Does Size Matter?

The Impact of College Football Player Size on Team Success

STAT GU 4001 - Probability & Statistics - Spring 2023

Sean Osier (smo2152@columbia.edu)

4/29/2023

Abstract

We explore the variability and importance of player size in top-level (FBS, formerly Division 1-A) college football. Looking at height, weight, and Body Mass Index (BMI), we demonstrate that there are statistically significant differences in the average size of players playing different positions. More importantly, we demonstrate that “team size” has a small, but statistically significant impact on a team’s ultimate success as measured by wins and losses. In doing so, we determine that weight is the measure of size most correlated with team success. Finally, we identify Tight End (TE), Defensive Line (DL), and Defensive Back (DB) as the positions where size differences matter most (simultaneously revealing that size differences at other positions don’t have a statistically significant impact).

Introduction & Problem Statement

College football is among the most popular sports in the United States, attracting millions of fans and generating billions of dollars in revenue each year. It’s a complex sport, with a multitude of different factors subtly influencing the outcome of games. Player skill, teamwork, athleticism, coaching, strategy, play calling, weather, crowd noise, officiating, and more all play a role.

One factor of debated importance is player size (here defined as player height and weight). Football is a contact sport, a game of force, *mass* × *acceleration* as given by Newton's second law. As such, when building their team coaches must consider when it’s worth it to tradeoff between player size and speed / athleticism, as larger players tend to be slower and less agile, and faster players tend to be smaller.

In this analysis, we seek to understand the impact of player size on college football teams’ success as measured by their final season win-loss record. We approach this problem by first trying to understand the variability in player size, investigating whether there are statistically significant differences in the average size of players playing different positions. We then seek to understand the impact of “team size” on a team’s performance, and in doing so identify which (if any) of our measures of player size is most correlated with team success. Finally, we identify the positions where size differences have the most significant impact.

After this introduction and problem statement, this report will include more details on the specific data used for this analysis. We will then describe the methodology and statistical analyses used in this study, present the results of our analysis, and finally conclude with a discussion of the implications and limitations of our findings. Additional analyses and other supporting documentation can be found in the appendix at the end of this report. For readers unfamiliar with the game of football, we’ve included a brief introduction to the game and common player positions in the appendix.

Data Sources

For this analysis we use a data set from the 2022-2023 season consisting of 15,475 top-level (FBS, formerly Division 1-A) football players from all 131 FBS teams. We retrieved this data using the “College Football Data (CFBD)” API available at collegefootballdata.com. The data contains identifier information (ID and first / last name) and the player’s team, height in inches, weight in pounds, and position. (*Table 1 in appendix*)

We combine this player level data with a separate data set from the same source containing the win-loss record for each of the 131 FBS teams during the 2022-2023 season. This data contains the year, team, games played, and number of wins / losses. (*Table 2 in appendix*)

Proposed Methodology

To analyze the data, we first create two derived metrics to condense player size and team success into single metrics:

1. *Body Mass Index (BMI)* based on a player's height and weight
2. *Win Percentage* based on the a team's wins and total games played

As a data cleaning step, we also clean and standardize the player positions. See the appendix for the full details on this cleaning procedure.

With the data cleaned, we begin our analysis of variability in player size, investigating whether there are statistically significant differences in the average size of players playing different positions. We first calculate the mean height, weight, and BMI by position group and graph the results. To determine if the size differences observed are meaningful we first run an ANOVA and find that the differences for each of our three size metrics is significant. Given this result, we then perform a Tukey test to check the significance of the differences between each of the individual position groups.

We then seek to understand the impact of "team size" on a team's performance. To do this, we calculate the mean height, weight, and BMI of the players on each team. Modelling a team's season using a binomial distribution where *n* is the number of games played and *p* is the probability of the team winning any given game, we perform three univariate binomial regressions, one for each of our three size metrics. In order to determine which of the size metrics is most predictive of team success, the size metric is our independent variable and wins / losses is our target variable in these regressions. We verify our results by generating scatter plots of the size metric vs. team win percentage and by plotting expected win percentage conditioned on different cuts of the size metric.

Finally, we calculate the mean height, weight, and BMI for each position group on each team. Again modelling the team's wins / losses as a binomial, we use binomial regression with the mean sizes by position group as the independent variables to identify the positions where size differences matter most. We then use backwards stepwise regression to remove model features (position groups) until we are left with only significant results. We again verify our results by generating scatter plots of the positional size metrics vs. team win percentage and by plotting expected win percentage conditioned on different cuts of the positional size metrics.

Key Results

Our analysis of the variability of player size by position concludes with extremely high confidence (>99.999%, $p < 2 \times 10^{-16}$) that the average size of players varies between position groups. This holds true for all three size metrics we studied (height, weight, and BMI; *Figures 1-3 in the appendix*). In fact, the differences are so strong that it's actually more informative to discuss the position groups between which there was *no* statistically significant difference in player size. A full listing of these not significantly different positions can be found in the appendix (*Tables 3-5*), but to highlight a few of the most interesting: Wide Receivers (WR), Defensive Backs (DB), and Place Kickers (PKs) form a triad of the three lightest positions. None are significantly different from eachother, but all weigh statistically significantly less than all other position groups.

Our analysis of overall "team size" suggests that of the three size metrics we studied weight is most significant predictor of team success (p = 0.0001). Height is not a significant predictor. BMI is a significant predictor (p = 0.0009), but is not as significant as weight likely because it's a

blended metric of weight combined with the non-significant metric height.
We estimate each pound above average a team weighs yields approximately 1.15% incremental win probability. Given most teams play at least 12 games, a team that weighs approximately 7.3 pounds above average would be expected to win about 1 game more per season than the average team. We can visually see this trend in the scatter plot and conditional win probability chart in the appendix (*Figures 4-9*).

Finally, our most detailed team and position level model reveals that Tight End (TE), Defensive Line (DL), and Defensive Back (DB) are the three position groups where weight significantly affects team success. (The BMI metric confirms these same position groups, though with less confidence.) Importantly, this also reveals that weight differences at other positions *don't* have a statistically significant impact. In our final model, Quarterback (QB) was the only position where height mattered, and interestingly it was inversely related to team success, that is teams with taller QBs on average performed worse than teams with shorter QBs.

We estimate each pound above average a team's TEs weigh yields approximately 0.67% incremental win probability. Each pound above average for DL is worth approximately 0.31% win probability. And, an additional pound for a DB is worth roughly 0.90% win probability. This win percentage yield is inversely related to how heavy the position group is on average, that is heavier positions on average need to increase their weight more to yield a similar increase in win probability. Ultimately, this is equivalent to needing to increase TE weights about 12.7 lbs above average for an extra win. The number is 27.4 lbs for DL, and 9.5 lbs for DBs for that extra expected win. We can again visually see this trend in the scatter plot and conditional win probability chart in the appendix (*Figures 10-17*).

Code for the analysis and complete results can be found in the appendix.

Conclusion

In conclusion, player size *does* matter in top-level college football, but it is not universally significant, and it is far from being the only determinant of team success. Weight is what matters, *not* height, and BMI is only relevant insofar as it incorporates weight. However, weight does not matter equally across all positions. It has a significant impact on Tight End (TE), Defensive Line (DL), and Defensive Back (DB), but it does not significantly affect other positions. Even at these positions, each incremental pound yields only a relatively small gain in expected win probability (<1% per pound). Nonetheless, a team with enough of these small incremental gains can expect to win one or even two more games than an average-weight team in expectation.

In many ways, this is good news for coaches and players because unlike height, weight is something that can be changed through exercise and nutrition regimes, both to increase or decrease weight. Given this, our analysis suggests that coaches should prioritize weight management plans to bulk up their TEs, DL, and DBs, while players at other positions could be encouraged to shed some weight if that helps them gain speed and/or agility.

Future analyses on this subject would ideally include other player metrics related to speed and agility to truly understand the tradeoffs between size and speed. A more comprehensive analysis would also try to understand and control for how size matters for starters vs. backups, perhaps by limiting the dataset to starters only or by weighting the averages based on playing time. Finally, future analyses could also be improved by simply including more historical data to ensure that the results are not biased by a single, potentially unrepresentative year.

Appendix

Football Basics / Positions

Football is a team sport played with an oval-shaped ball on a rectangular field. The objective of the game is to score points by advancing the ball into the opponent's end zone (called a touchdown) or kicking it through their goalposts.

There are two main teams in football: the offense and the defense. The offense is responsible for advancing the ball down the field and scoring points, while the defense is responsible for stopping the offense and preventing them from scoring.

Special teams are a separate unit that comes onto the field for kicking plays, such as punts and kickoffs. They are also responsible for field goal attempts and extra point attempts after a touchdown.

There are two main types of plays in football: running and passing. Running plays involve handing the ball to a player having them run with it, while passing plays involve a player throwing the ball to a receiver downfield who may then attempt to run with the ball after catching it. A play continues until the player with the ball is brought down (tackled), goes out of bounds, or scores. The play is also over if a pass is attempted and not completed, i.e. the ball touches the ground.

There are many different formations and strategies that teams can use to execute their plays, including different types of blocking and routes for receivers to run. Overall, football is a complex and dynamic sport that requires players with a variety of different skills and physical attributes to succeed. As such, players typically specialize in a single position or role. This study primary considers following position groups:

- **Offense:**
  - **Quarterback (QB):** Leader of the offensive team, responsible for calling plays in the huddle, receiving the snap from the center, and passing or handing off the ball to the running back
  - **Running Back:** A player primarily responsible for carrying the ball on running plays, who may also be asked to act as a blocker or receiver on passing plays
  - **Wide Receiver (WR):** A player primarily responsible for catching passes from the quarterback, though may be asked to block or more rarely be the ball carrier on running plays
  - **Tight End (TE):** A versatile player who is commonly serves as a blocker on some plays and and a pass catcher on others
  - **Offensive Line (OL):** A group of players who are solely responsible for blocking, both to protect the quarterback on passing plays and create running lanes on running plays
- **Defense:**
  - **Defensive Line (DL):** A group of players responsible for pressuring the quarterback on passing plays and stopping the ball carrier on running plays
  - **Linebacker (LB):** A versatile defensive player who on any given play may be responsible for stopping the run, covering receivers, and/or pressuring the quarterback

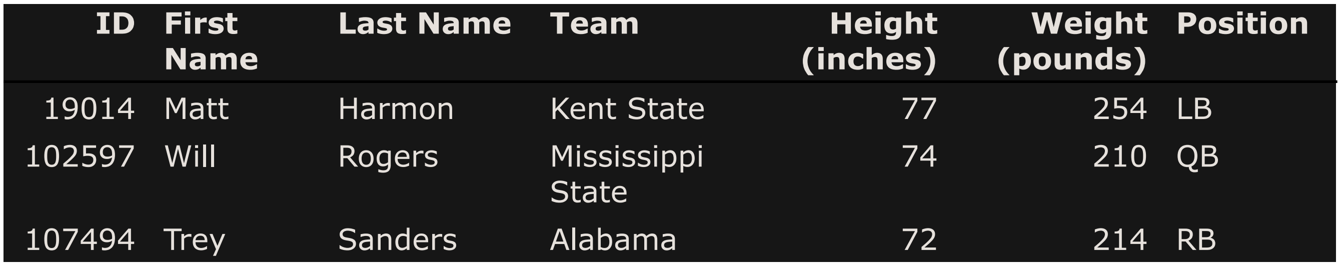
    - **Defensive Back (DB):** A player primarily responsible for covering receivers to prevent completed passes or to try to intercept passes from the quarterback, though they may some responsibility in stopping running plays as well
- **Special Teams:**
    - **Placekicker (PK):** Player who specializes in kicking field goals, extra points, and kickoffs
    - **Punter (P):** Player who specializes in kicking the ball downfield on punting plays to give the opposing team poor field position
    - **Long Snapper (LS):** Player who specializes in snapping the ball back to the punter or placekicker on special teams plays

Data Cleaning

To clean and standardize the player positions, we performed the following data cleaning operations:

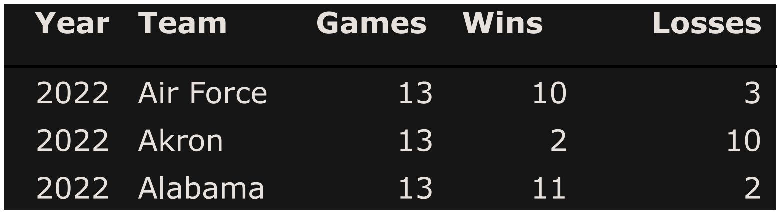
1. We remove the single player listed as a designated Punt Returner (PR), because one player is simply not enough to draw any conclusions from
2. We remove the small number of players listed as "Athletes" (ATH), i.e. players without a set / decided position, effectively players with missing data
3. Because teams differ in how exactly they list players by position, we standardize player positions into "position groups":
    a. To give a concrete example, not every team designates Offensive Linemen (OL) in specific offensive line positions such as Center (C), Guard (G), or Offensive Tackle (OT). As such, in order to ensure every team uses the same positions for their players, we standardize all specific offensive line positions to be just Offensive Linemen (OL)
    b. We do the same for the specific Defensive Line (DL) and Defensive Back (DB) positions
    c. Finally, we have to apply special logic for the Full Back (FB) position:
        i. The Full Back (FB) position is antiquated in many modern offenses and many teams will not even have a single FB on their roster
        ii. That said, there are still other teams where the FB position is a valued and distinct role
        iii. Given we want standardized positions that apply for all teams, we need to group FB's into a position group
        iv. Complicating matters, usage varies for teams that do use the position, and either Running Back (RB) or Tight End (TE) could be appropriate
        v. Ultimately, we assume on triple option teams like Air Force and Navy, RB is the closest position and that on all other teams TE is closest position

Tables



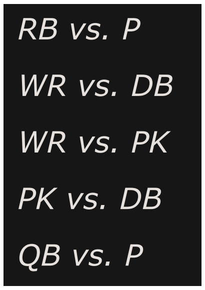
ID	First Name	Last Name	Team	Height (inches)	Weight (pounds)	Position
19014	Matt	Harmon	Kent State	77	254	LB
102597	Will	Rogers	Mississippi State	74	210	QB
107494	Trey	Sanders	Alabama	72	214	RB

Table 1: Selection of College Football Player Data



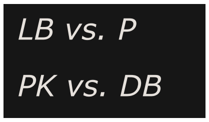
Year	Team	Games	Wins	Losses
2022	Air Force	13	10	3
2022	Akron	13	2	10
2022	Alabama	13	11	2

Table 2: Selection of Team Results Data for the 2022-2023 Season



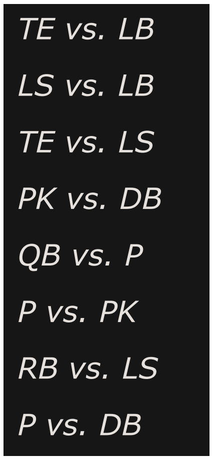
RB vs. P
WR vs. DB
WR vs. PK
PK vs. DB
QB vs. P

Table 3: Position Groups Whose Weights Do NOT Significantly Differ



LB vs. P
PK vs. DB

Table 4: Position Groups Whose Heights Do NOT Significantly Differ



TE vs. LB
LS vs. LB
TE vs. LS
PK vs. DB
QB vs. P
P vs. PK
RB vs. LS
P vs. DB

Table 5: Position Groups Whose BMIs Do NOT Significantly Differ

Figures

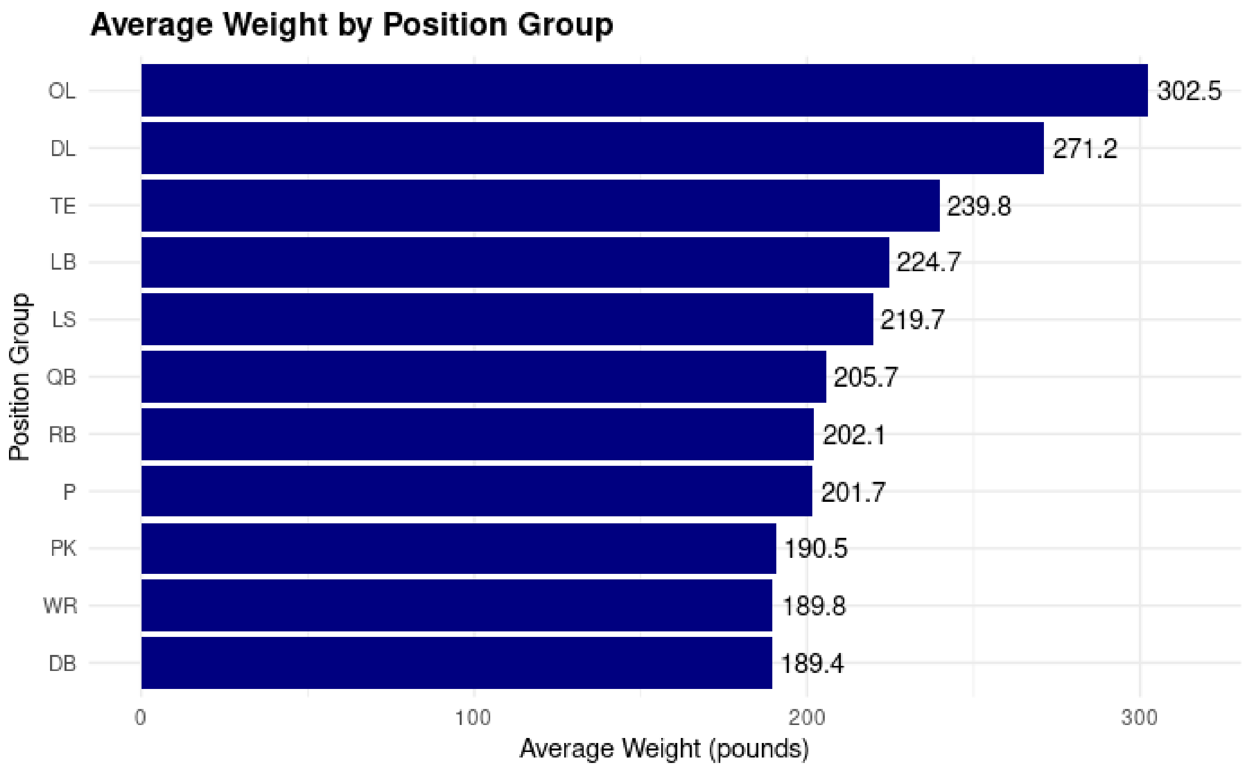


Figure 1: Average Weight by Position Group

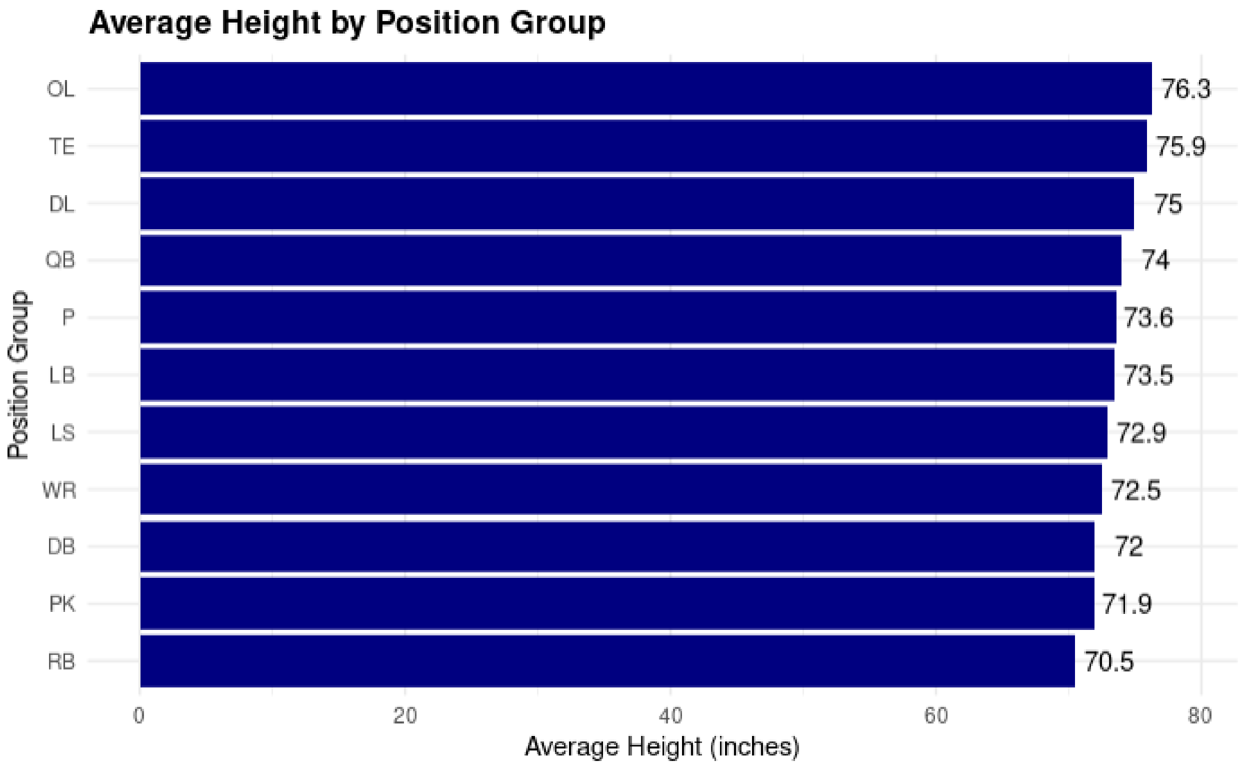


Figure 2: Average Height by Position Group

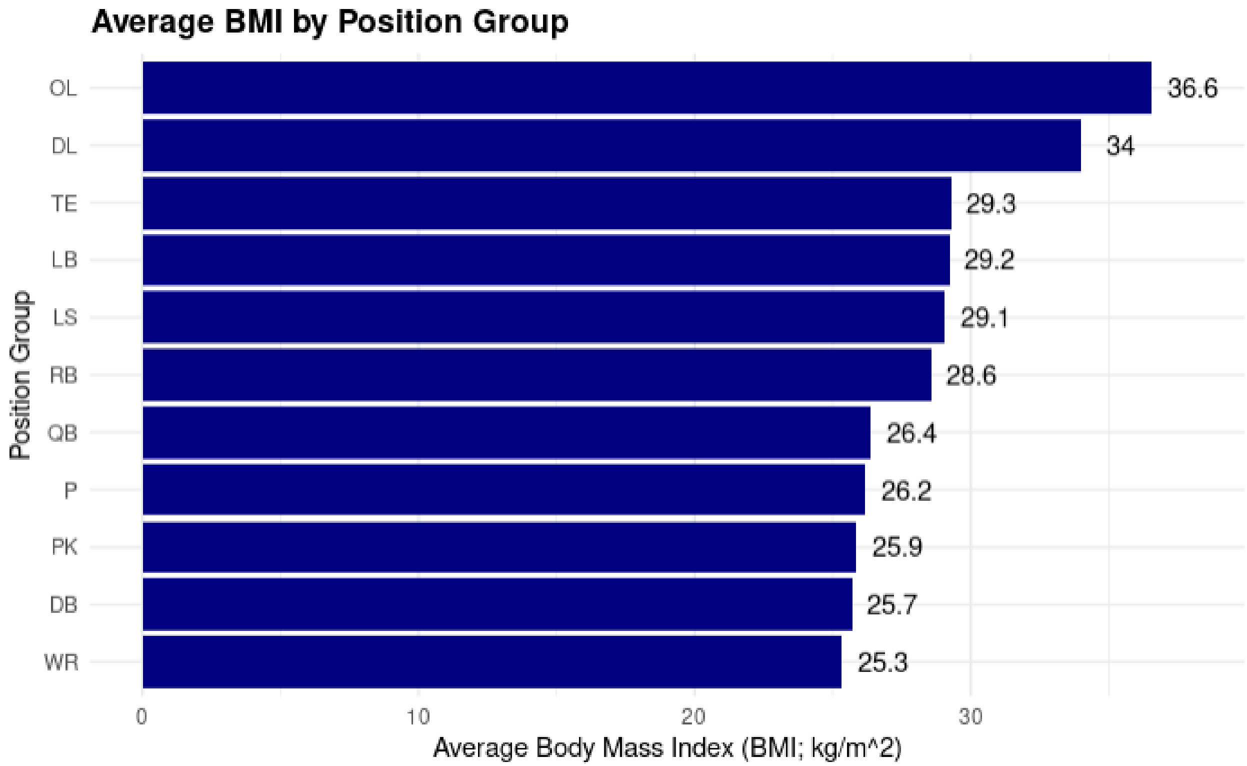


Figure 3: Average BMI by Position Group

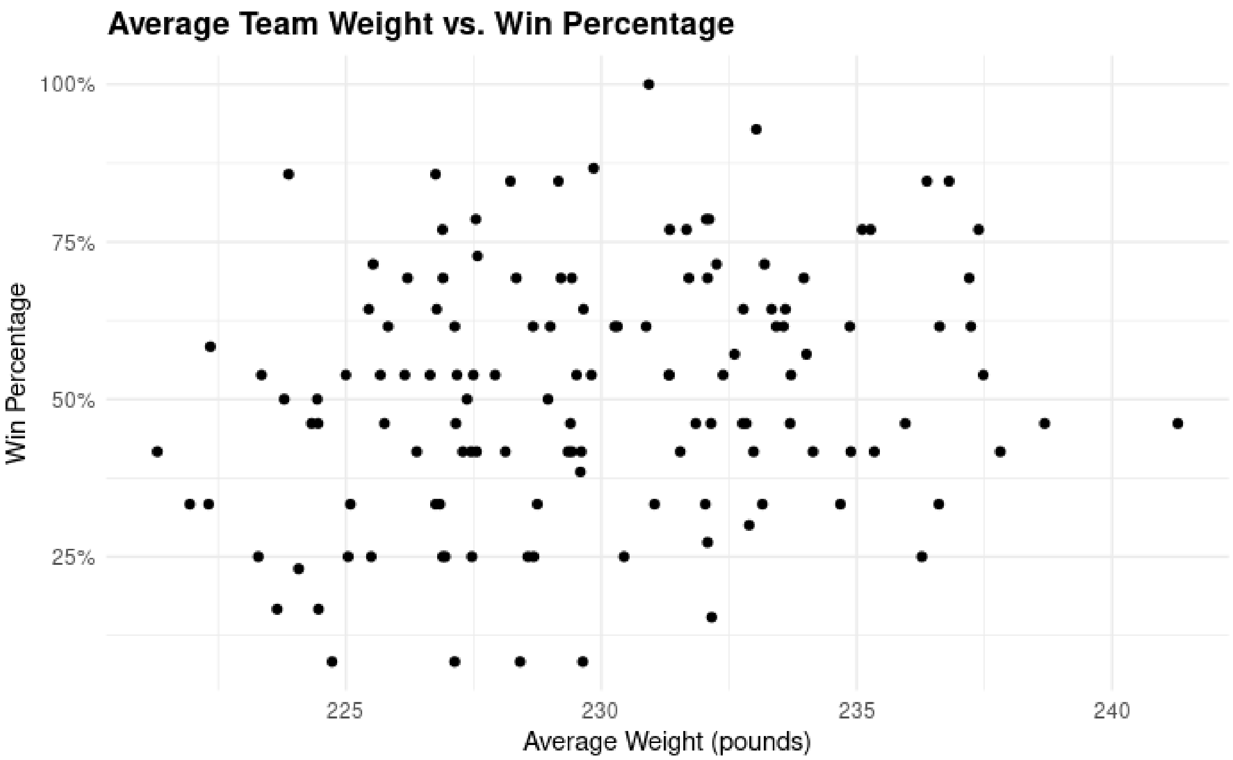


Figure 4: Average Team Weight vs. Win Percentage

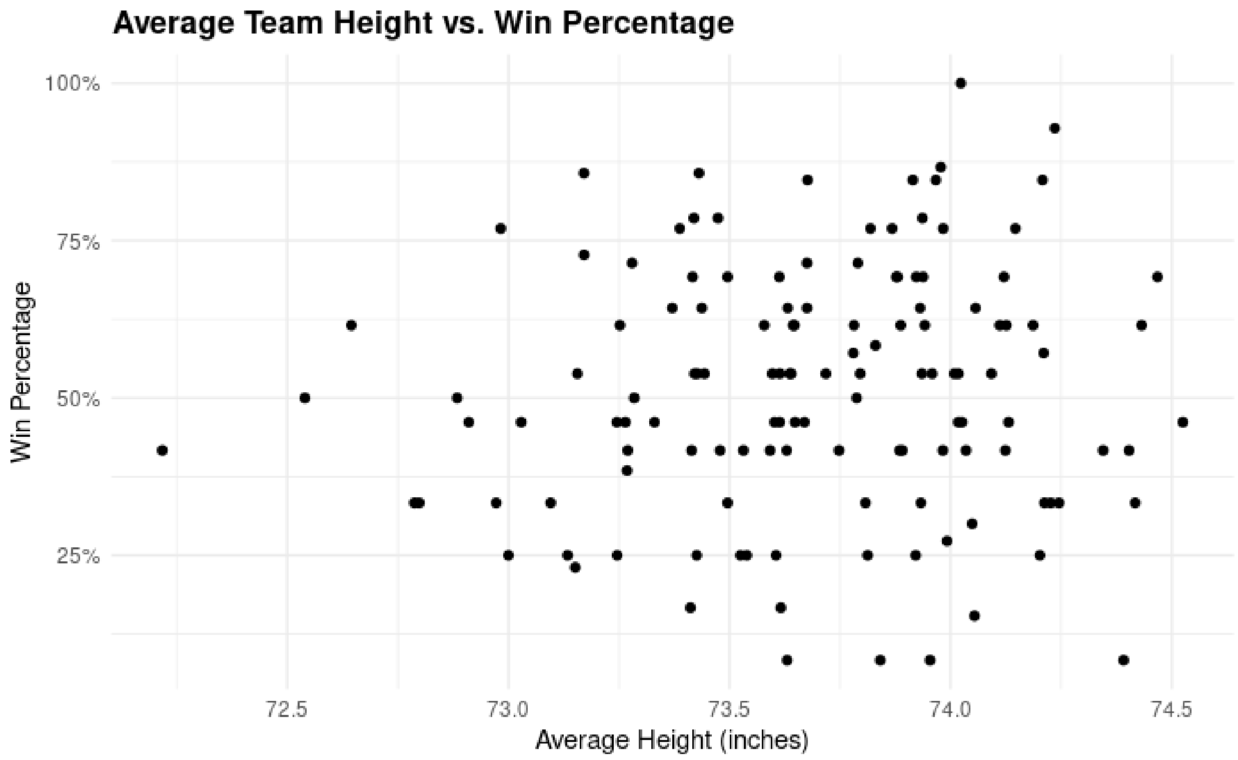


Figure 5: Average Team Height vs. Win Percentage

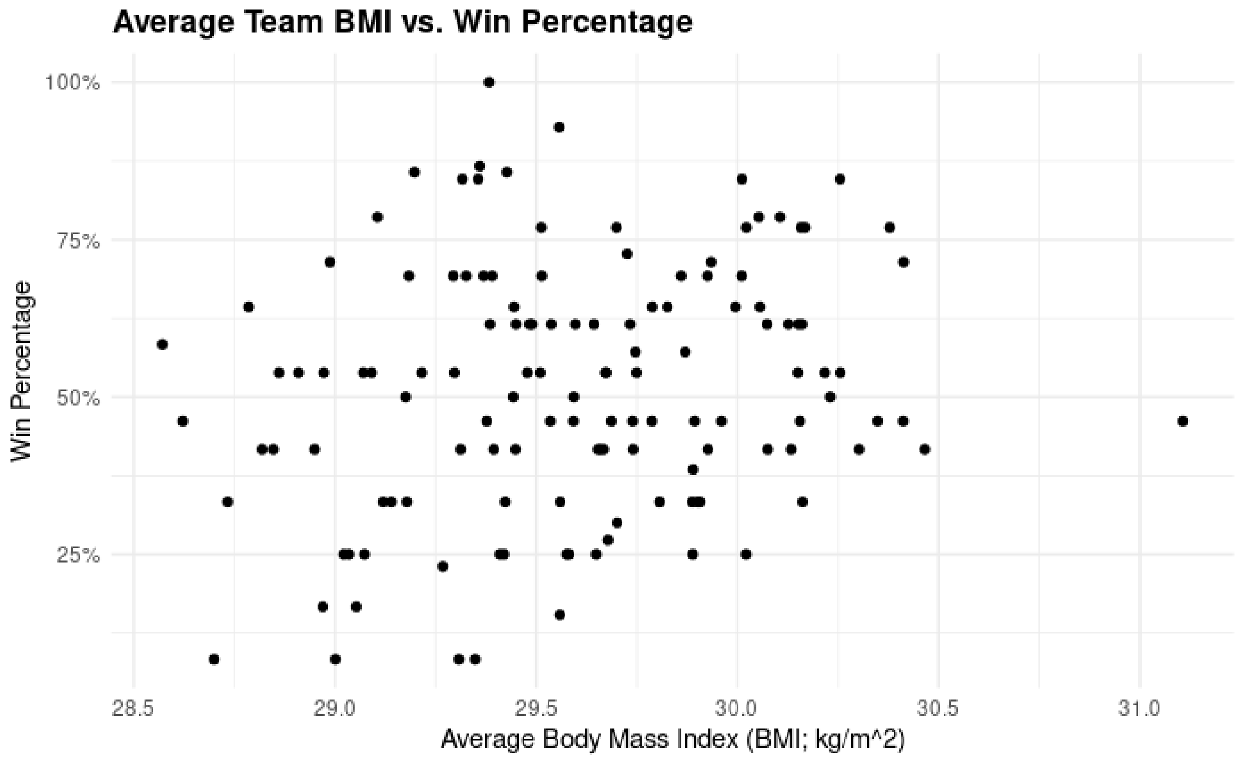


Figure 6: Average Team BMI vs. Win Percentage

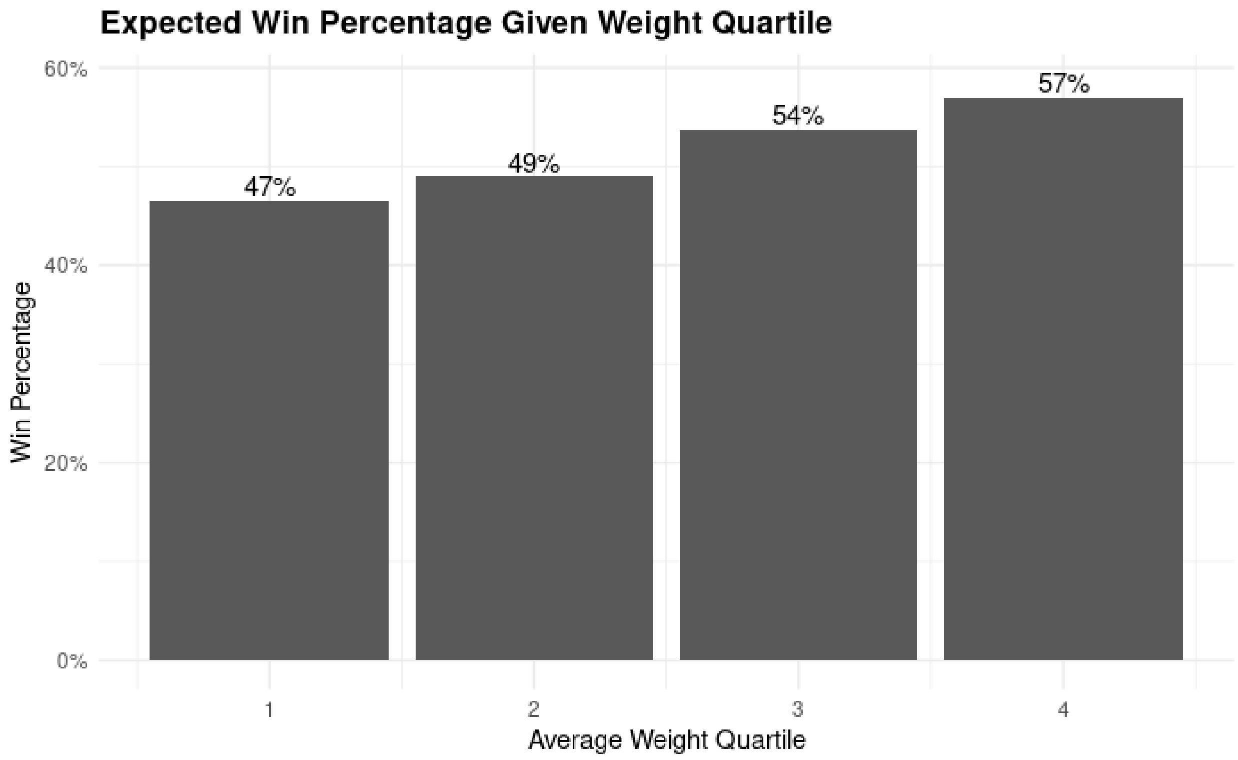


Figure 7: Expected Win Percentage Given Weight Quartile

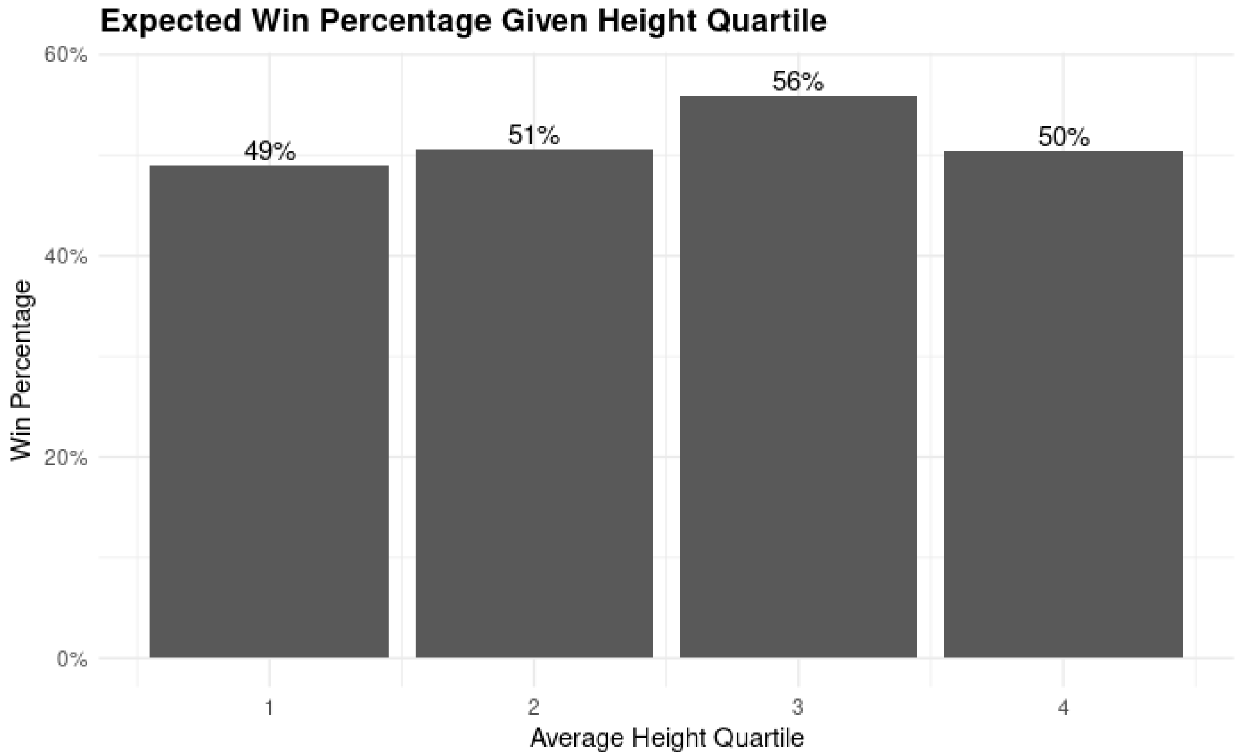


Figure 8: Expected Win Percentage Given Height Quartile

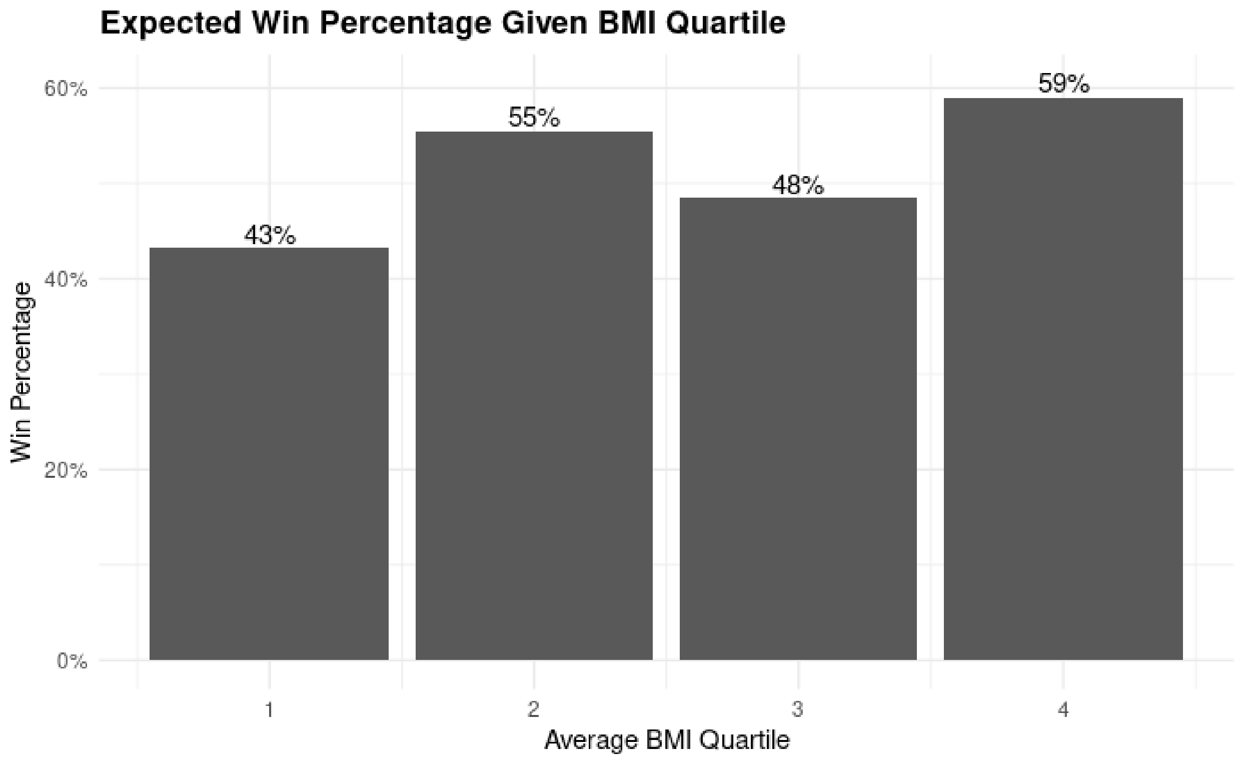


Figure 9: Expected Win Percentage Given BMI Quartile

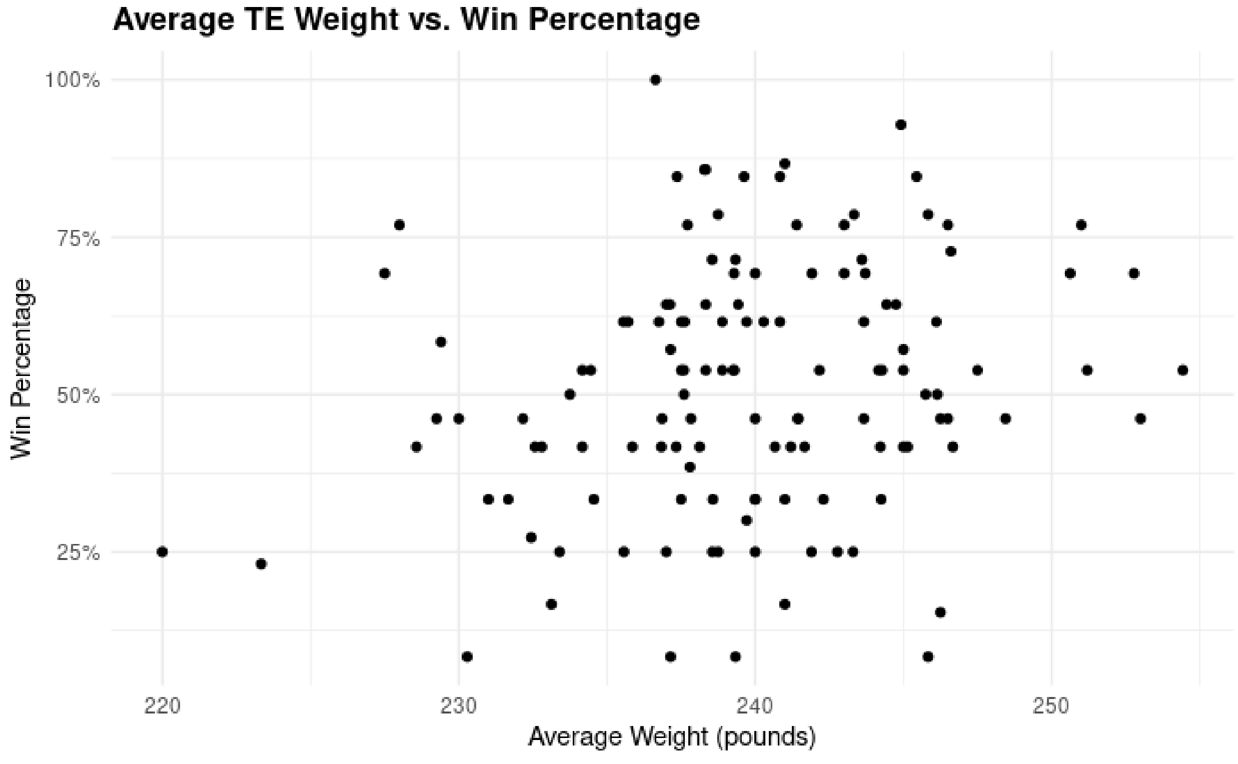


Figure 10: Average TE Weight vs. Win Percentage

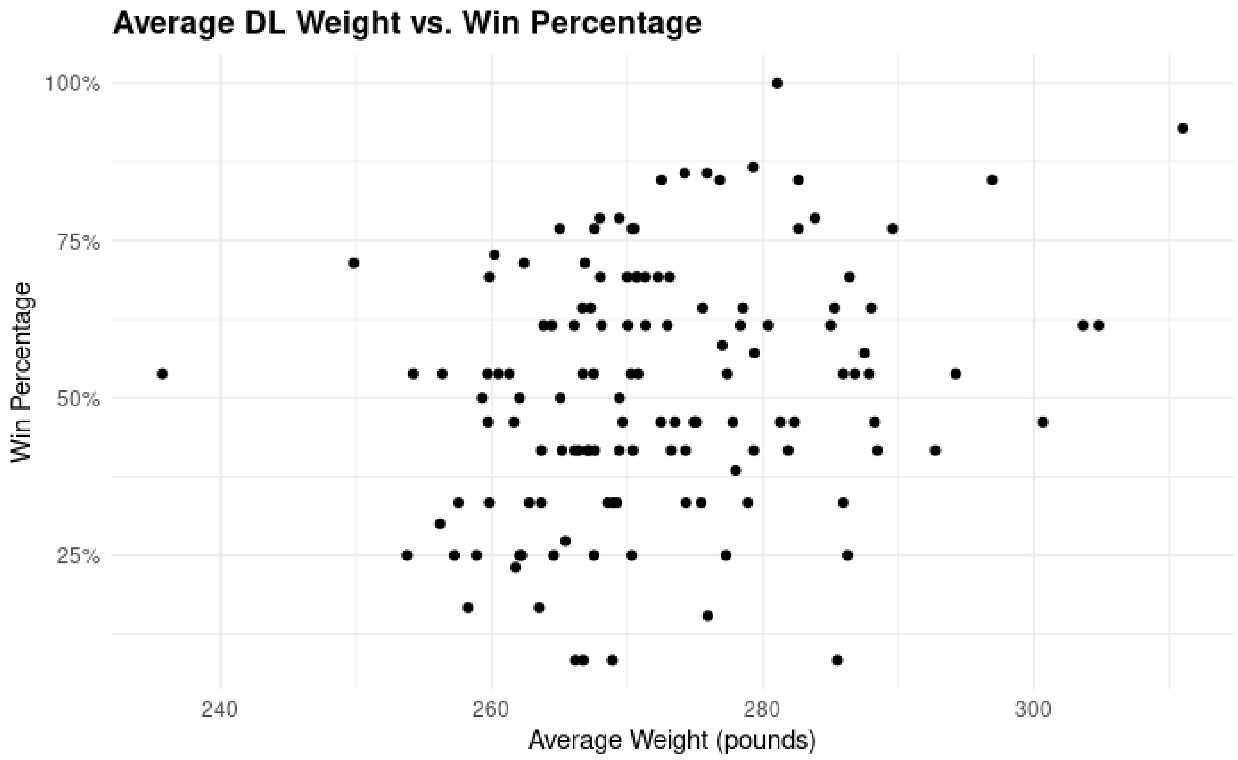


Figure 11: Average DL Weight vs. Win Percentage

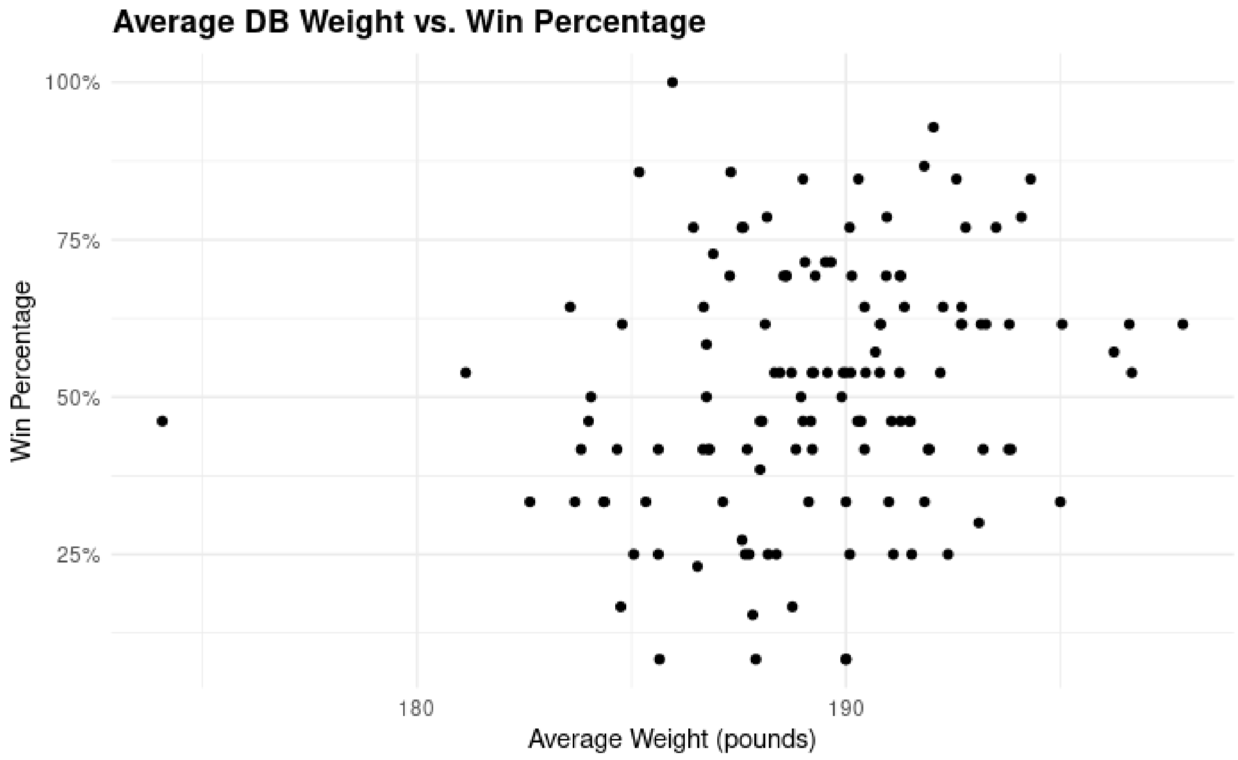


Figure 12: Average DB Weight vs. Win Percentage

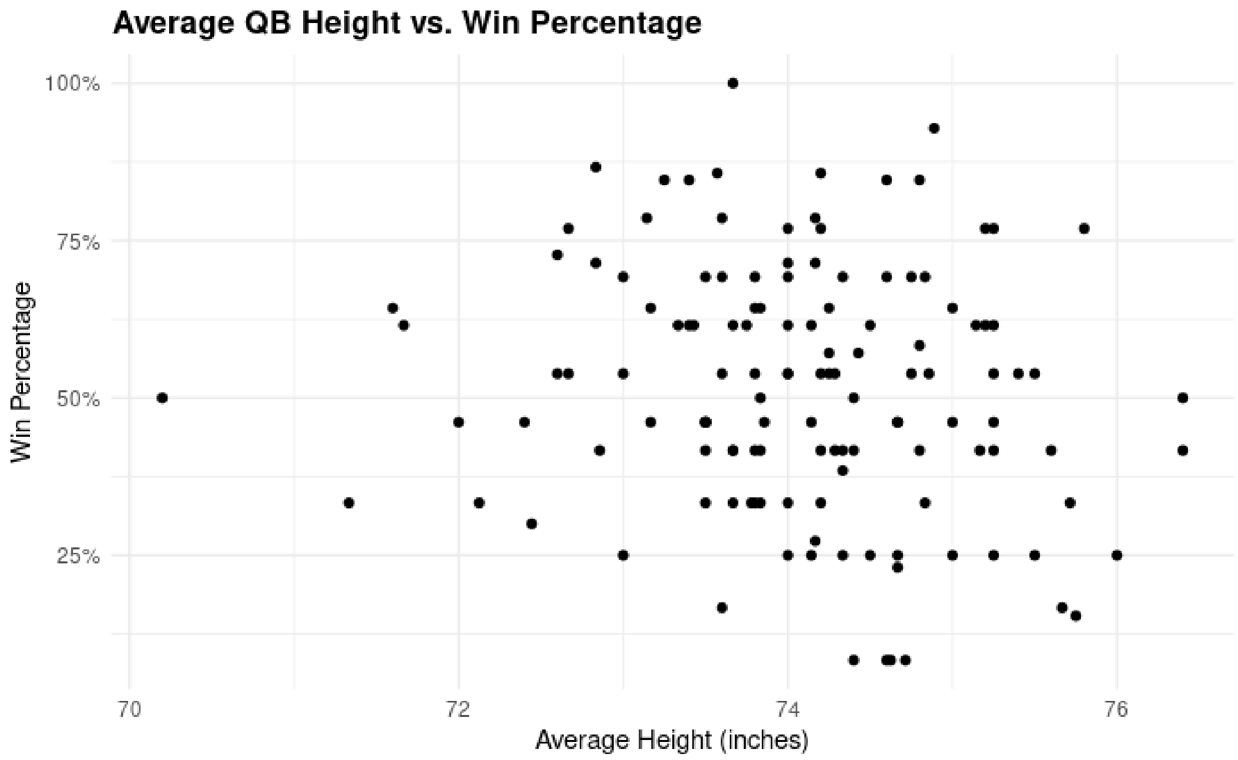


Figure 13: Average QB Height vs. Win Percentage

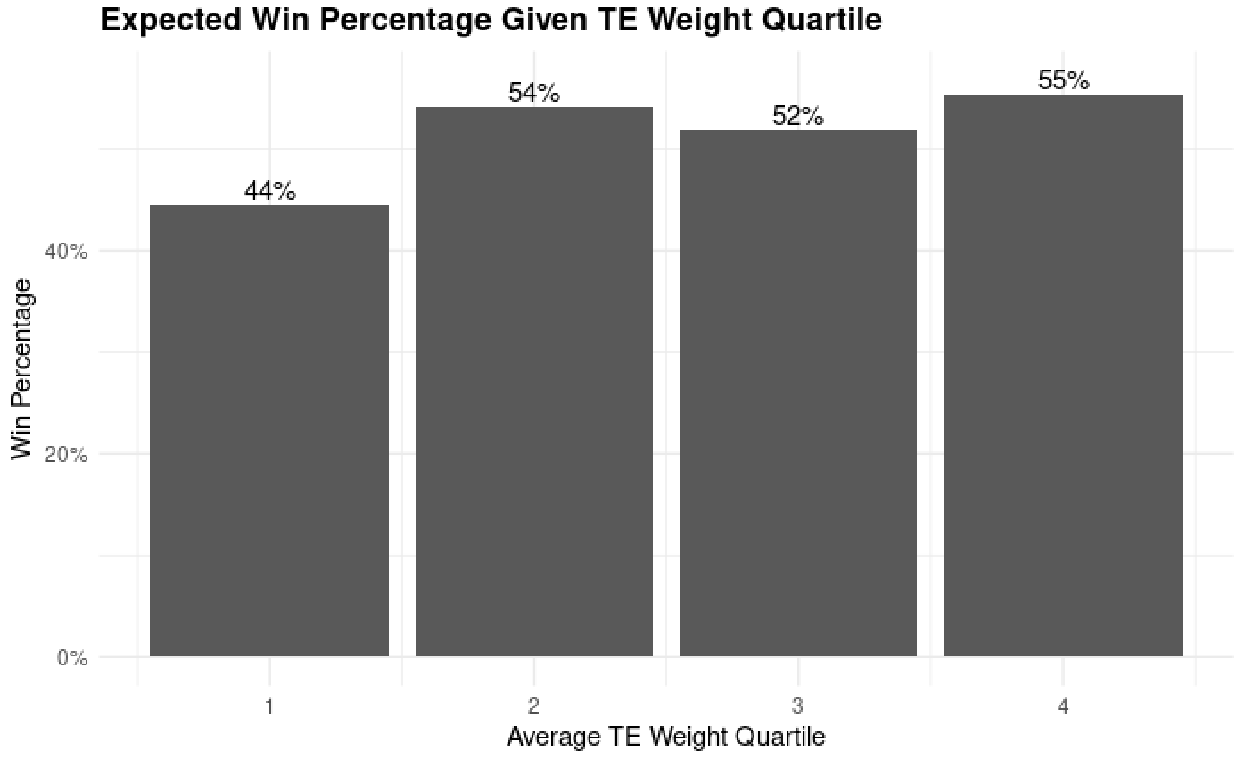


Figure 14: Expected Win Percentage Given TE Weight Quartile

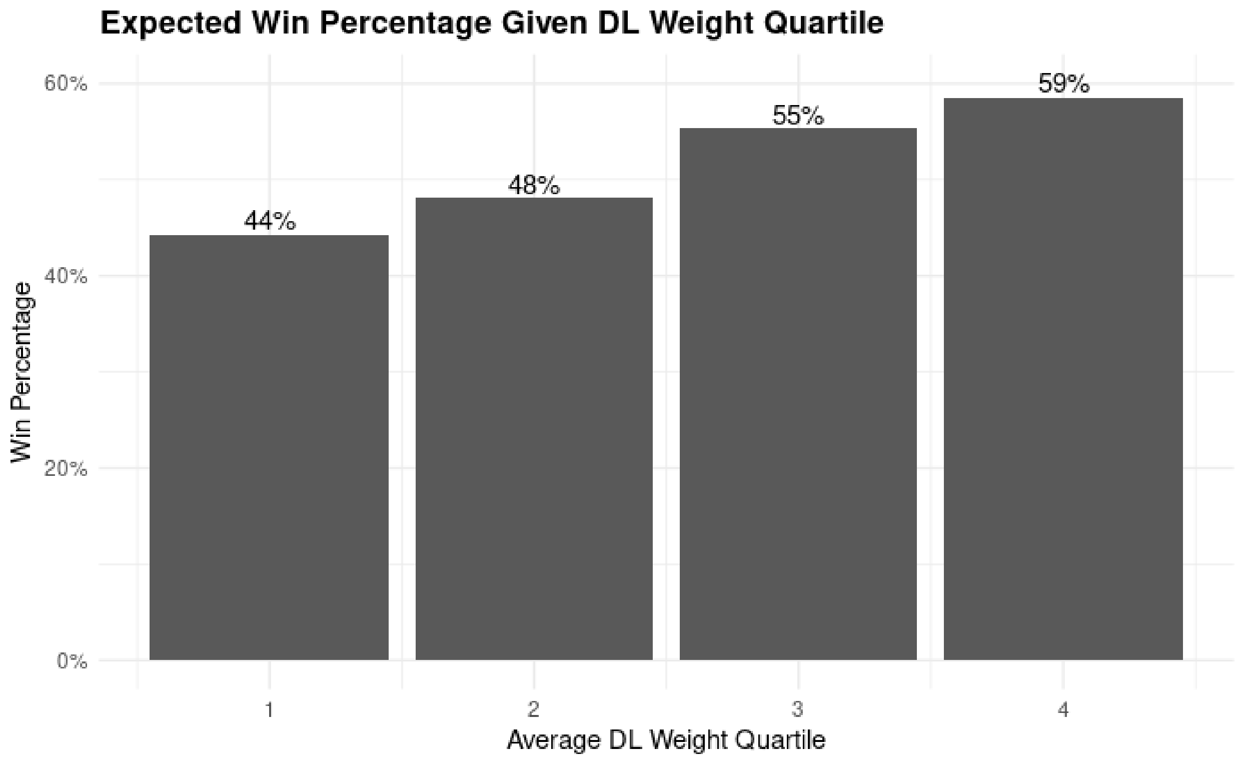


Figure 15: Expected Win Percentage Given DL Weight Quartile

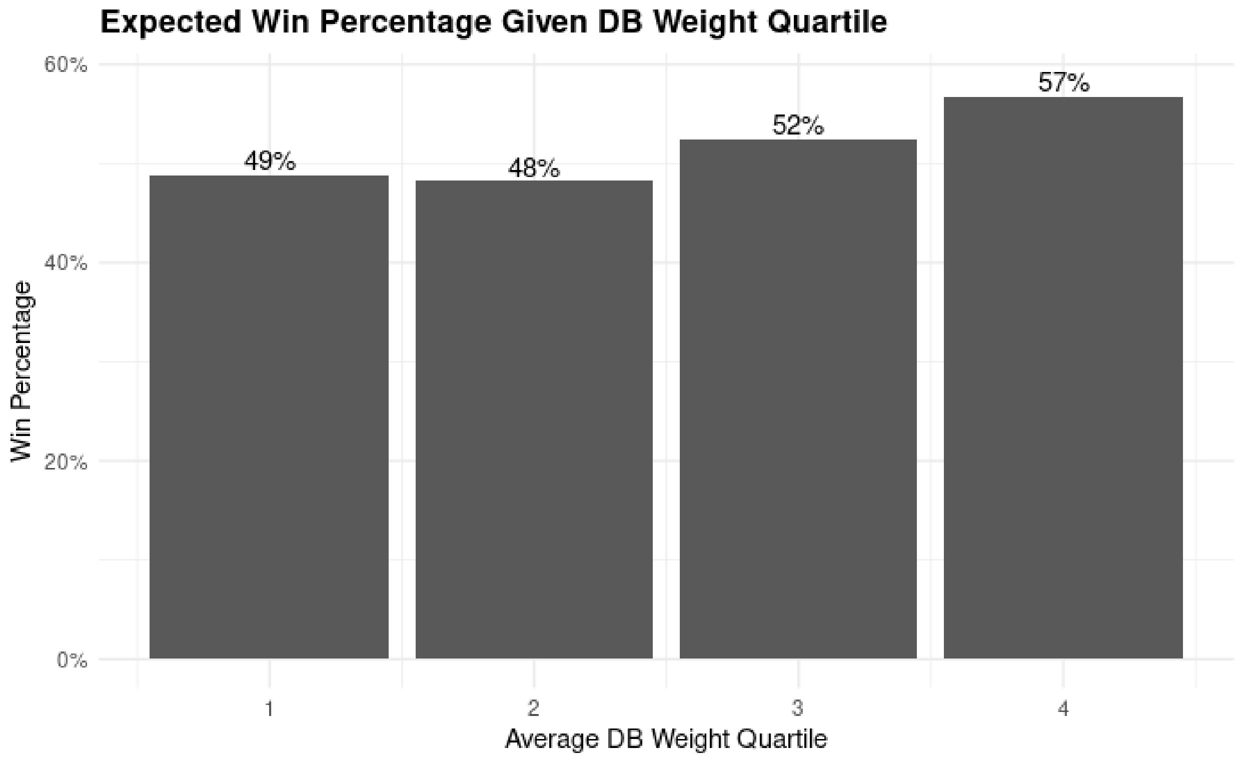


Figure 16: Expected Win Percentage Given DB Weight Quartile

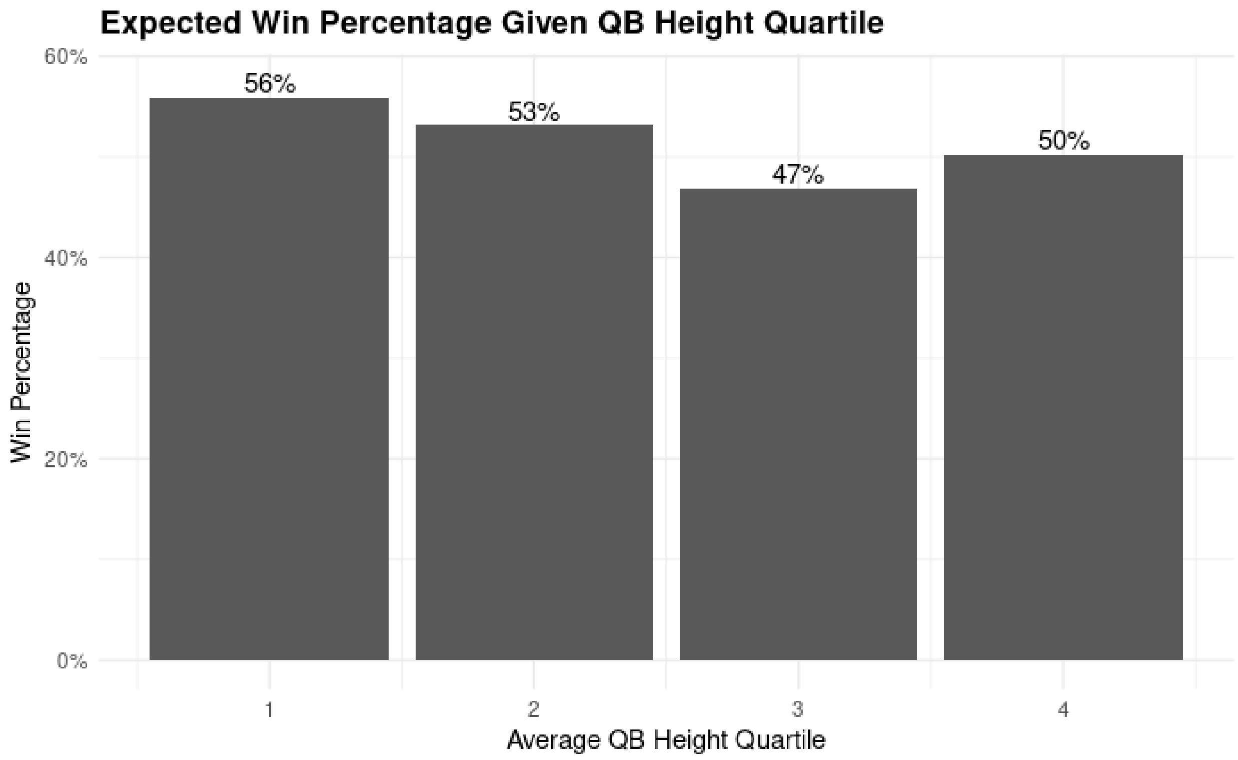
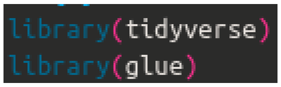


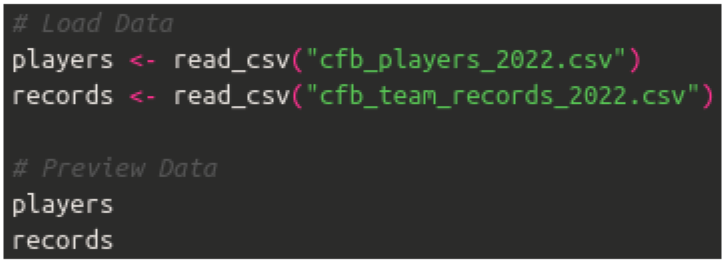
Figure 17: Expected Win Percentage Given QB Height Quartile

Selected Code Snippets



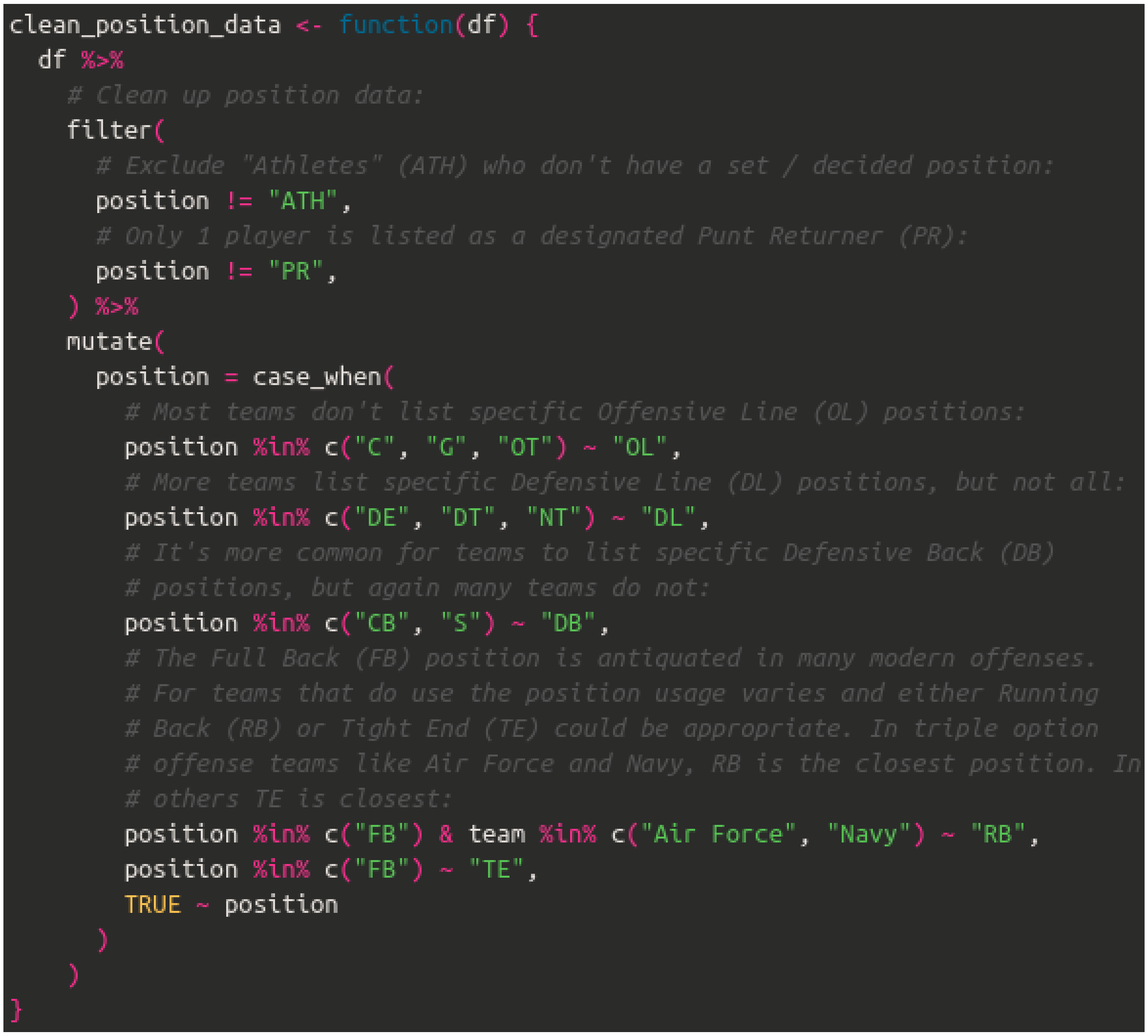
```
library(tidyverse)
library(glue)
```

Code Snippet 1: Helper Packages Used



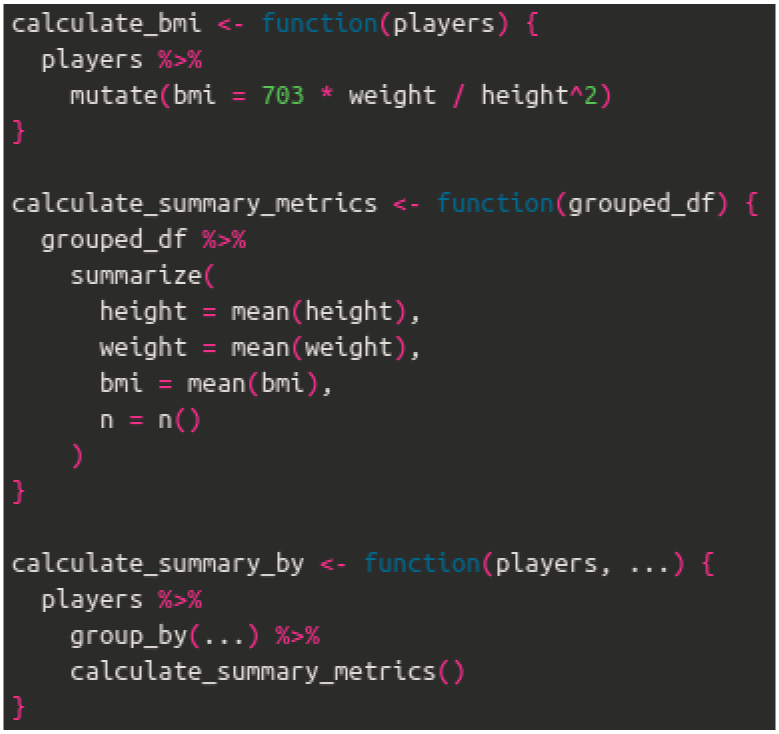
```
# Load Data
players <- read_csv("cfb_players_2022.csv")
records <- read_csv("cfb_team_records_2022.csv")

# Preview Data
players
records
```

Code Snippet 2: Loading Data



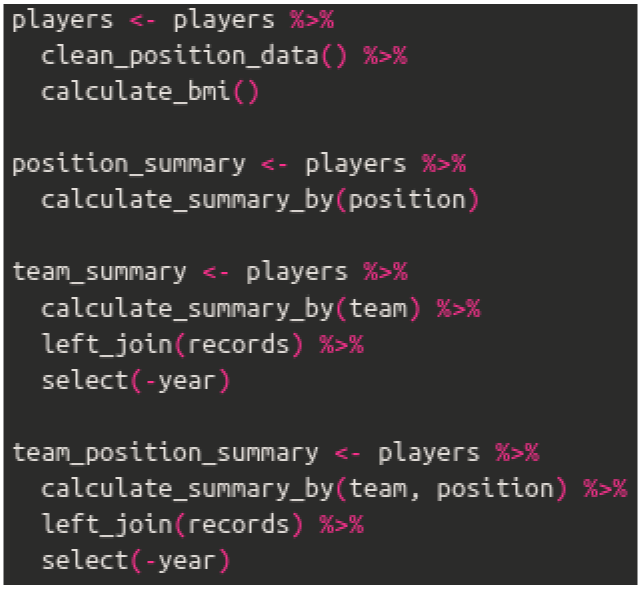
```
clean_position_data <- function(df) {
  df %>%
    # Clean up position data:
    filter(
      # Exclude "Athletes" (ATH) who don't have a set / decided position:
      position != "ATH",
      # Only 1 player is listed as a designated Punt Returner (PR):
      position != "PR",
    ) %>%
    mutate(
      position = case_when(
        # Most teams don't list specific Offensive Line (OL) positions:
        position %in% c("C", "G", "OT") ~ "OL",
        # More teams list specific Defensive Line (DL) positions, but not all:
        position %in% c("DE", "DT", "NT") ~ "DL",
        # It's more common for teams to list specific Defensive Back (DB)
        # positions, but again many teams do not:
        position %in% c("CB", "S") ~ "DB",
        # The Full Back (FB) position is antiquated in many modern offenses.
        # For teams that do use the position usage varies and either Running
        # Back (RB) or Tight End (TE) could be appropriate. In triple option
        # offense teams like Air Force and Navy, RB is the closest position. In
        # others TE is closest:
        position %in% c("FB") & team %in% c("Air Force", "Navy") ~ "RB",
        position %in% c("FB") ~ "TE",
        TRUE ~ position
      )
    )
}
```

Code Snippet 3: Function to Clean / Standardize Position Data



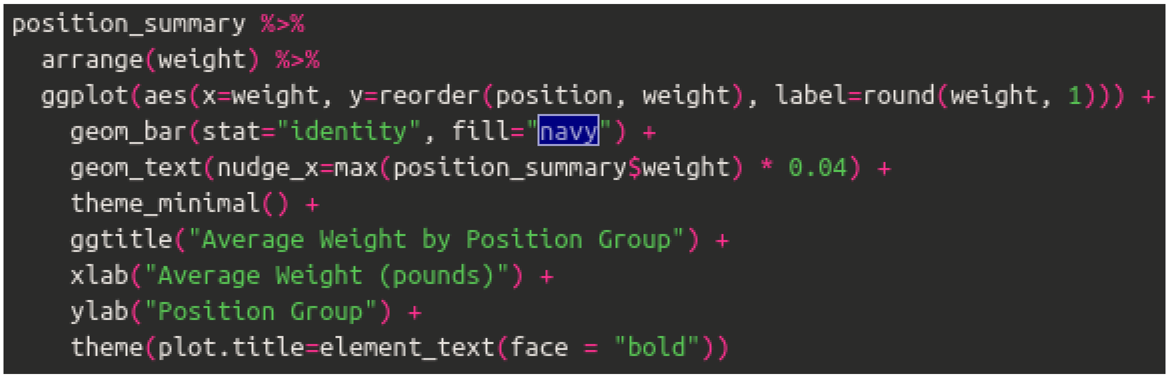
```
calculate_bmi <- function(players) {
  players %>%
    mutate(bmi = 703 * weight / height^2)
}

calculate_summary_metrics <- function(grouped_df) {
  grouped_df %>%
    summarize(
      height = mean(height),
      weight = mean(weight),
      bmi = mean(bmi),
      n = n()
    )
}

calculate_summary_by <- function(players, ...) {
  players %>%
    group_by(...) %>%
    calculate_summary_metrics()
}
```

Code Snippet 4: Other Helper Functions



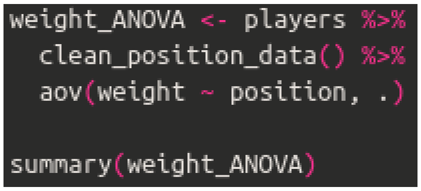
```
players <- players %>%
  clean_position_data() %>%
  calculate_bmi()

position_summary <- players %>%
  calculate_summary_by(position)

team_summary <- players %>%
  calculate_summary_by(team) %>%
  left_join(records) %>%
  select(-year)

team_position_summary <- players %>%
  calculate_summary_by(team, position) %>%
  left_join(records) %>%
  select(-year)
```

Code Snippet 5: Cleaning & Summarizing Data



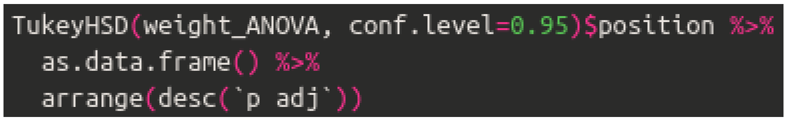
```r
position_summary %>%
  arrange(weight) %>%
  ggplot(aes(x=weight, y=reorder(position, weight), label=round(weight, 1))) +
    geom_bar(stat="identity", fill="navy") +
    geom_text(nudge_x=max(position_summary$weight) * 0.04) +
    theme_minimal() +
    ggtitle("Average Weight by Position Group") +
    xlab("Average Weight (pounds)") +
    ylab("Position Group") +
    theme(plot.title=element_text(face = "bold"))
```

Code Snippet 6: Plotting Position Size Comparison Example



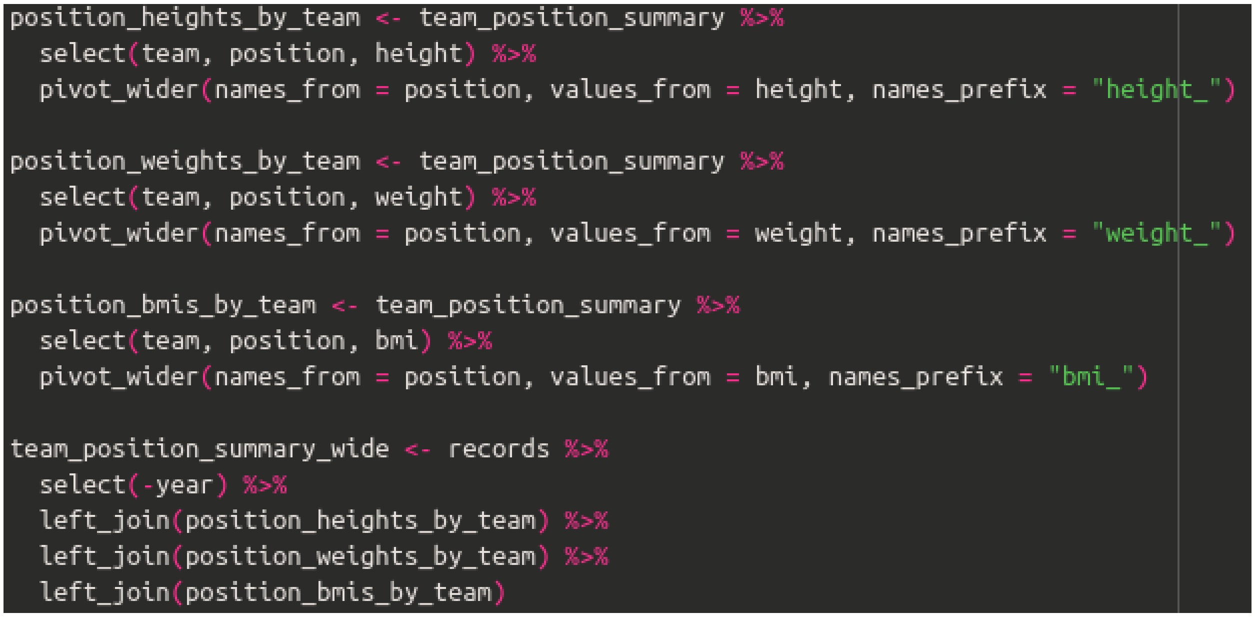
```r
weight_ANOVA <- players %>%
  clean_position_data() %>%
  aov(weight ~ position, .)

summary(weight_ANOVA)
```

Code Snippet 7: ANOVA Code Example



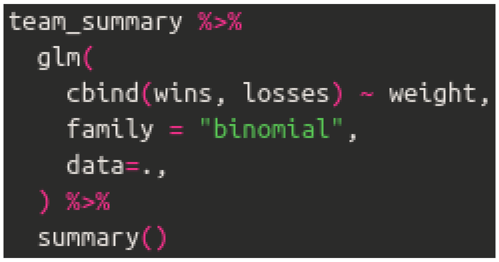
```r
TukeyHSD(weight_ANOVA, conf.level=0.95)$position %>%
  as.data.frame() %>%
  arrange(desc(`p adj`))
```

Code Snippet 8: Tukey Code Example



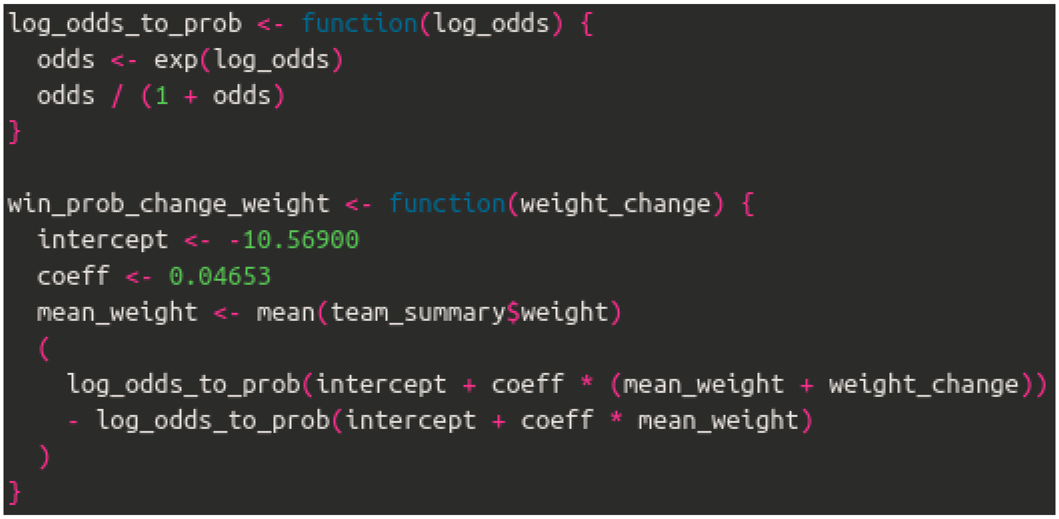
```r
position_heights_by_team <- team_position_summary %>%
  select(team, position, height) %>%
  pivot_wider(names_from = position, values_from = height, names_prefix = "height_")

position_weights_by_team <- team_position_summary %>%
  select(team, position, weight) %>%
  pivot_wider(names_from = position, values_from = weight, names_prefix = "weight_")

position_bmis_by_team <- team_position_summary %>%
  select(team, position, bmi) %>%
  pivot_wider(names_from = position, values_from = bmi, names_prefix = "bmi_")

team_position_summary_wide <- records %>%
  select(-year) %>%
  left_join(position_heights_by_team) %>%
  left_join(position_weights_by_team) %>%
  left_join(position_bmis_by_team)
```

Code Snippet 9: Formatting Data for Size by Position Level Regression



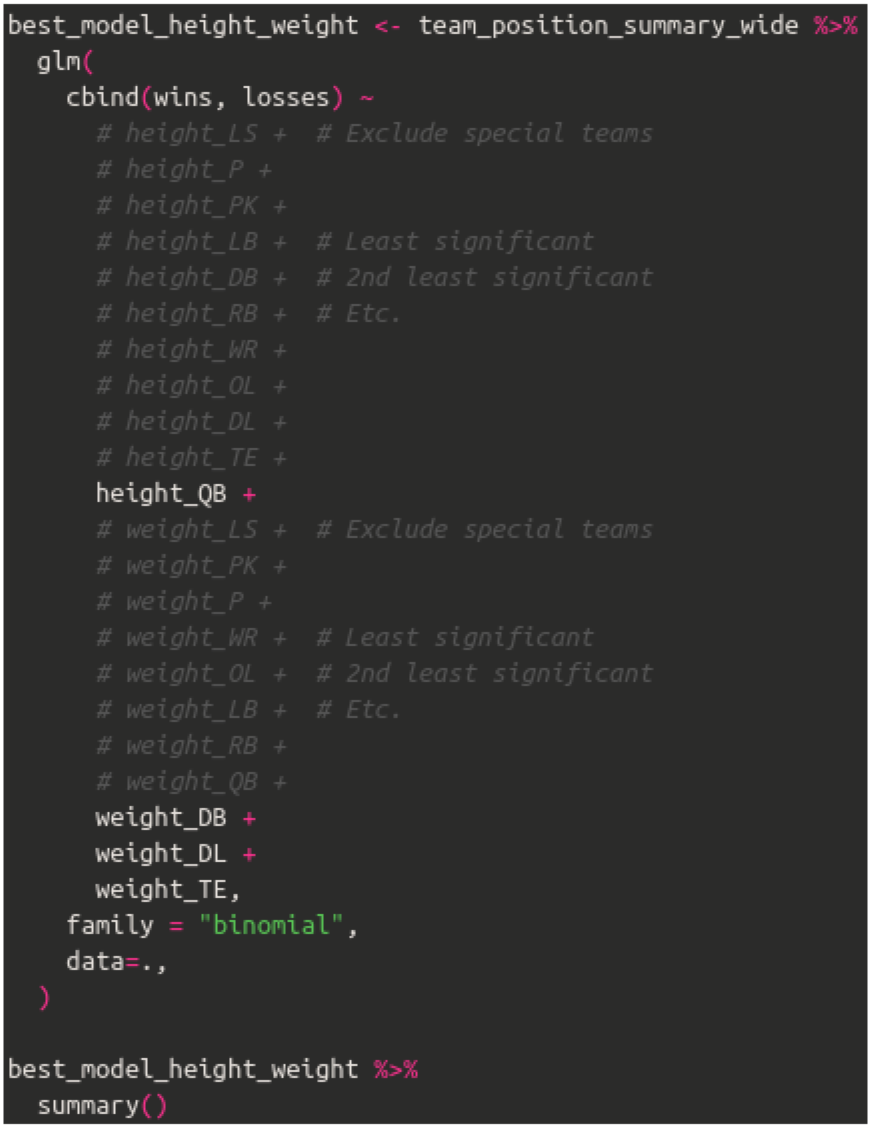
```r
team_summary %>%
  glm(
    cbind(wins, losses) ~ weight,
    family = "binomial",
    data=.,
  ) %>%
  summary()
```

Code Snippet 10: Univariate Binomial Regression



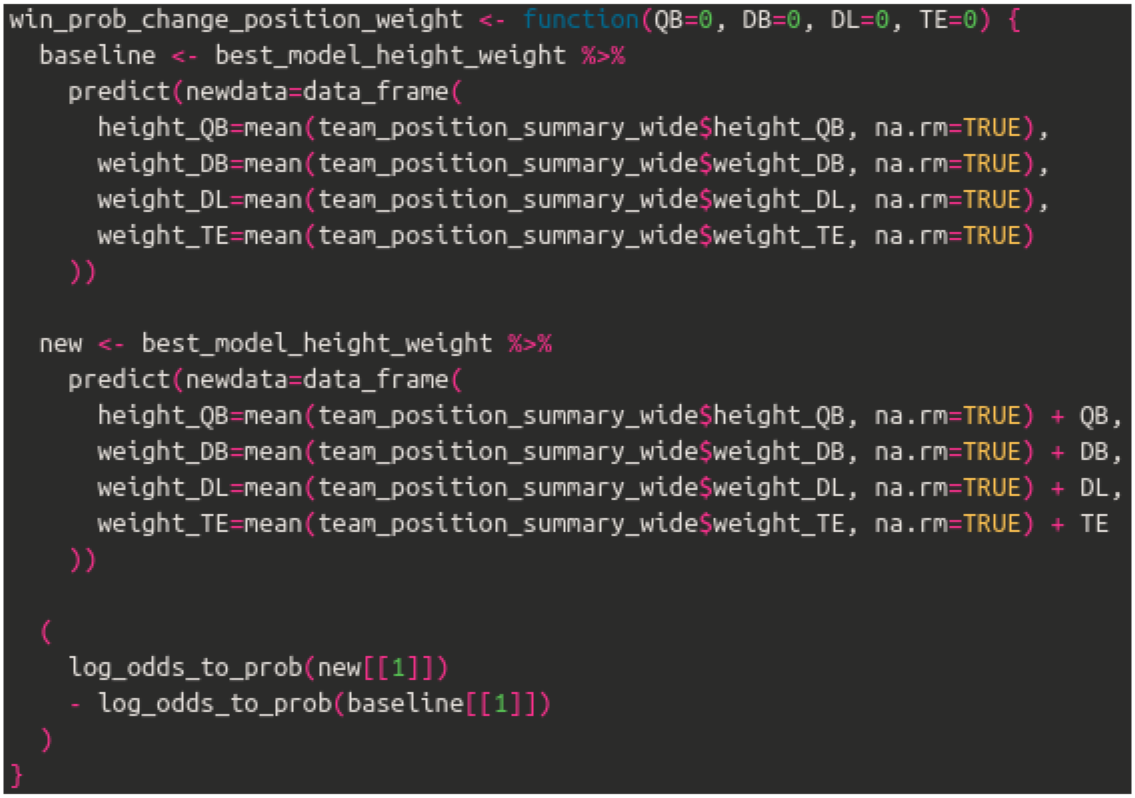
```r
log_odds_to_prob <- function(log_odds) {
  odds <- exp(log_odds)
  odds / (1 + odds)
}

win_prob_change_weight <- function(weight_change) {
  intercept <- -10.56900
  coeff <- 0.04653
  mean_weight <- mean(team_summary$weight)
  (
    log_odds_to_prob(intercept + coeff * (mean_weight + weight_change))
    - log_odds_to_prob(intercept + coeff * mean_weight)
  )
}
```

Code Snippet 11: Example Helpers to Assess the Impact of Size Changes in Univariate Model



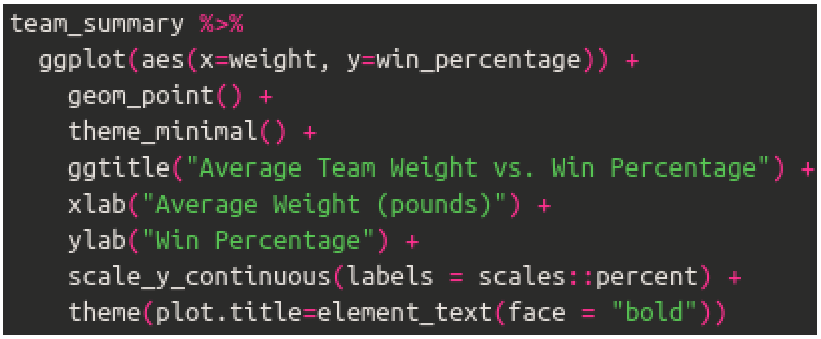
```
best_model_height_weight <- team_position_summary_wide %>%
  glm(
    cbind(wins, losses) ~
      # height_LS +  # Exclude special teams
      # height_P +
      # height_PK +
      # height_LB +  # Least significant
      # height_DB +  # 2nd least significant
      # height_RB +  # Etc.
      # height_WR +
      # height_OL +
      # height_DL +
      # height_TE +
      height_QB +
      # weight_LS +  # Exclude special teams
      # weight_PK +
      # weight_P +
      # weight_WR +  # Least significant
      # weight_OL +  # 2nd least significant
      # weight_LB +  # Etc.
      # weight_RB +
      # weight_QB +
      weight_DB +
      weight_DL +
      weight_TE,
    family = "binomial",
    data=.,
  )

best_model_height_weight %>%
  summary()
```

Code Snippet 12: Backwards Stepwise Position Level Binomial Regression



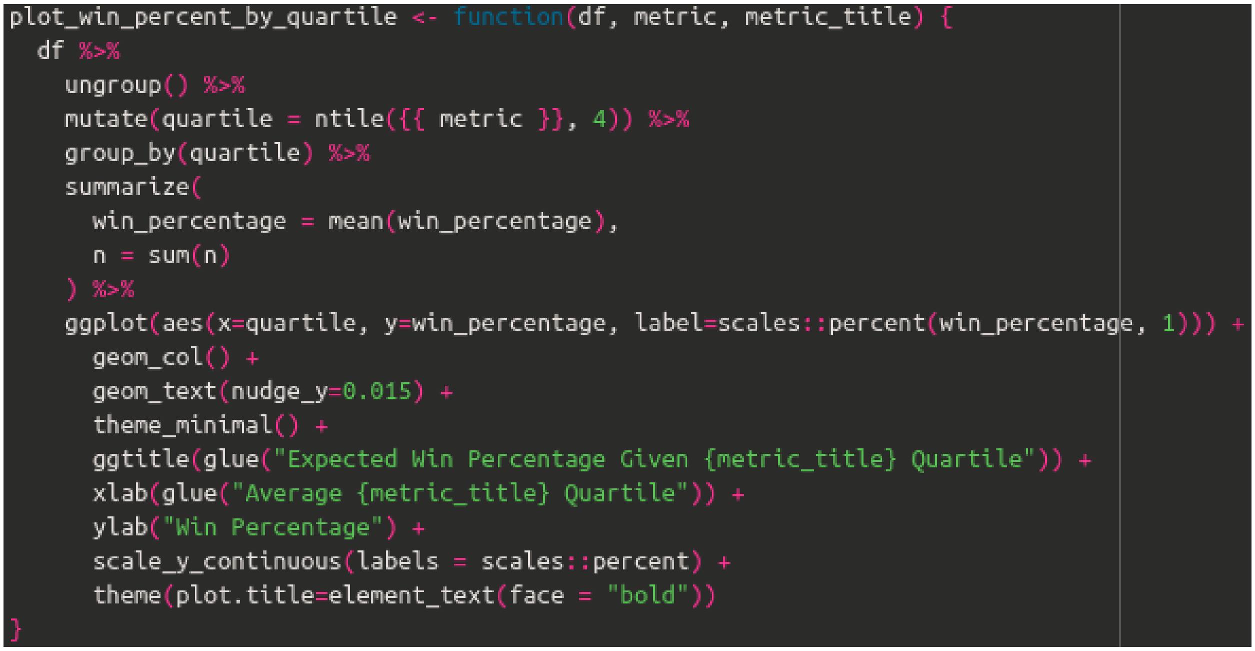
```r
win_prob_change_position_weight <- function(QB=0, DB=0, DL=0, TE=0) {
  baseline <- best_model_height_weight %>%
    predict(newdata=data_frame(
      height_QB=mean(team_position_summary_wide$height_QB, na.rm=TRUE),
      weight_DB=mean(team_position_summary_wide$weight_DB, na.rm=TRUE),
      weight_DL=mean(team_position_summary_wide$weight_DL, na.rm=TRUE),
      weight_TE=mean(team_position_summary_wide$weight_TE, na.rm=TRUE)
    ))

  new <- best_model_height_weight %>%
    predict(newdata=data_frame(
      height_QB=mean(team_position_summary_wide$height_QB, na.rm=TRUE) + QB,
      weight_DB=mean(team_position_summary_wide$weight_DB, na.rm=TRUE) + DB,
      weight_DL=mean(team_position_summary_wide$weight_DL, na.rm=TRUE) + DL,
      weight_TE=mean(team_position_summary_wide$weight_TE, na.rm=TRUE) + TE
    ))

  (
    log_odds_to_prob(new[[1]])
    - log_odds_to_prob(baseline[[1]])
  )
}
```

Code Snippet 13: Helper to Assess the Impact of Size Changes in Backwards Stepwise Binomial Regression Model



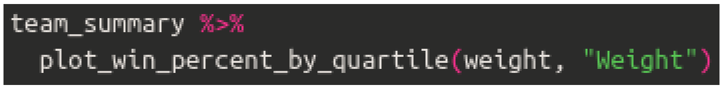
```r
team_summary %>%
  ggplot(aes(x=weight, y=win_percentage)) +
    geom_point() +
    theme_minimal() +
    ggtitle("Average Team Weight vs. Win Percentage") +
    xlab("Average Weight (pounds)") +
    ylab("Win Percentage") +
    scale_y_continuous(labels = scales::percent) +
    theme(plot.title=element_text(face = "bold"))
```

Code Snippet 14: Example Scatter Plot Code



```r
plot_win_percent_by_quartile <- function(df, metric, metric_title) {
  df %>%
    ungroup() %>%
    mutate(quartile = ntile({{ metric }}, 4)) %>%
    group_by(quartile) %>%
    summarize(
      win_percentage = mean(win_percentage),
      n = sum(n)
    ) %>%
    ggplot(aes(x=quartile, y=win_percentage, label=scales::percent(win_percentage, 1))) +
      geom_col() +
      geom_text(nudge_y=0.015) +
      theme_minimal() +
      ggtitle(glue("Expected Win Percentage Given {metric_title} Quartile")) +
      xlab(glue("Average {metric_title} Quartile")) +
      ylab("Win Percentage") +
      scale_y_continuous(labels = scales::percent) +
      theme(plot.title=element_text(face = "bold"))
}
```

Code Snippet 15: Function to Plot Conditional Win Percentages



```r
team_summary %>%
  plot_win_percent_by_quartile(weight, "Weight")
```

Code Snippet 16: Example Using Conditional Win Percentages Function

Bibliography and Credits

Data:

1. College Football Data (CFBD): https://collegefootballdata.com/
2. College Football Data API: https://api.collegefootballdata.com/api/docs/?url=/api-docs.json
3. cfbd-python library: https://github.com/CFBD/cfbd-python
4. BMI Calculation: https://www.registerednursern.com/bmi-calculation-formula-explained/

R:

1. Performing Tukey test: https://www.r-bloggers.com/2021/08/how-to-perform-tukey-hsd-test-in-r/
2. Binomial Regression in R:
   a. http://www.simonqueenborough.info/R/statistics/glm-binomial#:~:text=Group%2Dlevel%3A%20Proportion%20data
   b. https://cran.r-project.org/web/packages/ciTools/vignettes/ciTools-binomial-vignette.html#:~:text=Binomial%20regression%20in%20R