A Market Basket Analysis on the Sales Invoice of a General Merchandise

Arabella Mae M. Aduviso
BSCS-NS-3A
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila, Philippines
arabellamae.aduviso@tup.edu.ph

Luis Pocholo Caducio
BSCS-NS-3A
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila, Philippines
luispocholo.caducio@tup.edu.ph

Juliana Anne M. Capoquian
BSCS-NS-3A
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila, Philippines
julianaanne.capoquian@tup.edu.ph

Simone Arabella B. Caturla
BSCS-NS-3A
Technological University of the Philippines
Ayala Boulevard, Ermita, Manila, Philippines
simone.caturla@tup.edu.ph

It is an important objective of Market Basket Analysis to predict the probability of items being bought together by customers [1]. This project proposes a market basket analysis in sales data taken from the sales invoices issued by the case study. A sales invoice may be requested by a customer to be issued. CMC Rice Center and General Merchandise have been selected as the case study. The market basket analysis looks for items frequently found together in a market basket. The result obtained from it can help cross-sell products, and create product bundles for CMC Rice Center [2, 3].

Association Rule Mining, Apriori Algorithm, Market Basket Analysis, Sales Data, Sales Invoice.

I. INTRODUCTION

Market Basket Analysis, a cornerstone of data science in marketing, empowers businesses to decode consumer preferences and enhance sales strategies [4]. This research delves into the realm of market basket analysis, focusing on sales invoice data from a prominent general merchandise retailer, CMC Rice Center and General Merchandise.

Traditionally, marketers relied on intuition to create product combinations and marketing strategies. With the rise of data science, organizations now leverage Market Basket Analysis to identify items frequently purchased together, revealing correlations that escape human observation [5]. In the realm of general merchandise, where product diversity is vast, understanding these patterns becomes crucial for optimizing product placement, designing targeted promotions, and elevating the overall shopping experience.

At the core of Market Basket Analysis lies the Apriori algorithm, a key player in association rule mining [6]. By uncovering correlations between items purchased by users, businesses can tailor strategies to align with customer preferences. In the ensuing exploration, we delve into the nuances of Market Basket Analysis, its applications in general merchandise, and how businesses can leverage these insights for sustained growth.

II. BACKGROUND

A. Association Rule Mining: Apriori Algorithm

Behera, Fartale, Bhagat, and Sharma defined association rule mining as generally used to extract the interesting correlation, frequent pattern, association among sets of items in database. Also called market basket analysis when used in the retail industry, association rule mining is the main data mining technique. This technique looks for associations between data values. The data values are items bought by the customers [2]. This project looks for items frequently bought together in a market basket, hence its name. Association rule analysis generates many potential rules, and it is important to evaluate and select the most relevant rules [7].

The following measures are the three main components of the Apriori algorithm [7, 8]:

- Support:
    - Rules with high support are more significant as they occur more frequently in the dataset.
- Confidence:
    - Rules with high confidence are more reliable, as they have a higher probability of being true.
- Lift:
    - Rules with high lift indicate a strong association between the antecedent and consequent, as they occur together more frequently than expected by chance.

Apriori is one of the famous, most important, and scalable algorithms for mining frequent itemsets and association rule mining. Apriori was introduced by Agrawal and Srikant in 1993 [9, 10]. Research by Kaur and Kang [2016], knowledge has been provided about Apriori series approaches: AIS algorithm, Apriori algorithm, FP-Tree algorithm (Frequent Pattern-Tree algorithm), and RARM (Rapid Association Rule Mining) algorithm; but from these algorithms, Apriori is the biggest improvement from previous algorithms and easy to implement. A minimum threshold is set on the expert advice or user understanding [10].

1. Join Step: This step generates (K+1) itemset from K-item sets by joining each item with itself.
2. Prune Step: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent, and thus it is removed. This step is performed to reduce the size of the candidate item sets.

B. *Technology Used*

The electronic method of data processing, i.e. collection and translation of a data set into a more readable format, is done through a set of programs or software running on computers [11, 12].

