14.310x: Data Analysis for Social Scientists
Describing Data, Joint and Conditional distributions of Random Variables

Welcome to your third homework assignment! You will have about one week to work through the assignment. We encourage you to get an early start, particularly if you still feel you need more experience using R. We encourage you to get an early start, particularly if you still feel you need more experience using R. This assignment has 2 parts: we have provided with PDF copies of Part 1 and Part 2 of the assignment so that you can print and work through the assignment offline. You can also go online directly to complete the assignment. If you choose to work on the assignment using these PDFs, please go back to the online platform to submit your answers based on the output produced.

Good luck ☺!

PART 1:

In this part of the problem set, we will guide you through different ways of accessing real data sets and how to summarize and describe them properly. First, we will go through some of the data that is collected by the World Bank. We will do some cleaning on the data before we start analyzing it.

Let's start with the data sets of the World Bank. Please complete the following steps:

1. Go to the World Bank Datasets website: http://data.worldbank.org/.

2. Once you are there, use the data catalogue to find the Gender Stats data and download the file in csv format.

3. Save the file in your computer in a folder where you can get it easily. In my case, I have saved the files in the following directory: *"/Users/raz/Dropbox/14.31 edX Building the Course/Problem Sets/PSET 3/Gender_Stats_csv"*.

NOTE: It is important to work in the same directory that the files are or to use the whole path when you specify opening a data set. To know in which directory you are currently working, you can use the command getwd(). Similarly, in order to set a different directory, you can use the command setwd().

Analyzing the Data:
For the purpose of analyzing the data, we are going to use the package "utils". Once you have uploaded the data to R, you are going to see that there are multiple indicators of gender, countries and years in the data. In this case, we are just interested in analyzing the data for one indicator that is the *Adolescent Fertility Rate.*

The *Adolescent Fertility Rate* measures the annual number of births to women 15 to 19 years of age per 1,000 women in that age group. It represents the risk of childbearing among adolescent

women 15 to 19 years of age. It is also referred to as the age-specific fertility rate for women aged 15-19. Once you have completed this problem set, you'll have more information on how this rate has evolved over time and how it varies across different groups of countries.
Take a look at the following lines of code, whose main purpose is to upload the data in a data frame and to choose the proper indicator. Please, try to understand the code and then run it in your computer. Remember to set the directory accordingly to the folder where you saved the files.

```
#Preliminaries
rm(list = ls())
library("utils")
setwd("/Users/raz/Dropbox/14.31 edX Building the Course/Problem
Sets/PSET 3/Gender_Stats_csv")

#Getting the data
gender_data <- read.csv("GenderStat_Data.csv")
teenager_fr <- subset(gender_data, Indicator.Code == "SP.ADO.TFRT")
```

1. What is the purpose of the line **rm(list = ls())?**
   a. To remove all the current existing objects in R
   b. To change the current directory path
   c. To list all the files in the current directory
   d. To look in the web for the World Bank dataset.

2. What part of the code specifies the creation of a data frame that contains only the relevant information for the analysis that we are interested in?
   a. `setwd("/Users/raz/Dropbox/14.31 edX Building the Course/Problem Sets/PSET 3/Gender_Stats_csv")`
   b. `gender_data <- read.csv("GenderStat_Data.csv")`
   c. `teenager_fr <- subset(gender_data, Indicator.Code == "SP.ADO.TFRT")`
   d. `library("utils")`

3. If you had run in R the provided code, you should have found that now you have two different data frames in R. One named gender_data with 180,601 observations and 62 variables. The other one named teenager_fr with 263 observations and 62 variables. This seems to be kind of inefficient since we won't use the first data frame and it is kind of big. If you were interested in removing this object, write down the command that would allow you to do this?

   Answer: ____________________

Now that you have uploaded the data to R that we are interested in analyzing, it is time to get our hands dirty! First, let's try to explore the data a bit more. The command str() will allow you to see the structure of an object in R. Type str(teenager_fr) to get a sense of the variables that we are currently using. Likewise, the command head() and tail(), will allow you to see the first six and last six observations of your data frame, respectively. Try to add this into your R code and explore the data set a bit more by yourself.

