ROBERT H. SMITH
SCHOOL OF BUSINESS

All-Star Analytics -
NCAA Basketball Data Analysis

Data Processing and Analysis in Python

Analysis and Report Completed By:
Anna Farley
Wendy Guan
John Kinkead
Joanna Kwok
Pravah Malunjkar
Shane O’Brien

Introduction

Sport analysis is becoming increasingly prevalent in the modern professional sports world and can be attributed to the unprecedented access to big data and the capacity to employ sophisticated analytical methods to derive meaningful insights related to performance. This tendency is especially noticeable in the basketball sector, where data-driven insights are well-suited due to the dynamic character of the game and its complex strategies. Basketball teams are using sport analysis to estimate player performance, improve strategy, and ultimately predict game outcomes as they become more aware of its transformative potential. This exploration aims to utilize data analytics to uncover patterns and derive actionable insights from publicly accessible data, addressing key business questions related to performance improvement:

- Which leagues are performing best overall?
- How are specific basketball stats correlated with wins?
- How have performance factors changed pre and post-COVID?
- What is Maryland Basketball's impact within the Power 5?
- Modeling the winning percentage of Maryland Basketball.

Background

The landscape of basketball analytics has evolved significantly since gaining popularity in the 1990s. Fortune Business Insights projected the global sports analytics market to exhibit a CAGR of 28.7% from 2023 to 2030. Today, basketball analytics is more accessible than ever, offering live play-by-play data and real-time player and team stats. This is crucial for teams during the season, where they may face three completely different opponents in a week.

With NCAA Division I composed of over 350 teams playing around 30 games in a season, relying solely on video analysis presents a challenge. Our project aims to utilize a decade of NCAA Division I data to analyze key statistics that influence a team's winning percentage. First, through exploratory data analysis to uncover trends in various basketball statistics and provide insights, and then using those insights to create predictive models and conduct analysis.

Research Methods

To effectively understand, visualize, and make suggestions on these college basketball statistics, we have created a plan. This provides a general flow of information and ideas that work on its previous step to provide the most accurate and meaningful findings:

1. General Basketball and Variable Understanding
2. Data Cleaning and Interpretation
3. Exploratory Data Analysis and Scope Definition
4. Predictive Modeling and Analysis
5. Prediction, Interpretation, and Outtakes

Deliverables

The professional basketball sports world is witnessing remarkable growth, establishing itself as a crucial domain for both game competition and fan engagement. This thriving sector is a rich source of data that offers chances to identify trends and insights that can guide strategic choices. Our ultimate project objective involves crafting a comprehensive presentation and analysis report tailored for key stakeholders and enthusiasts in the professional basketball industry, including players, sports organizations, sponsors, and fans. This deliverable hopes to give teams, players, and stakeholders useful information by utilizing data-driven insights and trends, which will help them spot strategic and profitable opportunities in the sports world and make informed decisions. Our recommendations will be based on a combination of careful examination of relevant datasets and extensive research from various online resources.

Dataset

This dataset contains information on Division I college basketball seasons from 2013 to 2023. It consists of the following variables:

RK (Only in cbb20): The ranking of the team at the end of the regular season	TEAM: The Division I college basketball school DRB: Offensive Rebound Rate Allowed
ORB: Offensive Rebound Rate	DRB: Offensive Rebound Rate Allowed
FTR : Free Throw Rate	FTRD: Free Throw Rate Allowed
G: Number of games played	W: Number of games won
2P_O: Two-Point Shooting Percentage	2P_D: Two-Point Shooting Percentage Allowed
3P_O: Three-Point Shooting Percentage	3P_D: Three-Point Shooting Percentage Allowed
ADJOE: Adjusted Offensive Efficiency (points scored per 100 possessions)	ADJDE: Adjusted Defensive Efficiency (points allowed per 100 possessions)
WAB: Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)	ADJ_T: Adjusted Tempo (An estimate of the tempo (possessions per 40 minutes) a team would have against the team that wants to play at an average Division I tempo)
EFG_O: Effective Field Goal Percentage Shot	EFG_D: Effective Field Goal Percentage Allowed
WAB: Wins Above Bubble (The bubble refers to the cut off between making the NCAA March Madness Tournament and not making it)	POSTSEASON: Round where the given team was eliminated or where their season ended
TOR: Turnover Percentage Allowed (Turnover Rate)	TORD: Turnover Percentage Committed (Steal Rate)

