Final – Intro to Probability and Statistics

Gustavo Baroni - 353897

Instructions

1. **Due:**: 06/24 at 11:59PM.
2. Send both the `.Rmd` and the compiled `.pdf`. Make sure that you use the right places on e-Class.
3. No collaboration! The work should be yours.
4. The turn-it-in is going to be on.
5. If you use Stackoverflow, or any other Google resources, please make sure to cite it. Provide the link, so we can check.
6. Please double check if you picked the right question group.
7. Grading parameters:
   - `.Rmd:` have to compile to have positive grade.
   - If not compiling a `.pdf`, then 50% penalty in the final grade.
   - `.pdf` has to match with the `.Rmd` you submit.
   - Coding: has to be efficient and clean
   - Answer: has to be precise and meaningful
   - Make sure that you revise your answers before submit the problem set.

Question assignment

You should find your number and check which question groups you need to answer.

Class	Student	Question Groups	Class	Student	Question Groups	Class	Student	Question Groups
A	347647	5, 7, 11, 13	B	348494	5, 7, 12, 14	C	339473	3, 10, 12, 16
A	353063	5, 6, 12, 14	B	352887	4, 8, 11, 14	C	339549	3, 8, 12, 14
A	353066	5, 7, 12, 16	B	353071	2, 9, 12, 14	C	352757	4, 8, 11, 16
A	353070	2, 7, 12, 16	B	353098	5, 9, 12, 15	C	352759	3, 9, 12, 15
A	353087	2, 10, 12, 15	B	353119	5, 6, 11, 15	C	352888	3, 8, 11, 15
A	353096	4, 6, 12, 15	B	353122	2, 10, 12, 15	C	353080	3, 6, 12, 14
A	353101	3, 9, 11, 16	B	353124	5, 10, 11, 14	C	353088	4, 9, 12, 16
A	353105	1, 9, 11, 16	B	353541	5, 6, 11, 15	C	353094	2, 10, 11, 16
A	353106	5, 9, 11, 13	B	353553	2, 9, 11, 15	C	353104	2, 7, 11, 13
A	353108	4, 8, 12, 13	B	353668	5, 10, 11, 15	C	353111	3, 7, 11, 15
A	353117	4, 8, 12, 14	B	353673	2, 6, 12, 16	C	353112	1, 8, 11, 13
A	353120	4, 9, 11, 15	B	353675	4, 9, 11, 13	C	353118	1, 10, 12, 14
A	353131	2, 8, 12, 15	B	353682	3, 9, 11, 14	C	353120	3, 7, 11, 14

Class	Student	Question Groups	Class	Student	Question Groups	Class	Student	Question Groups
A	353542	3, 8, 11, 16	B	353683	1, 8, 12, 14	C	353126	4, 9, 12, 16
A	353555	3, 7, 11, 14	B	353885	1, 9, 11, 16	C	353127	2, 9, 12, 14
A	353667	3, 9, 11, 14	B	353886	2, 10, 11, 16	C	353532	5, 6, 12, 15
A	353685	4, 8, 11, 14	B	354107	2, 6, 12, 14	C	353554	3, 10, 11, 14
A	353687	2, 6, 11, 13	B	354430	4, 8, 12, 16	C	353878	1, 6, 11, 13
A	353879	4, 7, 11, 14	B	354434	5, 10, 11, 14	C	353893	5, 7, 12, 13
A	353897	3, 9, 12, 16	B	354435	4, 8, 11, 13	C	353895	4, 10, 12, 16
A	353906	1, 8, 11, 14	B	354438	2, 9, 12, 15	C	353900	1, 6, 11, 14
A	354106	5, 10, 11, 13	B	354440	4, 6, 12, 14	C	353908	4, 10, 11, 15
A	354108	2, 6, 11, 15	B	354443	5, 7, 11, 14	C	354093	3, 9, 11, 15
A	354495	5, 7, 11, 14	B	354577	5, 6, 12, 15	C	354094	2, 10, 12, 16
A	354583	1, 10, 11, 13	B	354653	4, 10, 12, 14	C	354104	3, 8, 11, 14
A	354651	5, 6, 11, 13				C	354431	4, 6, 12, 14
						C	354441	5, 6, 12, 16
						C	354498	2, 6, 12, 16
						C	354646	1, 7, 12, 14

Question Groups

Group 1 - Coethnic Voting in Africa

To explore whether a political candidate can utilize his wife's ethnicity to garner coethnic support (where a voter prefers to vote for a candidate of his/her own ethnic group, and a well-established phenomenon in many African democracies), a group of researchers used observational time-series cross sectional data from the Afrobarometer (a public attitude survey on democracy and governance in more than 35 countries in Africa, see Afrobarometer) to establish patterns of preferring a president based on a coethnic presidential wife. The researchers then conducted an experiment where they randomly reminded potential voters in Benin that the Beninoise President Yayi Boni's wife, Chantal, is of the ethnic Fon group and asked them whether they approve of Yayi Boni. This exercise is based on:

Adida, Claire, Nathan Combes, Adeline Lo, and Alex Verink. 2016. "The Spousal Bump: Do Cross-Ethnic Marriages Increase Political Support in Multiethnic Democracies?" *Comparative Political Studies*, Vol. 49, No. 5, pp. 635-661.

In the first dataset from Afrobarometer, the researchers focus on African democracies where information could be garnered about the ethnicities of the president and wife. For the purposes of this exercise, only African democracies where the president and wife are not of the same ethnicity are considered (i.e., the president and wife are not coethnic with one another), and the data is pre-subsetted to include only president non-coethnics. We will consider patterns of willingness to vote for the president amongst wife coethnics and non-coethnics across African democracies. Descriptions of the relevant variables in the data file `afb.csv` are:

Name	Description
`country`	Character variable indicating which country the respondent is from
`wifecoethnic`	`1` if respondent is same ethnicity as president's wife, and `0` otherwise
`oppcoethnic`	`1` if respondent is same ethnicity as main presidential opponent, and `0` otherwise

Name	Description
`ethnicpercent`	Respondent's ethnic group fraction in respondent country.
`vote`	`1` if respondent would vote for the president, `0` otherwise.

The second dataset is a survey experiment in Cotonou, Benin. Here the researchers randomly assigned survey respondents short biolographical passages on the then Beninoise president Yayi Boni that included no mention of his wife, included a mention of his wife, or included a mention of his Fon wife. Respondents were then asked whether they were willing to vote for Yayi Boni should an election be held and barring term limits. The goal of the experiment was to assess whether priming respondents about the president's Fon wife might raise support amongst wife coethnics for the president. Two pre-subsetted data from `benin.csv` are also provided: `coethnic.csv` which subsets `benin.csv` to only coethnic respondents with the wife, and `noncoethnic.csv` which subsets `benin.csv` to only noncoethnic respondents with the wife. Descriptions of the relevant variables in the data file `benin.csv` (and consequently `coethnic.csv` and `noncoethnic.csv`) are:

Name	Description
`sex`	`1` if respondent is female, and `0` otherwise
`age`	Age of the respondent
`ethnicity`	Ethnicity of the respondent
`fon`	`1` if respondent is Fon, and `0` otherwise.
`passage`	`Control` if respondent given control passage, `Wife` for wife passage, `FonWife` for Fon wife passage
`vote`	`1` if respondent would vote for the president, `0` otherwise.

Question 1

Load the `afb.csv` data set. Look at a summary of the `afb` data to get a sense of what it looks like. Obtain a list of African democracies that are in the data set. Create a new binary variable, which is equal to `1` if the `ethnicpercent` variable is greater than its mean and is equal to `0` otherwise. Call this new variable `ethnicpercent2`.

Question 2

What is the average willingness to vote for the president among all respondents? Now compute the average willingess separately for respondents who are coethnic with the presidential wife and respondents who are not. Given our initial hypothesis about how a president might be able to use his wife's ethnicity to get more support, how might we interpret the differences (or similarities) in the support amongst coethnics and non-coethnics?

Question 3

We might be concerned that we have not taken into account potentially confounding factors such as whether 1) the respondent is part of a proportionally larger or smaller ethnic group and 2) whether the respondent is also coethnic with the major opposition leader. This is because if a respondent's ethnic group is quite small, the members might be less able to put forth a candidate of their exact ethnic label and have more incentive to support a president who, while not the same ethnicity, has a wife who does (and who therefore might have the wife's ethnic group interests at heart). It may also be that should an opposition candidate hold the same ethnicity as the respondent, such a "wife effect" might be diminished.

To investigate this possibility, subset the `afb` data to adjust for potential confounding variable `ethnicpercent2` created in the previous question. Consider the group of individuals who are of smaller than average ethnic groups. What is the average willingness to vote between wife coethnics and wife non-coethnics? Next, consider only the group of individuals who are not only from smaller than average ethnic groups but are also not coethnic with the opponent. What is the difference in average willingness to vote between wife coethnics and wife non-coethnics now? What do these results tell us about the relationship between the "wife effect"?

Question 4

The Afrobarometer data, while rich and inclusive of many countries, is observational data. Thus, it is difficult to estimate the effect of *treatment*, which is coethnicity with the president's wife in the current application. To address this difficulty, the authors of the study conduct a survey experiment in Benin, a small democracy on the western coast of the African continent. It is also a country represented in the Afrobarometer data set. The president at the time of the survey was Yayi Boni, who is of two ethnicities, Nago and Bariba. His wife Chantal is Fon. For the experiment, the authors randomly surveyed adult walkers on the streets of Cotonou (the capital of Benin). Respondents were asked some personal information, such as gender and age, as well as their ethnicity. Then, respondents were randomly assigned to either the control or one of two treatment groups (*Wife* and *Fon Wife*):

In the control condition, respondents were read the following short biographical sketchof Yayi Boni, where there is no indication of the president's wife, Chantal:

> Yayi Boni became President of Benin on April 6, 2006 and was just re-elected for a second term. He has led a presidential campaign based on economic growth and suppressing corruption. However, some critics claim that the country's economic growth has been disappointing, and that Boni's administration is, itself, corrupt.

In the first treatment group, *Wife*, respondents were read the same passage as the control group, except the president's wife Chantal is explicitly mentioned at the beginning. That is, the above script is preceded with "Accompanied by his wife, Chantal". In the second treatment group, *Fon Wife*, respondents were read again the same passage, except the ethnicity of Chantal is explicitly mentioned with the script starting by "Accompnaied by his Fon wife, Chantal".

Now we turn to the `benin` dataset. Does being reminded that you are coethnic with the president's wife increase your willingness to vote for the president? The data has already been subsetted from the original experiment data so it contains only respondents who are not coethnic with the president (why would this be important to consider?). Take a closer look at the `ethnicity` variable by creating a table. How many ethnic groups are there represented in this dataset? Compare the mean willingness to vote for the president between the *Fon Wife* and control group. Briefly interpret the result. Was it important for the researchers to add a treatment with just the mention of the president's wife without her ethnicity? Why or why not?

Question 5

Now compare the mean willingness to vote for the president between the *Fon Wife* and control group for wife coethnics only (load `coethnic.csv` file). Briefly interpret the result. What happens when we compare wife coethnics in the *Fon Wife* to the *Wife* group? The *Wife* to the control group? Do these results apply to respondents who are NOT coethnic with the president's wife (load `noncoethnic.csv` file)?

Group 2 - The Mark of a Criminal Record

In this exercise, we analyze the causal effects of a criminal record on the job prospects of white and black job applicants. This exercise is based on:

Pager, Devah. (2003). "The Mark of a Criminal Record." *American Journal of Sociology* 108(5):937-975. You are also welcome to watch Professor Pager discuss the design and result here.

To isolate the causal effect of a criminal record for black and white applicants, Pager ran an audit experiment. In this type of experiment, researchers present two similar people that differ only according to one trait thought to be the source of discrimination. This approach was used in the resume experiment, where researchers randomly assigned stereotypically African-American-sounding names and stereotypically white-sounding names to otherwise identical resumes to measure discrimination in the labor market.

To examine the role of a criminal record, Pager hired a pair of white men and a pair of black men and instructed them to apply for existing entry-level jobs in the city of Milwaukee. The men in each pair were matched on a number of dimensions, including physical appearance and self-presentation. As much as possible, the only difference between the two was that Pager randomly varied which individual in the pair would indicate to potential employers that he had a criminal record. Further, each week, the pair alternated which applicant would present himself as an ex-felon. To determine how incarceration and race influence employment chances, she compared callback rates among applicants with and without a criminal background and calculated how those callback rates varied by race.

In the data you will use `ciminalrecord.csv` nearly all these cases are present, but 4 cases have been redacted. As a result, your findings may differ slightly from those in the paper. The names and descriptions of variables are shown below. You may not need to use all of these variables for this activity. We've kept these unnecessary variables in the dataset because it is common to receive a dataset with much more information than you need.

Name	Description
`jobid`	Job ID number
`callback`	1 if tester received a callback, 0 if the tester did not receive a callback.
`black`	1 if the tester is black, 0 if the tester is white.
`crimrec`	1 if the tester has a criminal record, 0 if the tester does not.
`interact city`	1 if tester interacted with employer during the job application, 0 if tester does not interact with employer. 1 is job is located in the city center, 0 if job is located in the suburbs.
`distance`	Job's average distance to downtown.
`custserv`	1 if job is in the costumer service sector, 0 if it is not.
`manualskill`	1 if job requires manual skills, 0 if it does not.

Question 1

Begin by loading the data into R and explore the data. How many cases are there in the data? Run `summary()` to get a sense of things. In how many cases is the tester black? In how many cases is he white?

Question 2

Now we examine the central question of the study. Calculate the proportion of callbacks for white applicants with and without a criminal record, and calculate this proportion for black applicants with and without a criminal record.

