Homework Assignment 1

DPI 610

Assignment Details

This homework assignment is due *February 20, 2026* before midnight. Everyone must complete their own code and assignment, though you may discuss the homework with classmates.

To access the code and data for the assignment, go to Posit.cloud or Canvas. To turn in the assignment, upload two files to Canvas: (1) your code file (.qmd), and (2) a PDF version (click Render to create PDF).

Technical Notes:

- Part 1 uses the North Carolina voter file data (`nc`) to explore voter file structure and basic data manipulation
- Part 2 uses the Florida experiment data (`fl`) to introduce randomized experiments and treatment effects

Setup

The following code loads the dataset and packages necessary to complete the problem set. Please make sure to run the following code before starting the problem set.

Part 1: Exploring the NC Voter File

Voter files are essential tools for political campaigns. They contain information about registered voters including demographics, voting history, and sometimes modeled predictions about voter behavior. In this section, you will explore the structure and content of a North Carolina voter file.

Question 1(a)

Use the `glimpse()` function to examine the structure of the `nc` dataset. How many observations (voters) are in the dataset? How many variables are available?

Answer 1(a)

```
glimpse(nc)
```

```
Rows: 10,402
Columns: 25
$ id                <int> 36974, 37362, 36635, 38352, 36619, 37731, 38205, 380~
$ state             <chr> "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC", "NC"~
$ age               <int> 30, 58, 65, 65, 52, 40, 55, 64, 72, 65, 47, 87, 74, ~
$ gender            <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F~
$ party             <chr> "DEM", "DEM", "DEM", "DEM", "DEM", "DEM", "DEM", "DE~
$ race              <chr> "C", "C", "C", "C", "C", "C", "C", "C", "C", "C", "C~
$ socioecon_bg      <chr> "VERY WEALTHY", "VERY WEALTHY", "VERY WEALTHY", "VER~
$ familyincome      <dbl> 135000, 83000, 73000, 176000, 138000, 137000, 100000~
$ density           <chr> "URBAN", "URBAN", "URBAN", "URBAN", "RURAL", "URBAN"~
$ ever_donor        <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1~
$ voted2008         <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1~
$ voted2010         <int> 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1~
$ voted2012         <int> 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ voted2014         <int> 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1~
$ vote2012_prob     <dbl> 49.5, 98.8, 93.7, 97.3, 91.0, 98.1, 80.5, 64.8, 46.2~
$ vote2014_prob     <dbl> 16.63, 84.80, 87.73, 34.64, 78.03, 84.39, 57.60, 64.~
$ ideology_score    <dbl> 80.1, 31.1, 38.8, 63.5, 72.5, 69.2, 67.1, 56.5, 62.2~
$ dem_score         <dbl> 70.65, 75.18, 69.74, 73.75, 94.15, 69.74, 83.76, 44.~
$ dem2012_score     <dbl> 92.67, 70.16, 85.89, 97.60, 96.62, 88.55, 92.80, 88.~
$ homeowner_score   <int> 9, 9, 7, 9, 9, 9, 9, 9, 9, 5, 9, 9, 9, 9, 9, 8, 9, 9~
$ college_grad_prob <dbl> 0.857, 0.697, 0.451, 0.650, 0.840, 0.730, 0.496, 0.6~
$ gun_owner_prob    <dbl> 0.140, 0.168, 0.206, 0.267, 0.176, 0.249, 0.160, 0.2~
```

```
$ hunter_prob       <dbl> NA, 0.046, 0.046, 0.231, 0.046, 0.046, 0.205, 0.046,~
$ married_prob      <dbl> 0.596, 0.173, 0.760, 0.726, 0.272, 0.957, 0.376, 0.6~
$ religion          <chr> "UNCODED CHRISTIAN", "CATHOLIC", "UNCODED", "PROTEST~
```

```r
nrow <- nc %>%
  nrow()
```

- There are 10402 observations (voters) in the dataset.
- There are 25 variables available.

Question 1(b)

Examine the variables related to past voting behavior: `voted2008`, `voted2010`, `voted2012`, and `voted2014`. What proportion of voters in the dataset voted in each of these elections? Create a summary table showing the turnout rate for each year. Which election had the highest turnout? Does this pattern match what you would expect based on the types of elections held in these years?

