Week_2_milestone

September 29, 2021

0.1 Provide a summary of the different descriptive statistics you looked at and why.

I focused my analysis in country, gender (sex), age, sports and medals to understand better this variables that are important to prove my hypothesis.

0.1.1 Importing libraries

[1]:
```
import pandas as pd
import psycopg2 as ps4

from sqlalchemy import create_engine


import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:
```
sns.set(style = 'darkgrid')

%matplotlib inline
```

[3]:
```
cnxn_string = ("postgresql+psycopg2://{username}:{pswd}"
               "@{host}:{port}/{database}")
print(cnxn_string)
```

```
postgresql+psycopg2://{username}:{pswd}@{host}:{port}/{database}
```

[4]:
```
engine = create_engine(cnxn_string.format(
    username="postgres",
    pswd="[redacted]",
    host="localhost",
    port=5432,
    database="CapStone"))
```

loading data

[5]:
```
athlete_events = pd.read_sql_table('athlete_events',engine)
athlete_events.head()
```

[5]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	No Medal
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	No Medal
2	3	Gunnar Nielsen Aaby	M	24.0	175.0	70.0	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	No Medal
3	4	Edgar Lindenau Aabye	M	34.0	175.0	70.0	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	No Medal

[6]:
```
athlete_info = pd.read_sql_table('athlete_events',engine).loc[:,"ID":"Weight"]
athlete_info.head()
```

[6]:

	ID	Name	Sex	Age	Height	Weight
0	1	A Dijiang	M	24.0	180.0	80.0
1	2	A Lamusi	M	23.0	170.0	60.0
2	3	Gunnar Nielsen Aaby	M	24.0	175.0	70.0
3	4	Edgar Lindenau Aabye	M	34.0	175.0	70.0
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0

I already cleaned the data in Week 1 assignment

[7]:
```
#clean data:
athlete_info.isnull().sum()
```

[7]:
```
ID        0
Name      0
Sex       0
Age       0
Height    0
Weight    0
dtype: int64
```

Group by gender

[8]:
```
gender_grp = athlete_info.groupby('Sex')
```

[9]:
```
age_by_sex = gender_grp['Age'].describe().loc[:,['count','mean','min','max']].
 ↪sort_values(by = 'Sex', ascending = False)
age_by_sex
```

[9]:

Sex	count	mean	min	max
M	187544.0	26.277562	10.0	97.0
F	74098.0	23.732881	11.0	74.0

[10]:
```
gender_grp['Height'].describe().loc[:,['count','mean','min','max']].
 ↪sort_values(by = 'Sex', ascending = False)
```

[10]:

Sex	count	mean	min	max
M	187544.0	177.963694	127.0	226.0
F	74098.0	168.494939	127.0	213.0

[11]:
```
gender_grp['Height'].describe().loc[:,['count','mean','min','max']].
 ↪sort_values(by = 'Sex', ascending = False)
```

[11]:

Sex	count	mean	min	max
M	187544.0	177.963694	127.0	226.0
F	74098.0	168.494939	127.0	213.0

[12]:
```
bin_query ="""

SELECT
  *,
  CASE
    WHEN"Age"BETWEEN 10 and 15
    THEN 'group_A'
    WHEN"Age"between 15 and 20
    THEN 'group_B'
    WHEN"Age"BETWEEN 20 and 25
    THEN 'group_C'
    WHEN"Age"BETWEEN 25 and 30
    THEN 'group_D'
    WHEN"Age"BETWEEN 30 and 35
    THEN 'group_E'
    WHEN"Age"BETWEEN 35 and 40
    THEN 'group_F'
    WHEN"Age"BETWEEN 40 and 50
    THEN 'group_G'
    WHEN"Age"BETWEEN 50 and 60
    THEN 'group_H'
    WHEN"Age"between 60 and 70
```

