Exploratory Data Analysis
Two Centuries of Ultra Marathon
Data Preprocessing-Gender based Comparison

February 27, 2024

```
[1]: import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
/var/folders/jf/719gm8g97dx90xtt2vrht47c0000gn/T/ipykernel_2037/3632437423.py:1:
DeprecationWarning:
Pyarrow will become a required dependency of pandas in the next major release of
pandas (pandas 3.0),
(to allow more performant data types, such as the Arrow string type, and better
interoperability with other libraries)
but was not found to be installed on your system.
If this would cause problems for you,
please provide us feedback at https://github.com/pandas-dev/pandas/issues/54466

  import pandas as pd
```

```
[2]: df= pd.read_csv('TWO_CENTURIES_OF_UM_RACES.csv')
```

```
/var/folders/jf/719gm8g97dx90xtt2vrht47c0000gn/T/ipykernel_2037/500617103.py:1:
DtypeWarning: Columns (11) have mixed types. Specify dtype option on import or
set low_memory=False.
  df= pd.read_csv('TWO_CENTURIES_OF_UM_RACES.csv')
```

```
[3]: df.head(10)
```

```
[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
     3           2018  06.01.2018  Selva Costera (CHI)                  50km
     4           2018  06.01.2018  Selva Costera (CHI)                  50km
     5           2018  06.01.2018  Selva Costera (CHI)                  50km
     6           2018  06.01.2018  Selva Costera (CHI)                  50km
     7           2018  06.01.2018  Selva Costera (CHI)                  50km
     8           2018  06.01.2018  Selva Costera (CHI)                  50km
     9           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
     5                         22           6:25:01 h                  NaN
     6                         22           6:28:00 h        Los Patagones
     7                         22           6:32:24 h       Reaktiva Chile
     8                         22           6:39:08 h    Puro Trail Osorno
     9                         22           6:45:11 h  Marlene Flores Team

       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
     3             ARG                 1976.0              M                  M40
     4             CHI                 1992.0              M                  M23
     5             ARG                 1974.0              M                  M40
     6             ARG                 1979.0              F                  W35
     7             CHI                 1967.0              F                  W50
     8             CHI                 1985.0              M                  M23
     9             CHI                 1976.0              M                  M40

       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
     5                 7.792           5
     6                 7.732           6
     7                 7.645           7
     8                 7.516           8
     9                 7.404           9
```

[4]:
```
df.shape
```

[4]:
```
(7461195, 13)
```

[5]:
```
df.dtypes
```

[5]:
```
Year of event                  int64
Event dates                   object
Event name                    object
Event distance/length         object
Event number of finishers      int64
Athlete performance           object
Athlete club                  object
Athlete country               object
Athlete year of birth        float64
Athlete gender                object
Athlete age category          object
Athlete average speed         object
Athlete ID                     int64
dtype: object
```

[6]:
```
#cleaning up data
```

[7]:
```
#only want 50km or 50 mi
```

[8]:
```
df[df['Event distance/length'] == '50km']
```

[8]:
```
         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)
…                  …          …                                  …
7461089           1995  07.01.1995  Centenary Lakes 50 Km Track Run (AUS)
7461090           1995  07.01.1995  Centenary Lakes 50 Km Track Run (AUS)
7461091           1995  07.01.1995  Centenary Lakes 50 Km Track Run (AUS)
7461092           1995  07.01.1995  Centenary Lakes 50 Km Track Run (AUS)
7461093           1995  07.01.1995  Centenary Lakes 50 Km Track Run (AUS)

        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
…                         …                          …                  …
7461089                  50km                          6           4:19:56 h
7461090                  50km                          6           4:28:57 h
7461091                  50km                          6           4:46:39 h
7461092                  50km                          6           4:47:39 h
7461093                  50km                          6           5:58:16 h

               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
```

```
…                     …                …                    …
7461089                *QLD             AUS                  1956.0
7461090                *QLD             AUS                  1954.0
7461091                *QLD             AUS                  1951.0
7461092                *QLD             AUS                  1939.0
7461093                *QLD             AUS                  1938.0

        Athlete gender Athlete age category Athlete average speed  Athlete ID
0                    M                  M35                10.286           0
1                    M                  M35                 9.501           1
2                    M                  M23                 9.472           2
3                    M                  M40                 8.976           3
4                    M                  M23                 8.469           4
…                …                   …                    …         …
7461089              F                  W35               11541.0     1046326
7461090              M                  M40               11154.0     1070007
7461091              M                  M40               10466.0      345672
7461092              M                  M55               10429.0     1082443
7461093              F                  W55                8374.0     1082581

[1522609 rows x 13 columns]
```

