Homework 4: Graphic Language

Econ 145

Overview

For this first homework assignment, you will be working with data from FiveThirtyEight, a statistical newspaper site. There are three different data sources you will be using: `covid_approval_polls_adjusted.csv`, `covid_approval_toplines.csv`, and `covid_concern_toplines.csv`. In this data, you will be analyzing the US sentiment on how president Joe Biden and former president Donald Trump have handled the Covid-19 pandemic or how concerned US citizens are with Covid-19. Each of these data sets have similar column names, so only `covid_approval_polls_adjusted.csv` will be described in what follows. Here are some of the most important columns from the `covid_approval_polls_adjusted.csv` data:

- `subject` - the president at the time (Trump/Biden).
- `modeldate` - the date that FiveThirtyEight created their weighted average of approvals. A full continuous range of dates from the beginning of the pandemic to June 2022.
- `pollster` - the entity that conducted the poll.
- `startdate` - the start date of the poll conducted.
- `enddate` - the end date of the poll conducted.
- `approve_adjusted` - a weighted approval rating (out of 100).
- `disappove_adjusted` - a weighted disapproval rating (out of 100).

To Receive Credit

- Save the scripting file (i.e. your R program file) as assignment_4.R. Make sure your capitalization is correct as the autograder is case-sensitive.
- Be sure to include your first name, last name, and perm number on your one page write-up.
- Make sure all changes to the original dataset are done within the R script.
- Your one page write-up must be submitted in a .pdf to receive credit.

Grading on Coding Questions

Grading on the coding portion of the homework will come in two types of questions: *Public Questions* and *Private Questions*. Public Questions can be submitted as many times as you like to the autograder, and the autograder will give detailed feedback. Additionally, the TAs will help you on Nectir and in office hours on the Public Questions. On the other hand, Private Questions can be thought of as a mini quiz within the homework. While you still have as many times to upload your answer as you want, the autograder will not provide any feedback, and the TAs and Professor Startz will not provide any guidance or assistance (but getting advice from classmates on Nectir or elsewhere is completely okay). Private Questions will be marked on the homework assignment.

Part 1: Coding Assignment

For each homework assignment, words colored in magenta indicate a variable/vector/tibble that will be graded by the autograder. Pay close attention to these colored texts and be sure not to miss any.

1. Read in the polling data using `readr::read_csv` and name the tibble polls_adjusted.

2. For this question, we will be demonstrating the power of using ggplot2 graphics and its layering grammar by creating a histogram of the average approval ratings. Throughout this problem, you will be creating new ggplot2 graphics that progressively add more layers of detail. Note that for this question to be considered correct by the autograder, you must follow the instructions exactly as written.

    a) Using the `dplyr::group_by`, `dplyr::mutate`, and `dplylr::ungroup` functions, group by the `subject` column and create a new column named `average_approval` which is the mean of the `approve_adjusted` column. Update the polls_adjusted tibble to reflect these changes.
    b) Let's create our first ggplot2 object/graphic. Using the polls_adjusted tibble, create a histogram of the `approval_adjusted` column. Label the x-axis "Approval", the y-axis "Count" and the title this "Adjusted Approval Ratings". Finally, use the `theme_minimal` layer to give the graph a cleaner look. Save this ggplot2 object as approval_graph.
    c) Notice that the graph created in (b) is not very informative outside of checking for outliers (which a boxplot could do better). Let's split the graph created in (b) by Biden and Trump to make it more informative. Use Figure 2 as your reference. To do this, you will need to use the `facet_wrap` layer, and also the `fill` argument in the `aes` function. Moreover, be sure your x-axis is labeled "Approval", y-axis "Count", title "American approval of Biden and Trump's response to coronavirus", subtitle "From 2020-2022", and the legend should be labeled exactly as shown in Figure 2. Use the `theme_minimal` layer to beautify the graph. Save this ggplot object as approval_graph_facet.
    d) While Figure 2 is a better graph, we can add layers additional layers to this to make it even better. In particular, it would be helpful to have a vertical line that shows the mean of both Biden and Trump's approvals. Using the `geom_vline` and the `average_approval` column, add a *dashed* verticle line on the approval_graph_facet object and save this new object as mean_line. Remember, you don't have to rewrite a ton of code with a ggplot2—it's easy to add on a layer to a ggplot2 object!
    e) Finally, while this graph is informative and pretty, the colors are incorrect—Biden is colored red and Trump is colored blue. Let's make these colors more aligned with their respective parties (i.e., Biden blue and Trump red). To do this, use the `scale_fill_manual` layer and the colors "#008FD5" and"#FF2700". Additionally, use the `theme` layer to move the legend to the bottom. Save this final ggplot2 object as approval_final. Refer to Figure 3 for how this should look.

