Big Data Analytics

10/03/2024

**Association Rule Mining on Groceries Dataset**

# **Abstract:**

This research paper delves into the application of association rule mining in the context of a grocery shopping dataset. We aim to identify frequent item co-occurrences to understand customer buying patterns and inform strategic decisions within the retail industry. The paper discusses the methodology employed, including data preparation, rule generation using popular algorithms like apriori algorithm, and evaluation metrics. Finally, we present insightful association rules discovered from the analysis and discuss their potential implications for retailers

# **Introduction:**

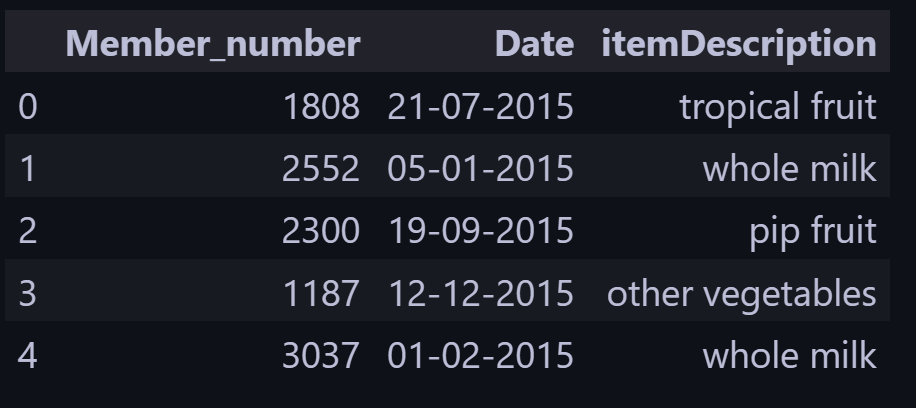
The retail industry, especially in the grocery sector, is witnessing a paradigm shift with the advent of big data analytics. The massive volumes of data generated by consumer transactions present an opportunity to extract valuable patterns, associations, and trends that can significantly impact business decisions. Association rule mining, a technique widely used in data mining, provides a means to discover interesting relationships between variables in large datasets.

For this assignment we have used the groceries dataset. The dataset has **38765** rows of the purchase orders of people from the grocery stores This dataset has only **three attributes** as follows

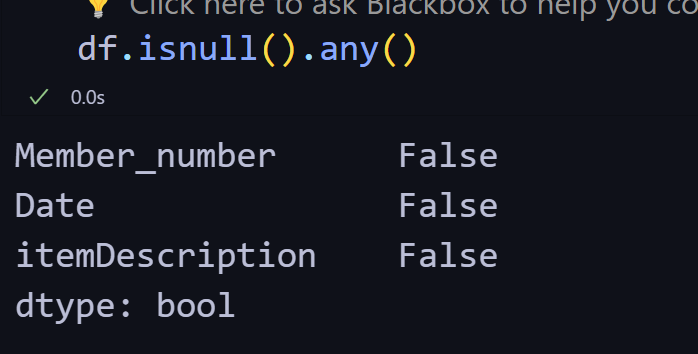
* Member\_number (id of member)
* Date
* Item\_description

# **Exploratory Data Analysis:**

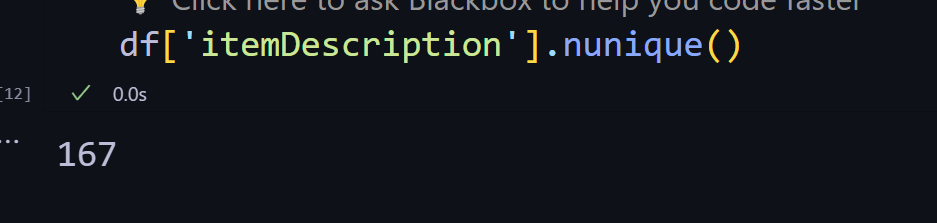
As already mentioned the dataset contains three attributes as given below



