OLYMPIC ATHLETES:

Using data to predict Olympic medalists

Problem

Countries engage in friendly international sports competition every two years in Olympic Games

Olympic committees scout for the best athletes to compete for medals

What if less money could be spent on scouting and more on training?

Hypothesis:

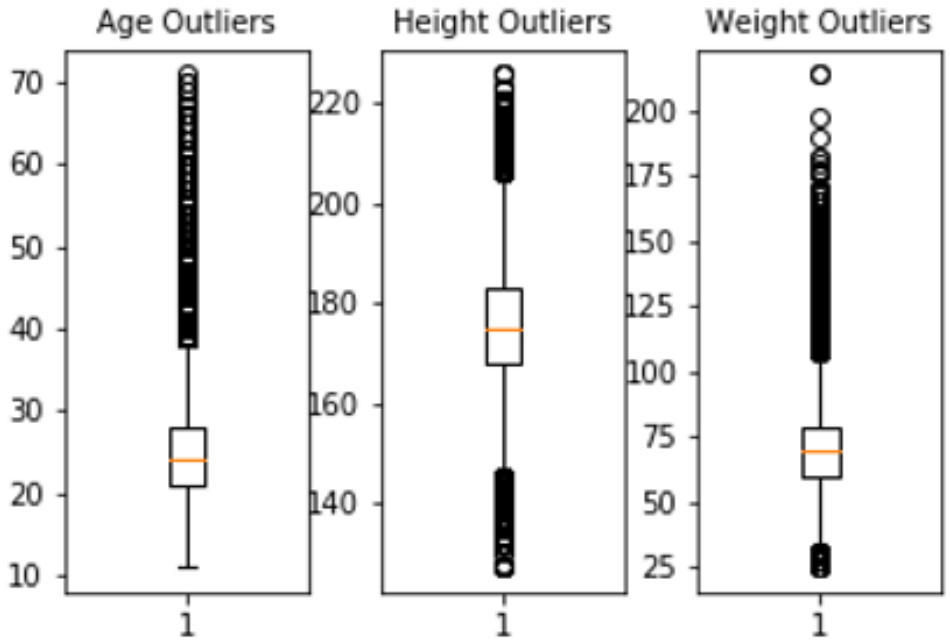
There is a relationship between age/weight/height that is unique to medalists

Data Analysis:

