**College Football Analysis:**

*What features of a college football team translates into wins?*

## Introduction

College football is a sport that is deeply rooted within American culture, captivating millions of individuals with its long-standing traditions, iconic rivalries and passionate supporters. Unlike many sports, its prominence is evident in its TV rankings during the fall season, illustrating the national fascination with the sport. Despite the popularity of the sport, having access to comprehensive team statistics across the past decade can be challenging due to scattered data, acronyms and structural changes within the sport.

Our project utilises the ‘College Football Team Stats Seasons 2013 to 2023’ dataset, taken from Kaggle, which contains detailed statistics for all Division 1 NCAA college football teams across a variety of measures. Such measures include, but are not limited to, offensive, defensive, turnover, and red zone performance. The college football Team Stats dataset offers over 145 variables per team from 2013 to 2023, which allows for a comprehensive analysis of the factors that drive success in college football.

With the consolidated dataset, we will be able to uncover trends and relationships that have been speculated upon, such as the role of defense for teams in winning championships as well as the characteristics that have been present in the most successful collegiate conferences over the past decade.

### Research Questions

Our primary focus is to answer the question: "What features (variables) of a football team translate into wins?" Additionally, we aim to explore:

1. Which collegiate conference is the ‘best’?
2. Does defense really affect win rate?

These research questions and aims should provide us with both broad insights and specific conclusions about the features that drive success in college football from teams to conferences and gameplans.

### Anticipated Findings

In this investigation, we anticipate to highlight the key variables that strongly correlate with team success including turnover margins, defensive efficiency and offensive yardages. By using a systematic approach, we aim to develop a parsimonious model that explains the likelihood of team wins and accounts for multicollinearity and missing data. Moreover, our analysis of collegiate conferences may highlight patterns of measures of success that are dominant and potentially lead to insights of the success of the conference. Last but not least, the investigation into the defensive performance of teams should lead to meaningful insights to answer the longstanding debate of whether “defense wins championships” particularly in the last decade.

## Results

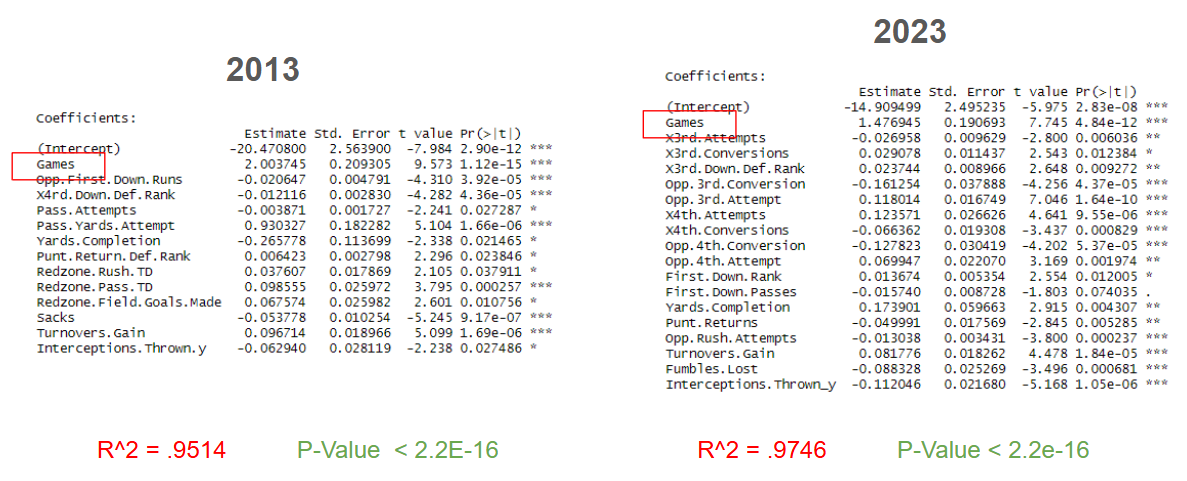
### 1. What features of a college football team translates into wins?

