In [1]:
```
import pandas as pd
```

In [3]:
```
import seaborn as sns
```

In [5]:
```
df = pd.read_csv("marathon.csv")
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\2664868931.py:1: DtypeWarning: Columns (11) have mixed types. Specify dtype option on import or set low_memory=False.
  df = pd.read_csv("marathon.csv")
```

In [7]:
```
df.head(10)
```

Out[7]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	Athlete country
0	2018	06.01.2018	Selva Costera (CHI)	50km	22	4:51:39 h	Tnfrc	CHI
1	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:15:45 h	Roberto Echeverría	CHI
2	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:16:44 h	Puro Trail Osorno	CHI
3	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:34:13 h	Columbia	ARG
4	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:54:14 h	Baguales Trail	CHI
5	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:25:01 h	NaN	ARG
6	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:28:00 h	Los Patagones	ARG
7	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:32:24 h	Reaktiva Chile	CHI
8	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:39:08 h	Puro Trail Osorno	CHI
9	2018	06.01.2018	Selva Costera (CHI)	50km	22	6:45:11 h	Marlene Flores Team	CHI

In [9]:
```
df.shape
```

Out[9]:
```
(7461195, 13)
```

In [11]:
```
df.dtypes
```

Out[11]:
```
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
```

In [13]:
```
# clean up data
#only want USA races, 50k or 50mi, 2020
```

In [15]:
```
#step 1 show 50mi,or 50km
```

In [19]:
```
df[df['Event distance/length'] == '50mi']
```

Out[19]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
55	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	9:53:05 h	*Middleville, MI
56	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:09:35 h	*Waterloo, ON
57	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:33:00 h	*Kitchener, ON
58	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:38:17 h	*Utica, MI
59	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:56:35 h	*Grass Lake, MI
...	...	...	...	...	...	...	...
7461181	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	11:59:37 h	NaN
7461182	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:01:41 h	NaN
7461183	1995	07.01.1995	Avalon Benefit 50-Mile	50mi	92	12:03:26 h	NaN

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
			Run (USA)				
7461184	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:03:26 h	NaN
7461185	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:05:59 h	NaN

352181 rows × 13 columns

In [21]:
```
# combine 50km and 50mi with isin
```

In [23]:
```
df[df['Event distance/length'].isin(['50km', '50mi'])]
```

Out[23]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	A cc
0	2018	06.01.2018	Selva Costera (CHI)	50km	22	4:51:39 h	Tnfrc	
1	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:15:45 h	Roberto Echeverría	
2	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:16:44 h	Puro Trail Osorno	
3	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:34:13 h	Columbia	
4	2018	06.01.2018	Selva Costera (CHI)	50km	22	5:54:14 h	Baguales Trail	
...	...	...	...	...	...	...	...	
7461181	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	11:59:37 h	NaN	
7461182	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:01:41 h	NaN	
7461183	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:03:26 h	NaN	
7461184	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:03:26 h	NaN	
7461185	1995	07.01.1995	Avalon Benefit 50-	50mi	92	12:05:59 h	NaN	

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	A cc
			Mile Run (USA)					

1874790 rows × 13 columns

In [25]:
```
#show race only of the year 2020
```

In [27]:
```
df[(df['Event distance/length'].isin(['50km', '50mi'])) & (df['Year of event']== 20
```

Out[27]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete
2538571	2020	07.-09.02.2020	Taipei 48hr Ultra Marathon - 50mi (TPE)	50mi	38	7:34:19 h	E
2538572	2020	07.-09.02.2020	Taipei 48hr Ultra Marathon - 50mi (TPE)	50mi	38	7:43:50 h	
2538573	2020	07.-09.02.2020	Taipei 48hr Ultra Marathon - 50mi (TPE)	50mi	38	8:04:40 h	
2538574	2020	07.-09.02.2020	Taipei 48hr Ultra Marathon - 50mi (TPE)	50mi	38	8:30:49 h	台灣大腸 跑
2538575	2020	07.-09.02.2020	Taipei 48hr Ultra Marathon - 50mi (TPE)	50mi	38	8:34:47 h	
...	...	...	...	...	...	...	
2762404	2020	03.10.2020	Bison Ultra-Trail 50 (POL)	50km	271	7:36:25 h	AKS Po War
2762405	2020	03.10.2020	Bison Ultra-Trail 50 (POL)	50km	271	7:36:27 h	*War
2762406	2020	03.10.2020	Bison Ultra-Trail 50 (POL)	50km	271	7:44:18 h	Ou Tra
2762407	2020	03.10.2020	Bison Ultra-Trail 50 (POL)	50km	271	8:04:50 h	PH By G

