USOPC Age-Focused Competitive Analysis

Stergios Koutrouvelis - MS Analytics
Georgia Tech Spring 2022

Table of Contents

Introduction

The scope of this analysis is for the United States Olympic and Paralympic Committee to gain a better understanding of the role that age plays in medal success for various Summer and Winter Olympic Sports. The original hypothesis is that peak age for winning medals varies by sport, event and gender. The goal is to evaluate this hypothesis, and if possible leverage this information (potentially in combination with other factors) to be able to more accurately predict which athletes will have a greater chance of winning medals in upcoming Olympic events. By leveraging the data provided by the USOPC, we will conduct a thorough analysis that can help us understand the health of Team USA's Olympic pipeline compared to those of the top medal-winning countries primarily through the lens of age.

Key Questions

The questions that this analysis will try to answer are the following:

- Can a country's percentage of athletes competing at an Olympic Games while in their peak age ranges in their respective sports predict medal success?
- Which athletes who competed in Tokyo 2021 (Beijing 2022) will be hitting their peak age in their respective sports in Paris 2024 (Milan 2026)?
- Can the number of young, "pre-peak" athletes that competed in Tokyo predict medal success for their countries in Paris 2024 and Milan 2026?
- Have many promising athletes who did not compete in Tokyo (and Beijing) but will reach peak age in Paris (and Milan) do Team USA and other countries have in their Pipelines?

At the end of this report, we will summarize our findings around the above questions as well as suggestions for extending the presented work for further improvements and adaptations.

The Data

The USOPC has provided an excellent and consistent set of datasets. Each dataset corresponds to one specific sport and covers multiple events of this sport. For example, the dataset for Snowboarding (which will be the main focus of this analysis) contains multiple different events such as Snowboardcross, Big Air, Slopestyle etc. Every row in the dataset represents one athlete's information and their performance in the specific event. Some of the key columns of these datasets are the following:

- Athlete Name and personal ID
- Athlete's nation
- Age of athlete at the time of the event
- Type of competition (e.g. Olympic Games, World Championship etc.)
- Sport/event or discipline
- Athlete Placement/Rank and Medal
- Result

One key characteristic of the datasets is the consistency of the schema. This enables the development of a framework that is scalable and can be used across sports in an automated way. For this reason, we opted for the development of an API that can ingest, transform and process the data and also plot consistent visualizations and develop machine learning models.

Project Approach and Phases

The approach of this project has emphasized not only on developing a standalone analysis but also on reproducibility and standardization by taking advantage of the data consistency as described previously.

For the purpose of this project we are going to be working with the **Snowboarding** dataset and more specifically with the **Snowboardcross** events, as this type of event has the largest number of data points and total athlete participation. The approach can be replicated and be applied and scaled to any dataset and event provided by the USOPC.

The progression can be split into the phases outlined below.

Processing Framework

During this phase, after some initial dive into the data, an API has been developed to make processing easier by developing functions and queries that can be used repeatedly for feature creation, visualizations and machine learning model development. In the initial stage, after going through the data, it was determined that the following columns contain useful information:

- **Class:** This is the competition class. This column can take the values of Elite, Juniors or Youth and is determined primarily by the age limitations of the particular event.
- **Competition Date:** As the name suggests, this is the date the competition has taken place.
- **Competition City:** This is the city where the competition has taken place. It is worth mentioning here that this does not include the country's name but only the city.
- **Competition Country:** This column is being generated from the framework. Based on the competition city column, API calls are being made through the Geopy library to get the competition country. The values are stored in a JSON file to speed up processing as depending on the size of the data, multiple calls can be very time consuming and computationally expensive.
- **Event/Athlete Gender:** As the name suggests this is the gender of the event that also matches the gender of the athlete.
- **Event:** This is the name of the event within the specific sport. For example, processing Snowboarding that is the focus of this analysis, individual events can be Snowboardcross, Slopestyle, Big Air etc.
- **Medal:** This column can take the values G, S, B or None indicating whether the athlete won a medal and if so what medal that was. In later stages of the processing a new

column will be created called won_medal which will be binary and will be used as one of our response variables.
- **Country:** This column represents the athlete's country of origin.
- **Home Competition:** This is one of the engineered columns based on the event country and the athlete's country of origin. If they match, the column value is true, indicating that the athlete is competing at an event held at home. We strongly believe that competing in the home country is an influencing factor in predicting success and we are going to validate this hypothesis based on both our exploratory analysis and the machine learning models.
- **Age:** This is the age of the athlete at the time of the event.
- **Rank:** The rank of the athlete at the event.

Exploratory Data Analysis

After the initial data cleanup, we performed exploratory data analysis across some major axes, in order to understand and identify patterns in the data. The main axes the analysis focused on are the following:

- Country
- Sport
- Event
- Age
- Gender
- Competition Location
- Medal Winning

Combining the factors above, the outline of the exploratory data analysis and graphs produced were the following:

- Exploratory Data Analysis by Athlete Country of Origin and Event to determine the following:
    - **Medal Rate:** Ratio of medals won over total athlete participation for a given country and given event. This factor provides some insight into how successful a country is in a particular type of event overall.
    - **Participation Count:** This parameter represents the total number of athletes from a particular country who have participated in the event under consideration. As countries vary in population and athletes sent to type of event, this will be a key factor in observing how many athletes participate in an event from each country and whether the sample is sufficiently large in order to extract conclusions.
    - **Participation Rate:** This parameter represents the distribution of a country's athletes across different events. It is computed as the ratio of a country's athletes participating in an individual event over the total number of a country's athletes.

This is insightful in terms of showcasing whether different countries participate in certain events more heavily than others.

- Exploratory Data Analysis by Athlete Country of Origin and Age to determine the following:
    - **Medal Rate:** This is the ratio of medals won across different ages. It is computed as the winning athletes for each age group over the total country's athletes in the particular age group. This is independent of the type of the event and focuses only on country of origin and age.
    - **Participation Count:** This represents the total number of athletes in the dataset broken down by country of origin and age group.
    - **Participation Rate:** This variable represents the proportion of a country's athlete that belong to different age groups irrespective of sport. It is computed as the number of a country's athletes in a given age group over the total number of a country's athletes. This is an indicator of whether there is a pattern of specific countries sending athletes of certain age groups to competitions.

- Exploratory Data Analysis for home competition success rate by country. In this analysis, we will be focusing on how different countries perform at home vs away. The graph will show the success rate for competitions held at home and away by computing the successful events over the total events. This will be insightful in determining how competing at home increases chances of success.

After developing the graphs we are going to discuss the observations and results.

Feature Engineering

Data preparation and feature engineering is needed for training and testing machine learning models. While evaluating our approach, two variations of training data have been considered.

In the initial phase, we opted for the simplest version of the data and treated each datapoint as an independent event. The feature set that has been utilized contained the following variables:

- **Gender:** Binary variable indicating the gender of the event and athlete. This is 1 if gender is male and 0 if gender is female
- **Country:** The athlete's country of origin after being processed with a one-hot encoding approach, where each column represents a country and is binary.
- **Is home competition:** Binary variable indicating whether the competition is being held at an athlete's home country or abroad.
- **Age:** The most critical factor for the purpose of this analysis indicating the age of the athlete at the time of the event.
- **Rank:** The rank of the athlete in the corresponding event (to be used as regression response variable).

- **Won_medal:** Binary variable indicating whether the athlete won a medal in the corresponding event (to be used as classification response variable)

In the second approach, we generated time series features, to leverage the power of past performance as an indicator of future performance. In order to achieve that, we compiled a dataset using the following procedure. For every event and athlete, we compiled the following information:

- Person ID
- Event date
- Event and sport name
- Home competition variable
- Athlete Rank in the event
- Won medal binary variable

For the attributes above and using the athlete-event data points defined above, we collected the three events the athlete participated for the same sport preceding the event in reference. This results in the following compilation:

- 4th preceding athlete event for given sport (feature prefix “ft_”)
- 3rd preceding athlete event for given sport (feature prefix “t_”)
- 2nd preceding athlete event for given sport (feature prefix “s_”)
- Event in reference (1st) athlete event for given sport. This is used as our prediction event (feature prefix “f_”).

So the time-series dataset looks as follows:

- **Person_ID:** The athlete identifier.
- **Country:** Athlete's country of origin, converted to one-hot-encoding variable based on a subset of medal winning countries.
- **Event:** event name (snowboardcross, halfpipe etc.).
- **Gender**: Binary variable indicating event and athlete gender.
- **4th latest competition date:** Date of the fourth to latest competition.
- **4th latest competition rank:** Rank of athlete in the respective competition.
- **4th latest competition athlete age:** Age of athlete during the respective competition.
- **4th latest competition won_medal:** Binary variable indicating whether the athlete won a medal in the respective competition.
- **3rd latest competition date**: Date of the third to latest competition.
- **3rd latest competition rank:** Rank of athlete in the respective competition.
  …
- **2nd latest competition won_medal:** Binary variable indicating whether the athlete won a medal in the respective competition.
- **Latest competition date:** Date of the latest competition.

- **Latest competition rank:** Rank of athlete in the latest competition (to be used as regression response variable).
- **Latest competition athlete age:** Age of athlete in the latest competition.
- **Latest competition won_medal:** Binary variable indicating whether the athlete won a medal in the latest competition (to be used as classification response variable).
- **Latest competition is_home_competition:** Binary variable indicating whether the competition was held in the home country.

The features will be used in both approaches for classification and regression with different response variables as described in the next section.

Machine Learning for Predictions

In this phase, we are going to develop several machine learning models for predicting the outcome of a specific athlete participating in an event, using both the **independent event data** and the **time series data** and comparing the results.

The prediction can be treated as a classification or a regression problem depending on the response variable.

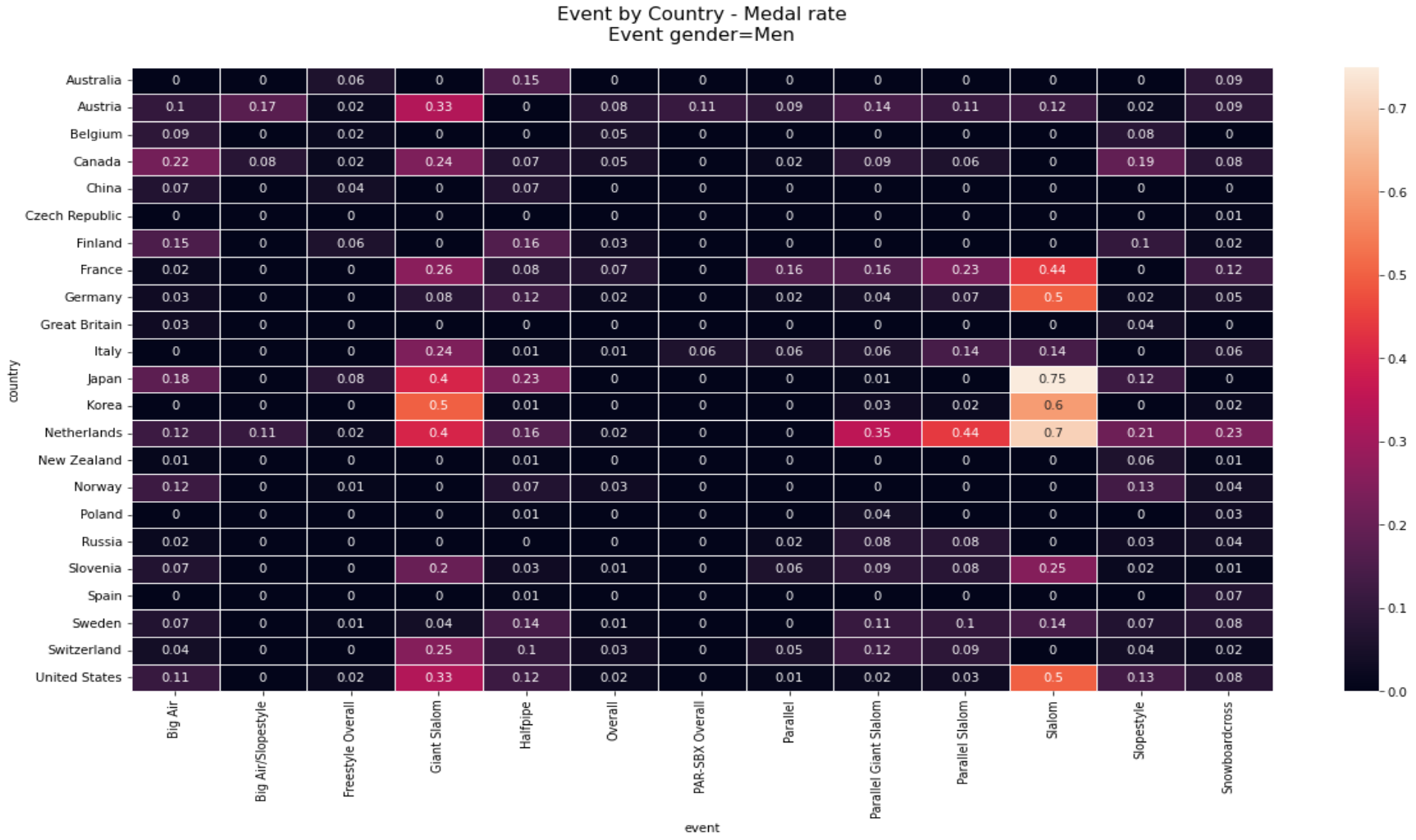
The first approach will focus on the **won_medal** variable as the response variable and treat the problem as classification. As a reminder, won_medal is 1 if the athlete won a medal in a given event or 0 if they did not. As expected this introduces heavy imbalance in the dataset, since for any given event, only three athletes win a medal and the majority does not. Through our analysis it has been determined that this ratio is around 1:12 for the majority of the events. The processing framework addresses this issue by passing class weights as hyperparameters when training the models. The alternative approach is to use the **rank** variable as our response variable and treat the problem as regression. Rank is the final ranking of the athlete in the event.

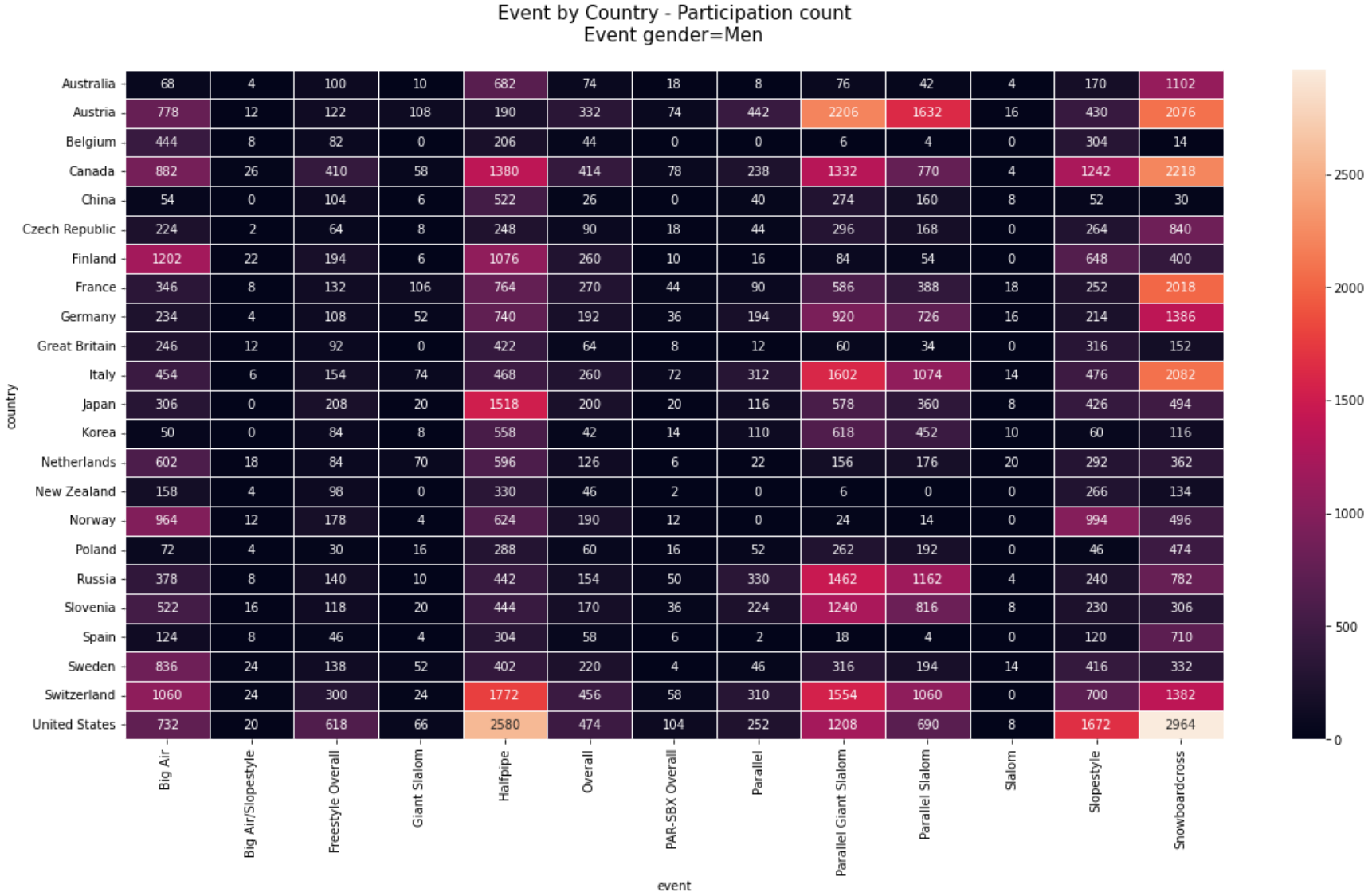
The features that will be used for prediction, and especially age, are highly nonlinear. For this reason we opted for ensemble methods that provide the best of-the-shelf accuracy and can handle nonlinear relationships and trends very well. The two algorithms that we will use are Random Forest and Gradient Boosting Decision Tree. In the hyperparameter tuning phase, we are going to employ cross-validation with 5 folds in order to determine the hyperparameters in an automated way and the best model's results will be reported in each case..

