Grocery Purchase Apriori Algorithm Analysis

BY GRISHMA VEMIREDDY

Introduction

Problem statement: Analyze customer shopping purchases to discover which products are frequently purchased together using Apriori algorithm.

The goal is to find association rules that can help optimize product placement, help with marketing and promotions, and cross selling strategies in retail sector.

What is Apriori?

It is an association rule learning algorithm used to mine frequent itemsets and derive association rules. Populare and widely used in market basket analysis. The core idea is to identify frequent itemsets. The principle is that if an itemset is frequent then all its subsets are frequent and vice versa. We can overcome overfitting by using higher minimum support and confidence thresholds.

```
RangeIndex: 999 entries, 0 to 998
Data columns (total 17 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    999 non-null    int64
 1   Apple         999 non-null    bool
 2   Bread         999 non-null    bool
 3   Butter        999 non-null    bool
 4   Cheese        999 non-null    bool
 5   Corn          999 non-null    bool
 6   Dill          999 non-null    bool
 7   Eggs          999 non-null    bool
 8   Ice cream     999 non-null    bool
 9   Kidney Beans  999 non-null    bool
 10  Milk          999 non-null    bool
 11  Nutmeg        999 non-null    bool
 12  Onion         999 non-null    bool
 13  Sugar         999 non-null    bool
 14  Unicorn       999 non-null    bool
 15  Yogurt        999 non-null    bool
 16  chocolate     999 non-null    bool
dtypes: bool(16), int64(1)
```

Dataset overview

Source: Kaggle (Market Basket Analysis Data)

Data description:

999 observation

17 features

No missing values

	Apple	Bread	Butter	Cheese	Corn	Dill	Eggs	
count	999.000000	999.000000	999.000000	999.000000	999.000000	999.000000	999.000000	99
mean	0.383383	0.384384	0.420420	0.404404	0.407407	0.398398	0.384384	
std	0.486454	0.486693	0.493874	0.491022	0.491598	0.489814	0.486693	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
75%	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

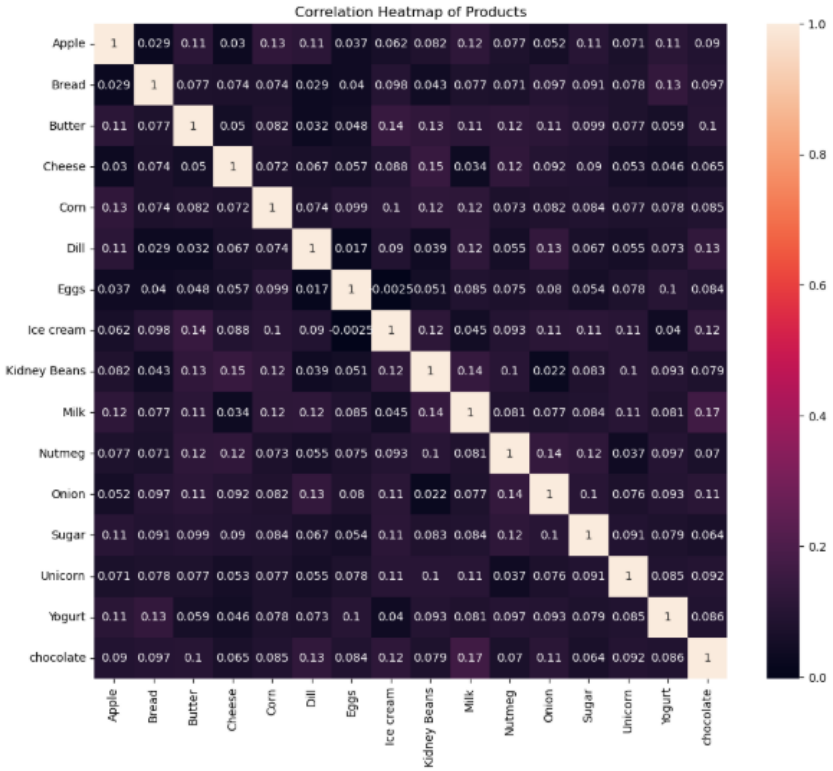
```
# Drop unnecessary index column
data_df = data_df.drop('Unnamed: 0', axis=1)
```

