# Association Rules for Dillard’s Retail Chain

## Executive Summary

As a major department store chain with 453 stores spread across the United States, Dillards should be utilizing data collected from years of sales to optimize the planograms of its stores. This would ensure that the store is keeping the items which sell together closer by to increase its sales numbers.

Association rules is a rule-based data mining method for discovering interesting relationships between variables. It is intended to identify strong rules discovered in datasets using some measure of interest. Market Basket Analysis is a vertical of Association Rules which deals with finding affinity between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets or stores.

Dillard’s has provided a POS dataset of ~120 million transactions spanning a year (1st August 2004 – 27th August 2005). Out of this, a subset of latest ~9 million transactions was selected for Market Basket Analysis.

The analysis yielded 52 rules to provide 100 SKUs for considering for movement. Off these 100 SKUs, the rules which would yield 20 best moves were selected based on high Lift and confidence values.

## Problem Statement

Dillard’s is a major retail chain with several stores. Their point-of-sales (POS) data over a period of time was provided. Due to limitations of maximum 20 moves across the entire store, our task was to identify 100 SKUs which would be best candidates to rearrange the planograms which would enable the chain to achieve better sales.

## Assumptions

1. The data is available for 332/453 stores. The data from these stores is representative of other stores. Customer behavior across these stores is the same.
2. Data for August 2005 is representative of the data for entire duration of the capture window. This assumption implies that there might be seasonal trends which will not be captured by the analysis. There is another implication in that there are ~366,116 SKUs which are purchased in this month from total 1,048,576 SKUs defined in the database. We assume that the most popular and non-seasonal SKUs will be captured in the subset of SKUs for August 2005.
3. SKUs are independent items. There are no 2 SKUs which differ only in size.
4. The returned items don’t matter in the analysis as the initial objective of the customer was to buy the item.
5. There is no data-entry error in the file database. The source of the data is reliable and the information in the dataset is correct.

## Methodology

The first step in the analysis was to do ETL (Extract, Transform and Loading) on the data provided in the database. Once the transaction table was explored and fields of interest identified, the data from the trnsact table in pos schema was copied over to a separate schema as follows:

1. trnsact.c1 -> trnsact\_baskets.sku (1:1)
2. trnsact.(c2, c3, c4, c6) -> trnsact\_baskets.basketId (1:1)

This defines a basket as a unique combination of Store, Register, Trannum and SaleDate

1. trnsact.c6 -> trnsact\_baskets.datetran

The saledate is used to subset the data.

1. trnsact.c2 -> trnsact\_baskets.store

Store is used to explore the data

1. All the ‘return’ transactions were ignored during copying over the data.

Once this new table was formed, indexes were added to make the queries perform faster. Once this was done, data was explored.

After the exploratory analysis, the data was subset by taking the transactional data for the month of August 2005. This was because of the following reasons:

1. There were ~120 million transactions in the trnsact table spanning over a year. This amount of data couldn’t fit the memory. The August 2005 data contained ~9 million transactions which is ~7.5% of all the transactions. This could fit the memory leaving space for calculations.
2. This data would capture the most recent and popular SKUs. These SKUs will be of most interest to the retail chain for increasing the sales.

The subset of data was then reshaped for association rule mining using orange3-associate Python library. The fpgrowth package was used to perform the analysis and yielded the association rules. Top 52 rules were then selected based on min\_support of 600 and minimum confidence of 0.2. Then top 31 among these were selected by looking for high values for lift, confidence and support.

## Analysis

The following association rules were discovered after doing the analysis:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Antecedent SKUs** | **Antecedent Dept** | **Antecedent Brand** | **Consequent SKUs** | **Consequent Dept** | **Consequent Brand** | **Support** | **Confidence** | **Lift** |
| ['341363'] | ['4400'] | ['SOUL MEN'] | ['7641362'] | [] | [] | 1111 | 0.872741555 | 1519.587 |
| ['7641362'] | [] | [] | ['341363'] | ['4400'] | ['SOUL MEN'] | 1111 | 0.444933921 | 1519.587 |
| ['959823'] | ['2200'] | ['LANCOME'] | ['4980033'] | ['2200'] | ['LANCOME'] | 611 | 0.96984127 | 1233.275 |
| ['869823'] | ['2200'] | ['LANCOME'] | ['4980033'] | ['2200'] | ['LANCOME'] | 661 | 0.953823954 | 1212.907 |
| ['2716578', '3908011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3988011'] | ['800'] | ['CLINIQUE'] | 630 | 0.801526718 | 1000.514 |
| ['3908011', '3988011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['2716578'] | ['800'] | ['CLINIQUE'] | 630 | 0.782608696 | 945.1498 |
| ['3908011'] | ['800'] | ['CLINIQUE'] | ['3998011', '2726578'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | 974 | 0.290486132 | 757.1603 |
| ['3998011', '2726578'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3908011'] | ['800'] | ['CLINIQUE'] | 974 | 0.583932854 | 757.1603 |
| ['2716578', '3988011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3908011'] | ['800'] | ['CLINIQUE'] | 630 | 0.576395242 | 747.3866 |
| ['3690654', '3898011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3968011'] | ['800'] | ['CLINIQUE'] | 825 | 0.825825826 | 629.1281 |
| ['994478'] | ['6400'] | ['DESIGNER'] | ['6931514'] | [] | [] | 679 | 0.591463415 | 543.8873 |
| ['6931514'] | [] | [] | ['944478'] | ['6400'] | ['DESIGNER'] | 1526 | 0.322758037 | 530.9314 |
| ['944478'] | ['6400'] | ['DESIGNER'] | ['6931514'] | [] | [] | 1526 | 0.577374196 | 530.9314 |
| ['3998011', '3908011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['2726578'] | ['800'] | ['CLINIQUE'] | 974 | 0.777955272 | 507.854 |
| ['3908011', '2726578'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3998011'] | ['800'] | ['CLINIQUE'] | 974 | 0.781074579 | 470.7985 |
| ['3968011', '3898011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3690654'] | ['800'] | ['CLINIQUE'] | 825 | 0.587188612 | 451.5234 |
| ['2801257'] | ['6400'] | ['ROMANCE'] | ['6931514'] | [] | [] | 1910 | 0.468596663 | 430.9037 |
| ['6931514'] | [] | [] | ['2801257'] | ['6400'] | ['ROMANCE'] | 1910 | 0.403976311 | 430.9037 |
| ['3968011', '3690654'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3898011'] | ['800'] | ['CLINIQUE'] | 825 | 0.674019608 | 412.6782 |
| ['3524026', '3978011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3898011'] | ['800'] | ['CLINIQUE'] | 1643 | 0.632409546 | 387.2018 |
| ['3898011'] | ['800'] | ['CLINIQUE'] | ['3524026', '3978011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | 1643 | 0.231375863 | 387.2018 |
| ['2716578'] | ['800'] | ['CLINIQUE'] | ['3988011'] | ['800'] | ['CLINIQUE'] | 1093 | 0.303611111 | 378.9856 |
| ['3988011'] | ['800'] | ['CLINIQUE'] | ['2716578'] | ['800'] | ['CLINIQUE'] | 1093 | 0.313809934 | 378.9856 |
| ['3988011'] | ['800'] | ['CLINIQUE'] | ['3908011'] | ['800'] | ['CLINIQUE'] | 805 | 0.231122595 | 299.6866 |
| ['3908011'] | ['800'] | ['CLINIQUE'] | ['3988011'] | ['800'] | ['CLINIQUE'] | 805 | 0.240083507 | 299.6866 |
| ['2716578'] | ['800'] | ['CLINIQUE'] | ['3908011'] | ['800'] | ['CLINIQUE'] | 786 | 0.218333333 | 283.1033 |
| ['3908011'] | ['800'] | ['CLINIQUE'] | ['2716578'] | ['800'] | ['CLINIQUE'] | 786 | 0.23441694 | 283.1033 |
| ['3908011'] | ['800'] | ['CLINIQUE'] | ['2726578'] | ['800'] | ['CLINIQUE'] | 1247 | 0.371905756 | 242.7824 |
| ['3898011', '3524026'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3978011'] | ['800'] | ['CLINIQUE'] | 1643 | 0.787254432 | 238.8679 |
| ['3908011'] | ['800'] | ['CLINIQUE'] | ['3998011'] | ['800'] | ['CLINIQUE'] | 1252 | 0.373396958 | 225.0678 |
| ['6601061'] | ['2200'] | ['LANCOME'] | ['5528349'] | ['2200'] | ['LANCOME'] | 798 | 0.59596714 | 200.6101 |
| ['3898011', '3978011'] | ['800', '800'] | ['CLINIQUE', 'CLINIQUE'] | ['3524026'] | ['800'] | ['CLINIQUE'] | 1643 | 0.658781075 | 179.0557 |
| ['3690654'] | ['800'] | ['CLINIQUE'] | ['3968011'] | ['800'] | ['CLINIQUE'] | 1224 | 0.216483905 | 164.9211 |
| ['3968011'] | ['800'] | ['CLINIQUE'] | ['3690654'] | ['800'] | ['CLINIQUE'] | 1224 | 0.214473454 | 164.9211 |
| ['3998011'] | ['800'] | ['CLINIQUE'] | ['2726578'] | ['800'] | ['CLINIQUE'] | 1668 | 0.231249134 | 150.9609 |
| ['2726578'] | ['800'] | ['CLINIQUE'] | ['3998011'] | ['800'] | ['CLINIQUE'] | 1668 | 0.25045045 | 150.9609 |
| ['3968011'] | ['800'] | ['CLINIQUE'] | ['3898011'] | ['800'] | ['CLINIQUE'] | 1405 | 0.246188891 | 150.7327 |
| ['3898011'] | ['800'] | ['CLINIQUE'] | ['3978011'] | ['800'] | ['CLINIQUE'] | 2494 | 0.351218138 | 106.5662 |
| ['8798636'] | [] | [] | ['5528349'] | ['2200'] | ['LANCOME'] | 1036 | 0.240874215 | 81.08131 |
| ['3898011'] | ['800'] | ['CLINIQUE'] | ['3524026'] | ['800'] | ['CLINIQUE'] | 2087 | 0.293902267 | 79.8822 |
| ['4010830'] | ['2200'] | ['LANCOME'] | ['4311399'] | ['2200'] | ['LANCOME'] | 1344 | 0.283963659 | 41.24221 |
| ['4010830'] | ['2200'] | ['LANCOME'] | ['4291399'] | ['2200'] | ['LANCOME'] | 1207 | 0.255017959 | 38.60511 |
| ['6618353'] | ['2200'] | ['LANCOME'] | ['4291399'] | ['2200'] | ['LANCOME'] | 624 | 0.230683919 | 34.92138 |
| ['6618353'] | ['2200'] | ['LANCOME'] | ['4311399'] | ['2200'] | ['LANCOME'] | 600 | 0.22181146 | 32.21537 |
| ['3537981'] | ['2200'] | ['LANCOME'] | ['4291399'] | ['2200'] | ['LANCOME'] | 648 | 0.212250246 | 32.13085 |
| ['3537981'] | ['2200'] | ['LANCOME'] | ['4311399'] | ['2200'] | ['LANCOME'] | 660 | 0.216180806 | 31.39759 |
| ['4108011'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 5115 | 0.256854474 | 17.15081 |
| ['3524026'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 3428 | 0.214303576 | 14.30958 |
| ['2072671'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 872 | 0.211650485 | 14.13243 |
| ['4138348'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 711 | 0.20917917 | 13.96741 |
| ['5681064'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 724 | 0.204577564 | 13.66015 |
| ['6318344'] | ['800'] | ['CLINIQUE'] | ['5650750'] | ['800'] | ['CLINIQUE'] | 1598 | 0.204452405 | 13.65179 |

There were some SKUs which were not found in the pos.skuinfo table. These suspicious SKUs were eliminated from further consideration in the analysis. Based on 20 moves, the rules in green were selected. Interestingly, most of these rules belong to the same department ‘CLINIQUE’. Also, the rules seem to involve a lot of repeating SKUs indicating a lot of correlation (high lift) as well.

The minimum support of the entire dataset was small, mainly due to a lot of SKUs and the number of transactions. A few rules were also selected based on high support values (>2000), but with low confidence (~0.25). They still have reasonably high Lift values which means they have good amount of correlation.

Overall, most of the rules have high confidence (>0.6). This means that there is a fairly high chance that the purchase of the Antecedent would trigger a purchase of the Consequent. Also, there are some rules for which the lift is in 4 digits. This indicates a very strong correlation between the two sides.

## Conclusions

Association Rules analysis yielded 31 association rules which have the most promise of triggering compulsive buying action from the consumer. These rules should be incorporated by Dillard’s retail chain to improve sales of its products.

Most of the association rules indicated restructuring planograms for CLINIQUE brand in the same department (800).

## Next Steps

Next steps in the analysis would be to calculate the optimized 20 moves which will be based on association rules and cost of movement. This will require going through the selected rules and collating these rules to find out the moves. Once the moves are decided, there could be a cost analysis and aesthetics of the store to be considered. This will require further data about the store and its operations.

Once the above analysis is done, for the moves which are possible, the items should be moved together in the stores. For the moves which are not possible, there could be small offer placed on the items to incentivize the consumers to walk through the stores. This might also trigger other compulsive purchases.