A second exploratory thing to do once we have organized a data set is to get basic summary statistics of the data. So, let's do this! One way to do this is to use the command mean(), which allows you to get the average of the variables in your data. For example, if you were interested in obtaining the sample mean of the Adolescent Fertility Rate in 1975, you can do this by running the following code.

```
mean(teenager_fr$X1975, na.rm = TRUE)
```

4. Why it is necessary to add the option "na.rm = TRUE" to the command? (Select all that apply)

    a. The default option of na.rm is set to FALSE. Therefore, if we don't specify this, R will try to calculate the mean using all the observations in the data.
    b. This part is necessary since otherwise R would duplicate some of the observations in the data set when it calculates the sample mean. In particular, the observations with missing values would have higher weights than the observations without missing values.
    c. It is not necessary to add this option to the command to obtain the mean of this variable.
    d. Otherwise we will obtain a missing value since not all the countries in the data have information on the adolescent fertility rate in 1975.
    e. This option is necessary since there are missing values in the data set. Thus, when R tries to calculate the mean it assumes that the result is not a number.

To calculate summary statistics for a group of variables, there are a few different commands. The command `mean()` is just one example of the different options available. Now, we ask you to go through the R documentation and explore some of the other commands by yourself.

This next set of questions you should be able to answer after you have taken sometime to explore the R documentation.

5. What is the sample mean and standard deviation of the Adolescent fertility rate in 1960?

    Sample Mean: __________   **Standard Deviation: _________________**

6. What is the sample mean and standard deviation of the Adolescent fertility rate in 2000?

    Sample Mean: ___________   **Standard Deviation**: ____________________

7. From the values that you have calculated above can you conclude that the Adolescent Fertility Rate has had a permanent decreasing trend over this period of time and that the dispersion of this variable has decreased?

a. Yes
b. No

Now, we are interested in plotting the evolution of the Adolescent Fertility Rate from 1960 to 2014. In addition, we are interested in having different information displayed in the same plot. First, we want to plot the sample mean of for the entire sample in the data set, but we also want to add more information such as the rate for low, middle and high income countries.

Someone has written the following R code to create a matrix with the information that is required for the plot. However, as you can see, we are missing some information that we need to be able to run it. Therefore, we need your help! We want to have a matrix whose first rows correspond to the years, and the next rows to the average rate for low income, middle, and high income countries, respectively. Here is the code and the missing pieces:

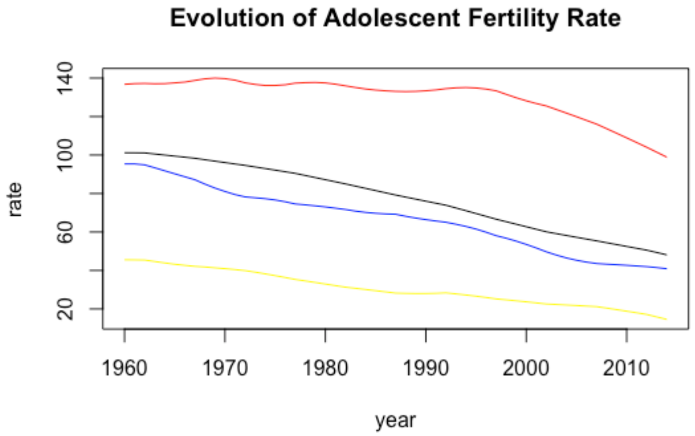
```
#Plotting the mean, and the value for low income, middle income, and
high income countries
low_income <- subset(teenager_fr, Country.Code == "□" )
middle_income <- subset(teenager_fr, Country.Code == "MIC")
high_income <- subset(teenager_fr, Country.Code == "HIC")

plot_frame <- matrix(NA, □, 55)
for (i in 5:59){
  k <- i - 4
  j <- i + □
  plot_frame[1, k] <- j
  plot_frame[2, k] <- mean(teenager_fr[, i], na.rm = □)
  plot_frame[3, k] <- low_income[, i]
  plot_frame[4, k] <- □ [, i]
  plot_frame[5, k] <- high_income[, i]
}
```

