Assessment 1 - Hricha Acharya (17107)

Group Name : Shanon

The goal of this assessment is to explore R programming using Gapminder dataset.Following are the data sets and the information they include:

children_per_woman_total_fertility : It includes the data of average children per woman for 194 different countries per year from the year 1800 to 2100.

child_mortality_0_5_year_olds_dying_per_1000_born : It includes the data of child mortality(children under 5 year's age dying per 1000) for 183 different countries per year from the year 1800 to 2100.

income_per_person_gdppercapita_ppp_inflation_adjusted : It includes the data of GDP per capita for 192 different countries per year from the year 1800 to 2040.

life_expectancy_years : It includes the data of Life expectancy for 186 different countries per year from the year 1800 to 2100.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

Loading data :

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
setwd("/Users/hrichaacharya/Desktop/R /dataset")
getwd()
```

```
## [1] "/Users/hrichaacharya/Desktop/R /dataset"
```

```
Children_per_woman <- read.csv("children_per_woman_total_fertility.csv", header =
T, check.names = F)
Life_expectancy <- read.csv("life_expectancy_years.csv", header = T, check.names =
F)
Child_mortality <- read.csv("child_mortality_0_5_year_olds_dying_per_1000_born.csv
", header = T, check.names = F)
Population_total <- read.csv("population_total.csv", header = T, check.names = F)
Income_per_person <- read.csv("income_per_person_gdppercapita_ppp_inflation_adjust
ed.csv", header = T, check.names = F)
```

Part 1

In Part 1 we consider ten countries each for three different sets of Countries (Developed, Developing and Under Developed).For each set of countries we take the mean of sample values for each year.We have considered mean of 10 countries because it will help us make a more accurate estimate of what trend a country from a set is likely to follow. Later we plot these means for Developed, Developing and Under Developed countries and compare their trends. We also make certain Hypothesis regarding the data we observe and try to prove or reject those Hypothesis with help of various Hypothesis testing methods we covered in this course so far. Countries that are considered 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

Underdeveloped 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/)).

Dataset : “Children_per_woman”

Subsetting dataframe into different sets of countries as described above

```
head(Children_per_woman)
```

	country <chr>	1800 <dbl>	1801 <dbl>	1802 <dbl>	1803 <dbl>	1804 <dbl>	1805 <dbl>	1806 <dbl>	1807 <dbl>
1	Afghanistan	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
2	Albania	4.60	4.60	4.60	4.60	4.60	4.60	4.60	4.60
3	Algeria	6.99	6.99	6.99	6.99	6.99	6.99	6.99	6.99
4	Angola	6.93	6.93	6.93	6.93	6.93	6.93	6.93	6.94
5	Antigua and Barbuda	5.00	5.00	4.99	4.99	4.99	4.98	4.98	4.97
6	Argentina	6.80	6.80	6.80	6.80	6.80	6.80	6.80	6.80

6 rows | 1-10 of 303 columns

Here when we look at our data, we can clearly see that we have one row for each country which represents sample data values for years 1800 to 2100 under the columns named by the respective year. We will now subset this dataset in the way we require i.e. we subset it into three smaller dataframes each of which will include the set of ten countries as selected above.

```
# Dataframe for Developed Countries:

Developed_Children_per_woman <- Children_per_woman[Children_per_woman$country %in%
c("Norway" , "Ireland" , "Switzerland" , "Finland" , "Iceland" , "Germany" , "Swed
en" , "Australia", "Netherlands" , "Denmark"),]
Developed_Children_per_woman
```

	country <chr>	1800 <dbl>	1801 <dbl>	1802 <dbl>	1803 <dbl>	1804 <dbl>	1805 <dbl>	1806 <dbl>	1807 <dbl>
8	Australia	6.50	6.48	6.46	6.44	6.42	6.40	6.38	6.36
46	Denmark	4.04	4.04	4.05	4.05	4.06	4.06	4.07	4.07
58	Finland	4.92	5.07	5.23	4.78	5.24	5.21	4.84	4.97
63	Germany	5.40	5.40	5.39	5.39	5.38	5.38	5.37	5.37
74	Iceland	4.88	4.88	4.88	4.88	4.88	4.88	4.88	4.88
79	Ireland	4.20	4.20	4.20	4.20	4.20	4.20	4.20	4.20
116	Netherlands	5.11	5.11	5.11	5.11	5.11	5.11	5.11	5.11
123	Norway	4.32	4.07	3.91	4.20	3.94	4.33	4.39	4.27
159	Sweden	4.07	4.26	4.50	4.45	4.52	4.50	4.36	4.42
160	Switzerland	4.14	4.14	4.14	4.14	4.14	4.14	4.14	4.14

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Developing Countries:

Developing_Children_per_woman <- Children_per_woman[Children_per_woman$country %in
% c("Algeria" , "Lebanon" , "Fiji" , "Moldova" , "Maldives" , "Tunisia" , "St. Vin
cent and the Grenadines" , "Suriname" , "Mongolia" , "Botswana"),]
Developing_Children_per_woman
```

	country <chr>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>
3	Algeria	6.99	6.99	6.99	6.99	6.99	6.99	6.99	6.99
22	Botswana	6.47	6.47	6.47	6.47	6.47	6.47	6.47	6.47
57	Fiji	6.45	6.45	6.45	6.45	6.45	6.45	6.45	6.45
92	Lebanon	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74
101	Maldives	5.98	5.98	5.98	5.98	5.98	5.98	5.98	5.98
108	Moldova	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39
109	Mongolia	5.94	5.94	5.94	5.94	5.94	5.94	5.94	5.94
156	St. Vincent and the Grenadines	6.54	6.54	6.54	6.54	6.54	6.54	6.54	6.54
158	Suriname	6.58	6.58	6.58	6.58	6.58	6.58	6.58	6.58
169	Tunisia	6.40	6.40	6.40	6.40	6.40	6.40	6.40	6.40

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Underdeveloped Countries:

Underdeveloped_Children_per_woman <- Children_per_woman[Children_per_woman$country
%in% c("Eritrea" , "Mozambique" , "Burkina Faso" , "Sierra Leone" , "Mali" , "Buru
ndi" , "South Sudan" , "Chad" , "Central African Republic" , "Niger"),]
Underdeveloped_Children_per_woman
```

	country <chr>	1800 <dbl>	1801 <dbl>	1802 <dbl>	1803 <dbl>	1804 <dbl>	1805 <dbl>	1806 <dbl>	1807 <dbl>
26	Burkina Faso	6.03	6.03	6.03	6.03	6.03	6.03	6.03	6.03
27	Burundi	6.80	6.80	6.80	6.80	6.80	6.80	6.80	6.80
32	Central African Republic	6.51	6.51	6.51	6.51	6.51	6.51	6.51	6.51
33	Chad	6.06	6.06	6.06	6.06	6.06	6.06	6.06	6.06
53	Eritrea	6.96	6.96	6.96	6.96	6.96	6.96	6.96	6.96
102	Mali	6.23	6.23	6.23	6.23	6.23	6.23	6.23	6.23

112	Mozambique	6.63	6.63	6.63	6.63	6.63	6.63	6.63	6.63
119	Niger	6.83	6.83	6.83	6.83	6.83	6.83	6.83	6.83
144	Sierra Leone	5.72	5.72	5.72	5.72	5.72	5.72	5.72	5.72
152	South Sudan	6.64	6.64	6.64	6.64	6.64	6.64	6.64	6.64

1-10 of 10 rows | 1-10 of 303 columns

Taking mean per column of subsets

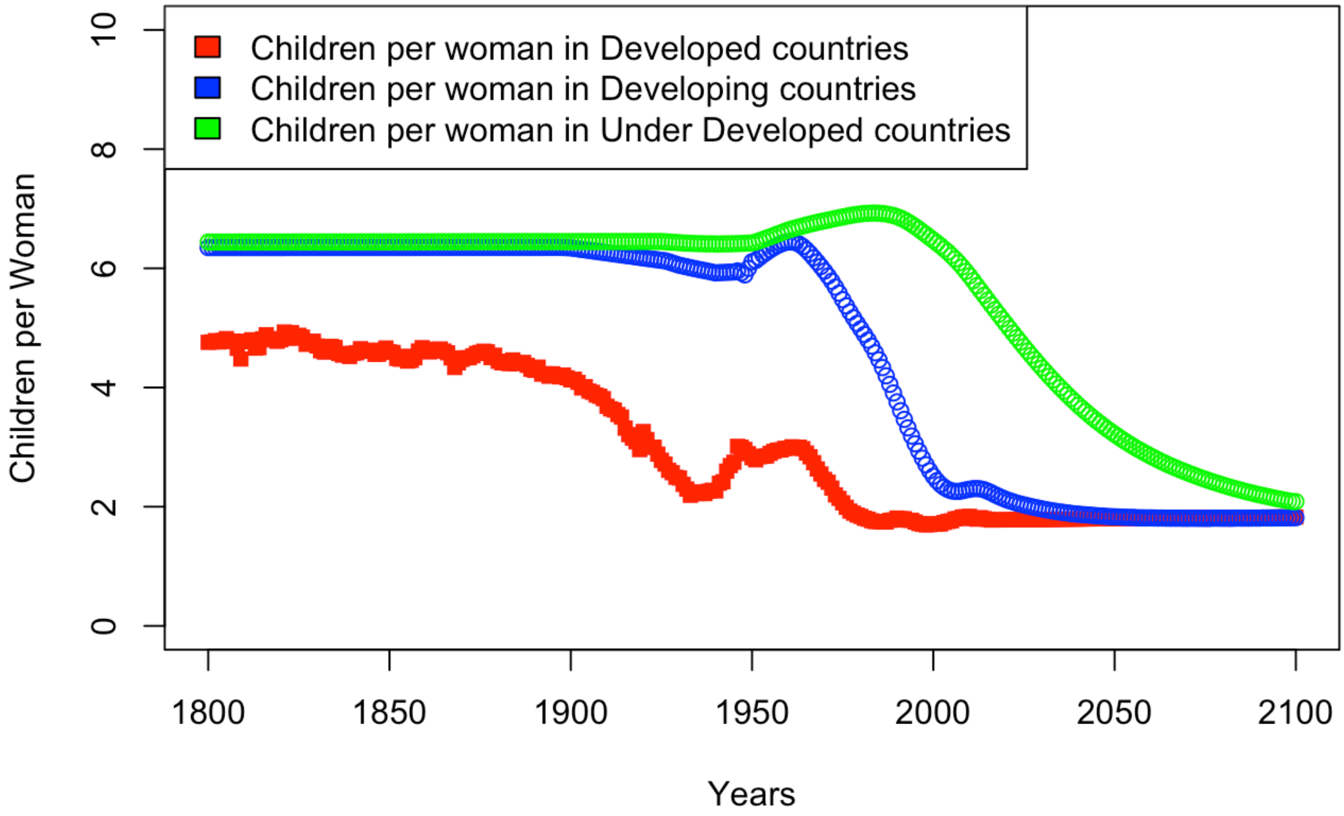
Now that we have created three subsets for sets of countries (Developed, Developing and Under Developed) we require for our analysis, we will take mean for values of each year for all the ten countries in the dataset.

```
# Taking mean of values for each set of countries and save it as a list:

Developed_Children_per_woman_Mean <-  apply(Developed_Children_per_woman[,2:302],2
, mean)
Developing_Children_per_woman_Mean <-  apply(Developing_Children_per_woman[,2:302]
,2, mean)
Underdeveloped_Children_per_woman_Mean <-  apply(Underdeveloped_Children_per_woman
[,2:302],2, mean)
```

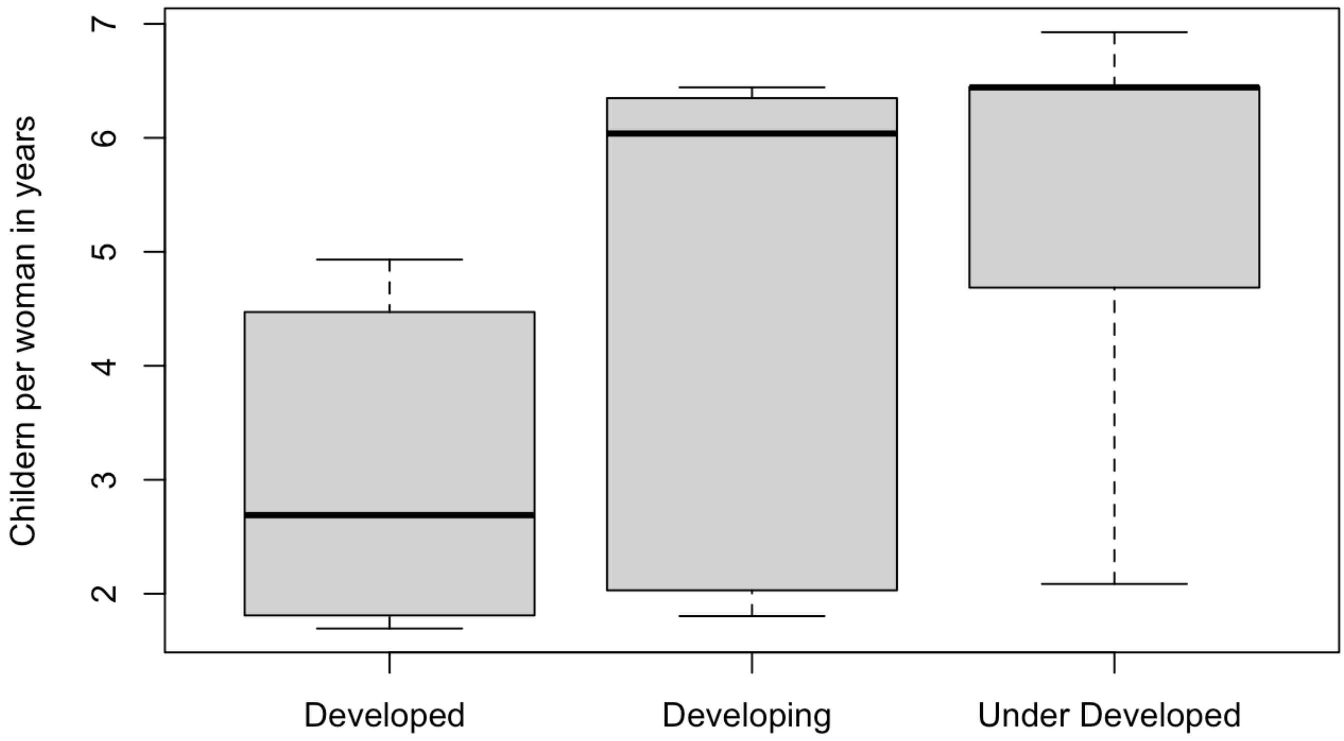
Plotting Graphs and making observations

To compare the trend of data values for Children per woman in developed, developing and under developed countries from the year 1800 to 2100 we will plot them on one graph and make certain hypothesis based on the observations we derive from that graph.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Children_per_woman_Mean,col="red",pch=15,ylim = c(0,10
),
     xlab="Years",ylab="Children per Woman")
points(c(1800:2100),Developing_Children_per_woman_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Children_per_woman_Mean,col="green")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Children per woman in Developed countries ","Children per woman
in Developing countries","Children per woman in Under Developed countries"),
       fill = c("red","blue","green"))
```

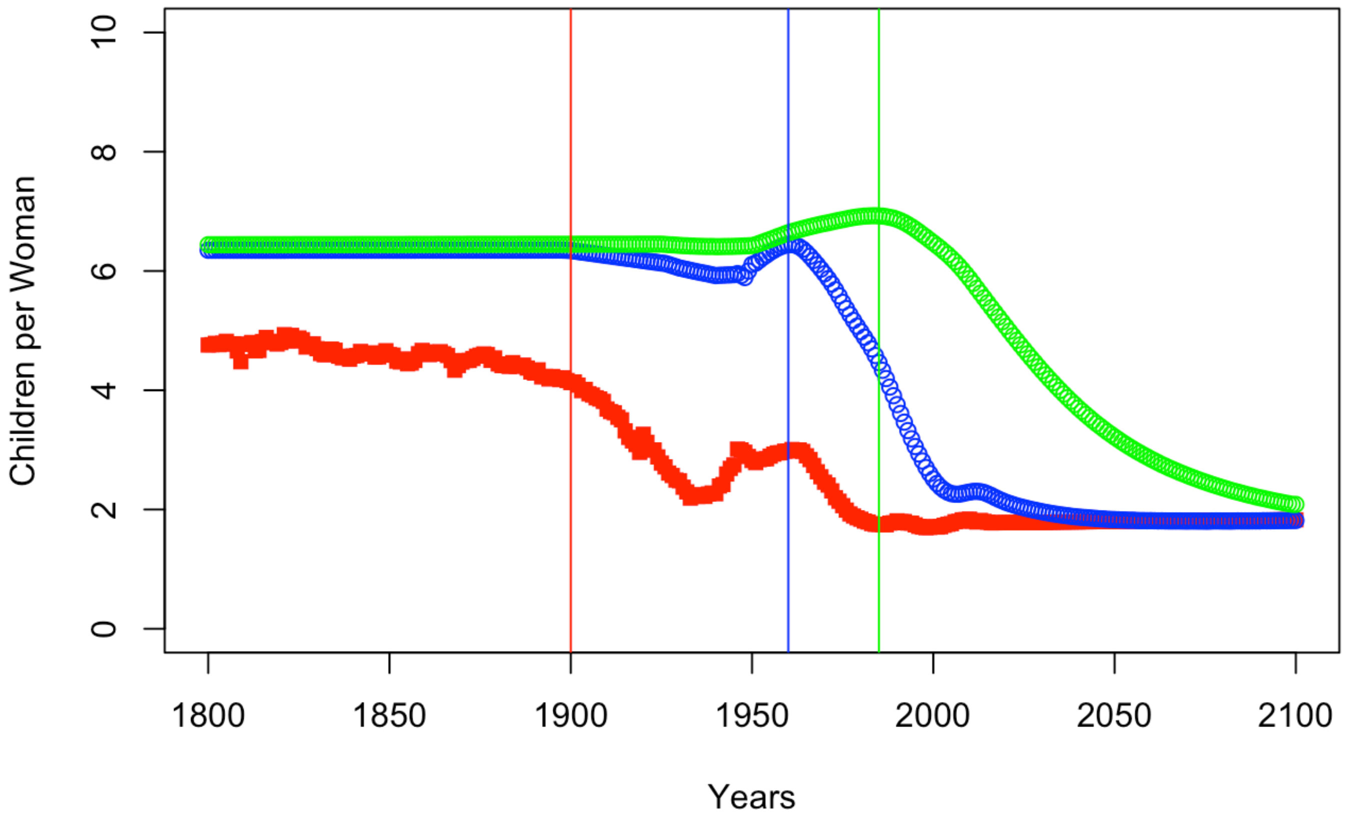



```
# Boxplots:
boxplot(Developed_Children_per_woman_Mean,Developing_Children_per_woman_Mean,Under
developed_Children_per_woman_Mean,
        ylab = "Childern per woman in years",
        names=c("Developed ","Developing ","Under Developed "))
```



In the first graph we observe that there is certain deviation in the trend that each of the set of countries follow. In the years 1800 to 1900 the trend for Children per women stay nearly the same but then it starts to drop. We can add vertical lines to our graph to check roughly where this drop occurs in different sets of countries.

```
# plotting data lits:
plot(c(1800:2100),Developed_Children_per_woman_Mean,col="red",pch=15,ylim = c(0,10
),xlab="Years",ylab="Children per Woman")
points(c(1800:2100),Developing_Children_per_woman_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Children_per_woman_Mean,col="green")

# Adding vertical lines which mark the beginning of deviation from general trend i
n years 1800s :

abline(v=1985,col="green")
abline(v=1960,col="blue")
abline(v=1900,col="red")
```



These are the following observations:

- Values for Under Developed countries remain roughly same from 1800 to 1980 but we see that it starts to drop in 1980s.
- Values for Developing countries remain roughly same from 1800 to 1960 but we see that it starts to drop in 1960s.
- Values for Developed countries remain roughly same from 1800 to 1900 but we see that it starts to drop in 1900s.

