

Unit 1 homework instructions

DKU Stats 101 Fall 2025 Session 2

2025-10-24

Table of contents

Scoring guide	2
Content	2
Technical	3
 Questions	 3
Question 1: Displaying and describing the data (25 points)	3
1a. Investigating offensive shots	4
1b. Investigating goals	5
1d. Thinking about your results	7
Question 2: Comparing groups (25 points)	8
2a. Compare <code>home_foul_advantage</code> by the variable <code>covid_match</code>	8
2b. Compare <code>home_yellow_advantage</code> by the variable <code>covid_match</code>	10
2c. Thinking about your results	11
Question 3: Considering deviations (25 points)	11
3a. Selecting your data	11
3b. Finding the average	12
3c. Normalizing the data	12
3d. Thinking about your results	13
Question 4: Your own investigation (25 points)	13
4a. Selecting your own question	13
4b. In summary	16

```
#| include: false  
#| echo: false  
  
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr    1.3.1
v purrr     1.0.4

-- Conflicts -----
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

# Make sure that your data file is in the same location as where you have saved this document
epl <- read.csv("E:/Github/dkustats101fall2025s2/homeworks/Unit 1 homework/epl.stats.cleaned")

# Put any other setup code here
```

Scoring guide

Content

- Getting the right answer is only a small part of the grade
- Good quality interpretation of your results is the name of the game
- If you see something that looks unusual in your data (outlier, some unusual distribution type) - investigate it!
- When explaining your results, say something interesting about them. Did it match your expectations? Why or why not?
- Brief explanations that simply repeat what I can visually see myself will not receive a good score
- On the other hand, filling the homework with pages of not very interesting description is not valuable either. The goal isn't to write the most words, but find the most interesting things in the data.
- You do not need to be an expert in football to get a good grade on this assignment, but I will expect you to look up basic information, such as “what is a shot in football” and “what kind of games were Covid games”? and so on to help you understand and set expectations your data.
- The information requested in the question prompts are only a starting point, if you find other interesting information along the way, please report that. You don't need to look at the data forever but if there is obviously something else interesting in the data you should report it.
- You must have up to Question 1 completed for the homework check on October 30th

Technical

- Make sure your graphs are produced using `ggplot()`, are well labeled, and are easy to read.
- Make sure your tables are produced with the `kable()` function from the `knitr` package, are well labeled, and are easy to read. You can make your tables prettier with the `kableExtra` package.
- Make sure you do not have anything rendered in your PDF file besides your results and, when asked for by a question, your code. That means no warnings, messages, or other output should appear in your final rendered PDF file.
- Make sure to accurately mark each page a question answer appears on when submitting on GradeScope.



Questions

Question 1: Displaying and describing the data (25 points)

For this investigation, we are going to examine the distribution of goals.

1a. Investigating offensive shots

Using the Think-Show-Tell framework from the textbook, investigate the distribution of `home_shots` and `away_shots`.

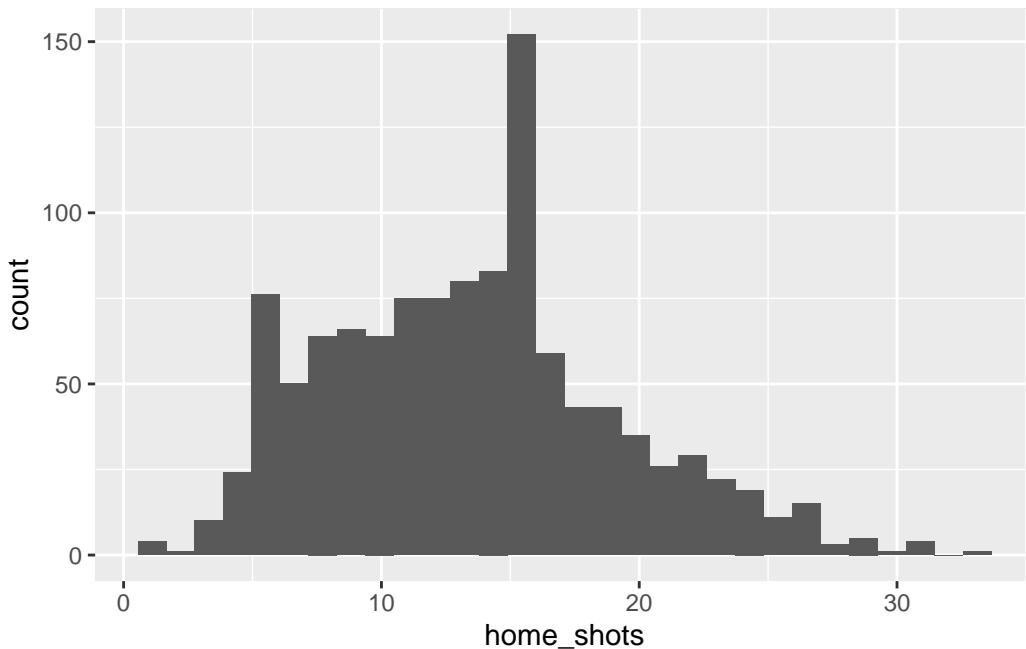
Note: I recommend you use the internet to look up how a shot is defined in football