Methodology

The main problem to overcome in feature selection was the number of variables prior to interacting with the dataset then performing a linear model on the data. That is, over 140 variables were reduced to a relatively parsimonious model containing less than a dozen variables. The first step was to eliminate the non-numerical and multicollinear variables (i.e. the variables that had strong correlations with each other). This was done by eliminating variables reaching a 0.8 correlation threshold or higher using a correlation matrix and the findCorrelation function within R. Additionally, we removed variables that had 0 variance. These two methods prepared the data for backwards elimination, which iteratively cycled through the variables to find the “best fit” for the model and the most relevant predictor variables. Lastly, we had Chat GPT manually exclude predictors with p-values above the 5% significance threshold. This made the model even more parsimonious and improved the probability that the remaining variables are statistically significant.

Results

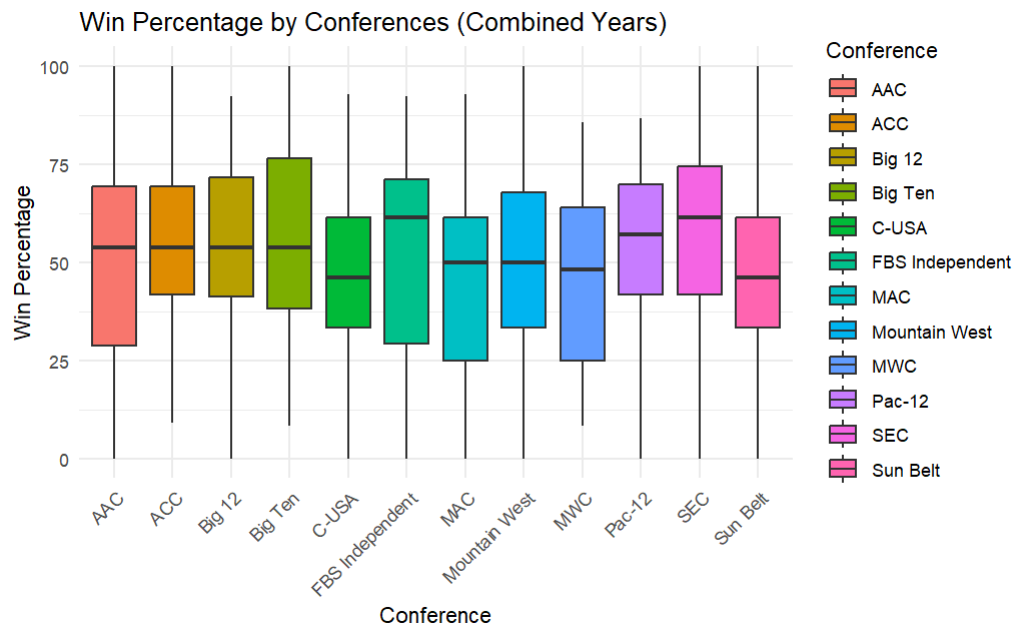
Below are the resulting models for 2013 and 2023 after following the process specified above. “Games” took the lead as the predictor variable, the R2 values were both greater than 95%, and all the p-values showed statistical significance. This shows that the number of games for a team is the most predictive variable out of the lot.



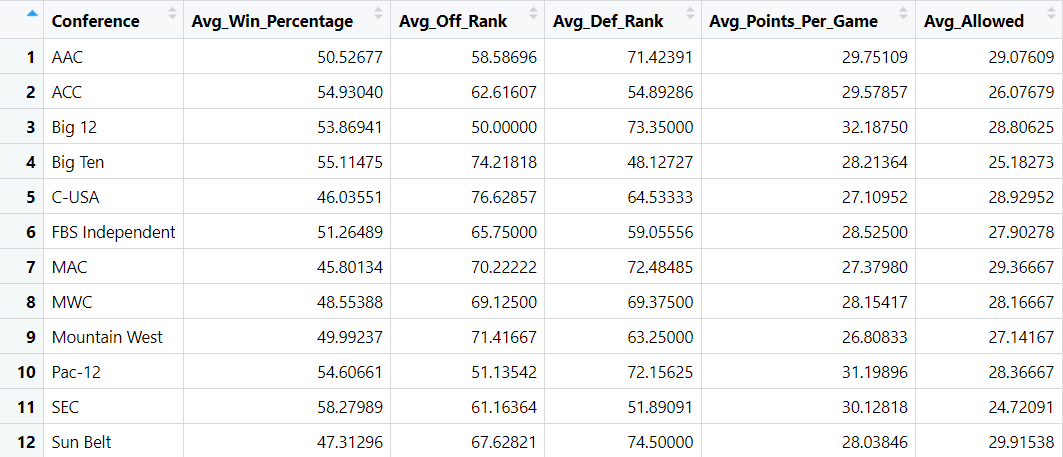
### 2. Which collegiate conference is the ‘best’?

