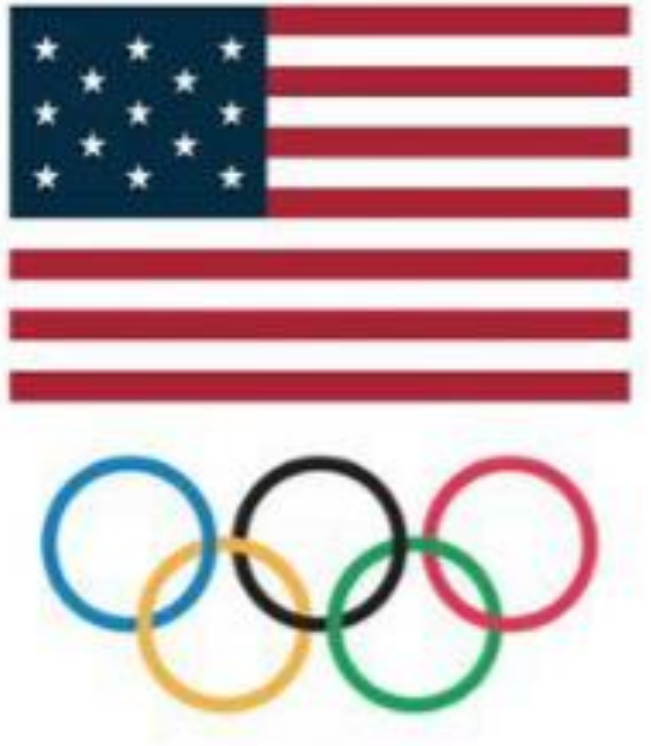


Stergios Koutrouvelis
USOPC
Age-Focused Olympic Competitive Analysis

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Project Summary:

- The USOPC would like to gain a better understanding of the role that age plays in medal success in various Summer and Winter Olympic sports
- We believe that peak age for winning medals varies by sport, event, discipline and gender.
- We would like to have 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 through the primary lens of age.

Key Questions:

- 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?
- 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?

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Data Summary:

- The data contains multiple sports but follows the same structure.
- For this initial analysis, the sport of Snowboarding has been selected. The analysis, framework and code can generalize and scale to any sport.
- Each row in the data contains information about a specific athlete and their result in a specific competition/event.
- The fields include information about:
  - Athlete and Personal ID
  - Athlete's Nation
  - Age of athlete at the time of the event
  - Competition type (e.g., Olympics, World Championships)
  - Sport/Event or discipline
  - Placement/rank
  - Result (e.g., distance, points, time)

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Problem Approach and Phases:

- Development of data processing framework for feature selection and schema standardization. Development of API that extracts new information, reads SQL queries to produce new data artifacts with specific statistics.
- Exploratory Data Analysis along the following axes:
  - Country
  - Sport
  - Event
  - Age
  - Competition Location
  - Medal Winning
- Feature Engineering for producing variations of datasets that will be used for ML
- Development of Machine Learning model that can predict the probability of an athlete winning a medal or the rank of the athlete based on preprocessed features.

Data Processing Framework

The framework is designed to extract/produce the following features from the raw data:

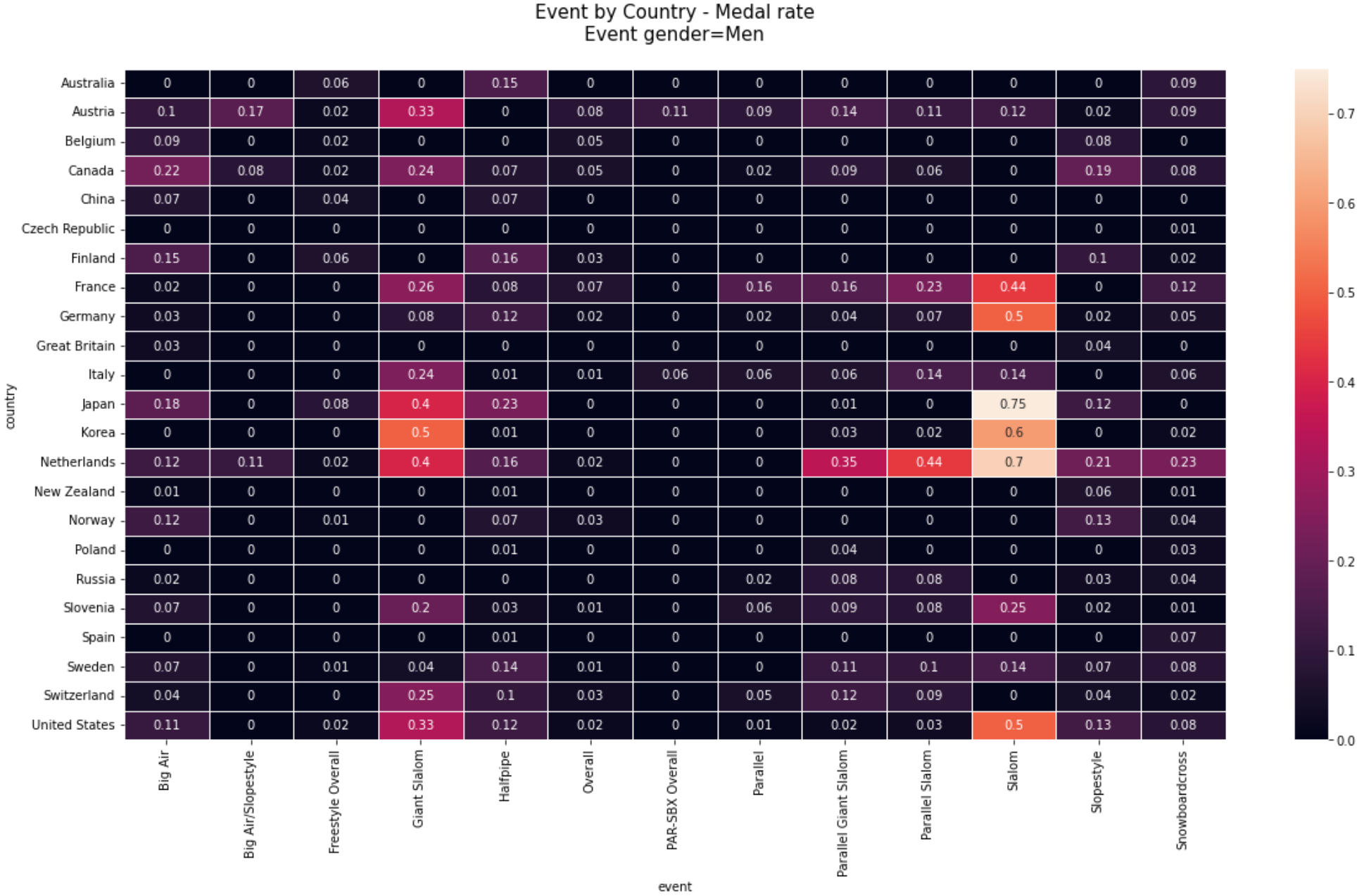
- **Class**: The competition class. Can be "Elite", "Juniors", "Youth"
- **Competition Date:** The date the competition was held
- **Competition City:** The city the competition was held at
- **Competition Country:** This column is being produced through the API that is using the geocoders library. It saves the matches to a JSON file to reduce computation.
- **Event Gender:** The gender of the event. Can be "Men" or "Women". This is further being processed for one-hot-encoding.
- **Event:** The event name (ex. "SnowboardCross").
- **Sport Name:** For this dataset it is just "Snowboard"
- **Medal:** This indicates whether the athlete has won a medal. Can be "G", "S", "B" or None. This is futher being processed to produce a binary won_medal column.
- **Country:** The country of origin for the athlete.
- **Is Home Competition:** Binary column indicating whether the competition was held at the athletes' home country
- **Age:** The age of the athlete in years.
- **Rank:** The rank of the athlete in the competition

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

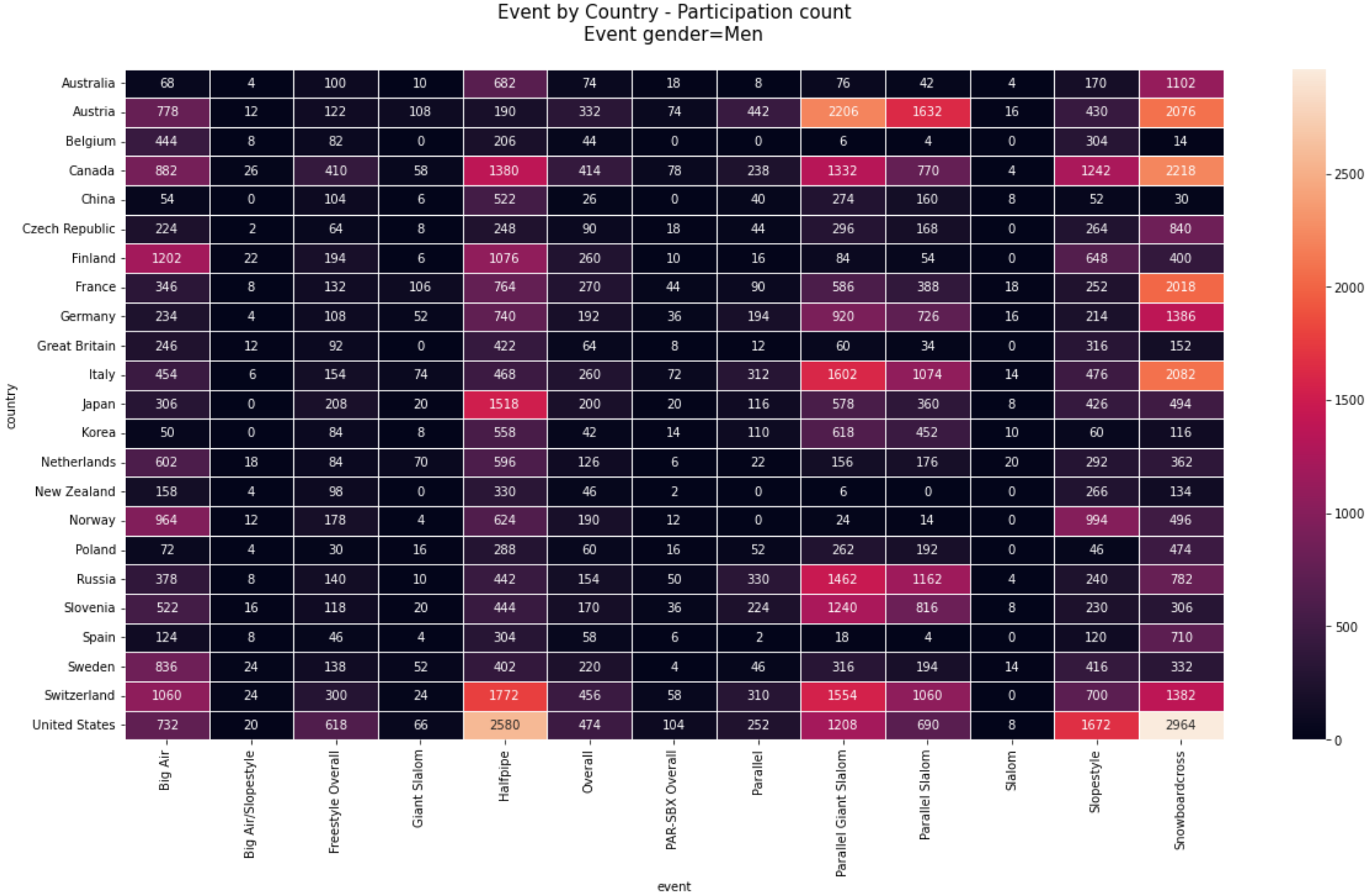
Exploratory Data Analysis

- **EDA by Athlete Country of Origin and Event:**
  - **Medal Rate:** Ratio of medals won over total participation by event and country.
  - **Participation Count:** Total Number of Participants by event and country.
  - **Participation Rate:** Proportion of country's total participation by event.
  - **Explicit Graphs for the United States**
- **EDA by Athlete Country of Origin and Age:**
  - **Medal Rate:** Ratio of medals won over total participation by country and age group.
  - **Participation Count:** Total Number of Participants by country and age group
  - **Participation Rate:** Proportion of countries total participation in each age group
  - **Explicit Graphs for the United States**
- **Home vs Away Competition Success Rate by Country:** Ratio of medals won to total participation for home vs away competitions

