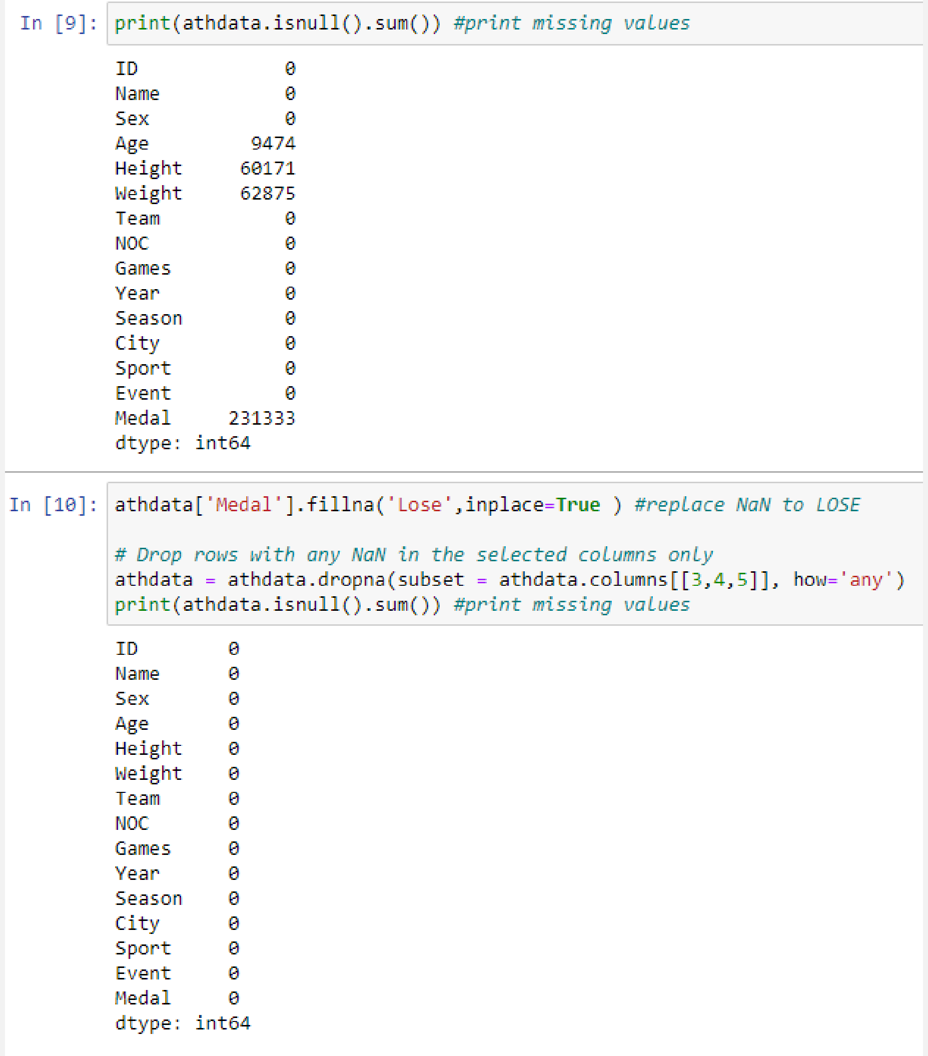
PROBLEM STATEMENT

Predicting Olympic medal winners using weight and height

DATA CLEANING

- Dropping NaN values in height, weight, age since the selected parameters need to be available
- Filling out Nan values for no medal winners

