In [215…
```
import pandas as pd
import numpy as np
from bokeh.layouts import column
from dask.array import invert
from holoviews.plotting.bokeh.styles import alpha
from streamlit import columns

df = pd.read_csv('../dataset/TWO_CENTURIES_OF_UM_RACES.csv')
```

```
/var/folders/q3/xgl4pwjd7lbg8skj81tsl0xr0000gn/T/ipykernel_44900/2166451878.py:7: DtypeWarning: Columns (11) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv('../dataset/TWO_CENTURIES_OF_UM_RACES.csv')
```

In [216…
```
df.shape
```

Out[216… (7461195, 13)

In [217…
```
df.columns
```

Out[217…
```
Index(['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'],
      dtype='object')
```

In [218…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7461195 entries, 0 to 7461194
Data columns (total 13 columns):
 #   Column                    Dtype
---  ------                    -----
 0   Year of event             int64
 1   Event dates               object
 2   Event name                object
 3   Event distance/length     object
 4   Event number of finishers  int64
 5   Athlete performance       object
 6   Athlete club              object
 7   Athlete country           object
 8   Athlete year of birth     float64
 9   Athlete gender            object
 10  Athlete age category      object
 11  Athlete average speed     object
 12  Athlete ID                int64
dtypes: float64(1), int64(3), object(9)
memory usage: 740.0+ MB
```

In [219…
```
df.head(5)
```

Out[219…

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2018 | 06.01.2018 | Selva Costera (CHI) | 50km | 22 | 4:51:39 h | Tnfrc | CHI | 1978.0 | M | M35 | 10.286 | 0 |
| **1** | 2018 | 06.01.2018 | Selva Costera (CHI) | 50km | 22 | 5:15:45 h | Roberto Echeverría | CHI | 1981.0 | M | M35 | 9.501 | 1 |
| **2** | 2018 | 06.01.2018 | Selva Costera (CHI) | 50km | 22 | 5:16:44 h | Puro Trail Osorno | CHI | 1987.0 | M | M23 | 9.472 | 2 |
| **3** | 2018 | 06.01.2018 | Selva Costera (CHI) | 50km | 22 | 5:34:13 h | Columbia | ARG | 1976.0 | M | M40 | 8.976 | 3 |
| **4** | 2018 | 06.01.2018 | Selva Costera (CHI) | 50km | 22 | 5:54:14 h | Baguales Trail | CHI | 1992.0 | M | M23 | 8.469 | 4 |

1. Filter combine year = 2020 and country = USA

## 2. Filter Event distance/length in 50km and 50 mi

In [220…
```
df_event_USA = df[df['Event name'].str.contains(r'\(USA\)')]
```

In [221…
```
df_event_USA.shape
```

Out[221… (1408416, 13)

In [222…
```
df_event_USA.head(10)
```

Out[222…

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **55** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 9:53:05 h | *Middleville, MI | USA | 1983.0 | M | M23 | 8.141 | 55 |
| **56** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 11:09:35 h | *Waterloo, ON | CAN | 1977.0 | F | W40 | 7.211 | 56 |
| **57** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 11:33:00 h | *Kitchener, ON | CAN | 1976.0 | M | M40 | 6.967 | 57 |
| **58** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 11:38:17 h | *Utica, MI | USA | 1986.0 | M | M23 | 6.914 | 58 |
| **59** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 11:56:35 h | *Grass Lake, MI | USA | 1988.0 | M | M23 | 6.738 | 59 |
| **60** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 12:32:16 h | *Olaton, KY | USA | 1995.0 | M | MU23 | 6.418 | 60 |
| **61** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 12:39:36 h | *Wyoming, MI | USA | 1979.0 | M | M35 | 6.356 | 61 |
| **62** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 12:39:36 h | *Grand Rapids, MI | USA | 1977.0 | F | W40 | 6.356 | 62 |
| **63** | 2018 | 06.01.2018 | Yankee Springs 50 Mile Winter Challenge (USA) | 50mi | 9 | 13:24:05 h | *Lansing, MI | USA | 1990.0 | F | W23 | 6.004 | 63 |
| **64** | 2018 | 06.01.2018 | Yankee Springs 50 km Winter Challenge (USA) | 50km | 36 | 5:09:40 h | *Okemos, MI | USA | 1991.0 | F | W23 | 9.688 | 64 |

In [223…
```
df_event_USA = df_event_USA[df_event_USA['Year of event'] == 2020 ]
df_event_USA.head(5)
```

Out[223…

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2538647** | 2020 | 07.-09.02.2020 | SC Ultra Running Festival - 48 Hour Race (USA) | 48h | 11 | 227.445 km | *Southern Shores, NC | USA | 1974.0 | F | W45 | 4.738 | 19078 |
| **2538648** | 2020 | 07.-09.02.2020 | SC Ultra Running Festival - 48 Hour Race (USA) | 48h | 11 | 208.648 km | NaN | USA | 1958.0 | F | W60 | 4.347 | 54142 |
| **2538649** | 2020 | 07.-09.02.2020 | SC Ultra Running Festival - 48 Hour Race (USA) | 48h | 11 | 183.008 km | *Hoover, AL | USA | 1938.0 | M | M80 | 3.813 | 19661 |
| **2538650** | 2020 | 07.-09.02.2020 | SC Ultra Running Festival - 48 Hour Race (USA) | 48h | 11 | 163.535 km | *Franklin, TN | USA | 1967.0 | M | M50 | 3.407 | 390210 |
| **2538651** | 2020 | 07.-09.02.2020 | SC Ultra Running Festival - 48 Hour Race (USA) | 48h | 11 | 163.535 km | *Maple Hill, NC | USA | 1970.0 | F | W45 | 3.407 | 265264 |

In [224…
```
df_event_USA = df_event_USA[df_event_USA['Event distance/length'].isin(['50mi','50km'])]
df_event_USA.head(5)
```

Out[224...

