STAT184 Final: NFL Team Rankings

Jay Cush, Nixon Kameen, Nathan O

The Qualities of a Super Bowl-Winning Team

Nixon Kameen, Jay Cush, Nathan O

Since 1967, with the merger of the National Football League (NFL), and the American Football League (AFL), the Super Bowl has become a major event in the United States, and for the sports industry as a whole. It marks finale of the NFL season between the best teams from each division: the American Football Conference (AFC), and the National Football Conference (NFC). Since the 60's, it has been every NFL team's top priority to win the championship game, and hold up the Lombardi Trophy, but how are they able to achieve this incredibly difficult task?

In this report, we are investigating two questions: a.) What trends are apparent between the offenses of the past 8 Super Bowl winners? b.) If there are consistent trends from our data, can we accurately predict the future and find the next Super Bowl winner for the 2024-25 NFL season? We will first state some relevant background information, followed by our strategy to tackle these questions. After that, we will state our findings and end with our closing remarks and statements.

Background Information

Over the years, predicting the Super Bowl has been a mix of statistics and fun. There have been many different cases throughout the years on how people predicted the Super Bowl. Some prediction methods that aren't considered mainstream have gained a lot of attention and even a bit of fame for being entertaining and (sometimes) outrageous. Animals like Paul the Octopus and Fiona the Hippo have famously picked Super Bowl winners by choosing between food options or team-branded items. Some fans have also enlisted astrology in their divination, consulting horoscopes or using celestial alignments to help divine a winner.

On the more conventional side, sports analysts depend on traditional data and expert wisdom to forecast who will win the Super Bowl. Their predictions are grounded in types of analysis

that many of us might recognize from school or the office. They look at all kinds of factors—some we might expect and others we might not, such as injuries. They also use all kinds of scientific strategies to come up with numbers that mean something (or not) in predicting the game's outcome. In fact, some would say that the Super Bowl may well be the most rigorously analyzed sporting event of the year, not just because of the large amount of money bet on it but also because of its considerable cultural cachet.

Methodology

We had two different tasks to accomplish: 1.) Finding the trends between the Super Bowl-winning teams' offense and 2.) Comparing those trends with the teams in the current NFL season to predict who will win it. For this experiment, Nixon was able to gather the offensive statistics of every team from each season from 2015-2023, respectfully, from "pro-football-reference.com". It's good to keep in mind that all of these data sets have the same named variables, but if we want to see which teams performed better in each individual statistic, we also need to find the correlation between winning the Super Bowl and each variable. Basically, we need to know if have more or less of the variable is a good statistic for winning.

After finding the correlation of each variable and winning the Super Bowl, we made a new data set that converted all the numeric values into rankings between each team for every variable in the season, and then combined all the seasons together to find different trends in the teams.

Next, we went through different offensive categories and found the different distributions between the winning and losing teams in said variables, creating different graphs to signify throughout the years how each winning team performed in that category compared to their competition. This helped us find the typical trends each Super Bowl team in the past 8 years tended to follow, and could help us with our predictions.

Finally, to make our predictions for the current NFL season, we grabbed the data of every team this season still eligible for making the playoffs and compared their different standings to the ones that were apparent with the previous Super Bowl winners. From our estimations thereafter, we can predict who will win the Super Bowl this season.

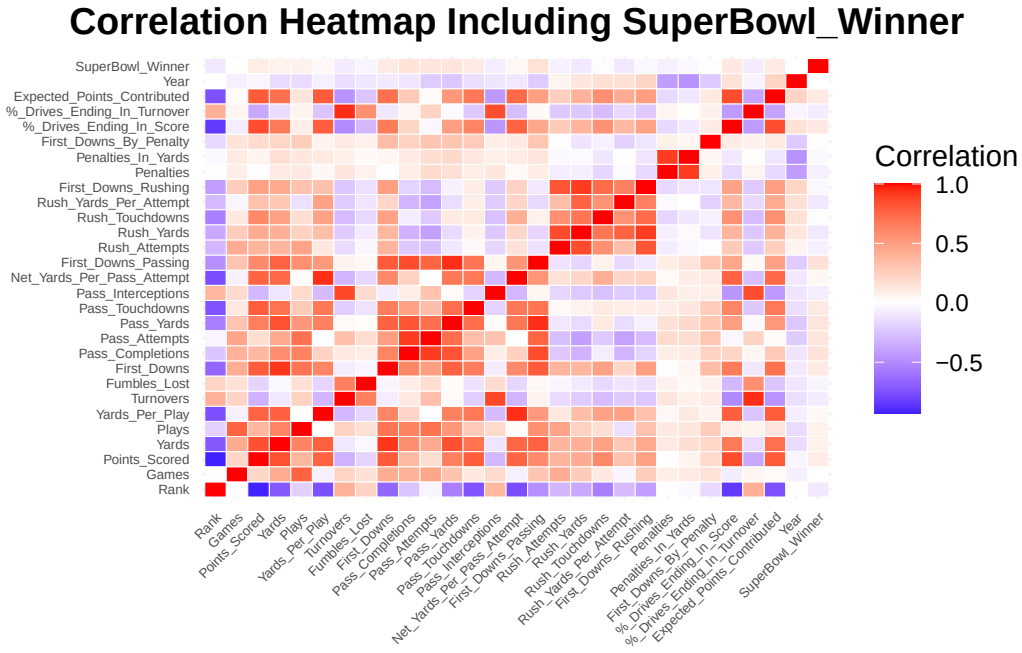
It is also good to keep in mind that the data is supported by the FAIR principles. All of the data is easily findable and accessible through the repository, it was interoperable with the combining of multiple data sets, and can be reused for future projects, if needed.

