CSCM35: Big Data and Data Mining Coursework 1

Andy Gray
445348

09/04/20

1 Introduction

We have a practical task assigned to us, that is related to the field of data mining. This task aims to use the association rule, a rule-based machine learning technique, to discover interesting relationships within the provided large dataset. We will be creating to code process the data, as well as analysing the results to see if any insights gains and if any possible reasons to why these might be the case.

Data mining is a necessary part of obtaining knowledge through discovery in databases (KDD). KDD is the term used for the overall process. Data mining tasks split into two main categories, which are predictive and descriptive tasks. However, these tasks split further into four core mining task, which is cluster analysis, predictive modelling, anomaly detection and association analysis [11]. We will be focusing on the association analysis within this paper.

We will be applying the apriori and association rule to look at the data as a whole, to see if any patterns emerge comparing the results on the confidence and lift metrics. As well as analysing the itemset for individual countries, these countries are the United Kingdom (Uk), Germany, France and the Republic of Ireland (Éire). We chose these countries as they are the top 4 counties with the number of transactions recorded. We found that France and Éire by there stock in big bulks, while the Uk buys in bulk but by not as much as Éire and France, while Germany tends to buy more individual items.

We will first look at the algorithms used within the proposed solution, explaining how they work and what is the maths formulas driving the algorithm. We will then explain the dataset, and the data preprocessing that occurred, followed by an explanation of the packages used and the parameters set for the algorithms. We will then explain an discuss the results to see if any insights are present. To end, we will be then concluding what we have found.

2 Proposed Solution

2.1 Understanding the Problem

What we need to do is take the provided dataset and perform appropriate data mining techniques on it. To try to find any patterns within it. We will achieve this by using appropriate data mining tools, techniques and algorithms.

To asses, if there are any distinctive patterns within the dataset, we will look at the results when focusing on different metrics as well as look at subsections, for example, country, within the data. We would expect to see a reduced list of items, displaying the antecedents items and the consequents of those items based on what metric we use.

2.2 Packages

We will be using the programming language Python 3 [9], as this allows us to use all the required algorithms needed to analyse the dataset, to check for any trends along with the Pandas library [8]. We will be using the library package MLXtend[10] to be able to get access to the apriori and the association rule algorithm. We will be using Matplotlib's [6] package library for visualising our data, to allow us to be able to get insights and spot possible trends.

2.3 Algorithms Used Explanation

The first algorithm that we used is one that is from the frequent itemset mining methods, called Apriori [5]. Apriori is an unsupervised learning machine learning algorithm proposed by R. Agrawal and R. Srikant in 1994 [2, 4]. The algorithm focuses on using boolean association rules [2] from using prior knowledge of itemsets that contain the frequent properties. Apriori uses a level-wise search, which operates an iterative approach, where k-itemsets get used for exploring $(k+1)$-itemsets [7]. In order to improve efficiency, which will reduce the search space, an important characteristic called the Apriori property needs to be applied [5].

The Apriori property has a two-step process which involves the join and prunes step. For this explanation, F_k represents the k-itemset where L_k represents the candidate for the k-itemset. The process of joining is to generate a new itemset, L_{k+1}, from the F_K itemset. While the pruning stage aims to identify the itemsets in L_{k+1} that are infrequent from k, and then remove them [7]. What indicates if the item is infrequent depends on the support count, which is predefined beforehand. Therefore what the algorithm does, is: Let us assume that $k = 1$ and a support count of 2, we generate a frequent itemset, at first 1, which we will refer to as F_1. What this is doing is scanning the dataset to figure out the count of each occurrence of each item. The next step is the merge, or join, the datasets. Using F_k we can then create L_{k+1}. We then prune the data based on the support count eliminating any data that is infrequent, therefore leaving any data that is classed as frequent, adding it to F_{k+1}. This process is repeated until F_k is empty [7, 5].

