Final Project Exploratory Data Analysis

Julia Vitale, Mia Zottoli, and Vishnu Lakshman

```
library(here)
library(tidyverse)
library(ggplot2)
library(corrr)
library(ggcorrplot)
library(FactoMineR)
library(factoextra)
library(ggfortify)
library(cluster)
```

```
ultimate_data <- read_csv(here("data", "ultimate_college_championship.csv")) %>%
  mutate(across(c(level, gender, division, team_name), as.factor)) %>%
  mutate(school = str_split_i(team_name, " ", 1)) %>%
  mutate(school = ifelse(str_detect(team_name, "St Olaf"), "St Olaf", school)) %>%
  mutate(school = ifelse(str_detect(team_name, "North Carolina"), "North Carolina", school))
  mutate(school = ifelse(str_detect(team_name, "Cal Poly"), "Cal Poly SLO", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "British Columbia"), "British Columbia", scho
  mutate(school = ifelse(str_detect(team_name, "Western Washington"), "Western Washington", s
  mutate(school = ifelse(str_detect(team_name, "Penn State"), "Penn State", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "San Diego"), "UCSD", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Lewis"), "Lewis & Clark", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Colorado State"), "Colorado State", school))
  mutate(school = ifelse(str_detect(team_name, "Oklahoma Christian"), "Oklahoma Christian", s
  mutate(school = ifelse(str_detect(team_name, "Binghamton"), "SUNY Binghamton", school))  %
  mutate(school = ifelse(str_detect(team_name, "Santa Bar"), "UCSB", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Santa Cr"), "USCS", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Colorado College"), "Colorado College", scho
  mutate(school = ifelse(str_detect(team_name, "Missouri S"), "Missouri S&T", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Holyoke"), "Mount Holyoke", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "NC State"), "NC State", school))  %>%
  mutate(school = ifelse(str_detect(team_name, "Oregon State"), "Oregon State", school))  %>
  mutate(school = ifelse(str_detect(team_name, "Washington University"), "Washington Universi
```

Exploring the Data

Answer the following questions:

- What is your outcome variable(s)? How well does it measure the outcome you are interested? How does it relate to your expectations?

Our outcome variable is plus_minus, which is the difference between the amount of points scored by an individual player's team while that player is on the field and the amount of points scored by the opposing team while that player is on the field. We are interested in the influence of an individual player on the success of the whole team, so this variable is a good measure of our outcome of interest.

Essentially, +/- for a select player = points scored by player's team (while player is on the field) - points scored by opposing team (while player is on the field)

The +/- score is used to track a player's overall effectiveness on the field and their impact on the game. A positive +/- score means the player's team scored more than the opposing team while the player was on the field, and a negative +/- score means the opposing team scored more while the player was on the field.

- What are your key explanatory variables?

  Turns (turnovers) thrown per game, points scored per game, Ds (defensive interceptions) per game, assists per game, level (Division 1 or 3), division (Men's or Women's) and school.

In addition, create a table of summary statistics for the variables you are planning to use.

```
ultimate_data %>% select(-c(player, team_name)) %>% gtsummary::tbl_summary()
```

Data Wrangling and Transformation

Answer the following question:

- What data cleaning did you have to do?
  - **The data was already pretty clean. We had to do some string manipulation to extract the school name from the team name, and had to convert some character variables to factors.**
- How did you wrangle the data?

