Case Study 2: Olympic Team Selection

2023-11-01

Introduction

As the Paris 2024 Olympic Games draw closer, the United States gymnastics program is getting ready to continue their dominance. The US women's artistic gymnastics team is one of the most decorated nations in the sport. The program has won three team gold medals at the 1996, 2012, and 2016 Olympics seven team gold medals at the World Championships. The program also trained the most decorated female gymnast - Simone Biles - who holds 37 world and Olympic medals. Beyond just their success in the gym, the program has created an international hype around them with their Olympic gold medal teams receiving the nicknames of the Magnificent Seven, the Fierce Five, and the Final Five. At their last Olympic showing, they attracted an average of 17.4 million total viewers for their appearance in the finals. While receiving less of a media frenzy than the women's team, the US men's artistic gymnastics team are still strong competitors. Although they received no medals in Tokyo, the program has achieved 6 medals in their history and look to make a strong appearance in Paris.

In order to make a strong showing from both teams and continue the legacy of the women's program, choosing the best possible team is essential. USA Gymnastics has nearly 200,000 members from their beginning training to their elite levels, but few can be taken to the international stage to represent their country. Both teams will only be allowed to bring five members to the Paris 2024 Olympics. The women's team will compete on four apparatuses: vault, uneven bars, balance beam, and floor exercise. The men's team will compete on six apparatuses: floor exercise, pommel horse, still rings, vault, parallel bars, and high bar. Both teams will also compete in the team and individual all-around, allowing for 6 opportunities to medal for the women and 8 for the men. Out of these 5 member teams, competition will follow a 5-4-3 format (4 members out of the 5 compete on each apparatus, and the top 3 scores count). In the team all-around this changes to a 5-3-3 format, so every score counts for the 3 members chosen to compete. For the individual competitions, only 2 gymnasts max per country can qualify to move on to the finals.

With all of these considerations in terms of format and competition stages, selecting a team is very challenging. Selection takes into consideration not only scores from previous competitions, but also protecting the ability to defend a medal or title. It is more than just looking at who has the highest score, since a team has to be selected that can compete across all of these events and do their best at each apparatus. This selection is normally done by USA Gymnastics (for the women it was done by the Karolyis before the recent scandal), but we will create a more analytical way to select the team. Our analysis is centered around the goal of selecting 5-member men and women's gymnastics teams that will maximize medal counts at the Paris 2024 Olympic Games.

To do this, we are assessing data from past performances at recent competitions. Each of these competitions have scores for a gymnast's performance at a competition on one apparatus. These scores are a combination of an execution score (E Score) and a difficulty score (D Score). There also is a possible penalty that can be applied to a gymnasts score.

Beginning Exploratory Data Analysis

We started by looking at the distribution of total scores (the combination of both the E and D scores) for all gymnasts of both genders across all appartuses. For most of the appartuses, the distributions had the largest number of outliers towards the lower end of the distribution. This means that there were many gymnasts

who had unusually low scores compared to the rest of the score distribution. Scores were more clustered towards the upper end of the distribution, with the average scores for all of the apparatuses being around 13 points. For both genders, the vault and floor events had the smallest range of scores.

Distribution of Score of Each Apparatus by Gender
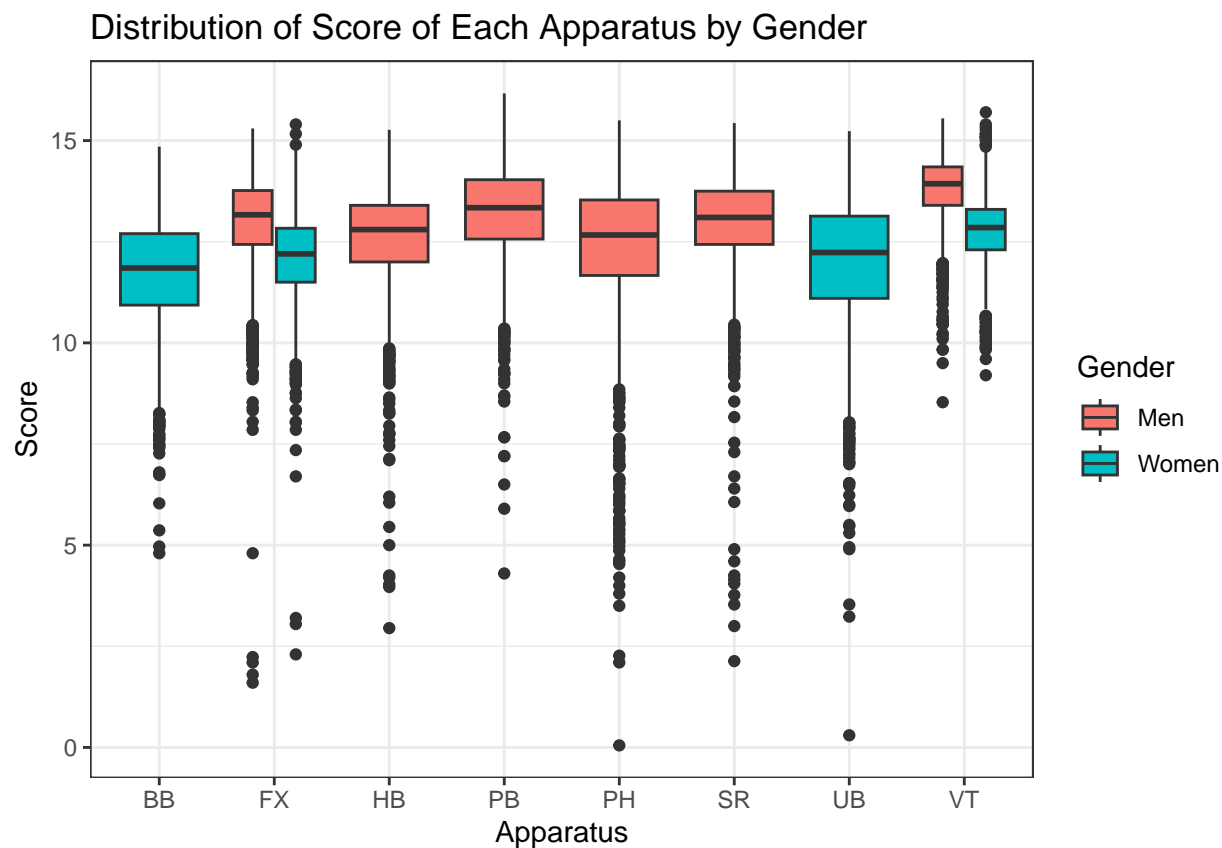
We also assessed the distribution of scores for US men and womens gymnasts against the distributions for competitors for other countries. For the women gymnasts, across all apparatuses the average score was higher for US women. This makes sense given their impressive history in the sport. The US mens distributions were on par with the distributions for gymnasts from other countries. The averages for each apparatus were mostly the same.

Distribution of Scores Across Appartus for US Women vs. Other Gymnasts
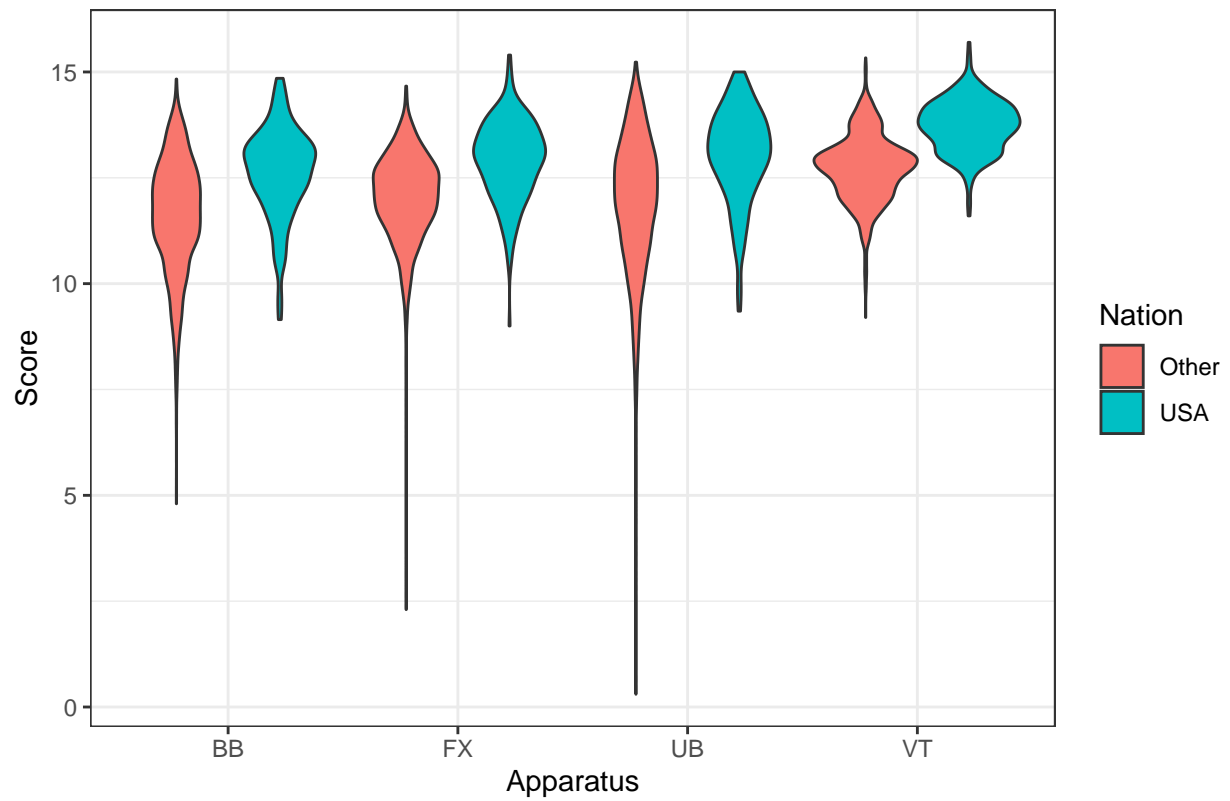
Distribution of Scores Across Appartus for US Men vs. Other Gymnasts
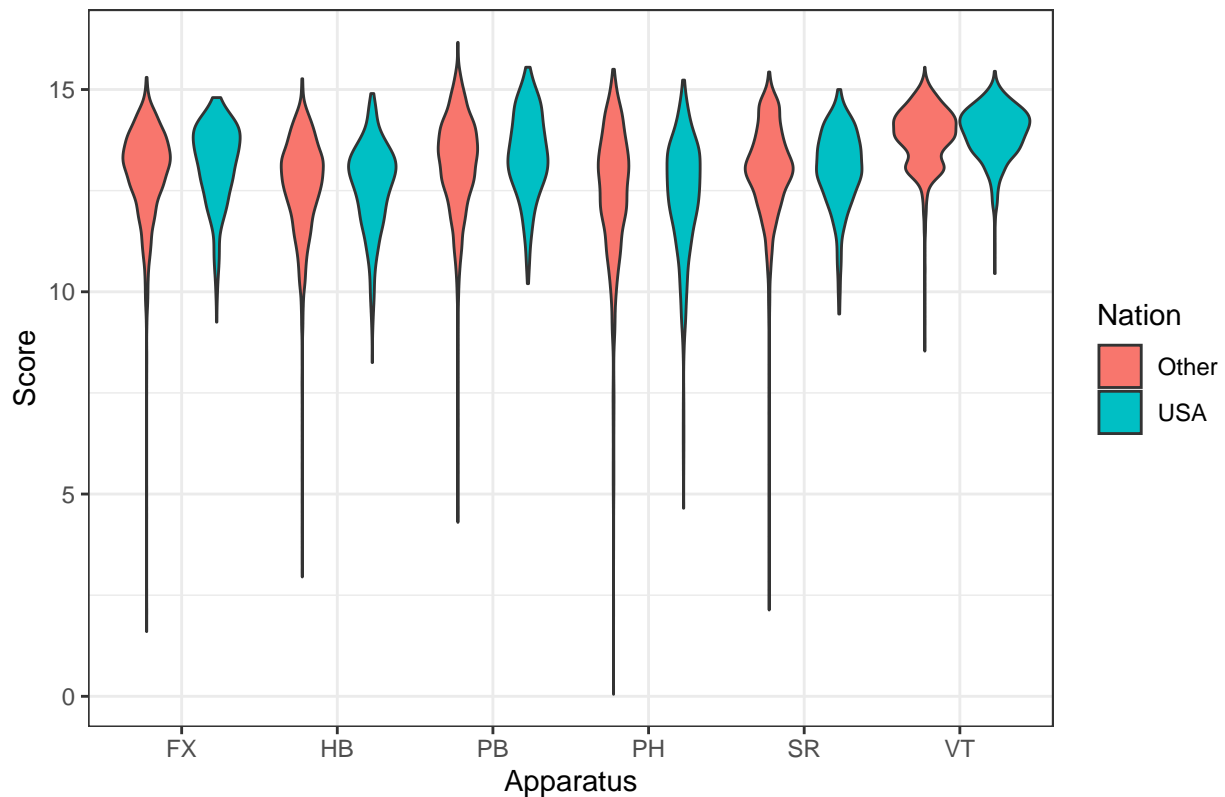
Methodology

1. Determining Constraint Level

In order to narrow down the field from the 163 men and 59 womens American gymnasts included in our dataset, we aimed to set a constraint level that a gymnast must meet to be considered. To do this, we first filtered out any gymnast that hadn't competed in 2023. This was due to both the need for relevant data to assess what state the gymnast will be in during the 2024 competitions, but also to weed out anyone that has been injured so couldn't recently compete or who has retired after the 2022 competitions. Taking out gymnasts who hadn't competed in the past year reduced the size of our selection pool to 142 men and 47 women.

We also filtered out gymnasts that had too low of an average score to warrant further consideration and even further reduce our selection pool. To do this, we calculated the average score across all competitions and apparatuses for each US gymnast (both male and female). We then found the first quartile value for each gender's average score. This means the average score that 25% of the average scores for male and female gymnasts respectively fell below this value. The 25th percentile score for US women gymnasts was 12.15156 and 12.15000 for men. After filtering out any competitors who had average scores below that, we were left with 106 men and 35 women.

2. Heuristically Analyzing Top Performers

Next, we aimed to heuristically narrow down the teams even more by considering how the process works in real life. The top two all-around finishers are automatically named to the Olympic team, which inspired our current analysis to find two individuals that should be part of a strong lineup for both the mens and womens team. We did this by taking the previous data for each performer and creating a weight metric and adjusted score that created an "adjusted" composite score to help us determine a narrower set of individuals that should, in expectation, be on the final 5.

First, we designed a formula that would weight certain historic results more/less than others based on it: 1) being in a final round, and 2) being in a more recent performance. Final round performances contribute 1.5x the weight and we defined a linear function where outcomes carry between an additional 1 to 2.5x weighing factor depending on the month and year of the tournament (January of 2022 was set as 1x and December of 2023 was set as 2.5x).

Second, we proposed a new composite score to create a new individual scoring system based on past performances that heavily increase the positive importance of top 5 ranks and also the importance (negatively) of ranking outside the top 10. We believe that consistently scoring within the top 10 is crucial for ensuring Olympic placements as the competition is much higher and even scoring within the 10-20 range will be insufficient for improving Team USA's score. For our adjusted score, gold, silver, and bronze medals contribute 10x, 7.5x, and 5x the weight of a rank outside the top 10, while placement within the 6th-10th range contributes a 2.5 weight.

With these adjusted metrics, we produce an adjusted score for each event & apparatus that a gymnast performed in, and consider three different methods of analyzing these metrics:

1. Total sum of adjusted scores: this metric will provide us with a comprehensive way to determine a ranking of the top all-around performers, based on the sum of their weighted adjusted scores in each apparatus they perform in, that we would like to be staples on any team.

2. Best individual apparatus score: this metric will provide a ranking of the highest weighted score a competitor has in any individual apparatus.

3. Best two-apparatus score: this metric will provide a ranking of the average of the two highest scores a competitor has in their two top apparatus.

We aimed to use the 2nd and 3rd metrics to determine individuals that would be very likely to excel in a few apparatuses, alongside the two best all-around performers.

3. Simulating Composite Scores

The simulation of composite team scores our methodology can be broken down into two parts. To prepare the data for our simulation of composite scores, we must fit a distribution to describe the qualifying scores of each gymnast. Because each gymnast in the dataset has varying amounts of qualifying data, we believe these individual distributions will be more descriptive than a summary statistic, such as just a mean score. We anticipate that an understanding of the standard deviation of scores will help optimize the accuracy of our simulation. To assess which distribution class would best fit our score data, we first displayed the histograms for both men's and women's qualifier scores.

Distribution of US Men's Gymnastics Scores
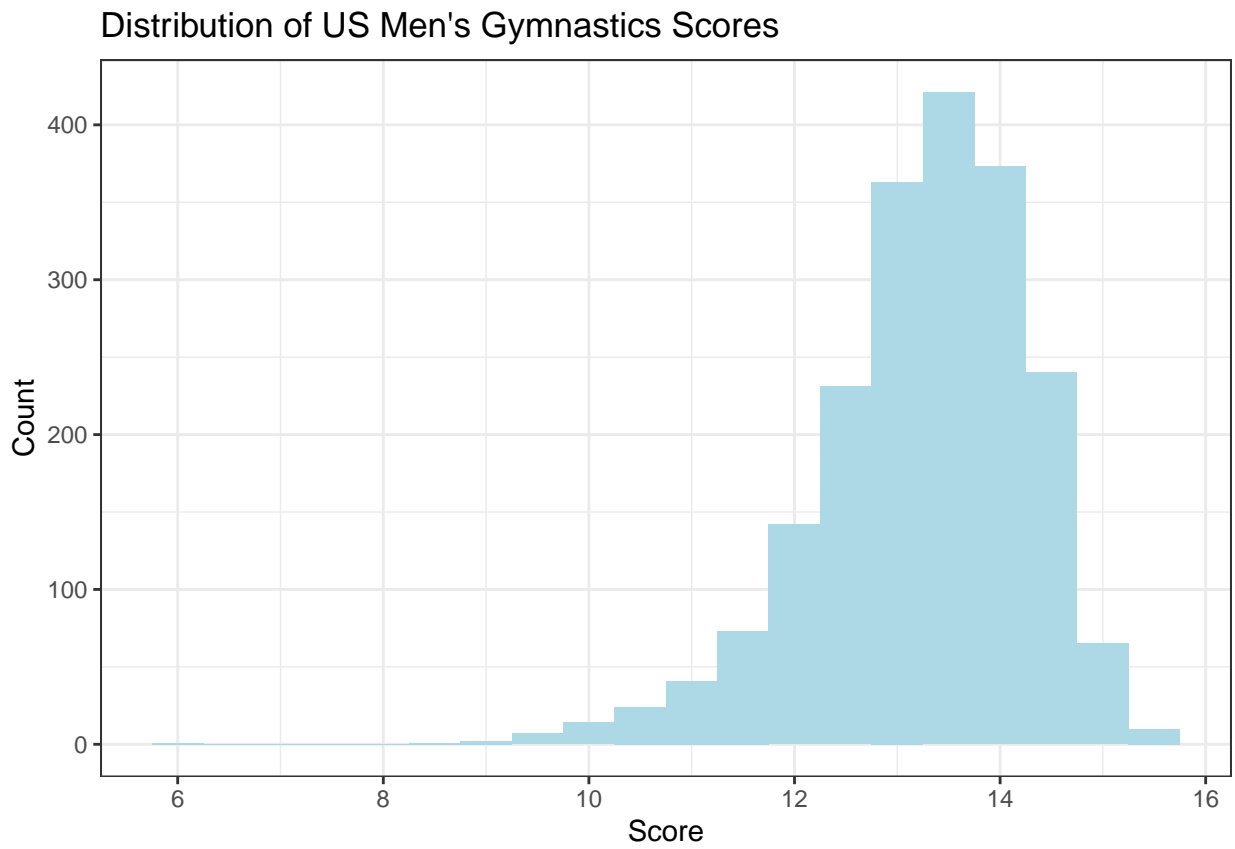
Distribution of US Women's Gymnastics Scores
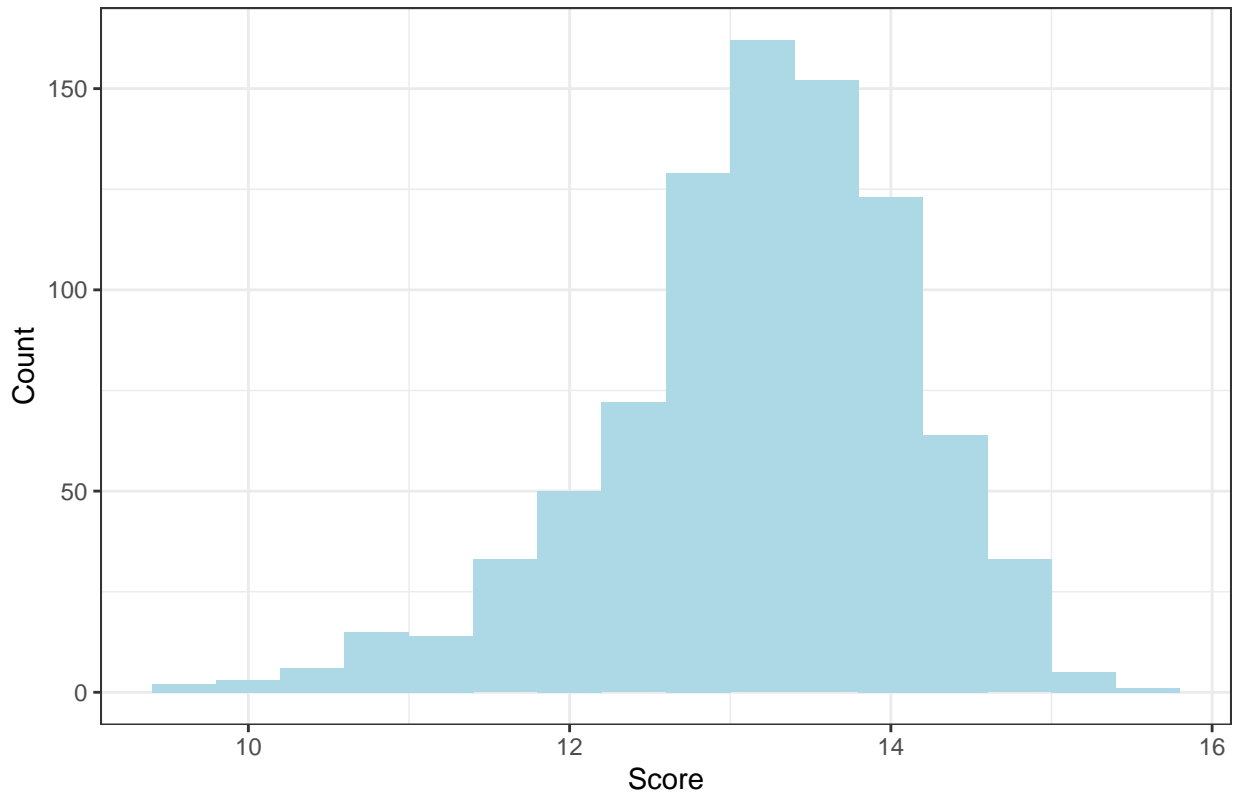
Despite the fact that our data is bounded (scores are between 0-16), the concentration of scores in our histogram revealed symmetry around approximately the 13 score range. We therefore elected to fit the scores of each gymnast to a normal distribution, as opposed to a beta distribution or a truncated normal.

To fit these distributions, we will start by iterating through the rows of our US Mens and USA Womens data separately. For each name in the data, we will take the corresponding subset of the data containing all qualifying scores. Using these scores, we will fit a distribution to the associated gymnast, and store the mean and standard deviation, as well as the name of the gymnast into a data frame.

To eliminate redundant information in our final datatable, we will first use regular expressions to extract the relevant name information (i.e. Simone-BILES instead of Simone-BILES [1]), and then use the distinct function so that each gymnast has just one row in our data frame. We will repeat this process so that we will have two separate data frames containing the names, mean scores, and standard deviation of scores for both men and women USA gymnasts. We can now use these data frames to run our composite score simulation.

The second step in our methodology is running our simulation to generate composite team scores for randomly sampled groups of six USA gymnasts. Using the output of the second step in our methodology, we can specifically sample teams from a smaller subset of 10 male gymnasts and 10 female gymnasts that were selected to either be the best all-around gymnasts, or gymnasts that were highly proficient in a few events. We start our simulation by randomly sampling five different gymnast names, means, and standard deviations without replacement (men and women separately). For each gymnast sampled, we will generate one score based on the gymnasts' mean and standard deviation of qualifying scores using the rnorm() function. We will then add this value to our variable for the composite score. In cases where our standard deviation was NA, we assigned a generic deviation of zero. When our standard deviation value was less than 1e-10, we simply assigned the mean score for that gymnast as our simulated score, as these scores often resulted in run-time errors. We will run this simulation process 5000 times, storing the composition of gymnasts and the associated composite score. Using this data frame, we can observe the combinations in descending order to reveal which gymnasts contributed to the highest composite score.

```
## # A tibble: 10 x 2
##    Gymnasts                                                   Composite_Score
##    <chr>                                                      <chr>
##  1 Taylor-BURKHART, Frederick-RICHARD, Paul-JUDA, Stephen-NEDOR~ 75.95991075774~
##  2 Donnell-WHITTENBURG, Taylor-BURKHART, Asher-HONG, Khoi-YOUNG~ 75.646142558567
##  3 Brody-MALONE, Stephen-NEDOROSCIK, Asher-HONG, Paul-JUDA, Tay~ 75.567457124974
##  4 Asher-HONG, Curran-PHILLIPS, Brody-MALONE, Taylor-BURKHART, ~ 75.56065179578~
##  5 Frederick-RICHARD, Asher-HONG, Stephen-NEDOROSCIK, Khoi-YOUN~ 75.49790103515~
##  6 Stephen-NEDOROSCIK, Donnell-WHITTENBURG, Paul-JUDA, Asher-HO~ 75.248167268251
##  7 Paul-JUDA, Curran-PHILLIPS, Taylor-BURKHART, Asher-HONG, Kho~ 75.23802877348~
##  8 Joshua-KARNES, Asher-HONG, Frederick-RICHARD, Brody-MALONE, ~ 75.10568583911~
##  9 Stephen-NEDOROSCIK, Curran-PHILLIPS, Donnell-WHITTENBURG, Br~ 75.062481528119
## 10 Joshua-KARNES, Brody-MALONE, Donnell-WHITTENBURG, Curran-PHI~ 75.00832985255~
```

```
## # A tibble: 10 x 2
##    Gymnasts                                                   Composite_Score
##    <chr>                                                      <chr>
##  1 Zoe-MILLER, Simone-BILES, Kayla-DICELLO, Skye-BLAKELY, Josce~ 74.58965770149~
##  2 Zoe-MILLER, Simone-BILES, Ashlee-SULLIVAN, Kaliya-LINCOLN, J~ 74.53404738220~
##  3 Ashlee-SULLIVAN, Jade-CAREY, Kaliya-LINCOLN, Skye-BLAKELY, S~ 74.28785641931~
##  4 Jade-CAREY, Simone-BILES, Kayla-DICELLO, Shilese-JONES, Josc~ 74.28090672985~
##  5 Kayla-DICELLO, Joscelyn-ROBERSON, Jordan-CHILES, Simone-BILE~ 74.08340802943~
##  6 Jade-CAREY, Simone-BILES, Zoe-MILLER, Jordan-CHILES, Kaliya-~ 74.08281136475
##  7 Kayla-DICELLO, Jordan-CHILES, Joscelyn-ROBERSON, Simone-BILE~ 73.94747630228~
##  8 Zoe-MILLER, Jade-CAREY, Jordan-CHILES, Kaliya-LINCOLN, Simon~ 73.89289633232~
##  9 Kayla-DICELLO, Simone-BILES, Kaliya-LINCOLN, Joscelyn-ROBERS~ 73.75254600163~
## 10 Kaliya-LINCOLN, Joscelyn-ROBERSON, Jade-CAREY, Simone-BILES,~ 73.73892088762~
```

Sources

Artistic Gymnastics. Paris 2024. (2023, June 1). https://www.paris2024.org/en/sport/artistic-gymnastics/

Wikimedia Foundation. (2023a, October 8). United States men's National Artistic Gymnastics Team. Wikipedia. https://en.wikipedia.org/wiki/United_States_men%27s_national_artistic_gymnastics_team

Wikimedia Foundation. (2023, November 13). United States women's National Artistic Gymnastics Team. Wikipedia. https://en.wikipedia.org/wiki/United_States_women%27s_national_artistic_gymnastics_team