2/3/2020

MARKET ANALYSIS IN BANKING DOMAIN

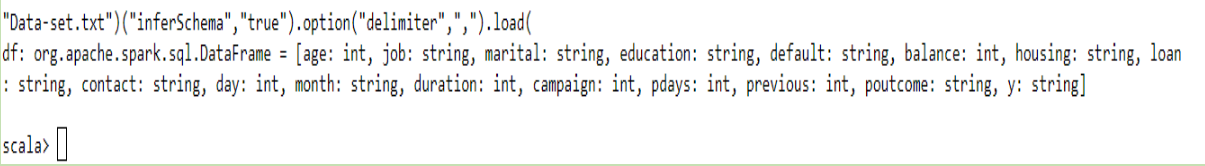
Project 4

Olwethu Matiwane

Question One

1 . Load data and create a Spark data frame

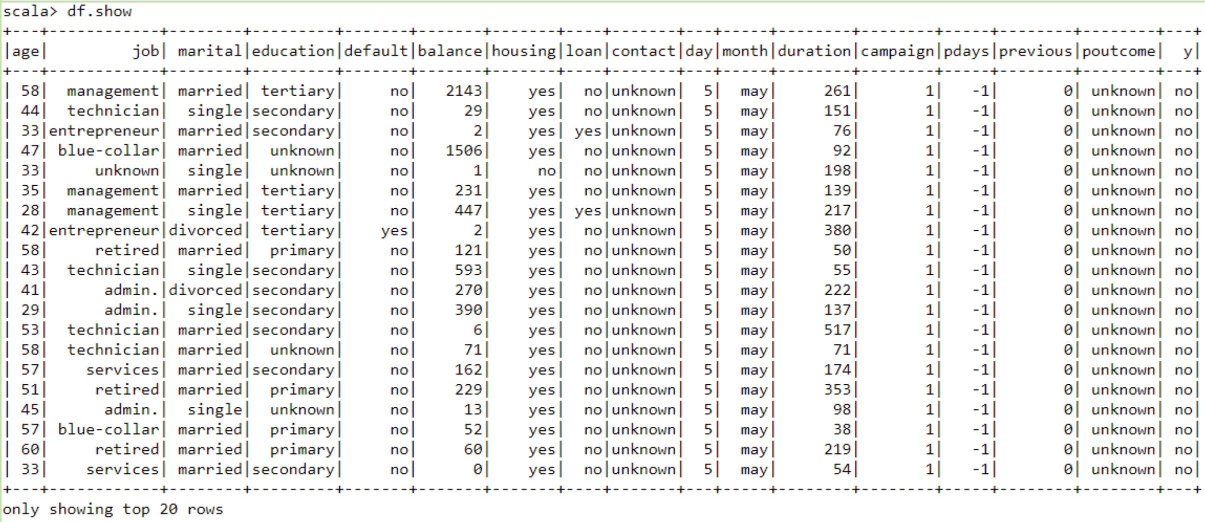
- val df = sqlContext.read.format("com.databricks.spark.csv").option("header","true").option("inferSchema","true").option("delimiter",",").load("Data-set.txt")
- **Output:**



```
"Data-set.txt")("inferSchema","true").option("delimiter",",").load(
df: org.apache.spark.sql.DataFrame = [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, y: string]

scala> 
```

-



```
scala> df.show
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
|age|         job| marital|education|default|balance|housing|loan|contact|day|month|duration|campaign|pdays|previous|poutcome|  y|
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5|  may|      55|       1|   -1|       0| unknown| no|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5|  may|     222|       1|   -1|       0| unknown| no|
| 29|      admin.|  single|secondary|     no|    390|    yes|  no|unknown|  5|  may|     137|       1|   -1|       0| unknown| no|
| 53|  technician| married|secondary|     no|      6|    yes|  no|unknown|  5|  may|     517|       1|   -1|       0| unknown| no|
| 58|  technician| married|  unknown|     no|     71|    yes|  no|unknown|  5|  may|      71|       1|   -1|       0| unknown| no|
| 57|    services| married|secondary|     no|    162|    yes|  no|unknown|  5|  may|     174|       1|   -1|       0| unknown| no|
| 51|     retired| married|  primary|     no|    229|    yes|  no|unknown|  5|  may|     353|       1|   -1|       0| unknown| no|
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5|  may|     219|       1|   -1|       0| unknown| no|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5|  may|      54|       1|   -1|       0| unknown| no|
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
only showing top 20 rows
```

-
- **Analysis:** The Dataframe shows different ages of different people and the marital status as well as personal details.

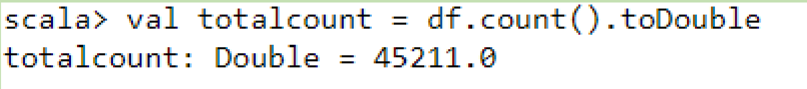
Question Two

2. Give marketing success rate (No. of people subscribed / total no. of entries)

Give marketing failure rate

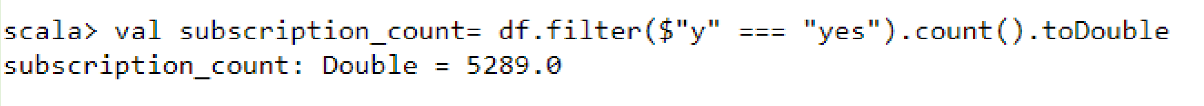
- val totalcount = df.count().toDouble
- val subscription_count= df.filter($"y" === "yes").count().toDouble
- val success_rate = subscription_count/totalcount

- **Output:**



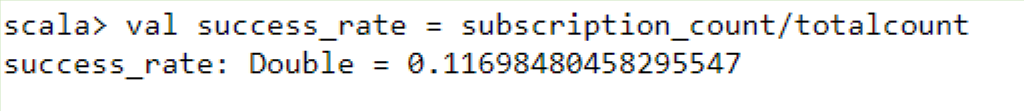
```
scala> val totalcount = df.count().toDouble
totalcount: Double = 45211.0
```

-



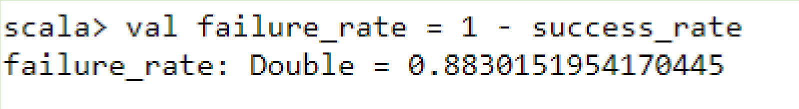
```
scala> val subscription_count= df.filter($"y" === "yes").count().toDouble
subscription_count: Double = 5289.0
```

-



```
scala> val success_rate = subscription_count/totalcount
success_rate: Double = 0.11698480458295547
```

-



```
scala> val failure_rate = 1 - success_rate
failure_rate: Double = 0.8830151954170445
```

-

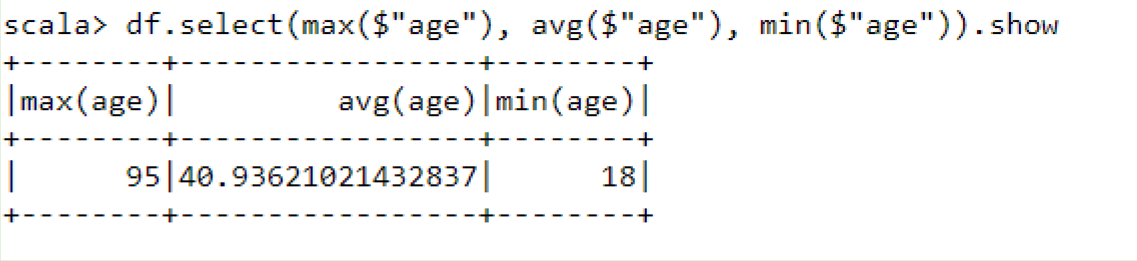
- **Analysis:**

- Number of entries = 45211.0, Bank success rate = 0.11 and Subscriptions = 5289.0 to the term deposit, this shows that the marketing campaign was not successful.

<u>Question three</u>

3. Give the maximum, mean, and minimum age of the average targeted customer

- **Output:**



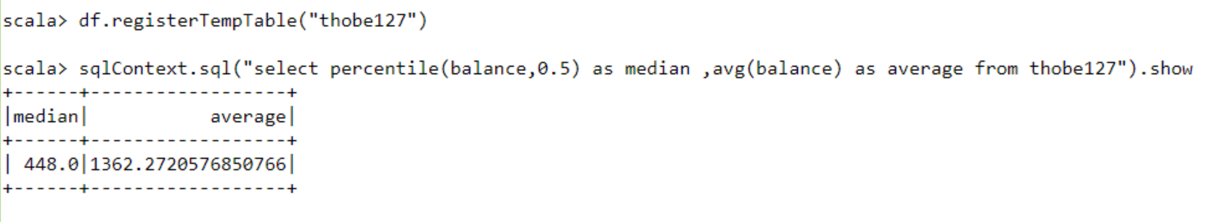
```
scala> df.select(max($"age"), avg($"age"), min($"age")).show
+--------+----------------+--------+
|max(age)|        avg(age)|min(age)|
+--------+----------------+--------+
|      95|40.93621021432837|      18|
+--------+----------------+--------+
```

-
- **Analysis:**
- This table shows the age of targeted customers, which displays the maximum age of 95 years, average age of 41 years and the minimum age of 18 years.

<u>Question four</u>

4. Check the quality of customers by checking average balance, median balance of customers

- **Output:**



```
scala> df.registerTempTable("thobe127")

scala> sqlContext.sql("select percentile(balance,0.5) as median ,avg(balance) as average from thobe127").show
+------+------------------+
|median|           average|
+------+------------------+
| 448.0|1362.2720576850766|
+------+------------------+
```

-

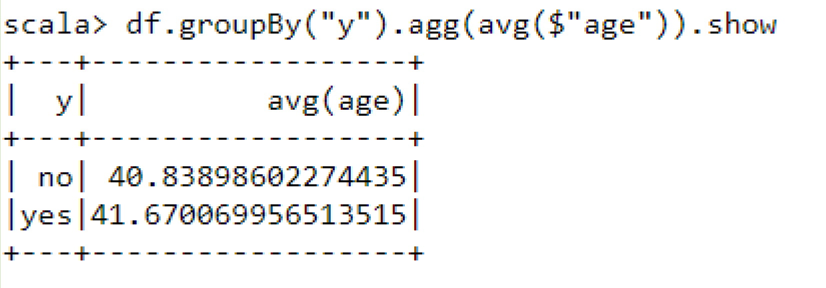
- **Analysis:**

The median is 448.0
The average is 1362.272

Question five

5. Check if age matters in marketing subscription for deposit

- **Output:**

-


```
scala> df.groupBy("y").agg(avg($"age")).show
+---+------------------+
|  y|          avg(age)|
+---+------------------+
| no| 40.83898602274435|
|yes|41.670069956513515|
+---+------------------+
```

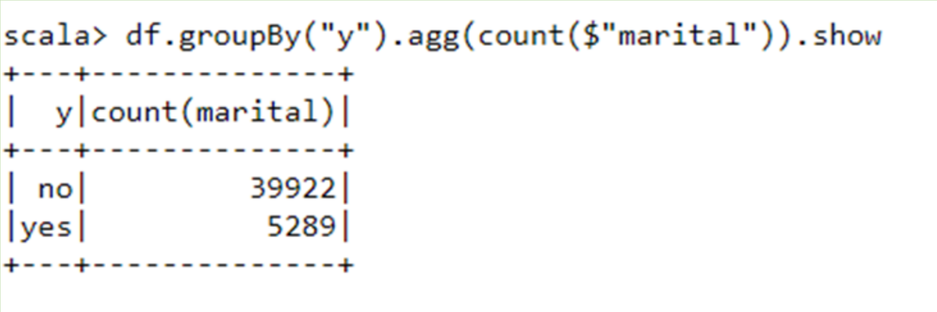
- **Analysis:**

People with age of 40 years and below, are not allowed to do subscriptions.
People with age of 41 years and above, do make subscriptions.

Question six

7. Check if marital status mattered for a subscription to deposit

- **Output:**

-


```
scala> df.groupBy("y").agg(count($"marital")).show
+---+--------------+
|  y|count(marital)|
+---+--------------+
| no|         39922|
|yes|          5289|
+---+--------------+
```

- **Analysis:**

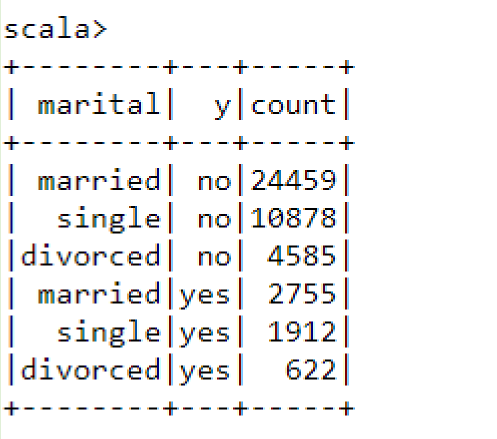
The count of 39922 says No, The marital status does not matter when it comes to subscriptions. While the 5289 says Yes, do matters.

Question seven

7. Check if age and marital status together mattered for a subscription to deposit scheme

- **df.groupBy("marital","y").count().sort($"count".desc).show**

- **Output:**



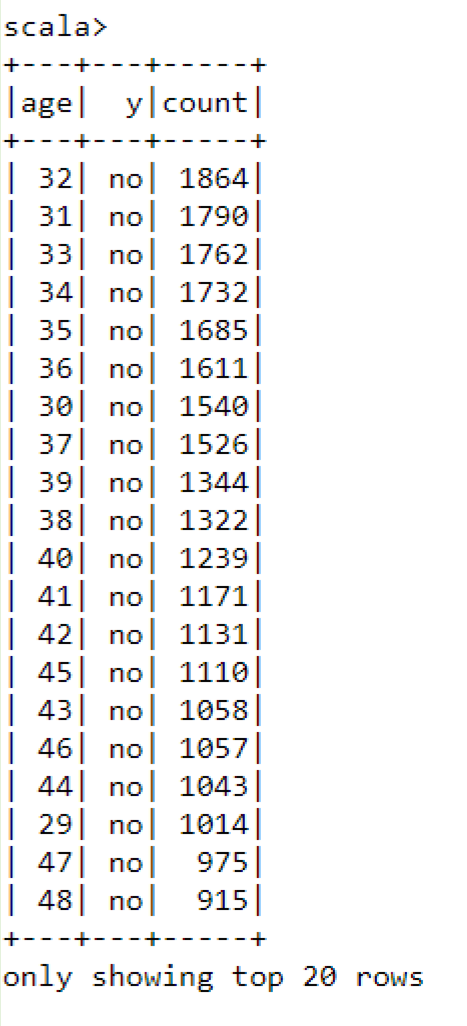
```
scala>
+--------+---+-----+
| marital|  y|count|
+--------+---+-----+
| married| no|24459|
|  single| no|10878|
|divorced| no| 4585|
| married|yes| 2755|
|  single|yes| 1912|
|divorced|yes|  622|
+--------+---+-----+
```

-
- **Analysis:**
- The analysis table above shows clearly that the age and marital status together do not matter when it comes to the subscriptions to deposit of the customers.

Question eight

- **Do feature engineering for the bank and find the right age effect on the campaign.**
- df.groupBy("age","y").count().sort($"count".desc).show

- **Output:**



```
scala>
+---+---+-----+
|age|  y|count|
+---+---+-----+
| 32| no| 1864|
| 31| no| 1790|
| 33| no| 1762|
| 34| no| 1732|
| 35| no| 1685|
| 36| no| 1611|
| 30| no| 1540|
| 37| no| 1526|
| 39| no| 1344|
| 38| no| 1322|
| 40| no| 1239|
| 41| no| 1171|
| 42| no| 1131|
| 45| no| 1110|
| 43| no| 1058|
| 46| no| 1057|
| 44| no| 1043|
| 29| no| 1014|
| 47| no|  975|
| 48| no|  915|
+---+---+-----+
only showing top 20 rows
```

-

- **Analysis:**

The table of analysis shows that there’s no age effect on the campaign