Transforming and Appending the day level dataset to the historical dataset for the dashboard

Reading the different daily datasets from S3

Importing and creating the SQL context

In [1]:
```
from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)
```

Importing the various functions for further usage.

In [2]:
```
from pyspark.sql.types import StructType, StringType
from pyspark import SparkConf, SparkContext
from pyspark import sql
from pyspark.sql.functions import lit
import numpy as np
import pandas as pd
```

Declaring a empty struct to later create an empty dataframe.

In [3]:
```
schema = StructType([])
```

Creating an empty dataframe to hold the graph numbers to include in the historical dataset to use in the dashboard.

In [4]:
```
empty = sqlContext.createDataFrame(sc.emptyRDD(), schema)
```

Declaring an array with the graph numbers

In [5]:
```
graph = np.array([1,2,3,4,5])
```

Creating the dataframe for the above created array

In [6]:
```
graphdf = pd.DataFrame(graph, columns = ['Graph'])
```

Registering the graph dataframe as a temp table to be used later for joining with historical datasets

In [7]:
```
graphDF = sqlContext.createDataFrame(graphdf)
```

In [8]:
```
graphDF.registerTempTable("grph_tbl")
```

In [9]:
```
graphDF.show()
```

```
+-----+
|Graph|
+-----+
|    1|
|    2|
|    3|
|    4|
|    5|
+-----+
```

Reading and formatting the different daily datasets to be appended to the historical data and then used in the dashboard

Reading the daily top bottom 5 cities dataset from S3 bucket

In [10]:
```
Dcity = sqlContext.read.format('com.databricks.spark.csv') \
    .option("inferSchema",True).option("header",False).load('s3://bigdataprjct/data_today/DATE_LVL_TOP_BOTTOM_CITIES')
```

Transforming the read city dataframe into RDD

In [11]:
```
DcityDataRDD = Dcity.rdd
```

Caching the city RDD for faster processing and avoid reading from the bucket everytime

In [12]:
```
DcityDataRDD.cache()
```

Out[12]:
```
MapPartitionsRDD[23] at javaToPython at NativeMethodAccessorImpl.java:0
```

Converting the city data RDD into a dataframe and providing the column names to the various columns as to the requirement for dashboard

In [13]:
```
DcityDataDF = DcityDataRDD.toDF(['Date_chr1','City1','CitySales1','TopBottom1','Rank1'])
```

Adding column for Graph# according to the Graph that will be created using this portion of the data and joining with the temp graph table created above.

In [14]:
```
DcityDataDF = DcityDataDF.withColumn('Graph',lit(1))
```

Registering the above created city dataframe as a temp table to join with the Graph table created earlier

In [15]:
```
DcityDataDF.registerTempTable("Dcity_g1")
```

Viewing the schema to check if it looks correct

In [16]:
```
DcityDataDF.printSchema()
```

```
root
 |-- Date_chr1: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: long (nullable = true)
 |-- Graph: integer (nullable = false)
```

Checking the data in the city dataframe

In [17]:
```
DcityDataDF.show()
```

```
+---------+------------+------------------+----------+-----+-----+
|Date_chr1|       City1|        CitySales1|TopBottom1|Rank1|Graph|
+---------+------------+------------------+----------+-----+-----+
|8/15/2017|      Cuenca| 35793.97299999999|       top|    3|    1|
|8/15/2017|    Guaranda| 9282.186999999998|    bottom|    5|    1|
|8/15/2017|   Guayaquil|101063.80600000006|       top|    2|    1|
|8/15/2017|      Ibarra| 7946.430999999999|    bottom|    4|    1|
|8/15/2017|     Machala| 28950.65900000001|       top|    5|    1|
|8/15/2017|      Playas| 5371.156000000001|    bottom|    1|    1|
|8/15/2017|        Puyo| 6917.787999999998|    bottom|    3|    1|
|8/15/2017|       Quito|341655.35799999995|       top|    1|    1|
|8/15/2017|     Salinas| 6522.787999999999|    bottom|    2|    1|
|8/15/2017|Santo Domingo| 30309.08099999999|       top|    4|    1|
+---------+------------+------------------+----------+-----+-----+
```

Joining the above formed city table with graph table

In [18]:
```
Dg1_data = sqlContext.sql("""
    SELECT gr.*,c1.Date_chr1, c1.City1, c1.CitySales1, c1.TopBottom1, c
1.Rank1
    from grph_tbl gr
    LEFT JOIN Dcity_g1 c1
```

```
    ON gr.Graph = c1.Graph
""")

Dg1_data.printSchema()
```

```
root
 |-- Graph: long (nullable = true)
 |-- Date_chr1: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: long (nullable = true)
```

In [19]:
```
Dg1_data.show()
```

```
+-----+---------+-------------+------------------+----------+-----+
|Graph|Date_chr1|        City1|        CitySales1|TopBottom1|Rank1|
+-----+---------+-------------+------------------+----------+-----+
|    5|     null|         null|              null|      null| null|
|    1|8/15/2017|       Cuenca| 35793.97299999999|       top|    3|
|    1|8/15/2017|     Guaranda| 9282.186999999998|    bottom|    5|
|    1|8/15/2017|    Guayaquil|101063.80600000006|       top|    2|
|    1|8/15/2017|       Ibarra| 7946.430999999999|    bottom|    4|
|    1|8/15/2017|      Machala| 28950.65900000001|       top|    5|
|    1|8/15/2017|       Playas| 5371.156000000001|    bottom|    1|
|    1|8/15/2017|         Puyo| 6917.787999999998|    bottom|    3|
|    1|8/15/2017|        Quito|341655.35799999995|       top|    1|
|    1|8/15/2017|      Salinas| 6522.787999999999|    bottom|    2|
|    1|8/15/2017|Santo Domingo| 30309.08099999999|       top|    4|
|    3|     null|         null|              null|      null| null|
|    2|     null|         null|              null|      null| null|
|    4|     null|         null|              null|      null| null|
+-----+---------+-------------+------------------+----------+-----+
```

Registering the above dataframe as a temp table to be joined later with other data parts.

```

```

In [20]:
```
Dg1_data.registerTempTable("Dg1_c")
```

Reading the day level items data for append and use in Graph 2

Reading the days level data for top and bottom items from S3 bucket for further processing

In [21]:
```
Ditem = sqlContext.read.format('com.databricks.spark.csv') \
    .option("inferSchema",True).option("header",False).load('s3://bigdataprjct/data_today/DATE_LVL_TOP_BOTTOM_ITEMS')
```

Converting the iteam dataset read into RDD for further processsing

In [22]:
```
DitemDataRDD = Ditem.rdd
```

In [23]:
```
DitemDataRDD.cache()
```

Out[23]:
```
MapPartitionsRDD[55] at javaToPython at NativeMethodAccessorImpl.java:0
```

Converting the RDD to dataframe and renaming the column as per the requirement for append and dashboard

In [24]:
```
DitemDataDF = DitemDataRDD.toDF(['Date_chr2','Family2','ItemSales2','TopBottom','Rank2'])
```

Adding the column Graph number and declaring all row values in the column to be 2 since item level graph in dashboard in number 2 and will further be joined with the above created final city dataset.

In [25]:
```
DitemDataDF = DitemDataDF.withColumn('Graph',lit(2))
```

In [26]:
```
DitemDataDF.registerTempTable("Ditem_g2")
```

Joining the above created item table with the final city dataset created earlier which also has the graph number data

In [27]:
```
Dg2_data = sqlContext.sql("""
    SELECT g1.*, ig.Date_chr2, ig.Family2, ig.ItemSales2, ig.TopBottom,
 ig.Rank2
    from Dg1_c g1
    LEFT JOIN Ditem_g2 ig
    ON g1.Graph = ig.Graph
""")

Dg2_data.printSchema()
```

```
root
 |-- Graph: long (nullable = true)
 |-- Date_chr1: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: long (nullable = true)
 |-- Date_chr2: string (nullable = true)
 |-- Family2: string (nullable = true)
 |-- ItemSales2: double (nullable = true)
 |-- TopBottom: string (nullable = true)
 |-- Rank2: long (nullable = true)
```

In [28]:
```
Dg2_data.show()
```

```
+-----+---------+-------------+------------------+----------+-----+----
-----+----------+----------+---------+-----+
|Graph|Date_chr1|        City1|        CitySales1|TopBottom1|Rank1|Date
_chr2|   Family2|ItemSales2|TopBottom|Rank2|
+-----+---------+-------------+------------------+----------+-----+----
-----+----------+----------+---------+-----+
|    5|     null|         null|              null|      null| null|
null|      null|      null|     null| null|
|    1|8/15/2017|       Cuenca| 35793.97299999999|       top|    3|
```

```
null|      null|      null|     null| null|
|    1|8/15/2017|     Guaranda| 9282.186999999998|    bottom|    5|
null|      null|      null|     null| null|
|    1|8/15/2017|    Guayaquil|101063.80600000006|       top|    2|
null|      null|      null|     null| null|
|    1|8/15/2017|       Ibarra| 7946.430999999999|    bottom|    4|
null|      null|      null|     null| null|
|    1|8/15/2017|      Machala| 28950.65900000001|       top|    5|
null|      null|      null|     null| null|
|    1|8/15/2017|       Playas| 5371.156000000001|    bottom|    1|
null|      null|      null|     null| null|
|    1|8/15/2017|         Puyo| 6917.787999999998|    bottom|    3|
null|      null|      null|     null| null|
|    1|8/15/2017|        Quito|341655.35799999995|       top|    1|
null|      null|      null|     null| null|
|    1|8/15/2017|      Salinas| 6522.787999999999|    bottom|    2|
null|      null|      null|     null| null|
|    1|8/15/2017|Santo Domingo| 30309.08099999999|       top|    4|
null|      null|      null|     null| null|
|    3|      null|         null|              null|      null| null|
null|      null|      null|     null| null|
|    2|      null|         null|              null|      null| null|8/1
5/2017|AUTOMOTIVE|     337.0|   bottom|    4|
|    2|      null|         null|              null|      null| null|8/1
5/2017| BABY CARE|       8.0|   bottom|    2|
|    2|      null|         null|              null|      null| null|8/1
5/2017|    BEAUTY|     339.0|   bottom|    5|
|    2|      null|         null|              null|      null| null|8/1
5/2017| BEVERAGES|  170773.0|      top|    2|
|    2|      null|         null|              null|      null| null|8/1
5/2017|  CLEANING|   58474.0|      top|    4|
|    2|      null|         null|              null|      null| null|8/1
5/2017|     DAIRY|   40707.0|      top|    5|
|    2|      null|         null|              null|      null| null|8/1
5/2017| GROCERY I|224208.125|      top|    1|
|    2|      null|         null|              null|      null| null|8/1
5/2017|  HARDWARE|      57.0|   bottom|    3|
+-----+---------+-------------+------------------+----------+-----+----
-----+----------+----------+---------+-----+
```

```
only showing top 20 rows
```

In [29]:
```
Dg2_data.registerTempTable("Dg2_ci")
```

Reading the daily store level data

Reading the daily store level data from AWS S3 bucket

In [30]:
```
Dstore = sqlContext.read.format('com.databricks.spark.csv') \
    .option("inferSchema",True).option("header",False).load('s3://bigda
taprjct/data_today/DATE_LVL_TOP_BOTTOM_STORES')
```

In [31]:
```
DstoreDataRDD = Dstore.rdd
```

In [32]:
```
DstoreDataRDD.cache()
```

Out[32]:
```
MapPartitionsRDD[91] at javaToPython at NativeMethodAccessorImpl.java:0
```

Converting the RDD to dataframe and renaming the column so that it can be used in the dashboard with ease

In [33]:
```
DstoreDataDF =
DstoreDataRDD.toDF(['Date_chr3','StoreNbr3','ItemSales3','TopBottom3','I
k3'])
```

Adding the Graph number to the store dataframe and populating the column with numeric 3 since the graph number for store level data is 3.

In [34]:
```
DstoreDataDF = DstoreDataDF.withColumn('Graph',lit(3))
```

In [35]:
```
DstoreDataDF.registerTempTable("Dstore_g3")
```

Joining the store data to the earlier created city and item dataset based on the graph number

In [36]:
```
Dg3_data = sqlContext.sql("""
    SELECT g1.*, sg.Date_chr3, sg.StoreNbr3, sg.ItemSales3, sg.TopBotto
m3, sg.Rank3
    from Dg2_ci g1
    LEFT JOIN Dstore_g3 sg
    ON g1.Graph = sg.Graph
""")

Dg3_data.printSchema()
```

```
root
 |-- Graph: long (nullable = true)
 |-- Date_chr1: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: long (nullable = true)
 |-- Date_chr2: string (nullable = true)
 |-- Family2: string (nullable = true)
 |-- ItemSales2: double (nullable = true)
 |-- TopBottom: string (nullable = true)
 |-- Rank2: long (nullable = true)
 |-- Date_chr3: string (nullable = true)
 |-- StoreNbr3: long (nullable = true)
 |-- ItemSales3: double (nullable = true)
 |-- TopBottom3: string (nullable = true)
 |-- Rank3: long (nullable = true)
```

In [37]:
```
Dg3_data.show()
```

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

```
|Graph|Date_chr1|        City1|        CitySales1|TopBottom1|Rank1|Date
_chr2|Family2|ItemSales2|TopBottom|Rank2|Date_chr3|StoreNbr3|        It
emSales3|TopBottom3|Rank3|
+-----+---------+-------------+------------------+----------+-----+----
-----+-------+----------+---------+-----+---------+---------+----------
--------+----------+-----+
|    5|     null|         null|              null|      null| null|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|       Cuenca| 35793.97299999999|       top|    3|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|     Guaranda| 9282.186999999998|    bottom|    5|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|    Guayaquil|101063.80600000006|       top|    2|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|       Ibarra| 7946.430999999999|    bottom|    4|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|      Machala| 28950.65900000001|       top|    5|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|       Playas| 5371.156000000001|    bottom|    1|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|         Puyo| 6917.787999999998|    bottom|    3|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|        Quito|341655.35799999995|       top|    1|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|      Salinas| 6522.787999999999|    bottom|    2|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
|    1|8/15/2017|Santo Domingo| 30309.08099999999|       top|    4|
null|   null|      null|     null| null|     null|     null|
   null|      null| null|
```

```
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       13| 6301.05099
9999999|    bottom|    3|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       25| 6522.78799
9999999|    bottom|    5|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       26|3694.897000
0000018|    bottom|    1|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|        3|30185.60200
0000006|       top|    4|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       32| 6504.91200
0000002|    bottom|    4|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       35| 5371.15600
0000001|    bottom|    2|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       44|33141.32199
9999986|       top|    1|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       45| 31562.9260
0000001|       top|    3|
|    3|     null|         null|               null|      null| null|
null|   null|      null|     null| null|8/15/2017|       47|31653.69100
0000006|       top|    2|
+-----+---------+-------------+-------------------+----------+-----+----
-----+-------+----------+---------+-----+---------+---------+----------
--------+----------+-----+
only showing top 20 rows
```

In [38]:
```
Dg3_data.registerTempTable("Dg3Final")
```

In [39]:
```
Dg3_data.registerTempTable('g3_cis')
```

Reading the day level transaction data for 4th dashboard

Reading the day level transaction data from the S3 bucket to join with previously created dataset

In [41]:
```
Ddate = sqlContext.read.format('com.databricks.spark.csv') \
    .option("inferSchema",True).option("header",False).load('s3://bigdataprjct/data_today/DATE_LVL_TODAY')
```

In [42]:
```
DdateDataRDD = Ddate.rdd
```

In [43]:
```
DdateDataRDD.cache()
```

Out[43]:
```
MapPartitionsRDD[142] at javaToPython at NativeMethodAccessorImpl.java:0
```

Converting the RDD into dataframe and providing it column names according to the dashboarding requirements.

In [49]:
```
DdateDataDF = DdateDataRDD.toDF(['Date_chr4','StoreNbr4','Item4','Sales4','Dcoil4','TrnsCount4'])
```

Adding the Graph number column to the above dataframe and populating the rows with numeric value 4 since the day level transaction data will be used for creating the 4th graph in the dashbaord.

In [50]:
```
DdateDataDF = DdateDataDF.withColumn('Graph',lit(4))
```

In [51]:
```
DdateDataDF.show()
```

```
+---------+---------+-----+----------+------+----------+-----+
|Date_chr4|StoreNbr4|Item4|    Sales4|Dcoil4|TrnsCount4|Graph|
+---------+---------+-----+----------+------+----------+-----+
|8/15/2017|       54| 3771|762661.936| 47.57|     86561|    4|
```

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

In [52]:
```
DdateDataDF.registerTempTable("date_g4")
```

Joining the day level data for Graph 4 with the previous data created for Graph 1, 2 and 3.

In [53]:
```
dtlvljoin = sqlContext.sql("""
        SELECT g1.*, dtl.Date_chr4, dtl.Sales4
    from g3_cis g1
    LEFT JOIN date_g4 dtl
    ON g1.Graph = dtl.Graph
        """)

dtlvljoin.printSchema()
```

```
root
 |-- Graph: long (nullable = true)
 |-- Date_chr1: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: long (nullable = true)
 |-- Date_chr2: string (nullable = true)
 |-- Family2: string (nullable = true)
 |-- ItemSales2: double (nullable = true)
 |-- TopBottom: string (nullable = true)
 |-- Rank2: long (nullable = true)
 |-- Date_chr3: string (nullable = true)
 |-- StoreNbr3: long (nullable = true)
 |-- ItemSales3: double (nullable = true)
 |-- TopBottom3: string (nullable = true)
 |-- Rank3: long (nullable = true)
 |-- Date_chr4: string (nullable = true)
 |-- Sales4: double (nullable = true)
```

In [54]:
```
dtlvljoin.registerTempTable("g4F")
```

Transforming the above dataset so to contain only necessary columns and appending to the historical dataset

In [61]:
```
Dg3F = sqlContext.sql("""
        SELECT Graph, COALESCE(Date_chr1, Date_chr2, Date_chr3, Date_ch
r4) AS Date,
        COALESCE(Date_chr1, Date_chr2, Date_chr3, Date_chr4) AS Date_ch
r,
        City1, CitySales1, TopBottom1, Rank1,
        Family2, ItemSales2, TopBottom, Rank2,
        StoreNbr3, ItemSales3, TopBottom3, Rank3,
        Sales4, "Term5", "Variable5", "Value5"
        FROM g4F
        """)
```

In [62]:
```
Dg3F.show()
```

```
+-----+---------+---------+-------------+------------------+----------+
-----+-------+----------+---------+-----+---------+------------------+-
---------+-----+------+-----+---------+------+
|Graph|     Date| Date_chr|        City1|        CitySales1|TopBottom1|
Rank1|Family2|ItemSales2|TopBottom|Rank2|StoreNbr3|        ItemSales3|T
opBottom3|Rank3|Sales4|Term5|Variable5|Value5|
+-----+---------+---------+-------------+------------------+----------+
-----+-------+----------+---------+-----+---------+------------------+-
---------+-----+------+-----+---------+------+
|    5|     null|     null|         null|              null|      null|
null|   null|      null|     null| null|     null|              null|
    null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|       Cuenca| 35793.97299999999|       top|
    3|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|     Guaranda| 9282.186999999998|    bottom|
    5|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|    Guayaquil|101063.80600000006|       top|
    2|   null|      null|     null| null|     null|              null|
```

```
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|       Ibarra| 7946.430999999999|    bottom|
    4|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|      Machala| 28950.65900000001|       top|
    5|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|       Playas| 5371.156000000001|    bottom|
    1|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|         Puyo| 6917.787999999998|    bottom|
    3|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|        Quito|341655.35799999995|       top|
    1|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|      Salinas| 6522.787999999999|    bottom|
    2|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    1|8/15/2017|8/15/2017|Santo Domingo| 30309.08099999999|       top|
    4|   null|      null|     null| null|     null|              null|
     null| null|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       13| 6301.050999999999|
  bottom|    3|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       25| 6522.787999999999|
  bottom|    5|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       26|3694.8970000000018|
  bottom|    1|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|        3|30185.602000000006|
     top|    4|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       32| 6504.912000000002|
  bottom|    4|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       35| 5371.156000000001|
```

```
  bottom|    2|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       44|33141.321999999986|
     top|    1|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       45| 31562.92600000001|
     top|    3|  null|Term5|Variable5|Value5|
|    3|8/15/2017|8/15/2017|         null|              null|      null|
null|   null|      null|     null| null|       47|31653.691000000006|
     top|    2|  null|Term5|Variable5|Value5|
+-----+---------+---------+-------------+------------------+----------+
-----+-------+----------+---------+-----+---------+------------------+-
---------+-----+------+-----+---------+------+
only showing top 20 rows
```

In [63]:
```
Dg3F.registerTempTable("gDaily")
```

Dump all the unnecessary columns

In [64]:
```
from functools import reduce
from pyspark.sql import DataFrame

daily = reduce(DataFrame.drop, ['Date_chr1','Date_chr2','Date_chr3','Da
te_chr4'], Dg3F)
```

Reading the historical dataset to append it to the daily one

Read the historical dataset created earlier from the S3 bucket

In [65]:
```
hist = sqlContext.read.format('com.databricks.spark.csv') \
    .option("inferSchema",True).option("header",True).\
    load('s3n://bigdataprjct/Miscellaneous/historical')
```

In [66]:
```
hist.printSchema()
```

```
root
 |-- Graph: integer (nullable = true)
 |-- Date: string (nullable = true)
 |-- Date_chr: string (nullable = true)
 |-- City1: string (nullable = true)
 |-- CitySales1: double (nullable = true)
 |-- TopBottom1: string (nullable = true)
 |-- Rank1: integer (nullable = true)
 |-- Family2: string (nullable = true)
 |-- ItemSales2: double (nullable = true)
 |-- TopBottom: string (nullable = true)
 |-- Rank2: integer (nullable = true)
 |-- StoreNbr3: integer (nullable = true)
 |-- ItemSales3: double (nullable = true)
 |-- TopBottom3: string (nullable = true)
 |-- Rank3: integer (nullable = true)
 |-- Sales4: double (nullable = true)
 |-- Term5: string (nullable = true)
 |-- Variable5: string (nullable = true)
 |-- Value5: double (nullable = true)
```

In [69]:
```
hist.registerTempTable("hs")
```

Union the historical and day level data to create the final set

In [67]:
```
Final = hist.union(daily)
```

In [68]:
```
Final.printSchema()
```

```
root
 |-- Graph: long (nullable = true)
 |-- Date: string (nullable = true)
 |-- Date_chr: string (nullable = true)
 |-- City1: string (nullable = true)
```