Unit 1 homework sample solution

DKU Stats 101 Fall 2025 Session 2

Anonymous

2025-11-02

Questions

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

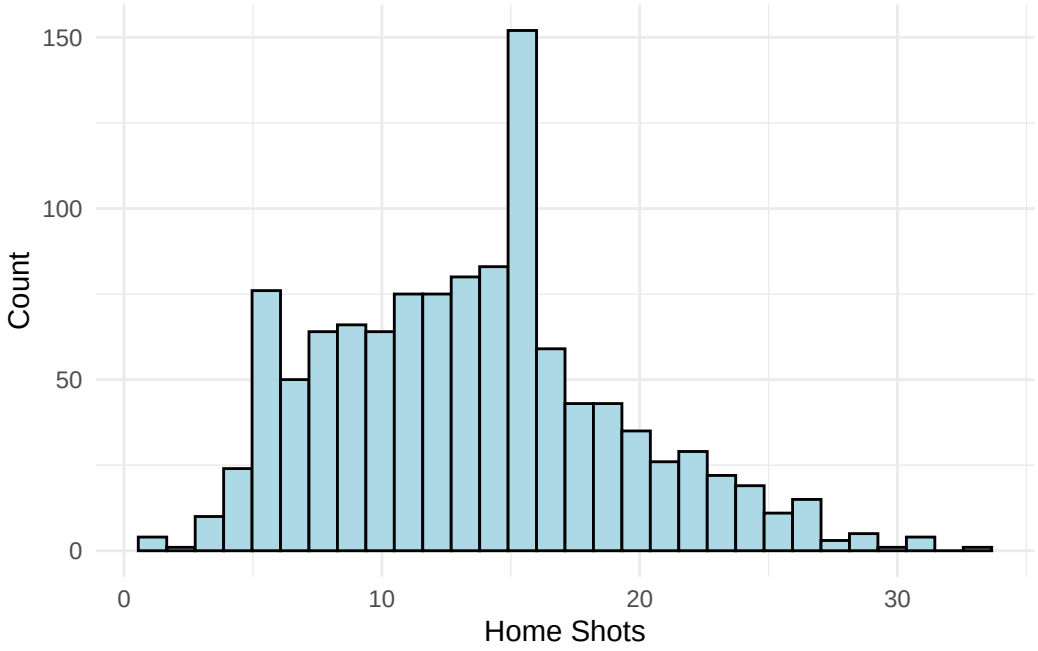
For this investigation, we are going to examine the distribution of goals.

1a. Investigating offensive shots

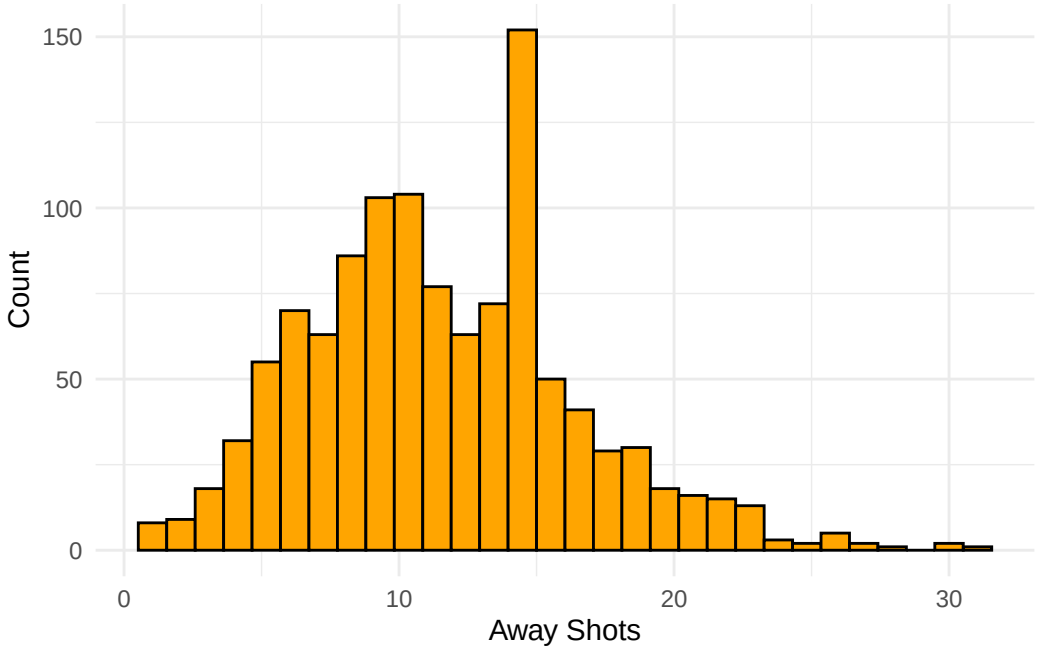
Using the Think-Show-Tell framework from the textbook, investigate the distribution of `home_shots` and `away_shots`.