Another interesting observation that can be drawn from the **boxplot** is that the median of values Children per woman for Developed countries is lower than that of Developing countries and mean of values Children per woman for Developing countries is lower than that of Under Developed countries. To check this we can perform various Hypothesis tests:

1. We check whether the mean of each type of countries have a normal distribution of not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used to compare means.
2. Since we want to compare the mean of three populations, we can use **ANOVA** only if in earlier part we get that our data is normally distributed. If not then we will use pairwise **Wilcoxon test** to compare the mean.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether the mean values for each set of countries has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(Developed_Children_per_woman_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developed_Children_per_woman_Mean
## W = 0.80031, p-value < 2.2e-16
```

```
shapiro.test(Developing_Children_per_woman_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developing_Children_per_woman_Mean
## W = 0.70657, p-value < 2.2e-16
```

```
shapiro.test(Underdeveloped_Children_per_woman_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Underdeveloped_Children_per_woman_Mean
## W = 0.69168, p-value < 2.2e-16
```

For each type of countries we have the p-value less than 0.05. Hence, we reject our Null Hypothesis (that the data vectors are normally distributed) and conclude that the set of mean values for each type of country is not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare means of our datasets.

Now we will need a non parametric test to compare the means of our datasets pairwise. We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2100.

1. Wilcoxon test between Developed and Developing Countries:

- Null Hypothesis, H_0 := The difference between mean value of Children per women in Developed countries and of Developing countries is zero.
- Alternate Hypothesis, H_a := The difference between mean value of Children per women in Developed countries and of Developing countries is less than zero (i.e.mean value of Children per women in Developed countries is less than that of Developing countries.)

```
wilcox.test(Developed_Children_per_woman_Mean,Developing_Children_per_woman_Mean,
paired=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developed_Children_per_woman_Mean and Developing_Children_per_woman_Mean
## V = 987, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is less than 0.05 hence we reject our Null hypothesis and get that **mean value of Children per women in Developed countries is less than that of Developing countries.**

2. Wilcoxon test between Developing and Under Developed Countries:

- Null Hypothesis, H_0 := The difference between mean value of Children per women in Developing countries and of Under Developed countries is zero.
- Alternate Hypothesis, H_a := The difference between mean value of Children per women in Developing countries and of Under Developed countries is less than zero (i.e.mean value of Children per women in Developing countries is less than that of Under Developed countries.)

```
wilcox.test(Developing_Children_per_woman_Mean,Underdeveloped_Children_per_woman_M
ean, paired=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developing_Children_per_woman_Mean and Underdeveloped_Children_per_woman
_Mean
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is less than 0.05 hence we reject our Null hypothesis and get that **mean value of Children per women in Developing countries is less than that of Under Developed countries.**

Hence we conclude that the mean of values for Children per woman in Developed countries is lower than that in Developing countries and mean of values for Children per woman in Developing countries is lower than that in Under Developed countries.

Dataset : “Child_mortality”

Subsetting dataframe into different sets of countries as described above

```
head(Child_mortality)
```

	country	1800	1801	1802	1803	1804	1805	1806	1807
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	Afghanistan	469	469	469	469	469	469	470	470
2	Albania	375	375	375	375	375	375	375	375
3	Algeria	460	460	460	460	460	460	460	460
4	Andorra	NA	NA	NA	NA	NA	NA	NA	NA
5	Angola	486	486	486	486	486	486	486	486
6	Antigua and Barbuda	474	470	466	462	458	455	451	447

6 rows | 1-10 of 303 columns

Here when we look at our data, we can clearly see that we have one row for each country which represents sample data values for years 1800 to 2100 under the columns named by the respective year. We will now subset this dataset in the way we require i.e. we subset it into three smaller dataframes each of which will include the set of ten countries as selected intially.

```
# Dataframe for Developed Countries:

Developed_Child_mortality <- Child_mortality[Child_mortality$country %in% c("Norwa
y" , "Ireland" , "Switzerland" , "Finland" , "Iceland" , "Germany" , "Sweden" , "A
ustralia", "Netherlands" , "Denmark"),]
Developed_Child_mortality
```

	country	1800	1801	1802	1803	1804	1805	1806	1807
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
9	Australia	391	391	391	391	391	391	391	391
47	Denmark	380	379	378	376	375	374	373	372
60	Finland	420	420	420	420	420	420	415	410
65	Germany	340	340	340	340	340	340	340	340
77	Iceland	412	412	412	412	412	412	412	412
82	Ireland	348	348	348	348	348	348	348	348
123	Netherlands	324	324	324	324	324	324	324	324
130	Norway	336	336	336	336	336	336	336	336
169	Sweden	381	329	283	275	272	266	333	278
170	Switzerland	345	345	345	345	345	345	345	345

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Developing Countries:

Developing_Child_mortality <- Child_mortality[Child_mortality$country %in% c("Alge
ria" , "Lebanon" , "Fiji" , "Moldova" , "Maldives" , "Tunisia" , "St. Vincent and
the Grenadines" , "Suriname" , "Mongolia" , "Botswana"),]
Developing_Child_mortality
```

	country <chr>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>
3	Algeria	460	460	460	460	460	460	460	460
23	Botswana	397	397	397	397	397	397	397	397
59	Fiji	499	499	499	499	499	499	499	499
95	Lebanon	448	448	448	448	448	448	448	448
105	Maldives	458	458	457	456	455	455	454	453
113	Moldova	397	396	395	395	394	393	392	392
115	Mongolia	420	420	420	420	420	420	420	420
166	St. Vincent and the Grenadines	463	461	458	456	453	450	448	445
168	Suriname	406	406	406	406	406	406	406	406
179	Tunisia	460	460	460	460	460	460	460	460

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Underdeveloped Countries:

Underdeveloped_Child_mortality <- Child_mortality[Child_mortality$country %in% c("
Eritrea" , "Mozambique" , "Burkina Faso" , "Sierra Leone" , "Mali" , "Burundi" , "
South Sudan" , "Chad" , "Central African Republic" , "Niger"),]
Underdeveloped_Child_mortality
```

	country <chr>	1800 <int>	1801 <int>	1802 <int>	1803 <int>	1804 <int>	1805 <int>	1806 <int>	1807 <int>
27	Burkina Faso	455	455	455	455	455	455	455	455
28	Burundi	424	424	424	424	424	424	424	424
33	Central African Republic	444	444	444	444	444	444	444	444
34	Chad	432	432	432	432	432	432	432	432
55	Eritrea	441	441	441	441	441	441	441	441
106	Mali	494	494	494	494	494	494	494	494

118	Mozambique	440	440	440	440	440	440	440	440
126	Niger	433	433	433	433	433	433	433	433
153	Sierra Leone	514	514	514	514	514	514	514	514
161	South Sudan	490	490	490	490	490	490	490	490

1-10 of 10 rows | 1-10 of 303 columns

Taking mean per column of subsets

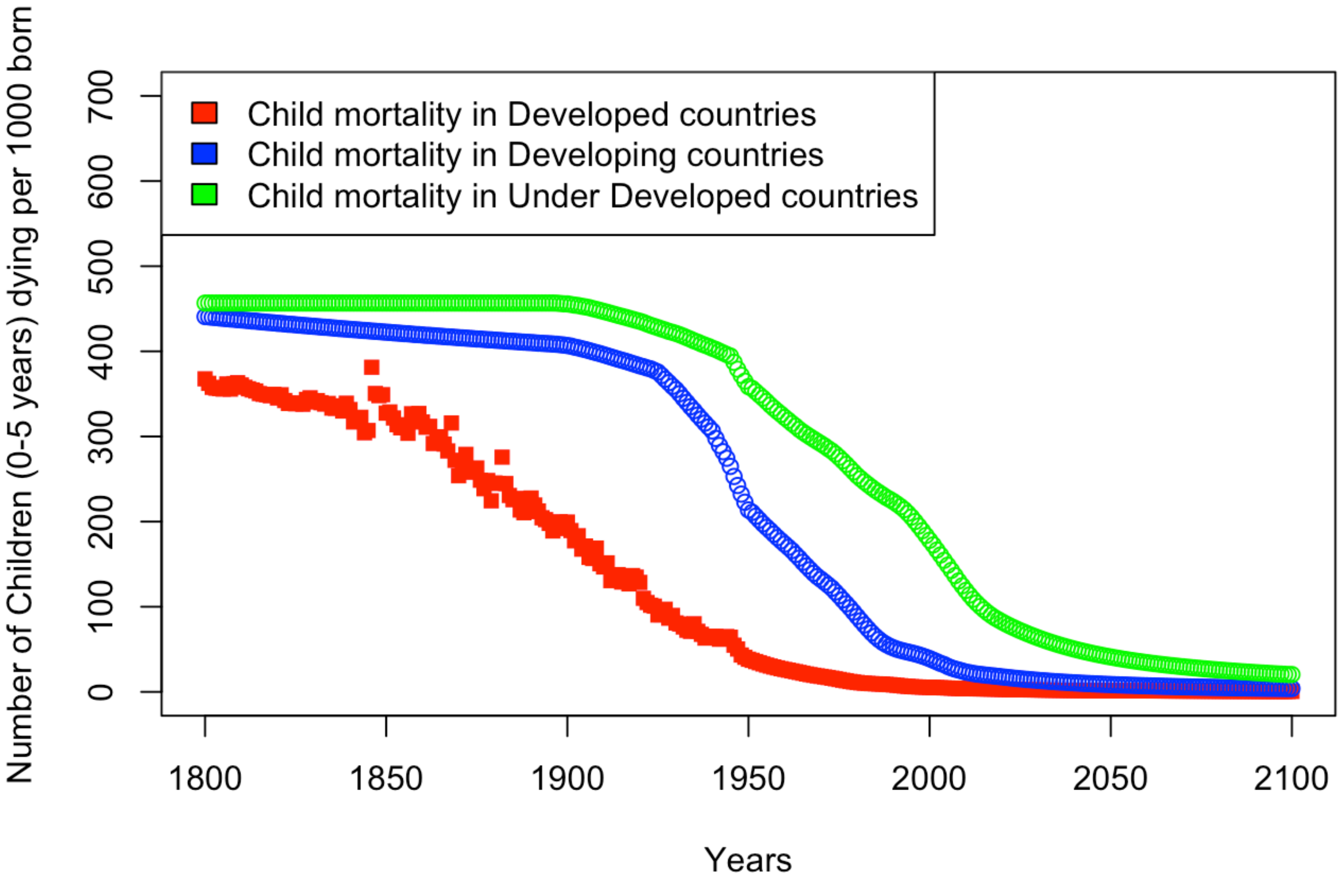
Now that we have created three subsets for sets of countries (Developed, Developing and Under Developed) we require for our analysis, we will take mean for values of each year for all the ten countries in the dataset.

```
# Taking mean of values for each set of countries and save it as a list:

Developed_Child_mortality_Mean <-  apply(Developed_Child_mortality[,2:302],2, mean
)
Developing_Child_mortality_Mean <-  apply(Developing_Child_mortality[,2:302],2, me
an)
Underdeveloped_Child_mortality_Mean <-  apply(Underdeveloped_Child_mortality[,2:30
2],2, mean)
```

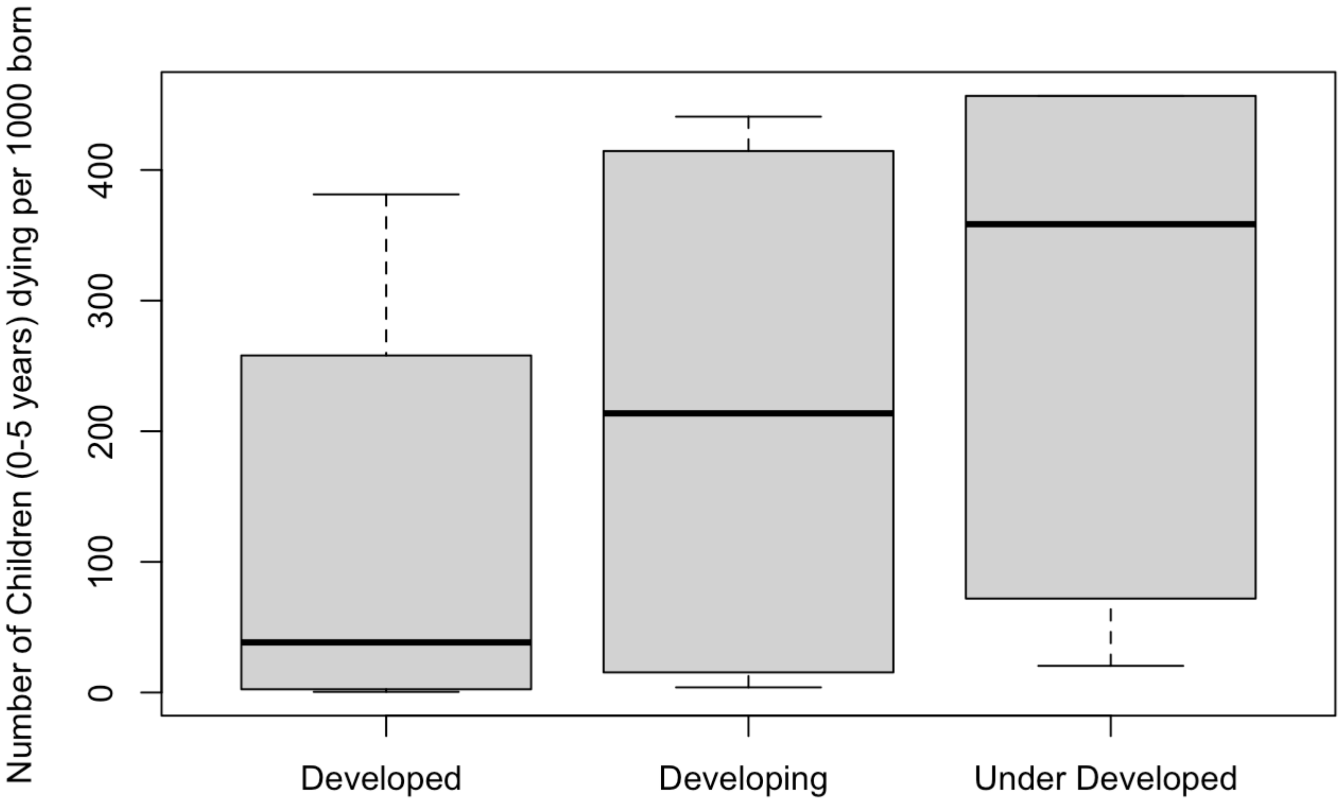
Plotting Graphs and making observations

To compare the trend of data values for Children mortality in developed, developing and under developed countries from the year 1800 to 2100 we will plot them on one graph and make certain hypothesis based on the observations we derive from that graph.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Child_mortality_Mean,col="red",pch=15,ylim = c(0,700),
     xlab=" Years ",ylab="Number of Children (0-5 years) dying per 1000 born")
points(c(1800:2100),Developing_Child_mortality_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Child_mortality_Mean,col="green")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Child mortality in Developed countries ","Child mortality in De
veloping countries","Child mortality in Under Developed countries"),
       fill = c("red","blue","green"))
```

