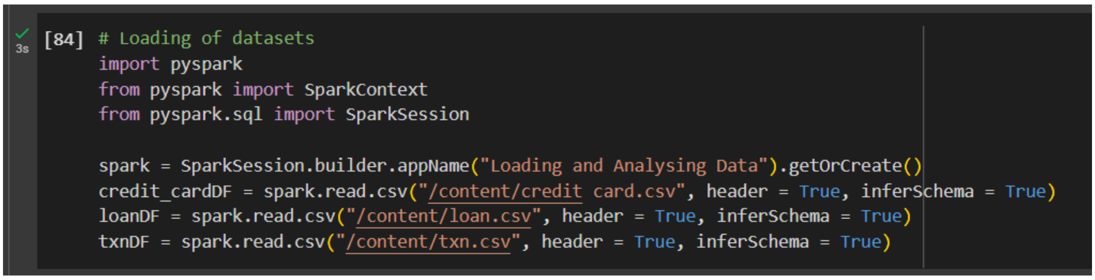
CASE STUDY – 1 (PySpark)

TRAINEE NAME: Swathi Baskaran

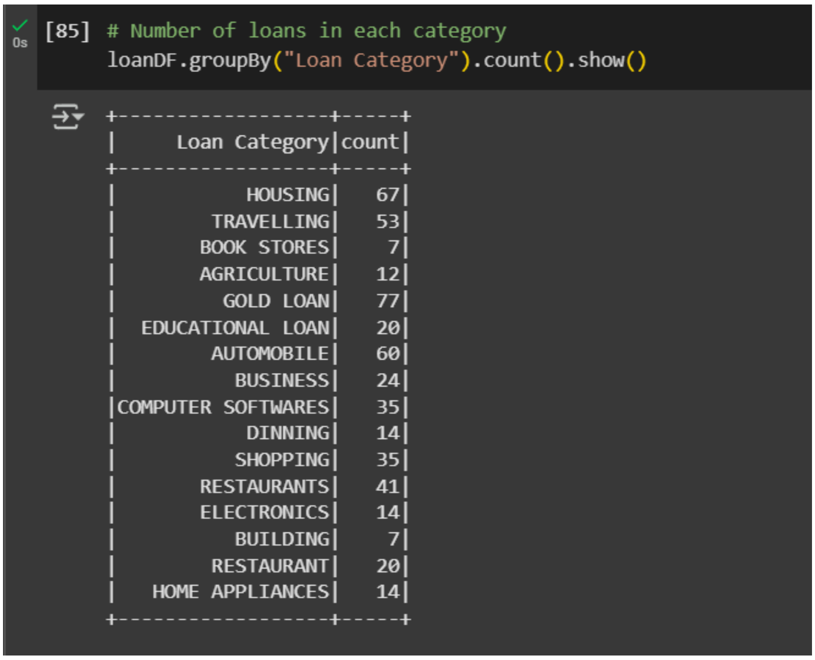
Loading of Datasets



```
[84] # Loading of datasets
     import pyspark
     from pyspark import SparkContext
     from pyspark.sql import SparkSession

     spark = SparkSession.builder.appName("Loading and Analysing Data").getOrCreate()
     credit_cardDF = spark.read.csv("/content/credit card.csv", header = True, inferSchema = True)
     loanDF = spark.read.csv("/content/loan.csv", header = True, inferSchema = True)
     txnDF = spark.read.csv("/content/txn.csv", header = True, inferSchema = True)
```

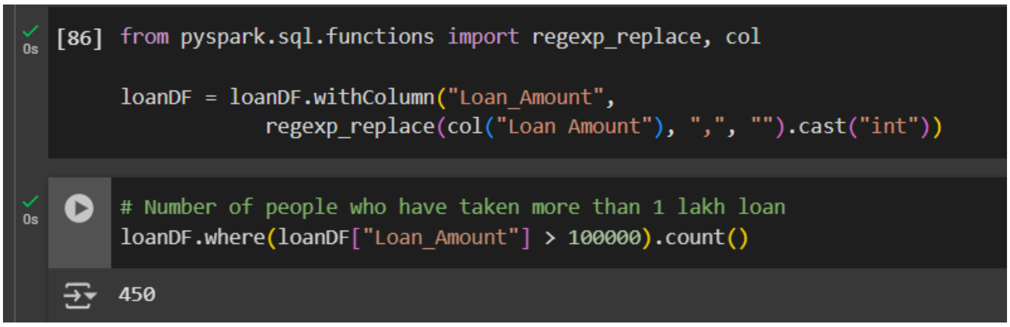
Number of loans in each category



```
[85] # Number of loans in each category
     loanDF.groupBy("Loan Category").count().show()
```

```
+------------------+-----+
|     Loan Category|count|
+------------------+-----+
|           HOUSING|   67|
|        TRAVELLING|   53|
|       BOOK STORES|    7|
|       AGRICULTURE|   12|
|         GOLD LOAN|   77|
|  EDUCATIONAL LOAN|   20|
|        AUTOMOBILE|   60|
|          BUSINESS|   24|
|COMPUTER SOFTWARES|   35|
|           DINNING|   14|
|          SHOPPING|   35|
|       RESTAURANTS|   41|
|       ELECTRONICS|   14|
|          BUILDING|    7|
|        RESTAURANT|   20|
|   HOME APPLIANCES|   14|
+------------------+-----+
```

Number of people who have taken more than 1 lakh loan



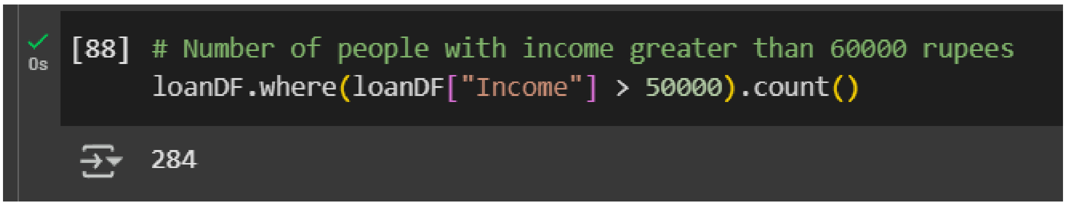
```
[86] from pyspark.sql.functions import regexp_replace, col

     loanDF = loanDF.withColumn("Loan_Amount",
                       regexp_replace(col("Loan Amount"), ",", "").cast("int"))
```

```
# Number of people who have taken more than 1 lakh loan
loanDF.where(loanDF["Loan_Amount"] > 100000).count()
```

```
450
```

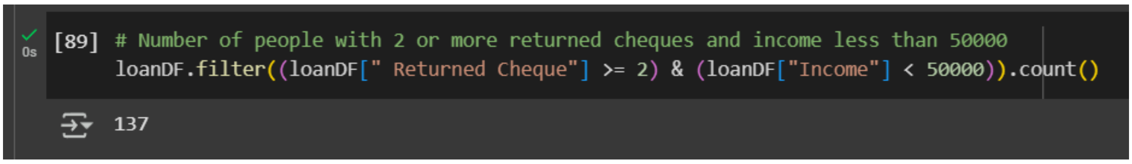
Number of people with income greater than 60000 rupees



```
[88] # Number of people with income greater than 60000 rupees
     loanDF.where(loanDF["Income"] > 50000).count()
```

```
284
```

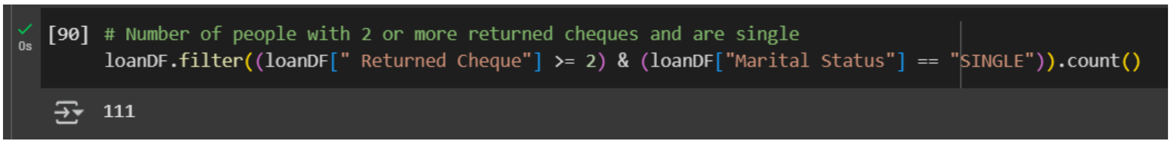
Number of people with 2 or more returned cheques and income less than 50000



```
[89] # Number of people with 2 or more returned cheques and income less than 50000
     loanDF.filter((loanDF[" Returned Cheque"] >= 2) & (loanDF["Income"] < 50000)).count()
```

```
137
```

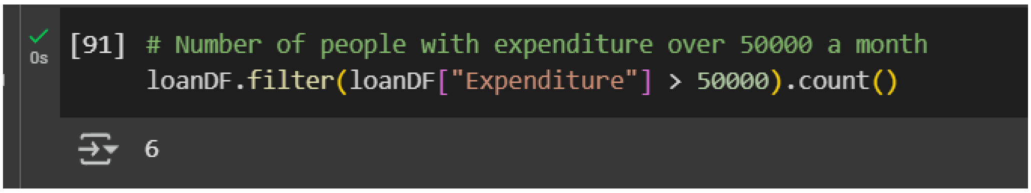
Number of people with 2 or more returned cheques and are single



```
[90] # Number of people with 2 or more returned cheques and are single
     loanDF.filter((loanDF[" Returned Cheque"] >= 2) & (loanDF["Marital Status"] == "SINGLE")).count()
```

```
111
```

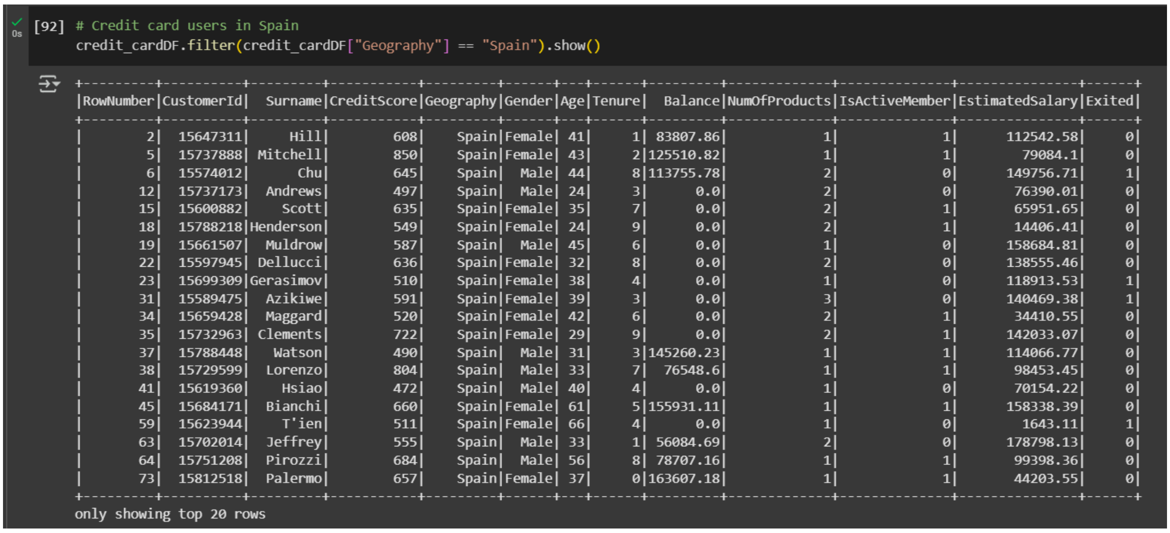
Number of people with expenditure over 50000 a month



```
[91] # Number of people with expenditure over 50000 a month
     loanDF.filter(loanDF["Expenditure"] > 50000).count()
```

```
6
```

