Problem Set 4

Multivariate Visualization and Analysis

[YOUR NAME]

Due Date: 2023-02-17

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]_ps4.Rmd` to your `code` folder.

Copy and paste the contents of this file into your `[LAST NAME]_ps4.Rmd` file. Then change the `author: [YOUR NAME]` (line 4) to your name.

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 five extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must both have the correct code **and include a comment describing what each line does**. In addition, some questions ask you to provide a written response in addition to the code. Unlike the first two problem sets, some of the code chunks are totally empty, requiring you to try writing the code from scratch. Make sure to comment each line, explaining what it is doing!

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 by midnight on 2023/02/17.

Good luck!

Question 0

Require `tidyverse` and load the `game_summary.rds` (https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Multivariate/data/game_summary.Rds?raw=true') data to an object called `games` . (Tip: use the `read_rds()` function with the link to the raw data.)

```
#LINK: https://github.com/jbisbee1/DS1000_S2023/blob/main/Lectures/4_Uni_Multivariate/data/game_summary.Rds?raw=true')
```

Question 1 [1 point]

How 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. EC: Draw two vertical lines for the average points at home and away and label them with the average points using `annotate(geom = 'text',...)` .

```
games %>%
  filter() %>% # Filter to the 2017 season (yearSeason) AND to the Boston Celtics (nameTeam)
  group_by() %>% # Group by the location of the game (locatoinGame)
  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 (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(), geo
m_bar(), etc.)
  labs() + # Add clear descriptions for the title, subtitle, axes, and legend
  geom_vline() + # EC: add vertical lines for the average points scored at home and away.
  annotate() # EC: label the vertical lines
```

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

> Write one to two sentences here.

Question 2 [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
games %>%
  filter() %>% # Filter to the Boston Celtics (nameTeam)
  group_by() %>% # Group by the location (locationGame) and the season (yearSeason)
  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(), geo
m_bar(), etc.)
  labs() + # Add clear descriptions for the title, subtitle, axes, and legend
  facet_wrap() + # Create separate panels for each season (facet_wrap())
  geom_vline() +
  geom_text()
```

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

```
# Over all seasons combined
games %>%
  filter() %>% # Filter to the Boston Celtics (nameTeam)
  group_by() %>% # Group by the location (locationGame)
  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(), geo
m_bar(), etc.)
  labs() + # Add clear descriptions for the title, subtitle, axes, and legend
  geom_vline() +
  geom_text()
```

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

> Write 3 to 4 sentences here.

Question 3 [2 points + 1 EC]

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! EC: Visualize your answer.

```
# Set the seed!
forBS <- games %>% # To make things easier, create a new data object that is filtered to just the Celtics
    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 (locationGame)
    summarise() %>% # Calculate the average points (pts)
    ungroup() %>% # Best practices!
    spread() %>% # Spread the data to get one column for average points at home and another for average poi
nts away
    mutate(diff = , # Calculate the difference between home and away points
           bsInd = ) %>% # 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() %>% : could not find function
"%>%"
```

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

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

```
# EC: Plot the result
```

> Write one to two sentences here.

Question 4 [2 point + 1 EC]

Re-do this analysis for three other statistics of interest to Brad: total rebounds (`treb`), turnovers (`tov`), and field goal percent (`pctFG`). EC: theorize about the seeming paradox in your answer to Brad Stevens.

```
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(locationGame) %>% # Group by the location of the game (locationGame)
    summarise(avg_reb = , # Calculate the average total rebounds (treb)
              avg_tov = , # Calculate the average turnovers (tov)
              avg_pctFG = ) %>% # Calculate the average field goal shooting percentage (pctFG)
    ungroup() %>% # Best practices!
    pivot_wider(names_from = , # Pivot wider to get each measure in its own column for home and away games.
                values_from = c('')) %>% # Use the values from the variables you created above
    mutate(diff_reb = , # Calculate the difference between home and away total rebounds
           diff_tov = , # Calculate the difference between home and away turnovers
           diff_pctFG = , # Calculate the difference between home and away field goal percentages
           bsInd = ) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 165
}
```

```
## Error in forBS %>% sample_n() %>% group_by(locationGame) %>% summarise(avg_reb = , : could not find func
tion "%>%"
```

```
# Calculate the confidence
bsRes %>%
  summarise(confidence_reb = , # Calculate the confidence for the home court advantage in rebounds
            confidence_tov = , # Calculate the confidence for the home court (dis)advantage in turnovers
            confidence_pctFG = ) # Calculate the confidence for the home court advantage in FG%
```

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

Write two to five sentences here.

Question 5 [2 point + 1 EC]

Now Brad is asking for a similar analysis of other teams. Calculate the difference between home and away turnovers 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 the difference is not zero. Based on these data, would you argue that there is an **overall** home court advantage in terms of turnovers across the NBA writ large? EC #1: 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. EC #2: How should we interpret confidence levels less than 50%?

```
bsRes <- NULL # Instantiate an empty object to store data from the loop
for(i in 1:1000) { # Loop 1,000 times
  bsRes <- games %>%
    group_by() %>% # Group by the team (nameTeam)
    sample_n() %>% # Sample the data with replacement using all possible rows FOR EACH TEAM (hint: use n())
    group_by() %>% # Group by the location of the game (locationGame) and the team (nameTeam)
    summarise(avg_tov = , # Calculate the average turnovers (tov)
              .groups = 'drop') %>% # Best practices! (Also reduces the messages!)
    pivot_wider(id_cols = , # Set the ID colunm to the team (nameTeam)
                names_from = , # Pivot wider to get each measure in its own column for home and away games
                values_from = c('')) %>% # Use the values from the average turnover measure you created abo
ve
    mutate(diff_tov = , # Calculate the difference between home and away turnovers
           bsInd = i) %>% # Save the bootstrap index
    bind_rows(bsRes) # Append the result to the empty object from line 165
}
```

```
## Error in games %>% group_by() %>% sample_n() %>% group_by() %>% summarise(avg_tov = , : could not find f
unction "%>%"
```

```
toplot <- bsRes %>% # If you want to attempt the EC#1, it helps to save the summarized results to a new obj
ect 'toplot'
  group_by() %>% # Group by the team (nameTeam)
  summarise(conf_tov = round(), # Calculate the confidence and round the number to two digits
            diff_tov = round()) # Calculate the average difference and round the number to two digits
```

```
## Error in bsRes %>% group_by() %>% summarise(conf_tov = round(), diff_tov = round()): could not find func
tion "%>%"
```

```
# EC #1: Visualize the results. Make sure to label clearly!
```

Write 2 to 5 sentences here.

Question 6 [2 points]

Redo question 5 but analyze the point difference instead. Do you think there is a systematic home court advantage in terms of points across the NBA writ large?

```
# Insert code here.
```

Write 1 to 3 sentences here