Interesting Eats at the Local Bakery using Association Rule Mining

Victoria Griffin
victoriakgriffin@lewisu.edu

DATA-51000-002
Data Mining and Analytics
Lewis University

I. INTRODUCTION

In the course of data mining education, it is necessary to learn the most common algorithms used in the profession. These algorithms aid in organizing data, solving complex problems, and, in the end, attaining the answers that are sought after. One common category of algorithms is called association rule mining, and the steps taken to prepare the data for association rule mining will be discussed throughout the entirety of this analysis. The purpose of this assignment was to generate association rules using a chosen dataset and rank them by various metrics to find the most interesting and useful rules based on a combination of said metrics.

For this paper, I used a dataset from the website github.com (originally found on Kaggle.com) called *The Bread Basket*[5]. Within the parameters that were given, the dataset is medium sized with a decent number of attributes that has not already been formatted for association rule mining. I was able to narrow the search down to a topic that I am interested in: purchases from a bakery in Edinburgh, Scotland. I enjoy spending my time in bakeries and cafes, and I wanted to see if there were any novel patterns that could be discovered from this data.

This dataset was chosen for its size: over 20,000 observations and five attributes. This is a good size set for someone very new to data mining, such as myself. The data is not perfectly tied to association rule mining, but can be easily applied, which makes for good practice. The dataset has a creative commons license for the public domain which allows anyone in the general public to use this data. It was originally published by Luis Alarcon in 2018 through MIT. As stated already, the origin of this data is from a bakery in Scotland. The bakery donated the data for public use sometime within the past five years (the timestamps of the purchases start in October 2016 and end in September 2017). Although I could have chosen a dataset closer to my concentration of bioinformatics/computational biology, I wanted to step away and do something more personal.

As a concept, association rule mining can aid in finding novel and/or interesting patterns not seen before, and can be applied to almost any dataset. It is most commonly known for analyzing grocery/retail purchases, but it is now increasingly used for analysis of genomic data, other biological data, survey data, textual data, and much more.

There are different algorithms that can be used in association rule mining. A few common ones to mention are: Apriori algorithm, PCY algorithm (which is similar to Apriori, but with a different first pass to take advantage of the computer's memory), and another similar version of Apriori called MSApriori (that utilizes a concept called minimum item support to make sure our frequent itemsets are novel and interesting)[1][2][3]. For this project, I used Orange's association rule widget, which applies the FP-growth algorithm for its data analysis[6].

When using association rule mining algorithms, we are looking for very specific metrics to aid in finding frequent itemsets. My analysis of The Bread Basket dataset used the following metrics: support, confidence, lift and leverage. We take into account the different metrics in order to find frequent itemsets that would be novel and interesting [4]. More will be said about this in Section IV.

The sections of this report cover different stages of the analysis. Within Section II, the dataset is described more in detail with information ranging from the types of variables, total missing values observed, distributions of variables, and other significant information. Section III explains the methodology of the analysis in detail and the tools utilized to carry it out. In Section IV, I dive into the results from the association rule mining and discuss the findings. Section V wraps up the analysis with final thoughts and future recommendations/implications.

II. DATA DESCRIPTION

The size of The Bread Basket dataset is approximately 976 kB[1]. The dataset is fairly new, and is a collection of purchase data from October of 2016 to September of 2017. I believe that the data is still relevant and useful for analysis, even though it is more than five years old. The data contains exactly 20,507 rows (of over 9000 transactions), and five different attributes. Overall, there was no missing data, which made the data very easy to work with. Due to the small number, none of the attributes were excluded from the analysis. Table 1 shows the attribute, data type, and gives an example with a more detailed description. Table 2 shows the descriptive statistics of the dataset.

Table 1: Summary of Bread Basket Dataset

Attribute	**Type**	**Example Value**	**Description**
Transaction	Nominal (primary key)	2	Record identifier
Item	Nominal (string)	bread	Bakery items sold - 94 different choices
date_time	Ordinal (string)	30-10-2016 09:58	Time stamp of purchase
period_day	Nominal (string)	morning	Morning, afternoon, evening, night
weekday_weekend	Binary (string)	weekend	Weekend or weekday

Table 2: Statistical Description of the Dataset

Attribute	**Percent Missing**	**Count**	**Mean**	**Standard Deviation**	**Min/Max**
Transaction	0%	20507	4976.2	2796.2	1/9684
Item	0%	20507	N/A	N/A	N/A
date_time	0%	20507	N/A	N/A	N/A
period_day	0%	20507	N/A	N/A	N/A
weekday_weekend	0%	20507	N/A	N/A	N/A

III. METHODOLOGY

In order to understand the specifics of the dataset, I conducted an initial dataset investigation using Python (version 3.9.8) through the IDE Spider (version 5.1.5) along with the Pandas library (version 1.5.0)[5]. I utilized the Pandas data analysis for attaining a surface level understanding of the data types and amounts of missing values present in the dataset. I also used Pandas for a statistical description that included mean, standard deviation, and min/max of the dataset.

