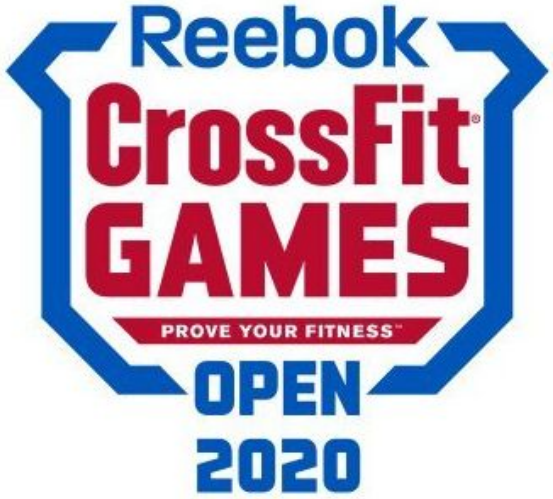
2020 Crossfit Open Analysis



Scott Graham
24th June, 2022

Data Summary

The Crossfit Open is the largest single sporting competition held worldwide. The goal, to find the fittest athletes on Earth.

Using the athlete metrics and workout results for 230,000+ for the five workouts we hope to be able to determine some key features of top performing athletes

Outline

- Purpose of Research and Hypothesis
- Data Understanding
- Data Cleaning
- Regression Analysis
- Further Analysis
- Conclusion

Purpose of Research and Hypothesis

To determine if athlete metrics, including age, height and weight are key determinants of better results

Difference between male and female body metrics

Comparison of the very top athletes vs the general population and difference in body metrics

Data Understanding

Two datasets were used and imported from Kaggle:

1. 2020 Athletes (393,000+ rows, 19 columns)
   Provided all the general information about the athlete including:
   a. Name
   b. Gender
   c. Age
   d. Weight
   e. Height
   f. Division they compete in
   g. Overall rank and score
2. 2020 Workouts (1965000+ rows, 13 columns)
   Provided information regarding:
   a. Gym they did the workouts in
   b. Their judge
   c. if the workout was scaled or as prescribed
   d. Results and ranking of each workout

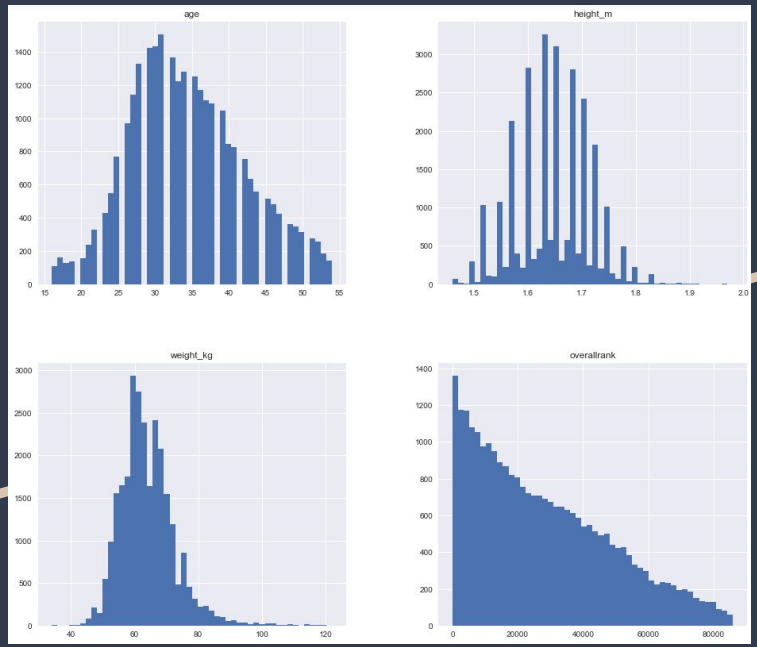
Data Cleaning

Removal of columns with no relevance to goal and duplicate rows

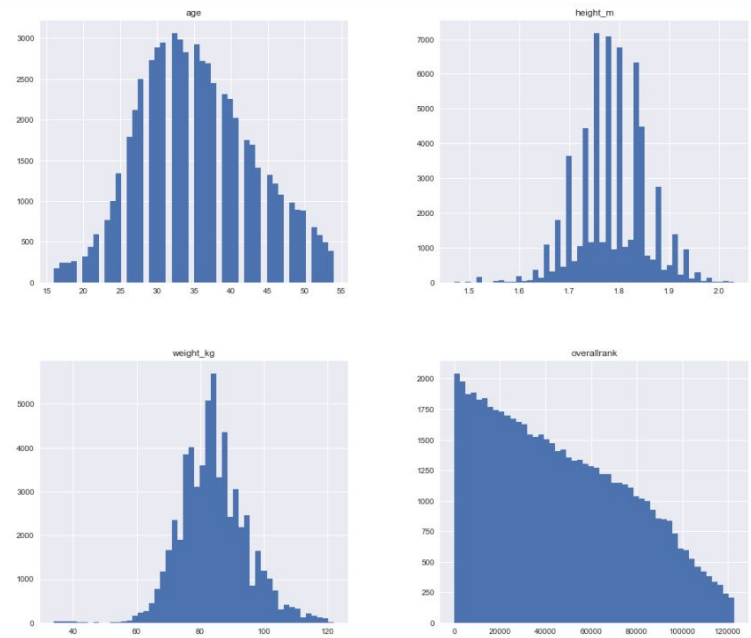
I removed all athletes that did not have both height and weights accurately entered (eg. 1kg body weight)