Think

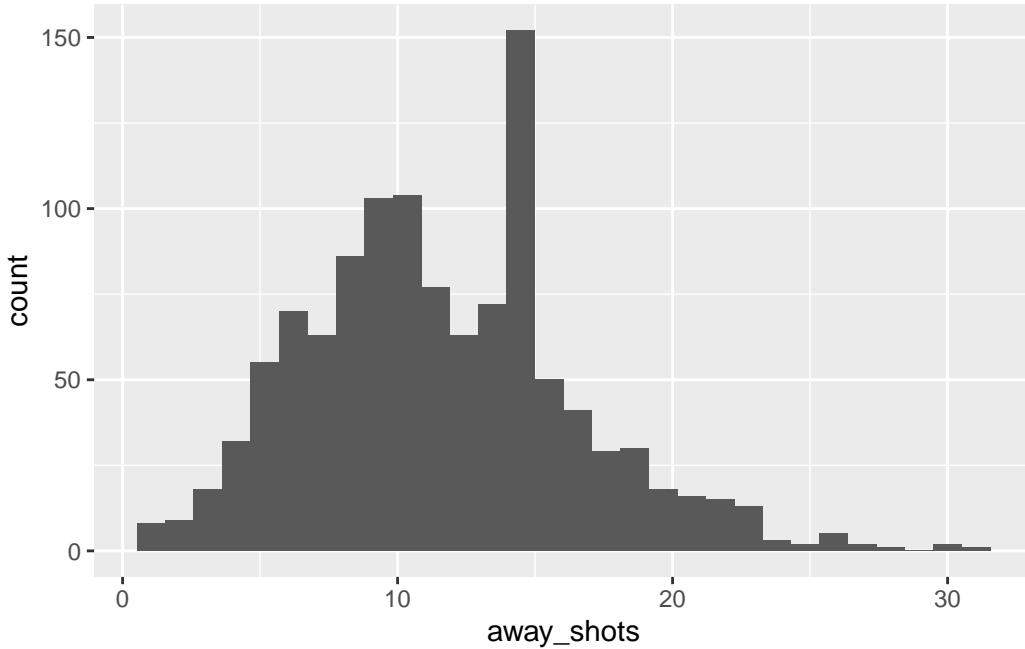
Expectation: home shots will be greater than away shots, even slightly. This is due to an advantage, as they are familiar with their field. This would mean a right-skewed distribution given the likelihood of high shot counts against weaker players. Variable meaning: `home_shots` is the variable for the number of shots made by the home team in a match. `away_shots` is the variable for the number of shots made by the away team in a match. Display: Due to the variable type, the best way to display this would be with a boxplot or histogram as they would be most helpful for comparing home v.s. away shots.

Show

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The tail extends to the right in `away_shots`, this is a right skewed distribution. The peak is at around 10-12 shots. In `home_shots`, the distribution is even further right skewed, with a peak around 12-15 shots.

Tell

The data shows that home teams generally take more shots than away teams, supporting the idea of a home-field advantage. The difference was smaller than expected, but still clear, as both are rightly skewed, but the `home_shots` distribution is even more skewed by 2-3 shots. Overall, playing at home appears to give teams a slight advantage.

1b. Investigating goals

Using the Think-Show-Tell framework from the textbook (example on page 71), investigate the distribution of `home_goals` and `away_goals`.

Think

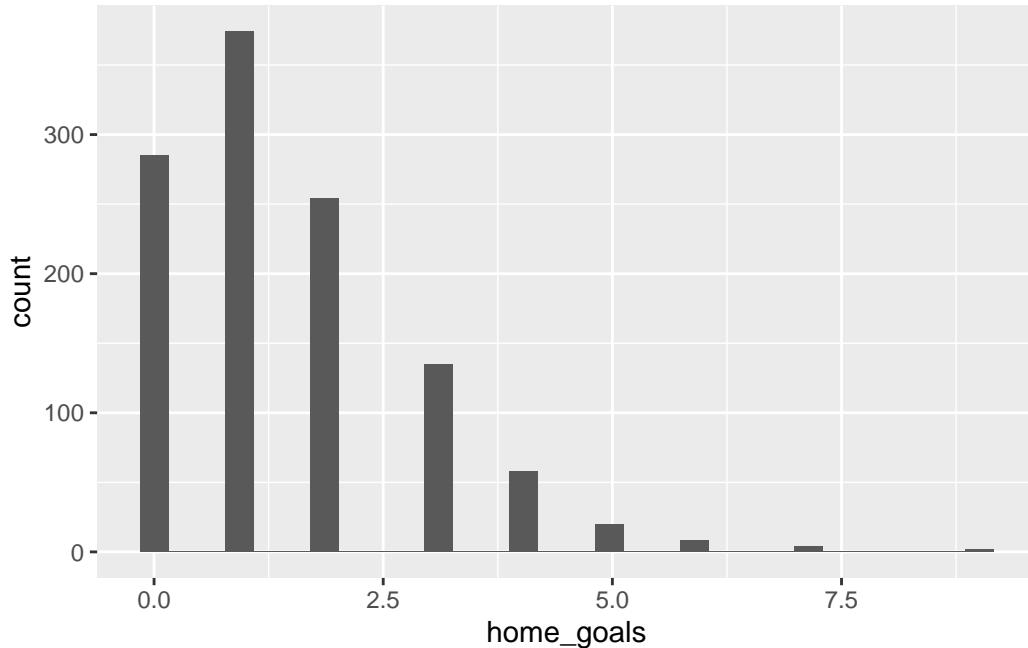
Expectation: The home team is generally expected to score more goals than the away team, although the difference might be small, as evidenced with the scores. Familiarity with the home field and crowd support give a slight edge, along with confidence, time zone, and other factors that can lead to higher home goal counts. This would result in a right-skewed distribution, since most games will have a modest number of goals, but occasional high-scoring matches can occur. Variable meaning: `home_goals`: Number of goals scored by the home team in a match.

away_goals: Number of goals scored by the away team in a match. Display: Since these are numeric count variables, the data is also best visualized with a boxplot or histogram, which shows the differences and spread between home and away goals.

Show

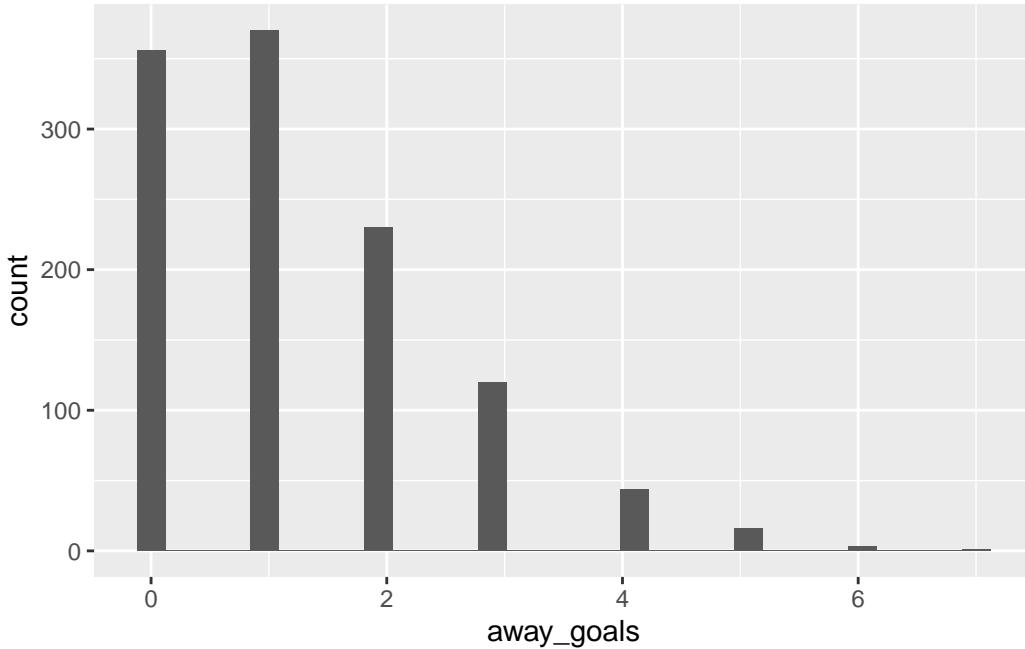
```
ggplot(ep1, aes(x=home_goals)) +  
  geom_histogram()
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



```
ggplot(ep1, aes(x=away_goals)) +  
  geom_histogram()
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



Tell

The data shows that home teams generally score more goals than away teams, supporting the idea of a home-field advantage. The difference was also smaller than expected, notably with how small the bars are. We can also note this difference in the way the data is graphed, as home goals extend passed 7.5, while the furthest point in away_goals is at 7 and very slight. Both are also rightly skewed.

1d. Thinking about your results

Consider the results of 1b. and 1c. together. What can we understand about these offensive statistics?

I don't see a "1c" question, however with both results we can see a correlation between the home_shots and home_goals in success, as well as with away_shots and home_goals. It is surprising that the difference was so slight, but this is understandable given the volume of factors that go into this process, as outlined in the "show" section. This can be the absence of a key player, the level of competition, the diversity in strength within a team's offense or defense, among other factors. Overall, there is always an explanation for my statistical data appears a certain way, and in this case we happened to have a correlation.

Question 2: Comparing groups (25 points)

One popular theory about why teams have a home field advantage is that the intensity of the crowd influences the officials to be more favorable to the home team (see [this article](#)). Let's see if our dataset supports this hypothesis.

First, create a variable called `covid_match` where the variable is TRUE if the match was conducted in an empty stadium during Covid-19 and FALSE if it was not. Next, create a variable called `home_foul_advantage` that subtracts `home_fouls` from `away_fouls` and another variable called `home_yellow_advantage` that subtracts `home_yellow` from `away_yellow`.

2a. Compare `home_foul_advantage` by the variable `covid_match`

Think

For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

```
names(ep1)
```

```
[1] "X.1"                      "X"  
[3] "date"                     "date_cleaned"  
[5] "year"                     "month"  
[7] "day"                      "day_of_week"  
[9] "match_start_time"         "match_start_time_cleaned"  
[11] "match_start_hour"        "match_start_minute"  
[13] "home_team"                "away_team"  
[15] "stadium"                  "attendance"  
[17] "home_team_league_rank"   "home_goals"  
[19] "away_team_league_rank"   "away_goals"  
[21] "home_possessions"        "away_possessions"  
[23] "home_shots"               "away_shots"  
[25] "home_on"                  "away_on"  
[27] "home_off"                 "away_off"  
[29] "home_blocked"             "away_blocked"  
[31] "home_pass"                 "away_pass"  
[33] "home_chances"              "away_chances"  
[35] "home_corners"              "away_corners"  
[37] "home_offside"              "away_offside"  
[39] "home_tackles"              "away_tackles"
```

```

[41] "home_duels"           "away_duels"
[43] "home_saves"           "away_saves"
[45] "home_fouls"           "away_fouls"
[47] "home_yellow"          "away_yellow"
[49] "home_red"              "away_red"
[51] "links"

epl$covid_match <- epl$attendance == 0
epl$home_foul_advantage <- epl$away_fouls - epl$home_fouls
epl$home_yellow_advantage <- epl$away_yellow - epl$home_yellow

tapply(epl$home_foul_advantage, epl$covid_match, mean, na.rm = TRUE)

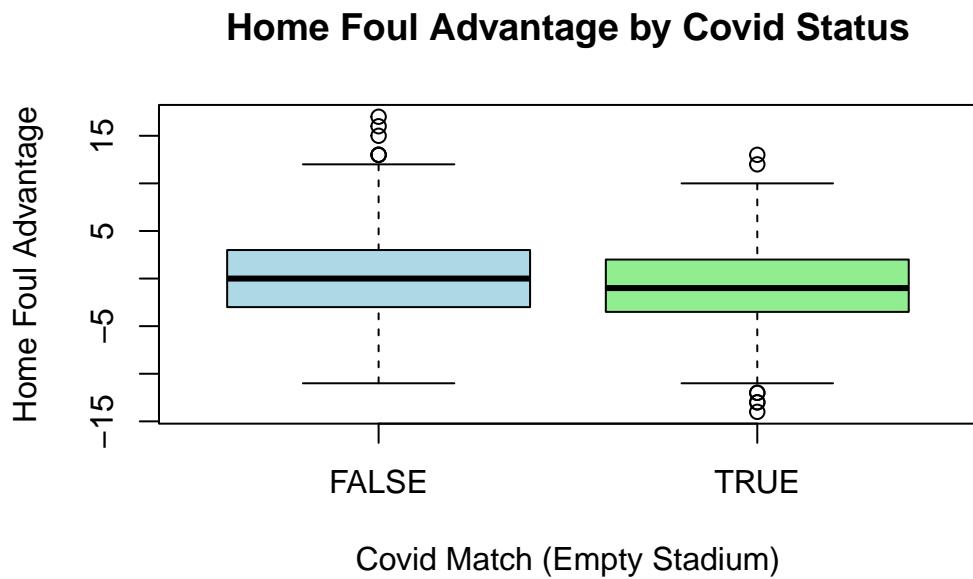
```

	FALSE	TRUE
mean	0.2289474	-0.6157895

```

boxplot(home_foul_advantage ~ covid_match, data = epl,
        main = "Home Foul Advantage by Covid Status",
        xlab = "Covid Match (Empty Stadium)",
        ylab = "Home Foul Advantage",
        col = c("lightblue", "lightgreen"))

```



```
t.test(home_foul_advantage ~ covid_match, data = epl)
```

Welch Two Sample t-test

```
data: home_foul_advantage by covid_match
t = 2.9349, df = 740.42, p-value = 0.00344
alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to zero
95 percent confidence interval:
 0.2796912 1.4097824
sample estimates:
mean in group FALSE mean in group TRUE
 0.2289474          -0.6157895
```

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2b. Compare home_yellow_advantage by the variable covid_match

Think

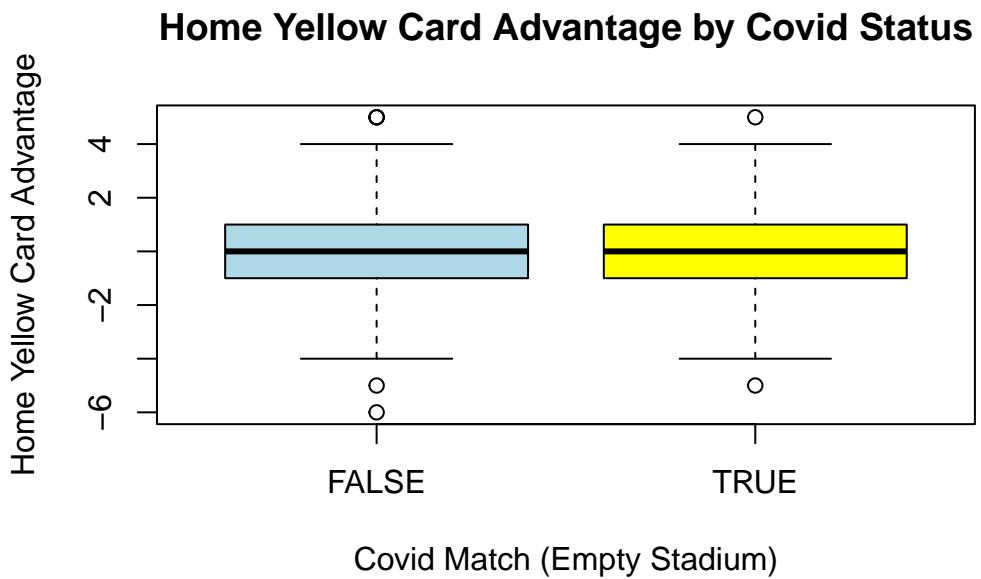
For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data

Show

```
tapply(epl$home_yellow_advantage, epl$covid_match, mean, na.rm = TRUE)
```

```
FALSE      TRUE
0.18552632 0.03421053
```

```
boxplot(home_yellow_advantage ~ covid_match, data = epl,
        main = "Home Yellow Card Advantage by Covid Status",
        xlab = "Covid Match (Empty Stadium)",
        ylab = "Home Yellow Card Advantage",
        col = c("lightblue", "yellow"))
```



For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

2c. Thinking about your results

Consider the results of 2b. and 2c. together. What can we learn about hypothesis for this question? What conclusion would you draw? And what other information do you think you would need to be more confident in your conclusion?

Question 3: Considering deviations (25 points)

3a. Selecting your data

Pick any team and create two subsets using the `filter` verb of just matches featuring that team, one subset for home, the other for away.

```

chelsea_home <- epl %>%
  filter(home_team == "Chelsea")

# Subset of Arsenal away matches
chelsea_away <- epl %>%
  filter(away_team == "Chelsea")

```

3b. Finding the average

Make a table of the averages of goals (home and away, separate columns), shots (home and away, separate columns), possession % (home and away, separate columns), and pass % (home and away, separate columns). Show your code using the `#| echo: true` code block option.

```

library(dplyr)

# summary table of averages
summary_table <- epl %>%
  summarise(
    avg_home_goals = mean(home_goals, na.rm = TRUE),
    avg_away_goals = mean(away_goals, na.rm = TRUE),
    avg_home_shots = mean(home_shots, na.rm = TRUE),
    avg_away_shots = mean(away_shots, na.rm = TRUE),
    avg_home_possessions = mean(home_possessions, na.rm = TRUE),
    avg_away_possessions = mean(away_possessions, na.rm = TRUE),
    avg_home_pass = mean(home_pass, na.rm = TRUE),
    avg_away_pass = mean(away_pass, na.rm = TRUE)
  )

#full table
summary_table

```

	avg_home_goals	avg_away_goals	avg_home_shots	avg_away_shots
1	1.502632	1.290351	13.55877	11.47456
	avg_home_possessions	avg_away_possessions	avg_home_pass	avg_away_pass
1	50.81675	49.20596	79.70737	78.9743

3c. Normalizing the data

Add a row to your table; find how many z units each of the averages are away from the overall averages in the dataset. Show your code using the `#| echo: true` code block option.

```

library(dplyr)

# summary table of averages
z_scores <- epl %>%
  summarise(
    avg_home_goals = mean(home_goals, na.rm = TRUE),
    avg_away_goals = mean(away_goals, na.rm = TRUE),
    avg_home_shots = mean(home_shots, na.rm = TRUE),
    avg_away_shots = mean(away_shots, na.rm = TRUE),
    avg_home_possessions = mean(home_possessions, na.rm = TRUE),
    avg_away_possessions = mean(away_possessions, na.rm = TRUE),
    avg_home_pass = mean(home_pass, na.rm = TRUE),
    avg_away_pass = mean(away_pass, na.rm = TRUE)
  )
#combines averages and z-scores
summary_table_full <- bind_rows(summary_table, z_scores)

#full table
summary_table_full

```

	avg_home_goals	avg_away_goals	avg_home_shots	avg_away_shots
1	1.502632	1.290351	13.55877	11.47456
2	1.502632	1.290351	13.55877	11.47456
	avg_home_possessions	avg_away_possessions	avg_home_pass	avg_away_pass
1	50.81675	49.20596	79.70737	78.9743
2	50.81675	49.20596	79.70737	78.9743

3d. Thinking about your results

Interpret your results - what do the z scores indicate about the offensive capabilities of the team? How does it vary by home vs. away? What other kind of data would you like to have to answer this question?

Question 4: Your own investigation (25 points)

4a. Selecting your own question

Similar to the previous questions, think of your own question that you would like to ask of the data, ideally one that goes deeper into one of the questions considered above. Use the Think-Show-Tell procedure to conduct your investigation. Think deeply about what your result means.

Think

Did Covid affect total goals per game? *For this section, please write down your expectations, why you expect it, the variable meaning, and, given the variable type, the best way to display the data color*

Show

```
# name assignment  
names(ep1)
```

```
[1] "X.1"                      "X"  
[3] "date"                     "date_cleaned"  
[5] "year"                     "month"  
[7] "day"                      "day_of_week"  
[9] "match_start_time"         "match_start_time_cleaned"  
[11] "match_start_hour"        "match_start_minute"  
[13] "home_team"                "away_team"  
[15] "stadium"                  "attendance"  
[17] "home_team_league_rank"   "home_goals"  
[19] "away_team_league_rank"   "away_goals"  
[21] "home_possessions"        "away_possessions"  
[23] "home_shots"               "away_shots"  
[25] "home_on"                  "away_on"  
[27] "home_off"                 "away_off"  
[29] "home_blocked"             "away_blocked"  
[31] "home_pass"                "away_pass"  
[33] "home_chances"              "away_chances"  
[35] "home_corners"              "away_corners"  
[37] "home_offside"              "away_offside"  
[39] "home_tackles"              "away_tackles"  
[41] "home_duels"                "away_duels"  
[43] "home_saves"                "away_saves"  
[45] "home_fouls"                 "away_fouls"  
[47] "home_yellow"                "away_yellow"  
[49] "home_red"                  "away_red"  
[51] "links"                     "covid_match"  
[53] "home_foul_advantage"      "home_yellow_advantage"
```

```
any(names(ep1) == "home_goals")
```

```
[1] TRUE
```

```
any(names(ep1) == "away_goals")
```

```
[1] TRUE
```

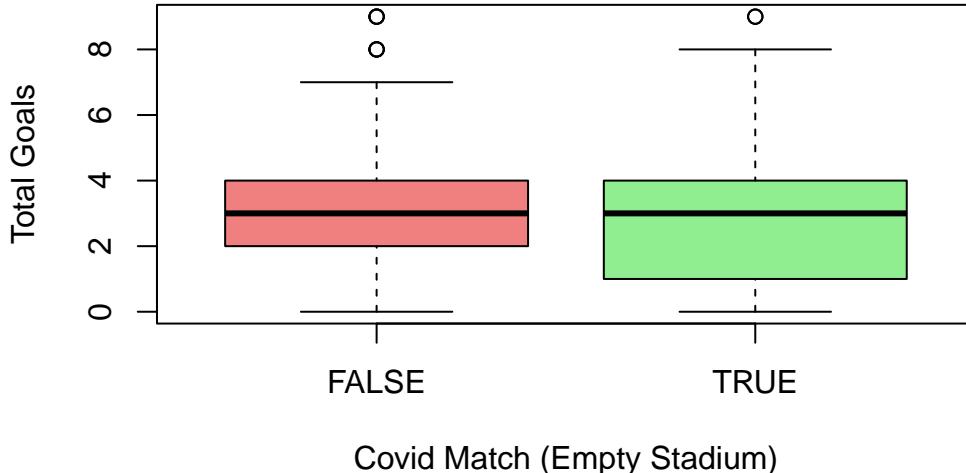
```
any(names(ep1) == "total_goals")
```

```
[1] FALSE
```

```
# total goals as variable for total goals per match  
ep1$total_goals <- ep1$home_goals + ep1$away_goals
```

```
# Graph as boxplot comparing total goals during Covid vs not  
boxplot(total_goals ~ covid_match, data = ep1,  
        main = "Total Goals per Match: Covid vs Non-Covid",  
        xlab = "Covid Match (Empty Stadium)",  
        ylab = "Total Goals",  
        col = c("lightcoral", "lightgreen"))
```

Total Goals per Match: Covid vs Non-Covid



```
head(ep1$total_goals)
```

[1] 5 3 1 2 2 1

For this section, please make an appropriate graph or table and briefly describe what you observe

Tell

Please interpret the meaning of your finding here, especially with respect to your expectation

4b. In summary

Sum up everything that you have learned from questions 1-4. Do not simply repeat/rephrase your previous results but try to say something larger that synthesizes the results together to draw a more meaningful general conclusion.