

Project 1: Screenshots

Project (Scala) Internal error: report

Attached: project File View: Code Permissions Run All Clear Results Publish Comments Revision history

```
val lines = sc.textFile("/FileStore/tables/0xajveat1493612980169/bankmarketingdata.csv")
```

lines: org.apache.spark.rdd.RDD[String] = /FileStore/tables/0xajveat1493612980169/bankmarketingdata.csv MapPartitionsRDD[1] at textFile at <console>:48
Command took 0.53 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:09:31 AM on project

```
val bank = lines.map(x => x.split(","))
```

bank: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[2] at map at <console>:50
Command took 0.28 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:09:36 AM on project

```
val bankf = bank.mapPartitionsWithIndex { (idx, iter) => if (idx == 0) iter.drop(1) else iter }
```

bankf: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[3] at mapPartitionsWithIndex at <console>:52
Command took 0.44 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:09:42 AM on project

```
case class Bank(age:Int, job:String, marital:String, education:String, defaults: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)
```

defined class Bank

```
val bankrdd = bankf.map(
  x => Bank(x(0).toInt,
    x(1).replaceAll("\\\\", ""),
    x(2).replaceAll("\\\\", ""),
    x(3).replaceAll("\\\\", ""),
    x(4).replaceAll("\\\\", ""),
    x(5).toInt,
    x(6).replaceAll("\\\\", ""),
    x(7).replaceAll("\\\\", ""),
    x(8).replaceAll("\\\\", ""),
    x(9).toInt,
    x(10).replaceAll("\\\\", ""),
    x(11).toInt,
    x(12).toInt,
    x(13).toInt,
    x(14).toInt,
    x(15).replaceAll("\\\\", ""),
    x(16).replaceAll("\\\\", ""))
)
```

bankrdd: org.apache.spark.rdd.RDD[Bank] = MapPartitionsRDD[4] at map at <console>:56
Command took 0.48 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:09:57 AM on project

```
val bankDF = bankrdd.toDF()
```

bankDF: org.apache.spark.sql.DataFrame = [age: int, job: string ... 15 more fields]
Command took 0.39 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:10:04 AM on project

```
val bankDF = bankrdd.toDF()
```

bankDF: org.apache.spark.sql.DataFrame = [age: int, job: string ... 15 more fields]
Command took 0.90 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:10:04 AM on project

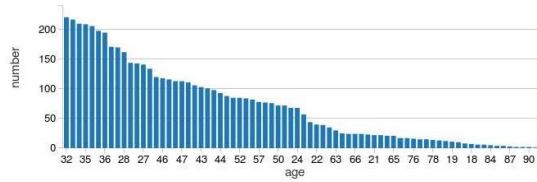
```
bankDF.registerTempTable("bank")
```

<console>:62: warning: method registerTempTable in class Dataset is deprecated: Use createOrReplaceTempView(viewName) instead.
bankDF.registerTempTable("bank")
A

Command took 0.20 seconds -- by sagarnildass@gmail.com at 5/1/2017, 3:10:11 AM on project

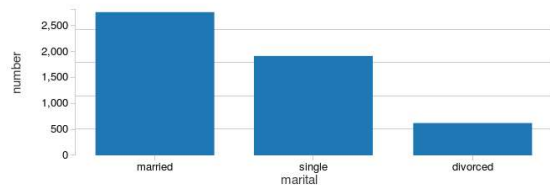
```
val age = sqlContext.sql("select age, count(*) as number from bank where y='yes' group by age order by number desc ")  
display(age)
```

▶ (1) Spark Jobs



```
val marital = sqlContext.sql("select marital, count(*) as number from bank where y='yes' group by marital order by number desc ")  
display(marital)
```

▶ (1) Spark Jobs

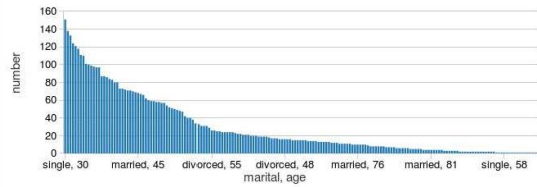


Plot Options...

Command took 1.28 seconds -- by sagarnildass@gmail.com at 4/30/2017, 9:40:30 PM on project

```
val age_marital = sqlContext.sql("select age, marital, count(*) as number from bank where y='yes' group by age,marital order by number desc ")  
display(age_marital)
```

▶ (1) Spark Jobs



Plot Options...

Command took 1.03 seconds -- by sagarnildass@gmail.com at 4/30/2017, 9:57:03 PM on project

```
import scala.reflect.runtime.universe
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.feature.Bucketizer
import org.apache.spark.ml.feature.Normalizer
import org.apache.spark.ml.feature.StringIndexer
import org.apache.spark.ml.feature.VectorAssembler
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics
import org.apache.spark.sql.DataFrame
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions.mean
```

```
val ageRDD = sqlContext.udf.register("ageRDD", (age: Int) => {
  if (age < 20)
    "Teen"
  else if (age > 20 && age <= 32)
    "Young"
  else if (age > 33 && age <= 55)
    "Middle Aged"
  else
    "Old"
})
```

ageRDD: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function1>,StringType,Some(List(IntegerType)))

Command took 0.26 seconds -- by sagarnidass@gmail.com at 5/1/2017, 3:10:39 AM on project

```
val banknewDF = bankDF.withColumn("age", ageRDD(bankDF("age")))
```

banknewDF: org.apache.spark.sql.DataFrame = [age: string, job: string ... 15 more fields]

Command took 0.22 seconds -- by sagarnidass@gmail.com at 5/1/2017, 3:10:45 AM on project

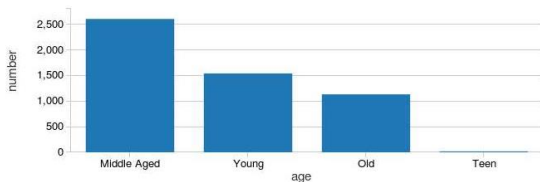
```
banknewDF.registerTempTable("bank_new")
```

<console>:68: warning: method registerTempTable in class Dataset is deprecated: Use createOrReplaceTempView(viewName) instead.
banknewDF.registerTempTable("bank_new")
^

Command took 0.13 seconds -- by sagarnidass@gmail.com at 5/1/2017, 3:11:02 AM on project

```
val age_target = sqlContext.sql("select age, count(*) as number from bank_new where y='yes' group by age order by number desc ")
display(age_target)
```

▶ (1) Spark Jobs



Plot Options...

Command took 1.48 seconds -- by sagarnidass@gmail.com at 5/1/2017, 3:11:57 AM on project

```
val ageInd = new StringIndexer().setInputCol("age").setOutputCol("ageIndex")
```

ageInd: org.apache.spark.ml.feature.StringIndexer = strIdx_bac9b8f8c4da

Command took 0.59 seconds -- by sagarnildass@gmail.com at 5/1/2017, 4:34:04 AM on project

```
var strIndModel = ageInd.fit(banknewDF)
```

▶ (1) Spark Jobs

strIndModel: org.apache.spark.ml.feature.StringIndexerModel = strIdx_bac9b8f8c4da

Command took 2.69 seconds -- by sagarnildass@gmail.com at 5/1/2017, 4:34:27 AM on project

```
strIndModel.transform(banknewDF).select("age", "ageIndex").show(5)
```

▶ (1) Spark Jobs

```
+-----+
|      age|ageIndex|
+-----+
|      Old|      2.0|
|Middle Aged|      0.0|
|      Old|      2.0|
|Middle Aged|      0.0|
|      Old|      2.0|
+-----+
```

only showing top 5 rows

Command took 1.29 seconds -- by sagarnildass@gmail.com at 5/1/2017, 4:34:43 AM on project