NAME: Srikar Choudary Katta   
NJIT UCID: sk3577  
Email Address: sk3577@njit.edu  
08/03/2024  
Professor: Yasser Abduallah  
CS 634 104- Data Mining

**Midterm Project Report**

Implementation and Code Usage

**Apriori Algorithm Implementation in Retail Data Mining**

Abstract:

In this project, I explore retail transaction data to find patterns using basic data analysis techniques. By looking at different retail datasets, I try to uncover connections and insights. I use simple methods to see how well they work and if they're useful in understanding transaction records. I also create some tools to help make sense of the data. The goal is to provide useful information for people making decisions in retail.

**Introduction:**

Data mining serves as a potent tool for unravelling concealed patterns and correlations nestled within vast datasets. In our project, we zero in on the Apriori Algorithm, a foundational technique for mining association rules, and explore its applicability within the realm of retail. This introduction delineates the fundamental data mining principles underpinning our endeavour, particularly focusing on the creation and exhibition of association rules.

The Apriori Algorithm operates on the premise of establishing associations among items within transactions. A pivotal initial step entails identifying the most frequently occurring items across transaction lists. Subsequently, the algorithm calculates the support for each item, gauging its occurrence frequency against user-specified support parameters. Items failing to meet the predefined support threshold are then eliminated, streamlining the association rule generation process.

This classic data mining algorithm employs a brute force methodology, iteratively expanding item sets and sieving out those failing to meet the minimum support criteria. Our implementation of the Apriori algorithm centres on a diverse array of retail datasets encompassing transactions from Amazon, Best Buy, Flipkart, Kmart, and Nike. Key procedural steps encompass initializing dictionaries for candidate and frequent item sets, loading dataset and itemset data from CSV files, preprocessing to ensure item order and uniqueness, and soliciting user input for minimum support and confidence thresholds.

Through this iterative process, we endeavour to extract meaningful insights from retail transaction data, unveiling frequent item sets and association rules. This exploration not only facilitates a deeper understanding of consumer behaviour but also furnishes valuable insights instrumental in informing strategic retail decisions.

**Core Concepts and Principles:**

Association Rule Mining:

Our project revolves around association rule mining, a crucial aspect of data mining focusing on discovering frequent item sets and extracting valuable insights into customer purchasing behavior and preferences. By identifying patterns of co-occurrence among items within retail transactions, we aim to uncover meaningful associations that can inform sales strategies and enhance customer experience.

Support and Confidence:

In our analysis, we rely on two fundamental metrics: support and confidence. Support measures the frequency of occurrence of item sets within transactions, while confidence assesses the strength of association between items. These metrics serve as guiding principles throughout our exploration, enabling us to filter out less significant patterns and focus on robust associations with actionable implications.

Apriori Algorithm Application:

Central to our project is the implementation of the Apriori Algorithm, a classic method for frequent itemset discovery. We follow a structured workflow, iteratively applying the algorithm to uncover frequent item sets and generate association rules from retail transaction data. This iterative approach involves progressively exploring larger itemset sizes, utilizing a "brute force" method to exhaustively search for significant associations.

Project Workflow:

Our project adheres to a systematic workflow, encompassing the following stages:

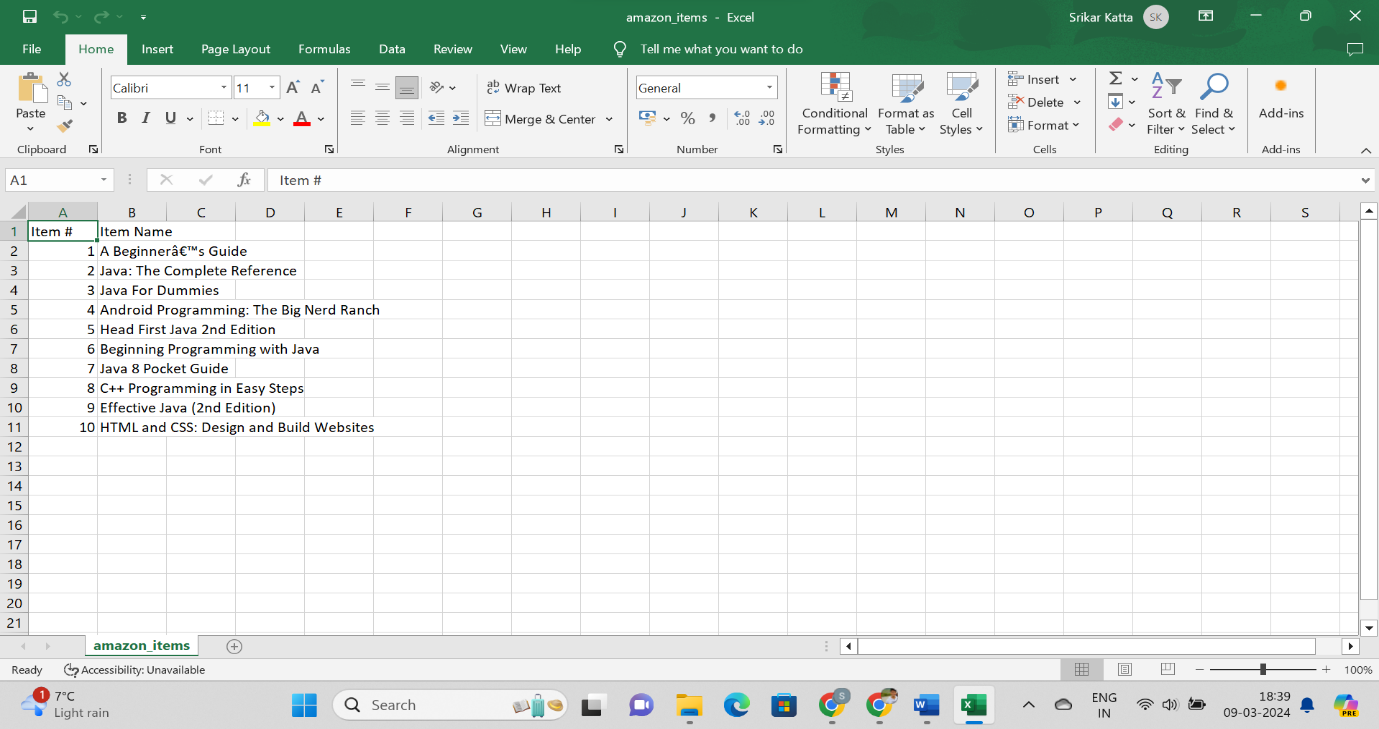
1. Data Loading and Preprocessing:
   * We start by loading transaction data from diverse retail sources, including Amazon, Best Buy, Flipkart, Kmart, and Nike. Each transaction comprises a list of items purchased by customers.
   * To ensure data integrity, we preprocess the dataset by filtering unique items and sorting them based on a predefined order.
2. Determination of Minimum Support and Confidence:
   * User input is solicited to define minimum support and confidence levels, guiding the identification of significant association rules.
3. Iteration Through Candidate Item sets:
   * The Apriori Algorithm is applied iteratively, generating candidate itemsets of increasing sizes (from single items to larger combinations) to uncover frequent item sets.
4. Support Count Calculation:
   * For each candidate itemset, we calculate its support by tallying the number of transactions containing the itemset. Item sets meeting the minimum support threshold are retained for further analysis.
5. Confidence Calculation:
   * We evaluate the confidence of association rules, comparing support values for individual items and item sets to gauge the strength of associations.
6. Association Rule Generation:
   * Association rules meeting predefined support and confidence requirements are extracted, providing insights into commonly co-purchased items and informing sales strategies.

Results and Evaluation:

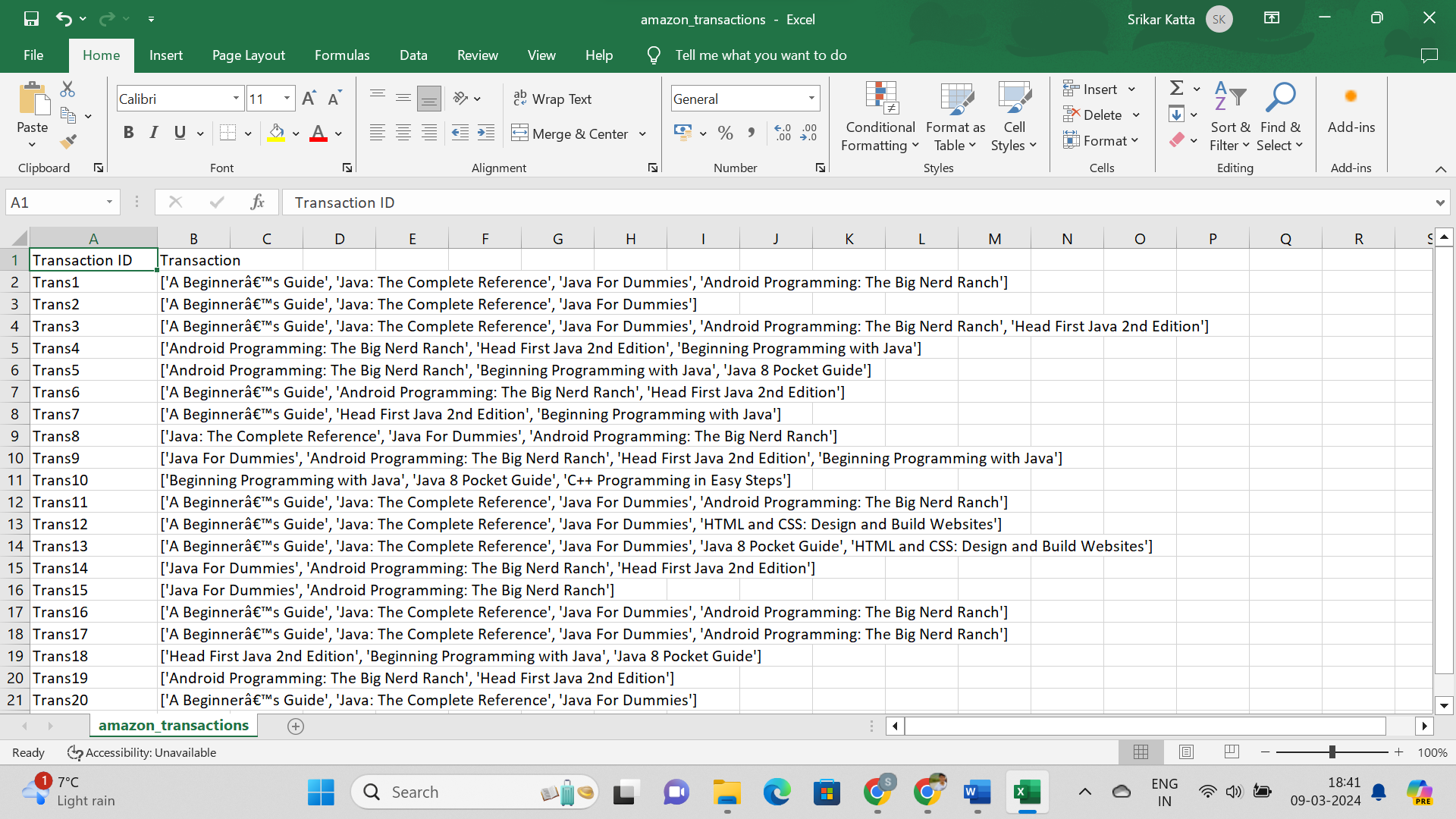
The effectiveness and efficiency of our project are assessed based on performance metrics such as support, confidence, and the resulting association rules. Additionally, we conduct a comparative analysis between our custom Apriori Algorithm implementation and existing libraries to evaluate reliability and performance.

Conclusion:

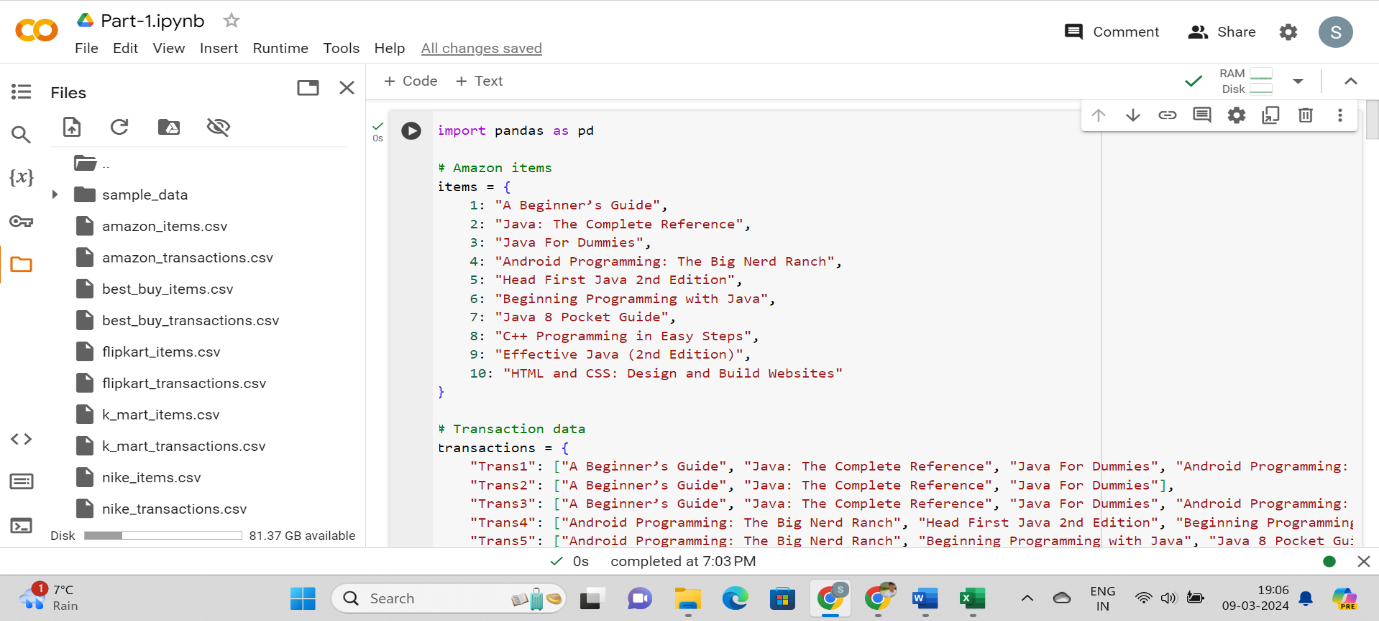
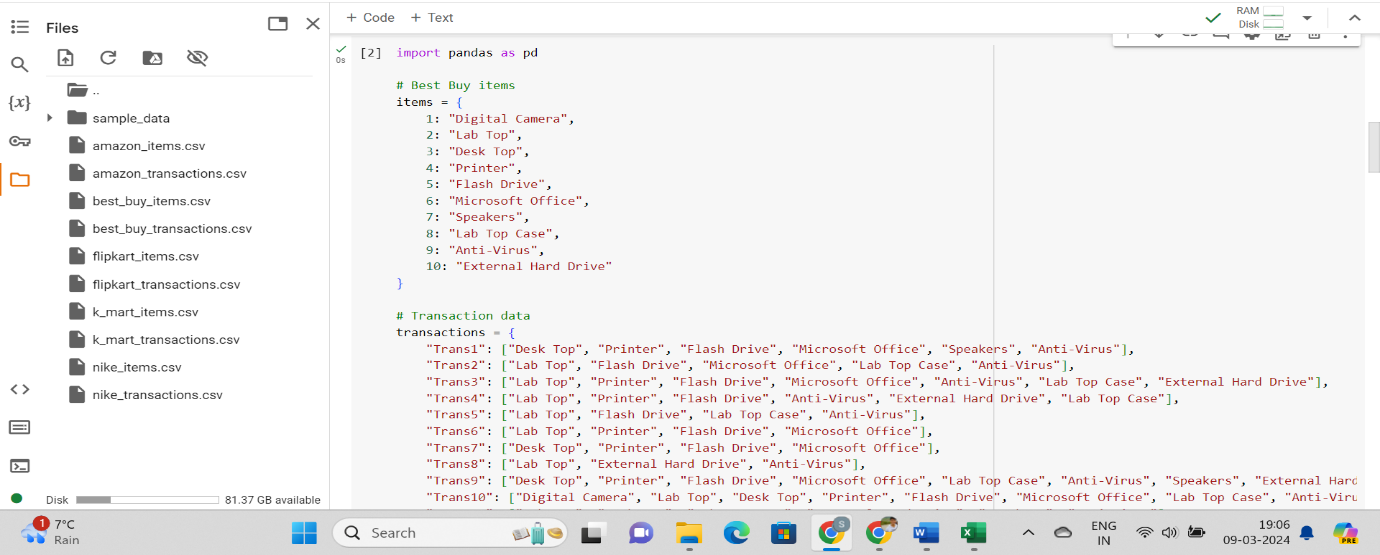
In conclusion, our project showcases the practical application of data mining principles, particularly the Apriori Algorithm, in uncovering valuable insights from retail transaction data. By adhering to a structured workflow and leveraging user-defined parameters, we demonstrate the power of data mining in informing decision-making and enhancing retail strategies.

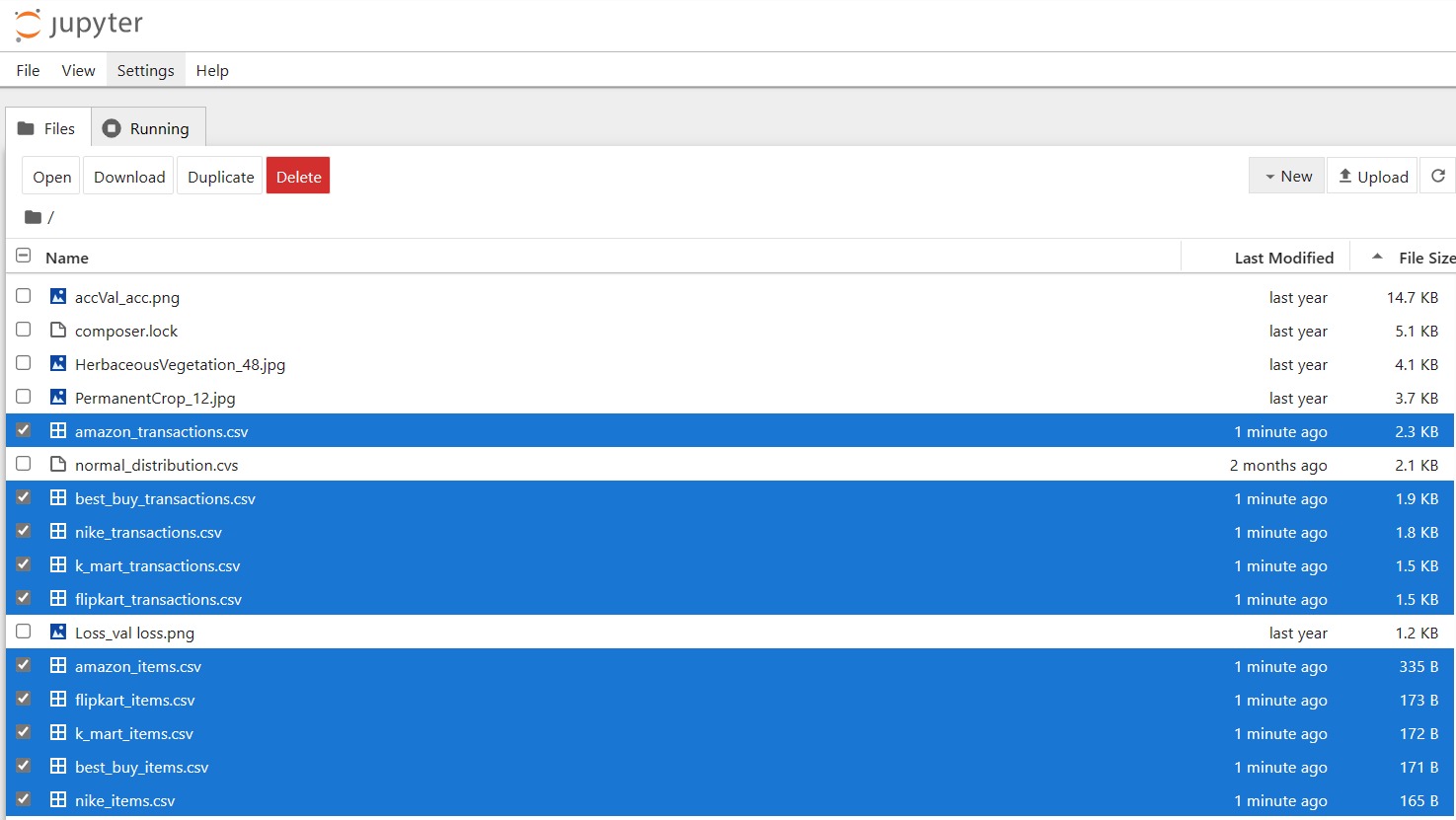
**Screenshots**

Amazon Item Names CSV File

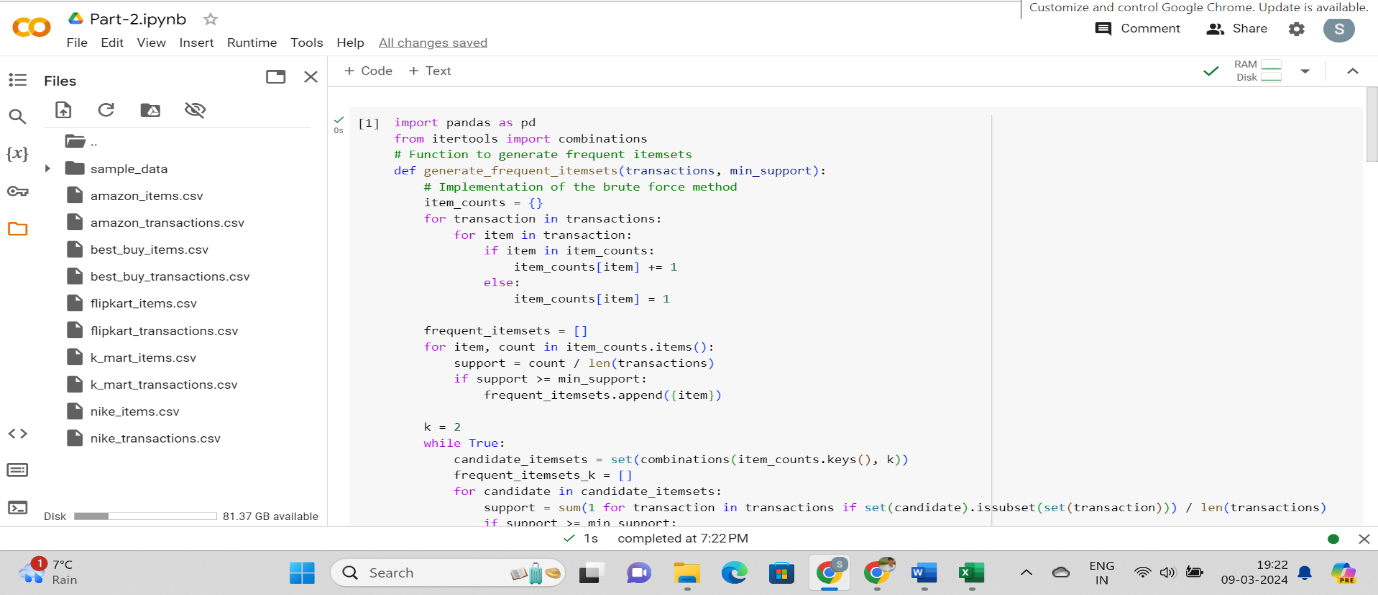
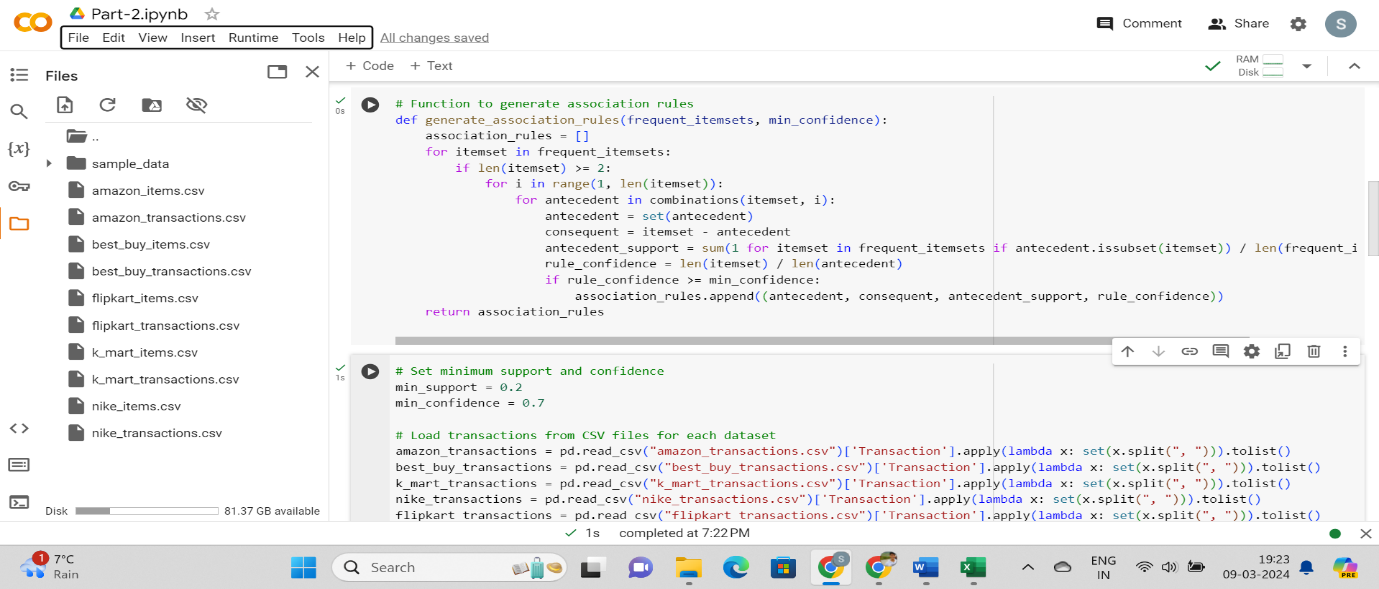


Amazon Transactions CSV File

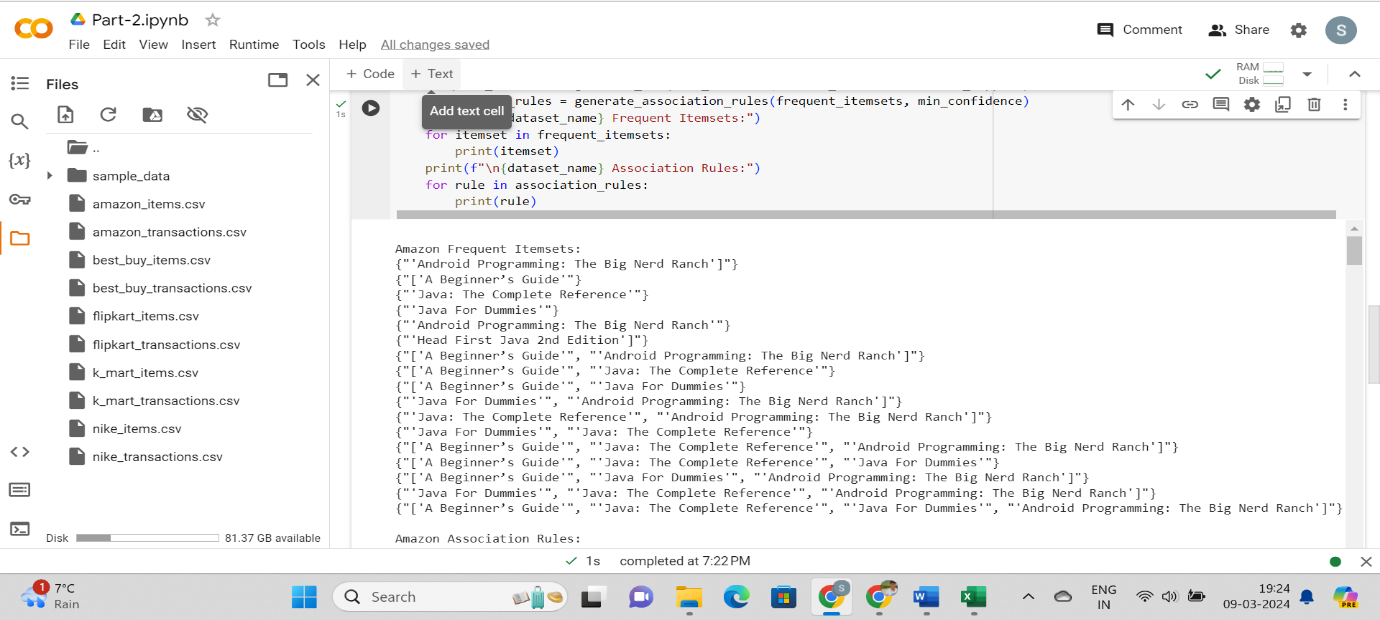
**Below are screenshots of the code from python file:**

PART – 1 CODE

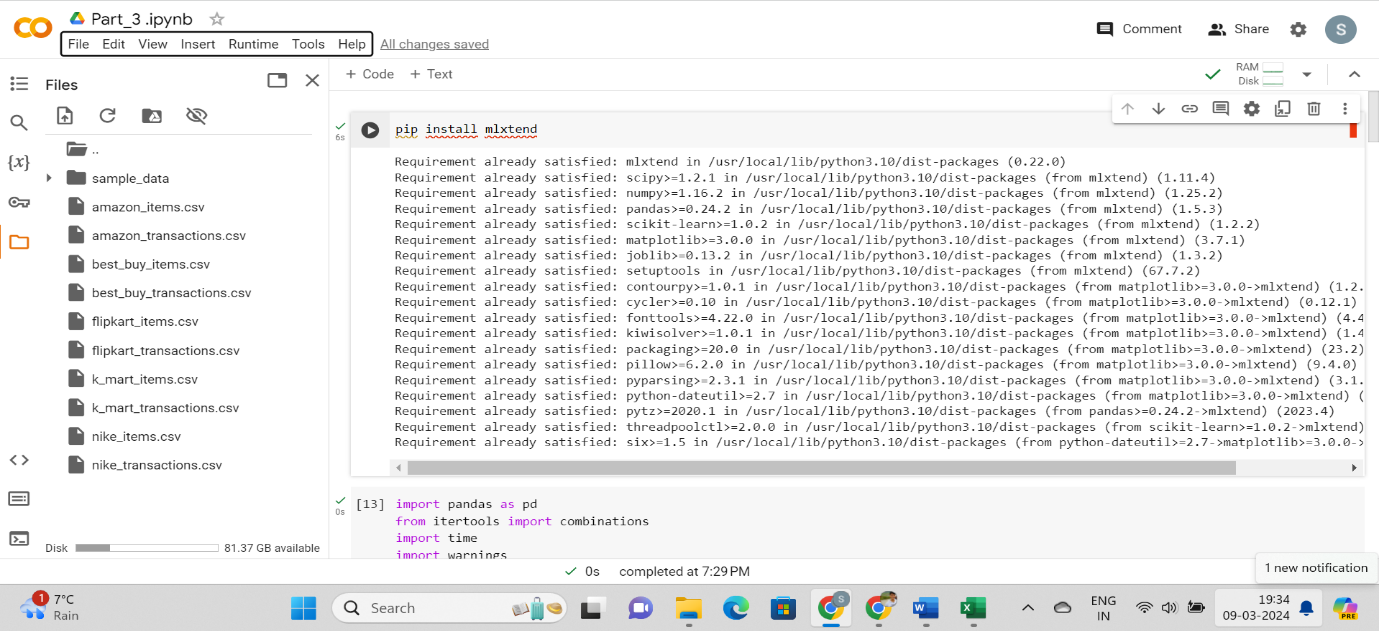
PART – 1 OUTPUT

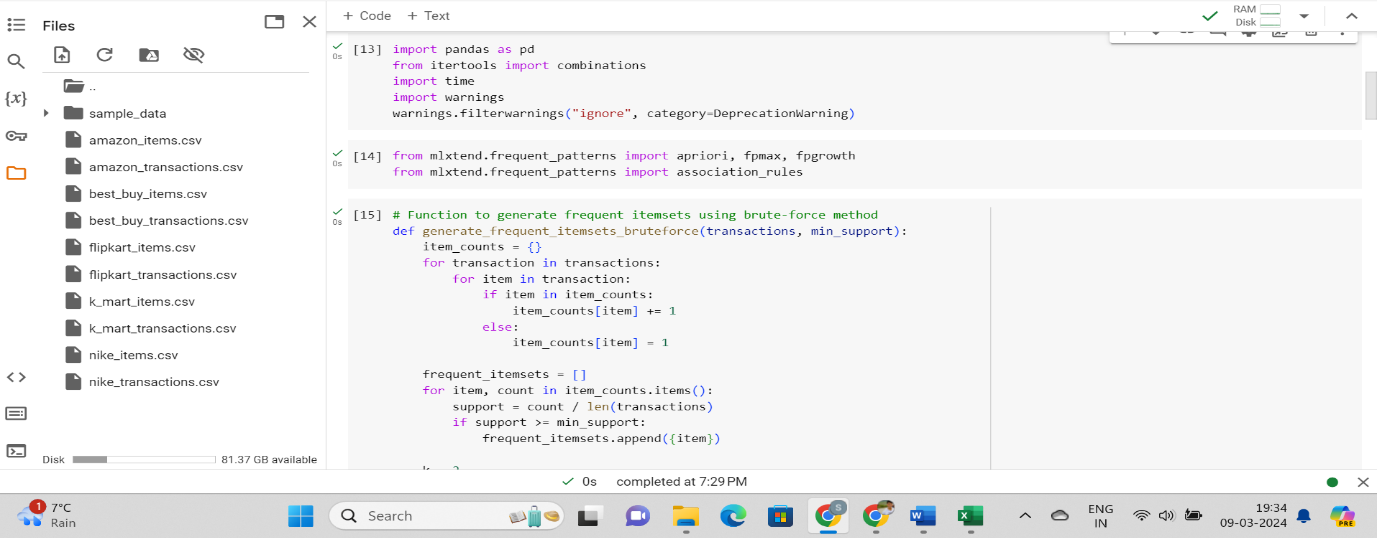


PART - 2 CODE

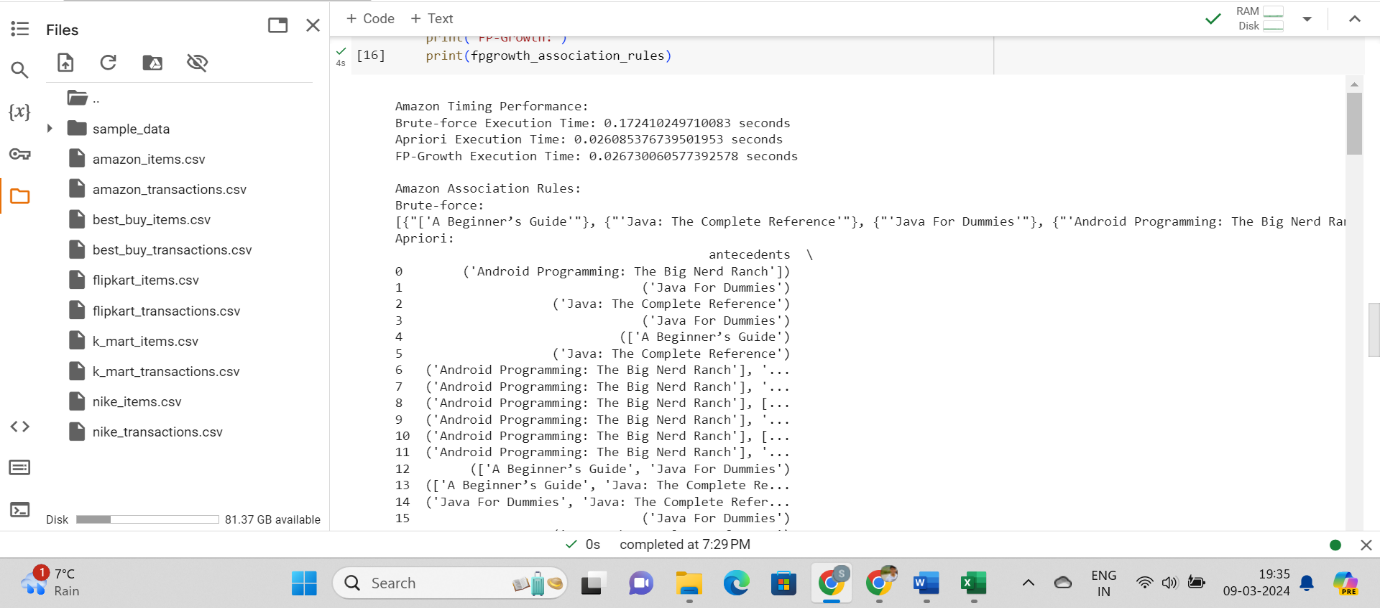


PART - 2 OUTPUT

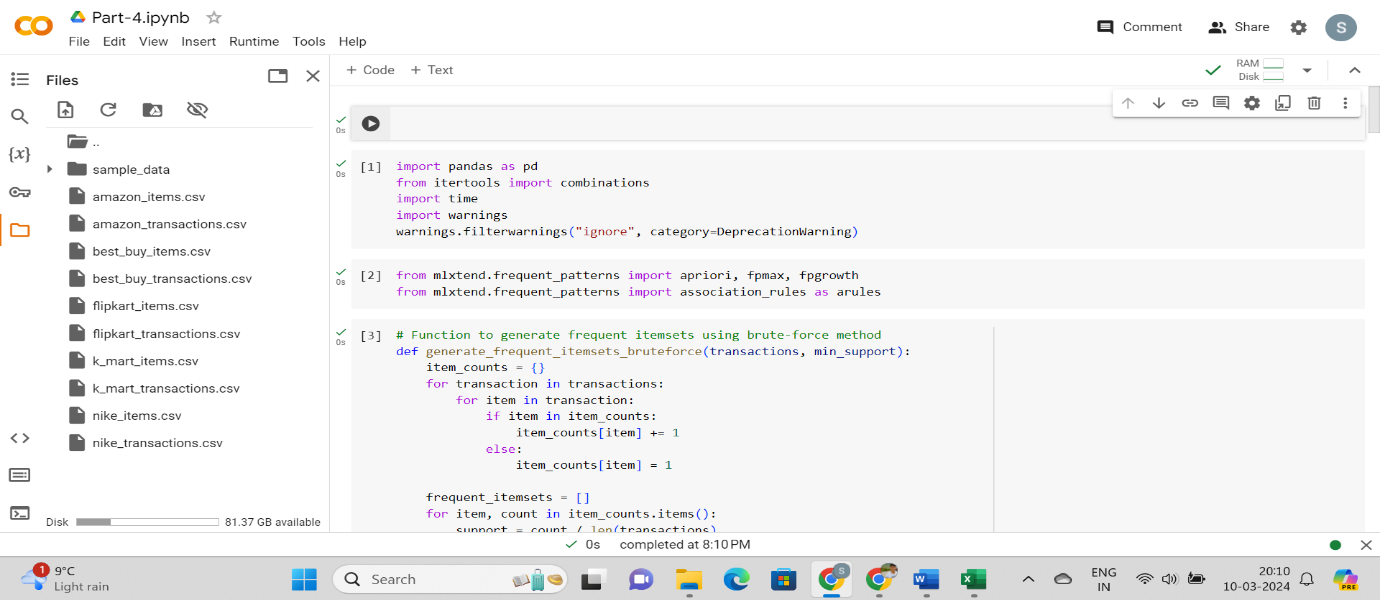


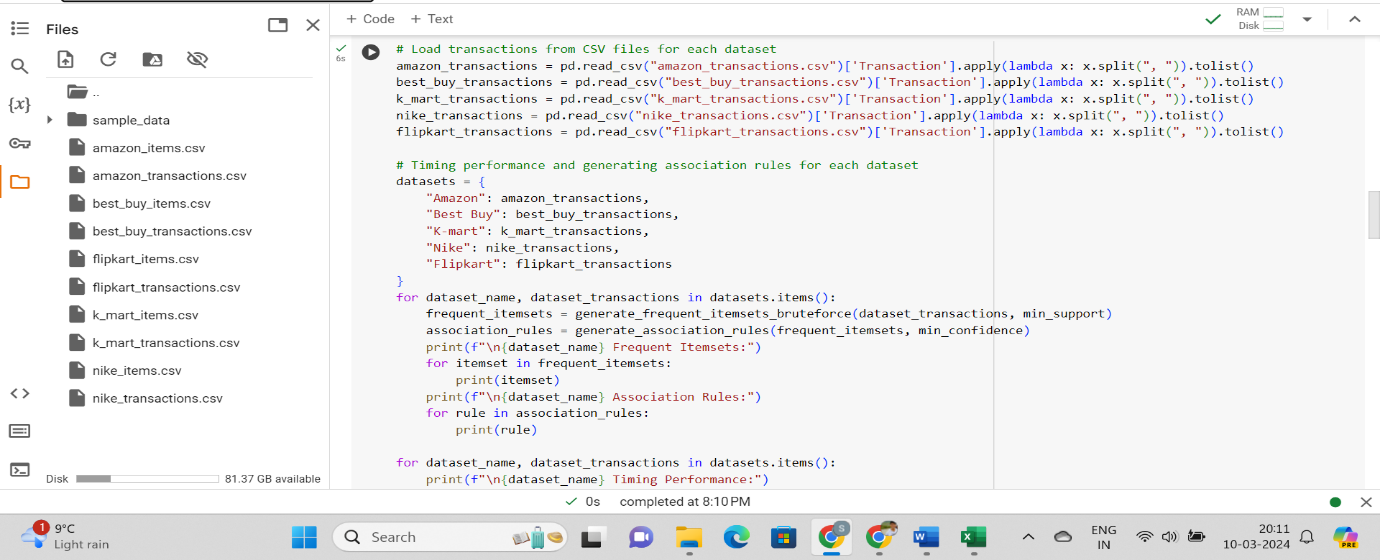


PART – 3 CODE

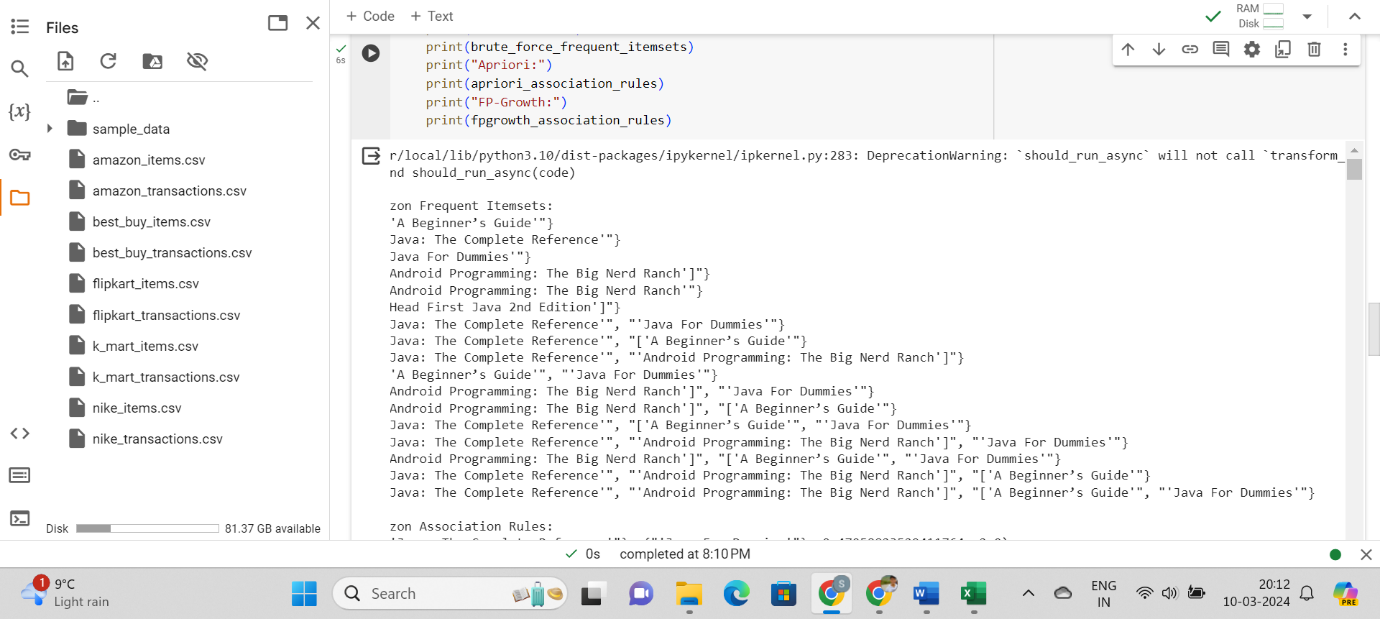


PART – 3 OUTPUT





PART – 4 CODE



PART – 4 OUTPUT

**Other:-**

The source code (.py files), the (.pynb files) and data sets (.csv files) will be attached to the zip file.

Link to Git Repository:

<https://github.com/SrikarKattanjit/katta_srikarchoudary_midtermproj.git>