CAN WE PREDICT THE NUMBER OF SEASONAL WINS IN THE NFL FROM A PREVIOUS SEASON?

Submitted by: Alicia Rene Jacobs

THE PURPOSE OF THIS REPORT IS TO ANSWER THE FOLLOWING QUESTION:

- **Can we predict the number of NFL team seasonal wins from a prior season's data?**

1. Introduction

The aim of this report is to analyze data from an unspecified NFL season, focusing on team season wins, with the goal of predicting the number of wins a team is likely to achieve in the upcoming season. To accomplish this objective, we will leverage the dataset "nfl_project2.csv," which was obtained from the source http://nflsavant.com/.

Our analytical process will begin by constructing scatter plots to visualize potential relationships between our independent variables and the dependent variable. Subsequently, we will compute correlation coefficients and p-values for each independent variable in relation to the dependent variable. Following this, we will employ the sklearn library to identify the most suitable features for our predictive models.

Our analysis will proceed with the development of several simple and multiple linear regression models. These models will be evaluated using various statistical metrics, including R-squared (R2), T-testing/F-testing, mean squared error (MSE), and residual mean squared error (RMSE), to refine our options down to two models.

The final model will be chosen through a comparative analysis of the two selected models, involving residual analysis. This analysis will include the presentation of residual plots to validate the assumption fit, calculation of variance inflation factors (VIF), and visual inspection of QQ plots/histograms for normality. Furthermore, we will reevaluate MSE and RMSE using both training and testing datasets.

Upon the selection of the final model, we will proceed to compute predictions and confidence intervals, enabling us to make informed forecasts about NFL team season wins.

In addition to the analytical process outlined, we will also develop an interactive program that allows users to input variable values. This program will provide users with the capability to obtain predictions for NFL team season wins and associated confidence intervals based on the selected final model. By doing so, we aim to enhance the practical utility of our analysis, making it more accessible and user-friendly.

2. Data Visualization

» Dataset & Assumptions:

	Team	of_yds_per_g	df_yds_per_g	of_pts_per_g	df_pts_per_g	wins
0	Atlanta Falcons	253.9	276.9	15.0	18.1	10
1	Baltimore Colts	256.7	369.4	14.9	26.3	6
2	Buffalo Bills	289.4	324.2	18.9	22.1	6
3	Chicago Bears	278.7	292.5	15.8	17.1	8
4	Cincinnati Bengals	304.5	289.5	15.8	17.8	5

Assumptions:

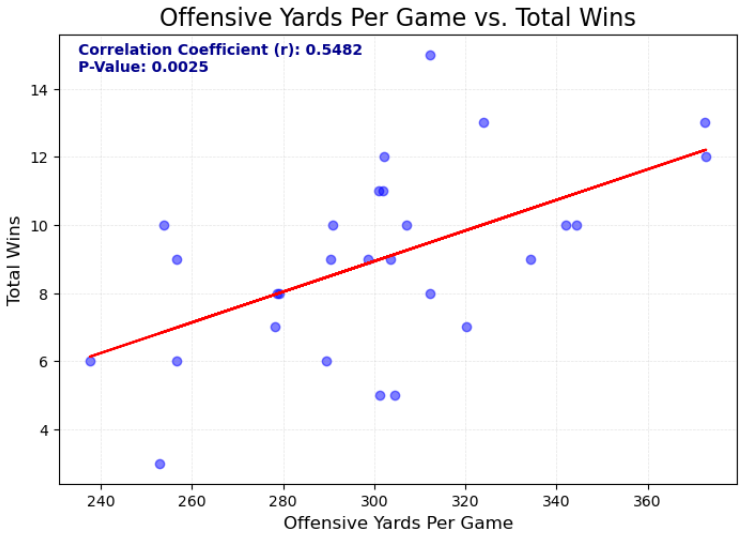
1. Linearity and Mean Centering: The residuals' mean value for various predictor variable sets is zero, implying that the response variable has a linear relationship with the predictor variables.
2. Independence: Observations are independent, and there are no discernible patterns in the residuals.
3. Normality: The residuals for each predictor variable set follow a normal distribution.
4. Homoscedasticity: The variance of regression errors remains constant for all predictor variable values (x).

