Final Project DSC520

Ori Rinaman

2021-03-7

# DSC520 Final Project

## Tell us why did you choose this research idea and the data set.

Over many years of experience in acadamia, I have discovered that when you are passionate about the topic it becomes fun to work on rather than a burden. I spent numerous hours browsing the web trying to find the perfect data set for this project. Late one night, I came across a data set of every NFL game from 1966 through this past Super Bowl. Growing up, I played football from PeeWee all the way until my senior year of high school. The game of football introduced me to some of my best friends along with memories that I will remember the rest of my life. I would also would watch every Pittsburgh Steeler football game with my father. Having this connection to football led me to choose this data set. The research idea that I am investigating with my data is to prove or disprove if homefield advantage is real. The data set clearly states each home and away team, so I thought this would be a very interesting investigation. To add on to that investigation, I am also intereste to see home homefield advantage affects teams in the playoffs

## Summarize the problem statement you addressed.

After selecting my data set, the next task was to develop questions that I would answer while analyzing my data. Having played sports all of my life, I had always wondered if homefield advantage is actually an advantage. In 2020, Covid-19 had a large impact on the world of sports, and especially the NFL. Covid-19 forced NFL teams to play games with little to no fans at games. Fans can play a very large roll in the success of a home team. Not only can they cheer on the team, but they can also cause disruptions to the other team. A good example of this can be seen with the Kansas City Chiefs. They play their games at Arrow Head Stadium which holds the World Record for the loudest stadium in the world. In games, opponents find it difficult to communicate with eachother on the field, and most of the time causes mishaps for the visiting team. Knowing that the full 2020 season played games with restrictions to fans at the games, I thought this would be a perfect oppurtunity to see if in fact homefield advantage is real or not. I first want to see if there is a statistical difference in winning percentage for the home team in the seasons without Covid-19 restrictions and games with Covid-19 restrictions. The second question I investigated is if homefield advantage is important in the playoffs. In most NFL seasons, teams take two approaches in the last few games if they have already secured a playoff spot. They can either rest their starters and risk losing one or two games, or teams can play to have guranteed home games in the playoffs excluding the Super Bowl because it is played at a neutral site which is already decided years in advance. The 2020 season was a bit different, because the Tampa Bay Buccaneers ended up playing in the Super Bowl as the visiting team at their home field. So is it worth risking your starters to get homefield or should you take a more relaxed approach and rest your starters for the last one or two games of the season.

## Summarize the methodology employed. Why do you think your method is appropriate?

After I created my questions, I soon found myself struggling to solve how I was going to prove my results significant. I knew that I would want to compare the winning percentages of the home teams in the two differetn data sets. After reading some extra chapters in Discovering Statistics Using R and some internet research, I made the decision that I was going to prove my results with the Two Proportions Test. The Two Proportions Test looks to see if two distinct populations will product different results. So my hypothesis for the test would be looking to see if a certain populations proportion is statistically different than the other.

## Summarize the implications to the consumer (target audience) of your analysis.

When I first discovered my data set, I thought I had found the perfect data set to use for my final project. The only real implication that I discovered was that it was hard to really relate the statistics of the project to those that we learned in class. I wanted to try and do some predictive analytics and see if I could predict the results of games using betting odds and homefield advantage. Soon, I came to the realization that the data was not in a form to do so, atleast from my thoughts. For one, the data set failed to note any significant charactoristics of the teams playing other than who was the home and away team, and who was favored by the people making the betting odds. It would have been nice to know the records of the teams and how their offenses and defenses ranked amongst other teams. I did make an attempt to see how good the bettors where at predicting the outcomes, but I kept running into issues with my code trying to work with the data.

