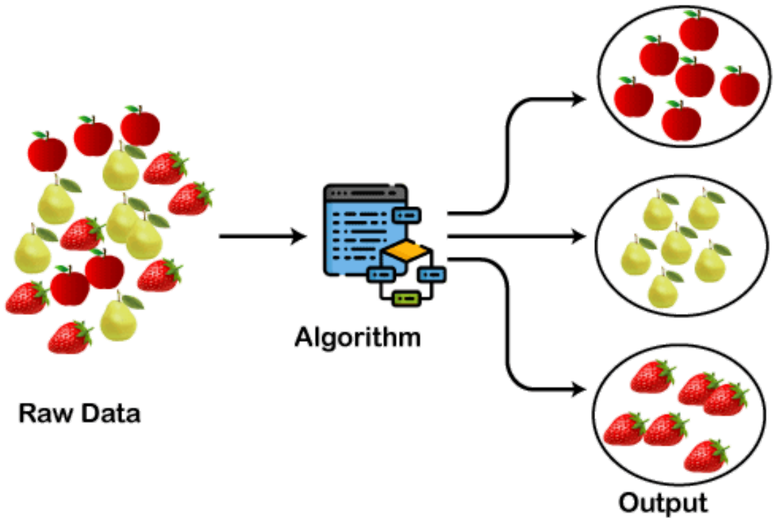


CLUSTERING ALGORITHM

Group Members:
Omkar Satupe (9232)
Ryan Valiaparambil (9237)
Mahek Intwala (9423)

PAPER DETAILS

- Title : A Hybrid Clustering Technique to Propose the Countries for HELP International
- Publication Year: February 2022
- Research paper Link: Click here
- Colab File Implementation: Click here

AGENDA

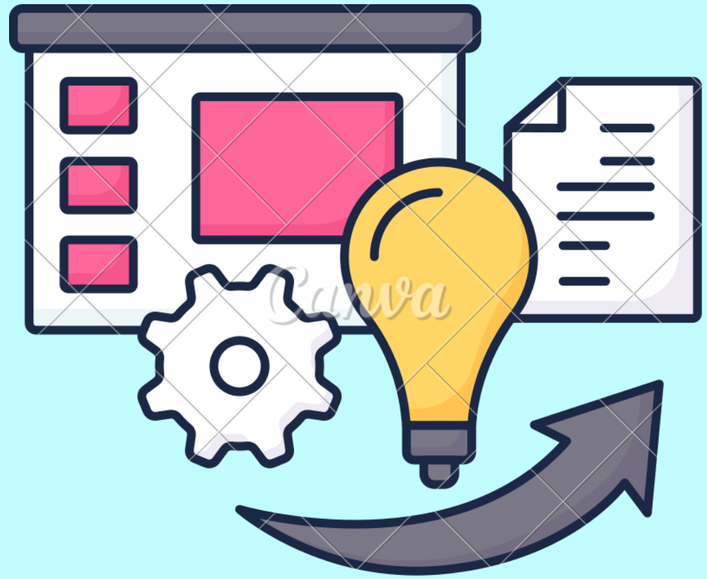
1. Problem Statement
2. Methodology & Algorithms
3. Results and Conclusion
4. Implementation

PROBLEM STATEMENT

- The problem statement for this study revolves around HELP International, a charitable NGO with $10 million in funds, seeking to strategically and effectively allocate these resources to countries in dire need of aid.
- The organization aims to make data-driven decisions on country selection based on socio-economic and health factors that influence overall development.
- To address this challenge, the paper proposes a hybrid clustering technique incorporating K-MEANS clustering and the Farthest First algorithm to cluster countries.
- This unsupervised learning approach identifies and recommends countries most in need of assistance, assisting the NGO's leadership in decision-making and enhancing the accuracy of country grouping based on economic and health factors.

