**Homework 02**

**ASSOCIATION RULE MINING AND CLUSTERING**

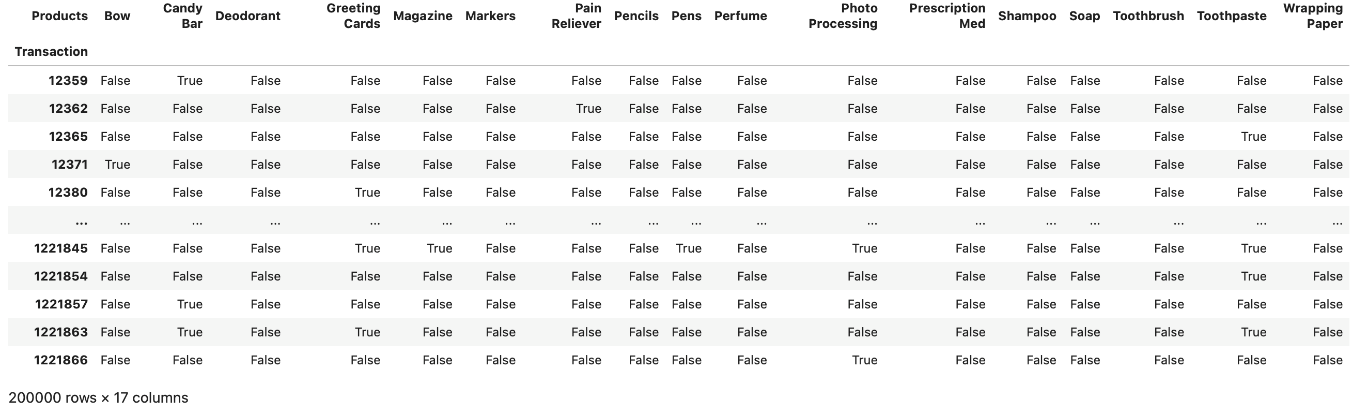
**BUAN 6383.001**

**Syam Menon**

Mihir Hirave (MDH230000)

**PART I: ANALYZING TRANSACTIONS**

1. **Read in the data and generate a file in which every row represents a transaction, with True identifying items that were part of that transaction, and False identifying items that were not (as in the example from class).**



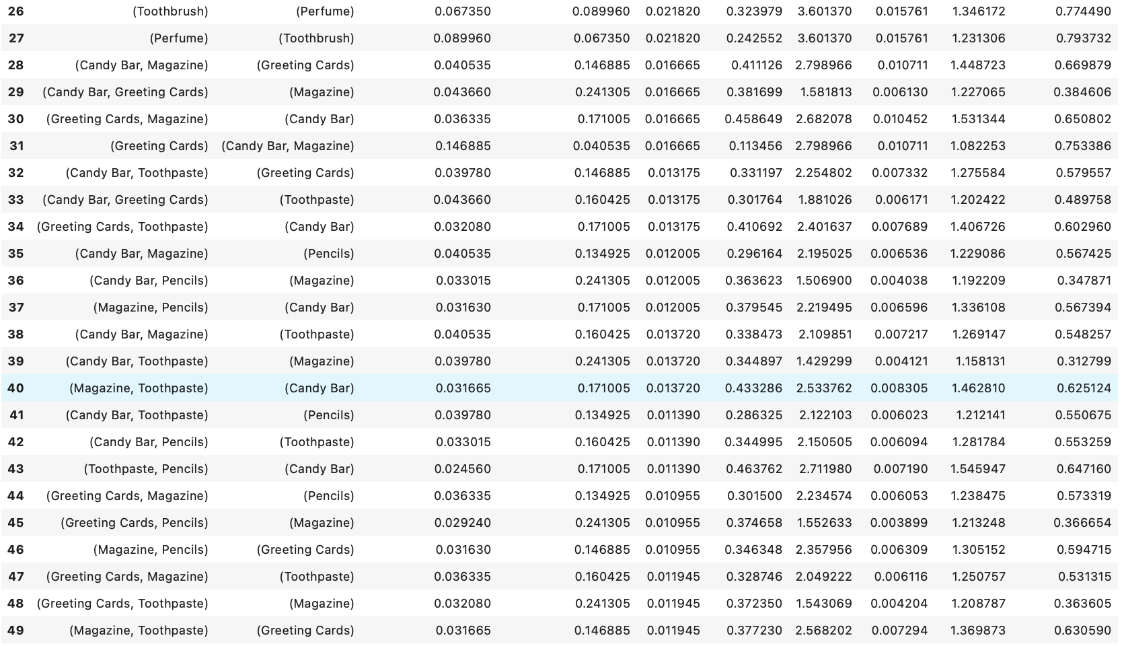
1. **Identify the frequent itemsets using a minimum support threshold of 1%. How many itemsets are frequent?**

* There are 40 itemsets which have a support threshold of 1%



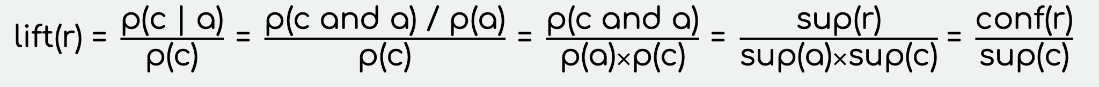
1. **Identify all association rules with a minimum confidence of 10%. How many rules are generated?**

* 50 association rules are generated with confidence threshold of 10%



1. **Which rules have the highest lift? Using the results from the previous questions, show exactly how this lift value was calculated for one of the rules with highest lift.**

* Rule 26 and 27 have highest lift



1. **For the same rule, show how leverage and conviction were obtained.**

**Leverage is same for both the rules**

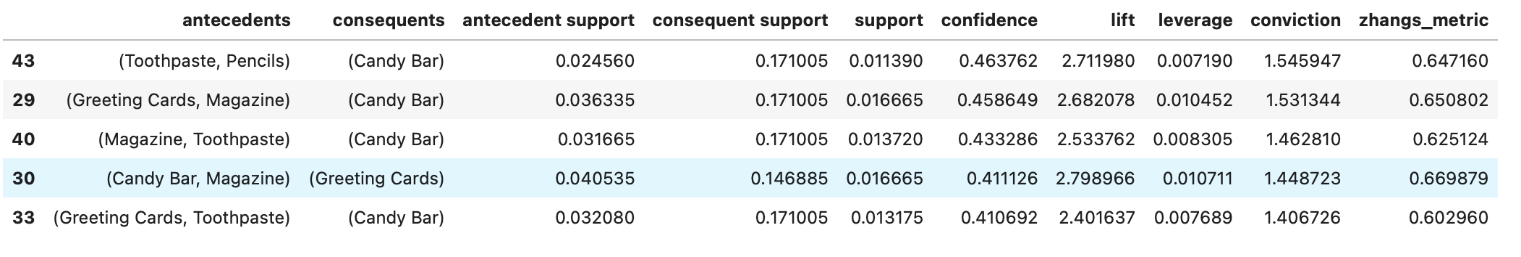
This suggests a slight positive relationship between perfume and toothbru

For rule 27:

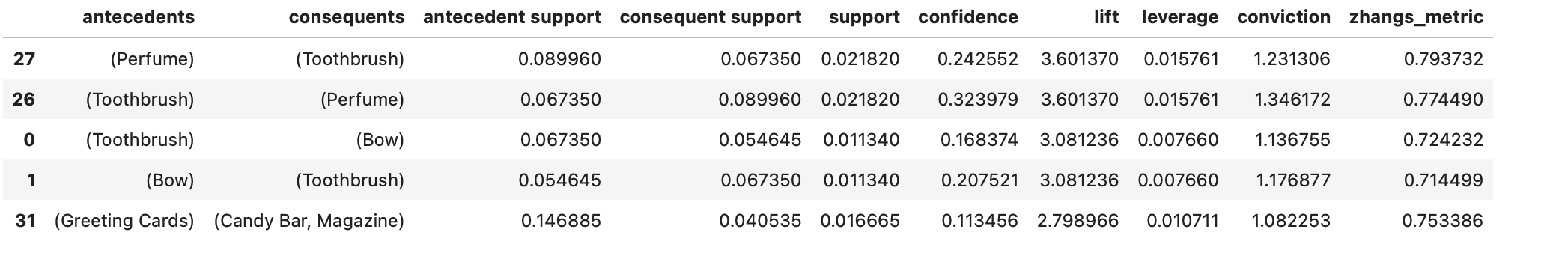
For rule 26:

Conviction = 0.91002/ 0.67602 = **1.346172**

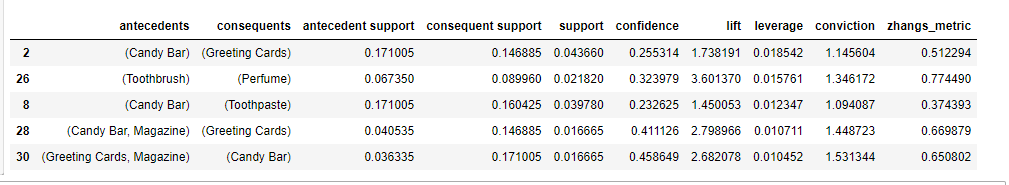
1. Interpret and discuss the 5 rules with
   1. the highest confidence



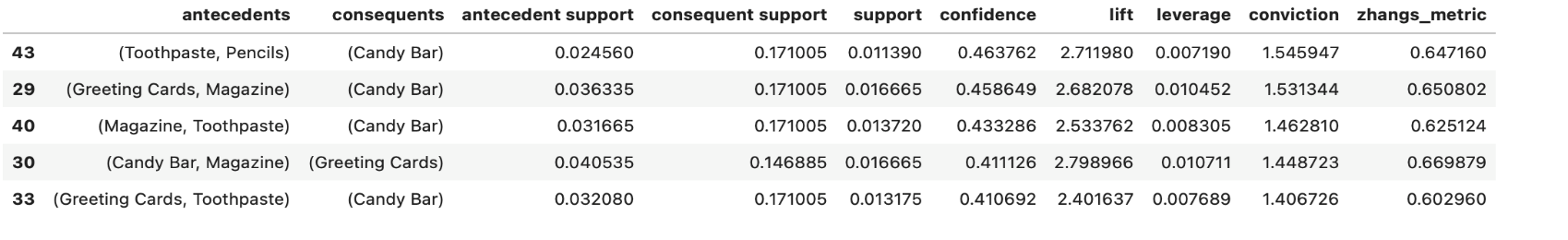
* In each rule, the confidence values are around 40-46%, indicating somewhat strong associations between the antecedents and the consequents.
* Customers who buy certain combinations of items (antecedents) have a good chance of also buying the specified consequent items.
* The same 4 items appear several times in the top 5 rules, sometimes simply switching up what is considered an antecedent versus a consequent.
* 5/5 of these rules include a Candy Bar. Candy Bars appear in 17% of the transactions.
* Magazines, Greeting Cards and Toothpaste each appear in 3/5 rules.
* The most frequently bought items in the store are Magazines (24%), Candy Bars (17%), Toothpaste (16%), and Greeting Cards (15%), in that order.
* While their frequency could be alarming, we can notice that the confidence of each rule (range: 0.41-0.46) is usually higher than the support of the consequents. So, these rules are likely to help us make correct associations and increase the chance of purchasing the consequents to ~40%.
* Lift is >1 in all rules – hence, we can assume they all have a positive correlation.
* Conviction ranges from 1.4-1.55. This tells us that, when we assume the antecedents and consequents of the rules above are related, if these two events are actually independent, we’d be wrong about 1.4-1.55 times more often.
  1. the highest lift



* Some insights we can gather from these results are:
* - Candy Bars appear in 3/5 rules as a consequent.
* - Toothbrushes appear in 2/5 rules.
* - Magazines appear in 2/5 rules.
* - Candy Bars and Magazines items are some of the most frequently sold items at the store.
* - Lift is >1 in all rules – hence, we can assume they all have a positive correlation.
* - Confidence is higher than the consequent support – this means that anytime the antecedents are purchased, the chance of the consequent getting purchased increases.
* - This is the second time that {pencils, toothpaste} --> {candy bar}, makes an appearance.
* - Rule where {greeting cards, magazines}-->{candy bar}, also made an appearance in the last sort
  1. the highest leverage, and



* Some insights we can gather from these results are:
* Candy Bar appears in 4/5 rules.
* Greeting Cards appear in 3/5 rules.
* {Perfume} --> {Toothbrush} is an interesting rule because it includes items, we have not encountered in our top association rules yet.
* Both items are less frequently purchased than the other products that consistently show up in our analysis. For instance, toothbrushes are bought about 7% of the time. But when someone has bought perfume, the chance of a toothbrush purchase more than triples, to 24%.
* - While we do not have access to the prices of these items, we must recognize that perfumes are often pricier than toothbrushes (pricier than the top 4 products, like candy bars or magazines). Hence, this rule is extra appealing due to its sale of a higher-end product.
* - Through a more practical lens, we noticed that toothpaste and perfume fall within the realm of “self-care” and “hygiene.” A person who takes care of their dental hygiene enough to get a new toothbrush is likely to also care about other aspects of their hygiene, like their perceived smell. Hence, a rational hypothesis is that the customers who buy perfume and toothbrushes could also be interested.
  1. the highest conviction.



Some insights we can gather from these results are:

* We have seen every single one of these rules before in our analysis. In fact, these rules include only the top 5 most frequently sold products in the store.
* Most of these rules have Candy Bar as their consequent.
* Confidence is higher than the consequent support for every single one of these rules. Hence, the consequents have a better chance of being purchased when they are bundled with the antecedents than when they are not.
* Overall, we reviewed a little less than 20 rules. Most of these 20 rules included the same 4 products (Magazine, Candy Bar, Toothpaste, Greeting Cards) which are the ones that are most frequently purchased. Hence, it feels redundant to utilize 10+ rules trying to predict the sale of a Candy Bar or a Magazine when it is already likely that they will purchase it.

1. **Do any of these metrics seem preferable to the others for this dataset? Discuss why or why not.**

The metrics that seem the most preferable to this dataset are lift and confidence. Confidence tells us the likelihood that we will purchase the consequent given we are also purchasing the antecedent. Conviction, on the other hand, indicates the influence of antecedent on our decision not to purchase the consequent. Thus, confidence is a better metric. Considering there is so much data and items have low support visa vis the entire dataset, lift is preferable to leverage, as it takes the ratio and shows more significant results.

1. **If you were in charge of these departments, how would you use the results of this analysis to come up with a strategic plan? Explain your reasoning. This question is open ended, and I am looking for innovative thinking.**

The initial phase of our strategy focuses on enhancing sales of our top-selling items by strategically co-locating them. Notably, Candy Bars and Magazines are strategically positioned at checkout areas to encourage impulse purchases of lower-priced items during waiting times. Thus, we propose, as both a test and preliminary action, to similarly position greeting cards adjacent to cashier stations. Should this trial boost greeting card sales, our next move would involve setting up comparable displays in other parts of the store experiencing less foot traffic. Locations for these displays include areas like photo development or the pharmacy, aiming to meet the same success criteria.

The latter phase of our strategy aims to spotlight bundles incorporating items with lower sales frequencies, such as perfumes, toothbrushes, or shampoos. Unlike rapidly consumed items like Candy Bars, Magazines, and Greeting Cards, which facilitate frequent purchases, self-care products like soaps, shampoos, or perfumes have a longer usage cycle. They depend not on spontaneous, minor buys but on dedicated consumers who regularly invest in these products. Consequently, we propose introducing coupons that incentivize buying one of these "durable" items for customers who purchase these "durable" goods. Our goal is to instill a routine among our patrons of committing to "durable" products, potentially prompting them to explore additional "durable" items.