From Stats to Standings: A Data-Driven Look at College Football Team Performance

Final Analysis Report BAIS:3250 Data Wrangling

Emily Sanders, Emily Nell, Chaitanya Patel

To Do Reminder

* Narrow down analysis questions
* Finish analysis
* Conclusion
* GitHub

**1 . Introduction**

College football is a cornerstone of university athletics in the United States, drawing widespread attention from students, alumni, fans, and media. Under the National Collegiate Athletic Association (NCAA), teams compete across multiple divisions, each facing different competitive landscapes shaped by variations in institutional size, funding, recruitment strength, and coaching philosophies. These factors result in a wide range of team performances, making it challenging to define a single formula for success. One of the most debated questions among coaches, sports analysts, and fans is whether high-performing teams maintain a balanced approach between both offensive and defensive play or whether teams can achieve success by strongly outperforming in just one of those areas.

We aim to investigate whether there is a measurable correlation between a team’s offensive yards, defensive yards, and overall success through win-loss records and average margin of victory. Our initial step involves identifying which teams had the highest average offensive and defensive yards per game and examining how these variables correlate with each other. Following this, we will combine our data sources into a single dataset that allows us to evaluate how variations in offensive and defensive output relate to key indicators of team success.

By analyzing these metrics, we hope to answer whether statistical dominance in one area can compensate for weaknesses in another, or whether true success lies in a balanced and consistent approach. Ultimately, our analysis will offer insights not only into what drives winning seasons but also into how coaches might prioritize training and strategy development based on data-driven performance indicators.

**2. Data**

*2.1 NCAA*

This analysis relies on data from two distinct sources related to college football performance statistics. Each source provides a different perspective on team performance, with the overall focus being on NCAA Division 1 teams. Our first dataset comes from the NCAA statistics portal[[1]](#footnote-1) for NCAA college athletics, including football programs from every division. The source provides rankings and detailed performance statistics at both the team and individual player level for all participating teams throughout seasons. The data is made available and updated as needed for each season. The source includes team data in key performance areas that can be utilized by coaches, analysts, and universities to gauge team performance. Each data entry is displayed in a ranking format, allowing for easy comparison across teams. The source allows users to filter by week, season, and performance category. This source provides as our core dataset within our analysis to evaluate patterns of team performance and explore the relationship between offense stats, defense stats, penalty stats, and other statistics that contribute to a team’s overall success. As this is an official NCAA-managed resource, the data is highly reliable. To collect the data from this website, we had to crawl through each URL provided on the page that provided specific statistics on a certain team characteristic (i.g. Offense, Defense, Passing, etc.). From each of the URLS crawled, we scraped the main data table on the webpage to store it into a data frame.

*2.2 BetIQ*

Our second data source comes from BetIQ[[2]](#footnote-2), a sports analytics platform that provides information for historical betting and winning percentages. The dataset also includes the average margin of victory (MOV) for each team. Sports betters use this statistic to make data-driven guesses as to the point margins when placing their bets. Additionally, coaches can utilize MOV to predict score outcomes and better plan against teams with higher MOVs. This source allows us to view these statistics and provides our dataset with additional valuable information for our analysis. To collect the data table from the website, we read in the html and uploaded the data table into a data frame. Below is the second source link, and the second source citation is in our references page.

*2.3 Combining NCAA and BetIQ*

For our analysis, we began by combining our two separate datasets into a single unified DataFrame to enable a more comprehensive examination of team performance. One dataset contained team statistics, such as offensive and defensive yards per game, while the other included performance metrics like wins, losses, and win percentage. To prepare for merging, we first ensured both datasets had a common column, which was team name. We had to clean the “Team” columns to troubleshoot any inconsistencies by column by stripping whitespace, adjusting capitalization, and formatting differences to avoid mismatches. We then used an inner join on the "Team" column with the Pandas merge() function, creating the final merged\_df DataFrame. This process is shown in *Project\_DataScraping.ipynb*. This merged dataset allowed us to analyze how statistical metrics, such as offensive yards per game, relate to actual performance outcomes like win percentage.

