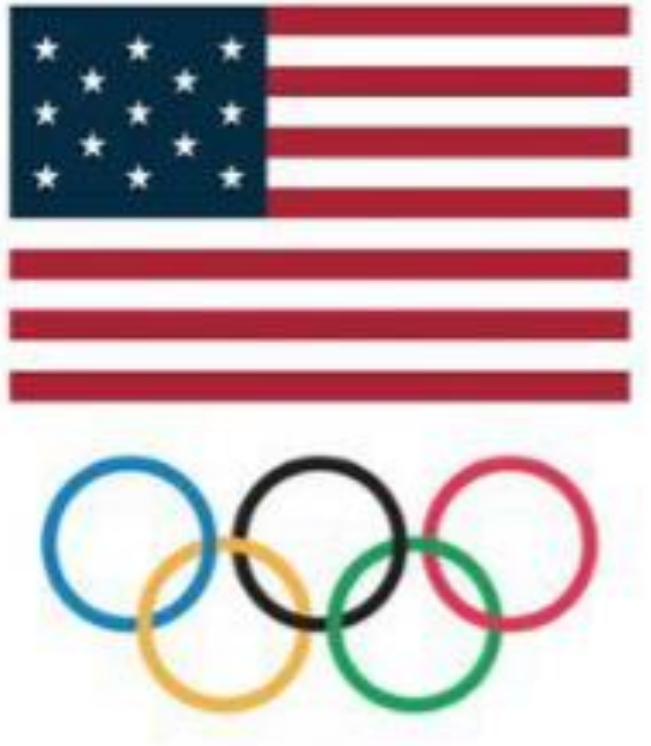


Stergios Koutrouvelis

USOPC
Age-Focused Olympic Competitive Analysis
Midterm Project Report

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?
- 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?

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 team names
  - Nation
  - Birth date
  - Competition type (e.g., Olympics, World Championships)
  - Sport, event, and/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
- Development of Machine Learning model that can predict the probability of an athlete winning a medal based on preprocessed features.

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

Midterm report items addressed:

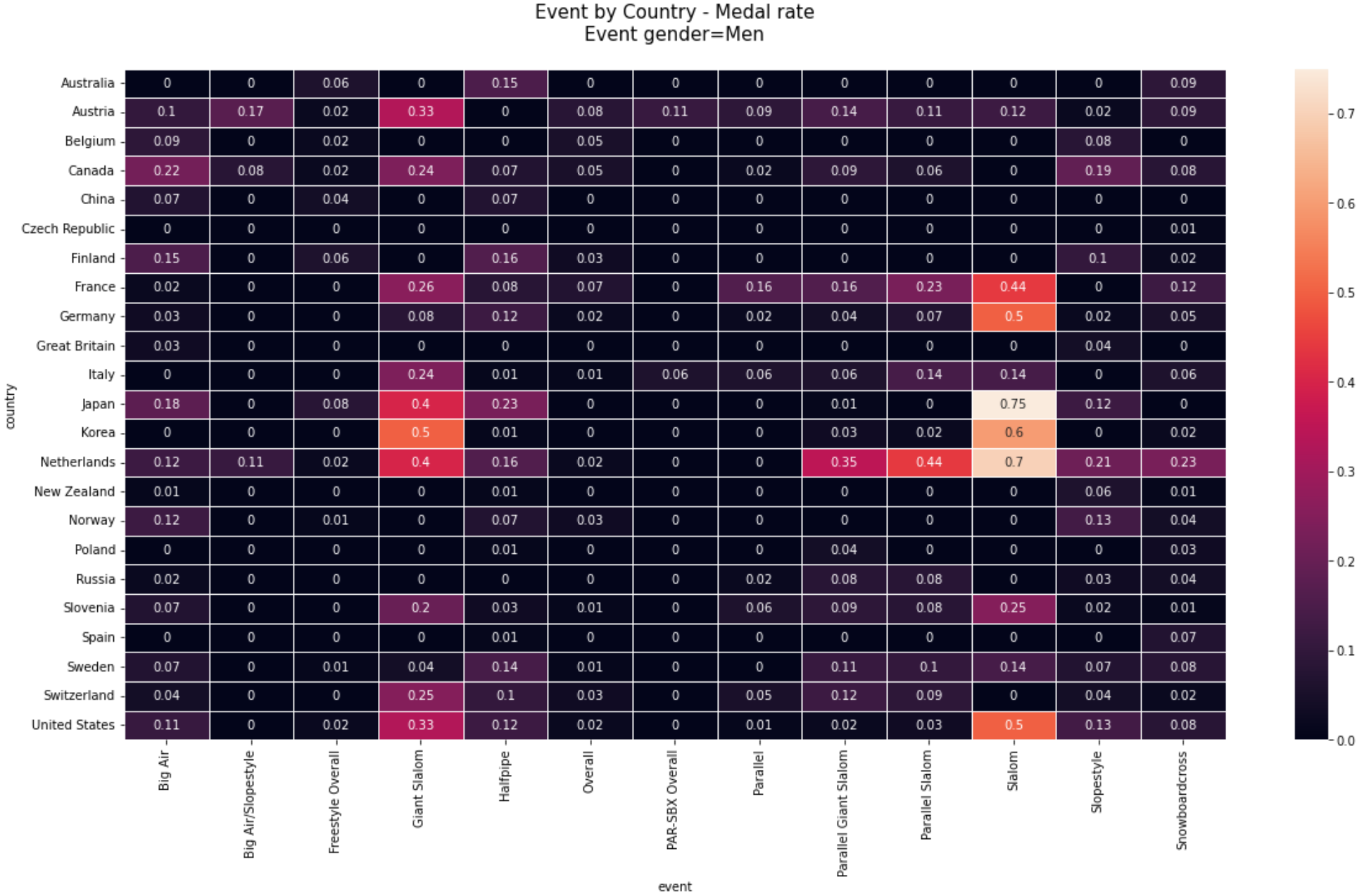
- Focused on Snowboarding (code is built to generalize to other sports).
- Data processing framework for raw datasets to create a structured dataset.
- API development that determines the competition country based on the competition city.
- Creation of new feature to factor in whether the competition is being held at the athlete's home country.
- Exploratory data analysis for country and events to determine:
  - Medal Rate: Success rate of athletes winning medals for this country by sport
  - Participation Count: Number of a country's athletes have participated in each one of the sports
  - Participation Rate: Proportion of a country's athletes that compete in each sport
- Exploratory data analysis for country and age to determine:
  - Medal Rate: Success rate of athletes winning medals for this country by age group
  - Participation Count: Number of a country's athletes that participated in each of the age groups
  - Participation Rate: Proportion of country's athletes by age group
- Analysis of medal rate when competitions are help at home vs away by country
- Development of preliminary Random Forest Classifier that predicts whether an athlete will win a medal