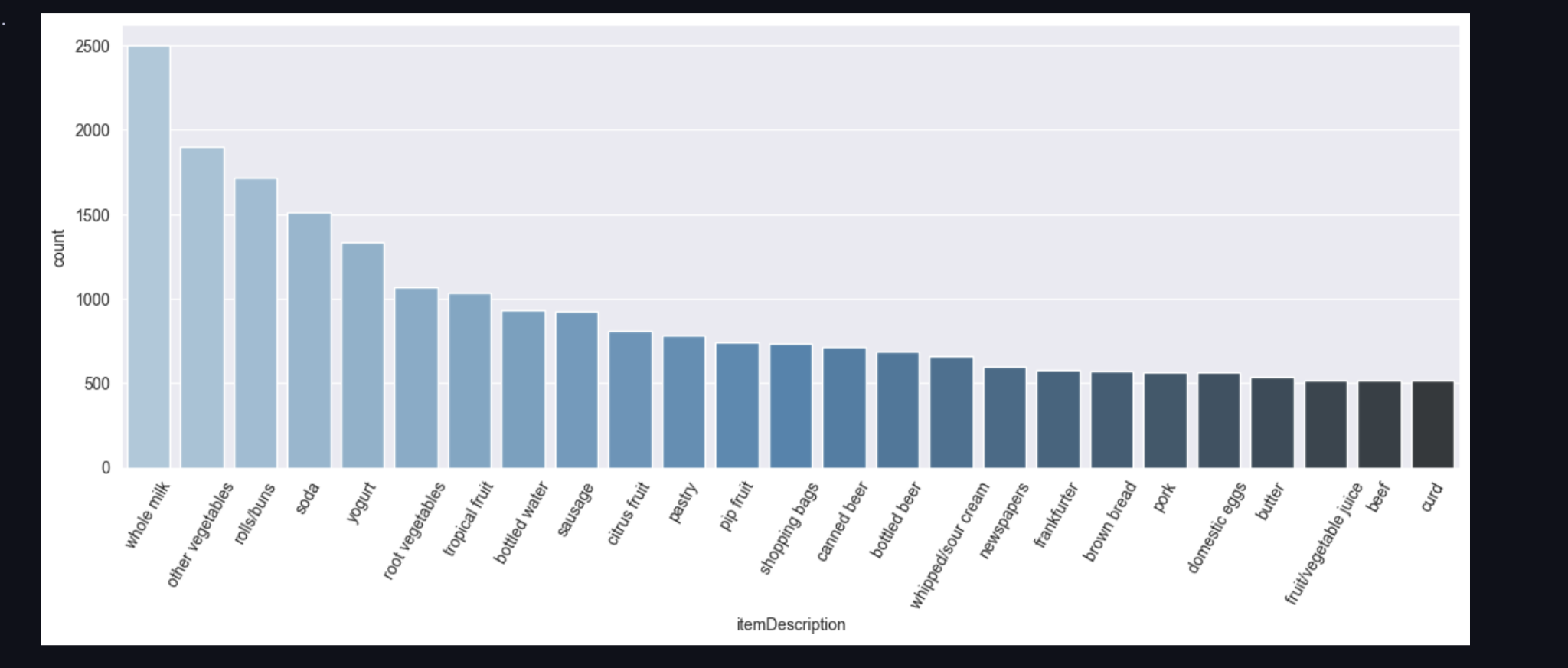
The dataset contains no null entries



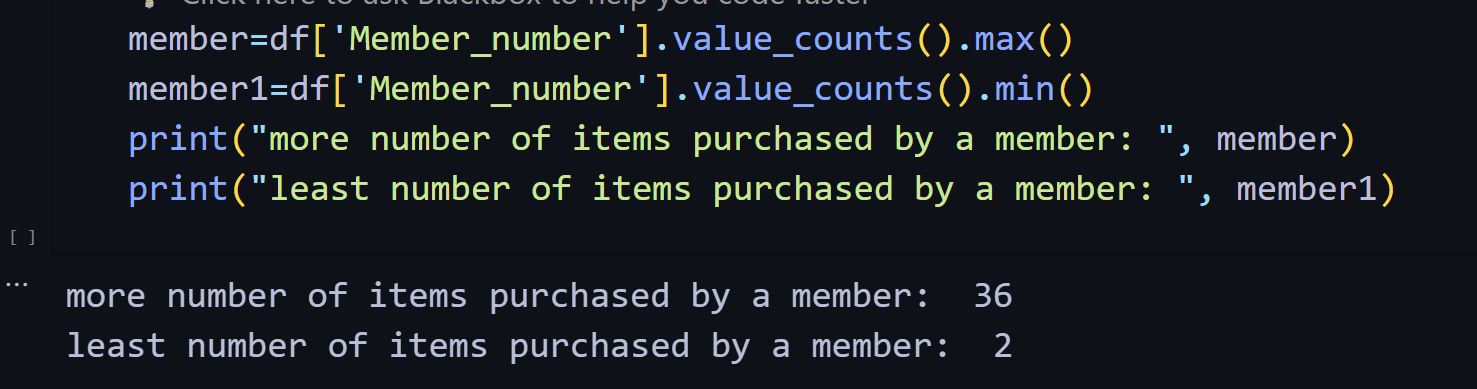
There are total 167 unique grocery items



