Case Study 2: Paris 2024 Team USA Olympics Gymnastics Team

Aimi Wen, Katie Nash, Nidhi Dhupati, Wendy Shi

Introduction

Background Information

The Olympic Games are an international multi-sport event held every four years, and are largely considered to be the most prestigious sporting competition in the world. These Games are therefore a popular spectator event across the world, with an average viewership of more than 15 million people in past years [1]. The 2024 Olympic Games will be held in Paris, France, and will include both Men's and Women's Artistic Gymnastics events. These Gymnastics events in particular are highly anticipated: as of March 2021, Artistic Gymnastics was the most followed Summer Olympics sport [2]. Winning a medal in an Olympic gymnastics event is an honor for both the individual athlete and for the country that athlete represents; the medal brings a sense of national pride to citizens across the country and promotes the global recognition of the winning country. The USA is known for its men and women gymnastics teams, which have consistently earned medals in past Olympic games. Because of the high honor and importance of upholding the legacy of the United States in these competitions, it is important to carefully select the men and women who will represent our country as Team USA. The importance of and interest surrounding the Olympics provides a rich foundation for statistical analysis of past athlete performances and an interesting platform for making predictions about who best represents Team USA in future Olympics. There have already been analyses, published in major news outlets, about who Team USA will consist of for non-gymnastics sports in 2024 [3]. The objective of this project is to extend this analysis to the Artistic Gymnastics events, and to select the 5 athletes for each of the men's and women's events that will optimize performance for Team USA. The format of the Olympic Artistic Gymnastics event includes a qualifying round and final round. Scores from qualifying rounds do not impact medal determinations, but rather decide which athletes advance to the finals. The final round, which is what determines medal placements, consists of team all-around, individual all-around, and individual apparatus events. Men's opportunities for medals are: 1) team all-around, 2) individual all-around, 3) floor exercise, 4) pommel horse, 5) still rings, 6) vault, 7) parallel bars, and 8) high bar. Women's opportunities for medals are: 1) team all-around, 2) individual all-around, 3) vault, 4) uneven bars, 5) balance beam, and 6) floor exercise. The athletes representing Team USA will compete in a pool of 96 men and 96 women, which will include teams of five from other countries as well as the option for individual entries for countries without full team qualifications.

Our hypothesis is:

Who are the 5 men and 5 women who can make Team USA achieve the highest level of success at the Paris 2024 Artistic Gymnastics Event?

Data Description

The dataset was given to us by UConn Sports Analytics Symposium (UCSAS). Major Gymnastics competitions from 2017-2023 were included in the dataset. These competitions were split into two separate datasets by the seasons leading up to the 2021 Tokyo Games and the seasons leading up to the 2024 Paris Games (the 2022 and 2023 seasons). Competitions were included from around the globe, including world competitions, non-world competitions that included the United States as a competitor, and non-world competitions that did not include the United States as a competitor. The competition data included scoring information, athlete

names, athlete demographic information, athlete scores, and competition details. Each entry in the dataset is one event from one athlete, which includes the country affiliation of the athlete, the competition and competition location, the difficulty, execution, penalty, and overall scores, the apparatus that the athlete competed on, and their ranking. Additionally, the countries who qualified as a team, which are the countries that will be allowed to compete in the Team All-Around event for the men's and women's Olympic events, were provided. The data set was provided to us as part of the UCSAS 2024 USOPC Data Challenge, and information for verification of athlete names and country affiliation was conducted by using Wikipedia [3] alongside the USA Gymnastics website [4]. The dataset we used for analysis was last synced with the central repository on November 2nd.

Methodology

Data Cleaning

Since our goal is to optimize performance in the 2024 Olympics, we chose to only include competition data from 2022 and 2023, while excluding data from the 2017-2021 season. This reflects the unprecedented global pandemic that even caused the Tokyo Olympics to be delayed, the scoring differences before 2021 versus now, and athletes who are most likely still competing. We assumed that competition data from the two most recent years would be the most useful in understanding 2024 Olympic performance.

From the original 2022-2023 competition dataset, we standardized variables and added columns. We standardized each athlete's name based on the first instance it appeared and created a column that included the full name of the athlete. Additionally, we made sure the country codes were consistent with International Olympic Committee Codes (for example, Scotland competes as part of the United Kingdom in the Olympics). We also separated the competition data into two datasets of men's and women's competitions, created a variable to track the order in the competitions that occurred by their dates, and added a competition scope variable to aid with data visualizations and athlete selection in the analysis approach. For our Team All-Around analysis approach, which is written in more detail below, we also created another dataframe, which accounted for new information on countries who have already qualified. Athletes who represent a qualifying country kept the country's label, but athletes who were not part of a qualifying country were labeled under "Other" for the country they represent. Taking a preliminary look at how each country's athletes have been performing recently, Figure 1 shows that each country has a broad range of performances. This brought us to create an athlete selection criteria to narrow down our possible pool of athletes for our analysis approach, which is detailed below.

Analysis Approach

In order to study which athletes could enable Team USA to achieve maximum success at the Paris Olympics, we decided to model possible Olympic scenarios by simulating the medal outcomes of different versions of Team USA (different combinations of 5 athletes) competing with teams from other countries. One iteration of our simulation would model one Olympic run for either the women or the men gymnasts.

Our metric of success was a weighted medal count: we assigned 3 points to a Gold medal, 2 points to a Silver medal, and 1 point to a Bronze medal. This reflects our belief in the relative prestige of each medal and will allow the highest score to reflect our interpretation of the highest level of success.

We conducted a separate simulation process for the Men's Artistic Gymnastics Event and the Women's Artistic Gymnastics Event. The methodology can be broken down into the following subsections: athlete selection, creating score distributions, team selection, Team All-Around event simulation, individual event simulation, and finding the top-performing USA teams.

