Performance of Arsenal and Tottenham Against Big 6 Clubs: A Goals-Based Analysis (2009-2019)

Younseo Kim

May 05, 2023

Executive Summary

The study was designed in order to explore the performance of Tottenham and Arsenal against the Premier League Big 6 over a decade of time. Using Tottenham and Arsenal, their location (Home and Away), their matchup (opponent and location), and 10 seasons of data, it was discovered that the interaction term of team and location had statistical significance. However, additional post hoc analysis revealed that no single individual combination of team and location was statistically significant, indicating the possible presence of a Type 1 Error. Several limitations of the design and suggestions for addressing them are discussed and offered.

Introduction and Background

Since the formation of the English Premier League in 1992, the idea of the "Big 6" has been one of constant controversy and evolution. The Big 6 represents the top 6 Premier League clubs that are competitive and dominant for a certain era. The initial Big 6 in the 90's contained the likes of Arsenal, Everton, Liverpool, Manchester United, Tottenham and Blackburn[1]. This listed has changed over time, with certain clubs leaving and entering. However, since 2009, the era that I began watching soccer, the Big 6 has remained consistent and dominant- formed by Arsenal, Tottenham, Manchester City, Chelsea, Liverpool, and Manchester United. The goal for this study is to see how two specific teams- Arsenal, and Tottenham performed against the big six in the modern era from 2009 to 2019.

Research Questions

For this study, there are multiple questions I hope to be able to answer.

1) Does location (either home or away) affect either team's performance against the big 6?
   - $H_{1,0}$: There is no statistically significant impact on performance due to location (home or away).
   - $H_{1,A}$: There is a statistically significant impact on performance due to location (home or away).
2) Does performance change significantly by season?
   - $H_{2,0}$: There is no statistically significant impact on performance due to season.
   - $H_{2,A}$: There is a statistically significant impact on performance due to season.

[1]Graham, M. (2022, June 21). Premier league big six: How did the balance of power in English football evolve? Planet-Sport. Retrieved May 5, 2023, from https://www.planetsport.com/soccer/news/premier-league-big-six-balance-power-english-football-evolve

3) Which team has performed better overall against the Big 6 clubs?

- $H_{3,0}$: There is no statistically significant impact on performance due to team (Tottenham or Arsenal).
- $H_{3,A}$: There is a statistically significant impact on performance due to team (Tottenham or Arsenal.)

Study Design and Methods

To answer these questions, I first collected a comprehensive set of data from all Premier League games from 2009 to 2019. I then cleaned the data to contain only games in which either Arsenal or Tottenham played against a big 6 club. To avoid interaction between the two, I elected not to include games where Arsenal and Tottenham played each other. I then had to decide how to quantify the idea of performance. I designed the study in such a way that the goals scored by either Tottenham or Arsenal in each individual game would quantify the team's performance. So, with a complete data set containing the season in which a game was played, which team played it, their location (home or away), the number of goals scored, and the specific matchup (who they played against), I had all the data necessary to begin the study. It is valuable to note that I do not have a random effect in the model, as I did not randomly select the teams, seasons or matchups. This was by design, as I wanted to look at these teams specifically in the era I grew up in, and there was not enough Big 6 data to random sample.

Analytical Methods

To analyze the data and answer the research questions I will use R and make use of ANOVA methods, in particular an adjusted version of a nested repeated measures design to fit the design of the study.

Appropriateness of ANOVA

Given that the response of interest is performance measured by goals, I have a continuous response. There is a case to be made that goals scored is a discrete variable, but I argue that goals scored can take on any value between 0 and the maximum number of goals that can be scored in a game, which is theoretically unlimited, so I can treat goals scored as a continuous response. Additionally, I have multiple factors- Location and Team- as well as the time point for repeated measures, Season. These factors are all categorical in nature.

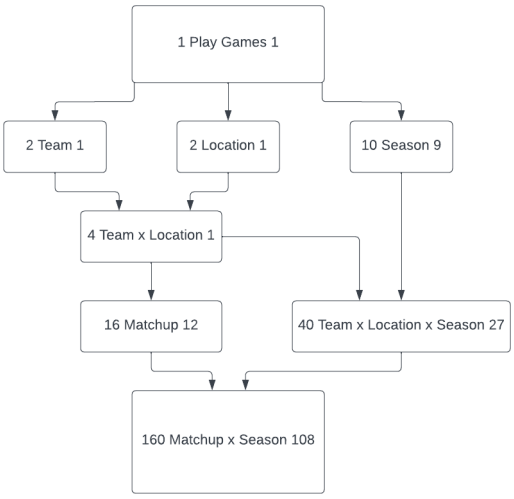

Figure 1: Hasse Diagram for the Premier League Performance Study

Figure 1 shows the Hasse diagram for the study. I can see the factors, Team and Location (fixed effects) as well as their interaction. Nested within this interaction is matchup, which represents the 16 total combinations of Arsenal and Tottenham against the other four Big 6 clubs, both home and away. Furthermore, given the sample size of 160 total games, I have sufficient *degrees of freedom* to estimate all main effects, interactions, and error terms. These elements point towards the appropriateness of ANOVA methods in answering the research questions. In particular, I will make use of a nested repeated measures model. This ANOVA model is the most appropriate as I have my factors of interest and their interaction, and nested in this interaction are the matchups which are being measured across 10 different seasons.

I made the choice to control the overall Type I risk at 5%. For multiple comparisons, I will control the False Discovery Rate at this level by using the Benjamini-Hochberg method. Within each hypothesis test, I will use an Unusualness Threshold equivalent to 5%. I elected to take this approach due to the fact that the experiment does not have any real-world implications, so I have no reason to be overly conservative and can take a relatively liberal approach.

Exploratory Data Analysis

Table 1: Summary Statistics for Premier League Performance Study- Team Goals by Location

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
Away:Arsenal	40	0	0.75	1	2	5	1.483	1.275	1.086	0.981	1.473
Away:Tottenham	40	0	0.00	1	2	3	1.483	0.950	1.012	0.531	-1.094
Home:Arsenal	40	0	0.00	1	2	4	1.483	1.225	1.121	0.524	-0.757
Home:Tottenham	40	0	0.75	1	2	5	1.483	1.525	1.320	0.697	-0.261