Jupyter Notebook is a notebook authoring application which combines two components [13]:

- A web application:
    - A browser-based editing program for interactive authoring of computational notebooks which provides a fast interactive environment for prototyping and explaining code, exploring and visualizing data, and sharing ideas with others.
- Computational Notebook documents:
    - A shareable document that combines computer code, plain language descriptions, data, rich visualizations like 3D models, charts, mathematics, graphs and figures, and interactive controls.

Notebooks extend the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results [13].

III. REVIEW OF RELATED LITERATURE

A. *Market Basket Analysis*

Market basket analysis is also known as association rule mining, which is a type of data mining procedure. It focuses on establishing the relationship between items in the market, which helps market analysts determine which items are frequently bought together [14]. Existing works in Market Basket Analysis were typically done to gain profit for the seller and improve the strategy in placement of products on the shelf [15].

B. *Apriori Algorithm in Data Mining*

Association rule typically uses the Apriori Algorithm, which is the best-known algorithm for mining [16]. However, it can’t identify changes in data because it can only analyze static data [14]. The theory behind the Apriori algorithm is that if a customer buys one thing, they will predictably buy other items too. By determining which items are frequently bought together, customers can reduce time in deciding which items to buy if they are already bundled together [9].

C. *Product Bundling*

Product pricing is different for separate sales compared to bundling. Which is why bundling products with a strong relationship with each other is a way to strategically promote products and improve sales [17]. This is proven by Giri et al. [18] in his study where it was shown that the profit of pure bundling sales in higher than separate sales because bundles can reduce customers’ cost.

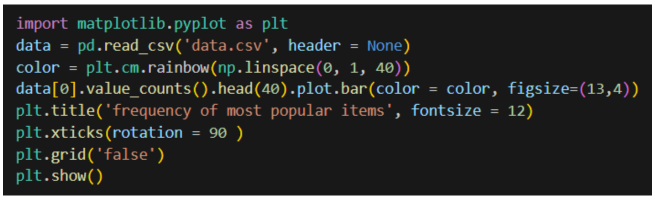
IV. EXPERIMENTS

A. *Data Sets*

The effectiveness of the Market Basket Analysis project for CMC Rice Center and General Merchandise hinges on acquiring and processing transactional data from sales invoices. The data, sourced directly from the store's paper-based system, requires meticulous manual transcription into a digital format, presenting a notable challenge. The project spans sales data from 2020 to 2022, offering a substantial historical dataset for analysis. This timeframe enables the identification of changing item associations, exploration of evolving trends, and assessment of factors such as shifting customer preferences. Ultimately, the project aims to leverage market basket analysis insights to adapt marketing and sales strategies in response to dynamic customer behaviors.

B. *Data Pre-processing*

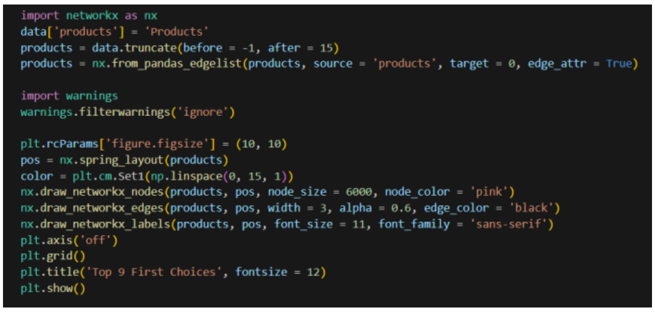
Succeeding to the digitalization of the sales invoice data of CMC Rice Center and General Merchandise, the data went through a rigorous data cleaning and pre-processing phase, including handling missing values and duplicates. Data pre-processing is a technique that ensures that the data is in a format suitable for analysis. The final dataset was then applied to the system for data mining.



```
import matplotlib.pyplot as plt
data = pd.read_csv('data.csv', header = None)
color = plt.cm.rainbow(np.linspace(0, 1, 40))
data[0].value_counts().head(40).plot.bar(color = color, figsize=(13,4))
plt.title('frequency of most popular items', fontsize = 12)
plt.xticks(rotation = 90 )
plt.grid('false')
plt.show()
```