What the data tells us

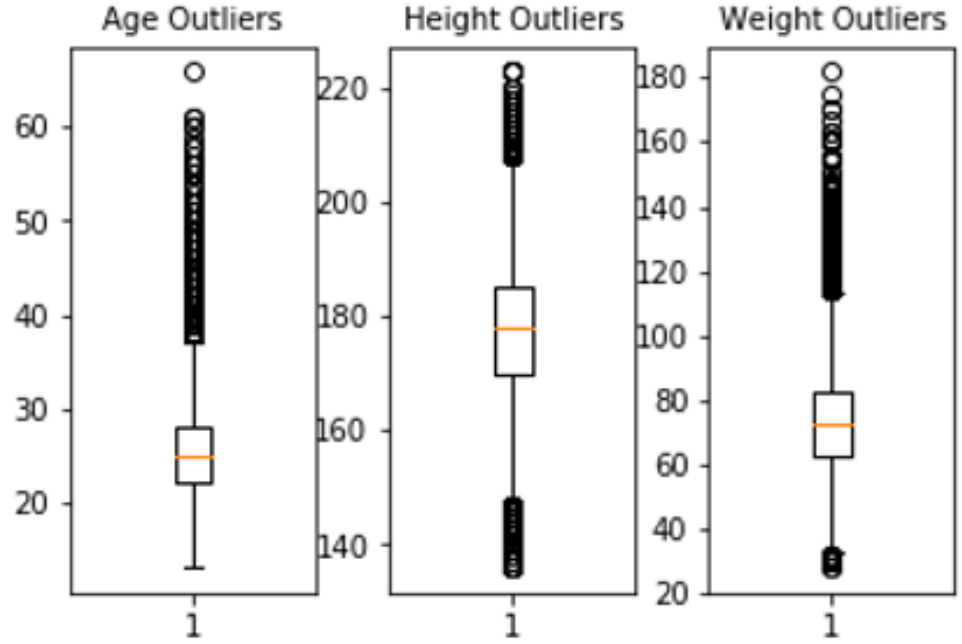
Variations and Outliers

Fig. 1: Age, Height, and Weight of Olympic Athletes



Median age: 24
Median Height: 175 cm
Median Weight: 70 kg

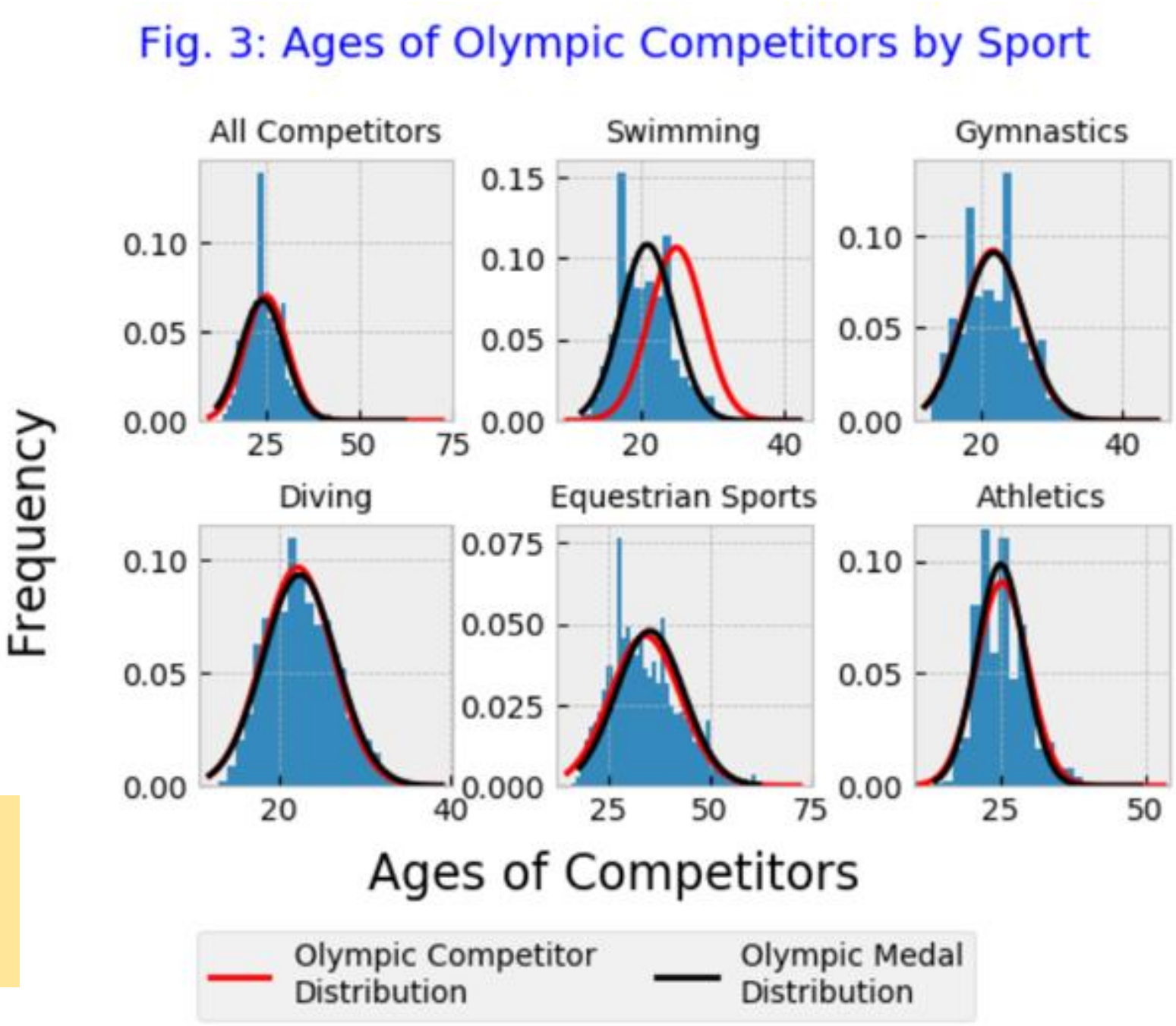
Fig. 2: Age, Height, and Weight of Olympic Medalists

