Final Project Report

Sam Andrews, Dom Bouchard, John Logsdon

Introduction

College basketball is a cornerstone of American sports culture, especially during the month of March and ensuing March Madness, drawing massive viewership and interest from fans, analysts, gamblers, media, and universities all around the country. Yet despite its popularity, there remain unanswered questions about how the game itself has evolved and what drives team success. This project will explore descriptive trends and patterns in NCAA Division 1 basketball over the past decade by analyzing an integrated dataset of team-level performance metrics as well as university information. In particular, we will examine how scoring and offensive efficiency have shifted over time, whether a conference or region steers a team's style of play and overall success, and what distinguishes the most successful programs in their approach to the game

To answer these questions, we will combine publicly available analytical statistics, downloaded from Kaggle, with conference and location data via web scraping. Then we will employ Python to clean, merge, visualize, and analyze the resulting dataset. By uncovering persistent trends and stylistic differences, this study aims to equip fans, analysts, and team strategists with data-driven insights into the modern college basketball landscape.

Data

This project uses two primary sources of data: Kaggle's dataset containing statistics surrounding a school's individual seasons' results[1] and NCSA's data about a school's conference, type of institution, and location[2] (City, State).

---

[1] https://www.kaggle.com/datasets/andrewsundberg/college-basketball-dataset

[2] https://www.ncsasports.org/mens-basketball/division-1-colleges

Individual Season Analytics

We downloaded the College Basketball dataset from Kaggle.com, which contains team-level metrics and box-score statistics for all NCAA Division 1 seasons from 2013 through 2024.

To import the data, we used Python to read the CSV into a pandas DataFrame and parse key fields such as Conference, ADJOE, ADJDE, EFG_O, etc. There was not much cleaning involved in importing this dataset, however, we did rename the 'TEAM' column to 'School' for future merging and integrating with supporting datasets.

Division 1 School Information

The recruiting website, NCSA, provides info about individual universities such as the conference they are affiliated with, the type of university, whether it be a private or public school, and the location of the school, notated in City, State format.

The information was contained in one large web page, so we wrote selenium-based web scraping Python code to pull the list of teams from the site. The script launches Chrome, waits for the list of teams to fully load, and then iterates through each entry to extract 'School', 'City and State', and 'Type'. All rows are collected into a list of dictionaries and then converted into a pandas DataFrame. Next, we wrote a block of code to split the 'Location' column, which contained City and State that a school resides in, into separate City and State columns. Lastly, we exported this cleaned pandas DataFrame into a CSV file named 'ncsa_scraped.csv' for easy access and use throughout the project.

Merging Datasets

The two sources – our cleaned team-level metrics (cbb.csv) and the NCSA-scraped list of Division 1 basketball programs (ncsa_scraped.csv) – both include a school field but use slightly different naming conventions (ex., 'Duke' vs. 'Duke University). To reconcile these, we pulled together a list of unique names from each source and then utilized AI to create a mapping

dictionary for each school and ensure the two sources matched. We then applied our name_mapping dictionary in Python using the .replace() syntax on the ncsa_df pandas DataFrame.

With the university names standardized across the two sources of data, we performed an inner join on 'School' to yield only the records present in both datasets. This merge results in a new pandas DataFrame, 'integrated_df', that contains 3,859 rows of data along with 27 columns. Following the integrated dataset, we had minimal cleaning to do. We had no problematic N/A's that needed to be dropped or filled, and we had 22 duplicate records existing. We dropped these duplicated records, leaving us with 3,837 rows of data and 27 columns. At this point, we felt we had done enough cleaning to answer our analysis questions efficiently and exported our pandas DataFrame into a CSV file named 'integrated_dataset.csv'.

Final Data Dictionary

Field	Type	Description
Team	Text	College name
Conference	Text	College conference (E.g., SEC, ACC)
G	Numeric	Number of games played
W	Numeric	Number of wins
ADJOE	Numeric	Adjusted offensive efficiency per 100 possessions
ADJDE	Numeric	Adjusted defensive efficiency per 100 possessions
BARTHAG	Numeric	Chance of beating an average division 1 team
EFG_O	Numeric	Effective field goal percent
EFG_D	Numeric	Effective field goal percent allowed
TOR	Numeric	Turnover rate per possession
TORD	Numeric	Turnover rate forced per possession
ORB	Numeric	Offensive rebound rate per possession
DRB	Numeric	Offensive rebound rate allowed per possession
FTR	Numeric	Free throw rate per possession
FTRD	Numeric	Free throw rate allowed per possession
2P_O	Numeric	Two-point shooting percent
2P_D	Numeric	Two-point shooting percent allowed
3P_O	Numeric	Three-point shooting percent
3P_D	Numeric	Three-point shooting percent allowed
ADJ_T	Numeric	Adjusted tempo (estimate of possessions per 40 minutes)

WAB	Numeric	Wins above bubble (cut off between making the NCAA Tournament and not)
POSTSEASON	Text	Round of the tournament a team was eliminated
SEED	Numeric	Seed in the NCAA Tournament
YEAR	Numeric	Year the season is from
City	Text	Name of city that an unique school resides in
State	Text	Name of state that an unique school resides in
Type	Text	Class of university (Private or Public)

Analysis

Q1 – Offensive Trends Over the Last Decade

We set out to look at how offense has evolved in college basketball over the last decade. We began by diving into year-by-year shooting trends to understand where those efficiency gains are coming from (Figure 1). Between 2013 and 2024, nationwide 2-point percentage inched upward from 47 percent to just over 50 percent. That improvement reflects smarter passing, better shot selection, and even the growing influence of 'positionless' basketball. Meanwhile, 3-point percentage climbed from around 33 percent to 35 percent by 2019, but settled back to about 34 percent more recently. That peak and slight drawback suggest a play on the law of averages, showing that once nearly every team focuses on shooting threes, it is not as efficient.

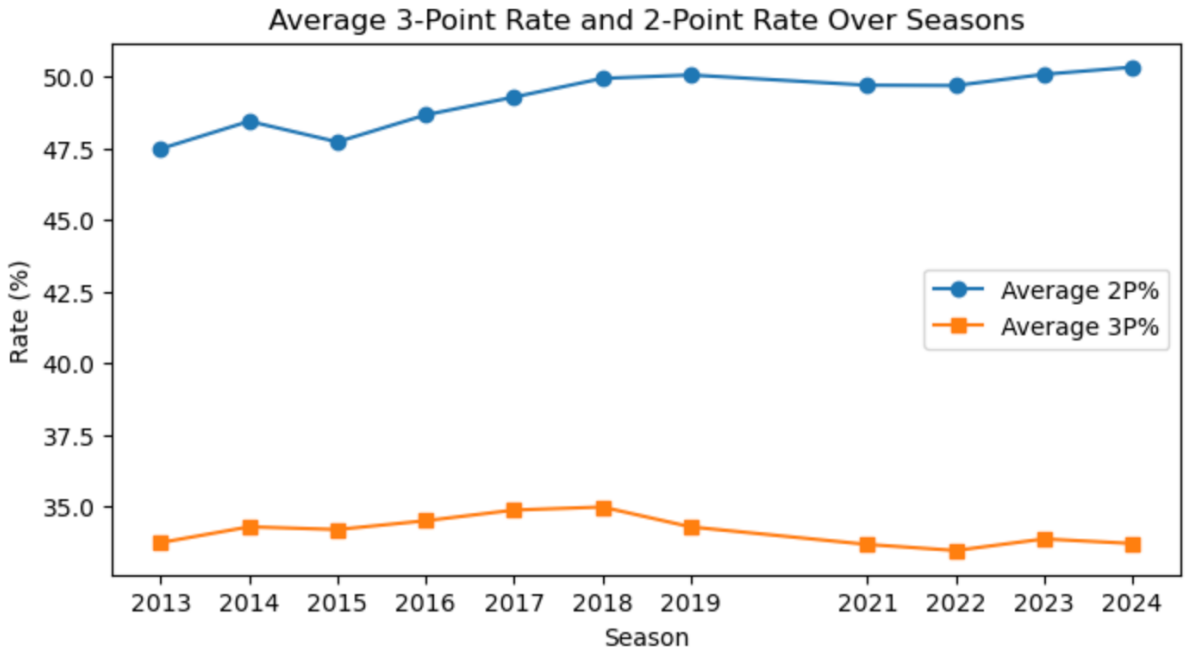


Figure 1

We also built a simple predictive model to pin down which factors drive offensive efficiency the most. We tested four different approaches – linear regression, lasso regression, KNN, and decision trees – and found that linear regression won out with an R^2 of about 0.83. Examining the coefficients of the model shows that effective field goal percentage is by far the biggest booster to ADJOE (1.47 pts per 1% increase), while turnover rate drags efficiency down (1.51 pts per 1% increase). Offensive rebounding rate, free throw rate, and an increase in year also contribute, whereas a faster pace of play comes with a slight reduction in efficiency. In short, shot quality and ball security are the clearest levers teams can utilize to improve their offense.

