NCAA Basketball Analysis

Zhe Jiang*
Carl Ge†

1 Abstract

This project delves into NCAA basketball data to explore the difference between scoring strategies, the analysis on offensive tactics for UC Davis, and the association between winning rate and team logo colors. Our results show that two-pointers are consistently the most efficient across divisions. For UC Davis, working on two-point and three-point shooting improved performance. Additionally, teams whose primary color is red have a small advantage in winning percentage.

2 Introduction

The National Collegiate Athletic Association (NCAA) serves as the primary governing body for collegiate athletics, organizing competitive sports programs across numerous colleges and universities in the United States. The NCAA's overarching structure is divided into three divisions — Division I, Division II, and Division III — each characterized by varying degrees of competitive intensity.

UC Davis is also a member of the NCAA basketball Division I. In this project, we delve into the trove of data curated by the NCAA Basketball, particularly focusing on the statistical records of men's basketball in the current season. The aim is to dissect and analyze various attributes across all teams and divisions, extracting insights into strategic gameplay and performance trends. Our analysis will pivot around the following core questions:

1. **Scoring Strategy Efficiency**: Which scoring strategy, 3-point shots, 2-point shots, or free throws, is statistically superior on average?
2. **Offensive Tactics for UC Davis**: Considering its divisional peers, which shooting practice should UC Davis Men's Basketball team prioritize to enhance their offensive prowess?
3. **Team Colors and Winning Rates**: How does the frequency of team colors associate with winning rates across different conferences?

*SID: 921067148
†SID: 921235886

3 Shooting Strategy Efficiency

Scoring in basketball is composed of three aspects: free throws, 2-point shots, and 3-point shots. A free throw, which is awarded following a foul by the opposing team, contributes one point to the team's score. A 2-point shot, including mid-range jumpers and layups, counts for two points. A 3-point shot, made from beyond the 3-point line, is worth three points.

Data Acquisition

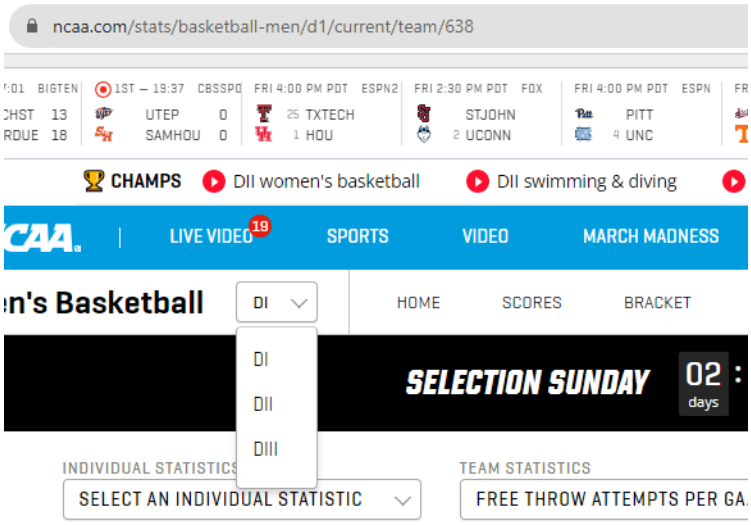
To evaluate the efficiency of various shooting strategies, we considered team data rather than individual player statistics since team data is consistent across all categories, but individuals are not. We identified three statistical categories correlating to the shooting aspects: Three Point Attempted Per Game, Field Goal Percentage,Free Throw Attempted Per Game. Due to the absence of specific 2-point shot data, we subtract 3-point goals from total field goals to determine 2-point statistics since total field goal points are the collection of scores made by two-point and three-point. We utilized Pandas DataFrames to collect and visualize this data, resulting in three separate DataFrames for each type of shot.

We have discovered that there are several pages for each category (Shown on the right side).

We went to the developer page and found that there will be a new link each time after clicking the arrow key(on the rightmost side of the number) on the developer page and the last page, there will not be an arrow key anymore, as well as the new link. Thus, we use a while loop to collect data for a dataframe and add the newly obtained dataframe to a list for each iteration and break the loop while no link is attained from the XPath of the next page class. After the loop, we got a list of dataframes for a category. Then, we concatenate all the elements, which are Dataframes, to a final Dataframe that contains the data across all pages. This procedure was replicated for each shot type.

Expansion to All Divisions

Initially, the data encompassed only Division I teams. To enhance our analysis, we also scraped data for Divisions II and III. See image below:



We thought it was a dropdown menu with a button to click. However, it was not a button and not intractable. When we manually clicked on the division, it jumped to the division homepage instead of the corresponding data page. (When we were on the "FREE THROW ATTEMPTS PER GAME" page, we expected it to jump to the "FREE THROW ATTEMPTS PER GAME" page for division II after clicking on DII from the dropdown menu.) Later, we found a more straightforward way to get to the page by altering the URL from "d1" to "d2" or "d3" (in the middle of the link), allowing access to the respective divisions' data. Thus, for each category, they concatenate the divisions to previous Dataframes.

Efficiency Calculation

To account for the different point values of the three shot types, we devised a formula to measure efficiency: Points Per Shot (PPS). PPS is calculated as follows:

$$PPS = \frac{\text{Total points from a shot type}}{\text{Total shot attempts}}$$

For example, for 2-point shots:

$$PPS_2 = \frac{2 \times \text{2-point goals made}}{\text{2-point attempts}}$$

After calculate the PPS for the 3 category and merge them together, we get the following Dataframe (shown the first 10 rows):

	Team	PPST	PPS2	PPS3	WinRate
0	Winthrop	0.702103	1.088180	1.024390	0.531
1	High Point	0.778802	1.081967	1.042998	0.758
2	Tulane	0.731857	1.088210	1.070444	0.452
3	Alabama A&M	0.719075	0.931715	0.854508	0.333
4	Arkansas	0.755530	1.036921	0.955200	0.485
5	Florida	0.708892	1.042433	1.018717	0.697
6	Nevada	0.721271	1.060884	1.094527	0.788
7	Purdue	0.720050	1.064039	1.231003	0.906
8	Grand Canyon	0.753846	1.073254	1.037258	0.875
9	SFA	0.706030	1.007015	1.003295	0.545

Then, we decide to compare the efficiency by the average of each PPS to find out on average which shooting strategy can make more point per shot, in other word, more efficiency.

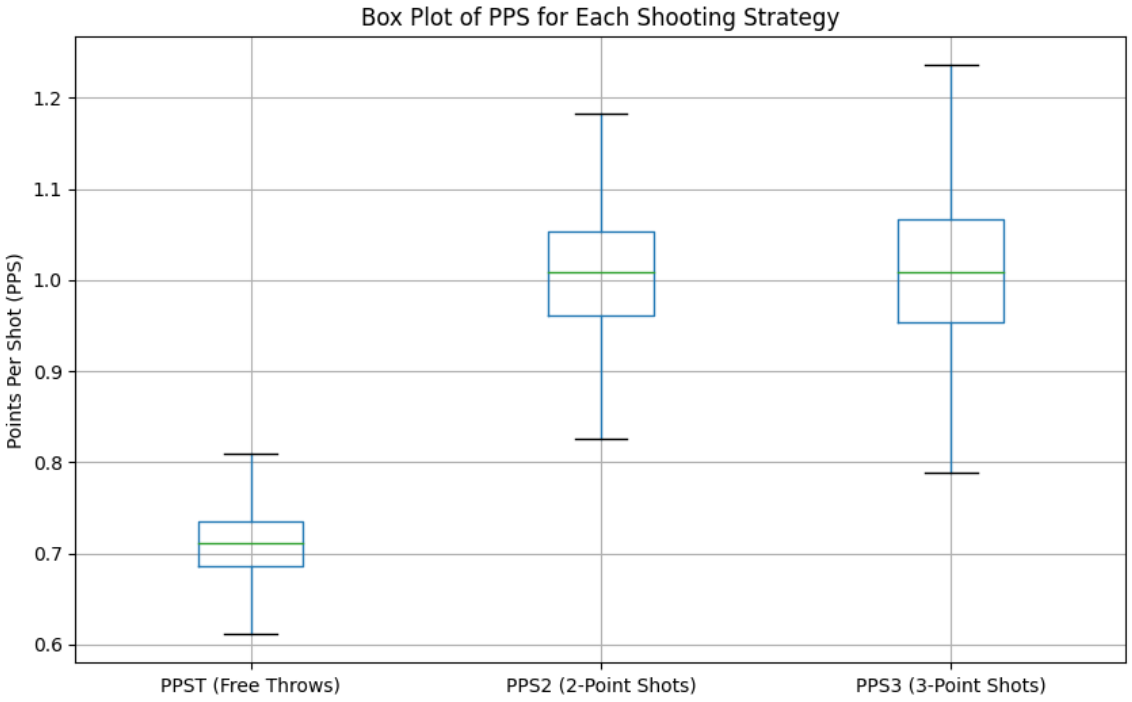
PPST (Free Throws)	PPS2 (2-Point Shots)	PPS3 (3-Point Shots)
0.709866	1.0096	1.008591

Table 1: Average Points Per Shot (PPS) for Each Shooting Strategy Across Divisions

From the table, we can discover that on average, 2-point shot is the most efficient scoring method, closely followed by 3-point shot, with the free throw being the least efficient.

Visualization

Despite the fact that using means to conclude that the 2-point shot is the most efficient scoring method on average, we can approach and analyze efficiency from another aspect. We decided to use a box plot to showcase the PPS since the box plot can cover some other important aspects of the data. Here is the plot created by matplotlib in Python:



The box plot clearly illustrates PPS's distribution ranges and median values for free throws, 2-point shots, and 3-point shots across all divisions. The median line within each box shows that the median efficiency of 2-point shots is marginally higher than that of 3-point shots, confirming our earlier conclusion that 2-point shots are the most efficient on average. However, a high value of the 3rd quartile(Q3) and the value after Q3 for PPS3 are greater than PPS2, suggesting that the top-performing teams in 3-point shooting score more efficiently with 3-pointers than 2-pointers.

4 Offensive Tactics for UC Davis

As a member of UC Davis, I am very interested in figuring out the offensive tactics for UC Davis compared to others in Division I. Currently, UC Davis ranks 112th out of 351 in terms of win rate in Division I—a position that certainly has room for improvement. To gain insight into their offensive performance, we've extracted UC Davis's data from the previous Dataframe and get the following:

PPST (Free Throws)	**PPS2 (2-Point Shots)**	**PPS3 (3-Point Shots)**
0.748212	1.012456	1.008772

Table 2: Average Points Per Shot (PPS) for UC Davis

If we compare the average PPS across divisions from Table 1, we can see that UC Davis is above average. However, there is still a discrepancy when placed in the context of the competitive Division I landscape. Further analysis was conducted by

scraping the team name from Division I using the same procedure in the previous part and using the team name to extract the PPS from the Dataframe created by the last section to get the PPS for the team only in Division I. We get the following:

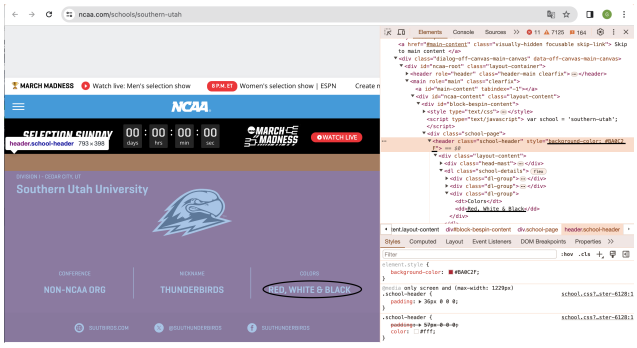
PPST (Free Throws)	PPS2 (2-Point Shots)	PPS3 (3-Point Shots)
0.717509	1.019049	1.017856

Table 3: Average Points Per Shot (PPS) for Division I

The comparison reveals that UC Davis's efficiency in both 2-point and 3-point shots is below the Division I average. This insight suggests a clear directive for the team's training focus. By improving their skills in 2-point and 3-point scoring, UC Davis can potentially boost their offensive efficiency and climb the ranks within their Division. With dedicated practice aimed at these critical areas, we hope UC Davis will give a better score.

5 Team Colors and Winning Rates

While investigating the factors that influence Shooting Strategy Efficiency, we found that many school team logos are composed of various colors and patterns:



In many cases, a school team's logo represents the school's tradition or spirit. In the NCAA arena, the team logo of each school can inspire the players to strive to win the game. In this regard, we are strongly interested in the relationship between team logos with different theme colors and the team's winning rate.

Webdriver

First, we try to set up the Webdriver environment on Google Colab (By reference URL Geeksforgeeks). Although the establishment is successful, due to the deficiencies of Colab, the operations of the Webdriver are not displayed on the screen (Use line 5: "–headless" to enable Webdriver to control web pages normally on Colab.)

```python
def web_driver():
    options = webdriver.ChromeOptions()
    options.add_argument("--verbose")
    options.add_argument('--no-sandbox')
    options.add_argument('--headless')
    options.add_argument('--disable-gpu')
    options.add_argument("--window-size=1920, 1200")
    options.add_argument('--disable-dev-shm-usage')
    driver = webdriver.Chrome(options=options)
    return driver
```

Then, we found that operating web pages without views would cause a lot of trouble when scraping information. After a group discussion, we decided to run Webdriver locally on Visual Studio Code.

In our efforts to automate the NCAA school search process, our process is blocked by an obstacle when trying to interact with the input text box on the official NCAA website. Initially, we used XPath to locate the element, but unfortunately we encountered a "non-interactive" error:

```
---------------------------------------------------------------------------
ElementNotInteractableException                Traceback (most recent call last)
Cell In[146], line 4
      2 driver.get('https://www.ncaa.com')
      3 input_element = driver.find_element(By.CLASS_NAME, 'school-finder')
----> 4 input_element.send_keys('test')
```

We then turned to CSS selectors, hoping to find a solution, but the same problem emerged again. We also tested the capability of the input text box by operating on familiar platforms such as Google and Yahoo search boxes to confirm that the NCAA's text box was disabled to interact. To overcome these challenges, we decided to adjust our strategy. We changed our current approach from "www.ncaa.com" to entering specific keywords through "google.com" to avoid that input textbox.

Finally, we successfully used google.com to automatically scrape the data of colors of each school team's logo and background colors, whereas the price was nearly 6 hours. The scraped data will be charted for easy visualization and analysis.

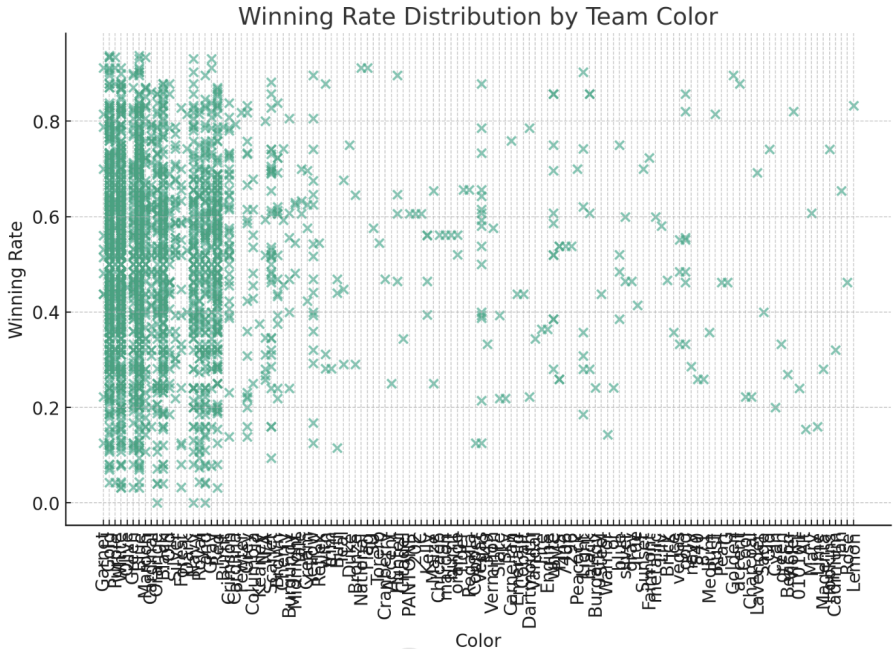
Visualization

The color of the school team logo is based on the school introduction on the website. It can have several colors (color names). The background color of the team logo exists in the form of RGB parameters. To be precise, there is only one color for the background color. Analyzing the relationship between 17 million colors and the winning rate is too complex. The number of samples is too sparse (1049 colleges in total), we decided to determine the primary color of the background color through the maximum number in RGB (If the RGB parameters are R: 255, G: 240, B: 20, then the primary color of this background is red).

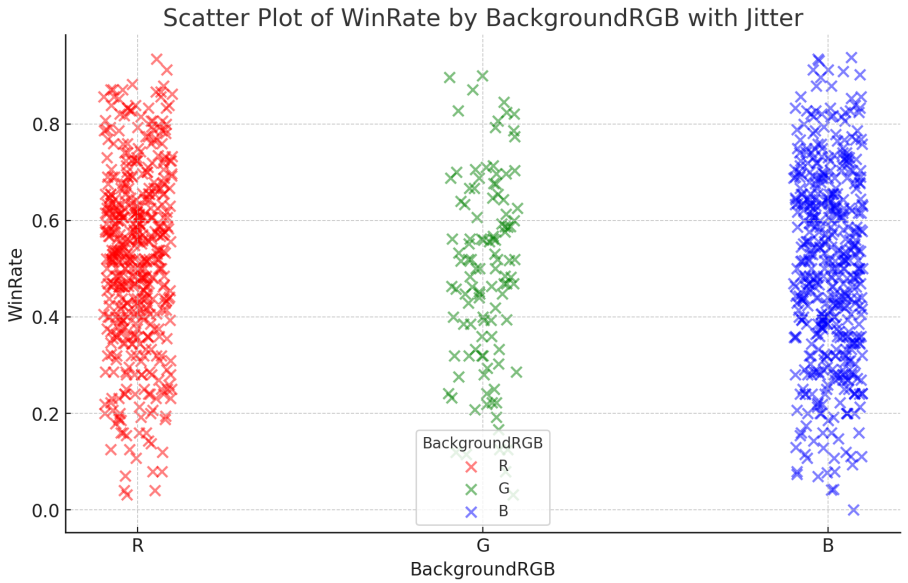
Num of Red	Num of Green	Num of Blue	N/A
465	110	457	17

Table 4: Number of Each Theme Color

We begin our analysis by creating a scatter plot that compares various colors extracted from all team logos to their respective winning percentages:

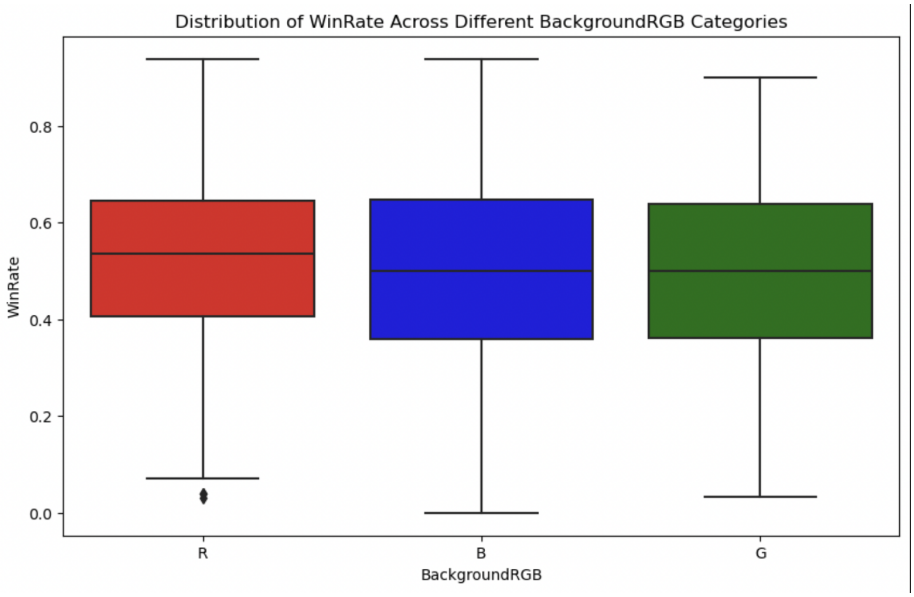


This initial approach did not form any clear patterns or trends suggesting that specific colors were more "symbolic" of victory. Instead, it only displayed the density and distribution of color usage among teams. Recognizing the limitations of this perspective, we turned to study the impact of varsity theme colors based on the basic RGB color model (red, green, or blue) based on their performance. This shift breaks our study down into the three primary colors, potentially providing clearer opinions on the correlation between team color and win rate. We continue to build a scatter plot between the school team's theme color and Winning Rate:



However, the number of team logos that have the main background color green is much smaller than red and blue, which makes the scatter plot still focus on density.

After careful consideration, we planned to draw a boxplot as our method of choice to showcase the distribution of theme colors(RGB) relative to winning rates:



This decision allowed us to effectively communicate the distribution of each color, avoiding problems caused by unequal amounts between the red, green, and blue categories. By taking this approach, we aim to provide a more reasonable comparison of how these primary theme colors relate to team winning rates.

The boxplot data indicates that teams with red as their primary color tend to cluster more closely around a higher median winning rate, suggesting both consistency and

a greater likelihood of success than teams with green or blue colors according to the IQRs and medians. This concise trend points towards red-themed teams having a potential winning advantage.

6 Conclusion

In this project, we explored NCAA basketball, focusing on scoring strategies, specific tactics for UC Davis, and team colors' impact on winning rates. Our analysis revealed that a team with high 3-point accuracy is most efficient among the three shooting strategies, but on average, 2-point shots are still the most efficient across all divisions. Besides, specifically for UC Davis, after evaluating their performance in Division I, emphasizing 2-point and 3-point shooting could boost their performance. Last but not least, the boxplot of the distribution of win rate across different background RGB categories has shown that a red background tends to have a higher winning rate.

Looking forward, the relationship between winning rate and PPS across different shooting categories offers a massive ground for further research. A regression approach could provide deeper insights into these dynamics, potentially incorporating additional performance metrics such as rebounds per game and defensive capabilities to create a more holistic model of basketball success. In addition, further research and analysis can be done by considering the psychological and cultural factors that contribute to the winning rate of background color.

7 Resource

https://www.ncaa.com/stats/basketball-men/d1
https://en.wikipedia.org/wiki/Field_goal_(basketball)
https://www.geeksforgeeks.org/how-to-do-web-scraping-using-selenium-and-google-colab/