Inferential Statistics

Braylee Cumro, Cameron Weisburg, Jett Moore

The data set we are looking at includes 20,179 observations that represent one game of soccer. These games are from January fourth, 2000, to September ninth, 2020. These data have nine variables: the date of the game, the home team, the away team, the home score, the away score, the tournament type, the city it was played in, the country it was played in, and whether the match was neutral or not. Neutral refers to whether the match was played at a neutral venue.

Are the average scores for the Oceania Nation Cup tournament type statistically different than the average home score for the Cyprus International Tournament? For this test we will be using a significance value of 0.05.

The null hypothesis is that the home score is the same for the two tournament types, Oceania Nation Cup and Cyprus International Tournament. The mean home score for teams in the Oceania Nation Cup is equal to the mean for teams in the Cyprus International Tournament.

The alternate hypothesis is that there is a statistically significant difference in the home score when the tournament type is Cyprus International Tournament rather than the Oceania Nation Cup. The mean for home score for teams in the Oceania Nation Cup tournaments is not equal to the mean of home scores for teams in the Cyprus International Tournament.

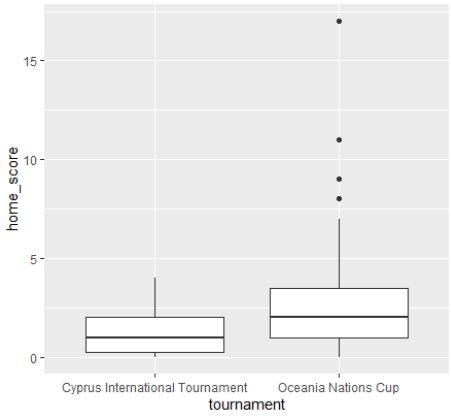
The first step to testing this with a t-test is to look at the difference in the mean visually. To do this I chose a boxplot. Boxplots show the difference in the mean of these data a lot easier than other visualization tools. To do this, I use the code below. There appears to be a difference between the two groups, but now we are going to test this statistically with a t test.

```
is_tourn <- SoccerData %>%

  filter((tournament == "Oceania Nations Cup") | (tournament ==
"Cyprus International Tournament"))


ggplot(is_tourn, aes(tournament,home_score))+

  geom_boxplot()
```



The parameter of interest is the means of these two groups, so we calculated this to see the true value of the mean. The mean for Cyprus is 1.444 and the Oceania Nations Cup has a mean home score of 2.661. We are using the t sampling distribution in this test.

```
tapply(is_tourn$home_score, is_tourn$tournament, mean)
```

```
Cyprus International Tournament             Oceania Nations Cup
                       1.444444                        2.661017
```

Next, we need to test the conditions for the t test. To do this we will need the size of the observations in these two groups. The first condition is normality. As seen below, we have 54 observations for Cyprus and 59 observations for Oceania. Since the data has observations over thirty, we can assume normality regardless of the underlying structure if there are no extreme outliers in the data.

```
tapply(is_tourn$home_score, is_tourn$tournament, length)
```

```
Cyprus International Tournament             Oceania Nations Cup
                             54                              59
```

Cyprus does not have outliers, so we only need to check the outliers for Oceania. The values we are going to be wary of are those that are more than Q3 + 3*IQR or less than Q1 – 3*IQR. This calculation gives us 9.5 as the value not to be larger then. We have two observations that fall higher than this value, so we will check our t test results using randomization to construct our sampling distribution.

```
check <- SoccerData %>%
  filter(tournament == "Oceania Nations Cup")
 hscore <- check$home_score
 2+3*IQR(hscore)
```

```
[1] 9.5

 max(hscore)

[1] 17
```

The second thing to check is that we have independence within and between the groups. This can be hard to check, but we can say that the outcome of a game in one group does not affect the outcome of a game in that same group or the other group. This can go towards saying these data have independence.

