Stats 15 Project – NBA Home Court Advantage

Group 4: Alex Andrada, Kyle Biagan, Tyler Chia, Gloria Thet, Daniel Wang

2022-12-05

Introduction

Research Question and Areas of Analysis

In the NBA, teams play half of their regular season games at their home stadium and play the other half as the away team at other teams' stadiums. Although the game of basketball remains the same no matter where the game is played, home-court advantage is believed to be a psychological and biased benefit that home teams experience due to factors like the familiarity/comfort with the arena and the encouragement from home fans. In this project, we will utilize data from the 2003-2004 to 2021-2022 NBA seasons to answer the question: **How important is home court advantage in the NBA?**

Background on the Game of Basketball

The game of basketball is played by two teams on an indoor rectangular court. On each end of the court lies a 10-foot hoop and net, and each team attempts to score by shooting the ball through the opponent's hoop. Each team gets to have five players on the court at a time and teams take turns playing offense and defense. When the team is on offense, the NBA features a 24 second shot clock in which the team must attempt a shot that hits the rim. If they do not attempt a shot or are unable to hit the rim within the 24 seconds, it is a shot clock violation and the defending team gets possession. Players can score either one, two, or three points during a possession. When in bounds, players score three points beyond the three-point line, two points for any basket inside the three-point line, or one point for a made free throw. A free throw is an uncontested shot at the free throw line a player takes when they get fouled while shooting or fouled while their team is "in the bonus". If a player gets fouled while attempting a three-point shot, they are awarded three free throw attempts, while if they get fouled while attempting a two-point shot, they get awarded two free throw attempts. If a player is fouled by the defense while taking a shot, and the shot is made, the shooter is awarded one free throw attempt in addition to those points being scored. Another way players can get free throws is if they are fouled in any form while their team is "in the bonus". An NBA team is "in the bonus" and is rewarded with two free throws when the opposing team has committed its fifth common foul of the quarter. The offense will continue to earn two free throw attempts for each additional common foul the defense commits until the quarter ends. Once a quarter ends and the next quarter begins, each team begins with zero common fouls for the quarter. In the NBA, each game has four quarters of regulation play and each quarter is twelve minutes long. If the score is tied at the end of regulation play, the teams will play for five more minutes in an "overtime" period to decide the winner. If the score remains tied after that five-minute overtime period, the teams will repeat the same overtime procedure until someone wins since NBA games are not allowed to end in a tie.

Background on the NBA

From the 2004/2005 season to the present day, there has been a total of 30 teams in the NBA. In the 2003/2004 season, there were 29 teams, but the addition of the Charlotte Bobcats (now the Hornets) for

the 2004/2005 season pushed the New Orleans Hornets (now the Pelicans) to the Western Conference and created a 30 team league. Within each conference, there are divisions. Since the addition of the Bobcats in 2004, there have been six divisions. Divisions are divided into teams in the same region. The Eastern Conference has three divisions called Atlantic, Central, and Southeast. The Western Conference also has three divisions, called the Northwest, Pacific, and Southwest. Teams within the same division typically play each other four times per year, twice at home and twice on the road. All 30 NBA teams play against the other 29 opponents at least twice during the regular season. The NBA regular season typically lasts from around October to May with 82 games, and each team plays 41 games at home and 41 games away. There have been some exceptions: the 2011-2012 season was delayed due to a 161-day lockout prior to the season that was centered around a conflict of revenue division between players and owners, the 2019-2020 season was abruptly halted due to the pandemic and later finished with fewer games, and the 2020-2021 season was shortened due to a late start as a result of the pandemic and the late finish of the prior season. Following the regular season, the top eight teams in each conference makes the playoffs. Beginning in the 2019/2020 season, the NBA began to implement a "play-in tournament" that featured the 7th, 8th, 9th, and 10th seeds in each conference and made them compete for the final two playoff spots. In the playoffs, there are four rounds and each round is a best-of-seven format. In the first round, the top ranked team plays the lowest ranked team, the #2 team plays the #7 team, the #3 team plays the #6 team, and the #4 team plays the #5 team. Within each series, the higher seeded team gets to play the first two games at home, and the lower seeded team plays the next two at home. If a series winner is not yet determined, the higher seeded team plays the fifth game at home, the lower seeded team plays the sixth game at home, and the higher seeded team would play Game 7 at home.

Explanation of the Variables in the Data

GAME_DATE_EST: The date of the game

SEASON: The season when the game occurred

HOME_TEAM: Home team city name

HOME_TEAM_W: The number of wins the home team had during the season at the time of the game

HOME_TEAM_L: The number of losses the home team had during the season at the time of the game

HOME_TEAM_WPCT: The win percentage of the home team during the season at the time of the game, calculated as HOME_TEAM_W divided by HOME_TEAM_L

PTS_home: Number of points scored by the home team

FG_PCT_home: Field goal percentage of the home team, calculated as field goals made divided by field goals attempted

FT_PCT_home: Free throw percentage of the home team, calculated as free throws made divided by free throws attempted

FG3_PCT_home: Three-point percentage of the home team, calculated as three point field goals made divided by three point field goals attempted

AST_home: Number of assists recorded by the home team. An assist is when a player passes the ball to a teammate that directly leads to a score by a made field goal

REB_home: Number of rebounds recorded by the home team. A rebound is when a player retrieves the ball after a missed field goal or free throw

AWAY_TEAM: Away team city name

AWAY_TEAM_W: The number of wins the away team had during the season at the time of the game

AWAY_TEAM_L: The number of losses the away team had during the season at the time of the game

AWAY_TEAM_WPCT: The win percentage of the away team during the season at the time of the game, calculated as AWAY_TEAM_W divided by AWAY_TEAM_L

PTS_away: Number of points scored by the away team

FG_PCT_away: Field goal percentage by the away team

FT_PCT_away: Free throw percentage of the away team

FG3_PCT_away: Three-point percentage of the away team

AST_away: Number of assists recorded by the away team

REB_away: Number of assists recorded by the away team

HOME_TEAM_WINS: If the home team won the game (1 if yes, 0 if no)

Each row is an observational unit, representing one NBA game. Basketball data is recorded officially at every game. The NBA uses STATS SportVU, which uses advanced technology and a system of cameras to track and measure the movements of all players and the ball on the court in order to record accurate statistics.

Data Loading + Cleanup

Installing Packages

We first will upload all necessary packages for the project.

```
library(tidyverse)
library(dplyr)
library(ggplot2)
library(magrittr)
library(knitr)
library(lubridate)
library(readr)
```

Data Loading

First, we will load the all of the data tables into the Rmd file. We will be using two different csv files with different statistics about the NBA and below are the top values in each dataset.

```
games <- read_csv("games.csv")
ranking <- read_csv("ranking.csv")

head(games)
```

```
## # A tibble: 6 x 21
##   GAME_DATE_EST  GAME_ID GAME_S~1 HOME_~2 VISIT~3 SEASON TEAM_~4 PTS_h~5 FG_PC~6
##   <date>           <dbl> <chr>      <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 2022-03-12    22101005 Final     1.61e9  1.61e9   2021  1.61e9     104   0.398
## 2 2022-03-12    22101006 Final     1.61e9  1.61e9   2021  1.61e9     101   0.443
## 3 2022-03-12    22101007 Final     1.61e9  1.61e9   2021  1.61e9     108   0.412
## 4 2022-03-12    22101008 Final     1.61e9  1.61e9   2021  1.61e9     122   0.484
## 5 2022-03-12    22101009 Final     1.61e9  1.61e9   2021  1.61e9     115   0.551
## 6 2022-03-12    22101010 Final     1.61e9  1.61e9   2021  1.61e9     134   0.558
```

```
## # ... with 12 more variables: FT_PCT_home <dbl>, FG3_PCT_home <dbl>,
## #   AST_home <dbl>, REB_home <dbl>, TEAM_ID_away <dbl>, PTS_away <dbl>,
## #   FG_PCT_away <dbl>, FT_PCT_away <dbl>, FG3_PCT_away <dbl>, AST_away <dbl>,
## #   REB_away <dbl>, HOME_TEAM_WINS <dbl>, and abbreviated variable names
## #   1: GAME_STATUS_TEXT, 2: HOME_TEAM_ID, 3: VISITOR_TEAM_ID, 4: TEAM_ID_home,
## #   5: PTS_home, 6: FG_PCT_home
```

```
head(ranking)
```

```
## # A tibble: 6 x 13
##      TEAM_ID LEAGUE_ID SEASON~1 STANDING~2 CONFE~3 TEAM      G     W     L W_PCT
##        <dbl> <chr>        <dbl> <date>     <chr>   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 1610612756 00           22021 2022-03-12 West    Phoe~    67    53    14 0.791
## 2 1610612744 00           22021 2022-03-12 West    Gold~    68    46    22 0.676
## 3 1610612763 00           22021 2022-03-12 West    Memp~    68    46    22 0.676
## 4 1610612762 00           22021 2022-03-12 West    Utah     67    42    25 0.627
## 5 1610612742 00           22021 2022-03-12 West    Dall~    67    41    26 0.612
## 6 1610612743 00           22021 2022-03-12 West    Denv~    68    40    28 0.588
## # ... with 3 more variables: HOME_RECORD <chr>, ROAD_RECORD <chr>,
## #   RETURNTOPLAY <dbl>, and abbreviated variable names 1: SEASON_ID,
## #   2: STANDINGSDATE, 3: CONFERENCE
```

Using Lubridate Package

We will first use the lubridate package to turn the STANDINGSDATE from character to numerical format. This will allow us to join the games and rankings tables.

```
library(lubridate)
ranking$STANDINGSDATE <- ymd(ranking$STANDINGSDATE)
```

Joining the Tables

We will now join the tables together to make one data table with all of the information we need. We will use the inner_join() function to do so. We will also select the specific columns we need, removing extra unnecessary columns such as CONFERENCE, RETURNTOPLAY, TEAM_ID_away, and TEAM_ID_home.

```
NBA <-
inner_join(games, ranking, by = c("GAME_DATE_EST" = "STANDINGSDATE", "HOME_TEAM_ID" =
↪  "TEAM_ID")) %>%
  inner_join(ranking, by = c("GAME_DATE_EST" = "STANDINGSDATE", "VISITOR_TEAM_ID" =
  ↪  "TEAM_ID"))  %>%
  select(GAME_DATE_EST, GAME_ID, SEASON, PTS_home, FG_PCT_home, FT_PCT_home,
  ↪  FG3_PCT_home, AST_home, REB_home, PTS_away, FT_PCT_away, FG_PCT_away, FG3_PCT_away,
  ↪  AST_away, REB_away, HOME_TEAM_WINS, TEAM.x, W.x, L.x, TEAM.y, W.y, L.y, G.x, G.y)
NBA
```

```
## # A tibble: 25,816 x 24
##    GAME_DATE_EST  GAME_ID SEASON PTS_h~1 FG_PC~2 FT_PC~3 FG3_P~4 AST_h~5 REB_h~6
##    <date>           <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 2022-03-12    22101005   2021     104   0.398   0.76    0.333      23      53
```

```
##  2 2022-03-12    22101006   2021     101   0.443   0.933   0.429      20      46
##  3 2022-03-12    22101007   2021     108   0.412   0.813   0.324      28      52
##  4 2022-03-12    22101008   2021     122   0.484   0.933   0.4        33      55
##  5 2022-03-12    22101009   2021     115   0.551   0.75    0.407      32      39
##  6 2022-03-12    22101010   2021     134   0.558   0.71    0.39       21      44
##  7 2022-03-12    22101011   2021     127   0.516   0.909   0.367      21      43
##  8 2022-03-11    22100995   2021     118   0.465   0.88    0.4        31      49
##  9 2022-03-11    22100996   2021     112   0.478   0.895   0.29       28      47
## 10 2022-03-11    22100997   2021     114   0.467   0.8     0.188      23      47
## # ... with 25,806 more rows, 15 more variables: PTS_away <dbl>,
## #   FT_PCT_away <dbl>, FG_PCT_away <dbl>, FG3_PCT_away <dbl>, AST_away <dbl>,
## #   REB_away <dbl>, HOME_TEAM_WINS <dbl>, TEAM.x <chr>, W.x <dbl>, L.x <dbl>,
## #   TEAM.y <chr>, W.y <dbl>, L.y <dbl>, G.x <dbl>, G.y <dbl>, and abbreviated
## #   variable names 1: PTS_home, 2: FG_PCT_home, 3: FT_PCT_home,
## #   4: FG3_PCT_home, 5: AST_home, 6: REB_home
```

Renaming Variables

The columns such as TEAM, W, L, W_PCT, etc. are in the table for both the home and away team; therefore, it is currently denoted by (.x) and (.y). We will now rename the columns so that each variable is clearly identified by either home or away and the statistic that specific row looks at.

```
library(dplyr)
NBA <-
rename(NBA, HOME_TEAM = TEAM.x)
```

```
NBA <-
rename(NBA, HOME_TEAM_W = W.x)
```

```
NBA <-
rename(NBA, HOME_TEAM_L = L.x)
```

```
NBA <-
  rename(NBA, HOME_TEAM_GAME = G.x)
```

```
NBA <-
rename(NBA, AWAY_TEAM = TEAM.y)
```

```
NBA <-
rename(NBA, AWAY_TEAM_W = W.y)
```

```
NBA <-
rename(NBA, AWAY_TEAM_L = L.y)
```

```
NBA <-
  rename(NBA, AWAY_TEAM_GAME = G.y)
```

Now that we have all of the columns renamed, we will calculate the win percentage for both the home and away team using the mutate() function.

```
NBA <- NBA %>%
  mutate(HOME_TEAM_WPCT = HOME_TEAM_W / HOME_TEAM_GAME * 100) %>%
  mutate(AWAY_TEAM_WPCT = AWAY_TEAM_W / AWAY_TEAM_GAME * 100)
```

We can also represent field goal percentage, three point field goal percentage, and free throw percentage as percentages rather than proportions by using the mutate() function and multiplying each value by 100

```
NBA <- NBA %>%
  mutate(FG_PCT_home = FG_PCT_home * 100) %>%  mutate(FG_PCT_away = FG_PCT_away * 100)
  ↪  %>%  mutate(FT_PCT_home = FT_PCT_home * 100) %>% mutate(FT_PCT_away = FT_PCT_away *
  ↪  100) %>% mutate(FG3_PCT_home = FG3_PCT_home * 100) %>%  mutate(FG3_PCT_away =
  ↪  FG3_PCT_away * 100)
NBA
```

```
## # A tibble: 25,816 x 26
##    GAME_DATE_EST  GAME_ID SEASON PTS_h~1 FG_PC~2 FT_PC~3 FG3_P~4 AST_h~5 REB_h~6
##    <date>           <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 2022-03-12    22101005   2021     104    39.8    76      33.3      23      53
##  2 2022-03-12    22101006   2021     101    44.3    93.3    42.9      20      46
##  3 2022-03-12    22101007   2021     108    41.2    81.3    32.4      28      52
##  4 2022-03-12    22101008   2021     122    48.4    93.3    40        33      55
##  5 2022-03-12    22101009   2021     115    55.1    75      40.7      32      39
##  6 2022-03-12    22101010   2021     134    55.8    71      39        21      44
##  7 2022-03-12    22101011   2021     127    51.6    90.9    36.7      21      43
##  8 2022-03-11    22100995   2021     118    46.5    88      40        31      49
##  9 2022-03-11    22100996   2021     112    47.8    89.5    29        28      47
## 10 2022-03-11    22100997   2021     114    46.7    80      18.8      23      47
## # ... with 25,806 more rows, 17 more variables: PTS_away <dbl>,
## #   FT_PCT_away <dbl>, FG_PCT_away <dbl>, FG3_PCT_away <dbl>, AST_away <dbl>,
## #   REB_away <dbl>, HOME_TEAM_WINS <dbl>, HOME_TEAM <chr>, HOME_TEAM_W <dbl>,
## #   HOME_TEAM_L <dbl>, AWAY_TEAM <chr>, AWAY_TEAM_W <dbl>, AWAY_TEAM_L <dbl>,
## #   HOME_TEAM_GAME <dbl>, AWAY_TEAM_GAME <dbl>, HOME_TEAM_WPCT <dbl>,
## #   AWAY_TEAM_WPCT <dbl>, and abbreviated variable names 1: PTS_home,
## #   2: FG_PCT_home, 3: FT_PCT_home, 4: FG3_PCT_home, 5: AST_home, ...
```

Filtering Out Games

This data set contains all NBA game logs from the 2003 season until the 2022 season, including preseason, regular season, and playoff games. Because preseason games are usually less competitive and generally don't count for anything, with most teams resting their best players or only playing them for limited minutes, they can skew the results of our analysis, so we will filter them out by only leaving only the dates of the regular season and playoffs.

```
NBA <- NBA %>%
  filter(GAME_DATE_EST >= "2003-10-28" & GAME_DATE_EST <= "2004-06-15" | GAME_DATE_EST >=
   ↪  "2004-11-02" & GAME_DATE_EST <= "2005-06-23" | GAME_DATE_EST >= "2005-11-01" &
   ↪  GAME_DATE_EST <= "2006-06-20" | GAME_DATE_EST >= "2006-10-31" & GAME_DATE_EST <=
   ↪  "2007-06-14" | GAME_DATE_EST >= "2007-10-30" & GAME_DATE_EST <= "2008-06-17" |
   ↪  GAME_DATE_EST >= "2008-10-28" & GAME_DATE_EST <= "2009-06-14" | GAME_DATE_EST >=
   ↪  "2009-10-27" & GAME_DATE_EST <= "2010-06-17" | GAME_DATE_EST >= "2010-10-26" &
   ↪  GAME_DATE_EST <= "2011-06-12" |  GAME_DATE_EST >= "2011-12-25" & GAME_DATE_EST <=
   ↪  "2012-06-21" | GAME_DATE_EST >= "2012-10-30" & GAME_DATE_EST <= "2013-06-20" |
   ↪  GAME_DATE_EST >= "2013-10-29" & GAME_DATE_EST <= "2014-06-15" | GAME_DATE_EST >=
   ↪  "2014-10-27" & GAME_DATE_EST <= "2015-06-16" | GAME_DATE_EST >= "2015-10-27" &
   ↪  GAME_DATE_EST <= "2016-06-19" | GAME_DATE_EST >= "2016-10-25" & GAME_DATE_EST <=
   ↪  "2017-06-12" | GAME_DATE_EST >= "2017-10-17" & GAME_DATE_EST <= "2018-06-08" |
   ↪  GAME_DATE_EST >= "2018-10-16" & GAME_DATE_EST <= "2019-06-13" | GAME_DATE_EST >=
   ↪  "2019-10-22" & GAME_DATE_EST <= "2020-10-11" | GAME_DATE_EST >= "2020-12-22" &
   ↪  GAME_DATE_EST <= "2021-07-20" | GAME_DATE_EST >= "2021-10-19" & GAME_DATE_EST <=
   ↪  "2022-06-16")
```

Removing Duplicates

During our analysis, we noticed a few duplicates between 2020-12-26 and 2020-12-29, so we want to be sure to get rid of all duplicates

```
NBA <- NBA %>%
distinct(GAME_ID, .keep_all = TRUE)
```

Reordering Columns

We will now select and reorder the necessary columns in the table so that all of the home team data is in one section and the away team data is in the other.

```
NBA <- NBA %>%
  select(GAME_DATE_EST, SEASON, HOME_TEAM, HOME_TEAM_W, HOME_TEAM_L, HOME_TEAM_WPCT,
   ↪  PTS_home, FG_PCT_home, FT_PCT_home, FG3_PCT_home, AST_home, REB_home, AWAY_TEAM,
   ↪  AWAY_TEAM_W, AWAY_TEAM_L, AWAY_TEAM_WPCT, PTS_away, FG_PCT_away, FT_PCT_away,
   ↪  FG3_PCT_away, AST_away, REB_away, HOME_TEAM_WINS)
```

The data is now loaded, joined, and cleaned. All necessary columns are now in one large dataset and below is a glimpse of the data.

```
NBA %>%
  glimpse()
```

```
## Rows: 24,091
## Columns: 23
## $ GAME_DATE_EST  <date> 2022-03-12, 2022-03-12, 2022-03-12, 2022-03-12, 2022-0~
## $ SEASON         <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2~
## $ HOME_TEAM      <chr> "Miami", "Chicago", "San Antonio", "Golden State", "Den~
## $ HOME_TEAM_W    <dbl> 45, 41, 26, 46, 40, 42, 26, 18, 32, 41, 45, 17, 46, 27,~
## $ HOME_TEAM_L    <dbl> 24, 26, 42, 22, 28, 25, 40, 50, 34, 27, 23, 50, 22, 40,~
## $ HOME_TEAM_WPCT <dbl> 65.21739, 61.19403, 38.23529, 67.64706, 58.82353, 62.68~
## $ PTS_home       <dbl> 104, 101, 108, 122, 115, 134, 127, 118, 112, 114, 117, ~
## $ FG_PCT_home    <dbl> 39.8, 44.3, 41.2, 48.4, 55.1, 55.8, 51.6, 46.5, 47.8, 4~
## $ FT_PCT_home    <dbl> 76.0, 93.3, 81.3, 93.3, 75.0, 71.0, 90.9, 88.0, 89.5, 8~
## $ FG3_PCT_home   <dbl> 33.3, 42.9, 32.4, 40.0, 40.7, 39.0, 36.7, 40.0, 29.0, 1~
## $ AST_home       <dbl> 23, 20, 28, 33, 32, 21, 21, 31, 28, 23, 28, 29, 23, 31,~
## $ REB_home       <dbl> 53, 46, 52, 55, 39, 44, 43, 49, 47, 47, 42, 39, 53, 40,~
## $ AWAY_TEAM      <chr> "Minnesota", "Cleveland", "Indiana", "Milwaukee", "Toro~
## $ AWAY_TEAM_W    <dbl> 39, 38, 23, 42, 37, 24, 29, 38, 35, 18, 38, 41, 28, 33,~
## $ AWAY_TEAM_L    <dbl> 30, 29, 45, 26, 30, 45, 37, 30, 34, 49, 28, 26, 39, 35,~
## $ AWAY_TEAM_WPCT <dbl> 56.52174, 56.71642, 33.82353, 61.76471, 55.22388, 34.78~
## $ PTS_away       <dbl> 113, 91, 119, 109, 127, 125, 118, 110, 106, 103, 105, 1~
## $ FG_PCT_away    <dbl> 42.2, 41.9, 48.9, 41.3, 47.1, 50.0, 47.0, 45.6, 48.8, 4~
## $ FT_PCT_away    <dbl> 87.5, 82.4, 100.0, 69.6, 76.0, 85.7, 96.3, 100.0, 82.4,~
## $ FG3_PCT_away   <dbl> 35.7, 20.8, 38.9, 38.6, 38.7, 39.4, 41.2, 33.3, 37.5, 2~
## $ AST_away       <dbl> 21, 19, 23, 27, 28, 27, 26, 24, 22, 21, 25, 20, 23, 41,~
## $ REB_away       <dbl> 46, 40, 47, 39, 50, 33, 35, 37, 36, 42, 40, 47, 54, 38,~
## $ HOME_TEAM_WINS <dbl> 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0~
```

Comparing Home vs Away Stats for all of the NBA from the 2003/2004 to 2021/2022 season

Wins

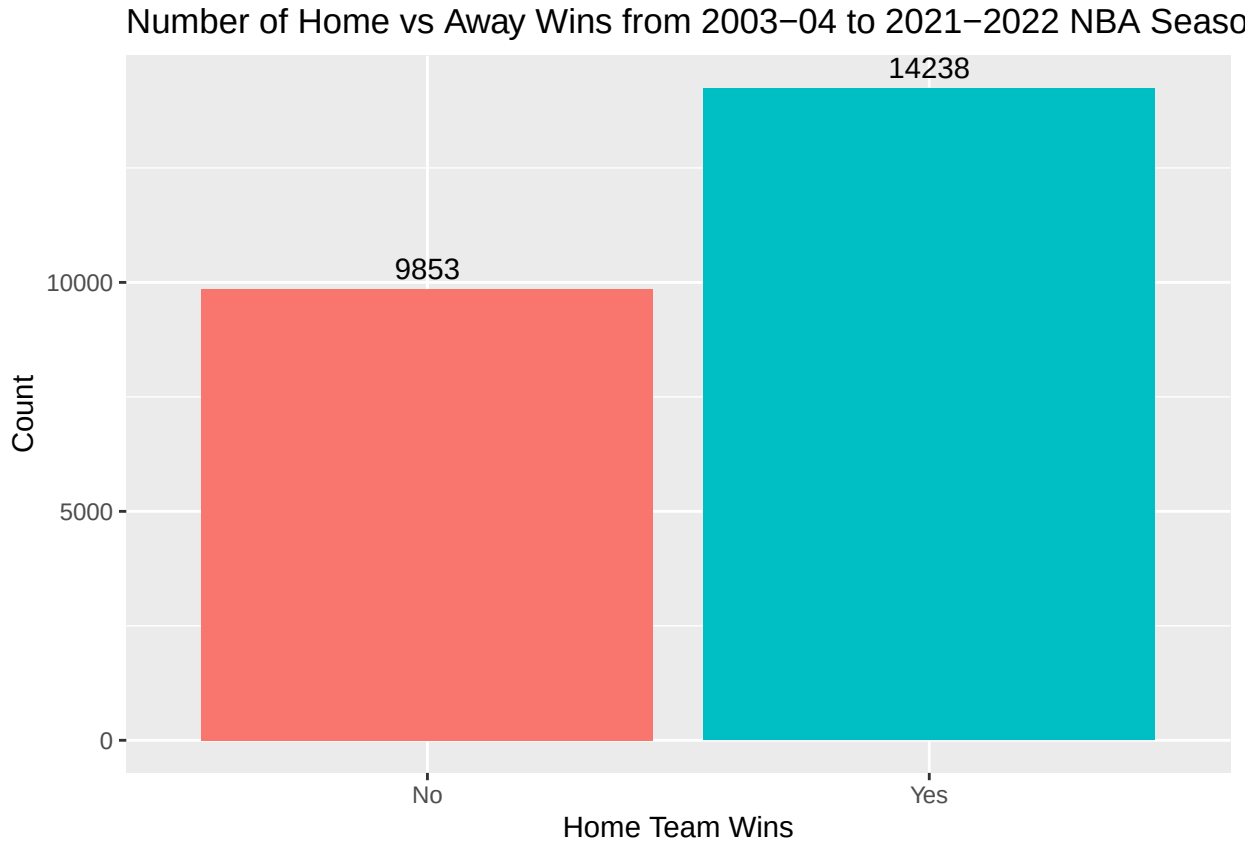
First, we can look at how many times the home team has won across the entire data set.

```
NBA %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0  9853
## 2              1 14238
```

We find that the home team has won 14238 out of 24091 games. If home court advantage did not exist, we would expect the home team to win half of the games and the away team to win the other half. We can create a bar chart to display the difference in win percentage and conduct a one-prop Z test to see if our result is statistically significant.

```
Home_Team_Wins <- c("No", "Yes")
Number <- c(9853, 14238)
Home_wins <- data.frame(Home_Team_Wins, Number)

Home_wins %>%
  ggplot(aes(x = Home_Team_Wins, y = Number)) + geom_col(aes(fill = Home_Team_Wins)) +
  ↪  geom_text(aes(label = Number), vjust = -0.5) + theme(legend.position = "none") +
  ↪  labs(x = "Home Team Wins", y = "Count", title = "Number of Home vs Away Wins from
  ↪  2003-04 to 2021-2022 NBA Season")
```




```
prop.test(14238, 24091, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  14238 out of 24091, null probability 0.5
## X-squared = 797.79, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5847659 0.5972231
## sample estimates:
##         p
## 0.5910091
```

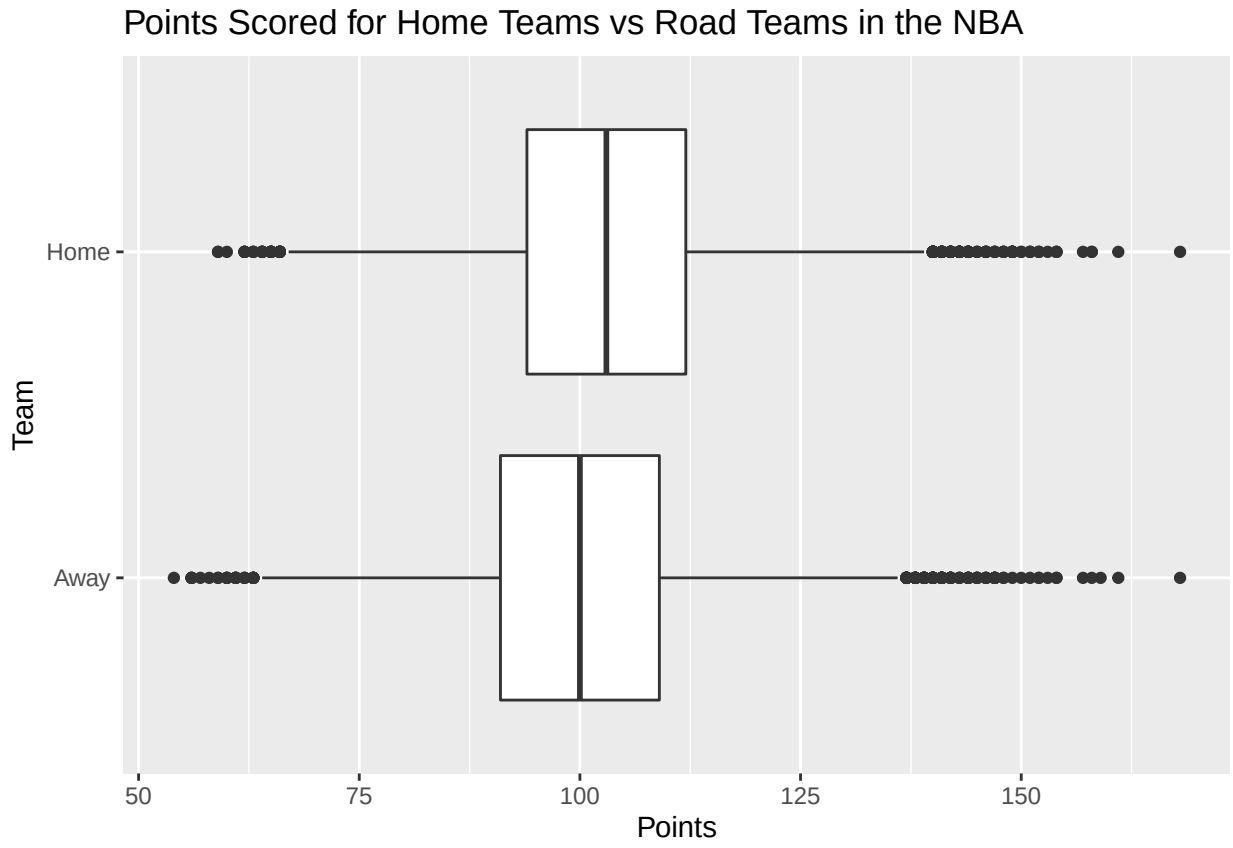
Based on the barchart created above, we see that there are many more instances where the home team wins versus when they lose. Since the 2003/2004 season, the home team has won 59.1 percent of the time. By conducting a one-proportion Z-Test, we get an extremely small p-value of less than 2.2 x 10^-16 (very close to zero), which suggests that there is a statistically significant difference between the sample proportion of home wins (0.591) with 0.5, which would be the expected proportion of home wins if home court advantage didn't exist. Since we can conclude that home teams have won more often than away teams since the 2003/2004 season, lets explore the specific statistics affected by home court advantage.

Individual Statistics

We can use the pivot_longer function to make our data easier to analyze and create box plots for each of the individual statistics to see the difference between home teams and road teams. If home court advantage didn't exist, we would expect the same box plot for home vs road teams. We will also be using a paired t-test in order to determine if the differences are statistically significant (whether or not the result is due to chance or an outside factor that influenced the result). The paired t-tests will be used because the data is in the form of matched pairs. The two samples are not independent because the home team plays against the away team, so the values in one sample affect the values in the other.

Points

```
NBA %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(y = "Team", title = "Points Scored for Home Teams vs Road
  ↪  Teams in the NBA")
```




```
t.test(NBA$PTS_home, NBA$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
```

```
## data:  NBA$PTS_home and NBA$PTS_away
## t = 33.081, df = 24090, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  2.693226 3.032480
## sample estimates:
## mean difference
##        2.862853
```

The box plot above suggests that home teams score more points than away teams as the first quartile, median, and third quartile all are higher compared to away teams. By conducting a paired T test, we can that the difference is statistically significant because the p-value is extremely small (less than 2.2 x 10^-16). From the T-test, we also obtain a 95% confidence interval of 2.693226 to 3.032480, which means that we are 95% confident that the true mean difference between points scored for home and away teams falls between this interval. This suggests that if two teams were the exact same, we would expect the home team to have an advantage within this interval, so around 2.862853 points, which is the mean difference in points scored between home and away teams.

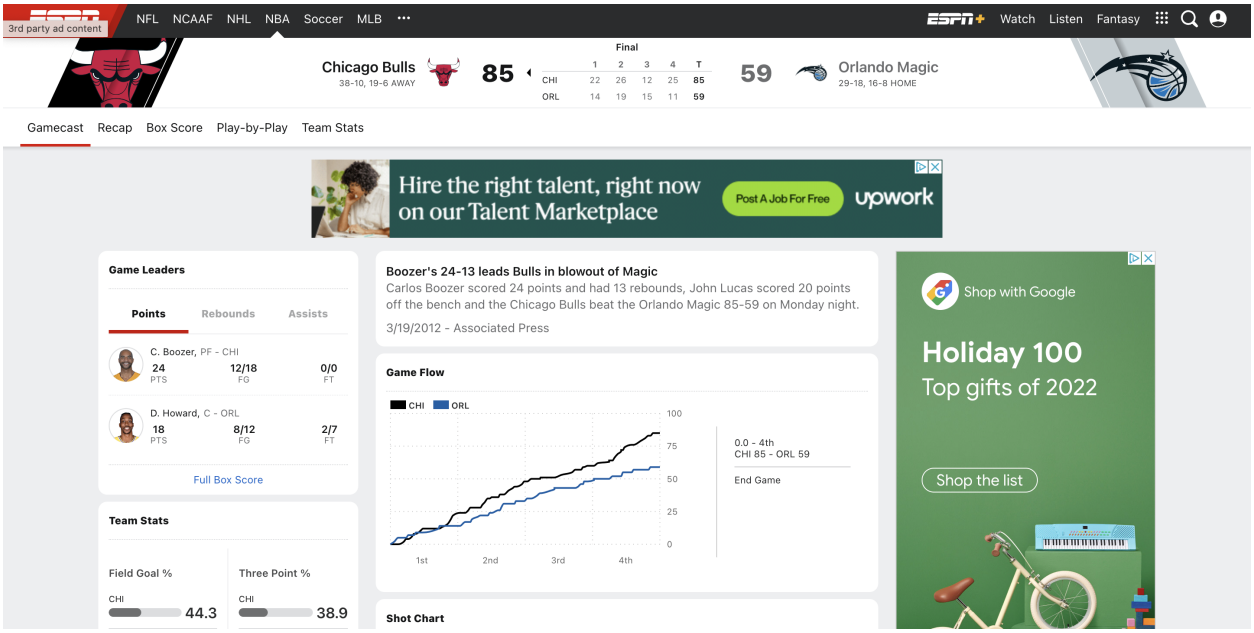
Verifying Outliers

Because we imported the dataset from Kaggle, we will now check for outlier values by looking at the highest and lowest values in points for home and away teams. The box plot shows us the outliers of the data, and we know that it's highly unusual to see a team score less than 60 points or more than 160 points in a game. We will verify that these values are correct by cross-checking them with official box scores from ESPN.

```
NBA %>%
  arrange(PTS_home) %>%
  select(HOME_TEAM, PTS_home, AWAY_TEAM, PTS_away, GAME_DATE_EST, SEASON) %>%
  head(5)
```

```
## # A tibble: 5 x 6
##   HOME_TEAM   PTS_home AWAY_TEAM    PTS_away GAME_DATE_EST SEASON
##   <chr>          <dbl> <chr>           <dbl> <date>         <dbl>
## 1 Orlando           59 Chicago            85 2012-03-19      2011
## 2 Atlanta           59 New Orleans       100 2011-01-21      2010
## 3 Utah              60 Indiana            84 2005-11-29      2005
## 4 New Orleans       62 Philadelphia       77 2012-11-07      2012
## 5 Charlotte         62 Boston             93 2010-12-11      2010
```

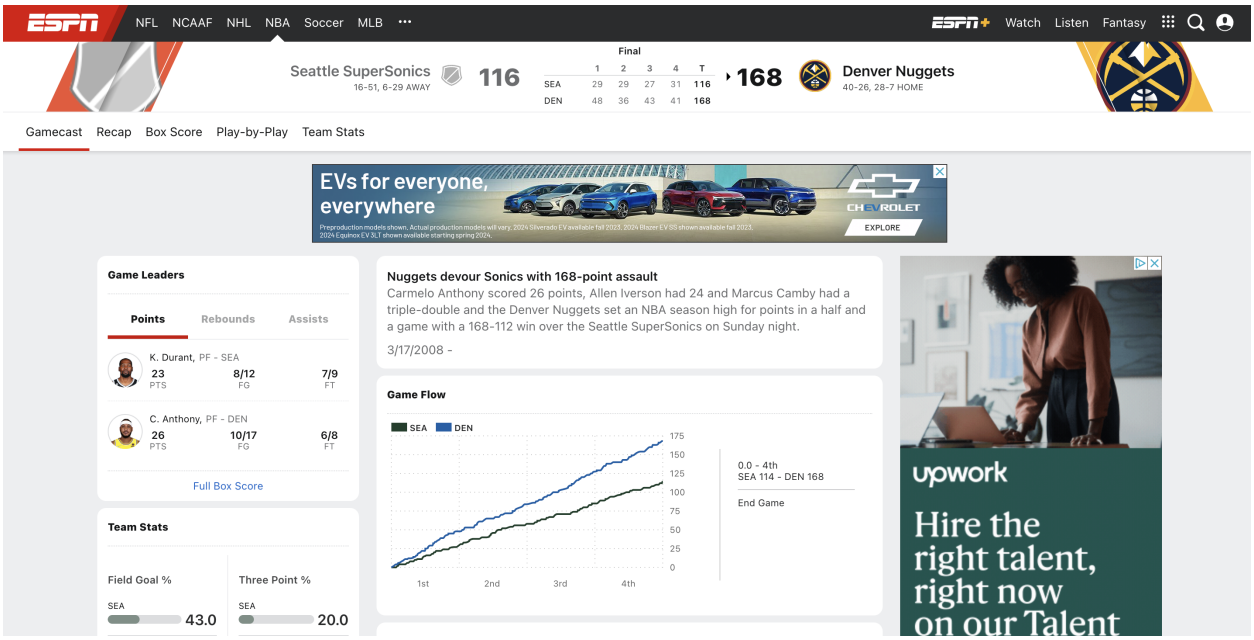
```
knitr::include_graphics("magic59.png")
```




```
NBA %>%
  arrange(desc(PTS_home)) %>%
  select(HOME_TEAM, PTS_home, AWAY_TEAM, PTS_away, GAME_DATE_EST, SEASON) %>%
  head(5)
```

```
## # A tibble: 5 x 6
##   HOME_TEAM  PTS_home AWAY_TEAM PTS_away GAME_DATE_EST SEASON
##   <chr>         <dbl> <chr>        <dbl> <date>          <dbl>
## 1 Denver          168 Seattle        116 2008-03-16       2007
## 2 Atlanta         161 Chicago        168 2019-03-01       2018
## 3 Houston         158 Atlanta        111 2019-11-30       2019
## 4 Washington      158 Houston        159 2019-10-30       2019
## 5 New Jersey      157 Phoenix        161 2006-12-07       2006
```

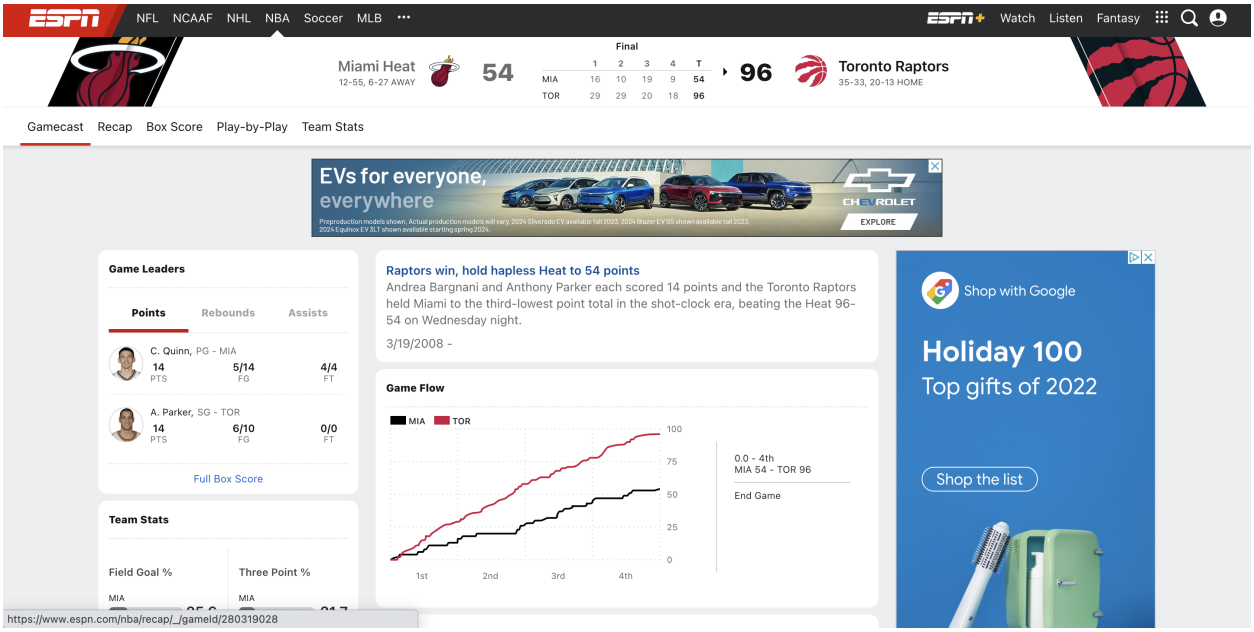
```
knitr::include_graphics("Nuggets168.png")
```




```
NBA %>%
  arrange(PTS_away) %>%
  select(HOME_TEAM, PTS_home, AWAY_TEAM, PTS_away, GAME_DATE_EST, SEASON) %>%
  head(5)
```

```
## # A tibble: 5 x 6
##   HOME_TEAM PTS_home AWAY_TEAM  PTS_away GAME_DATE_EST SEASON
##   <chr>        <dbl> <chr>         <dbl> <date>          <dbl>
## 1 Toronto         96 Miami            54 2008-03-19       2007
## 2 Boston          87 Orlando          56 2012-01-23       2011
## 3 Boston          87 Milwaukee        56 2011-03-13       2010
## 4 Detroit         78 New Jersey       56 2004-05-03       2003
## 5 Minnesota       73 Toronto          56 2003-11-01       2003
```

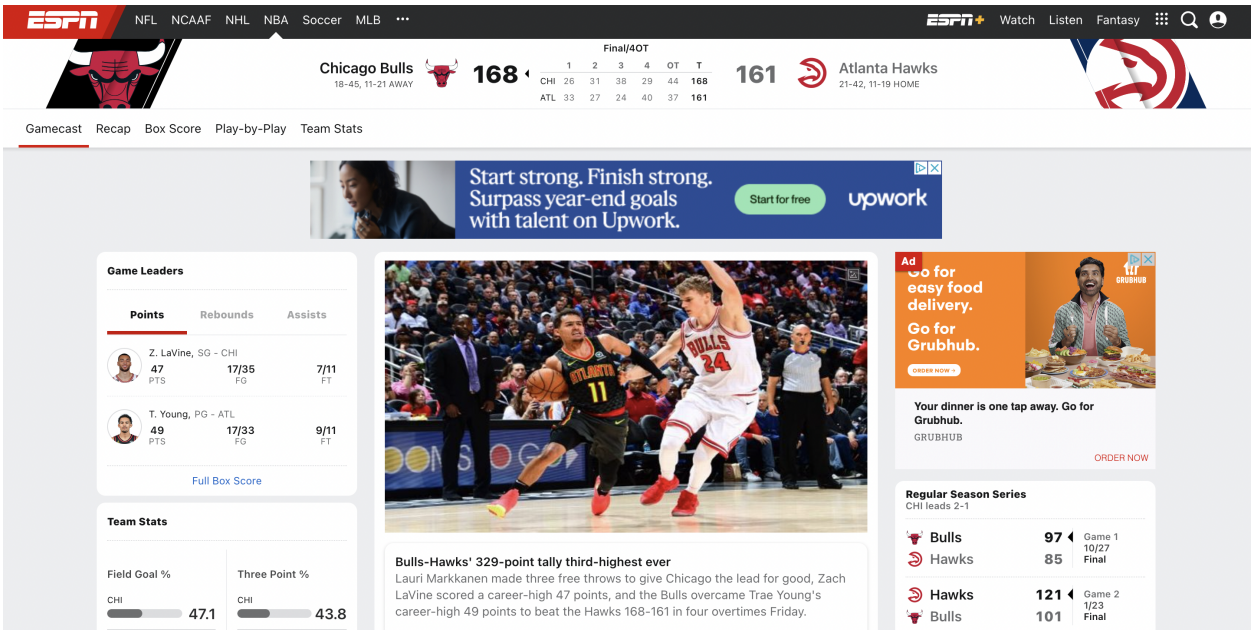
```
knitr::include_graphics("heat54.png")
```




```
NBA %>%
  arrange(desc(PTS_away)) %>%
  select(HOME_TEAM, PTS_home, AWAY_TEAM, PTS_away, GAME_DATE_EST, SEASON) %>%
  head(5)
```

```
## # A tibble: 5 x 6
##   HOME_TEAM  PTS_home AWAY_TEAM   PTS_away GAME_DATE_EST SEASON
##   <chr>         <dbl> <chr>          <dbl> <date>          <dbl>
## 1 Atlanta         161 Chicago          168 2019-03-01       2018
## 2 New Jersey      157 Phoenix          161 2006-12-07       2006
## 3 Washington      158 Houston          159 2019-10-30       2019
## 4 Indiana         126 Charlotte        158 2022-01-26       2021
## 5 Washington      153 San Antonio      157 2022-02-25       2021
```

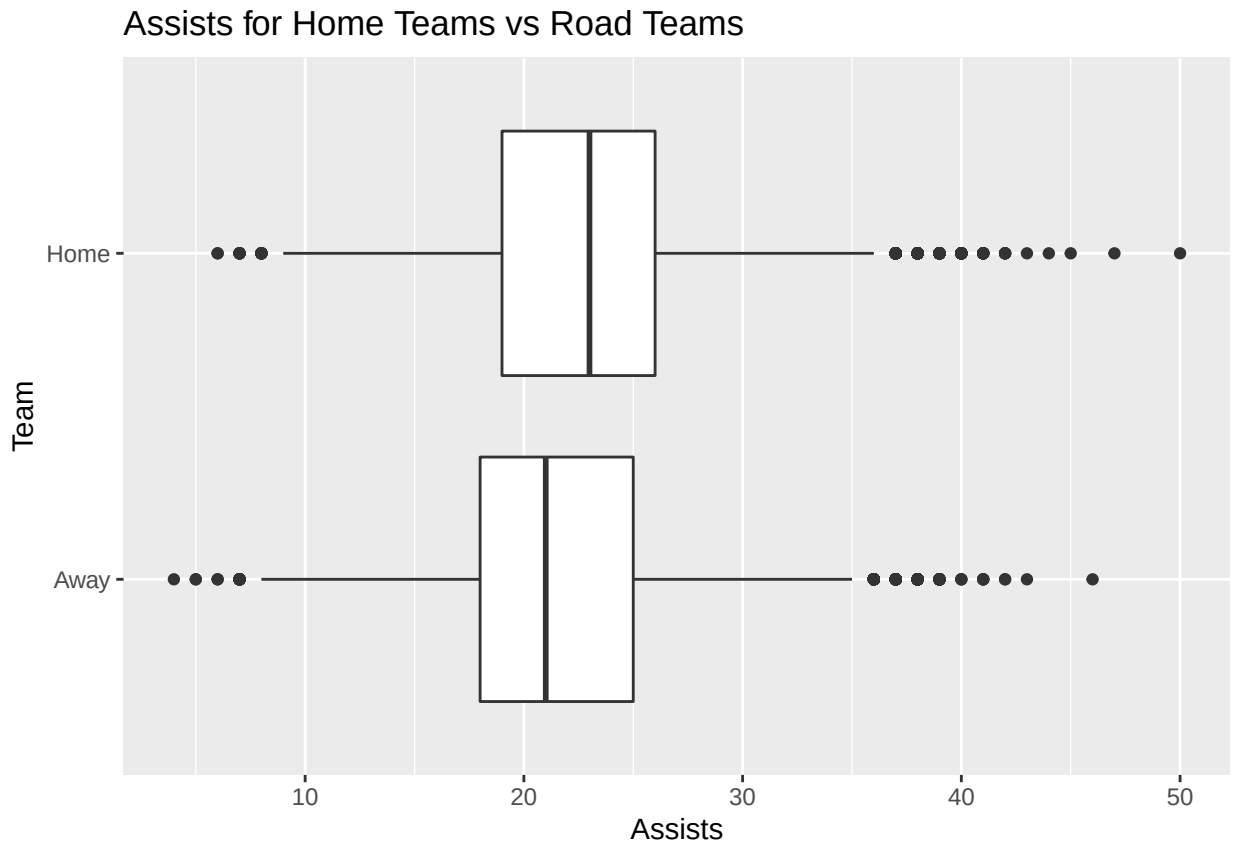
```
knitr::include_graphics("bulls168.png")
```



Now that we have verified that the outlier games did indeed occur, there's no reason to remove any of them and we can move forward with our analysis. Next, we will continue with our analysis by looking at assists, rebounds, field goal percentage, three point field goal percentage, and free throw percentage the same way we analyzed points to see if there are differences between home and away teams.

