Coding challenge

Python

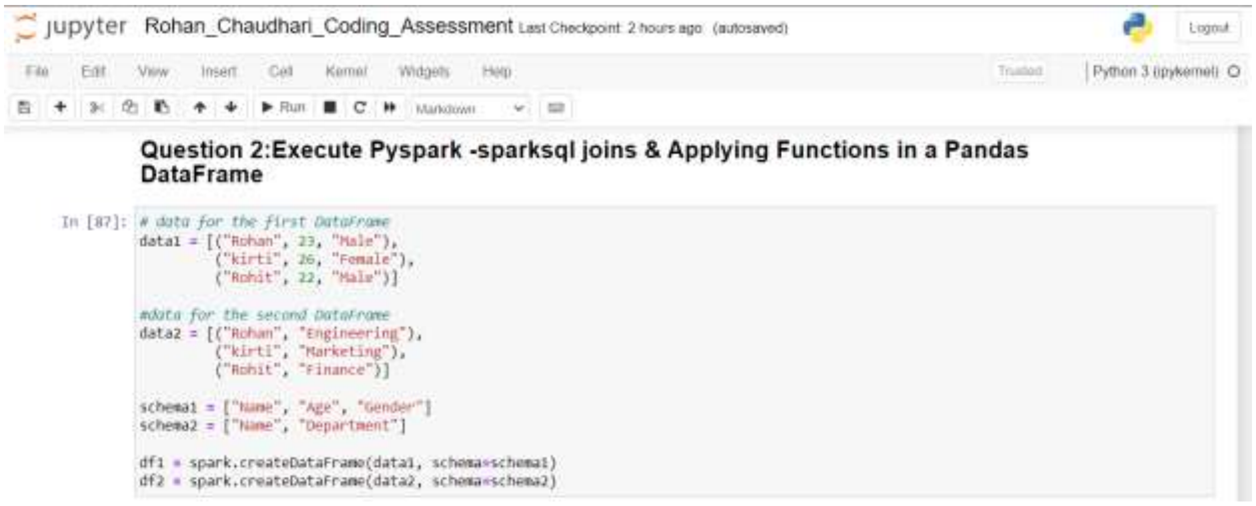
Name: Rohan Vinayak Chaudhari **Email:chaudharirohan24@gmail.com**

Batch: Data Engineering 1

Question2: Execute Pyspark -sparksql joins & Applying Functions in a Pandas DataFrame.

SparkSql:

Created 2 DataFrame df1 & df2 with common column Name to perform sql operations:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown

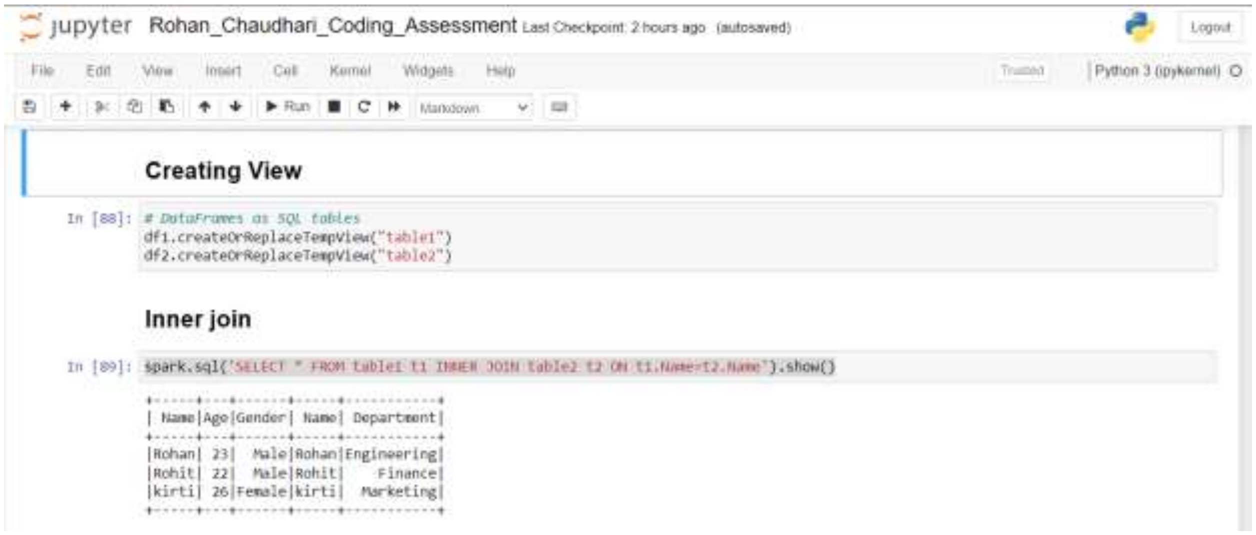
Question 2:Execute Pyspark -sparksql joins & Applying Functions in a Pandas DataFrame

In [87]:

```
# data for the first DataFrame
data1 = [("Rohan", 23, "Male"),
         ("kirti", 26, "Female"),
         ("Rohit", 22, "Male")]

#data for the second DataFrame
data2 = [("Rohan", "Engineering"),
         ("kirti", "Marketing"),
         ("Rohit", "Finance")]

schema1 = ["Name", "Age", "Gender"]
schema2 = ["Name", "Department"]

df1 = spark.createDataFrame(data1, schema=schema1)
df2 = spark.createDataFrame(data2, schema=schema2)
```

Created views To Perform SQL operations on DataFrame:

Performed Inner join with common Column Name:



jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)

Creating View

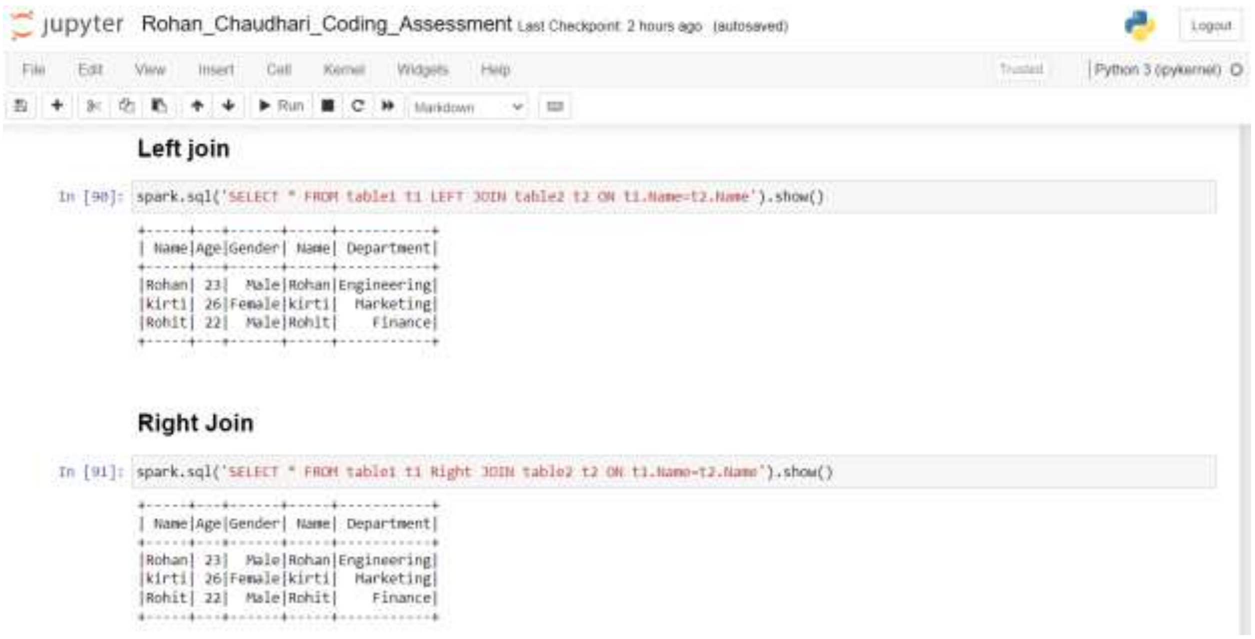
In [88]:
```
# DataFrames as SQL tables
df1.createOrReplaceTempView("table1")
df2.createOrReplaceTempView("table2")
```

Inner join

In [89]:
```
spark.sql('SELECT * FROM table1 t1 INNER JOIN table2 t2 ON t1.Name=t2.Name').show()
```

```
+-----+---+------+-----+-----------+
| Name|Age|Gender| Name| Department|
+-----+---+------+-----+-----------+
|Rohan| 23|  Male|Rohan|Engineering|
|Rohit| 22|  Male|Rohit|    Finance|
|kirti| 26|Female|kirti|  Marketing|
+-----+---+------+-----+-----------+
```

Performed Left & Right join on Both table:



jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)

Left join

In [90]:
```
spark.sql('SELECT * FROM table1 t1 LEFT JOIN table2 t2 ON t1.Name=t2.Name').show()
```

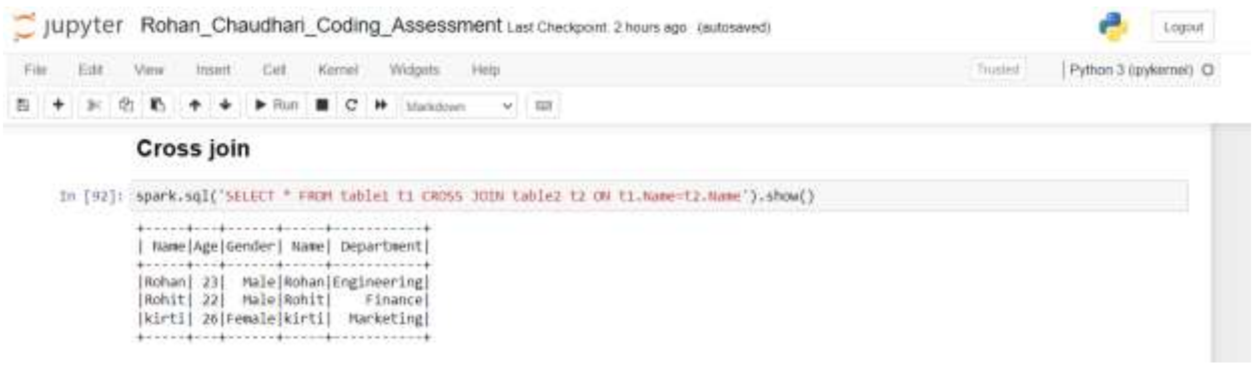
```
+-----+---+------+-----+-----------+
| Name|Age|Gender| Name| Department|
+-----+---+------+-----+-----------+
|Rohan| 23|  Male|Rohan|Engineering|
|kirti| 26|Female|kirti|  Marketing|
|Rohit| 22|  Male|Rohit|    Finance|
+-----+---+------+-----+-----------+
```

Right Join

In [91]:
```
spark.sql('SELECT * FROM table1 t1 Right JOIN table2 t2 ON t1.Name=t2.Name').show()
```

```
+-----+---+------+-----+-----------+
| Name|Age|Gender| Name| Department|
+-----+---+------+-----+-----------+
|Rohan| 23|  Male|Rohan|Engineering|
|kirti| 26|Female|kirti|  Marketing|
|Rohit| 22|  Male|Rohit|    Finance|
+-----+---+------+-----+-----------+
```

Performed Cross Join on Table 1 and table 2:


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown

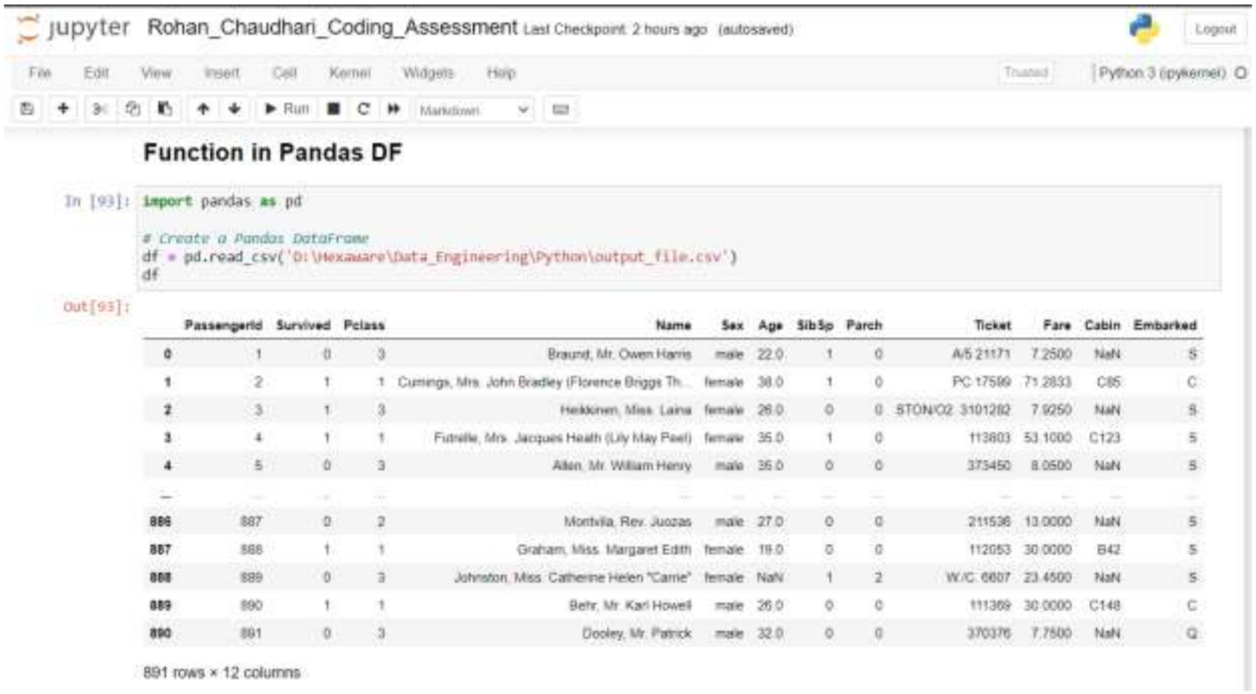
Cross join

In [92]:
```
spark.sql('SELECT * FROM table1 t1 CROSS JOIN table2 t2 ON t1.Name=t2.Name').show()
```

```
+-----+---+------+-----+-----------+
| Name|Age|Gender| Name| Department|
+-----+---+------+-----+-----------+
|Rohan| 23|  Male|Rohan|Engineering|
|Rohit| 22|  Male|Rohit|    Finance|
|kirti| 26|Female|kirti|  Marketing|
+-----+---+------+-----+-----------+
```

Pandas:

Retrieved a csv file Containing 891 rows:



jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown

Function in Pandas DF

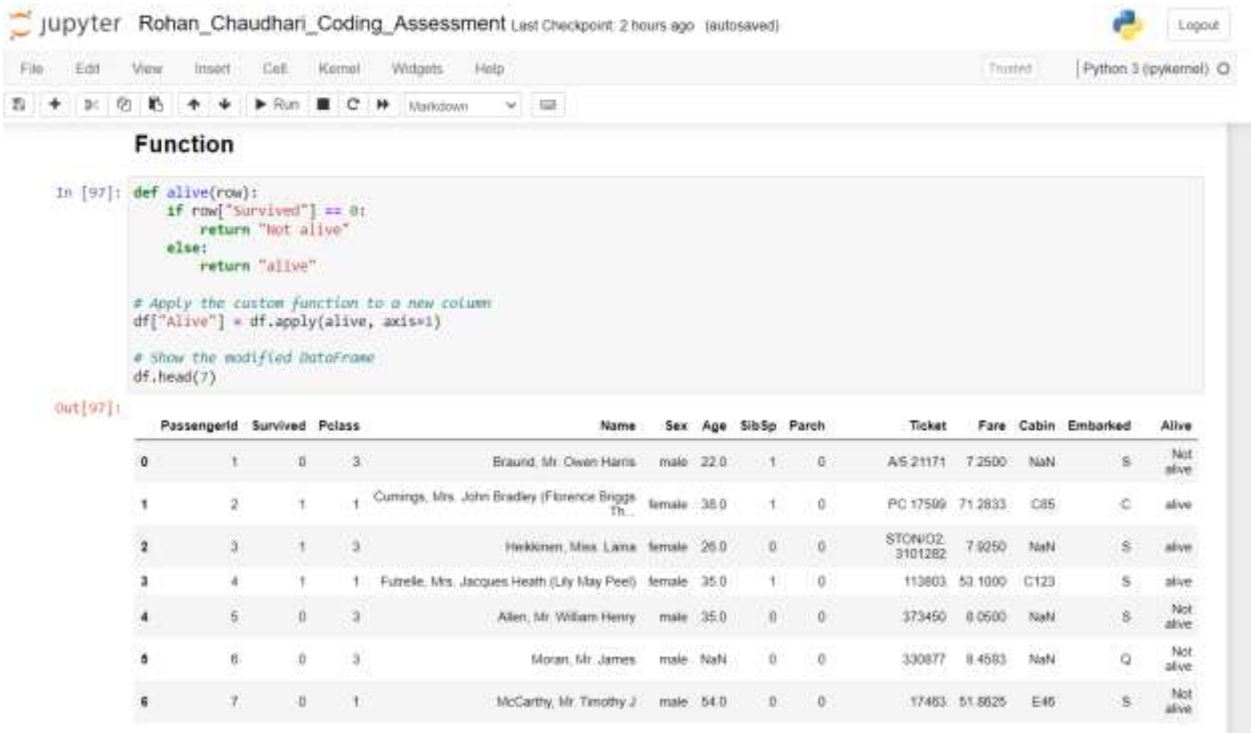
In [93]:
```python
import pandas as pd

# Create a Pandas DataFrame
df = pd.read_csv('D:\Hexaware\Data_Engineering\Python\output_file.csv')
df
```

Out[93]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

Created a function with help of Survived column where if row contains data as 1 then passenger is alive:



jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: 2 hours ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Markdown

Function

In [97]:
```python
def alive(row):
    if row["Survived"] == 0:
        return "Not alive"
    else:
        return "alive"

# Apply the custom function to a new column
df["Alive"] = df.apply(alive, axis=1)

# Show the modified DataFrame
df.head(7)
```

Out[97]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Alive
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	Not alive
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	alive
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	alive
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	alive
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	Not alive
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q	Not alive
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	Not alive