METHODOLOGY

- The paper's methodology involves utilizing clustering techniques, particularly K-means and K-means++, to address the problem of identifying and grouping countries in dire need of aid based on socio-economic and health factors.
- It begins by introducing data mining and the significance of clustering in statistical data analysis.
- The primary objective of clustering is explained as maximizing similarity within clusters while maximizing dissimilarity between them.
- The paper presents the K-means algorithm, detailing its steps, and introduces K-means++ as an enhanced version with a specific initialization method. Furthermore, it outlines a hybrid clustering technique that combines farthest-first and K-means++ to improve clustering accuracy based on socio-economic and health factors.
- This methodology provides a structured approach to help the NGO, HELP International, make data-driven decisions regarding aid distribution.



ALGORITHMS USED : K-means clustering, K++ means Clustering, Hybrid Clustering

PAPER'S RESULTS

Table 2: Number of Countries for each Cluster in Health Dataset

Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	66	52	42	7
Farthest First	35	99	32	1
Hybrid	**73**	**50**	**38**	**6**

Table 3: Number of Countries for each Cluster in Socio-Economic Dataset

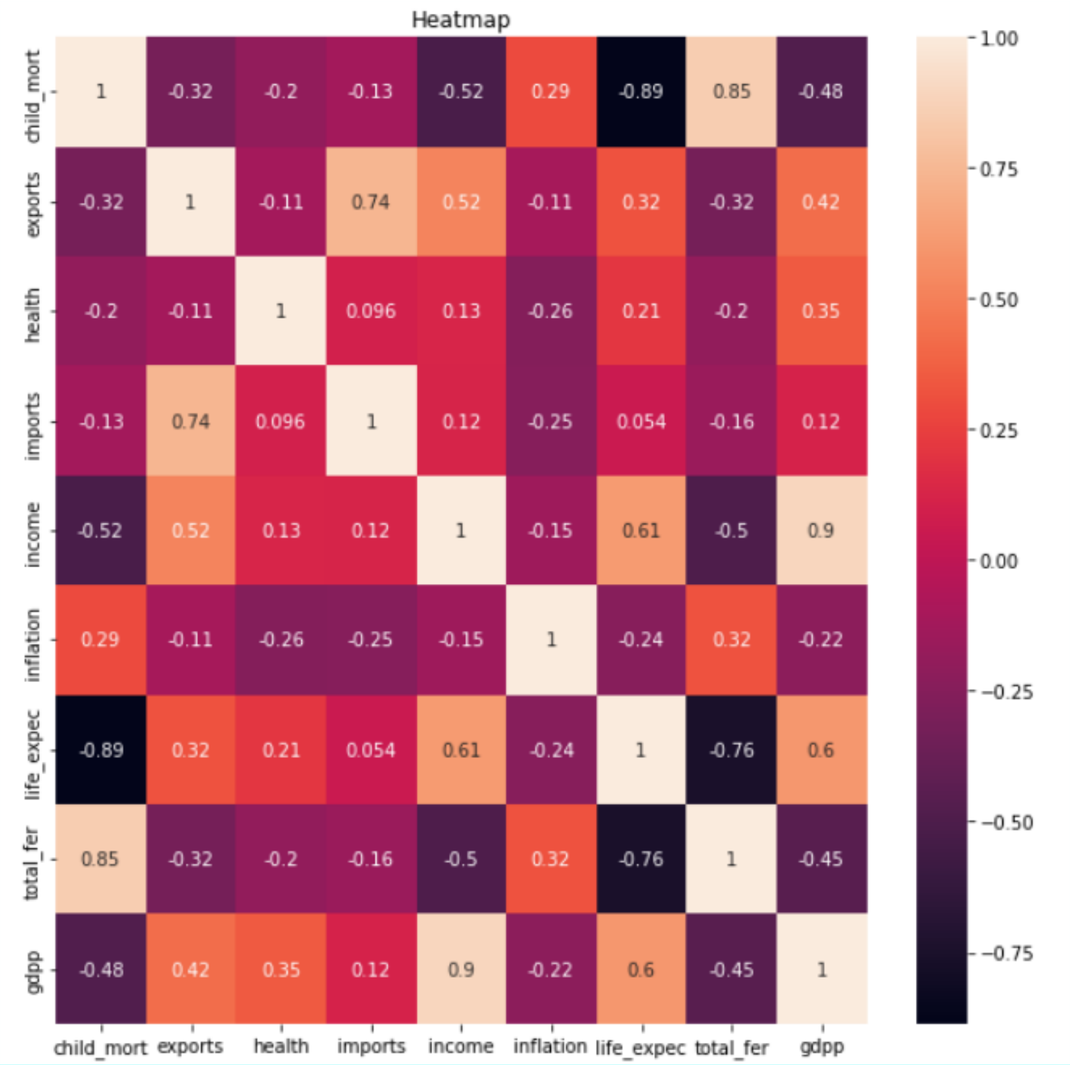
Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	23	35	42	67
Farthest First	3	1	44	119
Hybrid	**3**	**1**	**44**	**128**