Fig. 1. Frequency of most popular items bar graph Jupyter code

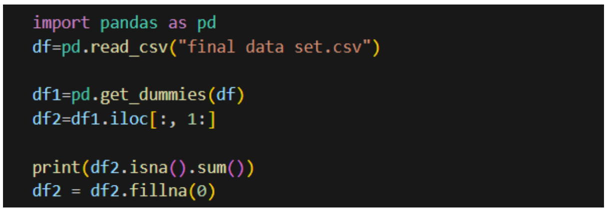
To better describe the data, a Python library *matplotlib* is used to show a bar graph of the frequency of the most popular items found in the data. This is an Exploratory Data Analysis (EDA) technique used to summarize and analyze datasets.



```
import networkx as nx
data['products'] = 'Products'
products = data.truncate(before = -1, after = 15)
products = nx.from_pandas_edgelist(products, source = 'products', target = 0, edge_attr = True)

import warnings
warnings.filterwarnings('ignore')

plt.rcParams['figure.figsize'] = (10, 10)
pos = nx.spring_layout(products)
color = plt.cm.Set1(np.linspace(0, 15, 1))
nx.draw_networkx_nodes(products, pos, node_size = 6000, node_color = 'pink')
nx.draw_networkx_edges(products, pos, width = 3, alpha = 0.6, edge_color = 'black')
nx.draw_networkx_labels(products, pos, font_size = 11, font_family = 'sans-serif')
plt.axis('off')
plt.grid()
plt.title('Top 9 First Choices', fontsize = 12)
plt.show()
```

Fig. 2. Top 9 first choice items diagram Jupyter code

Similarly, applying *networkx* library prompt to show the summarized top first choice items of the customers. In essence, this can help identify patterns in customers' purchasing behaviors. However, there is no guarantee that the market basket analysis will yield the first choice products, the graph is only shown to describe the trends in the purchase habits of customers.



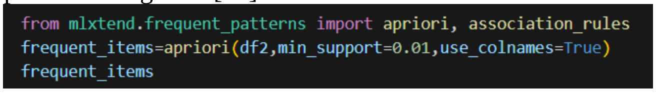
```
import pandas as pd
df=pd.read_csv("final data set.csv")

df1=pd.get_dummies(df)
df2=df1.iloc[:, 1:]

print(df2.isna().sum())
df2 = df2.fillna(0)
```

Fig. 3. Data pre-processing in Jupyter

Prior to data mining process, Python library *pandas* is used to read the dataset that is in CSV file format. Then, the system executes one-hot encoding on categorical variables where it converts these variables into binary columns for each category. The following line of codes selects relevant columns in the dataset, checks for missing values, and fills any missing values with 0. The dataset gone through this process to be executed for the data mining process flawlessly.

C. Data Mining

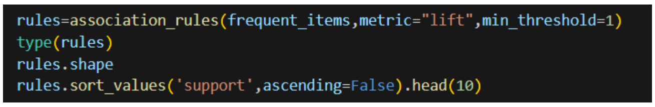
The aim of data mining is to identify patterns and relationship within the data. Market basket analysis is a data mining technique that analyzes patterns of co-occurrence and determines the strength of the link between products purchased together [19].



```
from mlxtend.frequent_patterns import apriori, association_rules
frequent_items=apriori(df2,min_support=0.01,use_colnames=True)
frequent_items
```

Fig. 4. Getting the frequent item sets using Apriori algorithm

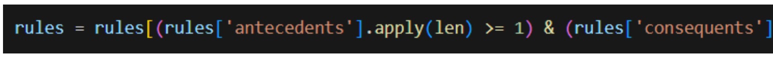
Association rules mining algorithm is used for MBA, whereas Apriori algorithm is utilized in the system. The algorithm identifies the sets of items that often appears together in a transaction—frequent item sets.



```
rules=association_rules(frequent_items,metric="lift",min_threshold=1)
type(rules)
rules.shape
rules.sort_values('support',ascending=False).head(10)
```

Fig. 5. Generating association rules

Association rules express the likelihood of one item being purchased given the purchase of another item [1]. The *association_rules* function is used to generate association rules from a set of frequent itemset. The parameters of the function include the metric used for rule evaluation, *lift*, and a minimum threshold for the specified metric, in this case *1*. Subsequently, the system examines the type and shape of the resulting rules DataFrame and displays the top 10 rules based on support in descending order.