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

> *Think*
>
> The variables `home_shots` and `away_shots` record the total number of attempts made by the home and away teams to score in each football match. A shot refers to any intentional attempt to score, regardless of whether it results in a goal. Both variables are quantitative counts.
>
> I expect home teams to take more shots on average than away teams because of the home advantage, which often comes from factors such as crowd support, familiarity with the field, and reduced travel fatigue.
>
> To investigate the distribution of these variables, histograms can show the overall distribution and possible skewness, while boxplots allow for a clear comparison of the central tendency and spread between home and away teams. A summary statistics table can also help highlight differences in their averages, variability, and range.
>
> *Show*



(a) Home Shots



(b) Away Shots

Figure 1: Histogram of shots taken by home and away teams

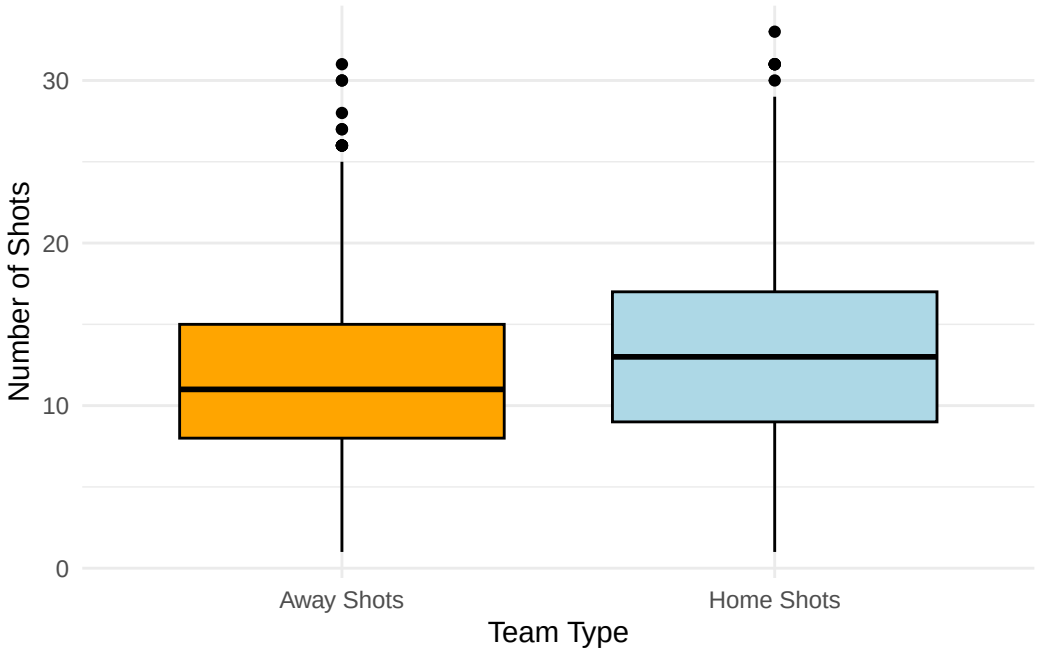


Figure 2: Boxplot of shots taken by home and away teams

Table 1: Statistics of variables: home_shots and away_shots

Statistic	Home_Shots	Away_Shots
Count	1140.00	1140.00
Std Dev	5.62	5.05
Mean	13.56	11.47
Min	1.00	1.00
25%	9.00	8.00
Median	13.00	11.00
75%	17.00	15.00
Max	33.00	31.00

Table 2: Outliers of the variables: home_shots and away_shots

Home Team	Away Team	Home Shots	Away Shots
brighton-and-hove-albion	brentford	33	7
crystal-palace	leicester-city	31	3
arsenal	bournemouth	31	4
brighton-and-hove-albion	norwich-city	31	6
manchester-city	leeds-united	31	6
arsenal	norwich-city	30	10
southampton	liverpool	15	30
leicester-city	tottenham-hotspur	14	27
aston-villa	leeds-united	12	27
watford	manchester-city	11	26
leeds-united	manchester-city	9	26
leeds-united	liverpool	9	30
west-bromwich-albion	liverpool	9	26
chelsea	brighton-and-hove-albion	8	26
newcastle-united	manchester-united	7	28
west-ham-united	manchester-city	6	31
brentford	newcastle-united	6	26

Tell

From Figure 1, we can observe that the distributions of `home_shots` and `away_shots` are unimodal and slightly right-skewed. Most matches fall between 5 and 20 shots, with home teams extending further to the right. On average, home teams take more shots (mean = 13.56, median = 13) than away teams (mean = 11.47, median = 11), which aligns with the expectation of a modest home advantage.

