Open in Colab

In [ ]:

```
from IPython import display
display.Image('/content/drive/MyDrive/Data_Engineering/ Olympic_Games_ Analytics _Project _With_ Apache Spark /Python-a
```

Out[ ]:

In [ ]:

```
#https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5285432114503862/17106
```

Olympic Games Analytics Project in Apache Spark

An Olympic Games Analytics Project in Apache Spark would involve the use of the Apache Spark framework to analyze and process large datasets related to the Olympic Games. This could include data such as athlete performance statistics, medal counts, and event schedules.

In [ ]:

```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.1.tar.gz (281.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 5.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 KB 17.3 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.3.1-py2.py3-none-any.whl size=281845512 sha256=76218b0ce5cbc2bd04fe3e67
271976066baa8c44aa01b312b7a957c5a0f4350a
  Stored in directory: /root/.cache/pip/wheels/43/dc/11/ec201cd671da62fa9c5cc77078235e40722170ceba231d7598
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.3.1
```

In [ ]:

```
#https://sparkbyexamples.com/pyspark/pyspark-groupby-agg-aggregate-explained/
#https://www.datacamp.com/cheat-sheet/pyspark-cheat-sheet-spark-dataframes-in-python
```

In [ ]:

```
#https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/5285432114503862/171066
```

In [ ]:

```
!pip install csv
import pandas as pd
from pandas import DataFrame
from typing import List
from datetime import datetime
import csv
from google.colab import files
from google.colab import drive
!pip install openpyxl
!install urllib
import urllib
from google.colab import drive
drive.mount('/content/drive')
!install seaborn
!install matplotlib
```

```
from IPython import display
!pip install numpy
import numpy as np
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
ERROR: Could not find a version that satisfies the requirement csv (from versions: none)
ERROR: No matching distribution found for csv
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: openpyxl in /usr/local/lib/python3.8/dist-packages (3.0.10)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.8/dist-packages (from openpyxl) (1.1.0)
install: missing destination file operand after 'urllib'
Try 'install --help' for more information.
Mounted at /content/drive
install: missing destination file operand after 'seaborn'
Try 'install --help' for more information.
install: missing destination file operand after 'matplotlib'
Try 'install --help' for more information.
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: numpy in /usr/local/lib/python3.8/dist-packages (1.21.6)
```

In [ ]:

```
!pip install seaborn
!pip install matplotlib

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

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: seaborn in /usr/local/lib/python3.8/dist-packages (0.11.2)
Requirement already satisfied: numpy>=1.15 in /usr/local/lib/python3.8/dist-packages (from seaborn) (1.21.6)
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.8/dist-packages (from seaborn) (1.7.3)
Requirement already satisfied: matplotlib>=2.2 in /usr/local/lib/python3.8/dist-packages (from seaborn) (3.2.2)
Requirement already satisfied: pandas>=0.23 in /usr/local/lib/python3.8/dist-packages (from seaborn) (1.3.5)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.8/dist-packages (from matplotlib>=2.2->se
aborn) (2.8.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.8/dist-packages (from
matplotlib>=2.2->seaborn) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.8/dist-packages (from matplotlib>=2.2->seabo
rn) (1.4.4)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.8/dist-packages (from matplotlib>=2.2->seaborn)
(0.11.0)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-packages (from pandas>=0.23->seaborn) (202
2.7)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-packages (from python-dateutil>=2.1->matplotli
b>=2.2->seaborn) (1.15.0)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: matplotlib in /usr/local/lib/python3.8/dist-packages (3.2.2)
Requirement already satisfied: numpy>=1.11 in /usr/local/lib/python3.8/dist-packages (from matplotlib) (1.21.6)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.8/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.8/dist-packages (from
matplotlib) (3.0.9)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.8/dist-packages (from matplotlib) (1.4.4)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.8/dist-packages (from matplotlib) (0.11.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-packages (from python-dateutil>=2.1->matplotli
b) (1.15.0)
```

In [ ]:

```
# Download Java Virtual Machine (JVM)
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
```

In [ ]:

```
# Download Spark
!wget -q https://dlcdn.apache.org/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz
# Unzip the file
!tar xf spark-3.2.1-bin-hadoop3.2.tgz
```

In [ ]:

```
# 1. Start by creating a new Google Colab notebook by going to https://colab.research.google.com/ and clicking on the "

# 2. Install pyspark by running !pip install pyspark in a code cell.

# 3. Import the necessary libraries by running the following code in a code cell:
```

