Introduction:

College football is vastly popular across the United States, as it rakes in several millions of spectators every year. With the high-energy and competitiveness of college football (or perhaps college sports in general), there's no doubt why it's so popular, especially amongst the students.

Our dataset, which was created by Jeff Gallini, provides a detailed description of the multitude of different statistics that college football has to offer. From a team's rank, to its kickoff return yardage, to turnover margin, this dataset provides plenty of statistics and data to work with.

The primary/overarching question that we asked ourselves was: What makes a winning college football team? Going more in-depth, we found ourselves falling down a rabbit hole of a range of questions that could help us in our project. We asked questions along the lines of:

- Does offense matter more than defense (or vice versa)?
- How does offensive yardage and touchdowns measure up with wins?
- How big of a factor does special teams play in winning games?
- Can we correlate outside aspects, such as conferences?

We anticipate to find that teams with high-powered offenses are more likely to prosper in college football than teams without a top-ranked offense, or even in comparison to teams with top-ranked defenses. This is our hypothesis because we've watched several teams with top-ranked offenses that will almost always win numerous games (i.e. Alabama, Clemson, Georgia, Louisiana, etc.).

Methods:

To test our hypothesis, as our questions in the introduction above have alluded to, we plan to observe the dataset according to how college football teams win games. In doing so, we split-up the observations to analyze the statistics among four groups: offense, defense, special teams, and other/outside aspects. This way, the analysis of winning games in college football will be ordered and placed in an organized fashion.

Splitting up the data will offer a more clear picture of how each side of the football (offense, defense, special teams, etc.) factors into winning games. Observing the

statistics, depending on each aspect of the sport, and creating plots and models to visualize trends and patterns will provide clear evidence to either prove or disprove our hypothesis that we'd made previously. Much of the methodology involved in this process belongs to data wrangling and tidying, which is largely done through packages like tidyverse, dplyr, ggplot2, and infer.

Then, qualitatively interpreting each plot or model will showcase the relative importance of some variable(s) to winning college football games, or at least lead to interesting observations that may not have been otherwise obvious.

Results: Make sure you're answering questions that are presented in the introduction. If questions that you're answering are not there, put them there.

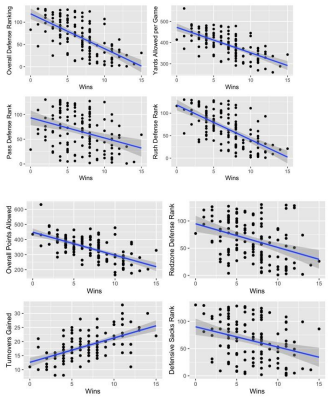
- Offense
    - In speculating how college football offenses attribute to winning games, we looked at several offensive statistics. In total, these statistics include offensive yardage, touchdowns, rank, plays, first downs, scoring rank, and turnover margin. Using these statistics, a range of different visualizations, along with a regression model, can be created to illustrate the impact that the offensive side of the ball has in winning games. In the box plot, we observe the range of yardage based on teams that are grouped together by the number of wins they have, where the size of the boxes represents the range in offensive yardage. Furthermore, with an increasing number of wins, we notice an upwards trend in the box plots, developing the idea that teams that win more are more of a threat offensively. Next, we wanted to see the correlation between offensive touchdowns and winning games. Once again, the teams with the same number of wins are categorized together to make the scatter plot easier to read. Moreover, a line of best-fit is drawn across the plot, displaying a strong linear relationship, telling us that winning teams score more touchdowns. Moreover, a similar linear relationship is shown when plotting the offensive rank against the number of wins. However, when observing the offensive rank of teams, you notice that there are some teams that don't win as much, but have a

decently-ranked offense. Lastly, we wanted to create regression models for each side of the ball. For offense, we create a formula to study the relationship between wins and offensive yards, touchdowns, rank, plays, first downs, scoring rank, and turnover margin. Next, we find the most parsimonious model using the step function to simplify the model and extract the most important variables (but all of them actually ended up being important). To illustrate our offensive regression model, we plot the residuals and see that there's an equal number of points above and below our horizontal line, telling us that our model represents a strong relationship between the number of wins that a team has and the variables that we utilized to create the model itself. Additionally, this regression model returned an adjusted-R-squared value of .84, which essentially tells us that our choice in variables are highly correlated to winning games. In observing all of these plots, we can take away the idea that the offense of a college football team has a substantial correlation with winning games.

- Defense
    - When looking at defense, there are a multitude of stats one could choose from. There were around 30 different stats in the data set that could be used when trying to find a relationship with a team's success. After creating scatter plots for about 20 of the more significant stats, it was determined that to find a relationship between wins and a defensive statistic, we would use a linear regression to try and find the highest level of correlation. 8 statistics were chosen to take a closer look at: Overall defensive rank, yards allowed per game, pass defense rank, rush defense rank, points allowed, red zone defense rank, turnovers gained and defensive sack rank.
    - To try and determine the best indicator of a winning team, scatter plots were created for each variable vs wins, with a linear regression model plotted as well including the standard error of the model. The R-squared values were also calculated to determine how effective the linear models are.

- Looking at the results, there were a few surprises and few obvious correlations. The obvious ones being defensive rank, yards allowed per game, rush defense rank and points allowed having the highest R values. They showed a moderate correlation with R values in the 60's. The most surprising outcome was pass defense rank having an R value of 0.34. Defense being primarily made up of 2 aspects, pass and rush, it was surprising to see that one correlates a lot more with winning than the other, meaning that a rush defense is a lot more important to winning than a pass defense. As for the remaining statistics, turnovers gained proved to be a factor with an R value of 0.57. The more chances you have to score, the more likely you are to succeed at scoring more points and ultimately winning. Redzone defense rank and sacks rank do not have much correlation with winning which makes sense because if the opposing teams makes it to the redzone, they're most likely to score whether it's a touchdown or field goal and that is not good for winning. When it comes to sacks, they're a big part of the game but they're a play that is easier for a team to overcome because it does not involve points.

Stat	R	R.Squared
<chr>	<dbl>	<dbl>
Defensive Rank	0.6495935	0.4219717
Yards per Game Allowed	0.6549270	0.4289294
Pass Defense Rank	0.3386502	0.1146840
Rush Defense Rank	0.6418760	0.4120047
Points Allowed	0.6276113	0.3938959
Redzone Defense Rank	0.3624258	0.1313525
Turnovers Gained	0.5698138	0.3246878
Defensive Sacks	0.3088951	0.0954162



- When you look at the phrase, defense wins championships, I would say that statement is true based off of this data. You need a good defense to win a lot of games. The data supports that but something important to note is that you don't need the best defense to win. The national champion in 2019, LSU, is the best example of having an above average defense, not great defense and still winning. The other 3 teams in the playoffs that year had great defensive rankings, leading them all to a lot of wins.

- Special Teams

    - Special Teams is the term used for the team formation used for kickoff plays, it consists of offense and defense players just for that single play, after that the team gets changed to its respective side, either defense or offense.
    - To answer the question “if special teams plays are a big factor to win games” we will go ahead and read our data for special teams. We will analyze and plot the Kickoff Return Rank data and the Kickoff Return Touchdowns.
    - The Kickoff Return Rank represents the overall performance of a team during this plays, it compresses Kickoff Return total Yards in the season, the Kickoff Return avg yards per game, and the Kickoff Return Touchdowns in the season into a single number, when reading and plotting the data we found out that overall performance of special teams during Kickoff Return plays is really beneficial on winning games, it show that good performance during this plays will increase your probability on winning.
    - I wanted to go in depth and see if achieving a miraculous touchdown during this plays will increase your chances of winning so I went ahead and analyzed and plotted the data for Kickoff Return Touchdowns and from what we can see from the boxplot graph is that achieving a touchdown won’t really have that much of an impact on winning probability we can see that there are more teams without achieving a Kickoff Return Touchdown with the same amount of wins than those who did achieve 1 or 2 kickoff return touchdowns on the season. Also, the team with the most wins didn’t achieve a kickoff return touchdown on the season.
- Other/outside aspects
    - Conference perspective
    - An athletic conference is a collection of sports teams, playing competitively against each other in a sports league. There are 11 conferences in college football which are AAC, ACC, Big 12, Big Ten, C-USA, FBS Independent, MAC, Mountain West, PAC-12, SEC, Sun belt.

In this project, I want to analyze the relationship between wins and total points scored by each team from a conference perspective, and see what's the difference between conferences or do they all have the same result.

- After adding the variable, I am going to create the plot that will show the relationship between wins and total points scored by each team from a conference perspective. From the plot I made, we can see that every conference has a clear positive linear relationship between wins and total points scored except conference Mountain West, which I can draw a conclusion that Every team has a positive linear relationship between wins and total points scored except conference Mountain West.
- Moreover, through the graph, I notice that there is a team in the Mountain West that only scored around 300 points but was able to win 10 games. I find out the team which is San Diego State University and compare them to other teams that also won 10 games. The result shows me that all other teams who also won 10 games scored a lot more points than San Diego St, also San Diego St has a defense rank of 5 which proves that Defense can win the game and you do not need to score many points.

Conclusion:

Through the analysis of this college football dataset, we were able to conclude that a team's offense and defense are the most important aspects in terms of winning football games. More importantly, however, the offense showed a much greater correlation in winning college football games more than the defense did, judging by the R-squared values. Additionally, we were able to present interesting and thoughtful observations that may have not been obvious to the average person. Creating graphs and models displaying the correlation between a multitude of different variables, and measuring the importance of each variable in determining how college football teams win games.

Our methods worked well with the goal that we were attempting to accomplish. As previously stated, it dealt with tidying the data and translating it through visualization

and modeling. After that, we produced thoughtful qualitative analysis of the data. The overall execution of analysis is appropriate due to the very reasonable conclusions that we are able to make from the results. Furthermore, the data that was collected initially, by Jeff Gallini, is reliable because he credits the use of the stats website from the National Collegiate Athletic Association (NCAA) itself.

Some improvements that we thought can be made involve the variety of plots and, perhaps, more questions to answer. We found that there is an incredible amount of diversity in the types of plots and the amount of detail you can add is expansive. More detail could've been crucial in ensuring that the plots are exciting, creative, and more professional. Moreover, more questions would've added more fun to the analysis. We thought this because answering more questions would've essentially led to more insight on our overarching question and more intriguing analytics.

However, there are of course some limitations that were encountered over the course of this analysis. Although we believe that our analysis was thoroughly completed, we still grappled with the amount of time that we were given to hit marks that we wanted to. This limited time would ultimately hinder us in making more observations and examining the data into more detail. Additionally, Gallini provides the college football stats for the years 2013-2020, but they're all separated. Even though our conclusions likely would still stand and be similar, the addition of several more years of data would've made for more accurate conclusions of the statistics if given more data. The addition of a "year" column may be an easy fix, but we figured it to be time consuming and not entirely important to get our questions answered. Also, an underlying limitation could also be the lack of skill or experience to gracefully answer some of our ambitious questions.

Some future questions that may be answered using these datasets and methods could be:

- How has Michigan State's football team fared over the 2013-2020 seasons?
- Can we quantify the competitiveness of the Big Ten Conference? (Conference that MSU, U of M, and OSU plays in)
- Over the 2013-2020 seasons, were any of the teams in the Big Ten potentially a good match-up for a top team(s) in the SEC?

References:

https://www.kaggle.com/jeffgallini/college-football-team-stats-2019

https://www.ncaa.com/stats/football/fbs

STT 180 Slides