Characteristic	**N = 1,665**[1]
level	
Division 1	973 (58%)
Division 3	692 (42%)
gender	
Men	893 (54%)
Women	772 (46%)
division	
Division 1 Men	521 (31%)
Division 1 Women	452 (27%)
Division 3 Men	372 (22%)
Division 3 Women	320 (19%)
Turns	2 (0, 7)
Ds	1.00 (0.00, 3.00)
Assists	1.0 (0.0, 3.0)
Points	1.0 (0.0, 4.0)
plus_minus	1 (0, 5)
team_games	
5	780 (47%)
6	738 (44%)
7	122 (7.3%)
8	25 (1.5%)
turns_per_game	0.40 (0.00, 1.17)
ds_per_game	0.20 (0.00, 0.50)
ast_per_game	0.17 (0.00, 0.60)
pts_per_game	0.20 (0.00, 0.71)
pls_mns_per_game	0.20 (0.00, 0.83)
school	
Alabama-Huntsville	23 (1.4%)
Bates	17 (1.0%)
Berry	26 (1.6%)
British Columbia	23 (1.4%)
Brown	25 (1.5%)
Cal	23 (1.4%)
Cal Poly SLO	24 (1.4%)
Carleton	101 (6.1%)
Claremont	24 (1.4%)
Colorado	49 (2.9%)
Colorado College	22 (1.3%)
Colorado State	21 (1.3%)
Davenport	31 (1.9%)
Franciscan	22 (1.3%)
Georgia	45 (2.7%)
Grinnell	15 (0.9%)
Haverford/Bryn	24 (1.4%)
Lewis & Clark	44 (2.6%)
Macalester	20 (1.2%)
Massachusetts	28 (1.7%)
Michigan	54 (3.2%)
Middlebury	53 (3.2%)

  - **We did not have to do significant data wrangling for this data set. If we choose an analysis method that requires standardization, we will have to standardize the numeric variables.**

- Are you deciding to exclude any observations? If so, why?
  - **No, we are not excluding any observations. There are no extreme outliers.**

- Did you have to create any new variables from existing variables? If so, how and why?
  - **We created a 'school name' variable which extracts the name of the college/university associated with the team name. Some schools have both men's and women's teams in this data set, and we are curious if school advantage transcends team-specific advantage.**

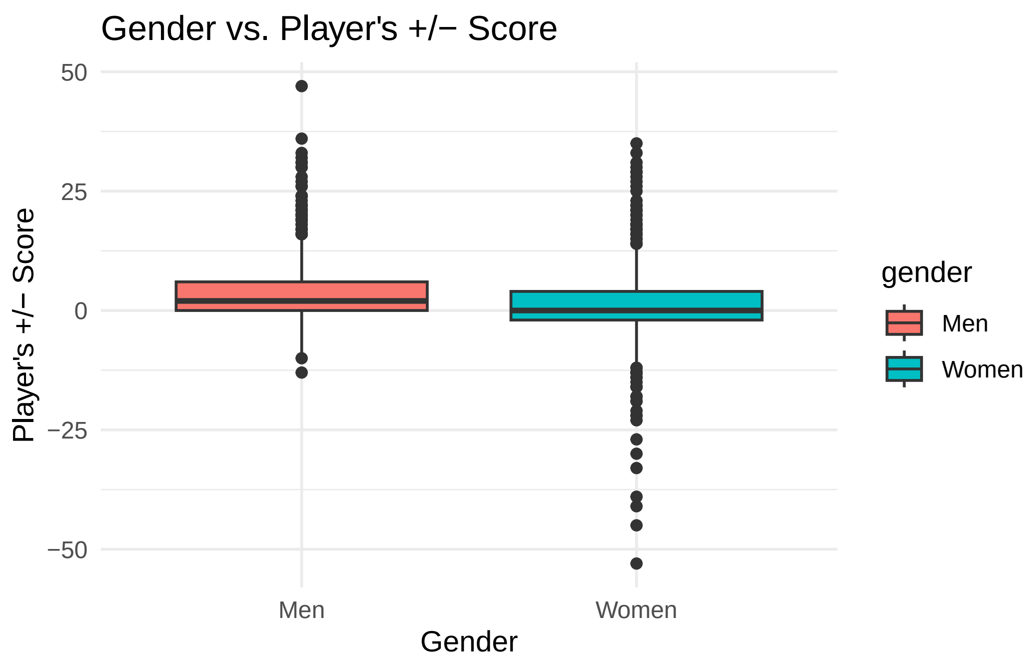
Codebook

You must add a *codebook* – a description of all variables you are using, including ones you are creating for this project – to the README.md page of the `data/` folder of your repo.

