Implementing Association Rules to Dillard's POS Transaction Data

IEMS 308 – HW 2

Alex Merryman

EXECUTIVE SUMMARY

Dillard's should incorporate the obtained association rules into their planogram rearrangement considerations. It is clear that certain products are frequently bought together, and Dillard's can potentially profit from taking advantage of these relationships by grouping products together on the store floors according to their association rules. Doing so would enable customers to more quickly find products that other customers have bought together, and also implicitly recommend other products to customers looking at antecedent products.

PROBLEM STATEMENT

Dillard's is considering rearranging the location of products on their store floors and wants to understand how their customers purchase items relative to one another in order to best position their products. However, budgetary constraints have limited the maximum number of products that can be moved to 20, so we are tasked with finding the top 100 candidate SKUs for rearrangement.

ASSUMPTIONS

Market Basket

Conducting association rule analysis requires analyzing market basket contents to uncover affinities between items and item groupings. As such, it is necessary to first define a market basket and translate the transaction data into useable market basket data.

For this project, a market basket was defined as a collection of two or more SKUs associated with a particular *store*, *register*, *trannum*, and *saledate* (also, *stype* = P since we want to look at purchases rather than returns).

Data

As the schema diagram and attribute description table were misaligned, it was necessary to make several assumptions regarding the data, particularly for the TRNSACT table. Below is a list of the TRNSACT table headers mapped to attribute names based on best judgement:

- c1 = SKU
- c2 = STORE
- c3 = REGISTER
- c4 = TRANNUM
- c5 = SEQ
- c6 = SALEDATE
- c7 = STYPE
- c8 = QUANTITY
- c9 = ORGPRICE
- c10 = SPRICE

c11 = unknown/redundant
c12 = INTERID
c13 = MIC
c14 = unknown/redundant

METHODOLOGY

Sampling Strategy

As the database is larger than 230 GB, it was necessary to devise a data sampling strategy in order to work with a more manageable subset of the data. Thus, 2005-07-12 (July 12, 2005) was chosen as a representative day for the following reasons:

Retail sales are highly seasonal in nature. In order to account for this seasonality, historical monthly department store sales figures were obtained from the US Census Bureau[1]. An analysis conducted on the 1992-2005 data determined that average monthly sales in July were closest to the overall average monthly sales (see Figure X below); thus, July could thus be used as a representative month upon which to conduct association rule analysis. July 2005 sales were also closest to average monthly sales in all of 2005, further validating its usage as a representative month.

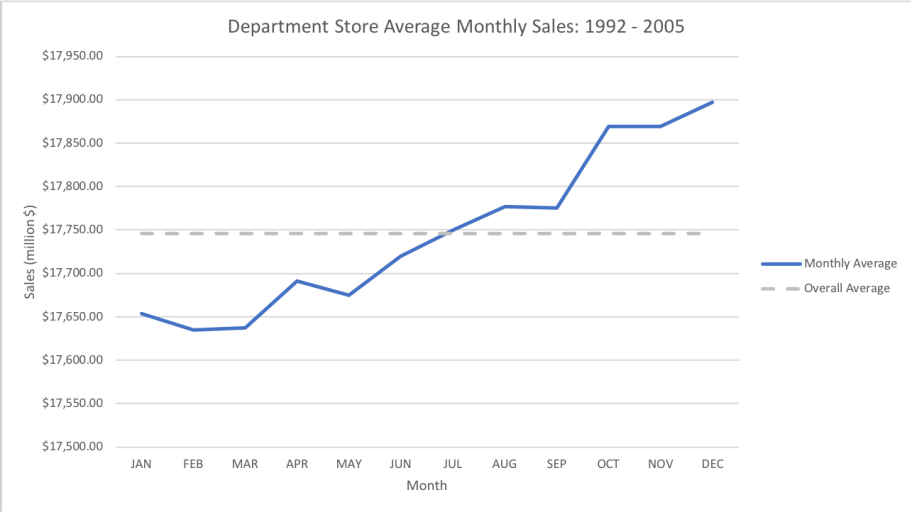


Figure X: Department Store Average Monthly Sales from 1992 – 2005

Department store revenues are also highly influenced by sales (planned product discounts), which are often aligned with major holidays. The only[2] major holiday in July is Independence Day (July 4), hence the decision to select July 12, which occurs more than 1 week after July 4. Further, July 12, 2005 lies on a Tuesday, which is likely more representative of sales during the overall week than a Saturday or Sunday.

[1] https://www.census.gov/retail/marts/www/timeseries.html
[2] https://www.timeanddate.com/holidays/us/

Finally, selecting July 12, 2005 avoids data from the time period of Hurricane Katrina (August 2005), a major US disaster that may have influenced product availability, prices, and consumer purchasing behavior, particularly in the Gulf Coast and Southern region where a large number of Dillard's stores are located.

Data Cleaning & Preprocessing

After selecting the sample date, the results were inserted into a table *trsnact* in my schema (afm900_*schema*) for faster analysis later. Then, the entire table (all 250,000+ records) were queried as a Pandas dataframe for faster manipulation.

Columns c11 and c14 were deemed unknown or redundant and ignored throughout the analysis.

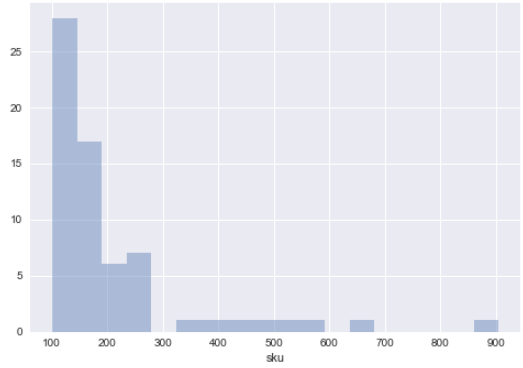
Data were cast to their proper types, most of which were string due to the primary key's immutable nature:

'sku': str,
'store': str,
'register': str
'trannum': str
'seq': str,
'stype': str
'qty': np.int
'orgprice': np.float
'sprice': np.float
'interid': str
'mic': str

The type of the column date was converted to date.

Data Exploration

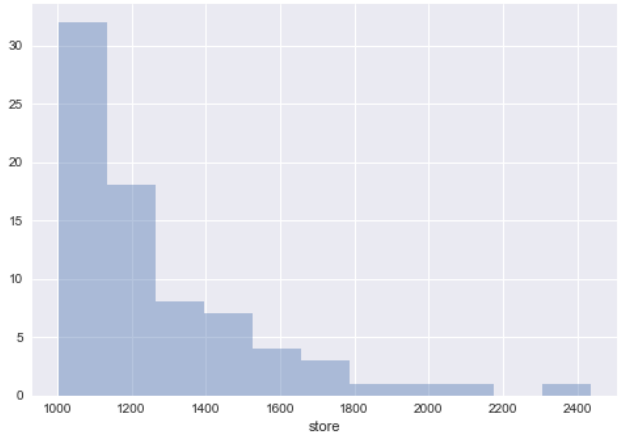
The resulting data were explored for overall insights.



SKU distribution among SKUs appearing in at least 100 records.

The most frequently occurring SKUs are summarized in the chart below:

SKU	Count	Brand
9460294	905	--
5278362	650	SUMMER S
3348362	554	SUMMER S
4108011	517	CLINIQUE
5528349	474	LANCOME
4610447	416	CLINIQUE
3524026	397	CLINIQUE
3978011	343	CLINIQUE
2783996	272	LANCOME
7228362	261	--



Store distribution among stores appearing in at least 1000 records.

Clearly, the vast majority of transactions take place at a few select stores, yielding a Pareto-like distribution.

In order to further shrink the sample to a more manageable size, only transactions at the top 20 highest-volume stores (by number of transactions) on 2005-07-12 were considered (top 10 shown below):

store	city	state	zip	Volume (# transactions)
9806	MABELVALE	AR	72103	2437
2707	MCALLEN	TX	78501	2050
504	LITTLE ROCK	AR	72205	1999
8402	METAIRIE	LA	70002	1823
9103	LOUISVILLE	KY	40207	1773
1607	DALLAS	TX	75225	1723

2007	SAN ANTONIO	TX	78216	1684
2907	BROWNSVILLE	TX	78526	1621
9304	OKLAHOMA CITY	OK	73118	1603
4307	LUBBOCK	TX	79414	1569

Market Basket Generation & Association Rules

Among these transactions, each unique pairing of store-register-trannum was found, and all unique SKUs associated with each of those unique pairings were added to that transaction's "market basket." Each market basket list of SKUs was joined together into a larger list, and then a one-hot encoded array was generated using the OnehotTransactions method from the mlxtend library. The dataframe of frequent itemsets was generated with the apriori function from mlxtend, and finally the association rules were generated using the association_rules function. The rules were sorted in descending order by the lift metric in order to find the items that occurred together in the most statistically significant manner, and SKUs in the top rules were selected as candidates to move within the store. The top 100 candidate SKUs are:

'1098342',	'4752472',	'697924',
'1196926',	'4772472',	'6989904',
'1322480',	'5079905',	'7108437',
'1332480',	'5096618',	'7190174',
'1426926',	'5146781',	'7258437',
'1696435',	'5196781',	'7439924',
'1706435',	'5256618',	'748635',
'1761637',	'5453561',	'7638632',
'1801637',	'5554617',	'7648632',
'1811637',	'5564617',	'7658632',
'2169036',	'5674617',	'7669501',
'2219036',	'5698434',	'7673560',
'2327335',	'5758473',	'7679501',
'2377335',	'5769904',	'768635',
'2397335',	'6082521',	'7838608',
'2476989',	'6152521',	'808695',
'2594124',	'6192521',	'8158459',
'2606394',	'6209962',	'818363',
'2616394',	'6292521',	'8283229',
'267924',	'6470353',	'828635',
'2928137',	'6490353',	'8357626',
'323561',	'6530353',	'8453229',
'3243090',	'6550353',	'8508322',
'3273090',	'6642521',	'8568532',
'3456396',	'668460',	'858363',
'3874124',	'6713560',	'8753044',
'4142521',	'6742521',	'878635',
'4462521',	'6752521',	'888635',
'4512521',	'6949904',	'8897755',
'4516618',	'6968434',	'9150547',

'9218572',	'9600684',	'9770531',
'9460196',	'9687626',	'9858694',
'9460294',	'9687846',	'998342'
'9507634',	'9706988',	

ANALYSIS

Perhaps more important than simply the raw list of candidate SKUs are the association rules for them. The top 10 rules (by the lift metric) are below:

Rule #	Antecedants	Consequents
1	['9858694', '8897755']	['2169036', '808695', '9218572', '2928137', '7439924']
2	['5758473', '9218572', '2928137', '8897755']	['9687846', '7838608', '2219036']
3	['9218572', '2928137', '9858694', '5758473', '9687846', '8897755']	['2169036', '808695', '7838608', '7439924']
4	['5758473', '8897755', '7439924']	['2169036', '808695', '2219036']
5	['5758473', '9218572', '9687846', '2928137', '7838608']	['2219036', '8897755']
6	['808695', '9687846', '2928137', '9858694']	['9218572', '7838608', '7439924']
7	['9687846', '9218572', '2219036']	['808695', '2928137']
8	['2169036', '9218572']	['9858694']
9	['5758473']	['808695', '9687846']
10	['9687846', '2928137']	['2219036']

Note: Each of these rules have Support=0.000450789, Confidence=1, and Lift=2218.333333. With confidence=1, that means that these rules hold true 100% of the time in the dataset.

Using Rule #10 as an example, customers who bought SKUs 9687846 and 2928137 also bought SKU 2219036.

CONCLUSIONS

While Dillard's now has 100 SKUs to consider for changing their planograms, we recommend referring to the obtained association rules, summarized in the 'assoc_rules.csv' file. This file contains rows of associations rules, such as those in the table above, which provide insights into how to best prioritize the planogram project. Without more in-depth information into Dillard's consumers' buying habits, we generally recommend arranging the store layout such that Consequent SKUs are near their Antecedent SKUs, although our concerns with these recommendations are addressed in Next Steps.

NEXT STEPS

Sampling

Ideally, the testing sample would contain 7 dates (one of each day of the week) from each fiscal quarter – or even better, every month – in a year. Sampling data in this manner, while time-consuming, would provide a fuller picture of the seasonality of retail sales. The SKUs obtained

from the association rules in this project, while valuable in their own right, are mostly summer clothing items, which are meaningless to extrapolate to colder months.

Furthermore, retail sales during the holiday months of November and December contribute on average nearly 17% of the sector's overall annual sales. Transaction data during these months is unlike transaction data during the rest of the year, and thus requires special attention beyond the scope of this project.

Similarly, additional association rule analysis could be conducted to discover insights into transactions occurring in the days leading up to major retail holidays such as Labor Day.

Finally, sampling could involve a clustering analysis to initially group stores that have similar characteristics (median income in surrounding zip codes, population size serviced, average/median transaction value, average/median number of daily customers/transactions, etc.). Then, association rule analysis could be conducted on a sample from each cluster to derive more granular insights and enable tailoring the planograms to the cluster's average customer.

Market Basket Item Quantity

The market baskets in this analysis only indicated whether a SKU appeared in at least one transaction, and ignored the quantity of SKUs purchased together. Further analysis should examine whether the quantity purchased of each basket item plays a role in predicting which items will be bought together.

Purchases vs. Returns

This project only looked at purchase transactions, not return transactions. However, returns may contain valuable insights as well. Thus, further analysis should be conducted on this. For example, SKUs that were later returned could be disregarded from the market basket. Additionally, association rule analysis could be conducted on returned items, for example: when bought with products X, Y, how often was Z returned? Were certain products returned together? These results could illustrate a more complete picture of customers' purchasing habits, and also indicate whether customers were unhappy with certain products they purchased.

Effect of Changing Planograms

From the association rules, the effect on revenue, profit, store traffic, etc. of rearranging products on the floor is unclear. While arranging similarly-bought products together may increase customer convenience, it may also prevent customers from finding new products and ultimately hurt Dillard's bottom line. For example, Walmart famously locates milk and eggs (common household staples, and likely two of their most frequently sold products) at the back of all of their stores, so customers are forced to walk past hundreds of other products on their way to pick up the milk and eggs they originally came for, and thus greatly increasing the likelihood that they will purchase more than just milk and eggs. A more thorough study into Dillard's data should determine whether Dillard's has similar types of "commodity" products

that are frequently purchased, and then examine the potential effects of locating those products in the store such that customers are exposed to additional products. The most frequently sold SKUs are good preliminary candidates for this exploration.