```
    THEN 'group_I'
    when "Age"between 70 and 80
    then 'group_J'
    when "Age"between 80 and 90
    then 'group_K'
    when "Age"between 90 and 100
    then 'group_L'
END as group_age
from athlete_events
"""
```

Binning the age:

[13]:
```
grouping_age = pd.read_sql_query(bin_query,engine)
grouping_age.head()
```

[13]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	group_age
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	No Medal	group_C
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	No Medal	group_C
2	3	Gunnar Nielsen Aaby	M	24.0	175.0	70.0	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	No Medal	group_C
3	4	Edgar Lindenau Aabye	M	34.0	175.0	70.0	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	group_E
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	No Medal	group_C

[14]:
```
df = pd.DataFrame(grouping_age['group_age'].value_counts())
age_group = df.sort_index()
```

[15]:
```
age_group
```

[15]:

	group_age
group_A	3280
group_B	44281
group_C	103280
group_D	68694
group_E	25723

```
group_F       8683
group_G       5763
group_H       1368
group_I        469
group_J         93
group_K          6
group_L          2
```

[16]:
```
Countries_query = """
SELECT
    nr."region",
    count(distinct ae."ID") as "number of players"
from
    athlete_events as ae
left join
    noc_regions as nr
on
    ae."NOC" = nr."NOC"
group by 1
order by 2 desc
limit 10

"""
```

[17]:
```
countries_with_most_players = pd.read_sql_query(Countries_query,engine)
countries_with_most_players
```

[17]:

	region	number of players
0	USA	9499
1	Germany	7475
2	UK	5789
3	Russia	5462
4	France	5200
5	Canada	4674
6	Italy	4654
7	Japan	3965
8	Australia	3800
9	Sweden	3784

[18]:
```
event_count = """
select
    "Sport",
    count(distinct "Event") as "Number of Event"
from
    athlete_events
group by 1
order by 2 desc
```

```
"""
```

[19]:
```
sport_by_event = pd.read_sql_query(event_count,engine)
sport_by_event
```

[19]:

	Sport	Number of Event
0	Athletics	83
1	Shooting	83
2	Swimming	54
3	Cycling	44
4	Sailing	37
..	…	…
61	Rugby	1
62	Tug-Of-War	1
63	Roque	1
64	Aeronautics	1
65	Softball	1

[66 rows x 2 columns]

[20]:
```
medals = """
select
    ng.region as REGION,
    ae."Medal" as Medal,
    count("ID") as number_of_medals
from
    athlete_events as ae
Left Outer Join
    noc_regions ng
on ae."NOC" = ng."NOC"
WHERE "Medal" != 'No Medal'
group by 1 ,2
order by 3 desc;
"""
```

[21]:
```
medal_by_country = pd.read_sql_query(medals,engine)
medal_by_country.head()
```

[21]:

	region	medal	number_of_medals
0	USA	Gold	2627
1	USA	Silver	1619
2	Russia	Gold	1599
3	USA	Bronze	1346
4	Germany	Gold	1293

[22]:
```
players = """
select
```

