Exploratory Data Analysis

STOR 320 - Group 15

November 03, 2023

Creator: Zaid Kamdar

Question 1. How does the Effective Field Goal Percentage (EFG_O, EFG_D) correlate with the win percentage for teams?

```
bball <- ball_data |>
  mutate(Win_Percentage = W/G)

q1_plot_offense <- ggplot(bball, aes(x = Win_Percentage, y = EFG_O)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(
    title = "Win Percentage vs. EFG_O",
    x = "Win_Percentage",
    y = "EFG_O"
  )
q1_plot_defense <- ggplot(bball, aes(x = Win_Percentage, y = EFG_D)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  labs(
    title = "Win Percentage vs. EFG_D",
    x = "Win_Percentage",
    y = "EFG_D"
  )
off_correlation <- cor(bball$Win_Percentage, bball$EFG_O)
def_correlation <- cor(bball$Win_Percentage, bball$EFG_D)
off_correlation
```

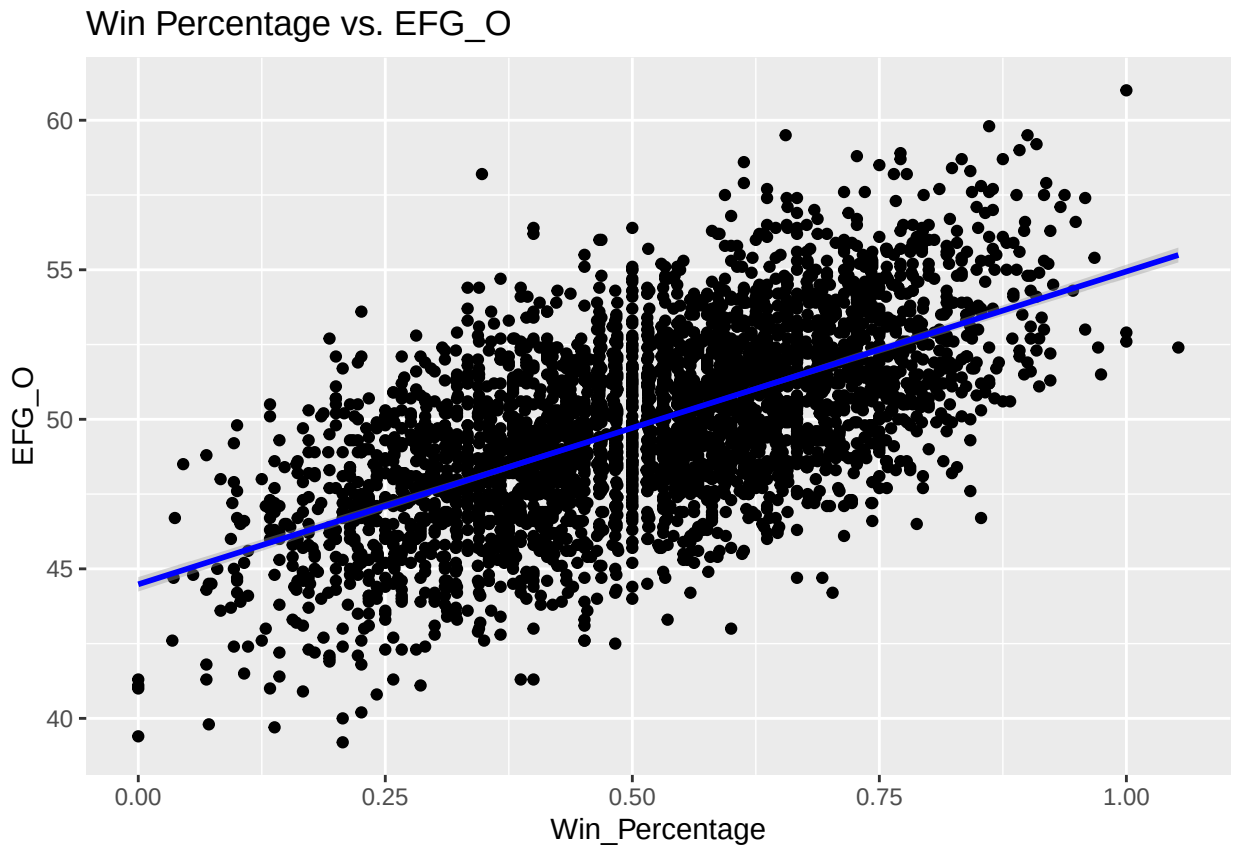
```
## [1] 0.6173164
```

```
def_correlation
```

```
## [1] -0.5741954
```

