Market Basket Analysis in Retail: Unsupervised Learning Methods for Association Rule Mining

Acknowledgement

Certificate

TABLE OF CONTENT

[1. Abstract 4](#_Toc446768916)

[2. Introduction 4](#_Toc1512110068)

[3. Literature review 4](#_Toc1591954126)

[4. Methodology 4](#_Toc237953708)

[5. Implementation 4](#_Toc2003491782)

[6. Analysis 4](#_Toc2028189829)

[7. Conclusion &Future work 4](#_Toc1162715407)

[8. Reference 4](#_Toc1900691088)

[Objective 6](#_Toc1319483938)

[Document structure 7](#_Toc275771467)

[Data Collection and Preprocessing: 10](#_Toc1715512322)

[Implementation Details: 11](#_Toc912044541)

[About The Algorithms 12](#_Toc1954905538)

[APRIORI 13](#_Toc1483294862)

[FP-GROWTH 15](#_Toc1528175366)

[Loading data and libraries 18](#_Toc261937524)

[Result 24](#_Toc466445163)

[Conclusion 27](#_Toc442742171)

[Future work 27](#_Toc1146227501)

ABSTRACT

Market basket analysis plays a crucial role in retail by uncovering hidden patterns in customer purchasing behaviour. This study explores the application of unsupervised learning methods for association rule mining in retail datasets. The research focuses on identifying frequent item sets and generating association rules to understand the relationships between products purchased together. A dataset containing transaction records from a retail store is used for analysis. There are several unsupervised learning methods for association rule mining, including the Apriori algorithm, FP-Growth algorithm and Eclat algorithm. In this paper we will compare algorithm . The Apriori algorithm is employed for mining association rules, and the results are evaluated based on support, confidence, and lift metrics. The findings reveal valuable insights into customer purchasing patterns, enabling retailers to optimize product placement, promotions, and inventory management strategies. The study contributes to the existing literature by demonstrating the effectiveness of unsupervised learning methods in market basket analysis and provides practical implications for retailers to enhance their marketing strategies.

INTRODUCTION

Market basket analysis (MBA) is a data mining technique used in the retail industry to analyse customer buying behaviour by identifying products that are frequently purchased together. It is a powerful tool that helps retailers discover hidden patterns and connections in sales data, helping them make informed decisions about product placement, pricing and marketing strategies. In this paper, we will explore the concept of market basket analysis and its importance in the retail industry. This paper aims to explore the application of unsupervised learning methods for association rule mining, which are commonly used to discover patterns in large data sets.

The process involves analysing transactional data, typically from point-of-sale systems, to find relationships between products. The most common metric used in market basket analysis is "support," which indicates the frequency with which a set of items appears together in transactions.

Another important metric is "confidence," which measures the likelihood that if a customer buys one item, they will also buy another item. "Lift" is a metric that compares the likelihood of two items being purchased together to the likelihood of them being purchased separately. A lift value greater than 1 indicates that the items are likely to be bought together, while a value less than 1 indicates that they are unlikely to be bought together.

Apriori is a popular algorithm for finding frequent item sets in transactional databases. It uses a "bottom-up" approach, where frequent subsets of items are extended one item at a time, and the algorithm prunes any item sets that are not frequent. This process continues until no further extensions are possible.

Apriori is widely used in market basket analysis because it efficiently identifies the frequent item sets needed to compute support, confidence, and lift values, which are essential for understanding purchase patterns and making informed business decisions.

# Objective

This study is to compare and evaluate different unsupervised learning algorithms for association rule mining, including Apriori and FP-growth, in terms of their efficiency and effectiveness in extracting meaningful associations from retail transaction data. The goal of market basket analysis is to uncover associations and patterns in consumer purchasing behaviour based on transactional data. By identifying which items are frequently bought together, retailers can:

1. Understand Customer Behaviour: Gain insights into the preferences and habits of their customers.

2. Optimize Product Placement: Place complementary products near each other to encourage additional purchases.

3. Improve Cross-Selling: Recommend related products to customers based on their purchasing history.

4. Enhance Promotional Strategies: Design targeted promotions and discounts to increase sales.

5. Optimize Inventory Management: Stock up on items that are frequently purchased together to meet demand more effectively.

