Final Project Exploratory Data Analysis

AUTHOR
Julia Vitale, Mia Zottoli, and Vishnu Lakshman

```
# Load necessary libraries
library(here)
library(tidyverse)
library(ggplot2)
library(corrr)
library(ggcorrplot)
library(FactoMineR)
library(factoextra)
library(ggfortify)
library(cluster)
```

```
# Load dataset
ultimate_data <- read_csv(here("data", "ultimate_college_championship.csv")) %
# Standardise school names
  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 Carol
  mutate(school = ifelse(str_detect(team_name, "Cal Poly"), "Cal Poly SLO", sc
  mutate(school = ifelse(str_detect(team_name, "British Columbia"), "British C
  mutate(school = ifelse(str_detect(team_name, "Western Washington"), "Western
  mutate(school = ifelse(str_detect(team_name, "Penn State"), "Penn State", sc
  mutate(school = ifelse(str_detect(team_name, "San Diego"), "UCSD", school))
  mutate(school = ifelse(str_detect(team_name, "Lewis"), "Lewis & Clark", scho
  mutate(school = ifelse(str_detect(team_name, "Colorado State"), "Colorado St
  mutate(school = ifelse(str_detect(team_name, "Oklahoma Christian"), "Oklahom
  mutate(school = ifelse(str_detect(team_name, "Binghamton"), "SUNY Binghamton
  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
  mutate(school = ifelse(str_detect(team_name, "Missouri S"), "Missouri S&T",
  mutate(school = ifelse(str_detect(team_name, "Holyoke"), "Mount Holyoke", sc
  mutate(school = ifelse(str_detect(team_name, "NC State"), "NC State", school
  mutate(school = ifelse(str_detect(team_name, "Oregon State"), "Oregon State"
  mutate(school = ifelse(str_detect(team_name, "Washington University"), "Wash
```

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.

```
# A table of summarizing statistics of the variables we are planning to use
ultimate_data %>% select(-c(player, team_name)) %>% gtsummary::tbl_summary()
```

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)

[1] n (%); Median (Q1, Q3)

Characteristic	N = 1,665[1]
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%)

[1] n (%); Median (Q1, Q3)

Characteristic	N = 1,665[1]
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%)
Minnesota	28 (1.7%)
Missouri S&T	15 (0.9%)
Mount Holyoke	12 (0.7%)
NC State	30 (1.8%)
North Carolina	61 (3.7%)
Oberlin	22 (1.3%)
Occidental	23 (1.4%)
Oklahoma Christian	24 (1.4%)
Oregon	51 (3.1%)
Oregon State	22 (1.3%)
Ottawa	16 (1.0%)
Penn State	24 (1.4%)
Pennsylvania	22 (1.3%)
Pittsburgh	28 (1.7%)
Portland	24 (1.4%)
Richmond	39 (2.3%)
Rochester	25 (1.5%)
St Olaf	51 (3.1%)
Stanford	22 (1.3%)
SUNY Binghamton	22 (1.3%)
Texas	26 (1.6%)
Tufts	21 (1.3%)

[1] n (%); Median (Q1, Q3)

Characteristic	N = 1,665[1]
UCSB	23 (1.4%)
UCSD	25 (1.5%)
Union	18 (1.1%)
USCS	21 (1.3%)
Utah	22 (1.3%)
Vermont	46 (2.8%)
Victoria	20 (1.2%)
Washington	26 (1.6%)
Washington University	29 (1.7%)
Wellesley	20 (1.2%)
Wesleyan	22 (1.3%)
Western Washington	23 (1.4%)
Whitman	22 (1.3%)
Williams	26 (1.6%)

[1] n (%); Median (Q1, Q3)

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?
    - **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' 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

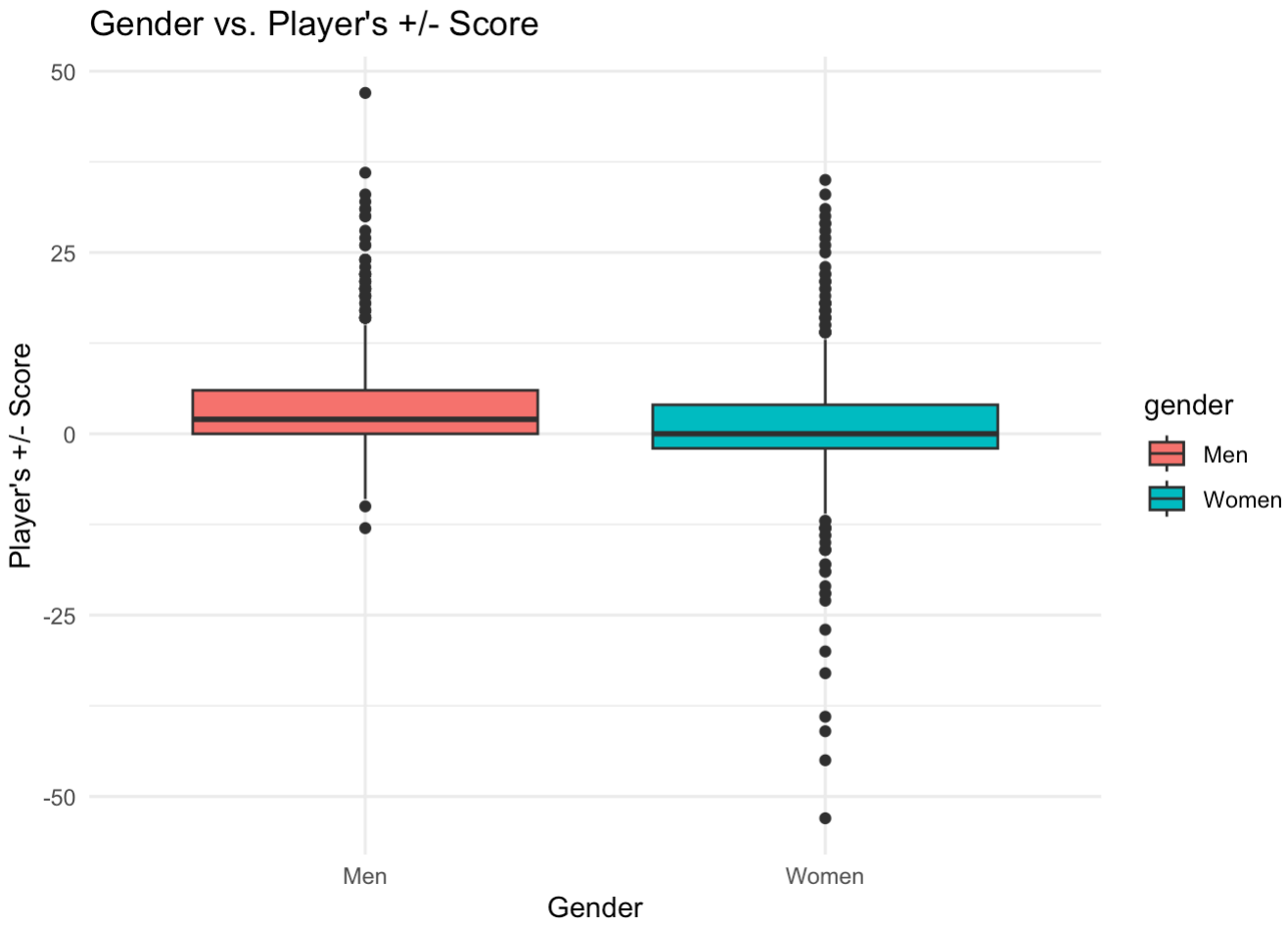
We have created a codebook in the README.md page in our repo on GitHub which contains a description of all the variables we are using.