## Overall, write a coherent narrative that tells a story with the data as you complete this section.

I started by bringing in the data set and loading in the R packages needed for this analysis.

nfl\_data <- read.csv("C:\\Users\\Rinaman\\Desktop\\spreadspoke\_scores.csv")  
library(tidyverse)

## -- Attaching packages ------------------------------------------------------------------------------------------------------------ tidyverse 1.3.0 --

## v ggplot2 3.3.1 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts --------------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(scales)

##   
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':  
##   
## discard

## The following object is masked from 'package:readr':  
##   
## col\_factor

head(nfl\_data)

## schedule\_date schedule\_season schedule\_week schedule\_playoff  
## 1 9/2/1966 1966 1 FALSE  
## 2 9/3/1966 1966 1 FALSE  
## 3 9/4/1966 1966 1 FALSE  
## 4 9/9/1966 1966 2 FALSE  
## 5 9/10/1966 1966 1 FALSE  
## 6 9/10/1966 1966 2 FALSE  
## team\_home score\_home score\_away team\_away team\_favorite\_id  
## 1 Miami Dolphins 14 23 Oakland Raiders   
## 2 Houston Oilers 45 7 Denver Broncos   
## 3 San Diego Chargers 27 7 Buffalo Bills   
## 4 Miami Dolphins 14 19 New York Jets   
## 5 Green Bay Packers 24 3 Baltimore Colts   
## 6 Houston Oilers 31 0 Oakland Raiders   
## spread\_favorite over\_under\_line stadium stadium\_neutral  
## 1 NA NA Orange Bowl FALSE  
## 2 NA NA Rice Stadium FALSE  
## 3 NA NA Balboa Stadium FALSE  
## 4 NA NA Orange Bowl FALSE  
## 5 NA NA Lambeau Field FALSE  
## 6 NA NA Rice Stadium FALSE  
## weather\_temperature weather\_wind\_mph weather\_humidity weather\_detail  
## 1 83 6 71   
## 2 81 7 70   
## 3 70 7 82   
## 4 82 11 78   
## 5 64 8 62   
## 6 77 6 82

As we all know, the first step to any analysis of data should be to clean the data. I first looked through the data and removed any NULL or NA values. Fortunatly, the data set had none of these value. In the begining of my analysis I decided right away that I was going to create two tables. The first table consisted of data from the seasons 1979 through 2019. This is what I considered to be “normal” because there was no pandemic happening during this time period. I then filtered out games that were played at a neutral site because for those games neither team had home field advantage. After that I had to use the mutate() function to create a column that included “KEEP” or “REMOVE” for games where teams that shared a stadium played eachother. Lastly I filtered out all of the playoff games so that all of the games that I focused on were from the regular season that way all teams would be included regardless if they were good enough to make the playoffs or not. The code and first rows of the data of non-Covid-19 games is seen below…

reg\_season <- nfl\_data %>%  
 filter(schedule\_season > 1978 & schedule\_season < 2020) %>%  
 filter(stadium\_neutral == FALSE) %>%  
 mutate(remove\_game = ifelse(team\_home == "New York Giants" & team\_away == "New York Jets", "REMOVE",   
 ifelse(team\_home == "New York Jets" & team\_away == "New York Giants", "REMOVE", "KEEP"))) %>%  
 filter(remove\_game == "KEEP") %>%  
 select(-remove\_game) %>%  
 filter(schedule\_playoff == FALSE)  
  
head(reg\_season)

## schedule\_date schedule\_season schedule\_week schedule\_playoff  
## 1 9/1/1979 1979 1 FALSE  
## 2 9/2/1979 1979 1 FALSE  
## 3 9/2/1979 1979 1 FALSE  
## 4 9/2/1979 1979 1 FALSE  
## 5 9/2/1979 1979 1 FALSE  
## 6 9/2/1979 1979 1 FALSE  
## team\_home score\_home score\_away team\_away  
## 1 Tampa Bay Buccaneers 31 16 Detroit Lions  
## 2 Buffalo Bills 7 9 Miami Dolphins  
## 3 Chicago Bears 6 3 Green Bay Packers  
## 4 Denver Broncos 10 0 Cincinnati Bengals  
## 5 Kansas City Chiefs 14 0 Baltimore Colts  
## 6 Los Angeles Rams 17 24 Oakland Raiders  
## team\_favorite\_id spread\_favorite over\_under\_line stadium  
## 1 TB -3 30.0 Houlihan's Stadium  
## 2 MIA -5 39.0 Ralph Wilson Stadium  
## 3 CHI -3 31.0 Soldier Field  
## 4 DEN -3 31.5 Mile High Stadium  
## 5 KC -1 37.0 Arrowhead Stadium  
## 6 LAR -4 36.5 Anaheim Stadium  
## stadium\_neutral weather\_temperature weather\_wind\_mph weather\_humidity  
## 1 FALSE 79 9 87  
## 2 FALSE 74 15 74  
## 3 FALSE 78 11 68  
## 4 FALSE 69 6 38  
## 5 FALSE 76 8 71  
## 6 FALSE 70 10 77  
## weather\_detail  
## 1   
## 2   
## 3   
## 4   
## 5   
## 6