3. For this question, we will be creating a scatter plot by political party on approvals of how the president is handling the covid-19 pandemic overtime. However, first there needs to be some light data wrangling done.

    a) Using the polls_adjusted data, create two new columns: `end_date` which is the `enddate` column but in the "date" datatype (use `lubridate::mdy`), and `approve_fraction` which is the `approve_adjusted` column divided by 100. Moreover, filter the data so that `party` only contains Democrats/Republicans/Independents (D/R/I). Save this tibble as polls_q3.
    b) Now, using polls_q3, recreate Figure 4 with `end_date` on the x-axis and `approve_fraction` on the y-axis. Note that the vertical line occurs at Joe Biden's inauguration date (2021-01-20). To replicate this graph, you will need to use the following layers: `geom_point`, `geom_smooth`, `geom_vline`, `scale_color_manual`, `scale_y_continuous`, `labs` and `theme_minimal`. Note that the colors used in this graph are "#008FD5","#77AB43", "#FF2700". Include this graph in your write-up on a its own page, be sure to give it the title "YOUR NAME: Approval of President's Handling of Covid-19 Pandemic". The graph will be graded on how close it is to the original. **HINT**: you do not need to put any arguments in the `geom_smooth` layer.

4. This question will be slightly different than most other questions you have seen in this course. The goal of this question will be to recreate Figure 5, which is very similar to a graph found on FiveThirtyEight's website. First, the question will guide you through some preprocessing/data wrangling. Finally, it will be your job to fill in the code to correctly replicate the graph. For this question, we will be using the `covid_approval_toplines.csv` data from FiveThirtyEight, which is similar to the previous data used, except with a couple of small tweaks that are not necessary for our purposes. The necessary columns in this question are the following:

- `subject` - same as before.
- `party` - same as before.
- `modeldate` - same as before.
- `approve_estimate` - our measure of approval we will be using.

a) First, read in the `covid_approval_toplines.csv` data using `readr::read_csv` and save this tibble as toplines.
b) Filter the `subject` to only include Biden, the `party` to only include D/R/I and, similar to question 2, use create a new column named `model_date` which will be the `modeldate` function but as a "date" datatype in YYYY-MM-DD form (using lubridate::mdy). Save this tibble as toplines_biden.
c) Create a new column named `party_description` which is equal to "Democrats", "Republicans", and "Independents" if the `party` column is "D", "R", and "I", respectively. Additionally create another column named `approve_estimate_frac` which is the `approve_estimate` column divided by 100. Be sure to update the toplines_biden data.
d) Finally, fill in the code to recreate the graph. Save the final ggplot2 object as final_graph.

```
## loading in libraries in case they are not loaded in.
## install if necessary
library(ggrepel)
library(scales)

toplines_biden %>%
  mutate(label = ifelse(model_date ==  max(model_date),
                        party_description, NA_character_)) %>% ## do not need to modify
  ggplot(aes()) + ## need to fill in
  geom_line() + ## do not need to modify
  geom_text_repel(aes(label = label),
                  nudge_x = 10, na.rm = T,
                  xlim = as_date(c("2022-07-01", "2022-10-01"))) + ## do not need to modify
  geom_vline() + ## need to fill in
  annotate("text", x = as_date("2021-01-20"), y = 0.05,
           label = "Biden sworn into office", size = 3,
           hjust = -0.1) + ## do not need to modify
  scale_color_manual(values = ) + ## need to fill in
  scale_y_continuous(labels = scales::percent) + ## do not need to modify
  coord_cartesian(ylim = c(.1,1), clip = "off") + ## do not need to modify
  scale_x_date(limits = c(as_date("2020-12-01"), as_date("2022-10-01"))) + ## do not need to modify
  labs() + ## need to fill in
  theme_minimal() + ## do not need to modify
  theme() ## need to fill in
```