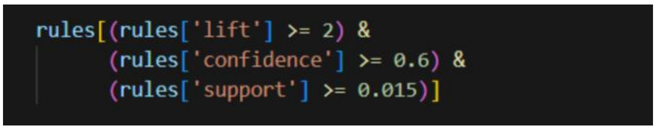


```
rules = rules[(rules['antecedents'].apply(len) >= 1) & (rules['consequents']
```

Fig. 6. Filtering association rules

An association rule consists of antecedent and consequent. The antecedent is the item or set of items that appear on the left-hand side of the rule, while the consequent is the item or set of items that appear on the right-hand side of the rule [19]. The generated association rules are filtered to include only those rules in which both the antecedent and consequent have a minimum length of 1.

Additionally, association rules look for certain patterns that associate two or more elements to a predetermined degree of confidence and support [20]. The association rules are then filtered based on the minimum values for lift, confidence, and support, providing a subset of rules that meet these criteria.



```
rules[(rules['lift'] >= 2) &
      (rules['confidence'] >= 0.6) &
      (rules['support'] >= 0.015)]
```

Fig. 7. Filtering association rules based on the defined minimum values of lift, confidence, and support

The defined minimum values of lift, confidence, and support are 2, 0.6, and 0.015 consecutively.

A value of 2 or higher for lift is often used to identify strong and meaningful associations. A lift of 2 indicates that the occurrence of the antecedent is twice as likely to result in the occurrence of the consequent compared to random chance.

A value of 0.5 or higher indicates a relatively reliable rule. A confidence of 0.6 means that the rule predicts the occurrence of the consequent with a reliability of 60%.

Support is the proportion of transactions that contain the items on both sides of the rule. A value of 0.015 or higher might be suitable, depending on the size of the dataset and the prevalence of the items.

The subset of rules provided after the filtering process, are the rules or item sets suitable for the Market Basket.

V. RESULTS AND DISCUSSION

A. *Data Visualization of the Dataset*

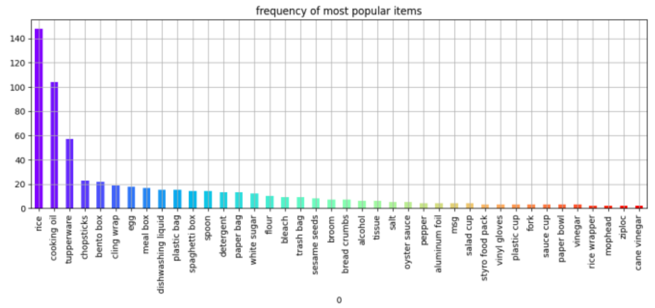


Fig. 8. Frequency of most popular items bar graph

The figure shows the most bought items in CMC Rice Center and General Merchandise and their corresponding frequency. It reveals that rice is the most popular item acquiring more than 140 sales, followed by cooking oil, Tupperware, chopsticks, and bento box.