# Document structure

The document consists of \_chapters

literature review

Numerous scholars have previously examined MBAs using a variety of methodologies. Rehman and Ghous's comprehensive review from 2021 addressed the subject of this study. The use of augmented reality and deep learning to MBA programs—more especially, online and offline retail—is the main focus.

In 2023, Apriori and FP-Growth algorithms were used to examine purchasing patterns based on real-world data. The results demonstrate that FP-Growth performs better in terms of memory and speed. In 2022, a new strategy suggested that recommended sets of products rather than single ones, increasing the likelihood that customers would make better decisions.

Apriori algorithms were employed in another study in 2022 to identify relationships in sales patterns, particularly in the auto parts and repair services industries. With certainty, the study found nine relationships. The algorithm known as Apriori In the same year, the Apriori algorithm was also used to create guidelines for product combinations in shopping baskets that increase revenue for firms.

In 2022, the FP Growth algorithm was used in another study to investigate group purchasing behaviour. The end result is a system that calculates the transactional data value of product relationships. In Pithoragarh, Uttarakhand, 45 consumers from different stores participated in a survey that same year. Three positive shopping patterns, each with a relationship strength value above 1, were discovered using the Apriori approach.

In 2021, a cool MBA study using smart Association Rules to analyse what customers purchase from a Vancouver Island grocery. 225 distinct items were examined using a tool named WEKA. They tested FP Growth and Apriori, but Apriori was less successful. Thus, using FP Growth, they determined the top 10 rules according to their level of certainty.

Another study conducted in 2020 at Maharani Supermarket examined when customers made purchases during holidays such as Christmas and fasting. For 12, 6, and 3 months, they advised serving milk with snacks. They also provided suggestions for how to arrange goods for the 2017 and 2018 New Year's festivities.

A research conducted that same year looked at the most popular items that customers purchased at Mart at various times. For example, a little Milo stick in the third quarter, an Alpha One 600ml in the second, and an Indomie Goreng 86gr in the first were all popular.

Imagine for a moment that the business uses the cool Association Rule to figure out what you buy when you're out shopping. They can learn what's popular and significant from this. However, it becomes increasingly difficult to determine what everyone like when more people shop. According to the brilliant experts, the best way to identify and

categorize these awesome shopping rules is to combine Association Rule with extremely intelligent deep learning (DL). But as of now, not many people are following through on this (Basysyar et al., 2021).

Methodology

# Data Collection and Preprocessing:

-The specific dataset that we used is from the UCI Machine Learning Repository containing 541909 Instances and 8 features (Invoice no., StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country).

-All transactions for a UK-based, registered non-store internet retailer that took place between January 12, 2010, and September 12, 2011, are included in this transnational data collection. The company mostly sells unusual presents suitable for any occasion. Retailers make up a large portion of the company's customer.

- Prior to analysis, the dataset will undergo through preprocessing for example removing irrelevant rows, duplicates, missing values and spaces etc.

- The dataset will then transform into a transaction format suitable for the analysis and the comparison of Apriori and FP-growth algorithm.

Algorithm:

- Two popular association rule mining algorithms, Apriori and FP-growth, will be implemented for comparison in this study.

- The Apriori algorithm IS implemented to mine association rules from the transaction dataset. The algorithm generates candidate item sets and prunes infrequent item sets based on a user-defined minimum support threshold.

- FP-tree (Frequent Pattern tree), is used to mine frequent item sets without the need for candidate generation, unlike the Apriori algorithm.

- Association rules are then generated based on a user-defined minimum confidence threshold.

Parameter Tuning:

- To achieve optimal performance and relevance, the parameters of the Association algorithm, such as the minimum support threshold, are tuned through iterative experimentation.

