Name: Vaibhav Bora

NJIT ucid: vb22

Email Address: vb22@njit.edu

13 October 2024

Professor: Yasser Abduallah

CS 634-101 Data Mining

Midterm Project Report

Project Technical Overview

In this project, I leveraged OpenAI's ChatGPT to generate a unique and random dataset alongside using the example datasets provided by the professor. This approach allowed me to model realistic retail transaction scenarios.

Dataset Creation and Management

The dataset was meticulously curated using ChatGPT, which provided a foundation for generating a diverse array of itemsets representative of typical retail purchases. These itemsets were carefully crafted and subsequently edited in Microsoft Excel, enabling precise manipulation and structuring of data to fit the requirements of our analysis. The final dataset was exported as a CSV file, ensuring compatibility and ease of integration with our data processing tools.

Development and Implementation Tools

The primary programming and editing I conducted occurred in Jupyter Notebook, a powerful tool that supports interactive data science and scientific computing. This environment boasted an interactive approach to coding and allowed for immediate feedback and continuous refinement of the python code. During the final verification, I ran the code in IDLE, Python's integrated development and learning environment. This step was very important for ensuring the reliability of my code.

Key Technologies Used

Microsoft Excel: Generated and downloaded all the datasets and CSV files from here.

Jupyter Notebook: Served as my main platform for developing, testing, and visualizing the algorithm.

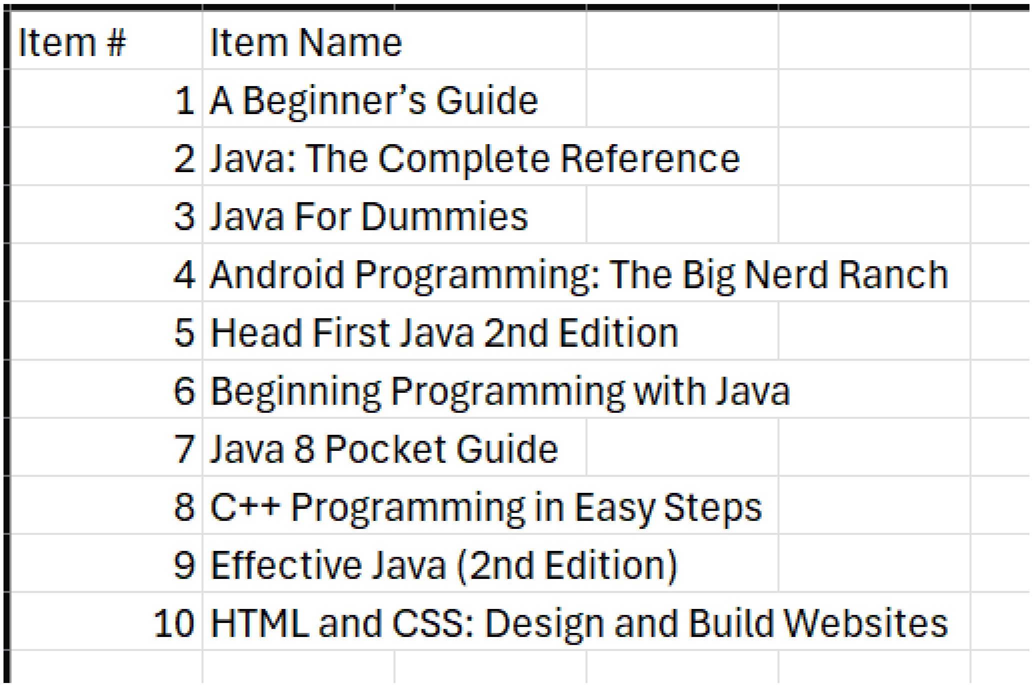
Python IDLE: Used in the final stages of the project to verify the completeness of the code.

ChatGPT: Used for generating a tailored dataset with randomized retail transactions and helping understanding the python libraries utilized for the verification.

Conclusion

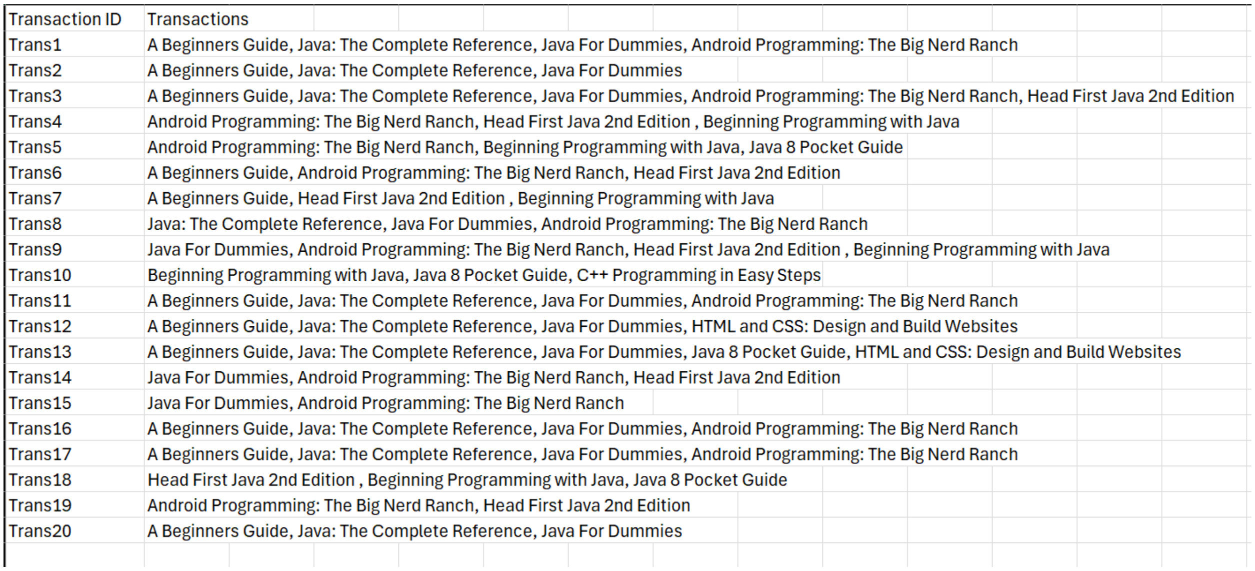
This project not only showed me the application of the Apriori Algorithm in getting helpful insights from a retail dataset but also the ability to utilize various different tools to work out the data mining process.