Athlete Selection Criteria

To create a group of "elite athletes" that might compete at the Olympics, we included athletes who placed in the top 10 at a World competition or in the top 3 at a non-world competition. This reflects two assumptions. One, the Olympic competition would only consist of athletes who are already competing at the highest level,
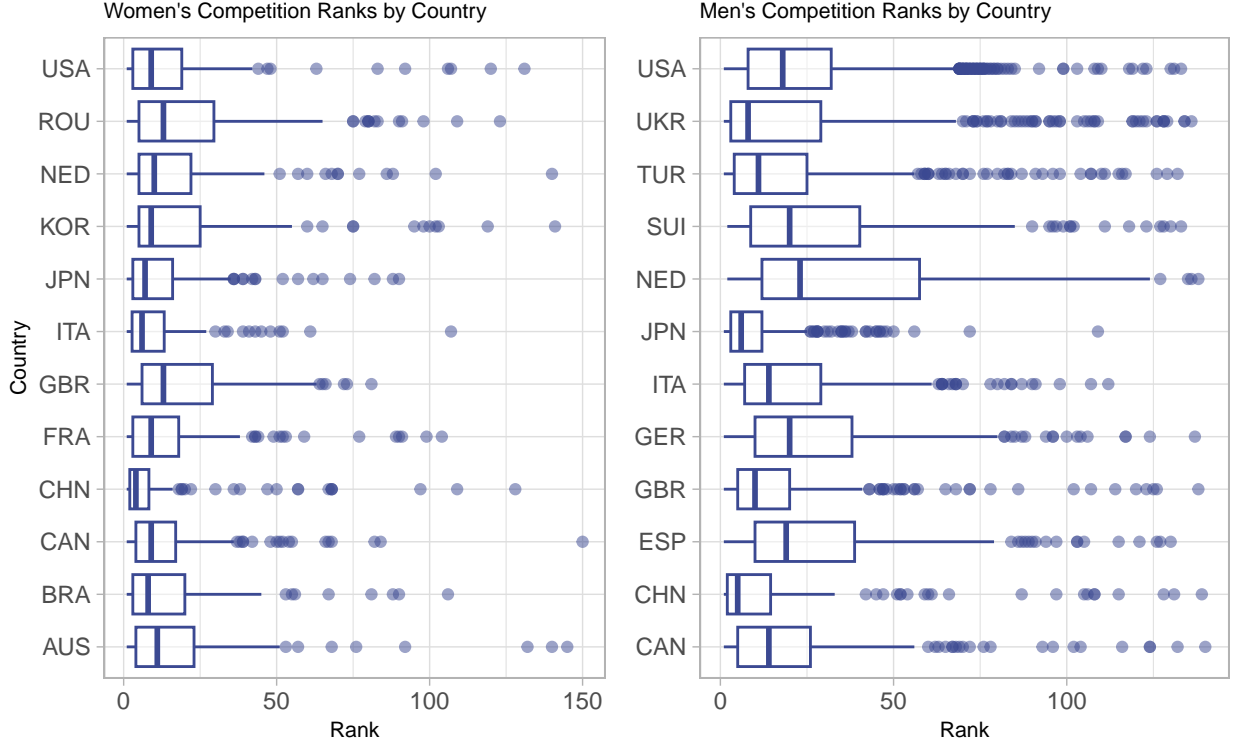
Figure 1: Each Athlete's Competition Ranks for Olympic Qualified Countries

thus we do this initial filtering. Two, world competitions are more competitive than non-world competitions so we allow top 10 vs top 3.

Creating Score Distributions

The next step is to simulate the score that each athlete receives on each apparatus. We created a score distribution for each athlete and apparatus based on the observed scores they received in past competitions on this apparatus, while adding noise and allowing higher or lower scores than the range of scores they have received in the past. We used the process of kernel density estimation (KDE), which provides a continuous representation of the score distribution estimated from their past scores. Note, we treated Vault_1 and Vault_2 as separate distributions because in real life the athletes perform different vaults for one and two. Mathematical details about the development of our KDE are included in the appendix.

Although we are only simulating scores for the final rounds, since those are the medal-earning events, we decided to include athlete scores from qualification and final rounds in our creation of the KDEs in order to have more data points to estimate from. Figure 2 shows the qualification and final round scores by apparatus. The distribution of qualification scores is slightly lower than the distribution of final round scores, which is consistent with the observation that the athletes who have low scores in the qualification rounds do not move on to the final rounds. Although there is still the possibility of a difference in qualification and final round scores for the same athlete, the inclusion of qualification rounds provides more data about each athlete's performance distribution, so we will include information from both rounds in our simulation.

Additionally, we only created the KDEs based on the total score for the athlete, instead of separating the scores by difficulty score, execution score, and penalty score. Since the total score is difficulty score + execution score - penalty score, it encompasses all of these sub-scores. Appendix Figure 1 additionally demonstrates that most penalty scores are near 0.
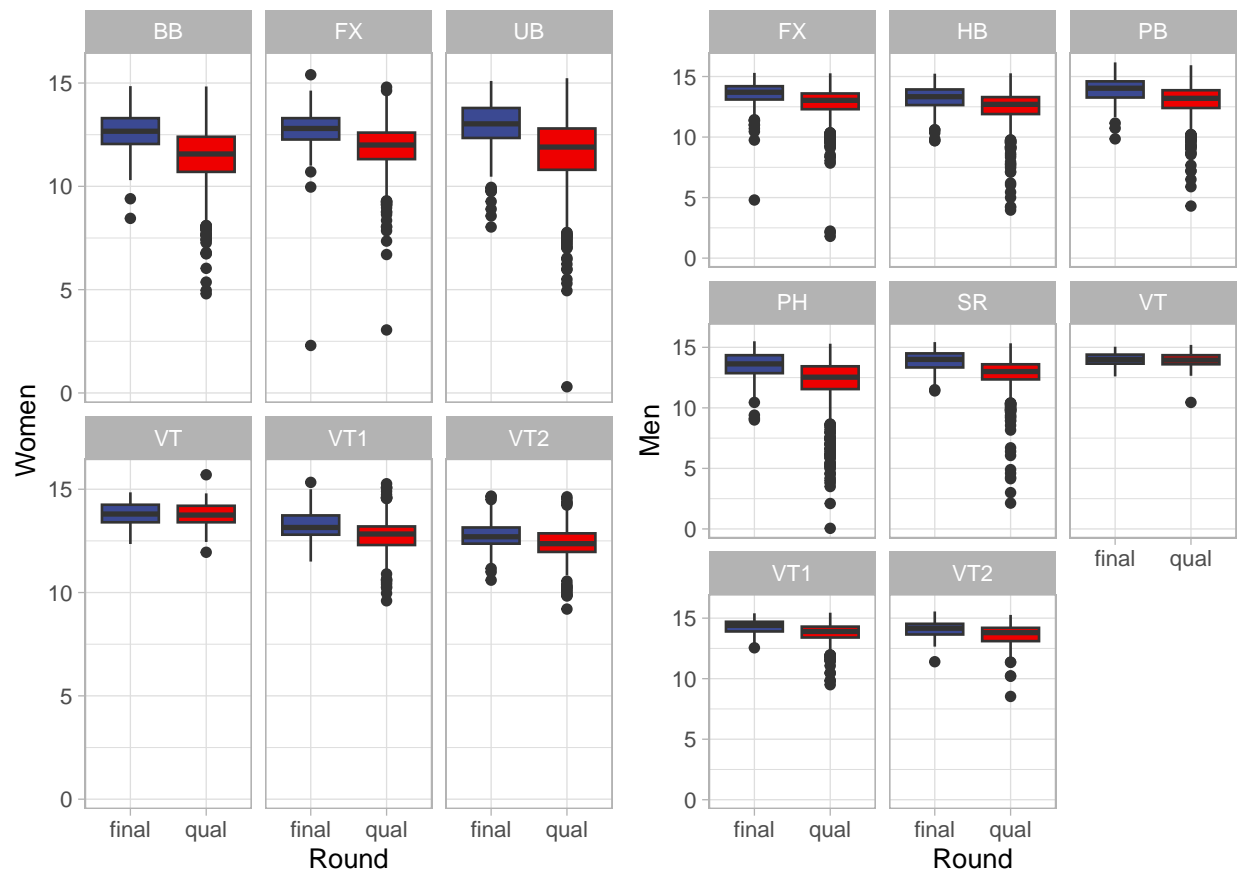
Figure 2: Distribution of the Score by Round and Apparatus

Team Selection

In order to simulate Olympics outcomes, we had to select a team of five gymnasts for each qualifying country in the Olympics, for each simulation round. For each qualifying country, we used weighted random sampling to select 5 gymnasts. Weighted random sampling was chosen rather than simple random sampling in order for the highest performing athletes to be sampled more often. This allows for the formation of teams to be more realistic, reflecting the assumption that each country would only select their top performers to represent their country in the Olympics. The weighting was based on the median of each athlete's kernel density for each apparatus, and reflects the observed past score and skill level of the athletes.

Mathematical details about the weighting process are included in the appendix, and further explanation and justification is included as well.

Team All-Around Simulation

To simulate the team all-around competition, three gymnasts would need to be selected to compete for each apparatus from the five gymnasts on each team. But to first choose which 3 athletes, we simulate scores for all of the 5 athletes across all events by sampling scores from the appropriate kernel density and choosing the athletes who scored the 3 highest scores on each apparatus to represent their country for that apparatus in the Team All Around. We believe this appropriately mimics a coach's selection criteria in real life.

Once every qualifying country had three athletes selected for each apparatus, the scores for the actual event with these athletes were again simulated from their kernel densities. Then, the scores for all apparatuses were summed by country to create the final scores for the Team All-Around event. The top 3 teams were given the appropriate medal counts.

Individual Events Simulation

Unlike the Team All-Around, individual events allow for athletes from countries that did not qualify as part of a team. Therefore, when simulating individual events, our dataset included five athletes who were selected (via the weighted random selection process) for each qualifying country, as well as all of the athletes who passed our athlete selection criteria but did not belong to a qualifying country.

For each individual event and athlete, we assigned scores by randomly sampling using the appropriate athlete and apparatus KDE score density. For individual events where multiple scores can be recorded (the individual all-around and vault events), we summed each score together to represent the final score for each event. The top 3 athletes for each event were given the appropriate medal counts.

Finding the Top USA Teams

We ran 3000 simulations for the Men's Artistic Gymnastics Event, and a separate 3000 simulations for the Women's Artistic Gymnastics Event. Each simulation included a separate selection process for the five athletes who would represent each qualifying country.

After running these simulations, we selected the top 10 simulations where the USA scored the highest weighted medal count for each gender. We extracted the athletes that represented Team USA in each of these top 10 simulations. Since the USA performed highly in each of these simulations, we decided to run another 300 simulations to narrow down these top 10 teams. With these additional 300 simulations, we randomly selected the athletes in other qualifying countries but kept the same athletes in every simulation for the United States.

These additional simulations address the potential variability in our simulated scores, as well as address the possibility that Team USA may have just scored highly in one of the "top 10" simulations due to poor performances of other countries.

After running the 300 simulations for each of the top 10 teams (so 3000 simulations total), we tracked how many times the United States had the highest weighted medal count of all countries for each team.

We chose the team that had the maximum weighted medal count of all countries for the highest percentage of times as our final suggested men's Team USA.

For the women's team, we repeated the same process up to this point. However, for women's gymnastics, there ended up being multiple teams tied with the same highest percentage.

Breaking the Tie for Women

There were five teams who tied amongst the women, which are included in the results section. We took the teams that tied, found the athletes who were chosen the most frequently from these teams, and created the second iteration of high-performing teams with these athletes. Additional details and the end result of these simulations are included in the results section.

Results

Men's Artistic Gymnastics Event

Of the 3000 simulations for both men and women, the 10 simulation with the highest weighted medal score for the United States were separated. Then, 300 more simulations were run with these 10 teams fixed, as described in the methodology. For each of these 10 teams, the percentage of simulations where the United States had the highest weighted medal count is included in the table below. The team where the USA won the highest weighted medal count for the largest percentage of the 300 simulations was taken to be the Men's Team USA.

Table 1: Men's Artistic Gymnastics - Simulations for Possible Top Teams

Teams 1-5				
Team 1	**Team 2**	**Team 3**	**Team 4**	**Team 5**
Athletes				
Vitaliy Guimaraes	Asher Hong	Riley Loos	Vitaliy Guimaraes	Asher Hong
Asher Hong	Michael Jaroh	Yul Moldauer	Asher Hong	Jose Lopez
Brody Malone	Paul Juda	Frederick Richard	Paul Juda	Brody Malone
Yul Moldauer	Donnell Whittenburg	Shane Wiskus	Ian Lasic	Yul Moldauer
Donnell Whittenburg	Khoi Young	Khoi Young	Khoi Young	Khoi Young
Percentage of Simulations With USA Having Highest Weighted Medal Count				
13.7%	*10.0%*	*6.7%*	*9.3%*	*15.3%*
Teams 6-10				
Team 6	**Team 7**	**Team 8**	**Team 9**	**Team 10**
Athletes				
Cameron Bock	Asher Hong	Taylor Burkhart	Raydel Gamboa	Vitaliy Guimaraes
Matt Cormier	Patrick Hoopes	Vitaliy Guimaraes	Vitaliy Guimaraes	Asher Hong
Vitaliy Guimaraes	Paul Juda	Brody Malone	Paul Juda	Paul Juda
Paul Juda	Brody Malone	Yul Moldauer	Riley Loos	Curran Phillips
Frederick Richard	Colt Walker	Khoi Young	Khoi Young	Khoi Young
Percentage of Simulations With USA Having the Highest Weighted Medal Count				
6.7%	*15.7%*	*8.7%*	*3.3%*	*13.7%*

As seen in the table above, Team USA is taken to be Asher Hong, Patrick Hoopes, Paul Juda, Brody Malone, and Colt Walker (Team 7 above). This Team USA won the highest weighted medal count of any country is 15.7% of the second round simulations.

Since this Team 7 was identified as the men's Team USA, we have included a table below to demonstrate a potential outcome of the Olympics with this team. This table portrays the *best* performance of Team 7 in the 300 simulations that were run. It is important to understand the limitations of showcasing only the

best score; the simulation process inherently included randomness and could, although at a low probability, include an unreasonable high score. The outcomes of each other simulation are included in the raw code.

Table 2: Men's Artistic Gymnastics - Best Simulated Outcome for Chosen Team USA

Event	Gold	Silver	Bronze
Team All-Around	China	USA	Italy
Individual All-Around	USA (Brody Malone)	USA (Asher Hong)	China (Xingyu Lan)
Vault	Great Britain (Jake Jarman)	Philippines (Carlos Yulo)	Ukraine (Nazar Chepurnyi)
Still Rings	USA (Asher Hong)	Greece (Eleftherios Petrounias)	China (Xingyu Lan)
Pommel Horse	USA (Patrick Hoopes)	Kazakhstan (Nariman Kurbanov)	Canada (Zachary Clay)
Parallel Bars	Philippines (Carlos Yulo)	Germany (Lukas Dauser)	Japan (Shinnosuke Oka)
High Bar	China (Hao Tian)	Cyprus (Marios Georgiou)	Japan (Shinnosuke Oka)
Floor Exercise	USA (Brody Malone)	Cuba (Alejandro De)	Philippines (Carlos Yulo)

Team USA Athletes: Brody Malone, Asher Hong, Patrick Hoopes, Paul Juda, Colt Walker
Team USA Medal Count: 6
Team USA Weighted Medal Count: 16 (where Gold = 3, Silver = 2, Bronze = 1)
Team USA Medal Breakdown: 4 Gold, 2 Silver, 0 Bronze

Women's Artistic Gymnastics Event

The same process above was repeated for the women's event, and the results are included in the table below, with the names of the athletes and the percentage of simulations where the United States had the highest weighted medal count.

Table 3: Women's Artistic Gymnastics - Simulations for Possible Top Teams

Teams 1-5				
Team 1	**Team 2**	**Team 3**	**Team 4**	**Team 5**
Athletes				
Simone Biles	Simone Biles	Simone Biles	Simone Biles	Simone Biles
Charlotte Booth	Jade Carey	Jade Carey	Jade Carey	Jade Carey
eMjae Frazier	eMjae Frazier	Karis German	Kayla DiCello	Jordan Chiles
Shilese Jones	Kaliya Lincoln	Shilese Jones	eMjae Frazier	Levi Jung
Leanne Wong	Zoe Miller	Konnor McClain	Konnor McClain	Konnor McClain
Percentage of Simulations With USA Having the Highest Weighted Medal Count				
98.7%	*99.0%*	*99.0%*	*99.3%*	*99.7%*
Teams 6-10				
Team 6	**Team 7**	**Team 8**	**Team 9**	**Team 10**
Athletes				
Simone Biles	Simone Biles	Simone Biles	Simone Biles	Simone Biles
eMjae Frazier	Jordan Chiles	Addison Fatta	Jordan Chiles	Charlotte Booth
Shilese Jones	eMjae Frazier	Nola Matthews	Silese Jones	Jade Carey
Konnor McClain	Karis German	Konnor McClain	Katelyn Jong	Addison Fatta
Elle Mueller	Konnor McClain	Elle Mueller	Marissa Neal	Konnor McClain
Percentage of Simulations With USA Having the Highest Weighted Medal Count				
99.7%	*99.7%*	*98.7%*	*99.7%*	*99.7%*

As seen above, Team 5, 6, 7, 9, and 10 are all tied for the percentage of times the USA had the highest weighted medal count (of 99.7%). To break the tie, the number of times each individual athlete appeared in one of the five teams with a 99.7% highest weighted medal count percentage was counted. Simone Biles appeared 5 times in a top-performing team, Konnor McClain appeared 4 times, and Jordan Chiles appeared 3 times. These three were chosen for Team USA. eMjae Frazier, Jade Carey, and Shilese Jones were all chosen 2 times (with the rest of the names appearing here

only chosen 1 time). Because Team USA can only consist of five athletes, we ran an additional 300 simulations with every combination of Simone, Konnor, and Jordan, and two of the three athletes that appeared 2 times.

In these additional simulations, the USA won the highest weighted medal count the same percentage of times with each team (99.7% of the time). Therefore, after confirming that these combinations of teams had a tied percentage in a second round of "top simulations", we decided to choose the Team USA that had the highest weighted medal count, on average, from these 300 simulations. The average weighted medal count is included in the table below, along with the athletes who achieved these counts:

Table 4: Women's Artistic Gymnastics - Simulations for Tied Top Teams

Team 1	Team 2	Team 3
Athletes		
Simone Biles	Simone Biles	Simone Biles
Konnor McClain	Konnor McClain	Konnor McClain
Jordan Chiles	Jordan Chiles.	Jordan Chiles
eMjae Frazier	eMjae Frazier	Jade Carey
Jade Carey	Shilese Jones	Shilese Jones
Percentage of Simulations With USA Having the Highest Weighted Medal Count		
99.7%	*99.7%*	*99.7%*
Average Weighted Medal Count		
16.8	*17.5*	*18.4*

Since Simone Biles, Konnor McClain, Jordan Chiles, Jade Carey, and Shilese Jones had the highest average weighted medal count in these last round of simulations, they were chosen as the women's Team USA.

Since this Team 3 was identified as the women's Team USA, we have included a table below to demonstrate the *best* performance of Team 3 in the 300 simulations that were run in the last iteration of simulations. It is again important to note the limitations of showcasing only the best score, although,in this case, the women's Team USA won the highest weighted medal count in 99.7% of the simulations (though not always by this wide of a margin).

Table 5: Women's Artistic Gymnastics - Best Simulated Outcome for Team USA

Event	Gold	Silver	Bronze
Team All-Around	USA	Brazil	China
Individual All-Around	USA (Simone Biles)	USA (Shilese Jones)	Brazil (Rebeca Andrade)
Balance Beam	USA (Simone Biles)	USA (Konnor McClain)	China (Ran Wu)
Vault	USA (Simone Biles)	Brazil (Rebeca Andrade)	USA (Jade Carey)
Floor Exercise	USA (Simone Biles)	USA (Shilese Jones)	Brazil (Rebeca Andrade)
Uneven Bars	USA (Shilese Jones)	Trinidad and Tobago (Annalise Newman)	Algeria (Kaylia Nemour)

Team USA Athletes: Simone Biles, Jade Carey, Jordan Chiles, Shilese Jones, Konnor McClain
Team USA Medal Count: 10
Team USA Weighted Medal Count: 25 (where Gold = 3, Silver = 2, Bronze = 1)
Team USA Medal Breakdown: 6 Gold, 3 Silver, 1 Bronze

Discussion

Brody Malone, Asher Hong, Patrick Hoopes, Paul Juda, and Colt Walker were selected as the men's Team USA in the simulation where the USA has the highest weighted medal score. All five of these athletes are also proven to be highly accomplished [4]:

- Brody Malone is the 2021 and 2022 U.S. National All-Around Champion
- Asher Hong was part of the bronze-winning team at the 2023 World Championships

- Patrick Hoopes was a 2023 NCAA All-American on the pommel horse
- Paul Juda was part of the bronze-winning team at the 2023 World Championships
- Colt Walker was part of the gold-winning team at the 2022 Pan-American Championships

The graphs below (Figure 3) show the scores for each athlete selected for Team USA on each apparatus through their 2022-2023 competitions. It can be seen which athletes may be "specialists" versus "all-arounders". These graphs also show the scores of each athlete on each apparatus compared to the others on Team USA, which can be helpful for strategy (for example, selecting which athletes will compete on which apparatus in the team event).

Some items of note are that Patrick Hoopes on the Men's team seems to specialize on the Pommel Horse (he also claims Pommel Horse is his favorite event because "it's not scary") [5]. Simone Biles had not scored below a 14 on any competition in the 2022-2023 dataset; her score distribution being on the higher end is consistent with her earning the title of the most decorated gymnast in history.

Additionally, we created heatmaps of the average country performance across the 300 simulations with team USA held constant and with the selected athletes from Results. Figure 4, below, shows that for men's gymnastics, individual athletes who do not belong to a qualifying country tend to dominate events such as Pommel Horse, Floor Exercise, Vault, and Still Rings. However, this could be unrealistic in the Olympics, as not as many athletes are allowed to compete for the individual events, and individual athletes still have to qualify in order to compete. In our simulations, we allowed for all individual athletes from non-qualifying countries who passed our athlete selection criteria to compete in individual events. For our selected USA men's team, Team All-Around and Individual All-Around appear to be USA's strong suit. The heatmap also shows that China and Japan are strong competitors, agreeing with past Olympic results. China and Japan have historically dominated many Olympic men gymnastic events [6].

Figure 4 shows that for women's gymnastics, our selected USA team tends to dominate most events, except uneven bars. This agrees with historical performances, as the USA has not won gold in Uneven Bars at the Olympics since 1984 [7]. USA's strong suit is Individual All-Around, which also agrees with USA's historical performances. USA has won gold in women's Individual All-Around at every single Olympics since 2004 [8].

An observation, noted in the results section, is that the selected women's Team USA receives the highest weighted medal number in the Olympic simulations more times than the selected men's Team USA. In addition, the best weighted medal score for the women's team is 25 and for the men's team is 16, despite the men having more medal events. This is consistent with the past performance of the Women's team; they are the 2023 reigning World Champions, and won silver at the last Olympics. The Men's team won bronze in the 2023 World Championships and did not place in the last Olympics; however, the individual athletes selected for Team USA have a strong track record described at the beginning of the discussion, and sports commentators believe they have a strong potential to place in the 2024 Games [9].

Although not included in the scope of this analysis, the simulation methods here can also provide ideas for Team USA alternates, who will still need to compete and train for the Olympics in case of an unanticipated emergency. For both the men's and women's alternates, it would be reasonable to select from the pool of athletes who made it into one of the "top 10 performing teams" in the first simulation round, but did not ultimately make the cut during the second round of simulations.

It is interesting to compare the athletes on Team USA who were selected from statistical simulations with general online gossip about who is expected to earn a spot on Team USA. There are rich discussions on online forums, including Reddit, that include people's opinions about who will represent Team USA. For example, there is a recent thread with more than 80 comments discussing the 2024 Women's Team USA [10]. The comment with the highest number of "upvotes", which are a form of endorsement other users give to a post, predicts that "Simone and Shilese are now locks", with "Suni, Jade, Jordan, Leanne, Joscelyn, Skye as the next level". Four of the five athletes in our recommended Team USA for women's gymnastics seem to also be popular "fan favorites" on social media. There is less consolidated information on social media for USA men's gymnastics team predictions.

There are limitations related to both our methodology chosen and inherent in the process of attempting to statistically simulate a future event. As athletes are human beings and there are more than six months until the start of the Olympic Games, uncertainty exists with the training schedules or unforeseen events with the athletes we chose for Team USA in this report. For example, Konnor McClain, who we included in the Women's Team USA, pulled out of a competition this summer due to her recent hand surgery [11]. Although she is still slated to pursue the Olympic Games, uncertainties like these could affect the outcome of the Games.

The variability in human performance also includes the competitions that have happened so far, since the statistical methods here only use data from past competitions in 2022 and 2023. Particularly, if an athlete has been injured
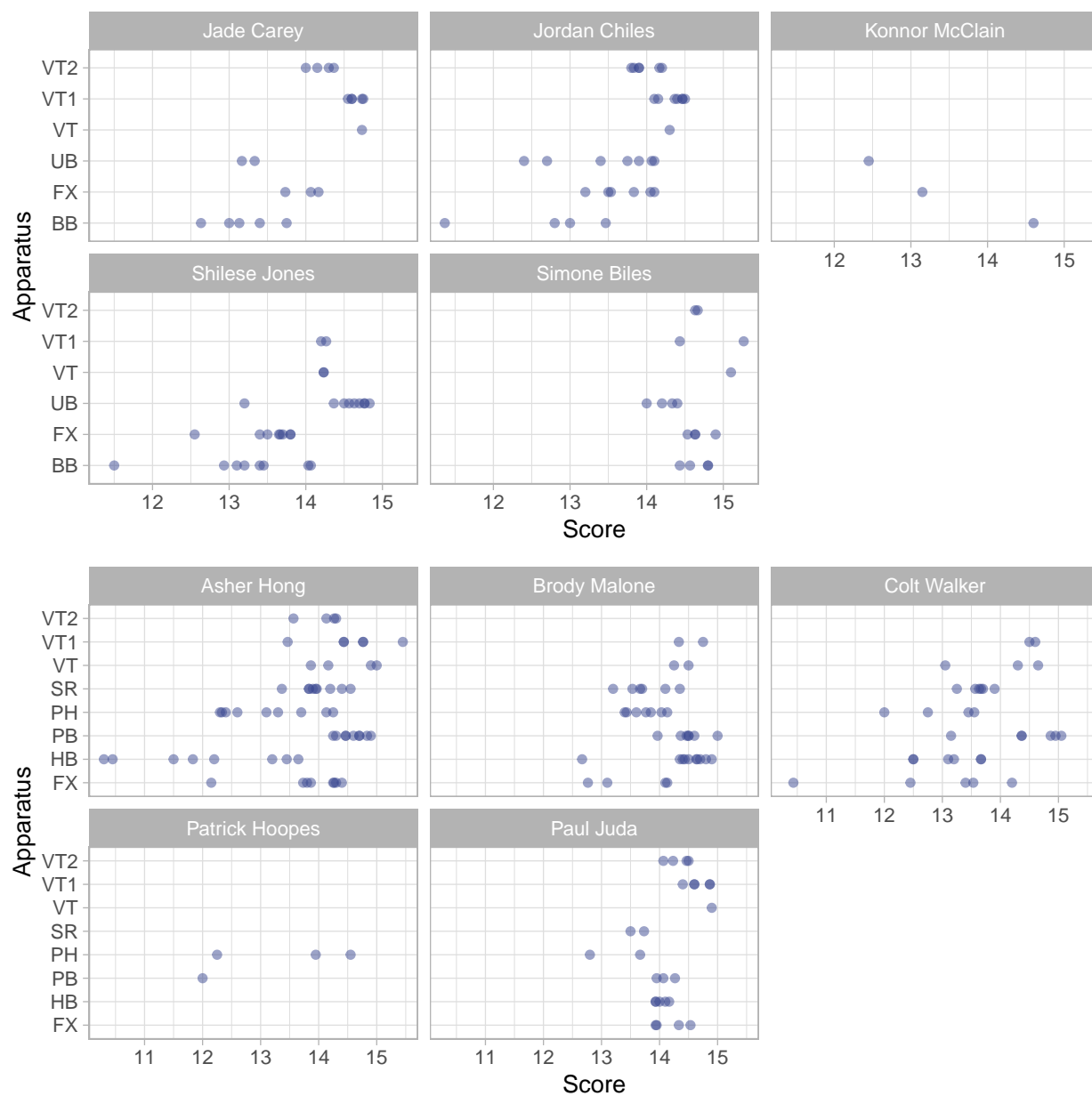
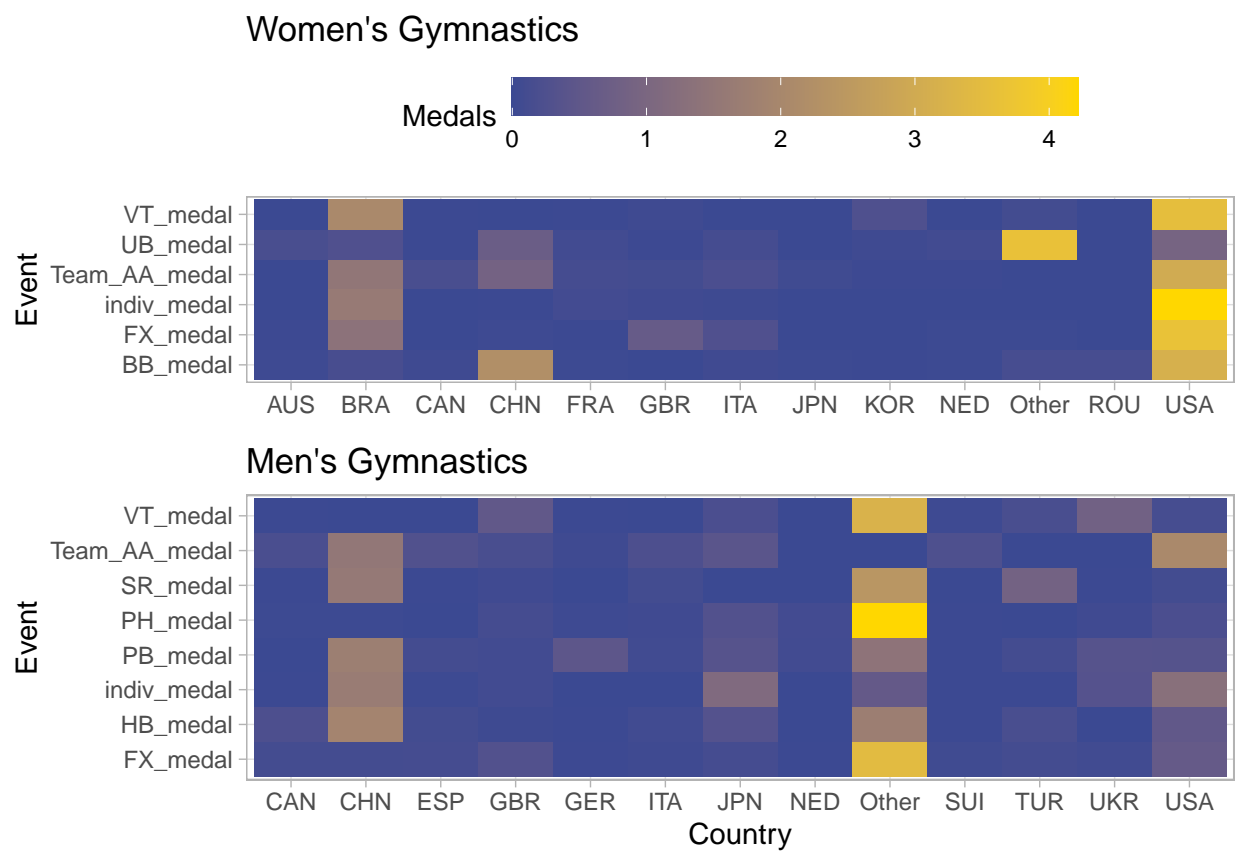Figure 3: Every Competition Score for each team USA Athlete by Apparatus

Figure 4: Average Country Performance Across Simulations

during the 2022-2023 season but is still planning on competing in the Olympics, they may not have shown up as a high performer in our dataset because of their limited or lack of past scores. An example from the United States is Sunisa Lee; although she was the Tokyo Olympics individual all-around champion and hopes to compete in the Paris Games, she has not competed in several competitions in 2023 due to a kidney issue [12]. She ultimately did not make the top simulated choice for Team USA despite winning gold medals at past Olympic Games.

More generally, since our method included taking the median of the kernel densities of each athlete and apparatus for team selection, the median may not provide a comprehensive summary of the data if there are only a small number of data points for an athlete.

Due to our method of creating a continuous distribution from the observed scores of the athletes, the simulation did not truncate the potential scores that athletes could receive at the maximum possible score for each apparatus. The score distributions for each athlete and apparatus potentially allow for technically impossible scores (such as a simulated score higher than the highest possible real-world score), albeit at a low probability.

Additionally, our simulation did not weight scores based on recency, which does not take into account if athletes experienced a growth trajectory during the 2022-2023 season-- who may have ranked poorly in the beginning competitions but are beginning to perform better-- had kernel densities that weighted their poor rankings the same as their more recent high rankings. More qualitative data about each athlete could also be important to include. For example, an athlete's future plans or competition data from past Olympics were not included; it could make sense to include whether a past Olympian is defending his/her title, or whether this is someone's last planned Olympics, as determining factors if two athletes have otherwise similar performance history. It could also be important to have feedback from coaches about the mental and physical readiness of each athlete. Including these additional variables could have added nuance to our selection process.

Our method of simulation included assumptions that defined what a "best performance" by Team USA meant. Particularly, the weighting of medals (with gold = 3, silver = 2, and bronze = 1) assumed an increase in prestige that could differ from other interpretations. If the United States is only interested in the total number of people on the podium, for example, a bronze medal may be weighted the same as a gold medal. Another assumption made was that each event medal was equally prestigious; a "gold" in vault was weighted the same as a "gold" in the floor exercise. Depending on the priorities or strategies of the USA Gymnastics, they may see certain events as more important to win (for example to defend a title), in which case it would be appropriate to apply a different level of weighting for weighted random sampling of teams. For example, if team USA wants to focus on winning gold for uneven bars, the weights could then be biased towards gymnasts who compete in this event. These alternative methods could be useful depending on the specific goals of USA Gymnastics.

We were also constrained by computational power and time limitations related to the number of simulations we were able to run. With more computational power, it would have also been possible to run additional iterations to narrow down which team combinations perform best under more circumstances.

A similar simulation methodology to the one used in this report could be applicable to a broader range of Olympic and non-Olympic sports, for sports officials and coaches as well as for individuals hoping to understand how a sports event will go. Our simulation methodology is applicable to many other sporting events which use a standardized scoring system, such as figure skating, archery, and diving. There is also a financial incentive to better understand these competitions and to be able to make predictions. The sports industry had a global revenue of almost $487 billion U.S. dollars in 2022, with this number expected to grow to $623 billion U.S. Dollars by 2027 [13]. There is a large incentive for sports officials to hold exciting competitions, and for coaches to select athletes who can achieve high results for their team. In addition, the most prestigious athletic events garner worldwide attention, with many fans betting monetarily as well as emotionally on the outcome of their teams. While this was not our primary focus, our methodology allows for prediction of end results across contestants. Therefore even though there is an inherent element of randomness to sports events, viewers interested in this market could use this simulation methodology to predict the outcomes of the teams.

In conclusion, this paper predicts the 5 men and 5 women who can make Team USA achieve the highest weighted medal score at the Paris 2024 Artistic Gymnastics Events. We recommend that Team USA for the Men's Event should include Brody Malone, Asher Hong, Patrick Hoopes, Paul Juda, and Colt Walker, and Team USA for the Women's Event should include Simone Biles, Shilese Jones, Jordan Chiles, Konnor McClain, and Jade Carey. These accomplished athletes have proven success in the 2022-2023 Gymnastics season, with a demonstrated potential to bring prestige to the United States in the 2024 Olympic Games.
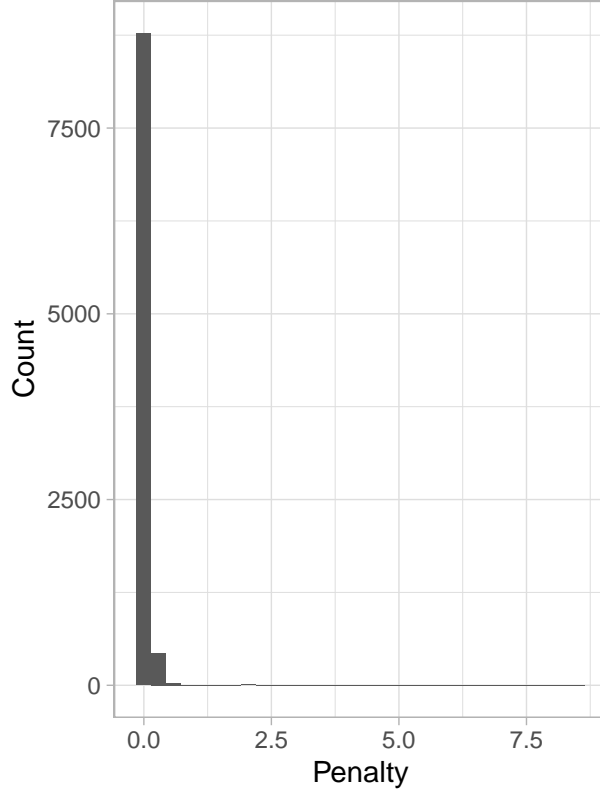
Citations

[1] Young, Jabari. "Simone Biles' Return Helped Tokyo Olympics Viewership, but the Average Stays at 16.8 Million Viewers." CNBC, CNBC, 5 Aug. 2021, www.cnbc.com/2021/08/04/simone-biles-return-helped-olympics-viewership-average-16point8-million.html.

[2] Published by Statista Research Department, and Dec 9. "Most Popular Sports in the Summer Olympics US 2021." Statista, 9 Dec. 2022, www.statista.com/statistics/1245746/summer-olympics-most-followed-sports-us/.

[3] James Herbert Oct 5. "2024 Paris Olympics: Projecting Team USA's 12-Man Roster with Joel Embiid Joining Lebron James, Stephen Curry." CBSSports.Com, 5 Oct. 2023, www.cbssports.com/nba/news/2024-paris-olympics-projecting-team-usas-12-man-roster-with-joel-embiid-joining-lebron-james-stephen-curry/.

[4] "List of IOC Country Codes." Wikipedia, Wikimedia Foundation, 3 Nov. 2023, en.wikipedia.org/wiki/List_of_IOC_country_codes.

[5] USA Gymnastics, 14 Aug. 2023, usagym.org/.

[6] "Patrick Hoopes - Men's Gymnastics." Air Force Academy Athletics, goairforcefalcons.com/sports/mens-gymnastics/roster/patrick-hoopes/17785. Accessed 19 Nov. 2023.

[7] "Olympic Games History - Men USA Gymnastics." USA Gymnastics, 28 July 2023, usagym.org/history/olympics-men/.

[8] "Uneven Bars at the Olympics." Wikipedia, Wikimedia Foundation, 14 Oct. 2021,

en.wikipedia.org/wiki/Uneven_bars_at_the_Olympics.

[9] "Individual All-around Artistic Gymnastics at the Olympics." Wikipedia, Wikimedia Foundation, 6 May 2023,

en.wikipedia.org/wiki/Individual_all-around_artistic_gymnastics_at_the_Olympics#:~:text=

The%20individual%20all%2Daround%20is,competition%20was%20added%20in%201952.

[10] Roenigk, Alyssa. "Takeaways from Gymnastics Worlds: Looking Ahead to the 2024 Olympics." ESPN, ESPN Internet Ventures, www.espn.com/olympics/gymnastics/story//id/38603995/2023-gymnastics-worlds-results-2024-olympics-predictions. Accessed 19 Nov. 2023.

[11] "Who Will Be on the USA Women's Team at the 2024 Olympics?" Reddit,

www.reddit.com/r/Gymnastics/comments/1700rmc/who_will_be_on_the_usa_womens_team_at_the_2024/. Accessed 19 Nov. 2023.

[12] Zaccardi, Nick. "Konnor Mcclain to Miss U.S. Gymnastics Meets after Hand Surgery." NBC Sports, NBC Sports, 13 July 2023, www.nbcsports.com/olympics/news/konnor-mcclain-gymnastics-surgery.

[13] Pathak, Manasi. "Olympic Champion Sunisa Lee Expects Paris 2024 Games to Be Stressful." Forbes, Forbes Magazine, 31 July 2023, www.forbes.com/sites/manasipathak-1/2023/07/24/olympic-champion-sunisa-lee-expects-paris-2024-games-to-be-stressful/?sh=53a9ae9c600c.

[14] Gough, Christina. "Global Sports Market Revenue 2027." Statista, 13 July 2023, www.statista.com/statistics/370560/worldwide-sports-market-revenue/.

[15] "Vault (Gymnastics)." Wikipedia, Wikimedia Foundation, 30 Oct. 2023, en.wikipedia.org/wiki/Vault(gymnastics).

[16] **"Chatgpt." ChatGPT, openai.com/chatgpt. Accessed 19 Nov. 2023.

**Note: We used ChatGPT to help with understanding and using S4 Objects, which we used to store our data, as well as general debugging.
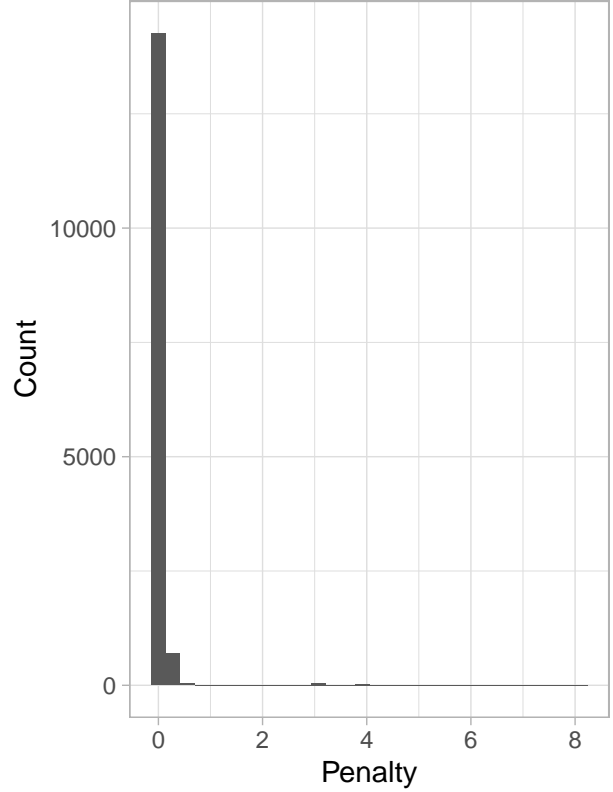
Appendix

Additional Figures

\setcounter{figure}{0}

Appendix 1: Penalty Distribution / Male Athletes

Mathematical Details

Creating Score Distributions

This Appendix section includes details about how the KDEs for a continuous representation of the score distribution for each athlete and apparatus were created.

KDE stands for Kernel Density Estimation, it is a nonparametric method to estimate the distribution of a certain set of points. In our case, we used the KDE to estimate the pdf of the score distribution for an athlete's performance on one apparatus. Delving into the mathematical details, the kernel function is evaluated for each past score separately, and these partial results are summed in order to form our KDE.

We start with a Gaussian kernel with zero mean and unit variance.

$K(x) = \frac{1}{\sqrt{2\pi}} e^{(-\frac{x^2}{2})}$

To shift the kernel to our data point x_i, we subtract the argument. And to make the kernel wider or narrower, we use a constant in the denominator, h, a kernel bandwidth. We also normalize pdf back to unit area by dividing by h. In general, a smaller h will mean a more narrow kernel pdf.

The kernel function for each score is therefore the following:

$\frac{1}{h} K\left(\frac{x-x_i}{h}\right)$

To get the overall KDE estimate $f(x)$ for a set of n points, $x_1 \ldots x_n$, we add all the kernel functions together and multiply by $\frac{1}{n}$:

$f(x) = \frac{1}{n} \left[\frac{1}{h} K\left(\frac{x-x_1}{h}\right) + \ldots + \frac{1}{h} K\left(\frac{x-x_n}{h}\right) \right]$

We used the density function in R to create the KDEs for most athletes. If an athlete only had one score for an event, we were unable to generate a realistic KDE estimate, so for these cases, we used a normal distribution with the mean as the athlete's score and standard deviation of 0.1 As mentioned before, we made gymnasts who compete

14

with two vaults have two score distributions for the event: one distribution for the first vault and one distribution for the second vault. We made a decision to have 2 separate score distributions because there could be a meaningful difference between the score distributions, since gymnasts use the opportunity for two vaults to demonstrate different techniques and execution styles [15].

Team Selection

To create weights for each athlete, we found the median of the KDE for each apparatus they competed in and took a "skewed average" of the medians. We chose this method rather than taking a simple average of the medians since simply averaging the medians would bias the weights toward the specialists in our dataset. For example, an athlete who was really good in only two events, but had zeroes in the other events would have a high average median score, because we are adding the scores and then dividing by 2 (for the two events). For the skewed average, instead of dividing by the number of events n they competed in, we divided by $n + 1$. The equation for weight w of i_{th} athlete is as below.

$$w_i = \frac{median\_apparatus_1 + ... + median\_apparatus_n}{n+1}$$

This biased our weights towards gymnasts who compete on more apparatuses. While we recognize that focusing on winning individual events could be a team strategy, we found with some testing that having too many specialists could cause more "unrealistic" simulations. For example, if specialists are sampled more often, this could mean cases where there would not be enough athletes able to compete on each apparatus for the Team All Around event. This is especially a concern for men's gymnastics, as they have more apparatuses. Even without weights being applied, we had instances where teams would only have 2 athletes that "could compete" (non zero past scores) for an apparatus. Taking this into consideration, we apply the above bias towards all-around athletes in our weighting scheme.

After calculating our pseudo average, we realized the weights were really similar to each other. Therefore to "stretch out" the scale, to emphasize the differences in skill level, we decided to use final weight as e^{w_i}. So, if the weight before was 13, the new weight would be e^{13}. This helped in differentiating between the athlete's weights and making sure again that the really top athletes were almost always chosen.