Written Report

Introduction and Data

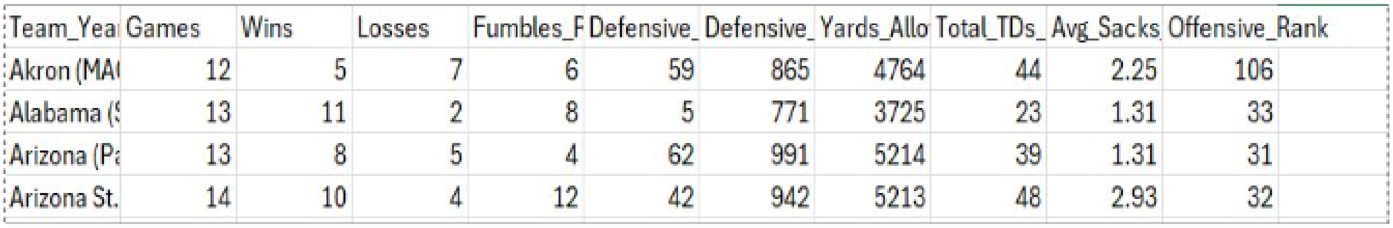
Introduction

Our project motivation stems from our interest in Georgia Tech's football team's recent losses. We want to be able to improve their chances of winning through using statistical analysis. There is a popular saying that "defense wins championships", so we decided that this idea would be something to investigate further. We want to verify through statistics that this saying is a guiding principle that teams should follow to improve their performance. This project will give statistics on much defense really matters in making wins happen.

Our research question is "Does defense win championships? and"What specific defense aspects matter?". Understanding these questions will give coaches a better understanding on where to improve. We also want to understand if specific aspects of defense matter more than others for winning.

Data

The dataset is from Kaggle. The data includes the team statistics for all of the Football Bowl Subdivision level teams from 2013 to 2023 with around 145 different features. We decided for our testing we would use these variables found in figure 1 The time period was cut down to 2013 to 2020 due to corruption of the columns in future years.



Team_Yea	Games	Wins	Losses	Fumbles_F	Defensive_	Defensive_	Yards_Allo	Total_TDs_	Avg_Sacks	Offensive_Rank
Akron (MA(	12	5	7	6	59	865	4764	44	2.25	106
Alabama (S	13	11	2	8	5	771	3725	23	1.31	33
Arizona (Pa	13	8	5	4	62	991	5214	39	1.31	31
Arizona St.	14	10	4	12	42	942	5213	48	2.93	32

Figure 1: Key variables used in analysis

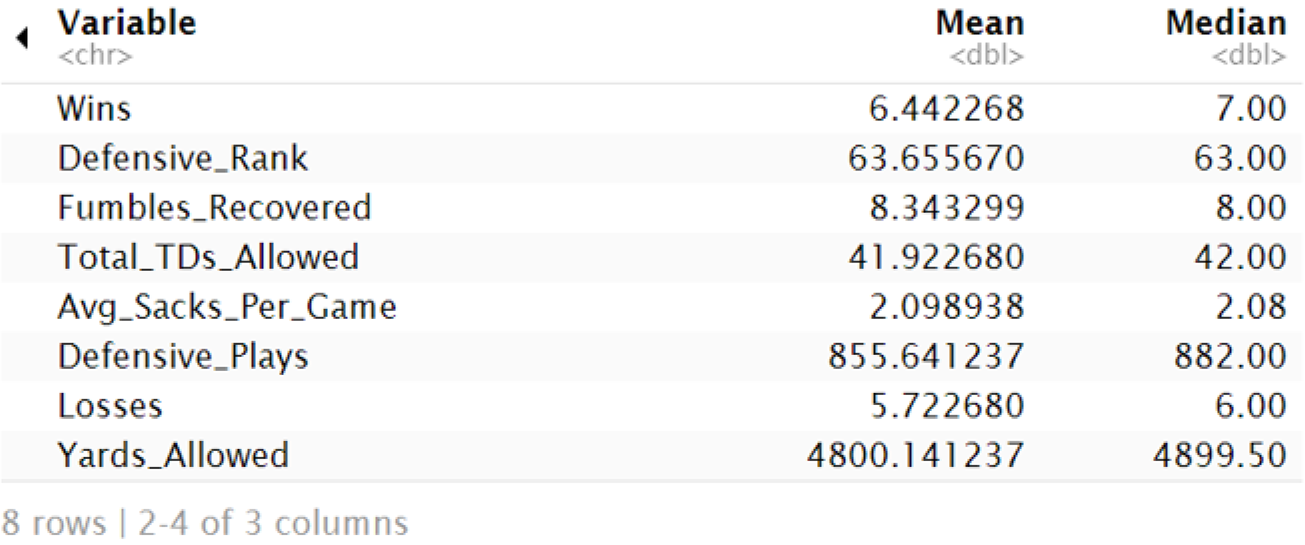
Below is a table of defintions of key variables