Assists

```
NBA %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(y = "Team", title = "Assists for Home Teams vs Road
  ↪  Teams")
```

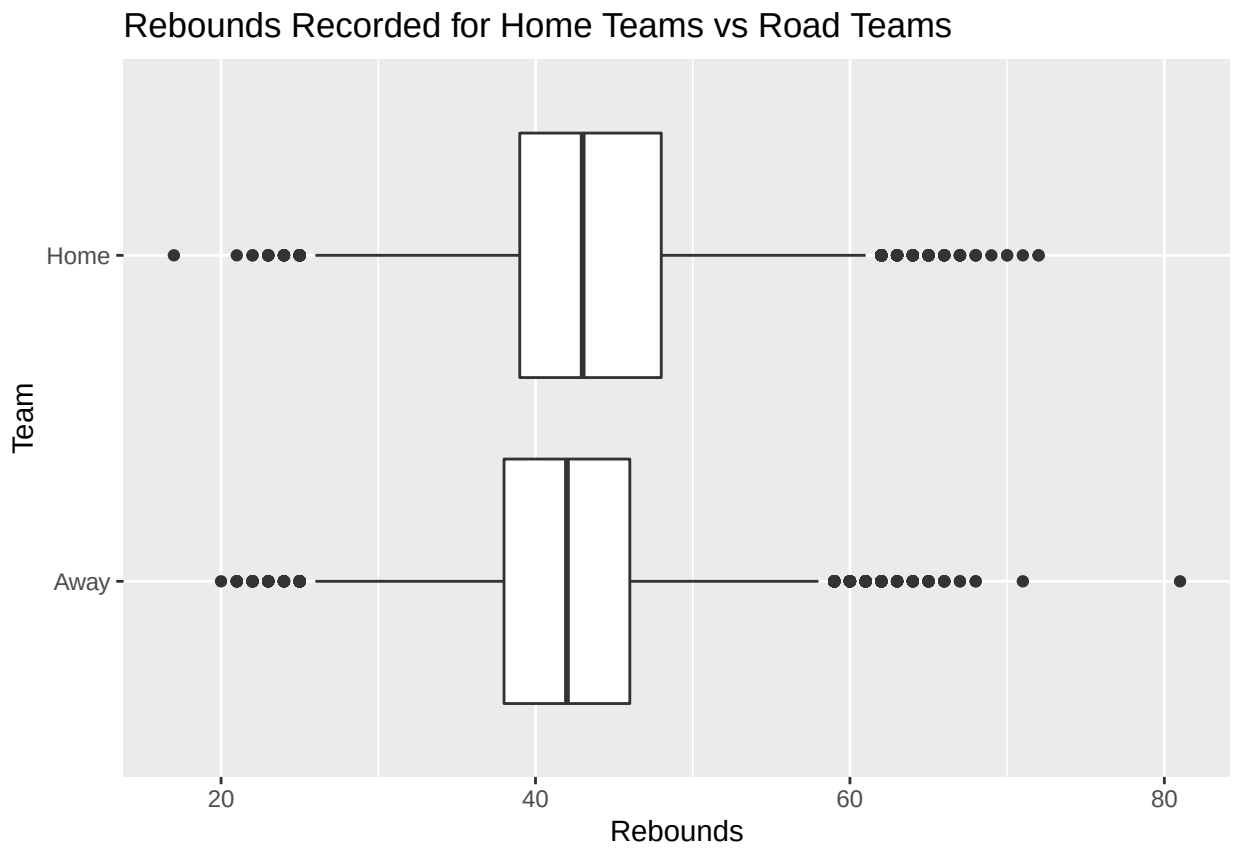



```
t.test(NBA$AST_home, NBA$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBA$AST_home and NBA$AST_away
## t = 30.7, df = 24090, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.254216 1.425291
## sample estimates:
## mean difference
##        1.339753
```

Rebounds

```
NBA %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(y = "Team", title = "Rebounds Recorded for Home Teams vs
  ↪  Road Teams")
```

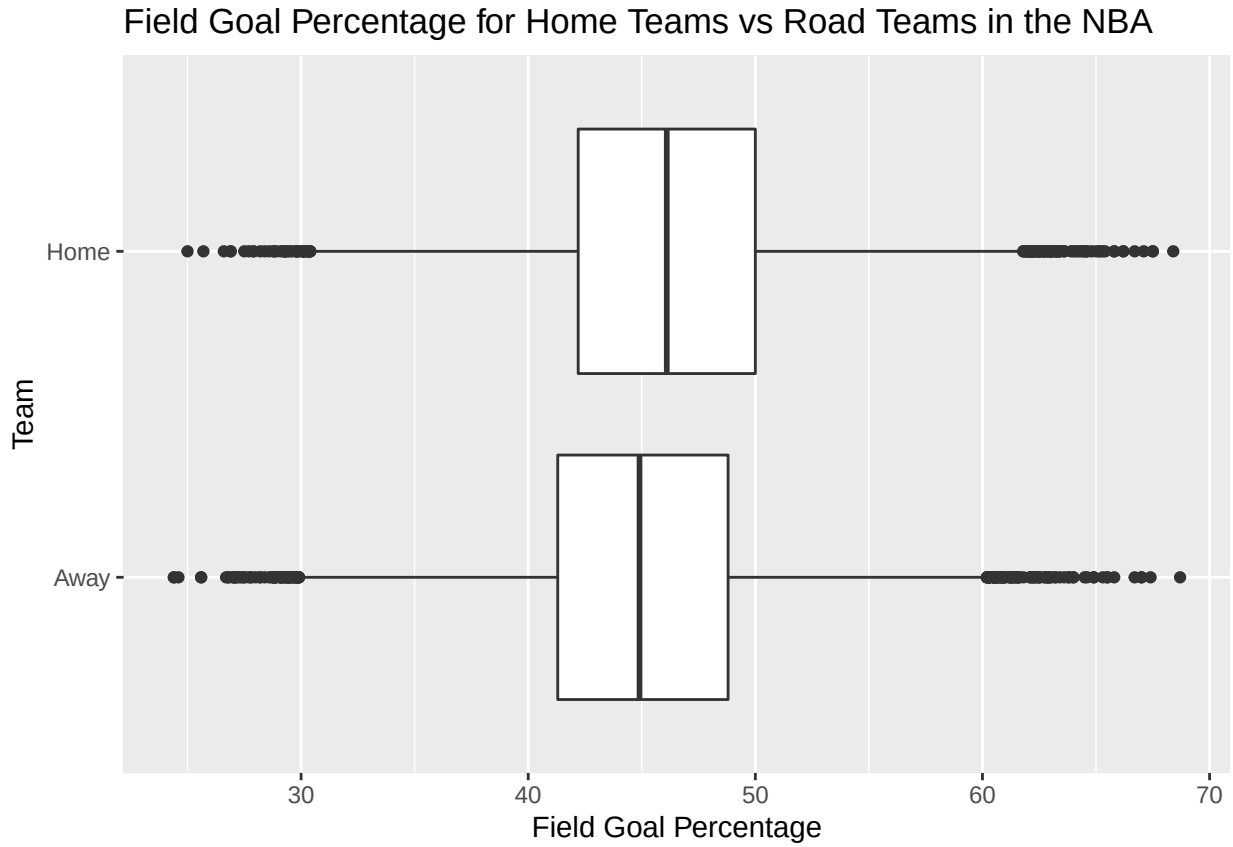



```
t.test(NBA$REB_home, NBA$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBA$REB_home and NBA$REB_away
## t = 22.334, df = 24090, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.177416 1.403962
## sample estimates:
## mean difference
##        1.290689
```

Field Goal Percentage

```
NBA %>%
  pivot_longer(c(FG_PCT_home, FG_PCT_away), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(x = "Field Goal Percentage", y
  ↪  = "Team", title = "Field Goal Percentage for Home Teams vs Road Teams in the NBA")
```

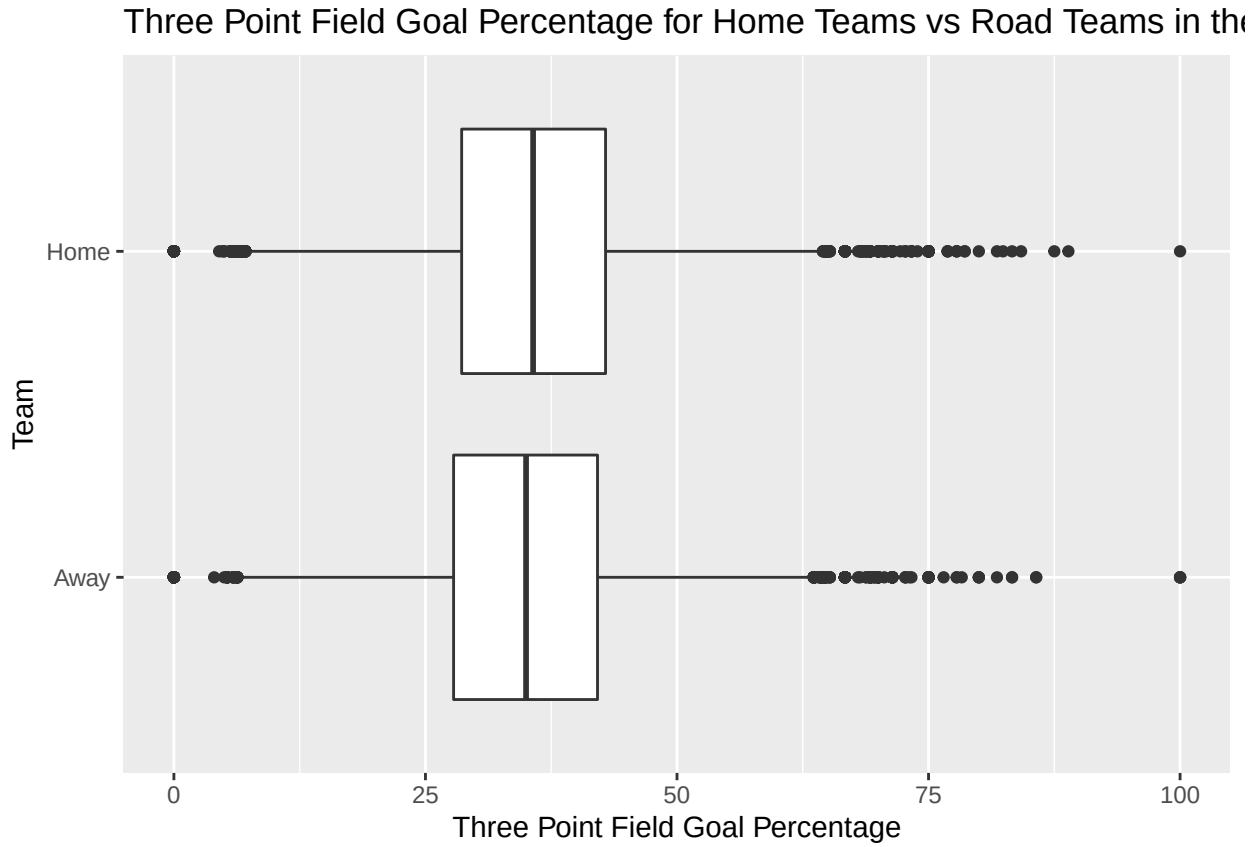



```
t.test(NBA$FG_PCT_home, NBA$FG_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBA$FG_PCT_home and NBA$FG_PCT_away
## t = 22.284, df = 24090, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.014673 1.210382
## sample estimates:
## mean difference
##        1.112527
```

Three Point Field Goal Percentage

```
NBA %>%
  pivot_longer(c(FG3_PCT_home, FG3_PCT_away), names_to = "Team", values_to =
  ↪  "PT3_Field_Goal_Percentage") %>%
  ggplot(aes(x = PT3_Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(x = "Three Point Field Goal
  ↪  Percentage", y = "Team", title = "Three Point Field Goal Percentage for Home Teams
  ↪  vs Road Teams in the NBA")
```




```
t.test(NBA$FG3_PCT_home, NBA$FG3_PCT_away, paired = TRUE)
```

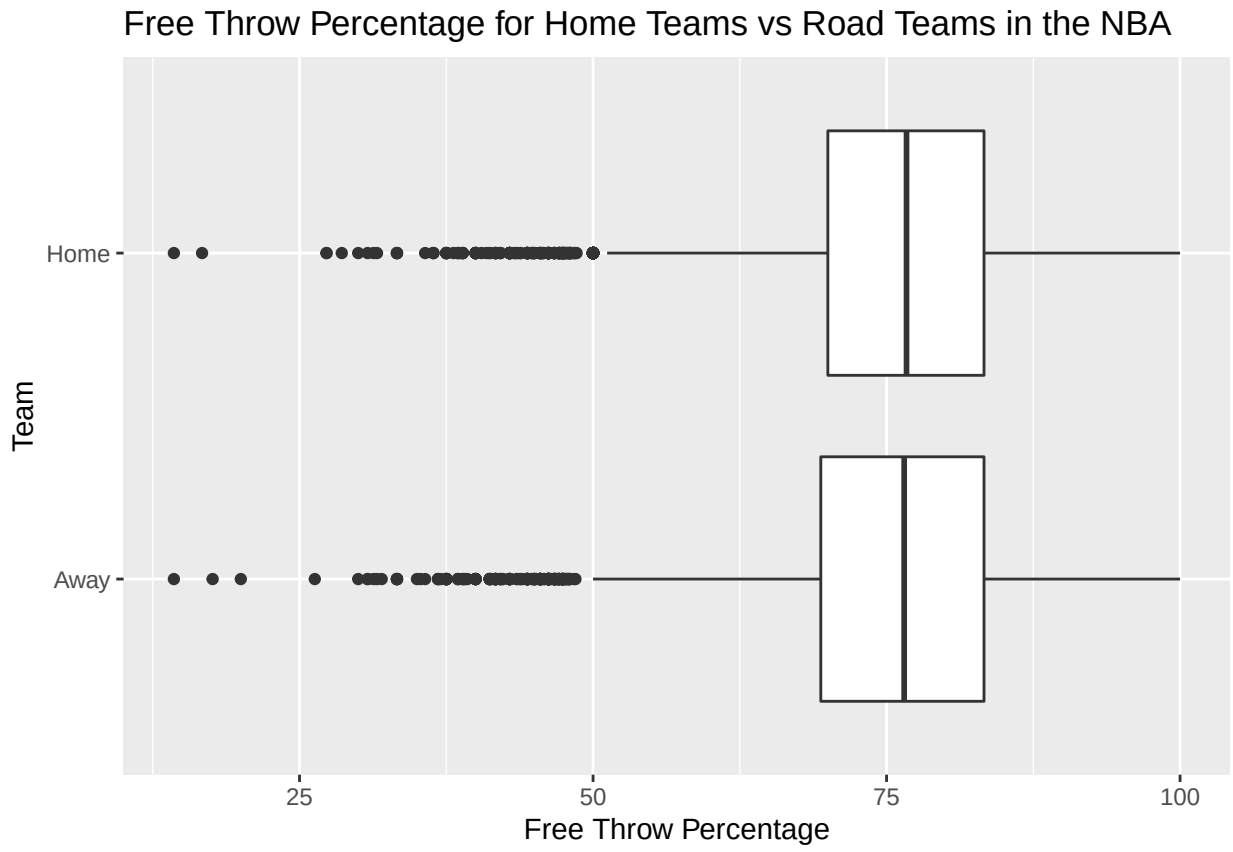
```
##
##  Paired t-test
##
## data:  NBA$FG3_PCT_home and NBA$FG3_PCT_away
## t = 6.4716, df = 24090, p-value = 9.881e-11
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.4556828 0.8516270
## sample estimates:
## mean difference
##       0.6536549
```

Similar to points, our box plots and paired T-test's suggest that home teams record more assists and rebounds than away teams and shoot a higher field goal percentage and three point field goal percentage as the first quartile, median, and third quartile are all higher for home teams as shown by the box plots and the T-tests generate a p-value less than 2.2 x 10^-16 for assists, rebounds, and field goal percentage, and a p-value of 9.881 x 10^-11 for three point percentage, which suggests a statistically significant result as these values are all very close to zero.

The differences are not significant, but it is clear that being the home team does warrant a slight advantage. Using the 95% confidence intervals generated by the T-tests, we are 95% confident that home teams record between 1.254216 and 1.425291 more assists, between 1.177416 and 1.403962 more rebounds, a field goal percentage between 1.014673% and 1.210382% higher, and a three point percentage between 0.4556828% 0.8516270% higher compared to away teams.

Free Throw Percentage

```
NBA %>%
  pivot_longer(c(FT_PCT_home, FT_PCT_away), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(x = "Free Throw Percentage", y
  ↪  = "Team", title = "Free Throw Percentage for Home Teams vs Road Teams in the NBA")
```




```
t.test(NBA$FT_PCT_home, NBA$FT_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBA$FT_PCT_home and NBA$FT_PCT_away
## t = 1.8627, df = 24090, p-value = 0.06252
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.008917128  0.350106784
## sample estimates:
## mean difference
##       0.1705948
```

Interestingly, there is not a statistically significant difference in free throw percentage between home and away teams. The box plots look like near mirror images of each other, and the T-test gives us a p-value of

0.0652 which is greater than the level of significance of 0.05, meaning there is not enough evidence to reject the null hypothesis that free throw percentage is equal between home and away teams. This result makes sense because the free throw is a very standard shot and is unopposed by defenses, hence why it's called a "free" throw, and many NBA players have practiced shooting so many free throws that the shot becomes muscle memory.

Overall, we can conclude that the home team generally scores more points, records more assists and rebounds, and shoots a higher field goal percentage and three point percentage than away teams. However, we do not see a statistically significant difference in free throw percentage between home and away teams. We can display this using bar charts.

Bar Charts Comparing General League Averages

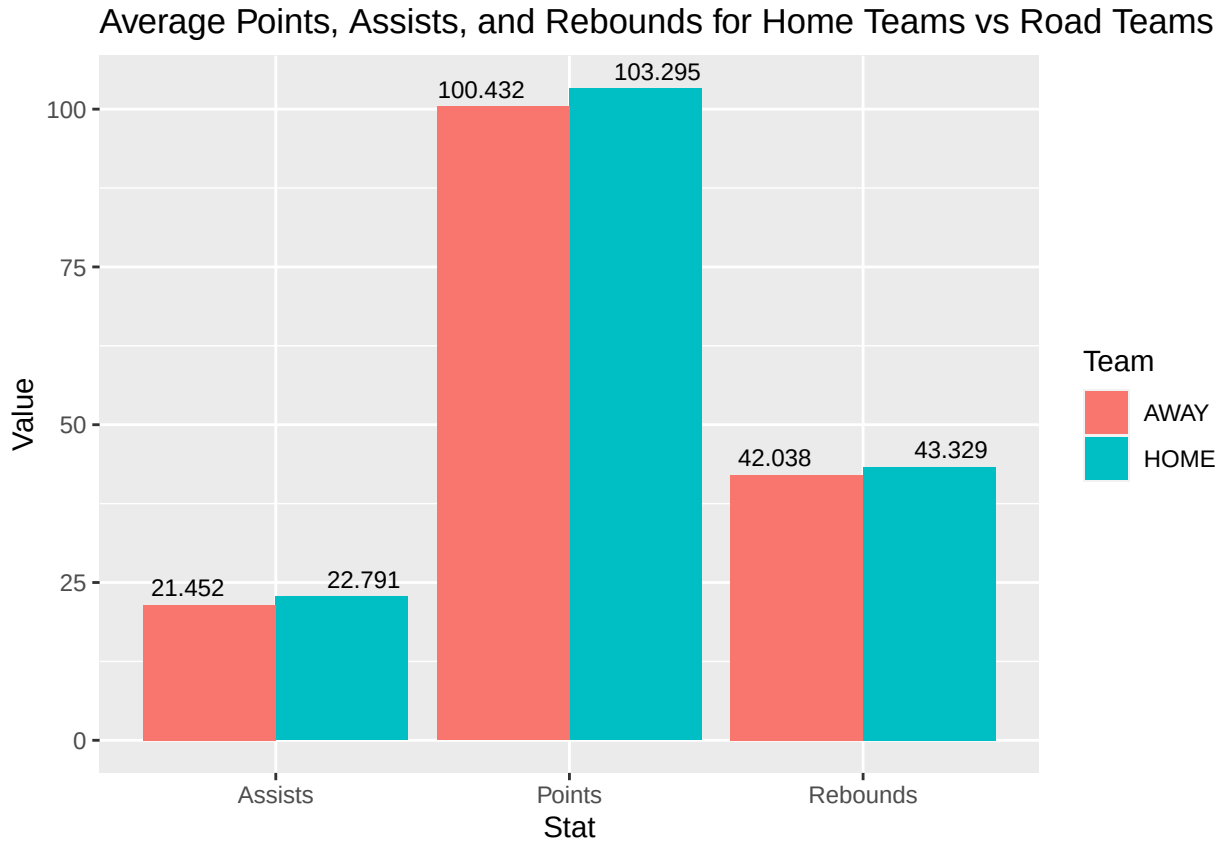
We will now use the NBA data and the summarize function to create a new NBA_SUMMARY dataset with all of the mean values for home and away teams in each statistic.

```
NBA_SUMMARY <- NBA %>%
  summarize(MEAN_AST_HOME = mean(AST_home), MEAN_AST_AWAY = mean(AST_away),
   ↪  MEAN_POINTS_HOME = mean(PTS_home), MEAN_POINTS_AWAY = mean(PTS_away), MEAN_REB_HOME
   ↪  = mean(REB_home), MEAN_REB_AWAY = mean(REB_away), MEAN_FGPCT_HOME =
   ↪  mean(FG_PCT_home), MEAN_FGPCT_AWAY = mean(FG_PCT_away), MEAN_FTPCT_HOME =
   ↪  mean(FT_PCT_home), MEAN_FTPCT_AWAY = mean(FT_PCT_away), MEAN_FG3PCT_HOME =
   ↪  mean(FG3_PCT_home), MEAN_FG3PCT_AWAY = mean(FG3_PCT_away))
```

Next, we will use the pivot_longer and select function to make the data tidy so that we can graph the data. We will make one bar chart to compare points, rebounds, and assists, and one bar chart to compare field goal percentage, three point percentage, and free throw percentage.

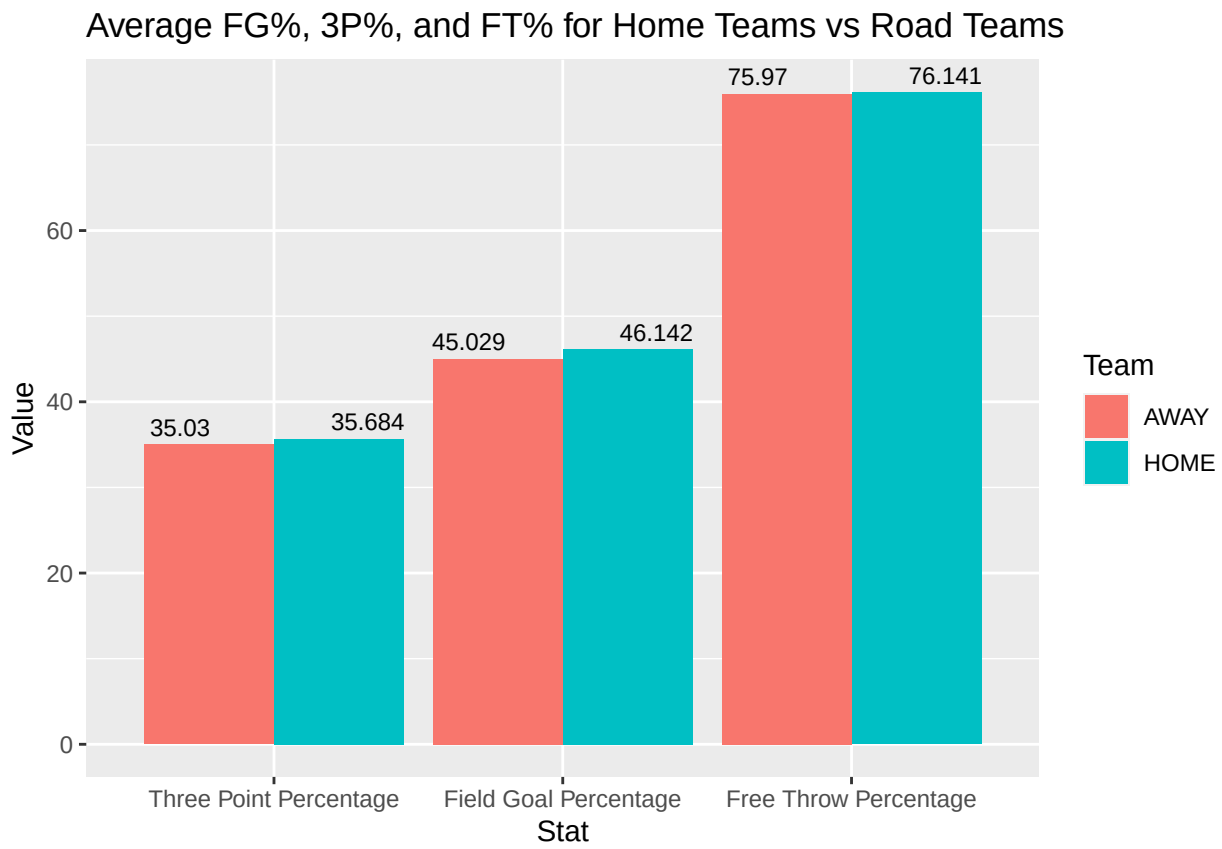
```
NBA_SUMMARY %>%
  pivot_longer(c(MEAN_AST_HOME,MEAN_AST_AWAY, MEAN_POINTS_HOME, MEAN_POINTS_AWAY,
   ↪  MEAN_REB_HOME, MEAN_REB_AWAY), names_to = c("Stat", "Team"), names_pattern =
   ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Team, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Team)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("Assists", "Points",
   ↪  "Rebounds")) + labs(title = "Average Points, Assists, and Rebounds for Home Teams
   ↪  vs Road Teams")
```

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




```
NBA_SUMMARY %>%
  pivot_longer(c(MEAN_FGPCT_HOME, MEAN_FGPCT_AWAY, MEAN_FTPCT_HOME, MEAN_FTPCT_AWAY,
  ↪  MEAN_FG3PCT_HOME, MEAN_FG3PCT_AWAY), names_to = c("Stat", "Team"), names_pattern =
  ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Team, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Team)) + geom_col(position = "dodge") +
  ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
  ↪  position_dodge(1.3)) + scale_x_discrete(labels = c("Three Point Percentage", "Field
  ↪  Goal Percentage", "Free Throw Percentage")) + labs(title = "Average FG%, 3P%, and
  ↪  FT% for Home Teams vs Road Teams")
```

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



Based on the bar charts above, we see that the values for each statistic is slightly higher for the home team than it is for the away team, except the difference for free throw percentage is extremely marginal. This shows us that there is a general home court advantage across the board for the main statistics recorded in the NBA except for free throw percentage.

Associations Between Variables

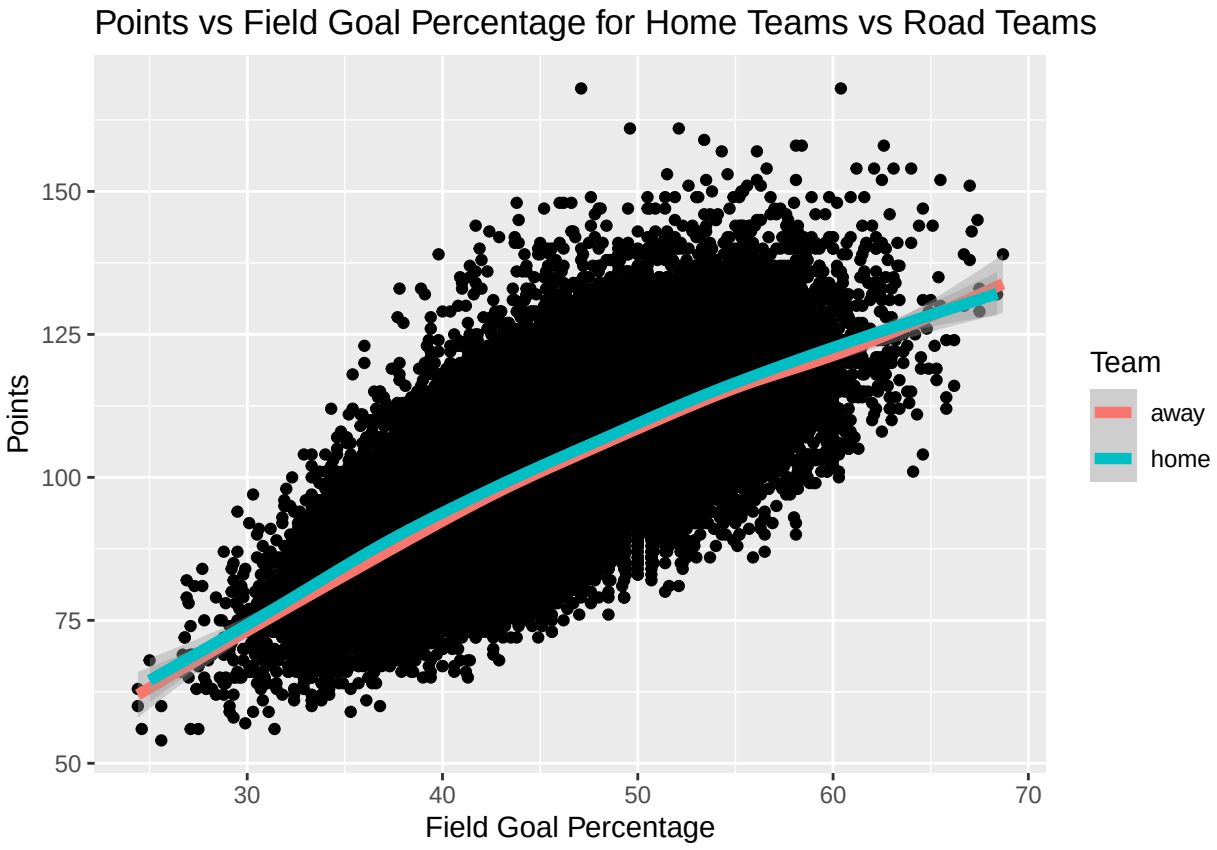
Now that we have concluded that there are statistically significant differences for home vs away teams in many of the different variables, we can analyze if there is any association between the variables. We will be creating scatter plots to compare the relationship between different variables for home vs away teams.

Points vs Field Goal Percentage

```
NBA %>%
 pivot_longer(c(PTS_home, PTS_away, FG_PCT_home, FG_PCT_away), names_to = c("stat",
  ↪  "name"), names_pattern = "(.*)(....)$", values_to = "Value") %>%
  pivot_wider(names_from = "stat", values_from = "Value", names_repair = "check_unique")
   ↪  %>%
  unnest(PTS_, FG_PCT_) %>%
  ggplot(aes(x = FG_PCT_, y = PTS_)) + geom_point() + geom_smooth(aes(color = name),
   ↪  size=2) + guides(color = guide_legend(title = "Team")) + labs(x = "Field Goal
   ↪  Percentage", y = "Points", title = "Points vs Field Goal Percentage for Home Teams
   ↪  vs Road Teams")
```

```
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(PTS_, FG_PCT_))`, with `mutate()` if needed
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



It makes sense that as teams make a higher percentage of their field goals, they score more points. Both

home and away teams show the same trend and are very similar, but the home trend line appears just slightly higher than the away trend line, which may suggest that they are scoring more three pointers or free throws which makes them score more points despite having the same field goal percentage. Since we previously concluded that home teams have an advantage when it comes to field goal percentage, it makes sense that they will generally score more points and have an advantage in this aspect too.

```
lm(PTS_home ~ FG_PCT_home, data = NBA)
```

```
##
## Call:
## lm(formula = PTS_home ~ FG_PCT_home, data = NBA)
##
## Coefficients:
## (Intercept)  FG_PCT_home
##      31.883        1.548
```

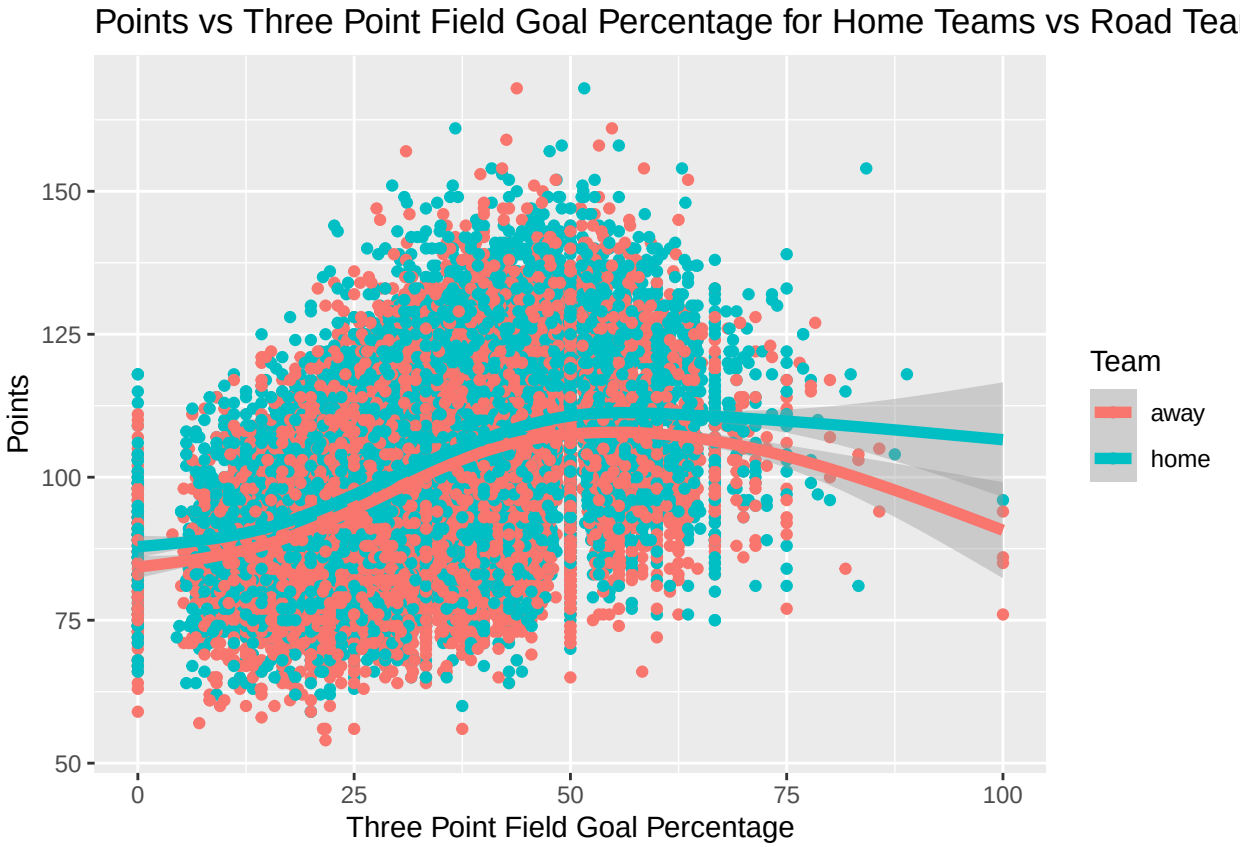
Based on the linear regression model, each 1% increase in field goal percentage for home teams corresponds to a 1.5477 point increase. Since our earlier matched pairs T-test found that the mean difference between home field goal percentage and away field goal percentage is 0.01112527, this model suggests that since home teams generally shoot 1.112527% better, they will score about 1.722 more points.

Points vs Three Point Shooting Percentage

```
NBA %>%
 pivot_longer(c(PTS_home, PTS_away, FG3_PCT_home, FG3_PCT_away), names_to = c("stat",
 ↪  "name"), names_pattern = "(.*)(....)$", values_to = "Value") %>%
  pivot_wider(names_from = "stat", values_from = "Value", names_repair = "check_unique")
  ↪  %>%
  unnest(PTS_, FG3_PCT_) %>%
  ggplot(aes(x = FG3_PCT_, y = PTS_, color = name)) + geom_point() + geom_smooth(size =
  ↪  2) + guides(color = guide_legend(title = "Team")) + labs(x = "Three Point Field
  ↪  Goal Percentage", y = "Points", title = "Points vs Three Point Field Goal
  ↪  Percentage for Home Teams vs Road Teams")
```

```
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(PTS_, FG3_PCT_))`, with `mutate()` if needed
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

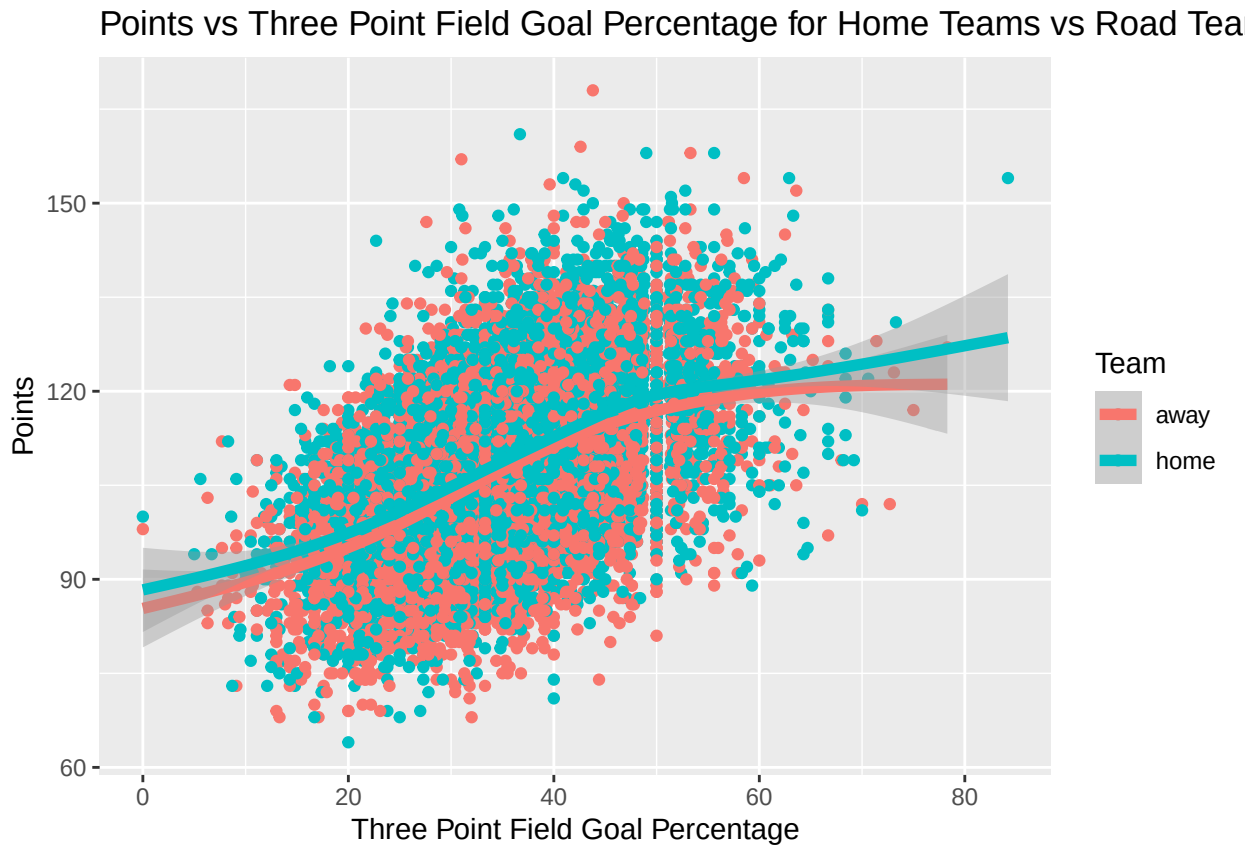


This graph interestingly shows that a lot of teams either shot 0% or 100% from three and these results skew the data. However, this is because the volume of three point attempts was much lower in the 2000s and early 2010s compared to the modern NBA game. On a lower volume of three point attempts, it is more likely for a team to record an unusual three point percentage like 0% or 100%. For looking at the relationship between three point field goal percentage and points, we will isolate the data to just the 2015-2016 season and after, because this is the point when the Warriors and their three-point reliant, small-ball lineups began to dominate the league and the volume of three-point shooting as a whole began to explode.

```
NBA %>%
  filter(SEASON >= 2015) %>%
 pivot_longer(c(PTS_home, PTS_away, FG3_PCT_home, FG3_PCT_away), names_to = c("stat",
 ↪  "name"), names_pattern = "(.*)(....)$", values_to = "Value") %>%
  pivot_wider(names_from = "stat", values_from = "Value", names_repair = "check_unique")
  ↪  %>%
  unnest(PTS_, FG3_PCT_) %>%
  ggplot(aes(x = FG3_PCT_, y = PTS_, color = name)) + geom_point() + geom_smooth(size =
  ↪  2) + guides(color = guide_legend(title = "Team")) + labs(x = "Three Point Field
  ↪  Goal Percentage", y = "Points", title = "Points vs Three Point Field Goal
  ↪  Percentage for Home Teams vs Road Teams")
```

```
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(PTS_, FG3_PCT_))`, with `mutate()` if needed
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



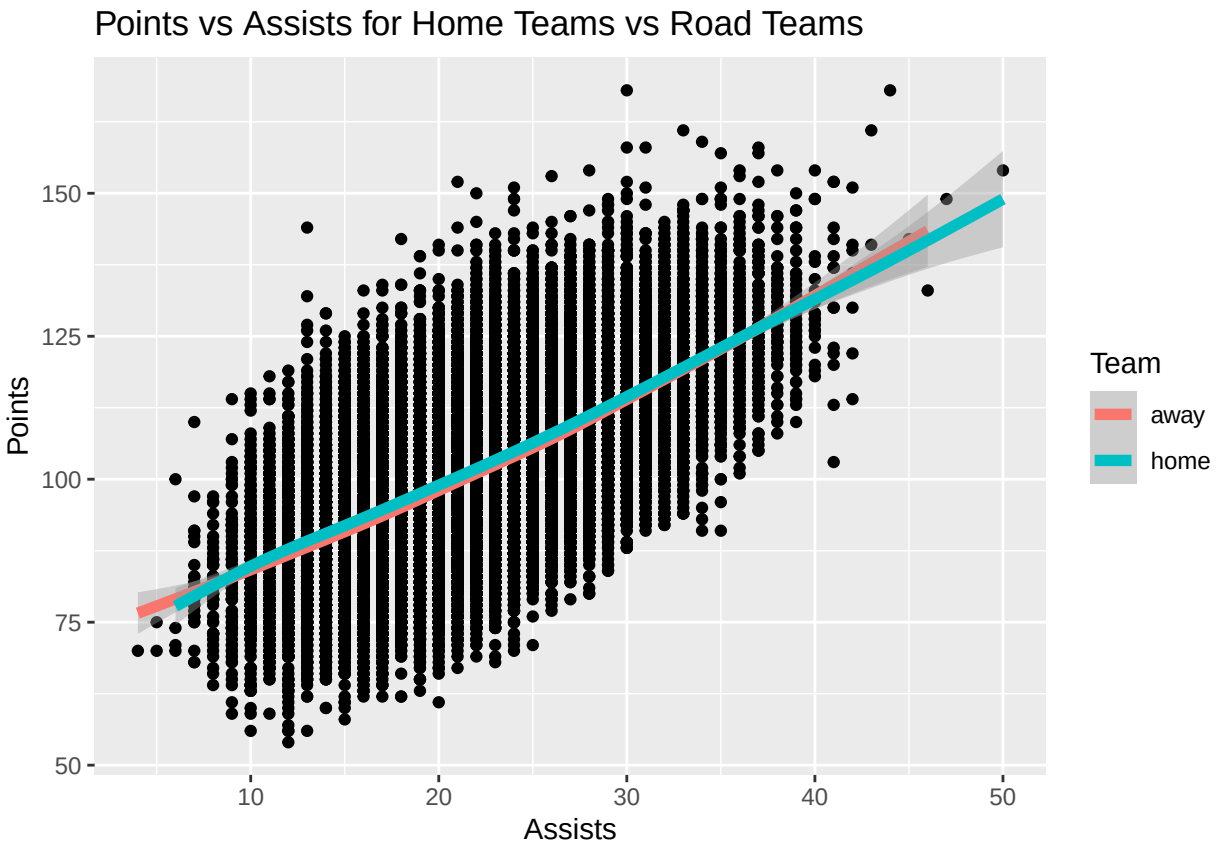
This graph shows a much clearer trend that as three point field goal percentage increases, points also increase, and the trend line for home teams is consistently above the away team which is interesting because even shooting the same three point percentage, they are scoring slightly more points. There could be many reasons for this; for example, home teams may score more two point field goals or free throws, or they may be attempting more shots in the game. Since our previous results concluded that home teams have an advantage when it comes to three point field goal percentage, it makes sense that they generally score more points as well.

Points vs Assists

```
NBA %>%
 pivot_longer(c(PTS_home, PTS_away, AST_home, AST_away), names_to = c("stat", "name"),
  ↪  names_pattern = "(.*)(....)$", values_to = "Value") %>%
  pivot_wider(names_from = "stat", values_from = "Value", names_repair = "check_unique")
   ↪  %>%
  unnest(PTS_, AST_) %>%
  ggplot(aes(x = AST_, y = PTS_)) + geom_point() + geom_smooth(aes(color = name), size =
   ↪  2) + guides(color = guide_legend(title = "Team")) + labs(x = "Assists", y =
   ↪  "Points", title = "Points vs Assists for Home Teams vs Road Teams")
```

```
## Warning: unnest() has a new interface. See ?unnest for details.
## Try `df %>% unnest(c(PTS_, AST_))`, with `mutate()` if needed
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



We see that when a team gets more assists, they generally score more points. This graph features vertical bands unlike the previous two graphs because assists can only be recorded as integers, where as field goal percentages are proportions. Because assists can only be recorded as a result of made field goals, this association makes sense since more assists means more field goals made, which results in more points scored. However, there are many instances where a team can score without an assist. For example, a player could dribble down the floor by himself and score a basket without a teammate assisting him, or a player could get fouled and score free throws, which cannot be assisted on. Thus, the results of the graph may suggest that teams with better ball movement and more assists are more productive on offense. Home and away teams feature the same trend and the lines of best fit are approximately the same. Since we previously concluded that home teams record more assists than away teams, it makes sense that home teams also score more points.

```
lm(PTS_home ~ AST_home, data = NBA)
```

```
##
## Call:
## lm(formula = PTS_home ~ AST_home, data = NBA)
##
## Coefficients:
## (Intercept)     AST_home
##      68.610        1.522
```

Our earlier matched pairs T-test for assists calculated that the mean difference between home and away assists is 1.339753, and since the linear regression model gives us a slope of 1.522, we get that home teams will score 2.039 points more because of their advantage in assists.