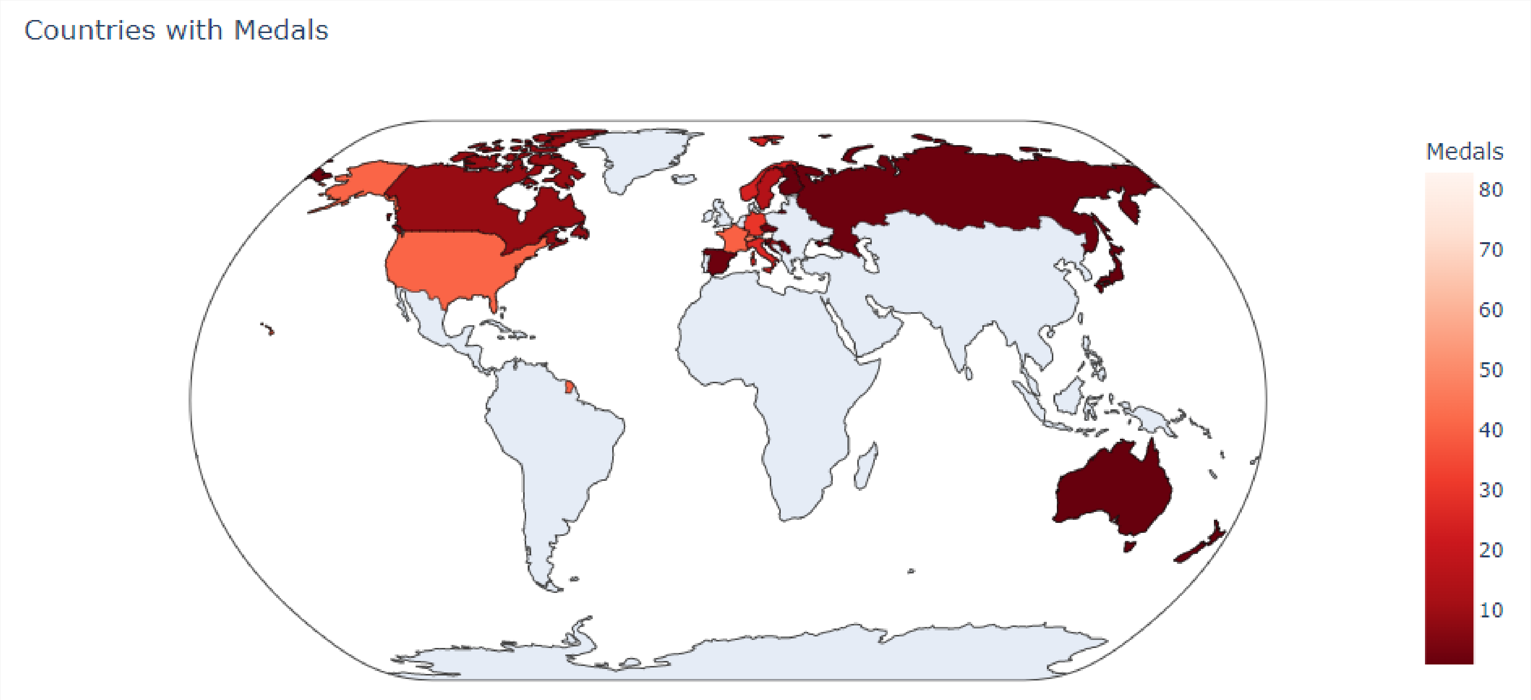


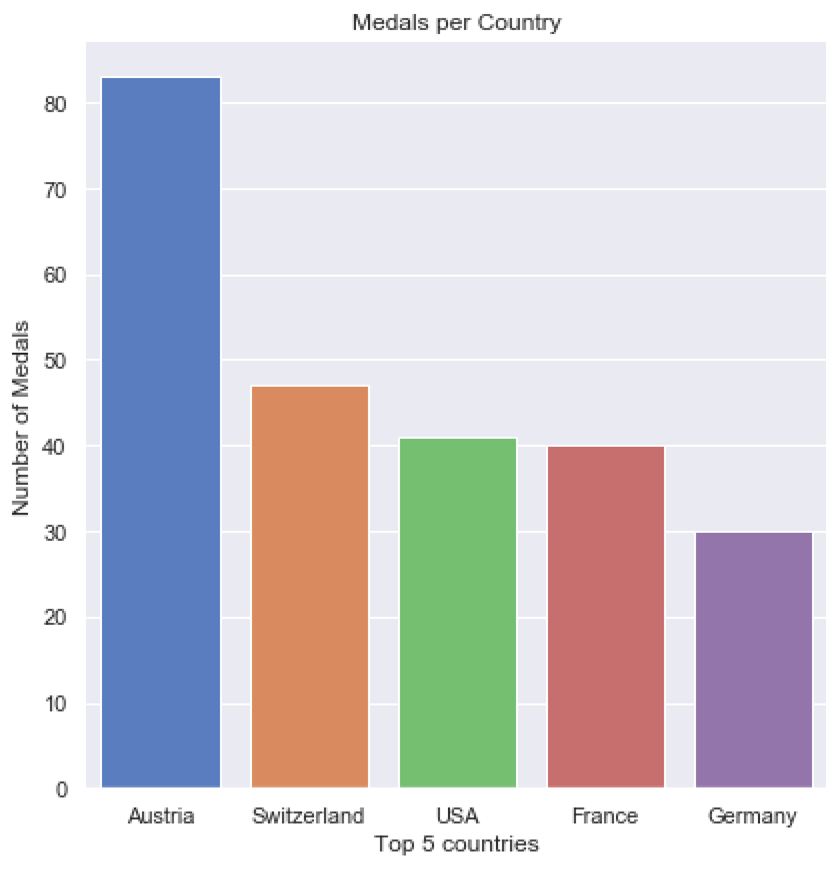
```
In [9]: print(athdata.isnull().sum()) #print missing values

        ID              0
        Name            0
        Sex             0
        Age          9474
        Height      60171
        Weight      62875
        Team            0
        NOC             0
        Games           0
        Year            0
        Season          0
        City            0
        Sport           0
        Event           0
        Medal      231333
        dtype: int64
```

```
In [10]: athdata['Medal'].fillna('Lose',inplace=True ) #replace NaN to LOSE

         # Drop rows with any NaN in the selected columns only
         athdata = athdata.dropna(subset = athdata.columns[[3,4,5]], how='any')
         print(athdata.isnull().sum()) #print missing values

         ID        0
         Name      0
         Sex       0
         Age       0
         Height    0
         Weight    0
         Team      0
         NOC       0
         Games     0
         Year      0
         Season    0
         City      0
         Sport     0
         Event     0
         Medal     0
         dtype: int64
```

MEDAL WINNERS

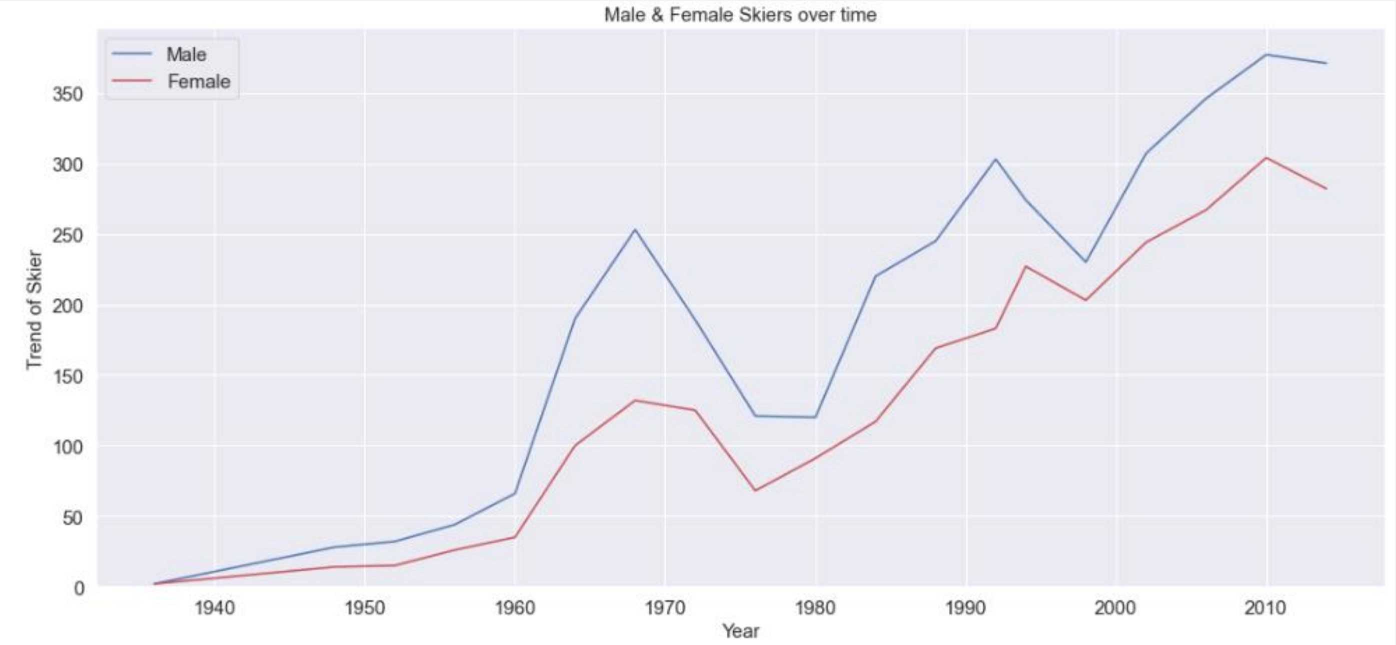
- By country, over the last 120 years from Y1896 to Y2016

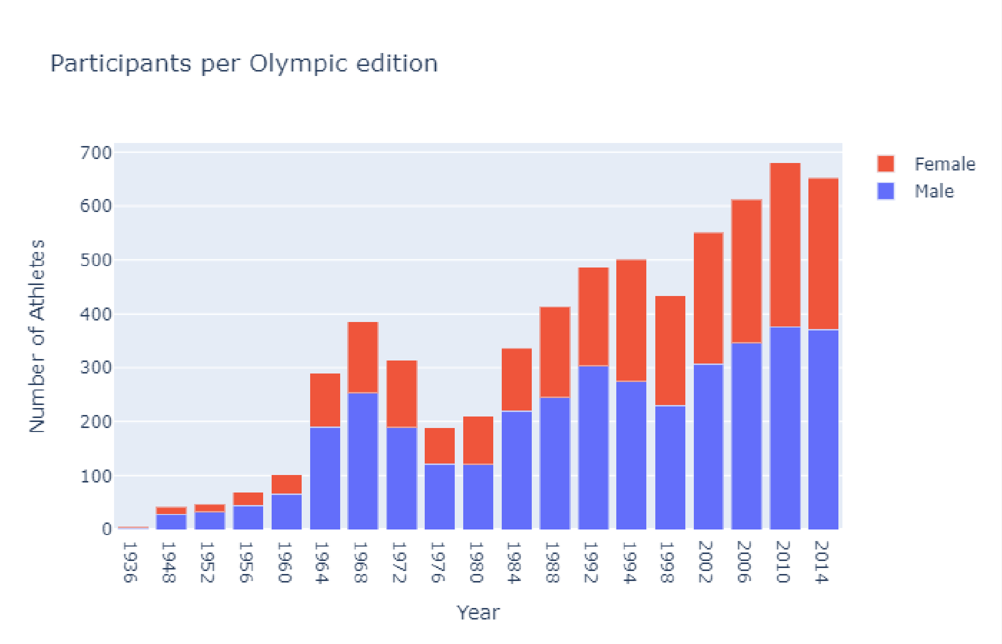




PARTICIPATION BY SEX

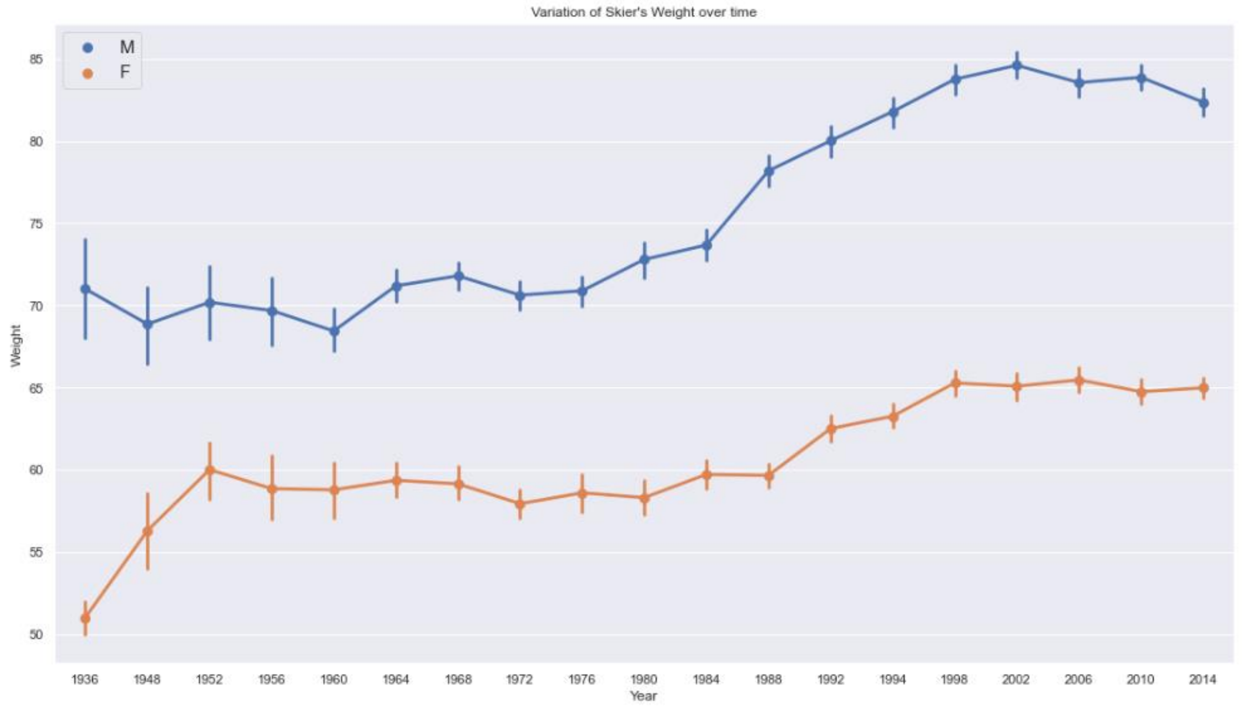
- Increasing trend over the years will give a much higher or better accuracy for finding the medal winners with chosen parameters

