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()
    job| marital|education|default|balance|housing|loan|contact|day|mon
    age
    th|duration|campaign|pdays|previous|poutcome|deposit|
    --+-----+
    | 59|
             admin. | married | secondary |
                                                2343
                                                         yes | no | unknown |
                                                                           5|
    may|
           1042
                                     0 | unknown |
                                                   yes
    | 56|
             admin. | married | secondary |
                                                  45|
                                                               no | unknown |
                                                          nol
                                                                           51
    may|
           1467
                       1|
                           -1|
                                     0 | unknown |
                                                   yes|
    | 41| technician| married|secondary|
                                                              no | unknown |
                                                                           51
                                          nol
                                                1270
                                                         yes|
                       1 |
                                                   yesl
    may |
           1389
                           -1|
                                     0 | unknown |
    | 55|
           services | married | secondary |
                                           no|
                                                2476|
                                                         yes|
                                                               no | unknown |
                                                                           5|
    may|
            579
                       1|
                           -1|
                                     0 | unknown |
                                                   yes
    | 54|
             admin. | married | tertiary |
                                           no|
                                                 184|
                                                             no|unknown|
                                                                           5 l
    may |
            673
                       21
                           -1|
                                     0 | unknown |
                                                   yes
    | 42| management| single| tertiary|
                                           nol
                                                   0|
                                                         yes | yes | unknown |
                                                                           5 l
            562 l
                       2|
                           -1 l
                                     0| unknown|
    may|
                                                   yesl
    | 56| management| married| tertiary|
                                                 830|
                                           nol
                                                         yes | yes | unknown |
                                                                           61
    may|
           1201
                       1|
                           -1|
                                     0 | unknown |
                                                   yes|
```

```
may|
             1030|
                         1|
                              -1|
                                        0 | unknown |
                                                        yes|
    | 37| technician| married|secondary|
                                                                    no | unknown |
                                                                                 6 l
                                              nol
                                                        1|
                                                              ves
             608
                         1|
                              -1|
    may|
                                        0 | unknown |
                                                        yes|
    1 281
            services | single | secondary |
                                              nol
                                                     5090 l
                                                              yes
                                                                    no | unknown |
                                                                                 6 I
                         3|
                              -1|
    may|
            1297 l
                                        0 | unknown |
                                                        yes|
    | 38|
              admin.
                       single|secondary|
                                                      100|
                                                                    no | unknown |
                                                                                 7 |
                                                              yes
    mavl
             786 l
                         1 l
                              -1 l
                                        0 | unknown |
                                                        yes|
    | 30|blue-collar| married|secondary|
                                                      309|
                                                              yes | no | unknown |
                                                                                 71
    may|
            1574
                         2|
                              -1|
                                        0 | unknown |
                                                        yes|
    | 29| management| married| tertiary|
                                                              yes | yes | unknown |
                                                                                 71
                                              nol
                                                      199|
            1689|
                         4|
                              -1|
                                        0 | unknown |
                                                        yes|
    | 46|blue-collar|
                        single | tertiary |
                                              no|
                                                                    no|unknown|
                                                                                 71
                                                      460|
                                                              yes|
    mavl
            1102
                         2|
                              -1 l
                                        0 | unknown |
                                                        yes
    | 31| technician|
                        single | tertiary |
                                                      703|
                                                              yes|
                                                                    no | unknown |
                         2|
    mavl
             943|
                              -1|
                                        0 | unknown |
                                                        yes|
    | 35| management|divorced| tertiary|
                                              nol
                                                     3837|
                                                              yes|
                                                                    no | unknown |
                                                                                 81
            1084
                         1|
                              -1|
    mavl
                                        0 | unknown |
                                                        yesl
    | 32|blue-collar| single| primary|
                                              no|
                                                      611|
                                                              yes|
                                                                    no|unknown|
                                                                                 81
    may|
             541 l
                         31
                              -1 l
                                        0 | unknown |
                                                        yes|
    | 49|
            services | married | secondary |
                                                       -8|
                                                              yes
                                                                    no | unknown |
                                                                                 81
    may|
            1119|
                         1 |
                              -1 l
                                        0 | unknown |
                                                        yes
    | 41|
              admin. | married | secondary |
                                                       55|
                                                              yes|
                                                                   no | unknown |
                                                                                 81
                                        0 | unknown |
    mayl
                         21
                              -1|
                                                        yes|
    | 49|
              admin. | divorced | secondary |
                                              nol
                                                      168
                                                              yes | yes | unknown |
                         1|
                              -1|
             513|
                                        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)
```

545 l

nol

ves| no|unknown|

6 l

I 60 I

retired|divorced|secondary|