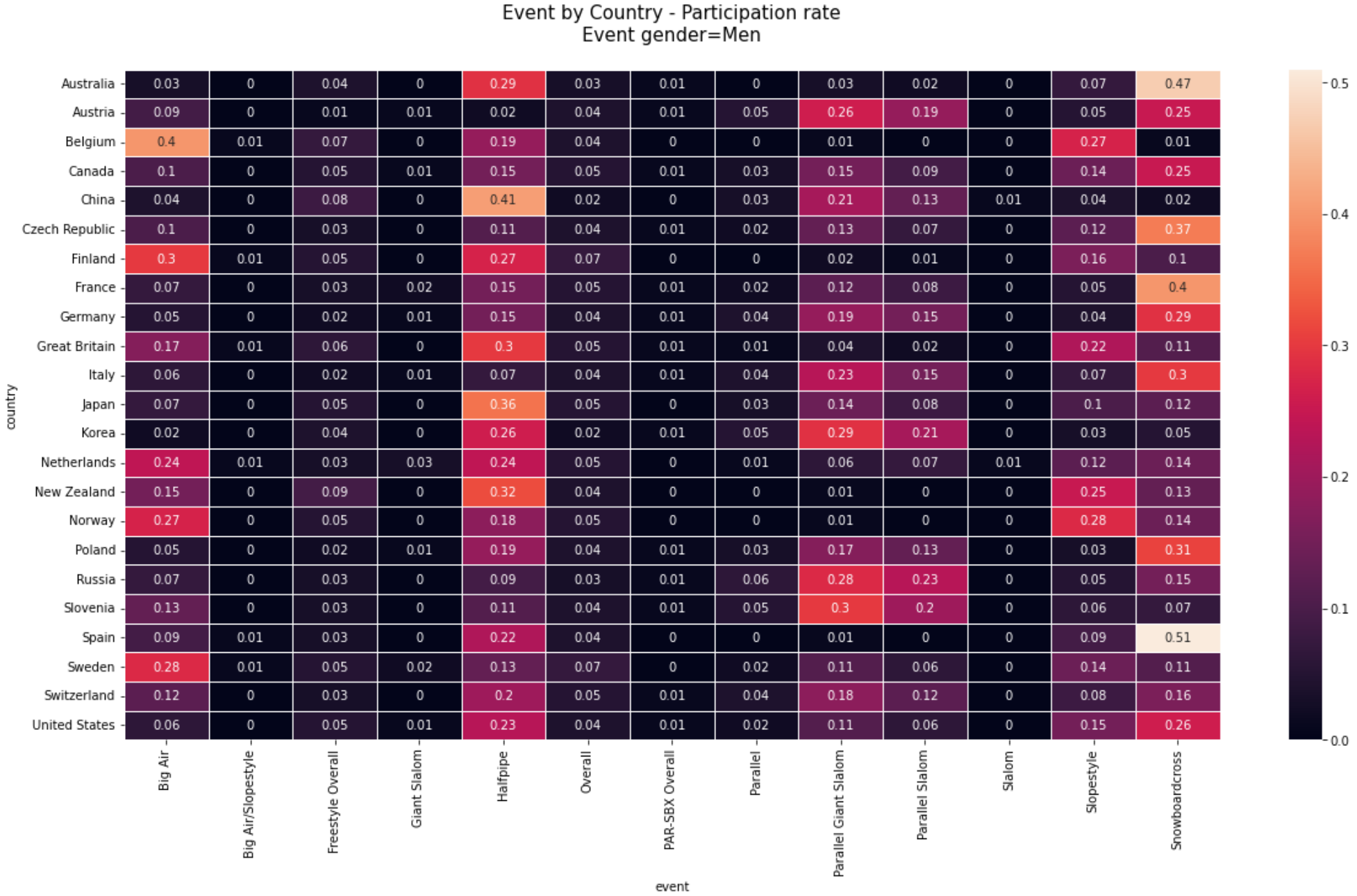
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



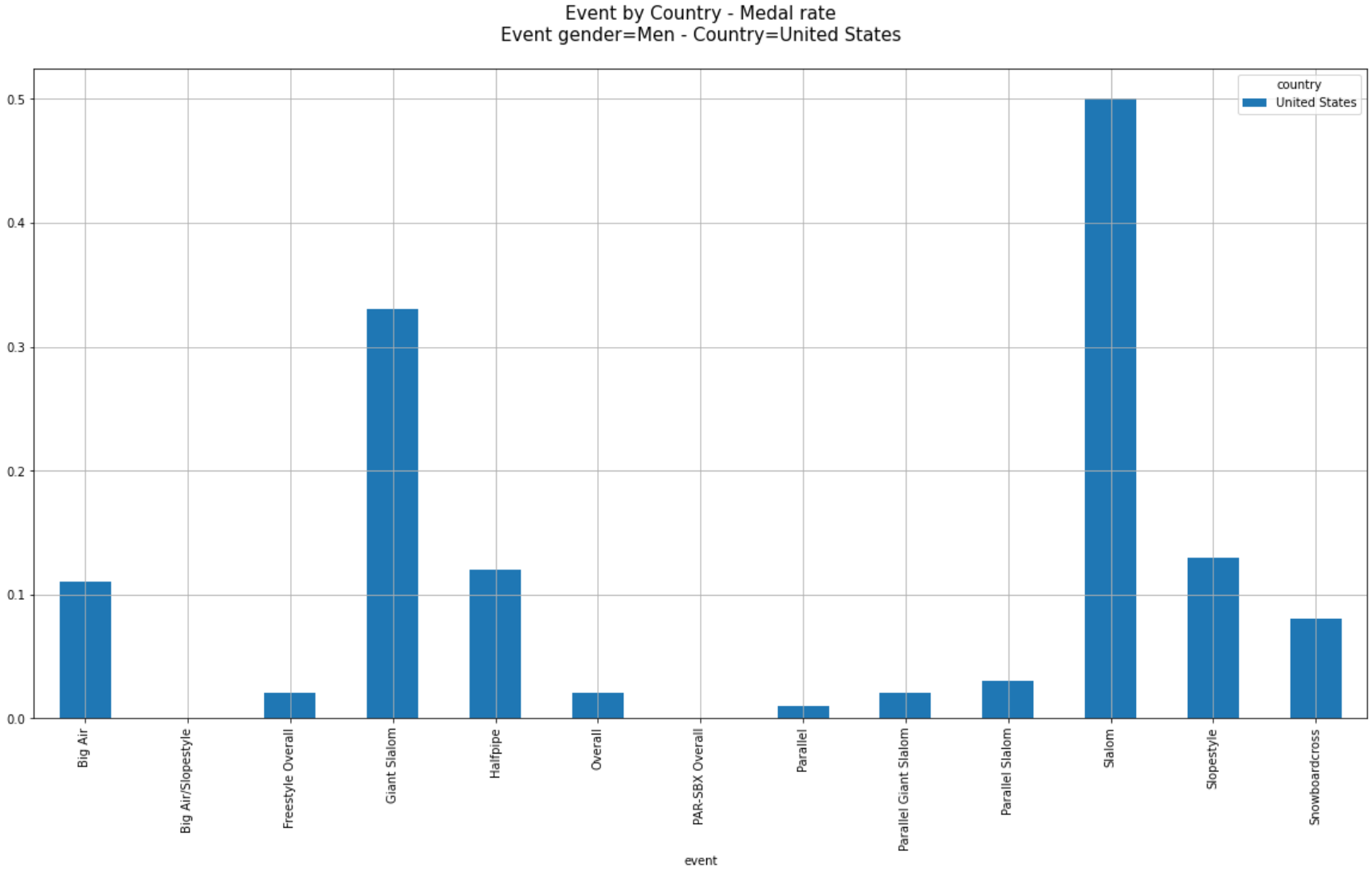
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



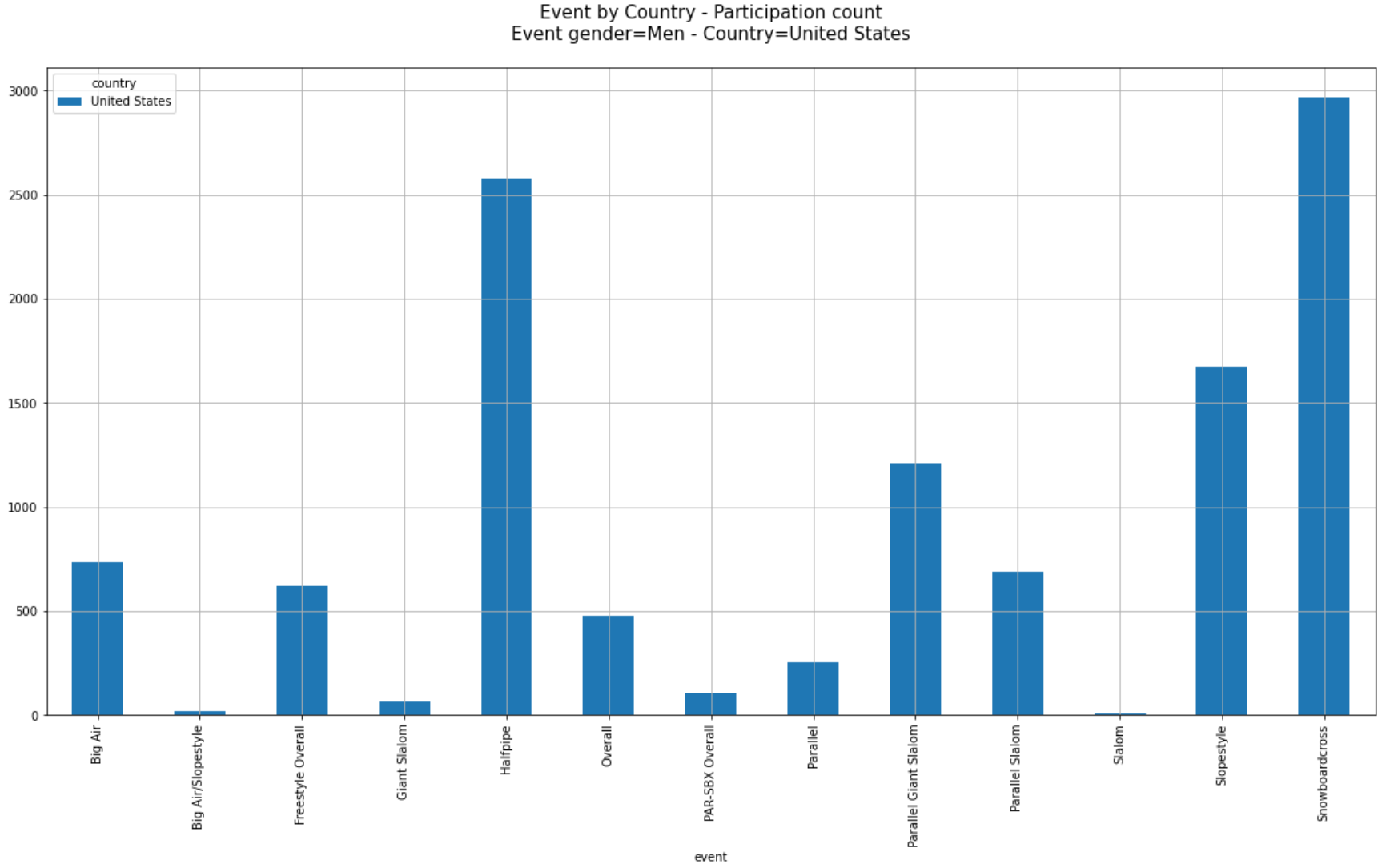
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



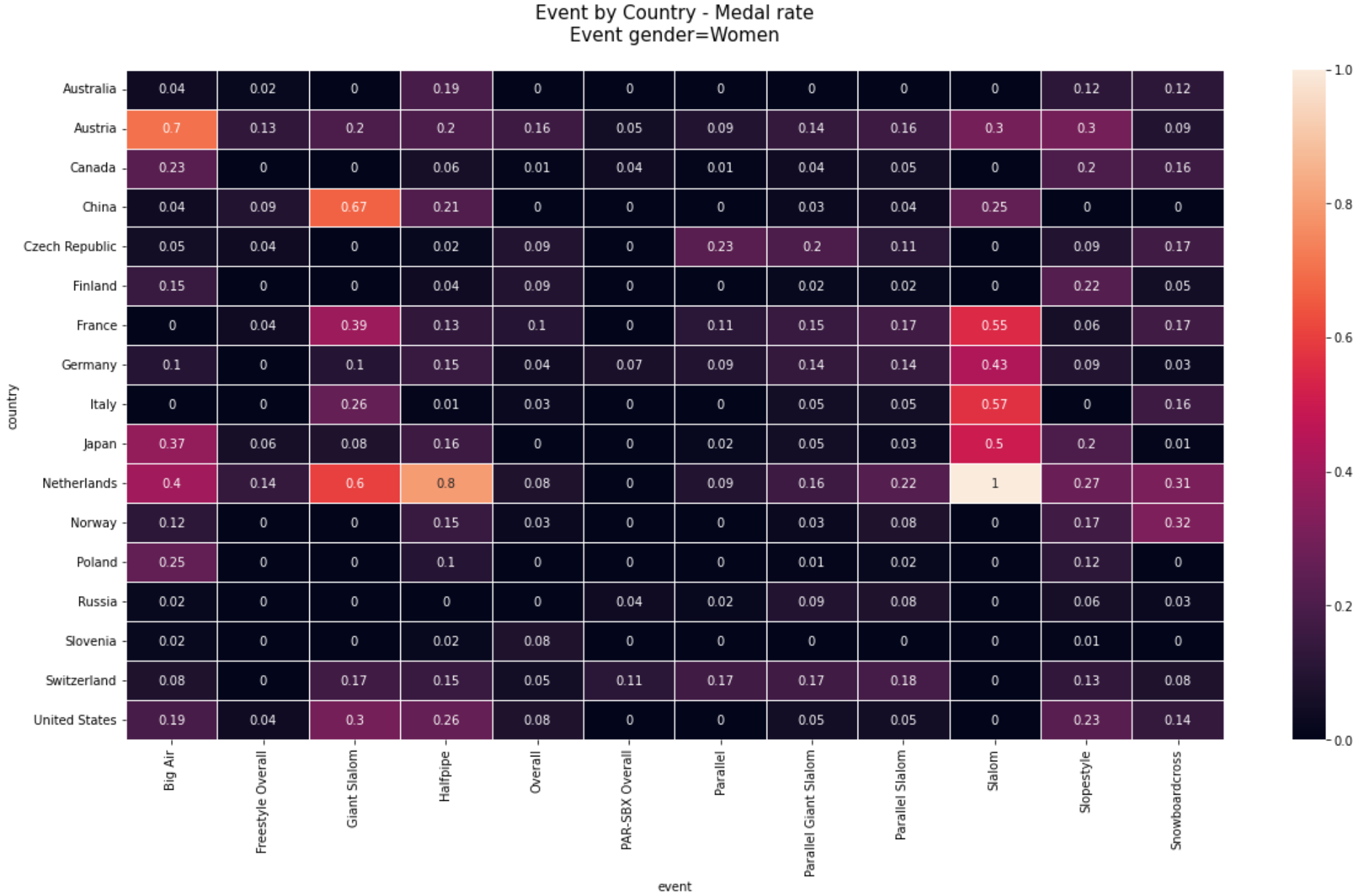
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



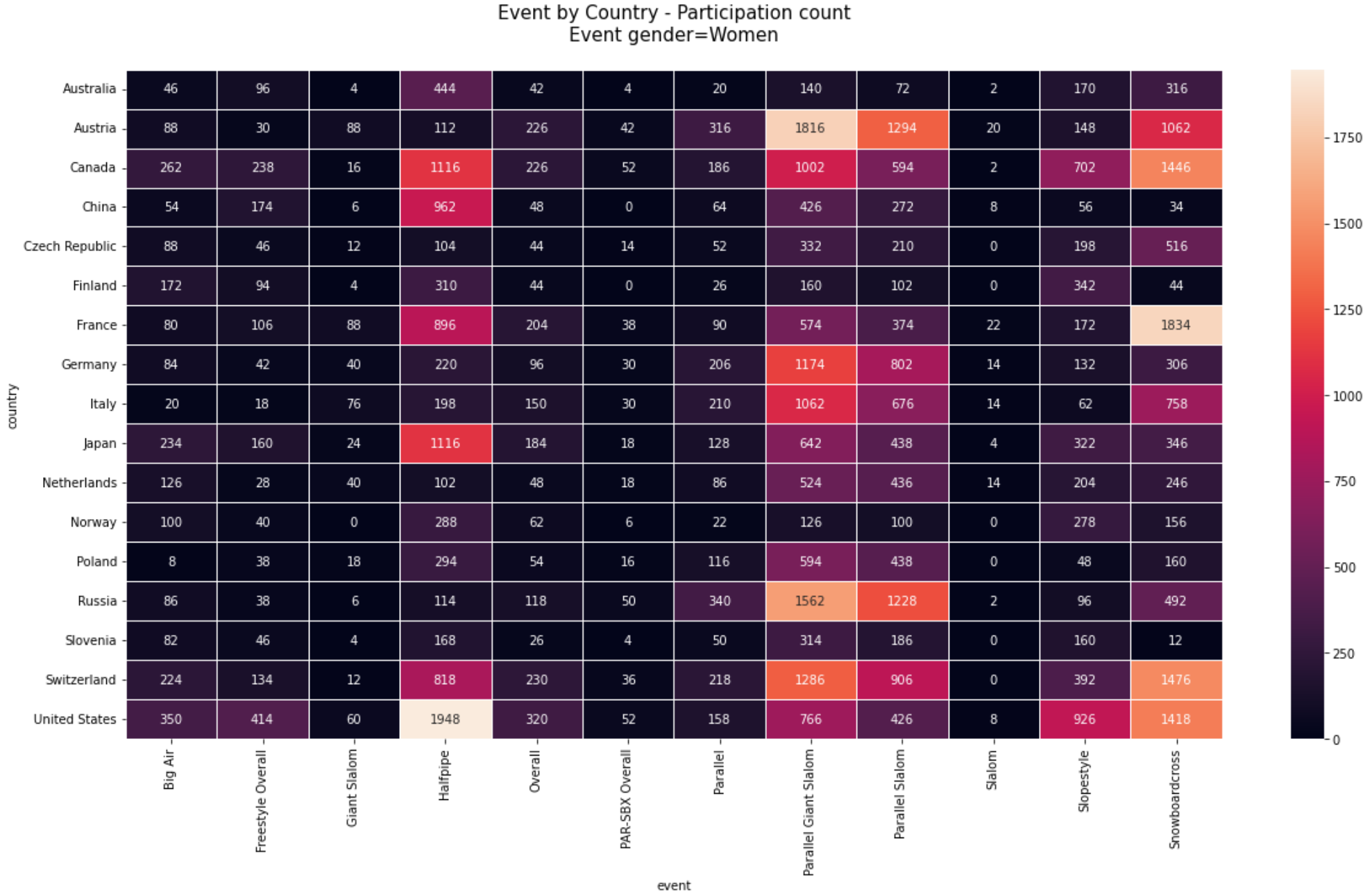
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



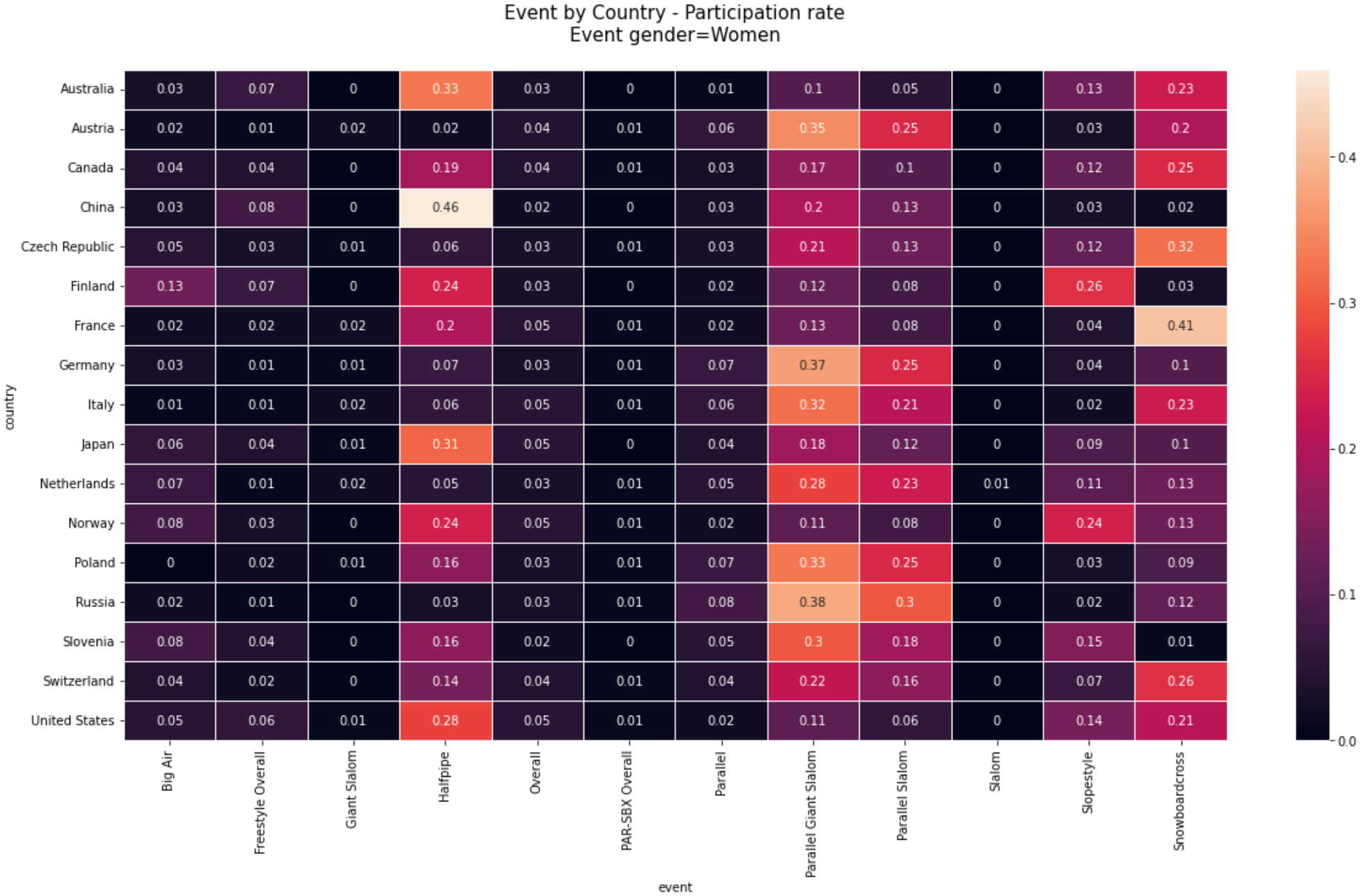
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



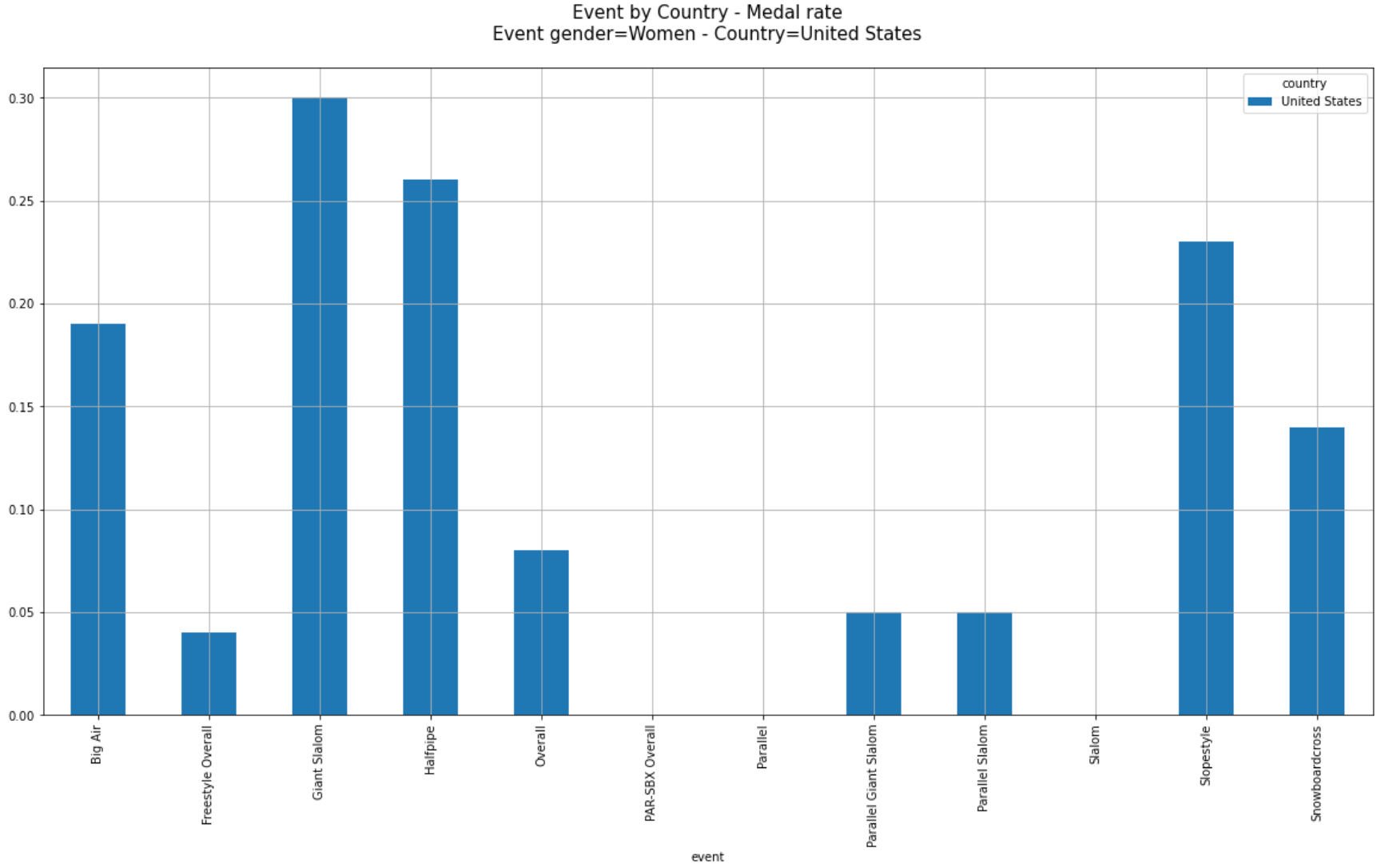
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