Exploring Factors that Influence Home Court Advantage

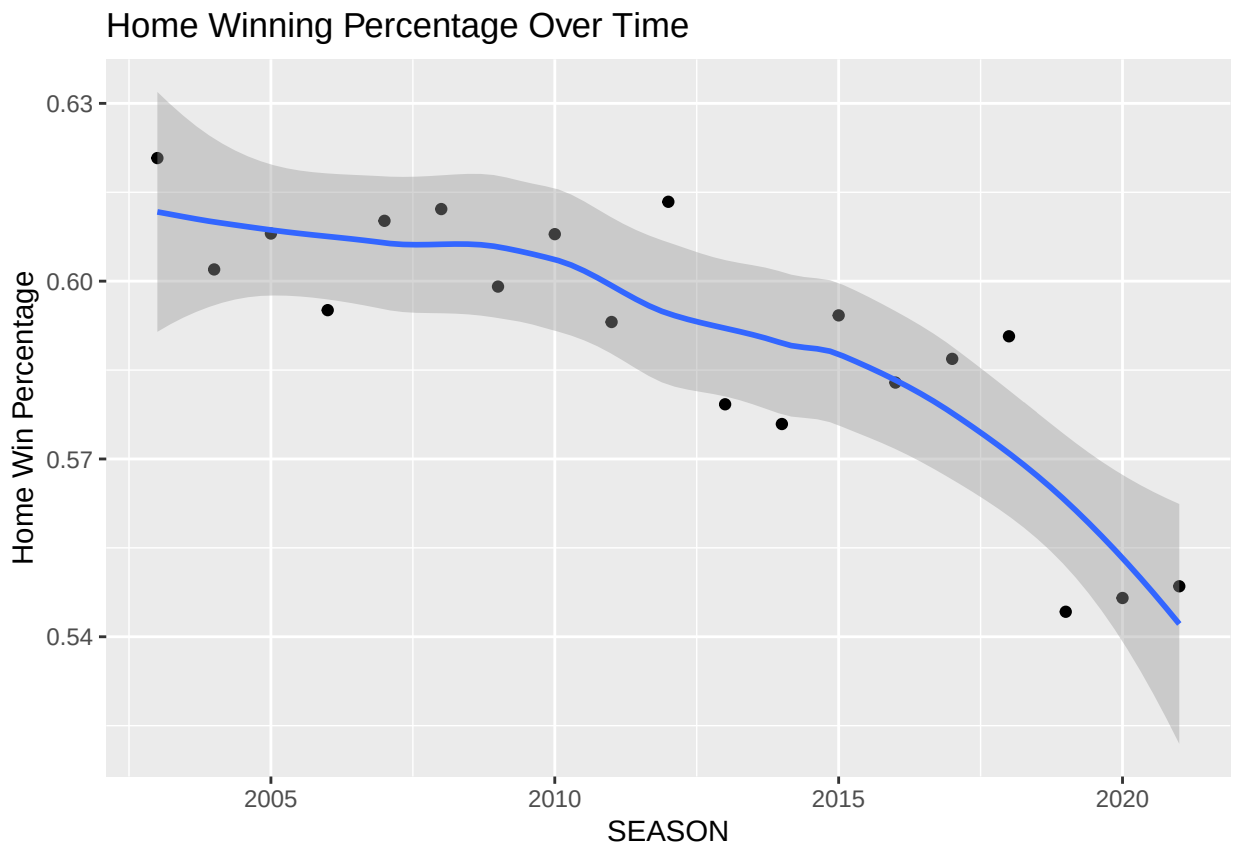
After determining that home court advantage does exist, we want to explore the factors that give home teams an advantage compared to away teams.

Home Team Win Percentage Over Time

We can group the data by season and find the average home winning percentage among all teams for each season from 2003/2004 to 2021/2022 using the summarize function. We can then display the change in home winning percentage over time using a scatterplot.

```
NBA %>%
  group_by(SEASON) %>%
  summarize(home_wpct = mean(HOME_TEAM_WINS)) %>%
  ggplot(aes(x = SEASON, y = home_wpct)) + geom_point() + geom_smooth() + labs(title =
  ↪  "Home Winning Percentage Over Time", y = "Home Win Percentage")
```

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



In the seasons 2019, 2020, and 2021, there is a sharp decline in home winning percentage. This can be attributed to the pandemic, which shut down the NBA midway through the 2019-2020 season and caused a number of disruptions including an NBA bubble and limited fan capacity.

Disney Bubble

During the 2020 COVID-19 pandemic, the NBA created a biosecure bubble at the Walt Disney World in Bay Lake, Florida. The bubble which is referred to as the Disney Bubble or the Orlando Bubble was effective in protecting players from the COVID-19 pandemic during the last eight games of the 2019-2020 regular season and throughout the 2020 NBA playoffs. Out of the 30 NBA teams, only 22 that were within reach of the playoffs were able to participate in the games held within the bubble. A team was determined to be in reach of the playoffs if they were in the playoff picture or within six games of the eighth seed. The other eight teams, known as the Deleted Eight, were not able to participate in the bubble. The 22 teams finished the regular season by playing a series of games and the bottom 6 teams were eliminated, leaving 16 teams for the playoffs. The bubble was a drastic change to players and impacted the way they played and trained. We will be analyzing how the bubble impacted home court advantage because all games were played on a neutral court, with no fans allowed physically in the stadium. Because one team is technically considered home and the other is considered away, we will investigate if a team being denoted as 'home' actually gets an advantage, despite playing on the same court as everyone else.

We will first use the NBA dataset we made and filter the dates of the Disney Bubble in order to make a new dataset titled 'Bubble'. This will contain all of the data from the Bubble.

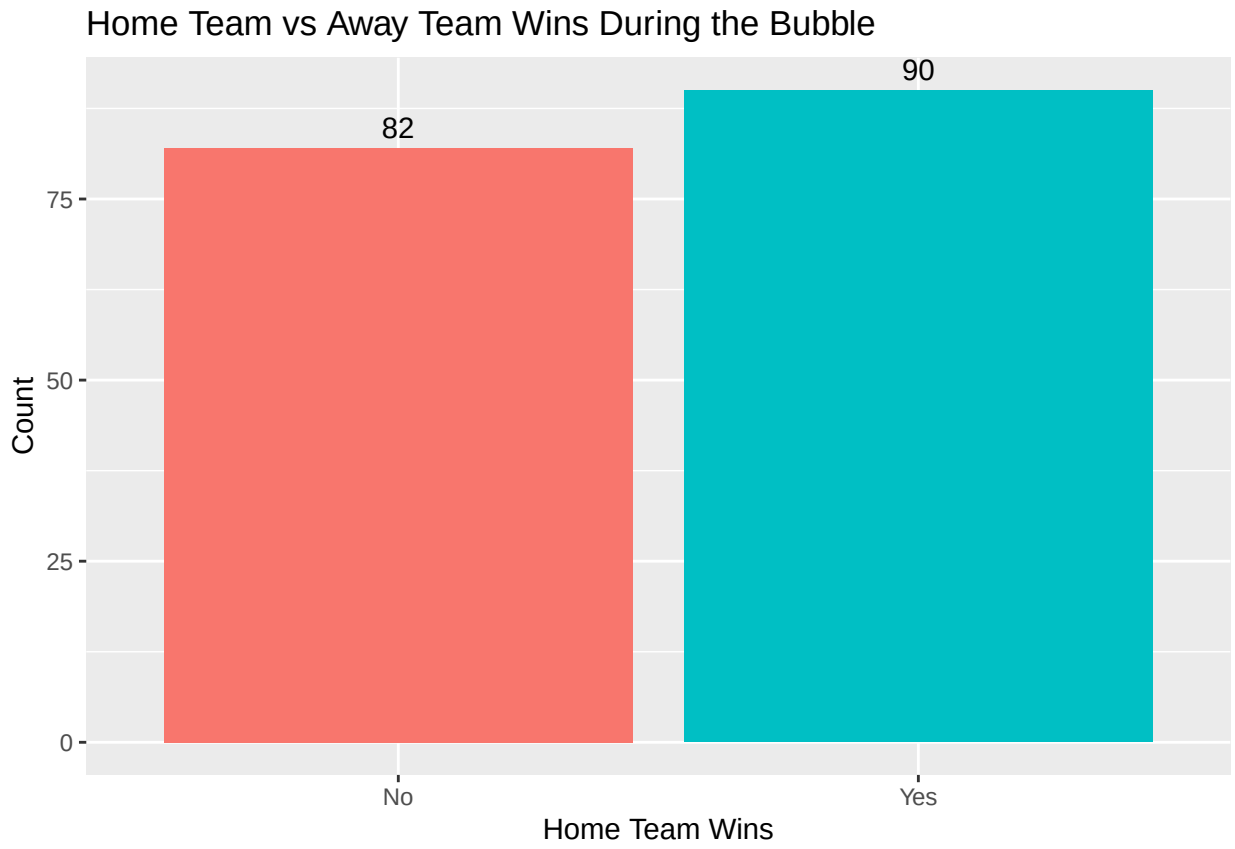
```
Bubble <- NBA %>%
  filter(GAME_DATE_EST >= "2020-07-30" & GAME_DATE_EST <= "2020-10-11")
```

Using this data, we can analyze if the "home" team actually had a higher win percentage compared to "away" teams despite the neutral settings.

```
Bubble %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0    82
## 2              1    90
```

```
Home_Team_Wins_Bubble <- c("No", "Yes")
Number_Bubble <- c(82, 90)
Home_wins_Bubble <- data.frame(Home_Team_Wins_Bubble, Number_Bubble)

Home_wins_Bubble %>%
  ggplot(aes(x = Home_Team_Wins_Bubble, y = Number_Bubble)) + geom_col(aes(fill =
  ↪  Home_Team_Wins_Bubble)) + geom_text(aes(label = Number_Bubble), vjust = -0.5) +
  ↪  labs(x = "Home Team Wins", y = "Count", title = "Home Team vs Away Team Wins During
  ↪  the Bubble") + theme(legend.position = "none")
```




```
prop.test(90, 172, p = 0.5)
```

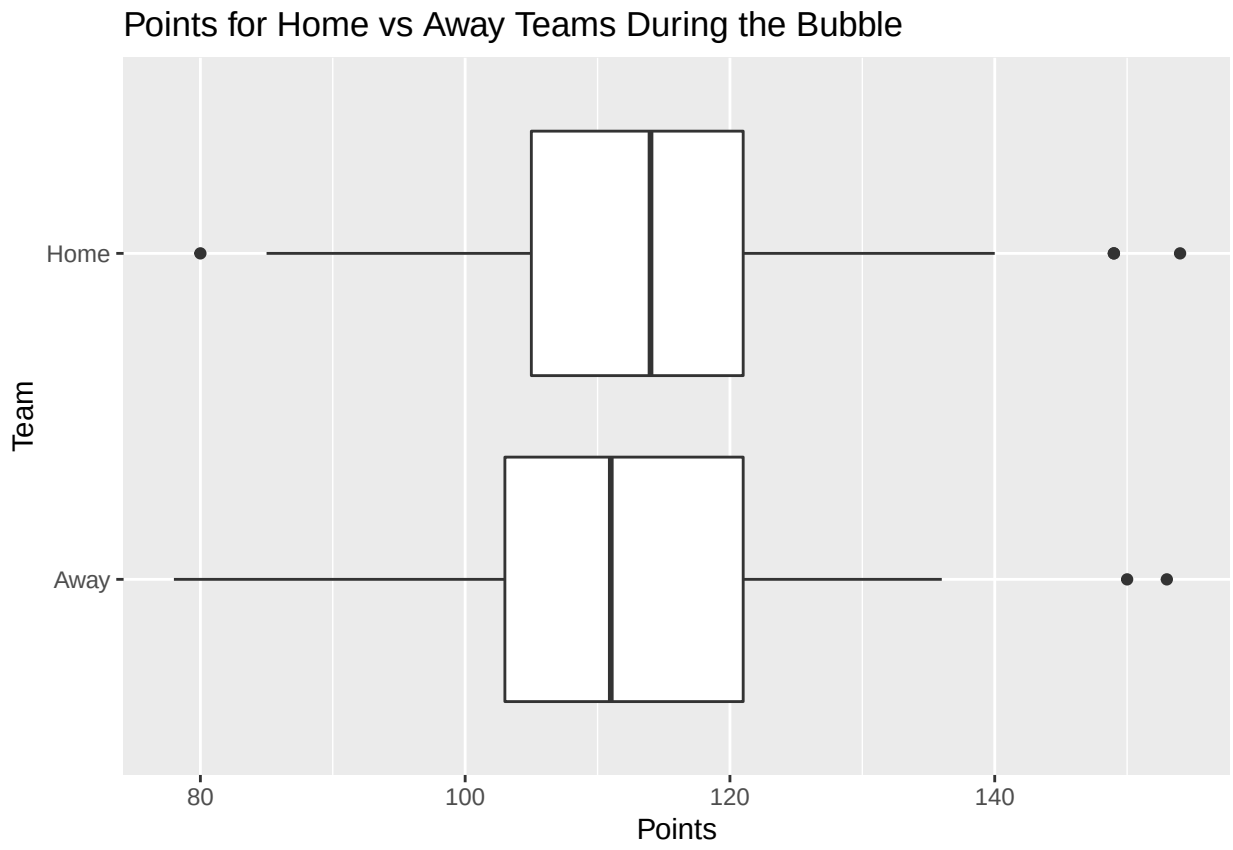
```
##
##  1-sample proportions test with continuity correction
##
## data:  90 out of 172, null probability 0.5
## X-squared = 0.28488, df = 1, p-value = 0.5935
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4460620 0.5993944
## sample estimates:
##         p
## 0.5232558
```

Although the home team has won a higher proportion of games, the proportion-test calculates a p-value of 0.5935, which is well above the 0.05 level of significance, so we cannot conclude that there is truly a difference between the win proportion of home and away teams during the Bubble.

We will now make boxplots for all of the statistics (such as points, field goal percentage, etc) to compare the home and away values. We will also run t-tests in order to determine if the differences are statistically significant.

Points

```
Bubble %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Points for Home vs Away Teams During the
  ↪  Bubble")
```

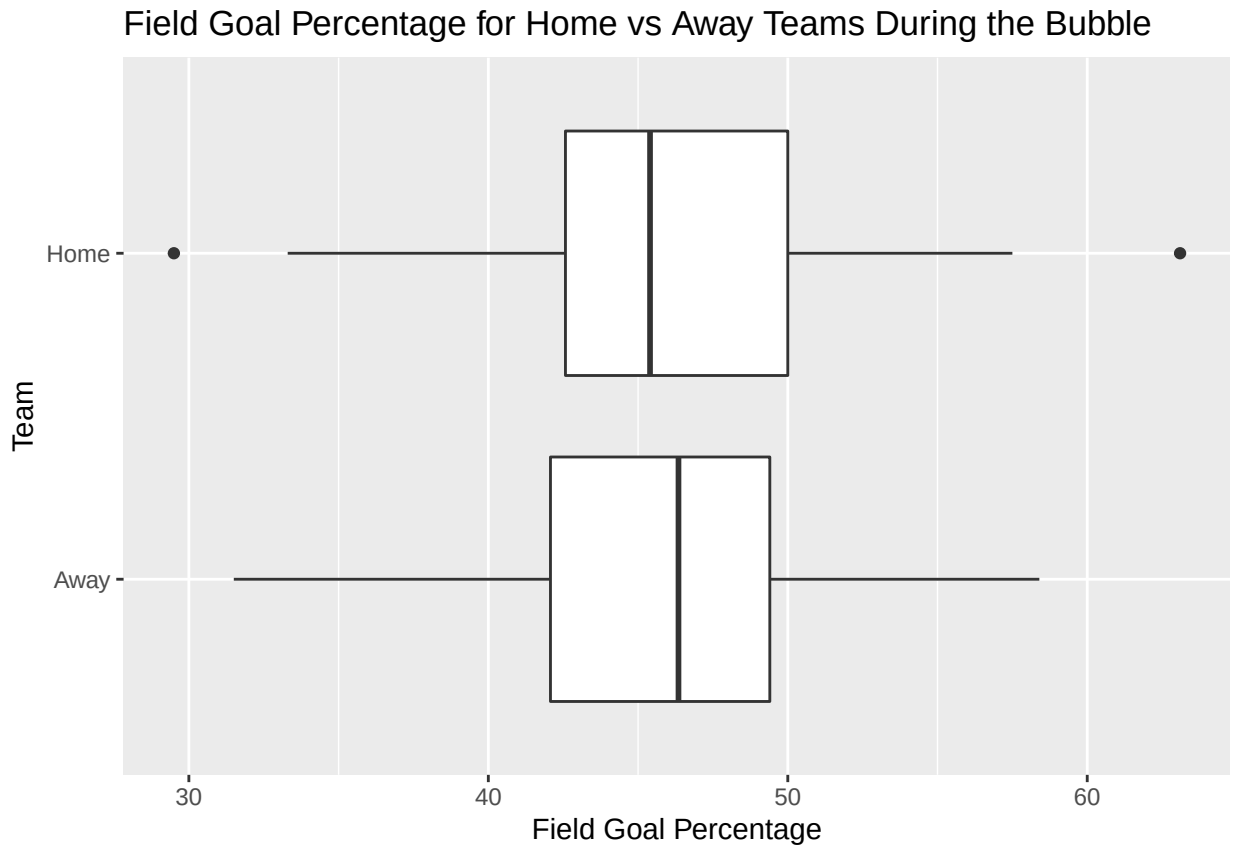



```
t.test(Bubble$PTS_home, Bubble$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$PTS_home and Bubble$PTS_away
## t = 1.321, df = 171, p-value = 0.1883
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.6666412  3.3643156
## sample estimates:
## mean difference
##        1.348837
```

Field Goal Percentage

```
Bubble %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Field Goal Percentage
  ↪  for Home vs Away Teams During the Bubble", x = "Field Goal Percentage")
```

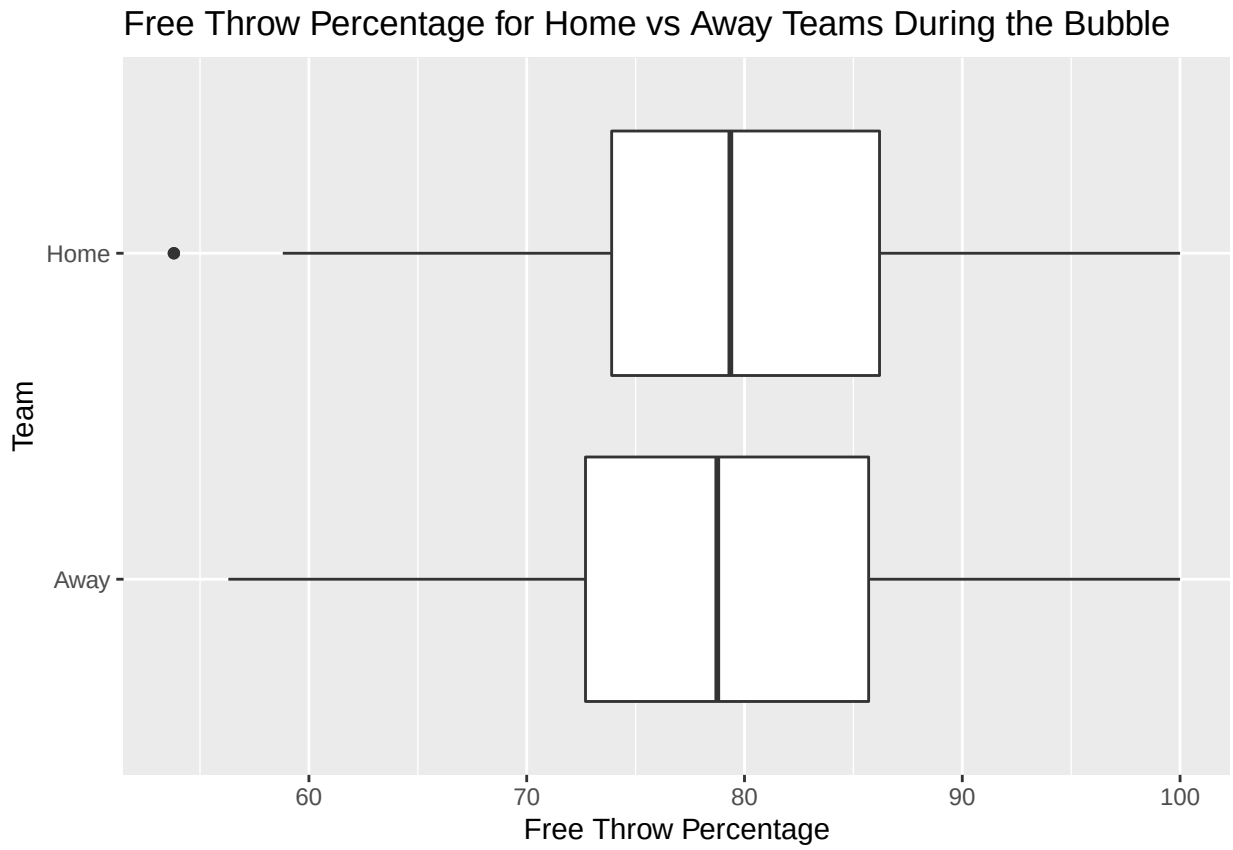



```
t.test(Bubble$FG_PCT_home, Bubble$FG_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$FG_PCT_home and Bubble$FG_PCT_away
## t = 0.35508, df = 171, p-value = 0.723
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.8853165  1.2736886
## sample estimates:
## mean difference
##        0.194186
```

Free Throw Percentage

```
Bubble %>%
  pivot_longer(c(FT_PCT_away, FT_PCT_home), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot()+
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Free Throw Percentage
  ↪  for Home vs Away Teams During the Bubble", x = "Free Throw Percentage")
```

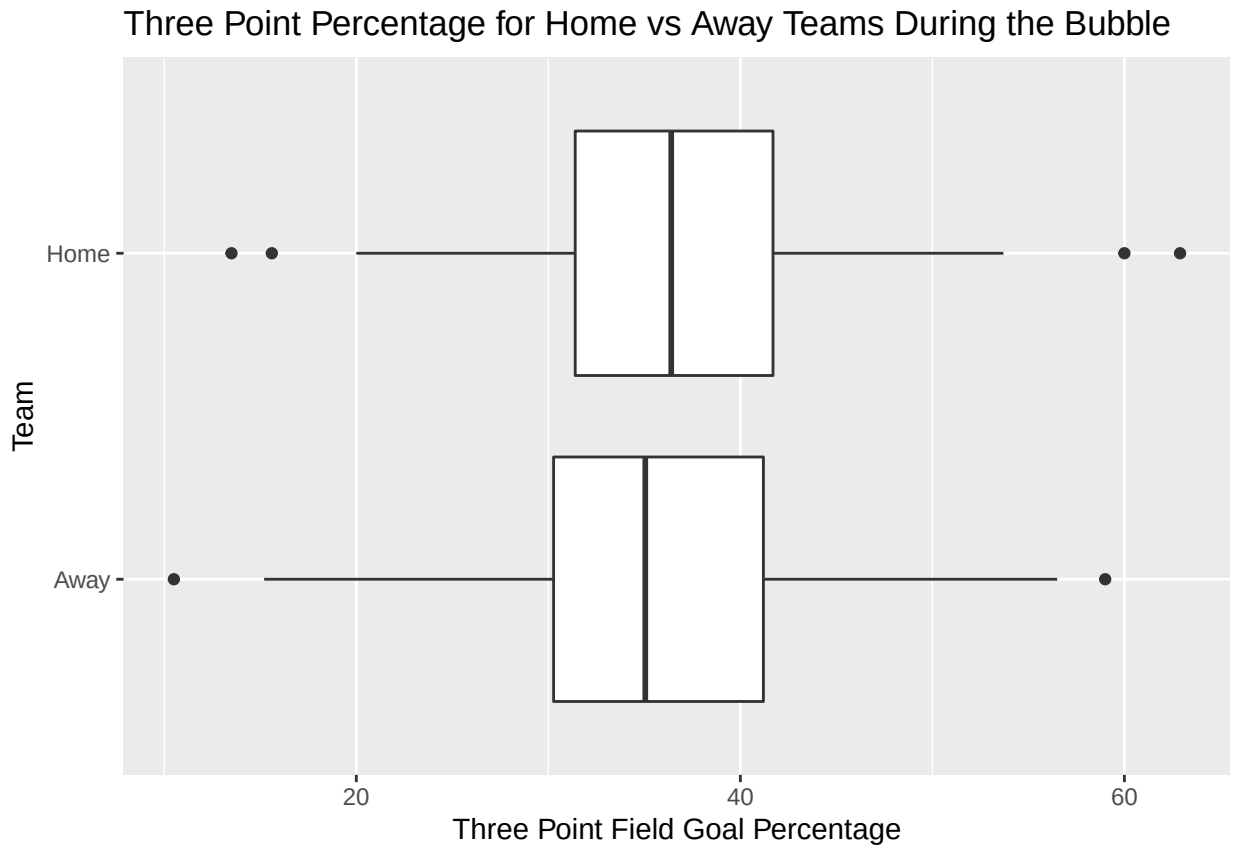



```
t.test(Bubble$FT_PCT_home, Bubble$FT_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$FT_PCT_home and Bubble$FT_PCT_away
## t = 0.58732, df = 171, p-value = 0.5578
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.401439  2.588648
## sample estimates:
## mean difference
##       0.5936047
```

Three Point Field Goal Percentage

```
Bubble %>%
  pivot_longer(c(FG3_PCT_away, FG3_PCT_home), names_to = "Team", values_to =
  ↪  "Three_Point_Percentage") %>%
  ggplot(aes(x = Three_Point_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Three Point Percentage
  ↪  for Home vs Away Teams During the Bubble", x = "Three Point Field Goal Percentage")
```

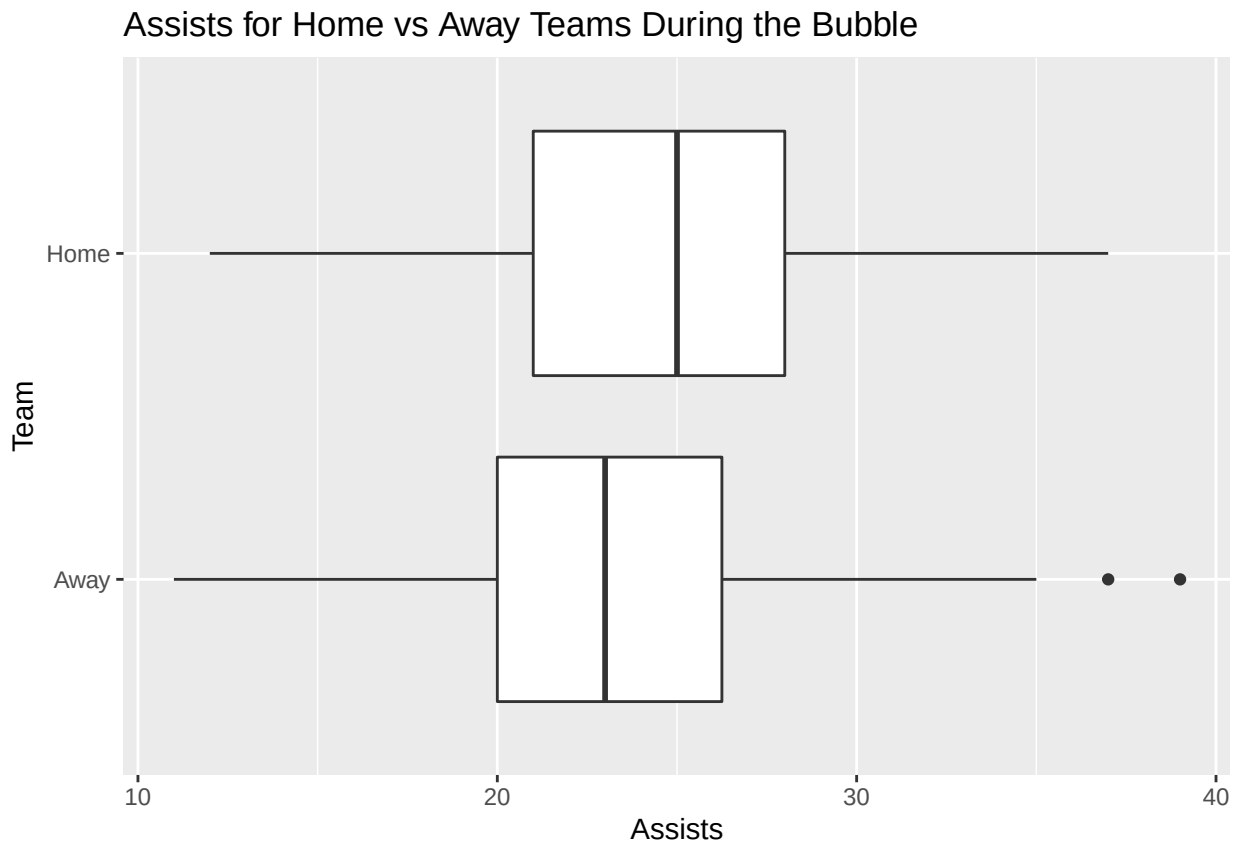



```
t.test(Bubble$FG3_PCT_home, Bubble$FG3_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$FG3_PCT_home and Bubble$FG3_PCT_away
## t = 1.2364, df = 171, p-value = 0.218
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.6316017  2.7490436
## sample estimates:
## mean difference
##        1.058721
```

Assists

```
Bubble %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Teams During the
  ↪  Bubble")
```

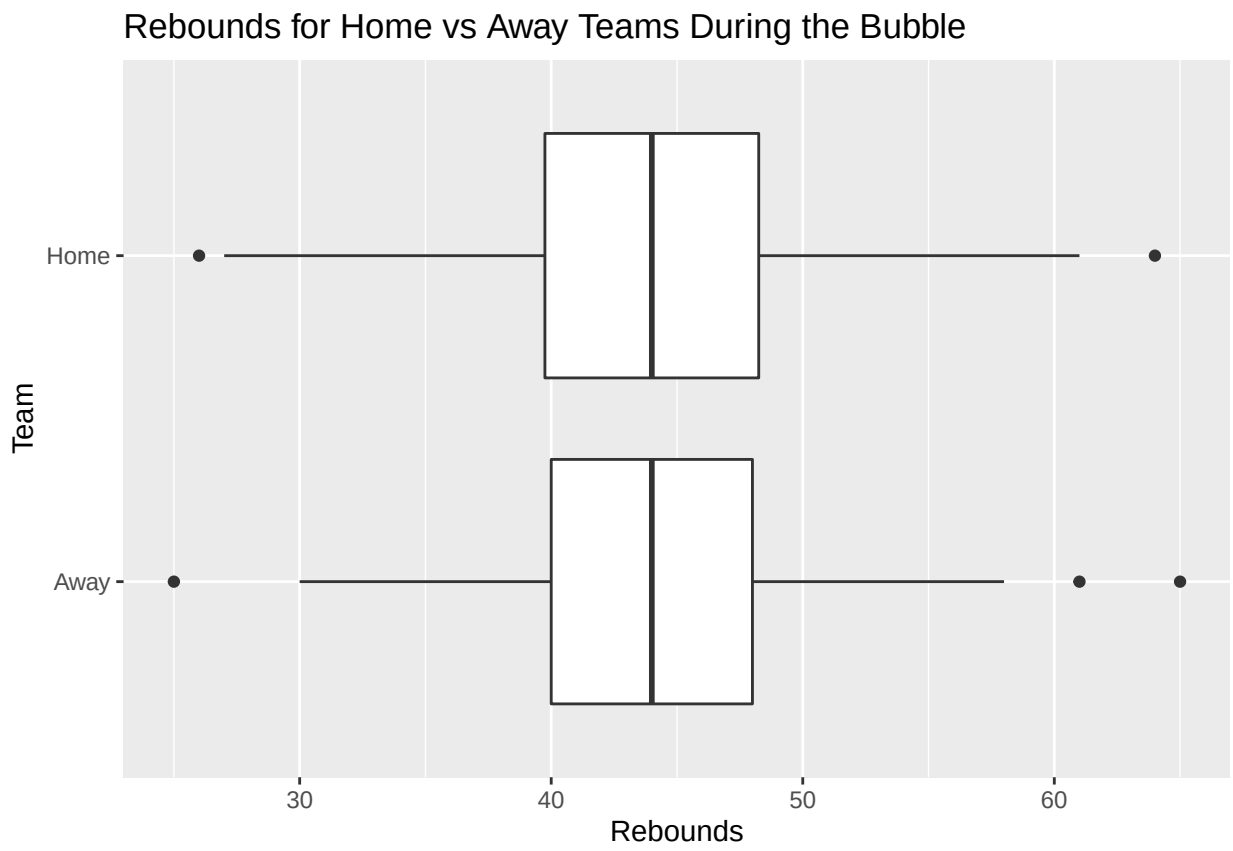



```
t.test(Bubble$AST_home, Bubble$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$AST_home and Bubble$AST_away
## t = 1.8959, df = 171, p-value = 0.05966
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.03662701  1.81569678
## sample estimates:
## mean difference
##       0.8895349
```

Rebounds

```
Bubble %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Teams During the
  ↪  Bubble")
```




```
t.test(Bubble$REB_home, Bubble$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Bubble$REB_home and Bubble$REB_away
## t = 0.46312, df = 171, p-value = 0.6439
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.081092  1.743882
## sample estimates:
## mean difference
##       0.3313953
```

Based on all of the boxplots above, we can see that the home team generally scores more points, has better field goal, free throw, and three-point percentages, and records more rebounds and assists. However, the

p-value for each of the tests are greater than 0.05 meaning that the differences are not statistically significant. We can display the differences between home and away teams in each of the statistics using bar charts.

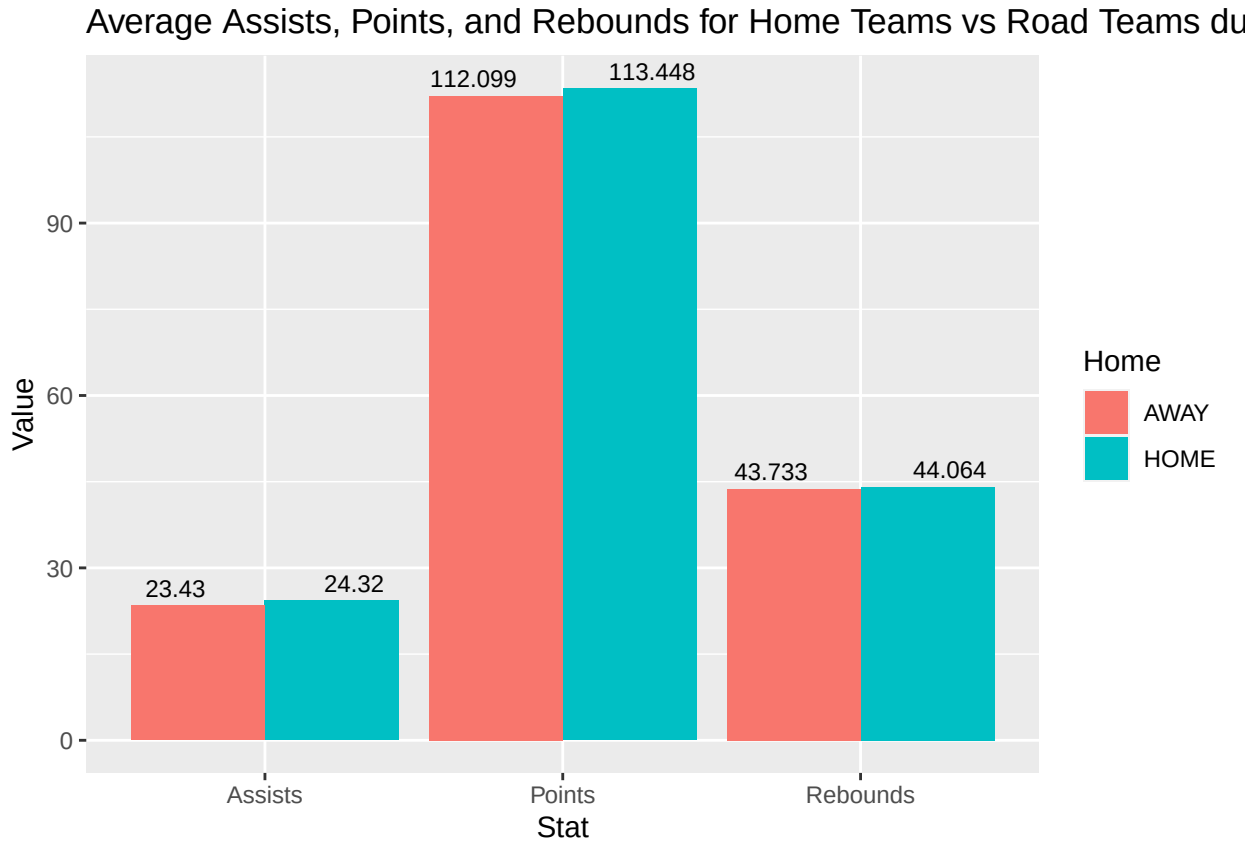
In order to make the barcharts, we will first summarize all of the statistics and calculate the mean values. We will do this by creating a new dataset called 'Bubble_summary' using the summarize function on the 'Bubble' dataset.

```
Bubble_summary <- Bubble %>%
  summarize(MEAN_AST_HOME = mean(AST_home), MEAN_AST_AWAY = mean(AST_away),
   ↪  MEAN_POINTS_HOME = mean(PTS_home), MEAN_POINTS_AWAY = mean(PTS_away), MEAN_REB_HOME
   ↪  = mean(REB_home), MEAN_REB_AWAY = mean(REB_away), MEAN_FGPCT_HOME =
   ↪  mean(FG_PCT_home), MEAN_FGPCT_AWAY = mean(FG_PCT_away), MEAN_FTPCT_HOME =
   ↪  mean(FT_PCT_home), MEAN_FTPCT_AWAY = mean(FT_PCT_away), MEAN_FG3PCT_HOME =
   ↪  mean(FG3_PCT_home), MEAN_FG3PCT_AWAY = mean(FG3_PCT_away))
```

Now that we have the new summarized dataset with all of the averages, we will now use the pivot longer function in order to make the data tidy and create the bar charts.

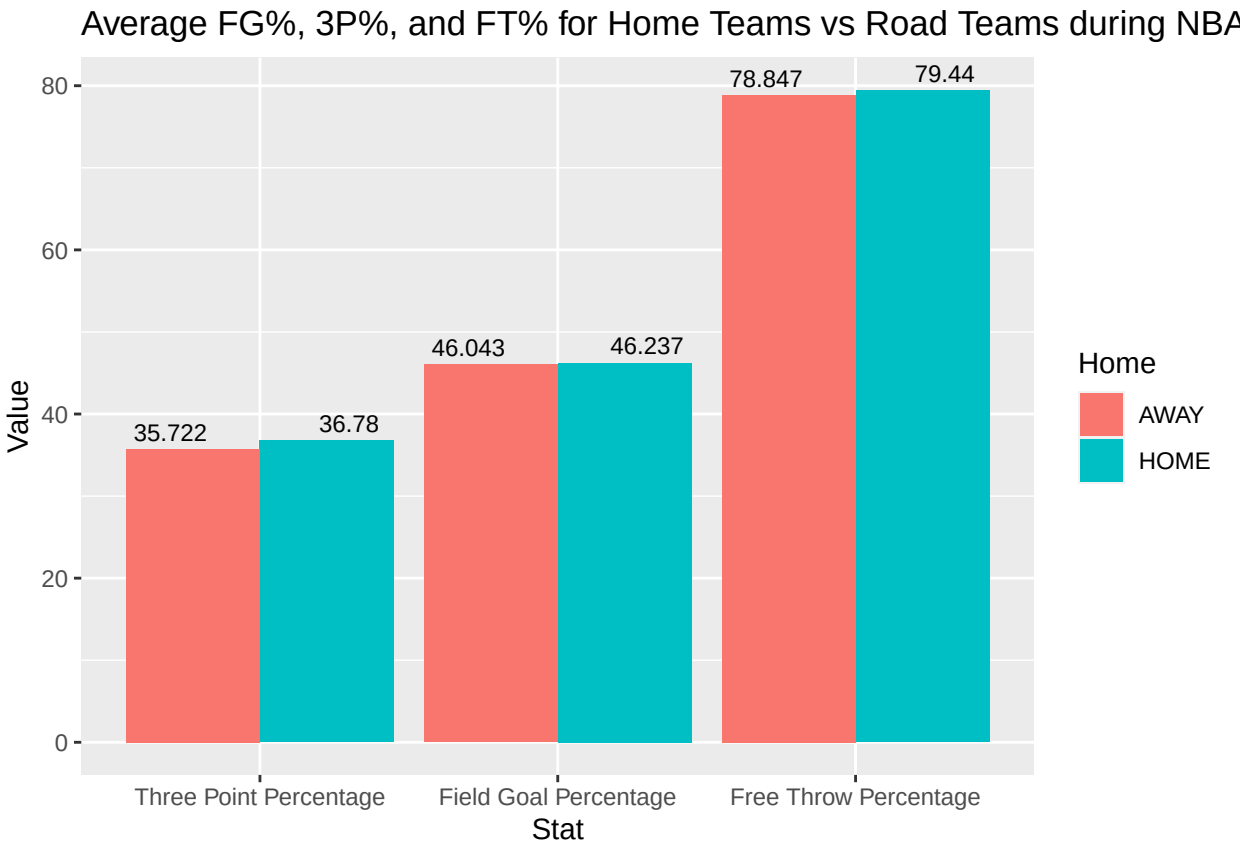
```
Bubble_summary %>%
 pivot_longer(c(MEAN_AST_HOME,MEAN_AST_AWAY, MEAN_POINTS_HOME, MEAN_POINTS_AWAY,
  ↪  MEAN_REB_HOME, MEAN_REB_AWAY), names_to = c("Stat", "Home"), names_pattern =
  ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Home, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("Assists", "Points",
   ↪  "Rebounds")) + labs(title = "Average Assists, Points, and Rebounds for Home Teams
   ↪  vs Road Teams during NBA Bubble")
```

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




```
Bubble_summary %>%
  pivot_longer(c(MEAN_FGPCT_HOME, MEAN_FGPCT_AWAY, MEAN_FTPCT_HOME, MEAN_FTPCT_AWAY,
   ↪  MEAN_FG3PCT_HOME, MEAN_FG3PCT_AWAY), names_to = c("Stat", "Home"), names_pattern =
   ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Home, Value)  %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("Three Point Percentage", "Field
   ↪  Goal Percentage", "Free Throw Percentage")) + labs(title = "Average FG%, 3P%, and
   ↪  FT% for Home Teams vs Road Teams during NBA Bubble")
```

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



From the graphs, we see that the difference between home and away teams is extremely minimal in each statistic. Still, it's interesting that the home stat is always marginally greater than the away stat, given the neutral setting of the bubble.

2020-2021 Season

The season after the Disney Bubble still had limited fan capacity during the regular season due to the Covid-19 pandemic. Most teams only allowed 10-25% fan capacity, with this figure rising to 25-50% later on in the season. We will analyze if the smaller fan capacity made an impact on home court advantage for all of the teams.

We will first filter the NBA dataset so that we only have the regular season games for the 2020-2021 season.

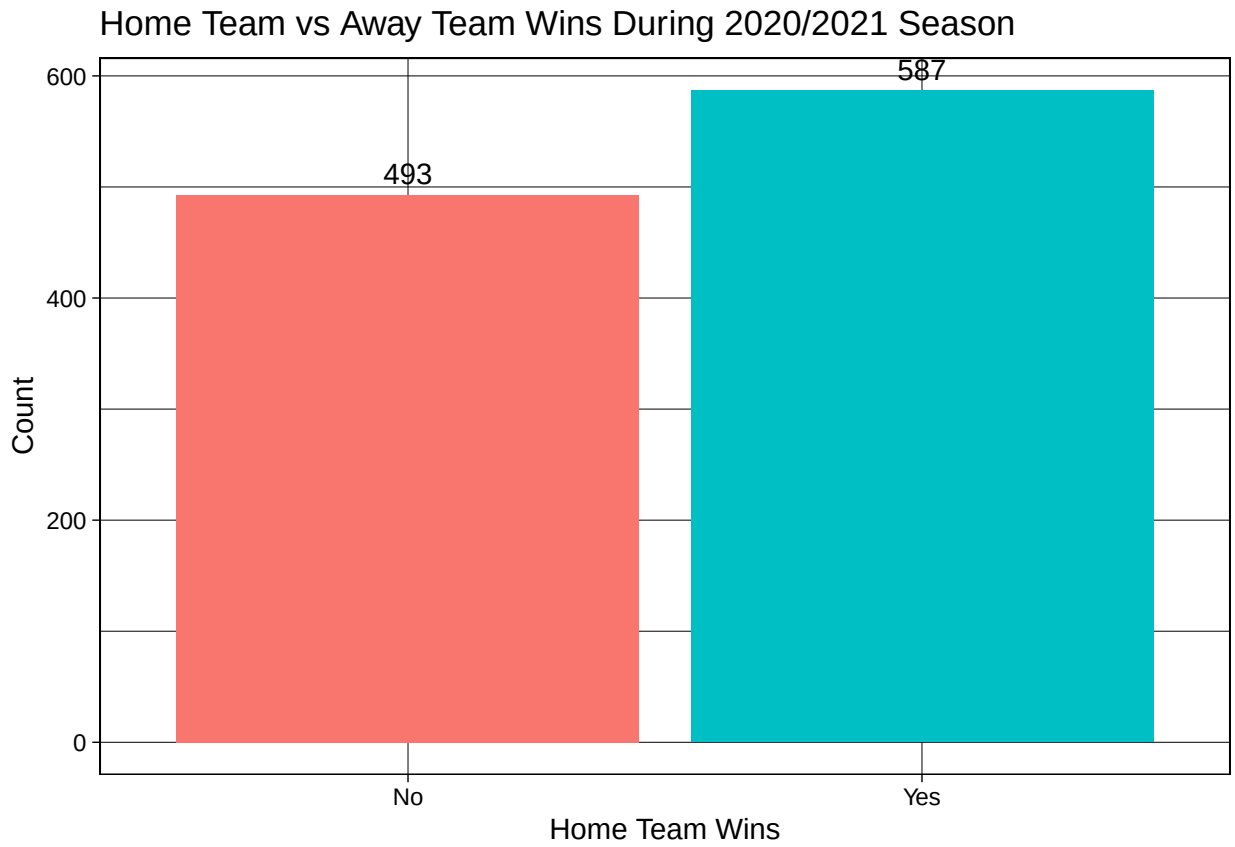
```
Season_2021 <-NBA %>%
  filter(GAME_DATE_EST >= "2020-12-22" & GAME_DATE_EST <= "2021-05-16")
```

We will now look at the amount of wins versus losses at home for all of the games during the 2021 season and use a bar chart and Z-test to analyze the result..

```
Season_2021 %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0   493
## 2              1   587
```

```
Home_Team_Wins_2021 <- c("No", "Yes")
Number_2021 <- c(493, 587)
Home_wins_2021 <- data.frame(Home_Team_Wins_2021, Number_2021)

Home_wins_2021 %>%
  ggplot(aes(x = Home_Team_Wins_2021, y = Number_2021)) + geom_col(aes(fill =
  ↪  Home_Team_Wins_2021)) + geom_text(aes(label = Number_2021), vjust = -0.5) +
  ↪  theme_linedraw() + labs(x = "Home Team Wins", y = "Count", title = "Home Team vs
  ↪  Away Team Wins During 2020/2021 Season") + theme(legend.position = "none")
```




```
prop.test(587, 1080, p = 0.5)
```

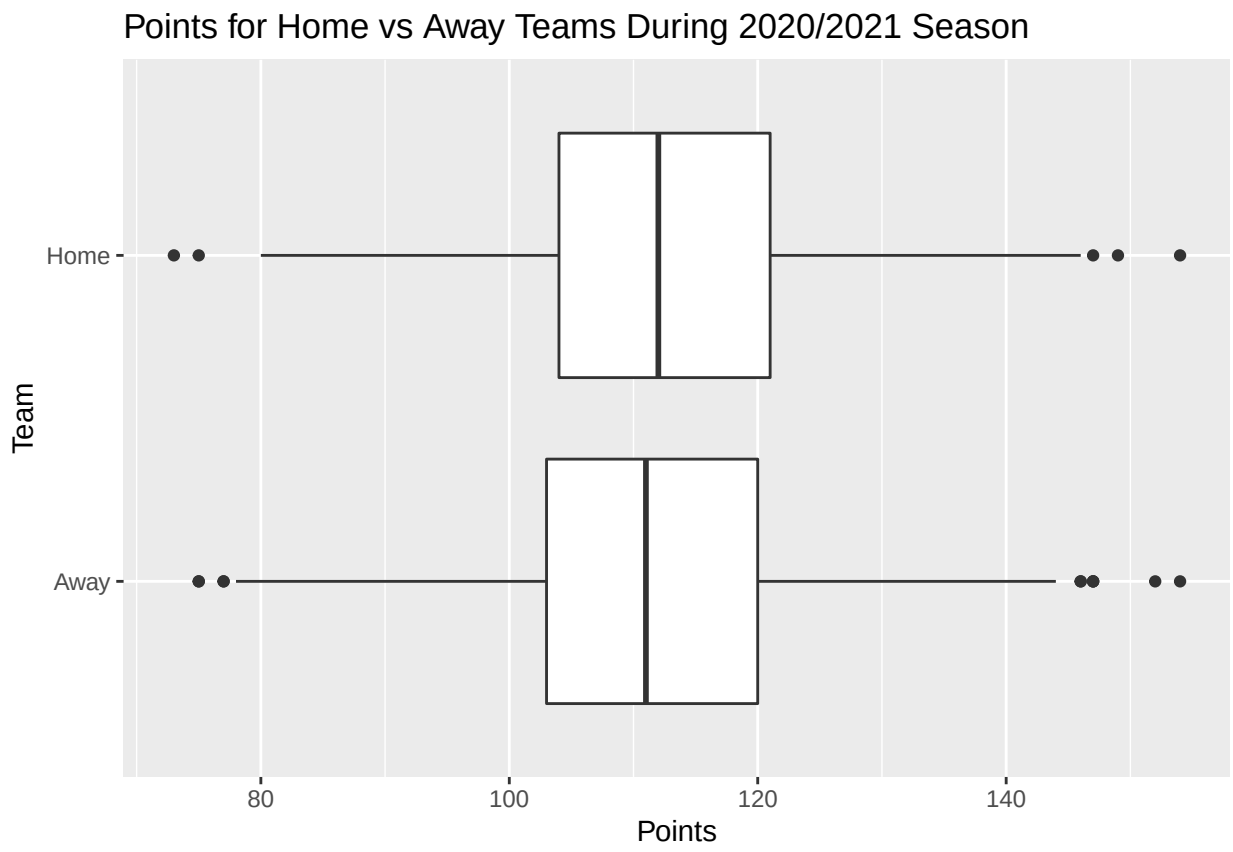
```
##
##  1-sample proportions test with continuity correction
##
## data:  587 out of 1080, null probability 0.5
## X-squared = 8.0083, df = 1, p-value = 0.004656
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5132461 0.5734776
## sample estimates:
##         p
## 0.5435185
```

From the proportion-test, we see that there is a statistically significant difference between the proportion of home wins and 0.5, the expected proportion if home court advantage didn't exist. The p-value of 0.004656 is extremely small and well below the 0.05 level of significance, so we can reject the null hypothesis that the proportion of home wins is equal to the proportion of away wins. Unlike the bubble, we do see an advantage that comes with playing on home court in 2020/2021 reflected in the proportion of home wins, despite the extremely limited fan capacity for the regular season.

We will now make boxplots for all of the statistics (such as points, field goal percentage, etc) for home vs away teams. We will also run T-tests in order to determine if the differences are statistically significant.

Points

```
Season_2021 %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Points for Home vs Away Teams During 2020/2021
  ↪  Season")
```

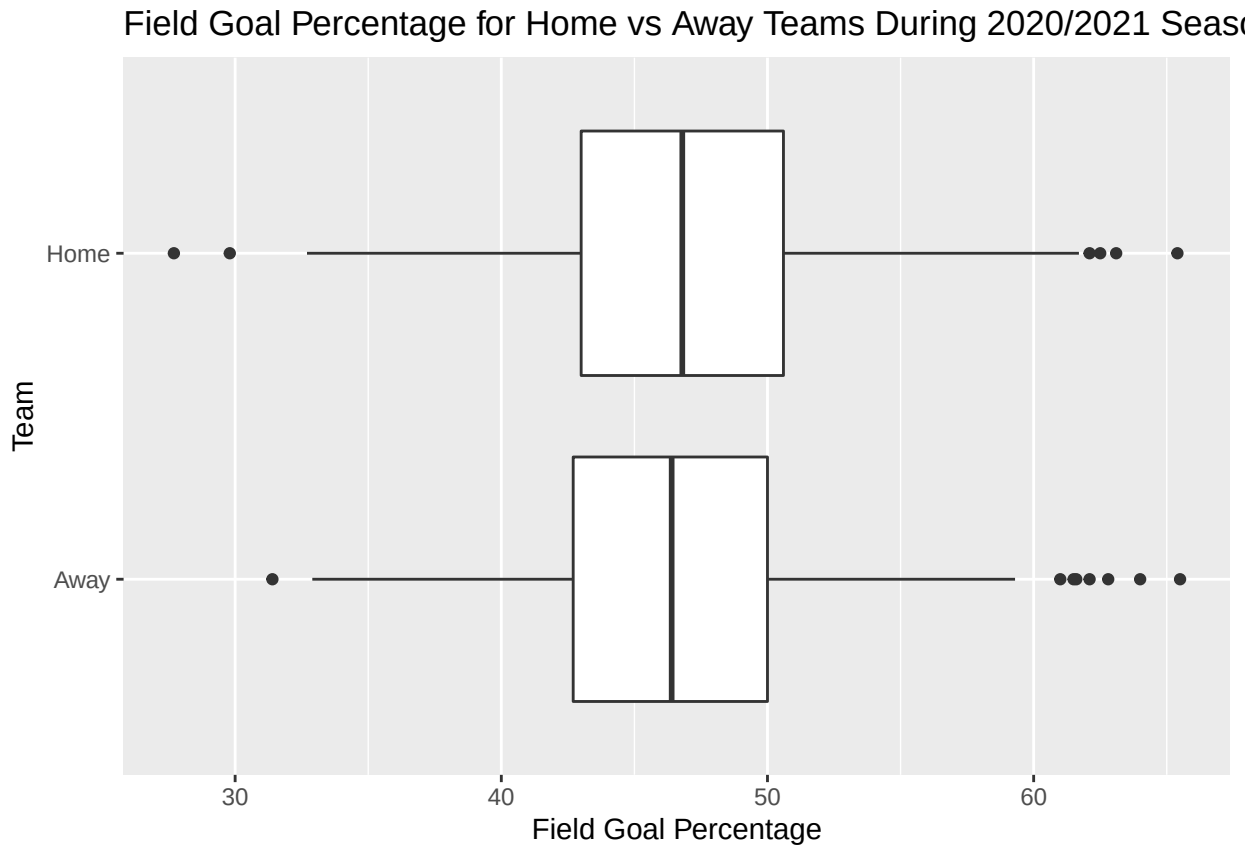



```
t.test(Season_2021$PTS_home, Season_2021$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$PTS_home and Season_2021$PTS_away
## t = 2.042, df = 1079, p-value = 0.04139
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.03690902 1.85012802
## sample estimates:
## mean difference
##       0.9435185
```

Field Goal Percentage

```
Season_2021 %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Field Goal Percentage
  ↪  for Home vs Away Teams During 2020/2021 Season", x = "Field Goal Percentage")
```

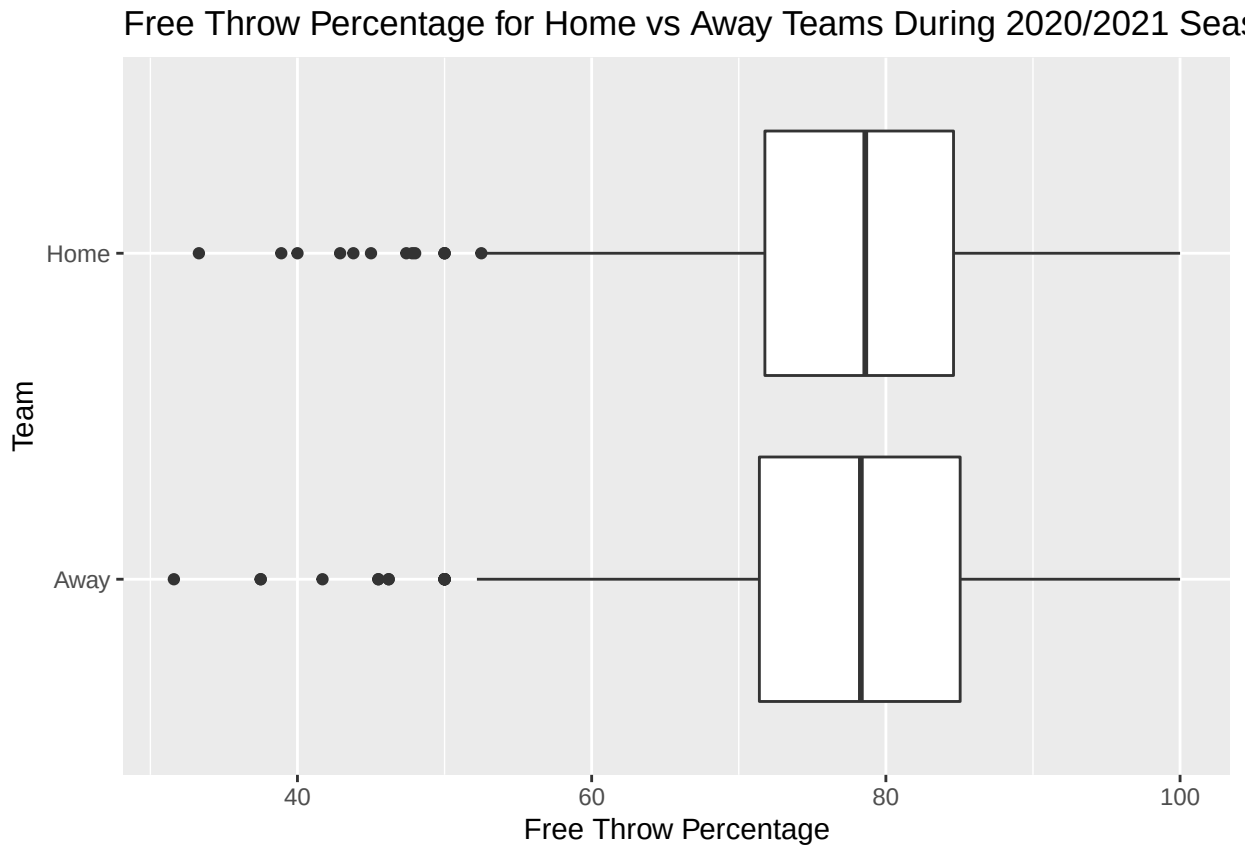



```
t.test(Season_2021$FG_PCT_home, Season_2021$FG_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$FG_PCT_home and Season_2021$FG_PCT_away
## t = 1.6801, df = 1079, p-value = 0.09324
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.06581315  0.84970204
## sample estimates:
## mean difference
##       0.3919444
```

Free Throw Percentage

```
Season_2021 %>%
  pivot_longer(c(FT_PCT_away, FT_PCT_home), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Free Throw Percentage
  ↪  for Home vs Away Teams During 2020/2021 Season", x = "Free Throw Percentage")
```

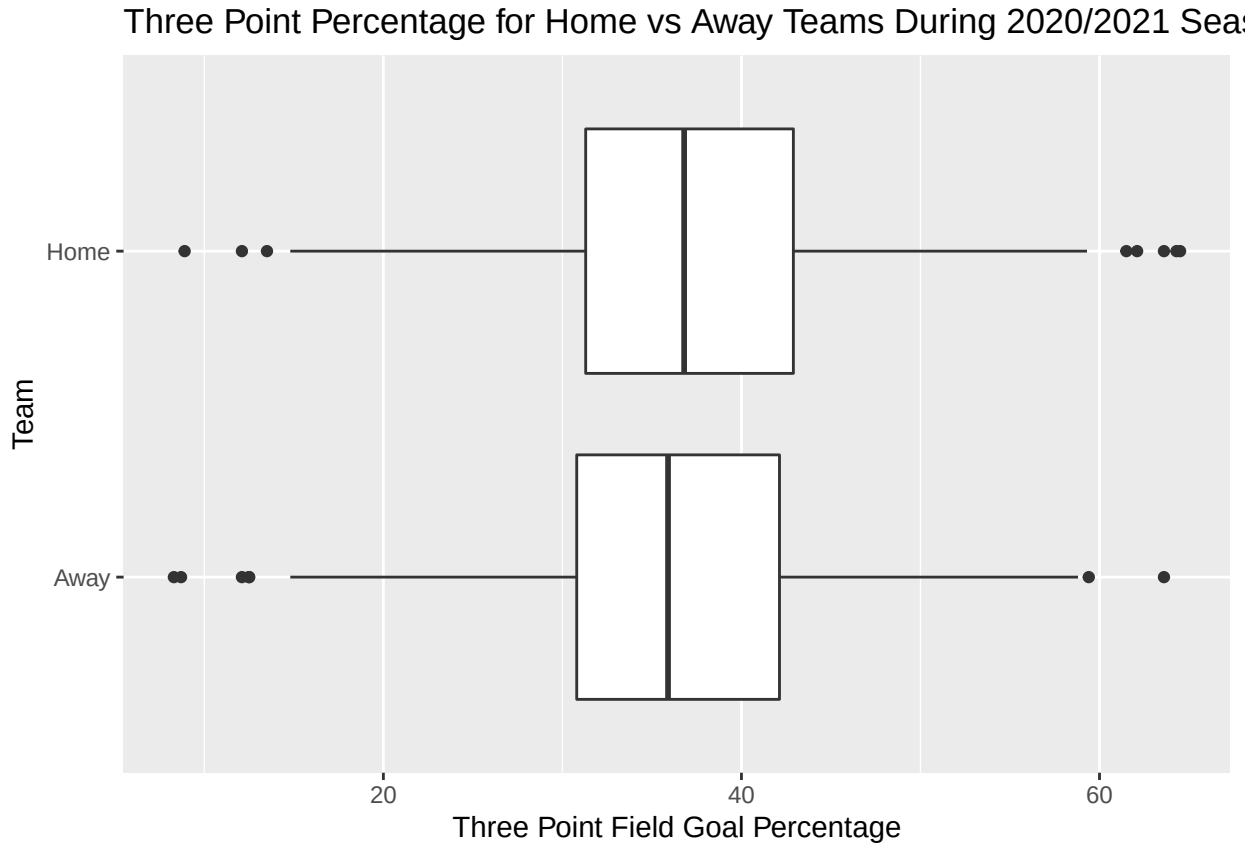



```
t.test(Season_2021$FT_PCT_home, Season_2021$FT_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$FT_PCT_home and Season_2021$FT_PCT_away
## t = 0.48579, df = 1079, p-value = 0.6272
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.6511521  1.0796706
## sample estimates:
## mean difference
##       0.2142593
```

Three Point Percentage

```
Season_2021 %>%
  pivot_longer(c(FG3_PCT_away, FG3_PCT_home), names_to = "Team", values_to =
  ↪  "Three_Point_Percentage") %>%
  ggplot(aes(x = Three_Point_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Three Point Percentage
  ↪  for Home vs Away Teams During 2020/2021 Season", x = "Three Point Field Goal
  ↪  Percentage")
```

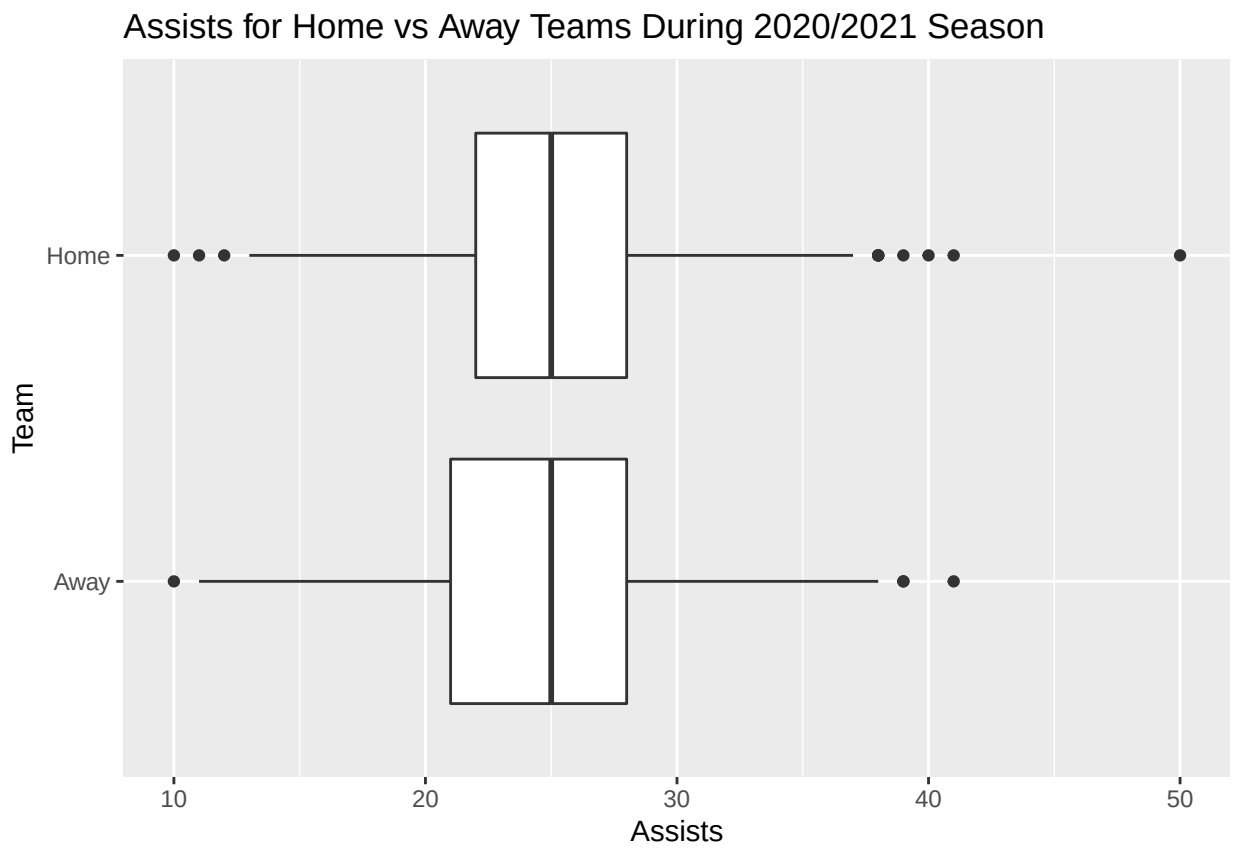



```
t.test(Season_2021$FG3_PCT_home, Season_2021$FG3_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$FG3_PCT_home and Season_2021$FG3_PCT_away
## t = 1.6372, df = 1079, p-value = 0.1019
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.1236873  1.3697984
## sample estimates:
## mean difference
##       0.6230556
```

Assists

```
Season_2021 %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Teams During 2020/2021
  ↪  Season")
```

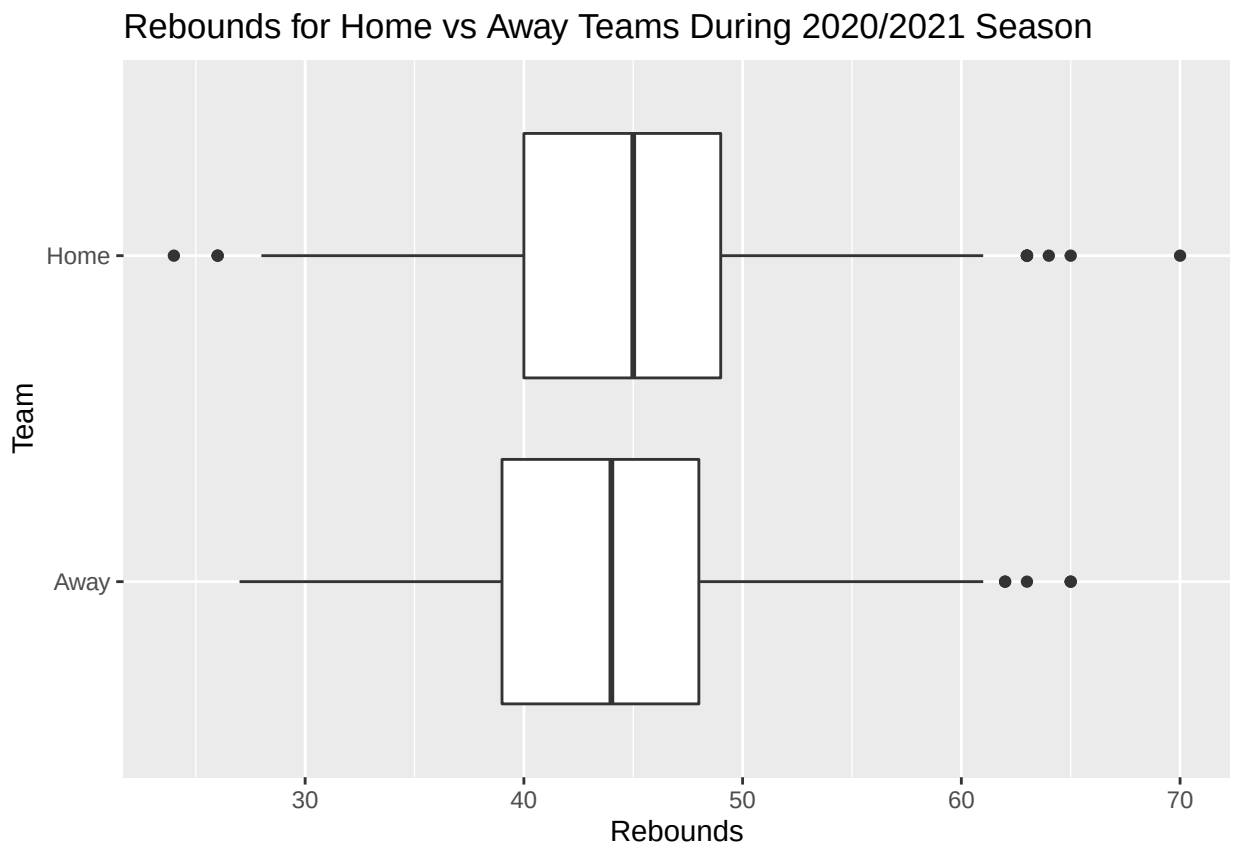



```
t.test(Season_2021$AST_home, Season_2021$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$AST_home and Season_2021$AST_away
## t = 1.6754, df = 1079, p-value = 0.09414
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.05815871  0.73778834
## sample estimates:
## mean difference
##       0.3398148
```

Rebounds

```
Season_2021 %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Teams During 2020/2021
  ↪  Season")
```




```
t.test(Season_2021$REB_home, Season_2021$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Season_2021$REB_home and Season_2021$REB_away
## t = 2.9931, df = 1079, p-value = 0.002825
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.2883041 1.3857700
## sample estimates:
## mean difference
##        0.837037
```

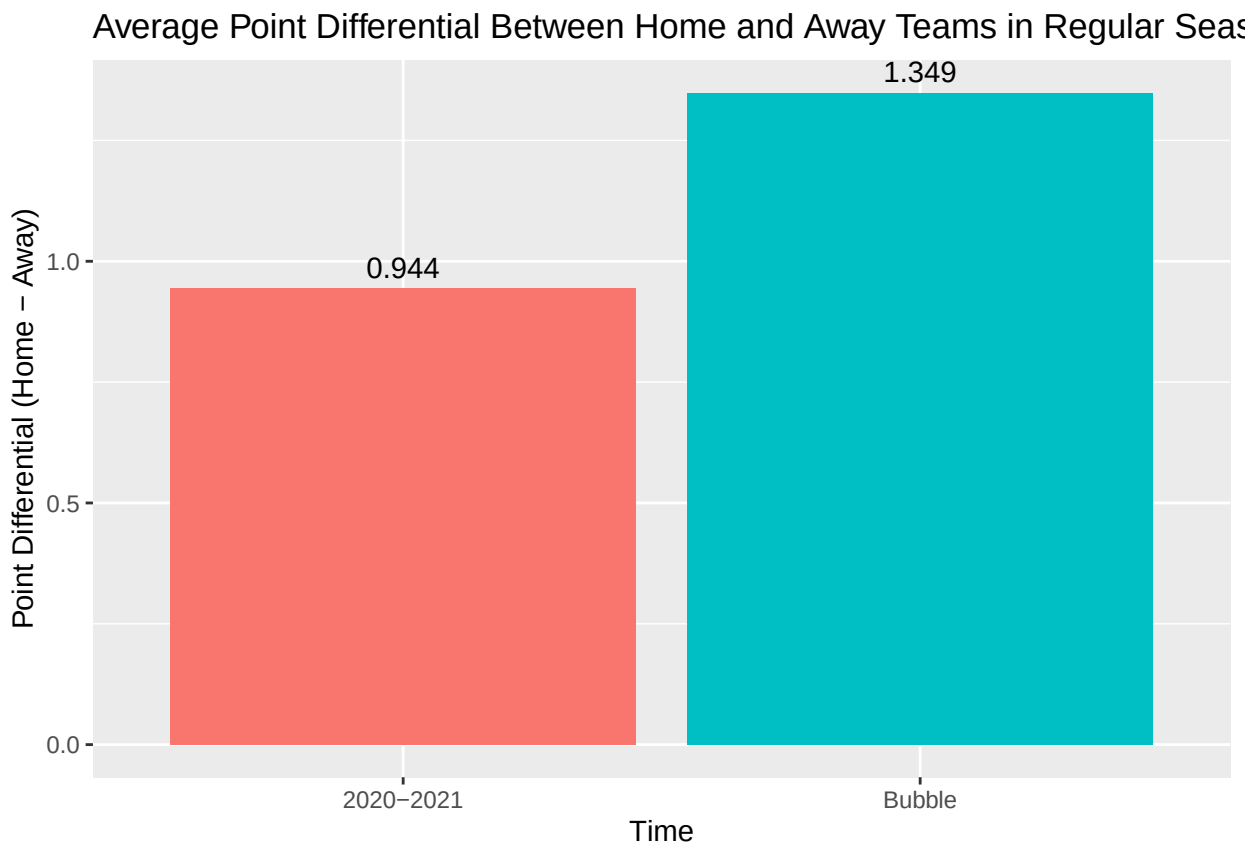
Based on the boxplots above, we can see that the first quartile, median, and third quartile are generally higher for home than away teams except for the third quartile of assists and free throw percentage. Looking

at the t-tests, we see that the p-values for all of the statistics except points and rebounds are greater than 0.05, meaning that the differences between home and away values are not statistically significant for those specific statistics. However, since there is a statistically significant difference in points, we can conclude that home teams generally score more than away teams during this season. Unlike the bubble, which saw no significant mean differences in any of the stats, there seems to be more of a home court advantage in this season which involved playing in different arenas despite having a low fan presence in the stadiums. While the differences in the results for home and away teams are not as significant as our analysis on the entire data, we do see that there is more of an impact of home court advantage when there involves playing in different arenas despite a lack of fan presence. We can also conclude that when there are less fans in the arena, there is less of a home court advantage. From our analysis on all of the data, we observed that playing at home gives a team an advantage of roughly 2.86 points. However, in the 2021 season, we are 95% confident that the true mean difference between home points and away points lies between 0.03690902 and 1.85012802 points, which is well below the 2.86. Additionally, our analysis on all of the data found that we can be 95% confident that proportion of wins for home teams is between 0.5847659 and 0.5972231. However, in the 2021 season, we are 95% confident that the proportion of home wins lies between 0.5132461 and 0.5734776.

Bubble vs 2020/2021 Season

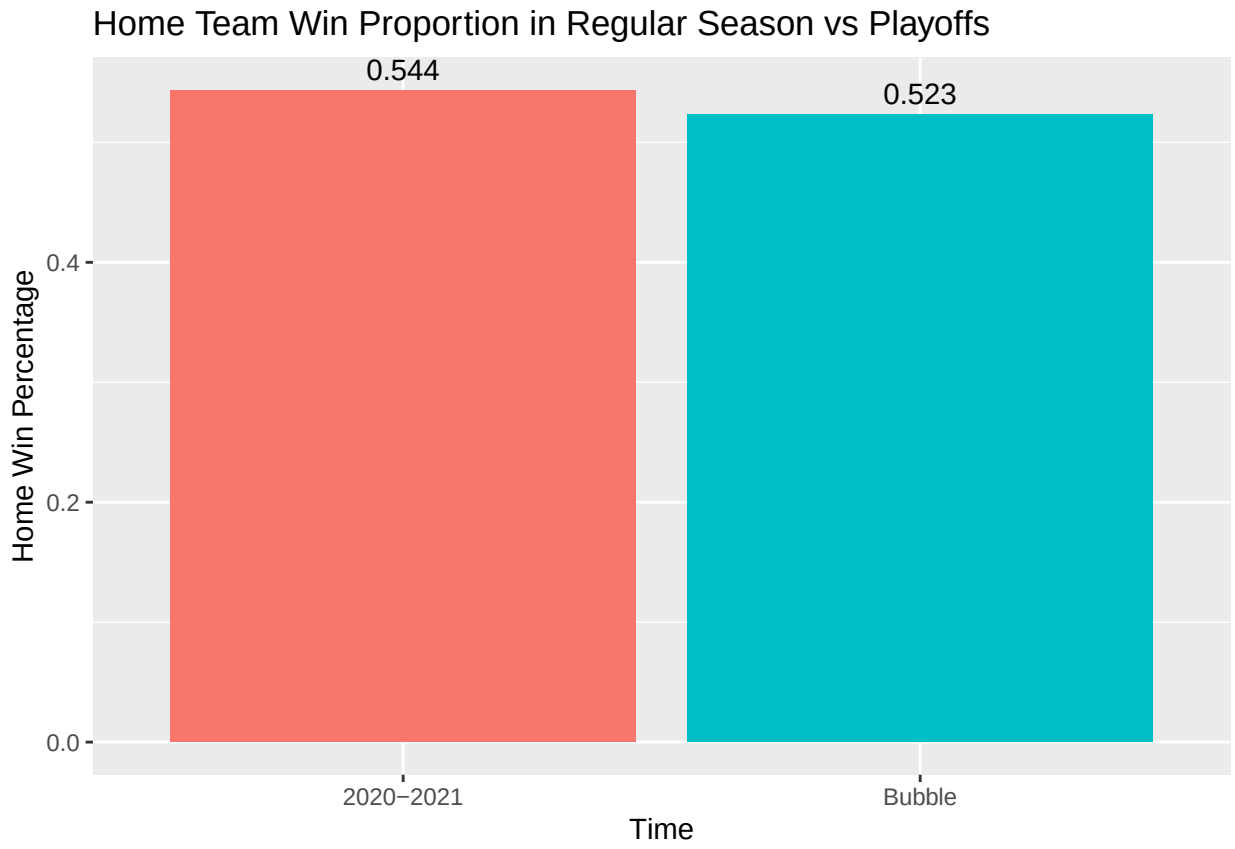
We can make bar graphs to compare the bubble with the 2020/2021 season by looking at the mean home win percentage, point differential, and field goal percentage differential. Since comparing two separate time periods does not involve matched, we cannot simply compare home points from the bubble with home points from the 2020/2021 season. This is because a team could score 120 points but lose 150-120, whereas another team may only score 100 points but win 100-80.

```
Bubble <- Bubble %>% mutate(Time = "Bubble") %>% mutate(pointdifferential = PTS_home -
↪  PTS_away) %>% mutate(fgpctdifferential = FG_PCT_home - FG_PCT_away)
Season_2021 <- Season_2021 %>% mutate(Time = "2020-2021") %>% mutate(pointdifferential =
↪  PTS_home - PTS_away) %>% mutate(fgpctdifferential = FG_PCT_home - FG_PCT_away)
total22 <- rbind(Bubble, Season_2021)

total22 %>%
  group_by(Time) %>%
  summarize(mean_point_differential = mean(pointdifferential)) %>%
  ggplot(aes(x = Time, y = mean_point_differential, fill = Time)) + geom_col() +
  ↪  theme(legend.position = "none") + labs(y = "Point Differential (Home - Away)",
  ↪  title = "Average Point Differential Between Home and Away Teams in Regular Season
  ↪  vs Playoffs") + geom_text(aes(label = round(mean_point_differential, 3)), vjust =
  ↪  -0.5, position = position_dodge(1))
```

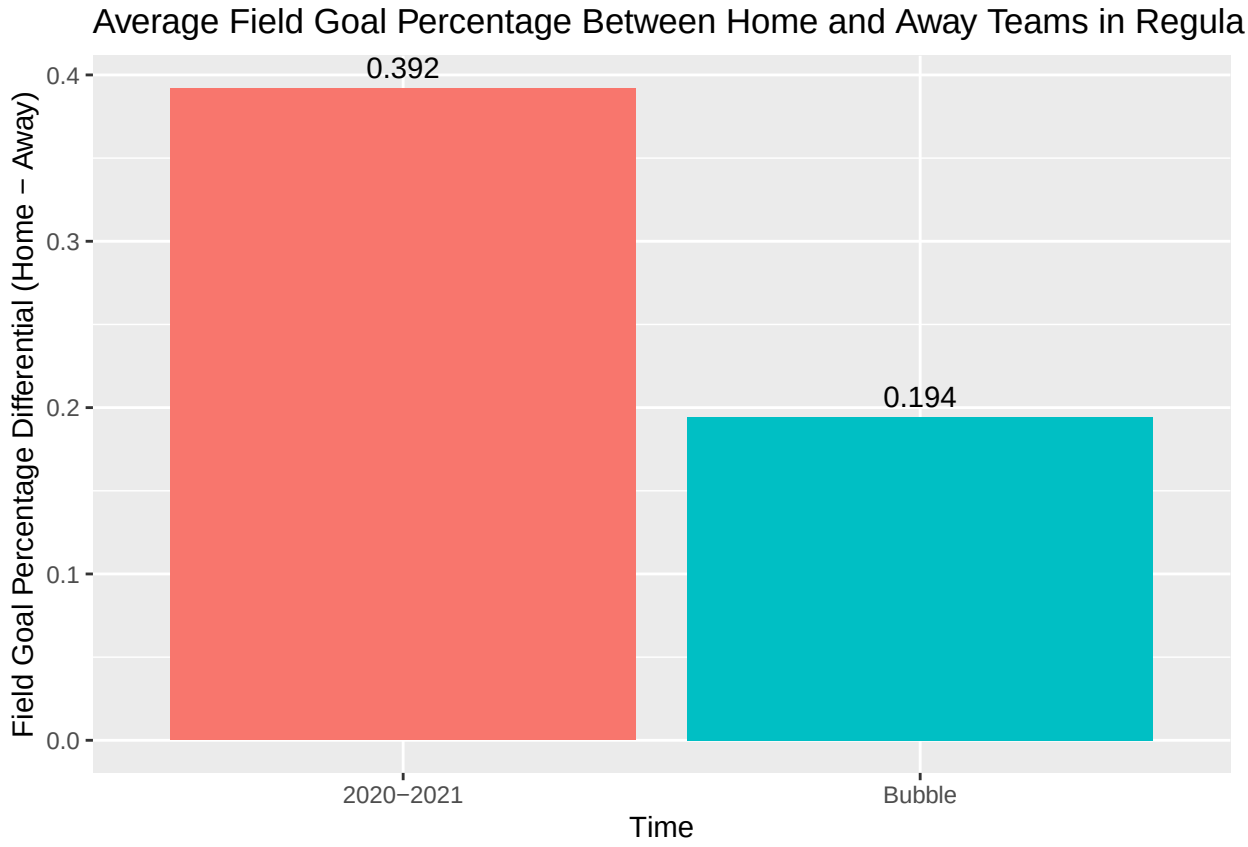



```
total22 %>%
  group_by(Time) %>%
```

```
  summarize(mean_wpct = mean(HOME_TEAM_WINS)) %>%
  ggplot(aes(x = Time, y = mean_wpct, fill = Time)) + geom_col() + theme(legend.position
  ↪  = "none") + labs(y = "Home Win Percentage", title = "Home Team Win Proportion in
  ↪  Regular Season vs Playoffs") + geom_text(aes(label = round(mean_wpct, 3)), vjust =
  ↪  -0.5, position = position_dodge(1))
```




```
total22 %>%
  group_by(Time) %>%
  summarize(mean_fgpct_differential = mean(fgpctdifferential)) %>%
  ggplot(aes(x = Time, y = mean_fgpct_differential, fill = Time)) + geom_col() +
  ↪  theme(legend.position = "none") + labs(y = "Field Goal Percentage Differential
  ↪  (Home - Away)", title = "Average Field Goal Percentage Between Home and Away Teams
  ↪  in Regular Season vs Playoffs") + geom_text(aes(label =
  ↪  round(mean_fgpct_differential, 3)), vjust = -0.5, position = position_dodge(1))
```




```
t.test(Bubble$pointdifferential, Season_2021$pointdifferential)
```

```
##
##  Welch Two Sample t-test
##
## data:  Bubble$pointdifferential and Season_2021$pointdifferential
## t = 0.36166, df = 246.57, p-value = 0.7179
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.802095  2.612732
## sample estimates:
## mean of x mean of y
## 1.3488372 0.9435185
```

```
t.test(Bubble$HOME_TEAM_WINS, Season_2021$HOME_TEAM_WINS)
```

```
##
##  Welch Two Sample t-test
##
## data:  Bubble$HOME_TEAM_WINS and Season_2021$HOME_TEAM_WINS
## t = -0.49307, df = 228.26, p-value = 0.6224
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10123598  0.06071057
```

```
## sample estimates:
## mean of x mean of y
## 0.5232558 0.5435185
```

```
t.test(Bubble$fgpctdifferential, Season_2021$fgpctdifferential)
```

```
##
##  Welch Two Sample t-test
##
## data:  Bubble$fgpctdifferential and Season_2021$fgpctdifferential
## t = -0.33261, df = 237.65, p-value = 0.7397
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3690394  0.9735226
## sample estimates:
## mean of x mean of y
## 0.1941860 0.3919444
```

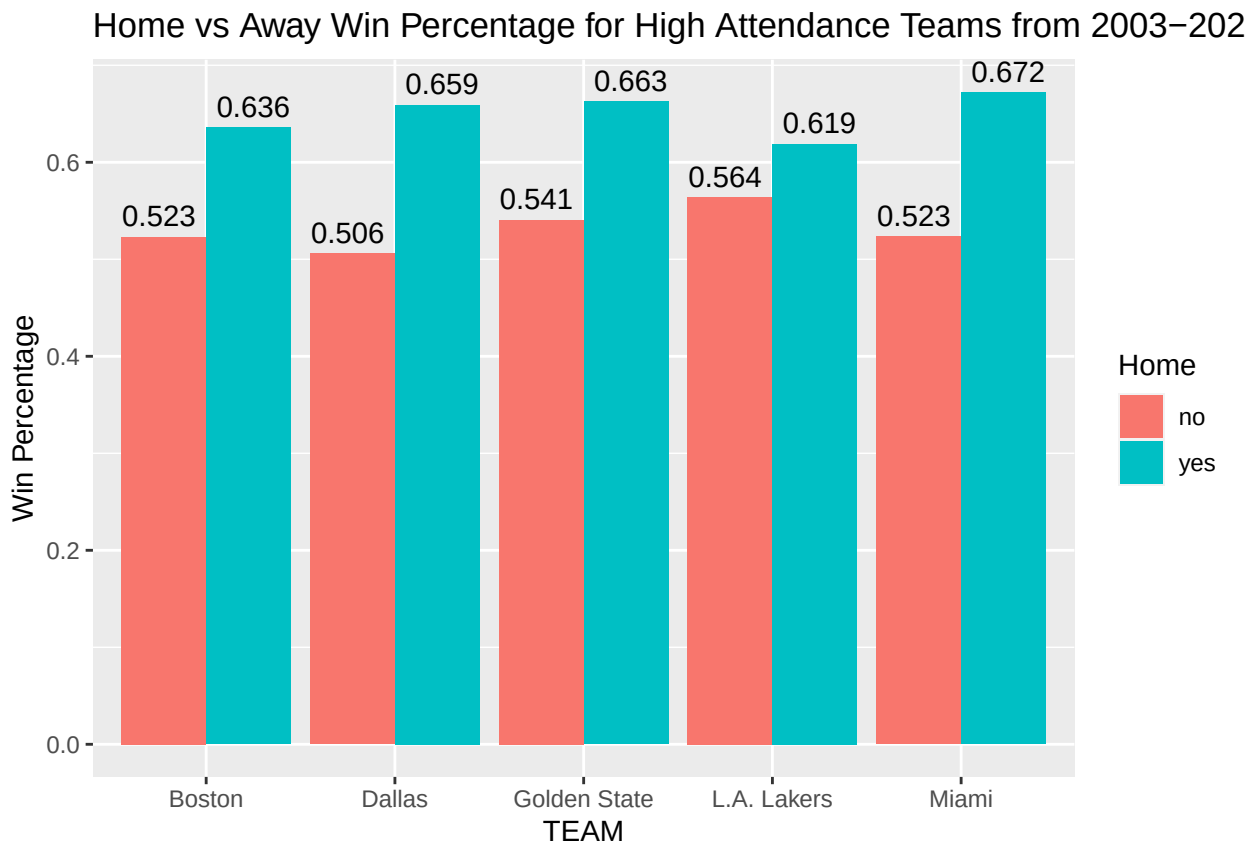
Comparing the bubble and the year after, there is no significant difference between points, win percentage, or field goal percentage, which suggests a lack of a difference between the seasons. So, while home court advantage was slightly noticeable in win percentage and points in the 2020/2021 season unlike the bubble, the added home court advantage in this season was not significantly different from the bubble. The bar graphs suggest that home teams won slightly more in 2020/2021 compared to the bubble, but they also won games by a smaller average margin, while shooting a better percentage from the field. Since these two seasons are quite similar and very far off from the general data, this suggests that crowds play a significant role in giving the home team an advantage, because even with playing at different arenas in 2020/2021, the differences between home and away stats did not jump back up to previous levels.

