Introduction

Market Basket Analysis (MBA) is a popular data mining technique used to uncover relationships between items in large transaction datasets. This technique is particularly useful in the retail industry, where understanding customer purchase behavior can lead to improved sales strategies, optimized inventory management, and enhanced customer experience.
Market Basket Analysis involves examining the items that customers place in their shopping baskets. By identifying sets of items that frequently occur together in transactions, retailers can gain insights into buying patterns and customer preferences.

Objectives:

1. **Understand Customer Behaviour-** Identify combinations of products that are frequently bought together.
2. **Optimize Product Placement-** Arrange store layouts and online product recommendations to increase sales.
3. **Improve Inventory Management-** Ensure that frequently bought together items are stocked adequately.
4. **Design Effective Marketing Strategies-** Create targeted promotions and cross-selling opportunities based on identified patterns.

For example, Consider a simple example where a store has the following transactions done by multiple customers:

- Transaction 1: {Milk, Bread, Butter}
- Transaction 2: {Bread, Butter}
- Transaction 3: {Milk, Bread}
- Transaction 4: {Milk, Butter}

Market Basket Analysis would help identify frequent itemsets such as:

- {Milk, Bread}
- {Bread, Butter}

These are the items which are frequently bought our might be bought by the customer. This way retail companies would be able to strategize to improve sales accordingly.

Benefits of Market Basket Analysis

1. **Increased Sales**: By placing frequently bought together items near each other, stores can encourage impulse purchases.
2. **Enhanced Customer Experience**: Customers find shopping more convenient when complementary items are easily accessible.
3. **Data-Driven Decisions**: Insights from MBA enable retailers to make informed decisions on product promotions, store layouts, and inventory management.
4. **Competitive Advantage**: Understanding customer behaviour better than competitors can provide a strategic edge.

The Apriori Algorithm

The Apriori algorithm is a foundational method in data mining used for discovering frequent itemsets and deriving association rules. It is widely used in Market Basket Analysis to identify items that frequently occur together in transactions and to establish rules that predict the likelihood of certain items being purchased together.

The Apriori algorithm operates on a database. The Apriori algorithm is a powerful tool for Market Basket Analysis, helping businesses uncover hidden patterns in transaction data. By identifying frequent itemsets and generating association rules, businesses can make data-driven decisions to enhance sales strategies, optimize inventory, and improve customer satisfaction.

Support:

The support of an itemset is the proportion of transactions in which the itemset appears. It measures the popularity of an itemset.

Support(A)=Number of transactions containing A/Total number of transctions

Confidence:

The confidence of a rule A→B is the proportion of transactions containing A also contain B. It measures the likelihood of purchasing item B given that item A is purchased.

$$\text{Confidence}(A \rightarrow B)=\text{Support}(A \cup B)/\text{Support}(A)$$

Lift:

Strength of the association rule. If lift>1, strong association.

$$\text{Lift}(A \rightarrow B)=\text{Support}(A \cup B)/\text{Support}(A)\times\text{Support}(B)$$

