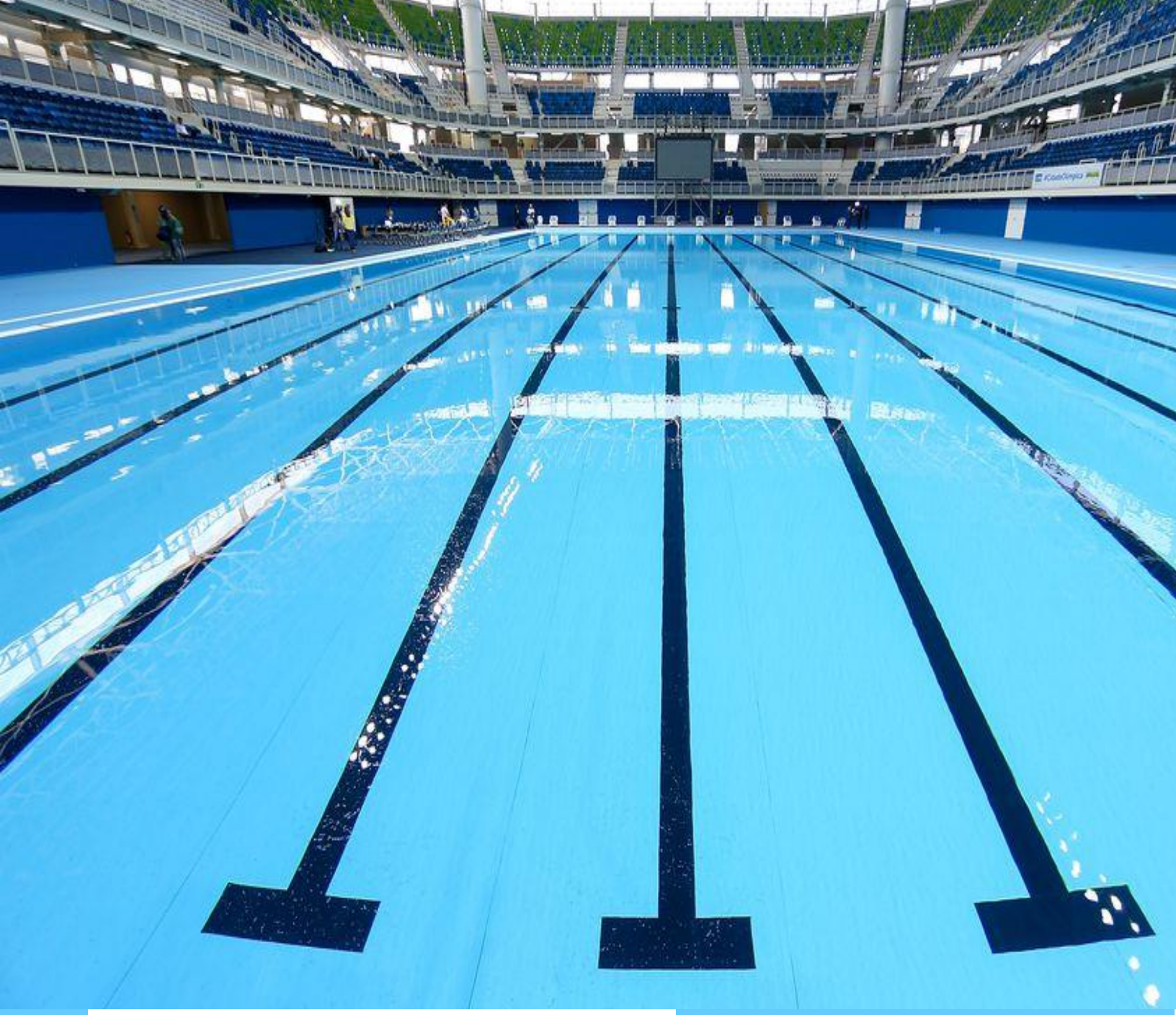

# Height and Weight in predicting the probability of winning an Olympic medal in swimming

SEPTEMBER 2, **2020**

**Amr Fahmy**

# Table of Contents

# 1. Introduction

## 1.1 Background

Competitive swimming is an Olympic sport where every minute detail counts. From the moment you start the season until you finish the race. Many variables count towards being a professional swimmer, some of those variables are: Genetics (bone structure, height, muscle composition), Training (duration & frequency of training, intensity & variety of trainings), and Nutrition. Some of these factors can't be controlled such as genetics, however some other factors can be controlled to improve swimmers' competitive edge and reduce their times by even factors of a second which can be the difference between a medal and no medal.

## 1.2 Problem

This report's objective is to determine the extent of Age, Height and Weight on an Olympic and to create a machine learning model that can predict the probability of winning an Olympic Medal in swimming based on genetic inputs such as the height and a controllable input such as the weight of a swimmer for each gender.

## 1.3 Interest

Target market for this model is competitive swimmers, swimming coaches and parents of young swimmers. During preparations of an important competition, it could potentially be used to determine the swimmer's physical condition compared to a past championship and determine whether he's more likely to be in a better physical state.

# 2. Data analysis

## 2.1 Data Source

The dataset related to the project was obtained from Kaggle and you can through this link. It contains 271116 rows and 15 columns of values and attributes of all the Olympic games from Athens 1896 to Rio 2016. For more information, click the link to view the metadata of the file.

## 2.2 Data cleaning

The dataset collected huge number of information regarding each Olympic athlete such as, their name, age, sex, height, weight, nationality, sport, event and whether they won a medal. Since we're only interested in Swimming for this model, the data was filtered to only contain rows that correspond to swimming.

Next step was to only select the columns that would contribute to the dataset output "Medal". The attributes of interested selected are Sex, Age, Height and Weight and the output column Medal.

By examining the output Medal. It contained a large number of NaN values. So by further examining it using the value_counts() method, it was concluded that values denoted as NaN referred having no medals won in that event, so all the NaN values were replaced in the column by "None" of type object

```
df["Medal"].value_counts()
```

```
]: Gold      13372
   Bronze    13295
   Silver    13116
   Name: Medal, dtype: int64
```

The Data was then split to 2 different datasets male and female swimmers through Sex attribute to perform separate analysis operations on both.

By examining the rest of the attributes (Height, Age, and Weight). It was observed that were missing values in each of them. So, all the NaN values were replaced by their average of each column. For example, Female Height NaN values were replaced by the average female Height of all the female swimmers

# 3. Methodology

## 3.1 Correlations

Now that the data is formatted and cleaned properly, some statistics and correlations will be explored to test the hypothesis that a relation exists between the input and output parameters

## 3.2 Machine Learning Algorithm

A Polynomial regression of degree 2 model will be implemented on the decided input (W/H) vs medal won to predict the likelihood of a medal won given an input (W/H) and a male or female.

The Medal column will be changed to represent 1 for a medal won or 0 for no medal won and the probability output of winning a medal will be between 0 and 1

The samples were split into training and testing samples with a testing size of 25%. Then the model training X set was transformed and fit using a degree 2 polynomial regression. After transformation, the model can be predicted using a simple linear regression model with x = “W/H”^2

# 4. Results

## 4.1 Correlations Results

Seaborn sns was imported to plot the data and gain some insights on our inputs vs outputs. The graphs below show sns pairplot for Male and Female swimmers