Screenshots



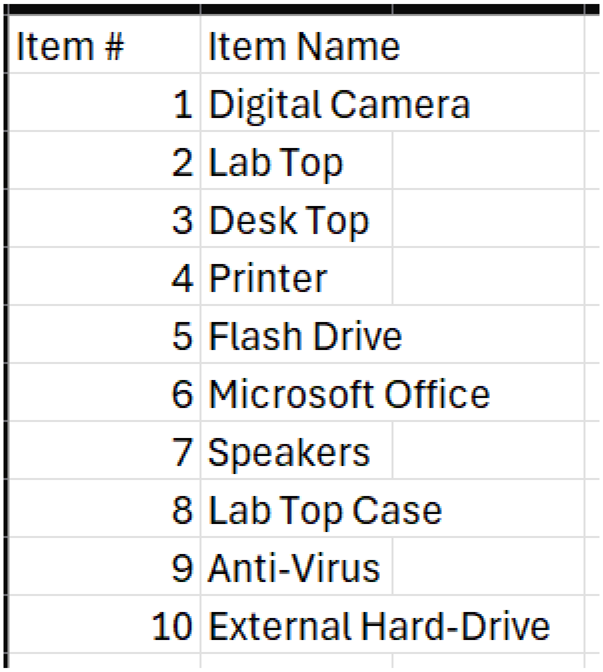
Item #	Item Name
1	A Beginner’s Guide
2	Java: The Complete Reference
3	Java For Dummies
4	Android Programming: The Big Nerd Ranch
5	Head First Java 2nd Edition
6	Beginning Programming with Java
7	Java 8 Pocket Guide
8	C++ Programming in Easy Steps
9	Effective Java (2nd Edition)
10	HTML and CSS: Design and Build Websites

Figure 1: Amazon Item Names.csv



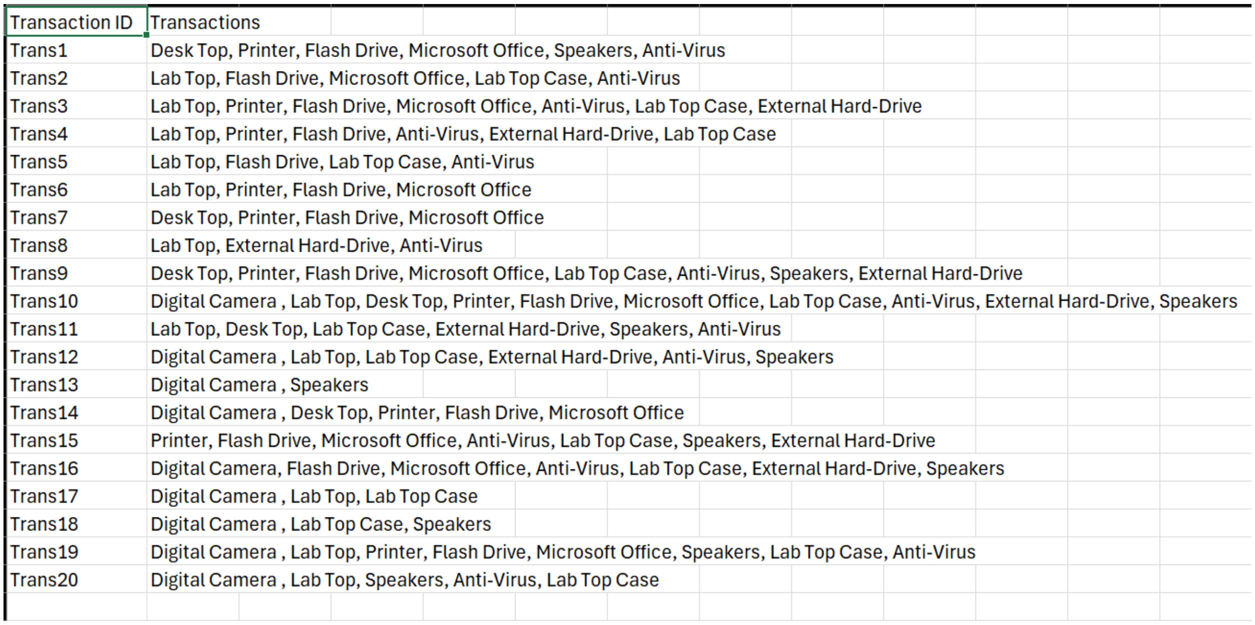
Transaction ID	Transactions
Trans1	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans2	A Beginners Guide, Java: The Complete Reference, Java For Dummies
Trans3	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans4	Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java
Trans5	Android Programming: The Big Nerd Ranch, Beginning Programming with Java, Java 8 Pocket Guide
Trans6	A Beginners Guide, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans7	A Beginners Guide, Head First Java 2nd Edition , Beginning Programming with Java
Trans8	Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans9	Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition , Beginning Programming with Java
Trans10	Beginning Programming with Java, Java 8 Pocket Guide, C++ Programming in Easy Steps
Trans11	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans12	A Beginners Guide, Java: The Complete Reference, Java For Dummies, HTML and CSS: Design and Build Websites
Trans13	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Java 8 Pocket Guide, HTML and CSS: Design and Build Websites
Trans14	Java For Dummies, Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans15	Java For Dummies, Android Programming: The Big Nerd Ranch
Trans16	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans17	A Beginners Guide, Java: The Complete Reference, Java For Dummies, Android Programming: The Big Nerd Ranch
Trans18	Head First Java 2nd Edition , Beginning Programming with Java, Java 8 Pocket Guide
Trans19	Android Programming: The Big Nerd Ranch, Head First Java 2nd Edition
Trans20	A Beginners Guide, Java: The Complete Reference, Java For Dummies

Figure 2: Amazon Transactions.csv



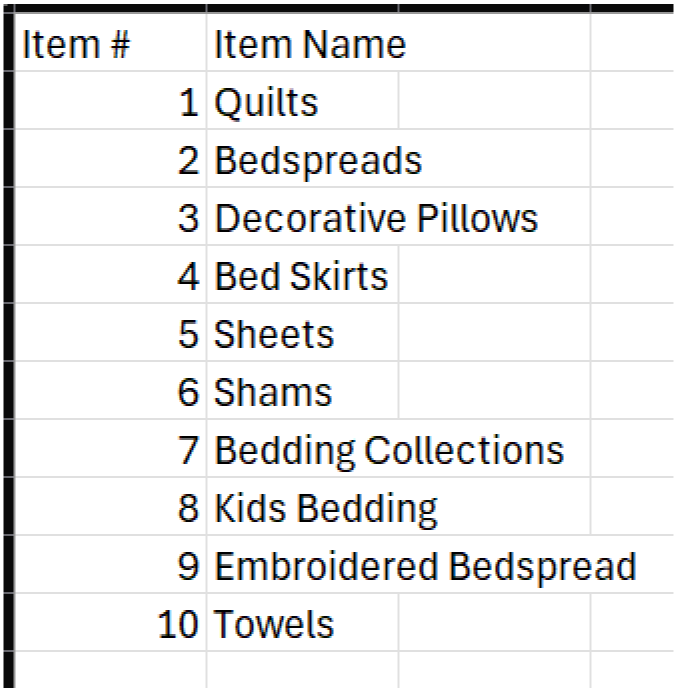
Item #	Item Name
1	Digital Camera
2	Lab Top
3	Desk Top
4	Printer
5	Flash Drive
6	Microsoft Office
7	Speakers
8	Lab Top Case
9	Anti-Virus
10	External Hard-Drive

Figure 3: Best Buy Item Names.csv



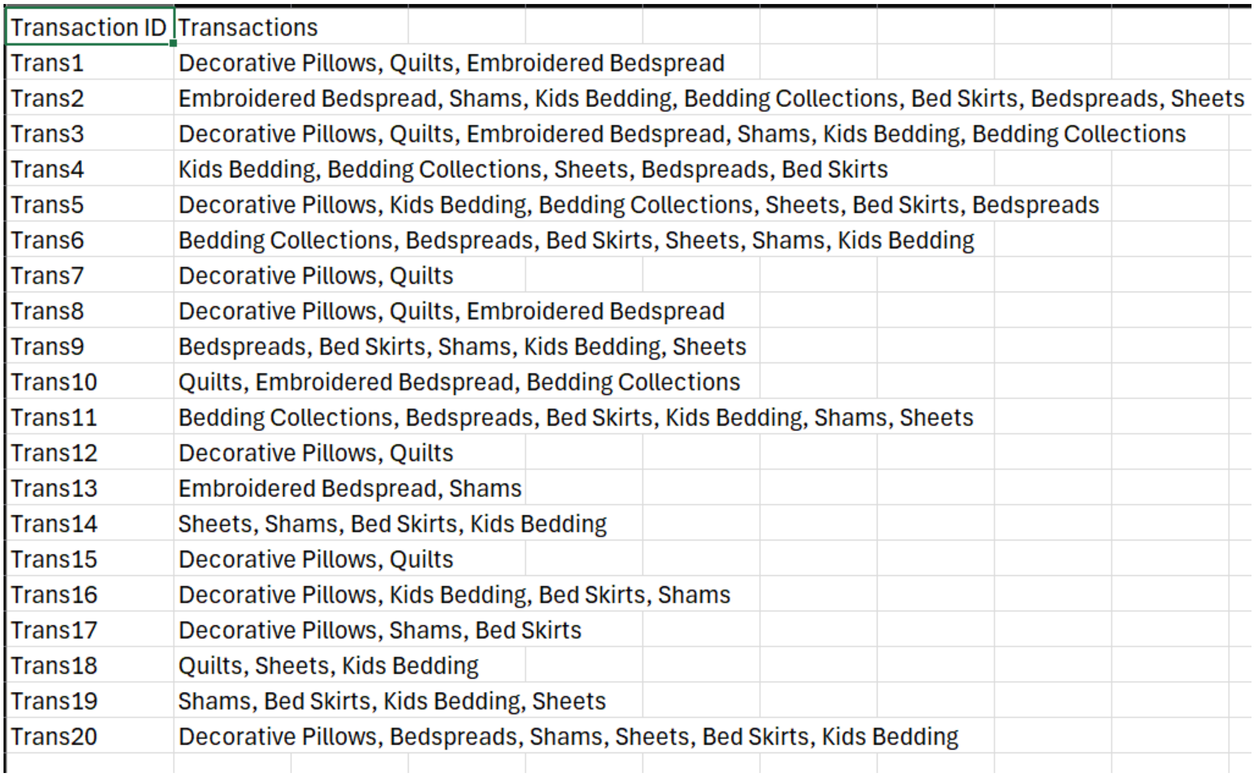
Transaction ID	Transactions
Trans1	Desk Top, Printer, Flash Drive, Microsoft Office, Speakers, Anti-Virus
Trans2	Lab Top, Flash Drive, Microsoft Office, Lab Top Case, Anti-Virus
Trans3	Lab Top, Printer, Flash Drive, Microsoft Office, Anti-Virus, Lab Top Case, External Hard-Drive
Trans4	Lab Top, Printer, Flash Drive, Anti-Virus, External Hard-Drive, Lab Top Case
Trans5	Lab Top, Flash Drive, Lab Top Case, Anti-Virus
Trans6	Lab Top, Printer, Flash Drive, Microsoft Office
Trans7	Desk Top, Printer, Flash Drive, Microsoft Office
Trans8	Lab Top, External Hard-Drive, Anti-Virus
Trans9	Desk Top, Printer, Flash Drive, Microsoft Office, Lab Top Case, Anti-Virus, Speakers, External Hard-Drive
Trans10	Digital Camera , Lab Top, Desk Top, Printer, Flash Drive, Microsoft Office, Lab Top Case, Anti-Virus, External Hard-Drive, Speakers
Trans11	Lab Top, Desk Top, Lab Top Case, External Hard-Drive, Speakers, Anti-Virus
Trans12	Digital Camera , Lab Top, Lab Top Case, External Hard-Drive, Anti-Virus, Speakers
Trans13	Digital Camera , Speakers
Trans14	Digital Camera , Desk Top, Printer, Flash Drive, Microsoft Office
Trans15	Printer, Flash Drive, Microsoft Office, Anti-Virus, Lab Top Case, Speakers, External Hard-Drive
Trans16	Digital Camera, Flash Drive, Microsoft Office, Anti-Virus, Lab Top Case, External Hard-Drive, Speakers
Trans17	Digital Camera , Lab Top, Lab Top Case
Trans18	Digital Camera , Lab Top Case, Speakers
Trans19	Digital Camera , Lab Top, Printer, Flash Drive, Microsoft Office, Speakers, Lab Top Case, Anti-Virus
Trans20	Digital Camera , Lab Top, Speakers, Anti-Virus, Lab Top Case