Comparing Home vs Away Win Percentage by Team Based on Attendance

We will look at five NBA teams with consistently high attendance in the top ten (Boston, Dallas, Golden State, L.A. Lakers, and Miami) and five NBA teams with consistently low attendance in the bottom ten (Detroit, Minnesota, Orlando, Sacramento, and Washington) from 2003-2004 to 2021-2022 according to ESPN's attendance statistics.

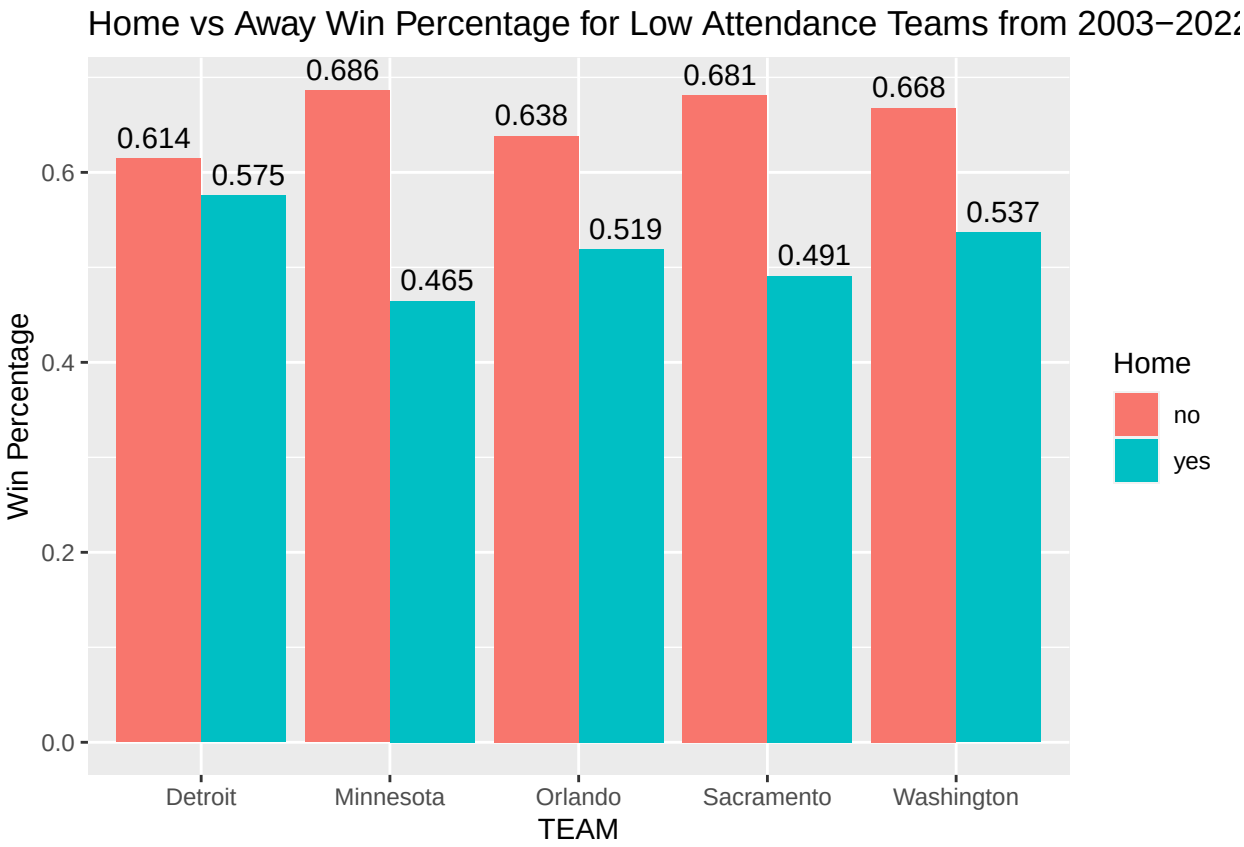
```
NBA_home <- NBA %>%
  group_by(HOME_TEAM) %>%
  summarize(avg_home_wpct = mean(HOME_TEAM_WINS))

NBA_away <- NBA %>%
  group_by(AWAY_TEAM) %>%
  summarize(avg_away_wpct = mean(HOME_TEAM_WINS))

NBA_wpct <- NBA_home %>%
  inner_join(NBA_away, by = c("HOME_TEAM" = "AWAY_TEAM")) %>%
  rename("TEAM" = "HOME_TEAM")

NBA_wpct %>%
  pivot_longer(c("avg_home_wpct", "avg_away_wpct"), names_to = "Home", values_to =
  ↪  "wpct") %>%
  filter(TEAM %in% c("Dallas", "Boston", "Golden State", "Miami", "L.A. Lakers")) %>%
  ggplot(aes(x = TEAM, y = wpct, fill = Home)) + geom_col(position = "dodge", stat =
  ↪  "identity") + labs(y = "Win Percentage", title = "Home vs Away Win Percentage for
  ↪  High Attendance Teams from 2003-2022")  + scale_fill_discrete(labels=c('no',
  ↪  'yes')) + geom_text(aes(label = round(wpct, 3)), vjust = -0.5, position =
  ↪  position_dodge(1))
```

```
## Warning: Ignoring unknown parameters: stat
```




```
NBA_wpct %>%
  pivot_longer(c("avg_home_wpct", "avg_away_wpct"), names_to = "Home", values_to =
  ↪  "wpct") %>%
  filter(TEAM %in% c("Sacramento", "Orlando", "Washington", "Detroit", "Minnesota")) %>%
  ggplot(aes(x = TEAM, y = wpct, fill = Home)) + geom_col(position = "dodge", stat =
  ↪  "identity") + labs(y = "Win Percentage", title = "Home vs Away Win Percentage for
  ↪  Low Attendance Teams from 2003-2022")  + scale_fill_discrete(labels=c('no', 'yes'))
  ↪  + geom_text(aes(label = round(wpct, 3)), vjust = -0.5, position =
  ↪  position_dodge(1))
```

```
## Warning: Ignoring unknown parameters: stat
```



These bar graphs show an interesting trend that the teams with higher attendance have a higher home winning percentage compared to away winning percentage, whereas teams with lower attendance all happened to have a higher away winning percentage compared to home winning percentage. This could suggest that aside from the arena at hand, the crowd environment plays a big role in providing home court advantage.

Analyzing Specific Team Matchups

Lakers vs Clippers matchups

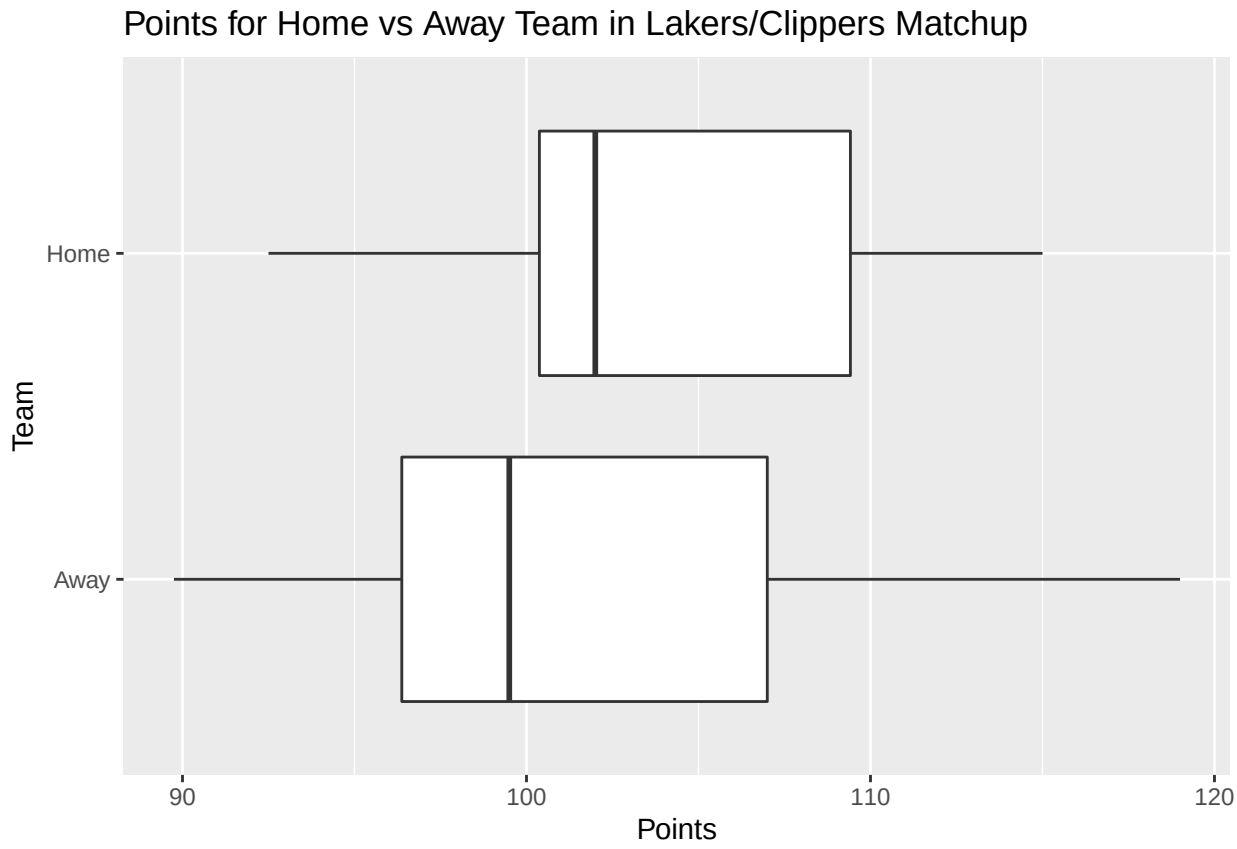
Another factor that is important when it comes to home court advantage is the arena and familiarity with the court, since many arenas have minor differences in lighting, environment, and depth perception, among others. During the period of time we are analyzing, the Los Angeles Lakers and Clippers have both played in the same arena, Staples Center (now Crypto.com Arena). Since they are both in the same division, they generally play four times per season, and they each play two as the home team and two as the away team. The only exceptions were the 2011-2012 shortened season due to lockout, and the 2020-2021 season due to less games being played as a consequence of the pandemic. We will analyze to see if there is a difference when one team is 'home' and the other team is 'away', despite playing in the same home arena with just minor cosmetic changes such as different court decorations.

We will first filter the NBA dataset so that we only get the matchups where the Lakers and Clippers play against each other. We will also group by season because teams generally gets new players between seasons and may become significantly better or worse, but within a season, teams generally only make minor changes. Grouping by season will ensure more consistency in our analysis. We will then use the summarize function to find the means of each stat for home and away teams within each individual season. Lastly, we will conduct paired t-test's for each of the statistics to conclude if the differences are statistically significant between home and away teams.

```
NBALA <- NBA %>%
  filter(HOME_TEAM %in% c("L.A. Lakers", "L.A. Clippers", "LA Clippers"), AWAY_TEAM %in%
  ↪  c("L.A. Clippers", "LA Clippers", "L.A. Lakers")) %>%
  group_by(SEASON) %>%
  summarize(home_ast_LA = mean(AST_home), away_ast_LA = mean(AST_away), home_points_LA =
  ↪  mean(PTS_home), away_points_LA = mean(PTS_away), home_reb_LA = mean(REB_home),
  ↪  away_reb_LA = mean(REB_away), home_fgpct_LA = mean(FG_PCT_home), away_fgpct_LA =
  ↪  mean(FG_PCT_away), home_ftpct_LA = mean(FT_PCT_home), away_ftpct_LA =
  ↪  mean(FT_PCT_away), home_fg3pct_LA = mean(FG3_PCT_home), away_fg3pct_LA =
  ↪  mean(FG3_PCT_away))
```

Points

```
NBALA %>%
  pivot_longer(c(home_points_LA, away_points_LA), names_to = "Team", values_to =
  ↪  "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Points for Home vs Away Team in Lakers/Clippers
  ↪  Matchup")
```

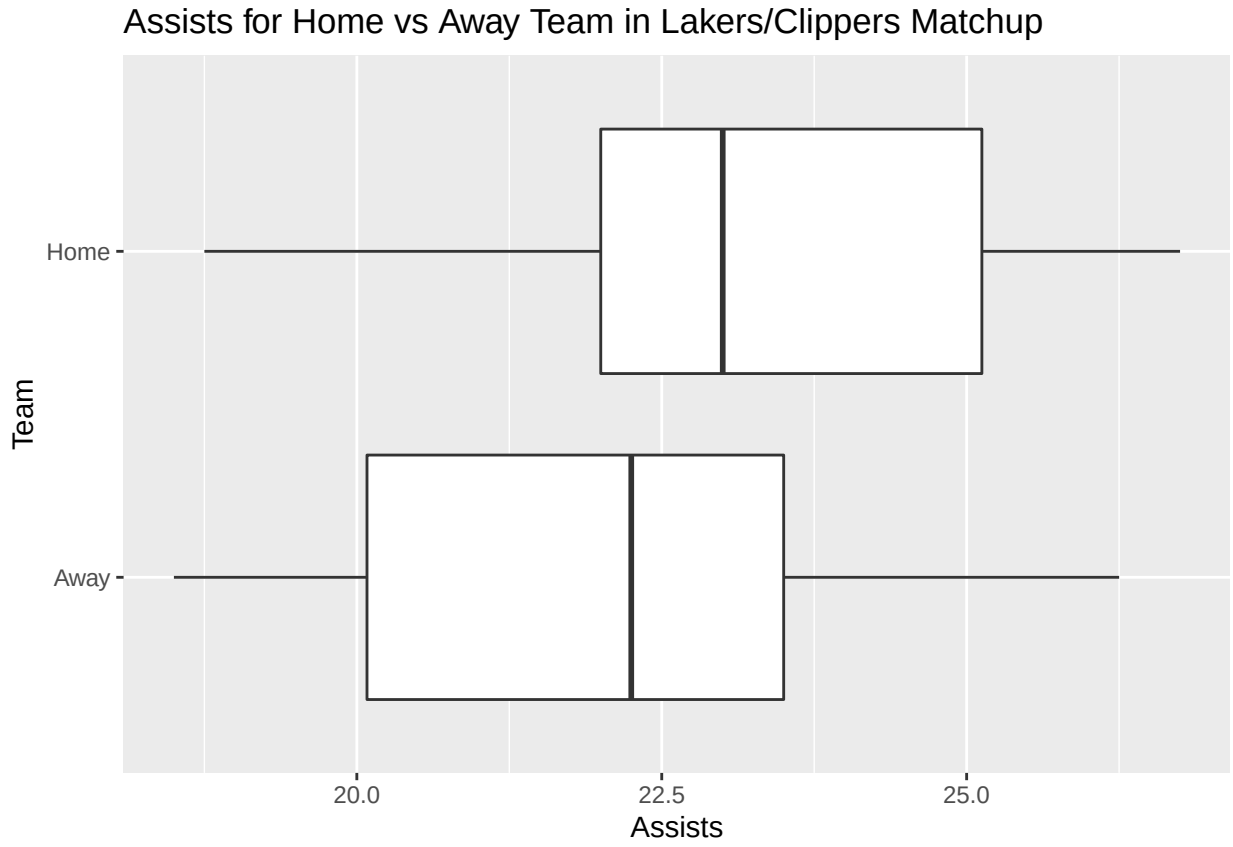



```
t.test(NBALA$home_points_LA, NBALA$away_points_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_points_LA and NBALA$away_points_LA
## t = 1.4225, df = 18, p-value = 0.172
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.043707  5.420900
## sample estimates:
## mean difference
##        2.188596
```

Assists

```
NBALA %>%
  pivot_longer(c(home_ast_LA, away_ast_LA), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Team in Lakers/Clippers
  ↪  Matchup")
```

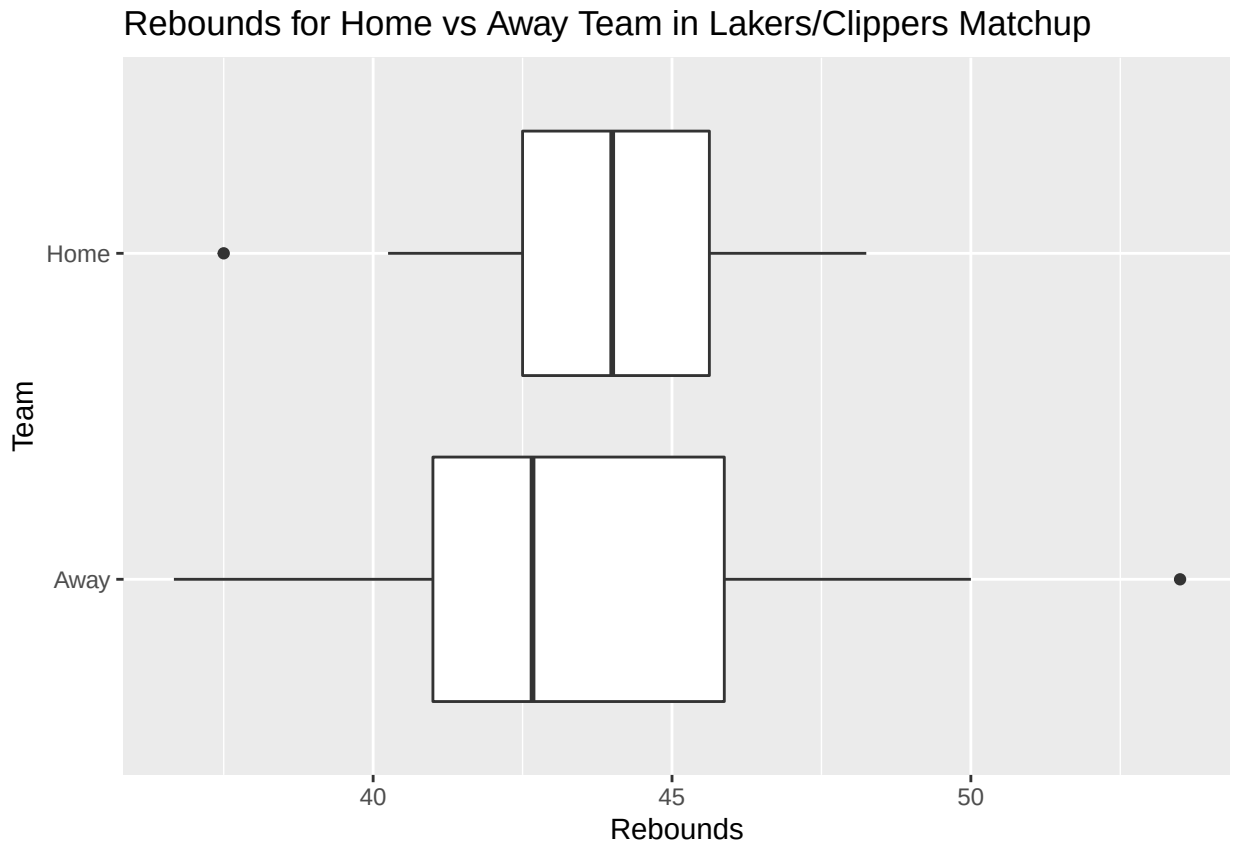



```
t.test(NBALA$home_ast_LA, NBALA$away_ast_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_ast_LA and NBALA$away_ast_LA
## t = 2.242, df = 18, p-value = 0.0378
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.07478907 2.30240391
## sample estimates:
## mean difference
##        1.188596
```

Rebounds

```
NBALA %>%
  pivot_longer(c(home_reb_LA, away_reb_LA), names_to = "Team", values_to = "Rebounds")
  ↪  %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Team in
  ↪  Lakers/Clippers Matchup")
```

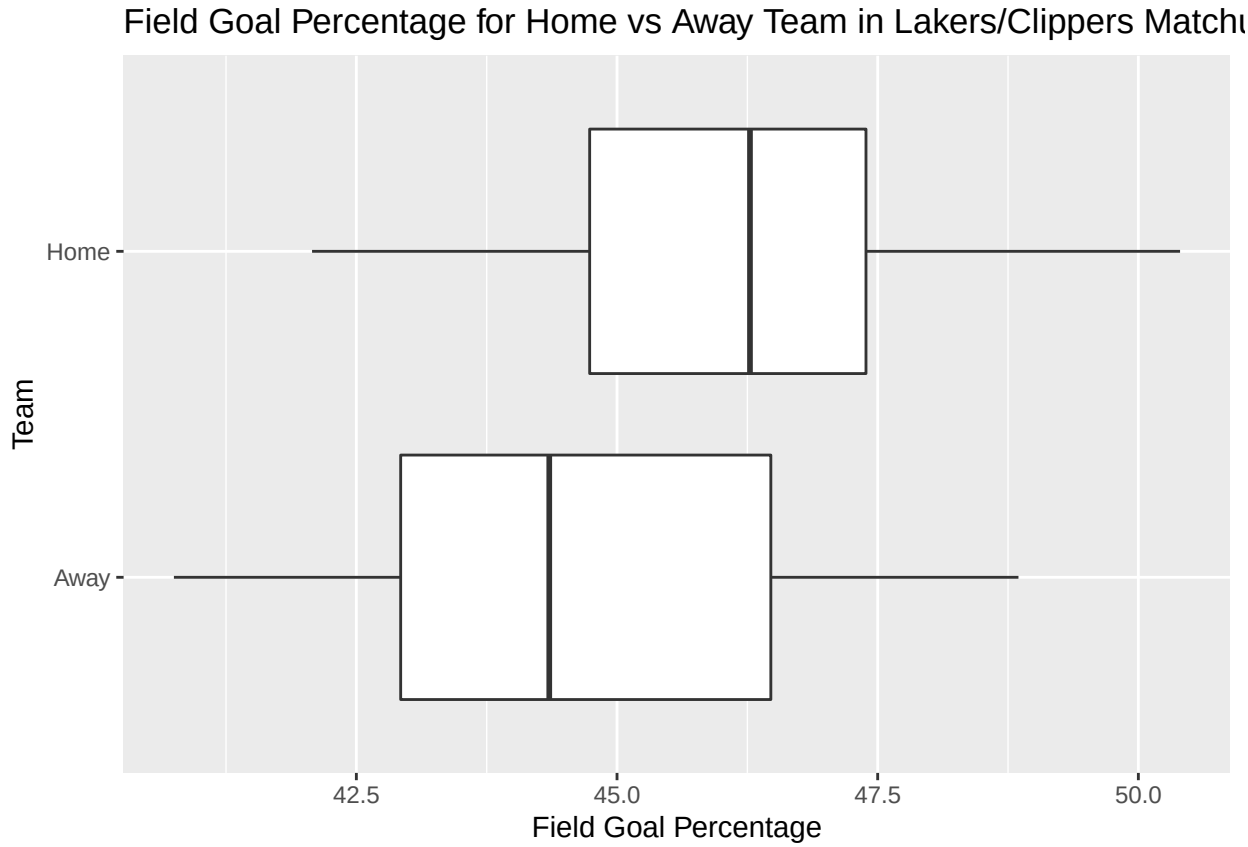



```
t.test(NBALA$home_reb_LA, NBALA$away_reb_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_reb_LA and NBALA$away_reb_LA
## t = 0.52029, df = 18, p-value = 0.6092
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.732183  2.872534
## sample estimates:
## mean difference
##       0.5701754
```

Field Goal Percentage

```
NBALA %>%
  pivot_longer(c(home_fgpct_LA, away_fgpct_LA), names_to = "Team", values_to = "FGPCT")
  ↪  %>%
  ggplot(aes(x = FGPCT, y = Team)) + geom_boxplot() + scale_y_discrete(labels = c("Away",
  ↪  "Home")) + labs(x = "Field Goal Percentage", title = "Field Goal Percentage for
  ↪  Home vs Away Team in Lakers/Clippers Matchup")
```

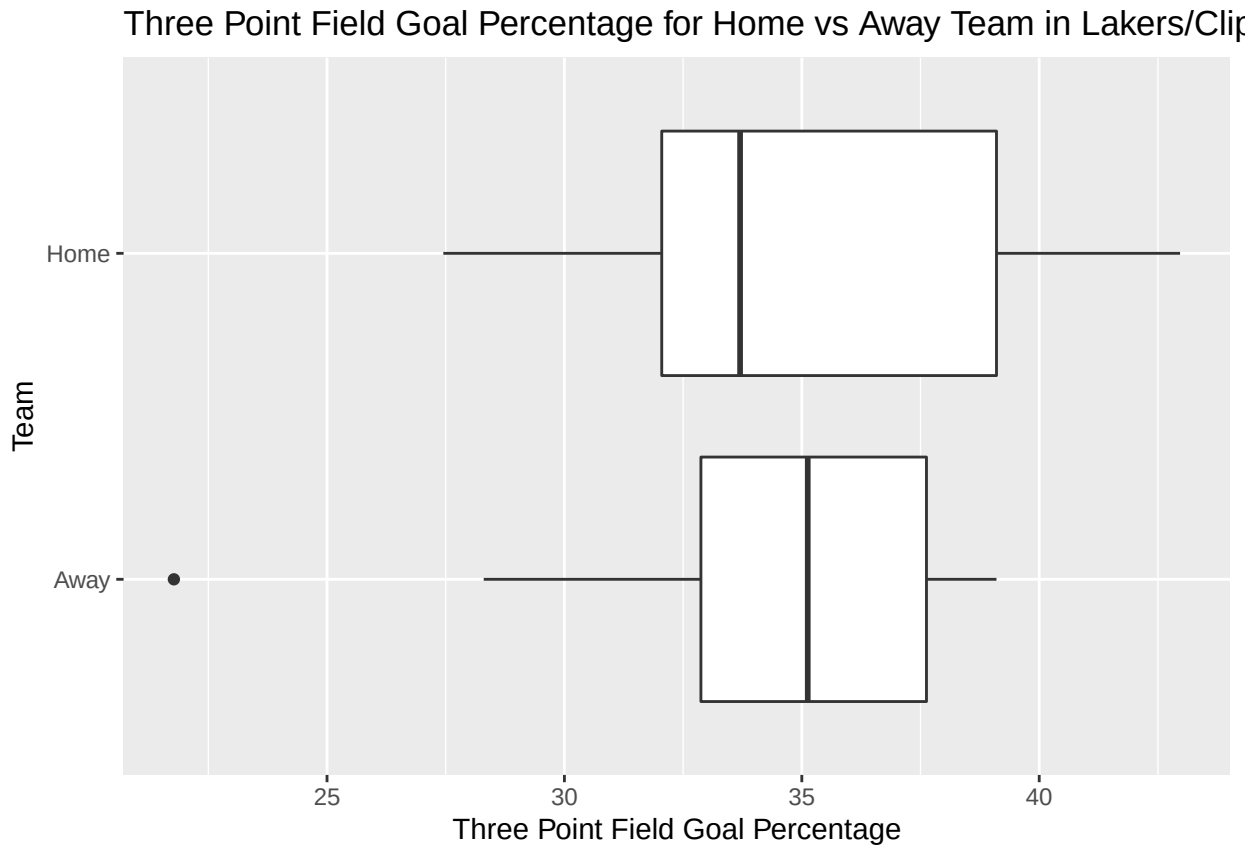



```
t.test(NBALA$home_fgpct_LA, NBALA$away_fgpct_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_fgpct_LA and NBALA$away_fgpct_LA
## t = 2.1645, df = 18, p-value = 0.04411
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.04753529 3.18755243
## sample estimates:
## mean difference
##        1.617544
```

Three Point Percentage

```
NBALA %>%
  pivot_longer(c(home_fg3pct_LA, away_fg3pct_LA), names_to = "Team", values_to =
  ↪  "FG3PCT") %>%
  ggplot(aes(x = FG3PCT, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(x = "Three Point Field Goal Percentage", title = "Three
  ↪  Point Field Goal Percentage for Home vs Away Team in Lakers/Clippers Matchup")
```

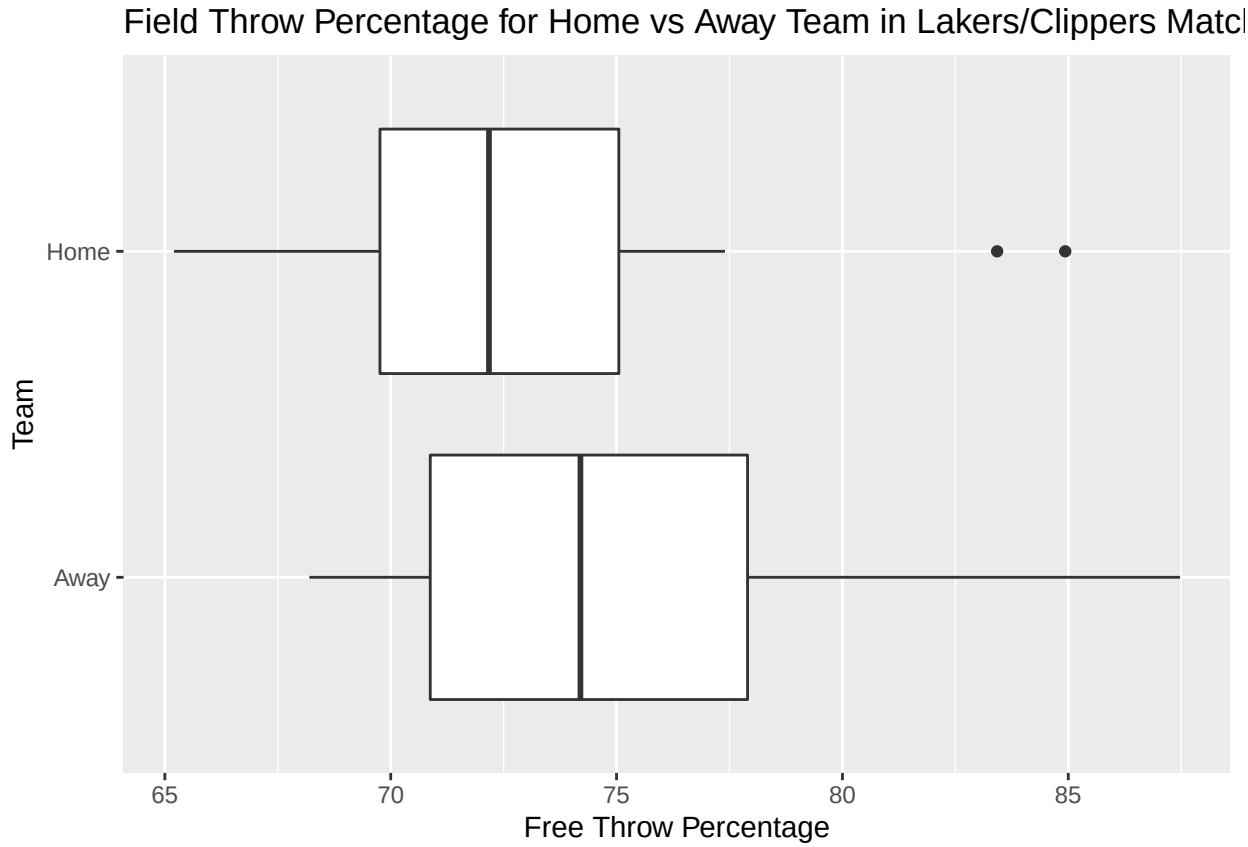



```
t.test(NBALA$home_fg3pct_LA, NBALA$away_fg3pct_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_fg3pct_LA and NBALA$away_fg3pct_LA
## t = 0.48511, df = 18, p-value = 0.6335
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2.229338  3.567935
## sample estimates:
## mean difference
##       0.6692982
```

Free Throw Percentage

```
NBALA %>%
  pivot_longer(c(home_ftpct_LA, away_ftpct_LA), names_to = "Team", values_to = "FTPCT")
  ↪  %>%
  ggplot(aes(x = FTPCT, y = Team)) + geom_boxplot() + scale_y_discrete(labels = c("Away",
  ↪  "Home")) + labs(x = "Free Throw Percentage", title = "Field Throw Percentage for
  ↪  Home vs Away Team in Lakers/Clippers Matchup")
```




```
t.test(NBALA$home_ftpct_LA, NBALA$away_ftpct_LA, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  NBALA$home_ftpct_LA and NBALA$away_ftpct_LA
## t = -1.3861, df = 18, p-value = 0.1827
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.885527  1.206579
## sample estimates:
## mean difference
##       -2.339474
```

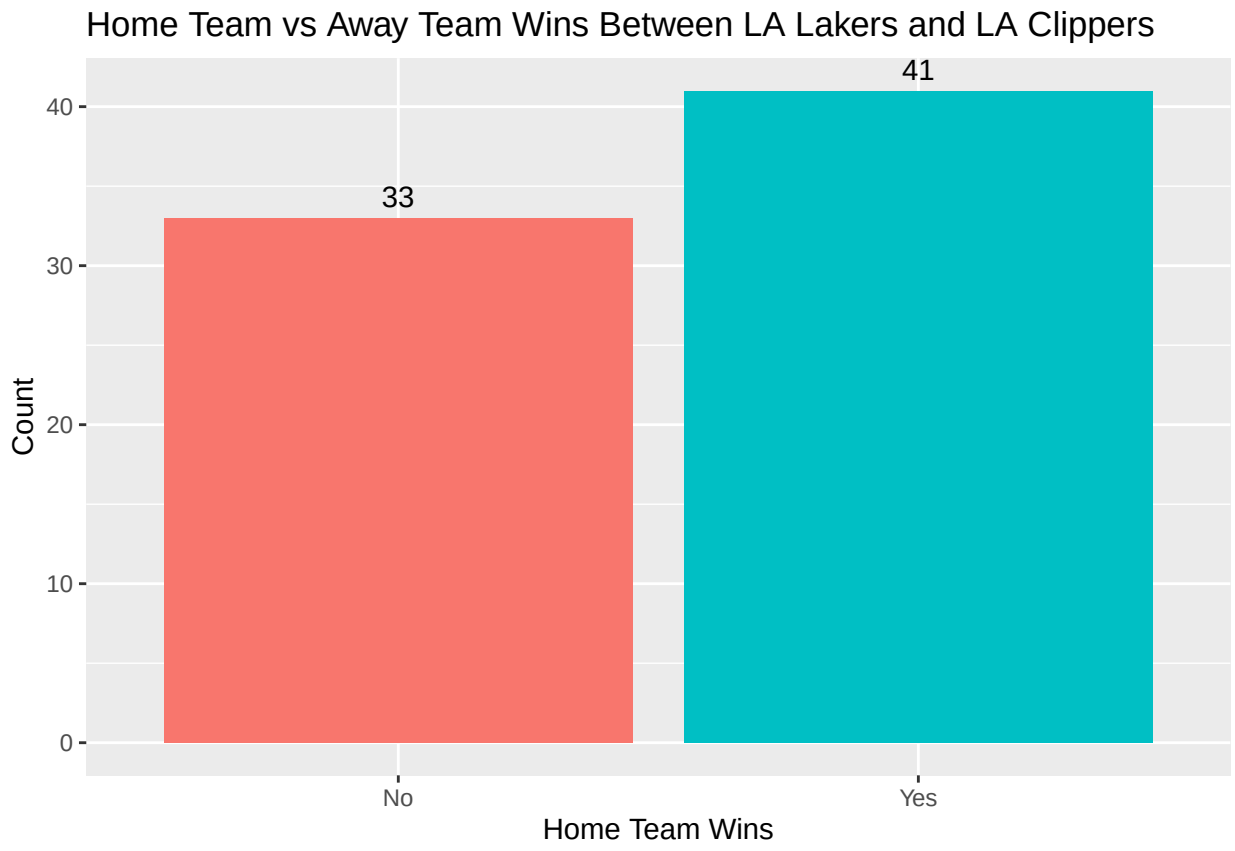
Wins

```
NBA %>%
  filter(HOME_TEAM %in% c("L.A. Lakers", "L.A. Clippers", "LA Clippers"), AWAY_TEAM %in%
  ↪  c("L.A. Clippers", "LA Clippers", "L.A. Lakers")) %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
```

```
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0    33
## 2              1    41
```

```
Home_Team_Wins_LA <- c("No", "Yes")
Number_LA <- c(33, 41)
Home_wins_LA <- data.frame(Home_Team_Wins_LA, Number_LA)

Home_wins_LA %>%
  ggplot(aes(x = Home_Team_Wins_LA, y = Number_LA)) + geom_col(aes(fill =
  ↪  Home_Team_Wins_LA)) + geom_text(aes(label = Number_LA), vjust = -0.5) + labs(x =
  ↪  "Home Team Wins", y = "Count", title = "Home Team vs Away Team Wins Between LA
  ↪  Lakers and LA Clippers") + theme(legend.position = "none")
```



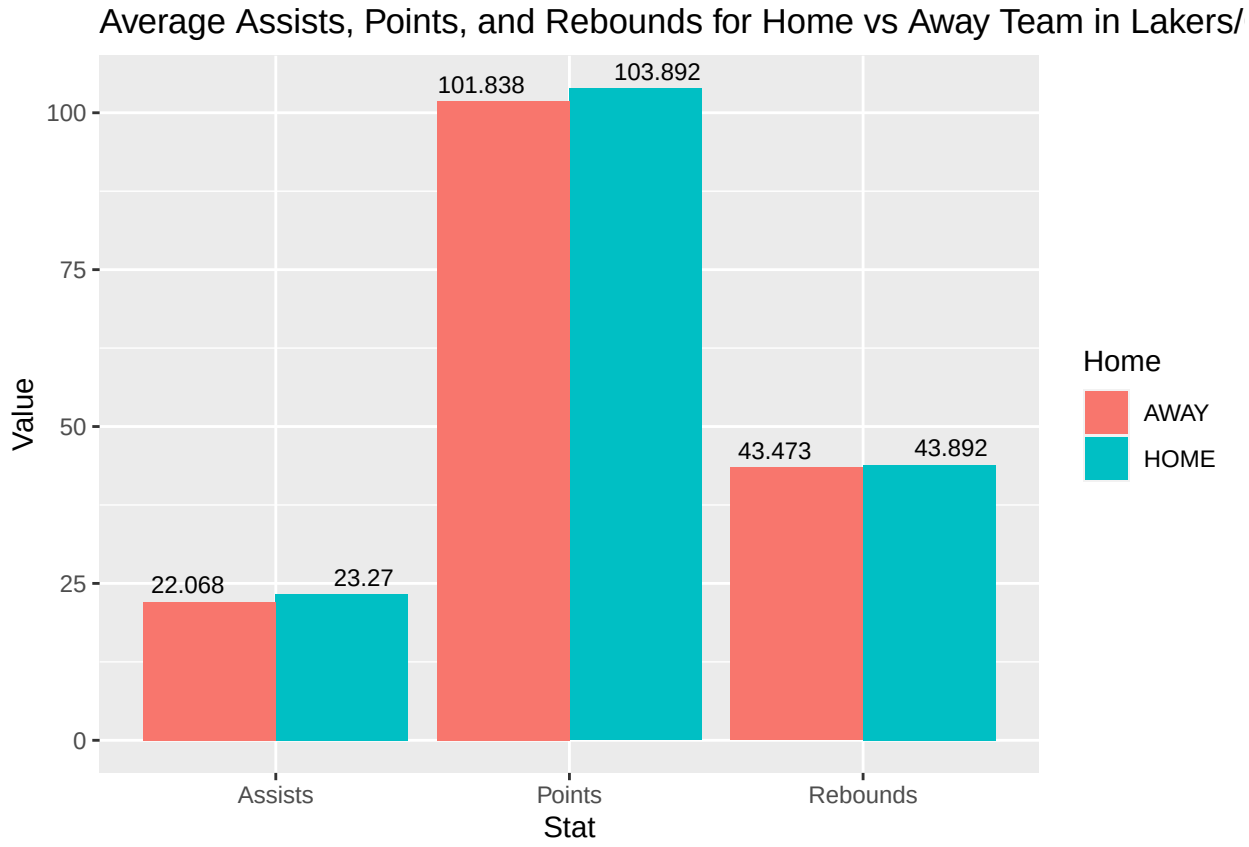

```
prop.test(41, 74, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  41 out of 74, null probability 0.5
## X-squared = 0.66216, df = 1, p-value = 0.4158
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.4343626 0.6681001
## sample estimates:
##         p
## 0.5540541
```

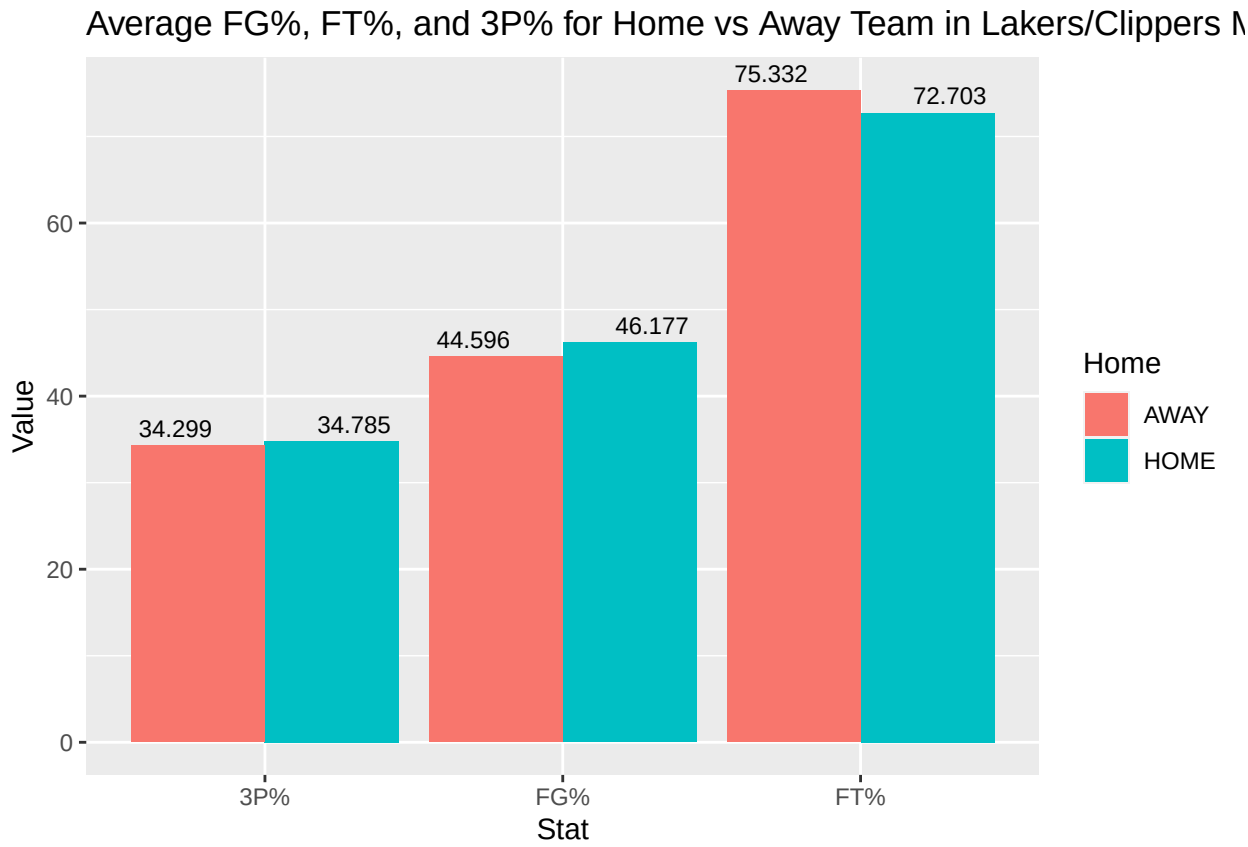
```
NBALASUM <- NBA %>%
  filter(HOME_TEAM %in% c("L.A. Lakers", "L.A. Clippers", "LA Clippers"), AWAY_TEAM %in%
   ↪  c("L.A. Clippers", "LA Clippers", "L.A. Lakers")) %>%
  summarize(LA_AST_HOME = mean(AST_home), LA_AST_AWAY = mean(AST_away), LA_PTS_HOME =
   ↪  mean(PTS_home), LA_PTS_AWAY = mean(PTS_away), LA_REB_HOME = mean(REB_home),
   ↪  LA_REB_AWAY = mean(REB_away), LA_FGPCT_HOME = mean(FG_PCT_home), LA_FGPCT_AWAY =
   ↪  mean(FG_PCT_away), LA_FTPCT_HOME = mean(FT_PCT_home), LA_FTPCT_AWAY =
   ↪  mean(FT_PCT_away), LA_FG3PCT_HOME = mean(FG3_PCT_home), LA_FG3PCT_AWAY =
   ↪  mean(FG3_PCT_away))

NBALASUM %>%
 pivot_longer(c(LA_AST_HOME,LA_AST_AWAY, LA_PTS_HOME, LA_PTS_AWAY,LA_REB_HOME,
  ↪  LA_REB_AWAY), names_to = c("Stat", "Home"), names_pattern = "(.*)(....)$", values_to
  ↪  = "Value") %>%
  select(Stat, Home, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("Assists", "Points",
   ↪  "Rebounds")) + labs(title = "Average Assists, Points, and Rebounds for Home vs Away
   ↪  Team in Lakers/Clippers Matchups")
```

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




```
NBALASUM %>%
 pivot_longer(c(LA_FGPCT_HOME,LA_FGPCT_AWAY, LA_FTPCT_HOME, LA_FTPCT_AWAY,LA_FG3PCT_HOME,
 ↪  LA_FG3PCT_AWAY), names_to = c("Stat", "Home"), names_pattern = "(.*)(....)$",
 ↪  values_to = "Value") %>%
  select(Stat, Home, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("3P%", "FG%", "FT%")) + labs(x =
   ↪  "Stat", title = "Average FG%, FT%, and 3P% for Home vs Away Team in Lakers/Clippers
   ↪  Matchup")
```

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



From the bar graphs and box plots, we see that the home team in the matchup tends to record higher values for each of the statistics except for free throw percentage, which the away team records a higher percentage for. However, from the T-tests, we get that only assists and field goal percentage have a statistically significant differences (p-value < 0.05) between the home and away teams within the Lakers/Clippers matchup. The mean difference point differential between home and away teams in the matchup is 2.188596 points, which is not within the 95% confidence interval of 2.693226 to 3.032480 when we analyzed the mean difference between home points scored with away points scored for the entire data. The p-value of 0.172 is also greater than the level of significance (0.05) which suggests that the difference in points could be due to random chance and does not suggest a true difference in mean points. This is interesting because we've found positive associations between points and assists and between points and field goal percentage, both variables that are greater for the home team in this matchup than the away team. Even though the home team in this matchup tends to get more assists and shoot a higher percentage, we cannot conclude that they've won a greater proportion of games since our proportion test gave us a p-value of 0.4158 or scored more points. Thus, we can conclude that the arena environment is a factor that impacts giving home teams an advantage, as we do not see significant differences between the home and away teams across all Lakers and Clippers matchups from the 2003/2004 to 2021/2022 season.

