Do Women Promote Different Policies than Men?
Part I: Loading and Making Sense of Data (with Solutions)

(Based on Raghabendra Chattopadhyay and Esther Duflo. 2004. “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica*, 72 (5): 1409–43.)

In a few problem sets, we will estimate the average causal effect of having a female politician on two different policy outcomes. For this purpose, we will analyze data from an experiment conducted in India, where villages were randomly assigned to have a female council head. The dataset we will use is in a file called “india.csv”. 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 how to load and make sense of data.

1. Use the function read.csv() to read the CSV file “india.csv” and use the assignment operator <- to store the data in an object called *india*. (Do not forget to set the working directory first.) Provide the R code you used (without the output). (10 points)

   R code:

   ```
   india <- read.csv("india.csv") # reads and stores data
   ```

   (Recall: to the left of the assignment operator <-, we specify the name of the object, india in this case; to the right of the assignment operator <-, we specify the contents, which, in this case, are produced by reading the CSV file “india.csv”. Also, we do not use quotes around the name of an object such as india or around the name of a function such as read.csv(), but we do use quotes around the name of a file: "india.csv".)

2. Use the function head() to view the first few observations of the dataset. Provide the R code you used (without the output). (10 points)

R code:

```
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
```

3. What does each observation in this dataset represent? (5 points)

   Answer: Each observation represents a village (Note: We know this because, as stated above, the unit of observation in this dataset is villages).

4. Please substantively interpret the first observation in the dataset. (5 points)

   Answer: The first observation in the dataset represents village 2 in Gram Panchayat 1, which was assigned a female politician, and it had 10 new or repaired drinking water facilities and 0 new or repaired irrigation facilities since politicians were randomly assigned. (Note: the first observation consists of the following values: *village*="GP1_village2", *female*=1, *water*=10, *irrigation*=0; we can interpret each value by using the description of the variables in Table 1.)

5. For each variable in the dataset, please identify the type of variable (character vs. numeric binary vs. numeric non-binary) (10 points)

   Answer: *village* is a character variable, *female* is numeric binary, *water* is numeric non-binary, and *irrigation* is numeric non-binary. (Recall: binary variables can only take two values, 0s and 1s, and non-binary variables can take more than two values.)

6. How many observations are in the dataset? In other words, how many villages were part of this experiment? (Hint: the function dim() might be helpful here.) Provide the R code you used (without the output) and provide the substantive answer. (10 points)

   R code:

   ```
   dim(india) # provides dimensions of dataframe: rows, columns
   ## [1] 322   4
   ```

   Answer: There were 322 villages in this experiment. (Recall: the first number provided by dim() corresponds to the number of observations in the dataframe, the second number corresponds to the number of variables. Based on the output of dim(), there are 322 observations in the dataframe *india*. Since the unit of observation is villages, the dataframe *india* has data on 322 villages in India.)