UCSAS 2024 Data Challenge

January 15, 2024

Anusha Bhat, Sarah Li, Shivani Ramalingam

Faculty Advisor: Ron Yurko (Department of Statistics and Data Science, Carnegie Mellon University)

1. Introduction

The objective of this data challenge is to strategically assemble the optimal group of five male and female athletes to comprise a USA gymnastics team, with the aim to maximize success in the Paris 2024 Olympics. The task involves crafting an analytical model capable of identifying and comparing the predicted medal count across the eight medal events for men and six medal events for women.

A total of 192 artistic gymnasts, evenly distributed between 96 men and 96 women, are slated to compete in Paris 2024. Men and women compete separately from their respective apparatus types. Women can compete on 4 apparatuses--vault, floor exercise, balance beam, and uneven bars--while men can compete on 6 apparatuses--vault, floor exercise, parallel bars, high bar, pommel horse, and still rings. The team events will feature 12 teams of five athletes, in both the men's and women's categories. In instances where countries do not secure a full team entry, a maximum of three individuals per country will have the opportunity to qualify for the Olympics. The determination of these additional 36 entries for each gender will be based on the results of the 2023 World Championships, the 2024 World Cup Series, and the 2024 Continental Championships.

Notably, three teams for each gender secured qualification at the 2022 World Championships in Liverpool, England. For the men, China, Japan, and Great Britain qualified, while the women's competition saw qualification for teams from the United States, Great Britain, and Canada. Furthermore, nine other countries in each gender earned team qualifications based on their performance at the 2023 World Championships in Antwerp, Belgium, scheduled from September 30, 2023, to October 8, 2023.

The intricacy of selecting a USA team is further compounded by the structure of the Olympic Competition, where athletes initially partake in a qualifying round. In this round, their scores determine team advancement, as well as individual qualification for the all-around and apparatus finals. An athlete must compete on all apparatuses to be eligible for the individual all-around final. The top 24 athletes qualify for the individual all-around final, with a maximum of two gymnasts per country, and the top eight athletes on each apparatus qualify for the apparatus final, again with a maximum of two gymnasts per country advancing. The top eight teams progress to the team final based on the sum of the top three all around scores for the four competing members.

In the team all-around, individual all-around, and individual apparatus finals, scores from the qualifying round are discarded. The team all-around final results are determined by the sum

of the all around scores for the three competing members. An athlete's scores in the team all-around final do not impact their individual all-around or apparatus final scores.

In summary, the overarching goal of this data challenge is to identify the most effective team of USA female and male gymnasts, with success defined as achieving the highest weighted medal count--with gold receiving the highest weighting and bronze the lowest. We also explore the relationship and correlation between different apparatuses.

2. Data

For constructing our model, we used data containing scoring and ranking information of all gymnastics competitions held from 2022-2023 that were meticulously gathered by the US Olympics and Paralympics Committee (USOPC) & University of Connecticut. Each row of the dataset represents a gymnast's performance in a specific event or apparatus at a particular competition. The data remains actively updated, with the latest information extending to the 2023 World Artistic Gymnastics Championships in Belgium. The variables in the dataset include information regarding dates and locations of competitions, event round, apparatus, score factors, penalty and athlete name. The most significant variables include the gymnast's name, gender, country, apparatus, round, and total score. To enhance data consistency, we pre-processed the data by merging first and last names, getting rid of middle names, and making apparatus and country names have uniform casing.

The dataset encapsulates scores for both men's and women's events drawing from 39 domestic and international competitions between 2022 and 2023, involving 109 countries and 1917 individuals. In our model, we created three sub-datasets consisting with one for US data, one for the 11 other qualifying countries, and one for the remaining countries that did not qualify full teams.