Answer 1(b)

```r
nc %>%
  summarize(
    turnout_2008 = mean(voted2008, na.rm = TRUE),
    turnout_2010 = mean(voted2010, na.rm = TRUE),
    turnout_2012 = mean(voted2012, na.rm = TRUE),
    turnout_2014 = mean(voted2014, na.rm = TRUE)
  ) %>%
  pivot_longer(
    cols = everything(),
    names_to = "Election Year",
    values_to = "Turnout Rate"
  )
```

```
# A tibble: 4 x 2
  `Election Year` `Turnout Rate`
  <chr>                    <dbl>
1 turnout_2008             0.769
2 turnout_2010             0.550
3 turnout_2012             0.879
4 turnout_2014             0.627
```

- The 2012 election had the highest turnout, followed by 2008. This is what I'd expect since these are presidential election years, which typically have higher turnout than midterm elections like 2010 and 2014.

Question 1(c)

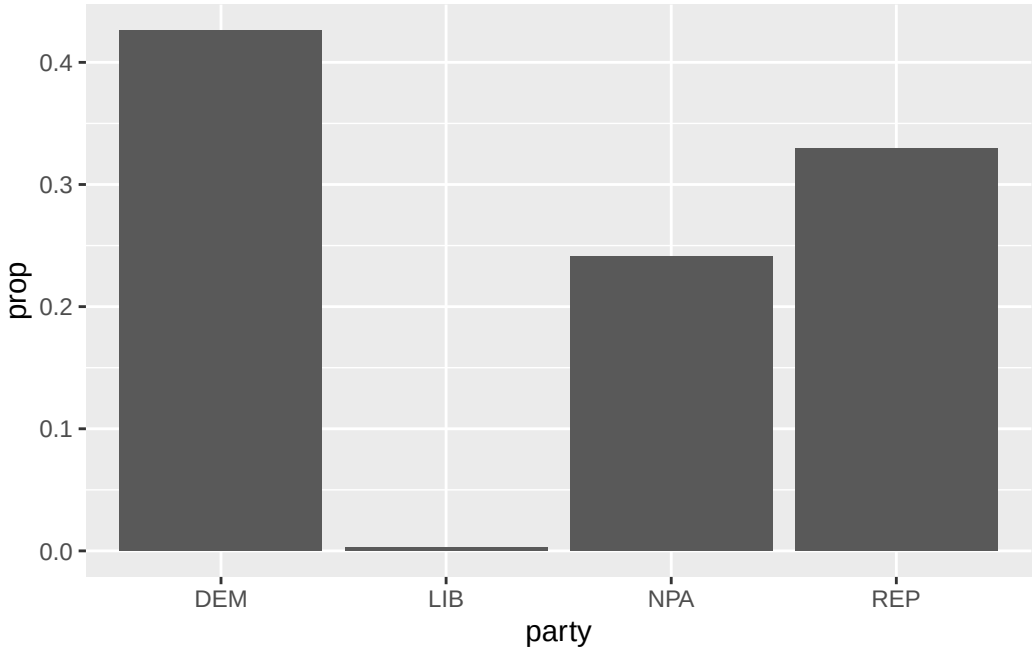
Examine the distribution of party registration in the `nc` data using the `party` variable. What proportion of voters are registered as Democrats (DEM), Republicans (REP), No Party Affiliation (NPA), and Libertarian (LIB)? Create a bar chart showing the distribution of party registration.

Answer 1(c)

```
# summary table of voter registration party proportions
nc %>%
  group_by(party) %>% # create groups of each party (DEM, LIB, NPA, REP)
  summarize(n = n()) %>% # count numb of obs in each group
  mutate(prop = round(n / sum(n), 4)) # find prop
```

```
# A tibble: 4 x 3
  party     n  prop
  <chr> <int> <dbl>
1 DEM    4431 0.426
2 LIB      31 0.003
3 NPA    2511 0.241
4 REP    3429 0.330
```

```
# create a bar graph of the distribution of party registration
nc %>%
  group_by(party) %>%
  summarize(n = n()) %>%
  mutate(prop = round(n / sum(n), 4)) %>%
  ggplot(aes(x = party, y = prop)) +
  geom_col()
```



Question 1(d)

Create a new variable called `consistent_voter` that equals 1 if a person voted in both 2008 and 2012 (both presidential elections), and 0 otherwise. What proportion of voters in the dataset are consistent voters by this definition? How does this proportion differ by party registration?

