Analyzing Weather's Impact on NFL Outcomes

In this project, we will analyze a data set labeled "NFL scores and betting data" This data set contains National Football League (NFL) game results since 1966. It also includes the weather conditions and betting data from each game. Weather data contained in the data set has been collected from the National Oceanic and Atmospheric Administration. It is a common consensus in the NFL that adverse weather benefits the home team because the home team is used to the weather, used to their own wet field conditions, etc. That consensus is the focus of this 10-page report.

Data Set URL:

https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data?resource=download&select=spreadspoke_scores.csv

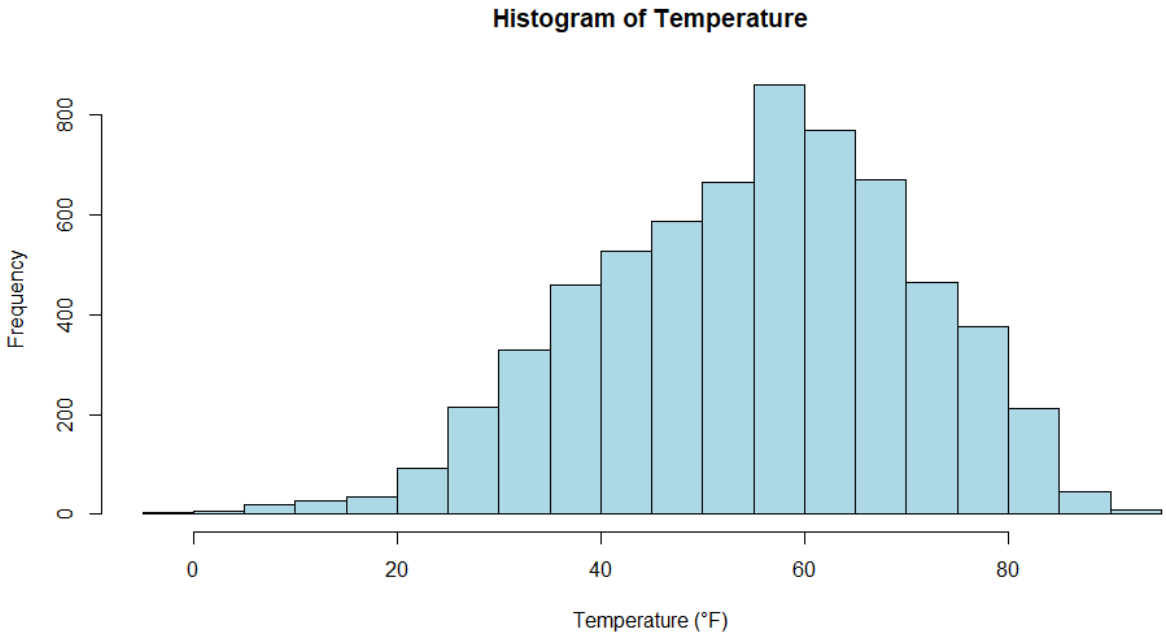
Inferential Questions:

1. How does the temperature affect the probability of the home team winning? Are home teams more likely to win in colder or warmer temperatures?
2. Does wind speed have an impact on the winning probability of the home team? Are home teams more likely to win in games with low or high wind speeds?
3. How does humidity influence the probability of the home team winning? Are home teams more likely to win in games with low or high humidity levels?

Variables of Interest:

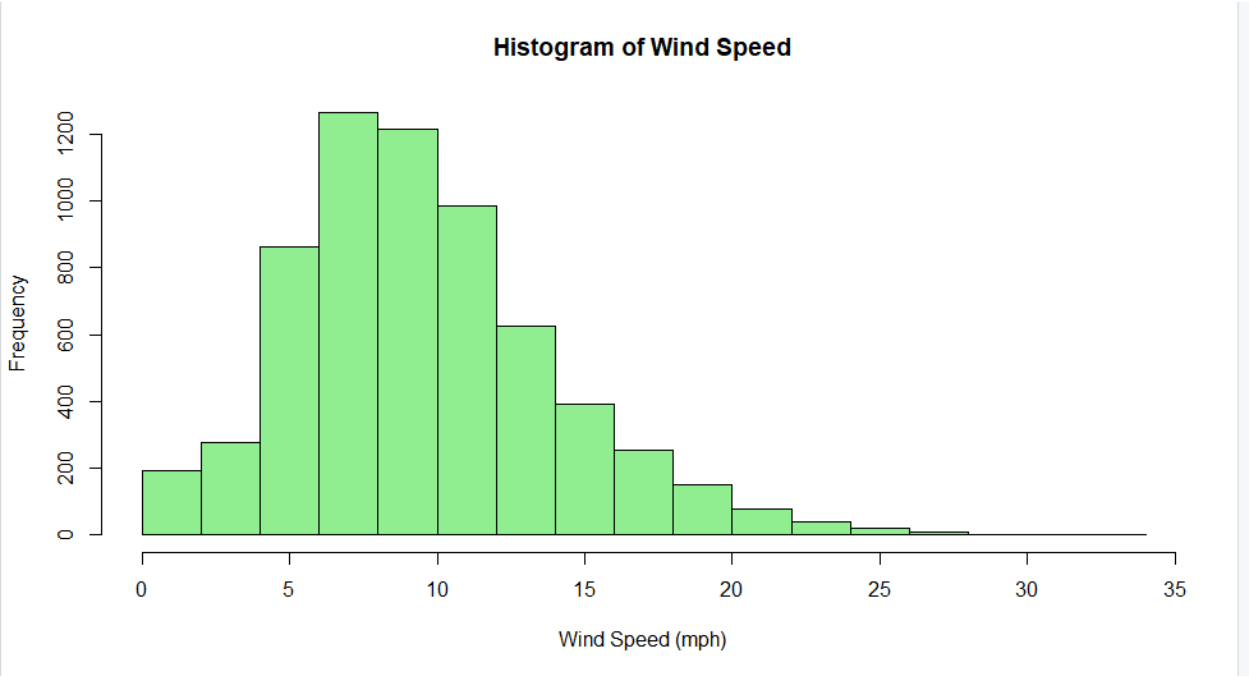
1. Temperature (Fahrenheit):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3	44	57	55.45	67	95



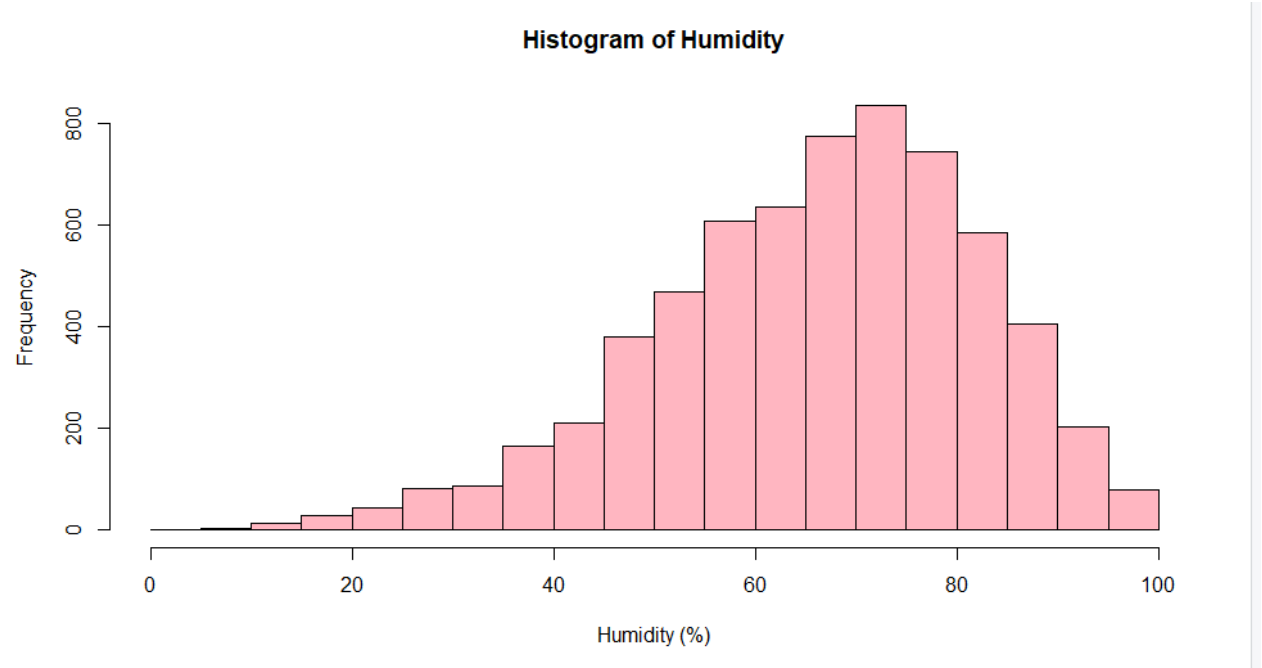
2. Wind (MPH):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	7	9	10.04	12	33



3. Humidity:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4	56	68	66.57	78	100



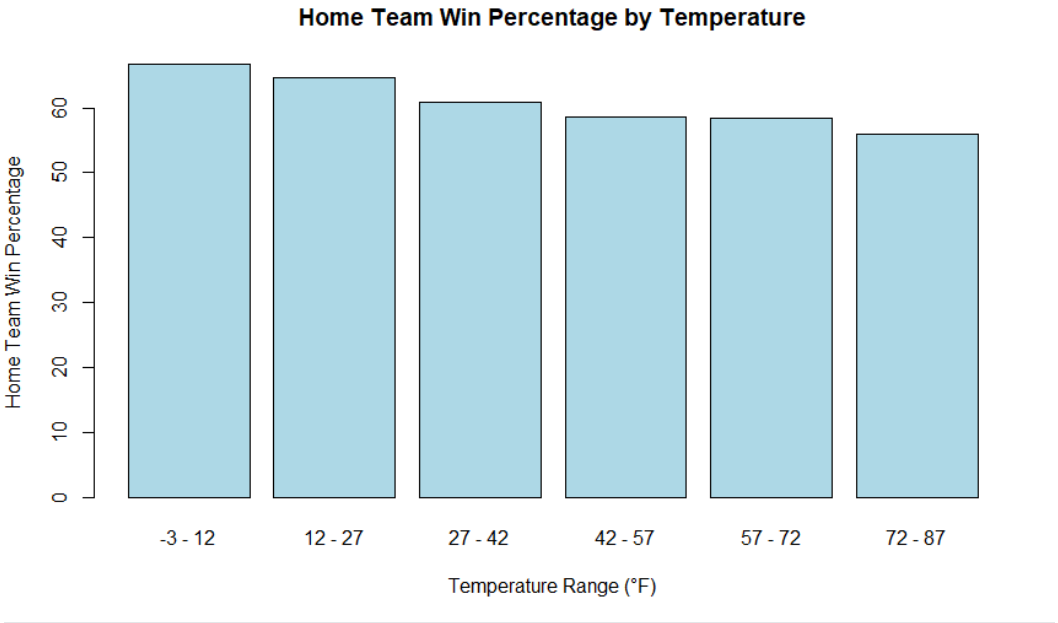
As seen above, the three variables of interest are temperature, wind speeds and humidity all recorded at the start of the NFL game. By studying these variables and their interactions, we

can gain insights into the potential impact of weather conditions on NFL game outcomes and the increasing of the home-field advantage during inclement weather.

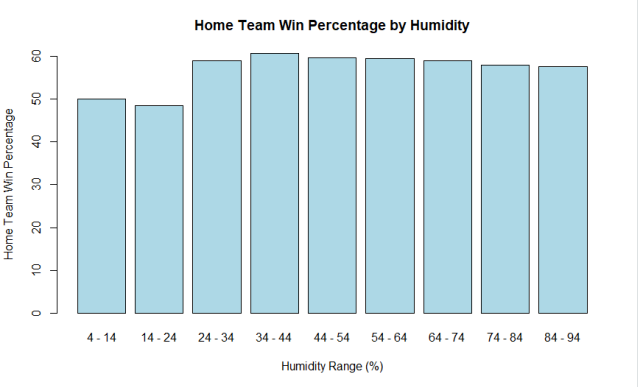
Data Analysis:

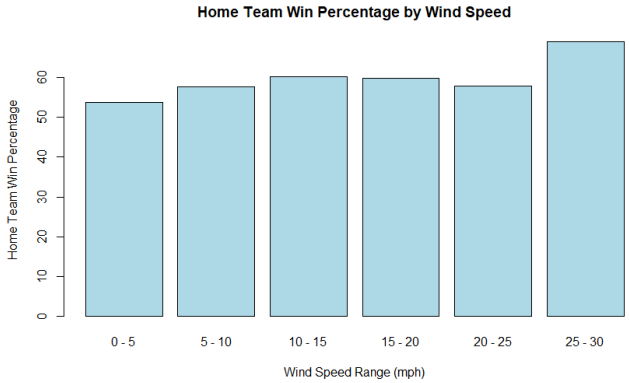
An extra column is created in the data set with a value of 1 if the home team wins and a value of 0 if the away team wins/home team loses. It is important to understand that in the NFL, the home team already has a big advantage due to no travel, home crowd, home locker room, etc. When Vegas comes up with their betting spreads, they provide the home team with a few points on the spread to cover this home-field-advantage. So, with the raw data it is found that the mean of this new column is .5859252. In other words, throughout NFL history the home team wins 58.59% of the time.

To analyze the data, we will divide the variables of interest into ranges and evaluate the win percentage of the home team at each range of temperature.



We can see that as the temperature decreases the winning percentage of the home team does increase. This appears to provide evidence to answer our first inferential question implying that home-team win percentage increases in colder weather.





No such trends are observed in the graphs involving the other two variables of interest. No evidence to answer our second or third inferential question.

Statistical Model:

A suitable statistical model for estimating population features and relationships from the data in this study is logistic regression. Logistic regression is a type of generalized linear model that is used to model the probability of a binary outcome based on one or more predictor variables. The binary outcome in this case is a home team win (1) or home team loss/away team win (0) and the predictor variables would be temperature, wind speed and humidity.

In the logistic regression model, we model the log-odds of the home team winning as a linear function of the predictor variables. The logistic function then transforms the log-odds back to probabilities, which range between 0 and 1. The model can be represented as:

$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 * \text{Temperature} + \beta_2 * \text{Wind} + \beta_3 * \text{Humidity}$

Where Y is the binary outcome indicating a win (1) or loss (0) for the home team, and β_0,

β_1, β_2, and β_3 are the coefficients that quantify the relationship between the predictor variables and the log-odds of the home team winning. By fitting a logistic regression model to the data, we can estimate the coefficients for the weather variables and assess their statistical significance.

Fitting the Data:

Logistic regression models are typically fitted by using MLE (maximum likelihood estimation). MLE procedure finds the parameter values that maximize the likelihood of observing the provided data. In logistic regression, the likelihood function measures how well the model's predicted probabilities match the observed binary outcomes in either a win or a loss. To fit the logistic regression model, we need to find parameter values that maximize the log-likelihood function. This process can be done using the Newton-Raphson method. This method finds the optimized parameter values by iteratively updating them using the gradient and the second derivative of the log-likelihood function. After performing this on our data, the parameters values found are:

β_0	β_1	β_2	β_3
0.5763426	-0.004659091	0.01160296	-0.001298435

Synthetic Data Generation:

To perform a simulation study to investigate the performance of the above estimation procedure, we first need to generate synthetic data. To generate synthetic data first we need to

establish the true population parameters which we have labeled above. Then, we need to generate data sets of predictor variables by simulating their values. This will be done by using the rnorm function in R.

We plug the generated values into the logistic regression model using the true population parameter values as the model's coefficients. This computes the log-odds for the home team winning at each synthetic data point. Convert this to probability through the logistic function and then generate binary outcomes. Combine the synthetic predictor variables and synthetic binary outcomes and there we have synthetic data generated that captures the true population features.

Simulation Study:

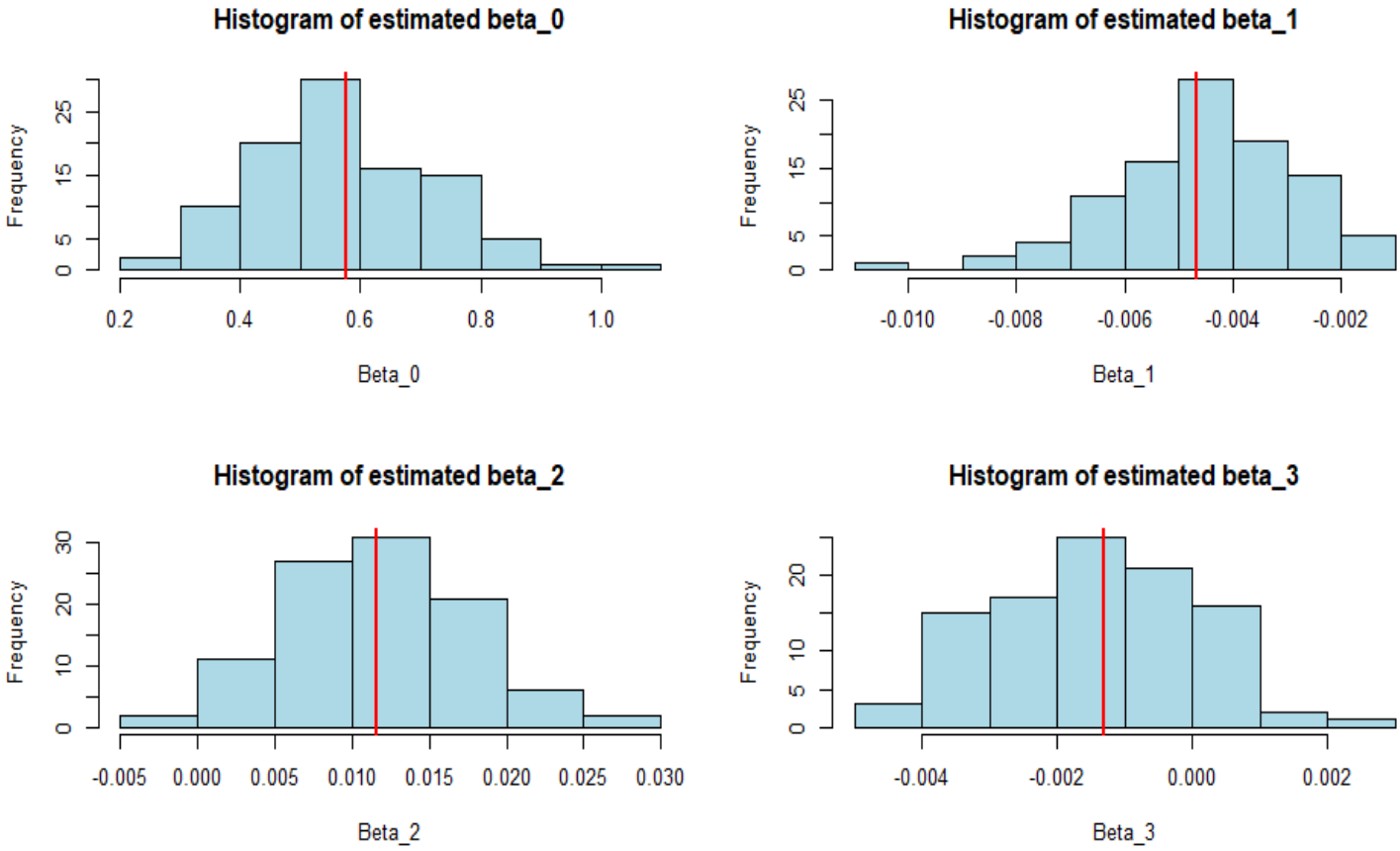
In the simulation study we aim to analyze the performance of our MLE estimation procedure using the Newton-Raphson method for estimating the parameters of a logistic regression model. It will be done following these steps:

1. Generate a synthetic data set using the procedure outlined in the previous section.
2. Implement Newton-Raphson method on the data set.
3. Evaluate the estimated model by calculating the Root Mean Square Error (RMSE) of the predicted and observed values.
4. Repeat steps 1-3 N times.
5. Use the output from 4 to produce plots for inference.