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete
2762408	2020	03.10.2020	Bison Ultra-Trail 50 (POL)	50km	271	8:11:43 h	* Aleksanc

63489 rows × 13 columns

In [46]:
```
df[df['Event name'].str.split('(').str.get(1).str.split(')').str.get(0)== "USA"]
```

Out[46]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
55	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	9:53:05 h	*Middleville, MI
56	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:09:35 h	*Waterloo, ON
57	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:33:00 h	*Kitchener, ON
58	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:38:17 h	*Utica, MI
59	2018	06.01.2018	Yankee Springs 50 Mile Winter Challenge (USA)	50mi	9	11:56:35 h	*Grass Lake, MI
...	...	...	...	...	...	...	...
7461181	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	11:59:37 h	NaN
7461182	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:01:41 h	NaN
7461183	1995	07.01.1995	Avalon Benefit 50-Mile	50mi	92	12:03:26 h	NaN

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
			Run (USA)				
7461184	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:03:26 h	NaN
7461185	1995	07.01.1995	Avalon Benefit 50-Mile Run (USA)	50mi	92	12:05:59 h	NaN

1398540 rows × 13 columns

In [48]:
```
# combine all filters together.
```

In [52]:
```
df[(df['Event distance/length'].isin(['50km', '50mi'])) & (df['Year of event']== 20
```

Out[52]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	3:17:55 h	*Normandy Park, WA
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:02:32 h	*Gold Bar, WA
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:07:57 h	*Vashon, WA
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:22:02 h	*Gig Harbor, WA
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:27:34 h	*Bainbridge Island, WA
...	...	...	...	...	...	...	...
2760957	2020	03.10.2020	Yankee Springs Fall Trail Run Festival (USA)	50km	30	7:07:48 h	*East Lansing, MI
2760958	2020	03.10.2020	Yankee Springs	50km	30	7:27:22 h	*Traverse City, MI

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	c
			Fall Trail Run Festival (USA)					
2760959	2020	03.10.2020	Yankee Springs Fall Trail Run Festival (USA)	50km	30	7:27:24 h	*Traverse City, MI	
2760960	2020	03.10.2020	Yankee Springs Fall Trail Run Festival (USA)	50km	30	7:38:30 h	*Mason, MI	
2760961	2020	03.10.2020	Yankee Springs Fall Trail Run Festival (USA)	50km	30	7:59:53 h	NaN	

26090 rows × 13 columns

In [54]:
```
df2 = df[(df['Event distance/length'].isin(['50km', '50mi'])) & (df['Year of event'
```

In [58]:
```
df2.head()
```

Out[58]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	A c
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	3:17:55 h	*Normandy Park, WA	
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:02:32 h	*Gold Bar, WA	
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:07:57 h	*Vashon, WA	
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:22:02 h	*Gig Harbor, WA	
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition (USA)	50km	20	4:27:34 h	*Bainbridge Island, WA	

In [60]:
```
# remove usa from events names
```

In [62]:
```
df2['Event name'].str.split('(').str.get(0)
```

Out[62]:
```
2539945    West Seattle Beach Run - Winter Edition
2539946    West Seattle Beach Run - Winter Edition
2539947    West Seattle Beach Run - Winter Edition
2539948    West Seattle Beach Run - Winter Edition
2539949    West Seattle Beach Run - Winter Edition
                          ...
2760957     Yankee Springs Fall Trail Run Festival
2760958     Yankee Springs Fall Trail Run Festival
2760959     Yankee Springs Fall Trail Run Festival
2760960     Yankee Springs Fall Trail Run Festival
2760961     Yankee Springs Fall Trail Run Festival
Name: Event name, Length: 26090, dtype: object
```

In [66]:
```
df2['Event name'] = df2['Event name'].str.split('(').str.get(0)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\3473829760.py:1: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['Event name'] = df2['Event name'].str.split('(').str.get(0)
```

df2.head()

In [69]:
```
df2.head()
```

Out[69]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	A c
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55 h	*Normandy Park, WA	
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32 h	*Gold Bar, WA	
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57 h	*Vashon, WA	
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02 h	*Gig Harbor, WA	
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34 h	*Bainbridge Island, WA	

In [71]:
```
# clean up athlete age by subtracting it to
```

In [79]:
```
df2['athlete_age']= 2020 - df2['Athlete year of birth']
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\1405298268.py:1: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['athlete_age']= 2020 - df2['Athlete year of birth']
```

df2.head()

In [81]:
```
#remove h in the athlete performance
```

In [89]:
```
df2['Athlete performance'] = df2['Athlete performance'].str.split(' ').str.get(0)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\2477507555.py:1: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['Athlete performance'] = df2['Athlete performance'].str.split(' ').str.get(0)
```

In [87]:
```
df2
```

Out[87]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55 h	*Normandy Park, WA
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32 h	*Gold Bar, WA
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57 h	*Vashon, WA
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02 h	*Gig Harbor, WA
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34 h	*Bainbridge Island, WA
...	...	...	...	...	...	...	...
2760957	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:07:48 h	*East Lansing, MI
2760958	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:27:22 h	*Traverse City, MI

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	c
2760959	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:27:24 h	*Traverse City, MI	
2760960	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:38:30 h	*Mason, MI	
2760961	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:59:53 h	NaN	

26090 rows × 14 columns

In [91]:
```
df2.head()
```

Out[91]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete club	A c
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55	*Normandy Park, WA	
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32	*Gold Bar, WA	
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57	*Vashon, WA	
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02	*Gig Harbor, WA	
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34	*Bainbridge Island, WA	

In [93]:
```
#drop columns: athlete club,athlete country,athlete year of birth, athlete age cate
```

In [97]:
```
df2 = df2.drop(['Athlete club', 'Athlete country', 'Athlete year of birth', 'Athlet
```

In [99]:
```
df2.head()
```

Out[99]:

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Athl avera spe
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55	M	15.
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32	M	12.
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57	M	12.
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02	M	11.
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34	M	11.

In [101…

```
df2
```

Out[101…

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Ath aver sp
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55	M	15.
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32	M	12.
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57	M	12.
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02	M	11.
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34	M	11.
...	...	...	...	...	...	...	...	
2760957	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:07:48	F	7.
2760958	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:27:22	F	6.
2760959	2020	03.10.2020	Yankee Springs	50km	30	7:27:24	F	6.

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Ath aver sp
			Fall Trail Run Festival					
2760960	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:38:30	F	6.
2760961	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:59:53	M	6.

26090 rows × 10 columns

In [103…
```
# clean up all null values:
```

In [105…
```
df2.isna().sum()
```

Out[105…
```
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
athlete_age                233
dtype: int64
```

In [107…
```
df2= df2.dropna()
```

In [109…
```
df2.shape
```

Out[109…
```
(25857, 10)
```

In [111…
```
df2.isna().sum()
```

Out[111…
```
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
athlete_age                 0
dtype: int64
```

In [113…
```
#check for duplicates
df2[df2.duplicated()== True]
```

In [115…
```
#reset_index
```

In [117…
```
df2.reset_index(drop = True)
```

Out[117...

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Athlet averag spee
0	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55	M	15.15
1	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32	M	12.36
2	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57	M	12.09
3	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02	M	11.44
4	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34	M	11.21
...	...	...	...	...	...	...	...	
25852	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:07:48	F	7.01
25853	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:27:22	F	6.70
25854	2020	03.10.2020	Yankee Springs	50km	30	7:27:24	F	6.70

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Athlet averag spee
			Fall Trail Run Festival					
25855	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:38:30	F	6.54
25856	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:59:53	M	6.25

25857 rows × 10 columns

In [119…
```
#fix index type
```

In [121…
```
df2['athlete_age'] = df2['athlete_age'].as(int)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\130840868.py:1: SettingWithCopyWarni
ng:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['athlete_age']= df2['athlete_age'].astype(int)
```

In [123…
```
df2['athlete_age'] = df2['athlete_age'].astype(int)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\2319086533.py:1: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['athlete_age'] = df2['athlete_age'].astype(int)
```

In [129…
```
df2
```

Out[129...

