**Executive Summary**

Store arrangement is an important factor that determines that revenue of a store. Some frequently used business strategies to improve revenue based on product position include 1).putting items that are purchased together close to each other, 2). putting items that are necessary for daily life, and also purchased together by customers far away from each other. For example, putting milk and eggs in two different corners of the store. Because customers will almost buy milk and eggs together, they will walk around to see other products when trying to get those two items.

Store transaction history provides a rich dataset to explore the market basket. Association rule is an unsupervised learning method that can automatically mine frequently appeared patterns in this case, item basket. It will help company to find combinations of items that occur together frequently, and this information will eventually help us decide the position of each item.

In this project, we are mostly interested in 1). finding the most frequently purchased item baskets in the store, rearranged them together 2). Investigating if the popular item basket differs by store 3). Searching for target items that are as frequently purchased but generates great profit.

Our result shows: 232 unique transactions and 336 unique SKU items in our 8000 row dataset. There are the top 5 most frequently appeared items in association rules: SKU9861016, 9859975, 9860180, 9860515. Our algorithm generates top 50 (more can be looked) association rules based on support, lift and confidence. Items that appeared in the same association rules should be placed close to each other.

**Problem Statement**

In-store layout and shelf design affects customer behavior which is directly linked to the final company revenue. A common used strategy is to put combination of products that co-occur in a transaction together. In this project, we will identify frequent market basket by using association rules and choose the top 100 SKUs to rearrange position based on the results from association rules.

**Assumptions**

No data entry error. Customer behavior change over time is not significant. Top 8000 rows is a large enough data set to represent the characteristics of the whole dataset. Missing data is missing completely at random.

**Methodology**

Data is extracted from pos.trnsact table in SQL. Data elements extracted from this table include SKU, tranum (transaction code), quantity and store. We will select the type of transaction as ‘purchased’ since that is what we are mostly interested in here. To limit the size of dataset, we only take the top 8000 rows for our analysis. We then performed data cleaning and tweaking in order to having the data in appropriate format for the arules . Our final data have all the SKUs during a same transaction for each row. Each SKU element in a row is separated by comma. We only keep the rows without missing data. After getting the right dataset format, we performed data exploration to get the basic information of our data. We then performed association rules by using the {arules} package in R and visualized the result using plot() function in {arulesViz} package.

**Analysis**

Our initial exploration of the data shows that there are 232 unique transactions among our 8000 row dataset, with 336 unique SKU items. The top 10 most frequently purchased items and their frequency are as the following:

SKU9861016 2389

SKU9860713 1184

SKU9860923 1004

SKU10049860515 337

SKU9860696 290

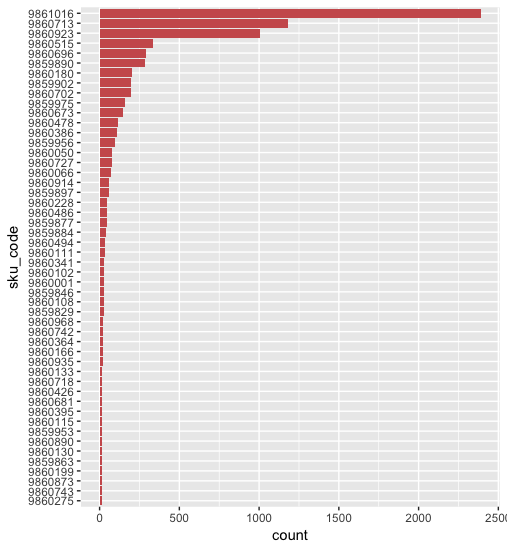
SKU9859890 283

SKU9860180 206

SKU9859902 199

SKU9860702 195

SKU9859975 157. We also visualized the distribution of top 50 most frequently purchased item here.



Association rule analysis shows most basket has a size of two, with median size of 2, and maximum size of 44. The top 5 most frequently used items are as the following:

most frequent items:

9861016 9859975 9860180 9860515

159 85 77 77 71

We chose our minimum support as 0.1 and confidence as 0.8 to reduce running time. It shows that the top 10 most frequently used rules are as the following:

lhs rhs support confidence lift

1 {9860111} => {9861016} 0.1034483 1 1.459119

5 {9860494} => {9861016} 0.1120690 1 1.459119

9 {9859877} => {9861016} 0.1206897 1 1.459119

15 {9859884} => {9861016} 0.1422414 1 1.459119

23 {9860228} => {9861016} 0.1551724 1 1.459119

58 {9860673} => {9861016} 0.1982759 1 1.459119

92 {9859902,9860494} => {9860180} 0.1034483 1 3.012987

93 {9860180,9860494} => {9859902} 0.1034483 1 3.267606

94 {9859902,9860494} => {9860515} 0.1034483 1 3.012987

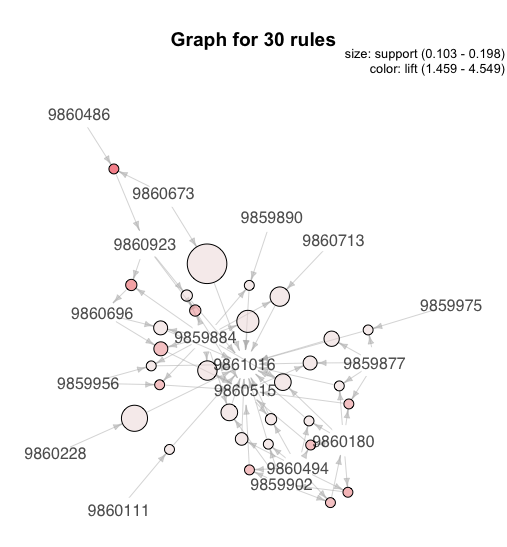
96 {9859902,9860494} => {9861016} 0.1034483 1 1.459119

Take rule 92 for example, customers who bought {{9859902,9860494} is also likely to buy {9860180}. They have three times higher chance to buy 9860180 then buying those two products independently.

We further order the rules by lift and find the top 50 rules that have the highest lift:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | lhs |  | rhs | support | confidence | lift |
| 151 | {9860486,9860673} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 176 | {9860486,9860696} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1722 | {9860180,9860486,9860673} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1725 | {9860486,9860515,9860673} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1728 | {9860486,9860673,9861016} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1774 | {9859890,9860486,9860696} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1789 | {9860180,9860486,9860696} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1792 | {9860486,9860515,9860696} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 1795 | {9860486,9860696,9861016} | => | {9860923} | 0.1034483 | 1 | 4.54902 |
| 2861 | {9859897,9860673,9860923} | => | {9860050} | 0.1034483 | 1 | 4.54902 |
| 2870 | {9859897,9860673,9860696} | => | {9860050} | 0.1077586 | 1 | 4.54902 |
| 3705 | {9859890,9860386,9860727} | => | {9860923} | 0.1077586 | 1 | 4.54902 |
| 3708 | {9860386,9860696,9860727} | => | {9860923} | 0.112069 | 1 | 4.54902 |
| 188 | {9860486,9860696} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 322 | {9859956,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 448 | {9859897,9860673} | => | {9859890} | 0.112069 | 1 | 3.932203 |
| 735 | {9860066,9860386} | => | {9859890} | 0.1206897 | 1 | 3.932203 |
| 958 | {9860050,9860386} | => | {9859890} | 0.1336207 | 1 | 3.932203 |
| 1773 | {9860486,9860696,9860923} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 1825 | {9860180,9860486,9860696} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 1828 | {9860486,9860515,9860696} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 1831 | {9860486,9860696,9861016} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2001 | {9860386,9860914,9860923} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2022 | {9860386,9860702,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2303 | {9859956,9860696,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2306 | {9859902,9859956,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2309 | {9859956,9860713,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2312 | {9859956,9860180,9860914} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2315 | {9859956,9860914,9861016} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2440 | {9860696,9860702,9860914} | => | {9859890} | 0.1163793 | 1 | 3.932203 |
| 2709 | {9859897,9860050,9860386} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2734 | {9859897,9859956,9860386} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2776 | {9859897,9860386,9860702} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2798 | {9859897,9860386,9860696} | => | {9859890} | 0.1163793 | 1 | 3.932203 |
| 2801 | {9859897,9859902,9860386} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2807 | {9859897,9860180,9860386} | => | {9859890} | 0.1163793 | 1 | 3.932203 |
| 2866 | {9859897,9860050,9860673} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2887 | {9859897,9859956,9860673} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2897 | {9859897,9860673,9860923} | => | {9859890} | 0.1034483 | 1 | 3.932203 |
| 2915 | {9859897,9860673,9860702} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2937 | {9859897,9860673,9860696} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2939 | {9859897,9859902,9860673} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2942 | {9859897,9860673,9860713} | => | {9859890} | 0.112069 | 1 | 3.932203 |
| 2945 | {9859897,9860180,9860673} | => | {9859890} | 0.112069 | 1 | 3.932203 |
| 2947 | {9859897,9860515,9860673} | => | {9859890} | 0.1077586 | 1 | 3.932203 |
| 2950 | {9859897,9860673,9861016} | => | {9859890} | 0.112069 | 1 | 3.932203 |
| 3144 | {9859897,9859956,9860702} | => | {9859890} | 0.1163793 | 1 | 3.932203 |
| 3251 | {9859897,9860702,9860923} | => | {9859890} | 0.112069 | 1 | 3.932203 |
| 3285 | {9859897,9860180,9860923} | => | {9859890} | 0.1163793 | 1 | 3.932203 |
| 114 | {9859884,9860923} | => | {9860696} | 0.1077586 | 1 | 3.741935 |

The rules graph generated by arulesViz is as the following. As we can see from the graph, SKU 9861016 appeared in most rules, rule



**Conclusion**

Top 50 rules that ranked by lift is found by our algorithm. Our support is set as 0.1 and confidence as 0.8. Among those 50 rules, 100 candidate items can be selected. Items that appeared in the same rule is suggested to be put close to each other.

**Next steps**

To make sure that our conclusion is not biased, validation should be performed to confirm: 1) does the characteristics of our sample dataset represent the whole dataset? 2) does association rules vary by stores? 3) does association rules vary by time?