Question 3

What is the difference in callback rates between individuals with and without a criminal record within each race. What do these specific results tell us? Consider both the difference in callback rates for records with and without a criminal record and the ratio of callback rates for these two types of records.

Question 4

Compare the callback rates of whites *with* a criminal record versus blacks *without* a criminal record. What do we learn from this comparison?

Question 5

When carrying out this experiment, Pager made many decisions. For example, she opted to conduct the research in Milwaukee; she could have done the same experiment in Dallas or Mexico City or Sao Paulo. She ran the study at a specific time: between June and December of 2001. But, she could have also run it at a different time, say 5 years earlier or 5 years later. Pager decided to hire 23-year-old male college students as her testers; she could have done the same experiment with 23-year-old female college students or 40-year-old male high school drop-outs. Further, the criminal record she randomly assigned to her testers was a felony convinction related to drugs (possession with intent to distribute, cocaine). But, she could have assigned her testers a felony conviction for assault or tax evasion. Pager was very aware of each of these decisions, and she discusses them in her paper. Now you should pick *one* of these decisions described above or another decision of your choosing. Speculate about how the results of the study might (or might not) change if you were to conduct the same study but alter this specific decision. This is part of thinking about the *external validity* of the study.

Group 3 - Efficiency of Small Classes in Early Education

The STAR (Student-Teacher Achievement Ratio) Project is a four year *longitudinal study* examining the effect of class size in early grade levels on educational performance and personal development.

This exercise is in part based on: Mosteller, Frederick. 1997. "The Tennessee Study of Class Size in the Early School Grades." *Bulletin of the American Academy of Arts and Sciences* 50(7): 14-25.

A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, collection of various measurements (e.g., performance on tests in eighth grade, overall high school GPA) continued through the end of participants' high school attendance.

We will analyze just a portion of this data to investigate whether the small class sizes improved performance or not. The data file name is `STAR.csv`, which is a CSV data file. The names and descriptions of variables in this data set are:

Name	Description
`race`	Student's race (White = 1, Black = 2, Asian = 3, Hispanic = 4, Native American = 5, Others = 6)

Name	Description
`classtype`	Type of kindergarten class (small = 1, regular = 2, regular with aid = 3)
`g4math`	Total scaled score for math portion of fourth grade standardized test
`g4reading`	Total scaled score for reading portion of fourth grade standardized test
`yearssmall`	Number of years in small classes
`hsgrad`	High school graduation (did graduate = 1, did not graduate = 0)

Note that there are a fair amount of missing values in this data set. For example, missing values arise because some students left a STAR school before third grade or did not enter a STAR school until first grade.

Question 1

Create a new factor variable called `kinder` in the data frame. This variable should recode `classtype` by changing integer values to their corresponding informative labels (e.g., change 1 to `small` etc.). Similarly, recode the `race` variable into a factor variable with four levels (`white`, `black`, `hispanic`, `others`) by combining Asians and Native Americans as the `others` category. For the `race` variable, overwrite the original variable in the data frame rather than creating a new one.

First of all, I need to install the following:

```
STAR <- read.csv("https://raw.githubusercontent.com/umbertomig/intro-prob-stat-FGV/master/datasets/STAR
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyverse_conflic
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Creating a new factor and recoding 'race':

```
STAR_kinder_race <- mutate(STAR,
                  kinder = ifelse(classtype == 1, "small",
                            ifelse(classtype == 2, "regular", "regaid")),
                  race = ifelse(race == 1, "white",
                           ifelse(race == 2, "black",
                             ifelse(race == 4, "hispanic","others"))))
summary(STAR_kinder_race)
```

```
##      race             classtype       yearssmall         hsgrad
##  Length:6325        Min.   :1.000   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000
##  Mode  :character   Median :2.000   Median :0.0000   Median :1.000
##                     Mean   :2.052   Mean   :0.9542   Mean   :0.833
##                     3rd Qu.:3.000   3rd Qu.:2.0000   3rd Qu.:1.000
```

```
##                    Max.   :3.000   Max.   :4.0000   Max.   :1.000
##                                                     NA's   :3278
##      g4math        g4reading        kinder
##  Min.   :487.0   Min.   :528.0   Length:6325
##  1st Qu.:688.0   1st Qu.:696.0   Class :character
##  Median :710.0   Median :723.0   Mode  :character
##  Mean   :708.8   Mean   :721.2
##  3rd Qu.:732.5   3rd Qu.:750.0
##  Max.   :821.0   Max.   :836.0
##  NA's   :3930    NA's   :3972
```

Question 2

How does performance on fourth grade reading and math tests for those students assigned to a small class in kindergarten compare with those assigned to a regular-sized class? Do students in the smaller classes perform better? Use means to make this comparison while removing missing values. Give a brief substantive interpretation of the results. To understand the size of the estimated effects, compare them with the standard deviation of the test scores.

Comparing reading grades and the standard deviation:

```
STAR_small_read <- select(STAR_kinder_race,
                      kinder, g4reading) %>%
  filter(kinder == "small")
mean(STAR_small_read$g4reading, na.rm = T)
```

```
## [1] 723.3912
```

```
sd(STAR_small_read$g4reading, na.rm = T)
```

```
## [1] 51.54494
```

```
STAR_reg_read <- select(STAR_kinder_race,
                      kinder, g4reading) %>%
  filter(kinder == "regular")
mean(STAR_reg_read$g4reading, na.rm = T)
```

```
## [1] 719.89
```

```
sd(STAR_reg_read$g4reading, na.rm = T)
```

```
## [1] 53.16788
```

As we can see, the mean of the reading grades is a bit higher for small class.

Comparing math grades and the standard deviation:

```
STAR_small_math <- select(STAR_kinder_race,
                      kinder, g4math) %>%
  filter(kinder == "small")
mean(STAR_small_math$g4math, na.rm = T)
```

```
## [1] 709.1851
```

```
sd(STAR_small_math$g4math, na.rm = T)
```

```
## [1] 43.57318
```

```
STAR_reg_math <- select(STAR_kinder_race,
                        kinder, g4math) %>%
  filter(kinder == "regular")
mean(STAR_reg_math$g4math, na.rm = T)
```

```
## [1] 709.5214
```

```
sd(STAR_reg_math$g4math, na.rm = T)
```

```
## [1] 41.02063
```

As we can see, the mean of the reading grades is a little bit higher for regular-sized class (but it is almost the same).

Question 3

Instead of comparing just average scores of reading and math tests between those students assigned to small classes and those assigned to regular-sized classes, look at the entire range of possible scores. To do so, compare a high score, defined as the 66th percentile, and a low score (the 33rd percentile) for small classes with the corresponding score for regular classes. These are examples of *quantile treatment effects*. Does this analysis add anything to the analysis based on mean in the previous question?

Comparing the high scores for reading:

```
quantile(STAR_small_read$g4reading, p=c(0.66), na.rm = T)
```

```
## 66%
## 741
```

```
quantile(STAR_reg_read$g4reading, p=c(0.66), na.rm = T)
```

```
## 66%
## 740
```

It means that the highest grades are almost the same, but in average, the grades in small classes are higher than in regular ones.

Comparing the high scores for math:

```
quantile(STAR_small_math$g4math, p=c(0.33), na.rm = T)
```

```
## 33%
## 694
```

```
quantile(STAR_reg_math$g4math, p=c(0.33), na.rm = T)
```

```
## 33%
## 696
```

However, the lowest grades are almost the same, as well as the means.

Question 4

We examine whether the STAR program reduced the achievement gaps across different racial groups. Begin by comparing the average reading and math test scores between white and minority students (i.e., Blacks and hispanics) among those students who were assigned to regular classes with no aid. Conduct the same comparison among those students who were assigned to small classes. Give a brief substantive interpretation of the results of your analysis.

Comparing the average of reading grades:

```
STAR_small_read_white <- select(STAR_kinder_race,
                      race, kinder, g4reading) %>%
  filter(kinder == "small", race == "white")
mean(STAR_small_read_white$g4reading, na.rm = T)
```

```
## [1] 727.8388
```

```
STAR_reg_read_white <- select(STAR_kinder_race,
                      race, kinder, g4reading) %>%
  filter(kinder == "regular", race == "white")
mean(STAR_reg_read_white$g4reading, na.rm = T)
```

```
## [1] 725.1158
```

The mean for white students is almost the same in both cases.

```
STAR_small_read_black <- select(STAR_kinder_race,
                      race, kinder, g4reading) %>%
  filter(kinder == "small", race == "black")
mean(STAR_small_read_black$g4reading, na.rm = T)
```

```
## [1] 698.614
```

```
STAR_reg_read_black <- select(STAR_kinder_race,
                      race, kinder, g4reading) %>%
  filter(kinder == "regular", race == "black")
mean(STAR_reg_read_black$g4reading, na.rm = T)
```

```
## [1] 689.3548
```

The mean for black students in small classes is higher than in regular ones.

```
STAR_small_read_hisp <- select(STAR_kinder_race,
                     race, kinder, g4reading) %>%
  filter(kinder == "small", race == "hispanic")
mean(STAR_small_read_hisp$g4reading, na.rm = T)
```

```
## [1] 737.5
```

```
STAR_reg_read_hisp <- select(STAR_kinder_race,
                     race, kinder, g4reading) %>%
  filter(kinder == "regular", race == "hispanic")
mean(STAR_reg_read_hisp$g4reading, na.rm = T)
```

```
## [1] NaN
```

As we can see, there is no hispanic students in regular-sized classes in the dataframe. Besides that, we can see that the average for hispanic students in small classes is higher than for white students.

Question 5

We consider the long term effects of kindergarden class size. Compare high school graduation rates across students assigned to different class types. Also, examine whether graduation rates differ by the number of years spent in small classses. Finally, as done in the previous question, investigate whether the STAR program has reduced the racial gap between white and minority students' graduation rates. Briefly discuss the results.

Comparing high school graduation rates across students assigned to different class types:

```
STAR_small_hs <- select(STAR_kinder_race,
                     kinder, hsgrad) %>%
  filter(kinder == "small")
mean(STAR_small_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8359202
```

```
STAR_reg_hs <- select(STAR_kinder_race,
                     kinder, hsgrad) %>%
  filter(kinder == "regular")
mean(STAR_reg_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8251619
```

```
STAR_regaid_hs <- select(STAR_kinder_race,
                     kinder, hsgrad) %>%
  filter(kinder == "regaid")
mean(STAR_regaid_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8392857
```

The mean is still close, in spite of the lowest value is the regular-sized class' one.

Examining whether graduation rates differ by the number of years spent in small classses:

```
STAR_0_hs <- select(STAR_kinder_race,
                    yearssmall, hsgrad) %>%
  filter(yearssmall == "0")
mean(STAR_0_hs$hsgrad, na.rm = T)
```

```
## [1] 0.828602
```

```
STAR_1_hs <- select(STAR_kinder_race,
                    yearssmall, hsgrad) %>%
  filter(yearssmall == "1")
mean(STAR_1_hs$hsgrad, na.rm = T)
```

```
## [1] 0.7910448
```

```
STAR_2_hs <- select(STAR_kinder_race,
                    yearssmall, hsgrad) %>%
  filter(yearssmall == "2")
mean(STAR_2_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8131868
```

```
STAR_3_hs <- select(STAR_kinder_race,
                    yearssmall, hsgrad) %>%
  filter(yearssmall == "3")
mean(STAR_3_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8324607
```

```
STAR_4_hs <- select(STAR_kinder_race,
                    yearssmall, hsgrad) %>%
  filter(yearssmall == "4")
mean(STAR_4_hs$hsgrad, na.rm = T)
```

```
## [1] 0.877551
```

As we can see, it is more likely get a high score spending no year than one year at small-sized classes(0.79 and 0.82 respectively). Besides that, it seems like spend four year in a small-sized class improves the likelihood to get the higher score (0.87 in average).

Investigating whether the STAR program has reduced the racial gap between white and minority students' graduation rates:

```
STAR_white_hs <- select(STAR_kinder_race,
                    race, hsgrad) %>%
  filter(race == "white")
mean(STAR_white_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8676272
```

```
STAR_black_hs <- select(STAR_kinder_race,
                        race, hsgrad) %>%
  filter(race == "black")
mean(STAR_black_hs$hsgrad, na.rm = T)
```

```
## [1] 0.7392901
```

```
STAR_hisp_hs <- select(STAR_kinder_race,
                       race, hsgrad) %>%
  filter(race == "hispanic")
mean(STAR_hisp_hs$hsgrad, na.rm = T)
```

```
## [1] NaN
```

```
STAR_others_hs <- select(STAR_kinder_race,
                         race, hsgrad) %>%
  filter(race == "others")
mean(STAR_others_hs$hsgrad, na.rm = T)
```