Data Exploration:

First, we wanted to see a broad overview of all the numeric values in the set, so we made a graph summarizing all of the data from the 2015-2023 seasons. This included the mean values and quartile split for different variables including Points_Scored, Yards, and Turnovers, which helped us understand the range, mean, and variability of key metrics.

As seen by the first graph, one of the most important factors to consider from now onward is the amount of games played. From transitioning from the 2020 to the 2021 season, the amount of weeks in the traditional season increased, meaning that team statistics (such as passing and rushing yards, turnovers, etc.) could increase, allowing for more variability to happen throughout the graphs. We also do have to consider that during the 2020-21 season, multiple teams had to shorten their seasons due to the spreading of COVID-19, meaning that some of the dips in the data below could have also resulted from that.

As mentioned above, we wanted to find the different correlations between each offensive variable and winning the Super Bowl. This helps readers get insight on the world of the NFL, by seeing how different variables affect teams and what they need to do to win the big game. In the graph below, and as seen by the key, having a red value represented a positive correlation between the two selected variables, and blue representing a negative one. The shading on the correlations also called for the strengths of each variable on each other, but more importantly, to winning the Super Bowl.



Since there are multiple variables that are needed to be considered in determining what a team needs to win a Super Bowl, it makes sense that there aren't any noticeably strong correlations between the variables and winning the big game. But we wanted to go more in depth on the concept of the strength of the variables on winning the big game, so we also gathered the absolute values of the correlation coefficient values and found which values were the most impactful. Below shows our findings:

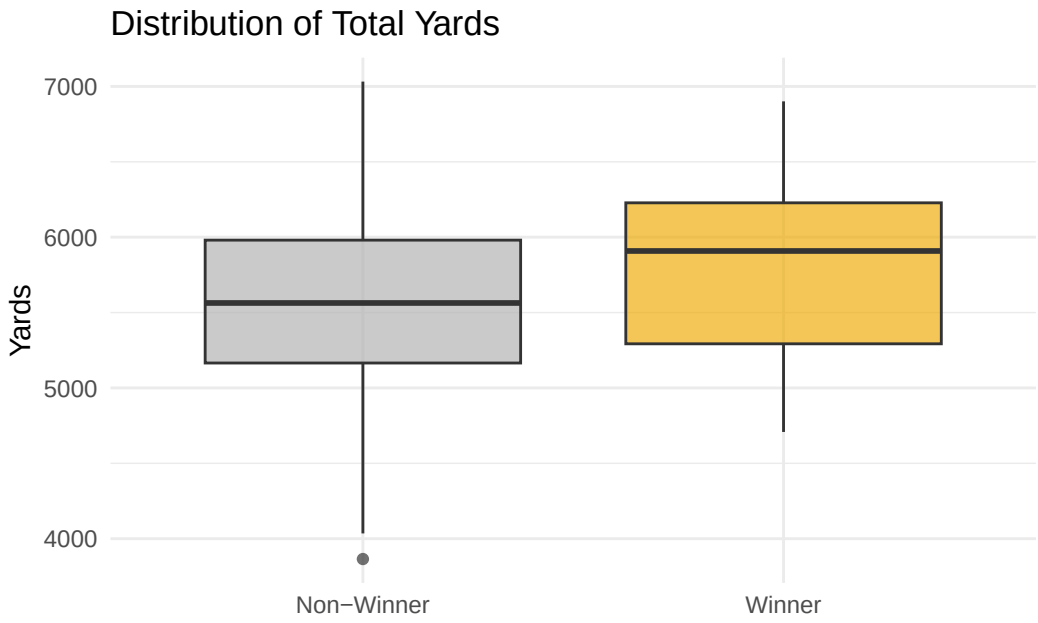
```
SuperBowl_Winner          First_Downs_Passing
```