The second algorithm that we have used is called the association rule. Rakesh Agrawal, Tomasz Imieliński and Arun Swami developed the algorithm in 1993 [1]. The association rule algorithm is an unsupervised machine learning algorithm [4]. What this algorithm focuses around is the support of the datasets' items and the confidence of the association. The math formula for the support is $support(A \Rightarrow B) = P(A \cup B)$, and the math formula for the confidence is $confidence(A \Rightarrow B) = P(B|A)$. Similar to the apriori, the support count will drop any relationships that do not meet the desired count. The formula to figure out if the relationships meet the support count is $confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}$ [7, 5]. However, the association rule relies on a procedure, like the apriori algorithm, to have been implemented on the dataset first before it can work effectively. While the association rule requires the support threshold, the confidence level, which we can use to make decisions based on the links, can be changed to additional met-

rics. The metric can be several different ones like conviction and leverage, but the other one to the confidence that we will focus on is the lift metric. The metric lift was introduced in 1997 by Sergey Brin, Rajeev Motwani, Jefferey D. Ullman and Shalom Tsur [3]. This metric figures out how the antecedent and consequent of a rule, $A \rightarrow C$, would occur together and not as statistically independent items. The lift score would indicate if A and C are independent by having a score of exactly 1. The math formula for lift is constructed as $lift(A \rightarrow C) = \frac{confidence(A \rightarrow C)}{support(C)}, range[0, \infty]$ [3, 7].

Overall the apriori algorithm will reduce the dataset by pruning it. The amount of pruning depends on the support count threshold that is applied. The output will create the required frequent itemset which the association rule requires. The association rule will then go through the frequent itemset to acquire any patterns of items based on the support count and the metric. In our case, this is the lift or confidence metric.

2.4 Dataset and Data Preprocessing

The dataset we have acquired is a shopping dataset. It is 44MB in size and is in the format of CSV. There are eight attributes, within the dataset, with 541,910 records. The attributes are InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerId, Country. There are 4,335 unique customers, 1,8405 individual invoices, 3,659 unique stock items and 37 unique countries. [Enter Individual countries here?]

The purpose of data preprocessing is to convert any raw data into a format that is appropriate for the following analysis of the data. Preprocessing can involve fusing data from several sources, as well as cleaning the raw data to remove any noise, duplicate observations or ambiguity [11]. The main aim of the preprocessing is to get data that is accurate, complete and consistent, but in the real world, we will usually get inaccurate, incomplete and inconsistent data [5]. The preprocessing stage can also involve just selecting the essential records and features that are desired and are relevant to the set data mining task [11]. We can now see that the main aim of data processing is to clean the data, we achieve this through filling in missing values, identifying or removing outliers, smoothing noisy data, and resolving and data inconsistencies [5].

The dataset had values missing in a number of the columns. The rows that had any missing values, within the features, were removed from the dataset. Also, any rows that had data that was an outlier, within its features, was removed from the dataset. These outliers included minus values. Once we had carried out these data cleaning actions, we then have 396,371 records remaining. The cleaning process indicates that we had removed a total of 145,539 records from the dataset.

Before we could give the apriori algorithm the dataset, we have to perform a data transfer on the dataset. First, we placed the required features into a basket and then performed the data transfer function on it, converting the values into binary values. Grouping the data by quantity using the InvoiceNo and Description feature and then index the values using the InvoiceNo. We then used this basket to feed into the apriori algorithm to create our frequent items dataset.

When we were analysing the data set based on country, we used the same process to transfer the data. However, we had another parameter for the basket that only selected the required data for that country, feeding that basket into the algorithm.

2.5 Parameters

When using the apriori algorithm on the whole dataset, we set the minimal support to 0.2, and we passed through no country filter. We then performed the association rule using both the lift and confidence metric. The minimum support for the lift was the value 5, and for the confidence, we used the value 0.5.

When analysing the dataset by individual countries, we looked at the countries the United Kingdom, Germany, France, Republic of Ireland (Éire) and Spain. When we performed the apriori algorithm, we used a minimum support level of 0.03 for all of the countries, except the UK, which had 0.02. When looking at the countries association rules using the lift metric, we used minimum support of 10 for all except for Germany, where we used a value of 5. For the confidence metric, we applied a minimum support level of 0.5 to all of the countries.

2.6 Visual and Statistical Analysis

When looking at the bar chart in appendix A, we can see that the United Kingdom is the most number of counts. The United Kingdom has a count of 345,005 which is then followed by Germany, with 8,659, France 8,034, Éire on 7,138 and Spain with 2,424 making up the top five. With the rest being between the ranges 2,326 and 9.

When applying the association rule with the lift metric (see appendix B), the antecedents with the hight lift value is 23.863 with the consequents of Roses Regency and the Green Regency teacup and saucer. This result demonstrates that there is a strong link between the antecedents and consequents. However, when looking at the whole dataset with the confidence metric used (see appendix C), the antecedents Roses and Pink Regency teacups and saucers had a confidence level of 0.894 that a consequent of Green Regency Teacup and saucer. This score indicates a high likelihood that some buying these items will buy the Green teacup set.

