Supplementary material for "A set-based visual analytics approach to analyze retail data"

1.0 Illustrations of visualizations in the analysis workflow (section 4.1)

The analysis workflow describes four summary (a-d) and three detailed (e-g) visualizations. The workflow overview (a) visualization has not been implemented yet. This section illustrates the remaining six visualizations (b-g) that are implemented in a visual analytics prototype tool, which we developed to investigate set-types data.

b) Product frequency histogram: This shows the distribution of the frequency of products. It is scalable to any number of unique products in the transaction database because it does not show the individual products but instead bins them based on their frequency.

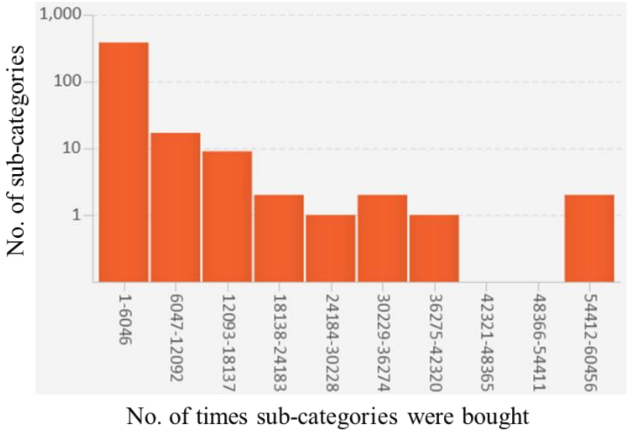

Figure 1: Product frequency histogram showing 417 unique sub-categories in 10 bins. The y-axis is set to a logarithmic scale.

c) Itemset length histogram: It allows analysts to see most common itemset lengths. In practice, it scales to any volume of data because transactions only contain a modest number of different products, and multiple lengths can be binned together. We used two versions of this chart in our convenience store case study. The first version plots the bins of the length of itemsets on the x-axis and the number of different itemsets on the y-axis (Figure 2a). We created the second version of this visualization in Microsoft Excel to display bins of the length of itemsets on the x-axis and the number of transactions on the y-axis (Figure 2b).

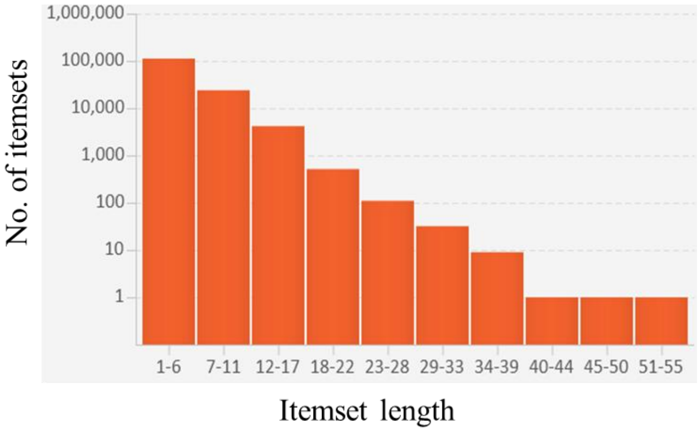

Figure 2a: Itemset length histogram showing 140,986 itemsets in 10 bins. The y-axis is set to a logarithmic scale.

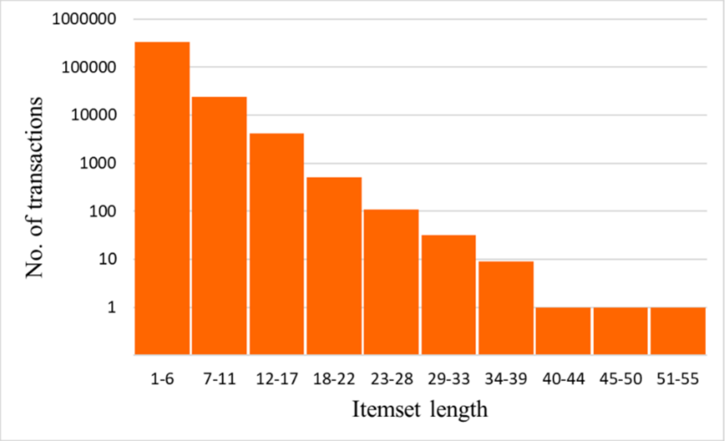

Figure 2b: Itemset length histogram showing 366,072 transactions in 10 bins. The y-axis is set to a logarithmic scale.

d) **Itemset frequency histogram:** It shows the distribution of the frequency of itemsets. This visualization also scales to any volume of data because the performance of histograms does not deteriorate with increasing numbers of observations.

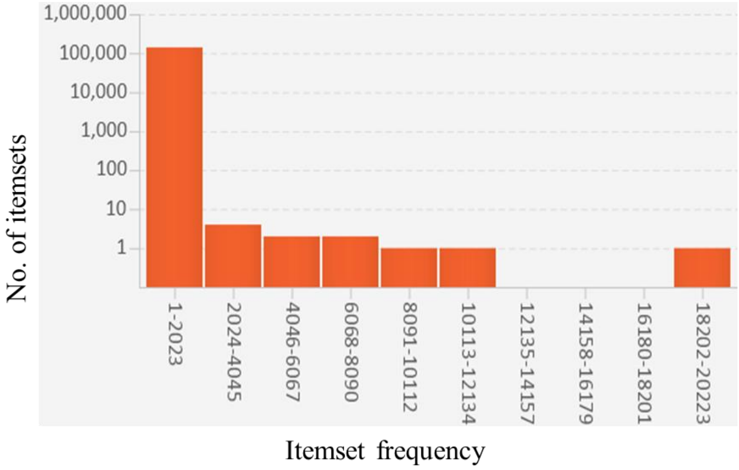

Figure 3: Itemset frequency histogram showing 140,986 itemsets in 10 bins. The y-axis is set to a logarithmic scale.

e) **Product frequency bar chart:** It allows analysts to see the number of times individual products were bought. While less scalable than the product frequency histogram, its scalability can be improved by automatically grouping low-frequency products as ‘other’.

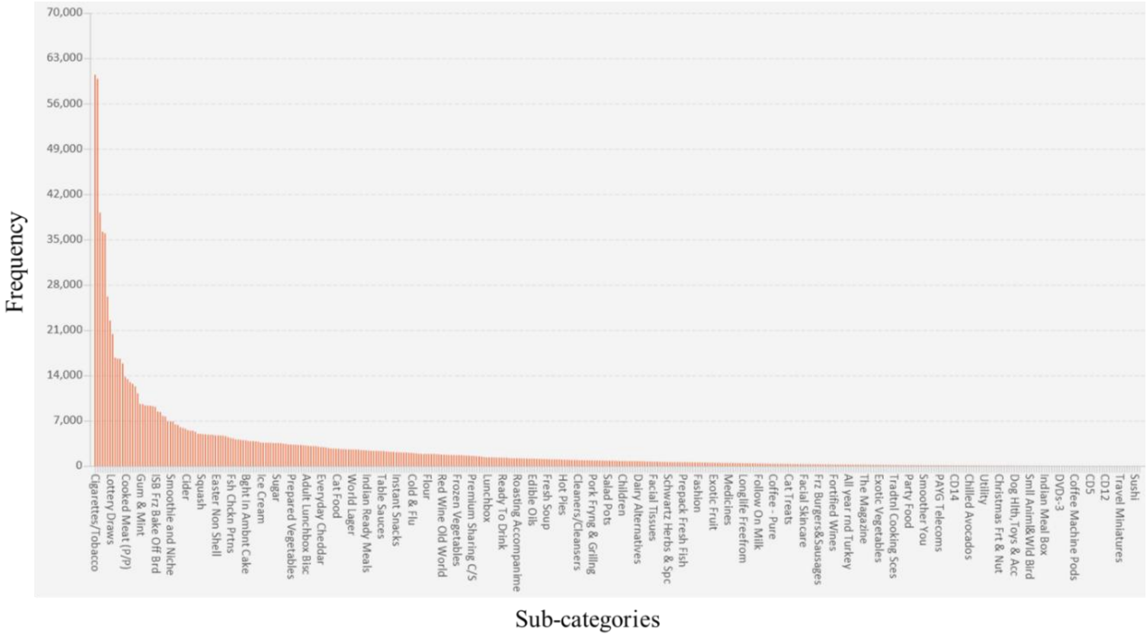

Figure 4: Product frequency bar chart showing the frequency of all 417 unique sub-categories in the dataset.

f) **Itemset frequency bar chart:** It shows the frequency of individual itemsets. While less scalable than the itemset frequency histogram, low-frequency itemsets can be grouped to improve its scalability.

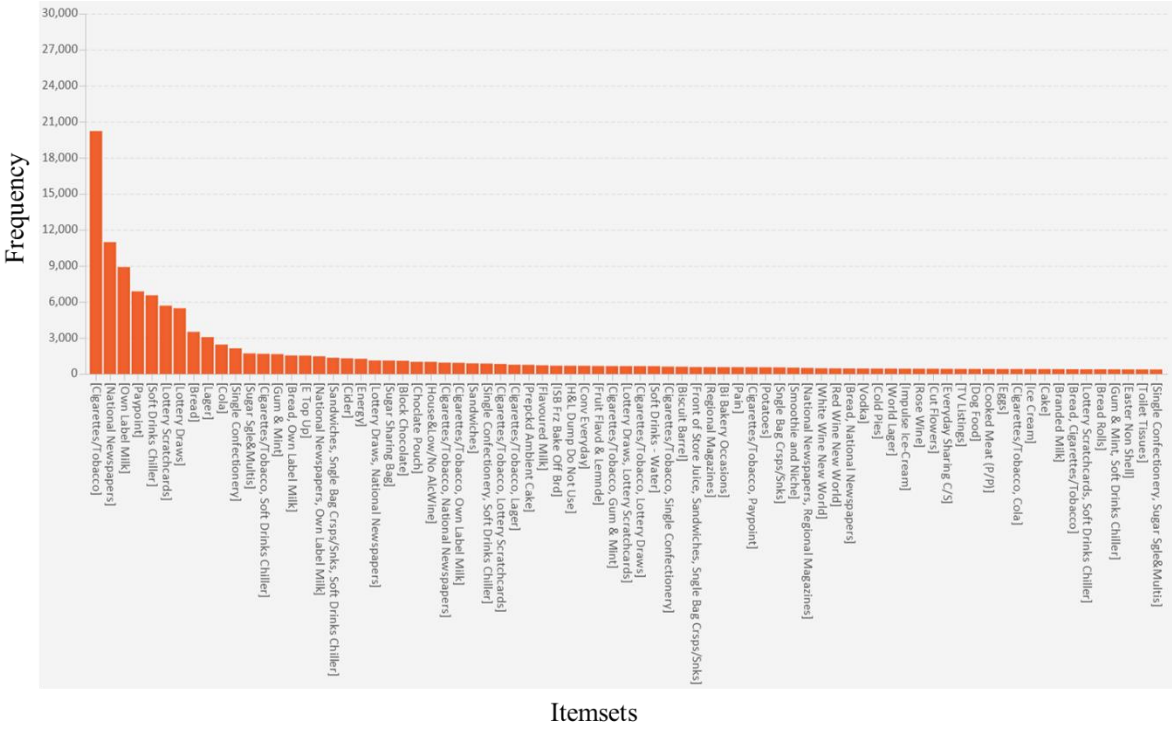

Figure 5: Itemset frequency bar chart showing the frequency of itemset with a support of 0.1 percent (i.e., 365 transactions).

g) **Itemset heatmap / matrix plot:** It allows analyst to see the frequency, length, and composition of each itemset. This visualization at most can only show a few thousand itemsets, and is more effective with many fewer itemsets (say, tens).

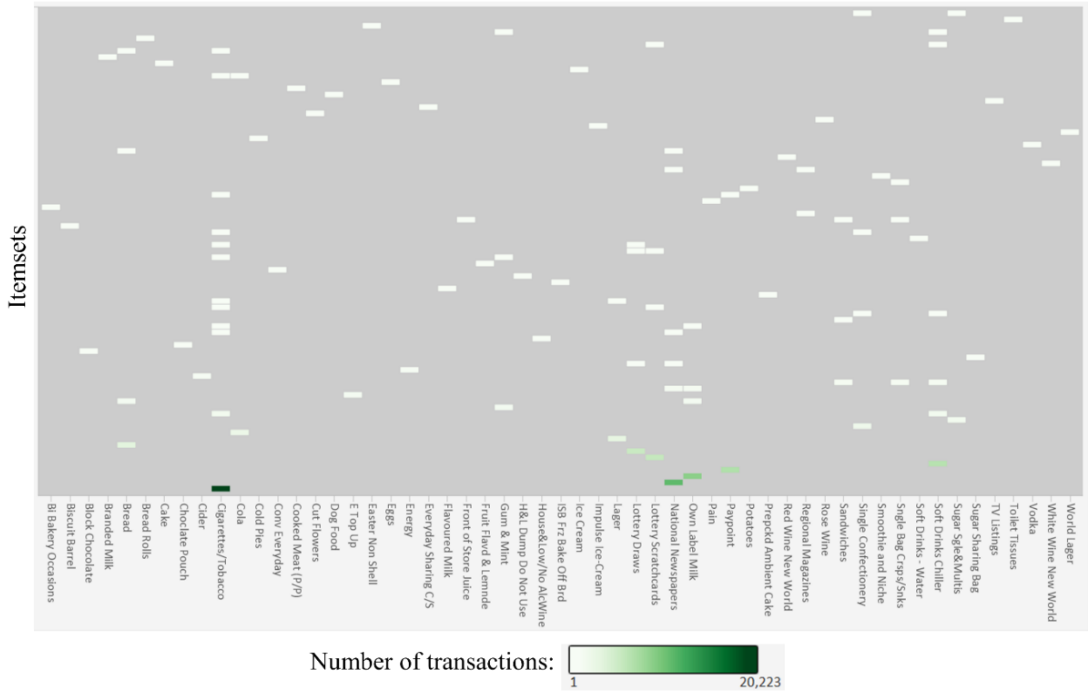

Figure 6: Itemset heatmap/matrix plot showing the frequency of itemset with a support of 0.1 percent (i.e., 365 transactions).

2.0 Details of analysis workflow

The first step in the proposed analysis workflow is data cleaning. Figure 7 illustrates the data cleaning process that was used in the convenience store case study, which began by removing the refunded items. Since the dataset includes a Boolean field for refunded items, this was a straightforward first step to clean the data. Next, we visually explored the data at different product hierarchies (i.e., product, sub-category, category, and super-category) and decided to perform the analysis at the level of sub-categories. The visual exploration of data also highlighted several products (e.g., shopping bags and tobacco think 25) that made sense to remove from the data. These products belonged to the sub-category named ‘system test’, which was removed from the data. The outcome of this process was a cleaned dataset that was used in the convenience store case study.

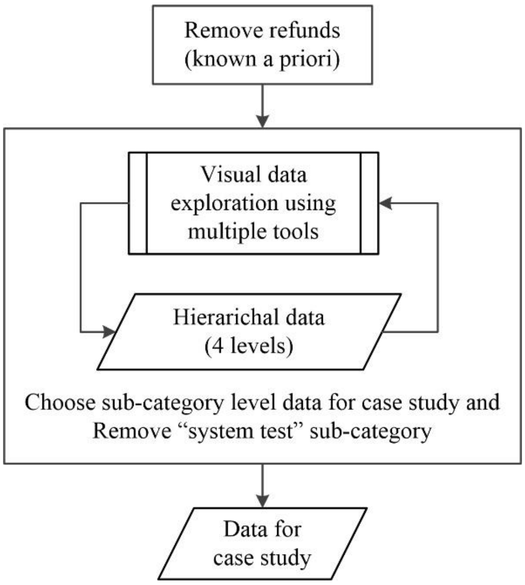
Remove refunds (known a priori)
Visual data exploration using multiple tools
Hierarichal data (4 levels)
Choose sub-category level data for case study and Remove “system test” sub-category
Data for case study

Figure 7: Illustration of the data cleaning process.

Data cleaning was followed by the computation of all itemsets using the ‘exclusive set intersections’ approach. Lastly, we iteratively analyzed the data by combining computational data processing with interactive visualizations (see Figure 8). The first iterative step of this analysis involved the exploration of data using interactive summary visualizations (i.e., product frequency histogram (Figure 1), itemset length histograms (Figures 2a and 2b), and itemset frequency histogram (Figure 3)). This led to the selection of a subset of data, containing length 1, 2 and 3 transactions, for further analysis.

The second step of the analysis involved three iterations (one each for transactions of length 1, 2, and 3) and then the creation of product frequency bar charts to compute the sub-categories that were bought most frequently in each of these lengths.

Next, we selected the length 2 and length 3 itemsets and performed an iteration for each length. In each iteration we created a itemset frequency histogram and bar chart to select the five most frequent itemsets, and created a heatmap to explore the composition of the five most frequent itemsets for the transaction length.

Lastly, we selected the most frequent length 2 itemset and computed the related occurrences of this itemset in the longer length transactions using Tableau. This showed that the selected itemset was bought 3,837 times with other sub-categories.

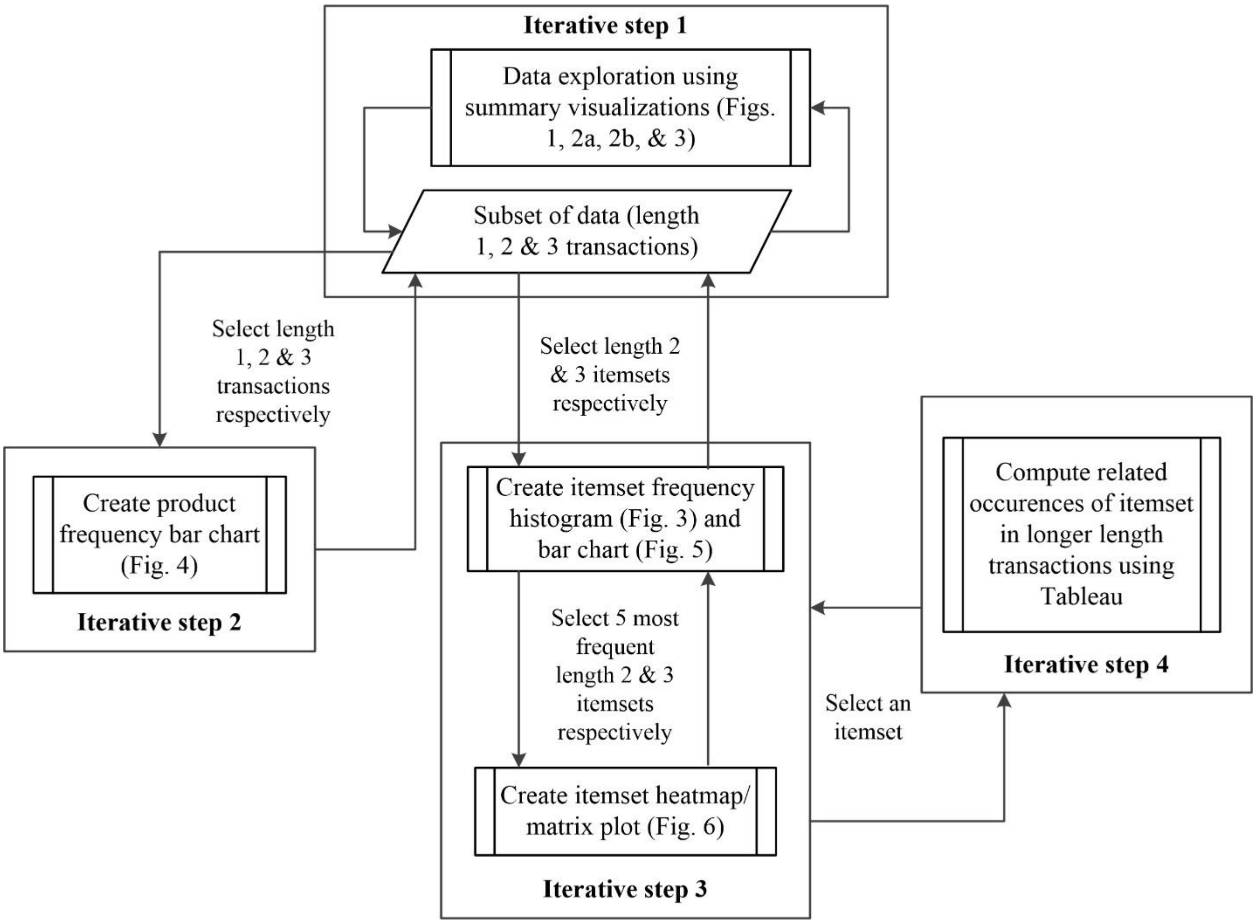
Iterative step 1
Data exploration using summary visualizations (Figs. 1, 2a, 2b, & 3)
Subset of data (length 1, 2 & 3 transactions)
Select length 1, 2 & 3 transactions respectively
Select length 2 & 3 itemsets respectively
Create product frequency bar chart (Fig. 4)
Iterative step 2
Create itemset frequency histogram (Fig. 3) and bar chart (Fig. 5)
Select 5 most frequent length 2 & 3 itemsets respectively
Create itemset heatmap/ matrix plot (Fig. 6)
Iterative step 3
Compute related occurences of itemset in longer length transactions using Tableau
Iterative step 4
Select an itemset

Figure 8: Workflow explaining the iterative analysis that was used in convenience store case study.