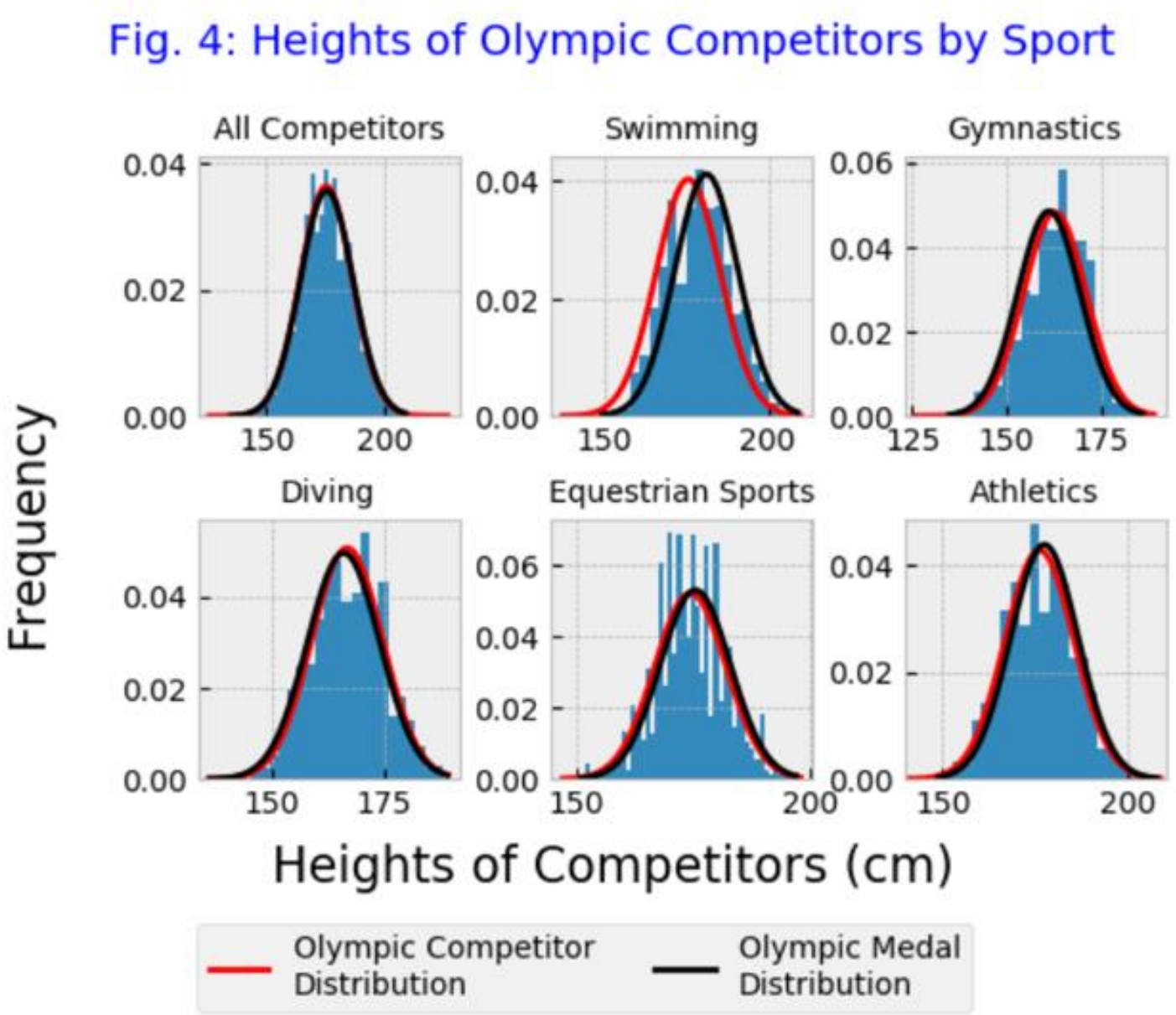


Median age: 25
Median Height: 178 cm
Median Weight: 73 kg

Ages of Competitors

Note difference in ages between medalists (black) and non-medalists, particularly in swimming, diving, and equestrian sports





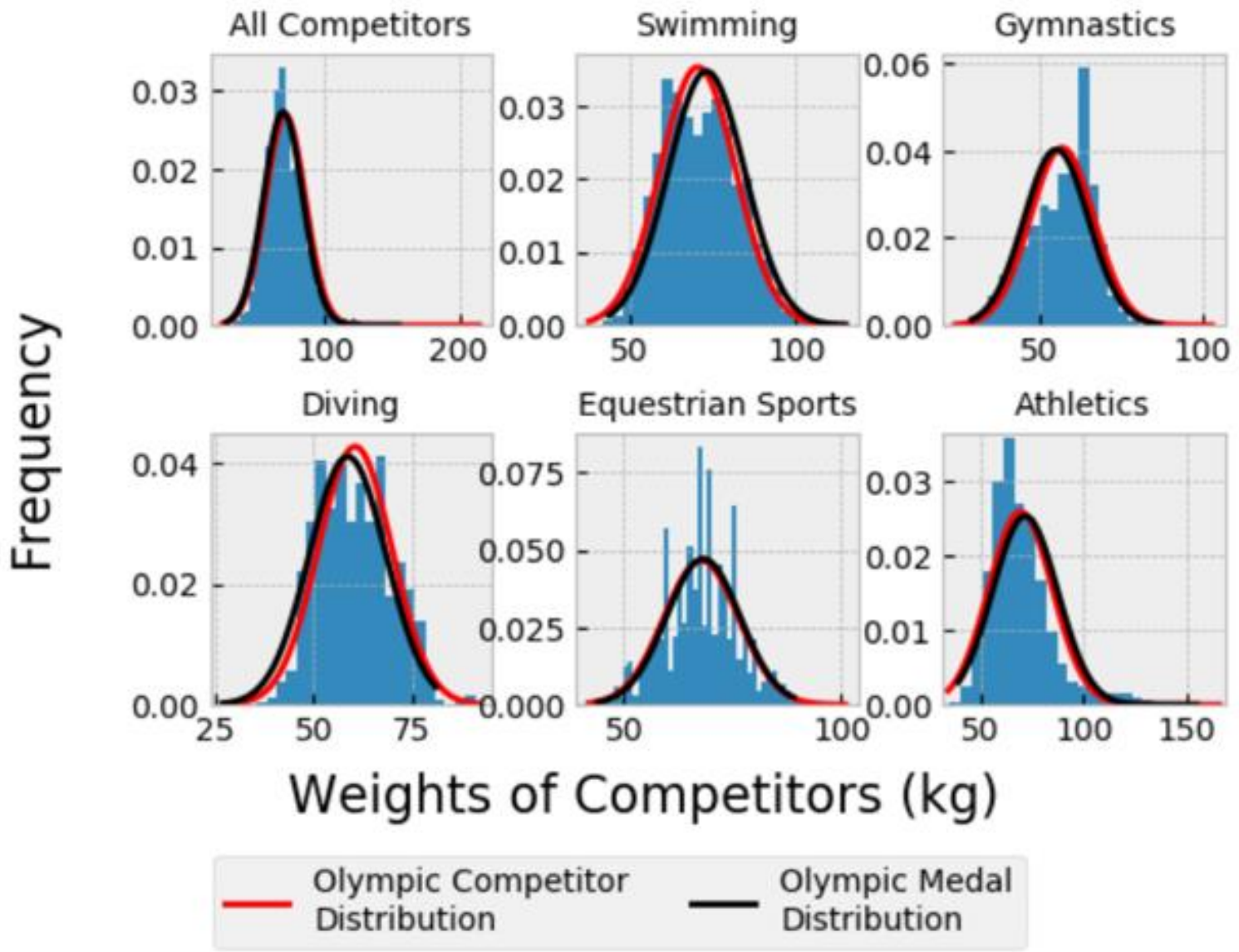
Heights of Olympic Competitors

Note differences in height between medalists (black) and non-medalists (red), particularly in swimming, gymnastics, and athletics

Weights of Olympic Competitors

Note weight difference between medalists (black) and non-medalists (red) in diving, swimming, and gymnastics