Answer 1(d)

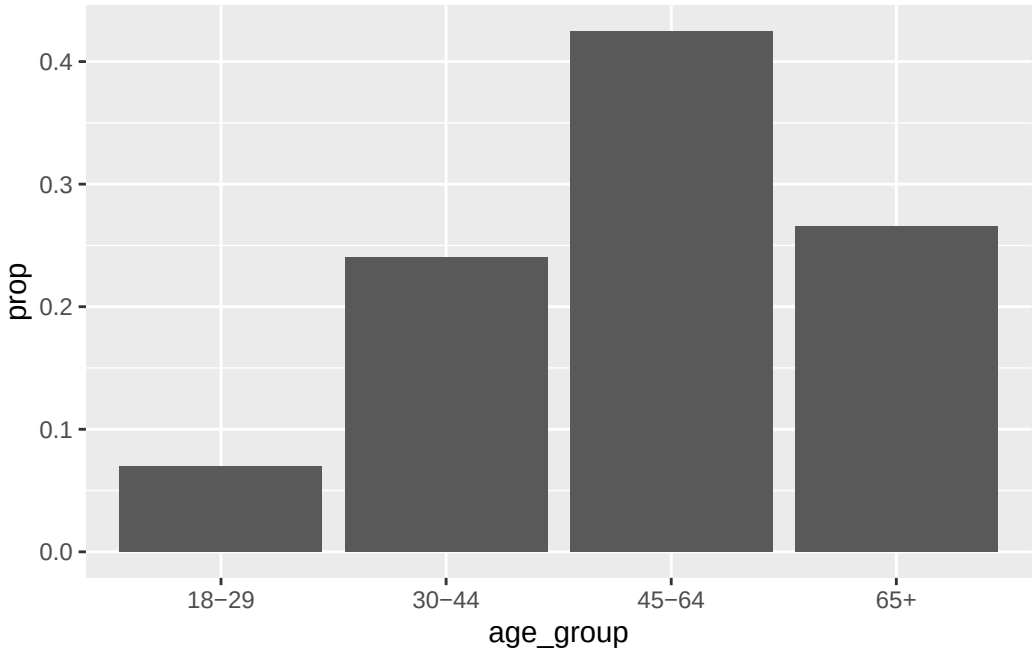
```
# create new var if voter voted in 2008 AND 2012
nc <- nc %>%
  mutate(consistent_voter = if_else(voted2008 + voted2012 == 2, 1, 0)) # %>%
  # select(id, voted2008, voted2012, consistent_voter)

#' since voter is a binary stored as 0 (no vote) or 1 (vote)
#' we can take the sum of the vars. voted2008 and voted2012
#' to determine if someone voted in both elections,
#' if sum is 2 they voted twice, 1 only in either election,
#' and if it's 0 then they didn't vote at all.
```

Question 1(e)

Examine the relationship between age and voter turnout in 2014. Group voters into the following age categories: 18-29, 30-44, 45-64, and 65+. Calculate the turnout rate in 2014 for each age group. Create a bar chart showing these turnout rates. What pattern do you observe?

Answer 1(e)

```
nc %>%
  mutate(age_group = case_when(age >= 18 & age <= 29 ~ "18-29",
                               age >= 30 & age <= 44 ~ "30-44",
                               age >= 45 & age <= 64 ~ "45-64",
                               age >= 65             ~ "65+")) %>%
  group_by(age_group) %>%
  summarize(n = n()) %>%
  mutate(prop = round(n / sum(n), 4)) %>%
  ggplot(aes(x = age_group, y = prop)) +
  geom_col()
```



- The bar graph suggests that as people get older they they vote more until retirement age (65+).

Part 2: Introduction to Randomized Experiments

You are employed at a political consulting firm working on several political races in Florida in 2020. Your boss has assigned you to look back over data gathered from an experiment conducted by your firm preceding the 2014 general election that (for some strange reason) was never analyzed (the data set is named `fl` and pre-loaded).

