**National Technology of Mexico**
**Technological Institute of Tijuana**

ACADEMIC SUBDIRECTION
Systems and Computing Department

SEMESTER
February - June 2020

ACADEMIC CAREER
Information and Communication Technologies Engineer

SUBJECT AND KEY:
Data Mining BDD-1703TI9A

STUDENT'S NAME AND REGISTRATION:
Velázquez Farrera César Alejandro 17212937
Camacho Manabe Juan Daniel 17210534

NAME OF THE JOB:
Exam

UNIT TO BE EVALUATED
Unit I

TEACHER'S NAME:
Mc José Christian Romero Hernández

# Exam

**Scenario**: The World Bank was very impressed with your delivery on the previous assignment and they have a new project for you.

You should generate a scatter-plot that shows the statistics of life expectancy (Life expectancy - y-axis) and fertility rate (Fertility Rate -x-axis) by country (Country).

The scatterplot should also be classified by Country Regions.

You have been provided data for 2 years: 1960 and 2013 and you are required to produce a visualization for each of these years.

Some data has been provided in a CVS file, some in R vectors. The CVS file contains combined data from both years. All data manipulation must be done in R (Not Excel) because this project can be audited at a later stage.

You have also been asked to provide information on how the two periods compare. (Hint: Basically the explanation of his observations)

### Data Vectors

1. To obtain the vectors of the countries and their respective fertility rates, a variable must be created to accommodate the data.

```
#Leer el archivo csv...
stats <- read.csv(file.choose())
stats
```

2. From the file "test-Vectors.R" the vectors must be copied to the R file that is being used and then executed to load the information.

```
Country_Code <-
c("ABW","AFG","AGO","ALB","ARE","ARG","ARM","ATG","AUS","AUT","AZE","
BDI","BEL","BEN","BFA","BGD","BGR","BHR","BHS","BIH","BLR","BLZ","BOL
","BRA","BRB","BRN","BTN","BWA","CAF","CAN","CHE","CHL",...más en los
archivos)

Life_Expectancy_At_Birth_1960 <-
c(65.5693658536586,32.328512195122,32.9848292682927,62.2543658536585,
52.2432195121951,65.2155365853659,65.8634634146342,61.7827317073171,7
0.8170731707317,68.5856097560976,60.836243902439,41.2360487804878,
...more on files)

Life_Expectancy_At_Birth_2013 <-
c(75.3286585365854,60.0282682926829,51.8661707317073,77.537243902439,
77.1956341463415,75.9860975609756,74.5613658536585,75.7786585365854,8
2.1975609756098,80.890243902439,70.6931463414634,56.2516097560976,
```

```
...more on files)
```

3. To filter the data of the countries by years (1960, 2013) a data frame must be created and the fields of the stats matrix and the vector of the corresponding year are joined. In the stats matrix, the data were selected where the year coincides with the year you want.

```
mydf2_1960 <- data.frame(stats[stats$Year==1960,],
Life_Expectancy_At_Birth_1960)
mydf2_1960


mydf1_2013 <- data.frame(stats[stats$Year==2013,],
Life_Expectancy_At_Birth_2013)
mydf1_2013
```

4. The "ggplot2" library is imported to create the scatter plots.

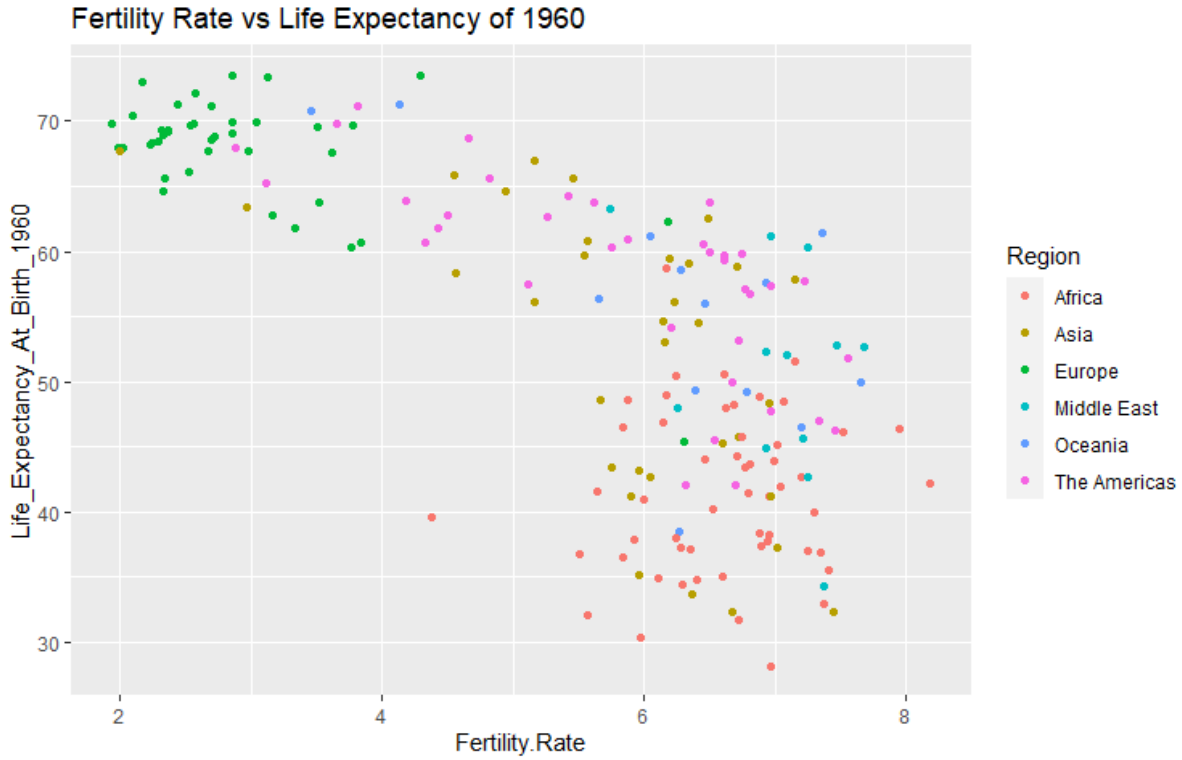
```
library("ggplot2")
```

5. To create the scatter diagram, the “qplot” method is used and the following parameters are sent (mydf2_1960, x = Fertility.Rate, y = LifeExpectancy).

```
#Diagrama de dispersión del 2013
qplot(data=mydf2_1960, x=Fertility.Rate,
y=Life_Expectancy_At_Birth_1960, color=Region,
     main="Fertility Rate vs Life Expectancy of 1960")

#Diagrama de dispersión de 2013
qplot(data=mydf1_2013, x=Fertility.Rate,
y=Life_Expectancy_At_Birth_2013, color=Region,
     main="Fertility Rate vs Life Expectancy of 2013")
```

## Graph 1: Fertility Rate vs Life Expectancy in 1960




## Graph 2: Fertility rate vs Life expectancy in 2013