Name: Rohan Vinayak Chaudhari

Batch: Data Engineering

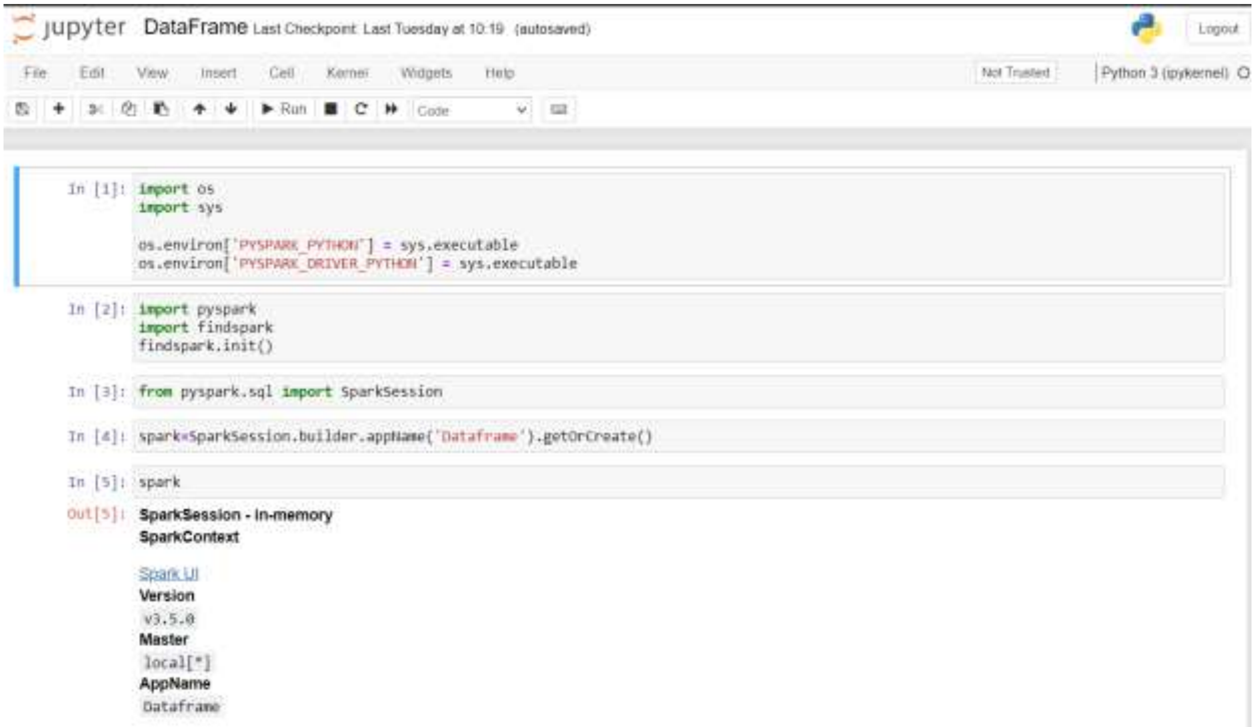
Date:09/02/2024

Topic: Pyspark

Solution:

1.Pyspark:

Creating session:



```
In [1]: import os
        import sys

        os.environ['PYSPARK_PYTHON'] = sys.executable
        os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
```

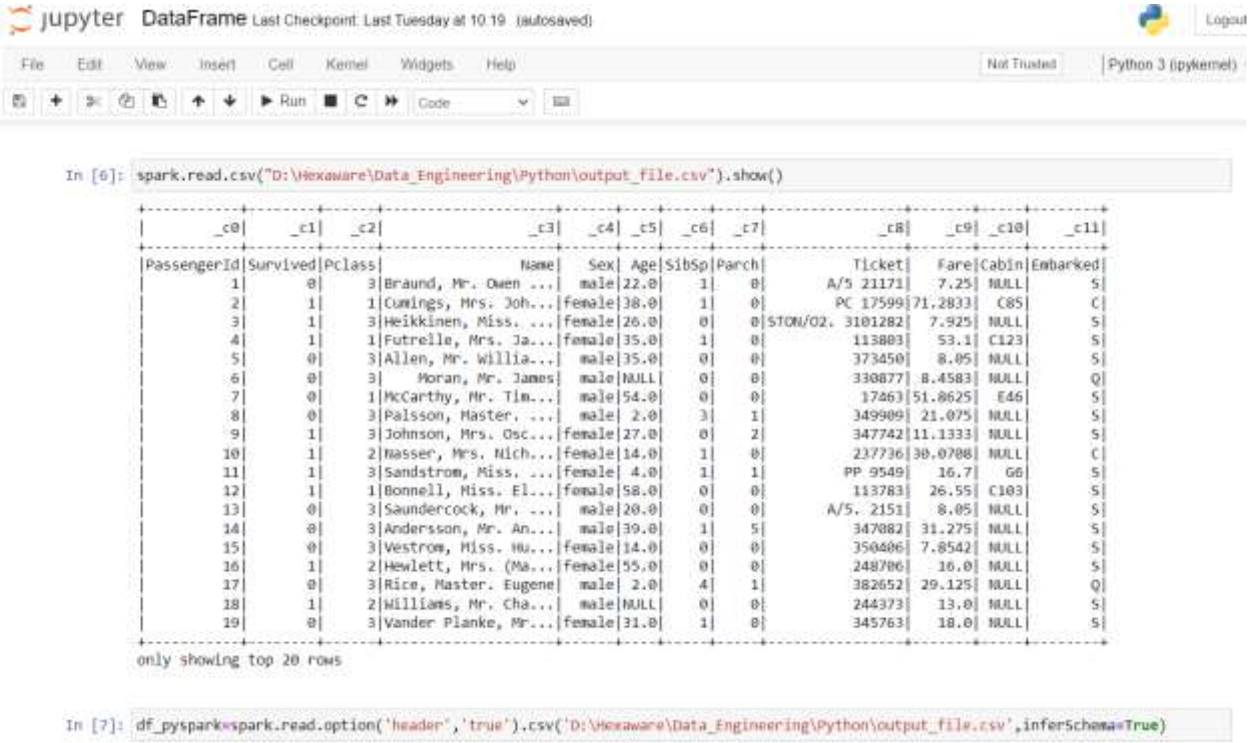
```
In [2]: import pyspark
        import findspark
        findspark.init()
```

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

```
In [4]: spark=SparkSession.builder.appName('Dataframe').getOrCreate()
```

```
In [5]: spark
```

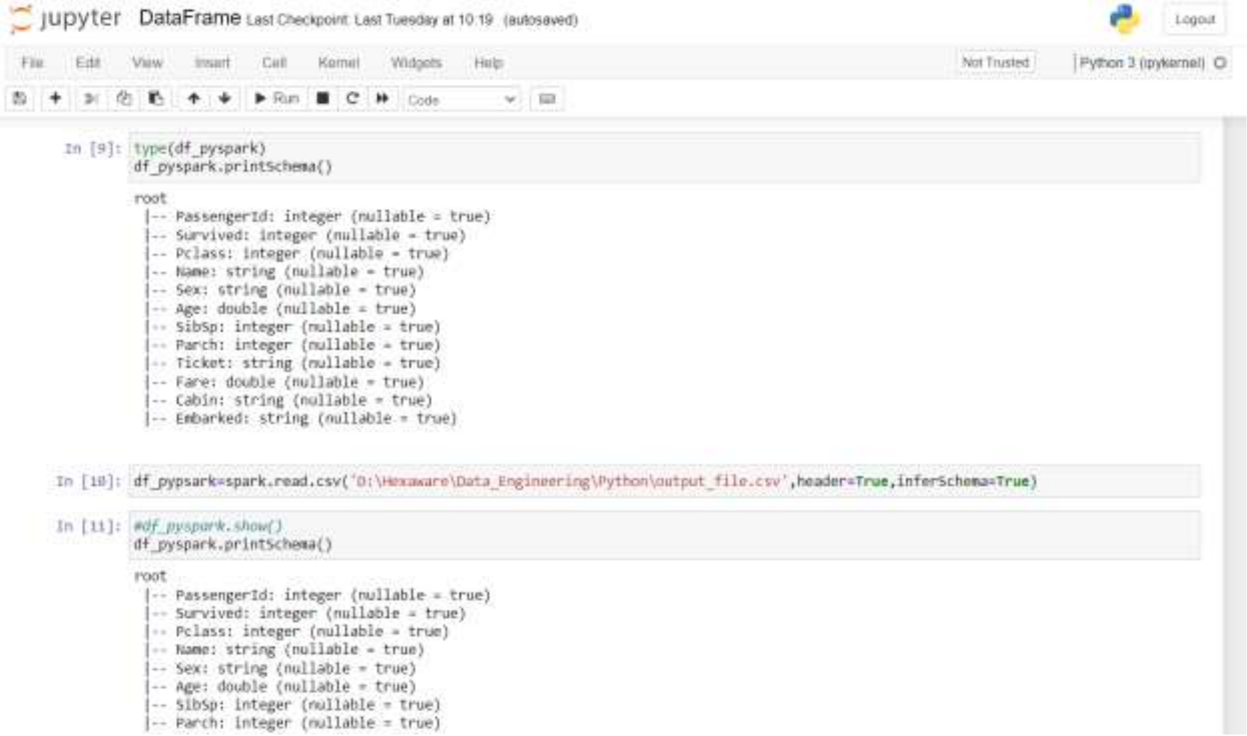
```
Out[5]: SparkSession - in-memory
        SparkContext

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


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

In [6]:
```
spark.read.csv("D:\Hexaware\Data_Engineering\Python\output_file.csv").show()
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|        _c0|     _c1|   _c2|                 _c3|   _c4| _c5|  _c6|  _c7|             _c8|    _c9| _c10|    _c11|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25| NULL|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| NULL|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| NULL|       S|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583| NULL|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| NULL|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| NULL|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| NULL|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| NULL|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| NULL|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| NULL|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| NULL|       S|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| NULL|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0| NULL|       S|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0| NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows
```

In [7]:
```
df_pyspark=spark.read.option('header','true').csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',inferSchema=True)
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

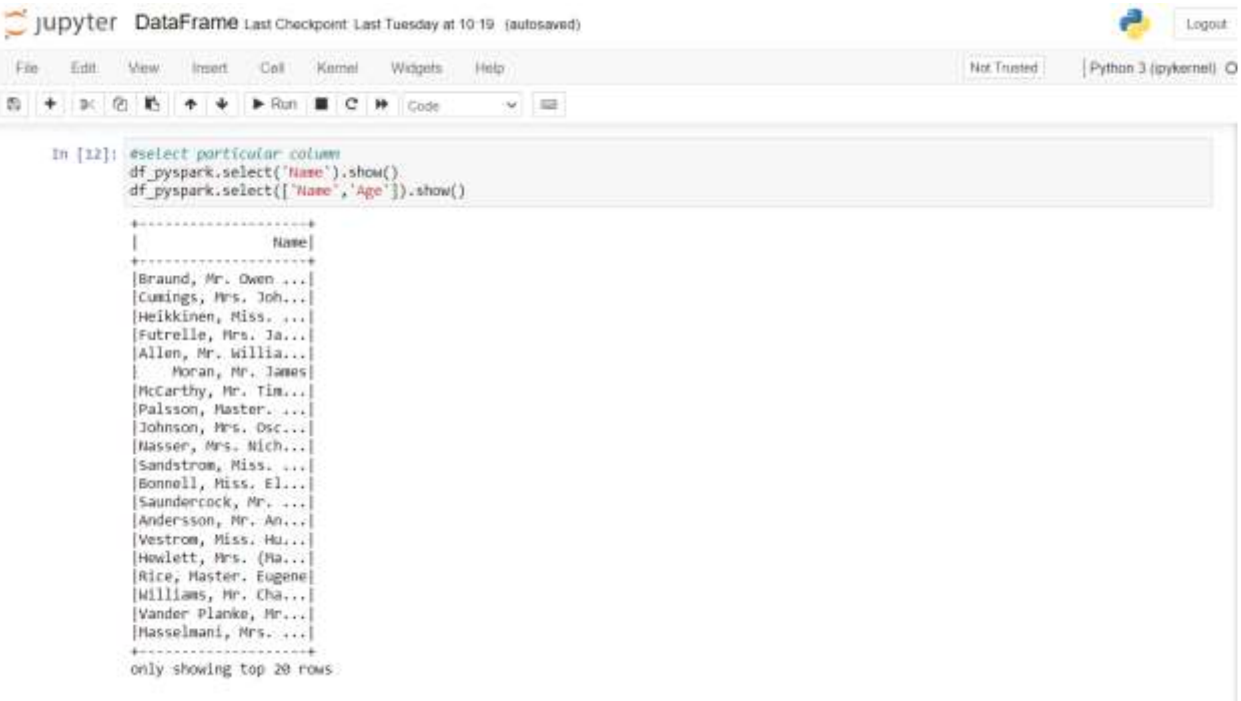
In [9]:
```
type(df_pyspark)
df_pyspark.printSchema()
```

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

In [10]:
```
df_pypsark=spark.read.csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',header=True,inferSchema=True)
```

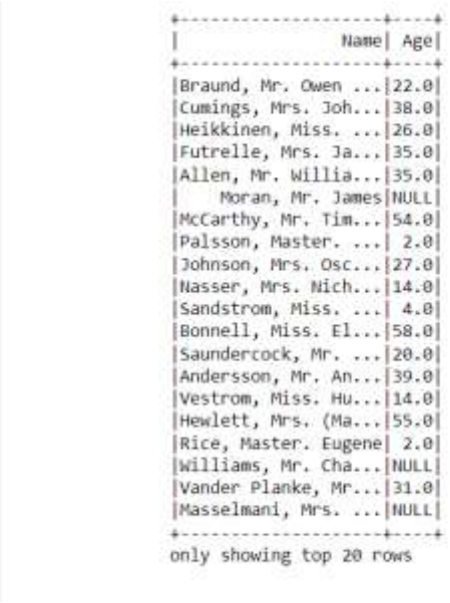
In [11]:
```
#df_pyspark.show()
df_pyspark.printSchema()
```

```
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
```

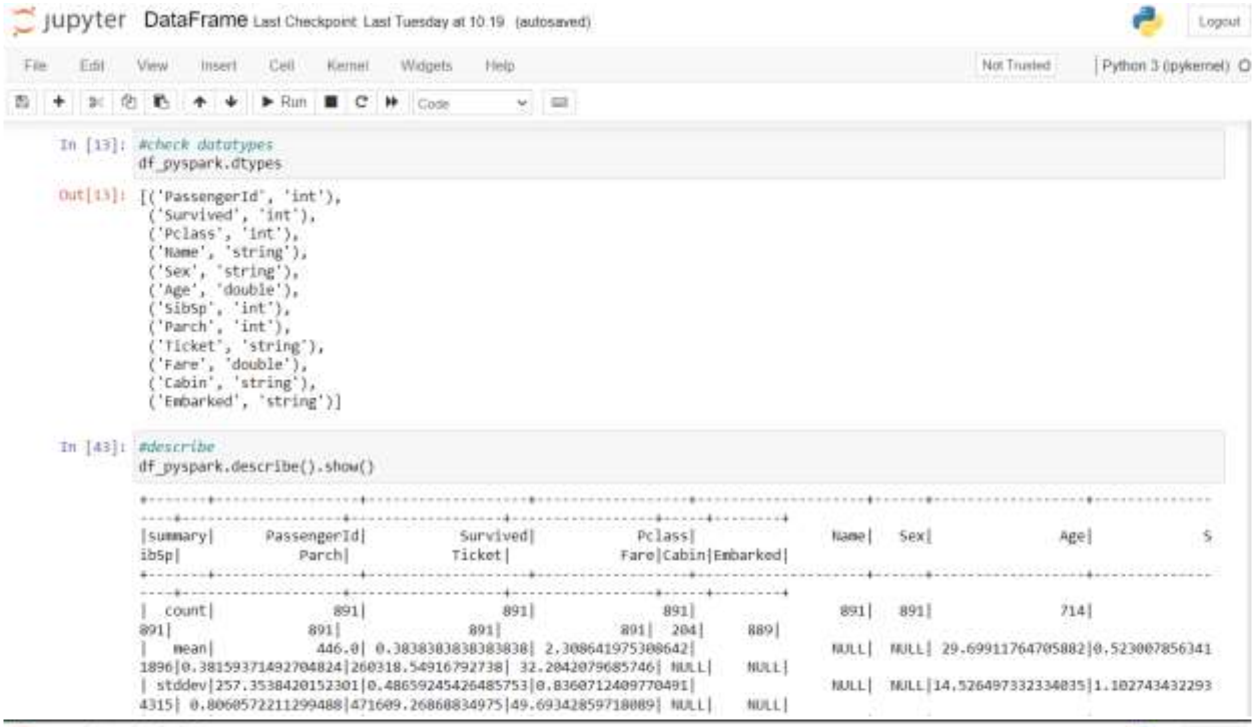


```
In [12]: #select particular column
         df_pyspark.select('Name').show()
         df_pyspark.select(['Name','Age']).show()
```

```
+--------------------+
|                Name|
+--------------------+
|Braund, Mr. Owen ...|
|Cumings, Mrs. Joh...|
|Heikkinen, Miss. ...|
|Futrelle, Mrs. Ja...|
|Allen, Mr. Willia...|
|    Moran, Mr. James|
|McCarthy, Mr. Tim...|
|Palsson, Master. ...|
|Johnson, Mrs. Osc...|
|Nasser, Mrs. Nich...|
|Sandstrom, Miss. ...|
|Bonnell, Miss. El...|
|Saundercock, Mr. ...|
|Andersson, Mr. An...|
|Vestrom, Miss. Hu...|
|Hewlett, Mrs. (Ma...|
|Rice, Master. Eugene|
|Williams, Mr. Cha...|
|Vander Planke, Mr...|
|Masselmani, Mrs. ...|
+--------------------+
only showing top 20 rows
```



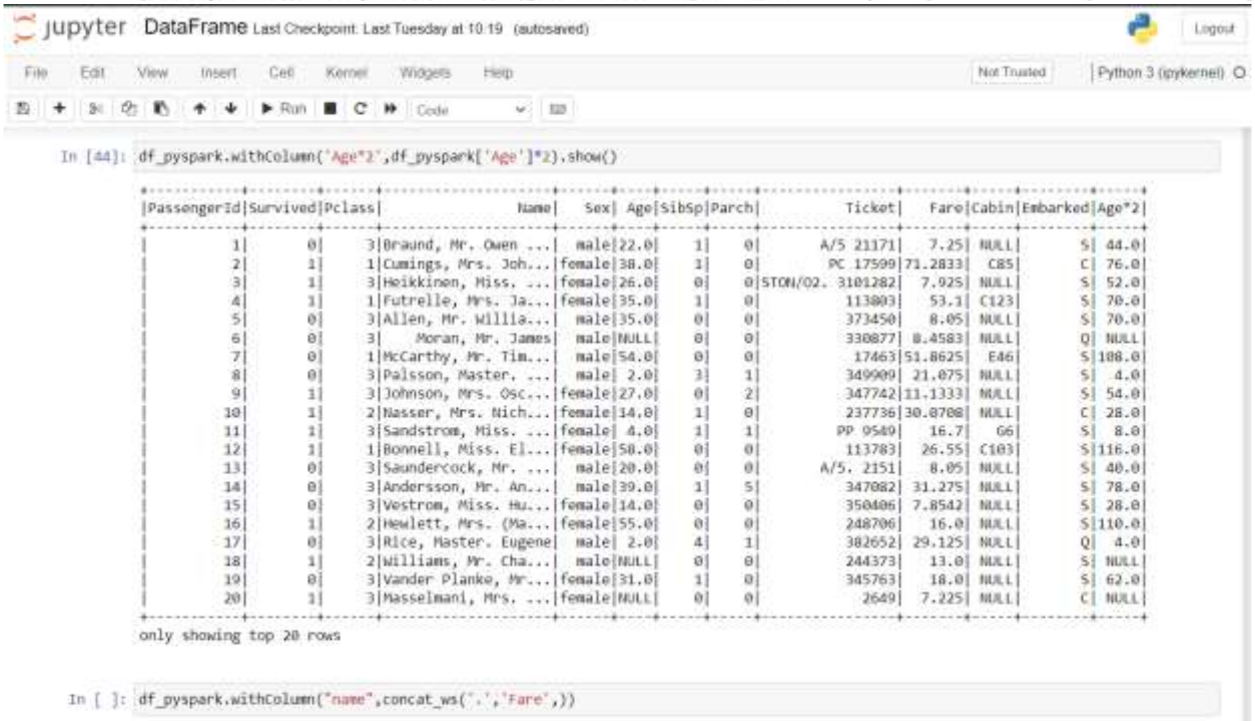
```
+--------------------+----+
|                Name| Age|
+--------------------+----+
|Braund, Mr. Owen ...|22.0|
|Cumings, Mrs. Joh...|38.0|
|Heikkinen, Miss. ...|26.0|
|Futrelle, Mrs. Ja...|35.0|
|Allen, Mr. Willia...|35.0|
|    Moran, Mr. James|NULL|
|McCarthy, Mr. Tim...|54.0|
|Palsson, Master. ...| 2.0|
|Johnson, Mrs. Osc...|27.0|
|Nasser, Mrs. Nich...|14.0|
|Sandstrom, Miss. ...| 4.0|
|Bonnell, Miss. El...|58.0|
|Saundercock, Mr. ...|20.0|
|Andersson, Mr. An...|39.0|
|Vestrom, Miss. Hu...|14.0|
|Hewlett, Mrs. (Ma...|55.0|
|Rice, Master. Eugene| 2.0|
|Williams, Mr. Cha...|NULL|
|Vander Planke, Mr...|31.0|
|Masselmani, Mrs. ...|NULL|
+--------------------+----+
only showing top 20 rows
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [13]: #check datatypes
         df_pyspark.dtypes

Out[13]: [('PassengerId', 'int'),
          ('Survived', 'int'),
          ('Pclass', 'int'),
          ('Name', 'string'),
          ('Sex', 'string'),
          ('Age', 'double'),
          ('SibSp', 'int'),
          ('Parch', 'int'),
          ('Ticket', 'string'),
          ('Fare', 'double'),
          ('Cabin', 'string'),
          ('Embarked', 'string')]
```

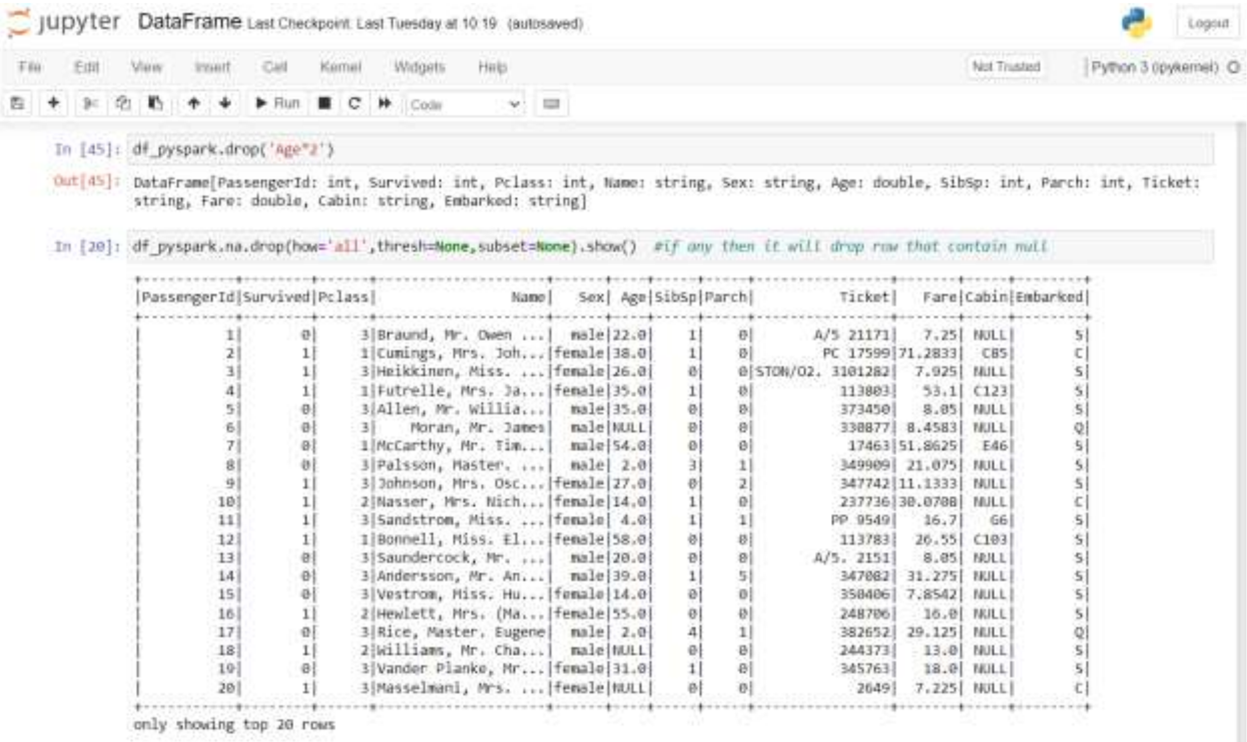
```
In [43]: #describe
         df_pyspark.describe().show()
```

```
+-------+------------------+-------------------+------------------+-----+-----+------------------+------------------
----+------------------+------------------+------------------+-----+--------+
|summary|       PassengerId|           Survived|            Pclass|           Name|  Sex|               Age|                 S
ibSp|             Parch|            Ticket|              Fare|Cabin|Embarked|
+-------+------------------+-------------------+------------------+-----+-----+------------------+------------------
----+------------------+------------------+------------------+-----+--------+
|  count|               891|                891|               891|            891|  891|               714|
891|               891|               891|               891|  204|     889|
|   mean|             446.0| 0.3838383838383838| 2.308641975308642|           NULL| NULL| 29.69911764705882|0.523007856341
1896|0.38159371492704824|260318.54916792738| 32.2042079685746| NULL|    NULL|
| stddev|257.3538420152301|0.48659245426485753|0.8360712409770491|           NULL| NULL|14.526497332334035|1.102743432293
4315| 0.8060572211299488|471609.26868834975|49.69342859718089| NULL|    NULL|
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [44]: df_pyspark.withColumn('Age*2',df_pyspark['Age']*2).show()

+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|Age*2|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----+
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25| NULL|       S| 44.0|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C| 76.0|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| NULL|       S| 52.0|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S| 70.0|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| NULL|       S| 70.0|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583| NULL|       Q| NULL|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|108.0|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| NULL|       S|  4.0|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| NULL|       S| 54.0|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| NULL|       C| 28.0|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|  8.0|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|116.0|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| NULL|       S| 40.0|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| NULL|       S| 78.0|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| NULL|       S| 28.0|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| NULL|       S|110.0|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| NULL|       Q|  4.0|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0| NULL|       S| NULL|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0| NULL|       S| 62.0|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225| NULL|       C| NULL|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+-----+
only showing top 20 rows
```

```
In [ ]: df_pyspark.withColumn("name",concat_ws('.','Fare',))
```

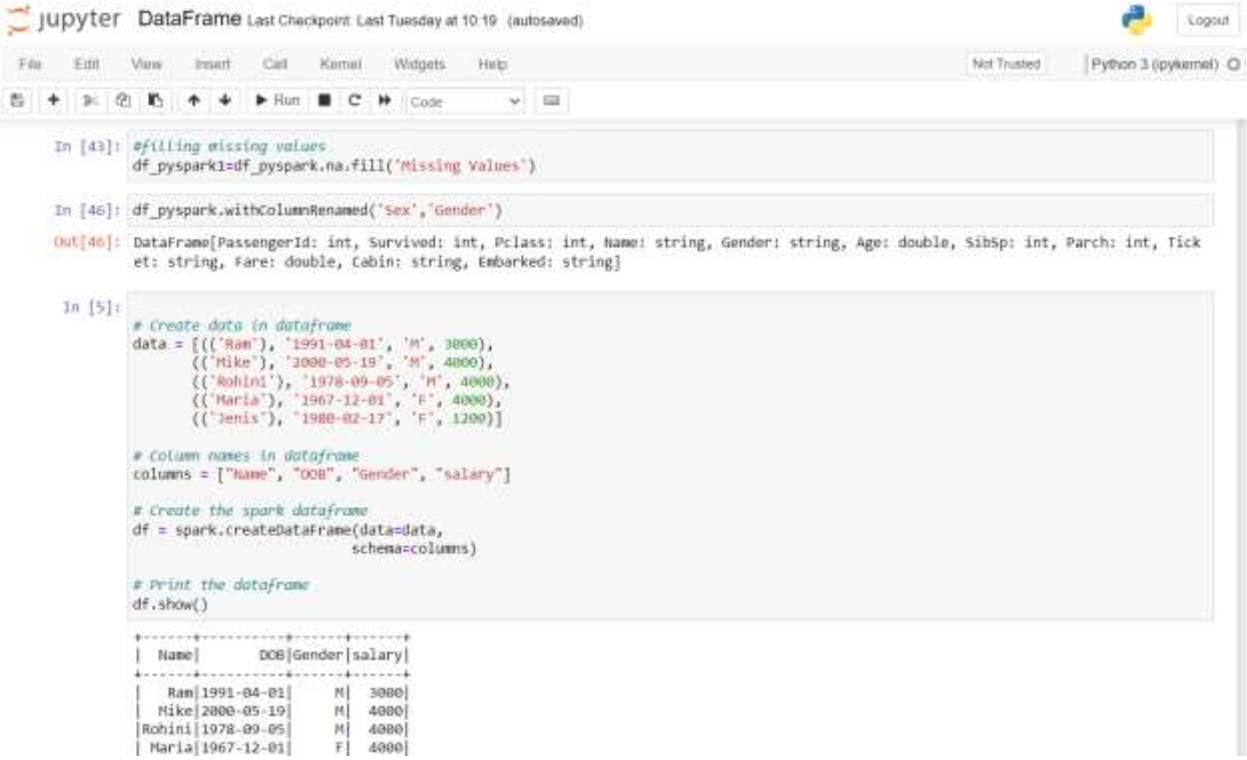


In [45]:
```
df_pyspark.drop('Age"2')
```

Out[45]:
```
DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Sex: string, Age: double, SibSp: int, Parch: int, Ticket:
string, Fare: double, Cabin: string, Embarked: string]
```

In [20]:
```
df_pyspark.na.drop(how='all',thresh=None,subset=None).show()  #if any then it will drop row that contain null
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25| NULL|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925| NULL|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05| NULL|       S|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583| NULL|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|  E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075| NULL|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333| NULL|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708| NULL|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|   G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55| C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05| NULL|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275| NULL|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542| NULL|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0| NULL|       S|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125| NULL|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0| NULL|       S|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0| NULL|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+-----+--------+
only showing top 20 rows
```



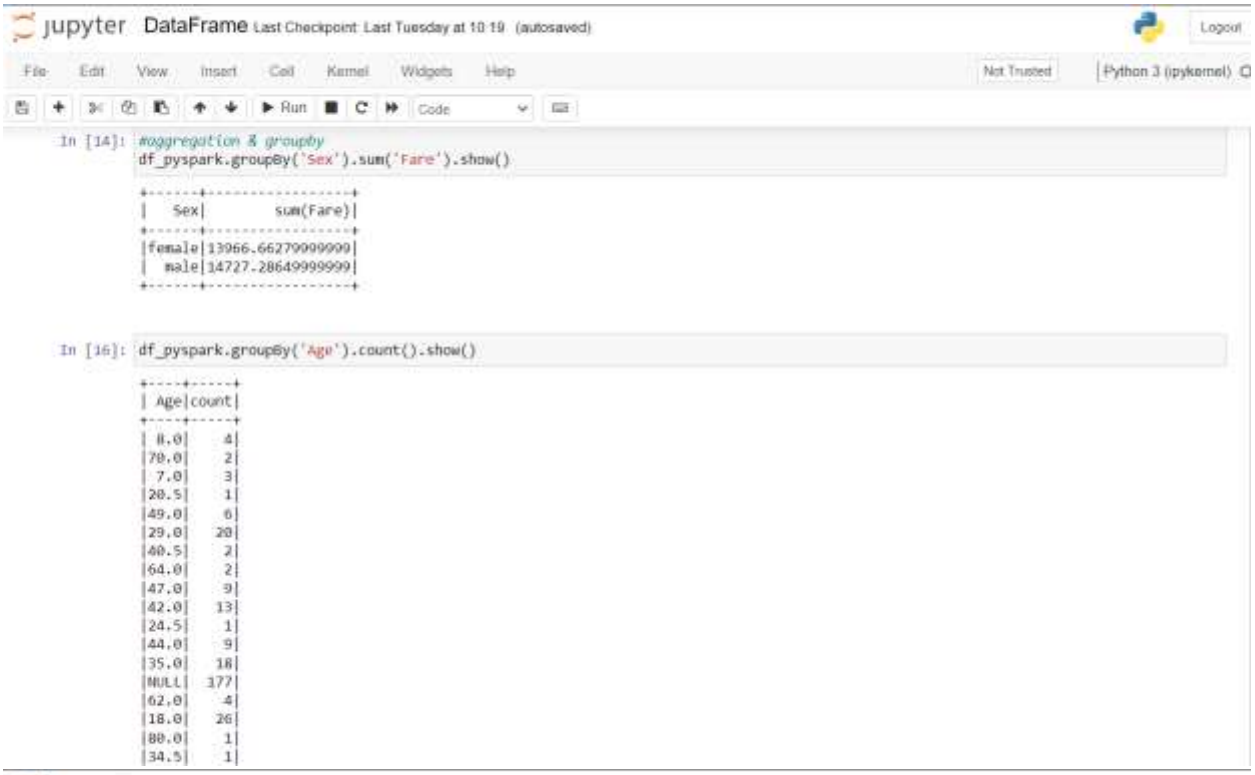
In [43]:
```
#filling missing values
df_pyspark1=df_pyspark.na.fill('Missing Values')
```

In [46]:
```
df_pyspark.withColumnRenamed('Sex','Gender')
```

Out[46]:
```
DataFrame[PassengerId: int, Survived: int, Pclass: int, Name: string, Gender: string, Age: double, SibSp: int, Parch: int, Tick
et: string, Fare: double, Cabin: string, Embarked: string]
```

In [5]:
```

# Create data in dataframe
data = [(('Ram'), '1991-04-01', 'M', 3000),
       (('Mike'), '2000-05-19', 'M', 4000),
       (('Rohini'), '1978-09-05', 'M', 4000),
       (('Maria'), '1967-12-01', 'F', 4000),
       (('Jenis'), '1980-02-17', 'F', 1200)]

# Column names in dataframe
columns = ["Name", "DOB", "Gender", "salary"]

# Create the spark dataframe
df = spark.createDataFrame(data=data,
                           schema=columns)

# Print the dataframe
df.show()
```

```
+------+----------+------+------+
|  Name|       DOB|Gender|salary|
+------+----------+------+------+
|   Ram|1991-04-01|     M|  3000|
|  Mike|2000-05-19|     M|  4000|
|Rohini|1978-09-05|     M|  4000|
| Maria|1967-12-01|     F|  4000|
```

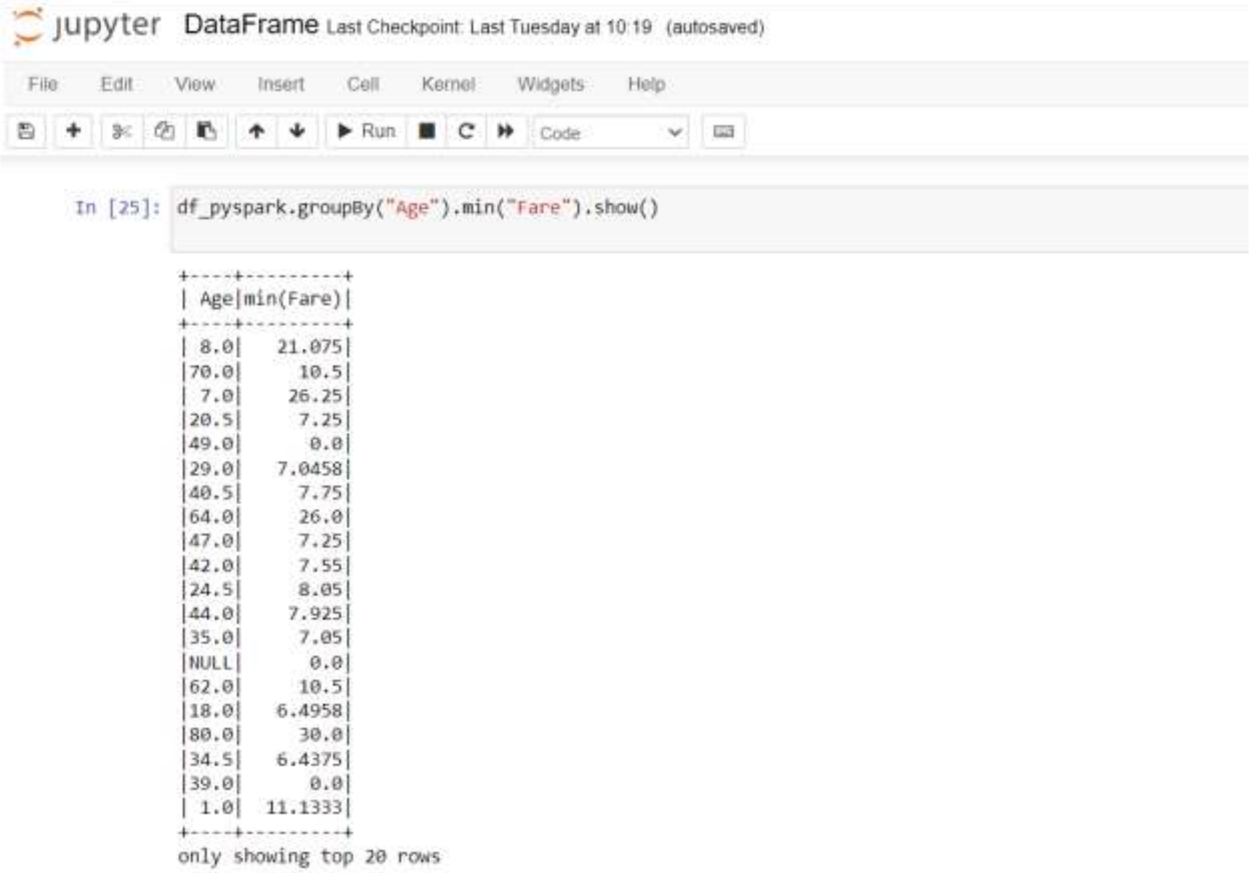

jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

In [14]:
```
#aggregation & groupby
df_pyspark.groupBy('Sex').sum('Fare').show()
```

```
+------+-----------------+
|   Sex|        sum(Fare)|
+------+-----------------+
|female|13966.66279999999|
|  male|14727.28649999999|
+------+-----------------+
```

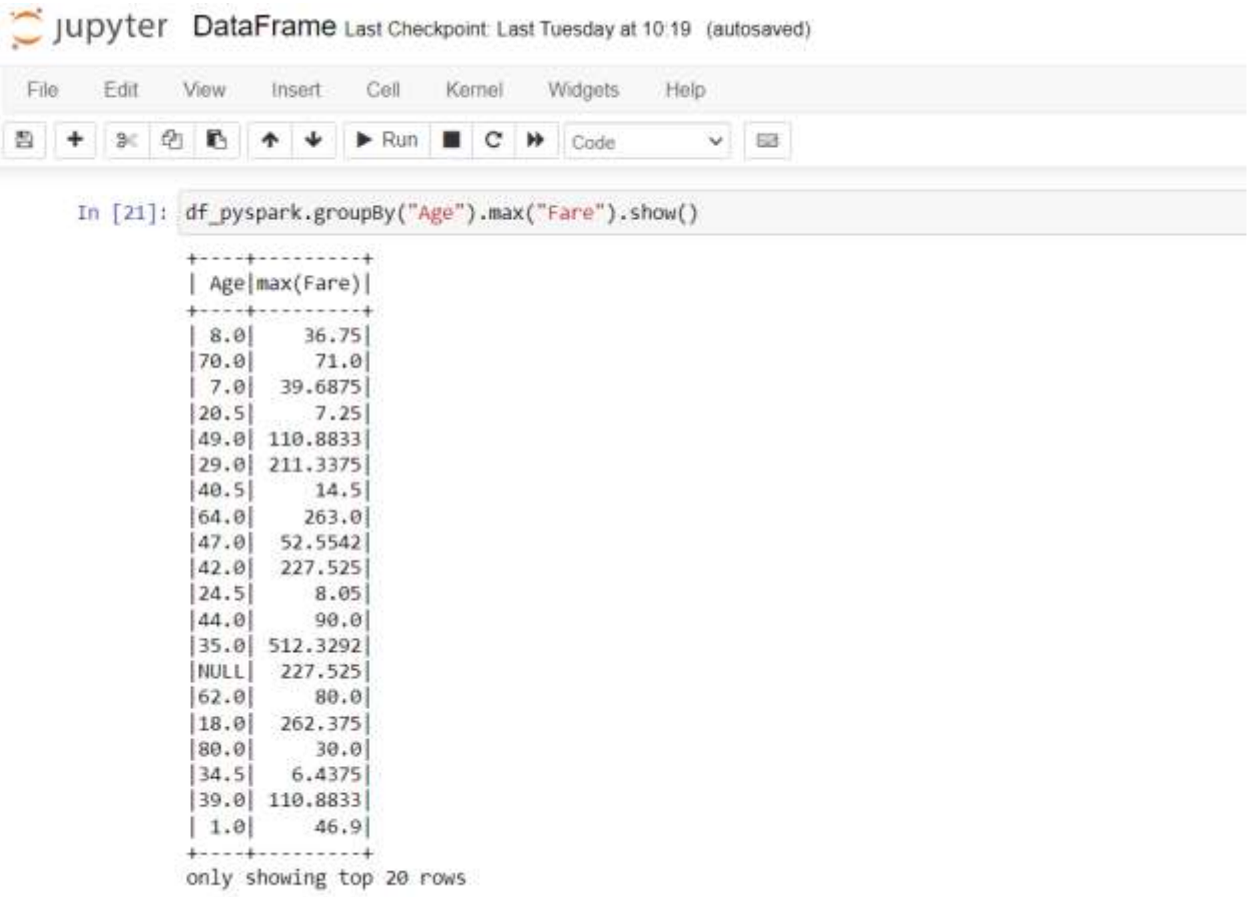
In [16]:
```
df_pyspark.groupBy('Age').count().show()
```

```
+----+-----+
| Age|count|
+----+-----+
| 8.0|    4|
|70.0|    2|
| 7.0|    3|
|20.5|    1|
|49.0|    6|
|29.0|   20|
|40.5|    2|
|64.0|    2|
|47.0|    9|
|42.0|   13|
|24.5|    1|
|44.0|    9|
|35.0|   18|
|NULL|  177|
|62.0|    4|
|18.0|   26|
|80.0|    1|
|34.5|    1|
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Code

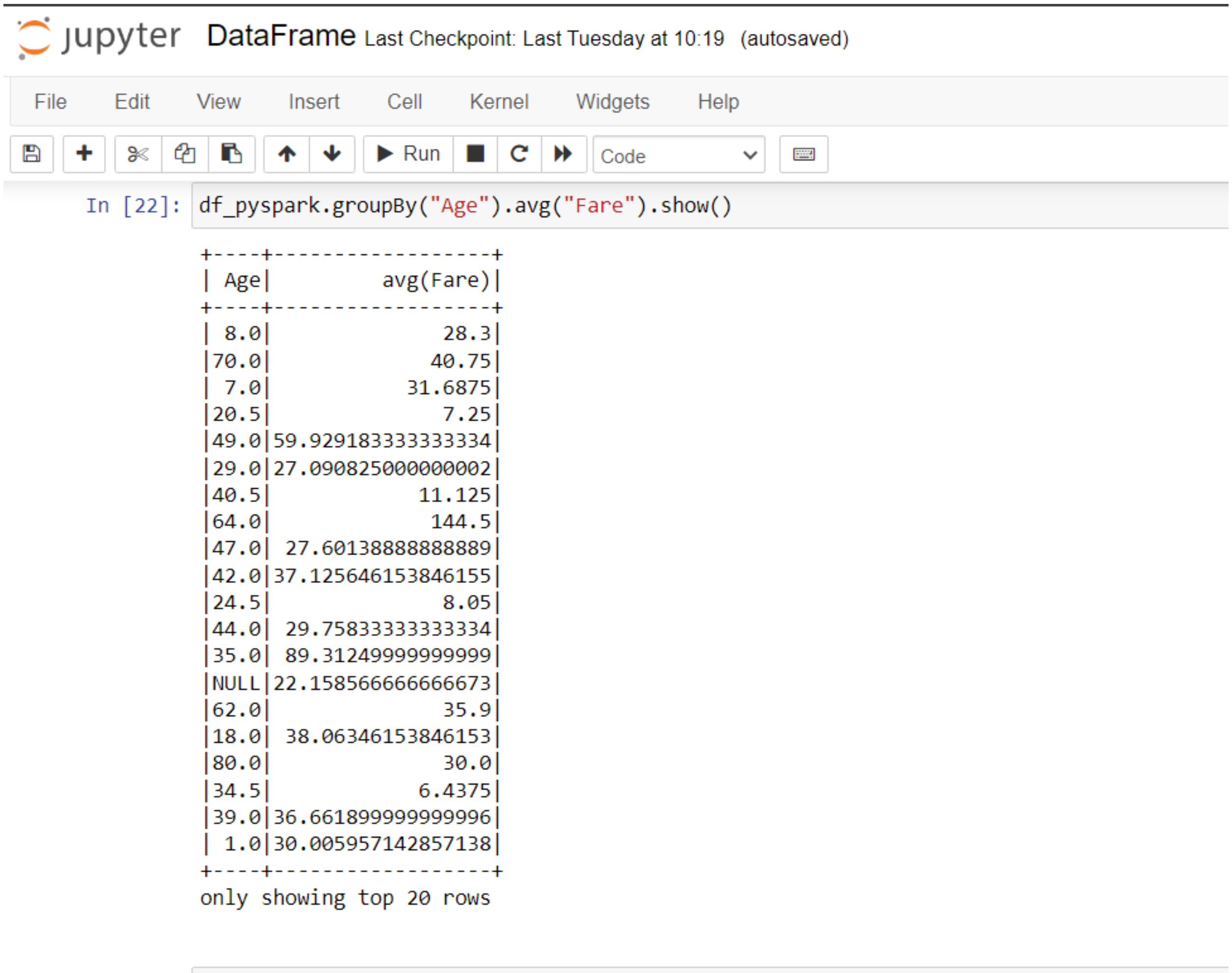
In [25]:
```
df_pyspark.groupBy("Age").min("Fare").show()
```

```
+----+---------+
| Age|min(Fare)|
+----+---------+
| 8.0|   21.075|
|70.0|     10.5|
| 7.0|    26.25|
|20.5|     7.25|
|49.0|      0.0|
|29.0|   7.0458|
|40.5|     7.75|
|64.0|     26.0|
|47.0|     7.25|
|42.0|     7.55|
|24.5|     8.05|
|44.0|    7.925|
|35.0|     7.05|
|NULL|      0.0|
|62.0|     10.5|
|18.0|   6.4958|
|80.0|     30.0|
|34.5|   6.4375|
|39.0|      0.0|
| 1.0|  11.1333|
+----+---------+
only showing top 20 rows
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run
Code

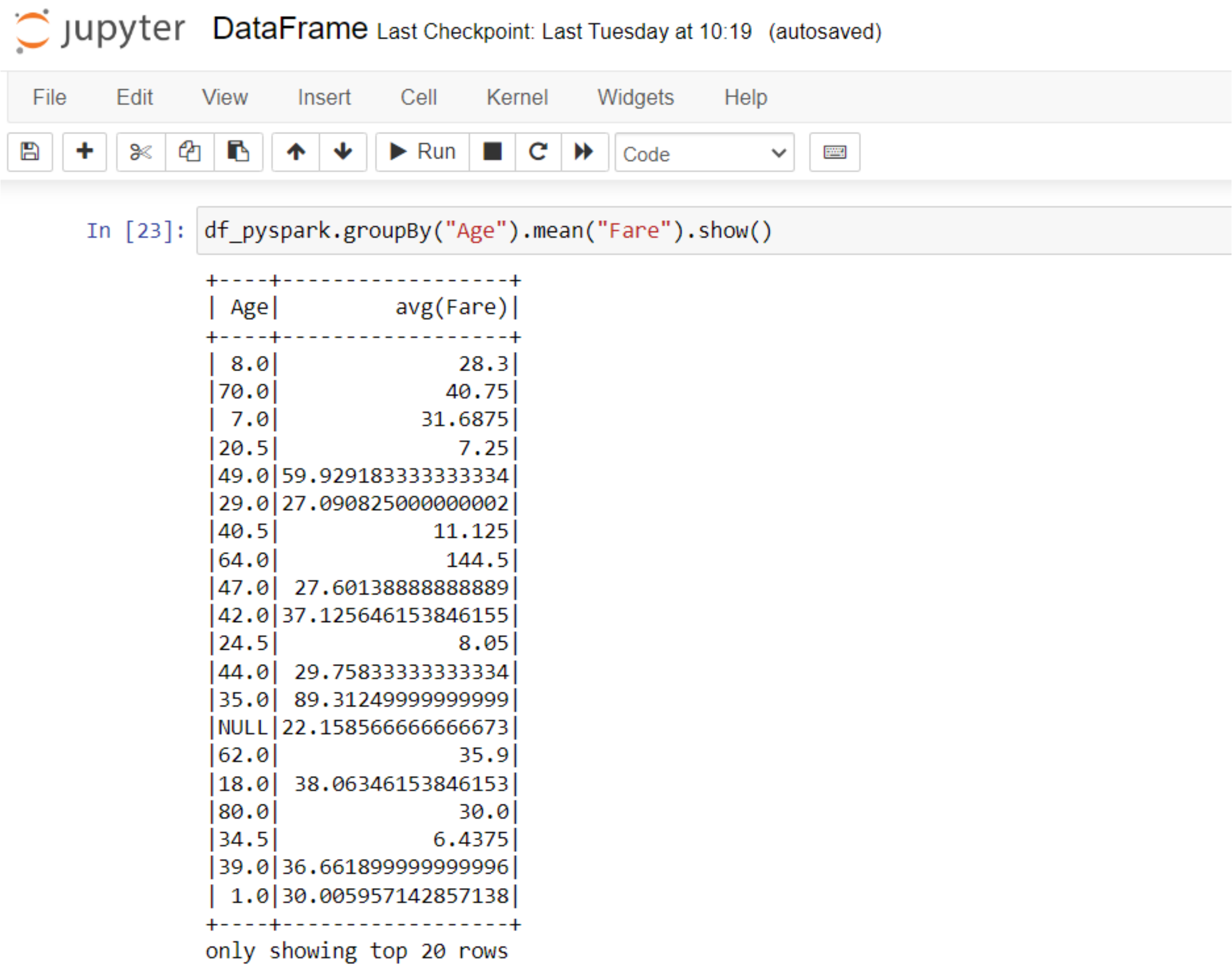
```
In [21]: df_pyspark.groupBy("Age").max("Fare").show()
```

```
+----+---------+
| Age|max(Fare)|
+----+---------+
| 8.0|    36.75|
|70.0|     71.0|
| 7.0|  39.6875|
|20.5|     7.25|
|49.0| 110.8833|
|29.0| 211.3375|
|40.5|     14.5|
|64.0|    263.0|
|47.0|  52.5542|
|42.0|  227.525|
|24.5|     8.05|
|44.0|     90.0|
|35.0| 512.3292|
|NULL|  227.525|
|62.0|     80.0|
|18.0|  262.375|
|80.0|     30.0|
|34.5|   6.4375|
|39.0| 110.8833|
| 1.0|     46.9|
+----+---------+
only showing top 20 rows
```


jupyter
DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

In [22]:
```
df_pyspark.groupBy("Age").avg("Fare").show()
```

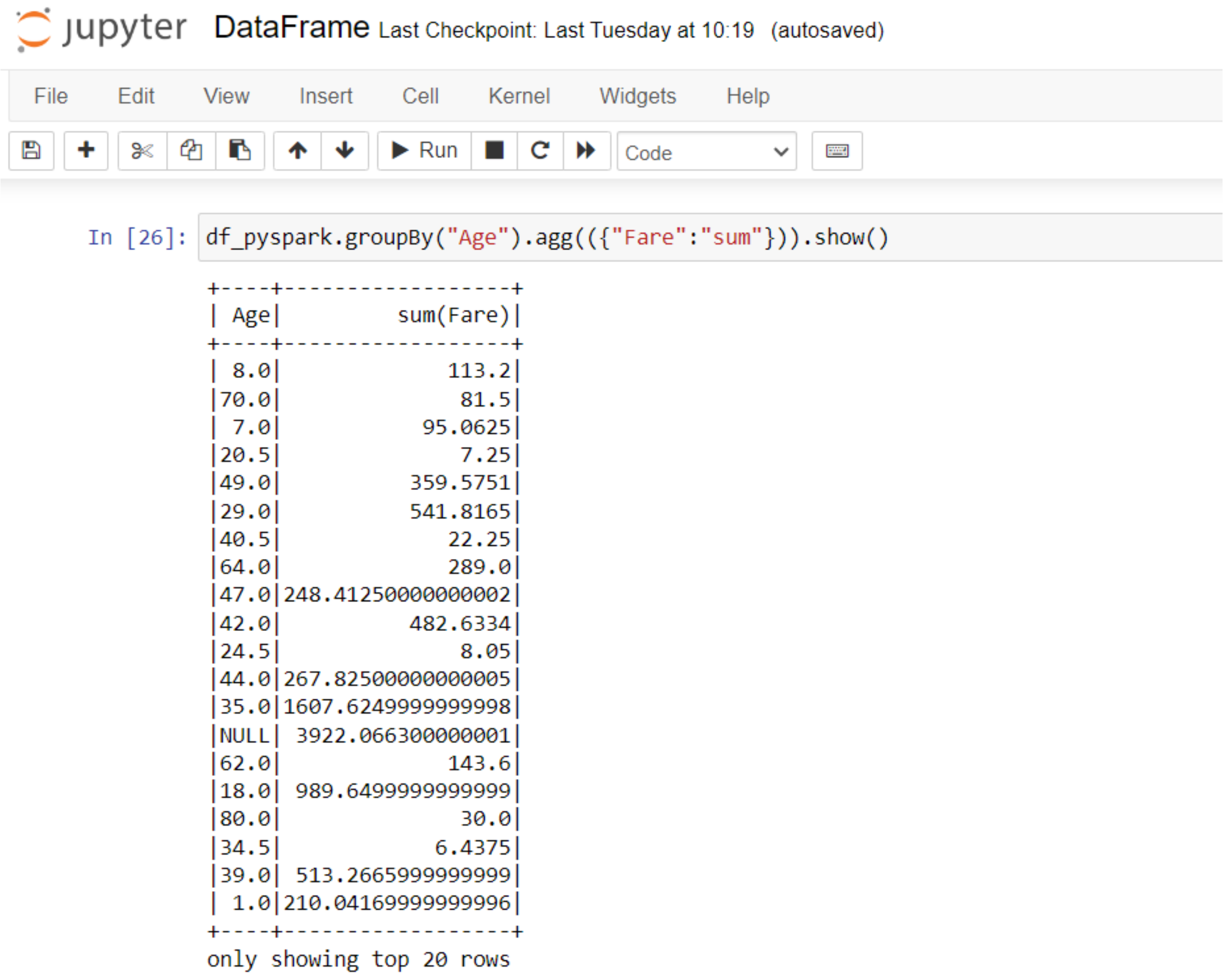
```
+----+------------------+
| Age|         avg(Fare)|
+----+------------------+
| 8.0|              28.3|
|70.0|             40.75|
| 7.0|           31.6875|
|20.5|              7.25|
|49.0|59.929183333333334|
|29.0|27.090825000000002|
|40.5|            11.125|
|64.0|             144.5|
|47.0| 27.60138888888889|
|42.0|37.125646153846155|
|24.5|              8.05|
|44.0| 29.75833333333334|
|35.0| 89.31249999999999|
|NULL|22.158566666666673|
|62.0|              35.9|
|18.0| 38.06346153846153|
|80.0|              30.0|
|34.5|            6.4375|
|39.0|36.661899999999996|
| 1.0|30.005957142857138|
+----+------------------+
only showing top 20 rows
```



jupyter
DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

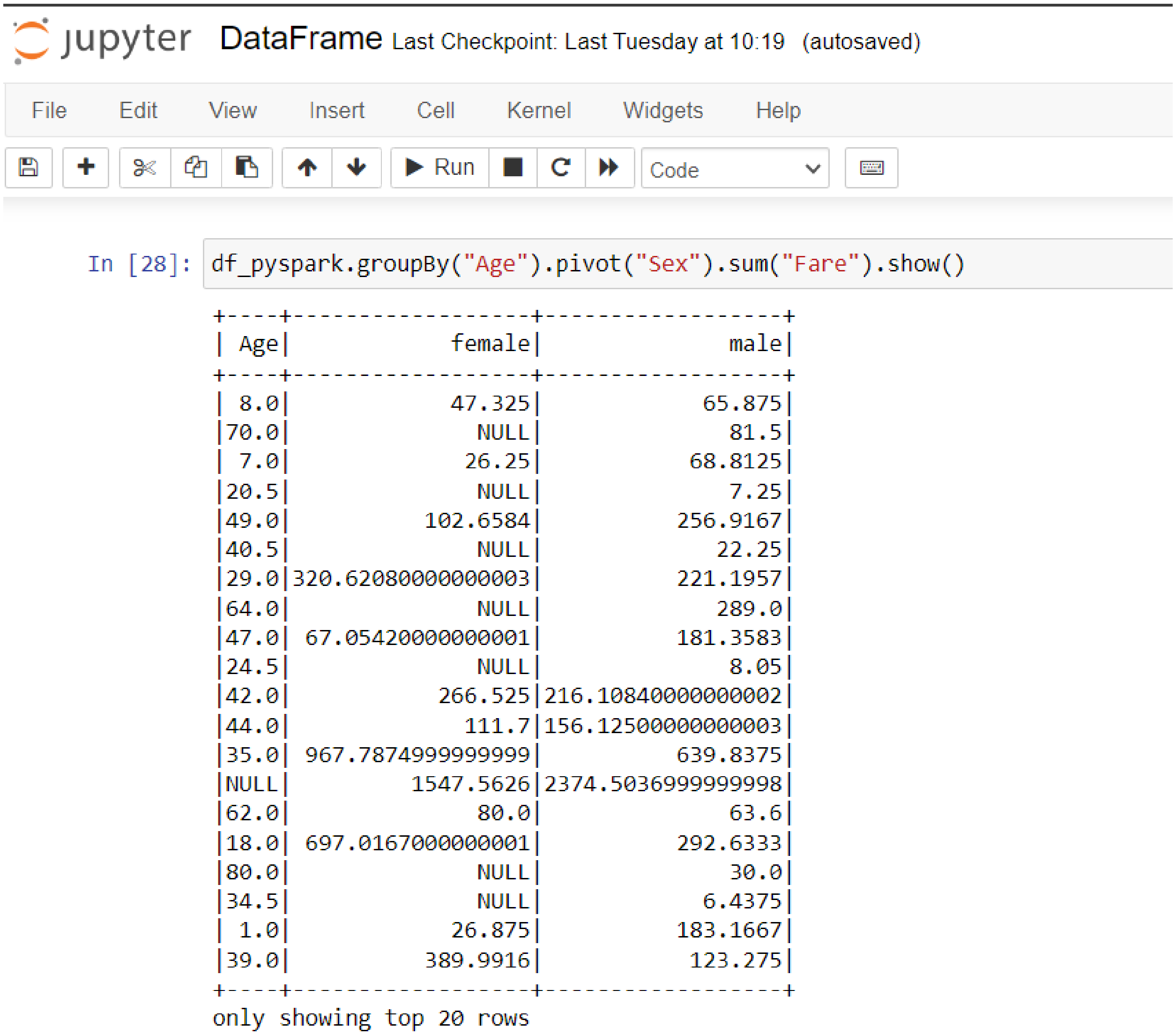
```
In [23]: df_pyspark.groupBy("Age").mean("Fare").show()
```

```
+----+------------------+
| Age|         avg(Fare)|
+----+------------------+
| 8.0|              28.3|
|70.0|             40.75|
| 7.0|           31.6875|
|20.5|              7.25|
|49.0|59.929183333333334|
|29.0|27.090825000000002|
|40.5|            11.125|
|64.0|             144.5|
|47.0| 27.60138888888889|
|42.0|37.125646153846155|
|24.5|              8.05|
|44.0| 29.75833333333334|
|35.0| 89.31249999999999|
|NULL|22.158566666666673|
|62.0|              35.9|
|18.0| 38.06346153846153|
|80.0|              30.0|
|34.5|            6.4375|
|39.0|36.661899999999996|
| 1.0|30.005957142857138|
+----+------------------+
only showing top 20 rows
```


jupyter
DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

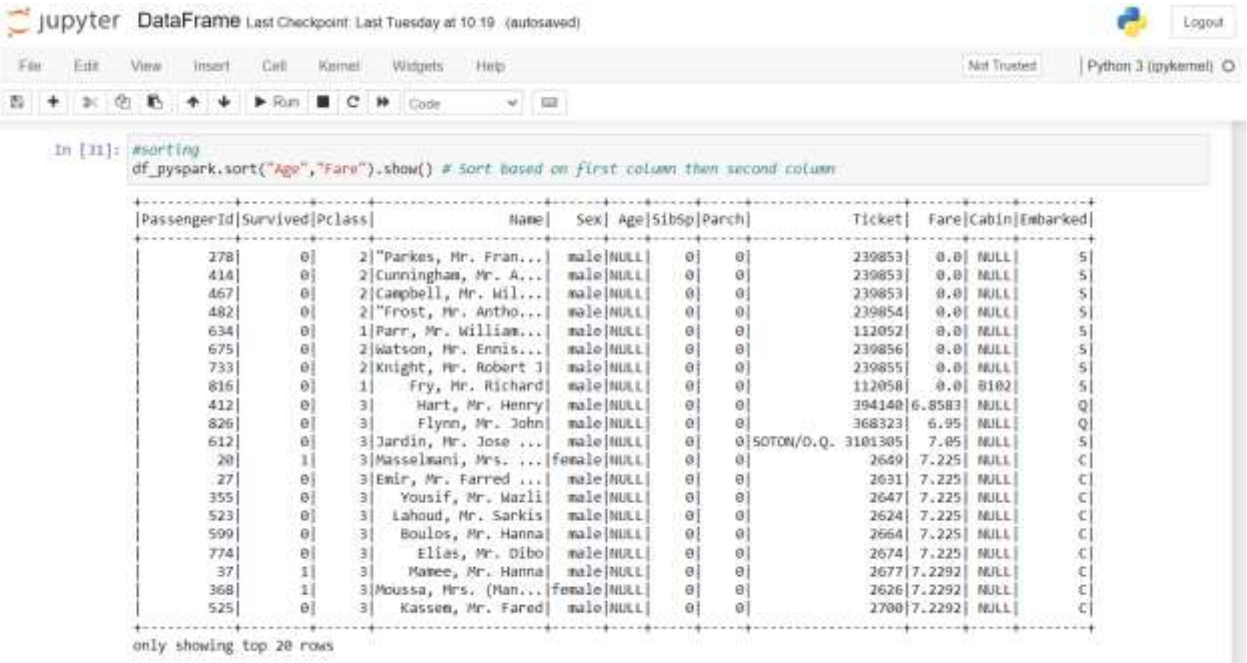
```
In [26]: df_pyspark.groupBy("Age").agg(({"Fare":"sum"})).show()
```

```
+----+-----------------+
| Age|        sum(Fare)|
+----+-----------------+
| 8.0|            113.2|
|70.0|             81.5|
| 7.0|          95.0625|
|20.5|             7.25|
|49.0|         359.5751|
|29.0|         541.8165|
|40.5|            22.25|
|64.0|            289.0|
|47.0|248.41250000000002|
|42.0|         482.6334|
|24.5|             8.05|
|44.0|267.82500000000005|
|35.0|1607.6249999999998|
|NULL| 3922.066300000001|
|62.0|            143.6|
|18.0| 989.6499999999999|
|80.0|             30.0|
|34.5|           6.4375|
|39.0| 513.2665999999999|
| 1.0|210.04169999999996|
+----+-----------------+
only showing top 20 rows
```


jupyter
DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code

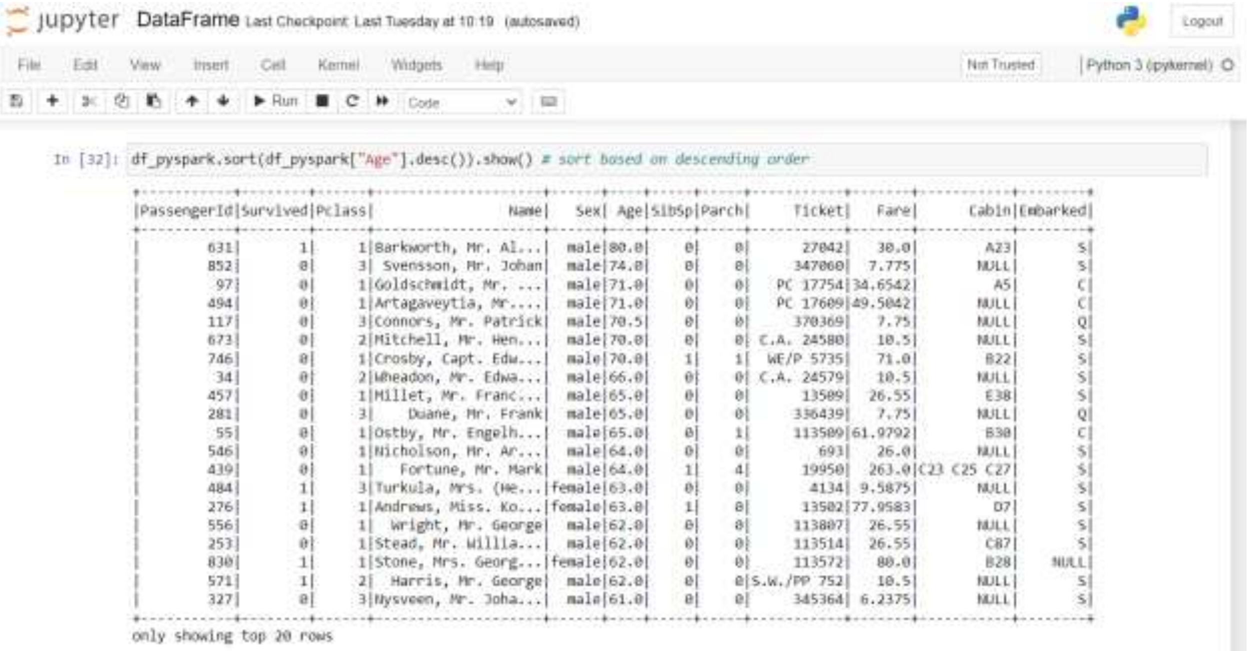
In [28]:
```
df_pyspark.groupBy("Age").pivot("Sex").sum("Fare").show()
```

```
+----+------------------+------------------+
| Age|            female|              male|
+----+------------------+------------------+
| 8.0|            47.325|            65.875|
|70.0|              NULL|              81.5|
| 7.0|             26.25|           68.8125|
|20.5|              NULL|              7.25|
|49.0|          102.6584|          256.9167|
|40.5|              NULL|             22.25|
|29.0|320.62080000000003|          221.1957|
|64.0|              NULL|             289.0|
|47.0| 67.05420000000001|          181.3583|
|24.5|              NULL|              8.05|
|42.0|           266.525|216.10840000000002|
|44.0|             111.7|156.12500000000003|
|35.0| 967.7874999999999|          639.8375|
|NULL|         1547.5626|2374.5036999999998|
|62.0|              80.0|              63.6|
|18.0| 697.0167000000001|          292.6333|
|80.0|              NULL|              30.0|
|34.5|              NULL|            6.4375|
| 1.0|            26.875|          183.1667|
|39.0|          389.9916|           123.275|
+----+------------------+------------------+
only showing top 20 rows
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

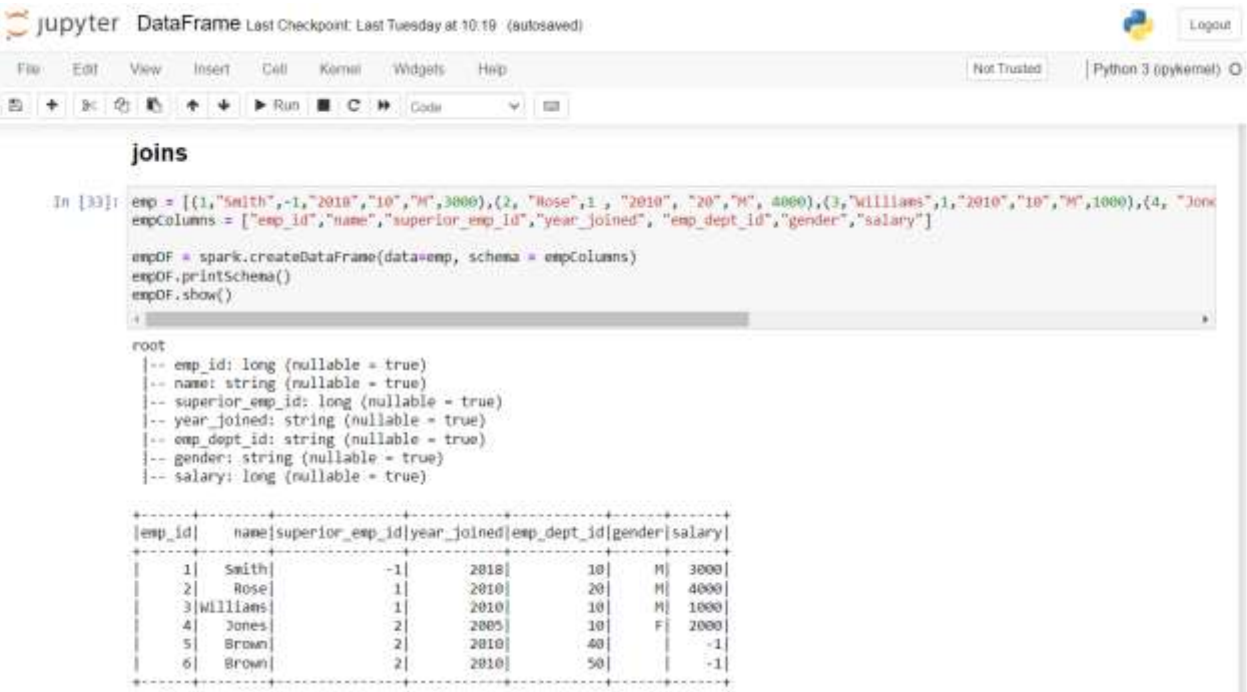
```
In [31]: #sorting
         df_pyspark.sort("Age","Fare").show() # Sort based on first column then second column
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+------------------+------+-----+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|            Ticket|  Fare|Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+------------------+------+-----+--------+
|        278|       0|     2|"Parkes, Mr. Fran...|  male|NULL|    0|    0|            239853|   0.0| NULL|       S|
|        414|       0|     2|Cunningham, Mr. A...|  male|NULL|    0|    0|            239853|   0.0| NULL|       S|
|        467|       0|     2|Campbell, Mr. Wil...|  male|NULL|    0|    0|            239853|   0.0| NULL|       S|
|        482|       0|     2|"Frost, Mr. Antho...|  male|NULL|    0|    0|            239854|   0.0| NULL|       S|
|        634|       0|     1|Parr, Mr. William...|  male|NULL|    0|    0|            112052|   0.0| NULL|       S|
|        675|       0|     2|Watson, Mr. Ennis...|  male|NULL|    0|    0|            239856|   0.0| NULL|       S|
|        733|       0|     2|Knight, Mr. Robert J|  male|NULL|    0|    0|            239855|   0.0| NULL|       S|
|        816|       0|     1|    Fry, Mr. Richard|  male|NULL|    0|    0|            112058|   0.0| B102|       S|
|        412|       0|     3|     Hart, Mr. Henry|  male|NULL|    0|    0|            394140|6.8583| NULL|       Q|
|        826|       0|     3|     Flynn, Mr. John|  male|NULL|    0|    0|            368323|  6.95| NULL|       Q|
|        612|       0|     3|Jardin, Mr. Jose ...|  male|NULL|    0|    0|SOTON/O.Q. 3101305|  7.05| NULL|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|              2649| 7.225| NULL|       C|
|         27|       0|     3|Emir, Mr. Farred ...|  male|NULL|    0|    0|              2631| 7.225| NULL|       C|
|        355|       0|     3|   Yousif, Mr. Wazli|  male|NULL|    0|    0|              2647| 7.225| NULL|       C|
|        523|       0|     3|  Lahoud, Mr. Sarkis|  male|NULL|    0|    0|              2624| 7.225| NULL|       C|
|        599|       0|     3|   Boulos, Mr. Hanna|  male|NULL|    0|    0|              2664| 7.225| NULL|       C|
|        774|       0|     3|     Elias, Mr. Dibo|  male|NULL|    0|    0|              2674| 7.225| NULL|       C|
|         37|       1|     3|    Mamee, Mr. Hanna|  male|NULL|    0|    0|              2677|7.2292| NULL|       C|
|        368|       1|     3|Moussa, Mrs. (Man...|female|NULL|    0|    0|              2626|7.2292| NULL|       C|
|        525|       0|     3|   Kassem, Mr. Fared|  male|NULL|    0|    0|              2700|7.2292| NULL|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+------------------+------+-----+--------+
only showing top 20 rows
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
In [32]: df_pyspark.sort(df_pyspark["Age"].desc()).show() # sort based on descending order
```

```
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+-----------+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|     Ticket|   Fare|      Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+-----------+--------+
|        631|       1|     1|Barkworth, Mr. Al...|  male|80.0|    0|    0|      27042|   30.0|        A23|       S|
|        852|       0|     3| Svensson, Mr. Johan|  male|74.0|    0|    0|     347060|  7.775|       NULL|       S|
|         97|       0|     1|Goldschmidt, Mr. ...|  male|71.0|    0|    0|   PC 17754|34.6542|         A5|       C|
|        494|       0|     1|Artagaveytia, Mr....|  male|71.0|    0|    0|   PC 17609|49.5042|       NULL|       C|
|        117|       0|     3|Connors, Mr. Patrick|  male|70.5|    0|    0|     370369|   7.75|       NULL|       Q|
|        673|       0|     2|Mitchell, Mr. Hen...|  male|70.0|    0|    0| C.A. 24580|   10.5|       NULL|       S|
|        746|       0|     1|Crosby, Capt. Edw...|  male|70.0|    1|    1|  WE/P 5735|   71.0|        B22|       S|
|         34|       0|     2|Wheadon, Mr. Edwa...|  male|66.0|    0|    0| C.A. 24579|   10.5|       NULL|       S|
|        457|       0|     1|Millet, Mr. Franc...|  male|65.0|    0|    0|      13509|  26.55|        E38|       S|
|        281|       0|     3|    Duane, Mr. Frank|  male|65.0|    0|    0|     336439|   7.75|       NULL|       Q|
|         55|       0|     1|Ostby, Mr. Engelh...|  male|65.0|    0|    1|     113509|61.9792|        B30|       C|
|        546|       0|     1|Nicholson, Mr. Ar...|  male|64.0|    0|    0|        693|   26.0|       NULL|       S|
|        439|       0|     1|   Fortune, Mr. Mark|  male|64.0|    1|    4|      19950|  263.0|C23 C25 C27|       S|
|        484|       1|     3|Turkula, Mrs. (He...|female|63.0|    0|    0|       4134| 9.5875|       NULL|       S|
|        276|       1|     1|Andrews, Miss. Ko...|female|63.0|    1|    0|      13502|77.9583|         D7|       S|
|        556|       0|     1|  Wright, Mr. George|  male|62.0|    0|    0|     113807|  26.55|       NULL|       S|
|        253|       0|     1|Stead, Mr. William...|  male|62.0|    0|    0|     113514|  26.55|        C87|       S|
|        830|       1|     1|Stone, Mrs. Georg...|female|62.0|    0|    0|     113572|   80.0|        B28|    NULL|
|        571|       1|     2|  Harris, Mr. George|  male|62.0|    0|    0|S.W./PP 752|   10.5|       NULL|       S|
|        327|       0|     3|Nysveen, Mr. Joha...|  male|61.0|    0|    0|     345364| 6.2375|       NULL|       S|
+-----------+--------+------+--------------------+------+----+-----+-----+-----------+-------+-----------+--------+
only showing top 20 rows
```



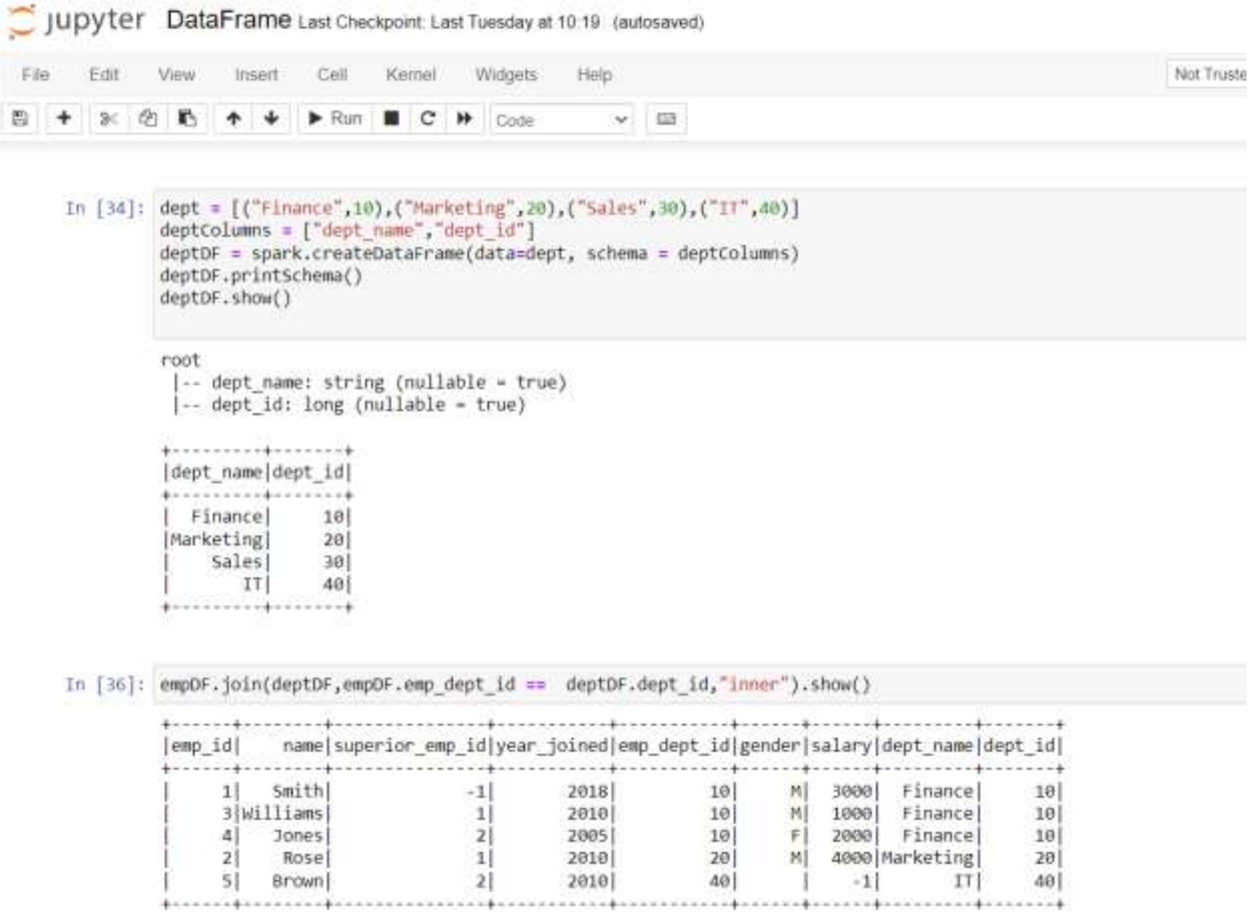
jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

joins

In [33]:

```
emp = [(1,"Smith",-1,"2018","10","M",3000),(2, "Rose",1 , "2010", "20","M", 4000),(3,"Williams",1,"2010","10","M",1000),(4, "Jone
empColumns = ["emp_id","name","superior_emp_id","year_joined", "emp_dept_id","gender","salary"]

empDF = spark.createDataFrame(data=emp, schema = empColumns)
empDF.printSchema()
empDF.show()
```

```
root
 |-- emp_id: long (nullable = true)
 |-- name: string (nullable = true)
 |-- superior_emp_id: long (nullable = true)
 |-- year_joined: string (nullable = true)
 |-- emp_dept_id: string (nullable = true)
 |-- gender: string (nullable = true)
 |-- salary: long (nullable = true)

+------+--------+---------------+-----------+-----------+------+------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|
+------+--------+---------------+-----------+-----------+------+------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|
|     2|    Rose|              1|       2010|         20|     M|  4000|
|     3|Williams|              1|       2010|         10|     M|  1000|
|     4|   Jones|              2|       2005|         10|     F|  2000|
|     5|   Brown|              2|       2010|         40|      |    -1|
|     6|   Brown|              2|       2010|         50|      |    -1|
+------+--------+---------------+-----------+-----------+------+------+
```



jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Truste
Run
Code

In [34]:

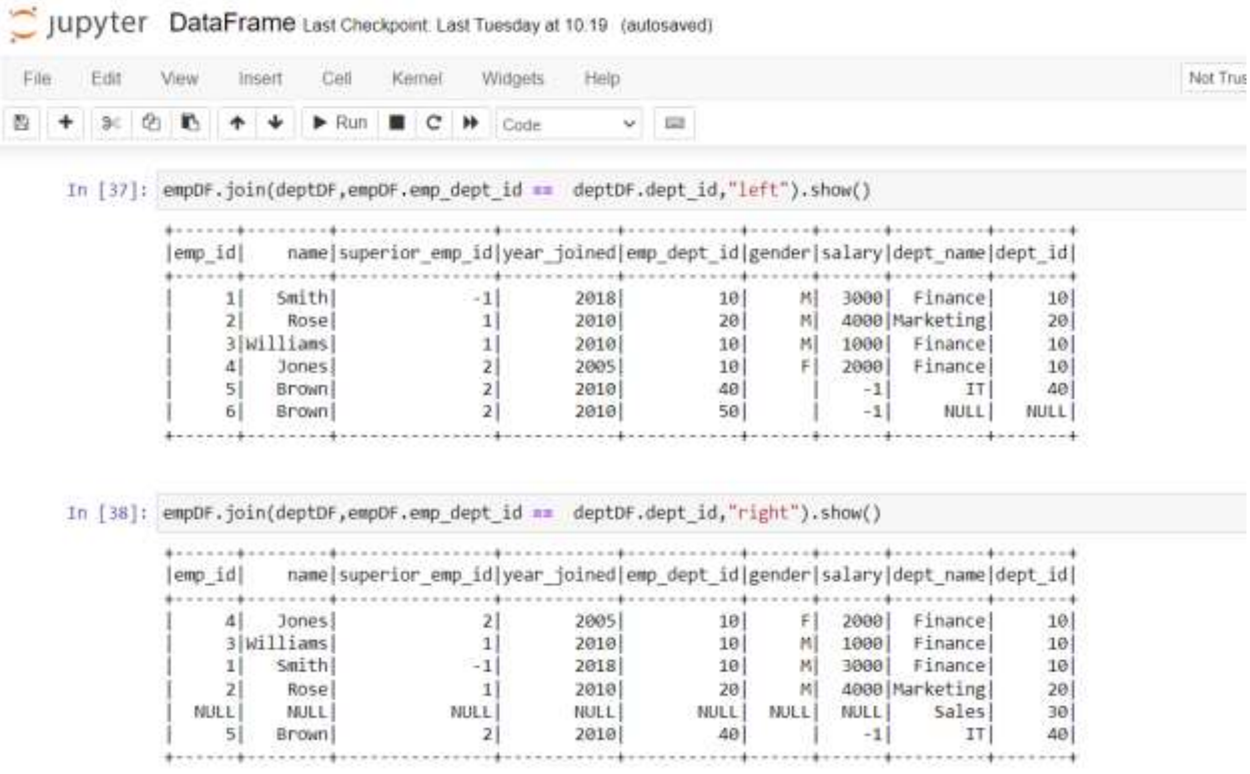
```
dept = [("Finance",10),("Marketing",20),("Sales",30),("IT",40)]
deptColumns = ["dept_name","dept_id"]
deptDF = spark.createDataFrame(data=dept, schema = deptColumns)
deptDF.printSchema()
deptDF.show()
```

```
root
 |-- dept_name: string (nullable = true)
 |-- dept_id: long (nullable = true)

+---------+-------+
|dept_name|dept_id|
+---------+-------+
|  Finance|     10|
|Marketing|     20|
|    Sales|     30|
|       IT|     40|
+---------+-------+
```

In [36]:

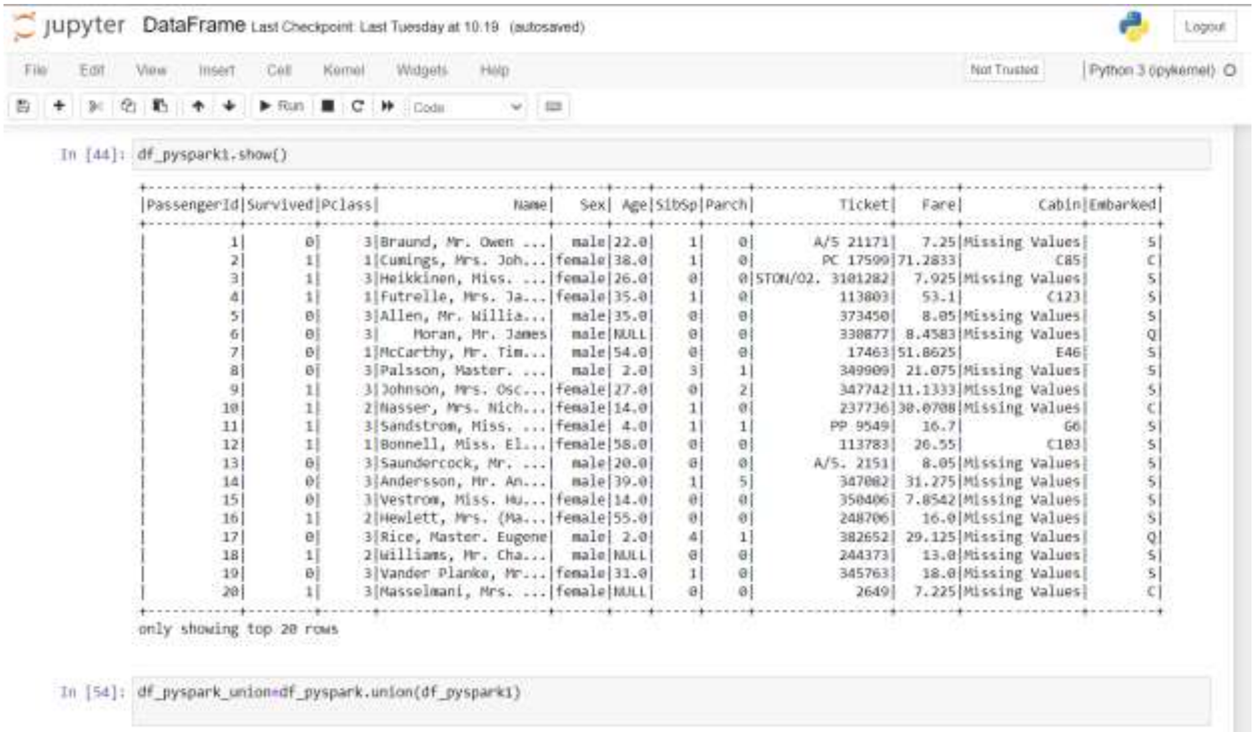
```
empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"inner").show()
```

```
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```


jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trus
Run
Code

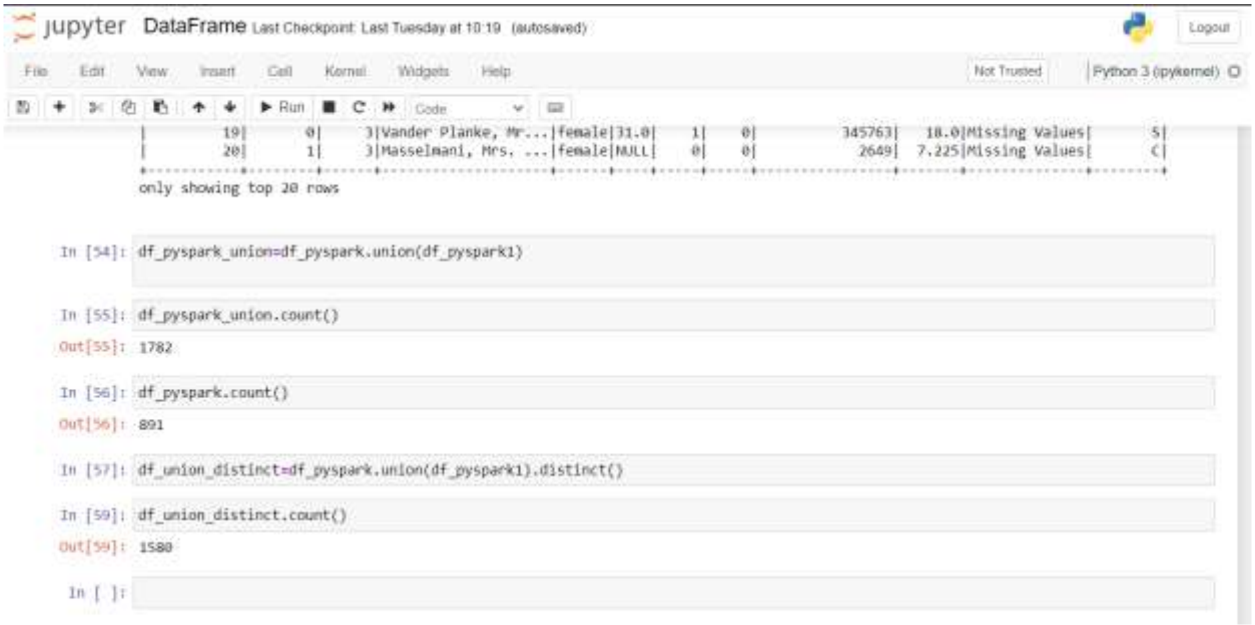
```
In [37]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"left").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
|     6|   Brown|              2|       2010|         50|      |    -1|     NULL|   NULL|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```

```
In [38]: empDF.join(deptDF,empDF.emp_dept_id ==  deptDF.dept_id,"right").show()

+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|emp_id|    name|superior_emp_id|year_joined|emp_dept_id|gender|salary|dept_name|dept_id|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
|     4|   Jones|              2|       2005|         10|     F|  2000|  Finance|     10|
|     3|Williams|              1|       2010|         10|     M|  1000|  Finance|     10|
|     1|   Smith|             -1|       2018|         10|     M|  3000|  Finance|     10|
|     2|    Rose|              1|       2010|         20|     M|  4000|Marketing|     20|
|  NULL|    NULL|           NULL|       NULL|       NULL|  NULL|  NULL|    Sales|     30|
|     5|   Brown|              2|       2010|         40|      |    -1|       IT|     40|
+------+--------+---------------+-----------+-----------+------+------+---------+-------+
```



```
In [44]: df_pyspark1.show()

+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+--------------+--------+
|PassengerId|Survived|Pclass|                Name|   Sex| Age|SibSp|Parch|          Ticket|   Fare|         Cabin|Embarked|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+--------------+--------+
|          1|       0|     3|Braund, Mr. Owen ...|  male|22.0|    1|    0|       A/5 21171|   7.25|Missing Values|       S|
|          2|       1|     1|Cumings, Mrs. Joh...|female|38.0|    1|    0|        PC 17599|71.2833|           C85|       C|
|          3|       1|     3|Heikkinen, Miss. ...|female|26.0|    0|    0|STON/O2. 3101282|  7.925|Missing Values|       S|
|          4|       1|     1|Futrelle, Mrs. Ja...|female|35.0|    1|    0|          113803|   53.1|          C123|       S|
|          5|       0|     3|Allen, Mr. Willia...|  male|35.0|    0|    0|          373450|   8.05|Missing Values|       S|
|          6|       0|     3|    Moran, Mr. James|  male|NULL|    0|    0|          330877| 8.4583|Missing Values|       Q|
|          7|       0|     1|McCarthy, Mr. Tim...|  male|54.0|    0|    0|           17463|51.8625|           E46|       S|
|          8|       0|     3|Palsson, Master. ...|  male| 2.0|    3|    1|          349909| 21.075|Missing Values|       S|
|          9|       1|     3|Johnson, Mrs. Osc...|female|27.0|    0|    2|          347742|11.1333|Missing Values|       S|
|         10|       1|     2|Nasser, Mrs. Nich...|female|14.0|    1|    0|          237736|30.0708|Missing Values|       C|
|         11|       1|     3|Sandstrom, Miss. ...|female| 4.0|    1|    1|         PP 9549|   16.7|            G6|       S|
|         12|       1|     1|Bonnell, Miss. El...|female|58.0|    0|    0|          113783|  26.55|          C103|       S|
|         13|       0|     3|Saundercock, Mr. ...|  male|20.0|    0|    0|       A/5. 2151|   8.05|Missing Values|       S|
|         14|       0|     3|Andersson, Mr. An...|  male|39.0|    1|    5|          347082| 31.275|Missing Values|       S|
|         15|       0|     3|Vestrom, Miss. Hu...|female|14.0|    0|    0|          350406| 7.8542|Missing Values|       S|
|         16|       1|     2|Hewlett, Mrs. (Ma...|female|55.0|    0|    0|          248706|   16.0|Missing Values|       S|
|         17|       0|     3|Rice, Master. Eugene|  male| 2.0|    4|    1|          382652| 29.125|Missing Values|       Q|
|         18|       1|     2|Williams, Mr. Cha...|  male|NULL|    0|    0|          244373|   13.0|Missing Values|       S|
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0|Missing Values|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225|Missing Values|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+--------------+--------+
only showing top 20 rows
```

```
In [54]: df_pyspark_union=df_pyspark.union(df_pyspark1)
```



jupyter DataFrame Last Checkpoint: Last Tuesday at 10:19 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

```
|         19|       0|     3|Vander Planke, Mr...|female|31.0|    1|    0|          345763|   18.0|Missing Values|       S|
|         20|       1|     3|Masselmani, Mrs. ...|female|NULL|    0|    0|            2649|  7.225|Missing Values|       C|
+-----------+--------+------+--------------------+------+----+-----+-----+----------------+-------+--------------+--------+
only showing top 20 rows
```

```
In [54]: df_pyspark_union=df_pyspark.union(df_pyspark1)
```

```
In [55]: df_pyspark_union.count()
Out[55]: 1782
```

```
In [56]: df_pyspark.count()
Out[56]: 891
```

```
In [57]: df_union_distinct=df_pyspark.union(df_pyspark1).distinct()
```

```
In [59]: df_union_distinct.count()
Out[59]: 1580
```

```
In [ ]:
```