~~~
---
title: "Exploratory Data Analysis"
author: "Ankit Sanghi and Bichuan Ren"
date: "11/2/2020"
output: pdf_document
---

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

## Introduction to Our Dataset

Our data consists of sports statistics for Major League Soccer for over 6000 matches,
ranging from years 1996 to 2020. Our key variables for analysis are venue, attendance,
teams, and points. Our data has a home column and an away column for the teams. To find
the team with the highest number of points, we will combine the home and away columns, as
well as the home score and away score columns to create an overall teams column and scores
column. This will allow us to better visualize the overall scores made by each team.

```{r, echo=FALSE}
matches <- read.csv("/Accounts/sanghia/STAT 120/Final Project/matches.csv")
```

## Fusing Our Dataset

We create a new dataframe with the columns Teams and Scores, where Teams consists of all
teams that played in matches, and Scores consists of the corresponding scores of those
teams.
```{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)
```

## Which Team in the League is the Best?

We have data consisting of all teams and their score from every match. We summed up the
scores for each team to identify which team scores the most number of points. We then
plotted a bar chart of these scores, as seen below:

```{r, fig.height=4, fig.width=5, echo=FALSE}
ggplot(teams_with_total_scores, aes(y = unique_teams, x = total_scores)) + geom_bar(stat =
"identity") + labs(title= "Overall Scores for Teams", y="Scores", x = "Teams") +
theme(legend.position = "none")
```

As we can see, LA Galaxy has scored the most number of points, with a whopping (enter
number here) points. This makes sense as if we look at the number of matches played by the
teams in our dataset, we can see that LA Galaxy has played over 200 more games than other
~~~

teams. This explains why they are able to score so many more points than other teams.

~~~~
```{r, include=FALSE}
summary(all_teams_with_scores)
```
~~~~

We want to also look at the average number of points each team scores in a given match. This will give us some more insight into which team scores the most points as if we find LA Galaxy to not have the highest average score per game, then we might suspect that they are on top simply because they played more games, and not because they are actually the best. To do so, we can calculate the mean score per game for every team and compare the results.

~~~~
```{r, include=FALSE}
score_means <- tapply(all_teams_with_scores$Scores, all_teams_with_scores$Teams, mean)
teams_with_mean_score <- data.frame(unique_teams, score_means)
# ggplot(teams_with_mean_score, aes(y = unique_teams, fill = score_means)) +
geom_bar(position = "dodge")
score_means
```
~~~~

Interestingly, we find that LA Galaxy scores an average of 1.6 points per match. This is much lower than many other teams, which leads us to believe that they are at the top simply because of the number of matches they've played, and not because of them actually being good.

## Comparing Attendance and Venues

Our data includes information about the venues for each match and their attendance. We want to find out which venue has the highest attendance overall. To do this, we can calculate the mean attendance for each venue. We filtered out the smaller attendance venues to reduce the sizes of our graphs.

~~~~
```{r, fig.height=3, fig.width=5, echo=FALSE}
cleaned <- drop_na(matches, attendance)
cleaned$attendance <- as.numeric(gsub(",", "", cleaned$attendance))
venue_means <- tapply(cleaned$attendance, cleaned$venue, mean, na.rm = T)
unique_venues <- unique(cleaned$venue)
unique_venues <- sort(unique_venues)

venues_with_mean <- data.frame(unique_venues, venue_means)
venues_with_mean <- na.omit(venues_with_mean)
venues_with_mean <- venues_with_mean %>% filter(venue_means > 30000)
ggplot(venues_with_mean, aes(y = unique_venues, x = venue_means)) + geom_bar(stat =
"identity") + labs(title= "Mean Attendance", y="Venues", x = "Attendance")
```
~~~~

From our bar chart, we can see that Mercedes-Benz Stadium, FedEx Field, Levi's Stadium, Stanford Stadium, and Rose Bowl are the most attended stadiums for MLB matches.

## Analysis Plans

We plan on conducting a chi-squared test on our data containing teams and average points scored by that team. We want to test whether there is an association between the team and the average points scored. It is possible that the teams who are ahead are only ahead because of the number of games they've played, so this should give us an idea about whether these two variables are actually related.

We also plan on doing a chi-squared test on our venues and attendance dataset to verify if the average attendance rate is associated with the stadium. This is to see if certain venues attract a larger crowd than other venues.