This dataset did not contain any missing or incorrect data, so cleaning of the dataset was not needed. It is important to mention that cleaning must happen before the next step can occur. Some data might already be structured for automatic application to the association rule mining algorithm, but unfortunately most data sets are not.

For the rest of my methodology I used Orange. In order to prepare a dataset to be used with association rule mining an extra step must occur. For this I used the discretize application to convert the data so it could be used correctly with the association rule mining algorithm (Fig. 1).

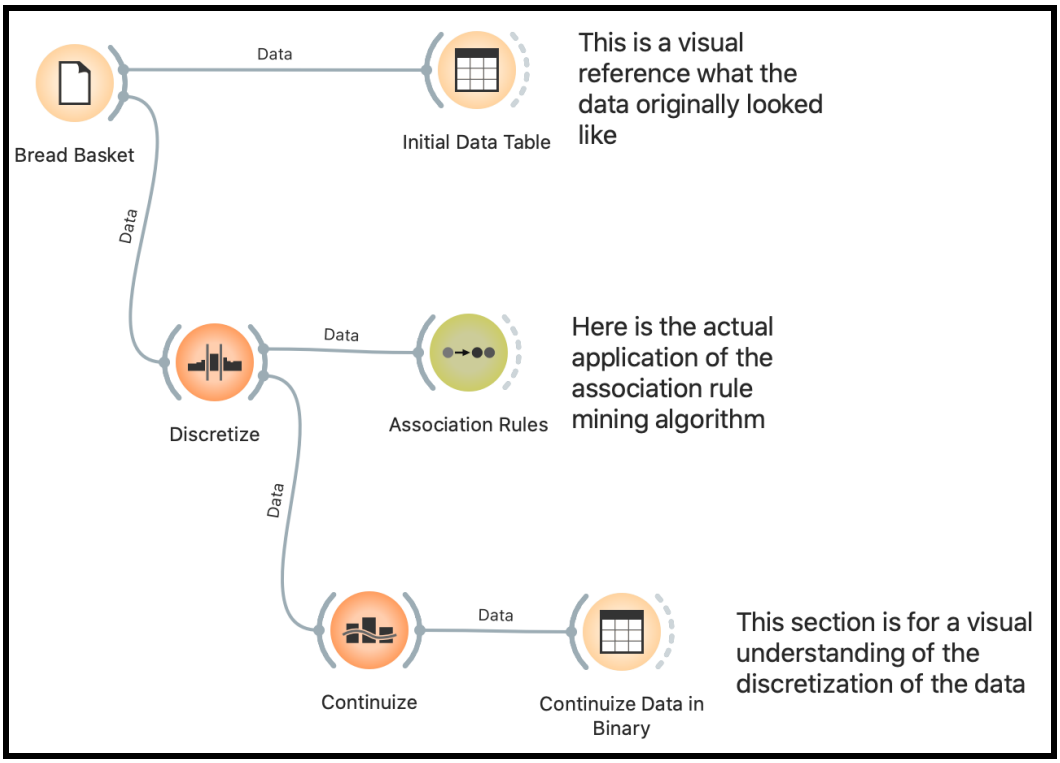

Figure 1: Visual Representation of Orange3 Process as Described in Methodology

To get a good idea of what the data looked like visually after applying the discretization, I used the continuize application within Orange and attached a data table app next to it to see the data manipulated into discrete categories (Fig. 1). The continuized data table showed 20,507 instances with 99 attributes and 1 meta attribute (the 'date_time' of the purchase). The discretize widget converts the data from its original form into discrete buckets of different values. For my data, most of the values in each bucket were binary, labeled either 0 or 1. The only exception was the transaction number, which was not binary.

After discretization of my data, I ran the data through the association rule mining algorithm. The association rule mining algorithm used by Orange is the FP-growth algorithm[6]. Essentially, the FP-growth algorithm uses two passes to analyze the data. During the first pass, the algorithm counts the occurrences of each item and stores the counts in a header table. During the second pass, it creates a frequent-pattern (FP) tree that organizes by descending order of frequency. If the minimum item support is too low, it will not be included in the tree. It is important to note that growth begins at the bottom of the tree.[6]

From the resulting FP-growth report, I carefully sorted through the following specific metrics: support, confidence, lift, and leverage[4].

Before stating the results of my report, I first must explain the four different metrics used and how each specifically aided in my results and final conclusions.

The first metric to consider is support, as seen in equation (1). Support is how popular an itemset is. It is a measurement of the proportion of how many baskets that item is found in versus the total number of baskets that exist. The result will be seen as a fraction. Of all the metrics used, I weighted support the least important when looking at all the metrics together.

$$\text{Support} = \frac{\text{\# of baskets containing item}}{\text{total \# of baskets}} \tag{1}$$

Equation 1: Mathematical equation for support

Confidence is the next metric to consider. Confidence is shown in equation (2) below. Confidence is how likely it is for item "A" (or group of items) to be associated with item "B" (or group of items). It is usually expressed as $\{A \rightarrow B\}$. This is measured by the proportion of times A also appears with B divided by the total instances of A. We are looking initially for a high confidence report. For my dataset, I was looking for at least 30% confidence (0.3+). For me a high confidence encouraged a deeper look into the itemset.

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Support } (A \cup B)}{\text{Support } (A)} \tag{2}$$

Equation 2: Mathematical equation for confidence

We must be careful in looking at only single metrics results. For example, confidence can be very useful as an initial indicator of association, but it could over-inflate the importance of the association. If the first item is very popular, and the second item is not, the calculation for confidence might make it look like the second item is just as popular. To help prevent overinflation of both items we can also look at another metric, called lift, as seen in equation (3).

Lift is how likely item B is to be associated with item A while also taking into account how popular items A and B are (so as not to overinflate popularity). A lift value higher than one means that item B is highly likely to be found with item A. A lift value less than one means it is highly unlikely those items would be found together.

$$\text{Lift } (A \rightarrow B) = \frac{\text{support } (A \cup B)}{\text{support } (A) \times \text{support } (B)} \tag{3}$$

Equation 3: Mathematical equation for lift

The last metric I used in my analysis was leverage, as seen in equation (4). Leverage measures the correlation between the items by comparing the support of items under independent assumption within the dataset. Leverage is very similar to lift with one important difference. Lift uses ratio, and leverage uses difference. It is good to look at both lift and leverage together, because leverage favors frequent itemsets with higher support, and lift will find associations between more rare itemsets. If leverage is zero, items A and B do not have any association or relationship. If leverage is greater than one, the observed support is greater than the expected support, which means the correlation between them is positive. If leverage is less than zero, the observed support is less than the expected support, and the correlation is negative.

$$\text{Leverage } (A \rightarrow B) = \text{Support } (A \cup B) - (\text{Support } (A) \times \text{Support } (B)) \tag{4}$$

Equation 4: Mathematical equation for leverage

It was through analysis of these metrics that I was able to find a few interesting and useful associations.

Fig. 2 shows my methodology in a step by step visualization. There are five main parts to my methodology with details for what was occurring at each step.

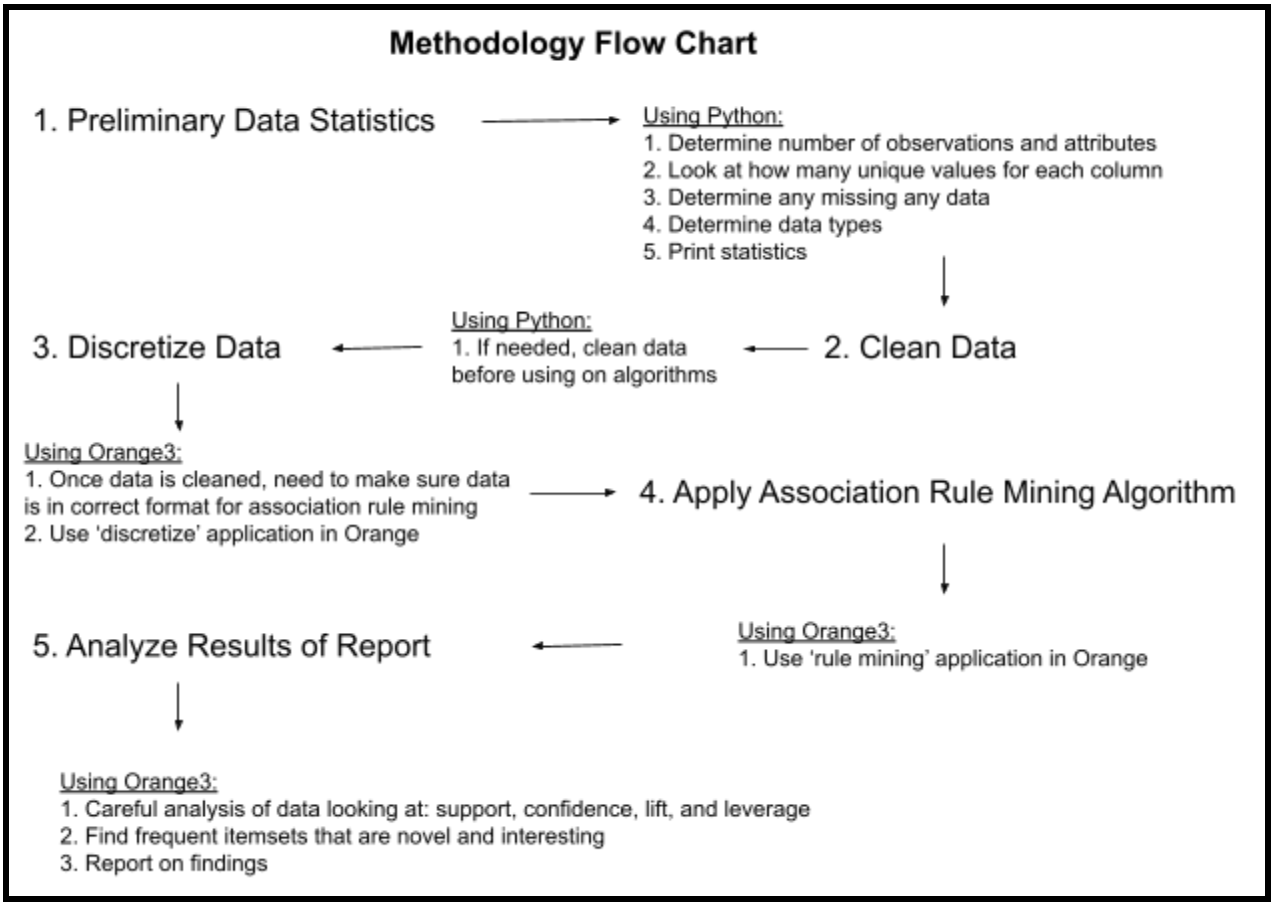

Figure 2: Methods Scheme Layout of beginning to end of Methodology

IV. RESULTS AND DISCUSSION

I played around with my metrics quite a bit. In the end I felt confident with a support of over 10% (I didn't think my data had quite as many observations as some other datasets - only 20k - so accepted a lower support). Of all four metrics, I looked least at support, since support only focuses on popularity of an item throughout the set. I think the interesting itemsets wouldn't necessarily need to have a high support. Although I kept my support on the lower side, I was looking for at least 30% confidence with a lift of more than 1.00. When comparing lift and leverage together, I was looking for an itemset that had a positive leverage (positive number) with a lift greater than 1.00. This would point to a strong correlation between items, as well as the possibility of the itemset being not very common (i.e. rare itemsets). The results of my top item sets can be seen in Table 3.

To make sure I investigated everything thoroughly, I analyzed more than just one item versus one item. I looked at two items to one item, three items to one, and two items to two. Sometimes I found some very interesting and useful results. Most of the time I did not.

Table 3: Top Frequent Itemsets

Frequent Itemset	Support	Confidence	Lift	Leverage
{coffee → weekday}	17.3%	64.8%	1.037	0.006
{afternoon → weekday}	35.5%	62.9%	1.007	0.002
{weekend, bread → morning}	3%	50.5%	1.233	0.006
{Sandwich, weekday → afternoon}	2.2%	86.3%	1.530	0.007
{cake, weekday → afternoon}	2.1%	71.4%	1.266	0.004
{coffee, morning → transaction < 3435.5, weekday}	3%	24%	1.146	0.004

Table 3: Shows a combination of different itemsets with listed support, confidence, lift, and leverage

Since my dataset included data from a bakery, I felt that most of the itemsets were intuitive to anyone that frequents bakeries and cafes often. There are a few interesting itemsets to point out. It would make sense that coffee is most strongly associated with weekdays, and looking further into Table 3 I could see that coffee purchased in the mornings had an association with cold weekdays (transactions <3435.5 means late autumn in the dataset). There was also an association between weekend bread purchases being in the morning. Also, sandwiches purchased during weekdays were associated with the afternoon. This makes sense in terms of what types of foods and beverages are consumed at different times of the day. Coffee tends to be a morning item, and sandwiches are usually eaten in the afternoon. My results confirmed that assumption.

Fig. 3 is a visual of my metrics results. I chose to show this scatter plot as confirmation that my analysis explained above was accurate and valid. Fig. 3 shows a good number of itemsets that have high confidence (greater than 20%), with a positive leverage (meaning positive correlation), as well as a good amount of green and yellow points (which indicates lift greater than 1.00). I chose not to include support in the scatter plot, as I think the other metrics hold higher weights than support.

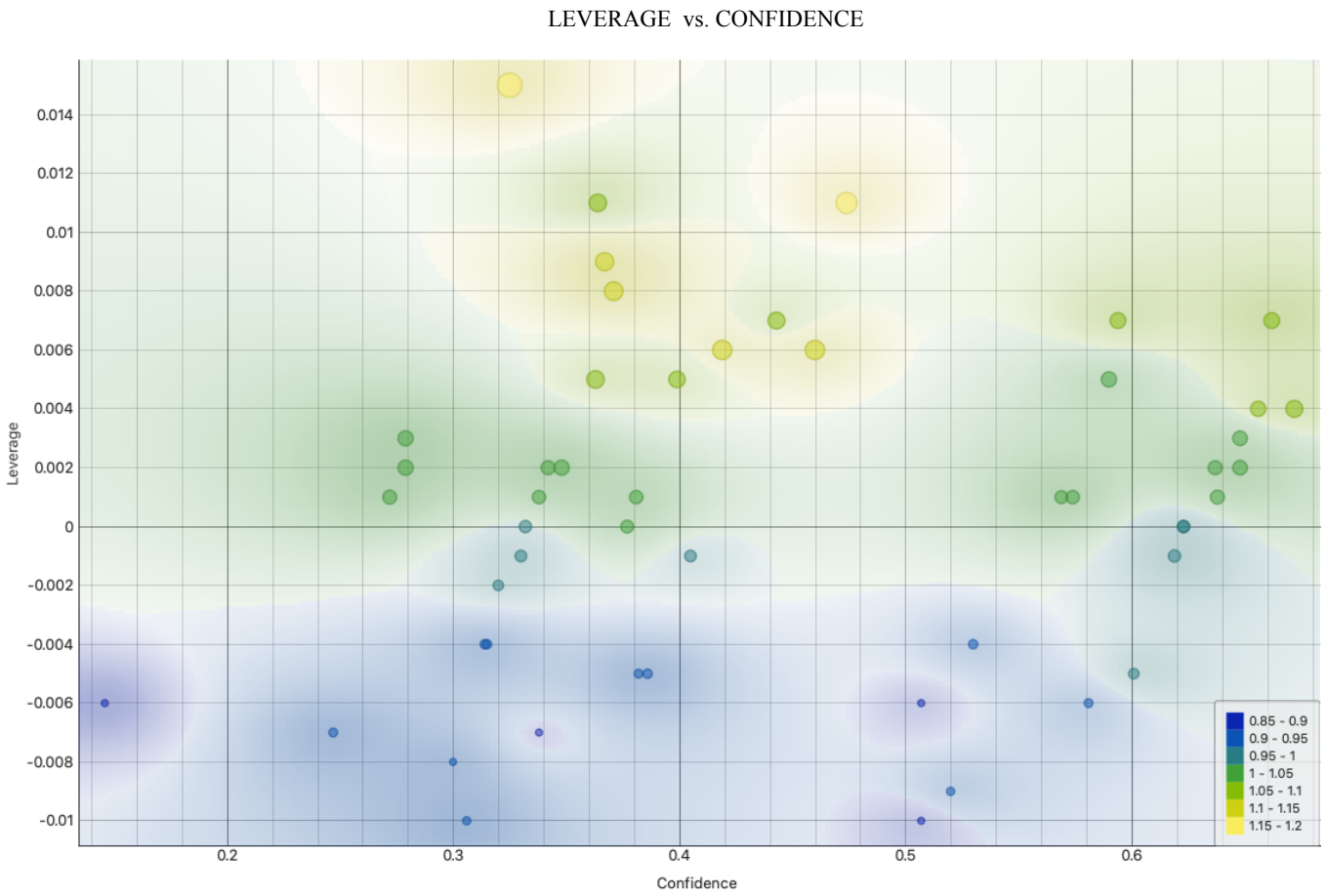

Figure 3: The following figure shows a scatter plot of association rule mining with leverage on the y-axis and confidence on the x-axis. Lift is used for the weights of the points. Lift is emphasized with color and size.

One surprising association for me personally was seeing cake purchased during the week associated with the afternoon. I would not have thought that cake would be a frequent item sold during the afternoons during weekdays. I would have thought it would be purchased more frequently on weekend mornings. This either could be that I have been purchasing cake wrong all my life or possibly my idea of what cake is might not be Scotland's idea of cake, or possibly their cafe culture is different than American cafe culture. As a bakery owner, I would make sure to have fresh cakes displayed after lunch during the week.

There were a few very strange and interesting associations I noticed that I would like to mention: t-shirts associated with weekday purchases, vegan feasts associated with the evening, and valentine day cards associated with weekday afternoons. The support and leverage were near nonexistent for these because the purchases were so few in comparison to the total purchases. With some of these more rare associations, I could have spent countless hours sifting through data to look more closely at those. I feel that an owner of a bakery would like to know these things as well. Although they aren't purchased often (low profit), maybe there could be a way to get more purchases in the future based on the associations that are found in this dataset.

Although most of these rules are not immensely interesting, they would be quite useful to the owner of the bakery/cafe.

V. CONCLUSION

The purpose of this assignment was to generate association rules using a chosen dataset and rank them by various metrics to find the most interesting and useful rules based on a combination of said metrics.

It took a while to find a dataset I felt comfortable using. Once found, the methodology was quite simple: preliminary data analysis, cleaning of data, discretization of the data, and application of the association rule algorithm. The most difficult part of this assignment was the analysis using the four metrics. Figuring out the best way to read the data, which metrics to pay more attention to, and which could be allowed a lower result was not easy. In the end I lowered the support in favor of a higher confidence as well as a positive combination of lift and leverage. This was my "recipe" for analysis and results. I looked at itemsets of different sizes and found the same associations in the simpler itemsets as well as the more complicated ones. I am confident that I successfully found interesting and useful itemsets within the data.

In terms of real-life application, I think this analysis can be useful for a bakery when it comes to many scenarios including: how to plan a menu, when to bake and/or prepare certain items for freshness, how much coffee or hot water to have ready at certain times of the day, when to schedule staff to make sure there are enough workers for the busy hours, how much to purchase when making weekly orders, what combos or meals to offer when writing the menu, what items to take off the menu and which ones to keep on, etc. The information can save the owner money, time, and effort if they had access to this information frequently.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *ACM SIGMOD International Conference on Management of Data,* 1994 [Online] Available https://rakesh.agrawal-family.com/papers/sigmod93assoc.pdf [Accessed: September 29, 2022]

[2] J.S. Park, M.S. Chen, P. Yu, "An Effective Hash Based Algorithm for Mining Association Rules," *SIGMOD Conference,* 1995 [Online] https://dl.acm.org/doi/pdf/10.1145/568271.223813 [Accessed: September 29, 2022]

[3] B.Ling, W. Hsu, Y. Ma, "Mining Association Rules with Multiple Minimum Supports", *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 1999 [Online] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.112.8948&rep=rep1&type=pdf [Accessed: September 29, 2022]

[4] A. Ng, "Association Rules and the Apriori Algorithm: A Tutorial," kdnuggets.com, 2016 [Online] https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html [Accessed: September 29, 2022]

[5] L. Alarcon "Bread Basket," github.com, 2018 [Online]. Available:https://github.com/luis-alarcon/Kaggle_BreadBasket/commit/3cbc16bfff52d121069e3576d15a11a092011ae5 [Accessed: September 26, 2022]

[6] J. Han, J. Pei, Y. Yin, R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," *Data Mining and Knowledge Discovery,* 2004 [Online] https://www.cs.sfu.ca/~jpei/publications/dami03_fpgrowth.pdf [Accessed: September 30, 2022]