I started with using randomization to construct the sampling distribution since we are lightly doubtful of there being normality in these two groups due to outliers. In this randomization I used seed 22 and went through 1000 simulations to get a p value. Long story short, this takes 1000 simulations and finds the difference of means for each one. 59 and 54 make sure that we are selecting 52 observations at random, and these numbers match the number of observations we have for each group. For this test we got a p value of 0.03. This is lower than our set significance value, so we reject the null hypothesis. This tells us there is a statistical difference in the home score means for the Cyprus and Oceania tournament types.

```
set.seed(22)

num_sim <- 1000

diffs <- numeric(num_sim)

for(i in 1:num_sim){

  ocean <- sample(1:59,54)

  o_mean <- mean(is_tourn$home_score[ocean])

  not_o_mean <- mean(is_tourn$home_score[-ocean])

  diffs[i] <- not_o_mean - o_mean

}


o_mean <- mean(is_tourn$home_score[is_tourn$tournament == "Cyprus
International Tournament"])

not_o_mean <- mean(is_tourn$home_score[is_tourn$tournament != "Cyprus
International Tournament"])

obs_diff <- not_o_mean - o_mean

pval <- mean((diffs >= obs_diff) | (diffs <= -obs_diff))

Pval
```

[1] 0.003

Once we had these randomized numbers, we decided to look at what a t test would give us. This code can be found below and gives us a p value of 0.008. while higher than what we got with the results using randomization to construct our sampling distribution, this p value will also say that there is statistically significant difference between these groups and cause us to reject the null hypothesis.

```
t.test(home_score~is_tourn$tournament, data=is_tourn)
```

```
Welch Two Sample t-test

data:  home_score by is_tourn$tournament

t = -2.6973, df = 73.276, p-value = 0.008668

alternative hypothesis: true difference in means between group Cyprus International Tournament and group Oceania Nations Cup is not equal to 0

95 percent confidence interval:

 -2.1154099 -0.3177351

sample estimates:

mean in group Cyprus International Tournament  mean in group Oceania Nations Cup

2.661017                          1.444444
```

If we look at this in context, the difference in how much a home team scores on average in these two tournaments can come down to the fact that one is friendly competition for national teams (Cyprus international tournament) while the other is more competitive. The friendly nature of the Cyprus tournament could make them be motivated to score lower because their top players are less likely to get hurt if they play more conservatively, and they can then be a factor in tournaments that hold more attention and prestige. It is also important to note that it could come from other factors such the demographics of the teams. One is mainly teams that come from countries that fall in the area around Australia (Oceania) while Cyprus has teams in and around the Mediterranean.

Do these data show a significant relationship between the tournament type and whether a game is played in a neutral setting? Since there are 79 tournament types, I narrowed this down to four that have around the same number of observations and had a range of teams that played. These are the CASAFA Cup, CFU Caribbean Cup Qualification, Gold Cup, and UEFA Nations League.

The null hypothesis is that there is no relationship between the tournament type and whether the game was played at a neutral venue.

The alternative hypothesis is that there is a relationship between the tournament type and whether a game is played at a neutral venue.

The parameter of interest will be the chi-squared value to answer this question. The sampling distribution will be the theoretical chi-squared distribution.

The threshold for significance set before doing the testing will be 0.05.

The requirements for a chi-squared test are that there is independence in the variables and the expected values are at least 5. Independence can be hard to prove when you were not the one collecting the data; however, since the venue of one tournaments game does not directly affect the venue of another tournament type, they are independent. As for the expected values, you can see in the table below generated by the code test2$expected that our expected values are at least 5.

To do the chi squared test I created a variable t2 which would filter out the four tournament types we decided to focus on. We then put this into a table of values that looks at how many of the observations for those tournaments are at a neutral venue(true) and not at a neutral venue (false). The results of this table are the observation values for these two categorical variables. Next, we put the chi squared test into the variable test 2 so that we could calculate the expected values and other pieces of info such as the residuals easier. The residuals can be found under the code test2$residuals. The residuals are the difference between the observed and expected values.

```
t2 <- SoccerData %>%

  filter((tournament == "COSAFA Cup") | (tournament == "CFU Caribbean
Cup qualification") |

           (tournament == "Gold Cup") | (tournament == "UEFA Nations
League"))


table1 <- table(t2$tournament, t2$neutral)
```