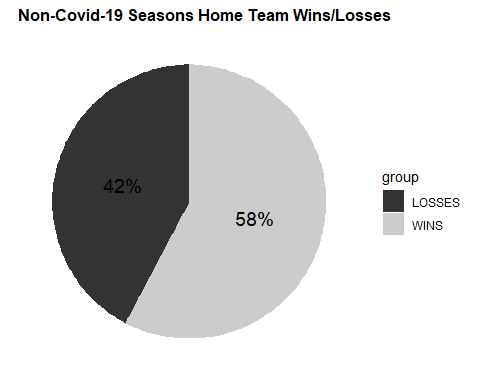
After looking at my data, I then decided to investigate the home teams record against the away team. This task involved using the mutate() function once again to create a column called “WIN” in my data set. Inside of the mutate function I added an ifelse() statement that compared the home score and away score of all the games. If the statement returned TRUE then it would be labled as a “WIN” for the home team. If the statement returned FALSE, then it would be labeled as a “LOSS” for the home team. After the column was created and the values were added, I then used the count() function to see the count of the home teams wins and losses.

win\_loss\_reg <- reg\_season %>%  
 mutate(WIN = ifelse(reg\_season$score\_home > reg\_season$score\_away, "WIN", "LOSS"))  
  
count(win\_loss\_reg, WIN)

## WIN n  
## 1 LOSS 4097  
## 2 WIN 5590

Since I was dealing with proportions, I decided that the best way to visualize this was to use a pie chart. A pie chart is a type of graph in which a circle is divided into sectors that wach represent a proportion of the whole. Even though R’s ggplot2 package does not have a pie chart add in, I discovered a way to do this using polar coordinates. To accomplish this I created a data frame that showed the counts of the “WINS” and “LOSSES” of the home team. Then I used ggplot to create a bar plot of the proportions of “WINS” and “LOSSES”. After I had my bar plot, I next added the function coord\_polar() to make the pie chart.

df\_pie <- data.frame(  
 group = c("WINS", "LOSSES"),   
 value = c(5590, 4097)  
)  
  
bp <- ggplot(df\_pie, aes(x = "", y = value, fill = group)) +  
 geom\_bar(width = 1, stat = "identity")  
  
blank\_theme <- theme\_minimal()+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=12, face="bold")  
 )  
  
pie\_chart <- bp +coord\_polar("y", start = 0)  
pie\_chart + scale\_fill\_grey() + blank\_theme +  
 theme(axis.text.x=element\_blank()) +   
 geom\_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),   
 label = percent(value/9687)), size=5) +  
 ggtitle("Non-Covid-19 Seasons Home Team Wins/Losses")

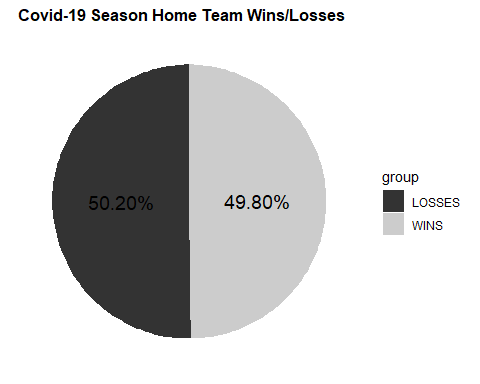


Earlier I had mentioned that I would be creating two tables for the analysis. The second table went through the exact same cleaning and aggregating steps as the first table, execpt we only kept games from the 2020 season. The 2020 season occured during the Covid-19 pandemic and little to no fans were allowed into games. I went through the same steps as the first table and was able to get counts for the wins and losses for the home team during this season. I also created a similar pie chart to visualize the proportions of wins and losses.