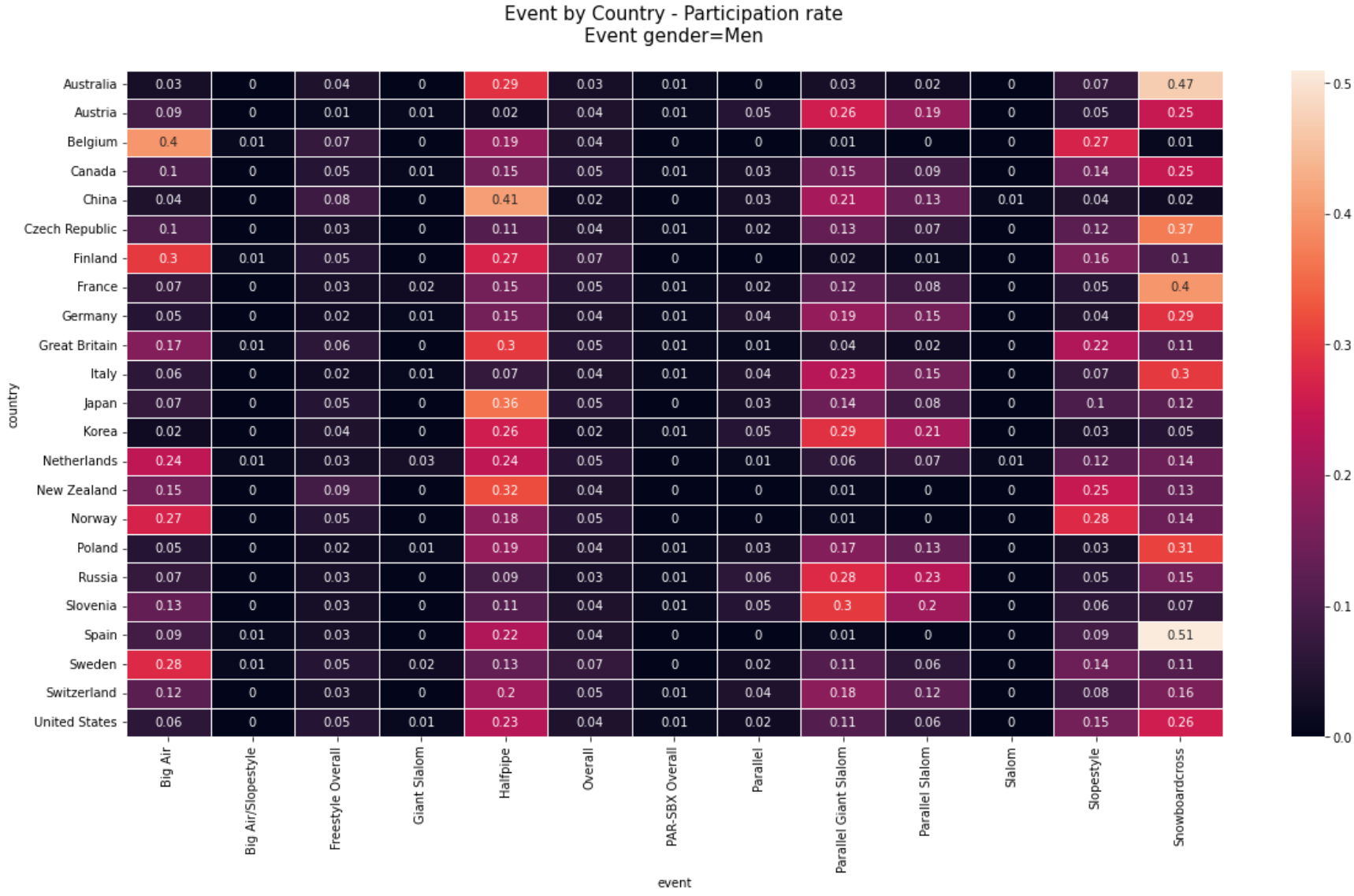
For the classification problem the metrics that we will use in evaluating the performance are precision, recall and f1-score. The regression algorithms will be evaluated on mean absolute error and mean squared error

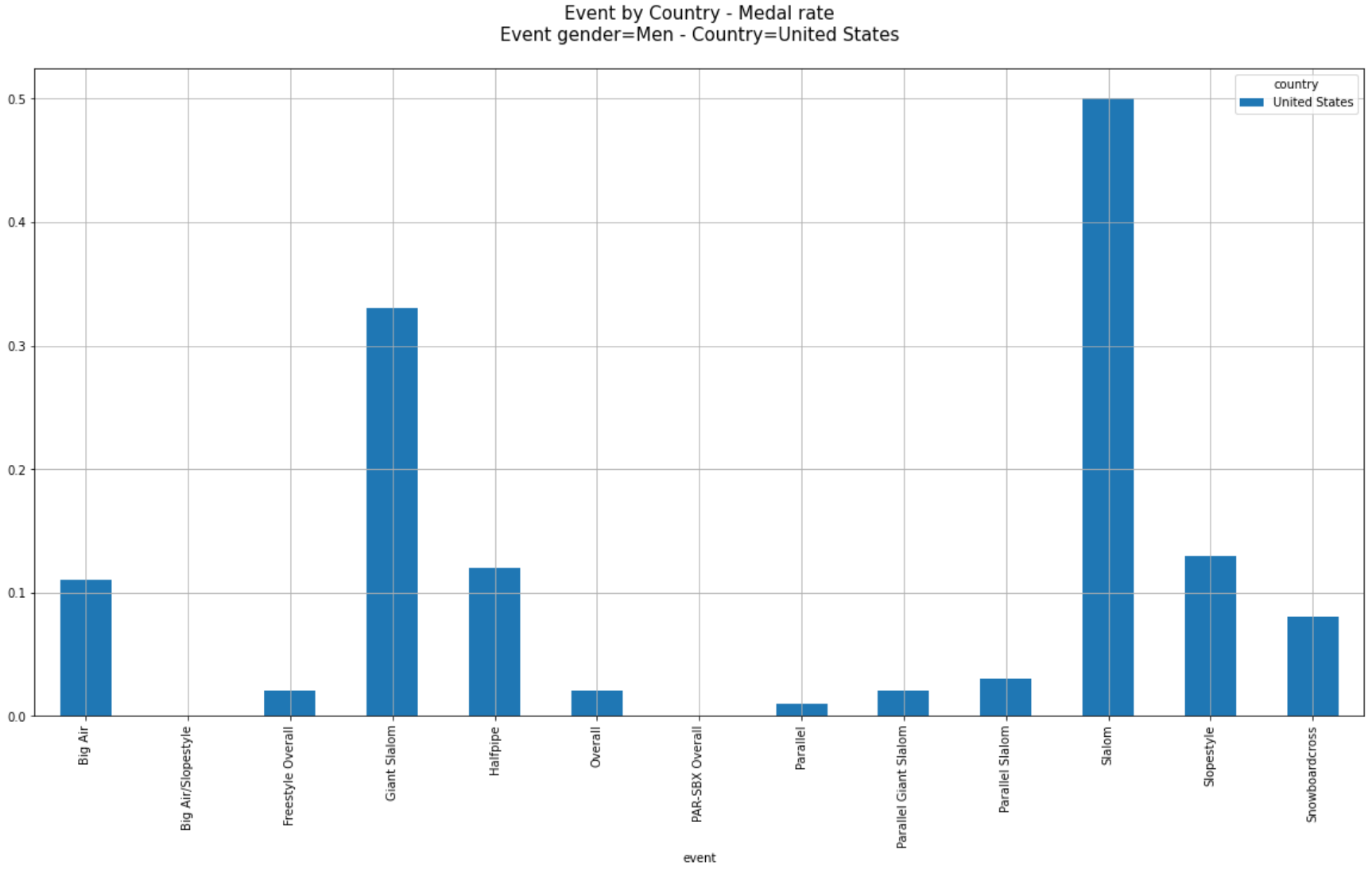
Exploratory Data Analysis

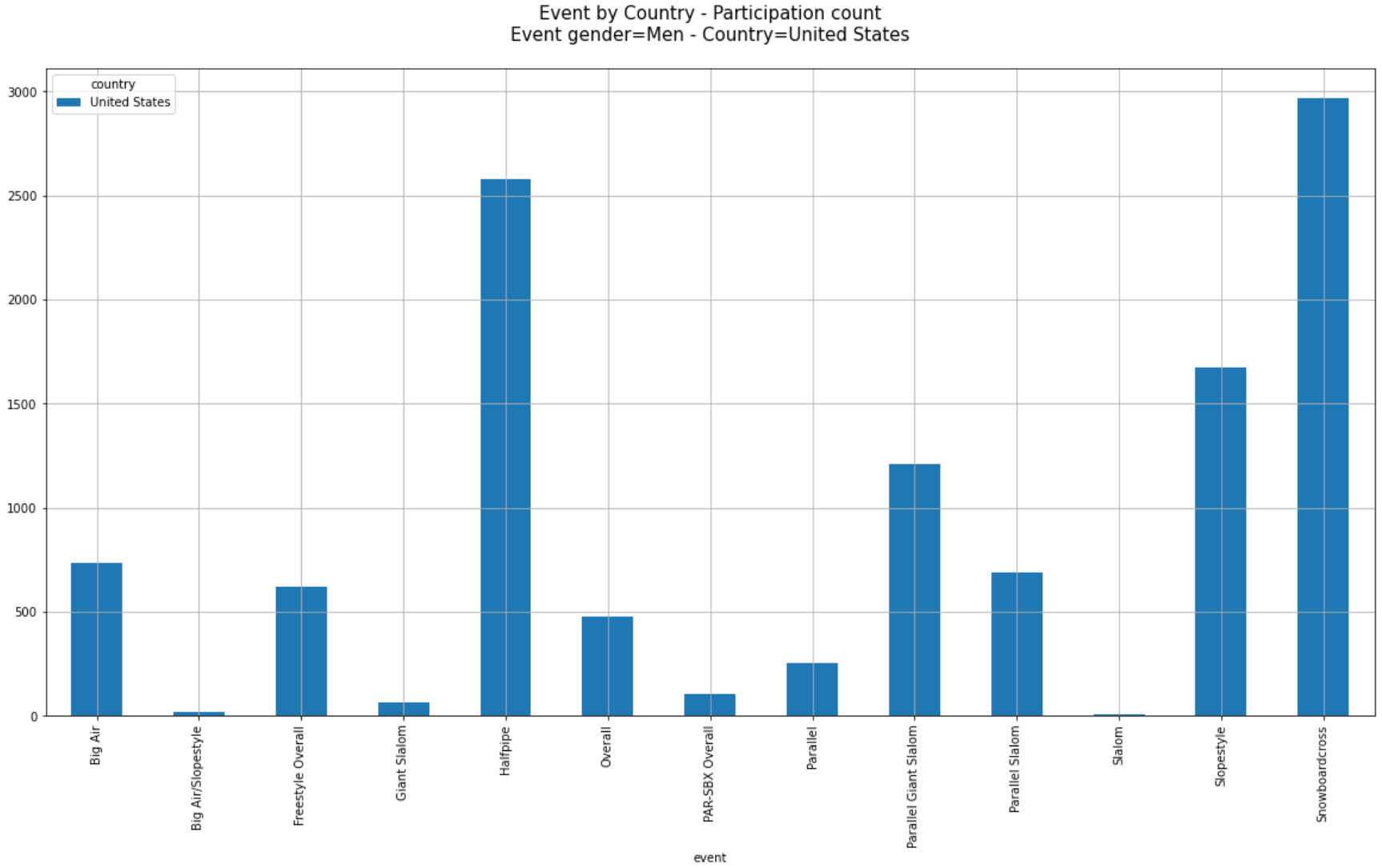
Analysis by Country of Origin and Event Type

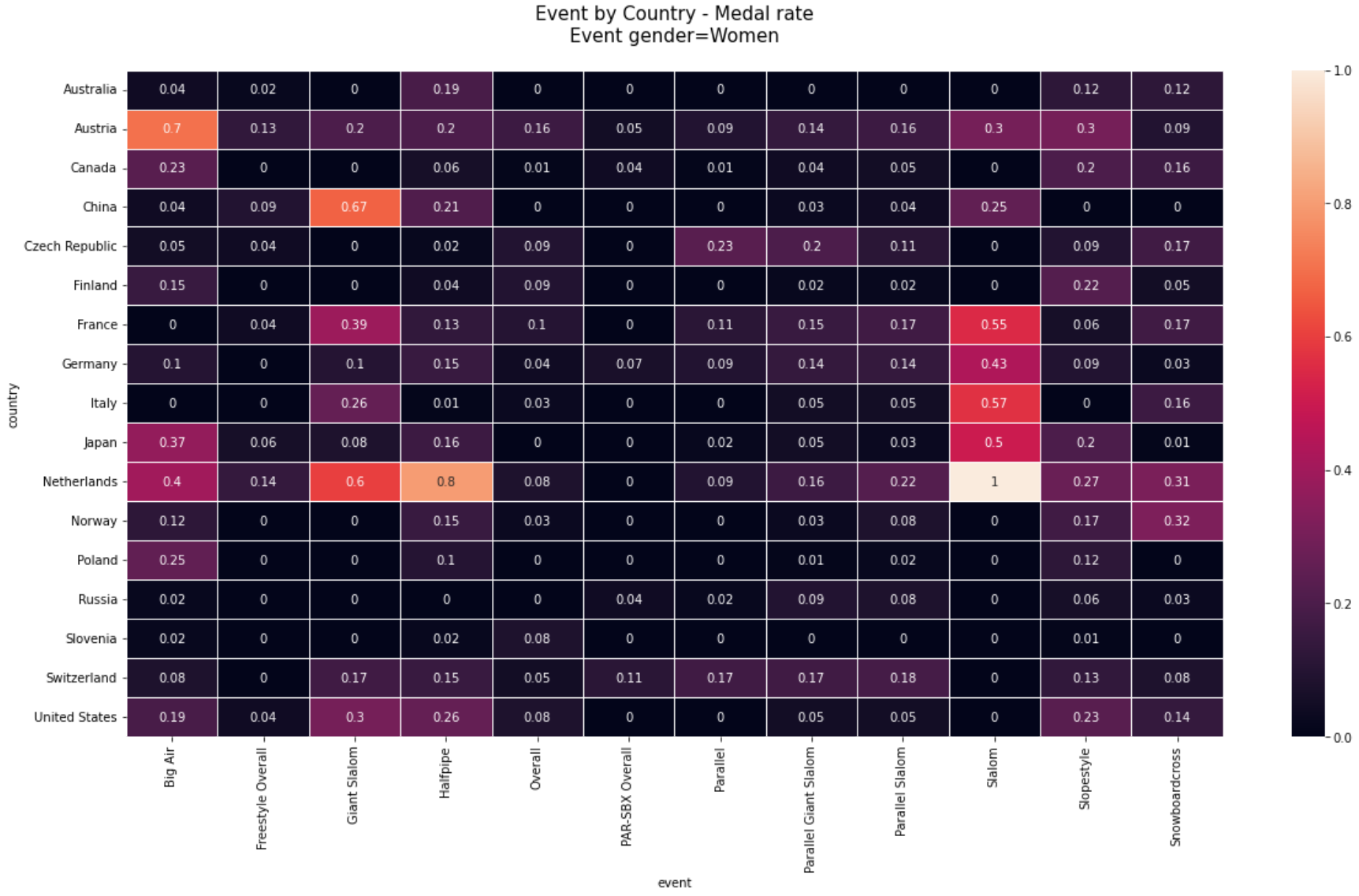


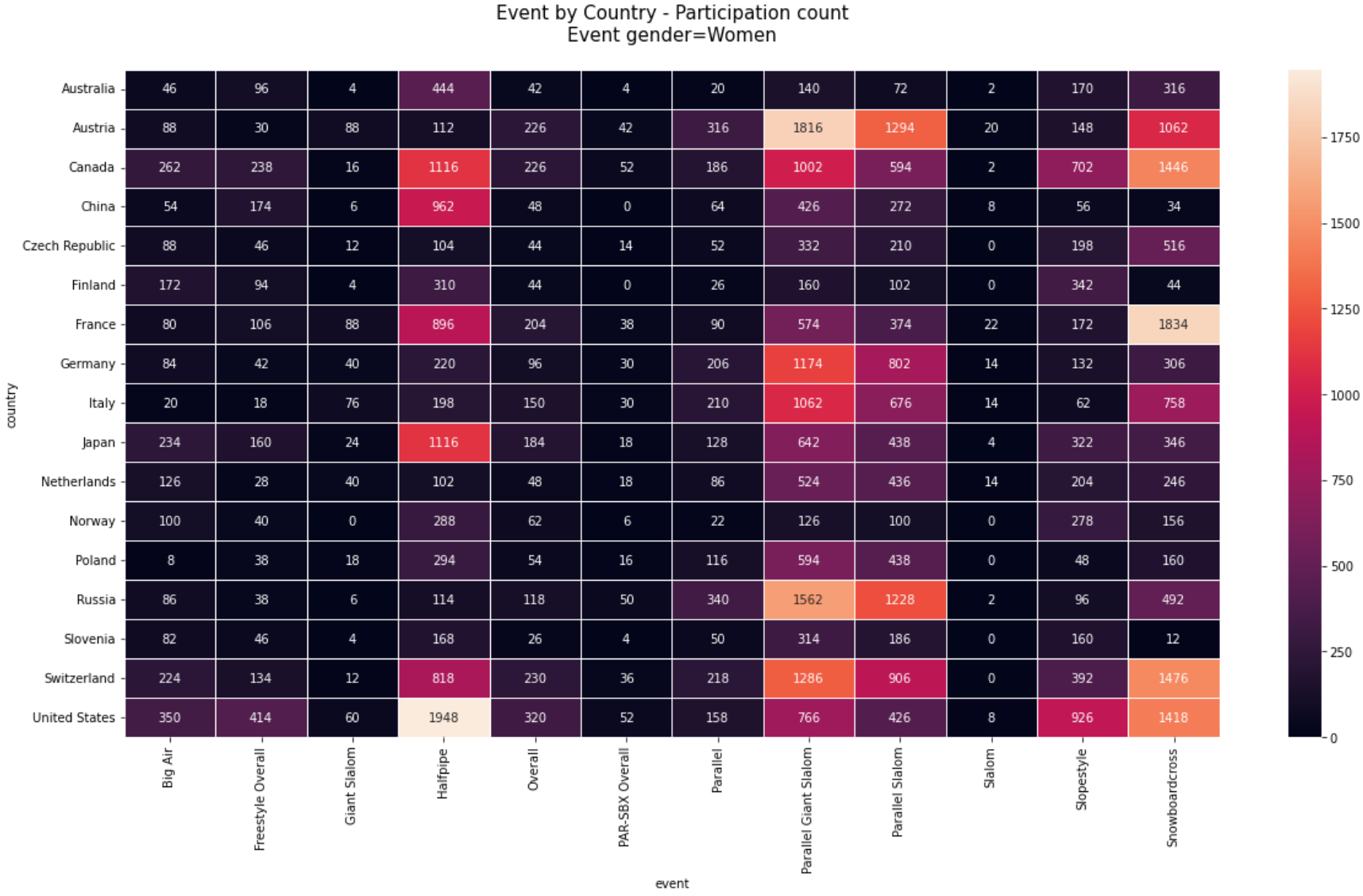

Event by Country - Participation count
Event gender=Men
country
Australia 68 4 100 10 682 74 18 8 76 42 4 170 1102
Austria 778 12 122 108 190 332 74 442 2206 1632 16 430 2076
Belgium 444 8 82 0 206 44 0 0 6 4 0 304 14
Canada 882 26 410 58 1380 414 78 238 1332 770 4 1242 2218
China 54 0 104 6 522 26 0 40 274 160 8 52 30
Czech Republic 224 2 64 8 248 90 18 44 296 168 0 264 840
Finland 1202 22 194 6 1076 260 10 16 84 54 0 648 400
France 346 8 132 106 764 270 44 90 586 388 18 252 2018
Germany 234 4 108 52 740 192 36 194 920 726 16 214 1386
Great Britain 246 12 92 0 422 64 8 12 60 34 0 316 152
Italy 454 6 154 74 468 260 72 312 1602 1074 14 476 2082
Japan 306 0 208 20 1518 200 20 116 578 360 8 426 494
Korea 50 0 84 8 558 42 14 110 618 452 10 60 116
Netherlands 602 18 84 70 596 126 6 22 156 176 20 292 362
New Zealand 158 4 98 0 330 46 2 0 6 0 0 266 134
Norway 964 12 178 4 624 190 12 0 24 14 0 994 496
Poland 72 4 30 16 288 60 16 52 262 192 0 46 474
Russia 378 8 140 10 442 154 50 330 1462 1162 4 240 782
Slovenia 522 16 118 20 444 170 36 224 1240 816 8 230 306
Spain 124 8 46 4 304 58 6 2 18 4 0 120 710
Sweden 836 24 138 52 402 220 4 46 316 194 14 416 332
Switzerland 1060 24 300 24 1772 456 58 310 1554 1060 0 700 1382
United States 732 20 618 66 2580 474 104 252 1208 690 8 1672 2964
Big Air
Big Air/Slopestyle
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event
2500
2000
1500
1000
500
0

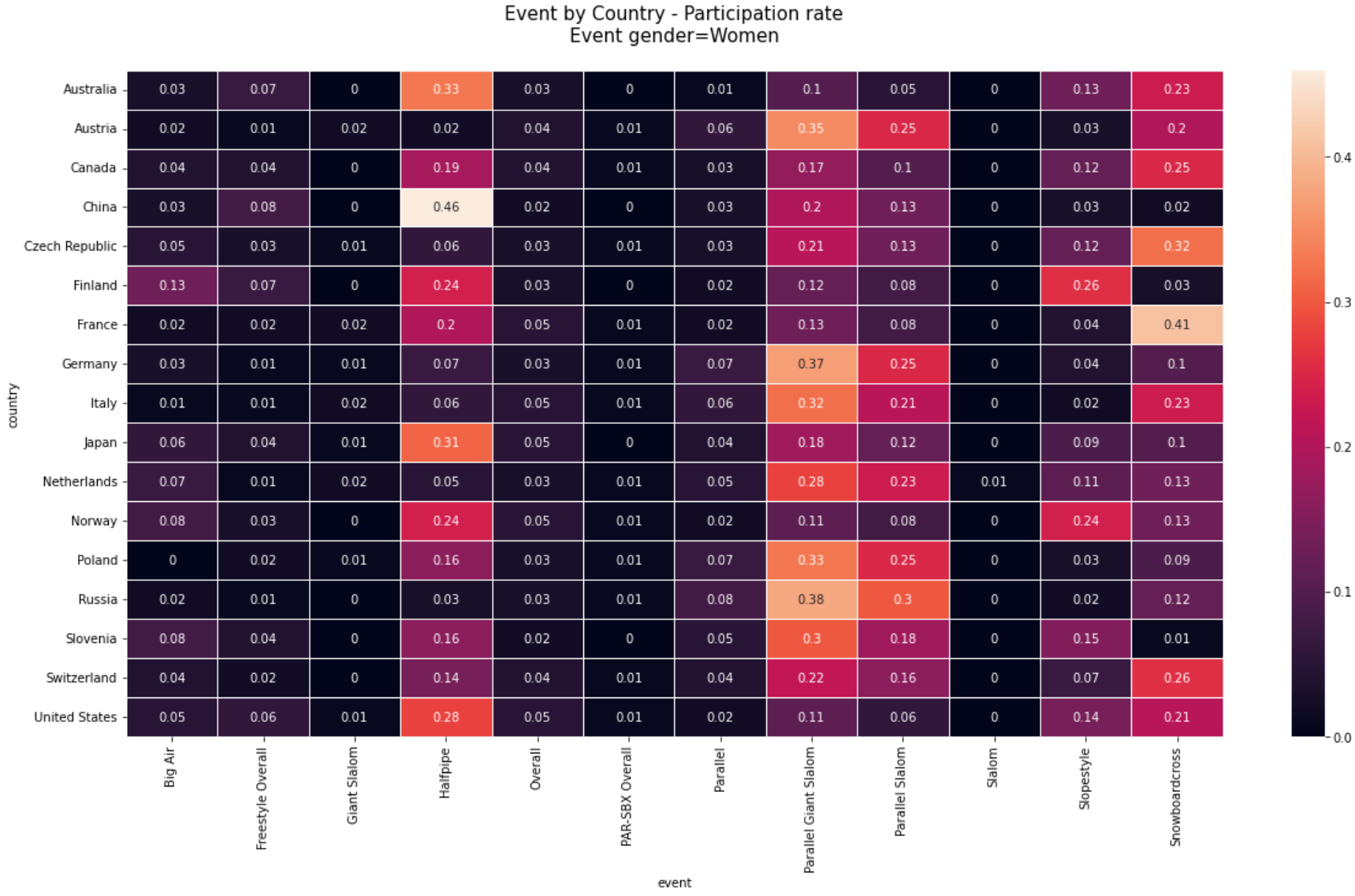

Event by Country - Participation rate
Event gender=Men
country
Australia 0.03 0 0.04 0 0.29 0.03 0.01 0 0.03 0.02 0 0.07 0.47
Austria 0.09 0 0.01 0.01 0.02 0.04 0.01 0.05 0.26 0.19 0 0.05 0.25
Belgium 0.4 0.01 0.07 0 0.19 0.04 0 0 0.01 0 0 0.27 0.01
Canada 0.1 0 0.05 0.01 0.15 0.05 0.01 0.03 0.15 0.09 0 0.14 0.25
China 0.04 0 0.08 0 0.41 0.02 0 0.03 0.21 0.13 0.01 0.04 0.02
Czech Republic 0.1 0 0.03 0 0.11 0.04 0.01 0.02 0.13 0.07 0 0.12 0.37
Finland 0.3 0.01 0.05 0 0.27 0.07 0 0 0.02 0.01 0 0.16 0.1
France 0.07 0 0.03 0.02 0.15 0.05 0.01 0.02 0.12 0.08 0 0.05 0.4
Germany 0.05 0 0.02 0.01 0.15 0.04 0.01 0.04 0.19 0.15 0 0.04 0.29
Great Britain 0.17 0.01 0.06 0 0.3 0.05 0.01 0.01 0.04 0.02 0 0.22 0.11
Italy 0.06 0 0.02 0.01 0.07 0.04 0.01 0.04 0.23 0.15 0 0.07 0.3
Japan 0.07 0 0.05 0 0.36 0.05 0 0.03 0.14 0.08 0 0.1 0.12
Korea 0.02 0 0.04 0 0.26 0.02 0.01 0.05 0.29 0.21 0 0.03 0.05
Netherlands 0.24 0.01 0.03 0.03 0.24 0.05 0 0.01 0.06 0.07 0.01 0.12 0.14
New Zealand 0.15 0 0.09 0 0.32 0.04 0 0 0.01 0 0 0.25 0.13
Norway 0.27 0 0.05 0 0.18 0.05 0 0 0.01 0 0 0.28 0.14
Poland 0.05 0 0.02 0.01 0.19 0.04 0.01 0.03 0.17 0.13 0 0.03 0.31
Russia 0.07 0 0.03 0 0.09 0.03 0.01 0.06 0.28 0.23 0 0.05 0.15
Slovenia 0.13 0 0.03 0 0.11 0.04 0.01 0.05 0.3 0.2 0 0.06 0.07
Spain 0.09 0.01 0.03 0 0.22 0.04 0 0 0.01 0 0 0.09 0.51
Sweden 0.28 0.01 0.05 0.02 0.13 0.07 0 0.02 0.11 0.06 0 0.14 0.11
Switzerland 0.12 0 0.03 0 0.2 0.05 0.01 0.04 0.18 0.12 0 0.08 0.16
United States 0.06 0 0.05 0.01 0.23 0.04 0.01 0.02 0.11 0.06 0 0.15 0.26
Big Air
Big Air/Slopestyle
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event
0.5
0.4
0.3
0.2
0.1
0.0

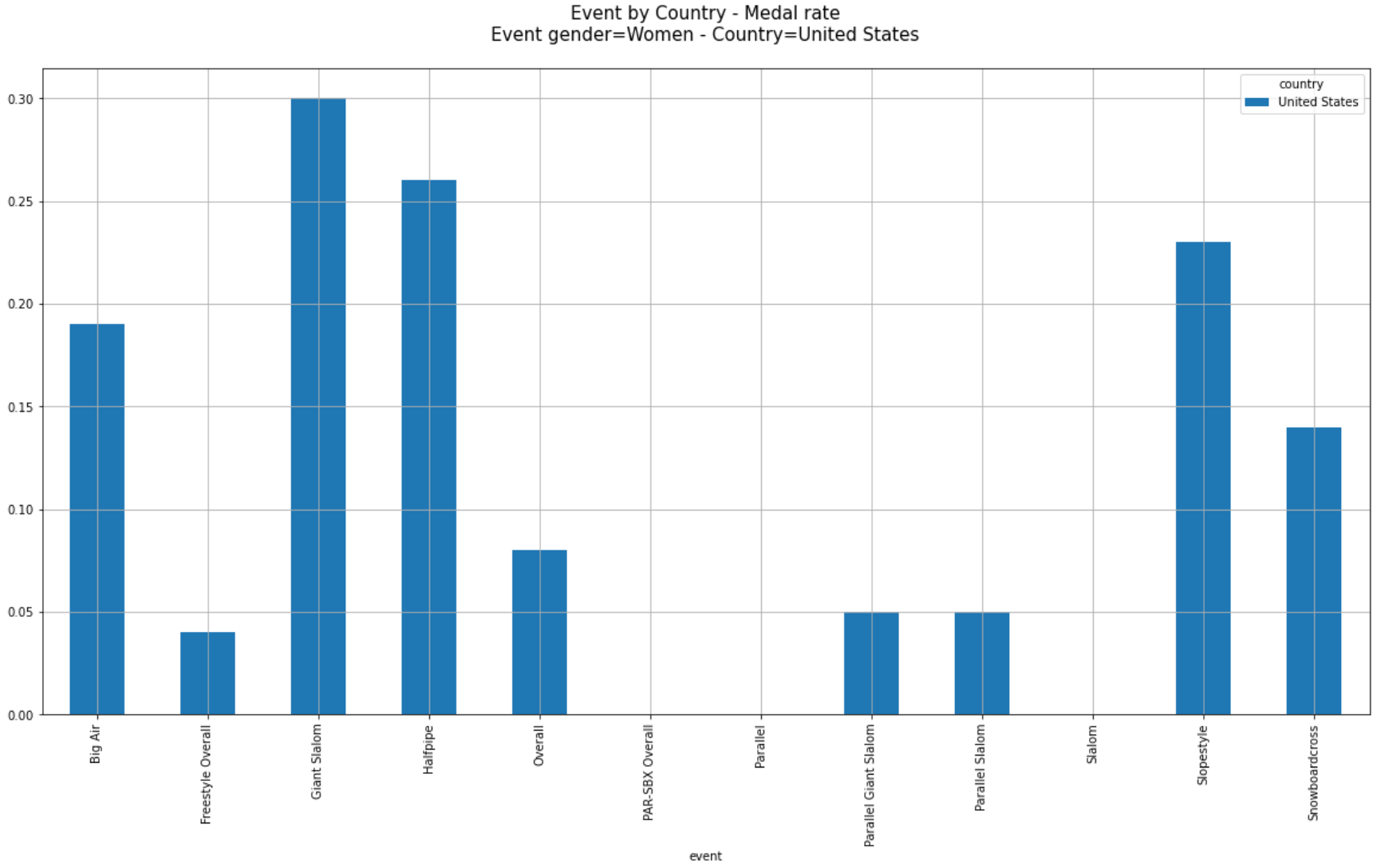

Event by Country - Medal rate
Event gender=Men - Country=United States
country
United States
0.5
0.4
0.3
0.2
0.1
0.0
Big Air
Big Air/Slopestyle
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event

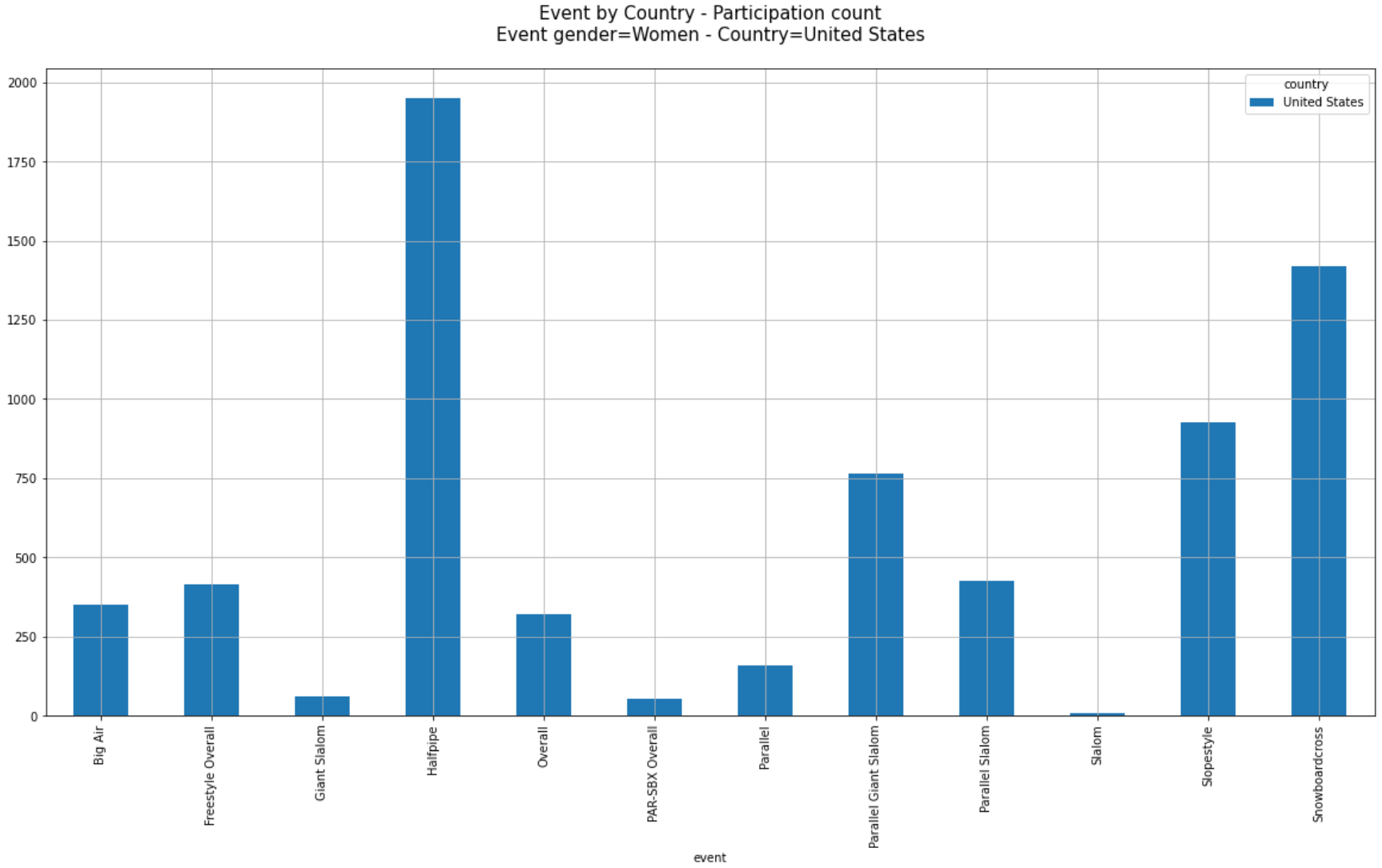

Event by Country - Participation count
Event gender=Men - Country=United States
country
United States
3000
2500
2000
1500
1000
500
0
Big Air
Big Air/Slopestyle
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event


Event by Country - Medal rate
Event gender=Women
country
Australia 0.04 0.02 0 0.19 0 0 0 0 0 0 0.12 0.12
Austria 0.7 0.13 0.2 0.2 0.16 0.05 0.09 0.14 0.16 0.3 0.3 0.09
Canada 0.23 0 0 0.06 0.01 0.04 0.01 0.04 0.05 0 0.2 0.16
China 0.04 0.09 0.67 0.21 0 0 0 0.03 0.04 0.25 0 0
Czech Republic 0.05 0.04 0 0.02 0.09 0 0.23 0.2 0.11 0 0.09 0.17
Finland 0.15 0 0 0.04 0.09 0 0 0.02 0.02 0 0.22 0.05
France 0 0.04 0.39 0.13 0.1 0 0.11 0.15 0.17 0.55 0.06 0.17
Germany 0.1 0 0.1 0.15 0.04 0.07 0.09 0.14 0.14 0.43 0.09 0.03
Italy 0 0 0.26 0.01 0.03 0 0 0.05 0.05 0.57 0 0.16
Japan 0.37 0.06 0.08 0.16 0 0 0.02 0.05 0.03 0.5 0.2 0.01
Netherlands 0.4 0.14 0.6 0.8 0.08 0 0.09 0.16 0.22 1 0.27 0.31
Norway 0.12 0 0 0.15 0.03 0 0 0.03 0.08 0 0.17 0.32
Poland 0.25 0 0 0.1 0 0 0 0.01 0.02 0 0.12 0
Russia 0.02 0 0 0 0 0.04 0.02 0.09 0.08 0 0.06 0.03
Slovenia 0.02 0 0 0.02 0.08 0 0 0 0 0 0.01 0
Switzerland 0.08 0 0.17 0.15 0.05 0.11 0.17 0.17 0.18 0 0.13 0.08
United States 0.19 0.04 0.3 0.26 0.08 0 0 0.05 0.05 0 0.23 0.14
Big Air
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event
1.0
0.8
0.6
0.4
0.2
0.0

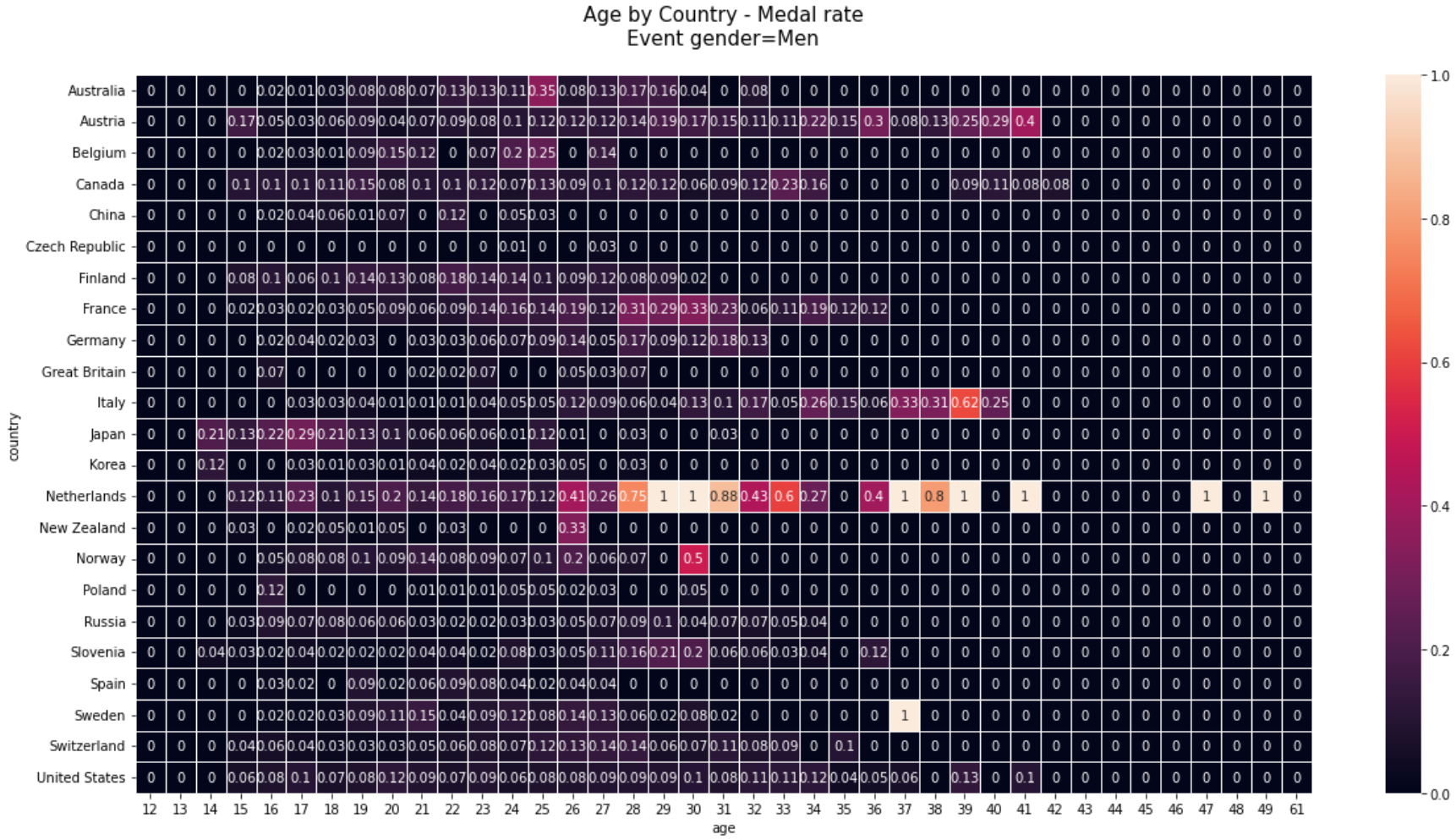

Event by Country - Participation count
Event gender=Women
country
Australia 46 96 4 444 42 4 20 140 72 2 170 316
Austria 88 30 88 112 226 42 316 1816 1294 20 148 1062
Canada 262 238 16 1116 226 52 186 1002 594 2 702 1446
China 54 174 6 962 48 0 64 426 272 8 56 34
Czech Republic 88 46 12 104 44 14 52 332 210 0 198 516
Finland 172 94 4 310 44 0 26 160 102 0 342 44
France 80 106 88 896 204 38 90 574 374 22 172 1834
Germany 84 42 40 220 96 30 206 1174 802 14 132 306
Italy 20 18 76 198 150 30 210 1062 676 14 62 758
Japan 234 160 24 1116 184 18 128 642 438 4 322 346
Netherlands 126 28 40 102 48 18 86 524 436 14 204 246
Norway 100 40 0 288 62 6 22 126 100 0 278 156
Poland 8 38 18 294 54 16 116 594 438 0 48 160
Russia 86 38 6 114 118 50 340 1562 1228 2 96 492
Slovenia 82 46 4 168 26 4 50 314 186 0 160 12
Switzerland 224 134 12 818 230 36 218 1286 906 0 392 1476
United States 350 414 60 1948 320 52 158 766 426 8 926 1418
Big Air
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event
1750
1500
1250
1000
750
500
250
0

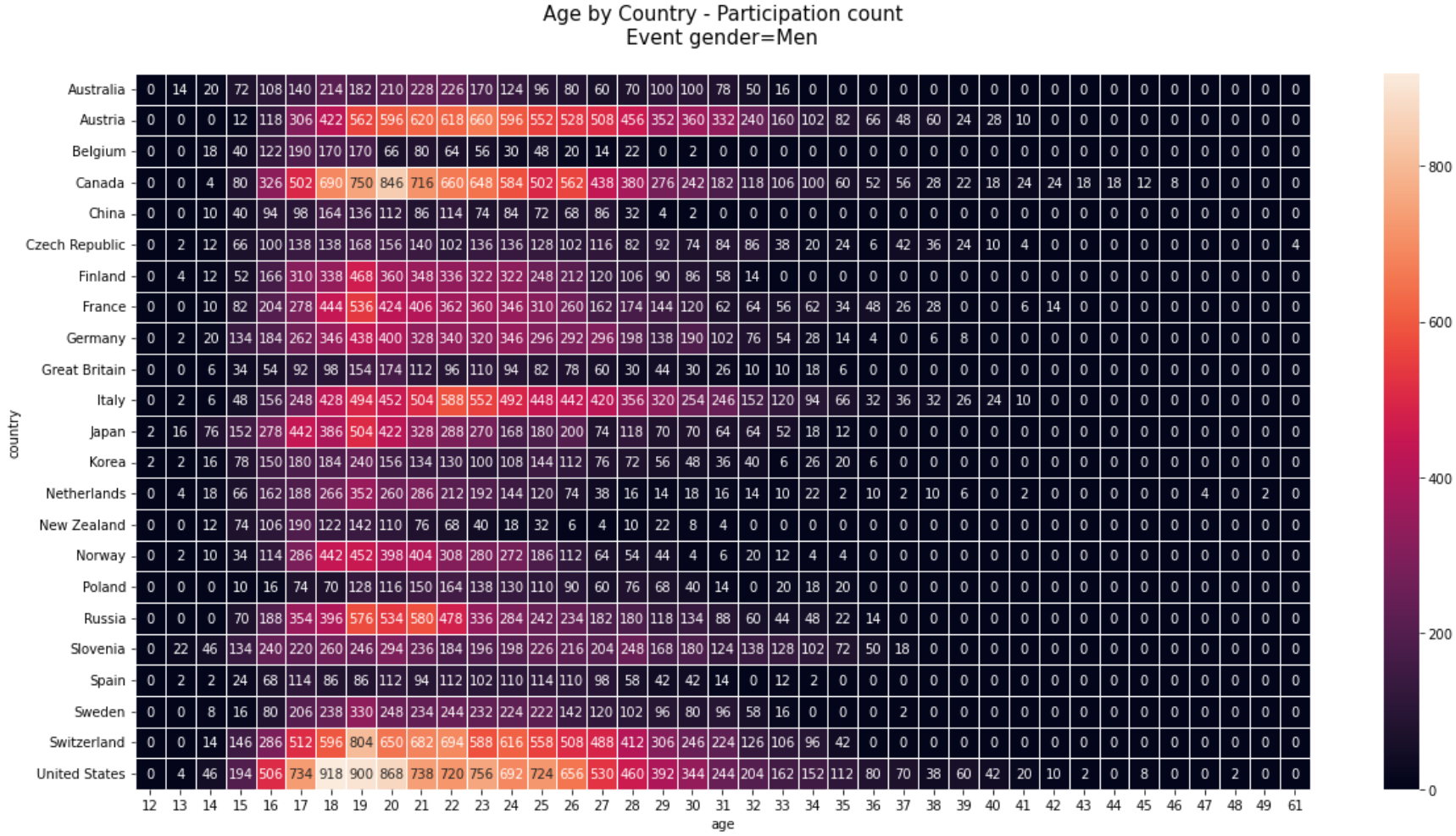

Event by Country - Participation rate
Event gender=Women
country
Australia 0.03 0.07 0 0.33 0.03 0 0.01 0.1 0.05 0 0.13 0.23
Austria 0.02 0.01 0.02 0.02 0.04 0.01 0.06 0.35 0.25 0 0.03 0.2
Canada 0.04 0.04 0 0.19 0.04 0.01 0.03 0.17 0.1 0 0.12 0.25
China 0.03 0.08 0 0.46 0.02 0 0.03 0.2 0.13 0 0.03 0.02
Czech Republic 0.05 0.03 0.01 0.06 0.03 0.01 0.03 0.21 0.13 0 0.12 0.32
Finland 0.13 0.07 0 0.24 0.03 0 0.02 0.12 0.08 0 0.26 0.03
France 0.02 0.02 0.02 0.2 0.05 0.01 0.02 0.13 0.08 0 0.04 0.41
Germany 0.03 0.01 0.01 0.07 0.03 0.01 0.07 0.37 0.25 0 0.04 0.1
Italy 0.01 0.01 0.02 0.06 0.05 0.01 0.06 0.32 0.21 0 0.02 0.23
Japan 0.06 0.04 0.01 0.31 0.05 0 0.04 0.18 0.12 0 0.09 0.1
Netherlands 0.07 0.01 0.02 0.05 0.03 0.01 0.05 0.28 0.23 0.01 0.11 0.13
Norway 0.08 0.03 0 0.24 0.05 0.01 0.02 0.11 0.08 0 0.24 0.13
Poland 0 0.02 0.01 0.16 0.03 0.01 0.07 0.33 0.25 0 0.03 0.09
Russia 0.02 0.01 0 0.03 0.03 0.01 0.08 0.38 0.3 0 0.02 0.12
Slovenia 0.08 0.04 0 0.16 0.02 0 0.05 0.3 0.18 0 0.15 0.01
Switzerland 0.04 0.02 0 0.14 0.04 0.01 0.04 0.22 0.16 0 0.07 0.26
United States 0.05 0.06 0.01 0.28 0.05 0.01 0.02 0.11 0.06 0 0.14 0.21
Big Air
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event
0.4
0.3
0.2
0.1
0.0

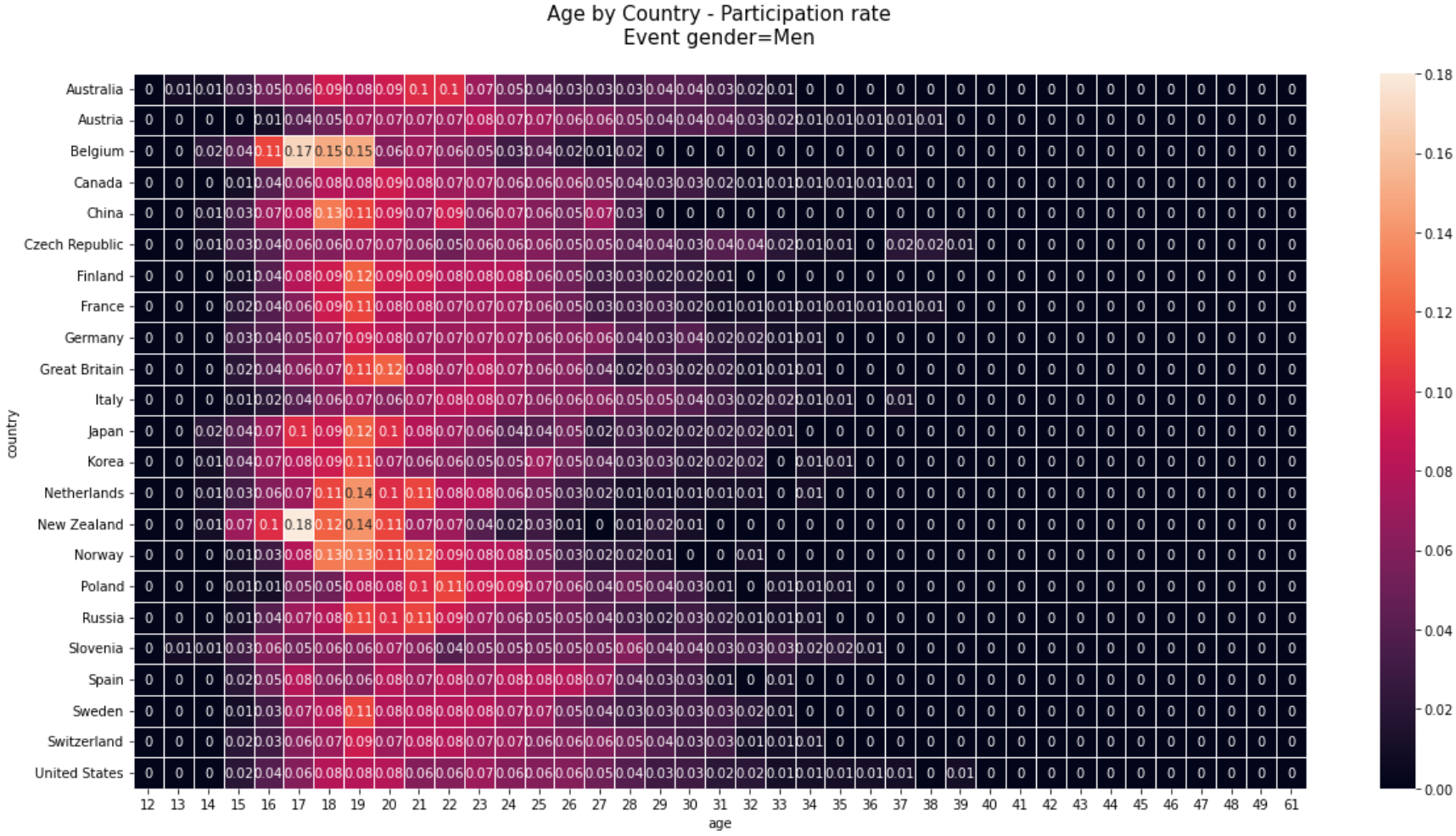

Event by Country - Medal rate
Event gender=Women - Country=United States
country
United States
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Big Air
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event

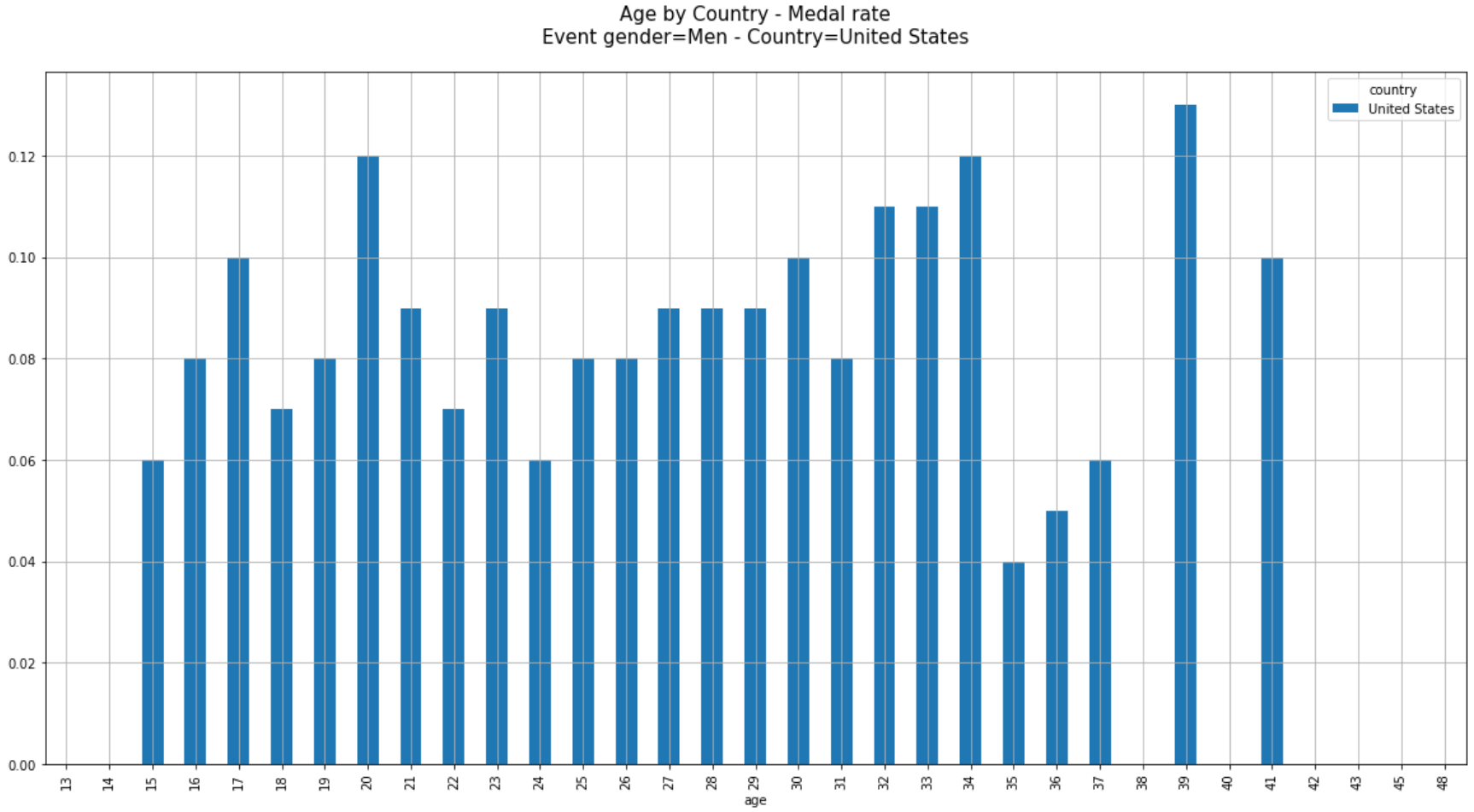

Event by Country - Participation count
Event gender=Women - Country=United States
country
United States
2000
1750
1500
1250
1000
750
500
250
0
Big Air
Freestyle Overall
Giant Slalom
Halfpipe
Overall
PAR-SBX Overall
Parallel
Parallel Giant Slalom
Parallel Slalom
Slalom
Slopestyle
Snowboardcross
event

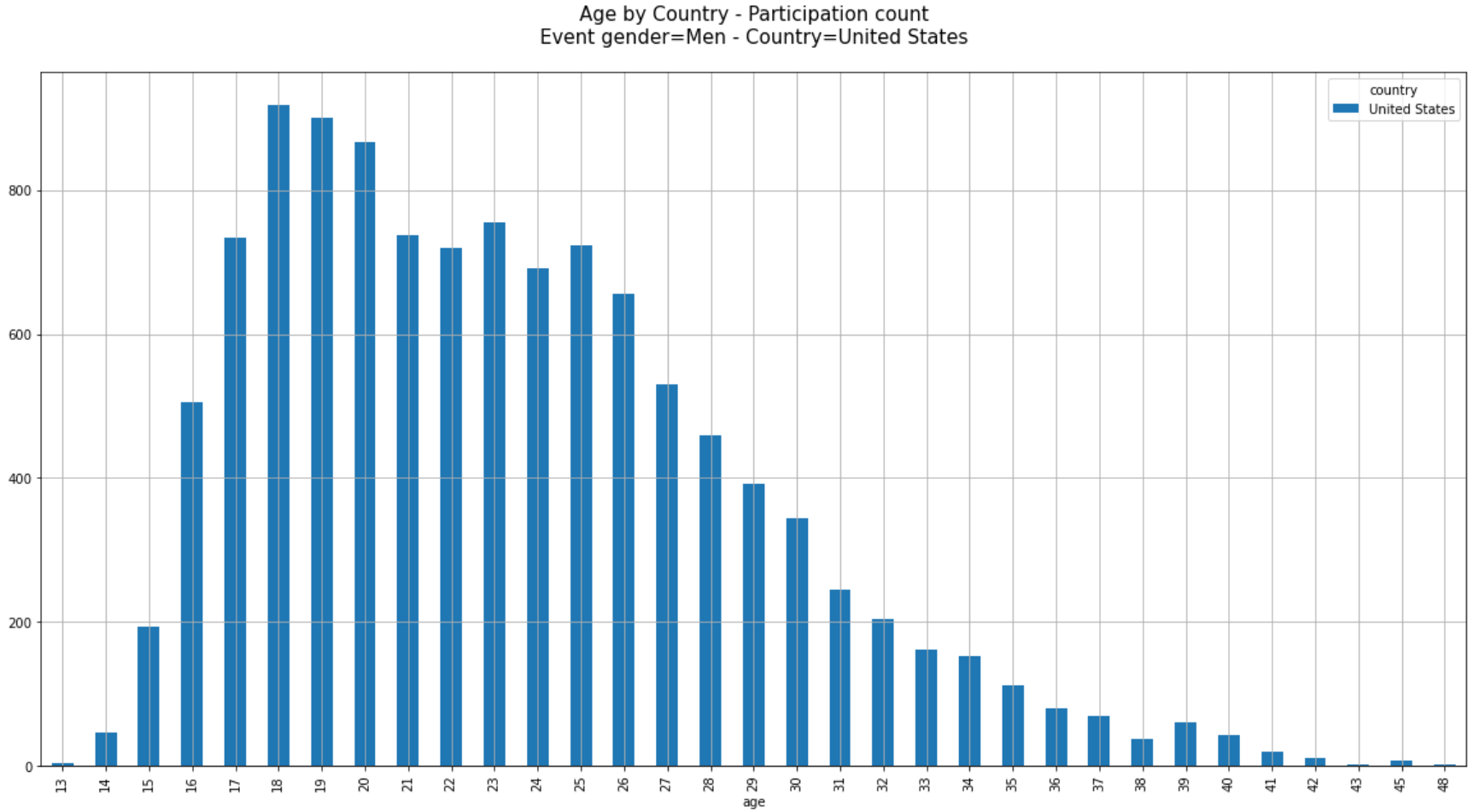
Analysis by Athlete Country of Origin and Age

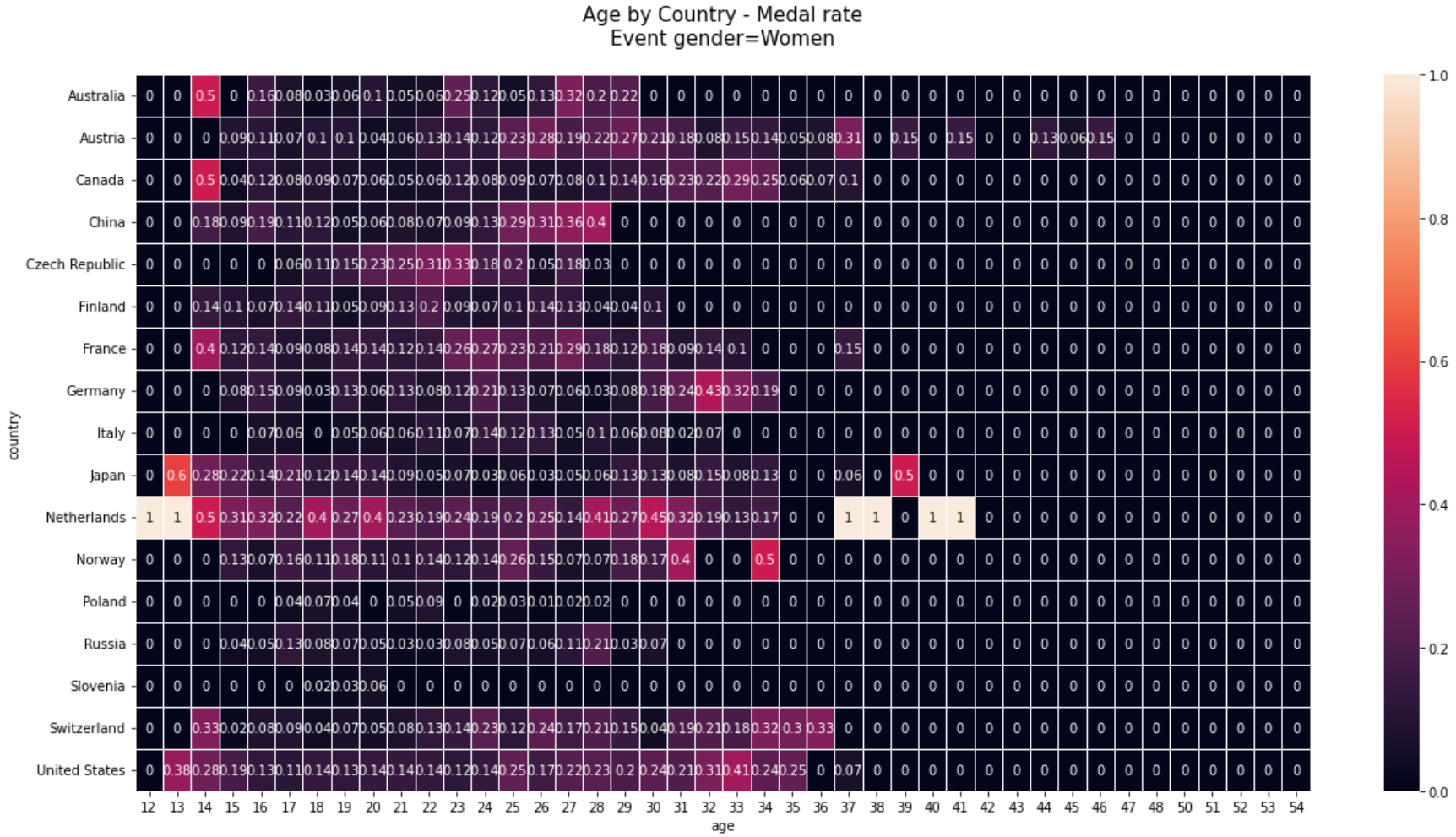


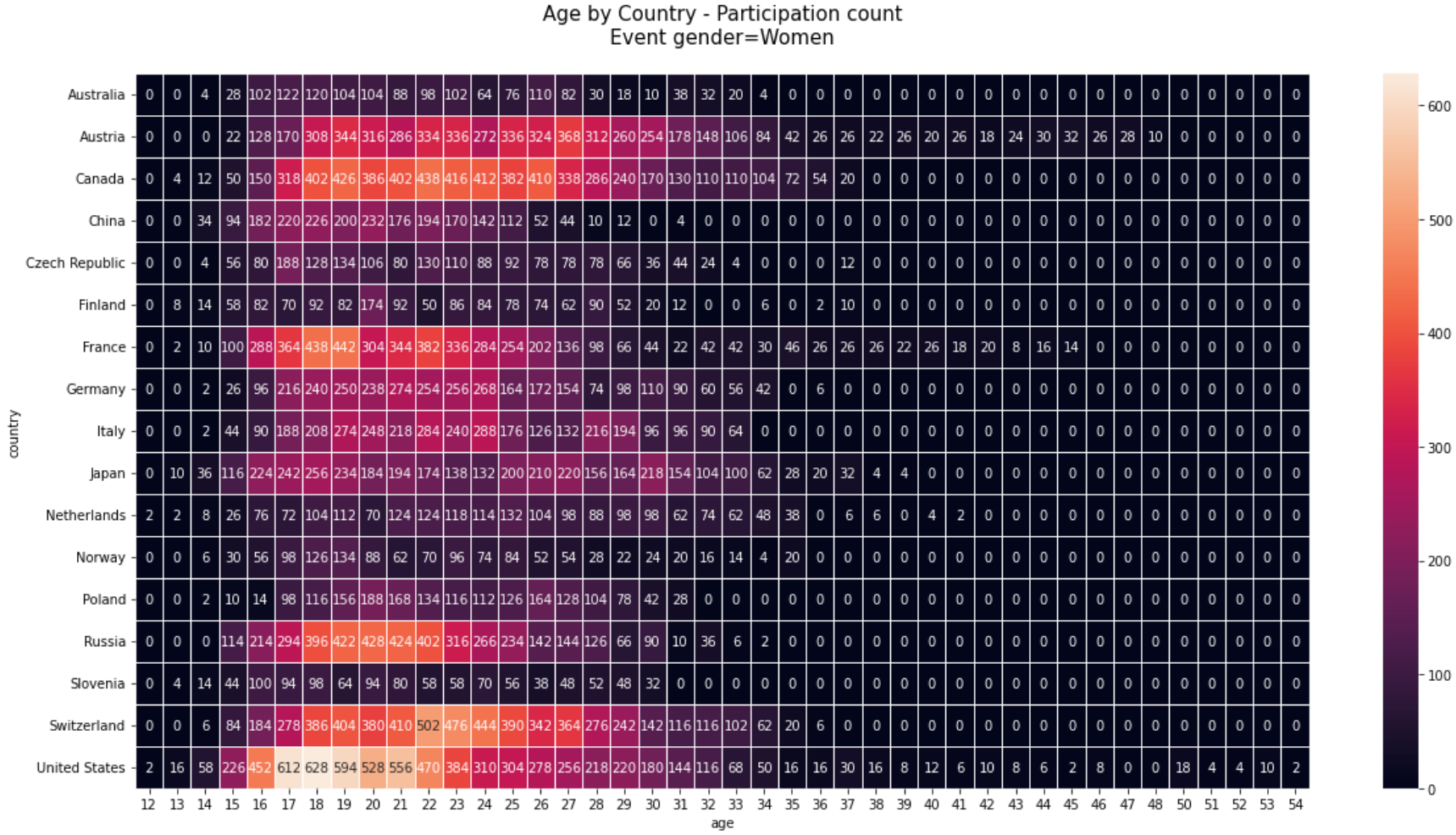

Age by Country - Participation count
Event gender=Men
country
Australia 0 14 20 72 108 140 214 182 210 228 226 170 124 96 80 60 70 100 100 78 50 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Austria 0 0 0 12 118 306 422 562 596 620 618 660 596 552 528 508 456 352 360 332 240 160 102 82 66 48 60 24 28 10 0 0 0 0 0 0 0 0
Belgium 0 0 18 40 122 190 170 170 66 80 64 56 30 48 20 14 22 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Canada 0 0 4 80 326 502 690 750 846 716 660 648 584 502 562 438 380 276 242 182 118 106 100 60 52 56 28 22 18 24 24 18 18 12 8 0 0 0 0
China 0 0 10 40 94 98 164 136 112 86 114 74 84 72 68 86 32 4 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Czech Republic 0 2 12 66 100 138 138 168 156 140 102 136 136 128 102 116 82 92 74 84 86 38 20 24 6 42 36 24 10 4 0 0 0 0 0 0 0 0 4
Finland 0 4 12 52 166 310 338 468 360 348 336 322 322 248 212 120 106 90 86 58 14 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
France 0 0 10 82 204 278 444 536 424 406 362 360 346 310 260 162 174 144 120 62 64 56 62 34 48 26 28 0 0 6 14 0 0 0 0 0 0 0 0
Germany 0 2 20 134 184 262 346 438 400 328 340 320 346 296 292 296 198 138 190 102 76 54 28 14 4 0 6 8 0 0 0 0 0 0 0 0 0 0
Great Britain 0 0 6 34 54 92 98 154 174 112 96 110 94 82 78 60 30 44 30 26 10 10 18 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Italy 0 2 6 48 156 248 428 494 452 504 588 552 492 448 442 420 356 320 254 246 152 120 94 66 32 36 32 26 24 10 0 0 0 0 0 0 0 0
Japan 2 16 76 152 278 442 386 504 422 328 288 270 168 180 200 74 118 70 70 64 64 52 18 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Korea 2 2 16 78 150 180 184 240 156 134 130 100 108 144 112 76 72 56 48 36 40 6 26 20 6 0 0 0 0 0 0 0 0 0 0 0 0 0
Netherlands 0 4 18 66 162 188 266 352 260 286 212 192 144 120 74 38 16 14 18 16 14 10 22 2 10 2 10 6 0 2 0 0 0 0 0 4 0 2 0
New Zealand 0 0 12 74 106 190 122 142 110 76 68 40 18 32 6 4 10 22 8 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Norway 0 2 10 34 114 286 442 452 398 404 308 280 272 186 112 64 54 44 4 6 20 12 4 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Poland 0 0 0 10 16 74 70 128 116 150 164 138 130 110 90 60 76 68 40 14 0 20 18 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Russia 0 0 0 70 188 354 396 576 534 580 478 336 284 242 234 182 180 118 134 88 60 44 48 22 14 0 0 0 0 0 0 0 0 0 0 0 0 0
Slovenia 0 22 46 134 240 220 260 246 294 236 184 196 198 226 216 204 248 168 180 124 138 128 102 72 50 18 0 0 0 0 0 0 0 0 0 0 0 0
Spain 0 2 2 24 68 114 86 86 112 94 112 102 110 114 110 98 58 42 42 14 0 12 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Sweden 0 0 8 16 80 206 238 330 248 234 244 232 224 222 142 120 102 96 80 96 58 16 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
Switzerland 0 0 14 146 286 512 596 804 650 682 694 588 616 558 508 488 412 306 246 224 126 106 96 42 0 0 0 0 0 0 0 0 0 0 0 0 0 0
United States 0 4 46 194 506 734 918 900 868 738 720 756 692 724 656 530 460 392 344 244 204 162 152 112 80 70 38 60 42 20 10 2 0 8 0 0 2 0 0
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 61
age
800
600
400
200
0

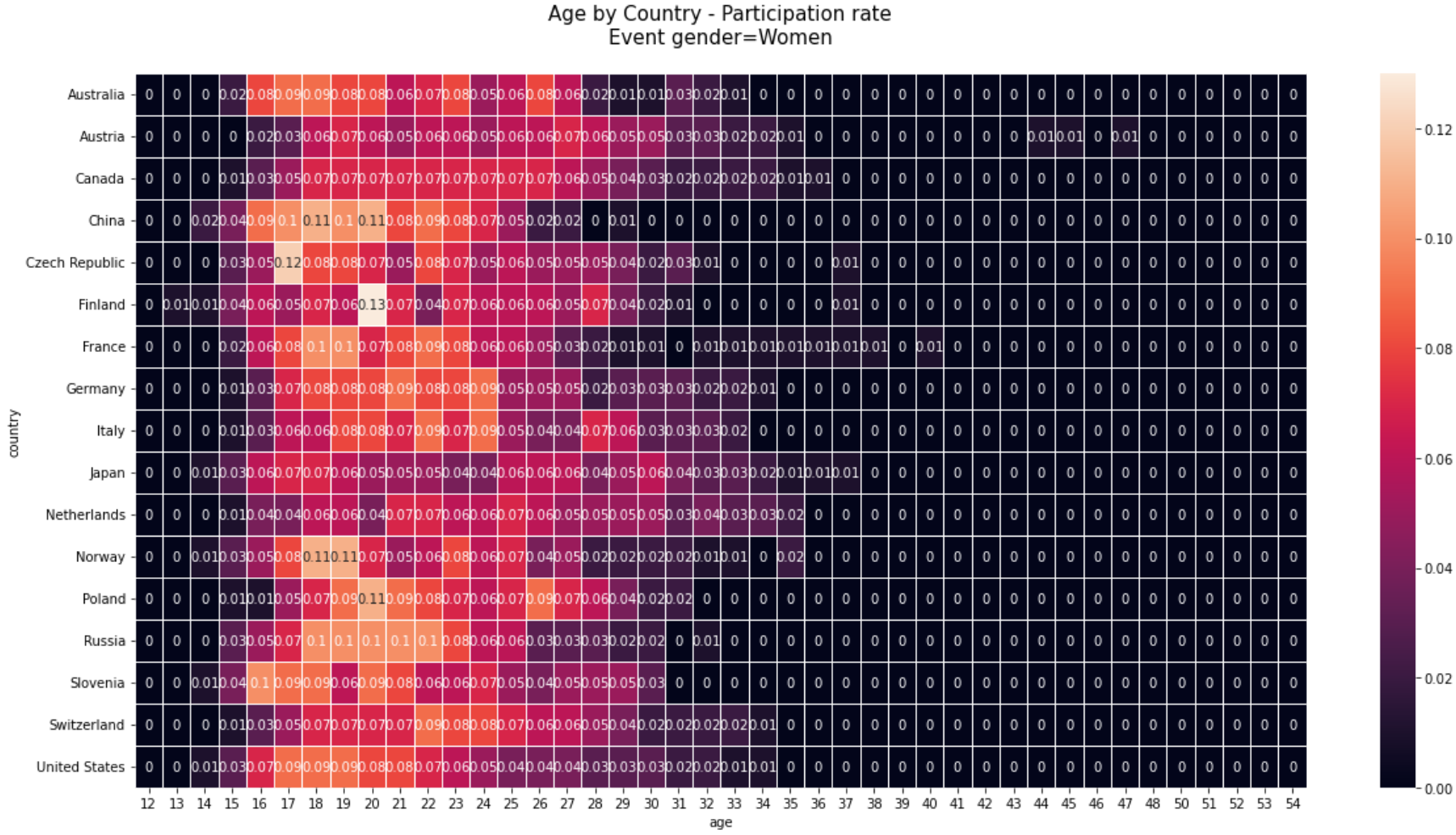

Age by Country - Participation rate
Event gender=Men
country
Australia 0 0.01 0.01 0.03 0.05 0.06 0.09 0.08 0.09 0.1 0.1 0.07 0.05 0.04 0.03 0.03 0.03 0.04 0.04 0.03 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Austria 0 0 0 0 0.01 0.04 0.05 0.07 0.07 0.07 0.07 0.08 0.07 0.07 0.06 0.06 0.05 0.04 0.04 0.04 0.03 0.02 0.01 0.01 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0
Belgium 0 0 0.02 0.04 0.11 0.17 0.15 0.15 0.06 0.07 0.06 0.05 0.03 0.04 0.02 0.01 0.02 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Canada 0 0 0 0.01 0.04 0.06 0.08 0.08 0.09 0.08 0.07 0.07 0.06 0.06 0.06 0.05 0.04 0.03 0.03 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0
China 0 0 0.01 0.03 0.07 0.08 0.13 0.11 0.09 0.07 0.09 0.06 0.07 0.06 0.05 0.07 0.03 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Czech Republic 0 0 0.01 0.03 0.04 0.06 0.06 0.07 0.07 0.06 0.05 0.06 0.06 0.06 0.05 0.05 0.04 0.04 0.03 0.04 0.04 0.02 0.01 0.01 0 0.02 0.02 0.01 0 0 0 0 0 0 0 0 0 0
Finland 0 0 0 0.01 0.04 0.08 0.09 0.12 0.09 0.09 0.08 0.08 0.08 0.06 0.05 0.03 0.03 0.02 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
France 0 0 0 0.02 0.04 0.06 0.09 0.11 0.08 0.08 0.07 0.07 0.07 0.06 0.05 0.03 0.03 0.03 0.02 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0
Germany 0 0 0 0.03 0.04 0.05 0.07 0.09 0.08 0.07 0.07 0.07 0.07 0.06 0.06 0.06 0.04 0.03 0.04 0.02 0.02 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Great Britain 0 0 0 0.02 0.04 0.06 0.07 0.11 0.12 0.08 0.07 0.08 0.07 0.06 0.06 0.04 0.02 0.03 0.02 0.02 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Italy 0 0 0 0.01 0.02 0.04 0.06 0.07 0.06 0.07 0.08 0.08 0.07 0.06 0.06 0.06 0.05 0.05 0.04 0.03 0.02 0.02 0.01 0.01 0 0.01 0 0 0 0 0 0 0 0 0 0 0 0
Japan 0 0 0.02 0.04 0.07 0.1 0.09 0.12 0.1 0.08 0.07 0.06 0.04 0.04 0.05 0.02 0.03 0.02 0.02 0.02 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Korea 0 0 0.01 0.04 0.07 0.08 0.09 0.11 0.07 0.06 0.06 0.05 0.05 0.07 0.05 0.04 0.03 0.03 0.02 0.02 0.02 0 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Netherlands 0 0 0.01 0.03 0.06 0.07 0.11 0.14 0.1 0.11 0.08 0.08 0.06 0.05 0.03 0.02 0.01 0.01 0.01 0.01 0.01 0 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
New Zealand 0 0 0.01 0.07 0.1 0.18 0.12 0.14 0.11 0.07 0.07 0.04 0.02 0.03 0.01 0 0.01 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Norway 0 0 0 0.01 0.03 0.08 0.13 0.13 0.11 0.12 0.09 0.08 0.08 0.05 0.03 0.02 0.02 0.01 0 0 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Poland 0 0 0 0.01 0.01 0.05 0.05 0.08 0.08 0.1 0.11 0.09 0.09 0.07 0.06 0.04 0.05 0.04 0.03 0.01 0 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Russia 0 0 0 0.01 0.04 0.07 0.08 0.11 0.1 0.11 0.09 0.07 0.06 0.05 0.05 0.04 0.03 0.02 0.03 0.02 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slovenia 0 0.01 0.01 0.03 0.06 0.05 0.06 0.06 0.07 0.06 0.04 0.05 0.05 0.05 0.05 0.05 0.06 0.04 0.04 0.03 0.03 0.03 0.02 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0
Spain 0 0 0 0.02 0.05 0.08 0.06 0.06 0.08 0.07 0.08 0.07 0.08 0.08 0.08 0.07 0.04 0.03 0.03 0.01 0 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Sweden 0 0 0 0.01 0.03 0.07 0.08 0.11 0.08 0.08 0.08 0.08 0.07 0.07 0.05 0.04 0.03 0.03 0.03 0.03 0.02 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Switzerland 0 0 0 0.02 0.03 0.06 0.07 0.09 0.07 0.08 0.08 0.07 0.07 0.06 0.06 0.06 0.05 0.04 0.03 0.03 0.01 0.01 0.01 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
United States 0 0 0 0.02 0.04 0.06 0.08 0.08 0.08 0.06 0.06 0.07 0.06 0.06 0.06 0.05 0.04 0.03 0.03 0.02 0.02 0.01 0.01 0.01 0.01 0.01 0 0.01 0 0 0 0 0 0 0 0 0 0 0
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 61
age
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00

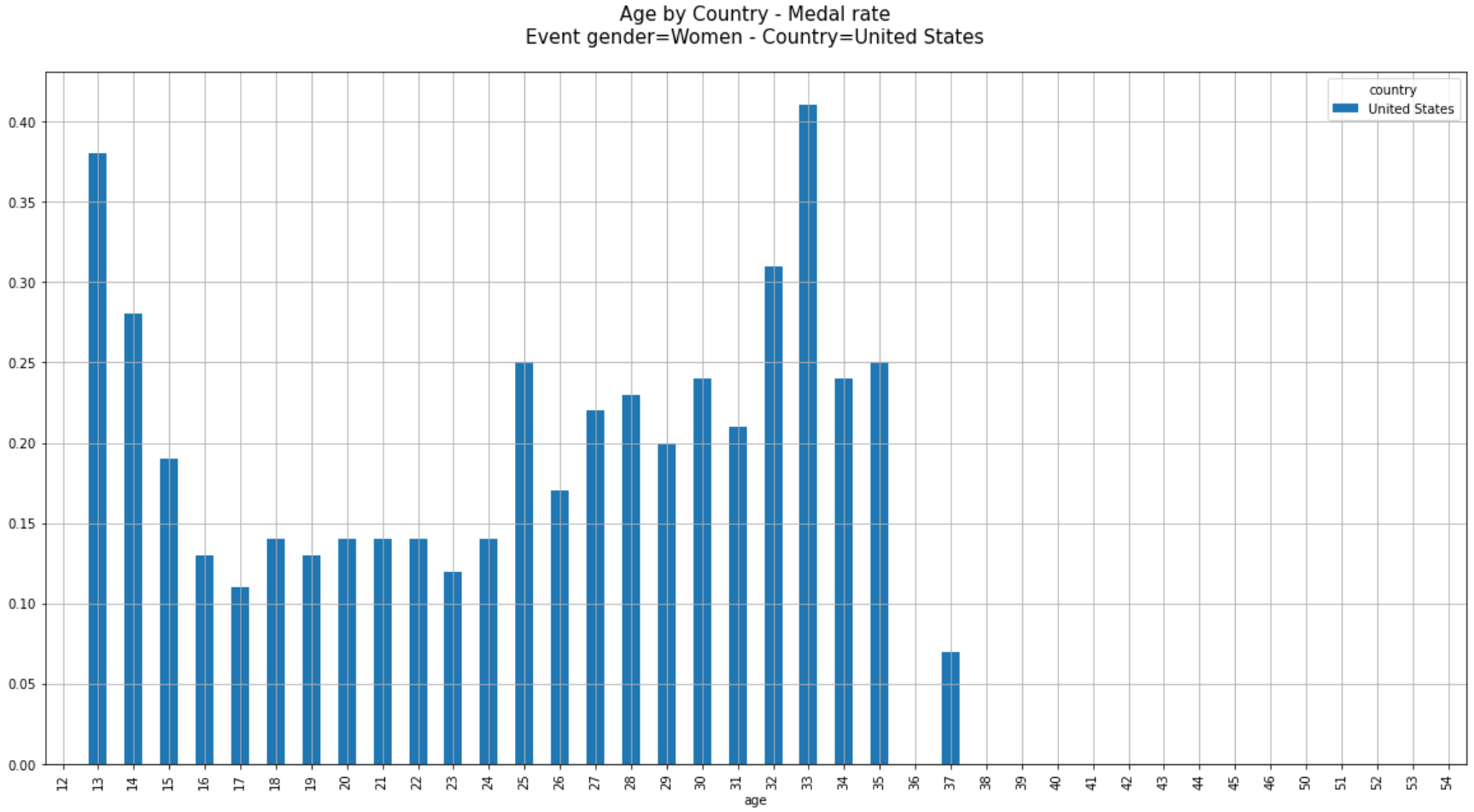

Age by Country - Medal rate
Event gender=Men - Country=United States
country
United States
0.12
0.10
0.08
0.06
0.04
0.02
0.00
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
45
48
age

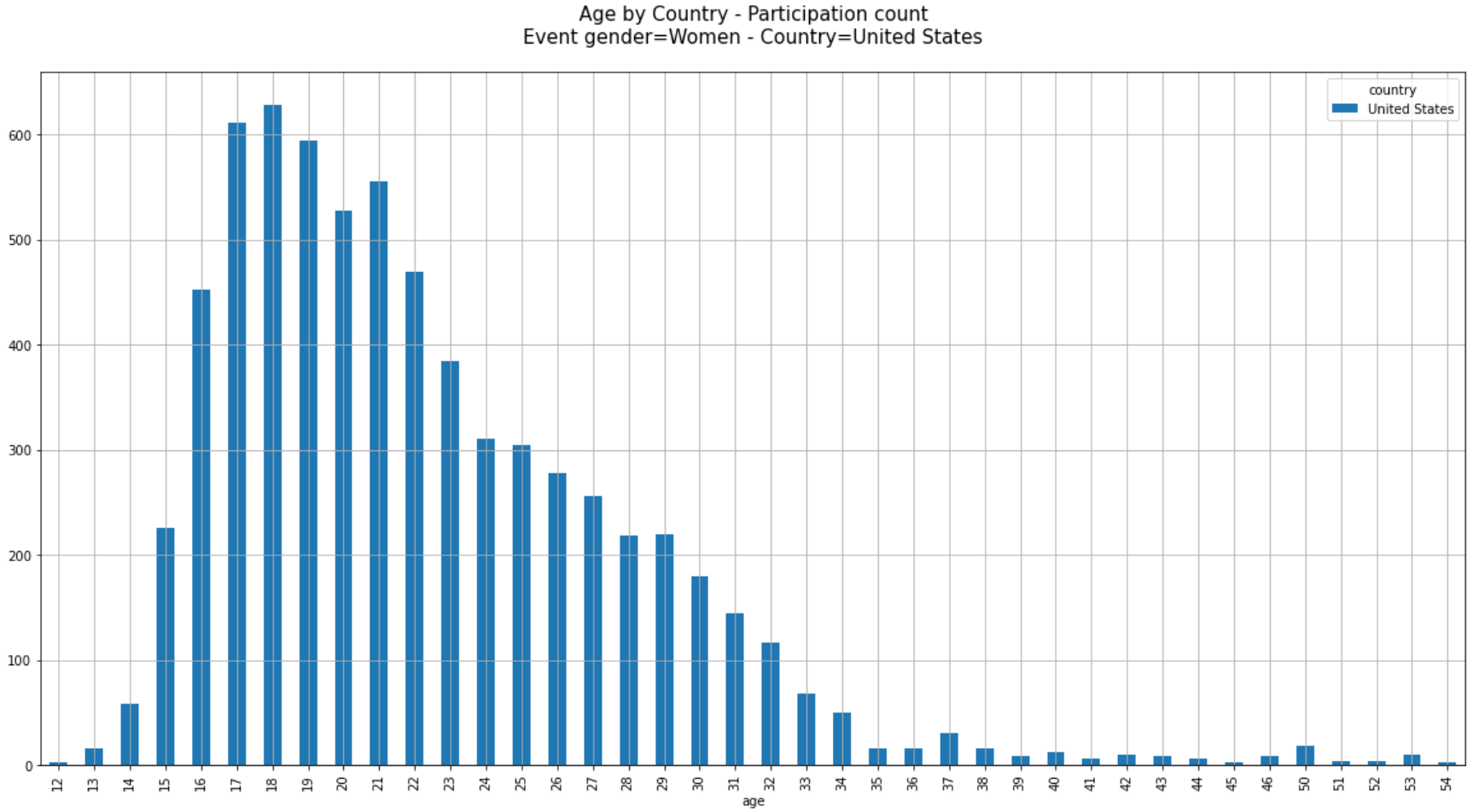

Age by Country - Participation count
Event gender=Men - Country=United States
country
United States
800
600
400
200
0
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
45
48
age


Age by Country - Medal rate
Event gender=Women
country
Australia
Austria
Canada
China
Czech Republic
Finland
France
Germany
Italy
Japan
Netherlands
Norway
Poland
Russia
Slovenia
Switzerland
United States
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50 51 52 53 54
age
1.0
0.8
0.6
0.4
0.2
0.0

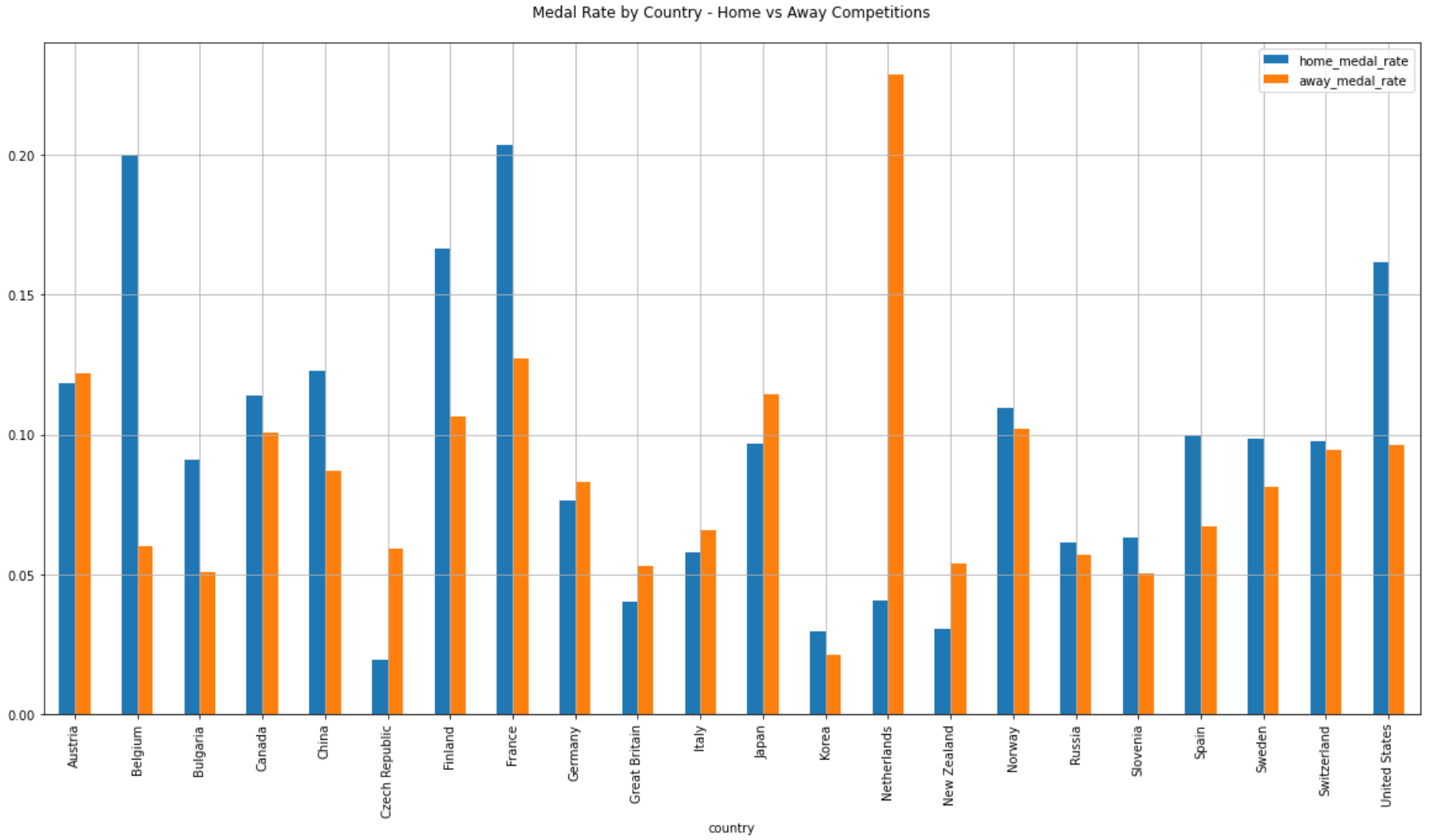

Age by Country - Participation count
Event gender=Women
country
Australia 0 0 4 28 102 122 120 104 104 88 98 102 64 76 110 82 30 18 10 38 32 20 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Austria 0 0 0 22 128 170 308 344 316 286 334 336 272 336 324 368 312 260 254 178 148 106 84 42 26 26 22 26 20 26 18 24 30 32 26 28 10 0 0 0 0 0
Canada 0 4 12 50 150 318 402 426 386 402 438 416 412 382 410 338 286 240 170 130 110 110 104 72 54 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
China 0 0 34 94 182 220 226 200 232 176 194 170 142 112 52 44 10 12 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Czech Republic 0 0 4 56 80 188 128 134 106 80 130 110 88 92 78 78 78 66 36 44 24 4 0 0 0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Finland 0 8 14 58 82 70 92 82 174 92 50 86 84 78 74 62 90 52 20 12 0 0 6 0 2 10 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
France 0 2 10 100 288 364 438 442 304 344 382 336 284 254 202 136 98 66 44 22 42 42 30 46 26 26 26 22 26 18 20 8 16 14 0 0 0 0 0 0 0 0 0 0
Germany 0 0 2 26 96 216 240 250 238 274 254 256 268 164 172 154 74 98 110 90 60 56 42 0 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Italy 0 0 2 44 90 188 208 274 248 218 284 240 288 176 126 132 216 194 96 96 90 64 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Japan 0 10 36 116 224 242 256 234 184 194 174 138 132 200 210 220 156 164 218 154 104 100 62 28 20 32 4 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Netherlands 2 2 8 26 76 72 104 112 70 124 124 118 114 132 104 98 88 98 98 62 74 62 48 38 0 6 6 0 4 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Norway 0 0 6 30 56 98 126 134 88 62 70 96 74 84 52 54 28 22 24 20 16 14 4 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Poland 0 0 2 10 14 98 116 156 188 168 134 116 112 126 164 128 104 78 42 28 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Russia 0 0 0 114 214 294 396 422 428 424 402 316 266 234 142 144 126 66 90 10 36 6 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Slovenia 0 4 14 44 100 94 98 64 94 80 58 58 70 56 38 48 52 48 32 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Switzerland 0 0 6 84 184 278 386 404 380 410 502 476 444 390 342 364 276 242 142 116 116 102 62 20 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
United States 2 16 58 226 452 612 628 594 528 556 470 384 310 304 278 256 218 220 180 144 116 68 50 16 16 30 16 8 12 6 10 8 6 2 8 0 0 18 4 4 10 2
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50 51 52 53 54
age
600
500
400
300
200
100
0


Age by Country - Participation rate
Event gender=Women
country
Australia
Austria
Canada
China
Czech Republic
Finland
France
Germany
Italy
Japan
Netherlands
Norway
Poland
Russia
Slovenia
Switzerland
United States
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 50 51 52 53 54
age
0.12
0.10
0.08
0.06
0.04
0.02
0.00


Age by Country - Medal rate
Event gender=Women - Country=United States
country
United States
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 50 51 52 53 54
age


Age by Country - Participation count
Event gender=Women - Country=United States
country
United States
600
500
400
300
200
100
0
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
50
51
52
53
54
age

Analysis of Success by Country for Home vs Away Competition



Observations and Conclusions

By observing the Country by Sport charts we can see the following patterns:

- There are wide variations in terms of event focus by country of origin as illustrated by the participation count by sport graph.
- The most popular sports by participation rate are Snowboardcross, Slopestyle, Parallel Giant Slalom, Halfpipe, Parallel Slalom and Big Air.
- Big Air has a significantly higher participation rate in Men and Women.
- Some countries show higher medal rates than others.
- The Netherlands appears to be the most successful using medal rate as the metric.
- The variation in country medal rate indicates the significance of country of origin as a factor that predicts success.
- Based on this analysis, countries and sporting events will be factored in when making predictions.

The patterns that arise from the Country by Age group statistics are the following:

- Peak ages, with highest medal rates show considerable variance with no explicit peak ages overall. This is an indicator that there is no clear peak age for any sport and event.
- Very high medal winning rates are observed by the Netherlands in higher ages.
- High medal rates appear to be more frequent in age ranges with lower participation counts, indicating potentially smaller competition.
- The highest participation rate by country follows a similar distribution across countries, but is centered around ages 19-20. This means most participants for this sport are around that age.
- The participation rate for women has a wider variance compared to the one for men and ages are a little bit more evenly distributed.

Finally, going over the home competition success plot, the following inferences can be made:

- Most countries are more successful when their athletes are competing at home versus when they are competing at a competition held in a foreign country.
- Netherlands exhibits an unusually high away winning rate but is generally an exception. This is likely due to the limited number of competitions held at home ground and the relatively small number of athletes the country has compared to other nations.
- For the United States in particular home competitions are much more successful.
- This is an indicator that a feature capturing whether an upcoming competition is being held at home versus away can provide valuable information when predicting the probability of winning a medal.

This concludes the exploratory data analysis portion of the report. In the next section we are going to discuss the Machine Learning methods used and the results of their predictive power.

Machine Learning for Predictions

As discussed earlier, for the machine learning phase we are going to treat the events in our dataset in two different ways. In the first approach we are going to utilize every event independently and in the second one we are going to leverage previous athlete performance by combining the events and forming a time series dataset where every row will represent the athlete's outcome in one event and the features of the three events that precede it.

In both cases we are going to predict the rank and the won_medal variable by using regression and classification respectively. The train/test split will be 80% and 20% respectively and we will use 5-fold Cross-Validation for hyperparameter tuning.

Once again, for the purpose of this study we are using the Snowboarding dataset and the event of Snowboardcross. The same process can be applied to any sport and event.

Classification

For the classification problem we developed two machine learning models and trained them on both the independent event dataset and the time series dataset and used the won_medal variable as our response variable. Rank of the final event was dropped from the dataset as it directly corresponds to the medal. The two models are a Random Forest Classifier and a Gradient Boosting Classifier. The performance results from both models are summarized below.

Independent Event Data

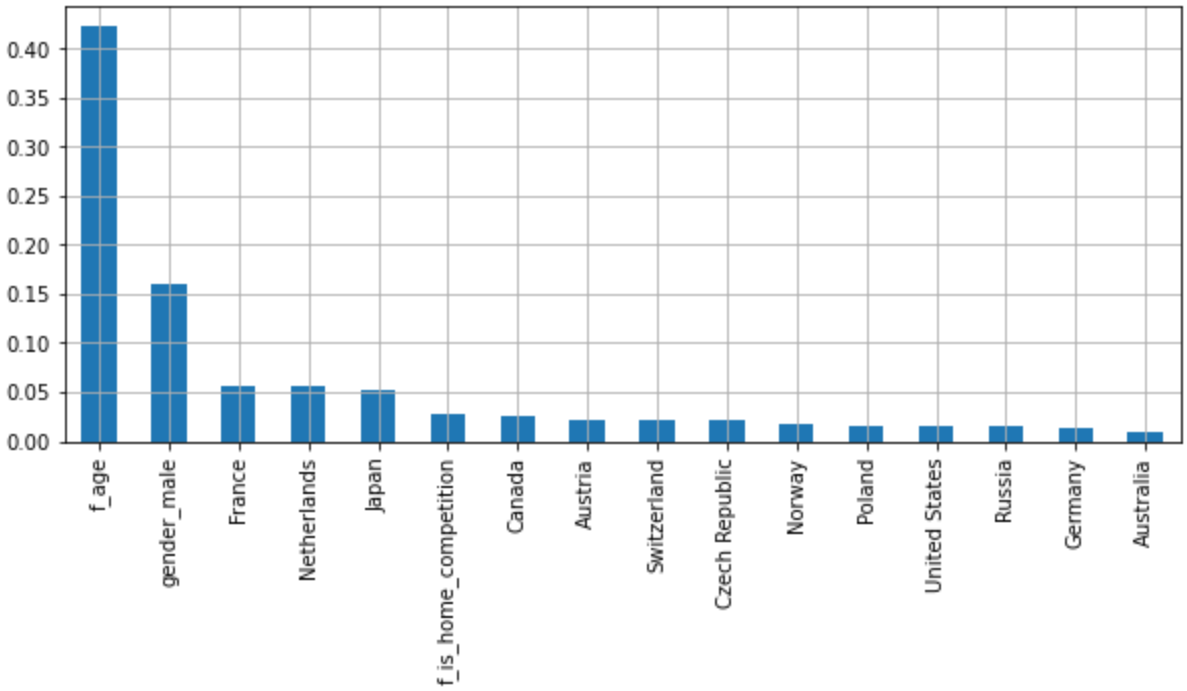
Random Forest Classifier - Independent Event Data					
Training Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.96	0.74	0.83	12689
	True - 1	0.18	0.66	0.29	1127
	Accuracy	-	-	0.73	13816
	Macro Average	0.57	0.70	0.56	13816
	Weighted Average	0.90	0.73	0.79	13816
Test Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.95	0.71	0.81	3152
	True - 1	0.16	0.57	0.25	302
	Accuracy	-	-	0.70	3454
	Macro Average	0.55	0.64	0.53	3454
	Weighted Average	0.88	0.70	0.76	3454

Training Set		Predicted N	Predicted P
	Actual Negative	9353	3336
	Actual Positive	379	748
Test Set		Predicted N	Predicted P
	Actual Negative	2253	899
	Actual Positive	130	172

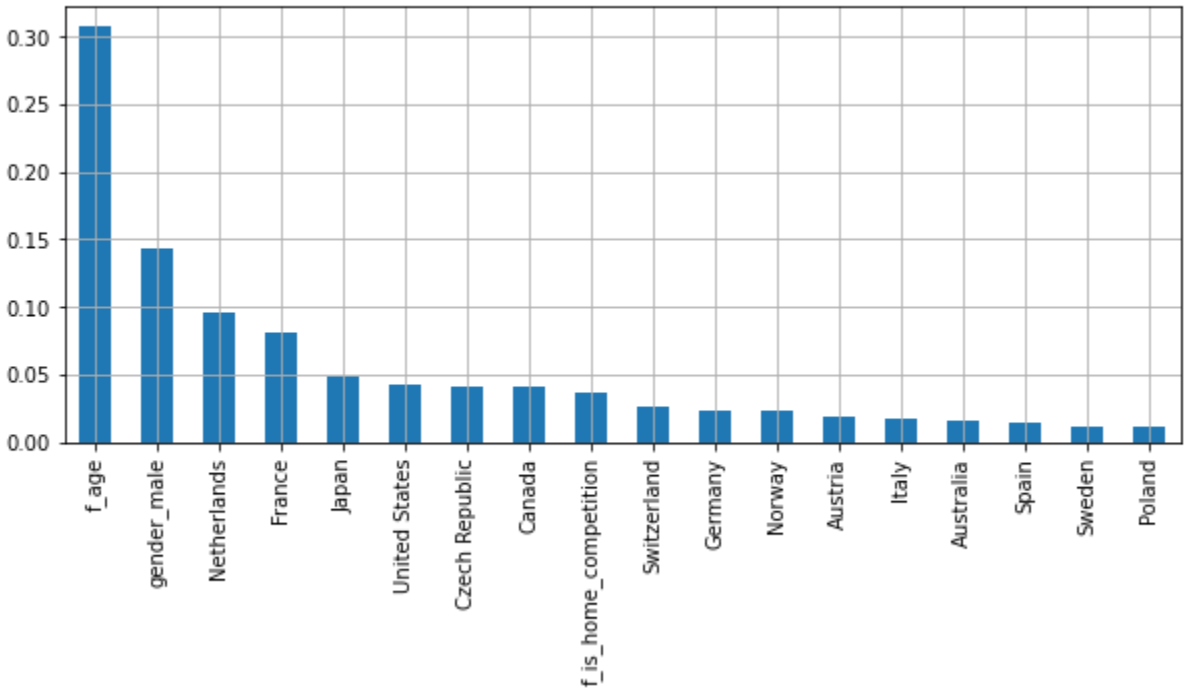
Gradient Boosting Classifier - Independent Event Data					
Training Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.96	0.71	0.82	12689
	True - 1	0.17	0.67	0.27	1127
	Accuracy	-	-	0.71	13816
	Macro Average	0.57	0.69	0.55	13816
	Weighted Average	0.90	0.71	0.77	13816
Test Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.95	0.70	0.80	3152
	True - 1	0.16	0.59	0.25	302
	Accuracy	-	-	0.69	3454
	Macro Average	0.55	0.64	0.53	3454
	Weighted Average	0.88	0.69	0.76	3454

Training Set		Predicted N	Predicted P
	Actual Negative	9063	3626
	Actual Positive	373	754
Test Set		Predicted N	Predicted P
	Actual Negative	2205	947
	Actual Positive	124	178

Random Forest Classifier - Independent Event Data - Feature Importances



Gradient Boosting Classifier - Independent Event Data - Feature Importances



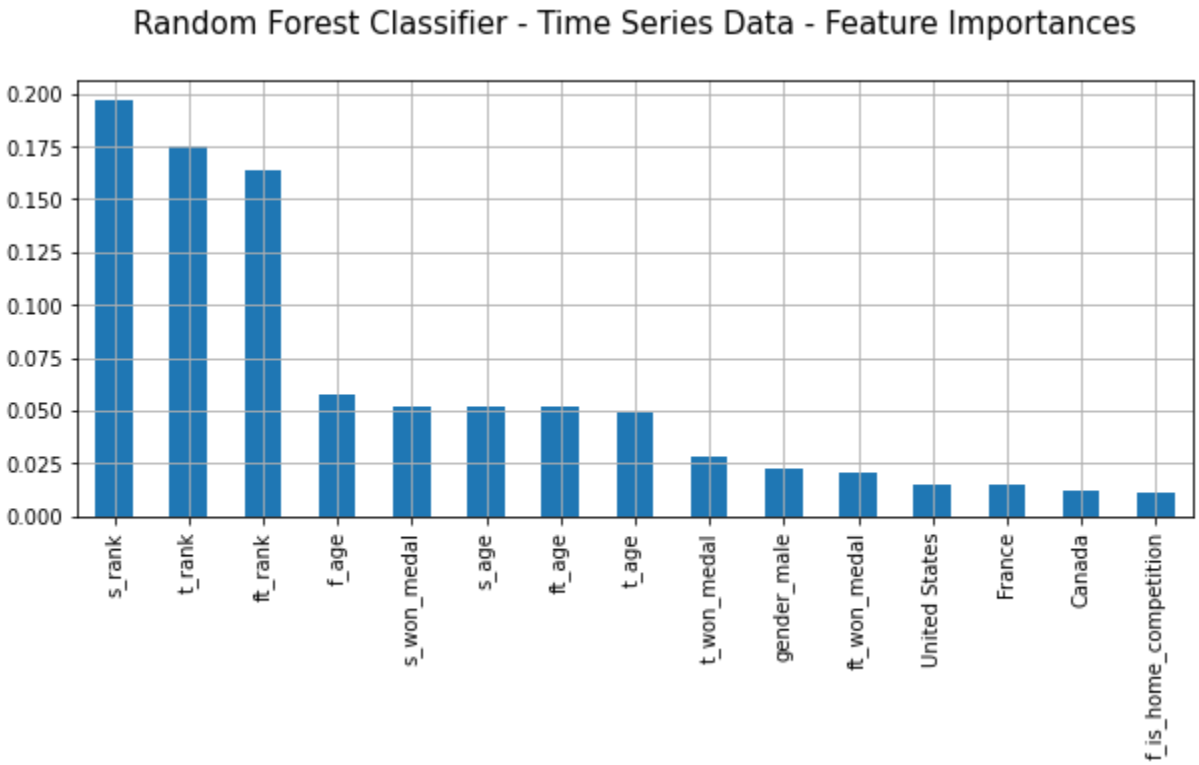
Time Series Data

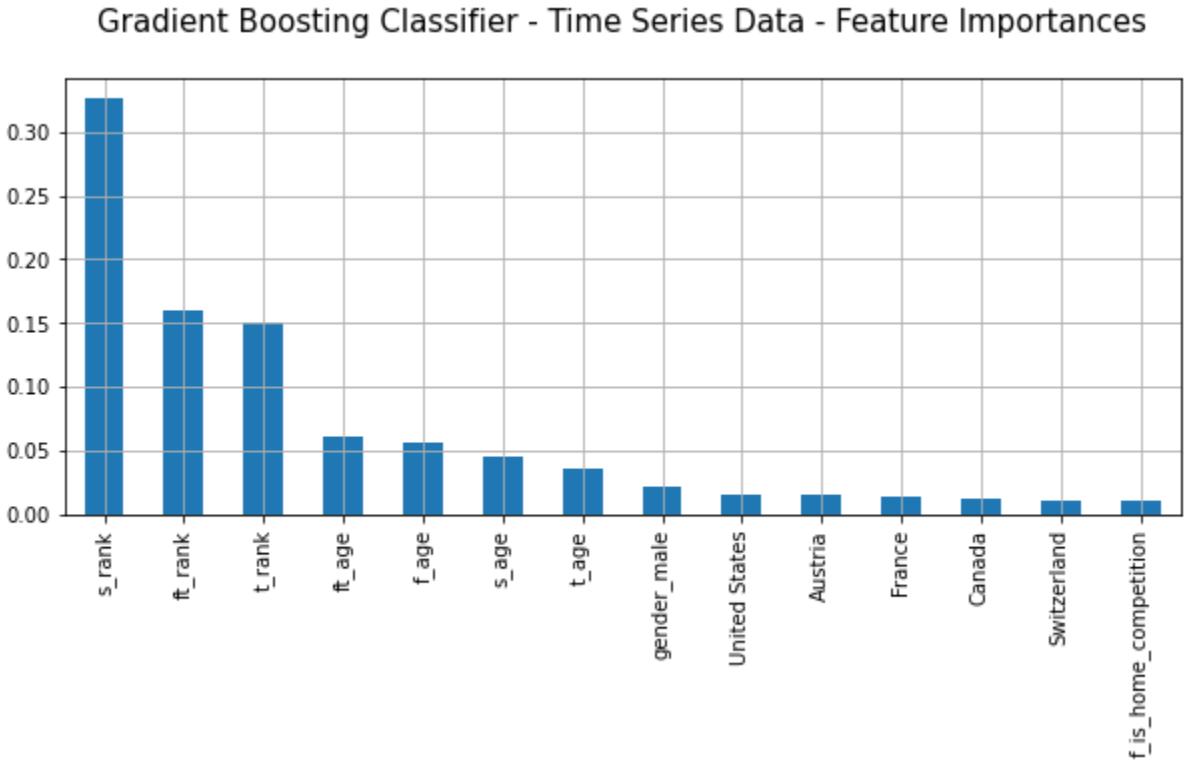
Random Forest Classifier - Time Series Data					
Training Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.99	0.97	0.98	11431
	True - 1	0.75	0.94	0.83	1097
	Accuracy	-	-	0.97	12528
	Macro Average	0.87	0.96	0.91	12528
	Weighted Average	0.97	0.97	0.97	12528
Test Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.95	0.95	0.95	2871
	True - 1	0.42	0.42	0.42	262
	Accuracy	-	-	0.90	3133
	Macro Average	0.68	0.68	0.68	3133
	Weighted Average	0.90	0.90	0.90	3133

Training Set		Predicted N	Predicted P
	Actual Negative	11085	346
	Actual Positive	64	1033
Test Set		Predicted N	Predicted P
	Actual Negative	2720	151
	Actual Positive	152	110

Gradient Boosting Classifier - Time Series Data					
Training Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	1.00	1.00	1.00	11431
	True - 1	0.97	1.00	0.98	1097
	Accuracy	-	-	1.00	12528
	Macro Average	0.98	1.00	0.99	12528
	Weighted Average	1.00	1.00	1.00	12528
Test Set Results	Medal Won	Precision	Recall	F1-Score	No. Datapoints
	False - 0	0.94	0.96	0.95	2871
	True - 1	0.39	0.27	0.32	262
	Accuracy	-	-	0.90	3133
	Macro Average	0.66	0.62	0.63	3133
	Weighted Average	0.89	0.90	0.90	3133

Training Set		Predicted N	Predicted P
	Actual Negative	11397	34
	Actual Positive	0	1097
Test Set		Predicted N	Predicted P
	Actual Negative	2758	113
	Actual Positive	191	71


Random Forest Classifier - Time Series Data - Feature Importances
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
s_rank
t_rank
ft_rank
f_age
s_won_medal
s_age
ft_age
t_age
t_won_medal
gender_male
ft_won_medal
United States
France
Canada
f_is_home_competition


Gradient Boosting Classifier - Time Series Data - Feature Importances
0.30
0.25
0.20
0.15
0.10
0.05
0.00
s_rank
ft_rank
t_rank
ft_age
f_age
s_age
t_age
gender_male
United States
Austria
France
Canada
Switzerland
f_is_home_competition

Regression

Similar to the classification approach, two machine learning models were developed for regression. The two models are a Random Forest Regressor and a Gradient Boosting Regressor. Both models were trained using the independent event data and the time series data in two separate phases and used the rank of the final event as the response variable. Of course, won_medal has been dropped from the dataset as it predicts the rank. The results from the two models are summarized below.

Independent Event Data

Random Forest Regressor		
	Mean Absolute Error	Mean Squared Error
Training Set Results	12.12	246.65
Testing Set Results	12.72	274.30

Gradient Boosting Regressor		
	Mean Absolute Error	Mean Squared Error
Training Set Results	11.81	237.68
Testing Set Results	12.74	278.71

Time Series Data

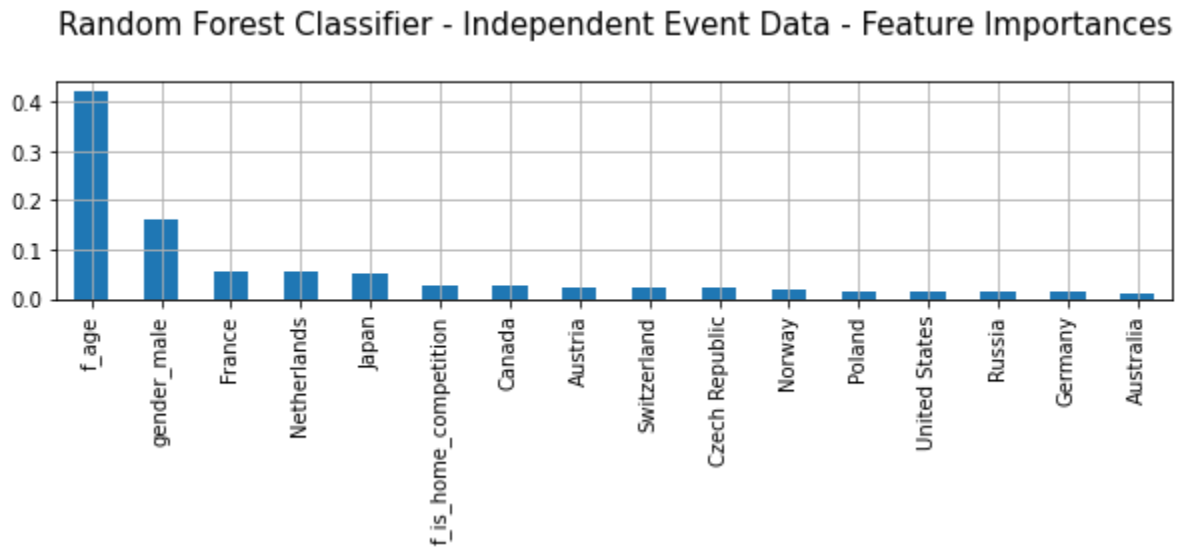
Random Forest Regressor		
	Mean Absolute Error	Mean Squared Error
Training Set Results	8.31	118.15
Testing Set Results	9.72	168.92

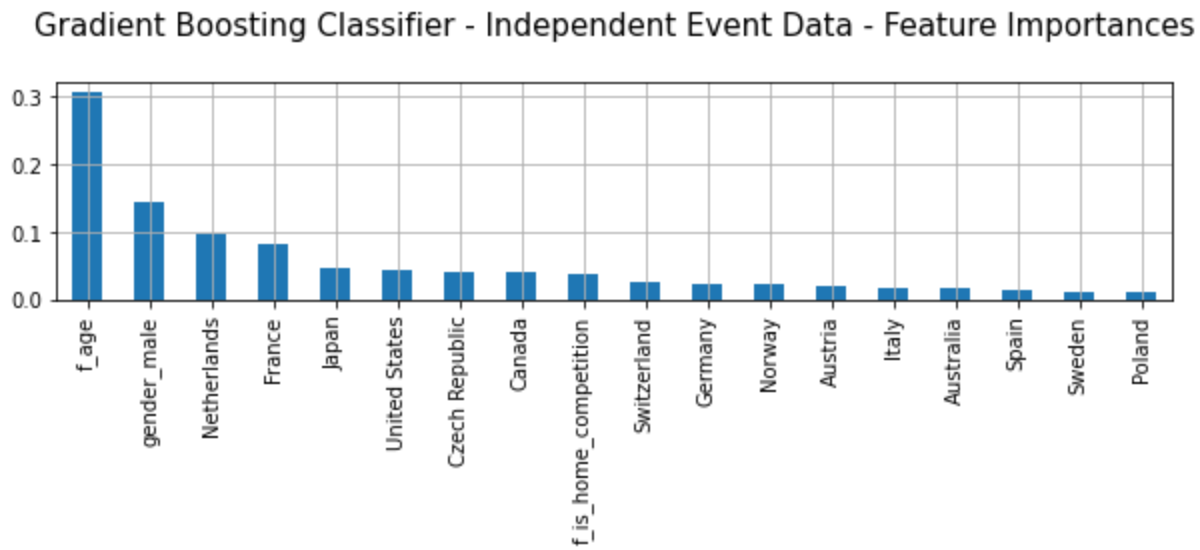
Gradient Boosting Regressor		
	Mean Absolute Error	Mean Squared Error
Training Set Results	9.53	141.57
Testing Set Results	10.78	188.68

Observations and Conclusions

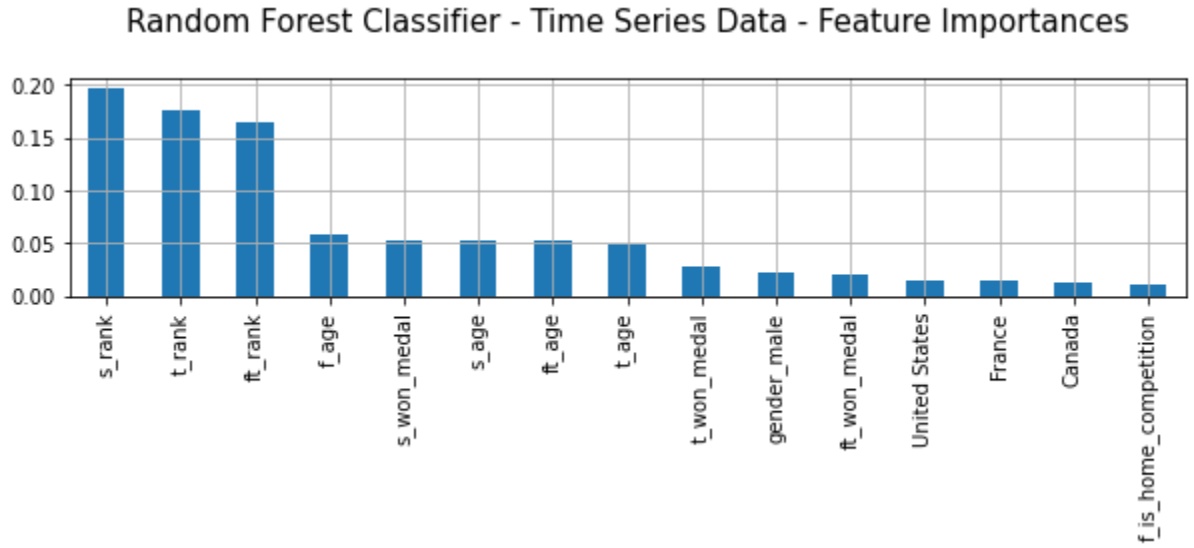
The analysis and model results showed that the Random Forest algorithm has outperformed the Gradient Boosting in both the regression and classification approaches. However the performance was significantly better when using the time series dataset as compared to the independent event dataset.

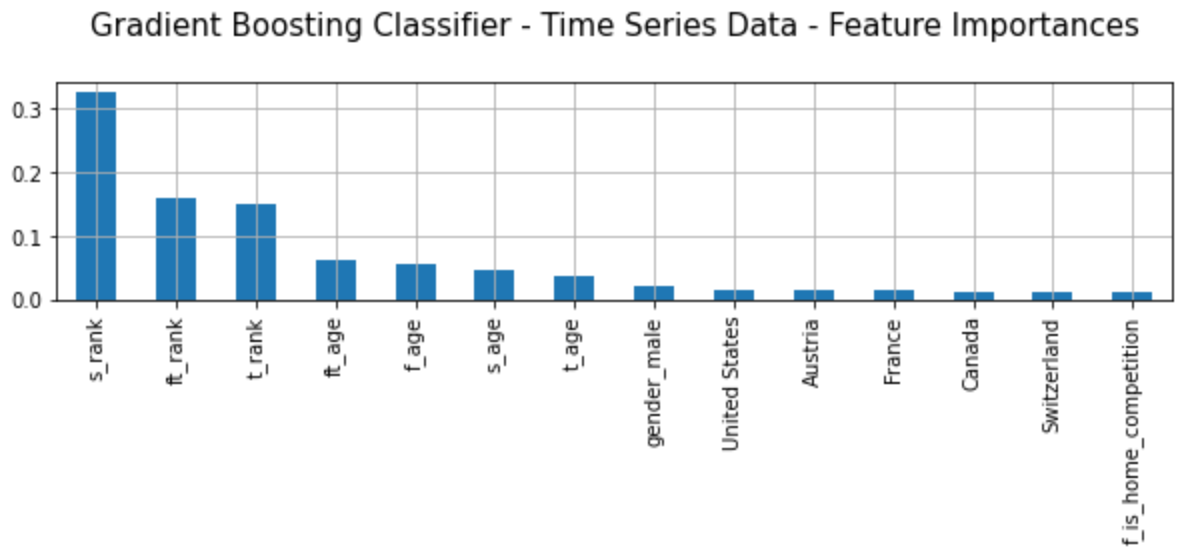
After training the model with the independent event data, both the regression and classification models showed relatively poor results. The best performing classifier was the Random Forest. The best precision score on the testing set after hyperparameter tuning was only 0.16 and the Mean Absolute Error from the regressor on the testing set was 12.72. The low performance in both the training and the testing set shows that the model is suffering from high bias. The feature importance graph illustrates that age is an important factor, followed by countries of origin that athletes are overall more successful. The high bias problem indicates that this dataset is overly simplistic and does not successfully capture the complexity of the dataset.





After training the model with the time series data, both the regression and classification models performed significantly better. The best performing classifier showed precision of 0.42 and recall of 0.42 in the testing set after hyperparameter tuning, which effectively translates to roughly a 42% chance of accurately predicting that an athlete will win a medal, whereas for non-medal winning athletes, both precision and recall is 0.95. The regressor results showed a Mean Absolute Error of 9.72 in accurately predicting the rank of the athlete. The feature importance analysis shows that performance and age in the preceding events have a very powerful effect in predicting the outcome, with country of origin for the most successful countries and the home competition variable coming after these.





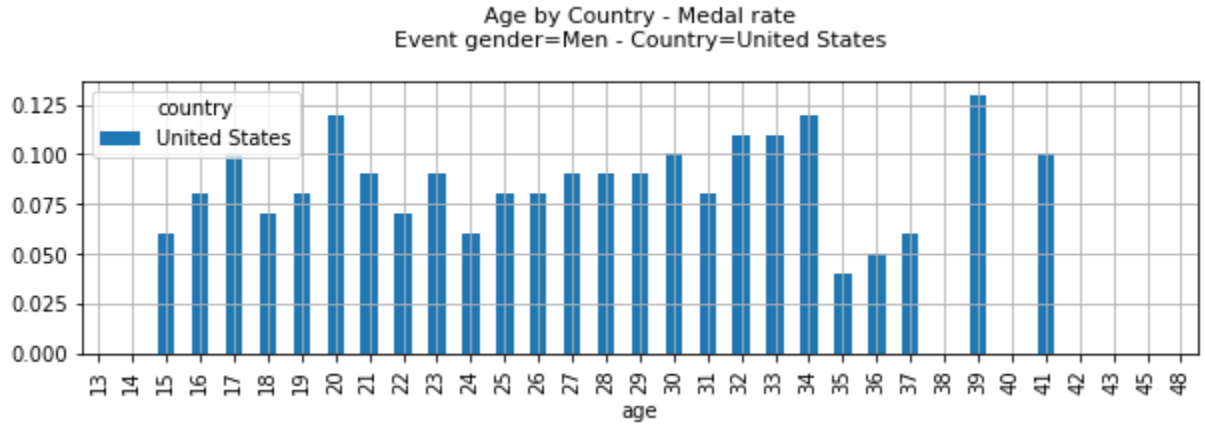
Overall, the Random Forest Classifier with the time series data, is the model of choice that provides the best predictive accuracy for the dataset of Snowboarding. The approach can be expanded in other sports as well. The predictions rely primarily on the performance and age of the athlete in the three events preceding the event we are trying to predict and the outcome is the probability of the athlete winning a medal. Alternatively, the Random Forest Regressor can be used in a similar way to predict the rank of the athlete for the event in question.

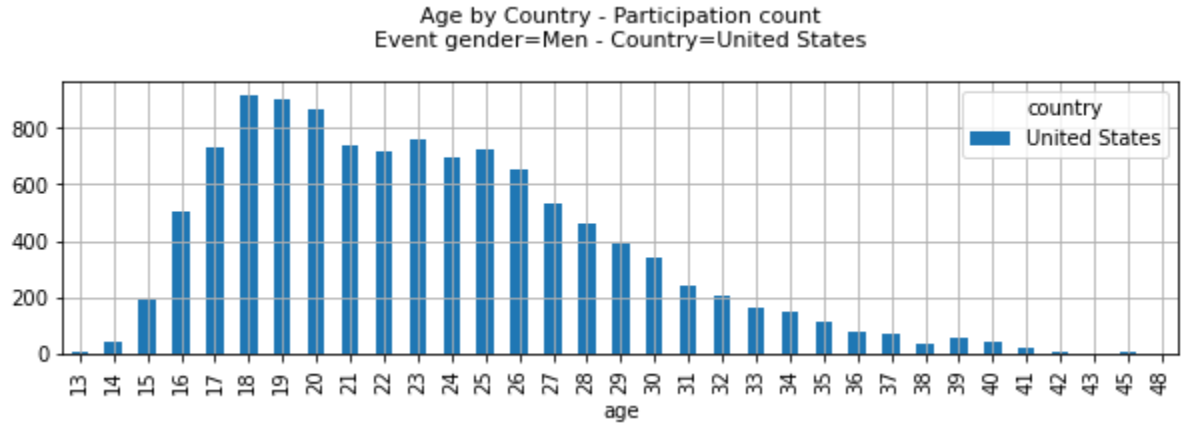
Key Questions and Answers

- Can a country's percentage of athletes competing at an Olympic Games while in their peak age ranges in their respective sports predict medal success?

Although medal success shows some higher rates in certain age groups by country, the difference is marginal and not significant enough to predict success. Thus, age cannot single-handedly be used to predict success and there is no clear cut peak age for medal success.

- Which athletes who competed in Tokyo 2021 (Beijing 2022) will be hitting their peak age in their respective sports in Paris 2024 (Milan 2026)?





Taking the figure above as a reference point for the United States with Snowboarding as the sport of reference, there is no specific peak age for athletes that are more successful. This is in line with previous conclusions of the analysis that even though there is a slightly higher success rate in some age ranges, there is no distinct "peak age" that athletes are more successful. Thus, with age alone as a factor it's nearly impossible to accurately predict success in the upcoming olympics based on age and success in previous ones.

- Can the number of young, “pre-peak” athletes that competed in Tokyo predict medal success for their countries in Paris 2024 and Milan 2026?

Using the classifiers or regressors built during this project it is very feasible to predict success in the upcoming Olympics. However, prediction is not limited to Tokyo only (though this is an option), but other preceding events and athlete performance during those can be used as well.

- How many promising athletes who did not compete in Tokyo (and Beijing) but will reach peak age in Paris (and Milan) do Team USA and other countries have in their pipelines?

Similarly to the previous question, any events that the athlete previously participated in along with their age and performance during those events, can be used with the models developed to predict their probability of success in the upcoming Olympics.

Final Outcome and Recommendations for future work

As a final outcome of this study, the time series approach is a robust and sustainable methodology for predicting the success of an athlete not only in Olympic Game events but also in World Championship events. The only requirement of the present state of the model is the age and rank of an athlete in three preceding events, as well as the country of origin of the athlete, the gender and the location of the competition in question to determine whether it takes place in the athlete's home country. This information can successfully predict medal winners with 42% and non-winners with 95% precision and recall.

Further enhancements can be achieved by accounting for biometric data of the athlete, as intuitively this information can provide extra predictive power.