Case 2: UCSAS 2024 USOPC Data Challenge

Sean Li, Christopher Tsai, Benjamin Thorpe, Jerry Xin

2023-11-06

Introduction

Background

The Olympics brings about a sense of national pride seldom felt by athletes and viewers alike. Not only national pride, but the sheer amount of viewers adds to the pressure and importance of an event like the Olympics for national athletic committees. In every Olympic games since 2008, over 3 billion unique viewers tuned in worldwide to view the games. To put that in perspective the Super Bowl, the most watched individual sporting event in America, garners around 115 million viewers. Americans, in particular, are fascinated by the sport of artistic gymnastics with the highest percentage of viewers claiming to be interested in the event. It follows that success in this sport has long been a source of national pride and international recognition for the United States. The breathtaking performances, grace, and unparalleled athleticism displayed by American gymnasts have not only brought numerous Olympic medals home but also served as an inspiration for generations of aspiring athletes. Simone Biles, Gabby Douglas, Nastia Liukin are all household names due to their success in Olympic gymnastics. As we approach the next Olympic Games in Paris 2024, it is extremely vital that team USA finds success and brings home as many medals as possible in artistic gymnastics. In the era of data-driven decision-making, the role of predictive analytics in the world of sports has gained unprecedented significance and the need for innovation and precision in athlete selection has never been more crucial. Our goal is to use data analytics to predict the best five-member lineups for both the USA Men's and Women's artistic gymnastics teams, optimizing for total medal count.

First, understanding Olympic gymnastics and its scoring system will be vital to understanding the decision making process later in this paper. Women's artistic gymnastics and Men's artistic gymnastics vary slightly with women competing in 4 apparatuses and men competing in 6. Each country, if they have qualified for the team all-around competition (which the US has), can send 5 athletes; otherwise a country can send a maximum of three individuals. The event begins with a qualification round which triples as qualifying for team all-around, individual all-around, and individual apparatus finals. A team will send 4 of their 5 athletes to compete for each of the apparatuses. For the team score, the top 3 of 4 scores will count. If an individual is one of the 4 of 5 team competitors on every apparatus they are then eligible for the individual all-around final (individuals not affiliated with a team that qualify for the individual all-around event are also eligible for the final). If an individual is in the top 8 in qualifying for the given apparatus they also qualify for that individual apparatus final. For all individual finals, however, a maximum of two athletes per country can qualify for the final. All things considered, this gives 6 medal opportunities on the women's side (team all-around, individual all-around, each of 4 apparatuses) and 8 on the men's side (team all-around, individual all-around, each of 6 apparatuses).

This selection process has, historically, relied heavily on the expertise and intuition of coaches and selection committees and thus research into data-driven decision making in gymnastics is sparse. While some prior studies have explored the use of data analytics in sports in general, they often fall short in the context of artistic gymnastics. Many of these studies are limited in scope and do not consider the nuanced aspects of gymnastics performance and individual variance from round to round. Also, we plan to address the fact that some apparatuses (like pommel horse) are more difficult to judge by taking these higher variances in

scores that earn medals into account when optimizing our team (Guston, 2023). We will explore how to mathematically account for this in our analysis.

Research Objective

This study aims to overcome these limitations by analyzing athlete-specific data and incorporating domain expertise from the gymnastics community. It will offer a comprehensive and customized approach that addresses the unique competition format presented by artistic gymnastics and provides the USA Olympics Committee with a robust tool to make informed and data-driven decisions. Our goal is to recommend a team of 5 women and 5 men for the Olympic Games in Paris 2024, with an expected medal count for each team.

Dataset Overview

The data is taken from major domestic and international competitions worldwide. We were provided 2 datasets, 1 from the competitions leading up to the Tokyo Games in 2020 (2017 to 2021), and 1 from the competitions leading up to the Paris Games in 2024 (2022 to 2023). We only used the 2022 to 2023 data, as the competition format changed between 2021 and 2022.

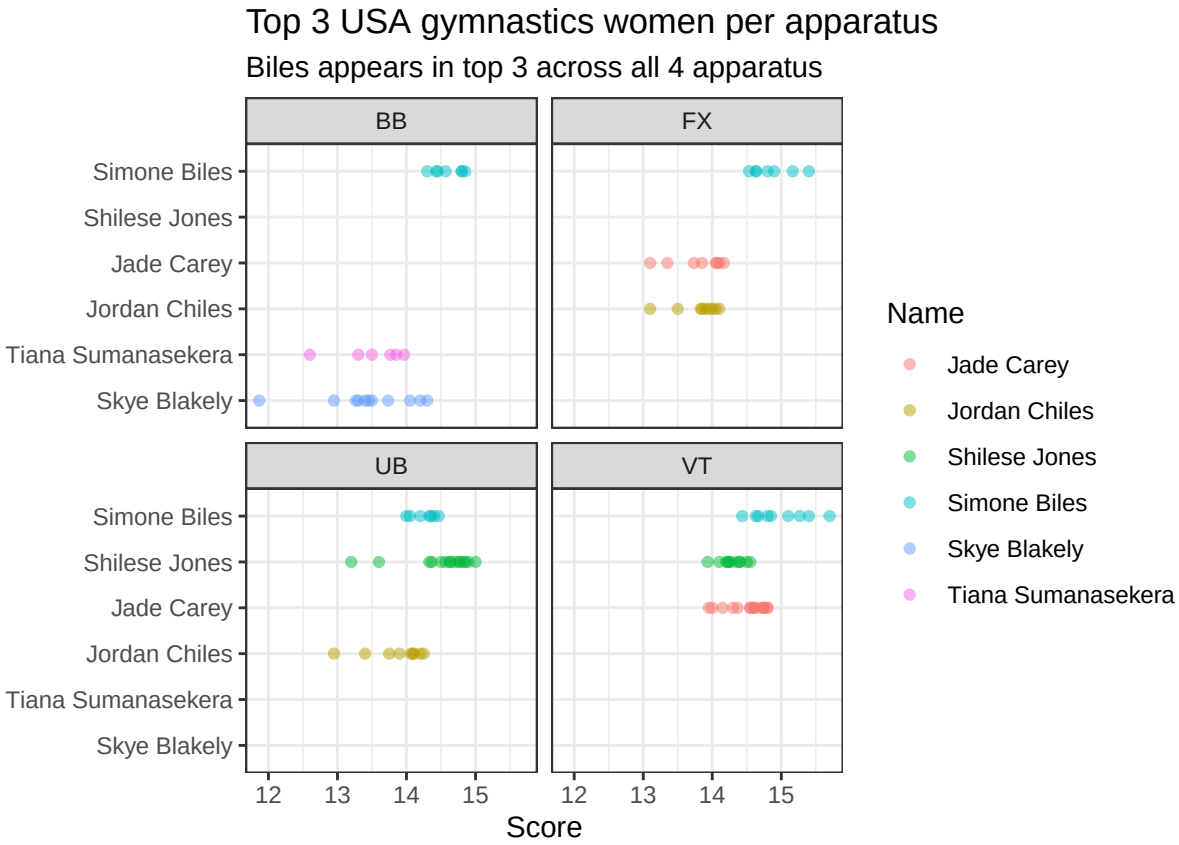
The data consisted of competition results from various competitors. Our categorical variables are FirstName and LastName, for the first and last names of the competitors, gender, country(of the competitor), date(date range of the competition), competition, round(either qualifier or final), location, and apparatus. Our numerical variables include rank(place in competition), D Score(Difficulty Score), E Score(Execution Score), Penalty, and Score. It is important to note that Score is simply the combination of D Score + E Score - Penalty.

The global data had 23891 observations, and the dataset, when filtered for the USA only, had 3362 observations. There were 12 features given in the dataset.

Exploratory Data Analysis

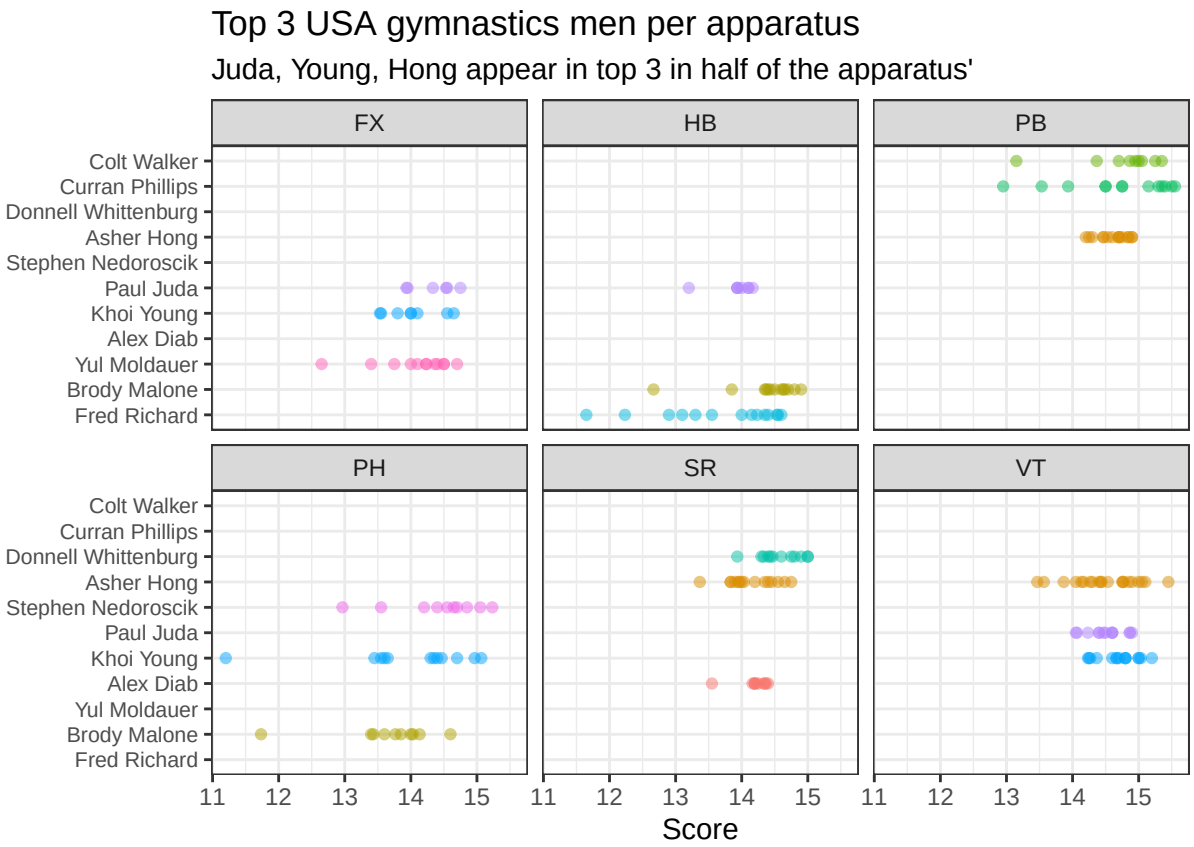
Our primary takeaways from our EDA (appendix A) were that the distribution of scores for each apparatus for men and women in the USA, and worldwide was roughly normal, with somewhat of a tail towards the left side/ left skew, and somewhat fat tails.

If our goal is to put forth the best lineup to maximize medal count, we explore who the top US candidates are in each apparatus and all-around. We graph dot plots of the top 3 US gymnasts in each event across all the events, using the mean score from their 2022-2023 data.



To clarify why some athletes appear but don't have score dots, that athlete was in the top 3 for another apparatus but not this one. We keep them in just to show the "depth" of US gymnastics. A small number of total athletes that appear in the top 3 across all apparatus signals that most elite athletes compete in, and are great at more than 1 event.

Taking a glance at the top US women, we can see that Simone Biles by far and away is the best USA gymnast. That is quite unsurprising, given that she has the most gymnastics medals in the world and has continued to perform at an elite level in all events. We can almost lock her in to make the Paris 2024 roster. Shilese Jones is the only gymnast have a higher mean score in an apparatus than Simone, specifically in uneven bars (UB). She's someone to keep an eye out for, especially considering her all-round Bronze medal in the World Championships in October 2023. There are only 6 women to place in the top 3 across 4 apparatus, a telling sign that the US womens team selections for Paris 2024 might be top heavy.



The US men's data is not as clear cut. There are a total of 11 gymnasts who place in the top 3 of at least one of the 6 events. Asher Hong, Khoi Young, Yul Moldauer place in the top 3 for 3 out of 6 events. Upon further investigation, Brody Malone places in the top 5 for 4 out of 6 events, and Fred Richard most recently won bronze in the 2023 World Championships in the all round. Those are athletes we might expect to be placed into our USA mens team.

Our EDA is important as it allows us to sanity check our decision making, and will provide a sturdy foundation to any USA team lineups.

Methodology

Breakdown of All Round + Specialists

Treating this as an optimization problem where we are maximizing for total medal count, we have developed nuanced heuristics based off analyzing past Olympic gymnastic event results to pick our top five-person lineups. Historically, we've seen that the USA, along with most other participating countries, send a combination of all-round gymnasts and event specialists in a five person lineup. Specifically, the USA Olympic Team Selection Procedure outlines that the top 2 overall all-round gymnasts will automatically qualify for the squad and then the committee will select the other 3 to complement them. A team can qualify a maximum of two individual gymnasts for each apparatus final as well as the all-around final. By this logic, it makes the most sense to send 2 all-around participants, while then having 3 specialists that would give that country the best shot at medaling in each individual apparatus. We wanted to mimic this approach, leading us to implement a two-tiered approach. First, we choose the top two overall all-round gymnasts, and then we select three more gymnasts to round out the team in an attempt to maximize medal count.

Selection Method

In selecting the two all-round gymnasts, we will choose the two candidates that have the simulated highest total score average across all the events they have participated in leading up to the 2024 Olympics. Specifically, we weighted each score by artificially determining its importance using competition date and "stage" and averaged the weighted scores across all the events a US gymnast has participated in. We then sum the average weighted scores in each apparatus to find a composite all-around score.

Clutch Factor and Recency Considerations

We chose to put weights on competition date and competition stage, specifically, because we believe a heavier weight should be placed on athletes with a history of performing better under pressure in the final rounds, and more emphasis on athletes which have great recent performances. This weighting is our attempt to measure the "clutch" factor and peak/prime performance of an individual athlete. The way our weights will work is we will duplicate rows of individual scores in the dataset given the assigned weight of that score. Competition date will be split into three categories by how recent the competition was: very recent ($<$ 6 months), somewhat recent (6 - 18 months), not recent (18+ months). A multiplier of 1 will be given to not recent competitions, a multiplier of 2 to somewhat recent competitions and a multiplier of 3 for very recent competitions. As for the competition stage, there are two stages of competition: qualifiers and finals. Qualifying round scores will receive a multiplier of 1 and final round scores will receive a multiplier of 2. For example, a row containing a score from a not recent competition and qualifying round will remain as one singular row. However, a row containing a qualifying round score, but from very recent, will be multiplied three times to get three rows containing that score. If that score was instead from the finals, it would be multiplied 2 x 3 = 6 times.

Taking the average of an individual athlete's scores after this duplication process will give us a mean expected score on each apparatus for that athlete. Still, we are aware that each gymnast will have variance in their scores for each apparatus and we wanted to take that into account. For each USA gymnast, we will also develop an individual total score distribution from their weighted scores to account for the variance for each individual on each apparatus. A density plot of all the scores after the duplication process will give us a distribution centered around the mean score for an athlete as well as a variance.

Obtaining Medaling Probabilities

Because the main goal of analysis is to determine the probability that an individual athlete medals, we also had to determine a way to get medal probabilities. We will accomplish this goal by determining medal thresholds, or thresholds necessary to achieve a certain medal. This will first require us to get a distribution of scores on individual apparatuses. We do this by creating a T-distribution from the top 24 individuals' weighted mean score. Of the 24 people, however, no more than 4 competitors from the same country will be included in this distribution as a maximum of 4 individuals from each country can compete in qualifying for individual apparatuses. From the EDA, we saw that the distribution of scores was roughly normal, with somewhat of a tail towards the left side of the distribution, and somewhat fat tails. This inspired us to take the top 24 people(the lower scoring competitors would be dropped), and from there create a T-distribution to sample from in our simulation, since we have normality and somewhat fat tails.

In order to find an individual's probability of earning a medal we will use a two step simulation and sampling process. We will simulate the following process 1,000 times. We first will sample from the distribution of the top 24 individuals 24 times. The top three scores of the 24 will represent the medal cutoffs with the 3rd highest sample score corresponding to a Bronze medal, the 2nd highest sample score corresponding to a Silver medal cutoff, and the highest sample score corresponding to a Gold medal cutoff. Because we are only worried about the placing of gymnasts from the USA, we will sample once from each of the top 8 USA gymnasts in each individual apparatus given their unique distributions. If that gymnast's sampled score surpasses one of the cutoffs they will have 'earned' that medal for the given simulation. For example, if when sampling from the 24 gymnasts distribution we get medal cutoffs of 14.5, 14.8, and 14.9 and gymnast A gets

a sampled score of 14.6 they will "earn" a Bronze medal. A sampled score of 14.81 would "earn" a Silver medal and a sampled score of 14.91 would "earn" a Gold medal. The only caveat is if multiple American gymnasts fall within the threshold for a certain medal. In that case the gymnasts would be ordered and the top score earns that medal and the next highest score would earn the next highest medal if it is available. We again simulate this two-tiered sampling approach 1,000 times to get each individual gymnast's probability of getting a specific medal in a specific apparatus. We will use this approach for the individual all-around as well, and the two men and two women with the highest medal probabilities will be automatically selected for our team. The assumptions for our simulation require an adequate amount of repeating sampling. 1000 is surely enough for each athlete, who on average has 10-20 observations. We also assume that past performance is somewhat indicative of future performance.

Optimization

From there we will use optimization to determine the 5-person lineup that maximizes expected medals earned including the team all-around medal. In determining the team all-around medal we will use a system identical to the way the Olympics will work in 2024. We will use many permutations of 3 specialists along with our two guaranteed all-around gymnasts and sample from every gymnast's individual distributions to get all-around scores. For each apparatus, the top three US gymnasts on the teams distributions will be sampled from (because only the top 3 gymnasts participate in the team all-around finals). The scores will be aggregated to get a total team all-around score. The same method of medal cutoffs will be used in the team all-around medal cutoffs, but using empirical team all-around score data to create a distribution to be sampled from. The 5-person team that maximizes total medals earned will be the team we recommend the USA send to Paris for the 2024 Olympics.

Results

Discussion

RESULT SUMMARY GOES HERE

Our goal was to recommend a team of 5 women and 5 men for the Olympic Games in Paris 2024, with an expected medal count for each team. We accomplish this by recommending Simone Biles, Skye Blakely, Shilese Jones, Jade Carey, and Zoe Miller for the women's team, and Asher Hong, Paul Juda, Yul Moldauer, Fred Richard, and Khoi Young for the men's team. Our predicted medal counts are 4.316 individual medals for the women's team and XXX for the men's team. We accomplished our goal of selecting atheletes by first selecting 2 all-around athletes, and then finding 3 specialists to complement them, while adjusting for 'clutch' factor and recency by duplicating observations based on final vs qualifying rounds, and recency of competition. We then account for individual variance by creating a T-distribution for each competitor for each apparatus, and sampling from that distrbution 1000 times. Lastly, we then find the medal thresholds using repeated sampling, and use those thresholds to find predicted medal counts.

The 'eye' test confirms some of the choices made by our model. The choices of Simone Biles, Skye Blakely, Shielese Jones, Joscelyn Roberson, and Leanne Wong align with the choices of the US team in the most recent World Championships where the US women won a gold medal. Both Jones and Biles also medaled in the individual all-around event justifying them being chosen as the two guaranteed all-around competitors. On the men's side the model's choices do align with the US men's team that won bronze in the most recent World Championship. All of Asher Hong, Paul Juda, Yul Moldauer, Fred Richard, and Khoi Young are both on our suggested Olympic team and the World Championship team. Hong and Richard were the selected all-around competitors also justifying our model's choices. Furthermore, an independent level 10 gymnast committed to a D1 university chose a team identical to that of the US women's world championship squad leading us to believe the eye test gives ground for our choices.

Still, there is important information that the USA olympic committee as well as some of the general public have that our model does not. For example, Konnor McClain may look like a strong choice, but she left

Elite (which is a program designed to be a pathway to the USA national team) in 2023 in favor of college. Although she plans to continue training for the Olympics, her non-Elite status likely keeps her from being in the USA olympic committee's player pool. A similar situation exists for Sunisa Lee, the 2020 Olympic all-around champion. Although her weighted scores are quite high and the model would select her as one of the all-around locks, she has not competed Elite since her enrollment in college in 2022 and a recent injury in which she gained over 50 pounds likely keeps her from making the USA national team. Injury concerns will also keep Brody Malone off the team on the men's side. Although he eyes a spring 2024 comeback from a knee injury suffered in the spring of 2023, he will likely not be ready or in form for the Olympics.

Limitations

There are limitations in the data that prevent us from creating as complete an analysis as possible. First, the data aggregates scoring data from a number of different competitions. Each of the different competitions lists competitors' names in a different way so for some individuals it was difficult to assign all scores to them given the multiple unique ways their names were inserted in the dataset. There were also some NAs in the dataset. Another issue with the dataset was the lack of a variable for age which could have a huge effect on future performance. If an individual has surpassed peak performance age, we would expect their performance to decline which may not be seen in the dataset. The opposite is true goes for young athletes that have not yet hit their prime.

There were also several limitations in our statistical methods. Firstly, we arbitrarily assigned weights to each of the decided 'cutoffs' it is more than possible that our weightings could result in the selection of a different team than if no weights were used. However, the weights used were for good reason given how recent performance affects future performance and our desire to pick more 'clutch' individuals. Another potential limitation in our method is when determining projected scores for athletes that have not competed in some time. Take Sunisa Lee as an example. Due to competing in college and injuries, Lee has very few scores since her Olympic performance in 2021. Because those performances account for so many of her competitions, her lower recent scores are severely overshadowed by her high, older scores. Even with weights her older scores account for a much greater proportion of her data. Thus in our methodology her projected score is unreasonably high for what recent scores suggest. This could remain true for many other athletes where their projected score does not resemble what their recent form might be. The final issue with our methodology and sampling in general is because there is technically no maximum score a gymnast could receive due to more challenging tricks being completed in every successive competition the sample could possibly get too high scores as there is no maximum cutoff in the distributions.

Some alternatives for methodology could include using a Kernal Density Estimation method to estimate the distribution, and then sampling, or obtaining all possible combinations of 5 athletes and then running an olympics simulation. We could also introduce more factors to weighting such as the age of the competitor, skill level of the competition, etc. For future work, and potentially in the final draft, we could also include a sensitivity analysis on our weights, since they were selected somewhat arbitarily.

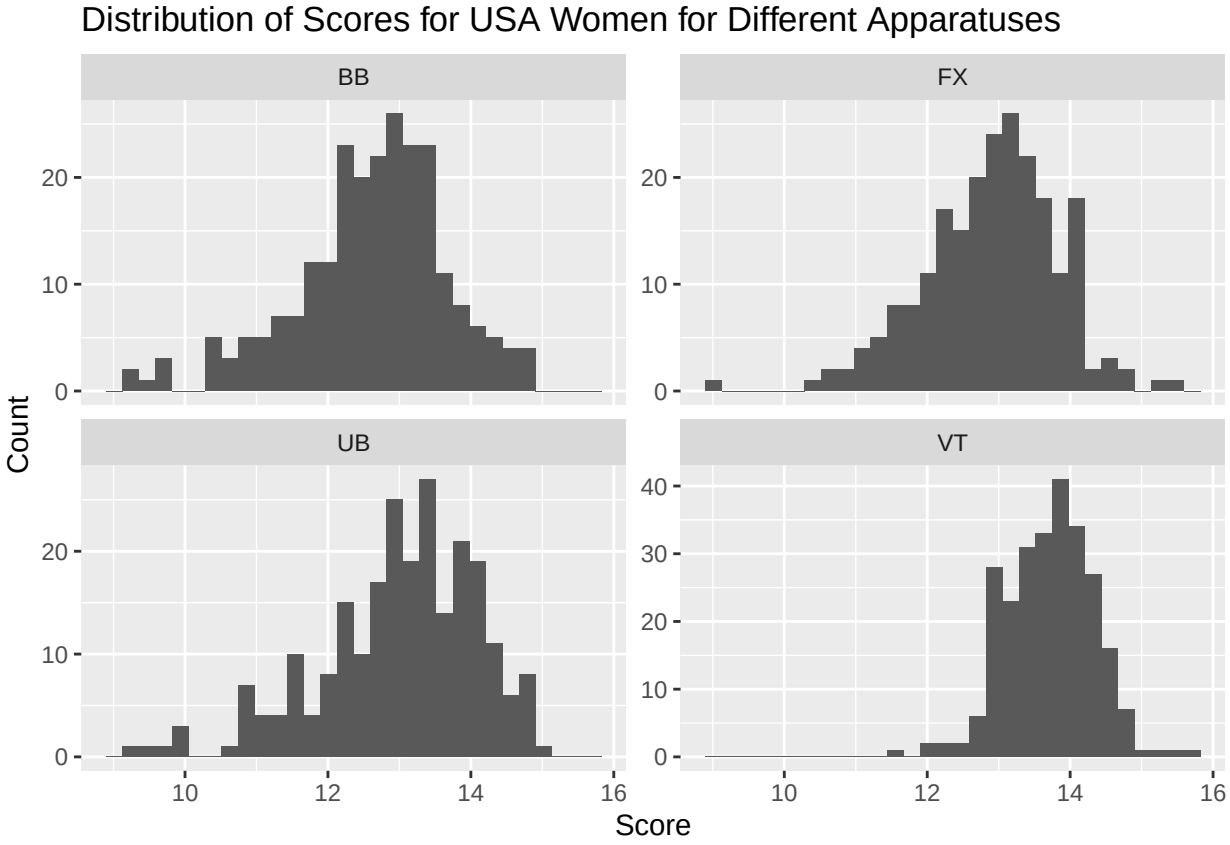
Implications

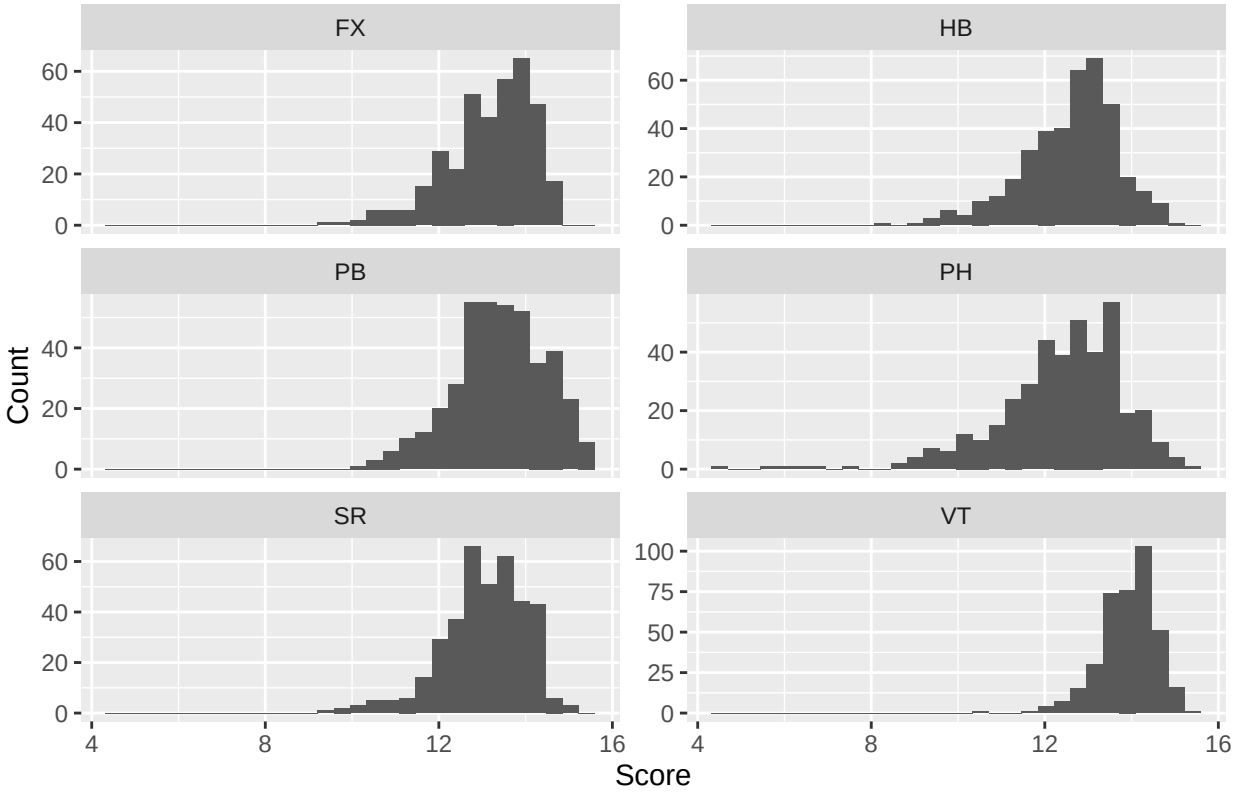
As stated earlier in the paper, several American viewers are concerned with the success of the USA Olympic gymnastic team. As the Olympics has become synonymous with national pride, success in Olympic gymnastics is vital to maintaining a sense of American pride. Given the high expectations laid out for the USA due to past success in the Olympics and as other countries have become more competitive on the gymnastics scene, it has become evermore important to select the team that gives the USA the best chance at being successful in Paris 2024. The statistical analysis done in this paper gives an objective way to determine that "best" team. As more advanced data is measured and the intersection between sports and data analytics continue to grow, there is much incentive to use data to create a competitive advantage over the rest of the field. While there are severe limitations to our analytic process, our optimization strategy combined with the more classical "eye test" method should give the USA Olympic committee the necessary tools to select the best 5 person gymnastics team to send to the 2024 Olympics.

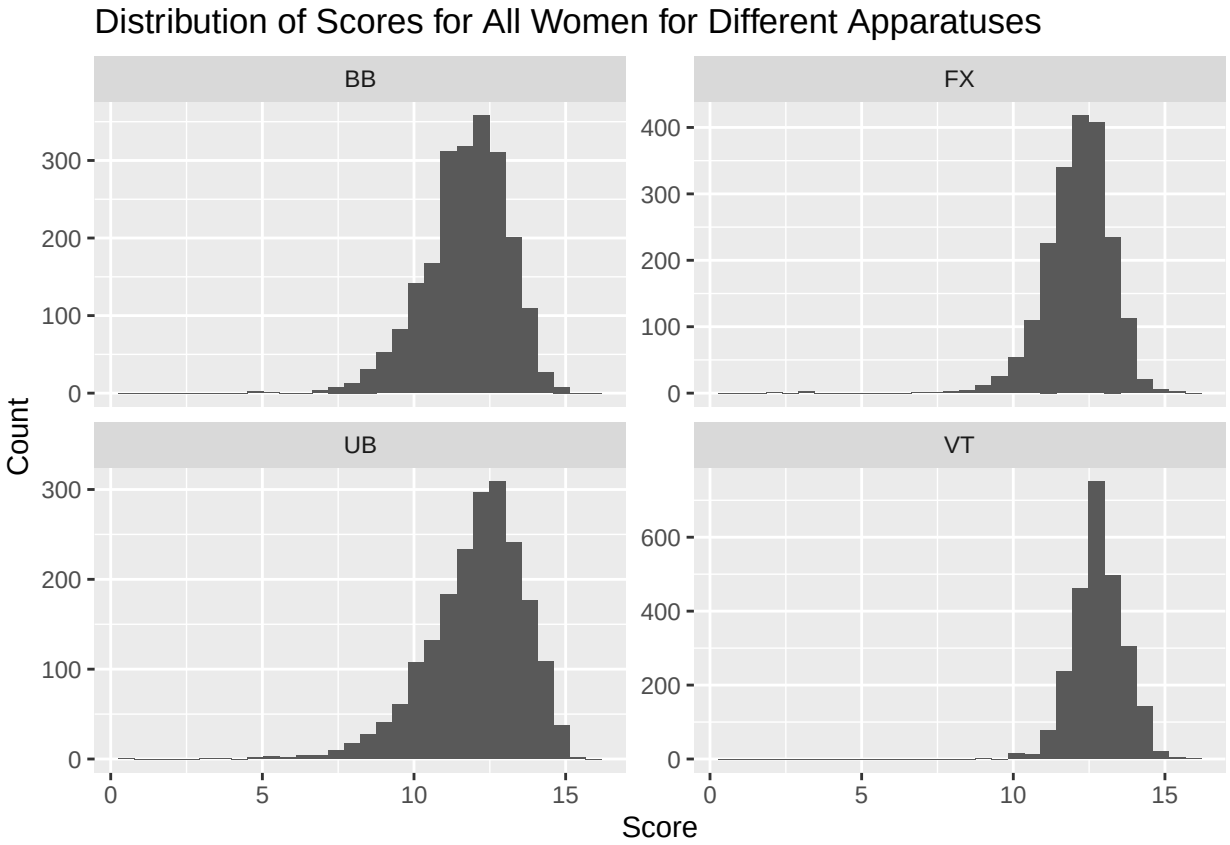
Future Research

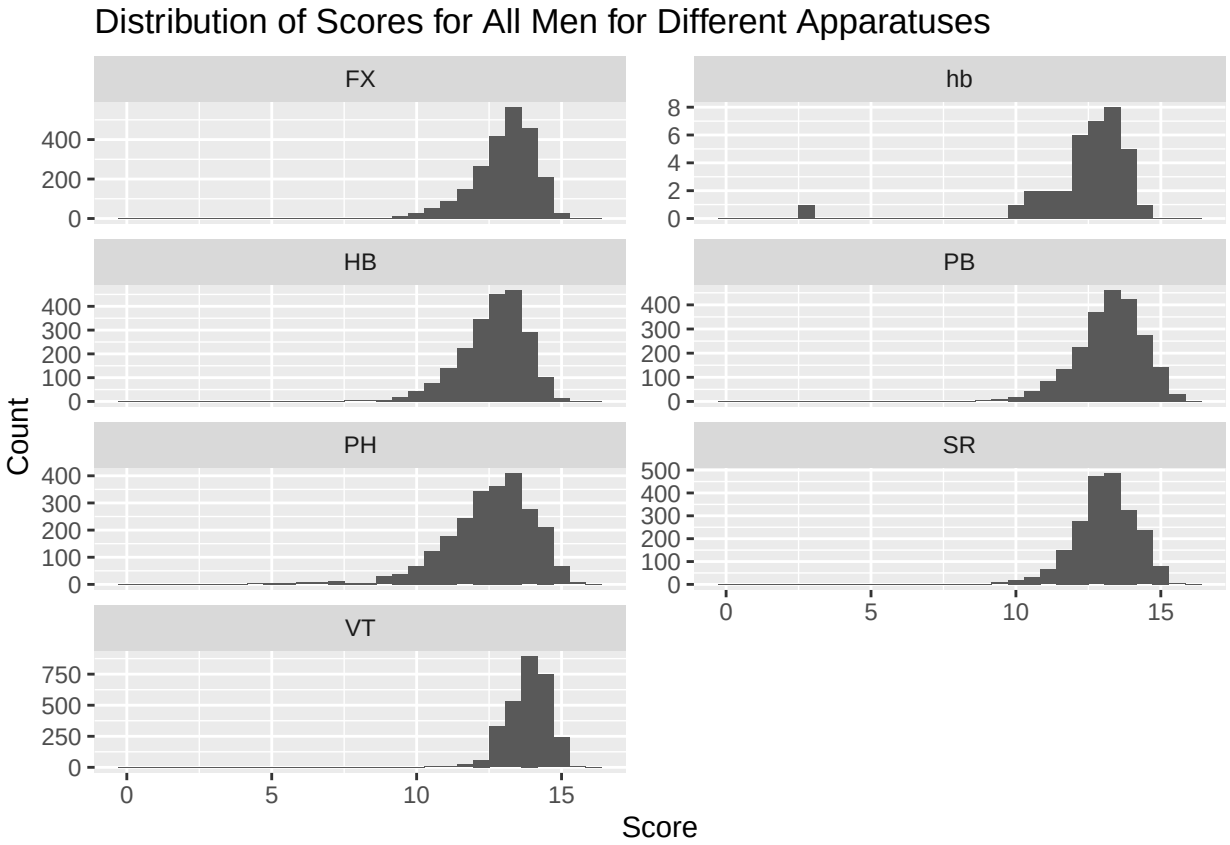
The potential for future research is endless. We hope to begin by looking into different ways to optimize lineups based on differing definitions of success. Here we optimize by total medal count, but perhaps the US wants to optimize by total gold medals, or by individual apparatus medals won, or by individual all-around gold. Our future research would allow the ideal team to change based on this definition of success. We also would like to look into optimizing lineups based on different criteria. That is, if a person wants to put a higher weighting on having the most consistent gymnasts or on the gymnast with highest potential than the model would be able to put out a lineup based on that criteria. We could also look into more data related to the olympics and gymnsatics, and maybe interview a couple of professional gymnasts or their coaches to get further insight into Olympics level gymnastics. Finally, in the future we hope to use the tools created during this data analysis to potentially predict or scout rising stars competing on the USA Elite pathway to keep an eye on for future Olympic cycles.

Appendix A


Distribution of Scores for USA Women for Different Apparatuses
BB
FX
UB
VT
Count
Score


Distribution of Scores for USA Men for Different Apparatuses
FX
HB
PB
PH
SR
VT
Count
Score


Distribution of Scores for All Women for Different Apparatuses
BB
FX
UB
VT
Count
Score
0
100
200
300
400
600
0
5
10
15



Works Cited

Gunston, Jo. "Judging the Judges – How Statistical Analysis Evaluates Fairness And . . . " Olympics.Com, 19 Oct. 2023, olympics.com/en/news/how-statistical-analysis-evaluates-fairness-accuracy-gymnastics. Accessed 06 Nov. 2023.

Most popular sports in the Summer Olympics US 2021. (n.d.). Statista. https://www.statista.com/statistics/1245746/summer-olympics-most-followed-sports-us/. Accessed 20 Nov. 2023.

Olympic Summer Games: global broadcast audience. (n.d.). Statista. https://www.statista.com/statistics/280502/total-number-of-tv-viewers-of-olympic-summer-games-worldwide/#:~:text=Between%202012%20and%202020%2C%. Accessed 20 Nov. 2023.

Brody Malone eyes spring 2024 return to gymnastics from leg surgeries. 17 May 2023. NBC Sports. https://www.nbcsports.com/olympics/news/brody-malone-gymnastics-injury-comeback. Accessed 20 Nov. 2023.

Olympic champ Sunisa Lee gained 45 pounds due to kidney issue. "It was so scary." 17 Nov. 2023. USA TODAY. https://www.usatoday.com/story/sports/olympics/2023/11/17/sunisa-lee-olympic-champion-kidney-health-paris/71616483007/. Accessed 20 Nov. 2023.

Duffy, P. Konnor McClain enrolling at LSU this fall with hopes of balancing NCAA and elite - Gymnastics Now. Gymnastics-Now.com. 13 Jul. 2023. https://gymnastics-now.com/konnor-mcclain-enrolling-at-lsu-this-fall-with-hopes-of-balancing-ncaa-and-elite/. Accessed 20 Nov. 2023.