Table1

```
   FALSE TRUE
  CFU Caribbean Cup qualification   162  131
  COSAFA Cup                        111  170
  Gold Cup                           74  223
  UEFA Nations League               296    8
test2 <- chisq.test(table1)
Pearson's Chi-squared test
     data:  table1
X-squared = 355.95, df = 3, p-value < 2.2e-16

test2$expected
                                  FALSE     TRUE
  CFU Caribbean Cup qualification 160.3396 132.6604
  COSAFA Cup                      153.7728 127.2272
  Gold Cup                        162.5285 134.4715
  UEFA Nations League             166.3591 137.6409
test2$observed
                                  FALSE TRUE
  CFU Caribbean Cup qualification   162  131
  COSAFA Cup                        111  170
  Gold Cup                           74  223
  UEFA Nations League               296    8
test2$residuals
                                     FALSE        TRUE
  CFU Caribbean Cup qualification   0.1311291  -0.1441613
  COSAFA Cup                       -3.4492734   3.7920783
  Gold Cup                         -6.9441386   7.6342794
  UEFA Nations League              10.0512142 -11.0501506
```

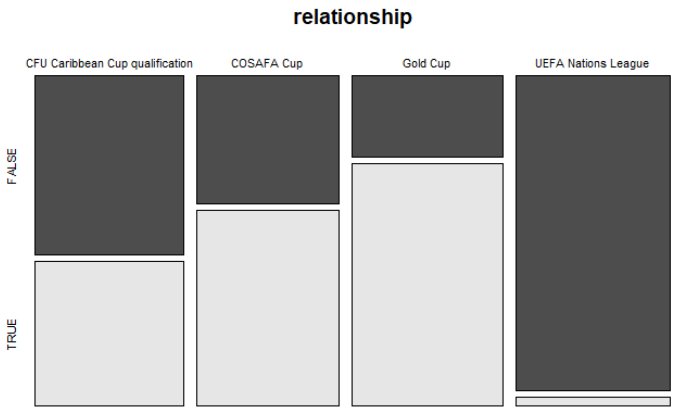
The chi squared value we calculated in this test was 355.95. The computed p-value for this test is 2.2e-16.

Since we set our level of significance at 0.05 this test causes us to reject the null hypothesis that there is no relationship in these variables. We can say that these results give us evidence that there is a relationship between the tournament type and whether the venue is a neutral one.

To visualize this relationship, I made a mosaic plot. True means the venue was a neutral one while false is not a neutral venue. The width of these boxes represents the number of observations under the tournament type they are labeled. The vertical height represents the proportion of these observations that fall within either a neutral venue (True) or not (false). We can see that the vertical height of these observations is very different, especially for the UEFA Nations League.

```
relationship <- table(t2$tournament, t2$neutral)

mosaicplot(relationship, color=TRUE)
```



The CFU Caribbean Cup qualification is very close to its expected values. This could come down to the fact that as a qualification round it is more important to get an idea of who will be in the final tournament, so the games are spread out not necessarily based on who is playing but rather on what fields could be utilized to go through the qualifying round most efficiently. Another factor that could come into play is the different regulations of the tournaments limiting what fields can be used for a certain type of tournament. Not all soccer fields are the same size, so if a tournament has its own rules on the size of the pitch or other pieces of the soccer field it could limit the number of available fields. This could give some teams a higher probability of playing on a home field in certain tournaments, especially if they make it through many rounds of the bracket.

Is the mean of the away_score for friendly tournaments statistically different than the mean of all the other tournaments put together? We will compare the mean of the away_score for friendly games against the mean of the away_Score for all the other tournament types. I choose the friendly game type because it is different than the very competitive nature of the other tournament types and there are a lot of observations of that type.

The parameter of interest is the means of these two groups. The mean for friendly games (true) is 1.006 while the mean for the rest of the tournament types is 1.155 (false). The code we used to calculate this can be found below. Friendly is a variable created to separate out the friendly tournament games.

```
friendly <- SoccerData$tournament == "Friendly"

tapply(SoccerData$away_score, friendly, mean)

   FALSE     TRUE

1.155474 1.006868
```

In this test we will be using the t distribution.

The null hypothesis is that there is no difference in the mean for away scores during friendly tournaments than the mean for the other tournament types combined. The mean away score for teams in the friendly tournaments is equal to the mean away score for the other tournament types.