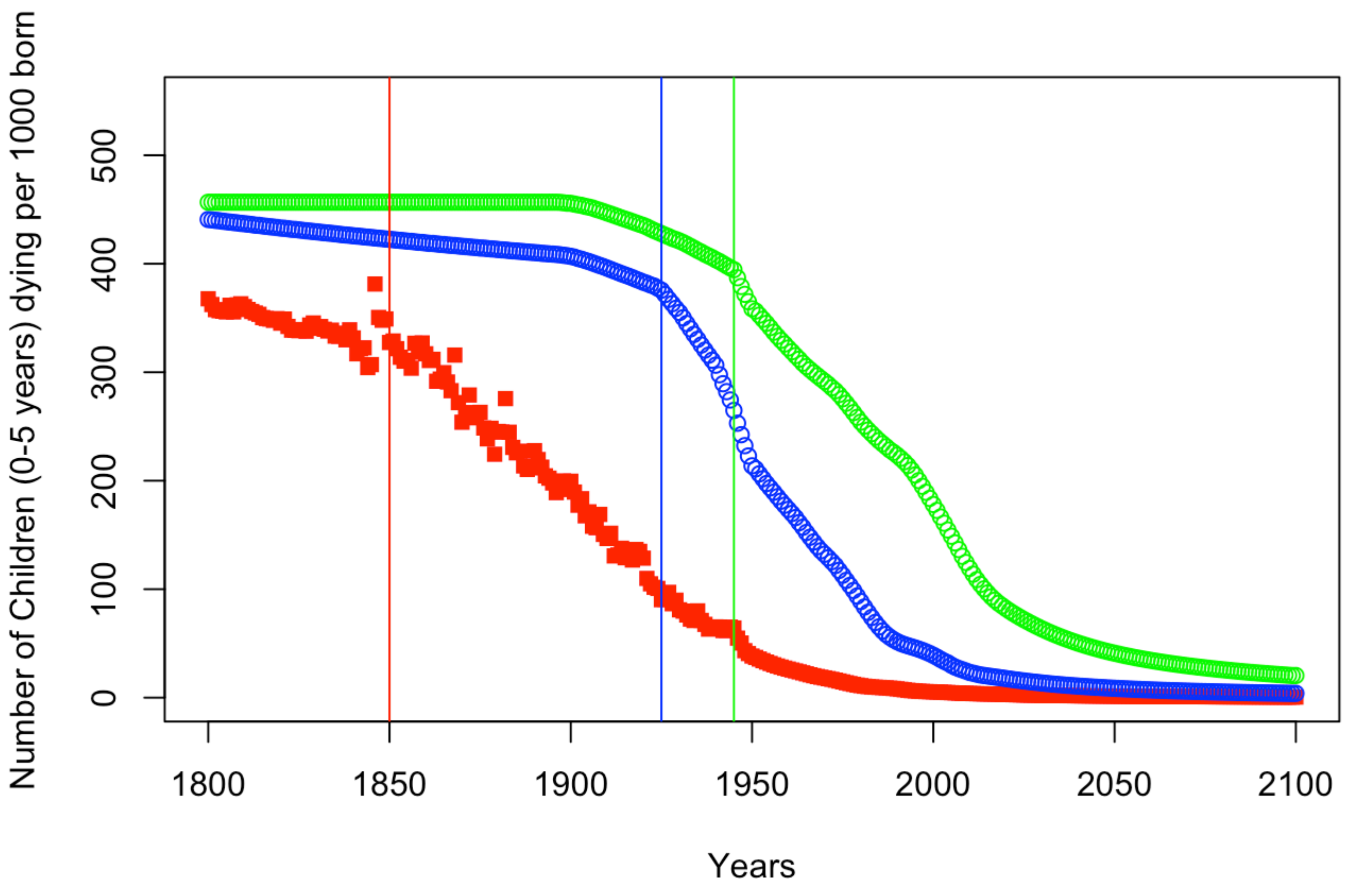



```
# Boxplots:
boxplot(Developed_Child_mortality_Mean,Developing_Child_mortality_Mean,Underdevelo
ped_Child_mortality_Mean,
        ylab = "Number of Children (0-5 years) dying per 1000 born",
        names=c("Developed ","Developing ","Under Developed "))
```



In the first graph we observe that there is certain deviation in the trend that each of the set of countries follow. In early 1800s Number of children dying per 1000 born is above 300 in case of all the three sets of countries, but we observe that it reduces significantly in later years. We can add vertical lines to our graph to check roughly where this drop occurs in different sets of countries.

```
# plotting data lits:
plot(c(1800:2100),Developed_Child_mortality_Mean,col="red",pch=15,ylim = c(0,550),
     xlab=" Years ",ylab="Number of Children (0-5 years) dying per 1000 born")
points(c(1800:2100),Developing_Child_mortality_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Child_mortality_Mean,col="green")

# Adding vertical lines which mark the beginning of deviation from general trend i
n early 1800s :

abline(v=1945,col="green")
abline(v=1925,col="blue")
abline(v=1850,col="red")
```



These are the following observations:

- Values for Under Developed countries remain roughly same from 1800 to 1940 but we see that it starts to drop in 1945s.
- Values for Developing countries remain roughly same from 1800 to 1920 but we see that it starts to drop in 1925.
- Values for Developed countries remain roughly same from 1800 to 1850 but we see that it starts to drop in 1850s.

Another interesting observation that can be drawn from the **boxplot** is that the median of Number of children dying per 1000 born for Developed countries is lower than that of Developing countries and median of Number of children dying per 1000 born in Developing countries is lower than that of Under Developed countries. To check this we can perform various Hypothesis tests:

1. We check whether the values of each type of countries have a normal distribution of not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used to compare our data.
2. Since we want to compare three populations, we can use **ANOVA** only if in earlier part we get that our data is normally distributed. If not then we will use pairwise **Wilcoxon test** to compare their median.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether values for each set of countries has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(Developed_Child_mortality_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developed_Child_mortality_Mean
## W = 0.78427, p-value < 2.2e-16
```

```
shapiro.test(Developing_Child_mortality_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developing_Child_mortality_Mean
## W = 0.77969, p-value < 2.2e-16
```

```
shapiro.test(Underdeveloped_Child_mortality_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Underdeveloped_Child_mortality_Mean
## W = 0.7866, p-value < 2.2e-16
```

For each type of countries we have the p-value less than 0.05. Hence, we reject our Null Hypothesis (that the data vectors are normally distributed) and conclude that the set of values for each type of country is not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare our datasets.

Now we will need a non parametric test to compare the means of our datasets pairwise. We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2100.

1. Wilcoxon test between Developed and Developing Countries:

- Null Hypothesis, H_0 := The difference between median value of Child mortality in Developed countries and of Developing countries is zero.
- Alternate Hypothesis, H_a := The difference between median value of Child mortality in Developed countries and of Developing countries is less than zero (i.e.median of Child mortality in Developed countries is less than that of Developing countries.)

```
wilcox.test(Developed_Child_mortality_Mean,Developing_Child_mortality_Mean, paired
=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developed_Child_mortality_Mean and Developing_Child_mortality_Mean
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is less than 0.05 hence we reject our Null hypothesis and get that **mediean of Child mortality in Developed countries is less than that of Developing countries.**

2. Wilcoxon test between Developing and Under Developed Countries:

- Null Hypothesis, H_0 := The difference between median value of Child mortality in of Developing countries and Under Developed countries is zero.
- Alternate Hypothesis, H_a := The difference between median value of Child mortality in Developing countries and Under Developed is less than zero (i.e.median of Child mortality in Developing countries is less than that of Under Developed countries.)

```
wilcox.test(Developing_Child_mortality_Mean,Underdeveloped_Child_mortality_Mean, p
aired=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developing_Child_mortality_Mean and Underdeveloped_Child_mortality_Mean
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is less than 0.05 hence we reject our Null hypothesis and get that **median of Child mortality in Developed countries is less than that of Developing countries.**

Hence we can conclude that the median of Number of children dying per 1000 born for Developed countries is lower than that of Developing countries and median of Number of children dying per 1000 born in Developing countries is lower than that of Under Developed countries.

Dataset : “Income_per_person”

Subsetting dataframe into different sets of countries as described above

```
head(Income_per_person)
```

	country	1800	1801	1802	1803	1804	1805	1806	1807
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	Afghanistan	603	603	603	603	603	603	603	603
2	Albania	667	667	667	667	667	668	668	668
3	Algeria	715	716	717	718	719	720	721	722
4	Andorra	1200	1200	1200	1200	1210	1210	1210	1210
5	Angola	618	620	623	626	628	631	634	637
6	Antigua and Barbuda	757	757	757	757	757	757	757	758

6 rows | 1-10 of 243 columns

Here when we look at our data, we can clearly see that we have one row for each country which represents sample data values for years 1800 to 2100 under the columns named by the respective year. We will now subset this dataset in the way we require i.e. we subset it into three smaller dataframes each of which will include the set of ten countries as selected intially.

```
# Dataframe for Developed Countries:

Developed_Income_per_person <- Income_per_person[Income_per_person$country %in% c(
"Norway" , "Ireland" , "Switzerland" , "Finland" , "Iceland" , "Germany" , "Sweden
" , "Australia", "Netherlands" , "Denmark"),]
Developed_Income_per_person
```

	country	1800	1801	1802	1803	1804	1805	1806	1807
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
9	Australia	817	822	826	831	836	841	845	850
47	Denmark	2010	2020	2020	2020	2030	2030	2030	2040
60	Finland	1230	1240	1240	1250	1250	1260	1260	1270
65	Germany	1990	2010	2020	2040	2050	2070	2080	2100
76	Iceland	926	926	927	927	927	927	927	927
81	Ireland	1460	1470	1480	1490	1500	1510	1520	1530
121	Netherlands	3330	3330	3330	3330	3330	3330	3330	3330
128	Norway	2530	2540	2540	2550	2550	2550	2560	2560
167	Sweden	1450	1440	1500	1490	1390	1480	1500	1430
168	Switzerland	2700	2700	2700	2700	2700	2700	2700	2700

1-10 of 10 rows | 1-10 of 243 columns

```
# Dataframe for Developing Countries:

Developing_Income_per_person <- Income_per_person[Income_per_person$country %in% c
("Algeria" , "Lebanon" , "Fiji" , "Moldova" , "Maldives" , "Tunisia" , "St. Vincen
t and the Grenadines" , "Suriname" , "Mongolia" , "Botswana"),]
Developing_Income_per_person
```

	country <chr>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>	1... <int>
3	Algeria	715	716	717	718	719	720	721	722
23	Botswana	397	397	397	397	397	398	398	398
59	Fiji	785	785	785	785	785	785	786	786
94	Lebanon	2150	2150	2150	2150	2160	2160	2160	2160
103	Maldives	842	843	843	843	843	843	843	843
111	Moldova	621	621	621	621	621	621	622	622
113	Mongolia	592	592	592	592	592	593	593	593
164	St. Vincent and the Grenadines	838	838	838	838	838	839	839	839
166	Suriname	1640	1640	1640	1640	1640	1640	1640	1640
177	Tunisia	715	715	716	716	716	716	716	716

1-10 of 10 rows | 1-10 of 243 columns

```
# Dataframe for Underdeveloped Countries:

Underdeveloped_Income_per_person <- Income_per_person[Income_per_person$country %i
n% c("Eritrea" , "Mozambique" , "Burkina Faso" , "Sierra Leone" , "Mali" , "Burund
i" , "South Sudan" , "Chad" , "Central African Republic" , "Niger"),]
Underdeveloped_Income_per_person
```

	country <chr>	1800 <int>	1801 <int>	1802 <int>	1803 <int>	1804 <int>	1805 <int>	1806 <int>	1807 <int>
27	Burkina Faso	480	480	480	480	480	480	480	480
28	Burundi	418	418	419	419	420	420	420	421
33	Central African Republic	424	424	424	424	424	424	424	425
34	Chad	418	418	418	418	418	418	418	419
55	Eritrea	532	532	532	532	532	532	532	533
104	Mali	603	603	603	603	603	603	603	603

116	Mozambique	390	391	391	391	391	392	392	392
124	Niger	446	446	446	446	446	446	446	447
151	Sierra Leone	734	734	734	734	734	734	735	735
159	South Sudan	507	507	507	507	508	508	508	508

1-10 of 10 rows | 1-10 of 243 columns

Taking mean per column of subsets

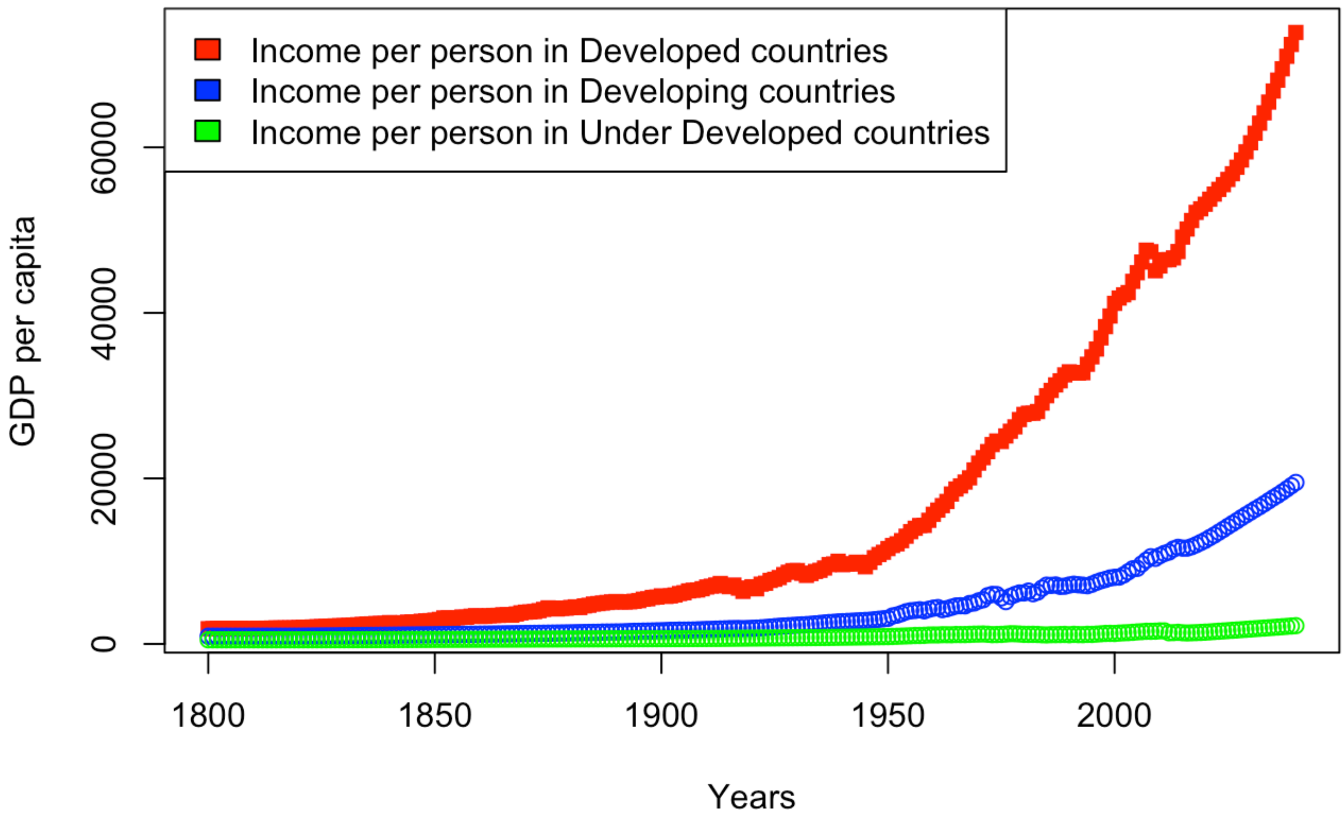
Now that we have created three subsets for sets of countries (Developed, Developing and Under Developed) we require for our analysis, we will take mean for values of each year for all the ten countries in the dataset.

```
# Taking mean of values for each set of countries and save it as a list:

Developed_Income_per_person_Mean <-  apply(Developed_Income_per_person[,2:242],2,
mean)
Developing_Income_per_person_Mean <-  apply(Developing_Income_per_person[,2:242],2
, mean)
Underdeveloped_Income_per_person_Mean <-  apply(Underdeveloped_Income_per_person[,
2:242],2, mean)
```

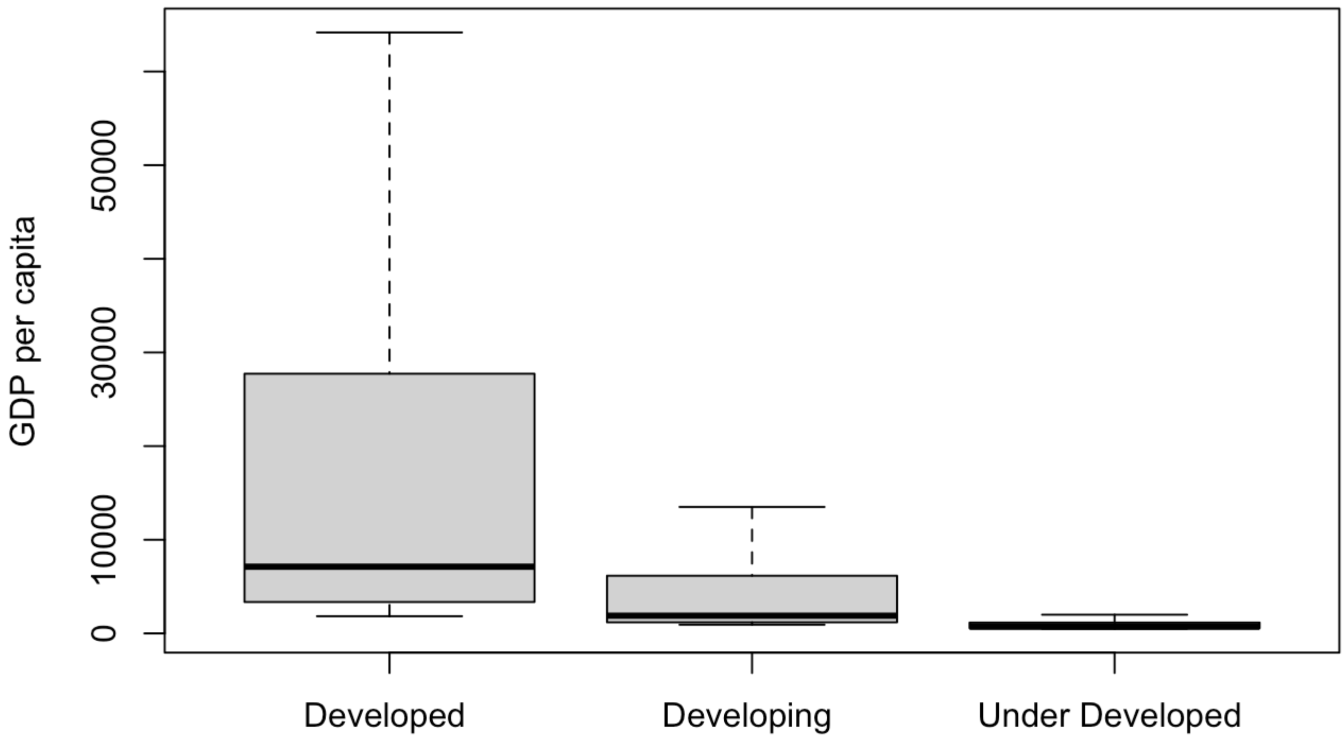
Plotting Graphs and making observations

To compare the trend of data values for Income per Person in developed, developing and under developed countries from the year 1800 to 2100 we will plot them on one graph and make certain hypothesis based on the observations we derive from that graph.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2040),Developed_Income_per_person_Mean,col="red",pch=15,
     xlab=" Years ",ylab="GDP per capita")
points(c(1800:2040),Developing_Income_per_person_Mean,col="blue")
points(c(1800:2040),Underdeveloped_Income_per_person_Mean,col="green")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Income per person in Developed countries ","Income per person i
n Developing countries","Income per person in Under Developed countries"),
       fill = c("red","blue","green"))
```




```
# Boxplots:
boxplot(Developed_Income_per_person_Mean,Developing_Income_per_person_Mean,Underde
veloped_Income_per_person_Mean,
        ylab = "GDP per capita",
        names=c("Developed ","Developing ","Under Developed "),outline=F)
```



An interesting observation that can be drawn from the **boxplot** is that the median of GDP per capita for Developed countries is greater than that of Developing countries and median of GDP per capita for Developing countries is greater than that of Under Developed countries. To check this we can perform various Hypothesis tests:

1. We check whether the values of each type of countries have a normal distribution of not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used to compare our data.
2. Since we want to compare three populations, we can use **ANOVA** only if in earlier part we get that our data is normally distributed. If not then we will use pairwise **Wilcoxon test** to compare their median.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether values for each set of countries has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(Developed_Income_per_person_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developed_Income_per_person_Mean
## W = 0.76903, p-value < 2.2e-16
```

```
shapiro.test(Developing_Income_per_person_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Developing_Income_per_person_Mean
## W = 0.74424, p-value < 2.2e-16
```

```
shapiro.test(Underdeveloped_Income_per_person_Mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Underdeveloped_Income_per_person_Mean
## W = 0.84715, p-value = 1.081e-14
```

For each type of countries we have the p-value less than 0.05. Hence, we reject our Null Hypothesis (that the data vectors are normally distributed) and conclude that the set of values for each type of country is not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare our datasets.

Now we will need a non parametric test to compare the means of our datasets pairwise. We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2100.

1. Wilcoxon test between Developed and Developing Countries:

- Null Hypothesis, H_0 := The difference between median value of GDP per capita in Developed countries and of Developing countries is zero.
- Alternate Hypothesis, H_a := The difference between median value of GDP per capita in Developed countries and of Developing countries is less than zero (i.e.median of GDP per capita in Developed countries is less than that of Developing countries.)

```
wilcox.test(Developed_Income_per_person_Mean,Developing_Income_per_person_Mean, pa
ired=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developed_Income_per_person_Mean and Developing_Income_per_person_Mean
## V = 29161, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is greater than 0.05 hence we cannot reject our Null hypothesis and get that **median of GDP per capita in Developed countries is greater than that of Developing countries.**

2. Wilcoxon test between Developing and Under Developed Countries:

- Null Hypothesis, H_0 := The difference between median value of GDP per capita in of Developing countries and Under Developed countries is zero.
- Alternate Hypothesis, H_a := The difference between median value of GDP per capita in Developing countries and Under Developed is less than zero (i.e.median of GDP per capita in Developing countries is less than that of Under Developed countries.)

```
wilcox.test(Developing_Income_per_person_Mean,Underdeveloped_Income_per_person_Mea
n, paired=TRUE,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  Developing_Income_per_person_Mean and Underdeveloped_Income_per_person_M
ean
## V = 29161, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

Since we get the p-value is greater than 0.05 hence we cannot reject our Null hypothesis and get that **median of GDP per capita in Developed countries is greater than that of Developing countries.**

Hence we can conclude that the median of GDP per capita for Developed countries is greater than that of Developing countries and median of GDP per capita for Developing countries is greater than that of Under Developed countries.

Dataset : “Life_expectancy”

Subsetting dataframe into different sets of countries as described above

```
head(Life_expectancy)
```

country	1800	1801	1802	1803	1804	1805	1806	1807
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>

1	Afghanistan	28.2	28.2	28.2	28.2	28.2	28.2	28.1	28.1
2	Albania	35.4	35.4	35.4	35.4	35.4	35.4	35.4	35.4
3	Algeria	28.8	28.8	28.8	28.8	28.8	28.8	28.8	28.8
4	Andorra	NA	NA	NA	NA	NA	NA	NA	NA
5	Angola	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0
6	Antigua and Barbuda	33.5	33.5	33.5	33.5	33.5	33.5	33.5	33.5

6 rows | 1-10 of 303 columns

Here when we look at our data, we can clearly see that we have one row for each country which represents sample data values for years 1800 to 2100 under the columns named by the respective year. We will now subset this dataset in the way we require i.e. we subset it into three smaller dataframes each of which will include the set of ten countries as selected intially.

```
# Dataframe for Developed Countries:

Developed_Life_expectancy <- Life_expectancy[Life_expectancy$country %in% c("Norwa
y" , "Ireland" , "Switzerland" , "Finland" , "Iceland" , "Germany" , "Sweden" , "A
ustralia", "Netherlands" , "Denmark"),]
Developed_Life_expectancy
```

	country <chr>	1800 <dbl>	1801 <dbl>	1802 <dbl>	1803 <dbl>	1804 <dbl>	1805 <dbl>	1806 <dbl>	1807 <dbl>
9	Australia	34.0	34.0	34.0	34.0	34.0	34.0	34.0	34.0
47	Denmark	37.4	38.5	44.4	44.8	42.8	43.0	43.8	42.6
60	Finland	36.6	40.3	39.2	28.5	35.9	39.8	38.8	36.6
65	Germany	38.4	38.4	38.4	38.4	38.4	38.4	38.4	38.4
76	Iceland	42.9	33.9	27.6	19.6	24.8	30.9	45.8	43.6
81	Ireland	38.3	38.3	38.3	38.3	38.3	38.3	38.3	38.3
119	Netherlands	39.9	39.9	39.9	39.9	39.9	39.9	39.9	39.9
126	Norway	37.9	35.8	38.4	38.7	40.5	44.3	43.8	41.8
162	Sweden	32.2	36.9	40.2	40.3	39.7	41.0	36.2	38.8
163	Switzerland	38.0	38.0	38.0	38.0	38.0	38.0	38.0	38.0

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Developing Countries:

Developing_Life_expectancy <- Life_expectancy[Life_expectancy$country %in% c("Alge
ria" , "Lebanon" , "Fiji" , "Moldova" , "Maldives" , "Tunisia" , "St. Vincent and
the Grenadines" , "Suriname" , "Mongolia" , "Botswana"),]
Developing_Life_expectancy
```

	country <chr>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>	1... <dbl>
3	Algeria	28.8	28.8	28.8	28.8	28.8	28.8	28.8	28.8
23	Botswana	33.6	33.6	33.6	33.6	33.6	33.6	33.6	33.6
59	Fiji	26.1	26.1	26.1	26.1	26.1	26.1	26.1	26.1
94	Lebanon	29.7	29.7	29.7	29.7	29.7	29.7	29.7	29.7
103	Maldives	32.6	32.6	32.6	32.6	32.6	32.6	32.6	32.6
111	Moldova	33.1	33.1	33.1	33.1	33.1	33.1	33.1	33.1
112	Mongolia	31.8	31.8	31.8	31.8	31.8	31.8	31.8	31.8
159	St. Vincent and the Grenadines	26.0	26.0	26.0	26.0	26.0	26.0	26.0	26.0
161	Suriname	32.9	32.9	32.9	32.9	32.9	32.9	32.9	32.9
172	Tunisia	28.8	28.8	28.8	28.8	28.8	28.8	28.8	28.8

1-10 of 10 rows | 1-10 of 303 columns

```
# Dataframe for Underdeveloped Countries:

Underdeveloped_Life_expectancy <- Life_expectancy[Life_expectancy$country %in% c("
Eritrea" , "Mozambique" , "Burkina Faso" , "Sierra Leone" , "Mali" , "Burundi" , "
South Sudan" , "Chad" , "Central African Republic" , "Niger"),]
Underdeveloped_Life_expectancy
```

	country <chr>	1800 <dbl>	1801 <dbl>	1802 <dbl>	1803 <dbl>	1804 <dbl>	1805 <dbl>	1806 <dbl>	1807 <dbl>
27	Burkina Faso	29.2	29.2	29.2	29.2	29.2	29.2	29.2	29.2
28	Burundi	31.5	31.5	31.5	31.5	31.5	31.5	31.5	31.5
33	Central African Republic	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
34	Chad	30.9	30.9	30.9	30.9	30.9	30.9	30.9	30.9
55	Eritrea	30.2	30.2	30.2	30.2	30.2	30.2	30.2	30.2
104	Mali	26.4	26.4	26.4	26.4	26.4	26.4	26.4	26.4

115	Mozambique	30.3	30.3	30.3	30.3	30.3	30.3	30.3	30.3
122	Niger	30.8	30.8	30.8	30.8	30.8	30.8	30.8	30.8
147	Sierra Leone	25.1	25.1	25.1	25.1	25.1	25.1	25.1	25.1
155	South Sudan	26.7	26.7	26.7	26.7	26.7	26.7	26.7	26.7

1-10 of 10 rows | 1-10 of 303 columns

Taking mean per column of subsets

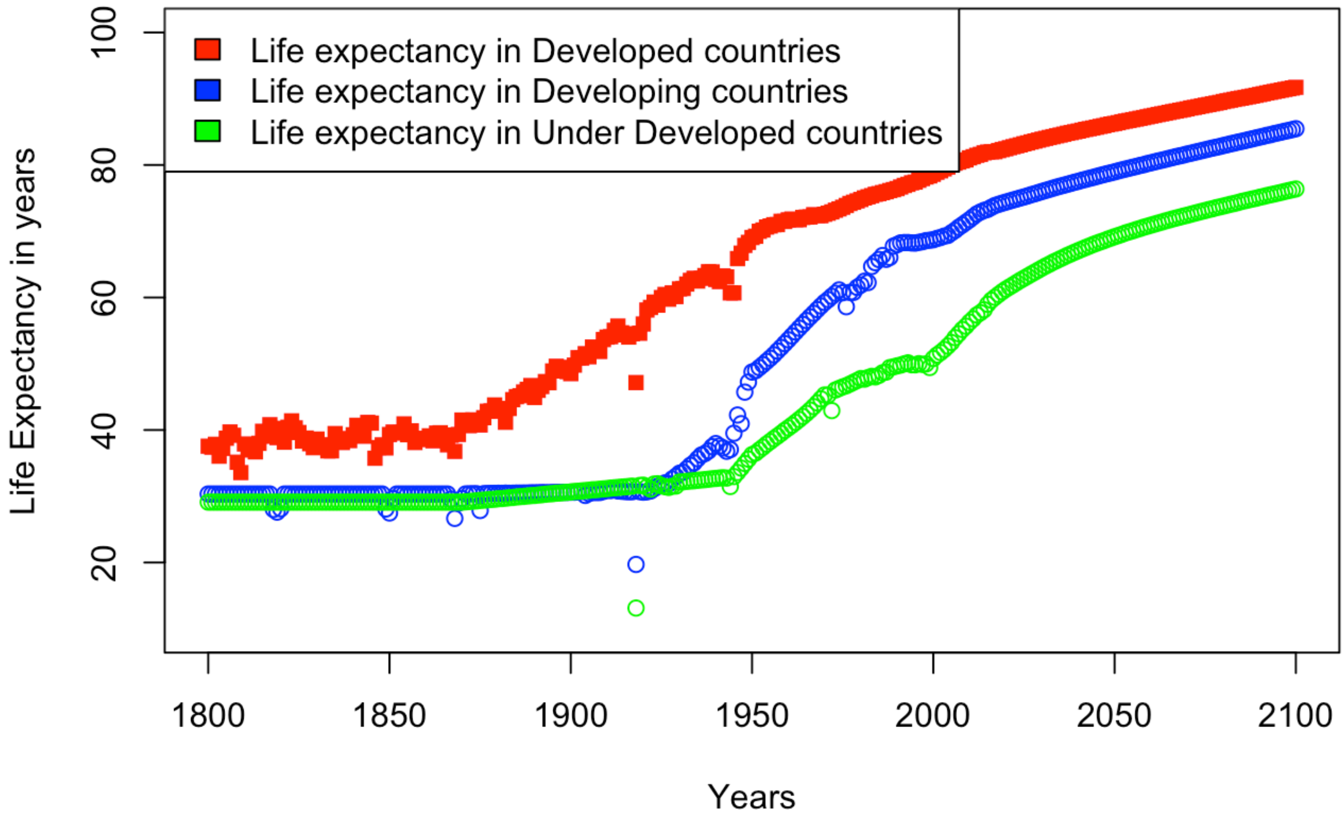
Now that we have created three subsets for sets of countries (Developed, Developing and Under Developed) we require for our analysis, we will take mean for values of each year for all the ten countries in the dataset.

```
# Taking mean of values for each set of countries and save it as a list:

Developed_Life_expectancy_Mean <-  apply(Developed_Life_expectancy[,2:302],2, mean
)
Developing_Life_expectancy_Mean <-  apply(Developing_Life_expectancy[,2:302],2, me
an)
Underdeveloped_Life_expectancy_Mean <-  apply(Underdeveloped_Life_expectancy[,2:30
2],2, mean)
```

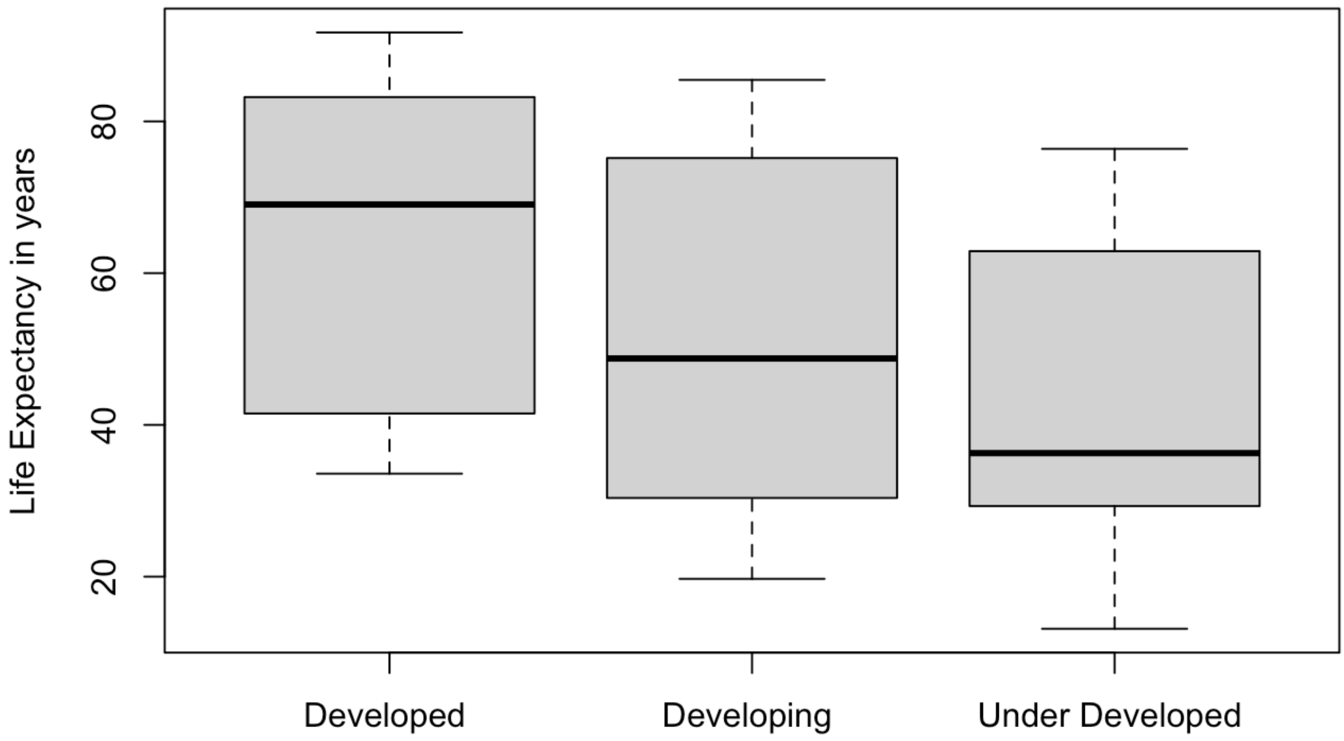
Plotting Graphs and making observations

To compare the trend of data values for Life Expectancy in developed, developing and under developed countries from the year 1800 to 2100 we will plot them on one graph and make certain hypothesis based on the observations we derive from that graph.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Life_expectancy_Mean,col="red",pch=15,ylim = c(10,100)
,
     xlab=" Years ",ylab="Life Expectancy in years")
points(c(1800:2100),Developing_Life_expectancy_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Life_expectancy_Mean,col="green")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Life expectancy in Developed countries ","Life expectancy in De
veloping countries","Life expectancy in Under Developed countries"),
       fill = c("red","blue","green"))
```

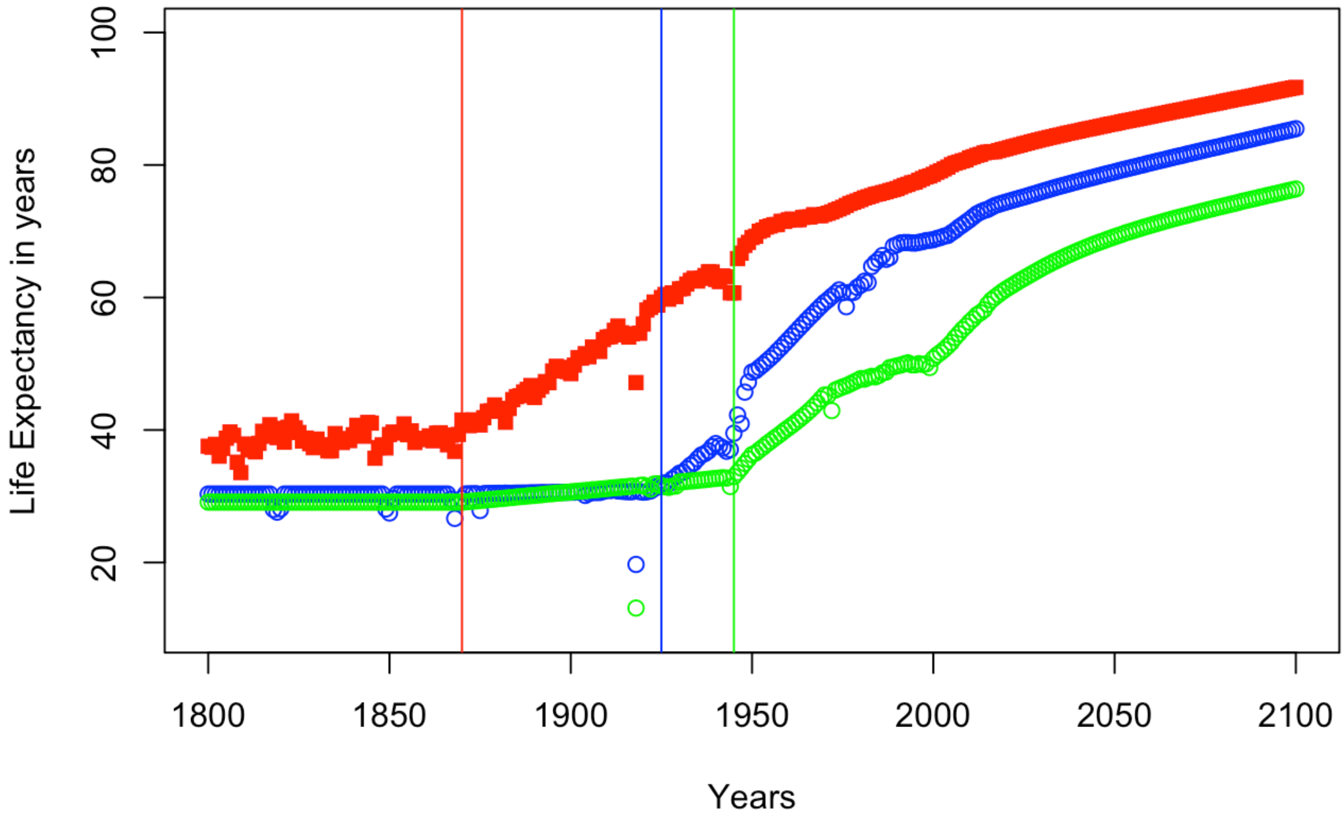



```
# Boxplots:
boxplot(Developed_Life_expectancy_Mean,Developing_Life_expectancy_Mean,Underdevelo
ped_Life_expectancy_Mean,
        ylab = "Life Expectancy in years",
        names=c("Developed ","Developing ","Under Developed "))
```



In the first graph we observe that there is certain deviation in the trend that each of the set of countries follow. In early 1800s Life expectancy is less than or equal to 40 years in case of all the three sets of countries, but we observe that it increases significantly in later years. We can add vertical lines to our graph to check roughly where this rise occurs in different sets of countries.

```
# plotting data lits:
plot(c(1800:2100),Developed_Life_expectancy_Mean,col="red",pch=15,ylim = c(10,100)
,
     xlab=" Years ",ylab="Life Expectancy in years")
points(c(1800:2100),Developing_Life_expectancy_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Life_expectancy_Mean,col="green")

# Adding vertical lines which mark the beginning of deviation from general trend i
n early 1800s :

abline(v=1945,col="green")
abline(v=1925,col="blue")
abline(v=1870,col="red")
```



These are the following observations:

- Values for Under Developed countries remain roughly same from 1800 to 1940 but we see that it starts to rise in 1945s.
- Values for Developing countries remain roughly same from 1800 to 1910s but we see that it starts to drop in 1925.
- Values for Developed countries remain roughly same from 1800 to 1860s but we see that it starts to drop in 1870.

Another interesting observation that can be drawn from the **boxplot** is that the median of Life expectancy in Developed countries is higher than that of Developing countries and median of Life expectancy in Developing countries is higher than that of Under Developed countries. To check this we can perform various Hypothesis tests:

1. We check whether the values of each type of countries have a normal distribution of not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used to compare our data.
2. Since we want to compare three populations, we can use **ANOVA** only if in earlier part we get that our data is normally distributed. If not then we will use pairwise **Wilcoxon test** to compare their median.

Checking Hypothesis

Part 2

Now that we have studied the general trends for Developed, Developing and Under Developed countries for each of the above gapminder dataset, we move on to comparing which trend does India follow in case of each of the discussed datasets.

Dataset : “Children_per_woman”

Getting vector of India Values

```
India_Children_per_woman <- na.omit(as.numeric(unlist(Children_per_woman[Children_
per_woman$country=="India",])))
```

```
## Warning in
## na.omit(as.numeric(unlist(Children_per_woman[Children_per_woman$country == : NA
s
## introduced by coercion
```

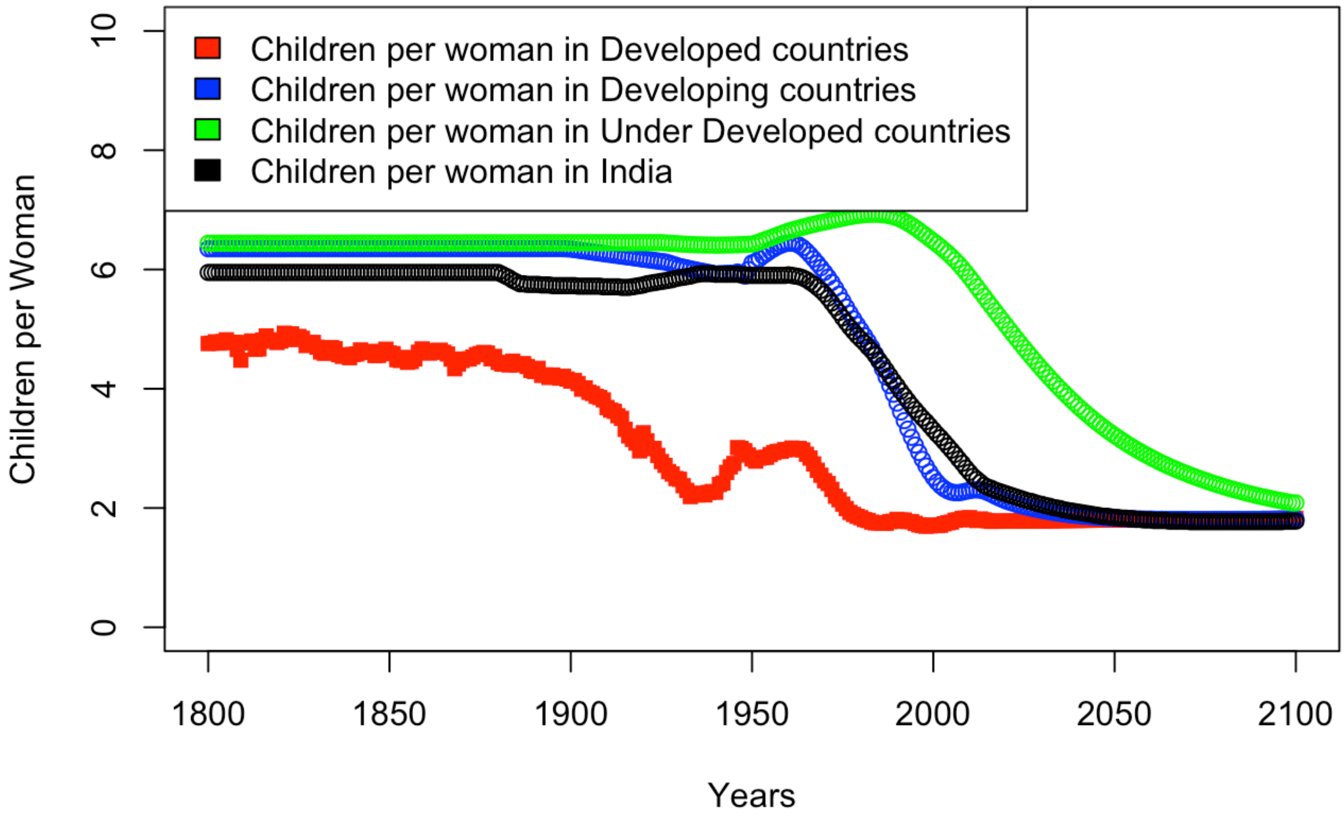
```
India_Children_per_woman
```

```
##   [1] 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.9
5
##  [16] 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.9
5
##  [31] 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.9
5
##  [46] 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.9
5
##  [61] 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.95 5.9
5
##  [76] 5.95 5.95 5.95 5.95 5.95 5.95 5.92 5.89 5.86 5.82 5.79 5.76 5.76 5.75 5.7
5
##  [91] 5.75 5.75 5.74 5.74 5.74 5.73 5.73 5.73 5.73 5.73 5.73 5.73 5.72 5.72 5.7
2
## [106] 5.72 5.72 5.72 5.72 5.71 5.71 5.71 5.71 5.71 5.70 5.70 5.70 5.71 5.72 5.7
3
## [121] 5.74 5.76 5.77 5.78 5.79 5.80 5.81 5.82 5.83 5.85 5.86 5.87 5.88 5.89 5.9
1
## [136] 5.92 5.93 5.93 5.93 5.93 5.92 5.92 5.92 5.92 5.92 5.92 5.92 5.91 5.91 5.9
0
## [151] 5.90 5.90 5.90 5.90 5.90 5.90 5.90 5.90 5.90 5.90 5.91 5.90 5.89 5.88 5.8
6
## [166] 5.83 5.79 5.75 5.70 5.65 5.59 5.52 5.44 5.36 5.28 5.19 5.11 5.03 4.96 4.8
9
## [181] 4.83 4.77 4.70 4.64 4.56 4.48 4.40 4.31 4.22 4.13 4.05 3.96 3.88 3.80 3.7
2
## [196] 3.65 3.58 3.51 3.45 3.38 3.31 3.24 3.18 3.11 3.04 2.97 2.90 2.82 2.75 2.6
7
## [211] 2.60 2.53 2.48 2.43 2.38 2.35 2.33 2.30 2.28 2.26 2.24 2.22 2.20 2.18 2.1
6
## [226] 2.14 2.12 2.11 2.09 2.07 2.06 2.04 2.03 2.01 2.00 1.99 1.97 1.96 1.95 1.9
4
## [241] 1.93 1.92 1.91 1.90 1.89 1.88 1.87 1.86 1.86 1.85 1.84 1.84 1.83 1.83 1.8
2
## [256] 1.82 1.81 1.81 1.80 1.80 1.80 1.79 1.79 1.79 1.79 1.78 1.78 1.78 1.78 1.7
8
## [271] 1.78 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.7
7
## [286] 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.77 1.78 1.78 1.7
8
## [301] 1.78
## attr(,"na.action")
## [1] 1
## attr(,"class")
## [1] "omit"
```

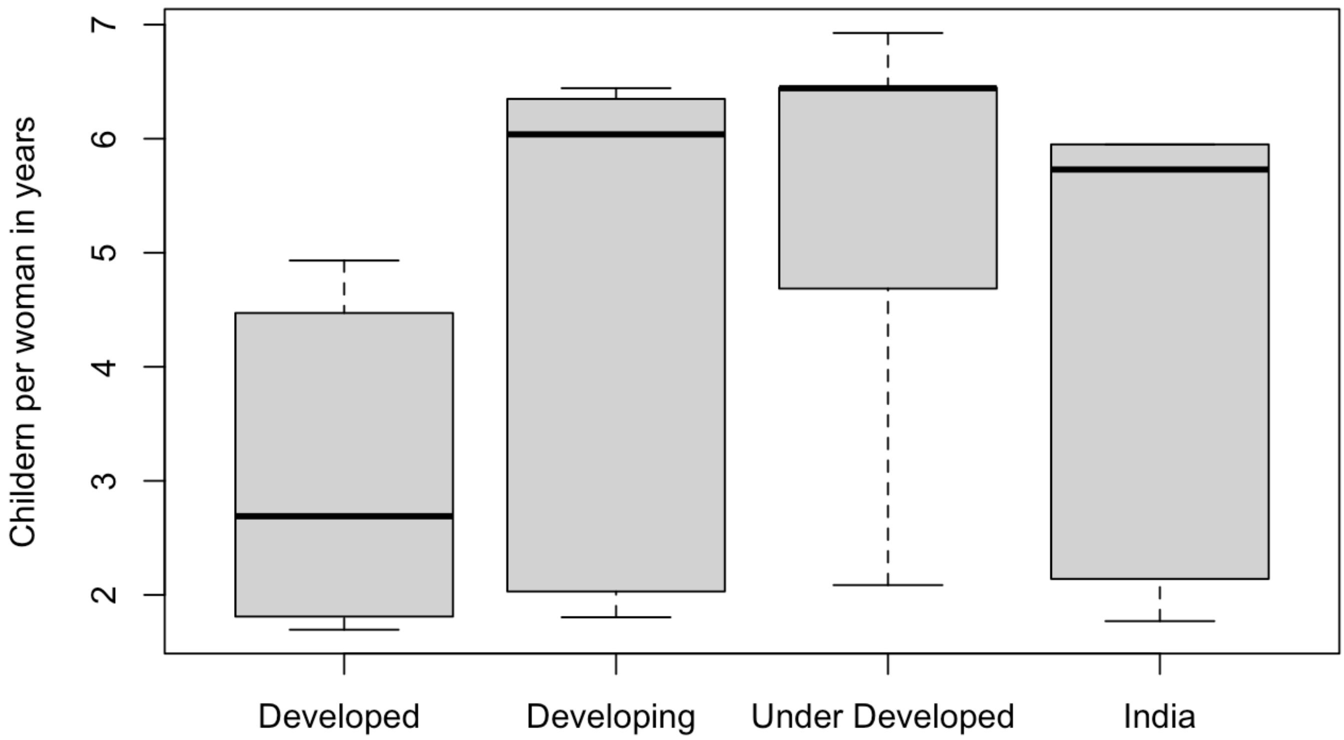
Plotting data vector of India along with different types of countries

We make a normal plot and boxplot for values of Children per woman in India from year 1800 to 2100 along with the values of Children per women in Developed, Developing and Under developed countries. Observing the plot we make certain observations.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Children_per_woman_Mean,col="red",pch=15,ylim = c(0,10
),
     xlab="Years",ylab="Children per Woman")
points(c(1800:2100),Developing_Children_per_woman_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Children_per_woman_Mean,col="green")
points(c(1800:2100),India_Children_per_woman,col="black")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Children per woman in Developed countries ","Children per woman
in Developing countries","Children per woman in Under Developed countries","Childr
en per woman in India"),
       fill = c("red","blue","green","black"))
```



```
# Boxplots:
boxplot(Developed_Children_per_woman_Mean,Developing_Children_per_woman_Mean,Under
developed_Children_per_woman_Mean,India_Children_per_woman,
        ylab = "Childern per woman in years",
        names=c("Developed ","Developing ","Under Developed ","India"))
```



Observations and Hypothesis

So from the first graph we can observe that the values of Children per woman in India closely follow the trend for Developing countries. In second graph (i.e Boxplot) we can observe that the mean of India is almost same as that for a Developing country. So we make our hypothesis that : India closly follows the trend of developing country and thus have a mean equal to that of a developing country. We check this hypothesis by following steps :

1. We check whether values of Children per woman in India are normally distributed or not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used in comparing mean of India and that of a developing country.
2. Depending on the result above we use a parametric or non-parametric test to determine whether thw mean of Children per woman in Developing countries is same as that for India.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether the values of Children per woman in India has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(India_Children_per_woman)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  India_Children_per_woman
## W = 0.7169, p-value < 2.2e-16
```

The p-value we get for Shapiro test is less than 0.05. Hence, we reject our Null Hypothesis that the values of Children per woman in India are normally distributed and conclude that they are not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare means of values of Children per Woman in India to that of Dveloping countries .

Now we will need a non parametric test to compare the means of our datasets . We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2100.

1. Wilcoxon test between India and Developing Countries:

- Null Hypothesis, H_0 := The difference between mean value of Children per women in India and of Developing countries is zero.
- Alternate Hypothesis, H_a := The difference between mean value of Children per women in India and of Developing countries is not zero.

```
wilcox.test(India_Children_per_woman,Developing_Children_per_woman_Mean,paired = T
)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Children_per_woman and Developing_Children_per_woman_Mean
## V = 9055, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

So we see that p value is much less than 0.05 so we reject the null hypothesis and must conclude that The difference between mean value of Children per women in Developed countries and of Developing countries is not zero. Now to get a better estimate in which range the value exactly lies we take one tail wilcoxon test for India in comparison with Developing country and Developed country.

```
wilcox.test(India_Children_per_woman,Developed_Children_per_woman_Mean,paired = T,
alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Children_per_woman and Developed_Children_per_woman_Mean
## V = 44190, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

```
wilcox.test(India_Children_per_woman,Developing_Children_per_woman_Mean,paired = T
,alternative = "greater")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Children_per_woman and Developing_Children_per_woman_Mean
## V = 9055, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

So clearly from the p-values we can conclude that the mean of Children per woman in India is greater than that of Developing countries but less than that of Developed countries.

Dataset : “Child_mortality”

Getting vector of India Values

```
India_Child_mortality <- na.omit(as.numeric(unlist(Child_mortality[Child_mortality
$country=="India",])))
```

```
## Warning in na.omit(as.numeric(unlist(Child_mortality[Child_mortality$country
## == : NAs introduced by coercion
```

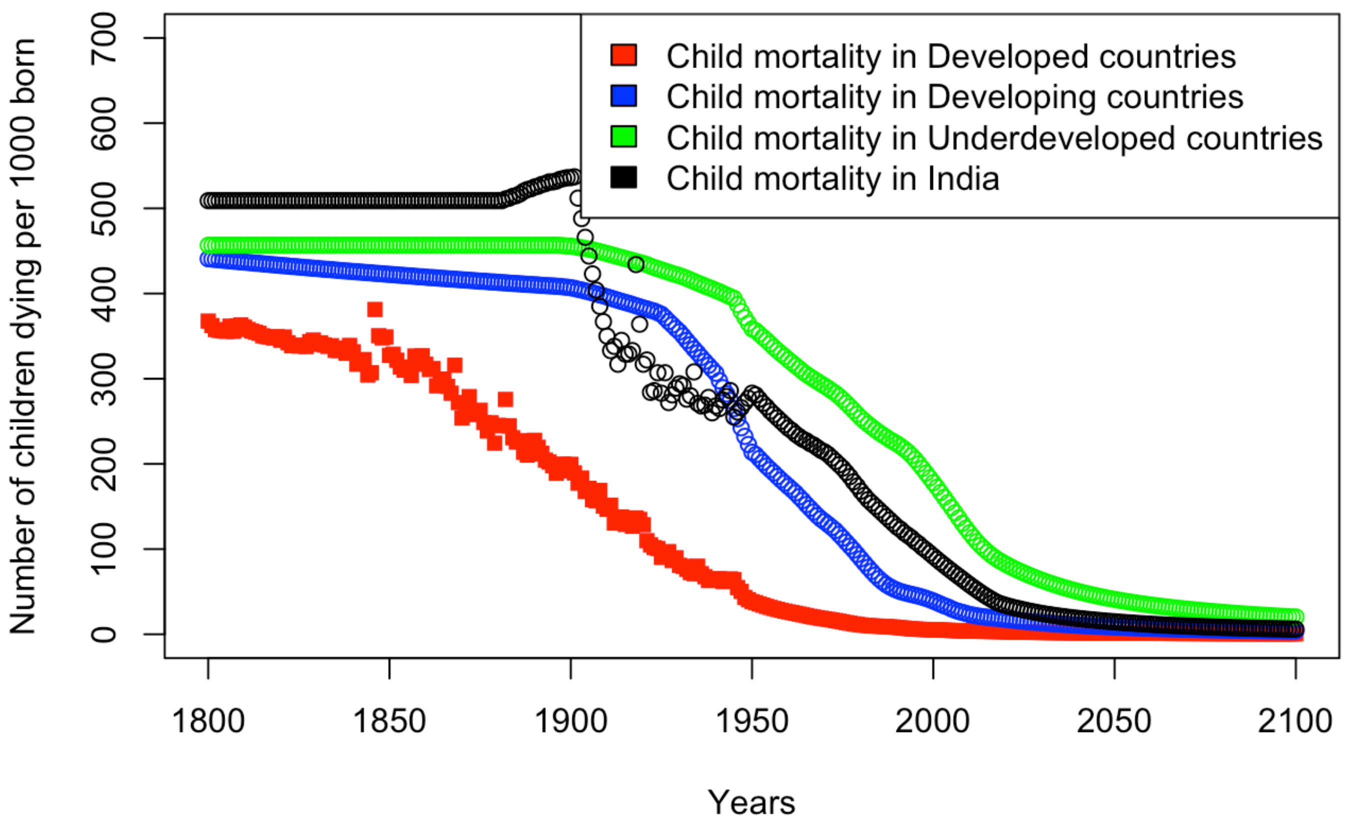
```
India_Child_mortality
```

```
##   [1] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [11] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [21] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [31] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [41] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [51] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [61] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [71] 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00 509.00
##  [81] 509.00 509.00 511.00 512.00 514.00 515.00 517.00 520.00 522.00 523.00
##  [91] 525.00 526.00 528.00 529.00 531.00 531.00 532.00 534.00 535.00 536.00
## [101] 536.00 537.00 512.00 488.00 466.00 444.00 423.00 404.00 385.00 367.00
## [111] 350.00 333.00 338.00 317.00 345.00 329.00 329.00 333.00 434.00 364.00
## [121] 317.00 322.00 284.00 286.00 307.00 283.00 307.00 272.00 281.00 289.00
## [131] 294.00 292.00 276.00 280.00 308.00 271.00 268.00 269.00 278.00 260.00
## [141] 268.00 265.00 275.00 279.00 286.00 255.00 261.00 266.00 272.00 277.00
## [151] 283.00 281.00 276.00 271.00 267.00 262.00 258.00 254.00 250.00 246.00
## [161] 242.00 239.00 235.00 232.00 229.00 227.00 224.00 222.00 219.00 216.00
## [171] 214.00 211.00 207.00 203.00 199.00 195.00 190.00 185.00 179.00 174.00
## [181] 168.00 163.00 158.00 154.00 150.00 146.00 142.00 138.00 134.00 130.00
## [191] 126.00 123.00 119.00 116.00 113.00 109.00 106.00 102.00  98.80  95.20
## [201]  91.60  88.00  84.50  81.10  77.70  74.40  71.10  67.90  64.70  61.40
## [211]  58.20  55.10  52.10  49.10  46.30  43.60  41.10  38.70  36.60  35.20
## [221]  33.90  32.70  31.60  30.60  29.60  28.60  27.70  26.80  26.00  25.20
## [231]  24.40  23.70  23.10  22.40  21.80  21.20  20.70  20.10  19.60  19.10
## [241]  18.60  18.20  17.70  17.30  16.90  16.50  16.10  15.70  15.30  15.00
## [251]  14.70  14.30  14.00  13.70  13.40  13.10  12.80  12.50  12.30  12.00
## [261]  11.80  11.50  11.30  11.10  10.80  10.60  10.40  10.20  10.10   9.87
## [271]   9.69   9.51   9.34   9.17   9.01   8.85   8.69   8.54   8.40   8.26
## [281]   8.13   8.00   7.87   7.75   7.63   7.51   7.40   7.28   7.18   7.07
## [291]   6.97   6.87   6.77   6.67   6.58   6.49   6.40   6.31   6.22   6.13
## [301]   6.13
## attr(,"na.action")
## [1] 1
## attr(,"class")
## [1] "omit"
```

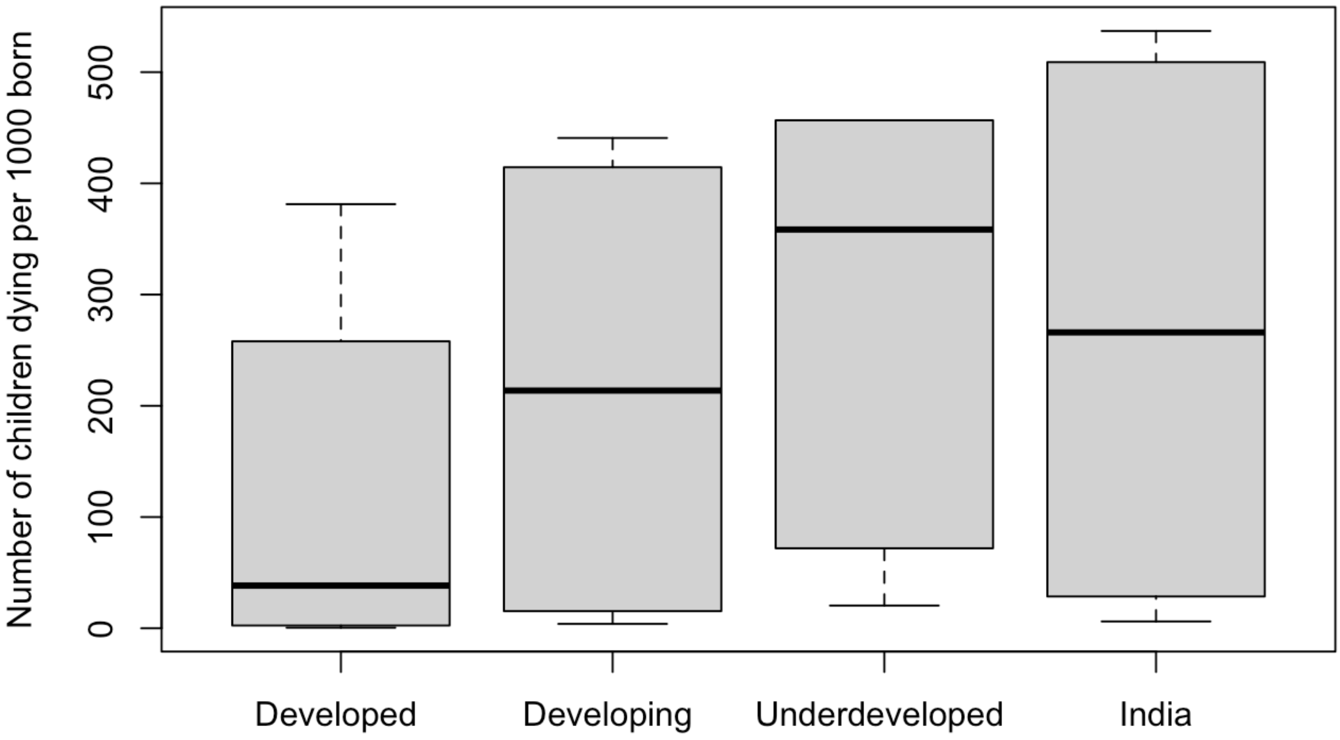
Plotting data vector of India along with different types of countries

We make a normal plot and boxplot for values of Number of children dying per 1000 born in India from year 1800 to 2100 along with the values of Number of children dying per 1000 born in Developed, Developing and Under developed countries. Observing the plot we make certain observations.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Child_mortality_Mean,col="red",pch=15,ylim = c(0,700),
     xlab="Years",ylab="Number of children dying per 1000 born")
points(c(1800:2100),Developing_Child_mortality_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Child_mortality_Mean,col="green")
points(c(1800:2100),India_Child_mortality,col="black")

# adding legends to graph:
legend(x = "topright",
       legend = c("Child mortality in Developed countries ","Child mortality in De
veloping countries","Child mortality in Underdeveloped countries","Child mortality
in India"),
       fill = c("red","blue","green","black"))
```




```
# Boxplots:
boxplot(Developed_Child_mortality_Mean,Developing_Child_mortality_Mean,Underdevelo
ped_Child_mortality_Mean,India_Child_mortality,
        ylab = "Number of children dying per 1000 born",
        names=c("Developed ","Developing ","Underdeveloped ","India"))
```



Observations and Hypothesis

So from the first graph we can observe that the values of Number of children dying per 1000 born in India does not follow any particular resemblance in trends of either Developing countries or Underdeveloped. So, in second graph (i.e Boxplot) we can observe that the mean of India is almost same as that for a Developing country. So we make our hypothesis that : India closly follows the trend of developing country and thus have a median equal to that of a developing country. We check this hypothesis by following steps :

1. We check whether values of Children per woman in India are normally distributed or not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used in comparing mean of India and that of a developing country.
2. Depending on the result above we use a parametric or non-parametric test to determine whether thw mean of Children per woman in Developing countries is same as that for India.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether the values of Number of children dying per 1000 born in India has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed

- Alternate Hypothesis, $H_a :=$ The population is **NOT** normally distributed

```
shapiro.test(India_Child_mortality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  India_Child_mortality
## W = 0.82893, p-value < 2.2e-16
```

The p-value we get for Shapiro test is less than 0.05. Hence, we reject our Null Hypothesis that the values of Number of children dying per 1000 born in India are normally distributed and conclude that they are not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare means of values of Number of children dying per 1000 born in India to that of Dveloping countries .

Now we will need a non parametric test to compare the means of our datasets . We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2100.

1. Wilcoxon test between India and Developing Countries:

- Null Hypothesis, $H_0 :=$ The difference between mean value of Number of children dying per 1000 born in India and of Developing countries is zero.
- Alternate Hypothesis, $H_a :=$ The difference between mean value of Number of children dying per 1000 born in India and of Developing countries is not zero.

```
wilcox.test(India_Child_mortality,Developing_Child_mortality_Mean,paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Child_mortality and Developing_Child_mortality_Mean
## V = 40379, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

So we see that p value is much less than 0.05 so we reject the null hypothesis and must conclude that The difference between mean value of umber of children dying per 1000 born in India and of Developing countries is not zero. Now to get a better estimate in which range the value exactly lies we take one tail wilcoxon test for India in comparison with Developing country and Under Developed country.

```
wilcox.test(India_Child_mortality,Underdeveloped_Child_mortality_Mean,paired = T,a
lternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Child_mortality and Underdeveloped_Child_mortality_Mean
## V = 15086, p-value = 2.016e-07
## alternative hypothesis: true location shift is less than 0
```

```
wilcox.test(India_Child_mortality,Developing_Child_mortality_Mean,paired = T,alter
native = "greater")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Child_mortality and Developing_Child_mortality_Mean
## V = 40379, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

So clearly from the p-values we can conclude that the median of umber of children dying per 1000 born in India is higher than that of Developing countries but less than that of Under Developed countries.

Dataset : “Income_per_person”

Getting vector of India Values

```
India_Income_per_person <- na.omit(as.numeric(unlist(Income_per_person[Income_per_
person$country=="India",])))
```

```
## Warning in na.omit(as.numeric(unlist(Income_per_person[Income_per_person$countr
y
## == : NAs introduced by coercion
```

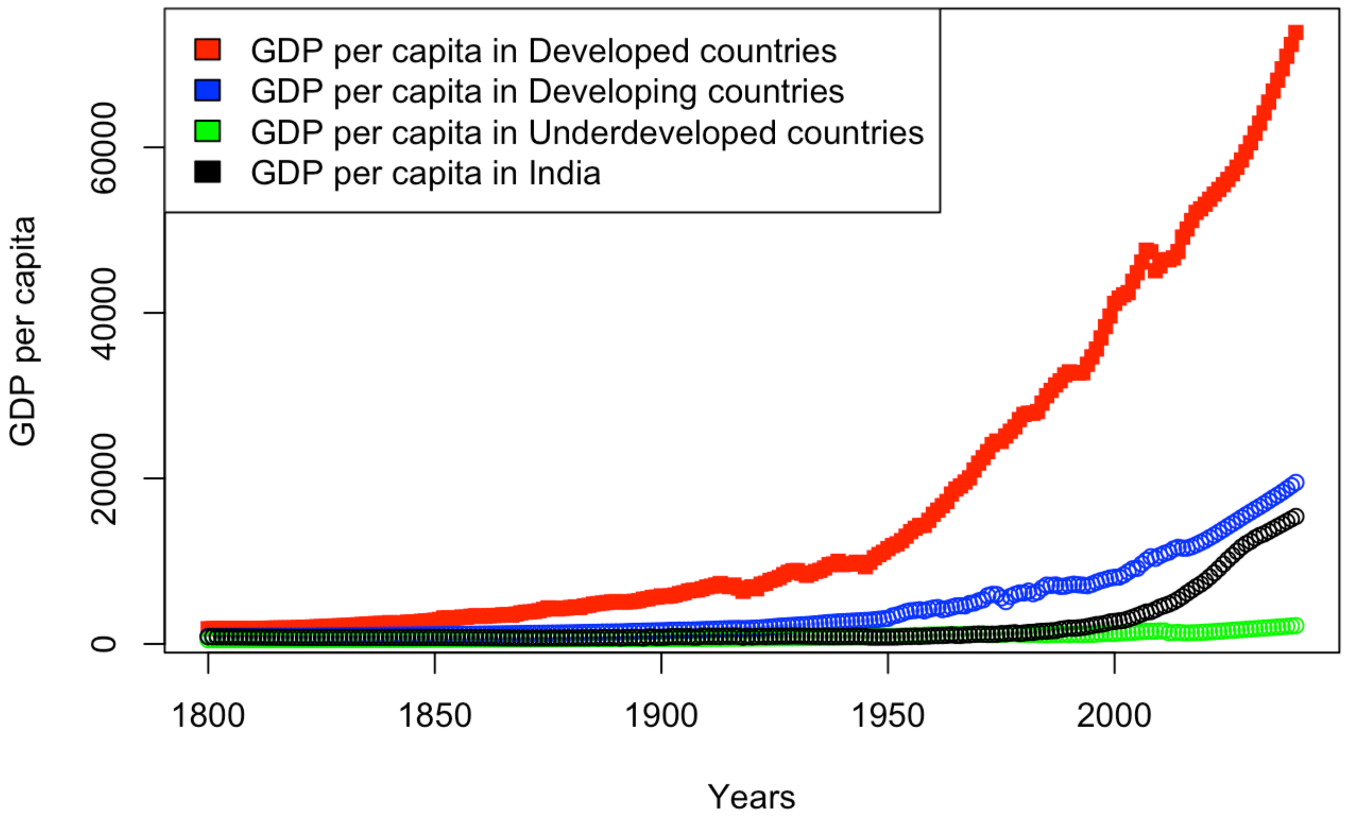
```
India_Income_per_person
```

```
##   [1]   863   862   858   854   849   845   841   837   833   829   825   821
##  [13]   817   813   809   805   801   797   794   790   786   782   783   783
##  [25]   784   785   785   786   787   788   788   789   789   789   789   789
##  [37]   788   788   788   788   788   788   788   788   789   789   789   789
##  [49]   789   790   790   790   786   781   777   773   768   764   760   755
##  [61]   751   747   743   739   734   730   726   722   718   714   710   712
##  [73]   713   715   716   718   720   721   723   725   726   728   730   731
##  [85]   733   754   729   761   766   744   777   705   760   778   788   768
##  [97]   710   838   838   772   797   809   871   878   877   856   875   817
## [109]   824   931   927   919   917   895   943   919   945   927   808   919
## [121]   845   904   933   893   928   929   949   940   940   969   966   946
## [133]   944   931   927   906   928   900   890   897   913   919   904   929
## [145]   909   884   828   822   821   831   824   829   838   874   894   899
## [157]   933   904   952   954  1000  1010  1020  1050  1110  1040  1030  1100
## [169]  1100  1160  1190  1180  1150  1180  1170  1250  1240  1310  1360  1260
## [181]  1330  1390  1400  1490  1520  1550  1590  1620  1760  1840  1910  1890
## [193]  1950  2000  2100  2210  2330  2380  2490  2660  2710  2790  2850  3020
## [205]  3210  3410  3630  3850  3910  4160  4450  4630  4820  5070  5380  5740
## [217]  6150  6520  6900  7230  7630  8100  8590  9100  9650 10200 10700 11200
## [229] 11700 12100 12400 12800 13100 13300 13600 13900 14200 14500 14800 15100
## [241] 15400
## attr(,"na.action")
## [1] 1
## attr(,"class")
## [1] "omit"
```

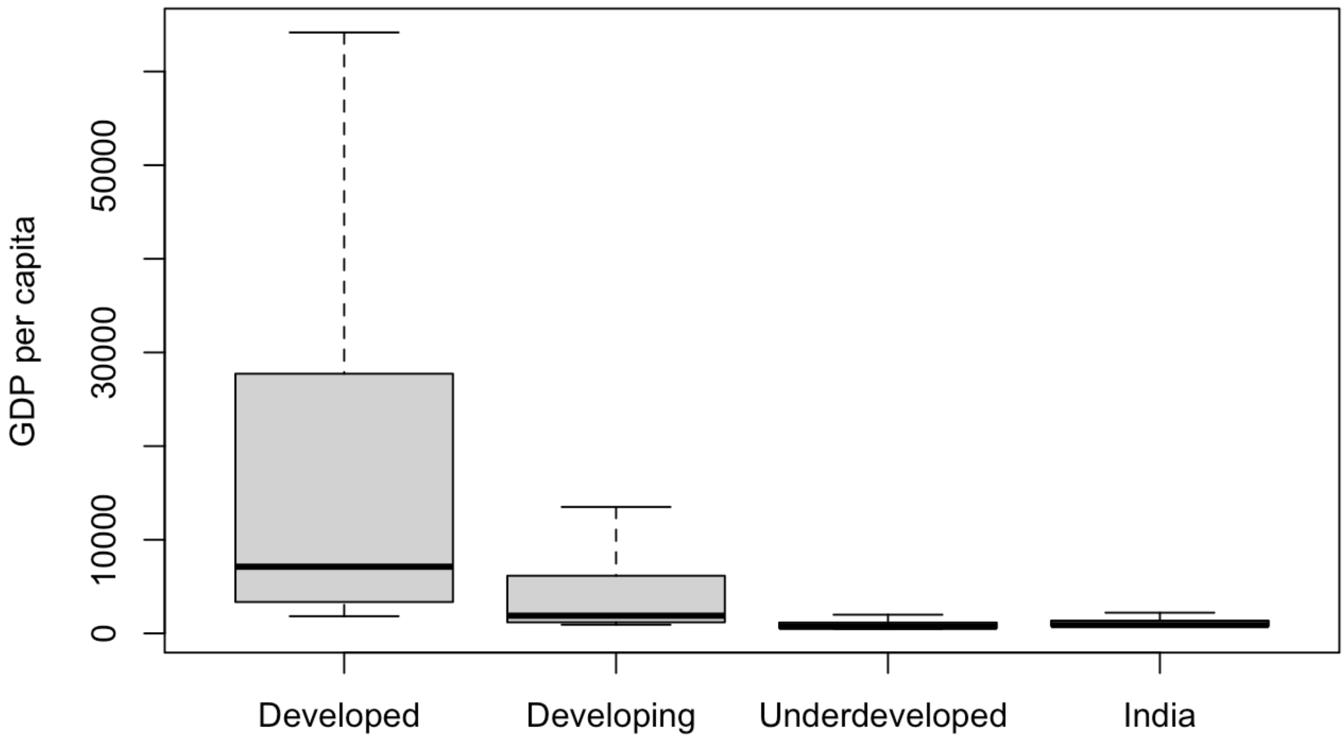
Plotting data vector of India along with different types of countries

We make a normal plot and boxplot for values of GDP per capita in India from year 1800 to 2100 along with the values of GDP per capita in Developed, Developing and Under developed countries. Observing the plot we make certain observations.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2040),Developed_Income_per_person_Mean,col="red",pch=15,
     xlab="Years",ylab="GDP per capita")
points(c(1800:2040),Developing_Income_per_person_Mean,col="blue")
points(c(1800:2040),Underdeveloped_Income_per_person_Mean,col="green")
points(c(1800:2040),India_Income_per_person,col="black")

# adding legends to graph:
legend(x = "topleft",
       legend = c("GDP per capita in Developed countries ","GDP per capita in Deve
loping countries","GDP per capita in Underdeveloped countries","GDP per capita in
India"),
       fill = c("red","blue","green","black"))
```




```
# Boxplots:
boxplot(Developed_Income_per_person_Mean,Developing_Income_per_person_Mean,Underdeveloped_Income_per_person_Mean,India_Income_per_person,
        ylab = "GDP per capita",
        names=c("Developed ","Developing ","Underdeveloped ","India"),outline = F)
```



Observations and Hypothesis

So from the first graph we can observe that the values of GDP per capita in India follows trend of Developing countries. In second graph (i.e Boxplot) we cannot clearly observe median of India is similar to that for a Developing country or Under developed country. So first we make our hypothesis that : India closly follows the trend of developing country and thus have a median equal to that of a developing country. We check this hypothesis by following steps :

1. We check whether values of GDP per capita in India are normally distributed or not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used in comparing mean of India and that of a developing country.
2. Depending on the result above we use a parametric or non-parametric test to determine whether the mean of GDP per capita in Developing countries is same as that for India.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether the values of GDP per capita in India has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(India_Income_per_person)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  India_Income_per_person
## W = 0.49894, p-value < 2.2e-16
```

The p-value we get for Shapiro test is less than 0.05. Hence, we reject our Null Hypothesis that the values of GDP per capita in India are normally distributed and conclude that they are not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare means of values of GDP per capita in India to that of Dveloping countries .

Now we will need a non parametric test to compare the means of our datasets . We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2040.

1. Wilcoxon test between India and Developing Countries:

- Null Hypothesis, H_0 := The difference between mean value of GDP per capita in India and of Developing countries is zero.
- Alternate Hypothesis, H_a := The difference between mean value of GDP per capita in India and of Developing countries is not zero.

```
wilcox.test(India_Income_per_person,Developing_Income_per_person_Mean,paired = T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Income_per_person and Developing_Income_per_person_Mean
## V = 0, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

So we see that p value is much less than 0.05 so we reject the null hypothesis and must conclude that The difference between mean value of GDP per capita in India and of Developing countries is not zero. Now to get a better estimate in which range the value exactly lies we take one tail wilcoxon test for India in comparison with Developing country and Under Developed country.

```
wilcox.test(India_Income_per_person,Underdeveloped_Income_per_person_Mean,paired =
T,alternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Income_per_person and Underdeveloped_Income_per_person_Mean
## V = 28601, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

```
wilcox.test(India_Income_per_person,Developing_Income_per_person_Mean,paired = T,a
lternative = "greater")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Income_per_person and Developing_Income_per_person_Mean
## V = 0, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

So clearly from the p-values we can conclude that the median of GDP per capita in India is less than that of Developing countries but greater than that of Under Developed countries.

Dataset : “Life_expectancy”

Getting vector of India Values

```
India_Life_expectancy <- na.omit(as.numeric(unlist(Life_expectancy[Life_expectancy
$country=="India",])))
```

```
## Warning in na.omit(as.numeric(unlist(Life_expectancy[Life_expectancy$country
## == : NAs introduced by coercion
```

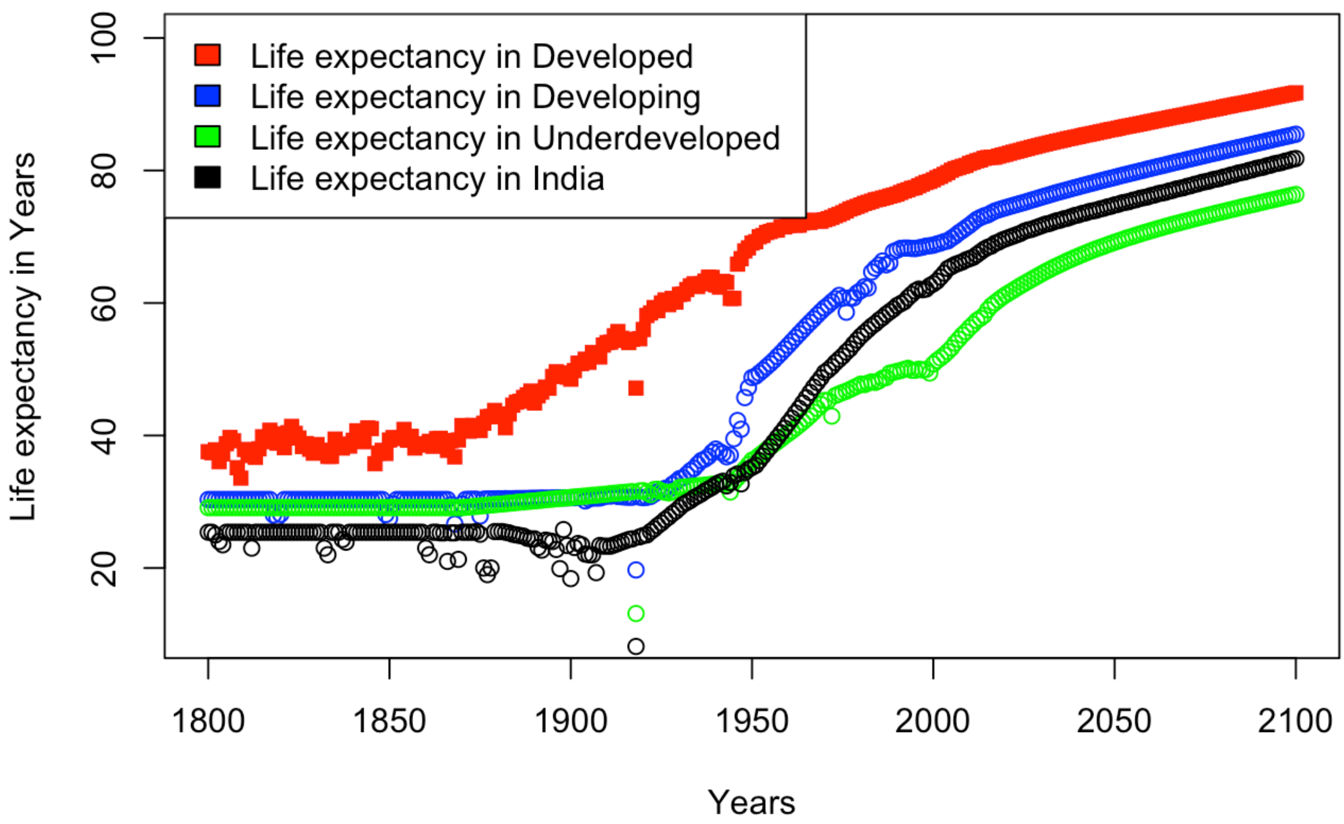
```
India_Life_expectancy
```

```
##   [1] 25.40 25.40 25.00 24.00 23.50 25.40 25.40 25.40 25.40 25.40 25.40 25.40
##  [13] 23.00 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40
##  [25] 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 23.00 22.00 25.40 25.40
##  [37] 25.40 24.30 23.90 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40
##  [49] 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40 25.40
##  [61] 23.00 22.00 25.40 25.40 25.30 25.30 21.00 25.30 25.30 21.30 25.40 25.40
##  [73] 25.40 25.40 25.50 25.10 20.00 19.00 20.00 25.50 25.50 25.50 25.40 25.30
##  [85] 25.10 25.00 24.90 24.80 24.50 24.40 24.40 23.10 22.70 24.20 24.10 24.00
##  [97] 22.80 19.90 25.80 23.40 18.40 23.10 23.70 23.50 22.10 22.00 22.00 19.30
## [109] 23.40 23.30 23.30 23.30 23.50 23.70 23.90 24.10 24.30 24.40  8.16 24.70
## [121] 24.90 25.00 25.50 25.90 26.40 26.80 27.30 27.70 28.20 28.60 29.10 29.60
## [133] 29.90 30.20 30.60 30.90 31.20 31.60 31.90 32.20 32.60 32.90 33.10 32.40
## [145] 32.90 33.90 34.20 32.70 34.40 34.90 35.20 35.50 36.20 36.90 37.60 38.30
## [157] 39.00 39.70 40.40 41.10 41.90 42.60 43.40 44.10 44.90 45.70 46.50 47.30
## [169] 48.00 48.70 49.50 49.90 50.40 51.00 51.50 52.00 52.60 53.10 53.80 54.40
## [181] 55.00 55.50 56.00 56.50 56.90 57.40 57.80 58.30 58.70 59.10 59.60 59.90
## [193] 60.20 60.80 61.30 61.80 62.10 62.00 62.10 62.60 62.90 63.30 63.90 64.50
## [205] 65.20 65.50 65.80 66.00 66.20 66.50 66.70 66.90 67.30 67.70 68.10 68.40
## [217] 68.60 69.00 69.20 69.50 69.70 69.90 70.10 70.30 70.50 70.80 71.00 71.10
## [229] 71.30 71.50 71.70 71.90 72.00 72.20 72.40 72.50 72.70 72.80 73.00 73.20
## [241] 73.30 73.50 73.60 73.80 73.90 74.00 74.20 74.30 74.50 74.60 74.80 74.90
## [253] 75.10 75.20 75.30 75.50 75.60 75.80 75.90 76.00 76.20 76.30 76.50 76.60
## [265] 76.70 76.90 77.00 77.20 77.30 77.40 77.60 77.70 77.90 78.00 78.20 78.30
## [277] 78.40 78.60 78.70 78.80 79.00 79.10 79.30 79.40 79.60 79.70 79.80 80.00
## [289] 80.10 80.30 80.40 80.50 80.70 80.80 81.00 81.10 81.20 81.40 81.50 81.70
## [301] 81.80
## attr(,"na.action")
## [1] 1
## attr(,"class")
## [1] "omit"
```

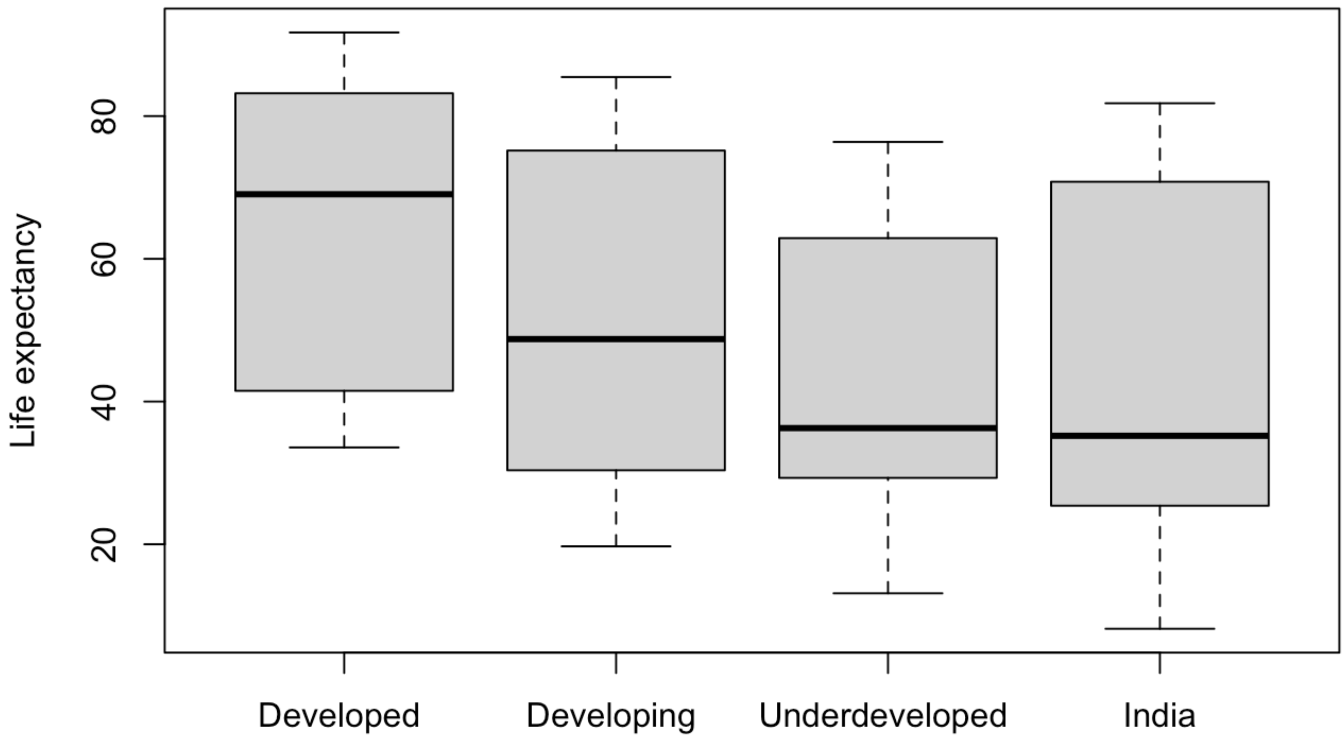
Plotting data vector of India along with different types of countries

We make a normal plot and boxplot for values of Life expectancy in India from year 1800 to 2100 along with the values of Life expectancy in Developed, Developing and Under developed countries. Observing the plot we make certain observations.

```
# plotting mean data vectors for each type of country:
plot(c(1800:2100),Developed_Life_expectancy_Mean,col="red",pch=15,ylim = c(10,100)
,
     xlab="Years",ylab="Life expectancy in Years")
points(c(1800:2100),Developing_Life_expectancy_Mean,col="blue")
points(c(1800:2100),Underdeveloped_Life_expectancy_Mean,col="green")
points(c(1800:2100),India_Life_expectancy,col="black")

# adding legends to graph:
legend(x = "topleft",
       legend = c("Life expectancy in Developed  ","Life expectancy in Developing
","Life expectancy in Underdeveloped ","Life expectancy in India"),
       fill = c("red","blue","green","black"))
```




```
# Boxplots:
boxplot(Developed_Life_expectancy_Mean,Developing_Life_expectancy_Mean,Underdevelo
ped_Life_expectancy_Mean,India_Life_expectancy,
        ylab = "Life expectancy",
        names=c("Developed ","Developing ","Underdeveloped ","India"),outline = F)
```



Observations and Hypothesis

So from the first graph we can observe that the values of Life Expectancy in India does not follow any particular resemblance in trends of either Developing countries or Underdeveloped. So, in second graph (i.e Boxplot) we can observe that the mean of India is almost same as that for a Underdeveloped country. So we make our hypothesis that : India cloesly follows the trend of Underdeveloped country and thus have a median equal to that of a Underdeveloped country. We check this hypothesis by following steps :

1. We check whether values of Life Expectancy in India are normally distributed or not using **Shapiro-Wilk Normality Test**. This is necessary since it will determine which kind of test (parametric or non parametric) can be used in comparing mean of India and that of a developing country.
2. Depending on the result above we use a parametric or non-parametric test to determine whether the mean of Life Expectancy in Developing countries is same as that for India.

Checking Hypothesis

First we will perform **Shapiro-Wilk Normality Test** to check whether the values of Life Expectancy in India has normal distribution.

- Null Hypothesis, H_0 := The population is normally distributed
- Alternate Hypothesis, H_a := The population is **NOT** normally distributed

```
shapiro.test(India_Life_expectancy)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  India_Life_expectancy
## W = 0.82507, p-value < 2.2e-16
```

The p-value we get for Shapiro test is less than 0.05. Hence, we reject our Null Hypothesis that the values of Life_expectancy in India are normally distributed and conclude that they are not normally distributed. Thus we cannot use **t test** or **ANOVA** to compare means of values of Life expectancy in India to that of Underdeveloped countries .

Now we will need a non parametric test to compare the means of our datasets . We will use **Wilcoxon test** to compare the mean. Note that we take our data vectors to be paired because the values have been taken under similar conditions from the years 1800 to 2040.

1. Wilcoxon test between India and Underdeveloped Countries:

- Null Hypothesis, $H_0 :=$ The difference between mean value of Life Expectancy in India and of Underdeveloped countries is zero.
- Alternate Hypothesis, $H_a :=$ The difference between mean value of Life Expectancy in India and of Underdeveloped countries is not zero.

```
wilcox.test(India_Life_expectancy,Underdeveloped_Life_expectancy_Mean,paired = T)
```

```
## 
##  Wilcoxon signed rank test with continuity correction
## 
## data:  India_Life_expectancy and Underdeveloped_Life_expectancy_Mean
## V = 27784, p-value = 0.0008098
## alternative hypothesis: true location shift is not equal to 0
```

So we see that p value is much less than 0.05 so we reject the null hypothesis and must conclude that The difference between mean value of Life_expectancy in India and of Underdeveloped countries is not zero. Now to get a better estimate in which range the value exactly lies we take one tail wilcoxon test for India in comparison with Developing country and Under Developed country.

```
wilcox.test(India_Life_expectancy,Underdeveloped_Life_expectancy_Mean,paired = T,a
lternative = "less")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Life_expectancy and Underdeveloped_Life_expectancy_Mean
## V = 27784, p-value = 0.9996
## alternative hypothesis: true location shift is less than 0
```

```
wilcox.test(India_Life_expectancy,Developing_Life_expectancy_Mean,paired = T,alter
native = "greater")
```

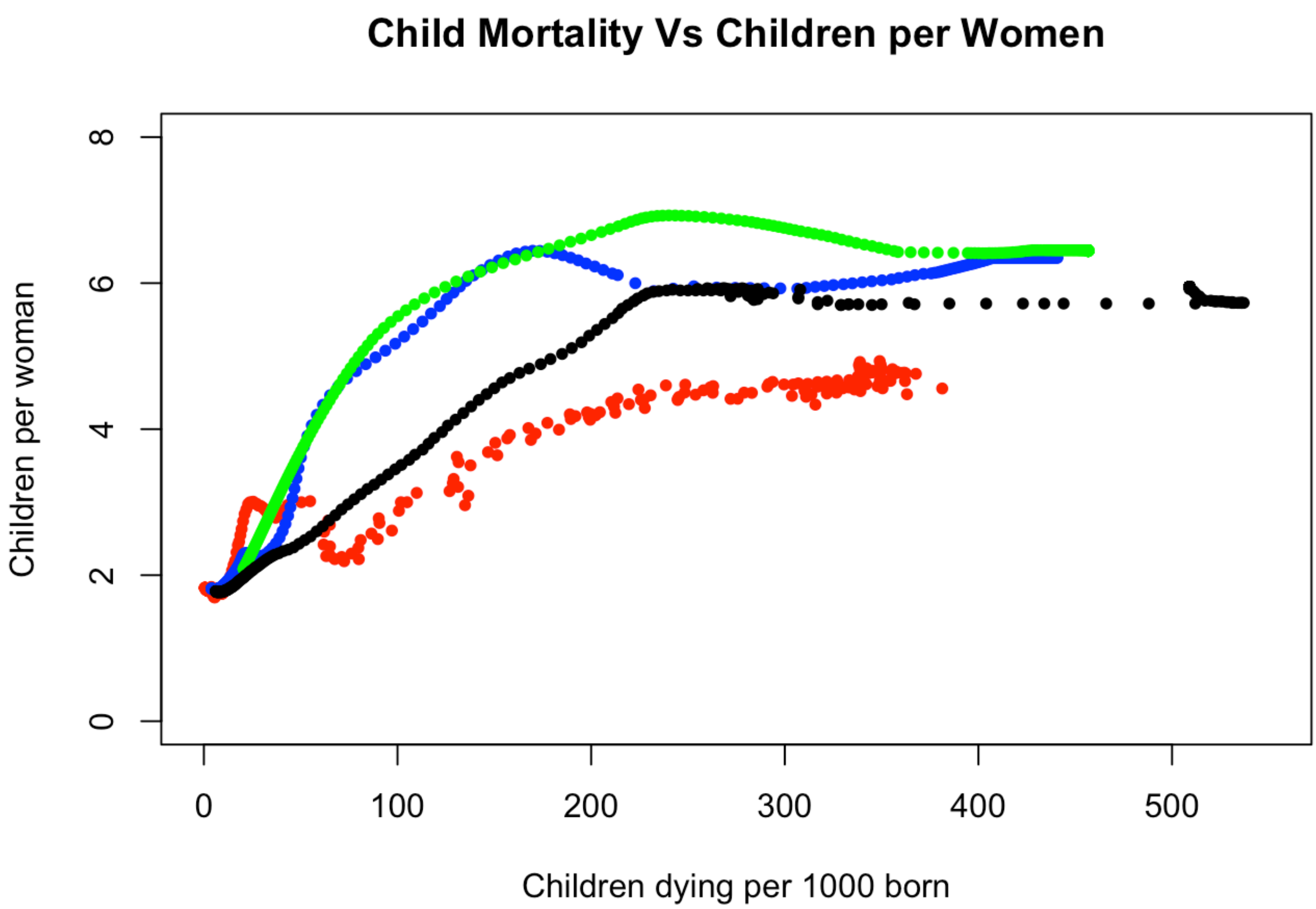
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  India_Life_expectancy and Developing_Life_expectancy_Mean
## V = 0, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

So clearly from the p-values we can conclude that the median of Life expectancy in India is less than that of Developing countries but greater than that of Under Developed countries.

Part 3

Child Mortality Vs Children per Women :

```
plot(Developed_Child_mortality_Mean, Developed_Children_per_woman_Mean, col = "red
", pch = 20,xlab = "Children dying per 1000 born", ylab = "Children per woman", ma
in = "Child Mortality Vs Children per Women",ylim = c(0,8), xlim = c(0,550) )
points(Developing_Child_mortality_Mean, Developing_Children_per_woman_Mean, col= "
blue", pch = 20)
points(Underdeveloped_Child_mortality_Mean, Underdeveloped_Children_per_woman_Mean
, col ="green", pch= 20)
points(India_Child_mortality, India_Children_per_woman, col = "black", pch = 20)
```



Observing the graph we can say that with increase in Children per women , Child mortality also increases. This Hypothesis can be tested using correlation test.

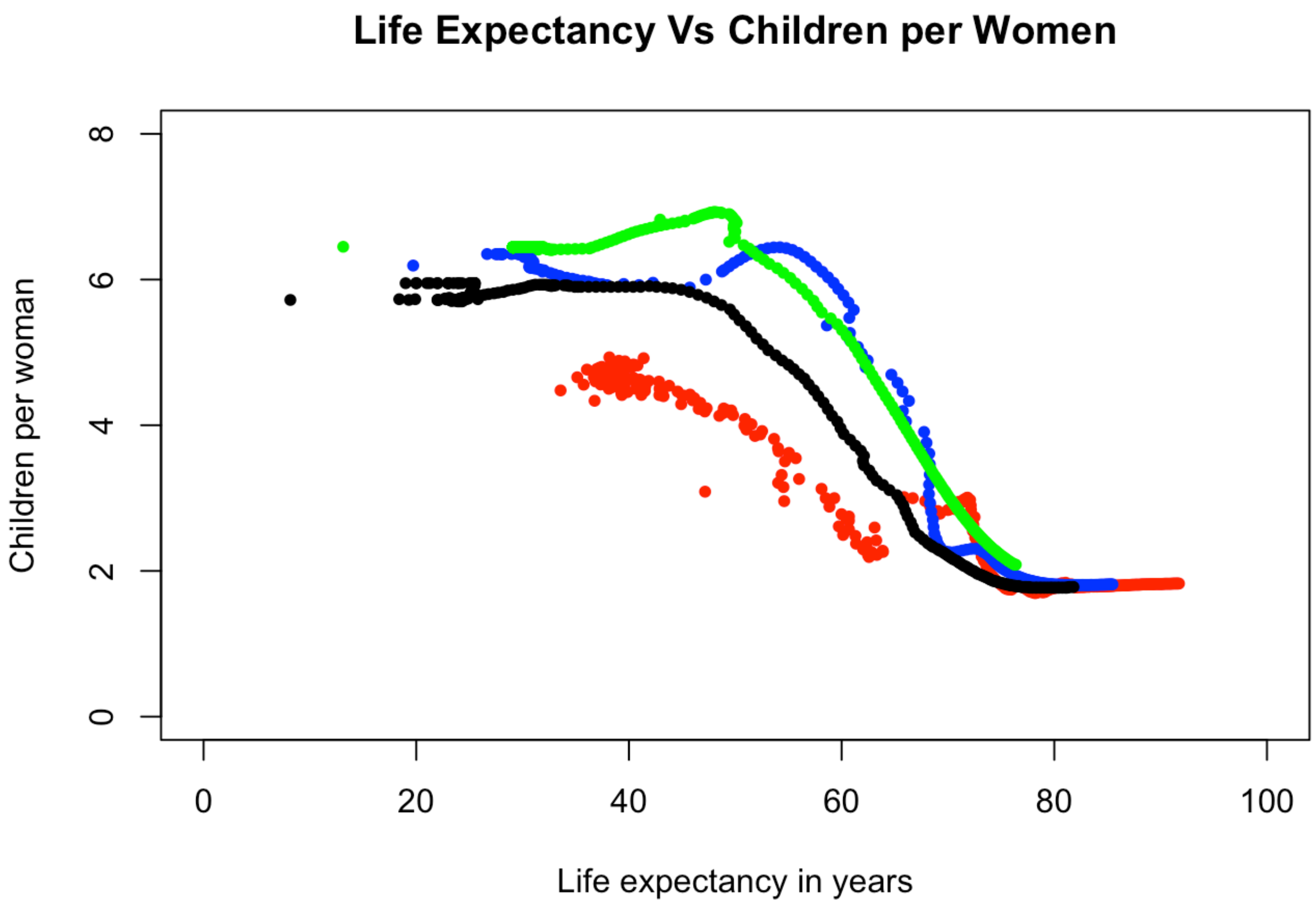
```
Children_per_woman_Mean <-  apply(Children_per_woman[,2:302],2, mean)
Child_mortality_Mean <-  apply(Child_mortality[,2:302],2, mean)

cor.test(Children_per_woman_Mean, Child_mortality_Mean )
```

```
##
##  Pearson's product-moment correlation
##
## data:  Children_per_woman_Mean and Child_mortality_Mean
## t = 46.582, df = 149, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9552002 0.9762289
## sample estimates:
##       cor
## 0.9673391
```

Since the correlation value is positive we say that with ncrease in Children per women , Child mortality also increases.

Life Expectancy Vs Children per Women :

```
plot(Developed_Life_expectancy_Mean, Developed_Children_per_woman_Mean, col = "red
", pch = 20,xlab = "Life expectancy in years", ylab = "Children per woman", main =
"Life Expectancy Vs Children per Women",ylim = c(0,8), xlim = c(0,100) )
points(Developing_Life_expectancy_Mean, Developing_Children_per_woman_Mean, col= "
blue", pch = 20)
points(Underdeveloped_Life_expectancy_Mean, Underdeveloped_Children_per_woman_Mean
, col ="green", pch= 20)
points(India_Life_expectancy, India_Children_per_woman, col = "black", pch = 20)
```



Observing the graph we can say that with decrease in Children per women , Life expectancy of individual also increases. This Hypothesis can be tested using correlation test.

```
Life_expectancy_Mean <- apply(Life_expectancy[,2:302],2, mean)

cor.test(Children_per_woman_Mean, Life_expectancy_Mean )
```

```
##
##  Pearson's product-moment correlation
##
## data:  Children_per_woman_Mean and Life_expectancy_Mean
## t = -28.833, df = 46, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9851012 -0.9528380
## sample estimates:
##        cor
## -0.9734314
```

As the correlation is negative we can say that with decrease in Children per women there is a increase in values of Life expectancy