[9]:
```
df[df['Event distance/length'] == '50mi']
```

[9]:
```
         Year of event Event dates  \
55                2018  06.01.2018
56                2018  06.01.2018
57                2018  06.01.2018
58                2018  06.01.2018
59                2018  06.01.2018
…                 …          …
7461181           1995  07.01.1995
7461182           1995  07.01.1995
7461183           1995  07.01.1995
7461184           1995  07.01.1995
7461185           1995  07.01.1995

                                        Event name Event distance/length  \
55       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
56       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
57       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
58       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
59       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
…                                                 …                   …
7461181               Avalon Benefit 50-Mile Run (USA)                  50mi
7461182               Avalon Benefit 50-Mile Run (USA)                  50mi
7461183               Avalon Benefit 50-Mile Run (USA)                  50mi
```

```
7461184               Avalon Benefit 50-Mile Run (USA)                  50mi
7461185               Avalon Benefit 50-Mile Run (USA)                  50mi

         Event number of finishers Athlete performance      Athlete club  \
55                               9           9:53:05 h  *Middleville, MI
56                               9          11:09:35 h     *Waterloo, ON
57                               9          11:33:00 h    *Kitchener, ON
58                               9          11:38:17 h        *Utica, MI
59                               9          11:56:35 h   *Grass Lake, MI
…                            …                 …               …
7461181                         92          11:59:37 h               NaN
7461182                         92          12:01:41 h               NaN
7461183                         92          12:03:26 h               NaN
7461184                         92          12:03:26 h               NaN
7461185                         92          12:05:59 h               NaN

        Athlete country  Athlete year of birth Athlete gender  \
55                  USA                 1983.0              M
56                  CAN                 1977.0              F
57                  CAN                 1976.0              M
58                  USA                 1986.0              M
59                  USA                 1988.0              M
…                 …                    …            …
7461181             USA                 1941.0              M
7461182             USA                 1932.0              M
7461183             USA                 1934.0              F
7461184             USA                 1951.0              F
7461185             USA                 1947.0              F

        Athlete age category Athlete average speed  Athlete ID
55                       M23                 8.141          55
56                       W40                 7.211          56
57                       M40                 6.967          57
58                       M23                 6.914          58
59                       M23                 6.738          59
…                      …                   …         …
7461181                  M50                6709.0     1045603
7461182                  M60                6690.0     1070463
7461183                  W60                6674.0      416139
7461184                  W40                6674.0     1098098
7461185                  W45                6650.0     1626367

[352181 rows x 13 columns]
```

```
[10]: #now we have to combine both 50 miles and 50 Km
```

```
[11]: df[df['Event distance/length'].isin(['50km','50mi'])]
```

```
[11]:          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)
…                  …         …                                …
7461181           1995  07.01.1995  Avalon Benefit 50-Mile Run (USA)
7461182           1995  07.01.1995  Avalon Benefit 50-Mile Run (USA)
7461183           1995  07.01.1995  Avalon Benefit 50-Mile Run (USA)
7461184           1995  07.01.1995  Avalon Benefit 50-Mile Run (USA)
7461185           1995  07.01.1995  Avalon Benefit 50-Mile Run (USA)

        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
…                        …                        …                …
7461181                  50mi                         92          11:59:37 h
7461182                  50mi                         92          12:01:41 h
7461183                  50mi                         92          12:03:26 h
7461184                  50mi                         92          12:03:26 h
7461185                  50mi                         92          12:05:59 h

               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
…                      …             …                      …
7461181                 NaN             USA                  1941.0
7461182                 NaN             USA                  1932.0
7461183                 NaN             USA                  1934.0
7461184                 NaN             USA                  1951.0
7461185                 NaN             USA                  1947.0

        Athlete gender Athlete age category Athlete average speed  Athlete ID
0                    M                  M35                10.286           0
1                    M                  M35                 9.501           1
2                    M                  M23                 9.472           2
3                    M                  M40                 8.976           3
4                    M                  M23                 8.469           4
…                 …                  …                   …         …
7461181              M                  M50                6709.0     1045603
```

```
7461182                  M                   M60                6690.0      1070463
7461183                  F                   W60                6674.0       416139
7461184                  F                   W40                6674.0      1098098
7461185                  F                   W45                6650.0      1626367

[1874790 rows x 13 columns]
```

[12]:
```
df[df['Event name'] == 'Everglades 50 Mile Ultra Run (USA)']['Event name'].str.
 ↪split('(').str.get(1).str.split(')').str.get(0)
```

[12]:
```
51923      USA
51924      USA
51925      USA
51926      USA
51927      USA
          …
6417091    USA
6417092    USA
6417093    USA
6417094    USA
6417095    USA
Name: Event name, Length: 338, dtype: object
```

[13]:
```
#this way we have all the races in USA
```

[14]:
```
df[df['Event name'].str.split('(').str.get(1).str.split(')').str.get(0) ==␣
 ↪'USA' ]
```

[14]:
```
         Year of event Event dates  \
55                2018  06.01.2018
56                2018  06.01.2018
57                2018  06.01.2018
58                2018  06.01.2018
59                2018  06.01.2018
…                  …          …
7461181           1995  07.01.1995
7461182           1995  07.01.1995
7461183           1995  07.01.1995
7461184           1995  07.01.1995
7461185           1995  07.01.1995

                                           Event name Event distance/length  \
55       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
56       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
57       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
58       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
59       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
```

```
…                                                 …                   …
7461181               Avalon Benefit 50-Mile Run (USA)                   50mi
7461182               Avalon Benefit 50-Mile Run (USA)                   50mi
7461183               Avalon Benefit 50-Mile Run (USA)                   50mi
7461184               Avalon Benefit 50-Mile Run (USA)                   50mi
7461185               Avalon Benefit 50-Mile Run (USA)                   50mi

         Event number of finishers Athlete performance      Athlete club  \
55                               9           9:53:05 h  *Middleville, MI
56                               9          11:09:35 h     *Waterloo, ON
57                               9          11:33:00 h    *Kitchener, ON
58                               9          11:38:17 h        *Utica, MI
59                               9          11:56:35 h   *Grass Lake, MI
…                            …                 …                …
7461181                         92          11:59:37 h               NaN
7461182                         92          12:01:41 h               NaN
7461183                         92          12:03:26 h               NaN
7461184                         92          12:03:26 h               NaN
7461185                         92          12:05:59 h               NaN

        Athlete country  Athlete year of birth Athlete gender  \
55                  USA                 1983.0              M
56                  CAN                 1977.0              F
57                  CAN                 1976.0              M
58                  USA                 1986.0              M
59                  USA                 1988.0              M
…                 …                     …             …
7461181             USA                 1941.0              M
7461182             USA                 1932.0              M
7461183             USA                 1934.0              F
7461184             USA                 1951.0              F
7461185             USA                 1947.0              F

        Athlete age category Athlete average speed  Athlete ID
55                       M23                 8.141          55
56                       W40                 7.211          56
57                       M40                 6.967          57
58                       M23                 6.914          58
59                       M23                 6.738          59
…                      …                   …          …
7461181                  M50                6709.0     1045603
7461182                  M60                6690.0     1070463
7461183                  W60                6674.0      416139
7461184                  W40                6674.0     1098098
7461185                  W45                6650.0     1626367

[1398540 rows x 13 columns]
```

[15]:

```
#combine all data filters
```

[16]:

```
df2 = df[(df['Event distance/length'].isin(['50km', '50mi'])) & (df['Event␣
↪name'].str.split('(').str.get(1).str.split(')').str.get(0) == 'USA')]
```

[17]:

```
df2
```

[17]:

```
         Year of event Event dates  \
55                2018  06.01.2018
56                2018  06.01.2018
57                2018  06.01.2018
58                2018  06.01.2018
59                2018  06.01.2018
…                …         …
7461181           1995  07.01.1995
7461182           1995  07.01.1995
7461183           1995  07.01.1995
7461184           1995  07.01.1995
7461185           1995  07.01.1995

                                            Event name Event distance/length  \
55       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
56       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
57       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
58       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
59       Yankee Springs 50 Mile Winter Challenge (USA)                  50mi
…                                                 …                   …
7461181               Avalon Benefit 50-Mile Run (USA)                  50mi
7461182               Avalon Benefit 50-Mile Run (USA)                  50mi
7461183               Avalon Benefit 50-Mile Run (USA)                  50mi
7461184               Avalon Benefit 50-Mile Run (USA)                  50mi
7461185               Avalon Benefit 50-Mile Run (USA)                  50mi

         Event number of finishers Athlete performance      Athlete club  \
55                               9           9:53:05 h  *Middleville, MI
56                               9          11:09:35 h     *Waterloo, ON
57                               9          11:33:00 h    *Kitchener, ON
58                               9          11:38:17 h        *Utica, MI
59                               9          11:56:35 h   *Grass Lake, MI
…                            …                 …               …
7461181                         92          11:59:37 h               NaN
7461182                         92          12:01:41 h               NaN
7461183                         92          12:03:26 h               NaN
7461184                         92          12:03:26 h               NaN
7461185                         92          12:05:59 h               NaN

        Athlete country  Athlete year of birth Athlete gender  \
55                  USA                 1983.0              M
56                  CAN                 1977.0              F
57                  CAN                 1976.0              M
58                  USA                 1986.0              M
59                  USA                 1988.0              M
…                 …                    …            …
7461181             USA                 1941.0              M
7461182             USA                 1932.0              M
7461183             USA                 1934.0              F
7461184             USA                 1951.0              F
7461185             USA                 1947.0              F

        Athlete age category Athlete average speed  Athlete ID
55                       M23                 8.141          55
56                       W40                 7.211          56
57                       M40                 6.967          57
58                       M23                 6.914          58
59                       M23                 6.738          59
…                      …                   …         …
7461181                  M50                6709.0     1045603
7461182                  M60                6690.0     1070463
7461183                  W60                6674.0      416139
7461184                  W40                6674.0     1098098
7461185                  W45                6650.0     1626367

[926241 rows x 13 columns]
```

```
[18]: df2.shape
```

```
[18]: (926241, 13)
```

```
[19]: #removing usa from eventsname
```

```
[20]: df2['Event name'] = df2['Event name'].str.split('(').str.get(0)
```

```
/var/folders/jf/719gm8g97dx90xtt2vrht47c0000gn/T/ipykernel_2037/3473829760.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df2['Event name'] = df2['Event name'].str.split('(').str.get(0)
```

```
[21]: df2.head()
```

[21]:

```
    Year of event Event dates                                Event name  \
55           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   
56           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   
57           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   
58           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   
59           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   

   Event distance/length  Event number of finishers Athlete performance  \
55                  50mi                          9           9:53:05 h   
56                  50mi                          9          11:09:35 h   
57                  50mi                          9          11:33:00 h   
58                  50mi                          9          11:38:17 h   
59                  50mi                          9          11:56:35 h   

        Athlete club Athlete country  Athlete year of birth Athlete gender  \
55  *Middleville, MI             USA                 1983.0              M   
56     *Waterloo, ON             CAN                 1977.0              F   
57    *Kitchener, ON             CAN                 1976.0              M   
58        *Utica, MI             USA                 1986.0              M   
59   *Grass Lake, MI             USA                 1988.0              M   

   Athlete age category Athlete average speed  Athlete ID  
55                  M23                 8.141          55  
56                  W40                 7.211          56  
57                  M40                 6.967          57  
58                  M23                 6.914          58  
59                  M23                 6.738          59  
```

[22]:

```
# lets remove the 'h' from that athlete performance
```

[23]:

```
df2['Athlete performance'] = df2['Athlete performance'].str.split(' ').str.
 ↪get(0)
```

```
/var/folders/jf/719gm8g97dx90xtt2vrht47c0000gn/T/ipykernel_2037/2477507555.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df2['Athlete performance'] = df2['Athlete performance'].str.split('
').str.get(0)
```

[24]:

```
df2.head(10)
```

[24]:

```
    Year of event Event dates                                Event name  \
55           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge   
```

```
56           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
57           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
58           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
59           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
60           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
61           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
62           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
63           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
64           2018  06.01.2018    Yankee Springs 50 km Winter Challenge

   Event distance/length  Event number of finishers Athlete performance  \
55                  50mi                          9             9:53:05
56                  50mi                          9            11:09:35
57                  50mi                          9            11:33:00
58                  50mi                          9            11:38:17
59                  50mi                          9            11:56:35
60                  50mi                          9            12:32:16
61                  50mi                          9            12:39:36
62                  50mi                          9            12:39:36
63                  50mi                          9            13:24:05
64                  50km                         36             5:09:40

         Athlete club Athlete country  Athlete year of birth Athlete gender  \
55   *Middleville, MI             USA                 1983.0              M
56      *Waterloo, ON             CAN                 1977.0              F
57     *Kitchener, ON             CAN                 1976.0              M
58         *Utica, MI             USA                 1986.0              M
59    *Grass Lake, MI             USA                 1988.0              M
60        *Olaton, KY             USA                 1995.0              M
61       *Wyoming, MI             USA                 1979.0              M
62  *Grand Rapids, MI             USA                 1977.0              F
63       *Lansing, MI             USA                 1990.0              F
64        *Okemos, MI             USA                 1991.0              F

   Athlete age category Athlete average speed  Athlete ID
55                  M23                 8.141          55
56                  W40                 7.211          56
57                  M40                 6.967          57
58                  M23                 6.914          58
59                  M23                 6.738          59
60                 MU23                 6.418          60
61                  M35                 6.356          61
62                  W40                 6.356          62
63                  W23                 6.004          63
64                  W23                 9.688          64
```

[25]:

```
#lets drop columns that arent important
```

[26]:
```
# athlete club, country, year of birth, age category
```

[27]:
```
df2 = df2.drop(['Athlete club','Athlete country','Athlete year of␣
 ↪birth','Athlete age category'], axis = 1)
```

[28]:
```
df2.head(10)
```

[28]:
```
    Year of event Event dates                              Event name  \
55           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
56           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
57           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
58           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
59           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
60           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
61           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
62           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
63           2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
64           2018  06.01.2018    Yankee Springs 50 km Winter Challenge

   Event distance/length  Event number of finishers Athlete performance  \
55                  50mi                          9             9:53:05
56                  50mi                          9            11:09:35
57                  50mi                          9            11:33:00
58                  50mi                          9            11:38:17
59                  50mi                          9            11:56:35
60                  50mi                          9            12:32:16
61                  50mi                          9            12:39:36
62                  50mi                          9            12:39:36
63                  50mi                          9            13:24:05
64                  50km                         36             5:09:40

   Athlete gender Athlete average speed  Athlete ID
55              M                 8.141          55
56              F                 7.211          56
57              M                 6.967          57
58              M                 6.914          58
59              M                 6.738          59
60              M                 6.418          60
61              M                 6.356          61
62              F                 6.356          62
63              F                 6.004          63
64              F                 9.688          64
```

[29]:
```
# lets check for null values in our table
```

[30]:
```
df2.isna().sum()
```

```
[30]: Year of event               0
      Event dates                 0
      Event name                  0
      Event distance/length       0
      Event number of finishers   0
      Athlete performance         0
      Athlete gender              0
      Athlete average speed       0
      Athlete ID                  0
      dtype: int64
```

```
[31]: # no null values--- our data is clean
```

```
[32]: # lets check for duplicate values
```

```
[33]: df2[df2.duplicated() == True]
```

```
[33]:
         Year of event     Event dates  \
1073758           2016  24.-25.09.2016
1087007           2016      17.09.2016
1103619           2016      10.09.2016
1317290           2017      18.03.2017
1399606           2017      13.05.2017
3540783           2022      17.12.2022
4238708           2005      30.07.2005
4290600           2005      26.03.2005
4666132           2009  19.-20.06.2009
4698963           2009      09.05.2009
4794364           2009      19.09.2009
5123404           2011      09.07.2011
5541799           2012      13.10.2012
5542079           2012      13.10.2012
5553711           2012      06.10.2012
5553791           2012      06.10.2012
5579518           2013      06.04.2013
5682076           2013      01.06.2013
5908250           2013      05.10.2013
6198571           2014      06.09.2014
6299545           2014      26.10.2014
6343970           2015      11.04.2015
6344012           2015      11.04.2015
6373940           2015      29.03.2015
7328184           1992      25.04.1992

                                             Event name Event distance/length  \
1073758                            Not Yo Momma's 50 km                  50km
1087007                        The Barkley Fall Classic                  50km
```

```
1103619                                  Los Pinos 50K                    50km
1317290            Lake Martin 50 Mile Ultra Trail Race                   50mi
1399606                              Quad Rock 50 Mile                    50mi
3540783                   Cave Creek Thriller 50K Race                    50km
4238708                  Pacific Crest 50 km Trail Run                    50km
4290600                             San Juan Trail 50K                    50km
4666132  Bighorn Mountain Wild & Scenic 50km Trail Run                    50km
4698963      Endurance Challenge - New York Trail 50km                    50km
4794364             Youngstown Ultra Trail Classic 50K                    50km
5123404                   Cuyamaca Three Peaks 50K Run                    50km
5541799                            Market to Market 50                    50km
5542079                                 Twin Peaks 50K                    50km
5553711                       Rock/Creek StumpJump 50K                    50km
5553791                       Rock/Creek StumpJump 50K                    50km
5579518            American River 50 Mile Endurance Run                   50mi
5682076     Rainier to Ruston Rail-Trail 50 Mile Ultra                    50mi
5908250                       Rock/Creek StumpJump 50K                    50km
6198571                                    Volcanic 50                    50km
6299545                       G.O.A.T.z 50km Trail Run                    50km
6343970                       Lake Sonoma 50 Mile Race                    50mi
6344012                       Lake Sonoma 50 Mile Race                    50mi
6373940                           Gorge Waterfalls 50k                    50km
7328184                      Cuyamaca 50 Km Trail Race                    50km

         Event number of finishers Athlete performance Athlete gender  \
1073758                         22            11:11:22              M
1087007                        120            10:56:16              F
1103619                         67             9:31:29              M
1317290                         77            10:08:39              M
1399606                        107            13:48:21              M
3540783                         76             5:51:02              M
4238708                        104             6:19:39              M
4290600                         62             8:12:37              M
4666132                        137             8:32:34              M
4698963                        157             7:44:54              M
4794364                         71             8:45:36              M
5123404                         65             6:01:00              M
5541799                        114             5:34:26              M
5542079                         23             9:39:42              M
5553711                        410             7:07:46              M
5553791                        410             7:44:47              M
5579518                        835            11:54:55              M
5682076                         60            10:46:21              M
5908250                        343             6:24:00              M
6198571                        192            11:07:59              M
6299545                        146             6:01:00              M
6343970                        278             8:08:40              M
```

```
6344012                        278            9:18:19              M
6373940                        284            6:52:32              M
7328184                        131            6:04:58              M

        Athlete average speed  Athlete ID
1073758                 4.468       26206
1087007                 4.571      333576
1103619                 5.249      366578
1317290                 7.932       34702
1399606                 5.828       33474
3540783                 8.546      863243
4238708                 7.902       94618
4290600                  6.09        4033
4666132                 5.853      683378
4698963                 6.453       84162
4794364                 5.708      337937
5123404                  8.31      297664
5541799                  8.97      510488
5542079                 5.175       13614
5553711                 7.013      387266
5553791                 6.455      118915
5579518                 6.753      544285
5682076                  7.47     1392324
5908250                 7.813       31507
6198571                 4.491      568468
6299545                  8.31      102947
6343970                  9.88      143697
6344012                 8.647       27559
6373940                 7.272        1381
7328184                8220.0     1085158
```

[34]:
```
# we have to drop duplicated data
```

[35]:
```
df2 = df2.drop_duplicates()
```

[36]:
```
df2.shape
```

[36]:
```
(926216, 9)
```

[37]:
```
# reset index
```

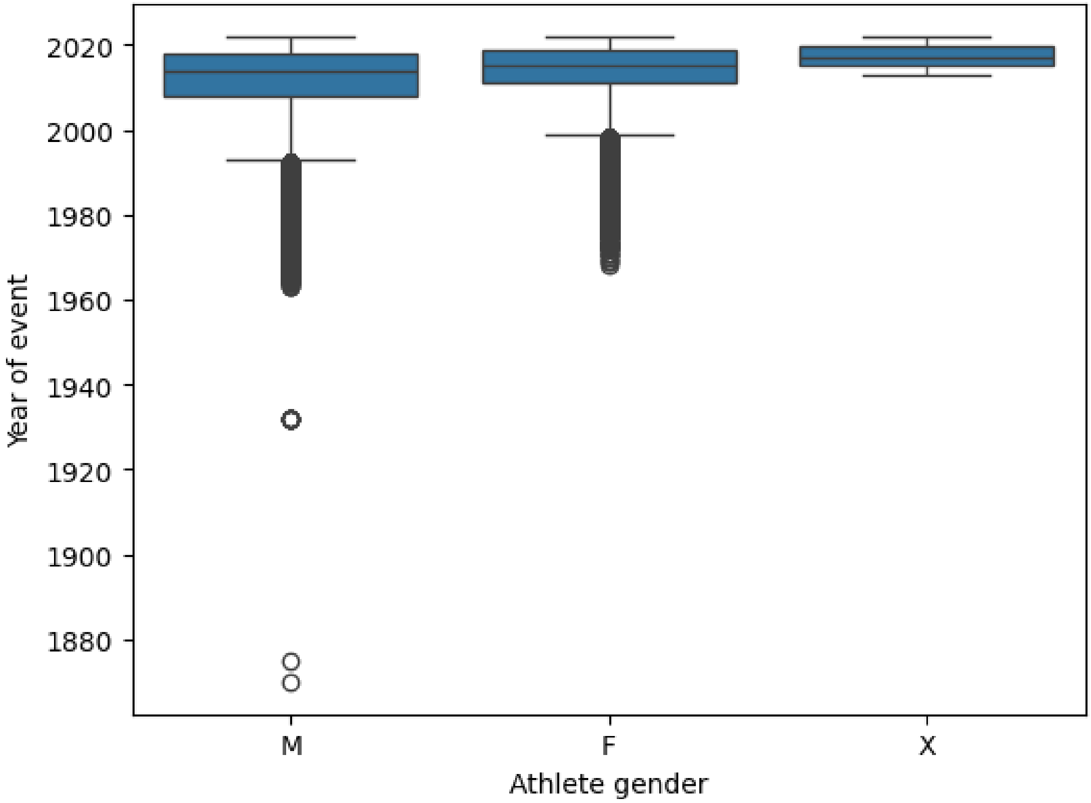
[38]:
```
df2.reset_index(drop = True)
```

[38]:
```
        Year of event Event dates                                Event name  \
0                2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
1                2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
2                2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
```

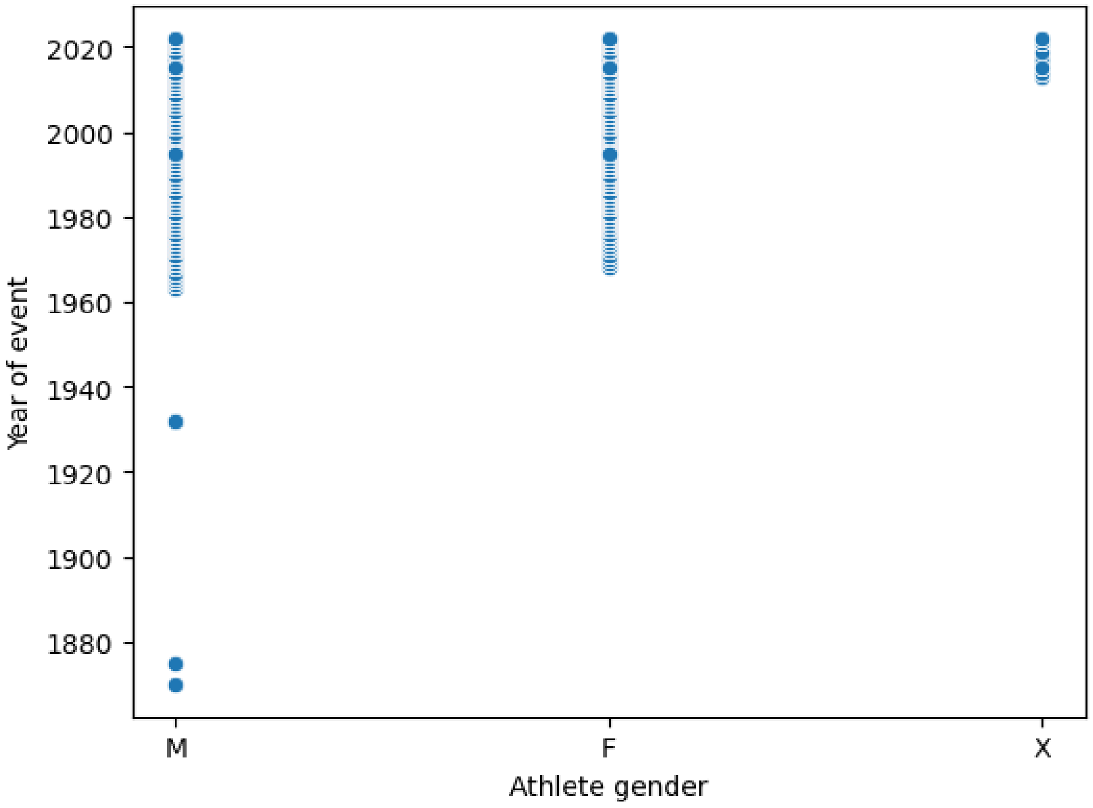
```
3                2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
4                2018  06.01.2018  Yankee Springs 50 Mile Winter Challenge
…               …         …                                       …
926211           1995  07.01.1995               Avalon Benefit 50-Mile Run
926212           1995  07.01.1995               Avalon Benefit 50-Mile Run
926213           1995  07.01.1995               Avalon Benefit 50-Mile Run
926214           1995  07.01.1995               Avalon Benefit 50-Mile Run
926215           1995  07.01.1995               Avalon Benefit 50-Mile Run

       Event distance/length  Event number of finishers Athlete performance  \
0                       50mi                          9             9:53:05
1                       50mi                          9            11:09:35
2                       50mi                          9            11:33:00
3                       50mi                          9            11:38:17
4                       50mi                          9            11:56:35
…                      …                        …                 …
926211                  50mi                         92            11:59:37
926212                  50mi                         92            12:01:41
926213                  50mi                         92            12:03:26
926214                  50mi                         92            12:03:26
926215                  50mi                         92            12:05:59

       Athlete gender Athlete average speed  Athlete ID
0                   M                 8.141          55
1                   F                 7.211          56
2                   M                 6.967          57
3                   M                 6.914          58
4                   M                 6.738          59
…               …                   …         …
926211              M                6709.0     1045603
926212              M                6690.0     1070463
926213              F                6674.0      416139
926214              F                6674.0     1098098
926215              F                6650.0     1626367

[926216 rows x 9 columns]
```

