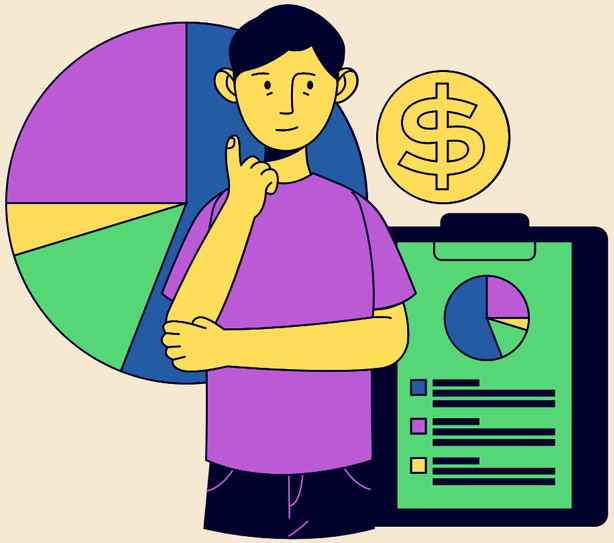
Timmerman Industries

Big data: analyzing spending habits

Introduction

Data preprocessing



1. CHECK FOR NULL
2. ADD NEW COLUMN(YEAR+SPENDING)
3. STRING INDEXER

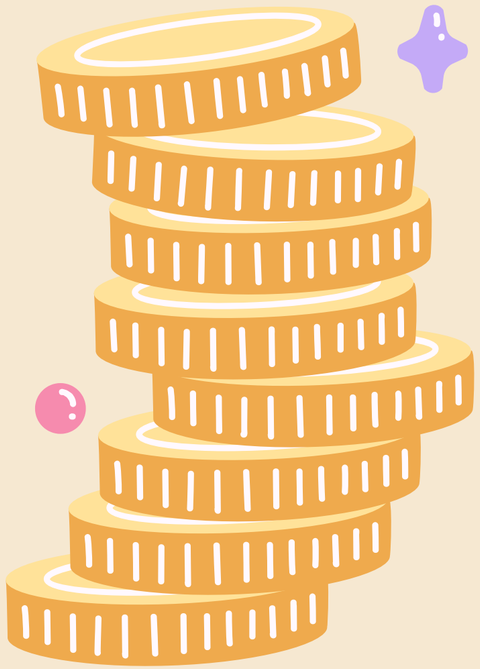
Before

index	City	Date	CardType	ExpType	Gender	Amount
0	Delhi	29-Oct-14	Gold	Bills	F	82475
1	Greater Mumbai	22-Aug-14	Platinum	Bills	F	32555
2	Bengaluru	27-Aug-14	Silver	Bills	F	101738
3	Greater Mumbai	12-Apr-14	Signature	Bills	F	123424
4	Bengaluru	5-May-15	Gold	Bills	F	171574
5	Delhi	8-Sep-14	Silver	Bills	F	100036
6	Delhi	24-Feb-15	Gold	Bills	F	143250
7	Greater Mumbai	26-Jun-14	Platinum	Bills	F	150980
8	Delhi	28-Mar-14	Silver	Bills	F	192247
9	Delhi	1-Sep-14	Platinum	Bills	F	67932
10	Delhi	22-Jun-14	Platinum	Bills	F	280061
11	Greater Mumbai	7-Dec-13	Signature	Bills	F	278036
12	Greater Mumbai	7-Aug-14	Gold	Bills	F	19226
13	Delhi	27-Apr-14	Signature	Bills	F	254359
14	Greater Mumbai	15-Aug-14	Signature	Bills	F	302834
15	Greater Mumbai	28-Nov-14	Platinum	Bills	F	647116
16	Greater Mumbai	14-Jun-14	Signature	Bills	F	421878
17	Greater Mumbai	30-Mar-15	Gold	Bills	F	986379
18	Greater Mumbai	15-Mar-14	Platinum	Bills	F	213047
19	Greater Mumbai	9-Nov-13	Platinum	Bills	F	735566

After

City	CardType	ExpType	Gender	Amount	year	spending	CityIndex	CardIndex	ExpIndex	genderIndex
Delhi	Gold	Bills	F	82475	2014	0	3.0	3.0	2.0	0.0
Greater Mumbai	Platinum	Bills	F	32555	2014	0	1.0	2.0	2.0	0.0
Bengaluru	Silver	Bills	F	101738	2014	0	0.0	0.0	2.0	0.0
Greater Mumbai	Signature	Bills	F	123424	2014	0	1.0	1.0	2.0	0.0
Bengaluru	Gold	Bills	F	171574	2015	1	0.0	3.0	2.0	0.0
Delhi	Silver	Bills	F	100036	2014	0	3.0	0.0	2.0	0.0
Delhi	Gold	Bills	F	143250	2015	0	3.0	3.0	2.0	0.0
Greater Mumbai	Platinum	Bills	F	150980	2014	0	1.0	2.0	2.0	0.0
Delhi	Silver	Bills	F	192247	2014	1	3.0	0.0	2.0	0.0
Delhi	Platinum	Bills	F	67932	2014	0	3.0	2.0	2.0	0.0
Delhi	Platinum	Bills	F	280061	2014	1	3.0	2.0	2.0	0.0
Greater Mumbai	Signature	Bills	F	278036	2013	1	1.0	1.0	2.0	0.0
Greater Mumbai	Gold	Bills	F	19226	2014	0	1.0	3.0	2.0	0.0
Delhi	Signature	Bills	F	254359	2014	1	3.0	1.0	2.0	0.0
Greater Mumbai	Signature	Bills	F	302834	2014	1	1.0	1.0	2.0	0.0
Greater Mumbai	Platinum	Bills	F	647116	2014	1	1.0	2.0	2.0	0.0
Greater Mumbai	Signature	Bills	F	421878	2014	1	1.0	1.0	2.0	0.0
Greater Mumbai	Gold	Bills	F	986379	2015	1	1.0	3.0	2.0	0.0
Greater Mumbai	Platinum	Bills	F	213047	2014	1	1.0	2.0	2.0	0.0
Greater Mumbai	Platinum	Bills	F	735566	2013	1	1.0	2.0	2.0	0.0

RDD operations



1. `CREDIT.TAKEORDERED(1)(ORDERING[INT].REVERSE.ON(X=>X.AMOUNT))`
2. `credit.filter(x=>x.cardType==”Gold”)`
3. `CREDIT.FILTER(_.AMOUNT>8000)`
4. `CREDIT.TAKESAMPLE(TRUE,50)`
5. `RDD[(String,Int)]=credit.map(m=>(m.City,1))`

Sql operations

1. CARDDF.GROUPBY(CARDDF.COL(“GENDER”)).AGG(AVG(“AMOUNT)).SHOW

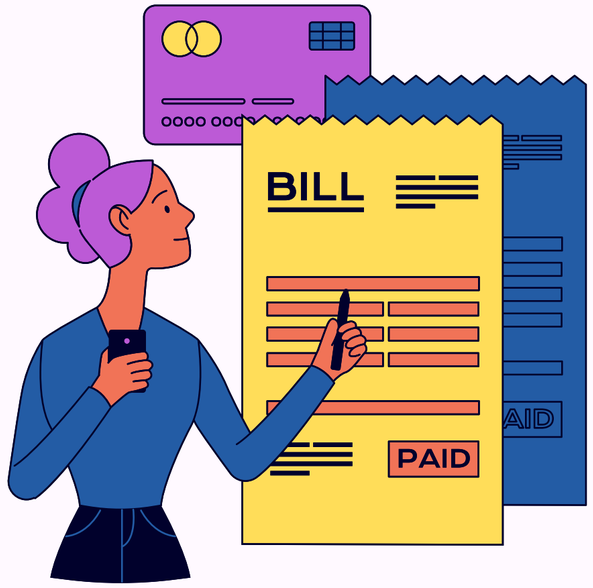
2. cardDF.where($” Amount” <156422).show

3. SPARK.SQL(“SELECT CITY, CARDTYPE FROM CARD WHERE CARDTYPE=’ GOLD’ ”)

4. CARDDF.GROUPBY(CARDDF.COL(“CITY”)).AGG(MIN(“AMOUNT)).SHOW

5. spark.sql(“SELECT Date,Amount FROM card WHERE Amount <6000”)

Machine learning



SELECTED METHOD :
DECISION TREE

input columns were cardIndex, genderIndex , year and ExpIndex

data into training and testing groups using the ratios of 0.7 and 0.3

Results

```
+----------+-----+--------------------+
|prediction|label|            features|
+----------+-----+--------------------+
|       0.0|  1.0|[3.0,1.0,2014.0,2.0]|
|       1.0|  1.0|[0.0,0.0,2013.0,4.0]|
|       0.0|  1.0|[2.0,0.0,2014.0,2.0]|
|       1.0|  1.0|[2.0,0.0,2015.0,2.0]|
|       1.0|  0.0|[0.0,0.0,2015.0,2.0]|
|       0.0|  1.0|[1.0,1.0,2014.0,2.0]|
|       1.0|  0.0|    (4,[2],[2015.0])|
|       1.0|  0.0|[1.0,0.0,2015.0,3.0]|
|       0.0|  0.0|[1.0,1.0,2014.0,0.0]|
|       0.0|  1.0|[0.0,1.0,2014.0,4.0]|
|       1.0|  0.0|[3.0,0.0,2015.0,2.0]|
|       0.0|  0.0|[3.0,0.0,2014.0,2.0]|
|       1.0|  1.0|[3.0,0.0,2015.0,2.0]|
|       1.0|  1.0|[3.0,0.0,2015.0,2.0]|
|       0.0|  1.0|[3.0,0.0,2013.0,2.0]|
|       0.0|  1.0|[3.0,0.0,2014.0,2.0]|
|       0.0|  1.0|[3.0,0.0,2014.0,2.0]|
|       0.0|  1.0|[3.0,0.0,2014.0,2.0]|
|       1.0|  1.0|[3.0,0.0,2015.0,2.0]|
|       0.0|  1.0|[3.0,0.0,2014.0,2.0]|
+----------+-----+--------------------+
```

```
[271] cm = confusion_matrix(y_orig, y_pred)
      print("Confusion Matrix:")
      print(cm)
```

```
Confusion Matrix:
[[1957  717]
 [1762  735]]
```

```
[264] print("Test Error = %g " % (1.0 - accuracy))
```

```
Test Error = 0.479404
```