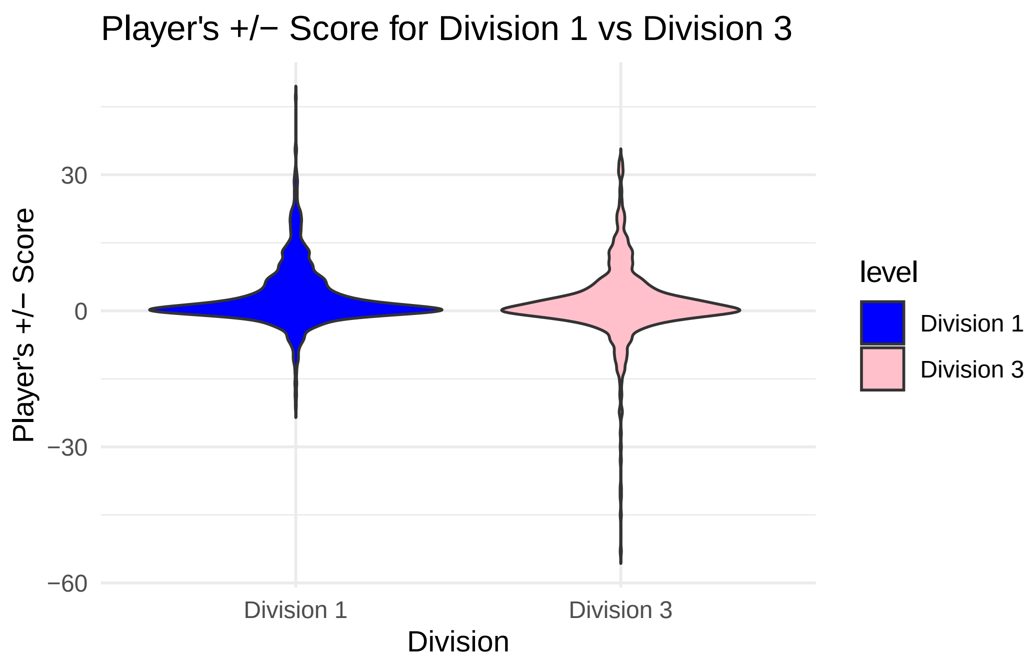
Data Visualization

You must include at least 4 visualizations of your data made in R. You must include your outcome variable in at least two plots and your key explanatory variable in at least two of these plots. You must use visualizations that are *appropriate* for the data type (categorical vs numeric, continuous vs discrete) of your outcome and explanatory variables. For example, you should not use a histogram to plot a categorical variable.

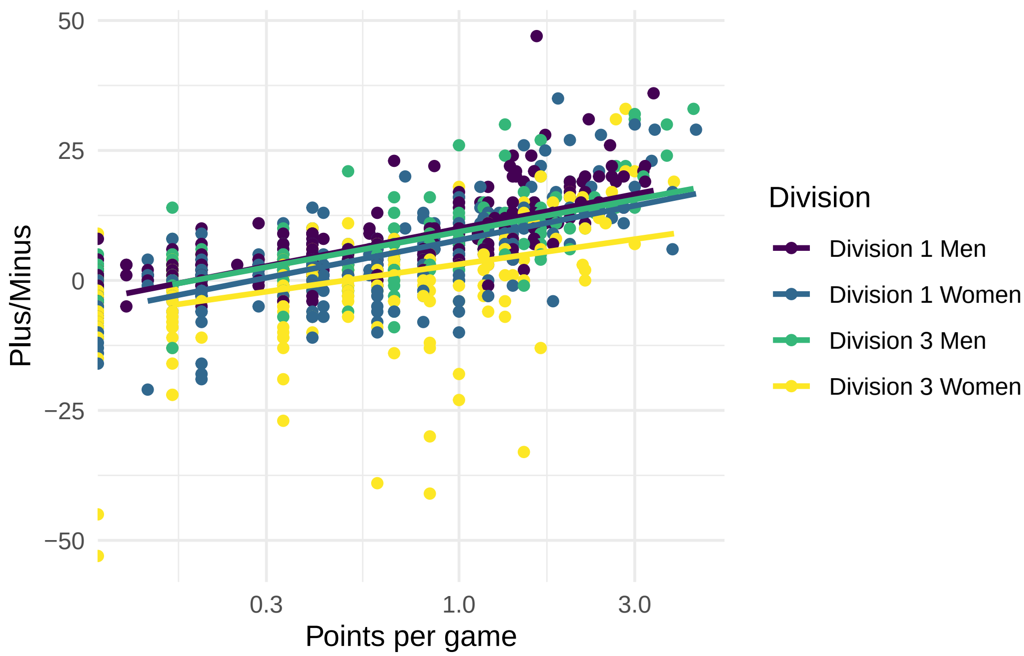
```r
ggplot(ultimate_data, aes(x = gender, y = plus_minus, fill = gender)) +
  geom_boxplot() +
  labs(
    title = "Gender vs. Player's +/- Score",
    x = "Gender",
    y = "Player's +/- Score"
  ) +
  theme_minimal()
```