Methodology

Determining the most skilled college football conference is a subjective question that we had to make quantitative in order to come up with an unbiased answer. We did this in R, and used the libraries tidyverse, dplyr, and ggplot. We started with all data sets from 2013 to 2023, and eventually got down to only using 2013-2020, as the last three years of the data were incompatible with the rest. In order to clean the data, we had to separate two columns that were joined: the university name and the conference. This was one of the first things we did, as the conference was the basis of our question. We calculated the win percentage of each conference for each year, and used that to compute a multiple regression model. The analysis of this model aimed to reveal the relationship between a college football team's win percentage and a set of predictor variables over several years of data. Hence, we determined how much variation in win percentage can be explained by the conference, rankings, points scored, and points allowed, per year. We chose these variables because they had high correlation with successful teams.



From there, we began to plot our results. We were able to create a multiple regression model function so we could run every year through the model, as one. We plotted the Win Percentage by conference for the combined years. This was done as a boxplot, and needed cleaning (was hard-coded).



I think the most useful thing to come out of this part of our project was the summary statistics for each conference. This outputted our chosen (highly correlated) variables with its data by conference. This table was easy to read, and very user-friendly, and allowed for a more sophisticated way to analyze it. We chose to do a weighted scoring system. We chose this method because we did not feel that average points allowed (for example) was as important as win percentage. Hence, we weighted the win percentage as 30% (0.3), due to its importance. Both rankings at 20% (0.2), and both PPG (Points per game) at 15% (0.15). We normalized these values, accounting for when a lower vs higher number was better. For example, we want rank to be lower, and points scored per game to be higher. We used the values from our summary, and from there were able to determine which conference was the “best”.

Results

Using our weighted scoring system, we got these results for the top four:

**SEC**: 0.800

**Pac-12**: 0.546

**Big Ten**: 0.418

**Big 12**: 0.350

We then concluded that the SEC was the ‘best’ collegiate football conference, followed by the Pac-12, and then the Big Ten.

### 3. Does defense really affect win rates

Methodology

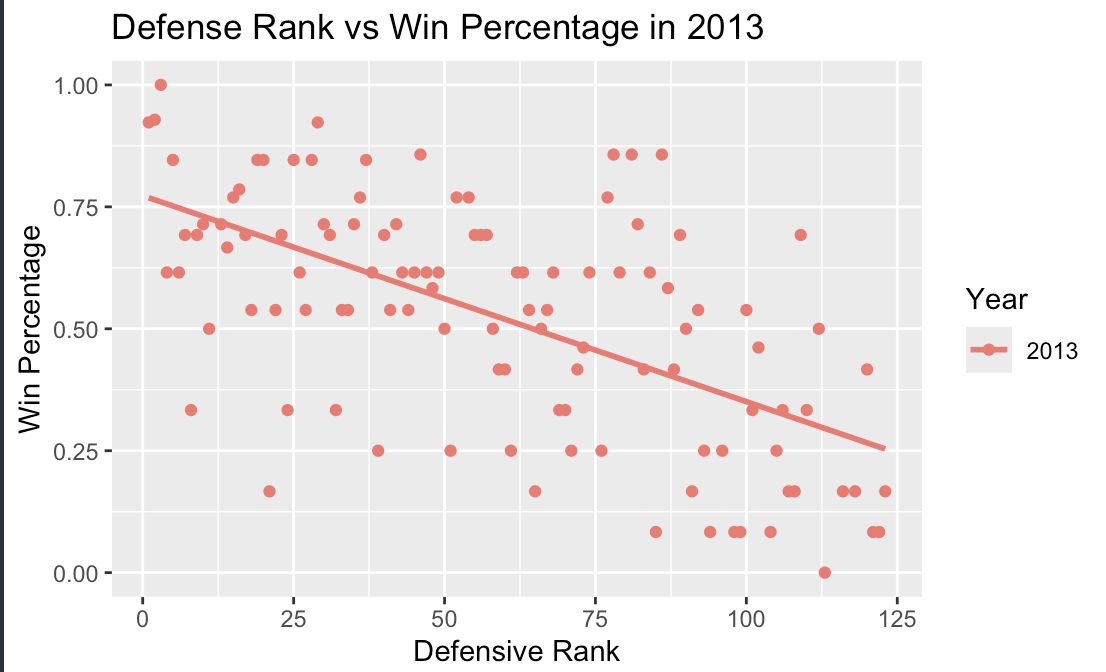
Due to the abundance of data we’re given, we have multiple ways of determining an answer to this question and that is exactly what we did. It’s important to note that the testing for this hypothesis only occurs from 2013-2020 because, as noted previously, the datasets from 2021 through 2023 were fundamentally different. What we first did for this was load in the data for each year 2013-2020 into their respective dataset and then create a win percentage(win\_pct) column for upcoming tests.



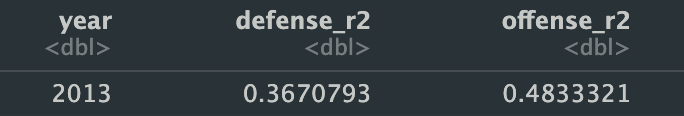
After this I made eight different plots comparing win percentage and defensive ranking as it is the most concrete way of showing what defense is better in this dataset. Over each year’s plot, I used the lm() method to fit the scatter plot data into a linear model to better show what the data ‘should’ look like. While this solution gives us a visual representation of the question, we also deemed it necessary to get concrete numbers for our results and decided to show each year’s regression values in a tidy table. It’s essentially using the same method from the first part but in a more numerical form along with giving us p-values. Lastly, we wanted to see how much of the win percentage is due to the defense and compare it to the offense to determine whether defense or offense is more important than the other. What we did there is create a linear regression again of defense ranking and win rate while also creating an offensive ranking and win rate regression. We put this in a dataset to each corresponding year and then pulled out the r-squared value individually in a tidy table.

Results

From our findings, it does appear that defense absolutely correlates to an increase in win percentage as shown from the table and graph below:



Additionally while defensive ranking does explain a portion of win percentage, offensive ranking seems to explain win-rate at a higher percentage:



## Conclusion

### Summarised Findings

In conclusion, our analysis throughout this project revealed key insights into the factors that contribute to college football success through the exploration of our three key research questions. In terms of discovering which features of a football team translates into wins, the number of games played and losses were revealed as the strongest predictors of the number of wins that a team achieves. While examining college football conferences, the SEC was discovered to rank as the highest conference, driven particularly by a combination of factors including their win percentage and defensive strength, with the Pac-12 and Big Ten completing the Top 3 conferences. Finally, the investigation into determining the role of defense proved that a higher defensive ranking strongly correlates with a higher win rate, supporting the belief that defense plays a critical role in winning championships. Our findings provided valuable insights into the dynamics and drivers of success in college football, offering a foundation for further exploration as well as decision-making within College Football in the US.

### Evaluation

Reliability and Validity of Data

All the data in the dataset was taken from the NCAA stats website, which is the most reliable source the data could have been obtained from. The data source thus is both comprehensive and consistent. Nevertheless, since the data is collected only from one source, which runs the college football games, the data could perhaps be independently verified or cross-validated. In terms of validity, the models did exhibit high R-squared values, indicating a good fit to the dataset data. Nevertheless, the validity of these models on unseen data is untested which limits its predictive validity.

Critique of Methods

In terms of our data selection and preprocessing, there are several strengths in our project. One of the major strengths of our methodology was the use of backward elimination to build parsimonious models. This ensured that the models were easier to interpret and more efficient. On the other hand, in the data preprocessing steps early in our methodology, although it is thorough, may have led to overfitting as indicated by the high R-squared values across multiple models. Whereas, for our modelling choices, linear regression models were utilised appropriately due to the continuous nature of our dependent variable, which was wins. However, other techniques perhaps could have provided additional insights and a better account for non-linear relationships. This is a suggestion for improvement in the future.