8. Which of the following options can be used to fill the blanks in order to achieve our purpose?

    a. LIC; 4; 1955; FALSE; middle_income
    b. OEDC; 3; 1956; TRUE; middle_income
    c. LIC; 5; 1955; FALSE; middle_income
    d. LIC; 5; 1955; TRUE; middle_income
    e. MIC; 5; 1956; TRUE; middle_income
    f. MIC; 4; 1955; TRUE; middle_income
    g. LIC; 5; 1955; TRUE; middle_income

9. Now that we have our matrix we can use it to construct our plot. Once we have run the code we have produced the following graph:



However, we have a problem! There is no legend in the figure, and we don't know which line corresponds to each category that we were trying to plot. Can you help us to figure out which of the following codes will produce this figure?

```
a. xlim <- range(c(plot_frame[1,]))
   ylim    <-    range(c(plot_frame[2,],    plot_frame[3,],
   plot_frame[4,], plot_frame[5,]))
   plot(plot_frame[1,],  plot_frame[2,],  type  =  "l",  col  =
   "black" xlim=xlim, ylim=ylim,
        main  =  "Evolution  of  Adolescent  Fertility  Rate",
   xlab = "year", ylab = "rate")
   lines(plot_frame[1,], plot_frame[3,], col = "red")
   lines(plot_frame[1,], plot_frame[4,], col = "yellow")
   lines(plot_frame[1,], plot_frame[5,], col = "blue")

b. xlim <- range(c(plot_frame[1,]))
   ylim    <-    range(c(plot_frame[2,],    plot_frame[3,],
   plot_frame[4,], plot_frame[5,]))
   plot(plot_frame[1,],  plot_frame[2,],  type  =  "l",  col  =
   "blue" xlim=xlim, ylim=ylim,
        main  =  "Evolution  of  Adolescent  Fertility  Rate",
   xlab = "year", ylab = "rate")
   lines(plot_frame[1,], plot_frame[3,], col = "yellow")
   lines(plot_frame[1,], plot_frame[4,], col = "black")
   lines(plot_frame[1,], plot_frame[5,], col = "red")

c. xlim <- range(c(plot_frame[1,]))
   ylim    <-    range(c(plot_frame[2,],    plot_frame[3,],
   plot_frame[4,], plot_frame[5,]))
   plot(plot_frame[1,],  plot_frame[2,],  type  =  "l",  col  =
   "yellow" xlim=xlim, ylim=ylim,
        main = "Evolution of Adolescent Fertility Rate",
   xlab = "year", ylab = "rate")
   lines(plot_frame[1,], plot_frame[3,], col = "blue")
   lines(plot_frame[1,], plot_frame[4,], col = "red")
   lines(plot_frame[1,], plot_frame[5,], col = "black")
```

```
d. xlim <- range(c(plot_frame[1,]))
   ylim <- range(c(plot_frame[2,], plot_frame[3,],
   plot_frame[4,], plot_frame[5,]))
   plot(plot_frame[1,], plot_frame[2,], type = "l", col =
   "black" xlim=xlim, ylim=ylim,
        main = "Evolution of Adolescent Fertility Rate",
   xlab = "year", ylab = "rate")
   lines(plot_frame[1,], plot_frame[3,], col = "red")
   lines(plot_frame[1,], plot_frame[4,], col = "blue")
   lines(plot_frame[1,], plot_frame[5,], col = "yellow")
```

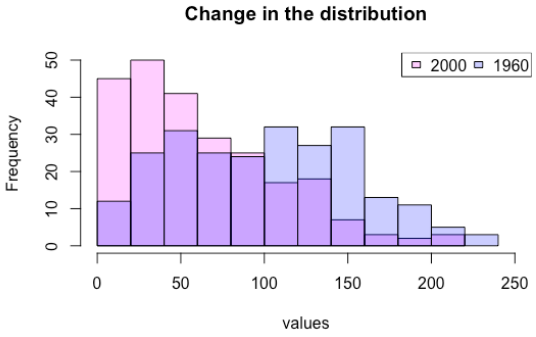
10. Which of the following statements can you conclude from the plot?