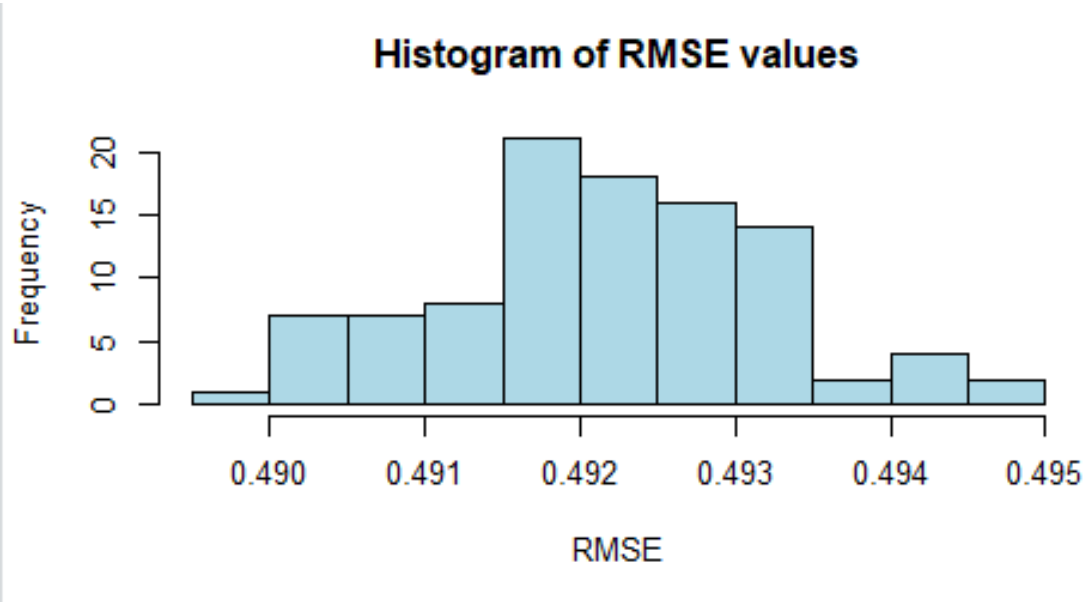
We have chosen N = 100 for this simulation study because with 100 simulations we will observe a reasonable amount of variability in the generated data and the computational power

needed to produce these simulations will be kept to a minimal. In addition, we chose the sample size of each data set to be 6366 because that is the number of values contained in the real data set. Lastly, the population features chosen were three because we have three predictor variables. Results are on the following page.

Results:



All four histograms are approximately centered around the red line (red line indicates the true parameter value) which suggests that our estimation procedure (MLE with Newton-Raphson method) performed well in estimating the true population features on our synthetic data.



Additionally, RMSE values are a measure of how well the estimated model predicted the actual model. Low RMSE values show the model's predictions are closer to the true values. The histogram shows a very small range of approximately .005 indicating a consistent performance across the synthetic data sets. Overall, I would conclude that the results of the simulation study do in fact lend confidence to the logistic regression model and MLE procedure using the Newton-Raphson method implemented on the real data.

Conclusion:

It is important to consider that there are limitations of our model. For example, NFL games do not have a binary result. Regular NFL season games can end in a tie. However, due to their rarity that was ignored (Sports King labels the odds of an NFL game ending in a tie at .2%). Additionally, the model assumes a linear relationship between the log-odds of the home team winning and the weather variables, which might not hold true in all cases. Lastly, the model does

not factor in team skill, experience, game strategies, injuries and anything else that could impact an NFL game.

It does appear that the broader implications and inferences learned from the model fit to the real data imply that weather conditions measured before an NFL game, such as temperature, wind speed, and humidity, may have an impact on the probability of the home team winning. I would suggest that the variable with the highest impact would be temperature. These implications seem reasonable and plausible because home teams are more accustomed to the weather of their hometown than their visiting opponents. They would not be accustomed to playing conditions of extreme wind speed or extreme humidity because those playing conditions are not practiced in while temperature is.