databrickshw2

Spark Assignment 2

Fubang ZHAO

```
from pyspark import SparkContext, SparkConf
from pyspark.sql import Row
from pyspark.sql import functions as F


sc = SparkContext.getOrCreate()
rdd = sc.wholeTextFiles("FileStore/tables/kl5f55dz1509040574264")
f1 = rdd.map(lambda kv: (kv[0].split("/")[-1].split(".txt")[0], kv[1]))
f2 = f1.flatMapValues(lambda v: v.split("\r\n"))
f3 = f2.map(lambda kv: (kv[0].split("_")[0], kv[0], kv[1].split(" ")[0],
kv[1].split(" ")[1]))# (city, city&num, month, revenue)


data = f3.map(lambda p: Row(city=p[0], store=p[1], month=p[2],
income=int(p[3])))

# Infer the schema, and register the DataFrame as a table.
revenue = spark.createDataFrame(data)
revenue.createOrReplaceTempView("Revenue")


spark.sql("select * from Revenue").show() #The dataframe that we got from the
data
```

```
+-----+------+-----+-----+
| city|income|month|store|
+-----+------+-----+-----+
|anger|    13|  JAN|anger|
|anger|    12|  FEB|anger|
|anger|    14|  MAR|anger|
|anger|    15|  APR|anger|
|anger|    12|  MAY|anger|
|anger|    15|  JUN|anger|
|anger|    19|  JUL|anger|
|anger|    15|  AUG|anger|
|anger|    13|  SEP|anger|
|anger|     8|  OCT|anger|
|anger|    14|  NOV|anger|
|anger|    16|  DEC|anger|
| lyon|    13|  JAN| lyon|
| lyon|    12|  FEB| lyon|
| lyon|    14|  MAR| lyon|
| lyon|    15|  APR| lyon|
| lyon|    12|  MAY| lyon|
```

```
| lyon|    15|  JUN| lyon|
| lyon|    19|  JUL| lyon|
| lyon|    25|  AUG| lyon|
+-----+------+-----+-----+
only showing top 20 rows
```

The First question:

Average monthly income of the shop in France (on 1 year data)

```
revenue.groupby("month").avg().show()
```

```
+-----+------------------+
|month|       avg(income)|
+-----+------------------+
|  APR| 20.23076923076923|
|  OCT| 26.53846153846154|
|  NOV| 24.53846153846154|
|  FEB|19.153846153846153|
|  SEP| 25.53846153846154|
|  JAN| 20.76923076923077|
|  AUG|23.076923076923077|
|  MAR| 17.53846153846154|
|  DEC|              29.0|
|  JUN|27.846153846153847|
|  JUL|21.692307692307693|
|  MAY| 22.46153846153846|
+-----+------------------+
```

The second question:

Total revenue per city per year

```
revenue.groupby("city").sum().show()
```

```
+----------+-----------+
|      city|sum(income)|
+----------+-----------+
|    nantes|        207|
|    troyes|        214|
|     paris|       1568|
|      lyon|        193|
|     anger|        166|
|marseilles|        515|
|      nice|        203|
|    orlean|        196|
|    rennes|        180|
|  toulouse|        177|
```

```
+----------+-----------+
```

The third question:

Average monthly income of the shop in each city

```
revenue.groupby("city", "month").avg().show()
```

```
+----------+-----+-----------+
|      city|month|avg(income)|
+----------+-----+-----------+
|    troyes|  JUN|       25.0|
|      nice|  MAY|       11.0|
|    rennes|  DEC|       20.0|
|  toulouse|  APR|       11.0|
|marseilles|  JAN|       16.0|
|      lyon|  MAR|       14.0|
|    orlean|  MAY|       12.0|
|      lyon|  APR|       15.0|
|      nice|  JUL|       19.0|
|     anger|  JAN|       13.0|
|     paris|  JUN|       55.0|
|  toulouse|  SEP|       23.0|
|     anger|  MAY|       12.0|
|      lyon|  JUL|       19.0|
|    orlean|  AUG|       25.0|
|      lyon|  AUG|       25.0|
|    rennes|  FEB|       18.0|
|marseilles|  JUL|       21.0|
|      nice|  DEC|       29.0|
|marseilles|  JUN|       25.0|
+----------+-----+-----------+
only showing top 20 rows
```

The 4th question:

Total revenue per store per year

```
revenue.groupby("store").sum().show()
```

```
+------------+-----------+
|       store|sum(income)|
+------------+-----------+
|      nantes|        207|
|      troyes|        214|
|        lyon|        193|
|marseilles_1|        284|
|     paris_2|        642|
```

```
|       anger|        166|
|     paris_3|        330|
|marseilles_2|        231|
|        nice|        203|
|      orlean|        196|
|      rennes|        180|
|     paris_1|        596|
|    toulouse|        177|
+------------+-----------+
```

The 5th question:

The store that achieves the best performance in each month

```
max_income = revenue.groupby("month").max().alias("max")
table_new = revenue.join(max_income,
revenue.month==max_income.month).select(revenue.store, revenue.income,
revenue.month, max_income["max(income)"]).collect()
df_new = spark.createDataFrame(table_new)
df_new.where(df_new["income"] == df_new["max(income)"]).select(df_new.month,
df_new.store, df_new.income).show()
```

```
+-----+-------+------+
|month|  store|income|
+-----+-------+------+
|  APR|paris_1|    57|
|  OCT|paris_1|    68|
|  NOV|paris_2|    64|
|  FEB|paris_2|    42|
|  SEP|paris_2|    63|
|  JAN|paris_1|    51|
|  AUG|paris_2|    45|
|  MAR|paris_2|    44|
|  DEC|paris_1|    71|
|  JUN|paris_2|    85|
|  JUL|paris_1|    61|
|  MAY|paris_2|    72|
+-----+-------+------+
```