As shown in Figure 2, home teams display a slightly wider spread and several upper outliers, reflecting a few high-shot matches. As shown in Table 1, the standard deviation is about 5.6 for home teams and 5.0 for away teams, indicating moderate variability. The interquartile ranges suggest that most matches fall within a consistent range of shots for both sides. Table 2 also shows that a small number of matches exceed 29 home or 25 away shots, representing genuine high-offense performances rather than data errors.

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`.

Table 3: Statistics of variables: home_goals and away_goals

Statistic	Home_Goals	Away_Goals
Count	1140.00	1140.00
Std Dev	1.36	1.23
Mean	1.50	1.29
Min	0.00	0.00
25%	0.50	0.00
Median	1.00	1.00
75%	2.00	2.00
Max	9.00	7.00

Think

The variables `home_goals` and `away_goals` represent the total number of goals scored by the home and away teams in each match. A goal is recorded when the ball completely crosses the goal line within the frame, following the official rules of football. Both variables are quantitative count variables.

I expect the distributions of both variables to be right-skewed, since most matches have only a few goals while high-scoring games are relatively rare. On average, home teams are expected to score slightly more goals than away teams, which would be consistent with the modest home advantage observed in the previous question.

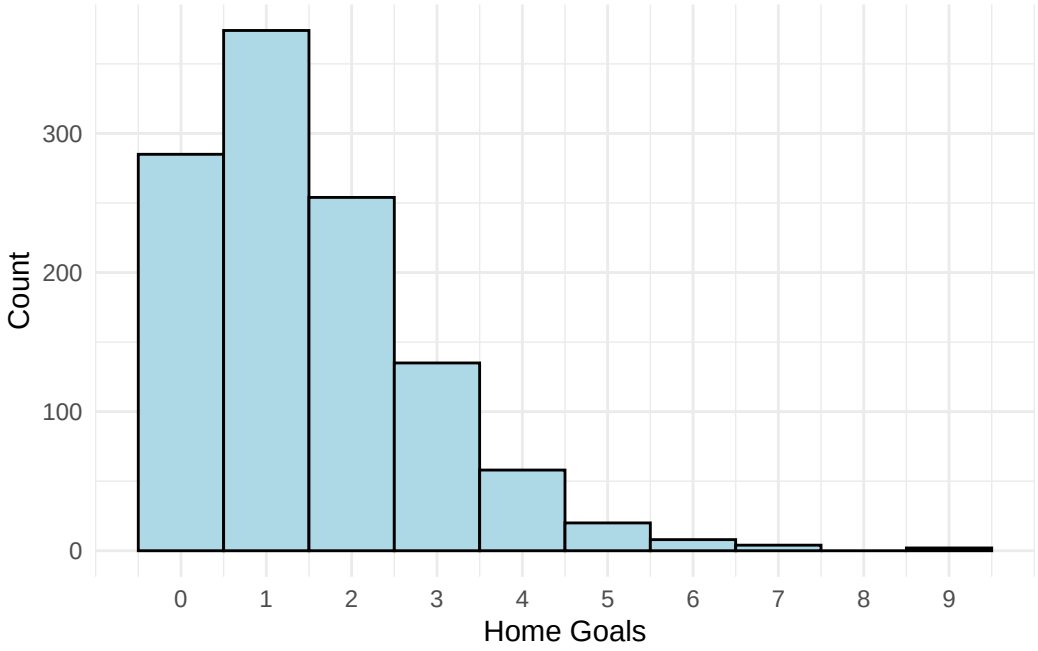
To investigate these variables, histograms can show the overall goal distribution, while boxplots can compare the central tendency and spread between home and away teams. A summary statistics table can also provide a clear numerical comparison of averages, variation, and range.