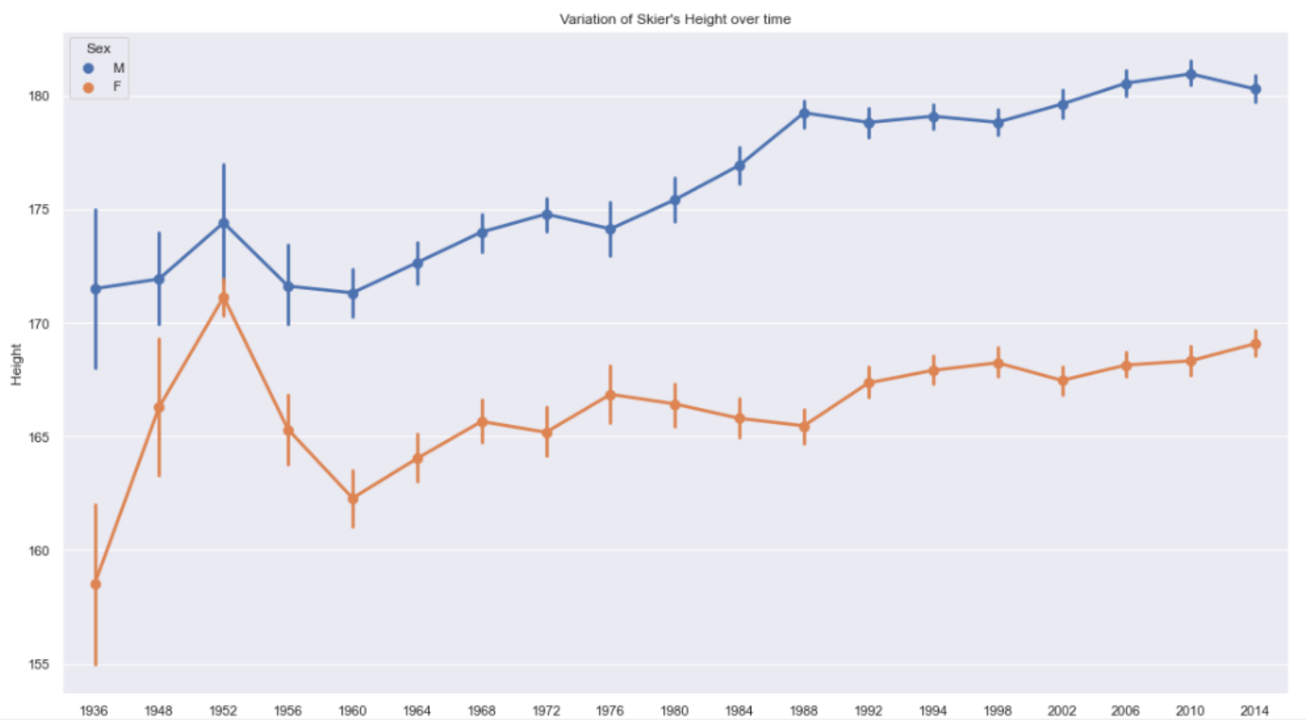




WEIGHT AND HEIGHT VARIATION

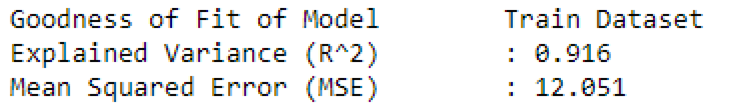
- There are many kinds of sports that the bigger they are, the higher chances of winning. Therefore, medal winners for sports are narrowed to alpine skiing, a sport that requires agility and speed.
- Increasing weight trend has been shown up till 21st century but from year 1952 to 1960 there seems to be a notion that any heavier is not good. But that was not the case due to athletes being heavier up to year 2014. In earlier years up till year 1956, weight and height of athletes were shown to be quite varied but has reduced over the years.



