**Blowing out the Competition**

An Analysis of how to win in the National Football League

**Francis Littig**  
Department of Computer Science  
 Kennesaw State University  
Marietta, GA, USA  
 [flittig@students.kennesaw.edu](mailto:flittig@students.kennesaw.edu)

**Ahsan Jamal**  
 Department of Computer Science  
 Kennesaw State University  
 Marietta, GA, USA  
 [ajamal@students.kennesaw.edu](mailto:ajamal@students.kennesaw.edu)

**Brendon Im**  
 Department of Computer Science

Kennesaw State University  
Marietta, GA, USA  
[bim@students.kennesaw.edu](mailto:bim@students.kennesaw.edu)

**ABSTRACT**

As the NFL continues to be popular in the United States, teams continue to look for ways to find factors of success, what has helped them win championships before, and what has cost them championships now. This paper will utilize linear and logistic regression in order to check the relationship of games won to factors like passing yards and rushing yards. From the data gathered, it can be concluded that rushing yards is the strongest individual factor that influences the number of wins a NFL team has in a season. Furthermore, a multiple linear regression model is better than single linear regression for modeling a complex and dynamic game like football.

**KEYWORDS**

linear regression, multiple linear regression, logistic regression

## 1 Research Statement & Conjecture

In all sports, every team has the goal to compete and become victorious over their peers. Nowadays, in order to compete in their maximum potential, teams hire a management staff in order to for the athletes to train their body, train their mind, watch their nutrition, manage their activities, In the end, all of the training and activities are done so that during a game, players can show the results of their training through the most common metric of their success: goals in soccer, runs in baseball, points in basketball, touchdowns in football, and other ways that points are scored for teams in these sports.

Even though a casual viewer may be interested in just the scores of games and a team’s match history to decide a favorite, sports teams must analyze more in depth into why they’ve achieved success, or had it taken from them by another. Many people have spent years accumulating data on every metric they can think of for football: yards passed, yards rushed, the number of attempts done for both, interceptions, the number of touchdowns from passing or rushing, and more. Teams now hire analysts and statisticians to gather and organize this data so that they may find their own optimal path to success. This data used in this paper will be from teams that participate in the National Football League (NFL) which is the most professional level of American Football available in the United States of America.

This paper is to showcase the potential that data analysis has in being able to obtain information that can either assist a team that is behind others in the standing of their season or help a team that is already ranked highly continue to improve in order to stay above their competitors in the league. By giving teams a sense of direction in focusing their efforts to improve on,by using the data of the highest level of football teams, other teams that are not in the NFL may potentially be able to raise their level of play to compete at that level.

The purpose of this paper is to provide an analysis of various statistics measured across multiple years in order to determine which metric would be most effective for a team to utilize in order to have a higher chance at success during the football season. This paper will look at four different metrics taken by Pro Football Reference [1] from the 2018 and 2019 seasons. These metrics are Pass Completion (%), Yards Passed, Yards Rushed, Team Turnovers, Games Won (per season), and Winning Season (1: yes, 0:no). These metrics will be analyzed using linear and logistic regression. As this style of research has been done before with logistic regression to analyze canadian football [2], logistic regression and a Markov Chain Model to analyze basketball for the National Collegiate Athletic Association (NCAA) [3], and more. Section 2 will take a closer look into the work of other researchers that have answered questions similar to this paper and Section 3 explores the methodology used to analyze the data. Section 4 goes into how the methodology was implemented, the results of that implementation, and the analysis of the results and Section 5 concludes the paper with how this research could’ve been further improved upon.

## 2 Related Work

Keith Willoughby in his paper analyzed canadian football and asked the question, “Which (if any) of the varied statistics produced in a football game are vital in determining a game’s outcome?” [2]. Basing his research on other papers that have done similar topics with other analytical tools, Willoughby focused his research on a few certain variables, notably the passing yardage, rushing yardage, number of interceptions, number of fumble recoveries, and number of quarterback sacks. Willoughby chooses to rather than perform his analysis on the entire Canadian Football League (CFL), he would focus on three specific teams that represent three areas of rankings within the CFL: Calgary for having the highest amount of wins, Ottawa for having the least amount of wins, and Saskatchewan for having the closest amount of wins and losses. Willoughby chooses to use logistic regression. The independent variables that Willoughby uses for his model are the difference in rushing yardage, difference in passing yardage, difference in the number of interceptions, the difference in the number of fumble recoveries, and the difference in the number of quarterback sacks. With these variables, Willoughby came to the results that the more wins than losses that a team in the CFL has, the more passing and rushing yardage matters. So, that means that in order to imitate the success that these higher ranked teams have, coaches would have to focus on both passing and rushing yardage which is correlated to their offensive plays.