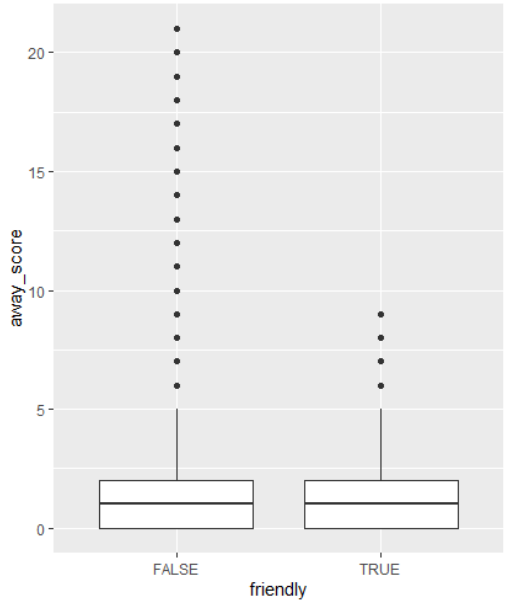
The alternate hypothesis is that there is a statistically significant difference in the mean away score when the tournament type is Friendly rather than any other type. The mean for away score for teams in friendly tournaments are not equal to the mean of away scores for teams in the other tournament types.

The threshold for significance that we are setting for this test is 0.05. This is stated before performing the test.

Now we will look at this problem in the form of a box plot and talk about the requirements for a t-test. A boxplot for this question is shown below where true records data for the friendly tournaments. We can see that there means appear very similar in a boxplot, but since we are talking about a large bulk of data 20179 observations small differences in values can be statistically significant.

```
ggplot(SoccerData, aes(friendly,away_score))+

  geom_boxplot()
```



The thing we need to check in our data to be more confident in a t tests results is if the data have independence within and between groups and that the data are nearly normal. To check for independence when we did not collect the data ourselves is a bit difficult. Since the outcome of one game away score does not affect the away score of another game, we can assume the independence is met. To check for normality, we need to look at how many observations we have first. As you can see below, the friendly tournaments have 7135 observations, and the other group has 13,044. This means that normality can be assumed if there are no extreme outliers in either group. Since the false group has so many outliers, I will also look at a randomization for this data to see if the two are close to matching.

```
tapply(SoccerData$away_score, friendly, length)

FALSE  TRUE

13044  7135

t.test(away_score~friendly, data=SoccerData)

Welch Two Sample t-test

data:  away_score by friendly

t = 8.1004, df = 18281, p-value = 5.82e-16

alternative hypothesis: true difference in means between group FALSE
and group TRUE is not equal to 0

95 percent confidence interval:

 0.1126472 0.1845653
```

```
sample estimates:
mean in group FALSE  mean in group TRUE
           1.155474            1.006868
```

With a simple t test, we get a p value of 5.82e-16 which is small enough to reject the null hypothesis and say that there is a statistically significant difference in the means for the away score at friendly tournaments versus the other types.

```
set.seed(22)
num_sim <- 1000
diffs <- numeric(num_sim)
for(i in 1:num_sim){
  fr <- sample(1:13044,7135)
  fr_mean <- mean(SoccerData$away_score[fr])
  not_fr_mean <- mean(SoccerData$away_score[-fr])
  diffs[i] <- not_fr_mean - fr_mean
}

fr_mean <- mean(SoccerData$away_score[friendly])
not_fr_mean <- mean(SoccerData$away_score[!friendly])
obs_diff <- not_fr_mean - fr_mean
pval <- mean((diffs >= obs_diff) | (diffs <= -obs_diff))
Pval
[1] 0
```

When we do this with randomized data, we get a p value of 0 which is very similar to the p value we got with our actual data. In this case both the regular t test and the randomization used to make our sampling distribution leads us to reject the null hypothesis. Our significance level was 0.05 which both tests give a p value much smaller than that.

If we put the results of this test in context, it means that the mean for the away scores is statistically different than the mean for other tournament types. The mean of the friendly matches is lower than the

other matches by a significant amount statistically. This could come down to the fact that the soccer clubs do not want their best players to get hurt for a friendly match so they do not play as aggressively as they would for qualifiers or cup matches. They play friendly matches to get time on the field against other teams, but the goal is not to go all out and get injured before more prestigious matches take place.

