Predicting match outcomes for Newcastle United F.C.

Introduction:

Our study aims to make the best possible model to predict the outcome of Soccer matches for Newcastle United F.C. in the Premier League and other outside leagues. In soccer, there are three possible outcomes of a given match, a loss, a tie, and a win. There are 11 players on the field for each team at a time, each player contributing towards team statistics which our study will use to predict Newcastle's performance. If all the players on the team are performing well, we would assume that the team is more likely to win. We would also assume the converse that a team with bad-performing players will most likely lose. Based on this knowledge, we think that by using statistics from both Newcastle and their opponents, we can build a regression model to accurately predict the outcome of a match for a given team. In our case, we want to focus on the Premier League team Newcastle United F.C. and analyze their statistics to determine if there is a relationship between when they are performing well and if they win, lose, or draw. We expect the use of our selected model to accurately determine these outcomes.

Methods:

For our project, we will gather data from the Sports Reference website. Since we are dealing with sports statistics, the data is mainly unbiased. This includes Newcastle's team metrics, such as goals, assists, possession, and shots on target, while also considering the opponent's formation and the number of goals they scored. We plan to use web scraping techniques to extract statistics from the 2022-2023 and 2023-2024 seasons to ensure we have enough data for our model.

Data description:

The web-scraped data includes the result of each match—whether it was a win, draw, or loss, the day of the week the match took place as categorical, and the attendance, which indicates the number of people who attended the game. Additionally, the data specifies the venue, noting whether the match was played at home or away. Key performance metrics are also included, such as goals against (GA), possession percentage, shots on target (SoT), and total shots (Sh). Furthermore, it contains tactical details like Newcastle's formation and that of their opponents, giving an expansive view of match dynamics and team performance.

Our project may have potential biases due to its limited scope, as it relies just on data from two recent seasons. This limited timeframe may not accurately represent long-term trends in soccer. The sport evolves over time, with changes in tactics, rules, and player performance, making it difficult to generalize patterns to future seasons. Furthermore, factors such as unbalanced competition and varying sample sizes for different formations could also impact the results. Lastly, the predictors were chosen based on years of soccer knowledge and experience.

Analysis:

For this section of the paper, we want to illustrate some of the most important aspects of soccer that determine game outcomes. From the various plots that can do this, we chose numerous plots for this study that we think best demonstrate these aspects. Before that happens, Some necessary information needs to be provided; soccer has a lot of variation in how it is played, so removing outliers from the data seems unnecessary. It is very realistic to see twenty shots on target in one game and only two in the next; this is also true with many other variables.

Plots:

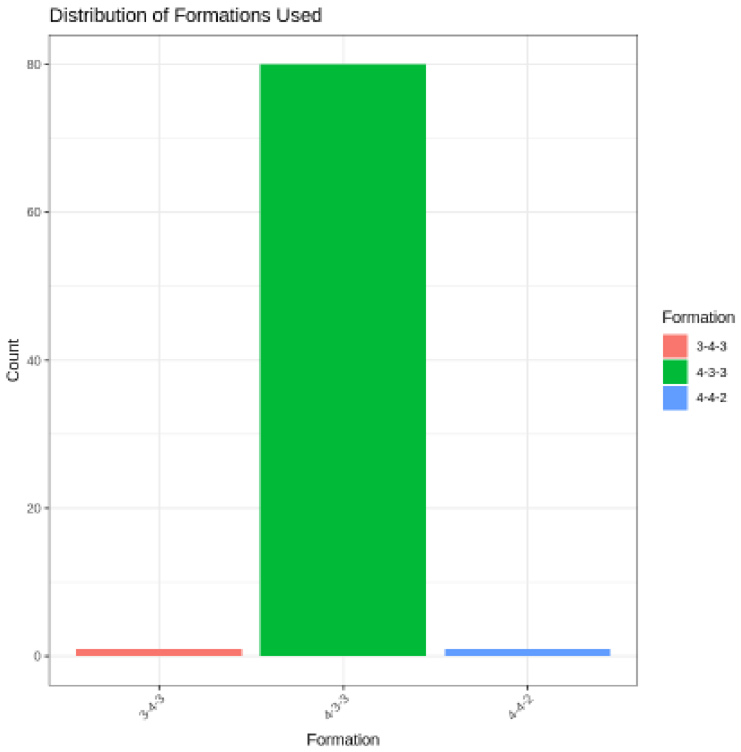


Figure 1 - Distribution of Newcastle Formations Used

To start considering formations, and to reconsider whether Newcastle's formation is a useful variable in the dataset, we decided to look at their formations. This plot is a total count of all the formations they ran within the two seasons our data consists of. We see that they use a total of three formations but heavily favor the 4-3-3. Knowing what formation we use and how that can contribute to our success as a team is good.

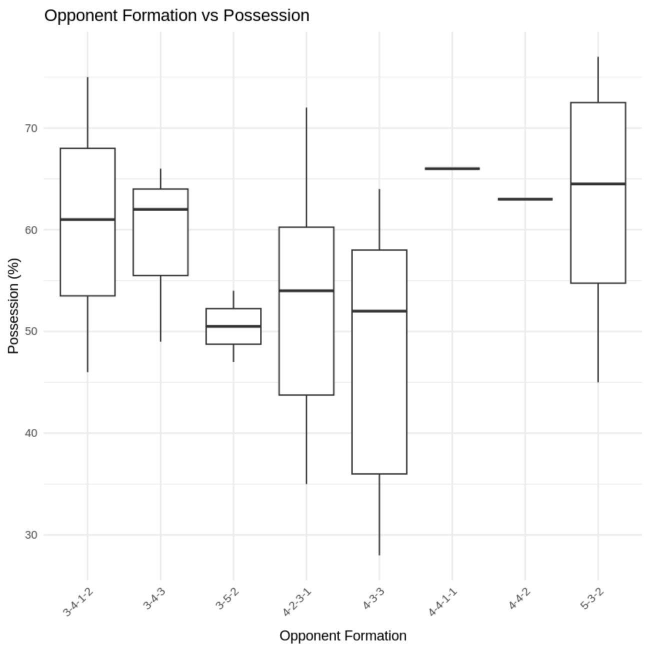


Figure 2 - Opponent's formation vs Newcastle's possession

This graph illustrates Newcastle's possession compared to various formations faced. They struggle in the 4-3-3 mirror matchup but perform well against most other formations and preform better against others.

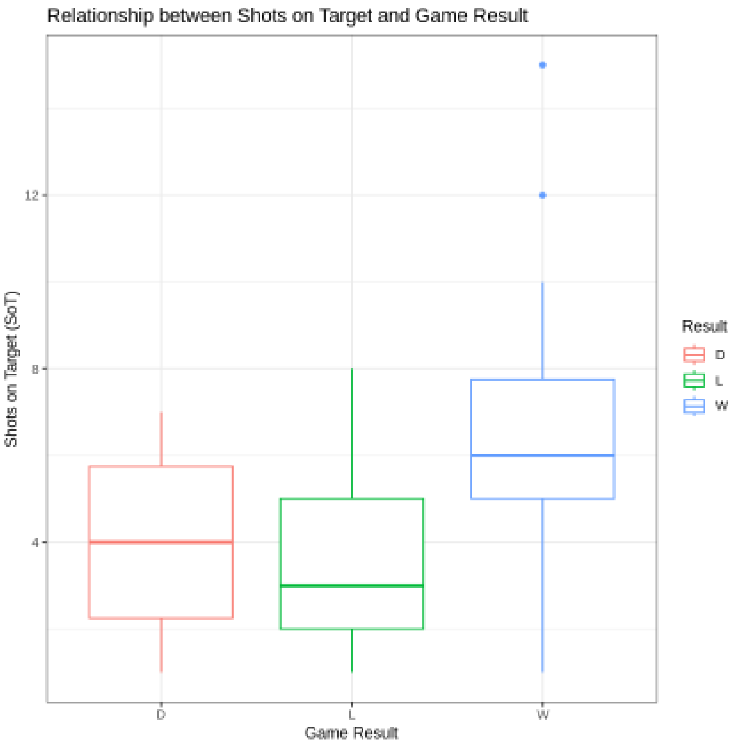


Figure 3 - Relationship of Shots on Target and Game Results

Shots on target are one of the most important statistics in soccer. They are used heavily to calculate the expected goals statistics seen almost everywhere in soccer these days. Knowing this, we wanted to see the relationship between the distribution of shots on target and each of our results. Looking at the plot, we can see that a higher number of shots on target leads to better results, as seen when comparing each box plot. The W box is higher than all the other plots, and this means that typically when Newcastle wins, they have more shots on target. The L box is right below then the D box.

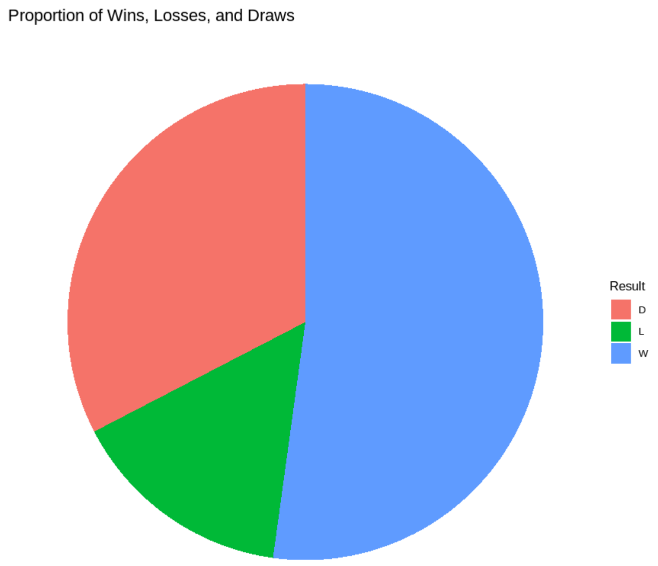


Figure 4 - distribution of match results in 2022-2024 seasons

The pie chart shows the distribution of match outcomes (wins, losses, and draws) for Newcastle United during the 2022-2023 and 2023-2024 seasons. Each slice of the pie represents a specific outcome, and the size of each slice indicates its relative frequency. The chart shows that wins (blue) are the most common result, followed by draws (red), while losses (green) are the least frequent. This means the team performed well, achieving many victories compared to losses and draws. Overall, this visualization effectively shows the team's performance distribution and how good they are.

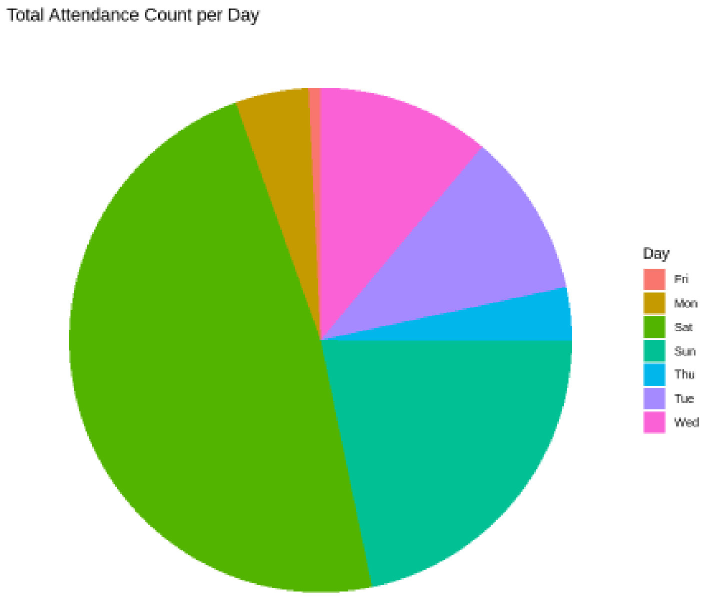


Figure 5 - Distribution of Attendance per Day

Soccer is played on many different days, and each match has a certain number of fans attending; we wanted to see which day typically has the most fans attending the match. Based on this pie chart, Saturday has the highest number of fans attending. This could be for several reasons, most likely because it is a weekend and the Premiere league plays Sat-Mon, so the fans are close in proximity and have the time to attend. As it turns out, attendance and day will be important variables to note when predicting results.

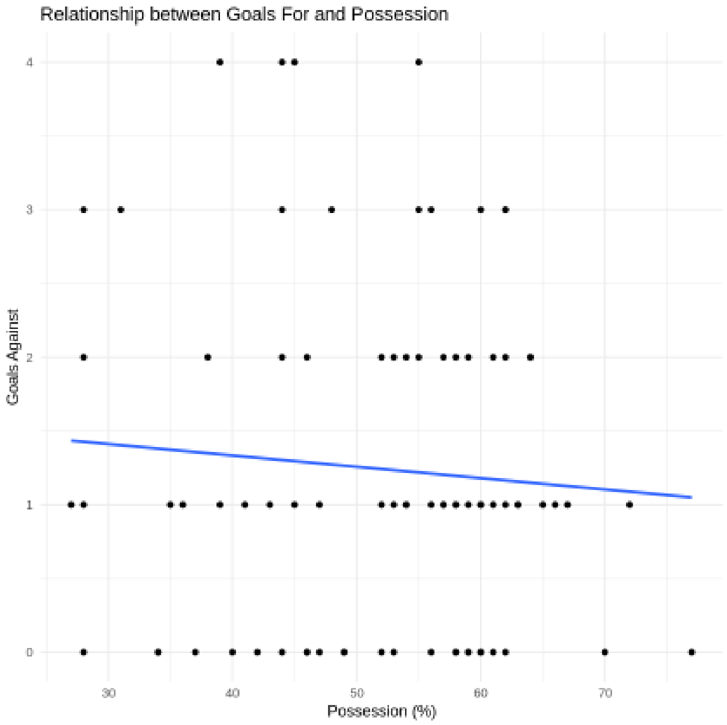


Figure 6 - Relationship of Goals Against and Newcastle Possession

For the final plot, we analyzed the relationship between Newcastle's possession percentages and the goals scored against them. This analysis is essential because, in soccer, a higher rate of possession typically correlates with fewer goals conceded. Our plot supports this notion; the regression plot created with ggplot indicates a negative relationship between possession and goals against. One interesting observation from the graph is that when Newcastle's possession is between 50% and 70%, we expect them to be winning and allowing fewer goals. However, the data shows in some observations that Newcastle had high possession percentages and many goals conceded in those games. This suggests the team may need better defense and their opponents effectively counterattacks. As Newcastle's possession percentage increased, the number of goals scored against them decreased. While we can only draw definitive conclusions by conducting a full regression analysis, this visual evidence suggests that possession should be included in the final model.

Model:

For this study, the model used is Multinomial Logistic Regression. This is because our response variable Result is Categorical with three distinct outcomes: Win, Loss, and Draw. All three of these outcomes need to be considered with individual probabilities, unlike Logistic Regression, where one could categorize a Loss and Draw together as not Winning. This is because, unlike many other sports, a match can end as a draw in Soccer. With this being said, the final model being used is the following:

Result ~ Day + Attendance + Venue + SoT + Sh + Possession + Opp_Formation

The model combines numerical and categorical predictors that contribute to the results of Newcastle United F.C. matches.

When working with Multinomial Regression, the probabilities are calculated along the same lines as Logistic Regression. Referring to (**Formula 1.**), the probability of each result is calculated using the same link function as Logistic Regression. Still, each outcome has to be calculated separately and needs to sum to one. This is important to note because when using R, the default coding output is treatment coding, For this study, R defaulted to Draw as the reference. To calculate the probability of drawing, sum up the Probability of Winning and Losing, then subtract from one (**Equation 1.**). After calculating each of the probabilities, One can interpret the result as follows: Given the entered match statistics for Newcastle United, The probabilities of the three results are a W% chance of winning, an L% chance of losing, and a D% chance of drawing said match.

With the Results model constructed, we want to get a general sense of probabilities of match results based on minimum, maximum, and average match statistics for Newcastle United. To do this, we made separate data frames composed of match statistics based on the summary statistics for each numerical predictor within the model (**Figure 6.**). To find the average, min, and max values for the categorical predictors, We used the table function in R to make a two-way table for each predictor with the result. This returns a count of each category corresponding to each result, we then noted how many observations were taken for each result and created the data frames from there. Since some categories had only one observation, we set a baseline of roughly greater than six observations to be put into the minimum data frame. We also created a home and away data frame for each because finding min, max, and average with only two categories is hard. After making the data frames, using the predict function in R, The predicted probabilities of the min, max, and average match statistics are:

Data Frame	**Predicted Probabilities: D, L, W**
Minimum, Home	.21% 1.01% 98.78%
Minimum, Away	5.31% 6.66% 88.03%
Maximum, Home	69.92% 1.28% 28.81%
Maximum, Away	98.08% .47% 1.44%
Average, Home	7.92% 4.23% 87.85%
Average, Away	65.05% 9.17% 25.77%

Table 1. Minimum, Maximum, Average, Home and Away Result Probabilities

The above table contains the predicted probabilities of each data frame we created. Some patterns observed from these probabilities are that Newcastle United F.C. has a significant advantage when playing at home. For all of the home data frames, the probabilities of winning are all above 80%; this shows that when playing at home, Newcastle has a very good chance of winning, no matter their performance. Another thing to note is that Newcastle performs well overall, home or away, they are going to end the game scoring points toward the league table. This is because the graphs of the distribution of results showed Newcastle winning for a higher proportion of the seasons from which the data comes. However, these data frames are not exact and are important to understand when interpreting these probabilities. There are a lot of factors that this study didn't take into account that could change these probabilities and some of the statistics don’t fully align with real-world examples. Although this is true, this is a good place to start formulating questions and answers for predicting match results for New Castle United F.C. using Multinomial Logistic Regression.

Future Plans and Conclusion:

This study aimed to predict the outcomes of Newcastle United F.C. matches through the use of Multinomial Logistic Regression. The model performed pretty well on the data that was provided and captured some notable patterns. Although this was the case we would need to check diagnostics for the model and restructure some of the questions being answered to fully apply it to the real world. Going forward, the Website also provides data for individual players, such as the number of shots, goals, possession, shots on target, etc. As such, we could consider individual player statistics in the model or determine which player’s performance best predicts the team’s success. We could also try predicting the current 2024-2025 season using our model, either improved by individual statistics or not. We could also extend our collection range to the past 2022 to obtain patterns that encapsulate a greater range other than the two seasons this study works with.

Appendix:

A.1 Loading packages

```r
# Load necessary libraries
library(rvest)
library(dplyr)
library(xml2)
library(ggplot2)
```

A.2 Web scraping and making one comprehensive table (2022-2023 season)

```r
#data from 2022-2023 season
# Define the URL
url <-
"https://fbref.com/en/squads/b2b47a98/2022-2023/matchlogs/all_comps/schedule/Newcastle-United-Scores-and-Fixtures-All-Competitions"
#reading
page <- read_html(url)
#extracting table
table <- page %>% html_node("table") %>% html_table(fill = TRUE)
# Replace 'Date', 'Time', etc., with the exact column names that you observe from the output
table <- table %>%
 rename(
   Date = `Date`,
   Time = `Time`,
   Comp = `Comp`,
   Round = `Round`,
   Day = `Day`,
   Venue = `Venue`,
   Result = `Result`,
   GF = `GF`,
   GA = `GA`,
   xG = `xG`,
   xGA = `xGA`,
   Possession = `Poss`,
   Attendance = `Attendance`,
   Captain = `Captain`,
   Formation = `Formation`,
   Opp_Formation = `Opp Formation`,
   Referee = `Referee`
 )
# Convert columns to appropriate types
table <- table %>%
 mutate(
   Date = as.Date(Date, format = "%Y-%m-%d"),
   Time = as.character(Time),
   Comp = as.character(Comp),
   Round = as.character(Round),
   Day = as.character(Day),
   Venue = as.character(Venue),
   Result = as.character(Result),
   GF = as.integer(GF),
   GA = as.integer(GA),
   xG = as.numeric(xG),
   xGA = as.numeric(xGA),
   Possession = as.integer(Possession),
   Attendance = as.integer(gsub(",", "", Attendance)), # Remove commas for integer conversion
   Captain = as.character(Captain),
   Formation = as.character(Formation),
   Opp_Formation = as.character(Opp_Formation),
   Referee = as.character(Referee)
 )

 #now looking getting the data from the same season but about shootings of games
url_standard <-
"https://fbref.com/en/squads/b2b47a98/2022-2023/matchlogs/all_comps/shooting/Newcastle-United-Match-Logs-All-Competitions"
page_standard <- read_html(url_standard)
data_standard <- page_standard %>% html_node("table") %>% html_table(fill = TRUE)

# Use the first row as column names
colnames(data_standard) <- data_standard[1, ]
data_standard <- data_standard[-1, ]
colnames(data_standard) <- trimws(colnames(data_standard))
# Convert columns
data_standard <- data_standard %>%
 mutate(
   Date = as.Date(Date, format = "%Y-%m-%d"),
   Gls = as.integer(Gls),
   Sh = as.integer(Sh),
   SoT = as.integer(SoT),
   `SoT%` = as.numeric(`SoT%`),
   `G/Sh` = as.numeric(`G/Sh`),
   `G/SoT` = as.numeric(`G/SoT`),
   Dist = as.numeric(Dist),
   FK = as.integer(FK),
   PK = as.integer(PK),
   PKatt = as.integer(PKatt)
 )

 merged_table <- table %>%
 inner_join(data_standard, by = "Date")

#dropping coulmns that are identical and unimportant columns
merged_table <- merged_table %>%
 select(-Time.y, -Comp.y, -Round.y, -Day.y, -Venue.y, -Result.y, -GF.y,
-GA.y, -Opponent.y, -`Match Report.y`,-xG.y, -Notes)

# Rename columns by removing the .y
s22_23_table <- merged_table %>%
 rename(
   Time = Time.x,
   Comp = Comp.x,
   Round = Round.x,
   Day = Day.x,
   Venue = Venue.x,
   Result = Result.x,
   GF = GF.x,
   GA = GA.x,
   Opponent = Opponent.x,
   `Match Report` = `Match Report.x`,
   xG = xG.x
 )
head(s22_23_table)
```

A.3 Exploring (charts and graphs)

```r
ggplot(combined_table, aes(x = Formation, fill = Formation)) +
 geom_bar() +  # Use geom_bar() for bar plot
 labs(title = "Distribution of Formations Used",
      x = "Formation",
      y = "Count") +
 theme_bw() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
ggplot(s22_23_table, aes(x = `Opp_Formation`, y = Possession)) +
 geom_boxplot() +
 labs(
   title = "Opponent Formation vs Possession",
   x = "Opponent Formation",
   y = "Possession (%)"
 ) +
 theme_minimal() +
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```r
ggplot(combined_table, aes(x = Result, y = SoT, color = Result)) +
 geom_boxplot() +
 labs(title = "Relationship between Shots on Target and Game Result",
      x = "Game Result",
      y = "Shots on Target (SoT)") +
 theme_bw() +
 scale_fill_manual(values = c("Win" = "darkgreen", "Loss" = "firebrick",
"Tie" = "steelblue"))
```

```r
# Aggregate attendance by day
attendance_by_day <- combined_table %>%
 group_by(Day) %>%
 summarize(TotalAttendance = sum(Attendance))

# Create the pie chart
ggplot(attendance_by_day, aes(x = "", y = TotalAttendance, fill = Day)) +
 geom_bar(stat = "identity", width = 1) +
 coord_polar("y", start = 0) +
 labs(title = "Total Attendance Count per Day", fill = "Day") +
 theme_void()
```

```
ggplot(combined_table, aes(x = Possession, y = GA)) +
 geom_point() +  # Scatter plot
 geom_smooth(method = "lm", se = FALSE) +  # Add a linear regression line
 labs(title = "Relationship between Goals For and Possession",
      x = "Possession (%)",
      y = "Goals Against") +
 theme_minimal()
```

A.4 Model Implementation

```
model_data$Result <- as.factor(model_data$Result)

multinom_model <- multinom(Result ~ Day + Attendance + Venue + GA +
Possession + SoT + Sh + Opp_Formation,data = model_data)

summary(multinom_model)
```

A.5 Prediction Data Frames and Predict Function

```
min_data_home <- data.frame(Day = 'Mon', Attendance =10419 , Venue =
'Home', GA = 0, Possession = 27 , SoT = 1, Sh = 3 , Opp_Formation =
'3-4-3')

min_data_away <- data.frame(Day = 'Mon', Attendance = 10419, Venue =
'Away', GA = 0, Possession = 27 , SoT = 1, Sh = 3, Opp_Formation =
'3-4-3')

max_data_home <- data.frame(Day = 'Sat', Attendance = 81365 , Venue =
'Home', GA = 4 , Possession = 77, SoT = 15, Sh = 27, Opp_Formation =
'4-2-3-1')

max_data_away <- data.frame(Day = 'Sat', Attendance = 81365, Venue =
'Away', GA = 4, Possession = 77, SoT = 15, Sh = 27, Opp_Formation =
'4-2-3-1')
```

```r
avg_data_home <- data.frame(Day = 'Sun', Attendance = 46981, Venue =
'Home', GA = 1, Possession = 52 , SoT = 5 , Sh = 14, Opp_Formation =
'4-2-3-1')

avg_data_away <- data.frame(Day = 'Sun', Attendance = 46981, Venue =
'Away', GA = 1, Possession = 52 , SoT = 5 , Sh = 14, Opp_Formation =
'4-2-3-1')
```

```r
predicted_avg_home <- predict(multinom_model, newdata = avg_data_home,
type = 'probs', se.fit = TRUE)

predicted_avg_home
```

Formula 1.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq}$$

$$p_i = e^{\eta_i} / 1 + e^{\eta_i}$$

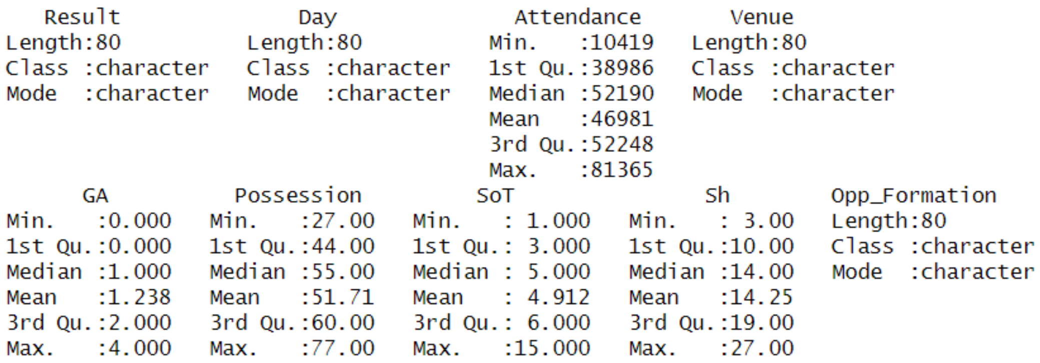
Where;

$$1 = p_W + p_L + P_D$$

Equation 1.

$$p_D = 1 - (P_w + P_L)$$

Figure 6.



```
    Result               Day              Attendance        Venue          
 Length:80          Length:80          Min.   :10419   Length:80         
 Class :character   Class :character   1st Qu.:38986   Class :character  
 Mode  :character   Mode  :character   Median :52190   Mode  :character  
                                       Mean   :46981                     
                                       3rd Qu.:52248                     
                                       Max.   :81365                     
       GA          Possession        SoT               Sh        Opp_Formation     
 Min.   :0.000   Min.   :27.00   Min.   : 1.000   Min.   : 3.00   Length:80         
 1st Qu.:0.000   1st Qu.:44.00   1st Qu.: 3.000   1st Qu.:10.00   Class :character  
 Median :1.000   Median :55.00   Median : 5.000   Median :14.00   Mode  :character  
 Mean   :1.238   Mean   :51.71   Mean   : 4.912   Mean   :14.25                     
 3rd Qu.:2.000   3rd Qu.:60.00   3rd Qu.: 6.000   3rd Qu.:19.00                     
 Max.   :4.000   Max.   :77.00   Max.   :15.000   Max.   :27.00                     
```