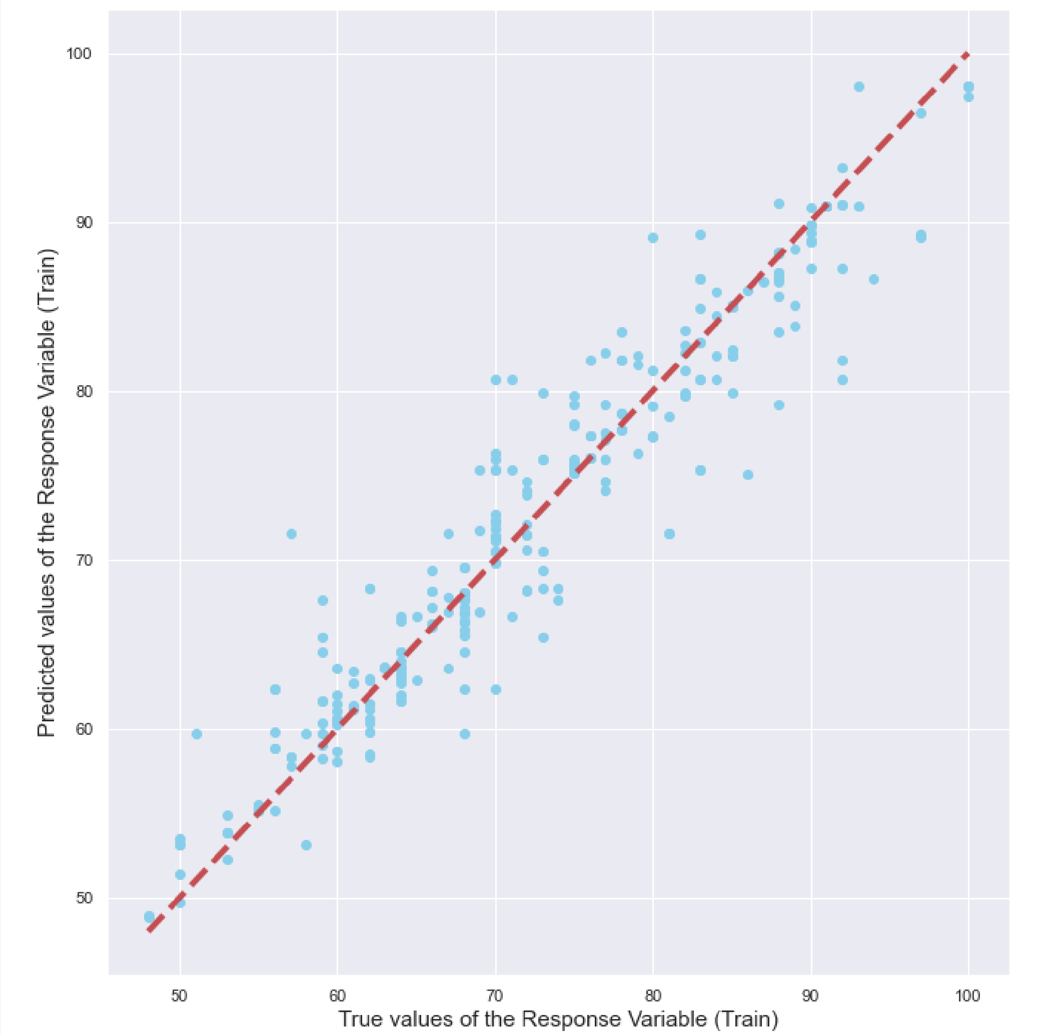


REGRESSION OF THE PARAMETERS

- The trained data resulted in a positive skew of 0.916 showing that bigger athletes in a speedy sport do make bigger impact for medal winners



```
Goodness of Fit of Model         Train Dataset
Explained Variance (R^2)         : 0.916
Mean Squared Error (MSE)         : 12.051
```


Predicted values of the Response Variable (Train)
True values of the Response Variable (Train)