NFL Rushing vs. Passing Analysis

Introduction and Background

This project is an analysis on which type of strategy correlates best with winning in the NFL. There are two main strategies pursued by NFL teams, rushing and passing. Passing is where a team prioritizes throwing the ball, whereas rushing prioritizes handing the football to a running back and having them run the football. The main question this analysis attempts to answer is which of these two strategies best correlates with winning in the NFL. In attempting to answer this question, I chose a dataset from Kaggle called NFL Team Data 2003-2023. This dataset was beneficial because it contained numerous variables that aided in answering the main research question. In addition, Kaggle rated this dataset's usability score 10.00, suggesting that the dataset is very reliable. The dataset contains 35 variables, but needs to be transformed to meet the requirements of this analysis. The transformed dataset is named "aggregated_data" and looks like this:

aggregated_data

Year	team_status	avg_win_pct	avg_rush_yds_per_att	avg_pass_yds_per_att	avg_int_per_pass_att	avg_fumble_per_rush_att	avg_passing_tds	avg_rushing_tds	avg_plays_offense	normalized_passing_tds_per_play	normalized_rushing_tds_per_play
2003	Good	0.680	4.1	6.0	0.035	0.026	19.8	15.6	1013.1	0.020	0.015
2003	Mid	0.500	4.4	6.0	0.031	0.026	24.1	13.3	1020.3	0.024	0.013
2003	Poor	0.328	3.9	5.5	0.036	0.024	17.4	10.4	966.6	0.018	0.011
2003	Terrible	0.250	4.3	5.2	0.034	0.035	16.0	10.5	996.5	0.016	0.011

The transferred data includes the following variables:

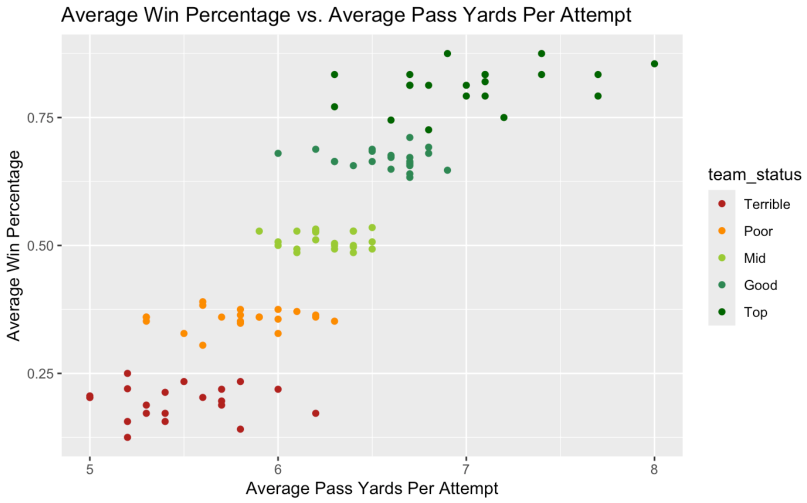
- year: A collective season's worth of data based on team status
- team_status: Divided into 5 categories - Top, Good, Mid, Poor, and Terrible. This variable represents the performance of teams during a given year. For example, Top represents the top 10% of teams during a given year. The top 10% of teams are aggregated into a single Top variable for each year from 2003-2023. This is done for the other four categories as well, albeit at different cutoff percentages.
- avg_win_pct: The average win percentage a team_status has for a given year
- avg_rush_yds_per_att: The average rushing yards per rushing attempt a team_status has for a given year
- avg_pass_yds_per_att: The average passing yards per passing attempt a team_status has for a given year
- avg_int_per_pass_att: The average number of interceptions thrown for every pass attempt a team_status has for a given year
- avg_fumble_per_rush_att: The average number of fumbles for every rushing attempt a team_status has for a given year
- avg_passing_tds: The average number of passing touchdowns a team_status has for a given year. This variable is only used to help calculate normalized_passing_tds_per_play
- avg_rushing_tds: The average number of rushing touchdowns a team_status has for a given year. This variable is only used to help calculate normalized_rushing_tds_per_play

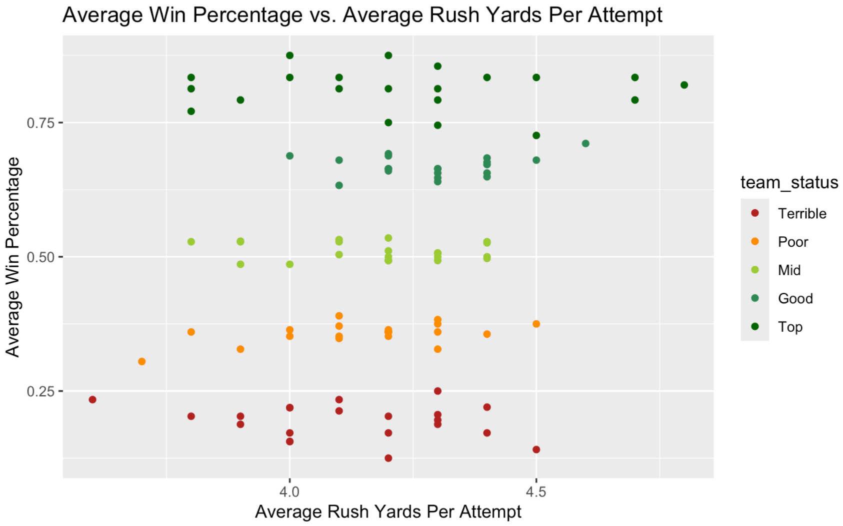
- avg_plays_offense: The average number of plays on offense a team_status has for a given year. This variable is only used to help calculate normalized_passing_tds_per_play and normalized_rushing_tds_per_play
- normalized_passing_tds_per_play: Every time an aggregated team_status group in a given year passes the football, this number represents the percentage that pass attempt results in a touchdown
- normalized_rushing_tds_per_play: Every time an aggregated team_status group in a given year rushes the football, this number represents the percentage that rush attempt results in a touchdown

The main research question can best be answered by breaking it into three subquestions. The first subquestion asks whether rushing net yards per attempt or passing net yards pet attempt is a bigger indicator of winning in the NFL. The second subquestion asks whether the normalized and average normalized rushing or passing touchdowns per play better correlates with winning in the NFL. Finally, the third subquestion asks whether fumbles, a rushing mistake, or turnovers, a passing mistake, correlate most with losing in the NFL. According to Dr. Ed Feng in "The surprising truth about passing and rushing in the NFL", Feng argues changes in the 1970's contribute to passing being more important than rushing. For example, the NFL changed rules related to guarding a wide receiver. Instead of being able to put his hands on a wide receiver throughout an entire route, a defensive back was now limited to only touching the wide receiver within the first 5 yards of his route, which is an argument in favor for passing being a greater indicator of success. In addition, Feng also argues that the rise of the West Coast Offense, that is the prioritizing of short to intermediate passes to skill position players rather than long deep passes, drops down the potential for interceptions and increases the likelihood that a wide receiver would gain just as many yards running after the catch than if a quarterback just threw a deep ball. (Feng, https://thepowerrank.com/2018/09/24/the-surprising-truth-about-passing-and-rushing-in-the-nfl/). In contrast, Nathan Wetmore, the author of, “How to Score Touchdowns: An Analytical Approach to Perfecting Red Zone Percentage”, states during his linear regression model analysis, “the clearly most important factor in predicting red zone percentage is rushing first down percentage” (Wetmore, https://www.bruinsportsanalytics.com/post/analytical-red-zone-percentage). His analysis showed that a rushing statistic best correlated with red zone percentage. Red zone percentage is a measure of how likely a team is to score when they are within 20 yards of the goal line. The positive correlation between rushing first down percentage and red zone percentage favors that rushing is more important to winning, at least under those circumstances. These analyses argue differing answers to the main research question. The following exploratory data analysis hopes to send light on the truth behind rushing vs passing variables and win percentage.

