College Football Analysis

Group 6: Maximus Sambucetti, Ricky Li, Hessa Alnuaimi

12/9/23

Introduction

College football is one of America's most popular and fastest-growing sports. With the introduction of official tracked football statistics in 1959, analysts and fans could better comprehend the game and make fundamental discoveries. This research provided insightful knowledge on patterns and critical elements essential to success. Our project explored last year's college football season (2022). The dataset contains 145 different statistics on all 131 teams in the NCAA Division I subdivision.
Our research questions were:

> What is the correlation between offense and defense?
> Which Collegiate Conference was the best for the 2022 season?
> Do better defensive teams get more wins?
> Which defensive features translate into wins?
> Do special teams contribute to wins?

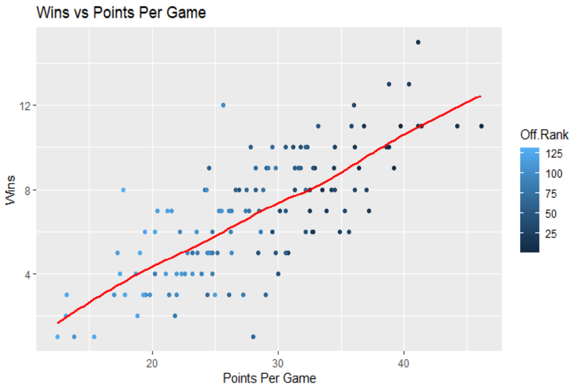
We implemented linear regression, multivariable regression, p-value testing, and graphical analysis to answer these questions. Using these techniques and methods, we hoped to gain a deeper understanding of which team aspects helped generate the most wins possible and unravel the dynamics within each team to gain valuable insights that can inform strategic decisions and enhance the overall competitiveness of college football.

Results

Question 1: What is the correlation between offense and defense?

When beginning our football data analysis, we decided to start by identifying the correlation between offense and defense. The goal was to get a sense of the broader impact each feature had when it came to getting wins. The easiest way to get straight to the point was to compare a defensive and offensive stat to wins. These two stats were points per game and yards allowed per play.

The first visualizations (Figure 1-2) are graphs that find the correlation between the points earned per game and the amount of wins. It is crucial to score points in football, especially when trying to win. It also further identifies the type of offensive rank each data point has, thus demonstrating that the higher the offensive rank, the more points they earned per game and the more wins they received. The first depiction shows the best curve fit line, which is linearly increasing. The best-fit line is then identified to emphasize further the pattern and trend that points mean more wins.



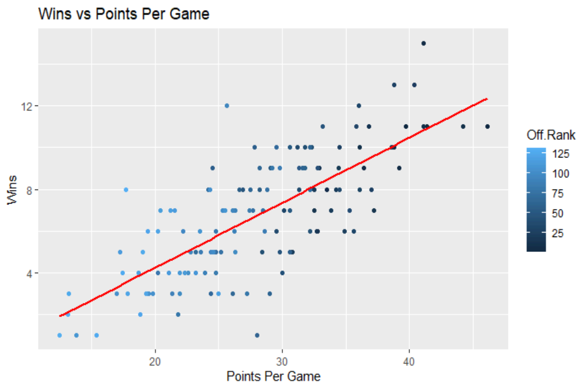
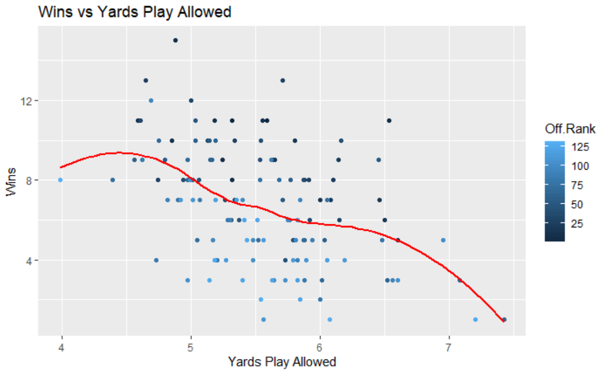


Figure 1 **Figure 2**

The second visualization (Figure 3-4) depicts the correlation between wins and the number of yards allowed per play. The most crucial stat when considering which football team has a solid defense is how many yards they are allowed per play. The fewer yards a team allowed an opponent to get, the more wins they could receive. The curve fit of the first depiction suggests a declining trend in which it plateaus for a little while in between points. Although the correlation could be stronger, the best-fit line shows a declining trend. Each point is also identified based on the team's defensive rank and how teams with a lower rank still had more wins.