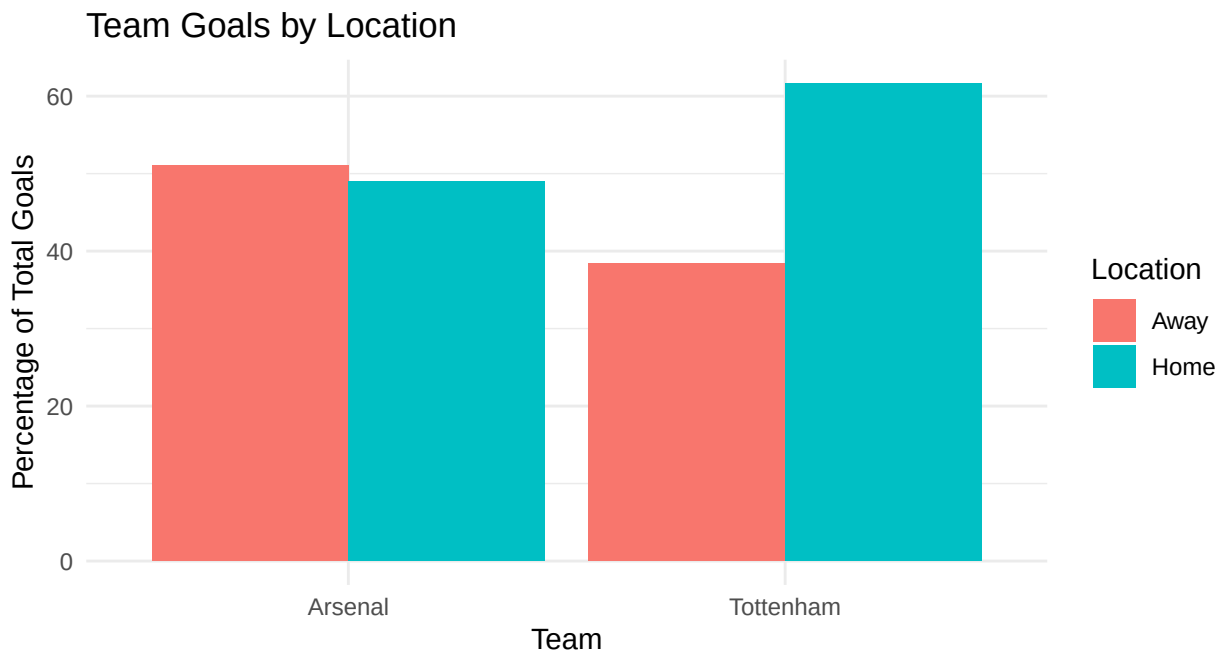

Figure 2: Bar Charts for Team Goals by Location

Table 1 shows the values of various descriptive statistics broken out by location. At a glance, using *Sample Arithmetic Mean* (*SAM*) as a measure of performance, it seems as though Tottenham performs worse away than at home, yet location does not seem to matter for Arsenal. This is not consistent with what I expected- that both teams would perform worse away. Comparing across teams, it is interesting to note that Arsenal seems to perform better away against the Big 6 than Tottenham does, yet Tottenham seems to perform better than Arsenal at home.

These interpretations are further confirmed by Figure 2, which gives a more visual representation of the statistics provided in Table 1.

Table 2: Summary Statistics for Premier League Performance Study- Team Goals by Season

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
2009-to-2010:Arsenal	8	0	0.00	1.0	1.25	2	1.483	0.875	0.835	0.182	-1.721
2009-to-2010:Tottenham	8	0	0.75	1.0	2.00	3	1.483	1.250	1.035	0.254	-1.377
2010-to-2011:Arsenal	8	0	0.00	1.0	1.50	3	1.483	1.125	1.246	0.575	-1.471
2010-to-2011:Tottenham	8	0	0.00	0.5	1.25	2	0.741	0.750	0.886	0.404	-1.754
2011-to-2012:Arsenal	8	0	0.00	1.0	2.00	5	1.483	1.375	1.685	1.051	-0.151
2011-to-2012:Tottenham	8	0	0.00	1.0	1.25	4	1.483	1.125	1.356	1.010	-0.276
2012-to-2013:Arsenal	8	0	1.00	1.0	1.25	2	0.000	1.125	0.641	-0.044	-0.943
2012-to-2013:Tottenham	8	1	1.75	2.0	2.25	3	0.741	2.000	0.756	0.000	-1.469
2013-to-2014:Arsenal	8	0	0.00	0.5	1.25	3	0.741	0.875	1.126	0.731	-1.107
2013-to-2014:Tottenham	8	0	0.00	0.5	1.25	2	0.741	0.750	0.886	0.404	-1.754
2014-to-2015:Arsenal	8	0	0.75	1.5	2.00	4	0.741	1.500	1.309	0.501	-0.894
2014-to-2015:Tottenham	8	0	0.00	0.0	1.25	5	0.000	1.000	1.773	1.346	0.316
2015-to-2016:Arsenal	8	0	0.00	2.0	2.25	3	1.483	1.500	1.309	-0.167	-1.915
2015-to-2016:Tottenham	8	0	0.00	1.5	2.25	4	2.224	1.500	1.512	0.326	-1.576
2016-to-2017:Arsenal	8	1	1.00	1.5	2.25	3	0.741	1.750	0.886	0.404	-1.754
2016-to-2017:Tottenham	8	0	0.75	1.5	2.00	2	0.741	1.250	0.886	-0.404	-1.754
2017-to-2018:Arsenal	8	0	0.00	1.0	1.25	3	1.483	1.000	1.069	0.614	-1.086
2017-to-2018:Tottenham	8	0	1.00	1.5	2.25	4	0.741	1.750	1.282	0.401	-1.221
2018-to-2019:Arsenal	8	0	1.00	1.5	2.00	2	0.741	1.375	0.744	-0.541	-1.269
2018-to-2019:Tottenham	8	0	0.00	0.5	1.50	3	0.741	1.000	1.309	0.668	-1.469

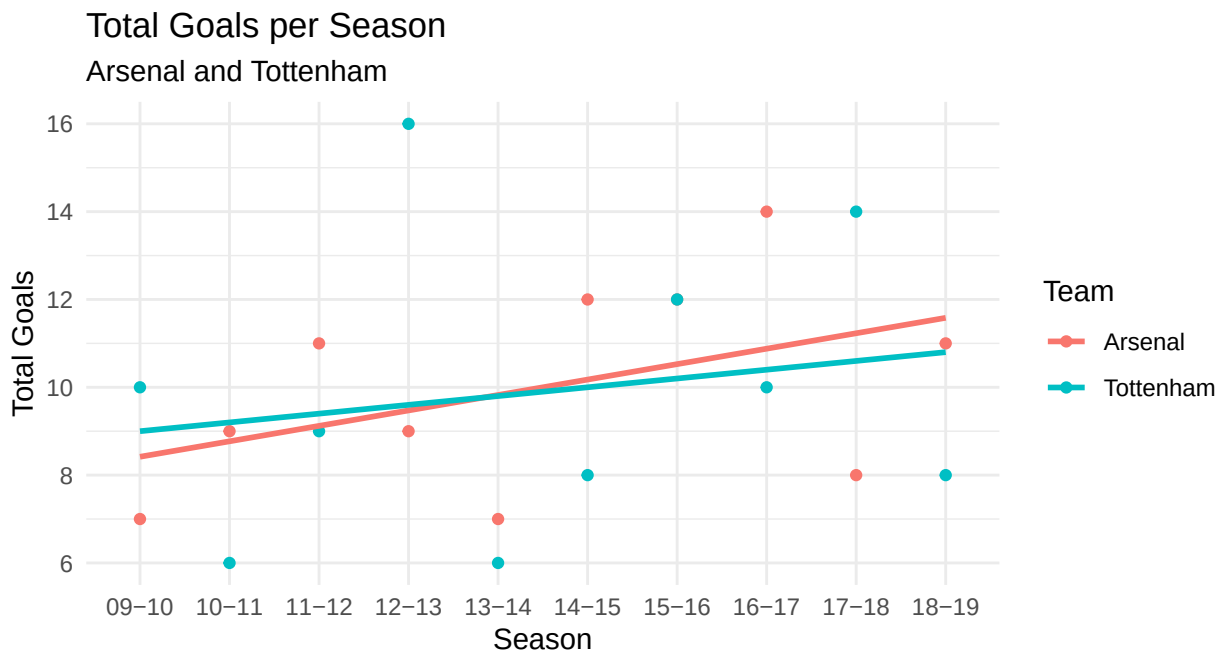

Figure 3: Bar Charts for Team Goals by Season

Table 2 shows the values of various descriptive statistics broken out by team and season. These statistics hope to provide valuable insight towards answering the second research question: Does performance change significantly by season? Upon initial inspection, it seems as though these teams had some seasons where they performed better than others. It is interesting to note that in 2012-2013, Tottenham had a *Sample Arithmetic Mean* value of 2 goals per game in their 8 matchups, which was the highest value across all seasons for both teams. Tottenham also had a *Sample Minimum* value of 1 in that same season, indicating they succeeded in scoring in all 8 of their matchups against the other four Big 6 teams.