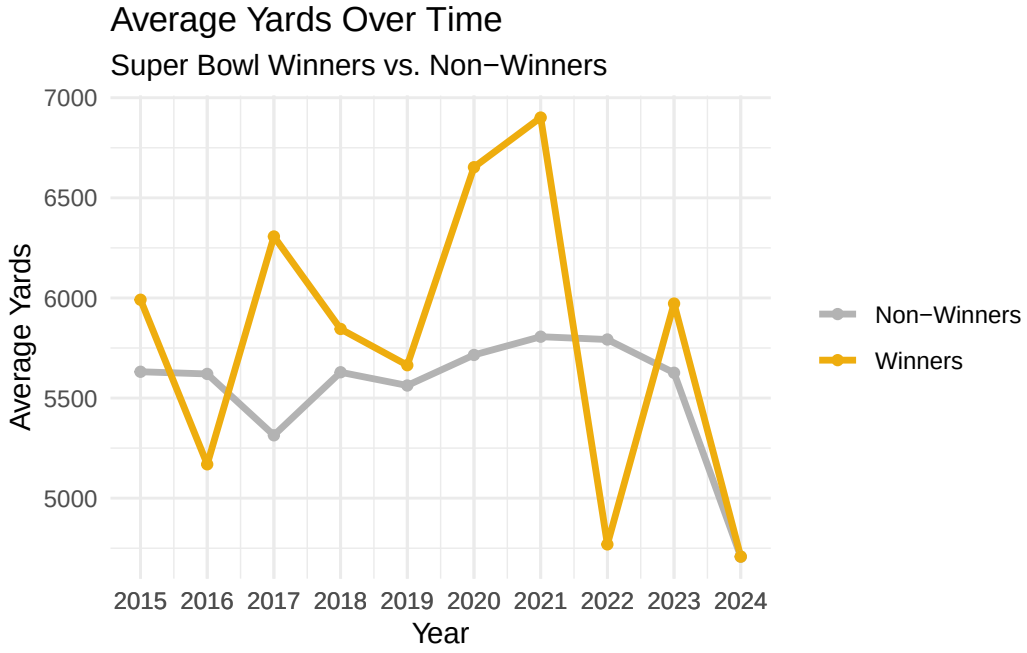
```
                1.000000000                 0.154382153
           Pass_Completions               Pass_Attempts
                0.151908329                 0.135002896
                 Pass_Yards    %_Drives_Ending_In_Score
                0.134854749                 0.112877869
           Pass_Touchdowns Expected_Points_Contributed
                0.108833726                 0.106365475
                First_Downs               Points_Scored
                0.106094272                 0.095343421
                      Yards                       Plays
                0.071614007                 0.068525540
 Net_Yards_Per_Pass_Attempt              Yards_Per_Play
                0.042754490                 0.034298546
     First_Downs_By_Penalty                       Games
                0.006807726                 0.001356333
                       Year             Rush_Touchdowns
                0.000000000                -0.011759799
         Penalties_In_Yards         First_Downs_Rushing
               -0.031373984                -0.032205343
               Fumbles_Lost               Rush_Attempts
               -0.041731982                -0.063464931
                  Penalties          Pass_Interceptions
               -0.068473443                -0.081778533
%_Drives_Ending_In_Turnover                   Turnovers
               -0.083588242                -0.084450050
                 Rush_Yards                        Rank
               -0.095634424                -0.099206959
     Rush_Yards_Per_Attempt
               -0.101957107
```

It was shown that First_Down_Passing, Pass_Completion, Pass_Attempts, Pass_Yards, and %_Drives_Ending_in_Score were the most impactful to winning, indicating that having more of a passing-based offense is more likely to win the Super Bowl than one focused around running the football.

However, something we noticed was that all of these coefficients were similarly weak in strength. The strongest coefficient value rounded to about 0.154, so even though all of the stronger variables were surrounding the pass game, the graph indicated that just having the best pass game wouldn't garuntee a Super Bowl ring. So that means that the rush game is still important to consider what makes a Super Bowl-winning team.