Lastly, we tracked how shooting efficiencies translated into overall output. Our annual ADJOE curve rose steadily from about 101 in 2013 to 105 in 2025, dipping briefly in 2020-2021 due to COVID, roster shuffling, and slower pace (Figure 2). To gauge future evolution and offensive style, we ran a Holt-Winters ten-year forecast, which predicts average ADJOE climbs into the high 105s by 2035. Even a modest uptick at the nationwide level will translate into several more made shots per game, which in a tight game in March can be the difference maker. Interestingly, our forecast model had projected a 2024 ADJOE of roughly 104, yet the actual value of 105 surpassed expectations, suggesting that offense is on a larger-than-anticipated uptick since coming out of the recent low point in 2020-2021.

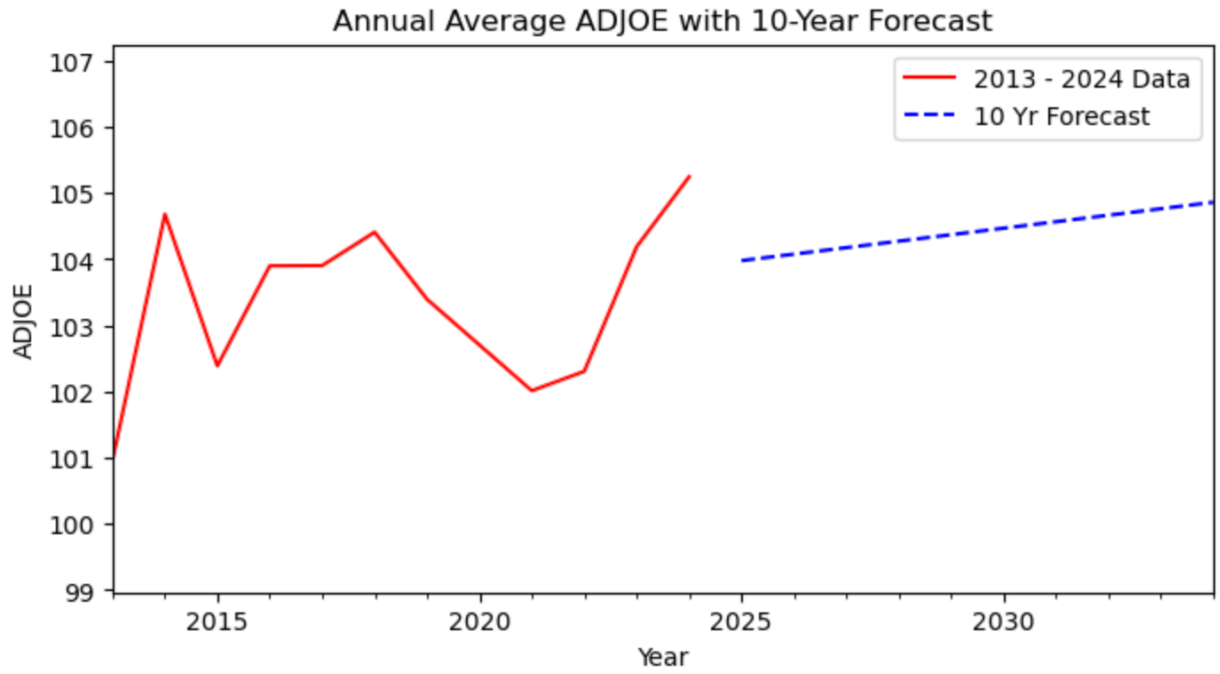


Figure 2

All together, these findings paint a clear picture: almost every program has gotten incrementally more efficient on offense, but there is a percentage of elite teams pulling away from the pack. Those top offenses, utilizing cutting-edge analytics, specialized shooting programs, and faster play, are redefining what modern offense looks like. The message has become clear: investing in smarter shot selection, individual player development, and an upper-tempo pace isn't just a trend, but the future of competing in Division 1 basketball.

Q2 – Effect of a Conference or Region

One question that arose from our dataset was why offensive efficiency varied across conferences. Among all Power Five conferences (including the Big East), the average adjusted offensive efficiency was lowest in the Pac-12 and highest in the Big 12.

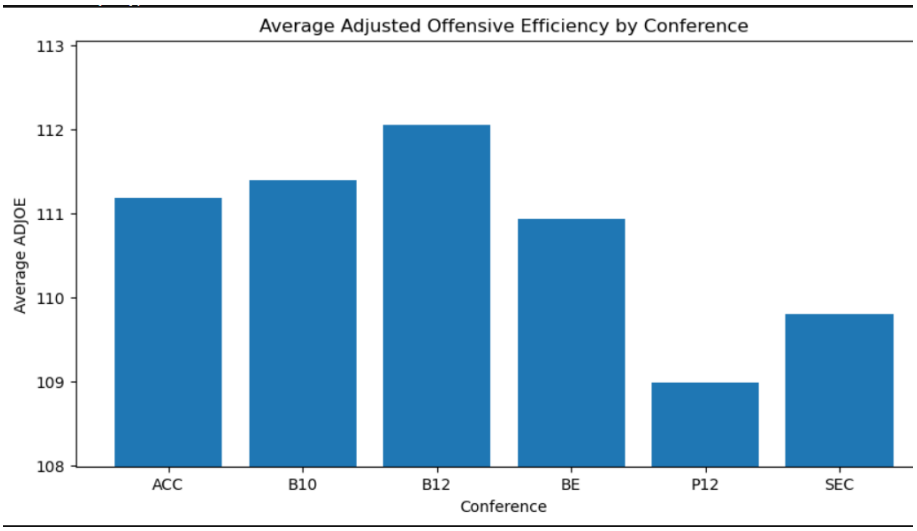


Figure 3

To investigate this, we analyzed conference-level averages for key statistics, including three-point percentage allowed, effective field goal percentage, three-point percentage, free throw rate per possession, turnover rate per possession, and offensive rebounds per possession. While there were some similarities across conferences, the Pac-12 showed a slight deviation in offensive rebounds per possession. However, the difference was not significant enough to explain its lower offensive efficiency. The Big Ten varied more noticeably in areas like free throw rate and offensive rebounds per possession.

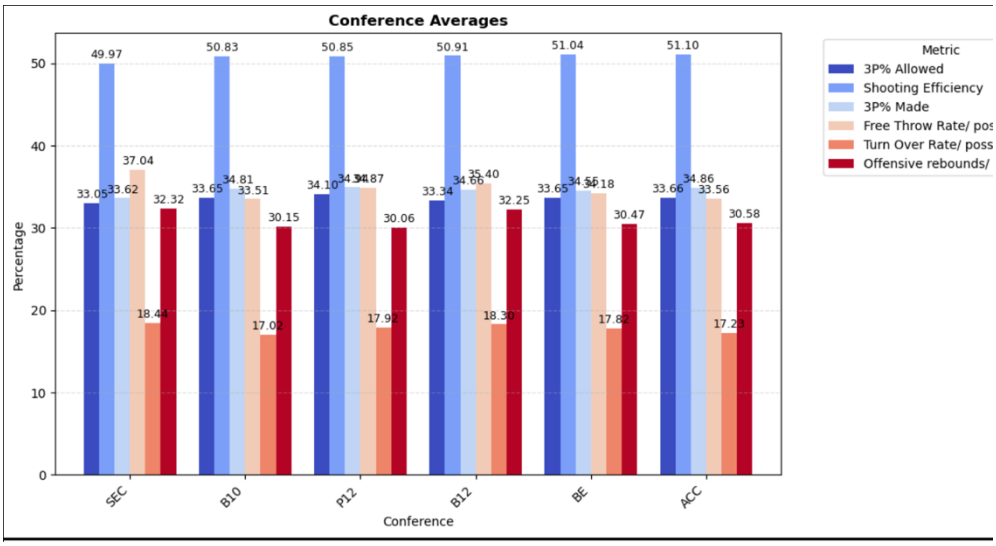


Figure 4

Next, we compared pace of play across conferences. The Pac-12 had one of the faster paces, while the Big Ten had the slowest. However, as shown in the bar chart, there were no substantial differences in pace that correlated with offensive efficiency, suggesting that pace alone does not explain the variation.

