SparkSQL PySpark-EDA

June 22, 2021

[142]:

```
#!pip install pyspark
import pyspark
from pyspark.sql import SparkSession
from pyspark.context import SparkContext
from pyspark.sql.functions import when
from pyspark.sql.functions import lit
```

[14]:

```
sc = SparkSession.builder.appName("SparkSQLExample")\
  .config("spark.sql.shuffle.partitions", "50")\
  .config("spark.driver.maxResultSize","5g")\
  .config ("spark.sql.execution.arrow.enabled", "true")\
  .getOrCreate()
```

[21]:

```
dataframe_csv = sc.read.csv('/content/drive/MyDrive/Bank_PySpark /bank.csv',␣
 ↪inferSchema=True, header=True)
```

[22]:

```
dataframe_csv.show()
```

```
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
|age|        job| marital|education|default|balance|housing|loan|contact|day|mon
th|duration|campaign|pdays|previous|poutcome|deposit|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
| 59|     admin.| married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|     admin.| married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41| technician| married|secondary|     no|   1270|    yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
| 55|   services| married|secondary|     no|   2476|    yes|  no|unknown|  5|
may|     579|       1|   -1|       0| unknown|    yes|
| 54|     admin.| married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
| 42| management|  single| tertiary|     no|      0|    yes| yes|unknown|  5|
may|     562|       2|   -1|       0| unknown|    yes|
| 56| management| married| tertiary|     no|    830|    yes| yes|unknown|  6|
may|    1201|       1|   -1|       0| unknown|    yes|
| 60|    retired|divorced|secondary|     no|    545|    yes|  no|unknown|  6|
may|    1030|       1|   -1|       0| unknown|    yes|
| 37| technician| married|secondary|     no|      1|    yes|  no|unknown|  6|
may|     608|       1|   -1|       0| unknown|    yes|
| 28|   services|  single|secondary|     no|   5090|    yes|  no|unknown|  6|
may|    1297|       3|   -1|       0| unknown|    yes|
| 38|     admin.|  single|secondary|     no|    100|    yes|  no|unknown|  7|
may|     786|       1|   -1|       0| unknown|    yes|
| 30|blue-collar| married|secondary|     no|    309|    yes|  no|unknown|  7|
may|    1574|       2|   -1|       0| unknown|    yes|
| 29| management| married| tertiary|     no|    199|    yes| yes|unknown|  7|
may|    1689|       4|   -1|       0| unknown|    yes|
| 46|blue-collar|  single| tertiary|     no|    460|    yes|  no|unknown|  7|
may|    1102|       2|   -1|       0| unknown|    yes|
| 31| technician|  single| tertiary|     no|    703|    yes|  no|unknown|  8|
may|     943|       2|   -1|       0| unknown|    yes|
| 35| management|divorced| tertiary|     no|   3837|    yes|  no|unknown|  8|
may|    1084|       1|   -1|       0| unknown|    yes|
| 32|blue-collar|  single|  primary|     no|    611|    yes|  no|unknown|  8|
may|     541|       3|   -1|       0| unknown|    yes|
| 49|   services| married|secondary|     no|     -8|    yes|  no|unknown|  8|
may|    1119|       1|   -1|       0| unknown|    yes|
| 41|     admin.| married|secondary|     no|     55|    yes|  no|unknown|  8|
may|    1120|       2|   -1|       0| unknown|    yes|
| 49|     admin.|divorced|secondary|     no|    168|    yes| yes|unknown|  8|
may|     513|       1|   -1|       0| unknown|    yes|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
only showing top 20 rows
```

[40]:
```
dataframe_csv.count() ##count # of rows
```

[40]:
```
11162
```

[41]:
```
len(dataframe_csv.columns) ##count # of columns
```

[41]:
```
17
```

[42]:
```
#Drop Duplicates (if any)
dataframe_dropduplicates = dataframe_csv.dropDuplicates()
dataframe_dropduplicates.count()
#No duplicate entries in this dataframe
```