Willoughby mentions the potential failings of utilizing logistic regression here due to the inability of the binary nature of logistic regression to determine the differences between a 1-point difference win and a 50-point difference win. This fault does have merit in how the actual games also ignore the point differential, as stated by Willoughby. In the end, his use of logistic regression was well implemented and provided insight into how a study of this kind can be performed.

Kvam and Sokol used not only logistic regression, but also a markov chain model in order to used various statistics gathered from the NCAA in order to be able to predict the outcome of games in the Division 1 men’s basketball tournament [3]. Kvam and Sokol chose to research this topic in order to answer how to accurately rank teams using basic input data. Their input data is based solely on games won and games lost for a team and the games won and lost for a team against specific teams. These four variables are then passed through the markov chain model with more probabilities obtained by using logistic regression. In the end Kvam and Sokol arimed to create a new model for predicting the victor of games in the NCAA, and they succeeded by having the model outperform other models designed for the same purpose and that a reason for this is that while other predictive models used a binary system of games won or lost, their model differentiated various levels of victories to provide a more accurate prediction.

Arnason et al. looked past the statistics of the game alone and investigated the correlations between the physical fitness of soccer athletes and their performance as a team as well as differences through the player positions [4]. The physical fitness of players was measured through peak O2 uptake, body composition or body mass index, the power in the legs of the athletes, jump height, flexibility, and comparison to team average. In these physical tests, the results were found using linear regression in order to find possible relationships between the various variables and their division in soccer.

The position of the players were also classified and found that the most popular formation was to have four defenders, four midfielders, and two attackers. The difference between the players in their positions were also analyzed between the defenders, midfielders, attackers, and goalkeepers. The research found that the major difference in team averages were only in height, but in terms of individual metrics, peak O2 uptake and body composition had the most significant differences between an elite league team and a first division team. However, when measured between their standings in the league, the variables with the biggest correlation to team success were found to be jump height having the most significance with leg power and body composition having an effect on team success as well as incurring less injuries throughout the season. The differences between goalkeepers and other players were also massive in flexibility, peak O2 uptime and time lost from injuries.

In the end, Amason et al. were able to find various relationships between the physical fitness of soccer athletes in jump height, body composition, and leg power to team success alongside a lower amount of injuries for the team which should lead the coaches towards focusing on power training as well as preventing future injuries and the rehabilitation of previous injuries in order to maximize their factors of success.

## 3 Methodology

Both linear and logistic regression models were created to show which metrics had the strongest correlation to games won in a season. For linear regression, rushing yards, passing yards, and pass completion are the x-values. Games won is the y-value. Single linear regression was done with each of these x-values individually, and multiple linear regression was also done with all three metrics at once. For logistic regression, rushing yards and turnovers are the x-values, and they were modeled individually with winning season (1:yes, 0:no) as the y-value.

## 4.1 Design

The data was gathered from Pro Football Focus which is a website that has free historical NFL data. Two CSV files were created called “2018data.csv” and “2019data.csv” that contain 32 rows representing all teams in the NFL and columns representing their respective data for that season. Ryds (rushing yards), Pyds (passing yards), Cmp (pass completion %), GW (games won), TO (turnovers), and Wseason (winning season) are the headers used.

The main framework for our linear and logistic regression was Python Pandas. Several libraries were used: Matplotlib for linear graphs, Seaborn for logistic graphs, Sklearn for linear and logistic regression models, and Statsmodels for regression statistics. Three Python programs were created to perform these functions: “linreg.py” for single linear regression, “multlinreg.py” for multiple linear regression, and “logreg.py” for logistic regression.