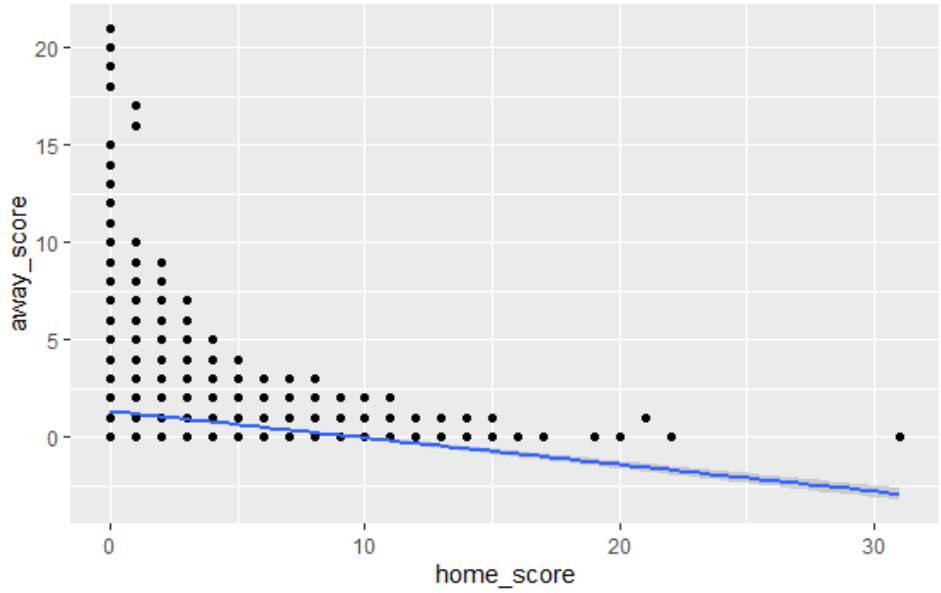
*The next question I wanted to answer with this data set was whether there is a dependence or correlation between the home scores and the away scores.

The null hypothesis is that there is no dependence or correlation between home scores and away scores. When the home team scores a lot, the away team is unaffected, meaning the other team does not score any more or any less depending on the other team's score.

The alternative hypothesis is that there is a dependence or correlation between home scores and away scores. If there is a negative correlation, then that would mean if one team scores a lot then the other team will score less. If there is a positive correlation, then that means when one team scores a lot of points, the other team will score a lot of points as well.

First, we visualized the data using a scatterplot. Scatter plots are great for showing how much one variable is affected by another, in other words their correlation. In this case it shows the correlation between home scores and away scores. To do this we used the code below.

```
ggplot(SoccerData, aes(home_score, away_score))+
  geom_point()+
  geom_smooth(method=lm)
```



The parameter of interest is the slope of the regression line, so we made a linear model. The slope estimate for this data comes out to be -0.136884 with a standard error of 0.005457 and a t-value of -25.09. This means there is a slight negative correlation between home scores and away scores. We tested this by making a linear model to show the residuals and coefficients. First the 95% confidence interval.

```
confint(my.lm, level=0.95)
```

```
2.5 %     97.5 %
(Intercept)  1.3020010  1.3528424
home_score  -0.1475796 -0.1261886
```

Now the code for the linear model below

```
my.lm <- lm(away_score~home_score, data=SoccerData)
summary(my.lm)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.3274 -1.0537 -0.1905  0.6726 19.6726
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  1.327422   0.012969  102.35   <2e-16 ***
home_score  -0.136884   0.005457  -25.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.333 on 20177 degrees of freedom
Multiple R-squared:  0.03025,    Adjusted R-squared:  0.0302
F-statistic: 629.3 on 1 and 20177 DF,  p-value: < 2.2e-16
```

With a significance level of 0.05 and a p-value of < 2.2e-16, we can reject the null hypothesis that there is no dependence or correlation between home scores and away scores. With these results we can say that there is a slight correlation between home scores and away scores. Since the slope of the data is −0.136884, this means that there is a negative correlation.