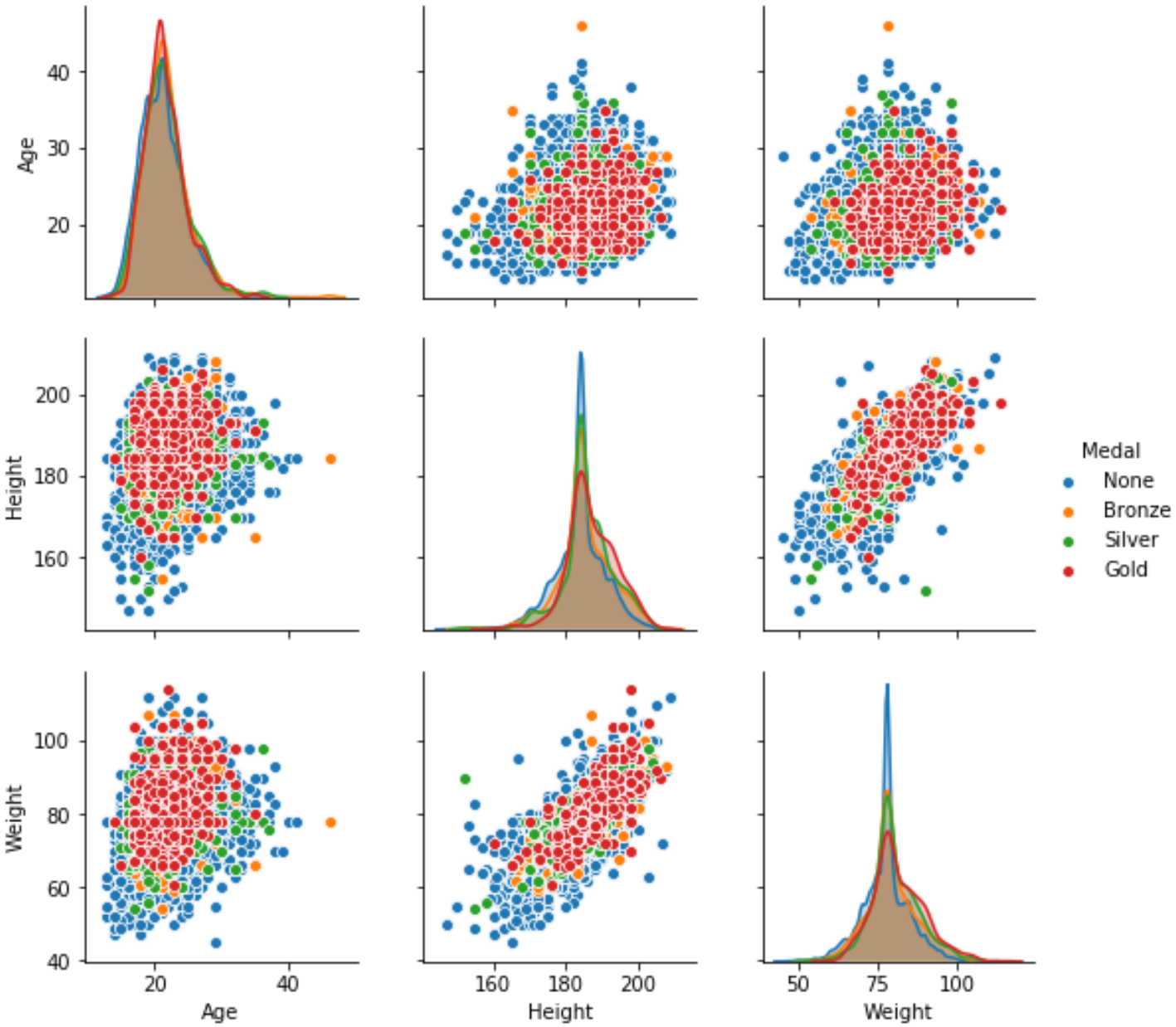


*Figure 1: Male Swimmers Pairplot*

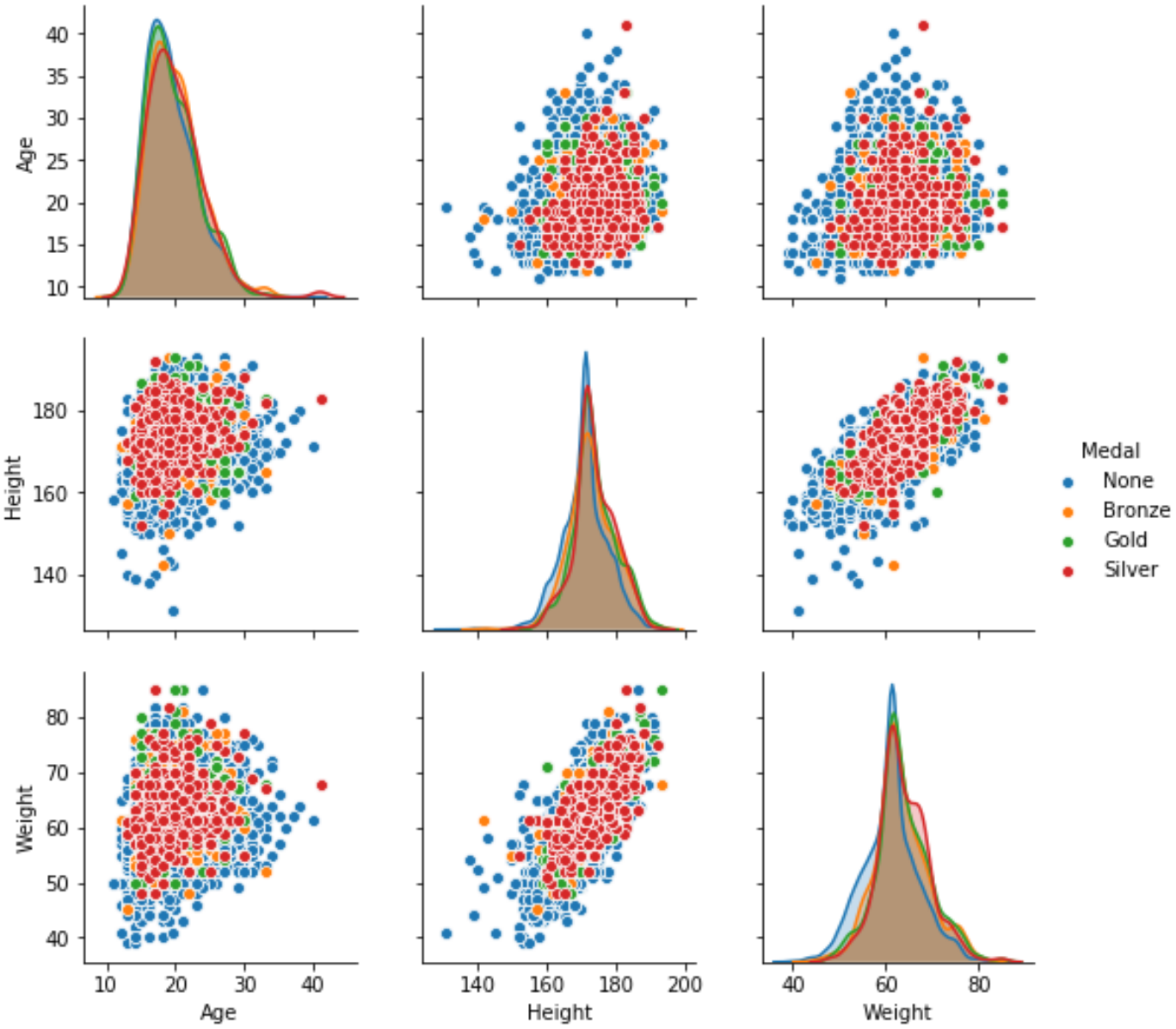


*Figure 2: Female Swimmers Pairplot*

It is observed that the scatter plots are less dispersed for medal winners vs for swimmers who didn't win medals, that's the first indication that a correlation between the inputs and output exists

By performing further analysis and plotting each input vs the output we obtain the following sns box plot graphs.

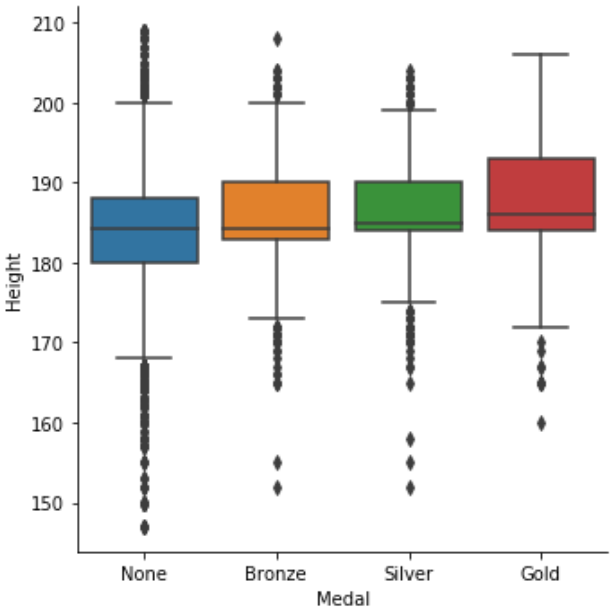


*Figure 3: Medal Vs Height Males*

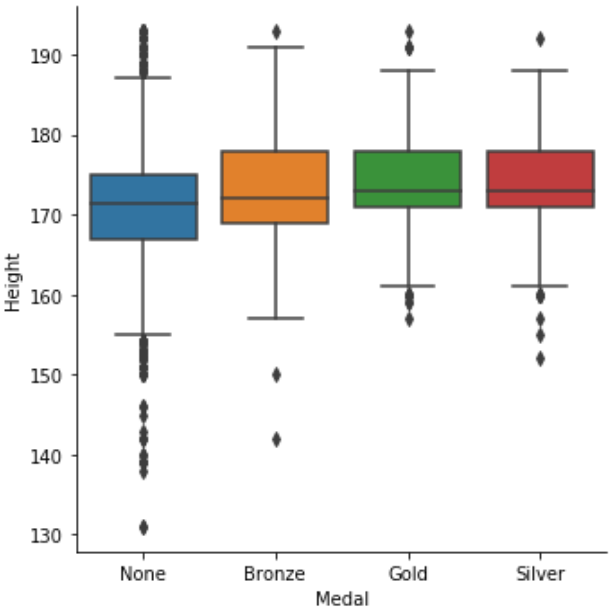


*Figure 4: Medal Vs Height Females*

As shown from the graphs above, height is directly proportional with medals won in the Olympics, we can examine that by looking at the mean of the box plots and observing its change for every type of medal won. As your height increases your probability of winning a gold medal increases. Optimal mean height appears to be 187 cm for males and 173 cm for females.

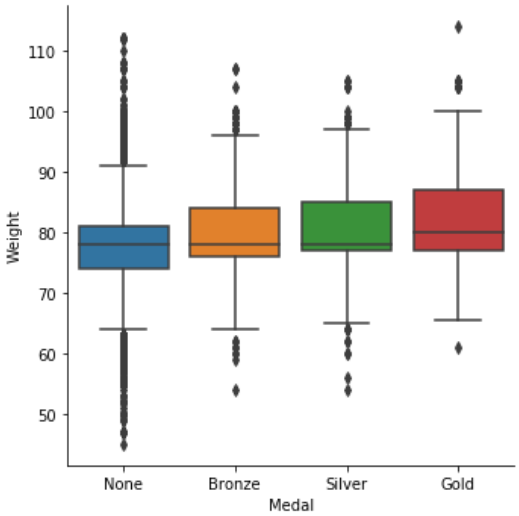


*Figure 5: Weight vs Medals Males*

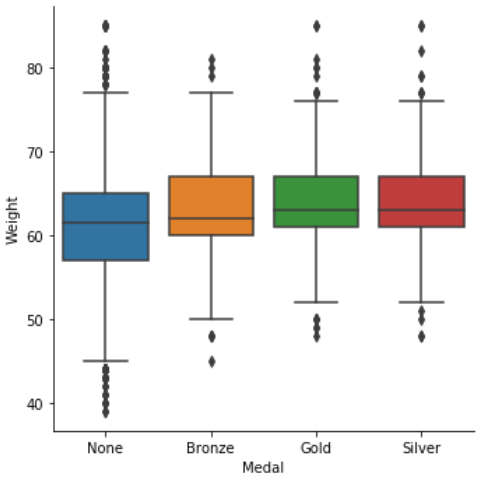


*Figure 6: Weight vs Medals Females*

Also as shown from the graphs above, Weight is directly proportional with medals won in the Olympics, we can examine that by looking at the mean of the box plots and observing its change for every type of medal won. As your Weight increases your probability of winning a gold medal increases up to a certain extent. Optimal mean Weight appears to be 80 Kg for males and 63 Kg for females.