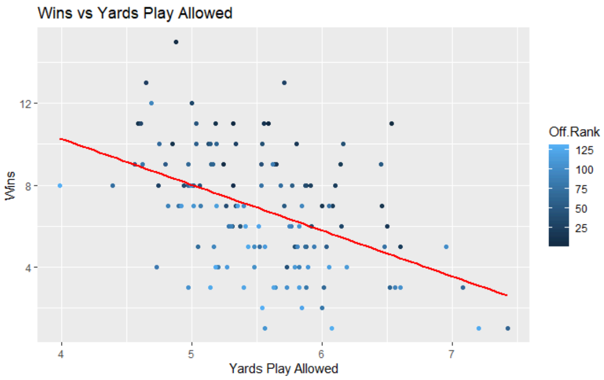
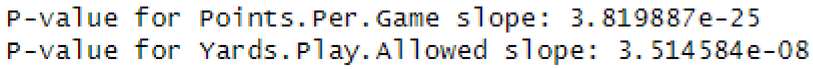


Figure 3 **Figure 4**

Finally, to further assess the correlations between the offensive stat and defensive stat to wins, we decided to identify the p-value of the best-fit line for both graphs. The p-value would help determine which stat was essential for gaining wins. Each p-value should be below 0.05 to show a significant impact, which was the case for both graphs.



```
P-value for Points.Per.Game slope: 3.819887e-25
P-value for Yards.Play.Allowed slope: 3.514584e-08
```

Figure 5: These are the two p-values for the best-fit lines of both graphs. States that the p-value of points per game was 3.8e-25 and the p-value for yards per play allowed was 3.5e-08, demonstrating a significant impact.

Question 2: Which Collegiate Conference was the best for the 2022 season?

Many factors come into play when determining the best collegiate conference, especially in the 2022 season. However, based on the previous visualizations (Figures 1-2), the best stat is to compare them to the points per game in question one. The correlation between points and wins is very high and demonstrates the probability of getting the most wins in a season.

The first visualization (Figure 6) is a more in-depth display of the correlation between points per game and wins, the offensive rank each team had, and what region each team was from. It is a positive trend that suggests again that more points per game meant more wins. However, there were a lot of SEC, Big Ten, and Pac-12 towards the upper right side of the figure. It is a well-organized depiction of each team, and where they are placed with the rest of the groups, but to get a deeper understanding, we decided to rely on the numbers.

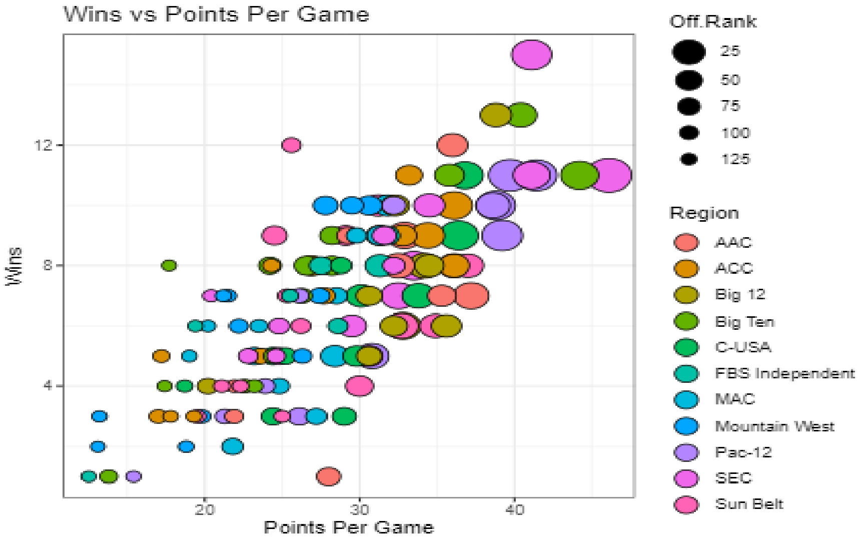
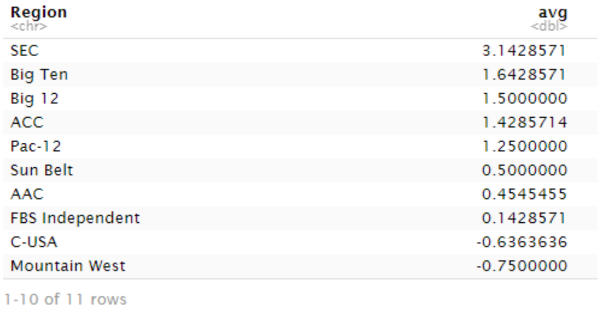