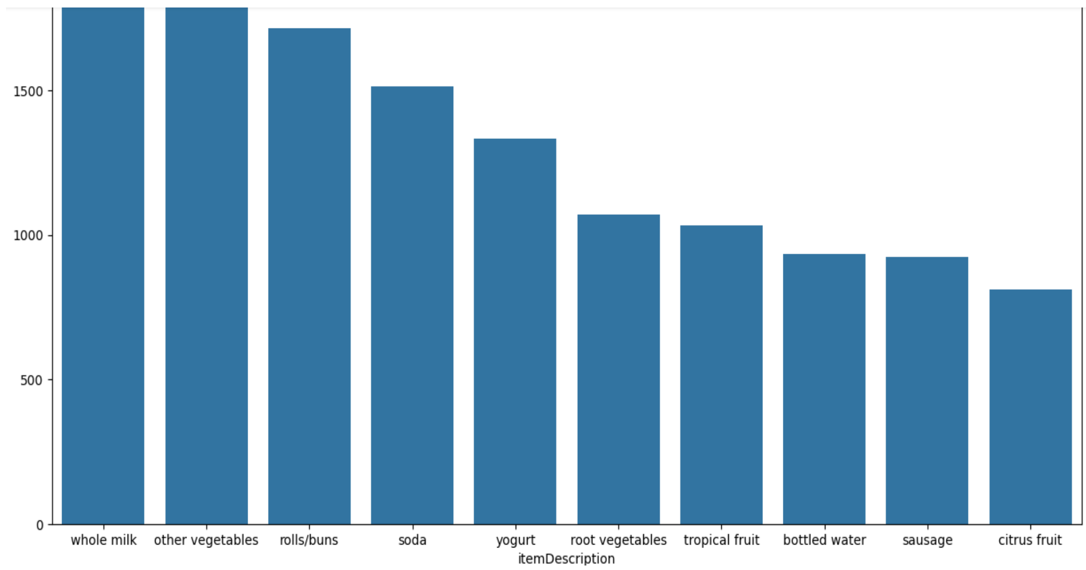
Methodology

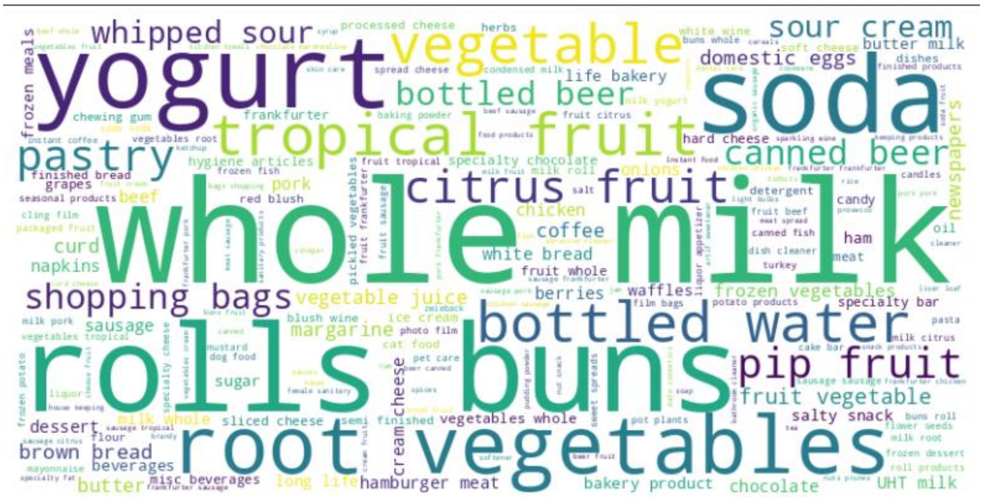
Data Preparation

1. **Dataset**: A transactional dataset where each row represents a transaction, and each column represents an item of a grocery store with 365 rows and 3 colums is taken from kaggle.

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk
...	...	...	...
38760	4471	08-10-2014	sliced cheese
38761	2022	23-02-2014	candy
38762	1097	16-04-2014	cake bar

2. **Preprocessing**: Convert the dataset into a format suitable for the Apriori algorithm, typically a binary matrix where rows are transactions and columns are items. Remove null values, filter the dataset. We sort the data according to most bought items and plot a graph and word cloud for the same.





In the graph and word cloud it was observed that whole milk, roll buns, yogurt, root vegetable were most frequently bought items.

3. We create a pivot table to see what all items are bought by a particular member and if any item is bought we assign it 1 else 0.

itemDescription	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	bags	baking powder	bathroom cleaner	beef	berries	...	turkey	vinegar	waffles	whipped/sour cream	whisky	white bread
Member_number																	
1000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
1001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	1.0
1002	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
1003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
1004	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4996	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
4997	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
4998	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0

4. **Apriori Algorithm**

The Apriori algorithm works in two main steps:

Frequent Itemset Generation: Identify all itemsets that meet a minimum support threshold. Support is the proportion of transactions in which an itemset appears.

Rule Generation: Generate association rules from the frequent itemsets. Rules are in the form of "If A then B," with confidence and lift metrics to evaluate their strength.

5. **Metrics**

Support: The frequency of occurrence of an itemset in the dataset.
Confidence: The likelihood that item B is purchased when item A is purchased.
Lift: The ratio of the observed support to that expected if A and B were

independent. A lift greater than 1 indicates a positive association between A and B.