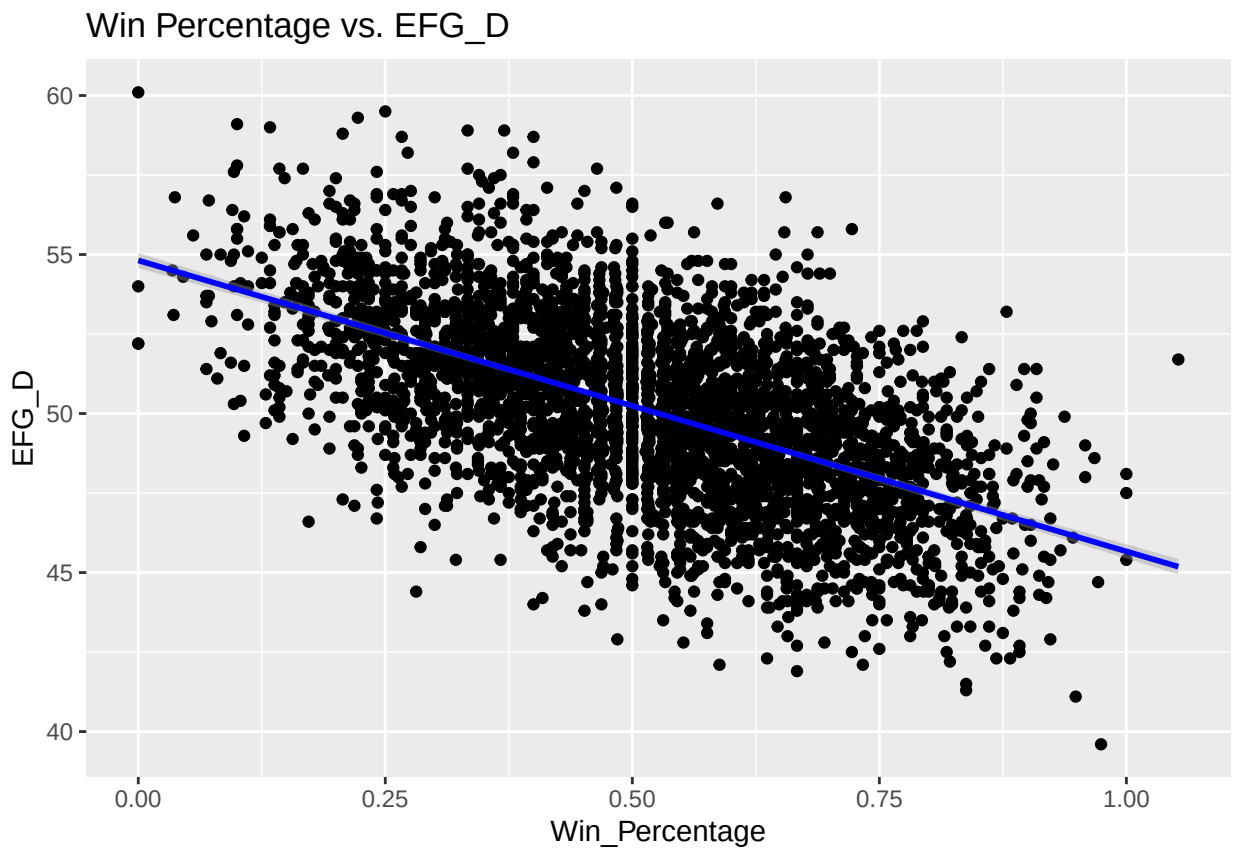
```
q1_plot_offense
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



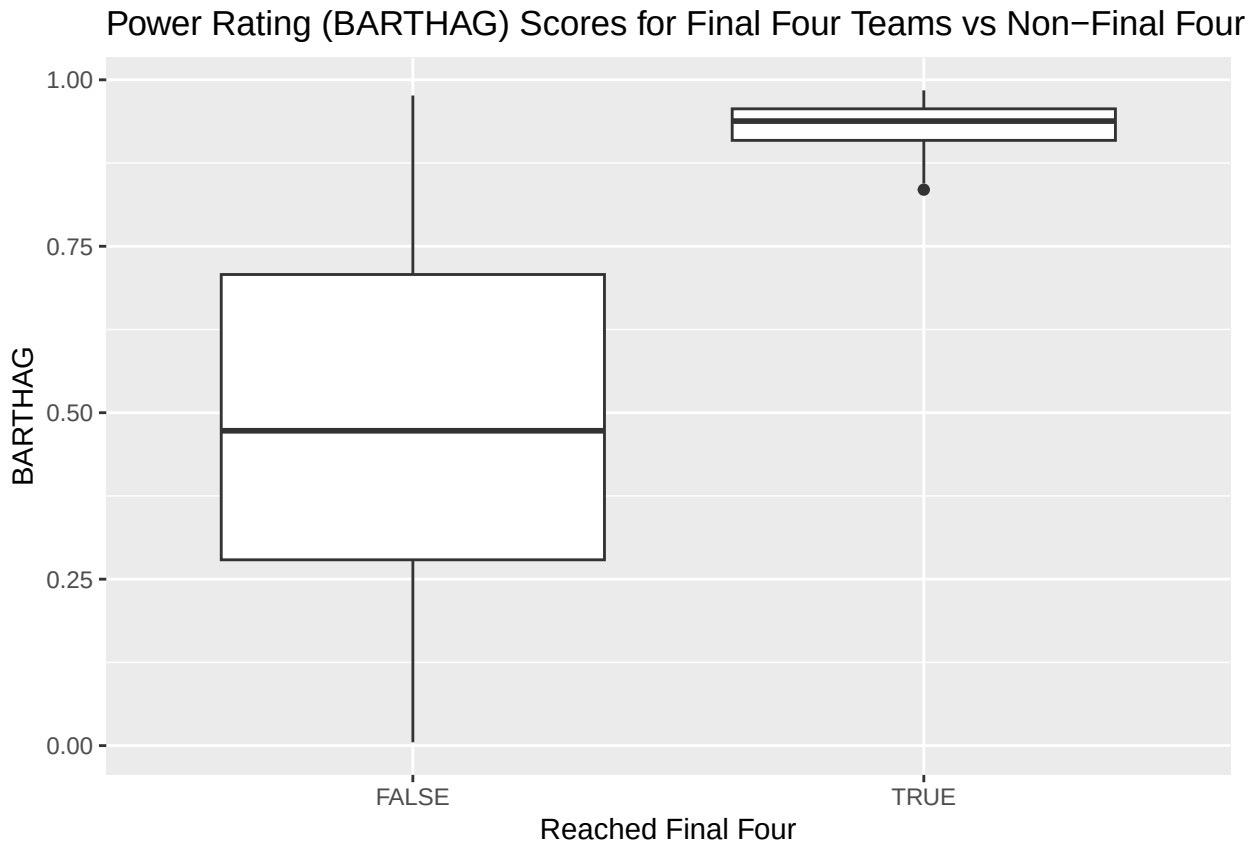

```
q1_plot_defense
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Question 2. Is the Power Rating (BARTHAG) an accurate indicator for a team reaching the final four in the postseason?

```
barthag_final_four <- ggplot(bball, aes(x = POSTSEASON %in% c("F4", "2ND", "Champion"), y = BARTHAG)) +
  geom_boxplot() +
  labs(
    title = "Power Rating (BARTHAG) Scores for Final Four Teams vs Non-Final Four Teams",
    x = "Reached Final Four",
    y = "BARTHAG"
  )
barthag_final_four
```

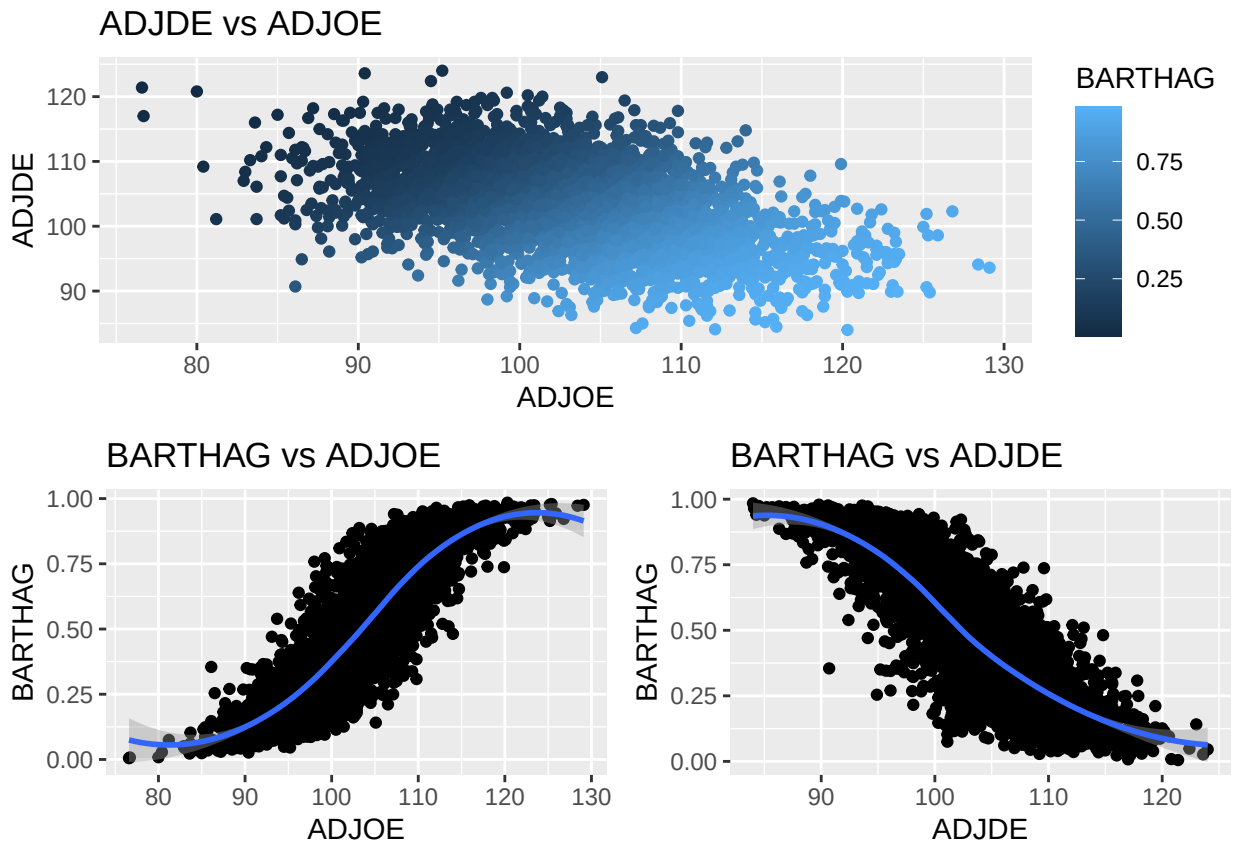


Interpreter: Rohan Phadke

Question 1. How do individual statistical categories (like ADJOE, ADJDE, etc.) contribute to a team's Barthag score (BARTHAG)?

```
adjoe_barthag_plot <-
  ggplot(ball_data, aes(x=ADJOE, y=BARTHAG)) +
  geom_point() +
  geom_smooth(method='loess') +
  labs(title="BARTHAG vs ADJOE", x="ADJOE", y="BARTHAG")
adjde_barthag_plot <-
  ggplot(ball_data, aes(x=ADJDE, y=BARTHAG)) +
  geom_point() +
  geom_smooth(method='loess') +
  labs(title="BARTHAG vs ADJDE", x="ADJDE", y="BARTHAG")
adjoe_vs_adjde_pot <-
  ggplot(ball_data, aes(x=ADJOE, y=ADJDE, color=BARTHAG)) +
  geom_point() +
  labs(title="ADJDE vs ADJOE", x="ADJOE", y="ADJDE")
graphic_layout = matrix(c(3, 1, 3, 2), ncol = 2)
grid.arrange(adjoe_barthag_plot, adjde_barthag_plot, adjoe_vs_adjde_pot, layout_matrix=graphic_layout)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

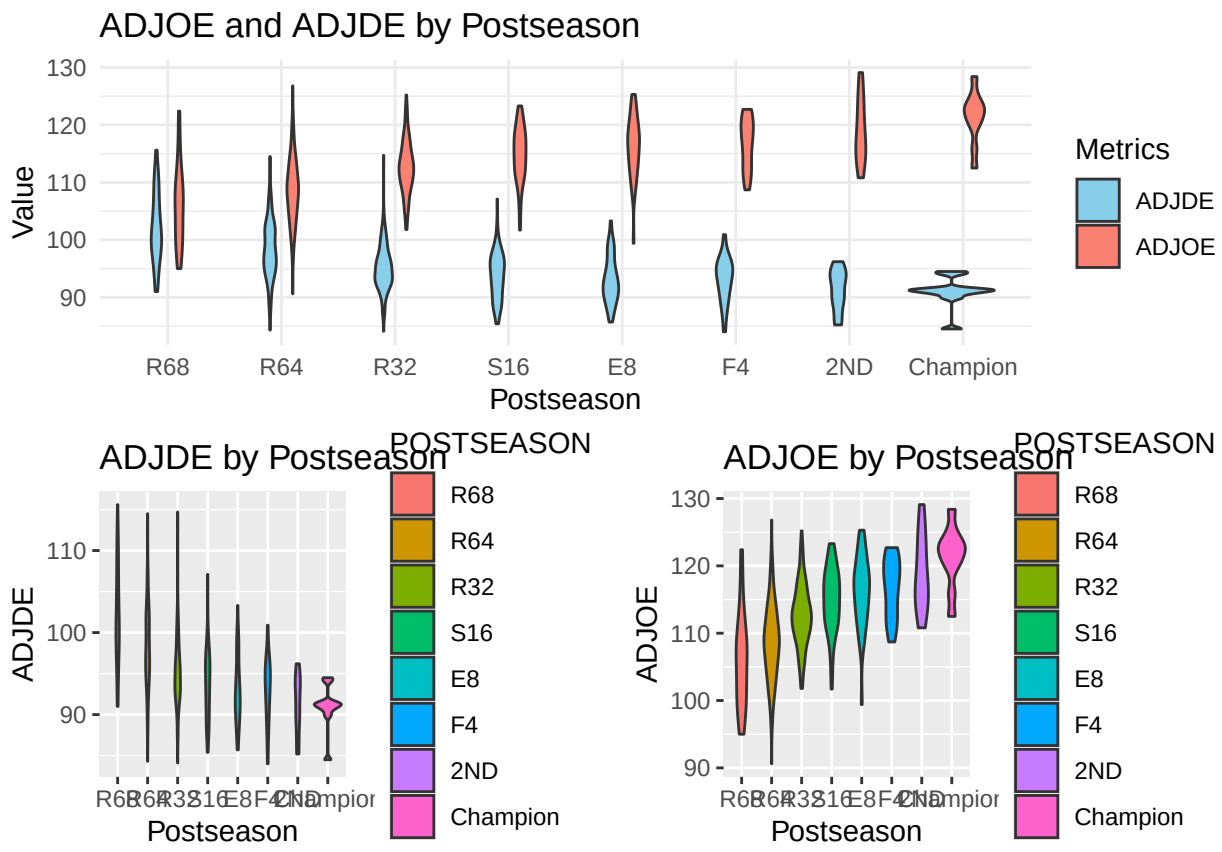


Question 2. Do ADJOE or ADJDE have an affect when determining postseason performance if at all?

```
bbData_postseason <-
  ball_data |>
  filter(!is.na(POSTSEASON) & POSTSEASON != "N/A")

bbData_postseason$POSTSEASON <- factor(bbData_postseason$POSTSEASON, levels = c("R68", "R64", "R32", "S
adjoe_plot <-
  ggplot(bbData_postseason, aes(x=POSTSEASON, y=ADJOE, fill=POSTSEASON)) +
  geom_violin() +
  labs(title="ADJOE by Postseason", x="Postseason", y="ADJOE")
adjde_plot <-
  ggplot(bbData_postseason, aes(x=POSTSEASON, y=ADJDE, fill=POSTSEASON)) +
  geom_violin() +
  labs(title="ADJDE by Postseason", x="Postseason", y="ADJDE")
data_long <-
  bbData_postseason |>
  select(POSTSEASON, ADJOE, ADJDE) |>
  pivot_longer(cols = c(ADJOE, ADJDE), names_to = "Metric", values_to = "Value")

combined_plot <-
  ggplot(data_long, aes(x=POSTSEASON, y=Value, fill=Metric)) +
  geom_violin(position=position_dodge(width=0.4), width=1.5) +
  labs(title="ADJOE and ADJDE by Postseason",
       x="Postseason",
       y="Value") +
```

```
  theme_minimal() +
  scale_fill_manual(values=c("skyblue", "salmon"), name="Metrics")

graphic_layout = matrix(c(3, 2, 3, 1), ncol = 2)
grid.arrange(adjoe_plot, adjde_plot, combined_plot,layout_matrix=graphic_layout)
```

