ZhaoCheng Yu zyr531

1. Business Target

My target is to find the most 100 suitable SKUs in a specific state to maximum the profit of the retailer chain.

2. Intuitive Approach

Based on the strinfo.csv, I make a data exploration of the location distribution of stores. The following figure depicts the result.

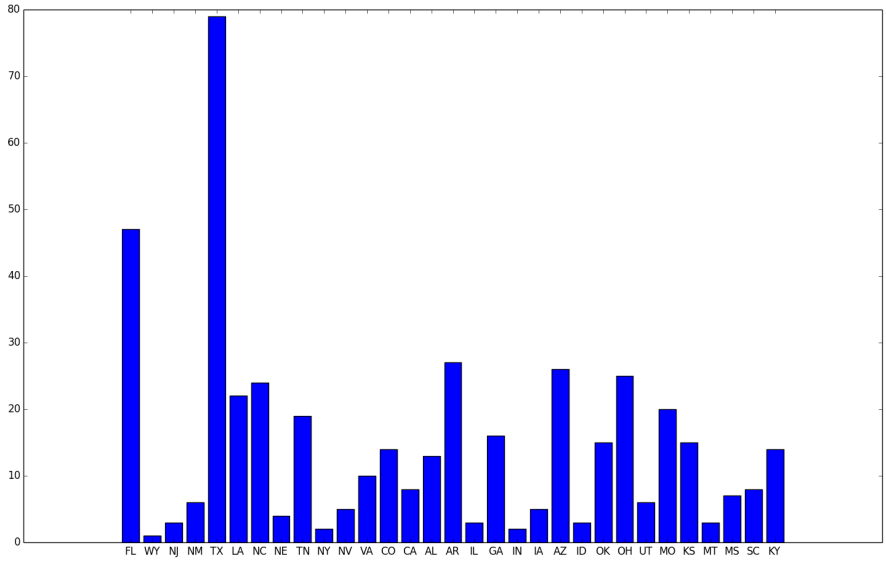

Figure 1 Store distribution

As the result shows, there are a lot stores in FL and TX. Considering the limitation of my laptop, I decided to pick LA as the target state. And assortments of SKUs of different states are little likely similar and thus we are able to draw a more promising result by focusing on a single state.

3. Data Preprocessing

By looking up to the STRINFO table, I am able to get the numbers of the stores in LA and then picking out LA's transaction records from TRNSACT table can be easily done. Compared with return items, it is more effective to focus on purchase items so that return items are flited. The most tricky part of this problem is how to identify an unique transaction. After several trial and error , I found that REGISTER , SALEDATE and TRANNUM could be used to mark an single transaction. Then all I have to do is to find items with same REGISTER, SALEDATE and TRANNUM and put them into the same basket. The transaction data is depicted as output_LA.csv.

4. Methodology and Result

Another tricky question is how to set the minimum support. When is comes to statistics, we always prefer the number of samples to exceed 30. Since the number of observations of my dataset is 348096 and therefore 30/348096 = 8.618312e-05 is the minimum support I chose. As for minimum confidence, I set it to 0.5 based on the rules of thumb. Following figures depict rules mining of the dataset.

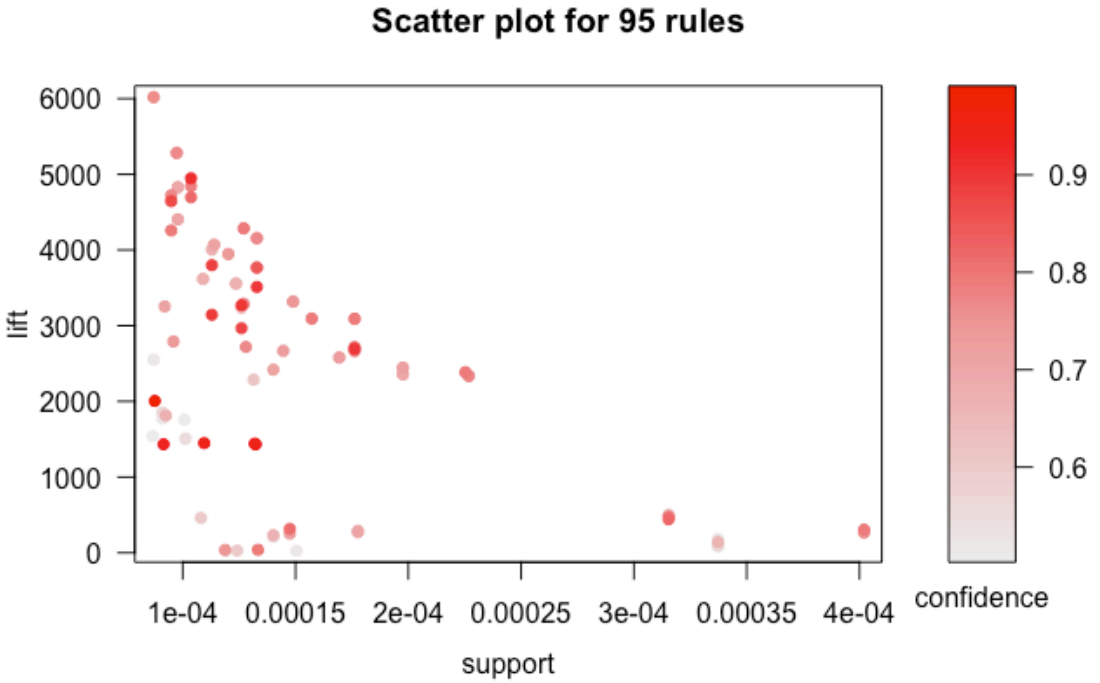

Figure 2 support=0.000086, confidence=0.5

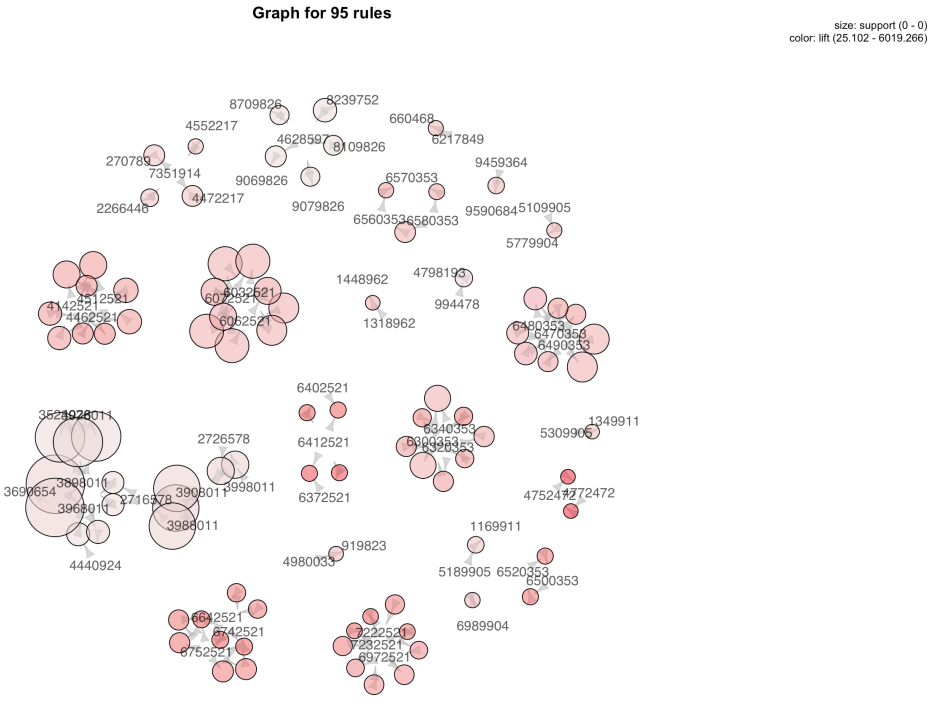

Figure 3 rules and SKU candidates

5. Conclusion

All rules and SKU candidates depicted could be used to solve this problem. However, there are still progress can be taken if given sufficient time and more advanced laptop. First, we can add some time variables of the dataset. For example, sunscreen and facial cleanser could have a strong association rule in stores of IL in summer but the rule might become very weak in winter. While such rule could always be strong in CA. By adding time variables, we are able to have a more persuading results to show how those SKUs should move in different time of the year. Additionally, it would be better if we can analyze more states instead of a single state.