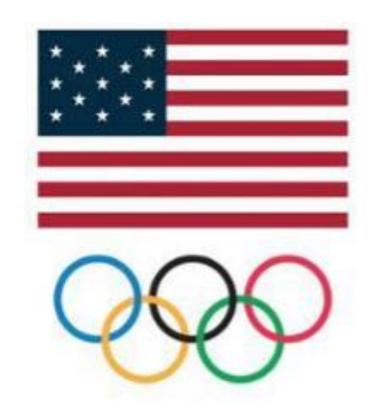
IEAM



Stergios Koutrouvelis USOPC

Age-Focused Olympic Competitive Analysis
Midterm Project Report





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?





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)





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.





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
- <u>Competition Country</u>: 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"
- <u>Medal:</u> 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.
- <u>Is Home Competition:</u> 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





- 0.7

- 0.6

- 0.5

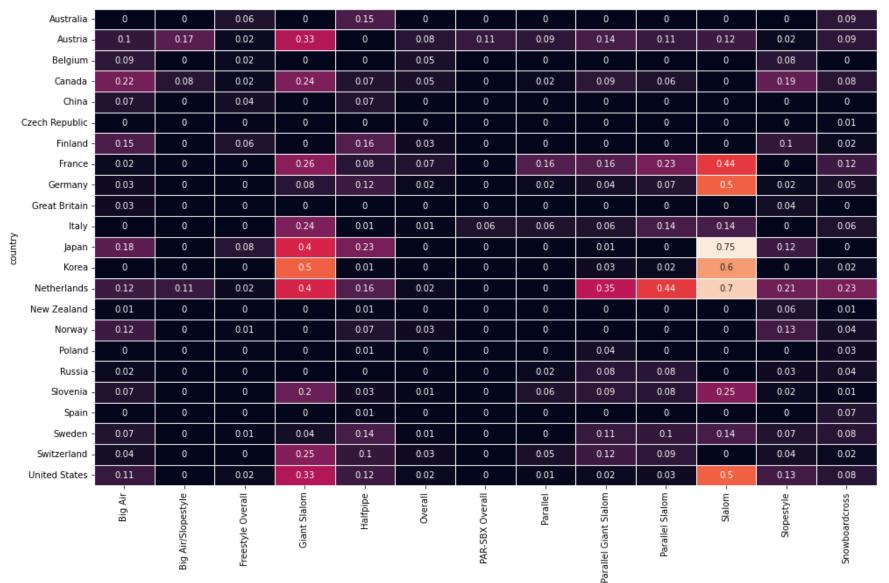