covid\_season <- nfl\_data %>%  
 filter(schedule\_season == 2020) %>%  
 filter(stadium\_neutral == FALSE) %>%  
 filter(schedule\_playoff == FALSE) %>%  
 mutate(WIN = ifelse(score\_home > score\_away, "WIN", "LOSS"))  
  
count(covid\_season, WIN)

## WIN n  
## 1 LOSS 128  
## 2 WIN 127

df\_pie2 <- data.frame(  
 group = c("WINS", "LOSSES"),   
 value = c(127, 128)  
)  
  
bp2 <- ggplot(df\_pie2, aes(x = "", y = value, fill = group)) +  
 geom\_bar(width = 1, stat = "identity")  
  
blank\_theme <- theme\_minimal()+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=12, face="bold")  
 )  
  
pie\_chart2 <- bp2 +coord\_polar("y", start = 0)  
pie\_chart2 + scale\_fill\_grey() + blank\_theme +  
 theme(axis.text.x=element\_blank()) +   
 geom\_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),   
 label = percent(value/255)), size=5) +  
 ggtitle("Covid-19 Season Home Team Wins/Losses")



At this step, I was lost as to what statistical test I should use to verify if the results of the two different proportions between the data sets are different. In the book, I discovered a 2-sample test for quality of proportions with continuity correction. This test is used to compare two observed proportions. To run this test I used the prop.test() function. Since I wanted to test to see of the proportion of wins in the non-Covid-19 seasons were greater than the Covid-19 season I added alternative = “greater” inside the function to make it a one tailed test. So the Null hypothesis would be that the two proportions are equal to eachother and the Alternative hypothesis would be that one of the proportions is greater than the other. The resulting p-value was a 0.007, which is below our significance level of 0.05 so we would reject the null and assume with confidence that statistically the proportion of wins in the non-Covid-19 seasons is greater than the proportion of wins during the Covid-19 season.

# Ho: Non\_Covid\_Reg\_Season <= Covid\_19\_Reg\_Season  
# Ha: Non\_Covid\_Reg\_Season > Covid\_19\_Reg\_Season  
prop.test(x = c(5590, 127),   
 n = c(9687, 255),   
 alternative = "greater")

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(5590, 127) out of c(9687, 255)  
## X-squared = 6.0299, df = 1, p-value = 0.007033  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## 0.02485088 1.00000000  
## sample estimates:  
## prop 1 prop 2   
## 0.5770620 0.4980392

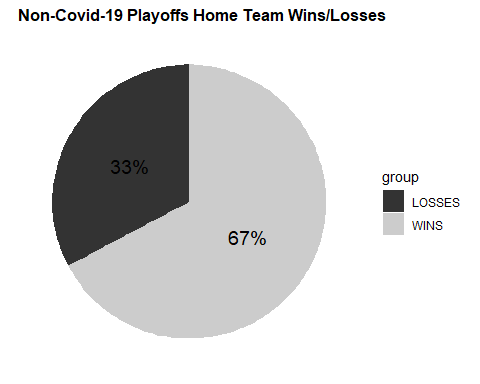
The second question that I tasked myself with answering is if homefield advantage is important in the playoffs? As I had mentioned earlier, at the end of the NFL season, teams prepare for the playoffs. Some teams make the decision to rest their start players risking taking a loss, while other teams play their starts to fight for a chance at homefield advantage. First, let me explain how a team makes the playoffs. In each conference there are four divisions. Each of these divisions contains four teams. In total there are eight divisions in the NFL, four in the AFC and four in the NFC. The winners of each conference are guarenteed a playoff spot, while there are two spots left over in each conference for wildcars or the two teams with the best records that did not make the playoffs. The two division winners with the best record are gifted homefield advatange. There was a change in the 2020 season though. The NFL added an extra wildcard spot in each conference. So now there are three wildcards in each division rather than two. Also, only the team with the best record in each conference gets homefield advantage. All other playoff games are played at the team wtih the best record between match-ups. Since, I showed in my first analysis that it is plausable that having homefield advatage with the fans improves your chances of winning when comparing the non-Covid-19 Seasons to the Covid-19 season. Now I am looking at the playoffs to see how they were effected by homefield advantage and if infact teams should try to get the best record possible for the playoffs.

