

```
In [2]: import findspark
findspark.init()
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()
from pyspark.sql.functions import *
```

## LOAN DATASET

```
In [14]: df = spark.read.csv("C:/Users/sidse/Downloads/big data course/ASSIGNMENTS/project
```



```
In [15]: df.printSchema()
```

```
root
|-- Customer_ID: string (nullable = true)
|-- Age: integer (nullable = true)
|-- Gender: string (nullable = true)
|-- Occupation: string (nullable = true)
|-- Marital Status: string (nullable = true)
|-- Family Size: integer (nullable = true)
|-- Income: integer (nullable = true)
|-- Expenditure: integer (nullable = true)
|-- Use Frequency: integer (nullable = true)
|-- Loan Category: string (nullable = true)
|-- Loan Amount: string (nullable = true)
|-- Overdue: integer (nullable = true)
|-- Debt Record: integer (nullable = true)
|-- Returned Cheque: integer (nullable = true)
|-- Dishonour of Bill: integer (nullable = true)
```

In [16]: `df.show(5)`

```

+-----+---+-----+-----+-----+-----+-----+-----+
--+-----+-----+-----+-----+-----+-----+-----+
--+-----+
|Customer_ID|Age|Gender| Occupation|Marital Status|Family Size|Income|Expendit
ure|Use Frequency|Loan Category|Loan Amount|Overdue| Debt Record| Returned Cheq
ue| Dishonour of Bill|
+-----+---+-----+-----+-----+-----+-----+-----+
--+-----+-----+-----+-----+-----+-----+-----+
--+-----+
| IB14001| 30| MALE|BANK MANAGER| SINGLE| 4| 50000| 22
199| 6| HOUSING| 10,00,000 | 5| 42898|
6| 9|
| IB14008| 44| MALE| PROFESSOR| MARRIED| 6| 51000| 19
999| 4| SHOPPING| 50000| 3| 33999|
1| 5|
| IB14012| 30|FEMALE| DENTIST| SINGLE| 3| 58450| 27
675| 5| TRAVELLING| 75000| 6| 20876|
3| 1|
| IB14018| 29| MALE| TEACHER| MARRIED| 5| 45767| 12
787| 3| GOLD LOAN| 6,00,000 | 7| 11000|
0| 4|
| IB14022| 34| MALE| POLICE| SINGLE| 4| 43521| 11
999| 3| AUTOMOBILE| 2,00,000 | 2| 43898|
1| 2|
+-----+---+-----+-----+-----+-----+-----+-----+
--+-----+-----+-----+-----+-----+-----+-----+
--+-----+
only showing top 5 rows

```

In [6]: `len(df.columns)`

Out[6]: 15

In [7]: `df.count()`

Out[7]: 500

In [8]: `df.distinct().count()`

Out[8]: 500

In [11]: *#number of Loans in each category*  
`df.groupby("Loan Category").count().orderBy("count", ascending = False).show()`

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

In [14]: *#number of people who have taken more than 1 Lack Loan*  
`df.filter(df["Loan Amount"]>"1,00,000").count()`

Out[14]: 379

In [18]: *#number of people with income greater than 60000 rupees*  
`df.filter(df["Income"]>"60000").count()`

Out[18]: 198

In [26]: *#number of people with 2 or more returned cheques and income Less than 50000*  
`df.filter((df["Returned Cheque"]>"1") & (df["Income"]<"50000")).count()`

Out[26]: 137

In [27]: *#number of people with 2 or more returned cheques and are single*  
`df.filter((df["Returned Cheque"]>"1") & (df["Marital Status"]<"SINGLE")).count()`

Out[27]: 283

In [6]: *#number of people with expenditure over 50000 a month*  
`df.filter((df["Expenditure"]>"50000")).show()`

```
+-----+---+-----+-----+-----+-----+-----+
|Customer_ID|Age|Gender|Occupation|Marital Status|Family Size|Income|Expenditure|Use Frequency|Loan Category|Loan Amount|Overdue|Debt Record|Returned Cheque|Dishonour of Bill|
+-----+---+-----+-----+-----+-----+-----+
|IB14158|54|MALE|AIRPORT OFFICER|MARRIED|6|80000|62541|2|AUTOMOBILE|20,45,789|1|16599|
|IB14176|54|MALE|AIRPORT OFFICER|MARRIED|6|80000|62541|2|HOUSING|20,45,789|1|16599|
|IB14204|54|MALE|AIRPORT OFFICER|MARRIED|6|81000|62541|2|DINNING|20,45,789|1|16599|
|IB14227|54|MALE|AIRPORT OFFICER|MARRIED|6|80000|62541|2|HOUSING|20,45,789|1|16599|
```

In [22]: `df.select("Customer_ID", "Age", "Occupation", "Marital Status", "Income").filter((df["Income"]>50000)).show()`

```
+-----+---+-----+-----+-----+
|Customer_ID|Age|Occupation|Marital Status|Income|
+-----+---+-----+-----+-----+
|IB14001|30|BANK MANAGER|SINGLE|50000|
|IB14012|30|DENTIST|SINGLE|58450|
|IB14085|30|ELECTRICIAN|MARRIED|30000|
|IB14187|30|DENTIST|SINGLE|58450|
|IB14220|30|DENTIST|SINGLE|58450|
|IB14304|30|ELECTRICIAN|MARRIED|30000|
|IB14497|30|ACCOUNT MANAGER|MARRIED|52568|
|IB14566|30|CHARTERED APPRAISER|MARRIED|81225|
|IB14633|30|SOFTWARE ENGINEER|MARRIED|62522|
|IB14667|30|AGRICULTURAL ENGI...|SINGLE|85289|
|IB14706|30|TEACHER|SINGLE|51564|
|IB14721|30|PILOT|SINGLE|76333|
|IB14800|30|TEACHER|SINGLE|49370|
|IB14804|30|BUSINESS|SINGLE|53957|
|IB14831|30|NAVY|SINGLE|51862|
|IB14976|30|CHARTERED APPRAISER|MARRIED|85225|
|IB15011|30|TEACHER|SINGLE|55564|
|IB15031|30|PILOT|SINGLE|86333|
+-----+---+-----+-----+-----+
```

```
In [20]: n = df.withColumn("Age", when(df.Age==30,40).otherwise(df.Age))
n.show()
```

```
+-----+---+-----+-----+-----+-----+-----+-----+
+-----+---+-----+-----+-----+-----+-----+-----+
+-----+---+-----+-----+-----+-----+-----+-----+
|Customer_ID|Age|Gender|Occupation|Marital Status|Family Size|Income|Expenditure|Use Frequency|Loan Category|Loan Amount|Overdue|Debt Record|Returned Cheque|Dishonour of Bill|
+-----+---+-----+-----+-----+-----+-----+-----+
+-----+---+-----+-----+-----+-----+-----+-----+
|IB14001|40|MALE|BANK MANAGER|SINGLE|4|50000|22199|6|HOUSING|10,00,000|5|42898|6|9|
|IB14008|44|MALE|PROFESSOR|MARRIED|6|51000|19999|4|SHOPPING|50000|3|33999|1|5|
|IB14012|40|FEMALE|DENTIST|SINGLE|3|58450|27675|5|TRAVELLING|75000|6|20876|3|1|
|IB14018|29|MALE|TEACHER|MARRIED|5|45767|12787|3|GOLD LOAN|6,00,000|7|11000|0|4|
|IB14022|34|MALE|POLICE|SINGLE|4|43521|11999|3|AUTOMOBILE|2,00,000|2|43898|1|2|
|IB14024|55|FEMALE|NURSE|MARRIED|6|34999|19888|4|AUTOMOBILE|47787|1|50000|0|3|
|IB14025|39|FEMALE|TEACHER|MARRIED|6|46619|18675|4|HOUSING|12,09,867|8|29999|6|8|
|IB14027|51|MALE|SYSTEM MANAGER|MARRIED|3|49999|19111|5|RESTAURANTS|60676|8|13000|2|5|
|IB14029|24|FEMALE|TEACHER|SINGLE|3|45008|17454|4|AUTOMOBILE|3,99,435|9|51987|4|7|
|IB14031|37|FEMALE|SOFTWARE ENGINEER|MARRIED|5|55999|23999|5|AUTOMOBILE|60999|2|0|5|3|
|IB14032|24|MALE|DATA ANALYST|SINGLE|4|60111|28999|6|AUTOMOBILE|35232|5|33333|1|2|
|IB14034|32|MALE|PRODUCT ENGINEER|MARRIED|6|null|29000|7|COMPUTER SOFTWARES|80660|6|4500|5|4|
|IB14037|54|FEMALE|TEACHER|MARRIED|5|48099|19999|4|RESTAURANTS|30999|1|12000|7|5|
|IB14039|45|MALE|ACCOUNT MANAGER|MARRIED|7|45777|18452|4|GOLD LOAN|9,87,611|7|39999|8|1|
|IB14041|59|FEMALE|ASSISTANT PROFESSOR|MARRIED|4|50999|22999|5|EDUCATIONAL LOAN|5,99,934|3|9000|
```

```

9|          9|
|  IB14042| 25| FEMALE|          DOCTOR|          SINGLE|          4| 60111|
27111|          5|          TRAVELLING| 12,90,929 |          4| 18000|
1|          0|
|  IB14045| 31|  MALE|          STORE KEEPER|          SINGLE|          5| 40999|
11999|          3|          BOOK STORES| 1,67,654 |          1| 4500|
0|          1|
|  IB14049| 49|  MALE|          BANK MANAGER|          MARRIED|          4| 45999|
14500|          4|          TRAVELLING|          79999|          4| 6700|
7|          3|
|  IB14050| 56|  MALE|          CIVIL ENGINEER|          MARRIED|          4| null|
13999|          3|          HOUSING| 10,65,577 |          6| 19999|
4|          2|
|  IB14054| 58| FEMALE|          DOCTOR|          MARRIED|          5| 60000|
25000|          5|          HOUSING| 9,00,000 |          5| 21000|
9|          0|
+-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+-----+-----+-----+-----+
-----+-----+-----+
only showing top 20 rows

```

```
In [28]: n.filter(df.Customer_ID=="IB14001").show()
```

```

+-----+-----+-----+-----+-----+-----+-----+
--+-----+-----+
|Customer_ID|Age|Gender| Occupation|Marital Status|Family Size|Income|Expendit
ure|Use Frequency|Loan Category|Loan Amount|Overdue| Debt Record| Returned Cheq
ue| Dishonour of Bill|
+-----+-----+-----+-----+-----+-----+-----+
--+-----+-----+
|  IB14001| 40|  MALE|BANK MANAGER|          SINGLE|          4| 50000|          22
199|          6|          HOUSING| 10,00,000 |          5|          42898|
6|          9|
+-----+-----+-----+-----+-----+-----+-----+
--+-----+-----+

```

```
In [30]: df.select("Marital Status").distinct().show()
```

```

+-----+
|Marital Status|
+-----+
|          SINGLE|
|          MARRIED|
+-----+

```

## CREDIT CARD DATASET

```
In [9]: dfc = spark.read.csv("C:/Users/sidse/Downloads/big data course/ASSIGNMENTS/projec
```

```
In [10]: dfc.printSchema()
```

```
root
|-- RowNumber: integer (nullable = true)
|-- CustomerId: integer (nullable = true)
|-- Surname: string (nullable = true)
|-- CreditScore: integer (nullable = true)
|-- Geography: string (nullable = true)
|-- Gender: string (nullable = true)
|-- Age: integer (nullable = true)
|-- Tenure: integer (nullable = true)
|-- Balance: double (nullable = true)
|-- NumOfProducts: integer (nullable = true)
|-- IsActiveMember: integer (nullable = true)
|-- EstimatedSalary: double (nullable = true)
|-- Exited: integer (nullable = true)
```

```
In [17]: len(dfc.columns)
```

```
Out[17]: 13
```

```
In [16]: dfc.count()
```

```
Out[16]: 10000
```

```
In [18]: dfc.distinct().count()
```

```
Out[18]: 10000
```

In [11]: dfc.show(5)

```

+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure|  Balanc
e|NumOfProducts|IsActiveMember|EstimatedSalary|Exited|
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|          1| 15634602|Hargrave|        619|   France|Female| 42|      2|      0.
0|          1|          1|    101348.88|      1|
|          2| 15647311|   Hill|        608|   Spain|Female| 41|      1| 83807.8
6|          1|          1|    112542.58|      0|
|          3| 15619304|   Onio|        502|   France|Female| 42|      8| 159660.
8|          3|          0|    113931.57|      1|
|          4| 15701354|   Boni|        699|   France|Female| 39|      1|      0.
0|          2|          0|    93826.63|      0|
|          5| 15737888|Mitchell|        850|   Spain|Female| 43|      2|125510.8
2|          1|          1|    79084.1|      0|
+-----+-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows

```

In [13]: *#number of members who are eligible for credit card*  
dfc.filter(dfc["CreditScore"]>700).count()

Out[13]: 3116

In [19]: *#number of members who are eligible and active in the bank*  
dfc.filter((dfc["IsActiveMember"]==1) & (dfc["CreditScore"]>700)).count()

Out[19]: 1637



```
In [21]: #credit card users in Spain
dfc.filter(dfc["Geography"]=="Spain").show()
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+
|RowNumber|CustomerId| Surname|CreditScore|Geography|Gender|Age|Tenure| Balan
ce|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| 7654
8.6|      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
```

```
In [24]: dfc.filter((dfc["EstimatedSalary"]>100000) & (dfc["Exited"]==1)).count()
```

```
Out[24]: 1044
```

```
In [25]: dfc.filter((dfc["EstimatedSalary"]<100000) & (dfc["NumOfProducts"]>1)).count()
```

```
Out[25]: 2432
```

## TRANSACTION DATASET

```
In [3]: txn = spark.read.csv("C:/Users/sidse/Downloads/big data course/ASSIGNMENTS/project")
```

```
In [4]: txn.printSchema()
```

```
root
 |-- Account No: string (nullable = true)
 |-- TRANSACTION DETAILS: string (nullable = true)
 |-- VALUE DATE: string (nullable = true)
 |-- WITHDRAWAL AMT : double (nullable = true)
 |-- DEPOSIT AMT : double (nullable = true)
 |-- BALANCE AMT: double (nullable = true)
```

```
In [5]: #COUNT OF TRANSACTION ON EVERY ACCOUNT
```

```
txn.groupBy("Account No").count().orderBy("count", ascending = False).show(20, Fa
```

```
+-----+-----+
|Account No|count|
+-----+-----+
|1196428'|48779|
|409000362497'|29840|
|409000438620'|13454|
|1196711'|10536|
|409000493210'|6014|
|409000438611'|4588|
|409000611074'|1093|
|409000493201'|1044|
|409000425051'|802|
|409000405747'|51|
+-----+-----+
```

In [10]: *#Maximum withdrawal amount*  
 txn.groupBy("Account No").max(" WITHDRAWAL AMT ").orderBy("max( WITHDRAWAL AMT )")

Account No	max( WITHDRAWAL AMT )
1196711'	4.594475464E8
409000438620'	4.0E8
409000425051'	3.54E8
409000438611'	2.4E8
409000405747'	1.7E8
1196428'	1.5E8
409000362497'	1.413662392E8
409000493210'	1.5E7
409000493201'	2500000.0
409000611074'	912000.0

In [24]: *#MINIMUM WITHDRAWAL AMOUNT OF AN ACCOUNT*  
 txn.groupBy("Account No").min(" WITHDRAWAL AMT ").orderBy("min( WITHDRAWAL AMT )")

Account No	min( WITHDRAWAL AMT )
409000493210'	0.01
409000438611'	0.2
1196711'	0.25
1196428'	0.25
409000438620'	0.34
409000362497'	0.97
409000425051'	1.25
409000493201'	2.1
409000405747'	21.0
409000611074'	120.0

In [28]: *#MAXIMUM DEPOSIT AMOUNT OF AN ACCOUNT*  
 txn.groupBy("Account No").max(" DEPOSIT AMT ").orderBy("max( DEPOSIT AMT )", asce

Account No	max( DEPOSIT AMT )
409000438620'	5.448E8
1196711'	5.0E8
1196428'	2.119594422E8
409000405747'	2.021E8
409000362497'	2.0E8
409000438611'	1.7025E8
409000425051'	1.5E7
409000493210'	1.5E7
409000611074'	3000000.0
409000493201'	1000000.0

In [29]: *#MINIMUM DEPOSIT AMOUNT OF AN ACCOUNT*  
 txn.groupBy("Account No").min(" DEPOSIT AMT ").orderBy("min( DEPOSIT AMT )").show

Account No	min( DEPOSIT AMT )
409000493210'	0.01
409000438611'	0.03
409000362497'	0.03
409000438620'	0.07
409000493201'	0.9
1196428'	1.0
409000425051'	1.0
1196711'	1.01
409000405747'	500.0
409000611074'	1320.0

```
In [9]: #sum of balance in every bank account
txn.groupBy("Account No").sum("BALANCE AMT").show()
```

```
+-----+-----+
| Account No | sum(BALANCE AMT) |
+-----+-----+
| 409000438611 | -2.49486577068339... |
| 1196711 | -1.60476498101275E13 |
| 1196428 | -8.1418498130721E13 |
| 409000493210 | -3.27584952132095... |
| 409000611074 | 1.615533622E9 |
| 409000425051 | -3.77211841164998... |
| 409000405747 | -2.43108047067000... |
| 409000362497 | -5.2860004792808E13 |
| 409000493201 | 1.0420831829499985E9 |
| 409000438620 | -7.12291867951358... |
+-----+-----+
```

```
In [32]: #Number of transaction on each date
txn.groupBy("VALUE DATE").count().orderBy("count", ascending = False).show()
```

```
+-----+-----+
| VALUE DATE | count |
+-----+-----+
| 27-Jul-17 | 567 |
| 13-Aug-18 | 463 |
| 8-Nov-17 | 402 |
| 7-Oct-17 | 382 |
| 10-Jul-18 | 374 |
| 12-Dec-17 | 367 |
| 12-Sep-18 | 365 |
| 9-Aug-18 | 360 |
| 19-Sep-17 | 358 |
| 16-Mar-17 | 353 |
| 10-Sep-18 | 344 |
| 14-Jul-17 | 333 |
| 7-Mar-18 | 319 |
| 11-Oct-18 | 303 |
| 22-Aug-17 | 301 |
| 9-Jan-18 | 299 |
| 9-Oct-18 | 297 |
+-----+-----+
```

```
In [11]: #List of customers with withdrawal amount more than 1 Lakh
txn.select("Account No","TRANSACTION DETAILS"," WITHDRAWAL AMT ").filter(txn[" W]
```

Account No	TRANSACTION DETAILS	WITHDRAWAL AMT
409000611074'	INDO GIBL Indiaforensic STL01071	133900.0
409000611074'	INDO GIBL Indiaforensic STL04071	195800.0
409000611074'	INDO GIBL Indiaforensic STL10071	143800.0
409000611074'	INDO GIBL Indiaforensic STL11071	331650.0
409000611074'	INDO GIBL Indiaforensic STL12071	129000.0
409000611074'	INDO GIBL Indiaforensic STL13071	230013.0
409000611074'	INDO GIBL Indiaforensic STL14071	367900.0
409000611074'	INDO GIBL Indiaforensic STL15071	108000.0
409000611074'	INDO GIBL Indiaforensic STL17071	141000.0
409000611074'	INDO GIBL Indiaforensic STL22071	206000.0
409000611074'	INDO GIBL Indiaforensic STL02081	242300.0
409000611074'	INDO GIBL Indiaforensic STL04081	113250.0
409000611074'	INDO GIBL Indiaforensic STL07081	206900.0
409000611074'	INDO GIBL Indiaforensic STL08081	276000.0
409000611074'	INDO GIBL Indiaforensic STL09081	171000.0
409000611074'	INDO GIBL Indiaforensic STL11081	189800.0
409000611074'	INDO GIBL Indiaforensic STL14081	271323.0
409000611074'	INDO GIBL Indiaforensic STL16081	200600.0
409000611074'	INDO GIBL Indiaforensic STL17081	176900.0
409000611074'	INDO GIBL Indiaforensic STL18081	150050.0

only showing top 20 rows