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition (USA) | 50km | 20 | 3:17:55 h | *Normandy Park, WA | USA | 1991.0 | M | M23 | 15.158 | 71287 |
| **2539946** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition (USA) | 50km | 20 | 4:02:32 h | *Gold Bar, WA | USA | 1981.0 | M | M35 | 12.369 | 629508 |
| **2539947** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition (USA) | 50km | 20 | 4:07:57 h | *Vashon, WA | USA | 1999.0 | M | MU23 | 12.099 | 64838 |
| **2539948** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition (USA) | 50km | 20 | 4:22:02 h | *Gig Harbor, WA | USA | 1983.0 | M | M35 | 11.449 | 704450 |
| **2539949** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition (USA) | 50km | 20 | 4:27:34 h | *Bainbridge Island, WA | USA | 1977.0 | M | M40 | 11.212 | 810281 |

In [225...
```
df_event_USA.shape
```

Out[225...
```
(26524, 13)
```

In [226...
```
df_event_USA['Event distance/length'].unique()
```

Out[226...
```
array(['50km', '50mi'], dtype=object)
```

3. In Event name, filter only name of event, no need USA

4. Create column 'event_month' only filter the month on column ' Event dates'

5. Delete columns unused

6. Check NA values

7. Delete NA values

8. Filter NA values

9. Change type data

In [227...
```
# 3. In Event name, filter only name of event, no need USA
df_event_USA['Event name'] = df_event_USA['Event name'].str.split('(').str.get(0)
```

In [228...
```
df_event_USA.head(5)
```

Out[228...

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition | 50km | 20 | 3:17:55 h | *Normandy Park, WA | USA | 1991.0 | M | M23 | 15.158 | 71287 |
| **2539946** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:02:32 h | *Gold Bar, WA | USA | 1981.0 | M | M35 | 12.369 | 629508 |
| **2539947** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:07:57 h | *Vashon, WA | USA | 1999.0 | M | MU23 | 12.099 | 64838 |
| **2539948** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:22:02 h | *Gig Harbor, WA | USA | 1983.0 | M | M35 | 11.449 | 704450 |
| **2539949** | 2020 | 02.02.2020 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:27:34 h | *Bainbridge Island, WA | USA | 1977.0 | M | M40 | 11.212 | 810281 |

In [229...
```
# 4. Create column 'event_month' only filter the month on column ' Event dates'
# Change type of columns Event date to datetime
df_event_USA['Event dates'] = pd.to_datetime(df_event_USA['Event dates'], dayfirst=True, errors='coerce')
```

In [230...
```
df_event_USA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 26524 entries, 2539945 to 2760961
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Year of event              26524 non-null  int64
 1   Event dates                25220 non-null  datetime64[ns]
 2   Event name                 26524 non-null  object
 3   Event distance/length      26524 non-null  object
 4   Event number of finishers  26524 non-null  int64
 5   Athlete performance        26524 non-null  object
 6   Athlete club               23394 non-null  object
 7   Athlete country            26524 non-null  object
 8   Athlete year of birth      26289 non-null  float64
 9   Athlete gender             26524 non-null  object
 10  Athlete age category       26306 non-null  object
 11  Athlete average speed      26524 non-null  object
 12  Athlete ID                 26524 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(3), object(8)
memory usage: 2.8+ MB
```

In [231…
```
#--> Change dates to month name
df_event_USA['month'] = df_event_USA['Event dates'].dt.month_name()
```

In [232…
```
df_event_USA.head(5)
```

Out[232…

| | 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 | month |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 20 | 3:17:55 h | *Normandy Park, WA | USA | 1991.0 | M | M23 | 15.158 | 71287 | February |
| **2539946** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:02:32 h | *Gold Bar, WA | USA | 1981.0 | M | M35 | 12.369 | 629508 | February |
| **2539947** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:07:57 h | *Vashon, WA | USA | 1999.0 | M | MU23 | 12.099 | 64838 | February |
| **2539948** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:22:02 h | *Gig Harbor, WA | USA | 1983.0 | M | M35 | 11.449 | 704450 | February |
| **2539949** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 20 | 4:27:34 h | *Bainbridge Island, WA | USA | 1977.0 | M | M40 | 11.212 | 810281 | February |

In [233…
```
## Delete columns unused
df_event_USA.drop(columns=['Athlete age category', 'Athlete club' , 'Event number of finishers', 'Athlete ID'], inplace=True)
```

In [234…
```
df_event_USA.head(5)
```

Out[234…

| | Year of event | Event dates | Event name | Event distance/length | Athlete performance | Athlete country | Athlete year of birth | Athlete gender | Athlete average speed | month |
|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 3:17:55 h | USA | 1991.0 | M | 15.158 | February |
| **2539946** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4:02:32 h | USA | 1981.0 | M | 12.369 | February |
| **2539947** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4:07:57 h | USA | 1999.0 | M | 12.099 | February |
| **2539948** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4:22:02 h | USA | 1983.0 | M | 11.449 | February |
| **2539949** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4:27:34 h | USA | 1977.0 | M | 11.212 | February |

In [235…
```
#Check NA values
#Delete NA values
#Filter NA values
```

In [236…
```
df_event_USA.isna().sum()
```

Out[236…
```
Year of event              0
Event dates             1304
Event name                 0
Event distance/length      0
Athlete performance        0
Athlete country            0
Athlete year of birth    235
Athlete gender             0
Athlete average speed      0
month                   1304
dtype: int64
```

In [237…
```
df_event_USA.shape
```

Out[237…
```
(26524, 10)
```

In [238…
```
#Drop only N/A value from Athlete year of birth column
df_event_USA.dropna(axis=0, how='any', inplace=True)
```

In [239…
```
df_event_USA.isna().sum()
```

Out[239…
```
Year of event           0
Event dates             0
Event name              0
Event distance/length   0
Athlete performance     0
Athlete country         0
Athlete year of birth   0
Athlete gender          0
Athlete average speed   0
month                   0
dtype: int64
```

In [240…
```
df_event_USA.shape
```

Out[240…
```
(24987, 10)
```

10. Change type data

11. Rearange order of columns for new data set

12. Rename column 'name' to 'event_name' because of mistaken

13. Drop column 'year_of_birth' because we already had the name column 'athlete_age

In [241…
```
#Change type data
df_event_USA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 24987 entries, 2539945 to 2760961
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Year of event         24987 non-null  int64
 1   Event dates           24987 non-null  datetime64[ns]
 2   Event name            24987 non-null  object
 3   Event distance/length 24987 non-null  object
 4   Athlete performance   24987 non-null  object
 5   Athlete country       24987 non-null  object
 6   Athlete year of birth 24987 non-null  float64
 7   Athlete gender        24987 non-null  object
 8   Athlete average speed 24987 non-null  object
 9   month                 24987 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(7)
memory usage: 2.1+ MB
```

In [242…
```
#Change type of Year event
df_event_USA['Year of event'] = pd.to_datetime(df_event_USA['Year of event'], format='%Y', errors='coerce').dt.year
```

In [243…
```
#Split h from Athlete performance
df_event_USA['Athlete performance'] = df_event_USA['Athlete performance'].str.split('h').str.get(0)
```

In [244…
```
df_event_USA['Athlete performance'] = pd.to_timedelta(df_event_USA['Athlete performance'])
```

In [245…
```
df_event_USA['Athlete year of birth'] = df_event_USA['Athlete year of birth'].astype(int)
```

In [246…
```
df_event_USA['Athlete average speed'] = df_event_USA['Athlete average speed'].astype(float)
```

In [247…
```
df_event_USA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 24987 entries, 2539945 to 2760961
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Year of event          24987 non-null  int32
 1   Event dates            24987 non-null  datetime64[ns]
 2   Event name             24987 non-null  object
 3   Event distance/length  24987 non-null  object
 4   Athlete performance    24987 non-null  timedelta64[ns]
 5   Athlete country        24987 non-null  object
 6   Athlete year of birth  24987 non-null  int64
 7   Athlete gender         24987 non-null  object
 8   Athlete average speed  24987 non-null  float64
 9   month                  24987 non-null  object
dtypes: datetime64[ns](1), float64(1), int32(1), int64(1), object(5), timedelta64[ns](1)
memory usage: 2.0+ MB
```

In [248…
```
#Rearange order of columns for new data set
df_event_USA.rename(columns={
    'Year of event' : 'Year_of_event',
    'Event dates' : 'Event_dates',
    'Event name' : 'Event_name',
    'Athlete performance' : 'Athlete_performance',
    'Athlete year of birth' : 'Athlete_year_of_birth',
    'Athlete average speed' : 'Athlete_average_speed',
    'Athlete ID' : 'Athlete_id',
    'Event distance/length' : 'Event_distance_length',
    'Athlete country' : 'Athlete_country',
    'Athlete gender' : 'Athlete_gender',
}), inplace=True)
```

In [249…
```
df_event_USA.head(4)
```

Out[249…

| | Year_of_event | Event_dates | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_year_of_birth | Athlete_gender | Athlete_average_speed | month |
|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 0 days 03:17:55 | USA | 1991 | M | 15.158 | February |
| **2539946** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 0 days 04:02:32 | USA | 1981 | M | 12.369 | February |
| **2539947** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 0 days 04:07:57 | USA | 1999 | M | 12.099 | February |
| **2539948** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 0 days 04:22:02 | USA | 1983 | M | 11.449 | February |

In [250…
```
#Change type of Athlete_performance to time hours
df_event_USA['Athlete_performance'] = df_event_USA['Athlete_performance'].dt.total_seconds() / 3600
```

In [251…
```
df_event_USA.head(3)
```

Out[251…

| | Year_of_event | Event_dates | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_year_of_birth | Athlete_gender | Athlete_average_speed | month |
|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 3.298611 | USA | 1991 | M | 15.158 | February |
| **2539946** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4.042222 | USA | 1981 | M | 12.369 | February |
| **2539947** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4.132500 | USA | 1999 | M | 12.099 | February |

In [252…
```
#Get athlete age
df_event_USA['Athlete_age'] = df_event_USA['Year_of_event'] - df_event_USA['Athlete_year_of_birth']
df_event_USA.head(4)
```

Out[252…

| | Year_of_event | Event_dates | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_year_of_birth | Athlete_gender | Athlete_average_speed | month | Athlete_age |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 3.298611 | USA | 1991 | M | 15.158 | February | 29 |
| **2539946** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4.042222 | USA | 1981 | M | 12.369 | February | 39 |
| **2539947** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4.132500 | USA | 1999 | M | 12.099 | February | 21 |
| **2539948** | 2020 | 2020-02-02 | West Seattle Beach Run - Winter Edition | 50km | 4.367222 | USA | 1983 | M | 11.449 | February | 37 |

In [253…
```
#Drop event date , Athlete year of birth
df_event_USA.drop(columns=(['Event_dates','Athlete_year_of_birth']),inplace=True, axis=1)
```

In [254…
```
df_event_USA.rename(columns={'month' : 'Event_month'}), inplace=True)
```

In [255…
```
df_event_USA['Event_month'] = pd.to_datetime(df_event_USA['Event_month'], format = '%B').dt.month
```

In [256…
```
df_event_USA.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 24987 entries, 2539945 to 2760961
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Year_of_event          24987 non-null  int32
 1   Event_name             24987 non-null  object
 2   Event_distance_length  24987 non-null  object
 3   Athlete_performance    24987 non-null  float64
 4   Athlete_country        24987 non-null  object
 5   Athlete_gender         24987 non-null  object
 6   Athlete_average_speed  24987 non-null  float64
 7   Event_month            24987 non-null  int32
 8   Athlete_age            24987 non-null  int64
dtypes: float64(2), int32(2), int64(1), object(4)
memory usage: 1.7+ MB
```

In [257…
```
df_event_USA.head(4)
```

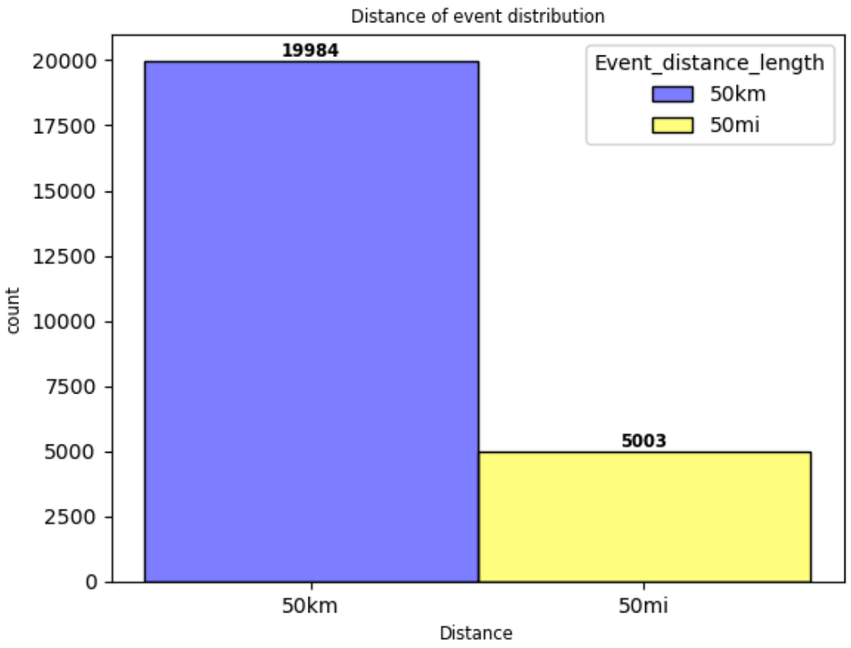
Out[257…

| | Year_of_event | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_gender | Athlete_average_speed | Event_month | Athlete_age |
|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 3.298611 | USA | M | 15.158 | 2 | 29 |
| **2539946** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.042222 | USA | M | 12.369 | 2 | 39 |
| **2539947** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.132500 | USA | M | 12.099 | 2 | 21 |
| **2539948** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.367222 | USA | M | 11.449 | 2 | 37 |

In [258…
```
#Vusualization with history chart bar
from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
```

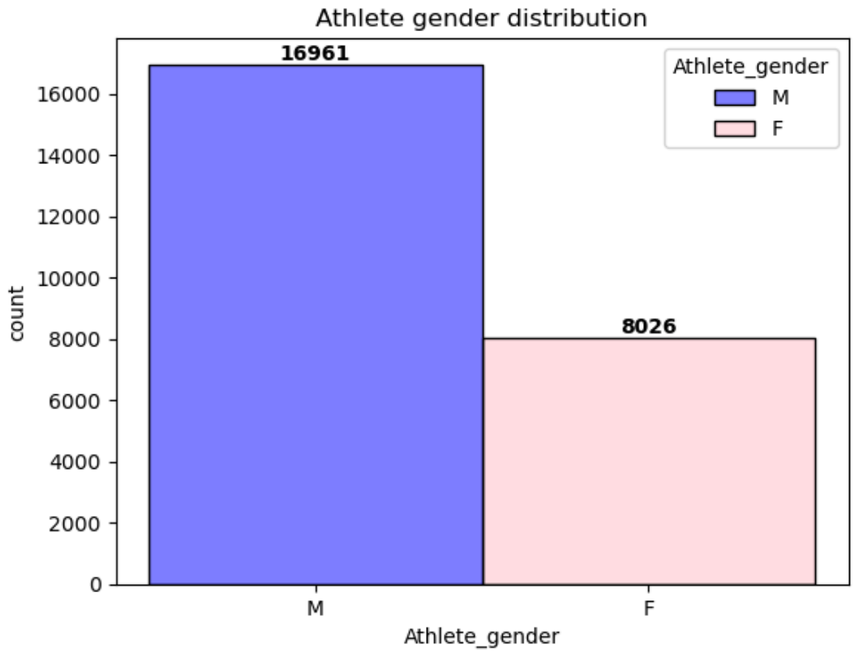
In [259…
```
##Vusualization with history chart bar Show chart to compare event distance length
sns.histplot(df_event_USA, x = 'Event_distance_length', hue='Event_distance_length', palette={'50km': 'blue', '50mi': 'yellow'}), bins=20)
