**Executive Summary**

The following study was done on Dillard’s which is a major retail chain with several stress in the US. Dillard’s is interested in rearranging the floors of the store in order to group SKU’s that are frequently bought together in order to increase profits. For budgetary reasons, at most 20 moves were allowed. The goal of the study was to determine the baskets that are most frequently purchased and therefore determine the 20 most profitable moves.

In the preliminary analysis profits were analyzed for various SKU as well as across regions to understand where Dillards should be focusing their energy. It was determined that the Northeast region be eliminated from the study since it had very few stores with lowest profits. The profits of most SKU’s were clustered around the mean suggesting that all products had similar profits. Outliers were eliminated. Next, analysis was conducted on brands to understand what were the most popular brands in terms of the number of transactions. It was seen that the top 10 brands controlled about 25% of all transactions and 3 of those top 10 brands are even more significant. Therefore for association, I decided to limit my data to transactions only across these top 10 brands. Basket ID data table was then created with meaningful features and the data was transformed into transactions and ready for association.

Using the apriori algorithm I initially created 100 rules using a very low confidence of 10%. Association was performed on brands and not on the individual items because brands are often grouped together in most stores incusing Dillards. It was noted that many of the rules appeared redundant and those with very low confidence did not give any useful insight. There was only a few rules with high supports and relatively high confidence that had high lifts. Therefore the rules were further pruned to finally determine the 20 most important rules. The confidence was increased to 30% for this step. It was noted there was a tradeoff between the support confidence and lift however the most important rule to make the rearrangements was grouping items that have high profitability.

It was seen that all the 20 rules contained the top 3 brands: Lancome, Clinique and Cabernet. Additionally, Clinique was never grouped with Lancome and both these brands were often grouped with Cabernet. This revealed the insight that Lancome and Clinique are competing prodicts each with their own loyal customer base and hence should not be located together. Instead, they should be located next to Cabernet. Also, it would be the most useful to focus on these top 3 brands and they have the strongest association.

**Problem statement**

Dillards is a major retail chain with several stores in the United States. Dillards is interested I rearranging the floors of the stores (change the planograms) in order to group SKU’s that are frequently bought together in order to increase storewide profits. For budgetary reasons, the store can make at most 20 moves across the entire chain. I have been supplied with point-of-sales data for analysis. My goal here is to determine the 100 SKU’s that are the best candidates to modify the planograms and hence determine the most profitable 20 moves.

Data Description:



The data contained 5 tables as depicted above. The tables included store information (STRINFO), prince information (SKSTINFO), SKU information (SKUINFO), transaction data (TRNSACT) and departments (DEPTINFO). Each table contained several attributes as shown above.

**Methodology:**

Before the data can be analyzed, the data had to be cleaned and prepared. The first step was to reduce the the data size for the TRNSACT table which currently had more than 120m rows. I decided to randomly sample 20% of this table which would still result in a significantly large table that well represents the overall data. Next the after labelling the columns in all tables, several columns were discarded as follows: INTERID, SEQ and MIC were all removed from TRNSACT, and UPC and VENDOR were removed from SKUINFO. This is because these attributes were not required for the analysis. Additionally, items that were returned were discarded from the data since I believe these items would negatively affect the analysis. Lastly, DEPTINFO table was discarded completely as it was not required in the study. Of course all rows with incomplete data were also discarded.