Variable	Description
Team_Year	Unique identifier combining team name and season year (e.g., "Alabama (SEC)_2013")
Games	Total number of games played in the season
Wins	Number of games won
Losses	Number of games lost
Fumbles_Recovered	Number of fumbles recovered by the defense

Variable	Description
Defensive_Rank	Team’s overall defensive ranking (lower number = better defense)
Defensive_Plays	Total number of defensive plays made
Yards_Allowed	Total yards allowed to opponents
Total_TDs_Allowed	Total touchdowns allowed to opponents
Avg_Sacks_Per_Game	Average number of quarterback sacks per game
Offensive_Rank	Team’s overall offensive ranking (lower number = better offense)

Descriptive Statistics

Here are some of the descriptive statistics highlights using the information regarding college football from the 2013 through 2020 seasons.

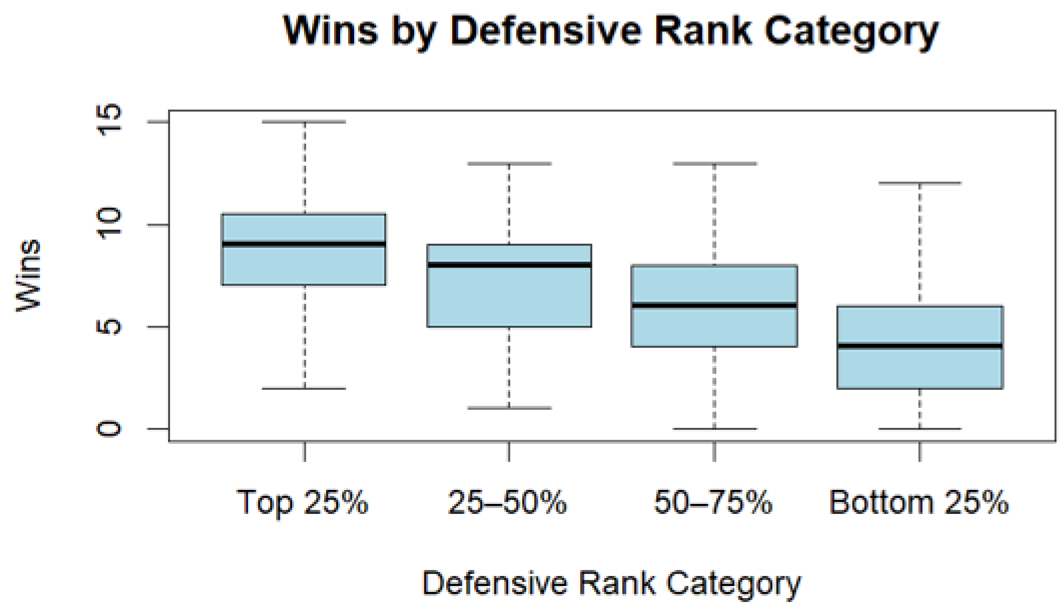


Variable <chr>	Mean <dbl>	Median <dbl>
Wins	6.442268	7.00
Defensive_Rank	63.655670	63.00
Fumbles_Recovered	8.343299	8.00
Total_TDs_Allowed	41.922680	42.00
Avg_Sacks_Per_Game	2.098938	2.08
Defensive_Plays	855.641237	882.00
Losses	5.722680	6.00
Yards_Allowed	4800.141237	4899.50

8 rows | 2-4 of 3 columns

Most of the medians and means for these values are pretty close, meaning the data is not very skewed in either direction. Wins and losses are an interesting data point as not every team played the same number of games. One data point with a somewhat noticeable skew is defensive plays, being skewed towards fewer defensive plays. This makes sense as “splash plays” are less common.

Here is a boxplot using quartiles of defensive rank.



BP

The median of wins gets progressively lower with each lower quartile. There is an apparent correlation between defensive quality and wins. There are some outliers in each category, showing that there are other factors involved in winning outside of defense, but that defensive rank does contribute.

Methodology

To test the strength of the relationship between the number of wins, the dependent variable, and defensive statistics, the independent variables, we used linear regression tests. The linear regression visualizations, regression equations, and the R-squared value helped answer the research question, “Does defense win championships,” by showing how defensive statistics impact the number of wins, providing the answer of whether or not defense correlates to winning championships. These tests quantify how much each defensive statistic, namely fumbles recovered, defensive rank, defensive plays, yards allowed, total touchdowns allowed, and average sacks per game, had the most effect on wins.

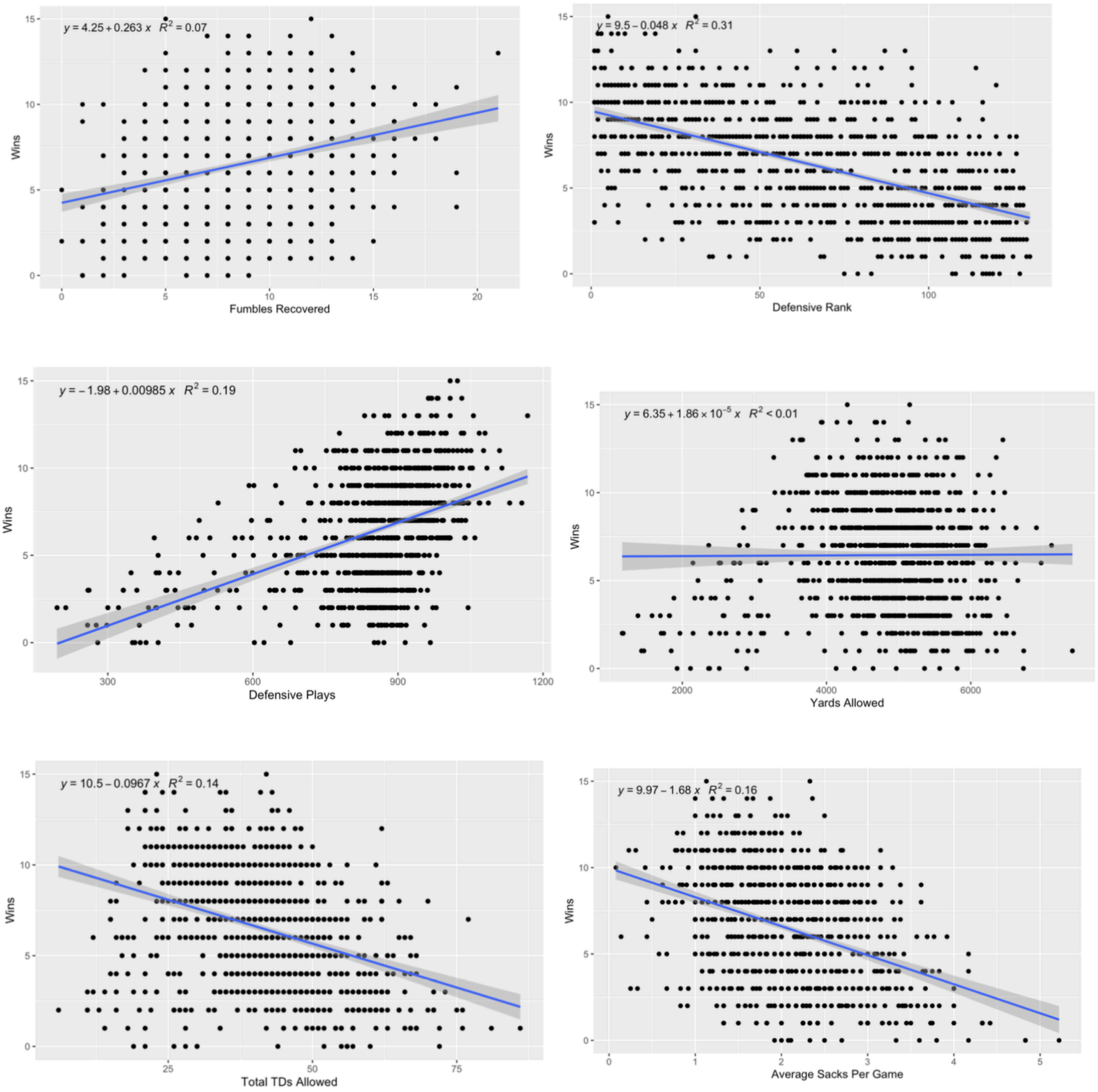
For context, the linear regression model visualizes the strength of each relationship, and the regression equations show mathematical relationships, quantifying how changes in the independent variable affects the dependent variable. Positive coefficients signify a direct relationship while negative coefficients represent indirect relationships.

On top of this, we used a t-test to determine the probability that the standard deviation of (defensive rank)/(wins) was equal to zero. To get the dataset for this test, we grouped 110 datapoints into groups of 5 (excluding the 111th datapoint that required dividing by zero) and calculated the standard deviation of each group, which resulted in a dataset of 10 values. Then, we used the built-in t-test to determine the probability that the mean standard deviation was equal to zero. Our null hypothesis for this test was that defensive rank affects all teams equally, and our alternative hypothesis was that defensive rank has a significantly different level of impact on each team.