distance_count = df_event_USA['Event_distance_length'].value_counts()
```

```
for distance, count in distance_count.items():
    plt.text(distance, count+0.1, str(count), ha='center', va='bottom', fontsize='small',fontweight='bold')
plt.xlabel('Distance', fontsize='small')
plt.ylabel('count', fontsize='small')
plt.title('Distance of event distribution', fontsize='small')
plt.show()
```

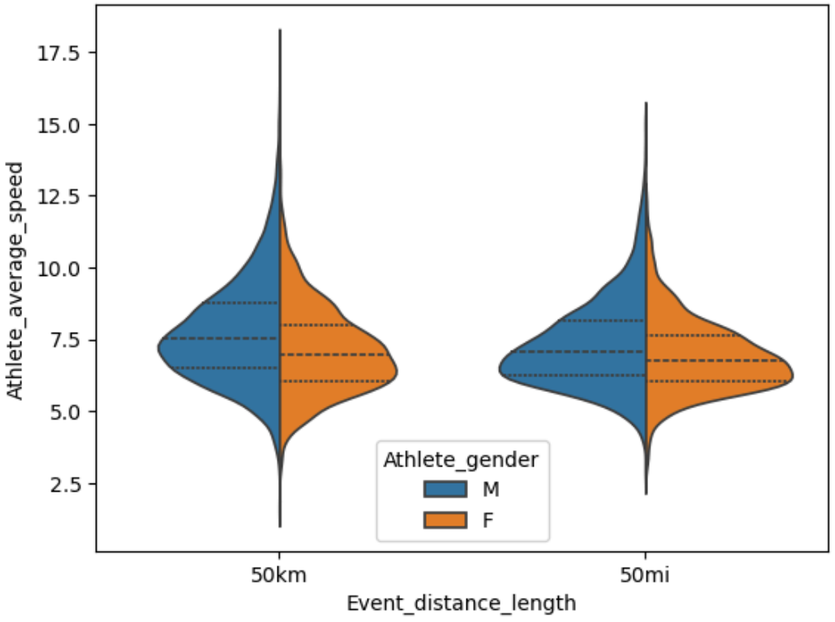



In [260…

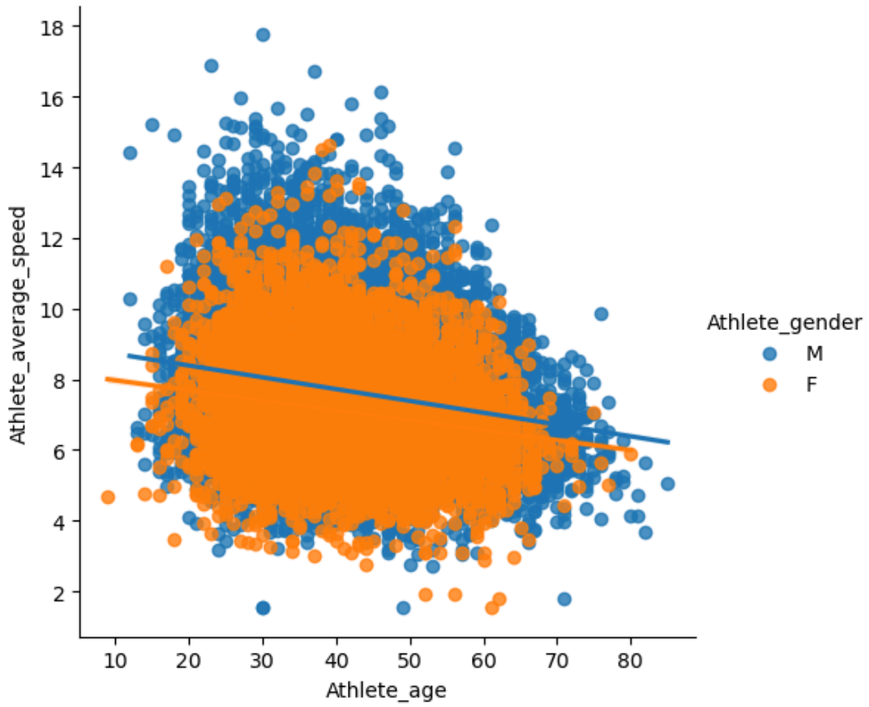
```
# History chart with gender
sns.histplot(df_event_USA, x='Athlete_gender', hue='Athlete_gender', palette={'M': 'blue', 'F': 'pink'}), bins=20 )
gender_count = df_event_USA['Athlete_gender'].value_counts()
for gender, count in gender_count.items():
    plt.text(gender, count+0.1, str(count), ha='center', va='bottom', fontweight='bold')
plt.xlabel('Athlete_gender')
plt.ylabel('count')
plt.title('Athlete gender distribution')
plt.show()
```



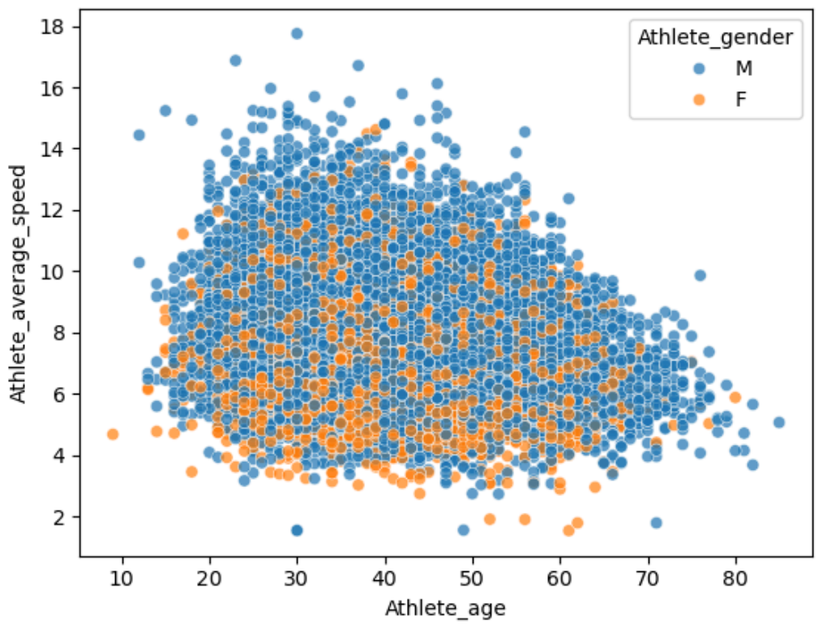

In [261…
```
#Visualization with violin chart from distance length in gender and average speed
sns.violinplot(df_event_USA, x='Event_distance_length', y='Athlete_average_speed',hue='Athlete_gender', split=True, inner='quart')
plt.title('Compare the distribution of gender in average speed in each kind of distance', fontsize='small')
plt.show()
```




In [262…
```
#Visualyze lm chart
ax = sns.lmplot(df_event_USA, x ='Athlete_age', y='Athlete_average_speed', hue = 'Athlete_gender')
ax
plt.show()
```




In [263…
```
#Try to compare with scatter plot for more clarify
sns.scatterplot(df_event_USA, x ='Athlete_age', y ='Athlete_average_speed', hue = 'Athlete_gender', alpha = 0.7)
plt.show()
```




In [264…
```
# Set up the period time for specify season column
season = ({
'Spring' : 1-3,
'Summer' : 4-6,
'Fall' : 7-9,
'Winter' : 10-12,
})
```

In [266…
```
#Create season name corresponding with Event_month
df_event_USA['season'] = df_event_USA['Event_month'].apply(lambda x: 'winter' if x < 1 else 'spring' if x < 4  else 'summer' if x < 7 else 'fall' if x < 10 else 'winter')
```

In [270…
```
df_event_USA.head(4)
```

Out[270…

| | Year_of_event | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_gender | Athlete_average_speed | Event_month | Athlete_age | season |
|---|---|---|---|---|---|---|---|---|---|---|
| **2539945** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 3.298611 | USA | M | 15.158 | 2 | 29 | spring |
| **2539946** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.042222 | USA | M | 12.369 | 2 | 39 | spring |
| **2539947** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.132500 | USA | M | 12.099 | 2 | 21 | spring |
| **2539948** | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.367222 | USA | M | 11.449 | 2 | 37 | spring |

