Problem Set 5

Multivariate Visualization

[YOUR NAME]

Due Date: 2024-02-16

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps5.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps5.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `Pres2020_PV.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/Pres2020_PV.Rds).

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 8 total points, plus two extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compiled the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Also note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _____

Question 0 [0 points]

Load `tidyverse` and the data from Github, which should be saved to an object called `pres` .

```
# INSERT CODE HERE
```

Question 1 [2 points]

Plot the total number of polls per start date in the data. NB: you will have convert `StartDate` to a `date` class with `as.Date()` . If you need help, see this post (https://www.r-bloggers.com/2013/08/date-formats-in-r/). Do you observe a noteworthy trend in the number of polls over time?

```
pres %>%
  mutate() %>% # Convert to a date class object
  ggplot() + # Put the start date on the x-axis
  geom_...() + # Choose the correct geom
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '')
```

```
## Error in pres %>% mutate() %>% ggplot(): could not find function "%>%"
```

Write answer here.

Question 2 [2 points]

Calculate the **prediction error** for Biden and Trump such that positive values mean that the poll *overestimated* the candidate's popular vote share (`DemCertVote` for Biden and `RepCertVote` for Trump). Plot the Biden and Trump prediction errors on a single plot using `geom_bar()` , with red indicating Trump and blue indicating Biden (make sure to set alpha to some value less than 1 to increase the transparency!). Add vertical lines for the average prediction error for both candidates (colored appropriately) as well as a vertical line indicating no prediction error. **HINT**: create a new object called `toplot` which adds the prediction error columns to `pres` via `mutate()` .

Do you observe a systematic bias toward one candidate or the other?

```
toplot <- pres %>%
  mutate(demErr = , # Calculate the prediction error per poll for Biden
         repErr = ) # Calculate the prediction error per poll for Trump
```

```
## Error in pres %>% mutate(demErr = , repErr = ): could not find function "%>%"
```

```
toplot %>%
  ggplot() + # Instantiate an empty plot
  geom_bar() +  # Add one set of blue bars for Biden
  geom_bar() +  # Add one set of red bars for Trump
  labs(title = '', # Make sure to give the plot intuitive labels!
       x = '',
       y = '') +
  theme_bw() + # Keep this to make the plot look fancy
  geom_vline() + # Add a vertical line at zero
  geom_vline() + # Add a blue vertical line at the average prediction error for Biden
  geom_vline() # Add a red vertical line for the average prediction error for Trump
```

```
## Error in toplot %>% ggplot(): could not find function "%>%"
```

Write answer here

Question 3 [2 points]

Plot the average prediction error for Trump (red) and Biden (blue) by start date using `geom_point()` and add a curvey line of best fit using `geom_smooth()`. What pattern do you observe over time, if any?

```
toplot %>%
  mutate() %>% # Convert start date to a date type variable
  group_by() %>% # Calculate the average prediction errors for Biden and Trump by start
date
  summarise() %>% # Calculate the average prediction errors for Biden and Trump by start
date
  ggplot() + # Create an empty plot
  geom_point() + # Add blue points for Biden's prediction error by start date
  geom_point() + # Add red points for Trump's prediction error by start date
  geom_smooth() + # Add a curvey blue line for Biden's prediction error over time
  geom_smooth() + # Add a curvey red line for Trump's prediction error over time
  labs(title = "", # Make sure to give the plot intuitive labels!
       x = "",
       y = "") +
  geom_hline() + # Add a dashed horizontal line at zero
  theme_bw() # Keep this to make the plot look fancy
```

```
## Error in toplot %>% mutate() %>% group_by() %>% summarise() %>% ggplot(): could not f
ind function "%>%"
```

Write answer here

Question 4 [2 points]

Calculate each poll's bias toward Biden (this should be the prediction error for Biden minus the prediction error for Trump) and plot the distribution. What proportion of polls' prediction error favored Biden over Trump? What does this mean about polling in the United States?

```
toplot %>%
  mutate() %>% # Calculate the poll's pro-Biden bias
  ggplot() +  # Put the Biden bias on the x-axis
  geom_...() + # Choose the correct geom
  labs(title = '', # Make sure to give the plot intuitive labels!
       subtitle = "",
       x = "",
       y = '') +
  geom_vline() + # Add a dashed vertical line at zero
  theme_bw() # Keep this to make the plot look fancy
```

```
## Error in toplot %>% mutate() %>% ggplot(): could not find function "%>%"
```

```
toplot %>%
  mutate() %>% # Calculate the poll's pro-Biden bias
  summarise() # Calculate the proportion of all polls that have a pro-Biden bias
```

```
## Error in toplot %>% mutate() %>% summarise(): could not find function "%>%"
```

Write answer here

Extra Credit [2 points]

Do polls that underestimate Trump's support overestimate Biden's support? Use a scatterplot to test, combined with a line of best fit. Then, calculate the proportion of polls that (1) underestimate both Trump and Biden, (2) underestimate Trump and overestimate Biden, (3) overestimate Trump and underestimate Biden, (4) overestimate both candidates. In these analyses, define "overestimate" as prediction errors greater than or equal to zero, whereas "underestimate" should be prediction errors less than zero.

```
# INSERT CODE HERE
```

Write answer here