the average, even though the highest fertility rate belongs to Saudi Arabia, while the lowest is Korea, which only 1 child born from every 1 woman.

- **GDP growth (annual %)**

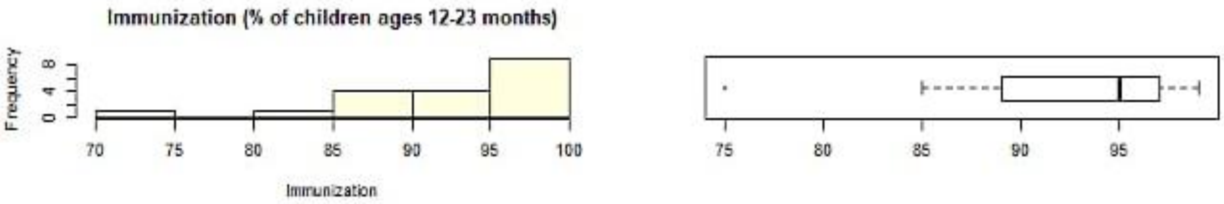


Average GDP growth of G-20 countries is 2.5% annually, with data distribution from -2% to 7%. Box plot above also shown that there are 2 outlier data in GDP variable, the one that the GDP growth is -4%, the other one is that GDP growth 8%.

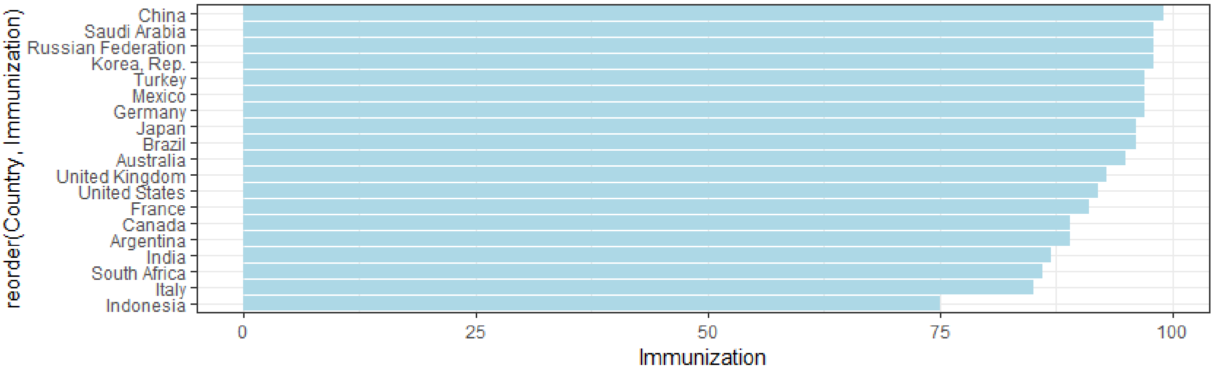


As developing country, Indonesia's GDP growth is in 4th position of G-20 countries. The highest GDP growth is India with 8% annual growth, and the lowest is Brazil which even decreasing 4% annually. Bar chart above shows that those outlier countries are India as the highest GDP growth, and Brazil as the lowest.

- **Immunization (% of children ages 12-23 months)**

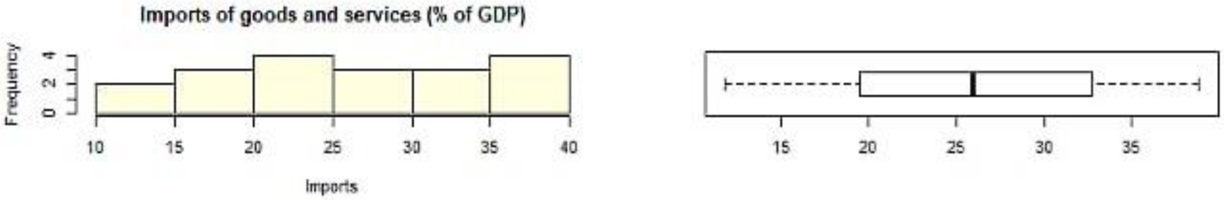


Histogram above shows that G-20 countries highly concern about the importance of immunization for children, since the graph is left skewed, with average of 92.5%. Even though there is still 1 country as outlier which has the lowest immunization rate.

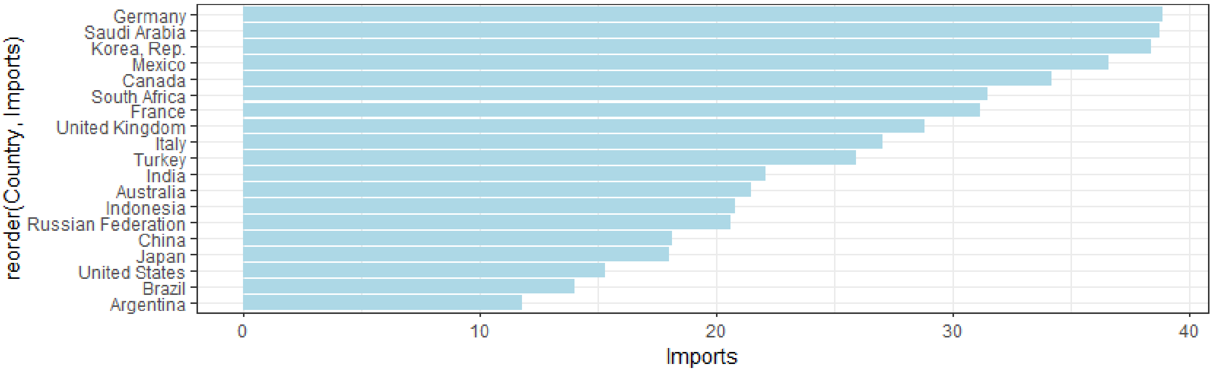


Indonesia is the lowest immunization rate of G-20 countries, which also makes Indonesia as the outlier country.

- **Imports of goods & services (% of GDP)**

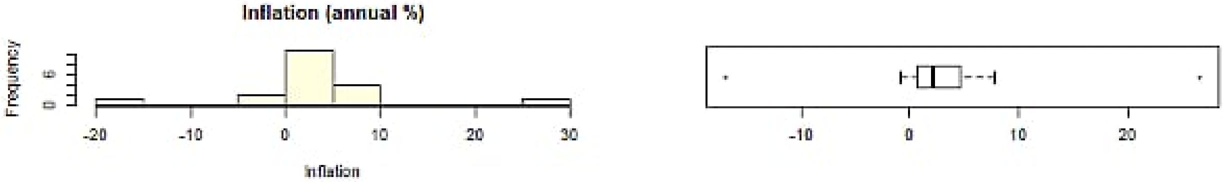


Imports portion of G-20 countries is varies from 10% to 40%, with average around 25.9% of GDP. Chart above also shown that the variance of imports data is quite high, which is 77.8.