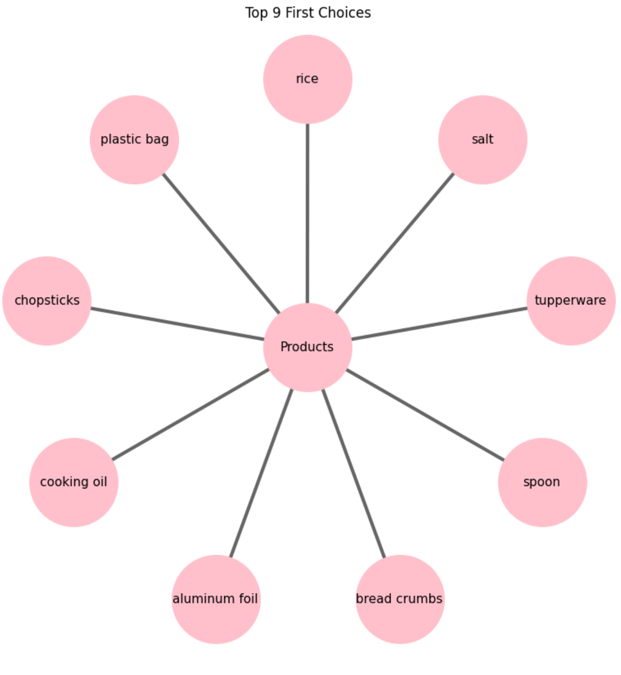


Fig. 9. Top 9 first choice items semantic map

The semantic map summarizes the top first choice items of the customers. These products are rice, salt, Tupperware, spoon, bread crumbs, aluminum foil, cooking oil, chopsticks, and plastic bag.

B. *Association Rules Data Visualization*

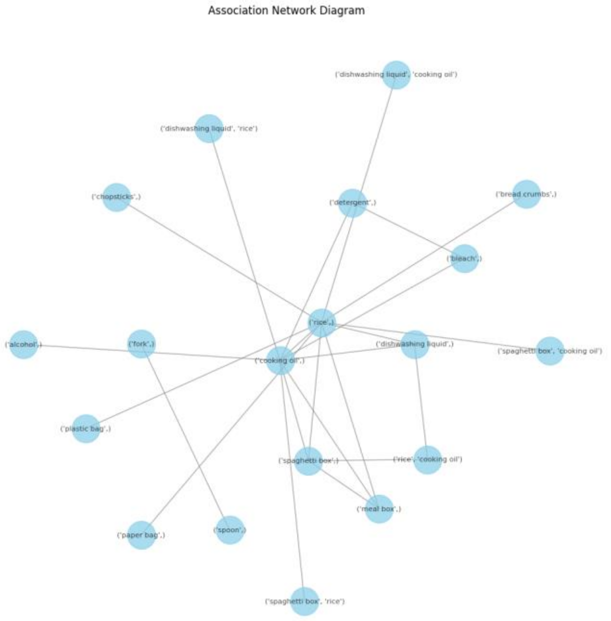
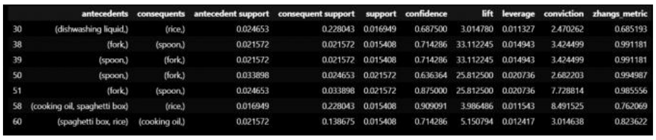


Fig. 10. Association Network Diagram

The figure is a network diagram that indicates the associations between each item in the dataset, prior to the association rules filtering process.

C. *Market Basket Analysis Result of CMC Rice Center and General Merchandise*

Following the association rules' filtering base on lift, confidence, and support, the association analysis has identified several rules suited for Market Basket.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
30	(dishwashing liquid,)	(rice,)	0.024653	0.228043	0.016949	0.687500	3.014780	0.011327	2.470262	0.685193
38	(fork,)	(spoon,)	0.021572	0.021572	0.015408	0.714286	33.112245	0.014943	3.424499	0.991181
39	(spoon,)	(fork,)	0.021572	0.021572	0.015408	0.714286	33.112245	0.014943	3.424499	0.991181
50	(spoon,)	(fork,)	0.033898	0.024653	0.021572	0.636364	25.812500	0.020736	2.682203	0.994987
51	(fork,)	(spoon,)	0.024653	0.033898	0.021572	0.875000	25.812500	0.020736	7.728814	0.985556
58	(cooking oil, spaghetti box)	(rice,)	0.016949	0.228043	0.015408	0.909091	3.986486	0.011543	8.491525	0.762069
60	(spaghetti box, rice)	(cooking oil,)	0.021572	0.138675	0.015408	0.714286	5.150794	0.012417	3.014638	0.823622

Fig. 11. Market Basket Analysis Result of CMC Rice Center and General Merchandise

The following are the association rules that can be extracted in the analysis as can be shown in Figure 10.

Association Rule 1:

- If a customer purchased 'dishwashing liquid,' then there is a 69% confidence that they will also purchase 'rice.'
- (Support = 0.017, Confidence = 69%, Lift = 3.014)
- Interpretation: This rule suggests that when customers buy 'dishwashing liquid,' there is a likelihood of 69% that they will also buy 'rice.' The lift of 3.014 suggests a high degree of dependency.

Association Rule 2:

- If a customer purchases 'spoon,' then there is a 71% confidence that they will also purchase 'fork,' and vice versa.

- (Support = 0.015, Confidence = 71%, Lift = 33)
- Interpretation: This rule indicates a strong relationship between the purchase of ‘spoon’ and ‘fork,’ The lift of 33 is exceptionally high, suggesting a strong relationship between the antecedent and consequent items.

Association Rule 3:

- There is a 91% confidence that customers will buy ‘cooking oil,’ ‘spaghetti,’ and ‘rice’ together in one purchase.
- (Support = 0.015, Confidence = 91%, Lift = 4)
- Interpretation: This rule indicates a strong association among the purchase of ‘cooking oil,’ ‘spaghetti,’ and ‘rice.’ The lift of 4 suggests a high degree of dependency.

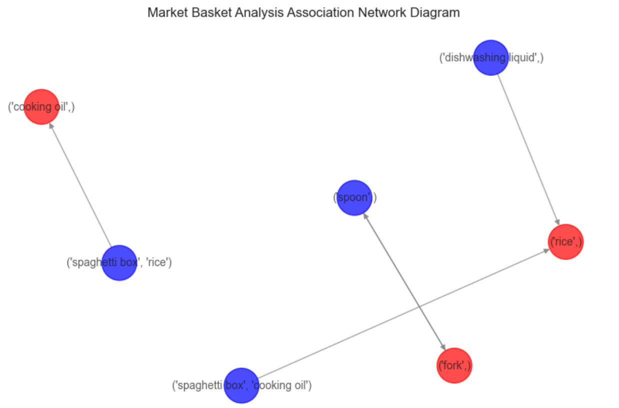

Market Basket Analysis Association Network Diagram
('dishwashing liquid',)
('cooking oil',)
('spoon',)
('rice',)
('spaghetti box', 'rice')
('fork',)
('spaghetti box', 'cooking oil')

Fig. 12. Market Basket Analysis Association Network Diagram of CMC Rice Center and General Merchandise

The figure shows the direct relationships of the association rules found after the Market Basket Analysis for CMC Rice Center and General Merchandise. The items in blue nodes are the antecedents. While the items in red nodes are the consequents. The arrow represents the items’ relationships, if they are antecedents and consequents to each other or not.

VI. RECOMMENDATION

MBA generates the frequent itemset, i.e., association rules, which can easily tell the customer buying behavior, and the retailer, with the help of these concepts, can easily develop the business in the future [7]. Apriori is the approach implemented by the proponents using Jupyter Notebook. From the results, the rules have sales data from 2020 to 2022 and from the retailer’s paper-based system [21]. For the results to be updated in near real-time, it is recommended that the retailer designs a point-of-sale system where the data collected from it may be integrated into the implementation of MBA [4, 22].

VII. REFERENCES

[1] S. Aby, “DATA SCIENCE IN MARKETING: Market Basket Analysis,” www.linkedin.com, Oct. 03, 2021. https://www.linkedin.com/pulse/data-science-marketing-market-basket-analysis-sara-aby-phd (accessed Nov. 27, 2023).

[2] N. Selvaraj, “How to Perform Market Basket Analysis,” 365 Data Science, Apr. 23, 2023. https://365datascience.com/tutorials/python-tutorials/market-basket-analysis/ (accessed Nov. 27, 2023).

[3] Y. Lim, “Data Mining: Market Basket Analysis with Apriori Algorithm,” Medium, Apr. 08, 2022. https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori-algorithm-970ff256a92c (accessed Nov. 27, 2023).

[4] Maheshwari, Data Analytics Made Accessible. Anil K. Maheshwari, Ph.D., 2020.

