This material was produced for instructors using Llaudet, Elena and Kosuke Imai.
Data Analysis for Social Science: A Friendly and Practical Introduction. (Princeton University Press)
and should not be shared beyond those who are enrolled in this class.

Does Having a Black Candidate Running Increase Black Turnout?

(Based on Bernard Fraga. 2016. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout." *American Journal of Political Science*, 60: 97-122.)

Some scholars have suggested that having a black candidate running increases black turnout. In this problem set, we explore the causal relationship between black candidates and black turnout using observational data from U.S. elections.

The dataset is in a file called "districts.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is district elections.

variable	description
year	year of the election
state	state where the district is located
district	district number, which is unique within states but not across states
proportion_black	proportion of the district's voting-age population that was black at the time (in percentages)
black_candidate	whether there was a black candidate running at the district-level election in that year: 1=there was a black candidate, 0=there was not
black_turnout	proportion of the district's black voting age population that voted in that year's election (in percentages)

Table 1: Variables in "districts.csv"

In this problem set, we practice fitting a linear model to compute the difference-in-means estimator and fitting a multiple linear regression model to statistically control for confounders.

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

```
## load and look at the data
districts <- read.csv("districts.csv") # reads and stores data
head(districts, n=8) # shows first eight observations
##   year state district proportion_black black_candidate black_turnout
## 1 2006    AK        1         3.178487               0       43.94295
## 2 2006    AL        1        26.122134               0       26.60873
## 3 2006    AL        2        29.478384               0       26.66497
## 4 2006    AL        3        30.843040               0       28.46918
## 5 2006    AL        4         4.996171               0       27.97092
## 6 2006    AL        5        16.597923               0       29.18517
## 7 2006    AL        6        10.682076               0       28.32708
## 8 2006    AL        7        61.952636               1       32.90645
```

In the code above, we asked R to show the first eight observations, instead of the default of six, by using the optional argument `n` in the function `head()`.

This material was produced for instructors using Llaudet, Elena and Kosuke Imai.
Data Analysis for Social Science: A Friendly and Practical Introduction. (Princeton University Press)
and should not be shared beyond those who are enrolled in this class.

1. First, let's make sure we understand the data and identify our X and Y variables.

    a. Substantively interpret the seventh and eight observations in the dataset and do not forget to include the unit of measurements. (2.5 points)

    b. Given that we are interested in estimating the average causal effect of having a black candidate running on black turnout: What should be our Y variable? In other words, which variable is the outcome variable? And, is this variable binary or non-binary? (2.5 points)

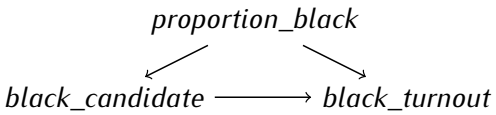
    c. Given that we are interested in estimating the average causal effect of having a black candidate running on black turnout: What should be our X variable? In other words, which variable is the treatment variable? And, is this variable binary or non-binary? (2.5 points)

2. For now, let's assume that the data we are analyzing came from a randomized experiment, where researchers were able to randomly assign black candidates to district elections. If this were true, then, we could estimate the average causal effect of having a black candidate running on black turnout by computing the difference-in-means estimator.

    a. Let's start by computing the average outcome for the treatment group, that is, the average black turnout in district elections with a black candidate running. Provide an interpretation of the result using a full sentences, and do not forget to include the unit of measurement. (2.5 points)
    b. Now, compute the average outcome for the control group, that is, the average black turnout in district elections without a black candidate running. Provide an interpretation of the result using a full sentences, and do not forget to include the unit of measurement. (2.5 points)
    c. Compute the difference-in-means estimator directly and report its value. (2.5 points)
    d. Now, let's use the `lm()` function to fit a line to the data in such a way that the $\widehat{\beta}$ coefficient will be equivalent to the difference-in-means estimator. What is the fitted line? In other words, provide the formula $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute $\widehat{\alpha}$ for the estimated value of the intercept coefficient, substitute $\widehat{\beta}$ for the estimated value of the slope coefficient, and substitute X for the name of the treatment variable.) (Hint: The `lm()` function requires an argument of the form `Y~X`) (5 points)
    e. Is the estimated slope coefficient ($\widehat{\beta}$) equivalent to the value of the difference-in-means estimator in this case? A yes/no answer will suffice. (5 points)
    f. If the data came from a randomized experiment, what would you conclude is the direction, size, and unit of measurement of the estimated average causal effect of having a black candidate running on black turnout? (5 points)

3. Since the data is observational, which means it did not come from a randomized experiment, there are many ways in which district elections with black candidates running might be different from district elections without black candidates running. In other words, there might be potential confounding variables present obscuring the causal relationship between *black_candidates* and *black_turnout*. A potential confounder is, for example, the proportion of the district's voting-age population that was black at the time of the election: *proportion_black*. Here is the reasoning: It is likely the case that districts with a higher proportion of black voting-age population are more likely to have black candidates running in the election ($Z \rightarrow X$). It is also likely the case that districts with a higher proportion of black voting-age population are more likely to have a higher black turnout in the elections ($Z \rightarrow Y$).

This material was produced for instructors using Llaudet, Elena and Kosuke Imai.
Data Analysis for Social Science: *A Friendly and Practical Introduction*. (Princeton University Press)
and should not be shared beyond those who are enrolled in this class.



a. To further explore the possibility that *proportion_black* is a confounder, compute the correlation between *proportion_black* and *black_candidate*. Are these two variables moderately to highly correlated with each other? A yes/no answer will suffice. (5 points)
b. Now, statistically control for *proportion_black* by running a multiple linear regression model and estimate the average causal effect of having a black candidate running on black turnout while keeping the proportion of black voting-age population in the district constant. Report the new direction, size, and unit of measurement of the estimated average causal effect of having a black candidate running on black turnout. (10 points)
c. Given this last analysis, would you conclude that having a black candidate running increases black turnout? A yes/no answer will suffice. (5 points)