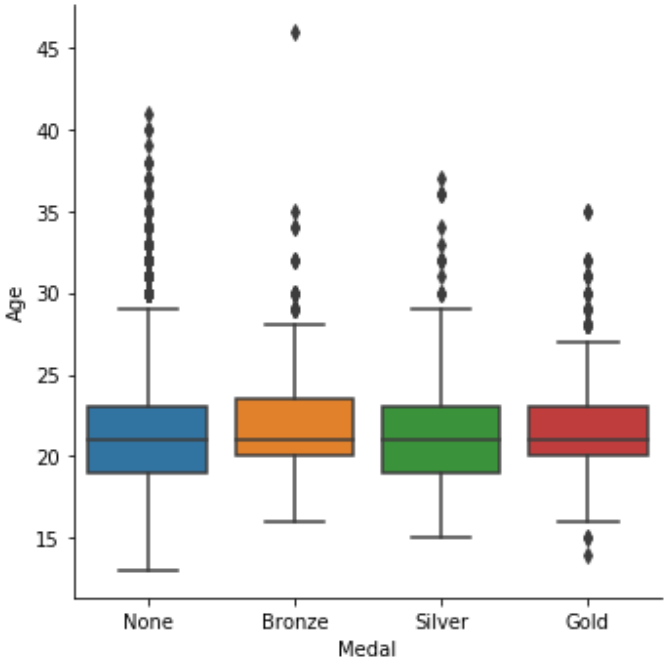


*Figure 7: Age Vs Medal Males*

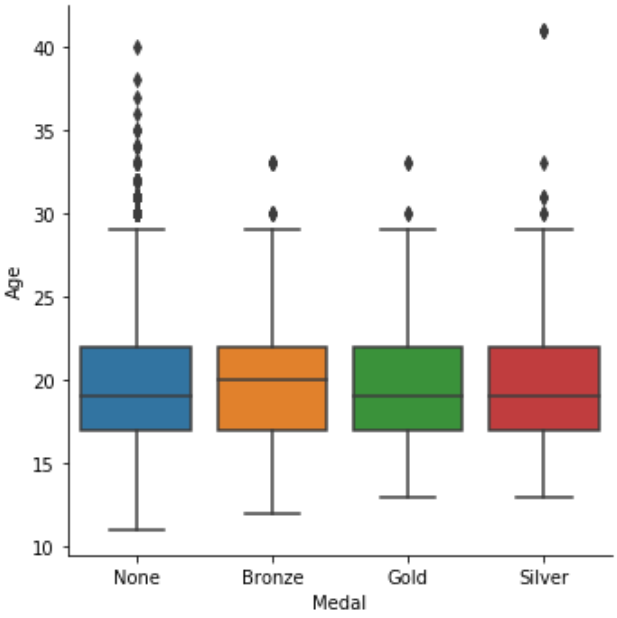


*Figure 8: Age vs Medal Females*

The above box plot shows the extent of age on swimmers' performance. It appears that most swimmers regardless of their skill level reach the peak of their career at a certain age. That peak for males is 22 years old and for females its 19 years old. A relation between age and medal won is not observed so that means it provides a weak input and can be omitted later in the analysis

Last step is group Height and Weight into one column as a ratio of Weight/Height and see its correlation with medals won. The sns box plots are represented in the graphs below