Figure 6

The next step was calculating and filtering the data to understand which conference performed the best overall. We decided to group by region and then summarize and organize the data based on the average win-to-loss ratio per region—this way, the best conferences could be ordered from most successful to least.



Region <chr>	avg <dbl>
SEC	3.1428571
Big Ten	1.6428571
Big 12	1.5000000
ACC	1.4285714
Pac-12	1.2500000
Sun Belt	0.5000000
AAC	0.4545455
FBS Independent	0.1428571
C-USA	-0.6363636
Mountain West	-0.7500000

1-10 of 11 rows

Figure 7: This is a descending list of each region's average win-to-loss ratios. It shows the SEC conference as the most successful, with Big Ten and Big 12 coming in second and third.

Question 3: Do better defensive teams get more wins?

Defense is essential to a successful football team, so do better defensive teams get more wins? For this question, we narrowed our observations down to just the college teams in the conference Michigan State is in, the Big 10.

The first visualization (Figure 8) we made to answer this question was a bar graph. This type of graph allows us to compare the number of wins for each team directly. Each bar is then colored based on the defensive rank and reordered from highest to lowest number of wins. With this, we can determine that teams ranked 1-30 (represented in blue) typically have more wins than higher-ranked teams (higher, in this case, means a defensive rank above 30).

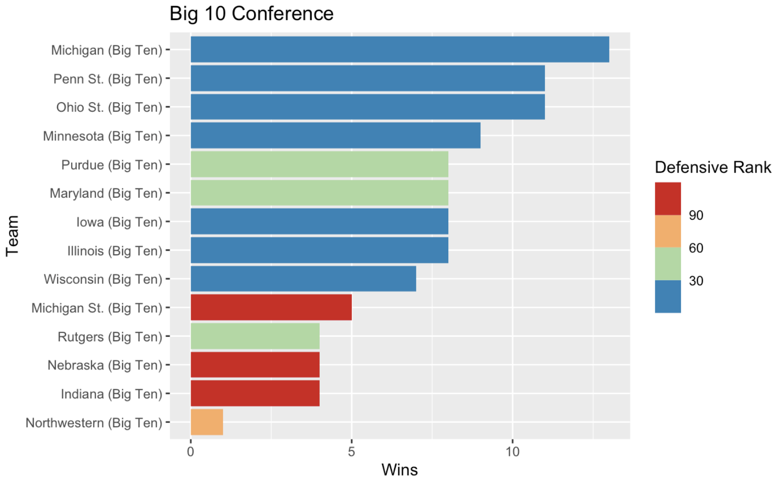


Figure 8

The second visualization (Figure 9) was a scatter plot with defense rank on the x-axis and win percentage on the y-axis. Each data point is colored based on the team, and the size of the points corresponds to the number of yards allowed per play (larger points mean more yards allowed per play, and smaller points represent fewer yards allowed). From this plot, we can see that lower-ranked teams (defensive rank of 1-25) have a higher win percentage than teams ranked higher, which is the same conclusion we came to from the bar graph. Additionally, we can see that lower-ranked teams typically allow fewer yards per play, as the data points corresponding to low ranking are smaller in size compared to data points higher in rank.