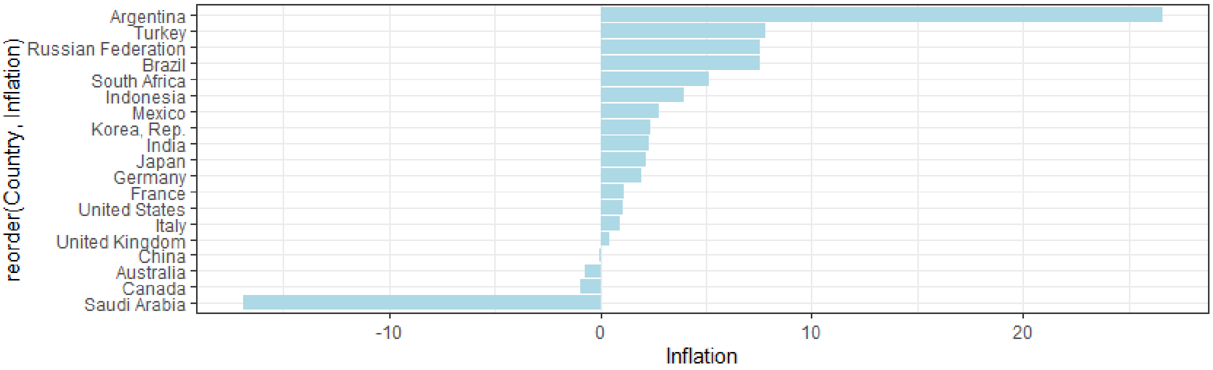


Similar with exports, country with the highest imports portion is also Germany, the lowest also Argentina. While Indonesia position is still below the average.

- **Inflation (annual %)**

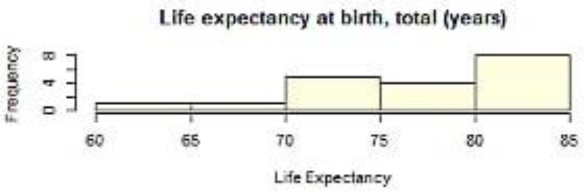


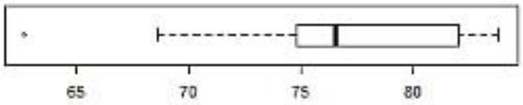
Inflation average of G-20 countries is around 2.9%, with distribution data from -5% to 10%. Based on box plot above also known that there 2 countries that become outlier, which is the one that has inflation -20%, and the one with inflation 30%.



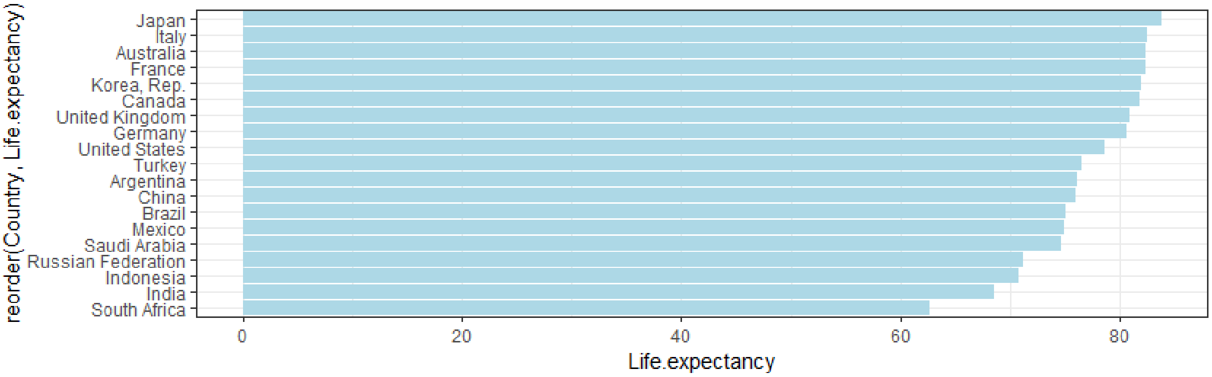
Bar chart above clearly shows that the outlier countries are Saudi Arabia as the lowest inflation, and Argentina as the highest inflation, which reach up to 26%.

- **Life expectancy at birth (years)**



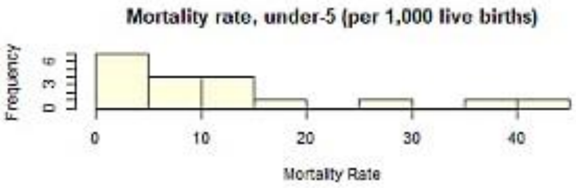


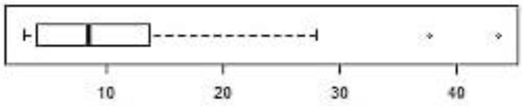
Similar with immunization, life expectancy histogram also left skewed distribution, with average of 76.9 years. Also, there is 1 outlier country that has lowest life expectancy of all, which is only around 60 years.



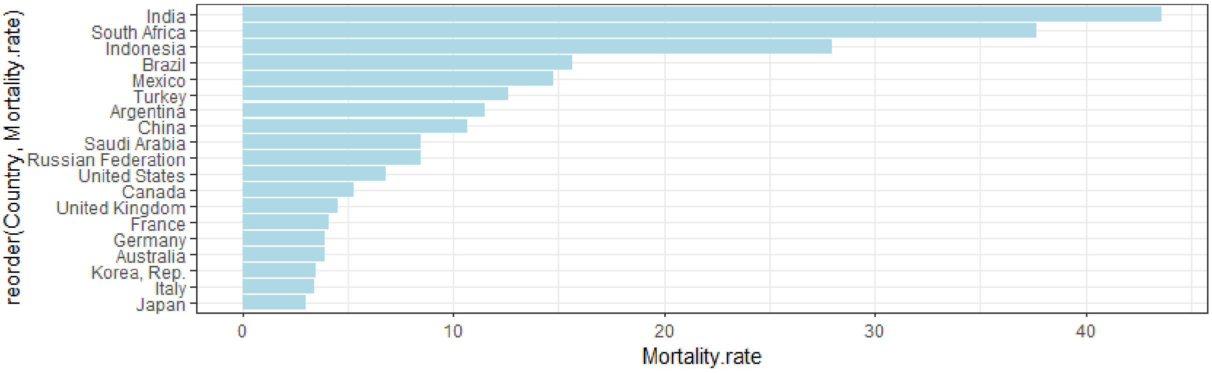
Country with the highest life expectancy is Japan, while the lowest is South Africa, which makes South Africa become the outlier country. Indonesia's life expectancy is actually still quite low, since the position is in number 3 from bottom.

- **Mortality rate, under 5 years (per 1000 live births)**



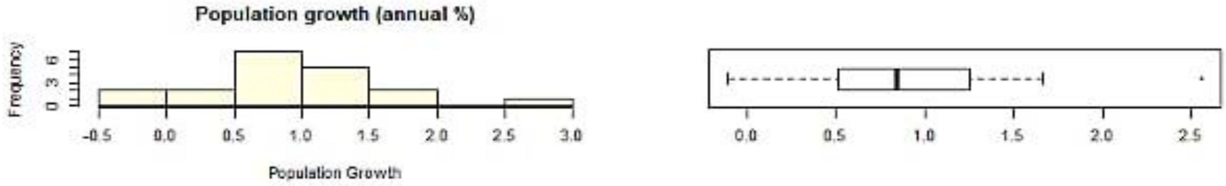


Average mortality rate of G-20 countries is 12 children of 1000 live births, with distribution data from 3 children to 44 children. There are 2 countries that become outliers based on box plot above, which has mortality rate around 40.