Fig. 5: Weights of Olympic Competitors by Sport



Weight vs. Height distributions

Fig. 6: Weight vs. Height for Olympic Competitors

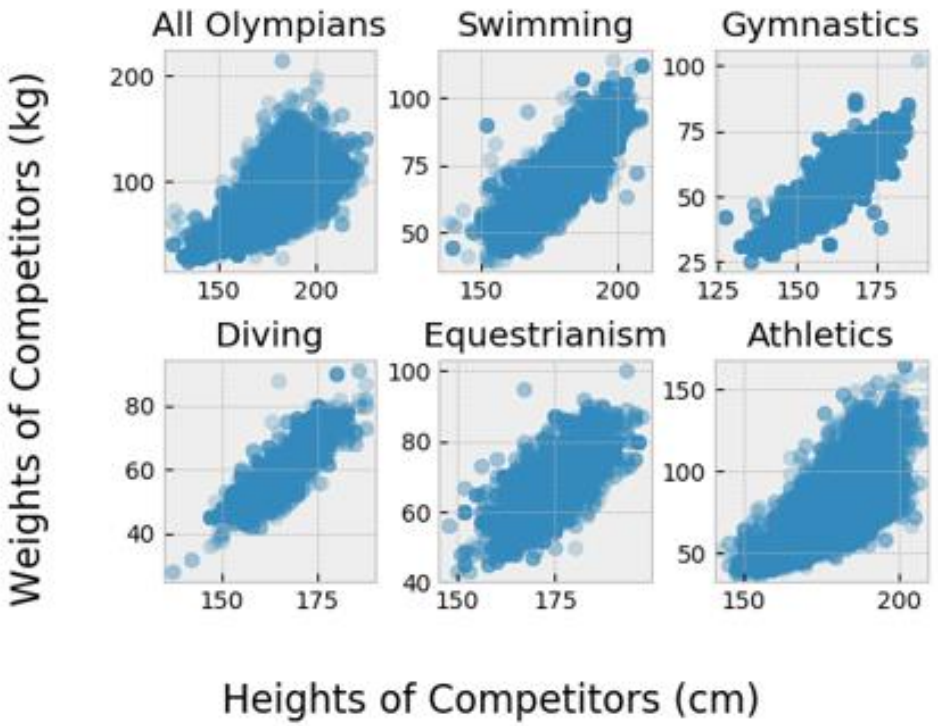
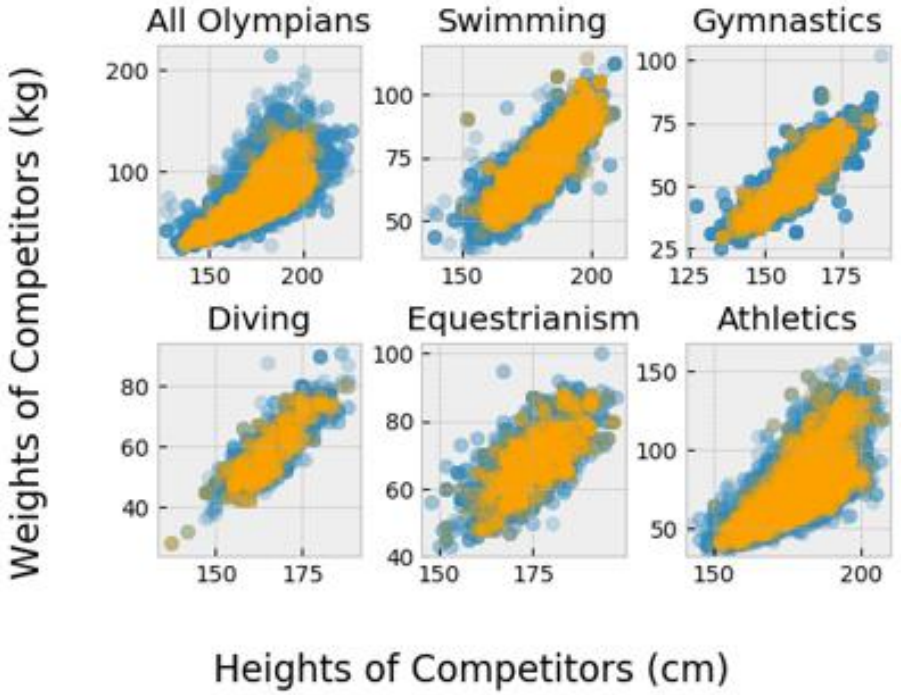