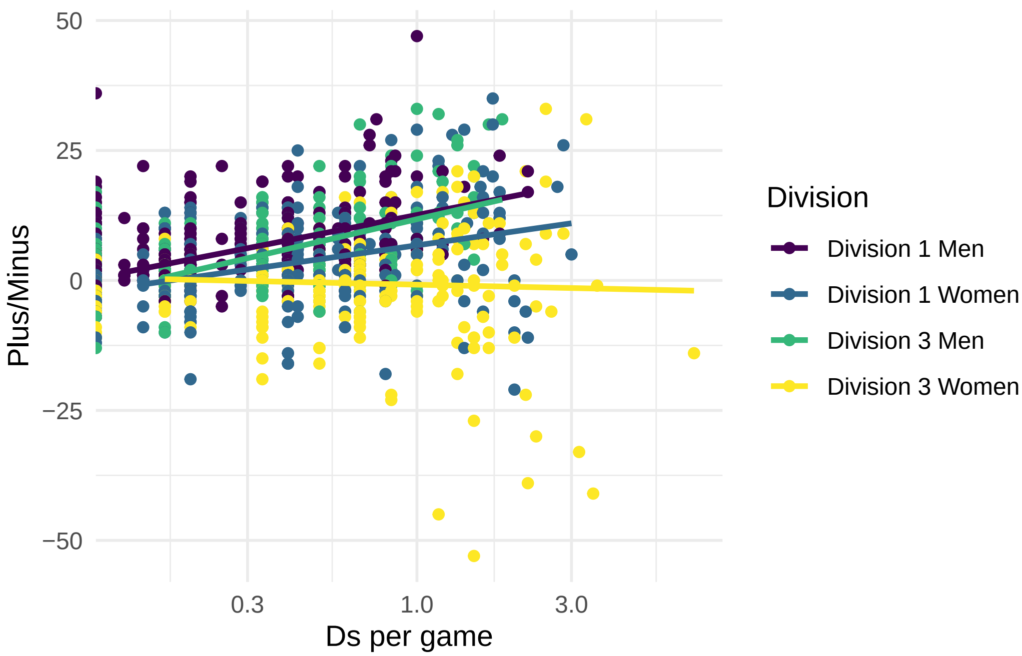

```r
ggplot(ultimate_data, aes(x = level, y = plus_minus, fill = level)) +
  geom_violin(trim = FALSE) +
  labs(
    title = "Player's +/- Score for Division 1 vs Division 3",
    x = "Division",
    y = "Player's +/- Score"
  ) +
  scale_fill_manual(values = c("Division 1" = "blue", "Division 3" = "pink")) +
  theme_minimal()
```



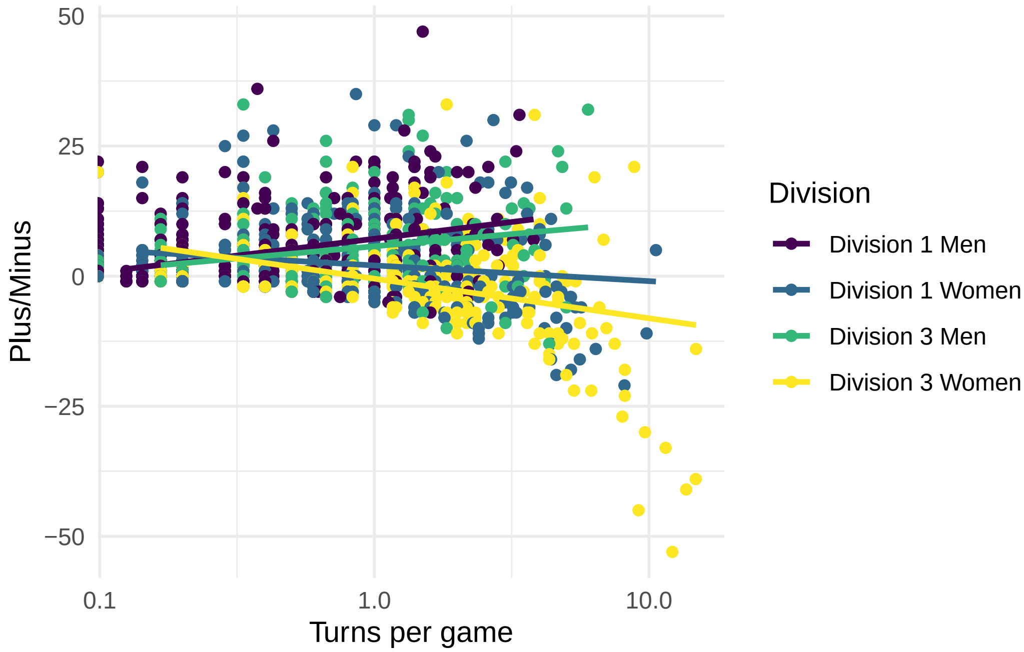

```
ultimate_data %>% ggplot(aes(x = pts_per_game, y = plus_minus, color = division)) +
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Points per game", y = "Plus/Minus", color = "Division") +
  theme_minimal() +
  scale_color_viridis_d()
```



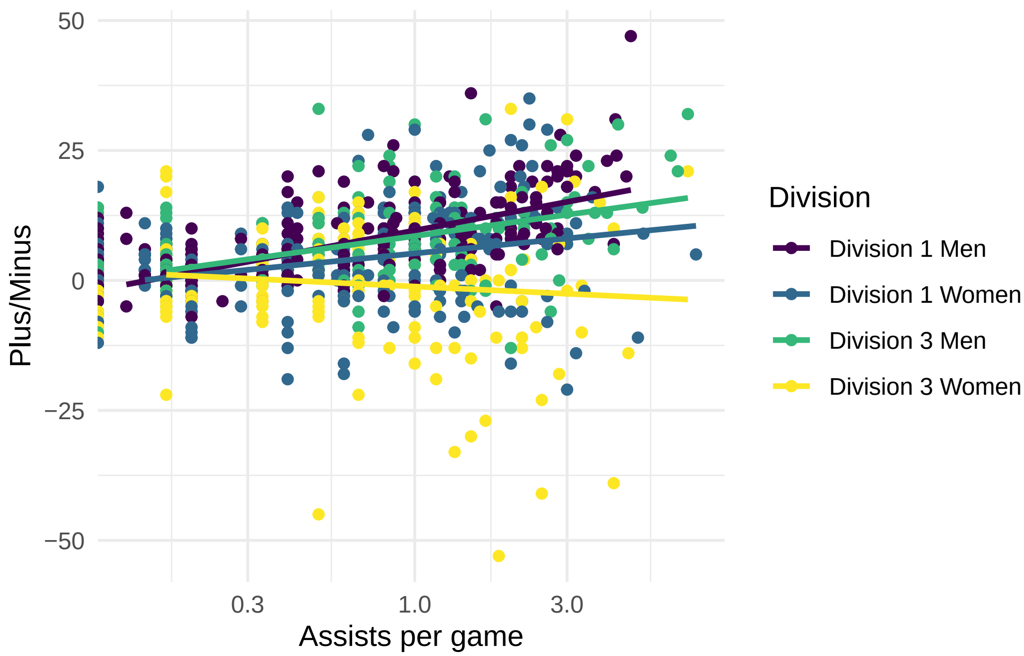

```r
ultimate_data %>% ggplot(aes(x = ds_per_game, y = plus_minus, color = division)) +
   geom_point() + geom_smooth(method = 'lm', se = F)+
  theme_minimal() + scale_x_log10() +
  labs(x = "Ds per game", y = "Plus/Minus", color = "Division") +
  scale_color_viridis_d()
```




```
ultimate_data %>% ggplot(aes(x = turns_per_game, y = plus_minus, color = division)) +
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Turns per game", y = "Plus/Minus", color = "Division") + theme_minimal() +
  scale_color_viridis_d()
```



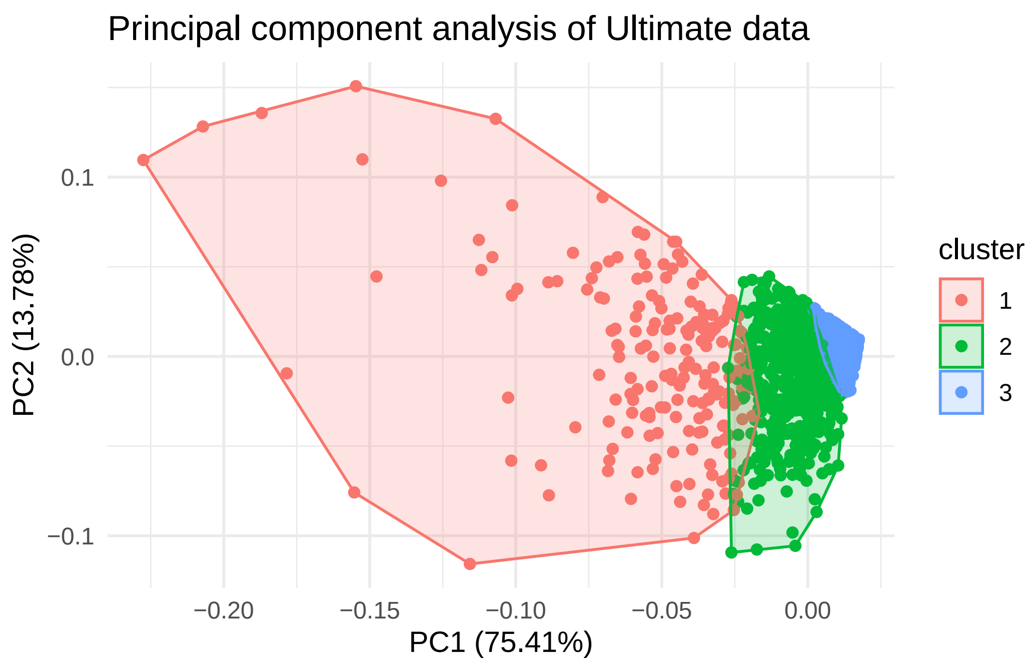

```
ultimate_data %>% ggplot(aes(x = ast_per_game, y = plus_minus, color = division)) +
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Assists per game", y = "Plus/Minus", color = "Division") + theme_minimal() +
  scale_color_viridis_d()
```




```r
df1 <- ultimate_data %>% select(c(
  turns_per_game, ds_per_game, pts_per_game, pls_mns_per_game, ast_per_game
))

pca <- (princomp(df1))

autoplot(pam(df1[-4], 3), frame = TRUE) + theme_minimal() +
  labs(frame = "Cluster", title = "Principal component analysis of Ultimate data")
```


Principal component analysis of Ultimate data
0.1
0.0
−0.1
PC2 (13.78%)
−0.20
−0.15
−0.10
−0.05
0.00
PC1 (75.41%)
cluster
1
2
3