Data Processing Framework

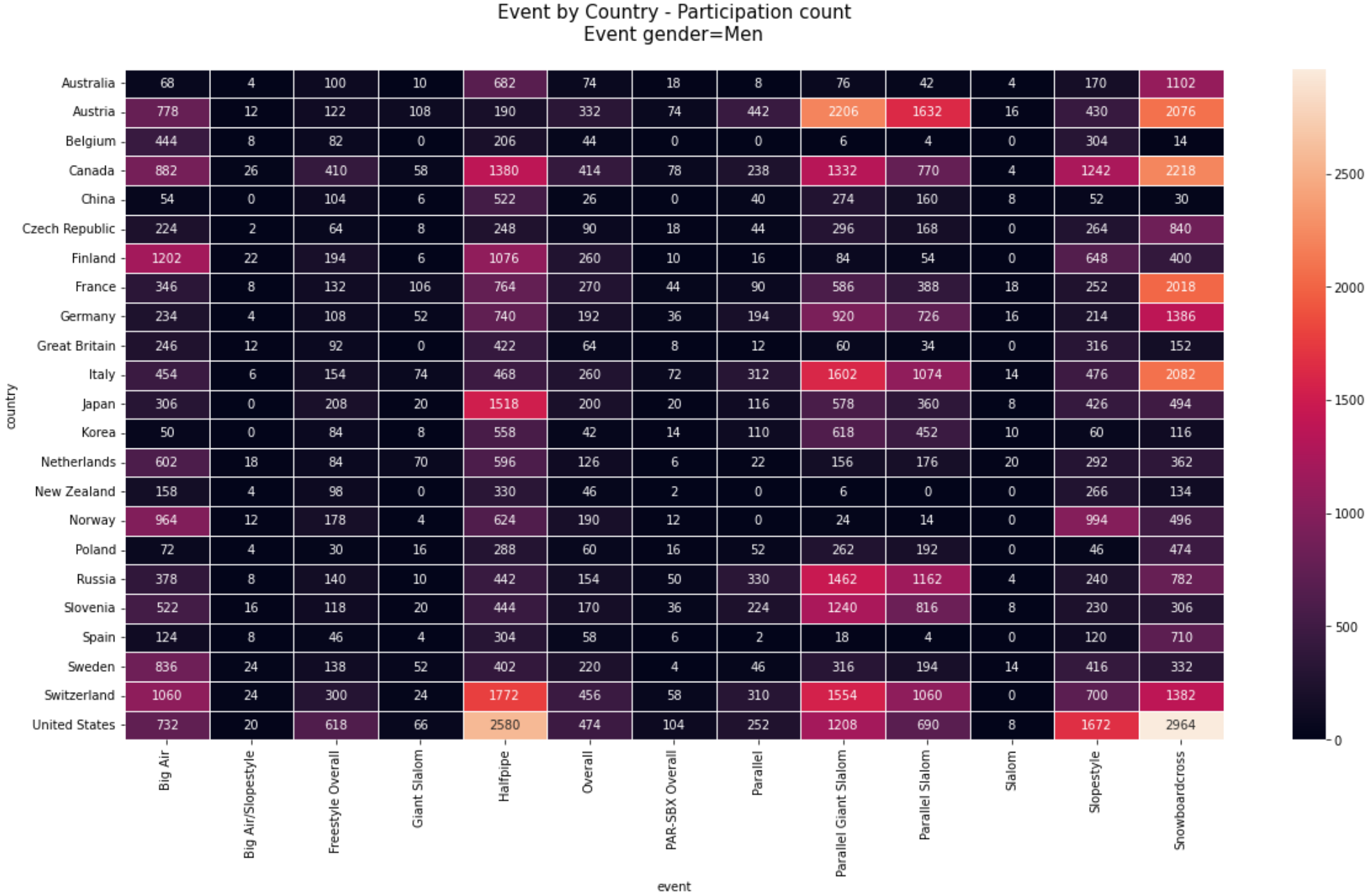
The framework is designed to extract/produce the following features from the raw data:

- **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

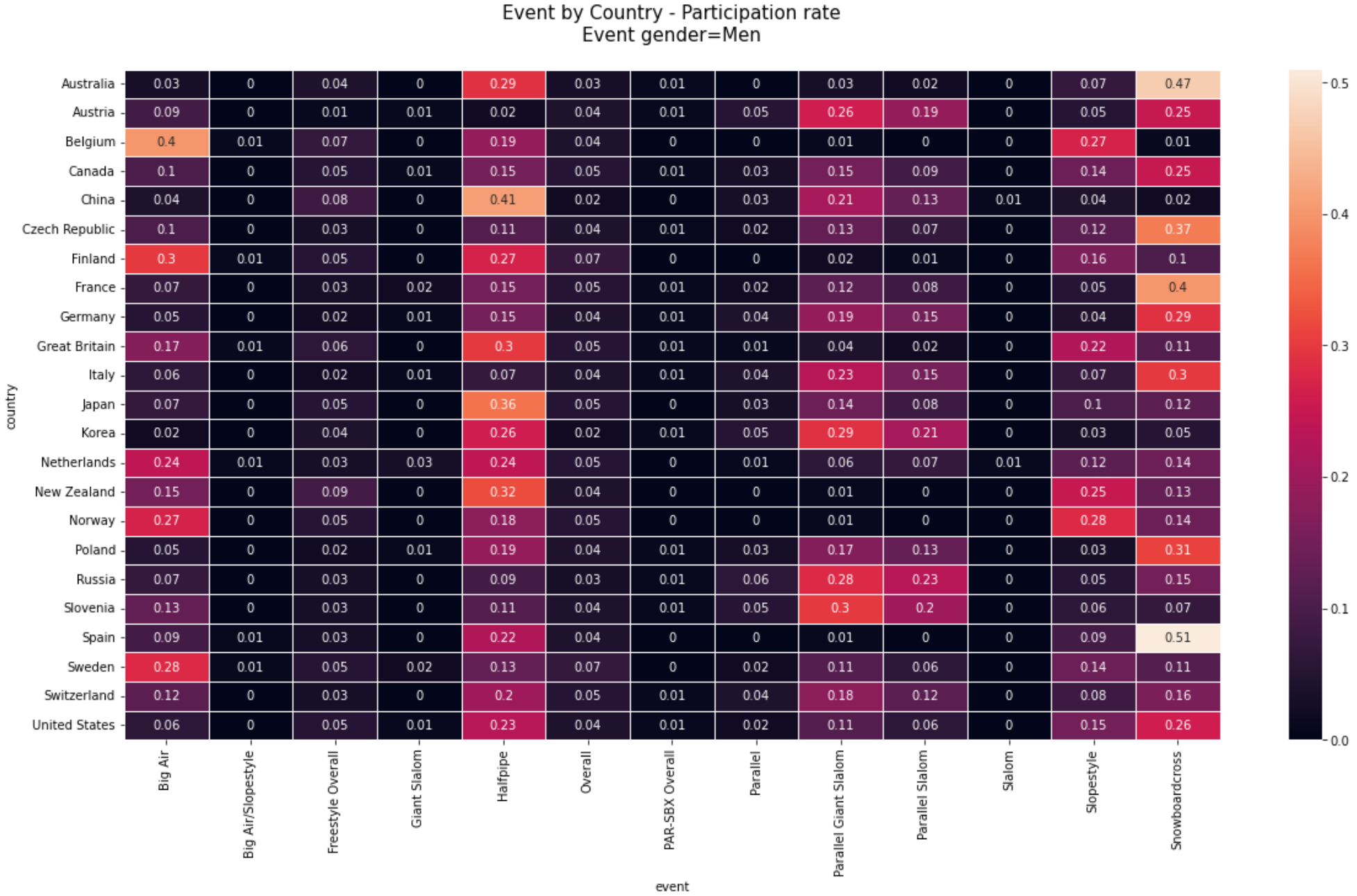
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