    a. Using all the data, the average of the rate is always below the rate of high and middle income countries, and below the one for low income countries.
    b. While the rate for high income countries has presented a decreasing trend for the entire period, the rate for low income countries is barely steady until the mid-nineties. From there onwards, the rate has decreased drastically.
    c. The gap between high and middle income countries is lower in 2014 than in 1960, while the gap between low and middle income countries is actually larger.
    d. Since the mid-nineties the rate for low income countries has decreased more than the rate for high and middle income countries.

Now, we are going to focus on how the distribution of the Adolescent Fertility Rate has changed from 1960 to 2000. The following code in R plots the histogram of these two variables in the same graph. Please take a look at the code and try to understand what it is doing.

```
p1 <- hist(teenager_fr$X1960)
p2 <- hist(teenager_fr$X2000)
plot( p2, col=rgb(1,0,1,1/4), xlim = c(0, 250), main = "Change
in the distribution", xlab = "values")
plot( p1, col=rgb(0,0,1,1/4), xlim = c(0, 250), add = TRUE)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
png("histogram")
```

Here is the figure that this code has produced:



11. The color of the bins was chosen using the option `rgb(0,0,1,1/4)`? What does the fourth argument in this vector represent in the plot?
    a. The red level in the color of the bin.
    b. The green level in the color of the bin.
    c. The blue level in the color of the bin.
    d. The level of transparency in the color of bin.

12. As you can see we have certain number of bins in the figure, go to the R documentation and look for the option in the command hist, that will allow you to change the number of bins in the figure.

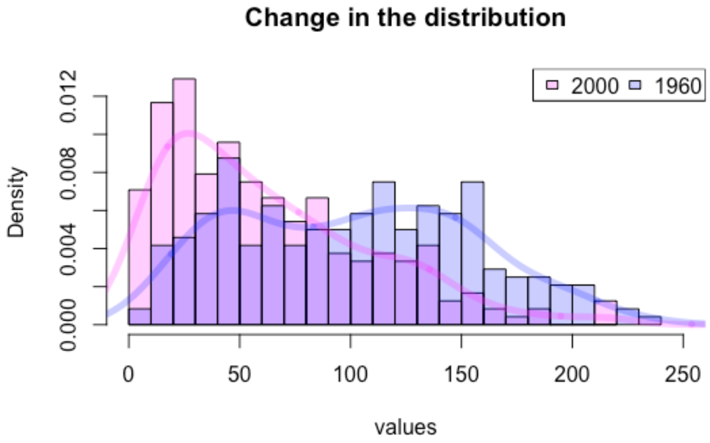
    Answer: __________________

Now, we are going to add some kernels to the histogram. The kernels were done using the command density and all the default options in R. Again, take a look at the code, run it on your computer and try to understand what it is doing.

```
p1 <- hist(teenager_fr$X1960, freq = FALSE, breaks = 20)
p2 <- hist(teenager_fr$X2000, freq = FALSE, breaks = 20)
p1$counts = p1$density
p2$counts = p2$density
p3 <- density(teenager_fr$X1960, na.rm = TRUE)
p4 <- density(teenager_fr$X2000, na.rm = TRUE)

plot( p2, col=rgb(1,0,1,1/4), xlim = c(0, 250), main = "Change in the
distribution", xlab = "values", ylab = "Density")
plot( p1, col=rgb(0,0,1,1/4), xlim = c(0, 250), add = TRUE)
lines( p4, col=rgb(1,0,1,1/4), xlim = c(0, 250), lwd = 5)
lines(p3, col=rgb(0,0,1,1/4), xlim = c(0, 250), lwd = 5)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
legend("topright", ncol = 2, legend = c("2000", "1960"),
fill=c(rgb(1,0,1,1/4), rgb(0,0,1,1/4)), text.width = 20)
```

