**Ultra Marathon Data set – Python**

***Step by Step guide of the project***

1. Imported the pandas and seaborn libraries
2. Imported the data into the data frame ‘df’

DATA CLEANING

1. Narrow down the dataset to Races in USA in the year 2020 which are at least 50km or 50mi
   1. Step 1 = Filter races with at least 50km or 50mi using isin, indexing and slicing
   2. Step 2 = filter the data to races held in the year 2020 using boolean condition and indexing
   3. Step 3 Filter the data to races held in USA using indexing, .str(), .split() and a boolean condition.

This is because the name of the hosting country is included in brackets at the end of the event name.

Eg. - Taipei 48hr Ultra Marathon - 24h Split (TPE) {(TPE) – Taiwan}

* 1. Step 4 = combine all the filters and assign the result into a new data frame 'df2'

1. Cleaning the Event name by removing '(USA)' from the end.
2. Calculate Athlete Age
3. Remove 'h' from Athlete performance
4. Deleting unwanted columns - Athlete club, Athlete year of birth, Athlete age category
5. Removing Null values
6. Checking for duplicate values
7. Reset Index
8. Change types of columns as necessary
9. Change Column Names for convivence
10. Reorder Columns according to needs and save it to new data frame 'df3'

DATA VIZUALISATION using Seaborn

1. Plotting race lengths on a histogram to see which race length was more frequent
2. Plotting race lengths on a histogram categorized by gender
3. Distribution of Average Speed of 50-mile races
4. A violin plot distribution of 50 miles vs 50 kms average race time categorized by gender.
5. Plotting average speed against age and categorizing it by gender on a scatter plot with a regression line
6. Plotting average speed against age and categorizing it by gender on a scatter plot with a regression line for 50-miles and 50kms separately
7. Answering some questions based of the data-
   1. Difference in average speed for 50k and 50m, Male to Female
   2. What age groups are the fastest in a 50-mile race (minimum 20 races) (Top 15)
   3. What age groups are the slowest in a 50-mile race (minimum 10 races)(Top 20)
   4. # Are athletes slower in the summer than winter? (Split between two decimals)

# spring - 3-5

# summer - 6-8

# Fall - 9-11

# Winter - 12-2