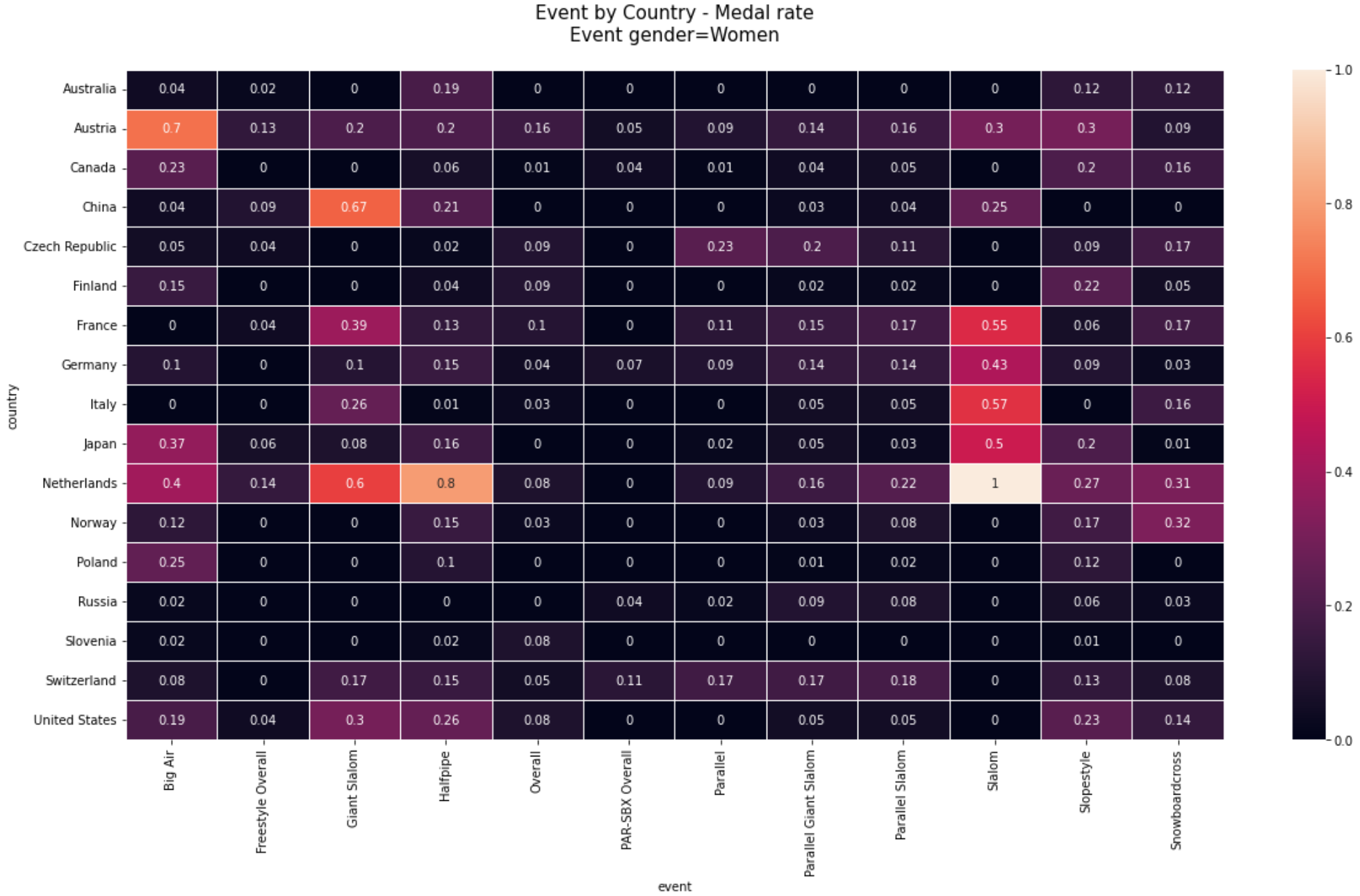
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



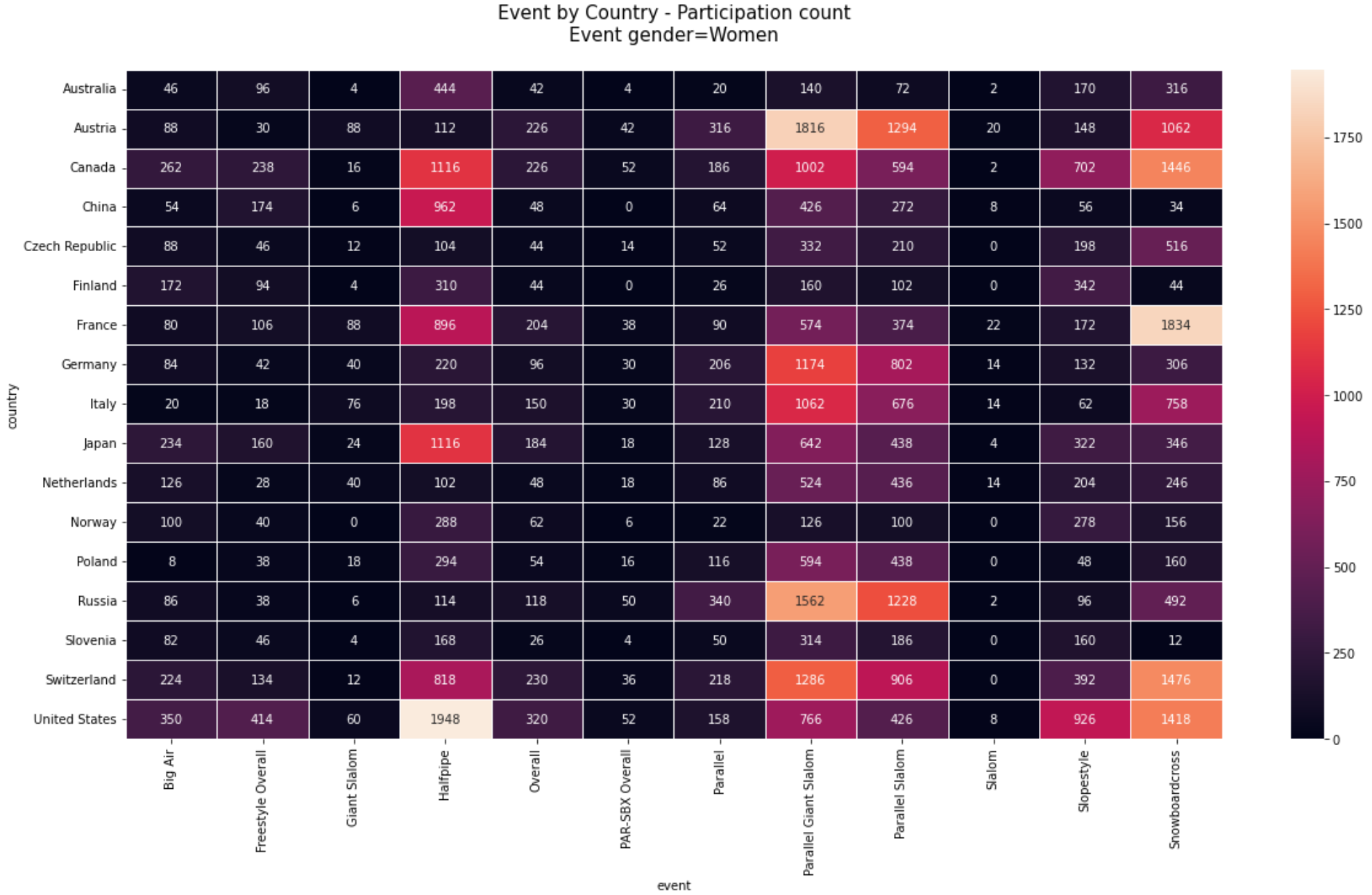
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



