Appendix: Data Set Descriptions

Ryan Morgan

November 9, 2017

Data Sets Introduction

In order to make college football data available to be easily analyzed and explored, data tables were scraped from the sports-reference website. The tables were then cleaned and compiled into manageable and formatted data sets and saved as CSV files. Once the CSV files were formatted, they were used to explore various ways to predict win probabilities of AP ranked teams.

The sports-reference website has detailed data available on every Division 1 College Football team. The page https://www.sports-reference.com/cfb/schools/ has a list of all the college football teams that sports-reference has information on. There is a total of 296 schools listed, along with the years that each school's football team was active. Users can also explore more detailed information on specific teams (For example, a page on the 117 years of Nebraska football can be viewed). These pages have the Year, Conference, Wins, Losses, Ties, Winning Percentage, SRS (Simple Rating System, which is a way of rating how good the team was that season), SOS (Strength of Schedule, a way of measuring how good a team's opponents were that season), the preseason Associated Press (AP) ranking, the highest AP ranking for that season, the postseason AP ranking, the head coach(es) from that season, the bowl game played in that season, and any notes on the that season. Users can also get a more detailed look at a specific season for a specific team (for example, a page with information on the 2016 Nebraska Football season can be viewed). These pages have a table that lists game averages for the offense, game averages for the defense, and the difference between the offensive and defensive averages (differences found by taking offensive average minus defensive average). For example, in the 2016 Nebraska football team averaged 15.5 completions per game on offense, while the defense surrendered an average of 19.5 completions per game to opponents, for a difference of -4.0 completions per game. These tables have game averages on 21 statistics, including Pass Completions, Rushing Attempts, and Total Turnovers. Users can also view the schedule for a specific season for a specific team (for example, a page with the schedule of the 2016 Nebraska football team can be viewed). The schedule pages have information on each game that a team played that season, including the date and location of the game, the opponent of the game, and the result (either Win, Loss, or Tie) of the game. Instead of just looking at the results of a game, users can also view pages with specific offensive and defensive statistics for each game in a season (for example, a page with specific game-by-game stats on each game Nebraska played in the

2016 season can be viewed). These pages have information on the offensive and defensive statistics of each game, including rushing and passing attempts, rushing and passing yards, and rushing and passing touchdowns.

CSV files were created by scraping data from the described sports-reference webpages. The rvest package was used to scrape the data from the described tables on the sports-reference website. The functions used to scrape tend to use the same basic process, in that the data is scraped, formatted into a data frame with the proper dimensions, and then the columns are named and formatted. Data was scraped for each available season for each available team. The data was then organized into five CSV files; Total_Team_History, Individual_Season_Results, Season_Averages, Game_Results, and Game_Logs.

Data Set Descriptions

Total_Team_History

File Description

The Total_Team_History file is a 296 by 19 file. Each row corresponds to a team that has played at least one season of division 1 college football, a total of 296 teams. The columns names, along with what they represent are:

- **Team:** The name of the Team
- **First_Season:** The first year that team played a Division 1 (D1) College football season.
- **Last_Season:** The last year that team played a Division 1 College football season. Teams that are still active in D1 football will have the last year be 2016.
- **Number_Of_Seasons:** The number of seasons the team has played D1 football. Note that it is not simply the difference between the Last_Year and the First_Year column, as some teams had a hiatus from playing D1 (For example, Appalachian State played D1 from 1972 through 1981 and 2014 through 2016, for a total of 14 years).
- **Games_Played:** The total number of D1 games played by the team in its history.
- **Wins:** The total number of D1 wins by the team in its history.
- **Losses:** The total number of D1 losses by the team in its history.
- **Winning_Pct:** The all time D1 winning percentage by the team.

- **Bowl_Games_Played:** The total number of bowl games in which the team played in throughout the team's history.
- **Bowl_Wins:** The total number of bowl wins by the team in its history.
- **Bowl_Losses:** The total number of bowl losses by the team in its history.
- **Bowl_Ties:** The total number of bowl ties by the team in its history.
- **Bowl_Winning_Pct:** The all time bowl winning percentage by the team.
- **Simple Rating System:** From the sports-reference glossary: "a rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average."
- **Strength of Schedule:** From the sports-reference glossary: " a rating of strength of schedule. The rating is denominated in points above/below average, where zero is average."
- **Number_Of_Seasons_Ranked_In_AP_Final_Poll:** Number of seasons ranked in the final Associated Press (AP) Poll.
- **Conference_Championships:** The number of D1 conference championships won in the Team's history.
- **Notes:** Any notes on the team's history. Every example of a note is a record adjustment by the NCAA due to sanctions or punishments.

Example Graphs

The Total_Team_History CSV file can be used to compare total history between teams. For example, Figure 1 compares the difference in total wins between a select number of teams. Figure 2 displays the relationship between a team's all time win total and the number of seasons the team has finished ranked in the AP Poll.

Individual_Season_Results

File Description

The Individual_Season_Results file is more detailed than the Total_Team_History file, in that instead of just looking at the total history for each team, this file looks at data on each season for each team. This file has data on each D1 football season played. The Individual_Season_Resulsts file is a 13226 by 16 file. Each row corresponds to a single Team's season. The column names, along with what they represent are:

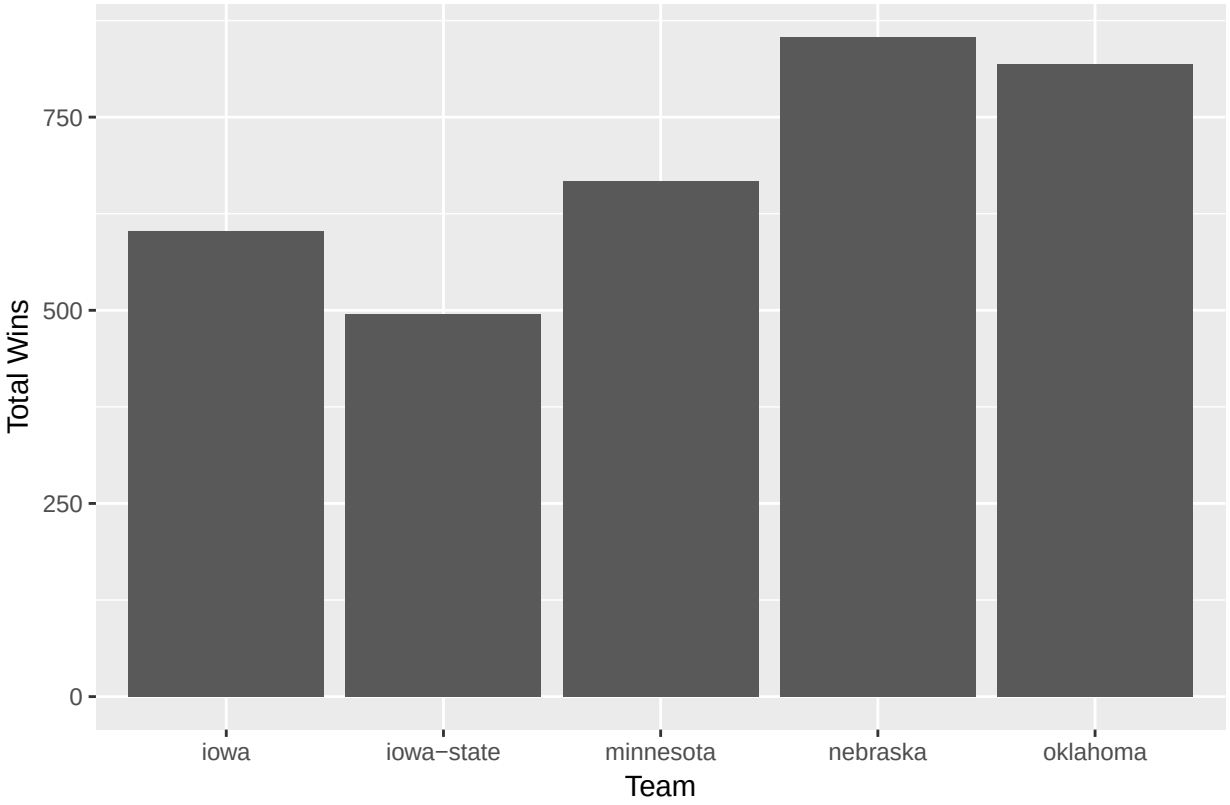

Figure 1: Comparing all time wins between the Nebraska, Iowa, Iowa State, Oklahoma, and Minnesota football programs

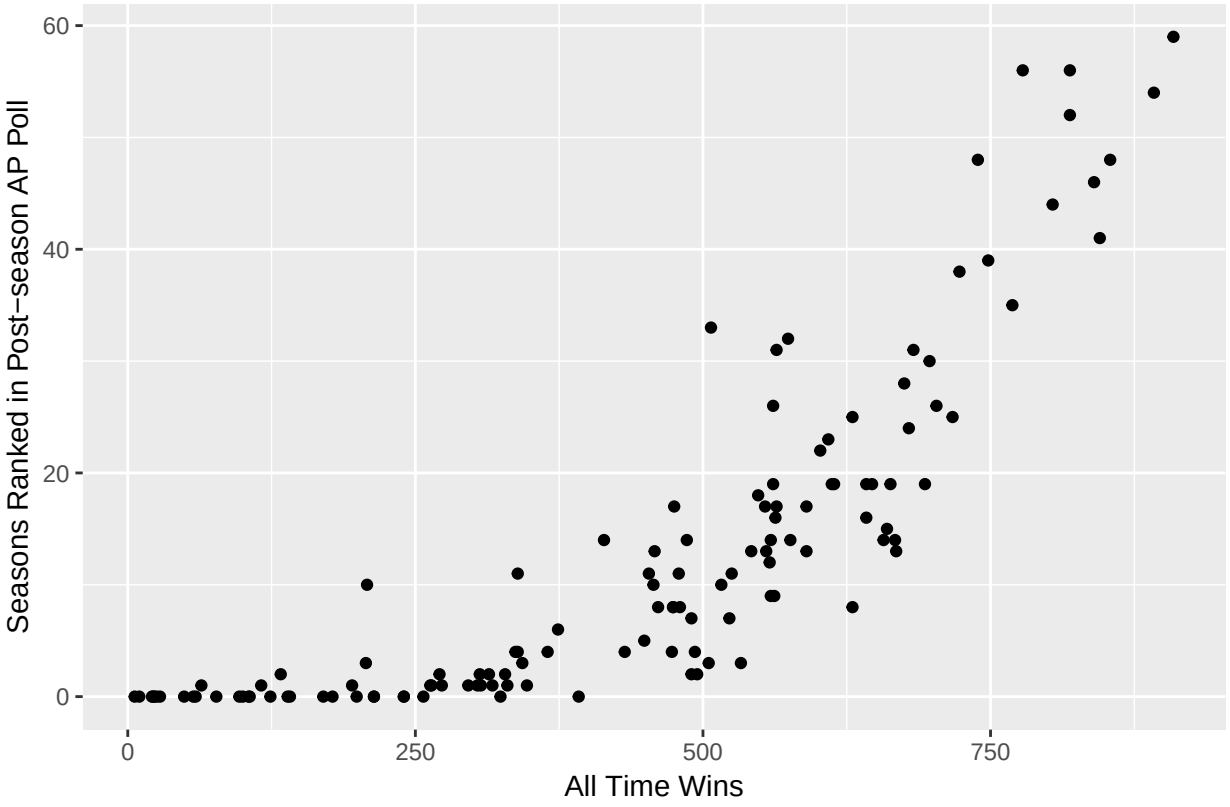

Figure 2: Comparing All Time Wins to Number of Seasons Ranked in the Post Season AP Poll for teams active in 2016. Programs with more all time wins tend to have more seasons in which they finished ranked.

- **Team:** The team name.
- **Season:** The season the rest of the row refers to.
- **Conference:** The Conference that the team was in during the given season.
- **Wins:** The number of wins by the given team during the given season.
- **Losses:** The number of losses by the given team during the given season.
- **Ties:** The number of ties by the given team during the given season.
- **Winning_Pct:** The winning percentage by the given team during the given season.
- **Simple Rating System:** From the sports-reference glossary, "a rating that takes into account average point differential and strength of schedule. The rating is denominated in points above/below average, where zero is average."
- **Strength of Schedule:** From the sports-reference glossary, " a rating of strength of schedule. The rating is denominated in points above/below average, where zero is average."
- **AP_Poll_Preseason_Rank:** The preseason Associated Press (AP) ranking of the given team for the given season. The AP ranks the 25 best college football teams before the season begins, each week of the season, and once the season has concluded. A ranking of 1 means the team is the highest ranked (or best) team in the poll. The AP only ranks what is viewed to be the 25 best teams.
- **AP_Poll_Postseason_Rank:** The postseason AP ranking of the given team for the given season.
- **AP_Poll_Highest_Rank:** The highest ranking the given team achieved for the given season.
- **Coach(es):** The name of the head coach (or coaches) for the given team during the given season. The record is also listed after the coaches name for that coach's record for that season (which is relevant, since teams sometimes fire a coach mid-season).
- **Bowl_Game:** The Bowl game played in by the given team in the given season. If the column is blank, the team did not appear in a bowl game for that season.
- **Bowl_Result:** The result of the bowl game for the given season. If the column is blank, the team did not appear in a bowl game for that season.
- **Notes:** Notes on the team's season.Typically a record adjustment by the NCAA.

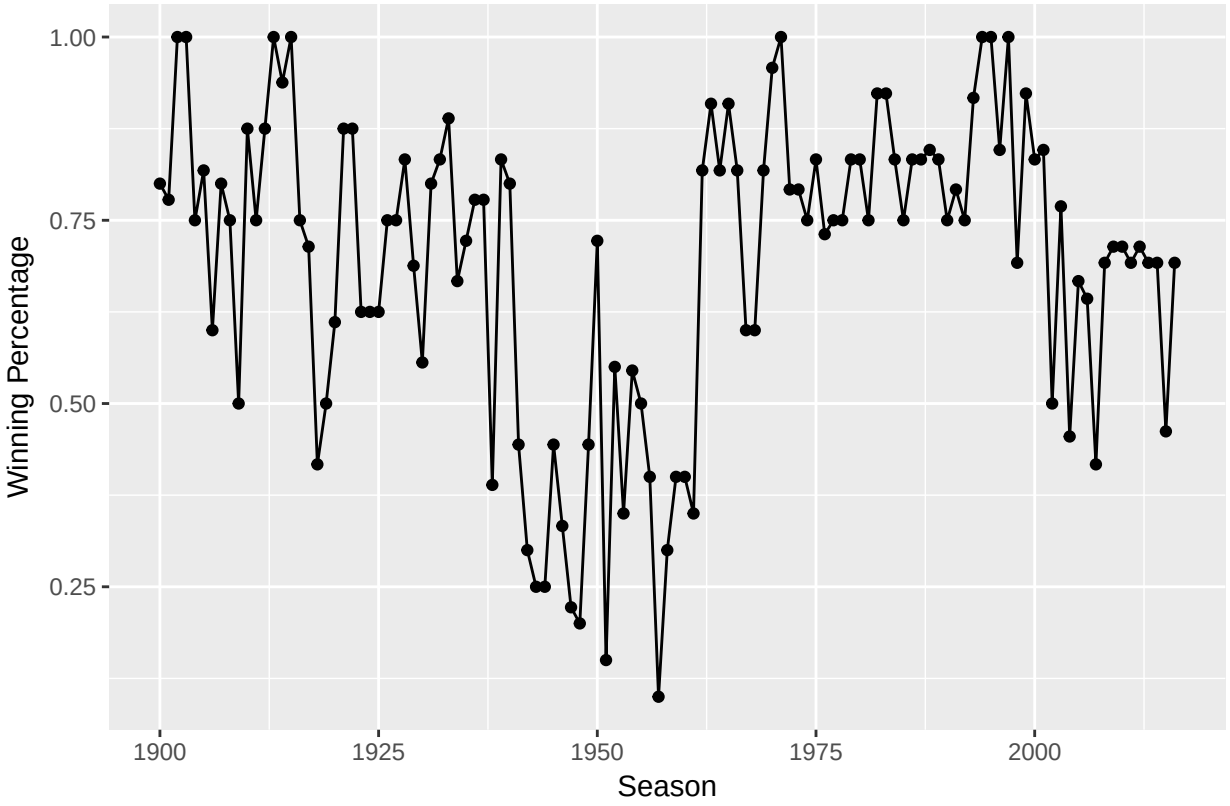

Figure 3: Nebraska Winning Percentage by Season

Example Graphs

The Individual_Season_Results CSV file has a lot more information than the Total_Team_History file, in that the Individual_Season_Results file allows users to look at each season. For example, Figure 3 displays how Nebraska's winning percentage has varied from season to season. This data set also allows for comparing variables across multiple schools. For example, Figure 4 displays how 3 SEC teams' simple rating system (SRS) is related to the team's winning percentage. This data set also makes it simple to compare conferences. Figure 5 display the number of bowl wins the SEC, Big 12, Big Ten, and Pac10/12 conferences have had per season since 2000.

Season_Averages

File Description

The sports-reference website has season averages for each team. The way the sports-reference page is laid out, there are 3 rows. The first row is the Offense's season averages, the second row is the Defense's season averages, and the third row is the Difference between the Offensive averages and Defensive Averages (found by taking Offense - Defense). The Tables have season averages on passing, rushing, total offense, first downs, penalties, and turnovers.

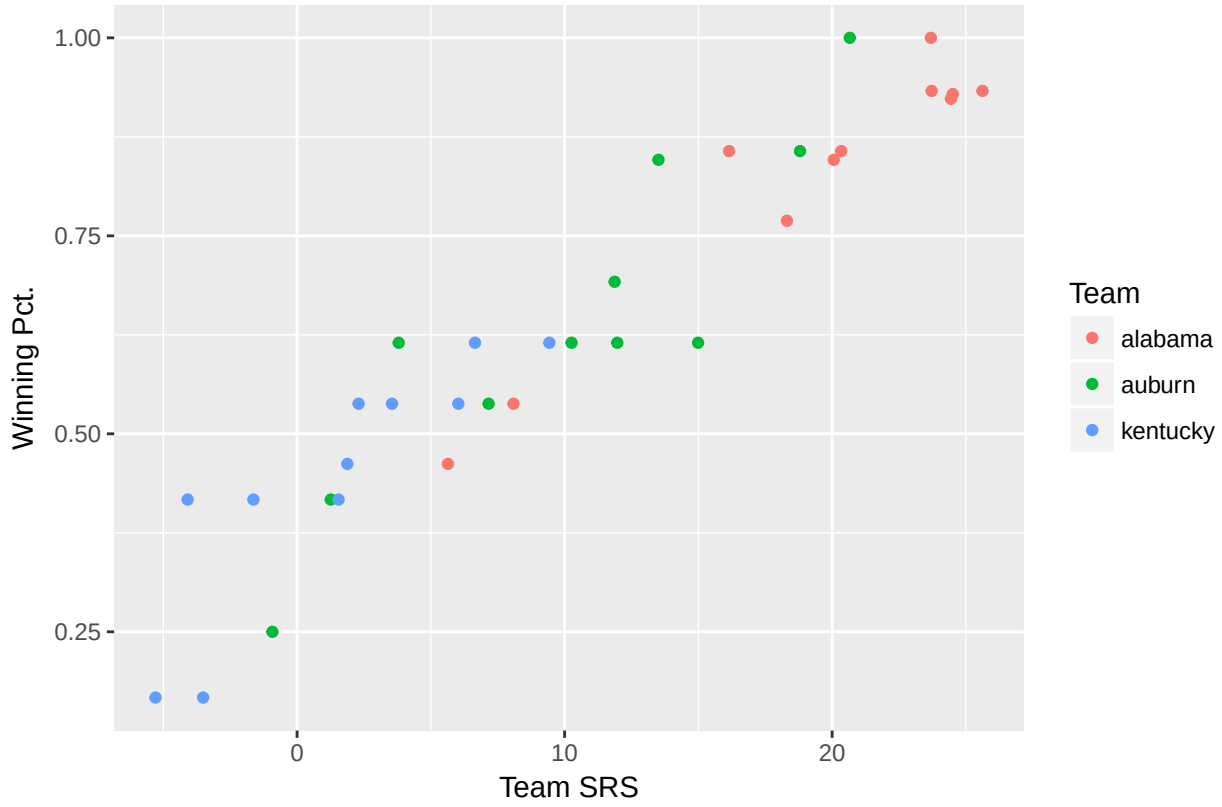

Figure 4: Relationship between SRS (Simple Rating System) and Winning Percentage for SEC team (2005-2016). The teams included show how some teams more often have higher SRS and winning percentages (Alabama), some teams often have lower SRS and winning percentages (Kentucky), and some teams have a lot of variation in SRS and winning percentage (Auburn)

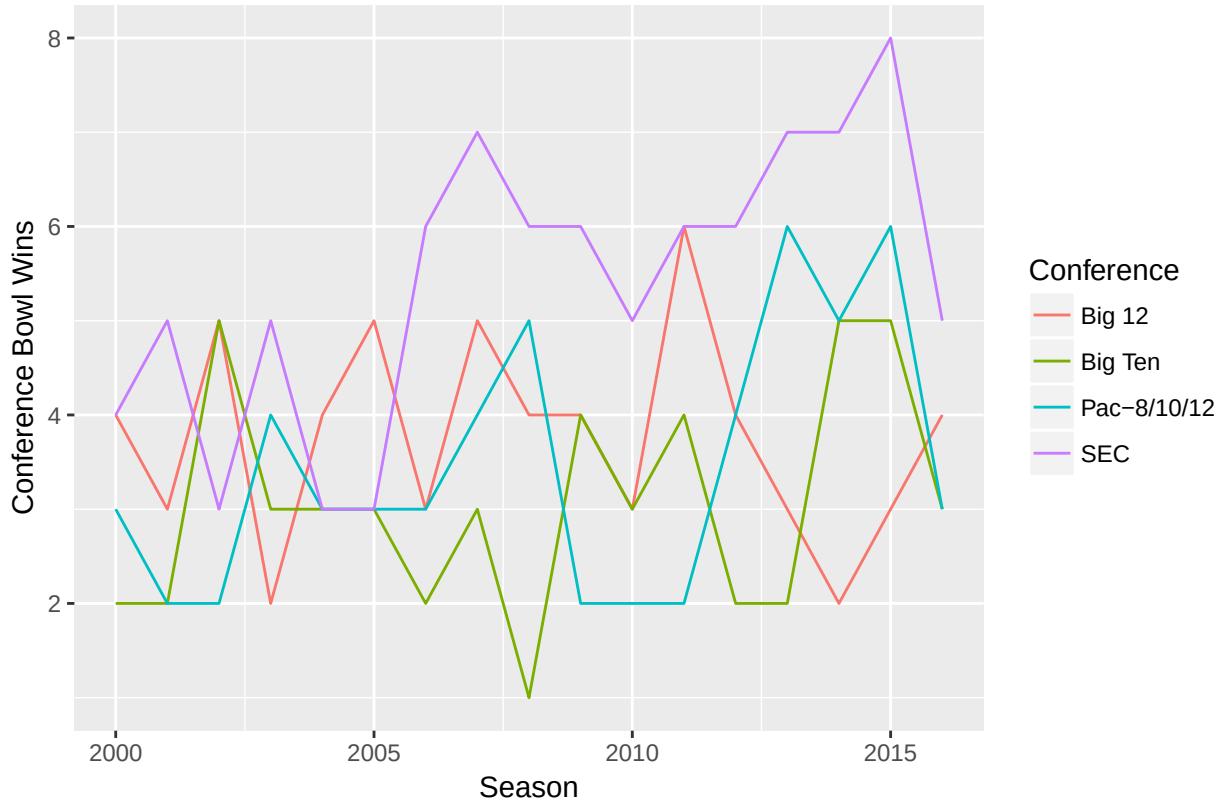

Figure 5: How a conference's number of bowl wins varies from season to season. This graph seems to confirm the reputation the SEC has for being the most sucessful conference in bowl games (despite the sharp drop off in 2016).

The tables of season statistics available on sports-reference are different depending on if the season is prior to or after the 2000 season. There are more season statistics available since the 2000 season, so different functions were used to scrape data from seasons prior to 2000 and seasons after 2000. Season averages are available for every team that has played a season of Division 1 football since 1956. The Season_Averages csv file has 21420 rows and 26 columns. Each row represents a Team's average Offensive, Defensive, or the difference between Offensive and Defensive statisitcs for a given season. This means that each team has three rows for each season (For example, the 2016 air-force football team has an offensive row, a defensive row, and a "difference" row). The column names and what they represent are:

- **Team:** The team name
- **Season:** The season, ranging from 1956 to 2016
- **Conference:** The Team's conference for the given Season.
- **Type:** The type of statistics that will appear in the rest of the row; either "Offense", "Defense", or "Difference"
- **Games:** The number of games played in the season.
- **Pass_Comp:** The average number of pass completions in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Pass_Att:** The average number of pass attempts in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Pass_Pct:** The average completion percentage in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Pass_Yds:** The average number of passing yards in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Pass_TD:** The average number of passing touchdowns in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Rush_Att:** The average number of rushing attempts in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Rush_Yds:** The average number of rushing yards in a game for the given team's offense, defense, or difference between the offense and defense for the given season.

- **Rush_Avg:** The average number of yards gained per rushing attempt in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Rush_TD:** The average number of rushing touchdowns in a game for the given team's offense, defense, or difference between the offense and defense for the given season.

Note: The remaining 12 columns are only available for seasons after 1999.

- **Total_Plays:** The average number of plays ran in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Total_Yds:** The average number of yards gained/given up in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Yards_Per_Play:** The average yards per play in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Pass_First_Down:** The average number of first downs via passing play in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Rush_First_Down:** The average number of first downs via rushing play in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Penalty_First_Down:** The average number of first downs via penalty in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Total_First_Down:** The average number of first downs in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Penalties:** The average number of penalties committed against (listed in the offense rows) or committed (listed in the defense rows) or the difference between penalties committed against and penalties committed (listed in the difference row) for a given team in a given season.
- **Penalty_Yards:** The average number yards via penalty either gained (listed in the offense rows), surrendered (listed in the defense rows), or difference between gained and surrendered (listed in the difference row) for a given team in a given season.
- **Fumbles:** The average number of fumbles in a game for the given team's offense, defense, or difference between the offense and defense for the given season. Note: This column only counts fumbles that resulted in a turnover. If the offense fumbles the ball and recovers it, it is not included in this column.

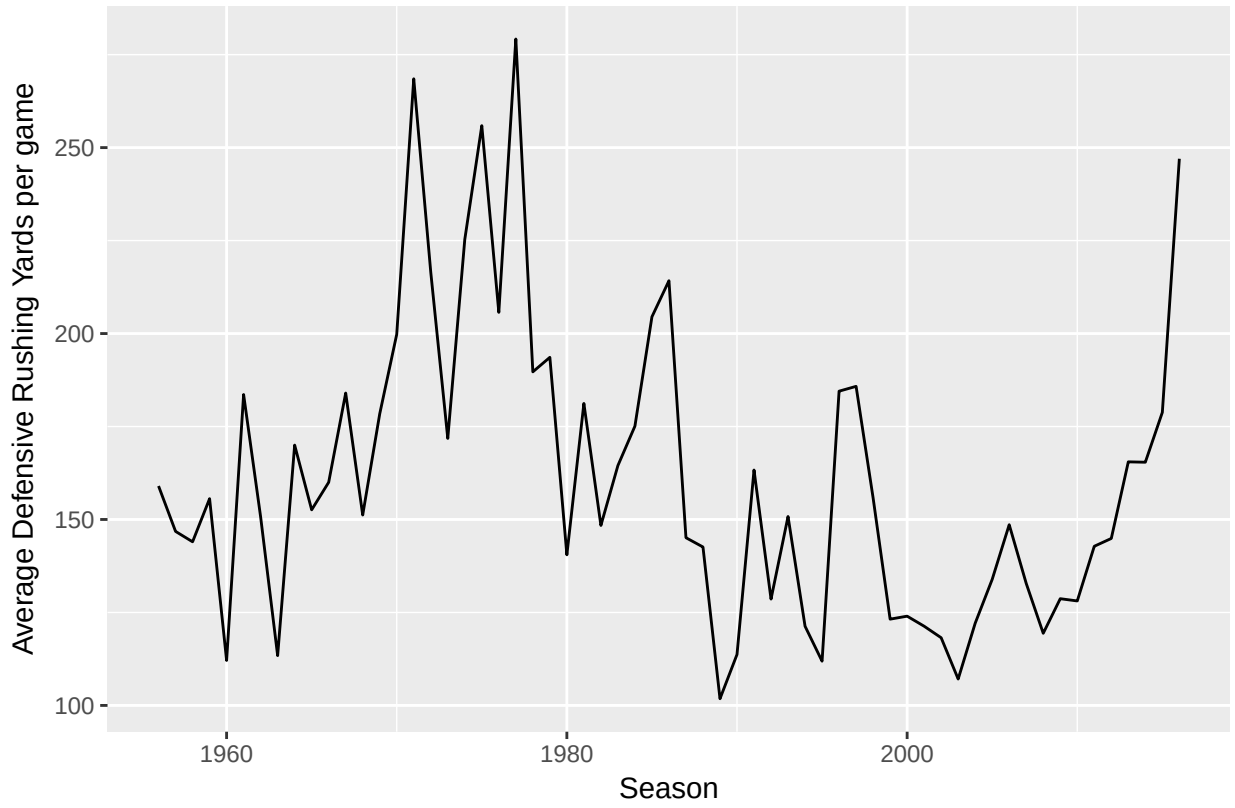

Figure 6: Oregon Defensive Rushing Yards averages. Something that is particularly interesting is the recent rise in rushing yards allowed by Oregon's defenses.

- **Interceptions:** The average number of interceptions in a game for the given team's offense, defense, or difference between the offense and defense for the given season.
- **Turnovers:** The average number of turnovers in a game for the given team's offense, defense, or difference between the offense and defense for the given season.

Example Graphs

The Season_Averages file gives a more detailed look at a Team's season, since it provides a look at offensive and defensive averages instead of just the winning percentage or total number of wins. For example, Figure 6 shows how Oregon's Defensive Rushing Yards per game has varied since the 1960's. One could also explore how two variables may be related to each other within a particular conference. For instance, Figure 7 shows the relationship between Pac10/12 team's offensive passing touchdowns (per game) and offensive completion percentages since 2000.This data set makes it easy to look at how teams and conferences differ from each other and change from year to year in more detailed ways. For example, Figure 8 displays side by side boxplots for the average passing yards per game for teams in the ACC and the Big 12.

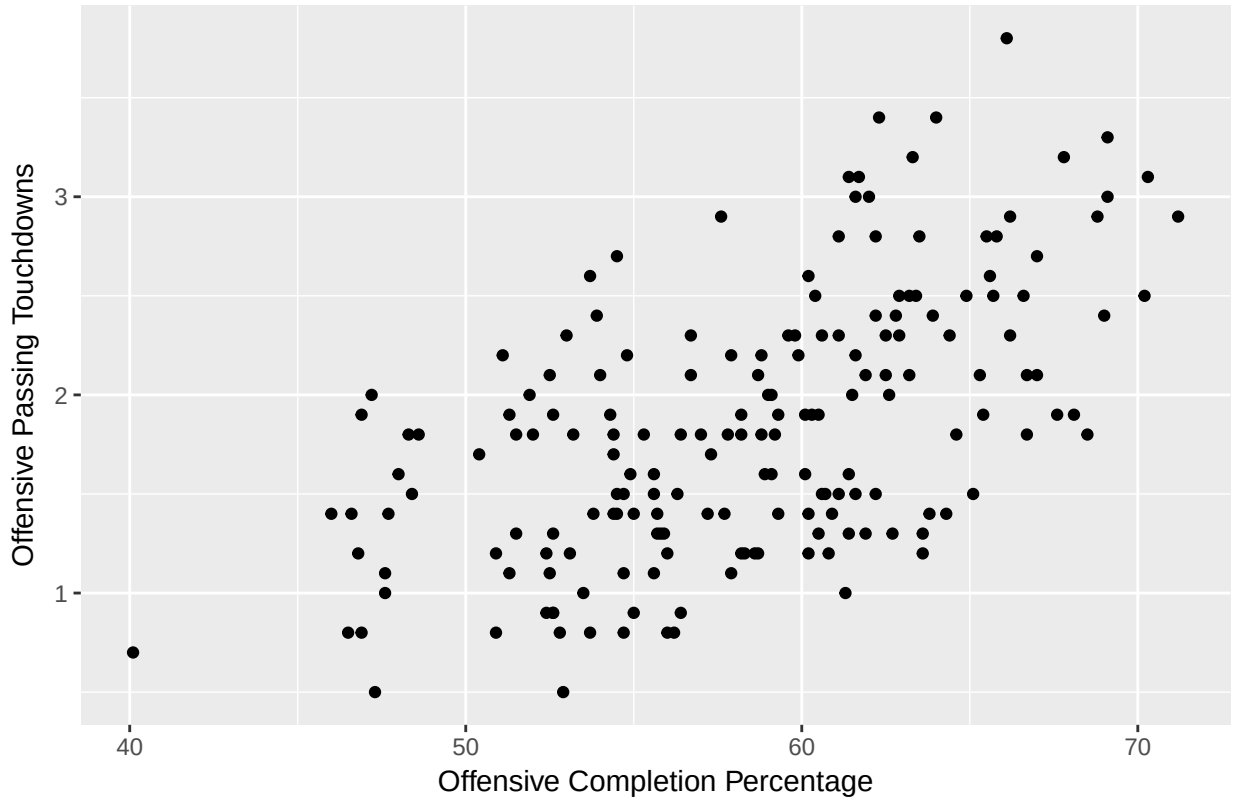

Figure 7: Relaiontip between Pac 10/12 Teams'(2000-2016) completion percentages and passing touchdowns per game. In a trend that should be expected, teams that have higher completion percentages tend to throw more touchdowns.

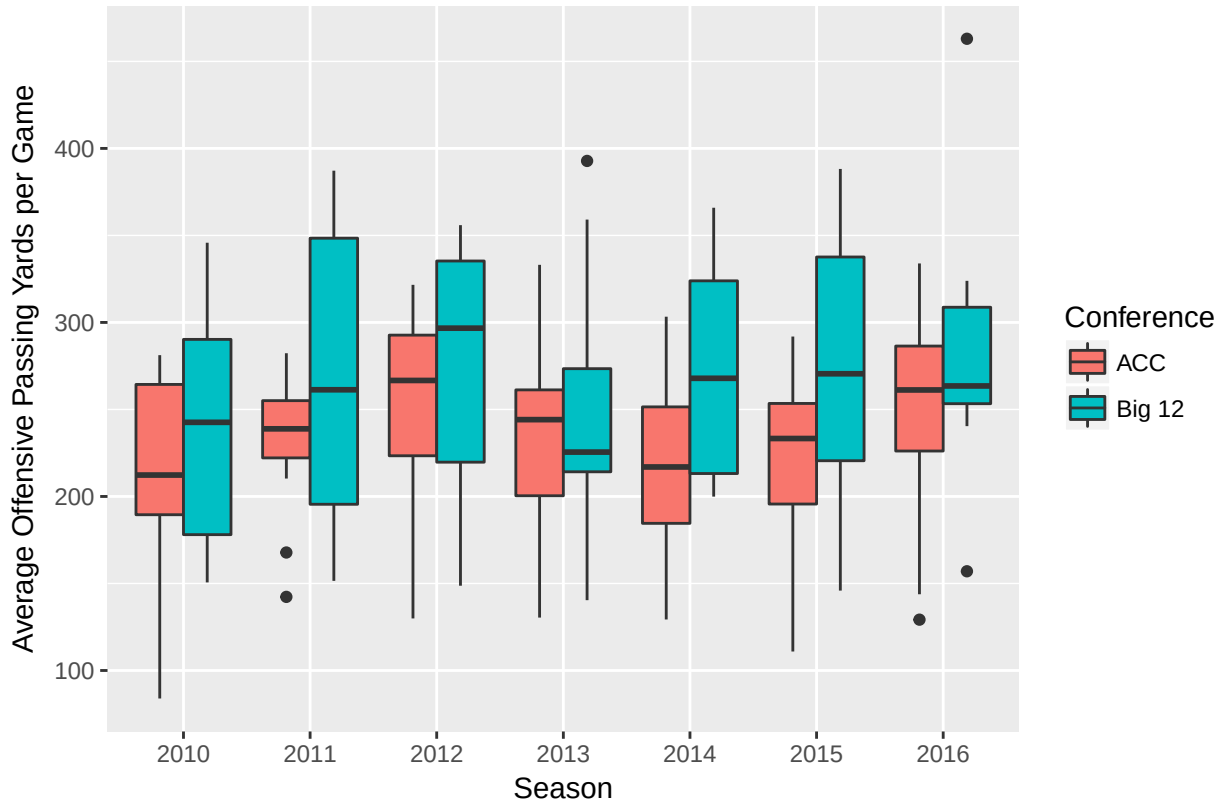

Figure 8: This graph shows how Average Offensive Passing Yards per Game differs between the ACC and Big 12, as well as how the averages have changed from season to season.

Game_Results

File Description

The sports-reference website has data on the schedules of every season in D1 college football history. This gives some information on every D1 cfb game played. The schedule tables on the sports-reference website have information on the date of the game, the location (either Home or Away), the points scored by both the team and the opponent, the ranking of the team and the opponent, as well as the result. The Game_Results csv file has 135354 rows and 18 columns. Each row represents a game played by a D1 football team. Games between two D1 football teams are accounted for twice in the csv file, once from each team's perspective. For example, the 2016 game between Iowa and Iowa State is listed once from Iowa State's perspective (where Iowa is the opponent) and once from Iowa's perspective (Where Iowa State is the opponent). The column names and what they represent are:

- **Team:** The name of the team. This is the team referred to as "Team" in the rest of the row.
- **Season:** The Season for which the game took place. Note, the Season doesn't always match the date. For example, the 2016 National Championship took place in 2017, but it was still a part of the 2016 season.
- **Conference:** The conference that the team was a member of at the time of the game.
- **Game_Number:** The game number for the team for the given season. The first game of the season would have a Game_Number of 1, the second a Game_Number of 2, etc.
- **Date:** The date of the given game.
- **Day:** The day of the week the given game took place.
- **Location:** The location of the game, from the "Team" perspective. The Locations are either Home, Away, or Neutral.
- **Team_Rank:** The AP poll ranking of the Team at the time of the game.
- **Opponent:** The opponent who played against "Team" in the given game.
- **Opponent_Rank:** The AP poll ranking of the Opponent at the time of the game.
- **Opponent_Conference:** The conference that the opponent was a member of at the time of the game.
- **Result:** The result of the game, from the perspective of the Team.
- **Points_Scored:** The number of points scored by the team in the given game.

- **Opponent_Points_Scored:** The number of points scored by the opponent in the given game.
- **Current_Wins:** The total number of games won up to that point of the season by the given team for the given season.
- **Current_Losses:** The total number of games lost up to that point of the season by the given team for the given season.
- **Current_Streak:** The current winning or losing streak by the given team for the given season. The streak is denoted by a "W" or an "L" to show if the team is on a current winning streak or losing streak, followed by a number to denote the number of games the current streak is at.
- **Notes:** Any additonal notes about the game. Usually the notes refer to if the game is a conference championship game, a bowl game, or the location of the game if it is at a neutral site.

Example Graphs

The Game Results csv file gives information on every single game played by a D1 football team. This data set makes it simple to observe the distribution of points scored in various games. For example, Figure 9 shows a scatterplot of Nebraska's points scored and Nebraska's points given up in conference games since 2000. Figure 10 shows a similar graph, but compares two teams. It is also easy to look at how many points are generally needed to win a football game. Figure 11 shows what the winning percentage is for a given number of points scored in a game (since 1990). Figure 12 shows how the winning percentage by points scored can vary between conferences. Another thing that is easy to do using this data set is look at what conferences records are against other conferences. For exmaple, Figure 13 shows the SEC's record in out of conference games. Figure 14 compares the record in out of confernce games between the ACC, Big 12, Big Ten, Pac 8/10/12, and SEC.

For example, for the past 10 years, commentators have claimed that the SEC is the best conference. One can easily look at the SEC's record versus other conferences in that time.

Looking at the graph, one can see that the SEC has had a winning record (well above .5) in out of conference games over the past 10 years, with a fairly sharp decline this past season (2016).

As a comparison, one could display what the Big Ten, ACC, Pac 10/12 and Big-12's out of conference record has been during that time, compared to the SEC.

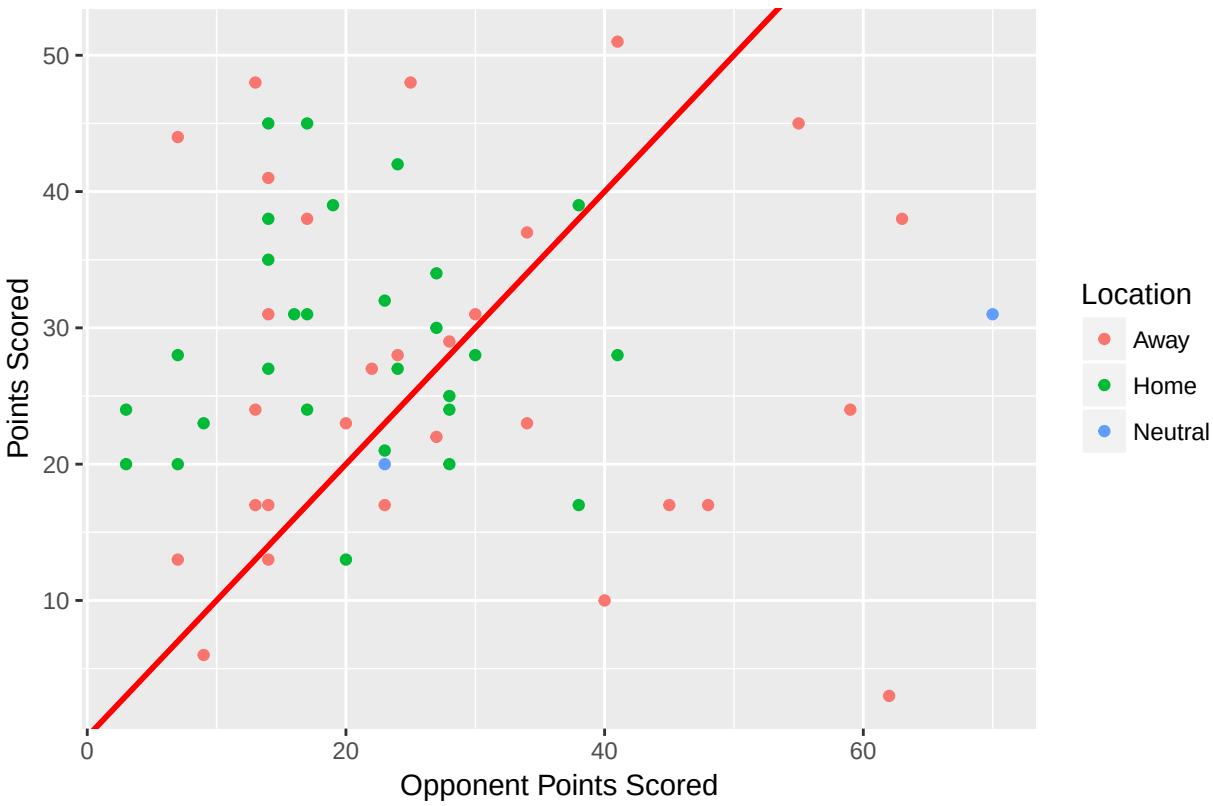

Figure 9: Scatter plot of Nebraska's Points scored and Nebraska's points given up in conference games since 2000. Points above the line signfiy a win (where the points scored are greater than points given up) while points below the line signify a loss. Furthermore, the points are colored by location (either Home, Away, or Neutral).

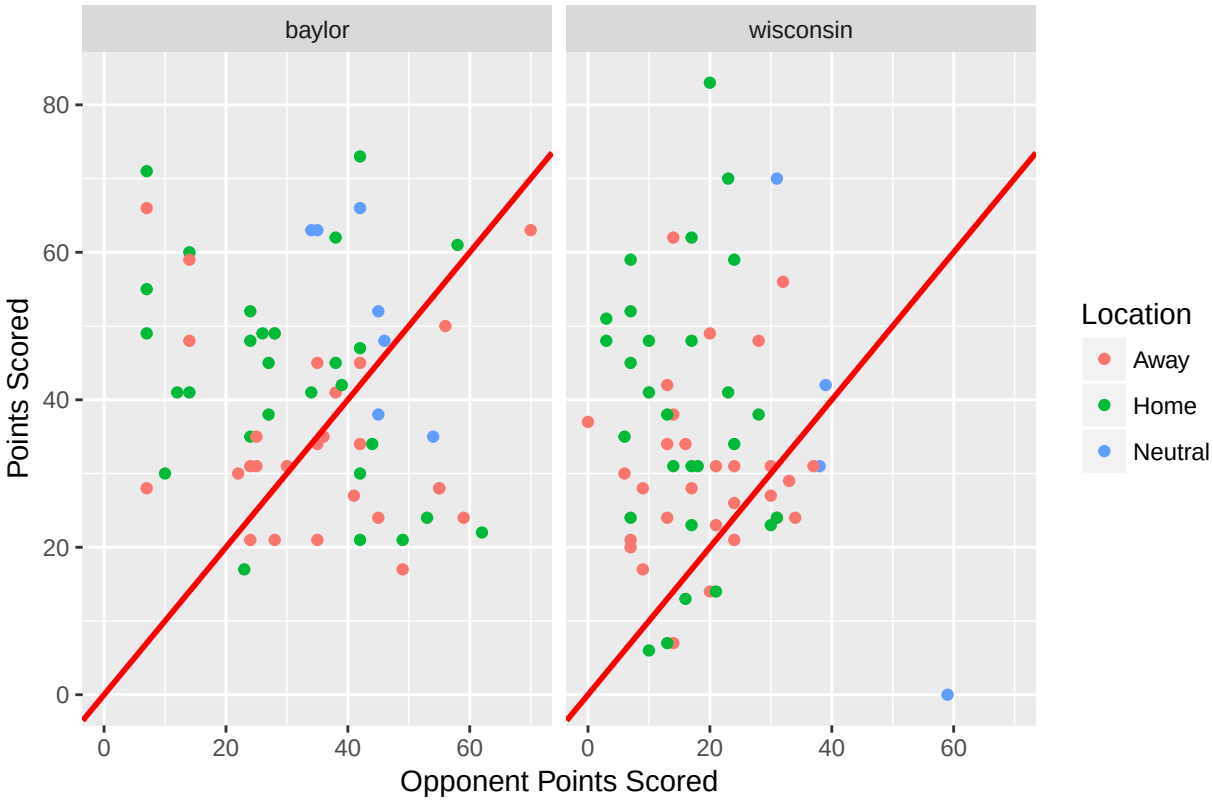

Figure 10: Scatterplots of team's poitns scored and team's points given up in conference games since 2000. This display makes it easy to get a quick snapshot of common scores for different teams. For example, Wisconsin's defense rarly gives up more than 40 points, and most losses generally have close scores. This is different than Baylor, who gives up more than 40 points quite frequently and whose losses aren't always in close scoring games.

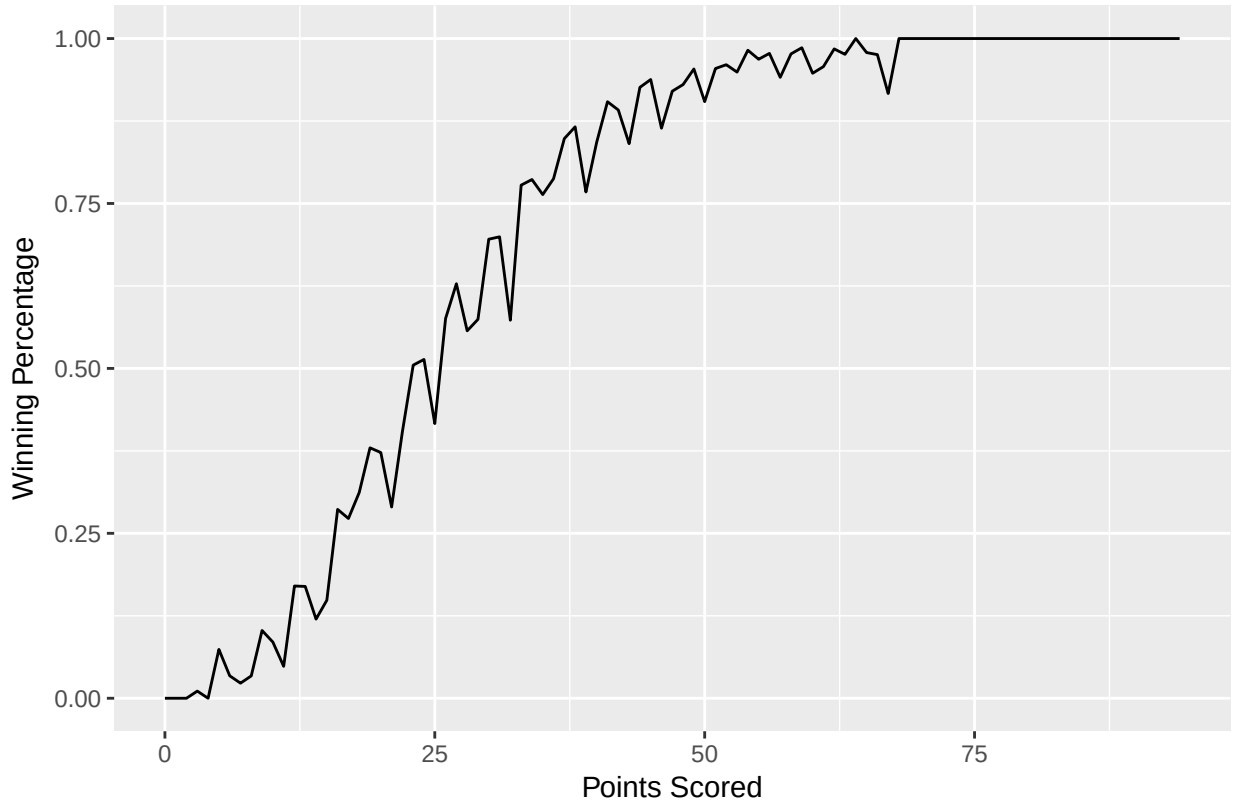

Figure 11: Winning Percentage by points scored in a game since 1990. The general trend seems to be that teams who score less than 25 points are more likely to lose, while teams who score more than 25 points are more likely to win.

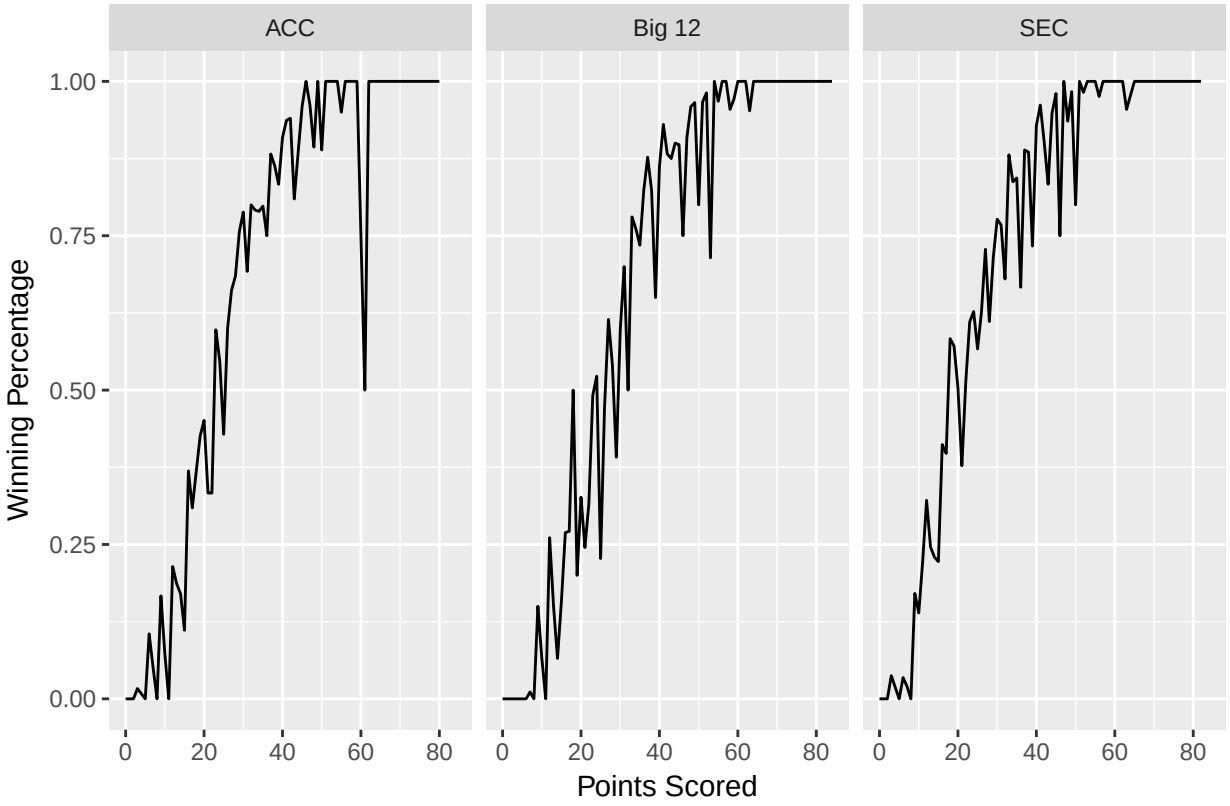

Figure 12: Winning percentage by points scored for ACC, Big 12, and SEC teams since 1990.

Game_Logs

File Description

The most detailed college football data that the sports-reference website has is individual game logs. Sports-reference has detailed information on most games since 2000. This information includes passing, rushing, total offense, first downs, penalties, and turnovers statistics for both the offense and defense for every game. Game logs that weren't available on sports-reference were found on foxsports's website. Scraping the game logs tables were more complicated than any of the previous files scraped. To skip over the details, an overview of the scraping will be provided. A list of teams on sports-reference was used to get a list of teams who have played a season since 2000. A function then found the years active for each of the teams who have played since 2000. From there, offensive game logs for every team for every season since 2000 was scraped. For whatever reason, the defense logs were not able to be scraped. Since the offense logs for a team in a game are the defense logs for the opponent in that game, a data frame of all the offense logs was used to compile another data frame of defense logs (for example, if Nebraska rushes for 200 yards against Wisconsin, that means Wisconsin's defense gave up 200 rushing yards).

It was at this point that several problems with sports-reference's data set were found. The first problem was there were 6 games that didn't have an opponent listed. This was easy to fix by substituting the blank in the

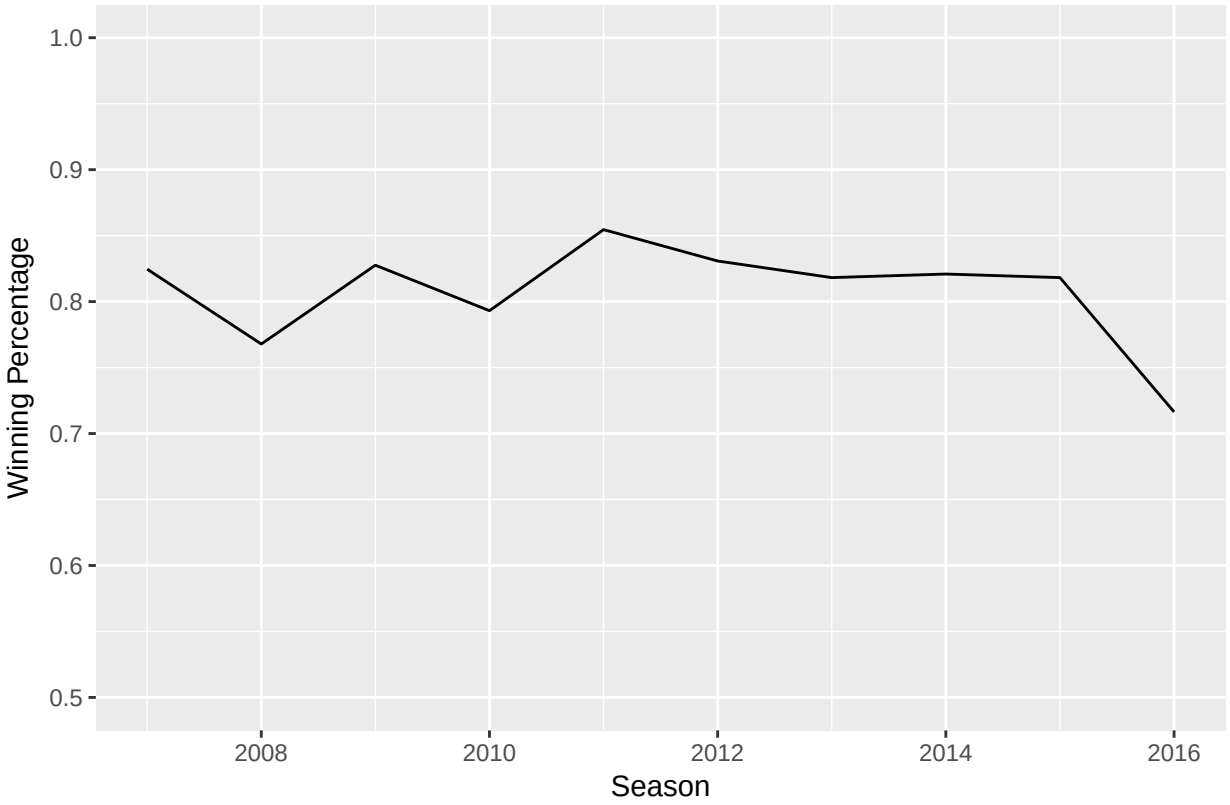

Figure 13: SEC winning percentage in non-conference games. Most people would argue that the SEC is the best conference in college football. This graph shows that the SEC has had a winning record (well above .5) in out of conference games over the past 10 years, with a fairly sharp decline in the 2016 season.

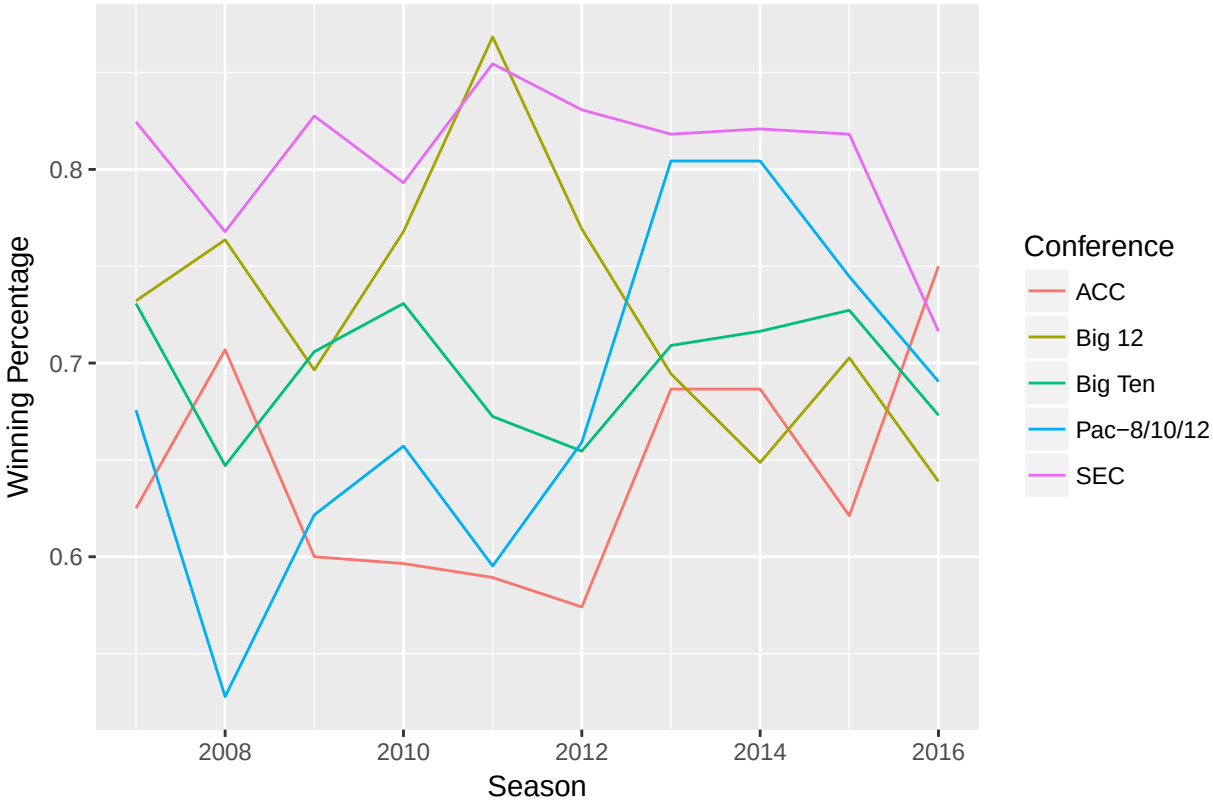

Figure 14: Comparing conference winning percentages in out of conference games. As seen in this graph,the SEC has had the best winning percentage in out of conference games except for two seasons: 2011 (when the Big 12 had the best winning percentage) and 2016 (when the ACC had the best winning percentage).

data set with the actual opponent name. The biggest problem was with the when teams play opponents that aren't division 1. For example, when Iowa State plays the University of Northern Iowa, sports-reference stores only the Iowa State game logs, since Iowa State is a divison 1 school and UNI is not. Since the offense logs were used to attain the defense logs, this means that the Northern Iowa offense logs would be needed in order to get Iowa State's defense logs. The problem is that sports-reference doesn't have a page with UNI's (or any non-division 1 team's) offense logs. To get around this problem, brute force was used and the missing defense logs was copied and pasted into a csv file. After already using brute force, a new website (fox sports) was found that also has game log data. If it were possible to go back, it would have been attempted to scrape from the fox sports website instead of copying and pasting game logs for more than 1300 games. Another smaller problem was that the bowl games for the 2000 and 2001 season didn't have game logs available on the sports-reference website. The game logs available on the fox sports website was used in order to attain these logs.

The (now completed) offensive game logs were used to create defensive game logs. The game logs were compared to the previously scraped allscheudles data frame to make sure each game since 2000 was accounted for. The information available in the game logs data frames was then combined with the information in the allscheudles data frame to give information on each game. The game logs data frame was then cleaned to fix any NA values and to try and catch any values that were potentially entered wrong (for instance, one game log claimed that there was more than 80 turnovers in a game). It was also verified that any relationships that should exist between columns were accounted for (for example, the completion percentage should be equal to 100*completions/attempts). After scraping and cleaning this data set, the file was stored as Game_Logs.

The Game_Logs file has 25602 rows and 60 columns. Each row corresponds to a game log for a single team in a game. Every D1 game since the 2000 season is accounted for. The column names along with what they represent are:

- **Team:** The name of the team referred to in the rest of the row.
- **Season:** The season in which the game took place.
- **Conference:** The given teams's conference for the given season
- **Game_Number:** The number of game for the given team during the given season.
- **Date:** The date in which the given game took place
- **Day:** The day of the week the given game took place
- **Location:** The location of the game (from the Team's perspective). Either "Home", "Away", or "Neutral".

- **Team_Rank:** The AP poll ranking of the given team at the time of the given game. Teams that were not ranked are listed as "Unranked"
- **Opponent:** The opponent played against by the given team in the given game.
- **Opponent_Rank:** The AP poll ranking of the given team at the time of the given game. Teams that were not ranked are listed as "Unranked"
- **Opponent_Conference:** The opponent's conference for the given season.
- **Result:** The result of the game (from the perspective of the "Team").
- **Points_Scored:** The number of points scored by the team in the given game.
- **Opponent_Points_Scored:** The number of points scored by the opponent in the given game.
- **Current_Wins:** The total number of games won up to that point of the season by the given team for the given season.
- **Current_Losses:** The total number of games lost up to that point of the season by the given team for the given season.
- **Current_Streak:** The current winning or losing streak by the given team for the given season. The streak is denoted by a "W" or an "L" to show if the team is on a current winning streak or losing streak, followed by a number to denote the number of games the current streak is at.
- **Notes:** Any additional notes about the game. Usually the notes refer to if the game is a conference championship game, a bowl game, or the location of the game if it is at a neutral site.
- **O_Pass_Comp:** The number of passes completed by the given team's offense in the given game.
- **O_Pass_Att:** The number of passes attempted by the given team's offense in the given game.
- **O_Pass_Pct:** The pass completion percentage by the given team's offense in the given game.
- **O_Pass_Yds:** The number of offensive passing yards by the given team's offense in the given game.
- **O_Pass_TD:** The number of offensive passing touchdowns scored by the given team's offense in the given game.
- **O_Rush_Att:** The number of rushing attempts by the given team's offense in the given game.
- **O_Rush_Yds:** The number of offensive rushing yards gained by the given team's offense in the given game.

- **O__Rush__Avg:** The average yards gained via rushing by the given team's offense in the given game.
- **O__Rush__TD:** The number of offensive rushing touchdowns scored by the given team's offense in the given game.
- **O__Total__Plays:** The total number of plays ran by the given team's offense in the given game.
- **O__Total__Yds:** The total number of yards gained by the given team's offense in the given game.
- **O__Yards__Per__Play:** The average yards gained per play by the given team's offense in the given game.
- **O__Pass__First__Down:** The number of first downs gained via pass plays by the given team's offense in the given game.
- **O__Rush__First__Down:** The number of first downs gained via rush plays by the given team's offense in the given game.
- **O__Penalty__First__Down:** The number of first downs gained via penalty by the given team's offense in the given game.
- **O__Total__First__Down:** The total number of first downs gained by the given team's offense in the given game.
- **Penalties__Against:** The number of penalties committed by the given team in the given game. Note: This includes penalties committed by the team's offense and defense.
- **Penalty__Yards__Against:** The total number of penalty yards as a result of the penalties commited by the given team in the given game.
- **O__Fumbles:** The number of fumbles lost by the given team's offense in the given game. Note: This is the the fumbles lost, not just offensive fumbles.
- **O__Interceptions:** The number of interceptions thrown by the given team's offense in the given game.
- **O__Turnovers:** The total number of turnovers by the given team's offense in the given game.
- **D__Pass__Comp:** The number of pass completions allowed by the given team's defense in the given game.
- **D__Pass__Att:** The number of passes attempted against the given team's defense in the given game.
- **D__Pass__Pct:** The passing percentage allowed by the given team's defense in the given game.
- **D__Pass__Yds:** The number of passing yards allowed by the given team's defense in the given game.

- **D__Pass__TD:** The number of passing touchdowns allowed by the given team's defense in the given game.
- **D__Rush__Att:** The number of rushing attempts against the given team's defense in the given game.
- **D__Rush__Yds:** The number of rushing yards allowed by the given team's defense in the given game.
- **D__Rush__Avg:** The average number of rushing yards per attempt given up by the given team's defense in the given game.
- **D__Rush__TD:** The number of rushing touchdowns allowed by the given team's defense in the given game.
- **D__Total__Plays:** The total number of plays ran against the given team's defense in the given game.
- **D__Total__Yds:** The total number of yards gained against the given team's defense in the given game.
- **D__Yards__Per__Play:** The average number of yards given up per play by the given team's defense in the given game.
- **D__Pass__First__Down:** The number of first downs surrendered via the pass by the given team's defense in the given game.
- **D__Rush__First__Down:** The number of first downs surrendered via the rush by the given team's defense in the given game.
- **D__Penalty__First__Down:** The number of first downs surrendered via penalty by the given team's defense in the given game.
- **D__Total__First__Down:** The total number of first downs surrendered by the given team's defense in the given game.
- **Penalties__For:** The total number of penalties committed by the given opponent in the given game.
- **Penalty__Yards__For:** The total number of penalty yards as a result of the penalties committed by the given opponent in the given game.
- **D__Fumbles:** The number of fumbles recovered by the given team's defense in the given game.
- **D__Interceptions:** The number of passes intercepted by the given team's defense in the given game.
- **D__Turnovers:** The total number of turnovers taken away by the given team's defense in the given game.

Note: The Defensive statistics for a team in a given game are the same as the Offensive statistics for the opponent in the given game.

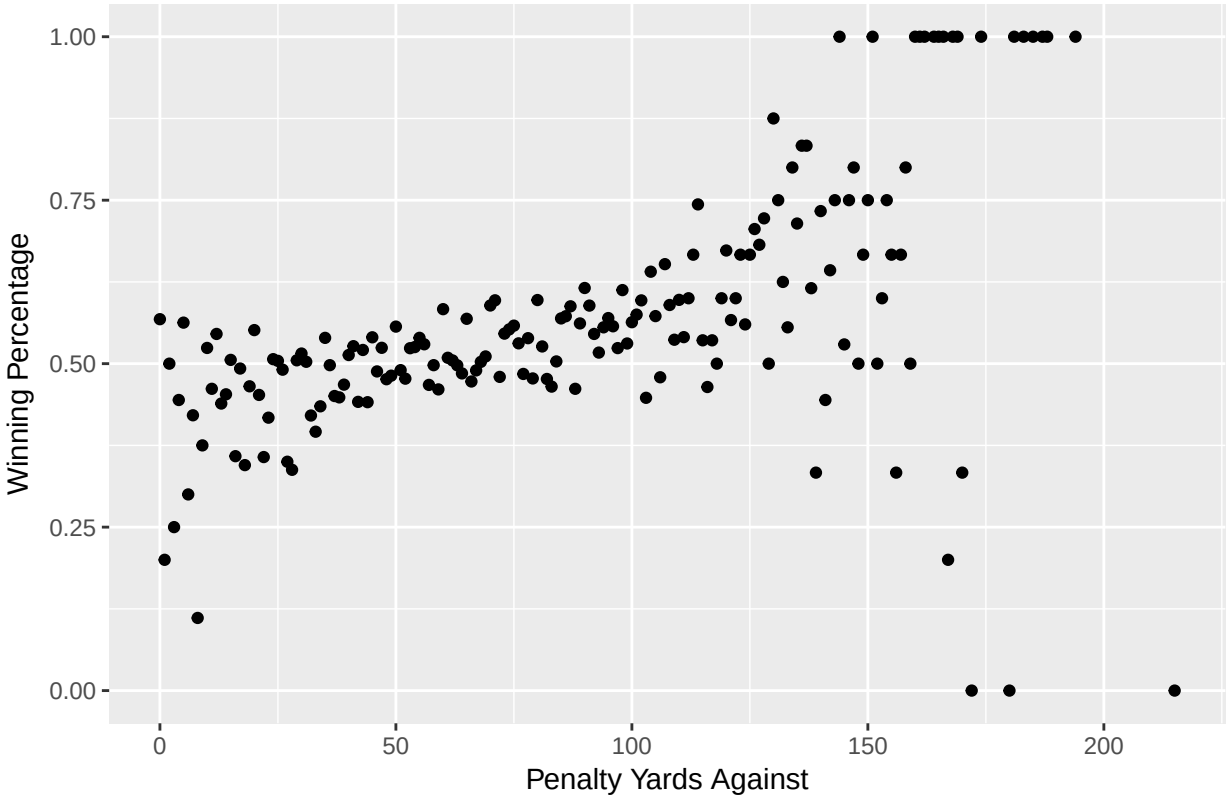

Figure 15: A scatterplot displaying how a team's penatly yards is related to winning percentage. It isn't surprsing that as team's gain more yards via penalty, the winning percentage increases. What is surprising is how flat the relationship appears to be (gaining 125 yards via penalty doesn't seem to drastically increase winning percentage from gaining only 50 yards via penalty).

Example Graphs

The Game_Logs data set allows for a more detailed look at each individual game. This allows for more detailed visualizations. One of the easier things to visualize is how certain variables may influence a team's winning percentage. Figure 15 shows how winning percentage varies by Penalty Yards. Figure 16 shows how winning percentage varies by turnover margin, and Figure 17 shows how winning percentage varies by Offensive yards per play. Another interesting analysis someone can do is look at scatter plots to observe the relationship between variables. For instance, Figure 18 shows the relationship between Offensive Yards per play and Points scored. It is also possible to look at the relationship between variables across weeks. As an example, Figure 19 shows the relationship between a defense's yards given up per rush and the previous week's rushing attempts against that defense.

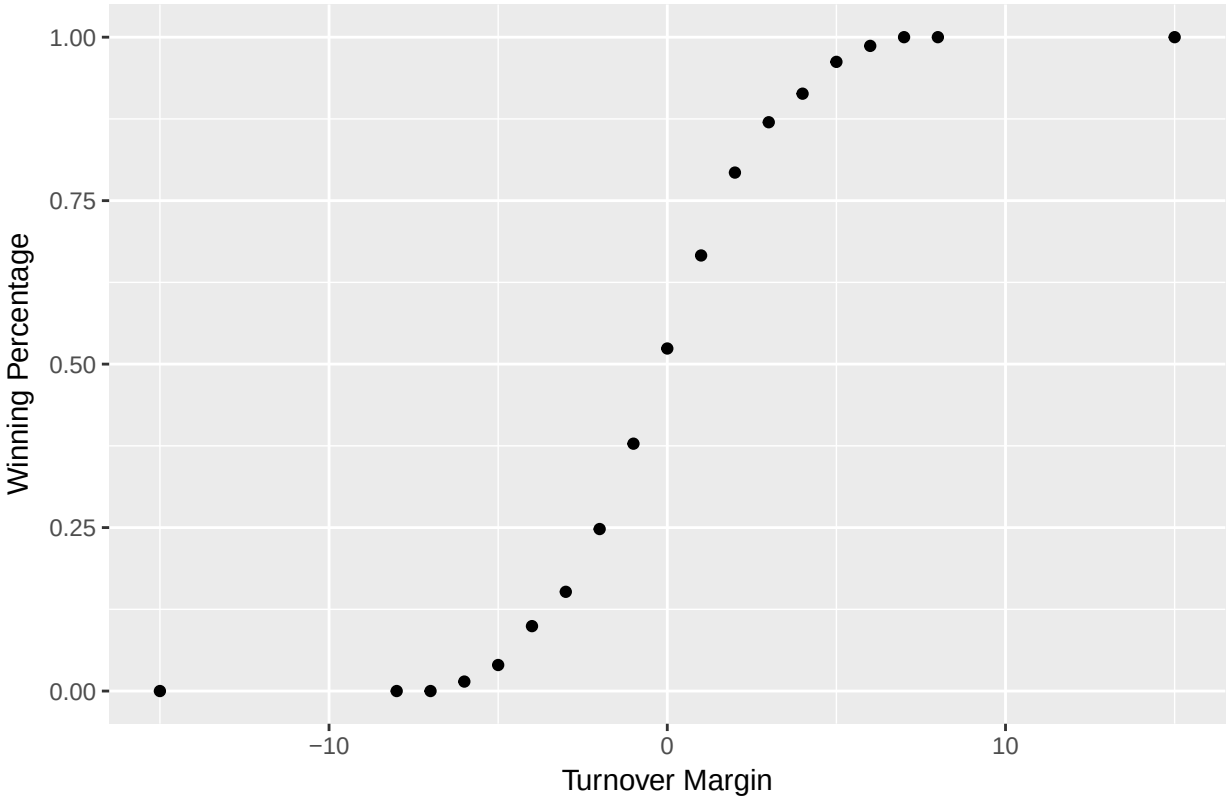

Figure 16: A scatterplot displaying how a team's turnover margin is related to winning percentage. Turnover margin is found by taking Opponent Turnovers minus the Team's Turnovers. It isn't a surprise that as a team commits less turnovers than their opponent (resulting in a positive turnover margin), the winning percentage increases.

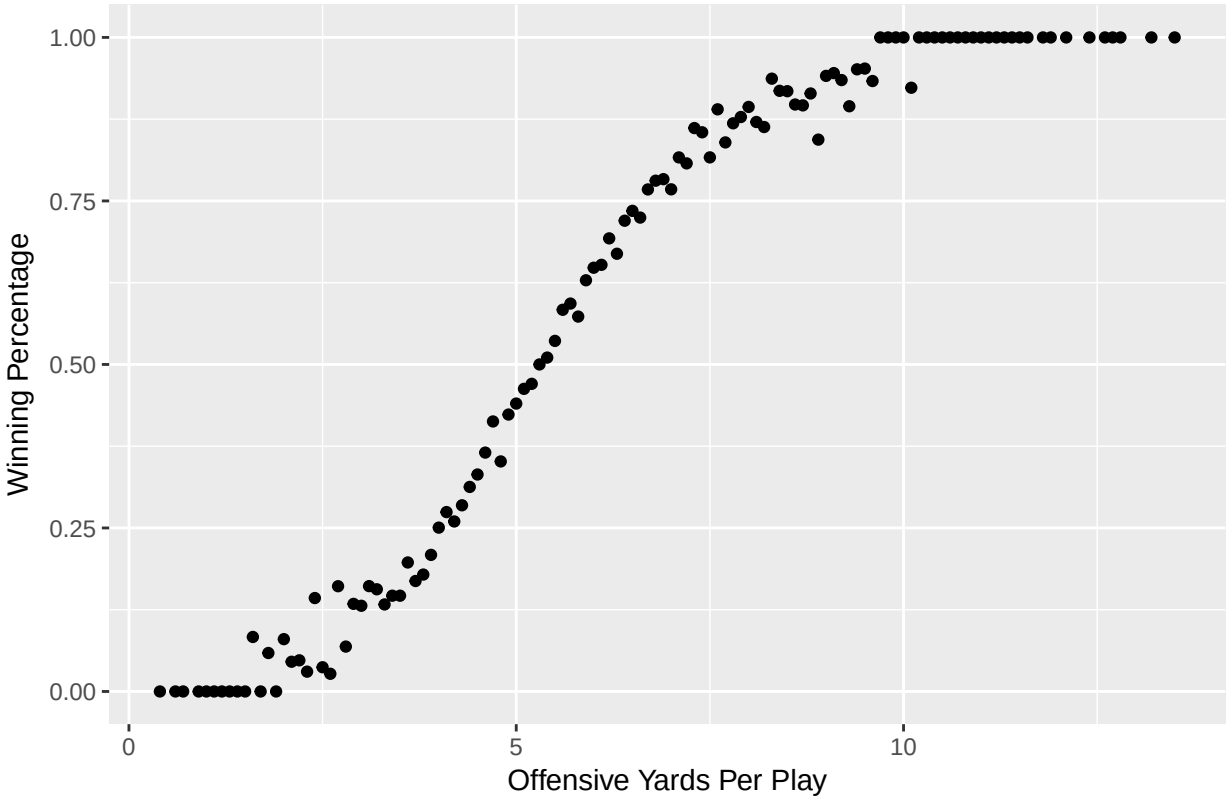

Figure 17: A scatterplot displaying how a team's offensive yards per play is realted to winning percentage. It isn't surprising that the more yards a team gains per play, the higher the winning percentage is. What is a little surprising is how strong the trend appears to be, following a fairly straight line.

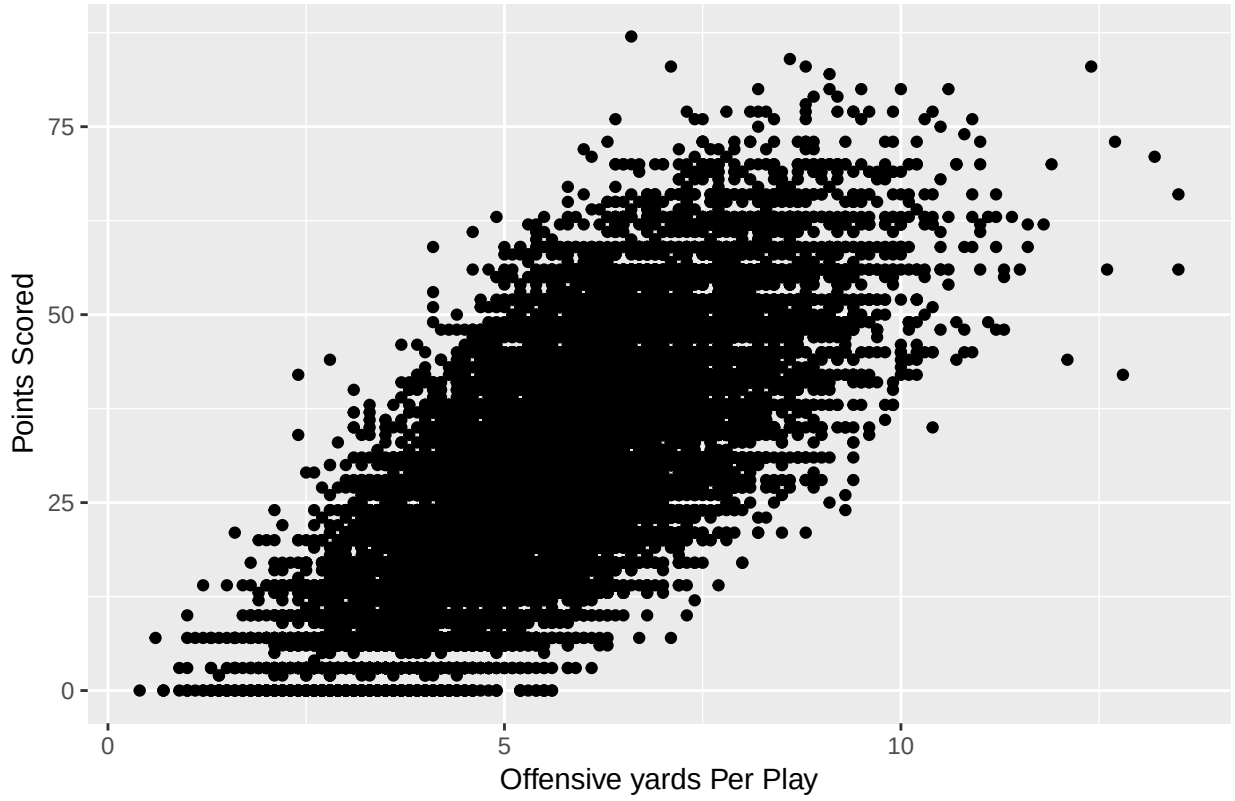

Figure 18: A scatterplot showing the relationship between a team's offensive yards per play and points scored. Unsurprisingly, offenses that gain more yards per play tend to score more points.

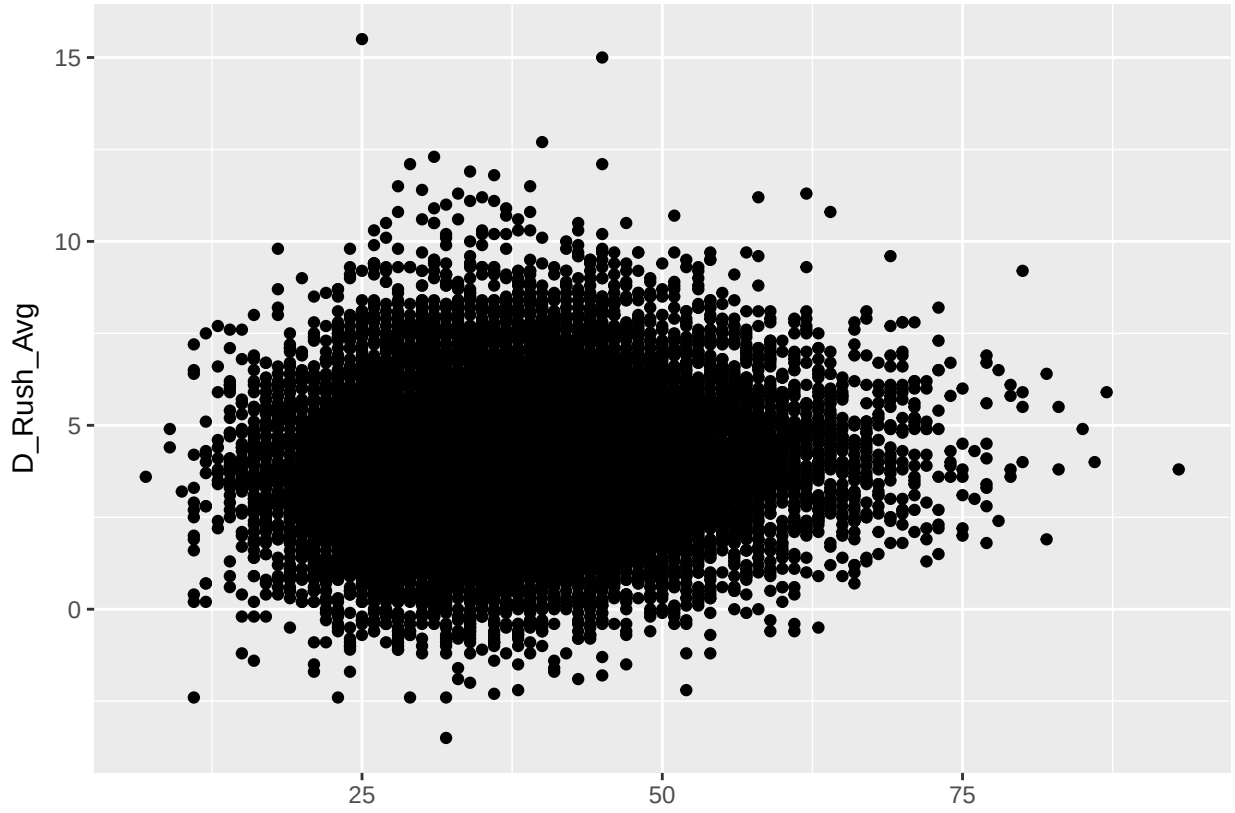

Figure 19: Relationshp between a defense's yards given up per rush and the last week's rushing attempts. Commentators oftentimes claim that a defense that had a lot of rush attempts the week before may be less successful in stopping the run the following week. If this was true, one would expect to see a relationship between the previous week's defensive rushing attempts and the yards given up per rush the following week. The scatterplot above does not seem to show an obvious relationship between these two variables.

Data Set Conclusion

The five csv files that were created (Total_Team_History, Individual_Season_Results, Season_Averages, Game_Results, and Game_Logs) provide clean, formatted, and complete sets of data. As demonstrated above, these data sets give the resources to explore many different trends or relationships seen in college football. One specific relationship that could be explored is how a team's Associated Press ranking is related to the team's win probability in a given game.