[5] N. Khadka, “The Ultimate Guide to Association Rule Analysis - DataAspirant,” Dataaspirant - A Data Science Portal For Beginners, Mar. 03, 2023. https://dataaspirant.com/association-rule-analysis/

[6] N. A. H. M. Rosli and N. H. I. Teo, “MARKET BASKET ANALYSIS USING APRIORI ALGORITHM: GROCERY ITEMS RECOMMENDATION,” ADVANCED INTERNATIONAL JOURNAL OF BUSINESS, ENTREPRENEURSHIP AND SME’s, vol. 4, no. 14, pp. 01–09, Dec. 2022, doi: 10.35631/aijbes.414001.

[7] Behera, Fartale, Bhagat, and Sharma, “Market Basket Analysis based on frequent Itemset Mining,” 2018. https://www.irjet.net/archives/V5/i2/IRJET-V5I297.pdf

[8] Great Learning Team, “What is Apriori Algorithm in Data Mining Implementation and Examples?,” Great Learning Blog: Free Resources What Matters to Shape Your Career!, Jul. 14, 2023. https://www.mygreatlearning.com/blog/apriori-algorithm-explained/

[9] “What is data processing in research? - Cint,” CintTM, Mar. 10, 2021. https://www.cint.com/blog/what-is-data-processing-in-research

[10] A. Banu, “What is Data Processing?,” EDUCBA, Jul. 2023, [Online]. Available: https://www.educba.com/what-is-data-processing/

[11] “The Jupyter Notebook — Jupyter Notebook 7.0.6 documentation.” https://jupyter-notebook.readthedocs.io/en/stable/notebook.html

[12] Manpreet Kaur, Shivani Kang, Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining, Procedia Computer Science, Volume 85, 2016, Pages 78-85, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.05.180.

[13] Raorane AA, Kulkarni RV, Jitkar BD. Association Rule – Extracting Knowledge Using Market Basket Analysis.Research Journal of Recent Sciences 2012:1(2):19-27.

[14] M. H. Santoso, “Application of Association Rule Method Using Apriori Algorithm to Find Sales Patterns Case Study of Indomaret Tanjung Anom,” Brilliance, vol. 1, no. 2, pp. 54–66, doi: 10.47709/brilliance.v1i2.1228.

[15] Shan, Haiyan, Chen Zhang, and Guo Wei. 2020. "Bundling or Unbundling? Pricing Strategy for Complementary Products in a Green Supply Chain" Sustainability 12, no. 4: 1331. https://doi.org/10.3390/su12041331

[16] Giri, R.N.; Mondal, S.K.; Maiti, M. Bundle pricing strategies for two complementary products with different channel powers. Ann. Oper. Res. 2017, 1–25

[17] S. Chaudhary, “Market Basket Analysis: Anticipating Customer Behavior,” Feb. 11, 2022. https://www.turing.com/kb/market-basket-analysis

[18] A. Kadlaskar, “Market Basket Analysis: A Comprehensive guide for businesses,” *Analytics Vidhya*, Nov. 16, 2023. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

[19] “How does the Apriori Algorithm work? | Data Basecamp,” *Data Basecamp*, Aug. 27, 2022. https://databasecamp.de/en/ml/apriori-algorithm

[20] Isa, Yusof, and Ramlan, “The Implementation of Data Mining Techniques for Sales Analysis using Daily Sales Data,” 2019. https://www.warse.org/IJATCSE/static/pdf/file/ijatcse16815sl2019.pdf

[21] G. C. Mooy and S. M. Isa, “Contextual Market Basket Analysis during Covid-19,” Journal of Social Science, vol. 4, no. 3, pp. 815–825, May 2023, doi: 10.46799/jss.v4i3.577.

[22] Simplilearn, “What is a data warehouse: overview, concepts and how it works,” Simplilearn.com, Aug. 10, 2023. https://www.simplilearn.com/data-warehouse-article#:~:text=Data%20Warehousing%20integrates%20data%20and,%2C%20website%2C%20and%20comment%20cards.