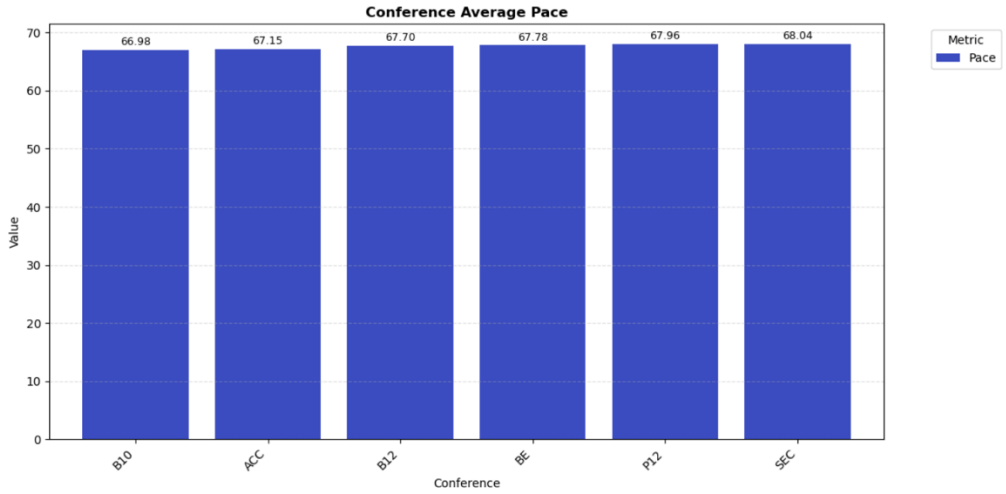


Figure 5

Finally, we analyzed average defensive efficiency. This is where we found a clearer pattern. The Pac-12 had one of the highest defensive efficiencies, which may help explain their lower offensive numbers. In contrast, the Big 12, which had the highest offensive efficiency, had one of the lowest defensive efficiencies. This inverse relationship supports the idea that a conference's defensive strength may impact offensive metrics. We concluded that there is a strong correlation between a conference's average offensive efficiency and its average defensive efficiency.

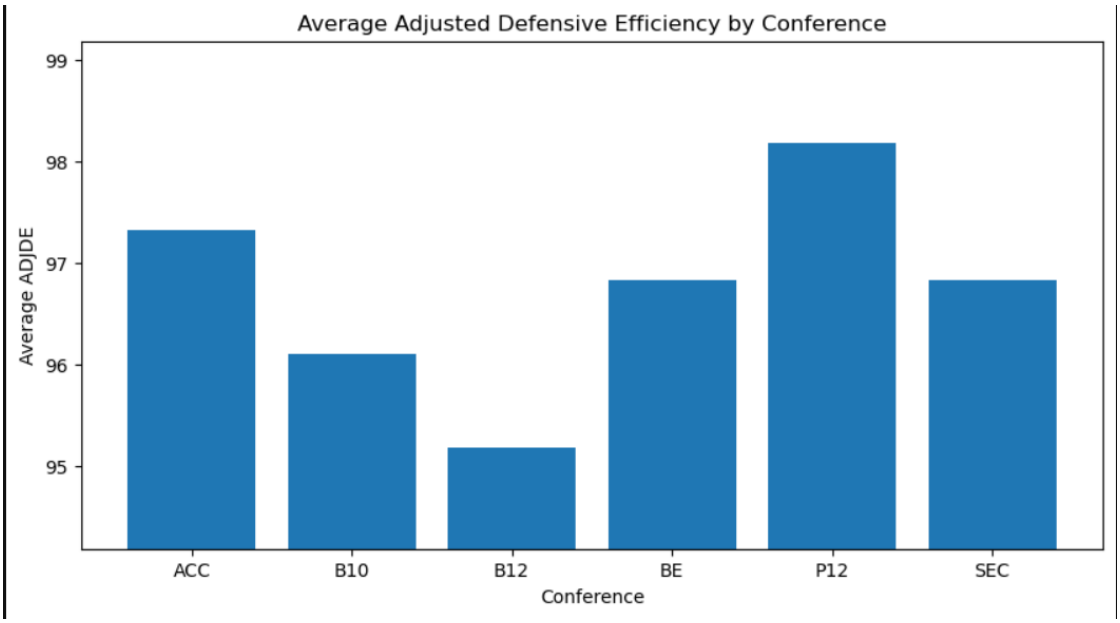


Figure 6

Q3 – Similarities & Differences of the Top Teams

We were searching for what key similarities and differences there were between the top and bottom teams, to see if there was a certain play style that gives a team an advantage. A top team is defined as any team that received a 5 seed or better for the NCAA Tournament. A bottom team is any team worse than a 5 seed, including teams that did not make the NCAA Tournament.

First, we used a scatter plot to show the difference between the top and bottom teams' offensive and defensive adjusted efficiencies. It is better to have higher offensive efficiency and lower defensive efficiency. As expected, from the plot, it is evident that to be a top team, you must consist of having a high adjusted offensive efficiency and a low adjusted defensive efficiency. To make sense of a few outliers, where teams that have a good adjusted offensive efficiency and good defensive efficiency, and are a bottom team, could be due to the conference or their strength of schedule, where these statistics could have been inflated from their opponents. It is also important to note that there are no top teams that are solely carried by their offense or defense. Ultimately, the trend is that to be a top team, you must have a high adjusted offensive efficiency and a low adjusted defensive efficiency.

