Problem Set 6

Multivariate Analysis and Uncertainty

[YOUR NAME]

Due Date: 2024-02-23

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]_ps6.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps6.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `game_summary.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/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 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

Require `tidyverse` and load the `game_summary.rds` (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/game_summary.Rds?raw=true') data to an object called `games` .

```
# INSERT CODE HERE
```

Question 1 [2 points]

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 2 [2 points]

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 3 [2 points]

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 another 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() %>% : could 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 4 [2 points]

Re-do this analysis for three other statistics of interest to Brad: total rebounds (treb), turnovers (tov), and field goal percent (pctFG). 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 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. How should we interpret confidence levels less than 50%?

```
# INSERT CODE HERE
```

> Write answer here