In [272…
```
#Reindex data sex
df_event_USA = df_event_USA.reset_index()
```

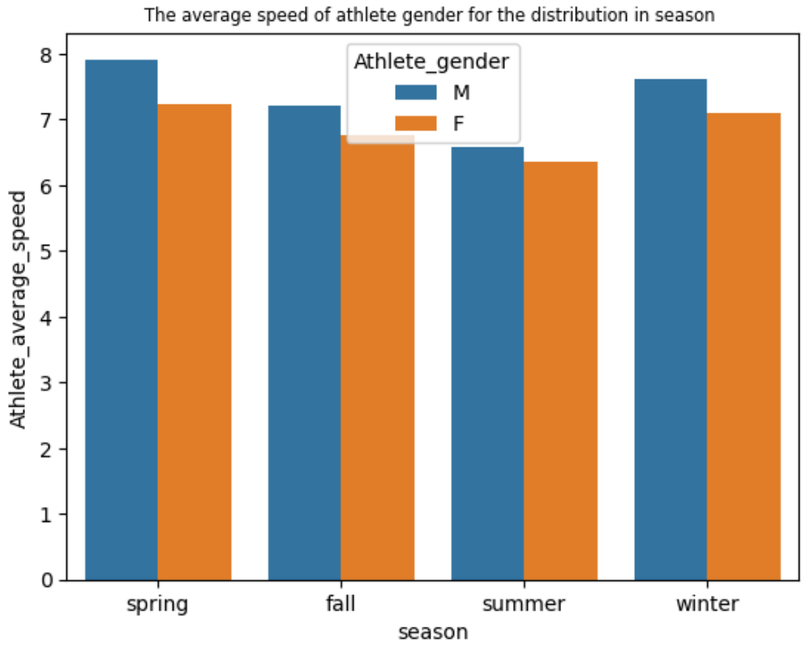
In [273…
```
df_event_USA.head(10)
```

Out[273…

| | index | Year_of_event | Event_name | Event_distance_length | Athlete_performance | Athlete_country | Athlete_gender | Athlete_average_speed | Event_month | Athlete_age | season |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2539945 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 3.298611 | USA | M | 15.158 | 2 | 29 | spring |
| **1** | 2539946 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.042222 | USA | M | 12.369 | 2 | 39 | spring |
| **2** | 2539947 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.132500 | USA | M | 12.099 | 2 | 21 | spring |
| **3** | 2539948 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.367222 | USA | M | 11.449 | 2 | 37 | spring |
| **4** | 2539949 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.459444 | USA | M | 11.212 | 2 | 43 | spring |
| **5** | 2539950 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.701667 | USA | F | 10.635 | 2 | 35 | spring |
| **6** | 2539951 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.822222 | USA | M | 10.369 | 2 | 59 | spring |
| **7** | 2539952 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.830556 | USA | M | 10.351 | 2 | 50 | spring |
| **8** | 2539953 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 4.850000 | USA | F | 10.309 | 2 | 45 | spring |
| **9** | 2539954 | 2020 | West Seattle Beach Run - Winter Edition | 50km | 5.043056 | USA | M | 9.915 | 2 | 41 | spring |

In [281…
```
average_speed = df_event_USA.groupby('Athlete_gender')['Athlete_average_speed'].mean()
```

In [284…
```
sns.barplot(df_event_USA, x='season', y='Athlete_average_speed', errorbar= None, hue='Athlete_gender')
plt.title('The average speed of athlete gender for the distribution in season ', fontsize='small')
plt.show()
```




In [ ]:
```
#Define top 5 country with highest average speed
```

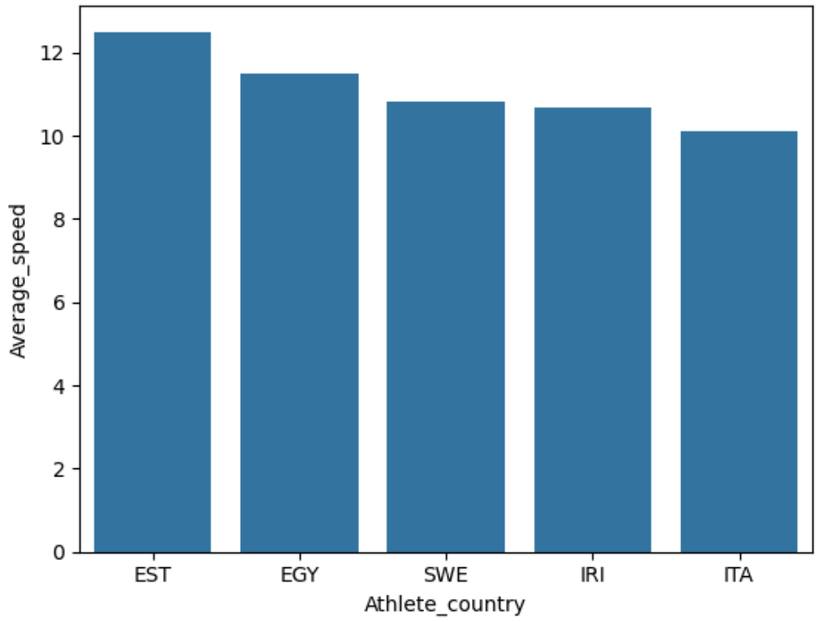
In [295…
```
get_average_top_country = df_event_USA.groupby('Athlete_country')['Athlete_average_speed'].mean().sort_values(ascending=False).head(5)
get_average_top_country = get_average_top_country.reset_index()
get_average_top_country.columns = ['Athlete_country', 'Average_speed']
get_average_top_country
```

Out[295…

| | Athlete_country | Average_speed |
|---|---|---|
| **0** | EST | 12.511000 |
| **1** | EGY | 11.510500 |
| **2** | SWE | 10.812750 |
| **3** | IRI | 10.690000 |
| **4** | ITA | 10.125687 |

In [297…
```
sns.barplot(get_average_top_country, x ='Athlete_country', y='Average_speed')

plt.show()
```


12
10
8
6
4
2
0
Average_speed
EST
EGY
SWE
IRI
ITA
Athlete_country

In [ ]:

In [ ]: