Coding challenge

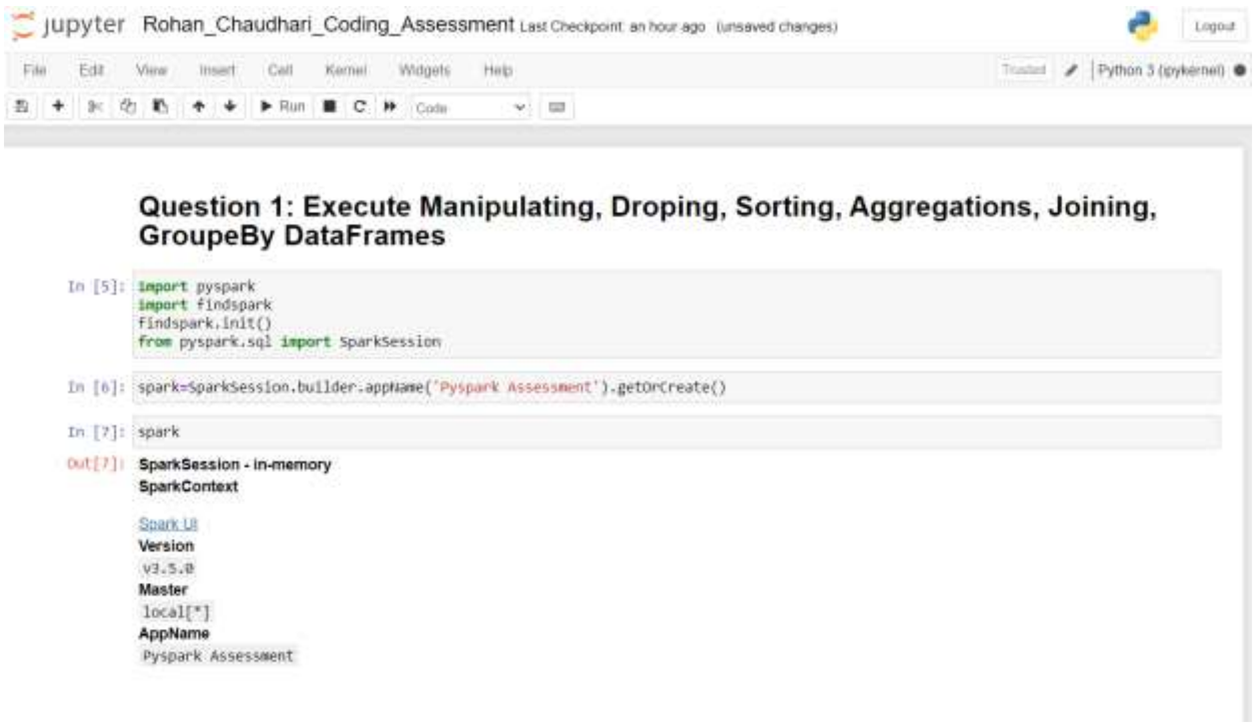
Python

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

Batch: Data Engineering 1

Question1: Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupeBy DataFrames

Created Spark Session


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Question 1: Execute Manipulating, Droping, Sorting, Aggregations, Joining, GroupeBy DataFrames

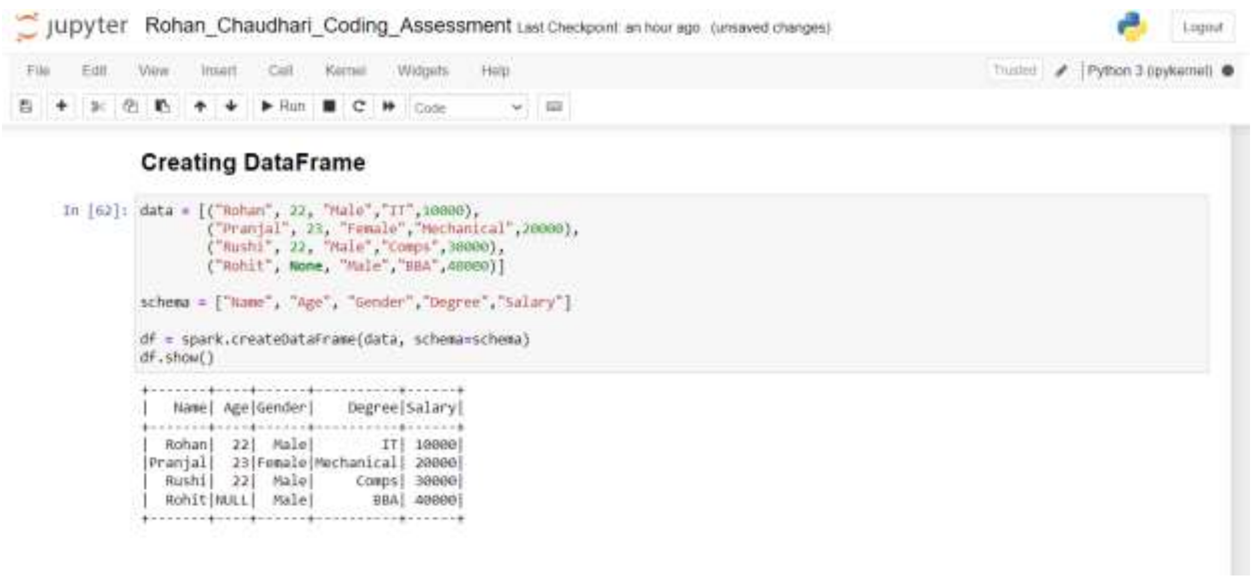
```
In [5]: import pyspark
        import findspark
        findspark.init()
        from pyspark.sql import SparkSession
```

```
In [6]: spark=SparkSession.builder.appName('Pyspark Assessment').getOrCreate()
```

```
In [7]: spark
Out[7]: SparkSession - in-memory
        SparkContext

        Spark UI
        Version
         v3.5.0
        Master
         local[*]
        AppName
         Pyspark Assessment
```

Created DataFrame


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Creating DataFrame

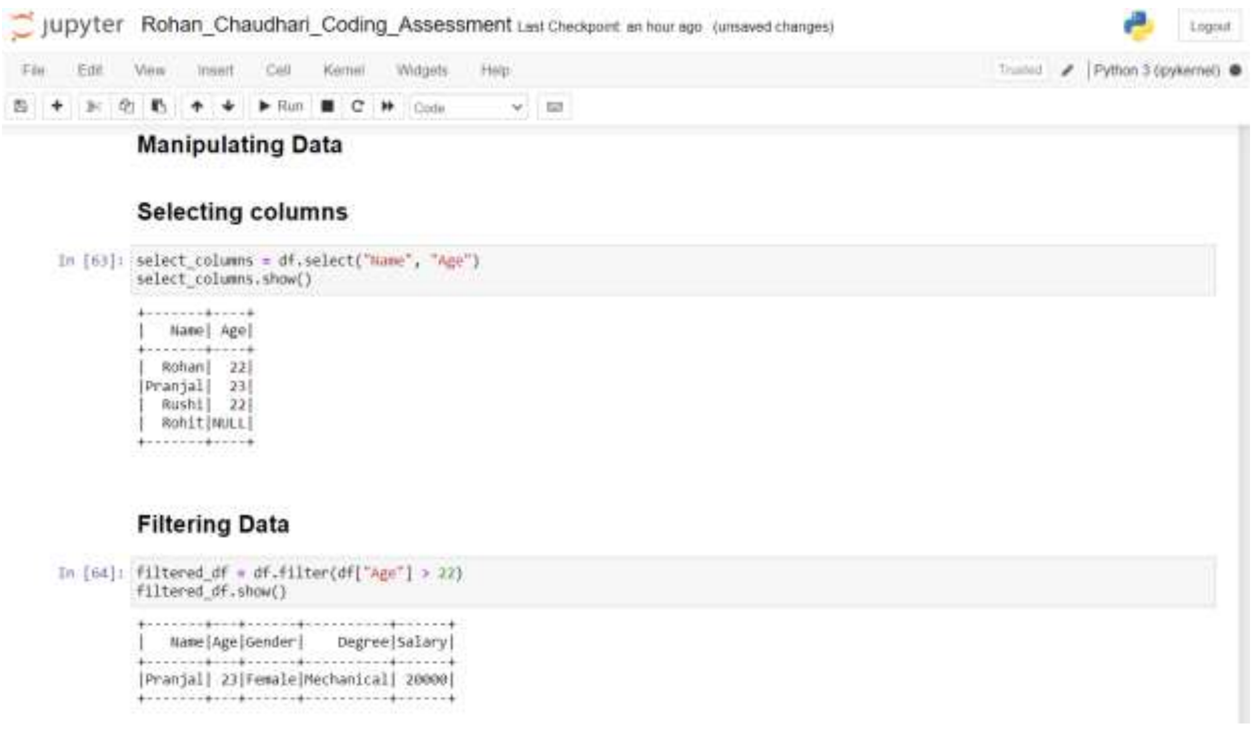
```
In [62]: data = [("Rohan", 22, "Male","IT",10000),
                ("Pranjal", 23, "Female","Mechanical",20000),
                ("Rushi", 22, "Male","Comps",30000),
                ("Rohit", None, "Male","BBA",40000)]

         schema = ["Name", "Age", "Gender","Degree","Salary"]

         df = spark.createDataFrame(data, schema=schema)
         df.show()
```

```
+-------+----+------+----------+------+
|   Name| Age|Gender|    Degree|Salary|
+-------+----+------+----------+------+
|  Rohan|  22|  Male|        IT| 10000|
|Pranjal|  23|Female|Mechanical| 20000|
|  Rushi|  22|  Male|     Comps| 30000|
|  Rohit|NULL|  Male|       BBA| 40000|
+-------+----+------+----------+------+
```

Performed Operation as seletecting 2 columns and Filtering data based on age:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Manipulating Data

Selecting columns

```
In [63]: select_columns = df.select("Name", "Age")
         select_columns.show()
```

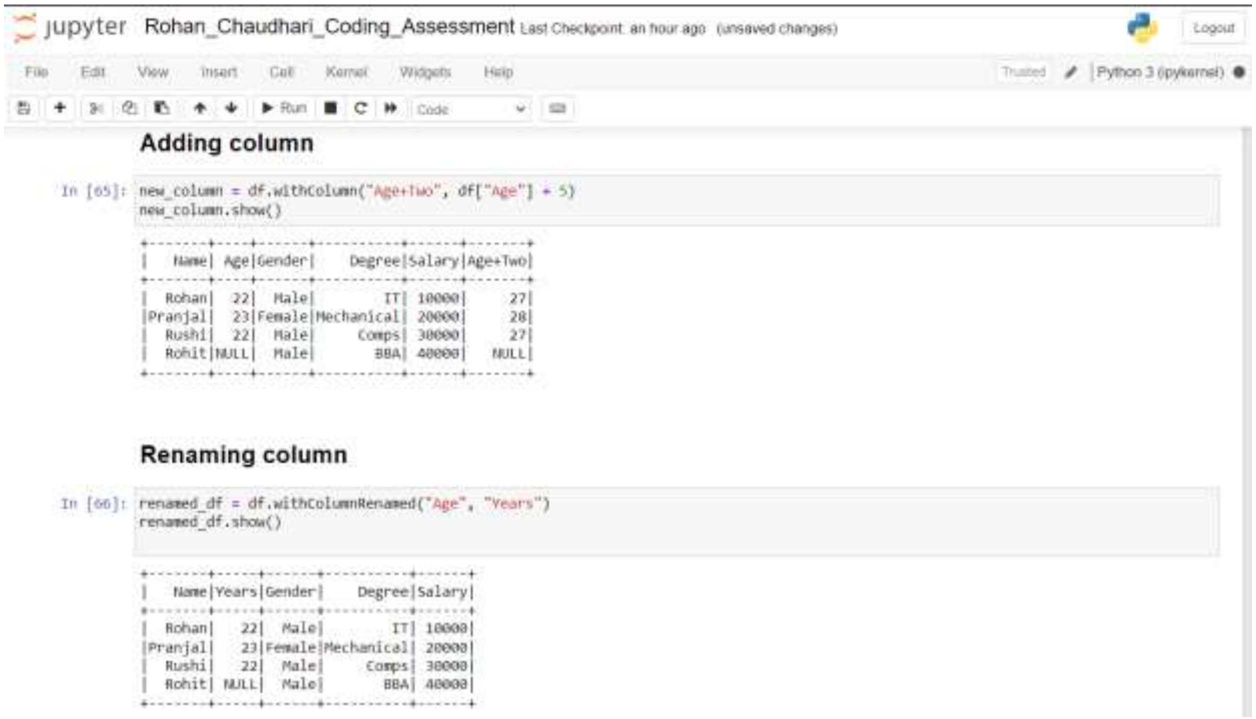
```
+-------+----+
|   Name| Age|
+-------+----+
|  Rohan|  22|
|Pranjal|  23|
|  Rushi|  22|
|  Rohit|NULL|
+-------+----+
```

Filtering Data

```
In [64]: filtered_df = df.filter(df["Age"] > 22)
         filtered_df.show()
```

```
+-------+---+------+----------+------+
|   Name|Age|Gender|    Degree|Salary|
+-------+---+------+----------+------+
|Pranjal| 23|Female|Mechanical| 20000|
+-------+---+------+----------+------+
```

Added a new column with 5 years ahed & Renamed the column:


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Adding column

In [65]:
```
new_column = df.withColumn("Age+Two", df["Age"] + 5)
new_column.show()
```

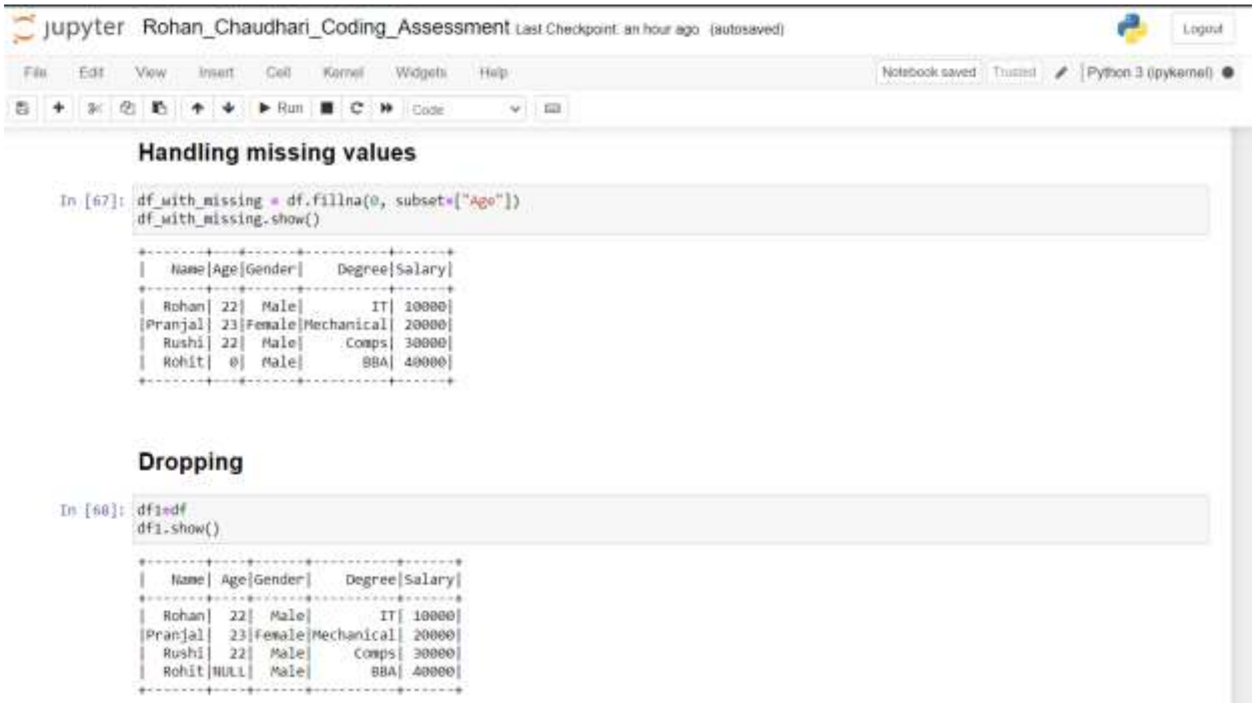
```
+-------+----+------+----------+------+-------+
|   Name| Age|Gender|    Degree|Salary|Age+Two|
+-------+----+------+----------+------+-------+
|  Rohan|  22|  Male|        IT| 10000|     27|
|Pranjal|  23|Female|Mechanical| 20000|     28|
|  Rushi|  22|  Male|     Comps| 30000|     27|
|  Rohit|NULL|  Male|       BBA| 40000|   NULL|
+-------+----+------+----------+------+-------+
```

Renaming column

In [66]:
```
renamed_df = df.withColumnRenamed("Age", "Years")
renamed_df.show()
```

```
+-------+-----+------+----------+------+
|   Name|Years|Gender|    Degree|Salary|
+-------+-----+------+----------+------+
|  Rohan|   22|  Male|        IT| 10000|
|Pranjal|   23|Female|Mechanical| 20000|
|  Rushi|   22|  Male|     Comps| 30000|
|  Rohit| NULL|  Male|       BBA| 40000|
+-------+-----+------+----------+------+
```

Handled the missing value of age and filled it with 0:


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Notebook saved
Trusted
Python 3 (ipykernel)
Run
Code

Handling missing values

In [67]:
```
df_with_missing = df.fillna(0, subset=["Age"])
df_with_missing.show()
```

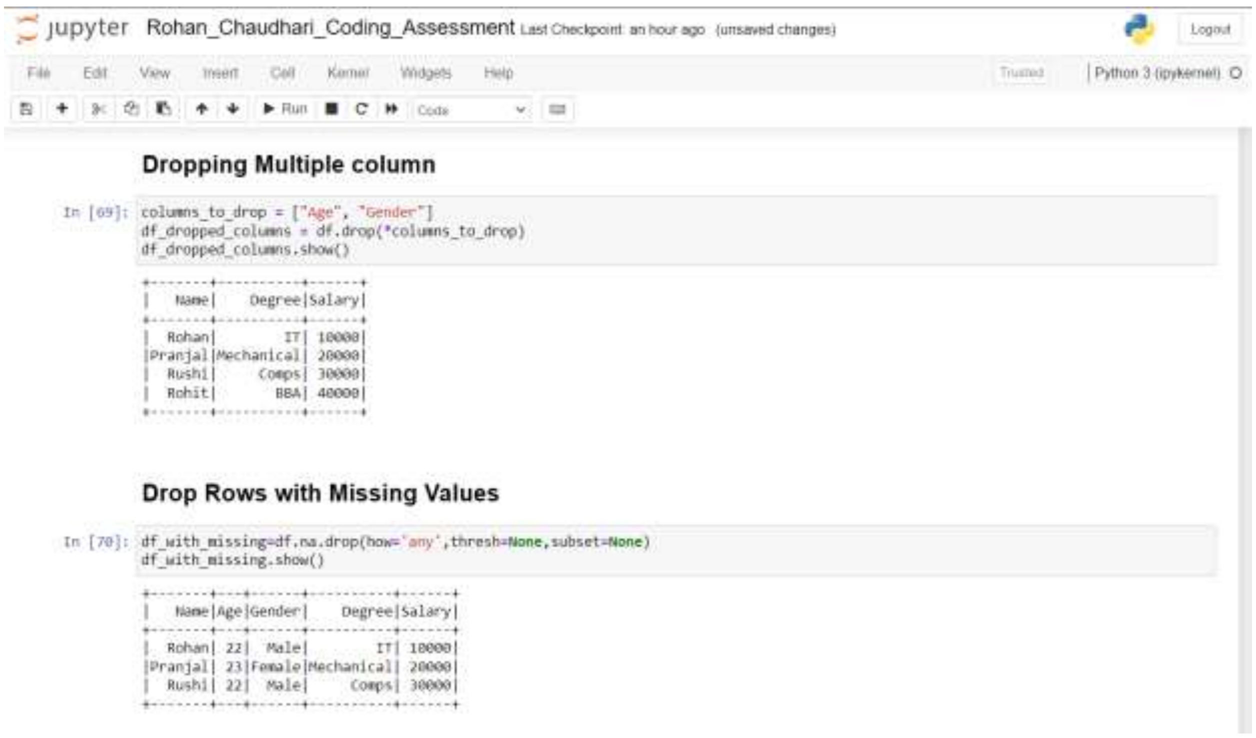
```
+-------+---+------+----------+------+
|   Name|Age|Gender|    Degree|Salary|
+-------+---+------+----------+------+
|  Rohan| 22|  Male|        IT| 10000|
|Pranjal| 23|Female|Mechanical| 20000|
|  Rushi| 22|  Male|     Comps| 30000|
|  Rohit|  0|  Male|       BBA| 40000|
+-------+---+------+----------+------+
```

Dropping

In [68]:
```
df1=df
df1.show()
```

```
+-------+----+------+----------+------+
|   Name| Age|Gender|    Degree|Salary|
+-------+----+------+----------+------+
|  Rohan|  22|  Male|        IT| 10000|
|Pranjal|  23|Female|Mechanical| 20000|
|  Rushi|  22|  Male|     Comps| 30000|
|  Rohit|NULL|  Male|       BBA| 40000|
+-------+----+------+----------+------+
```

Dropped More than 1 column & rows with missing value:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Dropping Multiple column

In [69]:
```
columns_to_drop = ["Age", "Gender"]
df_dropped_columns = df.drop(*columns_to_drop)
df_dropped_columns.show()
```

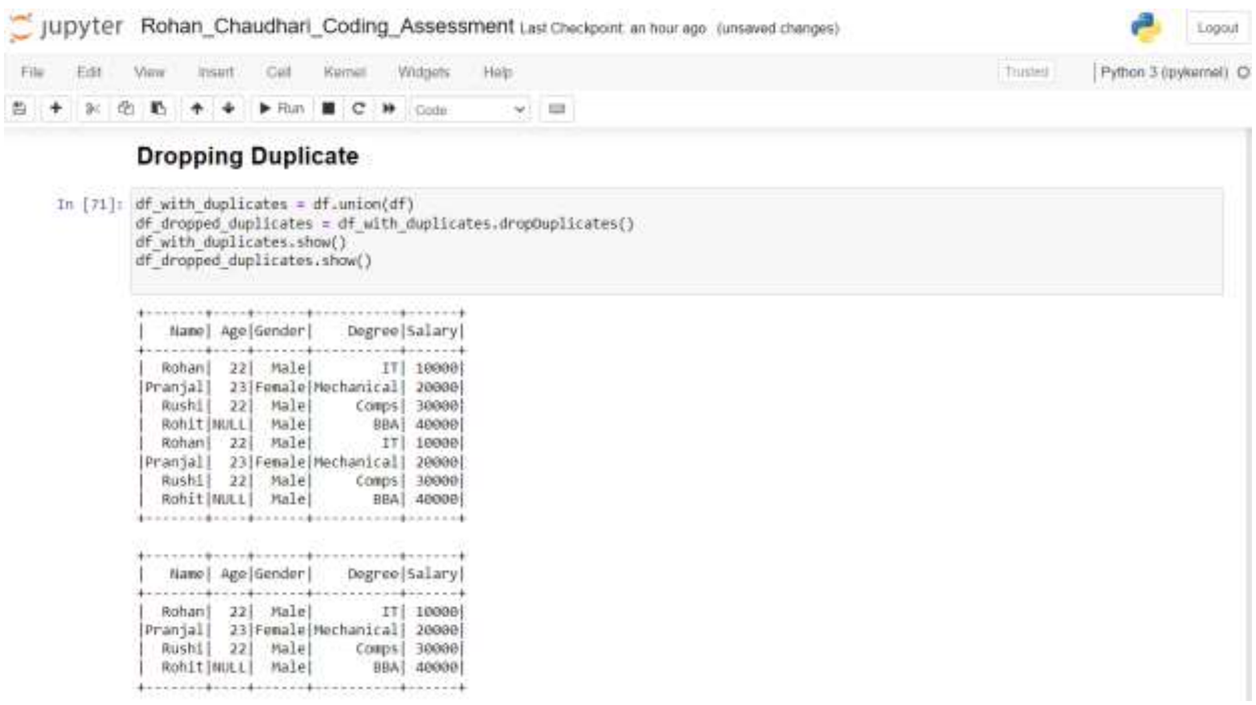
```
+------+----------+------+
|  Name|    Degree|Salary|
+------+----------+------+
| Rohan|        IT| 10000|
|Pranjal|Mechanical| 20000|
| Rushi|     Comps| 30000|
| Rohit|       BBA| 40000|
+------+----------+------+
```

Drop Rows with Missing Values

In [70]:
```
df_with_missing=df.na.drop(how='any',thresh=None,subset=None)
df_with_missing.show()
```

```
+-------+---+------+----------+------+
|   Name|Age|Gender|    Degree|Salary|
+-------+---+------+----------+------+
|  Rohan| 22|  Male|        IT| 10000|
|Pranjal| 23|Female|Mechanical| 20000|
|  Rushi| 22|  Male|     Comps| 30000|
+-------+---+------+----------+------+
```

Drop duplicate columns so for that first created duplicate rows with help of union:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Dropping Duplicate

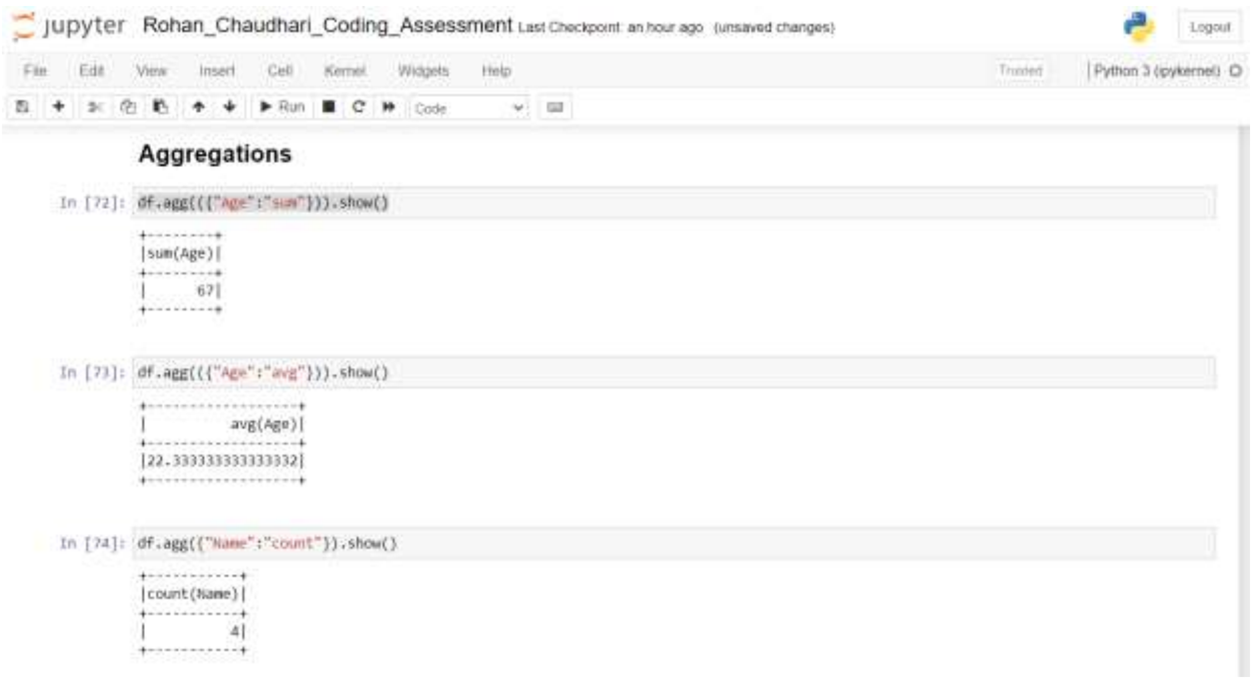
In [71]:
```
df_with_duplicates = df.union(df)
df_dropped_duplicates = df_with_duplicates.dropDuplicates()
df_with_duplicates.show()
df_dropped_duplicates.show()
```

```
+-------+----+------+----------+------+
|   Name| Age|Gender|    Degree|Salary|
+-------+----+------+----------+------+
|  Rohan|  22|  Male|        IT| 10000|
|Pranjal|  23|Female|Mechanical| 20000|
|  Rushi|  22|  Male|     Comps| 30000|
|  Rohit|NULL|  Male|       BBA| 40000|
|  Rohan|  22|  Male|        IT| 10000|
|Pranjal|  23|Female|Mechanical| 20000|
|  Rushi|  22|  Male|     Comps| 30000|
|  Rohit|NULL|  Male|       BBA| 40000|
+-------+----+------+----------+------+

+-------+----+------+----------+------+
|   Name| Age|Gender|    Degree|Salary|
+-------+----+------+----------+------+
|  Rohan|  22|  Male|        IT| 10000|
|Pranjal|  23|Female|Mechanical| 20000|
|  Rushi|  22|  Male|     Comps| 30000|
|  Rohit|NULL|  Male|       BBA| 40000|
+-------+----+------+----------+------+
```

Performed Aggregation Functions:

Sum,Average,Count:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Aggregations

```
In [72]: df.agg(({"Age":"sum"})).show()
```

```
+--------+
|sum(Age)|
+--------+
|      67|
+--------+
```

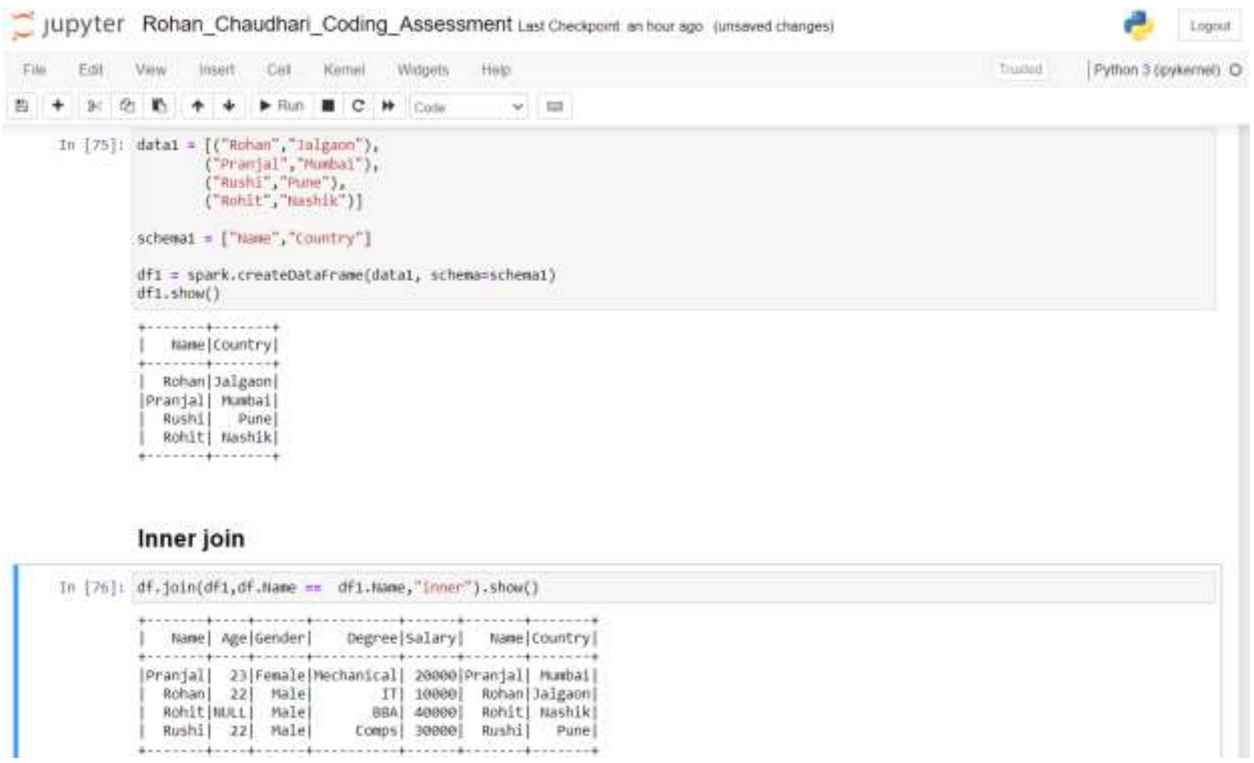
```
In [73]: df.agg(({"Age":"avg"})).show()
```

```
+------------------+
|          avg(Age)|
+------------------+
|22.333333333333332|
+------------------+
```

```
In [74]: df.agg({"Name":"count"}).show()
```

```
+-----------+
|count(Name)|
+-----------+
|          4|
+-----------+
```

Created another dataframe for performing joins operation here same column kept is of Name:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

```
In [75]: data1 = [("Rohan","Jalgaon"),
                 ("Pranjal","Mumbai"),
                 ("Rushi","Pune"),
                 ("Rohit","Nashik")]

         schema1 = ["Name","Country"]

         df1 = spark.createDataFrame(data1, schema=schema1)
         df1.show()
```

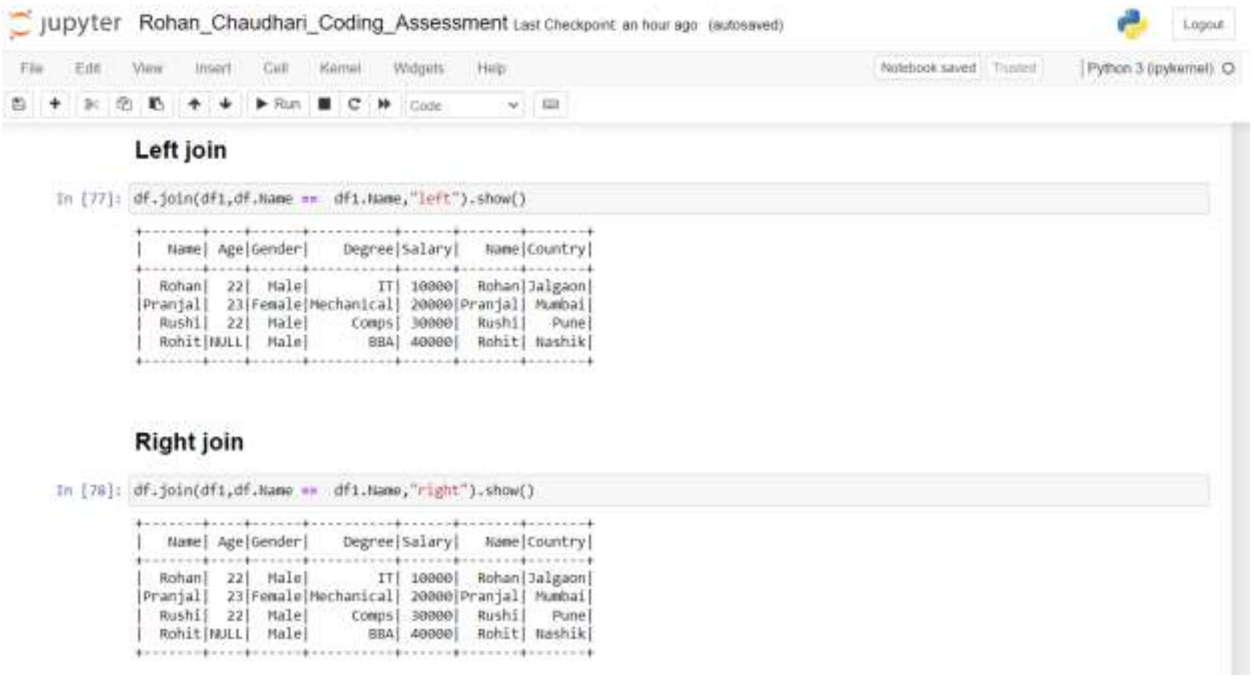
```
+-------+-------+
|   Name|Country|
+-------+-------+
|  Rohan|Jalgaon|
|Pranjal| Mumbai|
|  Rushi|   Pune|
|  Rohit| Nashik|
+-------+-------+
```

Inner join

```
In [76]: df.join(df1,df.Name ==  df1.Name,"inner").show()
```

```
+-------+----+------+-----------+------+-------+-------+
|   Name| Age|Gender|     Degree|Salary|   Name|Country|
+-------+----+------+-----------+------+-------+-------+
|Pranjal|  23|Female|Mechanical| 20000|Pranjal| Mumbai|
|  Rohan|  22|  Male|         IT| 10000|  Rohan|Jalgaon|
|  Rohit|NULL|  Male|        BBA| 40000|  Rohit| Nashik|
|  Rushi|  22|  Male|      Comps| 30000|  Rushi|   Pune|
+-------+----+------+-----------+------+-------+-------+
```

Performed Left & Right Join operation with common column as Name:



Left join

```
In [77]: df.join(df1,df.Name ==  df1.Name,"left").show()
```

```
+-------+----+------+----------+------+-------+-------+
|   Name| Age|Gender|    Degree|Salary|   Name|Country|
+-------+----+------+----------+------+-------+-------+
|  Rohan|  22|  Male|        IT| 10000|  Rohan|Jalgaon|
|Pranjal|  23|Female|Mechanical| 20000|Pranjal| Mumbai|
|  Rushi|  22|  Male|     Comps| 30000|  Rushi|   Pune|
|  Rohit|NULL|  Male|       BBA| 40000|  Rohit| Nashik|
+-------+----+------+----------+------+-------+-------+
```

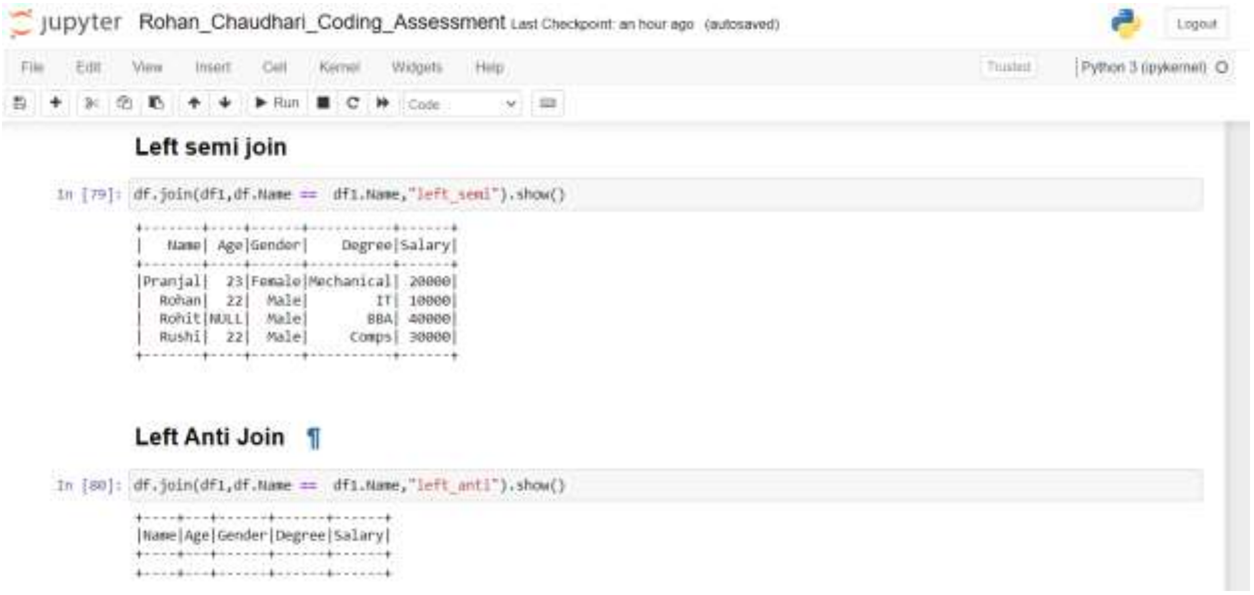
Right join

```
In [78]: df.join(df1,df.Name ==  df1.Name,"right").show()
```

```
+-------+----+------+----------+------+-------+-------+
|   Name| Age|Gender|    Degree|Salary|   Name|Country|
+-------+----+------+----------+------+-------+-------+
|  Rohan|  22|  Male|        IT| 10000|  Rohan|Jalgaon|
|Pranjal|  23|Female|Mechanical| 20000|Pranjal| Mumbai|
|  Rushi|  22|  Male|     Comps| 30000|  Rushi|   Pune|
|  Rohit|NULL|  Male|       BBA| 40000|  Rohit| Nashik|
+-------+----+------+----------+------+-------+-------+
```

Performed Left Semi & Anti Join operations:

Here Anti join output is null because every data in both table is matched:



Left semi join

```
In [79]: df.join(df1,df.Name ==  df1.Name,"left_semi").show()
```

```
+-------+----+------+----------+------+
|   Name| Age|Gender|    Degree|Salary|
+-------+----+------+----------+------+
|Pranjal|  23|Female|Mechanical| 20000|
|  Rohan|  22|  Male|        IT| 10000|
|  Rohit|NULL|  Male|       BBA| 40000|
|  Rushi|  22|  Male|     Comps| 30000|
+-------+----+------+----------+------+
```

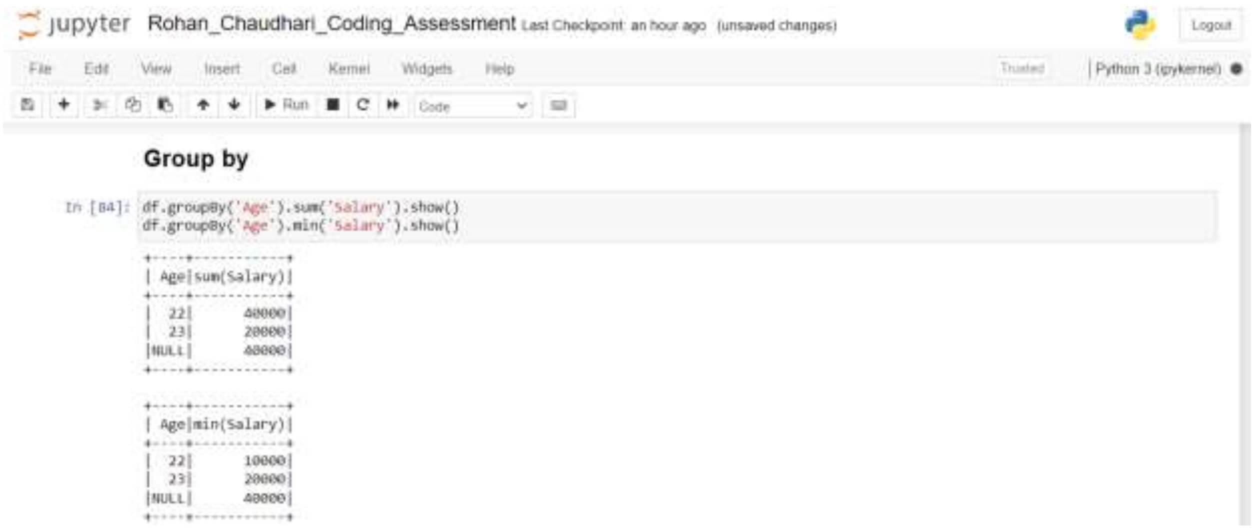
Left Anti Join ¶

```
In [80]: df.join(df1,df.Name ==  df1.Name,"left_anti").show()
```

```
+----+---+------+------+------+
|Name|Age|Gender|Degree|Salary|
+----+---+------+------+------+
+----+---+------+------+------+
```

GroupBY Operations:

Groupby with Sum & Min:


jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

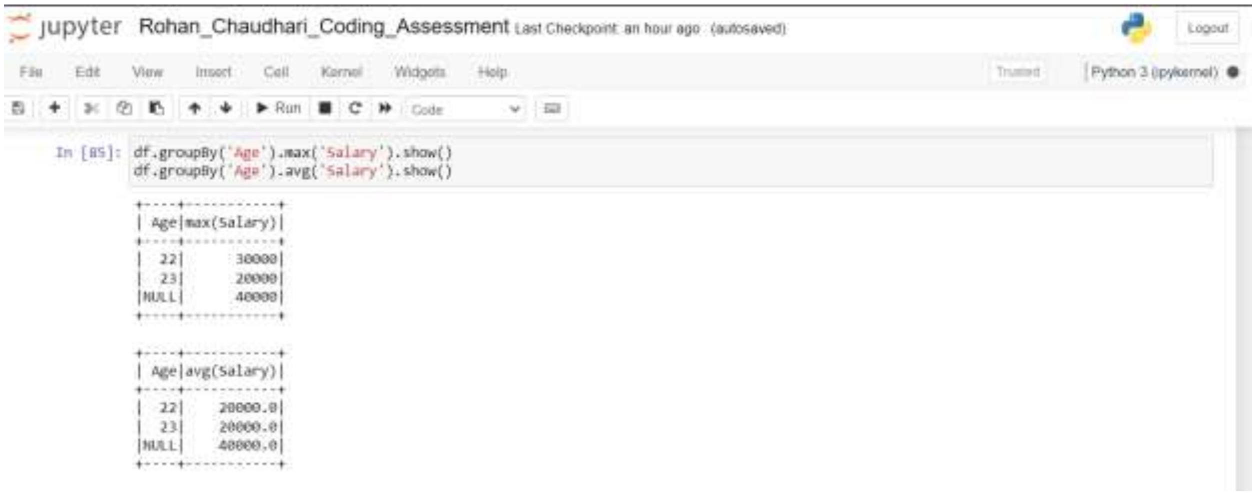
Group by

In [84]:
```
df.groupBy('Age').sum('Salary').show()
df.groupBy('Age').min('Salary').show()
```

```
+----+-----------+
| Age|sum(Salary)|
+----+-----------+
|  22|      40000|
|  23|      20000|
|NULL|      40000|
+----+-----------+

+----+-----------+
| Age|min(Salary)|
+----+-----------+
|  22|      10000|
|  23|      20000|
|NULL|      40000|
+----+-----------+
```

Groupby with Max& Avg:

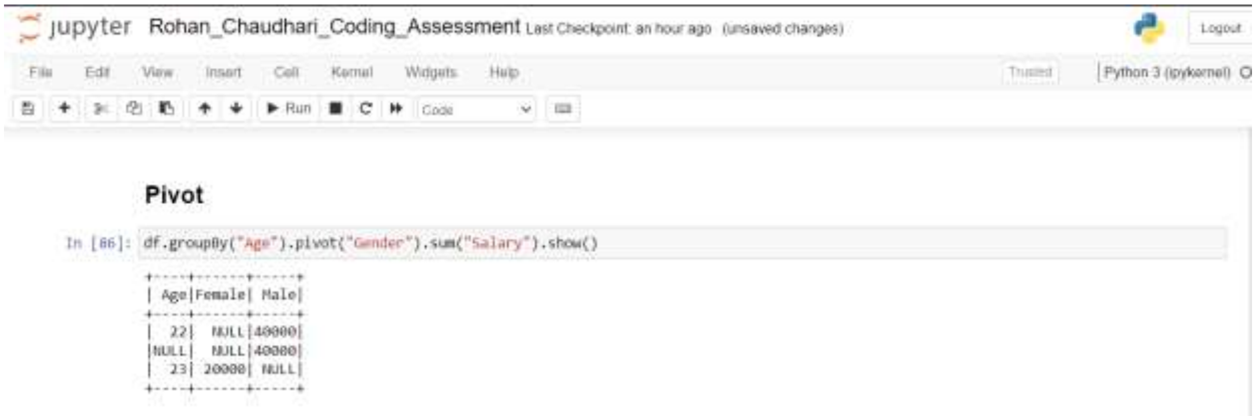

jupyter Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

In [85]:
```
df.groupBy('Age').max('Salary').show()
df.groupBy('Age').avg('Salary').show()
```

```
+----+-----------+
| Age|max(Salary)|
+----+-----------+
|  22|      30000|
|  23|      20000|
|NULL|      40000|
+----+-----------+

+----+-----------+
| Age|avg(Salary)|
+----+-----------+
|  22|    20000.0|
|  23|    20000.0|
|NULL|    40000.0|
+----+-----------+
```

Performed Pivot in groupBy with PIVOT kept as Gender:


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Pivot

In [86]:
```
df.groupBy("Age").pivot("Gender").sum("Salary").show()
```

```
+----+------+-----+
| Age|Female| Male|
+----+------+-----+
|  22|  NULL|40000|
|NULL|  NULL|40000|
|  23| 20000| NULL|
+----+------+-----+
```