Premier League Data Analysis (2024-25)

Alven Huang and Ben Rishel

2025-05-07

1 Introduction

The Premier League is a professional soccer league in England and is the highest level in the English system. Each season the 20 teams in the Premier League fight to stay out of the relegation zone and compete for the Premier League title. Soccer, like any other sport, can benefit from data analytics. In this sport, data can be used to find the most optimal strategies, whether it is playing the possession game or playing fast. While there are some cutting-edge statistics in soccer, like expected goals and expected points (more details to come), some of the timeless statistics, like possession and win rates, are useful as well. In this project we answered four important research questions regarding the 2025 Premier League Season (35/38 matchdays completed at the creation of the analysis). Below are the research questions and the attributes that they each focused on. If you don't understand what the question means or what the attributes are, do not worry; they will be fully explained in each section. The four research questions can each be classified as confirmatory data analysis questions. Given the vast dataset and possibilities, we felt it necessary to specify four questions of research in which we can find a definitive conclusion at the end. For complete transparency and to adhere to the Open Science Principles: Alven worked on questions 1 and 3, and Ben worked on questions 2 and 4 initially. However, as the project progressed, we each contributed to each others analysis.

1.1 Research Questions

1. Does Having a Higher Possession Percentage Translate to Wins? Attributes of Focus: Possession Percentage and Wins
2. Is Home-Field Advantage Real? Attributes of Focus: Points Gained at Home, xPoints Gained at Home
3. Which Teams Have the Best and Worst Shot Conversion Rates? Attribute of Focus: Shot Conversion Rate

4. How is the Premier League Different from the Other "Big 5" Leagues Financially? Attribute of Focus: Team Valuations

2 Data Provenance

2.1 Primary Dataset

Source: The primary dataset comes from fbref.com or Football Reference. However, Football Reference primarily gets their data from Opta.

Who Collected the Data: Opta uses a combination of human analysts and AI systems to track and collect data across world soccer.

Why Did They Collect the Data: Opta collects the data because they are paid to do so by teams and broadcast companies who use the data for decision-making and broadcasts, respectively. Football reference then displays this data through their partnership with Opta in order to generate website traffic and ad revenue.

Cases: Depending on the table in Football Reference, a case could be a player, a team, a specific match, a specific competition, or even an award. Needless to say, Football Reference contains an extraordinary amount of data. We specifically used data in which teams or specific matches were cases.

2.2 Secondary Dataset

Source: The secondary dataset comes from transfermarkt.com or Transfermarkt.

Who Collected the Data: Transfermarkt relies on user contributed data that is reviewed and moderated. In addition, they have employees who update data based upon publicly available sources like official club/league/federation websites and social media.

Why Did They Collect the Data: Transfermarkt collects all of this data to build a popular database for fans in order to gain website traffic for ad revenue and subscriptions.

Cases: Just like Football Reference, Transfermarkt contains many tables with different cases. For the purpose of this analysis, we only used data in which a specific player is a case.

2.3 Note on How We Used Data

For Football Reference, we used the worldfootballR package to scrape the data. To install the most updated version of worldfootballR package, refer to the beginning of the code appendix.

For Transfermarkt, we initially used the worldfootballR package as well. However, one day before the project was due, this part of the project was essentially destroyed as Transfermarkt changed their website and the worldfootballR package no longer worked. Last minute, we were able to just copy and paste the limited data that we needed from Transfermarkt into a google sheet. We then read this in using the googlesheets4 package.

2.4 Coding Style

The coding style that our code followed was the Tidyverse Style Guide. Because of the use of piping and the packages such as ggplot and the tidyverse family, we determined that this was the Tidyverse Style Guide.

3 FAIR Principles

Findable - Data is easily findable on fbref.com and transfermarkt.com

Accessible - The data is accessible as it is not behind a paywall and available through scraping. Easy CSV downloads are available on fbref.com

Interoperable - The data follows broadly applicable language for soccer data. Data is consistently formatted.

Reusable - The data includes many relevant, accurate attributes. In addition, they have terms of use listed.

4 CARE Principles

Collective-Benefit - This data helps benefit those who are sports fans, data analysts, etc.

Authority to Control - The data we collected is available for public use. We did not access any private information while collecting the data.

Responsibility - We did not include any offensive, sensitive, or any irregular information in our data.

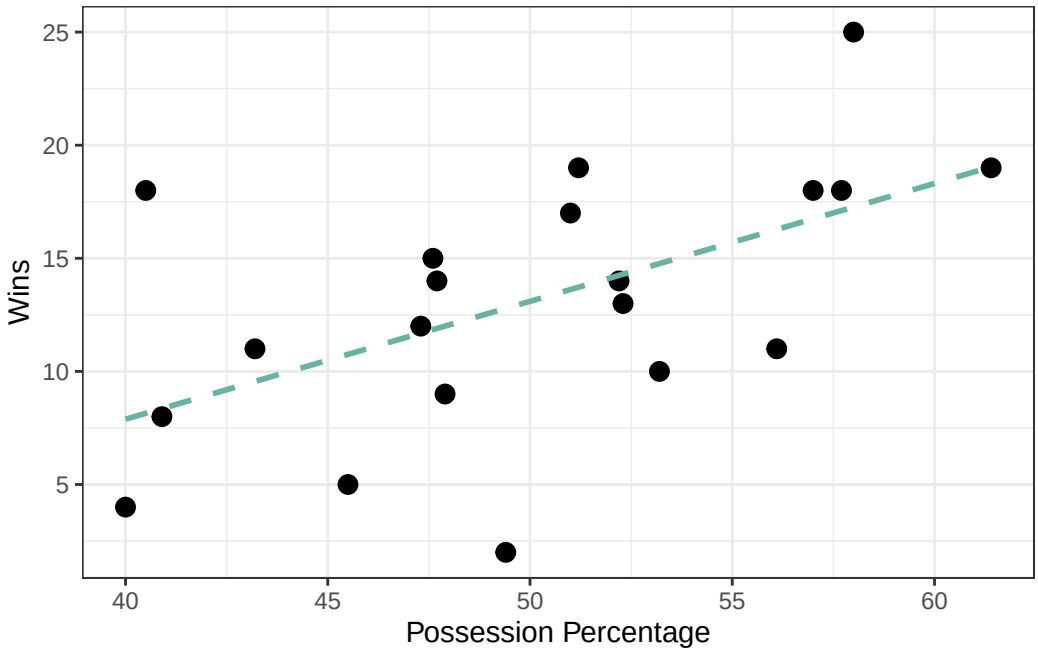
Ethics - We were fair and focused on all the principles. We remained unbiased.

