````
---
title: "MLS Popularity Analysis: Does Score and Attendance Measure Popularity?"
author: "Ankit Sanghi and Bichuan Ren"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(tidyr)
library(dplyr)
```

## Abstract

Our goal is to analyze whether average score and match attendance level are indicative of how popular a Major League Soccer team is, and we wanted to identify which teams are most popular. Using these metrics to analyze popularity, we found that mean score and attendance are indeed indicative of popularity, and the most popular teams according to these metrics are LA Galaxy, Portland Timbers, and Seattle Sounders FC.

## Introduction

Major League Soccer is a popular sport in North America, and we were interested in learning more about what the best teams are and which teams are most popular. We aim to measure this by using two metrics. We will be using the average number of points scored per team, as this is often an indicator of popularity, and we will be using attendance records for matches where a specific team plays, as if a team is more popular, more people tend to attend games.

```{r, echo=FALSE}
matches <- read.csv("~/Desktop/Ankit/Carleton/Classes/STAT 120/Final Project/matches.csv")
```

```{r, echo=FALSE}
teams <- select(matches, home, away)
all_teams <- data.frame(a=unlist(teams, use.names = FALSE))
names(all_teams)[1] <- "Teams"

scores <- select(matches, home_score, away_score)
all_scores <- data.frame(a=unlist(scores, use.names = FALSE))
names(all_scores)[1] <- "Scores"

all_teams_with_scores <- data.frame(all_teams, all_scores)
total_scores <- tapply(all_teams_with_scores$Scores, all_teams_with_scores$Teams, sum)
unique_teams <- unique(all_teams_with_scores$Teams)
unique_teams <- sort(unique_teams)
teams_with_total_scores <- data.frame(unique_teams, total_scores)
teams_with_total_scores <- teams_with_total_scores %>% filter(total_scores > 250)
```

## Data

Our data consists of sports statistics for Major League Soccer from over 6000 matches, ranging from years 1996 to 2020. It was publicly available on Kaggle, and was obtained through an observational study. Our key variables for analysis are attendance, teams, and points. Our data has a home column and an away column for the teams, which we have combined into a single column along with the corresponding scores and attendance records to make it easier for analysis. We believe that using historical sports data such as this is appropriate for answering our questions about which teams are most popular, as it includes the number of points scored by each team per match, as well as the number of attendees for each team per match, and our sample size is large enough for us to draw conclusions from it.
````

Results

Looking at Average Scores for Teams

First, let us look at the average number of points scored by each team. We plot the top 13 teams in the League by average score, and we find that our top 3 teams are East All-Stars, LAFC, and West All-Stars. This could be an indication of how popular these teams are. However, if we look at the number of games won overall, we find that East All-Stars and West All-Stars only played one game each, and did not win either game. However, they did manage to score 3 and 2 respectively, which is why their average is so high. Similarly, LAFC only played 90 matches, which would explain why they have such a high average score. The average number of matches played by a team is 315, so these teams have played significantly fewer matches compared to other teams. This makes us think that the team with the highest average number of points is not necessarily the one that is most popular.

```{r, echo=FALSE, fig.width=6, fig.height=3}
score_means <- tapply(all_teams_with_scores$Scores, all_teams_with_scores$Teams, mean)
teams_with_mean_score <- data.frame(unique_teams, score_means)
teams_with_mean_score_filtered <- teams_with_mean_score %>% filter(score_means > 1.5)
ggplot(teams_with_mean_score_filtered, aes(y = unique_teams, x = score_means)) +
geom_bar(stat = "identity") + labs(title= "Average Score Per Team", y="Teams", x =
"Average Scores")
```

We ran an ANOVA F-Test on our data to test whether the average scores for the teams are different. On running the test, we get a p-value of 2×10^{-16}, which tells us that there is in fact a difference in mean scores for the teams.

```{r, anova, include=FALSE}
summary(aov(Scores ~ Teams, data = all_teams_with_scores))
```

Since we found this discrepancy in the average scores, we decided to filter out all teams that played fewer than 200 games. This would control for teams that have high average scores simply due to playing fewer games than the average. As a result, the following graph is much more uniform, with little variation in average score per team. There are a few cases of above average scores, such as LA Galaxy, and below average scores, such as Chivas USA. This is a much better indication of team popularity, as we made sure teams that have played very few times and as a result have high average scores are controlled for.

```{r, echo=FALSE, fig.width=7, fig.height=4}
filtered_by_occurance <- all_teams_with_scores %>%
   group_by(Teams) %>%
   filter(n()>200)

score_means <- tapply(filtered_by_occurance$Scores, filtered_by_occurance$Teams, mean)
relavent_names <- names(score_means)
teams_with_mean_score <- data.frame(relavent_names, score_means)
teams_with_mean_score_filtered <- drop_na(teams_with_mean_score)
ggplot(teams_with_mean_score_filtered, aes(y = relavent_names, x = score_means)) +
geom_bar(stat = "identity") + labs(title= "Average Score for Teams that Played Over 200
Games", y="Teams", x = "Average Score")

```

We can test this finding by using an ANOVA F-Test to compare the mean scores of our teams. Our null hypothesis is that all teams have the same mean score, whereas our alternative hypothesis is that the teams have varying mean scores. Upon conducting the test, we get a p-value of 2.68×10^{-9}, which leads us to believe that we have enough evidence in favor of our alternative hypothesis, and that we can conclude that the mean scores for teams that played more than 200 games are different.

This means that according to mean scores, the three most popular teams are LA Galaxy,

Portland Timbers, and New York Red Bulls.

```{r, include=FALSE}
summary(aov(Scores ~ Teams, data = filtered_by_occurance))
```

Looking at Relationship between Attendance per match and Teams

The mean attendance for all games in the dataset is 19314.64. Based on this, we set a cutoff of 20000 for attendance in order to filter out all teams with a mean attendance lower than our cutoff. This allowed us to analyze the top 11 teams with the highest average attendance. We plot a boxplot of attendance of these teams to see their distributions. We find that the teams that draw the largest number of attendees are Atlanta United FC and Seattle Sounders FC. This would lead us to believe that these are the most popular teams when using attendance as a metric. Atlanta United FC is one of the best teams in the league, and are known to have some of the most supportive fans, as mentioned in a Forbes article written by Chris Smith. This would explain them having the highest attendance. Seattle Sounders FC is a similar case, with their supporters being very passionate, and the team set a new MLS record for average attendance in each of its first five seasons. This can be found on the Sounders website. We find that both of these teams are some of the most valued in the league, and their attendance metrics are a good indication of that.

```{r, echo=FALSE, fig.width=6, fig.height=3}
attendances <- select(matches, attendance)
attendances <- rbind(attendances, attendances)
teams_with_attendance <- data.frame(all_teams, attendances)
teams_with_attendance$attendance <- as.numeric(gsub(",", "",
teams_with_attendance$attendance))
teams_with_attendance <- drop_na(teams_with_attendance)

mean_attendance <- tapply(teams_with_attendance$attendance, teams_with_attendance$Teams,
mean)
required_names <- names(mean_attendance)
teams_with_mean_attendance <- data.frame(required_names, mean_attendance)
teams_with_mean_attendance <- teams_with_mean_attendance %>% filter(mean_attendance >
20000)
filtered_teams_with_attendance <- filter(teams_with_attendance, Teams %in%
teams_with_mean_attendance$required_names)
ggplot(filtered_teams_with_attendance, aes(x=Teams, y=attendance)) + geom_boxplot() +
coord_flip()
```

We can verify this statistic by using an ANOVA F-Test, where our null hypothesis is that the average attendance for all teams are equal, and our alternative hypothesis is that there is at least one team with average attendance that is not equal to the average attendance to the rest of the teams. Upon conducting the test, we observed a p-value of 2×10^{-16}, which is small enough for us to conclude that there is at least one team that has an average attendance that is not equal to the rest.

```{r, include=FALSE}
summary(aov(attendance ~ Teams, data = filtered_teams_with_attendance))
```

Comparing Results

We can now compare the best teams as identified by both of our metrics to come to a conclusion as to what the most popular team is. Our best teams by mean score were LA Galaxy, Portland Timbers, and New York Red Bulls. Out of these teams, only LA Galaxy and Portland Timbers made the cut for our top 11 teams as found by our attendance metric. These two teams are on par with the rest of our top 11, except for Atlanta United FC and Seattle Sounders FC, which draw more attendees. Conversely, when we analyze Atlanta United FC and Seattle Sounders FC by our mean score metric, only Seattle Sounders FC is present in our top 13 teams by mean score. This means that the only teams that score well on both

our metrics are LA Galaxy, Portland Timbers, and Seattle Sounders FC. It must be noted that although there are some more teams that exist on both our graphs, they do not do particularly well on either one, so we can take this to mean that they are not very popular.

Thus, we can conclude that according to our analysis, the most popular teams in Major League Soccer are LA Galaxy, Portland Timbers, and Seattle Sounders FC.

Discussion

We used mean score and attendance numbers as metrics to analyze which teams are the most popular in Major League Soccer. We found that LA Galaxy, Portland Timbers, and Seattle Sounders FC are the most popular teams according to these metrics. However, our analysis is limited by the variables we have available in our dataset. If we had some more variables about how much money each club makes, how famous the team members are, how much money they spend on advertising or club resources, and other such features, we could have done a more in-depth analysis of how popular every team is.
Some new questions that came up were whether how old a team is indicate how popular they are, or whether a team's home state determine their popularity, as soccer has varying interest levels in different states. We would have liked to have answered these questions as we found them interesting, but they were outside the scope of our analysis.

References

Mohr, Joseph. “Major League Soccer Dataset.” Kaggle, 9 Nov. 2020, www.kaggle.com/josephvm/major-league-soccer-dataset/.

Smith, Chris. “Major League Soccer's Most Valuable Teams 2019: Atlanta Stays On Top As Expansion Fees, Sale Prices Surge.” Forbes, Forbes Magazine, 9 Dec. 2019, www.forbes.com/sites/chrissmith/2019/11/04/major-league-soccers-most-valuable-teams-2019-atlanta-stays-on-top-as-expansion-fees-sale-prices-surge/?sh=160bb40051b5.

Matt Gaschk. “Sounders FC Announce Results of Landmark General Manager Vote.” Seattle Sounders FC, 10 Jan. 2014, www.soundersfc.com/post/2012/12/13/sounders-fc-announce-results-landmark-general-manager-vote.