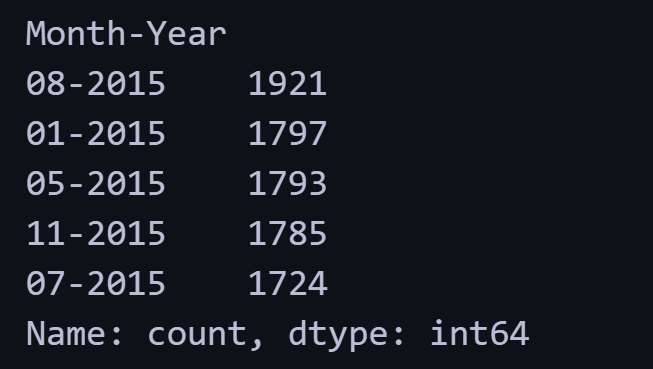
This bar chart provides a comprehensive view of the groceries dataset. The length of each bar reflects the frequency of an item being purchased, revealing the most popular items and the overall variety of products available in the store. We can see that **Whole Milk** is the most bought product that is bought **2502** times



Also the most numer of items purchased by a single member is 36 and the least number of items purchased is 2

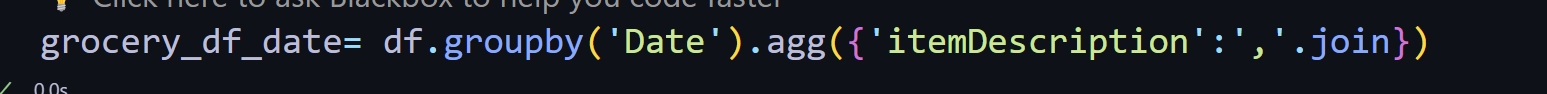


Exploring monthly sales, we find out that the most number of sales are made in **08-2015**

****

# **Methodology and Results:**

Our dataset contains only one item per row in the itemDescription column. So we need to convert the data into a list of transactions. We used the date column to combine all the grocery items bought on that date. We could join the items with Member\_number but it wouldn’t give wholistic information about the complete items that are bought together frequently





In this way we reduce our dataset to **728** transactions

Now we split our itemDescription column to a list of list of transactions and feed it to apriori algorithm imported from apyori library

We set the support to 5, confidence to 70% , and min\_lift to 3, and get the total 379 association rules. When we sort the association rules by lift we get the following top 10 association rules



# **Conclusion and Findings:**

Analyzing the association rules reveals some interesting customer buying habits within the grocery dataset. The frequent co-occurrences of **baking ingredients** like "curd," "baking powder," "chocolate," and "yeast" suggest that customers purchasing some of these staples are likely to buy others as well.

There's also an unexpected association between "pickled vegetables" and "spread cheese." While the reason behind this co-occurrence isn't immediately clear from the rules, it might be worth investigating further to uncover a specific recipe or regional preference driving this pairing. Additionally, several rules link condiments, such as "mustard" with "salt" and "light bulbs" with "cling film/bags." These co-occurrences potentially indicate planned meals or shopping habits where these items are frequently bought together

It's important to note that all the rules have a confidence level of 1.00, indicating a strong association between the items. However, this high confidence might be due to statistical quirks in smaller datasets. The more concerning aspect is the low support values (ranging from 0.005495 to 0.006868), which signify that these item co-occurrences happen in a small fraction of transactions. Depending on the overall dataset size, these low support values might limit the generalizability of the patterns to the broader customer base.

Finally, exploring co-occurrences between broader product categories (e.g., bakery items, condiments) might identify wider customer buying trends that can be leveraged for strategic decision-making within the retail space