Show

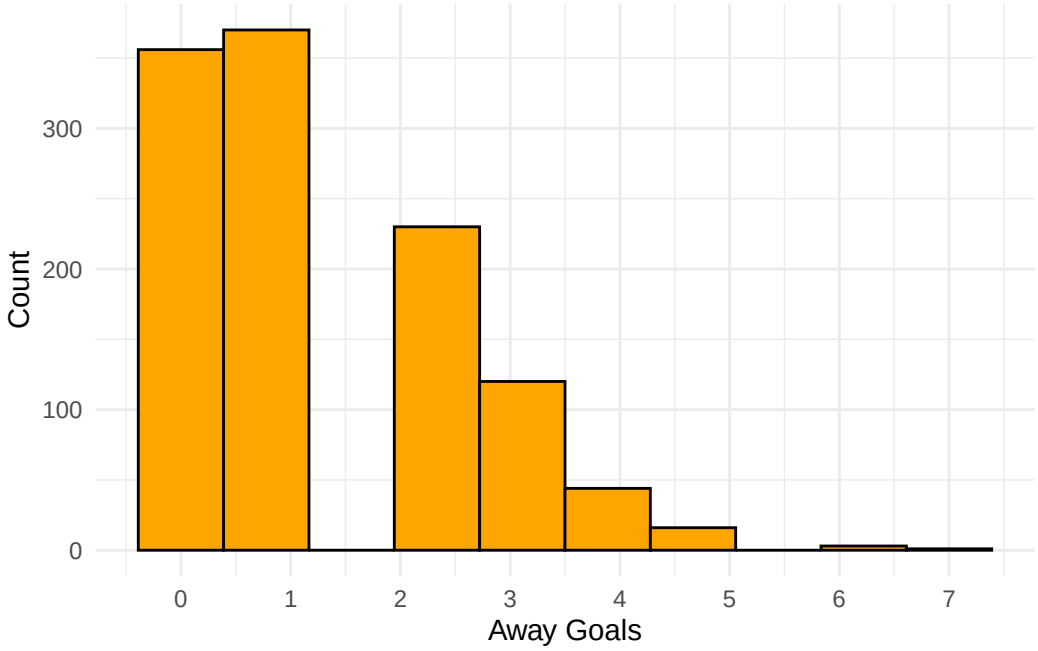
Tell

As shown in Figure 3, the distributions of `home_goals` and `away_goals` are unimodal and strongly right-skewed, matching the expectation that most matches have only a few goals while high-scoring games are rare. Home teams score slightly more on average (mean = 1.50, median = 1) than away teams (mean = 1.29, median = 1), reflecting the modest home advantage as I expected.

Figure 4 shows similar medians for both groups, but away goals have a slightly wider spread, while home goals include more extreme outliers. As shown in Table 3, standard deviations around 1.3 indicate limited variability, with most matches falling between 0 and 2 goals. Table 4 shows that games exceeding five goals are rare and mostly occur at home, suggesting that high-scoring matches are exceptional



(a) Home Goals



(b) Away Goals

Figure 3: Histogram of goals scored by home and away teams

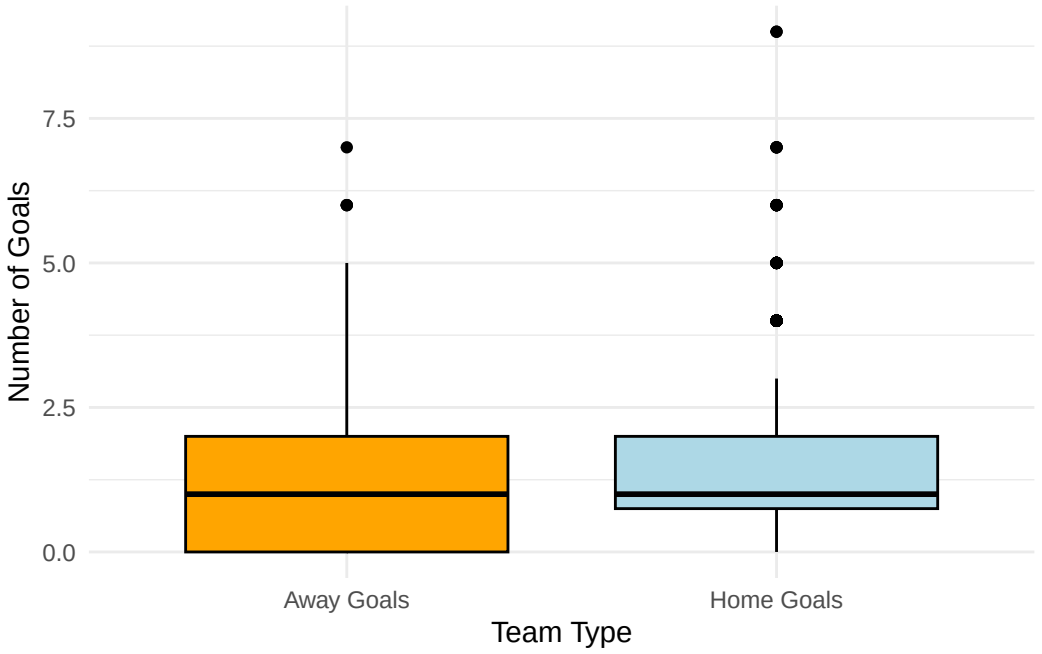


Figure 4: Boxplot comparing the distributions of home_goals and away_goals

Table 4: Possible outliers of the variables: home_goals (> 5) or away_goals (> 5)

