

ETL

April 6, 2024

1 Two Centuries of Ultra Marathon

In this project, I will conduct an Exploratory Data Analysis (EDA) on the dataset The big dataset of ultra-marathon running, which is available on Kaggle. Although my dataset is not exceptionally large (around 1 gigabyte in CSV format), it is still significant. My aim is to perform data manipulation and EDA on this large dataset using the Pandas library.

```
[ ]: import seaborn as sns
import pandas as pd
import numpy as np

[ ]: data = pd.read_csv("TWO_CENTURIES_OF_UM_RACES.csv", dtype={'Athlete average_
↳speed': 'object'})
data.head(3)
```

```
[ ]:      Year of event Event dates      Event name Event distance/length \
0          2018  06.01.2018  Selva Costera (CHI)          50km
1          2018  06.01.2018  Selva Costera (CHI)          50km
2          2018  06.01.2018  Selva Costera (CHI)          50km

      Event number of finishers Athlete performance      Athlete club \
0                22          4:51:39 h          Tnfrc
1                22          5:15:45 h  Roberto Echeverría
2                22          5:16:44 h   Puro Trail Osorno

      Athlete country  Athlete year of birth Athlete gender Athlete age category \
0                CHI          1978.0                M          M35
1                CHI          1981.0                M          M35
2                CHI          1987.0                M          M23

      Athlete average speed  Athlete ID
0                10.286            0
1                9.501            1
2                9.472            2
```

2 Rename columns

To make it easier to continue, I will rename all column titles to lowercase and remove any spaces

```
[ ]: data.columns = ["year_of_event", "event_dates", "event_name",
                    "event_distance/length", "event_number_of_finishers",
                    "athlete_performance", "athlete_club", "athlete_country",
                    "athlete_year_of_birth", "athlete_gender",
                    ↪ "athlete_age_category",
                    "athlete_average_speed", "athlete_id"]
```

```
[ ]: data.head(2)
```

```
[ ]:      year_of_event event_dates      event_name event_distance/length \
0          2018    06.01.2018  Selva Costera (CHI)              50km
1          2018    06.01.2018  Selva Costera (CHI)              50km

      event_number_of_finishers athlete_performance      athlete_club \
0                             22          4:51:39 h          Tnfrc
1                             22          5:15:45 h  Roberto Echeverría

      athlete_country athlete_year_of_birth athlete_gender athlete_age_category \
0                CHI          1978.0                M          M35
1                CHI          1981.0                M          M35

      athlete_average_speed athlete_id
0                10.286           0
1                9.501           1
```

```
[ ]: data.tail(2)
```

```
[ ]:      year_of_event event_dates      event_name \
7461193          1995    00.00.1995  Szombathely 24 hours running Race (HUN)
7461194          1995    00.00.1995  Szombathely 24 hours running Race (HUN)

      event_distance/length event_number_of_finishers athlete_performance \
7461193                24h                3          228.000 km
7461194                24h                3          224.000 km

      athlete_club athlete_country athlete_year_of_birth athlete_gender \
7461193      *Szeged                HUN          1959.0                M
7461194      *Pecs                HUN          1958.0                M

      athlete_age_category athlete_average_speed athlete_id
7461193                M35          9500.0          380150
7461194                M35          9333.0          1070482
```

3 Handling missing Values

```
[ ]: data.isna().sum()
```

```
[ ]: year_of_event          0
     event_dates           0
     event_name            0
     event_distance/length 1053
     event_number_of_finishers 0
     athlete_performance    2
     athlete_club          2826524
     athlete_country        3
     athlete_year_of_birth  588161
     athlete_gender         7
     athlete_age_category   584938
     athlete_average_speed  224
     athlete_id             0
     dtype: int64
```

It is better to try to calculate `athlete_year_of_birth` and `athlete_age_category` based on each other's values, unless both rows have NaN values. In that condition, we must remove those rows.

```
[ ]: data[(data["athlete_age_category"].isna() == True) &
         ↪(data["athlete_year_of_birth"].isna() == True)].shape
```

```
[ ]: (584740, 13)
```

As you can see, there are 584655 rows with both `athlete_year_of_birth` and `athlete_age_category` NaN values. The number of these rows is the same as the number of NaN values for `athlete_age_category`, so I will remove rows with NaN values for `athlete_age_category`. In another section, I will calculate `athlete_year_of_birth` based on the difference between `year_of_event` and `athlete_age_category`. (Go to the section “Filling Year of Birth” for more details).

```
[ ]: data.rename(columns={"athlete_age_category" : "athlete_age"}, inplace=True) #
     ↪renaming columns
data['athlete_club'] = data['athlete_club'].fillna('-') # fill missing values
     ↪of "athlete club" with "-"

data = data.dropna(subset=["athlete_age"])

# This is a critical value, if a row dose not include this variable, it is not
     ↪useful in next possible analysis, so i must drop these values.
data = data.dropna(subset=["event_distance/length"])
```

We must remove the ‘M’, ‘F’, and ‘W’ prefixes from the values in the `athlete_age` column.

```
[ ]: # get degite part of colum values, raw format of row values is like this M35 or W26
data["athlete_age"] = data["athlete_age"].str.extract('(\d+)').astype(int)
```

```
[ ]: data.head()
```

```
[ ]:
year_of_event event_dates      event_name event_distance/length \
0          2018  06.01.2018  Selva Costera (CHI)          50km
1          2018  06.01.2018  Selva Costera (CHI)          50km
2          2018  06.01.2018  Selva Costera (CHI)          50km
3          2018  06.01.2018  Selva Costera (CHI)          50km
4          2018  06.01.2018  Selva Costera (CHI)          50km

event_number_of_finishers athlete_performance      athlete_club \
0                        22          4:51:39 h          Tnfrc
1                        22          5:15:45 h  Roberto Echeverría
2                        22          5:16:44 h   Puro Trail Osorno
3                        22          5:34:13 h          Columbia
4                        22          5:54:14 h   Baguales Trail

athlete_country athlete_year_of_birth athlete_gender athlete_age \
0             CHI          1978.0             M          35
1             CHI          1981.0             M          35
2             CHI          1987.0             M          23
3             ARG          1976.0             M          40
4             CHI          1992.0             M          23

athlete_average_speed athlete_id
0             10.286           0
1              9.501           1
2              9.472           2
3              8.976           3
4              8.469           4
```

it looks good:))

4 Filling Year of Birth

Some rows lack the year of birth but contain the athlete's age and the year of the event. By calculating the difference between these two elements for each row, we can determine the birth year of each athlete.

```
[ ]: def calculate_year_of_birth(row):
      if pd.isna(row['athlete_year_of_birth']):
          return row['year_of_event'] - row['athlete_age']
```

```

else:
    return row['athlete_year_of_birth']

```

```

data["athlete_year_of_birth"] = data.apply(calculate_year_of_birth, axis=1)
data.head(4)

```

```

[ ]:   year_of_event event_dates      event_name event_distance/length \
0         2018  06.01.2018  Selva Costera (CHI)              50km
1         2018  06.01.2018  Selva Costera (CHI)              50km
2         2018  06.01.2018  Selva Costera (CHI)              50km
3         2018  06.01.2018  Selva Costera (CHI)              50km

      event_number_of_finishers athlete_performance      athlete_club \
0                             22      4:51:39 h      Tnfrc
1                             22      5:15:45 h  Roberto Echeverría
2                             22      5:16:44 h  Puro Trail Osorno
3                             22      5:34:13 h      Columbia

      athlete_country athlete_year_of_birth athlete_gender athlete_age \
0          CHI          1978.0          M          35
1          CHI          1981.0          M          35
2          CHI          1987.0          M          23
3          ARG          1976.0          M          40

      athlete_average_speed athlete_id
0          10.286          0
1           9.501          1
2           9.472          2
3           8.976          3

```

```

[ ]: data.isna().sum()

```

```

[ ]: year_of_event          0
     event_dates           0
     event_name            0
     event_distance/length  0
     event_number_of_finishers 0
     athlete_performance    0
     athlete_club           0
     athlete_country        0
     athlete_year_of_birth  0
     athlete_gender         0
     athlete_age            0
     athlete_average_speed  221
     athlete_id             0
     dtype: int64

```

5 Data Swap

In certain rows, there appears to be a mix-up in the recording of data between “athlete_performance” and “event_distance/length”. Specifically, some rows have the values intended for “athlete_performance” mistakenly recorded under “event_distance/length”, and vice versa.

To rectify this issue, we need to clean these rows by swapping the values appropriately between the two columns, ensuring that each column contains the correct data. This data cleaning process will help maintain the accuracy and integrity of our dataset.

```
[ ]: data[-5:]
```

```
[ ]:      year_of_event event_dates      event_name \
7461188      1995  00.00.1995      Les 24 heures de Fleurbaix (FRA)
7461189      1995  00.00.1995      Les 24 heures de Fleurbaix (FRA)
7461192      1995  00.00.1995  Szombathely 24 hours running Race (HUN)
7461193      1995  00.00.1995  Szombathely 24 hours running Race (HUN)
7461194      1995  00.00.1995  Szombathely 24 hours running Race (HUN)

      event_distance/length  event_number_of_finishers  athlete_performance \
7461188              24h              2              232.810 km
7461189              24h              2              221.374 km
7461192              24h              3              241.000 km
7461193              24h              3              228.000 km
7461194              24h              3              224.000 km

      athlete_club  athlete_country  athlete_year_of_birth  athlete_gender \
7461188      -              FRA              1958.0              M
7461189      -              BEL              1951.0              M
7461192      *Budapest              HUN              1950.0              M
7461193      *Szeged              HUN              1959.0              M
7461194      *Pecs              HUN              1958.0              M

      athlete_age  athlete_average_speed  athlete_id
7461188      35              9700.0      1069476
7461189      40              9224.0      1045647
7461192      40              10042.0      1047373
7461193      35              9500.0      380150
7461194      35              9333.0      1070482
```

```
[ ]: def swap_value(row):
      if ("km" in row['athlete_performance']) or ("mi" in_
      row['athlete_performance']):

      # Swap
      tanker = row['athlete_performance']
      row['athlete_performance'] = row['event_distance/length']
      row['event_distance/length'] = tanker
```

```
return row
```

```
data = data.apply(swap_value, axis=1)
data
```

```
[ ]:      year_of_event event_dates      event_name \
0          2018  06.01.2018      Selva Costera (CHI)
1          2018  06.01.2018      Selva Costera (CHI)
2          2018  06.01.2018      Selva Costera (CHI)
3          2018  06.01.2018      Selva Costera (CHI)
4          2018  06.01.2018      Selva Costera (CHI)
...          ...          ...          ...
7461188      1995  00.00.1995      Les 24 heures de Fleurbaix (FRA)
7461189      1995  00.00.1995      Les 24 heures de Fleurbaix (FRA)
7461192      1995  00.00.1995  Szombathely 24 hours running Race (HUN)
7461193      1995  00.00.1995  Szombathely 24 hours running Race (HUN)
7461194      1995  00.00.1995  Szombathely 24 hours running Race (HUN)

      event_distance/length  event_number_of_finishers  athlete_performance \
0              50km              22              4:51:39 h
1              50km              22              5:15:45 h
2              50km              22              5:16:44 h
3              50km              22              5:34:13 h
4              50km              22              5:54:14 h
...          ...          ...          ...
7461188      232.810 km              2              24h
7461189      221.374 km              2              24h
7461192      241.000 km              3              24h
7461193      228.000 km              3              24h
7461194      224.000 km              3              24h

      athlete_club  athlete_country  athlete_year_of_birth \
0          Tnfrc          CHI          1978.0
1  Roberto Echeverría          CHI          1981.0
2  Puro Trail Osorno          CHI          1987.0
3      Columbia          ARG          1976.0
4  Baguales Trail          CHI          1992.0
...          ...          ...          ...
7461188          -          FRA          1958.0
7461189          -          BEL          1951.0
7461192  *Budapest          HUN          1950.0
7461193  *Szeged          HUN          1959.0
7461194  *Pecs          HUN          1958.0

      athlete_gender  athlete_age  athlete_average_speed  athlete_id
0          M          35          10.286          0
```

1	M	35	9.501	1
2	M	23	9.472	2
3	M	40	8.976	3
4	M	23	8.469	4
...
7461188	M	35	9700.0	1069476
7461189	M	40	9224.0	1045647
7461192	M	40	10042.0	1047373
7461193	M	35	9500.0	380150
7461194	M	35	9333.0	1070482

[6875289 rows x 13 columns]

```
[ ]: data[-5:-3]
```

```
[ ]:      year_of_event event_dates      event_name \
7461188      1995  00.00.1995  Les 24 heures de Fleurbaix (FRA)
7461189      1995  00.00.1995  Les 24 heures de Fleurbaix (FRA)

      event_distance/length  event_number_of_finishers athlete_performance \
7461188      232.810 km      2      24h
7461189      221.374 km      2      24h

      athlete_club athlete_country  athlete_year_of_birth athlete_gender \
7461188      -      FRA      1958.0      M
7461189      -      BEL      1951.0      M

      athlete_age  athlete_average_speed  athlete_id
7461188      35      9700.0      1069476
7461189      40      9224.0      1045647
```

as you can see now values are in good form. here you can a capture of process we made: ##