- By adjusting these parameters, researchers can control the granularity of the association rules generated and tailor the analysis to specific business objectives or datasets.

Evaluation Metrics:

- The effectiveness of the association rules generated by the Apriori algorithm is evaluated using metrics such as support, confidence, and lift.

-FP- Growth includes the generation of the FP-tree, which is useful in maintaining the association among the item set.

- These metrics provide insights into the frequency, reliability, and strength of the associations discovered, enabling researchers to assess the practical relevance of the generated rules for retail decision-making.

# Implementation Details:

Now a days python libraries, framework and modules are used for more efficiency, flexibility and to facilitate the implementation of complex algorithms.

Python libraries that we will use:

- **PANDAS**

For data manipulation and preprocessing. Pandas provides data structures like DataFrame, which are ideal for handling transactional data.

- **NUMPY**

For numerical computing. While not directly related to the Apriori algorithm, NumPy can be useful for handling arrays and matrices, which may be needed in your data preprocessing or analysis.

- **MATPLOTLIB**

It provides a variety of plotting functions and customization options, making it suitable for a wide range of data visualization tasks.

- **SEABORN**

Seaborn is a Python data visualization library based on Matplotlib that provides a high-level interface for drawing attractive and informative statistical graphics.

- **MLXTEND**

The mlxtend.frequent\_patterns.apriori module for implementing the Apriori algorithm. This library offers a straightforward implementation of the algorithm and allows you to easily generate frequent item sets and association rules.

# About The Algorithms

The mining technique for frequent patterns and association rules was initially introduced by Agrawal et al. The association rules are typically derived using two measures: confidence and support. The support measure determines the frequency of item sets in transactional databases, or the number of times an item appears in a customer transaction. Frequent item sets are those that meet the user-specified minimum support threshold level. A pattern's degree of precision and reliability in generating strong association rules is called its confidence. For frequent itemset mining and association rule creation, the most popular data mining algorithms are Apriori and FP-Growth.

# APRIORI

A frequent item set meets a user-defined support threshold and has a support count equal to the number of transactions it appears in. The Apriori algorithm uses a property where any subset of a frequent item set is also frequent. This allows for the identification of frequent one-item sets, which are then used to discover candidate frequent (k+1)-item sets by joining with themselves. The algorithm tests these candidate sets to see if all of their k-item subsets are in the frequent k-item sets. This process continues until no successful (k-1) subset is found, at which point the set is discarded and the algorithm stops. The final result is the union of all the frequent k-item sets.

The strength of the associations—that is, the efficacy of the rules is measured using two metrics:

1.Confidence

2. Support.

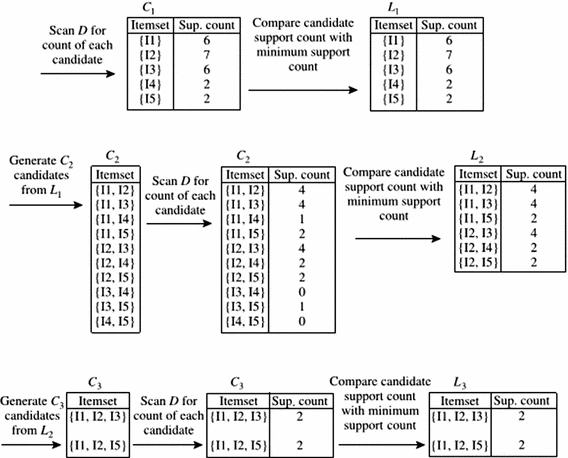
In a transactional database, the frequency of a rule is determined by the support team based on how frequently it arises in customer transactions.

The confidence quantifies the frequency with which the item in Y appears in client transactions containing the item X. It is the conditional likelihood that a transaction containing X will also contain Y.