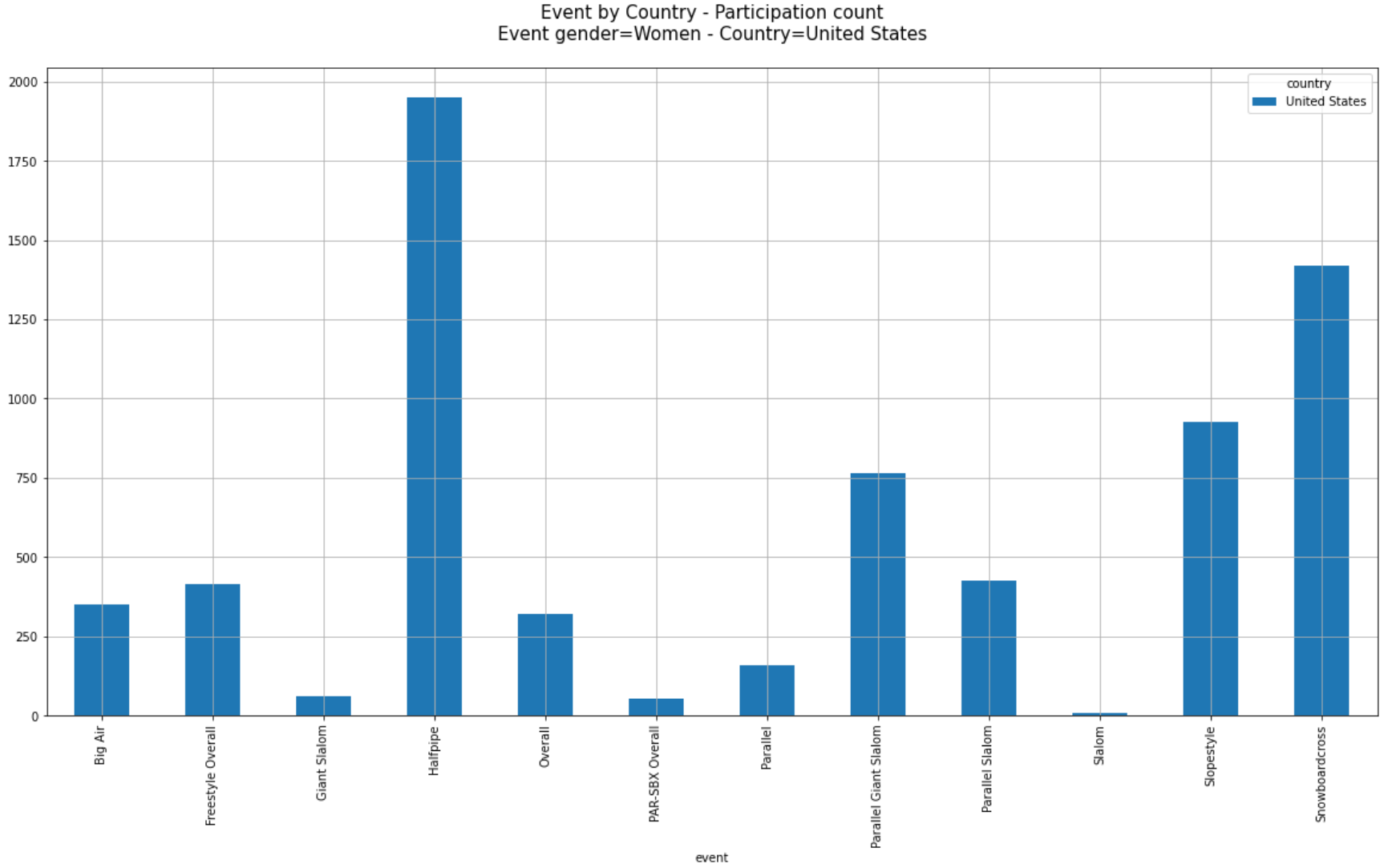
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



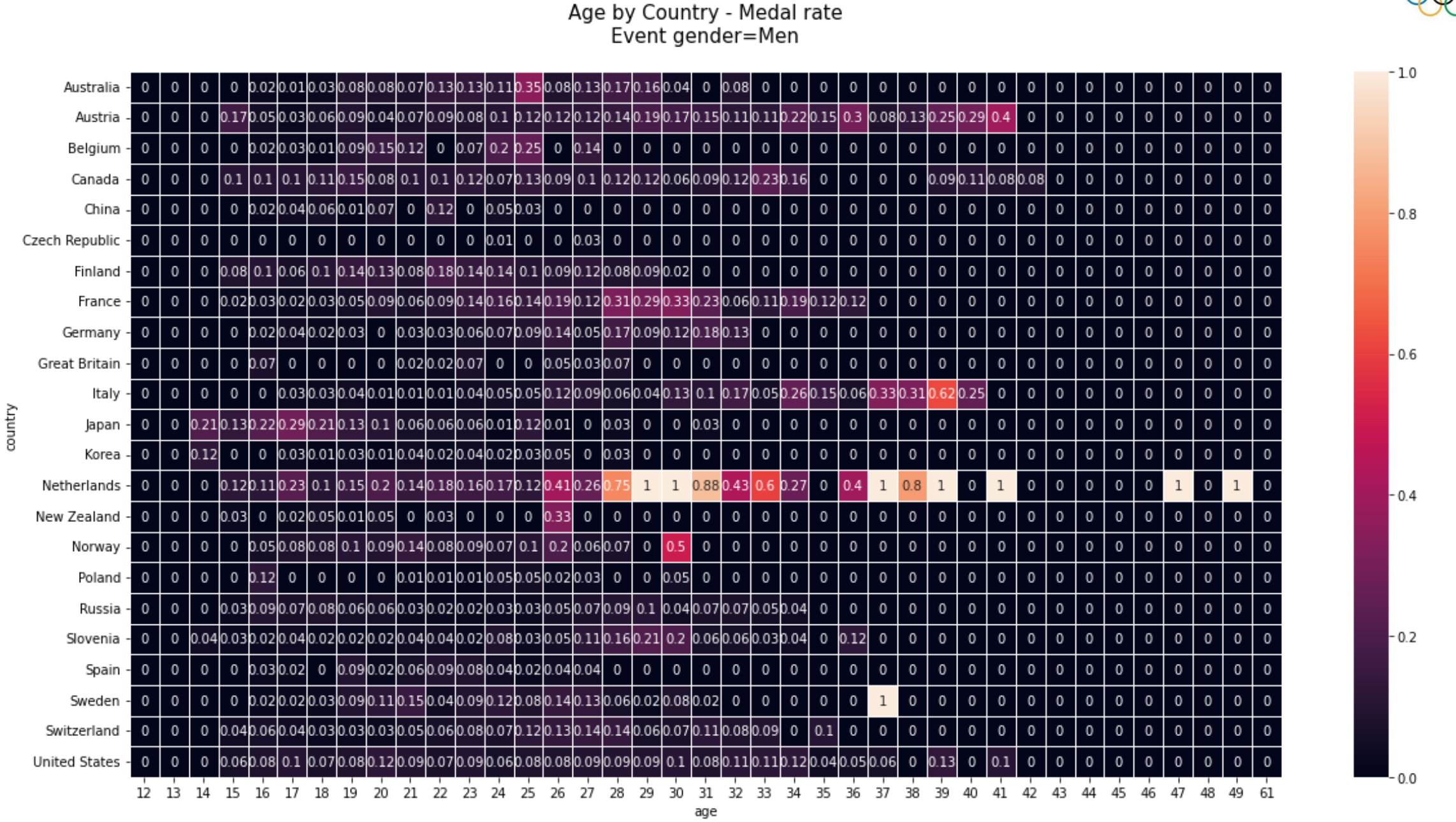
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



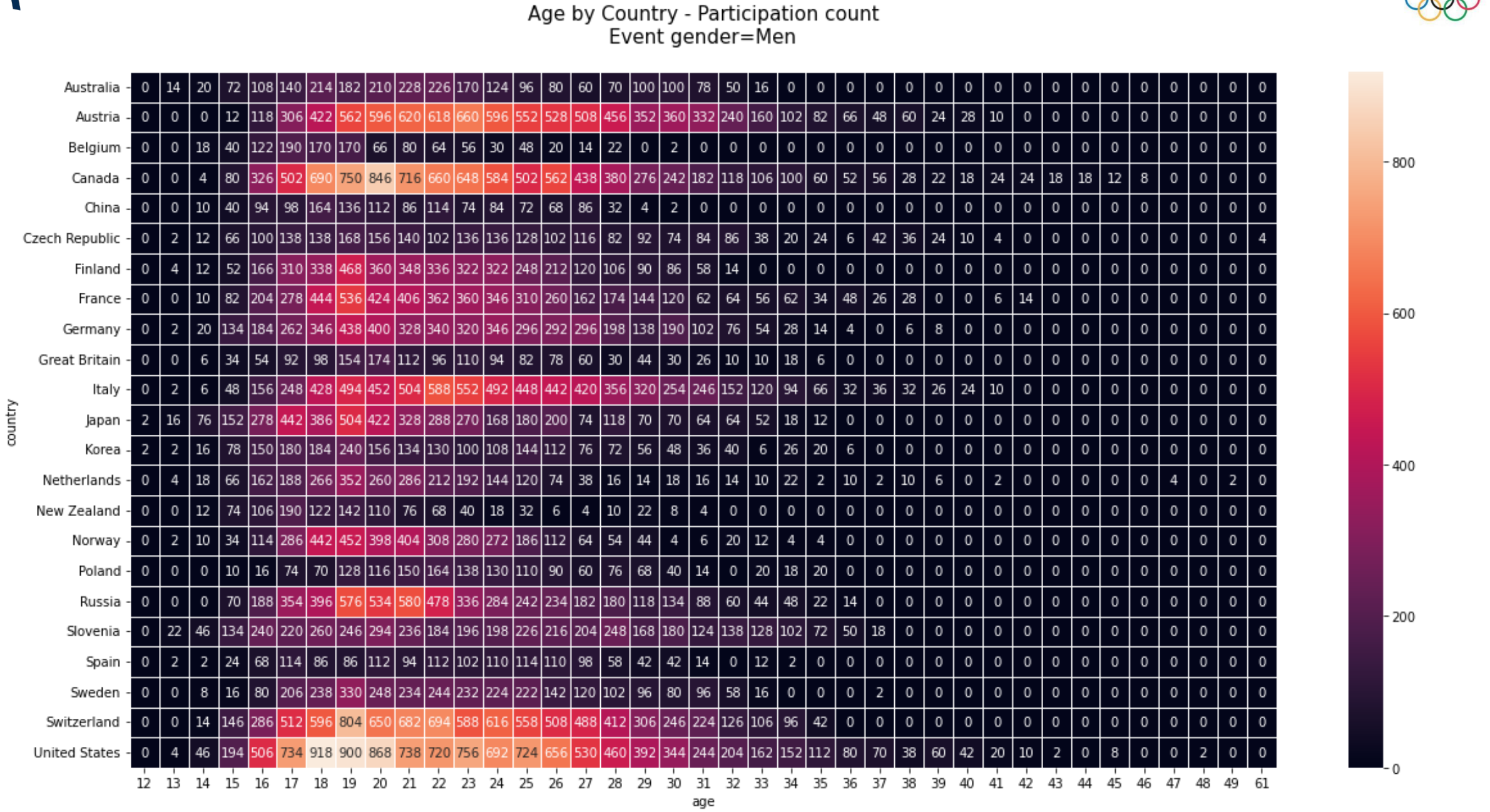
Country Statistics by Sport and Gender Conclusion:

- 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.

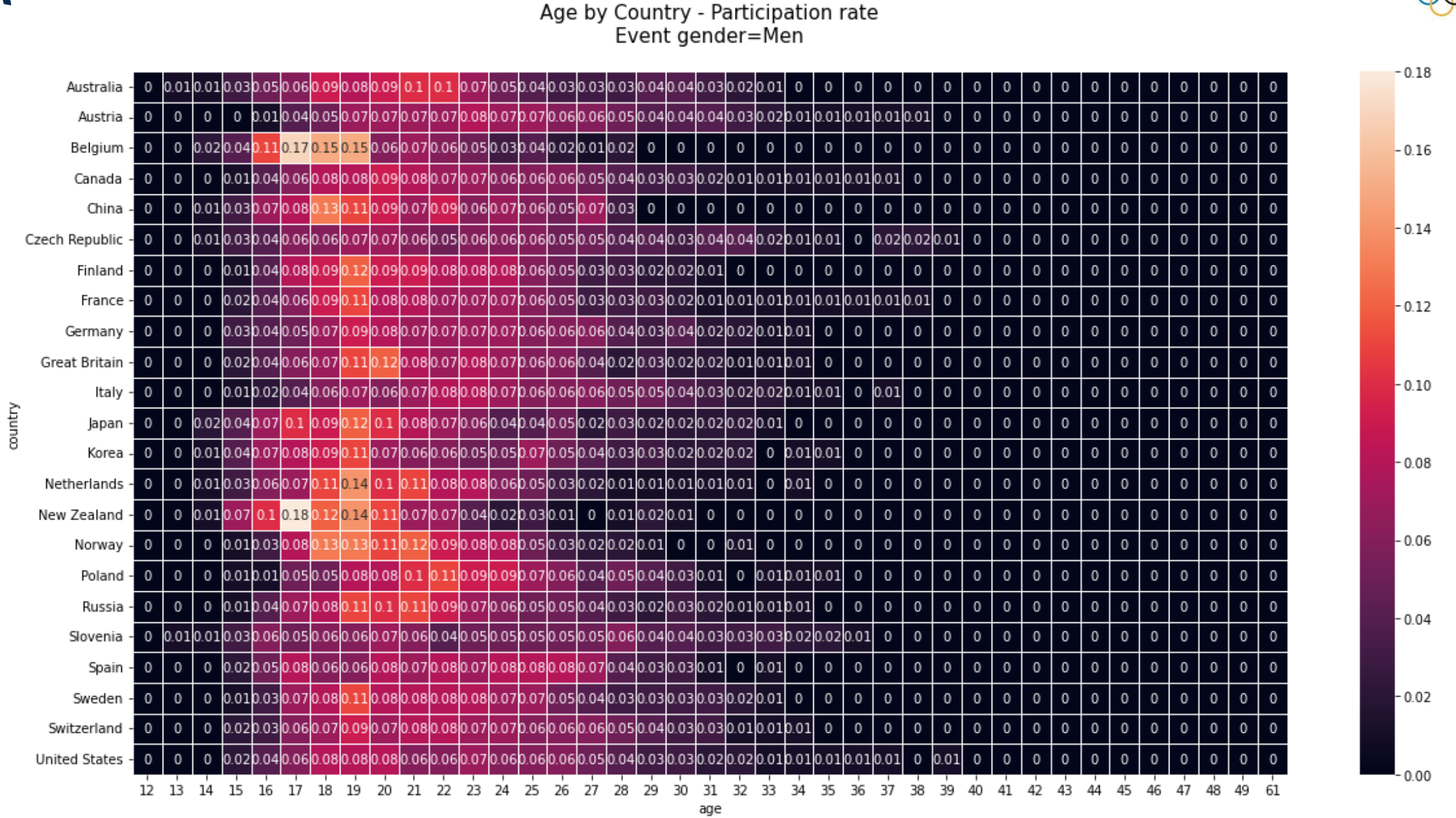
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



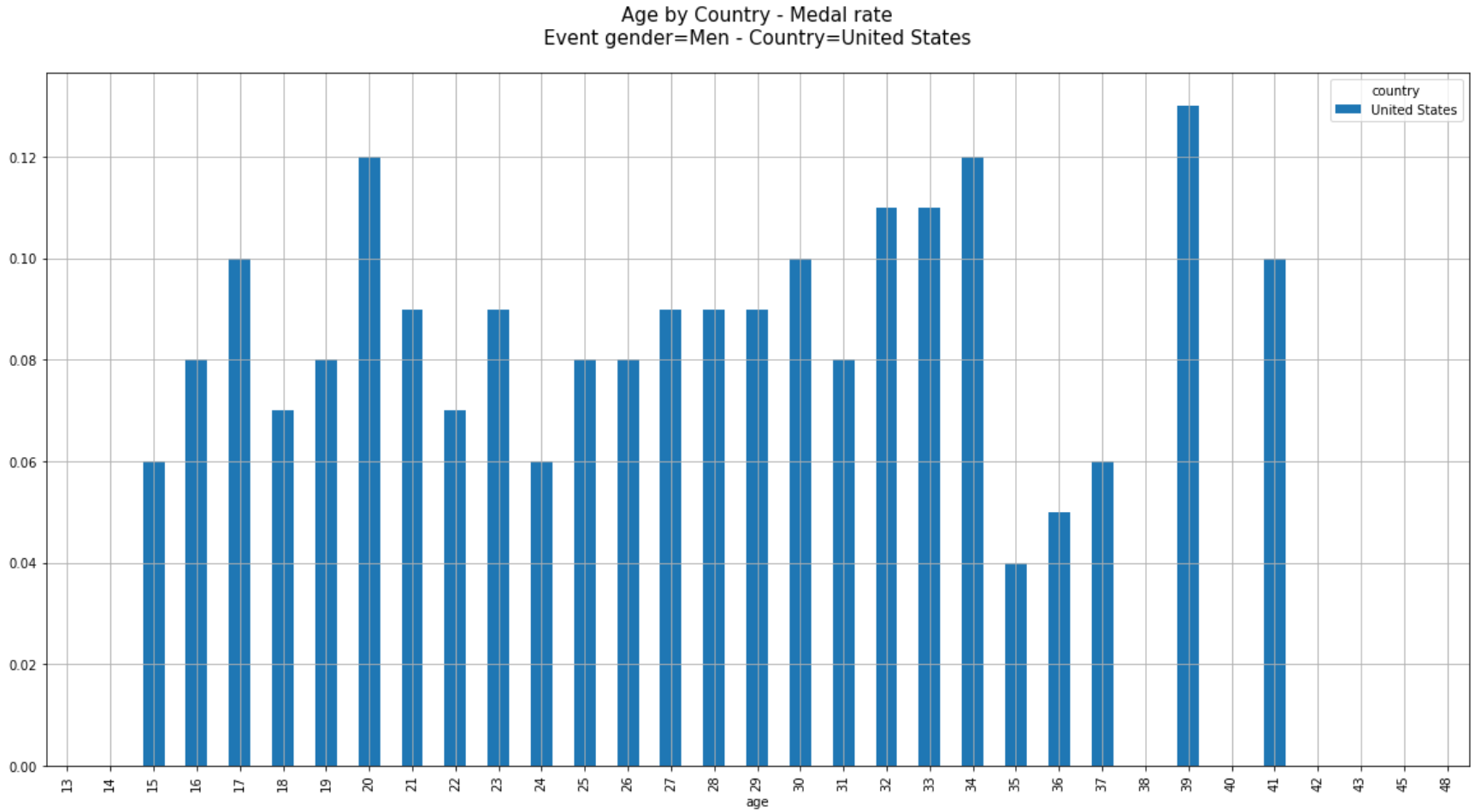
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



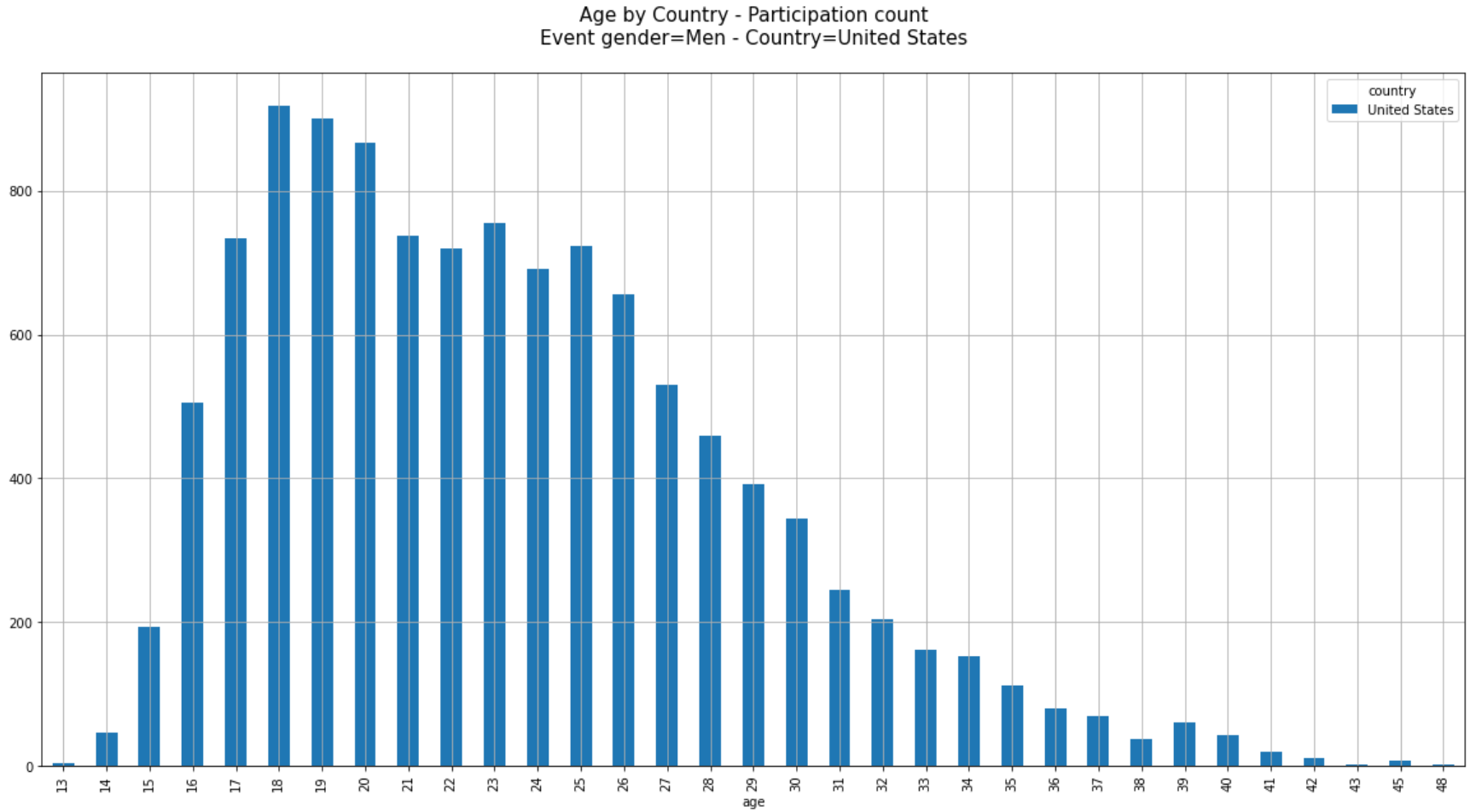
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



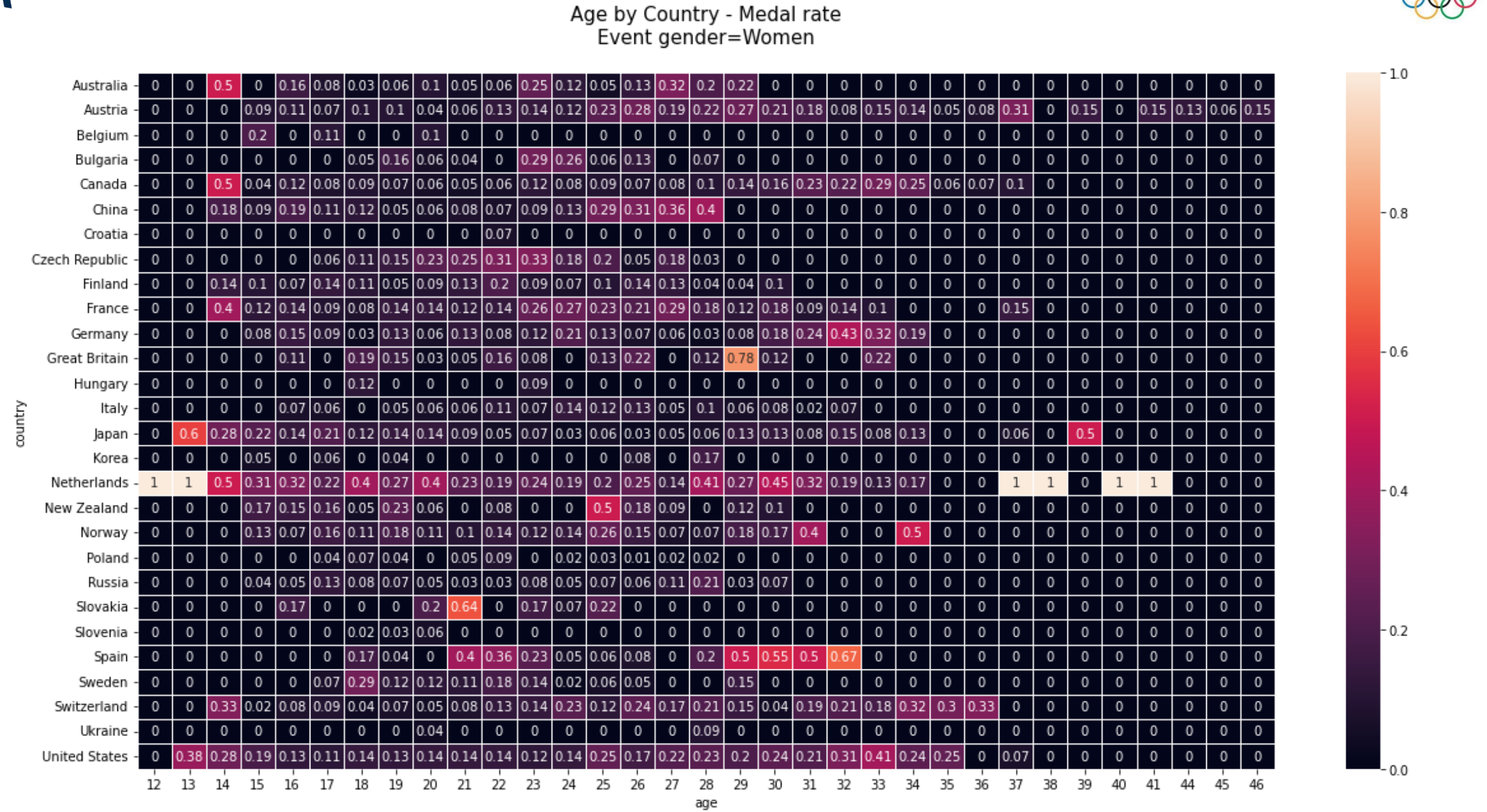
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



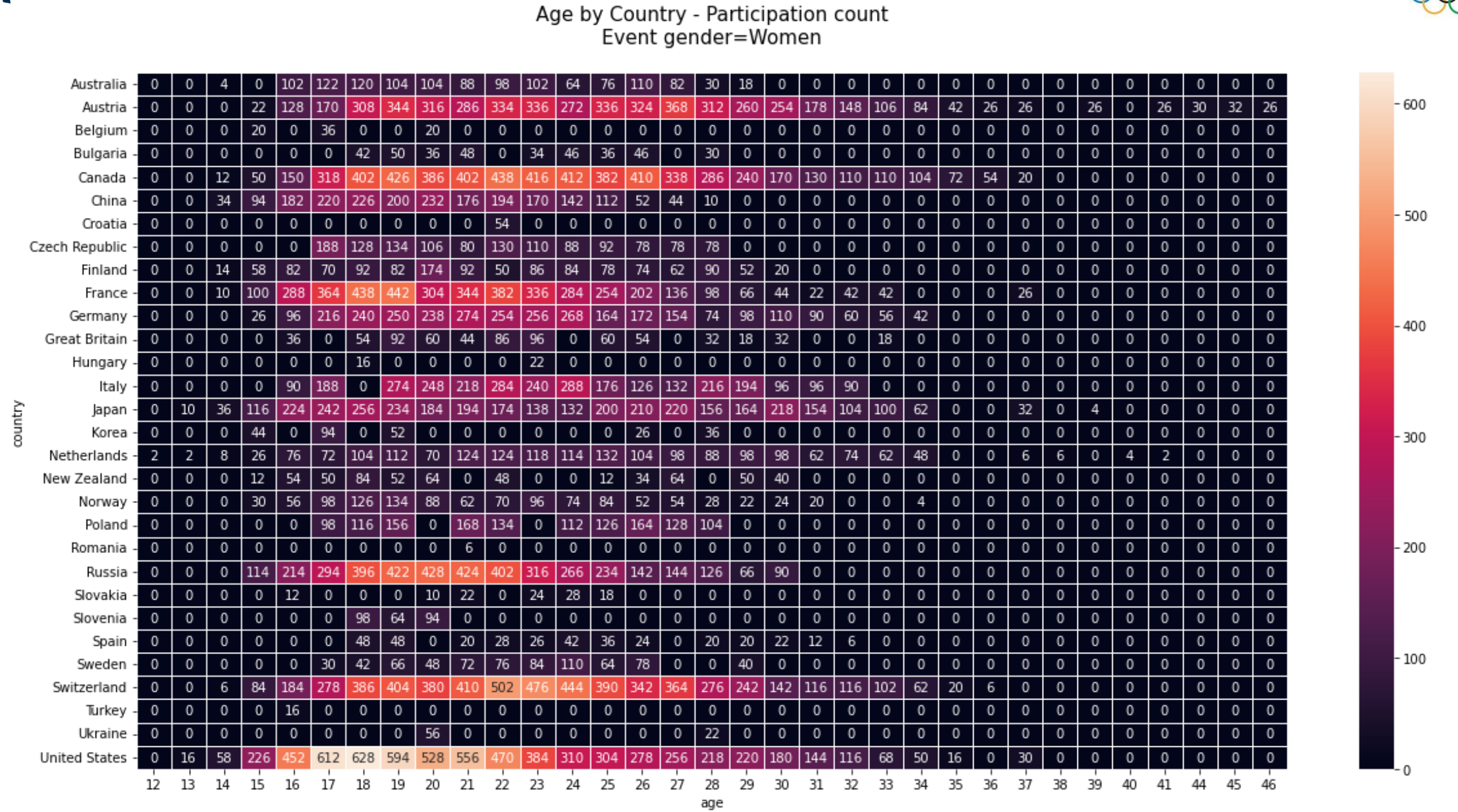
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