1 data[-5:]
✓ 0.0s

	year_of_event	event_dates	event_name	event_distance/length	event_number_of_finishers	athlete_performance	athlete_club
7461188	1995	00.00.1995	Les 24 heures de Fleurbaix (FRA)	232.810 km	2	24h	-
7461189	1995	00.00.1995	Les 24 heures de Fleurbaix (FRA)	221.374 km	2	24h	-

Before:

5.1 After:

```
1 data.tail()
```

	year_of_event	event_dates	event_name	event_distance/length	event_number_of_finishers	athlete_performance	athlete_club
7461188	1995	00.00.1995	Les 24 heures de Fleurbaix (FRA)	232.810 km	2	24h	-
7461189	1995	00.00.1995	Les 24 heures de Fleurbaix (FRA)	221.374 km	2	Looking goood. 24h	-

6 Set suitable Data types

6.1 event_distance/length DataCleaning

1 mile = 1.609344 km

```
[ ]: data.head(2)
```

```
[ ]:      year_of_event event_dates      event_name event_distance/length \
0          2018    06.01.2018  Selva Costera (CHI)          50km
1          2018    06.01.2018  Selva Costera (CHI)          50km

      event_number_of_finishers athlete_performance      athlete_club \
0                        22          4:51:39 h          Tnfrc
1                        22          5:15:45 h  Roberto Echeverría

      athlete_country  athlete_year_of_birth athlete_gender  athlete_age \
0             CHI          1978.0             M             35
1             CHI          1981.0             M             35

      athlete_average_speed  athlete_id
0             10.286             0
1             9.501             1
```

```
[ ]: def clean_distance(row):

      distance = row["event_distance/length"]

      if distance is None or not isinstance(distance, str):
          return None

      if "km" in distance:
          return pd.to_numeric(distance.replace("km", "").strip(),
                                errors='coerce')

      elif "mi" in distance:
```

```

        return pd.to_numeric(distance.replace("mi", "").strip(),
                                errors='coerce') * 1.609344

```

```

[ ]: data["event_distance/length"] = data.apply(clean_distance, axis=1)

```

```

[ ]: data.rename(columns={"event_distance/length": "event_distance/length(km)"},
                inplace=True)

```

```

[ ]: data.head(2)

```

```

[ ]:
   year_of_event  event_dates      event_name  event_distance/length(km) \
0             2018  06.01.2018  Selva Costera (CHI)                    50.0
1             2018  06.01.2018  Selva Costera (CHI)                    50.0

   event_number_of_finishers  athlete_performance      athlete_club \
0                          22          4:51:39 h              Tnfrc
1                          22          5:15:45 h  Roberto Echeverría

   athlete_country  athlete_year_of_birth  athlete_gender  athlete_age \
0                CHI                1978.0                M            35
1                CHI                1981.0                M            35

   athlete_average_speed  athlete_id
0                   10.286            0
1                   9.501            1

```

6.2 athlete_performance DataCleaning

```

[ ]: def clean_timing(row):

    time = row["athlete_performance"]

    if time is None or not isinstance(time, str):
        return None

    elif "h" in time:
        return pd.to_numeric(time.replace("h", "").strip(), errors='coerce')

    elif "d" in time:
        return pd.to_numeric(time.replace("d", "").strip(), errors='coerce')*24

    elif time in ["150km/3Etappen", "100km Split", "07:35", "6:40"]: # time/
        distance is not given. THESE ARE OUR CHUNKY ROWS
        return None # to drop all chunky rows in one action using dropna()

```

```

[ ]: data["athlete_performance"] = data.apply(clean_timing, axis=1)

```

```
[ ]: data.dropna(subset=["athlete_performance"], inplace=True)
```

```
[ ]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
Index: 485912 entries, 22 to 7461194  
Data columns (total 13 columns):  
#   Column                                Non-Null Count  Dtype  
---  -  
0   year_of_event                        485912 non-null  int64  
1   event_dates                          485912 non-null  object  
2   event_name                           485912 non-null  object  
3   event_distance/length(km)           485889 non-null  float64  
4   event_number_of_finishers            485912 non-null  int64  
5   athlete_performance                  485912 non-null  float64  
6   athlete_club                         485912 non-null  object  
7   athlete_country                      485912 non-null  object  
8   athlete_year_of_birth                485912 non-null  float64  
9   athlete_gender                       485912 non-null  object  
10  athlete_age                          485912 non-null  int64  
11  athlete_average_speed                485889 non-null  object  
12  athlete_id                           485912 non-null  int64  
dtypes: float64(3), int64(4), object(6)  
memory usage: 51.9+ MB
```

```
[ ]: data.to_csv('clean_dataset.csv', index=False)
```

```
[ ]:
```