Figure 4: Best Buy Transactions.csv



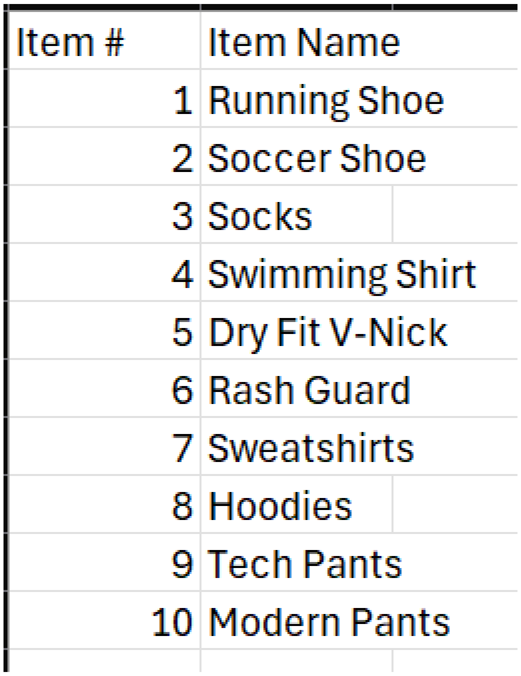
Item #	Item Name
1	Quilts
2	Bedspreads
3	Decorative Pillows
4	Bed Skirts
5	Sheets
6	Shams
7	Bedding Collections
8	Kids Bedding
9	Embroidered Bedspread
10	Towels

Figure 5: K-Mart Item Names.csv



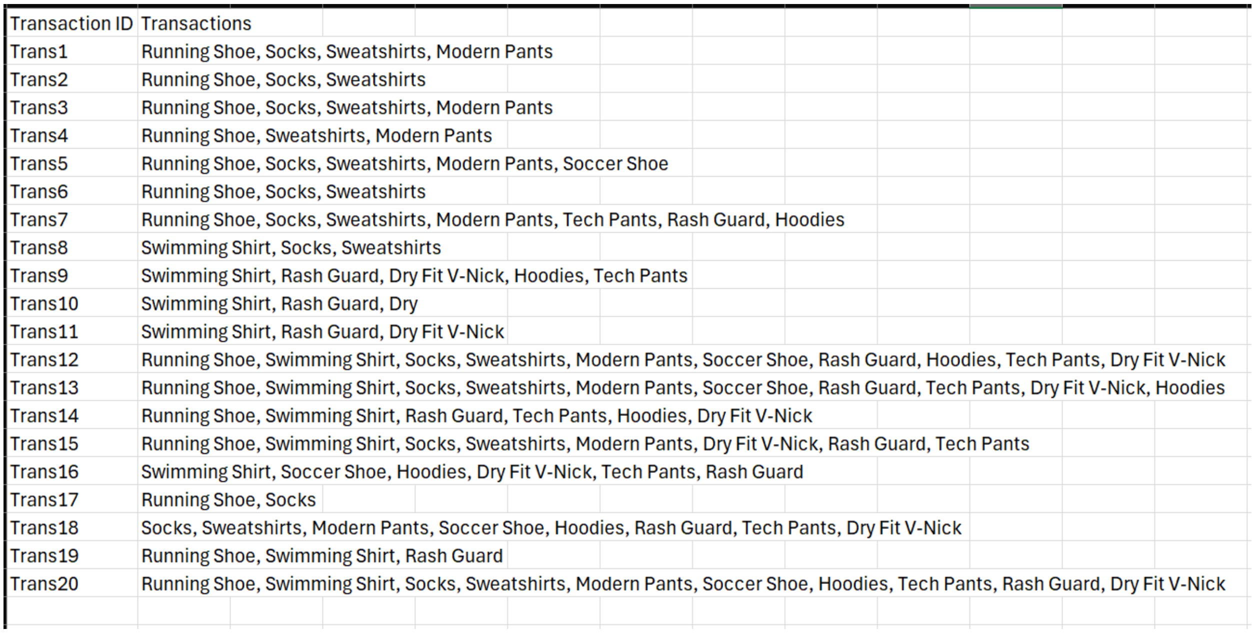
Transaction ID	Transactions
Trans1	Decorative Pillows, Quilts, Embroidered Bedspread
Trans2	Embroidered Bedspread, Shams, Kids Bedding, Bedding Collections, Bed Skirts, Bedspreads, Sheets
Trans3	Decorative Pillows, Quilts, Embroidered Bedspread, Shams, Kids Bedding, Bedding Collections
Trans4	Kids Bedding, Bedding Collections, Sheets, Bedspreads, Bed Skirts
Trans5	Decorative Pillows, Kids Bedding, Bedding Collections, Sheets, Bed Skirts, Bedspreads
Trans6	Bedding Collections, Bedspreads, Bed Skirts, Sheets, Shams, Kids Bedding
Trans7	Decorative Pillows, Quilts
Trans8	Decorative Pillows, Quilts, Embroidered Bedspread
Trans9	Bedspreads, Bed Skirts, Shams, Kids Bedding, Sheets
Trans10	Quilts, Embroidered Bedspread, Bedding Collections
Trans11	Bedding Collections, Bedspreads, Bed Skirts, Kids Bedding, Shams, Sheets
Trans12	Decorative Pillows, Quilts
Trans13	Embroidered Bedspread, Shams
Trans14	Sheets, Shams, Bed Skirts, Kids Bedding
Trans15	Decorative Pillows, Quilts
Trans16	Decorative Pillows, Kids Bedding, Bed Skirts, Shams
Trans17	Decorative Pillows, Shams, Bed Skirts
Trans18	Quilts, Sheets, Kids Bedding
Trans19	Shams, Bed Skirts, Kids Bedding, Sheets
Trans20	Decorative Pillows, Bedspreads, Shams, Sheets, Bed Skirts, Kids Bedding

