MSCI:3250 Final Project Report

2019 College Football Statistics:

A Journey to Discover the Hidden Statistics of College Football

Friday May 8th, 2020

Written and Presented By:

Ben Ahnen, Mark Conway, Jarrin Flores, Jonas Geerdes, and Natalie Lopez

1. Introduction

College football is a very popular sport to watch in America every fall. Every year, millions of fans tune in, attend games, and or support their college teams. This project will be exploring college football data in the United States and the differences in team play for the 2019 season. As University of Iowa students, we are very attached to our Iowa Hawkeye football team. Due to our love of the Hawkeyes, we became interested to see how many statistically significant patterns we could discover within college football.

In this project, we utilized college football data from two different sources and create queries and visuals to discover statistical significance in our data. We wanted to find smaller details that directed towards what geographic area of the country excels in college football, what elements contribute to that team's success, and if their ranking had an influence on the overall success of the team in 2019.

2. Data

We used two different datasets in this project. The first dataset used was 'College Football Team Stats 2019' that looks at the 130 FBS level teams from Kaggle. The data consists of statistics pertaining to offense, defense, and special teams.

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

The second dataset we used was taken from sportsreference.com using the rvest package and contained additional college football data including poll rankings which will be elaborated on in the analysis section of our report. The link below is the URL used in the code for web scraping.

https://www.sports-reference.com/cfb/years/2019-standings.html

2.1 Conference Win Percentage

In order to collect data solely on the conferences of the teams, we utilized the first data set of 130 teams and added our own columns for state based on the name and location of the university. We used data from the United States Census Bureau to dictate the state (https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf). The data used to predict the success of these conferences was based on the win percentage of the team.

2.1.5 Regional Win Percentage

In order to collect data solely on the team regions, we utilized the first data set of 130 teams and added our own columns for state and region based on the name and location of the university. We used data from the United States Census Bureau to dictate the regions (https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf). The data used to predict the success of these regions was based on the win percentage of the team.

2.2 Statistical Analysis on Time of Possession

The Kaggle dataset contained a statistic regarding the time of possession rank of the teams in the dataset. This variable paired with other data columns, such as offensive rank and defensive rank, which made for important statistical analysis that will be detailed in the third section, Analysis. All three selected data points were in a ranking style, assigned to each of the 130 collegiate teams in the dataset.

2.3 AP Top 25 Poll by Conference

A data frame contained the rankings of the teams in the AP Preseason and Final Rankings. Two bar plots were then utilized to compare and display the number of teams each conference had in the AP preseason and the final top 25 ranking for the 2019 season. Each bar plot shows how many teams as well as their ranking for the conference. This will be further analyzed into the conferences of importance in the Analysis section.

2.4 Data Types and Description

The data was processed in R by changing the names of certain teams in one dataset to match the other dataset. There were also columns that were dropped in R that were not needed for this project. Data types were also changed for some of the columns in the final data frame. Table 1 below is a data dictionary describing the columns of data in our dataset

Table 1:

Column	Type	Description
Team	text	The name of the team/college
Conference	factor	The conference in which a team plays
Games	numeric	Number of games played
Wins	numeric	Number of wins
Losses	numeric	Number of losses
Off.Rank	numeric	Rank of the Offense
Off.Plays	numeric	Total number of offensive plays
Off.Yards	numeric	Total number of offensive yards
Off.TDs	numeric	Total number of offensive touchdowns
Off.Yards.per.Game	numeric	Average number or offensive yards per game
Def.Rank	numeric	Rank of the Defense
Def.Plays	numeric	Total number of defensive plays
Yards.Allowed	numeric	Total number of yards allowed against a team
Off.TDs.Allowed	numeric	Total touchdowns that the defense allowed the opposing offense to score
Total.TDs.Allowed	numeric	Touchdowns that the other team scored against the team including special team scores
Yards.Per.Game.Allowed	numeric	The number of yards allowed per game by opposing team
First.Downs	numeric	Number of first downs the offense had
Interceptions.Thrown.x	numeric	Number of interceptions thrown by a team
Pass.Yards	numeric	Total number of passing yards for a team
Pass.Touchdowns	numeric	Total number of pass touchdowns for a team
Pass.Yards.Per.Game	numeric	Average number of passing yards per game by a team
Pass.Def.Rank	numeric	Rank of Defense against passes
Pass.Yards.Per.Game.Allowed	numeric	Average number of passing yards allowed by the opposing team per game
Rushing.Def.Rank	numeric	Rank based on number of rushing yards defense allowed opponent

Opp.Rush.Touchdowns.Allowed	numeric	Number of opponent rush touchdowns allowed by a team
Rushing.Off.Rank	numeric	Rank of Offense based on runs
Rushing.TD	numeric	Total rushing touchdowns scored by a team
Scoring.Def.Rank	numeric	Rank of Defense based on scoring
Touchdowns.Allowed	numeric	Total touchdowns scored by opposing offense
Points.Allowed	numeric	Total points allowed by opposing offense
Avg.Points.Per.Game.Allowed	numeric	Average number of points a team allowed an opponent per game
Scoring.Off.Rank	numeric	Rank based on offensive scoring
Touchdowns	numeric	Total number of touchdowns scored by a team
PAT	numeric	Total point after touchdown attempts
X2.Point.Conversions	numeric	Total number of two point conversions scored by a team
Field.Goals	numeric	Total number a field goals made by a team
Total.Points	numeric	Total number of points scored by a team
Points.Per.Game	numeric	Average points scored per game by a team
Time.of.Possession.Rank	numeric	Rank based on total time of possession
Turnover.Rank	numeric	Rank based on total number of turnovers by a team
Fumbles.Recovered	numeric	Total number of fumbles recovered by a team
Conference.Wins	numeric	Total number of Conference wins by a team
Conference.Losses	numeric	Total number of Conference losses by a team
Conference.Win.Loss.Percentage	numeric	Win-Loss percetange for a team based on Conference games
Simple.Rating.System	numeric	A rating that takes into account average point differential and strength of schedule
Strength.of.Schedule	numeric	A rating based on the strength of a teams schedule
AP.Preseason.Rank	numeric	Rank in the preseason AP poll
AP.Highest.Rank	numeric	Highest rank achieved in the AP poll
AP.Final.Rank	numeric	The final rank for the AP poll
State	text	The state in which a team is from
Region	Factor	The Region in which a team are from.

3. Analysis

3.1 Conference Win Percentage Analysis

The SEC is typically regarded as the most dominant conference in college football. We decided that we wanted to apply data analysis to determine if this assumption holds true. As University of Iowa students, we believe that the Big Ten is also up there as one of the country's best conferences. To assess this, we chose to analyze the conferences by average wins and win percentage. We pulled the conferences from 'College Football Team Stats 2019'. Once we had the teams broken into their conferences, we used the variables total games played by conference teams, total wins, total losses, average wins by the teams in the conference, and the win percentage of the conference. We used win percentage because conferences had varying amounts of bowl game participants, which gave average wins an unfair edge for conferences that played "extra" games. Through the data that we generated, we found that the SEC does have the highest win percentage, beating out the Big Ten by 2.4 percentage points. The average wins almost match the order of win percentage identically; the only difference is that the Pac-12 has a win percentage of .6 percentage points higher than the ACC, while the ACC average .12 more wins.

Conference	TotalGames	TotalWins	TotalLosses	MeanWins	WinPerc
SEC	180	107	73	7.643	59.4
Big Ten	179	102	77	7.286	57.0
AAC	153	87	66	7.250	56.9
Big 12	128	71	57	7.100	55.5
Mountain West	154	84	70	7.000	54.5
Pac-12	153	83	70	6.917	54.2
ACC	181	97	84	6.929	53.6
Sun Belt	127	66	61	6.600	52.0
C-USA	178	84	94	6.000	47.2
MAC	153	70	83	5.833	45.8
FBS Independent	76	34	42	5.667	44.7

This image shows the conferences of the 130 FBS teams, and is sorted by win percentage, high to low

3.1.5 Regional Win Percentage Analysis

The South is typically regarded as the most dominant football region in the United States. We decided that we wanted to analyze if that assumption holds true. As students at the University of Iowa, we believe that the Midwest, a region that has a majority of the Big Ten teams located in it, is comparable to the South. We used the United States Census Bureau to dictate the regions that we broke the teams into. Once we had the teams broken into their regions, we pulled the total games played by those teams, total wins, total losses, mean wins by teams, and the win percentage of the region. We used win percentage because regions have varying amounts of bowl game participants, as well as a wide range of total games played. Through the data that we generated, we found that the South does have the highest win percentage, beating out the Midwest by a mere by .3 percentage points. The average wins match the order of win percentage identically; the Midwest has .026 less wins than the South. While the South barely edges out the Midwest in terms of average wins by team and conference win percentage, the difference is so small that it is statistically insignificant.

Region	TotalGames	TotalWins	TotalLosses	MeanWins	WinPerc
South	845	459	386	6.955	54.3
Midwest	359	194	165	6.929	54.0
West	332	176	156	6.769	53.0
Northeast	126	56	70	5.600	44.4

This image shows the breakdown of which region the 130 FBS teams are located, and is sorted by win percentage, high to low

3.2 Time of Possession Statistical Analysis

Time of possession can be a determinant statistic in college football. The ability for a team to dominate the time of possession on offense often leads to many lopsided games in both the

collegiate and professional levels. The Iowa Hawkeyes are infamous for their ability to piece together long, time consuming drives that aids their ability to convert on both sides of the ball. However, as a group we wanted to dive into the effects of a time of possession rank versus a team's offensive and defensive performances, which are summarized with their overall offensive and defensive ranks.

In order to try and find the relationship between these variables, we ran correlation tests and fitted linear regression models using time of possession rank versus offensive and defensive rank that is available in the base R package.

The first test was between time of possession rank and the offensive rank. When running the correlation test, we were moderately surprised to find that there was extremely weak correlation between these two ranks. The R-value after correlation testing was .0545, indicating that there was little to no relation between the two variables.

Our hypothesis testing provided similar results. We hypothesized that time of possession would be a good predictor of offensive rank, represented by predicting the P-value < .05. The P-value from the linear fit testing was .5377. Because of these results, we rejected our hypothesis of the possession of time ranking being a good predictor of offensive ranking. We were surprised by these findings as we believed that a strong time of possession rank represented a team controlling the offensive side of things, giving them more time and opportunity to score points. Another theory is that the longer a defense is on the field against one of these strong time of possession teams, the more worn out they would become.

After testing the offensive side, we decided it would be best to also test the relationship between a team’s defensive prowess and their ability to possess the ball for long amounts of time on offense. Like our testing for offensive, we tested correlation and conducted a hypothesis test using linear regression. After running the correlation test, the R-value returned was .5272, which represented moderate correlation between the two variables.

We hypothesized that the P-value between the two ranks would be < .05, representing time of possession being a strong predictor of defensive rank. The resulting P-value was 1.16x 10^-10, showing that time of possession was a very strong predictor of defensive rank which led to us accept our null hypothesis. These results would represent that a higher time of possession game does predict a good defense.

Conclusively, the statistical analysis for the time of possession rank had important findings, as both the hypothesis testing revealed that the time of possession rank is not a good predictor of offensive ranking but is a good predictor of defensive rankings.

3.3 AP Top 25 Comparisons

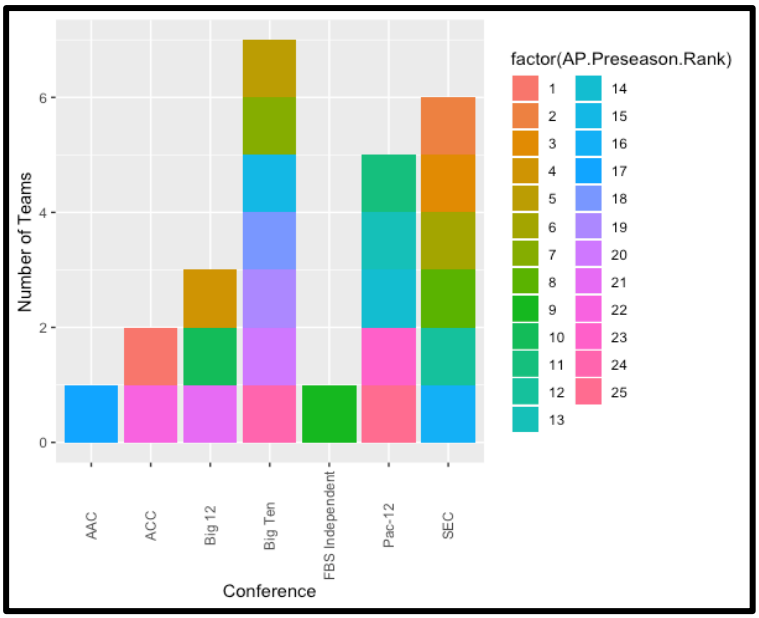
The final question we posed was how does the Top 25 AP Preseason Rankings compare to the final rankings at the end of the college football season? After placing information into a data frame containing relevant details of the teams, we became interested on comparison of the conferences.

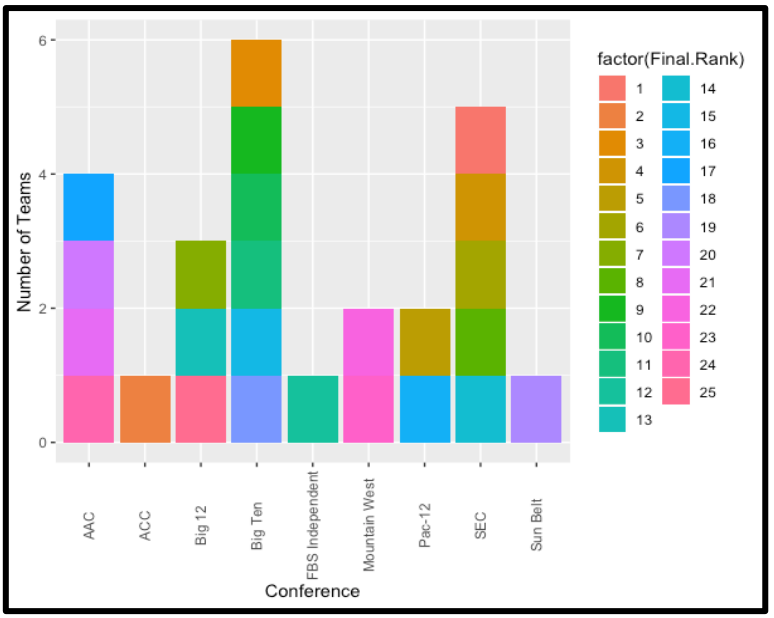
This is based upon the notion that the SEC leads the college football league based upon performance. Preseason and final ranking charts displayed the findings of the data and the truth behind how these teams perform.

	Team	Conference	AP.Preseason.Rank	Final.Rank	Win.Percentage	Offensive.Rank	Defensive.Rank	Schedule.Strength
1	Air Force	Mountain West	NA	22	84.61538	51	17	-2.22
2	Alabama	SEC	2	8	84.61538	6	20	2.81
3	Appalachian State	Sun Belt	NA	19	92.85714	39	26	-3.80
4	Auburn	SEC	16	14	69.23077	64	28	7.72
5	Baylor	Big 12	NA	13	78.57143	52	39	2.62
6	Boise State	Mountain West	NA	23	85.71429	48	33	-2.28
7	Cincinnati	AAC	NA	21	78.57143	80	40	2.51
8	Clemson	ACC	1	2	93.33333	5	6	2.70
9	Florida	SEC	8	6	84.61538	45	9	2.91
10	Georgia	SEC	3	4	85.71429	61	3	5.21
11	Iowa	Big Ten	20	15	76.92308	99	12	4.44
12	Iowa State	Big 12	21	NA	53.84615	27	45	4.53
13	LSU	SEC	6	1	100.00000	1	31	6.60
14	Memphis	AAC	NA	17	85.71429	10	61	2.09

The image above is a screenshot of part of the Ranking data frame

It is important to note that rankings are based upon more than final scores and wins. A voting committee made up of sports writers and broadcasters considers several different factors of a team's performance and then place their votes on where the teams should be ranked in the preseason and then week by week. Depending on the number of votes for a team's ranking depends on where it is placed. For example, 60 out of 61 votes for Iowa ranked at 13 will put Iowa's rank that week at 13.





The following charts are the total teams ranked by conference with their ranking during the preseason and final ranking displayed in the legend

The preseason and final ranking charts both show the Big Ten leading in number of teams ranked, with the SEC one team behind. If the dominant conference was based upon number of teams, wouldn't the Big Ten be the winner? Looking further into the data we can see that while the Big Ten has the most teams, they are not necessarily the highest ranking. The Big Ten has two teams within the top ten and the SEC has four of its five in the top ten. Exploring the final ranking subset, we learned that these four teams either have an offensive or defensive rank within the top ten of each category. These high rankings compared to the middle rankings of the majority of Big Ten teams in both the defensive and offensive rank sets these two conferences apart from the data displayed.

The findings of the data align with the saying, "quality over quantity." While the Big Ten may have more teams in the top 25, the question of their ability to compete with the SEC is still in question. The SEC may not dominate in quantity based upon number of rankings, but from the data displayed it does dominant in quality.

4. Conclusion

This data has been pulled Kaggle, which gave statistics and ranks of the 130 FBS College Football teams during the 2019 season; Sports Reference, which allowed us to look at the Pre-Season Top-25 and final rankings for the 2019 season; and the US Census Bureau, which allowed us to break our conferences into states and regions. From this information we were able to generate summary tables, visuals, graphs, and statistical tests. Theses tables, visuals, and graphs were then used to analyze aspects of college football. We found that SEC teams averaged more wins as well as a higher win percentage than any other conference. However, that can be attributed to the quality of the teams that the SEC had, having four of their teams in the final Top-10. Meanwhile, the Big Ten had more teams in the Top-25 to end the season. Additionally, we found that although the South is regarded as the most dominate region in the United States, the Midwest rivaled it very closely. Southern teams averaged only .026 more wins and had a .003 higher win percentage. From the statistical analysis we found little evidence of a team's time of possession rank affecting both their defensive and offensive rankings from the dataset. These tests upset some previously thought assumptions our group had about these rankings.

There are limitations to this project that should be acknowledged. The first limitation is that the data gathered and analyzed was only from the competition year 2019. More seasons statistics could add to or change each of the topics analyzed. Another limitation is that not all the variables were used in order to come to conclusions. From the original dataset, we deemed many of the variables provided irrelevant to our questions asked, but these variables could have added to the analysis.

References

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

https://www.sports-reference.com/cfb/years/2019-standings.html

https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

https://www.ncaa.com/news/football/article/2019-07-08/college-football-rankings-every-poll-explained-how-they-work