5 Does Having a Higher Possession Percentage Translate to Wins?

For this research question, we want to visualize the average possession percentage for all twenty Premier League teams in the 2024-25 season and compare it to how many wins each team has. In soccer, possession is essentially how long a team controls the ball for the entire game. So, the max possession percentage is 100, which is quite impossible. This will answer our question on if having higher possession means more wins.

We did this by first installing and calling the required packages. Then we took the scraped data and focused on the possession stat type. We then created a new function to obtain all the teams and their wins this season. We cleaned the data (FBref, 2025a) to focus on only wins, possession, and squads. Then, we combined the data and plotted it. On the plot, every team is a black dot. We figured that since this analysis is more focused on the location of the points, we do not need to identify each team with a specific color. We added a trend line to see who is above and below, which will also help us answer our question. We used a scatter plot to see each team's win total by their average possession percentage.

Figure 1: Wins vs. Possession Percentage



As we can see in the Figure 1, the trend line goes up and to the right, which represents a positive correlation between possession percentage and wins. We can also see that almost every team with an average possession of at least 57% has at least 17 wins, which is also above the trend line. The teams with a lower percentage can be seen towards the bottom of plot, where the wins are around 14 and less. However, there is one outlier that has a percentage of

just over 40, and they have about 17 wins. Other than the outlier, we can see that teams with a higher average possession rate tend to win more games. This is what we expected the result to be.

6 Is Home-Field Advantage Real?

Next, we will analyze whether or not home-field advantage is real in this Premier League season. Home-field advantage has long been a part of sports discourse. The theory is that teams perform better when they play in their home stadium/arena due to a multitude of factors. These factors include the support of their fans, the lack of travel necessary, and the comfort of being at home. To determine whether this is real or not in the Premier League, we will first analyze a stat we created called "Points Gained at Home Per Match". The point system in soccer is used to track how well teams have performed in a season and to organize the league table. If you have the most points at the end of the season, you win the league. If you win a game, you get three points. If you tie, you get one point. And, if you lose, you get zero points. "Points Gained at Home Per Match" is the difference between average points at home and average points at away matches. For example, if a team has a value 0.5 "Points Gained at Home Per Match", that means that they average 0.5 more points in home games as compared to away games. Let's take a look at a density chart of all 20 Premier League teams. For this section we will first use team stats data for the 2024-2025 season (FBref, 2025a) and then individual match data from the 2024-2025 season for a significance test (FBref, 2025b).

As you can see in Figure 2, the density is centered above 0, which shows that the average "Points Gained at Home Per Match" is positive. This shows some early evidence that home-field advantage may be real. However, we can see that there is some significant density to the left of 0, so not every team has the same benefit. There is also the argument that points is not necessarily the best metric to determine how well teams have performed. In soccer, scoring goals often includes luck. When scores are as low as 0-0, 1-0, or 1-1, a single lucky bounce can have a tremendous impact on the game. To help control for luck and to better measure the performance of teams in games, the expected goals (xG) stat has become quite popular. Expected goals measures how many goals a team would score historically given where the shots they took were and how many shots they took. As a result, a team that won 1-0 on a lucky shot may have just 0.05 xG, which is a better representation of how they performed. Using xG, we can determine the expected points (xPoints) of a team based on the results of their games given the xG they and their opposition had. Let's now take a look at a top-bottom density plot with "Points Gained at Home Per Match" and "xPoints Gained at Home per Match" on the bottom.

In Figure 3 you can see that "xPoints Gained at Home per Match" is centered at a higher value than "Points Gained at Home Per Match". However, there is also a slight peak of density below zero as well. From this chart, we can conclude that teams generally do gain points and xPoints at home. To see whether or not this is a statistically significant difference, let's run

Figure 2: Premier League Points Gained at Home

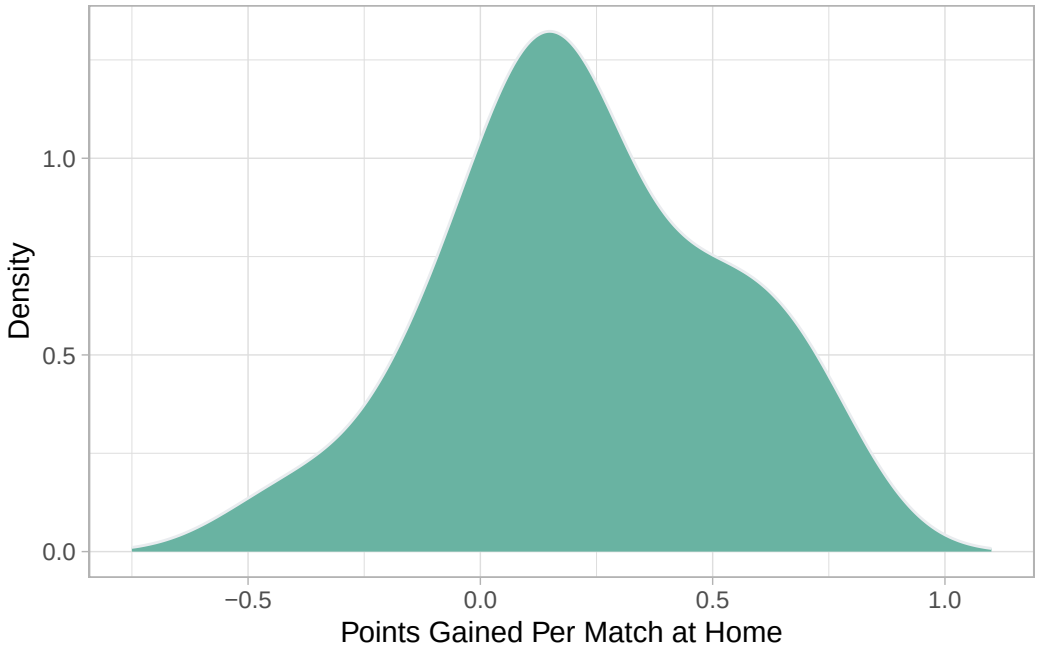
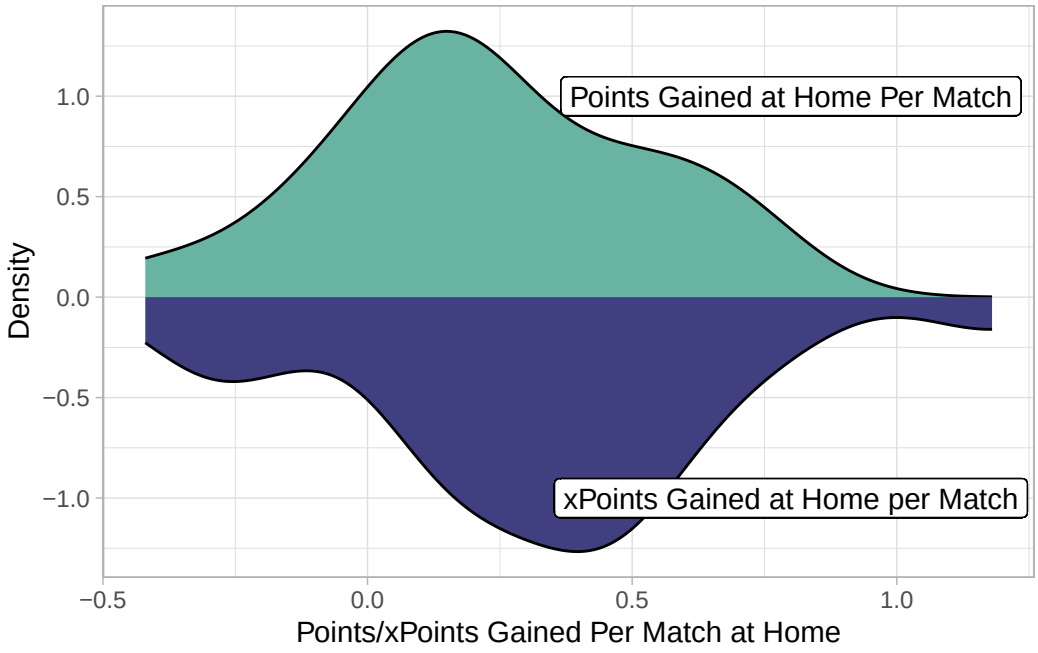


Figure 3: Premier League Points/Expected Points Gained at Home



a one-sided t-test with alpha=0.05. Our null hypothesis is that there are no points gained at home in the 2025 Premier League season.

Table 1: One-Sided T-Test Results 2025

Average Points Gained	T-Statistic	P-Value
0.223	1.61	0.054

As you can see from the Table 1, the p-value is just barely above the alpha of 0.05 at 0.054. Therefore, we cannot reject the null and do not have statistically significant evidence of home-field advantage for the Premier League 2025 season. Out of curiosity, let's check a bigger sample from 2022-2025 data (FBref, 2025c). We will run another one-sided t-test with alpha=0.05, and the null hypothesis is that there are no points gained at home in the 2022-2025 Premier League seasons.