Figure 3 shows the total goals scored in each season for Arsenal and Tottenham against the other big 6 clubs, from the 2009-2010 season to the 2018-2019 Premier League season. The reference line for each team represents the trend of performance across these seasons. I can observe that total goals by season varies and does not seem to follow any sort of trend, but the reference lines both have small positive slopes. Figure 3 supplements Table 2 quite well.

Table 3: Summary Statistics for Premier League Performance Study- Goals by Matchup

	n	Min	Q1	Median	Q3	Max	MAD	SAM	SASD	Sample Skew	Sample Ex. Kurtosis
Arsenal at Chelsea	10	0	0.00	0.5	2.00	3	0.741	1.1	1.287	0.400	-1.734
Arsenal at Liverpool	10	0	1.00	1.5	2.75	4	1.483	1.7	1.337	0.241	-1.402
Arsenal at Man City	10	0	0.00	0.5	1.75	2	0.741	0.8	0.919	0.340	-1.840
Arsenal at Man United	10	0	1.00	1.0	1.75	3	0.000	1.3	0.823	0.581	-0.445
Chelsea at Arsenal	10	0	0.00	0.0	1.00	5	0.000	0.9	1.595	1.623	1.448
Chelsea at Tottenham	10	0	0.00	0.5	1.75	3	0.741	0.9	1.100	0.621	-1.251
Liverpool at Arsenal	10	0	1.00	1.5	2.00	3	0.741	1.5	0.850	0.000	-0.963
Liverpool at Tottenham	10	0	0.00	1.0	2.00	2	1.483	1.0	0.943	0.000	-1.988
Man City at Arsenal	10	0	1.00	1.5	2.00	3	0.741	1.6	0.966	0.080	-1.297
Man City at Tottenham	10	0	0.25	1.0	1.75	2	1.483	1.0	0.817	0.000	-1.650
Man United at Arsenal	10	0	1.00	1.0	1.75	2	0.741	1.1	0.738	-0.120	-1.348
Man United at Tottenham	10	0	0.00	0.0	1.75	3	0.000	0.9	1.287	0.727	-1.384
Tottenham at Chelsea	10	0	1.00	1.5	2.00	5	0.741	1.8	1.398	0.974	0.114
Tottenham at Liverpool	10	0	0.25	1.5	2.00	4	1.483	1.6	1.506	0.443	-1.319
Tottenham at Man City	10	0	0.25	1.0	2.75	4	1.483	1.5	1.434	0.407	-1.470
Tottenham at Man United	10	0	0.25	1.0	2.00	3	1.483	1.2	1.033	0.196	-1.422

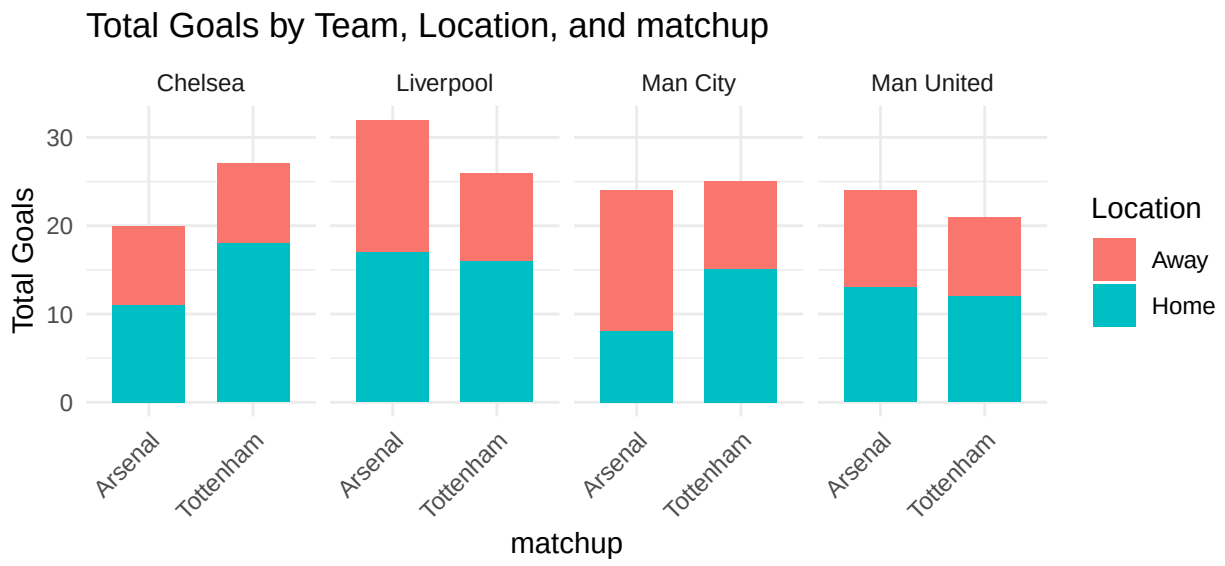

Figure 4: Stacked Bar Charts- Total Goals by Team, Locaion and Opponent

Table 3 shows the values of various descriptive statistics broken out by matchup. There are a few interesting elements of this data. Arsenal's away games against Chelsea have a sample arithmetic mean value of just .9, the second lowest among all of the matchups. Yet, at the same time, this matchup has the highest *Sample Maximum* value of 5! Figure 4 shows total goals, split by location, of Arsenal and Tottenham against the other big 6 clubs from 2009 to 2019 in the Premier League. From this, I can examine that Arsenal had scored more goals against Liverpool and Man United than Tottenham has, while Tottenham had scored more goals against Chelsea than Arsenal. As for Man City, both Arsenal and Tottenham scored a similar number of goals. In addition, I can observe that Tottenham had scored significantly more goals at Home when they are against Chelsea, and Arsenal had scored more goals Away when they are against Man City. Speaking generally, Table 3 and Figure 4 indicate that there may be something to explore regarding performance in these different matchups.