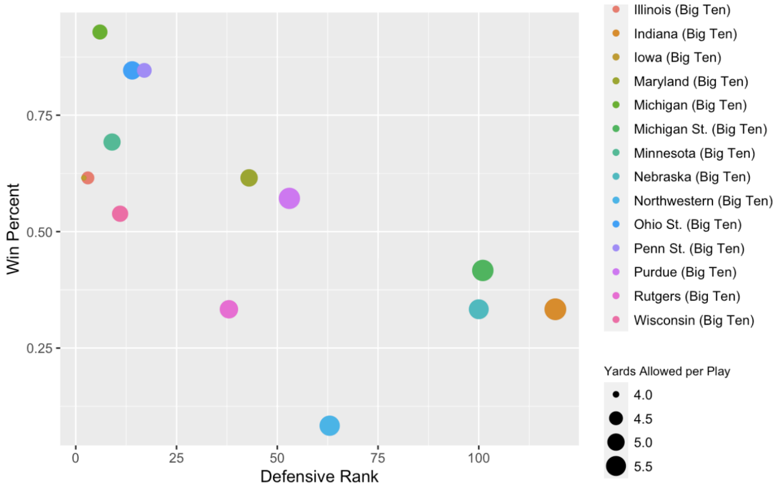


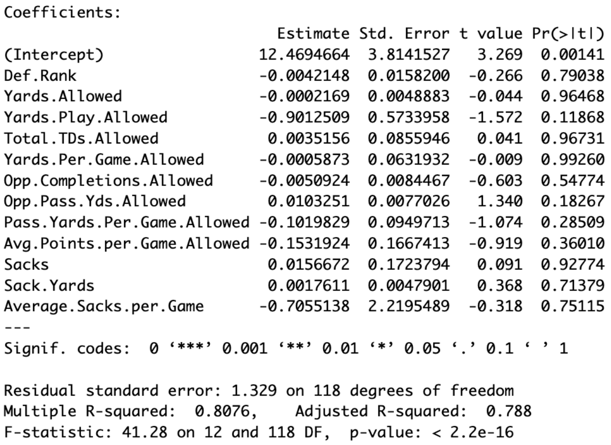
Figure 9

Question 4: Which defensive features translate into wins?

The first step in answering this question was to subset the main football data for only defensive statistics. After combing through all the features, we produced a defense data frame consisting of 12 defensive features for all 131 teams. These features included the total and averages for yards, touchdowns/points, completions, and sacks.

With this data frame, we can create a multivariable regression model that uses all the defensive features as independent variables to predict the number of wins, the dependent variable. After fitting using the lm() method, we get a model with an R-squared of 0.8076 and an adjusted R-squared of 0.788. Our values show that the model can capture around eighty percent of the variance in the dependent variable (wins). We can also see that the adjusted R-squared penalizes our explained variance because we use multiple features as our independent variables.

Another aspect to consider about this full feature model is the p-values. Each feature's p-values are considered not statistically significant (if we set the significance level at <= 0.05). This means that the defensive variables we used could be better predictors of wins. Even if the model can capture a high variance of the dependent variable, if the features used are not statistically significant, then the model is not considered viable.