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.

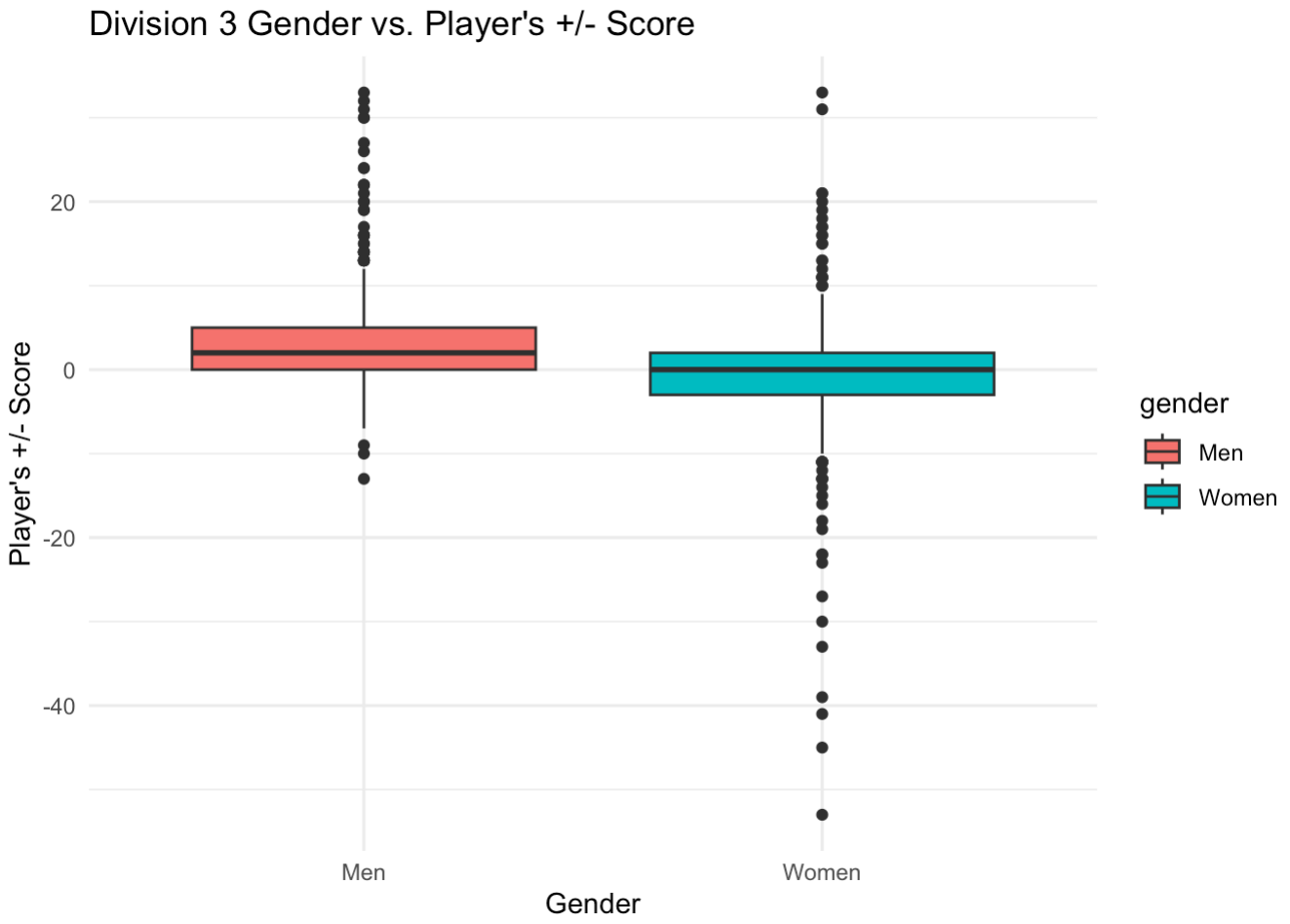
Data Visualization

We have 11 data visualizations of our data made in R. Our outcome variable, which is `plus_minus`, is used in multiple visualizations. There are visualizations involving various different data types (categorical, numerical, continuous and discrete).

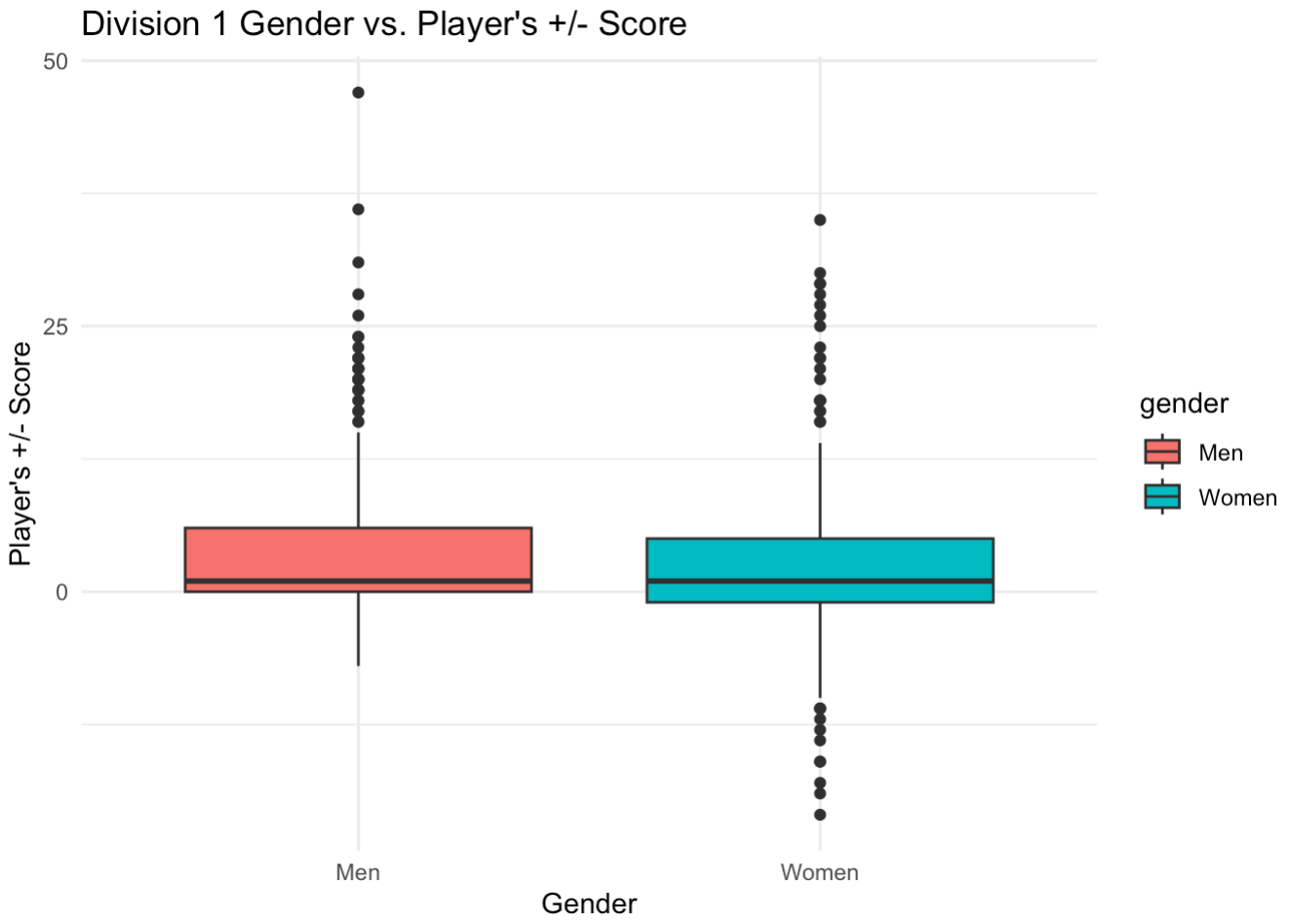
```
# Plot player's plus-minus scores and their gender
ggplot(ultimate_data, aes(x = gender, y = plus_minus, fill = gender)) +
  geom_boxplot() +
  labs(
    title = "Gender vs. Player's +/- Score",
    x = "Gender",
    y = "Player's +/- Score"
  ) +
  theme_minimal()
```



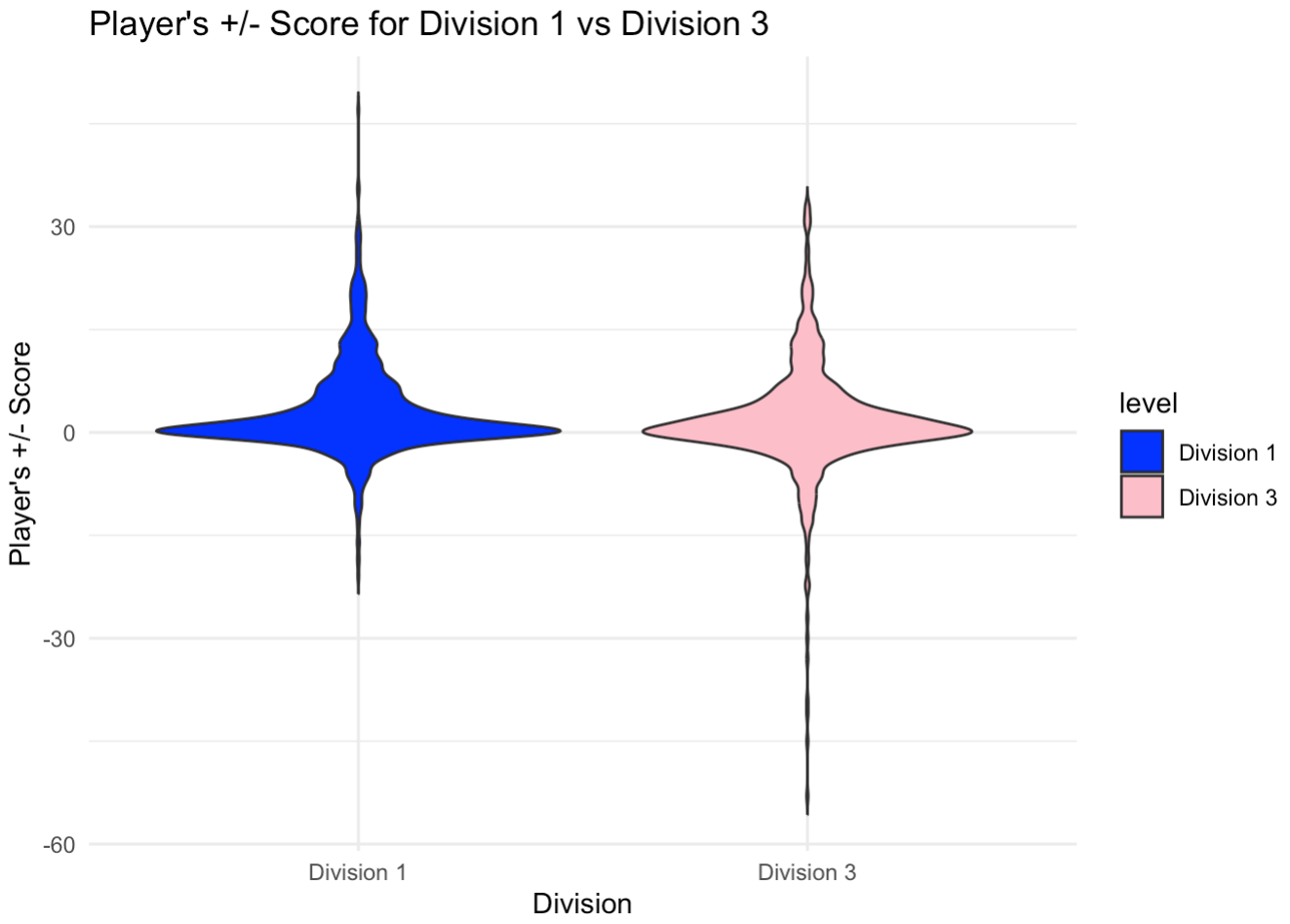

```
# Filter table to create a new table called d3 that only contains d3 level pla
d3 <- ultimate_data[ultimate_data$level == "Division 3",]

# Plot  Division 3 level player's plus-minus scores and their gender
ggplot(d3, aes(x = gender, y = plus_minus, fill = gender)) +
  geom_boxplot() +
  labs(
    title = "Division 3 Gender vs. Player's +/- Score",
    x = "Gender",
    y = "Player's +/- Score"
  ) +
  theme_minimal()
```



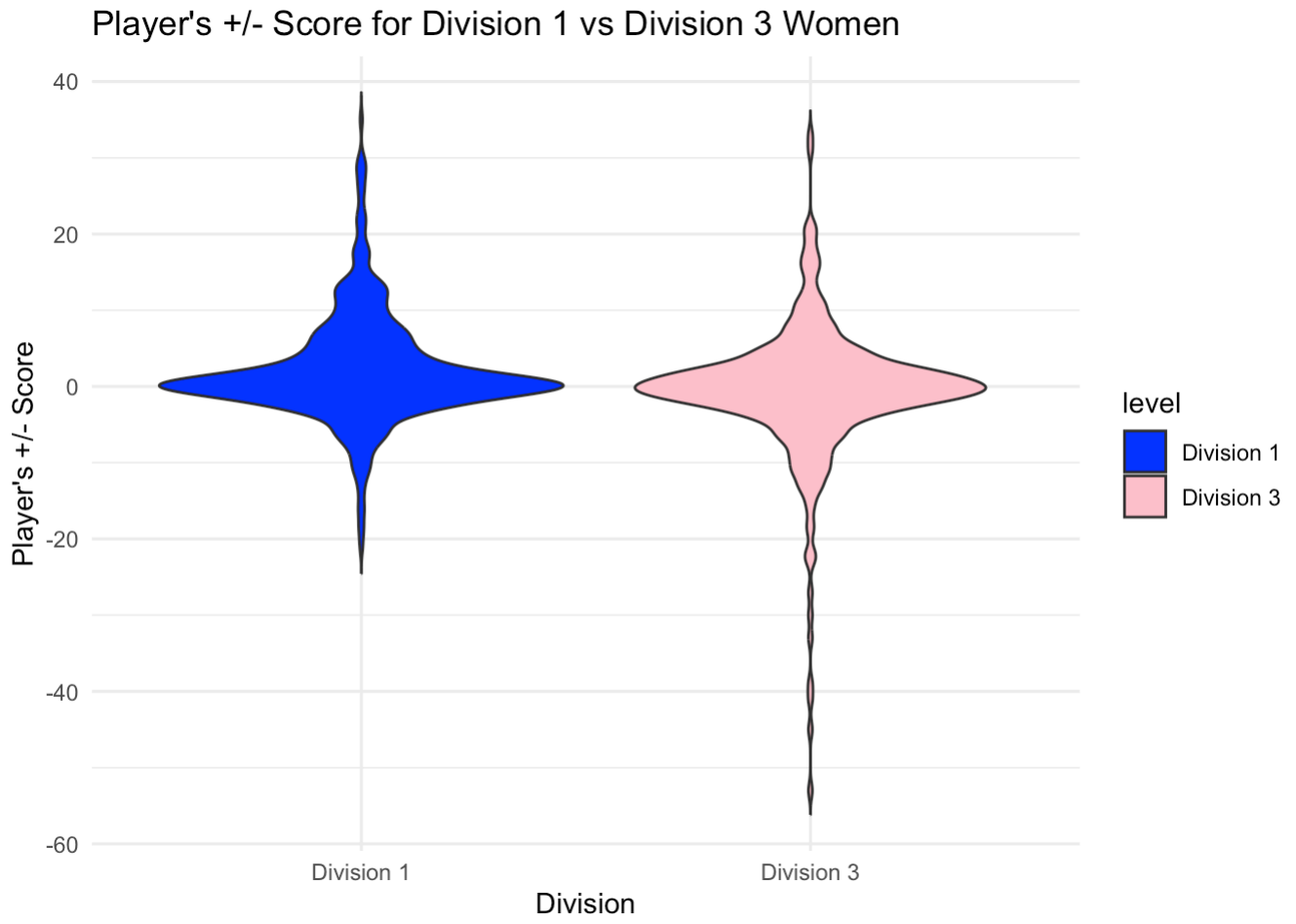

```
# Filter table to create a new table called d1 that only contains d3 level pla
d1 <- ultimate_data[ultimate_data$level == "Division 1",]

# Plot  Division 1 level player's plus-minus scores and their gender
ggplot(d1, aes(x = gender, y = plus_minus, fill = gender)) +
  geom_boxplot() +
  labs(
    title = "Division 1 Gender vs. Player's +/- Score",
    x = "Gender",
    y = "Player's +/- Score"
  ) +
  theme_minimal()
```



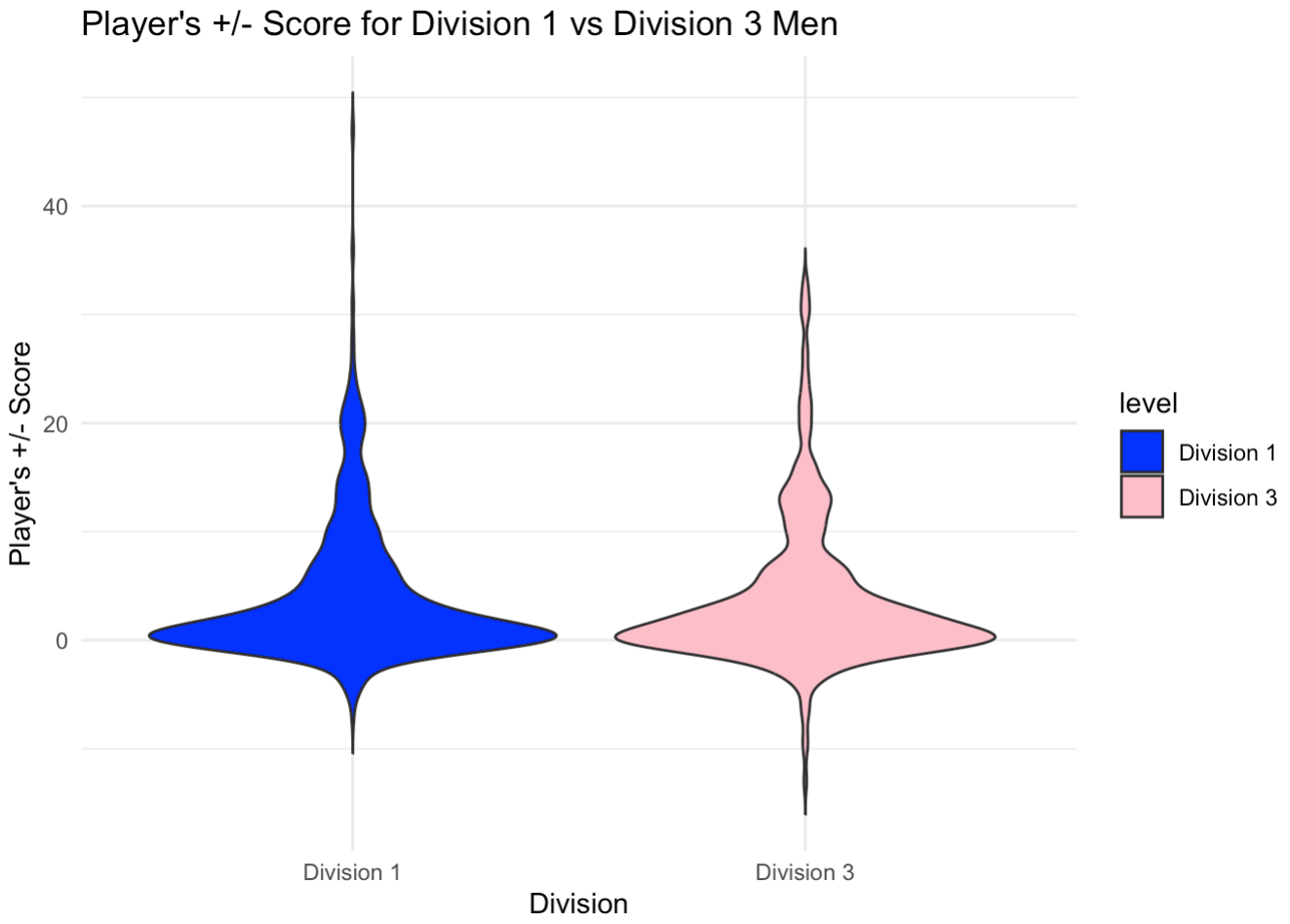

```
# Plot player's plus-minus score against the divisional level they are playing
ggplot(ultimate_data, aes(x = level, y = plus_minus, fill = level)) +
  geom_violin(trim = FALSE) +
  labs(
    title = "Player's +/- Score for Division 1 vs Division 3",
    x = "Division",
    y = "Player's +/- Score"
  ) +
  scale_fill_manual(values = c("Division 1" = "blue", "Division 3" = "pink")) +
  theme_minimal()
```



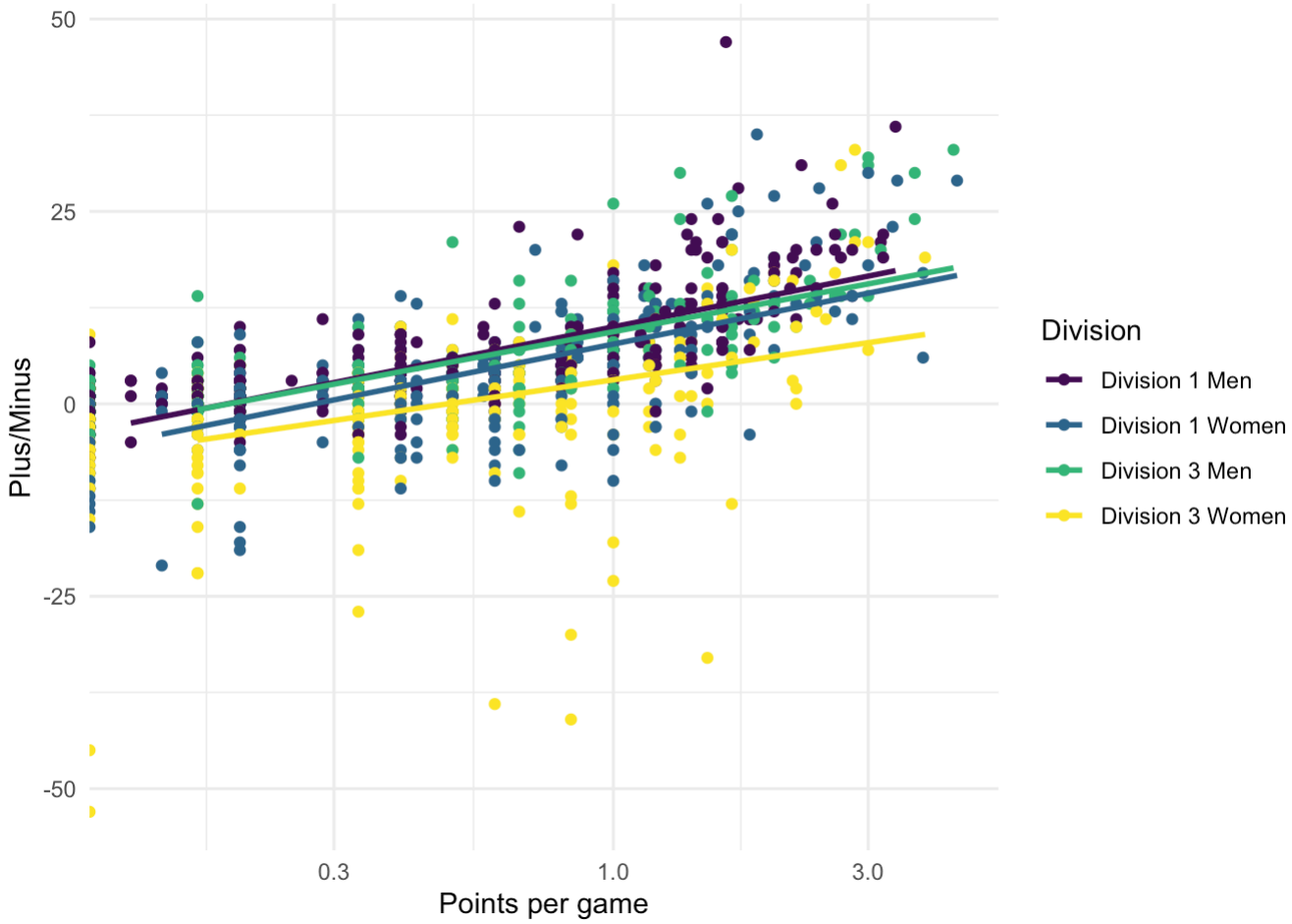

```r
# Filter table to create a new table called women_ultimate_data that only cont
women_ultimate_data <- ultimate_data %>%
  filter(gender == "Women")

# Plot female player's plus-minus score against the divisional level they are
ggplot(women_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 Women",
    x = "Division",
    y = "Player's +/- Score"
  ) +
  scale_fill_manual(values = c("Division 1" = "blue", "Division 3" = "pink"))
  theme_minimal()
```



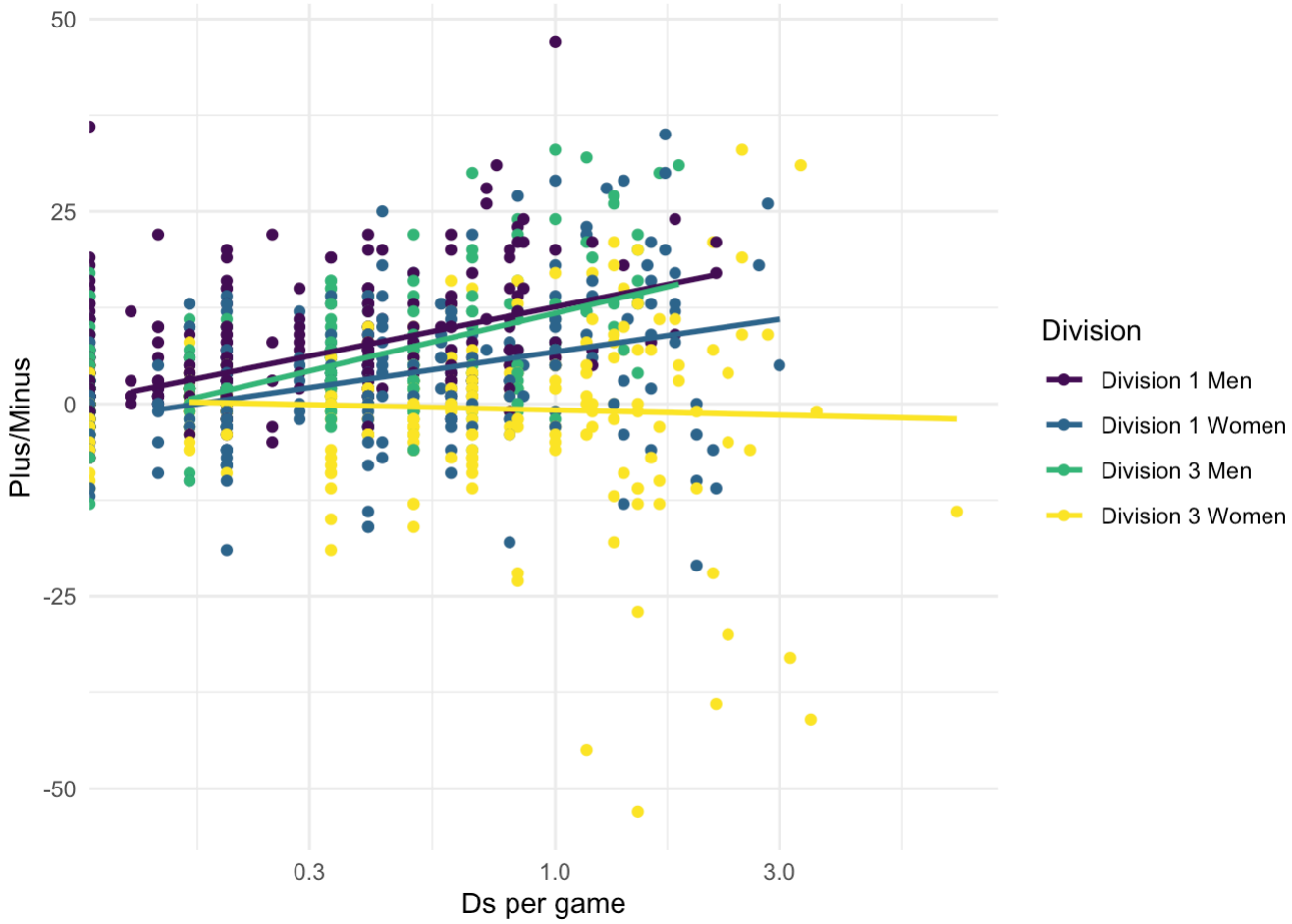

```
# Filter table to create a new table called men_ultimate_data that only contai
men_ultimate_data <- ultimate_data %>%
  filter(gender == "Men")

# Plot male player's plus-minus score against the divisional level they are pl
ggplot(men_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 Men",
    x = "Division",
    y = "Player's +/- Score"
  ) +
  scale_fill_manual(values = c("Division 1" = "blue", "Division 3" = "pink"))
  theme_minimal()
```



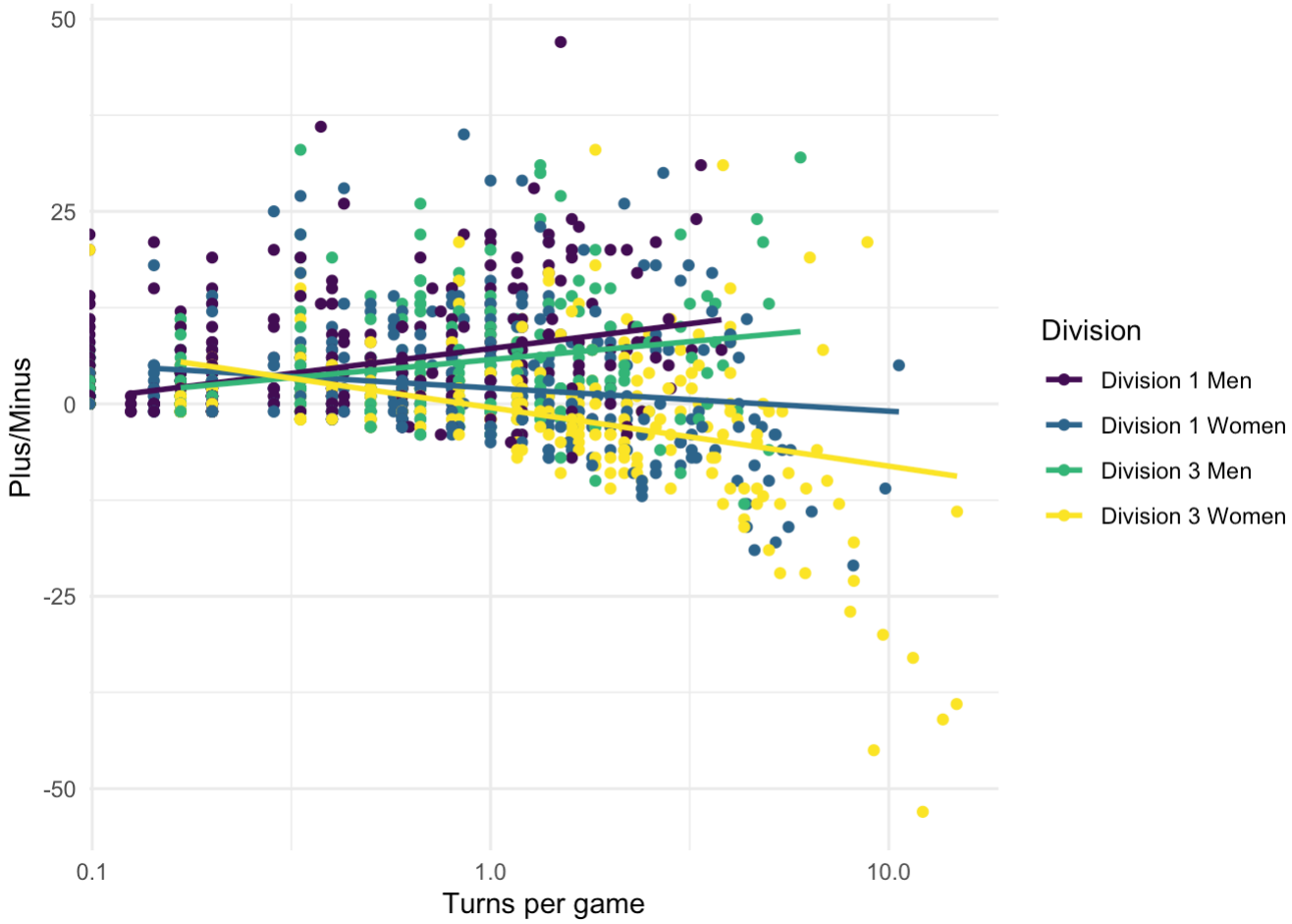

```
# Plot a player's plus-minus score and their points per game for the different
ultimate_data %>% ggplot(aes(x = pts_per_game, y = plus_minus, color = divisio
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Points per game", y = "Plus/Minus", color = "Division") +
  theme_minimal() +
  scale_color_viridis_d()
```



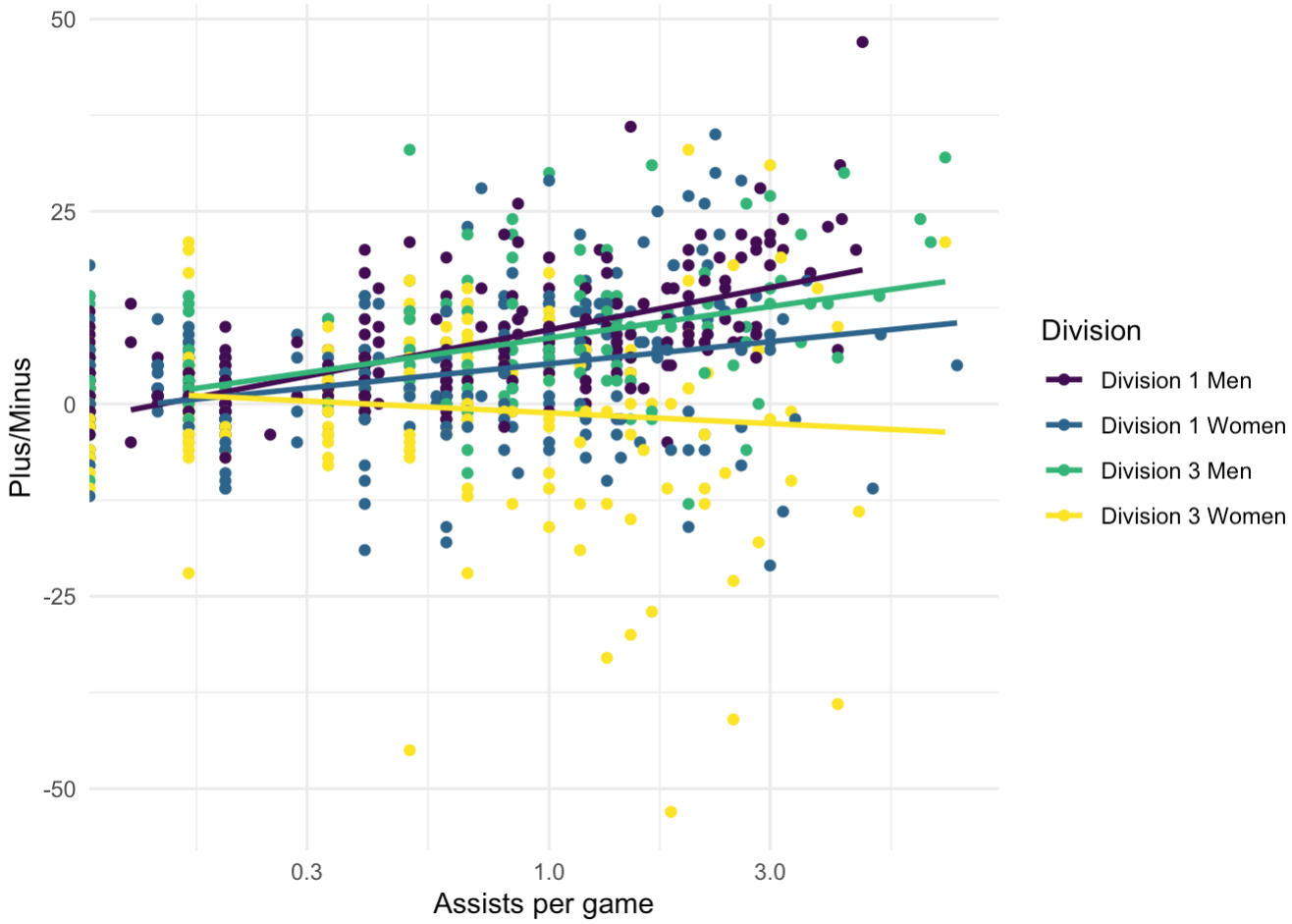

```
# Plot a player's plus-minus score and their D's per game for the different di
ultimate_data %>% ggplot(aes(x = ds_per_game, y = plus_minus, color = divisior
   geom_point() + geom_smooth(method = 'lm', se = F)+
  theme_minimal() + scale_x_log10() +
  labs(x = "Ds per game", y = "Plus/Minus", color = "Division") +
  scale_color_viridis_d()
```