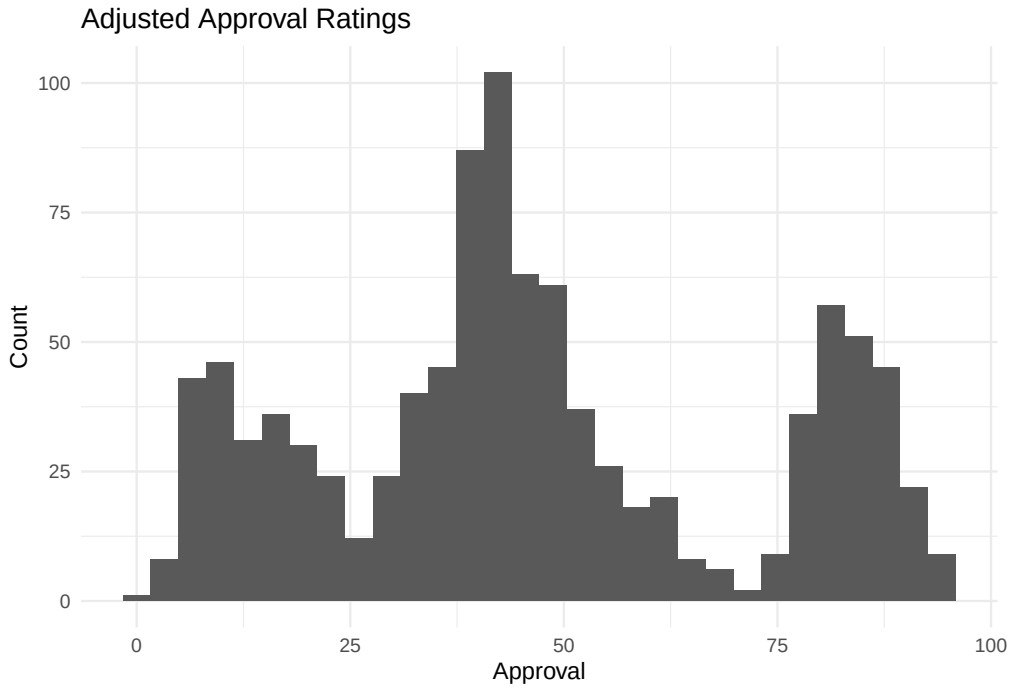

Figure 1: Graph to reproduce in 2b.