To conduct this analysis I created two more data sets that were deemed playoff data sets. One was for the Non-Covid-19 seasons, while the other was designated to the Covid-19 season. I cleaned and filtered the data almost the exact same as with the first analysis, but instead of filtering out playoff games, I only included playoff games. The Super Bowls were not included because they are played at a neutral site every year. Using similar steps as the first analysis, I found the count of wins of the home teams in the playoffs for both data sets and then also created pie charts for each. The reuslts can be seen below…

#Non-Covid-19 Playoffs data frame  
reg\_playoff\_df <- nfl\_data %>%  
 filter(schedule\_season > 1978 & schedule\_season < 2020) %>%  
 filter(stadium\_neutral == FALSE) %>%  
 mutate(remove\_game = ifelse(team\_home == "New York Giants" & team\_away == "New York Jets", "REMOVE",   
 ifelse(team\_home == "New York Jets" & team\_away == "New York Giants", "REMOVE", "KEEP"))) %>%  
 filter(remove\_game == "KEEP") %>%  
 select(-remove\_game) %>%  
 filter(schedule\_playoff == TRUE)  
  
win\_loss\_playoff\_reg <- reg\_playoff\_df %>%  
 mutate(WIN = ifelse(reg\_playoff\_df$score\_home > reg\_playoff\_df$score\_away, "WIN", "LOSS"))  
  
count(win\_loss\_playoff\_reg, WIN)

## WIN n  
## 1 LOSS 129  
## 2 WIN 265

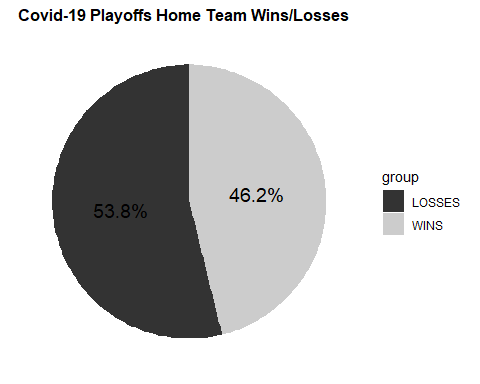
#Non-Covid-19 Playoffs Pie Chart  
df\_pie3 <- data.frame(  
 group = c("WINS", "LOSSES"),   
 value = c(265, 129)  
)  
  
bp2 <- ggplot(df\_pie3, aes(x = "", y = value, fill = group)) +  
 geom\_bar(width = 1, stat = "identity")  
  
blank\_theme <- theme\_minimal()+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=12, face="bold")  
 )  
  
pie\_chart2 <- bp2 +coord\_polar("y", start = 0)  
pie\_chart2 + scale\_fill\_grey() + blank\_theme +  
 theme(axis.text.x=element\_blank()) +   
 geom\_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),   
 label = percent(value/394)), size=5) +  
 ggtitle("Non-Covid-19 Playoffs Home Team Wins/Losses")



# Covid-19 Playoffs  
cov\_playoff\_df <- nfl\_data %>%  
 filter(schedule\_season == 2020) %>%  
 filter(stadium\_neutral == FALSE) %>%  
 mutate(remove\_game = ifelse(team\_home == "New York Giants" & team\_away == "New York Jets", "REMOVE",   
 ifelse(team\_home == "New York Jets" & team\_away == "New York Giants", "REMOVE", "KEEP"))) %>%  
 filter(remove\_game == "KEEP") %>%  
 select(-remove\_game) %>%  
 filter(schedule\_playoff == TRUE)  
  
win\_loss\_playoff2 <- cov\_playoff\_df %>%  
 mutate(WIN = ifelse(cov\_playoff\_df$score\_home > cov\_playoff\_df$score\_away, "WIN", "LOSS"))  
  
count(win\_loss\_playoff2, WIN)