Home Team	Away Team	Home Goals	Away Goals
liverpool	bournemouth	9	0
manchester-united	southampton	9	0
liverpool	manchester-united	7	0
manchester-city	leeds-united	7	0
chelsea	norwich-city	7	0
aston-villa	liverpool	7	2
brighton-and-hove-albion	wolverhampton-wanderers	6	0
newcastle-united	tottenham-hotspur	6	1
manchester-city	manchester-united	6	3
tottenham-hotspur	leicester-city	6	2
manchester-city	nottingham-forest	6	0
liverpool	leeds-united	6	0
manchester-city	leicester-city	6	3
manchester-united	leeds-united	6	2
leeds-united	liverpool	1	6
manchester-united	tottenham-hotspur	1	6
southampton	chelsea	0	6
crystal-palace	liverpool	0	7

but genuine. Overall, the results align with expectations: football matches are generally low-scoring, and home teams tend to achieve slightly higher goal totals.

1d. Thinking about your results

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

Answers will vary here, good quality effort to interpret investigation of this question is required.

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

The variable `home_foul_advantage` measures the difference in fouls between the away and home teams (`away_fouls - home_fouls`). Positive values indicate that away teams committed more fouls than home teams, which could suggest that referees were more lenient toward home teams.

During the Covid-19 period, all matches were played in empty stadiums without an audience. Since there was no crowd pressure, I expect the home foul advantage to be smaller on average in Covid matches than in normal matches. In other words, referees may have called fouls more evenly when there were no audiences in the stands.

Because `home_foul_advantage` is a quantitative variable and `covid_match` is a categorical variable with two groups (TRUE/FALSE), side-by-side boxplots or a summary statistics table comparing group means and spreads would be the best way to display and compare the data.

Show

Table 5

```r
epl_2 <- epl %>%
  mutate(
    covid_match = ifelse(attendance == 0 | is.na(attendance), TRUE, FALSE),
    home_foul_advantage = away_fouls - home_fouls,
    home_yellow_advantage = away_yellow - home_yellow
  )
```

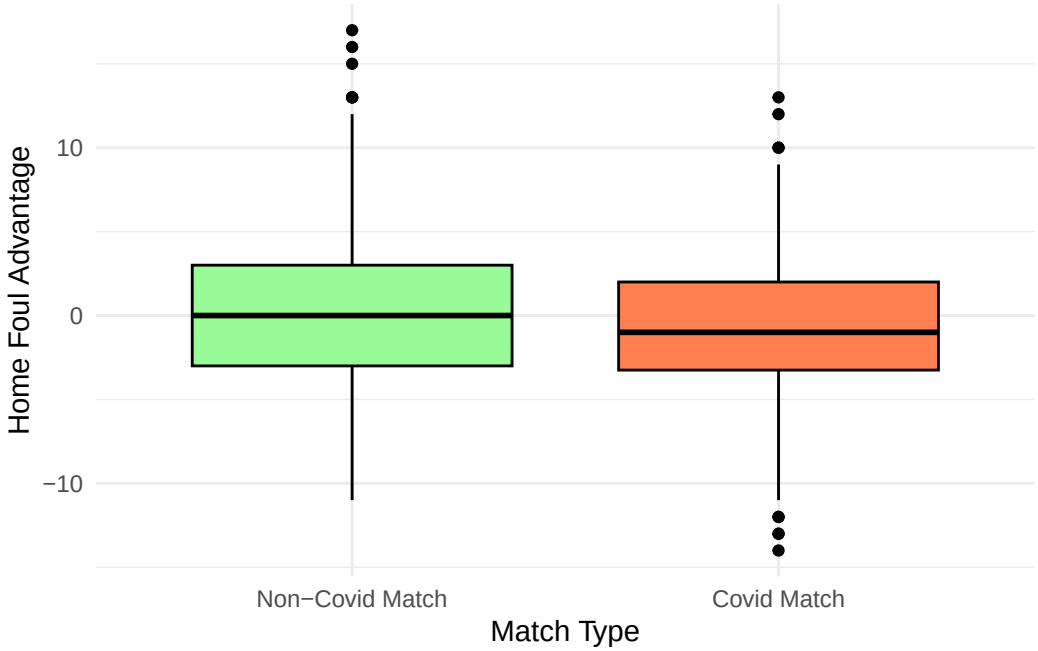


Figure 5: Boxplot of home foul advantage by audience presence

Table 6: Statistics of variable: home_foul_advantage by covid_match

Statistic	With_Audience	Without_Audience
Count	760.00	380.00
Std Dev	4.50	4.62
Mean	0.23	-0.62
Min	-11.00	-14.00
25%	-3.00	-3.50
Median	0.00	-1.00
75%	3.00	2.00
Max	17.00	13.00

Table 7: Possible outliers of home foul advantage by audience presence

Date	Home Team	Away Team	Home Foul Advantage	Audience
2023-05-06	manchester-city	leeds-united	13	With Audience
2022-12-26	everton	wolverhampton-wanderers	13	With Audience
2022-02-19	crystal-palace	chelsea	16	With Audience
2021-12-05	aston-villa	leicester-city	17	With Audience
2021-11-28	leicester-city	watford	15	With Audience
2021-11-07	arsenal	watford	13	With Audience
2021-02-28	leicester-city	arsenal	-12	Without Audience
2021-01-02	crystal-palace	sheffield-united	10	Without Audience
2021-01-01	manchester-united	aston-villa	-12	Without Audience
2020-12-28	chelsea	aston-villa	-13	Without Audience
2020-12-19	newcastle-united	fulham	10	Without Audience
2020-11-30	west-ham-united	aston-villa	-13	Without Audience
2020-11-30	leicester-city	fulham	12	Without Audience
2020-11-21	aston-villa	brighton-and-hove-albion	13	Without Audience
2020-09-27	sheffield-united	leeds-united	-14	Without Audience

Tell

As shown in Figure 3, matches played with audiences tend to have a slightly higher median and a wider IQR. Several large positive outliers indicate games where away teams committed noticeably more fouls than home teams. In contrast, matches without audiences have a lower and negative median, along with a few extreme values on both sides, suggesting that foul calls became more balanced, or even slightly against home teams, when no crowd was present. This result is quite surprising - home teams do not have an advantage with an audience and without an audience they have a disadvantage. It is possible there is some sort of selection effect happening, in which home teams know they have an advantage and therefore make riskier plays on the ball though this theory would need additional testing.

The outlier table also shows that most extreme positive values occurred with audiences, while negative or reversed cases appeared without audiences. None of these matches seem exceptional.

2b. Compare `home_yellow_advantage` by the variable `covid_match`

Think

The variable `home_yellow_advantage` measures the difference in yellow cards between the away and home teams (`away_yellow - home_yellow`). Positive values

mean that away teams received more yellow cards, while negative values mean that home teams were penalized more often.

Since yellow cards are a more serious referee decision than fouls, they may be more affected by psychological pressure from the crowd. I expect that with audiences, referees might be slightly more hesitant to give yellow cards to home players. Therefore, during the Covid-19 period when matches were played without crowds, I expect the home yellow advantage to be smaller on average, as referees were less influenced by audience reactions.

Because `home_yellow_advantage` is a quantitative variable and `covid_match` is categorical with two groups (With/Without Audience), side-by-side boxplots and summary tables are appropriate to compare their distributions and averages.