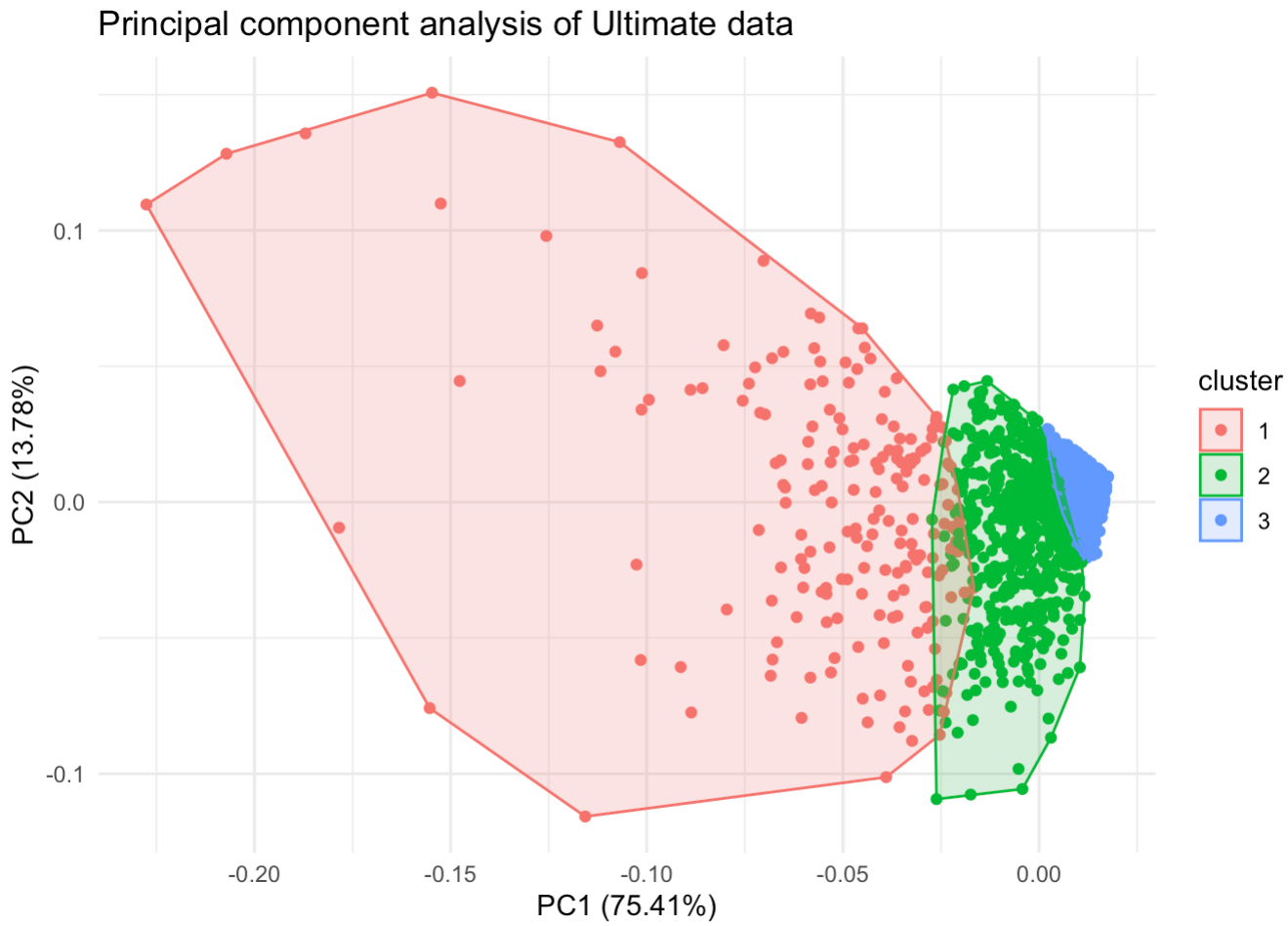

```
# Plot a player's plus-minus score and their turns per game for the different
ultimate_data %>% ggplot(aes(x = turns_per_game, y = plus_minus, color = divis
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Turns per game", y = "Plus/Minus", color = "Division") + theme_mir
  scale_color_viridis_d()
```




```
# Plot a player's plus-minus score and their assists per game for the differer
ultimate_data %>% ggplot(aes(x = ast_per_game, y = plus_minus, color = divisic
   geom_point() + geom_smooth(method = 'lm', se = F) + scale_x_log10() +
  labs(x = "Assists per game", y = "Plus/Minus", color = "Division") + theme_m
  scale_color_viridis_d()
```




```
# Create a dataset df1 by selecting the following the columns below
df1 <- ultimate_data %>% select(c(
  turns_per_game, ds_per_game, pts_per_game, pls_mns_per_game, ast_per_game
))

# Perform principle component analysis on df1
pca <- (princomp(df1))

# Plots the three clusters
autoplot(pam(df1[-4], 3), frame = TRUE) + theme_minimal() +
  labs(frame = "Cluster", title = "Principal component analysis of Ultimate da
```


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