Table 1 Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | | **Type** | **Description** |
| Teams | | Text | College Name |
| Conference | | Text | Conference in which the team belongs |
| Wins | | Numeric | Number of games won in the regular season |
| Losses | | Numeric | Number of games lost in the regular season |
| Win % | | Numeric | The percentage of a team’s wins compared to a team’s losses in the regular season |
| MOV | | Numerical | Average margin of victory (points) |
| ATS +/- | | Numerical | Against the spread record sports betting displayed as covered-not-covered-tie |
| MOV\_Status | Text | Categorized MOV based on whether a team has a MOV >=0 (Positive) or < 0 (Negative) | |
| G (Games played) | | Numerical | Total number of games played in the regular season |
| Offensive Rank | | Numerical | Determined rank based on a variety of offensive metrics and season performance |
| Offense TDs | | Numerical | Total number of combined offensive touchdowns in a regular season |
| Total Offense YPG | | Numerical | Season average of offensive yards per game |
| Rushing Rank | Numerical | Determined rank based on a variety of offensive rushing metrics and season performance | |
| Rushing TDs | Numerical | Total number of rushing touchdowns in a regular season | |
| Rush YPG | Numerical | Season average of offensive rushing yards per game | |
| Passing Rank | Numerical | Determined rank based on a variety of offensive passing metrics and season performance | |
| Pass YPG (Passing yard per game) | Numerical | Season AVG of Passing Yards / games | |
| Passing TDs | Numerical | Total number of passing touchdowns in a regular season | |
| Int (Interceptions) | Numerical | Total number of picked passes (interceptions) thrown from the offense in a regular season | |
| Total Defense YPG | | Numerical | Season average of defensive yards per game |
| Defensive Ranking | | Numerical | Determined rank based on a variety of defensive metrics and season performance |
| Opp TDs (Opponent Touchdowns) | Numerical | Total number of opponent’s touchdowns throughout the regular season | |
| Rushing Defensive Rank | Numerical | Determined rank based on a variety of metrics involving how a defense performs against an offense’s rushing plays | |
| Opp Rush TDs (Opponent Rushing Touchdowns) | Numerical | Total number of opponent’s rushing touchdowns throughout the regular season | |
| Opp Rush YPG (Opponent Rushing Yard per game) | | Numerical | Season AVG of the total number of yards gained by the opposing team through running plays during a game |
| Passing Defensive Rank | Numerical | Determined rank based on a variety of metrics involving how a defense performs against an offense’s passing plays | |
| Opp Pass TDs | Numerical | Total number of opponent’s passing touchdowns throughout the regular season | |
| Opp Pass YPG | Numerical | Season AVG of the total number of yards gained by the opposing team through passing plays during a game | |
| TM Rank | Numerical | Determined rank based on a variety of turnover metrics and season performance | |
| Fumbles Recovered | Numerical | Total number of fumbles recovered in a regular season | |
| Fumbles Lost | Numerical | Total number of fumbles lost in a regular season | |
| Turnover Margin | Numerical | Total number of giveaways (interceptions & fumbles lost) from the total number of takeaways (interceptions & opponent fumble recoveries) in a regular season | |
| PPG Rank (Penalties Per Game Rank) | Numerical | Determined rank based on a variety of penalty metrics and season performance | |
| Penalties | Numerical | Total number of penalties in a regular season | |
| Penalties Per Game | Numerical | Season AVG of the number of penalties a team receives in a game | |
| Penalty YPG Rank | Numerical | Determined rank based on a variety of penalty metrics and season performance | |
| Penalty YPG | Numerical | Season AVG of the total number of yards lost because of penalties | |
| TOP Rank | Numerical | Determined rank based on a variety of penalty metrics and season performance | |
| Avg Time of Possession | Time | Season AVG of a team’s offensive time of possession during games | |
| ATP Rank | Numerical | Determined rank based on a team’s average time of possession compared to other teams in the NCAA | |
| Season | | Numerical | The football season in which the team's data was collected |
| Season\_Date | Date | Dummy date variable for Time Series analysis with Season field | |

**3. Analysis**

*3.1 Descriptive Analytics of Yards and Win Percentage*

The dataset contains observations on multiple football teams, including average offensive yards per game, defensive yards allowed per game, and win percentage. A preliminary inspection of the data reveals that teams vary considerably in both offensive and defensive performance. According to the interquartile range offensive yardage typically ranges from 348.95 to 423.50 yards per game, with a mean of 384.37. Defensive yards allowed span a similar range from 335.95 to 404.10 with a mean of 372.52, but in the opposite direction as teams strive to allow fewer yards, and lower averages are typically indicative of a stronger defense.