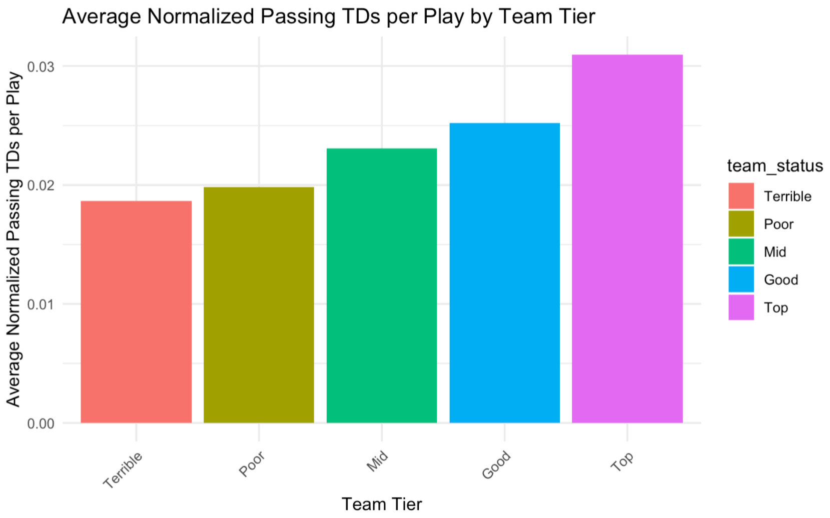
Exploratory Data Analysis

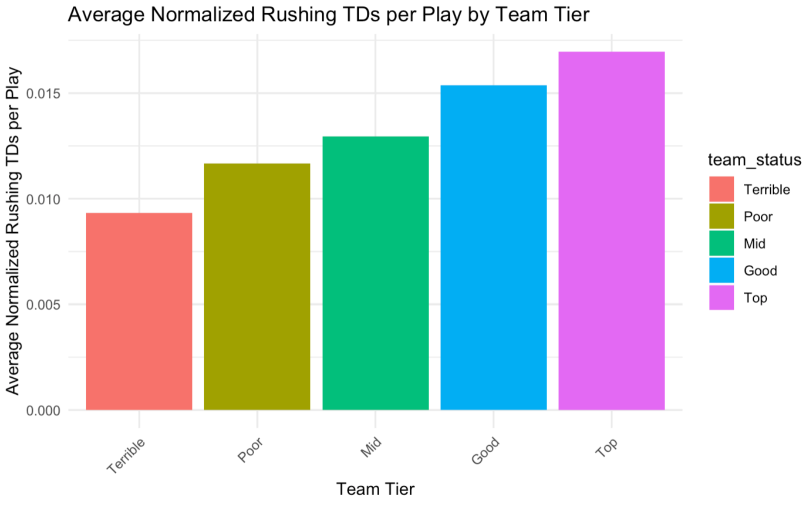
For the first subquestion, scatterplots were used to find insights. The scatterplot comparing win percentage to average passing yards per attempt show a near linear relationship. As win percentage increases, the cluster of dots representing each team tier also increases linearly. There is some greater variability in the Terrible and Top tiers for this scatterplot, but the scatterplot shows a general linear trend. On the other hand, win percentage against average rushing yards per attempt was much less linear. The clusters look more piled on top of each other rather than appearing linear, like the win percentage versus average pass yards per attempt scatterplot. The majority of points on the scatterplot, regardless of team status, fall between 4.0 and 4.5 rush yards per attempt, with some outliers extending beyond 4.5 for Top and Good tier teams. In addition, all tiers except the Good tier had plots existing below the 4.0 rush yards per attempt metric. This seems to suggest that passing correlates strongly with win percentage, whereas rushing does not due to the large variance in scatterplot data for each team tier.