In the experiment, the firm mailed postcards to a randomly selected set of registered voters, showing the recipient their past turnout history and comparing it to the level of participation of the typical person in their state.[1]

Your task is to determine what lessons, if any, can be drawn from this experiment and applied to future efforts at boosting turnout among target voters.

Question 2(a)

The 2014 American Community Survey (ACS) estimates that the demographic breakdown (in terms of Race) in Florida is the following:

```
race_fl <- c("White", "Black", "Hispanic", "Other")
pct_race_fl <- c("56.6", "15.4", "23.3", "4.7")

fl_acs <- tibble(race = race_fl,
                 pct_pop = as.numeric(pct_race_fl)
 )

fl_acs
```

```
# A tibble: 4 x 2
  race     pct_pop
  <chr>      <dbl>
1 White       56.6
2 Black       15.4
3 Hispanic    23.3
4 Other        4.7
```

How does the demographic breakdown of the sample of voters in our `fl` data compare to the ACS estimates in terms of race? Use the `race` variable to compute the proportion of racial categories. Briefly interpret the result.

[1] The data comes from the replication archive of the paper "The Generalizability of Social Pressure Effects on Turnout Across High-Salience Electoral Contexts: Field Experimental Evidence From 1.96 Million Citizens in 17 States" by Alan Gerber, Greg Huber, Albert Fang, and Andrew Gooch.

Answer 2(a)

```
fl %>%
  group_by(race) %>%
  summarize(n = n()) %>%
  mutate(pct_vote = (round(n / sum(n), 3))*100,
         race = recode_values(race, "B" ~ "Black", "W" ~ "White", "H" ~ "Hispanic", "O" ~
  select(!n) %>%
  left_join(fl_acs) %>%
  mutate(diff = pct_pop - pct_vote)
```

```
# A tibble: 4 x 4
  race     pct_vote pct_pop  diff
  <chr>       <dbl>   <dbl> <dbl>
1 Black        27.6    15.4 -12.2
2 Hispanic     32.4    23.3  -9.1
3 Other         5.5     4.7  -0.8
4 White        34.4    56.6  22.2
```

- Based on the proportion differentials, our sample under represented Whites by 22% and over represented Blacks by 12% and Hispanics by 9%.

Question 2(b)

In terms of gender, the 2014 ACS estimates were that 51.1 percent of the population in Florida is female. How does our sample compare?

Answer 2(b)

```
#Insert Answer Here

fl %>%
  summarize(round(mean(female), 3)*100)
```

```
  round(mean(female), 3) * 100
1                         67.7
```

Question 2(c)

In terms of turnout, the breakdown between 2006 and 2014 in Florida[2] was as follows:

Year	Pct
2006	47
2008	75
2010	49
2012	72
2014	51

How does our sample compare?

Answer 2(c)

```
#Insert Answer Here
```

Question 2(d)

One reason why randomized experiments allow us to estimate the causal impact of an intervention is because, through randomization, the observable and unobservable characteristics of the subjects in the experiment are independent of, or at least uncorrelated with, treatment assignment. One consequence of this is that covariates will be "balanced" between the treatment and control group. You can check for balance by calculating the mean of a variable for the treatment group and for the control group.

Is the turnout history of subjects in the experiment between 2006 and 2012 balanced between treatment and control?

Answer 2(d)

```
#Insert Answer Here
```

Question 2(e)

Are gender, marriage, age and race balanced between treatment and control?

[2] Available at https://dos.myflorida.com/elections/data-statistics/elections-data/voter-turnout

Answer 2(e)

```
#Insert Answer Here
```

Question 2(f)

What is the overall 2014 turnout rate for people in the sample? What is the 2014 turnout rate for people in the treatment group? In the control group? What is the estimate of the average treatment effect (ATE) from the experiment?

Answer 2(f)

```
#Insert answer here
```

Question 2(g)

Based on your answers to Questions 2(d), 2(e), and 2(f), what can you conclude about the effectiveness of the social pressure mailing intervention? Was the experiment properly randomized? Did the treatment have an effect on voter turnout?

Answer 2(g)

Insert answer here.