```
    distinct ae."ID",
    ae."Name",
    ng.region as REGION,
    count(ae."Medal") as number_of_medals
from
    athlete_events as ae
Left Outer Join
    noc_regions ng
on ae."NOC" = ng."NOC"
WHERE "Medal" != 'No Medal'
group by 1,2,3
having count(ae."Medal") > 1
order by 4 desc;
"""
```

[23]:
```
athlete_medal = pd.read_sql_query(players,engine)
athlete_medal.head()
```

[23]:

	ID	Name	region	number_of_medals
0	94406	Michael Fred Phelps, II	USA	28
1	67046	Larysa Semenivna Latynina (Diriy-)	Russia	18
2	4198	Nikolay Yefimovich Andrianov	Russia	15
3	11951	Ole Einar Bjrndalen	Norway	13
4	74420	Edoardo Mangiarotti	Italy	13

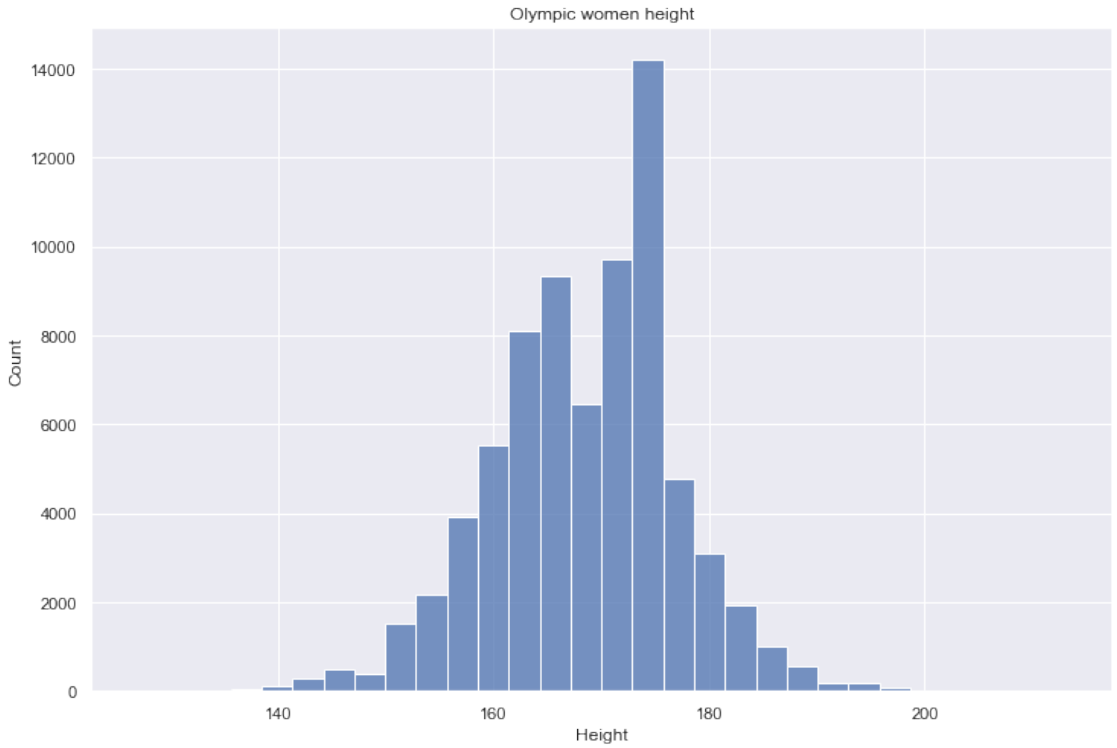
0.2 Submit 2-3 key points you may have discovered about the data, e.g. new relationships?

0.3 Did you come up with additional ideas for other things to review?

Variables gender (sex), height and weight:

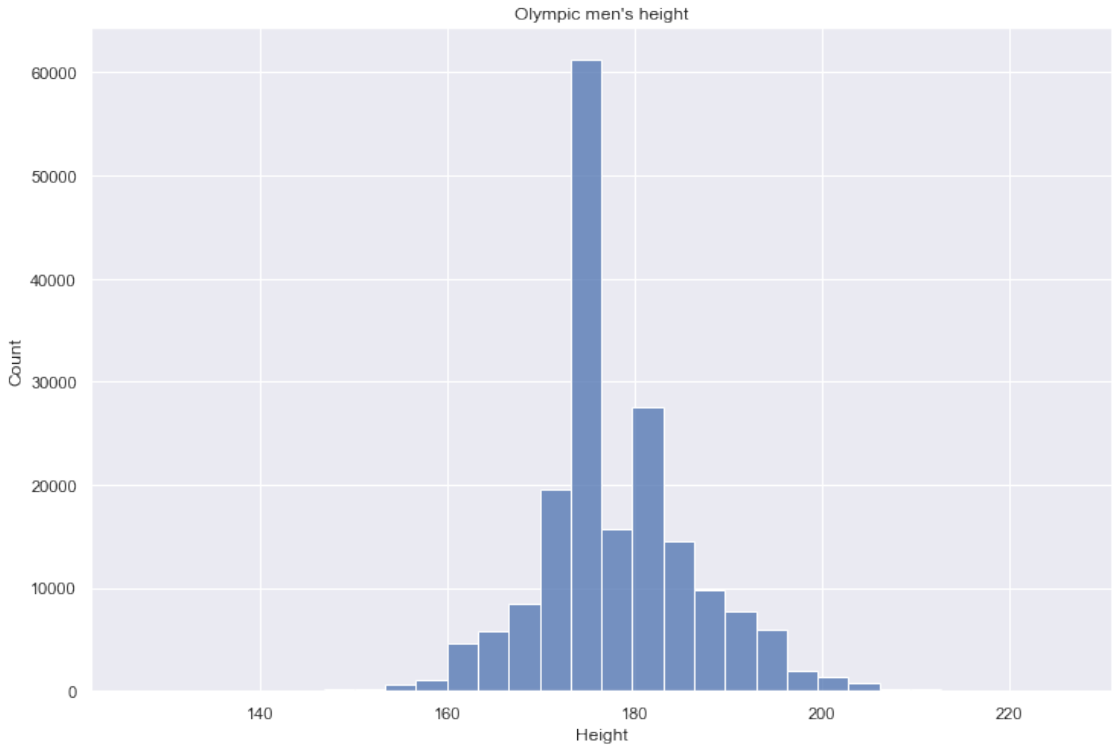
[24]:
```
plt.figure(figsize = (12,8))
sns.histplot(data = athlete_info , x = athlete_info['Height'].
 ↪loc[athlete_info['Sex'] == 'F'],bins = 30)
plt.title('Olympic women height')
```

[24]:
```
Text(0.5, 1.0, 'Olympic women height')
```



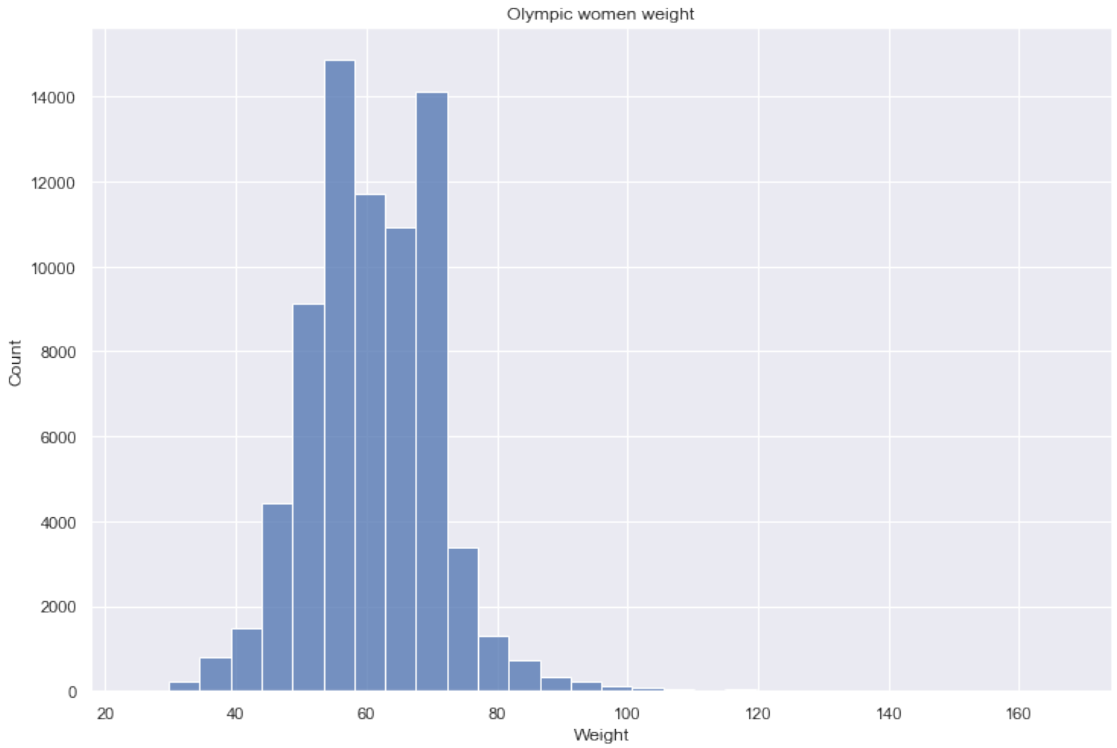
[25]:
```
plt.figure(figsize = (12,8))
sns.histplot(data = athlete_info , x = athlete_info['Height'].
 ↪loc[athlete_info['Sex'] == 'M'],bins = 30)
plt.title('Olympic men\'s height')
```

[25]:
```
Text(0.5, 1.0, "Olympic men's height")
```



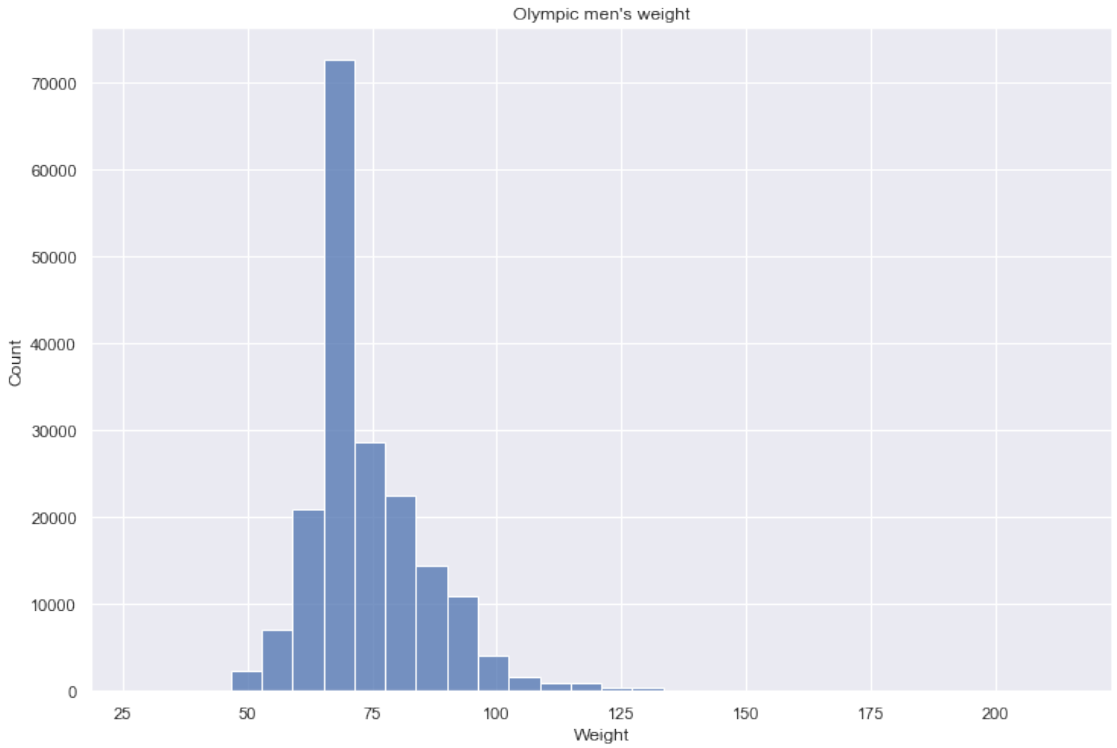
[26]:
```
plt.figure(figsize = (12,8))
sns.histplot(data = athlete_info , x = athlete_info['Weight'].
 ↪loc[athlete_info['Sex'] == 'F'],bins = 30)
plt.title('Olympic women weight')
```

[26]:
```
Text(0.5, 1.0, 'Olympic women weight')
```



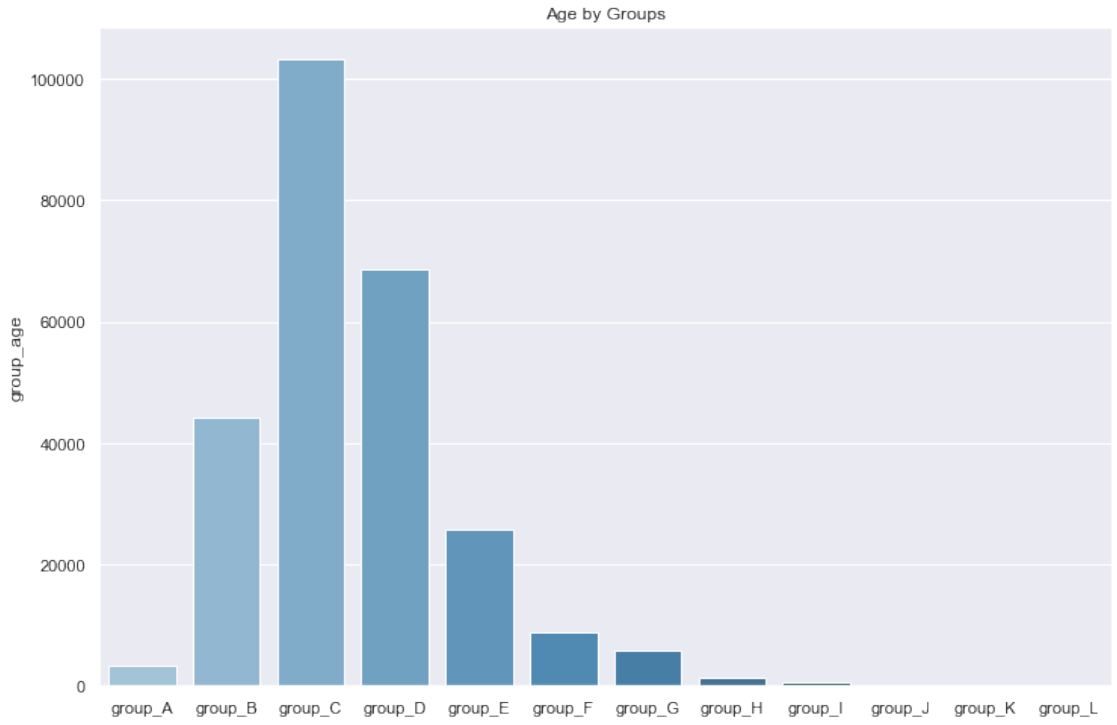
[27]:
```
plt.figure(figsize = (12,8))
sns.histplot(data = athlete_info , x = athlete_info['Weight'].
 ↪loc[athlete_info['Sex'] == 'M'],bins = 30)
plt.title('Olympic men\'s weight')
```

[27]:
```
Text(0.5, 1.0, "Olympic men's weight")
```



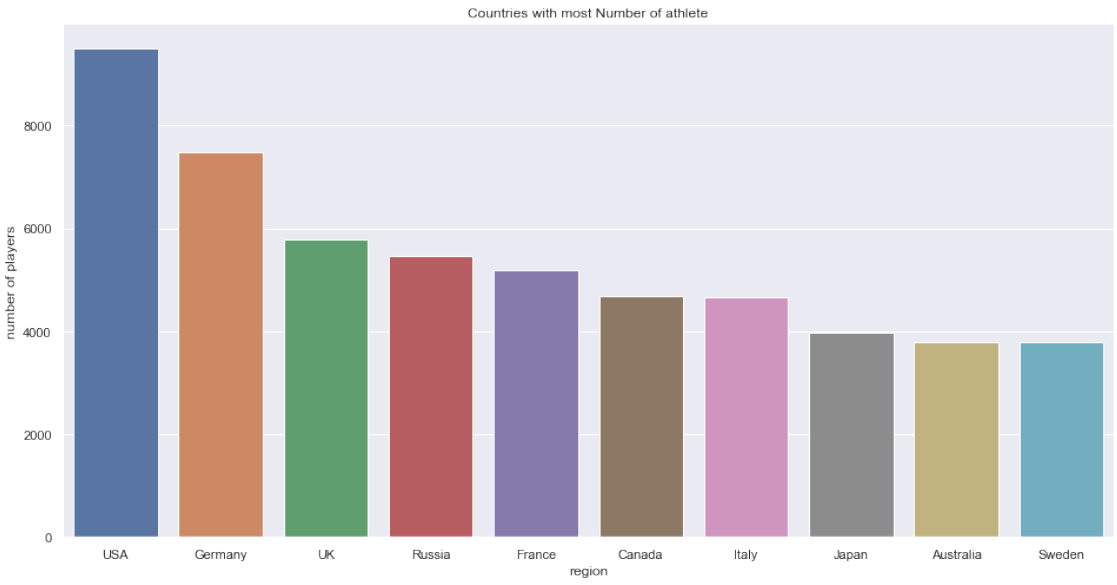
[28]:
```
plt.figure(figsize=(12,8))
sns.barplot(data = age_group , x = age_group.index , y = 'group_age'␣
 ↪,palette="Blues_d",dodge=False)
plt.title("Age by Groups")
```

[28]: Text(0.5, 1.0, 'Age by Groups')



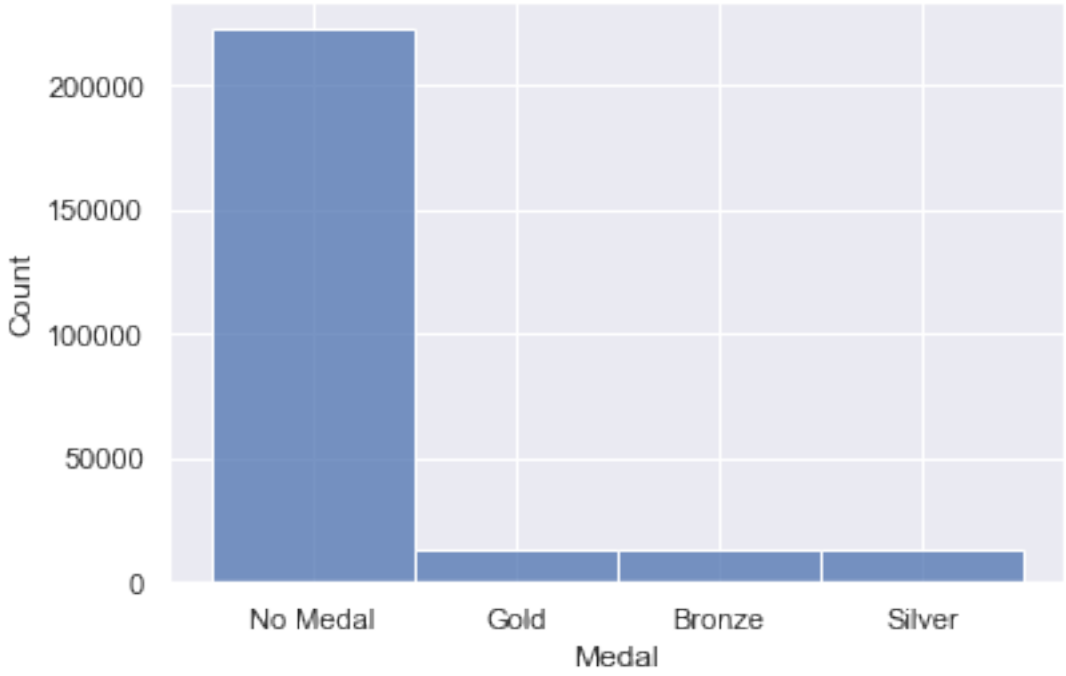
[29]:
```
plt.figure(figsize=(16,8))
sns.barplot(data = countries_with_most_players , x = 'region' , y = 'number of␣
 ↪players')
plt.title('Countries with most Number of athlete')
```

[29]: Text(0.5, 1.0, 'Countries with most Number of athlete')



[30]:
```
sns.histplot(data = athlete_events , x = 'Medal')
```

[30]: <AxesSubplot:xlabel='Medal', ylabel='Count'>



0.4 Did you prove or disprove any of your initial hypotheses? If so, which one and what do you plan to do next?

Hypothesis week 1 Q1: - The age group 20-25 is the most represented?

Confirmed

Q2:- US is the most regular country along the years (measure: number of participants and medals)?

- Partially confirmed. The USA is the country with the most participants and medals. However, it is necessary to check over time.

Q3: - Women in developed countries participate more and get better results (won more medals) + Need more work.

0.5 What additional questions are you seeking to answer?

1:- Which country (NOC) have the most medals for each sport? Has there been a shift over the time?

2:- Are athletes taller today than they were in the past? Is height an advantage in some sports? Is height just as much of an advantage for women as men?

[ ]: