Problem Set 4

Univariate Analysis

[YOUR NAME]

Due Date: 2024-02-09

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 `.Rmd` file into your `[LAST NAME]_ps4.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using the `nba_players_2018.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/nba_players_2018.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!

ChatGPT Link [Optional]

*Copy the link to ChatGPT you used here: _______________________.

Question 0

Require `tidyverse` and an additional package called `haven` (remember to `install.packages("haven")` if you don't have it yet), and then load the `nba_players_2018.Rds` (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/nba_players_2018.Rds?raw=true') data to an object called `nba` .

```
# INSERT CODE HERE
```

Question 1 [2 points]

Plot the distribution of points scored by all NBA players in the 2018-2019 season. Explain why you chose the visualization that you did. Then add a vertical line indicating the mean and median number of points in the data. Color the median line blue and the mean line red. Why is the median lower than the mean?

```
nba %>%
  ggplot() + # Put the pts variable on the x-axis of a ggplot.
   geom_...() + # Choose the appropriate geom function to visualize.
  labs(title = '',# Write a clear title explaining the plot
       subtitle = '',# Write a clear subtitle describing the data
       x = '',# Write a clear x-axis label
       y = '') + # Write a clear y-axis label
      geom_vline(xintercept = median(),color = '') + # Median vertical line (blue)
      geom_vline(xintercept = mean(),color = '') # Mean vertical line (red)
```

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

> Write answer here.

Question 2 [2 points]

Now examine the `country` variable. Visualize this variable using the appropriate `geom_...`, and justify your reason for choosing it. Tweak the plot to put the country labels on the y-axis, ordered by frequency. Which country are most NBA players from? What is weird about your answer, and what might explain it?

```
nba %>%
  count() %>% # count the number of players by country
  ggplot() + # place the country on the y-axis, reordered by the number of players. Put the number of players on the x-axis
  geom_bar() + # Set stat = 'identity' because we are setting both x and y axes
  labs(title = '', # Clear title
       subtitle = '', # Clear subtitle
       x = '', # Clear x-axis label
       y = '') # Clear y-axis label
```

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

> Write answer here.

Question 3 [2 points]

Perform a thorough univariate description of the variable `agePlayer`. Start by determining what type of measure it is (i.e., continuous, ordered categorical, etc.). Then, based on this conclusion, summarize it with either `summary()` or `count()`. Finally, visualize it. In the write-up, explain each part of this process and defend your choice of the `geom_...` used to visualize the data. Make sure to label the plot!

```
glimpse() # Look at the variable first
```

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

```
summary() # Summarize the variable with either summary() or count()
```

```
## Error in is.factor(object): argument "object" is missing, with no default
```

```
nba %>% # Visualize with ggplot() (don't forget to label the plot!)
  ggplot() + # Put the agePlayer variable on the x-axis of a ggplot.
  geom_...() + # Choose the appropriate geom function to visualize.
  labs(title = "", # Write a clear title explaining the plot
       subtitle = '', # Write a clear subtitle describing the data
       x = '', # Write a clear x-axis label
       y = '') # Write a clear y-axis label
```

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

> Write answer here.

Question 4 [2 points]

Consider the following research question: do coaches give more minutes to younger players? Hypothesize an answer to this question, and describe your thought process (theory).

> Write answer here.

Extra Credit [2 points]

Generate a multivariate visualization that provides an answer to this question. To do this, first calculate the average minutes played by player age, then plot the resulting data. Does the data support your hypothesis?

```
# INSERT CODE HERE
```

> Write answer here.