We use **“ mlxtend”** library to form association rule and find support , confidence, and lift against each antecedent and consequent which tells us what combination of items the customer prefers to buy.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(beef)	(whole milk)	0.119548	0.458184	0.064135	0.536481	1.170886	0.009360	1.168919	0.165762
1	(whole milk)	(beef)	0.458184	0.119548	0.064135	0.139978	1.170886	0.009360	1.023754	0.269364
2	(bottled beer)	(other vegetables)	0.158799	0.376603	0.068497	0.431341	1.145345	0.008692	1.096257	0.150857
3	(other vegetables)	(bottled beer)	0.376603	0.158799	0.068497	0.181880	1.145345	0.008692	1.028212	0.203563
4	(bottled beer)	(rolls/buns)	0.158799	0.349666	0.063109	0.397415	1.136555	0.007582	1.079240	0.142829

6. Filter out values according to lift which tells how strong the association is. Confidence>0.5 and lift>1.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(beef)	(whole milk)	0.119548	0.458184	0.064135	0.536481	1.170886	0.009360	1.168919	0.165762
6	(bottled beer)	(whole milk)	0.158799	0.458184	0.085428	0.537964	1.174124	0.012669	1.172672	0.176297
14	(bottled water)	(whole milk)	0.213699	0.458184	0.112365	0.525810	1.147597	0.014452	1.142615	0.163569
18	(brown bread)	(whole milk)	0.135967	0.458184	0.069779	0.513208	1.120091	0.007481	1.113034	0.124087
20	(butter)	(whole milk)	0.126475	0.458184	0.066188	0.523327	1.142176	0.008239	1.136661	0.142501
26	(canned beer)	(whole milk)	0.165213	0.458184	0.087224	0.527950	1.152268	0.011526	1.147795	0.158299
37	(curd)	(whole milk)	0.120831	0.458184	0.063622	0.526539	1.149188	0.008259	1.144374	0.147663
39	(domestic eggs)	(whole milk)	0.133145	0.458184	0.070292	0.527938	1.152242	0.009287	1.147766	0.152421
47	(newspapers)	(whole milk)	0.139815	0.458184	0.072345	0.517431	1.129310	0.008284	1.122775	0.133115
66	(other vegetables)	(whole milk)	0.376603	0.458184	0.191380	0.508174	1.109106	0.018827	1.101643	0.157802
74	(pastry)	(whole milk)	0.177527	0.458184	0.091072	0.513006	1.119651	0.009732	1.112572	0.129931
80	(pip fruit)	(whole milk)	0.170600	0.458184	0.086968	0.509774	1.112598	0.008801	1.105239	0.122020
82	(pork)	(whole milk)	0.132376	0.458184	0.066957	0.505814	1.103955	0.006305	1.096381	0.108533
95	(rolls/buns)	(whole milk)	0.349666	0.458184	0.178553	0.510638	1.114484	0.018342	1.107190	0.157955
107	(sausage)	(whole milk)	0.206003	0.458184	0.106978	0.519303	1.133394	0.012591	1.127146	0.148230

These are the most frequent bought itemsets that the customer buys from a grocery store. If a customer buys beef he prefers to buy whole milk as well. Brown bread followed by whole milk.

Results

The output consists of frequent itemsets and association rules. For instance:

1. **Frequent Itemsets**:

- {rolls/buns, yogurt, whole milk} Support = 0.07
- {whole milk, yogurt, buns/rolls }: Support = 0.06

2. **Association Rules**:

- If {rolls/buns, yogurt} then {whole milk}: Confidence = 0.59, Lift = 1.3
- If {whole milk, yogurt} then {buns/rolls}: Confidence = 0.5, Lift = 1.29

Analysis

1. **Frequent Itemsets**: Identify combinations of items that frequently appear together in transactions. The results came out to be itmesets like roll/buns, yogurt followed by whole milk or whole milk, yogurt followed by buns/rolls.
2. **Association Rules**: Determine rules that have high confidence and lift, indicating a strong relationship between items.
3. **Business Implications**:
    - **Product Placement**: Place frequently bought together items near each other to increase convenience for customers.
    - **Cross-Selling**: Create promotions that bundle items that are often bought together.
    - **Inventory Management**: Ensure that frequently associated items are always in stock together to avoid losing sales opportunities.

Conclusion & Future Scope

Conclusion:

Market Basket Analysis using the Apriori algorithm provides valuable insights into customer purchasing behavior. By understanding which items are frequently bought together, retailers can make data-driven decisions to enhance the shopping experience, optimize store layout, and increase sales through targeted marketing strategies.

Future Work:

1. **Dynamic Analysis**: Regularly update the analysis to capture seasonal trends and changes in purchasing behavior.
2. **Advanced Algorithms**: Explore other algorithms like FP-Growth for larger datasets.
3. **Integration with Other Data**: Combine transaction data with demographic or psychographic data for deeper insights.