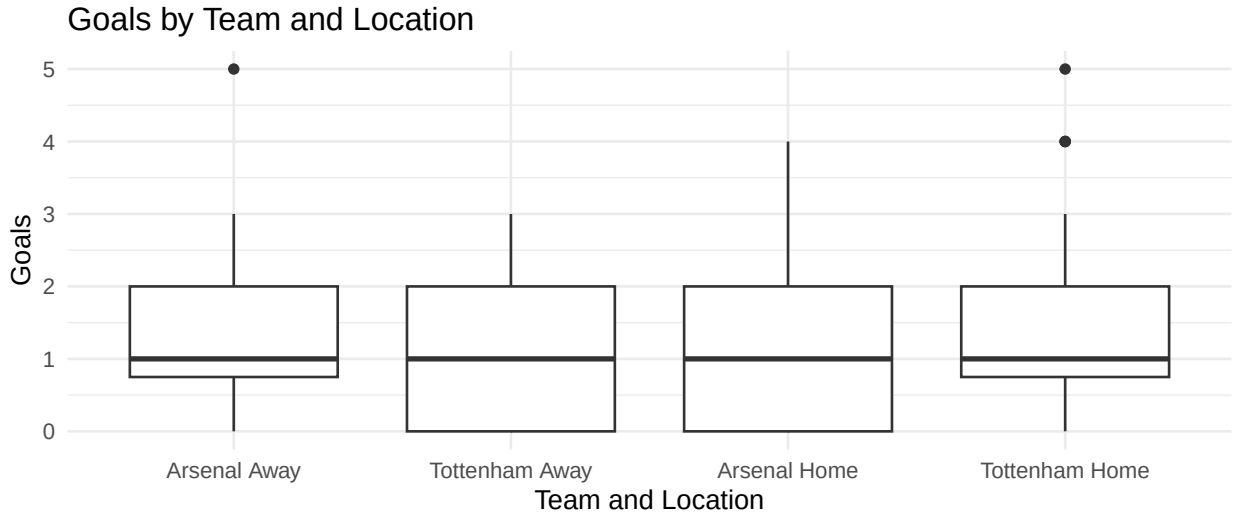

Figure 5: Side-by-side Box Plots of Team goals by Location

Boxplots are a great way to visualize data. Figure 5 shows side-by-side boxplots of goals by team and

location. The plots show some potential outliers involving Arsenal away games and Tottenham home games. Given the nature of the data, I should proceed with these outliers. While it is indeed uncommon for a team to score five goals in one game, it is not so outlandish that I would consider removing them from the dataset.

Results

I present the results in three sections. First, I will discuss the assumptions of the parametric shortcut. Then I will move on to answering the omnibus questions before ending this section with post hoc analysis, if appropriate.

Assumptions

To use a parametric shortcut for the model, I need to satisfy four assumptions: the model residuals need to follow a Gaussian distribution, I need homoscedasticity (around the model), I need independent subjects, and I need sphericity, which indicates having the same level of variance among treatment differences. If any of these assumptions are violated, then the inference results will not be trustworthy.

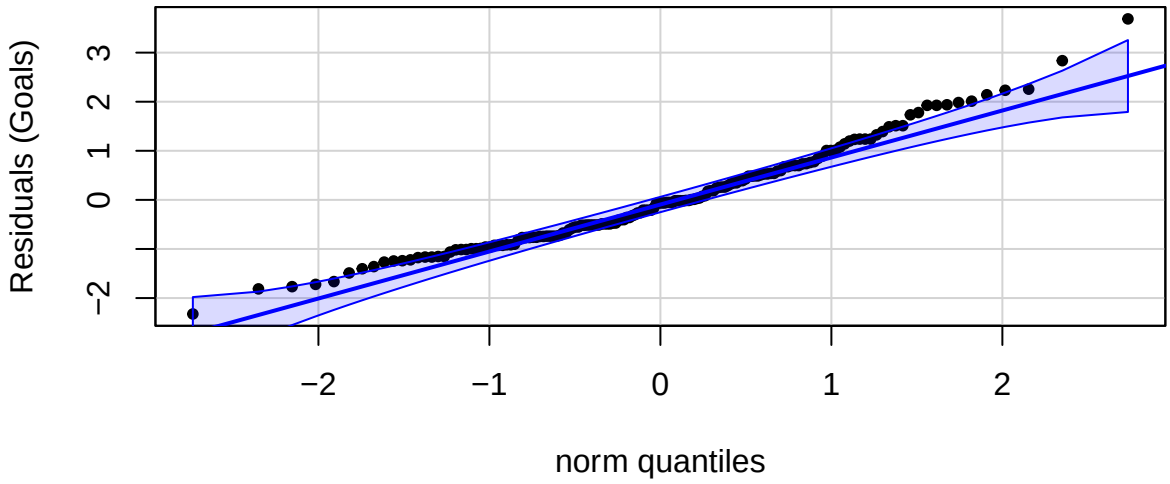

Figure 6: QQ Plot for Premier League Performance Study

The first assumption I will discuss is that of the residuals following a Gaussian distribution. Figure 6 shows the quantile-quantile plot of the residuals against a Gaussian distribution. I have included a 90% confidence envelope to help identify points which might stray too far from the reference line of perfect fit. While I do have some points that lie outside of this confidence envelope, the vast majority of the points lie comfortably within the envelope, and the percentage of points outside of the envelope is certainly less than 10%. I can say that the Gaussian Residuals assumption is satisfied.

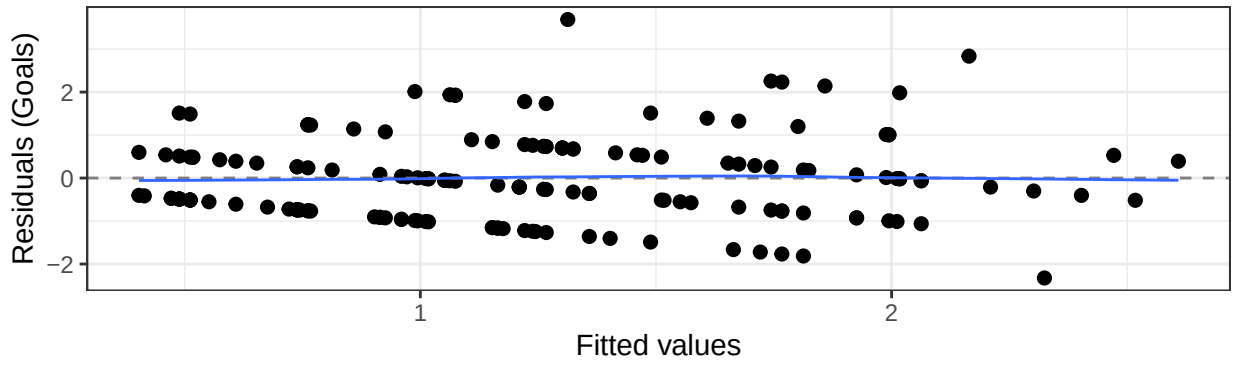

Figure 7: Tukey-Anscombe Plot for Premier League Performance Study

To assess whether I have homoscedasticity around the model, I will use a Tukey-Anscombe plot shown in Figure 7. As I look across the plot, I want to assess whether there is a pattern that relates the fitted value to the residuals. This plot looks quite concerning at a glance. There are clear distinct parallel lines that are indicative of a pattern relating the fitted values to residuals. However, when thinking about the study design, this makes sense. Since I only have 6 distinct integer values for the response (goals), a pattern like this is likely. It is more important to note that the blue line is relatively smooth and horizontal, indicating their is no true major issue regarding this assumption. With this in mind, I will say, cautiously, that the homoscedasticity assumption is satisfied.

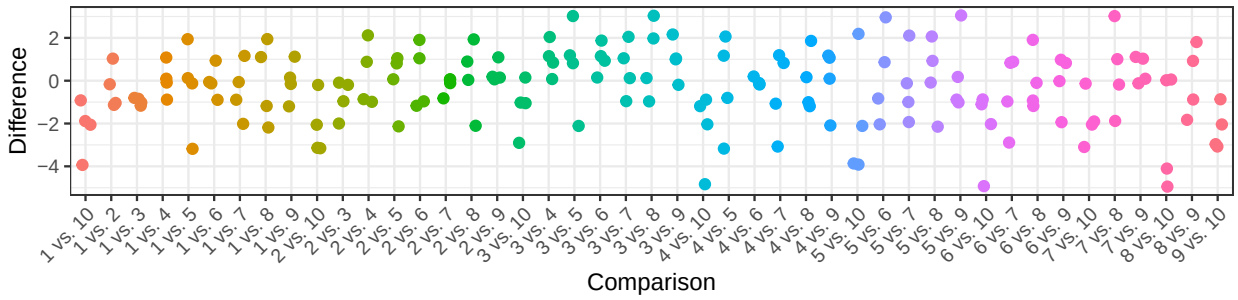

Figure 8: Sphericity Plot for Premier League Performance Study

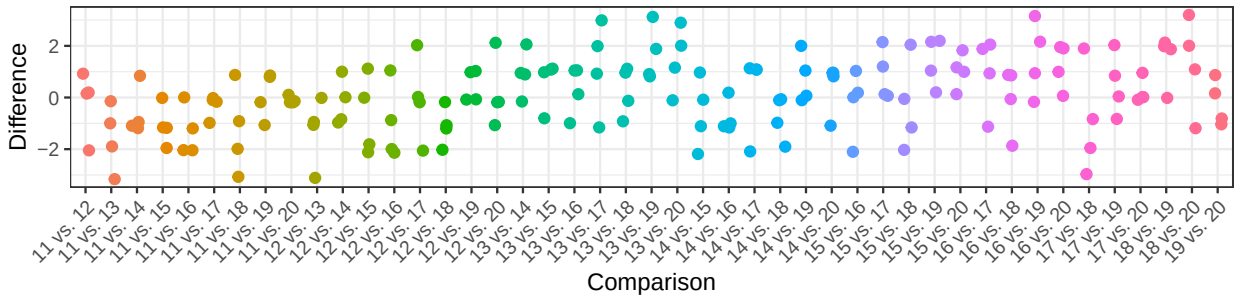

Figure 9: Sphericity Plot for Premier League Performance Study

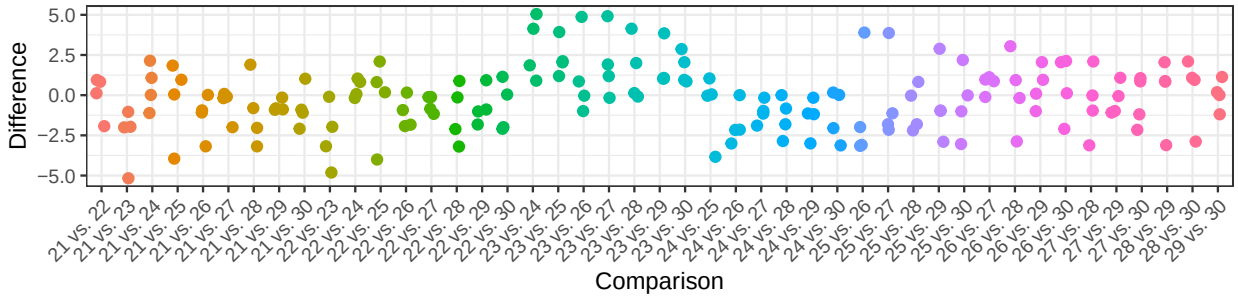

Figure 10: Sphericity Plot for Premier League Performance Study

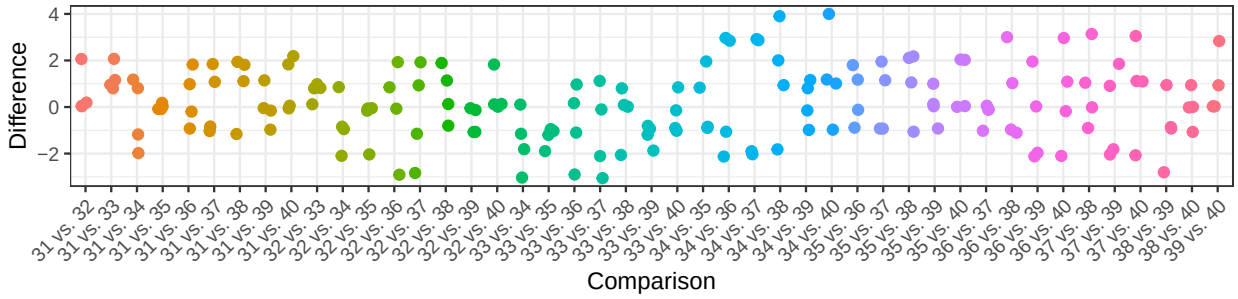

Figure 11: Sphericity Plot for Premier League Performance Study

When assessing sphericity, I will want to see if any difference has excessively different variation than another difference. To do this, I will use a sphericity plots shown above. Due to the large number of interactions, I split the sphericity plot into four to improve readability. Within these four plots, there are no clusters of points that indicate an excessively different variation from another. This allows to say that the sphericity assumption is satisfied.

When assessing independence of subjects, I need to turn to the design of the study. The question at hand is whether or not each of the specified matchups are independent of each other. The EPL ensures that each matchup has a time gap of at least one week, which allows changes in factors such as player injuries that may affect a game to not carryover between games. Moreover, the result of an individual matchup is highly unpredictable due to many variables such as refereeing and weather condition. Thus, the unpredictability contributes to the indepdencne of each matchup. Considering these factors, I believe that the independence of subjects assumption is satisfied.

Omnibus Results

Table 4 shows the ANOVA table for the model after using the parametric shortcut. Starting in the first row, I see that the main effect of team accounts for approximately .008 times as much variation as the residuals. Similarly, the main effect of location accounts for approximately 3.47 times as much variation. The interaction of these main effects, team and location, accounts for roughly 4.92 times as much variation as what is left unexplained by the model. For team, I would anticipate observing a F-ratio at least as large as .008 around 93% of the time. For location, I would anticipate observing a F-ratio at least as large as 3.47 around 8.7% of the time. For the interaction, I would anticipate observing a F-ratio at least as large as 4.92 around 4.67% of the time. Of these, only the interaction term appears to have statistical significance (p-values less than the Unusualness Threshold of 0.05). It is also valuable to note that season, the time point, accounts for approximately .70 times as much variation as the residuals. The table shows that I would

anticipate observing a F-ratio at least as large as .70 around 70.6% of the time. This is useful in answering the second SRQ- Does performance change significantly by season?

Table 4: ANOVA Table for Premier League Study

Source	df	SS	MS	F	p-value
team	1	0.0063	0.0063	0.0079	0.9308
location	1	2.7563	2.7563	3.4724	0.0870
team:location	1	3.9063	3.9063	4.9213	0.0466
matchup	12	9.5250	0.7937	0.5540	0.8739
season	9	9.0563	1.0063	0.7024	0.7055
team:location:season	27	29.5187	1.0933	0.7631	0.7879
matchup:season	108	154.7250	1.4326		

Point Estimates

Table 5: Point Estimates from Premier League Performance Study- Team and Location

Term	Estimate
Grand Mean	1.24
Arsenal	0.01
Home	-0.13
Arsenal:Home	0.16

Table 5 tells that both Arsenal and Tottenham score 1.24 goals per game in matches against the other top 6 clubs (Chelsea, Liverpool, Manchester City, and Manchester United). This is the estimate for baseline performance. I can also observe treatment effects. Arsenal at Home had scored additional 0.16 goals than the baseline performance. It is also important to note that treatment effects of Tottenham, Away, and any interactions including them had point estimate values of 0, and are not included in the above table.

Table 6: Point Estimates from Premier League Performance Study- Season

Term	Estimate
Grand Mean	1.24
2009-2010	0.47
2010-2011	-0.78
2011-2012	0.22
2012-2013	0.47
2013-2014	-0.52
2014-2015	-0.27
2015-2016	0.22
2016-2017	0.23
2017-2018	0.48
2018-2019	0.16

Table 6 tells that, in general, Arsenal and Tottenham will score 1.24 goals per game against the other big 6 clubs (Chelsea, Liverpool, Man City, Man United). This is the estimate for baseline performance. I can

also observe effects of the time point estimates. For the 2010-2011 season, Arsenal and Tottenham scored -0.78 goals per game than the baseline performance, which is the worst performed season across 2009 to 2019. Additionally, for 2017-2018 season, Arsenal and Tottenham scored +0.48 goals above the baseline performance, which is the best performance in a season across the decade of interest.

Post Hoc

Given that I have a statistically significant impact regarding the interaction of team and location, I can begin the Post Hoc Analysis. This entails analyzing all pairwise comparisons of team and location in order to give a better understanding of the first research question research question- Does location (either home or away) affect either team's performance against the big 6?

Table 7: Pairwise Comparisons of Teams by Location

Comparison	Estimate	SE	DF	t ratio	p-value
Arsenal Away - Tottenham Away	0.325	0.2959	12	1.0984	0.4976
Arsenal Away - Arsenal Home	0.050	0.2959	12	0.1690	0.8686
Arsenal Away - Tottenham Home	-0.250	0.2959	12	-0.8449	0.4976
Tottenham Away - Arsenal Home	-0.275	0.2959	12	-0.9294	0.4976
Tottenham Away - Tottenham Home	-0.575	0.2959	12	-1.9433	0.4548
Arsenal Home - Tottenham Home	-0.300	0.2959	12	-1.0139	0.4976

Table 7 shows the post hoc pairwise comparisons of all team and location interactions. I have used the Benjamini & Hochberg method for controlling the False Discovery Rate. It is interesting to note that while the location and team interaction term showed significance through the omnibus test, not a single individual combination of location and team is statistically significant. I will discuss this in detail in the following section.

Discussion

Moving through the initial research questions, I can now draw conclusions.

1) Does location (either home or away) affect either team's performance against the big 6?

- $H_{1,0}$: There is no statistically significant impact on performance due to location (home or away).
- $H_{1,A}$: There is a statistically significant impact on performance due to location (home or away).

This question is very difficult to answer. Through the omnibus test, the interaction of team and location showed statistical significance. However, I saw through post hoc analysis that no individual of combination of team and location was statistically significant. There is a legitimate concern that I committed a type 1 error here- especially considering that both main effect terms that formed the interaction were not statistically significant through the omnibus test, yet the interaction term was. It is still quite possible that location really does have a true, significant effect on performance- yet there could be limitations within the study design that masked this. One potential example of this is that selecting only two teams limited the data in such a way that I did not have enough to show the true effect. Overall, I cannot make a decision on this research question at this time, but work can be done in the future change this.

2) Does performance change significantly by season?

- $H_{2,0}$: There is no statistically significant impact on performance due to season.
- $H_{2,A}$: There is a statistically significant impact on performance due to season.

Given the omnibus test results and season point estimates, I fail to reject the null hypothesis here. It appears as though season does not have a significant impact on performance for either team. This appeared to be the case when I was exploring the data, as there were no clear indicators through the graphs or summary statistics that indicated there may be something significant in that regard. The absence of an extreme change in performance by season for both teams is certainly related to their status as a Big 6 club, as these teams are historically consistent and dominant and it would be out of character to see stark change in performance through seasons.

3) Which team has performed better overall against the Big 6 clubs?

   - $H_{3,0}$: There is no statistically significant impact on performance due to team (Tottenham or Arsenal).
   - $H_{3,A}$: There is a statistically significant impact on performance due to team (Tottenham or Arsenal.)

I once again fail to reject the null hypothesis. These teams performed too similarly across the decade of interest for there to be a statistically significant difference in their performance. This was hinted at in the exploratory data analysis section when I noticed that Arsenal scored 100 goals to Tottenham's 99. Given that I was quantifying performance as goals in individual games, and these goal sum values are just the sum of goals across all games, it makes sense why team had no statistically significant impact on performance. Perhaps comparing a Big 6 Club to a non Big 6 club would yield a different result.

Limitations

There were quite a few limitations in the study design. One such limitation was the absence of a random effect. This results in restrictions regarding the ability to generalize findings to a sample population. Without having a random effect, I could only draw conclusions regarding the two specific teams I selected and the 10 seasons I selected. Selecting only two teams was another limitation of the study design, as mentioned previously. This is perhaps the underlying reason that the main effect "team" did not have a statistically significant impact on performance. Perhaps comparing Tottenham and Arsenal to more teams, or across more seasons good have given a better understanding of who performs better. A final potential limitation in the study design is the lack of control for confounding variables. There are many confounding variables- injuries, change in stadium, change in staff, and transfer of ownership, to name a few- that could effect a teams performance from season to season. The fact that these could not be accounted for in the model is a limitation.

Future Work

These limitations can be addressed in future work. For a future study design, I could elect to randomize the seasons I analyze, the teams, or both in an effort to incorporate a random effect into the model. I could also add more teams to the model, giving more data and mitigating the shortcomings that resulted from the choice to only analyze two teams. Future work may also entail brainstorming ways to account for confounding variables in a future model.

