Entire SAS script comprises of 3 sections –

1. Section A – PriceTrac data # Due : 10-01-14 Done on date
2. Section B – Sales data # Due : 15-01-14 Done on date
3. Section C – Content data(reviews and image) # Due :31-01-14

# Section A – PriceTrac Data

1. Import PriceTrac data
2. Calculate week number and discount
3. Prepare individual variables for each retailer
   1. Create separate files for each retailer
   2. Sort the data by week number, product id and client identification number (TCIN)
   3. Keep only one observation for week number and product id if there are multiple TCIN
   4. It is advisable to keep only required column and rename the columns with respect to each retailer
   5. Merge the files prepared for each retailer
4. We need to make sure that for each week there is only one product id.
5. Calculate additional variables such as
   1. Price difference variables(price difference between 2 retailers)
   2. Availability(own retailer and competitor)
   3. Length of the product description(own retailer and competitor)
6. Calculate Recency of the product in months and weeks. We need first extraction date of each product id. Then subtract today’s date from first extraction date. Convert the values in weeks and months
7. If the product has variants then it is captured in client identification number(TCIN). Take week, product id wise count of distinct TCIN in order to get number of variants. Similarly calculate maximum and minimum price
8. Final PriceTrac data set will have following list of variables –
   1. PriceTrac basic variables for client(own retailer)
   2. PriceTrac basic variables for competitor
   3. Price difference variables
   4. Availability and product description length(own and competitors)
   5. Number of variants

# Section B – Sales Data

1. Import sales data
2. Determine gross sales
3. Aggregate data on product id level
4. Transpose the data in such a way that data per week and product ids available in respective week will be stacked one below the other. For example –

|  |  |  |
| --- | --- | --- |
| Week Number | Product Id | Sales |
| 1 | P1 | 10 |
| 1 | P2 | 36 |
| 1 | P3 | 78 |
| 2 | P3 | 12 |
| 2 | P7 | 41 |
| . |  |  |
| . |  |  |
| 40 | P1 | 37 |
| 40 | P12 | 49 |

1. This has to be done for net and gross sales
2. Finally prepare consolidated file which will have week number, product id, gross sales and net sales
3. Merge sales data with PriceTrac. Do not exclude any observation from PriceTrac dataset. Keep the observation even if it does not have sales. PriceTrac file will be master file here

## Calculating Number of substitute

1. After merging PriceTrac and sales data, extract unique list of product names
2. Prepare excel file which will have product names and individual column for each features. Number of features should be decided manually looking at the product names
3. If the feature is present in the product name description then mark it as 1 else 0
4. Upload this excel file on SAS. We can create individual files for each segment
5. Number of substitute is calculated for own retailer and competitors
   1. **Own retailer**
      1. First determine feature wise count
      2. Keep this result in separate table
      3. Then calculate product name and feature wise sum of the count(calculated in step 1) and store the result in separate table
      4. Merge the 2 tables using each feature. If there are more than one features then join them using “OR” conditions
      5. For example, If we are joining , table1 and table2 tat has 2 features then conditions would be

table1.feature1 = table2.feature1 and feature1 = 1 OR

table2.feature2 = table2.feature2 and feature2 = 1

* + 1. Subtract 1 from the summation as it considers own product name for which we are finding substitute
  1. **Competitor retailer**
     1. Determine number of products having final price > 0. It ensures product was available on the retailer’s site
     2. In order to perform step i, we need to use PriceTrac and Sales merged file
     3. Remaining steps are same as calculating substitute for own retailer

1. Merge these variables with PriceTrac and sales data

# Section C – Content Data

## Section C.1 – Amazon Review Data

1. Import Amazon review data
2. Calculate week number using review creation date
3. Calculate rating score variables such as if rating score is 1 then amz\_rating\_1 = 1 and so on
4. **Review velocity variables**
   1. We need following variables to calculate velocity
      1. To date - week ending
      2. From ending - week beginning
      3. Month back - 30 days back from week beginning
      4. Two months back - 60 days back from week beginning
      5. Three months back - 90 days back from week beginning
      6. Six months back **-** 180 days back from week beginning
   2. Merge above variables with review file
   3. For week less than 1, we get next week’s beginning as week beginning for current week. In order to rectify this, update from date as **from date = from date – 7** if week number<1 and review creation date is not Monday
   4. Calculate week number and product id wise sum of Amazon’s individual scale rating, average review score, total reviews etc
   5. Find number of reviews present between week ending and month back,2 month back, 3 month back and six month back
5. Calculate additional variables using Amazon rating1, Amazon rating2, Amazon rating4, Amazon rating5 and reviews count
6. Merge all these variables back to review file

## Section C.1.1 – Sentiment Analysis

* 1. Run sentiment analysis on R. Import the file on SAS. This file will have sentiment score for each review
  2. Calculate variables such as average and total positive as well as negative score

## Section C.1.2 – Helpfulness of review

1. Calculate average review helpful votes by week number and product id
2. Divide sum of review helpful votes by sum of review total votes. We will be using this variable as weighted variable
3. Merge aggregated variables from sentiment and helpfulness of review
4. Compute additional variables such as
   1. Average weighted score – average sentiment score \* weighted variable
   2. Total weighted score – total sentiment score \* weighted variable
   3. Positive average weighted score – positive average sentiment score \* weighted variable
   4. Positive total weighted score – positive total sentiment score \* weighted variable
   5. Negative average weighted score – negative average sentiment score \* weighted variable
   6. Negative total weighted score – negative total sentiment score \* weighted variable

## Section C.2 – Amazon Image Data

1. Import image file
2. Calculate product id (Main product id)wise count of URL
3. Determine total number of instances where UGC = Yes. This will tell you number of images generated by user
4. With UGC = No, we can determine number of images not generated by user
5. Merge these variable with the file that has Amazon review(section c.1) data
6. After merging we will have content file for Amazon
7. Merge it back with PriceTrac, sales file

## Section C.3 – Kohl Image Data

***Note – Review file is not available due to insufficient information***

1. Import image file for Kohl
2. Determine same set of variables that are present for Amazon
3. Merge all these variables with the file that has information for PriceTrac, Sales, Amazon content
4. Sort the data by week number and productid

# Model Data

1. Drop unnecessary columns from the data file
2. Take log of price and sales variables
3. Data is now ready for modeling