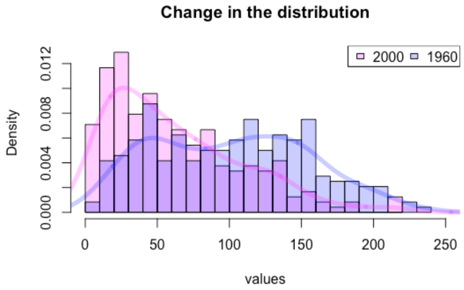
The figure that is produced by running this code is presented next:



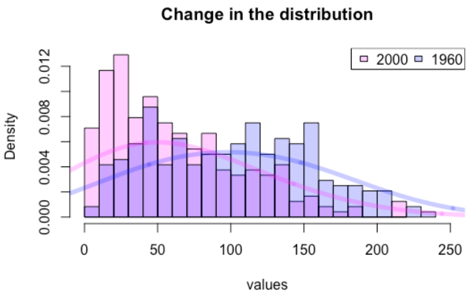
13. As it was stated before, the plot was done using the default options in R. For the kernel, the default option is to use gaussian. There are other options that the user can state when running the *density* command in R. Of the following list, which one doesn't have a bell-shaped? In other words, which one doesn't weight less the observations in the extremes of the bandwidth?
    a. gaussian
    b. epanechnikov
    c. rectangular
    d. triangular
    e. biweight
    f. cosine
    g. optcosine

14. The following plots were done changing the bandwidth of the kernel function in R. Which one of them was done with the largest bandwidth?

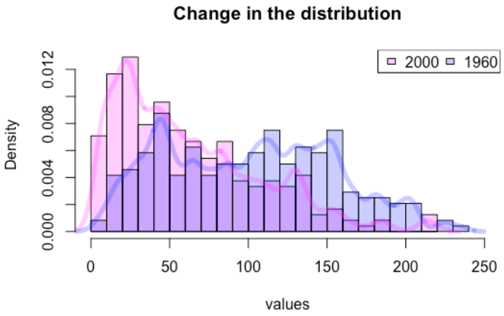
    a. It is not possible to tell just by looking at the figure.
    b.



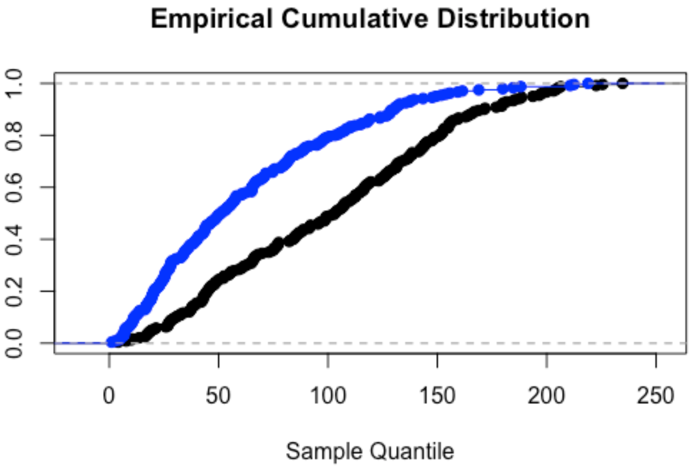
    c.



d.



One of the things that Professor Duflo also discussed in the lecture, was the construction of the Empirical Cumulative Distribution (ECD). The following figures shows the ECD for the Adolescent Fertility Rate in the World in 1960 and in 2000. However, as you can see the person who made the graph forgot to properly labeled it.



15. Can you infer from the histograms that were plotted before, which one corresponds to the Adolescent Fertility Rate in 2000 and which one to the same indicator in 1960. (Select all that apply)

   a. Blue corresponds to 2000
   b. Black corresponds to 2000

c. Blue corresponds to 1960
d. Black corresponds to 1960
e. It is not possible to tell from the plot

16. Can you infer from the figure, whether the distribution used to construct the black series satisfy the First Order Stochastic Dominance property over the distribution used to construct the blue series?

    a. Yes
    b. No