Seperating male and female athletes

Female



Male

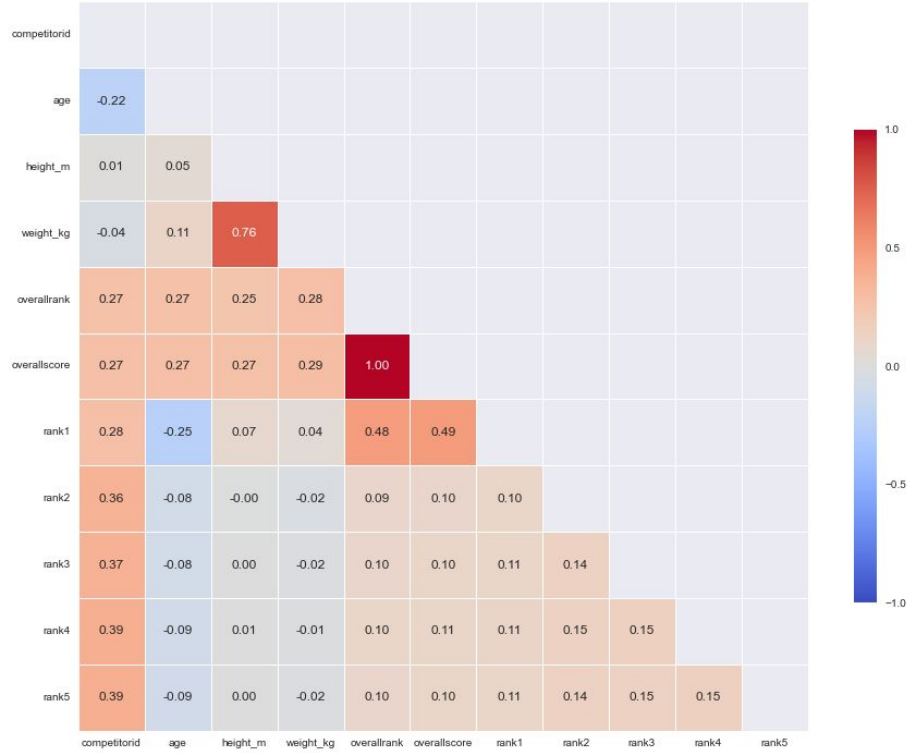


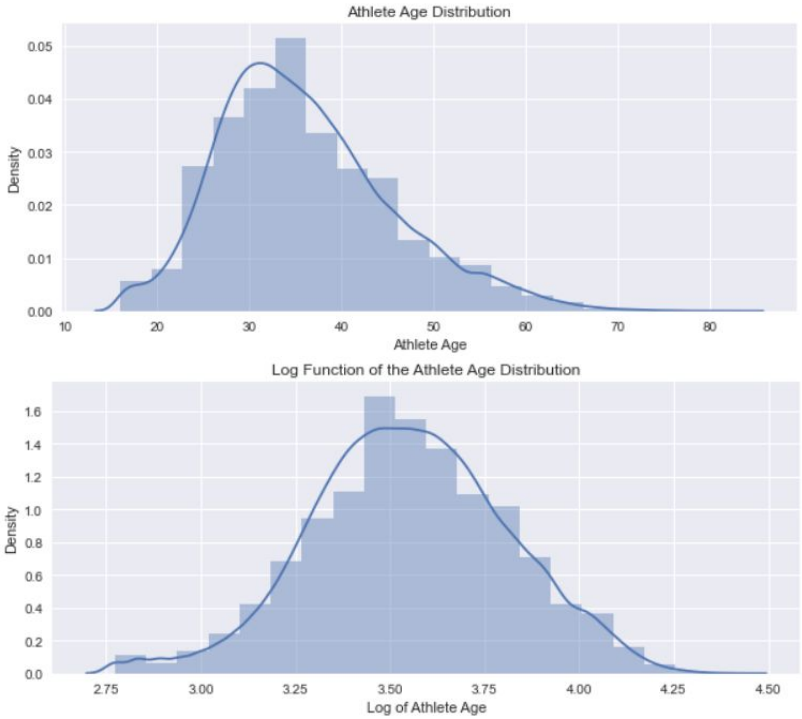
Data Cleaning

Removal of outliers using Z-Score

Normalise age data for easier use

Compare data against each other using a correlation matrix





Regression Analysis

Train-Test-Split done with 70/30 split

Train and Test Mean Square Error converged:

- Train Mean Squarred Error: 0.8786019
- Test Mean Squarred Error: 0.8900844
- Meaning that our data was producing similar results in the training model to the test model

However our R-Squared value was only 0.155, meaning our data had only a 15.5% chance of predicting the athletes overall rank from their age, weight, height and gender

Further Analysis

Split data into the 4 quartiles based on overall ranking of athletes

Take the mean values of age, height and weight of each quartiles and compare to Games athletes

```
Female Top Quartile
 age              32.39
height_m           1.64
weight_kg         62.95
overallrank    4,837.77
dtype: float64

Female Second Quartile
 age              34.17
height_m           1.65
weight_kg         63.09
overallrank   16,763.24
dtype: float64

Female Third Quartile
 age              35.43
height_m           1.65
weight_kg         63.76
overallrank   32,737.57
dtype: float64

Female Bottom Quartile
 age              35.84
height_m           1.65
weight_kg         66.27
overallrank   57,667.13
dtype: float64
```

```
Games Female Athletes
 age              29.35
height_m           1.64
weight_kg         64.21
overallrank    2,933.21
dtype: float64
```

```
Male Top Quartile
 age              31.86
height_m           1.78
weight_kg         83.64
overallrank    9,554.44
dtype: float64

Male Second Quartile
 age              34.63
height_m           1.78
weight_kg         83.02
overallrank   30,798.78
dtype: float64

Male Third Quartile
 age              36.74
height_m           1.79
weight_kg         84.20
overallrank   56,091.27
dtype: float64

Male Bottom Quartile
 age              38.25
height_m           1.79
weight_kg         86.17
overallrank   90,082.63
dtype: float64
```

```
Games Male Athletes
 age              28.99
height_m           1.77
weight_kg         86.43
overallrank    4,010.39
dtype: float64
```

Conclusions

Unable to produce a predictive model based on athlete metrics, this is obviously due to many other parameters around fitness performance

The performance of athletes could be seen to decrease as they aged. With Games athletes being the youngest on average

Future Work

Better use of predictive modelling with expansion on workout information would be a great way to advance this analysis

Thank You!

Email: scottgraham14@gmail.com
GitHub: @scottgraham1
LinkedIn: linkedin.com/in/scott-graham/