The negative correlation between home scores and away scores means that if the home team scores more, then the away team will score less and vice versa. In the context of the actual soccer games being played, this can be interpreted as the team that wins the game. If the home team wins the game, then the away team will obviously have a lower score than the home team. The same can be said about the away team. If the away team wins the game, then the home team obviously has a lower score than the away team.

We chose to compare the differences of the means of the total population and a given country. The country we chose to compare to the given mean was France. We chose France because it has the most observations, 176, whereas the other countries have less. To start the process, below of the hypotheses.

The null hypothesis is that the average home score for all countries is the same to Frances’s average home score.

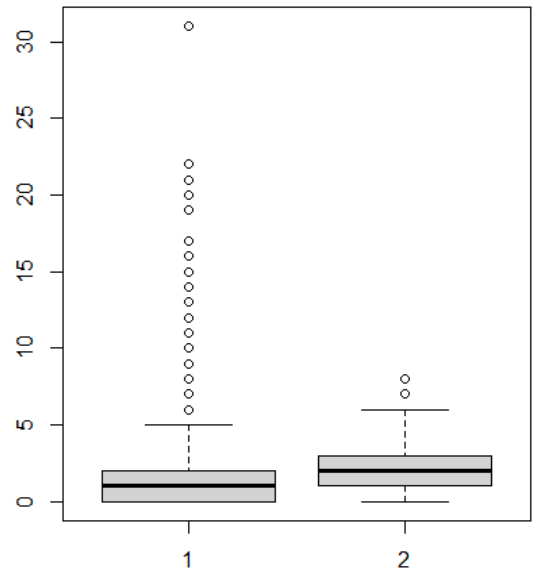
The alternate hypothesis is that there is a statistically significant difference in the average home score for all countries as compared to the average home score of France.

To first see the prediction of the hypothesis test, we produced a visual test. As talked about above, a boxplot is a good starting test to just see the difference between the means. We compared the average home scores of France (#2) and the total home score average (#1). There seems to be a difference between the two variables. The code that was produced is shown below.

```
homescore <- SoccerData$home_score

francescore <- SoccerData$home_score[SoccerData$home_team =="France"]

boxplot(homescore,francescore)
```



The parameter of interest is the mean of the two populations. The mean for the average home_scores was found to be 1.64. The mean home score for France was found to be 1.98. The sampling distribution that will be used is a t sampling distribution.

```
mean(SoccerData$home_score)

mean(SoccerData$home_score[SoccerData$home_team == "France"])
```

To begin the t-test procedure, certain measures must be in order. The normality is the first step. By looking at the two populations we can assume normality. France, having 176 observations, and the total population has way more than enough: 20,179 observations. Since both populations are above 30, we can assume normality. Even though the average of all home scores has multiple outliers, we can assume that they do not affect the data because there are so many observations. Having 17 outliers will not change the results of the 20,179 observations. The home scores of France also have 2 outliers. Unlike the total population, they are closer to the actual mean, so they do

not affect the data too much. However, we will still check our t-test results using randomization to construct our sample distribution.

The second check is to see if the two variables are independent. This is a true statement. The home scores of France will not influence the scores of all other countries.

Again, we are checking for randomization, so everything said previously is related except for the variable names and the population sizes. We will be using 20179 and 176 instead of 59 and 54. However, after taking the randomization test, the results that were expected were not provided. A p-value of 1 was produced explaining that there was no significant difference. Since there are a lot of outliers, especially some extreme ones that are more than three IQR's away from the mean, in both sets of data, it would be hard to take a test a get a good result, even if the test-taking involved randomization. So doing a t-test would be unwise because it would also provide a p-value of 1. A p-value of 1 means that the difference between the groups was due to chance. In context, this shows that team scores are not planned out. The French home scores cannot be predicted. Whether or not France has a good soccer team, predicting the average score per game is extremely difficult and therefore random.

```
set.seed(22)

num_sim <- 1000

diffs <- numeric(num_sim)

for(i in 1:num_sim){

  france <- sample(1:20179,176)

  france_mean <- mean(SoccerData$home_score[france])

  total_mean <- mean(SoccerData$home_score[-france])

  diffs[i] <- total_mean - france_mean

}

france_mean <- mean(SoccerData$home_score[SoccerData$home_team ==
"France"])

total_mean <- mean(SoccerData$home_score)

obs_diff <- total_mean - france_mean

pval <- mean((diffs >= obs_diff) | (diffs <= -obs_diff))

Pval

[1] 1
```

In the future, it would be wise to choose a set of data that has fewer outliers or choose to get rid of the outliers after more detail-oriented searching. A visual test can be a good starter, but clearly it cannot determine the actual results.