Warriors vs Cavaliers Rivalry

For about four seasons, the Golden State Warriors and Cleveland Cavaliers dominated the NBA. This recent rivalry primarily focused on the four consecutive NBA Finals appearances between the two teams from 2015 to 2018. The four Finals matchups saw the Warriors win three (2015, 2017, and 2018) and the Cavaliers win one (2016). The rivalry was considered to have ended in 2018 following LeBron James' departure from the Cavaliers to the Los Angeles Lakers. Whenever the Warriors and Cavaliers played each other, the stadiums were always nearly if not completely full given the magnitude of the games. We will analyze if home court advantage played an impact in the Finals matchups between these two elite teams by comparing the home team's stats with the away team's stats and looking at the proportion of home wins.

We will first filter the NBA dataset in order to only look at the games where the Warriors and Cavs played each other from the 2014-2015 season to the 2017-2018 season, then filter out their regular season matchups to leave only their Finals games.

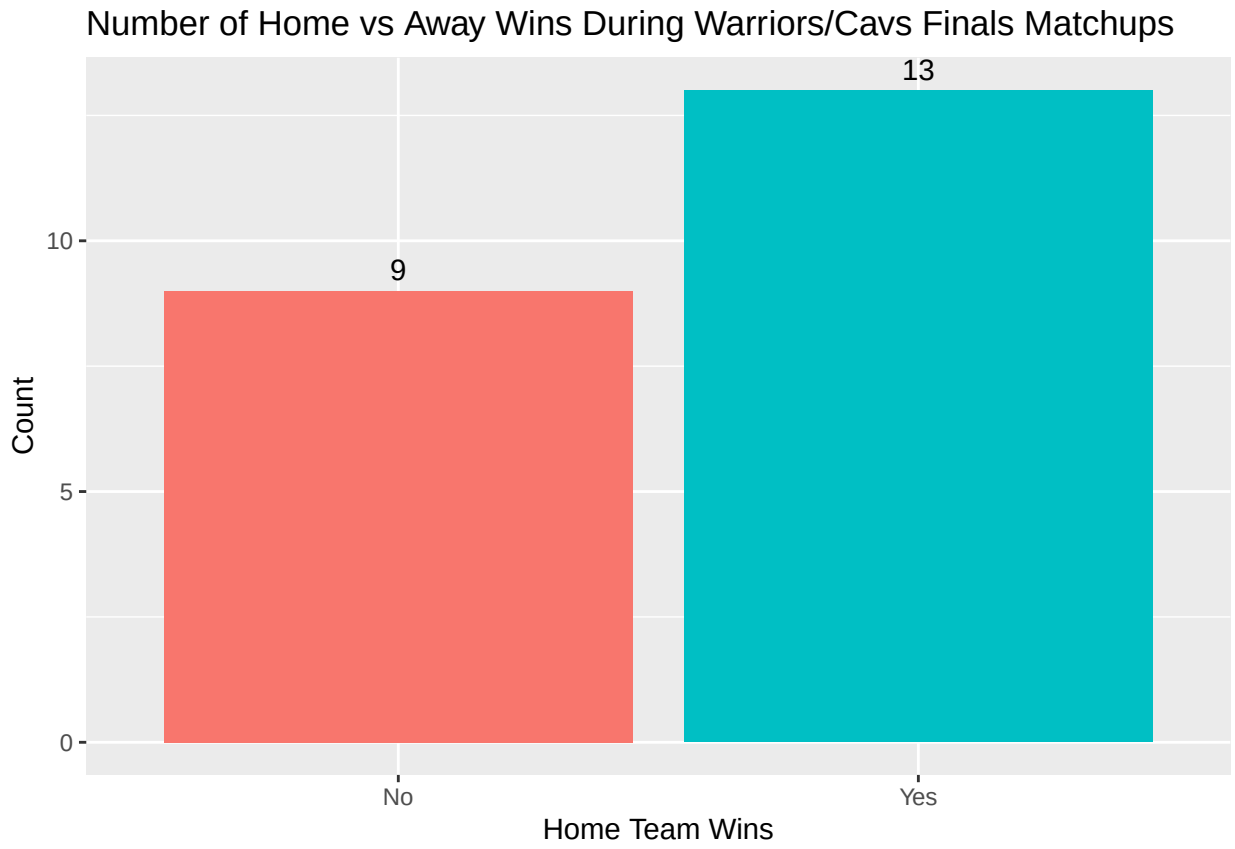
```
Warriors_Cavs <- NBA %>%
  filter(SEASON >= 2014, SEASON <= 2017) %>%
  filter(HOME_TEAM %in% c("Golden State", "Cleveland"), AWAY_TEAM %in% c("Golden State",
  ↪  "Cleveland")) %>%
  filter(GAME_DATE_EST != "2018-01-15") %>% filter(GAME_DATE_EST != "2017-12-25") %>%
  ↪  filter(GAME_DATE_EST != "2017-01-16") %>% filter(GAME_DATE_EST != "2016-12-25") %>%
  ↪  filter(GAME_DATE_EST != "2016-01-18") %>% filter(GAME_DATE_EST != "2015-12-25") %>%
  ↪  filter(GAME_DATE_EST != "2015-02-26") %>% filter(GAME_DATE_EST != "2015-01-09")
```

First, we can look at the proportion of home vs away wins in this matchup.

```
Warriors_Cavs %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0     9
## 2              1    13
```

```
Home_Team_Wins_WC <- c("No", "Yes")
Number_WC <- c(9, 13)
Home_wins_WC <- data.frame(Home_Team_Wins_WC, Number_WC)

Home_wins_WC %>%
  ggplot(aes(x = Home_Team_Wins_WC, y = Number_WC)) + geom_col(aes(fill =
  ↪  Home_Team_Wins_WC)) + geom_text(aes(label = Number_WC), vjust = -0.5) +
  ↪  theme(legend.position = "none") + labs(x = "Home Team Wins", y = "Count", title =
  ↪  "Number of Home vs Away Wins During Warriors/Cavs Finals Matchups")
```



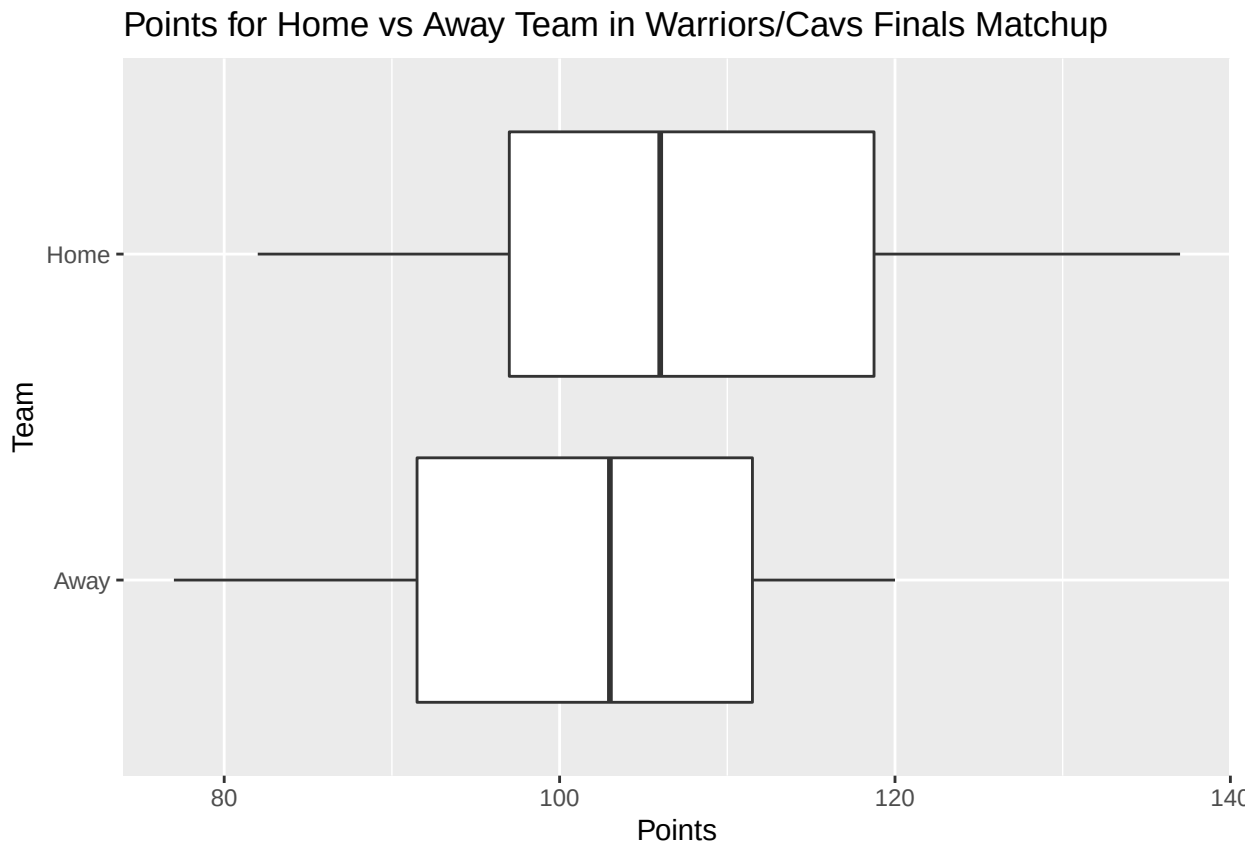

```
prop.test(13, 22, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  13 out of 22, null probability 0.5
## X-squared = 0.40909, df = 1, p-value = 0.5224
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3667993 0.7852364
## sample estimates:
##         p
## 0.5909091
```

While the bar graph above shows that the home team has won 13 games of the 22 matchups between these two teams in the Finals, the proportion test gives us a p-value of 0.5224, so we cannot conclude that there is a statistically significant difference between home and away win percentage. Next, we can make box plots and conduct matched pairs T-tests for each individual statistic to analyze the difference between home and away stats.

Points

```
Warriors_Cavs %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labels(title = "Points for Home vs Away Team in Warriors/Cavs
  ↪  Finals Matchup")
```

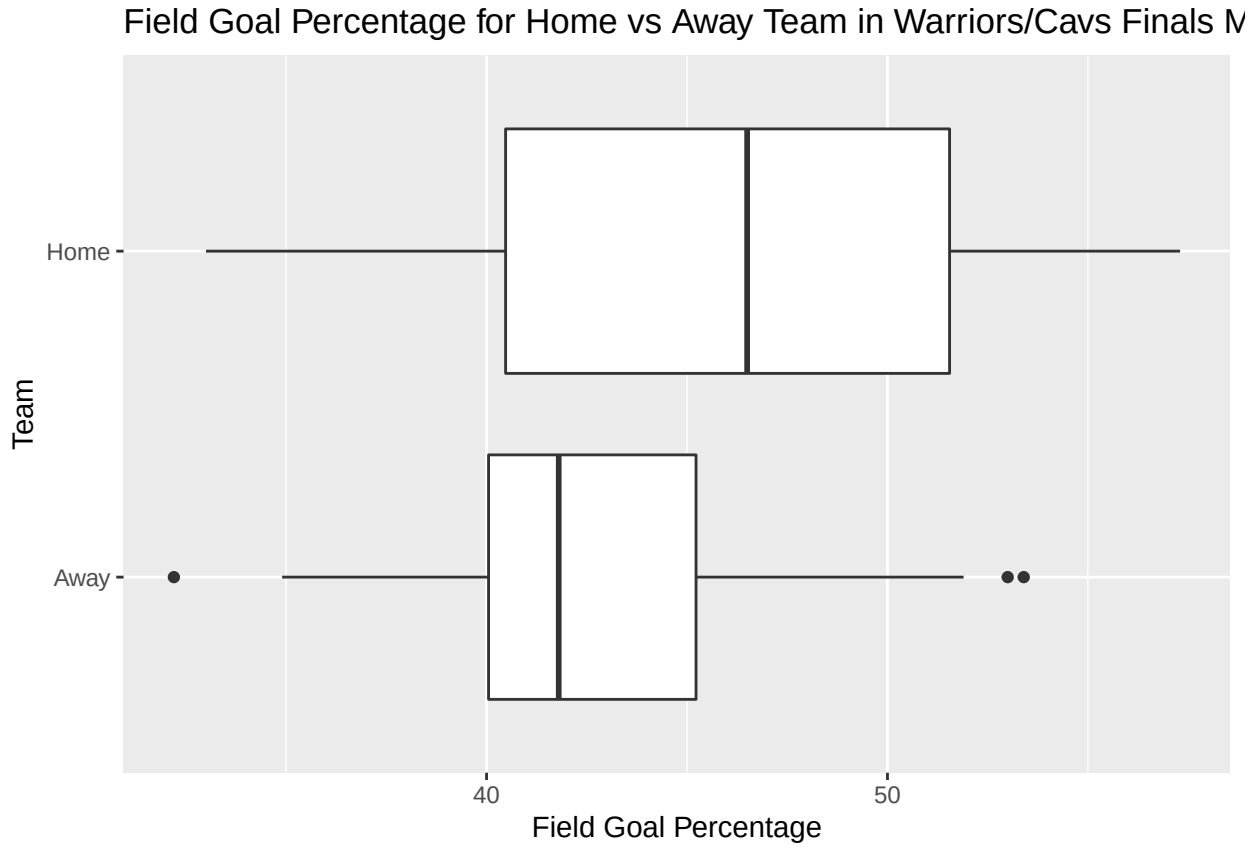



```
t.test(Warriors_Cavs$PTS_home, Warriors_Cavs$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Warriors_Cavs$PTS_home and Warriors_Cavs$PTS_away
## t = 1.6209, df = 21, p-value = 0.12
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.556626 12.556626
## sample estimates:
## mean difference
##             5.5
```

Field Goal Percentage

```
Warriors_Cavs %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Field Goal Percentage
  ↪  for Home vs Away Team in Warriors/Cavs Finals Matchup", x = "Field Goal
  ↪  Percentage")
```

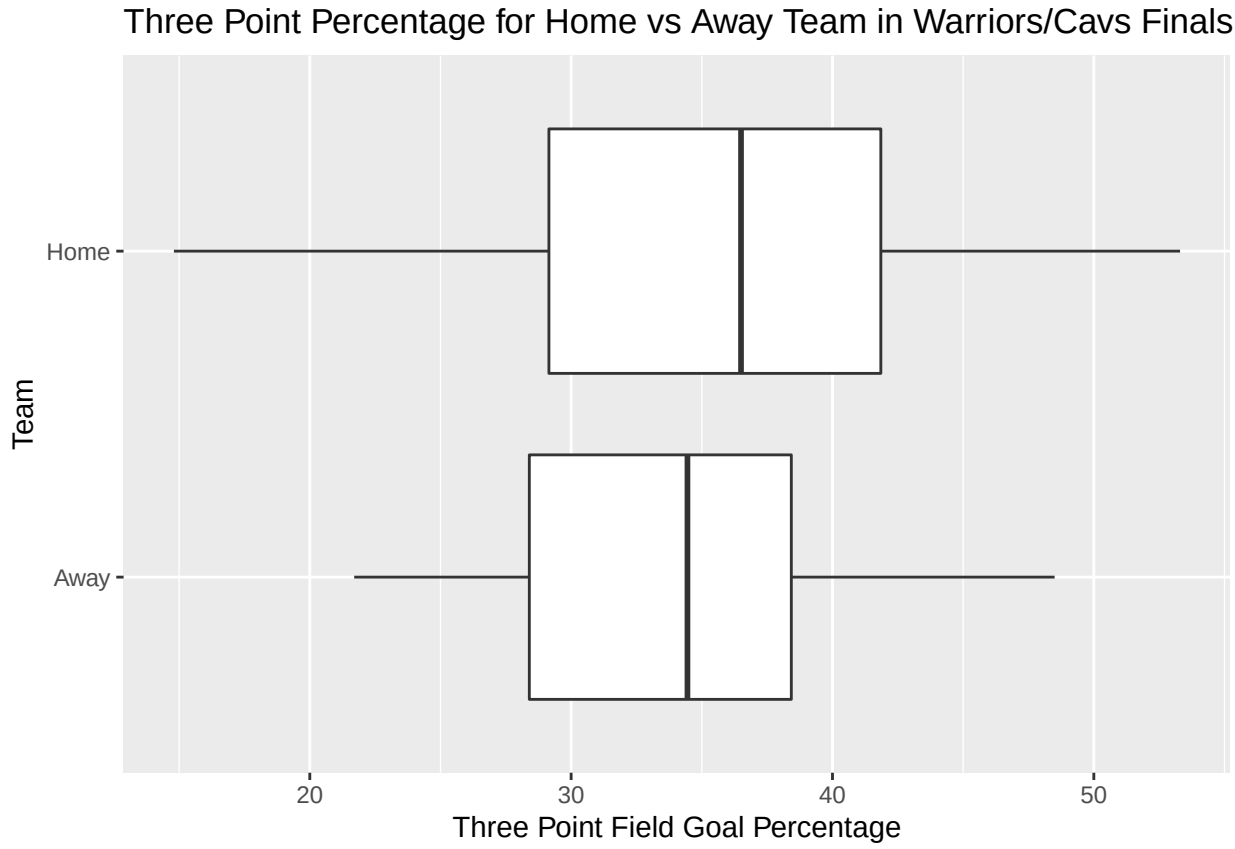



```
t.test(Warriors_Cavs$FG_PCT_home, Warriors_Cavs$FG_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Warriors_Cavs$FG_PCT_home and Warriors_Cavs$FG_PCT_away
## t = 1.5008, df = 21, p-value = 0.1483
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.177987  7.287078
## sample estimates:
## mean difference
##        3.054545
```

Three Point Percentage

```
Warriors_Cavs %>%
  pivot_longer(c(FG3_PCT_away, FG3_PCT_home), names_to = "Team", values_to =
  ↪  "Three_Point_Percentage") %>%
  ggplot(aes(x = Three_Point_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Three Point Percentage
  ↪  for Home vs Away Team in Warriors/Cavs Finals Matchup", x = "Three Point Field Goal
  ↪  Percentage")
```

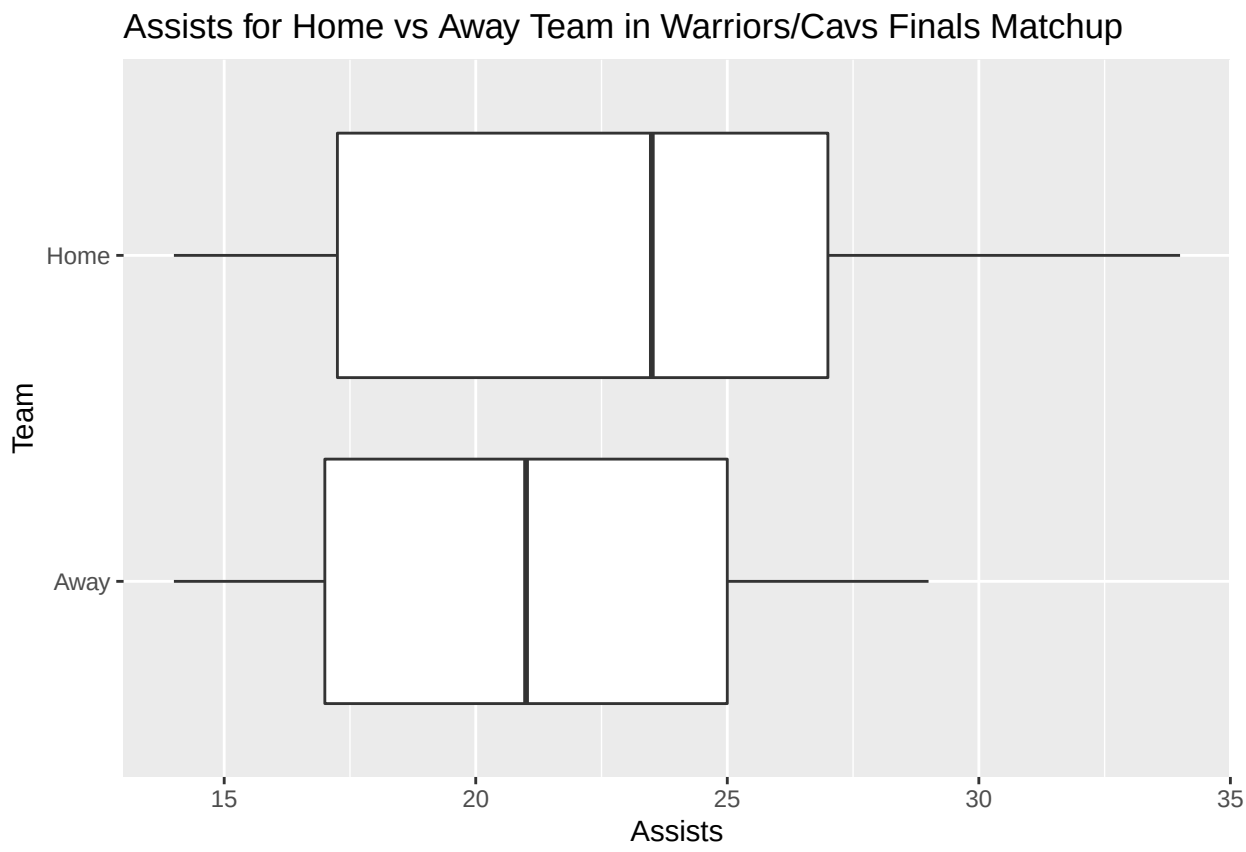



```
t.test(Warriors_Cavs$FG3_PCT_home, Warriors_Cavs$FG3_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Warriors_Cavs$FG3_PCT_home and Warriors_Cavs$FG3_PCT_away
## t = 0.21455, df = 21, p-value = 0.8322
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -5.887503  7.242048
## sample estimates:
## mean difference
##       0.6772727
```

Assists

```
Warriors_Cavs %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Team in Warriors/Cavs
  ↪  Finals Matchup")
```

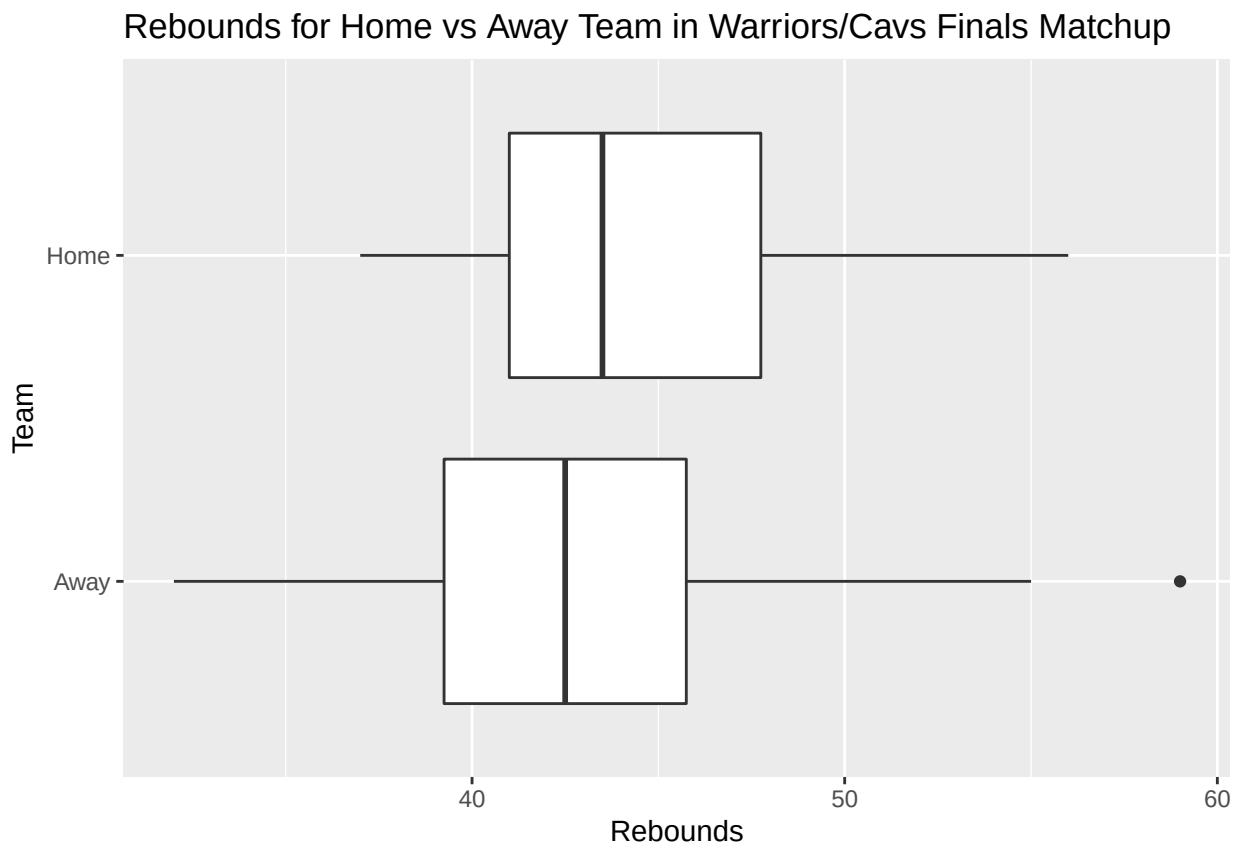



```
t.test(Warriors_Cavs$AST_home, Warriors_Cavs$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Warriors_Cavs$AST_home and Warriors_Cavs$AST_away
## t = 1.0077, df = 21, p-value = 0.3251
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -1.885823  5.431278
## sample estimates:
## mean difference
##        1.772727
```

Rebounds

```
Warriors_Cavs %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Team in Warriors/Cavs
  ↪  Finals Matchup")
```

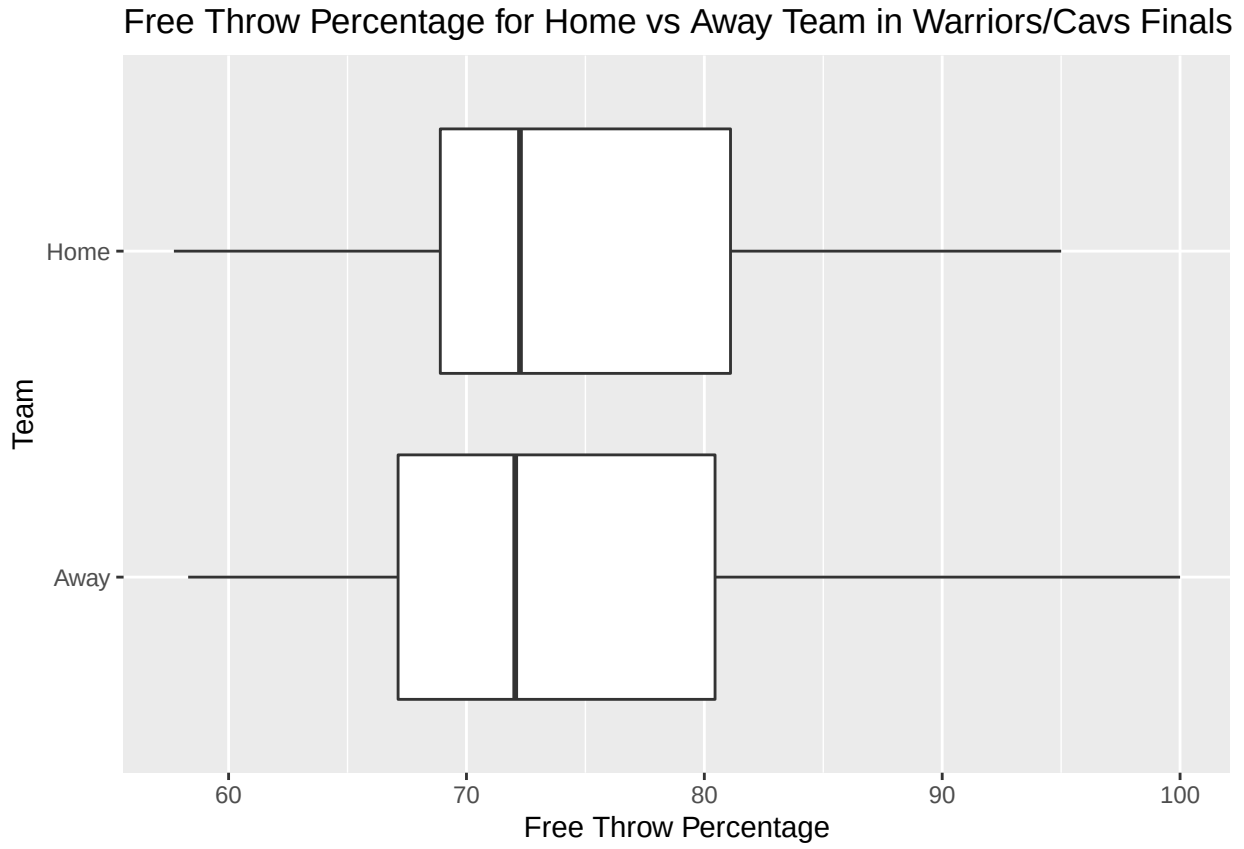



```
t.test(Warriors_Cavs$REB_home, Warriors_Cavs$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  Warriors_Cavs$REB_home and Warriors_Cavs$REB_away
## t = 0.85567, df = 21, p-value = 0.4018
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -2.405639  5.769275
## sample estimates:
## mean difference
##        1.681818
```

Free Throw Percentage

```
Warriors_Cavs %>%
  pivot_longer(c(FT_PCT_away, FT_PCT_home), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Free Throw Percentage
  ↪  for Home vs Away Team in Warriors/Cavs Finals Matchup",  x = "Free Throw
  ↪  Percentage")
```




```
t.test(Warriors_Cavs$FT_PCT_home, Warriors_Cavs$FT_PCT_away, paired = TRUE)
```

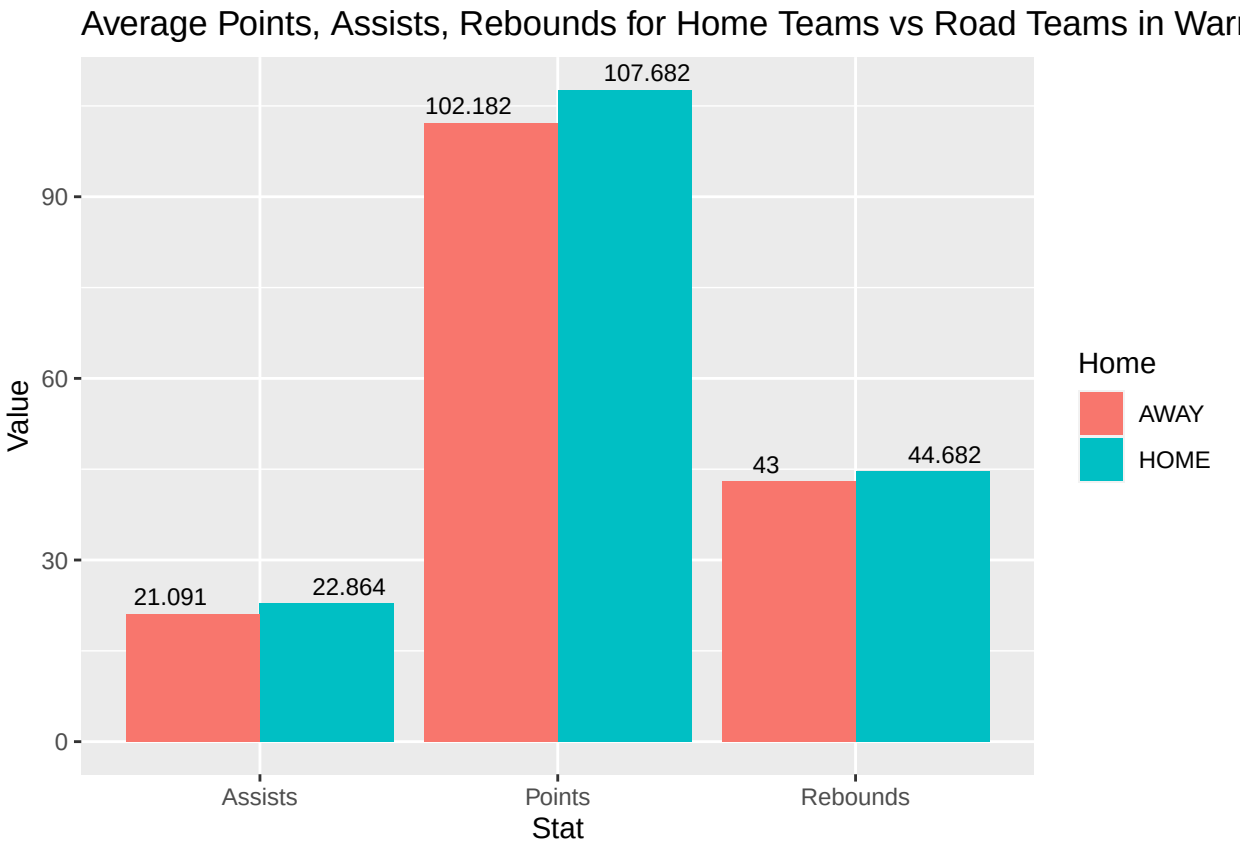
```
##
##  Paired t-test
##
## data:  Warriors_Cavs$FT_PCT_home and Warriors_Cavs$FT_PCT_away
## t = 0.031088, df = 21, p-value = 0.9755
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -6.289932  6.480841
## sample estimates:
## mean difference
##      0.09545455
```

Creating Bar Charts

```
WARRIORS_CAVS_SUMMARY <- Warriors_Cavs %>%
  summarize(MEAN_AST_HOME = mean(AST_home), MEAN_AST_AWAY = mean(AST_away),
  ↪  MEAN_POINTS_HOME = mean(PTS_home), MEAN_POINTS_AWAY = mean(PTS_away), MEAN_REB_HOME
  ↪  = mean(REB_home), MEAN_REB_AWAY = mean(REB_away), MEAN_FGPCT_HOME =
  ↪  mean(FG_PCT_home), MEAN_FGPCT_AWAY = mean(FG_PCT_away), MEAN_FTPCT_HOME =
  ↪  mean(FT_PCT_home), MEAN_FTPCT_AWAY = mean(FT_PCT_away), MEAN_FG3PCT_HOME =
  ↪  mean(FG3_PCT_home), MEAN_FG3PCT_AWAY = mean(FG3_PCT_away))
```

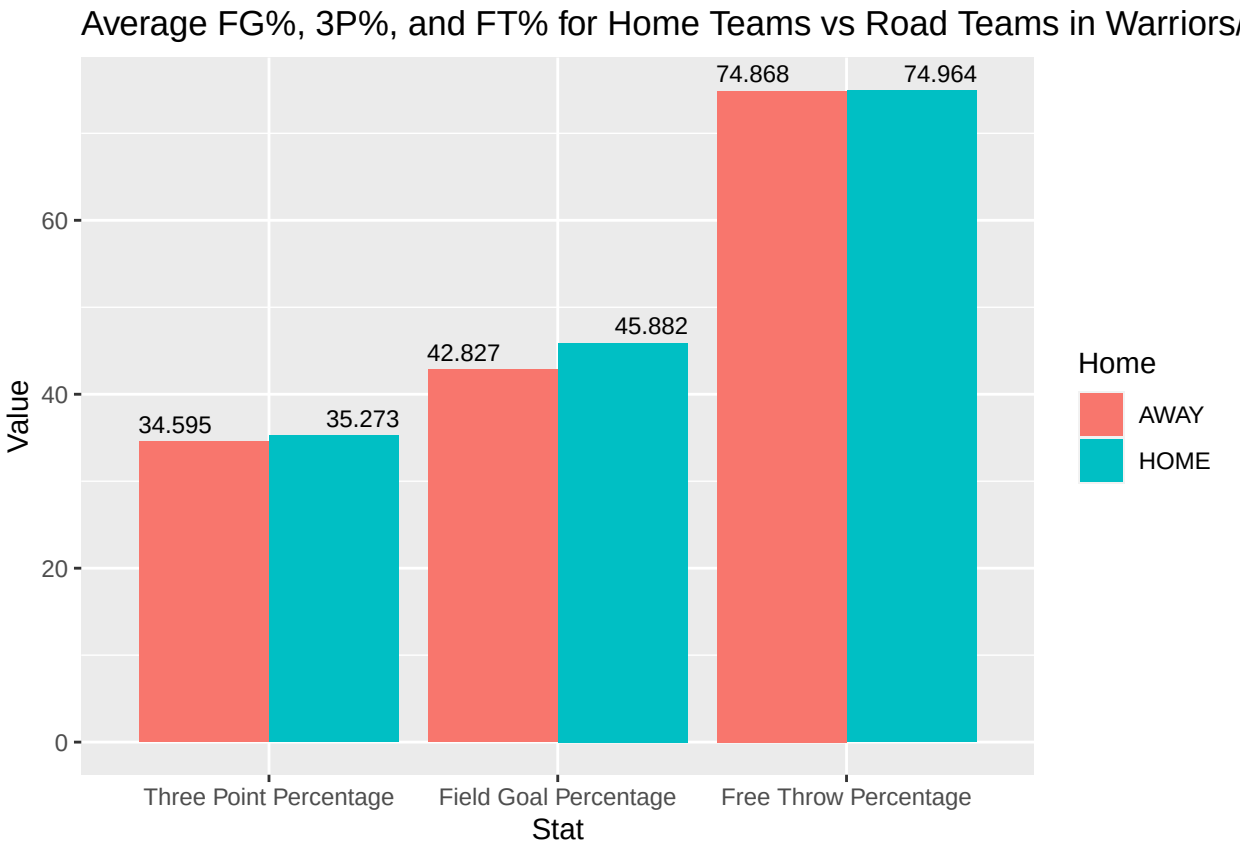
```
WARRIORS_CAVS_SUMMARY %>%
  pivot_longer(c(MEAN_AST_HOME,MEAN_AST_AWAY, MEAN_POINTS_HOME, MEAN_POINTS_AWAY,
  ↪  MEAN_REB_HOME, MEAN_REB_AWAY), names_to = c("Stat", "Home"), names_pattern =
  ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Home, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
  ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
  ↪  position_dodge(1.2)) + scale_x_discrete(labels = c("Assists", "Points",
  ↪  "Rebounds")) + labs(title = "Average Points, Assists, Rebounds for Home Teams vs
  ↪  Road Teams in Warriors/Cavs Finals Matchups")
```

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



```
WARRIORS_CAVS_SUMMARY %>%
  pivot_longer(c(MEAN_FGPCT_HOME, MEAN_FGPCT_AWAY, MEAN_FTPCT_HOME, MEAN_FTPCT_AWAY,
   ↪  MEAN_FG3PCT_HOME, MEAN_FG3PCT_AWAY), names_to = c("Stat", "Home"), names_pattern =
   ↪  "(.*)(....)$", values_to = "Value") %>%
  select(Stat, Home, Value) %>%
  ggplot(aes(x = Stat, y = Value, fill = Home)) + geom_col(position = "dodge") +
   ↪  geom_text(aes(label = round(Value, 3)), vjust = -0.5, size = 3, position =
   ↪  position_dodge(1.3)) + scale_x_discrete(labels = c("Three Point Percentage", "Field
   ↪  Goal Percentage", "Free Throw Percentage")) + labs(title = "Average FG%, 3P%, and
   ↪  FT% for Home Teams vs Road Teams in Warriors/Cavs Finals Matchups")
```

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



Looking at the boxplots created above, we can see that the first quartile, median, and third quartile for all of the statistics are higher for the home team versus the away team. Taken together with the bar charts, we can also see that the gap between the home and away teams for many of the stats is larger than the general league averages computed earlier on in the report. For example, from the T-test, we see that the mean difference in points between home and away teams is 5.5 points in this matchup compared to around 2.86 points which we calculated for the entire data set. Home teams in the matchup also shoot about 3% better from the field, compared to a league average of just 1.11% from our earlier analysis on the entire data. However, the p-values generated from all of the t-tests are far greater than 0.05, so we cannot conclude that there is a statistically significant difference between the stats of home teams and away teams in this matchup. However, since we see that the mean difference between home and away teams in points and field goal percentage are significantly greater in this matchup compared to our results from analyzing the entire data, we want to investigate whether home court advantage is even more important in the playoffs. Since

the NBA playoffs is a best of 7 series for each matchup, the higher seeded team has "home court advantage" as they would play Game 7 at home if the series goes the distance. Is the idea of home court advantage even greater in the playoffs?

