Assignment 1: Group Shannon

Code

Prithvi Poddar (17191)

Hide

```
library(ggplot2)
library(tidyverse)
library(dplyr)
library(tidyr)
library(Hmisc)
library(plyr)
library(readr)
library(GGally)
library(caret)
```

Setting the working directory

Hide

```
setwd("~/Documents/iiser works/ECS/8th_sem/Bio_stats")
```

Comparison among developed, developing and under developed countries

Here, we take 10 countries from each **Developed**, **Developing** and **Under Developed** countries. We then find the year-wise mean for the 10 countries and plot the values from the year **1800 to 2100**. Then we plot the graphs for these three groups and find out some patterns or test various hypothesis on them.
The countries in the lists are:
Developed countries : Norway , Ireland , Switzerland , Hong Kong , Iceland , Germany , Sweden , Australia, Netherlands , Denmark
Developing countries : Algeria , Lebanon , Fiji , Moldova , Maldives , Tunisia , Saint Vincent and the Grenadines , Suriname , Mongolia , Botswana
Under Developed Countries : Eritrea , Mozambique , Burkina Faso , Sierra Leone , Mali , Burundi , South Sudan , Chad , Central African Republic, Niger
Above set of countries have been selected based on their Human Development Index rankings published in year 2020 by United Nations Development Program (http://www.hdr.undp.org/ (http://www.hdr.undp.org/)).

First we load all the datasets

Hide

```
read.csv("child_mortality_0_5_year_olds_dying_per_1000_born.csv", header = T, check.n
ames = F) -> ChildMortality
read.csv("children_per_woman_total_fertility.csv", header = T, check.names = F) -> Ch
ildrenPerWomen
read.csv("income_per_person_gdppercapita_ppp_inflation_adjusted.csv", header = T, che
ck.names = F)-> IncomePerPerson
read.csv("life_expectancy_years.csv", header = T, check.names = F)-> LifeExpectancy
read.csv("population_total.csv", header = T, check.names = F)-> Population
```

Next, we extract the grouped data of the countries based on whether they are developed, developing or under developed. To do that, we first write a funtion as follows:

Hide

```
Get_country_data <- function(Dataset, a, b, c, d, e, f, g, h, i, j) {
  Dataset[Dataset$country %in% c(a, b, c, d, e, f, g, h, i, j), ]}
```

Getting the data for **Developed Countries**

Hide

```
a<-"Norway"
b<-"Ireland"
c<-"Switzerland"
d<-"Finland"
e<-"Iceland"
f<-"Germany"
g<-"Sweden"
h<-"Australia"
i<-"Netherlands"
j<-"Denmark"

TopLife <- Get_country_data(LifeExpectancy, a, b, c, d, e, f, g, h, i, j)
TopPopulation <- Get_country_data(Population, a, b, c, d, e, f, g, h, i, j)
TopIncome <- Get_country_data(IncomePerPerson, a, b, c, d, e, f, g, h, i, j)
TopChildren <- Get_country_data(ChildrenPerWomen, a, b, c, d, e, f, g, h, i, j)
TopMortality <- Get_country_data(ChildMortality, a, b, c, d, e, f, g, h, i, j)
```

Getting the data for **Developing Countries**

Hide

```
a<-"Algeria"
b<-"Lebanon"
c<-"Fiji"
d<-"Moldova"
e<-"Maldives"
f<-"Tunisia"
g<-"Saint Vincent and the Grenadines"
h<-"Suriname"
i<-"Mongolia"
j<-"Botswana"

MidLife <- Get_country_data(LifeExpectancy, a, b, c, d, e, f, g, h, i, j)
MidPopulation <- Get_country_data(Population, a, b, c, d, e, f, g, h, i, j)
MidIncome <- Get_country_data(IncomePerPerson, a, b, c, d, e, f, g, h, i, j)
MidChildren <- Get_country_data(ChildrenPerWomen, a, b, c, d, e, f, g, h, i, j)
MidMortality <- Get_country_data(ChildMortality, a, b, c, d, e, f, g, h, i, j)
```

Getting the data for **Under Developed Countries**

Hide

```
a<-"Eritrea"
b<-"Mozambique"
c<-"Burkina Faso"
d<-"Sierra Leone"
e<-"Mali"
f<-"Burundi"
g<-"South Sudan"
h<-"Chad"
i<-"Central African Republic"
j<-"Niger"

LowLife <- Get_country_data(LifeExpectancy, a, b, c, d, e, f, g, h, i, j)
LowPopulation <- Get_country_data(Population, a, b, c, d, e, f, g, h, i, j)
LowIncome <- Get_country_data(IncomePerPerson, a, b, c, d, e, f, g, h, i, j)
LowChildren <- Get_country_data(ChildrenPerWomen, a, b, c, d, e, f, g, h, i, j)
LowMortality <- Get_country_data(ChildMortality, a, b, c, d, e, f, g, h, i, j)
```

DataSet: Life Expentency

Finding the year wise mean for all the countries

Hide

```
TopLifeMean <- sapply(TopLife[,2:302], mean)
MidLifeMean <- sapply(MidLife[,2:302], mean)
LowLifeMean <- sapply(LowLife[,2:302], mean)
```

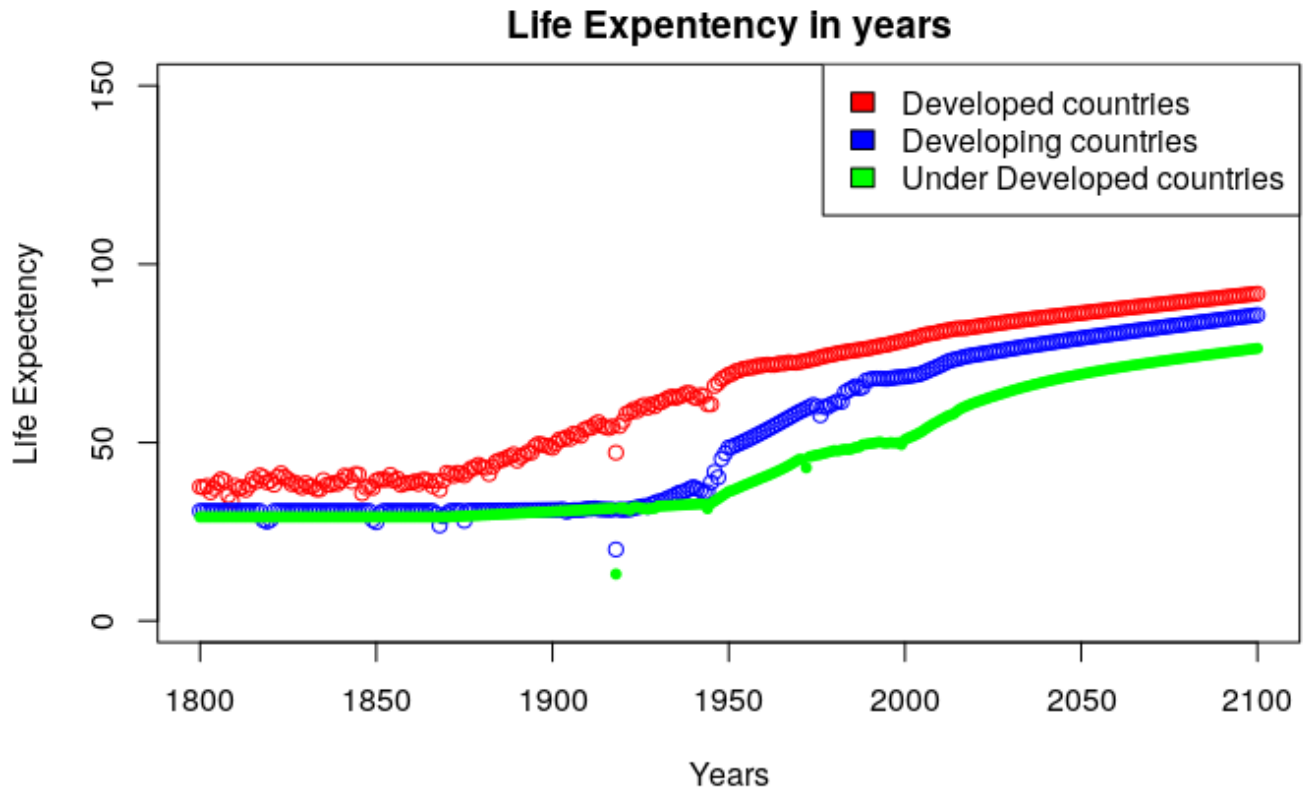
Plotting the means for all the countries for all the 3 groups in the same plot, to compare the the trends

Hide

```
plot(c(1800:2100), TopLifeMean, col="red", main = "Life Expentency in years", pch=1,
 ylim = c(0,150), xlab = "Years", ylab = "Life Expectency")
points(c(1800:2100), MidLifeMean, col="blue")
```

Hide

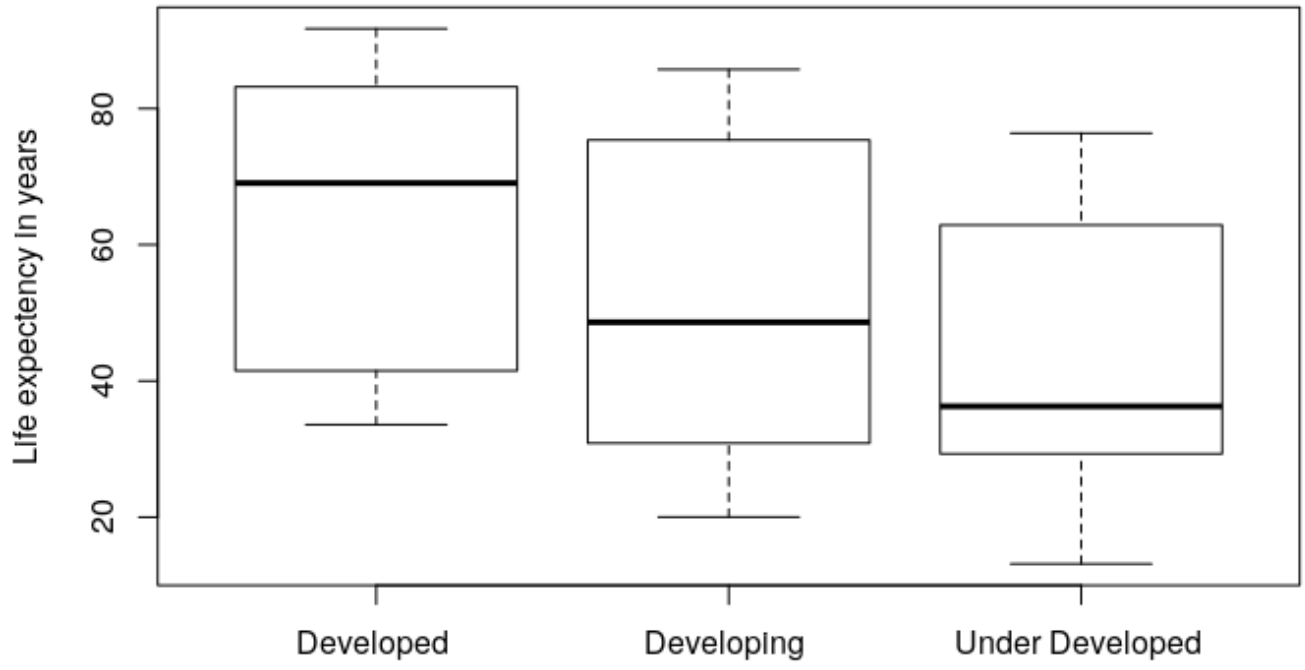
```
points(c(1800:2100), LowLifeMean, pch=20, col="green")
legend(x = "topright", legend = c("Developed countries ","Developing countries","Unde
r Developed countries"), fill = c("red","blue","green"))
```



Making box plot to compare the means over the course of years

Hide

```
boxplot(TopLifeMean, MidLifeMean, LowLifeMean, ylab = "Life expectency in years", nam
es=c("Developed ","Developing ","Under Developed "))
```



Analysis:

In the initial years, the life expectancy for developing and under developed countries were the same with the developed countries slightly above them. From around 1860, the life expectency of developed countries started to increase. From 1945 onwards, the life expectancy for developing and under developed countries also started to increase as the World War 2 just ended.

Hide

```
plot(c(1800:2100), TopLifeMean, col="red", main = "Life Expentency in years", pch=1,
 ylim = c(0,150), xlab = "Years", ylab = "Life Expectency")
points(c(1800:2100), MidLifeMean, col="blue")
```

Hide

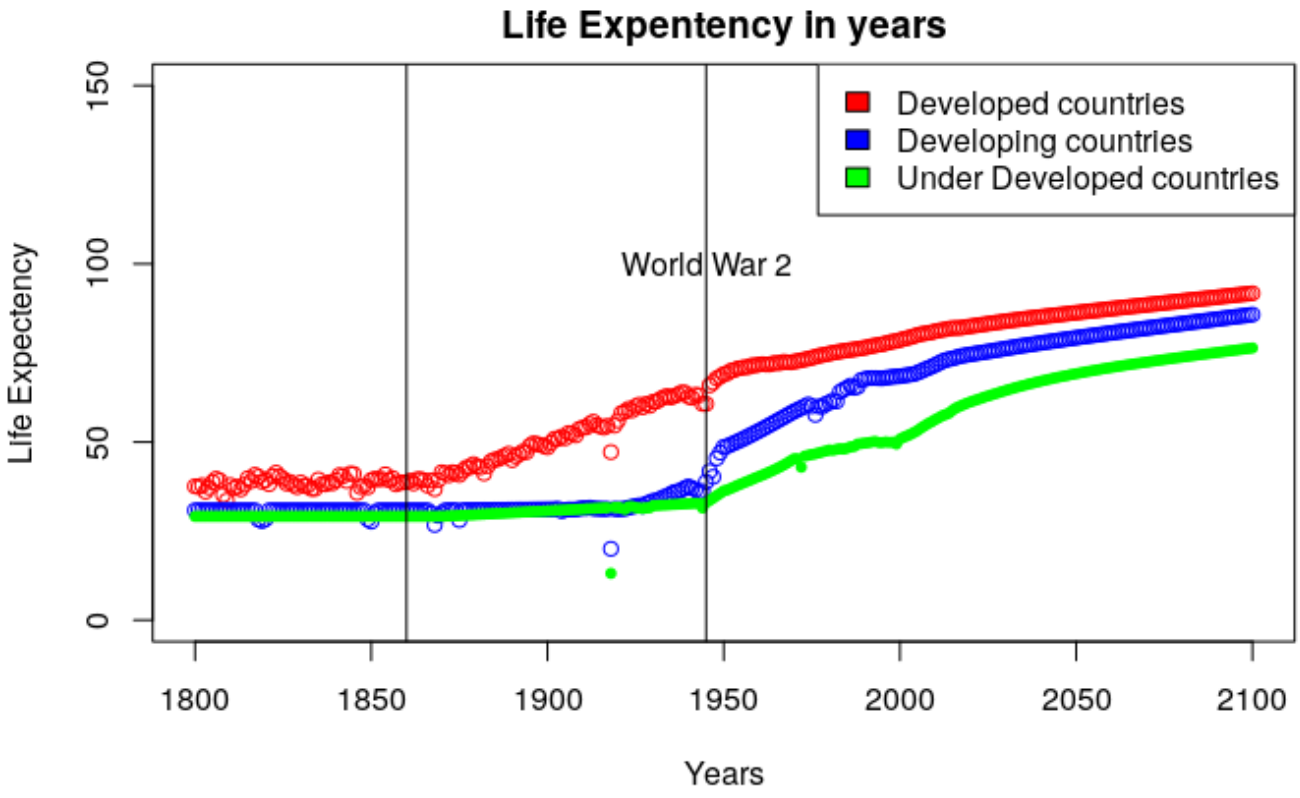
```
points(c(1800:2100), LowLifeMean, pch=20, col="green")
abline(v= 1860)
```

Hide

```
abline(v= 1945)
text(1945, 100, "World War 2")
```

Hide

```
legend(x = "topright", legend = c("Developed countries ","Developing countries","Unde
r Developed countries"), fill = c("red","blue","green"))
```



DataSet: Total population

Finding the year wise mean for all the countries

Hide

```
TopPopulationMean <- sapply(TopPopulation[,2:302], mean)
MidPopulationMean <- sapply(MidPopulation[,2:302], mean)
LowPopulationMean <- sapply(LowPopulation[,2:302], mean)
```

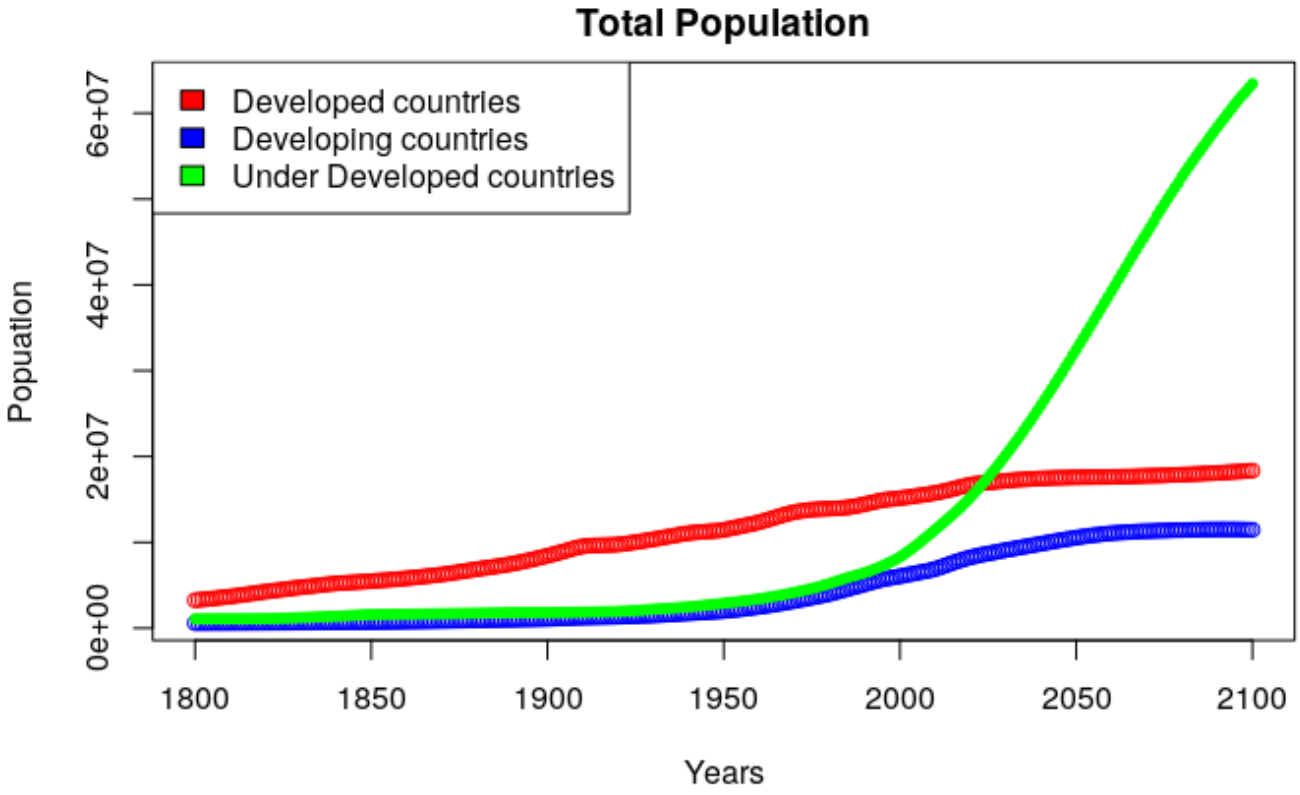
Plotting the means for all the countries for all the 3 groups in the same plot, to compare the the trends

Hide

```
plot(c(1800:2100), TopPopulationMean, col="red", main = "Total Population", pch=1, yl
im = c(1076000,63436000), xlab = "Years", ylab = "Popuation")
points(c(1800:2100), MidPopulationMean, col="blue")
```

Hide

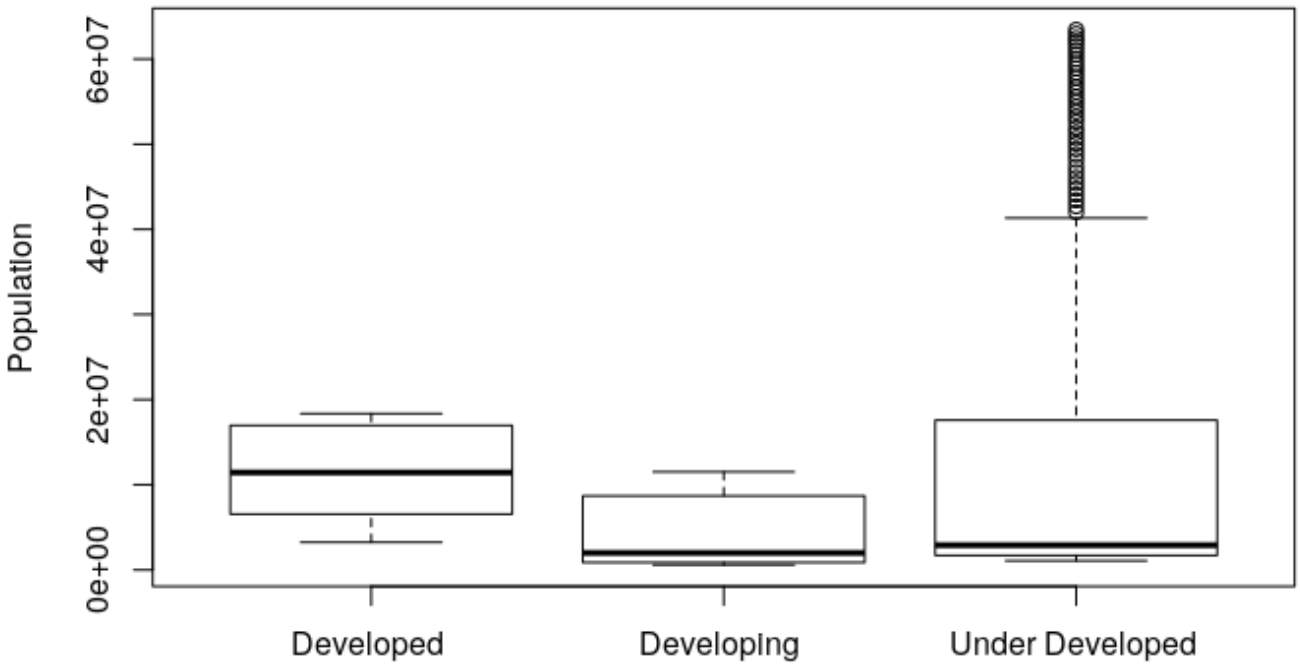
```
points(c(1800:2100), LowPopulationMean, pch=20, col="green")
legend(x = "topleft", legend = c("Developed countries ","Developing countries","Under
Developed countries"), fill = c("red","blue","green"))
```



Making box plot to compare the means over the course of years

Hide

```
boxplot(TopPopulationMean, MidPopulationMean, LowPopulationMean, ylab = "Population",
names=c("Developed ","Developing ","Under Developed "))
```



Analysis:
We see an exponential increase in the population of the under developed coountries in the plot. But the box plot shows us that there are many outliers and the average population is still considerably lower than that of the developed countries. This exponential increase might be due to poverty and lack of education awareness among the people in these countries. Before 1990, the developing and under developed countries had the similar trend in their population.

Hide

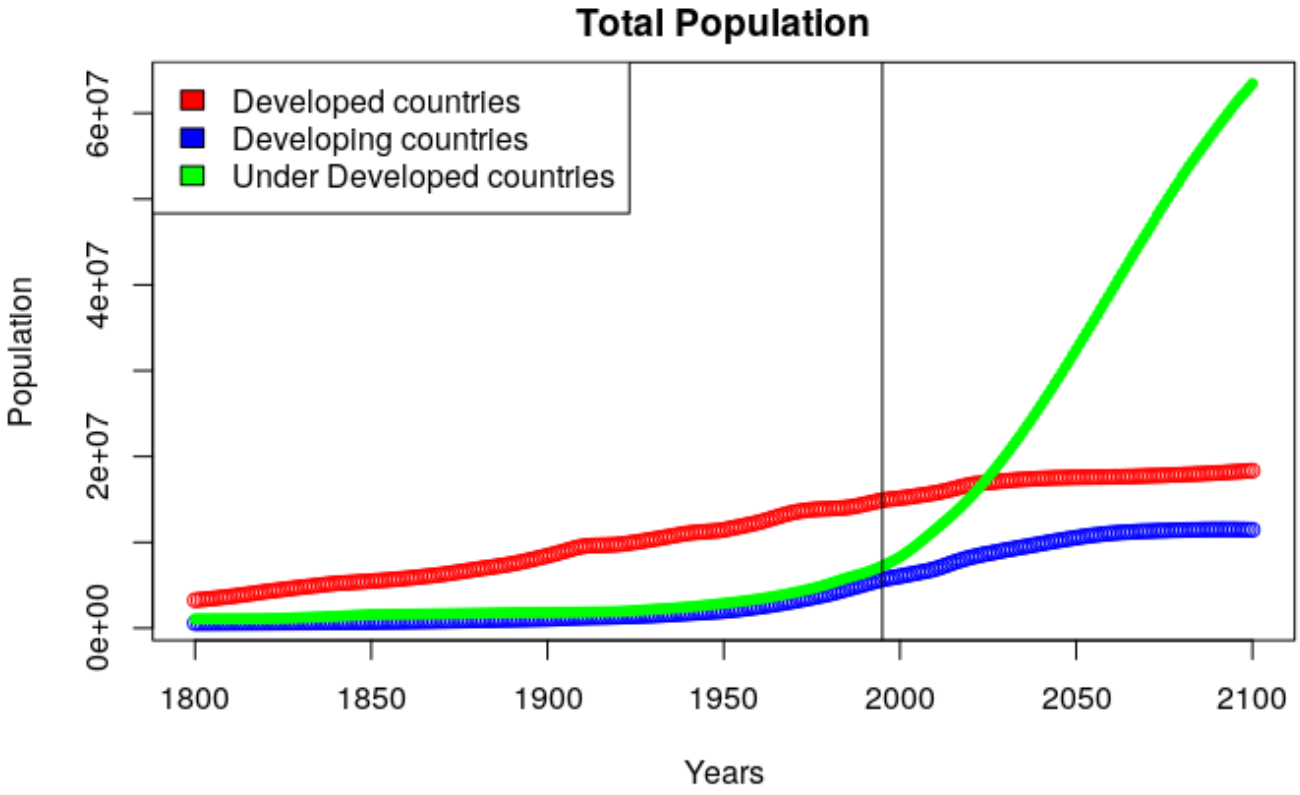
```
plot(c(1800:2100), TopPopulationMean, col="red", main = "Total Population", pch=1, yl
im = c(1076000,63436000), xlab = "Years", ylab = "Population")
points(c(1800:2100), MidPopulationMean, col="blue")
```

Hide

```
points(c(1800:2100), LowPopulationMean, pch=20, col="green")
abline(v = 1995)
```

Hide

```
legend(x = "topleft", legend = c("Developed countries ","Developing countries","Under
Developed countries"), fill = c("red","blue","green"))
```



DataSet: Income per person

Finding the year wise mean for all the countries

Hide

```
TopIncomeMean <- sapply(TopIncome[,2:242], mean)
MidIncomeMean <- sapply(MidIncome[,2:242], mean)
LowIncomeMean <- sapply(LowIncome[,2:242], mean)
```

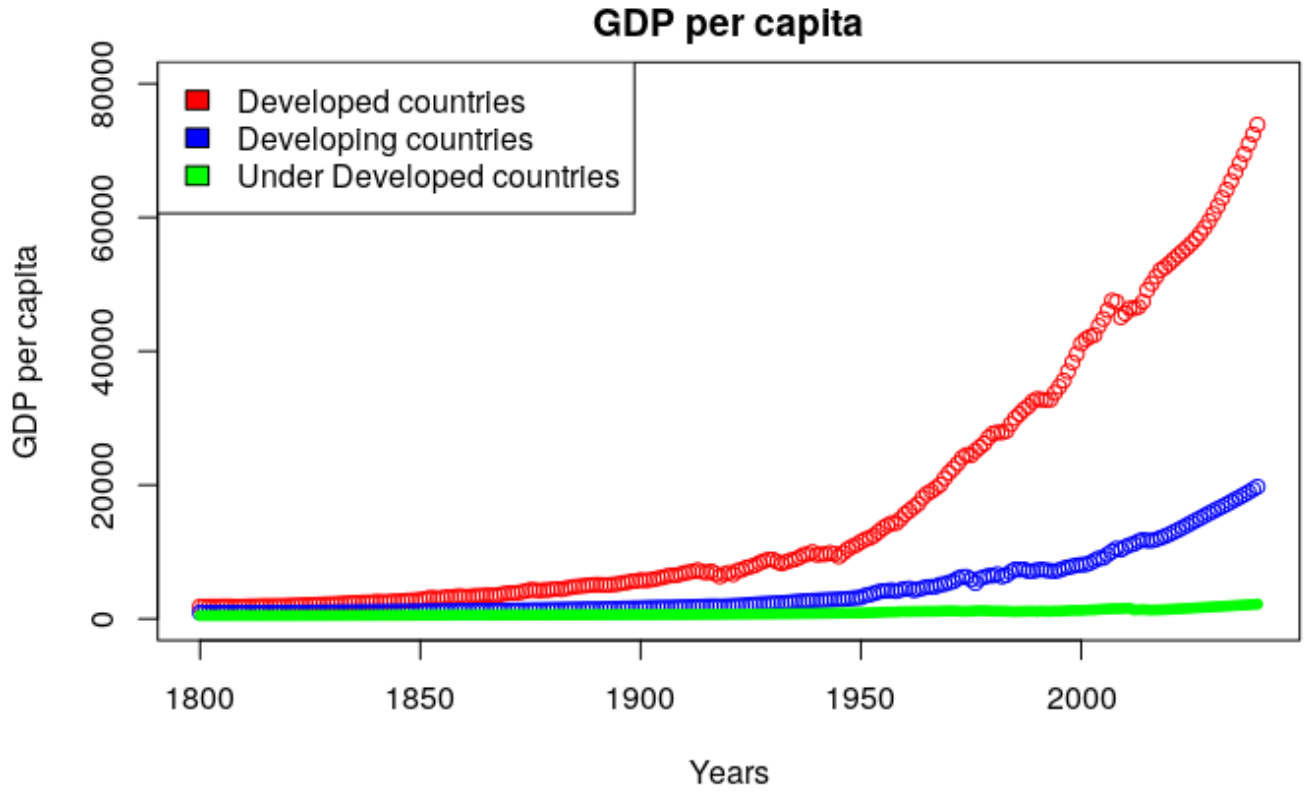
Plotting the means for all the countries for all the 3 groups in the same plot, to compare the the trends

Hide

```
plot(c(1800:2040), TopIncomeMean, col="red", main = "GDP per capita", pch=1, ylim = c
(0,80000), xlab = "Years", ylab = "GDP per capita")
points(c(1800:2040), MidIncomeMean, col="blue")
```

Hide

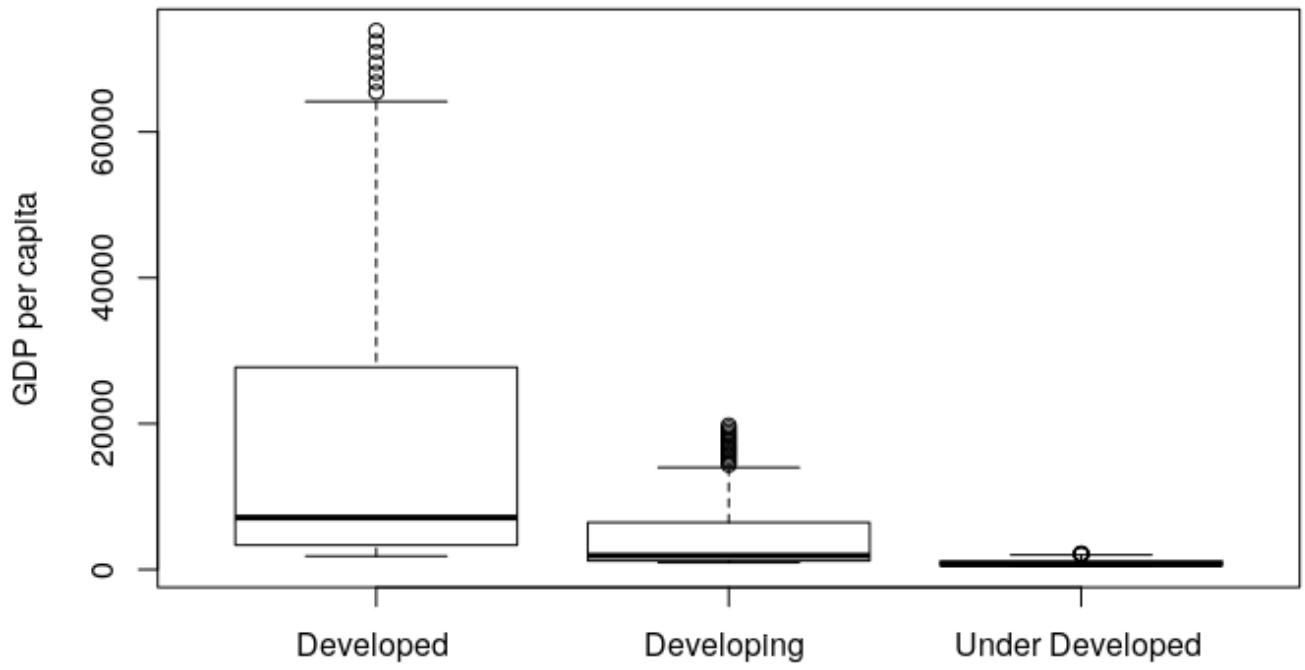
```
points(c(1800:2040), LowIncomeMean, pch=20, col="green")
legend(x = "topleft", legend = c("Developed countries ","Developing countries","Under
Developed countries"), fill = c("red","blue","green"))
```



Making box plot to compare the means over the course of years

Hide

```
boxplot(TopIncomeMean, MidIncomeMean, LowIncomeMean, ylab = "GDP per capita", names=c
("Developed ","Developing ","Under Developed "))
```



Analysis:

We observe that the GDP dips during World War I and World War II

Hide

```
plot(c(1800:2040), TopIncomeMean, col="red", main = "GDP per capita", pch=1, ylim = c
(0,80000), xlab = "Years", ylab = "GDP per capita")
points(c(1800:2040), MidIncomeMean, col="blue")
```

Hide

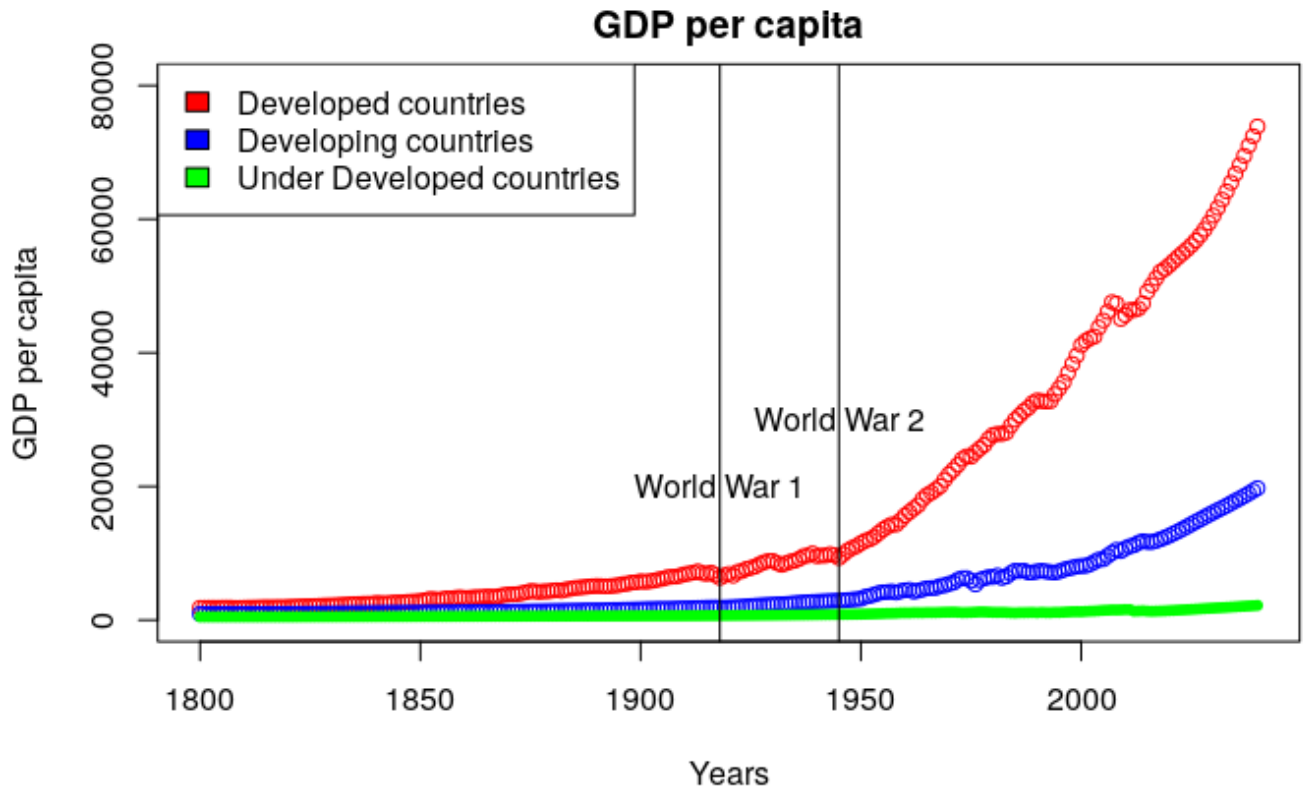
```
points(c(1800:2040), LowIncomeMean, pch=20, col="green")
abline(v= 1918)
```

Hide

```
abline(v= 1945)
text(1918, 20075, "World War 1")
```

Hide

```
text(1945, 30075, "World War 2")
legend(x = "topleft", legend = c("Developed countries ","Developing countries","Under
Developed countries"), fill = c("red","blue","green"))
```



DataSet: Children per woman

Finding the year wise mean for all the countries

Hide

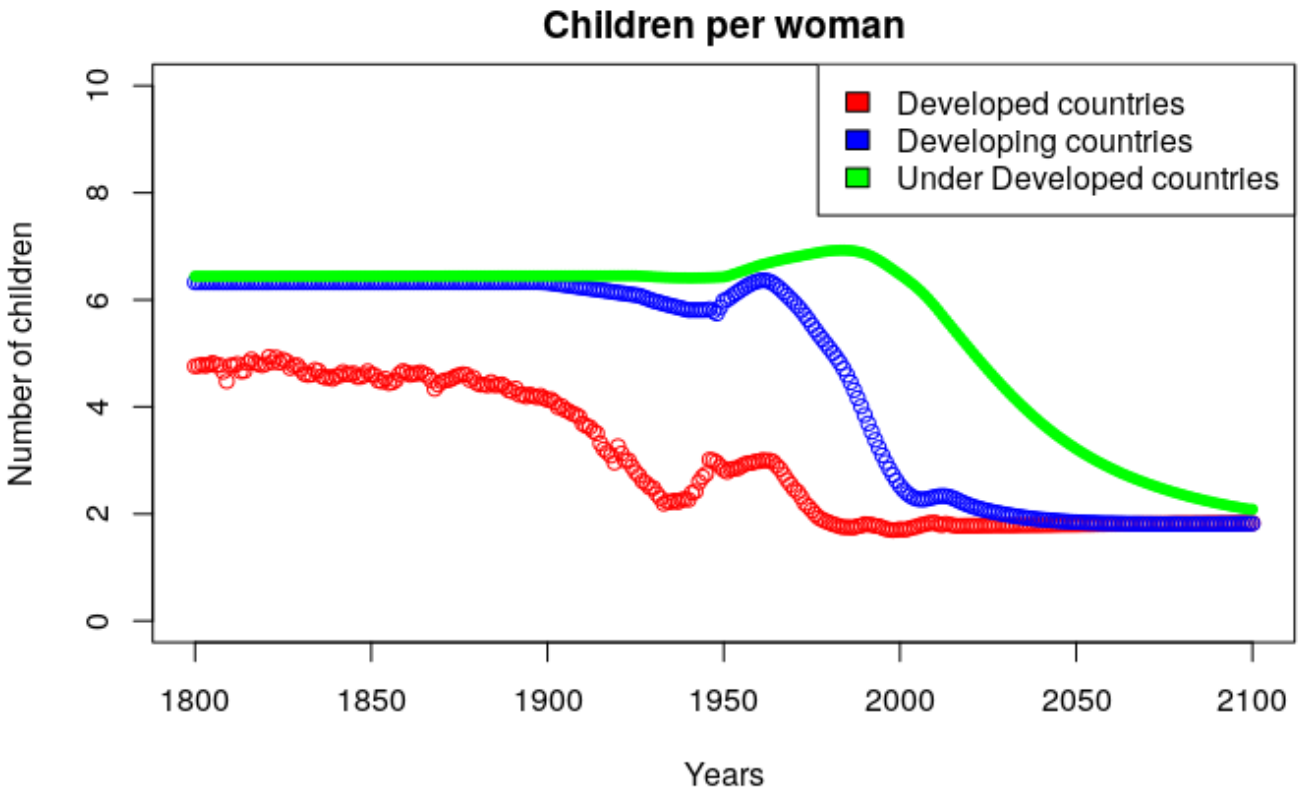
```
TopChildrenMean <- sapply(TopChildren[,2:302], mean)
MidChildrenMean <- sapply(MidChildren[,2:302], mean)
LowChildrenMean <- sapply(LowChildren[,2:302], mean)
```

Hide

```
plot(c(1800:2100), TopChildrenMean, col="red", main = "Children per woman", pch=1, yl
im = c(0,10), xlab = "Years", ylab = "Number of children")
points(c(1800:2100), MidChildrenMean, col="blue")
```

Hide

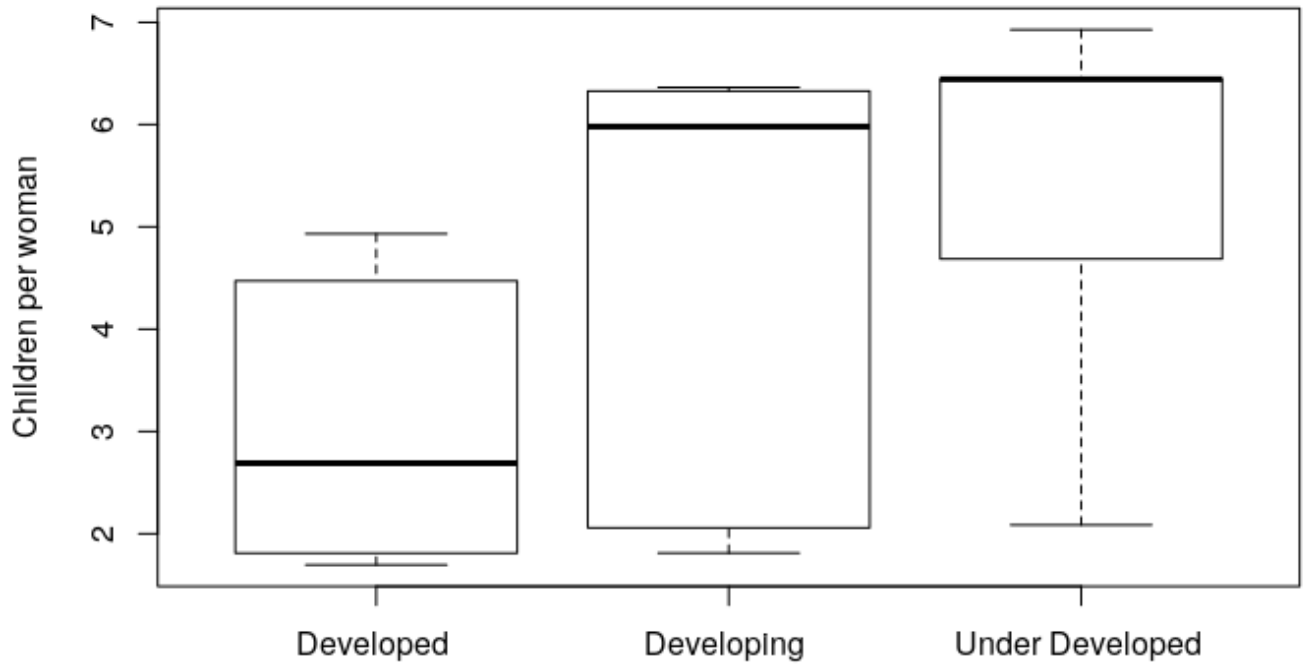
```
points(c(1800:2100), LowChildrenMean, pch=20, col="green")
legend(x = "topright", legend = c("Developed countries ","Developing countries","Unde
r Developed countries"), fill = c("red","blue","green"))
```



Making box plot to compare the means over the course of years

Hide

```
boxplot(TopChildrenMean, MidChildrenMean, LowChildrenMean, ylab = "Children per woma
n", names=c("Developed ","Developing ","Under Developed "))
```



Analysis: We see that on an average, under developed countries have more children per woman than developed and developing countries

DataSet: Child Mortality of 0-5 year olds dying per 1000 born

Finding the year wise mean for all the countries

Hide

```
TopMortalityMean <- sapply(TopMortality[,2:302], mean)
MidMortalityMean <- sapply(MidMortality[,2:302], mean)
LowMortalityMean <- sapply(LowMortality[,2:302], mean)
```

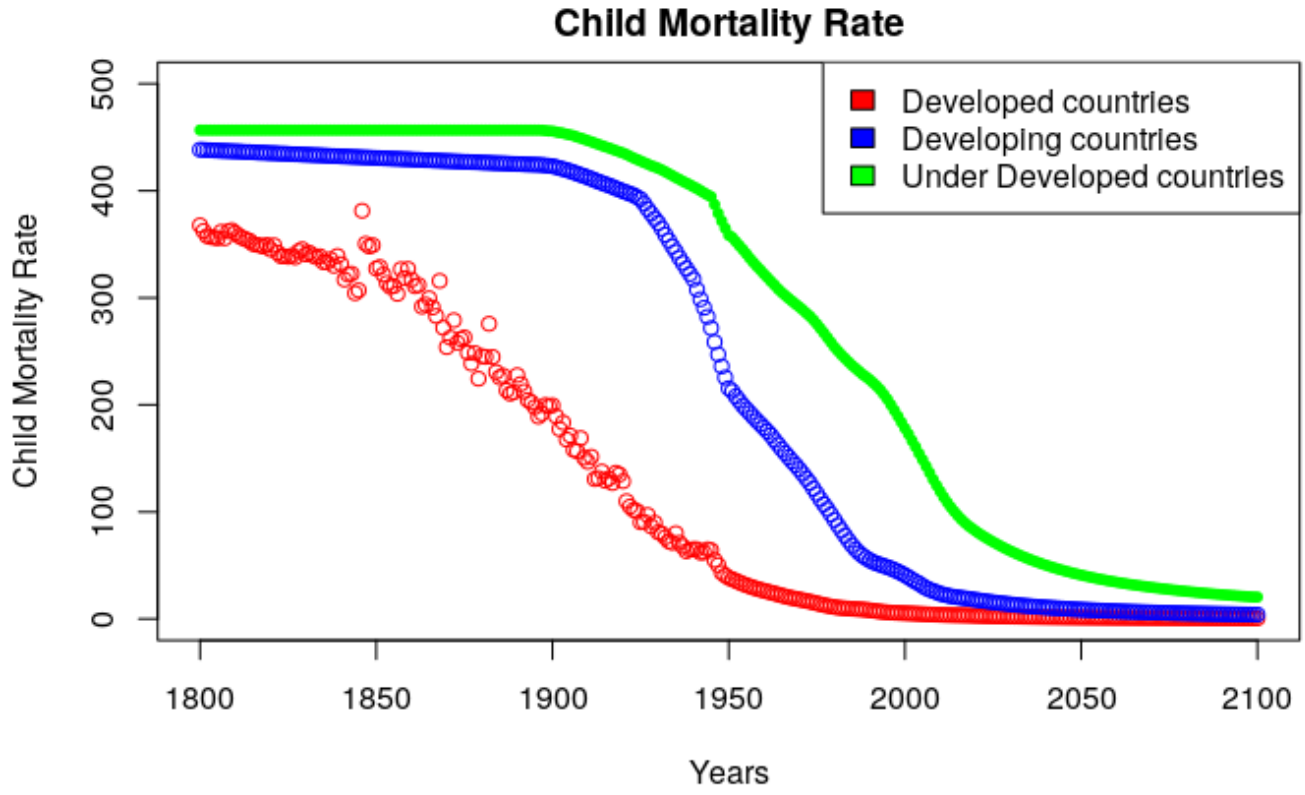
Plotting the means for all the countries for all the 3 groups in the same plot, to compare the the trends

Hide

```
plot(c(1800:2100), TopMortalityMean, col="red", main = "Child Mortality Rate", pch=1,
ylim = c(0,500), xlab = "Years", ylab = "Child Mortality Rate")
points(c(1800:2100), MidMortalityMean, col="blue")
```

Hide

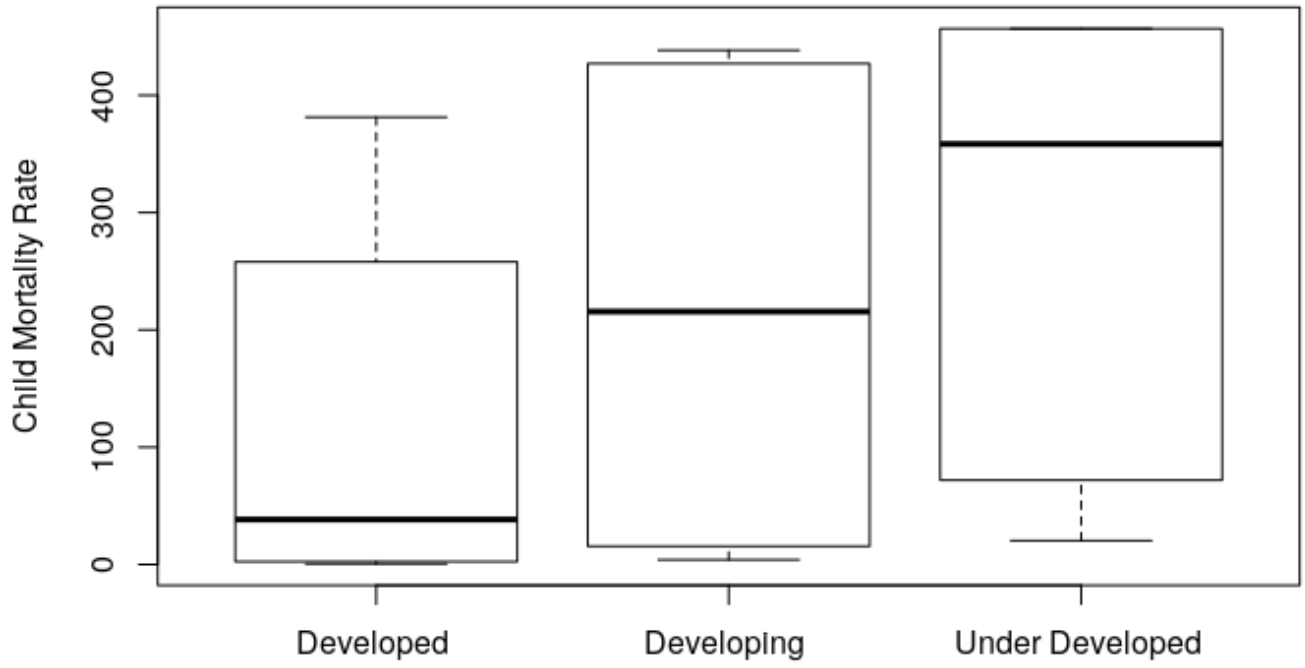
```
points(c(1800:2100), LowMortalityMean, pch=20, col="green")
legend(x = "topright", legend = c("Developed countries ","Developing countries","Unde
r Developed countries"), fill = c("red","blue","green"))
```



Making box plot to compare the means over the course of years

Hide

```
boxplot(TopMortalityMean, MidMortalityMean, LowMortalityMean, ylab = "Child Mortality
Rate", names=c("Developed ","Developing ","Under Developed "))
```



Analysis:

The child mortality rate for all the countries have significantly gone down over the years.

Now we plot all the datasets in a group wise manner and try to see the effect of one parameter in the others

For developed countries

We will normalize all the values using min-max normalization so that we can compare the values.

Hide

```
normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
```

Hide

```
TopLifeMeanNorm <- normalize(TopLifeMean)
TopPopulationMeanNorm <- normalize(TopPopulationMean)
TopIncomeMeanNorm <- normalize(TopIncomeMean)
TopChildrenMeanNorm <- normalize(TopChildrenMean)
TopMortalityMeanNorm <- normalize(TopMortalityMean)
```

Hide

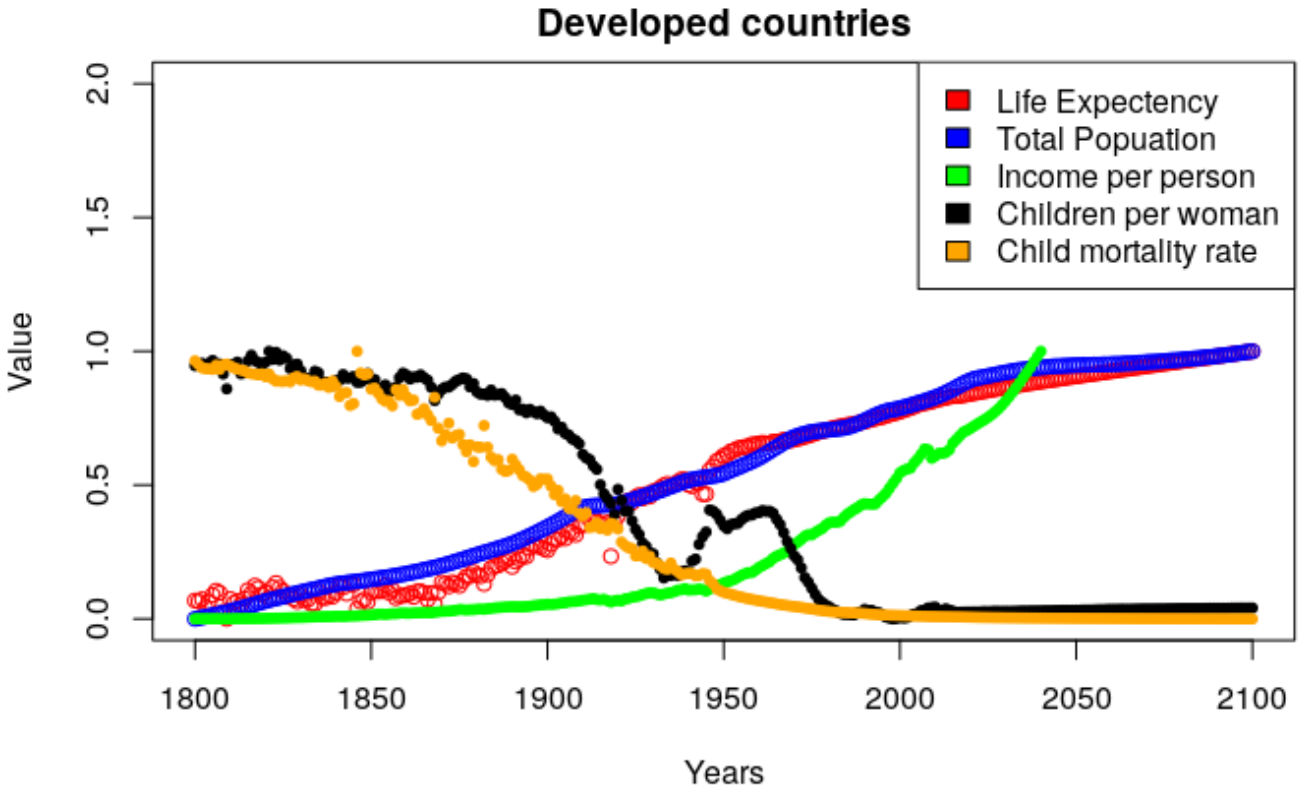
```
plot(c(1800:2100), TopLifeMeanNorm, col="red", main = "Developed countries", pch=1, y
lim = c(0,2), xlab = "Years", ylab = "Value")
points(c(1800:2100), TopPopulationMeanNorm, col="blue")
```

Hide

```
points(c(1800:2040), TopIncomeMeanNorm, pch=20, col="green")
points(c(1800:2100), TopChildrenMeanNorm, pch=20, col="black")
```

Hide

```
points(c(1800:2100), TopMortalityMeanNorm, pch=20, col="orange")
legend(x = "topright", legend = c("Life Expectency ","Total Popuation","Income per pe
rson", "Children per woman", "Child mortality rate"), fill = c("red","blue","green",
"black", "orange"))
```



For developing countries

Hide

```
MidLifeMeanNorm <- normalize(MidLifeMean)
MidPopulationMeanNorm <- normalize(MidPopulationMean)
MidIncomeMeanNorm <- normalize(MidIncomeMean)
MidChildrenMeanNorm <- normalize(MidChildrenMean)
MidMortalityMeanNorm <- normalize(MidMortalityMean)
```