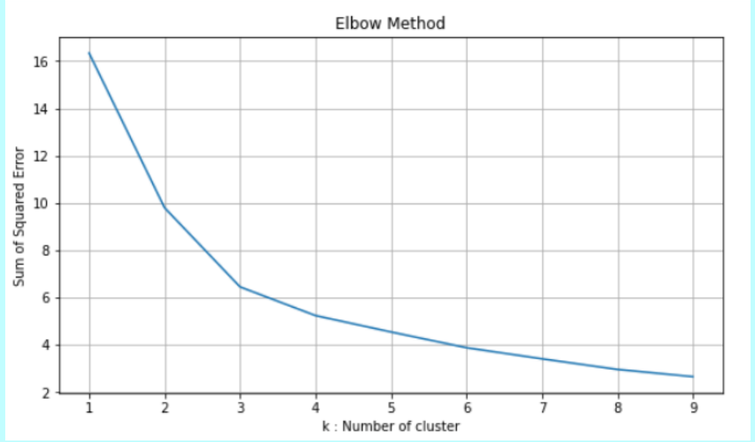
Table 4: Number of Countries for each Cluster in Socio-Economic and Health Datasets

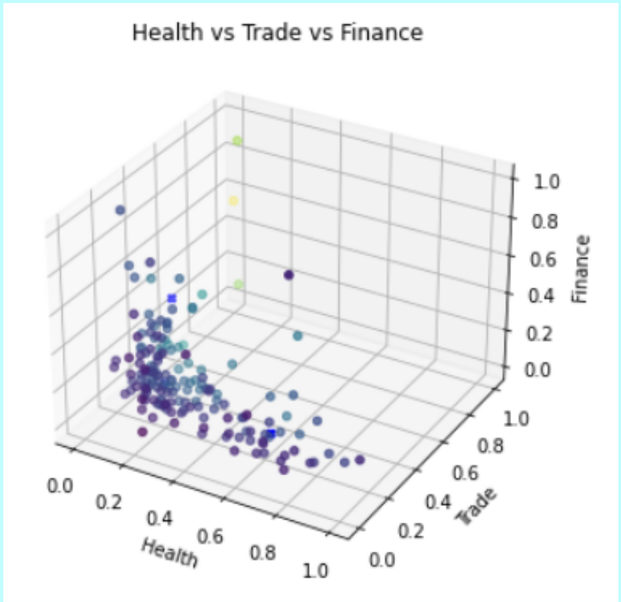
Techniques/clusters	Without risk	Low risk	Medium risk	High risk
k-means++	31	61	2	73
Farthest First	7	92	28	40
Hybrid	**32**	**46**	**27**	**62**

- The paper introduces a novel hybrid clustering technique that combines the K-means algorithm and the farthest-first algorithm to cluster countries based on socio-economic and health datasets obtained from Kaggle.com.
- The study compares the performance of this hybrid approach with that of K-means++ and farthest-first when clustering countries individually.
- The experimental results reveal that the hybrid technique outperforms both K-means++ and farthest-first, significantly enhancing the accuracy of clustering while reducing the associated risks related to socio-economic and health data for each country.