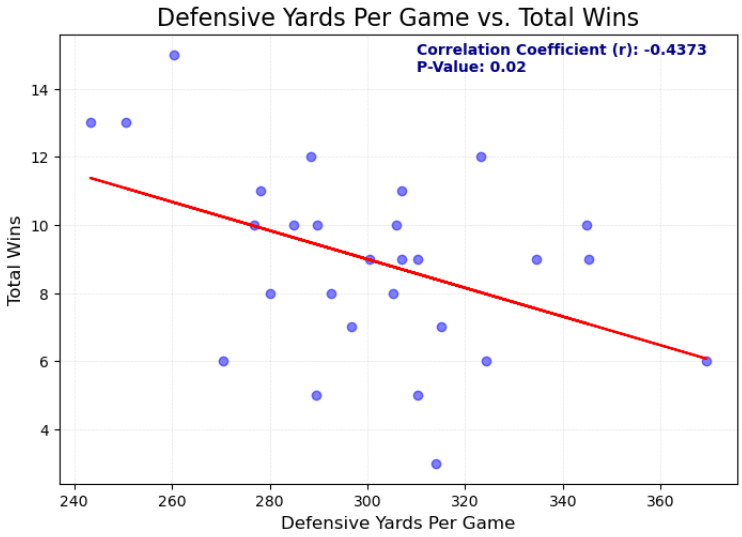
» Potential independent variables for predicting “wins”.

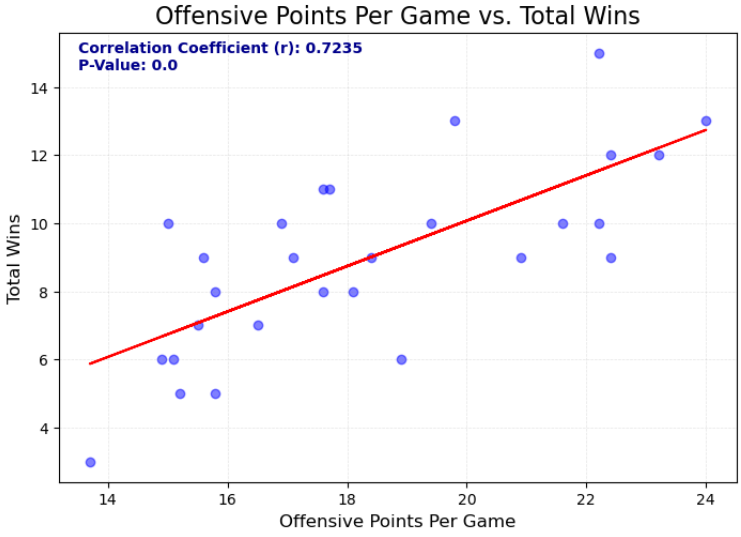
Predicting NFL Team Season Wins	
Potential Independent Variable	**Representation**
of_yds_per_g	Offensive yards per game
df_yds_per_g	Defensive yards per game given up
of_pts_per_g	Offensive points per game
df_pts_per_g	Defensive points per game given up

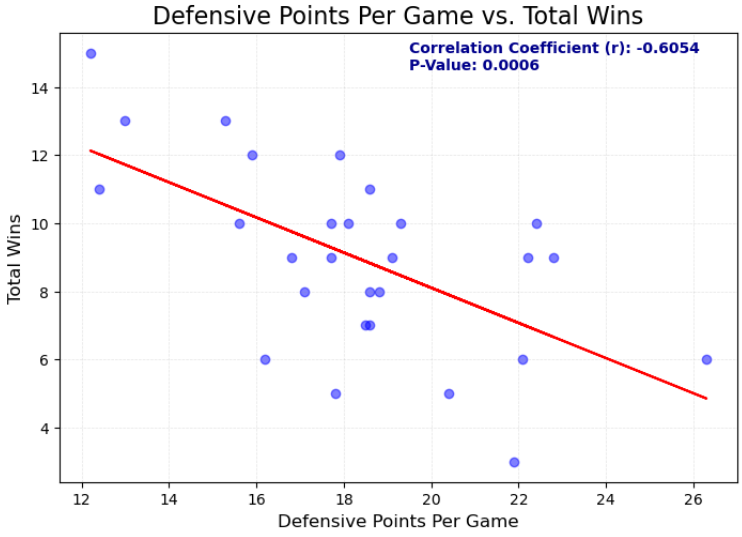
» Scatter Plots for Correlation Visual

The purpose of a scatter plot is to visually examine the relationship between the data points for predictor and response variables in order to identify any discernible patterns. In this context, the focus is on determining the presence of a linear correlation. Each scatter plot will be analyzed to identify patterns that suggest the existence of such a linear correlation.









