University of East London School of Architecture, Computing and Engineering (ACE)
Author: Dr Fahimeh Jafari (f.jafari@uel.ac.uk) 25-26

Tutorial 6: Apache Spark SQL

CN7031 - Big Data Analytics

Dr Fahimeh Jafari (f.jafari@uel.ac.uk)

LEARNING OUTCOMES: After completing this tutorial, you should:

- Have gotten a hands-on experience in deploying PySpark
- Practice Spark commands
- Be able to analyse data using Spark SQL queriesp
- Be able to visualise the output in Python

Tutorial Submission [Mandatory]

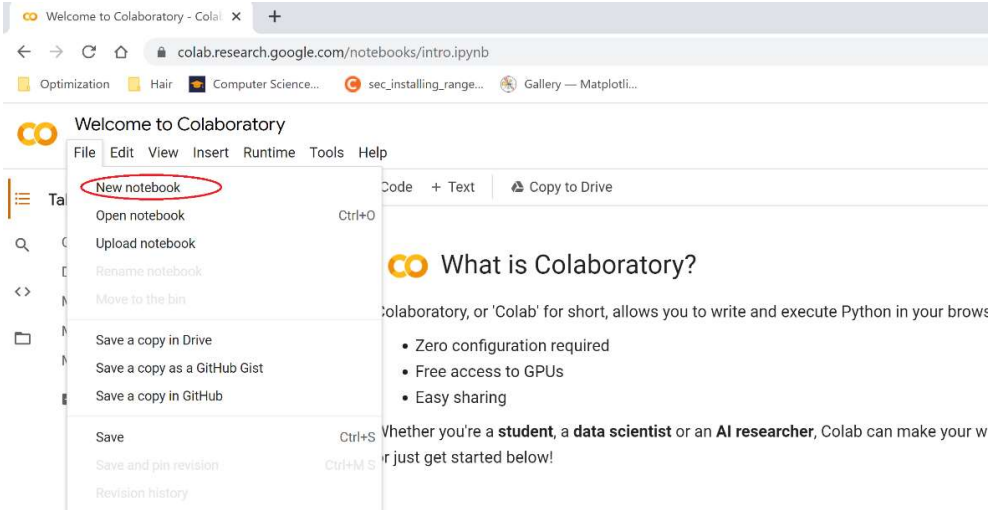
You must submit your `.ipynb` file through the submission link provided in Moodle. In jupyter, your `.ipynb` file will be created automatically. In Google Colab, you can download it through `File-> download .ipynb`

Task 1: Getting Ready

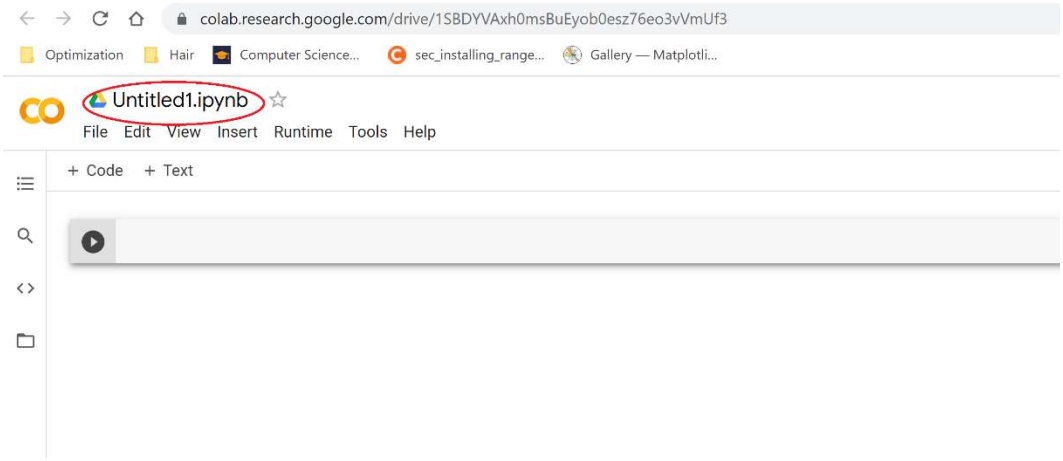
You can execute PySpark commands using Jupyter in VMWare OR Google *Colab* which is an online platform. Please go to the correct section due to your notebook.

Google Colab

- Open the Google Colab through https://colab.research.google.com/
- Open a new notebook to type the commands.



- Click on the name box (see the red circle) and rename your file to `FIFA18`.

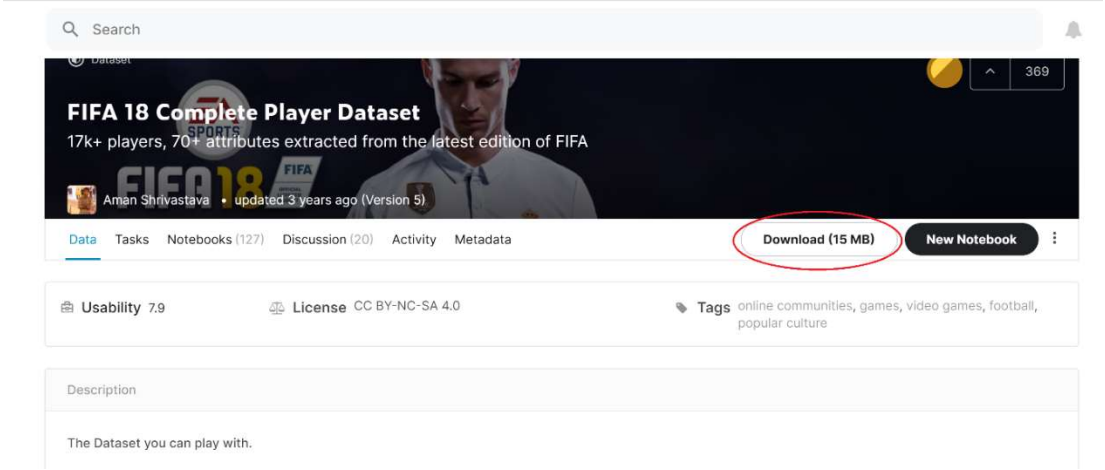


- Download data (FIFA 18 Player Dataset) through the following link. You may need to copy & past the link into your browser to work.

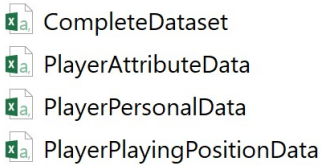
  https://tinyurl.com/y57wxuht

  Alternatively, if you are currently in the lab rooms, you can access the files through the following directory:

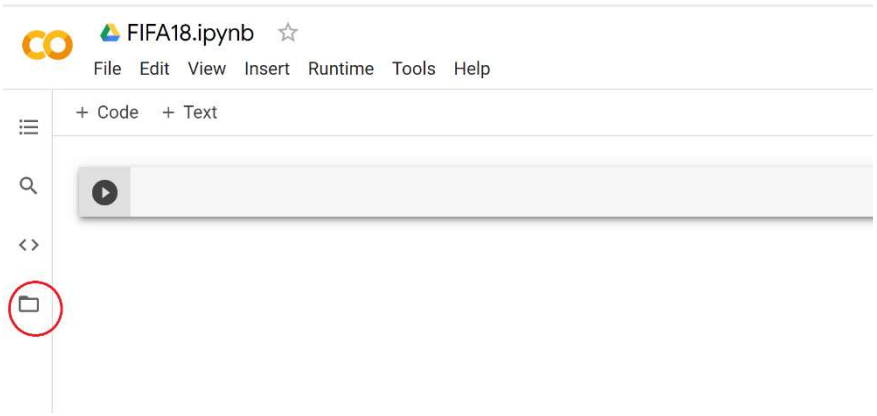
  C:\acesource\fifa18



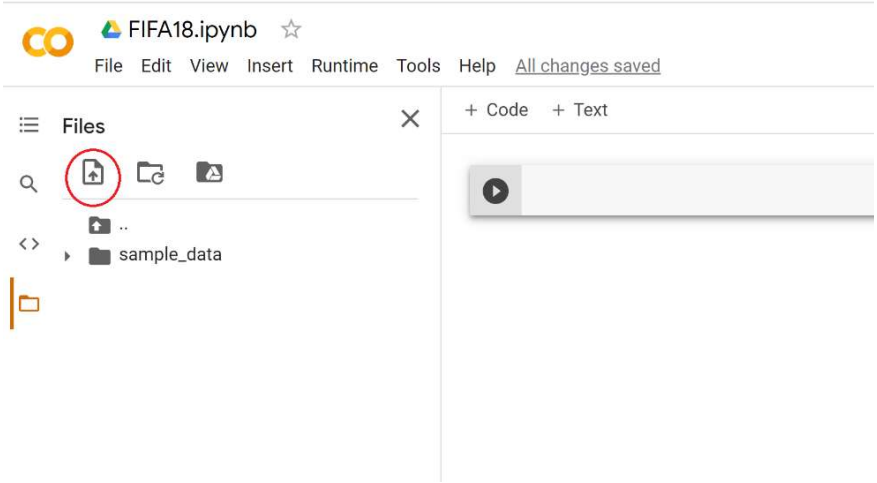
- Extract "`Archive.zip`" by Right-clicking and selecting "`Extract All…`". Now, you will have four `csv` files as below. In this tutorial, we are going to work with `CompleteDataset.csv` file.



- Upload `CompleteDataset.csv` file into Google Colab. Click on `Files` on the left side (red cycle in the screenshot).



- Then, click on `upload to session storage` (red cycle in the screenshot) and upload `CompleteDataset.csv` file from your computer.



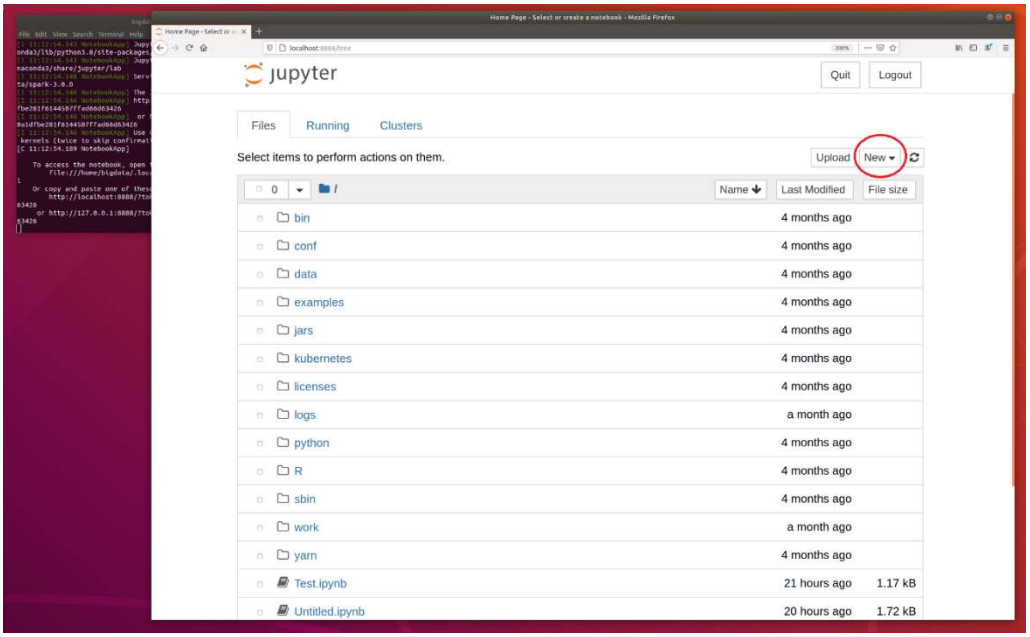
- Type and run the following commands in the editor to install and lunch Spark.

```
# [1] download and install pyspark in Google Colab
!pip3 install pyspark
```

Now, go to Task 2

Jupyter in VMWare

- Run VMWare machine with `Big Data-Ubuntu`
- Open the terminal and go to Spark directory by typing `cd $SPARK_HOME`
- Type `pyspark` in the terminal to launch PySpark and Jupyter.
- Open a new notebook in Jupyter through `New->Python3` (red cycle in the screenshot). Now, your editor is ready for programming.

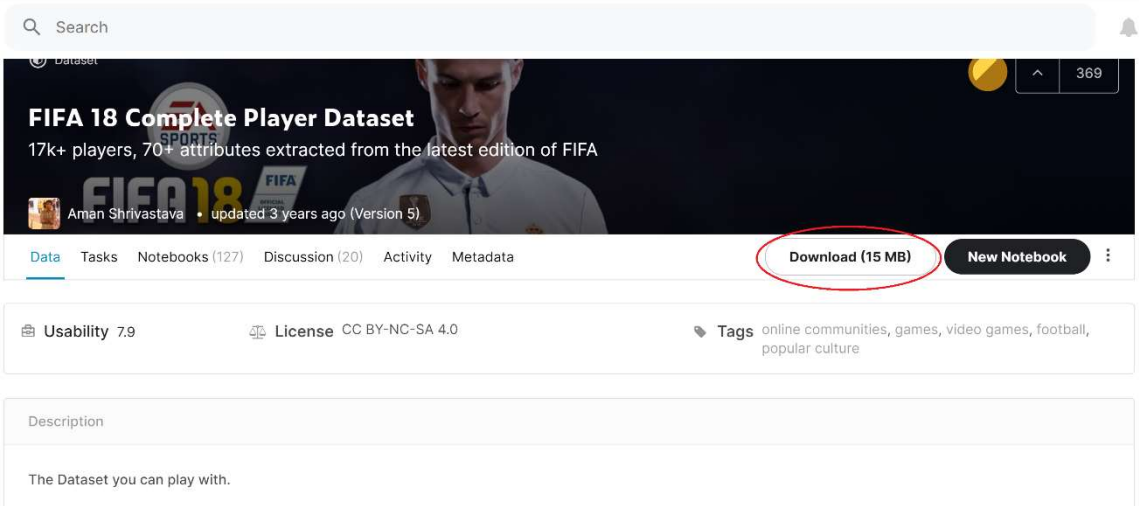


- Download data (FIFA 18 Player Dataset) through the following link. You may need to copy & past the link into your browser to work.

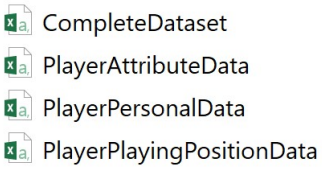
  https://tinyurl.com/y57wxuht

  Alternatively, if you are currently in the lab rooms, you can access the files through the following directory:

  C:\acesource\fifa18



- Extract "`Archive.zip`" by Right-clicking and selecting "`Extract All…`". Now, you will have four `csv` files as below. In this tutorial, we are going to work with `CompleteDataset.csv` file.



- Drag and drop `CompleteDataset.csv` file into your *Desktop*

Task 2: Reading Data from the csv file

Go to the editor and follow the steps to create DataFrame based on the content of `CompleteDataset.csv` file.

The entry point into all functionality in Spark is the `SparkSession` class. So, type the following commands to create a basic `SparkSession`.

```
from pyspark.sql import SparkSession
spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

You can create DataFrame and display the content using the following commands.

NOTE: Please type and run the commands due to your editor.

Google Colab

```
fifa_df = spark.read.load("CompleteDataset.csv", format="csv", inferSchema=True,
 header=True)

fifa_df.show()
```

```
+---+----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------------------+-----
|_c0|            Name|Age|               Photo|Nationality|                Flag|Overall|Potential|                Club|           Club Logo| Valu
+---+----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------------------+-----
|  0|Cristiano Ronaldo| 32|https://cdn.sofif...|   Portugal|https://cdn.sofif...|     94|       94|     Real Madrid CF|https://cdn.sofif...|€95.5
|  1|        L. Messi| 30|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     93|       93|       FC Barcelona|https://cdn.sofif...| €105
|  2|          Neymar| 25|https://cdn.sofif...|     Brazil|https://cdn.sofif...|     92|       94|Paris Saint-Germain|https://cdn.sofif...| €123
|  3|       L. Suárez| 30|https://cdn.sofif...|    Uruguay|https://cdn.sofif...|     92|       92|       FC Barcelona|https://cdn.sofif...|  €97
|  4|        M. Neuer| 31|https://cdn.sofif...|    Germany|https://cdn.sofif...|     92|       92|   FC Bayern Munich|https://cdn.sofif...|  €61
|  5|  R. Lewandowski| 28|https://cdn.sofif...|     Poland|https://cdn.sofif...|     91|       91|   FC Bayern Munich|https://cdn.sofif...|  €92
|  6|          De Gea| 26|https://cdn.sofif...|      Spain|https://cdn.sofif...|     90|       92| Manchester United|https://cdn.sofif...|€64.5
|  7|       E. Hazard| 26|https://cdn.sofif...|    Belgium|https://cdn.sofif...|     90|       91|             Chelsea|https://cdn.sofif...|€90.5
|  8|        T. Kroos| 27|https://cdn.sofif...|    Germany|https://cdn.sofif...|     90|       90|     Real Madrid CF|https://cdn.sofif...|  €75
|  9|      G. Higuaín| 29|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     90|       90|           Juventus|https://cdn.sofif...|  €77
| 10|    Sergio Ramos| 31|https://cdn.sofif...|      Spain|https://cdn.sofif...|     90|       90|     Real Madrid CF|https://cdn.sofif...|  €52
| 11|    K. De Bruyne| 26|https://cdn.sofif...|    Belgium|https://cdn.sofif...|     89|       92|    Manchester City|https://cdn.sofif...|  €83
| 12|     T. Courtois| 25|https://cdn.sofif...|    Belgium|https://cdn.sofif...|     89|       92|             Chelsea|https://cdn.sofif...|  €59
| 13|      A. Sánchez| 28|https://cdn.sofif...|      Chile|https://cdn.sofif...|     89|       89|             Arsenal|https://cdn.sofif...|€67.5
| 14|       L. Modrić| 31|https://cdn.sofif...|    Croatia|https://cdn.sofif...|     89|       89|     Real Madrid CF|https://cdn.sofif...|  €57
| 15|         G. Bale| 27|https://cdn.sofif...|      Wales|https://cdn.sofif...|     89|       89|     Real Madrid CF|https://cdn.sofif...|€69.5
| 16|       S. Agüero| 29|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     89|       89|    Manchester City|https://cdn.sofif...|€66.5
| 17|   G. Chiellini| 32|https://cdn.sofif...|      Italy|https://cdn.sofif...|     89|       89|           Juventus|https://cdn.sofif...|  €38
| 18|       G. Buffon| 39|https://cdn.sofif...|      Italy|https://cdn.sofif...|     89|       89|           Juventus|https://cdn.sofif...|  €4.5
| 19|       P. Dybala| 23|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     88|       93|           Juventus|https://cdn.sofif...|  €79
+---+----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------------------+-----
only showing top 20 rows
```

Jupyter in VMWare

```
fifa_df = spark.read.load("/home/bigdata/Desktop/CompleteDataset.csv",
format="csv", inferSchema=True, header=True)
fifa_df.show()
```

```
+---+-----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------
------------+------+-----+-------+------------+----------+-------+-------+------------+---------+--------+-----+---------+-----
----+-----------------+--------+----------+-----------+---------+-------------+---------+--------------+-----------+-------+----
--------+----------+-------+---------+-----------+---------+------------+---------+--------------+-----------+------+------
---------+--------+------+--------+----+----+----+----+----+------+----+----+----+----+----+----+----+----+----+----+----------
--------+----+----+----+----+----+----+----+----+----+----+----+
|_c0|             Name|Age|               Photo|Nationality|               Flag|Overall|Potential|                Club|
Club Logo| Value| Wage|Special|Acceleration|Aggression|Agility|Balance|Ball control|Composure|Crossing|Curve|Dribbling|Finishin
g|Free kick accuracy|GK diving|GK handling|GK kicking|GK positioning|GK reflexes|Heading accuracy|Interceptions|Jumping|Long pa
ssing|Long shots|Marking|Penalties|Positioning|Reactions|Short passing|Shot power|Sliding tackle|Sprint speed|Stamina|Standing
tackle|Strength|Vision|Volleys| CAM|  CB| CDM|  CF|  CM|    ID| LAM|  LB| LCB| LCM| LDM|  LF|  LM|  LS|  LW| LWB|Preferred Posi
tions| RAM|  RB| RCB| RCM| RDM|  RF|  RM|  RS|  RW| RWB|  ST|
+---+-----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------
------------+------+-----+-------+------------+----------+-------+-------+------------+---------+--------+-----+---------+-----
----+-----------------+--------+----------+-----------+---------+-------------+---------+--------------+-----------+-------+----
--------+----------+-------+---------+-----------+---------+------------+---------+--------------+-----------+------+------
---------+--------+------+--------+----+----+----+----+----+------+----+----+----+----+----+----+----+----+----+----+----------
--------+----+----+----+----+----+----+----+----+----+----+----+
|  0|Cristiano Ronaldo| 32|https://cdn.sofif...|   Portugal|https://cdn.sofif...|     94|       94|     Real Madrid CF|https://
cdn.sofif...|€95.5M|€565K|   2228|          89|        63|     89|     63|          93|       95|      85|   81|       91|
94|               76|       7|        11|       15|          14|         11|             88|           29|     95|
77|       92|     22|       85|        95|       96|           83|        94|            23|          91|     92|
31|      80|    85|     88|89.0|53.0|62.0|91.0|82.0| 20801|89.0|61.0|53.0|82.0|62.0|91.0|89.0|92.0|91.0|66.0|             ST LW
|89.0|61.0|53.0|82.0|62.0|91.0|89.0|92.0|91.0|66.0|92.0|
|  1|         L. Messi| 30|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     93|       93|       FC Barcelona|https://
cdn.sofif...| €105M|€565K|   2154|          92|        48|     90|     95|          95|       96|      77|   89|       97|
95|               90|       6|        11|       15|          14|          8|             71|           22|     68|
87|       88|     13|       74|        93|       95|           88|        85|            26|          87|     73|
28|      59|    90|     85|92.0|45.0|59.0|92.0|84.0|158023|92.0|57.0|45.0|84.0|59.0|92.0|90.0|88.0|91.0|62.0|                RW
|92.0|57.0|45.0|84.0|59.0|92.0|90.0|88.0|91.0|62.0|88.0|
|  2|           Neymar| 25|https://cdn.sofif...|     Brazil|https://cdn.sofif...|     92|       94|Paris Saint-Germain|https://
cdn.sofif...| €123M|€280K|   2100|          94|        56|     96|     82|          95|       92|      75|   81|       96|
89|               84|       9|         9|       15|          15|         11|             62|           36|     61|
```

Task 3: Work with DataFrame Operations

In this step, we practice some of DataFrame operations explained in the lecture.

a) See the structure of the DataFrame by typing:

```
fifa_df.printSchema()
```

```
root
 |-- _c0: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Photo: string (nullable = true)
 |-- Nationality: string (nullable = true)
 |-- Flag: string (nullable = true)
 |-- Overall: integer (nullable = true)
 |-- Potential: integer (nullable = true)
 |-- Club: string (nullable = true)
 |-- Club Logo: string (nullable = true)
 |-- Value: string (nullable = true)
 |-- Wage: string (nullable = true)
 |-- Special: integer (nullable = true)
 |-- Acceleration: string (nullable = true)
 |-- Aggression: string (nullable = true)
 |-- Agility: string (nullable = true)
 |-- Balance: string (nullable = true)
 |-- Ball control: string (nullable = true)
 |-- Composure: string (nullable = true)
 |-- Crossing: string (nullable = true)
 |-- Curve: string (nullable = true)
 |-- Dribbling: string (nullable = true)
 |-- Finishing: string (nullable = true)
 |-- Free kick accuracy: string (nullable = true)
 |-- GK diving: string (nullable = true)
 |-- GK handling: string (nullable = true)
 |-- GK kicking: string (nullable = true)
 |-- GK positioning: string (nullable = true)
 |-- GK reflexes: string (nullable = true)
 |-- Heading accuracy: string (nullable = true)
 |-- Interceptions: string (nullable = true)
 |-- Jumping: string (nullable = true)
 |-- Long passing: string (nullable = true)
 |-- Long shots: string (nullable = true)
 |-- Marking: string (nullable = true)
 |-- Penalties: string (nullable = true)
```

b) Collect some information about columns through the following commands.

```
fifa_df.columns
```

```
['_c0',
 'Name',
 'Age',
 'Photo',
 'Nationality',
 'Flag',
 'Overall',
 'Potential',
 'Club',
 'Club Logo',
 'Value',
 'Wage',
 'Special',
 'Acceleration',
 'Aggression',
 'Agility',
 'Balance',
 'Ball control',
 'Composure',
 'Crossing',
 'Curve',
 'Dribbling',
 'Finishing',
 'Free kick accuracy',
 'GK diving',
 'GK handling',
 'GK kicking',
 'GK positioning',
 'GK reflexes',
 'Heading accuracy',
 'Interceptions',
 'Jumping',
 'Long passing',
```

```
fifa_df.count()
```

17981

```
len (fifa_df.columns)
```

75

```
fifa_df.select('Name','Nationality','club').show()
```

```
+----------------+-----------+------------------+
|            Name|Nationality|              club|
+----------------+-----------+------------------+
|Cristiano Ronaldo|   Portugal|     Real Madrid CF|
|         L. Messi|  Argentina|       FC Barcelona|
|           Neymar|     Brazil|Paris Saint-Germain|
|        L. Suárez|    Uruguay|       FC Barcelona|
|         M. Neuer|    Germany|   FC Bayern Munich|
|   R. Lewandowski|     Poland|   FC Bayern Munich|
|          De Gea|      Spain|  Manchester United|
|        E. Hazard|    Belgium|           Chelsea|
|         T. Kroos|    Germany|     Real Madrid CF|
|       G. Higuaín|  Argentina|          Juventus|
|     Sergio Ramos|      Spain|     Real Madrid CF|
|     K. De Bruyne|    Belgium|    Manchester City|
|      T. Courtois|    Belgium|           Chelsea|
|       A. Sánchez|      Chile|           Arsenal|
|        L. Modrić|    Croatia|     Real Madrid CF|
|          G. Bale|      Wales|     Real Madrid CF|
|        S. Agüero|  Argentina|    Manchester City|
|     G. Chiellini|      Italy|          Juventus|
|        G. Buffon|      Italy|          Juventus|
|        P. Dybala|  Argentina|          Juventus|
+----------------+-----------+------------------+
only showing top 20 rows
```

```
fifa_df.select('Name','Long shots').distinct().show()
```

```
+----------------+----------+
|            Name|Long shots|
+----------------+----------+
|Cristiano Ronaldo|        92|
|      J. Cuadrado|        80|
|      M. Brozović|        79|
|          A. Rami|        58|
|       D. Abraham|        65|
|     Borja Bastón|        73|
|       J. Montero|        68|
|      T. Barnetta|        74|
|          Wallace|        26|
|       A. Barreca|        42|
|    Y. Benalouane|        39|
|          Juankar|        64|
|        D. Appiah|        38|
|   Rafael Martins|        69|
|          Granell|        77|
|     A. Cornelius|        68|
|         J. Henry|        75|
|        M. Ozdoev|        69|
|            Fábio|        58|
|       T. Dingomé|        60|
+----------------+----------+
only showing top 20 rows
```

c) Apply a filter on `fifa_df` DataFrame to select people older than 21.

```
fifa_df.filter(fifa_df['age'] > 21).show()
```

```
+---+-----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------
-----------+------+-----+-------+------------+----------+-------+-------+-----------+---------+--------+-----+---------+-----
----+-----------------+----------+---------+-----------+-------------+----------+--------------+------------+-------+------
--------+----------+-------+---------+-----------+---------+-------------+----------+--------------+------------+-------+------
---------+--------+------+-------+----+----+----+----+----+------+----+----+----+----+----+----+----+----+----+----+-----------
--------+----+----+----+----+----+----+----+----+----+----+----+----+
|_c0|             Name|Age|               Photo|Nationality|               Flag|Overall|Potential|                Club|
Club Logo| Value| Wage|Special|Acceleration|Aggression|Agility|Balance|Ball control|Composure|Crossing|Curve|Dribbling|Finishin
g|Free kick accuracy|GK diving|GK handling|GK kicking|GK positioning|GK reflexes|Heading accuracy|Interceptions|Jumping|Long pa
ssing|Long shots|Marking|Penalties|Positioning|Reactions|Short passing|Shot power|Sliding tackle|Sprint speed|Stamina|Standing
tackle|Strength|Vision|Volleys| CAM|  CB| CDM|  CF|  CM|   ID| LAM|  LB| LCB| LCM| LDM|  LF|  LM|  LS|  LW| LWB|Preferred Posi
tions| RAM|  RB| RCB| RCM| RDM|  RF|  RM|  RS|  RW| RWB|  ST|
+---+-----------------+---+--------------------+-----------+--------------------+-------+---------+--------------------+--------
-----------+------+-----+-------+------------+----------+-------+-------+-----------+---------+--------+-----+---------+-----
----+-----------------+----------+---------+-----------+-------------+----------+--------------+------------+-------+------
--------+----------+-------+---------+-----------+---------+-------------+----------+--------------+------------+-------+------
---------+--------+------+-------+----+----+----+----+----+------+----+----+----+----+----+----+----+----+----+----+-----------
--------+----+----+----+----+----+----+----+----+----+----+----+----+
|  0|Cristiano Ronaldo| 32|https://cdn.sofif...|   Portugal|https://cdn.sofif...|     94|       94|     Real Madrid CF|https://
cdn.sofif...|€95.5M|€565K|   2228|          89|        63|    89|     63|          93|       95|      85|   81|       91|
94|               76|        7|         11|         15|           14|         11|           88|           29|     95|
77|       92|     22|       85|         95|       96|          83|        94|           23|          91|     92|
31|      80|    85|     88|89.0|53.0|62.0|91.0|82.0| 20801|89.0|61.0|53.0|82.0|62.0|91.0|89.0|92.0|91.0|66.0|             ST LW
|89.0|61.0|53.0|82.0|62.0|91.0|89.0|92.0|91.0|66.0|92.0|
|  1|        L. Messi| 30|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     93|       93|       FC Barcelona|https://
cdn.sofif...| €105M|€565K|   2154|          92|        48|    90|     95|          95|       96|      77|   89|       97|
95|               90|        6|         11|         15|           14|          8|           71|           22|     68|
87|       88|     13|       74|         93|       95|          88|        85|           26|          87|     73|
28|      59|    90|     85|92.0|45.0|59.0|92.0|84.0|158023|92.0|57.0|45.0|84.0|59.0|92.0|90.0|88.0|91.0|62.0|                RW
|92.0|57.0|45.0|84.0|59.0|92.0|90.0|88.0|91.0|62.0|88.0|
|  2|           Neymar| 25|https://cdn.sofif...|     Brazil|https://cdn.sofif...|     92|       94|Paris Saint-Germain|https://
cdn.sofif...| €123M|€280K|   2100|          94|        56|    96|     82|          95|       92|      75|   81|       96|
89|               84|        9|          9|         15|           15|         11|           62|           36|     61|
75|       77|     21|       81|         90|       88|          81|        80|           33|          90|     78|
24|      53|    80|     83|88.0|46.0|59.0|88.0|79.0|190871|88.0|59.0|46.0|79.0|59.0|88.0|87.0|84.0|89.0|64.0|                LW
|88.0|59.0|46.0|79.0|59.0|88.0|87.0|84.0|89.0|64.0|84.0|
```

- Count the number of players by age.

```
fifa_df.groupBy("age").count().show()
```

```
+---+-----+
|age|count|
+---+-----+
| 29| 1121|
| 30|  804|
| 34|  272|
| 28| 1051|
| 22| 1324|
| 35|  191|
| 16|   13|
| 47|    1|
| 43|    2|
| 31|  671|
| 18|  672|
| 27| 1152|
| 17|  258|
| 26| 1202|
| 19| 1069|
| 23| 1394|
| 41|    3|
| 38|   36|
| 40|    8|
| 25| 1522|
+---+-----+
only showing top 20 rows
```

Task 4: SQL Queries and Visualisation

To create SQL queries, you need to first register the DataFrame as a SQL temporary view and then define the queries on the view.

```
# Register the DataFrame as a SQL temporary view
fifa_df.createOrReplaceTempView("FifaView")
```

Query 1:

Retrieve content of the dataset.

```
sqlDF = spark.sql("SELECT * FROM FifaView")
sqlDF.show()
```

```
15|   14|       30|       13|                11|       91|         90|         95|           91|         89|
25|           30|     78|           59|        16|     10|       47|         12|       85|           55|        25|
11|           61|     44|             10|       83|    70|     11|null|null|null|null|null|167495|null|null|null|nul
l|null|null|null|null|null|null|                 GK |null|null|null|null|null|null|null|null|null|null|null|
|  5|   R. Lewandowski| 28|https://cdn.sofif...|     Poland|https://cdn.sofif...|     91|       91|   FC Bayern Mur
ich|https://cdn.sofif...|  €92M|€355K|   2143|          79|        80|     78|     80|          89|       87|
62|   77|       85|       91|                84|       15|          6|        12|             8|         10|
85|           39|     84|           65|        83|     25|       81|         91|       91|           83|        88|
19|           83|     79|             42|       84|    78|     87|84.0|57.0|62.0|87.0|78.0|188545|84.0|58.0|57.0|78.
0|62.0|87.0|82.0|88.0|84.0|61.0|                 ST |84.0|58.0|57.0|78.0|62.0|87.0|82.0|88.0|84.0|61.0|88.0|
|  6|           De Gea| 26|https://cdn.sofif...|      Spain|https://cdn.sofif...|     90|       92|  Manchester Uni
ted|https://cdn.sofif...|€64.5M|€215K|   1458|          57|        38|     60|     43|          42|       64|
17|   21|       18|       13|                19|       90|         85|         87|             86|         90|
21|           30|     67|           51|        12|     13|       40|         12|       88|           50|        31|
13|           58|     40|             21|       64|    68|     13|null|null|null|null|null|193080|null|null|null|nul
l|null|null|null|null|null|null|                 GK |null|null|null|null|null|null|null|null|null|null|null|
|  7|        E. Hazard| 26|https://cdn.sofif...|    Belgium|https://cdn.sofif...|     90|       91|             Chel
sea|https://cdn.sofif...|€90.5M|€295K|   2096|          93|        54|     93|     91|          92|       87|
80|   82|       93|       83|                79|       11|         12|          6|             8|          8|
```

Query 2:

Count the number of players by age.

```
sqlDF = spark.sql("SELECT age, count(*) as count from FifaView GROUP BY age")
sqlDF.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
```

Type and run the following commands to show the output as a bar graph.

- Import required libraries

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

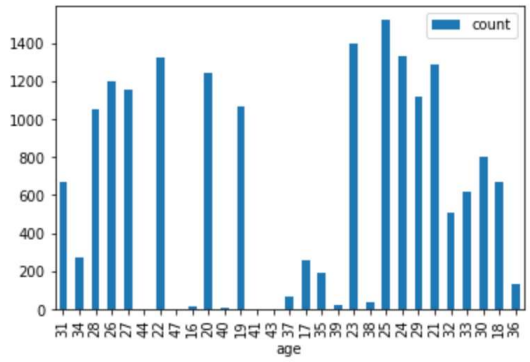
- Convert SQL dataframe to Pandas dataframe

```
pandas_df = sqlDF.toPandas()
```

- Plot the bar chart

```
pandas_df.plot(x ='age', y='count', kind = 'bar')
```

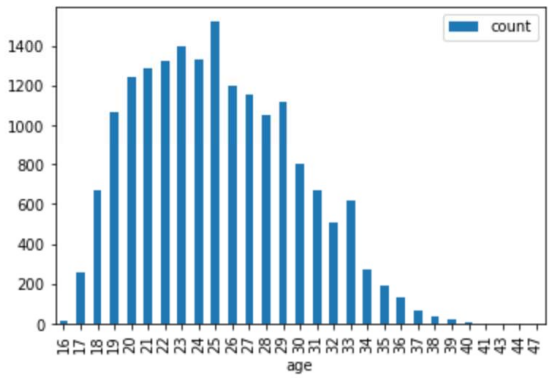
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa24f2c7580>
```



You can order the output by adding the `sort.value ()`

```
pandas_df.sort_values(by='age',ascending=True).plot(x ='age', y='count', kind = 'bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7303ebb6a0>
```



Query 3:

Count the number of players in each club.

```
sqlDF = spark.sql("SELECT club, count(*) FROM FifaView GROUP BY club")
sqlDF.show()
```

```
+--------------------+--------+
|                club|count(1)|
+--------------------+--------+
|             Palermo|      28|
|         Yeovil Town|      21|
|  1. FC Union Berlin|      27|
|  Santiago Wanderers|      20|
|               Carpi|      30|
|Evkur Yeni Malaty...|      30|
|          Sagan Tosu|      25|
|            FC Basel|      25|
|  Argentinos Juniors|      28|
|       Karlsruher SC|      27|
|  Lorca Deportiva CF|      29|
|     SC Paderborn 07|      28|
|     Cheltenham Town|      28|
|San Lorenzo de Al...|      28|
|         SC Freiburg|      32|
|  SpVgg Unterhaching|      28|
```

Query 4:

Count the number of players in each club and displays those have more than 33 members.

```
sqlDF = spark.sql("SELECT club, count(*) FROM FifaView GROUP BY club HAVING Count(*) > 33")
sqlDF.show()
```

```
+-----------------+--------+
|             club|count(1)|
+-----------------+--------+
| Manchester United|      34|
|     UD Las Palmas|      34|
|              null|     248|
|Olympique Lyonnais|      34|
|     VfL Wolfsburg|      34|
|          OGC Nice|      34|
|     Villarreal CF|      35|
|         FC Nantes|      34|
| Borussia Dortmund|      34|
+-----------------+--------+
```

Type and run the following commands to show the output as a bar graph. You don't need to import the libraries as you have done it before.

```
pandas_df = sqlDF.toPandas()
```

```
pandas_df.plot(x ='club', y='count(1)', kind = 'pie')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2ace8480d0>
```