Fig. 7: Weight vs. Height for Olympic Competitors and Medalists



Although there are outliers, the medalist data (orange) clusters together, suggesting a trend in weight/height for medalists

Predictive Modeling

Models used

Random Forest Classifier

Support Vector Classifier

K-Nearest Neighbors Classifier

Results (full dataset)

Random Forest	SVC	KNN
Confusion Matrix: [[34090 463 421 460] [1674 186 191 133] [1185 132 588 172] [1492 106 200 184]]	Confusion Matrix: [[35268 12 144 10] [2012 72 87 13] [1651 17 401 8] [1828 7 131 16]]	Confusion Matrix: [[34903 140 272 119] [1809 174 149 52] [1436 86 502 53] [1643 61 165 113]]

Looking at the true positives (located along the diagonal from top left to bottom right), the Random Forest classifier performed the best at predicting the medal status of Olympic athletes in the full data set.

Results (Swimming dataset)

	Random Forest	SVC	KNN
Swimming Dataset	Confusion Matrix: [[3125 60 51 49] [112 10 16 6] [64 15 77 22] [89 12 31 17]]	Confusion Matrix: [[3198 0 87 0] [121 0 23 0] [70 0 108 0] [107 0 42 0]]	Confusion Matrix: [[3217 3 54 11] [130 0 13 1] [89 2 84 3] [112 1 31 5]]

The Random Forest classifier performed the best in the Swimming dataset while SVC performed the worst (only predicting non-medalists and gold medalists correctly)

Results (Gymnastics dataset)

	Random Forest	SVC	KNN
Gymnastics Dataset	Confusion Matrix: [[3396 7 19 7] [77 2 2 2] [59 2 10 3] [59 2 7 1]]	Confusion Matrix: [[3429 0 0 0] [83 0 0 0] [74 0 0 0] [69 0 0 0]]	Confusion Matrix: [[3425 0 3 1] [81 1 1 0] [70 1 3 0] [64 0 5 0]]

Random Forest performed best here, with SVC failing to accurately predict any medalists and KNN failing to correctly predict any Silver Medalists

Results (Diving dataset)

	Random Forest	SVC	KNN
Diving Dataset	Confusion Matrix: [[308 9 6 4] [16 3 2 3] [5 3 10 4] [13 5 8 4]]	Confusion Matrix: [[324 0 3 0] [21 0 3 0] [6 0 16 0] [22 0 8 0]]	Confusion Matrix: [[326 0 1 0] [23 0 1 0] [10 0 12 0] [25 0 5 0]]

Random Forest performed the best; SVC and KNN both failed to accurately predict Bronze and Silver medalists

Results (Equestrianism dataset)

	Random Forest	SVC	KNN
Equestrianism Dataset	Confusion Matrix: [[526 19 14 18] [31 1 1 1] [18 2 8 3] [29 4 8 0]]	Confusion Matrix: [[577 0 0 0] [34 0 0 0] [31 0 0 0] [41 0 0 0]]	Confusion Matrix: [[573 0 4 0] [32 1 1 0] [28 1 2 0] [38 0 3 0]]

Random Forest performed the best; however all classifiers failed to accurately predict Silver medalists. This is likely due to the high variance in athletes' age, weight, and height at the time of competition

Results (Athletics dataset)

	Random Forest	SVC	KNN
Athletics Dataset	Confusion Matrix: [[5565 66 50 69] [199 9 18 7] [190 6 51 17] [183 16 24 5]]	Confusion Matrix: [[5750 0 0 0] [233 0 0 0] [264 0 0 0] [228 0 0 0]]	Confusion Matrix: [[5723 7 18 2] [228 0 5 0] [247 0 14 3] [221 1 6 0]]

Looking at the true positives (located along the diagonal), the Random Forest classifier performed the best at predicting the medal status of Olympic athletes while SVC failed to accurately predict any medalists.

Conclusion

- There is a difference between medalists and non-medalists
- Using country, age, height, weight, sport, and sex, it is possible to predict Olympic medalists

Recommendations:

- Improve prediction by:
  - Adjusting for imbalanced data
  - Add more biometrics
  - Restrict country fields to country of interest
  - Perform analysis again on two categories: medalists vs. non-medalists