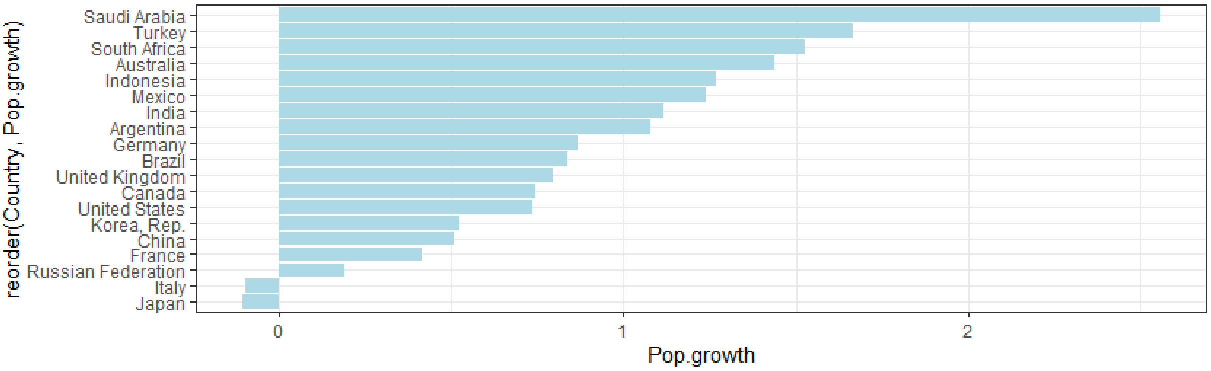


The highest mortality rate belongs to India, and followed by South Africa and Indonesia. Bar chart above shows that Indonesia's mortality rate is over than twice of the average mortality rate of all G-20 countries.

- **Population growth (annual %)**

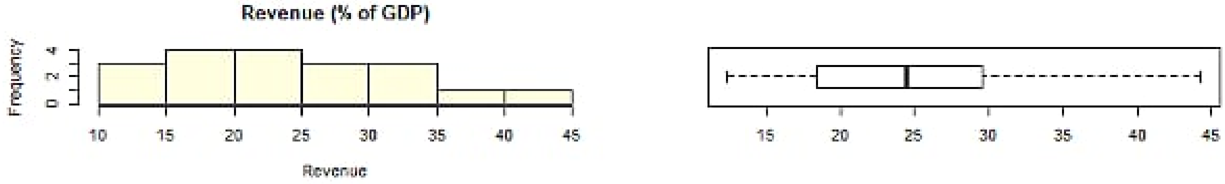


Annual population growth of G-20 countries in average is around 0.91%, with distribution from -0.5% to 2%, and 1 outlier data, which has population growth above 2.5%.

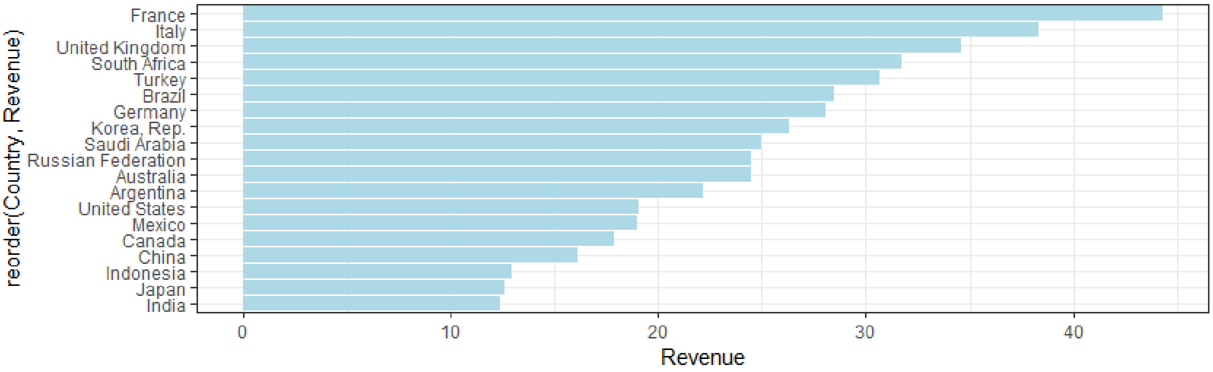


Saudi Arabia is the highest population growth country of all, and the growth is much higher than other, which makes it become an outlier.

- **Revenue (% of GDP)**



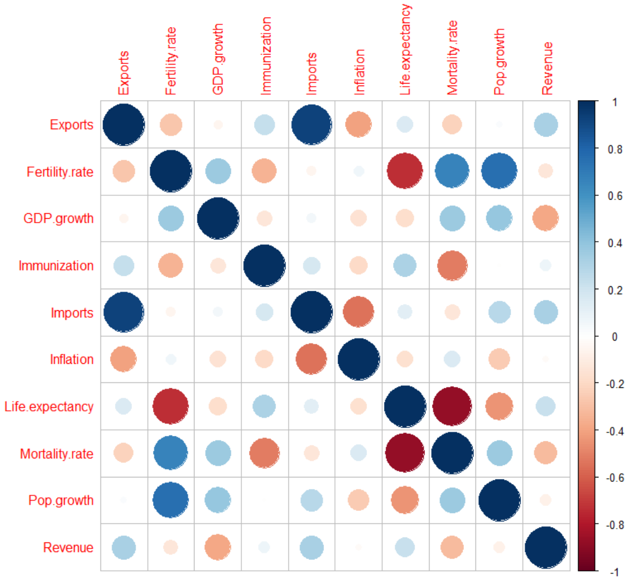
Countries' revenue portion to GDP is distribute varies, from 10% to 45%, with average around 24.67% to GDP.



Top 3 highest revenue belongs to European countries, which are France, Italy, and United Kingdom. Indonesia's revenue is on the lowest 3 position, above Japan and India.