Table 2: One-Sided T-Test Results 2022-2025

Average Points Gained	T-Statistic	P-Value
0.377	5.582	0

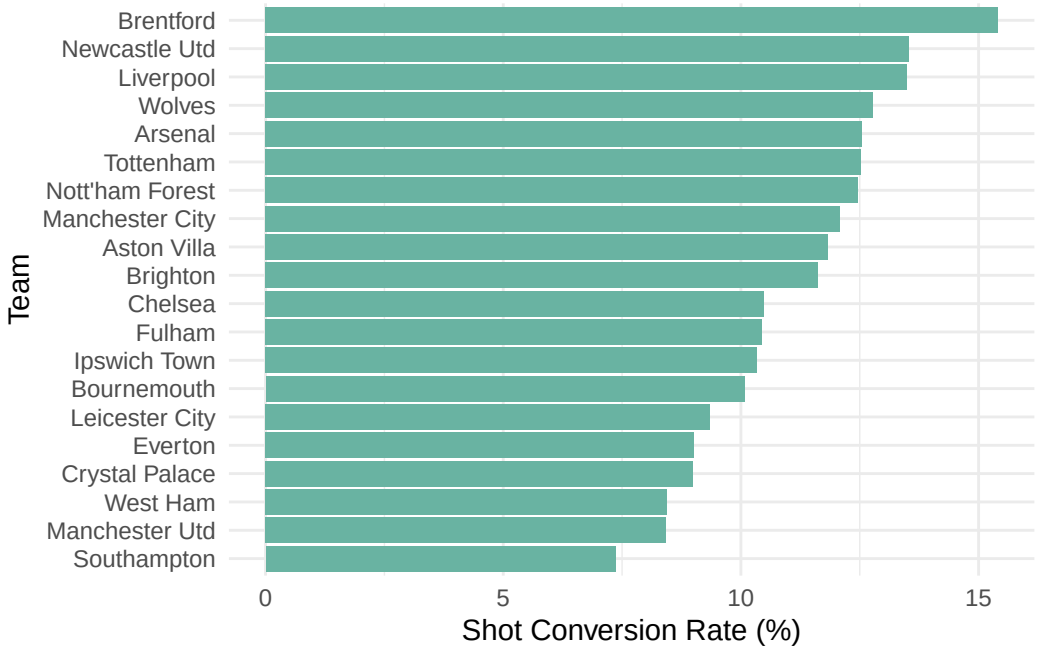
From Table 2, we can see that the p-value is absolutely below the alpha of 0.05; we can safely reject the null hypothesis in favor of the alternative which is that there are positive points gained at home per match in the 2022-2025 Premier League season. Therefore, home-field advantage existed during this time period.

7 Which Teams Have the Best and Worst Shot Conversion Rates?

In this analysis, we took a look at which teams have the best shot conversion rate and which teams have the worst. Shot conversion rate is the rate at which a player scores when they shoot. Every time they shoot, the rate changes based on whether the ball goes in or not. We believe this is a great analytic to look at to see which teams are attack heavy and which teams are most consistent on the attack.

To do this, we used our scraped data (FBref, 2025a) and called it where we focused on the shooting statistics. We then cleaned the data and added the shot conversion rate formula to the function. The formula is easy: the amount of goals scored divided by the amount of shots taken. Then, you divide that number by 100 to find the percentage. Then we plotted this data using a horizontal bar chart, where the teams are on the left and the percent is on the bottom. This way, We can easily observe the data and answer our question.

Figure 4: Premier League Teams Shot Conversion Rate (2024-25)



From this Figure 4, we can discover that Brentford has the highest shot conversion rate with 15 percent. We can also observe that Southampton has the worst shot conversion rate with a rate about 6.5 percent. This data does not necessarily tell us which teams are the worst and best teams but rather which teams tend to be the most consistent when having an opportunity to score. As we are soccer fans ourselves, we recognize each team's skill level, and we can confidently say that Brentford is not the team we expected to see at the top. However, we predicted Southampton to be on the bottom.

8 How is the Premier League Different from the Other "Big 5" Leagues Financially?

In soccer, there are five leagues known as the "Big 5" leagues that are considered to be superior to the rest of the leagues in the world. These five leagues are the Bundesliga in Germany, La Liga in Spain, Ligue 1 in France, Serie A in Italy, and the Premier League in England. For this part of the analysis, we are going to take a look at how the Premier League compares to these leagues financially. The data we will be using to measure the value of teams and leagues is the teams estimated valuation via Transfermarkt. Out of curiosity, let's first take a look at the teams that have the highest and lowest valuations in the data (Transfermarkt, 2025).

Table 3: Top Ten Team Valuations (Big 5 Leagues)

Squad Name	Big 5 League	Valuation
Manchester City	Premier League	$1,493,400,000
Real Madrid	LaLiga	$1,447,800,000
Arsenal FC	Premier League	$1,288,200,000
FC Barcelona	LaLiga	$1,162,800,000
Liverpool FC	Premier League	$1,132,590,000
Paris Saint-Germain	Ligue 1	$1,052,790,000
Chelsea FC	Premier League	$1,051,080,000
Bayern Munich	Bundesliga	$979,260,000
Tottenham Hotspur	Premier League	$953,154,000
Manchester United	Premier League	$791,445,000

As you can see from Table 3 , the top 10 is dominated by the "Big 6" Premier League clubs: Manchester City, Arsenal, Liverpool, Chelsea, Tottenham, and Manchester United. The rest of the top 10 is rounded out by the dominant clubs of La Liga, Ligue 1, and the Bundesliga. Real Madrid, Barcelona, PSG, and Bayern Munich are expectedly in this list, as they are hugely successful clubs in some of the biggest cities in the world. It is important to note that there are no Italian clubs (Serie A) in this list, which is a bit surprising. While this gives an idea that the Premier League may have more valuable teams in general, it is possible that the Premier League is top heavy. It is also possible that the other leagues are top heavy and their representation in the top 10 are outliers in their respective leagues. Let's check the bottom ten teams to see if we can learn anything else.

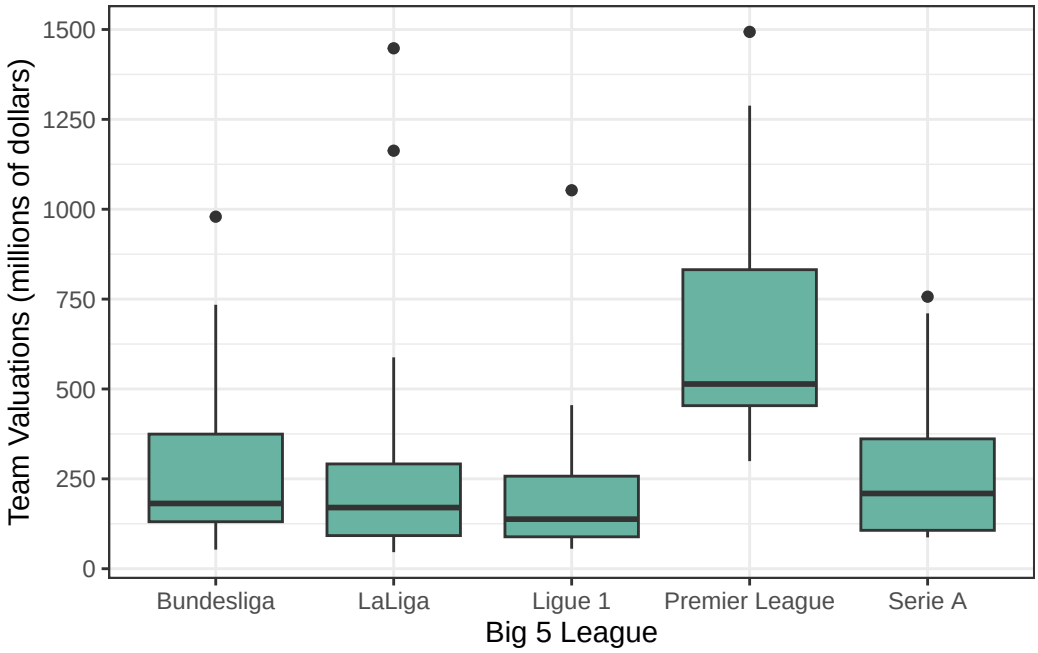
Table 4: Bottom Ten Team Valuations (Big 5 Leagues)

Squad Name	Big 5 League	Valuation
CD Leganés	LaLiga	$46,056,000
Real Valladolid CF	LaLiga	$49,362,000
Holstein Kiel	Bundesliga	$53,158,200
Angers SCO	Ligue 1	$55,575,000
Montpellier HSC	Ligue 1	$63,555,000
AS Saint-Étienne	Ligue 1	$65,436,000

VfL Bochum	Bundesliga	$67,773,000
FC St. Pauli	Bundesliga	$69,574,200
Le Havre AC	Ligue 1	$74,271,000
1.FC Heidenheim 1846	Bundesliga	$76,129,200

The first most notable observation from Table 4 is just how less valuable these teams are as compared to the top ten. At $46 million, CD Leganés is about just 3% of the total valuation of Manchester City. This shows just how financially unbalanced the sport of soccer is in Europe where there is no salary cap like most American sports. These teams at the bottom have virtually no chance of ever competing with the top teams. Again, there is an absence of Serie A teams from this list which is quite interesting. Serie A has no teams in the bottom ten or top ten which means that Serie A must have more financial balance between the teams within the league. The Premier League has no representation here, which shows that maybe the Premier League really is more valuable. Ligue 1 and the Bundesliga are most represented here, which show that it is possible that Bayern Munich and PSG may have been outliers being so highly valued. Let's take a look at how the distributions as a whole compare across the leagues with a box plot.

Figure 5: Distribution of Team Valuations by Big 5 League



As you can see in Figure 5, the Premier League is indeed much more valuable compared to the other four "Big 5" leagues. In fact, the minimum valued Premier League team would be above

the average value in any other of the leagues. In terms of outliers, it is interesting that Serie A does in fact have a positive outlier but that it did not crack the top ten most valuable teams. It is also interesting that each league has at least one outlier that shows that each of the five leagues have a dominant teams, with La Liga being different with two dominant teams. While it might sound like a bit of a leap to say that just because a team is more valuable that they are a dominant team, being a successful team is the main way to accrue value. While branding and financial assets like a stadium are important, at the end of the day having greater value than others in your league comes from being more successful than other teams. In conclusion, we can see that the Premier League is the most valuable of the "Big 5" leagues.

9 Conclusion

This project allowed us to dive deep into the statistics for the 2024-25 season of the Premier League. We were able to smoothly answer all of our research questions and discover new information about the sport. While we were already familiar with professional soccer, we have unlocked new trends in the realm of soccer analytics. Not only are these visualizations useful for soccer fans, but they can also be helpful for those who do not watch soccer or for those who are trying to get into the sport.

Our project consisted of four main questions about the Premier League:

1. Does Having a Higher Possession Percentage Translate to Wins?
2. Is Home-Field Advantage Real?
3. Which Teams Have the Best and Worst Shot Conversion Rates?
4. How is the Premier League Different from the Other "Big 5" Leagues Financially?

Let's quickly recap what we discovered while analyzing each question:

1. From what the visual showed, we saw that most teams with a higher average possession percentage had more wins than others.
2. While there are some signs that home-field advantage may be real for the 2025 season, in the end it was not statistically significant. However, when looking at data from 2022-2025, there was a statistically significant home-field advantage.
3. Brentford had the best shot conversion rate and Southampton had the worst.
4. The Premier League is cemented as the most valuable of the "Big 5" leagues by a wide margin.