```
-+----+
          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 |
may|
                  1|
                      -1|
                               0 | unknown |
                                            yes
        admin. | married | secondary |
l 561
                                   nol
                                           45 l
                                                  nol
                                                      no | unknown |
                                                                  5 I
                                            yes|
may|
       1467
                 1|
                      -1 l
                               0 | unknown |
| 41|technician| married|secondary|
                                   no|
                                         1270
                                                 yes|
                                                      no | unknown |
                                                                  5 I
       1389|
                               0 | unknown |
may|
                  1 |
                      -1|
                                            yes|
| 55|
      services | married | secondary |
                                   nol
                                         2476
                                                 yes|
                                                      no | unknown |
                                                                  5 I
                  1|
may|
        579
                      -1|
                               0 | unknown |
                                            yes|
| 54|
        admin. | married | tertiary |
                                   nol
                                          184
                                                      no | unknown |
may |
        673|
                  21
                      -1|
                               0 | unknown |
                                            yesl
| 42|management| single| tertiary|
                                                 yes| yes|unknown|
                                   no|
                                            01
                                                                  5 I
        5621
                 2|
                      -1|
                               0 | unknown |
                                            yesl
may |
| 56|management| married| tertiary|
                                          830|
                                                 yes | yes | unknown |
                                                                  61
                                   nol
       1201
                 1|
                      -1|
may |
                               0 | unknown |
                                            yesl
| 60|
       retired | divorced | secondary |
                                                                  6 I
                                   nol
                                          545
                                                 yes | no | unknown |
                  1|
                      -1 l
                               0 | unknown |
                                            yes|
| 37|technician| married|secondary|
                                   nol
                                            1|
                                                      no | unknown |
                                                 yes|
may|
        608 l
                 11
                      -1 l
                               0 | unknown |
                                            yesl
| 28|
               single|secondary|
                                         5090|
                                                 yesl
      services|
                                   no|
                                                      no|unknown|
       1297 l
                 31
                      -1 l
                               0 | unknown |
may|
                                            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|

```
[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 |
                                  45|
                          yesl
    |technician| married|
                          yesl
                                 1270|
      services | married |
                          yes|
                                 2476
        admin. | married |
                          yes|
                                 184
    |management| single|
                                   01
                          yes|
    |management| married|
                                 830|
                          yesl
       retired|divorced|
                                 545 l
                          yes|
    |technician| married|
                          yes|
                                   1|
      services | single |
                                 5090|
                          yes|
    +----+
   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 l
    | married|
                                                       1|
    | married|
                                                       1 |
    | married|
    | single|
                                                       01
    | married|
    |divorced|
                                                       0|
    | married|
                                                       11
    | single|
                                                       01
   only showing top 10 rows
[82]: dataframe_csv[dataframe_csv.job.isin("admin.",
    "management")].show()
```

```
+----+
            job|marital|education|default|balance|housing|loan|contact|day|month
    |duration|campaign|pdays|previous|poutcome|deposit|
   +----+
          admin. | married | secondary |
                                           yes| no|unknown| 5|
                                nol
   may|
         10421
                  1 l
                             0 | unknown |
                                        yesl
          admin. | married | secondary |
   | 56|
                                no|
                                      45|
                                            nol
                                               no|unknown|
                                                         5|
                      -1|
                                        yesl
   may|
         1467
                  1|
                             0 | unknown |
    | 54|
          admin. | married | tertiary |
                                      184|
                                                         5 l
                                no|
                                            no| no|unknown|
          673|
   may|
                  2|
                      -1|
                             0 | unknown |
                                        yes|
    | 42|management| single| tertiary|
                                       0|
                                no|
                                           yes| yes|unknown|
                                        yesl
          562
                  2|
                      -1|
                             0 | unknown |
   mavl
    | 56|management|married| tertiary|
                                nol
                                      830|
                                            yes| yes|unknown|
         1201
                             0 | unknown |
   mavl
                  1|
                      -1 l
                                        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|
                          truel
| 54| married|unknown|
                          truel
| 42| single|unknown|
                          true
| 56| married|unknown|
                          truel
| 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|
                                 falsel
|secondary| 56|
                                 falsel
|secondary| 41|
                                 falsel
|secondary| 55|
                                 false
| tertiary| 54|
                                 false
| tertiary| 42|
                                 falsel
| tertiary| 56|
                                 false
|secondary| 60|
                                 falsel
|secondary| 37|
                                 false
|secondary| 28|
                                 false
|secondary| 38|
                                 false
|secondary| 30|
                                 false
| tertiary| 29|
                                 false
| tertiary| 46|
                                 false
| tertiary| 31|
                                 falsel
| tertiary| 35|
                                 falsel
| primary| 32|
                                 falsel
|secondary| 49|
                                 falsel
```

```
|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|
                                   truel
    | tertiary| 31|
                                   truel
    | tertiary| 31|
                                   truel
    | 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|
                                   truel
    |secondary| 36|
                                   true
    +----+
    only showing top 20 rows
[129]: ##ENDSWITH
     df4 = dataframe_csv.select("education", "age", dataframe_csv.poutcome.
```

falsel

|secondary| 41|

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

```
+----+
|education|age|endswith(poutcome, ure)|
|secondary| 59|
                                falsel
|secondary| 56|
                                false
|secondary| 41|
                                false
|secondary| 55|
                                false
| tertiary| 54|
                                falsel
| tertiary| 42|
                                false
| tertiary| 56|
                               falsel
|secondary| 60|
                                falsel
|secondary| 37|
                                falsel
|secondary| 28|
                                falsel
|secondary| 38|
                               false
|secondary| 30|
                                false
| tertiary | 29 |
                                false
                                false
| tertiary| 46|
| tertiary| 31|
                                false
| tertiary| 35|
                                falsel
| primary| 32|
                                false
|secondary| 49|
                                falsel
|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|
                                 truel
|secondary| 30|
                                 true
| tertiary| 46|
                                 truel
| tertiary| 38|
                                 true
|secondary| 32|
                                 truel
|secondary| 31|
                                 true
| primary | 50 |
                                 truel
|secondary| 47|
                                 truel
| tertiary| 59|
                                 true
```

```
|secondary| 31|
                          truel
   |secondary| 53|
                          truel
   |secondary| 31|
                          truel
   | tertiary| 40|
                          true
   |secondary| 44|
                          truel
   |secondary| 43|
                          truel
   |secondary| 54|
   +----+
   only showing top 20 rows
[129]: 1228
[145]: #UPDATING Columns
    dataframe_csv.show(5)
    dataframe updatecol = dataframe csv.withColumnRenamed('education', 'Education',
    dataframe_updatecol.show(5)
   +----+
           job|marital|Education Level|default|balance|housing|loan|contact|day
   lagel
   |month|duration|campaign|pdays|previous|poutcome|deposit|
   +----+----+----+
                                      2343|
   | 59|
         admin. | married |
                        secondary|
                                  no|
                                            yes | no | unknown |
   5| mav|
           1042
                              0 | unknown |
                    1|
                       -1|
                                        yes
   | 56|
         admin. | married |
                        secondary
                                  nol
                                       45|
                                             no | no | unknown |
           1467 l
                              0 | unknown |
   5| may|
                   11
                       -1|
                                        yes|
                                      1270|
                                            yes| no|unknown|
   | 41|technician|married|
                       secondary
                                  no|
                                        yes|
   5| may|
           1389 l
                   1|
                       -1|
                              0 | unknown |
   | 55| services|married|
                                      2476
                       secondary
                                  no|
                                            yes| no|unknown|
   5| may|
            579 l
                       -1 l
                              0 | unknown |
                   1 |
                                        yes
   | 54|
         admin. | married |
                        tertiary
                                  no|
                                      184|
                                             no | no | unknown |
            673
                   2|
                              0 | unknown |
                       -1|
                                        yes|
   +----+
   only showing top 5 rows
   +----+
           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|
                    -1 l
   may|
         1042
                 11
                           0 | unknown |
                                     yes|
```

no|

l 561

admin. | married | secondary |

45 l

no| no|unknown| 5|

```
-1|
           1467
                                 0 | unknown |
    mavl
                    1|
                                             yesl
    | 41|technician|married|secondary|
                                    nol
                                         1270|
                                                yes| no|unknown| 5|
                         -1|
                                 0 | unknown |
                                             yes|
    may|
           1389
                    1|
          services|married|secondary|
                                    no|
    | 55|
                                         2476
                                                yes| no|unknown|
                                                                5|
    may|
           579 l
                    11
                         -1 l
                                 0 | unknown |
                                             yes|
    | 54|
           admin. | married | tertiary |
                                          184|
                                    no|
                                                 no | no | unknown | 5 |
                    2|
                                 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()
    __+____
               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 |
                                 0 | unknown |
    may|
                    1|
                         -1|
           1042|
                                             yesl
    | 56|
            admin. | married | secondary |
                                             45|
                                      no|
                                                   no| no|unknown| 5|
    mavl
                    1|
                         -1|
                                 0 | unknown |
                                             yesl
    | 41| technician| married|secondary|
                                      no|
                                           1270
                                                  yesl
                                                       no|unknown| 5|
    may|
           1389|
                    11
                         -1 l
                                 0 | unknown |
                                             yesl
    | 55|
           services | married | secondary |
                                           2476
                                                  yes| no|unknown| 5|
                                      nol
    may|
           579 l
                    1 |
                       -1|
                             0| unknown|
                                             yes|
```

```
l 541
         admin. | married | tertiary |
                                                   no | no | unknown |
                                     nol
                                           184 l
                                                                   5 l
may|
        673|
                  2|
                      -1|
                               0 | unknown |
                                             yes|
| 42| management| single| tertiary|
                                                  yes | yes | unknown |
                                                                   5 l
                                     nol
                                             0|
        562|
                  2|
                      -1|
                               0 | unknown |
may|
                                             yes|
| 56 | management | married | tertiary |
                                           830 l
                                                  yes | yes | unknown |
                                                                   6 l
       1201 l
may|
                  1|
                      -1 l
                               0 | unknown |
                                             yes|
| 60|
        retired|divorced|secondary|
                                           545|
                                                  yes | no | unknown |
                                                                    6|
may|
       1030 l
                  1 l
                      -1 l
                               0 | unknown |
                                             yes|
| 37| technician| married|secondary|
                                             1|
                                                        no | unknown |
                                                                    61
                                                  yes|
may|
        608 l
                  1|
                      -1|
                               0 | unknown |
                                             yes|
| 28|
       services|
                 single|secondary|
                                     no|
                                          5090|
                                                       no|unknown|
                                                                    61
                                                  yesl
may |
       1297
                  3|
                      -1|
                               0 | unknown |
                                             yes|
| 38|
         admin.
                 single|secondary|
                                           100|
                                                  yes|
                                                        no | unknown |
                                                                   7 |
        786 l
                  1 |
                      -1 l
                               0 | unknown |
mavl
                                             yes
| 30|blue-collar| married|secondary|
                                           3091
                                                  yes | no | unknown |
                                                                   71
may |
       1574
                  21
                      -1 l
                               0 | unknown |
                                             yesl
| 29| management| married| tertiary|
                                           199|
                                                  yes | yes | unknown |
                                                                   7|
                                     no|
       1689|
may|
                  4|
                      -1|
                               0 | unknown |
                                             yes|
| 46|blue-collar| single| tertiary|
                                     no|
                                           460|
                                                        no|unknown|
                                                                   7 |
                                                  yes
       1102|
                  21
                      -1 l
may|
                               0 | unknown |
                                             yes|
| 31| technician|
                 single | tertiary |
                                           703|
                                                  yes
                                                        no | unknown |
                                                                   81
        943 l
                  2|
                      -1 l
                               0 | unknown |
may|
                                             yes|
| 35| management|divorced| tertiary|
                                          3837 l
                                                  yes
                                                        no | unknown |
                                                                    81
may|
       1084
                  1|
                      -1|
                               0 | unknown |
                                             yesl
| 32|blue-collar| single| primary|
                                                                   81
                                     nol
                                           611
                                                  yes|
                                                       no | unknown |
        541
                  3|
                      -1|
may |
                               0 | unknown |
                                             yes
| 49|
       services | married | secondary |
                                     no|
                                            -8|
                                                  yes|
                                                        no | unknown |
                                                                   81
may|
       1119|
                  1|
                      -1|
                               0 | unknown |
                                             yes|
| 41|
         admin. | married | secondary |
                                     nol
                                            55 l
                                                  ves | no | unknown |
                  2|
                      -1|
                               0 | unknown |
may|
                                             yes|
| 49|
         admin. | divorced | secondary |
                                     nol
                                           168 l
                                                  yes | yes | unknown |
        513|
                  1 l
                      -1 l
                               0 | unknown |
mavl
                                             yes|
-----
only showing top 20 rows
_+_____
______
|summary|
                             job | marital | education | default |
                     age|
balance|housing| loan| contact|
                                         day | month |
                                                           duration|
                   pdays|
                                  previous|poutcome|deposit|
campaign|
_____+
                    11162 | 11162 |
                                   11162
                                            11162 | 11162 |
11162 | 11162 | 11162 |
                    11162
                                      11162 | 11162 |
                                                            11162
11162
                 111621
                                  11162|
                                          11162 | 11162 |
```

```
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|
     2.2920072186705081
                            null
                                     null
                               18 | admin. | divorced | primary |
          min
     -6847 l
                 nol
                       no|cellular|
                                                      1|
                                                          apr|
                                                                                21
     1|
                                             0| failure|
                        -1|
                                                              nol
                                             single| unknown|
          max
                               95 | unknown |
     81204
                yes| yes| unknown|
                                                         sepl
                                                                             3881
                                                     31|
     63 l
                        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')
```

nulll

nulll

null

mean | 41.231947679627304 |

```
[158]: # 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]: #GROUP BY
     dataframe_csv.groupBy("education").count().show(10)
    +----+
    |education|count|
    +----+
    | tertiary| 3689|
    | primary| 1500|
     | unknown| 497|
    |secondary| 5476|
    +----+
[163]: #SPARK SQL
     dataframe_csv.registerTempTable("dataframe_csv")
     sc.sql("select * from dataframe_csv").show(3)
    +----+
               job|marital|education|default|balance|housing|loan|contact|day|month
    |duration|campaign|pdays|previous|poutcome|deposit|
    ----+----+
    | 59|
            admin. | married | secondary |
                                                    yes| no|unknown| 5|
                                      no|
                                            2343
    may|
           1042
                      1|
                          -1|
                                   0 | unknown |
                                                yes|
    | 56|
           admin. | married | secondary |
                                      no|
                                             45 l
                                                    no| no|unknown| 5|
                                   0| unknown|
    may|
           1467 l
                      1 l
                        -1|
                                                yes
```

```
yes| no|unknown| 5|
    | 41|technician|married|secondary|
                                nol
                                    1270|
   may|
         1389|
                  1|
                      -1|
                             0 | unknown |
                                        yesl
    ----+
   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)
   +----+
            job|marital|education|default|balance|housing|loan|contact|day|month
    |duration|campaign|pdays|previous|poutcome|deposit|
   +----+
          admin. | married | tertiary |
                                nol
                                     184|
                                            no | no | unknown | 5 |
                                        yesl
          673
                             0 | unknown |
   may|
                  2|
                      -1|
    | 42|management| single| tertiary|
                                nol
                                       0|
                                           yes | yes | unknown | 5 |
          562|
                  2|
                      -1 l
                             0 | unknown |
   may|
                                        yes|
    | 56|management|married| tertiary|
                                    830 | yes | yes | unknown |
                                no|
```

```
1 -1
                            0 | unknown |
                                       yes|
   +----+
   only showing top 3 rows
[197]: sc.sql("select * from dataframe_csv order by balance desc").show(3)
   ___+___
            job|marital|education|default|balance|housing|loan|
   contact|day|month|duration|campaign|pdays|previous|poutcome|deposit|
   +----+
   l 841
          retired | married | secondary |
                                no| 81204|
                                           no| no|telephone| 28|
   dec
          679|
                 1 | 313 |
                                other|
                                       yesl
   1841
          retired | married | secondary |
                                no| 81204|
                                           no| no|telephone| 1|
          390|
                     94|
                 1|
                            3| success|
                                       yes|
   | 52|blue-collar|married| primary|
                                no| 66653|
                                           no | no | cellular | 14 |
                     -1 l
                            0 | unknown |
   aug
                                       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
                  3743281
      management |
                 4602541
     blue-collar|
                 23404331
        retired
                 1880621
     entrepreneur|
                  531997|
         admin.
                 1595286
      technician|
                 2837125
      unemployed|
                  469355
        unknown
                  1361821
[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)
```

```
-----+
       job | marital | education | default | balance | housing | loan | contact | day | mont
h|duration|campaign|pdays|previous|poutcome|deposit|
-+----+
60
    retired | divorced | secondary |
                                yes| no|unknown| 6|
                       no|
                           545
may|
    1030
           1|
              -1|
                    0 | unknown |
                             yes|
| 35|management|divorced| tertiary|
                       nol
                           3837|
                                yes | no | unknown |
                             yesl
may|
    1084
           1|
              -1|
                    0 | unknown |
| 49|
     admin.|divorced|secondary|
                       nol
                           168
                                yes| yes|unknown|
may|
     513|
           1|
              -1|
                    0 | unknown |
                             yesl
_+_____
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| 92|
| 775|retired| 92|
| 775|retired| 92|
| 712|retired| 90|
| 553|retired| 89|
| 648|retired| 88|
| 433|retired| 88|
```

```
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)
    +----+
             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 |
                                          yesl
    | 56|
          admin. | married | secondary |
                                  no|
                                        45|
                                                  no | unknown |
                                                            5|
                                              no|
    may|
          1467 l
                   1 |
                       -1 l
                               0 | unknown |
                                          yes|
    | 41|technician|married|secondary|
                                  no|
                                      1270
                                             yes|
                                                  no | unknown |
    may|
          13891
                   11
                       -1 l
                               0 | unknown |
                                          yes
         services | married | secondary |
                                  nol
                                      24761
    | 55|
                                             yesl
                                                  no | unknown |
          579|
                   1|
                       -1 l
                               0 | unknown |
    may|
                                          yes|
                                       184|
    | 54|
          admin. | married | tertiary |
                                  nol
                                              no| no|unknown| 5|
                   2|
                       -1 l
                               0 | unknown |
                                          yesl
    +----+
    only showing top 5 rows
[238]: sc.sql("select count(*) from dataframe csv where duration >200").show()
    +----+
    |count(1)|
    +----+
       6768 l
```

433|retired| 87| 2190|retired| 87|

+----+

[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 |
                851 unknown1
       1 l
               248 | unknown |
       11
               215 | unknown |
       1 l
               173 | unknown |
       11
       11
               167 | unknown |
       1|
               395 | unknown |
               102 | unknown |
       1|
               528 | unknown |
       1|
               210 | unknown |
               102 | unknown |
       1 |
               535 | unknown |
       1|
       1|
               77 | unknown |
               506 | unknown |
       1|
       2|
               182 | unknown |
       2|
               147 | unknown |
               1941 unknown
               703 | unknown |
       2|
               450 | unknown |
       21
           -----+
only showing top 20 rows
```

File colab_pdf.py already there; not retrieving.

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