When we look at the UK's data (see appendix D), we can see that the items with the highest lift are 'Green Regency Teacup and Saucer' with a consequence of 'Pink Regency Teacup and Saucer and Roses Regency Teacup and Saucer'. However, this only has a confidence level of 0.557, while if they are the other way round the confidence level of the Green set being the consequence is 0.89. The table shows that there is a strong link between these items whatever the order they are in, all being around the 0.85 confidence but the most like combination is the one stated previously. These results are evident within the results focusing on the data when the confidence metric is applied. [create appendix for these two results]

When we look at the data for Germany, it is clear that the antecedents and

consequents are one to one item mapping (see appendix F). The antecedents and conquests are also very similar items, 'Spaceboy Cup' matched with 'Spaceboy Bowl', as well as 'Set/6 Red spotty paper cups' with Set/6 red spotty paper plates'. They are showing a strong link between item similarity. However, when looking at the confidence (see appendix G), the antecedents items 'Red retrospot Charlotte bag' and 'Round snack boxes set of 4 Woodland' has a value of 1.00 for the consequent 'Woodland Charlotte bag' but only has a lift value of 7.63, which is about mid-level compared to the others.

The data for France (see appendix H) shows that all the links for the lift are all around the items 'set of 20 red retrospot paper napkins', 'pack of 6 skull paper plates', 'set of 6 spotty paper cups' for antecedents. With the consequence of 'set of 6 spotty paper plates', 'pack of 20 skull paper napkins' which has a lift value of 24.989. The rest are based around these items, alternating between being the antecedents and consequents. Again when reviewing the confidence metric, The same items are near the top with a confidence value of 1 (see appendix I).

In regards to Ireland, the items at the top of the lift table (see appendix J), with a value of 25.283, is dominated by the same items that can interchange between antecedents and consequents. These are 'green regency teacup and saucer', 'regency sugar bowl green', 'regency cakestand 3 tier', 'regency tea plate pink', 'Regency tea plate green', 'Pink regency teacup and saucer', 'regency milk jug pink'. These items are also what dominates the confidence table (see appendix K), with confidence support value of 1. support value of 1.

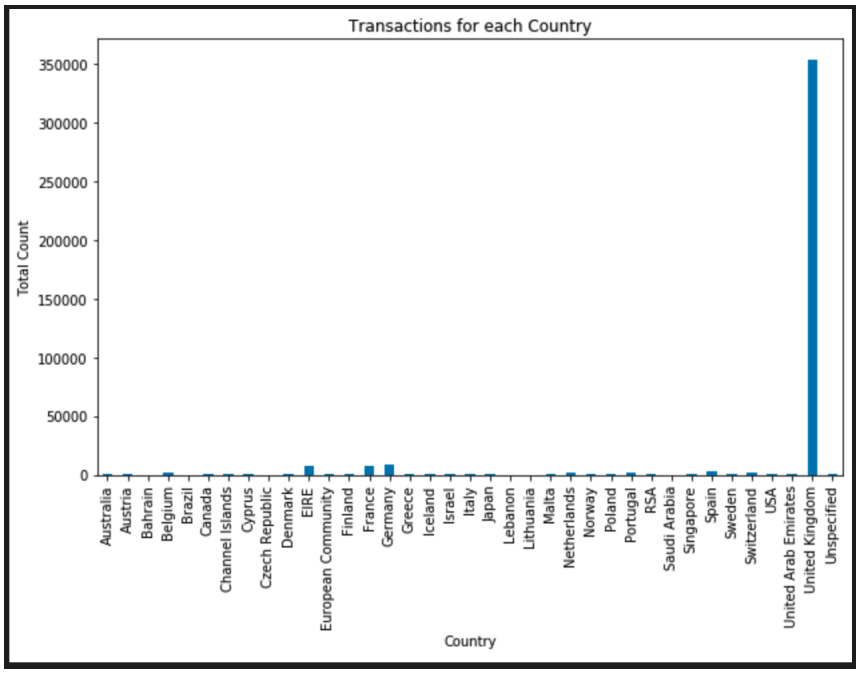
3 Discussion and Conclusion

When comparing the countries transactions, Germany's results indicate that many items get bought individually, or in small bulks, as there are not many items required to provide a consequent, especially for the lift. However, places like France and especially Éire will have about 3 to 5 antecedents to provide a consequent, hinting that many items get bought together often probably in bulk. Compared to the other countries, Germany has very different types of items being bought compared to the rest. The other countries have trends related to teacups and saucers, which is similar to the dataset as a whole, but Éire items are the same items but green. While Germany is more based around woodland bags and children related items like paper cups and plates.