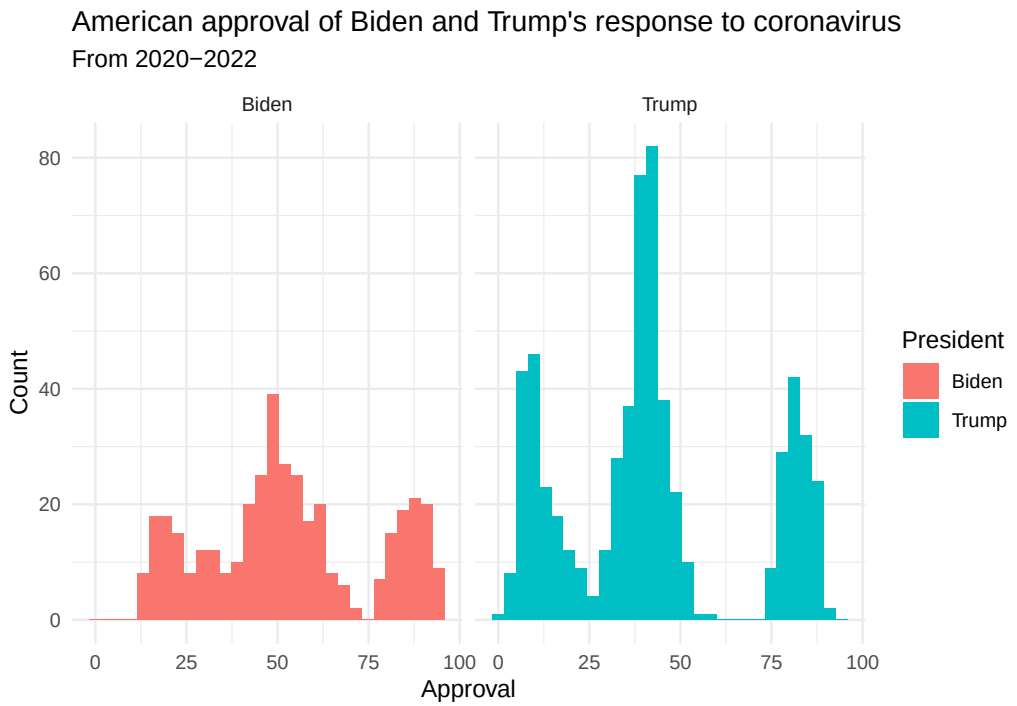

Figure 2: Graph to reproduce in 2c.

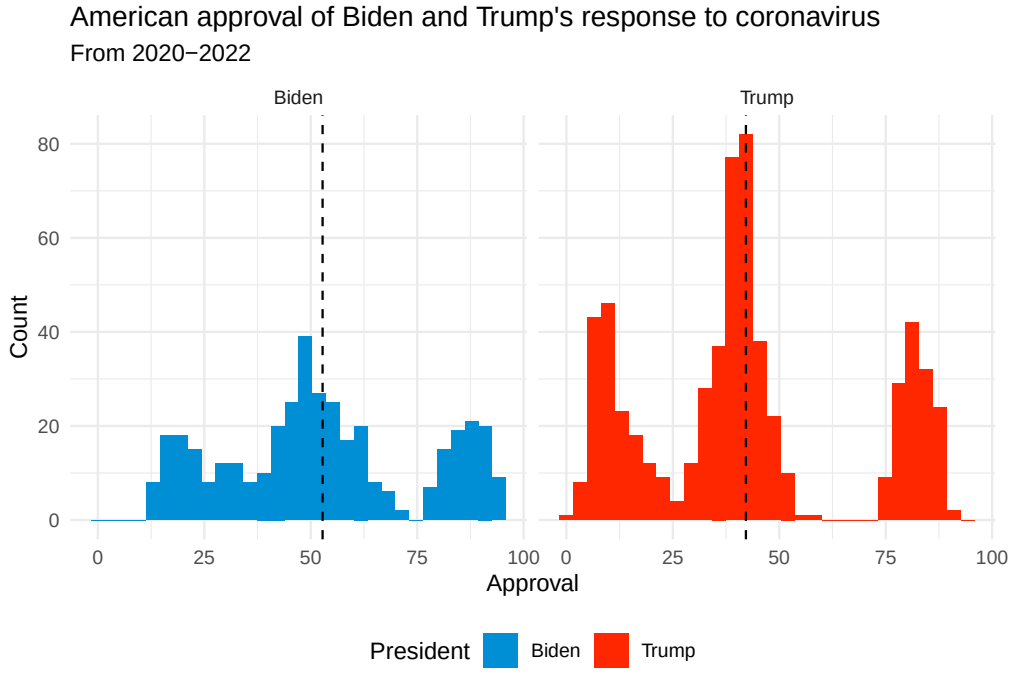

Figure 3: Graph to reproduce in 2e.