```
## Warning: `position_dodge()` requires non-overlapping x intervals
```



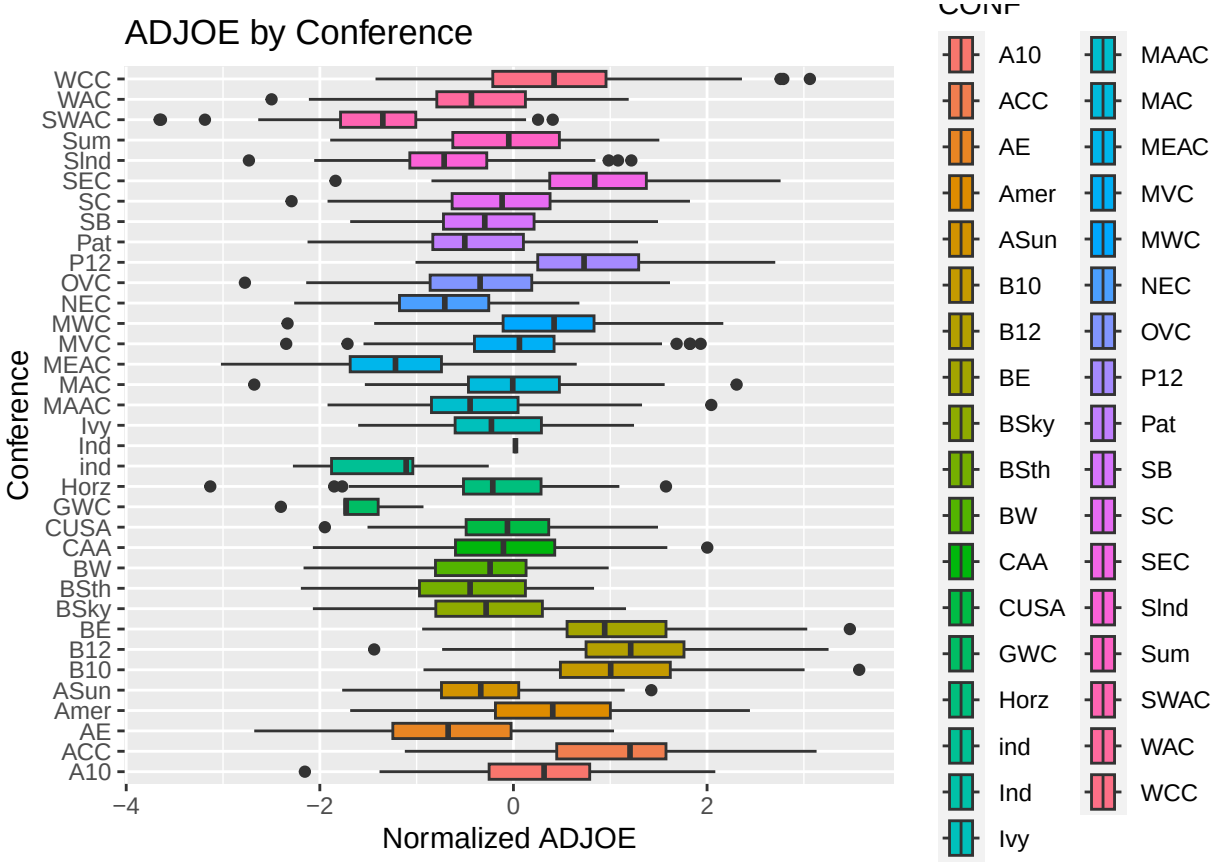
Orator: Mason Lennon

Question 1. Do teams from certain conferences (CONF) consistently outperform teams from other conferences in terms of Adjusted Offensive Efficiency (ADJOE)?

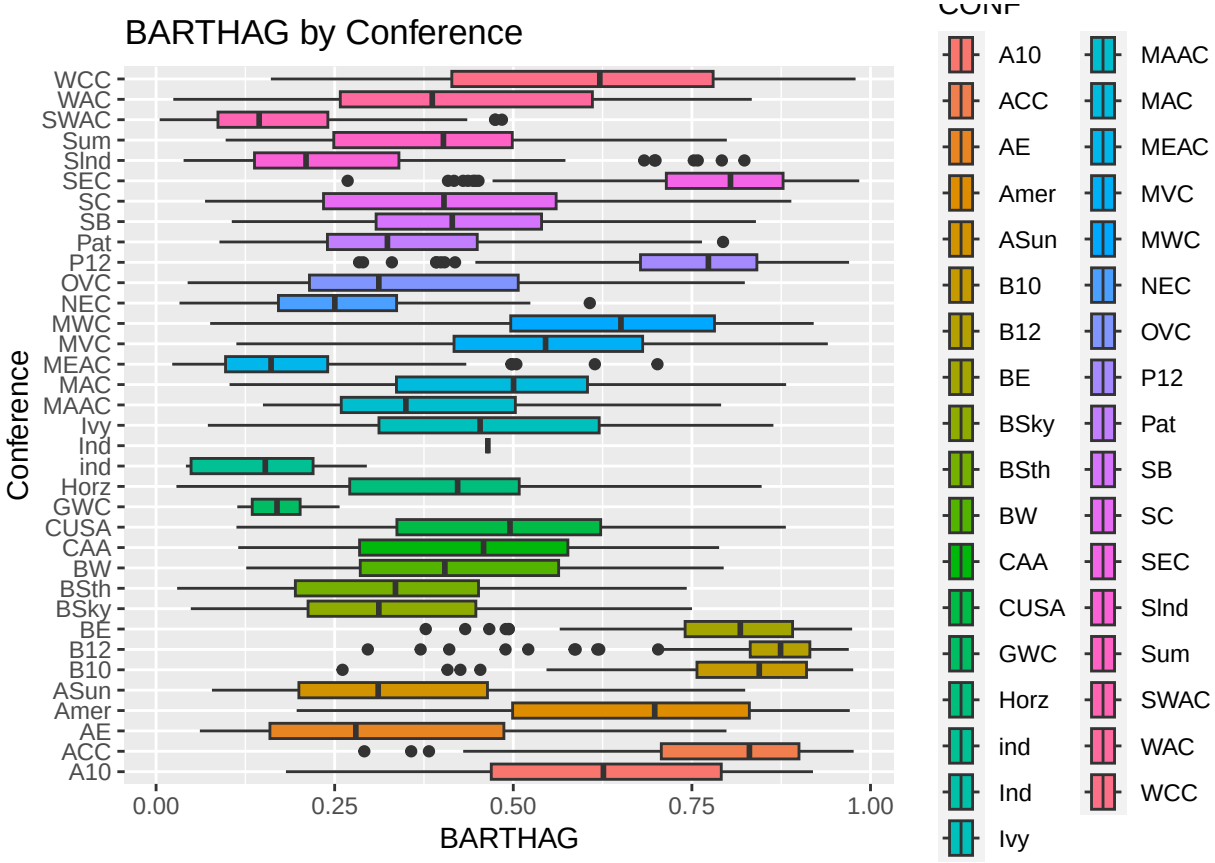
```
filtered_data <- ball_data |>
  select(CONF, ADJOE)

conference_summary <- filtered_data |>
  group_by(CONF) |>
  summarise(mean_ADJOE = mean(ADJOE)) |>
  arrange(mean_ADJOE, descending = TRUE)

kable(conference_summary)
```

CONF	mean_ADJOE
GWC	91.24000
SWAC	93.12212
ind	93.62000
MEAC	94.15478
NEC	97.78218
Slnd	98.31417
AE	98.39565
BSth	99.76396
Pat	100.17245
WAC	100.22842
MAAC	100.35688
ASun	100.67474
OVC	100.85948
BW	100.86701
BSky	100.96140
SB	101.35470
Horz	101.73824
SC	102.06408
Ivy	102.07639
CUSA	102.43212
Sum	102.53596
CAA	102.79417
MAC	103.22083
Ind	103.30000
MVC	103.55294
A10	105.13521
MWC	105.38519
Amer	106.33800
WCC	106.63737
P12	108.74917
SEC	109.34214
BE	110.65556
B10	111.10368
ACC	111.11905
B12	111.78700

```
ball_data |>
  mutate(Normalized_ADJOE = (ADJOE - mean(ADJOE)) / sd(ADJOE)) |>
  ggplot(aes(x=Normalized_ADJOE, y=CONF, fill=CONF)) +
  geom_boxplot() +
  labs(title="ADJOE by Conference", x="Normalized ADJOE", y="Conference")
```

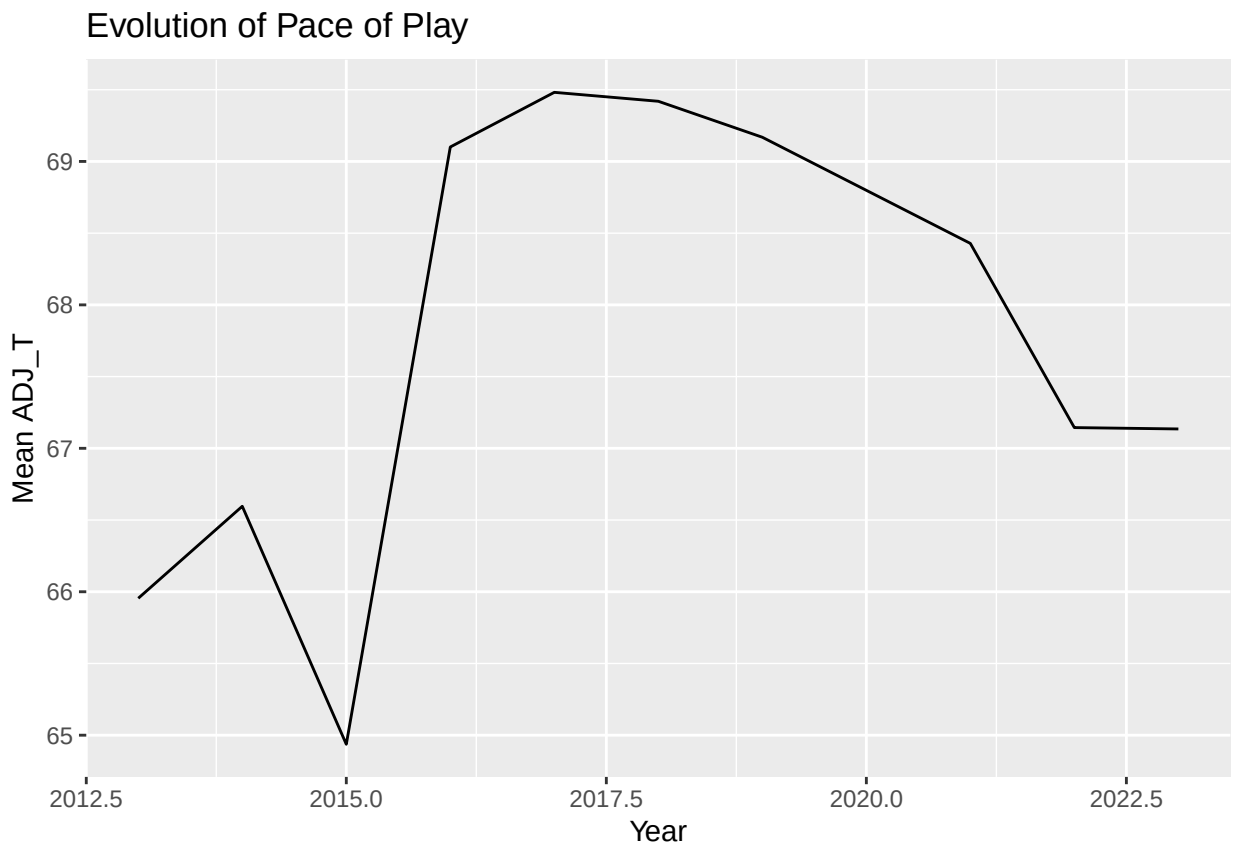



```
ball_data |>

ggplot(aes(x=BARTHAG, y=CONF, fill=CONF)) +
  geom_boxplot() +
  labs(title="BARTHAG by Conference", x="BARTHAG", y="Conference")
```



Question 2. How has the pace of play (ADJ_T) evolved over the years (YEAR)? Are teams playing faster or slower?

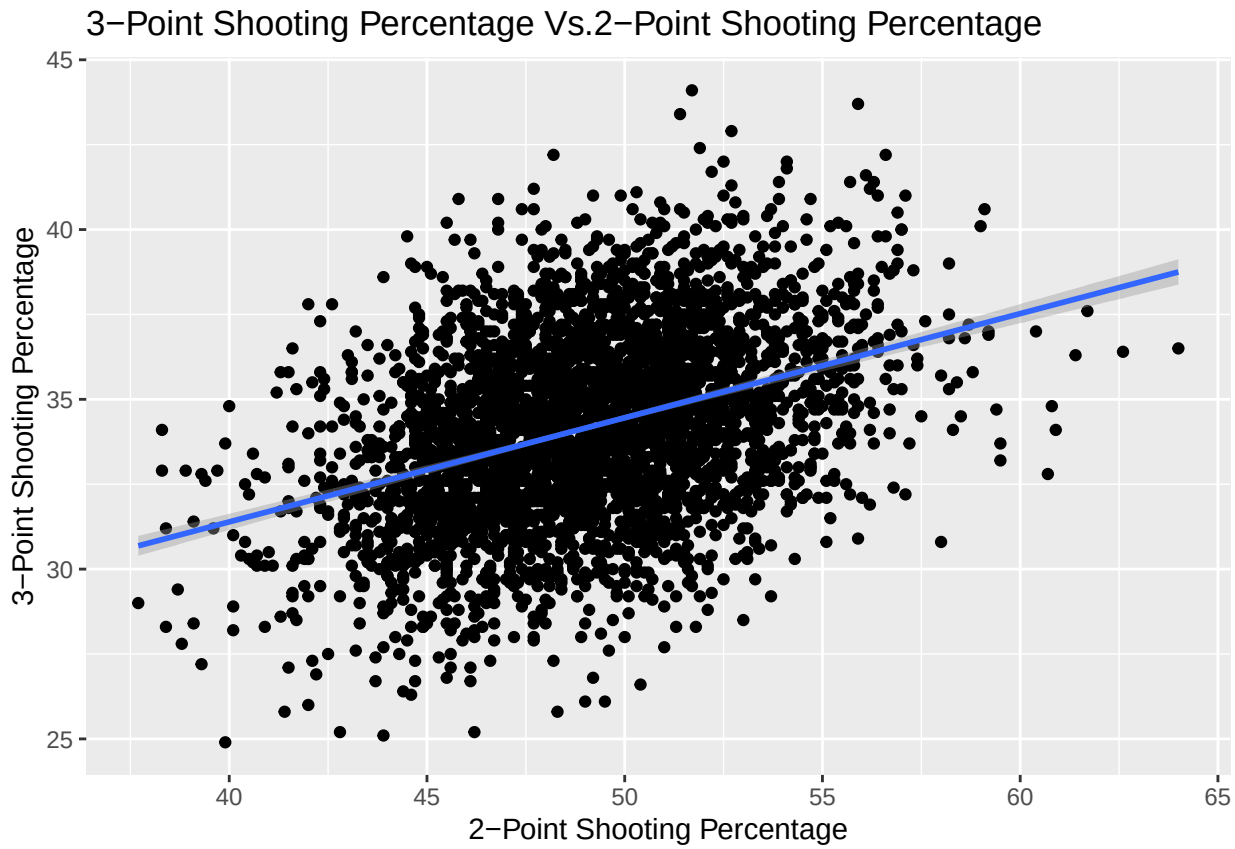
```
pace_summary <- ball_data |>
  group_by(YEAR) |>
  summarise(mean_ADJ_T = mean(ADJ_T))

ggplot(pace_summary, aes(x = YEAR, y = mean_ADJ_T)) +
  geom_line() +
  labs(x = "Year", y = "Mean ADJ_T", title = "Evolution of Pace of Play")
```



Deliverer: Nicholas Poon

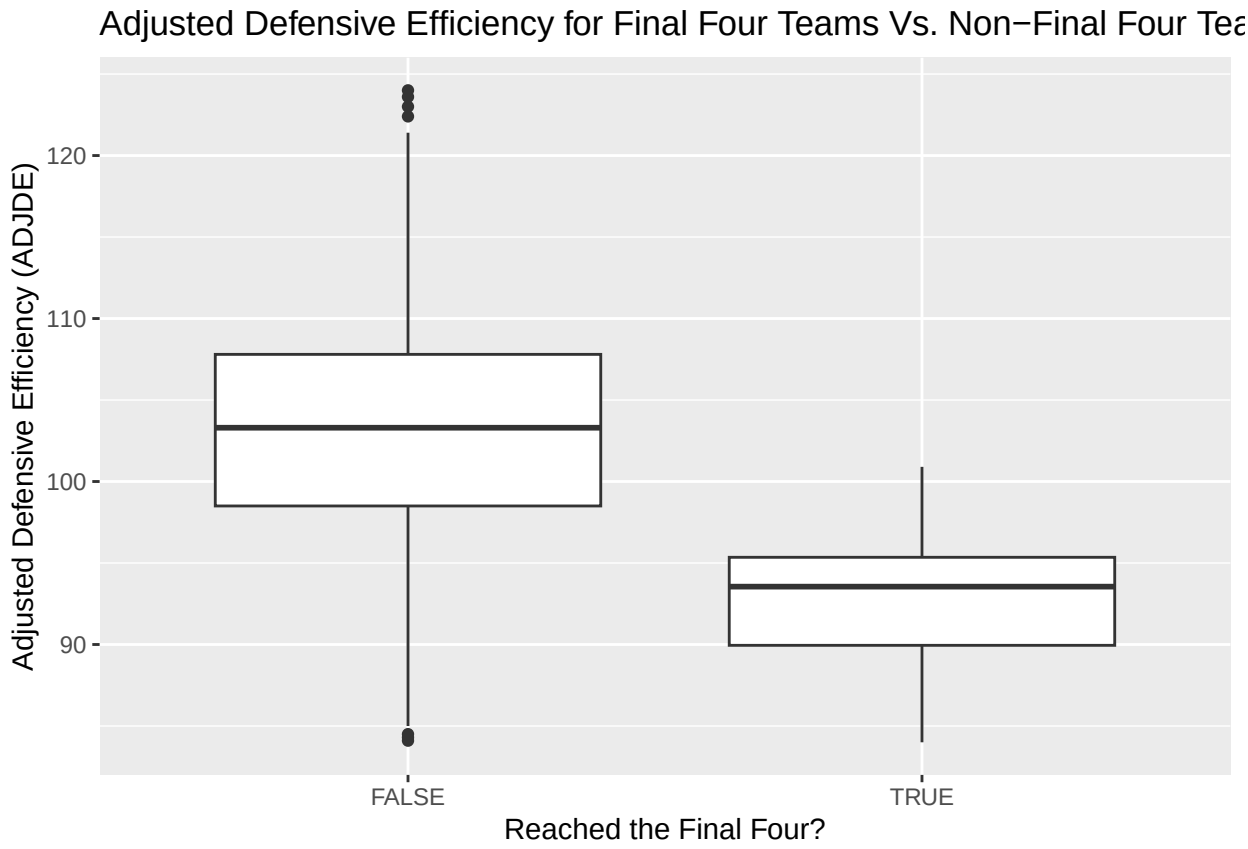
Question 1. Do teams with higher 2-point shooting percentages (2P_O) have lower 3-point shooting percentages (3P_O) or vice versa?

```
team_shooting_data <- ball_data|>
  select(TEAM, X2P_O, X3P_O)
team_shooting_data |>
  ggplot(mapping = aes(x= X2P_O, y= X3P_O)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y~x)+
  xlab("2-Point Shooting Percentage") +
  ylab("3-Point Shooting Percentage") +
  ggtitle("3-Point Shooting Percentage Vs.2-Point Shooting Percentage")
```



Question 2. Is Adjusted Defensive Efficiency (ADJDE) a reliable predictor for postseason success?

```
ADJDE_final_four <- ggplot(ball_data, aes(x = POSTSEASON %in% c("F4", "2ND", "Champion"), y = ADJDE)) +
  geom_boxplot() +
  xlab("Reached the Final Four?") +
  ylab("Adjusted Defensive Efficiency (ADJDE)") +
  ggtitle("Adjusted Defensive Efficiency for Final Four Teams Vs. Non-Final Four Teams")
ADJDE_final_four
```



Follow-up Questions

New Questions Based Off Initial Investigation

Question 1. Is it more important for a team to have a better ADJOE or ADJDE if they are trying to win a championship? If ADJOE is more important, which has more of an effect on Adjusted Offensive Efficiency, a team's 3-point percentage or their 2-point percentage?

Question 2. Where do turnover and rebound metrics rank compared to offense and defense metrics in predicting post-season success?

Question 3. Considering that EFG_O and EFG_D are an indicator in win percentage; and a higher Power Rating (BARTHAG) indicates a higher likelihood of a team reaching the final four, Is there a particular range for EFG_O and EFG_D that, when combined with a certain BARTHAG score, boosts a team's chances of making it to the final four?

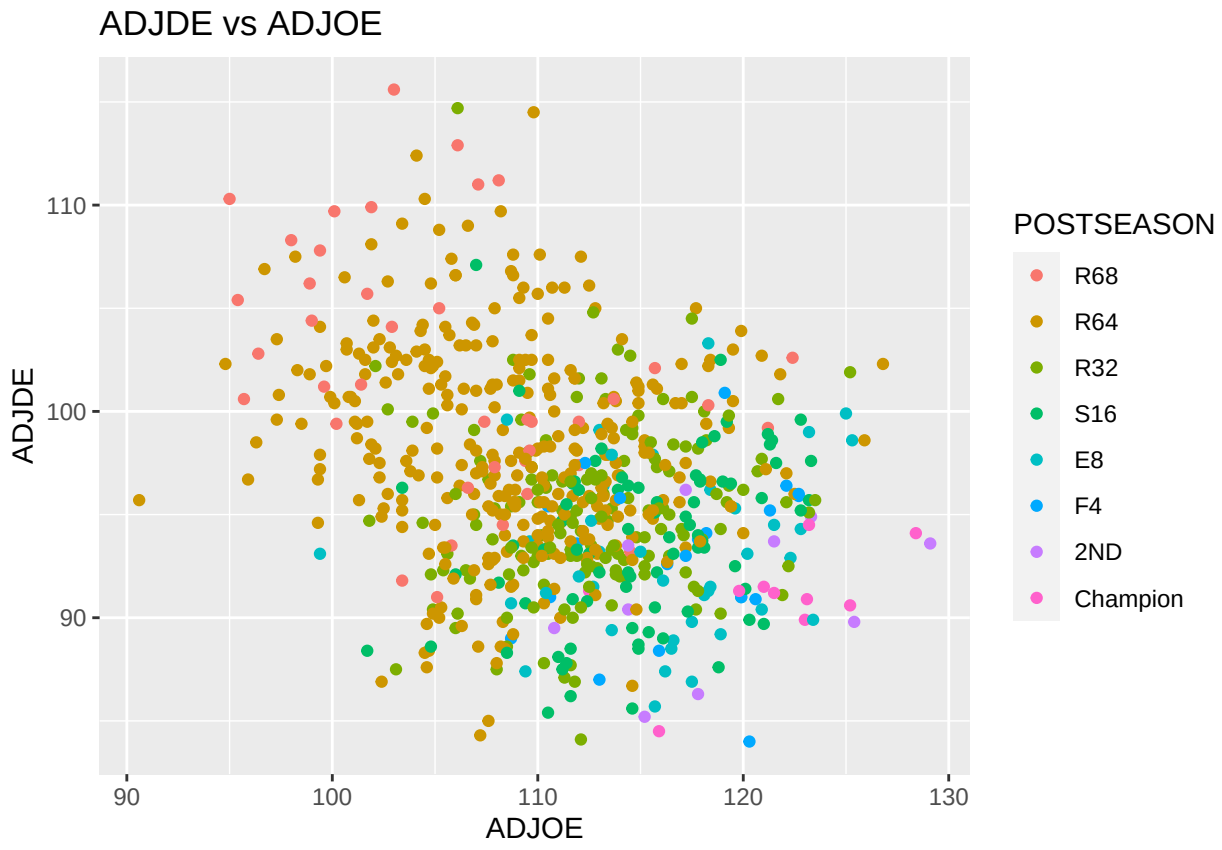
Question 4. Is there a correlation between the pace of play of a team and their final ranking? Do teams with a higher pace of play do better in the postseason?

Investigation of Follow-up Questions

- Which questions did you decide to answer? Follow up questions 1 & 3
- Show at least two tables or figures below that explore answers to those questions

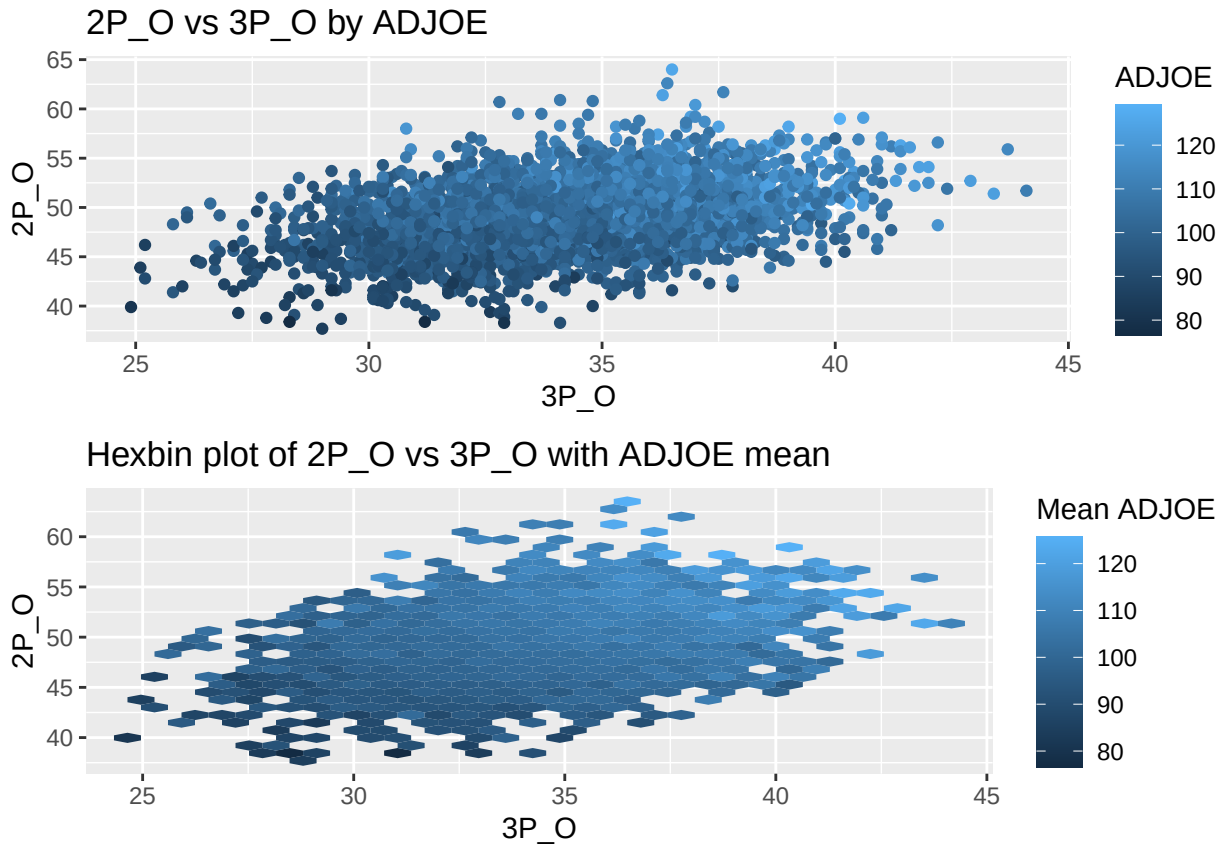
Question 1. Is it more important for a team to have a better ADJOE or ADJDE if they are trying to win a championship? If ADJOE is more important, which has more of an effect on Adjusted Offensive Efficiency, a team's 3-point percentage or their 2-point percentage?

```
#Figuring out which of the two ADJOE or ADJDE is more important to win a championship
ggplot(bbData_postseason, aes(x=ADJOE, y=ADJDE, color=POSTSEASON)) +
  geom_point() +
  labs(title="ADJDE vs ADJOE", x="ADJOE", y="ADJDE")
```




```
#If it is ADJOE which type of shooting is more important to win a championship
scatterplot_adjoe <-
  ggplot(ball_data, aes(x=`X3P_O`, y=`X2P_O`, color=ADJOE)) +
  geom_point() +
  labs(title="2P_O vs 3P_O by ADJOE", x="3P_O", y="2P_O")
hexbin_plot_adjoe <-
  ggplot(ball_data, aes(x=`X3P_O`, y=`X2P_O`, z=ADJOE)) +
  stat_summary_hex(
    fun = mean,
    bins = 30
  ) +
  labs(
    title = "Hexbin plot of 2P_O vs 3P_O with ADJOE mean",
    x = "3P_O",
    y = "2P_O",
    fill = "Mean ADJOE"
  )
```

```
graphic_layout = matrix(c(1, 2), ncol = 1)
grid.arrange(scatterplot_adjoe, hexbin_plot_adjoe, layout_matrix=graphic_layout)
```



Question 3. Considering that EFG_O and EFG_D are an indicator in win percentage; and a higher Power Rating (BARTHAG) indicates a higher likelihood of a team reaching the final four, Is there a particular range for EFG_O and EFG_D that, when combined with a certain BARTHAG score, boosts a team's chances of making it to the final four?

```
efg_o_threshold <- median(ball_data$EFG_O)
efg_d_threshold <- median(ball_data$EFG_D)
barthag_threshold <- median(ball_data$BARTHAG)

bball <- ball_data |>
  mutate(EFG_O_Group = ifelse(EFG_O >= efg_o_threshold, "High", "Low"),
         EFG_D_Group = ifelse(EFG_D <= efg_d_threshold, "Low", "High"),
         BARTHAG_Group = ifelse(BARTHAG >= barthag_threshold, "High", "Low"),
         Final_Four = POSTSEASON %in% c("F4", "2ND", "Champion"))

final_four_proportions <- bball |>
  group_by(EFG_O_Group, EFG_D_Group, BARTHAG_Group) |>
  summarise(Final_Four = sum(POSTSEASON %in% c("F4", "2ND", "Champion")),
            Total = n()) |>
  mutate(Proportion = Final_Four / Total)
```

```
## `summarise()` has grouped output by 'EFG_O_Group', 'EFG_D_Group'. You can
```

```
## override using the `.groups` argument.
```

```
final_four_proportions|>kable()
```

EFG_O_Group	EFG_D_Group	BARTHAG_Group	Final_Four	Total	Proportion
High	High	High	3	340	0.0088235
High	High	Low	0	487	0.0000000
High	Low	High	24	872	0.0275229
High	Low	Low	0	104	0.0000000
Low	High	High	0	112	0.0000000
Low	High	Low	0	805	0.0000000
Low	Low	High	3	438	0.0068493
Low	Low	Low	0	365	0.0000000

Summary

From investigating the initial questions, we gathered valuable insights into the dynamics of college basketball. We found a clear positive correlation between Effective Field Goal Percentage (EFG_O) and win percentage, highlighting the importance of offensive efficiency. Conversely, a negative correlation between Defensive Effective Field Goal Percentage (EFG_D) and win percentage underscored the significance of a solid defense. We also found that the Power Rating (BARTHAG) is a reliable indicator for a team's likelihood of reaching the final four in the postseason. Furthermore, we observed that teams from the ACC, B10, and B12 conferences consistently outperformed others in terms of Adjusted Offensive Efficiency (ADJOE) and Adjusted Defensive Efficiency (ADJDE). An interesting trend was noted in the pace of play, where teams played faster between 2015 and 2017, but the pace has slightly declined since. The relationship between 2-point and 3-point shooting percentages revealed a weak correlation, indicating that teams do not necessarily compensate for lower 2-point percentages with higher 3-point percentages. Adjusted Defensive Efficiency (ADJDE) proved to be a reliable predictor for postseason success, with teams having lower ADJDE more likely to make it to the final four. Additionally, we saw that higher ADJOE and lower ADJDE contribute to a higher BARTHAG score, and that teams making it to the final four typically have higher ADJOE and lower ADJDE. These insights from the initial questions were crucial in understanding the key factors that contribute to a team's success in college basketball.

Having explored the importance of Offensive Efficiency (ADJOE) and Defensive Efficiency (ADJDE), we now delve into the nuanced aspects of these metrics. The insights from the violin plot of adjusted offense (ADJOE) and adjusted defense (ADJDE) across postseason success were fascinating. There is a very sharp gap between ADJOE and ADJDE as teams went deeper into the playoffs. Teams that made it past the final four round had a clear positive difference between ADJOE and ADJDE and this was a very pronounced trend in the data. From this insight, we were interested in learning more about these adjusted offense and adjusted defense metrics. While both seem to play a big part in a team's postseason success, we were interested in trying to assess which one had a bigger impact as well as to understand the drivers of these metrics themselves. We were initially quite surprised that there was limited association between two-point and three-point shooting so we were interested in exploring how those metrics affected the ADJOE. This led us to our first follow up question.

The other offense and defense metrics also piqued our interest from the initial exploration. We saw strong associations between offensive and defensive effective field goal percentage (EFG_O and EFG_D) and win percentages for teams. These combined with the team's power rating seemed to have a high impact on a team's postseason success. We wanted to try to quantify this impact and find an effective range of values for which teams performed better. This led us to our second follow up question.

After answering our initial questions, we formulated four essential follow-up questions to gain a deeper understanding of our basketball dataset and lay the groundwork for informed conclusions on how this data could predict march madness outcomes.

For instance, we explored the importance of Offensive Efficiency (ADJOE) and Defensive Efficiency (ADJDE) in the quest for a championship. We discovered that the significance of these metrics depends on several factors, and when it comes to ADJOE, a team's 3-point percentage has a more substantial influence on Adjusted Offensive Efficiency than their 2-point percentage. This nuanced understanding of offensive success is compelling as it provides actionable information for teams aspiring to reach the championship level.

In addition, our investigation shed light on the relative importance of turnover and rebound metrics in comparison to traditional offensive and defensive statistics when predicting post-season success. These findings challenge conventional wisdom and highlight the complexity of the game. Furthermore, our research identified specific ranges for Effective Field Goal Percentage (EFG_O and EFG_D) and their combination with Power Rating (BARTHAG) scores that can significantly boost a team's chances of making it to the final four, offering valuable strategic insights. Lastly, we explored the relationship between a team's pace of play and their final ranking, revealing a complex correlation that challenges preconceived notions about how playing style impacts postseason success. Our findings collectively provide a comprehensive view of the multifaceted nature of basketball performance, offering valuable insights for teams and enthusiasts alike. We identified two of our follow-up questions as pivotal in enhancing our comprehension of the dataset's predictive capabilities. Consequently, we made the decision to craft visual representations to aid in their interpretation.

Out of the four follow-up questions we found questions one and three to be the most interesting and decided to investigate these questions further. For question one we chose to use multiple figures. To start we needed to figure out which of the two, ADJOE or ADJDE, is more important to win a championship. The plot we made to investigate this was a scatter plot of all the team's ADJOE on the x-axis and their ADJDE on the y-axis. We also colored the points in the plot based on where each team made it in the postseason. From this plot we were able to learn that the ADJOE seems to be more important to postseason success, as all of the more successful teams had a high ADJOE, but not necessarily a high ADJDE. From there we went to investigate which was more important to the ADJOE, a team's 3-point percentage or 2-point percentage. We figured this out using two different plots, a scatterplot and a hexbin plot, both showing the relationship between the teams' 3-point percentages and 2-point percentages, with 3P_O on the x-axis and 2P_O on the y-axis. The difference between the two plots was that the scatterplot's color was based on just the ADJOE and the hexbin plot's color was based on the mean ADJOE. Looking at these plots, the data leans towards the higher 3-point percentages having a higher ADJOE rather than the 2-point percentage. This leads us to conclude that a team's 3-point percentage is more important to a team's ADJOE, thus for a team to generally have postseason success it must have a high ADJOE and a high 3-point percentage.

For our investigation on question three we decided to go a different route by using a table instead of a plot to find our answer. We chose this because we thought it would represent the data we wanted to see in a clear way to get our desired answer. To make our table we first created median thresholds for EFG_O, EFG_D, and BARTHAG to be able to determine whether or not a team has a high or low value for each of these metrics. Our table then tells us different statistics for teams with each combination of high and low values for each metric. By defining postseason success as making the Final Four, we add three columns that show the number of teams in each combination that make the Final Four, how many teams are in each combination, and the proportion of teams in each combination that make the Final Four (Final Four teams / Total teams). Looking at the described table, it's clear that having a good balance of offense and defense, along with a strong power rating, really makes a difference when it comes to a team's chances of making it to the final four. Teams that score high in offensive efficiency (EFG_O), keep their defensive efficiency low (EFG_D), and have a high power rating (BARTHAG) are the ones that are most likely to succeed in the postseason, with about 2.75% of them making it to the final four. On the other hand, teams that don't do well in these areas are really at a disadvantage, and the data shows they don't really stand a chance of making it to the final four. So, it's safe to say that these metrics are super important for a team's success in the postseason. Overall, our investigation of college basketball has revealed essential insights into the variables influencing team success. Key metrics, such as Effective Field Goal Percentage, Power Rating, and Offensive Efficiency, have emerged as pivotal factors in predicting postseason outcomes. These findings paint a vivid picture of the intricate web of basketball performance, providing practical guidance for teams aspiring to thrive in the college basketball arena.