

Quiz 5

MSAN 694, Diane Woodbridge

Description

Use the given *app_name*, *file_name_1*, *file_name_2* and *supervisor_id* in “user_definition.py” to complete the python code “quiz5.py”.

```
app_name = "sparksql_basic"
file_name_1 = "filtered_registered_business_sf.csv"
file_name_2 = "supervisor_sf.csv"
supervisor_id = 1
```

Requirement:

Using the input files, generate a DataFrame with the format of

```
zip(integer, not- nullable)
supervisor_id( string,nullable)
business_name( string,nullable)
street(string, nullable)
city (string, nullable)
state (string, nullable)
```

input example..



zip, supervisor_id

supervisor_sf.csv

```
94102,8
94102,6
94102,3
94102,5
94103,8
94103,0
```



filtered_registered_business_sf.csv

zip, business_name, street, city, state

```
94124,Stephens Institute Inc,2225 Jerrold Ave,San Francisco,CA
94105,Stephens Institute Inc,180 New Montgomery St,San Francisco,CA
94108,Stephens Institute Inc,540 Powell St,San Francisco,CA
94107,Stephens Institute Inc,460 Townsend St,San Francisco,CA
94109,Stephens Institute Inc,1835-49 Van Ness Ave,San Francisco,CA
94102,Stephens Institute Inc,620 Sutter St,San Francisco,CA
94102,Stephens Institute Inc,655 Sutter St,San Francisco,CA
94109,Stephens Institute Inc,1055 Pine St,San Francisco,CA
94107,Stephens Institute Inc,121 Wisconsin St,San Francisco,CA
94102,Stephens Institute Inc,150 Hayes St,San Francisco,CA
94133,Stephens Institute Inc,2300 Stockton St,San Francisco,CA
94133,Stephens Institute Inc,2801 Leavenworth St,San Francisco,CA
```

The created DataFrame should **have only 5 digit zipcode values and should have no duplicate rows.**

“quiz5.py” should

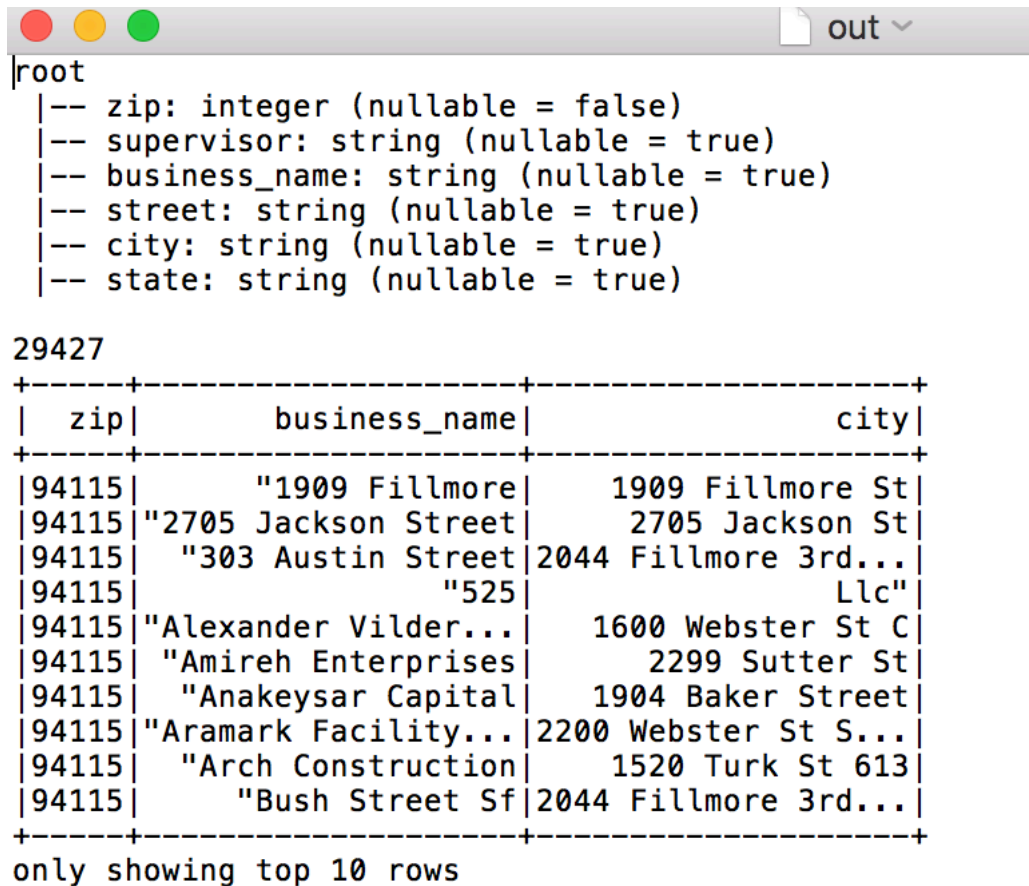
print the schema of your DataFrame (orders of the columns/rows do not matter.) ,

print the number of rows of the given supervisor_id in the joined DataFrame and

print 10 zip, business_name, city of the given supervisor_id sorted in ascending order by zip, business_name and city.

Submit the quiz5.py file (**ONLY**)- the name of your file should be quiz5_LastName_Firstname.py on the link below.

If you run spark-submit quiz5_Woodbridge_Diane.py > out, the output should look like as below.

A terminal window with a title bar containing three colored circles (red, yellow, green) and a tab labeled 'out'. The terminal displays the output of a Spark job. It starts with a 'root' prompt followed by a list of schema fields: zip (integer, nullable=false), supervisor (string, nullable=true), business_name (string, nullable=true), street (string, nullable=true), city (string, nullable=true), and state (string, nullable=true). Below this is a large number '29427'. Then, a table of 10 rows is shown, each with three columns: zip, business_name, and city. The data is truncated in the last two columns. At the bottom, it says 'only showing top 10 rows'.

```
root
|-- zip: integer (nullable = false)
|-- supervisor: string (nullable = true)
|-- business_name: string (nullable = true)
|-- street: string (nullable = true)
|-- city: string (nullable = true)
|-- state: string (nullable = true)

29427
+-----+-----+-----+
|  zip|      business_name|      city|
+-----+-----+-----+
|94115|      "1909 Fillmore|    1909 Fillmore St|
|94115|"2705 Jackson Street|    2705 Jackson St|
|94115|    "303 Austin Street|2044 Fillmore 3rd...|
|94115|              "525|              Llc"|
|94115|"Alexander Vilder...|    1600 Webster St C|
|94115|    "Amireh Enterprises|    2299 Sutter St|
|94115|    "Anakeysar Capital|    1904 Baker Street|
|94115|"Aramark Facility...|2200 Webster St S...|
|94115|    "Arch Construction|    1520 Turk St 613|
|94115|    "Bush Street Sf|2044 Fillmore 3rd...|
+-----+-----+-----+
only showing top 10 rows
```