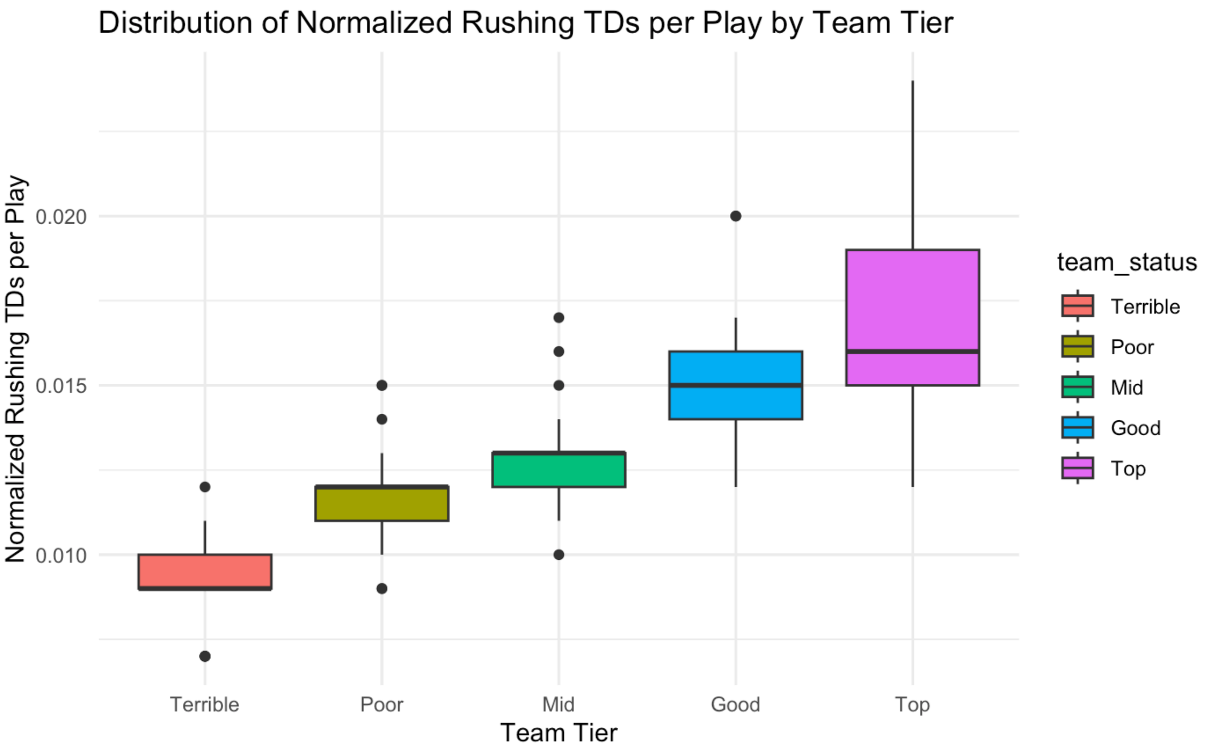


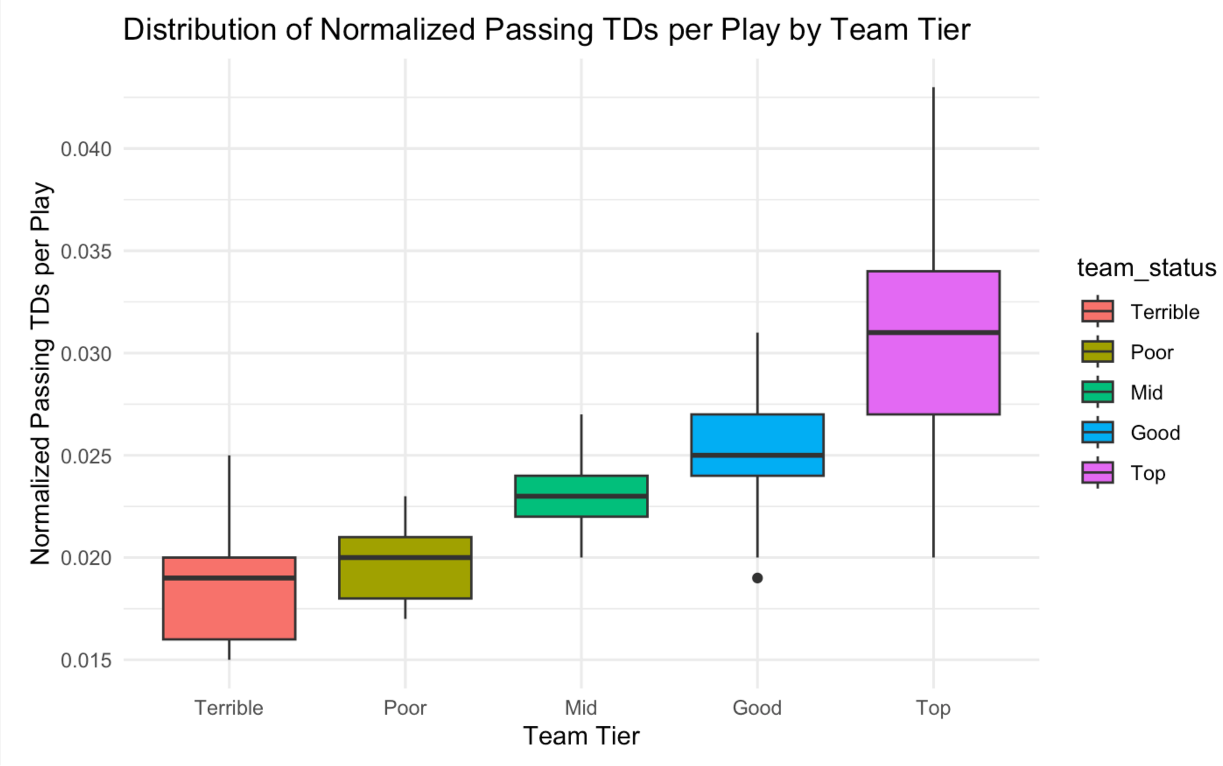
The second subquestion was explored using bar plots and box plots. Firstly, bar plots were used to get a general understanding of how average normalized passing/rushing touchdowns per play related to each team tier. Both rushing and passing average normalized touchdowns increases as the tiers increased, showing a general linear increase. Average normalized passing touchdowns per play started at around 0.019 for Terrible teams, whereas average normalized rushing touchdowns per play started at around 0.017, indicating that regardless of team status, teams general score more passing touchdowns per play on average compared to rushing touchdowns per play on average. In addition, the overall range for rushing touchdowns is much less than the overall range for passing touchdowns.


Average Normalized Passing TDs per Play by Team Tier


Average Normalized Rushing TDs per Play by Team Tier

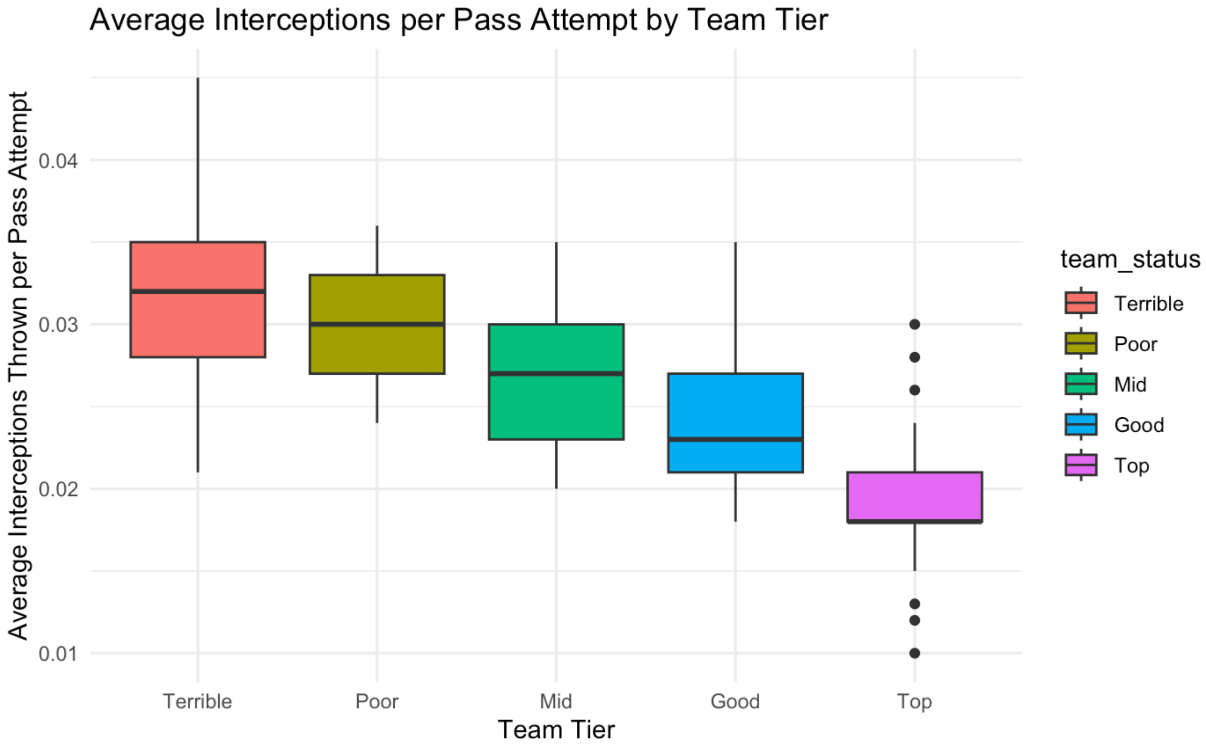
Second, box plots were used to help analyze question 2. The box plots show that top tier teams exhibit an extremely high normalized passing touchdown rate and a wide range of performance. In contrast, lower-tiered teams, like the Terrible tier, have an extremely difficult time scoring touchdowns and have a tighter variation in performance. This is further supported by the fact that the Top tier median line is almost twice that of the Terrible tier median line. Rushing touchdowns vary from passing touchdowns in that all tiers show more consistent distribution. Although the Top tier has a higher median and quartile values, the difference between tiers is not as pronounced as normalized passing touchdowns.

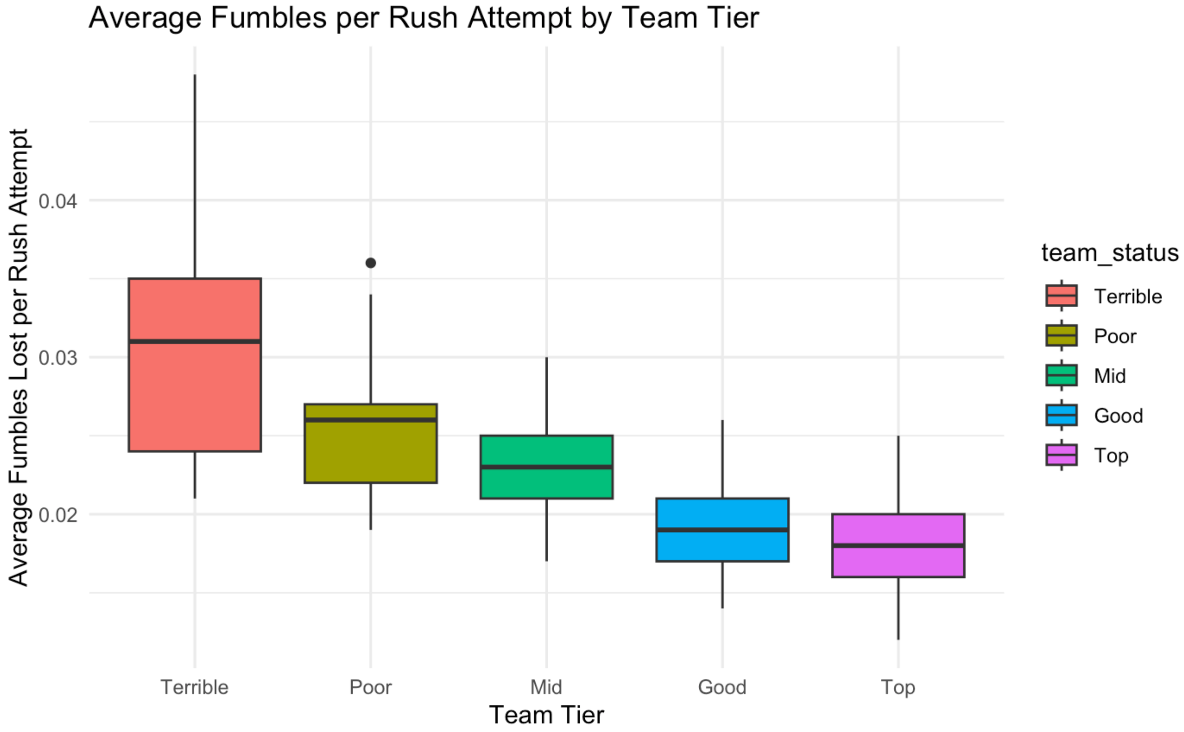




The large variance in the Top tier for normalized passing plays interprets the importance of explosive passing plays. Explosive passing plays generate seem to translate to higher degrees of success. The lower tier teams struggle to generate high normalized passing touchdowns per play, thus resulting in less explosive plays and less opportunity to take control of the game. In contrast, the more consistent distribution for normalized rushing touchdowns per play suggests that rushing performance is not a great indicator of performance difference between team tiers. Overall, this suggests that passing is a greater indicator or winning than rushing.

The third subquestion utilized box plots to visualize the relationship team tiers had with fumbles and interceptions. The box plots show that both fumbles and interceptions correlate negatively with win percentage, as expected. However, there is no discernible difference between the level of negativity. In other words, one type of mistake is not as costly as another type of mistake. This suggests, that fumbles and interceptions are equally as detrimental to a team's win





percentage. Furthermore, it would be more appropriate to state that mistakes in general, not just rushing or passing, are detrimental to win percentage.

Inferential Analysis

Data Transformation

I chose to perform an inferential analysis on my first subquestion. That is, do rushing net yards per attempt or passing net yards pet attempt correlate greater with winning in the NFL? For data cleaning and transformations, I transformed the original Kaggle dataset into the modified aggregated_data dataset, containing all of the necessary variables. The main transformation I performed was aggregating the data by team category (Top, Good, Mid, Poor, and Terrible). For each category, I calculated the average rushing yards per attempt, passing yards per attempt, and win percentage. This allowed myself to evaluate the performance of teams in each category based on their rushing/passing yards per attempt and how those two variables related to win percentage. The main goal of the inferential analysis was on comparing these key statistics, rushing yards per attempt, passing yards per attempt, and win percentage, across these five tiers of teams. This helped determine which type of game is most effective at winning, whether rushing or passing efficiency were better indicators of success for different tiers of teams.

Methods and Assumptions

Outliers were not removed in this analysis. This was because outliers within the dataset may have held meaningful information, and removing them may have withheld valuable insights about NFL offensive efficiency. The decision to keep outliers may have had an unintended consequence on certain statistical methods (e.g., correlations and regression). This could lead to skewed results. However, the decision to retain the outliers provides a more complete view on the efficiency of both the air and ground games within the league and how they influence the win percentage. Visualizations such as scatterplots will emphasize and lead to assessments on the influence of these outliers. Any potential influences by these outliers on the correlation and regression statistical methods will be included in the results section.

Statistical Methods

There were four statistical methods that were explored during this inferential analysis: correlation analysis, linear regression, hypothesis testing, and confidence intervals. Pearson's correlation was used to assess the strength and direction of the linear relationship between rushing yards per attempt, passing yards per attempt, and win percentage. A coefficient near +1 or -1 indicated a strong relationship, while a value near 0 suggested a weak relationship.

A multiple linear regression model was used to determine contributions to win percentage from rushing yards per attempt and passing yards per attempt. This model estimated the relationship between the dependent variable (win percentage) and the independent variables (passing yards per attempt and rushing yards per attempt). The goal of the regression was to determine which independent variable had a better correlation with the dependent variable and thus which type of game is more effective for team success in the NFL. The regression provided coefficient estimates and p-values, which were beneficial in determining the strength of relationship between the independent variables and the dependent variable.

The null hypothesis (H0) states that the independent variables are not significantly correlated with the dependent variable. In other words, rushing and passing yards per attempt do not have a significant impact on winning NFL games. The alternative hypothesis (H1) states that the independent variables are significant correlated with the dependent variable, that is, rushing and passing yards per attempt do have a significant impact on winning NFL games.

A 95% confidence interval was calculated for both rushing and passing yards per attempt in regards to their relationship with win percentage. The goal of the confidence interval was to better understand the uncertainty around the estimated effect of the two variables.

Outline of Assumptions

The relationship between win percentage and rushing/passing yards per attempt must be linear. I checked for linearity by inspecting scatter plots of the data, using the Pearson correlation, and using a multiple linear regression model to inspect p-values and coefficients. I assumed that the relationship between rushing yards per attempt and win percentage was linear. However, the analysis showed that rushing yards per attempt had a weak correlation with win percentage ($r = 0.24$) and was not statistically significant to linear regression ($p = 0.334$). This suggested that either other factors would be better contributors to win percentage, or the relationship between rushing yards per attempt and win percentage is not strictly linear.

The observations in this analysis are independent, as we are analyzing data aggregated by team categories over different years. There is no relationship between data for different years or team categories.

The errors, defined as the differences between the observed win percentages and the predicted values, must follow a normal distribution. This is important for hypothesis testing and constructing confidence intervals. I assessed normality by visually inspecting the residual plots (e.g., a histogram and Q-Q plot). Finally, I confirmed normality by doing a Shapiro-Wilk test. If the histogram follows a bell-shaped curve, the Q-Q plot's values exist mostly on the red line, and Shapiro-Wilk test returns a p-value greater than 0.05, these three tests confirm the residuals are normally distributed.

The variance of the residuals must be constant across all levels of the independent variable (rushing or passing yards). I visually inspected residual plots for any patterns that suggest unequal variance. If the plot showed a random scatter of residuals with no discernible pattern (e.g., no funnel shapes or trends), this suggested that the assumption of equal variance holds. Furthermore, I conducted a Breusch-Pagan Test to formally test for Equal Variance. If the p-value returned by the test was greater than 0.05, then equal variance was assumed.