For illustration, let's examine the table that contains the five items' transaction data

|  |  |
| --- | --- |
| Transaction ID | List of items |
| T100 | I1, I2, I5 |
| T200 | I2, I4 |
| T300 | I2, I3 |
| T400 | I1, I2, I4 |
| T500 | I1, I3 |
| T600 | I2, I3 |
| T700 | I1, I3 |
| T800 | I1, I2, I3, I5 |
| T900 | I1, I2, I3 |

Apriori counts the support of each item as an itemset in the database based on how frequently it occurs. This technique is known as frequent itemset mining. Itemsets that have support that is at least as high as the minimal threshold are identified. As a result, the algorithm must search through the whole database until it finds no more itemsets that have greater support than the minimal threshold before extracting each frequently occurring itemset.



Since items with support of two are taken into consideration for the further steps of the algorithms in the first step, as can be seen in the graphic above, the minimal support criterion is two.

Additionally, it sends item sets with a minimum support count of two for additional processing in the subsequent steps.

# FP-GROWTH

The popular and original Apriori approach for frequent pattern mining has a major drawback in that it necessitates many transaction dataset scans and produces an enormous amount of candidate itemsets, which lowers the program's efficiency. It works through unnecessary candidate item sets and repeatedly evaluates large item sets by pattern matching the database. Candidate generation can dominate the cost of scanning the database when the number of items and pattern length increase.

The most widely used, quicker, and less memory-intensive tree-based technique for frequent pattern recognition in sizable transactional databases is called FP-Growth.

FP-Growth is a method for mining frequent patterns. It uses a prefix-tree structure called the frequent pattern tree to store compressed information. The tree consists of a header table and the frequent patterns. It also sorts the items in the frequent pattern or conditional pattern base.

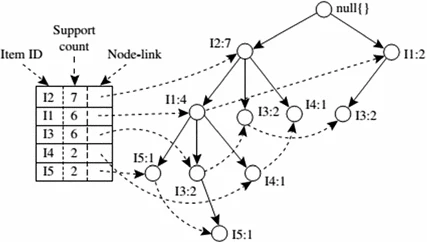
Now let’s illustrate FP-growth Algorithm.

The Apriori method generated candidates for creating item sets, as we have observed. This approach uses a tree structure to represent the data in the FP-Growth process. FP-tree is the name we give to this lexicographic tree construction. It takes the responsibility of maintaining the association data among the frequently used things.

Each frequent item's set of conditional FP-Trees is separated from the FP-Tree after it has been created. Further mining and measurement of a set of conditional FP-Trees is possible. The dataset we utilized in the apriori technique, for instance, is comparable to the database. This will result in the following table for each conditional FP-Tree.

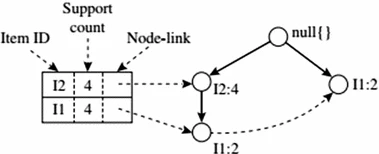
|  |  |  |  |
| --- | --- | --- | --- |
| item | Conditional pattern base | Conditional FP-Tree | Frequent pattern generated |
| I5 | {{I2, I1: 1}, {I2, I1, I3: 1}} | {I2: 2, I1: 2} | {I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2} |
| I4 | {{I2, I1: 1}, {I2: 1}} | {I2: 2} | {I2, I4: 2} |
| I3 | {{I2, I1: 2}, {I2: 2}, {I1: 2}} | {I2: 4, I1: 2}, {I1: 2} | {I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2} |
| I1 | {{I2: 4}} | {I2: 4} | {I2, I1: 4} |

The following picture illustrates the frequently occurring things based on the table above.

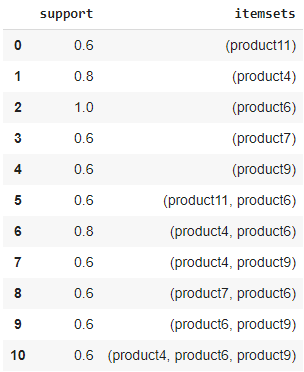


The support for I2 is seven, as we can see here. There it appeared four times with I1, twice with I3, and once with I4. A single scan is enough to construct the itemsets in this case, as opposed to the apriori algorithm's repeated table scanning that produced the frequent set.

It will resemble the following conditional FP-Tree for I3.



If we implement both algorithms using python, we can observe that in the apriori comparison, the data frame series in input had a similar series for the frequent itemset, but in the FP-growth case, the series is arranged in descending order of support value.

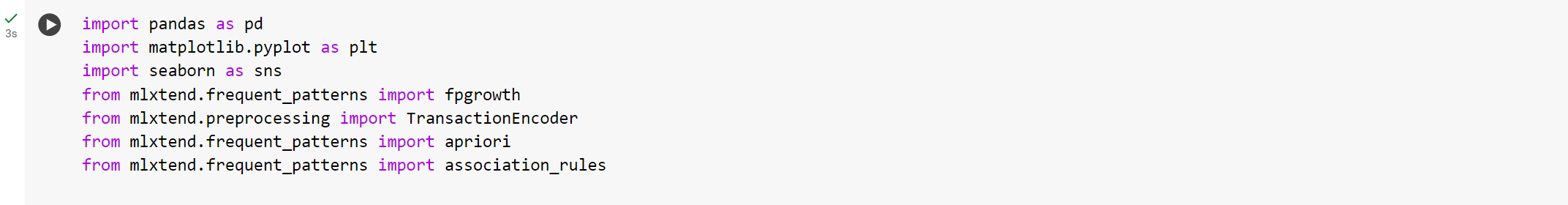
 

1. In case of Apriori 2. In case of FP-growth

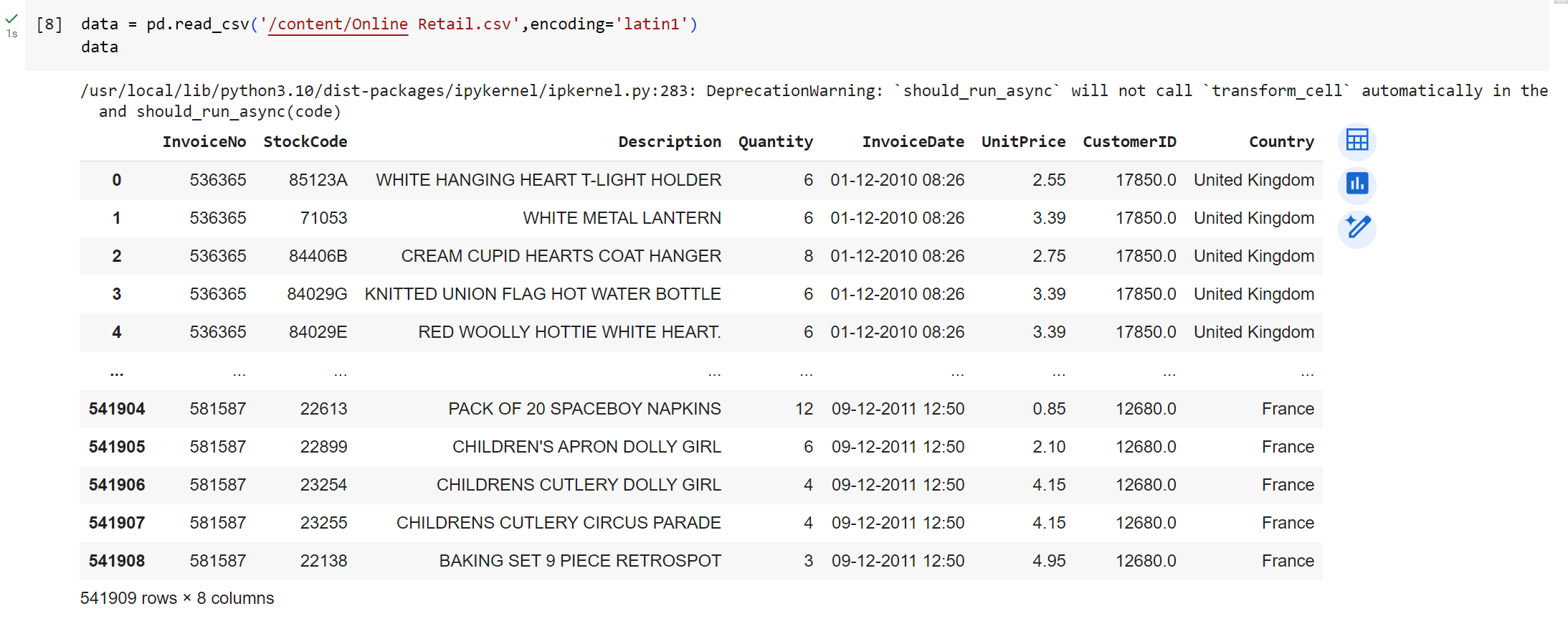
Model implementation

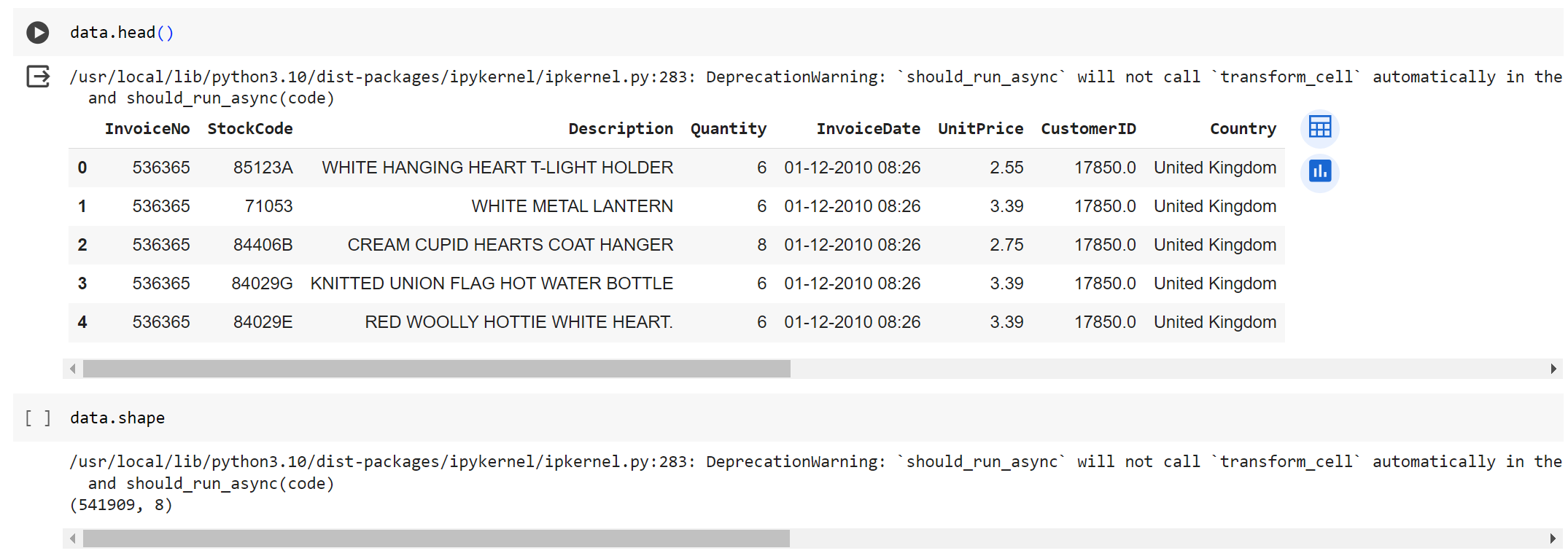
Having discussed the theoretical background and methodology of market basket analysis using the Apriori and FP-growth algorithms, this section focuses on the practical implementation of the algorithms on real-world retail transaction data. The implementation aims to demonstrate the application of Apriori and FP-growth in extracting meaningful association rules from transaction datasets and deriving actionable insights for retailers.

