Chapter 4: Spark SQL & Dataframe

Ex 1: Fifa2018

Cho tập tin Fifa2018_dataset.csv

Yêu cầu:

1. Đọc tập tin Fifa2018_dataset.csv vào fifa_df.
2. In schema của fifa_df. Hiển thị 2 dòng đầu tiên của dữ liệu. Cho biết dữ liệu có bao nhiêu dòng?
3. Tạo view 'fifa_table' từ fifa_df
4. Hãy thực hiện SQL Query để lấy cột Age của các vận động viên có Nationality là "Germany" => fifa_germany_age. Hiển thị 3 dòng đầu của dữ liệu. In thống kê dữ liệu
5. Trực quan hóa dữ liệu fifa_germany_age. Nhận xét biểu đồ.
6. Từ fifa_df, cho biết mỗi độ tuổi có bao nhiêu cầu thủ. Độ tuổi trung bình của cầu thủ mỗi quốc gia là bao nhiêu?
7. Từ fifa_df, cho biết "Age" nhỏ nhất, "Age" lớn nhất, "Strength" nhỏ nhất, "Strength" lớn nhất
8. Liệt kê danh sách các "Club" (duy nhất) theo 2 cách với Dataframe fifa_df và SQL query với fifa_table.
9. Từ fifa_df, sắp xếp dữ liệu giảm dần theo Age => fifa_df_desc.
10. Có bao nhiêu cầu thủ trong "Name" có "Cristiano" theo 2 cách với Dataframe fifa_df và SQL query với fifa_table. In tên các cầu thủ này.

In [1]:
```
import findspark
findspark.init()
```

In [2]:
```
import pyspark
```

In [3]:
```
from pyspark import SparkContext
from pyspark.conf import SparkConf
from pyspark.sql import SparkSession
```

In [4]:
```
sc =SparkContext()
```

In [5]:
```
spark = SparkSession(sc)
```

In [6]:
```
#1. Create a DataFrame from CSV file. # Load the Dataframe
fifa_df = spark.read.csv("data/Fifa2018_dataset.csv", header=True,
                         inferSchema=True)
```

In [7]:
```
#2. Check the schema of columns
# fifa_df.printSchema()
```

In [8]:
```
# Show the first 3 observations
for row in fifa_df.head(2):
    print(row)
    print('\n')
#fifa_df.show(3)
```

```
Row(_c0=0, Name='Cristiano Ronaldo', Age=32, Photo='https://cdn.sofifa.org/48/18/players/20801.png', Nationality='Portugal', Flag='https://cdn.sofifa.org/flags/38.png', Overall=94, Potential=94, Club='Real Madrid CF', Club Logo='https://cdn.sofifa.org/24/18/teams/243.png', Value='€95.5M', Wage='€565K', Special=2228, Acceleration='89', Aggression='63', Agility='89', Balance='63', Ball control='93', Composure='95', Crossing='85', Curve='81', Dribbling='91', Finishing='94', Free kick accuracy='76', GK diving='7', GK handling='11', GK kicking='15', GK positioning='14', GK reflexes='11', Heading accuracy='88', Interceptions='29', Jumping='95', Long passing='77', Long shots='92', Marking='22', Penalties='85', Positioning='95', Reactions='96', Short passing='83', Shot power='94', Sliding tackle='23', Sprint speed='91', Stamina='92', Standing tackle='31', Strength='80', Vision='85', Volleys='88', CAM=89.0, CB=53.0, CDM=62.0, CF=91.0, CM=82.0, ID=20801, LAM=89.0, LB=61.0, LCB=53.0, LCM=82.0, LDM=62.0, LF=91.0, LM=89.0, LS=92.0, LW=91.0, LWB=66.0, Preferred Positions='ST LW ', RAM=89.0, RB=61.0, RCB=53.0, RCM=82.0, RDM=62.0, RF=91.0, RM=89.0, RS=92.0, RW=91.0, RWB=66.0, ST=92.0)


Row(_c0=1, Name='L. Messi', Age=30, Photo='https://cdn.sofifa.org/48/18/players/158023.png', Nationality='Argentina', Flag='https://cdn.sofifa.org/flags/52.png', Overall=93, Potential=93, Club='FC Barcelona', Club Logo='https://cdn.sofifa.org/24/18/teams/241.png', Value='€105M', Wage='€565K', Special=2154, Acceleration='92', Aggression='48', Agility='90', Balance='95', Ball control='95', Composure='96', Crossing='77', Curve='89', Dribbling='97', Finishing='95', Free kick accuracy='90', GK diving='6', GK handling='11', GK kicking='15', GK positioning='14', GK reflexes='8', Heading accuracy='71', Interceptions='22', Jumping='68', Long passing='87', Long shots='88', Marking='13', Penalties='74', Positioning='93', Reactions='95', Short passing='88', Shot power='85', Sliding tackle='26', Sprint speed='87', Stamina='73', Standing tackle='28', Strength='59', Vision='90', Volleys='85', CAM=92.0, CB=45.0, CDM=59.0, CF=92.0, CM=84.0, ID=158023, LAM=92.0, LB=57.0, LCB=45.0, LCM=84.0, LDM=59.0, LF=92.0, LM=90.0, LS=88.0, LW=91.0, LWB=62.0, Preferred Positions='RW ', RAM=92.0, RB=57.0, RCB=45.0, RCM=84.0, RDM=59.0, RF=92.0, RM=90.0, RS=88.0, RW=91.0, RWB=62.0, ST=88.0)
```

In [9]:
```
# Print the total number of rows
print("There are {} rows in the fifa_df DataFrame".format(fifa_df.count()))
```

```
There are 17981 rows in the fifa_df DataFrame
```

In [10]:
```
#3. Create a temporary view of fifa_df
fifa_df.createOrReplaceTempView('fifa_table')
```

In [11]:
```
# 4.
# Construct the "query"
query = '''SELECT Age FROM fifa_table WHERE Nationality == "Germany"'''
# Apply the SQL "query"
fifa_germany_age = spark.sql(query)
```

In [12]:
```
fifa_germany_age.show(3)
```

```
+---+
|Age|
+---+
| 31|
| 27|
| 28|
+---+
only showing top 3 rows
```

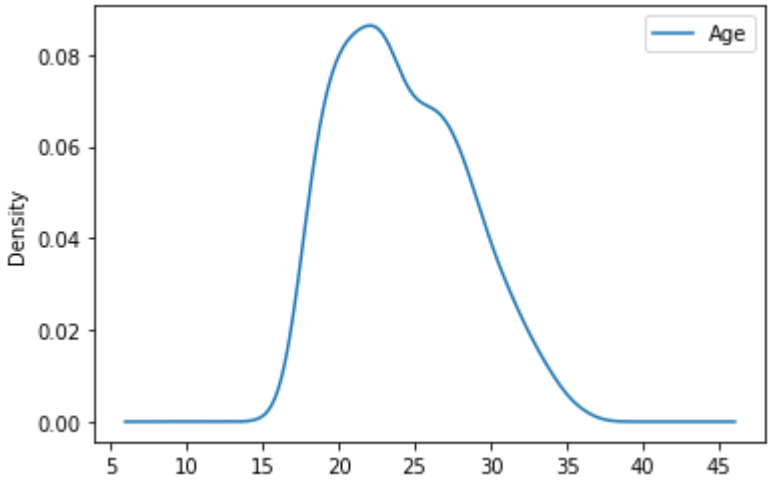
In [13]:
```
# Generate basic statistics
fifa_germany_age.describe().show()
```

```
+-------+-----------------+
|summary|              Age|
+-------+-----------------+
|  count|             1140|
|   mean|24.20263157894737|
| stddev|4.197096712293752|
|    min|               16|
|    max|               36|
+-------+-----------------+
```

In [14]:
```
#5.
# Convert fifa_germany_age to fifa_germany_age_pandas DataFrame
fifa_germany_age_pandas = fifa_germany_age.toPandas()
```

In [15]:
```
import matplotlib.pyplot as plt
```

In [16]:
```
# Plot the 'Age' density of Germany Players
fifa_germany_age_pandas.plot(kind='density')
plt.show()
```



In [17]:
```
#6.
fifa_df.groupBy("Age").count().show()
```

```
+---+-----+
|Age|count|
+---+-----+
| 31|  671|
| 34|  272|
| 28| 1051|
| 26| 1202|
| 27| 1152|
| 44|    2|
| 22| 1324|
| 47|    1|
| 16|   13|
| 20| 1245|
| 40|    8|
| 19| 1069|
| 41|    3|
| 43|    2|
| 37|   69|
| 17|  258|
| 35|  191|
| 39|   20|
| 23| 1394|
| 38|   36|
+---+-----+
only showing top 20 rows
```

In [18]:
```
fifa_df.groupBy("Nationality").avg("Age").show()
```

```
+--------------+------------------+
|   Nationality|          avg(Age)|
+--------------+------------------+
|          Chad|              25.0|
|        Russia| 25.23202614379085|
|      Paraguay| 26.10144927536232|
|       Senegal|25.046511627906977|
|        Sweden|25.119565217391305|
|        Guyana|              28.0|
|       Eritrea|              32.0|
|   Philippines|25.666666666666668|
|          Fiji|              29.0|
|        Turkey|25.127147766323024|
|          Iraq|              26.0|
|       Germany| 24.20263157894737|
|St Kitts Nevis|26.666666666666668|
|       Comoros| 27.11111111111111|
|   Afghanistan|              22.0|
|   Ivory Coast| 24.10891089108911|
|         Sudan|              22.5|
|        France|24.634969325153374|
|        Greece|24.418367346938776|
|        Kosovo|           23.9375|
+--------------+------------------+
only showing top 20 rows
```

In [19]:
```
#7.
from pyspark.sql import functions as F
fifa_df.groupBy("Nationality").agg(F.min("Age"),
                                   F.max("Age"),
                                   F.min("Strength"),
                                   F.max("Strength")).show()
```

```
+--------------+--------+--------+-------------+-------------+
|   Nationality|min(Age)|max(Age)|min(Strength)|max(Strength)|
+--------------+--------+--------+-------------+-------------+
|          Chad|      24|      26|           73|           79|
|      Paraguay|      18|      37|           33|           91|
|        Russia|      17|      37|           26|           93|
|       Senegal|      18|      34|           37|           94|
|        Sweden|      17|      37|           21|           91|
|        Guyana|      25|      34|           47|           75|
|       Eritrea|      32|      32|           85|           85|
|   Philippines|      22|      28|           42|           76|
|          Fiji|      29|      29|           57|           57|
|        Turkey|      17|      39|           30|           90|
|          Iraq|      21|      30|           55|           86|
|       Germany|      16|      36|           28|           94|
|St Kitts Nevis|      23|      32|           32|           84|
|       Comoros|      23|      32|           28|           82|
|   Afghanistan|      19|      27|           40|           58|
|   Ivory Coast|      17|      34|           38|           92|
|         Sudan|      22|      23|           41|           62|
|        France|      16|      40|           26|           93|
|        Greece|      18|      38|           32|           91|
|        Kosovo|      18|      33|           30|           90|
+--------------+--------+--------+-------------+-------------+
only showing top 20 rows
```

In [20]:
```
#8.
fifa_df.select("Club").distinct().show()
```

```
+--------------------+
|                Club|
+--------------------+
|             Palermo|
|         Yeovil Town|
|  1. FC Union Berlin|
|  Santiago Wanderers|
|               Carpi|
|Evkur Yeni Malaty...|
|          Sagan Tosu|
|            FC Basel|
|  Argentinos Juniors|
|       Karlsruher SC|
|  Lorca Deportiva CF|
|     SC Paderborn 07|
|     Cheltenham Town|
|San Lorenzo de Al...|
|         SC Freiburg|
|  SpVgg Unterhaching|
|Atletico Nacional...|
|Universidad Católica|
|         GFC Ajaccio|
|           FC Luzern|
+--------------------+
only showing top 20 rows
```

In [21]:
```
query = '''SELECT DISTINCT Club FROM fifa_table'''
# Apply the SQL "query"
fifa_clubs = spark.sql(query)
fifa_clubs.show()
```

```
+--------------------+
|                Club|
+--------------------+
|             Palermo|
|         Yeovil Town|
|  1. FC Union Berlin|
|  Santiago Wanderers|
|               Carpi|
|Evkur Yeni Malaty...|
|          Sagan Tosu|
|            FC Basel|
|  Argentinos Juniors|
|       Karlsruher SC|
|  Lorca Deportiva CF|
|     SC Paderborn 07|
|     Cheltenham Town|
|San Lorenzo de Al...|
|         SC Freiburg|
|  SpVgg Unterhaching|
|Atletico Nacional...|
|Universidad Católica|
|         GFC Ajaccio|
|           FC Luzern|
+--------------------+
only showing top 20 rows
```

In [22]:
```
#9.
fifa_df_desc = fifa_df.orderBy(fifa_df["Age"].desc())
```

In [23]:
```
fifa_df_desc.select("Name", "Age", "Strength").show(3)
```

```
+-------------+---+--------+
|         Name|Age|Strength|
+-------------+---+--------+
|B. Richardson| 47|      47|
| E. El Hadary| 44|      73|
|     O. Pérez| 44|      66|
+-------------+---+--------+
only showing top 3 rows
```

In [24]:
```
#10.
people_with_Cristiano = fifa_df.where(fifa_df["Name"].contains("Cristiano"))
people_with_Cristiano.count()
```

Out[24]: 3

In [25]:
```
people_with_Cristiano.select("Name").show()
```

```
+----------------+
|            Name|
+----------------+
|Cristiano Ronaldo|
|        Cristiano|
|        Cristiano|
+----------------+
```

In [26]:
```
query = '''SELECT * FROM fifa_table WHERE Name like "%Cristiano%"'''
people_with_Cristiano_2 = spark.sql(query)
people_with_Cristiano_2.count()
```

Out[26]: 3

In [27]:
```
people_with_Cristiano_2.select("Name").show()
```

```
+----------------+
|            Name|
+----------------+
|Cristiano Ronaldo|
|        Cristiano|
|        Cristiano|
+----------------+
```