Descriptive Statistics

Braylee Cumro, Cameron Weisburg, Moore

Our data set was collected by Mart Jürisoo. He started creating this data set on November ninth of 2017. Since then, he updates the data every month with any new game data. This data was collected so that there was an easy to access and read collection of international football matches. The creator of this data set wanted to be able to answer questions like what teams dominated different eras of football or what trends could be found in geopolitics of the football fixtures. The cases in our data set are one match of an international football game. After we subset the data to include only matches from 2000 and later, we have 20,179 observations of nine variables. The variables that we picked include home_team, away_team, home_score, away_score, and tournament.

For the first variable, home_team, it is a categorical variable. The possible values that can be included are country names. The name used for the home teams is kept consistent with the team's current name so there are no issues dealing with country name changes as the guy who collected this data made those changes for us. There are 296 different teams in this dataset. The countries who show up in this categorical variable are Mexico with 240, the United States with 230, Japan with 216, and South Korea with 195. These numbers were found with the code bellow.

```
ht <- SoccerData %>%

  count(home_team)

print(ht, n=nrow(ht))
```

In this part of the project, we choose to narrow down the vast amounts of teams to four of the most popular teams for men's soccer. These are England with 146 observations, France with 126 observations, Germany with 175 observations and Argentina with 141 observations. If we look at the proportions for these teams, we get the following table of values. Argentina and England are close to each other in proportions and France and Germany are close together. This does not show their proportions in relation to any other team in this data set, but rather their proportions when these data only contain these four teams.

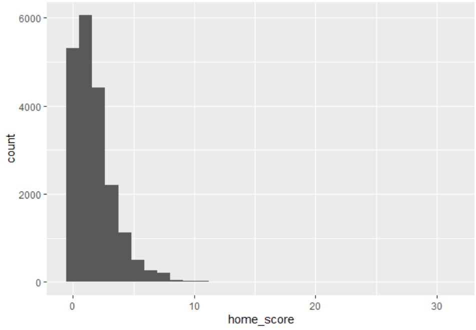
```
table_pop <- table(popular$home_team)
```

```
prop.table(table_pop)

Argentina   England    France   Germany

0.2210031 0.2288401 0.2758621 0.2742947
```

Home_score is a numerical variable in this data. Its units are points or goals in this case. The minimum value of this variable is zero and the max in this data set is 31. The median is 1 and the mean is 1.64. It is important to note that these scores are the full scores of the home team including overtime but excluding any penalty shootouts. The standard deviation is 1.72 goals.

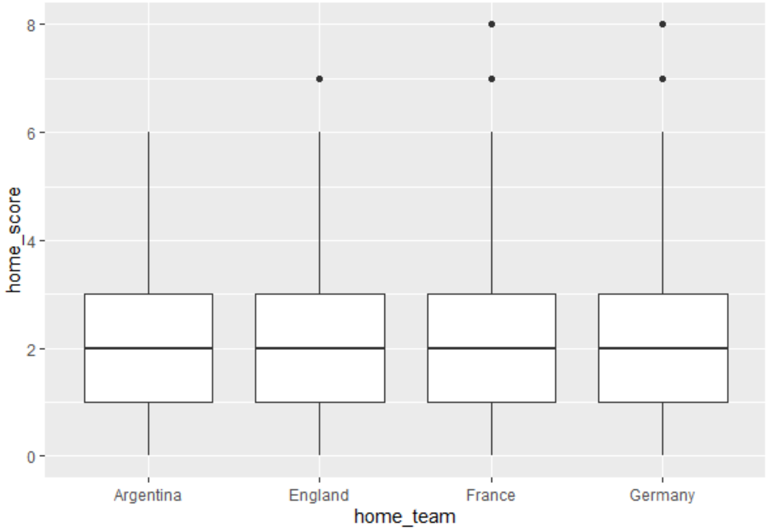
Home_score is right skewed

```
ggplot(SoccerData, aes(x = home_score))+

  geom_histogram(binwidth = 1)
```

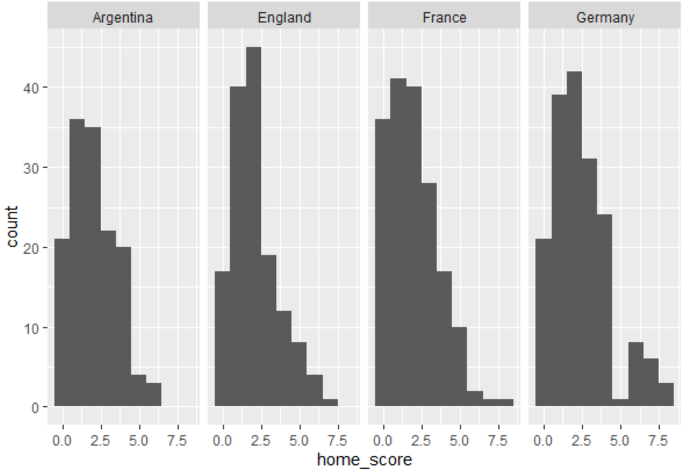


When we take the home_score and look at it individually for some of the most popular countries in men’s soccer, the histograms become more normal shaped as shown below. The countries I chose in this case are Argentina, France, England, and Germany. The histograms still show right skew, but they are more towards the normal end than the unfiltered data. Argentina has 141 observations as a home team. France has 176 observations as a home team. England has 146 observations as a home team. Germany has 175 observations as a home team.

```
popular <- SoccerData %>%

  filter(home_team == "England"  | home_team == "Argentina"
         | home_team == "France" | home_team == "Germany")
ggplot(popular, aes(x=home_team, y=home_score))+
  geom_boxplot()
```



```
ggplot(popular, aes(x=home_score))+
  geom_histogram(binwidth = 1)+
  facet_grid(~home_team)
```



Away_team is a categorical variable that is very similar to the home_team variable. The only difference here is that they were playing as the away team in this case. In this case, 292 different countries have been classified as an away team. Some of the countries with the most games as an away team are Zambia with 182, Uruguay with 158, Paraguay with 166 and Panama with 154. If we continue to look at the countries of Germany, England, France, and Argentia like we did for the home_team, their observation numbers are 121, 114, 115, and 136 respectively. These numbers can be looked at with the code below.

```
at <- SoccerData %>%

  count(away_team)

print(at, n=nrow(at))
```

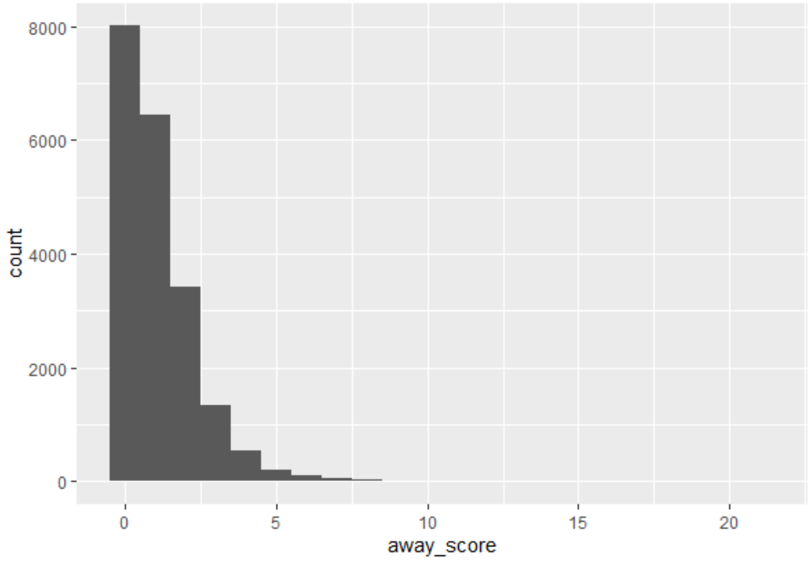
We can calculate the proportions of observations for these four teams the same way we did for the home_team proportions. The code and the table of values can be found below.

```
Argentina   England    France   Germany

0.2798354 0.2345679 0.2366255 0.2489712
```

Away_score is another numerical variable in this data set. The minimum value this variable has is zero, the max is 21 goals. The median is 1 and the mean is 1.103. The standard deviation is 1.354. This variable has the unit of score or goals like home_score. It includes the final score of the away team of a game including overtime goals but excluding penalty-shootout goals.

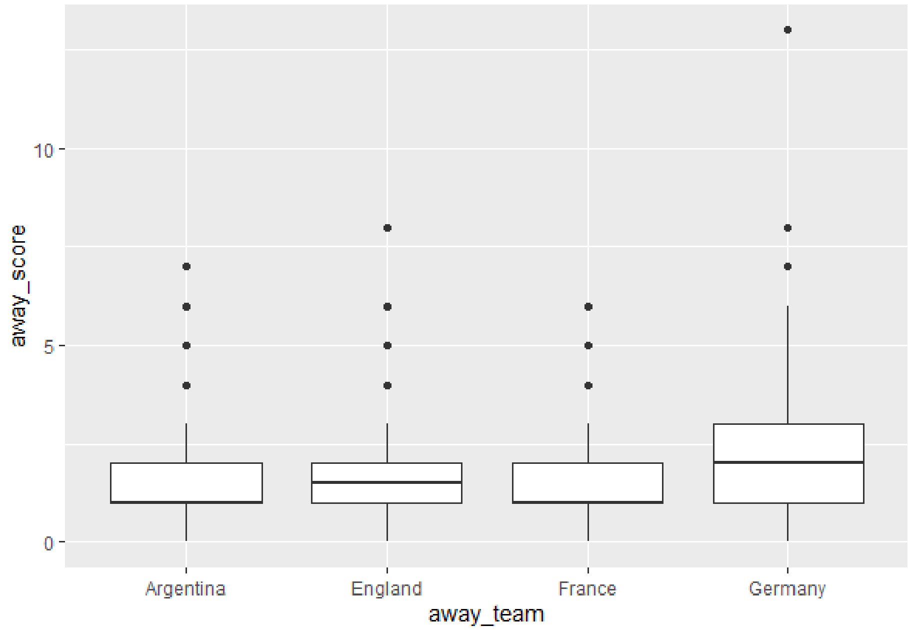
The away_score is right skewed like the home_score. This can be seen in the following histogram.

```
ggplot(SoccerData, aes(x = away_score))+

  geom_histogram(binwidth = 1)
```

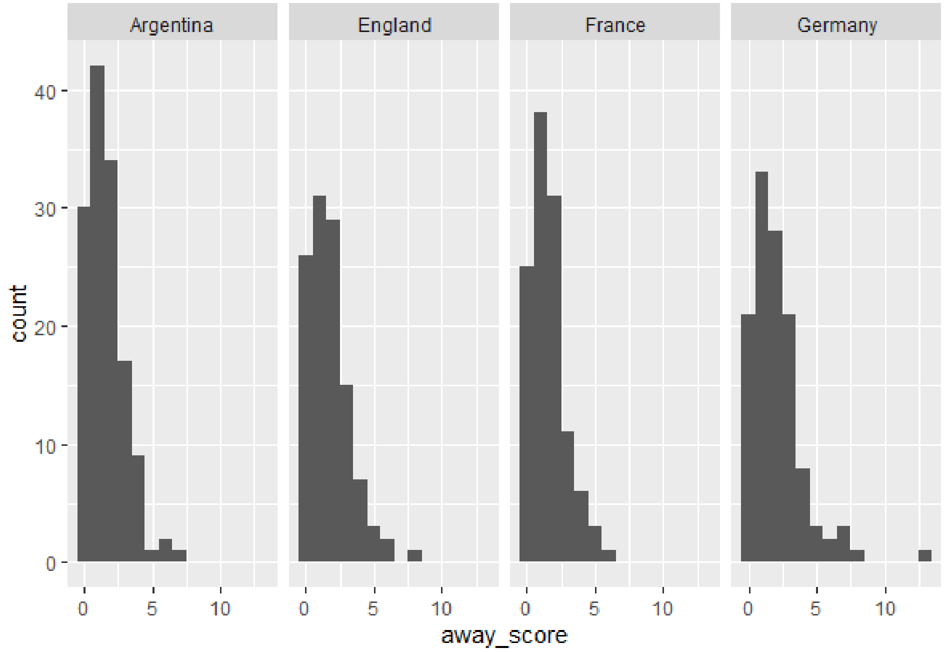


We took the away_score and decided to pair it with some away_teams again to look at some boxplots and histograms. In this case I decided to look at Argentina, England, France, and Germany again. The resulting histograms and boxplots are below. We cans see that the histograms become more normal when we look at one teams away_score rather than all the observations for away_score. The boxplots show less variance and a few more outliers than the boxplots for home_team and home_score.

```
four <- SoccerData %>%

  filter(away_team == "England"  | away_team == "Argentina"

         | away_team == "France" | away_team == "Germany")

ggplot(four, aes(x=away_team, y=away_score))+

  geom_boxplot()
```



```
ggplot(four, aes(x=away_score))+
  geom_histogram(binwidth = 1)+
  facet_grid(~away_team)
```



Tournament is a categorical variable. There are 79 different tournament types that can be found in our data. UEFA Euro qualification has 1282 observations. There are 7135 observations for friendly matches. The third highest type is the FIFA World Cup qualification with 4503 observations.

For the proportions for the tournament type, I found the proportion of games under each tournament type for the four countries: England, Argentina, France, Germany. These proportions were calculated for when these teams were both the home and away team. To do this I took the data filtered to include only these teams as the home team (held in the variable popular created above) and made a table of their tournaments. I then made the prop.table on this table.

```
table_tourn <- table(popular$tournament)
```

```
prop.table(table_tourn)
```

```
         Confederations Cup                Copa América
                0.025078370                  0.050156740
              FIFA World Cup    FIFA World Cup qualification
                0.095611285                  0.191222571
                    Friendly    King Hassan II Tournament
                0.437304075                  0.001567398
                   UEFA Euro      UEFA Euro qualification
                0.068965517                  0.105015674
         UEFA Nations League
                0.025078370
```

I did the same thing when they were the home team. The difference in the code is the use of the variable four in place of popular. This variable is created above, and it holds the data filtered down to when the away team is one of these four teams. One note to make here is that these teams participated in the Kirin Cup as an away team but there are no observations of them in the Kirin cup as a home team.

```
table_t <- table(four$tournament)
```

```
prop.table(table_t)
```

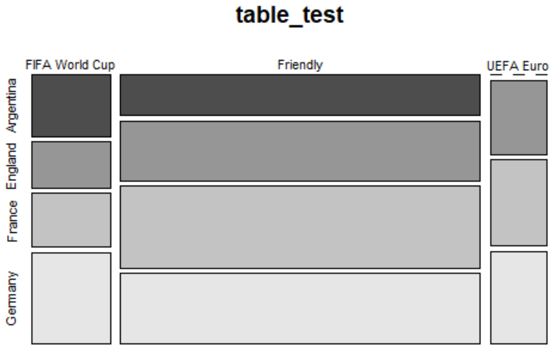
```
          Confederations Cup                Copa América
                 0.018518519                  0.018518519
              FIFA World Cup      FIFA World Cup qualification
                 0.088477366                  0.251028807
                    Friendly    King Hassan II Tournament
                 0.380658436                  0.002057613
                   Kirin Cup                    UEFA Euro
                 0.002057613                  0.067901235
     UEFA Euro qualification          UEFA Nations League
                 0.137860082                  0.03292181
```

The home_team and home_score variables are related to one another. The home_team gives us a label for who scored what points as a home_team. This opens the ability to compare the average scores of home_teams. The relationship that exists between these two variables is also present between the two variables away_team and away_score.

Home_team and tournament type are related because you can look at the proportions of tournaments a particular team has played in as a home_team. We created a mosaic plot to show the relationships that can be found in these two variables. The width of these boxes represents the number of observations under the tournament type they are labeled. This means that for these four teams they had the most observations for friendly tournaments, then FIFA World Cup and lastly the least amount for the UEFA Euro tournaments. The vertical height represents the proportion of these observations that fall within the different teams. We can see here that Argentina is a team that has not participated in the UEFA Euro tournaments for the length of time that this data covers. Germany has participated in the most FIFA World cup tournaments as the home_team and France has been in more matches labeled friendly.

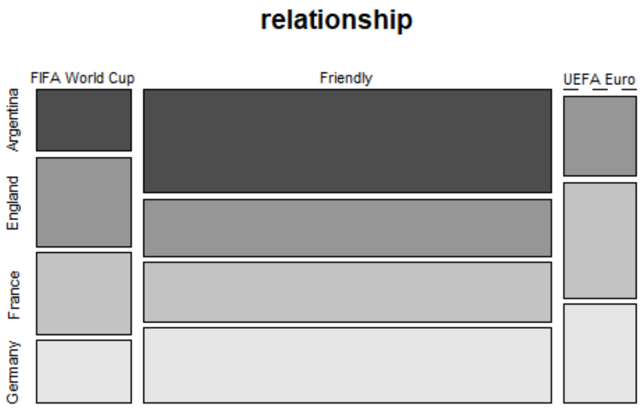
```
tourn <- popular %>%
  filter((tournament == "Friendly") | (tournament == "FIFA World
             Cup") | (tournament == "UEFA Euro"))
table_test <- table(tourn$tournament, tourn$home_team)
```

```
mosaicplot(table_test, color=TRUE)
```



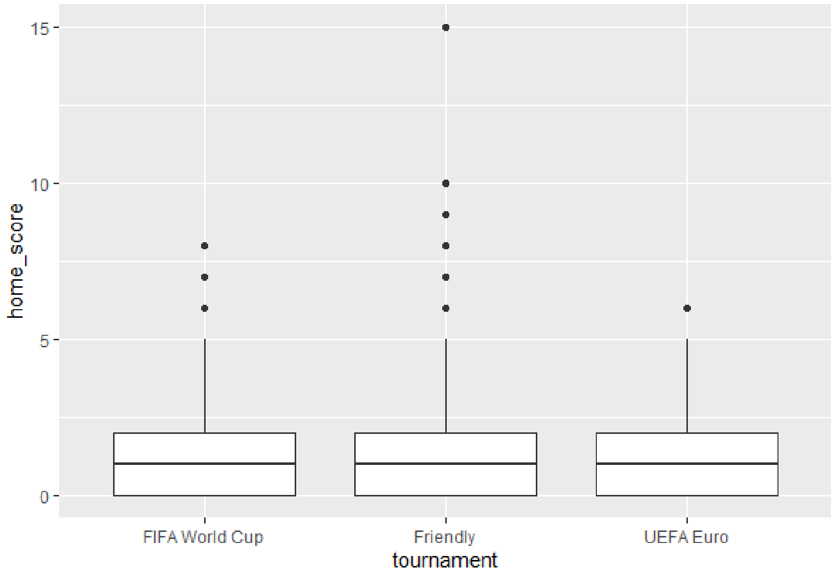
We can create a similar mosaic plot for these teams when they were the away team. We can see that Argentina has not participated in the UEFA Euro as an away team either; however, they have participated in the most matches labeled friendly as an away team. France has had the most UEFA Euro matches as an away team between these four and England did the most FIFA world cup matches as an away team.

```
away_tourn <- four %>%

  filter((tournament == "Friendly") | (tournament == "FIFA World
          Cup") | (tournament == "UEFA Euro"))

relationship <-table(away_tourn$tournament,
away_tourn$away_team)

mosaicplot(relationship, color=TRUE)
```



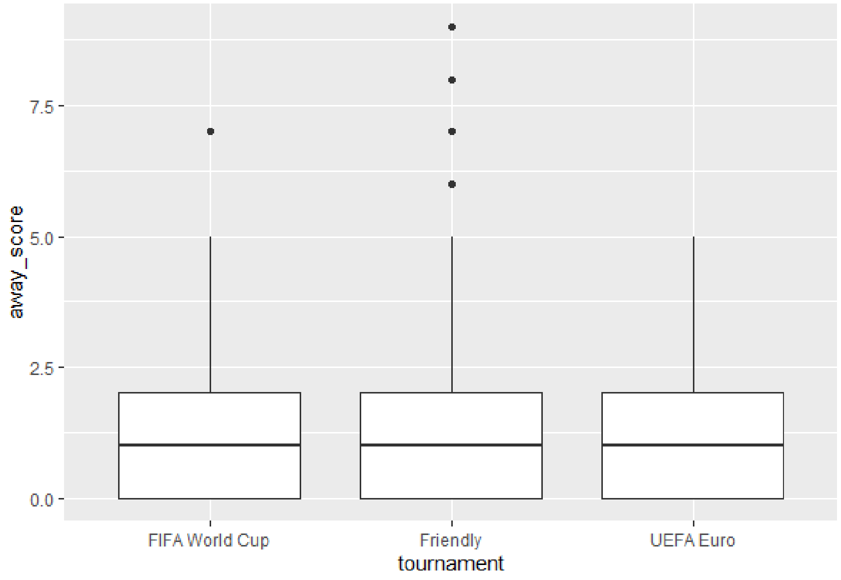
We can use side by side boxplots to look at the relationships between home_score and tournament and away_score and tournament. The side-by-side boxplot can show us the median value of home_scores for different tournament types as well as the variance in these data. Looking at the side by side for home_score, we can see that friendly tournaments have the most outliers, but the median and the variance are similar for the three tournaments. These plots show right skew in this relationship.

```
three_t <- SoccerData %>%

  filter((tournament == "Friendly") | (tournament == "FIFA World Cup") |

           (tournament == "UEFA Euro"))

ggplot(three_t, aes(x=tournament, y=home_score))+

  geom_boxplot()
```



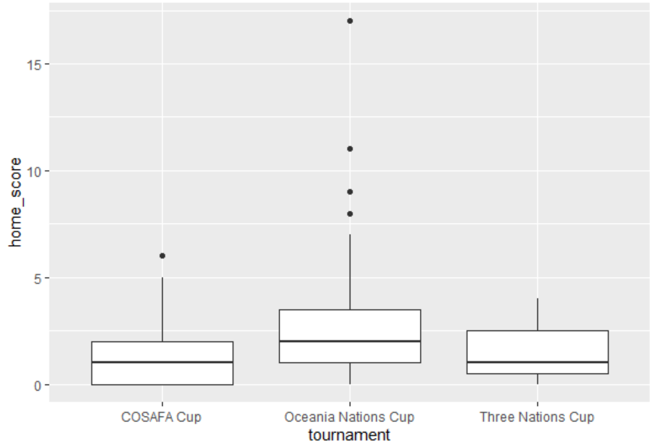
We can do the same thing for the away score. Friendly tournaments still have the most outliers with FIFA having one reported outlier and UEFA Euro does not show any outliers on this boxplot.

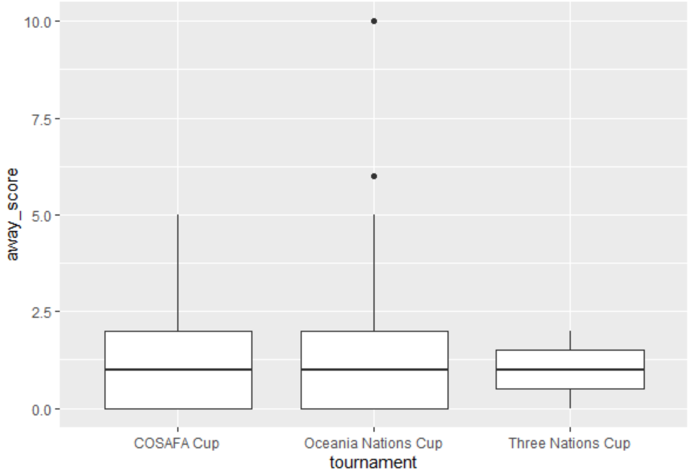
```
ggplot(three_t, aes(x=tournament, y=away_score))+
  geom_boxplot()
```



The box plots for these three tournament types looked similar so we looked at the boxplots for three other lesser-known tournaments. The first plot will show the home_score data for these tournaments the second will show the away_score for these tournaments.

```
ran_t <- SoccerData %>%
  filter((tournament == "Three Nations Cup") | (tournament ==
     "Oceania Nations Cup") | (tournament == "COSAFA Cup"))
ggplot(ran_t, aes(x=tournament, y=home_score))+
  geom_boxplot()
ggplot(ran_t, aes(x=tournament, y=away_score))+
  geom_boxplot()
```





There are outliers in our data. Considering how many observations we have for our data set, most of the data are not outliers. Since the tests we will perform on these data are robust and are not extremely thrown off by the presence of outliers in the data, we can leave these outliers in the data. We will keep these outliers in mind while doing the rest of the project and take notes of when outliers show up in the data we are working on at the time.

There are no missing values for our data set. We can also note that the home-team and away-team names take on the most current name, so we do not need to look for other variations of a country or team name. The creator of this data made sure that the data was filled in for every observation and that the team names stayed consistent even though those have changed throughout the years.