Show

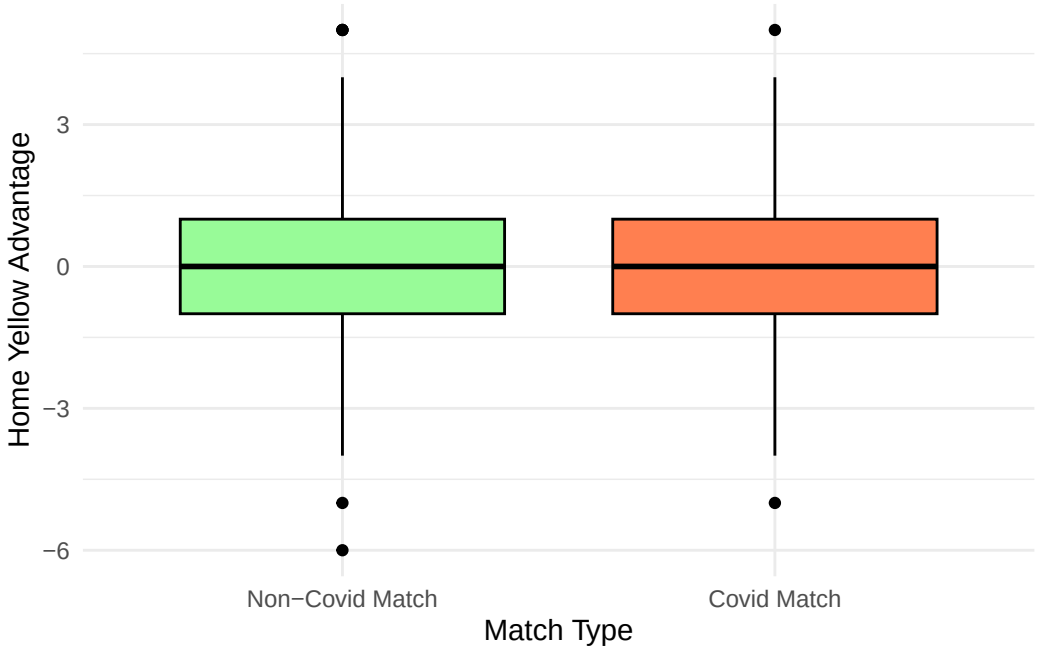


Figure 6: Boxplot of home yellow advantage by audience condition

Tell

As shown in Figure 5 and Table 4, the distributions of home_yellow_advantage are nearly identical across match types. Both with-audience and without-audience games have a median of 0 and an interquartile range from –1 to 1, indicating that referees typically issue similar numbers of yellow cards to home and away teams.

Table 8: Statistics of variable: Home Yellow Advantage

Statistic	With_Audience	Without_Audience
Count	760.00	380.00
Std Dev	1.65	1.55
Mean	0.19	0.03
Min	-6.00	-5.00
25%	-1.00	-1.00
Median	0.00	0.00
75%	1.00	1.00
Max	5.00	5.00

Table 9: Possible outliers of home_yellow_advantage by match type

Date	Home Team	Away Team	Home Yellow Adv.	Audience
2023-02-18	brighton-and-hove-albion	fulham	5	With Audience
2022-12-26	everton	wolverhampton-wanderers	5	With Audience
2022-11-13	brighton-and-hove-albion	aston-villa	5	With Audience
2022-08-28	nottingham-forest	tottenham-hotspur	-5	With Audience
2021-10-24	manchester-united	liverpool	-6	With Audience
2021-02-14	west-bromwich-albion	manchester-united	-5	Without Audience
2020-10-18	tottenham-hotspur	west-ham-united	5	Without Audience

The mean values (0.19 vs 0.03) and standard deviations (1.65 vs 1.55) differ only slightly, suggesting minimal change in officiating behavior during the Covid period.

According to Table 5, when audiences were present, away teams occasionally received several more yellow cards than home teams, suggesting a mild referee bias consistent with the home advantage effect. In contrast, matches played without audiences show fewer and more balanced extremes. Overall, the results partly meet my expectation: while the home yellow advantage appears slightly smaller without crowds, the difference is minimal. This suggests that although crowd pressure may influence referees in rare cases, yellow-card decisions remain largely consistent across match conditions. This result is overall a surprise - there does not appear to be any meaningful effect in terms of yellow cards.

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?

Answers will vary here, good quality effort to interpret investigation of this question is required.

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.

```
# Create subsets for Liverpool
liverpool_home <- epl %>%
  filter(home_team == "liverpool")

liverpool_away <- epl %>%
  filter(away_team == "liverpool")

# Create subsets for Manchester City
mancity_home <- epl %>%
  filter(home_team == "manchester-city")

mancity_away <- epl %>%
  filter(away_team == "manchester-city")
```

Table 10: Average goals, shots, possession, and pass % for Liverpool and Manchester City (home vs away)

Team	H Goals	A Goals	H Shots	A Shots	H Poss	A Poss	H Pass	A Pass
Liverpool	2.20	1.98	18.62	15.67	63.94	60.69	85.26	83.76
Manchester City	2.79	1.98	17.47	15.89	66.64	64.83	89.69	88.46