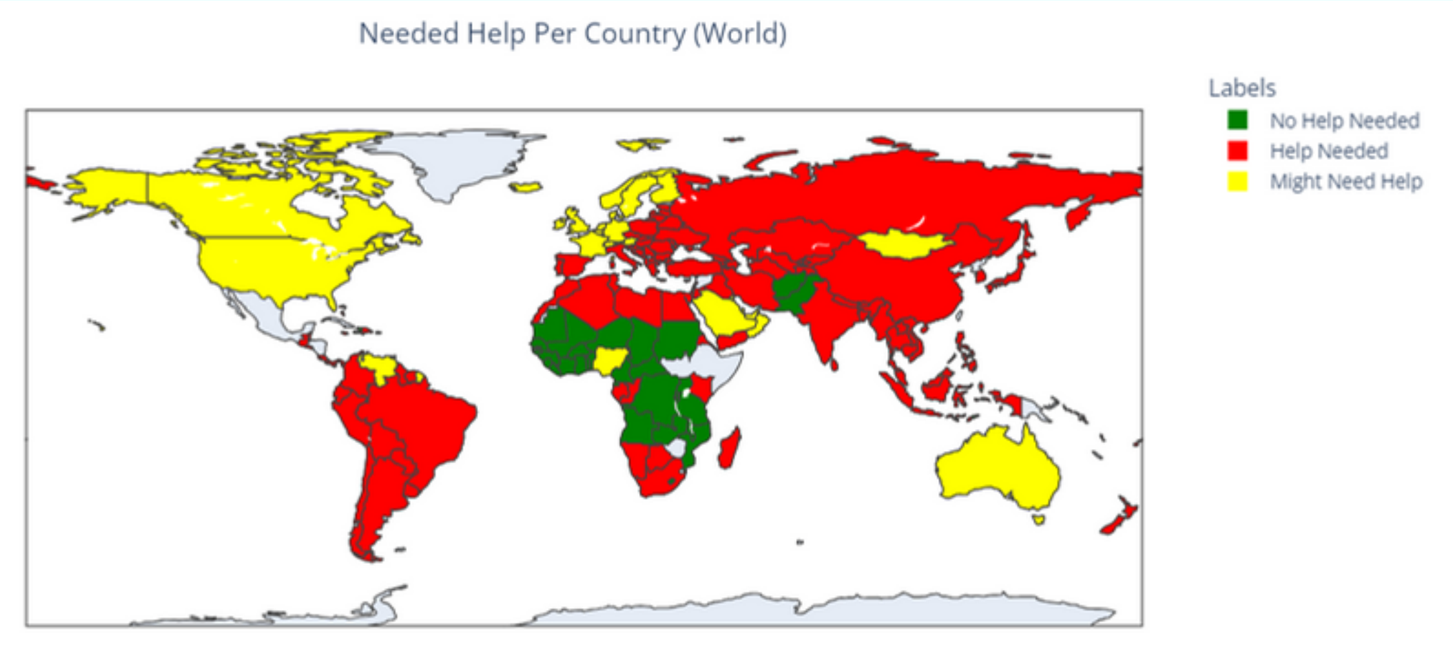
OUR IMPLEMENTATION







RESULTS

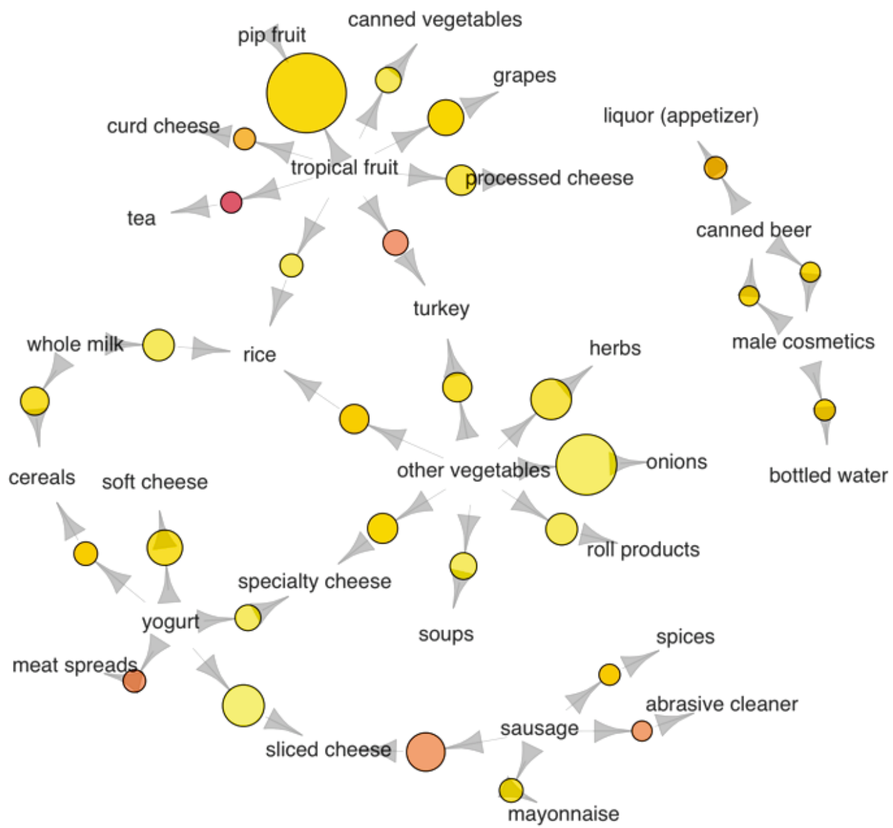



```
class_counts = df1['Class'].value_counts()

class_counts_with_labels = df1['Class'].map({0: 'No Help Needed', 1: 'Help Needed', 2: 'Might Need Help'}).value_counts()

print(class_counts)
print(class_counts_with_labels)
```

```
Help Needed        102
No Help Needed      36
Might Need Help     29
Name: Class, dtype: int64