Hide

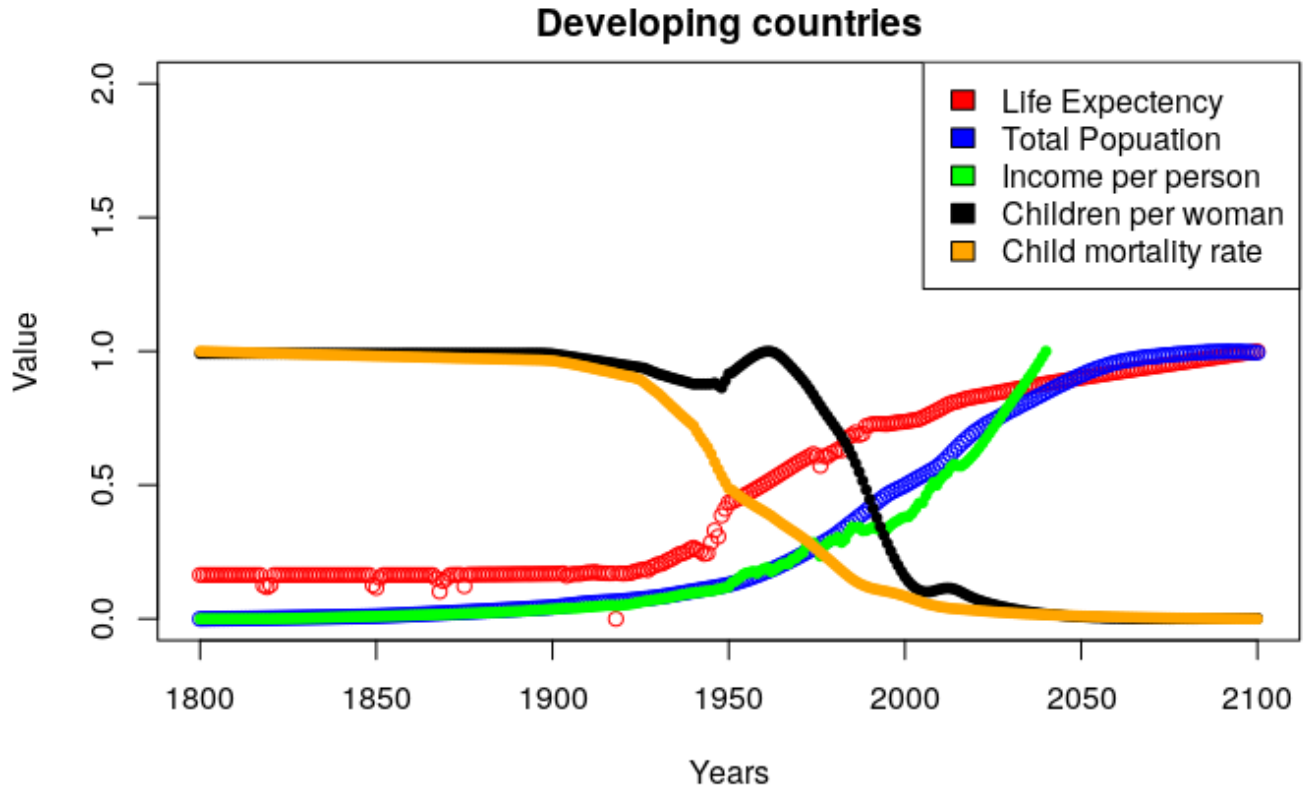
```
plot(c(1800:2100), MidLifeMeanNorm, col="red", main = "Developing countries", pch=1,
 ylim = c(0,2), xlab = "Years", ylab = "Value")
points(c(1800:2100), MidPopulationMeanNorm, col="blue")
```

Hide

```
points(c(1800:2040), MidIncomeMeanNorm, pch=20, col="green")
points(c(1800:2100), MidChildrenMeanNorm, pch=20, col="black")
```

Hide

```
points(c(1800:2100), MidMortalityMeanNorm, pch=20, col="orange")
legend(x = "topright", legend = c("Life Expectency ","Total Popuation","Income per pe
rson", "Children per woman", "Child mortality rate"), fill = c("red","blue","green",
"black", "orange"))
```



For under developed countries

Hide

```
LowLifeMeanNorm <- normalize(LowLifeMean)
LowPopulationMeanNorm <- normalize(LowPopulationMean)
LowIncomeMeanNorm <- normalize(LowIncomeMean)
LowChildrenMeanNorm <- normalize(LowChildrenMean)
LowMortalityMeanNorm <- normalize(LowMortalityMean)
```

Hide

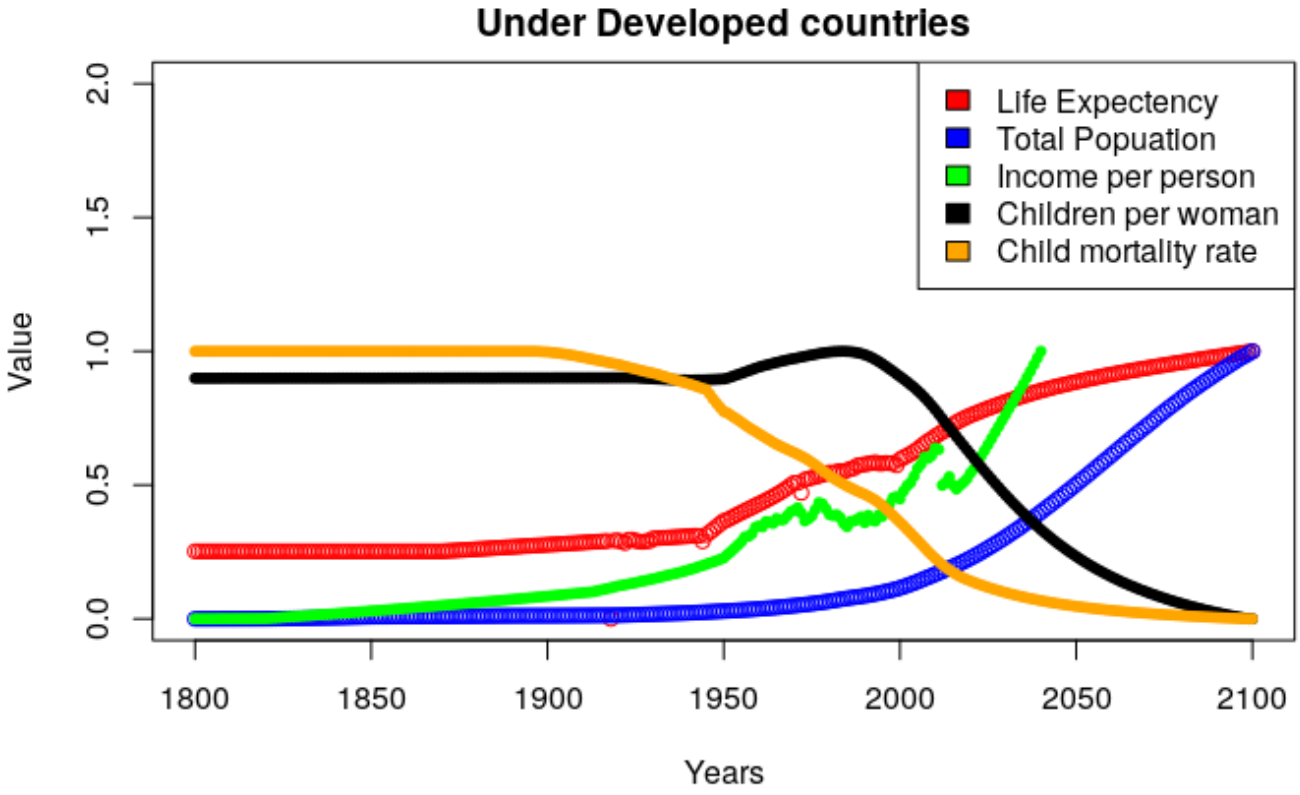
```
plot(c(1800:2100), LowLifeMeanNorm, col="red", main = "Under Developed countries", pc
h=1, ylim = c(0,2), xlab = "Years", ylab = "Value")
points(c(1800:2100), LowPopulationMeanNorm, col="blue")
```

Hide

```
points(c(1800:2040), LowIncomeMeanNorm, pch=20, col="green")
points(c(1800:2100), LowChildrenMeanNorm, pch=20, col="black")
```

Hide

```
points(c(1800:2100), LowMortalityMeanNorm, pch=20, col="orange")
legend(x = "topright", legend = c("Life Expectency ","Total Popuation","Income per pe
rson", "Children per woman", "Child mortality rate"), fill = c("red","blue","green",
"black", "orange"))
```



Analysis:
In all the 3 groups of countries, we see that as the income per person goes up, the child mortality rate and children per woman decrease while life expectancy and total population increase.

Testing the correlation between income per person and the rest of the datasets

Income vs. Life expectancy

For developed countries

Hide

```
cor.test(TopIncomeMean, TopLifeMean[1:241])
```

```
    Pearson's product-moment correlation

data:  TopIncomeMean and TopLifeMean[1:241]
t = 27.529, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8378711 0.8992099
sample estimates:
      cor
0.8719197
```

For developing countries

Hide

```
cor.test(MidIncomeMean, MidLifeMean[1:241])
```

```
    Pearson's product-moment correlation

data:  MidIncomeMean and MidLifeMean[1:241]
t = 38.19, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9067796 0.9428586
sample estimates:
      cor
0.9269321
```

For under developed countries

Hide

```
cor.test(LowIncomeMean, LowLifeMean[1:241])
```

```
    Pearson's product-moment correlation

data:  LowIncomeMean and LowLifeMean[1:241]
t = 59.406, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9586344 0.9749082
sample estimates:
      cor
0.9677664
```

For all the categories, the correlation was close to 1, which means as the per person income increases, life expectancy increases.

Income vs. Children per woman

For developed countries

Hide

```
cor.test(TopIncomeMean, TopChildrenMean[1:241])
```

```
    Pearson's product-moment correlation

data:  TopIncomeMean and TopChildrenMean[1:241]
t = -21.941, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8554544 -0.7707017
sample estimates:
       cor
-0.8174555
```

For developing countries

Hide

```
cor.test(MidIncomeMean, MidChildrenMean[1:241])
```

```
    Pearson's product-moment correlation

data:  MidIncomeMean and MidChildrenMean[1:241]
t = -41.098, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9499785 -0.9182058
sample estimates:
       cor
-0.9359711
```

For under developed countries

Hide

```
cor.test(LowIncomeMean, LowChildrenMean[1:241])
```

```
    Pearson's product-moment correlation

data:  LowIncomeMean and LowChildrenMean[1:241]
t = -14.504, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7460669 -0.6106354
sample estimates:
       cor
-0.6842059
```

For all the categories, the correlation was negative, which means as the per person income increases, children per woman decreases.

Income vs. Child mortality rate

For developed countries

Hide

```
cor.test(TopIncomeMean, TopMortalityMean[1:241])
```

```
    Pearson's product-moment correlation

data:  TopIncomeMean and TopMortalityMean[1:241]
t = -17.296, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7968641 -0.6836156
sample estimates:
       cor
-0.7455747
```

For developing countries

Hide

```
cor.test(MidIncomeMean, MidMortalityMean[1:241])
```

```
    Pearson's product-moment correlation

data:  MidIncomeMean and MidMortalityMean[1:241]
t = -30.502, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9151850 -0.8628663
sample estimates:
       cor
-0.8919742
```

For under developed countries

Hide

```
cor.test(LowIncomeMean, LowMortalityMean[1:241])
```

```
    Pearson's product-moment correlation

data:  LowIncomeMean and LowMortalityMean[1:241]
t = -58.256, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9739537 -0.9570744
sample estimates:
       cor
-0.9665449
```

For all the categories, the correlation was negative, which means as the per person income increases, child mortality rate decreases.

Income vs. Popuation

For developed countries

Hide

```
cor.test(TopIncomeMean, TopPopulationMean[1:241])
```

```
    Pearson's product-moment correlation

data:  TopIncomeMean and TopPopulationMean[1:241]
t = 30.744, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8646570 0.9163233
sample estimates:
      cor
0.8934066
```

For developing countries

Hide

```
cor.test(MidIncomeMean, MidPopulationMean[1:241])
```

```
    Pearson's product-moment correlation

data:  MidIncomeMean and MidPopulationMean[1:241]
t = 96.699, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9838616 0.9902600
sample estimates:
    cor
0.98746
```

For under developed countries

Hide

```
cor.test(LowIncomeMean, LowPopulationMean[1:241])
```

```
    Pearson's product-moment correlation

data:  LowIncomeMean and LowPopulationMean[1:241]
t = 39.463, df = 239, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9120458 0.9461442
sample estimates:
     cor
0.931101
```

For all the categories, the correlation was close to 1, which means as the per person income increases, population increases linearly.

Testing a few hypothesis on the datasets

Child Mortality Rate

Shapiro-Wilk normality test
We first check if the data is normally distributed.
-> Null hypothesis H_0 : The data is normally distributed
-> Alternative hypothesis H_d : The data is not normally distributed

For developed countries

Hide

```
shapiro.test(TopMortalityMean)
```

```
    Shapiro-Wilk normality test

data:  TopMortalityMean
W = 0.78427, p-value < 2.2e-16
```

For developing countries

Hide

```
shapiro.test(MidMortalityMean)
```

```
    Shapiro-Wilk normality test

data:  MidMortalityMean
W = 0.7671, p-value < 2.2e-16
```

For under developed countries

Hide

```
shapiro.test(LowMortalityMean)
```

```
    Shapiro-Wilk normality test

data:  LowMortalityMean
W = 0.7866, p-value < 2.2e-16
```

For all the groups, the P-values are too low, suggesting that our alternate hypothesis is true, meaning that the data is not normally distrbuted.

Wilcoxon test

Now we will do a non parametric test to compare the means of our datasets pairwise. For that, we use the Wilcoxon test to compare the means.

Case 1

-> Null hypothesis H_0 : The difference between the mean child mortality rate of developing and developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean child mortality rate of developed and developing countries is less than zero i.e. developed countries are lower in child mortality rate.

Hide

```
wilcox.test(TopMortalityMean, MidMortalityMean, paired = TRUE, alternative = "less")
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopMortalityMean and MidMortalityMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

Case 2

-> Null hypothesis H_0 : The difference between the mean child mortality rate of developed and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean child mortality rate of developed and under developed countries is less than zero i.e. developed countries are lower in child mortality rate.

Hide

```
wilcox.test(TopMortalityMean, LowMortalityMean, paired = TRUE, alternative = "less")
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopMortalityMean and LowMortalityMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

Case 3

-> Null hypothesis H_0 : The difference between the mean child mortality rate of developing and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean child mortality rate of developing and under developed countries is less than zero i.e. developing countries are lower in child mortality rate.

Hide

```
wilcox.test(MidMortalityMean, LowMortalityMean, paired = TRUE, alternative = "less")
```

```
    Wilcoxon signed rank test with continuity correction

data:  MidMortalityMean and LowMortalityMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is less than 0
```

Life Expectancy

Shapiro-Wilk normality test

We first check if the data is normally distributed.

-> Null hypothesis H_0 : The data is normally distributed

-> Alternative hypothesis H_d : The data is not normally distributed

For developed countries

Hide

```
shapiro.test(TopLifeMean)
```

```
    Shapiro-Wilk normality test

data:  TopLifeMean
W = 0.88127, p-value = 1.582e-14
```

For developing countries

Hide

```
shapiro.test(MidLifeMean)
```

```
    Shapiro-Wilk normality test

data:  MidLifeMean
W = 0.80681, p-value < 2.2e-16
```

For under developed countries

Hide

```
shapiro.test(LowLifeMean)
```

```
    Shapiro-Wilk normality test

data:  LowLifeMean
W = 0.81815, p-value < 2.2e-16
```

For all the groups, the P-values are too low, suggesting that our alternate hypothesis is true, meaning that the data is not normally distrbuted.

Wilcoxon test

Now we will do a non parametric test to compare the means of our datasets pairwise. For that, we use the Wilcoxon test to compare the means.

Case 1

-> Null hypothesis H_0 : The difference between the mean life expectancy of developing and developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean life expectancy of developed and developing countries is more than zero i.e. developed countries are higher in life expectancy.

Hide

```
wilcox.test(TopLifeMean, MidLifeMean, paired = TRUE, alternative = "greater")
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopLifeMean and MidLifeMean
V = 45451, p-value < 2.2e-16
alternative hypothesis: true location shift is greater than 0
```

Case 2

-> Null hypothesis H_0 : The difference between the mean mean life expectancy of developed and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean mean life expectancy of developed and under developed countries is more than zero i.e. developed countries are greater in mean life expectancy.

Hide

```
wilcox.test(TopMortalityMean, LowMortalityMean, paired = TRUE, alternative = "greate
r")
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopMortalityMean and LowMortalityMean
V = 0, p-value = 1
alternative hypothesis: true location shift is greater than 0
```

Case 3

-> Null hypothesis H_0 : The difference between the mean life expectancy of developing and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean life expectancy of developing and under developed countries is more than zero i.e. developing countries are greater in child mortality rate.

Hide

```
wilcox.test(MidMortalityMean, LowMortalityMean, paired = TRUE, alternative = "greate
r")
```

```
    Wilcoxon signed rank test with continuity correction

data:  MidMortalityMean and LowMortalityMean
V = 0, p-value = 1
alternative hypothesis: true location shift is greater than 0
```

Population

Shapiro-Wilk normality test

We first check if the data is normally distributed.

-> Null hypothesis H_0 : The data is normally distributed

-> Alternative hypothesis H_d : The data is not normally distributed

For developed countries

Hide

```
shapiro.test(TopPopulationMean)
```

```
    Shapiro-Wilk normality test

data:  TopPopulationMean
W = 0.90351, p-value = 6.032e-13
```

For developing countries

Hide

```
shapiro.test(MidPopulationMean)
```

```
    Shapiro-Wilk normality test

data:  MidPopulationMean
W = 0.79085, p-value < 2.2e-16
```

For under developed countries

Hide

```
shapiro.test(LowPopulationMean)
```

```
    Shapiro-Wilk normality test

data:  LowPopulationMean
W = 0.68812, p-value < 2.2e-16
```

For all the groups, the P-values are too low, suggesting that our alternate hypothesis is true, meaning that the data is not normally distrbuted.

Wilcoxon test

Now we will do a non parametric test to compare the means of our datasets pairwise. For that, we use the Wilcoxon test to compare the means.

Case 1
-> Null hypothesis H_0 : The difference between the mean population of developing and developed countries is zero
-> Alternative hypothesis H_d : The difference between the mean population of developed and developing countries is not zero.

Hide

```
wilcox.test(TopPopulationMean, MidPopulationMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopPopulationMean and MidPopulationMean
V = 45451, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Case 2

-> Null hypothesis H_0 : The difference between the mean mean population of developed and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean mean population of developed and under developed countries is not zero.

Hide

```
wilcox.test(TopPopulationMean, LowPopulationMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopPopulationMean and LowPopulationMean
V = 27855, p-value = 0.0006892
alternative hypothesis: true location shift is not equal to 0
```

Case 3

-> Null hypothesis H_0 : The difference between the mean population of developing and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean population of developing and under developed countries is not zero.

Hide

```
wilcox.test(MidPopulationMean, LowPopulationMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  MidPopulationMean and LowPopulationMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Children per woman

Shapiro-Wilk normality test

We first check if the data is normally distributed.

-> Null hypothesis H_0 : The data is normally distributed

-> Alternative hypothesis H_d : The data is not normally distributed

For developed countries

Hide

```
shapiro.test(TopChildrenMean)
```

```
    Shapiro-Wilk normality test

data:  TopChildrenMean
W = 0.80031, p-value < 2.2e-16
```

For developing countries

Hide

```
shapiro.test(MidChildrenMean)
```

```
    Shapiro-Wilk normality test

data:  MidChildrenMean
W = 0.70919, p-value < 2.2e-16
```

For under developed countries

Hide

```
shapiro.test(LowChildrenMean)
```

```
    Shapiro-Wilk normality test

data:  LowChildrenMean
W = 0.69168, p-value < 2.2e-16
```

For all the groups, the P-values are too low, suggesting that our alternate hypothesis is true, meaning that the data is not normally distrbuted.

Wilcoxon test

Now we will do a non parametric test to compare the means of our datasets pairwise. For that, we use the Wilcoxon test to compare the means.

Case 1
-> Null hypothesis H_0 : The difference between the mean children per woman of developing and developed countries is zero
-> Alternative hypothesis H_d : The difference between the mean children per woman of developed and developing countries is not zero.

Hide

```
wilcox.test(TopChildrenMean, MidChildrenMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopChildrenMean and MidChildrenMean
V = 677.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Case 2

-> Null hypothesis H_0 : The difference between the mean mean children per woman of developed and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean mean children per woman of developed and under developed countries is not zero.

Hide

```
wilcox.test(TopChildrenMean, LowChildrenMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  TopChildrenMean and LowChildrenMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Case 3

-> Null hypothesis H_0 : The difference between the mean children per woman of developing and under developed countries is zero

-> Alternative hypothesis H_d : The difference between the mean children per woman of developing and under developed countries is not zero.

Hide

```
wilcox.test(MidChildrenMean, LowChildrenMean, paired = TRUE)
```

```
    Wilcoxon signed rank test with continuity correction

data:  MidChildrenMean and LowChildrenMean
V = 0, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Note:

For each category of countries, the shapiro test told us that the datas aren't normally distributed. Hence we cannot use **t test or ANOVA** to compare the means of our datasets.