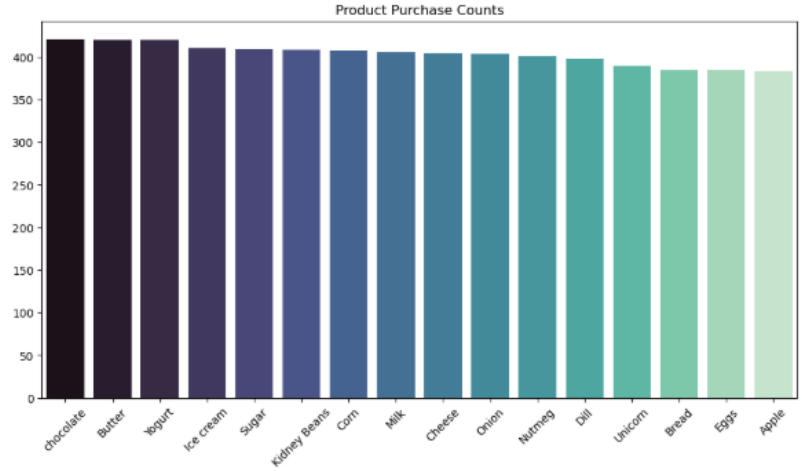
```
#convert bool feature into int
data_df = data_df.astype(int)
```

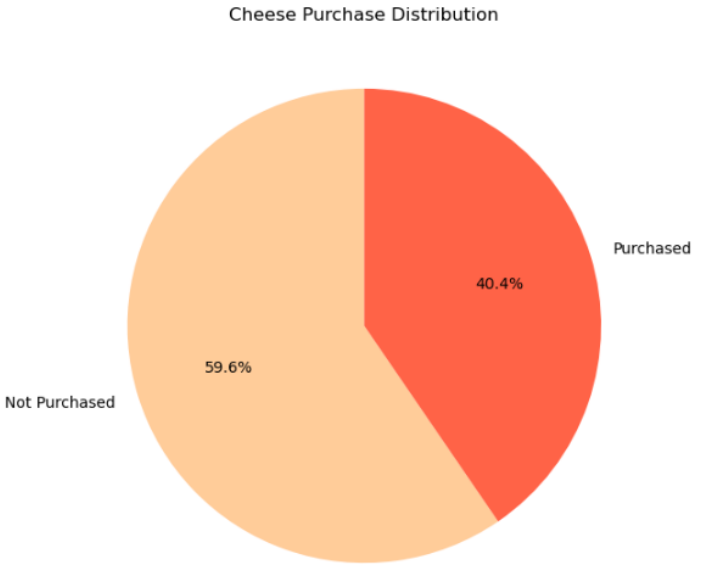
Methodology

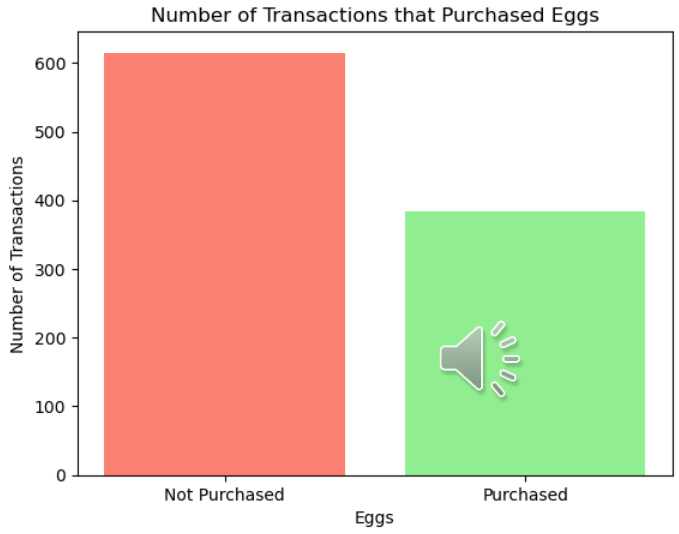
Data exploration and data preprocessing:

Encoded categorical variables and Dropped some irreverent features









```
#Apply apriori
frequent_itemsets = apriori(data_df, min_support=0.05, use_colnames=True)
```

```
#generating association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
```

```
#print association rules
print(frequent_itemsets)
print(rules)
```

```
# filtering rules with high support and high confidence
filtered_rules = rules[(rules['support'] > 0.05) & (rules['confidence'] > 0.
print(filtered_rules)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity
2844	(Dill, Unicorn)	(chocolate)	0.168168	0.421421	0.101101	0.601190	1.426578	1.0
3574	(Apple, Bread, Sugar)	(Yogurt)	0.082082	0.420420	0.050050	0.609756	1.450348	1.0
3657	(Apple, Ice cream, Sugar)	(Butter)	0.091091	0.420420	0.056056	0.615385	1.463736	1.0
3771	(Corn, Dill, Sugar)	(Apple)	0.084084	0.383383	0.052052	0.619048	1.614696	1.0
3796	(Apple, Corn, Eggs)	(Sugar)	0.082082	0.409409	0.051051	0.621951	1.519142	1.0
...	...	...	...	...	...	...	...	...
6639	(Onion, chocolate, Unicorn)	(Dill)	0.099099	0.398398	0.060060	0.606061	1.521243	1.0
6640	(Onion, Dill, Unicorn)	(chocolate)	0.093093	0.421421	0.060060	0.645161	1.530917	1.0
6781	(Ice cream, Milk, Sugar)	(Onion)	0.091091	0.403403	0.055055	0.604396	1.498241	1.0
6864	(Onion, Ice cream, Unicorn)	(chocolate)	0.088088	0.421421	0.053053	0.602273	1.429146	1.0
6890	(Nutmeg, Milk, Unicorn)	(Kidney Beans)	0.081081	0.408408	0.050050	0.617284	1.511438	1.0

Modeling – Apriori

Used Apriori with min support 5%

Association rules mined with lift > 1

Filtered to show some top strong rules

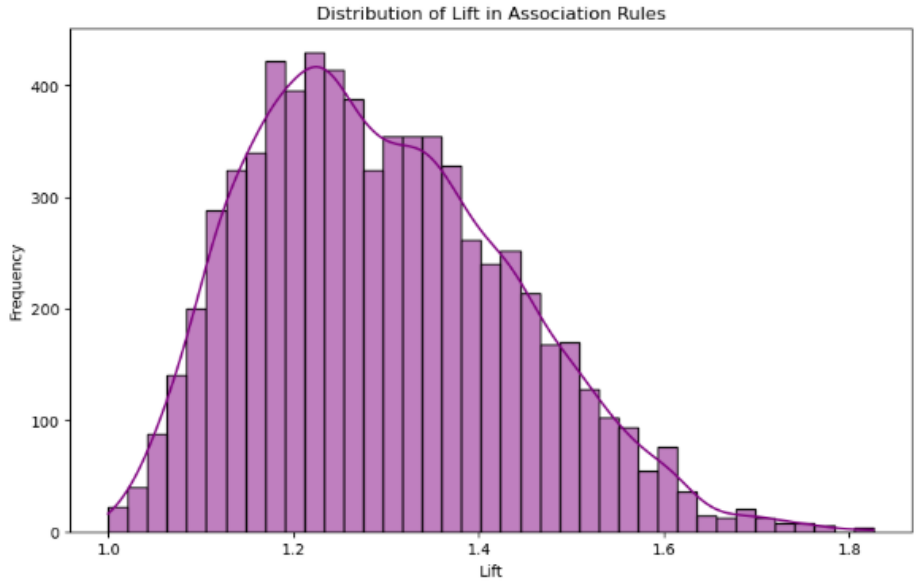
Support : shows how frequently a rule's itemset appears in the dataset

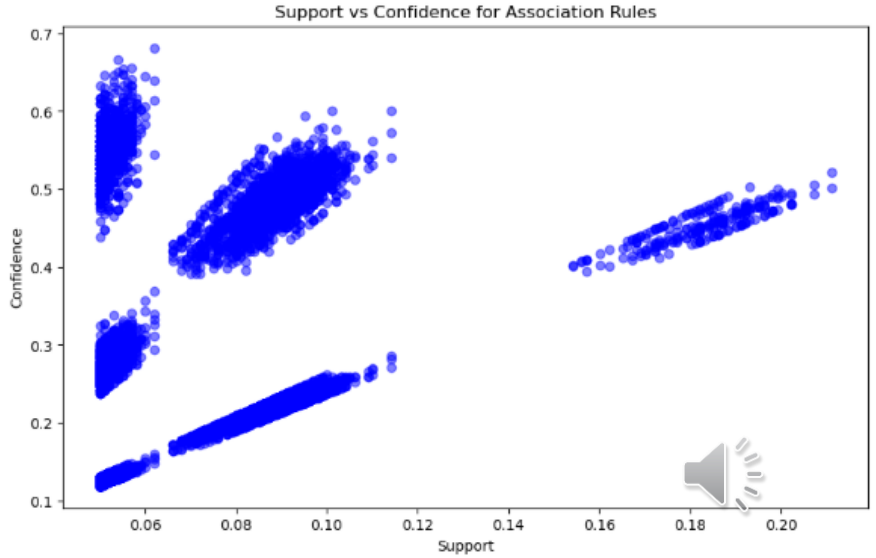
Confidence: indicates the likelihood of the consequent item being purchased when the antecedent item is purchased.

Results

	antecedents	consequents	support	confidence	lift
5637	(Cheese, Dill)	(Onion, Kidney Beans)	0.055055	0.310734	1.826022
5636	(Onion, Kidney Beans)	(Cheese, Dill)	0.055055	0.323529	1.826022
6647	(Dill, Unicorn)	(Onion, chocolate)	0.060060	0.357143	1.820335
6642	(Onion, chocolate)	(Dill, Unicorn)	0.060060	0.306122	1.820335
5662	(chocolate, Cheese)	(Dill, Kidney Beans)	0.057057	0.306452	1.779914
5667	(Dill, Kidney Beans)	(chocolate, Cheese)	0.057057	0.331395	1.779914
3930	(Corn, Sugar)	(Apple, Unicorn)	0.055055	0.294118	1.770021