Series([], Name: Class, dtype: int64)
```



ASSOCIATION APRIORI ALGORITHM

Group Members:
Omkar Satupe (9232)
Ryan Valiaparambil (9237)
Mahek Intwala (9423)

PAPER DETAILS

- Title : MARKET BASKET ANALYSIS FOR A SUPERMARKET
- Research paper Link: Click here
- Dataset: Click here
- Colab File Implementation: Click here

AGENDA

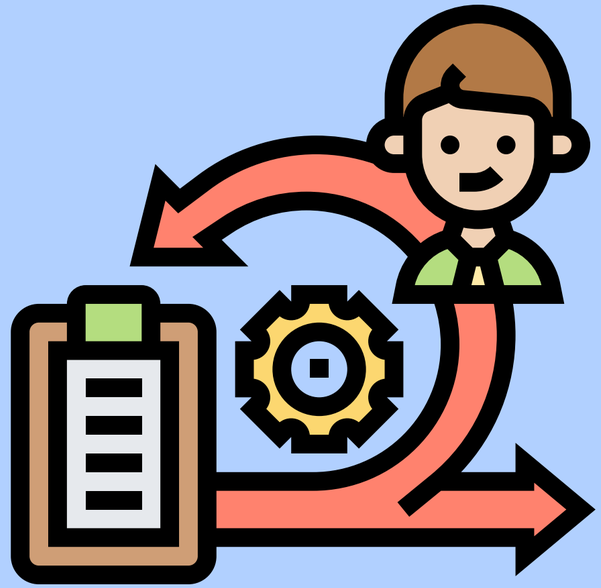
1. Problem Statement
2. Methodology & Algorithms
3. Paper Results and Conclusion
4. Implementation & Results