Appropriateness of Statistical Analysis

The stepwise regression approach, using backwards elimination, aligns with the objective of identifying the key predictors, especially when determining what features of a football team predicts win and determining what is the best conference. However, this method may sometimes lead to the overfitting of data, especially with our dataset that contains numerous predictors. Additionally, the decision to focus on p-values for the elimination of variables might overlook predictors that may be meaningful and may indirectly contribute to success through interactions with other variables.

Limitations

In evaluating our project, there are several limitations that particularly stood out to us. Firstly, our analysis at several points leaves out data from 2020-2023 due to formatting issues, restricting the scope of the data. Expanding the temporal scope could enhance the analysis and offer more comprehensive insights. Moreover, the utilisation of the linear regression approach assumes linear relationships between variables which may not be accurate across all the predictors. Upon later reflection, the variables prior to modeling were not scaled which would have affected the performance of the model. Lastly, more rigorous statistical tests could strengthen our claims about the performance of college football conferences compared to our assigned weight scale for the key predictors.

Suggestions for Improvement

Despite the strengths of our project, there are also several areas for improvement that can be done. On one hand, our analysis could benefit from deeper background research to improve our elimination of variables in our models. Specifically, we should investigate why ‘games’ emerged as the leading predictor and whether this is driven by unaccounted factors or an underlying trend.

Secondly, our analysis of the three key questions was limited by the incomplete utilisation of the dataset. Specifically, there were several times in which data from 2020-2023 were not incorporated due to issues in different formatting and missing information. Including all the available years of data would improve the comprehensive nature of the analysis and ensure the results are more representative of the entire time period.

External factors, such as changes in the NCAA rules, game-playing styles and advancement in technology may have influenced our findings, but this was not explicitly accounted for in our analysis. There have also been conference realignments that have occurred several times between 2013 and 2023 which could have impacted competition levels and the metrics we explored. Incorporating all the contextual changes could refine our interpretations and provide a more nuanced understanding of the college football trends observed.

Last but not least, we can consider exploring additional predictors such as team-specific metrics to improve the generalizability of our findings. Cross-validation to better assess the robustness of our model would also be useful to reduce the risk of overfitting. It would also be interesting and beneficial if machine learning models were incorporated to not only improve predictive accuracy but also perhaps reveal hidden interactions between variables.

### Future exploration

Besides the improvements that can be done in future iterations of the project, there are also several intriguing questions that could be explored using our dataset along with similar datasets to further explore trends in College Football. The three future exploration questions we came up with are:

1. What is the impact of coaching changes on team success?
2. How does recruitment quality correlate with team performance?
3. What role does home-field advantage play in determining outcomes?

The first question that could be explored in the future relates to the impact of coaching changes to teams. This can investigate whether teams perform better following a change in their head coach and or key coaching staff members.

The next question we believe will be interesting for further exploration relates to how the recruiting of football players to college football teams correlates with team performance. In doing so, we would be able to analyse the relationship between recruitment rankings, such as player star ratings, and on-field success over time.

Last but not least, the final question we suggested for further research surrounds investigating the degree to which home-field games give, or not, an advantage to teams for success. By examining the impact of the location and attendance of games on win percentages across conferences and teams over the last decade, we would be able to gain a variety of insights into College football and the culture surrounding it.

## 

## References

ESPN. (n.d.). *College football conference realignment tracker: 2023* season. Retrieved December 3, 2024, from<https://www.espn.com/college-football/story/_/id/35804245/cfb-conference-realignment-tracker-2023>

Gallini, J. (2019). *College football team stats: Seasons 2013 to 2023* [Dataset]. Kaggle. Retrieved from<https://www.kaggle.com/datasets/jeffgallini/college-football-team-stats-2019>

GeeksforGeeks. (n.d.). *Calculate the interquartile range in R programming – IQR function*. Retrieved December 3, 2024, from<https://www.geeksforgeeks.org/calculate-the-interquartile-range-in-r-programming-iqr-function/>

OpenAI. (2024). *ChatGPT (December 3, 2024 version) [Large language model]*. Retrieved from<https://chat.openai.com>

Syracuse New Times. (n.d.). *What makes college football an integral part of U.S. culture?* Retrieved November 15, 2024, from<https://syracusenewtimes.com/what-makes-college-football-an-integral-part-of-us-culture/>

Stack Overflow. (n.d.). [https://stackoverflow.com](https://stackoverflow.com/)