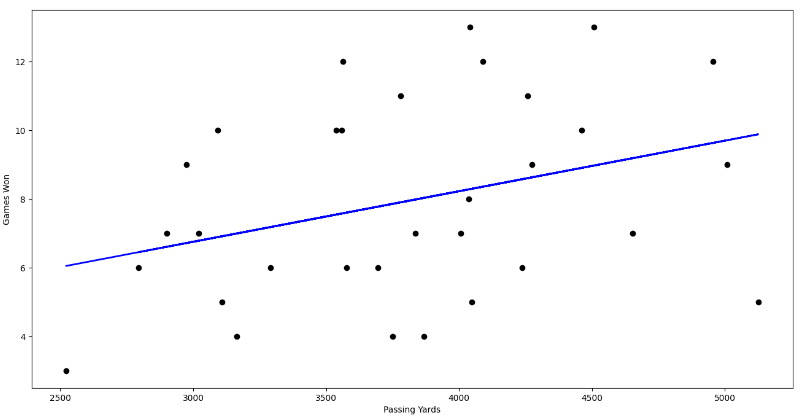
### 

### 

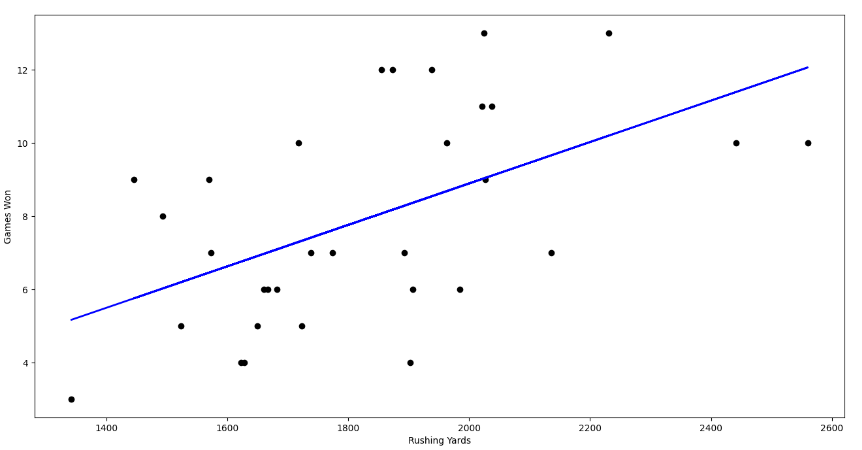
### 

### 

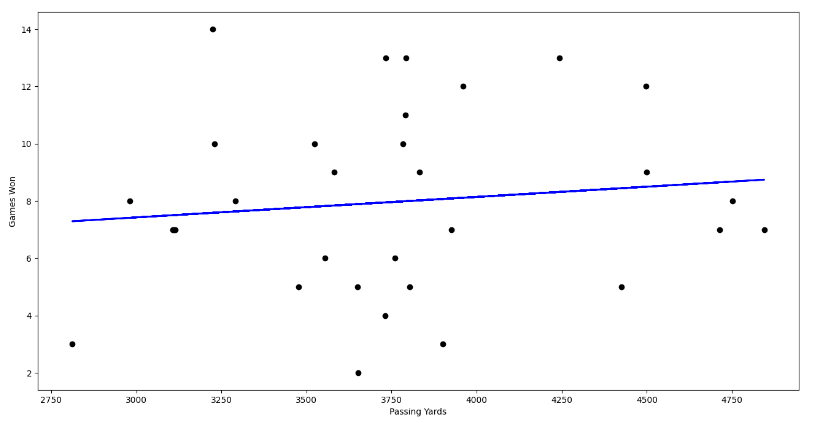
### 4.2 Results



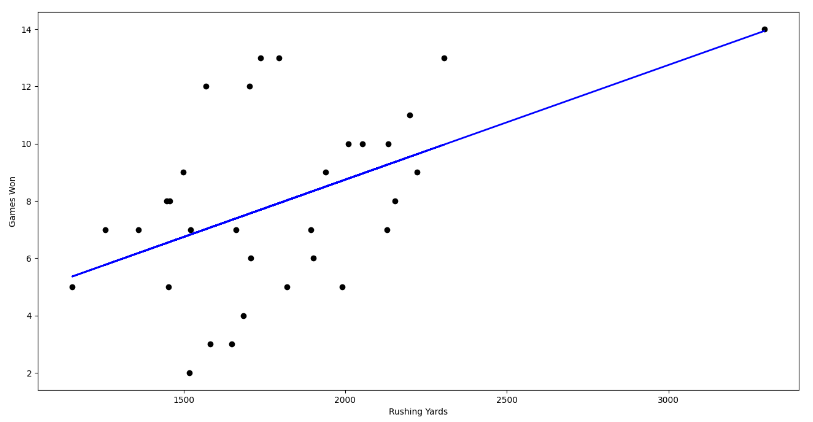
### Figure 1: Linear Regression Passing