PROBLEM STATEMENT

- Market Basket Analysis (MBA) is a pivotal data mining technique used across various industries, including retail, to uncover item associations within large datasets.
- This analysis serves to enhance cross-selling strategies, optimize product placement, detect fraudulent activities, and gain insights into consumer purchasing behaviors and preferences.
- The prevalence of MBA in international corporations underscores its significance in contemporary business operations, aligning with the evolving landscape of technology and industry trends. As consumer expectations continue to evolve, it has become imperative for businesses to refine the precision of their operations. In this context, the problem at hand focuses on a neighborhood grocery store, where the primary objective is to scrutinize and compare the runtime performance of two key algorithms, Apriori and FP Growth.
- This study further aims to identify the combinations of products frequently purchased together, facilitating informed decision-making for the store's product stocking and marketing strategies.

METHODOLOGY

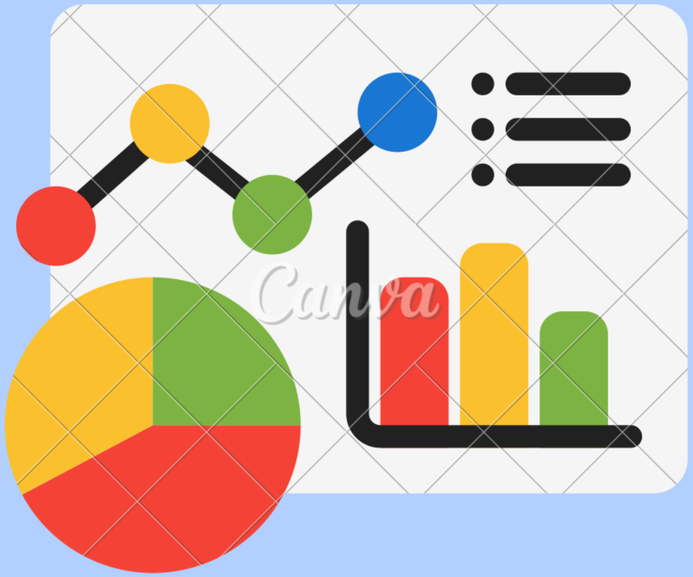
- This research adopts a structured methodology to investigate market basket analysis and association rule mining.
- The study begins with data collection and preprocessing, where a substantial dataset of 7501 transactions is obtained and cleaned for analysis.
- Subsequently, frequent itemsets are identified, and the support and confidence thresholds are set to generate meaningful association rules. Pattern generation and association rule extraction are performed.
- Furthermore, a comprehensive evaluation is conducted to measure the efficiency and performance of association rule mining in market basket analysis. This includes runtime analysis and assessing the ability of the algorithms to discover significant item associations.
- The study concludes with insights into the best-suited methods for practical applications of market basket analysis.



ALGORITHMS USED : Apriori, ECLAT, and FP Growth

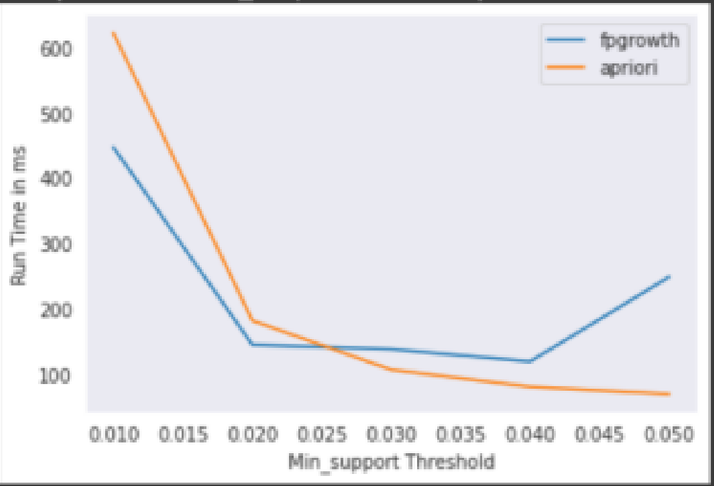
PAPER'S RESULTS

- The study's results provide valuable insights into the shopping habits and associations among products in the dataset. It was revealed that mineral water is a highly popular product, implying a consistent demand that retailers should heed. Using the Apriori algorithm with a lift threshold of 1.3, intriguing connections were discovered.
- It showed that 22% of transactions containing mineral water also included chocolate, indicating a potential product pairing. Moreover, 32% of transactions containing chocolate also contained mineral water. A comparison of lift, leverage, and conviction metrics between {spaghetti and mineral water} and {chocolate and mineral water} revealed that the chances of a transaction involving {spaghetti and mineral water} were higher than {chocolate and mineral water}.
- Additionally, when employing the FP Growth algorithm with a lift threshold above 1.3, it was evident that spaghetti and mineral water are likely to be purchased together. This data underscores the importance of understanding and leveraging such associations for strategic product placement and marketing.
- The comparison of runtimes demonstrated that FP Growth outperforms Apriori in terms of efficiency, being five times faster. Consequently, businesses are encouraged to stock mineral water to meet customer demands, and they can harness the power of FP Growth for streamlined market basket analysis.



```
[ ]  sns.lineplot(x=l,y=f,label="fpgrowth")
     sns.lineplot(x=l,y=t,label="apriori")
     plt.xlabel("Min_support Threshold")
     plt.ylabel("Run Time in ms")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe16d3b5790><mat
```



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(eggs)	(mineral water)	0.179709	0.238368	0.050927	0.283383	1.188845	0.008090	1.062815
1	(mineral water)	(eggs)	0.238368	0.179709	0.050927	0.213647	1.188845	0.008090	1.043158
2	(spaghetti)	(mineral water)	0.174110	0.238368	0.059725	0.343032	1.439085	0.018223	1.159314
3	(mineral water)	(spaghetti)	0.238368	0.174110	0.059725	0.250559	1.439085	0.018223	1.102008
4	(chocolate)	(mineral water)	0.163845	0.238368	0.052660	0.321400	1.348332	0.013604	1.122357
5	(mineral water)	(chocolate)	0.238368	0.163845	0.052660	0.220917	1.348332	0.013604	1.073256

OUR IMPLEMENTATION & RESULTS

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(yogurt)	(other vegetables)	0.328667	0.417693	0.156775	0.477002	1.141991	0.019493	1.113401
1	(whole milk, yogurt)	(other vegetables)	0.328667	0.417693	0.156775	0.477002	1.141991	0.019493	1.113401
2	(yogurt)	(other vegetables, whole milk)	0.328667	0.417693	0.156775	0.477002	1.141991	0.019493	1.113401
3	(other vegetables)	(yogurt)	0.417693	0.328667	0.156775	0.375335	1.141991	0.019493	1.074708
4	(other vegetables, whole milk)	(yogurt)	0.417693	0.328667	0.156775	0.375335	1.141991	0.019493	1.074708
5	(other vegetables)	(whole milk, yogurt)	0.417693	0.328667	0.156775	0.375335	1.141991	0.019493	1.074708
6	(other vegetables)	(rolls/buns)	0.417693	0.389698	0.179171	0.428954	1.100736	0.016397	1.068745
7	(rolls/buns)	(other vegetables)	0.389698	0.417693	0.179171	0.459770	1.100736	0.016397	1.077887
8	(other vegetables, whole milk)	(rolls/buns)	0.417693	0.389698	0.179171	0.428954	1.100736	0.016397	1.068745
9	(rolls/buns, whole milk)	(other vegetables)	0.389698	0.417693	0.179171	0.459770	1.100736	0.016397	1.077887
10	(other vegetables)	(rolls/buns, whole milk)	0.417693	0.389698	0.179171	0.428954	1.100736	0.016397	1.068745

```
frequently_bought_together('salty snack')
```

```
Items frequently bought together with salty snack
array([frozenset({'bottled beer'}), frozenset({'salty snack'}),
       frozenset({'bottled water'}), frozenset({'brown bread'}),
       frozenset({'butter'}), frozenset({'canned beer'})], dtype=object)
```

THANK YOU !