References and Materials Consulted

FIFA 2022 dataset CSVS (19k+ players, 100+ attributes). Sports Statistics Sports Data SportsStatisticscom. (n.d.). Retrieved May 5, 2023, from https://sports-statistics.com/soccer/

Graham, M. (2022, June 21). Premier league big six: How did the balance of power in English football evolve? PlanetSport. Retrieved May 5, 2023, from https://www.planetsport.com/soccer/news/premier-league-big-six-balance-power-english-football-evolve

Code Appendix

```
# Setting Document Options ----
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  message = FALSE,
  fig.align = "center",
  dpi = 300 # helps create higher quality graphics
)

# Add additional packages by name to the following list ----
packages <- c(
  "tidyverse", "knitr", "kableExtra", "hasseDiagram",
  "psych", "car", "parameters", "lme4"
  )
lapply(X = packages, FUN = library, character.only = TRUE, quietly = TRUE)
library(boastUtils)
# Loading Helper Files and Setting Global Options ----
options(knitr.kable.NA = "")
options("contrasts" = c("contr.sum", "contr.poly"))

source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/ANOVATools.R")

source("https://raw.github.com/neilhatfield/STAT461/master/rScripts/shadowgram.R")

library(readxl)
goalData <- read_excel("data/goalData.xlsx")
View(goalData)

# filter only Arsenal and Tottenham
targetTeams <- c("Arsenal", "Tottenham", "Liverpool", "Man City", "Man United", "Chelsea")
at <- c("Arsenal", "Tottenham")
goalData <- goalData %>%
  filter(HomeTeam %in% targetTeams & AwayTeam %in% targetTeams) %>%
  filter(HomeTeam %in% at | AwayTeam %in% at) %>%
  filter(!(HomeTeam %in%  at & AwayTeam %in% at)) %>%
  mutate(
    Source.Name = str_remove(string = Source.Name, pattern = "england-premier-league-"),
    Source.Name = str_remove(string = Source.Name, pattern = ".csv"),
    team = case_when(
      HomeTeam == "Arsenal" | AwayTeam == "Arsenal" ~ "Arsenal",
      HomeTeam == "Tottenham" | AwayTeam == "Tottenham" ~ "Tottenham",
      TRUE ~ "ERROR"
    ),
    location = case_when(
      team == HomeTeam ~ "Home",
      team == AwayTeam ~ "Away",
      TRUE ~ "ERROR"
    ),
    goals = case_when(
      location == "Home" ~ FTHG,
      location == "Away" ~ FTAG,
```

```
      TRUE ~ NA_real_
    ),
    matchup = paste(HomeTeam, AwayTeam, sep = " at ")
  ) %>%
  rename(season = Source.Name)

goalData$team <- as.factor(goalData$team)
goalData$location <- as.factor(goalData$location)
goalData$matchup <- as.factor(goalData$matchup)
goalData$season <- as.factor(goalData$season)
goalData=goalData[,-2:-6]

## Create Wide Data

goalWide <- goalData %>%
  mutate(
    teamXloc = paste(team, location, sep = "x")
  ) %>%
  group_by(team, location) %>%
  mutate(
    game_id = row_number()
  ) %>%
  ungroup() %>%
  dplyr::select(teamXloc, goals, game_id) %>%
  pivot_wider(
    id_cols = teamXloc,
    names_from = game_id,
    values_from = goals
  )

knitr::include_graphics("Blank diagram.png", error = FALSE)

# Descriptive statistics on Team Goals by Location ----
goalStats <- psych::describeBy(
  x = goalData$goals,
  group = goalData$location:goalData$team,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)

goalStats %>%
   tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
```

```
  knitr::kable(
    caption = "Summary Statistics for Premier League Performance Study- Team Goals by Location",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis"),
    booktabs = TRUE
  )  %>%
  kableExtra::kable_styling(
    font_size = 6.5,
    latex_options = c("scale_down", "HOLD_position")
  )

# Cleaning data for EDA graphs
## Delete the distinction of home game and away game from matchup column
no_locData <- goalData %>%
  mutate(matchup = case_when(
    matchup == "Chelsea at Arsenal" ~ "Chelsea",
    matchup == "Liverpool at Arsenal" ~ "Liverpool",
    matchup == "Man City at Arsenal" ~ "Man City",
    matchup == "Man United at Arsenal" ~ "Man United",
    matchup == "Chelsea at Tottenham" ~ "Chelsea",
    matchup == "Liverpool at Tottenham" ~ "Liverpool",
    matchup == "Man City at Tottenham" ~ "Man City",
    matchup == "Man United at Tottenham" ~ "Man United",
    matchup == "Arsenal at Chelsea" ~ "Chelsea",
    matchup == "Arsenal at Liverpool" ~ "Liverpool",
    matchup == "Arsenal at Man City" ~ "Man City",
    matchup == "Arsenal at Man United" ~ "Man United",
    matchup == "Tottenham at Chelsea" ~ "Chelsea",
    matchup == "Tottenham at Liverpool" ~ "Liverpool",
    matchup == "Tottenham at Man City" ~ "Man City",
    matchup == "Tottenham at Man United" ~ "Man United",
    TRUE ~ as.character(matchup)
  )) %>%
    mutate(season = case_when(
      season == "2009-to-2010" ~ "09-10",
      season == "2010-to-2011" ~ "10-11",
      season == "2011-to-2012" ~ "11-12",
      season == "2012-to-2013" ~ "12-13",
      season == "2013-to-2014" ~ "13-14",
      season == "2014-to-2015" ~ "14-15",
      season == "2015-to-2016" ~ "15-16",
      season == "2016-to-2017" ~ "16-17",
      season == "2017-to-2018" ~ "17-18",
      season == "2018-to-2019" ~ "18-19",
      TRUE ~ as.character(season)
    ))

sum_goals <- no_locData %>%
  group_by(team,location,matchup,season) %>%
  summarise(total_goals = sum(goals, na.rm = TRUE))
```

```r
# Histogram comparing total goals of Arsenal and Tottenham in percentage
## Calculate total goals by team and location
total_goals_by_team_location <- goalData %>%
  group_by(team, location) %>%
  summarise(total_goals = sum(goals))

## Calculate total goals for each team
total_goals_by_team <- total_goals_by_team_location %>%
  group_by(team) %>%
  summarise(team_total_goals = sum(total_goals))

## Calculate percentage of total goals for each team by location
goals_percentage <- total_goals_by_team_location %>%
  left_join(total_goals_by_team, by = "team") %>%
  mutate(percentage = total_goals / team_total_goals * 100)

## Create the histogram
ggplot(goals_percentage, aes(x = team, y = percentage, fill = location)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Team Goals by Location",
       x = "Team",
       y = "Percentage of Total Goals",
       fill = "Location")
# Descriptive statistics on Team Goals by Season ----
goalStats <- psych::describeBy(
  x = goalData$goals,
  group = c(goalData$season:goalData$team),
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)

goalStats %>%
   tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for Premier League Performance Study- Team Goals by Season",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis"),
    booktabs = TRUE
```

```
  )  %>%
  kableExtra::kable_styling(
    font_size = 6.5,
    latex_options = c("scale_down", "HOLD_position")
  )