```
## [1] 0.8888889
```

As we can see, minoritys in 'others', such as Asian students, have a higher mean grade than white students.

Group 4 - Bias in Self-reported Turnout

Surveys are frequently used to measure political behavior such as voter turnout, but some researchers are concerned about the accuracy of self-reports. In particular, they worry about possible *social desirability bias* where in post-election surveys, respondents who did not vote in an election lie about not having voted because they may feel that they should have voted. Is such a bias present in the American National Election Studies (ANES)? The ANES is a nation-wide survey that has been conducted for every election since 1948. The ANES conducts face-to-face interviews with a nationally representative sample of adults. The table below displays the names and descriptions of variables in the `turnout.csv` data file.

Name	Description
`year`	Election year
`VEP`	Voting Eligible Population (in thousands)
`VAP`	Voting Age Population (in thousands)
`total`	Total ballots cast for highest office (in thousands)
`felons`	Total ineligible felons (in thousands)
`noncitizens`	Total non-citizens (in thousands)
`overseas`	Total eligible overseas voters (in thousands)
`osvoters`	Total ballots counted by overseas voters (in thousands)

Question 1

Calculate the turnout rate based on the voting age population or VAP. Note that for this data set, we must add the total number of eligible overseas voters since the *VAP* variable does not include these individuals in the count. Next, calculate the turnout rate using the voting eligible population or VEP. What difference do you observe?

Question 2

Compute the difference between *VAP* and *ANES* estimates of turnout rate. How big is the difference on average? What is the range of the difference? Conduct the same comparison for the VEP and ANES estimates of voter turnout. Briefly comment on the results.

Question 3

Compare the VEP turnout rate with the ANES turnout rate separately for presidential elections and midterm elections. Note that the data set excludes the year 2006. Does the bias of the ANES vary across election types?

Question 4

Divide the data into half by election years such that you subset the data into two periods. Calculate the difference between the VEP turnout rate and the ANES turnout rate separately for each year within each period. Has the bias of the ANES increased over time?

Question 5

The ANES does not interview overseas voters and prisoners. Calculate an adjustment to the 2008 VAP turnout rate. Begin by subtracting the total number of ineligible felons and non-citizens from the VAP to calculate an adjusted VAP. Next, calculate an adjusted VAP turnout rate, taking care to subtract the number of overseas ballots counted from the total ballots in 2008. Compare the adjusted VAP turnout with the unadjusted VAP, VEP, and the ANES turnout rate. Briefly discuss the results.

Group 5 - Effect of Demographic Change on Exclusionary Attitudes

A researcher conducted a randomized field experiment assessing the extent to which individuals living in suburban communities around Boston, Massachusetts, and their views were affected by exposure to demographic change.

This exercise is based on: Enos, R. D. 2014. "Causal Effect of Intergroup Contact on Exclusionary Attitudes." *Proceedings of the National Academy of Sciences* 111(10): 3699–3704.

Subjects in the experiment were individuals riding on the commuter rail line and overwhelmingly white. Every morning, multiple trains pass through various stations in suburban communities that were used for this study. For pairs of trains leaving the same station at roughly the same time, one was randomly assigned to receive the treatment and one was designated as a control. By doing so all the benefits of randomization apply for this dataset.

The treatment in this experiment was the presence of two native Spanish-speaking 'confederates' (a term used in experiments to indicate that these individuals worked for the researcher, unbeknownst to the subjects) on the platform each morning prior to the train's arrival. The presence of these confederates, who would appear as Hispanic foreigners to the subjects, was intended to simulate the kind of demographic change anticipated for the United States in coming years. For those individuals in the control group, no such confederates were present on the platform. The treatment was administered for 10 days. Participants were asked questions related to immigration policy both before the experiment started and after the experiment had ended. The names and descriptions of variables in the data set `boston.csv` are:

Name	Description
`age`	Age of individual at time of experiment
`male`	Sex of individual, male (1) or female (0)
`income`	Income group in dollars (not exact income)
`white`	Indicator variable for whether individual identifies as white (1) or not (0)
`college`	Indicator variable for whether individual attended college (1) or not (0)
`usborn`	Indicator variable for whether individual is born in the US (1) or not (0)
`treatment`	Indicator variable for whether an individual was treated (1) or not (0)
`ideology`	Self-placement on ideology spectrum from Very Liberal (1) through Moderate (3) to Very Conservative (5)
`numberim.pre`	Policy opinion on question about increasing the number immigrants allowed in the country from Increased (1) to Decreased (5)
`numberim.post`	Same question as above, asked later
`remain.pre`	Policy opinion on question about allowing the children of undocumented immigrants to remain in the country from Allow (1) to Not Allow (5)
`remain.post`	Same question as above, asked later
`english.pre`	Policy opinion on question about passing a law establishing English as the official language from Not Favor (1) to Favor (5)
`english.post`	Same question as above, asked later

Question 1

Load the dataset and summarize the variables. Use the function `summary(.)`.

Question 2

The benefit of randomly assigning individuals to the treatment or control groups is that the two groups should be similar, on average, in terms of their covariates. This is referred to as 'covariate balance.' Show that the treatment and control groups are balanced with respect to the income variable (`income`) by comparing its distribution between those in the treatment group and those in the control group. Also, compare the proportion of males (`male`) in the treatment and control groups. Interpret these two numbers.

Question 3

Individuals in the experiment were asked a series of questions both at the beginning and the end of the experiment. One such question was "Do you think the number of immigrants from Mexico who are permitted to come to the United States to live should be increased, left the same, or decreased?" The response to this question prior to the experiment is in the variable `numberim.pre`. The response to this question after the experiment is in the variable `numberim.post`. In both cases the variable is coded on a $1 - 5$ scale. Responses with values of 1 are inclusionary ('pro-immigration') and responses with values of 5 are exclusionary ('anti-immigration'). Compute the average treatment effect on the change in attitudes about immigration. That

is, how does the mean change in attitudes about immigration policy for those in the control group compare to those in the treatment group. Interpret the result.

Question 4

Does having attended college influence the effect of being exposed to 'outsiders' on exclusionary attitudes? Another way to ask the same question is this: is there evidence of a differential impact of treatment, conditional on attending college versus not attending college? Calculate the necessary quantities to answer this question and interpret the results. Consider the average treatment effect for those who attended college and then those who did not.

Question 5

Repeat the same analysis as in the previous question but this time with respect to age and ideology. For age, divide the data based on its quartile and compute the average treatment effect within each of the resulting four groups. For ideology, compute the average treatment effect within each value. What patterns do you observe? Give a brief substantive interpretation of the results.

Group 6 - Sources of Empathy in the Circuit Courts

In this exercise, we will analyze the relationship between various demographic traits and pro-feminist voting behavior among circuit court judges. In a recent paper, Adam N. Glynn and Maya Sen argue that having a female child causes circuit court judges to make more pro-feminist decisions. The paper can be found at:

Glynn, Adam N., and Maya Sen. (2015). "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?." *American Journal of Political Science* Vol. 59, No. 1, pp. 37–54.

The dataset `dbj.csv` contains the following variables about individual judges:

Name	Description
`name`	The judge's name
`child`	The number of children each judge has.
`circuit.1`	Which federal circuit the judge serves in.
`girls`	The number of female children the judge has.
`progressive.vote`	The proportion of the judge's votes on women's issues which were decided in a pro-feminist direction.
`race`	The judge's race (`1` = white, `2` = African-American, `3` = Hispanic, `4` = Asian-American).
`religion`	The judge's religion (`1` = Unitarian, `2` = Episcopalian, `3` = Baptist, `4` = Catholic, `5` = Jewish, `7` = Presbyterian, `8` = Protestant, `9` = Congregationalist, `10` = Methodist, `11` = Church of Christ, `16` = Baha'i, `17` = Mormon, `21` = Anglican, `24` = Lutheran, `99` = unknown).
`republican`	Takes a value of `1` if the judge was appointed by a Republican president, `0` otherwise. Used as a proxy for the judge's party.
`sons`	The number of male children the judge has.
`woman`	Takes a value of `1` if the judge is a woman, `0` otherwise.
`X`	Indicator for the observation number.
`yearb`	The year the judge was born.

Question 1

Load the `dbj.csv` file. Find how many judges there are in the dataset, as well as the gender and party composition of our dataset. Is the party composition different for male and female judges? Additionally, note that our outcome in this exercise will be the proportion of pro-feminist rulings. What is the range of this variable (`progressive.vote`)?

Question 2

Next, we consider differences between some groups. For each of the four groups (Republican men/women, Democratic men/women) defined by gender and partisanship, create a boxplot (using a single command) that illustrates the differences in `progressive.vote`. Briefly interpret the results of the analysis. For example, do any of the results surprise you? Does it appear that partisanship, gender, or both contribute to progressive voting patterns? Should we interpret any of these effects causally? Why or why not?

Question 3

Create a new binary variable which takes a value of `1` if a judge has *at least* one child (that is, any children at all), `0` otherwise. Then, use this variable to answer the following questions. Are Republicans and Democrats equally likely to be parents (that is, have at least one child)? Do judges with children vote differently than judges without? If so, how are they different? Do republican and democratic parents vote differently on feminist issues?

Question 4

What is the difference in the proportion of pro-feminist decisions between judges who have at least one daughter and those who do not have any? Compute this difference in two ways; (1) using the entire sample, (2) separately by the number of children judges have (only considering judges that have 3 children or less). What assumptions are required for us to interpret these differences as causal estimates?

Question 5

Next, we are going to consider the design of this study. The original authors assume that conditional on the number of children a judge has the number of daughters is random (as we did in the previous question). Indeed, this is the assumption that would justify the analysis of the previous question. For example, among the judges who have two children, the number of daughters – either `0`, `1`, or `2` – has nothing to do with the (observed or unobserved) pre-treatment characteristics of judges. Is this assumption reasonable? Is there a scenario under which this assumption can be violated? Do the data support the assumption?

Group 7 - Diverse Mechanisms of Migration

Scholars across disciplines have identified several mechanisms that cause people to migrate. Some propose an "income maximizer" hypothesis and argue that individuals migrate because they are drawn to higher wages in receiving countries. Others argue that it is risk and uncertainty in the sending countries–such as low-wages and lack of market opportunities– that is driving migration patterns. They offer a "risk diversifier" hypothesis. Others still hypothesize that growing ties among individuals in receiving and sending countries

fosters immigration, and advocate for analyses that focus on "network migrants." In this exercise, rather than examining them as competing hypotheses, we examine these theories together and test whether each represents the profile of a different stream of migrants from Mexico to the U.S. in recent decades. Using cluster analysis, we attempt to discover the "configurations of various attributes that characterize different migrant types." This exercise is based on the following article:

Garip, Filiz. 2012. "Discovering Diverse Mechanisms of Migration: The Mexico–US Stream 1970–2000." *Population and Development Review*, Vol. 38, No. 3, pp. 393-433.

The data come from the **Mexican Migration Project**, a survey of Mexican migrants from 124 communities located in major migrant-sending areas in 21 Mexican states. Each community was surveyed once between 1987 and 2008, during December and January, when migrants to the U.S. are mostly likely to visit their families in Mexico. In each community, individuals (or informants for absent individuals) from about 200 randomly selected households were asked to provide demographic and economic information and to state the time of their first and their most recent trip to the United States.

The data set is the file `migration.csv`. Variables in this dataset can be broken down into three categories:

INDIVIDIUAL LEVEL VARIABLES

Name	Description
`year`	Year of respondent's first trip to the U.S.
`age`	Age of respondent
`male`	1 if respondent is male, 0 if respondent is female
`educ`	Years of education: secondary school in Mexico is from years 7 to 12

HOUSEHOLD LEVEL VARIABLES

Name	Description
`log_nrooms`	Logged number of rooms across all properties owned by respondent's household
`log_landval`	Logged value of all land owned by respondent's household (U.S. dollars)
`n_business`	Number of businesses owned by respondent
`prop_hhmig`	Proportion of respondent's household who are also U.S. migrants

COMMUNITY LEVEL VARIABLES

Name	Description
`prop_cmig`	Proportion of respondent's community who are also U.S. migrants
`log_npop`	Logged size of respondent's community.
`prop_self`	Proportion of respondent's community who are self-employed
`prop_agri`	Proportion of respondent's community involved in agriculture
`prop_lessminwage`	Proportion of respondent's community who earn less than the U.S. minimum wage

Question 1

Examine the mean values for the individual level, household level, and community level characteristics in the dataset. Briefly interpret your answers.

Question 2

Use scatterplots to investigate the relationship between `prop_self` and `prop_agri`, as well as the relationship between `prop_self` and `log_npop`. Briefly interpret these scatter plots and what they imply about self-employed workers. Do these relationships appear to be independent? What does knowing that a migrant is self-employed tell us about them? Then calcuate the correlation for all possible interactions of the four community level variables: `prop_self`, `prop_agri`, `prop_lessminwage`, and `log_npop`. Use these correlations to help with your interpretation of the scatter plots. Does adding the `prop_lessminwage` variable add anything to your interpretation?

Question 3

We'll focus on the variables: `year`, `educ`, `log_nrooms`, `log_landval`, `n_business`, `prop_hhmig`, `prop_cmig`, `log_npop`, `prop_self`, `prop_agri`, and `prop_lessminwage`. Remove observations with missing values. Then, subset your dataset to all of your variables **except** `year`, and use the `scale()` function to standardize the variables in your subsetted dataset so that they are comparable. Compare the means and standard deviations before and after scaling. Standardizing substracts the mean of a variable from each observation and divides by the standard deviation.

Question 4

Fit the k-means clustering algorithm with *three* clusters, using the scaled variables from the data set with no missing values. Insert the code `set.seed(2016)` right before your cluster analysis so that you can compare your results from the kmeans clustering to exercise solutions later. How many observations are assigned to each cluster? Each cluster has a center. What do the centers of these clusters represent? Interpret the type of migrant described by cluster 1. To help witih interpretability, you can also calculate the mean value of the variables for each cluster, using their original scale. Repeat the cluster analysis. This time with *four* centers. How are the two results different? Is there one you prefer?

Question 5

Do these different clusters represent different temporal trends in migration from Mexico to the US? Use a time-series plot to graph the proportions of migrants in each of the four clusters from Question 4 over time (variable `year`). Briefly describe the major trends you discover.

Group 8 - Oil, Democracy, and Development

Researchers have theorized that natural resources may have an inhibiting effect on the democratization process. Although there are multiple explanations as to why this might be the case, one hypothesis posits that governments in countries with large natural resource endowments (like oil) are able to fund their operations without taxing civilians. Since representation (and other democratic institutions) are a compromise offered

by governments in exchange for tax revenue, resource-rich countries do not need to make this trade. In this exercise, we will not investigate causal effects of oil on democracy. Instead, we examine whether the association between oil and democracy is consistent with the aforementioned hypothesis.

This exercise is in part based on Michael L. Ross. (2001). 'Does Oil Hinder Democracy?' *World Politics*, 53:3, pp.325-361.

The data set is in the csv file `resources.csv`. The names and descriptions of variables are:

Name	Description
`cty_name`	Country name
`year`	Year
`logGDPcp`	Logged GDP per capita
`regime`	A measure of a country's level of democracy: -10 (authoritarian) to 10 (democratic)
`oil`	Amount of oil exports as a percentage of the country's GDP
`metal`	Amount of non-fuel mineral exports as a percentage of the country's GDP
`illit`	Percentage of the population that is illiterate
`life`	Life expectancy in the country

Question 1

Load the dataset and provide summary statistics for all variables.

Question 2

Use scatterplots to examine the bivariate relationship between logged GDP per capita and life expectancy as well as between logged GDP per capita and illiteracy. Be sure to add informative axis labels. Also, compute the correlation separately for each bivariate relationship. Briefly comment on the results.

Question 3

We focus on the following subset of the variables: `regime`, `oil`, `logGDPcp`, and `illit`. Remove observations that have missing values in any of these variables. Using the `scale()` function, scale these variables so that each variable has a mean of zero and a standard deviation of one. Fit the k-means clustering algorithm with two clusters. How many observations are assigned to each cluster? Using the original unstandardized data, compute the means of these variables in each cluster.

Question 4

Using the clusters obtained above, modify the scatterplot between logged GDP per capita and illiteracy rate in the following manner. Use different colors for the clusters so that we can easily tell the cluster membership of each observation. In addition, make the size of each circle proportional to the `oil` variable so that oil-rich countries stand out. Briefly comment on the results.

Question 5

Repeat the previous two questions but this time with three clusters instead of two. How are the results different? Which clustering model would you prefer and why?

Group 9 - Poverty and Economic Decision-Making

Do changes in one's financial circumstances affect one's decision-making process and cognitive capacity? In an experimental study, researchers randomly selected a group of US respondents to be surveyed before their payday and another group to be surveyed after their payday. Under this design, the respondents of the `Before Payday` group are more likely to be financially strained than those of the `After Payday` group. The researchers were interested in investigating whether or not changes in people's financial circumstances affect their decision making and cognitive performance. Other researchers have found that scarcity induce an additional mental load that impedes cognitive capacity. This exercise is based on:

Carvalho, Leandro S., Meier, Stephen, and Wang, Stephanie W. (2016). "Poverty and economic decision-making: Evidence from changes in financial resources at payday." *American Economic Review*, Vol. 106, No. 2, pp. 260-284.

In this study, the researchers administered a number of decision-making and cognitive performance tasks to the `Before Payday` and `After Payday` groups. We focus on the *numerical stroop task*, which measures cognitive control. In general, taking more time to complete this task indicates less cognitive control and reduced cognitive ability. They also measured the amount of cash the respondents have, the amount in their checking and saving accounts, and the amount of money spent. The data set is in the CSV file `poverty.csv`. The names and descriptions of variables are given below:

Name	Description
`treatment`	Treatment conditions: `Before Payday` and `After Payday`
`cash`	Amount of cash respondent has on hand
`accts_amt`	Amount in checking and saving accounts
`stroop_time`	Log-transformed average response time for cognitive stroop test
`income_less20k`	Binary variable: `1` if respondent earns less than 20k a year and `0` otherwise

Question 1

Load the `poverty.csv` data set. Look at a summary of the `poverty` data set to get a sense of what its variables looks like. Use histograms to examine the univariate distributions of the two financial resources measures: `cash` and `accts_amt`. What can we tell about these variables' distributions from looking at the histograms? Evaluate what the shape of these distributions could imply for the authors' experimental design.

Loading the `poverty.csv` data set:

```
library(ggplot2)
poverty <- read.csv("https://raw.githubusercontent.com/umbertomig/intro-prob-stat-FGV/master/datasets/p
summary(poverty)
```

```
##   treatment             cash          accts_amt      stroop_time
##  Length:2670        Min.   :   0.0   Min.   :    0   Min.   :5.356
##  Class :character   1st Qu.:  15.0   1st Qu.:  176   1st Qu.:7.436
##  Mode  :character   Median :  49.5   Median : 1000   Median :7.564
##                     Mean   : 169.0   Mean   : 6211   Mean   :7.545
##                     3rd Qu.: 136.2   3rd Qu.: 5000   3rd Qu.:7.692
```

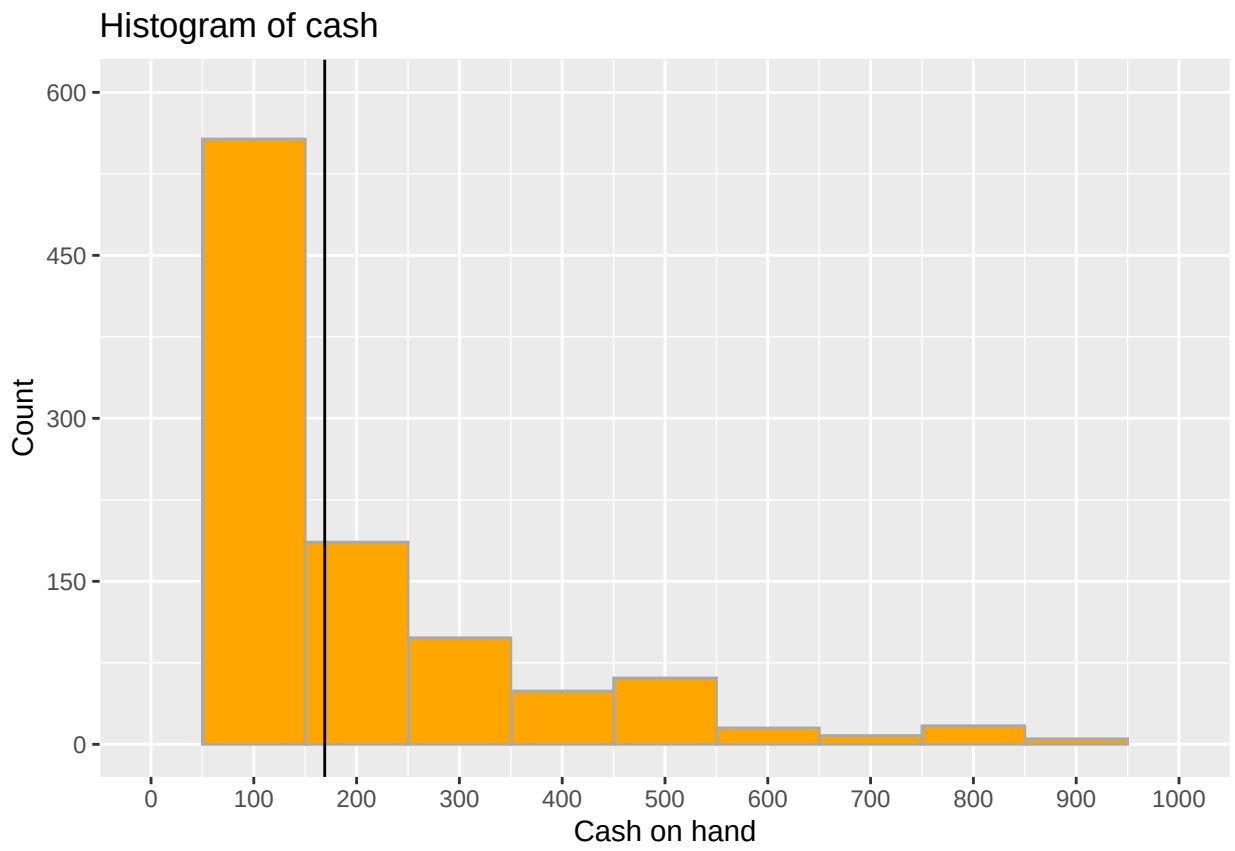
```
##                     Max.   :9000.0   Max.   :95000   Max.   :8.517
##                     NA's   :226      NA's   :433
##  income_less20k
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.4139
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Using histograms:

```
ggplot(poverty, aes(x = cash, na.rm = T)) +
  geom_histogram(color = 'darkgray', fill = 'orange', binwidth = 100) +
  scale_x_continuous(limits = c(0, 1000), breaks = seq(0, 1000, 100)) +
  scale_y_continuous(limits = c(0, 600), breaks = seq(0, 600, 150)) +
  labs(x = 'Cash on hand',
       y = 'Count',
       title = 'Histogram of cash') +
  geom_vline(aes(xintercept = mean(cash, na.rm = T, color = 'mean')))
```

```
## Warning: Removed 296 rows containing non-finite values (stat_bin).
```

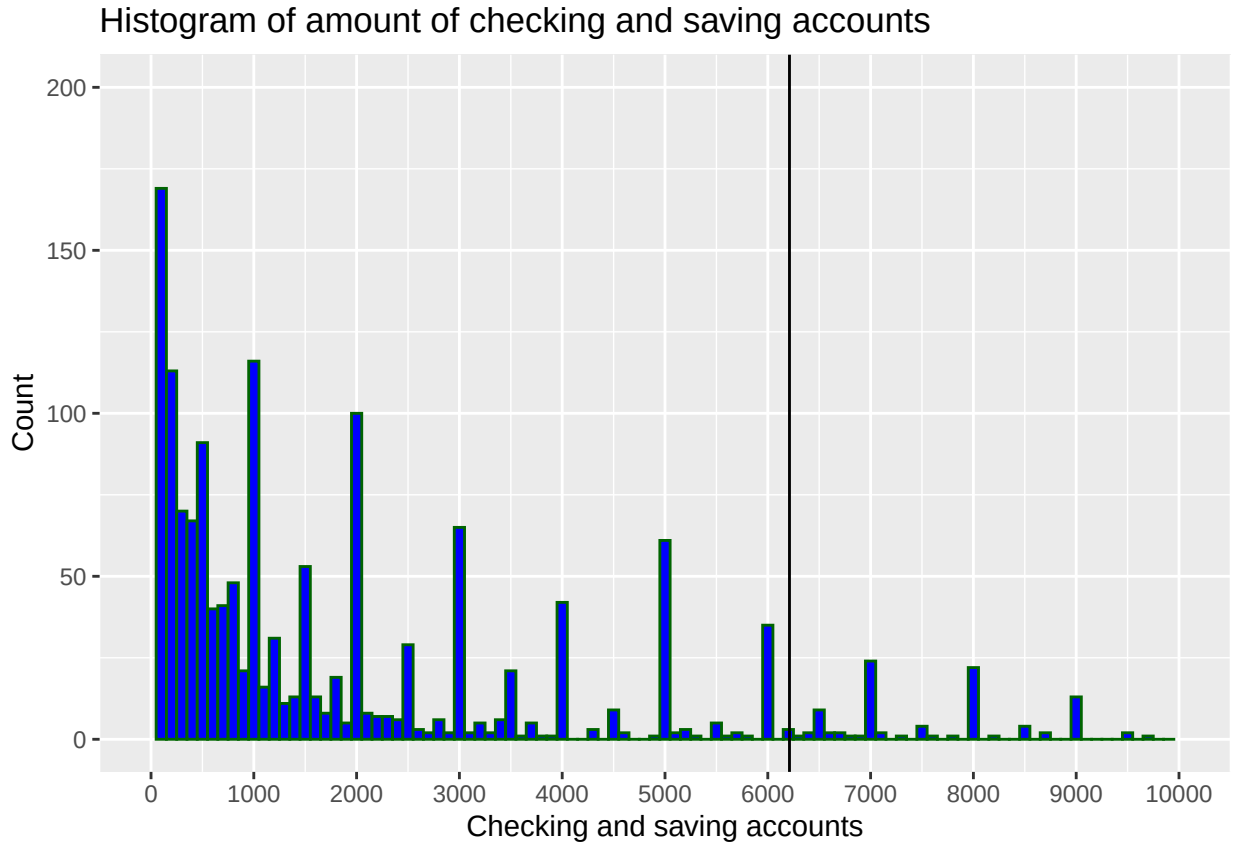
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



```
ggplot(poverty, aes(x = accts_amt, na.rm = T)) +
  geom_histogram(color = 'darkgreen', fill = 'blue', binwidth = 100) +
  scale_x_continuous(limits = c(0, 10000), breaks = seq(0, 10000, 1000)) +
  scale_y_continuous(limits = c(0, 200), breaks = seq(0, 200, 50)) +
  labs(x = 'Checking and saving accounts',
       y = 'Count',
       title = 'Histogram of amount of checking and saving accounts') +
  geom_vline(aes(xintercept = mean(accts_amt, na.rm = T, color = 'mean')))
```

```
## Warning: Removed 754 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



As we can see, people had about 170 dollars on hand and the majority had less than 1000 dollars in the checking and saving accounts.

Now, take the *natural logarithm* of these two variables and plot the histograms of these tranformed variables. How does the distribution look now? What are the advantages and disadvantages of transforming the data in this way?

NOTE: Since the natural logarithm of 0 is undefined, researchers often add a small value (in this case, we will use \$1 so that $\log 1 = 0$) to the 0 values for the variables being transformed (in this case, `cash` and `accts_amt`) in order to successfully apply the `log()` function to all values. Be sure to do this recoding only for the purposes of taking the logarithmic transformation – keep the original variables the same.

poverty %>% mutate(cash = ifelse(poverty$cash == 0, "1", $ifelse($poverty accts_amt == 0, "1",)))

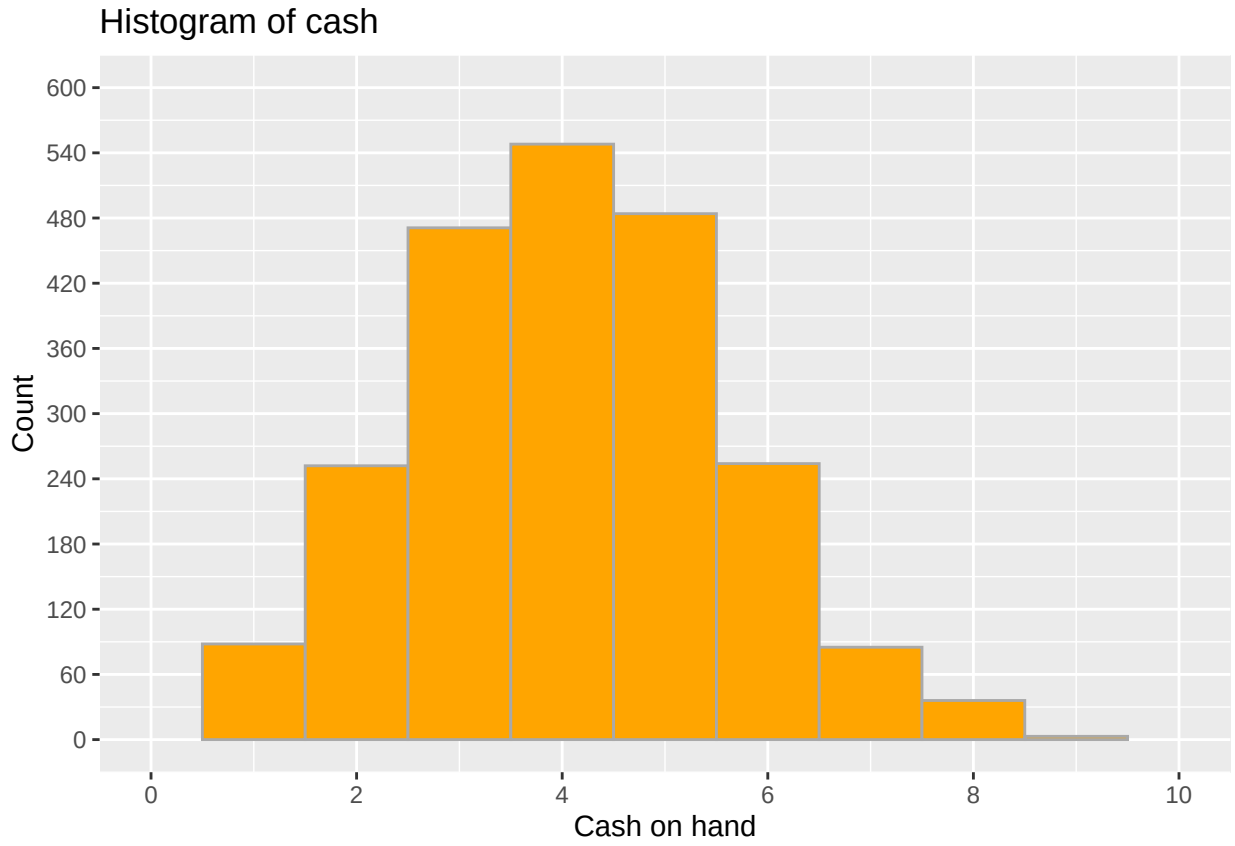
New histograms:

```r
poverty_log <- poverty %>%
  mutate(cash = log(poverty$cash),
         accts_amt = log(poverty$accts_amt))

ggplot(poverty_log, aes(x = cash, na.rm = T)) +
  geom_histogram(color = 'darkgray', fill = 'orange', binwidth = 1) +
  scale_x_continuous(limits = c(0, 10), breaks = seq(0, 10, 2)) +
  scale_y_continuous(limits = c(0, 600), breaks = seq(0, 600, 60)) +
  labs(x = 'Cash on hand',
       y = 'Count',
       title = 'Histogram of cash')
```

```
## Warning: Removed 411 rows containing non-finite values (stat_bin).
```

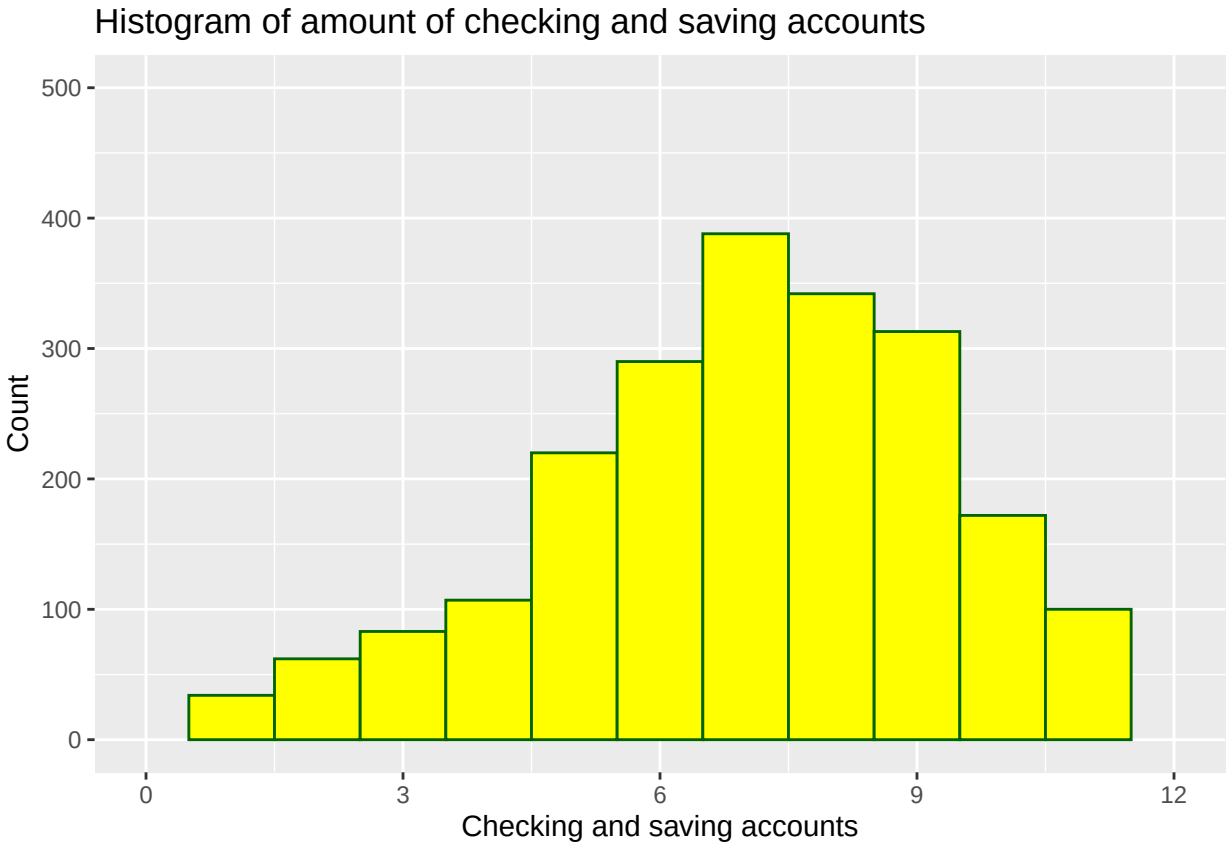
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```r
ggplot(poverty_log, aes(x = accts_amt, na.rm = T)) +
  geom_histogram(color = 'darkgreen', fill = 'yellow', binwidth = 1) +
  scale_x_continuous(limits = c(0, 12), breaks = seq(0, 12, 3)) +
  scale_y_continuous(limits = c(0, 500), breaks = seq(0, 500, 100)) +
  labs(x = 'Checking and saving accounts',
       y = 'Count',
       title = 'Histogram of amount of checking and saving accounts')
```

```
## Warning: Removed 534 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



As we can see, the histograms from logged numbers are better to visualize the distributions, but they are not so precise as the previous ones.

Question 2

Now, let's examine the primary outcome of interest for this study– the effect of a change in financial situation (in this case, getting paid on payday) on economic decision-making and cognitive performance. Begin by calculating the treatment effect for the `stroop_time` variable (a log-transformed variable of the average response time for the stroop cognitive test), using first the mean and then the median. What does this tell you about differences in the outcome across the two experimental conditions?

Calculating the treatment effect for the `stroop_time` variable:

```
bpd <- select(poverty,
                      treatment, stroop_time) %>%
  filter(treatment == 'Before Payday') %>%
  mutate(stroop_time = log(stroop_time))

apd <- select(poverty,
                      treatment, stroop_time, cash, accts_amt) %>%
  filter(treatment == 'After Payday') %>%
  mutate(stroop_time = log(stroop_time))
```

```r
mean(bpd$stroop_time)
```

```
## [1] 2.019319
```

```r
mean(apd$stroop_time)
```

```
## [1] 2.020914
```

```r
median(bpd$stroop_time)
```

```
## [1] 2.022446
```

```r
median(apd$stroop_time)
```

```
## [1] 2.024337
```

As we can see, the mean for 'Before Payday' is lower than for 'After Payday', as well as the median.

Secondly, let's look at the relationship between finanical circumstances and the cognitive test variable. Produce two scatter plots side by side (hint: use the par(mfrow)) before your plot commands to place graphs side-by-side), one for each of the two experimental conditions, showing the bivariate relationship between your *log-transformed* `cash` variable and the amount of time it took subjects to complete the stroop cognitive test administered in the survey (`stroop_time`). Place the `stroop_time` variable on the y-axis. Be sure to title your graphs to differentiate between the `Before Payday` and `After Payday` conditions. Now do the same, for the *log-transformed* `accts_amt` variable.

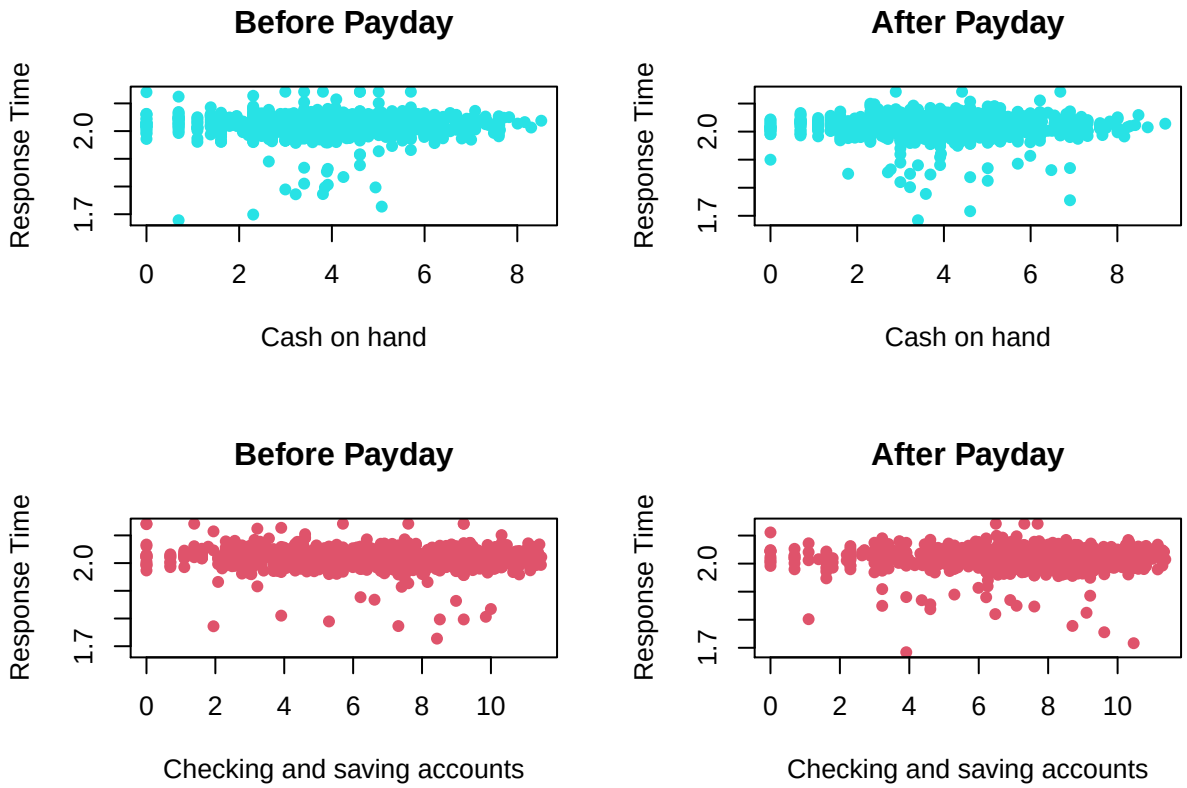
The graphs are the following ones:

```r
poverty_bpd <- select(poverty,
                      treatment, stroop_time, cash, accts_amt) %>%
  filter(treatment == 'Before Payday') %>%
  mutate(stroop_time = log(stroop_time),
         cash = log(cash),
         accts_amt = log(accts_amt))

poverty_apd <- select(poverty,
                      treatment, stroop_time, cash, accts_amt) %>%
  filter(treatment == 'After Payday') %>%
  mutate(stroop_time = log(stroop_time),
         cash = log(cash),
         accts_amt = log(accts_amt))

par(mfrow = c(2,2))
plot(poverty_bpd$cash, poverty_bpd$stroop_time,
     pch = 16, col = 5, cex.lab = 1.02,
     ylab = "Response Time",
     xlab = "Cash on hand",
     main = "Before Payday")
plot(poverty_apd$cash, poverty_apd$stroop_time,
     pch = 16, col = 5, cex.lab = 1.02,
     ylab = "Response Time",
```

```
     xlab = "Cash on hand",
     main = "After Payday")
plot(poverty_bpd$accts_amt, poverty_bpd$stroop_time,
     pch = 16, col = 2, cex.lab = 1.02,
     ylab = "Response Time",
     xlab = "Checking and saving accounts",
     main = "Before Payday")
plot(poverty_apd$accts_amt, poverty_apd$stroop_time,
     pch = 16, col = 2, cex.lab = 1.02,
     ylab = "Response Time",
     xlab = "Checking and saving accounts",
     main = "After Payday")
```



Briefly comment on your results in light of the hypothesis that changes in economic circumstances will influence cognitive performance.

The results look quite similar to me, so I do not think there is a conclusion from this plots besides that the payday does not change a cognitive reaction.

Question 3

Now, let's take a closer look at whether or not the `Before Payday` versus `After Payday` treatment created measurable differences in financial circumstances. What is the effect of payday on participants' financial resources? To help with interpretability, use the original variables `cash` and `accts_amt` to calculate this effect. Calculate both the mean and median effect. Does the measure of central tendency you use affect your perception of the effect?

Question 4

Compare the distributions of the `Before Payday` and `After Payday` groups for the *log-transformed* `cash` and `accts_amt` variables. Use quantile-quantile plots to do this comparison, and add a 45-degree line in a color of your choice (not black). Briefly interpret your results and their implications for the authors' argument that their study generated variation in financial resources before and after payday. When appropriate, state which ranges of the outcome variables you would focus on when comparing decision-making and cognitive capacity across these two treatment conditions.

Question 5

In class (first slides, before corona break), we covered the difference-in-difference design for comparing average treatment effects across treatment and control groups. This design can also be used to compare average treatment effects across different ranges of a *pre-treatment variable*- a variable that asks about people's circumstances before the treatment and thus could not be affected by the treatment. This is known as *heterogeneous treatment effects* – the idea that the treatment may have differential effects for different subpopulations. Let's look at the pre-treatment variable `income_less20k`. Calculate the treatment effect of Payday on amount in checking and savings accounts separately for respondents earning more than 20,000 dollars a year and those earning less than 20,000 dollars. Use the original `accts_amt` variable for this calculation. Then take the difference between the effects you calculate. What does this comparison tell you about how payday affects the amount that people have in their accounts? Are you convinced by the authors' main finding from Question 2 in light of your investigation of their success in manipulating cash and account balances before and after payday?

Group 10 - Election and Conditional Cash Transfer Program in Mexico

In this exercise, we analyze the data from a study that seeks to estimate the electoral impact of 'Progresa', Mexico's *conditional cash transfer program* (CCT program).

This exercise is based on the following article: Ana de la O. (2013). 'Do Conditional Cash Transfers Affect Voting Behavior? Evidence from a Randomized Experiment in Mexico.' *American Journal of Political Science*, 57:1, pp.1-14. and Kosuke Imai, Gary King, and Carlos Velasco. (2015). 'Do Nonpartisan Programmatic Policies Have Partisan Electoral Effects? Evidence from Two Large Scale Randomized Experiments.' Working Paper.

The original study relied on a randomized evaluation of the CCT program in which eligible villages were randomly assigned to receive the program either 21 (Early *Progresa*) or 6 months (Late *Progresa*) before the 2000 Mexican presidential election. The author of the original study hypothesized that the CCT program would mobilize voters, leading to an increase in turnout and support for the incumbent party (PRI in this case). The analysis was based on a sample of precincts that contain at most one participating village in the evaluation.

The data we analyze are available as the CSV file `progresa.csv`. The names and descriptions of variables in the data set are:

Name	Description
`treatment`	Whether an electoral precinct contains a village where households received Early *Progresa*
`pri2000s`	PRI votes in the 2000 election as a share of precinct population above 18
`pri2000v`	Official PRI vote share in the 2000 election

Name	Description
t2000	Turnout in the 2000 election as a share of precinct population above 18
t2000r	Official turnout in the 2000 election
pri1994	Total PRI votes in the 1994 presidential election
pan1994	Total PAN votes in the 1994 presidential election
prd1994	Total PRD votes in the 1994 presidential election
pri1994s	Total PRI votes in the 1994 election as a share of precinct population above 18
pan1994s	Total PAN votes in the 1994 election as a share of precinct population above 18
prd1994s	Total PRD votes in the 1994 election as a share of precinct population above 18
pri1994v	Official PRI vote share in the 1994 election
pan1994v	Official PAN vote share in the 1994 election
prd1994v	Official PRD vote share in the 1994 election
t1994	Turnout in the 1994 election as a share of precinct population above 18
t1994r	Official turnout in the 1994 election
votos1994	Total votes cast in the 1994 presidential election
avgpoverty	Precinct Avg of Village Poverty Index
pobtot1994	Total Population in the precinct
villages	Number of villages in the precinct

Each observation in the data represents a precinct, and for each precinct the file contains information about its treatment status, the outcomes of interest, socioeconomic indicators, and other precinct characteristics.

Question 1

Estimate the impact of the CCT program on turnout and support for the incumbent party (PRI or Partido Revolucionario Institucional) by comparing the average electoral outcomes in the 'treated' (Early *Progresa*) precincts versus the ones observed in 'control' (Late *Progresa*) precincts. Next, estimate these effects by regressing the outcome variable on the treatment variable. Interpret and compare the estimates under these approaches. Here, following the original analysis, use the turnout and support rates as shares of the voting eligible population (`t2000` and `pri2000s`, respectively). Do the results support the hypothesis? Provide a brief interpretation.

Question 2

In the original analysis, the authors fit a linear regression model that includes, as predictors a set of pre-treatment covariates as well as the treatment variable. Here, we fit a similar model for each outcome that includes the average poverty level in a precinct (`avgpoverty`), the total precinct population in 1994 (`pobtot1994`), the total number of voters who turned out in the previous election (`votos1994`), and the total number of votes cast for each of the three main competing parties in the previous election (`pri1994` for PRI, `pan1994` for Partido Acción Nacional or PAN, and `prd1994` for Partido de la Revolución Democrática or PRD). Use the same outcome variables as in the original analysis that are based on the shares of the voting age population. According to this model, what are the estimated average effects of the program's availability on turnout and support for the incumbent party? Are these results different from what you obtained in the previous question?

Question 3

Next, we consider an alternative, and more natural, model specification. We will use the original outcome variables as in the previous question. However, our model should include the previous election outcome variables measured as shares of the voting age population (as done for the outcome variables `t1994`, `pri1994s`, `pan1994s`, and `prd1994s`) instead of those measured in counts. In addition, we apply the natural logarithm transformation to the precinct population variable when including it as a predictor. As in the original model, our model includes the average poverty index as an additional predictor. Are the results based on these new model specifications different from what we obtained in the previous question? If the results are different, which model fits the data better?

Question 4

We examine the balance of some pre-treatment variables used in the previous analyses. Using boxplots, compare the distributions of the precinct population (on the original scale), average poverty index, previous turnout rate (as a share of the voting age population), and previous PRI support rate (as a share of the voting age population) between the treatment and control groups. Comment on the patterns you observe.

Question 5

We next use the official turnout rate `t2000r` (as a share of the registered voters) as the outcome variable rather than the turnout rate used in the original analysis (as a share of the voting age population). Similarly, we use the official PRI's vote share `pri2000v` (as a share of all votes cast) rather than the PRI's support rate (as a share of the voting age population). Compute the average treatment effect of the CCT program using a linear regression with the average poverty index, the log-transformed precinct population, and the previous official election outcome variables (`t1994r` for the previous turnout; `pri1994v`, `pan1994v`, and `prd1994v` for the previous PRI, PAN, and PRD vote shares). Briefly interpret the results.

Group 11 - Predicting Race Using Demographic Information

In this exercise group, we investigate the problem of predicting the ethnicity of individual voters given their surname and residence location using Bayes' rule. This exercise is based on the following article: Kosuke Imai and Kabir Khanna. (2016). "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24(2): 263-272.

In this exercise, we attempt to improve that prediction by taking into account demographic information such as age and gender. As done earlier, we validate our method by comparing our predictions with the actual race of each voter.

Name	Description
`county`	County census id of voting district.
`VTD`	Voting district census id (only unique within county)
`total.pop`	Total population of voting district

Other variables are labeled in three parts, each separated by a period. See below for each part. Each column contains the proportion of people of that gender, age group, and race in the voting district.

Name	Description
gender	Male or female
age groups	Age groups as defined by U.S. Census (see table below)
race	Different racial categories (see table below)

Below is the table for variables describing racial categories:

Name	Description
whi	non-Hispanic whites in the voting district
bla	non-Hispanic blacks in the district
his	Hispanics
asi	non-Hispanic Asian and Pacific Islanders
oth	other racial categories
mix	non-Hispanic people of two or more races.

Below is the table for age-group variables, as defined by the U.S. Census:

Name	Description
1	18–19
2	20–24
3	25–29
4	30–34
5	35–39
6	40–44
7	45–49
8	50–54
9	55–59
10	60–64
11	65–69
12	70–74
13	75–79
14	80–84
15	85+

We use three data sets in this exercises, two of which were already introduced in Section 6.1. The first data set is a random sample of 10,000 registered voters contained in the csv file, `FLVoters.csv`. Table 6.1 presents the names and descriptions of variables for this data set. The second data set is a csv file, `cnames.csv`, containing a modified version of the original data set, `names.csv`, after making appropriate adjustments about a special value as done in Section 6.2. Table 6.3 presents the names and descriptions of variables in this data set. Finally, the third data set, `FLCensusDem`, contains the updated census data with two additional demographic variables – gender and age. Unlike the other census data we analyzed earlier, each observation of this data set consists of one voting district and the proportion of each demographic by age, gender, and race within that district. The tables above present the names and descriptions of variables in this data set of Florida districts. There is also a table that contains the age groupings used in the variable names of the `FLCensusDem.csv` file.

Question 1

Use Bayes' Rule to find a formula for the probability that a voter belongs to a given racial group conditional on their age, gender, surname, and residence location. Given the data sets we have, can we use this formula to predict each voter's race? If the answer is yes, briefly explain how you would make the prediction. If the answer is no, explain why you cannot apply the formula you derived.

Question 2

Assume that, given the person's race, the surname is conditionally independent from residence, age, and gender. Express this assumption mathematically and also substantively interpret. Show that under this assumption, the probability that a voter belongs to a given racial group conditional on their age and gender as well as their surname and residence location is given by the following formula.

$$\frac{P(\text{residence}, \text{age}, \text{gender} \mid \text{race}) P(\text{race} \mid \text{surname})}{P(\text{residence}, \text{age}, \text{gender} \mid \text{surname})}$$

Question 3

Using the formula derived in the previous question, we wish to compute the predicted probability that a voter belongs to a given racial group, conditional on their age and gender as well as their surname and residence location. Provide a step-by-step explanation of how to do this computation using the data. Hint: you will need to modify the formula without invoking an additional assumption such that all quantities can be computed from the data sets we have. The definition of conditional probability and the law of total probability might be useful.

Question 4

Use the procedure described in the previous question, compute the predicted probability for each voter in the `FLVoters.csv` that the voter belongs to a given racial group conditional on their age, gender, surname and residence location. Exclude the voters with missing data from your analysis. Also, note that the csv file `cnames.csv` has been processed from `names.csv` using the code from Section 6.1. Thus, there is no need to re-adjust the values to account for negligibly small race percentages, but the racial proportions by surname are initially expressed as percentages rather than as decimals.

Question 5

Given the results in the previous question, identify the most likely race for each individual in `FLVoters.csv`, given their surname, residence, age, and gender.

Group 12 - The Mathematics of Enigma

The Enigma machine is the most famous cipher machine to date. Nazi Germany used it during World War II to encrypt messages so that enemies could not understand them. The story of the British cryptanalysts who successfully deciphered Enigma has become the subject of multiple movies *Enigma* (2001); *The Imitation*

Game (2014). In this exercise, we will focus our attention on a simplified version of the Enigma machine, which we name "Little Enigma." Like the real Enigma machine shown in the picture above, this machine consists of two key components. First, the Little Enigma machine has 5 different *rotors*, each of which comes with 10 pins with numbers ranging from 0 to 9. Second, a component called the *plugboard* contains 26 holes, corresponding to the 26 letters of the alphabet. In addition, 13 cables connect all possible pairs of letters. Since a cable has two ends, one can connect, for example, the letter A with any other of the other 25 letters present in the plugboard.

To either encode a message or decode an encrypted message, one must provide the Little Enigma machine with a correct five-digit passcode to align the rotors and a correct configuration of the plugboard. The rotors are set up just like many combination locks. For example, the passcode 9–4–2–4–9 means that five rotors display the numbers 9, 4, 2, 4, and 9 in that order. In addition, the 13 cables connecting the letters in the plugboard must be appropriately configured. The purpose of the plugboard is thus to scramble the letters. For example, if B is connected to W, the Little Enigma machine will switch B with W and W with B to encode a message or decode an encoded message. Thus, a sender types a message on the keyboard, the plugboard scrambles the letters, and the message is sent in its encrypted form. A receiver decodes the encrypted message by re-typing it on a paired Little Enigma machine that has the same passcode and plugboard configuration.

Question 1

How many different five-digit passcodes can be set out of the 5 rotors?

Since there are 10 different and possible numbers for each 5 rotors, there are 100000 different combinations.

Question 2

How many possible configurations does the plugboard provide? In other words, how many ways can 26 letters be divided into 13 pairs?

From the text above, we can organize mathematically as

$$\frac{26!}{2^{13}13!}$$

There are thus 7905853580625 possible configurations.

Question 3

Based on the previous two questions, what is the total number of possible settings for the Little Enigma machine?

The total number of possible settings for the Little Enigma machine is

$$7905853580625.10^5$$

Question 4

Five cryptanalytic machines have been developed to decode 1,500 messages encrypted by the Little Enigma machine. The table below presents information on the number of messages assigned to each machine and the machine's failure rate (i.e., the percentage of messages the machine was unable to decode). Aside from this information, we do not know anything about the assignment of each message to a machine or whether the machine was able to correctly decode the message.

Machine	Number of messages	Failure Rate
`Banburismus`	300	10%
`Bombe`	400	5%
`Herivel`	250	15%
`Crib`	340	17%
`Hut 6`	210	20%

Suppose that we select one message at random from the pool of all 1,500 messages but found out this message was not properly decoded. Which machine is most likely responsible for this mistake?

Since Crib generates more errors (almost 58) than the others, it is the most likely responsible for this mistake.

Question 5

Write an R function that randomly configures the plugboard. This function will take no input but randomly selects a set of 13 pairs of letters. The output object should be a 2×13 matrix for which each column represents a pair of letters. You may use the built-in R object `letters`, which contains the 26 letters of the alphabet as a character vector. Name the function `plugboard`.

Then, write an R function that encodes and decodes a message given a plugboard configuration set by the `plugboard` function from the previous question. This function should take the output of the `plugboard` function as well as a message to be encoded (decoded) as inputs, and return an encoded (decoded) message. You may wish to use the `gsub` function, which replaces a pattern in a character string with another specified pattern. The `tolower` function, which makes characters in a character vector lowercase, and `toupper` function, which capitalizes characters in a character vector, can also help.

Group 13 - Immigration attitudes: the role of economic and cultural threat

Why do the majority of voters in the U.S. and other developed countries oppose increased immigration? According to the conventional wisdom and many economic theories, people simply do not want to face additional competition on the labor market (*economic threat* hypothesis). Nonetheless, most comprehensive empirical tests have failed to confirm this hypothesis and it appears that people often support policies that are against their personal economic interest. At the same time, there has been growing evidence that immigration attitudes are rather influenced by various deep-rooted ethnic and cultural stereotypes (*cultural threat* hypothesis). Given the prominence of workers' economic concerns in the political discourse, how can these findings be reconciled?

This exercise is based in part on Malhotra, N., Margalit, Y. and Mo, C.H., 2013. "Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact." *American Journal of Political Science*, Vol. 38, No. 3, pp. 393-433.

The authors argue that, while job competition is not a prevalent threat and therefore may not be detected by aggregating survey responses, its *conditional* impact in selected industries may be quite sizable. To test their hypothesis, they conduct a unique survey of Americans' attitudes toward H-1B visas. The plurality of H-1B visas are occupied by Indian immigrants, who are skilled but ethnically distinct, which enables the authors to measure a specific skill set (high technology) that is threatened by a particular type of immigrant (H-1B visa holders). The data set `immig.csv` has the following variables:

Name	Description
`age`	Age (in years)
`female`	`1` indicates female; `0` indicates male
`employed`	`1` indicates employed; `0` indicates unemployed
`nontech.whitcol`	`1` indicates non-tech white-collar work (e.g., law)
`tech.whitcol`	`1` indicates high-technology work
`expl.prejud`	Explicit negative stereotypes about Indians (continuous scale, 0-1)
`impl.prejud`	Implicit bias against Indian Americans (continuous scale, 0-1)
`h1bvis.supp`	Support for increasing H-1B visas (5-point scale, 0-1)
`indimm.supp`	Support for increasing Indian immigration (5-point scale, 0-1)

The main outcome of interest (`h1bvis.supp`) was measured as a following survey item: "Some people have proposed that the U.S. government should increase the number of H-1B visas, which are allowances for U.S. companies to hire workers from foreign countries to work in highly skilled occupations (such as engineering, computer programming, and high-technology). Do you think the U.S. should increase, decrease, or keep about the same number of H-1B visas?" Another outcome (`indimm.supp`) similarly asked about the "the number of immigrants from India." Both variables have the following response options: `0` = "decrease a great deal", `0.25` = "decrease a little", `0.5` = "keep about the same", `0.75` = "increase a little", `1` = "increase a great deal".

To measure explicit stereotypes (`expl.prejud`), respondents were asked to evaluate Indians on a series of traits: capable, polite, hardworking, hygienic, and trustworthy. All responses were then used to create a scale lying between `0` (only positive traits of Indians) to `1` (no positive traits of Indians). Implicit bias (`impl.prejud`) is measured via the *Implicit Association Test* (IAT) which is an experimental method designed to gauge the strength of associations linking social categories (e.g., European vs Indian American) to evaluative anchors (e.g., good vs bad). Individual who are prejudiced against Indians should be quicker at making classifications of faces and words when *European American* (*Indian American*) is paired with *good* (*bad*) than when *European American* (*Indian American*) is paired with *bad* (*good*). If you want, you can test yourself here.

Question 1

Start by examining the distribution of immigration attitudes (as factor variables). What is the proportion of people who are willing to increase the quota for high-skilled foreign professionals (`h1bvis.supp`) or support immigration from India (`indimm.supp`)?

Now compare the distribution of two distinct measures of cultural threat: explicit stereotyping about Indians (`expl.prejud`) and implicit bias against Indian Americans (`impl.prejud`). In particular, create a scatterplot, add a linear regression line to it, and calculate the correlation coefficient. Based on these results, what can you say about their relationship?

Question 2

Compute the correlations between all four policy attitude and cultural threat measures. Do you agree that cultural threat is an important predictor of immigration attitudes as claimed in the literature?

If the labor market hypothesis is correct, opposition to H-1B visas should also be more pronounced among those who are economically threatened by this policy such as individuals in the high-technology sector. At the same time, tech workers should not be more or less opposed to general Indian immigration because of any *economic* considerations. First, regress H-1B and Indian immigration attitudes on the indicator variable for tech workers (`tech.whitcol`). Do the results support the hypothesis? Is the relationship different from the one involving cultural threat and, if so, how?

Question 3

When examining hypotheses, it is always important to have an appropriate comparison group. One may argue that comparing tech workers to everybody else as we did in Question 2 may be problematic due to a variety of confounding variables (such as skill level and employment status). First, create a single factor variable `group` which takes a value of `tech` if someone is employed in tech, `whitecollar` if someone is employed in other "white-collar" jobs (such as law or finance), `other` if someone is employed in any other sector, and `unemployed` if someone is unemployed. Then, compare the support for H-1B across these conditions by using the linear regression. Interpret the results: is this comparison more or less supportive of the labor market hypothesis than the one in Question 2?

Now, one may also argue that those who work in the tech sector are disproportionately young and male which may confound our results. To account for this possibility, fit another linear regression but also include `age` and `female` as pre-treatment covariates (in addition to `group`). Does it change the results and, if so, how?

Finally, fit a linear regression model with all threat indicators (`group`, `expl.prejud`, `impl.prejud`) and calculate its R^2. How much of the variation is explained? Based on the model fit, what can you conclude about the role of threat factors?

Question 4

Besides economic and cultural threat, many scholars also argue that gender is an important predictor of immigration attitudes. While there is some evidence that women are slightly less opposed to immigration than men, it may also be true that gender conditions the very effect of other factors such as cultural threat. To see if it is indeed the case, fit a linear regression of H-1B support on the interaction between gender and implicit prejudice. Then, create a plot with the predicted level of H-1B support (y-axis) across the range of implicit bias (x-axis) by gender. Considering the results, would you agree that gender alters the relationship between cultural threat and immigration attitudes?

Age is another important covariate. Fit two regression models in which H-1B support is either a linear or quadratic function of age. Compare the results by plotting the predicted levels of support (y-axis) across the whole age range (x-axis). Would you say that people become more opposed to immigration with age?

Question 5

To corroborate your conclusions with regard to cultural threat, create separate binary variables for both prejudice indicators based on their median value (`1` if `>` than the median) and then compare average H-1B and Indian immigration attitudes (as numeric variables) depending on whether someone is implicitly or explicitly prejudiced (or both). What do these comparisons say about the role of cultural threat?

What about the role of economic threat? One may argue that tech workers are simply more or less prejudiced against Indians than others. To account for this possibility, investigate whether economic threat is in fact distinguishable from cultural threat as defined in the study. In particular, compare the distribution of cultural threat indicator variable using the Q-Q plot depending on whether someone is in the high-technology sector. Would you conclude that cultural and economic threat are really distinct?

Group 14 - Co-ethnic Candidates and Voter Turnout

For these problems, we will analyze data from the following article:

Fraga, Bernard. (2015) "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout," *American Journal of Political Science*, Vol. 60, No. 1, pp. 97–122.

Fraga assesses the theory that minority voters are more likely to vote in elections featuring co-ethnic candidates. He shows that the association between minority voter turnout and co-ethnic candidates disappears once we take into account district-level racial composition. In particular, he demonstrates that in districts where blacks make up a greater share of the voting-age population, blacks in that district are more likely to vote in elections *regardless* of candidate race.

A description of the variables is listed below:

Name	Description
`year`	Year the election was held
`state`	State in which the election was held
`district`	District in which the election was held (unique within state but not across states)
`turnout`	The proportion of the black voting-age population in a district that votes in the general election
`CVAP`	The proportion of a district's voting-age population that is black
`candidate`	Binary variable coded "1" when the election includes a black candidate; "0" when the election does not include a black candidate

Question 1

Fraga analyzes turnout data for four different racial and ethnic groups, but for this analysis we will focus on the data for black voters. Load `blackturnout.csv`. Which years are included in the dataset? How many different states are included in the dataset?

Question 2

Create a boxplot that compares turnout in elections with and without a co-ethnic candidate. Be sure to use informative labels. Interpret the resulting graph.

Question 3

Run a linear regression with black turnout as your outcome variable and candidate co-ethnicity as your predictor. Report the coefficient on your predictor and the intercept. Interpret these coefficients. Do not merely comment on the direction of the association (i.e., whether the slope is positive or negative). Explain what the value of the coefficients mean in terms of the units in which each variable is measured. Based on these coefficients, what would you conclude about blacks voter turnout and co-ethnic candidates?

Question 4

You decide to investigate the results of the previous question a bit more carefully because the elections with co-ethnic candidates may differ from the elections without co-ethnic candidates in other ways. Create a scatter plot where the x-axis is the proportion of co-ethnic voting-age population and the y-axis is black voter turnout. Color your points according to candidate co-ethnicity. That is, make the points for elections featuring co-ethnic candidates one color, and make the points for elections featuring no co-ethnic candidates a different color. Interpret the graph.

Question 5

Run a linear regression with black turnout as your outcome variable and with candidate co-ethnicity and co-ethnic voting-age population as your predictors. Report the coefficients, including the intercept. Interpret the coefficients on the two predictors. Explain what each coefficient represents in terms of the units of the relevant variables.

Group 15 - Durably Reducing Transphobia

A recent paper by Broockman and Kalla asked if individual minds could be changed with respect to contentious political topics, particularly transgender rights. Broockman and Kalla developed a field experiment where individuals were randomly assigned to receive different treatments and their attitudes toward transgender rights were measured before and after the treatment. The paper that describes their findings will be the basis of this excercise:

Broockman, David, and Joshua Kalla (2016). "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science* 352(6282): 220-224.

Broockman and Kalla were focused on whether perspective-taking – "imagining the world from another's vantage point" – might be particularly effective in reducing intergroup prejudice. Therefore, participants in the treatment group were visited by a canvasser—some of whom were transgender and some of whom were not—and asked to think about a time when they were judged unfairly and were guided to translate that experience to a transgender individual's experience. A placebo group had conversation with canvassers about recycling.

The data you will use, `broockman_kalla_2016.csv`, contains all the variables you will need. The names and descriptions of variables are shown below. You may not need to use all of these variables for this activity. We've kept these unnecessary variables in the dataset because it is common to receive a dataset with much more information than you need. Some of the items appeared on multiple surveys; the `#` sign below will be replaced with the survey number in your analysis:

- the baseline (pre-treatment) survey is survey 0;
- the 3-day survey is survey 1;
- the 3-week survey is survey 2;
- the 6-week survey is survey 3, and;
- the 3-month survey is survey 4.

Name	Description
`vf_age`	Voter file age
`vf_racename`	Voter file race
`vf_female`	`1` if the voter is female, `0` if the voter is male.
`vf_democrat`	`1` if the voter is a Democrat, `0` if the voter is not a Democrat.
`respondent_t#`	`1` if voter responded to the survey at wave 0, 1, 2, 3, or 4 of the study, `0` if voter does not.
`miami_trans_law_t#`	Support or opposition to antidiscrimination law (measured from `-3` (opposed) to `3` (in support))
`therm_trans_t#`	Feeling thermometer towards trans people (0-100).
`treat_ind`	`1` if individual was assigned to treatment, `0` if assigned to placebo.
`exp_actual_convo`	`1` if treatment was actually delivered, `0` if it was not.
`canvasser_trans`	`1` if canvasser identified as transgender or gender non-conforming, `0` if canvasser did not.
`canvasser_id`	Canvasser identifier.
`hh_id`	Household identifier.

Question 1

First, load `broockman_kalla_2016.csv` into R and explore the data. How many observations are there? This file includes all of the people that the researchers *attempted* to contact. As we've learned, experiments also experience this noncompliance. However, many of these people did not even respond to the baseline survey before the experiment was conducted. As such, we are not be able to estimate how the treatment changed these individuals' attitudes, since we don't know where they were to begin with. To deal with these people who were never reached, subset your data so that you are only looking at people who have a valid (non-NA) answer for `therm_trans_t0`, which was collected in the pre-treatment baseline study.

How many observations are you left with? How many are assigned to treatment? Placebo? What do these numbers tell us about the randomization process? You should use this subset for the rest of this questions in this part of the exam.

Question 2

To examine the effect of these conversations, Broockman and Kalla compared how people in the treatment and placebo groups evaluated transgender people on a specific type of survey question called a "feeling thermometer," where respondents indicate where their feelings fall on a scale of 0 (very cold) to 100 (very warm). Further, to measure whether the changes persisted, Broockman and Kalla also conducted follow-up surveys up to 3 months after the original treatment. Thus, Brookman and Kalla's study required reaching participants in multiple ways: both for a face-to-face interaction where the treatment would be delivered and through multiple follow-up surveys where the long-term effects of the treatment could be measured.

We would like to further check that noncompliance did not meaningfully affect our randomization, since the basis of calculating our treatment effect depends on this. To do this, we can conduct a *placebo test* to ensure that the average level of support (our outcome) is not meaningfully different between the treatment and control groups *before* treatment was administered (during the baseline survey). Conduct a t-test on support before treatment was administered ($t = 0$) between the treatment and control groups among the subset of people you calculated in question 1. Construct a 95% confidence interval around your estimate. If the null hypothesis is no difference between the two groups, can you reject the null hypothesis?

Question 3

To simplify our analysis, we will focus on the average treatment effects among the treated (ATT). Kalla and Broockman experienced noncompliance in their study, where individuals assigned to the treatment group refused to engage in conversations with canvassers. Because accounting for this noncompliance statistically is beyond the scope of this class, we will look only at the cases where treatment was actually delivered. In contrast, Broockman and Kalla estimate the average treatment effect of compliers (adjusting for compliance rates). They also include many covariates; we will only include a few. For these reasons and others, our estimates may differ slightly.

Next, let's see if the attitudes in the treatment and placebo changed after the treatment was administered and for how long these differences lasted. Perform four difference-in-means calculations—one for each wave ($t = 1, 2, 3, 4$). When performing each test, remove any NAs in that wave. (These NAs might arise from participants failing to complete different waves of the surveys—a process that social scientists call "attrition"—or for other reasons.) For example, for wave 1, use all cases that don't have missing outcome data in wave 1, and in wave 2, use all cases that don't have missing outcome data in wave 2 (in the `t.test()` function, you can accomplish this with the `na.action = na.omit` argument). Construct 95% confidence intervals for your estimates. At each stage, do we reject or retain the null hypothesis that no difference between the two groups exists?

Question 4

Let's focus on the difference-in-means test at time $t = 1$. If the null hypothesis is that there is no difference between the groups, what would it mean to make a type I error? What would it mean to make a type II error? If we did something to increase the statistical power of the study, would we increase or decrease the probability of a type II error? Hint: if yoy need a refresh on type I and type II errors, check the following wikipedia page.

Question 5

Finally, let's approximate the estimation strategy used by Broockman and Kalla in their analysis. To estimate the average treatment effect, they use a regression framework. Regress the feeling thermometer dependent variable (measure at time $t = 3$) on treatment assignment. To further alleviate concerns of imbalances between treatment and control groups, adjust for an individual's feeling thermometer scores at time $t = 0$, her age, gender, race, and political party. Further, to account for possible differences between the more than 50 canvassers, please include in your model a fixed effect (i.e., dummy variable) for each canvasser. What is the estimated treatment effect (be sure to mention units)? Is this estimate statistically significant at the 0.05 level? Conduct the same analysis for surveys three months after treatment ($t = 4$) and compare the effects and statistical significance. Provide a substantive interpretation of the results.

Group 16 - Sex Ratio and the Price of Agricultural Crops in China

In this group, we consider the effect of a change in the price of agricultural goods whose production and cultivation are dominated by either men or women.

These exercises are based on: Qian, Nancy. 2008. "Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance." *Quarterly Journal of Economics* 123(3): 1251–85.

Our data come from China, where centrally planned production targets during the Maoist era led to changes in the prices of major staple crops. We focus here on tea, the production and cultivation of which required a large female labor force, as well as orchard fruits, for which the labor force was overwhelmingly male. We use price increases brought on by government policy change in 1979 as a proxy for increases in sex-specific income, and ask the following question: Do changes in sex-specific income alter the incentives for Chinese families to have children of one gender over another? The CSV data file, `chinawomen.csv`, contains the variables shown in the table below, with each observation representing a particular Chinese county in a given year. Note that `post` is an indicator variable that takes takes 1 in a year following the policy change and 0 in a year before the policy change.

Name	Description
`birpop`	Birth population in a given year
`biryr`	Year of cohort (birth year)
`cashcrop`	Amount of cash crops planted in county
`orch`	Amount of orchard-type crops planted in county
`teasown`	Amount of tea sown in county
`sex`	Proportion of males in birth cohort
`post`	Indicator variable for introduction of price reforms

Question 1

We begin by examining sex ratios in the post-reform period (that is, the period after 1979) according to whether or not tea crops were sown in the region. Estimate the mean sex ratio in 1985, which we define as the proportion of male births, separately for tea-producing and non-tea-producing regions. Compute the 95% confidence interval for each estimate by assuming independence across counties within a year (We will maintain this assumption throughout this exercise). Furthermore, compute the difference-in-means between the two regions and its 95% confidence interval. Are sex ratios different across these regions? What assumption is required in order for us to interpret this difference as causal?

NB: I do not know how to make the analysis by region. What variable could give me that information? admin, for instance, is not helpful to get a tea-producing or non-tea-producing region.

However, I can do the following:

```
chinaw <- read.csv("https://raw.githubusercontent.com/umbertomig/intro-prob-stat-FGV/master/datasets/ch

library("Rmisc")
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
china_85 <- chinaw %>%
  filter(biryr == '1985')
```

Estimating the mean sex ratio in 1985 and computing the confidence interval:

```
mean(china_85$sex, na.rm = T)
```

```
## [1] 0.5211646
```

```
CI(china_85$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5269486 0.5211646 0.5153805
```

Estimating the mean for tea-producing in 1985 and computing the confidence interval:

```
mean(china_85$teasown, na.rm = T)
```

```
## [1] 0.03301009
```

```
CI(china_85$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04114475 0.03301009 0.02487543
```

Estimating the mean for non-tea-producing in 1985 and computing the confidence interval:

```
mean(china_85$orch, na.rm = T)
```

```
## [1] 0.1488621
```

```
CI(china_85$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1683254 0.1488621 0.1293988
```

Computing the difference-in-means:

t.test(orch, data = china_85, conf.level= 0.95) t.test(teasown, data = china_85, conf.level= 0.95)

Question 2

Repeat the analysis in the previous question for subsequent years, i.e., 1980, 1981, 1982, . . . , 1990. Create a graph which plots the difference-in-means estimates and their 95% confidence intervals against years. Give a substantive interpretation of the plot.

Repeating the analysis:

```
china_80 <- chinaw %>%
  filter(biryr == '1980')
mean(china_80$sex, na.rm = T)
```

```
## [1] 0.5215332
```

```
CI(china_80$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5275400 0.5215332 0.5155264
```

```r
mean(china_80$teasown, na.rm = T)
```

```
## [1] 0.0337587
```

```r
CI(china_80$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04215586 0.03375870 0.02536154
```

```r
mean(china_80$orch, na.rm = T)
```

```
## [1] 0.1529814
```

```r
CI(china_80$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1730821 0.1529814 0.1328808
```

```r
china_81 <- chinaw %>%
  filter(biryr == '1981')
mean(china_81$sex, na.rm = T)
```

```
## [1] 0.5175298
```

```r
CI(china_81$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5231263 0.5175298 0.5119333
```

```r
mean(china_81$teasown, na.rm = T)
```

```
## [1] 0.03400932
```

```r
CI(china_81$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04244613 0.03400932 0.02557252
```

```r
mean(china_81$orch, na.rm = T)
```

```
## [1] 0.1522261
```

```r
CI(china_81$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1722977 0.1522261 0.1321545
```

```r
china_82 <- chinaw %>%
  filter(biryr == '1982')
mean(china_80$sex, na.rm = T)
```

```
## [1] 0.5215332
```

```r
CI(china_82$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5307538 0.5252309 0.5197080
```

```r
mean(china_82$teasown, na.rm = T)
```

```
## [1] 0.03365052
```

```r
CI(china_82$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04200128 0.03365052 0.02529975
```

```r
mean(china_82$orch, na.rm = T)
```

```
## [1] 0.1508247
```

```r
CI(china_82$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1706928 0.1508247 0.1309566
```

```r
china_83 <- chinaw %>%
  filter(biryr == '1983')
mean(china_83$sex, na.rm = T)
```

```
## [1] 0.5221599
```

```r
CI(china_83$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5283830 0.5221599 0.5159368
```

```r
mean(china_83$teasown, na.rm = T)
```

```
## [1] 0.03367783
```

```r
CI(china_83$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04203815 0.03367783 0.02531751
```

```r
mean(china_83$orch, na.rm = T)
```

```
## [1] 0.1517148
```

```r
CI(china_83$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1717129 0.1517148 0.1317167
```

```r
china_84 <- chinaw %>%
  filter(biryr == '1984')
mean(china_84$sex, na.rm = T)
```

```
## [1] 0.5244052
```

```r
CI(china_84$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5301525 0.5244052 0.5186580
```

```r
mean(china_84$teasown, na.rm = T)
```

```
## [1] 0.03351382
```

```r
CI(china_84$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04185379 0.03351382 0.02517386
```

```r
mean(china_84$orch, na.rm = T)
```

```
## [1] 0.1515668
```

```r
CI(china_84$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1715260 0.1515668 0.1316076
```

```r
china_86 <- chinaw %>%
  filter(biryr == '1986')
mean(china_86$sex, na.rm = T)
```

```
## [1] 0.5186162
```

```
CI(china_86$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5245858 0.5186162 0.5126467
```

```
mean(china_86$teasown, na.rm = T)
```

```
## [1] 0.03218033
```

```
CI(china_86$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04011386 0.03218033 0.02424679
```

```
mean(china_86$orch, na.rm = T)
```

```
## [1] 0.1465847
```

```
CI(china_86$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1656004 0.1465847 0.1275690
```

```
china_87 <- chinaw %>%
  filter(biryr == '1987')
mean(china_87$sex, na.rm = T)
```

```
## [1] 0.5159377
```

```
CI(china_87$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5214729 0.5159377 0.5104026
```

```
mean(china_87$teasown, na.rm = T)
```

```
## [1] 0.03212766
```

```
CI(china_87$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04004842 0.03212766 0.02420690
```

```
mean(china_87$orch, na.rm = T)
```

```
## [1] 0.1458538
```

```r
CI(china_87$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1648301 0.1458538 0.1268775
```

```r
china_88 <- chinaw %>%
  filter(biryr == '1988')
mean(china_88$sex, na.rm = T)
```

```
## [1] 0.5278146
```

```r
CI(china_88$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5332797 0.5278146 0.5223495
```

```r
mean(china_88$teasown, na.rm = T)
```

```
## [1] 0.03210498
```

```r
CI(china_88$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04004147 0.03210498 0.02416848
```

```r
mean(china_88$orch, na.rm = T)
```

```
## [1] 0.146129
```

```r
CI(china_88$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1651448 0.1461290 0.1271133
```

```r
china_89 <- chinaw %>%
  filter(biryr == '1989')
mean(china_89$sex, na.rm = T)
```

```
## [1] 0.5241544
```

```r
CI(china_89$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5293601 0.5241544 0.5189488
```

```r
mean(china_89$teasown, na.rm = T)
```

```
## [1] 0.03190244
```

```r
CI(china_89$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.03977259 0.03190244 0.02403229
```

```r
mean(china_89$orch, na.rm = T)
```

```
## [1] 0.1450081
```

```r
CI(china_89$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1638721 0.1450081 0.1261441
```

```r
china_90 <- chinaw %>%
  filter(biryr == '1990')
mean(china_90$sex, na.rm = T)
```

```
## [1] 0.5291878
```

```r
CI(china_90$sex, ci = 0.95)
```

```
##     upper      mean     lower
## 0.5357967 0.5291878 0.5225788
```

```r
mean(china_90$teasown, na.rm = T)
```

```
## [1] 0.03219298
```

```r
CI(china_90$teasown, ci = 0.95)
```

```
##      upper       mean      lower
## 0.04015088 0.03219298 0.02423508
```

```r
mean(china_90$orch, na.rm = T)
```

```
## [1] 0.1457182
```

```r
CI(china_90$orch, ci = 0.95)
```

```
##     upper      mean     lower
## 0.1647582 0.1457182 0.1266782
```

Question 3

Next, we compare tea-producing and orchard-producing regions before the policy enactment. Specifically, we examine the sex ratio and the proportion of Han Chinese in 1978. Estimate the mean difference, its standard error, and 95% confidence intervals for each of these measures between the two regions. What do the results imply about the interpretation of the results given in Question 1?

Question 4

Repeat the analysis for the sex ratio in the previous question for each year before the reform, i.e., from 1962 until 1978. Create a graph which plots the difference-in-means estimates between the two regions and their 95% confidence intervals against years. Give a substantive interpretation of the plot.

Question 5

We will adopt the difference-in-differences design by comparing the sex ratio in 1978 (right before the reform) with that in 1980 (right after the reform). Focus on a subset of counties that do not have missing observations in these two years. Compute the difference-in-differences estimate and its 95% confidence interval.