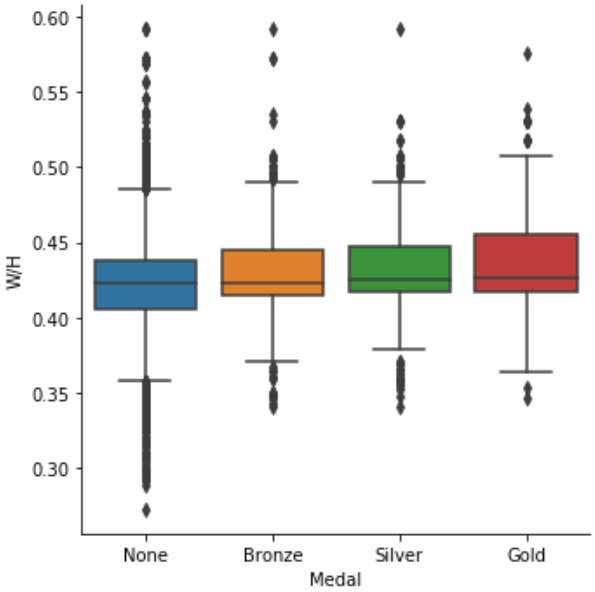


*Figure 10: W/H vs Medal Males*

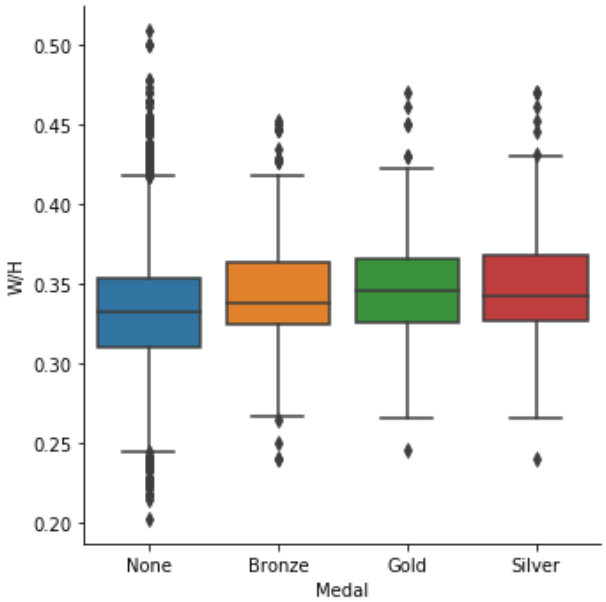


*Figure 9: W/H vs Medal Females*

From the graphs above we can still see the direct correlation between W/H vs medals for both males and females. A higher ratio is not necessarily good because it depends on factors such as muscle structure and body fat, however assuming all else constant a high swimmers muscle mass ratio vs height is very desirable

## 4.2 Machine Learning Algorithm Results and Evaluation

Through the transformation of the polynomial model to a simple linear model, yhat was predicted by fitting the training data through x = “W/H”^2 and y= “Medal”. The results can be expressed on as a swimmer either won a medal or no model in binary “0” or “1”. The correlation is represented in a box plot as following

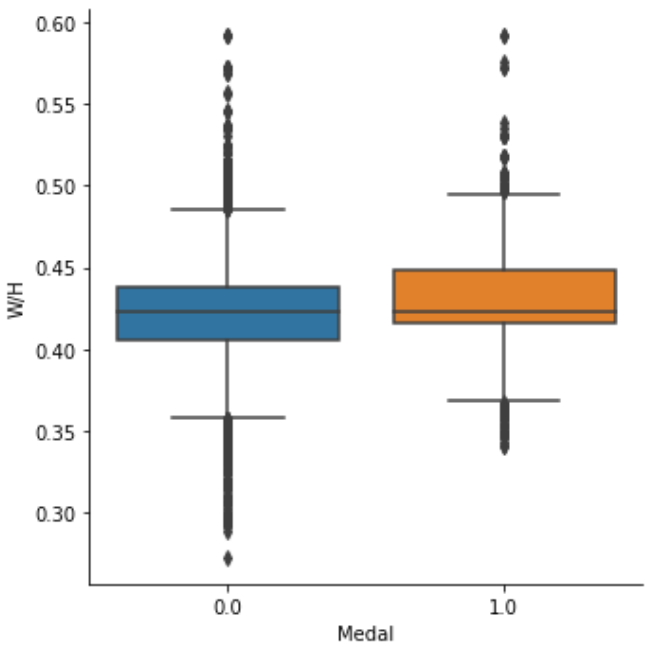


*Figure 12: "W/H" vs "Medal" Male*

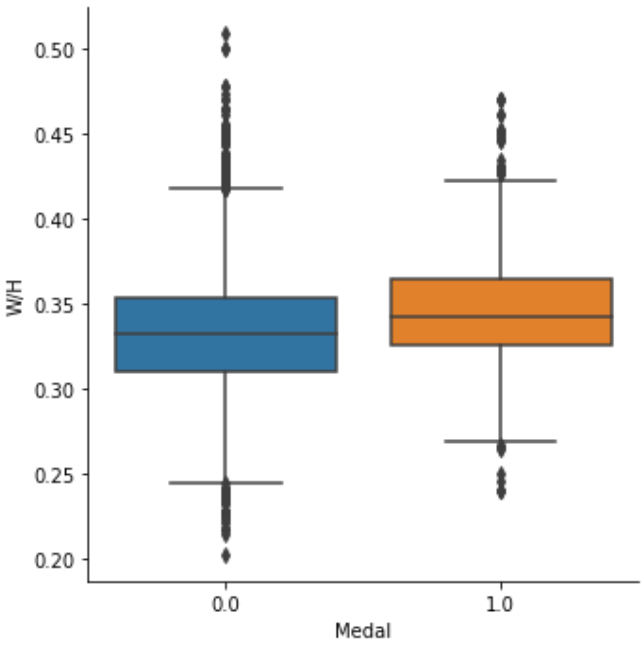


*Figure 11: "W/H" vs "Medal" Female*

### 4.2.1 Male Dataset

The coefficients and y-intercepts of the polynomial function for male swimmers have been calculated as:

```
Coefficients:  [ 0.         -1.24282176  2.84911369]
Intercept:  0.1382658802127123
```

Which we can use to plug into the equation of the polynomial model of degree 2 and find the probability of winning a medal in the Olympics for Male swimming.

The testing set was used to evaluate the accuracy of the model and the mean absolute error and the residual sum of squares were calculated based on the predicted probability vs actual value. Both are shown below

```
Mean absolute error: 0.22
Residual sum of squares (MSE): 0.11
```

### 4.2.2 Female Dataset

The coefficients and y-intercepts of the polynomial function for Female swimmers have been calculated as:

```
Coefficients:  [ 0.          4.92711844 -5.33186148]
Intercept:  -0.9071710384507015
```

Which we can use to plug into the equation of the polynomial model of degree 2 and find the probability of winning a medal in the Olympics for Male swimming.

The testing set was used to evaluate the accuracy of the model and the mean absolute error and the residual sum of squares were calculated based on the predicted probability vs actual value. Both are shown below

```
Mean absolute error: 0.24
Residual sum of squares (MSE): 0.13
```

# 5. Discussion

As shown in the results section (4.2) The coefficients and intercepts of the polynomial functions were determined and can be used to determine to manually determine the probability of winning a medal given a set of inputs by plugging in the coefficients and intercepts into the following equation

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2$$

Where: $\theta_0$ = y-intercept of the curve of best fit.
$\{\theta_1, \theta_2\}$ = the coefficients of $x$ and $x^2$ respectively
$\hat{y}$ = estimated probability of winning a medal through the model

# 6. Conclusion

In This study, the relationship between Weight, Height and Probability of winning a medal was observed. A regression model was also created and evaluated to estimate the probability of winning a medal in the Olympics given height and weight. It is observed that the mean of the prediction values through the test sample is hovering around 14% for males and 18% for females. This is due to having significantly more samples of swimmers having won no medals than sample of swimmers that have. The reason for keeping it that way is because I have tried to keep the sample ratio of medals won vs no medals won at 3/16 since in each event in the Olympics only 3 people out of 16 in the semifinals and finals win a medal.

# 7. Future directions

For future reference, more in-depth analysis can be done with more inputs such as body fat ratio, multi-Olympic medalists, nationality. And find more relationship outputs such as Olympic records,