Findings and Interpretation of Results

From the Pearson's correlation analysis, I discovered that the relationship between passing yards per attempt and win percentage was strong and positive (r = 0.88). This indicates a strong correlation between the two variables, suggesting teams that pass more efficiently win more games. On the other hand, I discovered that the relationship between rushing yards per attempt and win percentage was weak (r = 0.24). This suggest that rushing efficiency alone is not a good indicator of winning, and that more variables need to be considered alongside examining rushing efficiency. These unknown factors may explain the reasoning behind NFL teams continuing to utilize the ground game.

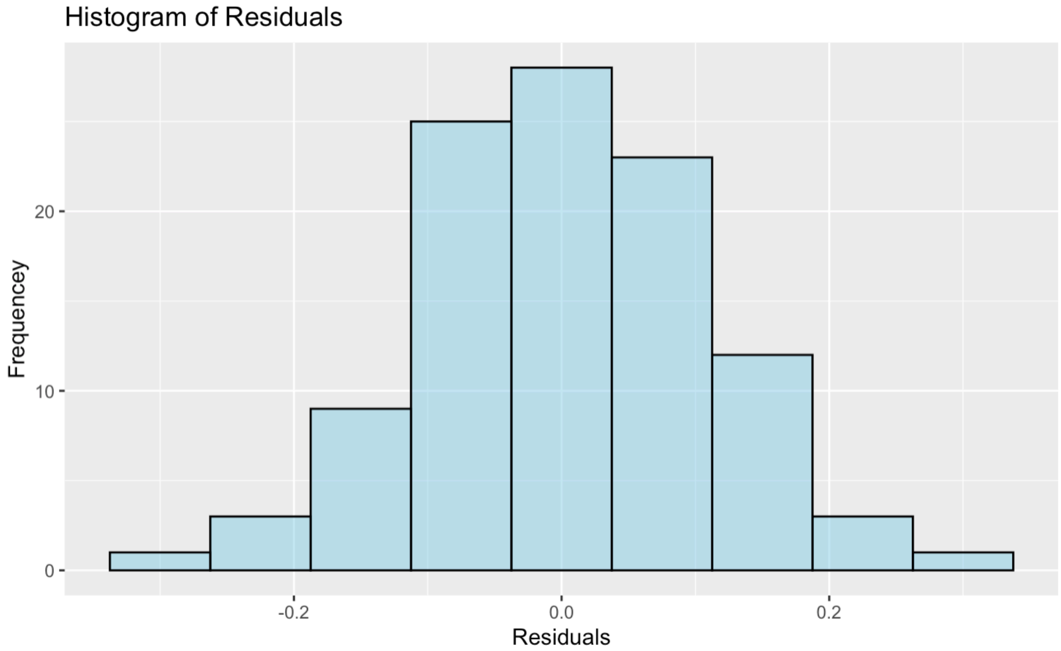
The regression coefficient for average passing yards per attempt was B1 = 0.312 and had a p-value well below 0.05 (p = 2e-16). These two factors indicate that for every one yard gained from passing, win percentage increases. In particular the model predicated that each additional passing yard per attempt resulted in an increase in win percentage anywhere from 0.28% and 0.35% (based on the confidence interval). However, the regression coefficient for average rushing yards per attempt was B2 = -0.049, with a p-value of 0.334. These values indicate that average rushing yards per attempt was not statistically significant to winning in the NFL. While other factors, like field position and offensive line may contribute greatly to the overall ground game, run efficiency alone is not a good indicator of winning NFL games.

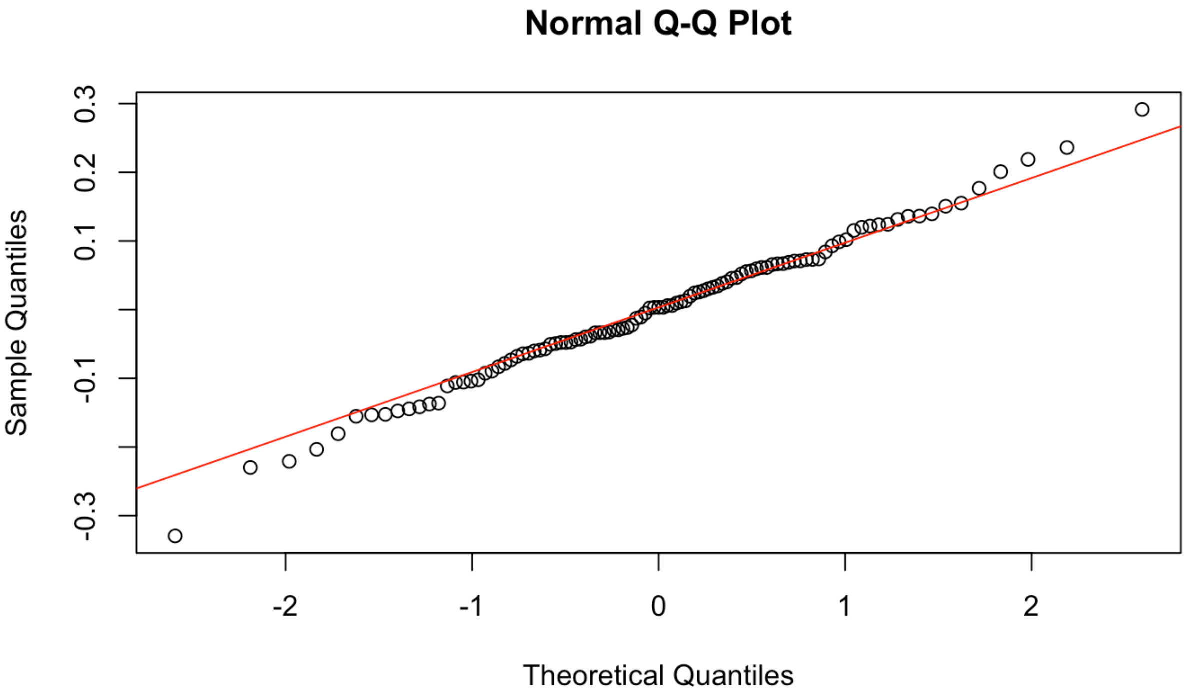
The analysis shows that, of the two types of strategy, the air game (passing the football) is a greater predictor of winning NFL games than the ground game (running the football). This result is further backed by the increased reliance of the passing game, dynamic quarterback play, and high-powered offenses, like the West Coast Offense, in recent years. Although rushing is still utilized in the NFL, it seems that rushing efficiency alone, unlike passing efficiency, is an unreliable contributor to winning in the NFL.

Recall the null hypothesis (H0), that there is no significant relationship between rushing yards per attempt and win percentage. The alternative hypothesis (H1) was that there is a significant relationship. The same null and alternative hypothesis conditions exist for passing yards per attempt when compared to win percentage. The hypothesis tests further reinforces our earlier finding from the regression analysis. For average rushing yards per attempt, the p-value was 0.334, which is greater than the common threshold of 0.05. This means that we fail to reject the null hypothesis for rushing yards per attempt. In other words, there is no strong evidence to suggest rushing yards per attempt is a good indicator of win percentage. Thus, the null hypothesis stands for rushing yards per attempt. For average passing yards per attempt, the p-value was 2e-16, which is well below the common threshold of 0.05. This means that we reject the null hypothesis, and there exists a significant relationship between passing efficiency and win percentage. Thus, the null hypothesis is rejected for passing yards per attempt.