Still, to follow with our data, we decided to choose Total Yards, Average Points Scored Over Time, and Average Yards Over Time to see any trends:



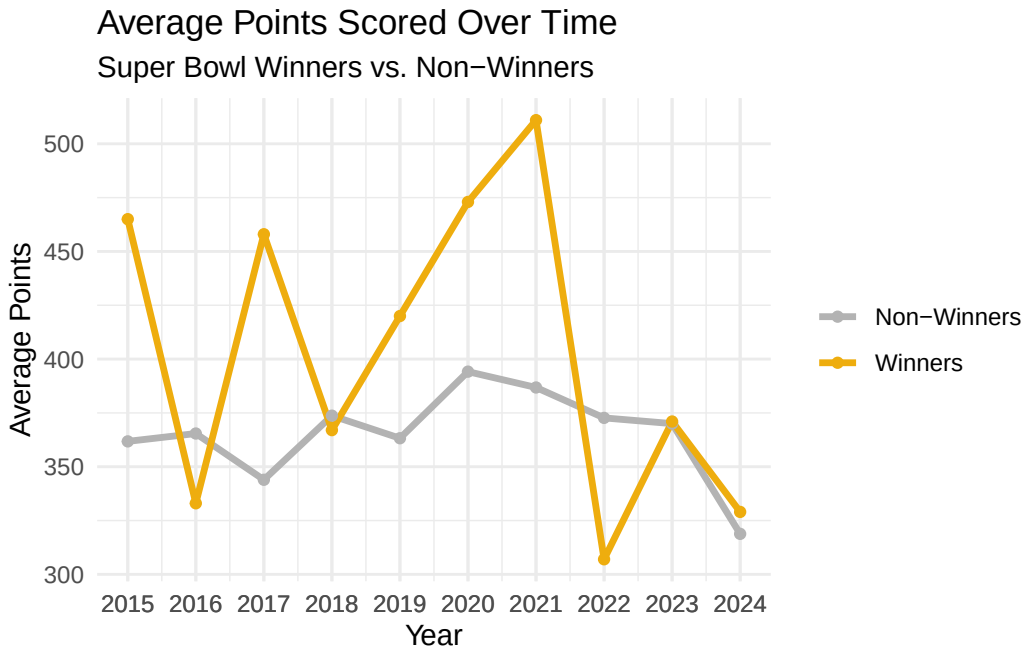


As seen above from both graphs, it's noticeable that the winning teams tend to have a higher average of yards (passing and rushing combined) when being compared to the other teams in the league. We decided to compare the average of all the teams that lost each season in the second graph with each winner since it would be more readable. It also indicated that

higher yards compared the the rest of the league meant a better chance of winning the Super Bowl. We do however see some outliers from the trend, noticeably in 2022 (LA Rams) and 2016 (Denver Broncos). Both teams throughout their respected seasons didn't perform at the top, but still made an amazing push throughout the playoffs, resulting in them lifting the trophy. These outliers also help indicate that although these trends are beneficial to look at, the winning team doesn't always has to follow them.

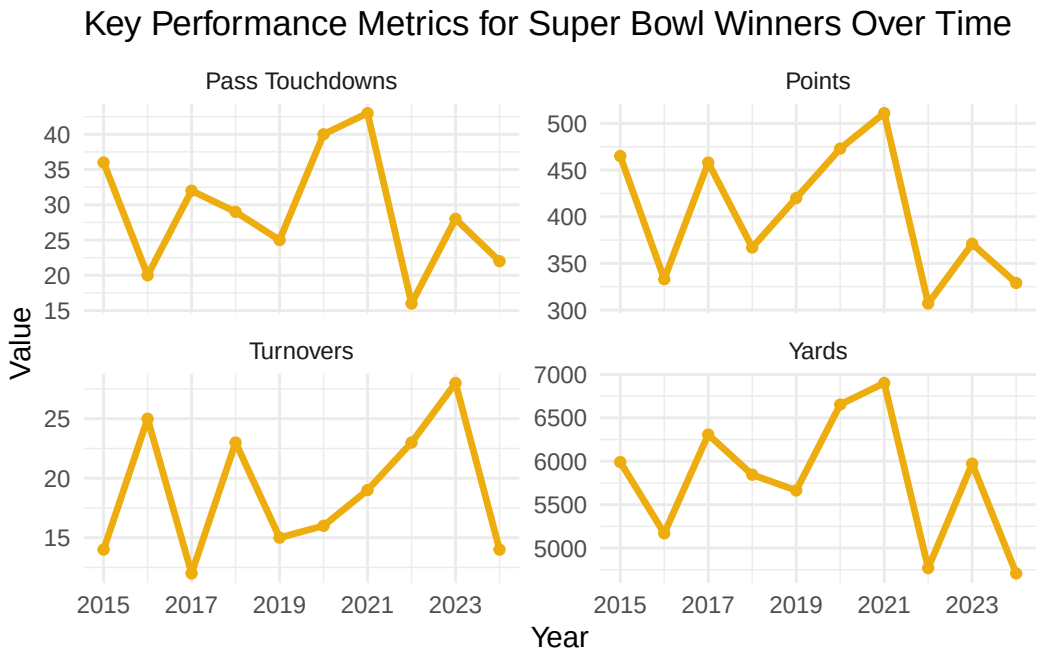
We then wanted to look at the total points scored over time between the winning team and the average of the remaining 31 teams to see a trend through that. Our graph below shows more insight:

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



As you can see, both 2022 and 2016 are outliers again, proving that the winning team doesn't always have to over-perform above the average. Yet, it is still apparent that it is an indicator that most Super Bowl winners outshine their competition immensely in scoring points.

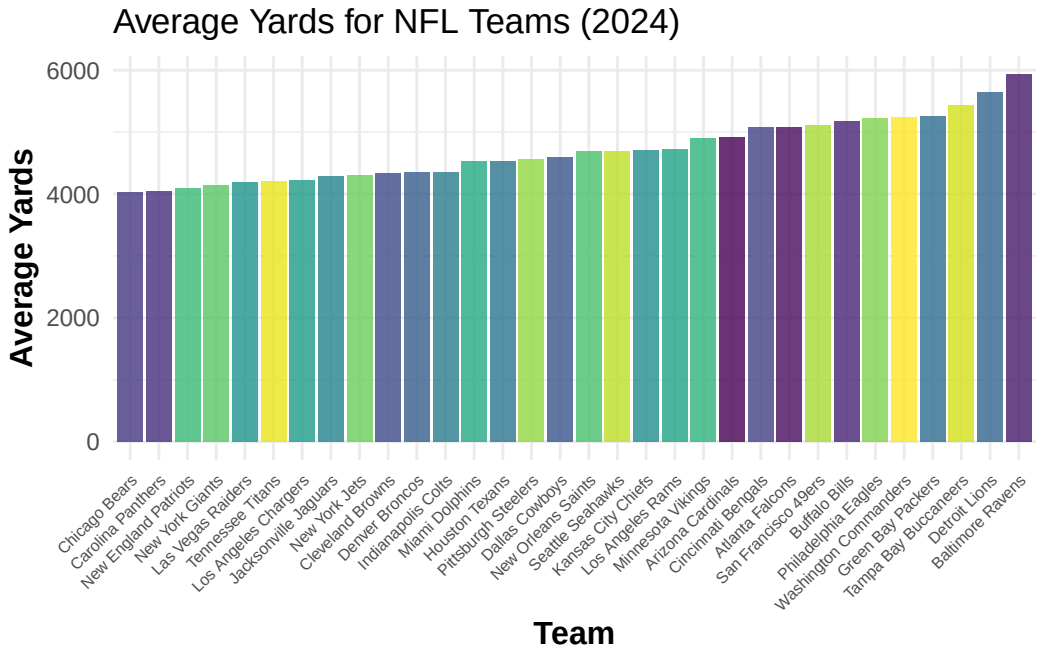
Before taking a look at the 2024 data and making our prediction, we also wanted to see the direction of different categories over time for all the Super Bowl winners. This included Passing Touchdowns, Points, Turnovers, and Yards.