[39]:
```
#since the data is clean... we can state the visualization
```

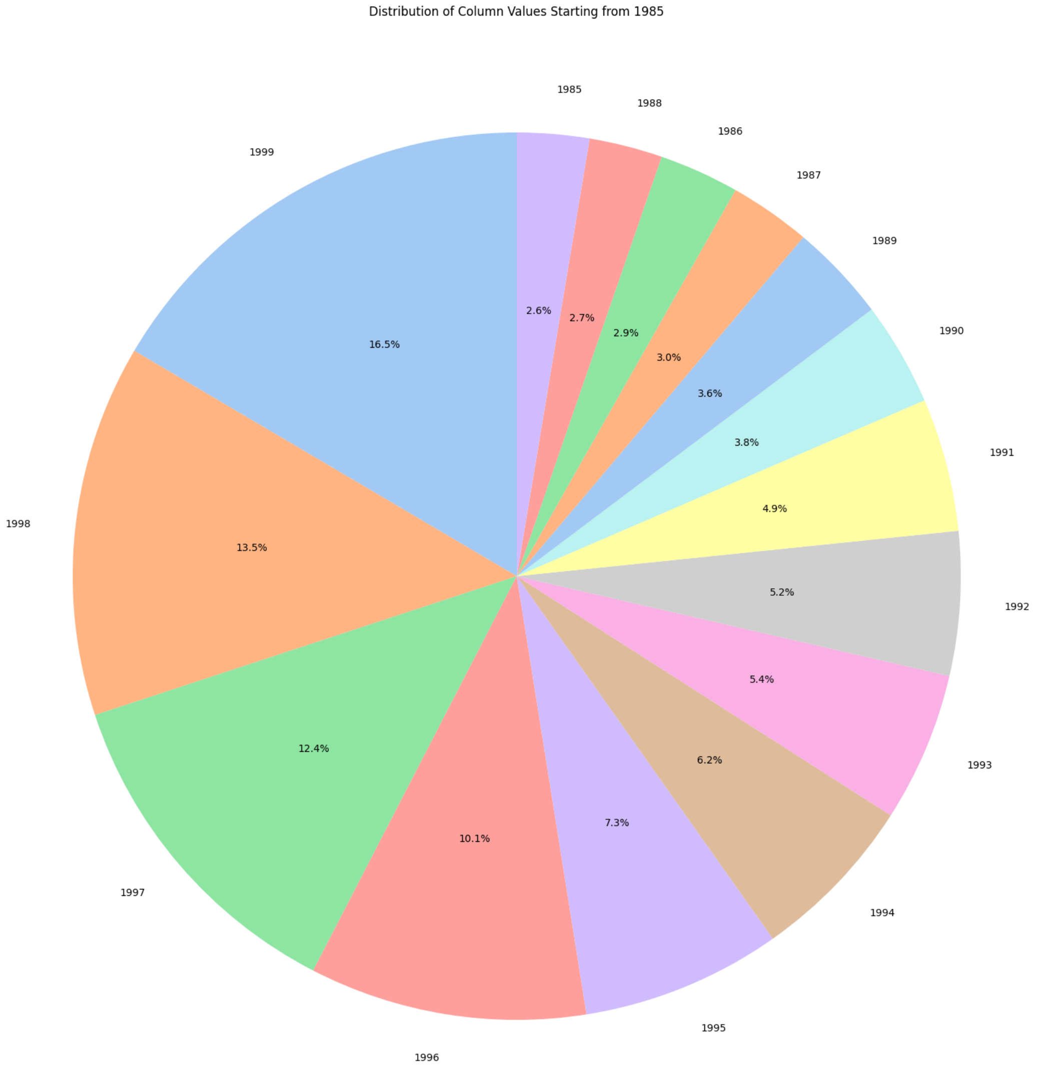
[40]:
```
sns.boxplot(data=df2, x='Athlete gender', y='Year of event')

plt.show()
```



```
[41]: # Create the scatter plot
      sns.scatterplot(data=df2, x='Athlete gender', y='Year of event')
      plt.show()
```

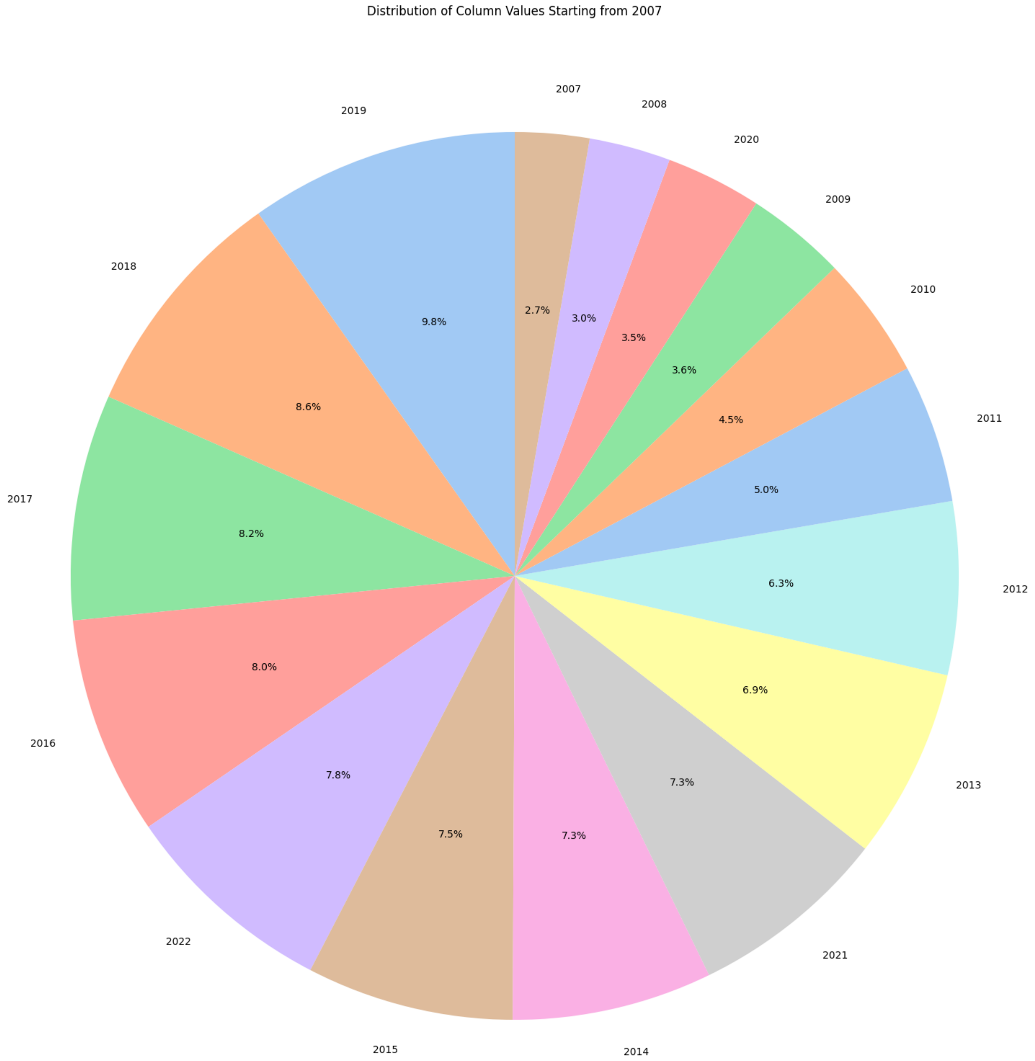


```
[42]: data = df2['Year of event'].value_counts()

      # Filter data to start from 1985 to 1999
      filtered_data = data[(data.index >= 1985) & (data.index <= 1999)]

      # Create a pie chart
      plt.figure(figsize=(20, 20))
      plt.pie(filtered_data, labels=filtered_data.index, autopct='%1.1f%%',␣
       ↪startangle=90, colors=sns.color_palette('pastel'))
      plt.title('Distribution of Column Values Starting from 1985')
      plt.show()
```



[43]:

```
# Assuming 'Year of event' is the column you want to visualize
# You may replace it with the actual column name in your DataFrame
data = df2['Year of event'].value_counts()

# Filter data to start from 2000 to 2024
filtered_data = data[(data.index >= 2007) & (data.index <= 2024)]

# Create a pie chart
plt.figure(figsize=(20, 20))
plt.pie(filtered_data, labels=filtered_data.index, autopct='%1.1f%%',␣
 ↪startangle=90, colors=sns.color_palette('pastel'))
plt.title('Distribution of Column Values Starting from 2007')
plt.show()
```


Distribution of Column Values Starting from 2007
2007
2.7%
2008
3.0%
2020
3.5%
2009
3.6%
2010
4.5%
2011
5.0%
2012
6.3%
2013
6.9%
2021
7.3%
2014
7.3%
2015
7.5%
2022
7.8%
2016
8.0%
2017
8.2%
2018
8.6%
2019
9.8%

[ ]:

```

```

[ ]:

```

```

[ ]:

```

```