Correlation Coefficient (r) and P-Values

Variable Correlation to "Wins"		
Variable	**Correlation Coefficient (r)**	**P-Value**
of_yds_per_g	0.5482	0.0025
df_yds_per_g	-0.4373	0.0200
of_pts_per_g	0.7235	0.0000
df_pts_per_g	-0.6054	0.0006

Now we will consider each plot with its associated correlation coefficient and determine it the correlation is statistically significant at a significant level of 5%. We will utilize Parvez Ahammad's interpretation of correlation strength (Jaadi, 2019).

Value of r	Strength Interpretation
|.90| to |1.00|	Very High Correlation
|.70| to |.90|	High Correlation
|.50| to |.70|	Moderate Correlation
|.30| to |.50|	Low Correlation
|.00| to |.30|	Negligible Correlation

Offensive Yards per Game

The relationship between offensive yards gained per game and total season wins is evident in the scatter plot, displaying a discernible positive trend characterized by the upward trajectory of the regression line. This observation is supported by a positive correlation coefficient of 0.5482. It is worth noting, however, that, as highlighted by Ahammad, this correlation coefficient signifies only a moderately positive correlation, a characterization reinforced by the noticeable dispersion of data points around the regression line.

Defensive Yards per Game

The inverse correlation between defensive yards conceded per game and total season wins is evident in the scatter plot, manifested by the descending orientation of the regression line and underscored by a correlation coefficient of -0.4373. It is noteworthy, however, that Ahammad characterizes this correlation coefficient as indicative of a low negative correlation, a classification that aligns with the discernible spread of data points surrounding the regression line. Furthermore, the dispersion of data points appears notably concentrated toward the center of the plot rather than being skewed towards the outer edges.

Offensive Points per Game

The scatter plot depicting offensive points per game versus total wins vividly portrays a positive relationship, evident through the ascending trajectory of the regression line and substantiated by a robust correlation coefficient of 0.723. It is noteworthy that this correlation coefficient, standing as the highest among all variables examined thus far, is categorized as a high positive correlation by Ahammad.

What sets this apart is not just the numerical scale, but the discernible cohesiveness reflected in the comparatively narrower distribution of data points. This contrast from the broader scatter in other analyses accentuates the noticeable relationship between offensive points per game and overall wins, elevating its importance in evaluating team success.

Defensive Yards per Game

Upon scrutinizing the scatter plot illustrating defensive yards per game, a conspicuous negative relationship unfolds. This is underscored by the descending trajectory of the regression line and the negative correlation coefficient of -0.6054, solidifying this discernment. Intriguingly, this correlation coefficient, standing as the second-highest among the quartet of variables under consideration, is qualified by Ahammad as a moderate correlation. The dispersion of data points exhibits a certain breadth, with a notable concentration towards the center of the plot.

Statistical Significance Hypothesis Tests

To assess the statistical significance of this correlation, a hypothesis test has been conducted. The following outlines the step-by-step procedure for the hypothesis test pertaining to the correlation between offensive yards and total wins. Subsequent variables will be addressed with a brief presentation of their respective conclusions.

Offensive Yards per Game

Significance: 0.05

Null: There is no statistically significant correlation between offensive yards per game and total season wins.
Alternative: There is a statistically significant correlation between offensive yards per game and total season wins.

H_0: $\rho = 0$
H_1: $\rho \neq 0$

T-Statistic: 3.34
P-Value: 0.0025

Conclusion: Since the p-value is less than the significance of 0.05 we reject the null hypothesis. The correlation between offensive yards per game and total wins is statically significant.

Defensive Yards per Game

Conclusion: Since the p-value is less than the significance of 0.05 we reject the null hypothesis. The correlation between defensive yards per game and total wins is statically significant.

Offensive Points per Game

Conclusion: Since the p-value is less than the significance of 0.05 we reject the null hypothesis. The correlation between offensive points per game and total wins is statically significant.

Defensive Points per Game

Conclusion: Since the p-value is less than the significance of 0.05 we reject the null hypothesis. The correlation between defensive points per game and total wins is statically significant.

Best Feature Selection

As per the results obtained from sklearn's SelectKBest function, the top two features identified are 'of_pts_per_g' and 'df_pts_per_g.' Expanding our selection to the three best features, we have 'of_yds_per_g,' 'of_pts_per_g,' and 'df_pts_per_g.' Nevertheless, it is important to note that we will not solely rely on this initial feature selection. Our approach will involve constructing multiple models that encompass a diverse range of potential variables. This comprehensive analysis, in conjunction with additional investigations, will guide our final feature selection process.