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Ath aver sp
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	3:17:55	M	15.
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:02:32	M	12.
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:07:57	M	12.
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:22:02	M	11.
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	4:27:34	M	11.
...	...	...	...	...	...	...	...	
2760957	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:07:48	F	7.
2760958	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:27:22	F	6.
2760959	2020	03.10.2020	Yankee Springs	50km	30	7:27:24	F	6.

	Year of event	Event dates	Event name	Event distance/length	Event number of finishers	Athlete performance	Athlete gender	Ath aver sp
			Fall Trail Run Festival					
2760960	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:38:30	F	6.
2760961	2020	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	7:59:53	M	6.

25857 rows × 10 columns

In [131…
```
df2['athlete_age'] = df2['athlete_age'].astype(int)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\2319086533.py:1: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['athlete_age'] = df2['athlete_age'].astype(int)
```

In [133…
```
df2.dtypes
```

Out[133…
```
Year of event                int64
Event dates                 object
Event name                  object
Event distance/length       object
Event number of finishers    int64
Athlete performance         object
Athlete gender              object
Athlete average speed       object
Athlete ID                   int64
athlete_age                  int32
dtype: object
```

In [135…
```
df2['Athlete average speed'] = df2['Athlete average speed'].astype(float)
```

```
C:\Users\User\AppData\Local\Temp\ipykernel_4424\501852820.py:1: SettingWithCopyWarni
ng:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  df2['Athlete average speed'] = df2['Athlete average speed'].astype(float)
```

In [137…
```
df2.dtypes
```

Out[137…
```
Year of event                 int64
Event dates                  object
Event name                   object
Event distance/length        object
Event number of finishers     int64
Athlete performance          object
Athlete gender               object
Athlete average speed       float64
Athlete ID                    int64
athlete_age                   int32
dtype: object
```

In [141…
```
#rename columns
```

In [143…
```
df2= df2.rename(columns = {'Year of event':'year_of_race' ,
                'Event dates':'race_day' ,
                'Event name':'race_name' ,
                'Event distance/length':'race_distance' ,
                'Event number of finishers':'race_numbers_of_finishers',
                'Athlete performance':'athlete_performance' ,
                'Athlete gender':'athlete_gender' ,
                'Athlete average speed':'athlete_average_speed',
                'Athlete ID':'athlete_id',
                 })
```

In [145…
```
df2.head()
```

Out[145…

	year_of_race	race_day	race_name	race_distance	race_numbers_of_finishers	ath
2539945	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	
2539946	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	
2539947	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	
2539948	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	
2539949	2020	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	

In [147…
```
#reoder columns
```

In [191…
```
df3= df2[['race_day', 'race_name', 'race_distance', 'race_numbers_of_finishers', 'a
```

In [151…
```
df3.head()
```

Out[151…

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2539945	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	71287	
2539946	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	629508	
2539947	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	64838	
2539948	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	704450	
2539949	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	810281	

In [193…

```
df3
```

Out[193...

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2539945	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	71287	
2539946	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	629508	
2539947	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	64838	
2539948	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	704450	
2539949	02.02.2020	West Seattle Beach Run - Winter Edition	50km	20	810281	
...	...	...	...	...	...	
2760957	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	816361	
2760958	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	326469	
2760959	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	372174	
2760960	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	860349	

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2760961	03.10.2020	Yankee Springs Fall Trail Run Festival	50km	30	770097	

25857 rows × 9 columns

In [155…
```
# find the 2 races he ran in 2020
```

In [161…
```
df3[df3['race_name']== "Everglades 50 Mile Ultra Run "]
```

Out[161...

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2591476	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820757	
2591477	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	46432	
2591478	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	813617	
2591479	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820758	
2591480	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	647115	
2591481	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	696063	
2591482	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	222509	
2591483	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820759	
2591484	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	359359	
2591485	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	103020	
2591486	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	202097	
2591487	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820760	
2591488	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820761	
2591489	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820762	

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2591490	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820763	
2591491	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	39534	
2591492	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	190251	
2591493	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	202096	
2591494	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	695761	
2591495	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	696060	
2591496	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	47609	
2591497	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	347938	
2591498	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	55213	
2591499	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	393456	
2591500	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	197745	
2591501	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	248521	
2591502	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820764	
2591503	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	733944	

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2591504	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	140763	
2591505	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	47616	
2591506	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	820765	
2591507	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	311648	
2591508	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	211177	
2591509	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	369928	
2591510	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	198005	
2591511	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	695776	
2591512	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	647125	
2591513	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	12531	
2591514	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	19081	
2591515	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	34961	

In [175…
```
df3[df3['athlete_id'] == 222509]
```

Out[175…

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlete_perform

In [177…
```
df3[df3['athlete_id'] == 222509]
```

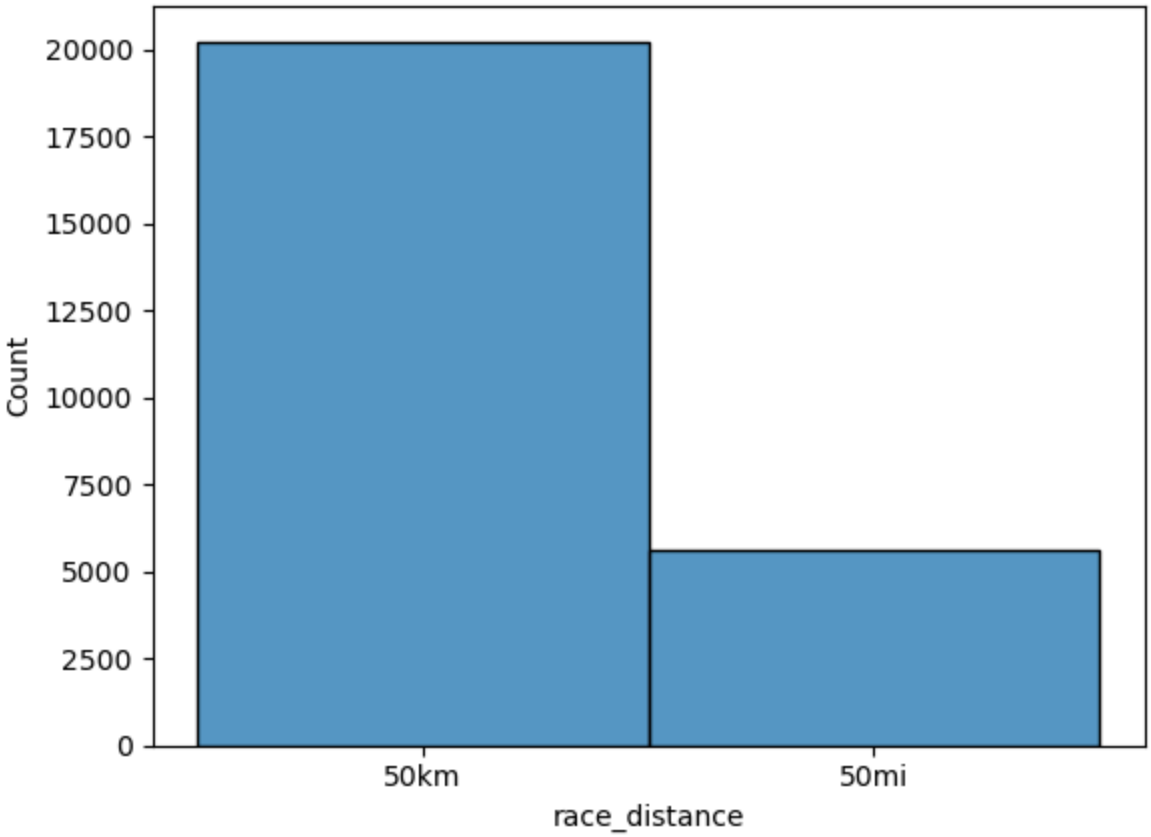
Out[177…

	race_day	race_name	race_distance	race_numbers_of_finishers	athlete_id	athlet
2591482	14.03.2020	Everglades 50 Mile Ultra Run	50mi	40	222509	
2616900	22.02.2020	Manasota Track Club 50K	50km	36	222509	

In [179…
```
# chart and graph
```

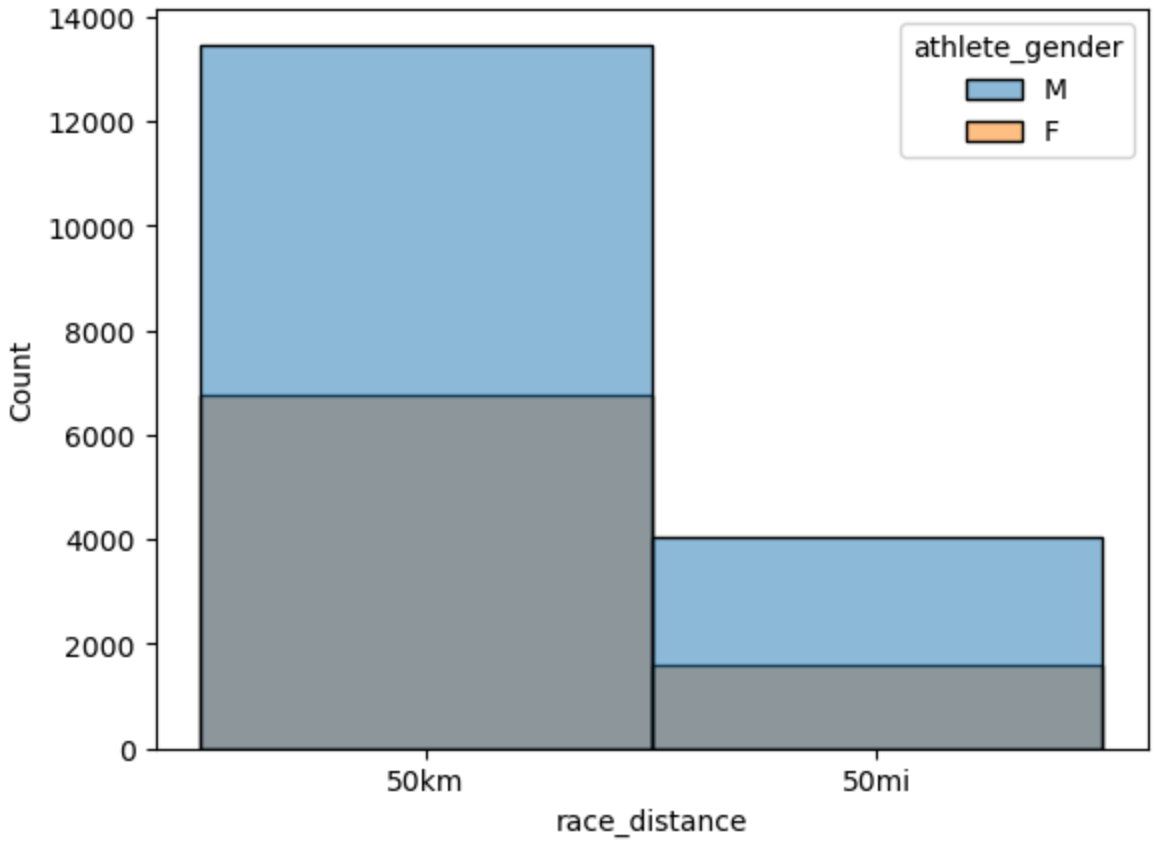
In [183…
```
sns.histplot(df3['race_distance'])
```

Out[183…
```
<Axes: xlabel='race_distance', ylabel='Count'>
```



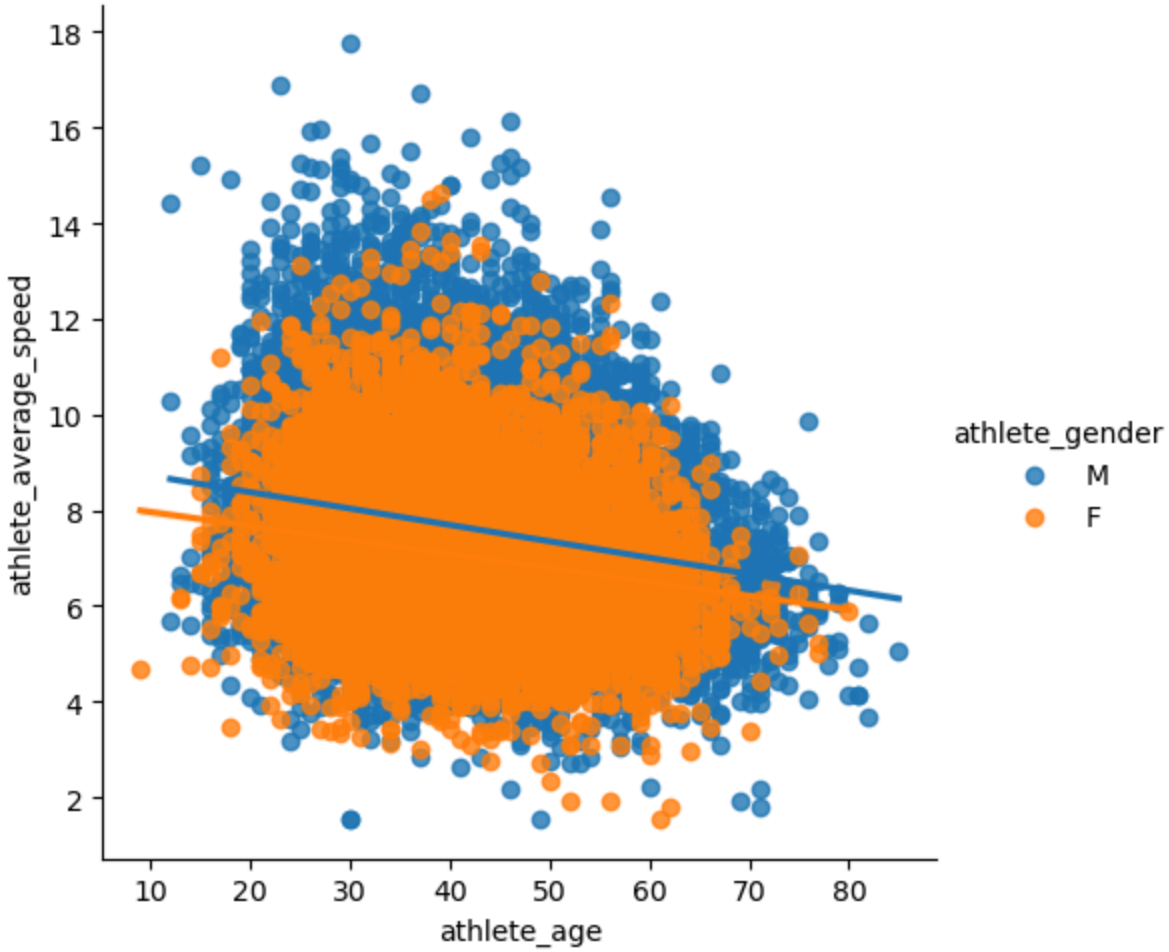
In [195…
```
sns.histplot(df3, x = 'race_distance', hue = 'athlete_gender')
```

Out[195…
```
<Axes: xlabel='race_distance', ylabel='Count'>
```



In [203…
```
sns.lmplot(df3, x ='athlete_age', y= 'athlete_average_speed', hue = 'athlete_gender
```

Out[203…
```
<seaborn.axisgrid.FacetGrid at 0x2df28112660>
```



In [205…
```
df3.dtypes
```

Out[205…
```
race_day                     object
race_name                    object
race_distance                object
race_numbers_of_finishers     int64
athlete_id                    int64
athlete_performance          object
athlete_age                   int32
athlete_average_speed       float64
athlete_gender               object
dtype: object
```

In [209…
```
df3.groupby(['race_distance', 'athlete_gender'])['athlete_average_speed'].mean()
```

Out[209…
```
race_distance  athlete_gender
50km           F                7.083011
               M                7.738985
50mi           F                6.834371
               M                7.257633
Name: athlete_average_speed, dtype: float64
```

In [ ]: