**MARKET BASKET INSIGHTS**

**PHASE 5: PROJECT DOCUMENTATION & SUBMISSION**

**TABLE OF CONTENTS:**

|  |  |
| --- | --- |
| 1 | INTRODUCTION |
| 2 | PROBLEM STATEMENT |
| 3 | PROBLEM DEFINITION |
| 4 | DESIGN THINKING  (I)DATA COLLECTION  (II)DATA PREPROCESSING  (III)FEATURE SELECTION  (IV)MODEL SELECTION  (V)EVALUATION  (VI)ITERATIVE IMPROVEMENT |
| 5 | ANALYZING THE PROGRAM |
| 6 | IMPORTING NECESSARY LIBRARIES |
| 7 | LOAD THE DATASET |
| 8 | DATA CLEANING |
| 9 | DATA ANALYSIS  (I)GENERATING FREQUENT ITEMSETS  (II)SUPPORT COUNTING  (III)GENERATING ASSOCIATION RULE |
| 10 | DATA VISUALIZATION  (I)DATA PREPARATION  (II)SELECTING VISUALIZATION TOOLS AND TYPES  (III)DATA ENCODING  (IV)CREATE VISUALIZATION |
| 11 | INNOVATIVE TECHNIQUES  (I)GRAPH-BASED METHODS  (II)GEOSPACIAL ANALYSIS  (III)CUSTOMER SEGMENTATION |
| 12 | APROPRI ALGORITHM  (I)PROS  (II)CONS |
| 13 | MODELING BASED ON APRIORI ALGORITHM  (I)COMPUTING SUPPORT AND CONFIDENCE  (II)COMPUTING FREQUENT ITEMS USING APRIORI ALGORITHM  (III)ASSOCIATION RULES ARE USED FOR FREQUENT ITEMS  (IV)COMPUTING ZHANG’S RULE  (V)TRANSFORMING ANTECEDENT, CONSEQUENT AND SUPPORT  (VI)GENERATING HEAT-MAP  (VII)GENERATING SCATTER-PLOT  (VIII)GENERATING PARALLEL COORDINATES PLOT |
| 14 | OUTPUT |
| 15 | CONCLUSION |

**INTRODUCTION:**

* Market Basket Analysis is a powerful data mining technique that enables businesses to gain valuable insights from customer transaction data.
* It is widely used in retail, e-commerce, and various other industries to understand customer purchasing patterns and optimize marketing and sales strategies.
* One of the most popular algorithms for conducting Market Basket Analysis is the Apriori algorithm.

**PROBLEM STATEMENT:**

**Unveiling the customer behavior through associationanalysis:** Utilize the Market Basket Analysis on the provided dataset to uncover hidden patterns and associations between products, aiming to understand customer purchasing behavior and identify potential cross selling opportunities for the retail business.

**PROBLEM DEFINITION:**

The problem is to perform market basket analysis on a provided dataset to unveil hidden patterns and association between products. The goal is to understand customers purchasing behavior and identify potential cross-selling opportunities for a retail business. This project involves using association analysis techniques, such as Apropri algorithm, to find frequently co-occurring products and generate insights for business optimization.

**DESIGN THINKING:**

1. **DATA COLLECTION:**

1. Choose a relevant data source such as POS systems or e-commerce platforms.

2. Collect transaction data with essential fields (Transaction ID, Date/Time, Customer ID, Items Purchased), validate, clean, and format it.

3. Ensure data security, compliance, and regular updates to facilitate market basket analysis for uncovering insights.

**DATASET LINK:**

[**https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis**](https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis)

1. **DATA PREPROCESSING:**

1. Clean and remove duplicates, correct errors, and address missing values in transaction data.

2. Encode data into a binary format, where 1 represents item purchase, 0 represents non-purchase.

3. Optionally, aggregate data, filter using support and confidence thresholds, and ensure privacy compliance before analysis in market basket analysis.

**III. FEATURE SELECTION:**

1. Data Reduction: Feature selection in market basket analysis aims to reduce the number of items or attributes under consideration. By focusing on a subset of relevant features, it simplifies the analysis and makes it computationally more efficient.

2. Association Rule Mining: Market basket analysis typically employs association rule mining techniques to discover patterns and relationships between items purchased together. Feature selection helps identify the most important items for generating meaningful rules.

3. Key Metrics: Common metrics like support, confidence, lift, and information gain are used to evaluate the importance of items. Feature selection involves ranking or filtering items based on these metrics to retain the most significant ones.

4. Model Efficiency: Selecting the right features improves the efficiency of the market basket analysis model. It reduces processing time and memory usage, making the analysis more practical for large retail datasets.

5. Actionable Insights: Effective feature selection leads to more focused and actionable insights for retailers. By concentrating on the most influential items, businesses can optimize product recommendations, design effective promotions, and enhance inventory management strategies to boost sales and customer satisfaction.

**IV.MODEL SELECTION:**

Model selection in market basket analysis is a critical step to choose the appropriate technique or algorithm to extract meaningful insights from transaction data.

Here are key considerations for model selection in market basket analysis:

1. Apriori Algorithm

2. FP-Growth Algorithm

3. Association Rule Mining

4. Recurrent Neural Networks (RNNs)

5. Long Short-Term Memory Networks (LSTMs)

6. Hybrid Models

**V. EVALUATION:**

Evaluation in market basket analysis assesses the quality of association rules. Metrics like support, confidence, and lift measure rule significance. High support signifies common rules, while confidence reflects reliability. Lift gauges the strength of associations, while leverage and conviction assess rule dependency. Effective evaluation guides retailers in optimizing product recommendations and marketing strategies.

**VI. ITERATIVE IMPROVEMENT:**

Iterative improvement is a process of gradually refining a solution or approach through a series of repetitive cycles.

In various contexts, such as data analysis or software development, it involves making incremental changes and adjustments based on feedback and outcomes from previous iterations.

This iterative approach allows for continuous enhancement and optimization, often resulting in more effective and efficient results.

**ANALYZING THE PROGRAM:**

Analyzing a program involves examining its code, structure, and behavior to understand its functionality and performance. This process aims to identify issues, improve efficiency, and ensure the program meets its intended objectives.

**Key aspects of program analysis include:**

* Code Review
* Debugging
* Profiling
* Testing
* Security Analysis
* Documentation
* Scalability Analysis
* Maintenance Planning

**IMPORTING NECESSARY LIBRARIES:**

* **Numerical computing -** import numpy as np
* **Data manipulation** -import pandas as pd
* **Data visualization-**import matplotlib.pyplot as plt
* **Machine learning -**import tensorflow as tf

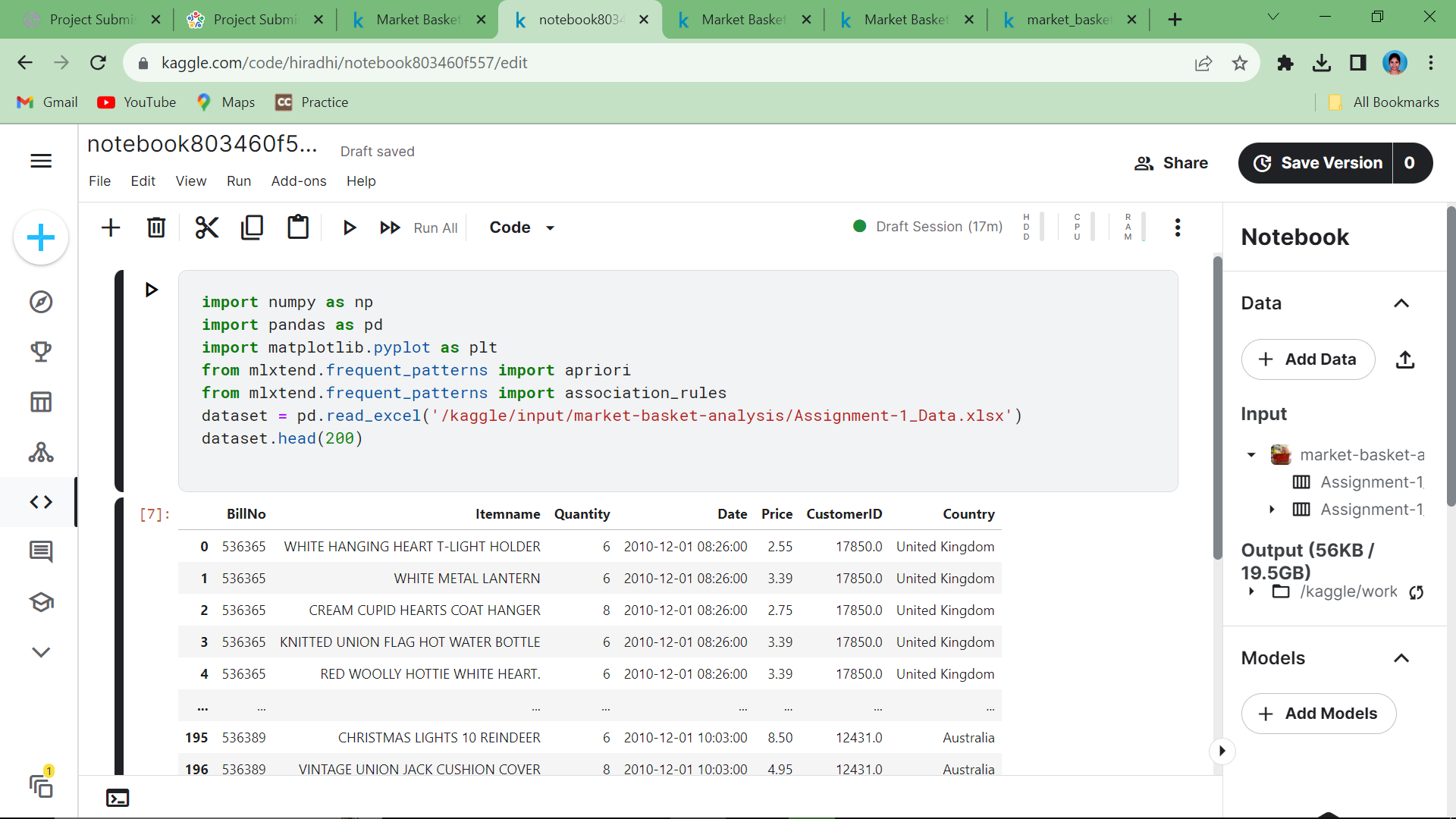
**LOAD THE DATASET:**

Use Pandas or other relevant libraries to load the data into your Python environment. Here's how to load a CSV file, for instance: **data = pd.read\_csv('your\_data.csv')**

**EXPLORE THE DATA:**

After importing the data, you can explore it. Here are some common exploration tasks:

**Check the first few rows of the dataset: data.head()**

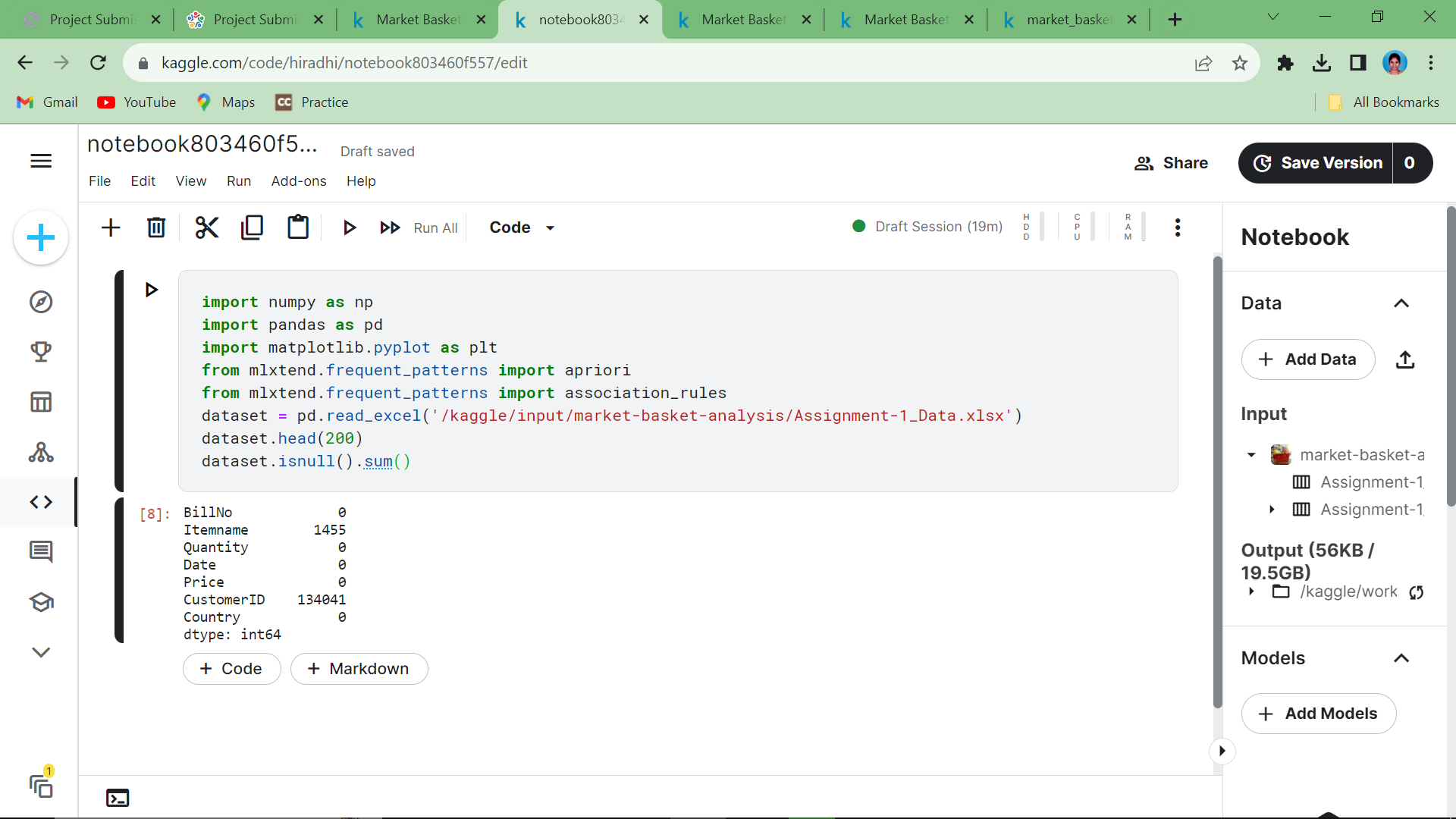
****

**DATA CLEANING:**

Data cleaning is a time-consuming but necessary process to prepare data for meaningful analysis and modeling. The goal is to ensure that the data accurately reflects the real-world phenomena it represents and that it is free from errors and inconsistencies that could lead to misleading results.

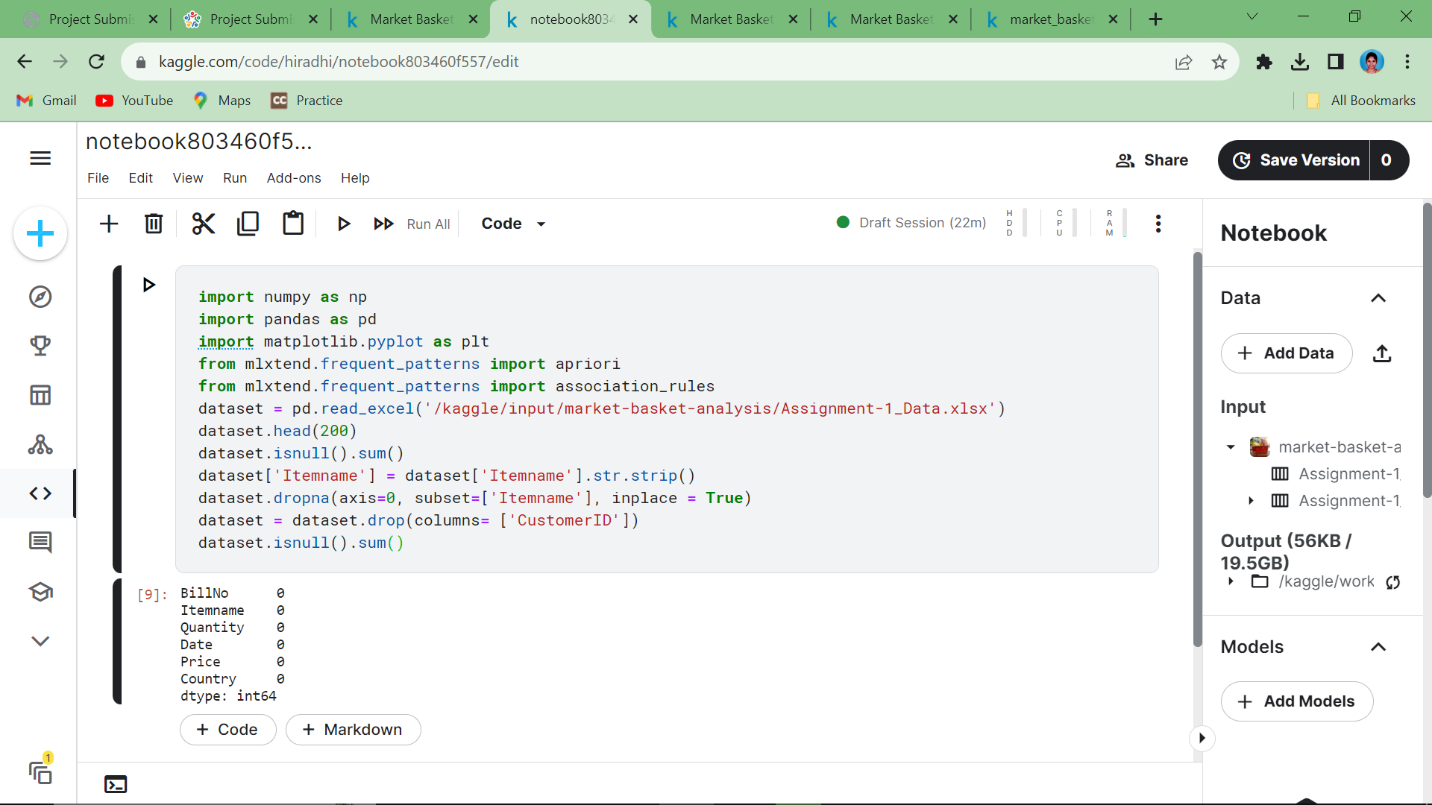
**Data cleaning is achieved by,**

* Removing duplicates
* Adding missing elements



**Code Explanation:**

**dataset.isnull().sum()** is used to check the missing data.



**Code Explanation:**

**dataset.dropna(axis=0,subset=[‘Itemname’],(inplace=True )** is used to drop rows that has missing values.

**DATA ANALYSIS:**

Data analysis is the process of inspecting, cleaning, transforming, and interpreting data to discover meaningful insights, patterns, and trends.

1. **GENERATING FREQUENT ITEMSETS:**

It starts by finding frequent individual items (1-itemsets) by scanning the dataset and counting their support. Then, it iteratively generates frequent itemsets of increasing size (2-itemsets, 3-itemsets, etc.) by joining pairs of frequent (k-1)-itemsets.

1. **SUPPORT COUNTING:**

It is used to determine the frequency of itemsets (combinations of items) within a dataset, which helps identify how often specific item combinations appear in a collection of transactions. Support count is an important metric because it allows to filter out infrequent or less significant itemsets and focus on those that are more prevalent.

**Support counting involves the following steps:**

1. Counting Occurrences.
2. Calculating Support.

|  |
| --- |
| **Support(X)= Total number of transactions/Transactions containing itemset X** |

1. Threshold Setting.
2. **GENERATING ASSOCIATION RULE:**

Association rules are generated from frequent itemsets by examining the confidence of rule X => Y, where X and Y are itemsets. The confidence of a rule is calculated as Support(X ∪ Y) / Support(X).

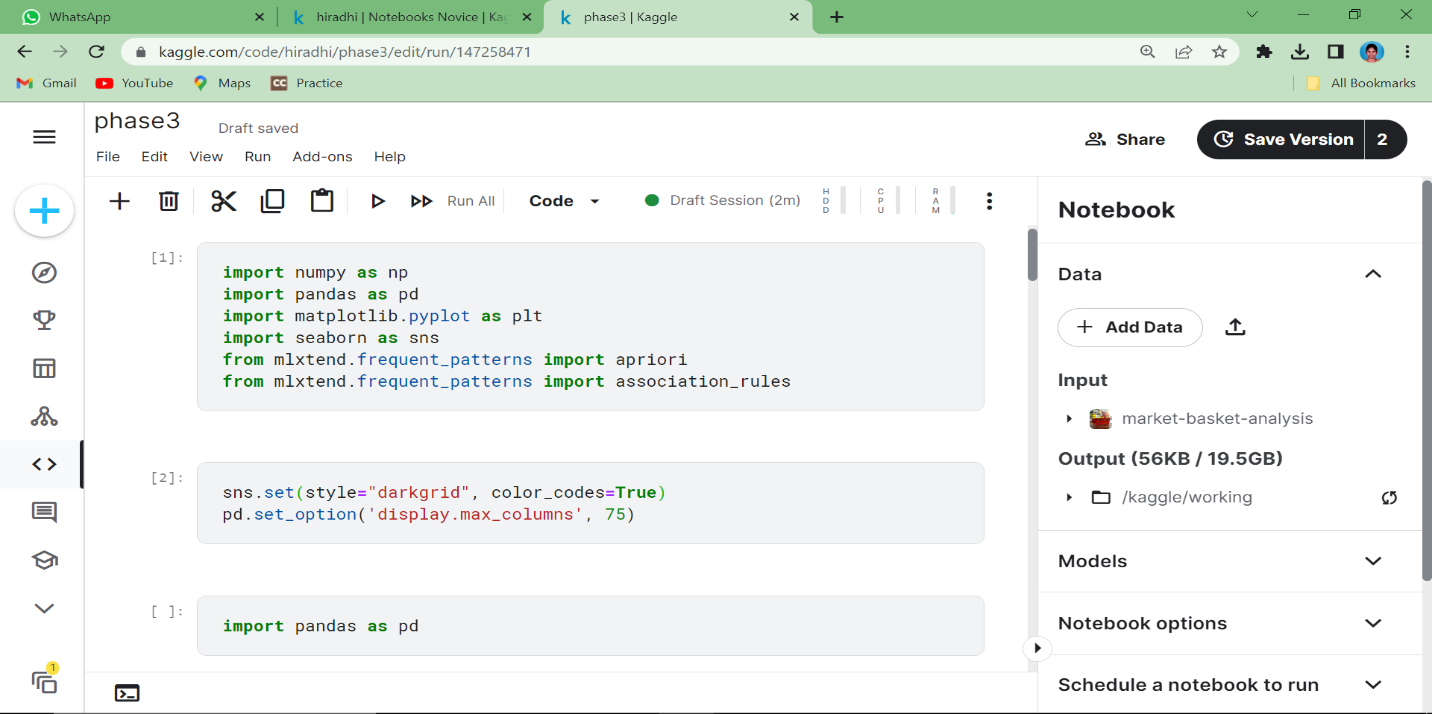
* Frequent Itemset Mining
  + Support Calculation
  + Confidence Calculation
  + Rule Generation
  + Pruning and Filtering
  + Presentation and Interpretation

**DATA VISUALIZATION:**

Data visualization plays a crucial role in market basket analysis, which is a technique used to discover patterns and relationships among items that customers purchase together. Effective data visualization can help us to understand customer behavior, identify association rules, and make data-driven decisions.

1. **DATA PREPARATION:**

Data preparation is a foundational and essential step in the data analysis process. It involves the meticulous cleaning, transformation, and organization of raw data to ensure its accuracy, reliability, and suitability for analysis. At the outset, data is collected from various sources, be it databases, spreadsheets, external datasets, or APIs, and it is crucial to verify that the data is complete and free of errors.

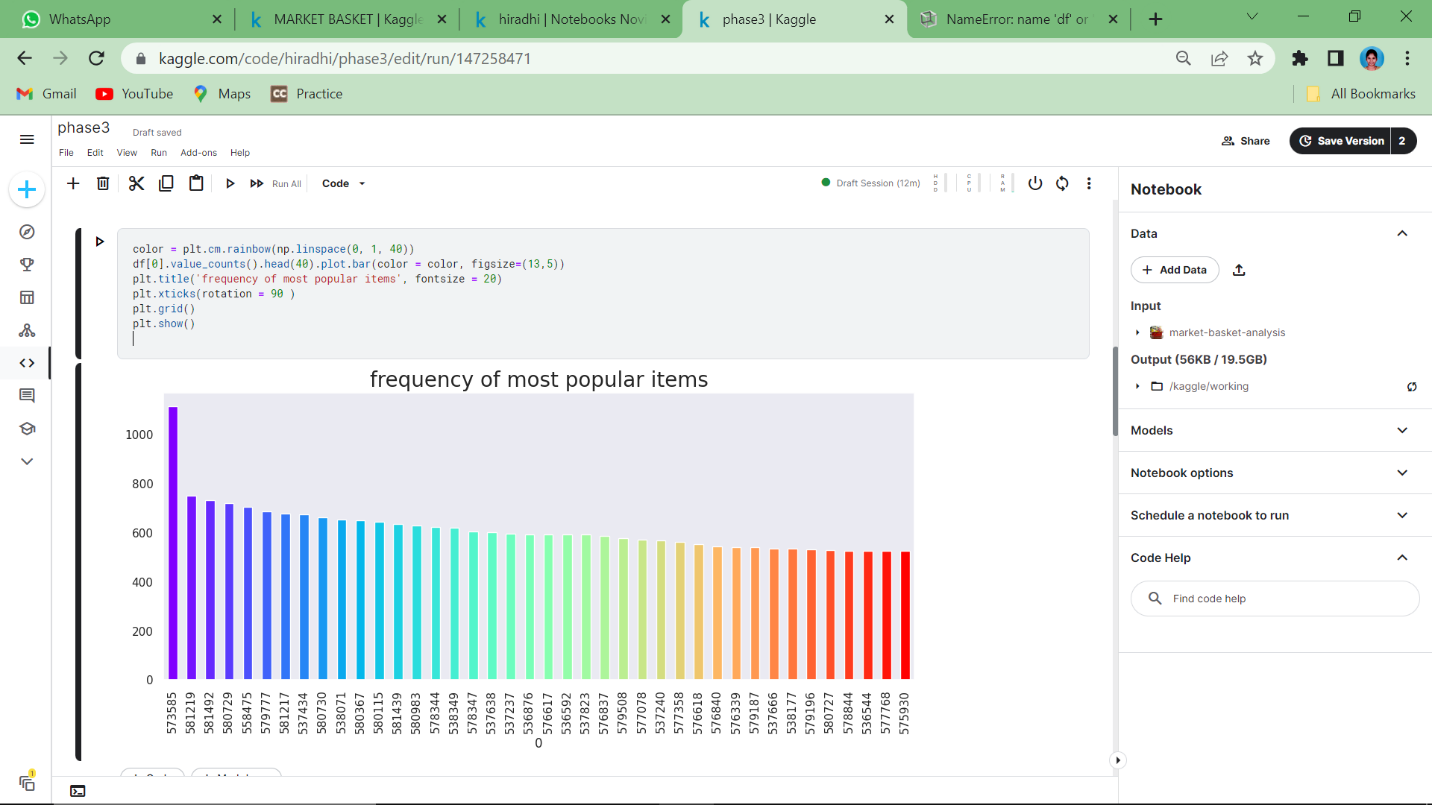


1. **SELECTING VISUALIZATION TOOLS AND TYPES:**

Choose visualization tools that match the data's nature, analysis objectives, and audience familiarity. Opt for interactive platforms like Tableau for complex data and select appropriate visualization types, such as bar charts for categories or line charts for time-series data. Consider data size and the publication medium to effectively convey insights.

**1.Select the Items and Frequencies:**Choosing the frequent items to visualize.

**2.Create a Bar Chart:**Use data visualization software or tools such as Excel, Python with libraries like Matplotlib or Seaborn, R, or specialized data visualization tools like Tableau.

**3.Label and Format the Chart:**Adding the appropriate labels to the x and y-axes, a title to the chart, and any other relevant information 

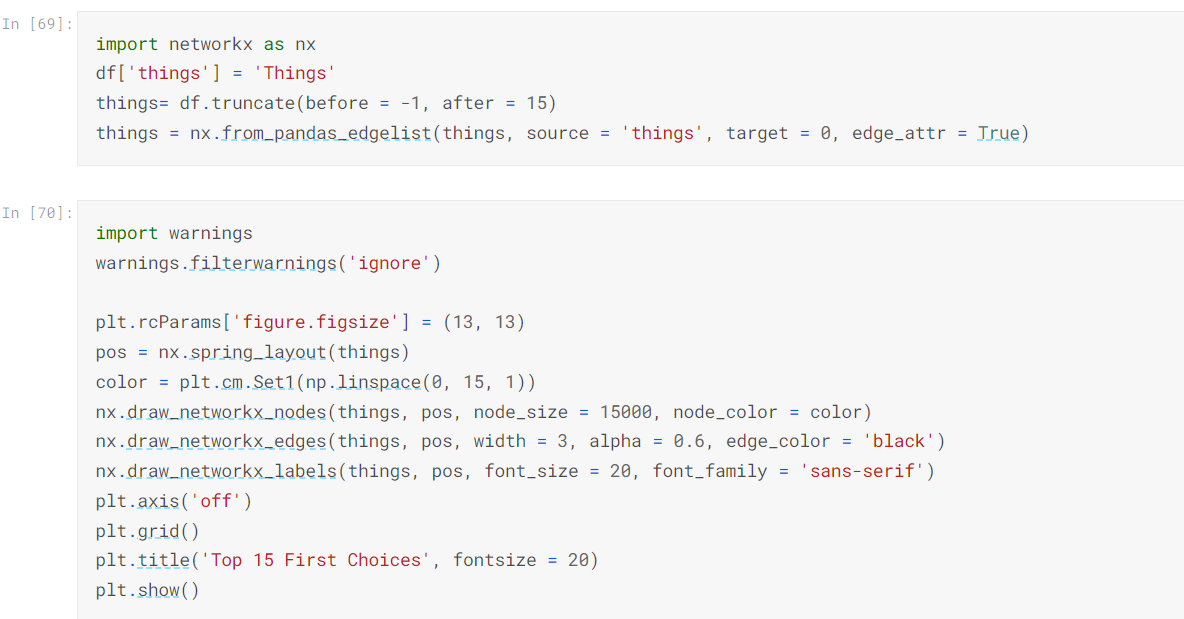
**DATA ENCODING:**

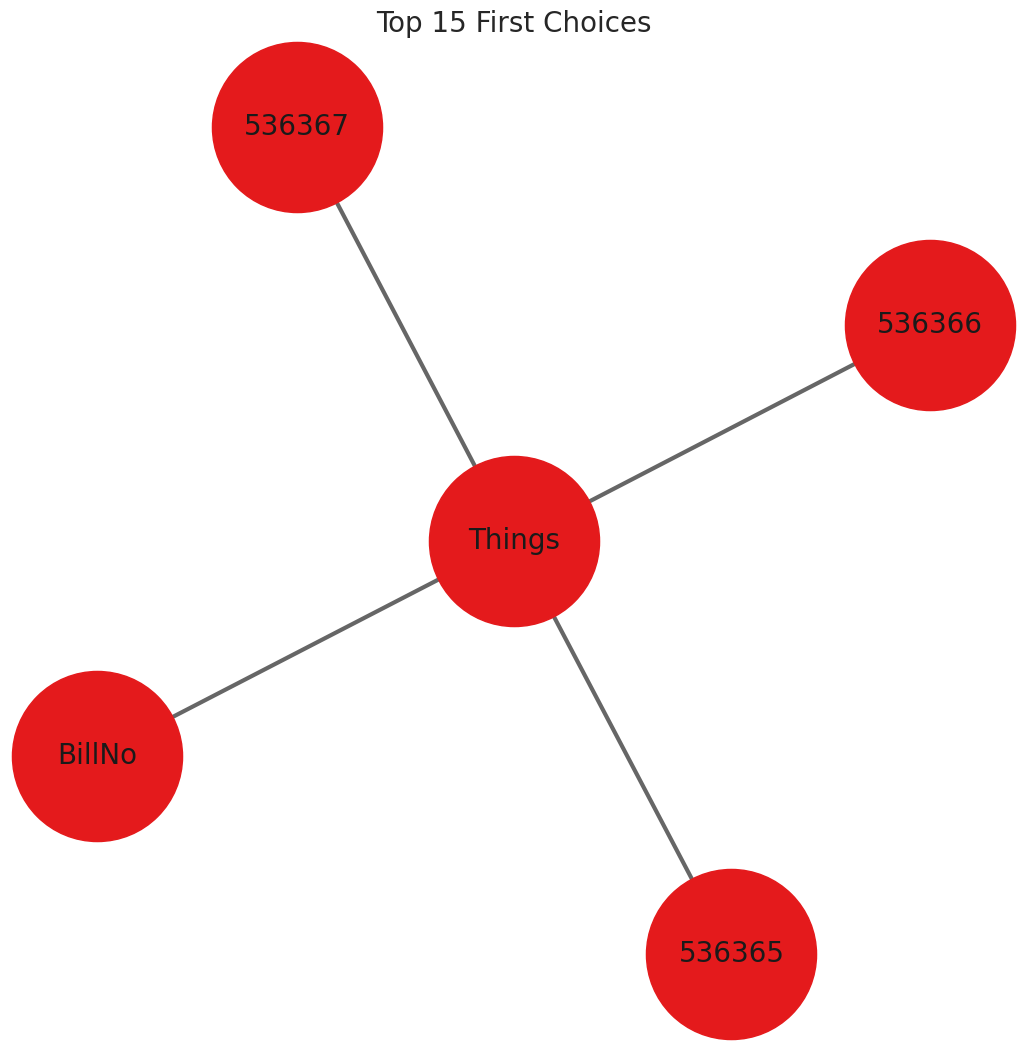
Data encoding is often used to convert non-numeric data into a numerical format that can be processed by machine learning algorithms. Common encoding techniques include one-hot encoding, label encoding, and binary encoding, each with its own advantages and use cases. Data encoding is essential to make diverse types of data compatible with analytical and modeling tools, facilitating meaningful insights and predictions.

**CREATE VISUALIZATION:**

Use visualization tools or libraries such as Matplotlib, Seaborn, Tableau, or Power BI to generate the graphical representation of the data.Discovering frequent itemsets is a crucial step in identifying patterns and associations between products in a retail or e-commerce dataset.

The given below diagram shows the top priority choices in the dataset.





**INNOVATIVE TECHNIQUES:**

1. **GRAPH-BASED METHODS:**

* In market basket analysis, graph-based methods can be used to represent and analyze relationships between products or items in transactions.
* Techniques such as association rule mining and graph algorithms can then be applied to extract meaningful patterns and associations between products, aiding in product recommendations, inventory management, and marketing strategies for retailers.

1. **GEOSPACIAL ANALYSIS:**

**Geospatial analysis in market basket analysis:**

1. Incorporates location data: Adds a spatial dimension to transaction records.

2. Reveals regional insights: Helps identify location-specific purchasing trends and preferences.

3. Informs inventory and marketing strategies: Enables retailers to optimize stock levels and tailor promotions based on geographic patterns.

4. Enhances customer experiences: Delivers personalized shopping recommendations and strategies to specific regions.

1. **CUSTOMER SEGMENTATION:**

**Customer segmentation in market basket analysis:**

1. Utilizes transaction data to categorize customers into distinct groups.

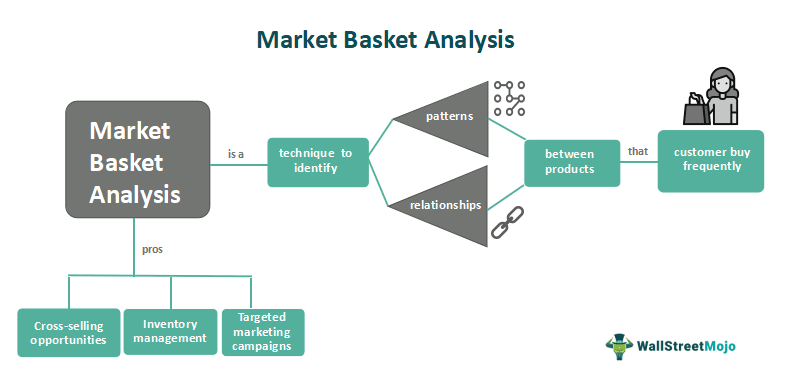
2. Enables tailored marketing strategies, promotions, and recommendations for each segment.

3. Optimizes inventory management based on the preferences of customer segments.

**APRIORI ALGORITHM:**

The Apriori algorithm is a data mining method applied in market basket analysis to find patterns in transaction data. It identifies associations and relationships between items frequently purchased together by customers.

By setting support and confidence thresholds, it generates rules that inform decisions on product recommendations, inventory optimization, and marketing strategies, enhancing sales and customer experiences.

****

1. **PRO’S OF THE APRIORI ALGORITHM:**
   1. Pattern Discovery
   2. Customizable Thresholds
   3. Enhanced Decision-Making
   4. Improved Sales
   5. Customer Satisfaction
   6. Efficiency
   7. Widespread Use

These pros make the Apriori algorithm a valuable tool for retailers looking to harness the power of association rules to optimize their operations and enhance customer interactions.

1. **CON’S OF THE APRIORI ALGORITHM:**

1. Computational Intensity

2. Large Number of Candidate Itemsets

3. Infrequent Itemsets

4. Static Thresholds

5. Limited to Binary Data

6. Limited to Market Basket Analysis

7. Doesn't Consider Temporal or Sequential Data

While the Apriori algorithm offers valuable insights, these cons should be considered when choosing it for specific data mining tasks, particularly in scenarios where computational resources are constrained or when working with large and complex datasets.

**MODELING BASED ON APRIORI ALGORITHM:**

**I. COMPUTING SUPPORT AND CONFIDENCE:**

**SUPPORT:**

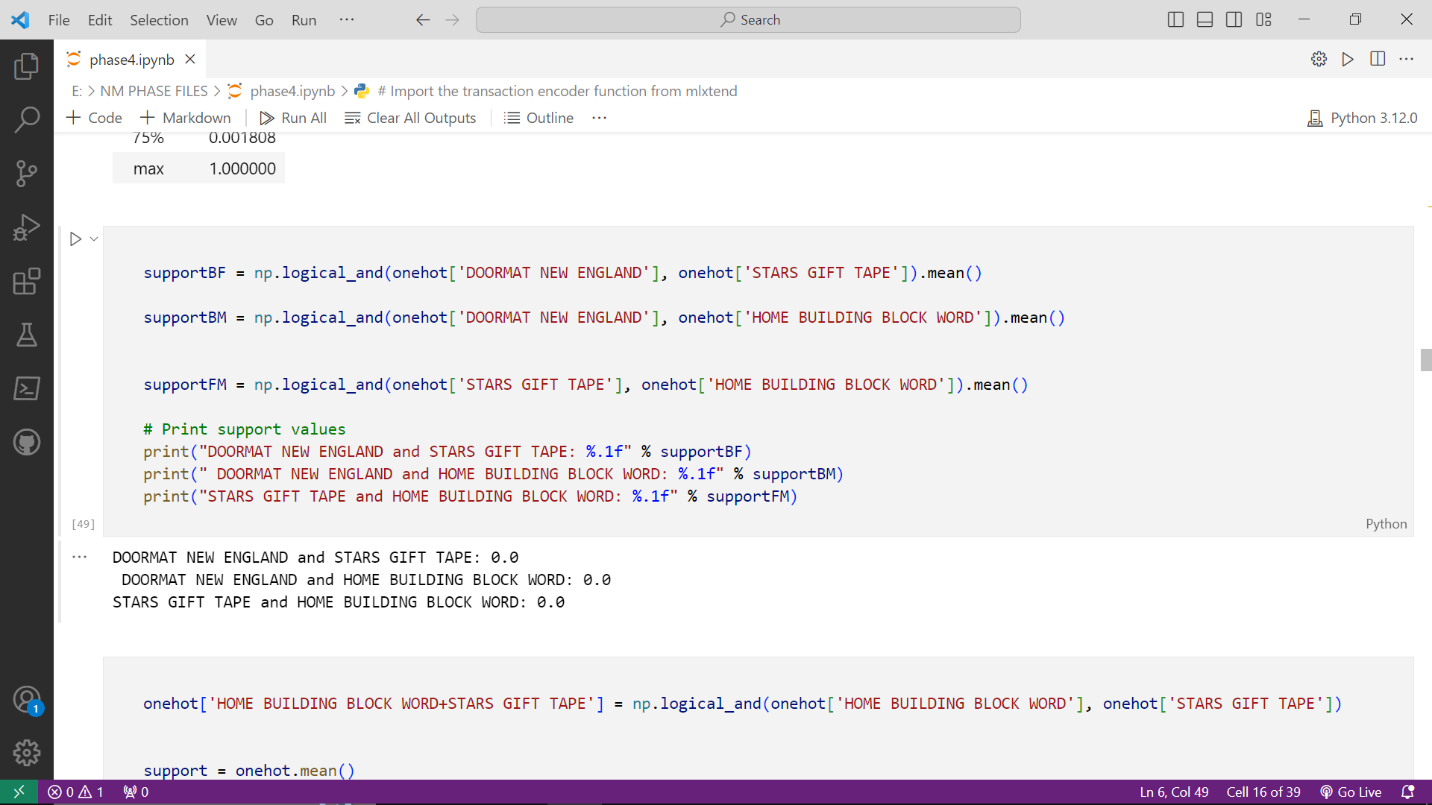
In market basket analysis, "support" is a metric used to measure the frequency or popularity of a specific item or itemset within a given dataset of transactions.

|  |
| --- |
| **Support(A)=Total number of transactions / Number of transactions containing itemset A** |

**CONFIDENCE:**

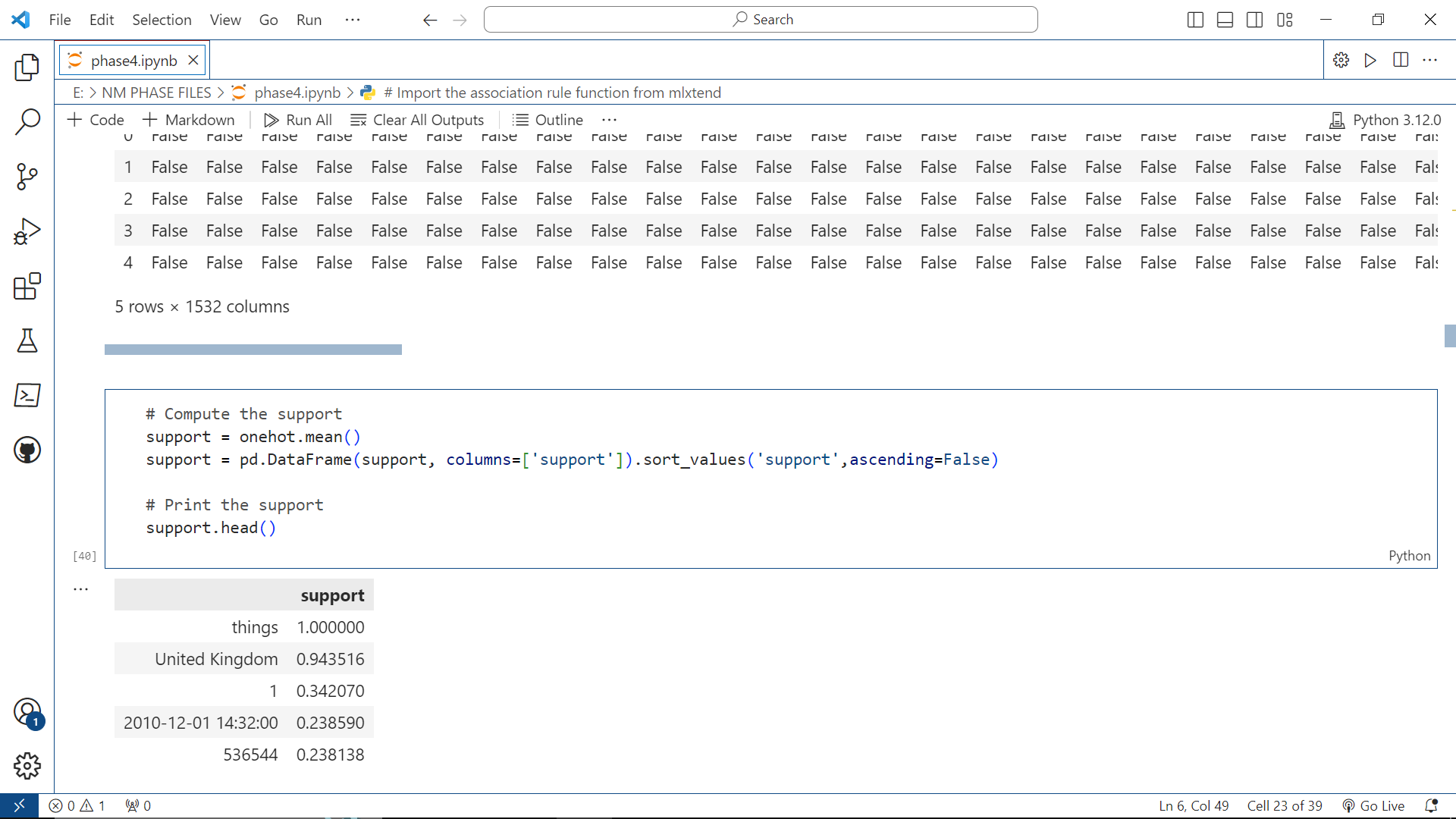
Confidence is a metric that quantifies the strength of an association between two items in a dataset of transactions.

|  |
| --- |
| **Confidence(A→B)=Support(A∪B)/Support(A)** |



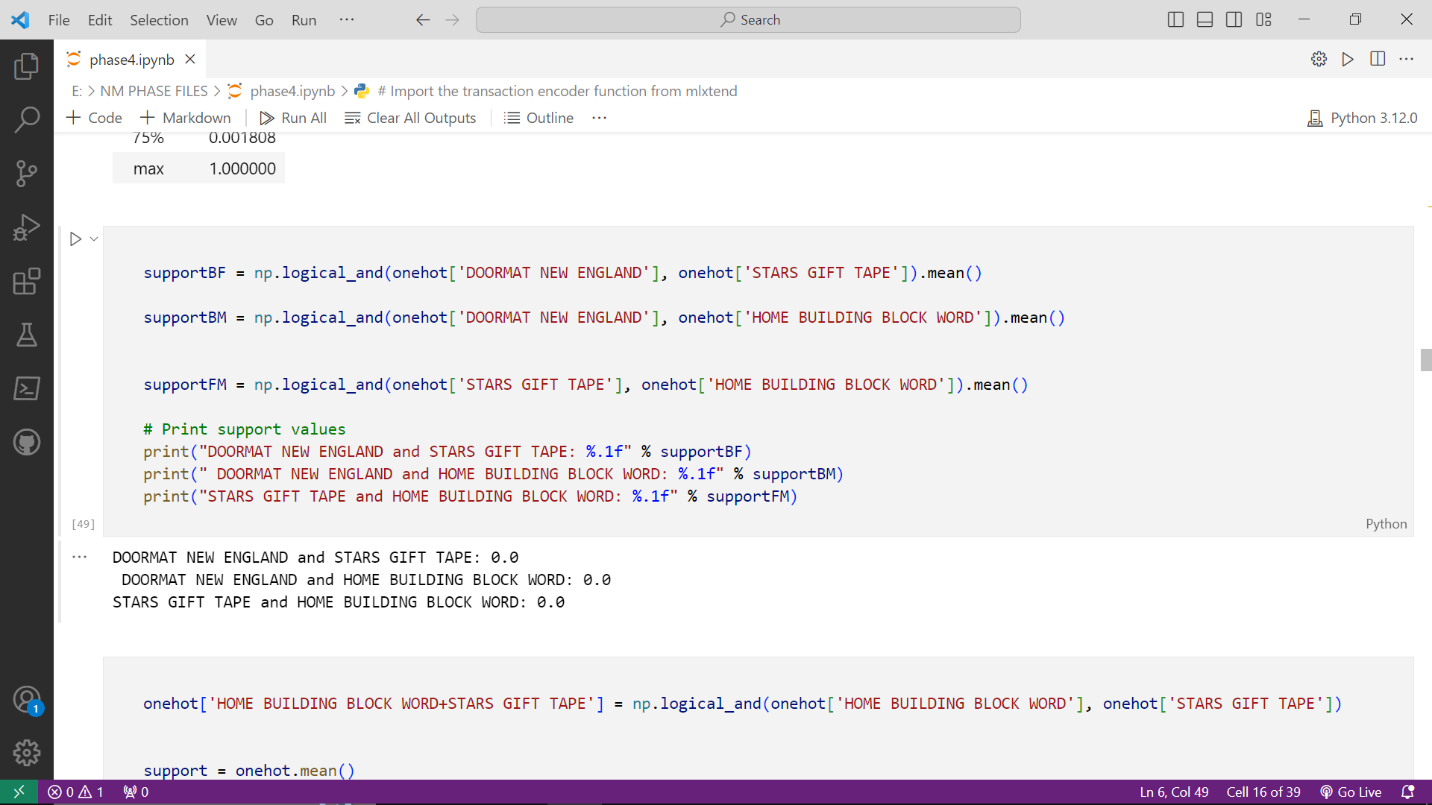
**COMPUTING FREQUENT ITEMS USING APRIORI ALGORITHM:**

* Create a List of Candidate 1-Itemsets
* Generate Candidate 2-Itemsets
* Scan Database and Calculate Support for Candidate 2-Itemsets
* Generate Candidate 3-Itemsets
* Continue the Process
* Output Frequent Item sets

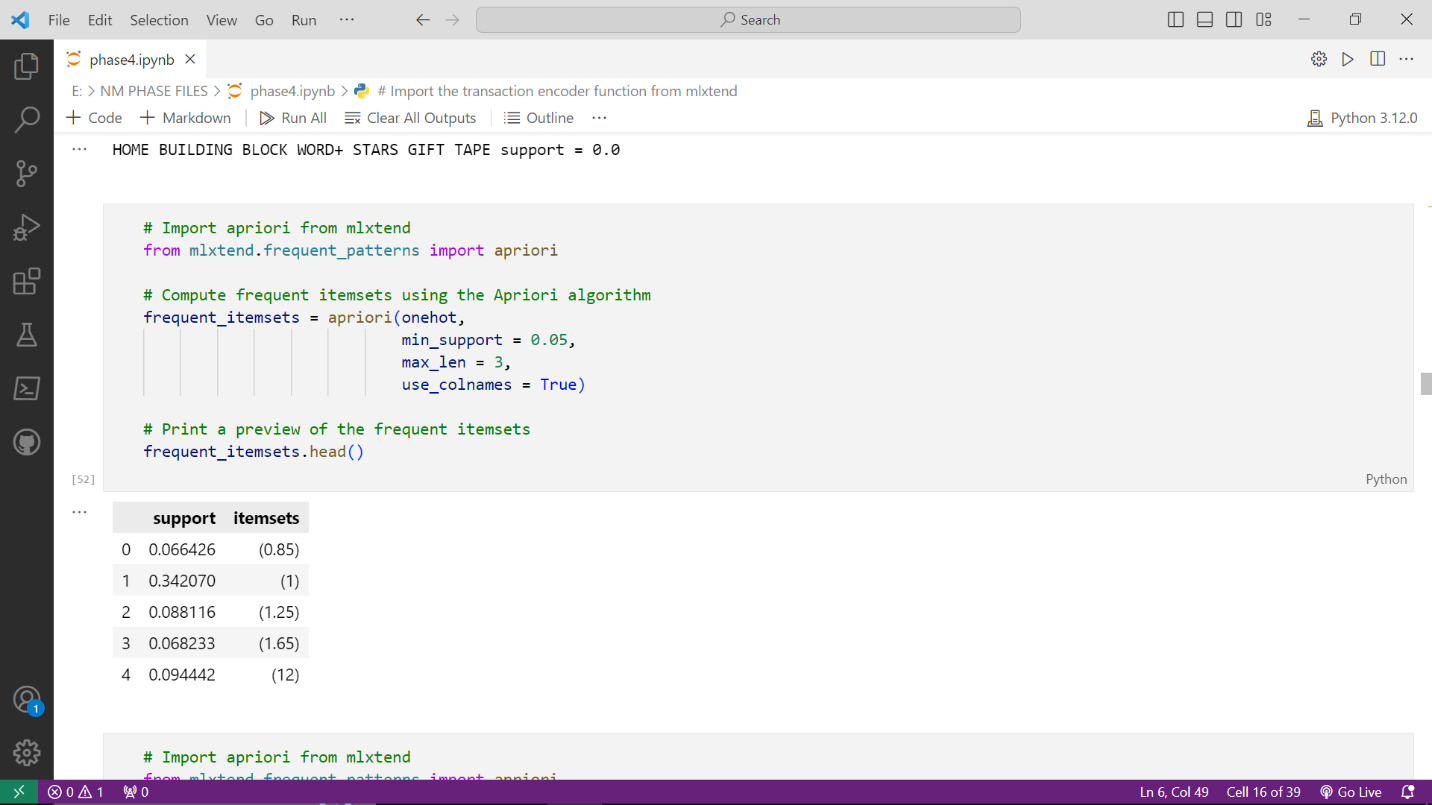


**ASSOCIATION RULES ARE USED FOR FREQUENT ITEMS:**

Association rules are used to discover interesting relationships between items in large datasets, often in the context of market basket analysis. These rules help identify associations and patterns between items that are frequently purchased or used together. Frequent item sets are an essential component of association rule mining.



**COMPUTING ZHANG’S RULE:**



**TRANSFORMING ANTECEDENT, CONSEQUENT AND SUPPORT:**

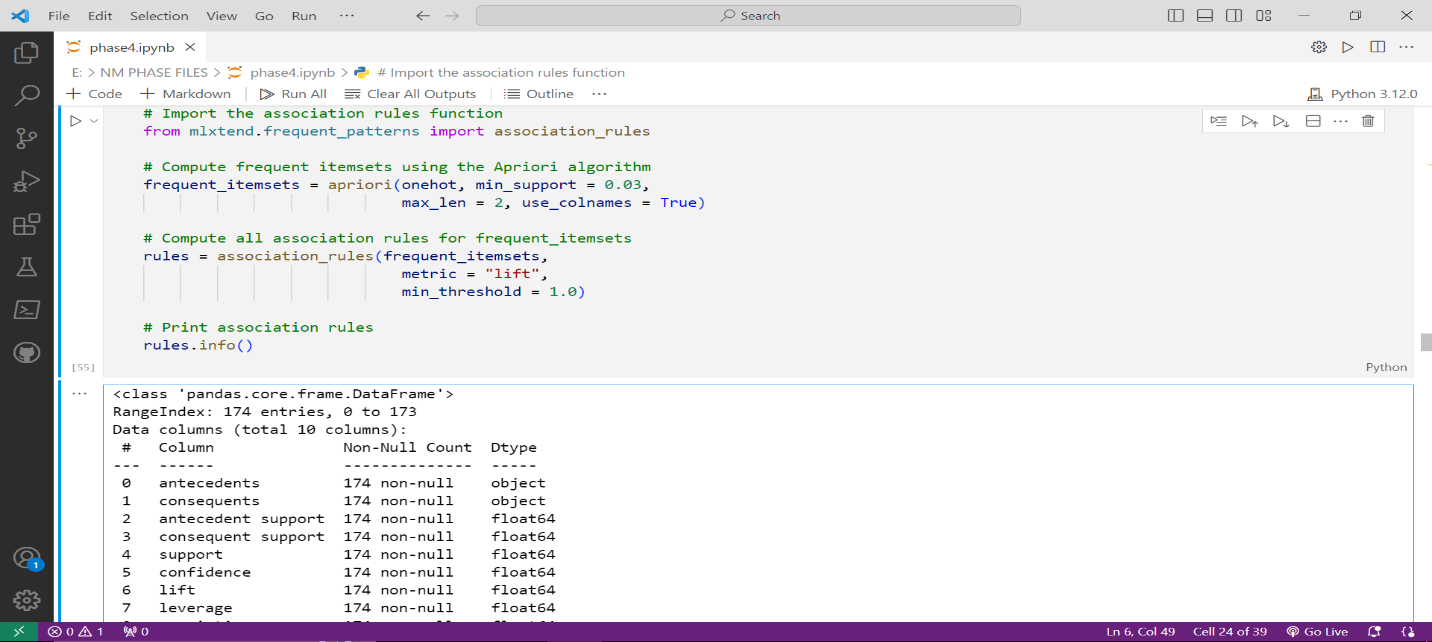
**1. Transforming the Consequent:**

**Consequent Expansion:**

For example, if have a rule {A} => {B}, can expand the consequent to {A} => {B, C, D}, where B, C, and D are related items.

**Consequent Reduction:**

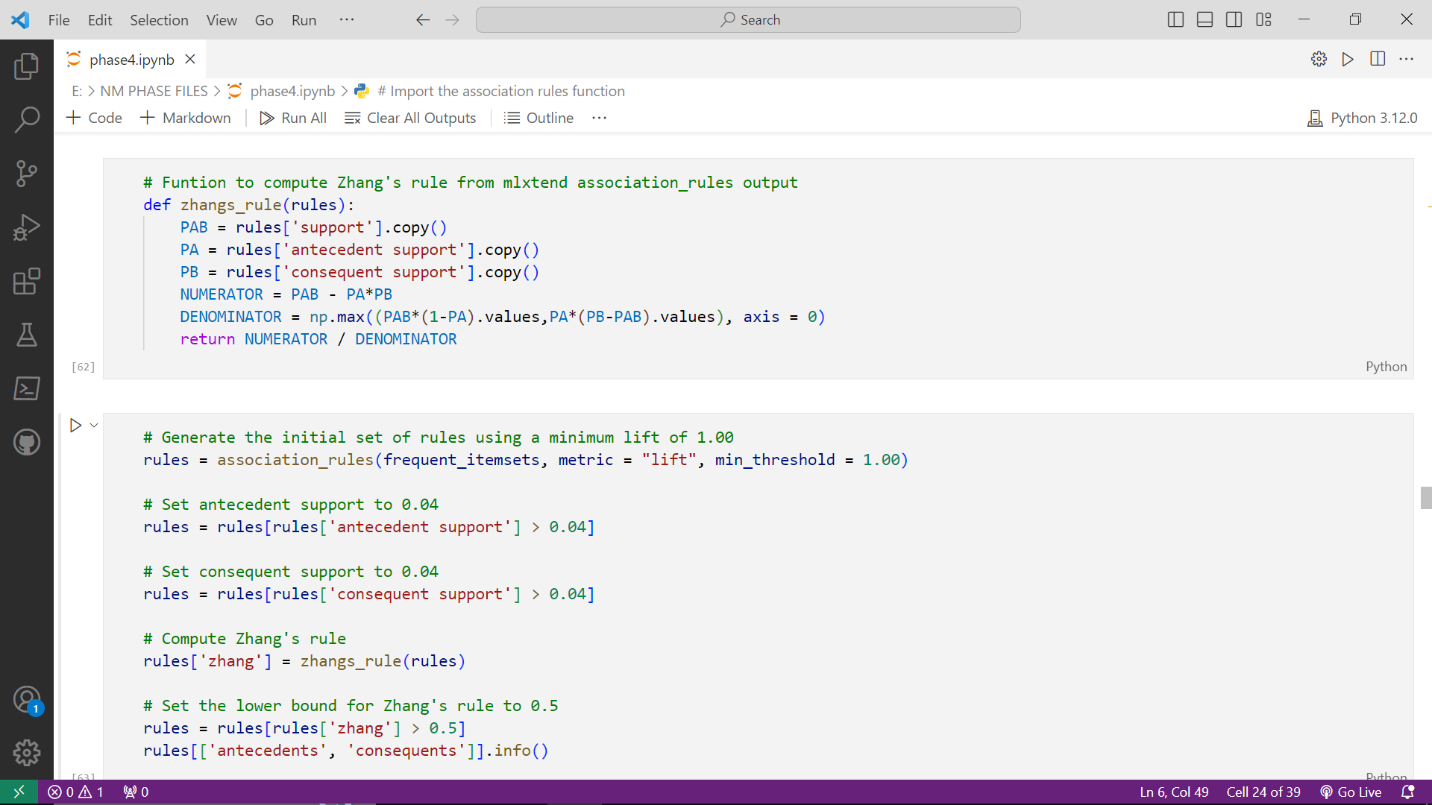
For example, if have a rule {A, B, C} => {D, E}, reduce the consequent to {A, B, C} => {D}.

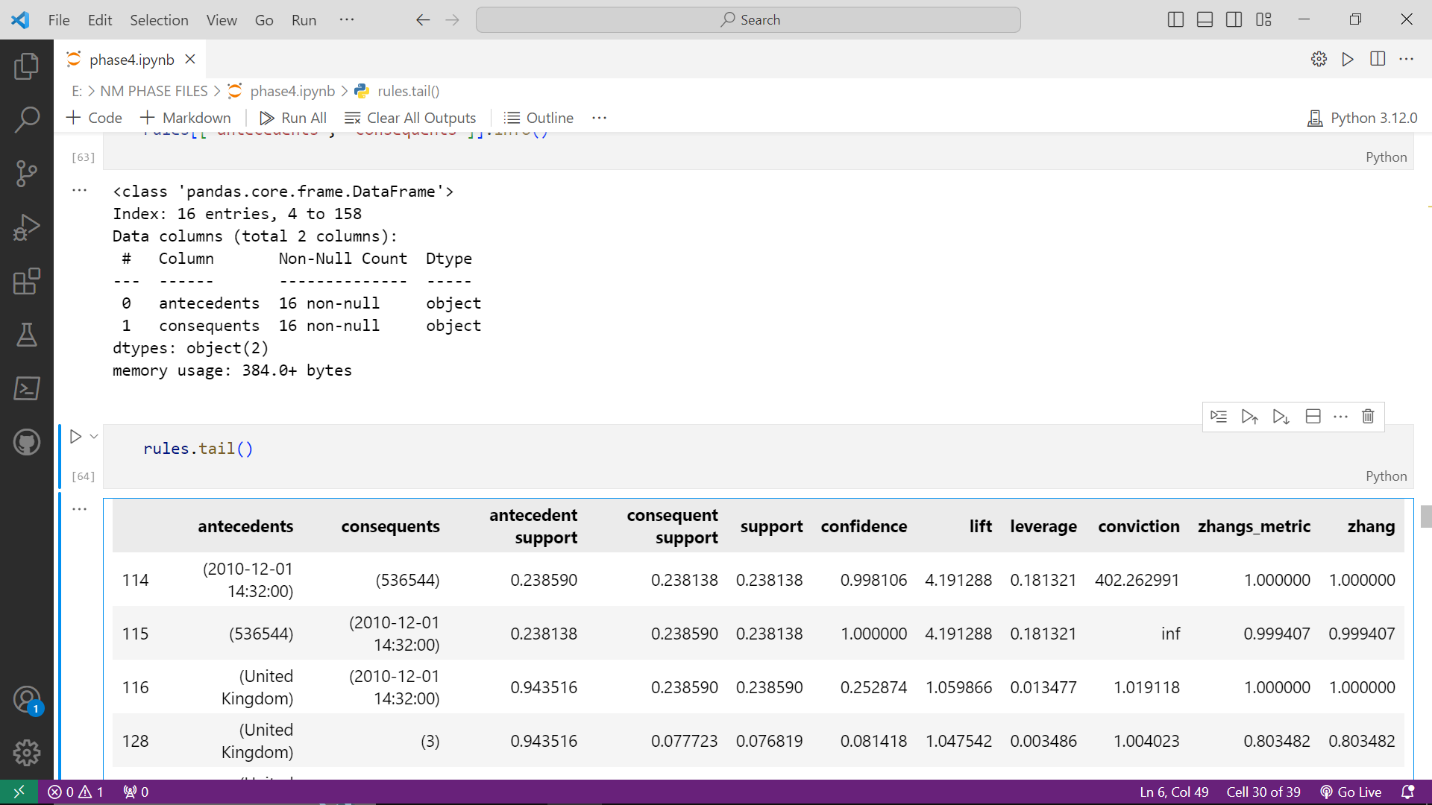


**2. Adjusting Support:**

**Lowering Minimum Support:** By lowering the minimum support threshold, you can discover associations that may not be considered frequent by a more stringent threshold.

**Raising Minimum Support:**Conversely, by increasing the minimum support threshold, you can focus on discovering only the most significant, highly frequent associations, filtering out less common patterns.





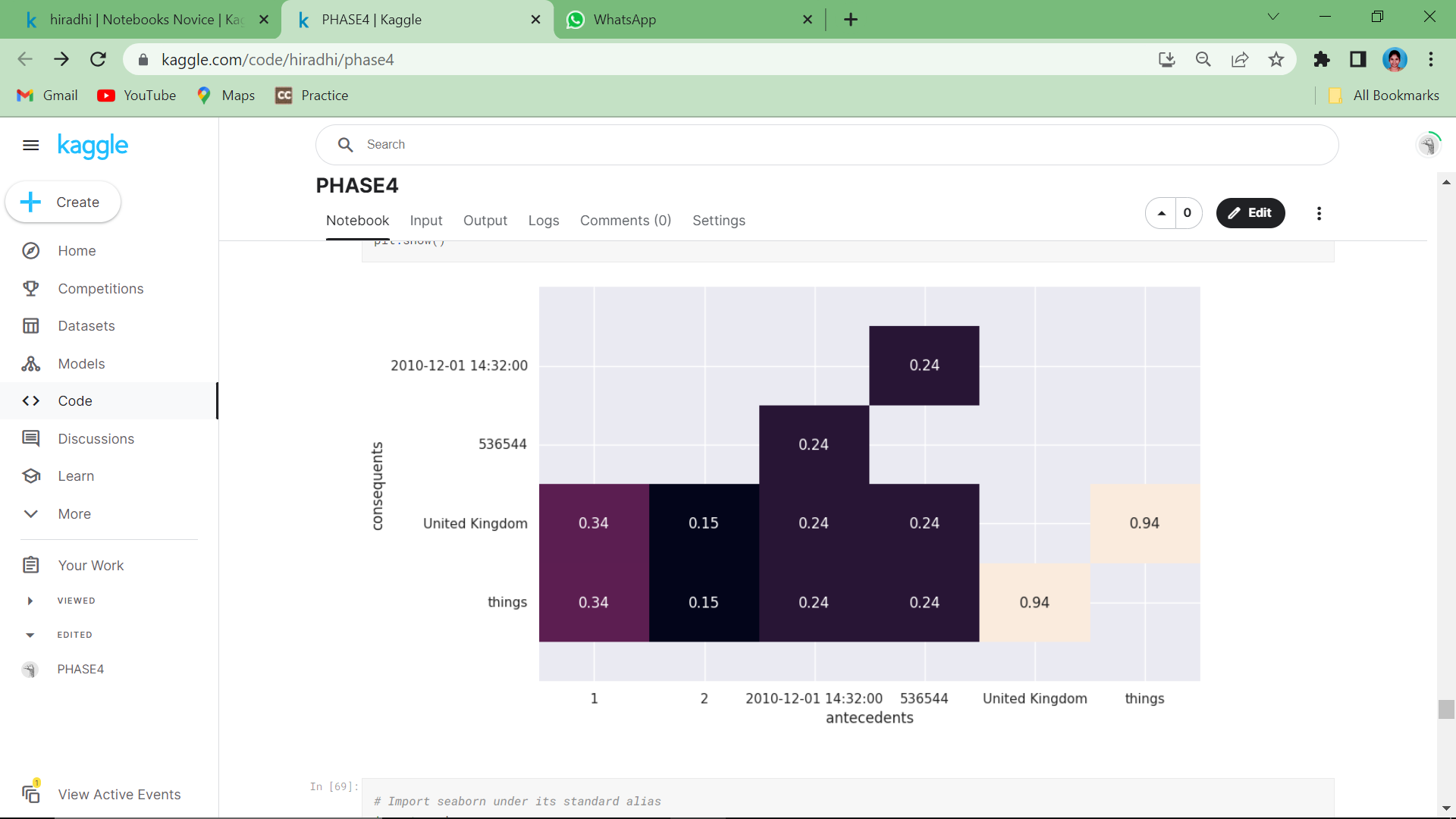
**OUTPUT:**

**I.GENERATING HEAT MAP:**

To generate a heatmap, use various tools and programming languages, but one of the most commonly used libraries for creating heatmaps is Python's `seaborn` library, which is built on top of `matplotlib`.

**Here's a step-by-step guide on how to generate a heatmap using `seaborn`:**

* Install the Necessary Libraries
* Import the Libraries
* Prepare Your Data
* Create the Heatmap
* Display the Heatmap



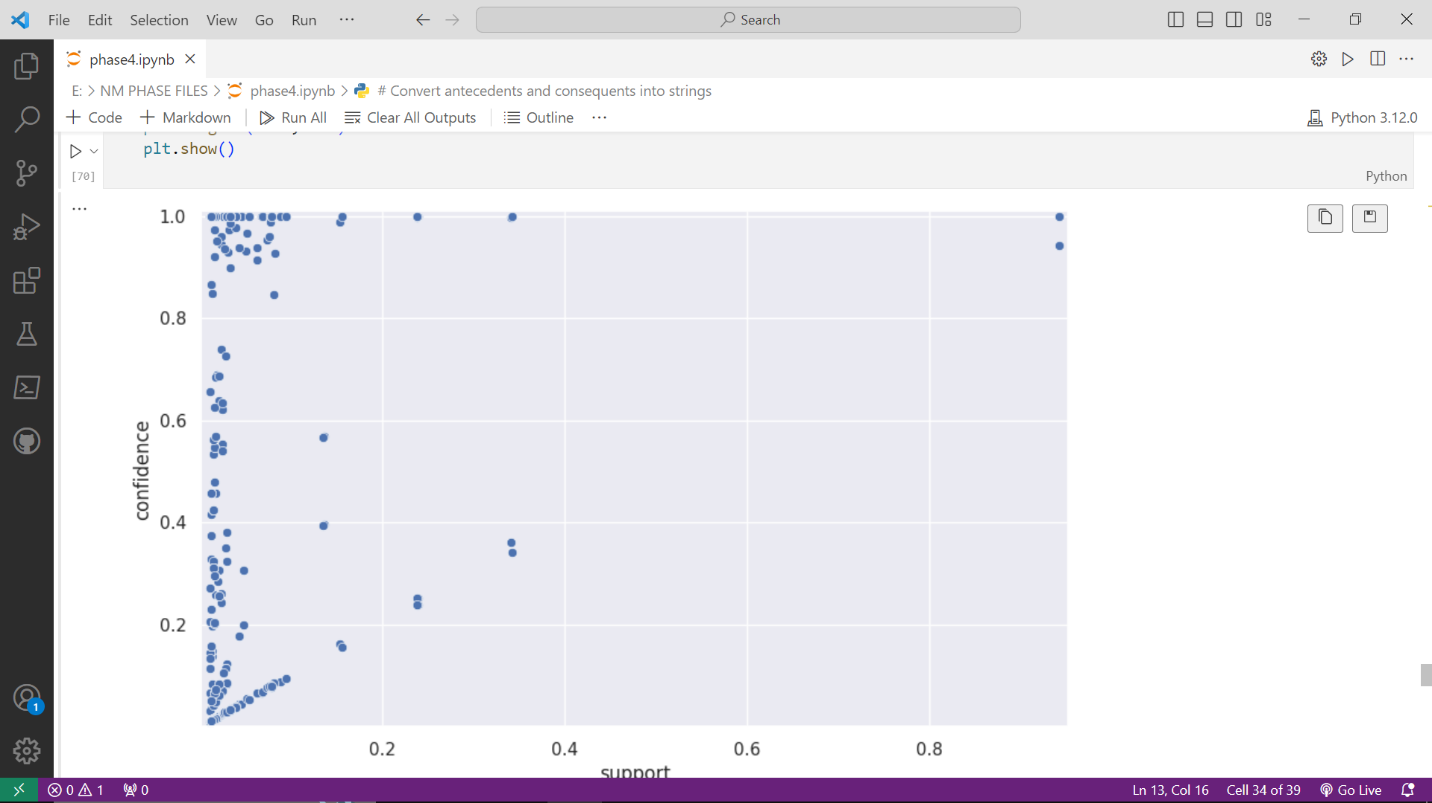
**II.GENERATING SCATTER PLOT:**

**A scatterplot displays pairs of values.**

* + Antecedent and consequent support.
  + Confidence and lift.

**Can provide starting point for pruning.**

* + Identify patterns in data and rules.



**II.GENERATING PARALLEL COORDINATES PLOT:**

The parallel coordinates plot will allow us to visualize whether a relationship exist between an antecedent and consequent.

**When to use parallel coordinate plots**

**Parallel coordinates vs. heatmap.**

* + Don't need intensity information.
  + Only want to know whether rule exists.
  + Want to reduce visual clutter.

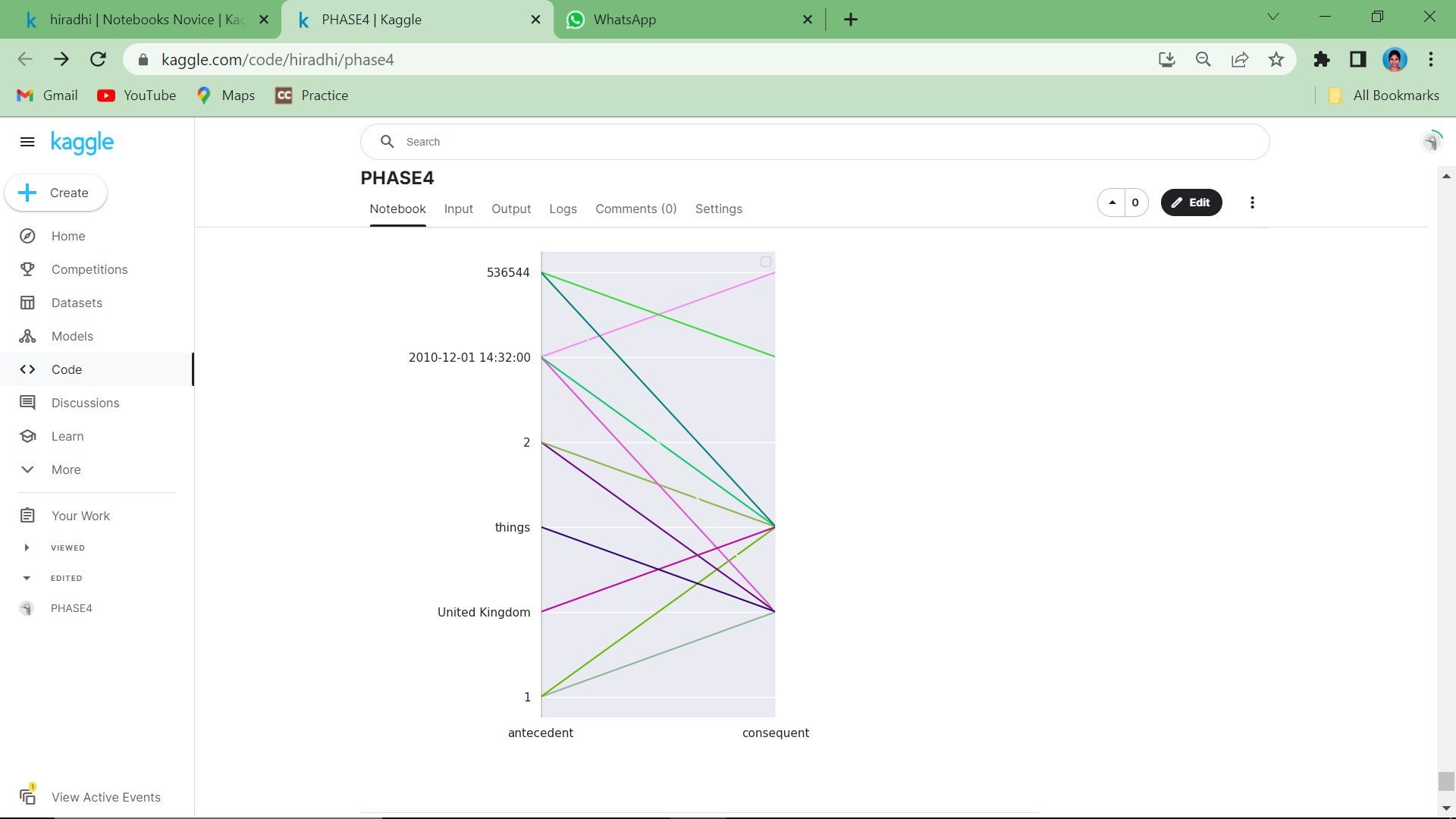
**Parallel coordinates vs. scatterplot.**

* + Want individual rule information.
  + Not interested in multiple metrics.
  + Only want to examine final rules.



**III.USING PARALLEL COORDINATES TO VISUALIZE RULES:**

Parallel coordinates are a data visualization technique used to visualize multivariate data, making it suitable for representing association rules with multiple conditions and measures.



**CODE FILE:**



**CONCLUSION:**

The insights we've gained through this market basket analysis not only have the potential to boost revenue and customer satisfaction but also to streamline our inventory management and marketing efforts. It's imperative that we regularly revisit and update our analysis to remain agile in response to shifting market dynamics and evolving consumer behaviors.