Yards vs Games Won 2018



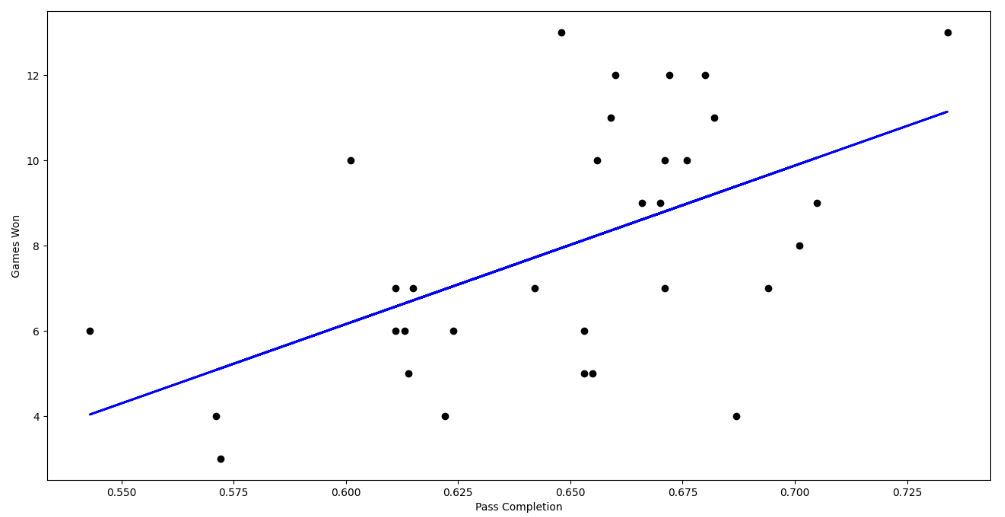
### Figure 2: Linear Regression Rushing Yards vs Games Won 2018



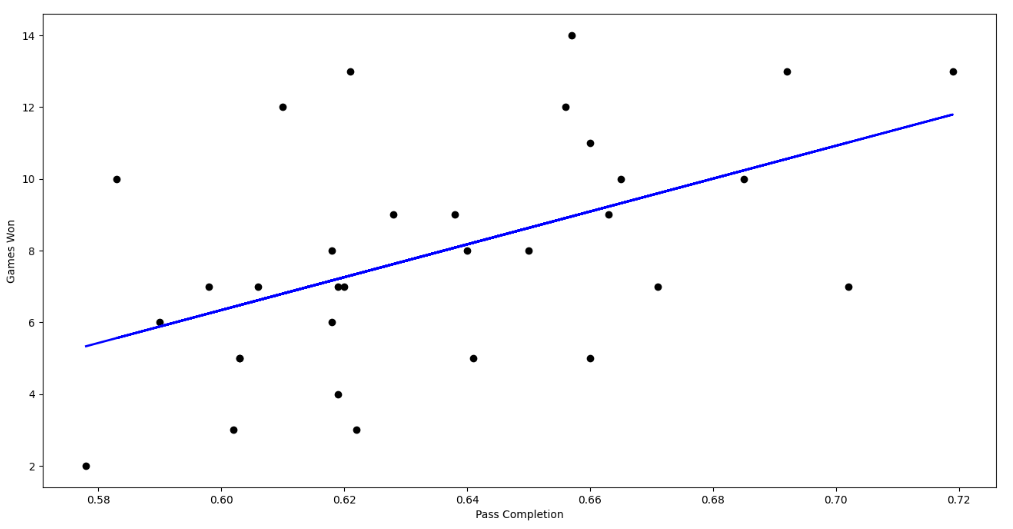
### Figure 3: Linear Regression Passing Yards vs Games Won 2019



