Basic Retail Analysis in 1010data
Contents

Retail Sales Analysis ................................................................................................................................. 3
Most Profitable Weekday in Month ........................................................................................................... 7
Retail Sales Analysis

The 1010data Quick Start Guide breaks down basic operations in 1010data and how each can individually be used to perform a basic analysis. If you've read that guide you should have a basic idea on how to do things like select row, create metrics with computed columns, and summarize data sets with a tabulation or cross-tabulation. This guide will walk you through how to take those atomized operations and combine them to produce meaningful insights and analysis. The first project we'll look at is how to perform a basic market basket analysis.

Market basket analysis is broad term for the kinds of analyses used to better understand sales patterns and shopper behavior. Like any good analysis, this one starts with a simple questions: How much, in terms of dollars, am I selling on each day of a given month? In this instance, we want to look at a month's worth of data, and then summarize the whole month with a day by day breakdown of sales in dollars. The first step is to specify the actual month we are interested in. But first, let's take a quick look at the data we're going to analyze.

If you read the 1010data Quick Start Guide this might look familiar to you. This data contains all the same columns as the very small (35 rows) data set that was used in that tutorial. However, in this version, you can see that the number of rows is much higher (3.3 billion!). Regardless of the size of the table, both these data sets share basic information. They contain the transaction number, date of the transaction, the SKU (product identification number) for each item purchased, and the date of the transaction, among others. Now that we understand the data a little better, we're going to revisit out central question: How much in sales is my organization doing, day by day, in a month?

Since we are primarily interested in a single month’s worth of data, the first step is to narrow down the rows in the table to only those that take place during the month we’re interested in. If you’d like to use the GUI for this, feel free. But part of this guide is helping you get comfortable with the 1010data Macro Language. So we'll provide both the GUI and Macro Language way of doing things for the first few tutorials. But don't be shocked if at some point we're only working with the Macro Language. Here's our date selection in the GUI:
As you can see, we're going to look at sales for the organization in the month of January, 2011. Here's what the Macro Language code looks like:

```
<sel value="between(date;20110101;20110131)"/>
```

As you may have noticed, we're using the `between(X;Y;Z)` function to select our date range. `between(X;Y;Z)` is great for date selections, and has many other useful purposes. Once the selection is made, we'll go from our initial 3.3 billion rows of data to a more manageable 85 million, as shown below:

```
After the row selection executes, you should see that every remaining row (more than 85 million) falls within the date range specified. Now we can see every item sold in every basket for the entire month. But remember, we're interested in total sales dollars for each day of the month, and don't particularly care about specific items or transactions (for now). If you remember the basics of performing a tabulation, then it should be fairly clear that we want to group by the date and view the sum of sales for each day. Let's take a look at the tabulation settings. Go to Analysis > Tabulation.... Here are our requirements for the tabulation:

- Group by the date
- View the sum of sales
- View the total costs for the items sold on each day
- View the number of transactions on each day

Let's take a look at the Tabulation... dialog to understand how to create the aggregations we're interested in. Then, we'll quickly run through the Macro Language code. Here's the dialog:
Now that we’ve defined our grouping metric and the summarizations we’re interested in, click Submit to get the results:
As you can see, we've produced a tidy summarization that clearly lists sales totals for each date within January, 2011. And just in case you're interested, here's a nicely commented Macro Language version of this analysis:

```
<note>First step is to select the month (or other time period) of interest</note>
<sel value="between(date;20110101;20110131)"/>

<note>Next, create a tabulation with the tabu element. The breaks attribute specifies the column to group by.</note>
<tabu label="Monthly Sales by Date" breaks="date">

  <note>The break element tells the system how to sort the values in the group-by column</note>
  <break col="date" sort="up"/>

  <note>This tcol element is the first summarization in the tabulation. It produces the sum of extended sales</note>
  <tcol name="sumsales" source="xsales" fun="sum" label="Sum of Extended Sales"/>

  <note>This tcol creates a sum of the total cost for each date</note>
  <tcol name="sumcost" source="cost" fun="sum" label="Sum of Cost"/>

  <note>This tcol counts the number of unique transaction IDs for each date</note>
  <tcol name="numtransid" source="transid" fun="count" label="Num of Unique Values in Trans ID"/>
</tabu>
```
Basic Retail Analysis in 1010data | Retail Sales Analysis | 7

And just in case you don't care about the comments, here's the code the system produces (or which a savvy 1010data user might write):

```xml
<tabu label="Monthly Sales by Date" breaks="date">
  <break col="date" sort="up"/>
  <tcol name="sumsales" source="xsales" fun="sum" label="Sum of Extended Sales"/>
  <tcol name="sumcost" source="cost" fun="sum" label="Sum of Cost"/>
  <tcol name="transucnt" source="transid" fun="ucnt" label="Num of Unique Values in Trans ID"/>
</tabu>
```

Producing a basic summary of sales by date for a given time period is a very common way of measuring retail sales performance. In 1010data it can be performed on truly massive data sets in a matter for a couple minutes. Try your hand and playing with these values and altering the analysis to answer your own questions. For instance, as a follow up, try to determine the margin for each day of the month.

Next up, we'll take this analysis one step further by determining what the most profitable weekday of the month is.

**Most Profitable Weekday in Month**

The next step in our basic analysis of sales data is to determine what the most profitable weekday was for a given month (or whatever period of time you prefer to evaluate). We'll start with the tabulation we have already produced, which give us totals for sales and cost for each date in January, 2011, as follows:

![Monthly Sales by Date Table]

Our next objective is to determine what week day (Sunday - Saturday) was the most profitable overall. To do this we only need to perform three basic steps:

1. Create a computed column that assigns a week day for each date in the table
2. Create a tabulation that groups by week day and summarizes sales and cost for each one
3. Create a computed column that calculates margin (aka profit!)

Step 1: Assign a week day to each date. While we could do this by going to **Columns > Create Computed Column...**, we're actually going to do everything for this part of the analysis in the Macro Language. Start with the code we generated last time and then add the last line in bold:
<sel value="between(date;20110101;20110131)="/>
<tabu label="Monthly Sales by Date" breaks="date">
  <break col="date" sort="up"/>
  <tcol name="sumsales" source="xsales" fun="sum" label="Sum of Extended Sales"/>
  <tcol name="sumcost" source="cost" fun="sum" label="Sum of Cost"/>
  <tcol name="transucnt" source="transid" fun="ucnt" label="Num of Unique Values in Trans ID"/>
</tabu>

<note>Create a computed column that assigns a week day to each date using the sdayofwk(X) function</note>
<willbe name="dayofwk" label="Day of Week" value="sdayofwk(date)="/>

The line above produces the following results:

![Image of sales data table]

Step 2, perform a tabulation and use the new column we just created as the group by column:

<sel value="between(date;20110101;20110131)="/>
<tabu label="Monthly Sales by Date" breaks="date">
  <break col="date" sort="up"/>
  <tcol name="sumsales" source="xsales" fun="sum" label="Sum of Extended Sales"/>
  <tcol name="sumcost" source="cost" fun="sum" label="Sum of Cost"/>
  <tcol name="transucnt" source="transid" fun="ucnt" label="Num of Unique Values in Trans ID"/>
</tabu>

<note>Create a computed column that assigns a week day to each date using the sdayofwk(X) function</note>
<willbe name="dayofwk" label="Day of Week" value="sdayofwk(date)="/>

<note>Tabulate using sdayofwk as the grouping column. Summarize both sales and cost</note>
<tabu label="Sales by Weekday" breaks="dayofwk">
  <tcol name="sumsalesbyday" source="sumsales" fun="sum" label="Total Sales By Day"/>
  <tcol name="sumcostbyday" source="sumcost" fun="sum" label="Total Cost By Day"/>
</tabu>

Notice here that we are tabulating a tabulation. No problem. Tabulations are just regular old 1010data worksheets, and can be manipulated the exact same ways as any other table or worksheet. Here are the results of the last tabulation:
Step 3, calculate margin (profits rule!!). To do this, create another computed column. Again, feel free to do this in the GUI, but we'll stick to the Macro Language for this example:

```
<sel value="between(date;20110101;20110131)"/>
<tableu label="Monthly Sales by Date" breaks="date">
  <break col="date" sort="up"/>
  <tcol name="sumsales" source="xsales" fun="sum" label="Sum of Extended Sales"/>
  <tcol name="sumcost" source="cost" fun="sum" label="Sum of Cost"/>
  <tcol name="transucnt" source="transid" fun="ucnt" label="Num of Unique Values in Trans ID"/>
</tableu>

<note>Create a computed column that assigns a week day to each date using the sdayofwk(X) function</note>
<willbe name="dayofwk" label="Day of Week" value="sdayofwk(date)"/>

<note>Tabulate using sdayofwk as the grouping column. Summarize both sales and cost</note>
<tableu label="Sales by Weekday" breaks="dayofwk">
  <tcol name="sumsalesbyday" source="sumsales" fun="sum" label="Total Sales By Day"/>
  <tcol name="sumcostbyday" source="sumcost" fun="sum" label="Total Cost By Day"/>
</tableu>

<note>Create a computed column to calculate margin by weekday</note>
<willbe name="marginbyweekday" label="Margin by Weekday" value="sumsalesbyday - sumcostbyday"/>
```

Run the code and you should get the following results:
Our results tell us that Saturday and Sunday are by far our most profitable days of the week. This makes a lot of sense, since that's when most people actually have time to go grocery shopping. Still, it's nice to not only know for sure, but be able to attach hard numbers to the phenomenon. But it does raise another question. It is intuitive that our highest margins, in terms of total dollars, are on the same days we have the highest sales, also in terms of total dollars. However, if we want to think about this same relationship in terms of percentages, does the same dynamic hold true?

Just for fun, let's create one last computed column. This time we're going to calculate margin as a percent of cost, to understand if we're profiting by a higher or lower percentage of our cost on a given day of the week. Remember, this will only apply to January 2011, but we could easily use this same process to understand these relationships across any time period we like.

Click the Submit button to run the code. Let's take a close look at the results:
Now this is interesting. While Saturday and Sunday are our two highest margin days in terms of total dollars, they actually represent the two days where our margin represents the least profit as a percentage of total cost. What could account for this apparent inversion of the relationship?

If you're interested in taking this process one step further to discover what drives these results, here are some questions you might ask:

- Are higher margin items sold in a higher percentage of the total volume of items during week days, as opposed to weekends?
- Are lower margin items sold in a higher percentage of the total volume of items during weekend days, as opposed to regular weekdays?
- Do these results reflect a seasonal result or do they hold their integrity throughout the entire year?

You can answer all the questions above with the basics building blocks we've already covered in this tutorial. However, as always, the most interesting questions are the ones you invent yourself.