The Relationship Between Stadium Attendance and City Population In Relation to Winning Percentage of NFL Teams (2010-2019)

Jake Sturges, William Bressner, Donovan Carmichael

1 Literature Review

The success of a National Football League (NFL) team is influenced by a variety of factors, including team performance, market size, fan support, and financial resources. Among these, stadium attendance and city population are frequently cited as significant contributors towards a team's success. This literature review explores the relationship between stadium attendance, city population, and the success of NFL teams between 2010 and 2019.

Overall, a team in a big city is more likely to be successful than one in a small city. However, this trend does not always hold true, and there are notable exceptions where teams in smaller cities can outperform competitors located in larger cities. For example, the Pittsburgh Steelers have the most titles and are in the bottom half of NFL city populations. Similarly, the Green Bay Packers play in the smallest city of any professional sports team, yet are still considered one of the most successful NFL franchises.

When analyzing the relationship between stadium attendance and NFL win percentage, the more a team wins, the more people will show up to the game to watch, and if a team is winning less, less people will want to go to the game. An increased fan turnout contributes to higher attendance rates, which in turn generate more revenue for the team. This revenue, whether from ticket sales, merchandise, or concessions, enhances the financial resources available to the franchise, providing them with the ability to invest in better players, improved facilities, and more advanced training programs, all of which contribute to better success. On the other hand, teams that are losing regularly face the opposite effect, with declining attendance leading to reduced revenue that can make it harder for them to improve their roster and win more.

There are some confounding variables that could influence the data. One possible confounding variable is home field advantage, while another is team management.

In summary, smaller cities with adequate attendance have proven that size does not always correlate with performance. Future research could explore the relationship between these variables in greater depth.

2 Methodology

To investigate whether there is a correlation between stadium attendance and win percentage, we will use data collected from ESPN and US Census. We are analyzing our data to explore our research questions, “Between the years 2010 and 2019, is there a relationship between city population and regular season record in the NFL?” and “In the NFL, is there a relationship between the average stadium attendance per game and the regular season record between the years 2010 and 2019?” In order to analyze our data and answer these questions, we will find descriptive statistics, perform a t-test, and create data visualizations. Our two explanatory variables in this study will be stadium attendance and city population while our response variable will be a team’s season winning percentage.

3 FARE and CARE Principles

The data we collected is pretty findable, as the data we analyzed from the ESPN website was easily accessible. The data is found in an appropriate repository. The meta data does a fairly good job of using language that we can understand and information we can gain for our data analysis. The data provides data usage licenses and sources on the websites.

The data ecosystems function in ways that do not harm Indigenous Peoples. The data clearly respects Indigenous people and their rights. The data clearly outlines the inclusion and support of indigenous people. The data includes the NFL Team’s name change from the Redskins to the Washington Commanders in 2022. The US Census data from 2010 to 2019 abides by the CARE Principles developed by the Global Indigenous Data Alliance.

4 Data Exploration

We began by making visualizations that would help us understand the spread of our data, before drawing any conclusions. We chose to do this to identify potential outliers and observations which could skew our data. To analyze the distribution of each variable, Yearly Win Percentage, Yearly City Population, and Yearly Average Stadium Attendance we made histograms.

Figure 1: Figure 1

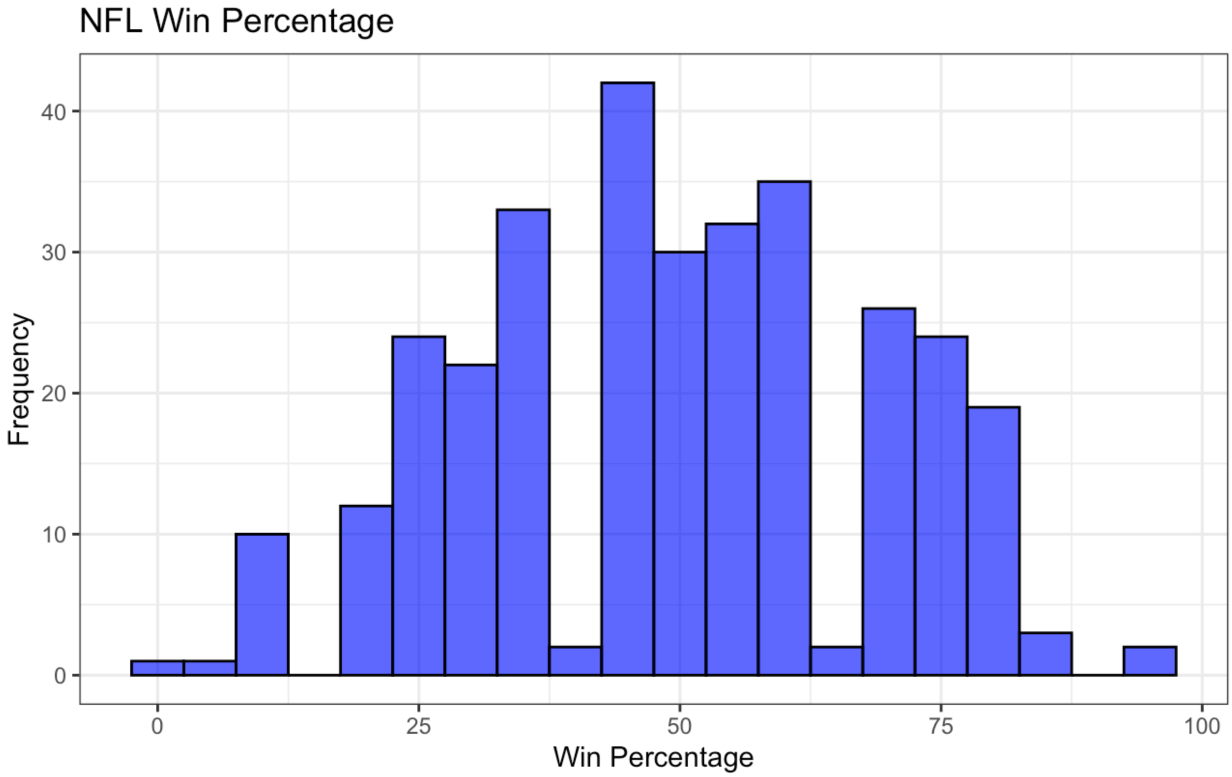


Figure 1, displays a relatively normal spread of Win Percentages. Although there are a few deviations in the spread, since it is mostly normal, we are able to use trust this data more so than if there were obvious outliers.

Figure 2: Figure 2

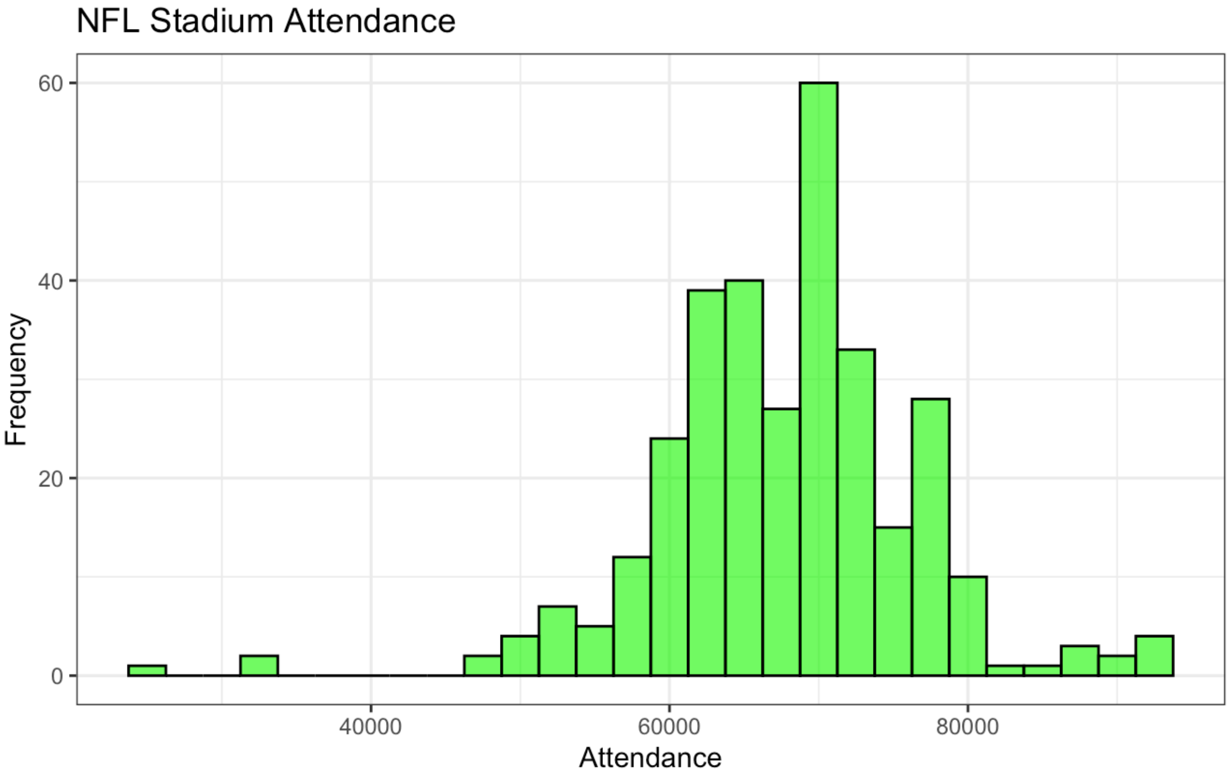


Figure 2 is a display of the yearly average attendance for each stadium. This data does appear relatively normal as well, but there are some extremely low outliers which we will keep an eye on. These attendance numbers represent the years in which the Rams first moved to Los Angeles. This makes sense that there attendance would be low since they had not yet established a fan base. We will keep an eye on this as we continue to produce more results.