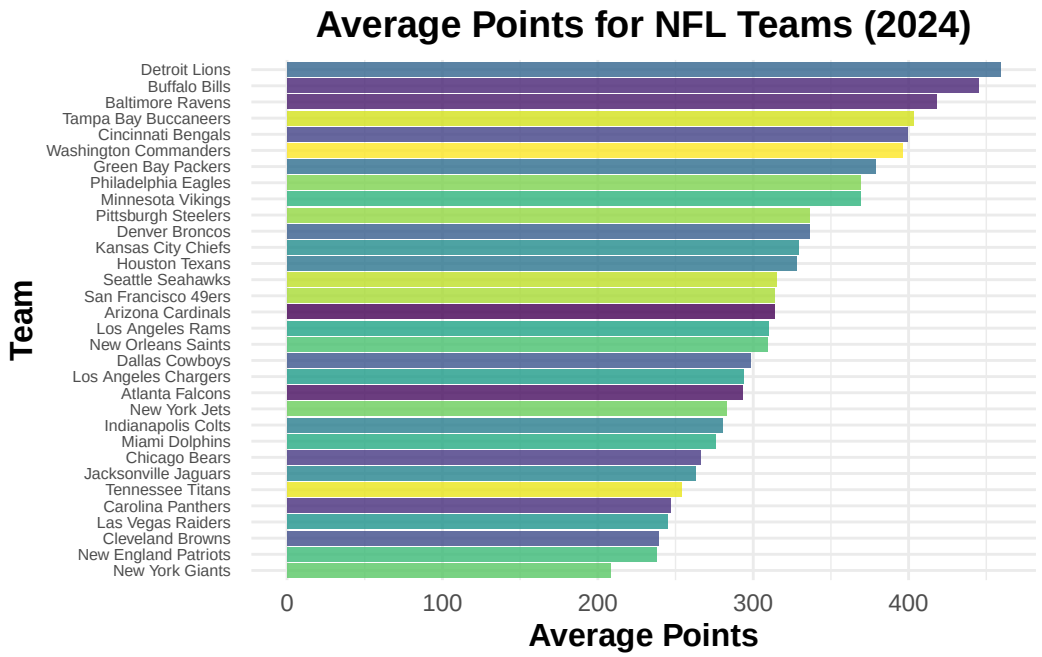


From these four graphs, we were able to find a lot of inconsistency between each winner, which resulted in no definite trends to exist. This can also be supported by the correlation coefficients above, which indicated weak correlations between these variables and winning the Super Bowl.

From our findings so far, we were able to discover that the winning team most likely had to have above 5500 total yards of offense and above 350 total points on the year. We also have to reiterate that these trends aren't 100% with all the Super Bowl winners; there are always outliers.

As the 2024 NFL regular season nears its conclusion, the playoff picture is coming into focus, showing intense debates about potential Super Bowl contenders. We will quickly dive into the standout performances of the season.





The Detroit Lions have emerged as formidable contenders, leading the league in points scored, showcasing their strong offense. Meanwhile, the Baltimore Ravens have dominated in total yards gained, reflecting their ability to move the ball consistently throughout the season.

When we analyze the data and compare it with the current NFL landscape and past Super

Bowl winners, all signs point to the Ravens as the team most poised to claim the title. Their high-powered offense and consistent performance make them the strongest candidates to bring home the championship.

Discussion

From the different data sets we were able to create, we were able to discover that Super Bowl winners in the past 8 years tended to usually have above 5500 total yards of offense and above 350 total points on the year. Looking at the different graphs showing the average points and yards in 2024, the two most stand-out teams that follow these trends the best are the Detroit Lions and the Baltimore Ravens. Since Total_Yards does however have a higher correlation coefficient that points, we predict that the Baltimore Ravens will win the Super Bowl for the 2024 season.

Work Cited

"2016 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2016/index.htm.

"2018 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2018/index.htm.

"2020 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2020/index.htm.

"2021 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2021/index.htm.

"2022 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2022/index.htm.

"2023 NFL Standings & Team Stats." *Pro-Football-Reference.com*, www.pro-football-reference.com/years/2023/index.htm.

Pro Football Reference. "2015 NFL Standings & Team Stats | Pro-Football-Reference.com." *Pro-Football-Reference.com*, 2015, www.pro-football-reference.com/years/2015/index.htm. Accessed 19 Dec. 2024.

—. "2017 NFL Standings & Team Stats | Pro-Football-Reference.com." *Pro-Football-Reference.com*, 2017, www.pro-football-reference.com/years/2017/index.htm.

—. "2019 NFL Standings & Team Stats | Pro-Football-Reference.com." *Pro-Football-Reference.com*, 2019, www.pro-football-reference.com/years/2019/index.htm.

—. "2024 NFL Standings & Team Stats | Pro-Football-Reference.com." *Pro-Football-Reference.com*, 2024, www.pro-football-reference.com/years/2024/index.htm.