2. Social Economy Indicator Correlation Checking

Based on the descriptive data above, suspected that there is inline pattern in several pairs of social economy indicator data, which leads to correlation between the variables. To figuring out the correlation pattern, here is the correlation matrix chart from output of R:



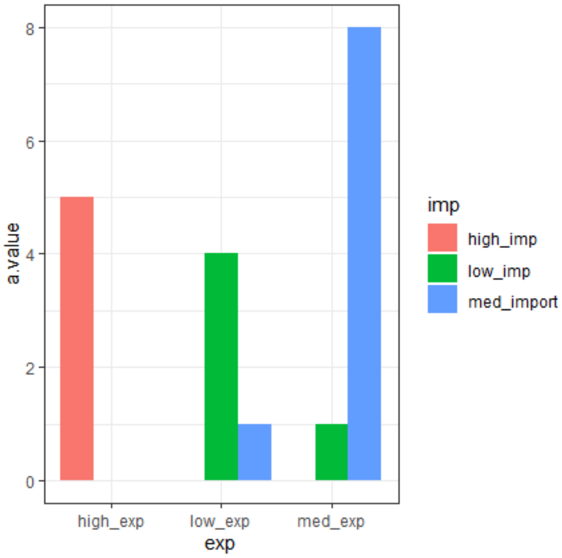
The ticker color of the dot, and the bigger size of the dot, shows that the correlation between 2 variables are stronger. Blue color represents for positive correlation, and red color for negative correlation. Correlation matrix chart above shows that there is correlation between variables, such as:

- Imports & Exports: positive correlation
- Life expectancy & Fertility rate: negative correlation
- Mortality rate & Fertility rate: positive correlation
- Population growth & Fertility rate: positive correlation
- Mortality rate & Life expectancy: negative correlation

Overall there are 6 of 10 variables that has dependency with the other variables.

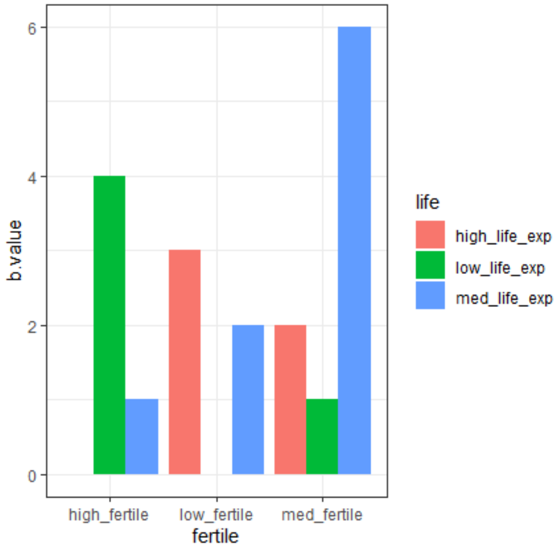
This dependency also can be shown through group bar chart as below:

- **Imports & Exports**



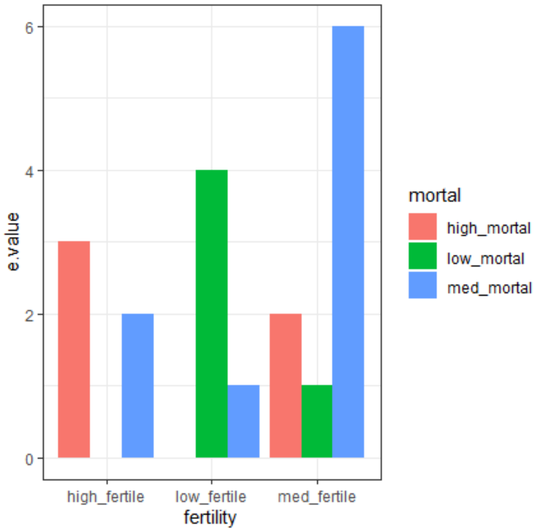
Bar chart above shows that the higher export of a country, the higher also import activity that it has. Those condition is reflecting that the correlation between export and import is positive correlation.

- **Life expectancy & Fertility rate**



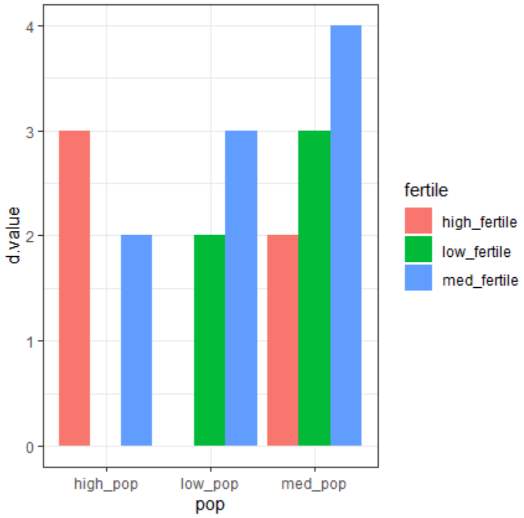
A country which has high fertility rate, tend to has low life expectancy. So does for a country which has low fertility rate, tend to has high life expectancy.

- **Mortality rate & Fertility rate**



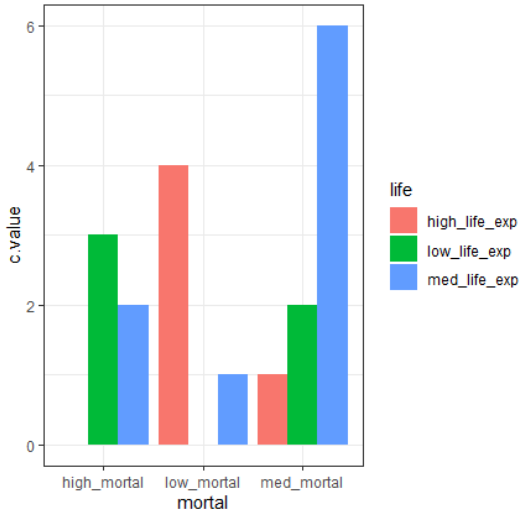
A country which has high fertility rate also tend to has high mortality rate. It shows a positive correlation pattern for those 2 data.

- **Population growth & Fertility rate**



As a common understanding, higher population growth of a country also has higher fertility rate, which indicating positive correlation between those data.

- **Mortality rate & Life expectancy**



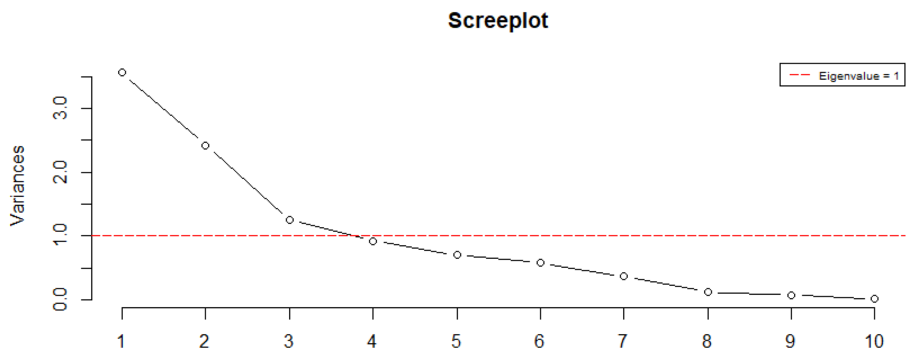
The last data that has correlation are mortality rate and life expectancy, which has negative correlation. The bar chart above also shows that the higher mortality rate of a country, the life expectancy ratio tends to be low.

3. Classification G-20 Countries through PCA

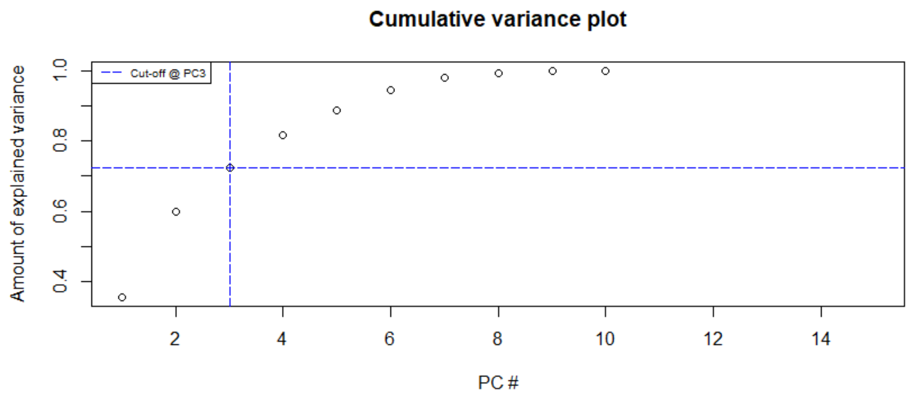
In order to understand the condition of each country of G-20 member in a better way, a classification is needed to mapping all of the country based on its social economy condition. One of statistical method that supports for the classification purpose is Principal Component Analysis (PCA), with the concept of reducing size of variables into smaller one, so that easier to plot the position of each observation (in this case is country).

The sufficient amount of the new group of variables is defined by Eigenvalue of each Principal Component (PC), which the Eigenvalue must be higher than 1.

The Eigenvalue results of social economy indicator of G-20 countries are showed in scree plot below:

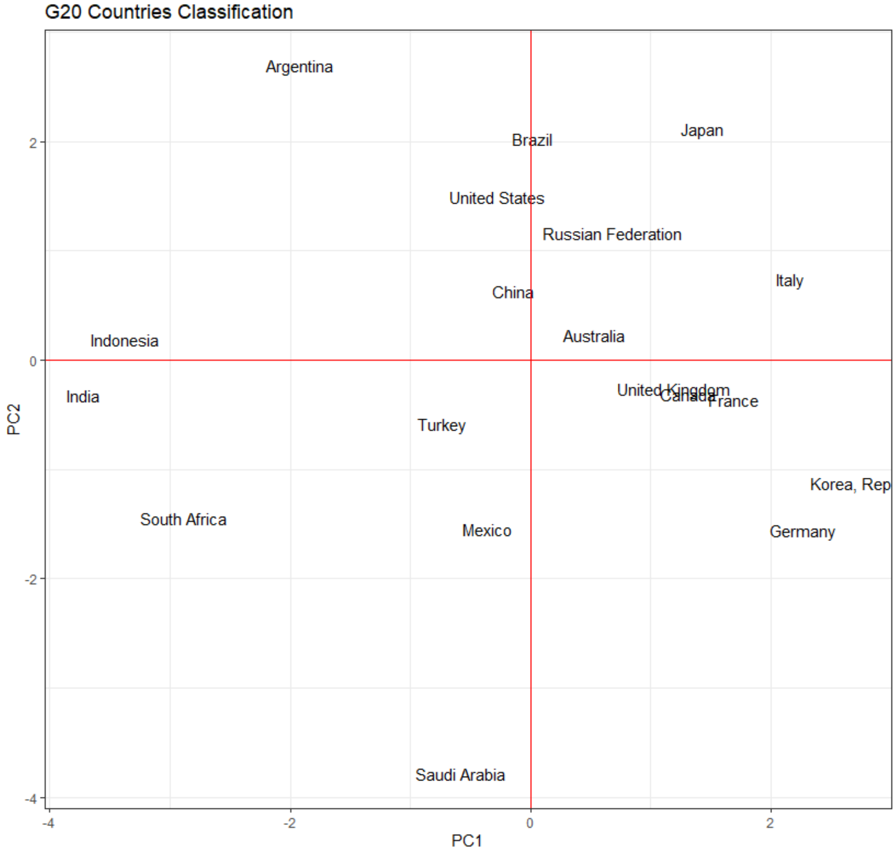


Based on the scree plot above, the sufficient amount of PC should be 3 PCs.
Besides scree plot of Eigenvalue, cumulative variance also useful to decide the sufficient amount of the PC, as below:

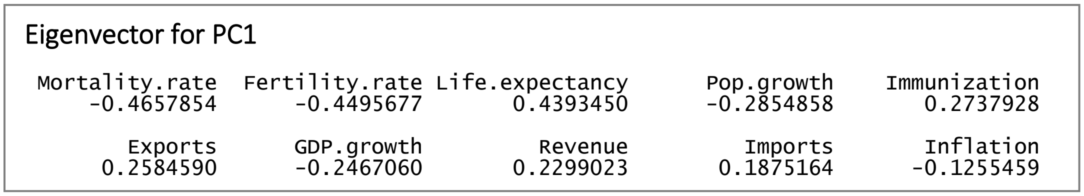


Cumulative variance represents the portion that can be explained through the number of PC. On the graph above shows than 3 PCs can represents for around 70% of the variables, which the value is very sufficient. But actually, even only with 2 PCs, 60% of variables also can be explained. So, for this paper the amount of PC that is chosen is 2 PCs.

The mapping results of G-20 countries based on PC1 and PC2 is presents through scatter plot below:



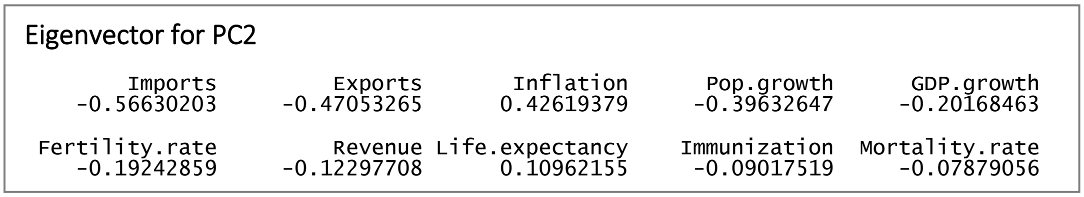
Based on above graph, now there are 4 groups of country that is formed into 4 quadrants. To interpretant the character of every quadrant, need to check eigenvectors result of each variable to PC1 and PC 2.
The eigenvectors results, which already sorted from the most impactful to the least impactful, are shown as below:



Eigenvector for PC1

```
  Mortality.rate  Fertility.rate Life.expectancy      Pop.growth    Immunization
      -0.4657854      -0.4495677       0.4393450      -0.2854858       0.2737928

         Exports      GDP.growth         Revenue         Imports       Inflation
       0.2584590      -0.2467060       0.2299023       0.1875164      -0.1255459
```



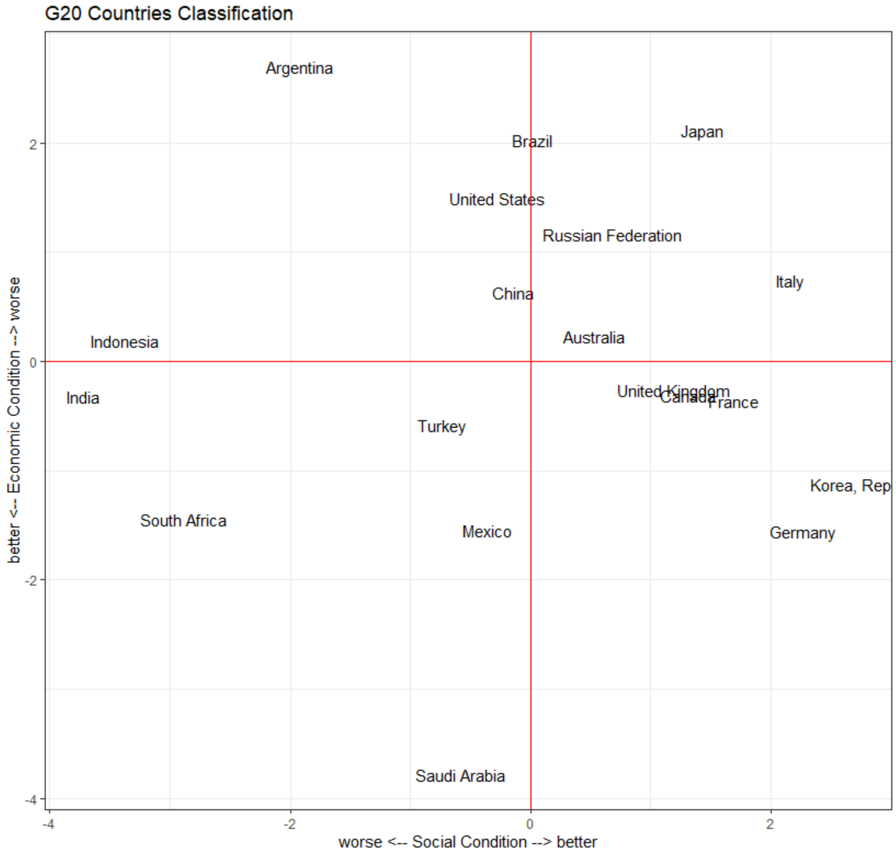
Eigenvector for PC2

```
         Imports         Exports       Inflation      Pop.growth      GDP.growth
     -0.56630203     -0.47053265      0.42619379     -0.39632647     -0.20168463

  Fertility.rate         Revenue Life.expectancy    Immunization  Mortality.rate
     -0.19242859     -0.12297708      0.10962155     -0.09017519     -0.07879056
```

Output of R above shows that PC1 stands for social indicator that is explained mostly by mortality rate, fertility rate, and life expectancy. Left side means worse social indicator, and right side means better social indicator.
While PC2 stands for economy indicator that is explained mostly by imports, exports, and inflation rate of its country. Upper side means worse economy condition (low imports, exports, and high inflation rate), while lower side means better economy condition.

Hence, based on those interpretation of PCs, the classification figure can be completed as below:



The graph above provide information that Indonesia condition of economy in term of exports, imports, and inflation is in good condition, while still need to improve the social condition a lot.

In overall figure, condition of G-20 countries are mostly in good condition for social indicator, while varies condition of economy indicator.

D. Enclosure

1. Dataset

Country	Exports	Fertility rate	GDP growth	Immunization	Imports	Inflation	Life expectancy	Mortality rate	Pop growth	Revenue
Argentina	10.705652	2.301	2.7311598	89	11.780574	26.579992	76.068	11.5	1.0780013	22.14908
Australia	19.987417	1.814	2.3360755	95	21.512295	-0.701819	82.4	3.9	1.4392167	24.464169
Brazil	12.900191	1.753	-3.5457634	96	14.053435	7.5661564	74.994	15.7	0.8388483	28.520095
Canada	31.680302	1.563	0.6899065	89	34.158162	-0.9024047	81.9	5.3	0.7463395	17.892394
China	21.443274	1.665	6.9053167	99	18.185733	0.0626994	75.928	10.7	0.5081367	16.110033
France	30.592622	1.96	1.1129123	91	31.159072	1.1382488	82.321951	4.1	0.417226	44.287474
Germany	46.859744	1.5	1.7389676	97	38.852227	1.9790965	80.641463	3.9	0.8657026	28.09473
India	19.813189	2.295	7.9962538	87	22.109725	2.2795881	68.607	43.6	1.1166331	12.41521
Indonesia	21.160179	2.389	4.8763223	75	20.777461	3.9802427	70.768	28	1.2674656	12.966699
Italy	29.914943	1.35	0.9239925	85	27.016364	0.9330718	82.543902	3.4	-0.0963761	38.320779
Japan	17.610976	1.45	1.222921	96	18.030045	2.1453968	83.793902	3	-0.106125	12.626727
Korea, Rep.	45.33669	1.239	2.7902362	98	38.375432	2.3946589	82.02439	3.5	0.5272885	26.303608
Mexico	34.563331	2.215	3.2879916	97	36.602984	2.787393	74.904	14.8	1.2411777	19.038663
ssian Federat	28.707262	1.777	-2.3077335	98	20.643292	7.5876917	71.183415	8.5	0.1925586	24.470014
Saudi Arabia	33.321171	2.507	4.1064089	98	38.753825	-16.908525	74.651	8.5	2.5567844	24.97212
South Africa	30.156046	2.484	1.2795493	86	31.444103	5.1213719	62.649	37.7	1.5289263	31.696948
Turkey	23.34593	2.101	6.0858866	97	25.953935	7.826942	76.532	12.6	1.6660517	30.653235
nited Kingdo	27.407918	1.8	2.3491214	93	28.831246	0.4354304	80.956098	4.5	0.7949679	34.609479
United States	12.432131	1.8435	2.8809105	92	15.294009	1.0693417	78.690244	6.8	0.7373354	19.077025

2. R Programming Command

```
x=read.csv('C:/Users/Ardya Novi/OneDrive/Documents/Back up ID & Doc/MMT Business Analytics/Big
Data Analysis/Task 1 - EDA/WorldIndicatorG20.csv')

#Statistics Descriptive
rata=cbind(0)
for (i in 2:length(x))
{
  rata[i-1]<-mean(x[,i])
}
varians=cbind(0)
for (i in 2:length(x))
{
  varians[i-1]<-var(x[,i])
}
maks=cbind(0)
for (i in 2:length(x))
{
  maks[i-1]<-max(x[,i])
}
min=cbind(0)
for (i in 2:length(x))
{
  min[i-1]<-min(x[,i])
}
c1<-c("var","exports","fertility","GDP Growth","immunization","imports","inflation","life
expect","mortality","pop growth","revenue")
c2<-c("rata",rata)
c3<-c("varians",varians)
c4<-c("maks",maks)
c5<-c("min",min)
m1<-matrix(c(c1,c2,c3,c4,c5), nrow=5, ncol=11, byrow=TRUE)
print(m1)

#histogram & boxplot
par(mfrow=c(5,2))
grid<-seq(9.5,49.5,by=5)
hist(x[,2],breaks=grid,main="Exports of goods and services (% of GDP)",xlab="Exports",col="light
yellow")
boxplot(x[,2],horizontal=TRUE)
grid<-seq(1.0,3.0,by=0.5)
hist(x[,3],breaks=grid,main="Fertility rate (births per woman)",xlab="Fertility Rate",col="light
yellow")
boxplot(x[,3],horizontal=TRUE)
```

```
grid<-seq(-4.0,8.5,by=1)
hist(x[,4],breaks=grid,main="GDP growth (annual %)",xlab="GDP Growth",col="light yellow")
boxplot(x[,4],horizontal=TRUE)
grid<-seq(70.0,100.0,by=5)
hist(x[,5],breaks=grid,main="Immunization (% of children ages 12-23
months)",xlab="Immunization",col="light yellow")
boxplot(x[,5],horizontal=TRUE)
grid<-seq(10.0,40.5,by=5)
hist(x[,6],breaks=grid,main="Imports of goods and services (% of GDP)",xlab="Imports",col="light
yellow")
boxplot(x[,6],horizontal=TRUE)

par(mfrow=c(5,2))
grid<-seq(-20.0,30.0,by=5)
hist(x[,7],breaks=grid,main="Inflation (annual %)",xlab="Inflation",col="light yellow")
boxplot(x[,7],horizontal=TRUE)
grid<-seq(60.0,85.0,by=5)
hist(x[,8],breaks=grid,main="Life expectancy at birth, total (years)",xlab="Life
Expectancy",col="light yellow")
boxplot(x[,8],horizontal=TRUE)
grid<-seq(0.0,45.0,by=5)
hist(x[,9],breaks=grid,main="Mortality rate, under-5 (per 1,000 live births)",xlab="Mortality
Rate",col="light yellow")
boxplot(x[,9],horizontal=TRUE)
grid<-seq(-0.5,3.0,by=0.5)
hist(x[,10],breaks=grid,main="Population growth (annual %)",xlab="Population Growth",col="light
yellow")
boxplot(x[,10],horizontal=TRUE)
grid<-seq(10.0,46.0,by=5)
hist(x[,11],breaks=grid,main="Revenue (% of GDP)",xlab="Revenue",col="light yellow")
boxplot(x[,11],horizontal=TRUE)

#bar chart
library(ggplot2)
p1<-ggplot(x, aes(x=reorder(Country, Exports), y=Exports))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p2<-ggplot(x, aes(x=reorder(Country, Fertility.rate), y=Fertility.rate))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p3<-ggplot(x, aes(x=reorder(Country, GDP.growth), y=GDP.growth))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p4<-ggplot(x, aes(x=reorder(Country, Immunization), y=Immunization))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p5<-ggplot(x, aes(x=reorder(Country, Imports), y=Imports))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p6<-ggplot(x, aes(x=reorder(Country, Inflation), y=Inflation))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p7<-ggplot(x, aes(x=reorder(Country, Life.expectancy), y=Life.expectancy))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p8<-ggplot(x, aes(x=reorder(Country, Mortality.rate), y=Mortality.rate))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p9<-ggplot(x, aes(x=reorder(Country, Pop.growth), y=Pop.growth))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
p10<-ggplot(x, aes(x=reorder(Country, Revenue), y=Revenue))+
  geom_bar(stat="identity", fill="light blue")+
  theme_bw()+
  coord_flip()
library(gridExtra)
grid.arrange(p1,p2,p3, ncol=1, nrow=3)
grid.arrange(p4,p5,p6, ncol=1, nrow=3)
grid.arrange(p7,p8,p9, ncol=1, nrow=3)
grid.arrange(p10, ncol=1, nrow=3)

#correlation checking
par(mfrow=c(1,1))
library(corrplot)
mcor=cor(x[-1])
corrplot(mcor)

#coding data into interval
q_exp <- cbind(quantile(x$Exports))
exp_code<-c(1:19)
for (i in 1:19){
  if (x$Exports[i] < q_exp[2]){
    exp_code[i] = 1}
  else if ((x$Exports[i] > q_exp[2]) & (x$Exports[i]) < q_exp[4]){
    exp_code[i] = 2}
  else {exp_code[i] = 3}
}

q_imp <- cbind(quantile(x$Imports))
imp_code<-c(1:19)
for (i in 1:19){
  if (x$Imports[i] < q_imp[2]){
    imp_code[i] = 1}
  else if ((x$Imports[i] > q_imp[2]) & (x$Imports[i]) < q_imp[4]){
    imp_code[i] = 2}
  else {imp_code[i] = 3}
}

q_fertile <- cbind(quantile(x$Fertility.rate))
fertile_code <- c(1:19)
for (i in 1:19){
  if (x$Fertility.rate[i] < q_fertile[2]){
    fertile_code[i] = 1}
  else if ((x$Fertility.rate[i] > q_fertile[2]) & (x$Fertility.rate[i]) < q_fertile[4]){
    fertile_code[i] = 2}
  else {fertile_code[i] = 3}
}

q_life <- cbind(quantile(x$Life.expectancy))
life_code <- c(1:19)
for (i in 1:19){
  if (x$Life.expectancy[i] < q_life[2]){
    life_code[i] = 1}
  else if ((x$Life.expectancy[i] > q_life[2]) & (x$Life.expectancy[i]) < q_life[4]){
    life_code[i] = 2}
  else {life_code[i] = 3}
}

q_mortal <- cbind(quantile(x$Mortality.rate))
mortal_code <- c(1:19)
for (i in 1:19){
  if (x$Mortality.rate[i] < q_mortal[2]){
    mortal_code[i] = 1}
  else if ((x$Mortality.rate[i] > q_mortal[2]) & (x$Mortality.rate[i]) < q_mortal[4]){
    mortal_code[i] = 2}
  else {mortal_code[i] = 3}
}

q_pop <- cbind(quantile(x$Pop.growth))
pop_code <- c(1:19)
for (i in 1:19){
  if (x$Pop.growth[i] < q_pop[2]){
    pop_code[i] = 1}
  else if ((x$Pop.growth[i] > q_pop[2]) & (x$Pop.growth[i]) < q_pop[4]){
    pop_code[i] = 2}
  else {pop_code[i] = 3}
}

#grouping chart -> export + import
library(dplyr)
a <- data.frame(cbind(exp_code, imp_code))
exp <- c(rep("low_exp",3), rep("med_exp",3), rep("high_exp",3))
imp <- rep(c("low_imp","med_import","high_imp"),3)
a.value <- c(sum(a$exp_code == 1 & a$imp_code == 1),sum(a$exp_code == 1 & a$imp_code == 2),
sum(a$exp_code == 1 & a$imp_code == 3), sum(a$exp_code == 2 & a$imp_code == 1), sum(a$exp_code ==
2 & a$imp_code == 2), sum(a$exp_code == 2 & a$imp_code == 3), sum(a$exp_code == 3 & a$imp_code ==
1), sum(a$exp_code == 3 & a$imp_code == 2), sum(a$exp_code == 3 & a$imp_code == 3))
```

```
a.df <- data.frame(exp, imp, a.value)
ggplot(a.df, aes(fill=imp, y=a.value, x=exp)) +
  geom_bar(position="dodge", stat="identity")+
  theme_bw()

#grouping chart -> life expectancy + fertility rate
b <- data.frame(cbind(fertile_code, life_code))
fertile <- c(rep("low_fertile",3), rep("med_fertile",3), rep("high_fertile",3))
life <- rep(c("low_life_exp","med_life_exp","high_life_exp"),3)
b.value <- c(sum(b$fertile_code == 1 & b$life_code == 1),sum(b$fertile_code == 1 & b$life_code ==
2), sum(b$fertile_code == 1 & b$life_code == 3), sum(b$fertile_code == 2 & b$life_code == 1),
sum(b$fertile_code == 2 & b$life_code == 2), sum(b$fertile_code == 2 & b$life_code == 3),
sum(b$fertile_code == 3 & b$life_code == 1), sum(b$fertile_code == 3 & b$life_code == 2),
sum(b$fertile_code == 3 & b$life_code == 3))
b.df <- data.frame(fertile, life, b.value)
ggplot(b.df, aes(fill=life, y=b.value, x=fertile)) +
  geom_bar(position="dodge", stat="identity")+
  theme_bw()

#grouping chart -> mortality rate + life expectancy
c <- data.frame(cbind(mortal_code, life_code))
mortal <- c(rep("low_mortal",3), rep("med_mortal",3), rep("high_mortal",3))
life <- rep(c("low_life_exp","med_life_exp","high_life_exp"),3)
c.value <- c(sum(c$mortal_code == 1 & c$life_code == 1),sum(c$mortal_code == 1 & c$life_code ==
2), sum(c$mortal_code == 1 & c$life_code == 3), sum(c$mortal_code == 2 & c$life_code == 1),
sum(c$mortal_code == 2 & c$life_code == 2), sum(c$mortal_code == 2 & c$life_code == 3),
sum(c$mortal_code == 3 & c$life_code == 1), sum(c$mortal_code == 3 & c$life_code == 2),
sum(c$mortal_code == 3 & c$life_code == 3))
c.df <- data.frame(mortal, life, c.value)
ggplot(c.df, aes(fill=life, y=c.value, x=mortal)) +
  geom_bar(position="dodge", stat="identity")+
  theme_bw()

#grouping chart -> population growth + fertility rate
d <- data.frame(cbind(pop_code, fertile_code))
pop <- c(rep("low_pop",3), rep("med_pop",3), rep("high_pop",3))
fertile <- rep(c("low_fertile","med_fertile","high_fertile"),3)
d.value <- c(sum(d$pop_code == 1 & d$fertile_code == 1),sum(d$pop_code == 1 & d$fertile_code ==
2), sum(d$pop_code == 1 & d$fertile_code == 3), sum(d$pop_code == 2 & d$fertile_code == 1),
sum(d$pop_code == 2 & d$fertile_code == 2), sum(d$pop_code == 2 & d$fertile_code == 3),
sum(d$pop_code == 3 & d$fertile_code == 1), sum(d$pop_code == 3 & d$fertile_code == 2),
sum(d$pop_code == 3 & d$fertile_code == 3))
d.df <- data.frame(pop, fertile, d.value)
ggplot(d.df, aes(fill=fertile, y=d.value, x=pop)) +
  geom_bar(position="dodge", stat="identity") +
  theme_bw()

#grouping chart -> mortality rate + fertility rate
e <- data.frame(cbind(mortal_code, fertile_code))
fertility <- rep(c("low_fertile", "med_fertile", "high_fertile"),3)
e.value <- c(sum(e$mortal_code == 1 & e$fertile_code == 1),sum(e$mortal_code == 1 & e$fertile_code
== 2), sum(e$mortal_code == 1 & e$fertile_code == 3), sum(e$mortal_code == 2 & e$fertile_code ==
1), sum(e$mortal_code == 2 & e$fertile_code == 2), sum(e$mortal_code == 2 & e$fertile_code == 3),
sum(e$mortal_code == 3 & e$fertile_code == 1), sum(e$mortal_code == 3 & e$fertile_code == 2),
sum(e$mortal_code == 3 & e$fertile_code == 3))
e.df <- data.frame(mortal, e.fertile, e.value)
ggplot(c.df, aes(fill=mortal, y=e.value, x=fertility)) +
  geom_bar(position="dodge", stat="identity") +
  theme_bw()

#PCA
x.pca<-prcomp(x[,c(2:11)],center=TRUE, scale=TRUE)
summary(x.pca)
print(x.pca)

par(mfrow=c(2,1))

screeplot(x.pca, type = "l", npcs = 10, main = "Screeplot")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"),
       col=c("red"), lty=5, cex=0.6)

cumpro <- cumsum(x.pca$sdev^2 / sum(x.pca$sdev^2))
plot(cumpro[0:15], xlab = "PC #", ylab = "Amount of explained variance", main = "Cumulative
variance plot")
abline(v = 3, col="blue", lty=5)
abline(h = 0.7239, col="blue", lty=5)
legend("topleft", legend=c("Cut-off @ PC3"),
       col=c("blue"), lty=5, cex=0.6)

#PCA grouping plot
par(mfrow=c(1,1))
library(ggplot2)
pca.data<-data.frame(Sample=x[,1],X=x.pca$x[,1],Y=x.pca$x[,2])
ggplot(data=pca.data,aes(x=X,y=Y,label=Sample))+
geom_text()+
xlab(paste("PC1"))+
ylab(paste("PC2"))+
theme_bw()+
ggtitle("G20 Countries Classification")+
geom_vline(xintercept=0, colour="red")+
geom_hline(yintercept=0, colour="red")

#checking interpretation
loading_score<-x.pca$rotation[,1]
var_score<-abs(loading_score)
var_score_rank<-sort(var_score, decreasing=TRUE)
top_10_var<-names(var_score_rank[1:10])
top_10_var
x.pca$rotation[top_10_var,1]

loading_score<-x.pca$rotation[,2]
var_score<-abs(loading_score)
var_score_rank<-sort(var_score, decreasing=TRUE)
top_10_var<-names(var_score_rank[1:10])
top_10_var
x.pca$rotation[top_10_var,2]

#PCA grouping plot with interpretation
par(mfrow=c(1,1))
library(ggplot2)
pca.data<-data.frame(Sample=x[,1],X=x.pca$x[,1],Y=x.pca$x[,2])
ggplot(data=pca.data,aes(x=X,y=Y,label=Sample))+
  geom_text()+
  xlab(paste("worse <-- Social Condition --> better"))+
  ylab(paste("better <-- Economic Condition --> worse"))+
  theme_bw()+
  ggtitle("G20 Countries Classification")+
  geom_vline(xintercept=0, colour="red")+
  geom_hline(yintercept=0, colour="red")
```