Search and Filter DataFrames in PySpark HW

Now it's time to put what you've learn into action with a homework assignment!

In case you need it again, here is the link to the documentation for the full list available function in pyspark.sql.functions library: http://spark.apache.org/docs/latest/api/python/pyspark.sql.html#module-pyspark.sql.functions

First set up your Spark Session!

Alright so first things first, let's start up our pyspark instance.

In [70]:
```
import findspark
findspark.init()
```

In [7]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName("DfApp").getOrCreate()
```

Read in the DataFrame for this Notebook

We will be continuing to use the fifa19.csv file for this notebook. Make sure that you are writting the correct path to the file.

In [12]:
```
df=spark.read.csv("D:\Fifa.csv",header=True, inferSchema=True)
df.show()
```

```
|_c0|    ID|            Name|Age|               Photo|Nationality|                Flag|Overall|Potential|               Club|          Club Logo| Value| Wage|Special|Preferred Foot|International Reputation|Weak Foot|Skill Moves|     Work Rate| Body Type|Real Face|Position|Jersey Number|    Joined|Loaned From|Contract Valid Until|Height|Weight|  LS|  ST|  RS|  LW|  LF|  CF|  RF|  RW| LAM| CAM| RAM|  LM| LCM|  CM| RCM|  RM| LWB| LDM| CDM| RDM| RWB|  LB| LCB|  CB| RCB|  RB|Crossing|Finishing|HeadingAccuracy|ShortPassing|Volleys|Dribbling|Curve|FKAccuracy|LongPassing|BallControl|Acceleration|SprintSpeed|Agility|Reactions|Balance|ShotPower|Jumping|Stamina|Strength|LongShots|Aggression|Interceptions|Positioning|Vision|Penalties|Composure|Marking|StandingTackle|SlidingTackle|GKDiving|GKHandling|GKKicking|GKPositioning|GKReflexes|Release Clause|
|  0|158023|        L. Messi| 31|https://cdn.sofif...|  Argentina|https://cdn.sofif...|     94|       94|       FC Barcelona|https://cdn.sofif...|€110.5M|€565K|   2202|          Left|                       5|        4|          4|Medium/ Medium|     Messi|      Yes|      RF|           10| 1-Jul-04|       null|                2021|   5'7|159lbs|88+2|88+2|88+2|92+2|93+2|93+2|93+2|92+2|93+2|93+2|93+2|91+2|84+2|84+2|84+2|91+2|64+2|61+2|61+2|61+2|64+2|59+2|47+2|47+2|47+2|59+2|      84|       95|             70|          90|     86|       97|   93|        94|         87|         96|          91|         86|     91|       95|     95|       85|     68|     72|      59|       94|        48|           22|         94|    94|       75|       96|     33|            28|           26|       6|        11|       15|           14|         8|       €226.5M|
|  1| 20801|Cristiano Ronaldo| 33|https://cdn.sofif...|   Portugal|https://cdn.sofif...|     94|       94|           Juventus|https://cdn.sofif...|   €77M|€405K| [illegible]
[illegible]
|  ...| 94|            85|          80|     84|       80|   78|        68|         82|         84|          68|         72|     71|       91|     71|       88|     78|     89|      84|       85|        76|           35|         93|    80|       90|       89|     56|            36|           38|       8|        10|       11|           14|        11|       €160.7M|
| 17|194765|    A. Griezmann| 27|https://cdn.sofif...|     France|https://cdn.sofif...|     89|       90|    Atlético Madrid|https://cdn.sofif...|   €78M|€145K|   2246|          Left|                       4|        3|          4|   High/ High|      Lean|      Yes|     CAM|            7|28-Jul-14|       null|                2023|   5'9|161lbs|86+3|86+3|86+3|87+3|87+3|87+3|87+3|87+3|86+3|86+3|86+3|86+3|80+3|80+3|80+3|86+3|70+3|67+3|67+3|67+3|70+3|67+3|61+3|61+3|61+3|67+3|      82|       90|             84|          83|     87|       88|   84|        78|         76|         90|          88|         85|     90|       90|     80|       80|     90|     83|      62|       82|        69|           35|         91|    83|       79|       87|     59|            47|           48|      14|         8|       14|           13|        14|       €165.8M|
| 18|192448|   M. ter Stegen| 26|https://cdn.sofif...|    Germany|https://cdn.sofif...|     89|       92|       FC Barcelona|https://cdn.sofif...|   €58M|€240K|   1328|         Right|                       3|        4|          1|Medium/ Medium|    Normal|      Yes|      GK|           22| 1-Jul-14|       null|                2022|   6'2|187lbs|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|      15|       14|             11|          36|     14|       17|   18|        12|         42|         18|          38|         50|     37|       85|     43|       22|     79|     35|      79|       10|        43|           22|         11|    69|       25|       69|     25|            13|           10|      87|        85|       88|           85|        90|       €123.3M|
| 19|192119|      T. Courtois| 26|https://cdn.sofif...|    Belgium|https://cdn.sofif...|     89|       90|        Real Madrid|https://cdn.sofif...| €53.5M|€240K|   1311|          Left|                       4|        2|          1|Medium/ Medium|  Courtois|      Yes|      GK|            1| 9-Aug-18|       null|                2024|   6'6|212lbs|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|null|      14|       14|             13|          33|     12|       13|   19|        20|         35|         23|          46|         52|     61|       84|     45|       36|     68|     38|      70|       17|        23|           15|         13|    44|       27|       66|     20|            18|           16|      85|        91|       72|           86|        88|       €113.7M|
only showing top 20 rows
```

About this dataframe

The **fifa19.csv** dataset includes a list of all the FIFA 2019 players and their attributes listed below:

- **General**: Age, Nationality, Overall, Potential, Club
- **Metrics:** Value, Wage
- **Player Descriptive:** Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight
- **Possition:** LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB,
- **Other:** Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes, and Release Clause.

Source: https://www.kaggle.com/karangadiya/fifa19

Use the .toPandas() method to view the first few lines of the dataset so we know what we are working with.

In [14]:
```
df.toPandas()
```

Out[14]:

	_c0	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	...	Composure	Marking	StandingTackle
0	0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona	...	96.0	33.0	28.0
1	1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus	...	95.0	28.0	31.0
2	2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain	...	94.0	27.0	24.0
3	3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United	...	68.0	15.0	21.0
4	4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City	...	88.0	68.0	58.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18202	18202	238813	J. Lundstram	19	https://cdn.sofifa.org/players/4/19/238813.png	England	https://cdn.sofifa.org/flags/14.png	47	65	Crewe Alexandra	...	45.0	40.0	48.0
18203	18203	243165	N. Christoffersson	19	https://cdn.sofifa.org/players/4/19/243165.png	Sweden	https://cdn.sofifa.org/flags/46.png	47	63	Trelleborgs FF	...	42.0	22.0	15.0
18204	18204	241638	B. Worman	16	https://cdn.sofifa.org/players/4/19/241638.png	England	https://cdn.sofifa.org/flags/14.png	47	67	Cambridge United	...	41.0	32.0	13.0
18205	18205	246268	D. Walker-Rice	17	https://cdn.sofifa.org/players/4/19/246268.png	England	https://cdn.sofifa.org/flags/14.png	47	66	Tranmere Rovers	...	46.0	20.0	25.0
18206	18206	246269	G. Nugent	16	https://cdn.sofifa.org/players/4/19/246269.png	England	https://cdn.sofifa.org/flags/14.png	46	66	Tranmere Rovers	...	43.0	40.0	43.0

18207 rows × 89 columns

Now print the schema of the dataset so we can see the data types of all the varaibles.

In [13]:
```
df.printSchema()
```

```
root
 |-- _c0: integer (nullable = true)
 |-- ID: 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)
 |-- Preferred Foot: string (nullable = true)
 |-- International Reputation: integer (nullable = true)
 |-- Weak Foot: integer (nullable = true)
 |-- Skill Moves: integer (nullable = true)
 |-- Work Rate: string (nullable = true)
 |-- Body Type: string (nullable = true)
 |-- Real Face: string (nullable = true)
 |-- Position: string (nullable = true)
 |-- Jersey Number: integer (nullable = true)
[illegible]
```

Now let's get started!

First things first.... import the pyspark sql functions library

Since we know we will be using it a lot.

In [14]:
```
import pyspark.sql.functions as f
```

1. Select the Name and Position of each player in the dataframe

```
df.createOrReplaceTempView("tempview")
res=spark.sql("select Name,Position from tempview")
res.toPandas()
```

	Name	Position
0	L. Messi	RF
1	Cristiano Ronaldo	ST
2	Neymar Jr	LW
3	De Gea	GK
4	K. De Bruyne	RCM
...	...	...
18202	J. Lundstram	CM
18203	N. Christoffersson	ST
18204	B. Worman	ST
18205	D. Walker-Rice	RW
18206	G. Nugent	CM

18207 rows × 2 columns

1.1 Display the same results from above sorted by the players names

```
res=spark.sql("select Name,Position from tempview order by Name")
res.toPandas()
```

	Name	Position
0	A. Abang	ST
1	A. Abdellaoui	LB
2	A. Abdennour	CB
3	A. Abdi	CM
4	A. Abdu Jaber	ST
...	...	...
18202	[illegible]	[illegible]
18203	[illegible]	[illegible]
18204	[illegible]	[illegible]
18205	[illegible]	[illegible]
18206	[illegible]	[illegible]

18207 rows × 2 columns

2. Select only the players who belong to a club begining with FC

```
res=spark.sql("select Name,Club from tempview where Club like 'FC%'")
res.toPandas()
```

	Name	Club
0	L. Messi	FC Barcelona
1	L. Suárez	FC Barcelona
2	R. Lewandowski	FC Bayern München
3	M. ter Stegen	FC Barcelona
4	Sergio Busquets	FC Barcelona
...	...	...
1902	[illegible]	[illegible]
1903	[illegible]	[illegible]
1904	[illegible]	[illegible]
1905	[illegible]	[illegible]
1906	[illegible]	[illegible]

1907 rows × 2 columns

3. Who is the oldest player in the dataset and how old are they?

Display only the name and age of the oldest player.

```
res=spark.sql("select ID,Name,Age from tempview order by Age desc")
res.show(1)
```

```
[illegible]
only showing top 1 row
```

4. Select only the following players from the dataframe:

- L. Messi
- Cristiano Ronaldo

```
res=spark.sql("select Name from tempview where Name like '%L. Messi%' or Name like '%Cristiano Ronaldo%'")
res.toPandas()
```

	Name
0	L. Messi
1	Cristiano Ronaldo

5. Can you select the first character from the Release Clause variable which indicates the currency used?

```
df.select("Release Clause",df["Release Clause"].substr(1,1)).toPandas()
```

	Release Clause	substring(Release Clause, 1, 1)
0	[illegible]	[illegible]
1	[illegible]	[illegible]
2	[illegible]	[illegible]
3	[illegible]	[illegible]
4	[illegible]	[illegible]
...	...	...
18202	[illegible]	[illegible]
18203	[illegible]	[illegible]
18204	[illegible]	[illegible]
18205	[illegible]	[illegible]
18206	[illegible]	[illegible]

18207 rows × 2 columns

6. Can you select the players who are over the age of 40?

```
res=spark.sql("select Name,Age from tempview where Age>40")
res.toPandas()
```

	Name	Age
0	[illegible]	[illegible]
1	[illegible]	[illegible]
2	[illegible]	[illegible]
3	[illegible]	[illegible]
4	[illegible]	[illegible]
5	[illegible]	[illegible]
6	[illegible]	[illegible]
7	[illegible]	[illegible]
8	[illegible]	[illegible]
9	[illegible]	[illegible]

That's it for now.... Great Job!