To conclude, in order to run the apriori algorithm, we first needed to prepare the dataset and carry out preprocessing, which involved cleaning the data from any records that were deemed inappropriate for analysing. Once we have done this, we can then place the data into buckets to then encode it and then pass it through the apriori algorithm which then, by using minimum support of 0.5, give us a frequent itemset. Using the frequent itemset with the association rule algorithm provided us with a list of antecedents and consequents items, depending on what metric, lift or confidence, was assigned. From these results, we could see that three of the countries had very similar outcomes as that of the dataset as a whole. However, Germany was different. Mainly have single items rather than several items providing a likely consequence.

Appendices

A Total count for Country



B Lift Table of Items Whole Dataset

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
22	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER)	0.038209	0.029394	0.021190	0.701439	23.863183	0.020302	3.250945
19	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.029394	0.038209	0.021190	0.720887	23.863183	0.020302	3.474549
23	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER)	0.037544	0.023689	0.021190	0.564399	23.825164	0.020301	2.241298
18	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.023689	0.037544	0.021190	0.894495	23.825164	0.020301	9.122408
8	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.038209	0.037544	0.024993	0.827338	22.036408	0.023859	5.574223
9	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.037544	0.038209	0.024993	0.665702	22.036408	0.023859	2.980976
21	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER)	0.042543	0.024993	0.021190	0.498084	19.928786	0.020127	1.942571
20	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.024993	0.042543	0.021190	0.847826	19.928786	0.020127	6.291862
14	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.042543	0.038209	0.023689	0.556833	18.432564	0.022404	2.188318
15	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.038209	0.042543	0.023689	0.784173	18.432564	0.022404	4.436218
10	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.042543	0.037544	0.029394	0.690932	18.403197	0.027797	3.114062
11	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.037544	0.042543	0.029394	0.782923	18.403197	0.027797	4.410686
5	(SPACEBOY LUNCH BOX)	(DOLLY GIRL LUNCH BOX)	0.038250	0.033469	0.023837	0.602273	17.994853	0.021757	2.430135

C Confidence Table of Items Whole Dataset

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
27	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.023689	0.037544	0.021190	0.894495	23.825164	0.020301	9.122408
29	(PINK REGENCY TEACUP AND SAUCER, GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.024993	0.042543	0.021190	0.847826	19.928786	0.020127	6.291862
7	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.038209	0.037544	0.024993	0.827338	22.036408	0.023859	5.574223
22	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.038209	0.042543	0.023689	0.784173	18.432564	0.022404	4.436218
11	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.037544	0.042543	0.029394	0.782923	18.403197	0.027797	4.410686
5	(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.034501	0.041076	0.025156	0.729134	17.750937	0.023739	3.540214
28	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.029394	0.038209	0.021190	0.720887	23.863183	0.020302	3.474549
30	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER)	0.038209	0.029394	0.021190	0.701439	23.863183	0.020302	3.250945
10	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.042543	0.037544	0.029394	0.690932	18.403197	0.027797	3.114062
3	(DOLLY GIRL LUNCH BOX)	(SPACEBOY LUNCH BOX)	0.033469	0.038250	0.023837	0.688312	17.994853	0.021757	3.085613
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.042869	0.047650	0.028797	0.671736	14.097273	0.026754	2.981174
23	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.036892	0.107098	0.024722	0.670103	6.257355	0.020771	2.706632
8	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.037544	0.038209	0.024993	0.665702	22.036408	0.023859	2.980976

D Lift Table of United Kingdom Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
71	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	0.036910	0.023099	0.020566	0.557190	24.122088	0.019713	2.206139
70	(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.023099	0.036910	0.020566	0.890339	24.122088	0.019713	8.782466
72	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	0.029733	0.028708	0.020566	0.691684	24.094329	0.019712	3.150310
69	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.028708	0.029733	0.020566	0.716387	24.094329	0.019712	3.421090
4	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.036910	0.029733	0.024365	0.660131	22.202084	0.023268	2.854025
5	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.029733	0.036910	0.024365	0.819473	22.202084	0.023268	5.334871
73	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER)	0.040890	0.024365	0.020566	0.502950	20.642108	0.019569	1.962850
68	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.024365	0.040890	0.020566	0.844059	20.642108	0.019569	6.150482
6	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.036910	0.040890	0.028708	0.777778	19.021141	0.027198	4.315994
7	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.040890	0.036910	0.028708	0.702065	19.021141	0.027198	3.232551
60	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.029733	0.040890	0.023099	0.776876	18.999094	0.021883	4.298556
61	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.040890	0.029733	0.023099	0.564897	18.999094	0.021883	2.229970
2	(GARDENERS KNEELING PAD KEEP CALM)	(GARDENERS KNEELING PAD CUP OF TEA)	0.044750	0.037014	0.027622	0.617251	16.323179	0.025930	2.513079

E Confidence Table of United Kingdom Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
26	(PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.023099	0.036910	0.020566	0.890339	24.122088	0.019713	8.782466
24	(GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.024365	0.040890	0.020566	0.844059	20.642108	0.019569	6.150482
5	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.029733	0.036910	0.024365	0.819473	22.202084	0.023268	5.334871
6	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.036910	0.040890	0.028708	0.777778	19.021141	0.027198	4.315994
18	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.029733	0.040890	0.023099	0.776876	18.999094	0.021883	4.298556
3	(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.037814	0.044750	0.027622	0.730463	16.323179	0.025930	3.544034
25	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.028708	0.029733	0.020566	0.716387	24.094329	0.019712	3.421090
7	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.040890	0.036910	0.028708	0.702065	19.021141	0.027198	3.232551
28	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	0.029733	0.028708	0.020566	0.691684	24.094329	0.019712	3.150310
20	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.038719	0.113624	0.025813	0.666667	5.867304	0.021413	2.659128
4	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.036910	0.029733	0.024365	0.660131	22.202084	0.023268	2.854025
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.041614	0.045715	0.027381	0.657971	14.392899	0.025478	2.790071
17	(PAPER CHAIN KIT VINTAGE CHRISTMAS)	(PAPER CHAIN KIT 50'S CHRISTMAS)	0.040838	0.056993	0.026355	0.645495	11.325873	0.024028	2.660066

F Lift Table of Germany Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
39	(SPACEBOY CHILDRENS CUP)	(SPACEBOY CHILDRENS BOWL)	0.045147	0.042889	0.038375	0.850000	19.818421	0.036438	6.380737
38	(SPACEBOY CHILDRENS BOWL)	(SPACEBOY CHILDRENS CUP)	0.042889	0.045147	0.038375	0.894737	19.818421	0.036438	9.071106
6	(CHILDRENS CUTLERY SPACEBOY)	(CHILDRENS CUTLERY DOLLY GIRL)	0.049661	0.051919	0.040632	0.818182	15.758893	0.038054	5.214447
7	(CHILDRENS CUTLERY DOLLY GIRL)	(CHILDRENS CUTLERY SPACEBOY)	0.051919	0.049661	0.040632	0.782609	15.758893	0.038054	4.371558
35	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.054176	0.058691	0.047404	0.875000	14.908654	0.044224	7.530474
34	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.058691	0.054176	0.047404	0.807692	14.908654	0.044224	4.918284
10	(JAM JAR WITH PINK LID)	(JAM JAR WITH GREEN LID)	0.065463	0.036117	0.033860	0.517241	14.321121	0.031496	1.996614
11	(JAM JAR WITH GREEN LID)	(JAM JAR WITH PINK LID)	0.036117	0.065463	0.033860	0.937500	14.321121	0.031496	14.952596
8	(COFFEE MUG APPLES DESIGN)	(COFFEE MUG PEARS DESIGN)	0.063205	0.040632	0.036117	0.571429	14.063492	0.033549	2.238525
9	(COFFEE MUG PEARS DESIGN)	(COFFEE MUG APPLES DESIGN)	0.040632	0.063205	0.036117	0.888889	14.063492	0.033549	8.431151
30	(WHITE SPOT RED CERAMIC DRAWER KNOB)	(RED STRIPE CERAMIC DRAWER KNOB)	0.054176	0.047404	0.033860	0.625000	13.184524	0.031292	2.540256
31	(RED STRIPE CERAMIC DRAWER KNOB)	(WHITE SPOT RED CERAMIC DRAWER KNOB)	0.047404	0.054176	0.033860	0.714286	13.184524	0.031292	3.310384
12	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.036117	0.081264	0.033860	0.937500	11.536458	0.030925	14.699774

G Confidence Table of Germany Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
66	(RED RETROSPOT CHARLOTTE BAG, ROUND SNACK BOXES SET OF4 WOODLAND)	(WOODLAND CHARLOTTE BAG)	0.031603	0.130926	0.031603	1.000000	7.637931	0.027465	inf
67	(SPACEBOY LUNCH BOX , ROUND SNACK BOXES SET OF 4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.040632	0.252822	0.038375	0.944444	3.735615	0.028102	13.449210
14	(JAM JAR WITH GREEN LID)	(JAM JAR WITH PINK LID)	0.036117	0.065463	0.033860	0.937500	14.321121	0.031496	14.952596
15	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.036117	0.081264	0.033860	0.937500	11.536458	0.030925	14.699774
52	(CHARLOTTE BAG APPLES DESIGN, ROUND SNACK BOXES SET OF 4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.033860	0.252822	0.031603	0.933333	3.691667	0.023042	11.207675
49	(SPACEBOY CHILDRENS BOWL)	(SPACEBOY CHILDRENS CUP)	0.042889	0.045147	0.038375	0.894737	19.818421	0.036438	9.071106
11	(COFFEE MUG PEARS DESIGN)	(COFFEE MUG APPLES DESIGN)	0.040632	0.063205	0.036117	0.888889	14.063492	0.033549	8.431151
47	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.054176	0.058691	0.047404	0.875000	14.908654	0.044224	7.530474
64	(PLASTERS IN TIN WOODLAND ANIMALS, ROUND SNACK BOXES SET OF 4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.049661	0.252822	0.042889	0.863636	3.415998	0.030334	5.479388
70	(WOODLAND CHARLOTTE BAG, ROUND SNACK BOXES SET OF 4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.047404	0.252822	0.040632	0.857143	3.390306	0.028647	5.230248
50	(SPACEBOY CHILDRENS CUP)	(SPACEBOY CHILDRENS BOWL)	0.045147	0.042889	0.038375	0.850000	19.818421	0.036438	6.380737
32	(RED RETROSPOT CHARLOTTE BAG)	(WOODLAND CHARLOTTE BAG)	0.072235	0.130926	0.060948	0.843750	6.444504	0.051491	5.562077
40	(ROUND SNACK BOXES SET OF 4 FRUITS)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.162528	0.252822	0.135440	0.833333	3.296131	0.094350	4.483070

H Lift Table of France Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
613	(SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS)	0.034301	0.036939	0.031662	0.923077	24.989011	0.030395	12.519789
616	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS)	(SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	0.036939	0.034301	0.031662	0.857143	24.989011	0.030395	6.759894
473	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS)	0.042216	0.036939	0.036939	0.875000	23.687500	0.035380	7.704485
575	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER CUPS, PACK OF 20 SKULL PAPER NAPKINS)	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	0.031662	0.042216	0.031662	1.000000	23.687500	0.030326	inf
582	(PACK OF 6 SKULL PAPER CUPS, PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER PLATES)	0.031662	0.042216	0.031662	1.000000	23.687500	0.030326	inf
587	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER PLATES)	(PACK OF 6 SKULL PAPER CUPS, PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	0.042216	0.031662	0.031662	0.750000	23.687500	0.030326	3.873351
624	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS)	0.042216	0.031662	0.031662	0.750000	23.687500	0.030326	3.873351
594	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER CUPS, PACK OF 20 SKULL PAPER NAPKINS)	0.042216	0.031662	0.031662	0.750000	23.687500	0.030326	3.873351
617	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER PLATES)	(SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	0.042216	0.031662	0.031662	0.750000	23.687500	0.030326	3.873351
612	(SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER PLATES)	0.031662	0.042216	0.031662	1.000000	23.687500	0.030326	inf
468	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS)	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	0.036939	0.042216	0.036939	1.000000	23.687500	0.035380	inf
605	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS)	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	0.031662	0.042216	0.031662	1.000000	23.687500	0.030326	inf
606	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 6 SKULL PAPER PLATES)	(PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	0.034301	0.039578	0.031662	0.923077	23.323077	0.030305	12.485488

I Confidence Table of France Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
415	(SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.031662	0.131926	0.031662	1.0	7.580000	0.027485	inf
403	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS, SET/6 RED SPOTTY PAPER CUPS)	(PACK OF 6 SKULL PAPER PLATES)	0.036939	0.050847	0.036939	1.0	17.227273	0.034795	inf
406	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 20 SKULL PAPER NAPKINS)	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	0.036939	0.042216	0.036939	1.0	23.687500	0.035380	inf
160	(DOLLY GIRL CHILDRENS CUP, SPACEBOY CHILDRENS BOWL)	(DOLLY GIRL CHILDRENS BOWL)	0.034301	0.047493	0.034301	1.0	21.055556	0.032672	inf
158	(PLASTERS IN TIN WOODLAND ANIMALS, CHARLOTTE BAG DOLLY GIRL DESIGN)	(PLASTERS IN TIN SPACEBOY)	0.036939	0.139842	0.036939	1.0	7.150943	0.031774	inf
154	(ALARM CLOCK BAKELIKE RED , PLASTERS IN TIN WOODLAND ANIMALS)	(PLASTERS IN TIN SPACEBOY)	0.031662	0.139842	0.031662	1.0	7.150943	0.027235	inf
152	(PLASTERS IN TIN WOODLAND ANIMALS, ALARM CLOCK BAKELIKE PINK)	(PLASTERS IN TIN SPACEBOY)	0.034301	0.139842	0.034301	1.0	7.150943	0.029504	inf
413	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RETROSPOT PAPER NAPKINS , PACK OF 20 SKULL PAPER NAPKINS)	(SET/6 RED SPOTTY PAPER CUPS)	0.031662	0.142480	0.031662	1.0	7.018519	0.027151	inf
283	(PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.042216	0.131926	0.042216	1.0	7.580000	0.036647	inf
138	(ALARM CLOCK BAKELIKE ORANGE, ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.031662	0.097625	0.031662	1.0	10.243243	0.028571	inf
136	(PLASTERS IN TIN WOODLAND ANIMALS, ALARM CLOCK BAKELIKE GREEN)	(PLASTERS IN TIN SPACEBOY)	0.034301	0.139842	0.034301	1.0	7.150943	0.029504	inf
440	(SET/6 RED SPOTTY PAPER PLATES, PACK OF 6 SKULL PAPER CUPS, PACK OF 6 SKULL PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.036939	0.142480	0.036939	1.0	7.018519	0.031676	inf
443	(PACK OF 6 SKULL PAPER CUPS, PACK OF 6 SKULL PAPER PLATES, SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.036939	0.131926	0.036939	1.0	7.580000	0.032066	inf

J Lift Table of Erie Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
15406	(GREEN REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
19108	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK , ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
19158	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK, ROSES REGENCY TEACUP AND SAUCER , REGENCY SUGAR BOWL GREEN)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
15449	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
19029	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK, ROSES REGENCY TEACUP AND SAUCER , REGENCY SUGAR BOWL GREEN)	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
19079	(GREEN REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK , ROSES REGENCY TEACUP AND SAUCER)	0.035156	0.035156	0.031250	0.888889	25.283951	0.030014	8.683594
19157	(REGGNCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK, ROSES REGGNCY TEACUP AND SAUCER , REGENCY MILK JUG PINK)	0.042969	0.031250	0.031250	0.727273	23.272727	0.029907	3.552083
15448	(REGGNCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN)	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY MILK JUG PINK , REGENCY TEA PLATE PINK)	0.042969	0.031250	0.031250	0.727273	23.272727	0.029907	3.552083
19581	(REGENCY TEA PLATE GREEN , REGENCY TEA PLATE ROSES , REGENCY MILK JUG PINK , PINK REGENCY TEACUP AND SAUCER)	(REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK, ROSES REGENCY TEACUP AND SAUCER)	0.031250	0.042969	0.031250	1.000000	23.272727	0.029907	inf
19830	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE PINK, ROSES REGENCY TEACUP AND SAUCER , REGENCY MILK JUG PINK)	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY SUGAR BOWL GREEN)	0.031250	0.042969	0.031250	1.000000	23.272727	0.029907	inf

K Confidence Table of Eire Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
12461	(GREEN REGENCY TEACUP AND SAUCER, REGENCY TEA PLATE GREEN , REGENCY CAKESTAND 3 TIER, PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER , REGENCY MILK JUG PINK)	(REGENCY SUGAR BOWL GREEN)	0.031250	0.093750	0.031250	1.0	10.666667	0.028320	inf
11961	(REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE ROSES , REGENCY TEAPOT ROSES , ROSES REGENCY TEACUP AND SAUCER)	(REGENCY TEA PLATE GREEN , REGENCY SUGAR BOWL GREEN)	0.039062	0.058594	0.039062	1.0	17.066667	0.036774	inf
7824	(GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE ROSES , PINK REGENCY TEACUP AND SAUCER, REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN)	0.042969	0.082031	0.042969	1.0	12.190476	0.039444	inf
15325	(REGENCY TEA PLATE GREEN , REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE ROSES , REGENCY TEA PLATE PINK, REGENCY TEAPOT ROSES)	(REGENCY SUGAR BOWL GREEN, ROSES REGENCY TEACUP AND SAUCER)	0.031250	0.078125	0.031250	1.0	12.800000	0.028809	inf
10681	(REGENCY TEA PLATE ROSES , REGENCY TEAPOT ROSES , PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(REGENCY TEA PLATE GREEN , REGENCY CAKESTAND 3 TIER)	0.031250	0.070312	0.031250	1.0	14.222222	0.029053	inf
6527	(REGENCY CAKESTAND 3 TIER, REGENCY TEAPOT ROSES , REGENCY TEA PLATE PINK, ROSES REGENCY TEACUP AND SAUCER)	(REGENCY SUGAR BOWL GREEN)	0.035156	0.093750	0.035156	1.0	10.666667	0.031860	inf
2773	(REGENCY TEA PLATE ROSES , REGENCY TEAPOT ROSES , REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN)	0.035156	0.082031	0.035156	1.0	12.190476	0.032272	inf
1773	(REGENCY TEA PLATE GREEN , PINK REGENCY TEACUP AND SAUCER, REGENCY MILK JUG PINK)	(ROSES REGENCY TEACUP AND SAUCER)	0.035156	0.171875	0.035156	1.0	5.818182	0.029114	inf
10678	(REGENCY CAKESTAND 3 TIER, REGENCY TEA PLATE ROSES , REGENCY TEAPOT ROSES , PINK REGENCY TEACUP AND SAUCER)	(REGENCY TEA PLATE GREEN , ROSES REGENCY TEACUP AND SAUCER)	0.031250	0.062500	0.031250	1.0	16.000000	0.029297	inf
10674	(REGENCY TEA PLATE GREEN , REGENCY TEA PLATE ROSES , REGENCY TEAPOT ROSES , PINK REGENCY TEACUP AND SAUCER)	(REGENCY CAKESTAND 3 TIER, ROSES REGENCY TEACUP AND SAUCER)	0.031250	0.113281	0.031250	1.0	8.827586	0.027710	inf
988	(GREEN REGENCY TEACUP AND SAUCER, REGENCY TEA PLATE ROSES , REGENCY TEA PLATE PINK)	(PINK REGENCY TEACUP AND SAUCER)	0.042969	0.109375	0.042969	1.0	9.142857	0.038269	inf
6528	(REGENCY SUGAR BOWL GREEN, REGENCY CAKESTAND 3 TIER, REGENCY TEAPOT ROSES , REGENCY TEA PLATE PINK)	(ROSES REGENCY TEACUP AND SAUCER)	0.035156	0.171875	0.035156	1.0	5.818182	0.029114	inf

References

[1] Agrawal, R., Imieliński, T., and Swami, A. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (1993), pp. 207–216.

[2] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[3] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data* (1997), pp. 255–264.

[4] Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

[5] Han, J., Pei, J., and Kamber, M. *Data mining: concepts and techniques*. Elsevier, 2011.

[6] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in science & engineering 9*, 3 (2007), 90–95.

[7] jingjingslides. jingjingslides slides. In *Slide titles* (2020).

[8] McKinney, W. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (2010), S. van der Walt and J. Millman, Eds., pp. 51 – 56.

[9] Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, Vienna, Austria, 2020.

[10] Raschka, S. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software 3*, 24 (Apr. 2018).

[11] Tan, P.-N., Steinbach, M., and Kumar, V. *Introduction to data mining*. Pearson Education India, 2016.