Problem Set 3

Data Wrangling

[YOUR NAME]

Due Date: 2024-07-12

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 `[YOUR NAME]_ps3.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[YOUR NAME]_ps3.Rmd` file. Then change the `author: [Your Name]` on line 2 to your name.

We will be using two different files. First is the `Pres2020_PV.Rds` data from the course github page (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/data/Pres2020_PV.Rds). Second is the `game_summary.Rds` data, which is also on the course github page (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/data/game_summary.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 10 total points, plus **two** extra credit questions, each worth **two** 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, email the knitted output to Eun Ji Kim (kej990804@snu.ac.kr (mailto:kej990804@snu.ac.kr)) **as a PDF** by the start of class on Friday, July 12th. If you need help converting to a PDF, see this tutorial (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/Psets/ISP_pset_0_HELPER.pdf).

Good luck!

*Copy the link to ChatGPT you used here: _____

Question 0

Require `tidyverse` and load the `Pres2020_PV.Rds` data to an object called `pres`. ALSO load a new package called `labelled`, which will allow us to read the labels for our variables. Remember, if you don't have this package yet, you need to use `install.packages("labelled")` in the `Console` window.

```
require()
```

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

```
require()
```

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

```
pres <- read_rds("")
```

```
## Error in read_rds(""): could not find function "read_rds"
```

Question 1 [1 point]

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 pattern in the number of polls over time? Why do you think this is?*

```
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 [1 point]

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 [1 point]

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()` *(allow it to be curved!). 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 1 [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. What do you conclude?

```
# INSERT CODE HERE
```

Write answer here

Question 5 [1 point]

Now let's load a different dataset to practice multivariate visualization and confidence. Open `game_summary.Rds` *from the github page and save it to a new object called* `games`*. This dataset contains information on basketball games in the NBA from the 2017-2019 seasons. The codebook for it can be found in homework 5, which is also on github (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/Homeworks/ISP_hw_5.pdf).*

```
games <- read_rds("") # Insert link here
```

```
## Error in read_rds(""): could not find function "read_rds"
```

We are interested in the concept of "home court advantage", which predicts that teams play better at home than on the road. We are focusing on one team in the dataset (meaning you will need to `filter()`), my home team the Boston Celtics. How many points, on average, did the Boston Celtics score at home and away games in the 2017 season? Calculate this answer and also plot the multivariate relationship. Explain why your chosen visualization is justified. Draw two vertical lines for the average points at home and away.

```
# Create extra object to plot vertical lines for average points at home and away
vertLines <- games %>%
filter() %>% # Filter to the 2017 season (yearSeason) AND to the Boston Celtics (nameTea
m)
  group_by() %>% # Group by the location of the game
  summarise() # Calculate the average points (pts)
```

```
## Error in games %>% filter() %>% group_by() %>% summarise(): could not find function
"%>%"
```

```
games %>%
  filter() %>% # Filter to the 2017 season (yearSeason) AND to the Boston Celtics (nameT
eam)
  ggplot() + # Create a multivariate plot comparing points scored between home and away
games
  geom_...() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(),
geom_density(), geom_bar(), etc.)
  labs(title = '', # Add clear descriptions for the title, subtitle, axes, and legend
       subtitle = '',
       x = '',
       y = '',
       color = '') +
  geom_vline() # add vertical lines for the average points scored at home and away.
```

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

Write answer here

Question 6 [1 point]

Now recreate the same plot for the 2018, 2019, and combined seasons. Imagine that you work for the Celtics organization and Brad Stevens (the GM), asks you if the team scores more points at home or away? Based on your analysis, what would you tell him?

```
# By season
vertLines <- games %>%
filter() %>% # Filter to the Boston Celtics (nameTeam)
  group_by() %>% # Group by the location and the season
  summarise() # Calculate the average points (pts)
```

```
## Error in games %>% filter() %>% group_by() %>% summarise(): could not find function
"%>%"
```

```
games %>%
  filter() %>% # Filter to the Boston Celtics (nameTeam)
  ggplot() + # Create a multivariate plot comparing points scored between home and away
games
  geom_...() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(),
geom_density(), geom_bar(), etc.)
  labs(title = '', # Add clear descriptions for the title, subtitle, axes, and legend
       subtitle = '',
       x = '',
       y = '',
       color = '') +
  facet_wrap() + # Create separate panels for each season (facet_wrap())
  geom_vline() # add vertical lines for the average points scored at home and away.
```

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

```
# Over all seasons combined
vertLines <- games %>%
filter() %>% # Filter to the Boston Celtics (nameTeam)
  group_by() %>% # Group by the location
  summarise() # Calculate the average points (pts)
```

```
## Error in games %>% filter() %>% group_by() %>% summarise(): could not find function
"%>%"
```

```
games %>%
  filter() %>% # Filter to the Boston Celtics (nameTeam)
  ggplot() + # Create a multivariate plot comparing points scored between home and away
games
  geom_...() + # Choose the appropriate geom_... for this plot (i.e., geom_histogram(),
geom_density(), geom_bar(), etc.)
  labs(title = '', # Add clear descriptions for the title, subtitle, axes, and legend
       subtitle = '',
       x = '',
       y = '',
       color = '') +
  geom_vline() # add vertical lines for the average points scored at home and away.
```

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

Write answer here

Question 7 [1 point]

Brad Stevens thanks you for your answer, but is a well-trained statistician in his own right, and wants to know how confident you are in your claim. Bootstrap sample the data 1,000 times to provide him with a more sophisticated answer. How confident are you in your conclusion that the Celtics score more points at home games than away games? Make sure to `set.seed(123)` *to ensure you get the same answer every time you* `knit` *your code!*

```
set.seed(123) # Set the seed!
forBS <- games %>% # To make things easier, create a new data object that is filtered to
just the Celtics so we don't have to do this every time in the loop
    filter() # Filter to the Celtics (nameTeam)
```

```
## Error in games %>% filter(): could not find function "%>%"
```

```
bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- forBS %>%
    sample_n() %>% # Sample the data with replacement using all possible rows
    group_by() %>% # Group by the location of the game
    summarise() %>% # Calculate the average points (pts)
    ungroup() %>% # Best practices!
    spread() %>% # Spread the data to get one column for average points at home and anot
her for average points away
    mutate(, # Calculate the difference between home and away points
           ) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 133
}
```

```
## Error in forBS %>% sample_n() %>% group_by() %>% summarise() %>% ungroup() %>% : coul
d not find function "%>%"
```

```
# Calculate the confidence
bsRes %>%
  summarise(, # Calculate the proportion of bootstrap simulations where the home points
are greater than the away points
            ) # Calculate the overall average difference
```

```
## Error in bsRes %>% summarise(, ): could not find function "%>%"
```

Write answer here

Question 8 [2 points]

Re-do this analysis for three other statistics of interest to Brad: total rebounds (treb), turnovers (tov), and field goal percent (pctFG). **NOT GRADED:** *Do you notice anything strange in these results? What might explain it?*

```
bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- forBS %>%
    sample_n() %>% # Sample the data with replacement using all possible rows
    group_by() %>% # Group by the location of the game
    summarise(, # Calculate the average total rebounds (treb)
              , # Calculate the average turnovers (tov)
              ) %>% # Calculate the average field goal shooting percentage (pctFG)
    ungroup() %>% # Best practices!
    pivot_wider(, # Pivot wider to get each measure in its own colunm for homme and away
games
                ) %>% # Use the values from the variables you created above
    mutate(, # Calculate the difference between home and away total rebounds
           , # Calculate the difference between home and away turnovers
           , # Calculate the difference between home and away field goal percentages
           ) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 165
}
```

```
## Error in forBS %>% sample_n() %>% group_by() %>% summarise(, , ) %>% ungroup() %>% :
could not find function "%>%"
```

```
# Calculate the confidence
bsRes %>%
  summarise(, # Calculate the confidence for rebounds being greater than zero
            , # Calculate the confidence for turnovers being greater than zero
            )
```

```
## Error in bsRes %>% summarise(, , ): could not find function "%>%"
```

Write answer here

Extra Credit 2 [2 points]

Now Brad is asking for a similar analysis of other teams. Calculate the difference between home and away points for every team in the league and prepare a summary table that includes both the average difference for each team, as well as your confidence about whether the difference is not zero. Based on these data, would you argue that there is an **overall** *home court advantage in terms of points across the NBA writ large? Visualize these summary results by plotting the difference on the x-axis, the teams (reordered) on the y-axis, and the points colored by whether you are more than 90% confident in your answer. What does it mean to have less than 50% confidence?*

```
# INSERT CODE HERE
```

Write answer here