3. Simple Linear Regression Models: $\hat{y} = b_0 + b_1x$

In a simple linear regression equation, "y" denotes the dependent variable under investigation, which we aim to predict or understand. "x," on the other hand, signifies the independent variable, employed to make predictions regarding the dependent variable "y." In this context, "b_0" serves as the y-intercept, representing the value of "y" when "x" equals zero. It essentially defines the baseline value of the dependent variable when the independent variable exerts no influence. "b_1," meanwhile, denotes the slope of the regression coefficient and quantifies the change in the dependent variable "y" for each one-unit alteration in the independent variable "x."

So, the equation "$\hat{y} = b_0 + b_1x$" essentially defines a linear relationship between the dependent variable “wins” and the independent variable “df_yds_per_g.” The "b_0" term is the baseline value, and "b_1" represents the rate of change in "wins" associated with changes in "df_yds_per_g."

SLR Model Analysis

Single Linear Regression Models and Comparisons									
Model	**Variable**	**Equation**	**R^2**	**T-Stat**	**P-Value**	**F-Stat**	**F-Test**	**MSE**	**RMSE**
SLR Model 1	df_yds_per_g	y = 21.62 - 0.04x	0.191	-2.480	0.020	6.1480	0.0200	5.9202	2.4331
SLR Model 2	of_pts_per_g	y = -3.27 + 0.67x	0.524	5.345	0.000	28.5700	0.0000	3.4879	1.8676
SLR Model 3	df_pts_per_g	y = 18.43 - 0.52x	0.367	-3.879	0.001	15.0400	0.0006	4.6372	2.1534
SLR Model 4	of_yds_per_g	y = -4.56 + 0.05x	0.301	3.342	0.003	11.1700	0.0025	5.1204	2.2628

SLR Model 1

Defensive Yards per Game:

Regression Equation: $\hat{y} = 21.62 – 0.04x$

The hypothesis testing procedure for each model follows a uniform structure, with a shared null and alternative hypothesis presented comprehensively within our analysis. In the forthcoming sections, we will intricately explore the model-specific t-statistics, p-values, and draw corresponding conclusions for the initial model. Subsequent models will be summarized succinctly, focusing solely on the derived conclusions. This streamlined approach is designed for clarity and efficiency, facilitating a concise yet comprehensive presentation of results for each model. For easy reference, the comprehensive listing of test statistics and p-values for all models is conveniently captured in the Single Linear Regression Models Comparisons Table.

Hypothesis F-Tests

Significance: 0.05

Null: No liner relationship exists between defensive yards per game and total wins.
Alternative: A liner relationship exists between defensive yards per game and total wins.

H_0: $b_1 = 0$
H_1: $b_1 \neq 0$

T-Statistic: -2.480
F-Statistic: 6.0480
P-Value: 0.0200

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is statistically significant evidence to support the claim that a linier relationship exists between defensive yards per game and total wins.

Prediction: What is the predicted number of wins given that the Defensive Yards per game for a team is 290 yards?

$\hat{y} = 21.62 - 0.04x$
$\hat{y} = 21.62 - 0.04(290)$
$\hat{y} = 21.62 - 11.6$
$\hat{y} = 10.02$

The predicted number of wins for a team given that the defensive yards given up per game were 290, is 10.02.

Coefficient of Determination:

When examining the relationship between defensive yards per game and total wins, the coefficient of determination reveals a modest value of 0.191. This signifies that approximately 19% of the variability in wins can be explained by the regression model. In simpler terms, the variation in defensive yards per game accounts for only about 19% of the overall variability observed in team wins. This insight underscores the limited explanatory power of defensive yards per game in predicting the diverse factors influencing total victories.

SLR Model 2

Offensive Points per Game F-Test

Regression Equation: $\hat{y} = -3.27 + 0.67x$

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is statistically significant evidence to support the claim that a linier relationship exists between offensive points per game and total wins.

SLR Model 3

Defensive Points per Game F-Test

Regression Equation: $\hat{y} = 18.43 - 0.52x$

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is statistically significant evidence to support the claim that a linier relationship exists between defensive points per game and total wins.

SLR Model 4

Offensive Yards per Game F-Test

Regression Equation: $\hat{y} = -4.56 + 0.05x$

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is statistically significant evidence to support the claim that a linier relationship exists between offensive yards per game and total wins.

- Upon a comprehensive examination of the models, considering their conclusions, R2 values, and other integral components, none of the simple linear regression models stand out as particularly compelling. It is noteworthy to mention that SLR Model 2 exhibited a noteworthy strength during presentation and testing. Nevertheless, a more in-depth exploration into potential multiple regression models becomes imperative for a more nuanced understanding and robust analysis.

4. Multiple Linear Regression Models: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$

In a multiple linear regression equation, " $\hat{y}$ " denotes the dependent variable under investigation, which we aim to predict or understand. " $x_1, x_2, \ldots x_n$" represent the independent variables being used to make predictions regarding the dependent variable. In this context, "b_0" still serves as the y-intercept, representing the value of " $\hat{y}$ " when all independent variables equal zero. It essentially defines the baseline value of the dependent variable when the independent variables exert no influence. "$b_1 \ldots b_n$" denotes the slope of the regression coefficient and quantifies the change in the dependent variable for each one-unit alteration in the independent variable while all other variables remain consistent.

- MLR Model Analysis

Multiple Linear Regression Models and Comparisons

Model	Variables	Equation	R^2	R^2_{adj}	F-Stat	Overall F-Test P-Value	T-Stat	T-Test P-Value	MSE	RMSE
MLR Model 1	df_yds_per_g	y = 7.11 - 0.03x1 + 0.04x2	0.418	0.372	8.985	0.00115	-2.249	0.034	4.2589	2.0637
	of_yds_per_g						3.123	0.004		
MLR Model 2	of_yds_per_g	y = 1.60 + 0.04x1 + 0.02x2 - 0.63x3	0.589	0.537	11.46	7.42E-05	3.501	0.002	3.0101	1.735
	df_yds_per_g						1.087	0.288		
	df_pts_per_g						-3.155	0.004		
MLR Model 3	of_yds_per_g	y = 5.67 -0.003x1 + 0.005x2 + 0.60x3 - 0.44x4	0.747	0.703	16.98	-1.3E-06	-0.234	0.817	1.8514	1.3607
	df_yds_per_g						0.238	0.814		
	of_pts_per_g						3.794	0.001		
	df_pts_per_g						-2.678	0.013		
MLR Model 4	of_yds_per_g	y = -2.97 - 0.002x1 + 0.69x2	0.524	0.486	13.75	9.39E-05	-0.117	0.908	3.486	1.8971
	of_pts_per_g						3.424	0.002		
MLR Model 5	of_yds_per_g	y = 6.47 - 0.41x1 + 0.61x2 -0.004x3	0.746	0.715	23.55	2.49E-07	-4.591	0.000	1.8559	1.3623
	of_pts_per_g						4.104	0.000		
	df_pts_per_g						-0.313	0.757		
MLR Model 6	of_pts_per_g	y = 5.86 - 0.41x1 + 0.58x2	0.745	0.725	36.6	3.74E-08	-4.668	0.000	1.8651	1.3651
	df_pts_per_g						6.100	0.000		

MLR Model 1

Defensive Yards per Game & Offensive Yards per Game:

Regression Equation: $\hat{y} = 7.11 - 0.03x_1 + 0.04x_2$

Coefficient of Determination: R^2: 0.418

R^2_{adj}: 0.372

The adjusted coefficient of determination reveals that 37.2% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis Testing:

The hypothesis testing procedure for each model follows a uniform structure, with a shared null and alternative hypothesis presented comprehensively within our analysis. In the forthcoming sections, we will intricately explore the model-specific t-statistics, p-values, and draw corresponding conclusions for the initial model. Subsequent models will be summarized succinctly, focusing solely on the derived conclusions. This streamlined approach is designed for clarity and efficiency, facilitating a concise yet comprehensive presentation of results for each model. For easy reference, the comprehensive listing of test statistics and p-values for all models is conveniently captured in the Multiple Linear Regression Models Comparisons Table.

Hypothesis F-Tests

Significance: 0.05
Null: None of the independent variables, defensive yards per game or offensive yards per game are significantly different from 0 and therefore a significant linear relationship does not exist.
Alternative: At least one of the independent variables, defensive yards per game or offensive yards per game are significantly different from 0 and therefore a significant linear relationship does exist between at least one of the independent variables.

H_0: $b_1 = b_2 = 0$
H_1: At least one $b_i \neq 0$ for $i = 1$ & 2

F-Statistic: 8.985
P-Value: 0.00115

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables—defensive yards per game or offensive yards per game.

Hypothesis T-Tests

defensive yards per game
Significance: 0.05
Null: The independent variable, defensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive yards per game and total wins.
Alternative: The independent variable, defensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive yards per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_1 = 0$
H_1: $b_1 \neq 0$

T-Statistic: -2.279
P-Value: 0.034

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, defensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive yards per game and total wins when the remaining predictor variables in the model are fixed.

Hypothesis T-Tests

offensive yards per game

Significance: 0.05

Null: The independent variable, offensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between offensive yards per game and total wins.

Alternative: The independent variable, offensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive yards per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_2 = 0$

H_1: $b_2 \neq 0$

T-Statistic: 3.123

P-Value: 0.004

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive yards per game and total wins when the remaining predictor variables in the model are fixed.

Prediction: What is the predicted number of wins given that the offensive yards per game for a team is 307 and Defensive Yards per game is 290?

$\hat{y} = 7.11 - 0.03x_1 + 0.04x_2$

$\hat{y} = 7.11 - 0.03(290) + 0.04(307)$

$\hat{y} = 7.11 – 8.7 + 12.28$

$\hat{y} = 10.69$

The predicted number of wins for a team given that offensive yards per game is 307 and defensive yards per game is 290 is 11 wins.

Prediction Results and Confidence Intervals:

Mean Expected: 9.57

Expected Mean Interval: (8.66, 10.48)

Prediction Interval: (4.98, 14.16)

MLR Model 2

Offensive Yards per Game, Defensive Yards per Game & Defensive Points per Game:

Regression Equation: $\hat{y} = 1.60 + 0.04x1 + 0.02x2 - 0.63x3$

Coefficient of Determination: R^2: 0.589

R^2_{adj}: 0.537

The adjusted coefficient of determination reveals that 53.7% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis Testing:

Hypothesis F-Tests

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables— offensive yards per game, defensive yards per game & defensive points per game.

Hypothesis T-Tests

Offensive Yards per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive yards per game, is significantly different from 0 and therefore a significant linear relationship does exist between offensive yards per game and total wins when the remaining predictor variables in the model are fixed.

Defensive Yards per Game

Conclusion: Since the p-value is more than the significance level of 0.05, fail to reject the null hypothesis. The independent variable, defensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive yards per game and total wins when the remaining predictor variables in the model are fixed. This variable should be dropped from the model.

Defensive Points per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, defensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive points per game and total wins when the remaining predictor variables in the model are fixed.

MLR Model 3

Offensive Yards per Game, Defensive Yards per Game, Offensive Points per Game & Defensive Points per Game:

Regression Equation: $\hat{y} = 5.67 - 0.003x_1 + 0.005x_2 + 0.60x_3 - 0.44x_4$

Coefficient of Determination: R^2: 0.747

R^2_{adj}: 0.703

The adjusted coefficient of determination reveals that 70.3% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis F-Tests

Significance: 0.05

Null: None of the independent variables, defensive yards per game or offensive yards per game are significantly different from 0 and therefore a significant linear relationship does not exist.

Alternative: At least one of the independent variables, defensive yards per game or offensive yards per game are significantly different from 0 and therefore a significant linear relationship does exist between at least one of the independent variables.

H_0: $b_1 = b_2 = b_3 = b_4 = 0$
H_1: At least one $b_i \neq 0$ for $i = 1, 2, 3, 4$

F-Statistic: 16.98
P-Value: 1.31e-06 (0.0000)

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables— offensive yards per game, defensive yards per game, offensive points per game & defensive points per game.

Hypothesis T-Tests
offensive yards per game
Significance: 0.05
Null: The independent variable, offensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between offensive yards per game and total wins.
Alternative: The independent variable, offensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive yards per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_1 = 0$
H_1: $b_1 \neq 0$

T-Statistic: -0.234
P-Value: 0.817

Conclusion: Since the p-value is more than the significance level of 0.05, fail to reject the null hypothesis. The independent variable, offensive yards per game, is not significantly different from 0 and therefore a significant linear relationship does not exist between offensive yards per game and total wins when the remaining predictor variables in the model are fixed. This variable should be dropped from the model.

Hypothesis T-Tests
defensive yards per game
Significance: 0.05
Null: The independent variable, defensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive yards per game and total wins.
Alternative: The independent variable, defensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive yards per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_2 = 0$
H_1: $b_2 \neq 0$

T-Statistic: 0.238
P-Value: 0.814

Conclusion: Since the p-value is more than the significance level of 0.05, fail to reject the null hypothesis. The independent variable, defensive yards per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive yards per game and total wins when the remaining predictor variables in the model are fixed. This variable should be dropped from the model.

Hypothesis T-Tests
defensive points per game
Significance: 0.05
Null: The independent variable, defensive points per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive points per game and total wins.
Alternative: The independent variable, defensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive points per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_3 = 0$
H_1: $b_3 \neq 0$

T-Statistic: 3.794
P-Value: 0.001

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, defensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive points per game and total wins when the remaining predictor variables in the model are fixed.

Hypothesis T-Tests
offensive points per game
Significance: 0.05
Null: The independent variable, offensive points per game, is not significantly different from 0 and therefore no significant linear relationship exists between offensive points per game and total wins.
Alternative: The independent variable, offensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive points per game and total wins when the remaining predictor variables in the model are fixed.

H_0: $b_4 = 0$
H_1: $b_4 \neq 0$

T-Statistic: -2.678
P-Value: 0.013

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive points per game and total wins when the remaining predictor variables in the model are fixed.

Note: Despite the failure of two of the individual variable T-Test, the predictions are provided below for review purposes only.

Prediction: What is the predicted number of wins given that the offensive yards per game for a team is 307 and Defensive Yards per game is 290, offensive points per game is 17 and defensive points is 20?

$\hat{y} = 5.67 - 0.003x_1 + 0.005x_2 + 0.60x_3 - 0.44x_4$
$\hat{y} = 5.67 - 0.003(307) + 0.005(290) + 0.60(17) - 0.44(20)$
$\hat{y} = 5.67 – 0.921 + 1.45 + 10.2 – 8.8$
$\hat{y} = 7.599$

The predicted number of wins for a team given that offensive yards per game is 307 and defensive yards per game is 290 is 8 wins.

Prediction Results and Confidence Intervals:
Mean Expected: 7.35
Expected Mean Interval: (6.22, 8.47)
Prediction Interval: (4.04, 10.65)

MLR Model 4

Offensive Yards per Game & Offensive Points per Game:
Regression Equation: $\hat{y} = -2.97 - 0.002x_1 + 0.69x_2$

Coefficient of Determination: R^2: 0.524
R^2_{adj}: 0.486

The adjusted coefficient of determination reveals that 48.6% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis Testing:

Hypothesis F-Tests

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables— offensive yards per game & offensive points per game.

Hypothesis T-Tests

Offensive Yards per Game

Conclusion: Since the p-value is more than the significance level of 0.05, fail to reject the null hypothesis. The independent variable, offensive yards per game, is not significantly different from 0 and therefore a significant linear relationship does not exist between offensive yards per game and total wins when the remaining predictor variables in the model are fixed. This variable should be dropped from the model.

Offensive Points per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive points per game and total wins when the remaining predictor variables in the model are fixed.

MLR Model 5

Offensive Yards per Game, Offensive Points per Game & Defensive Points per Game:

Regression Equation: $\hat{y} = 6.47 - 0.41x_1 + 0.61x_2 - 0.004x_3$

Coefficient of Determination: R^2: 0.746

R^2_{adj}: 0.715

The adjusted coefficient of determination reveals that 71.5% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis Testing:

Hypothesis F-Tests

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables— offensive yards per game, offensive yards per game & defensive points per game.

Hypothesis T-Tests

Offensive Yards per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive yards per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive yards per game and total wins when the remaining predictor variables in the model are fixed.

Offensive Points per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive points per game and total wins when the remaining predictor variables in the model are fixed.

Defensive Points per Game

Conclusion: Since the p-value is more than the significance level of 0.05, fail to reject the null hypothesis. The independent variable, defensive points per game, is not significantly different from 0 and therefore no significant linear relationship exists between defensive points per game and total wins when the remaining predictor variables in the model are fixed. This variable should be dropped from the model.

MLR Model 6

Offensive Points per Game & Defensive Points per Game:

Regression Equation: $\hat{y} = 5.86 - 0.41x_1 + 0.58x_2$

Coefficient of Determination: R^2: 0.745

R^2_{adj}: 0.725

The adjusted coefficient of determination reveals that 72.5% of the variability in total team season wins can be effectively explained by this regression model. This statistic highlights the model's capacity to account for the variations observed in the outcome variable, providing valuable insight into its explanatory power.

Hypothesis Testing:

Hypothesis F-Tests

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. There is sufficient evidence to support that a significant linear relationship does exist between at least one of the independent variables— offensive yards per game & defensive points per game.

Hypothesis T-Tests

Offensive Yards per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, offensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between offensive points per game and total wins when the remaining predictor variables in the model are fixed.

Defensive Points per Game

Conclusion: Since the p-value is less than the significance level of 0.05, reject the null hypothesis. The independent variable, defensive points per game, is significantly different from 0 and therefore a significant linear relationship exists between defensive points per game and total wins when the remaining predictor variables in the model are fixed.

Multiple Linear Regression Models Progress

Model	Model Progress
MLR Model 1	MLR Model 1 passed the F-Test and all individual T-Test. We move forward to analyze this model further.
MLR Model 2	MLR Model 2 passed the F-Test but failed one individual T-Test. We will not move forward with this model.
MLR Model 3	MLR Model 3 passed the F-Test but failed two individual T-Test. We will not move forward with this model.
MLR Model 4	MLR Model 4 passed the F-Test but failed one individual T-Test. We will not move forward with this model.
MLR Model 5	MLR Model 5 passed the F-Test but failed one individual T-Test. We will not move forward with this model.
MLR Model 6	MLR Model 6 passed the F-Test and all individual T-Test. We move forward to analyze this model further.

5. Model Selection

Multiple Linear Regression Models and Comparisons

Model	Variables	Equation	R^2	R^2_{adj}	F-Stat	Overall F-Test P-Value	T-Stat	T-Test P-Value	MSE	RMSE
MLR Model 1	df_yds_per_g	y = 7.11 - 0.03x1 + 0.04x2	0.418	0.372	8.985	0.00115	-2.249	0.034	4.2589	2.0637
	of_yds_per_g						3.123	0.004		
MLR Model 6	of_pts_per_g	y = 5.86 - 0.41x1 + 0.58x2	0.745	0.725	36.6	3.74E-08	-4.668	0.000	1.8651	1.3651
	df_pts_per_g						6.100	0.000		

Single Linear Regression Models and Comparisons

Model	Variable	Equation	R^2	T-Stat	P-Value	F-Stat	F-Test	MSE	RMSE
SLR Model 2	of_pts_per_g	y = -3.27 + 0.67	0.524	5.345	0.000	28.5700	0.0000	3.4879	1.8676

Upon thorough examination of the proposed models, the top three have been identified for further in-depth analysis and ultimate selection. Notably, MLR Model 6 emerges as the superior choice, boasting not only the highest R^2 value but also the lowest mean squared residual (MSE) and Residual Standard Error (RMSE) values. An intriguing observation lies in the fact that the variables incorporated into MLR Model 6 align precisely with the two best features predicted by sklearn's SelectKBest function. This convergence adds a layer of validation to the model's efficacy, reinforcing its standing as the prime candidate for final selection.

CONCLUSION:

My preference among the analyzed models leans towards MLR Model 6 for several compelling reasons. Firstly, the two variables incorporated into this model, offensive points per game and defensive points per game, boast the highest correlation coefficient values of 0.72 and 0.61, respectively. This correlation is visually affirmed by their respective scatter plots, where data points exhibit a more concentrated alignment along the regression line.

Secondly, MLR Model 6 not only successfully passed the overall F-Test but also excelled in individual variable T-Tests. Furthermore, it demonstrated the highest R^2_{adj} value at 0.745, coupled with the lowest RMSE value at 1.37, solidifying its standing as the most robust choice.

Finally, after dividing the data into training and testing sets, MLR Model 6 exhibited noteworthy performance by producing the lowest RMSE values on both the training and testing sets, measuring at 1.48 and 1.12, respectively. This consistency in performance across different data sets adds another layer of confidence in the model's predictive capabilities.

However, despite the model's potential to generate predictions for total team season wins, it's crucial to acknowledge the limitations of relying solely on one year's data for creating an effective and accurate predictive model. To enhance the model's reliability, incorporating data from multiple years and exploring additional relevant variables in future analyses would be imperative.

References

Jaadi, Z. (2019, October 16). *Eveything you need to know about interpreting correlations*. Medium. https://towardsdatascience.com/eveything-you-need-to-know-about-interpreting-correlations-2c485841c0b8