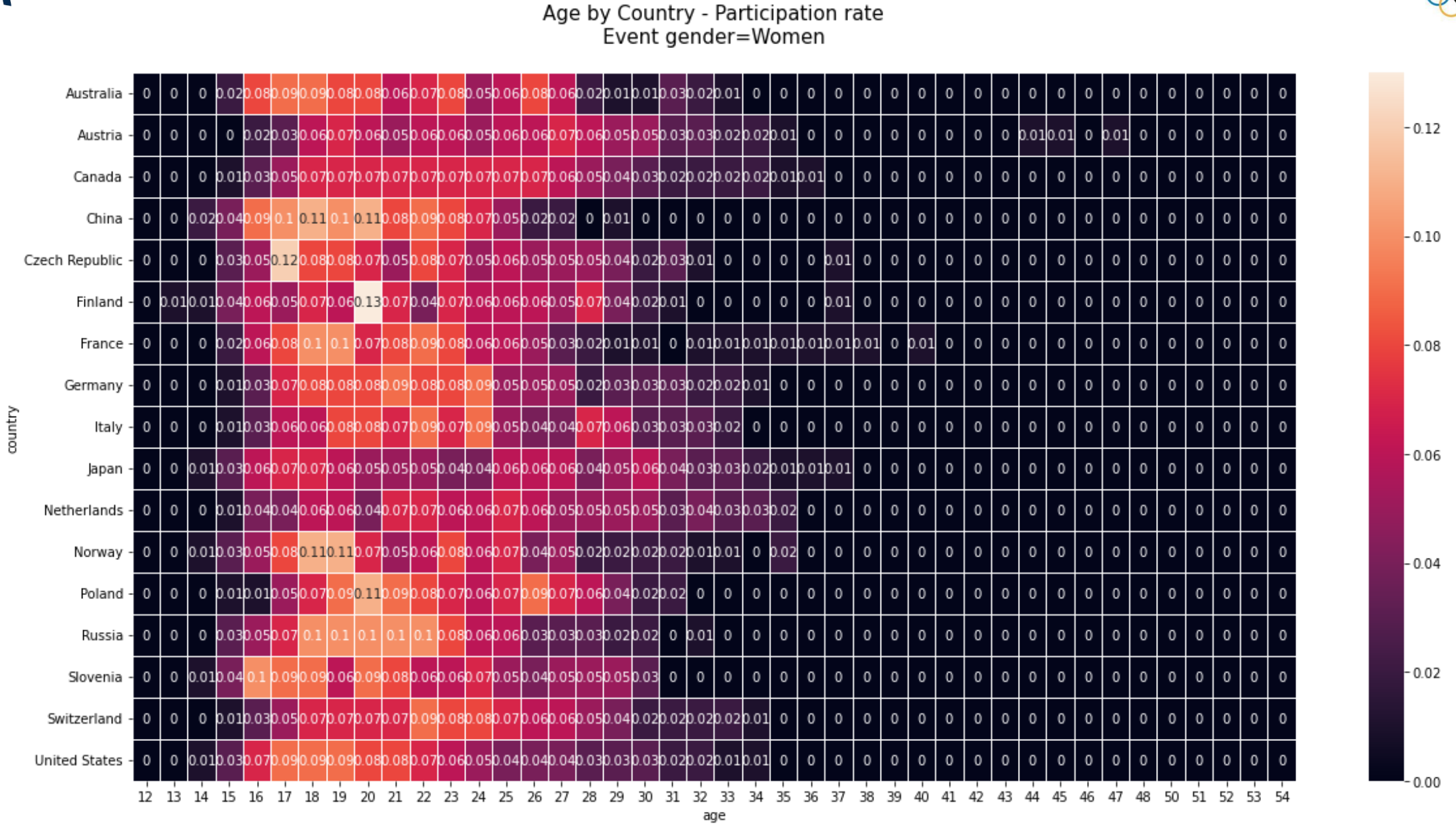
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



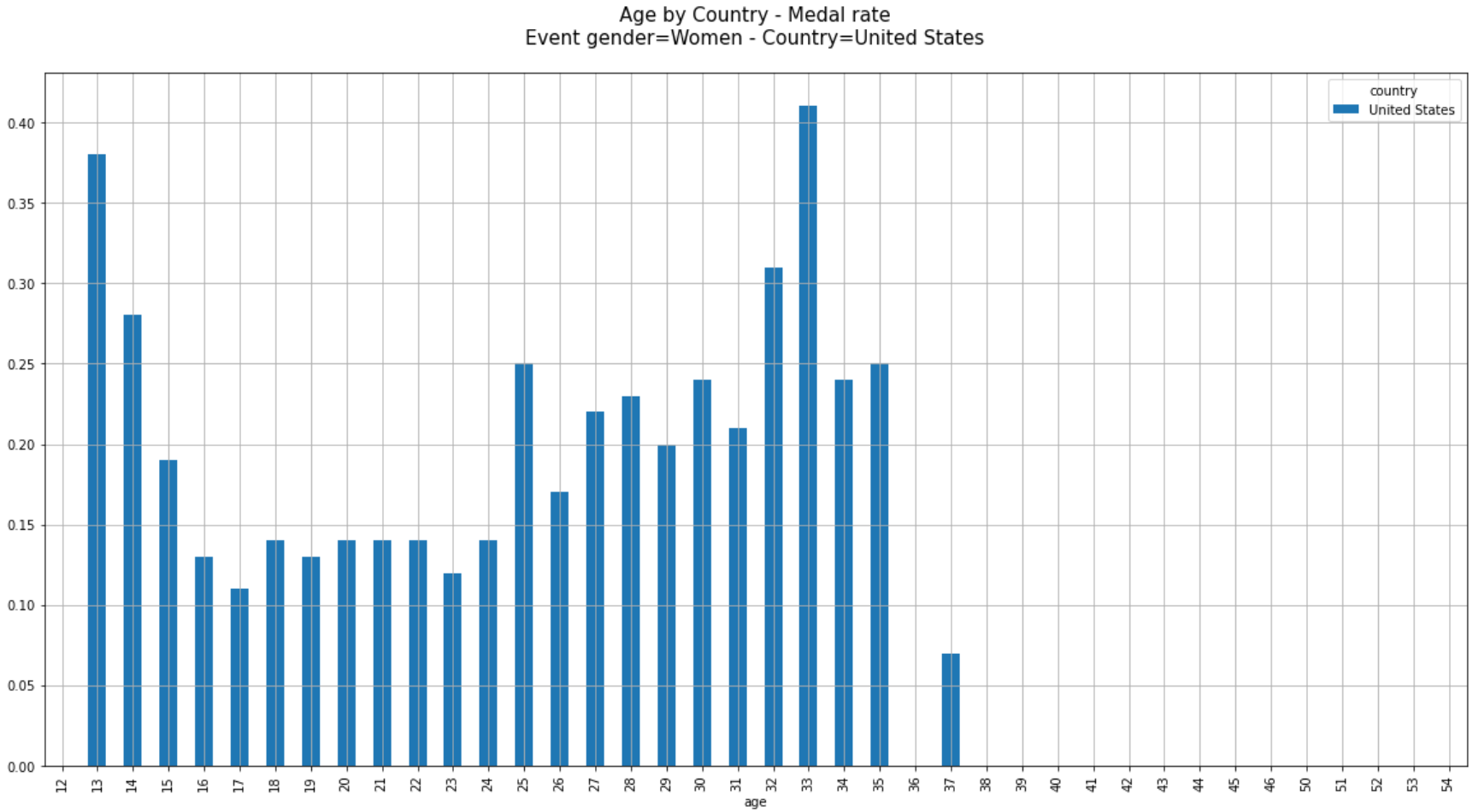
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



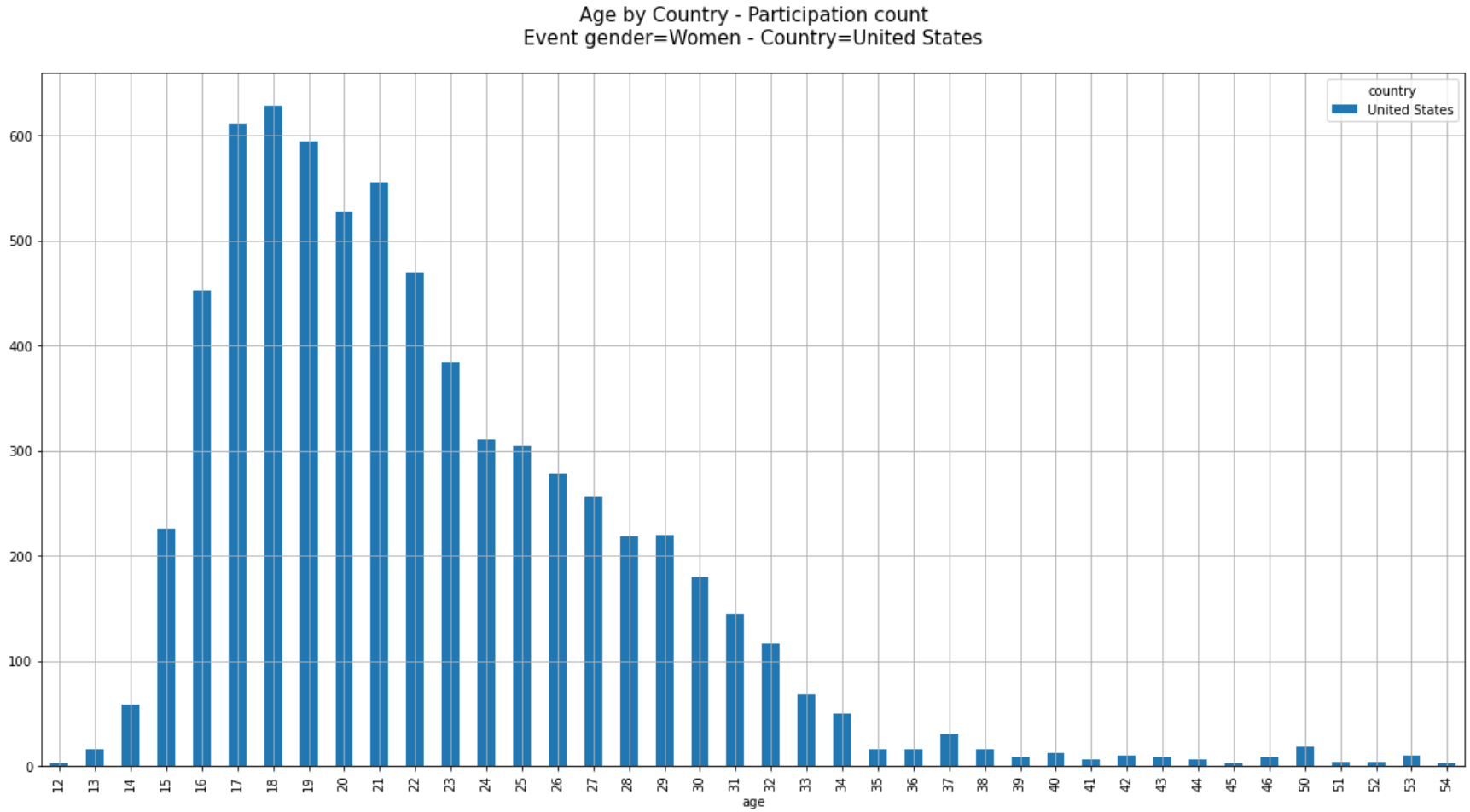
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



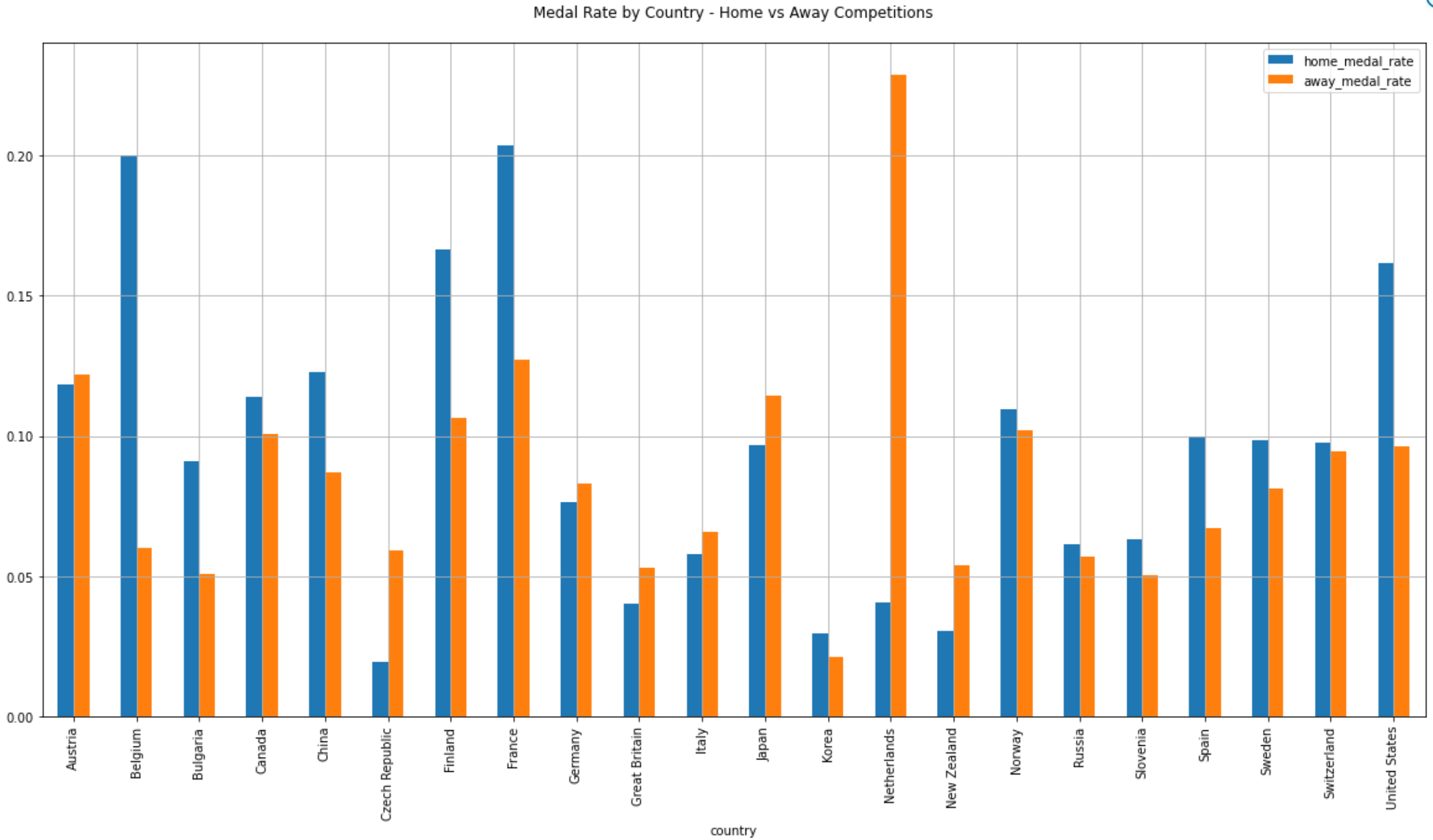
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Country Statistics by Age and Gender Conclusion:

- 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.

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Home vs Away Competitions medal rate conclusion:

- 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.

Feature Engineering and Machine Learning:

- Two different Approaches in feature engineering:
  - Independent Event Dataset
  - Time Series Dataset

- Two different predictions:
  - Won medal (Classification)
  - Rank (Regression)

- Two Machine Learning Models:
  - Random Forest
  - Gradient Boosting Decision Tree

Independent Event Dataset:

- Treating each athlete event independently:
  - **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)

Time Series Dataset with Past Performance as feature:

- As an alternative, we will combine the power of past performance.
- The way of compiling the dataset involves the following steps:
  - For every event and athlete we compile the following features:
    - Athlete ID
    - Event Date
    - Event and Sport Name
    - Is home competition variable as defined previously
    - Rank in the event
    - Won Medal binary variable as defined previously
  - Now for the athlete-event datapoints that we defined previously, we are going to find the 3 events the athlete participated for the same sport that preceded the event in reference.
  - This results in the following compilation
    - 4th preceding athlete event for given sport
    - 3rd preceding athlete event for given sport
    - 2nd preceding athlete event for given sport
    - Event in reference (1st) athlete event for given sport. This is used as our prediction event.

Time Series with Past Performance Dataset:

- **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 competion_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
- **Latest competition_ahtlete_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
- **Latest competition_is_home_competition**: Binary variable indicating whether the competition was held in the home country

Machine Learning – Regression vs Classification Approach:

- The prediction problem can be approached both as a classification and regression problem:
  - **Classification:**
    - In this approach we will be using the "won_medal" binary variable from the current event as defined before to be our response variable.
    - Rank from the current event will be dropped from the dataset.
    - All other variables will be used as predictor variables
    - The model will be evaluated with the following classification metrics:
      - Precision
      - Recall
      - F1-Score
    - The inherent class imbalance will be addressed by adjusting the classifier to account for class weights.
  - **Regression:**
    - In this approach, we will be using "rank" from the current event as our response variable.
    - Won_medal for the current event will be dropped from the dataset.
    - All other variables will be used as predictor variables.
    - The model will be evaluated with the following regression metrics:
      - Mean Absolute Error
      - Mean Squared Error

Machine Learning Models:

- As illustrated in the exploratory analysis phase, age factor is highly nonlinear in predicting success
- Linear Regression and Logistic Regression models have difficulty capturing these types of relationships
- Thus, in this analysis we will opt for tree-based ensemble methods that exhibit high of-the-shelf predictive power and capture nonlinearities in the data.
- The models used are the following:
  - Classification:
    - Random Forest Classifier
    - Gradient Boosting Decision Tree Classifier
  - Regression:
    - Random Forest Regressor
    - Gradient Boosting Decision Tree Regressor
- Cross-Validation and parameter tuning
  - For both methods, parameter tuning will be performed using five-fold cross-validation.
  - The best classifier/regressor will be selected and results will be reported.

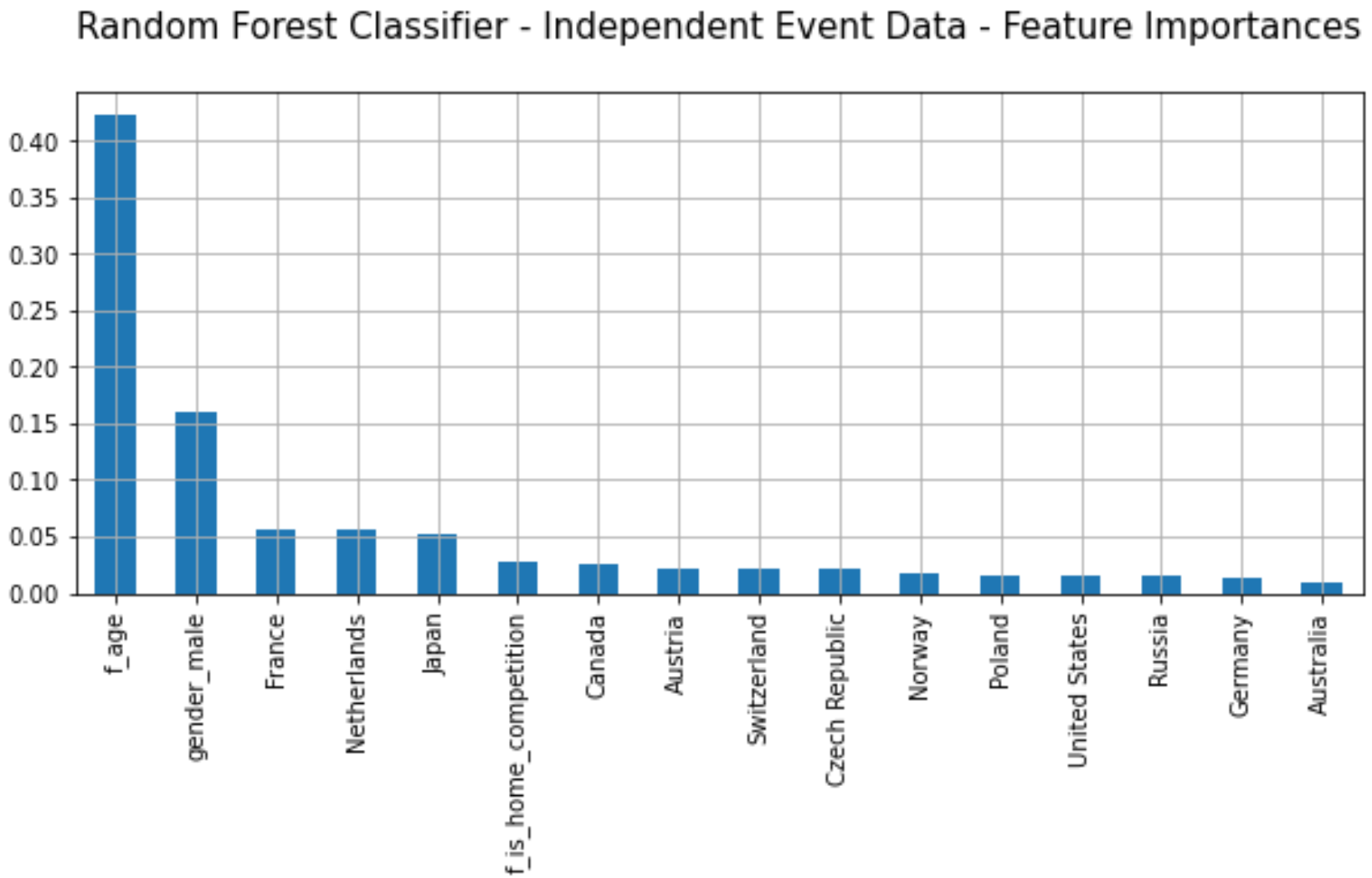
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Random Forest Classifier with Independent Event Data:

	Medal Won	Precision	Recall	F1-Score
Training Set Results	False – 0	0.96	0.74	0.83
	True - 1	0.18	0.66	0.29
	Accuracy	-	-	0.73
	Macro Average	0.57	0.70	0.56
	Weighted Average	0.90	0.73	0.79

	Medal Won	Precision	Recall	F1-Score
Test Set Results	False – 0	0.95	0.71	0.81
	True - 1	0.16	0.57	0.25
	Accuracy	-	-	0.70
	Macro Average	0.55	0.64	0.53
	Weighted Average	0.88	0.70	0.76

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



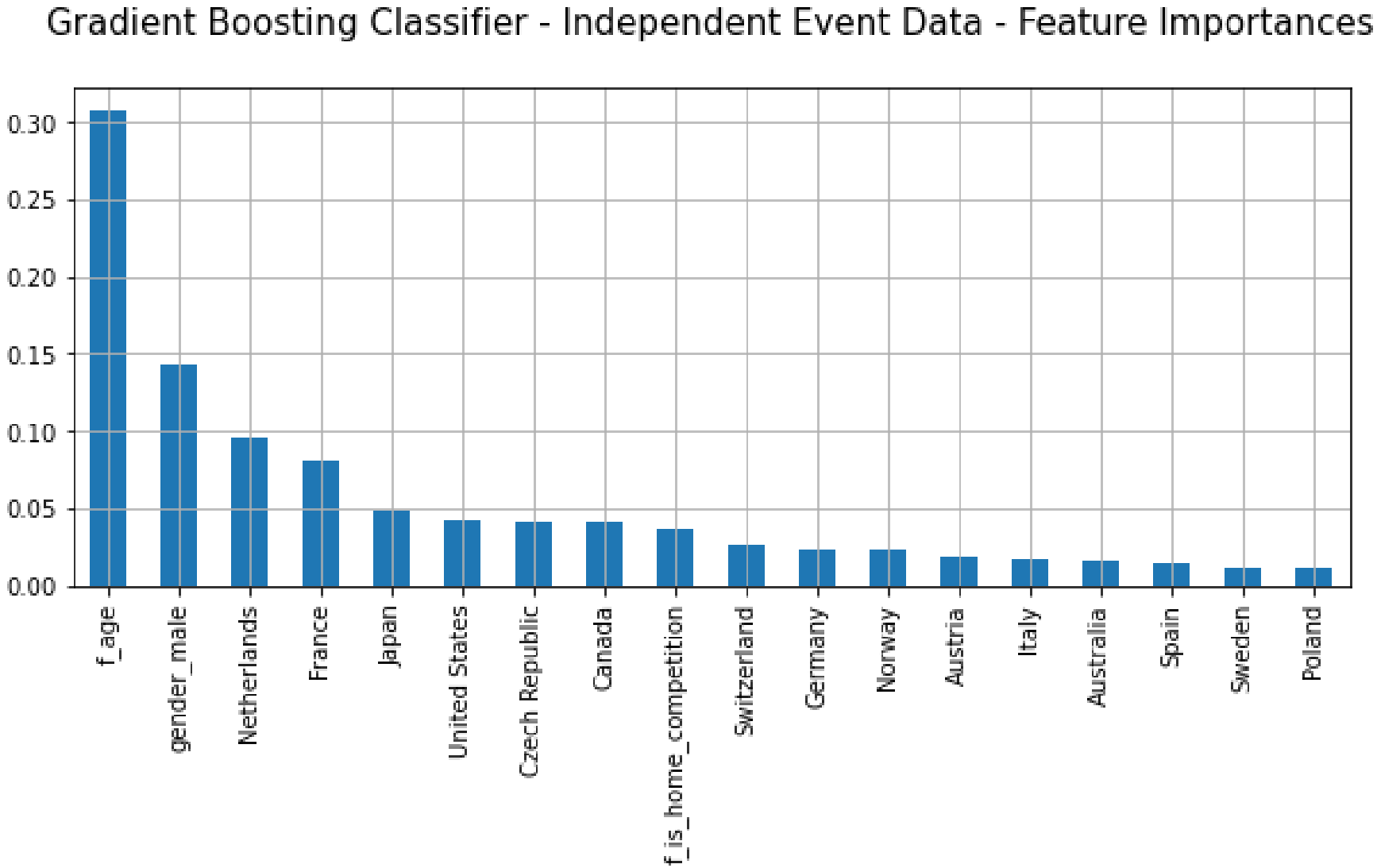
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Gradient Boosting Classifier with Independent Event Data:

	Medal Won	Precision	Recall	F1-Score
Training Set Results	False – 0	0.96	0.71	0.82
	True - 1	0.17	0.67	0.27
	Accuracy	-	-	0.71
	Macro Average	0.57	0.69	0.55
	Weighted Average	0.90	0.71	0.77

	Medal Won	Precision	Recall	F1-Score
Test Set Results	False – 0	0.95	0.70	0.80
	True - 1	0.16	0.59	0.25
	Accuracy	-	-	0.69
	Macro Average	0.55	0.64	0.53
	Weighted Average	0.88	0.69	0.76

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



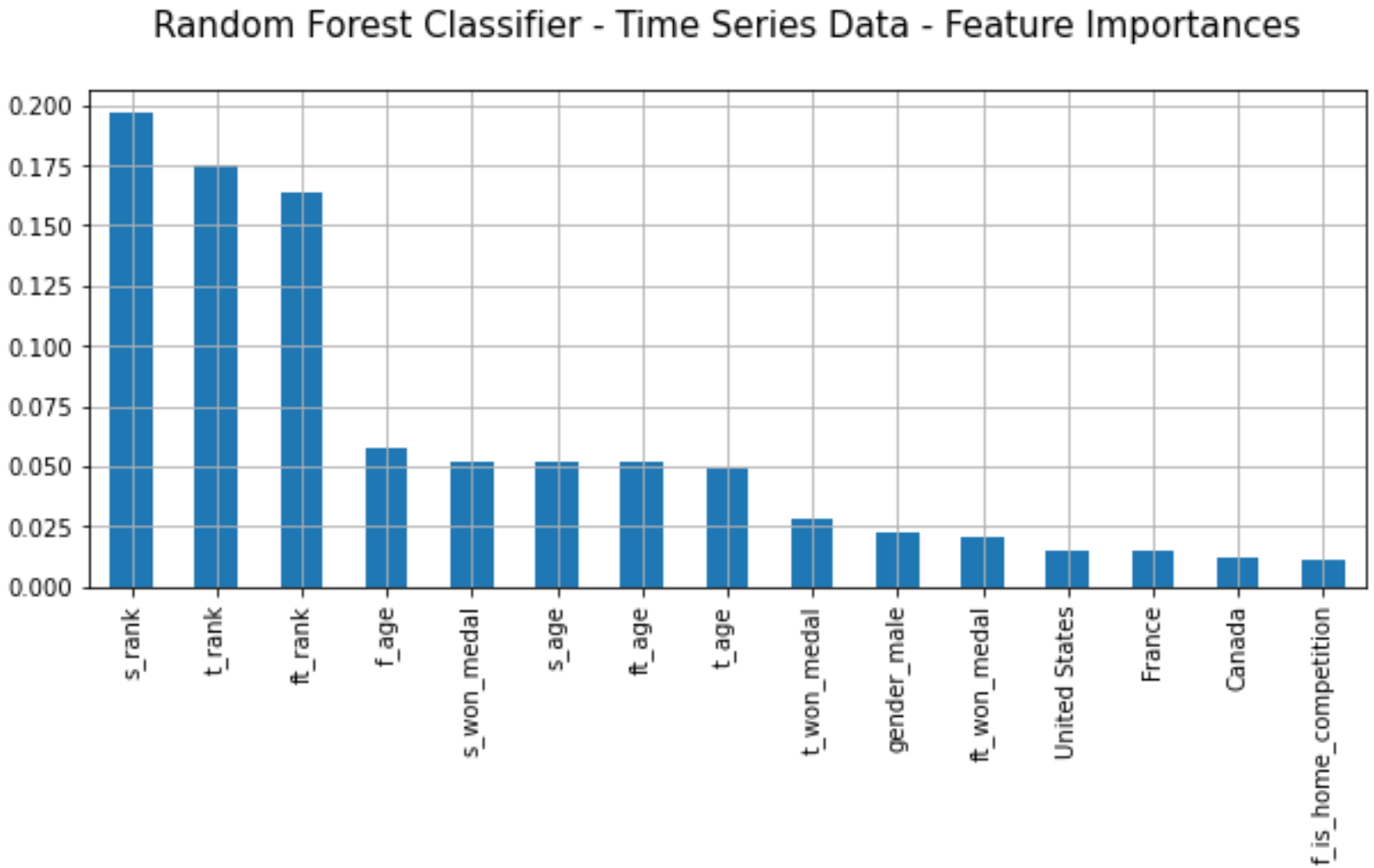
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Random Forest Classifier with Time Series Data:

	Medal Won	Precision	Recall	F1-Score
Training Set Results	False – 0	0.99	0.97	0.98
	True - 1	0.75	0.94	0.83
	Accuracy	-	-	0.97
	Macro Average	0.87	0.96	0.91
	Weighted Average	0.97	0.97	0.97

	Medal Won	Precision	Recall	F1-Score
Test Set Results	False – 0	0.95	0.95	0.95
	True - 1	0.42	0.42	0.42
	Accuracy	-	-	0.90
	Macro Average	0.68	0.68	0.68
	Weighted Average	0.90	0.90	0.90

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



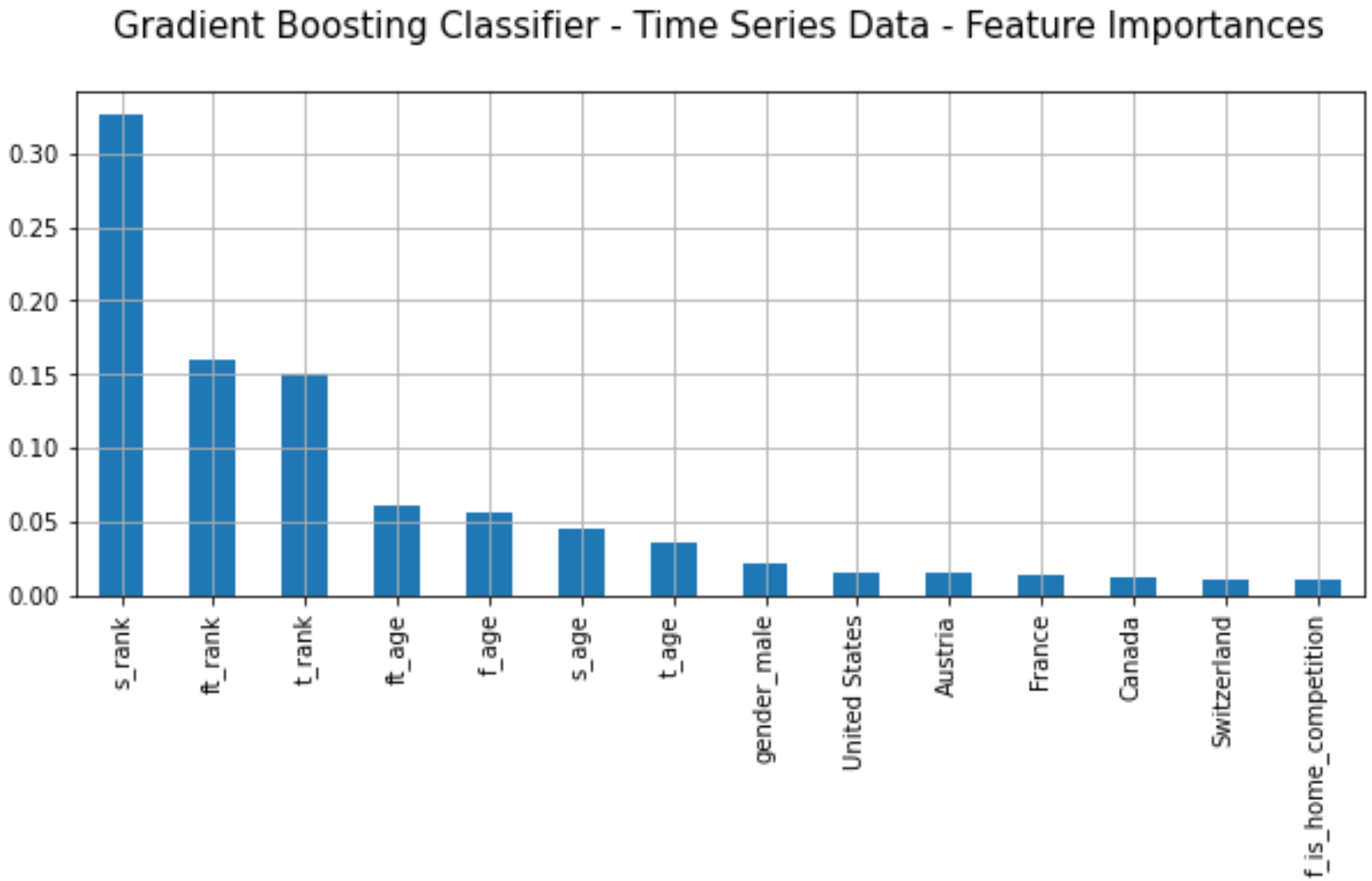
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Gradient Boosting Classifier with Time Series Data:

	Medal Won	Precision	Recall	F1-Score
Training Set Results	False – 0	1.00	1.00	1.00
	True - 1	0.97	1.00	0.98
	Accuracy	-	-	1.00
	Macro Average	0.98	1.00	0.99
	Weighted Average	1.00	1.00	1.00

	Medal Won	Precision	Recall	F1-Score
Test Set Results	False – 0	0.94	0.96	0.95
	True - 1	0.39	0.27	0.32
	Accuracy	-	-	0.90
	Macro Average	0.66	0.62	0.63
	Weighted Average	0.89	0.90	0.90

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Regressors with 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

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Regressors with 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:

- Independent Event Data:
  - The best performing classifier was the Random Forest
  - Best Precision Score on the test set was 0.16
  - Best performing regressor Mean Absolute Error on the test set was 12.72
- Time Series Data:
  - The best performing classifier was the Random Forest
  - The best performing classifier showed precision of 0.42 and recall of 0.42 in the testing set
  - Best performing regressor Mean Absolute Error on the test set was 9.72

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

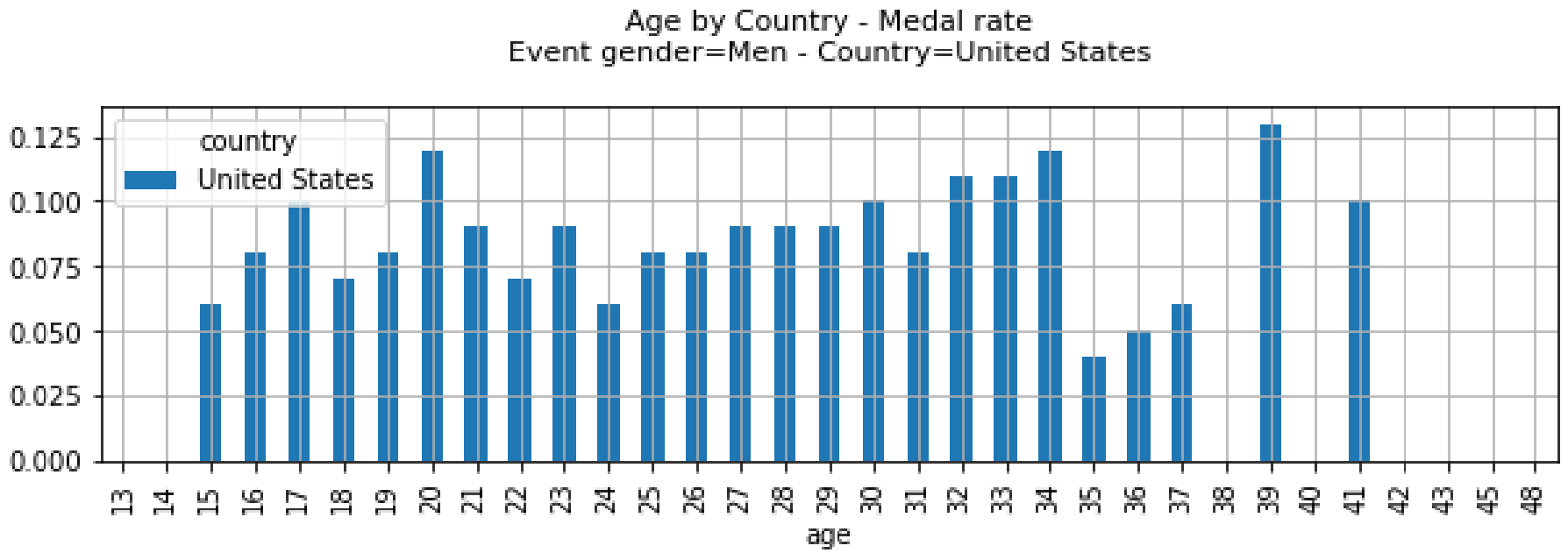
Key Questions:

- 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
  - Age cannot single-handedly be used to predict success
  - No clear-cut peak age

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Key Questions:

- Which athletes who competed in Tokyo 2021 (Beijing 2022) will be hitting their peak age in their respective sports in Paris 2024 (Milan 2026)?
  - No specific peak age for athletes that are more successful



Key Questions:

- Can the number of young, "pre-peak" athletes that competed in Tokyo predict medal success for their countries in Paris 2024 and Milan 2026?
  - Feasible to predict success in the upcoming Olympics
  - Prediction is not limited to Tokyo only
  - Other preceding events and athlete performance during those can be used as well

Key Questions:

- 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?
  - 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.

THANK YOU