- 0.4

- 0.3

- 0.2

- 0.1

Event by Country - Medal rate Event gender=Men







- 2500

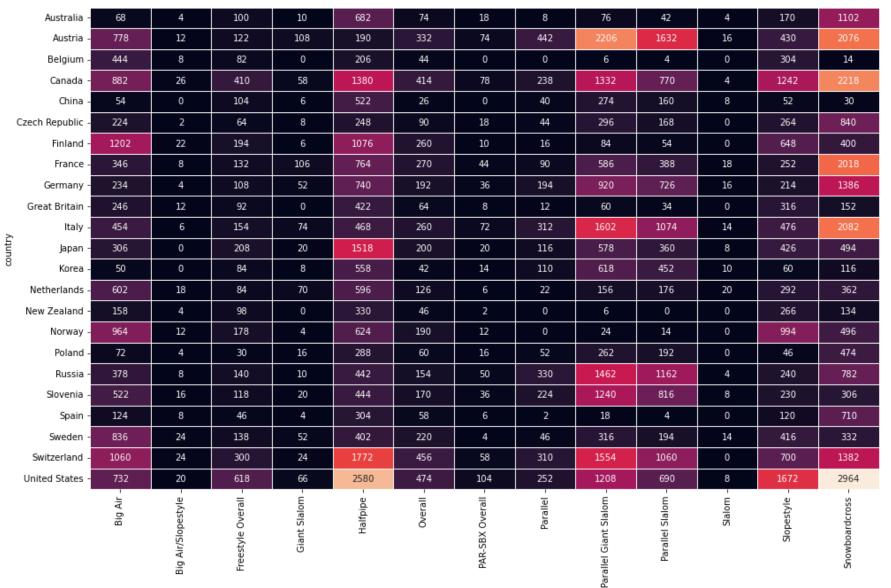
- 2000

- 1500

- 1000

- 500

Event by Country - Participation count Event gender=Men







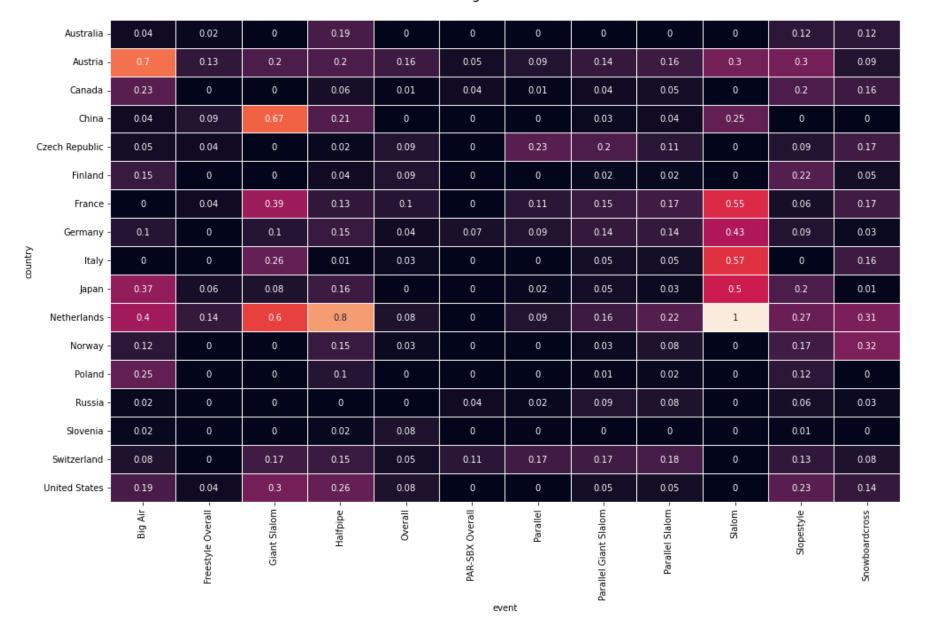
Event by Country - Participation rate Event gender=Men







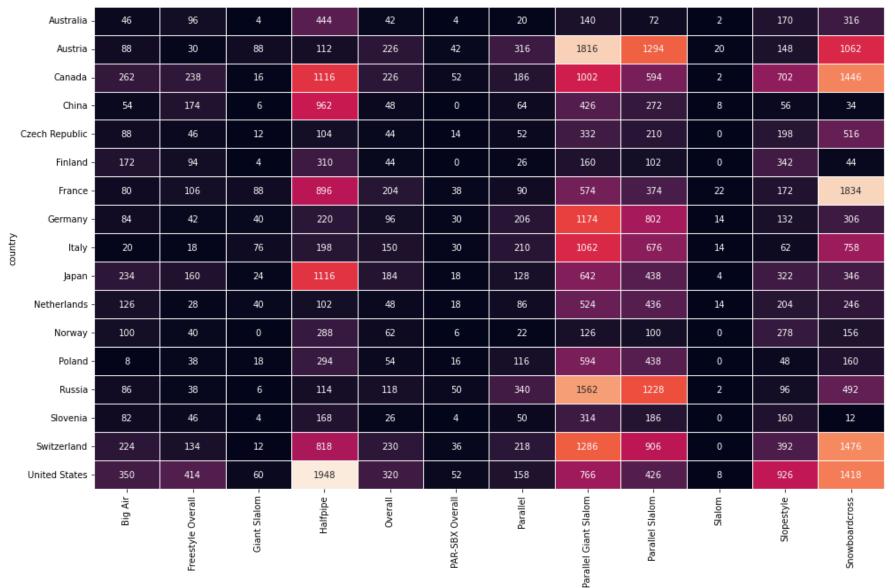
Event by Country - Medal rate Event gender=Women







Event by Country - Participation count Event gender=Women



- 1750 - 1500 - 1250 - 1000 - 750 - 500 - 250





Event by Country - Participation rate Event gender=Women







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.