[42]:
```
11162
```

[45]:
```
##SQL Queries
#Show all columns/features of dataframe with 10 entries
dataframe_csv.select("*").show(10)
```

```
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
```

```
-+--------+--------+-----+--------+--------+-------+
|age|       job| marital|education|default|balance|housing|loan|contact|day|mont
h|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
-+--------+--------+-----+--------+--------+-------+
| 59|    admin.| married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.| married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41|technician| married|secondary|     no|   1270|    yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
| 55|  services| married|secondary|     no|   2476|    yes|  no|unknown|  5|
may|     579|       1|   -1|       0| unknown|    yes|
| 54|    admin.| married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
| 42|management|  single| tertiary|     no|      0|    yes| yes|unknown|  5|
may|     562|       2|   -1|       0| unknown|    yes|
| 56|management| married| tertiary|     no|    830|    yes| yes|unknown|  6|
may|    1201|       1|   -1|       0| unknown|    yes|
| 60|   retired|divorced|secondary|     no|    545|    yes|  no|unknown|  6|
may|    1030|       1|   -1|       0| unknown|    yes|
| 37|technician| married|secondary|     no|      1|    yes|  no|unknown|  6|
may|     608|       1|   -1|       0| unknown|    yes|
| 28|  services|  single|secondary|     no|   5090|    yes|  no|unknown|  6|
may|    1297|       3|   -1|       0| unknown|    yes|
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
-+--------+--------+-----+--------+--------+-------+
only showing top 10 rows
```

[46]:

```
#Show top 10 entries for feature "job" in dataframe_csv
dataframe_csv.select("job").show(10)
```

```
+----------+
|       job|
+----------+
|    admin.|
|    admin.|
|technician|
|  services|
|    admin.|
|management|
|management|
|   retired|
|technician|
|  services|
+----------+
only showing top 10 rows
```

[48]:

```
#Show a subset of features for dataframe_csv
dataframe_csv.select("job", "marital", "deposit", "balance").show(10)
```

```
+----------+--------+-------+-------+
|       job| marital|deposit|balance|
+----------+--------+-------+-------+
|    admin.| married|    yes|   2343|
|    admin.| married|    yes|     45|
|technician| married|    yes|   1270|
|  services| married|    yes|   2476|
|    admin.| married|    yes|    184|
|management|  single|    yes|      0|
|management| married|    yes|    830|
|   retired|divorced|    yes|    545|
|technician| married|    yes|      1|
|  services|  single|    yes|   5090|
+----------+--------+-------+-------+
only showing top 10 rows
```

[87]:

```
dataframe_csv.select("marital",when(dataframe_csv.marital == 'married',
1).otherwise(0)).show(10)
```

```
+--------+-----------------------------------------------+
| marital|CASE WHEN (marital = married) THEN 1 ELSE 0 END|
+--------+-----------------------------------------------+
| married|                                              1|
| married|                                              1|
| married|                                              1|
| married|                                              1|
| married|                                              1|
|  single|                                              0|
| married|                                              1|
|divorced|                                              0|
| married|                                              1|
|  single|                                              0|
+--------+-----------------------------------------------+
only showing top 10 rows
```

[82]:

```
dataframe_csv[dataframe_csv.job.isin("admin.",
"management")].show()
```

```
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+-------+-------+
|age|       job|marital|education|default|balance|housing|loan|contact|day|month
|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+-------+-------+
| 59|    admin.|married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.|married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 54|    admin.|married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
| 42|management| single| tertiary|     no|      0|    yes| yes|unknown|  5|
may|     562|       2|   -1|       0| unknown|    yes|
| 56|management|married| tertiary|     no|    830|    yes| yes|unknown|  6|
may|    1201|       1|   -1|       0| unknown|    yes|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+-------+-------+
only showing top 5 rows
```

[89]:

```
#Show Distinct Values
dataframe_csv.select("contact").distinct().show()
```

```
+---------+
|  contact|
+---------+
| cellular|
|  unknown|
|telephone|
+---------+
```

[130]:

```
#Show DISTINCT Values
dataframe_csv.select("poutcome").distinct().show()
```

```
+--------+
|poutcome|
+--------+
| failure|
|   other|
| success|
| unknown|
+--------+
```

[101]:

```
#LIKE Operator
dataframe_csv.select("age", "marital","contact",
dataframe_csv.contact.like("u%")).show(10)
```

```
+---+--------+-------+---------------+
|age| marital|contact|contact LIKE u%|
+---+--------+-------+---------------+
| 59| married|unknown|           true|
| 56| married|unknown|           true|
| 41| married|unknown|           true|
| 55| married|unknown|           true|
| 54| married|unknown|           true|
| 42|  single|unknown|           true|
| 56| married|unknown|           true|
| 60|divorced|unknown|           true|
| 37| married|unknown|           true|
| 28|  single|unknown|           true|
+---+--------+-------+---------------+
only showing top 10 rows
```

[134]:

```
#STARTSWITH
df3 = dataframe_csv.select("education", "age", dataframe_csv.poutcome.
 ↪startswith("su"))
df3.show()
```

```
+---------+---+-----------------------+
|education|age|startswith(poutcome, su)|
+---------+---+-----------------------+
|secondary| 59|                  false|
|secondary| 56|                  false|
|secondary| 41|                  false|
|secondary| 55|                  false|
| tertiary| 54|                  false|
| tertiary| 42|                  false|
| tertiary| 56|                  false|
|secondary| 60|                  false|
|secondary| 37|                  false|
|secondary| 28|                  false|
|secondary| 38|                  false|
|secondary| 30|                  false|
| tertiary| 29|                  false|
| tertiary| 46|                  false|
| tertiary| 31|                  false|
| tertiary| 35|                  false|
|  primary| 32|                  false|
|secondary| 49|                  false|
|secondary| 41|                   false|
|secondary| 49|                   false|
+---------+---+------------------------+
only showing top 20 rows
```

[135]:

```
df3.printSchema()
```

```
root
 |-- education: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- startswith(poutcome, su): boolean (nullable = true)
```

[136]:

```
df3_filter = df3.where(df3['startswith(poutcome, su)'] == 'true')
df3_filter.show()
```

```
+---------+---+------------------------+
|education|age|startswith(poutcome, su)|
+---------+---+------------------------+
|secondary| 56|                    true|
| tertiary| 53|                    true|
|secondary| 46|                    true|
| tertiary| 40|                    true|
| tertiary| 31|                    true|
| tertiary| 31|                    true|
| tertiary| 33|                    true|
|secondary| 40|                    true|
| tertiary| 30|                    true|
| tertiary| 44|                    true|
|secondary| 37|                    true|
| tertiary| 52|                    true|
| tertiary| 32|                    true|
|secondary| 53|                    true|
|  primary| 50|                    true|
|secondary| 29|                    true|
|secondary| 38|                    true|
| tertiary| 36|                    true|
|secondary| 47|                    true|
|secondary| 36|                    true|
+---------+---+------------------------+
only showing top 20 rows
```

[129]:

```
##ENDSWITH
df4 = dataframe_csv.select("education", "age", dataframe_csv.poutcome.
 ↪endswith("ure"))
```

```python
df4.show()
df4_filter = df4.where(df4['endswith(poutcome, ure)'] == 'true')
df4_filter.show()
df4_filter.count()
```

```
+---------+---+-----------------------+
|education|age|endswith(poutcome, ure)|
+---------+---+-----------------------+
|secondary| 59|                  false|
|secondary| 56|                  false|
|secondary| 41|                  false|
|secondary| 55|                  false|
| tertiary| 54|                  false|
| tertiary| 42|                  false|
| tertiary| 56|                  false|
|secondary| 60|                  false|
|secondary| 37|                  false|
|secondary| 28|                  false|
|secondary| 38|                  false|
|secondary| 30|                  false|
| tertiary| 29|                  false|
| tertiary| 46|                  false|
| tertiary| 31|                  false|
| tertiary| 35|                  false|
|  primary| 32|                  false|
|secondary| 49|                  false|
|secondary| 41|                  false|
|secondary| 49|                  false|
+---------+---+-----------------------+
only showing top 20 rows

+---------+---+-----------------------+
|education|age|endswith(poutcome, ure)|
+---------+---+-----------------------+
|secondary| 33|                   true|
| tertiary| 34|                   true|
|secondary| 37|                   true|
|secondary| 45|                   true|
| tertiary| 32|                   true|
|secondary| 30|                   true|
| tertiary| 46|                   true|
| tertiary| 38|                   true|
|secondary| 32|                   true|
|secondary| 31|                   true|
|  primary| 50|                   true|
|secondary| 47|                   true|
| tertiary| 59|                   true|
|secondary| 31|                    true|
|secondary| 53|                    true|
|secondary| 31|                    true|
| tertiary| 40|                    true|
|secondary| 44|                    true|
|secondary| 43|                    true|
|secondary| 54|                    true|
+---------+---+------------------------+
only showing top 20 rows
```

[129]: 1228

[145]:

```
#UPDATING Columns
dataframe_csv.show(5)
dataframe_updatecol = dataframe_csv.withColumnRenamed('education', 'Education␣
 ↪Level')
dataframe_updatecol.show(5)
```

```
+---+----------+-------+---------------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+-------+
|age|       job|marital|Education Level|default|balance|housing|loan|contact|day
|month|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+-------+
| 59|    admin.|married|      secondary|     no|   2343|    yes|  no|unknown|
5|  may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.|married|      secondary|     no|     45|     no|  no|unknown|
5|  may|    1467|       1|   -1|       0| unknown|    yes|
| 41|technician|married|      secondary|     no|   1270|    yes|  no|unknown|
5|  may|    1389|       1|   -1|       0| unknown|    yes|
| 55|  services|married|      secondary|     no|   2476|    yes|  no|unknown|
5|  may|     579|       1|   -1|       0| unknown|    yes|
| 54|    admin.|married|       tertiary|     no|    184|     no|  no|unknown|
5|  may|     673|       2|   -1|       0| unknown|    yes|
+---+----------+-------+---------------+-------+-------+-------+----+-------+---
+-----+--------+--------+-----+--------+--------+-------+
only showing top 5 rows

+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
|age|       job|marital|education|default|balance|housing|loan|contact|day|month
|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
| 59|    admin.|married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.|married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41|technician|married|secondary|      no|   1270|     yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
| 55|  services|married|secondary|      no|   2476|     yes|  no|unknown|  5|
may|     579|       1|   -1|       0| unknown|    yes|
| 54|    admin.|married| tertiary|      no|    184|      no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
only showing top 5 rows
```

[150]:

```
#Removing Columns
dataframe_csv_remove = dataframe_updatecol.drop("month", "previous")
print(len(dataframe_csv.columns))
print(len(dataframe_csv_remove.columns))
```

```
17
15
```

[152]:

```
# Displays the content of dataframe
dataframe_csv.show()
# Return first n rows
dataframe_csv.head()
# Return first n rows
dataframe_csv.take(5)
# Computes summary statistics
dataframe_csv.describe().show()
# Counts the number of rows in dataframe
dataframe_csv.count()
# Counts the number of distinct rows in dataframe
dataframe_csv.distinct().count()
```

```
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
|age|        job| marital|education|default|balance|housing|loan|contact|day|mon
th|duration|campaign|pdays|previous|poutcome|deposit|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
| 59|     admin.| married|secondary|      no|   2343|     yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|     admin.| married|secondary|      no|     45|      no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41| technician| married|secondary|      no|   1270|     yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
| 55|   services| married|secondary|      no|   2476|     yes|  no|unknown|  5|
may|     579|       1|   -1|       0| unknown|    yes|
| 54|     admin.| married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
| 42| management|  single| tertiary|     no|      0|    yes| yes|unknown|  5|
may|     562|       2|   -1|       0| unknown|    yes|
| 56| management| married| tertiary|     no|    830|    yes| yes|unknown|  6|
may|    1201|       1|   -1|       0| unknown|    yes|
| 60|    retired|divorced|secondary|     no|    545|    yes|  no|unknown|  6|
may|    1030|       1|   -1|       0| unknown|    yes|
| 37| technician| married|secondary|     no|      1|    yes|  no|unknown|  6|
may|     608|       1|   -1|       0| unknown|    yes|
| 28|   services|  single|secondary|     no|   5090|    yes|  no|unknown|  6|
may|    1297|       3|   -1|       0| unknown|    yes|
| 38|     admin.|  single|secondary|     no|    100|    yes|  no|unknown|  7|
may|     786|       1|   -1|       0| unknown|    yes|
| 30|blue-collar| married|secondary|     no|    309|    yes|  no|unknown|  7|
may|    1574|       2|   -1|       0| unknown|    yes|
| 29| management| married| tertiary|     no|    199|    yes| yes|unknown|  7|
may|    1689|       4|   -1|       0| unknown|    yes|
| 46|blue-collar|  single| tertiary|     no|    460|    yes|  no|unknown|  7|
may|    1102|       2|   -1|       0| unknown|    yes|
| 31| technician|  single| tertiary|     no|    703|    yes|  no|unknown|  8|
may|     943|       2|   -1|       0| unknown|    yes|
| 35| management|divorced| tertiary|     no|   3837|    yes|  no|unknown|  8|
may|    1084|       1|   -1|       0| unknown|    yes|
| 32|blue-collar|  single|  primary|     no|    611|    yes|  no|unknown|  8|
may|     541|       3|   -1|       0| unknown|    yes|
| 49|   services| married|secondary|     no|     -8|    yes|  no|unknown|  8|
may|    1119|       1|   -1|       0| unknown|    yes|
| 41|     admin.| married|secondary|     no|     55|    yes|  no|unknown|  8|
may|    1120|       2|   -1|       0| unknown|    yes|
| 49|     admin.|divorced|secondary|     no|    168|    yes| yes|unknown|  8|
may|     513|       1|   -1|       0| unknown|    yes|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+---
--+--------+--------+-----+--------+--------+-------+
only showing top 20 rows

+-------+------------------+-------+--------+---------+-------+-----------------
-+-------+-----+--------+------------------+-----+------------------+-----------
-------+------------------+------------------+--------+-------+
|summary|               age|    job| marital|education|default|
balance|housing| loan| contact|               day|month|          duration|
campaign|             pdays|          previous|poutcome|deposit|
+-------+------------------+-------+--------+---------+-------+-----------------
-+-------+-----+--------+------------------+-----+------------------+-----------
-------+------------------+------------------+--------+-------+
|  count|             11162|  11162|   11162|    11162|  11162|
11162|  11162|11162|   11162|             11162|11162|             11162|
11162|             11162|             11162|   11162|  11162|
|   mean|41.231947679627304|   null|    null|     null|
null|1528.5385235620856|   null| null|    null|15.658036194230425|
null|371.99381831213043| 2.508421429851281|
51.33040673714388|0.8325568894463358|    null|   null|
| stddev|11.913369192215518|   null|    null|     null|   null|
3225.413325946149|   null| null|    null| 8.420739541006462|
null|347.12838571630687|2.7220771816614824|108.75828197197717|
2.292007218670508|    null|   null|
|    min|                18| admin.|divorced|  primary|     no|
-6847|     no|   no|cellular|                 1|  apr|                 2|
1|                -1|                 0| failure|     no|
|    max|                95|unknown|  single|  unknown|    yes|
81204|    yes|  yes| unknown|                31|  sep|              3881|
63|               854|                58| unknown|    yes|
+-------+------------------+-------+--------+---------+-------+-----------------
-+-------+-----+--------+------------------+-----+------------------+-----------
-------+------------------+------------------+--------+-------+
```

[152]:

```
11162
```

[154]:

```
# Returns columns of dataframe
dataframe_csv.columns
```

[154]:

```
['age',
 'job',
 'marital',
 'education',
 'default',
 'balance',
 'housing',
 'loan',
 'contact',
 'day',
 'month',
 'duration',
 'campaign',
 'pdays',
 'previous',
 'poutcome',
 'deposit']
```

[157]:

```
# Returns first row
dataframe_csv.first()
```

[157]:

```
Row(age=59, job='admin.', marital='married', education='secondary',
default='no', balance=2343, housing='yes', loan='no', contact='unknown', day=5,
month='may', duration=1042, campaign=1, pdays=-1, previous=0,
poutcome='unknown', deposit='yes')
```

[158]:

```python
# Returns dataframe column names and data types
dataframe_csv.dtypes
```

[158]:

```
[('age', 'int'),
 ('job', 'string'),
 ('marital', 'string'),
 ('education', 'string'),
 ('default', 'string'),
 ('balance', 'int'),
 ('housing', 'string'),
 ('loan', 'string'),
 ('contact', 'string'),
 ('day', 'int'),
 ('month', 'string'),
 ('duration', 'int'),
 ('campaign', 'int'),
 ('pdays', 'int'),
 ('previous', 'int'),
 ('poutcome', 'string'),
 ('deposit', 'string')]
```

[160]:

```python
#GROUP BY
dataframe_csv.groupBy("education").count().show(10)
```

```
+---------+-----+
|education|count|
+---------+-----+
| tertiary| 3689|
|  primary| 1500|
|  unknown|  497|
|secondary| 5476|
+---------+-----+
```

[163]:

```python
#SPARK SQL
dataframe_csv.registerTempTable("dataframe_csv")
sc.sql("select * from dataframe_csv").show(3)
```

```
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
|age|       job|marital|education|default|balance|housing|loan|contact|day|month
|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
| 59|    admin.|married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.|married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41|technician|married|secondary|     no|   1270|    yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
only showing top 3 rows
```

[165]:

```
sc.sql("select marital from dataframe_csv").show(3)
```

```
+-------+
|marital|
+-------+
|married|
|married|
|married|
+-------+
only showing top 3 rows
```

[170]:

```
sc.sql("select distinct education from dataframe_csv").show()
sc.sql("select distinct education from dataframe_csv").count()
```

```
+---------+
|education|
+---------+
| tertiary|
|  primary|
|  unknown|
|secondary|
+---------+
```

[170]: 4

[196]:

```
sc.sql("select * from dataframe_csv where education =='tertiary'").show(3)
```

```
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
|age|       job|marital|education|default|balance|housing|loan|contact|day|month
|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
| 54|    admin.|married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
| 42|management| single| tertiary|     no|      0|    yes| yes|unknown|  5|
may|     562|       2|   -1|       0| unknown|    yes|
| 56|management|married| tertiary|     no|    830|    yes| yes|unknown|  6|
may|    1201|       1|   -1|       0| unknown|    yes|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
only showing top 3 rows
```

[197]:

```
sc.sql("select * from dataframe_csv order by balance desc").show(3)
```

```
+---+-----------+-------+---------+-------+-------+-------+----+---------+---+--
---+--------+--------+-----+--------+--------+-------+
|age|        job|marital|education|default|balance|housing|loan|
contact|day|month|duration|campaign|pdays|previous|poutcome|deposit|
+---+-----------+-------+---------+-------+-------+-------+----+---------+---+--
---+--------+--------+-----+--------+--------+-------+
| 84|    retired|married|secondary|     no|  81204|     no|  no|telephone| 28|
dec|     679|       1|  313|       2|   other|    yes|
| 84|    retired|married|secondary|     no|  81204|     no|  no|telephone|  1|
apr|     390|       1|   94|       3| success|    yes|
| 52|blue-collar|married|  primary|     no|  66653|     no|  no| cellular| 14|
aug|     109|       3|   -1|       0| unknown|     no|
+---+-----------+-------+---------+-------+-------+-------+----+---------+---+--
---+--------+--------+-----+--------+--------+-------+
only showing top 3 rows
```

[208]:

```
sc.sql("select job, sum(balance) from dataframe_csv group by job").show()
```

```
+-------------+------------+
|          job|sum(balance)|
+-------------+------------+
|      student|      540282|
|self-employed|      755476|
|     services|      997921|
|    housemaid|      374328|
|   management|     4602541|
|  blue-collar|     2340433|
|      retired|     1880621|
| entrepreneur|      531997|
|       admin.|     1595286|
|   technician|     2837125|
|   unemployed|      469355|
|      unknown|      136182|
+-------------+------------+
```

[217]:

```
sc.sql("select max(balance) AS MaximumBalance from dataframe_csv").show()
sc.sql("select min(balance) AS Minimum_Balance from dataframe_csv").show()
```

```
+--------------+
|MaximumBalance|
+--------------+
|         81204|
+--------------+

+---------------+
|Minimum_Balance|
+---------------+
|          -6847|
+---------------+
```

[218]:

```
sc.sql("select * from dataframe_csv where marital like 'd%'").show(3)
```

```
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
-+--------+--------+-----+--------+--------+-------+
|age|       job| marital|education|default|balance|housing|loan|contact|day|mont
h|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
-+--------+--------+-----+--------+--------+-------+
| 60|   retired|divorced|secondary|     no|    545|    yes|  no|unknown|  6|
may|    1030|       1|   -1|       0| unknown|    yes|
| 35|management|divorced| tertiary|     no|   3837|    yes|  no|unknown|  8|
may|    1084|       1|   -1|       0| unknown|    yes|
| 49|    admin.|divorced|secondary|     no|    168|    yes| yes|unknown|  8|
may|     513|       1|   -1|       0| unknown|    yes|
+---+----------+--------+---------+-------+-------+-------+----+-------+---+----
-+--------+--------+-----+--------+--------+-------+
only showing top 3 rows
```

[220]:

```
sc.sql("select balance, job, age from dataframe_csv where balance between 100␣
 ↪and 3000 order by age desc").show()
```

```
+-------+-------+---+
|balance|    job|age|
+-------+-------+---+
|   2282|retired| 95|
|    775|retired| 93|
|    775|retired| 93|
|    775|retired| 92|
|    775|retired| 92|
|    712|retired| 90|
|    553|retired| 89|
|    648|retired| 88|
|    433|retired| 88|
|    433|retired| 87|
|   2190|retired| 87|
|    230|retired| 87|
|   1255|retired| 86|
|    157|retired| 86|
|    614|retired| 86|
|   1255|retired| 85|
|   1934|retired| 85|
|    639|retired| 84|
|   1965|retired| 83|
|    425|retired| 83|
+-------+-------+---+
only showing top 20 rows
```

[236]:
```
sc.sql("SELECT * FROM dataframe_csv").show(5)
```

```
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
|age|       job|marital|education|default|balance|housing|loan|contact|day|month
|duration|campaign|pdays|previous|poutcome|deposit|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
| 59|    admin.|married|secondary|     no|   2343|    yes|  no|unknown|  5|
may|    1042|       1|   -1|       0| unknown|    yes|
| 56|    admin.|married|secondary|     no|     45|     no|  no|unknown|  5|
may|    1467|       1|   -1|       0| unknown|    yes|
| 41|technician|married|secondary|     no|   1270|    yes|  no|unknown|  5|
may|    1389|       1|   -1|       0| unknown|    yes|
| 55|  services|married|secondary|     no|   2476|    yes|  no|unknown|  5|
may|     579|       1|   -1|       0| unknown|    yes|
| 54|    admin.|married| tertiary|     no|    184|     no|  no|unknown|  5|
may|     673|       2|   -1|       0| unknown|    yes|
+---+----------+-------+---------+-------+-------+-------+----+-------+---+-----
+--------+--------+-----+--------+--------+-------+
only showing top 5 rows
```

[238]:
```
sc.sql("select count(*) from dataframe_csv where duration >200").show()
```

```
+--------+
|count(1)|
+--------+
|    6768|
+--------+
```

[241]:

```
sc.sql("select balance, duration, poutcome from dataframe_csv where marital␣
 ↪=='married' and poutcome=='unknown' and balance > 0 order by balance").show()
```

```
+-------+--------+--------+
|balance|duration|poutcome|
+-------+--------+--------+
|      1|     608| unknown|
|      1|      55| unknown|
|      1|      85| unknown|
|      1|     248| unknown|
|      1|     215| unknown|
|      1|     173| unknown|
|      1|     167| unknown|
|      1|     395| unknown|
|      1|     102| unknown|
|      1|     528| unknown|
|      1|     210| unknown|
|      1|     102| unknown|
|      1|     535| unknown|
|      1|      77| unknown|
|      1|     506| unknown|
|      2|     182| unknown|
|      2|     147| unknown|
|      2|     194| unknown|
|      2|     703| unknown|
|      2|     450| unknown|
+-------+--------+--------+
only showing top 20 rows
```

[ ]:

```
!wget -nc https://raw.githubusercontent.com/brpy/colab-pdf/master/colab_pdf.py
from colab_pdf import colab_pdf
colab_pdf('SparkSQL PySpark-EDA.ipynb')
```

```
File colab_pdf.py already there; not retrieving.


WARNING: apt does not have a stable CLI interface. Use with caution in scripts.
```