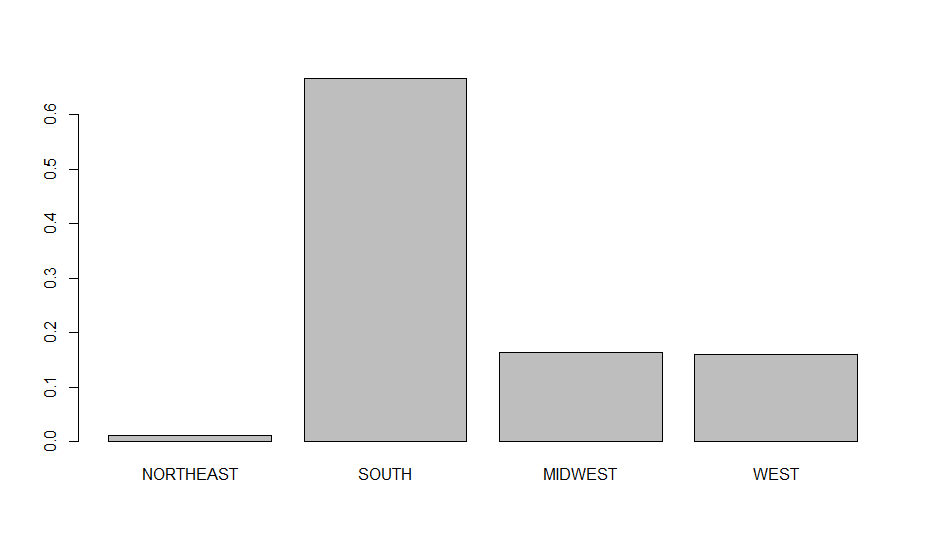
The next step was to convert data into factors such that is was ready for association analysis. Also, the states were divided into 4 regions: the NORTHEAST, WEST, MIDWEST and SOUTH to make it easier to understand the presence of the retail stores and further clean the data.

Now SQL commands were used in R to extract sales and cost information on all SKU’s in order to create a new table that contained the profit for each item. The profit analysis was then extended to regions to understand which regions were most profitable. This was done primarily to evaluate the results at the end. Brands were then analyzed further to determine which were the most popular. It seemed most appropriate to run associating on the top performing brands since these are part of most baskets.

Now to carry out association, basket id’s need to be created which will identify items bought in certain transactions. The features of the basket ID are are all selected from the TRNSCT table and will include sales, date, register, transaction number and SKU. From the brand analysis done earlier, I decided to run the association only on the top 10 brands as they represented about 20% of all transactions which created not only a representative sample but also made association more meaningful since items under these brands are much more prevalent. Finally, the data was ready for association.

**Analysis:**

The first step of the analysis was summarizing the most important data in order to make inferences that were to be included in association. Regions were analyzed to begin with to determine where the stores were most prevalent.



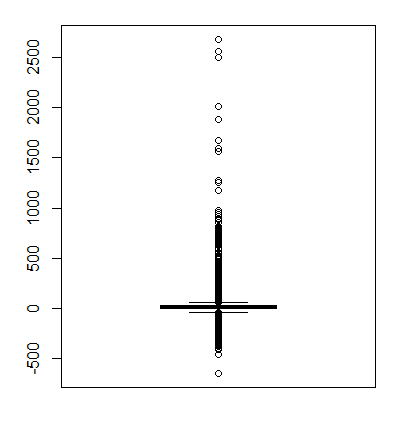
As shown above, the NORTHEAST has the lowest number of stores by far while the SOUTH has the largest number of stores. Additionally, there were many missing data points for the NORTHEAST region and therefore it was concluded that these stores will be removed from the analysis. This helps reduce the dataset further and make it more manageable.

**Profit analysis:**

First the profit was calculated for each SKU. The results are as follows:



It is surprising that there were several SKUS’s that result in negative profits and perhaps this is something to research further in the future.



Also as depicted on the left, while there is a large spread of profits across the SKU’s, most of the profits are close to $20 which is around the mean. There are also very few outliers. These outliers were removed to make the analysis more accurate since they would skew results.

Profit by region:

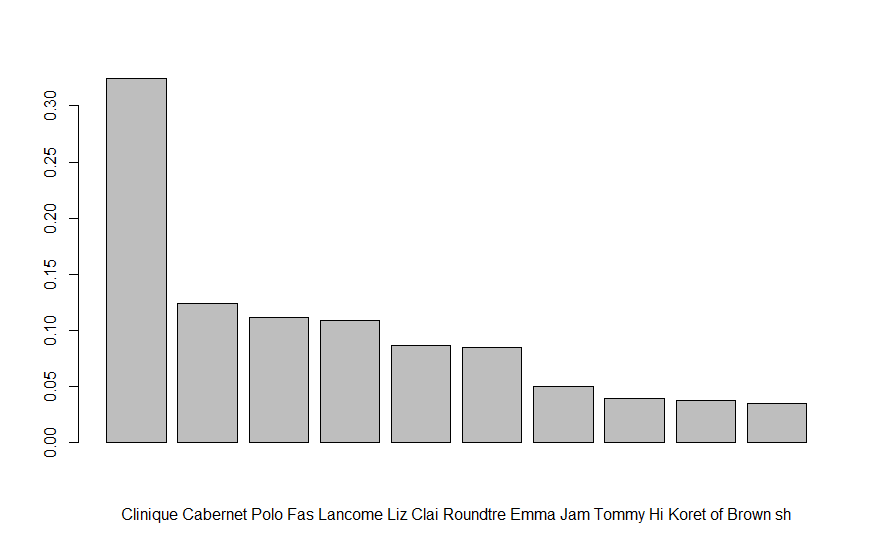
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Northeast | Midwest | West | South |
| Mean profit | 15.08 | 19.18 | 19.61 | 20.89 |

Overall there is a almost uniform distribution of profits across all 4 regions.

**Brand analysis**

It was determined there were 1830 brands amongst my reduced data set. Out of this the top 10 brands made more than 25% of the data by counting the number of transactions in each brand. This is depicted below. It made sense to continue association only using these top 10 brands since they are more prevalent in the data and will ensure better analysis. Clinique, Cabarnet and Polo were amongst the top 3 performing brands and perhaps Dillard’s should pay more attention to them.

**# of transactions of top 10 versus other brands**



TOP 10

REST



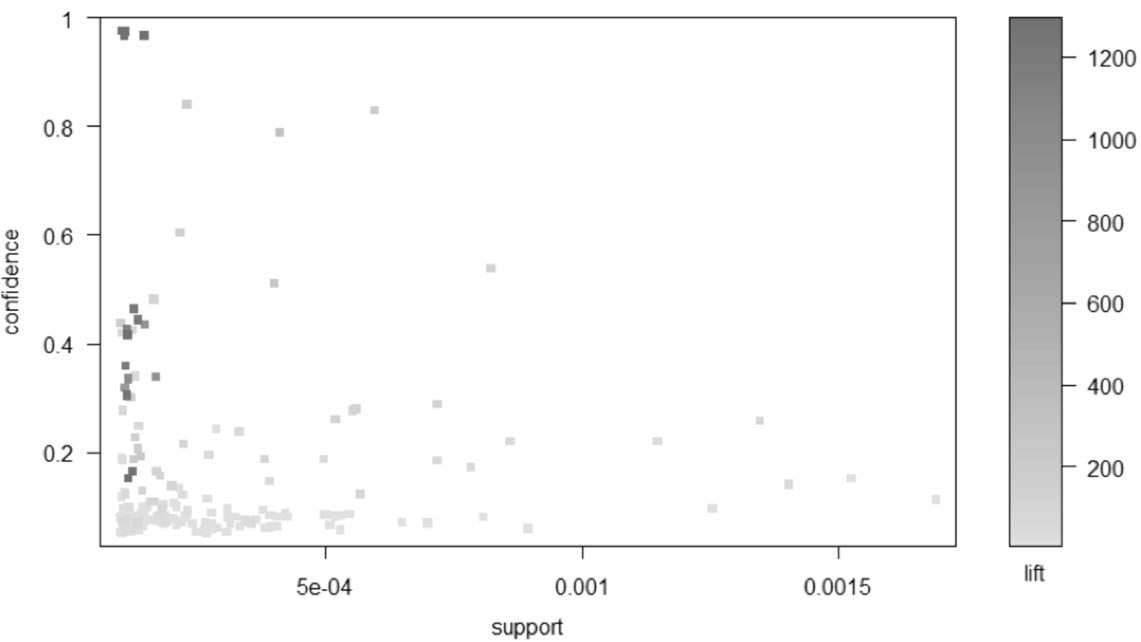
**Association**

To begin with association a basket table was created with basket ID’s that will contain the items that were purchased in given transactions. The basket ID’s contained the attributes Sales, DATE, Register, Transaction number and SKU. The table was then transformed into transaction form.

Now the Apriori function in R was used. I initially set a confidence of 0.1 and support of 0.001 which were inputted into this function. The reason for choosing a low support is because of the large number of baskets analyzed however this could have been much lower if more data points were included. Confidence of 10% seemed reasonable especially since we had limited our data to top 10 brands.

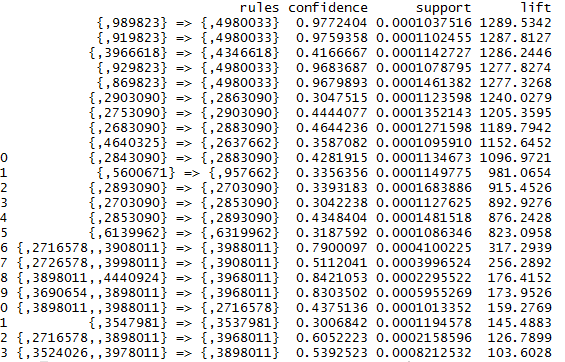
NOTE: Association was done on brands as it seemed more reasonable in stores like Dillards where brands are often grouped together for most types of products. There are of course some exceptions of this and perhaps would be a limitation of the study. However given that only 20 moves were allowed, it seemed more reasonable to associate by brands.

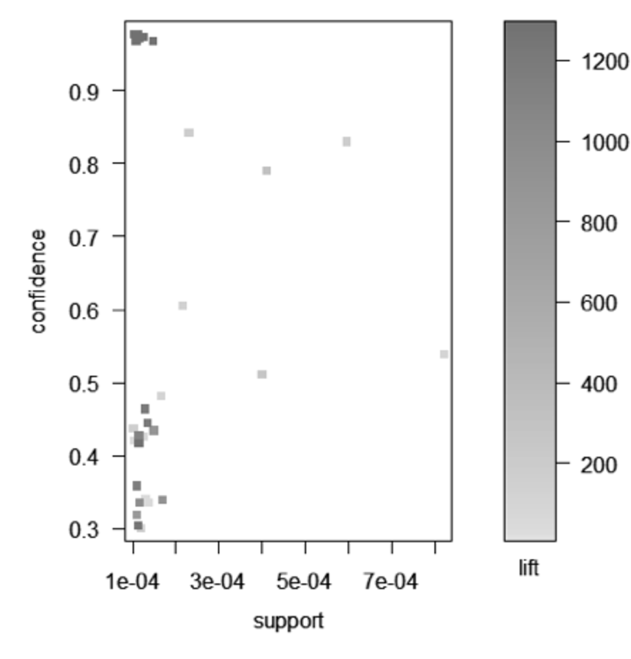
As explained in the problem statement we wanted to begin with 100 moves and therefore determine the top 100 rules. This will be pruned further in the next steps to determine the top 20 rules. These 100 rules are visualized as follows:

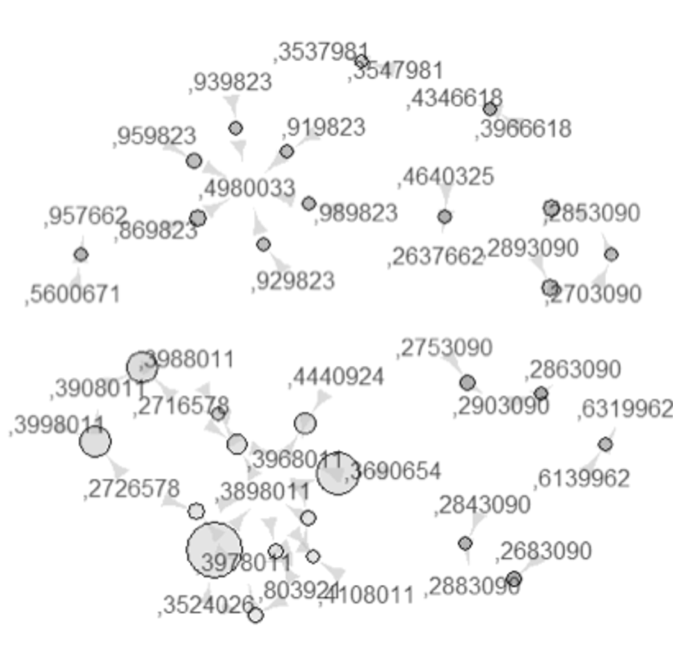


Looking at this graph we can see that the highest lift occurs for items that relatively high confidence and minimum level of support. Looking at the spread, there are a few rules on the bottom right corner that have high support (meaning the items are present in many baskets) but very low confidence (meaning an item purchased doesn’t not imply the purchase of the other). These rules are of course of no meaning since there is no association. Therefore, as we further prune the rules we will look for rules with high confidence and relatively high support.

Now to reduce the number of rules to 20 I set a new confidence of 0.30 after testing various values. The support remains the same. This higher confidence is important to identify rules that show stronger association of products. The lift in this case is greater than 100 suggesting that it is 100 times more likely that a certain item is purchased given the purchase of another item. The 20 resulting rules are as follows:

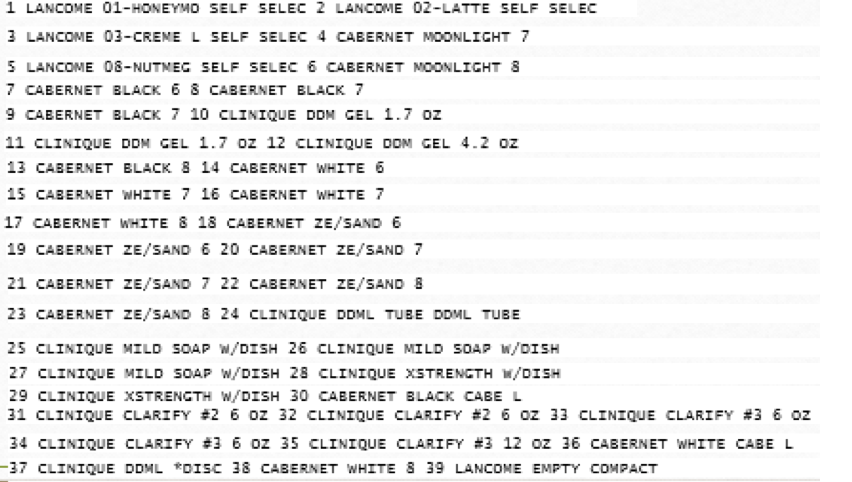






Support is higher for bigger bubbles and and lift is higher for darker bubbles. Unfortunately this graph suggests that there is often a tradeoff between lift, support and confidence. While lift is the most important rule, it does not necessarily correlate to the highest profits. Items profit margins need to be studied and those items with large profits and relatively large lifts and confidence with a reasonable support should be grouped together. Since we don’t have data on cost of all the SKU’s it is impossible to create a dynamic minsup.

Furthermore, looking at the 20 rules output it can be seen that a majority of the baskets only contain 2 items. There are a few with 3 items. Therefore most moves to be made in the planogram will require 2 items to be grouped together. Also since support of the rules with 3 items is much lower, this suggests that the store should focus its moves on 2 items.



Here we finally have a list of 20 groupings that should be made to increase sales and profitability. It is interesting to note that the top 3 brands I mentioned before appear in all of the rules suggesting that they are often bought together. However, Lancôme and Clinique are never bought together. Tis makes sense because both brands are competitors providing the same range of products. Consumers most probably are loyal to either of the brands. Additionally all the above products appear to be toiletry or skin care products suggesting they are often bought together.

**Conclusion and Recommendations**

The key insights we drew out were as follows. Firstly, most of the baskets contained 2 items and a few with 3. Therefore, during rearrangement, it will be important on focusing on the bringing 2 products together most of the time. Secondly, It was noted that all of the 20 rules contained only the 3 most popular brands: Lancome, Clinique and Cabernet. Additionally, Lancome was always seen with a Cabernet and so was Clinique but Lancome and Clinique were never seen together. This supports the hypothesis that most consumers have strong brand preferences and will not purchase some products from Lancome and others from Clinique. Instead they may supplement their purchase with Cabernet that most likely has a different product range. Therefore Dillard’s should not group these top 2 brands together but instead group them near Cabernet products. Overall Dillards must focus on the top 3 brands.