Figure 3: Figure 3

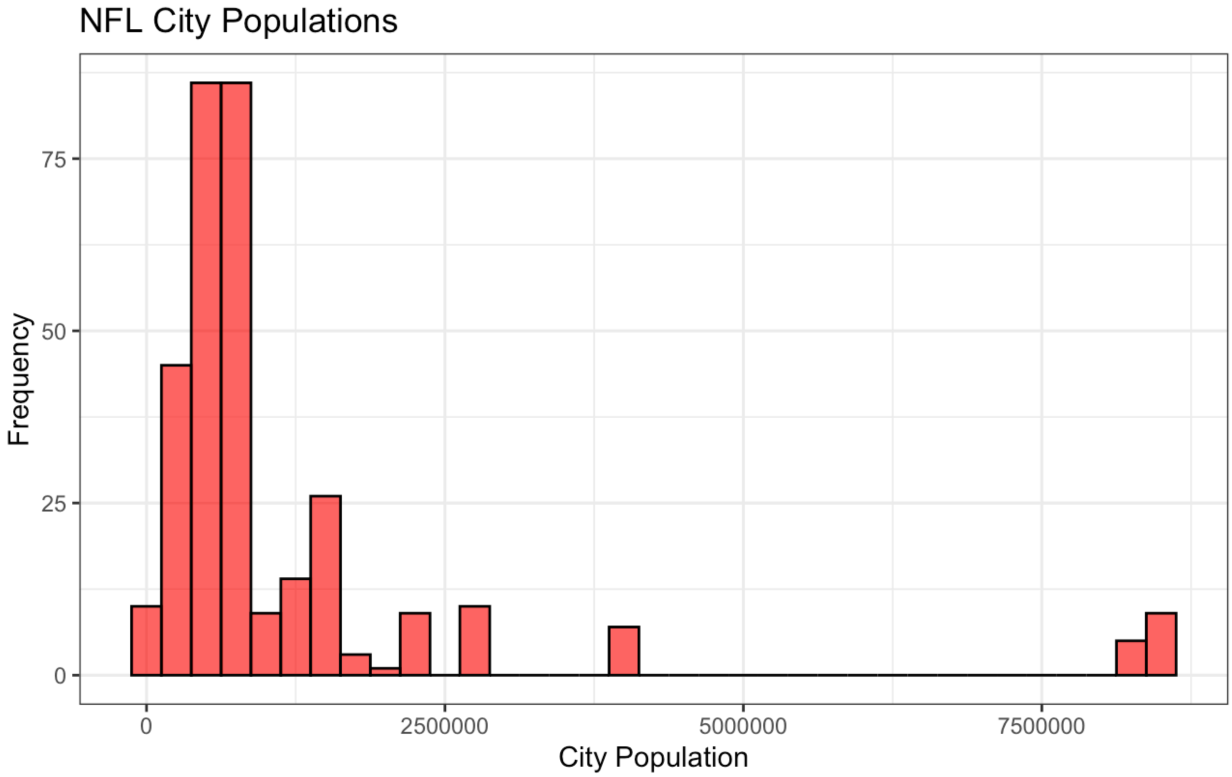
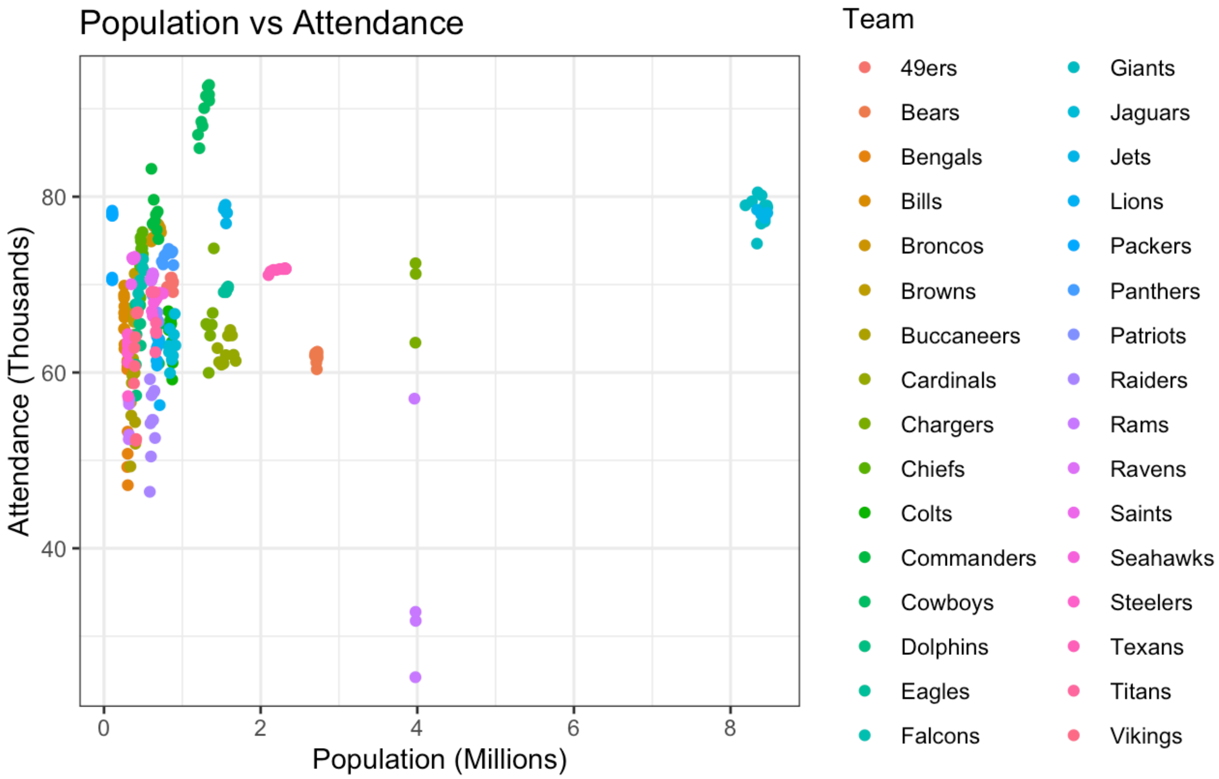


Figure 3 represents the population reports of the cities with NFL teams according to the US Census between the years 2010 to 2019. There is not as obvious variation in this data, largely because the scale is thrown off by outlier cities of New York and Los Angeles. We will keep an eye on the New York outlier since they have such a high population, and will therefore likely be influential on the correlation between population and win percentage.

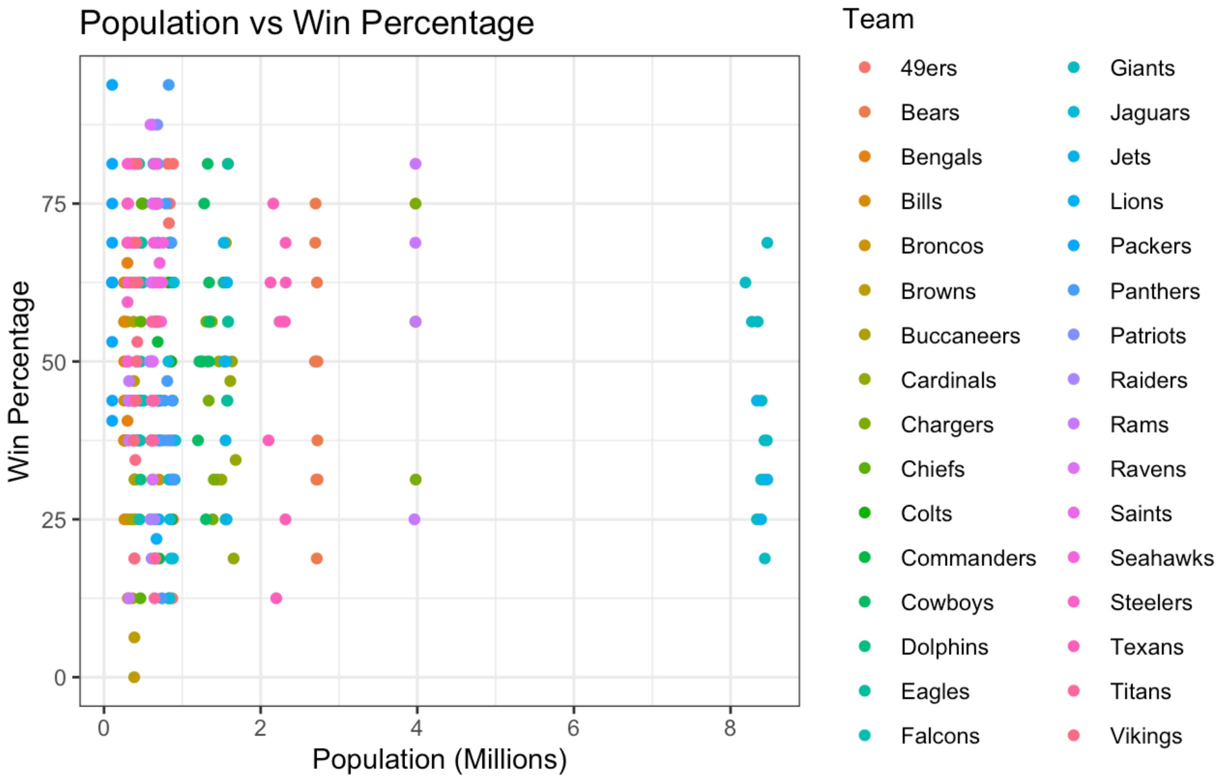
After we did our preliminary exploration of the variables in the data, we began to explore the correlations between them. We did this using scatter plots since trends tend to be clearer in them.

Figure 4: Figure 4



In the above graph, Figure 4, the relationship between population is both made clear but also confusing. In the first cluster, there appears to be a strong upward trend in the data. However, this is ruined by the teams in Los Angeles and New York which don't have significantly higher attendance, but they do have extremely high populations, masking the correlation.

Figure 5: Figure 5



In our preliminary graph to explore the population vs. win percentage, Figure 5, we were very surprised at the seemingly negative correlation. This went against our initial assumption that a larger population would be beneficial to the teams win percentage. However we will continue to test since there may be no correlation at all, or we may be being misled by this scatter plot.

Figure 6: Figure 6

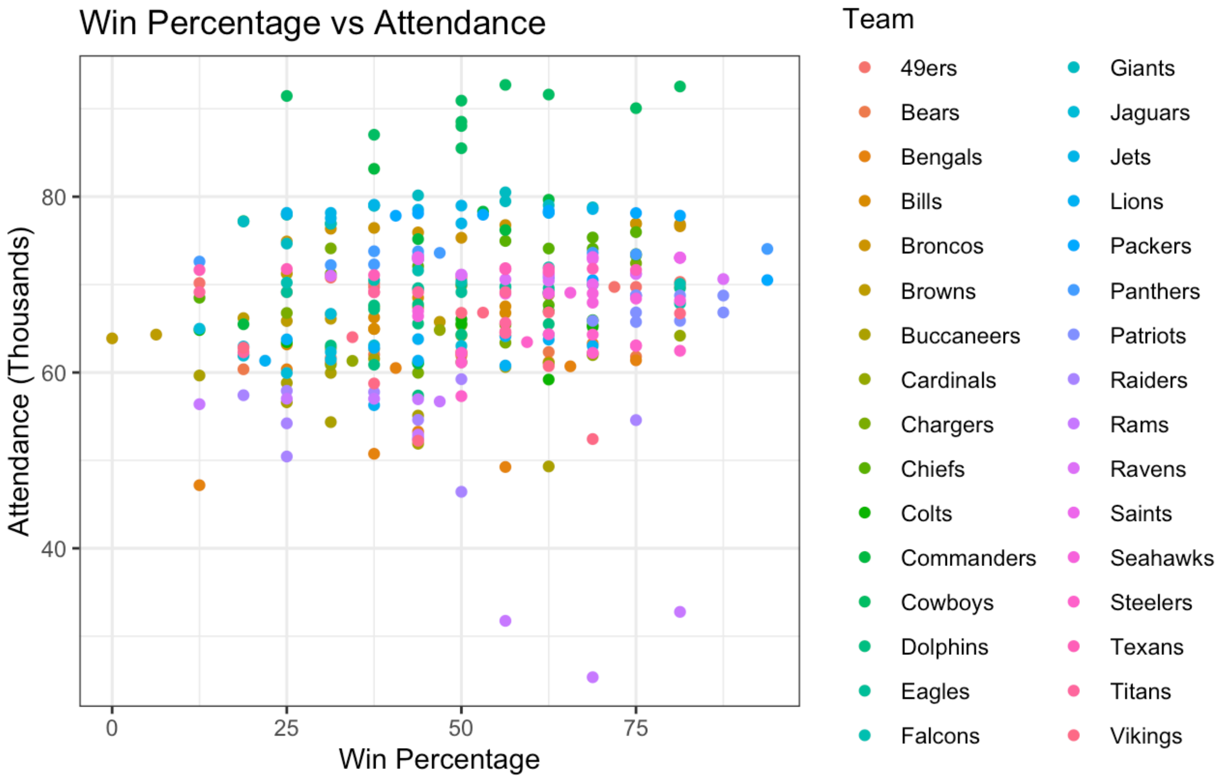
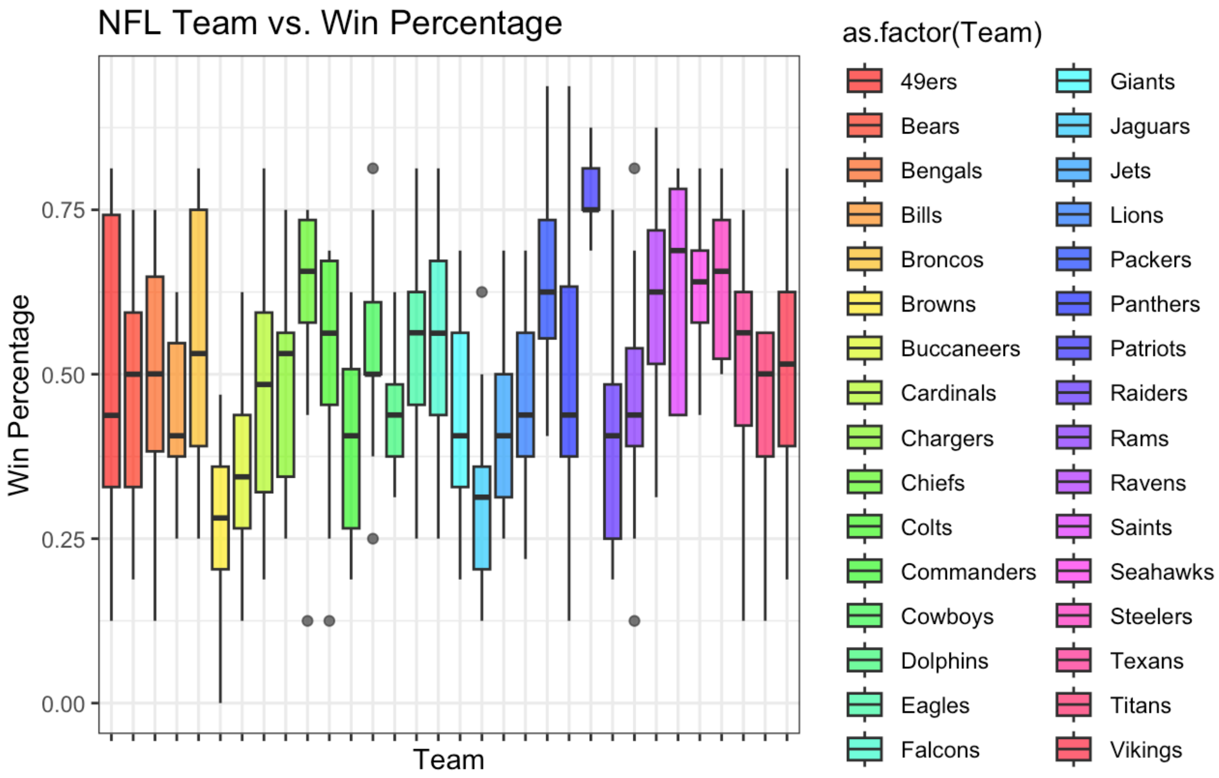


Figure 6, is giving us exactly what we would expect. There is a small upward trend in the data, which means that a higher win percentage, results in a higher attendance. This makes sense as a better team would attract more fans because people want to see the team win. Still the trend is not steep since the fan bases are loyal and mush of the time will show up no matter the expected result.

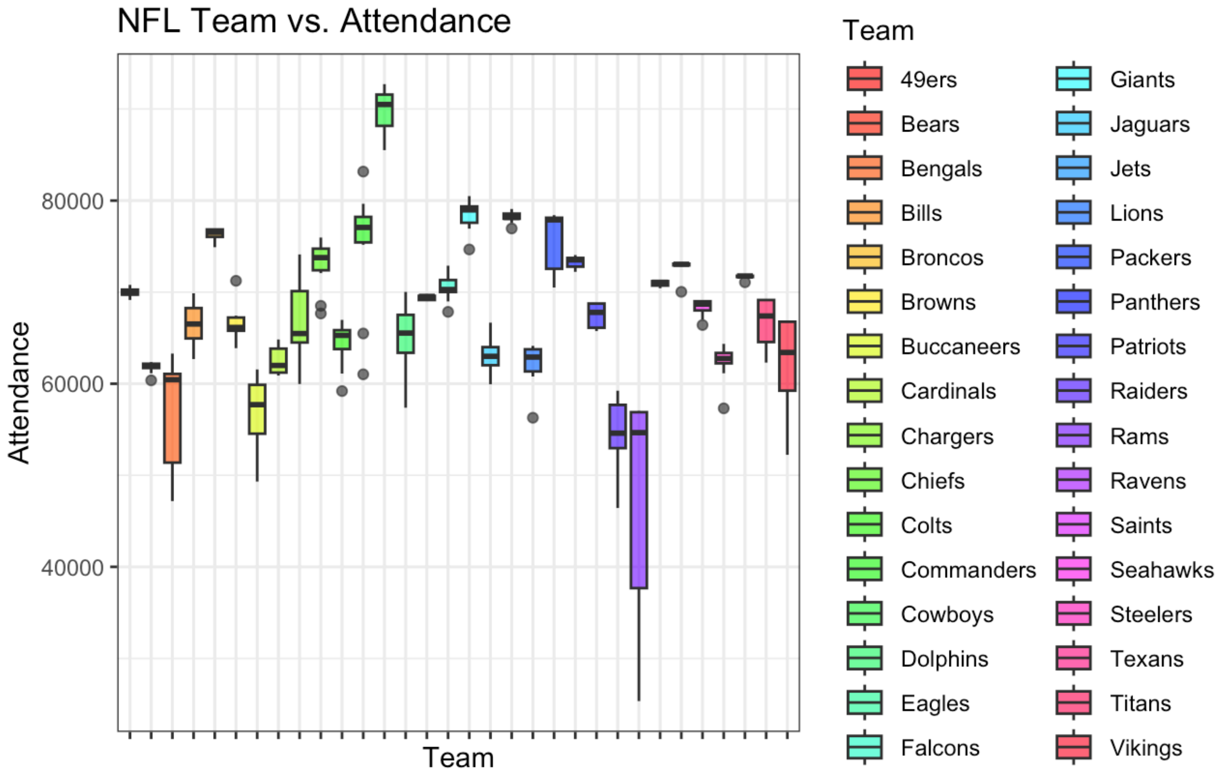
Following this we looked at how the team could be a strongly correlated with each of these variables, and therefore may be a confounding one. To do this we used a series of box plots

Figure 7: Figure 7



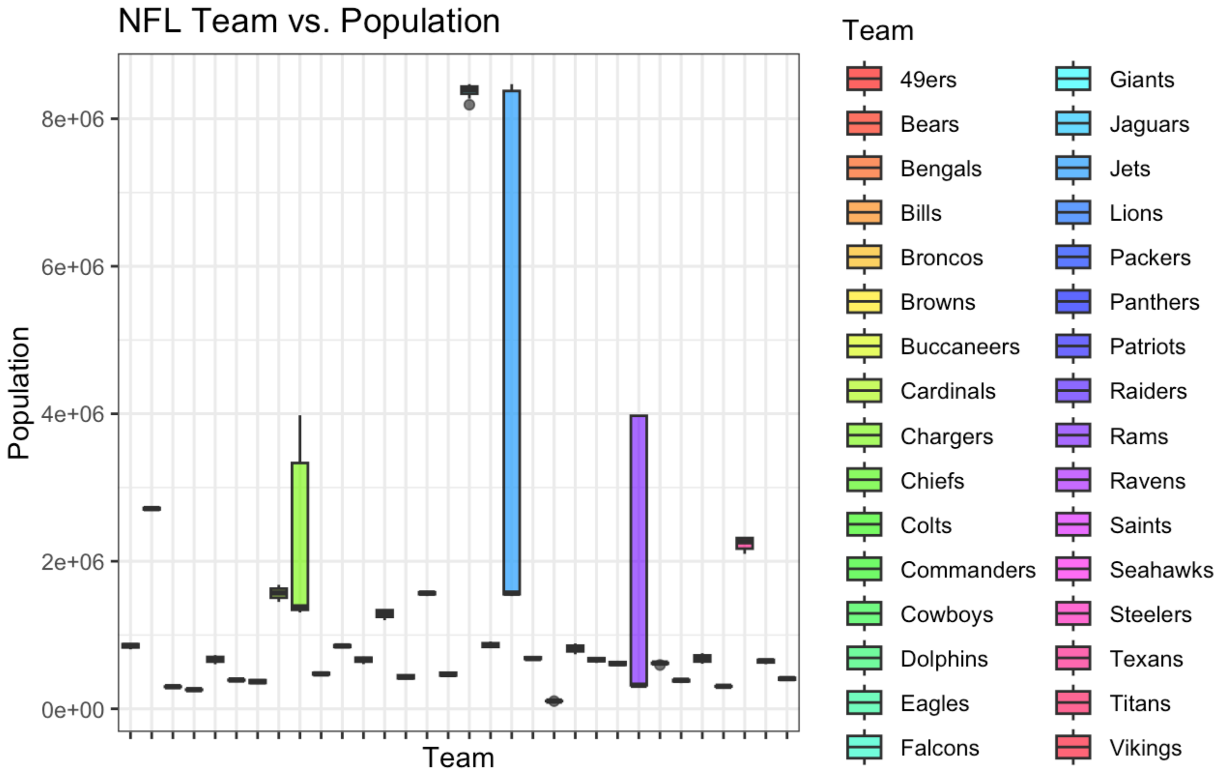
Based on Figure 7, there is team by team trends, however, there is a general line that is followed. There are some outliers, like the Patriots which have a much higher win percentage on average.

Figure 8: Figure 8



According to the above image, the stadium attendance is largely based on the team, much more so than the win percentage was, which suggests that there is a strong correlation between team and attendance. This has two main reasons: stadium size and fan base loyalty. You can only average as much as your stadium can seat, so the bigger stadiums will almost always seat more, like the cowboys. Also loyal fan bases will consistently show out to games. Teams with large spreads show a team that moved, like the Rams which show large variation.

Figure 9: Figure 9



In the above image, Figure 9, we can clearly see that populations among teams are relatively constant, except for teams that have moved, and a few exceptions. This team specific population data is something we will need to keep in mind as it can skew anything where correlation with population is involved.

5 Analysis

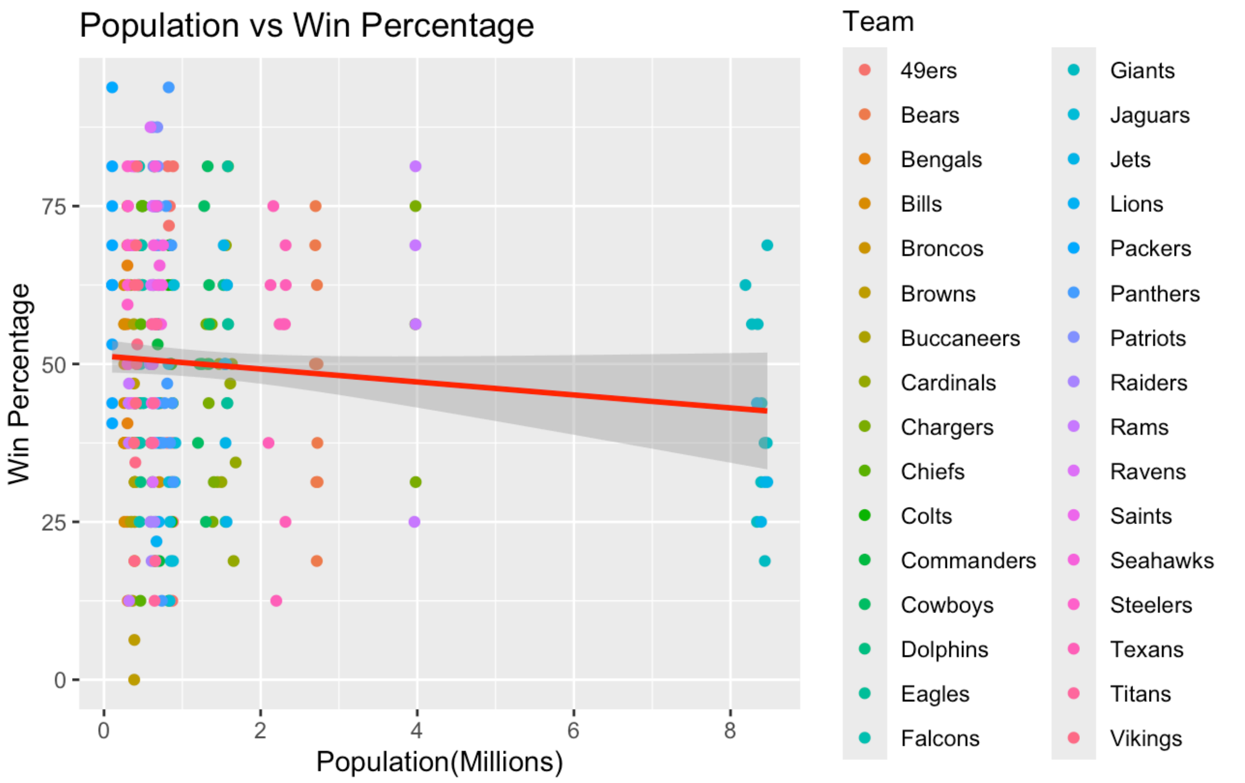
To explore our research questions we first must define our null hypothesis. For both questions we are assuming that there is no correlation between attendance and win percentage or population and win percentage, therefore our null hypothesis follows:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

for both attendance and win percentage. Beta 1 represents the slope for the regression line and we will be testing if it is significant with an alpha level of 0.05.

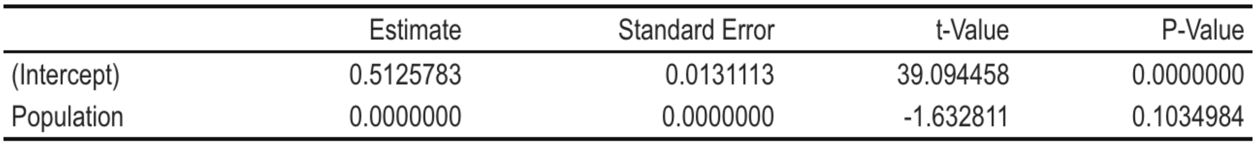
For population vs win percentage we created the following scatter plot:

Figure 10: Figure 10



We chose to make each team a different color to show how each teams city population remained relatively similar in each year from 2010-2019. The line of best fit(in red) has a negative slope which is contrary to what we originally thought the data would show. One thing we would like to point out is the New York teams(in blue on the right side of the above graph) are high leverage points and influence the regression line more than the rest of the data, this is a leading cause for the negative regression line since the New York teams had a below average winning percentage from 2010-2019.