Win percentage across the dataset varies from 0% to 92.9%, with the average winning percentage landing at 52.44%. Our regression displayed in Figure 1 shows that teams achieving higher offensive yardage tend to win more frequently. Likewise, Figure 2 displays those that allow fewer yards on defense also tend to win more games. However, raw yardage alone is unlikely to be the sole determinant of winning. It is important to quantify these relationships more formally.

However, the Margin of Victory (MOV) also played an important role in our analysis. Unlike yardage metrics, which reflect field movement, MOV directly captures scoreboard outcomes, integrating both offensive and defensive performance. In this dataset, MOV ranges from -30.20 to 25.7, with an average of 1.47. Shown in Figure 3, teams with higher MOVs generally outperform opponents consistently and are more likely to win by large margins. While two teams may share similar win percentages, a higher MOV often indicates greater game control or consistency. MOV is included in the machine learning models to test whether it strengthens predictive performance when paired with yardage metrics.

This analysis is show in *Project\_Analysis.ipynb.*

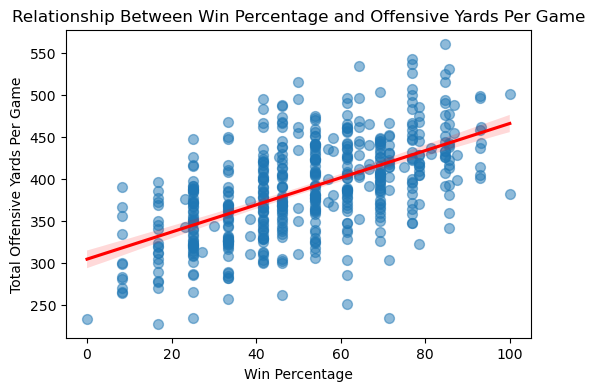


Figure 1 Offensive Yards Regression

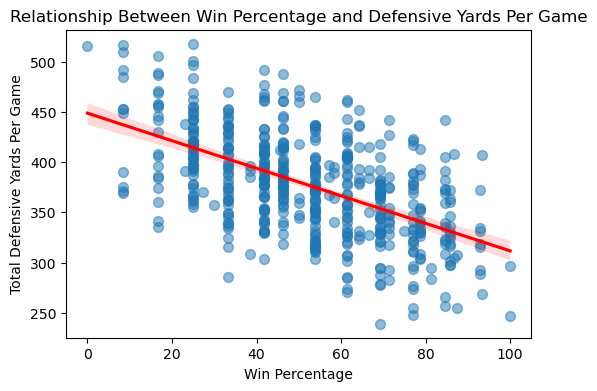


Figure 2 Defensive Yards Regression

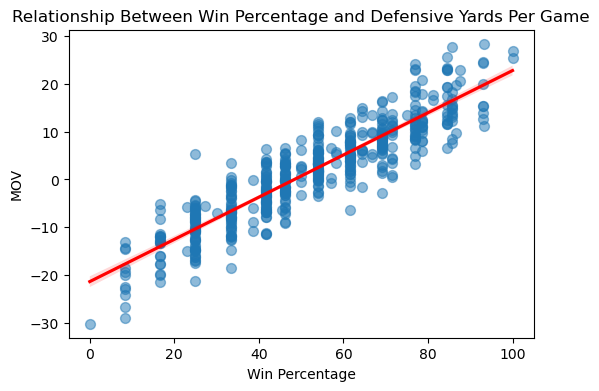


Figure 3 MOV Regression

*3.2 Hypothesis Testing of Yards and Win Percentage*

We conducted a Pearson correlation test to evaluate the linear relationship between offensive yards per game and win percentage. The correlation coefficient was found to be 0.522, indicating a moderate positive relationship. The p-value associated with this test was 0.008, which is less than the standard significance threshold of 0.05. Therefore, we reject the null hypothesis and conclude that offensive yardage has a statistically significant association with winning outcomes. In other terms, this means that teams who are more effective in gaining yards on offense tend to win more frequently. However, the relationship is not perfect; teams with high offensive yardage may still lose if they struggle in other aspects of the game.

To evaluate the relationship between defensive yards allowed per game and win percentage, a Pearson correlation test was conducted. The correlation coefficient was -0.6013, indicating a moderate negative association as expected, teams that allow fewer yards on defense tend to win more often. However, the p-value for this correlation was 0.1114, which is greater than 0.05. Therefore, we fail to reject the null hypothesis suggesting that the observed relationship is not statistically significant at the conventional 5% level. This means that, while there appears to be a downward trend between defensive yardage and win percentage, we do not have sufficient statistical evidence to conclude that this relationship is statistically significant. In other terms, this may indicate that defensive yardage alone is not a reliable predictor of team success, and that other factors may play a more critical role than overall yards allowed.

*3.3 Machine Learning Predicting*

In this section, we applied regression-based machine learning models to predict the Win Percentage of college football teams using features such as Total Offense Yards Per Game and Total Defense Yards Per Game. Our goal was to determine how well different regression models can explain and predict team success based on offensive and defensive performance metrics. We used a 70/30 train-test split and the model to the same training and testing data for consistency. Model performance was evaluated using the **R²** score, which reflects the proportion of variance in the dependent variable explained by the independent variables. The r-squared score 0.5192 indicating that approximately 51.92% of the variance in college football teams' win percentages can be explained by the model using Total Offense Yards Per Game and Total Defense Yards Per Game as predictors. This analysis is present in *Project\_Analysis.ipynb*. However, this also means that 48.08% of the variance remains unexplained, highlighting that other factor such as turnovers influence win outcomes as well. While the model begins to create an understanding the impact of offensive and defensive yardage, we were able to make improvements by incorporating additional variables.

To incorporate additional variables, we scraped other factors from the NCAA source for seasons 2021-2024 and merged them into one DataFrame. This merge and scrape are presented in *scraped\_data.ipynb*.

*3.4 Time Series Forecasting*

In this section, we start aggregating each conference’s average marking of victory (MOV) by season from 2021 to 2024 and how it impacts the series and seasonality. We chose to run a classical season decomposition additive, which revealed a modest upward trend of about 0.3 points per season. Based on the analysis, there are no strong repeating seasonal swings. We used Holt’s linear exponential-smoothing model on each conference series using 2021-2023 for training and validating on 2024 to project and quantify the 2025-26 conference average MOV. Residuals yielded an average mean absolute error (MAE) of 0.4 and a root mean square error (RMSE) of 0.60, which indicates the model captures most of the year-to-year movement. Forecasted projections for the two seasons ahead league-wide MOV, for example, the SEC’s average MOV is projected to climb from .63 points in 2024 to approximately 8.85 in 2025 and 9.6 in 2026, while the Big Ten moves from 3.7 to about 3.5 and then 3.39. These modest but consistent increases suggest that, absent significant rule changes or strategic shifts, winning margins across FBS Independent conferences are likely to continue their gradual rise.

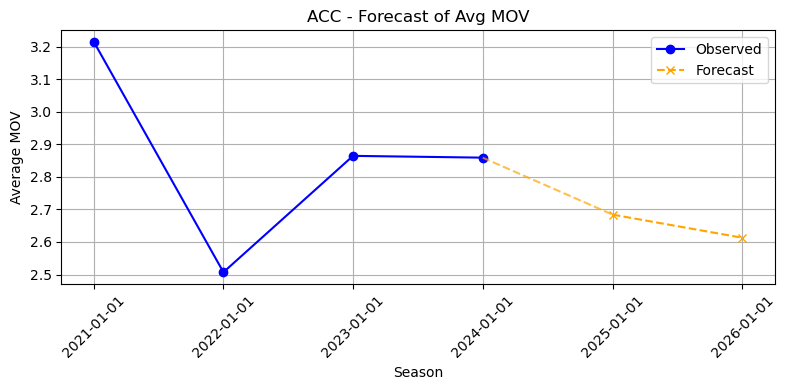


Figure 4 ACC Forecasted MOV

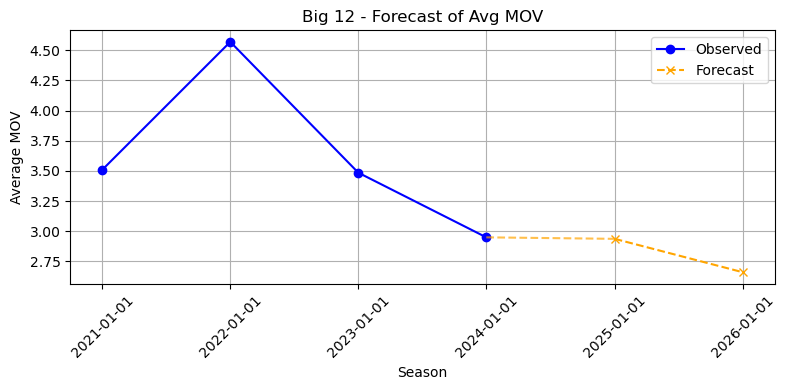


Figure 5 Big 12 Forecasted MOV

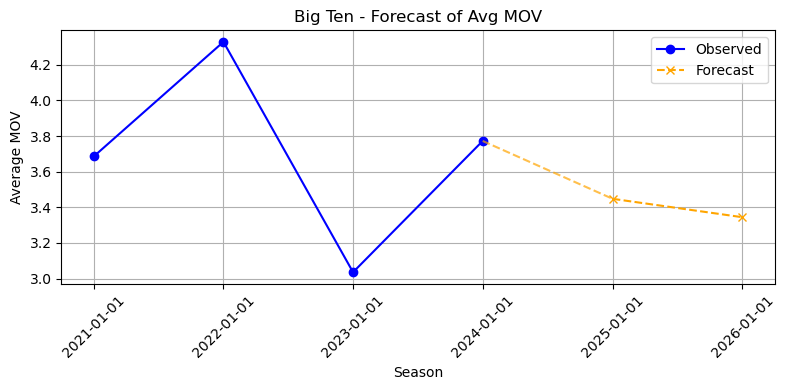


Figure 6 Big Ten Forecasted MOV

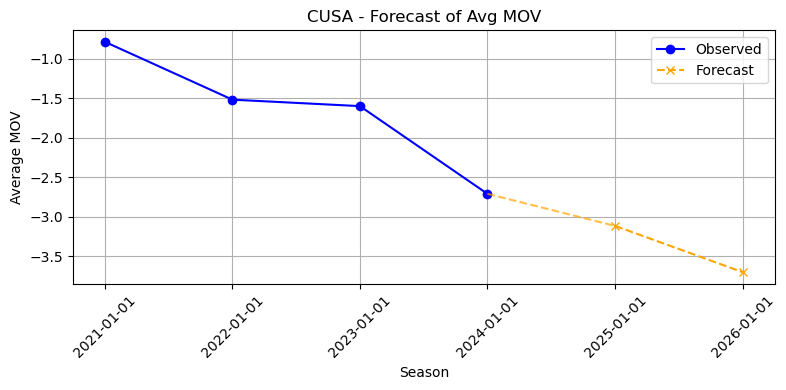


Figure 7 CUSA Forecasted MOV

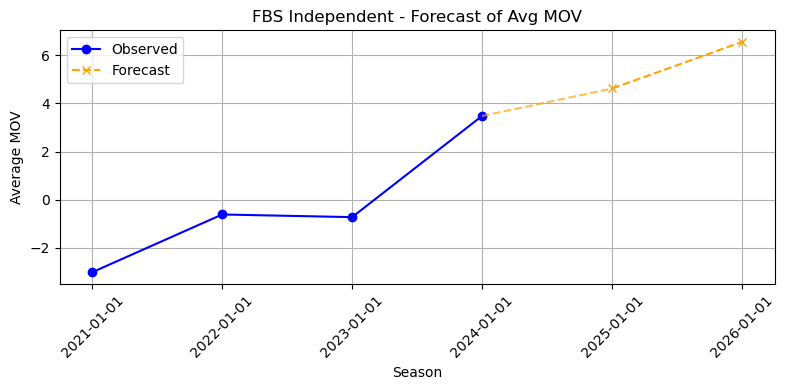


Figure 8 FBS Independent Forecasted MOV

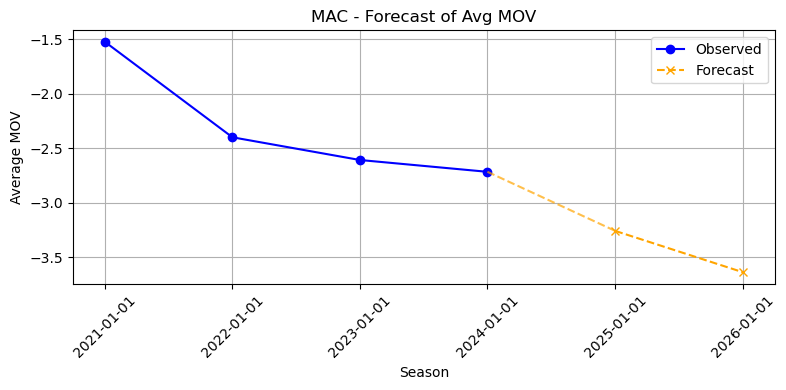


Figure 9 MAC Forecasted MOV

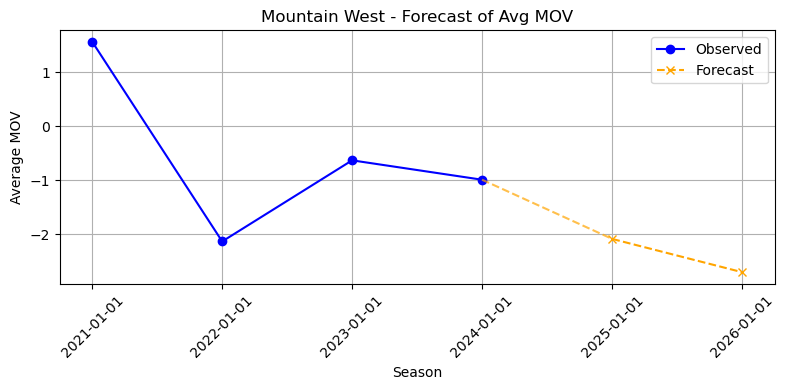


Figure 10 Mountain West Forecasted MOV

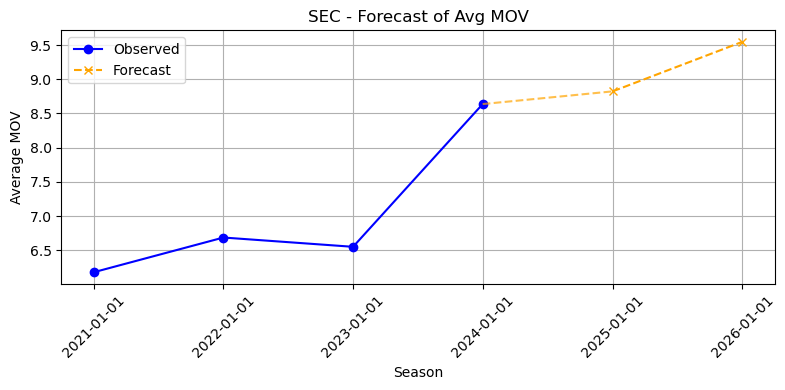


Figure 11 SEC Forecasted MOV

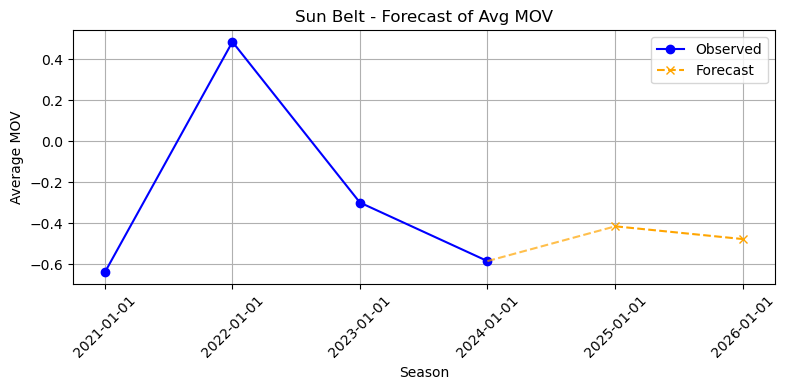


Figure 12 Sun Belt Forecasted MOV

**Your analysis should include, but is not limited to:**

* + Basic descriptive statistics and visualizations (univariate and bivariate)
  + Hypothesis Tests
  + Machine Learning – must have at least 1 machine learning analysis (regression or classification) With each, test 3-4 different models with that analysis (regression – linear, ridge, lasso, etc.; classification – logistic regression, decision tree, knn, svm, etc.) and report on R^2; accuracy/F1 score
  + Text analytics or time series forecasting (if you cannot complete with your data, you’ll do an additional machine learning analysis)
  + At least 4 visualizations

**4. Conclusion**

Briefly present your overall conclusions, limitations, and suggestions for future work (0.5 - 1 page)

In this project we analyzed blah blah blah. From the analysis questions presented in our proposal, we found the following conclusions.

1. *Which statistical categories have the greatest correlation with wins among teams?*

Based on the Pearson correlation analysis, we can determine the statistical category with the strongest correlation to win percentage among college football teams is Margin of Victory, with a correlation coefficient of 0.8975, indicating a strong positive linear relationship. This means that teams that win by larger margins tend to have significantly higher win percentages. Total Offense Yards Per Game also shows a moderate positive correlation (0.5220), suggesting that gaining more yards on offense contributes to winning more games. On the other hand, Total Defense Yards Per Game has a moderate negative correlation (-0.6013), indicating that teams allowing fewer yards on defense are more likely to succeed. These findings display the importance of both offensive strength and defensive efficiency, but MOV proves to be the most powerful indicator of team performance.

1. *Do teams with balanced offensive and defensive performance achieve higher margins of victory than teams with one-sided strengths?*

Balanced teams that exceed the total offense season’s mean and their total defense recognize fewer years than the season’s mean. Everything else is one-sided, which seems to have a massive impact on the on-field performance, resulting in promising outcomes. Balance team average of 11.94 MOV per game compared to -1.95 MOV for one-sided teams. Giving us a 14-point difference, whether or not a team would excel on both offense and defense, also means winning comfortably.

1. *Do any teams have a high count of games with a statistically average or below-average offense/defense?*

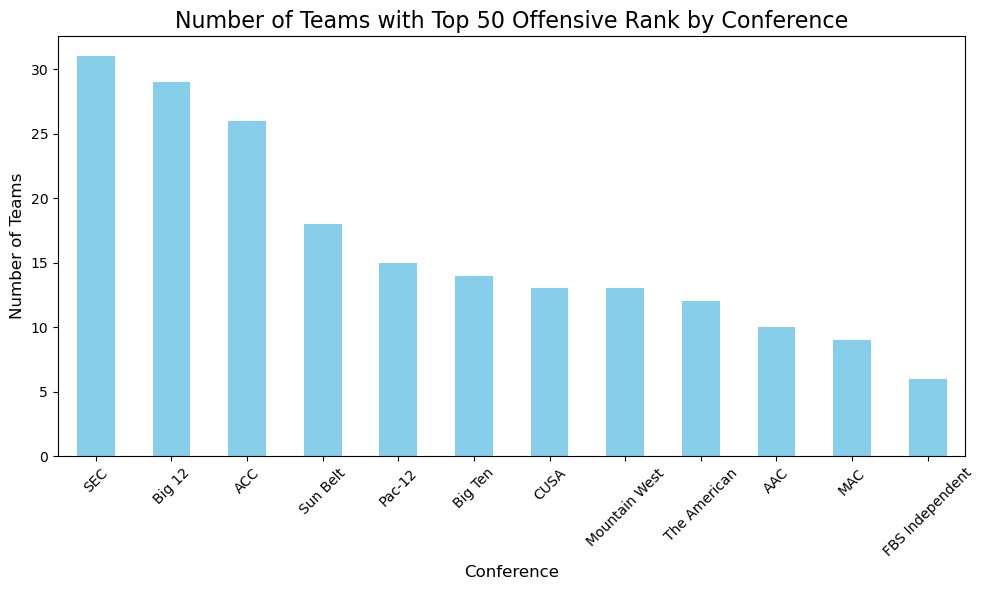
To identify the teams that struggled to reach the least average performance on a side of the ball, we decided to flag every game where teams' total offense YPG was at or below the overall mean offense, and similarly for total defense YPG, which was at or below the overall defense. Across all 4 seasons, there were about 54 teams that gave below-average performance more than 3 times through those seasons. Teams like North Texas, Buffalo, U Mass, and Ball State are a few amongst those 54 teams. Each team failed at least 4 times to exceed the average offensive or defensive performance. These teams have struggled in at least one phase of the game and have stayed in the bottom standings during those seasons.

1. *Are there any outliers of teams with statistically below-average offense/defense that obtained more wins than ones with above-average numbers?*

When we started comparing each team's total offense and defense against their season means, Marshall stood out as a solo outlier. Although they performed below average in their offense with only 382.8 YPG and defense of 378 YPG, they still managed to win 10 times, which showed they did better than the average win throughout the season. This shows that they can outperform major players with factors that are not comparable to the metrics of YPG and win games against above-average offense and defense teams.

1. *Are there certain conferences that present higher overall statistics. For example, does one conference have multiple teams with highly ranked offenses compared to all other conferences?*

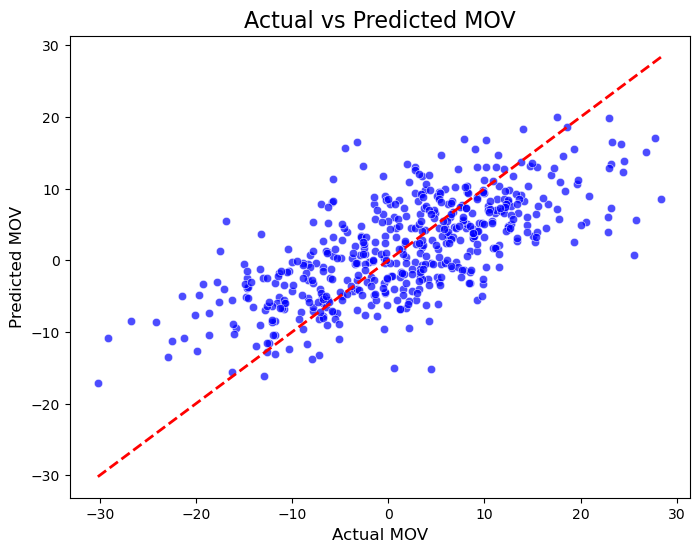
The top 50 offensive teams are heavily skewed towards the Power-Five conference, with the SEC leading the pack with 31 teams in the top 50, with the Big 10 closely behind with 29 teams, and the ACC with 26. Followed by the Sun Belt with 18, the Pac-12 with 15, then the Big Ten with 14, trailed by CUSA with 13, tied with Mountain West, and The American at 12, ACC with 10, MAC with nine, and FBS Independents with six teams.

Figure 13 Number of Teams in top 50 Offensive rank

1. *Can we predict a team’s potential MOV based on a conference’s average performance throughout Seasons 2021 through 2024?*

MOV = -44.70 x (Total Offense YPG) - 7.10 x (Rush YPG) - 7.15 x (Pass YPG)

Utilizing each team's conference season average YPG for offense and defense as predictors in linear regression. About 65 % of the variation in MOVfor per game, with a MAE of 1.1 and RMSE of 1.4. Model Coefficients shows that stronger offensive conferences boost the expected MOV. Better Defense can decrease the results based on how they perform in each game. For better and accurate prediction, having sufficient information on recruiting rankings and the Coach they hire, or what kind of plays coaches have planned, can close the gap for variation in the prediction.

Figure 14 Actual MOV vs Predicted MOV

1. <https://stats.ncaa.org/rankings?sport_code=MFB&division=11> [↑](#footnote-ref-1)
2. <https://betiq.teamrankings.com/college-football/betting-trends/win-loss-records/> [↑](#footnote-ref-2)