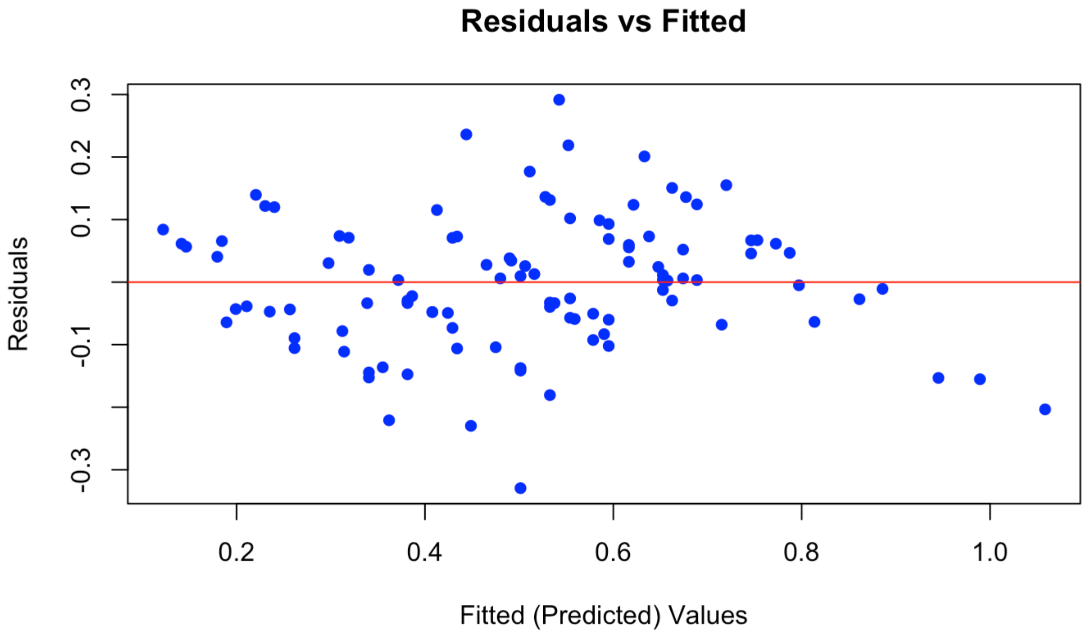
I calculated 95% confidence intervals for both average passing yards per attempt and average rushing yards per attempt, and found that only passing yards per attempt contained a confidence interval that held 0 within its range, with the interval being [0.2775, 0.3468]. This coincides with the rest of our findings because a confidence interval that does not have 0 within its range is statistically significant. The rushing yards per attempt confidence interval did contain a 0 within its range. That confidence interval was [-0.1491, 0.0511]. This supports my previous findings that rushing yards per attempt is not significantly related to win percentage.

My findings showed that the residuals follow a normal distribution. This was evident with the histogram of residuals displaying a bell-shaped curve. Furthermore, my Q-Q Plot had residuals at or near the red line. Finally, I formalized my findings with a Shapiro-Wilk test, and found that the p-value was 0.954, indicating that normality can be safely assumed.


Histogram of Residuals
Frequencey
20
10
0
-0.2
0.0
0.2
Residuals


Normal Q-Q Plot
Sample Quantiles
0.3
0.2
0.1
-0.1
-0.3
-2
-1
0
1
2
Theoretical Quantiles

I found that the graph displaying the residuals on a plot were relatively randomly scattered. Furthermore, I did a formal Breusch-Pagan test and found that my p-value was 0.62, well above the standard 0.05 and indicating that equal variance can be assumed.



Conclusion and Future Questions

This project sought to determine what type of strategy was most effective at winning NFL games, a passing game versus a rushing game. The analysis determined that passing was a more effective contributor to winning games than rushing. Furthermore, the explosive play potential of passing plays is a major contributor to win percentage. In contrast, rushing was not well correlated to win percentage. The linear regression model determined that average rushing yards per play did not correlate to win percentage. While rushing is still utilized in the NFL, rushing efficiency alone is not a great indicator of winning games, and other variables, like strong offensive line, may be needed to be analyzed in tandem with average rushing yards per play to explain why rushing the ball is still utilized by NFL teams. Lastly, turnovers, whether fumbles or interceptions, contribute negatively to winning games. While interception are shown on the graph to have a slightly more negative impact than fumbles, I considered this negative impact to be negligible due to the closeness in y-axis data between fumbles and interceptions.

In the future, I am hoping to analyze red zone efficiency and how well that correlates with the ground and air games. Determining which type of strategy would work best within 20 yards of the end zone could illuminate whether coaches more often try for touchdowns or settle for field goals depending on which strategy is implemented. Furthermore, determining whether rushing or passing contributes to more touchdowns or field goals could greater illuminate which type of strategy correlates best with winning.

The dataset provided valuable insights, but was limited in the sense that it did not have diverse variables for specific positions. In the future, I would like to examine how an offensive line variable, like average offensive line efficiency, compares with average rushing yards per attempt and determine how those two variables combined correlate with win percentage. In addition, finding a dataset that contains individual games with play-by-play data could provide a more in-depth analysis on which type of strategy is best utilized at specific times within the game. For example, perhaps the air game is best utilized in general, but the ground game is specifically utilized in the red zone or when teams have a lead.

Code Appendix

```
# Load libraries I probably need

library(dplyr)
library(ggplot2)
library(tidyr)
library(readr)
library(lmtest)
```

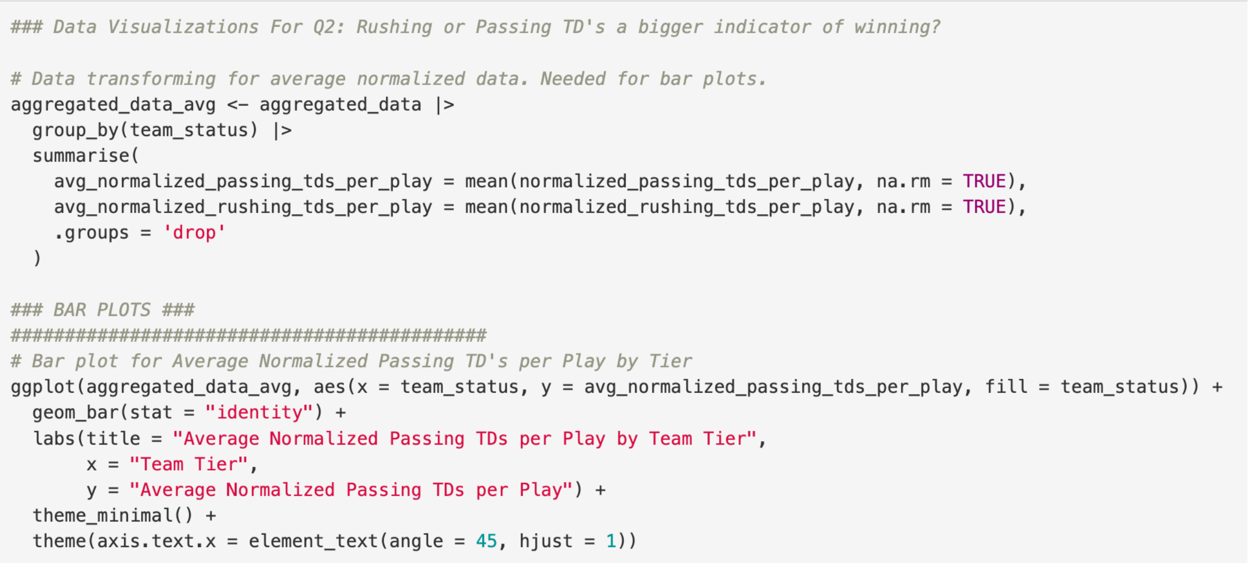
```
# Load NFL Dataset
NFL_Data <- read_csv("/Users/RyanMiller/Documents/MATH167R/team_stats_2003_2023.csv")
```

```
### Data Cleaning: mov and ties variables

# Calculate mov for N/A values
NFL_Data_Clean <- NFL_Data |>
  mutate(
    ties = ifelse(is.na(ties), 0, ties),
    mov = round((points - points_opp) / g, 1)
  )

### Data Cleaning: Creating a 'rank' column
NFL_Data_Clean <- NFL_Data_Clean |>
  group_by(year) |>
  arrange(desc(win_loss_perc)) |>
  mutate(rank = row_number()) |>
  ungroup()

### Data Cleaning: Cutoff Teams

# Define cutoff points for 5 categories (Top, Good, Mid, Poor, Terrible)
top_cutoff <- floor(0.10 * 32)     # Top Teams
good_cutoff <- floor(0.35 * 32)    # Good Teams
mid_cutoff <- floor(0.65 * 32)     # Mid Teams
poor_cutoff <- floor(0.90 * 32)    # Poor Teams
terrible_cutoff <- 32              # Terrible Teams

# Create new 'team_status' column for categories above
NFL_Data_Clean$team_status <- NA

# Assign team categories based on rank
NFL_Data_Clean$team_status[NFL_Data_Clean$rank <= top_cutoff] <- "Top"
NFL_Data_Clean$team_status[NFL_Data_Clean$rank > top_cutoff & NFL_Data_Clean$rank <= good_cutoff] <- "Good"
NFL_Data_Clean$team_status[NFL_Data_Clean$rank > good_cutoff & NFL_Data_Clean$rank <= mid_cutoff] <- "Mid"
NFL_Data_Clean$team_status[NFL_Data_Clean$rank > mid_cutoff & NFL_Data_Clean$rank <= poor_cutoff] <- "Poor"
NFL_Data_Clean$team_status[NFL_Data_Clean$rank > poor_cutoff & NFL_Data_Clean$rank <= terrible_cutoff] <- "Terrib
le"
```

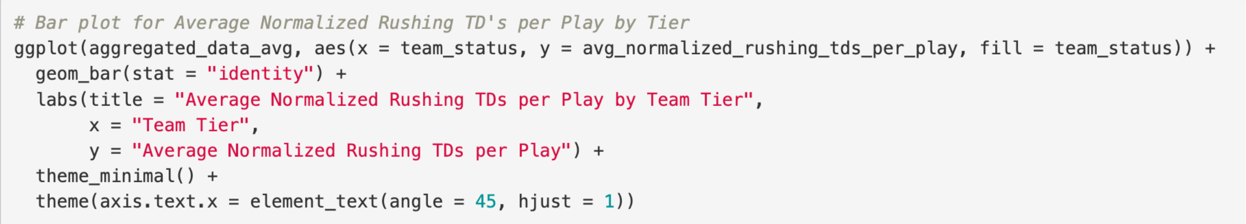
```
# Filter data by team status. Pre-aggregating set up
top_teams <- NFL_Data_Clean |> filter(team_status == "Top")
good_teams <- NFL_Data_Clean |> filter(team_status == "Good")
mid_teams <- NFL_Data_Clean |> filter(team_status == "Mid")
poor_teams <- NFL_Data_Clean |> filter(team_status == "Poor")
terrible_teams <- NFL_Data_Clean |> filter(team_status == "Terrible")

# Aggregate data
combined_teams <- bind_rows(top_teams, good_teams, mid_teams, poor_teams, terrible_teams)

aggregated_data <- combined_teams |>
  group_by(year, team_status) |>
  summarise(
    avg_win_pct = round(mean(win_loss_perc, na.rm = TRUE), 3),
    avg_rush_yds_per_att = round(mean(rush_yds_per_att, na.rm = TRUE), 1),
    avg_pass_yds_per_att = round(mean(pass_net_yds_per_att, na.rm = TRUE), 1),
    avg_int_per_pass_att = round(mean(pass_int / pass_att, na.rm = TRUE), 3),
    avg_fumble_per_rush_att = round(mean(fumbles_lost / rush_att, na.rm = TRUE), 3),
    avg_passing_tds = round(mean(pass_td, na.rm = TRUE), 1),
    avg_rushing_tds = round(mean(rush_td, na.rm = TRUE), 1),
    avg_plays_offense = round(mean(plays_offense, na.rm = TRUE), 1),
    normalized_passing_tds_per_play = round(avg_passing_tds / avg_plays_offense, 3),
    normalized_rushing_tds_per_play = round(avg_rushing_tds / avg_plays_offense, 3),
    .groups = 'drop'
  )

head(aggregated_data)
```

```
### Exploratory Data Analysis
# Data Visualizations for Q1: Relationship between avg_rush_yds_per_att/avg_pass_yds_per_att and avg_win_pct

aggregated_data$team_status <- factor(aggregated_data$team_status,
                                      levels = c("Terrible", "Poor", "Mid", "Good", "Top"))

############################################
### SCATTERPLOTS ###
# Scatterplots: avg_win_pct vs avg_rush_yds_per_att
ggplot(aggregated_data, aes(x = avg_rush_yds_per_att, y = avg_win_pct, color = team_status)) +
  geom_point() +
  scale_color_manual(values = c("Top" = "#006400", "Good" = "#2e8b57",
                                "Mid" = "#9acd32", "Poor" = "#ff8c00",
                                "Terrible" = "#b22222")) +
  labs(title = "Average Win Percentage vs. Average Rush Yards Per Attempt",
       x = "Average Rush Yards Per Attempt", y = "Average Win Percentage")
```

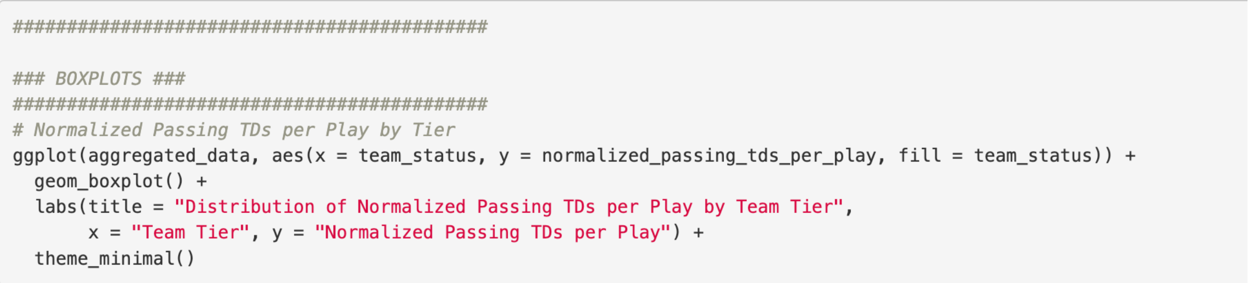
```
# Scatterplots: avg_win_pct vs avg_pass_yds_per_att
ggplot(aggregated_data, aes(x = avg_pass_yds_per_att, y = avg_win_pct, color = team_status)) +
  geom_point() +
  scale_color_manual(values = c("Top" = "#006400", "Good" = "#2e8b57",
                                "Mid" = "#9acd32", "Poor" = "#ff8c00",
                                "Terrible" = "#b22222")) +
  labs(title = "Average Win Percentage vs. Average Pass Yards Per Attempt",
       x = "Average Pass Yards Per Attempt", y = "Average Win Percentage")
```



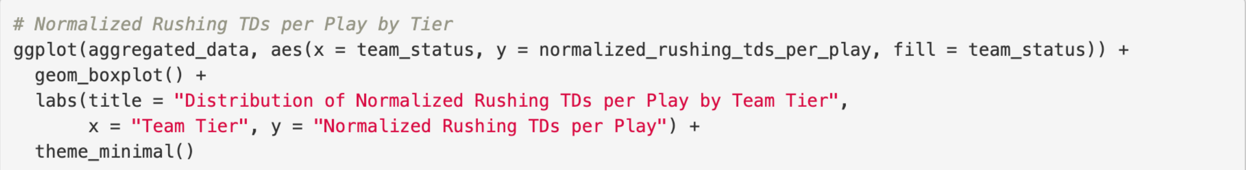
```
### Data Visualizations For Q2: Rushing or Passing TD's a bigger indicator of winning?

# Data transforming for average normalized data. Needed for bar plots.
aggregated_data_avg <- aggregated_data |>
  group_by(team_status) |>
  summarise(
    avg_normalized_passing_tds_per_play = mean(normalized_passing_tds_per_play, na.rm = TRUE),
    avg_normalized_rushing_tds_per_play = mean(normalized_rushing_tds_per_play, na.rm = TRUE),
    .groups = 'drop'
  )

### BAR PLOTS ###
#############################################
# Bar plot for Average Normalized Passing TD's per Play by Tier
ggplot(aggregated_data_avg, aes(x = team_status, y = avg_normalized_passing_tds_per_play, fill = team_status)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Normalized Passing TDs per Play by Team Tier",
       x = "Team Tier",
       y = "Average Normalized Passing TDs per Play") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



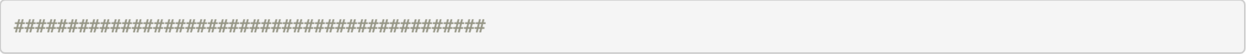
```
# Bar plot for Average Normalized Rushing TD's per Play by Tier
ggplot(aggregated_data_avg, aes(x = team_status, y = avg_normalized_rushing_tds_per_play, fill = team_status)) +
  geom_bar(stat = "identity") +
  labs(title = "Average Normalized Rushing TDs per Play by Team Tier",
       x = "Team Tier",
       y = "Average Normalized Rushing TDs per Play") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



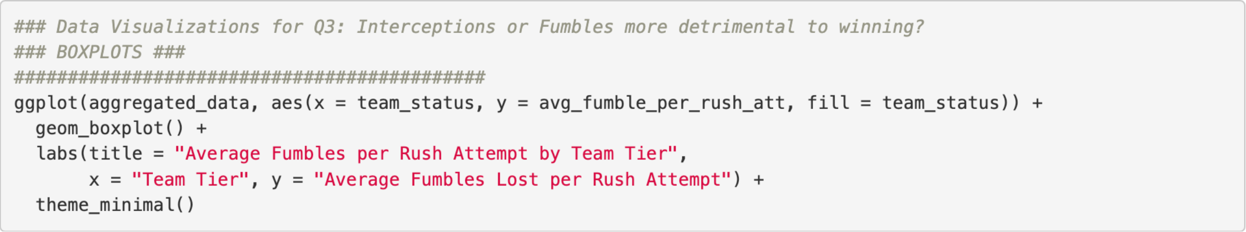
```
#############################################

### BOXPLOTS ###
#############################################
# Normalized Passing TDs per Play by Tier
ggplot(aggregated_data, aes(x = team_status, y = normalized_passing_tds_per_play, fill = team_status)) +
  geom_boxplot() +
  labs(title = "Distribution of Normalized Passing TDs per Play by Team Tier",
       x = "Team Tier", y = "Normalized Passing TDs per Play") +
  theme_minimal()
```



```
# Normalized Rushing TDs per Play by Tier
ggplot(aggregated_data, aes(x = team_status, y = normalized_rushing_tds_per_play, fill = team_status)) +
  geom_boxplot() +
  labs(title = "Distribution of Normalized Rushing TDs per Play by Team Tier",
       x = "Team Tier", y = "Normalized Rushing TDs per Play") +
  theme_minimal()
```



```
#############################################
```



```
### Data Visualizations for Q3: Interceptions or Fumbles more detrimental to winning?
### BOXPLOTS ###
#############################################
ggplot(aggregated_data, aes(x = team_status, y = avg_fumble_per_rush_att, fill = team_status)) +
  geom_boxplot() +
  labs(title = "Average Fumbles per Rush Attempt by Team Tier",
       x = "Team Tier", y = "Average Fumbles Lost per Rush Attempt") +
  theme_minimal()
```

```
ggplot(aggregated_data, aes(x = team_status, y = avg_int_per_pass_att, fill = team_status)) +
  geom_boxplot() +
  labs(title = "Average Interceptions per Pass Attempt by Team Tier",
       x = "Team Tier", y = "Average Interceptions Thrown per Pass Attempt") +
  theme_minimal()
```

```
### Inferential Analysis for Sub-question 1

# Correlation Matrix
cor_matrix <- cor(aggregated_data |>
                    select(avg_win_pct, avg_rush_yds_per_att, avg_pass_yds_per_att),
                  use = "complete.obs")

print(cor_matrix)
```

```
##                      avg_win_pct avg_rush_yds_per_att avg_pass_yds_per_att
## avg_win_pct            1.0000000            0.2408511            0.8774247
## avg_rush_yds_per_att   0.2408511            1.0000000            0.3239736
## avg_pass_yds_per_att   0.8774247            0.3239736            1.0000000
```

```
# Linear Regression Model
l_model <- lm(avg_win_pct ~ avg_rush_yds_per_att + avg_pass_yds_per_att, data = aggregated_data)
summary(l_model)
```

```
## 
## Call:
## lm(formula = avg_win_pct ~ avg_rush_yds_per_att + avg_pass_yds_per_att, 
##     data = aggregated_data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.32945 -0.06009  0.00331  0.06700  0.29154 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          -1.22823    0.20443  -6.008 2.92e-08 ***
## avg_rush_yds_per_att -0.04897    0.05046  -0.970    0.334    
## avg_pass_yds_per_att  0.31215    0.01747  17.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1068 on 102 degrees of freedom
## Multiple R-squared:  0.772,  Adjusted R-squared:  0.7675 
## F-statistic: 172.7 on 2 and 102 DF,  p-value: < 2.2e-16
```

```
### Check Normality of Residuals

# Get residuals
residuals <- residuals(l_model)

# Create histogram of residuals
ggplot(data = aggregated_data, aes(x = residuals)) +
  geom_histogram(binwidth = 0.075, fill = "skyblue", color = "black", alpha = 0.5) +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequencey")
```

```
# Create Q-Q Plot to further check normality
qqnorm(residuals)
qqline(residuals, col = "red")
```

```
# Shapiro-Wilk Test to definitively check normality
shapiro.test(residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.99455, p-value = 0.954
```

```
### Equal Variance Test

# Visual Test
fitted_values <- fitted(l_model)

# Create plot
plot(fitted_values, residuals(l_model),
     main = "Residuals vs Fitted",
     xlab = "Fitted (Predicted) Values",
     ylab = "Residuals",
     pch = 16,
     col = "blue")
abline(h = 0, col = "red")
```

```
# Equal Variance Formal Test
bptest(l_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  l_model
## BP = 0.95608, df = 2, p-value = 0.62
```

```
### Calculate Confidence Interval
confint(l_model)
```

```
##                          2.5 %      97.5 %
## (Intercept)          -1.6337165 -0.82274349
## avg_rush_yds_per_att -0.1490604  0.05111638
## avg_pass_yds_per_att  0.2775093  0.34680041
```