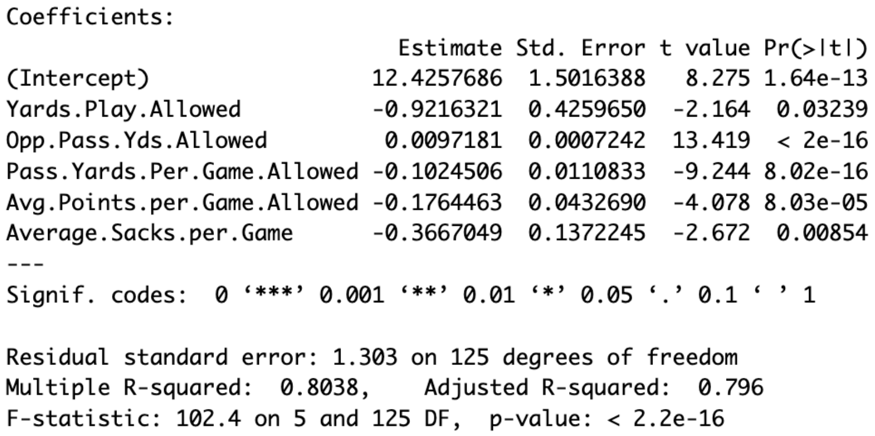


```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               12.4694664  3.8141527   3.269  0.00141
Def.Rank                  -0.0042148  0.0158200  -0.266  0.79038
Yards.Allowed             -0.0002169  0.0048883  -0.044  0.96468
Yards.Play.Allowed        -0.9012509  0.5733958  -1.572  0.11868
Total.TDs.Allowed          0.0035156  0.0855946   0.041  0.96731
Yards.Per.Game.Allowed    -0.0005873  0.0631932  -0.009  0.99260
Opp.Completions.Allowed   -0.0050924  0.0084467  -0.603  0.54774
Opp.Pass.Yds.Allowed       0.0103251  0.0077026   1.340  0.18267
Pass.Yards.Per.Game.Allowed -0.1019829  0.0949713  -1.074  0.28509
Avg.Points.per.Game.Allowed -0.1531924  0.1667413  -0.919  0.36010
Sacks                      0.0156672  0.1723794   0.091  0.92774
Sack.Yards                 0.0017611  0.0047901   0.368  0.71379
Average.Sacks.per.Game    -0.7055138  2.2195489  -0.318  0.75115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.329 on 118 degrees of freedom
Multiple R-squared:  0.8076,    Adjusted R-squared:  0.788
F-statistic: 41.28 on 12 and 118 DF,  p-value: < 2.2e-16
```

Figure 10: Summary output of full feature lm() model. Note that all p-values for all features are greater than 0.05 and, therefore, are not statistically significant.

Since we are addressing the question of which defensive features are the most important, it makes sense to implement a feature reduction method to determine the most essential features that contribute to wins. We used backward selection on the full feature model from above. By using backward selection, we can remove the least contributive feature for each step until the model reaches a point where all features are statistically significant. After applying this method, we get a reduced feature model with an R-squared of 0.8038 and an adjusted R-squared of 0.796. Comparing these values to the full feature model, the reduced feature model's R-squared is only slightly smaller (a 0.038 difference is negligible). However, its adjusted R-squared is higher due to fewer predictor variables. This means that the reduced feature model is just as good or even better at capturing the variance in the dependent variable (wins).

If we look at the reduced feature model's p-values, we can see that every feature's p-value is less than 0.05 and, therefore, statistically significant. This means that the reduced feature model has captured the most critical defensive features: yards per play allowed, opponent pass yards allowed, passing yards allowed per game, average points allowed, and the average sacks per game.



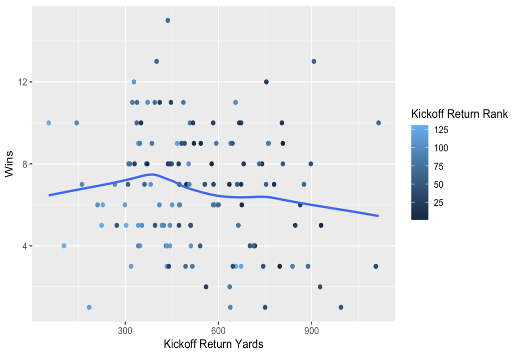
```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                12.4257686  1.5016388   8.275 1.64e-13
Yards.Play.Allowed         -0.9216321  0.4259650  -2.164  0.03239
Opp.Pass.Yds.Allowed        0.0097181  0.0007242  13.419  < 2e-16
Pass.Yards.Per.Game.Allowed -0.1024506  0.0110833  -9.244 8.02e-16
Avg.Points.per.Game.Allowed -0.1764463  0.0432690  -4.078 8.03e-05
Average.Sacks.per.Game     -0.3667049  0.1372245  -2.672  0.00854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.303 on 125 degrees of freedom
Multiple R-squared:  0.8038,    Adjusted R-squared:  0.796
F-statistic: 102.4 on 5 and 125 DF,  p-value: < 2.2e-16
```