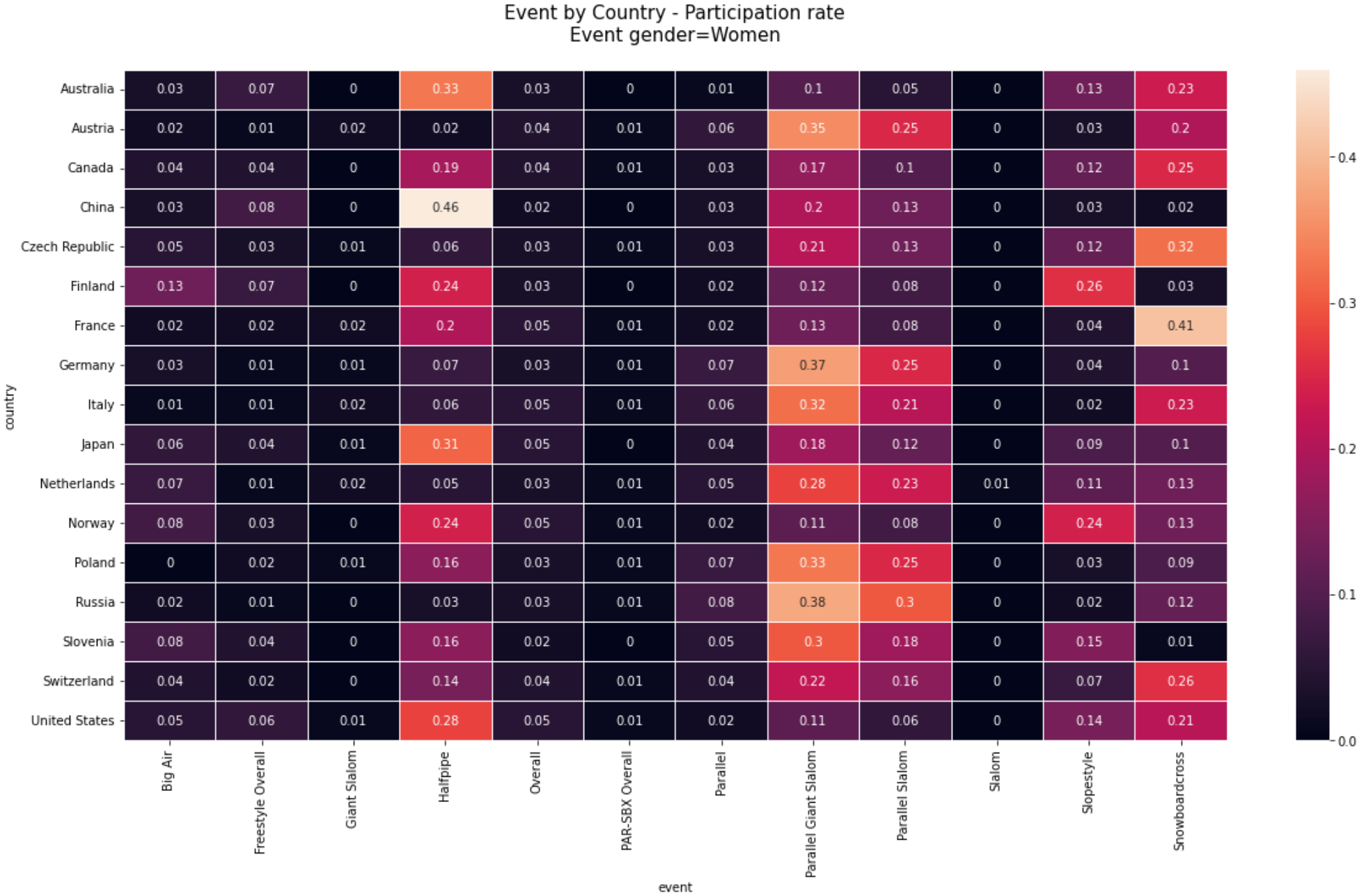
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



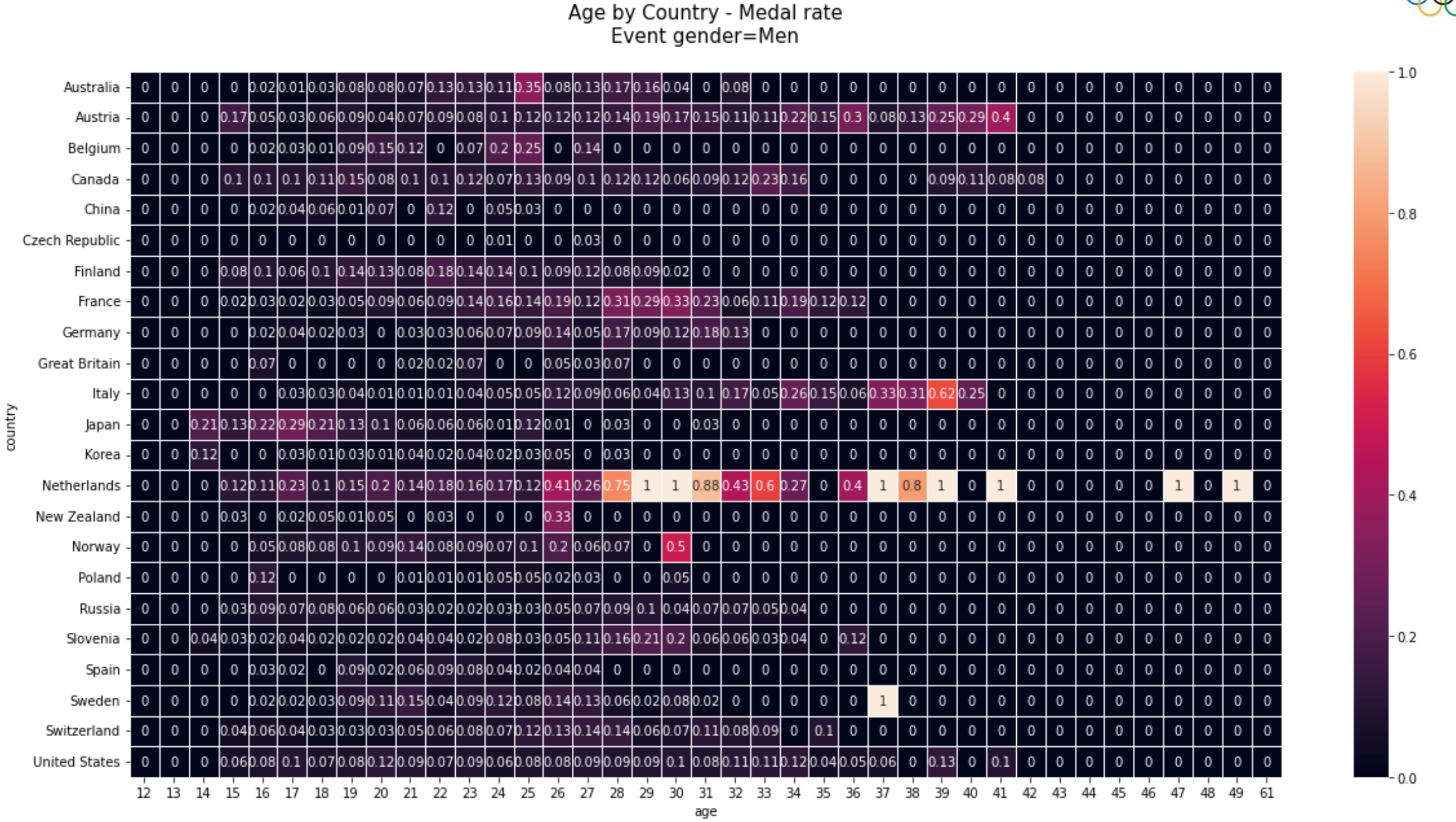
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