Figure 6: K-Mart Transactions.csv



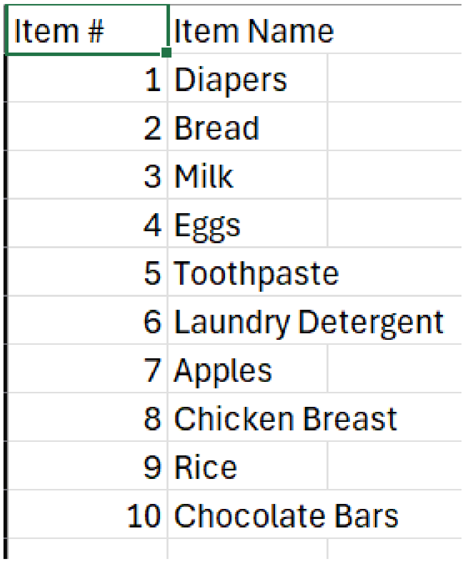
Item #	Item Name
1	Running Shoe
2	Soccer Shoe
3	Socks
4	Swimming Shirt
5	Dry Fit V-Nick
6	Rash Guard
7	Sweatshirts
8	Hoodies
9	Tech Pants
10	Modern Pants

Figure 7: Nike Item Names.csv



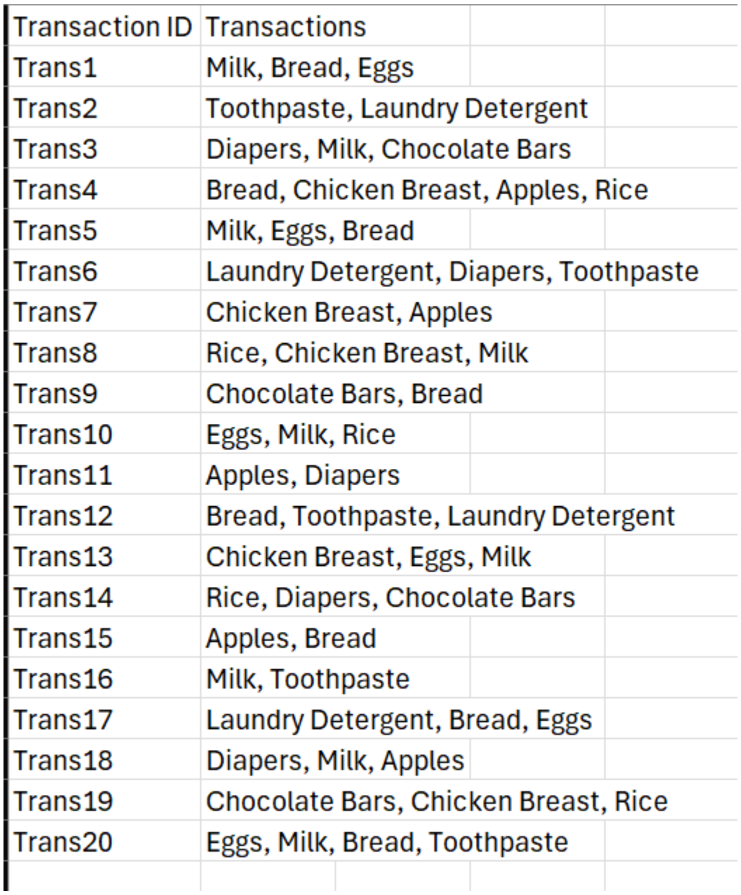
Transaction ID	Transactions
Trans1	Running Shoe, Socks, Sweatshirts, Modern Pants
Trans2	Running Shoe, Socks, Sweatshirts
Trans3	Running Shoe, Socks, Sweatshirts, Modern Pants
Trans4	Running Shoe, Sweatshirts, Modern Pants
Trans5	Running Shoe, Socks, Sweatshirts, Modern Pants, Soccer Shoe
Trans6	Running Shoe, Socks, Sweatshirts
Trans7	Running Shoe, Socks, Sweatshirts, Modern Pants, Tech Pants, Rash Guard, Hoodies
Trans8	Swimming Shirt, Socks, Sweatshirts
Trans9	Swimming Shirt, Rash Guard, Dry Fit V-Nick, Hoodies, Tech Pants
Trans10	Swimming Shirt, Rash Guard, Dry
Trans11	Swimming Shirt, Rash Guard, Dry Fit V-Nick
Trans12	Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Rash Guard, Hoodies, Tech Pants, Dry Fit V-Nick
Trans13	Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Rash Guard, Tech Pants, Dry Fit V-Nick, Hoodies
Trans14	Running Shoe, Swimming Shirt, Rash Guard, Tech Pants, Hoodies, Dry Fit V-Nick
Trans15	Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Dry Fit V-Nick, Rash Guard, Tech Pants
Trans16	Swimming Shirt, Soccer Shoe, Hoodies, Dry Fit V-Nick, Tech Pants, Rash Guard
Trans17	Running Shoe, Socks
Trans18	Socks, Sweatshirts, Modern Pants, Soccer Shoe, Hoodies, Rash Guard, Tech Pants, Dry Fit V-Nick
Trans19	Running Shoe, Swimming Shirt, Rash Guard
Trans20	Running Shoe, Swimming Shirt, Socks, Sweatshirts, Modern Pants, Soccer Shoe, Hoodies, Tech Pants, Rash Guard, Dry Fit V-Nick

Figure 8: Nike Transactions.csv



Item #	Item Name
1	Diapers
2	Bread
3	Milk
4	Eggs
5	Toothpaste
6	Laundry Detergent
7	Apples
8	Chicken Breast
9	Rice
10	Chocolate Bars

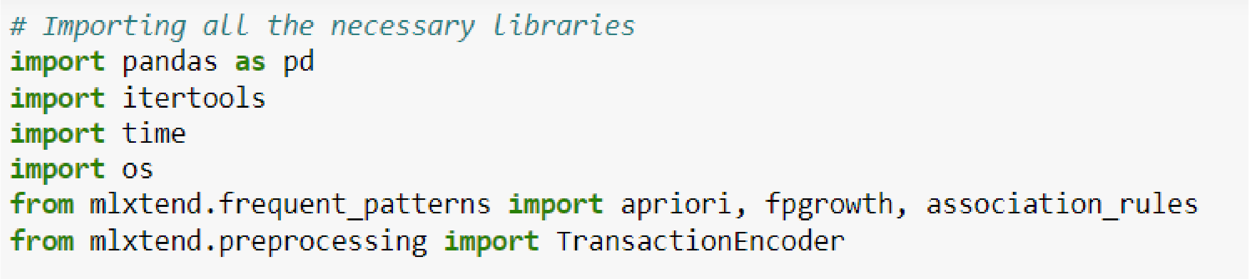
Figure 9: SuperMarket Item Names.csv



Transaction ID	Transactions
Trans1	Milk, Bread, Eggs
Trans2	Toothpaste, Laundry Detergent
Trans3	Diapers, Milk, Chocolate Bars
Trans4	Bread, Chicken Breast, Apples, Rice
Trans5	Milk, Eggs, Bread
Trans6	Laundry Detergent, Diapers, Toothpaste
Trans7	Chicken Breast, Apples
Trans8	Rice, Chicken Breast, Milk
Trans9	Chocolate Bars, Bread
Trans10	Eggs, Milk, Rice
Trans11	Apples, Diapers
Trans12	Bread, Toothpaste, Laundry Detergent
Trans13	Chicken Breast, Eggs, Milk
Trans14	Rice, Diapers, Chocolate Bars
Trans15	Apples, Bread
Trans16	Milk, Toothpaste
Trans17	Laundry Detergent, Bread, Eggs
Trans18	Diapers, Milk, Apples
Trans19	Chocolate Bars, Chicken Breast, Rice
Trans20	Eggs, Milk, Bread, Toothpaste

Figure 10: SuperMarket Transactions.csv

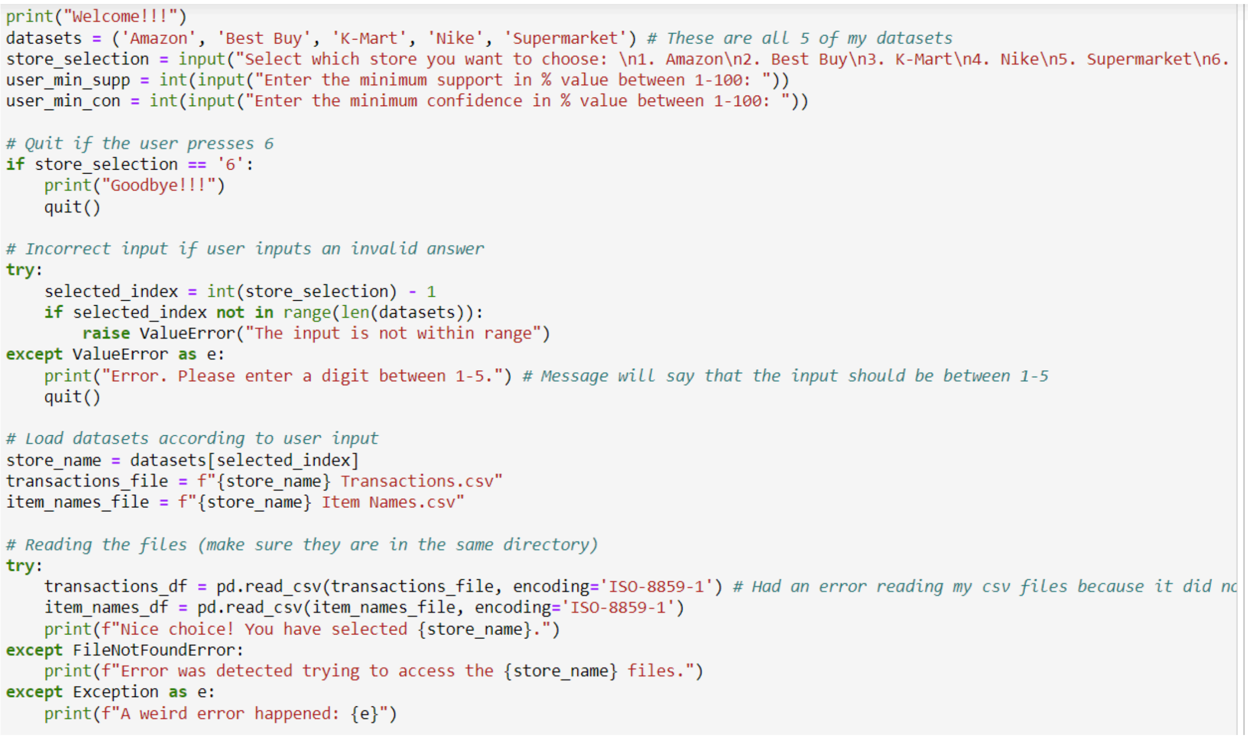
Below are the screenshots of the source code from Jupyter:



```
# Importing all the necessary libraries
import pandas as pd
import itertools
import time
import os
from mlxtend.frequent_patterns import apriori, fpgrowth, association_rules
from mlxtend.preprocessing import TransactionEncoder
```

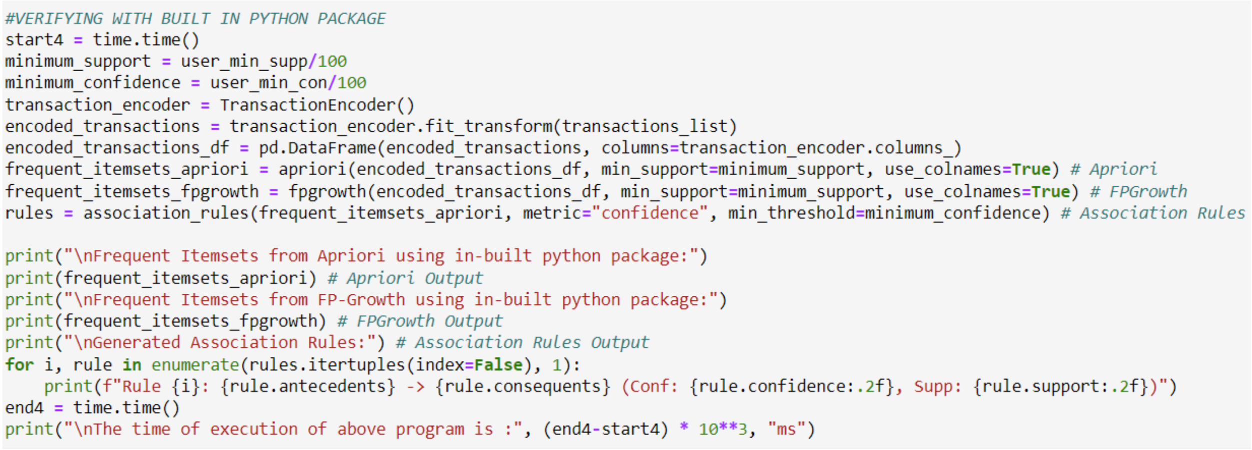
Figure 11: All the libraries necessary to run the program

In order to install these libraries such as pandas or mlxtend, just type ‘pip install pandas’ or ‘pip install mlxtend’ respectively in your command prompt.



```
print("Welcome!!!")
datasets = ('Amazon', 'Best Buy', 'K-Mart', 'Nike', 'Supermarket') # These are all 5 of my datasets
store_selection = input("Select which store you want to choose: \n1. Amazon\n2. Best Buy\n3. K-Mart\n4. Nike\n5. Supermarket\n6.
user_min_supp = int(input("Enter the minimum support in % value between 1-100: "))
user_min_con = int(input("Enter the minimum confidence in % value between 1-100: "))

# Quit if the user presses 6
if store_selection == '6':
    print("Goodbye!!!")
    quit()

# Incorrect input if user inputs an invalid answer
try:
    selected_index = int(store_selection) - 1
    if selected_index not in range(len(datasets)):
        raise ValueError("The input is not within range")
except ValueError as e:
    print("Error. Please enter a digit between 1-5.") # Message will say that the input should be between 1-5
    quit()

# Load datasets according to user input
store_name = datasets[selected_index]
transactions_file = f"{store_name} Transactions.csv"
item_names_file = f"{store_name} Item Names.csv"

# Reading the files (make sure they are in the same directory)
try:
    transactions_df = pd.read_csv(transactions_file, encoding='ISO-8859-1') # Had an error reading my csv files because it did nc
    item_names_df = pd.read_csv(item_names_file, encoding='ISO-8859-1')
    print(f"Nice choice! You have selected {store_name}.")
except FileNotFoundError:
    print(f"Error was detected trying to access the {store_name} files.")
except Exception as e:
    print(f"A weird error happened: {e}")
```

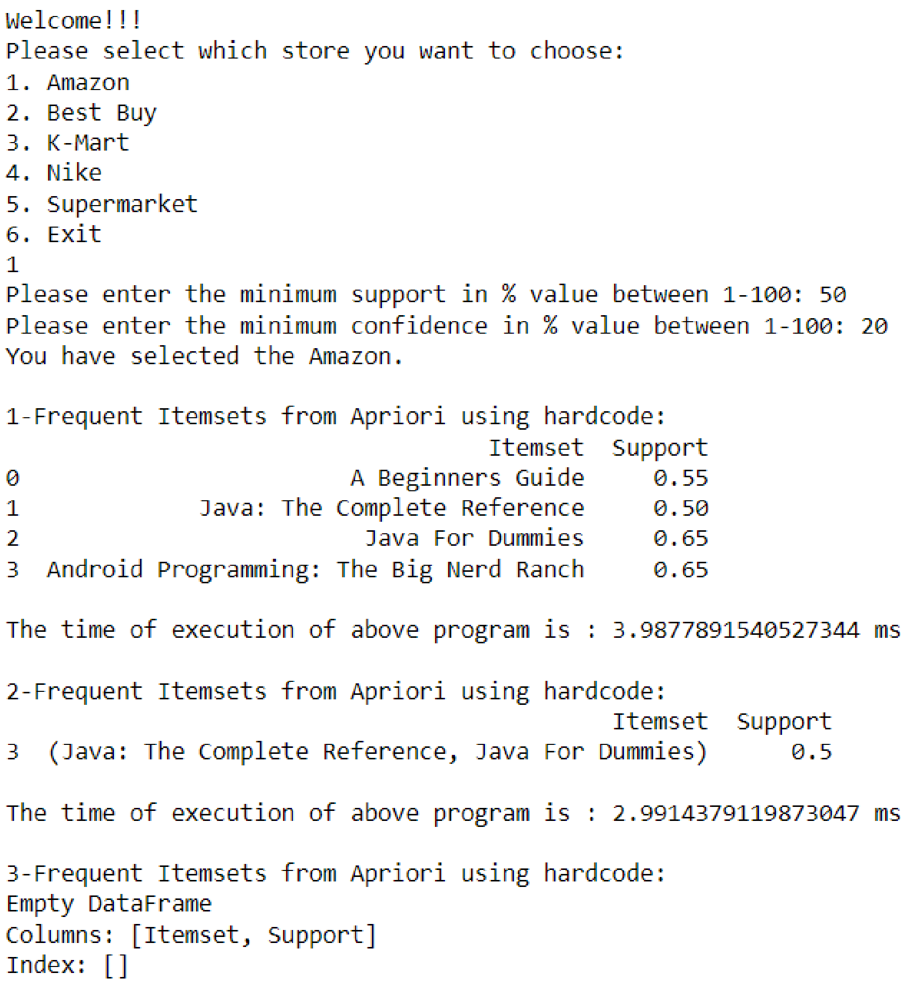
Figure 12: Welcome message



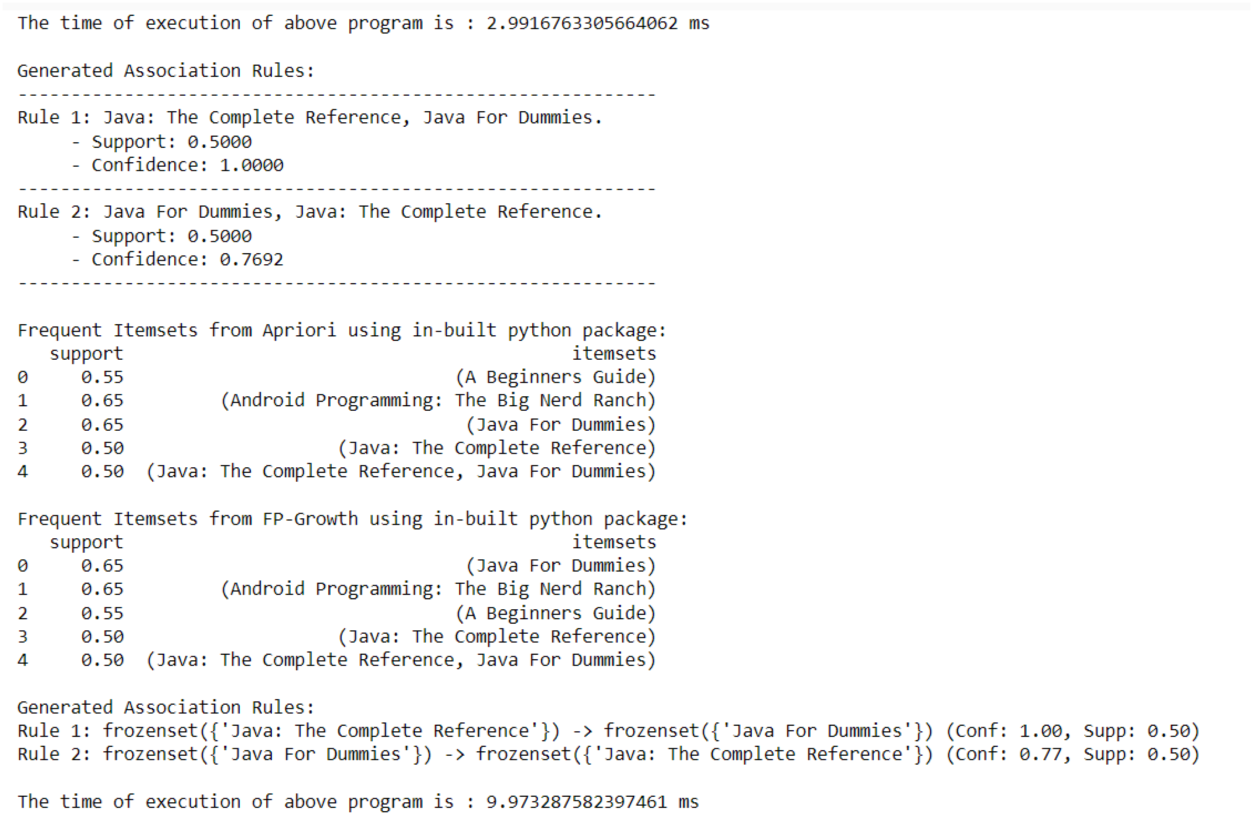
```python
#VERIFYING WITH BUILT IN PYTHON PACKAGE
start4 = time.time()
minimum_support = user_min_supp/100
minimum_confidence = user_min_con/100
transaction_encoder = TransactionEncoder()
encoded_transactions = transaction_encoder.fit_transform(transactions_list)
encoded_transactions_df = pd.DataFrame(encoded_transactions, columns=transaction_encoder.columns_)
frequent_itemsets_apriori = apriori(encoded_transactions_df, min_support=minimum_support, use_colnames=True) # Apriori
frequent_itemsets_fpgrowth = fpgrowth(encoded_transactions_df, min_support=minimum_support, use_colnames=True) # FPGrowth
rules = association_rules(frequent_itemsets_apriori, metric="confidence", min_threshold=minimum_confidence) # Association Rules

print("\nFrequent Itemsets from Apriori using in-built python package:")
print(frequent_itemsets_apriori) # Apriori Output
print("\nFrequent Itemsets from FP-Growth using in-built python package:")
print(frequent_itemsets_fpgrowth) # FPGrowth Output
print("\nGenerated Association Rules:") # Association Rules Output
for i, rule in enumerate(rules.itertuples(index=False), 1):
    print(f"Rule {i}: {rule.antecedents} -> {rule.consequents} (Conf: {rule.confidence:.2f}, Supp: {rule.support:.2f})")
end4 = time.time()
print("\nThe time of execution of above program is :", (end4-start4) * 10**3, "ms")
```

Figure 13: Output code using built in python libraries

Here is one example of the output:



```
Welcome!!!
Please select which store you want to choose:
1. Amazon
2. Best Buy
3. K-Mart
4. Nike
5. Supermarket
6. Exit
1
Please enter the minimum support in % value between 1-100: 50
Please enter the minimum confidence in % value between 1-100: 20
You have selected the Amazon.

1-Frequent Itemsets from Apriori using hardcode:
                                     Itemset  Support
0                          A Beginners Guide     0.55
1              Java: The Complete Reference     0.50
2                           Java For Dummies     0.65
3  Android Programming: The Big Nerd Ranch     0.65

The time of execution of above program is : 3.9877891540527344 ms

2-Frequent Itemsets from Apriori using hardcode:
                                            Itemset  Support
3  (Java: The Complete Reference, Java For Dummies)      0.5

The time of execution of above program is : 2.9914379119873047 ms

3-Frequent Itemsets from Apriori using hardcode:
Empty DataFrame
Columns: [Itemset, Support]
Index: []
```



```
The time of execution of above program is : 2.9916763305664062 ms

Generated Association Rules:
------------------------------------------------------------
Rule 1: Java: The Complete Reference, Java For Dummies.
     - Support: 0.5000
     - Confidence: 1.0000
------------------------------------------------------------
Rule 2: Java For Dummies, Java: The Complete Reference.
     - Support: 0.5000
     - Confidence: 0.7692
------------------------------------------------------------

Frequent Itemsets from Apriori using in-built python package:
   support                                          itemsets
0     0.55                               (A Beginners Guide)
1     0.65         (Android Programming: The Big Nerd Ranch)
2     0.65                                (Java For Dummies)
3     0.50                    (Java: The Complete Reference)
4     0.50  (Java: The Complete Reference, Java For Dummies)

Frequent Itemsets from FP-Growth using in-built python package:
   support                                          itemsets
0     0.65                                (Java For Dummies)
1     0.65         (Android Programming: The Big Nerd Ranch)
2     0.55                               (A Beginners Guide)
3     0.50                    (Java: The Complete Reference)
4     0.50  (Java: The Complete Reference, Java For Dummies)

Generated Association Rules:
Rule 1: frozenset({'Java: The Complete Reference'}) -> frozenset({'Java For Dummies'}) (Conf: 1.00, Supp: 0.50)
Rule 2: frozenset({'Java For Dummies'}) -> frozenset({'Java: The Complete Reference'}) (Conf: 0.77, Supp: 0.50)

The time of execution of above program is : 9.973287582397461 ms
```

Github link: https://github.com/vaibhavbora11/Data-Mining-Midterm-Project