When we were looking at the measures of the spread and centers of our data, we used boxplots. These representations make it easier to see outliers and variance because outliers fall outside the whiskers on the box plot and the variance is shown by the height of the box. Since we were often looking at the meaning of the numerical data the boxplots were also helpful in seeing the differences in means side by side. The solid black line in the box plots show you the mean while something like a histogram requires a guess as to where the mean is in the data. Boxplots are straighter forward for the types of testing we did. A histogram is better at showing the normality of data, but in most cases, we were dealing with t-tests where we had enough observations that normality depended on outliers and not the sampling distribution itself.

When we did the mosaic plot, I chose a mosaic plot to view the data. The mosaic plot was a quick visual. I knew how to make for our data that shows easily the differences in proportions of the categorical data we were looking at. I keep the number of observations per type of tournament as similar as possible, so it was easy to see how the tournaments compared to each other in how many had neutral venues. It was the best reorientation I knew how to make that would show the proportion of neutral matches and non-neutral matches in an easy-to-read manner.

This data was collected by the person who made the data set to answer questions about who dominated what eras of international football and to answer questions about the geopolitics of soccer. They also wanted to be able to look at the difference that participating in friendly matches has on a team's performance in competitive tournaments. This data in the hands of people with more ideas on how to work with the date field and some more advanced techniques that do not rely so heavily on normally distributed data could be used to answer the questions it was made for. This data was a bit hard to work with because it didn’t follow a normal distribution as nicely as some other data we could have used, but it does answer questions it was created to answer. If I were to extend this data, I would want to add a categorical variable about whether a game went to a shootout at the end or not. This could let you look at whether there was a trend in the tournament type that goes to shootouts or if there was a certain team that has a higher tendency to go to shootouts. It would also be

nice to include how many people were thrown out of a game by a red card and how many left for injuries. In some instances, this could be a reason for some of the more extreme outliers. In some cases, teams could lose some of their best players to these instances and that could change the outcome of the game drastically.

In this lab we used t-tests over ANOVA. This comes down to the lack of normality in our data set that would make ANOVA not the best choice to look at the differences in multiple means. ANOVA is robust to the lack of normality, but we were not sure how safely we could stray from a normal model and still get results that meant something statistically. We stuck to the t-tests for the most part because we were able to do randomization to check our test when we were not sure about the normality of our data. We were not sure how to do the same thing for an ANOVA test.

In the case of the linear model, we did a linear model simply because this is the one, we learned in class. When you look at the relationship in the scatterplot, the relationship better follows some sort of exponential model. If we knew how to do this test correctly with other forms of best fit lines, we would have used one that fits the relationship between our two numerical variables, home score and away score.

In the case of the chi squared test, I chose this test again to avoid needing to worry as much about normality as I would have had to with other tests like a hypothesis test. We are really limited in a lower division statistics course to models that work best with normal distributions, so we decided to be careful with which ones to do in this assignment.

In our t-tests, we used randomized distributions to when we were unsure of our data matching the conditions for the test. Randomization helped us make sure that outliers were not too extreme to mess up our assumption that the data had normality. In the case of the linear model, our data does not fit the linear relationship we were really looking for. We were kind of stuck with a linear line as our line of best fit since it was the only one, we learned how to work with in this course. The chi squared test appears to for the conditions well. We had more than 5 expected observations and we were able to assume independence. Independence for all these tests is hard to prove but we felt good about assuming that there was independence within and between groups in this data set. Most of the tests we learned in this test are robust and can handle little deviations from the conditions you need to meet to use them. We tried to avoid places where the conditions would be too big of a stretch in our opinion.

The arguments we have in favor of our t tests come down to them in most cases matching a sampling distributions result that we made using randomization. The randomization allowed us to step away from our

actual data and make sure outliers were not pulling too hard on the results of our testing.