Figure 11: Summary output of reduced feature lm() model. Note that after feature reduction using backward selection, the R-squared decreases slightly, and the adjusted R-squared increases (meaning more explained variance of the dependent variable (wins) is captured by the model). P-values for all features are less than 0.05, which means all features are significantly significant.

Question 5: Do special teams contribute to wins?

One essential part of football games often overlooked is the contribution of special teams to a team's performance. This is mainly because these units are only on the field for a small fraction of time compared to the offensive/defensive. Special teams comprise many units, but our research only focuses on the kickoff and punt return teams.

For the kickoff return team, we wanted to look at the relationship between kickoff return yards (the total number of kickoff return yards obtained by a team) and kickoff return rank compared to the number of wins a team gets. To get a visual of these relationships, a scatter plot was created with the total number of kickoff return yards on the x-axis and the total number of wins on the y-axis. A line of best fit is plotted to model the relationship between these variables. The data points are also colored based on kickoff return rank, with darker blue corresponding to a higher kickoff return rank and lighter blue to a lower kickoff return rank.



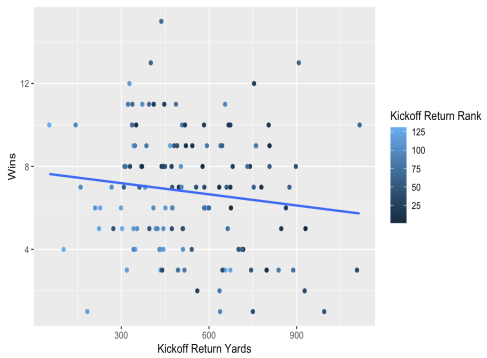
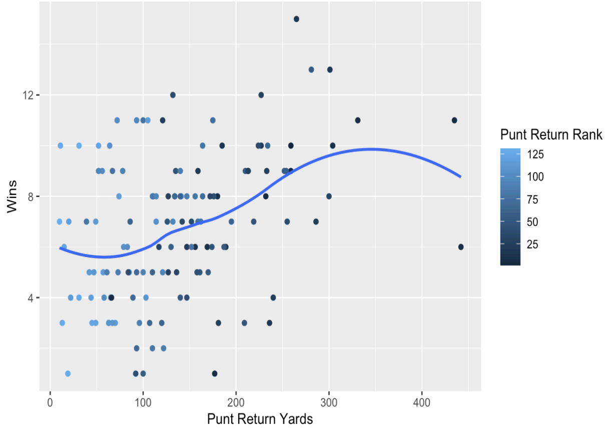


Figure 12 Figure 13

From the scatter plot, we can see that the curve of best fit (Figure 12) increases from yards 75 to 375 and then decreases for yards greater than 375. This suggests that teams with a low number of kickoff return yards (< 375) would get more wins and teams with a higher number of kickoff return yards would get fewer wins, a very counter-intuitive finding as it is often observed that more yards (in an offensive perspective) would correlate with more wins, not the other way around. If we look at the line of best fit (Figure 13), this negative relationship is supported due to the negative slope present.

However, if we incorporate kickoff return rank into the discussion, we can observe that teams that get more yards are typically ranked higher in kickoff return rank. This can be seen on the scatter plots where darker dots appear more often on the right side of the plot, where the x-axis corresponds to more kickoff return yards. Using these observations, the kickoff return teams that get more yards negatively correlate with the number of wins a team gets.

The second special team we considered in our project was the punt return team. Using a scatter plot, we can visualize the relationship between punt return yards and punt return rank versus the number of wins. Punt return yards are plotted on the x-axis, and the number of wins is plotted on the y-axis, with the color of the data points corresponding to the punt return rank for each observation (team). Like the previous plot, darker blue colors correspond to a higher punt return ranking and lighter blue to a lower ranking.



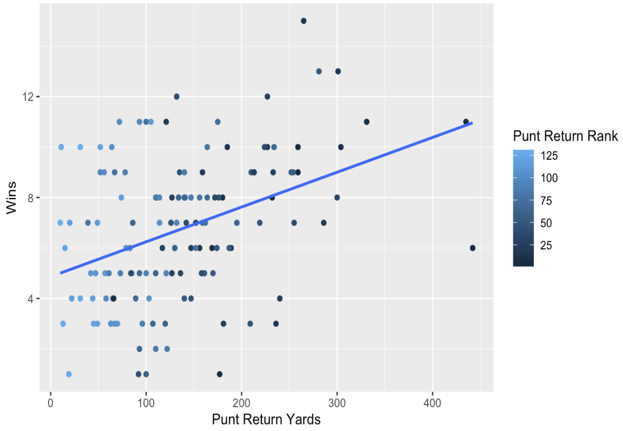


Figure 14 Figure 15

From this scatter plot, we can see that the curve of best fit (Figure 14) increases as the number of punt return yards increases. This means that teams with more punt return yards typically get more wins than teams with small return yards. This trend follows true until the total yards reach around 350, where the number of wins decreases with more yards. If we look at the line of best fit (Figure 15), we can see that the trend is only positive. This means that a team gets more wins as the number of yards increases. Incorporating punt return rank into our analysis, we can determine that teams who get more yards on punt returns are associated with a higher rank. This is represented by the darker blue dots appearing more often on the right side of the plots. This also means that teams ranked higher in punt return will win more. This is supported by the fact that darker blue teams are the only data points right of the 300-yard mark, and from our previous analysis, we concluded that more yards equals more wins. If teams ranked 50 or lower only appear in high-yardage areas (right of the 300-yard mark), we can say that teams ranked higher are more likely to get more wins.

Conclusion

Summarization of Results

After a thorough analysis of the collegiate football season of 2022, it was determined that although there was high evidence that defense does increase win percentage, the offense also played a significant role and increased the chances of having a higher win percentage.

Each offensive and defensive stat played a tremendous role in increasing the amount of wins a team received. Especially when it came to points per game, many teams that scored higher gained even more wins. When comparing this feature to each region, we found supporting evidence that a specific region performed better during the 2022 season—after further assessing the average win-to-loss ratio of each area, the SEC had the most competitive and successful conference. On top of this, after an even deeper look into the defensive stats, it was concluded that the most important stats for a defensive team to win were the passing yards allowed per game, the number of sacks, and the average points allowed per game. In conclusion, when it came to special teams performances, such as punt return yards, they too helped gain more wins in the long run and were crucial to a team's success.

Based on all the data collected and the answers found, many different features of a college football team impact their performance for that season. These team features increase a team's win percentage and are one of the many reasons why having a well-balanced team in college football is essential.

Critiques, Limitations, Validity of Data

The only limitation we saw when organizing and analyzing the data was that any models created to predict wins for the next season would be flawed. Although a model could predict the possible wins based on the current team data, too many unpredictable variables could come into play for the model to be accurate. These variables include the unpredictability of injuries, team dynamics, fouls, etc. There are also no individual player stats for each football team. This is crucial when identifying how good a conference is and how good a team will be. Players play a massive role in how well the team performs for the season. Overall, these were the limitations we ran into, but the validity of the data could also be improved if we used the data sets from the other seasons. It would boost the conclusion evidence and give a more well-rounded and solid answer to some questions.

Suggestions for Future Research

To improve our analysis of the football data and how important defensive features are for winning games, we could have compiled the previous seasons before 2022 into more accurate and precise comparisons. This way, the more data used, the more efficient a model will be and the more accurate the conclusions will be.

References

"College Football Team Stats Seasons 2013 to 2022", https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019/code

"College Football Exploratory Analysis", https://www.kaggle.com/code/nicholaseby/college-football-exploratory-analysis