3927	(Apple, Unicorn)	(Corn, Sugar)	0.055055	0.331325	1.770021
4253	(Bread, Kidney Beans)	(Corn, Milk)	0.057057	0.341317	1.766715
4248	(Corn, Milk)	(Bread, Kidney Beans)	0.057057	0.295337	1.766715

	antecedents	consequents	support	confidence	lift
207	(Milk)	(chocolate)	0.211211	0.520988	1.236263
206	(chocolate)	(Milk)	0.211211	0.501188	1.236263
67	(Butter)	(Ice cream)	0.207207	0.492857	1.200889
66	(Ice cream)	(Butter)	0.207207	0.504878	1.200889
83	(Butter)	(chocolate)	0.202202	0.480952	1.141262
181	(chocolate)	(Ice cream)	0.202202	0.479810	1.169098
180	(Ice cream)	(chocolate)	0.202202	0.492683	1.169098
69	(Kidney Beans)	(Butter)	0.202202	0.495098	1.177626
68	(Butter)	(Kidney Beans)	0.202202	0.480952	1.177626
82	(chocolate)	(Butter)	0.202202	0.479810	1.141262





	antecedents	consequents	support	confiden
637	(Cheese, Dill)	(Onion, Kidney Beans)	0.055055	0.3107
636	(Onion, Kidney Beans)	(Cheese, Dill)	0.055055	0.3235
647	(Dill, Unicorn)	(Onion, chocolate)	0.060060	0.3571
642	(Onion, chocolate)	(Dill, Unicorn)	0.060060	0.3061
662	(chocolate, Cheese)	(Dill, Kidney Beans)	0.057057	0.3064
667	(Dill, Kidney Beans)	(chocolate, Cheese)	0.057057	0.3313
930	(Corn, Sugar)	(Apple, Unicorn)	0.055055	0.2941
927	(Apple, Unicorn)	(Corn, Sugar)	0.055055	0.3313
253	(Bread, Kidney Beans)	(Corn, Milk)	0.057057	0.3413
248	(Corn, Milk)	(Bread, Kidney Beans)	0.057057	0.2953

	lift
637	1.826022
636	1.826022
647	1.820335
642	1.820335
662	1.779914
667	1.779914
930	1.770021
927	1.770021
253	1.766715
248	1.766715

	antecedents	consequents	support	confidence	lift
207	(Milk)	(chocolate)	0.211211	0.520988	1.236263
206	(chocolate)	(Milk)	0.211211	0.501188	1.236263
67	(Butter)	(Ice cream)	0.207207	0.492857	1.200889
66	(Ice cream)	(Butter)	0.207207	0.504878	1.200889
83	(Butter)	(chocolate)	0.202202	0.480952	1.141262
181	(chocolate)	(Ice cream)	0.202202	0.479810	1.169098
180	(Ice cream)	(chocolate)	0.202202	0.492683	1.169098
69	(Kidney Beans)	(Butter)	0.202202	0.495098	1.177626
68	(Butter)	(Kidney Beans)	0.202202	0.480952	1.177626
82	(chocolate)	(Butter)	0.202202	0.479810	1.141262

Conclusion

The bottom picture captures the result of the Apriori algorithm on the dataset showing frequent but less surprising rules that are common co-occurrences but the lift of ~1.14 – 1.2 does not represent novel association.

The less frequent but much more interesting rules are shown by the rules with higher lift

Thank You!