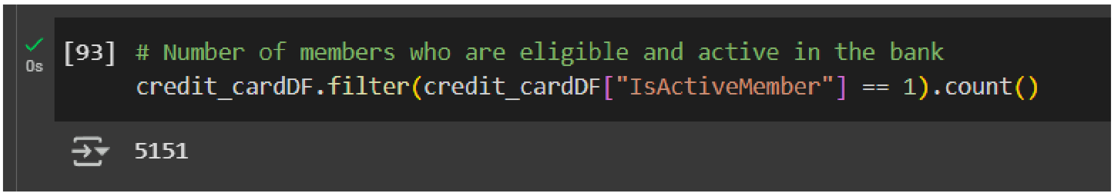
Credit card users in Spain



```
[92] # Credit card users in Spain
     credit_cardDF.filter(credit_cardDF["Geography"] == "Spain").show()
```

```
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|RowNumber|CustomerId|  Surname|CreditScore|Geography|Gender|Age|Tenure|  Balance|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
|        2|  15647311|     Hill|        608|    Spain|Female| 41|     1| 83807.86|            1|             1|      112542.58|     0|
|        5|  15737888| Mitchell|        850|    Spain|Female| 43|     2|125510.82|            1|             1|        79084.1|     0|
|        6|  15574012|      Chu|        645|    Spain|  Male| 44|     8|113755.78|            2|             0|      149756.71|     1|
|       12|  15737173|  Andrews|        497|    Spain|  Male| 24|     3|      0.0|            2|             0|       76390.01|     0|
|       15|  15600882|    Scott|        635|    Spain|Female| 35|     7|      0.0|            2|             1|       65951.65|     0|
|       18|  15788218|Henderson|        549|    Spain|Female| 24|     9|      0.0|            2|             1|       14406.41|     0|
|       19|  15661507|  Muldrow|        587|    Spain|  Male| 45|     6|      0.0|            1|             0|      158684.81|     0|
|       22|  15597945| Dellucci|        636|    Spain|Female| 32|     8|      0.0|            2|             0|      138555.46|     0|
|       23|  15699309|Gerasimov|        510|    Spain|Female| 38|     4|      0.0|            1|             0|      118913.53|     1|
|       31|  15589475|  Azikiwe|        591|    Spain|Female| 39|     3|      0.0|            3|             0|      140469.38|     1|
|       34|  15659428|  Maggard|        520|    Spain|Female| 42|     6|      0.0|            2|             1|       34410.55|     0|
|       35|  15732963| Clements|        722|    Spain|Female| 29|     9|      0.0|            2|             1|      142033.07|     0|
|       37|  15788448|   Watson|        490|    Spain|  Male| 31|     3|145260.23|            1|             1|      114066.77|     0|
|       38|  15729599|  Lorenzo|        804|    Spain|  Male| 33|     7|  76548.6|            1|             1|       98453.45|     0|
|       41|  15619360|    Hsiao|        472|    Spain|  Male| 40|     4|      0.0|            1|             0|       70154.22|     0|
|       45|  15684171|  Bianchi|        660|    Spain|Female| 61|     5|155931.11|            1|             1|      158338.39|     0|
|       59|  15623944|    T'ien|        511|    Spain|Female| 66|     4|      0.0|            1|             0|        1643.11|     1|
|       63|  15702014|  Jeffrey|        555|    Spain|  Male| 33|     1| 56084.69|            2|             0|      178798.13|     0|
|       64|  15751208|  Pirozzi|        684|    Spain|  Male| 56|     8| 78707.16|            1|             1|       99398.36|     0|
|       73|  15812518|  Palermo|        657|    Spain|Female| 37|     0|163607.18|            1|             1|       44203.55|     0|
+---------+----------+---------+-----------+---------+------+---+------+---------+-------------+--------------+---------------+------+
only showing top 20 rows
```

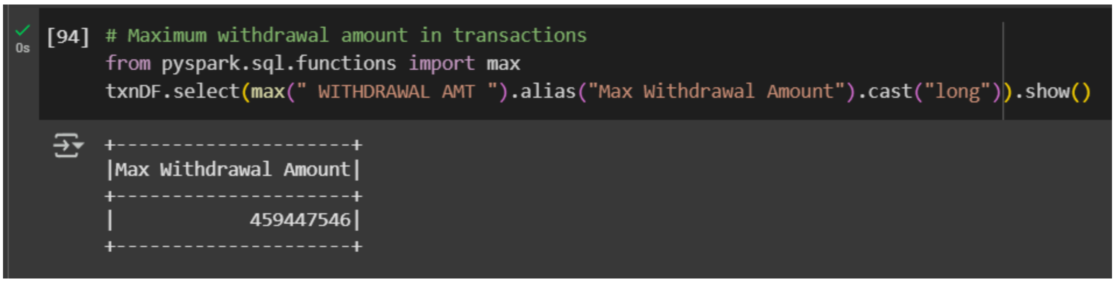
Number of members who are eligible and active in the bank



```
[93] # Number of members who are eligible and active in the bank
     credit_cardDF.filter(credit_cardDF["IsActiveMember"] == 1).count()
```

```
5151
```

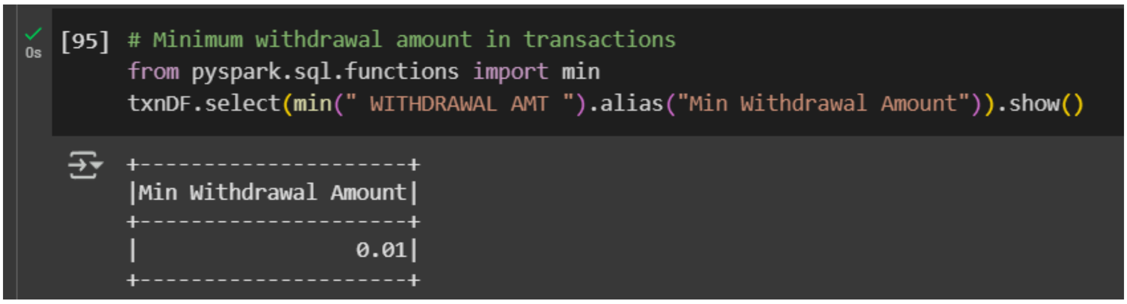
Maximum withdrawal amount in transactions



```
[94] # Maximum withdrawal amount in transactions
     from pyspark.sql.functions import max
     txnDF.select(max(" WITHDRAWAL AMT ").alias("Max Withdrawal Amount").cast("long")).show()
```

```
+---------------------+
|Max Withdrawal Amount|
+---------------------+
|            459447546|
+---------------------+
```

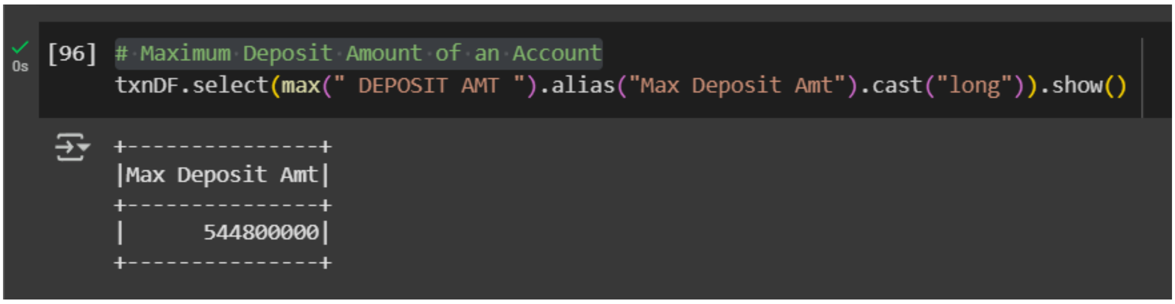
Minimum withdrawal amount in transactions



```
[95] # Minimum withdrawal amount in transactions
     from pyspark.sql.functions import min
     txnDF.select(min(" WITHDRAWAL AMT ").alias("Min Withdrawal Amount")).show()
```

```
+---------------------+
|Min Withdrawal Amount|
+---------------------+
|                 0.01|
+---------------------+
```

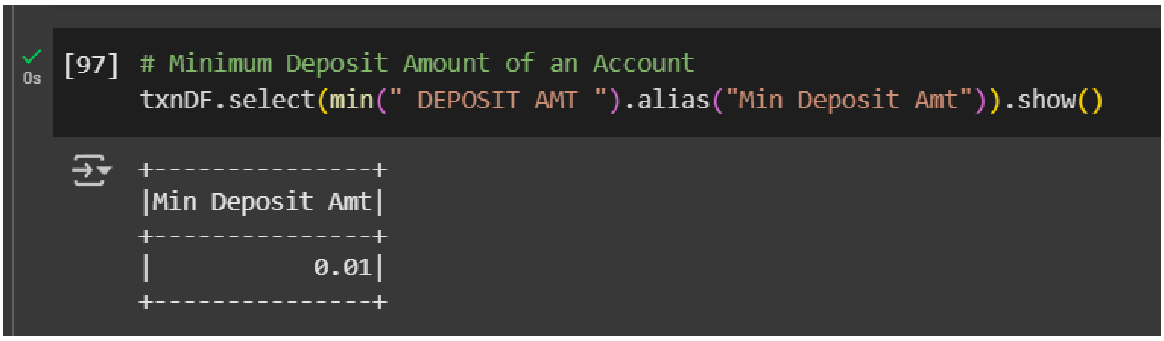
Maximum Deposit Amount of an Account



```
[96] # Maximum Deposit Amount of an Account
     txnDF.select(max(" DEPOSIT AMT ").alias("Max Deposit Amt").cast("long")).show()
```

```
+---------------+
|Max Deposit Amt|
+---------------+
|      544800000|
+---------------+
```

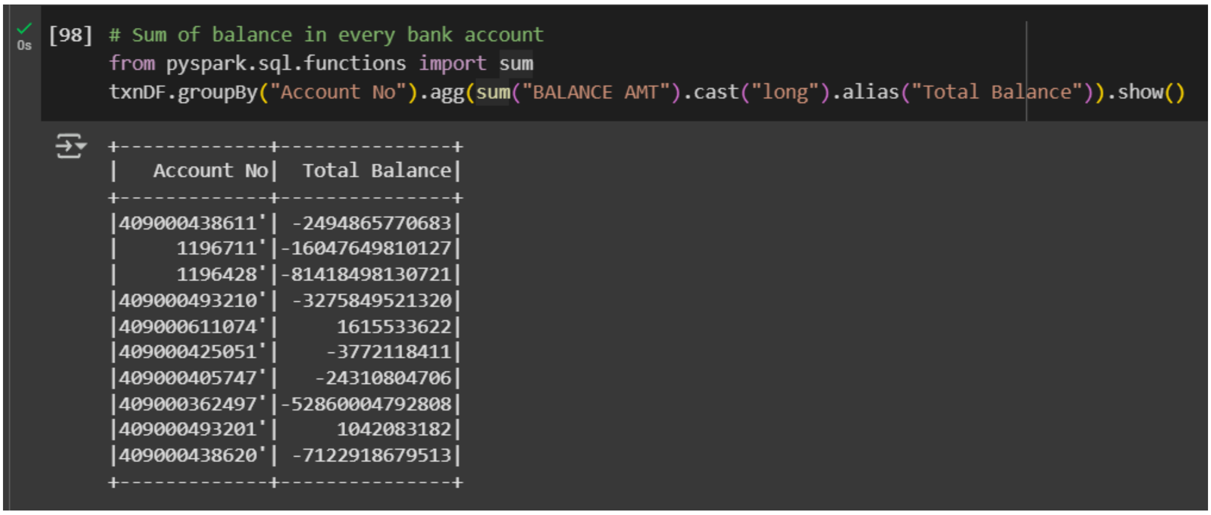
Minimum Deposit Amount of an Account



```
[97] # Minimum Deposit Amount of an Account
     txnDF.select(min(" DEPOSIT AMT ").alias("Min Deposit Amt")).show()
```

```
+---------------+
|Min Deposit Amt|
+---------------+
|           0.01|
+---------------+
```

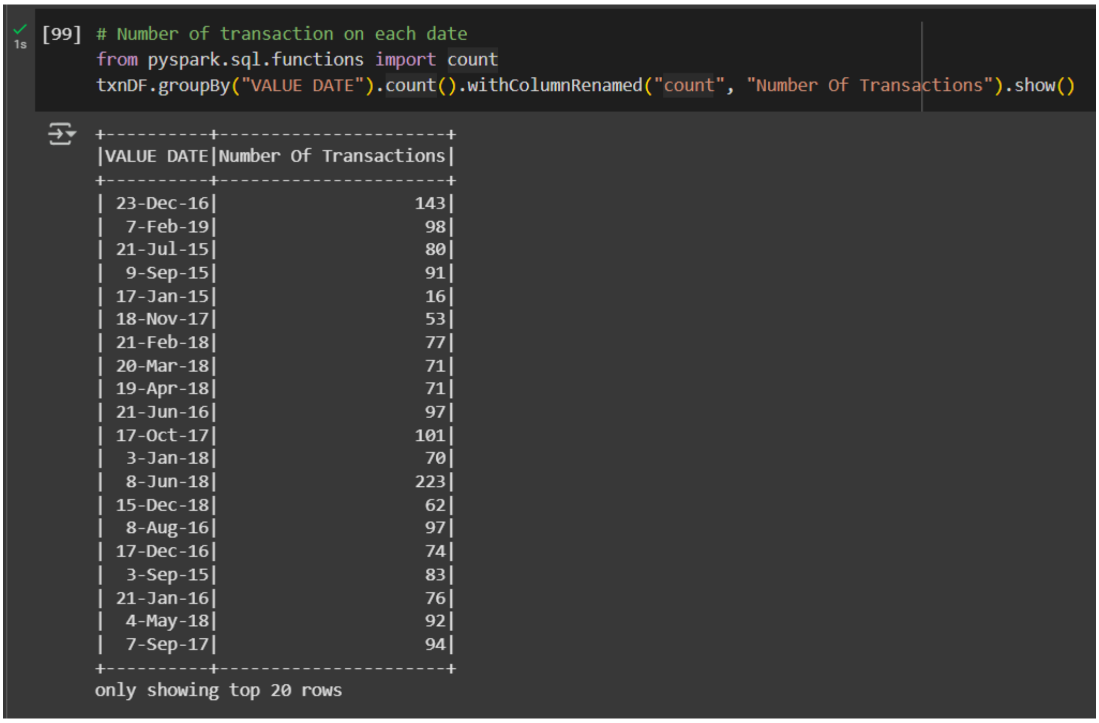
Sum of balance in every bank account



```
[98] # Sum of balance in every bank account
     from pyspark.sql.functions import sum
     txnDF.groupBy("Account No").agg(sum("BALANCE AMT").cast("long").alias("Total Balance")).show()
```

```
+------------+--------------+
|   Account No| Total Balance|
+------------+--------------+
|409000438611'| -2494865770683|
|     1196711'|-16047649810127|
|     1196428'|-81418498130721|
|409000493210'| -3275849521320|
|409000611074'|     1615533622|
|409000425051'|    -3772118411|
|409000405747'|   -24310804706|
|409000362497'|-52860004792808|
|409000493201'|     1042083182|
|409000438620'| -7122918679513|
+------------+--------------+
```

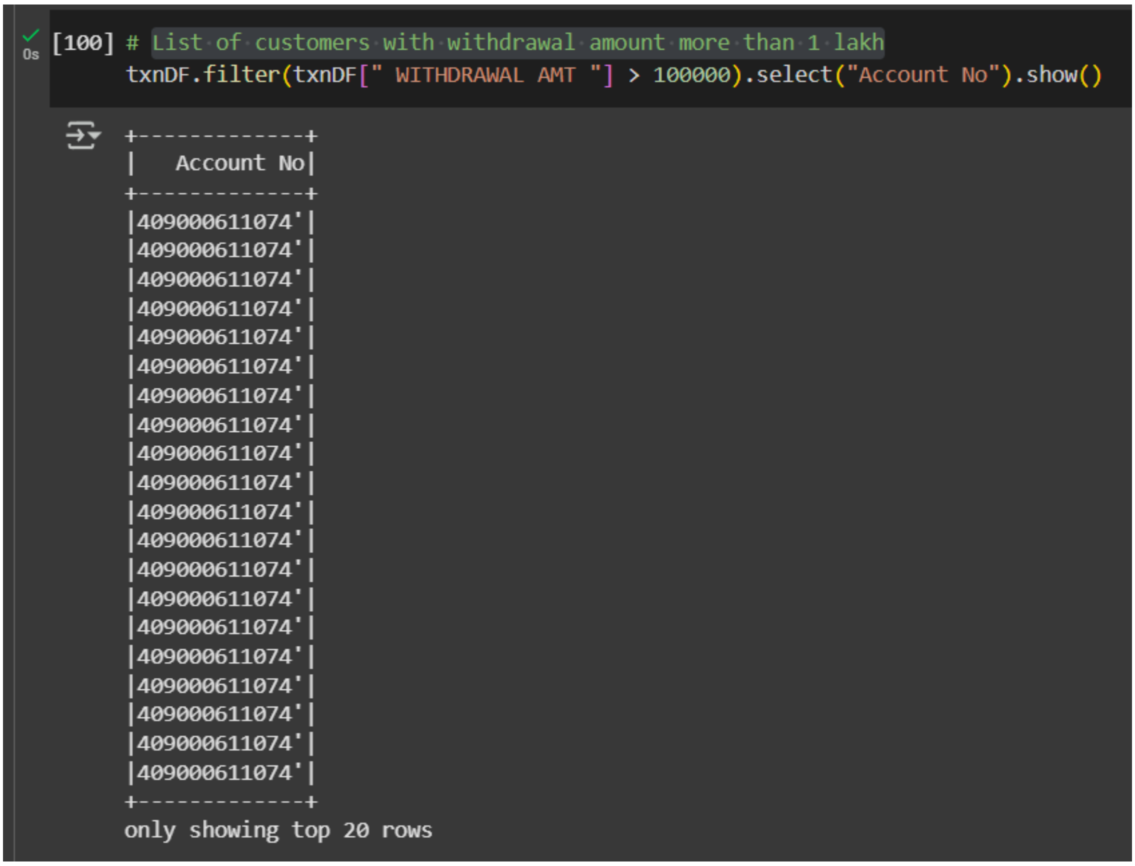
Number of transactions on each date



```
[99] # Number of transaction on each date
     from pyspark.sql.functions import count
     txnDF.groupBy("VALUE DATE").count().withColumnRenamed("count", "Number Of Transactions").show()
```

```
+----------+----------------------+
|VALUE DATE|Number Of Transactions|
+----------+----------------------+
| 23-Dec-16|                   143|
|  7-Feb-19|                    98|
| 21-Jul-15|                    80|
|  9-Sep-15|                    91|
| 17-Jan-15|                    16|
| 18-Nov-17|                    53|
| 21-Feb-18|                    77|
| 20-Mar-18|                    71|
| 19-Apr-18|                    71|
| 21-Jun-16|                    97|
| 17-Oct-17|                   101|
|  3-Jan-18|                    70|
|  8-Jun-18|                   223|
| 15-Dec-18|                    62|
|  8-Aug-16|                    97|
| 17-Dec-16|                    74|
|  3-Sep-15|                    83|
| 21-Jan-16|                    76|
|  4-May-18|                    92|
|  7-Sep-17|                    94|
+----------+----------------------+
only showing top 20 rows
```

List of customers with withdrawal amount more than 1 lakh



```
[100] # List of customers with withdrawal amount more than 1 lakh
      txnDF.filter(txnDF[" WITHDRAWAL AMT "] > 100000).select("Account No").show()
```

```
+-------------+
|   Account No|
+-------------+
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
|409000611074'|
+-------------+
only showing top 20 rows
```