## WIN n  
## 1 LOSS 7  
## 2 WIN 6

# Covid-19 Playoffs Pie Chart  
df\_pie4 <- data.frame(  
 group = c("WINS", "LOSSES"),   
 value = c(6, 7)  
)  
  
bp2 <- ggplot(df\_pie4, aes(x = "", y = value, fill = group)) +  
 geom\_bar(width = 1, stat = "identity")  
  
blank\_theme <- theme\_minimal()+  
 theme(  
 axis.title.x = element\_blank(),  
 axis.title.y = element\_blank(),  
 panel.border = element\_blank(),  
 panel.grid=element\_blank(),  
 axis.ticks = element\_blank(),  
 plot.title=element\_text(size=12, face="bold")  
 )  
  
pie\_chart2 <- bp2 +coord\_polar("y", start = 0)  
pie\_chart2 + scale\_fill\_grey() + blank\_theme +  
 theme(axis.text.x=element\_blank()) +   
 geom\_text(aes(y = value/2 + c(0, cumsum(value)[-length(value)]),   
 label = percent(value/13)), size=5) +  
 ggtitle("Covid-19 Playoffs Home Team Wins/Losses")



Similar to the first quesion I answered, I also leaned on the two-proportion test for statistical significance. After running a one sided proportion test, the resuling p-value was 0.09 whcih is above our significance level of 0.05. The tells us that we need to reject the alternative, and assume the null hypothesis to be true. So there was no statistical differnce in winning proportions of home teams of the different season with or without Covid-19.

# Ho: Non\_Covid\_Playoff\_Seasons <= Covid\_19\_Playoff\_Season  
# Ha: Non\_Covid\_Playoff\_Seasons > Covid\_19\_Playoff\_Season  
prop.test(x = c(265, 6),   
 n = c(394, 13),   
 alternative = "greater")

## Warning in prop.test(x = c(265, 6), n = c(394, 13), alternative = "greater"):  
## Chi-squared approximation may be incorrect

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(265, 6) out of c(394, 13)  
## X-squared = 1.6601, df = 1, p-value = 0.09879  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.05940512 1.00000000  
## sample estimates:  
## prop 1 prop 2   
## 0.6725888 0.4615385

## Summarize the interesting insights that your analysis provided.

This was an extreamly long analysis because I had do make sure my data was filtered perfectly to make the results unbiased. After some hard work, I finally was able to produce results which were suprising. The insight, I discovered was that homefield advantage is in fact important in the NFL. Looking at the pie charts, it is easy to see visually that when the home team did not have their fans present (eliminating most advantages of the home team) the winning percentage of the home team dropped by almost 10%. I was shocked at these results, because it shows how important having homefield with your fans is to the team’s success in regular season games. Aother interesting insight I gained from my analysis is that homefield advantage was significant in the non-playoff game data sets, but it was not significant in the playoff data sets.

An insight gained from other analyses of the data, focused on who the bet makers tended to favor when looking at home or away teams. Suprisingly, the home team was favored in about 58% of the games which is roughly similar to the percentage of games won by the home teams in the non Covid Regular season games. Maybe the book keepers new that homefield advantage existed after all.