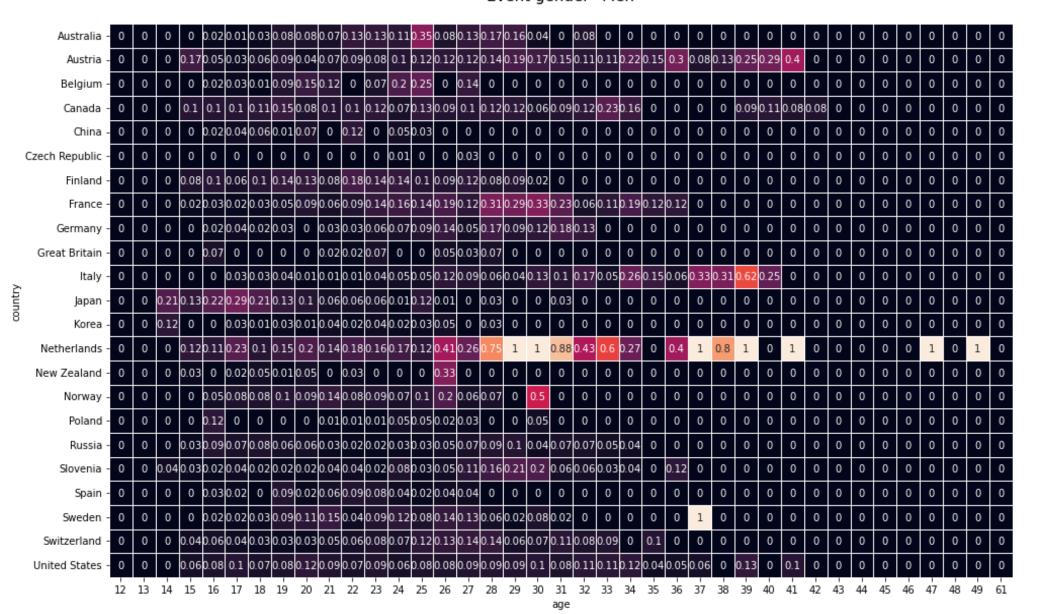
- 0.8

- 0.6

- 0.4

- 0.2

Age by Country - Medal rate Event gender=Men







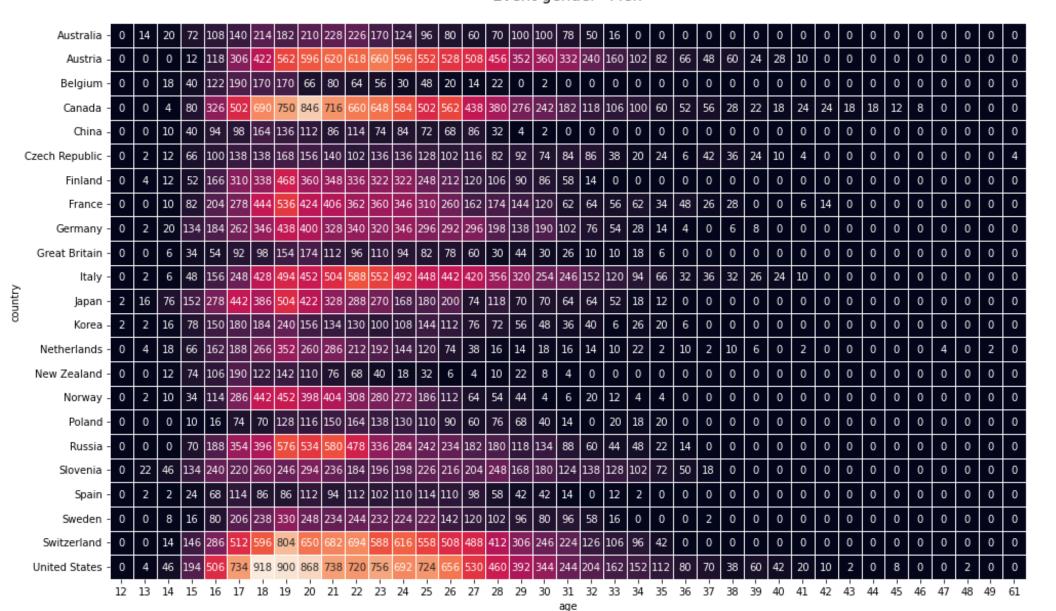
- 800

- 600

- 400

- 200

Age by Country - Participation count Event gender=Men







-0.18

-0.16

-0.14

-0.12

-0.10

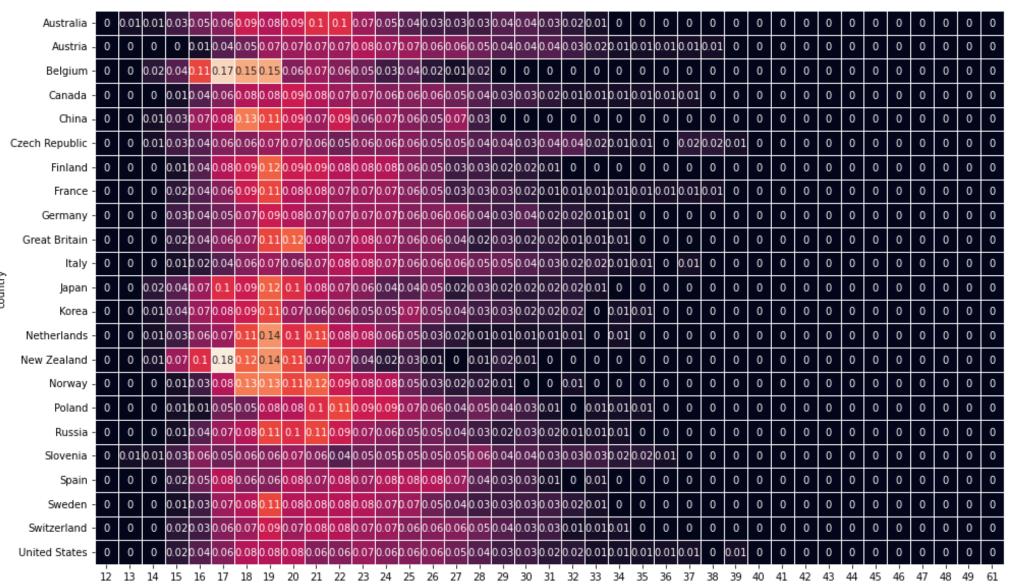
- 0.08

-0.06

-0.04

-0.02

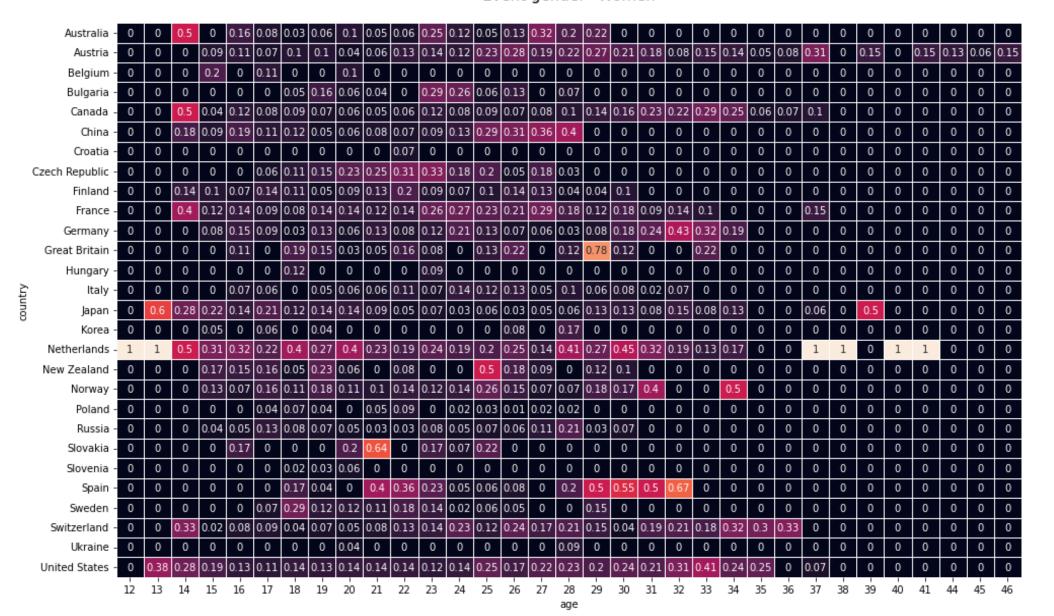
Age by Country - Participation rate Event gender=Men







Age by Country - Medal rate Event gender=Women



- 1.0

- 0.8

- 0.6

- 0.4

- 0.2

- 0 0





Age by Country - Participation count Event gender=Women

	Australia -	0	0	4	0	102	122	120	104	104	88	98	102	64	76	110	82	30	18	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	0	26	0	26	30	32	26
	Belgium -	0	0	0	20	0	36	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bulgaria -	0	0	0	0	0	0	42	50	36	48	0	34	46	36	46	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Canada -	0	0	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
	China -	0	0	34	94	182	220	226	200	232	176	194	170	142	112	52	44	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Croatia -	0	0	0	0	0	0	0	0	0	0	54	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	0	0	0	188	128	134	106	80	130	110	88	92	78	78	78	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Finland -	0	0	14	58	82	70	92	82	174	92	50	86	84	78	74	62	90	52	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	France -	0	0	10	100	288	364	438	442	304	344	382	336	284	254	202	136	98	66	44	22	42	42	0	0	0	26	0	0	0	0	0	0	0
	Germany -	0	0	0	26	96	216	240	250	238	274	254	256	268	164	172	154	74	98	110	90	60	56	42	0	0	0	0	0	0	0	0	0	0
	Great Britain -	0	0	0	0	36	0	54	92	60	44	86	96	0	60	54	0	32	18	32	0	0	18	0	0	0	0	0	0	0	0	0	0	0
	Hungary -	0	0	0	0	0	0	16	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Italy -	0	0	0	0	90	188	0	274	248	218	284	240	288	176	126	132	216	194	96	96	90	0	0	0	0	0	0	0	0	0	0	0	0
itry	Japan -	0	10	36	116	224	242	256	234	184	194	174	138	132	200	210	220	156	164	218	154	104	100	62	0	0	32	0	4	0	0	0	0	0
country	Korea -	0	0	0	44	0	94	0	52	0	0	0	0	0	0	26	0	36	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	0	0	6	6	0	4	2	0	0	0
	New Zealand -	0	0	0	12	54	50	84	52	64	0	48	0	0	12	34	64	0	50	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Norway -	0	0	0	30	56	98	126	134	88	62	70	96	74	84	52	54	28	22	24	20	0	0	4	0	0	0	0	0	0	0	0	0	0
	Poland -	0	0	0	0	0	98	116	156	0	168	134	0	112	126	164	128	104	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Romania -	0	0	0	0	0	0	0	0	0	6	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	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Slovakia -	0	0	0	0	12	0	0	0	10	22	0	24	28	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Slovenia -	0	0	0	0	0	0	98	64	94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Spain -	0	0	0	0	0	0	48	48	0	20	28	26	42	36	24	0	20	20	22	12	6	0	0	0	0	0	0	0	0	0	0	0	0
	Sweden -	0	0	0	0	0	30	42	66	48	72	76	84	110	64	78	0	0	40	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
	Turkey -	0	0	0	0	16	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	0	0	0
	Ukraine -	0	0	0	0	0	0	0	0	56	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	United States -	0	16	58	226	452	612	628	594	528	556	470	384	310	304	278	256	218	220	180	144	116	68	50	16	0	30	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	44	45	46

- 600

- 500

400

300

200

100

- 0





-0.12

-0.10

- 0.08

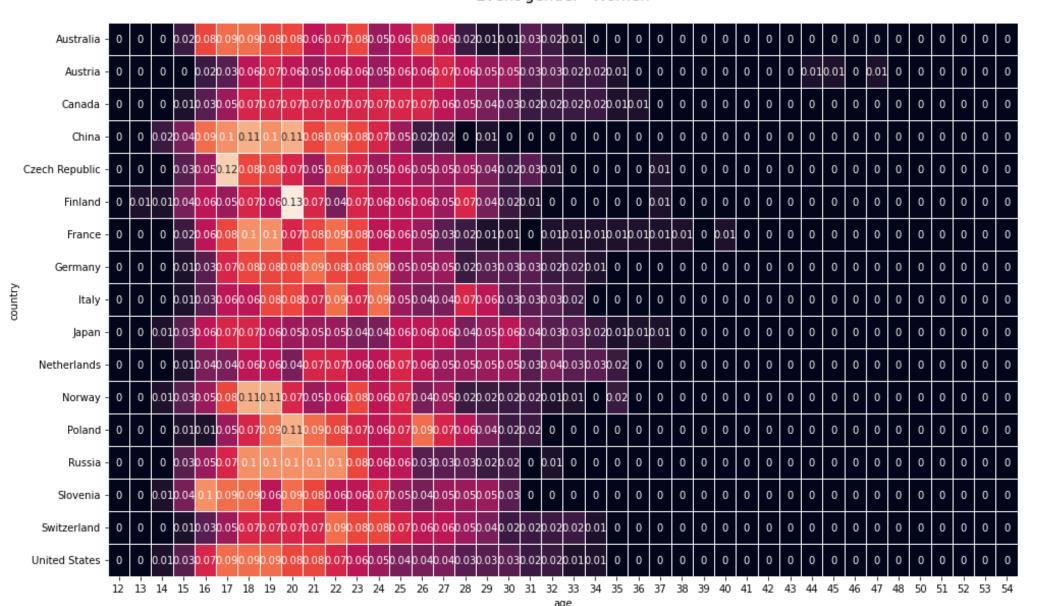
-0.06

- 0.04

-0.02

-0.00

Age by Country - Participation rate Event gender=Women







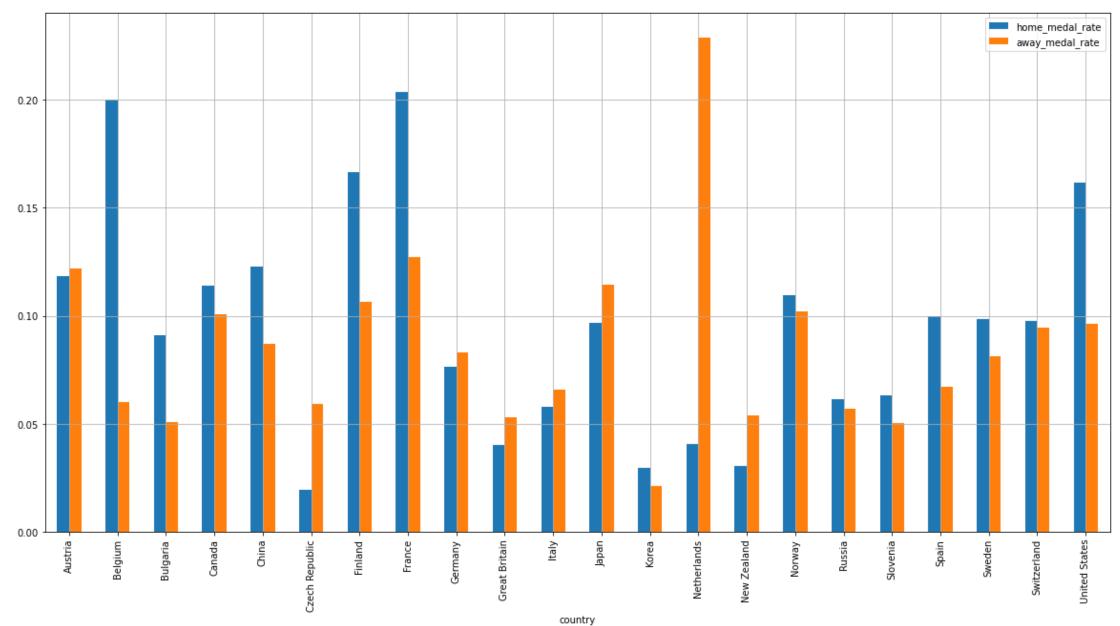
Country Statistics by Age and Gender Conclusion:

- 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.





Medal Rate by Country - Home vs Away Competitions







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.





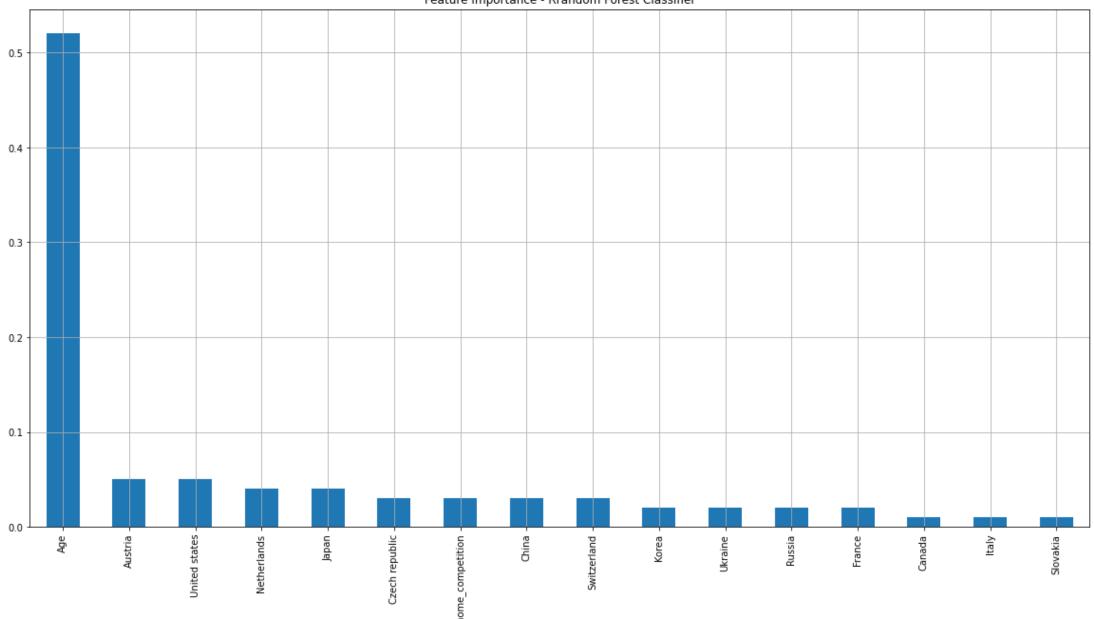
	Medal Won	Precision	Recall	F1-Score
	False – 0	0.98	0.66	0.79
Training	True - 1	0.17	0.83	0.28
Set				
Results	Accuracy	-	-	0.67
	Macro Average	0.57	0.75	0.54
	Weighted Average	0.92	0.67	0.75

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





Feature Importance - Rrandom Forest Classifier







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