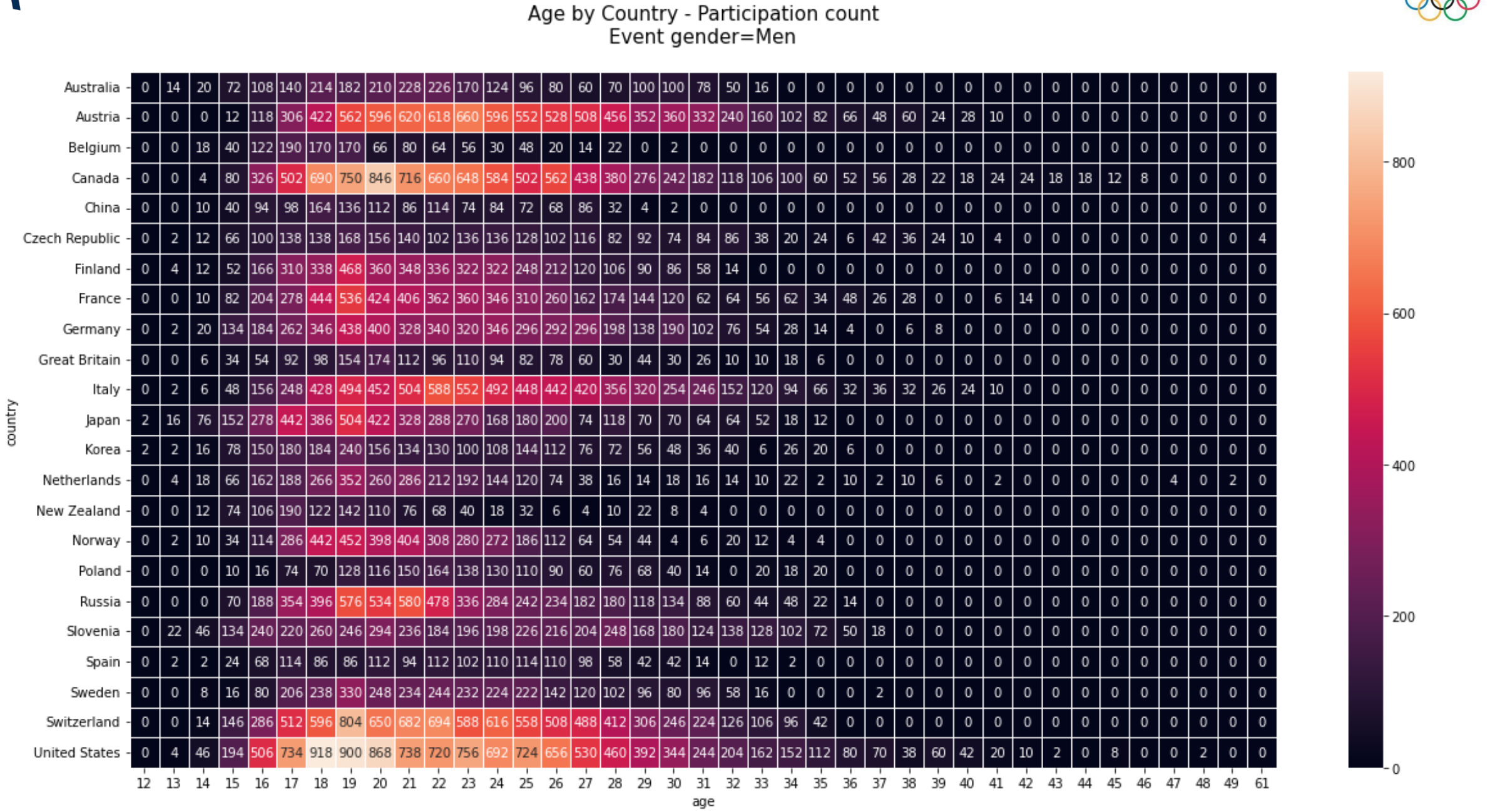
Country Statistics by Sport and Gender Conclusion:

- There are wide variations in terms of event focus by country of origin as illustrated by the participation rate by sport.
- The most popular events by participation rate are:
  - Snowboardcross
  - Slopestyle
  - Parallel Giant Slalom
  - Halfpipe
  - Parallel Slalom
  - Big Air
- Big Air has significantly higher participation rate in Men than 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, country and sporting event will be factored in when making predictions.

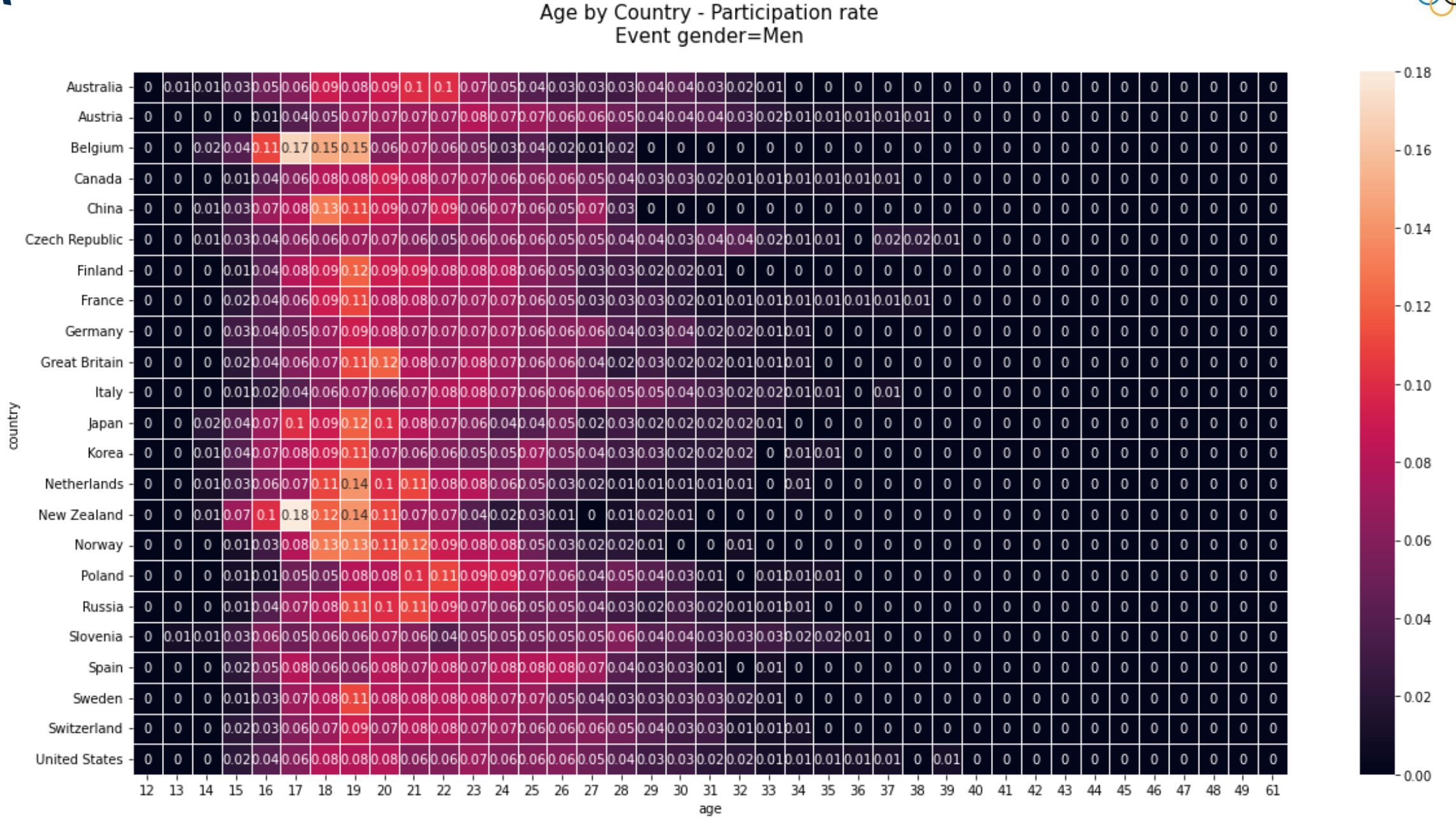
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



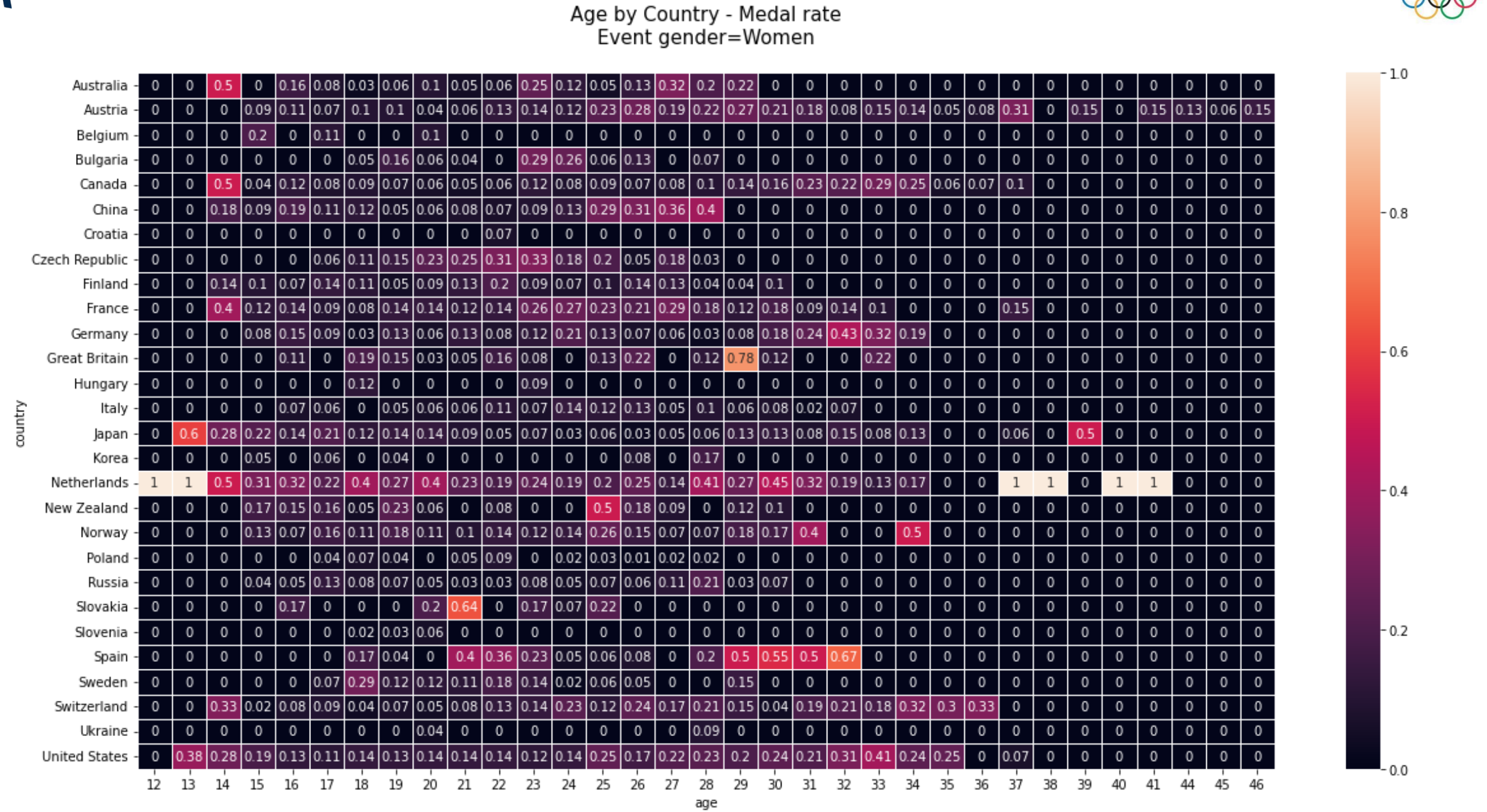
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



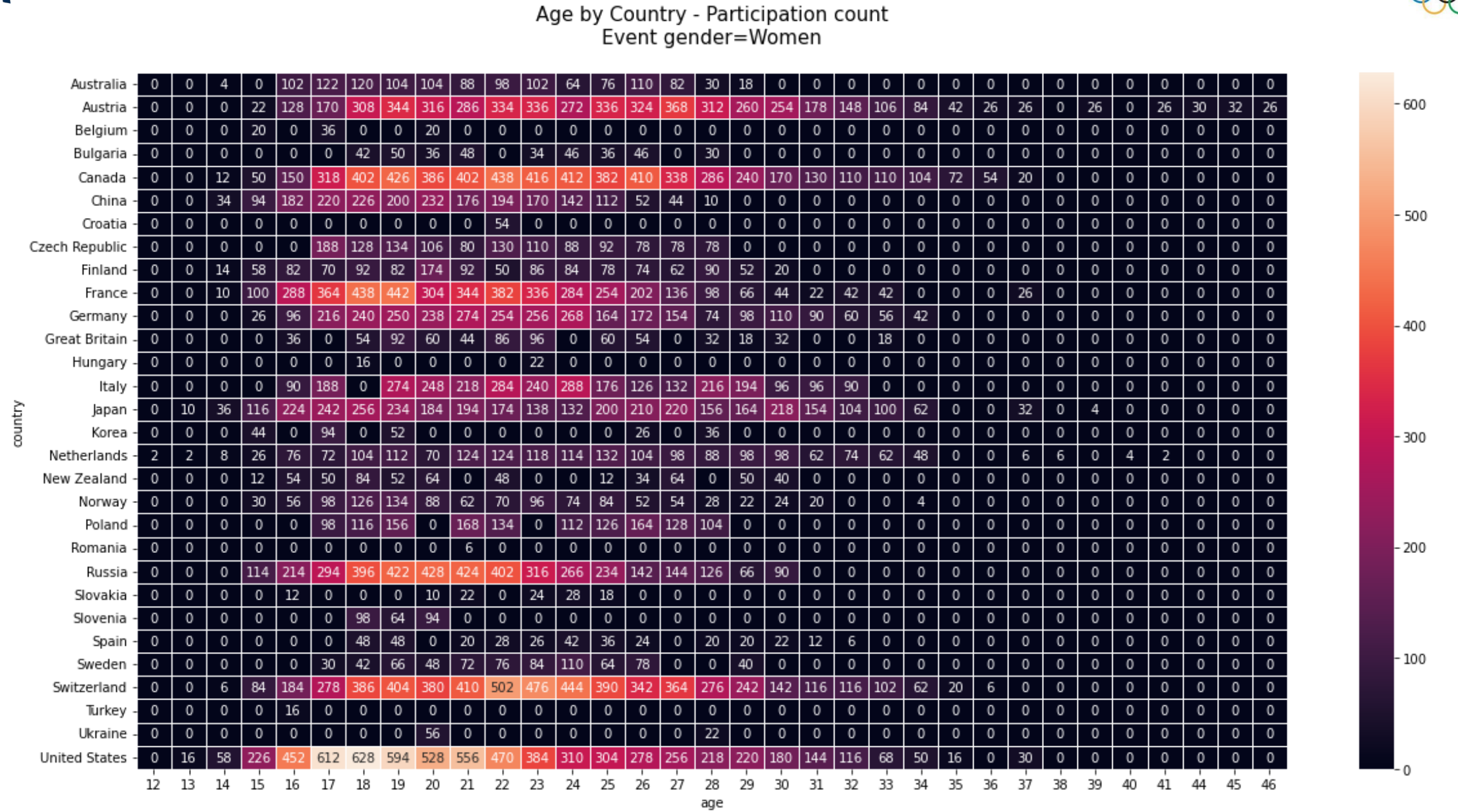
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



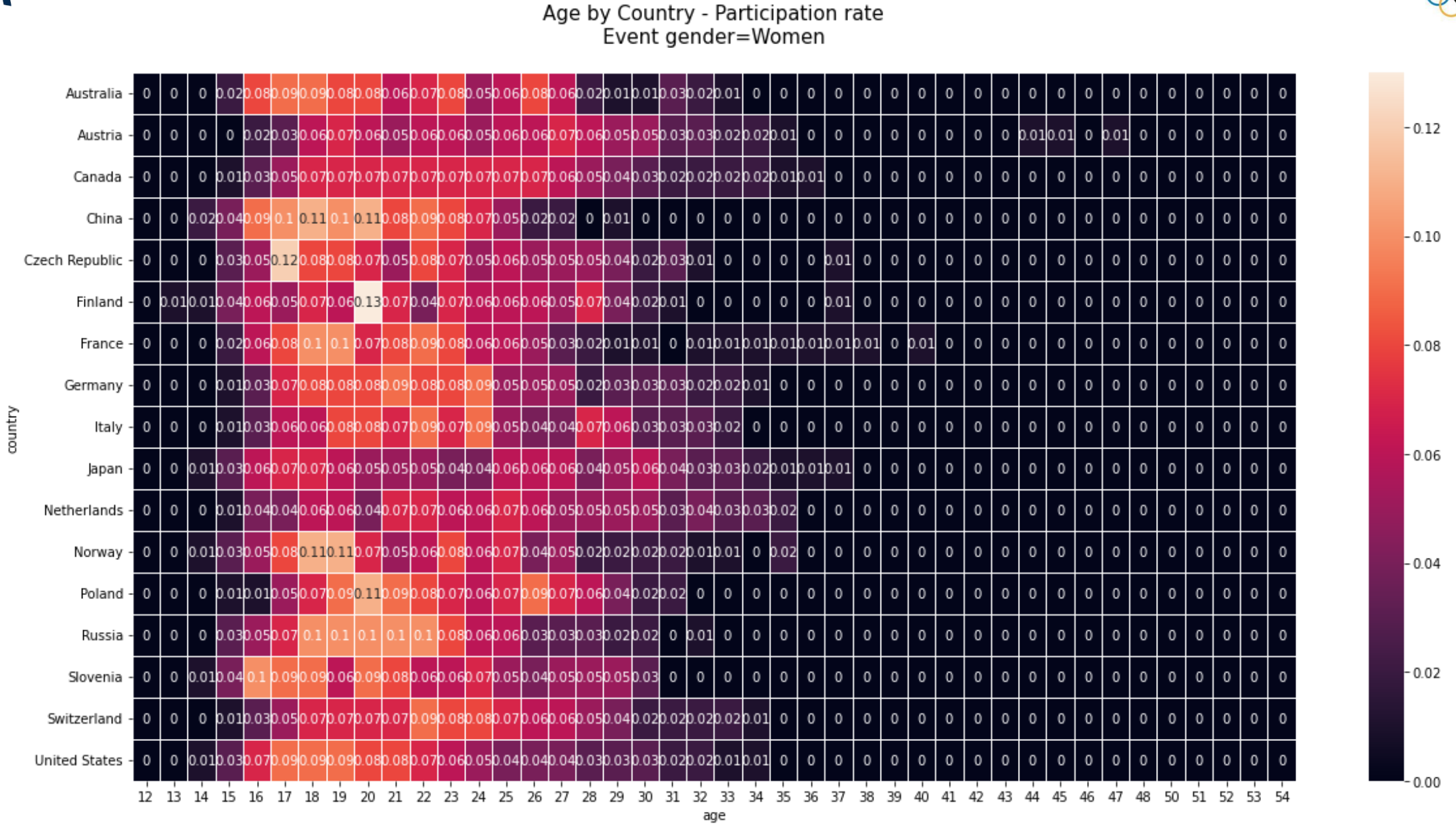
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



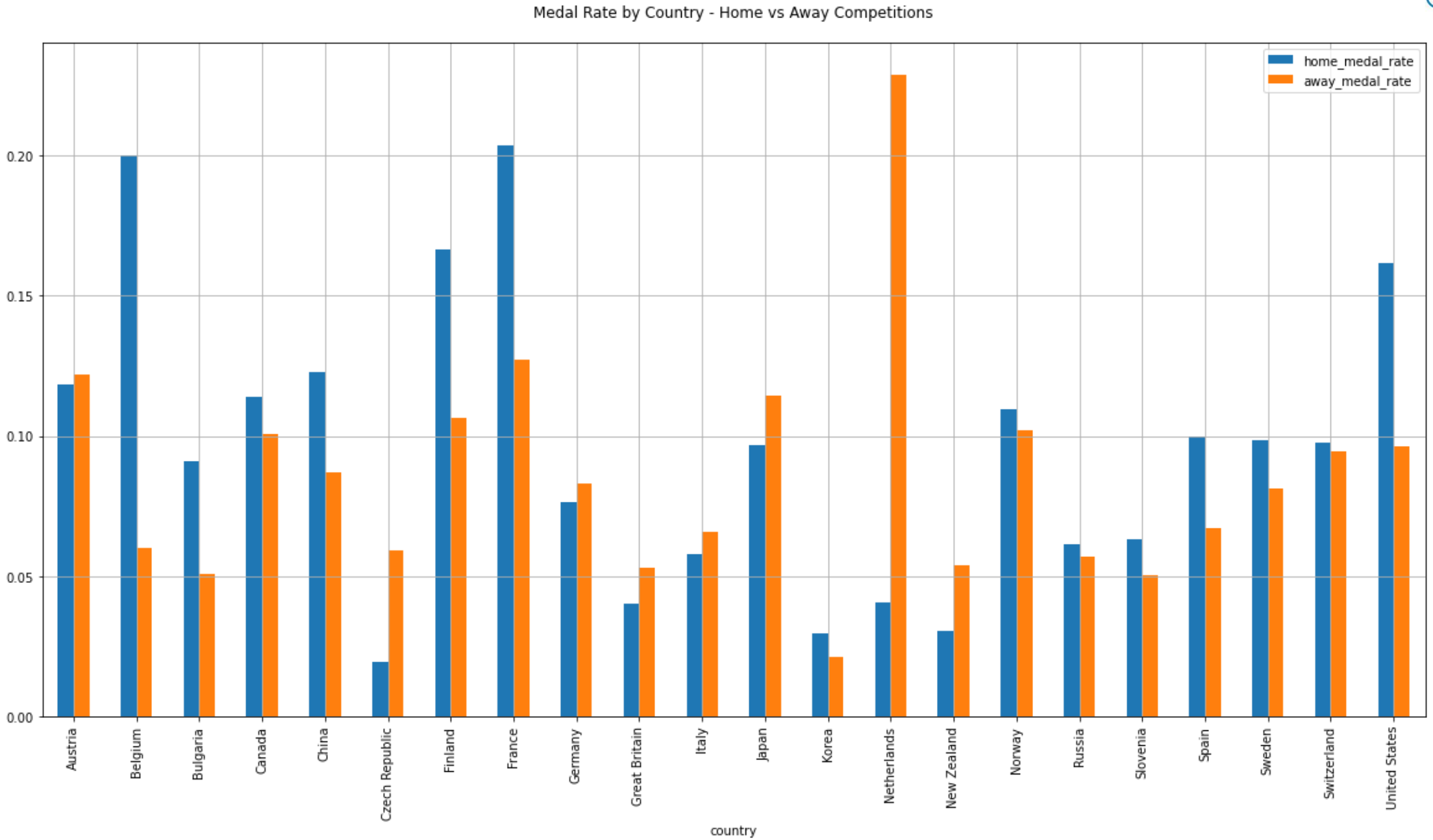
Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Country Statistics by Age and Gender Conclusion:

- Different countries exhibit different medal winning rates depending on age and gender.
- Very high medal winning rates are being observed by the Netherlands in higher ages.
- High medal rates appear to be more frequent in age ranges will 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 that most participants for this sport are around that age.
- There is a wider variation of the participation rate metric for women compared to men.
- For women in the United States in particular, there is a high medal winning rate for the age of early 30s while for men it is more evenly distributed.