Results

The results of the linear regression testing were as follows:

The regression equation for fumbles recovered is Wins = 4.2453 + 0.2633x, and the R-squared value is 0.074. The regression equation for defensive rank is Wins = 9.4977 + -0.048x, with 0.3122 as the R-squared value. For defensive plays, the equation is Wins = -1.9824 + 0.0098x, and the R-squared value is 0.1934. Wins = 6.3529 is the regression equation for yards allowed, with an R-squared value of zero. For total touchdowns allowed, the regression equation is Wins = 10.4973 + -0.0967x, and the R-squared value is 0.1397. Lastly, the regression equation for average sacks per game is Wins = 9.9669 + -1.6792x, and the R-squared value is 0.1593. The graphs with all three components are shown below:



Above is an image of the linear regressions graphs from the linear_regressions.qmd file.

After looking at these graphs, the results point to several interpretations about the defensive statistics' effects on wins. Each visualization and regression equation displays the expected direct or indirect relationships, except for yards allowed, which reveals a flat line with no coefficient in the equation. The line is flat not due to a coding issue or data error but rather because almost no linear relationship exists between yards allowed and wins. Meanwhile, the other variables show logical relationships. For example, defensive plays have a positive equation coefficient and apositive visualization slope, revealing its direct relationship with number of wins. This finding is logical because more defensive plays results in fewer points awarded to the opposing team, reducing the likelihood of being outscored, which results in a higher chance of winning the game.

Furthermore, no R-squared value reached 0.5, meaning the defensive statistics lack a strong correlation to the number of wins. In other words, though they play a role, defensive statistics alone do not determine wins. Fumbles recovered has the second least R-squared value because fumble recovery mostly relies on chance. The most shocking result, however, is yards allowed having an R-squared value of zero, but there are explanations. First, in NCAA football, games are won by points. Two defenses could allow the same yards but different numbers of points, resulting in varying numbers of wins. Additionally, yardage poorly measures defensive effectiveness because number of points and situational plays matter more. Because of these factors, the R-squared value, linear regression visualization, and regression equation become null.

Ultimately, the linear regression visualizations and regression equations show that better defensive statistics possess a strong relationship with a higher number of wins, but the R-squared values show that they do not correlate strongly enough to be the driving force.

The results of the t-test were as follows: p-value = 0.00005986 95% confident that the true standard deviation is between 15.97672 and 32.21666 Because the p-vale was so low, we were able to reject the null that defensive rank impacts each team's win rate equally. In fact, because the true standard deviation is likely between 15.97672 and 32.21666, defensive rank likely has an incredibly variable impact on each football team.

Discussion

Based on the statistical tests run on the data, it seems that good defense does positively impact wins in college football, but it is not the only factor that determines wins. The linear regressions show strong correlation between good defense and a high number of wins, however the R^2 values show that the correlation is too weak to conclude that defense is the main factor influencing whether a college football team will win their game. It is good for teams to work on a strong defense, but it seems there are other areas of the team that can't be ignored. Football data presents a challenge in seperating variables, since variables such as yards allowed and touch downs allowed are definitely connected to each other. There is also no way to claim causation as there are many unnacounted for variables in football that could affect performance, such as loud fans or bad weather. The only thing we can look at regarding college football is correlation. One way we could improve this limitation is by looking further into which variables have the highest correlation to number of wins and testing those. Another thing we could do is test more football datasets to ensure that our results are as strong as possible. Potential issues pertaining to the reliability of the data are that it excludes the most recent years of 2021 to present due to a corrupted dataset for later years, and only goes up to 2023. Football is always developing as a sport in different ways, so there may be factors that weren’t present in the past, like they are now. The 2020 season was also affected by COVID, so that data may not be reflective of football generally. The thought that our statistical analysis using linear regression was appropriate allowed us to find out how useful each individual defensive metric was in terms of getting wins. This allowed us to find out information like yards allowed, not being a useful metric for wins.

Conclusion

For future work, we could test other variables such as weather, which conference the team is in, and whether the game was close or a blowout. We could look at offensive metrics and doing linear regressions on them and see if they have the similar trends as the defensive metrics. We could also investigate even more of the features in the dataset that appear to be related to defense such as redzone defensive rank. We could also adjust the wins to be a different metric so that it correctly weighs how much win is worth based on the strength of the overall team so that the win metric isn’t inflated from weaker team wins.