```
[ ]  1  # Name: Jason Weng jyw320
```

```
1  # (10 pts)
2  # 1. What are the top 10 nations in terms of athlete participation?
3  nations_count = all_nations.reduceByKey(lambda a, b: a+b).sortBy(lambda r: -r[1])
4  nations_count.take(10)
```

```
[('United States', 567),
 ('Brazil', 485),
 ('Germany', 441),
 ('Australia', 431),
 ('France', 410),
 ('China', 404),
 ('United Kingdom', 374),
 ('Japan', 346),
 ('Canada', 321),
 ('Spain', 313)]
```

```
1  # (15 pts)
2  # 2. Find the top 3 athletes who won the most gold medals at the 2016 Olympics
3  #HINT: Write an SQL Query.
4  #No credit will be received unless runnable code gets you the result
5  athletes_nations.registerTempTable("athletes_nations")
6  query = "SELECT name, sport, gold FROM athletes_nations ORDER BY gold DESC"
7  gm = spark.sql(query)
8  gm.show(3)
```

```
+--------------+----------+----+
|          name|     sport|gold|
+--------------+----------+----+
|Michael Phelps|  aquatics|   5|
| Katie Ledecky|  aquatics|   4|
|  Simone Biles|gymnastics|   4|
+--------------+----------+----+
only showing top 3 rows
```

```
[71]  1  # (20 pts)
      2  # 3. Find the amount of athletes that participated in each sport, in decending order.
      3  query = "SELECT sport, COUNT(name) AS num_athletes FROM athletes_nations \
      4  WHERE sport == 'aquatics' OR sport == 'gymnastics' \
      5  GROUP BY sport ORDER BY num_athletes DESC"
      6  num_ath = spark.sql(query)
      7  num_ath.show()
```

```
+----------+------------+
|     sport|num_athletes|
+----------+------------+
|  aquatics|        1397|
|gymnastics|         319|
+----------+------------+
```

```
# (20 pts)
# 4. Find the number of total medals that each country won, from most to least.
query = "SELECT country, SUM(gold) as gold, SUM(silver) as silver, SUM(bronze) as bronze, \
(SUM(gold) + SUM(silver) + SUM(bronze)) as total_medals \
FROM athletes_nations GROUP BY country ORDER BY total_medals DESC"
all_medals = spark.sql(query)
all_medals.show(30)
```

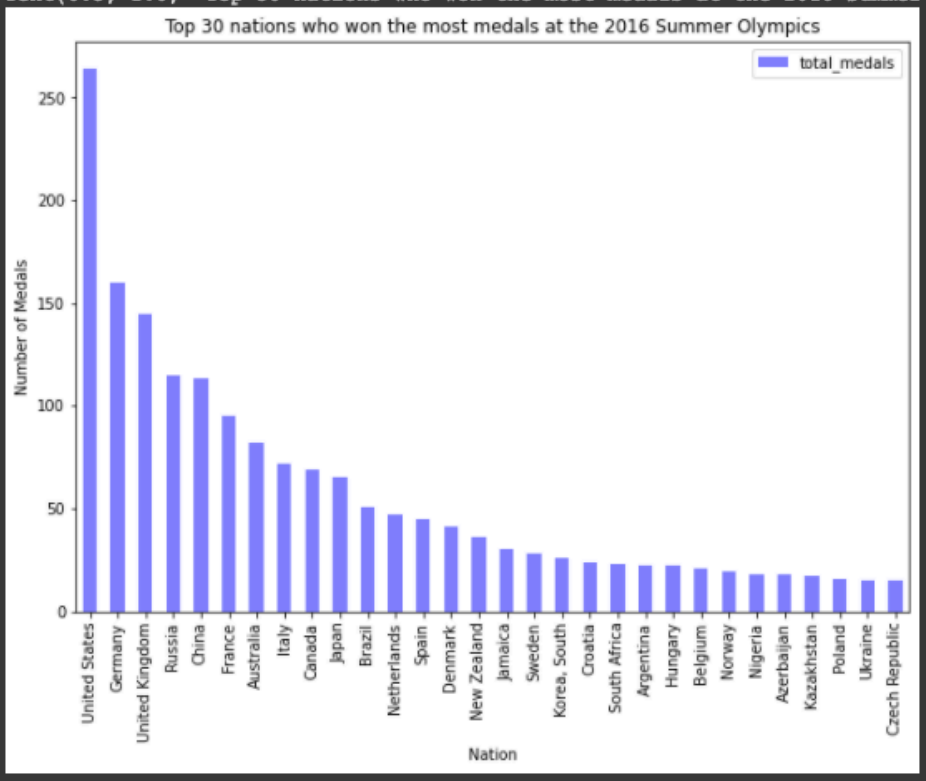
country	gold	silver	bronze	total_medals
United States	139	54	71	264
Germany	49	44	67	160
United Kingdom	64	55	26	145
Russia	52	29	34	115
China	46	30	37	113
France	20	54	21	95
Australia	23	34	25	82
Italy	8	40	24	72
Canada	4	4	61	69
Japan	17	13	35	65
Brazil	37	8	6	51
Netherlands	9	25	13	47
Spain	9	19	17	45
Denmark	15	10	16	41
New Zealand	6	25	5	36
Jamaica	11	17	2	30
Sweden	2	23	3	28
Korea, South	13	3	10	26
Croatia	7	15	2	24
South Africa	2	7	14	23
Hungary	12	3	7	22
Argentina	21	1	0	22
Belgium	2	17	2	21
Norway	0	0	19	19
Azerbaijan	1	7	10	18
Nigeria	0	0	18	18
Kazakhstan	3	5	9	17
Poland	3	3	10	16
Ukraine	2	8	5	15
Czech Republic	1	2	12	15

only showing top 30 rows

[65]
```
# (20 pts)
# 5. Convert the spark dataframe you just created in Q4 to Pandas, and plot a bar graph
#    of the top 30 nations who won the most medals in 2016. Make sure the graph has an
#    x label, y label, and a title
#    Hint: use slicing after you convert the dataframe into Pandas

total_medals_pd = all_medals.toPandas()[:30]
pl = total_medals_pd.plot(kind="bar",
                          x="country", y="total_medals",
                          figsize=(10, 7),  alpha=0.5, color="blue")
pl.set_xlabel("Nation")
pl.set_ylabel("Number of Medals")
pl.set_title("Top 30 nations who won the most medals at the 2016 Summer Olympics")
```

```
Text(0.5, 1.0, 'Top 30 nations who won the most medals at the 2016 Summer Olympics')
```



```
[66]  1  # (15 pts)
      2  # 6. Any other interesting analysis?
      3  query = "SELECT country, SUM(gold) as gold, SUM(silver) as silver, SUM(bronze) as bronze \
      4  FROM athletes_nations GROUP BY country ORDER by silver DESC"
      5  only_bronze = spark.sql(query)
      6  only_bronze.show(5)
      7  #ONLY TOTAL MEDALS THAT THE USA DOES NOT LEAD IN IS SILVER MEDALS
```

```
+--------------+----+------+------+
|       country|gold|silver|bronze|
+--------------+----+------+------+
|United Kingdom|  64|    55|    26|
|        France|  20|    54|    21|
| United States| 139|    54|    71|
|       Germany|  49|    44|    67|
|         Italy|   8|    40|    24|
+--------------+----+------+------+
only showing top 5 rows
```