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Home vs Away Competitions medal rate conclusion:

- Medal rate is higher when most countries compete at home vs away.
- Netherlands is the one of the only exceptions with a high away medal rate.
- For the United States in particular home competitions are much more successful.
- This is an indicator that a feature indicating whether an upcoming competition is being held at home versus away can provide valuable information when predicting the probability of winning a medal.

Random Forest Classifier for medal winning prediction:

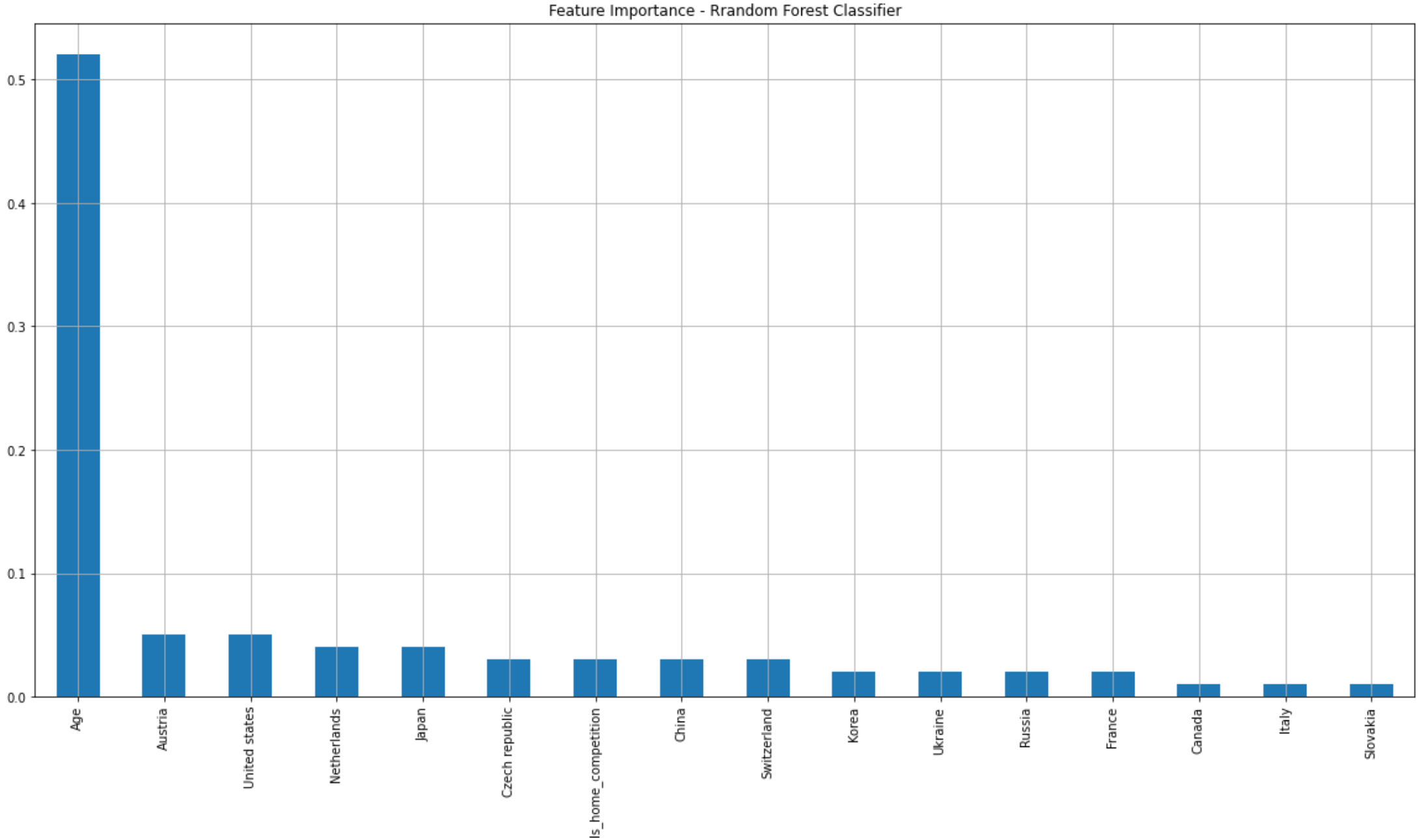
- A preliminary classifier has been developed to evaluate the power of raw features in predicting medal success.
- The classifier is being developed separately for each sporting event. For this example, the event of Parallel Giant Slalom has been evaluated.
- The features used are the following:
  - Gender: Binary variable denoted the gender of the athlete and event
  - Country: Country of origin of the athlete. One-hot-encoding has been performed here.
  - Is home competition: Binary variable indicating if the competition has been held at athlete's home country.
  - Age: Year of age for the respective athlete
- Response variable:
  - Won medal: Binary variable indicating whether the athlete won a medal.
  - The raw dataset is highly imbalanced. The reason for that is that on any given competition most athletes do not win medals. The ratio of datapoints winning medals to non-winning medals is ~1:14.
- The classifier has been evaluated for its performance on the training and test set.
- Feature importance has been extracted from the model parameters to determine the most critical parameters.

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis

	Medal Won	Precision	Recall	F1-Score
Training Set Results	False – 0	0.98	0.66	0.79
	True - 1	0.17	0.83	0.28
	Accuracy	-	-	0.67
	Macro Average	0.57	0.75	0.54
	Weighted Average	0.92	0.67	0.75

	Medal Won	Precision	Recall	F1-Score
Test Set Results	False – 0	0.97	0.65	0.78
	True - 1	0.16	0.75	0.26
	Accuracy	-	-	0.66
	Macro Average	0.56	0.70	0.52
	Weighted Average	0.90	0.66	0.74

Stergios Koutrouvelis - USOPC - Age-Focused Olympic Competitive Analysis



Random Forest Classifier Performance and Conclusion:

- Feature importance analysis indicates that age is affecting the medal outcome far more than other parameters.
- Country of origin is the second most important parameter here, with certain countries having much higher success rates.
- The United States, Austria, Netherlands and Japan are countries of origin that have higher success rates and athletes from these countries have higher probability of success.
- The metrics that the model has been evaluated against are :
  - Precision
  - Recall
  - Accuracy
  - F1-Score
- The results indicated that the model is suffering from high bias. More specifically, performance was low in the medal-winning class both in the training and testing set.
- More features will be required to address the high bias problem in the model.

Next Steps and Suggested Future Work:

- Deep dive into the US Team Statistics and expansion of the Exploratory Data Analysis for the United States.
- Evaluation of potential additional features for improving classifier performance.
- Evaluation of medal winning performance based on individual athlete historical competition record, introducing a time-series based approach by using past performance to predict future performance.
- Evaluation of additional Machine Learning algorithms and models:
  - Logistic Regression
  - Decision Tree Classifier
  - Gradient Boosting Decision Trees
- Exploration of other winter sports like Freestyle Skiing.
- Expansion of API functionality for use in other sports:
  - Exploratory Data Analysis
  - Automated Statistics
  - Feature engineering
  - Parameter Selection
  - Machine Learning model development