Figure 11: Table 1

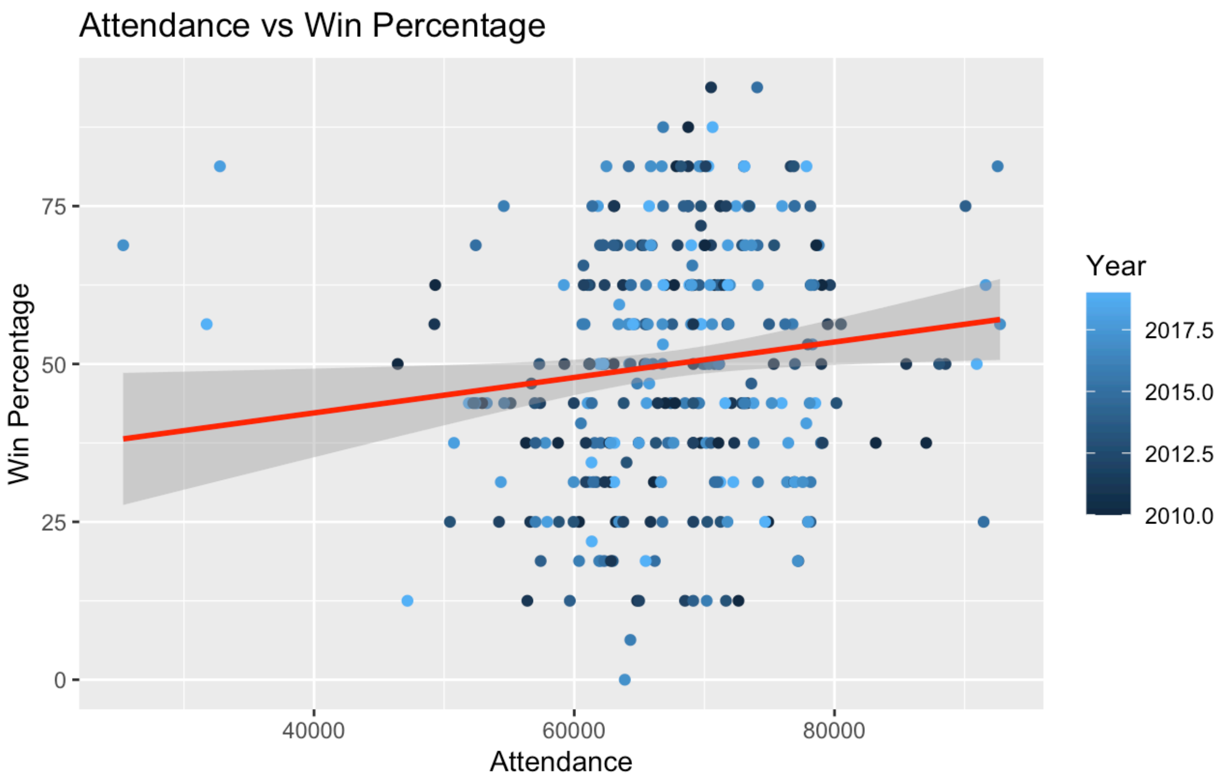


	Estimate	Standard Error	t-Value	P-Value
(Intercept)	0.5125783	0.0131113	39.094458	0.0000000
Population	0.0000000	0.0000000	-1.632811	0.1034984

Above is our regression summary statistics for the population vs win percentage data. With a p-value of 0.1034984 greater than our required alpha level of 0.05, we cannot say that the slope of the regression line differs from 0 and it is not a significant predictor of win percentage. Since we fail to reject our null hypothesis, we cannot however say that the slope of city population vs win percentage is 0, only that we can't say it is different from 0 with certainty.

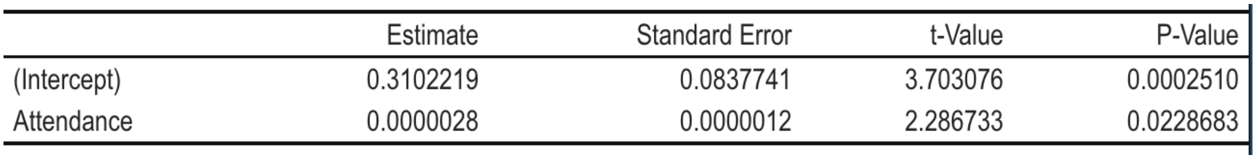
For attendance vs win percentage we created the following scatter plot:

Figure 12: Figure 11



For the above graph we can see a much more visible correlation compared to the population graph. The regression line for stadium attendance vs win percentage(above red line) has a positive slope confirming our original belief that attendance would be correlated with win percentage. The 3 points below 40,000 attendance are the 3 years after the rams moved from St. Louis to Los Angeles and these points decrease the slope of the regression line since they are influential points(extreme in both the x and y direction).

Figure 13: Table 2



	Estimate	Standard Error	t-Value	P-Value
(Intercept)	0.3102219	0.0837741	3.703076	0.0002510
Attendance	0.0000028	0.0000012	2.286733	0.0228683

Above is the regression summary statistics for the attendance vs win percentage data. With a p-value of 0.0228683 less than 0.05 we can say that the slope of the regression line is greater than 0 at the 5% significance level and attendance is a significant predictor of win percentage. The attendance slope of 0.000028 can be interpreted as for every increase of a person in attendance per game, there is a 0.000028% increase in a teams predicted win percentage.

6 Conclusion

In conclusion, stadium attendance is a significant predictor of win percentage with a positive slope, meaning that the higher a teams stadium attendance is, the higher their predicted end of season win percentage is. For our other research question, we found that city population is not a significant predictor of win percentage and we cannot reject the null that the slope of city population vs win percentage is 0.

7 Works Cited

Bureau, U. C. (2022, February 16). City and town population totals: 2010-2019. Census.gov. https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html

ESPN Internet Ventures. (n.d.-a). ESPN. https://www.espn.com/nfl/attendance

ESPN Internet Ventures. (n.d.). NFL standings - 2024 season. ESPN. https://www.espn.com/nfl/standings

Oravec, T. (2024, April 25). How does state/city population affect sports teams success?. ArcGIS StoryMaps. https://storymaps.arcgis.com/stories/71b6dfbc606047f69fe60fab8b591091

Yahoo! (n.d.). Farewell to NFL home-field advantage, as home teams have a losing record through 5 weeks. Yahoo! Sports. https://sports.yahoo.com/farewell-to-nfl-home-field-advantage-as-home-teams-have-a-losing-record-through-5-weeks-153759757.html Analysis/Results

```r
### CODE is done in BOAST style
# Goal: Find, Tidy, Clean, Wrangle Data into usable tables
# PLAN
# 1) Load necessary data packages: {rvest}, {tidyverse}
#
# 2) Find the needed website urls for the data
# a] US Census for city population
# b] NFl Website for Season Records
# c] ESPN for stadium attendance over the last
#
# 3) Create population table
# a] Read Census data into a table
# b] Tidy Data so city is case with attributes of:
#.      population and year
# c] Clean the data to remove all n/a or 0 values
# d] Remove all cities without an NFL team
#
# 4) Create NFL Win Percentage by Year Table
# a] Read the data into a table using read_html function
# b] Tidy the data so that each case is a Team with attributes:
#.      year and win percentage
# c] Clean any data that does not follow table guidelines
#
# 5) Create Average NFL Stadium Attendance by year table
# a] Read the yearly tables into R using read_html
# b] Join the data into one table
# c] Clean and tidy the table, making each case a team
#.      with attributes: year and average stadium attendance
# d] Clean any n/a or 0 values out of the data set

library(tidyverse)
library(rvest)
library(readxl)


# Step 3)
#Create original data set for NFL city populations
#Need list of NFL locations according to the data set
#Use Census Data to create US population data set from 2010-2019
US_Populations <- read_excel("US_City_Populations.xlsx") %>%
  select(c(2,5:14)) %>% #Isolate city and yearly population data
  slice(-(n=1)) %>% #Remove First Row of n/a values
  set_names("City", "2010", "2011","2012","2013","2014","2015",
            "2016","2017","2018","2019")%>% #Rname cols
  #Select only the cities with NFL teams using filter
  filter(City %in% c("Phoenix city, Arizona", "Atlanta city, Georgia",
                     "Baltimore city, Maryland", "Buffalo city, New York",
                     "Charlotte city, North Carolina",
                     "Chicago city, Illinois",
                     "Cincinnati city, Ohio", "Cleveland city, Ohio",
                     "Dallas city, Texas", "Denver city, Colorado",
                     "Detroit city, Michigan", "Green Bay city,
                     Wisconsin",
                     "Houston city, Texas",
                     "Indianapolis city (balance), Indiana",
                     "Jacksonville city, Florida",
                     "Kansas City city, Missouri",
                     "Las Vegas city, Nevada",
                     "Los Angeles city, California",
                     "Miami city, Florida", "Minneapolis city, Minnesota",
                     "Nashville-Davidson metropolitan government (balance),
                     Tennessee",
                     "New Orleans city, Louisiana",
                     "New York city, New York",
                     "Philadelphia city, Pennsylvania",
                     "Pittsburgh city, Pennsylvania",
                     "San Francisco city, California",
                     "Seattle city, Washington", "Tampa city, Florida",
                     "Washington city, District of Columbia",
                     "Boston city, Massachusetts",
                     "St. Louis city, Missouri",
                     "San Diego city, California"))
#Must have St. Louis and San Diego because stadiums moved 2016 and 2017
#respectively for rams and chargers

#Cast all the 2010 values to numbers since for some reason they weren't
US_Populations$"2010" <- as.numeric(as.character(US_Populations$"2010"))

#Reshape and rename the table with case an NFL city in a certain year
#Attribute would be population
NFL_Yearly_Populations <- US_Populations %>%
  pivot_longer(
    #Make every yearly column into one column of years
    cols = starts_with("20"),
    names_to = "Year", #All years in the year column
    values_to = "Population" #All yearly populations in this column
  ) %>%
  arrange(City)
NFL_Yearly_Populations$Year <- as.numeric((NFL_Yearly_Populations$Year))
#Make year a number
NFL_Yearly_Populations <- NFL_Yearly_Populations%>%
  #remove cases of San Diego after the Chargers moved
  filter(!(City == "San Diego city, California" & Year > 2016))%>%
  #remove cases of San Diego after the Chargers moved
  filter(!(City == "St. Louis city, Missouri" & Year > 2015)) %>%
  #remove cases of LA before a team was there
  filter(!(City == "Los Angeles city, California" & Year < 2016))
#Rearrange the data so it's alphabetical


#Step 4)
## the goal of this piece of code is to harvest data from espn and put
# together a table with each case being
## a team in a given year from the timespan of 2010-2019 and the variable
# being win percentage.

## this code reads the data in from the ESPN website

ESPNWINPCT <- read_html(x = "https://www.espn.com/nfl/standings") %>%
  html_elements(css = "table") %>%
  html_table()

## since the data is read in in 4 tibbles, the 1st and
## 3rd being the team names for their respective conferences(AFC,NFC)
## and the 2nd and 4th being the statistics associated with each team, I
## combined all the AFC teams with their data and
## the NFC teams with their data, then combined all this data with the
## bind_rows method.

AFCWINPCT <- bind_cols(ESPNWINPCT[[1]],ESPNWINPCT[[2]])

NFCWINPCT <- bind_cols(ESPNWINPCT[[3]],ESPNWINPCT[[4]])

## after we have all the necessary data we remove all data other than team
## name and win percentage with the select method
## and we get rid of filler rows by slicing only the rows with teams
NFLWINPCT <- bind_rows(NFCWINPCT,AFCWINPCT) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,
        28,29,30,32,33,34,35,37,38,39,40) %>%
  ## I separated by spaces to get all the team names in one column then
  ## I got rid of the rest of the columns so we
  ## are left with just team name and win pct
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  ## to add the year I mutated a column onto the end with 2024 for every
  ## team
  mutate(
    Year = c(2024,2024,2024,2024,2024,2024,2024,2024,2024,2024,2024,2024
             ,2024,2024,2024,2024,2024,2024,2024,2024,2024,2024,2024,
             2024,2024,2024,2024,2024,2024,2024,2024,2024)
  )

## Now that we can do this for one year, we just need to do this for the
## years 2010-2019 then use bind_rows for
## all 10 data frames to get our finished product



#2010
ESPNWINPCT2010 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2010") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2010 <- bind_cols(ESPNWINPCT2010[[1]],ESPNWINPCT2010[[2]])

NFCWINPCT2010 <- bind_cols(ESPNWINPCT2010[[3]],ESPNWINPCT2010[[4]])

NFLWINPCT2010 <- bind_rows(NFCWINPCT2010,AFCWINPCT2010) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,
        25,27,28,29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2010,2010,2010,2010,2010,2010,2010,2010,2010,2010,2010,
             2010,2010,2010,2010,2010,2010,2010,2010,2010,2010,2010,
             2010,2010,2010,2010,2010,2010,2010,2010,2010,2010)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )
```

```r
#2011
ESPNWINPCT2011 <- read_html(x =
            "https://www.espn.com/nfl/standings/_/season/2011") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2011 <- bind_cols(ESPNWINPCT2011[[1]],ESPNWINPCT2011[[2]])

NFCWINPCT2011 <- bind_cols(ESPNWINPCT2011[[3]],ESPNWINPCT2011[[4]])

NFLWINPCT2011 <- bind_rows(NFCWINPCT2011,AFCWINPCT2011) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,
        25,27,28,29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2011,2011,2011,2011,2011,2011,2011,2011,2011,2011,
             2011,2011,2011,2011,2011,2011,2011,2011,2011,2011,2011,
             2011,2011,2011,2011,2011,2011,2011,2011,2011,2011,2011)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2012
ESPNWINPCT2012 <- read_html(x =
            "https://www.espn.com/nfl/standings/_/season/2012") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2012 <- bind_cols(ESPNWINPCT2012[[1]],ESPNWINPCT2012[[2]])

NFCWINPCT2012 <- bind_cols(ESPNWINPCT2012[[3]],ESPNWINPCT2012[[4]])

NFLWINPCT2012 <- bind_rows(NFCWINPCT2012,AFCWINPCT2012) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,
        27,28,29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2012,2012,2012,2012,2012,2012,2012,2012,2012,2012,2012,
             2012,2012,2012,2012,2012,2012,2012,2012,2012,2012,2012,
             2012,2012,2012,2012,2012,2012,2012,2012,2012,2012)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2013
ESPNWINPCT2013 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2013") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2013 <- bind_cols(ESPNWINPCT2013[[1]],ESPNWINPCT2013[[2]])

NFCWINPCT2013 <- bind_cols(ESPNWINPCT2013[[3]],ESPNWINPCT2013[[4]])

NFLWINPCT2013 <- bind_rows(NFCWINPCT2013,AFCWINPCT2013) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,28,
        29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2013,2013,2013,2013,2013,2013,2013,2013,2013,2013,2013,
             2013,2013,2013,2013,2013,2013,2013,2013,2013,2013,2013,
             2013,2013,2013,2013,2013,2013,2013,2013,2013,2013)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2014
ESPNWINPCT2014 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2014") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2014 <- bind_cols(ESPNWINPCT2014[[1]],ESPNWINPCT2014[[2]])

NFCWINPCT2014 <- bind_cols(ESPNWINPCT2014[[3]],ESPNWINPCT2014[[4]])

NFLWINPCT2014 <- bind_rows(NFCWINPCT2014,AFCWINPCT2014) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,28,
        29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,
             2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,2014,
             2014,2014,2014,2014,2014,2014,2014,2014)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2015
ESPNWINPCT2015 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2015") %>%
  html_elements(css = "table") %>%
  html_table()
```

```r
AFCWINPCT2015 <- bind_cols(ESPNWINPCT2015[[1]],ESPNWINPCT2015[[2]])

NFCWINPCT2015 <- bind_cols(ESPNWINPCT2015[[3]],ESPNWINPCT2015[[4]])

NFLWINPCT2015 <- bind_rows(NFCWINPCT2015,AFCWINPCT2015) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,28,29,
        30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2015,2015,2015,2015,2015,2015,2015,2015,2015,2015,2015,
             2015,2015,2015,2015,2015,2015,2015,2015,2015,2015,2015,
             2015,2015,2015,2015,2015,2015,2015,2015,2015,2015)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2016
ESPNWINPCT2016 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2016") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2016 <- bind_cols(ESPNWINPCT2016[[1]],ESPNWINPCT2016[[2]])

NFCWINPCT2016 <- bind_cols(ESPNWINPCT2016[[3]],ESPNWINPCT2016[[4]])

NFLWINPCT2016 <- bind_rows(NFCWINPCT2016,AFCWINPCT2016) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,28,29,
        30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2016,2016,2016,2016,2016,2016,2016,2016,2016,2016,2016,
             2016,2016,2016,2016,2016,2016,2016,2016,2016,2016,2016,
             2016,2016,2016,2016,2016,2016,2016,2016,2016,2016)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2017
ESPNWINPCT2017 <- read_html(x =
              "https://www.espn.com/nfl/standings/_/season/2017") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2017 <- bind_cols(ESPNWINPCT2017[[1]],ESPNWINPCT2017[[2]])

NFCWINPCT2017 <- bind_cols(ESPNWINPCT2017[[3]],ESPNWINPCT2017[[4]])

NFLWINPCT2017 <- bind_rows(NFCWINPCT2017,AFCWINPCT2017) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,
        28,29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2017,2017,2017,2017,2017,2017,2017,2017,2017,2017,2017,
             2017,2017,2017,2017,2017,2017,2017,2017,2017,2017,2017,
             2017,2017,2017,2017,2017,2017,2017,2017,2017,2017)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2018
ESPNWINPCT2018 <- read_html(x =
            "https://www.espn.com/nfl/standings/_/season/2018") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2018 <- bind_cols(ESPNWINPCT2018[[1]],ESPNWINPCT2018[[2]])

NFCWINPCT2018 <- bind_cols(ESPNWINPCT2018[[3]],ESPNWINPCT2018[[4]])

NFLWINPCT2018 <- bind_rows(NFCWINPCT2018,AFCWINPCT2018) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,28,
        29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2018,2018,2018,2018,2018,2018,2018,2018,2018,2018,2018,2018
             ,2018,2018,2018,2018,2018,2018,2018,2018,2018,2018,
             2018,2018,2018,2018,2018,2018,2018,2018,2018,2018)
  )%>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "Redskins" ~ "Commanders",
      .default = Team
    )
  )

#2019
ESPNWINPCT2019 <- read_html(x =
            "https://www.espn.com/nfl/standings/_/season/2019") %>%
  html_elements(css = "table") %>%
  html_table()

AFCWINPCT2019 <- bind_cols(ESPNWINPCT2019[[1]],ESPNWINPCT2019[[2]])

NFCWINPCT2019 <- bind_cols(ESPNWINPCT2019[[3]],ESPNWINPCT2019[[4]])

NFLWINPCT2019 <- bind_rows(NFCWINPCT2019,AFCWINPCT2019) %>%
  select(1,5)%>%
  set_names("TeamName", "WinPercentage") %>%
  slice(2,3,4,5,7,8,9,10,12,13,14,15,17,18,19,20,22,23,24,25,27,
        28,29,30,32,33,34,35,37,38,39,40) %>%
  separate_wider_delim(
    cols = "TeamName",
    delim = " ",
    names = c("a","b","c","Team"),
    too_few = "align_end"
  ) %>%
  select(4,5) %>%
  mutate(
    Year = c(2019,2019,2019,2019,2019,2019,2019,2019,2019,2019,2019,
             2019,2019,2019,2019,2019,2019,2019,2019,2019,2019,2019,
             2019,2019,2019,2019,2019,2019,2019,2019,2019,2019)
  ) %>% mutate( #Change Washington team name for consistency and CARE
    Team = case_match(
      .x = Team,
      "WSHWashington" ~ "Commanders",
      .default = Team
    )
  )

## adds all the different years data together to make a final dataset
NFLWINPCTFINAL <- bind_rows(NFLWINPCT2010,NFLWINPCT2011,
                            NFLWINPCT2012,NFLWINPCT2013,
                            NFLWINPCT2014,NFLWINPCT2015,
                            NFLWINPCT2016,NFLWINPCT2017,
                            NFLWINPCT2018,NFLWINPCT2019)


#Step 5)
#Read in URLs: https://www.espn.com/nfl/attendance/_/year/2010
#              https://www.espn.com/nfl/attendance/_/year/2011
#              https://www.espn.com/nfl/attendance/_/year/2012
#              https://www.espn.com/nfl/attendance/_/year/2013
#              https://www.espn.com/nfl/attendance/_/year/2014
#              https://www.espn.com/nfl/attendance/_/year/2015
#              https://www.espn.com/nfl/attendance/_/year/2016
#              https://www.espn.com/nfl/attendance/_/year/2017
#              https://www.espn.com/nfl/attendance/_/year/2018
#              https://www.espn.com/nfl/attendance/_/year/2019
#Creating a data frame for the stadium attendance data for each year
#Gather 2010 Stadium data from espn
stadiumRaw_2010 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2010") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for 2010
stadium_2010 <- stadiumRaw_2010[[1]] %>%
  mutate(
    Year = c(2010) #Add year column for 2010
  )

#Gather 2011 Stadium data from espn
stadiumRaw_2011 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2011") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame 2011
stadium_2011 <- stadiumRaw_2011[[1]] %>%
  mutate(
    Year = c(2011) #Add year column for 2011
  )
#Gather 2012 Stadium data from espn
stadiumRaw_2012 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2012") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for 2012
stadium_2012 <- stadiumRaw_2012[[1]]%>%
  mutate(
    Year = c(2012) #Add year column for 2012
  )
#Gather 2013 Stadium data from espn
stadiumRaw_2013 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2013") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for 2013
stadium_2013 <- stadiumRaw_2013[[1]] %>%
  mutate(
    Year = c(2013) #Add year column for 2013
  )
#Gather 2014 Stadium data from espn
stadiumRaw_2014 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2014") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame 2014
stadium_2014 <- stadiumRaw_2014[[1]]%>%
  mutate(
    Year = c(2014) #Add year column for 2014
  )
```

```r
#Gather 2015 Stadium data from espn
stadiumRaw_2015 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2015") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for 2015
stadium_2015 <- stadiumRaw_2015[[1]]%>%
  mutate(
    Year = c(2015) #Add year column for 2015
  )
#Gather 2016 Stadium data from espn
stadiumRaw_2016 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2016") %>%
  html_elements(css = "table") %>%
  html_table()#List of data frames read in
#Make one data frame for 2016
stadium_2016 <- stadiumRaw_2016[[1]]%>%
  mutate(
    Year = c(2016) #Add year column for 2016
  )
#Gather 2017 Stadium data from espn
stadiumRaw_2017 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2017") %>%
  html_elements(css = "table") %>%
  html_table()#List of data frames read in
#Make one data frame for 2017
stadium_2017 <- stadiumRaw_2017[[1]]%>%
  mutate(
    Year = c(2017) #Add year column for 2017
  )
#Gather 2018 Stadium data from espn
stadiumRaw_2018 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2018") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for 2018
stadium_2018 <- stadiumRaw_2018[[1]] %>%
  mutate(
    Year = c(2018) #Add year column for 2018
  )
#Gather 2019 Stadium data from espn
stadiumRaw_2019 <- read_html(
  "https://www.espn.com/nfl/attendance/_/year/2019") %>%
  html_elements(css = "table") %>%
  html_table() #List of data frames read in
#Make one data frame for stadium 2019
stadium_2019 <- stadiumRaw_2019[[1]] %>%
  mutate(
    Year = c(2019) #Add year column for 2019
  )

#Sequence of unneeded rows
unneeded_rows_1 <- seq(from = 1, to = 340, by = 34)
unneeded_rows_2 <- seq(from = 1, to = 330, by = 33) #Rest of bad rows

#Join the yearly data into one data frame then tidy
yearly_Stadium_Data <- bind_rows(
  stadium_2010,stadium_2011,stadium_2012,stadium_2013,
  stadium_2014,stadium_2015,stadium_2016,stadium_2017,
  stadium_2018,stadium_2019) %>%
  select(c(2,5,12)) %>%
  set_names("Team","Attendance","Year")%>%
  slice(-c(unneeded_rows_1)) %>% #Removes rows that aren't teams
  slice(-c(unneeded_rows_2)) #Removes rows that aren't teams
#LA rows have same name, using source data, identify chargers rows
chargers_rows <- c(21,(1*32+19),(2*32+28),(3*32+22),(4*32+21),
                   (5*32+19),(6*32+7),(7*32+26),(8*32+10),(9*32+11))
#Rename the chargers rows from LA to Chargers
yearly_Stadium_Data[chargers_rows, 1] <- "Chargers"


# Goal: Combine the three tidied datasets into one
# Case: team and year
# Attributes: win percentage, city, population
#Plan:
# Step 1) collect all 3 data frames into this file
#Must have the run the collection code in the same Environment
# Step 2) Wrangle tables to make case matching easy
# a] Duplicate certain city results for the cities with multiple teams
# b] Add/edit Team column so that it corresponds correctly with each city:
#.       NY - Giants/Jets, LA - Rams/Chargers
# c] Rename teams and cities of all tables so all have team name
# Step 3) Create Attendance/Win PCT table
# a] join_by year and team
# Step 4) make the final table by adding the population in
# a] join_by year and team
# Step 4) Clean and Tidy any unforeseen loose ends
# a] Make sure all datatypes match and are usable
# b] Check for correct number of rows after joining

#Arrange Win% data by alphabetically by team
Win_PCT <- NFLWINPCTFINAL %>%
  arrange(Team)
#Make the percetage a numerical variable to make data usable
```

```r
Win_PCT$WinPercentage <- as.numeric((Win_PCT$WinPercentage))

#used for duplicating city data with two teams associated to it
NY_duplicates <- c(1,1,1,1,1,1,1,1,1,1)
LA_duplicates <- c(1,1,1)
#Copy alphabetized city data
City_Populations <- NFL_Yearly_Populations
#Duplicate cities with 2 teams (LA & NY)
#Id the rows you want to dup, and how many dups
LA_duplications <- rep(172:174, LA_duplicates) #Los angeles rows x2
#Dup rows in new set
LA_Population_dup <- City_Populations[LA_duplications,]%>%
  mutate(#Rename this duplicate data to make it unique to same city data
    City = case_match(
      .x = City,
      "Los Angeles city, California" ~ "Chargers",
      .default = "Chargers"
    )
  )
NY_duplications <- rep(221:230, NY_duplicates) #New york rows x2
NY_Population_dup <- City_Populations[NY_duplications,] %>%
  mutate(#Rename this duplicate data to make it unique to same city data
    City = case_match(
      .x = City,
      "New York city, New York" ~ "Jets",
      .default = "Jets"
    )
  )
#Add the duplicate city data to the City Populations and rename cities
Team_Populations <- bind_rows(#Combine data
  City_Populations, LA_Population_dup, NY_Population_dup
) %>% mutate( #Set all the names to the team name, not city
  City = case_match(
    .x = City,
    "Phoenix city, Arizona" ~ "Cardinals",
    "Atlanta city, Georgia" ~ "Falcons",
    "Baltimore city, Maryland" ~ "Ravens",
    "Buffalo city, New York" ~ "Bills",
    "Charlotte city, North Carolina" ~ "Panthers",
    "Chargers" ~ "Chargers",
    "Chicago city, Illinois" ~ "Bears",
    "Cincinnati city, Ohio" ~ "Bengals",
    "Cleveland city, Ohio" ~ "Browns",
    "Dallas city, Texas" ~ "Cowboys",
    "Denver city, Colorado" ~ "Broncos",
    "Detroit city, Michigan" ~ "Lions",
    "Green Bay city, Wisconsin" ~ "Packers",
    "Houston city, Texas" ~ "Texans",
    "Indianapolis city (balance), Indiana" ~ "Colts",
    "Jacksonville city, Florida" ~ "Jaguars",
    "Kansas City city, Missouri" ~ "Chiefs",
    "Las Vegas city, Nevada" ~ "Raiders",
    "Los Angeles city, California" ~ "Rams",
    "Miami city, Florida" ~ "Dolphins",
    "Minneapolis city, Minnesota" ~ "Vikings",
    "Boston city, Massachusetts" ~ "Patriots",
    "New Orleans city, Louisiana" ~ "Saints",
    "New York city, New York" ~ "Giants",
    "Jets" ~ "Jets",
    "Philadelphia city, Pennsylvania" ~ "Eagles",
    "Pittsburgh city, Pennsylvania" ~ "Steelers",
    "San Francisco city, California" ~ "49ers",
    "Seattle city, Washington" ~ "Seahawks",
    "Tampa city, Florida" ~ "Buccaneers",
    "Nashville-Davidson metropolitan government (balance),
    Tennessee" ~ "Titans",
    "Washington city, District of Columbia" ~ "Commanders",
    "St. Louis city, Missouri" ~ "Rams",
    "San Diego city, California" ~ "Chargers",
    .default = "missing"
  )
)%>% set_names(
  "Team", #Make it to team to make things easier when joining
  "Year",
  "Population"
)%>% arrange(Team)
#Change year type to make it compatable for joining, set it to number


#Create the stadium data to have the names of each team the way the Win%
#does
Stadium_Data <- yearly_Stadium_Data%>%
  pivot_wider(
    id_cols = Team,
    names_from = Year,
    values_from = Attendance
  ) %>% arrange (#Alphabetize the data to make renaming easier
    Team
  )%>% mutate( #Set all the names to the team name, not city
    Team = case_match(
      .x = Team,
      "Arizona" ~ "Cardinals",
      "Atlanta" ~ "Falcons",
      "Baltimore" ~ "Ravens",
      "Buffalo" ~ "Bills",
      "Carolina" ~ "Panthers",
      "Chargers" ~ "Chargers",
      "Chicago" ~ "Bears",
      "Cincinnati" ~ "Bengals",
      "Cleveland" ~ "Browns",
      "Dallas" ~ "Cowboys",
      "Denver" ~ "Broncos",
      "Detroit" ~ "Lions",
      "Green Bay" ~ "Packers",
      "Houston" ~ "Texans",
      "Indianapolis" ~ "Colts",
      "Jacksonville" ~ "Jaguars",
      "Kansas City" ~ "Chiefs",
      "Las Vegas" ~ "Raiders",
      "Los Angeles" ~ "Rams",
      "Miami" ~ "Dolphins",
      "Minnesota" ~ "Vikings",
      "New England" ~ "Patriots",
      "New Orleans" ~ "Saints",
      "NY Giants" ~ "Giants",
      "NY Jets" ~ "Jets",
      "Philadelphia" ~ "Eagles",
      "Pittsburgh" ~ "Steelers",
      "San Francisco" ~ "49ers",
      "Seattle" ~ "Seahawks",
      "Tampa Bay" ~ "Buccaneers",
      "Tennessee" ~ "Titans",
      "Washington" ~ "Commanders",
      .default = "missing"
    )
  )%>% pivot_longer(#make case team and year and attribute attendance
    cols = starts_with("20"), #Make every yearly column into one column of
    #years
    names_to = "Year", #All years in the year column
    values_to = "Attendance" #All yearly attendance avgs in this column
  ) %>% arrange(Team) #Alphabetize them to make it easier to combine data
#sets
#Set the years to numbers not strings, since other datasets have num
#datatype
Stadium_Data$Year <- as.numeric(as.character(Stadium_Data$Year))
#Make the attendance numbers to make data usable, must remove commas
Stadium_Data$Attendance <- as.numeric(gsub(",","",Stadium_Data$Attendance))

#Combine the Stadium and win% now that they have the same format
Attendance_Win_PCT_Table <- full_join(
  x = Win_PCT,
  y = Stadium_Data,
  by = join_by(Year == Year, Team == Team) #Match cases of year and team
)%>% relocate(WinPercentage, .after = Year) #Rearrange cols:
#team|year|Win|Attendance

#Add in the Population data to the table for the final table
Final_Table <- full_join(
  x = Attendance_Win_PCT_Table,
  y = Team_Populations,
  by = join_by(Year == Year, Team == Team) #Match cases of year and team
) %>% arrange(Year)%>%
  arrange(Team)

head(Final_Table)


### Goal - Create simple displays to understand the spread of variables
###        and the potential relationships between them
# Plan -
# 1) Install necessary packages (ggplot,kable)
# 2) Create histograms of each variable to see general spread
## a) use ggplot package and Final_Table previously made
## b) Win Percentage, Population, and Attendace should be graphed
# 3) Create basic plots to show potential correlation between the data
## a) Win Pct vs. Population (Scatter)
## b) Attendance vs. Win Pct (Scatter)
## c) Population vs. Attendance (Scatter) Will not statistically prove,
#but maybe useful
## d) Win Pct vs. Team (Boxplot)
## e) Population vs. Team (Boxplot))
## f) Attendance vs. Team (Boxplot)

# Step 1) Load Packages
library(ggplot2)
library(kableExtra)
library(tidyverse)

# Step 2) Create Histograms
# Create histogram for WinPercentage
ggplot(
  data = Final_Table,
  mapping = aes(
    x = WinPercentage)) + # x-axis is win percentage
  geom_histogram(binwidth = 0.05, #new bar every 5% difference in win pct
                 fill = "blue",
                 color = "black",
                 alpha = 0.7)+
  labs(title = "NFL Win Percentage",
       x = "Win Percentage",
       y = "Frequency") +
  theme_bw()


# Create histogram for Attendance with x as attendance and y as frequency
ggplot(
  data = Final_Table,
  mapping = aes(
    x = Attendance)) + #Graphing average season attendance
  geom_histogram(binwidth = 2500, #Bars measure segments of 2500
                 fill = "green",
                 color = "black",
                 alpha = 0.7)+
  labs(title = "NFL Stadium Attendance",
       x = "Attendance",
       y = "Frequency") +
  theme_bw()


# Create histogram for Population with x as population and y as frequency
ggplot(
  data = Final_Table,
  mapping = aes(
    x = Population)) + #Population is the x axis
  geom_histogram(binwidth = 250000, #For every 250 thousand split bars
                 fill = "red",
                 color = "black",
                 alpha = 0.7)+
  labs(title = "NFL City Populations",
       x = "City Population",
       y = "Frequency") +
  theme_bw()

# Step 3) Discovering potential correlation
# Population vs. Attendance Scatterplot to show likely relationship there
ggplot(
  data = Final_Table,
  mapping = aes(x = Population / 1000000,# sets x values to Population
                y = Attendance/1000, #y values to Attendance
                color = Team)) + # sets color to match team
  geom_point(shape = 19) + # sets shape of the points to a closed circle
  labs(
    x = "Population (Millions)", # sets labels for the graph
    y = "Attendance (Thousands)",
    title = "Population vs Attendance"
  ) +
  theme_bw()

# Population vs. Win Pct Scatterplot to show likely relationship there
ggplot(
  data = Final_Table,
  mapping = aes(x = Population / 1000000,# sets x values to Population
                y = WinPercentage*100, #y values to winPercentage
                color = Team)) + # sets color to match team
  geom_point(shape = 19) + # sets shape of the points to a closed circle
  labs(
    x = "Population (Millions)", # sets labels for the graph
    y = "Win Percentage",
    title = "Population vs Win Percentage"
  ) +
  theme_bw()

# Win Percentage vs. Attendance Scatterplot to show likely relationship there
ggplot(
  data = Final_Table,
  mapping = aes(x = WinPercentage*100,# sets x values to attendance
                y = Attendance/1000, #y values to winPercentage
                color = Team)) + # sets color to match team
  geom_point(shape = 19) + # sets shape of the points to a closed circle
  labs(
    x = "Win Percentage", # sets labels for the graph
    y = "Attendance (Thousands)",
    title = "Win Percentage vs Attendance"
  ) +
  theme_bw()

# Graph the teams and win percentages with boxplot since it may be
# confounding
ggplot(
  data = Final_Table,
  aes(x=as.factor(Team), # x-axis is based on teams
      y=WinPercentage*100, # y-axis is win percentage
      fill = as.factor(Team))) + # Fill based on which team
  geom_boxplot(
    alpha=0.7) +
  labs(title = "NFL Team vs. Win Percentage",
       x = "Team",
       y = "Win Percentage") +
  theme_bw() +
  theme(
    axis.text.x = element_blank() # remove names from x-axis for
    # readability
  )+
  scale_fill_manual(
    # Assign teams colors
    values = rainbow(length(unique(Final_Table$Team)))
  )

# Graph the teams and Attendance with boxplot since it may be confounding
ggplot(
  data = Final_Table,
  aes(x=as.factor(Team), # x-axis is based on teams
      y= Attendance, # y-axis is Attendnace
      fill = as.factor(Team))) + # Fill based on which team
  geom_boxplot(
    alpha=0.7) +
  labs(title = "NFL Team vs. Attendance",
       x = "Team",
       y = "Attendance") +
  theme_bw() +
  theme(
    # remove names from x-axis for readability
    xis.text.x = element_blank()
  )+
  scale_fill_manual(
    # Assign teams colors
    values = rainbow(length(unique(Final_Table$Team)))
  )

# Graph the teams and Population with boxplot since it may be confounding
ggplot(
  data = Final_Table,
  aes(x=as.factor(Team), # x-axis is based on teams
      y= Population, # y-axis is Population
      fill = as.factor(Team))) + # Fill based on which team
  geom_boxplot(
    alpha=0.7) +
  labs(title = "NFL Team vs. Population",
       x = "Team",
       y = "Population") +
  theme_bw() +
  theme(
    # remove names from x-axis for readability
    axis.text.x = element_blank()
  )+
  scale_fill_manual(
    values = rainbow(length(unique(Final_Table$Team))) # Assign teams colors
  )
```

```r
## plan
## create 2 regression models, city population vs win percentage and avg
## stadium attendance vs win percentage
## next test to see if the regression coefficients are significant

library(ggplot2)

ggplot(data = Final_Table,
       mapping = aes(# sets x values to attendance,
         x = Population / 1000000,
         #  y values to winPercentage
                     y = WinPercentage * 100,
                     color = Team)) + # sets color to match team
  geom_point(shape = 19) + # sets shape of the points to a closed circle
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  # creates a linear line of best fit for the scatterplot
  labs(
    x = "Population(Millions)", # sets labels for the graph
    y = "Win Percentage",
    title = "Population vs Win Percentage"
  )

# sets regression model for population vs winpercentage'
PopulationRegModel <- lm(WinPercentage~Population, data = Final_Table)
# creates summary statistics for the above regression model
PopSummary <- summary(PopulationRegModel)

ggplot(data = Final_Table,
       # sets x values to attendance, y values to winPercentage
       mapping = aes(x = Attendance,
                     y = WinPercentage * 100,
                     color = Year)) + # codes each point color by year
  geom_point(shape = 19) + # sets shape of the points to a closed circle
  # create a linear line of best fit for the scatterplot
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    x = "Attendance",   # sets labels for the graph
    y = "Win Percentage",
    title = "Attendance vs Win Percentage"
  )

StadiumAttendanceRegModel <- lm(WinPercentage~Attendance, data = Final_Table)
# sets regression model for population vs winpercentage
# create summary statistics for the above regression model
AttendanceSummary <- summary(StadiumAttendanceRegModel)

# next I want to create a table showing summary statistics for the NFL
# final data

library(kableExtra)
library(dplyr)
library(htmltools)

Summary_Table <- PopSummary$coefficients %>%
  kable() %>%
  kable_classic()

#Create summary tables for population on Win Pct
Summary_Table_Population <- PopSummary$coefficients %>%
  kable(
    col.names = c("Estimate", "Standard Error", "t-Value", "P-Value")
  ) %>%
  kable_classic()
Summary_Table_Population

#Create summary tables of Win PCT on Attendance
Summary_Table_Attendance <- AttendanceSummary$coefficients %>%
  kable(
    col.names = c("Estimate", "Standard Error", "t-Value", "P-Value")
  ) %>%
  kable_classic()
Summary_Table_Attendance
```