### Figure 4: Linear Regression Rushing Yards vs Games Won 2019



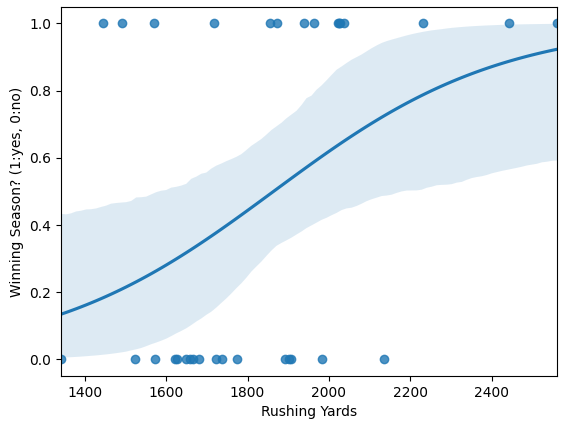
### Figure 5: Linear Regression Pass Completion vs Games Won 2018



### Figure 6: Linear Regression Pass Completion vs Games Won 2019

### 2018 turnovers vs winning season

### Figure 7: Logistic Regression Turnovers vs Winning Season 2018

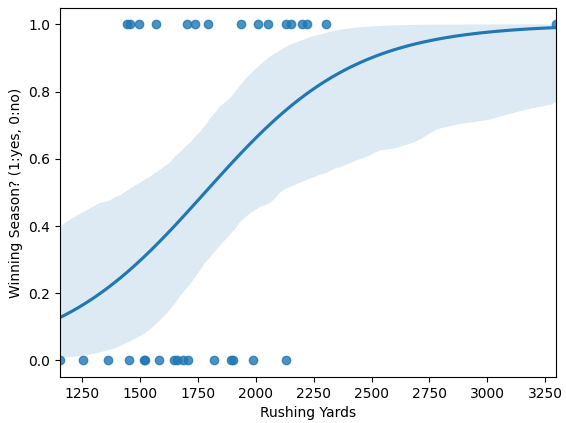


### Figure 8: Logistic Regression Rushing Yards vs Winning Season 2018

### 

### 2019 turnovers vs winning season

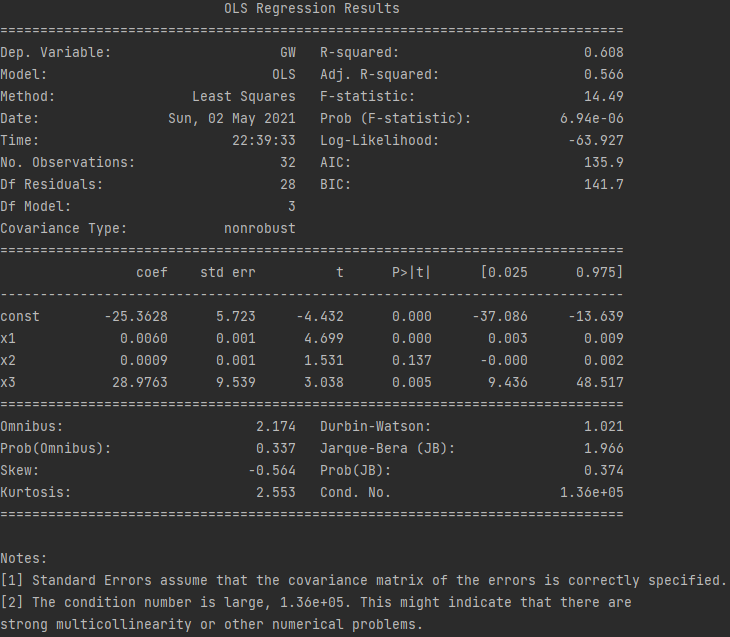
### Figure 9: Logistic Regression Turnovers vs Winning Season 2019



### Figure 10: Logistic Regression Rushing Yards vs Winning Season 2019

### 4.3 Analysis

At first glance, it can be determined that rushing yards has a stronger linear correlation to games won due to the slope of the line in the graphs. More specifically, the R2 value is consistently the highest for rushing yards. This value indicates how closely the model fits the data and is obtained from the Statsmodels package. The model for multiple linear regression for the 2018 season proved to be the best fit overall with a R2 value of 0.608.



**Figure 11: Multiple Linear Regression Ryds, Pyds, Cmp vs Winning Season 2018**

## 4.4 Comparison

In comparison with Dr. Willoughby’s analysis of Canadian Football we can determine that our two separate analyses have similar results. Dr. Willoughby focused on the difference between interceptions, the difference between fumbles, the difference between sacks, the difference between rushing yards, and the difference between passing yards. Dr. Willoughby also primarily wanted to determine the winner of a single game. Our analysis turns the focus on season long results. A team’s success is determined based on a winning season, not necessarily the outcome of a single game. For the purposes of our preliminary testing we also reduced sacks, interceptions, and fumbles into the single statistic known as turnovers. This single statistic encapsulated all meaningful plays on defense as well as mistakes made on offense. Furthering the differences between our two tests, we wanted to perform both a logistic regression and linear regression, Dr. Willoughby purely makes use of the logistic regression. Even with these apparent differences our two analyses ended with similar results, an opportunistic defense and solid running offense will lead to more wins than just about any other combination of team traits.

## 5 Conclusion

Our findings show that there are indeed traits of a team that lead to winning seasons more so than others. There is a stronger linear correlation between rushing yards and winning than passing yards and winning. Essentially meaning that rushing yards have a greater effect on the outcome of a game than any other statistic.

The shortcomings of this model include the dataset, which only consists of two seasons of data. In a future analysis we would expand the dataset to include more seasons. This would allow the model to take into consideration long term success of teams as well as play calling trends in the NFL. Furthermore, we could include a more robust dataset, one which dives deeper into the minutiae and attempts to single out which plays or series of plays led to bigger success in the NFL.

For this model we made use of logistic regression and linear regressions. Another option we could have made use of is PCA, or principle component analysis. PCA would essentially allow us to take a large dataset, one that could include advanced statistics such as short, intermediate, and long pass completions, or short, intermediate, and long runs, or even interceptions that led to points being scored and reduce it into a smaller dataset, with less variable while maintaining all the advantages of the larger, more variable containing set.

In order to successfully implement PCA we would firstly need to standardize the data. This would be done in order to limit the effect of large differences in range between variables. This will allow the data to be presented in a comparable scale to one another. Next we would compute covariance matrix, eigenvectors, and eigenvalues. We would then eliminate eigenvectors of lesser significance and form feature vectors. Finally, we would take the feature vector and multiply it by the standardized original data set, creating our final PCA reduced dataset [5].

However, this model did not need the use of PCA to determine that rushing yards have a stronger correlation to winning football than passing yards or any other studied statistic in the NFL.

**REFERENCES**

[1] "Pro Football Statistics and History", *Pro Football Raference*. [Online]. Available: https://www.pro-football-reference.com/. [Accessed: 26- Apr- 2021].

[2] K. Willoughby, "Winning Games in Canadian Football: A Logistic Regression Analysis", *The College Mathematics Journal*, vol. 33, no. 3, p. 215, 2002.

[3] P. Kvam and J. Sokol, "A logistic regression/Markov chain model for NCAA basketball", Naval Research Logistics, vol. 53, no. 8, pp. 788-803, 2006. Available: 10.1002/nav.20170.

[4] A. ARNASON, S. SIGURDSSON, A. GUDMUNDSSON, I. HOLME, L. ENGEBRETSEN and R. BAHR, "Physical Fitness, Injuries, and Team Performance in Soccer", Medicine & Science in Sports & Exercise, vol. 36, no. 2, pp. 278-285, 2004. Available: 10.1249/01.mss.0000113478.92945.ca.

[5] “A Step-By-Step Explanation of Principal Component Analysis”, Builtin. [Online]. Available: [https://builtin.com](https://builtin.com/data-science/step-step-explanation-principal-component-analysis). [Accessed 1- May- 2021].