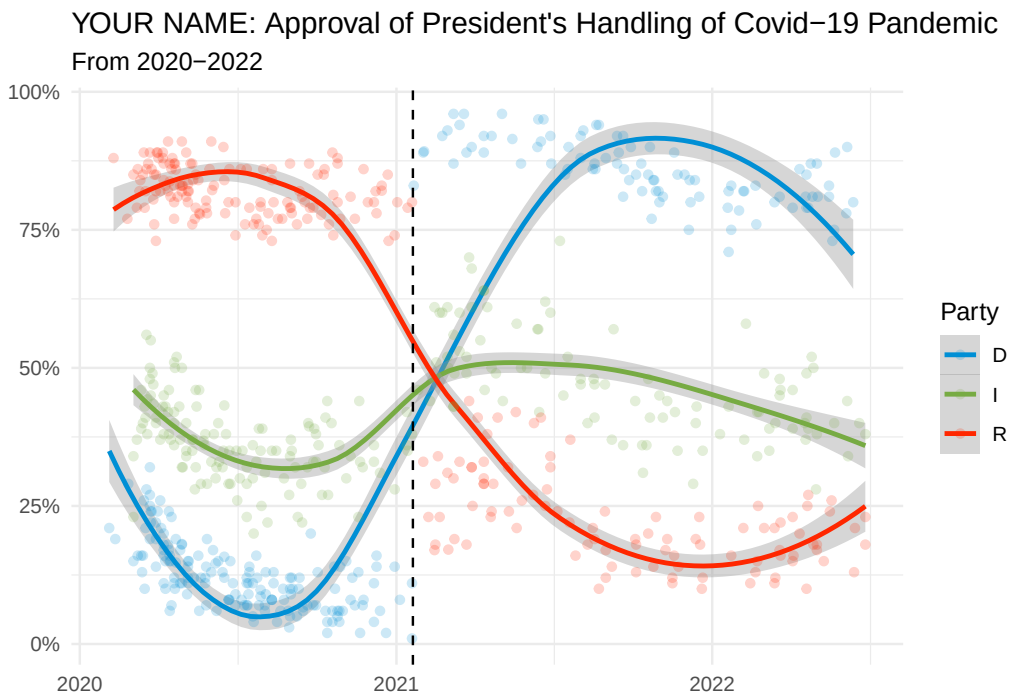

Figure 4: Graph to reproduce in 3b.

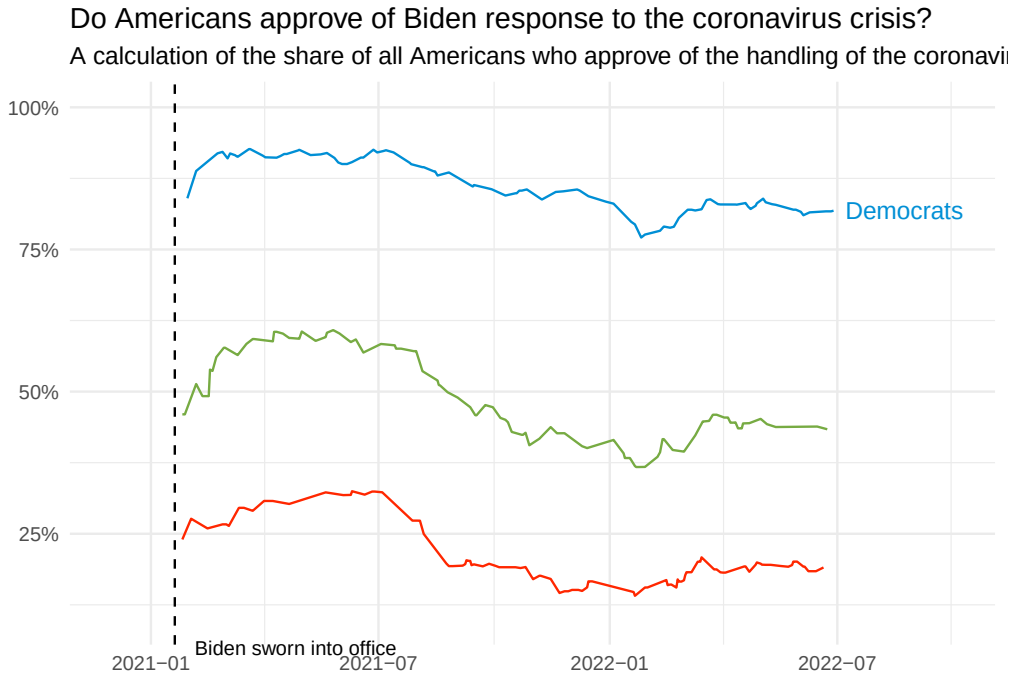

Figure 5: Graph to reproduce in 4b.

Part 2: Write-up

For this write-up, you will be using the `covid_concern_toplines.csv` data from FiveThirtyEight. This data contains information on how concerned people are with covid-19 over time. Your job is to pretend as if you are a journalist covering a very brief (no more than one-page!) article on the Covid-19 pandemic titled "How much fear does Covid-19 have left?". Your write-up should have at least one graph that is clearly labeled and some statistics which help describe your analysis. As seen in the homework, it is always more interesting to break down graphs/statistics by groups or show trends over time.