2020-2021 English Premier League Season

Jeannie Hinton

The Data

- Wanted a large dataset with match statistics that I believed could contribute to the outcome of a match
- Originally, a betting dataset on Kaggle
- Eliminated betting variables I did not find necessary
- Started with 106 variables, refined it to 24 variables, and finally ended with 39 variables
- Started with 380 rows with each row representing a match, refined it to 297 matches by filtering out draws
- Chose to filter out draws as I thought it would negatively impact the logistic regression I planned to do
- Transition from Covid Season

```
## $ Date     <chr> "12/09/2020", "12/09/2020", "12/09/202
## $ Time     <time> 12:30:00, 15:00:00, 17:30:00, 20:00:0
## $ HomeTeam <chr> "Fulham", "Crystal Palace", "Liverpool
## $ AwayTeam <chr> "Arsenal", "Southampton", "Leeds", "Ne
## $ FTHG     <dbl> 0, 1, 4, 0, 0, 0, 1, 0, 5, 4, 1, 2, 2,
## $ FTAG     <dbl> 3, 0, 3, 2, 3, 1, 3, 2, 2, 3, 3, 1, 5,
## $ FTR      <chr> "A", "H", "H", "A", "A", "A", "A", "A"
## $ HTHG     <dbl> 0, 1, 3, 0, 0, 0, 0, 0, 2, 2, 0, 1, 1,
## $ HTAG     <dbl> 1, 0, 2, 0, 0, 0, 1, 2, 1, 1, 1, 1, 1,
## $ HTR      <chr> "A", "H", "H", "D", "D", "D", "A", "A"
## $ Referee  <chr> "C Kavanagh", "J Moss", "M Oliver", "S
## $ HS       <dbl> 5, 5, 22, 15, 7, 9, 13, 9, 17, 10, 13,
## $ AS       <dbl> 13, 9, 6, 15, 13, 15, 10, 11, 6, 14, 1
## $ HST      <dbl> 2, 3, 6, 3, 1, 5, 3, 2, 7, 7, 4, 3, 7,
## $ AST      <dbl> 6, 5, 3, 2, 7, 4, 5, 4, 4, 6, 5, 3, 6,
```

Predictors of Interest: HalfTimeWinner, HST, AST, Fouls, Yellow & Red Cards, Crowds, PostInt

Research Question: Who will win a match, the home club or the away club, based on match predictors?

Data Wrangling

```
library(tidyverse)
library(lubridate)
library(tidymodels)
library(caret)
```

- Working with the filtered dataset, had to create several new variables using the mutate function
- Some of the new variables include **Winner, MoreRedCards, MoreYellowCards, More Fouls,** and even **HomeShotsOnTarget**
- **Winner** was needed because the original dataset only had home club and away club names
  - Needed a response variable that evaluated to a binary outcome of home or away club for the prediction model I was planning
- **MoreRedCards, MoreYellowCards,** and **MoreFouls** were needed because they were more helpful in the visualizations than the original variables
  - After working on the project proposal, I didn't need the very precise amount of fouls or cards for each match and who received them, instead needed the amount of matches when one club received more red or yellow cards than the other
- Even though HST existed, for purposes of others who are not football fans, making a new variable was needed to more quickly work with and explain the data

```
refined_epl <- refined_epl %>%
  mutate(Winner = case_when(FTHG > FTAG & FTHG != FTAG ~ "Home",
                            FTAG > FTHG & FTAG != FTHG ~ "Away",
                            FTHG == FTAG ~ "Draw"
                           )) %>%
```

Data Wrangling continued…

- New **Crowds** variable
- Required outside research using the EPL's website and watching film
- Fans were allowed to return to stadiums at reduced capacity for the last two weeks of the season.
- Even though I use lubridate later in the project, when I first created the **Crowds** variable during the proposal, I had to work with the mutate function in a very cumbersome way

```
mutate(Crowds = case_when(HomeTeam == "Man United" & AwayTeam == "Fulham" & Referee == "L Mason" ~ "Fans present",
                          HomeTeam == "Southampton" & AwayTeam == "Leeds" & Referee == "P Bankes" ~ "Fans present",
                          HomeTeam == "Brighton" & AwayTeam == "Man City" & Referee == "S Attwell" ~ "Fans present",
                          HomeTeam == "Chelsea" & AwayTeam == "Leicester" & Referee == "M Dean" ~ "Fans present",
                          HomeTeam == "Everton" & AwayTeam == "Wolves" & Referee == "A Madley" ~ "Fans present",
                          HomeTeam == "Newcastle" & AwayTeam == "Sheffield United" & Referee == "R Jones" ~ "Fans present",
                          HomeTeam == "Tottenham" & AwayTeam == "Aston Villa" & Referee == "C Pawson" ~ "Fans present",
                          HomeTeam == "Crystal Palace" & AwayTeam == "Arsenal" & Referee == "A Taylor" ~ "Fans present",
```

Data Wrangling continued…

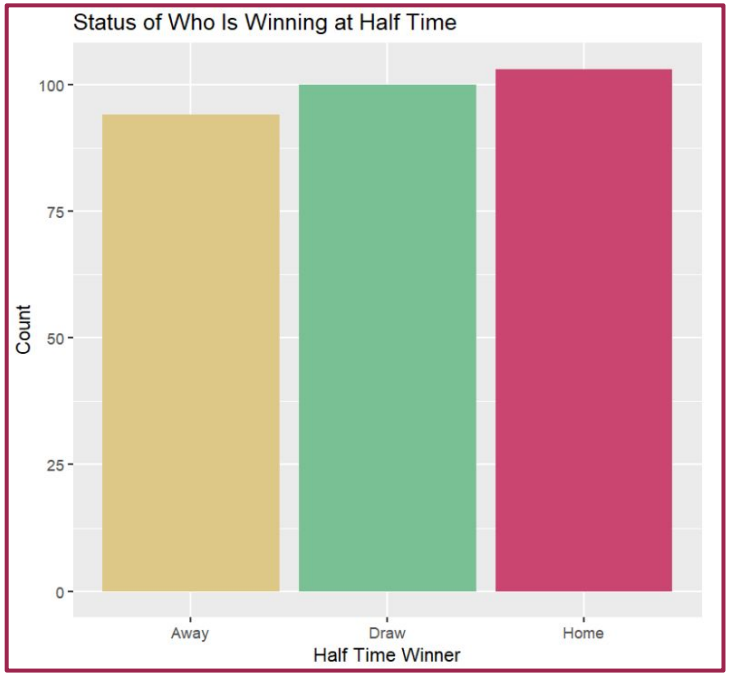
- Created new date variable called **NewDate**
  - Old date was a character -> new date is a date data type using lubridate
- Used **NewDate** variable to create another new variable called **PostInt**
  - In order to determine if a match fell within a week of the international break, used mutate and the **NewDate** variable to determine if it was post an international break or not

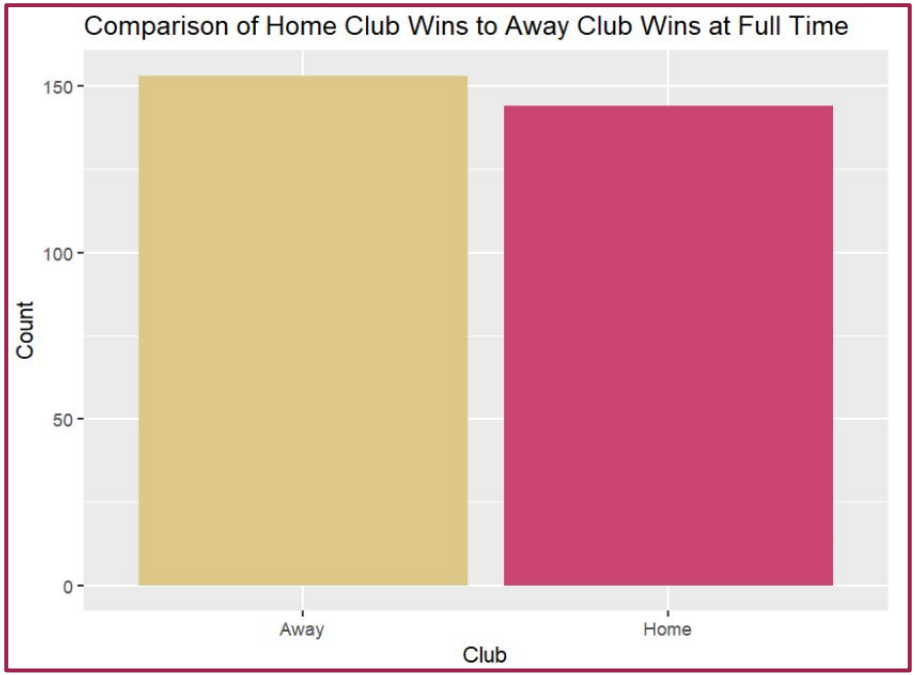
```
```{r date work for international breaks, echo = FALSE}
refined_epl$NewDate <- dmy(refined_epl$Date)

post_aug_break <- interval(ymd("2020-9-9"), ymd("2020-9-16"))
post_oct_break <- interval(ymd("2020-10-14"), ymd("2020-10-21"))
post_nov_break <- interval(ymd("2020-11-17"), ymd("2020-11-24"))
post_jan_break <- interval(ymd("2021-2-2"), ymd("2021-2-9"))
post_mar_break <- interval(ymd("2021-3-30"), ymd("2021-4-6"))

refined_epl <- refined_epl %>%
  mutate(PostInt = case_when(NewDate %within% post_aug_break ~ "Yes",
                             NewDate %within% post_oct_break ~ "Yes",
                             NewDate %within% post_nov_break ~ "Yes",
                             NewDate %within% post_jan_break ~ "Yes",
                             NewDate %within% post_mar_break ~ "Yes",
                             TRUE ~ "No"))
```
```

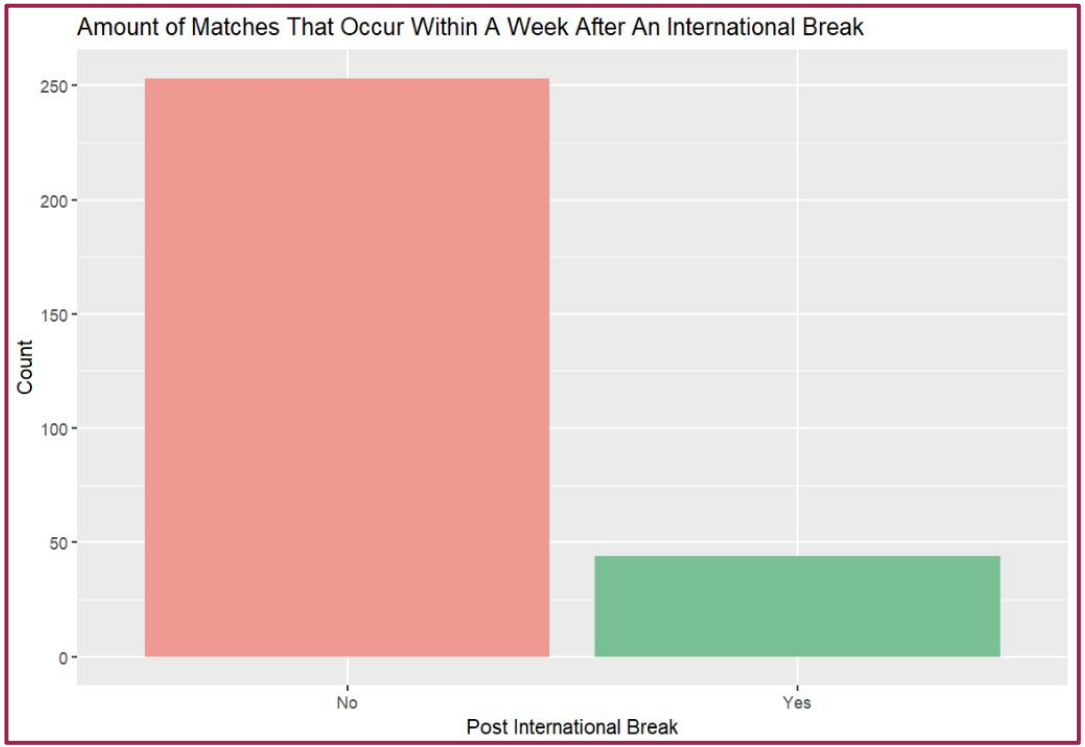
Predictors of Interest Individually





These two visualizations helped me better understand my data and how there really is something to each half of a match. One club could be winning at half time and then loses the whole match. All in all, seeing the differences between the two halves and helped me realize I needed to further study how HalfTimeWinner or HTR would impact the full time outcome.

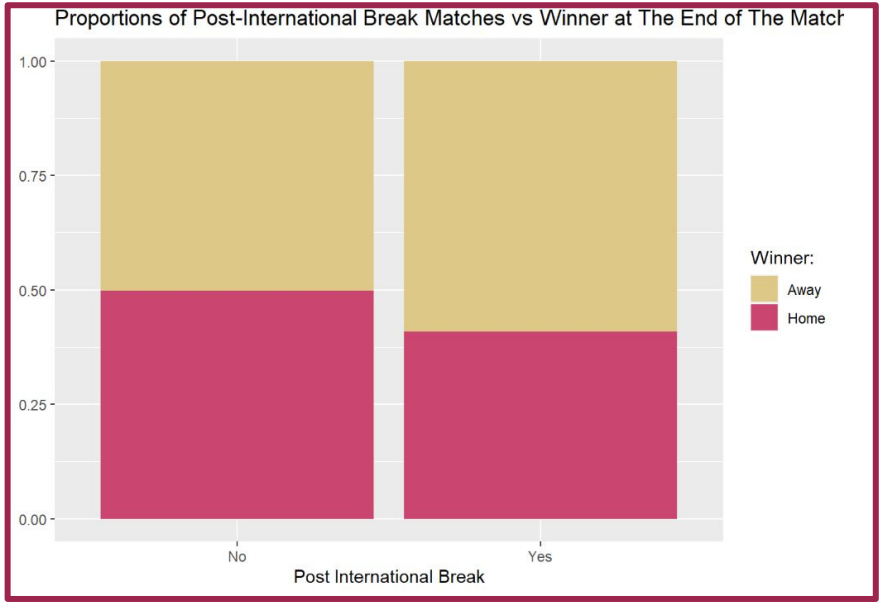
Predictors of Interest Individually



From this visualization, I learned that there are a lot more matches that fall right after an international break than I thought. With almost 50 matches that fall within a week of international breaks, I wanted to compare this predictor to my response.

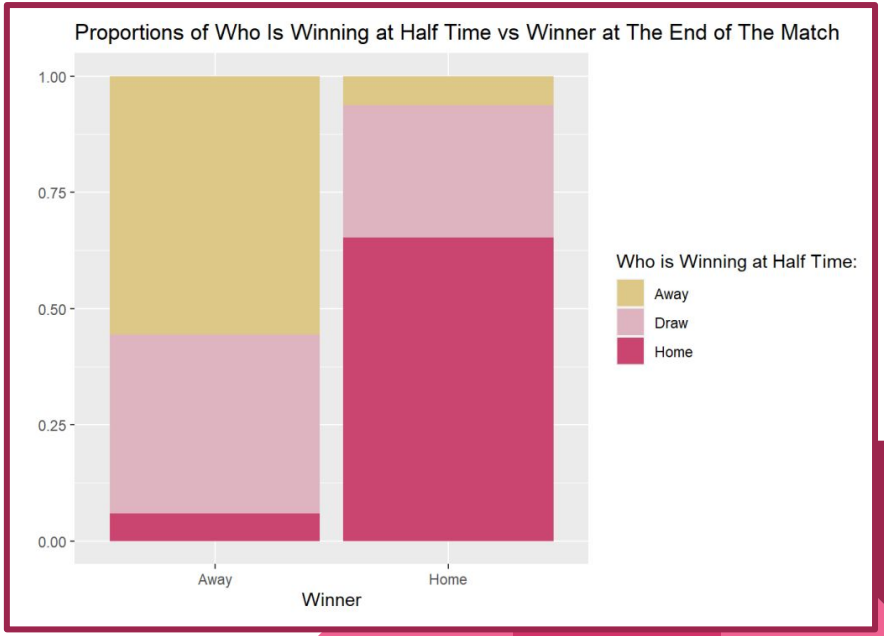
Categorical Predictors vs. Response

Response is **Winner**

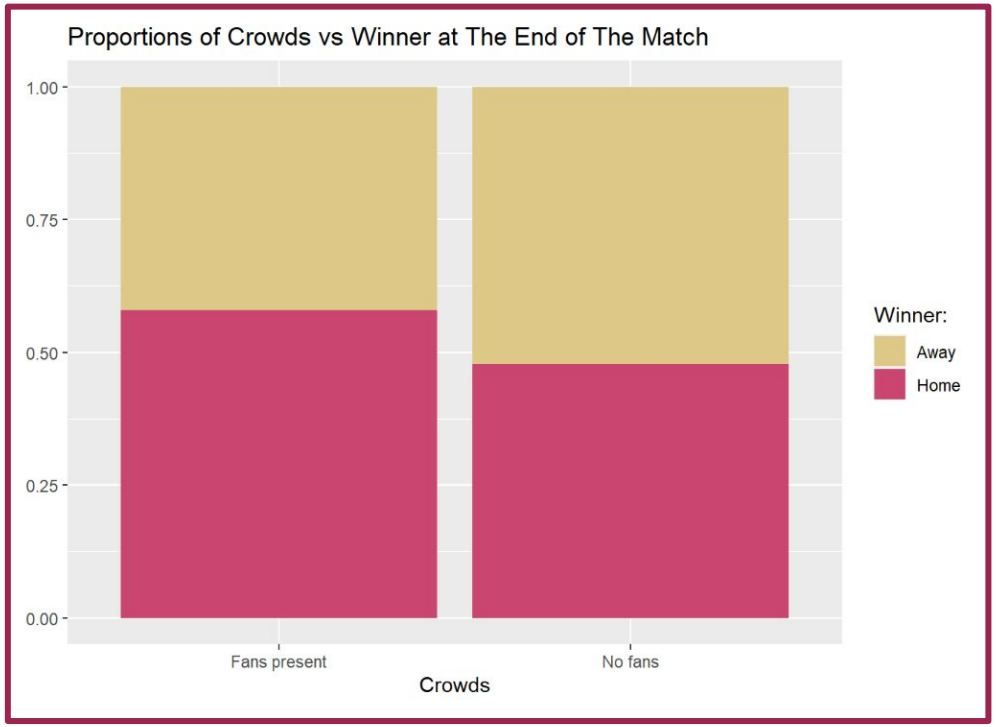


When it's regular season play, home clubs and away clubs win in equal proportions, but for the matches that followed an international break, around 60% of the matches resulted in an away club winning and around 40% of the matches resulted in a home club winning. Thus, there is a difference in who wins based on when a match occurs, and this helped me choose this predictor to be a part of the model I was planning.

For matches when the away club wins, the away club is winning at half time for about 60% of the matches, around 10% have the home club winning at half time, and 30% have the two clubs at a draw at half time. In contrast, when the home club wins, for around 65% of the matches, the home club is winning at half time and around 10% of the matches, the away club is winning at half time. For the remaining 25% of the matches when the home club wins, the match is at a draw at half time. Since there is a difference between the proportions in this visualization, I decided to add **HalfTimeWinner** to the model also.

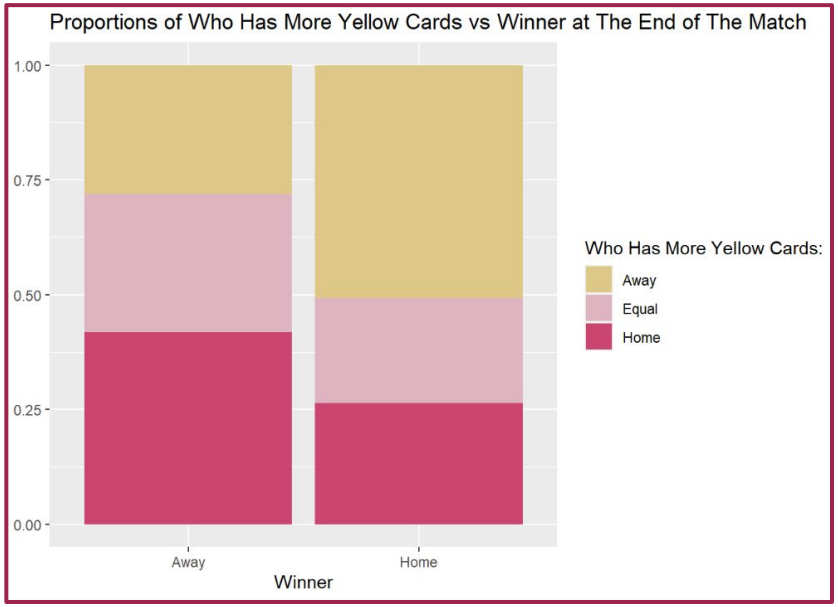


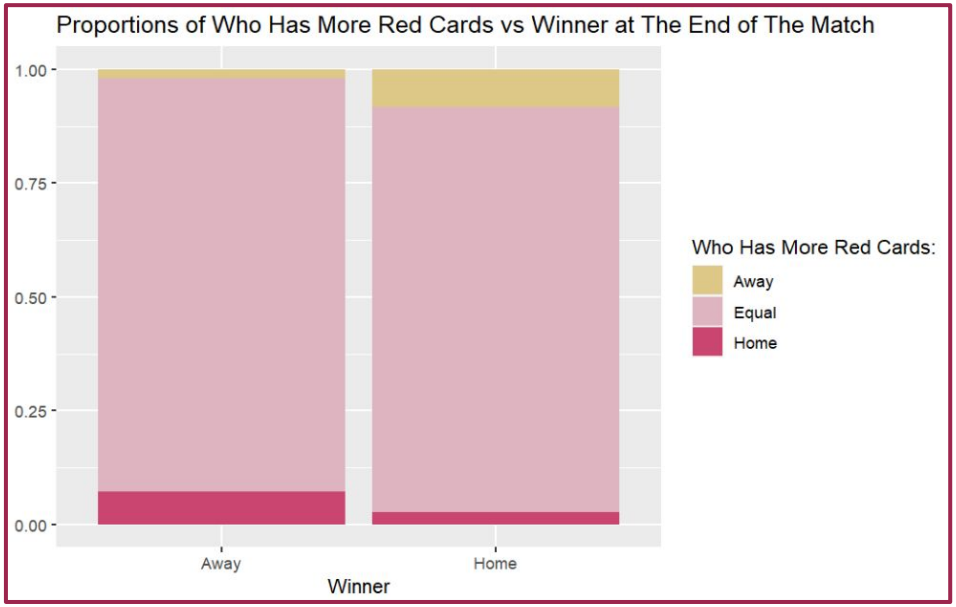
Categorical Predictors vs. Response



The visualization, “Proportions of Crowds vs Winner at The End of The Match,” reveal an almost identical set of proportions. However, it appears that for matches when fans are present, the home club wins more often than the away club. This slight difference is enough evidence to at least test this variable in my predictive model, but I did not expect it to have a significant impact on the predictions themselves.

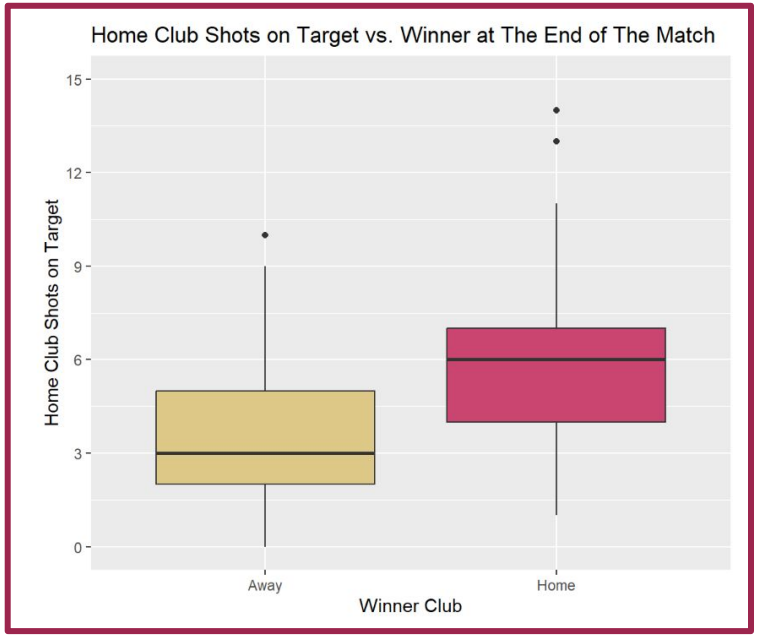
Categorical Predictors vs. Response





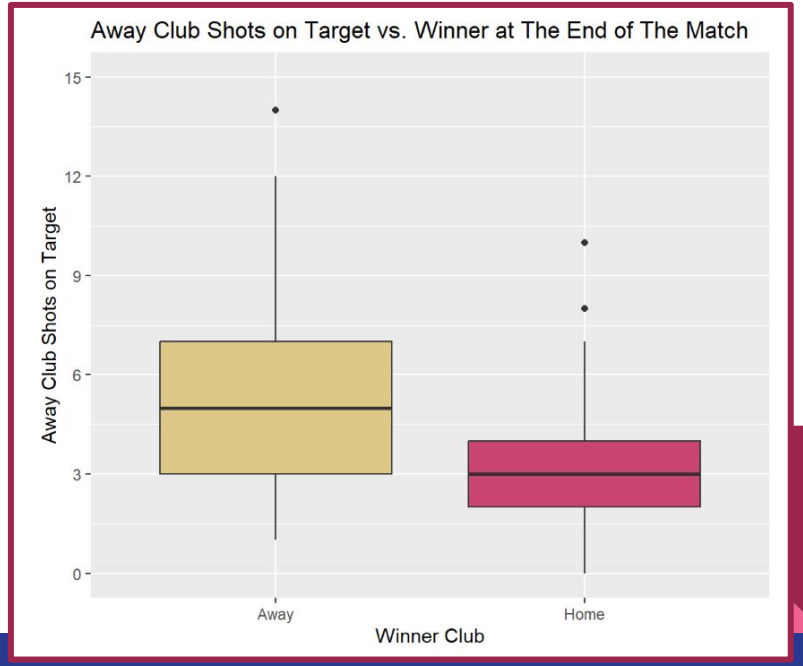
- Mirrored proportions are clear in these two visualizations, almost identical in magnitude but opposite in what they represent

Quantitative Predictors vs. Response



- Middle 50% of matches when the away club wins, the home club has approximately 2-5 shots on target with a range of 0 to 10 shots on target
- Median amount of home shots on target is 3 when the away club wins
- When the home club wins, the middle 50% of the matches have the home club shooting around 4 to 7 shots on target with a range of 1 shot to 14 shots
- Median amount of home shots on target is 6 when the home club wins

- Middle 50% of matches when the away club wins, the away club has approximately 3 to 7 shots on target with a range of 1 to 13 shots on target
- Median amount of away shots on target is 5 when the away club wins
- When the home club wins, the middle 50% of the matches have the away club shooting around 2 to 4 shots on target with a range of 0 to 14 shots
- Median amount of home shots on target is 3 when the home club wins



Logistic Regression for Prediction

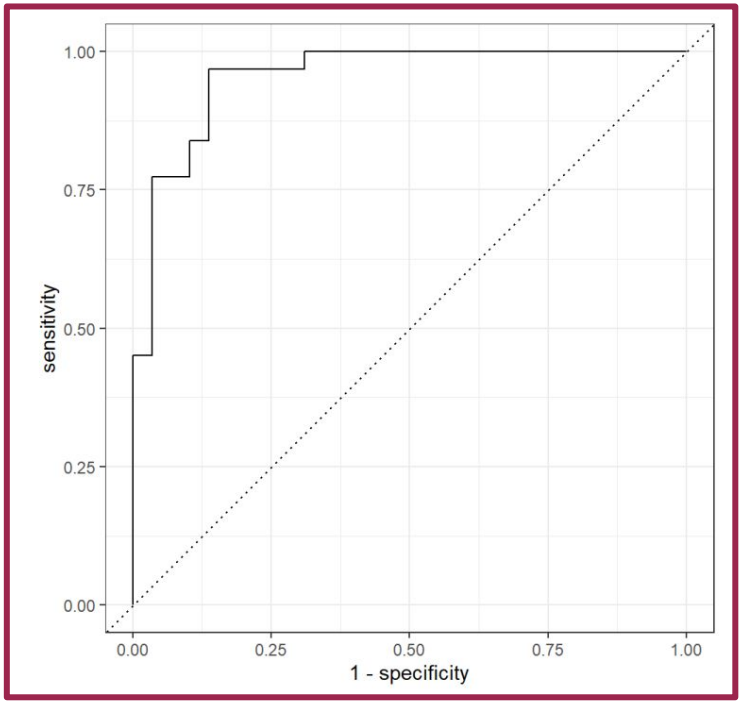
```
winner_fit <- glm(Winner01 ~ MoreFouls + MoreYellowCards + MoreRedCards + HalfTimeWinner + PostI
nt + HomeShotsOnTarget + AwayShotsOnTarget + Crowds, data = train_data, family = "binomial")
```

```
## # A tibble: 13 × 5
##    term                 estimate std.error statistic  p.value
##    <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)            -1.06      1.57     -0.674 5.00e- 1
##  2 MoreFoulsEqual         -0.715     0.848    -0.844 3.99e- 1
##  3 MoreFoulsHome          -0.197     0.465    -0.424 6.72e- 1
##  4 MoreYellowCardsEqual   -0.190     0.553    -0.343 7.32e- 1
##  5 MoreYellowCardsHome    -0.429     0.533    -0.805 4.21e- 1
##  6 MoreRedCardsEqual      -1.04      1.05     -0.990 3.22e- 1
##  7 MoreRedCardsHome       -2.16      1.36     -1.59  1.11e- 1
##  8 HalfTimeWinnerDraw      2.18      0.623     3.50  4.68e- 4
##  9 HalfTimeWinnerHome      4.68      0.713     6.56  5.50e-11
## 10 PostIntYes              0.684     0.608     1.12  2.61e- 1
## 11 HomeShotsOnTarget       0.481     0.110     4.37  1.25e- 5
## 12 AwayShotsOnTarget      -0.404     0.122    -3.30  9.56e- 4
## 13 CrowdsNo fans          -0.786     1.07     -0.737 4.61e- 1
```

- Classified **Winner** response in such a way that a home club win is represented by a 1 and a not home club win, or in other words, an away club win is a 0
- Predicted probability threshold of 0.5
- Training data is 80% of my refined dataset
- Testing data is 20% of my refined dataset
- Tried removing some predictors and eventually tried one predictor at a time, but this model produced the best results

Testing the Model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 26  6
##          1  3 25
##
##                Accuracy : 0.85
##                  95% CI : (0.7343, 0.929)
##     No Information Rate : 0.5167
##     P-Value [Acc > NIR] : 6.136e-08
##
##                   Kappa : 0.7007
##
##  Mcnemar's Test P-Value : 0.505
##
##             Sensitivity : 0.8966
##             Specificity : 0.8065
##          Pos Pred Value : 0.8125
##          Neg Pred Value : 0.8929
##              Prevalence : 0.4833
##          Detection Rate : 0.4333
##    Detection Prevalence : 0.5333
##       Balanced Accuracy : 0.8515
##
##        'Positive' Class : 0
##
```



ROC Curve Statistics:

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.954
```

The accuracy for this model was found using a Confusion Matrix reporting 85% accuracy with an ROC Curve area of 0.95, making the model, a good predictor of the probability of who will win a match based on the predictors of interest.

Conclusion

- Through creating a logistic regression model and using the Confusion Matrix and ROC curve, I conclude that the ideal combination of match predictors for predicting the probability of who will win a match are **MoreFouls**, **MoreYellowCards**, **MoreRedCards**, **HalfTimeWinner**, **PostInt**, **Crowds**, and **HomeShotsOnTarget** and **AwayShotsOnTarget**.
- **Limitation:** Since this data came from a year of transition from Covid, I had hoped this dataset would provide statistical or visual evidence of the impact of crowds. However, it was very difficult to see any strong relationship between **Crowds** and **Winner** outcome. I believe this is due to the drastically different sample sizes of fans present and no fans present.
    - When building a logistic regression model with **Crowds** as my only predictor for **Winner**, the model produced a Confusion Matrix accuracy of less than 50%.
- **Limitation:** Classifying the response to only have a home club win or an away club win does not accurately reflect football. Since there are draws in football, I think K Nearest Neighbors would probably be a better predictive model.
- **Future Research:**
    - Using an entire covid season with no fans and entire season of all fans would probably be able to address the issue with **Crowds**
    - Using new datasets with different predictors and merging the two datasets could possibly lead to an even better model