Playoffs vs Regular Season

From the previous example, we observed that their may be a more significant difference between home and away performance in the playoffs compared to the regular season, so we want to investigate in more detail. This topic is especially important because a lot of NBA teams are resting their stars for several games in the regular season, prioritizing their health over playoff seeding. This raises the question of the actual importance of home court advantage in the playoffs.

First, we will filter for the dates of the playoffs and the regular season separately.

```
NBAPLAYOFFS <- NBA %>%
  filter(GAME_DATE_EST >= "2004-04-17" & GAME_DATE_EST <= "2004-06-15" | GAME_DATE_EST >=
   ↪  "2005-04-23" & GAME_DATE_EST <= "2005-06-23" | GAME_DATE_EST >= "2006-04-22" &
   ↪  GAME_DATE_EST <= "2006-06-20" | GAME_DATE_EST >= "2007-04-21" & GAME_DATE_EST <=
   ↪  "2007-06-14" | GAME_DATE_EST >= "2008-04-19" & GAME_DATE_EST <= "2008-06-17" |
   ↪  GAME_DATE_EST >= "2009-04-18" & GAME_DATE_EST <= "2009-06-14" | GAME_DATE_EST >=
   ↪  "2010-04-17" & GAME_DATE_EST <= "2010-06-17" | GAME_DATE_EST >= "2011-04-16" &
   ↪  GAME_DATE_EST <= "2011-06-12" |  GAME_DATE_EST >= "2012-04-28" & GAME_DATE_EST <=
   ↪  "2012-06-21" | GAME_DATE_EST >= "2013-04-20" & GAME_DATE_EST <= "2013-06-20" |
   ↪  GAME_DATE_EST >= "2014-04-19" & GAME_DATE_EST <= "2014-06-15" | GAME_DATE_EST >=
   ↪  "2015-04-18" & GAME_DATE_EST <= "2015-06-16" | GAME_DATE_EST >= "2016-04-16" &
   ↪  GAME_DATE_EST <= "2016-06-19" | GAME_DATE_EST >= "2017-04-15" & GAME_DATE_EST <=
   ↪  "2017-06-12" | GAME_DATE_EST >= "2018-04-14" & GAME_DATE_EST <= "2018-06-08" |
   ↪  GAME_DATE_EST >= "2019-04-13" & GAME_DATE_EST <= "2019-06-13" | GAME_DATE_EST >=
   ↪  "2020-08-17" & GAME_DATE_EST <= "2020-10-11" | GAME_DATE_EST >= "2021-05-22" &
   ↪  GAME_DATE_EST <= "2021-07-20" | GAME_DATE_EST >= "2022-04-16" & GAME_DATE_EST <=
   ↪  "2022-06-16") %>%
  mutate(Playoffs = "Yes")
```

```
NBAREGULARSEASON <- NBA %>%
filter(GAME_DATE_EST >= "2003-10-28" & GAME_DATE_EST <= "2004-04-14" | GAME_DATE_EST >=
↪  "2004-11-02" & GAME_DATE_EST <= "2005-04-20" | GAME_DATE_EST >= "2005-11-01" &
↪  GAME_DATE_EST <= "2006-04-19" | GAME_DATE_EST >= "2006-10-31" & GAME_DATE_EST <=
↪  "2007-04-18" | GAME_DATE_EST >= "2007-10-30" & GAME_DATE_EST <= "2008-04-16" |
↪  GAME_DATE_EST >= "2008-10-28" & GAME_DATE_EST <= "2009-04-16" | GAME_DATE_EST >=
↪  "2009-10-27" & GAME_DATE_EST <= "2010-04-14" | GAME_DATE_EST >= "2010-10-26" &
↪  GAME_DATE_EST <= "2011-04-13" |  GAME_DATE_EST >= "2011-12-25" & GAME_DATE_EST <=
↪  "2012-04-26" | GAME_DATE_EST >= "2012-10-30" & GAME_DATE_EST <= "2013-04-17" |
↪  GAME_DATE_EST >= "2013-10-29" & GAME_DATE_EST <= "2014-04-16" | GAME_DATE_EST >=
↪  "2014-10-27" & GAME_DATE_EST <= "2015-04-15" | GAME_DATE_EST >= "2015-10-27" &
↪  GAME_DATE_EST <= "2016-04-13" | GAME_DATE_EST >= "2016-10-25" & GAME_DATE_EST <=
↪  "2017-04-12" | GAME_DATE_EST >= "2017-10-17" & GAME_DATE_EST <= "2018-04-11" |
↪  GAME_DATE_EST >= "2018-10-16" & GAME_DATE_EST <= "2019-04-10" | GAME_DATE_EST >=
↪  "2019-10-22" & GAME_DATE_EST <= "2020-08-14" | GAME_DATE_EST >= "2020-12-22" &
↪  GAME_DATE_EST <= "2021-05-16" | GAME_DATE_EST >= "2021-10-19" & GAME_DATE_EST <=
↪  "2022-04-10") %>%
  mutate(Playoffs = "No")
```

Side Note: We aren't counting play-in games for this particular analysis because the NBA does not count them as regular season or playoff games

```
NBAPLAYIN <- NBA %>%
  filter(GAME_DATE_EST == "2020-08-15" | GAME_DATE_EST >= "2021-05-18" & GAME_DATE_EST <=
  ↪  "2021-05-21" | GAME_DATE_EST >= "2022-04-12" & GAME_DATE_EST <= "2022-04-15")
```

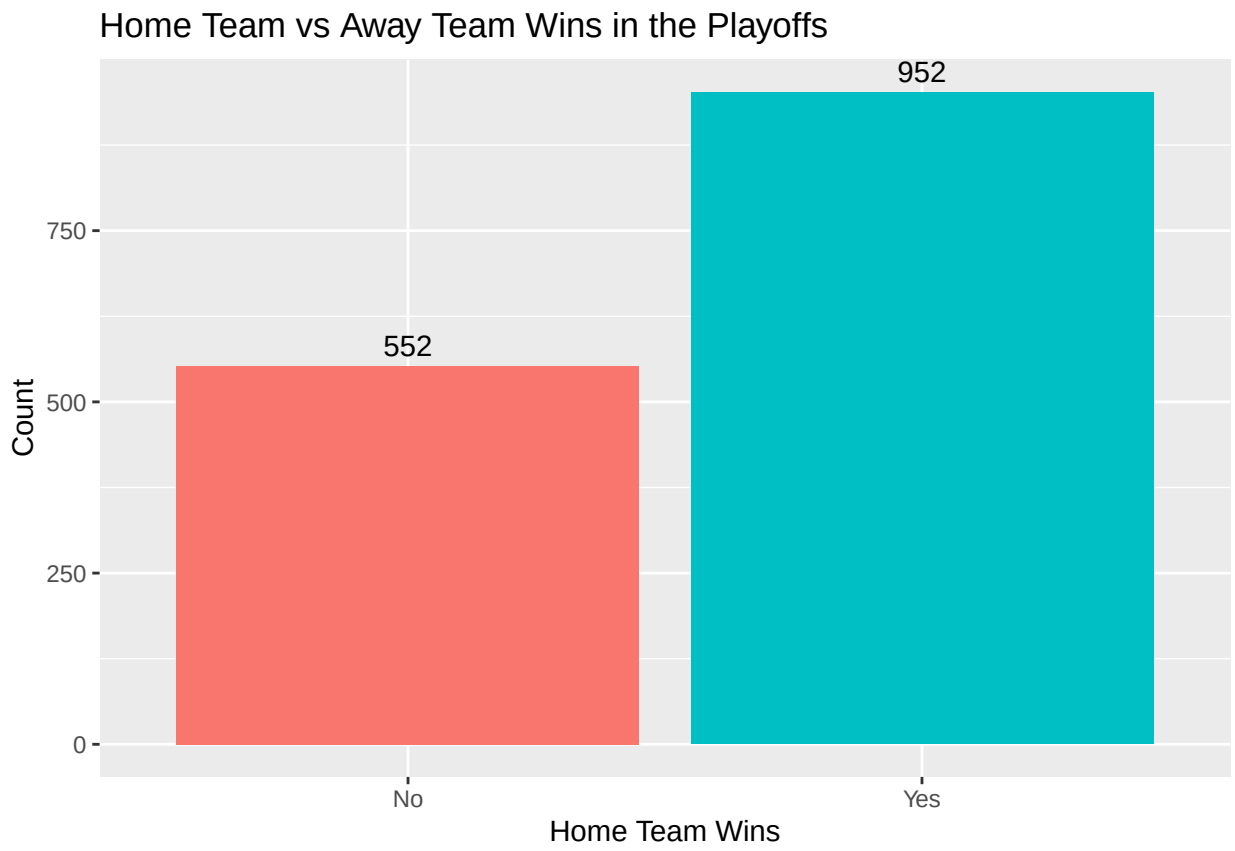
Now, we can compare the home and away stats for the regular season and playoffs separately, then compare the results.

Wins

```
NBAPLAYOFFS %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0   552
## 2              1   952
```

```
P_Home_Team_Wins <- c("No", "Yes")
P_Number <- c(552, 952)
P_Home_wins <- data.frame(P_Home_Team_Wins, P_Number)

P_Home_wins %>%
  ggplot(aes(x = P_Home_Team_Wins, y = P_Number, fill = P_Home_Team_Wins)) + geom_col() +
  ↪  geom_text(aes(label = P_Number), vjust = -0.5) + labs(x = "Home Team Wins", y =
  ↪  "Count", title = "Home Team vs Away Team Wins in the Playoffs") +
  ↪  theme(legend.position = "none")
```




```
prop.test(952, 1504, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  952 out of 1504, null probability 0.5
## X-squared = 105.85, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6079731 0.6572975
## sample estimates:
##         p
## 0.6329787
```

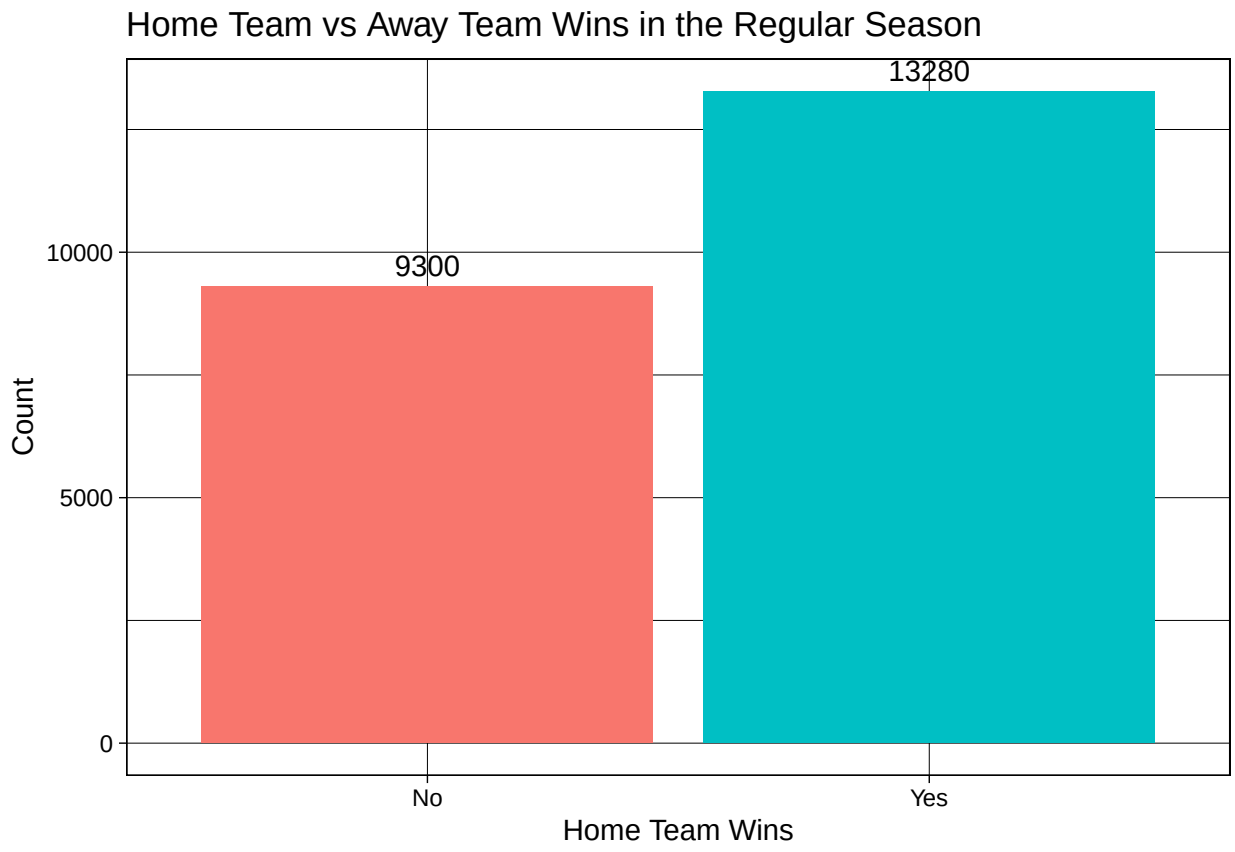
```
NBAREGULARSEASON %>%
  count(HOME_TEAM_WINS)
```

```
## # A tibble: 2 x 2
##   HOME_TEAM_WINS     n
##            <dbl> <int>
## 1              0  9300
## 2              1 13280
```

```r
prop.test(13280, 22580, p = 0.5)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  13280 out of 22580, null probability 0.5
## X-squared = 701.17, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5816749 0.5945572
## sample estimates:
##         p
## 0.5881311
```

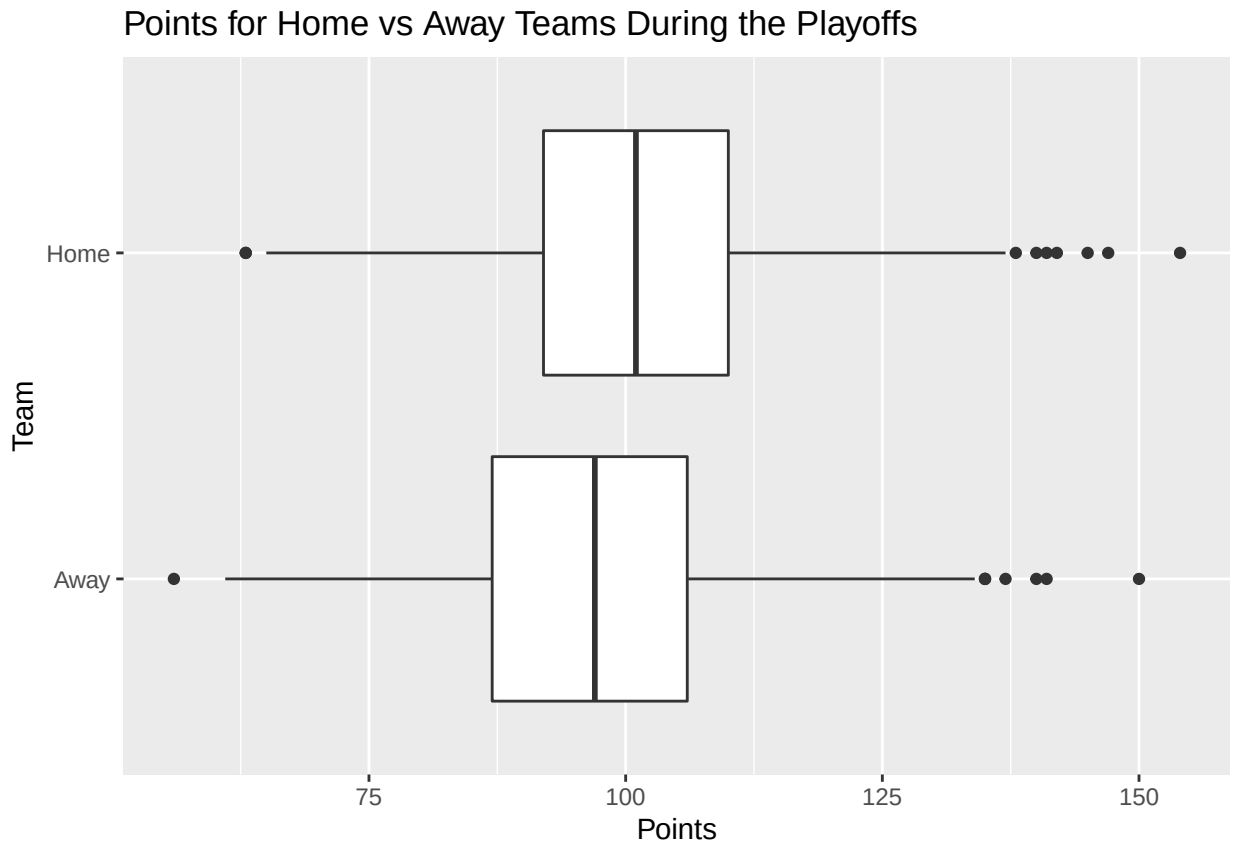
```r
Home_Team_Wins <- c("No", "Yes")
Number <- c(9300, 13280)
Home_wins <- data.frame(Home_Team_Wins, Number)

Home_wins %>%
  ggplot(aes(x = Home_Team_Wins, y = Number)) + geom_col(aes(fill = Home_Team_Wins)) +
  ↪  geom_text(aes(label = Number), vjust = -0.5) + theme_linedraw() + labs(x = "Home
  ↪  Team Wins", y = "Count", title = "Home Team vs Away Team Wins in the Regular
  ↪  Season") + theme(legend.position = "none")
```



Both during the regular season and playoffs, there is a statistically significant difference between the proportion of home wins from the expected proportion of 0.5 if home court advantage didn't exist as both of the proportion tests generate a p-value below 2.2 x 10^-16. However, the difference is exemplified in the playoffs, as we are 95% confident that the true proportion of home wins lies between 0.6079731 and 0.6572975, whereas we are 95% confident that the true proportion of home wins lies between 0.5816749 and 0.5945572 in the regular season.

Points

```
NBAPLAYOFFS %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Points for Home vs Away Teams During the
  ↪  Playoffs")
```

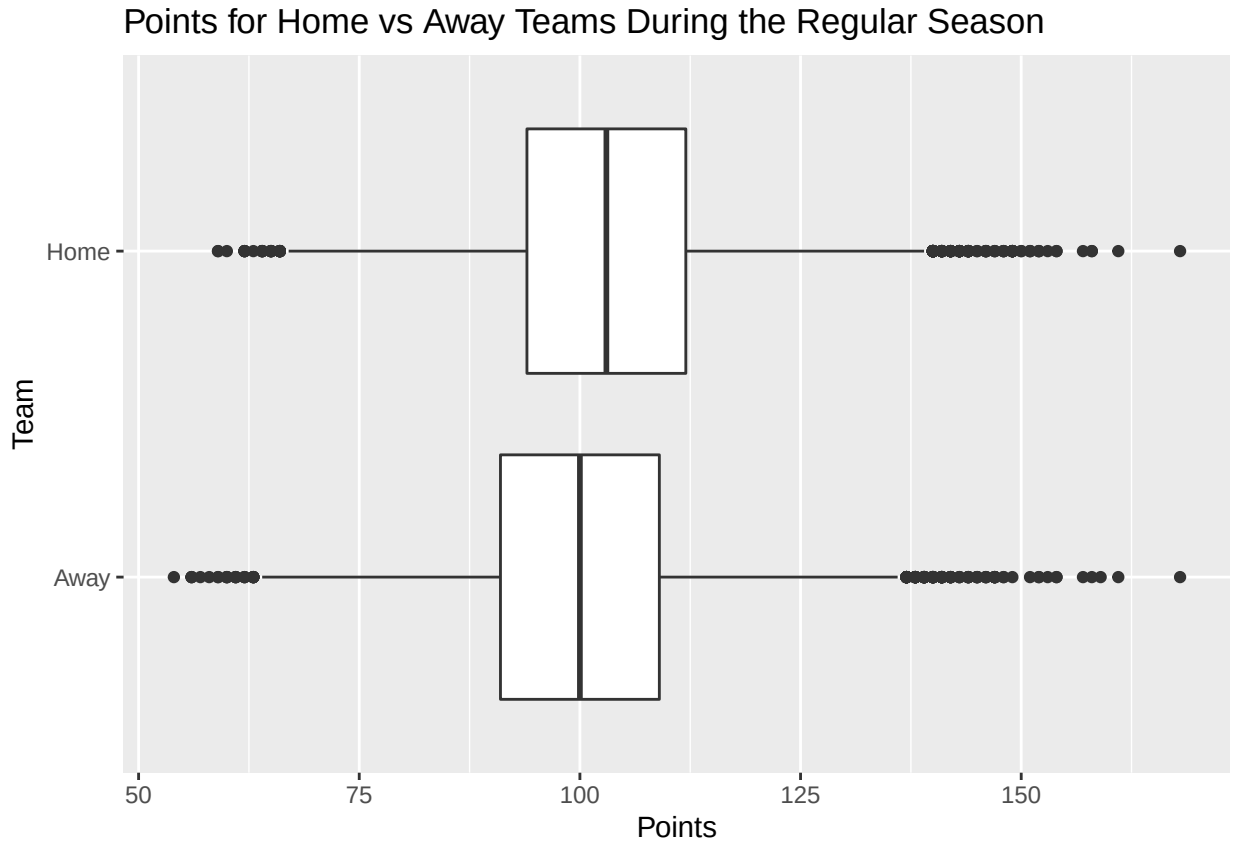



```
t.test(NBAPLAYOFFS$PTS_home, NBAPLAYOFFS$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAPLAYOFFS$PTS_home and NBAPLAYOFFS$PTS_away
## t = 12.479, df = 1503, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
```

```
## 95 percent confidence interval:
##  3.706941 5.089602
## sample estimates:
## mean difference
##        4.398271
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  ggplot(aes(x = Points, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Points for Home vs Away Teams During the Regular
  ↪  Season")
```



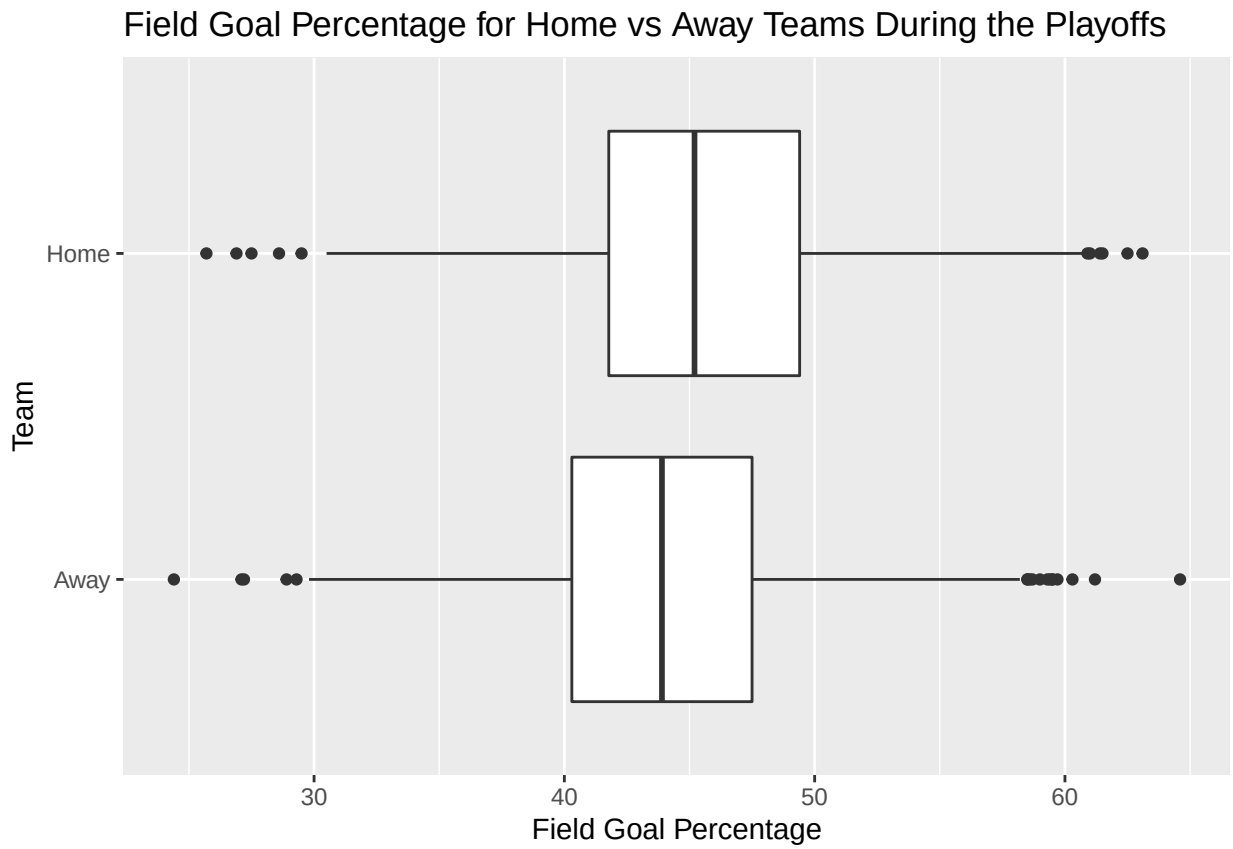

```
t.test(NBAREGULARSEASON$PTS_home, NBAREGULARSEASON$PTS_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$PTS_home and NBAREGULARSEASON$PTS_away
## t = 30.905, df = 22579, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  2.583093 2.932938
## sample estimates:
## mean difference
##        2.758016
```

Both during the regular season and playoffs, there is a statistically significant difference between the points scored for home vs away teams as both T-tests generate a p-value below 2.2 x 10^-16. However, the difference is once again exemplified in the playoffs, as we are 95% confident that the true difference between points for home vs away teams lies between 3.706941 and 5.089602, whereas we are 95% confident that the true difference between points for home vs away teams lies between 2.583093 and 2.932938 in the regular season. While we previously estimated a general advantage of around 2.86 points for home teams, our analysis suggests that in the playoffs specifically, the home team would have an advantage of around 4.398271 points. This can be attributed to the fact that most playoff games are sold out and feature a wild crowd environment because of the magnitude of the games.

Field Goal Percentage

```
NBAPLAYOFFS %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Field Goal Percentage
  ↪  for Home vs Away Teams During the Playoffs", x = "Field Goal Percentage")
```

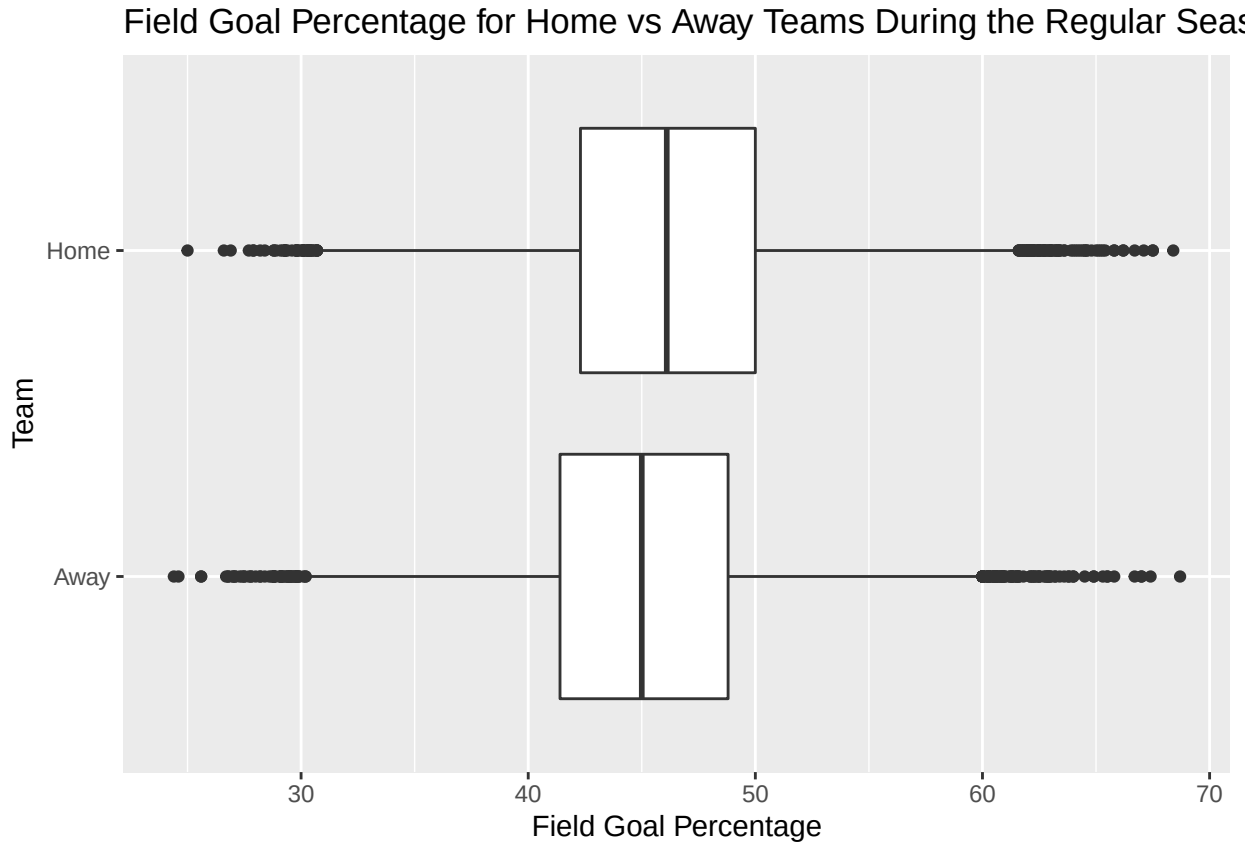



```
t.test(NBAPLAYOFFS$FG_PCT_home, NBAPLAYOFFS$FG_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
```

```
##
## data:  NBAPLAYOFFS$FG_PCT_home and NBAPLAYOFFS$FG_PCT_away
## t = 7.1134, df = 1503, p-value = 1.746e-12
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.041539 1.834658
## sample estimates:
## mean difference
##        1.438098
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to =
  ↪  "Field_Goal_Percentage") %>%
  ggplot(aes(x = Field_Goal_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Field Goal Percentage
  ↪  for Home vs Away Teams During the Regular Season", x = "Field Goal Percentage")
```




```
t.test(NBAREGULARSEASON$FG_PCT_home, NBAREGULARSEASON$FG_PCT_away, paired = TRUE)
```

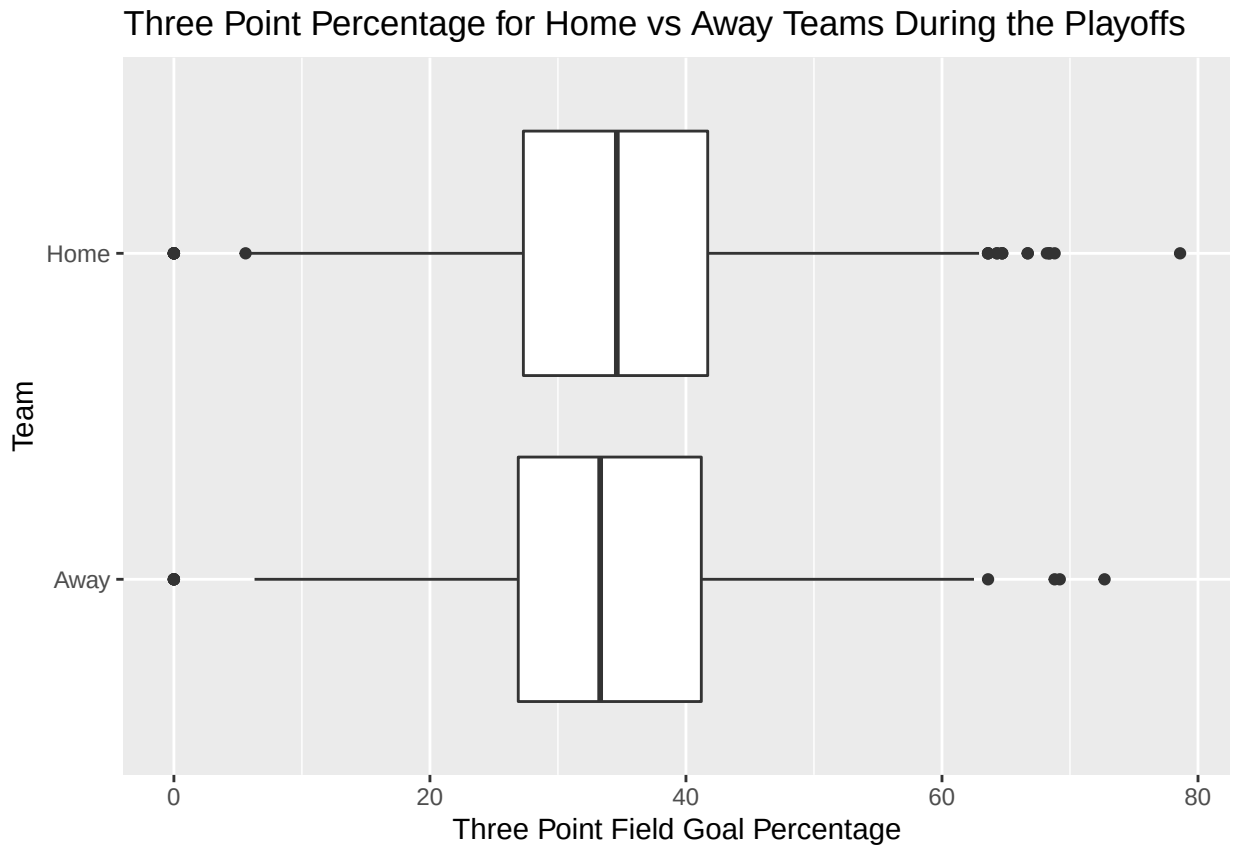
```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$FG_PCT_home and NBAREGULARSEASON$FG_PCT_away
## t = 21.154, df = 22579, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.9889438 1.1909233
## sample estimates:
## mean difference
##        1.089934
```

Once again, we see that there is a statistically significant difference between field goal percentage for home vs away teams in both the regular season and playoffs because both T-tests generate a p-value that is nearly 0 (1.746 x 10ˆ-12 for playoffs, less than 2.2 x 10ˆ-16 for regular season). However, the difference is greater in the playoffs as we are 95% confident that the true mean difference for field goal percentage between home and away teams in the playoffs is between 1.041539% and 1.834658%, whereas in the regular season, we are 95% confident that the true mean difference for field goal percentage between home and away teams is between 0.9889438% and 1.1909233%. While there is some overlap in the intervals, the interval for the playoffs is slightly greater.

Three Point Percentage

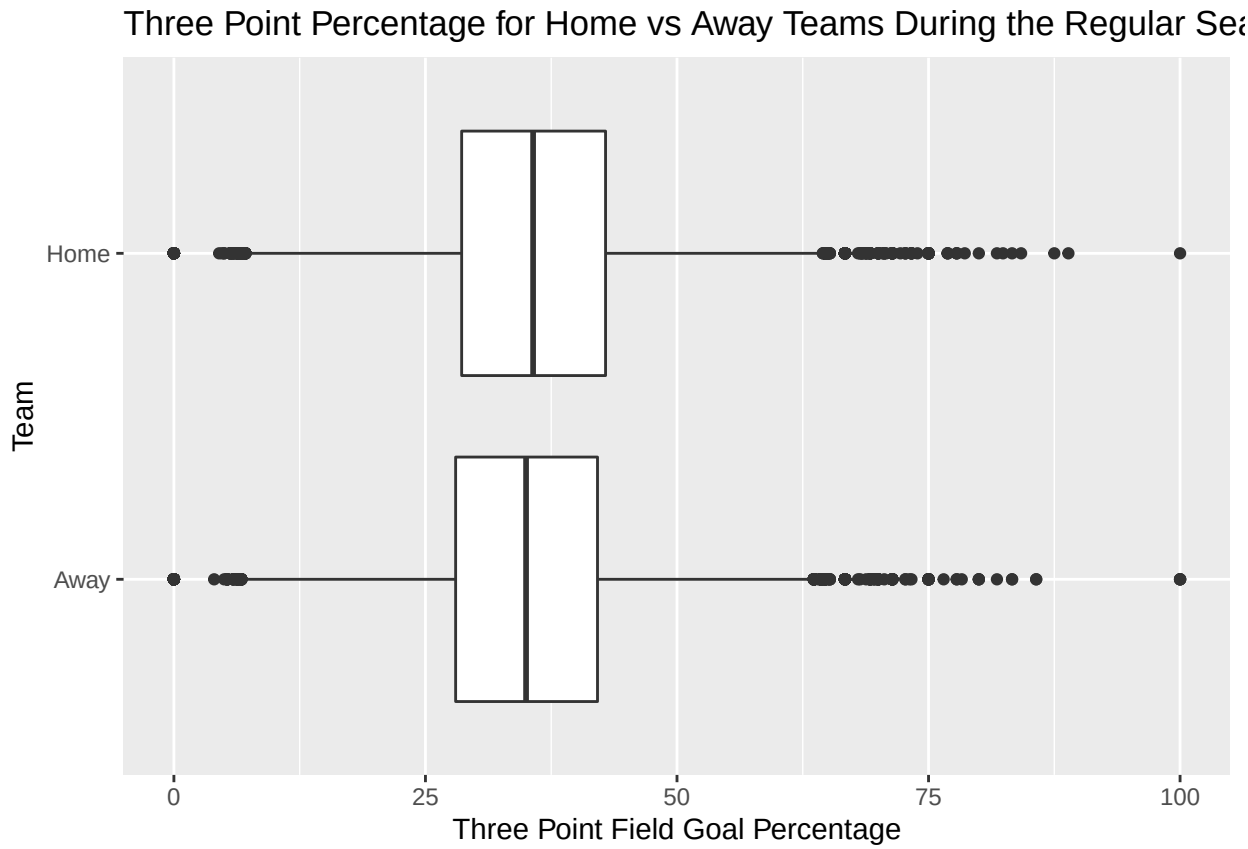
```
NBAPLAYOFFS %>%
  pivot_longer(c(FG3_PCT_away, FG3_PCT_home), names_to = "Team", values_to =
  ↪  "Three_Point_Percentage") %>%
  ggplot(aes(x = Three_Point_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Three Point Percentage
  ↪  for Home vs Away Teams During the Playoffs", x = "Three Point Field Goal
  ↪  Percentage")
```




```
t.test(NBAPLAYOFFS$FG3_PCT_home, NBAPLAYOFFS$FG3_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAPLAYOFFS$FG3_PCT_home and NBAPLAYOFFS$FG3_PCT_away
## t = 1.408, df = 1503, p-value = 0.1593
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.2151566  1.3098374
## sample estimates:
## mean difference
##       0.5473404
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(FG3_PCT_away, FG3_PCT_home), names_to = "Team", values_to =
  ↪  "Three_Point_Percentage") %>%
  ggplot(aes(x = Three_Point_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Three Point Percentage
  ↪  for Home vs Away Teams During the Regular Season", x = "Three Point Field Goal
  ↪  Percentage")
```



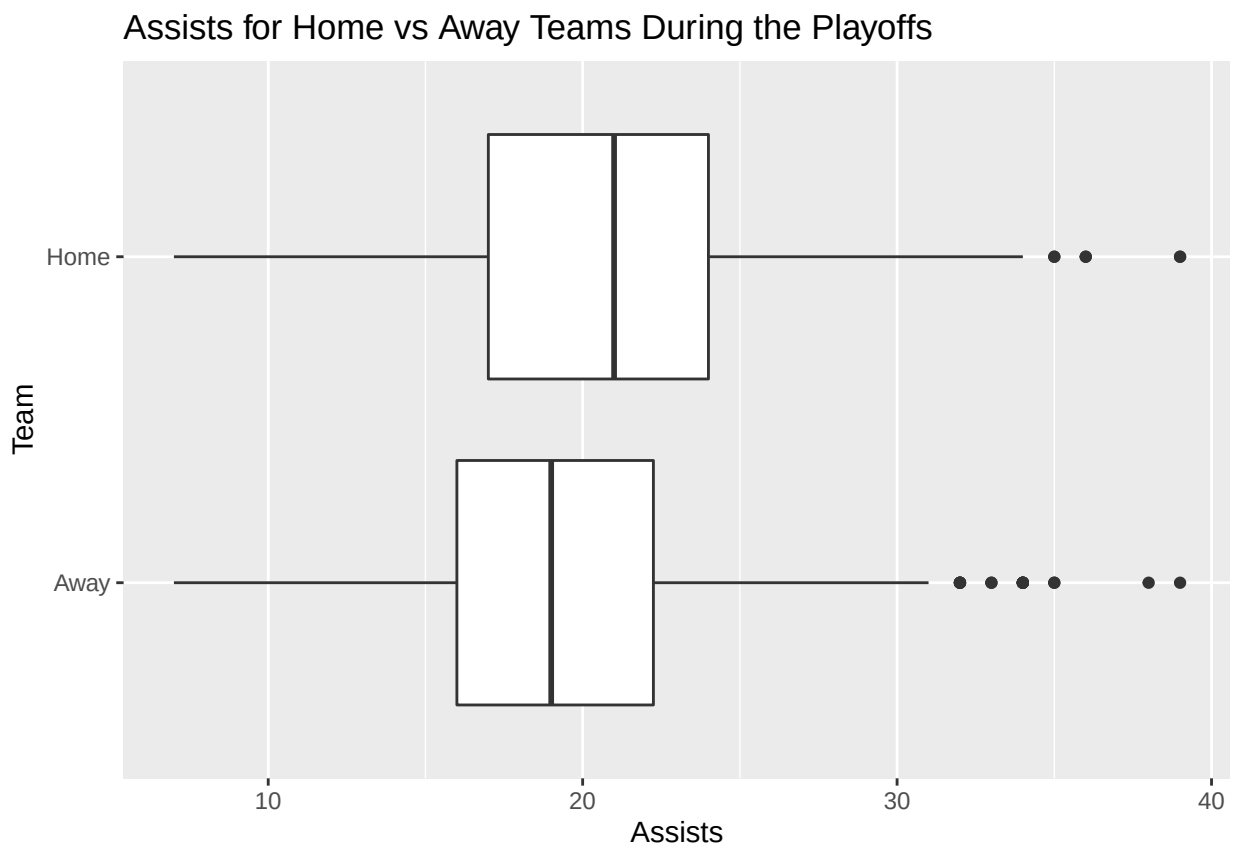

```
t.test(NBAREGULARSEASON$FG3_PCT_home, NBAREGULARSEASON$FG3_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$FG3_PCT_home and NBAREGULARSEASON$FG3_PCT_away
## t = 6.3167, df = 22579, p-value = 2.722e-10
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.4556785 0.8657032
## sample estimates:
## mean difference
##       0.6606909
```

While we see a statistically significant difference in three point percentage for home vs away teams in the regular season (p-value = 2.722 x 10^-10), we do not see a statistically significant difference in three point percentage for home vs away teams in the playoffs (p-value = 0.1593).

Assists

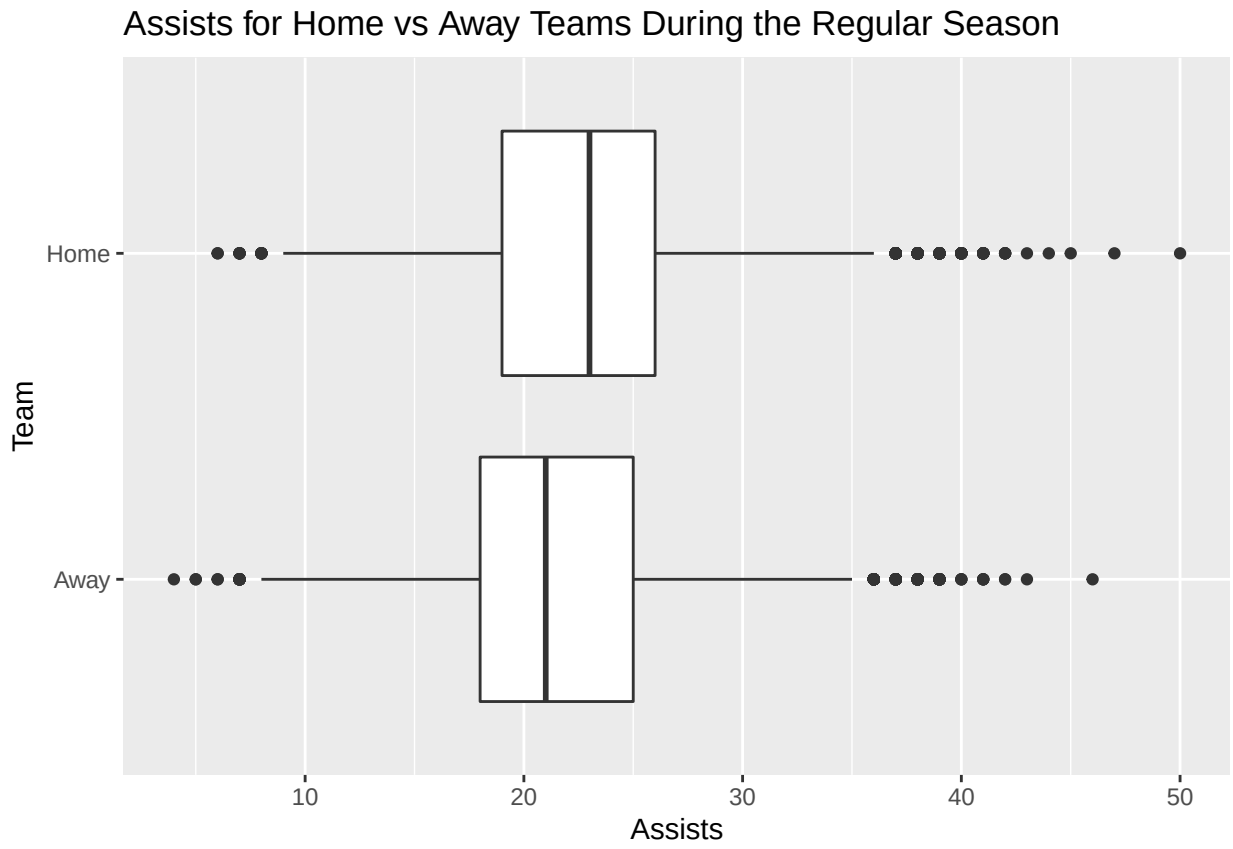
```
NBAPLAYOFFS %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Teams During the
  ↪  Playoffs")
```




```
t.test(NBAPLAYOFFS$AST_home, NBAPLAYOFFS$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAPLAYOFFS$AST_home and NBAPLAYOFFS$AST_away
## t = 9.8929, df = 1503, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.305463 1.951186
## sample estimates:
## mean difference
##        1.628324
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(AST_away, AST_home), names_to = "Team", values_to = "Assists") %>%
  ggplot(aes(x = Assists, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Assists for Home vs Away Teams During the
  ↪  Regular Season")
```



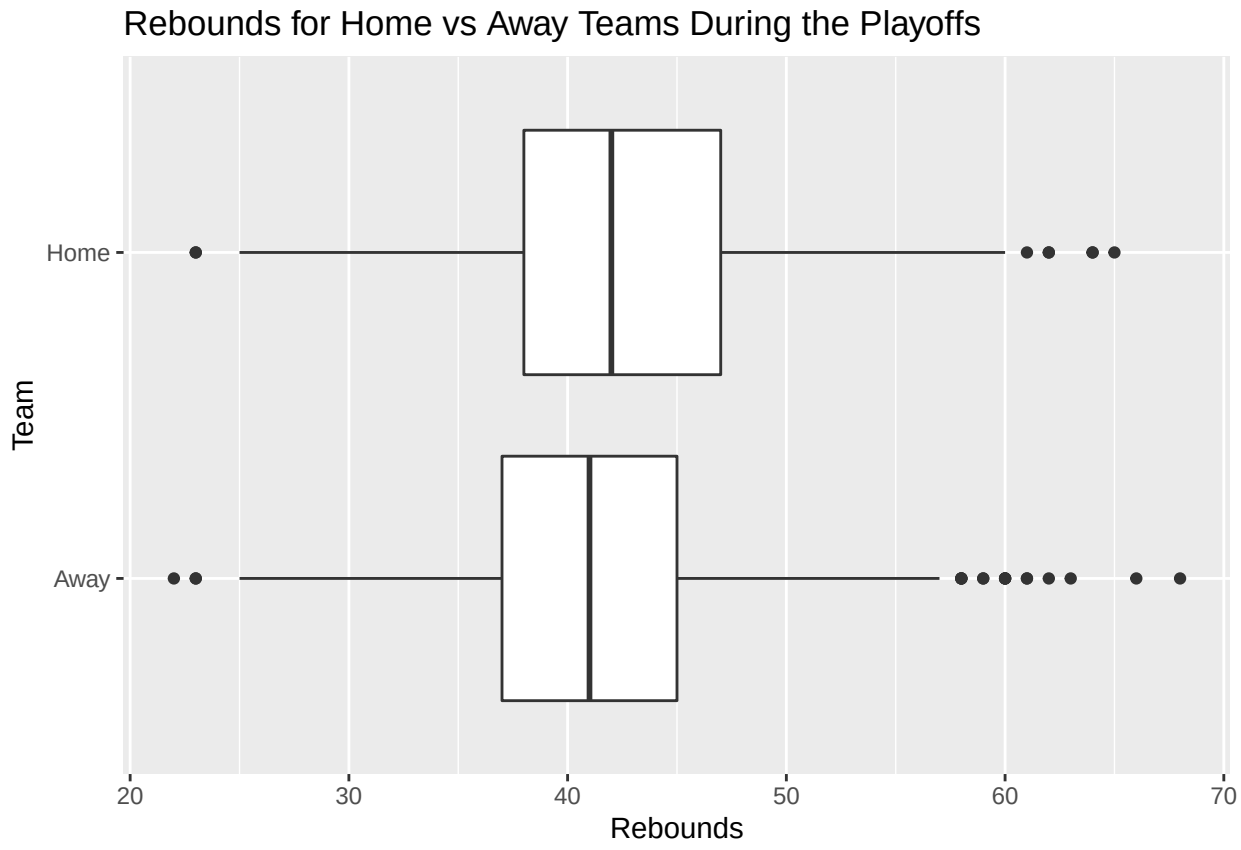

```
t.test(NBAREGULARSEASON$AST_home, NBAREGULARSEASON$AST_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$AST_home and NBAREGULARSEASON$AST_away
## t = 29.177, df = 22579, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.231383 1.408741
## sample estimates:
## mean difference
##        1.320062
```

Like points and field goal percentage, we see that there is a statistically significant difference between assists for home vs away teams in both the regular season and playoffs because both T-tests generate a p-value less than 2.2 x 10^-16. However, the difference is greater in the playoffs as we are 95% confident that the true mean difference for assists between home and away teams in the playoffs is between 1.305463 and 1.951186, whereas in the regular season, we are 95% confident that the true mean difference for assists between home and away teams is between 1.231383 and 1.408741.

Rebounds

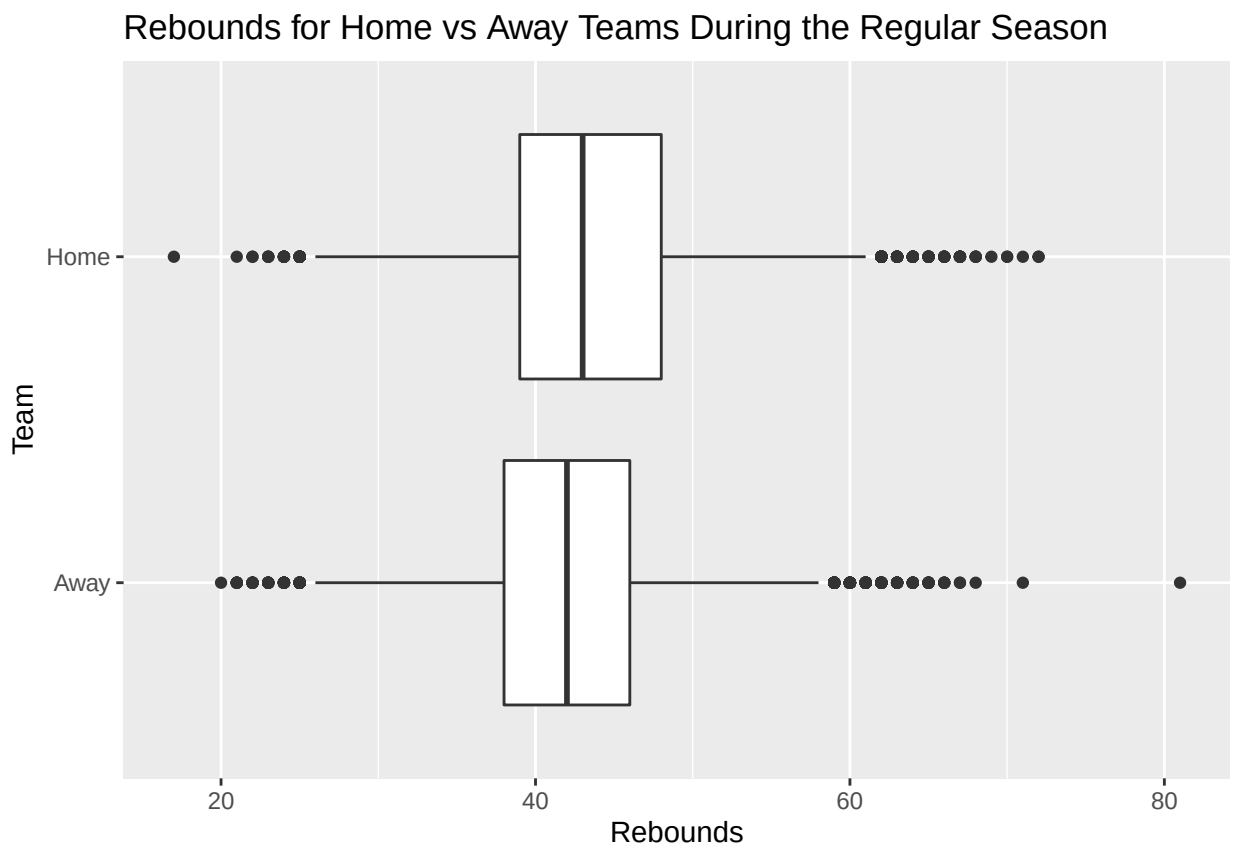
```
NBAPLAYOFFS %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Teams During the
  ↪  Playoffs")
```




```
t.test(NBAPLAYOFFS$REB_home, NBAPLAYOFFS$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAPLAYOFFS$REB_home and NBAPLAYOFFS$REB_away
## t = 7.292, df = 1503, p-value = 4.918e-13
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.203918 2.089965
## sample estimates:
## mean difference
##        1.646941
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(REB_away, REB_home), names_to = "Team", values_to = "Rebounds") %>%
  ggplot(aes(x = Rebounds, y = Team)) + geom_boxplot() + scale_y_discrete(labels =
  ↪  c("Away", "Home")) + labs(title = "Rebounds for Home vs Away Teams During the
  ↪  Regular Season")
```



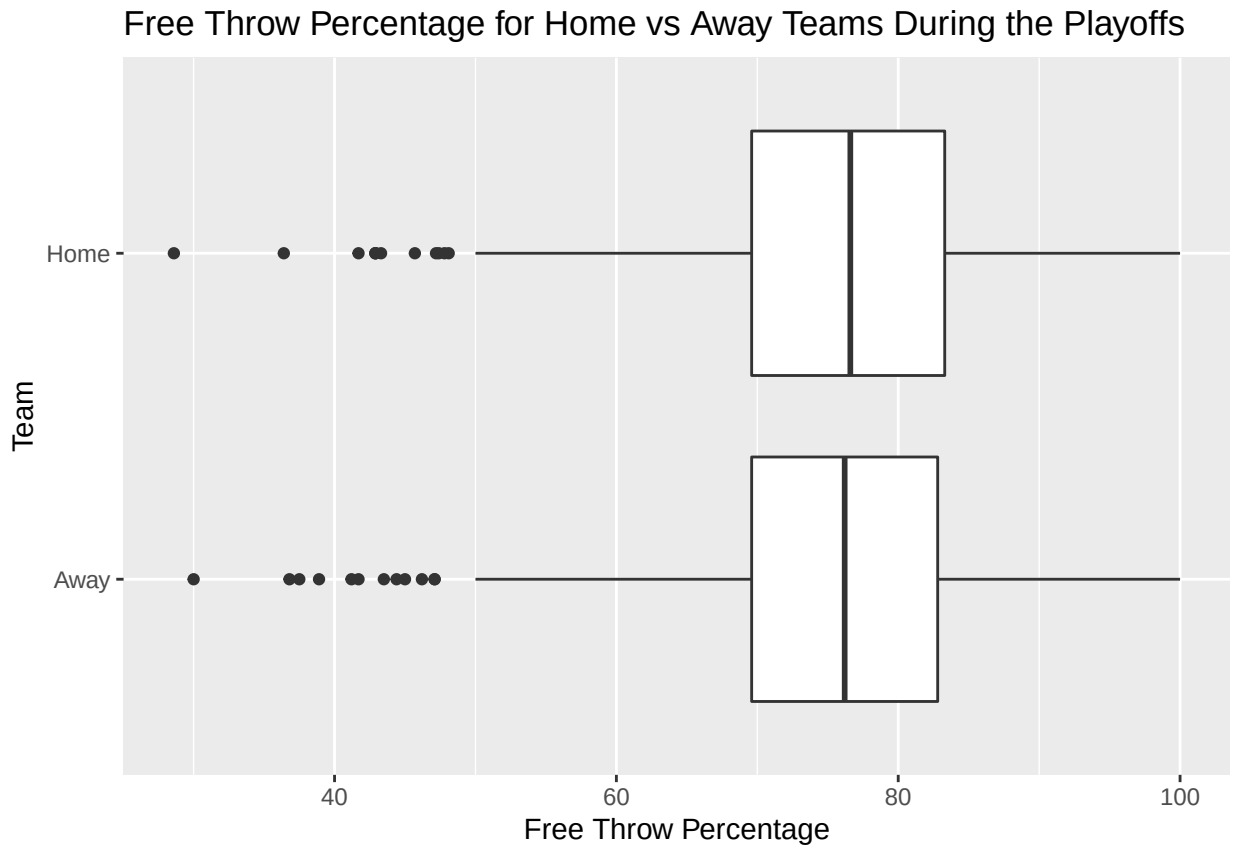

```
t.test(NBAREGULARSEASON$REB_home, NBAREGULARSEASON$REB_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$REB_home and NBAREGULARSEASON$REB_away
## t = 21.181, df = 22579, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  1.149073 1.383433
## sample estimates:
## mean difference
##        1.266253
```

With rebounds as well, we see that there is a statistically significant difference between rebounds for home vs away teams in both the regular season and playoffs because both T-tests generate a p-value that is nearly 0 (4.918 x 10^-13 for playoffs, less than 2.2 x 10^-16 for regular season). However, the difference is greater in the playoffs as we are 95% confident that the true mean difference for rebounds between home and away teams in the playoffs is between 1.203918 and 2.089965, whereas in the regular season, we are 95% confident that the true mean difference for assists between home and away teams is between 1.149073 and 1.383433.

Free Throw Percentage

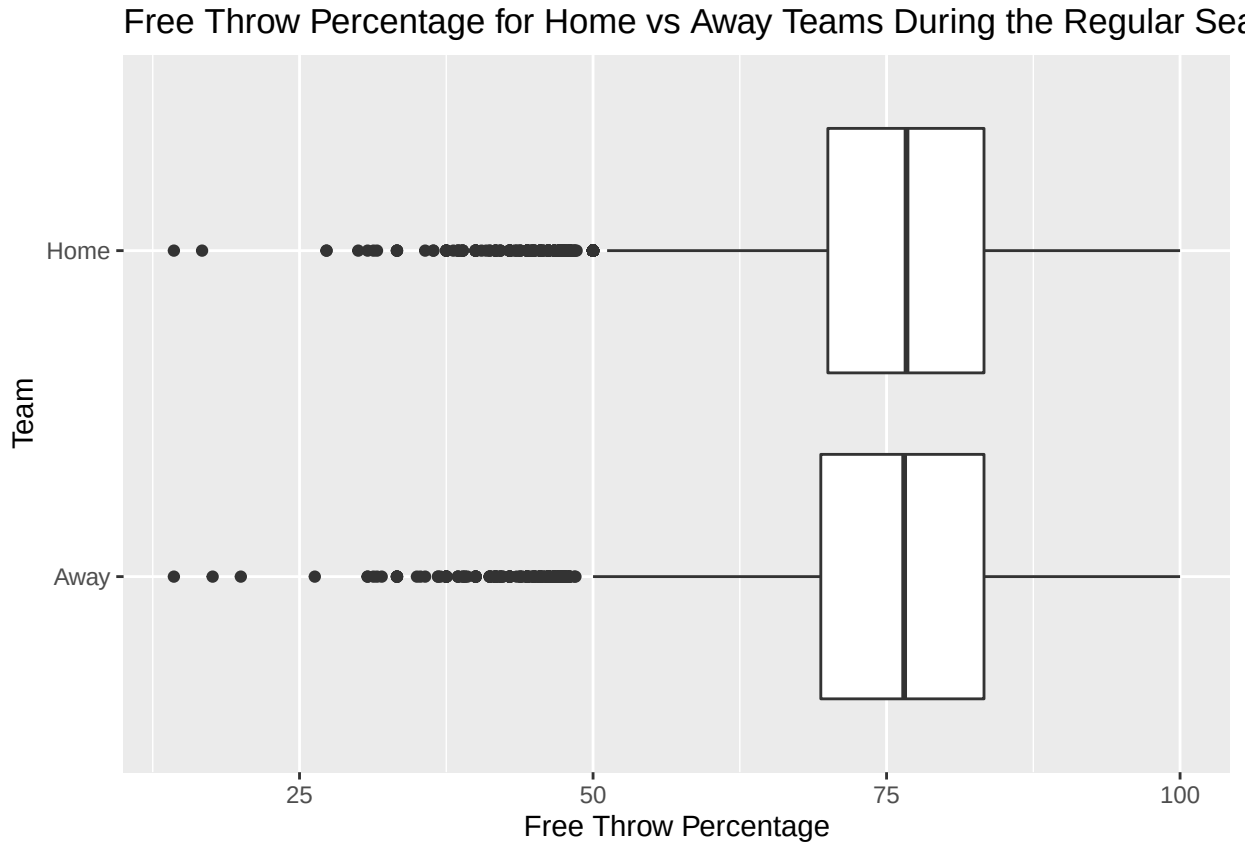
```
NBAPLAYOFFS %>%
  pivot_longer(c(FT_PCT_away, FT_PCT_home), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Free Throw Percentage
  ↪  for Home vs Away Teams During the Playoffs", x = "Free Throw Percentage")
```




```
t.test(NBAPLAYOFFS$FT_PCT_home, NBAPLAYOFFS$FT_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAPLAYOFFS$FT_PCT_home and NBAPLAYOFFS$FT_PCT_away
## t = 0.74304, df = 1503, p-value = 0.4576
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.4548928  1.0096800
## sample estimates:
## mean difference
##       0.2773936
```

```
NBAREGULARSEASON %>%
  pivot_longer(c(FT_PCT_away, FT_PCT_home), names_to = "Team", values_to =
  ↪  "Freethrow_Percentage") %>%
  ggplot(aes(x = Freethrow_Percentage, y = Team)) + geom_boxplot() +
  ↪  scale_y_discrete(labels = c("Away", "Home")) + labs(title = "Free Throw Percentage
  ↪  for Home vs Away Teams During the Regular Season", x = "Free Throw Percentage")
```




```
t.test(NBAREGULARSEASON$FT_PCT_home, NBAREGULARSEASON$FT_PCT_away, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  NBAREGULARSEASON$FT_PCT_home and NBAREGULARSEASON$FT_PCT_away
## t = 1.7461, df = 22579, p-value = 0.08081
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.02021673  0.35019016
## sample estimates:
## mean difference
##       0.1649867
```

Like we saw with our analysis of the entire data, there is no statistically significant difference between free throw percentage for home vs away teams in both the regular season and playoffs as both of our T-tests generate a p-value greater than the 0.05 level of significance.

Overall, we see that for the playoffs, the sample mean differences for each of the stats is slightly greater compared to the regular season, and we see statistically significant differences between home and away teams for points, assists, rebounds, and field goal percentage. This would suggest that home court advantage not only exists but may be even more important in the playoffs, which should incentivize teams to continue pushing for higher playoff seeding in order to secure home court advantages. The mean difference for points between home and away teams in the playoffs is 4.398271, which would suggest that if two teams were equal, the team at home would have an advantage of about 4.4 points. Especially given how close playoff teams can be, every advantage counts, and 4.4 points is a large edge.

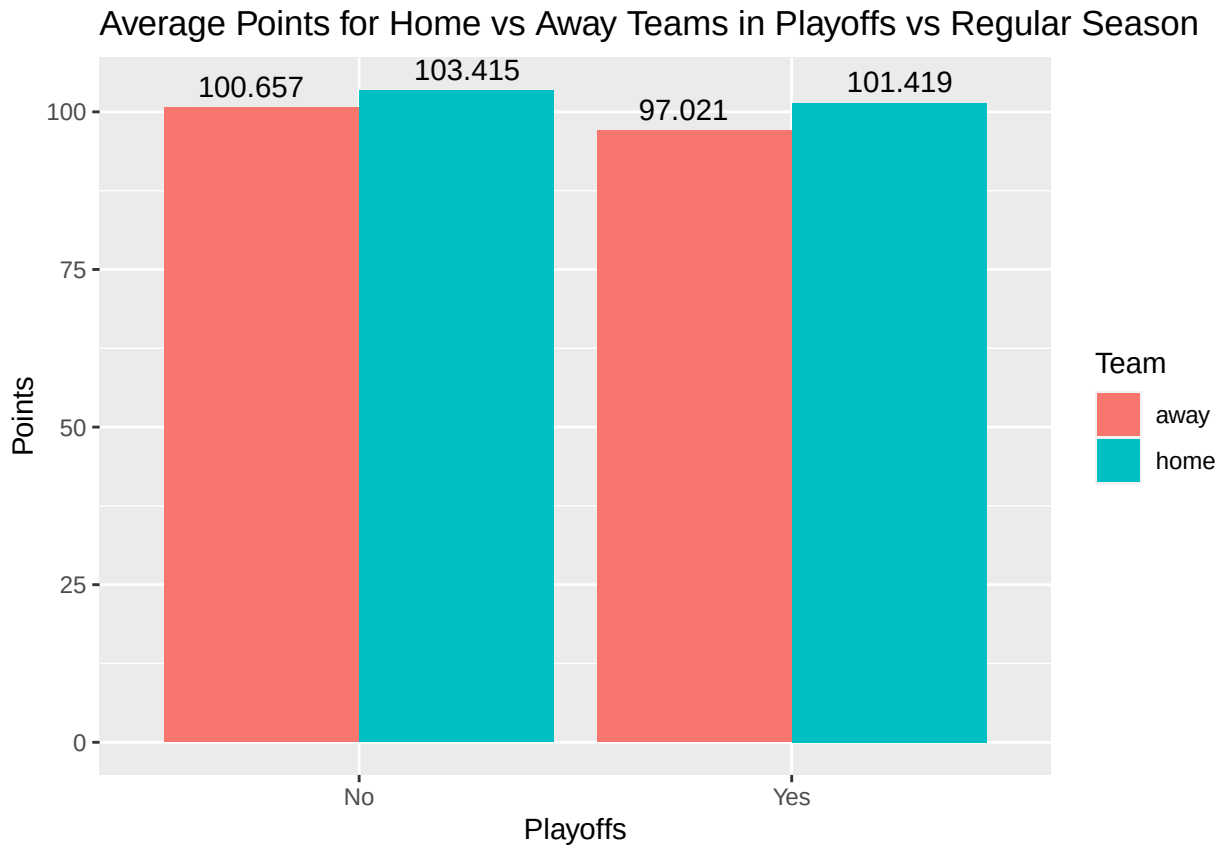
Comparing Regular Season and Playoffs with Bar Charts

To further highlight the differences between home and away stats in the regular season vs playoffs, we can use bar charts to visualize the differences.

We earlier used the mutate function to create a column indicating whether a game was a playoff or regular season game for each of the separate data sets. We can now join them back together to analyze the differences in different stats for home vs away teams in the playoffs vs regular season.

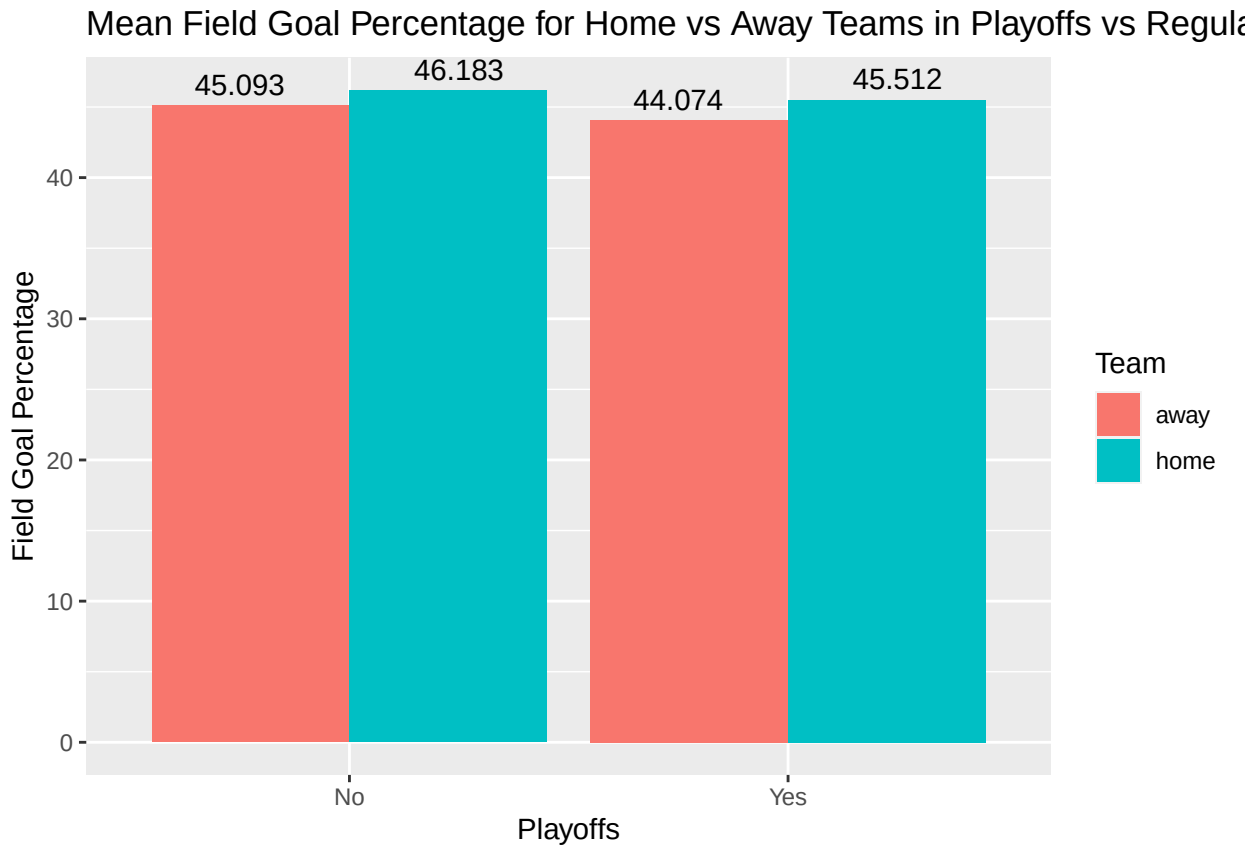
```
NBAREGULARVSPLAYOFF <- rbind(NBAREGULARSEASON, NBAPLAYOFFS)
NBAREGULARVSPLAYOFF %>%
  pivot_longer(c(PTS_away, PTS_home), names_to = "Team", values_to = "Points") %>%
  group_by(Playoffs, Team) %>%
  summarize(mean_pts = mean(Points)) %>%
  ggplot(aes(x = Playoffs, y = mean_pts, fill = Team)) + geom_col(position = "dodge") +
  ↪  geom_text(aes(label = round(mean_pts, 3)), vjust = -0.5, position =
  ↪  position_dodge(1)) + labs(y = "Mean Points") + labs(y = "Points", title = "Average
  ↪  Points for Home vs Away Teams in Playoffs vs Regular Season") +
  ↪  scale_fill_discrete(labels = c("away", "home"))
```

```
## `summarise()` has grouped output by 'Playoffs'. You can override using the
## `.groups` argument.
```



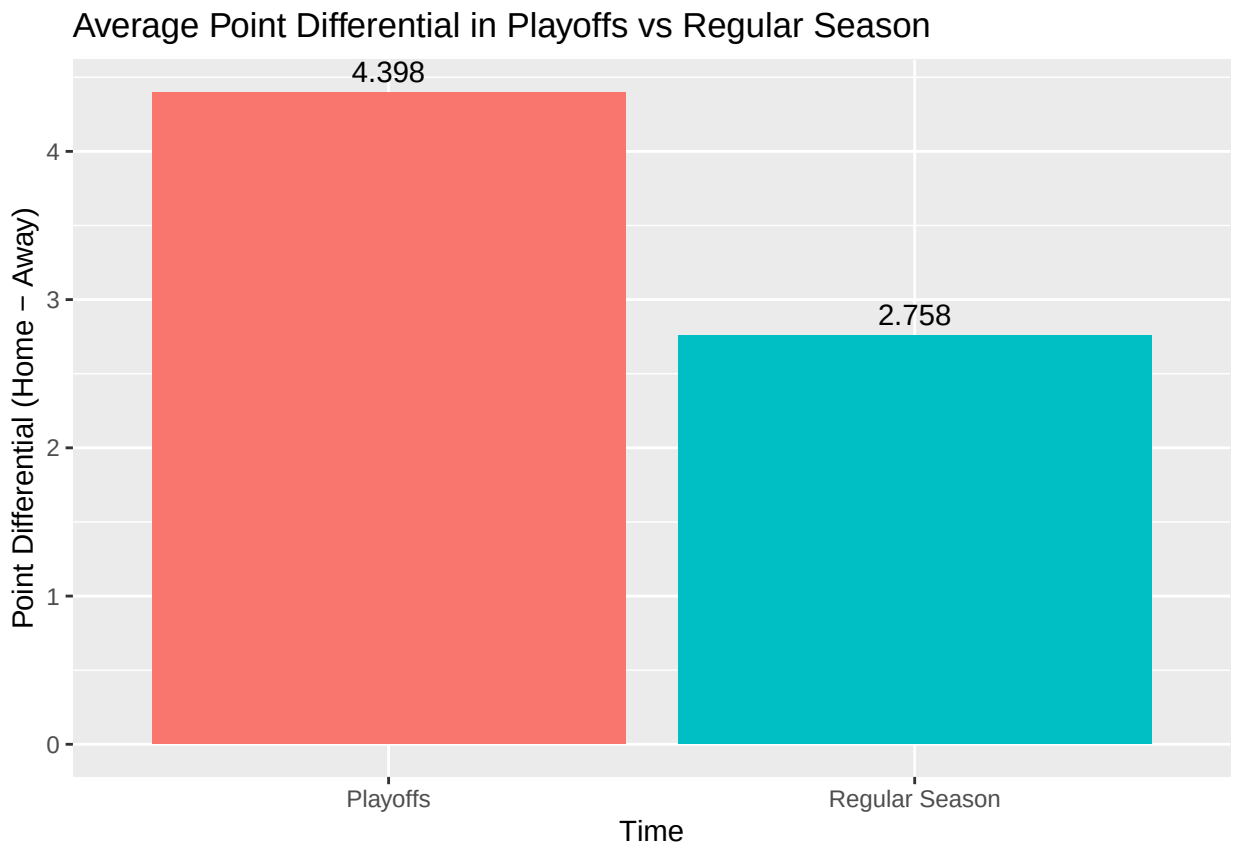

```
NBAREGULARVSPLAYOFF %>%
  pivot_longer(c(FG_PCT_away, FG_PCT_home), names_to = "Team", values_to = "FG_PCT") %>%
  group_by(Playoffs, Team) %>%
  summarize(mean_fgpct = mean(FG_PCT)) %>%
  ggplot(aes(x = Playoffs, y = mean_fgpct, fill = Team)) + geom_col(position = "dodge") +
  ↪  geom_text(aes(label = round(mean_fgpct, 3)), vjust = -0.5, position =
  ↪  position_dodge(1)) + labs(y = "Field Goal Percentage", title = "Mean Field Goal
  ↪  Percentage for Home vs Away Teams in Playoffs vs Regular Season") +
  ↪  scale_fill_discrete(labels = c("away", "home"))
```

```
## `summarise()` has grouped output by 'Playoffs'. You can override using the
## `.groups` argument.
```

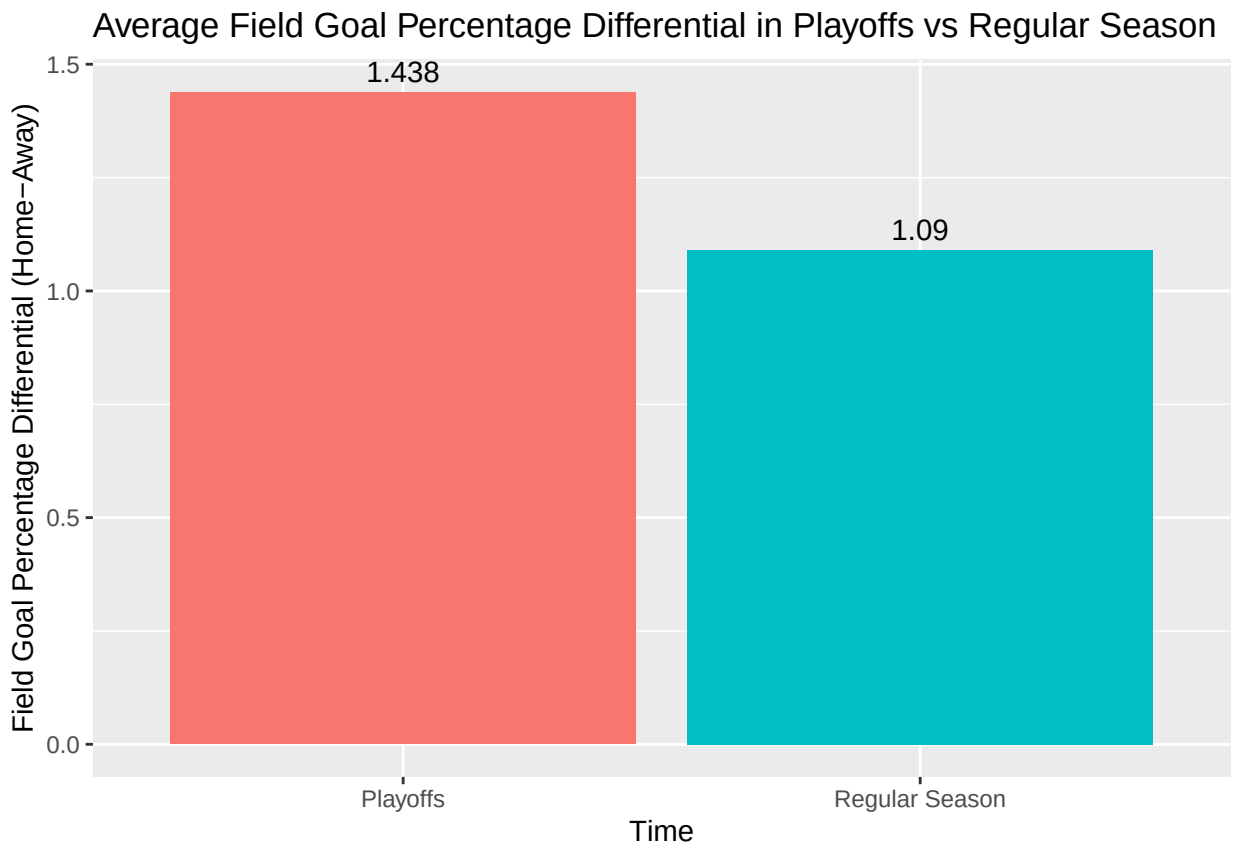


We can now take a more clear look at the point and field goal differential between home and away teams in the playoffs vs regular season by using the mutate() function to create a column for point differential and field goal percentage differential. We can also compare the home win proportion of teams during the playoffs vs regular season.

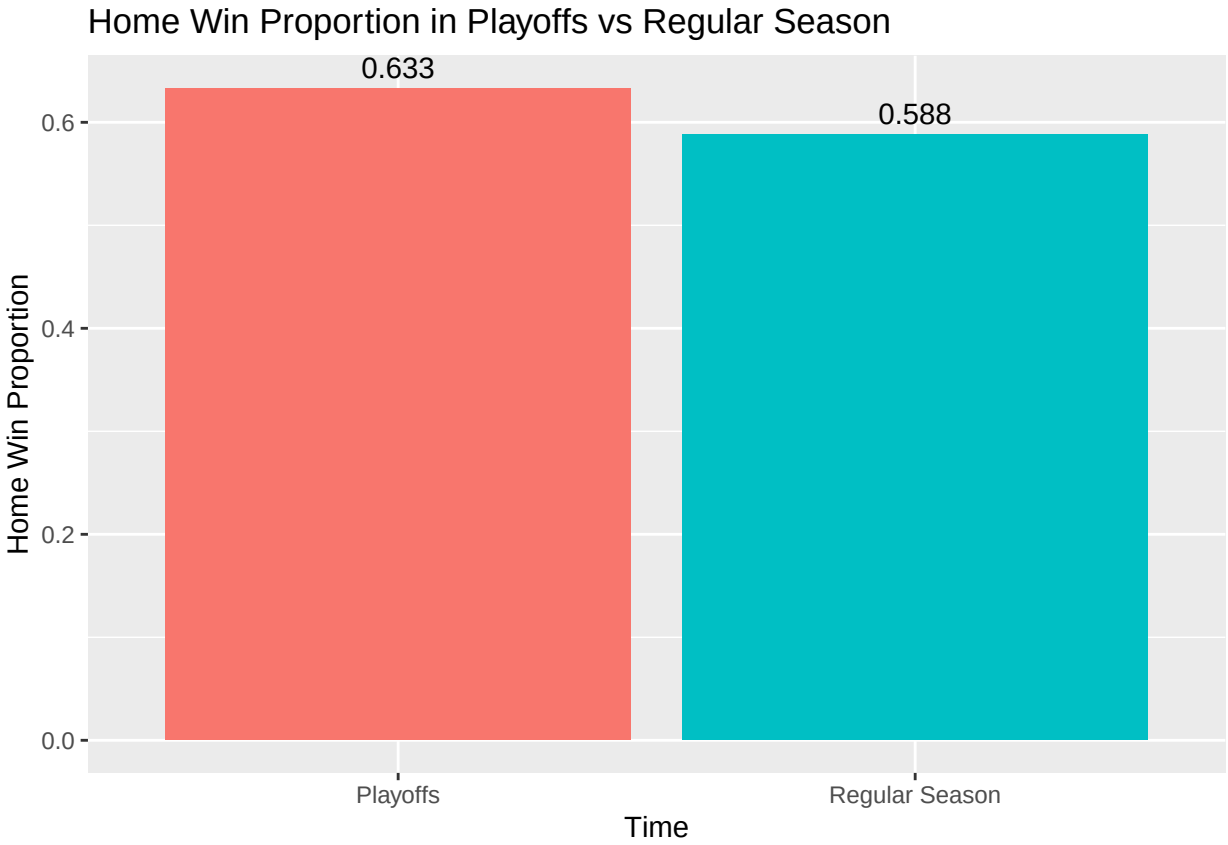
```
NBAREGULARSEASON <- NBAREGULARSEASON %>% mutate(pointdifferential = PTS_home - PTS_away)
↪  %>% mutate(fgpctdifferential = FG_PCT_home - FG_PCT_away) %>% mutate(Time = "Regular
↪  Season")
NBAPLAYOFFS <- NBAPLAYOFFS %>% mutate(pointdifferential = PTS_home - PTS_away) %>%
↪  mutate(fgpctdifferential = FG_PCT_home - FG_PCT_away) %>% mutate(Time = "Playoffs")
TOTALNBA <- rbind(NBAREGULARSEASON, NBAPLAYOFFS)

TOTALNBA %>%
  group_by(Time) %>%
  summarize(mean_point_differential = mean(pointdifferential)) %>%
  ggplot(aes(x = Time, y = mean_point_differential, fill = Time)) + geom_col() +
  ↪  geom_text(aes(label = round(mean_point_differential, 3)), vjust = -0.5, position =
  ↪  position_dodge(1)) + theme(legend.position = "none") + labs(y = "Point Differential
  ↪  (Home - Away)", title = "Average Point Differential in Playoffs vs Regular Season")
  ↪
```




```
TOTALNBA %>%
  group_by(Time) %>%
  summarize(mean_fgpct_differential = mean(fgpctdifferential)) %>%
  ggplot(aes(x = Time, y = mean_fgpct_differential, fill = Time)) + geom_col() +
  ↪  geom_text(aes(label = round(mean_fgpct_differential, 3)), vjust = -0.5, position =
  ↪  position_dodge(1)) + theme(legend.position = "none") + labs(y = "Field Goal
  ↪  Percentage Differential (Home-Away)", title = "Average Field Goal Percentage
  ↪  Differential in Playoffs vs Regular Season")
```




```
TOTALNBA %>%
  group_by(Time) %>%
  summarize(mean_wpct = mean(HOME_TEAM_WINS)) %>%
  ggplot(aes(x = Time, y = mean_wpct, fill = Time)) + geom_col() + geom_text(aes(label =
  ↪  round(mean_wpct, 3)), vjust = -0.5, position = position_dodge(1)) +
  ↪  theme(legend.position = "none") + labs(y = "Home Win Proportion", title = "Home Win
  ↪  Proportion in Playoffs vs Regular Season")
```



These bar graphs reinforce our earlier results that the differences in points, field goal percentage, and proportion of wins is even greater for home teams compared to away teams in the playoffs compared to the regular season. We can now conduct two sample T-tests to test for statistical significance.

```
t.test(NBAREGULARSEASON$pointdifferential, NBAPLAYOFFS$pointdifferential)
```

```
##
##  Welch Two Sample t-test
##
## data:  NBAREGULARSEASON$pointdifferential and NBAPLAYOFFS$pointdifferential
## t = -4.5116, df = 1701.4, p-value = 6.875e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.3533375 -0.9271732
## sample estimates:
## mean of x mean of y
##  2.758016  4.398271
```

```
t.test(NBAREGULARSEASON$HOME_TEAM_WINS, NBAPLAYOFFS$HOME_TEAM_WINS)
```

```
##
##  Welch Two Sample t-test
##
## data:  NBAREGULARSEASON$HOME_TEAM_WINS and NBAPLAYOFFS$HOME_TEAM_WINS
## t = -3.4882, df = 1718.3, p-value = 0.0004984
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07006423 -0.01963104
## sample estimates:
## mean of x mean of y
## 0.5881311 0.6329787
```

```
t.test(NBAREGULARSEASON$fgpctdifferential, NBAPLAYOFFS$fgpctdifferential)
```

```
##
##  Welch Two Sample t-test
##
## data:  NBAREGULARSEASON$fgpctdifferential and NBAPLAYOFFS$fgpctdifferential
## t = -1.6688, df = 1704.1, p-value = 0.09534
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.75736175  0.06103208
## sample estimates:
## mean of x mean of y
##  1.089934  1.438098
```

There is strong evidence to suggest that there home court advantage is even more important in the playoffs compared to the regular season, as the p-value comparing the point differential between home and away teams in the playoffs vs regular season is 6.875 x 10^-6, and the p-value comparing the difference in home team win proportion in the playoffs vs regular season is 0.0004984. However, the p-value for the difference in field goal percentage differential is 0.09534, so we cannot conclude that there is a statistically significant difference in field goal percent differential between home and away teams in the playoffs vs regular season. Still, we see that home teams tend to win more and by a larger margin in the playoffs compared to the regular season.

Conclusion

Home court advantage is definitely real in the NBA and we estimate that it provides an advantage of about 2.86 points overall, with this advantage becoming even more important and jumping to about 4.40 points in the playoffs. From our analysis, we can see that home teams generally score more points, record more assists and rebounds, and shoot higher field goal and three point percentages compared to away teams. The main factors at play with home court advantage are arena familiarity and the momentum provided by a crowd, which we can observe from looking at the NBA Bubble, the 2020/2021 season, Lakers/Clippers matchups, Warriors/Cavs Finals matchups, and the home vs away win percentage for teams with varying levels of average attendance. Based on our analysis, the momentum and encouragement provided by a crowd at sports games is the most significant factor in giving home teams an advantage, while arena familiarity is a smaller but still present factor. Looking at our graph of home win percentage over time, there is a sharp downward trend starting in 2019 due to the COVID-19 pandemic, but now that fan capacity is back to normal levels and we are moving past the pandemic, we expect the proportion of home wins and the home advantage in each of the individual statistics to go back up to pre-2019 levels. While many people that attempt to predict the outcome of a game between two teams will look at how they stack up talent-wise, home-court advantage is a crucial, underrated factor in determining the outcome of a game. Still, it's important to remember that the game of basketball remains the same no matter where its played, so even though home court advantage can certainly swing a game between closely-matched teams, the impact of home court advantage should not be exaggerated and can be nullified by a mismatch in team levels.