SEED: Seed in the NCAA March Madness Tournament	YEAR: Season
CONF: The Athletic Conference in which the school participates in	PRE-COVID: Whether or not the data came from a year before the coronavirus

Data Cleaning

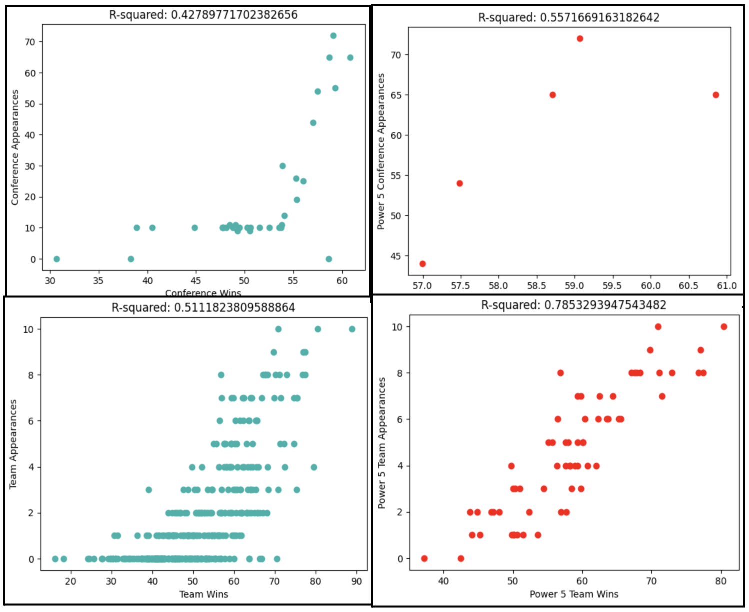
The data cleaning process has been split into four parts: interpretation, consolidation, deletion, and addition. With our group having varying knowledge of basketball and some of the specific statistics that they recorded, it was important for each of us to understand what each attribute in the dataset meant. The second part was consolidation. The data itself was generally clean, but was split by year; to accurately model year over year, we consolidated each year of data into a single master file that we all used when performing our analysis and further manipulation to the dataset.

Our data consolidation choices prompted us to add a YEAR column to indicate the season that the data is from. Additionally, we added a PRE_COVID column to indicate whether the season of data was before the coronavirus had impacted their season, or post to reflect our intentions to analyze performance factors categorized this way as mentioned in the introduction. Importantly, the 2020 season did not have a March Madness tournament, and therefore columns related to such (WAB, POSTSEASON, and SEED) were omitted.

Data Limitations and Considerations

When conducting our exploratory data analysis and modeling, we consistently look at the wins and win percentage of teams and whole conferences as these are likely to be the variables with the highest interest for those looking to spot strategic and profitable opportunities in the sports world and make informed decisions. Win percentage is calculated by dividing the total number of games that the team won in a given year by the total number of games that they played in a given year.

However, the interpretation of win percentage and general wins is more nuanced and complex when accounting for the variations in difficulty among conferences and the number of games played, respectively. Some more competitive conferences, such as those in the Power 5, produce a significantly larger number of teams that appear in the March Madness tournament. This tournament appearance, seeding, and final placement are the best gauge of national ranking and success, rather than only looking at the performance of teams within their conference. The scatter plots and accompanying R-squared values help better visualize the nuance, showing the win percentage by the number of March Madness appearances both by conference and individual teams. You would assume that teams with a higher win percentage have more appearances in the tournament, but this is simply not the case. Looking at all conferences, we see an R-squared value of 0.428 and an R-squared value of 0.511 among teams. However, when comparing conferences and individual teams that are a part of the Power 5, we see a higher correlation with an R-squared value of 0.557 for conferences and 0.785 for individual teams.