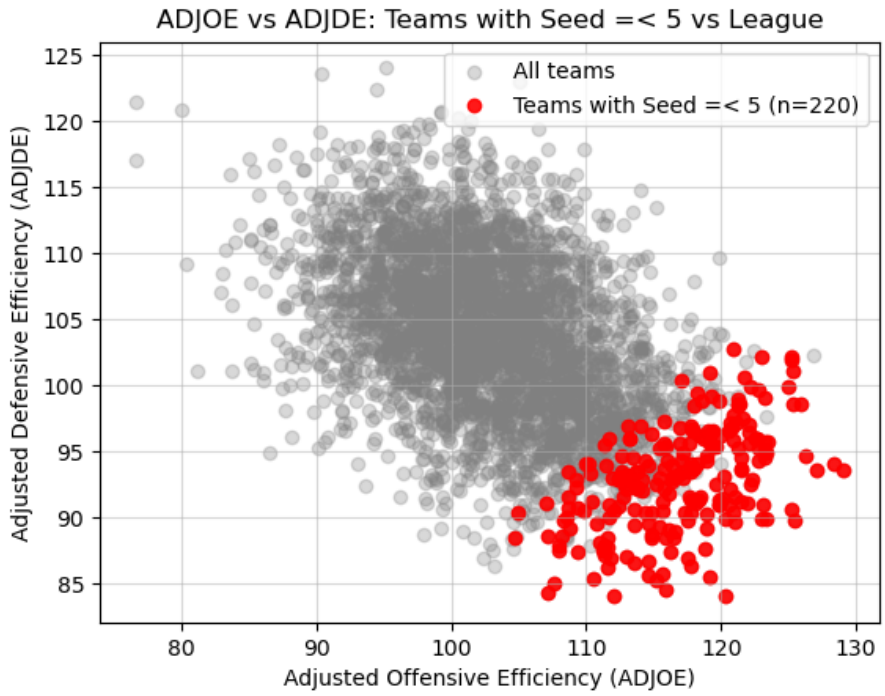


Figure 7

After finding that it takes a good offense and defense to be a top team, we decided to look at the effective field goal percentage (eFG) and effective field goal percentage allowed. The box plots revealed that the top teams had a much better effective field goal percentage, shown by the median eFG being higher than the bottom teams by about 3%. The box plots also show that the top teams allow a lower eFG than the bottom teams by about 3% as well. This further supports that to be a top team, you must have a good offense and defense.

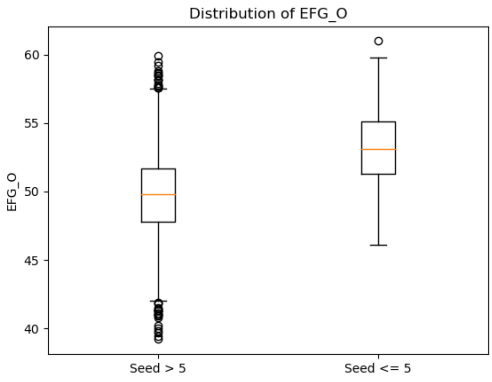


Figure 8

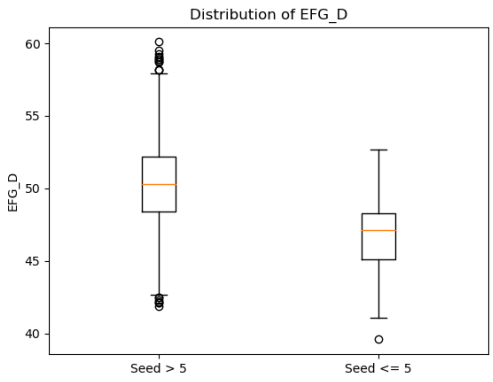


Figure 9

We also used a box plot to look at whether the top and bottom teams' adjusted tempo was similar or different. The box plot reveals that there is no significant difference between their adjusted tempos, with the majority of teams having an adjusted tempo between 60 and 70 possessions per 40 minutes. It is important to note that there are outliers in the bottom teams and no outliers in the top teams, which can be due to more teams being in the bottom team category.

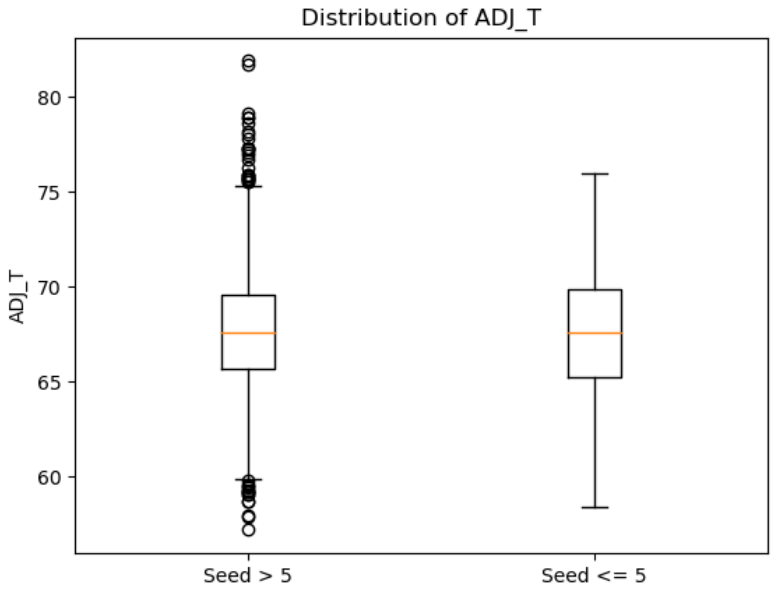


Figure 10

Next, we decided to go further in depth with offensive statistics so we could find any similarities and differences between the top and bottom teams using a grouped bar graph. The graph shows that the free-throw rate per possession is the only similarity that the top and bottom teams have. This comes as a surprise because it's expected that a top team would shoot more free throws, however, it is not the case. They could perhaps shoot a higher percentage from the free-throw line, but that is inconclusive with the data we collected. Differences from these offensive statistics include shooting percentages, offensive rebound rate, and turnover rate. The top teams shoot about a percentage or two better from two and three. Other differences include top teams turning the ball over at a lower rate and offensive rebounding at a higher rate than bottom teams. Overall, top teams perform better at all aspects of offense except free-throw rate.

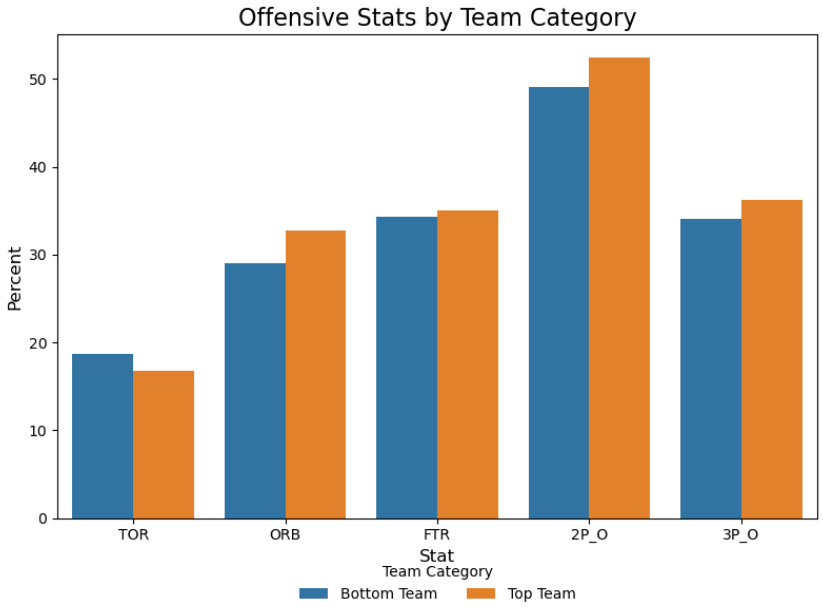


Figure 11

We also used a grouped bar graph to examine the similarities of offensive and defensive metrics for top and bottom teams. It was revealed that the turnover rate forced is the only defensive statistic that is similar for both teams. This is a surprise because it is expected that top teams would force their opponent to turn the ball over more, however, that is not the case. The difference in shooting is that the top teams allow their opponent to shoot worse from two and three, but there is a more significant gap in two-point field goals. Other differences include top teams allowing their opponents to shoot a lower rate of free throws and allowing a lower offensive rebounding rate. Overall, a top team's defense performs better in every aspect except at turning their opponent over, which is at approximately the same rate as bottom teams.

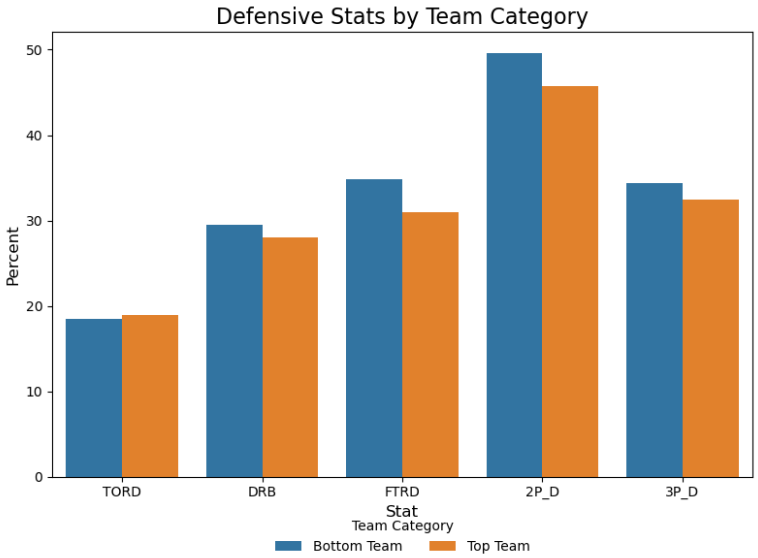
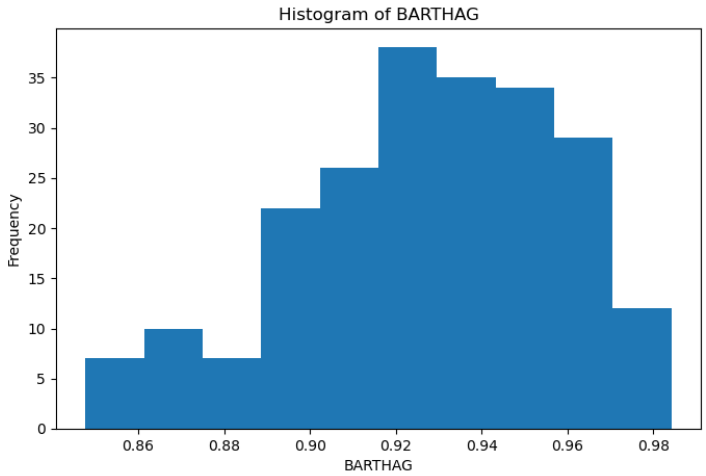


Figure 12

Lastly, we used a histogram to examine how the top teams performed against an average Division 1 opponent. Looking at the histogram, it reveals that the majority of the top teams beat an average team 90%-97% of the time. You would expect that the higher the percentage of the time a team beats an average team, the better seed they are in the NCAA Tournament, and vice versa.



Conclusion

In this project, we analyzed three aspects of modern-day college basketball: offensive evolution and trends, impact of a region or type of school, and the common denominators of the

most successful teams over the last decade. Based on the analysis questions provided in our project proposal, we found the following results:

1. *How have scoring/offensive trends evolved over the last decade?*
    a. Offensive efficiency has climbed steadily over the last decade, from 101 to 105, with 2024's output even outpacing our forecast. Teams embracing analytics, shooting development, and more modern styles of play have pulled ahead and will continue to set the standard in Division 1 basketball.
2. *Does the conference a school is a member of or the region it resides in determine the team's style of play and success?*
    a. While pace and box score stats showed only modest conference differences, a clear inverse pattern emerged when we looked at defense: the Pac-12's strong defensive efficiency coincided with its lower offensive output, and the Big 12's weaker defenses paired with the highest scoring. This highlights a strong negative correlation between conference-level offensive and defensive efficiency.
3. *In what ways do the most successful teams compare and contrast in their style of play?*
    a. Overall, the top teams perform better on both offense and defense than the bottom teams. However, a few things stuck out. The first is that top and bottom teams play at the same adjusted tempo, making the tempo not a factor in determining a team's success. In conclusion, to be a top team, one favors playing balanced and not relying solely on their offense or defense to carry their team's output.

This report is subject to several limitations: our analysis only covers the past ten seasons and relies on aggregate box score stats rather than richer play-by-play or player tracking data, conference realignments over the period introduce noise in cross-year comparisons, and our scraped conference information may contain gaps or inconsistencies. Future work could extend the timeframe by tapping historical sources, layer in recruiting rank and roster experience variables, and leverage machine-learning models on play-by-play analysis. Additionally, incorporating text analytics to evaluate NCAA champions, runners-up, and other deep-run teams on our integrated dataset would reveal how title contenders' offensive and defensive styles evolve as they ascend.