Code Appendix

```
# Load Required Libraries
library(dplyr)
library(readxl)
library(esquisse)
library(ggplot2)
library(tidyr)
library(scales)
library(reshape2)
library(corrplot)

# Data Download and Preparation
links <- list(
  "2015" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2016" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2017" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2018" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2019" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2020" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2021" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2022" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2023" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
  "2024" = "https://raw.githubusercontent.com/Stat184-Fall2024/Sec4_FP_NixonKameen_JayCush_Na
)

# Read all data and combine
all_data <- lapply(names(links), function(year) {
  temp_file <- tempfile(fileext = ".xlsx")
  download.file(links[[year]], temp_file, mode = "wb")
  read_excel(temp_file) %>%
    mutate(Year = as.integer(year))
}) %>%
  bind_rows()

# Define standard column names
standard_columns <- c(
```

```
  "Rank", "Team", "Games", "Points_Scored", "Yards", "Plays", "Yards_Per_Play",
  "Turnovers", "Fumbles_Lost", "First_Downs", "Pass_Completions", "Pass_Attempts",
  "Pass_Yards", "Pass_Touchdowns", "Pass_Interceptions", "Net_Yards_Per_Pass_Attempt",
  "First_Downs_Passing", "Rush_Attempts", "Rush_Yards", "Rush_Touchdowns",
  "Rush_Yards_Per_Attempt", "First_Downs_Rushing", "Penalties", "Penalties_In_Yards",
  "First_Downs_By_Penalty", "%_Drives_Ending_In_Score", "%_Drives_Ending_In_Turnover",
  "Expected_Points_Contributed", "Year"
)

# Rename columns to standard
colnames(all_data) <- standard_columns[1:ncol(all_data)]

# Data Cleaning
cleaned_data <- all_data %>%
  filter(!is.na(Rank)) %>%
  filter(!grepl("Avg|League|Total|Tm", Team)) %>%
  mutate(across(!c("Team"), as.numeric)) %>%
  mutate(
    SuperBowl_Winner = case_when(
      (Team == "Kansas City Chiefs" & Year == 2024) ~ 1,
      (Team == "Kansas City Chiefs" & Year == 2023) ~ 1,
      (Team == "Los Angeles Rams" & Year == 2022) ~ 1,
      (Team == "Tampa Bay Buccaneers" & Year == 2021) ~ 1,
      (Team == "Kansas City Chiefs" & Year == 2020) ~ 1,
      (Team == "New England Patriots" & Year == 2019) ~ 1,
      (Team == "Philadelphia Eagles" & Year == 2018) ~ 1,
      (Team == "New England Patriots" & Year == 2017) ~ 1,
      (Team == "Denver Broncos" & Year == 2016) ~ 1,
      (Team == "New England Patriots" & Year == 2015) ~ 1,
      TRUE ~ 0
    )
  )

# Ranking Within Each Year
ranked_data <- cleaned_data %>%
  group_by(Year) %>%
  mutate(across(
    where(is.numeric) & !c(SuperBowl_Winner), # Exclude SuperBowl_Winner
    ~ rank(-., ties.method = "min")
  )) %>%
  ungroup() %>%
  select(-Games, -Rank)
```

```r
# Output the final dataset to check the structure
head(ranked_data)

# Basic summary statistics of numeric columns
overall_summary <- cleaned_data %>%
  select(where(is.numeric)) %>%
  summary()

overall_summary


# Create a column to indicate if a team has ever won a Super Bowl
aggregated_data <- cleaned_data %>%
  group_by(Team) %>%
  summarise(
    Total_Points_Scored = sum(Points_Scored, na.rm = TRUE),
    Ever_SuperBowl_Winner = ifelse(any(SuperBowl_Winner == "1"), "Winner", "Non-Winner"),
    .groups = "drop"
  )

# Plot with the updated column
ggplot(aggregated_data, aes(x = reorder(Team, Total_Points_Scored), y = Total_Points_Scored,
  geom_bar(stat = "identity", alpha = 0.8) +
  scale_fill_manual(
    values = c("Non-Winner" = "grey70", "Winner" = "darkgoldenrod2")
  ) +
  labs(
    title = "Total Points Scored by Team",
    subtitle = "Super Bowl Winners vs. Non-Winners",
    x = "Team",
    y = "Total Points Scored",
    fill = "Super Bowl Status"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

  ggplot(cleaned_data, aes(x = factor(SuperBowl_Winner), y = Yards, fill = factor(SuperBowl_W
  geom_boxplot(alpha = 0.7) +
  scale_fill_manual(
    values = c("0" = "grey70", "1" = "darkgoldenrod2"),
    labels = c("Non-Winner", "Winner")
  ) +
```

```
  scale_x_discrete(
    labels = c("0" = "Non-Winner", "1" = "Winner")
  ) +
  labs(
    title = "Distribution of Total Yards",
    x = "",
    y = "Yards"
  ) +
  theme_minimal() +
  theme(legend.position = "none")

  yearly_stats <- cleaned_data %>%
  group_by(Year, SuperBowl_Winner) %>%
  summarise(avg_points = mean(Points_Scored, na.rm = TRUE), .groups = "drop")

ggplot(yearly_stats, aes(x = Year, y = avg_points, color = factor(SuperBowl_Winner))) +
  geom_line(size = 1.2) +
  geom_point() +
  scale_color_manual(values = c("0" = "grey70", "1" = "darkgoldenrod2"),
                     labels = c("Non-Winners", "Winners")) +
  scale_x_continuous(breaks = yearly_stats$Year) +
  labs(title = "Average Points Scored Over Time",
       subtitle = "Super Bowl Winners vs. Non-Winners",
       x = "Year",
       y = "Average Points") +
  theme_minimal() +
  theme(legend.title = element_blank())

  yearly_stats <- cleaned_data %>%
  group_by(Year, SuperBowl_Winner) %>%
  summarise(avg_yards = mean(Yards, na.rm = TRUE), .groups = "drop")

ggplot(yearly_stats, aes(x = Year, y = avg_yards, color = factor(SuperBowl_Winner))) +
  geom_line(size = 1.2) +
  geom_point() +
  scale_color_manual(values = c("0" = "grey70", "1" = "darkgoldenrod2"),
                     labels = c("Non-Winners", "Winners")) +
  scale_x_continuous(breaks = yearly_stats$Year) +
  labs(title = "Average Yards Over Time",
       subtitle = "Super Bowl Winners vs. Non-Winners",
       x = "Year",
       y = "Average Yards") +
```

```r
  theme_minimal() +
  theme(legend.title = element_blank())

  winners_over_time <- cleaned_data %>%
  filter(SuperBowl_Winner == 1) %>%
  group_by(Year) %>%
  summarise(
    median_points = median(Points_Scored, na.rm = TRUE),
    median_yards = median(Yards, na.rm = TRUE),
    median_turnovers = median(Turnovers, na.rm = TRUE),
    median_pass_tds = median(Pass_Touchdowns, na.rm = TRUE),
    .groups = "drop"
  )

winners_long <- winners_over_time %>%
  pivot_longer(cols = c("median_points", "median_yards", "median_turnovers", "median_pass_tds
               names_to = "Statistic",
               values_to = "Value") %>%
  mutate(Statistic = recode(Statistic,
                            "median_points" = "Points",
                            "median_yards" = "Yards",
                            "median_turnovers" = "Turnovers",
                            "median_pass_tds" = "Pass Touchdowns"))

ggplot(winners_long, aes(x = Year, y = Value)) +
  geom_line(color = "darkgoldenrod2", size = 1.2) +
  geom_point(color = "darkgoldenrod2") +
  facet_wrap(~Statistic, scales = "free_y") +
  scale_x_continuous(breaks = winners_long$Year) +
  labs(title = "Key Performance Metrics for Super Bowl Winners Over Time",
       x = "Year",
       y = "Value") +
  theme_minimal()

  # Convert SuperBowl_Winner to numeric if not already done
cleaned_data <- cleaned_data %>%
  mutate(SuperBowl_Winner = as.numeric(SuperBowl_Winner))

numeric_data <- cleaned_data %>%
  select(where(is.numeric))

corr_matrix <- cor(numeric_data, use = "complete.obs", method = "pearson")
```

```r
# Correlation Heatmap
melted_corr <- melt(corr_matrix)
ggplot(melted_corr, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0,
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()) +
  labs(title = "Correlation Heatmap Including SuperBowl_Winner")

  melted_corr <- melt(corr_matrix)

sbw_correlations <- melted_corr %>%
  filter(Var1 == "SuperBowl_Winner" | Var2 == "SuperBowl_Winner") %>%
  filter(Var1 != Var2) %>%
  arrange(desc(abs(value)))

head(sbw_correlations, 10)

# Filter data for the 2024 season
teams_2024 <- cleaned_data %>%
  filter(Year == 2024) %>%
  group_by(Team) %>%
  summarise(
    Avg_Yards = mean(Yards, na.rm = TRUE),
    Avg_Points = mean(Points_Scored, na.rm = TRUE)
  )

# Create scatter plot with unique colors for each team
# Create bar chart
ggplot(teams_2024, aes(x = reorder(Team, Avg_Yards), y = Avg_Yards, fill = Team)) +
  geom_bar(stat = "identity", alpha = 0.8, show.legend = FALSE) +
  scale_fill_viridis_d() + # Optional: Use a visually appealing color palette
  labs(
    title = "Average Yards for NFL Teams (2024)",
    x = "Team",
    y = "Average Yards"
  ) +
  theme_minimal() +
  theme(
```

```
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10), # Rotate team names for rea
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold")
  )

  # Create horizontal bar chart
ggplot(teams_2024, aes(x = reorder(Team, Avg_Points), y = Avg_Points, fill = Team)) +
  geom_bar(stat = "identity", alpha = 0.8, show.legend = FALSE) +
  scale_fill_viridis_d() +
  labs(
    title = "Average Points for NFL Teams (2024)",
    x = "Average Points",
    y = "Team"
  ) +
  theme_minimal() +
  coord_flip() + # Flips the axes
  theme(
    axis.text.y = element_text(size = 10, margin = margin(r = 10), lineheight = 1.2), # Adjus
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold")
  )
```