10 References

FBref. (2025a). *Premier League Team Stats: 2024-2025 season*. Retrieved May 7, 2025, from https://fbref.com/

FBref. (2025b). *Premier League Match Results: 2024-2025 seasons*. Retrieved May 7, 2025, from https://fbref.com/

FBref. (2025c). *Premier League Match Results: 2022-2025 seasons*. Retrieved May 7, 2025, from https://fbref.com/

Transfermarkt. (2025). *Squad Valuations: 2025*. Retrieved May 6, 2025, from https://www.transfermarkt.com/

11 Code Appendix

```
#### Chunk1 - How to Install Updated worldfootballR

#devtools::install_github("JaseZiv/worldfootballR")
#installs most updated version of worldfootballR package
#This package allows easy scraping of FBref and Transfermarkt
#https://jaseziv.github.io/worldfootballR/articles/extract-fbref-data.html

#### Chunk2 - Does Having a Higher Possession Percentage Translate to Wins?

### Plan: Goal, Needs, Steps

## Goal: Use scraped data and clean data to create a data visualization that
# displays and answers our research questions,
# "Does having a higher average possessionpercentage translate to wins?"

## Needs:
# Verbs: filter, select, inner_join, mutate
# Nouns: worldFootballR package, all necessary libraries, team stats

## Steps:
# 1. Install the packages and call them
# 2. Call needed libraries to scrape data
# 3. Call all functions from scraping data
# 4. Create new R file to answer research question
# 5. Call all needed libraries
```

```r
# 6. Create a function to obtain the possession stat for male first
# tier teams in England only, most recent year.
# 7. Create a function to obtain the table of the male first tier league in
# England, in the most recent year.
# 8. Clean the possession data, with necessary verbs.
# 9. Clean the league table data
# 10. Join the two cleans data and put into one
# 11. Create visualization for the conjoined data.


### Calls all needed libraries
library(worldfootballR)
library(dplyr)
library(tidyr)
library(tidyverse)
library(ggplot2)

### Focuses on possession for all the teams in English Mens' First Tier
pl_possession <- fb_season_team_stats(
  country = "ENG",
  gender = "M",
  season_end_year = 2025,
  tier = "1st",
  stat_type = "possession"
)

### Focuses on the teams on the table in most recent year
pl_table <- fb_season_team_stats(
  country = "ENG",
  gender = "M",
  season_end_year = 2025,
  tier = "1st",
  stat_type = "league_table"
)

### Cleans the possession data
pl_possession_clean <- pl_possession %>%
  filter(!str_starts(Squad, "vs ")) %>% # Gets rid of average opposing possession
  select(Squad, Poss) %>% # Focuses on team and the possession percentage
  mutate(Poss = as.numeric(str_remove(Poss, "%")))
```

```r
### Cleans the league table data
pl_table_clean <- pl_table %>%
  filter(!str_starts(Squad, "vs ")) %>% # Filters out average opposing once again
  select(Squad, W) # Focuses on the team and the wins


### Combines the two clean datas
pl_poss_wins <- pl_possession_clean %>%
  inner_join(pl_table_clean, by = "Squad")


### Creates scatter plot for the combined data
ggplot(
  data = pl_poss_wins,
  mapping = aes(
    x = Poss,
    y = W
  )
) +
  geom_point(size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "#69b3a2", linetype = "dashed") +
  labs(
    x = "Possession Percentage",
    y = "Wins"
  ) +
  theme_bw() +
  theme(
    legend.position = "bottom"
  )

#### Chunk3 - Is Home-Field Advantage Real?

### Plan: Goal, Needs, Steps

## Goal: Research whether or not home-field advantage is real in world football

## Needs:
# Nouns: Data, worldfootballR package, dplyr, tidyr, ggplot2, knitr
# Verbs: worldfootballR functions, data wrangling verbs, ggplot functions

## Steps:
# 1. Import Packages
```

```r
# 2. Load in Data
# 3. Wrangle data to just contain home/away stats
# 4. Visualize the difference in home/away stats for all teams
# 5. Run a t-test to see if the difference is statistically significant

### Import Packages
library(worldfootballR)
library(dplyr)
library(tidyr)
library(ggplot2)
library(knitr)


### Load in Data
pl_team_stats_raw <- fb_season_team_stats(
  country = "ENG",
  gender = "M",
  season_end_year = 2025,
  tier = "1st",
  stat_type = "league_table_home_away"
)

### Data Wrangling

## Selecting attributes for analysis, and making them numeric values
pl_pts_per_match <- pl_team_stats_raw %>%
  select(
    Squad,
    Pts_per_MP_Home,
    Pts_per_MP_Away,
    xGD_per_90_Home,
    xGD_per_90_Away
  ) %>%
  mutate( #Force stats to be numeric
    across(c(Pts_per_MP_Home, Pts_per_MP_Away, xGD_per_90_Home, xGD_per_90_Away), as.numeric)
  )

## Creating Summary Table with Points Gained Metrics
pl_more_pts_home <- pl_pts_per_match %>%
  summarize(
    Squad = Squad,
    Pts_gained_at_home_per_match = Pts_per_MP_Home - Pts_per_MP_Away,
```

```
    xPts_gained_at_home_per_match = xGD_per_90_Home - xGD_per_90_Away
  )

### Data Visualizations

## Creating Density Chart to Analyze How Much Home Field Advantage Matters
ggplot(
  data = pl_more_pts_home,
  aes(
    x = Pts_gained_at_home_per_match
    )
  ) +
  geom_density( #density plot
    fill = "#69b3a2",
    color = "#e9ecef"
  ) +
  labs(
    x = "Points Gained Per Match at Home",
    y = "Density"
  ) +
  scale_x_continuous(
    limits = c(-.75, 1.1) #Making sure all data fits by setting x bounds
    ) +
  theme_light()

#### Chunk3.1 - Is Home-Field Advantage Real?

## Top-Bottom Density Plot that Compares Points vs. Expected Points
ggplot(
  data = pl_more_pts_home,
  aes(x = Pts_gained_at_home_per_match)
) +
  # Top
  geom_density(
    aes(y = ..density..),
    fill = "#69b3a2"
  ) +
  geom_label( #adds label to top chart
    aes(x = .8, y = 1, label = "Points Gained at Home Per Match")
  ) +
  # Bottom (flipped)
  geom_density(
```

```
    aes(x = xPts_gained_at_home_per_match, y = -..density..),
    fill = "#404080"
  ) +
  geom_label( #adds label to bottom chart
    aes(x = .8, y = -1, label = "xPoints Gained at Home per Match")
  ) +
  labs(
    x = "Points/xPoints Gained Per Match at Home",
    y = "Density"
  ) +
  theme_light()

#### Chunk3.2 - Is Home-Field Advantage Real?

### Testing League-Wide Home Field Advantage Statistical Significance

pl_2025_match_raw <- fb_match_results(
  country = "ENG",
  gender = "M",
  season_end_year = 2025,
  tier = "1st"
  )

## Selecting attributes for analysis, and making them numeric values
pl_2025_match <- pl_2025_match_raw %>%
  select(
    HomeGoals,
    AwayGoals
  ) %>%
  mutate( #force stats to be numeric
    across(c(HomeGoals, AwayGoals), as.numeric)
  ) %>%
  drop_na(HomeGoals, AwayGoals) #dropping nullified matches

## Creating Summary Table with Match as Case, attributes homePts, awayPts
pl_homeAwayPts <- pl_2025_match %>%
  mutate(
    homePts = case_when( #when home scores more - win, if same - tie
      HomeGoals  > AwayGoals ~ 3,
      HomeGoals  == AwayGoals ~ 1,
      HomeGoals  < AwayGoals ~ 0
    ),
```

```
    awayPts = case_when(
      AwayGoals  > HomeGoals ~ 3,
      AwayGoals  == HomeGoals ~ 1,
      AwayGoals < HomeGoals ~ 0
    )
  )

## Creating table to be used for t-test
pl_homePtsGained <- pl_homeAwayPts %>%
  mutate(
    homePtsGained = homePts - awayPts
  )

## Running t-test
test <- t.test(
        pl_homePtsGained$homePtsGained,
        mu = 0,          #Null Hypothesis: no home advantage
        alternative = "greater" ) # One sided t-test

## Building Dataframe with info from t-test
test_info <- data.frame(
  `Average Points Gained` = mean(pl_homePtsGained$homePtsGained),
  `T-Statistic` = unname(test$statistic),
  `P-Value`     = test$p.value,
  check.names   = FALSE  # keep our nice column names
)

## Create Nice Table with Kable - "One-Sample T-Test: Home vs. Away Point Difference"
test_info %>% kable(
      digits  = 3,
      align = c(rep("c", 3))
      ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 16,
    full_width = FALSE,
    position = "center"
  )
#### Chunk3.4 - Is Home-Field Advantage Real?

### Testing League-Wide Home Field Advantage Statistical Significance 2022-2025
```

```r
pl_2022_2025_match_raw <- fb_match_results(country = "ENG", gender = "M", season_end_year = c

## Selecting attributes for analysis, and making them numeric values
pl_2022_2025_match <- pl_2022_2025_match_raw %>%
  select(
    HomeGoals,
    AwayGoals
  ) %>%
  mutate( #Force stats to be numeric
    across(c(HomeGoals, AwayGoals), as.numeric)
  ) %>%
  drop_na(HomeGoals, AwayGoals) #dropping nullified matches

## Creating Summary Table with Match as Case, attributes homePts, awayPts
pl_homeAwayPts <- pl_2022_2025_match %>%
  mutate(
    homePts = case_when( #When home scores more - win, if same - tie
      HomeGoals  > AwayGoals ~ 3,
      HomeGoals  == AwayGoals ~ 1,
      HomeGoals  < AwayGoals ~ 0
    ),
    awayPts = case_when(
      AwayGoals  > HomeGoals ~ 3,
      AwayGoals  == HomeGoals ~ 1,
      AwayGoals < HomeGoals ~ 0
    )
  )

## Creating table to be used for t-test
pl_homePtsGained <- pl_homeAwayPts %>%
  mutate(
    homePtsGained = homePts - awayPts
  )

## Running t-test
test <- t.test(
  pl_homePtsGained$homePtsGained,
  mu = 0,          #Null Hypothesis: no home advantage
  alternative = "greater" ) # One sided t-test

## Building Dataframe with info from t-test
test_info <- data.frame(
```

```
  `Average Points Gained` = mean(pl_homePtsGained$homePtsGained),
  `T-Statistic` = unname(test$statistic),
  `P-Value`     = test$p.value,
  check.names   = FALSE  # keep our nice column names
)

## Create Nice Table with Kable - "One-Sample T-Test: Home vs. Away Point Difference"
test_info %>% kable(
  digits  = 3,
  align = c(rep("c", 5))
) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 16,
    full_width = FALSE,
    position = "center"
  )

#### Chunk4 - Which Teams Have the Best and Worst Shot Conversion Rates?

# Goal: Use scraped and clean data to find each Premier League teams'
# shot conversion rate in the 2024-25 season. This will help answer our
# question of which team has the best shot conversion rate and which has the worst.

## Needs:
# Verbs: filter, select, mutate, arrange
# Nouns: worldFootballR package, all necessary libraries, team stats

# Steps:
# 1. Install the packages and call them
# 2. Call needed libraries to scrape data
# 3. Call all functions from scraping data
# 4. Create new R file to answer research question
# 5. Call all needed libraries
# 6. Create a function to obtain the shooting stats for each premier league team
# 7. Create a function that cleans the shooting stats and uses the shot conversion rate
# 8. Clean the possession data, with necessary verbs.
# 9. Create visualization for the conjoined data.


### Calls all needed libraries
library(ggplot2)
```

```r
library(worldfootballR)
library(dplyr)


### Focuses on shooting for all the teams in English Mens' First Tier
pl_shooting <- fb_season_team_stats(
  country = "ENG",
  gender = "M",
  season_end_year = 2025,
  tier = "1st",
  stat_type = "shooting"
)


### Cleans shooting data and uses the shot conversion formula
pl_conversion <- pl_shooting %>%
  filter(!str_starts(Squad, "vs ")) %>%
  select(Squad, Goals = Gls_Standard, Shots = Sh_Standard) %>%
  mutate(
    ConversionRate = (Goals / Shots) * 100
  ) %>%
  arrange(desc(ConversionRate))

### Creates a horizontal bar chart to visualize the shot conversion rate for each team
ggplot(
  data = pl_conversion,
  mapping = aes(
    x = reorder(Squad, ConversionRate),
    y = ConversionRate
  )
) +
  geom_col(fill = "#69b3a2") +
  coord_flip() +
  labs(
    x = "Team",
    y = "Shot Conversion Rate (%)"
  ) +
  theme_minimal()

#### Chunk5 - How is the Premier League Different from the Other "Big 5" Leagues Financially

### Plan: Goal, Needs, Steps
```

```r
## Goal: Use Transfermarket data to find differences of big 5 leagues

## Needs:
# Nouns: Data, worldfootballR package, dplyr, tidyr, ggplot2, knitr
# Verbs: worldfootballR functions, data wrangling verbs, ggplot functions

## Steps
# 1. Import Packages
# 2. Load in Data
# 3.  data
# 4. Visualize difference in team/player valuations across leagues

### Import Packages

library(dplyr)
library(tidyr)
library(ggplot2)
library(knitr)
library(kableExtra)
library(scales)     # dollar()
library(googlesheets4)
library(readr) #for parse_number function

### Load in Data (google sheet)

sheet_url <- "https://docs.google.com/spreadsheets/d/1S80k3N9s6KpMXTTkIb2YCQnrlBAdc8YoeB6YkqI
gs4_deauth()
big_5_valuations <- read_sheet(sheet_url)


### Wrangle Data - add column with valuation as numeric figure
## €1.31bn -> 1310000000

big_5_valuations_wrangled <- big_5_valuations %>%
  mutate(
    valuation = case_when(
      str_detect(Valuation, "bn") ~ parse_number(Valuation) * 1e9,
      str_detect(Valuation, "m") ~ parse_number(Valuation) * 1e6,
    )
  ) %>%
  select(
    "Club",
```

```
    "League",
    "valuation"
  )

### Visualize difference in team valuations across big 5 leagues

## Create Sorted Table

# Table of ten teams with highest valuation

top_ten <- big_5_valuations_wrangled %>% #selecting top 10
  slice_max(
    order_by = valuation,
    n = 10
    ) %>%
  mutate(              # 1.14 dollars/euro conversion
    valuation = dollar(1.14*valuation) #for formatting in table in dollars
  )

top_ten %>% #creating table
  kable(
    booktabs = TRUE,
    align = c(rep("l", 3)),
    col.names = c("Squad Name", "Big 5 League", "Valuation")
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 16
  )

#### Chunk5.1 - How is the Premier League Different from the Other "Big 5" Leagues Financial

# Table of ten lowest average player values

bottom_ten <- big_5_valuations_wrangled %>% #selecting top 10
  slice_min(
    order_by = valuation,
    n = 10
    ) %>%
  mutate(              # 1.14 dollars/euro conversion
    valuation = dollar(1.14*valuation) #for formatting in table in dollars
  )
```

```r
bottom_ten %>% #creating table
  kable(
    booktabs = TRUE,
    align = c(rep("l", 3)),
    col.names = c("Squad Name", "Big 5 League", "Valuation")
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c("striped", "condensed"),
    font_size = 16
  )

#### Chunk5.2 - How is the Premier League Different from the Other "Big 5" Leagues Financial

### Visualize difference in league valuations across big 5 leagues

## Wrangle data to be displayed in a better scale

big_5_valuations_wrangled <- big_5_valuations_wrangled %>%
  mutate(
    valuation = (1.14*valuation)/1000000
  )

## Create Viz (Box plot of 5 leagues valuations distributions)

ggplot(
  big_5_valuations_wrangled,
  aes(
    x = League,
    y = valuation
    )
  ) +
  geom_boxplot(
    fill = "#69b3a2"
  ) +
  labs(
    x = "Big 5 League",
    y = "Team Valuations (millions of dollars)"
  ) +
  scale_y_continuous(
    breaks = seq(0, 1500, by = 250)
  ) +
  theme_bw()
```