# Total Goals by Team, Season
## Calculate total goals by team and season
total_goals_by_team_season <- no_locData %>%
  group_by(team, season) %>%
  summarise(total_goals = sum(goals))

## Create the scatterplot
ggplot(total_goals_by_team_season, aes(x = season, y = total_goals, color = team, group = team)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +  # Add a linear trend line
  theme_minimal() +
  labs(title = "Total Goals per Season",
       subtitle = "Arsenal and Tottenham",
       x = "Season",
       y = "Total Goals",
       color = "Team")
# Descriptive statistics on Goals by Matchup ----
goalStats <- psych::describeBy(
  x = goalData$goals,
  group = goalData$matchup,
  na.rm = TRUE,
  skew = TRUE,
  ranges = TRUE,
  quant = c(0.25, 0.75),
  IQR = FALSE,
  mat = TRUE,
  digits = 4
)

goalStats %>%
   tibble::remove_rownames() %>%
  tibble::column_to_rownames(
    var = "group1"
  ) %>%
  dplyr::select(
    n, min, Q0.25, median, Q0.75, max, mad, mean, sd, skew, kurtosis
  ) %>%
  knitr::kable(
    caption = "Summary Statistics for Premier League Performance Study- Goals by Matchup",
    digits = 3,
    format.args = list(big.mark = ","),
    align = rep('c', 11),
    col.names = c("n", "Min", "Q1", "Median", "Q3", "Max", "MAD", "SAM", "SASD",
                  "Sample Skew", "Sample Ex. Kurtosis"),
    booktabs = TRUE
  )  %>%
  kableExtra::kable_styling(
```

```
    font_size = 6.5,
    latex_options = c("scale_down", "HOLD_position")
  )

ggplot(sum_goals, aes(x = team, y = total_goals, fill = location)) +
  geom_bar(stat = "identity", position = "stack", width = 0.7) +
  labs(title = "Total Goals by Team, Location, and matchup",
       x = "matchup",
       y = "Total Goals",
       fill = "Location") +
  facet_grid(. ~ matchup, scales = "free_x", space = "free_x") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
boxplotData <- goalData %>%
  filter(team %in% c('Arsenal', 'Tottenham')) %>%
  mutate(team_location = interaction(team, location, sep = ' '))

# Goals Distribution by Team and Location (box plot)
ggplot(boxplotData, aes(x = team_location, y = goals)) +
  geom_boxplot() +
  labs(title = "Goals by Team and Location",
       x = "Team and Location",
       y = "Goals") +
  theme_minimal()

# Fit the models ----
EPLModel <- aov(
  formula = goals ~ team*location + season + season:team:location + Error(matchup %in% team:location),
  data = goalData
)

EPLAssumptions <- blme::blmer(
  data = goalData,
  formula = goals ~ team*location + season + season:team:location + (1|matchup)
)

## QQPlot for Gaussian Residuals
car::qqPlot(
  x = residuals(EPLAssumptions),
  distribution = "norm",
  envelope = 0.90,
  id = FALSE,
  pch = 20,
  ylab = "Residuals (Goals)"
)

## Tukey-Anscombe Plot for Homoscedasticity Assumption ----

ggplot(
  data = data.frame(
    residuals = residuals(EPLAssumptions),
    fitted = fitted.values(EPLAssumptions)
  ),
```

```
  mapping = aes(x = fitted, y = residuals)
) +
  geom_point(size = 2) +
  geom_hline(
    yintercept = 0,
    linetype = "dashed",
    color = "grey50"
  ) +
  geom_smooth(
    formula = y ~ x,
    method = stats::loess,
    method.args = list(degree = 1),
    se = FALSE,
    linewidth = 0.5
  ) +
  theme_bw() +
  xlab("Fitted values") +
  ylab("Residuals (Goals)")
## Sphericity Plots for Sphericity Assumption

sphericityPlot(
  dataWide = goalWide[,c(1:11)],
  subjectID = "teamXloc"
)

sphericityPlot(
  dataWide = goalWide[,c(1,12:21)],
  subjectID = "teamXloc"
)

sphericityPlot(
  dataWide = goalWide[,c(1,22:31)],
  subjectID = "teamXloc"
)

sphericityPlot(
  dataWide = goalWide[,c(1,32:41)],
  subjectID = "teamXloc"
)
# Modern ANOVA Table ----
EPLTemp <- summary(EPLModel)
EPLOmni <- rbind(
  EPLTemp$`Error: matchup:team:location`[[1]],
  EPLTemp$`Error: Within`[[1]]
)
row.names(EPLOmni) <- c("team", "location", "team:location","matchup", "season","team:location:season",
EPLOmni["matchup", "F value"] <- EPLOmni["matchup", "Mean Sq"] /
  EPLOmni["matchup:season", "Mean Sq"]
EPLOmni["matchup", "Pr(>F)"] <- pf(
  q = EPLOmni["matchup", "F value"],
  df1 = EPLOmni["matchup", "Df"],
  df2 = EPLOmni["matchup:season", "Df"],
  lower.tail = FALSE
```

```
)
EPLOmni %>%
  tibble::rownames_to_column(
    var = "Source"
  ) %>%
  dplyr::mutate(
    `Pr(>F)` = ifelse(
      test = is.na(`Pr(>F)`),
      yes = NA,
      no = pvalRound(`Pr(>F)`)
    )
  ) %>%
  knitr::kable(
    digits = 4,
    col.names = c("Source", "df", "SS", "MS", "F", "p-value"),
    caption = "ANOVA Table for Premier League Study",
    align = c('l',rep('c',5)),
    booktab = TRUE,
    format.args = list(big.mark = ",")
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12,
    latex_options = c("HOLD_position")
  )

temp1 <- as.data.frame(fixef(EPLAssumptions, add.dropped = TRUE)) %>%
  rownames_to_column("term")
temp1$term[1] <- "Grand Mean"
temp1$term[2] <- "Arsenal"
temp1$term[3] <- "Home"
temp1$term[13] <- "Arsenal:Home"
temp1= temp1[-(14:49),]
temp1= temp1[-(4:12),]
colnames(temp1) <- c('Term','Estimate')
rownames(temp1) <- 1:nrow(temp1)

data.frame(temp1) %>%
  knitr::kable(
    digits = 2,
    caption = "Point Estimates from Premier League Performance Study- Team and Location",
    booktabs = TRUE,
    align = "c"
  ) %>%
  kableExtra::kable_styling(
    font_size = 12,
    latex_options = c("HOLD_position")
  )
point2= temp1 <- as.data.frame(fixef(EPLAssumptions, add.dropped = TRUE)) %>%
  rownames_to_column("term")

point2= point2[-(14:49),]
point2= point2[-(2:3),]
```

```r
point2= point2[-13,]
rownames(point2) <- 1:nrow(point2)
point2$term[1] <- "Grand Mean"
point2$term[2] <- "2009-2010"
point2$term[3] <- "2010-2011"
point2$term[4] <- "2011-2012"
point2$term[5] <- "2012-2013"
point2$term[6] <- "2013-2014"
point2$term[7] <- "2014-2015"
point2$term[8] <- "2015-2016"
point2$term[9] <- "2016-2017"
point2$term[10] <- "2017-2018"
point2$term[11] <- "2018-2019"
colnames(point2) <- c('Term','Estimate')

data.frame(point2) %>%
  knitr::kable(
    digits = 2,
    caption = "Point Estimates from Premier League Performance Study- Season",
    booktabs = TRUE,
    align = "c"
  ) %>%
  kableExtra::kable_styling(
    font_size = 10,
    latex_options = c("HOLD_position")
  )

teamPH2 <- emmeans::emmeans(
  object = EPLAssumptions,
  specs = pairwise ~ team:location,
  adjust = "BH",
  level = 0.95
)

as.data.frame(teamPH2$contrasts) %>%
  knitr::kable(
    digits = 4,
    col.names = c("Comparison", "Estimate", "SE", "DF", "t ratio", "p-value"),
    align = c("l", rep("c", 5)),
    caption = "Pairwise Comparisons of Teams by Location",
    booktabs = TRUE
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 12
    #latex_options = c("HOLD_position")
  )
```