reg\_season\_bet <- reg\_season %>%  
 mutate(fav\_id = ifelse(team\_favorite\_id == "GB", "Green Bay Packers",  
 ifelse(team\_favorite\_id == "ARI", "Arizona Cardinals",   
 ifelse(team\_favorite\_id == "ATL", "Atlanta Falcons",   
 ifelse(team\_favorite\_id == "BAL", "Baltimore Ravens",   
 ifelse(team\_favorite\_id == "BUF", "Buffalo Bills",   
 ifelse(team\_favorite\_id == "CAR", "Carolina Panthers",   
 ifelse(team\_favorite\_id == "CHI", "Chicago Bears",   
 ifelse(team\_favorite\_id == "CIN", "Cincinnati Bengals",   
 ifelse(team\_favorite\_id == "CLE", "Cleveland Browns",   
 ifelse(team\_favorite\_id == "DAL", "Dallas Cowboys",   
 ifelse(team\_favorite\_id == "DEN", "Denver Broncos",   
 ifelse(team\_favorite\_id == "DET", "Detroit Lions",   
 ifelse(team\_favorite\_id == "HOU", "Houston Texans",   
 ifelse(team\_favorite\_id == "IND", "Indianapolis Colts",   
 ifelse(team\_favorite\_id == "JAX", "Jacksonville Jaguars",   
 ifelse(team\_favorite\_id == "KC", "Kansas City Chiefs",   
 ifelse(team\_favorite\_id == "MIA", "Miami Dolphins",   
 ifelse(team\_favorite\_id == "MIN", "Minnesota Vikings",   
 ifelse(team\_favorite\_id == "NE", "New England Patriots",   
 ifelse(team\_favorite\_id == "NO", "New Orleans Saints",   
 ifelse(team\_favorite\_id == "NYG", "New York Giants",   
 ifelse(team\_favorite\_id == "NYJ", "New York Jets",   
 ifelse(team\_favorite\_id == "OAK", "Oakland Raiders",   
 ifelse(team\_favorite\_id == "LVR", "Las Vegas Raiders",   
 ifelse(team\_favorite\_id == "PHI", "Philadelphia Eagles",   
 ifelse(team\_favorite\_id == "PIT", "Pittsburgh Steelers",   
 ifelse(team\_favorite\_id == "SD", "San Diego Chargers",   
 ifelse(team\_favorite\_id == "LAC", "Los Angeles Chargers",   
 ifelse(team\_favorite\_id == "SF", "San Francisco 49ers",   
 ifelse(team\_favorite\_id == "SEA", "Seattle Seahawks",   
 ifelse(team\_favorite\_id == "LAR", "Los Angeles Rams",   
 ifelse(team\_favorite\_id == "TB", "Tampa Bay Buccaneers",   
 ifelse(team\_favorite\_id == "TEN", "Tennesee Titans",   
 ifelse(team\_favorite\_id == "WAS", "Washington Redskins",   
 ifelse(team\_favorite\_id == "PICK", "NO FAV.", "K")))))))))))))))))))))))))))))))))))) %>%  
 mutate(is\_home\_fav = ifelse(fav\_id == team\_home, "HOME TEAM FAVORED",   
 ifelse(fav\_id == "NO FAV.", "NO FAVORITE", "AWAY TEAM FAVORED")))  
count(reg\_season\_bet, is\_home\_fav)

## is\_home\_fav n  
## 1 AWAY TEAM FAVORED 3890  
## 2 HOME TEAM FAVORED 5656  
## 3 NO FAVORITE 141

## Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

My novice experience to the world of data science I feel had some impact on my analysis. It really limited me, especially from an R skill perspective, to manipulate the dataset into a form that was usable for different purposes. I have no doudt that if I had better R skills and more experience on working with real life data sets (not just acadamia data sets) I could have been able to rework the data set into a form that was suitable for predictive analytics. The data set from the outside looked very promising because it contained multiple rows of data, but the form of the data set handicapped it from gaining any further insight. Looking forward in ways someone else or myself could improve this analysis would possibly be to combine this data set with another data set that included statistics, rankings, and records of the teams playing. Also, I think that the data set can be manipulated in some way to make it useful for logistical regression, but it would need someone with better R skills than myself to accomplish this. With added data and better R skills I think that it would make a very interesting project to see if a model can be created to predict the winner of games using homefield advantage as a variable. Another limitation came from the second half of my analysis from the Covid-19 Playoff data set. Since it only included one playoff season there were only 13 games. I feel that the reuslts would be better if I had more observations like the regular season.

Overall I would like to conclude that homefield advantage does exist because there is statistical evidence from my analysis that the proportion of home team wins in normal stadium conditions is greater than the proportion of home wins in restricted stadium condiontions. In the playoffs homefield advantage proportions between the different stadiu conditions statistically did not differ between teams. I would like to add thatI learned an abundance of useful skills for cleaning and manipulating data to make it into a useable form. I also learned how to create a pie chart visualization that is not directly in the ggplot2 package. I had to learn how to use polar coordinates to manipulate a barchart to accomlish this. Finally I would like to thank you professor for the enjoyable semester and all of the help along the way. Thank you and I am looking forward to my career as a data scientist.