In [ ]:

```
from pyspark import SparkConf, SparkContext
from pyspark.sql import SparkSession
```

In [ ]:

```
# 4. Create a SparkSession by running the following code in a code cell:
```

In [ ]:

```
spark = SparkSession.builder.appName("CSV Processing with PySpark").getOrCreate()
```

In [ ]:

```
# 5. Now you can read a CSV file and create a dataframe by running the following code in a code cell:
```

In [ ]:

```
df = spark.read.format("csv").options(header="true", inferSchema="true").load("/content/drive/MyDrive/Data_Engineering/
```

In [ ]:

```
# 6. Perform operations on the dataframe like selection, filtering, and aggregation using the DataFrame API or SQL.

# 7. To save the dataframe in csv format you can use the following code: df.write.format("csv").save("")
```

In [ ]:

```
df.show()
```

```
+---+--------------------+---+---+------+------+--------------+---+-----------+----+------+-----------+----------------
----+--------------------+-----+
| ID|                Name|Sex|Age|Height|Weight|          Team|NOC|      Games|Year|Season|       City|               S
port|               Event|Medal|
+---+--------------------+---+---+------+------+--------------+---+-----------+----+------+-----------+----------------
----+--------------------+-----+
|  1|           A Dijiang|  M| 24|   180|    80|         China|CHN|1992 Summer|1992|Summer|  Barcelona|          Basket
ball|Basketball Men's ...|   NA|
|  2|            A Lamusi|  M| 23|   170|    60|         China|CHN|2012 Summer|2012|Summer|     London|                
Judo|Judo Men's Extra-...|   NA|
|  3| Gunnar Nielsen Aaby|  M| 24|    NA|    NA|       Denmark|DEN|1920 Summer|1920|Summer|  Antwerpen|            Foot
ball|Football Men's Fo...|   NA|
|  4|Edgar Lindenau Aabye|  M| 34|    NA|    NA|Denmark/Sweden|DEN|1900 Summer|1900|Summer|      Paris|          Tug-Of
-War|Tug-Of-War Men's ...| Gold|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988|Winter|    Calgary|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988|Winter|    Calgary|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992|Winter|Albertville|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992|Winter|Albertville|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994|Winter|Lillehammer|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994|Winter|Lillehammer|       Speed Ska
ting|Speed Skating Wom...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994|Winter|Lillehammer|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994|Winter|Lillehammer|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994|Winter|Lillehammer|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994|Winter|Lillehammer|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
iing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992|Winter|Albertville|Cross Country Sk
```

```
iing|Cross Country Ski...|   NA|
+---+--------------------+---+---+------+------+--------------+---+-----------+----+------+-----------+----------------
----+--------------------+-----+
only showing top 20 rows
```

In [ ]:

```python
# You can print the schema of a Spark DataFrame in PySpark by using the .printSchema() method.
df.printSchema()
```

```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- Height: string (nullable = true)
 |-- Weight: string (nullable = true)
 |-- Team: string (nullable = true)
 |-- NOC: string (nullable = true)
 |-- Games: string (nullable = true)
 |-- Year: string (nullable = true)
 |-- Season: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Sport: string (nullable = true)
 |-- Event: string (nullable = true)
 |-- Medal: string (nullable = true)
```

In [ ]:

```python
# Convert String Type to Double Type
#In PySpark, the "double" data type is used to represent decimal numbers,
#while the "string" data type is used to represent text.
```

In [ ]:

```python
from pyspark.sql.functions import col

# Assume that your DataFrame is called "df" and the column you want to convert is called "column_name"
#df = df.withColumn("column_name", col("column_name").cast("double"))
```

In [ ]:

```python
from pyspark.sql.functions import col, cast
```

In [ ]:

```python
# Assume you have a DataFrame called df

# List of columns to be modified to double datatype
columns_to_change = ['ID', 'Age', 'Height','Weight','Year']
```

```
for column in columns_to_change:
    df1 = df.withColumn(column, col(column).cast("double"))
```

In [ ]:

```
#In PySpark, the "double" data type is used to represent decimal numbers,
#while the "string" data type is used to represent text.
df1.printSchema()
```

```
root
 |-- ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- Height: string (nullable = true)
 |-- Weight: string (nullable = true)
 |-- Team: string (nullable = true)
 |-- NOC: string (nullable = true)
 |-- Games: string (nullable = true)
 |-- Year: double (nullable = true)
 |-- Season: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Sport: string (nullable = true)
 |-- Event: string (nullable = true)
 |-- Medal: string (nullable = true)
```

In [ ]:

```
df1.show()
```

```
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
| ID|                Name|Sex|Age|Height|Weight|          Team|NOC|      Games|  Year|Season|       City|
Sport|               Event|Medal|
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
|  1|           A Dijiang|  M| 24|   180|    80|         China|CHN|1992 Summer|1992.0|Summer|  Barcelona|          Bask
etball|Basketball Men's ...|   NA|
|  2|            A Lamusi|  M| 23|   170|    60|         China|CHN|2012 Summer|2012.0|Summer|     London|
Judo|Judo Men's Extra-...|   NA|
|  3| Gunnar Nielsen Aaby|  M| 24|    NA|    NA|       Denmark|DEN|1920 Summer|1920.0|Summer|  Antwerpen|            Fo
otball|Football Men's Fo...|   NA|
|  4|Edgar Lindenau Aabye|  M| 34|    NA|    NA|Denmark/Sweden|DEN|1900 Summer|1900.0|Summer|      Paris|          Tug-
Of-War|Tug-Of-War Men's ...| Gold|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988.0|Winter|    Calgary|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988.0|Winter|    Calgary|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992.0|Winter|Albertville|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992.0|Winter|Albertville|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994.0|Winter|Lillehammer|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994.0|Winter|Lillehammer|       Speed S
kating|Speed Skating Wom...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
Skiing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country 
```

```
Skiing|Cross Country Ski...|   NA|
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
only showing top 20 rows
```

```
In [ ]: #In PySpark, a temporary view can be created using the createOrReplaceTempView method of a DataFrame. This method create
```

```
In [ ]: df1.createOrReplaceTempView("df1_temp_table")
```

```
In [ ]: spark.sql("SELECT * FROM df1_temp_table").show()
```

```
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
| ID|                Name|Sex|Age|Height|Weight|          Team|NOC|      Games|  Year|Season|       City|
Sport|               Event|Medal|
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
|  1|           A Dijiang|  M| 24|   180|    80|         China|CHN|1992 Summer|1992.0|Summer|  Barcelona|          Bask
etball|Basketball Men's ...|   NA|
|  2|            A Lamusi|  M| 23|   170|    60|         China|CHN|2012 Summer|2012.0|Summer|     London|
Judo|Judo Men's Extra-...|   NA|
|  3| Gunnar Nielsen Aaby|  M| 24|    NA|    NA|       Denmark|DEN|1920 Summer|1920.0|Summer|  Antwerpen|            Fo
otball|Football Men's Fo...|   NA|
|  4|Edgar Lindenau Aabye|  M| 34|    NA|    NA|Denmark/Sweden|DEN|1900 Summer|1900.0|Summer|      Paris|          Tug-
Of-War|Tug-Of-War Men's ...| Gold|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988.0|Winter|    Calgary|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 21|   185|    82|   Netherlands|NED|1988 Winter|1988.0|Winter|    Calgary|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992.0|Winter|Albertville|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 25|   185|    82|   Netherlands|NED|1992 Winter|1992.0|Winter|Albertville|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994.0|Winter|Lillehammer|       Speed S
kating|Speed Skating Wom...|   NA|
|  5|Christine Jacoba ...|  F| 27|   185|    82|   Netherlands|NED|1994 Winter|1994.0|Winter|Lillehammer|       Speed S
kating|Speed Skating Wom...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 31|   188|    75| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country
Skiing|Cross Country Ski...|   NA|
|  6|     Per Knut Aaland|  M| 33|   188|    75| United States|USA|1994 Winter|1994.0|Winter|Lillehammer|Cross Country
Skiing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
Skiing|Cross Country Ski...|   NA|
|  7|        John Aalberg|  M| 31|   183|    72| United States|USA|1992 Winter|1992.0|Winter|Albertville|Cross Country
```

```
Skiing|Cross Country Ski...|   NA|
+---+--------------------+---+---+------+------+--------------+---+-----------+------+------+-----------+--------------
------+--------------------+-----+
only showing top 20 rows
```

In [ ]:

```
dff = spark.read.format("csv").options(header="true", inferSchema="true").load("/content/drive/MyDrive/Data_Engineering
dff.show()
```

```
+---+--------------+--------------------+
|NOC|        region|               notes|
+---+--------------+--------------------+
|AFG|   Afghanistan|                null|
|AHO|       Curacao|Netherlands Antilles|
|ALB|       Albania|                null|
|ALG|       Algeria|                null|
|AND|       Andorra|                null|
|ANG|        Angola|                null|
|ANT|       Antigua| Antigua and Barbuda|
|ANZ|     Australia|         Australasia|
|ARG|     Argentina|                null|
|ARM|       Armenia|                null|
|ARU|         Aruba|                null|
|ASA|American Samoa|                null|
|AUS|     Australia|                null|
|AUT|       Austria|                null|
|AZE|    Azerbaijan|                null|
|BAH|       Bahamas|                null|
|BAN|    Bangladesh|                null|
|BAR|      Barbados|                null|
|BDI|       Burundi|                null|
|BEL|       Belgium|                null|
+---+--------------+--------------------+
only showing top 20 rows
```

In [ ]:

```
dff.createOrReplaceTempView("dff_temp_table")
```

In [ ]:

```
spark.sql("SELECT * FROM dff_temp_table").show()
```

```
+---+--------------+--------------------+
|NOC|        region|               notes|
+---+--------------+--------------------+
|AFG|   Afghanistan|                null|
|AHO|       Curacao|Netherlands Antilles|
|ALB|       Albania|                null|
|ALG|       Algeria|                null|
|AND|       Andorra|                null|
|ANG|        Angola|                null|
|ANT|       Antigua| Antigua and Barbuda|
|ANZ|     Australia|         Australasia|
|ARG|     Argentina|                null|
|ARM|       Armenia|                null|
|ARU|         Aruba|                null|
|ASA|American Samoa|                null|
|AUS|     Australia|                null|
|AUT|       Austria|                null|
|AZE|    Azerbaijan|                null|
|BAH|       Bahamas|                null|
|BAN|    Bangladesh|                null|
|BAR|      Barbados|                null|
|BDI|       Burundi|                null|
|BEL|       Belgium|                null|
+---+--------------+--------------------+
only showing top 20 rows
```

In [ ]:

```
# Distribution of the age of gold medalists
```

In [ ]:

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget -q https://www-us.apache.org/dist/spark/spark-2.4.6/spark-2.4.6-bin-hadoop2.7.tgz
!tar xf spark-2.4.6-bin-hadoop2.7.tgz
!pip install -q findspark
import os
import findspark
findspark.init()
```

```
tar: spark-2.4.6-bin-hadoop2.7.tgz: Cannot open: No such file or directory
tar: Error is not recoverable: exiting now
```

In [ ]:

```
from pyspark.sql.functions import count
```

In [ ]:

```
from pyspark.sql import SQLContext
from pyspark.sql.functions import sum, col, desc
from pyspark.sql import *
```

In [ ]:

```
query1 = spark.sql("SELECT COUNT(Medal)  as Medals, Age  FROM df1_temp_table WHERE Medal = 'Gold' group By Age ORDER BY
```

Distribution of the age of gold medalists ORDER BY Age DESC

In [ ]:

```
query1.show()
```

```
+------+---+
|Medals|Age|
+------+---+
|   148| NA|
|     2| 64|
|     4| 63|
|     4| 60|
|     2| 59|
|     3| 58|
|     2| 57|
|    10| 56|
|     1| 55|
|    15| 54|
|     6| 53|
|    12| 52|
|     4| 51|
|    12| 50|
|    15| 49|
|    21| 48|
|    24| 47|
|    24| 46|
|    20| 45|
|    38| 44|
+------+---+
only showing top 20 rows
```

In [ ]:

```
df_pandas1 = query1.toPandas()
df_pandas1
```

Out[ ]:

	Medals	Age
0	148	NA
1	2	64
2	4	63
3	4	60
4	2	59
5	3	58
6	2	57
7	10	56
8	1	55
9	15	54
10	6	53
11	12	52
12	4	51
13	12	50
14	15	49
15	21	48
16	24	47
17	24	46
18	20	45
19	38	44
20	32	43
21	41	42
22	43	41
23	74	40
24	65	39

	Medals	Age
25	89	38
26	81	37
27	131	36
28	174	35
29	217	34
30	289	33
31	354	32
32	396	31
33	523	30
34	647	29
35	797	28
36	859	27
37	970	26
38	1045	25
39	1125	24
40	1126	23
41	1087	22
42	910	21
43	666	20
44	457	19
45	278	18
46	189	17
47	113	16
48	75	15
49	27	14

	Medals	Age
50	7	13

In [ ]:

```
# Convert PySpark DataFrame to Pandas DataFrame
#df_pandas = df.toPandas()
```

In [ ]:

```
import seaborn as sns
#pd(result1['Medals'])
```

In [ ]:

```
```

Distribution of the age of gold medalists

In [ ]:

```
query2 = spark.sql("Select count(Medal),Age from df1_temp_table  where Medal='Gold' group by Age order by Age;")
query2.show()
```

```
+------------+---+
|count(Medal)|Age|
+------------+---+
|           7| 13|
|          27| 14|
|          75| 15|
|         113| 16|
|         189| 17|
|         278| 18|
|         457| 19|
|         666| 20|
|         910| 21|
|        1087| 22|
|        1126| 23|
|        1125| 24|
|        1045| 25|
|         970| 26|
|         859| 27|
|         797| 28|
|         647| 29|
|         523| 30|
|         396| 31|
|         354| 32|
+------------+---+
only showing top 20 rows
```

In [ ]:

```
df_pandas2 = query2.toPandas()
```

In [ ]:

```
df_pandas2
```

Out[ ]:

	count(Medal)	Age
0	7	13
1	27	14
2	75	15
3	113	16
4	189	17
5	278	18
6	457	19
7	666	20
8	910	21
9	1087	22
10	1126	23
11	1125	24
12	1045	25
13	970	26
14	859	27
15	797	28
16	647	29
17	523	30
18	396	31
19	354	32
20	289	33
21	217	34
22	174	35
23	131	36
24	81	37

	count(Medal)	Age
25	89	38
26	65	39
27	74	40
28	43	41
29	41	42
30	32	43
31	38	44
32	20	45
33	24	46
34	24	47
35	21	48
36	15	49
37	12	50
38	4	51
39	12	52
40	6	53
41	15	54
42	1	55
43	10	56
44	2	57
45	3	58
46	2	59
47	4	60
48	4	63
49	2	64

	count(Medal)	Age
50	148	NA

In [ ]:

```

```

In [ ]:

```
#df_pandas1_sns = sns.load_dataset("df_pandas1")
#sns.displot(df_pandas1, x = 'Age',y ='count(Medal)') #binwidth=3
#plt.figure(figsize=(30,15))
#sns.set(figure_size=(8, 6))
plt.figure(figsize = (20,15))
sns.distplot(df_pandas2, x = df_pandas2['Age'].replace('NA', np.nan, inplace=True), hist=True )
```

In [ ]:

```

```

In [ ]:

```
#sns.barplot(
#sns.barplot(df_pandas1.iloc[0:10], x = df_pandas1['Age'].value_counts().index)
```

Gold Medals for Athletes Over 50 based on Sports

In [ ]:

```
query3 = spark.sql("Select Sport, Age from df1_temp_table where Medal='Gold'and  Age > 50;")
query3.show()
```

In [ ]:

```
df_pandas3 = query3.toPandas()
df_pandas3
```

Women medals per edition(Summer Season) of the Games

In [ ]:

```
query4 = spark.sql("""Select count(Medal),Year from df1_temp_table where Sex='F' and Season ="Summer" and Medal in('Gol
query4.show()
```

In [ ]:

```
df_pandas4 = query4.toPandas()
df_pandas4
```

Top 5 Gold Medal Countries

In [ ]:

```
query5 = spark.sql("""Select count(Medal) as MedalCount ,region from df1_temp_table DF1 JOIN dff_temp_table  NR ON DF1.
query5.show()
```

```
+----------+-------+
|MedalCount| region|
+----------+-------+
|      2535|    USA|
|      1597| Russia|
|      1300|Germany|
|       677|     UK|
|       575|  Italy|
+----------+-------+
```

In [ ]:

```
df_pandas5 = query5.toPandas()
```

In [ ]:

```
df_pandas5
```

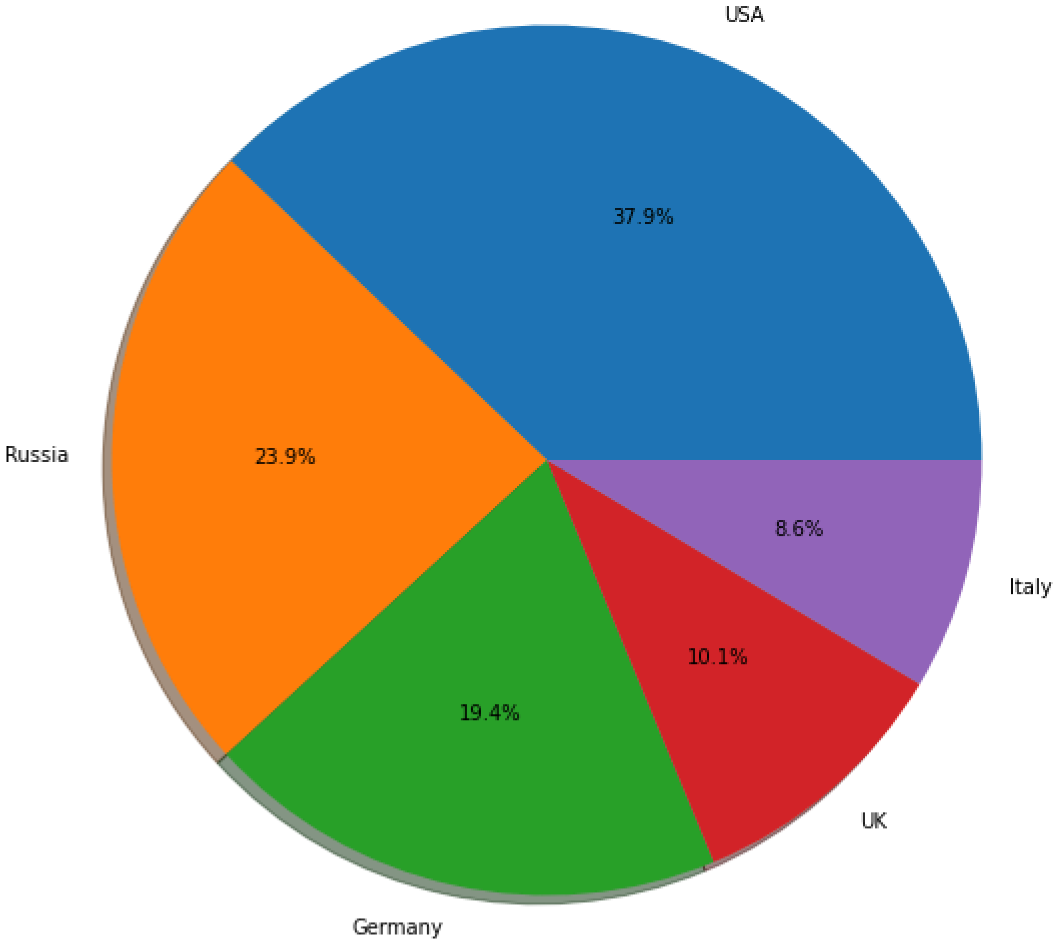
Out[ ]:

	MedalCount	region
0	2535	USA
1	1597	Russia
2	1300	Germany
3	677	UK
4	575	Italy

In [ ]:

```
keys = ['USA', 'Russia', 'Germany', 'UK', 'Italy']
plt.figure(figsize = (20,10))
plt.pie(data=df_pandas5,x = df_pandas5['MedalCount'],labels=keys,autopct='%1.1f%%',shadow = True)
```

```
Out[ ]: ([<matplotlib.patches.Wedge at 0x7f4af796c100>,
          <matplotlib.patches.Wedge at 0x7f4af796c7f0>,
          <matplotlib.patches.Wedge at 0x7f4af78fb100>,
          <matplotlib.patches.Wedge at 0x7f4af78fb9d0>,
          <matplotlib.patches.Wedge at 0x7f4af79072e0>],
         [Text(0.4073010443520564, 1.0218149828954968, 'USA'),
          Text(-1.0999648851301842, 0.008789282140262408, 'Russia'),
          Text(-0.23698985105652282, -1.0741674964809758, 'Germany'),
          Text(0.7187469935411899, -0.8327080876726856, 'UK'),
          Text(1.0600717540749174, -0.2936798873135993, 'Italy')],
         [Text(0.22216420601021256, 0.5573536270339073, '37.9%'),
          Text(-0.5999808464346458, 0.004794153894688585, '23.9%'),
          Text(-0.12926719148537608, -0.5859095435350776, '19.4%'),
          Text(0.3920438146588308, -0.4542044114578284, '10.1%'),
          Text(0.5782209567681367, -0.1601890294437814, '8.6%')])
```



Disciplines with the greatest number of Gold Medals

In [ ]:

```
query6 = spark.sql("""Select Count(Medal) Medals_Wins,Event from df1_temp_table where Medal='Gold' group by Event order
result = query6.show()
```

```
+-----------+--------------------+
|Medals_Wins|               Event|
+-----------+--------------------+
|        414|Football Men's Fo...|
|        404|Ice Hockey Men's ...|
|        360| Hockey Men's Hockey|
|        286|Water Polo Men's ...|
|        249|Gymnastics Men's ...|
|        238|Rowing Men's Coxe...|
|        227|Basketball Men's ...|
|        194|Handball Men's Ha...|
|        166|Volleyball Men's ...|
|        157|Hockey Women's Ho...|
|        156|Volleyball Women'...|
|        155|Handball Women's ...|
|        145|Swimming Men's 4 ...|
|        134|Fencing Men's epe...|
|        131|Basketball Women'...|
|        130|Fencing Men's Sab...|
|        129|Gymnastics Women'...|
|        125|Swimming Women's ...|
|        117|Fencing Men's Foi...|
|        112|Baseball Men's Ba...|
+-----------+--------------------+
only showing top 20 rows
```

In [ ]:

```
df_pandas6 = query6.toPandas()
df_pandas6
```

Out[ ]:

	Medals_Wins	Event
0	414	Football Men's Football
1	404	Ice Hockey Men's Ice Hockey
2	360	Hockey Men's Hockey
3	286	Water Polo Men's Water Polo
4	249	Gymnastics Men's Team All-Around
...	...	...
745	1	Gymnastics Men's Tumbling
746	1	Equestrianism Mixed Hacks And Hunter Combined
747	1	Snowboarding Men's Parallel Slalom
748	1	Snowboarding Women's Giant Slalom
749	1	Aeronautics Mixed Aeronautics

750 rows × 2 columns

Disciplines with the greatest number of Gold Medals for Usa

In [ ]:

```
query7 = spark.sql("""Select Count(Medal) Medals_Wins,Event from df1_temp_table DF1 JOIN dff_temp_table  NR ON DF1.NOC
query7.show()
```

```
+-----------+--------------------+
|Medals_Wins|               Event|
+-----------+--------------------+
|        179|Basketball Men's ...|
|        105|Swimming Men's 4 ...|
|        103|Swimming Men's 4 ...|
|        102|Rowing Men's Coxe...|
|         95|Basketball Women'...|
|         77|Swimming Women's ...|
|         76|Athletics Men's 4...|
|         74|Swimming Women's ...|
|         64|Football Women's ...|
|         61|Athletics Men's 4...|
|         57|Swimming Men's 4 ...|
|         48|Athletics Women's...|
|         45|Softball Women's ...|
|         38|Athletics Women's...|
|         36|Volleyball Men's ...|
|         36|Rowing Women's Co...|
|         33|   Rugby Men's Rugby|
|         33|Ice Hockey Men's ...|
|         33|Swimming Women's ...|
|         25|Water Polo Women'...|
+-----------+--------------------+
only showing top 20 rows
```

In [ ]:

```
df_pandas7 = query7.toPandas()
df_pandas7
```

Out[ ]:

	Medals_Wins	Event
0	179	Basketball Men's Basketball
1	105	Swimming Men's 4 x 100 metres Medley Relay
2	103	Swimming Men's 4 x 200 metres Freestyle Relay
3	102	Rowing Men's Coxed Eights
4	95	Basketball Women's Basketball
...	...	...
307	1	Freestyle Skiing Men's Moguls
308	1	Gymnastics Men's Tumbling
309	1	Athletics Men's 5,000 metres
310	1	Swimming Women's 50 metres Freestyle
311	1	Shooting Men's Small-Bore Rifle, Three Positio...

312 rows × 2 columns

Height vs Weight of Olympic Medalists

In [ ]:

```
query8 = spark.sql("""select Weight, Height from df1_temp_table where  Medal = 'Gold'AND Weight IS NOT NULL AND Height 
query8.show()
```

```
+------+------+
|Weight|Height|
+------+------+
|    NA|    NA|
|    64|   175|
|    64|   175|
|    64|   175|
|    85|   176|
|    85|   176|
|    85|   176|
|    85|   176|
|    NA|   163|
|    NA|    NA|
|    NA|    NA|
|    NA|    NA|
|    83|   180|
|    86|   182|
|    86|   182|
|    82|   185|
|    83|   186|
|    82|   181|
|    85|   190|
|    96|   188|
+------+------+
only showing top 20 rows
```

In [ ]:

```
df_pandas8 = query8.toPandas()
df_pandas8
```

Out[ ]:

	Weight	Height
0	NA	NA
1	64	175
2	64	175
3	64	175
4	85	176
...	...	...
13249	90	182
13250	60	167
13251	93	200
13252	93	197
13253	80	168

13254 rows × 2 columns

In [ ]:

```

```

Variation of Male Athletes over time

In [ ]:

```
query9 = spark.sql("""select count(Sex) as Males, Year from df1_temp_table where Sex = 'M' and Season = 'Summer' group
query9.show()
```

```
+-----+------+
|Males|  Year|
+-----+------+
|  380|1896.0|
| 1901|1900.0|
| 1278|1904.0|
| 1721|1906.0|
| 3039|1908.0|
| 3944|1912.0|
| 4149|1920.0|
| 4978|1924.0|
| 4574|1928.0|
| 2609|1932.0|
| 6023|1936.0|
| 5743|1948.0|
| 6743|1952.0|
| 4208|1956.0|
| 6660|1960.0|
| 6326|1964.0|
| 6786|1968.0|
| 8090|1972.0|
| 6457|1976.0|
| 5435|1980.0|
+-----+------+
only showing top 20 rows
```

In [ ]:

```
df_pandas9 = query9.toPandas()
df_pandas9
```

Out[ ]:

	Males	Year
0	380	1896.0
1	1901	1900.0
2	1278	1904.0
3	1721	1906.0
4	3039	1908.0
5	3944	1912.0
6	4149	1920.0
7	4978	1924.0
8	4574	1928.0
9	2609	1932.0
10	6023	1936.0
11	5743	1948.0
12	6743	1952.0
13	4208	1956.0
14	6660	1960.0
15	6326	1964.0
16	6786	1968.0
17	8090	1972.0
18	6457	1976.0
19	5435	1980.0
20	6984	1984.0
21	8473	1988.0
22	8832	1992.0
23	8760	1996.0
24	8386	2000.0

	Males	Year
25	7895	2004.0
26	7783	2008.0
27	7099	2012.0
28	7462	2016.0

Variation of Female Athletes over time

```
In [ ]: query10 = spark.sql("""select count(Sex) as Females, Year from df1_temp_table where Sex = 'F' and Season = 'Summer' grou
        query10.show()
```

```
+-------+------+
|Females|  Year|
+-------+------+
|     32|1900.0|
|     16|1904.0|
|     11|1906.0|
|     47|1908.0|
|     87|1912.0|
|    133|1920.0|
|    243|1924.0|
|    401|1928.0|
|    337|1932.0|
|    459|1936.0|
|    624|1948.0|
|   1484|1952.0|
|    891|1956.0|
|   1422|1960.0|
|   1336|1964.0|
|   1767|1968.0|
|   2179|1972.0|
|   2164|1976.0|
|   1755|1980.0|
|   2442|1984.0|
+-------+------+
only showing top 20 rows
```

In [ ]:

```
df_pandas10 = query10.toPandas()
df_pandas10
```

Out[ ]:

	Females	Year
0	32	1900.0
1	16	1904.0
2	11	1906.0
3	47	1908.0
4	87	1912.0
5	133	1920.0
6	243	1924.0
7	401	1928.0
8	337	1932.0
9	459	1936.0
10	624	1948.0
11	1484	1952.0
12	891	1956.0
13	1422	1960.0
14	1336	1964.0
15	1767	1968.0
16	2179	1972.0
17	2164	1976.0
18	1755	1980.0
19	2442	1984.0
20	3535	1988.0
21	4114	1992.0
22	4998	1996.0
23	5430	2000.0
24	5545	2004.0

	Females	Year
25	5816	2008.0
26	5815	2012.0
27	6223	2016.0

```
In [ ]:
```

```
In [ ]:
```