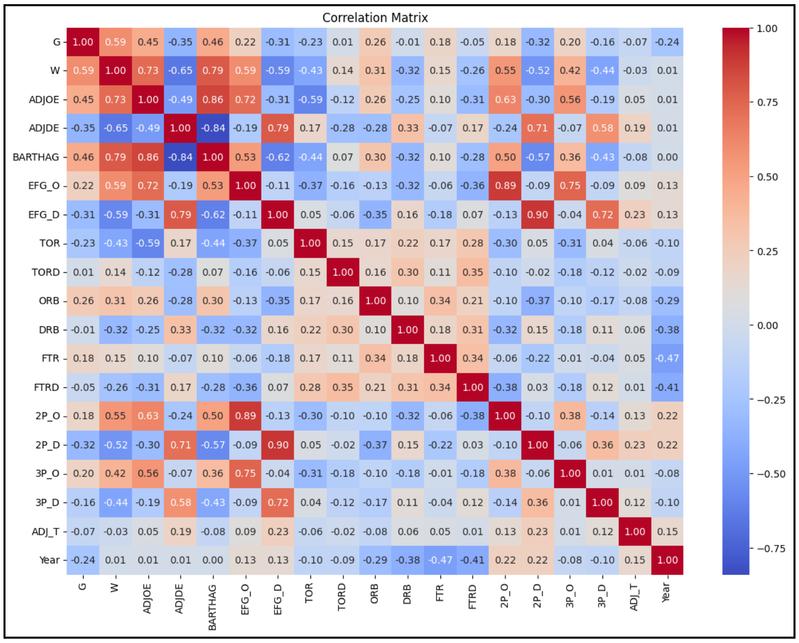
This means that within the Power 5, both by conference and by team, that win percentage better explains March Madness appearances compared to looking at all conferences. Keeping this information in mind, while specific analysis and modeling may seem close in results for all conferences and Power 5, we have determined that looking at Power 5 solely can provide more telling information.

Exploratory Data Analysis

In this exploratory data analysis (EDA) project, we delved into a comprehensive dataset capturing Division I college basketball seasons from 2013 to 2023. Our focus was on understanding team performance metrics, particularly in the context of Power 5 conferences (ACC, SEC, B10, B12, P12). The dataset encompassed key variables such as Adjusted Offensive Efficiency (ADJOE), Adjusted Defensive Efficiency (ADJDE), Win Percentage (WIN_PERCENTAGE), and more.

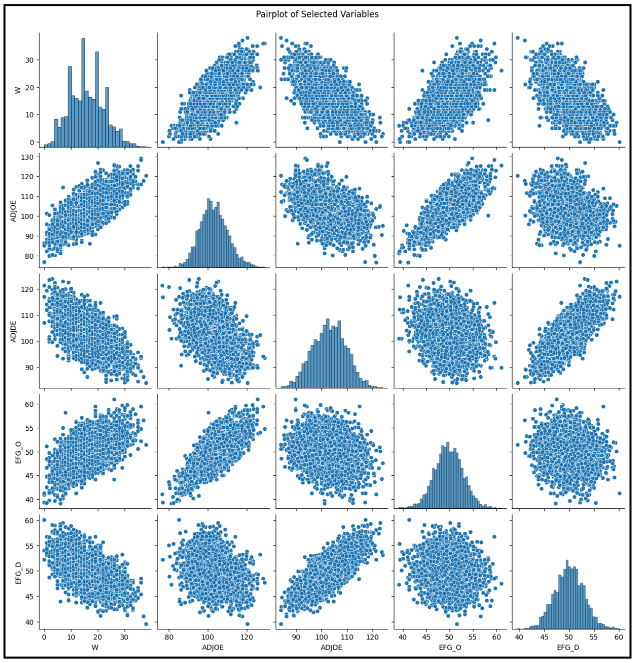
Correlation Heatmap:

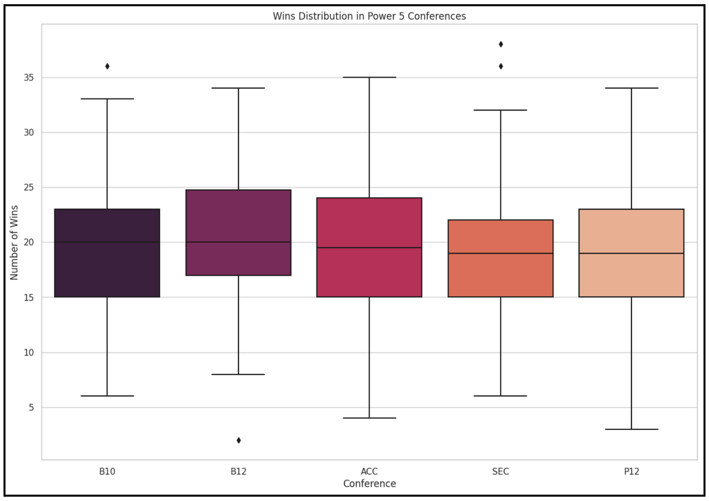
A correlation heatmap showed which variables had a significant impact on the number of wins and offered a quantitative assessment of the relationships between the variables. This insightful analysis of the interactions between various metrics advances our comprehensive knowledge of the factors that influence college basketball success.



Pair Plot Exploration:

To capture comprehensive insights, we constructed a pair plot encompassing all variables in the dataset, with a specific focus on their relationships with the 'Wins' variable. This visual representation allowed for the identification of potential patterns, correlations, and outliers, contributing to a holistic understanding of the dataset's dynamics.



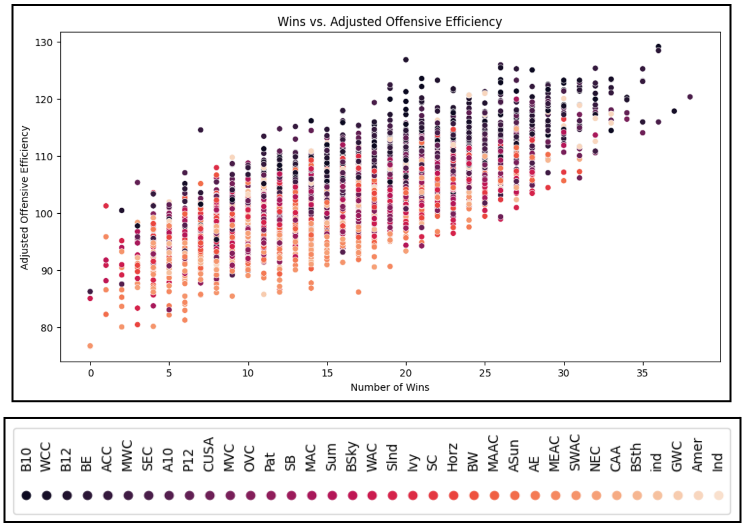


Distribution of Wins in Power 5 Conferences:

A boxplot can be used to see how the Power 5 conferences' wins are distributed. There is some variability among the Big Ten (B10) and Big Twelve (B12) teams, but there is also a more concentrated distribution in the Atlantic Coast Conference (ACC).

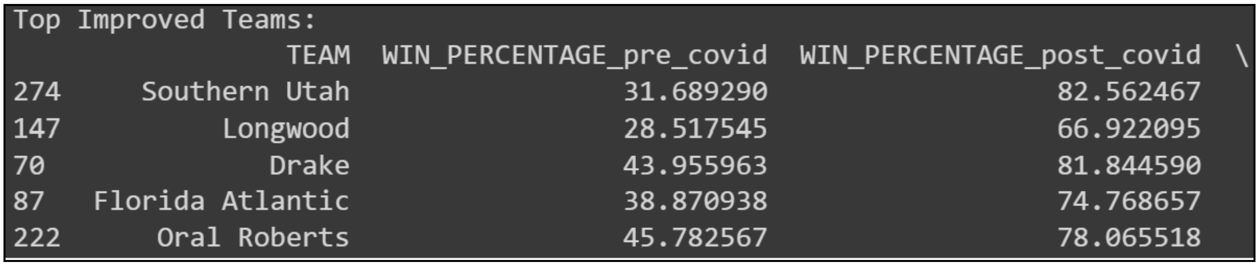
Relationship between Wins and Adjusted Offensive Efficiency:

We can learn more about the relationship between wins and Adjusted Offensive Efficiency by looking at a scatter plot that colors teams according to conferences. With the help of this visualization, we can see patterns in each conference. To improve legibility, the legend has been thoughtfully positioned at the bottom.



Top Teams Showing Improvement or Decline:

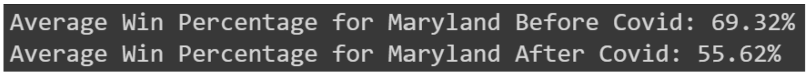
The five teams that improved the most and those whose win percentage significantly decreased both before and after the COVID outbreak have been determined. This analysis has shed light on the evolution of team performance over the given time frame.



```
Top Improved Teams:
                 TEAM  WIN_PERCENTAGE_pre_covid  WIN_PERCENTAGE_post_covid  \
274     Southern Utah                 31.689290                  82.562467
147          Longwood                 28.517545                  66.922095
70              Drake                 43.955963                  81.844590
87   Florida Atlantic                 38.870938                  74.768657
222      Oral Roberts                 45.782567                  78.065518
```

Considering the Impact of Pre-Covid Factors:

Now we analyze the same for Maryland as their performance in the Big 10 matters to us. So we measure their pre and post-Covid success below.



```
Average Win Percentage for Maryland Before Covid: 69.32%
Average Win Percentage for Maryland After Covid: 55.62%
```

This comprehensive EDA provides a nuanced understanding of team dynamics, highlighting conference-wise variations and the impact of external factors such as the COVID-19 pandemic on performance metrics. The visualizations presented offer valuable insights for further investigation and strategic decision-making in the realm of college basketball.

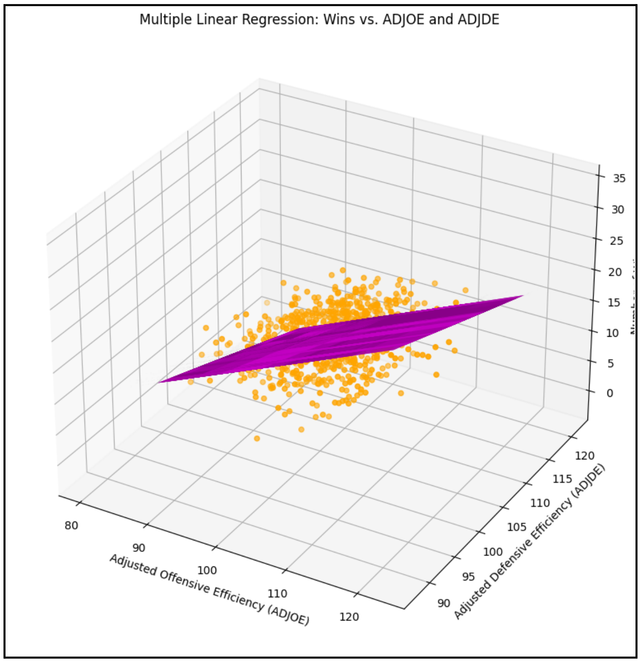
On average, with all other variables remaining constant, the ADJOE, BARHTAG, EFG_O, 2P_O, and 3P_O make the most positive impact on a league's/team's win percentage. With this newfound information in mind, we move towards data modeling to predict win percentages with the above variables as our independent variables.

Data Analysis and Modeling

The problem that we chose to address is the issue of predicting future college basketball performance, measured by win percentage, for all teams, the Power 5 conferences, and eventually, the University of Maryland. To do this, we planned to use our analysis to create a model that could accurately predict this utilizing variables involving offensive and defensive strategy. Before building our model, however, it was necessary to analyze the data to determine the significance and implications of variables.

To begin, we choose a basic multiple regression model based on two variables: Adjusted Offensive Efficiency and Adjusted Defensive Efficiency to predict Wins. Then we calculate the R^2 and Adjusted R^2 values to check if the model is well fit to predict our dependent variable.

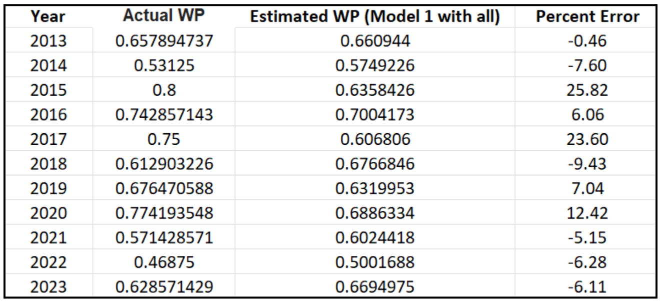
Here, we see that the win percentage sees an upward trend with ADJOE but a downward trend with ADJDE. Now even though this claim is supported by our correlation matrix, the $R^2 = 0.6281$ and Adjusted $R^2 = 0.6272$ values suggest that the model is not that well fit to predict wins and is rather moderate in its accuracy. Hence, we look into building a model that factors in more independent variables to better predict wins.



We then used an ordinary least squares (OLS) model, a type of linear regression that is useful for estimating a dependent variable using the relationship between one or more independent variables, as our base model because of the linear relationship between many of the variables in our dataset, specifically tied with WIN_PERCENTAGE. Utilizing statsmodels.api, we created our first model, regressing WIN_PERCENTAGE on every independent variable in our updated dataset (ADJOE, ADJDE, EFG_O, EFG_D, TOR, TORD, ORB, DRB, FTR, FTRD, ADJ_T). The resulting output displayed an R^2 value of 0.839 and low p-values for almost every variable, indicating both a relatively good fit for the model and high significance for each of the predictors. The Durbin-Watson test also displayed a value of 1.772, which is fairly close to 2.0 and indicates low autocorrelation. The model also stated that there was a large condition number, which could indicate the presence of multicollinearity. We then examined the correlation matrix and created similar models that now omitted some of the highly correlated variables, but after determining that these models were even worse predictors (based on the R² and p-values), we chose to remain with our base model.

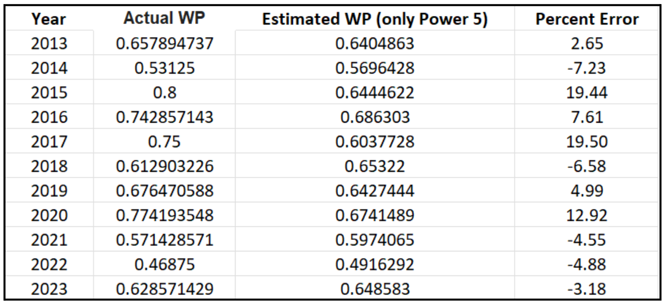
We then chose to estimate a second model using the same equation, solely on the data from teams in Power 5 conferences. Our resulting output was a slightly better fit, based on the R² value of 0.848, the Durbin-Watson test of 1.965, and similar highly significant p-values of 0.000. Because the Power 5 basketball teams may represent a more homogenous or distinct subset of the data compared to all conferences, the resulting increase in model performance may be attributed to the fact that it focuses on a specific group, rather than trying to generalize across all conferences where there may be more outliers, leading to a higher R-squared.

To determine which model was preferred, we calculated the estimated win percentage of the University of Maryland's basketball program using data from previous years and compared them to their respective actual WIN_PERCENTAGE values.



Year	Actual WP	Estimated WP (Model 1 with all)	Percent Error
2013	0.657894737	0.660944	-0.46
2014	0.53125	0.5749226	-7.60
2015	0.8	0.6358426	25.82
2016	0.742857143	0.7004173	6.06
2017	0.75	0.606806	23.60
2018	0.612903226	0.6766846	-9.43
2019	0.676470588	0.6319953	7.04
2020	0.774193548	0.6886334	12.42
2021	0.571428571	0.6024418	-5.15
2022	0.46875	0.5001688	-6.28
2023	0.628571429	0.6694975	-6.11

All Conferences Model



Year	Actual WP	Estimated WP (only Power 5)	Percent Error
2013	0.657894737	0.6404863	2.65
2014	0.53125	0.5696428	-7.23
2015	0.8	0.6444622	19.44
2016	0.742857143	0.686303	7.61
2017	0.75	0.6037728	19.50
2018	0.612903226	0.65322	-6.58
2019	0.676470588	0.6427444	4.99
2020	0.774193548	0.6741489	12.92
2021	0.571428571	0.5974065	-4.55
2022	0.46875	0.4916292	-4.88
2023	0.628571429	0.648583	-3.18

Power 5 Model

Our models were excellent predictors of Maryland's win percentage, with average percent errors of 3.63 and 3.70 for the models in their respective order. While the Power 5 Model was a better fit for the University of Maryland, which can likely be attributed to the fact that it is a part of the Big Ten, other universities may have differing results, especially if they do not play in the Power 5 themselves.

Strategy Recommendation

While the models performed well when it came to predicting the win percentage for Maryland and college basketball as a whole, it is important to note that to predict the win percentage for a specific year using this model, we must have data on the offensive, defensive, and overall success rate of a team for that particular year; e.g. the year must have passed already. If a team wishes to predict their future win percentage, the best course of action would be to use data from the past year and to compare it with other factors such as changes in players (as most college careers last roughly four years), coaches, funding, game strategies, and other outside factors. Although this new model would include an entirely different set of variables, our analysis is still valuable for teams as it provides insight into how specific game strategies can impact a team's record. Furthermore, it can allow universities to look into other programs and their success rate, understand their competition, and determine what they might need to work on to be able to win against another which is extremely valuable in the long run.

References

Fortune Business Insights. (2023, May). Sports Analytics Market. Fortune Business Insights. https://www.fortunebusinessinsights.com/sports-analytics-market-102217

NCAA. (2023, December). Men's College Basketball Standings. NCAA. https://www.ncaa.com/standings/basketball-men/d1

Sundberg, A. (2023, July 7). College basketball dataset. Kaggle. https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset/data?select=cbb.csv

Zhongbo Bai, Xiaomei Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges", *Complexity*, vol. 2021, Article ID 6676297, 11 pages, 2021. https://doi.org/10.1155/2021/6676297