In the gymnastics competition, the notation in the "Apparatus" column distinguishes between "VT," "VT1," and "VT2", which each represent vault performances. The key distinction arises when athletes participate in individual apparatus events as some gymnasts opt to execute the vault twice. In such cases, the first result is designated as "VT1," while the second result is marked as "VT2." Consequently, "VT1" may indicate either the sole performance of a vault or specifically denote the first of two vaults in an individual apparatus event. Conversely, in the all-around competition, there is typically only one vault that contributes to the overall score (denoted as "VT"). To sum up, "VT" signifies the performance of a single vault, "VT1" may

indicate either one vault or the initial vault of a pair, and "VT2" specifies the second vault when two are executed, with the notation adjusting based on the competition format.

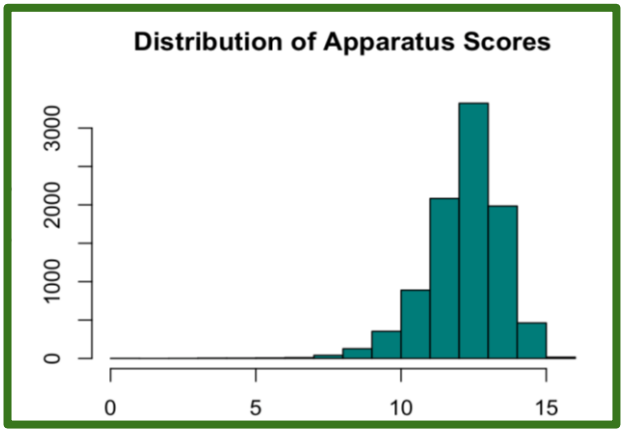

Fig. 1: Total scores are left skewed and range between ~7 to ~15, with the max possible being 16.

Across all apparatuses, we observe, on average, that the USA men's and women's teams score in the higher range of the score distribution displayed in Figure 1. We can use this knowledge to build intuition for our model as we can expect that the scores for each athlete should be relatively high in our final output.

3. Methods

To achieve our ultimate goal of predicting medal outcomes for the most successful Team USA, we implemented a simulation model that leverages historical data and fixed nations based on the assumption of the top 4 athletes from each country. We decided to create a simulation model using linear mixed effects regression over alternatives such as random forest or simple linear regression models since we have to consider the intricate nature of the gymnastics competition, which involves multiple rounds and apparatuses. For example, our dataset lacked sufficient labeled training data, and the available quantitative variables were limited to total score, difficulty score, and execution score, with total score being a derivative of the latter two. This limitation on our data renders random forests or regression inappropriate for addressing the nature of the competition progression. However, The simulation model allows us to tailor the code to accommodate the specific rules, flow of the Olympics competition, and complexities inherent in the gymnastics scoring system, making it a more fitting choice for our analysis. Additionally, we only used the 2022-2023 dataset since these years are part of the

most recent Olympic cycle with the most up to date rules and regulations compared to the 2017-2021 dataset.

3.1 Model Assumptions

For our simulation to run, we must first make several assumptions regarding the data and the competition structure:

1. Fixed nations:
    a. 3 countries qualified full teams of five athletes at the 2022 World Championships and 9 more countries qualified at the 2023 World Championships for both genders. The women's teams are from the **U.S., Great Britain, Canada, China, Brazil, Italy, the Netherlands, France, Japan, Australia, Romania, and South Korea**. The men's teams are from **China, Japan, Great Britain, the U.S., Canada, Germany, Italy, Switzerland, Spain, Turkey, the Netherlands, and Ukraine**.
    b. We must assume which players the countries (excluding the U.S.) are sending out. Fixing these players allows us to model how the U.S. team may perform compared to the other teams at the Olympics.
    c. We averaged the scores of the teams sent to the 2023 World Championships by these qualifying countries. We then selected the 4 highest scoring athletes from each team since these were the most recent teams composed by these nations in our dataset. This allows us to fix 44 of the 84 actively competing athletes at the Olympics (12 of the 96 players are alternates).
2. Fixing the 36 individuals:
    a. Currently only 20 individuals have qualified for the women's competition and 15 for the men's. The remaining 16 women and 21 men will qualify through various competitions in early 2024. We fixed the known individuals who have qualified in our model. We then selected the remaining individuals based on the highest scoring athletes who can qualify through criterias 3, 6-7, and the universality spot. During our selection, we ensured that the individual players were not from a qualifying country that is sending a full team.
3. Team Composition:

    a. Although teams consist of five members, four compete for consideration. Unable to account for substitutes in our model, we only simulate scores for teams of 4.
4. Score and Rank:
    a. Medals are determined by rank, however, since rank and scores are correlated (the higher the score, the better an athlete will rank), we can use scores as a proxy for modeling rank outcomes for each player. This is more decisive than simply using rank since score is a decimal value not a whole number, allowing for clearer separation between athlete ranks.
5. Tie Breaking:
    a. In the Olympics, ties are broken using the execution and difficulty scores. During preliminary testing, we did not observe the occurrences of ties, so we currently assume that there is no need for tie breaking.
6. VT1 & VT2
    a. Our dataset contains scores for two vaults (VT1 and VT2), which are the first and second routines performed respectively by athletes competing in the vault individual apparatus event. There is also a vault (VT) score for those who competed on vault but only performed one routine (e.g. for individual all around). We collapsed these three into a singular vault event since there was no difference in the distribution of the three vaults.
7. Final Round Score Distributions:
    a. In our dataset, the final round scores are stored based on which final event they were performed for (e.g. team All Around, individual All Around, or apparatus finals). We observed that there was no difference in the score distributions for each apparatus for each final event, so we collapsed them into one final round. This allows us to have more data to sample from during our simulation.
8. No Correlations Between Apparatuses:
    a. To address our research question regarding the relationship between individual events, we constructed a PCA biplot shown in Figure 2. This plot reveals that there may be a positive correlation between the apparatuses, so our simulation may have to sample the scores together.
    b. Further investigation with a paired plot reveals that the positive correlation is explained by individual athlete clusters (e.g. Simpson's Paradox). More successful athletes tend to perform better in each event. However, within an

individual athlete's cluster, there is no correlation between the scores. A sample of our pairs plots for the women's team is included in Figure 3.

c. Since there is no correlation between apparatuses at the individual athlete level, we do not have to sample scores together and can treat the apparatuses independently. This result applies for all apparatus pairs for both genders.

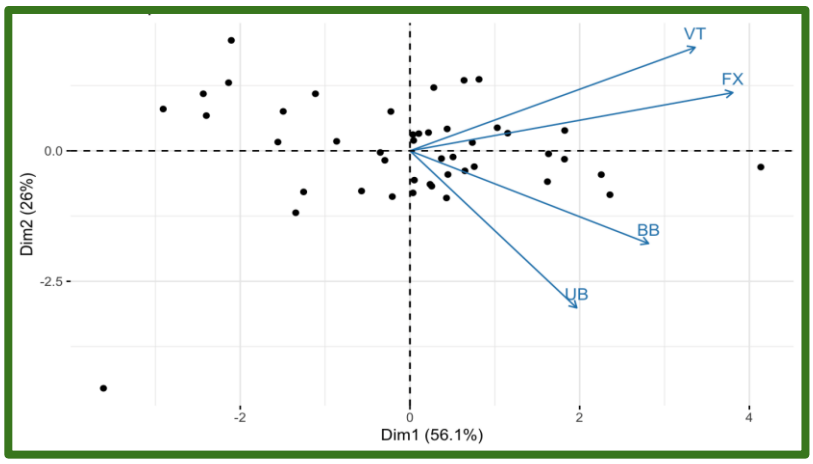

Fig. 2: The apparatuses are positively correlated with one another.

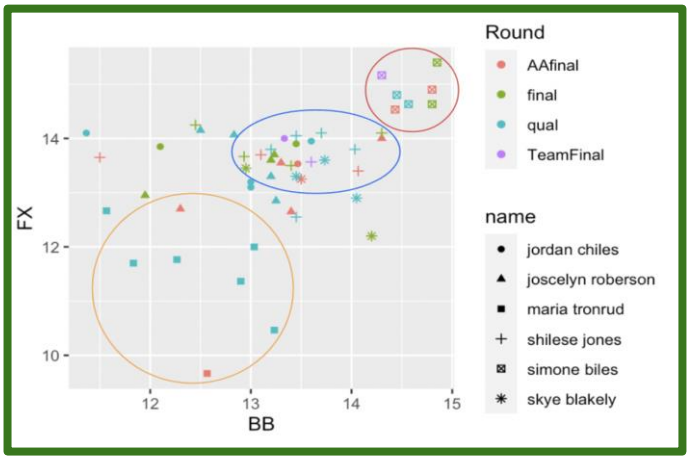

Figure 3: Individual athlete clusters do not have a positive correlation between scores, therefore we can treat the apparatuses as independent.

3.2 Constructing the Model

The simulation operates by conducting a series of intricate steps to predict medal outcomes for a team of five athletes. After preprocessing our data for consistency, we construct

a new dataset using a linear mixed effects regression model. For this regression, we specify the athlete's full name, gender, and country variables as fixed effects and the apparatus variable as a random effect to predict an athlete's score. These are essentially varying intercepts that shift the intercept term instead of the slope of the total score variable for an athlete. Constructing this regression model allows us to take into account the mixed effects in our original dataset to effectively move our scores towards the average score. The model also allows us to simulate 500 scores for an athlete in a particular apparatus, creating more data to sample from compared to the low amount of scores available for a particular athlete-apparatus pair in the original dataset.

After constructing the new dataset based on the linear mixed effects regression model, we begin the simulation part of our analytical model. In this part, we first determine the players from the other qualifying nations and the 36 individuals as mentioned in our assumptions, then we pass in a combination of US athletes and simulate a qualification then a final round of the Olympics using these fixed individuals and US players. A single round of the simulation is comprised of the following key stages

1. Qualification Round:
    a. For each apparatus, we randomly sample a score for each of the 84 players in the competition.
    b. After sample the scores, we determine which players advance to each of the final rounds based on Olympic Rules.
        i. Individual Apparatuses: top 8 players advance for each apparatus with a maximum of 2 athletes from one country allowed to advance.
        ii. Individual All Around: For an individual player, we sum their scores across all of the apparatuses. The top 24 players advance to the finals, with a maximum of 2 athletes from one country allowed to advance. A player must compete in all events in order to be considered for this event.
        iii. Team All Around: For one team, calculate the individual all around scores for each team member and then sum the top 3 scores to determine the team all around score. The top 8 teams will advance to the finals, retaining the 3 members used for the team score calculation.
2. Final Round:
    a. Scores from the qualification round are discarded.

b. For each apparatus in each final event (individual apparatuses, individual all around, and team all around), sample the scores for the qualifying athletes.
c. Calculate the individual and team all around scores. Since these final events are sampled separately, the individual all around scores and the individual apparatus scores do not affect the team all around score.
d. Determine the top 3 gymnasts for each apparatus final and the individual all around, and the top 3 countries for the team all around. These placements correspond to the gold, silver, and bronze medalists for these events.

3. Weighted Medal Count
   a. For the athletes in the input combination, determine the weighted medal count achieved by the combination. For one country, the women's team can earn a maximum of 11 medals (2 for each of the four apparatuses, 2 medals for individual all around, and 1 for team all around), whereas, the men's team can earn a maximum of 15 medals (2 for each of the six apparatuses, 2 medals for individual all around, and 1 for team all around). We count how many medals were won across the individual all around, team all around, and individual apparatus final events, assigning 3 points to a gold medal, 2 points to a silver medal, and one point to a bronze medal. The highest possible weighted medal count for the women's team is 28 and 38 for the men's team.

In Figure 4, we visualize the flow of the Olympics competition. We note that one qualification round is used to determine the players that advance to each of the final events (rather than three qualification rounds). There is also a different number of medals that are awarded by each final event.

Fig. 4: The team all around and individual all around finals each award 3 medals. Each individual apparatus final awards 3 medals, resulting in a total of 12 medals for the women's competition and 18 for the men's.

One iteration of the model outputs a predicted weighted medal count for a given combination of US athletes. We average over 100 iterations to determine the expected weighted medal count for the combination and repeat this process for all possible combinations in our dataset. This allows us to determine the best combination robustly, while incorporating the randomness of real competitions. We also only considered combinations of 4 athletes among the top 10 women's and men's USA gymnasts in the original dataset based on average total score across entries. We note that for the women's combinations, we fixed Simone Biles as a player for each combination since she is the current top gymnast in the world and is actively vying for a spot on the national team. For the men's combinations, we fixed Curran Phillips as a player for each combination since we found that he had the best score trend throughout past competitions during our elementary data analysis. This allows us to improve our time complexity by reducing the number of possible player combinations (total of 168 combinations with 84 per gender) to pass into our model while ensuring we are considering the best contenders. For 100 iterations on one combination of athletes, we observe small error bars in our results, therefore, we opted not to perform 1,000 simulations in our model.

4. Results

Our simulation runs 100 iterations for each combination of women's and men's teams. The results of several women's teams are plotted in Figure 5, and the breakdown of their medals by event are in Figure 6.

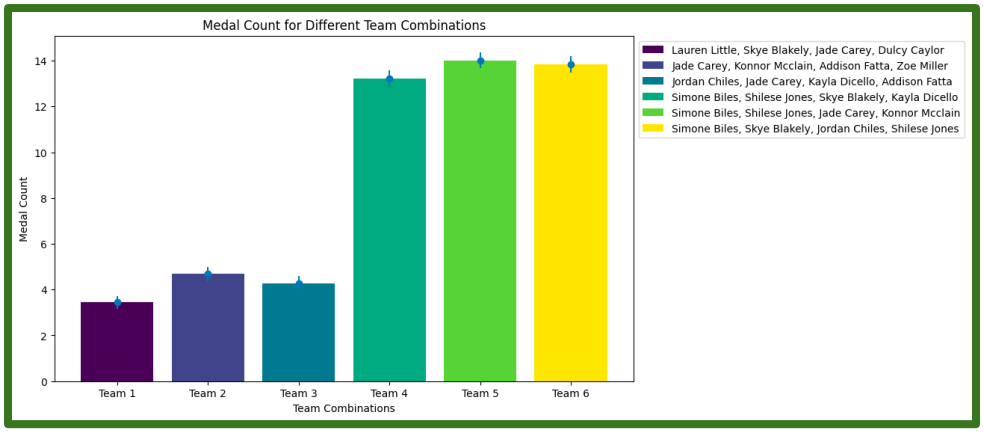

Fig.5 : Overall weighted performance of 6 different combinations of US female gymnasts.

The optimal lineup identified for the women's gymnastics team is composed of the following athletes: **Simone Biles, Shilese Jones, Jade Carey, and Konnor Mcclain**. This team averaged 13.8 out of the highest weightage of 28.

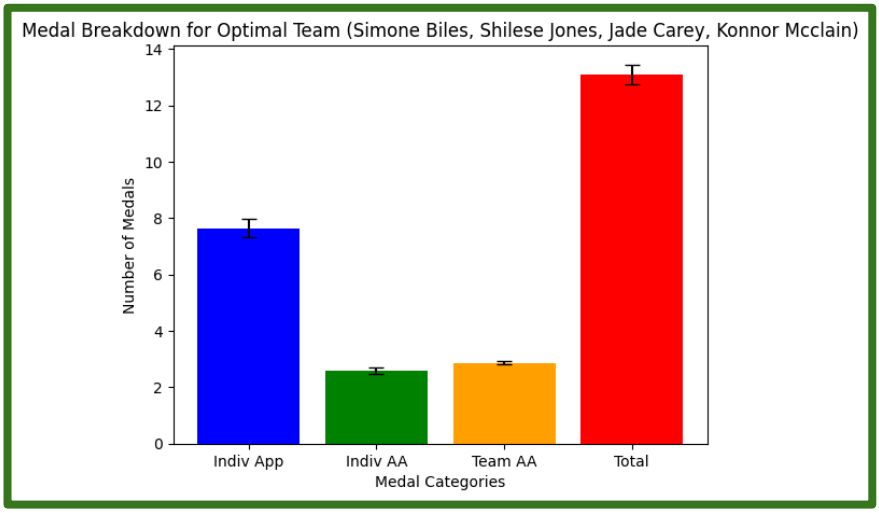

Fig. 6: The best performing women's team placed in all events.

From Figure 6, we can see that this team averages close to a 3 for the weighted team All Around score, indicating that they are winning more gold medals on average for this event--the hallmark of a successful team. We can additionally observe that a majority of the medals are being won in individual apparatuses and at least one or two of these women won an individual All Around placement. Furthermore, this women's team is meddling in almost all of the final events which is similar to the performance that we typically see by the US women's gymnastics team at past Olympics.

Overall, these results make intuitive sense. The USA women's team has consistently medaled in the team all around event since 1992 and averages around 4-6 medals in each of the Olympics competition since then. This parallels our simulation predictions for the women's teams, suggesting that our model may be approximately correct.

For specifying the alternative player, we performed elementary data analysis inspecting the rolling medians of several top women's athletes in our dataset as pictured in Figure 7.

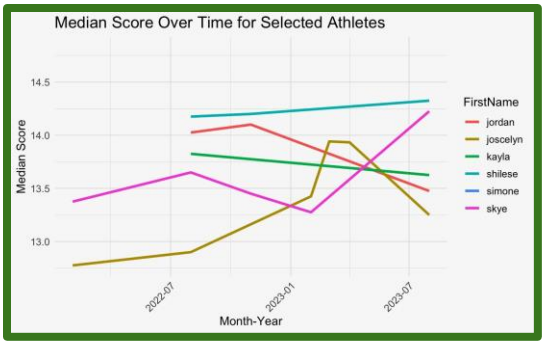

Fig. 7: Rolling median scores over time for top US Women's gymnasts.

From Figure 7, we can see that Sky Blakely has the best score performance and trend of the top female gymnasts that are not in the optimal team. Due to this, we will designate Skye Blakely as the alternative, and fifth athlete of our women's team.

The results of several men's teams are plotted in Figure 8, and the breakdown of their medals by event are in Figure 9.

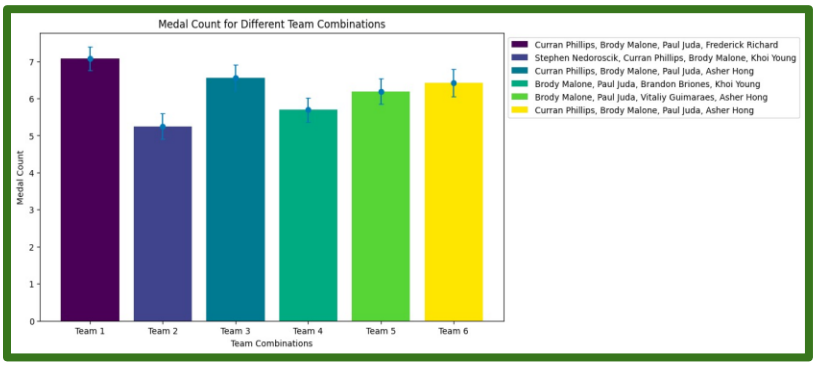

Fig. 8: Overall weighted performance of 6 different combinations of US male gymnasts.

The optimal lineup identified for the men's gymnastics team is composed of the following athletes: **Curran Phillips, Brody Malone, Paul Juda, and Frederick Richard**. This team averaged 7.08 out of the highest weightage of 38.

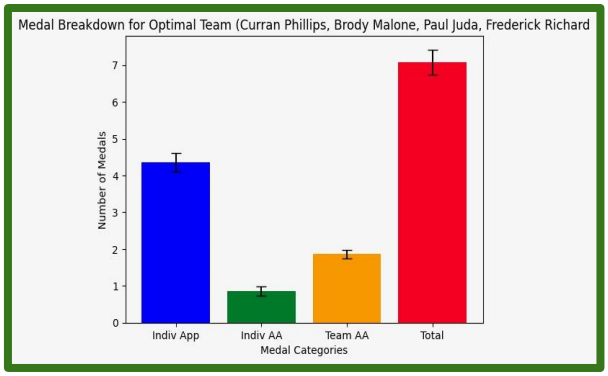

Fig. 9: The best performing men's team placed in all events.

From Figure 9, we can see that this team averages close to a 2 for the weighted team All Around score, indicating that they are winning more silver medals on average for this event. We can additionally observe that this team won between 2-4 medals in the individual apparatuses and at least one of these men won a bronze in the individual All Around placement. Furthermore, this men's team is meddling in almost all of the final events. This is a marked improvement to the performance that we typically see by the US men's gymnastics team at past Olympics. This could potentially be due to the fact that we weigh all athletes equally, giving equal weightage to emerging athletes and established athletes. For the men's team, we specify Asher Hong as the alternate since he recently emerged as the youngest US male individual all around champion in 34 years at the 2023 U.S. Championships.

5. Discussion

In addressing our research question and endeavoring to model the most successful women's and men's gymnastics USA teams for the Paris 2024 Olympics, we opted to construct a simulation model using linear mixed effects regression due to its robust capacity to account for the intricate competition structure. It is essential to acknowledge the inherent limitations of our approach. Our model cannot account for unpredictable human behaviors, such as athletes dropping out of the competition, and does not consider the specific order of gymnasts on apparatuses or how they are grouped in the competition progression. Moreover, the model is contingent on athletes being present in the dataset, limiting predictions for potential rising stars who are not included. Lastly, since we randomly sample from the simulated scores constructed by the linear mixed effects regression model, the optimal team may vary from run to run.

However, we note that the teams we selected were based off of several runs of the model and an observation of player trends in the most successful teams.

Our proposed women's team, consisting of **Simone Biles, Shilese Jones, Jade Carey, Konnor Mcclain, and Skye Blakely as the alternative** emerged as the optimal choice through the simulation model, which considered various athlete combinations. For the men's team we found **Curran Phillips, Brody Malone, Paul Juda, and Frederick Richard and Asher Hong as the alternate** as the optimal choice. These 5 female athletes and 5 male athletes will comprise the most successful USA gymnastics team at the upcoming Paris Olympics out of all possible teams in the dataset.

6. References

Armour, Nancy. “Simone Biles Finishes with Four Golds at 2023 Gymnastics World Championships.” *USA Today*, Gannett Satellite Information Network, www.usatoday.com/story/sports/olympics/2023/10/08/simone-biles-at-2023-gymnastics-world-championships-live-updates/71099947007/#:~:text=Simone%20Biles%20finishes%20with%20four%20golds%20at%202023%20Gymnastics%20World%20Championships&text=ANTWERP%2C%20Belgium%20%E2%80%94%20Five%20medals%2C,Biles%20can%20pull%20that%20off.

“Gymnastics at the 2024 Summer Olympics – Qualification.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Gymnastics_at_the_2024_Summer_Olympics_%E2%80%93_Qualification

TOKYO 2020 ARTISTIC GYMNASTICS MEN’S VAULT RESULTS, olympics.com/en/olympic-games/tokyo-2020/results/artistic-gymnastics.

“List of Olympic Medalists in Gymnastics (Women).” *Wikipedia*, Wikimedia Foundation, 29 Nov. 2023, en.wikipedia.org/wiki/List_of_Olympic_medalists_in_gymnastics_(women).

.