Spark Dataframes

Tom Bohbot

May 2021

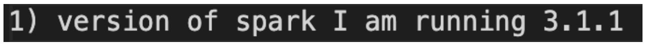
1 Introduction

Questions 0, 4, and 6 do not display photos in the report as no output is necessary for these questions. Additionally, question 2 uses a dataframe instead of dataset (as discussed with Dr. Leff). I did not include code in screenshots as the code is pushed to Github. If necessary please see code in file called SparkDF1.py.

2 0) Initialize the Spark session for your application in "local" mode.

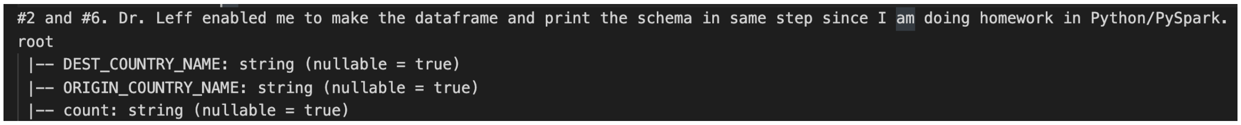
No output necessary.

3 1) Print the version of Spark on which this application is running.



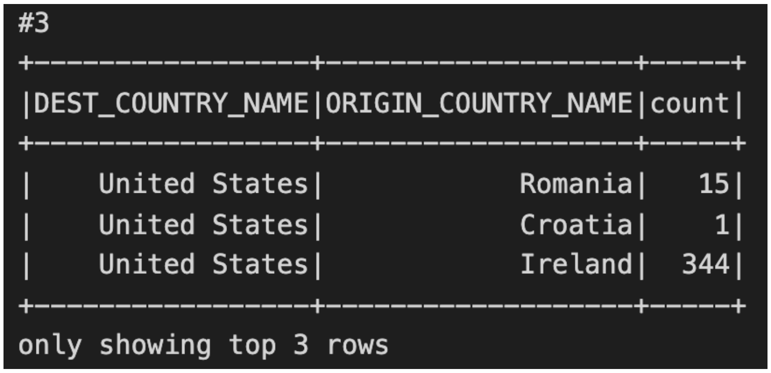
```
1) version of spark I am running 3.1.1
```

4 2) Load the 2015 csv flight data, and display the schema of the resulting Dataset Row



```
#2 and #6. Dr. Leff enabled me to make the dataframe and print the schema in same step since I am doing homework in Python/PySpark.
root
 |-- DEST_COUNTRY_NAME: string (nullable = true)
 |-- ORIGIN_COUNTRY_NAME: string (nullable = true)
 |-- count: string (nullable = true)
```

5 3) Print three tuples (rows) of the data

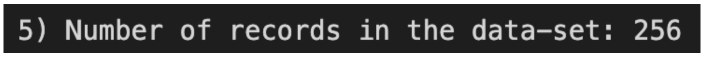


```
#3
+-----------------+-------------------+-----+
|DEST_COUNTRY_NAME|ORIGIN_COUNTRY_NAME|count|
+-----------------+-------------------+-----+
|    United States|            Romania|   15|
|    United States|            Croatia|    1|
|    United States|            Ireland|  344|
+-----------------+-------------------+-----+
only showing top 3 rows
```

6 4) Save the flight data to a postgreSQL relational db

No output necessary.

7 5) Number of records in the data-set.

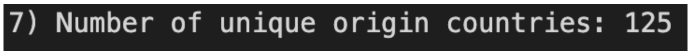


```
5) Number of records in the data-set: 256
```

8 6) Create a temporary view using the csv DataFrame

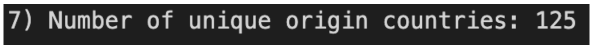
no output necessary

9 7) Number of unique origin countries



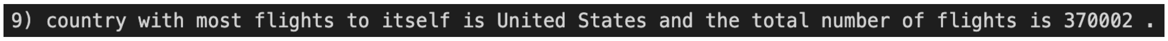
```
7) Number of unique origin countries: 125
```

10 8) How many rows are associated with the destination country that has the largest number of rows in the data-set?



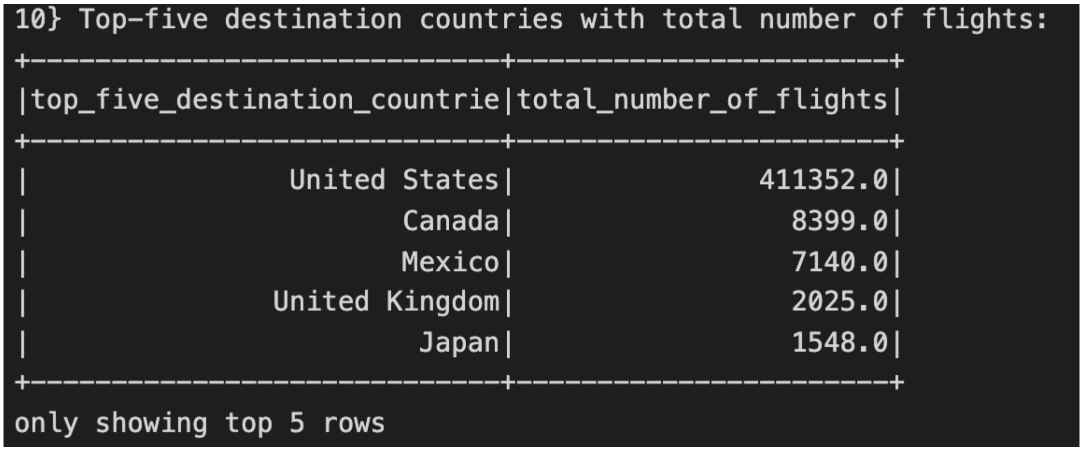
```
7) Number of unique origin countries: 125
```

11 9) Which country has the most flights to itself, and what are the number of such flights?



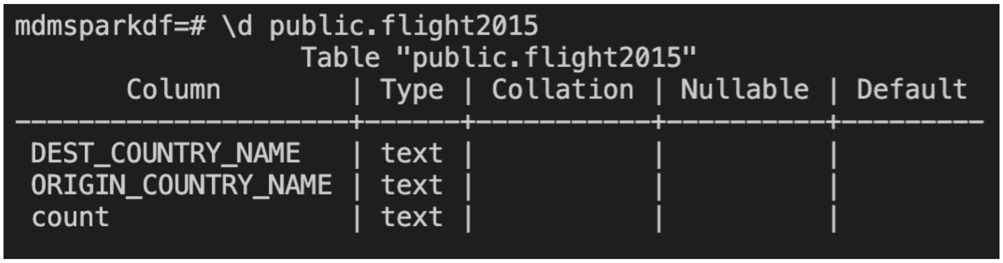
```
9) country with most flights to itself is United States and the total number of flights is 370002 .
```

12 10)Top-five destination countries (by total number of flights). Report both the destination countries and the total number of flights.

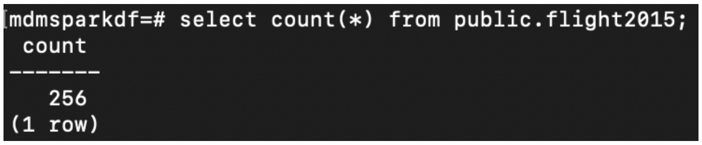


```
10} Top-five destination countries with total number of flights:
+-----------------------------+-----------------------+
|top_five_destination_countrie|total_number_of_flights|
+-----------------------------+-----------------------+
|                United States|               411352.0|
|                       Canada|                 8399.0|
|                       Mexico|                 7140.0|
|               United Kingdom|                 2025.0|
|                        Japan|                 1548.0|
+-----------------------------+-----------------------+
only showing top 5 rows
```

13 11)Show that your Spark program actually created a database table from the csv data, by including a screen-shot of a psql-based inter- action with the flight-data table in which you execute the following commands:



```
mdmsparkdf=# \d public.flight2015
                  Table "public.flight2015"
       Column        | Type | Collation | Nullable | Default
---------------------+------+-----------+----------+---------
 DEST_COUNTRY_NAME   | text |           |          |
 ORIGIN_COUNTRY_NAME | text |           |          |
 count               | text |           |          |
```



```
mdmsparkdf=# select count(*) from public.flight2015;
 count
-------
   256
(1 row)
```