Do Women Promote Different Policies than Men? Part IV: Visualizations and Correlations (with Solutions)

Let's continue working with the data from the experiment in India. As a reminder, Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is villages.

variable	description
village	village identifier ("Gram Panchayat number _ village number")
female	whether village was assigned a female politician: 1=yes, 0=no
water	number of new (or repaired) drinking water facilities in the village since random assignment
irrigation	number of new (or repaired) irrigation facilities in the village since random assignment

Table 1: Variables in "india.csv"

In this problem set, we practice (1) how to create and make sense of visualizations and (2) how to compute and interpret the correlation between two variables, among other things.

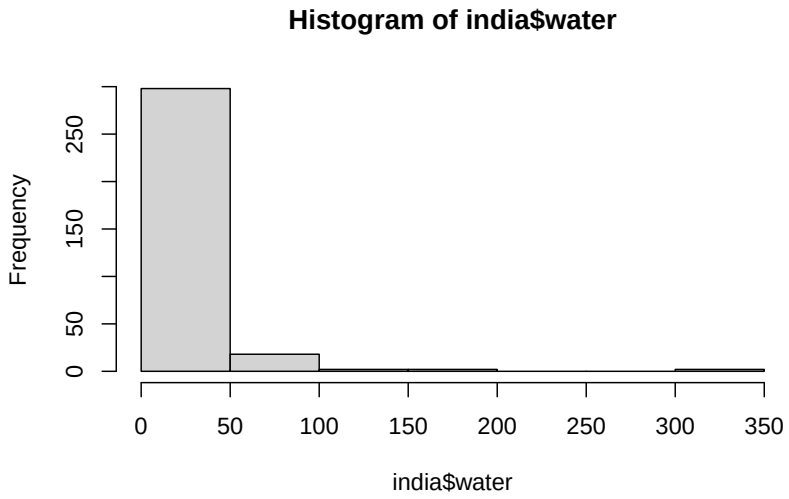
As always, we start by loading and looking at the data:

```
## load and look at the data
india <- read.csv("india.csv") # reads and stores data
head(india) # shows first observations
##          village female water irrigation
## 1 GP1_village2      1    10          0
## 2 GP1_village1      1     0          5
## 3 GP2_village2      1     2          2
## 4 GP2_village1      1    31          4
## 5 GP3_village2      0     0          0
## 6 GP3_village1      0     0          0
```

1. Create a visualization of the distribution of the variable *water*.
    a. Does this variable look normally distributed? (5 points)
    b. Approximately how many villages in this experiment had about 250 new (or repaired) drinking water facilities since the randomization of politicians? (5 points)

    R code:

    ```
    ## create a histogram to visualize the distribution of water
    hist(india$water)
    ```



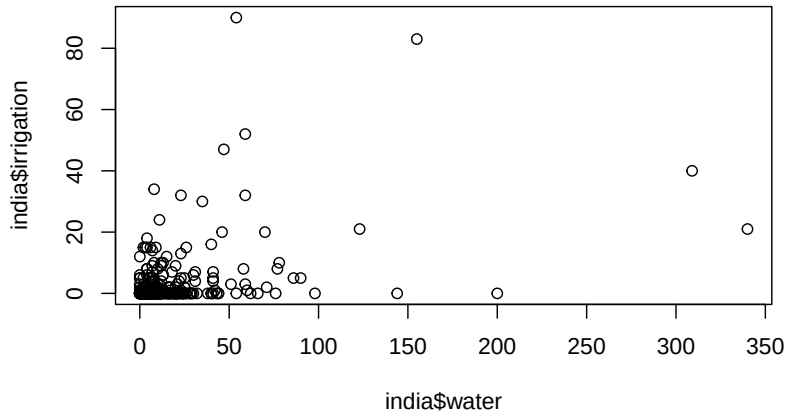
(Recall: the histogram of a variable is the visual representation of its distribution. The function in R to create a histogram is hist(). The only required argument is the code identifying the variable.)

Answers: (a) No, the variable doesn't look normally distributed. (Note: The distribution is not symmetric.) (b) Fewer than 10 villages in this experiment had about 250 new (or repaired) drinking water facilities.

2. Create a visualization of the relationship between *water* and *irrigation*.
   a. Does the linear relationship between these two variables look positive or negative? A positive/negative answer will suffice. (5 points)
   b. Does the relationship between these two variables look strongly linear? A yes/no answer will suffice. (5 points)

   R code:

```
## create a scatter plot to visualize the relationship
plot(india$water, india$irrigation)
```



(Recall: a scatter plot is the graphical representation of the relationship between two variables. The function in R to create a scatter plot is plot(). It requires two arguments (separated by a comma) and in this particular order: (1) the code identifying the variable to be plotted along the x-axis, and (2) the the code identifying the variable to be plotted along the y-axis.)

Answers: (a) The linear relationship looks positive. (Note: the slope of the line of best fit is positive.) (b) No, the relationship does not look strongly linear. (Note: the observations are pretty far away from the line of best fit.)

3. Compute the correlation between *water* and *irrigation*.
   a. Are you surprised by the sign of the correlation? Provide your reason. (5 points)
   b. And are you surprised by the absolute value of the correlation? Provide your reason. (5 points)

   R code:

```
##  compute the correlation
cor( india$water,  india$ irrigation )
##  [1]  0.4073307
```

   (Recall: the function in R to compute a correlation coefficient is cor(). It requires two arguments (separated by a comma) and in no particular order: the code identifying each of the two variables.)

   Answers: (a) No, I am not surprised that the correlation is positive because, in the scatter plot above, I observed that the line that best summarizes the relationship between these two variables has a positive slope. (b) No, I am not surprised that the absolute value of the correlation coefficient is 0.4 because, in the scatter plot above, I observed taht the relationship between *water* and *irrigation* does not look strongly linear. (Note: the observations are pretty far away from the line of best fit.)

4. If we wanted to use the sample of villages in this dataset to infer the characteristics of all villages in India, we would have to make sure that the sample is _____ of the population. (Please provide the missing word). (10 points)

   Answer: representative.

5. What would have been the best way of selecting the villages for the sample to ensure that the statement above was true? (10 points)

   Answer: The best way to make the sample of villages representative of all villages in India would have been to select the villages through random sampling.