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.

```
team_avg <- data.frame(
  Team = c("Liverpool", "Manchester City"),
  Home_Goals = c(mean(liverpool_home$home_goals, na.rm = TRUE),
                 mean(mancity_home$home_goals, na.rm = TRUE)),
  Away_Goals = c(mean(liverpool_away$away_goals, na.rm = TRUE),
                 mean(mancity_away$away_goals, na.rm = TRUE)),
  Home_Shots = c(mean(liverpool_home$home_shots, na.rm = TRUE),
                 mean(mancity_home$home_shots, na.rm = TRUE)),
  Away_Shots = c(mean(liverpool_away$away_shots, na.rm = TRUE),
                 mean(mancity_away$away_shots, na.rm = TRUE)),
  Home_Possession = c(mean(liverpool_home$home_possessions, na.rm = TRUE),
                      mean(mancity_home$home_possessions, na.rm = TRUE)),
  Away_Possession = c(mean(liverpool_away$away_possessions, na.rm = TRUE),
                      mean(mancity_away$away_possessions, na.rm = TRUE)),
  Home_Pass = c(mean(liverpool_home$home_pass, na.rm = TRUE),
                mean(mancity_home$home_pass, na.rm = TRUE)),
  Away_Pass = c(mean(liverpool_away$away_pass, na.rm = TRUE),
                mean(mancity_away$away_pass, na.rm = TRUE))
)


kbl(team_avg, digits = 2,
    col.names = c("Team", "H Goals", "A Goals", "H Shots", "A Shots", "H Poss",
                  "A Poss", "H Pass", "A Pass")) %>%
  kable_styling()
```

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.

```
league_mean <- epl %>%
  summarise(
    Home_Goals = mean(home_goals, na.rm = TRUE),
    Away_Goals = mean(away_goals, na.rm = TRUE),
    Home_Shots = mean(home_shots, na.rm = TRUE),
    Away_Shots = mean(away_shots, na.rm = TRUE),
    Home_Possession = mean(home_possessions, na.rm = TRUE),
    Away_Possession = mean(away_possessions, na.rm = TRUE),
    Home_Pass = mean(home_pass, na.rm = TRUE),
    Away_Pass = mean(away_pass, na.rm = TRUE)
  )

league_sd <- epl %>%
  summarise(
    Home_Goals = sd(home_goals, na.rm = TRUE),
    Away_Goals = sd(away_goals, na.rm = TRUE),
    Home_Shots = sd(home_shots, na.rm = TRUE),
    Away_Shots = sd(away_shots, na.rm = TRUE),
    Home_Possession = sd(home_possessions, na.rm = TRUE),
    Away_Possession = sd(away_possessions, na.rm = TRUE),
    Home_Pass = sd(home_pass, na.rm = TRUE),
    Away_Pass = sd(away_pass, na.rm = TRUE)
  )

z_row <- (team_avg[1, -1] - league_mean) / league_sd
z_row2 <- (team_avg[2, -1] - league_mean) / league_sd

team_avg_z <- rbind(
  team_avg,
  c("Liverpool (z-score)", as.numeric(z_row)),
  c("Manchester City (z-score)", as.numeric(z_row2))
)

team_avg_z[, -1] <- lapply(team_avg_z[, -1], as.numeric)

kbl(team_avg_z, digits = 2,
    col.names = c("Team", "H Goals", "A Goals", "H Shots", "A Shots", "H Poss",
```

Table 11: Average goals, shots, possession, pass %, and z-score row relative to league averages

Team	H Goals	A Goals	H Shots	A Shots	H Poss	A Poss	H Pass	A Pass
Liverpool	2.20	1.98	18.62	15.67	63.94	60.69	85.26	83.76
Manchester City	2.79	1.98	17.47	15.89	66.64	64.83	89.69	88.46
Liverpool (z-score)	0.51	0.56	0.90	0.83	1.02	0.89	0.75	0.65
Manchester City (z-score)	0.95	0.56	0.70	0.87	1.23	1.21	1.34	1.29

```
                  "A Poss", "H Pass", "A Pass")) %>%
  kable_styling()
```

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?

> Both Liverpool and Manchester City have positive z-scores across all offensive metrics, showing that they perform above the league average in goals, shots, possession, and pass accuracy. Manchester City has higher z-scores overall, especially in possession and passing, suggesting stronger overall control.
>
> Liverpool's shot z-scores are higher at home (0.90) than away (0.83), suggesting they create more shooting chances at home. In contrast, Manchester City's away shot z-score (0.87) is slightly higher than at home (0.70), indicating they maintain strong offensive pressure even when playing away. Overall, both teams demonstrate balanced performance across venues, with Manchester City showing the strongest overall offensive profile, particularly in possession and passing.
>
> To better understand their offensive capabilities, it would be helpful to have additional data such as shot accuracy or conversion rate. These could reveal whether high shot volume effectively translates into goals.

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

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

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

Answers will vary here, good quality effort to interpret investigation of this question is required.

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.

Need to think deeply about what information this dataset provides for full points.