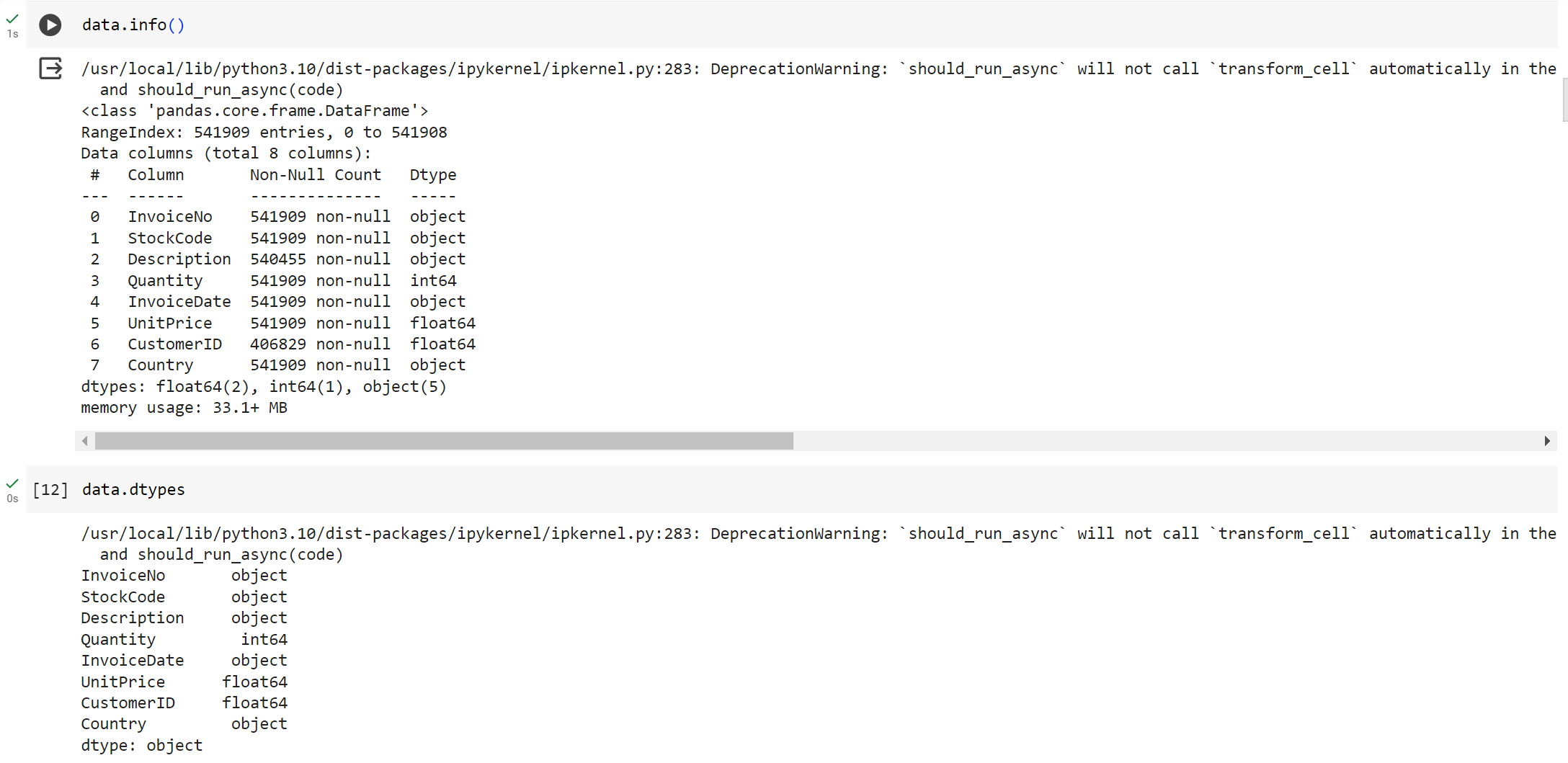
# Loading data and libraries

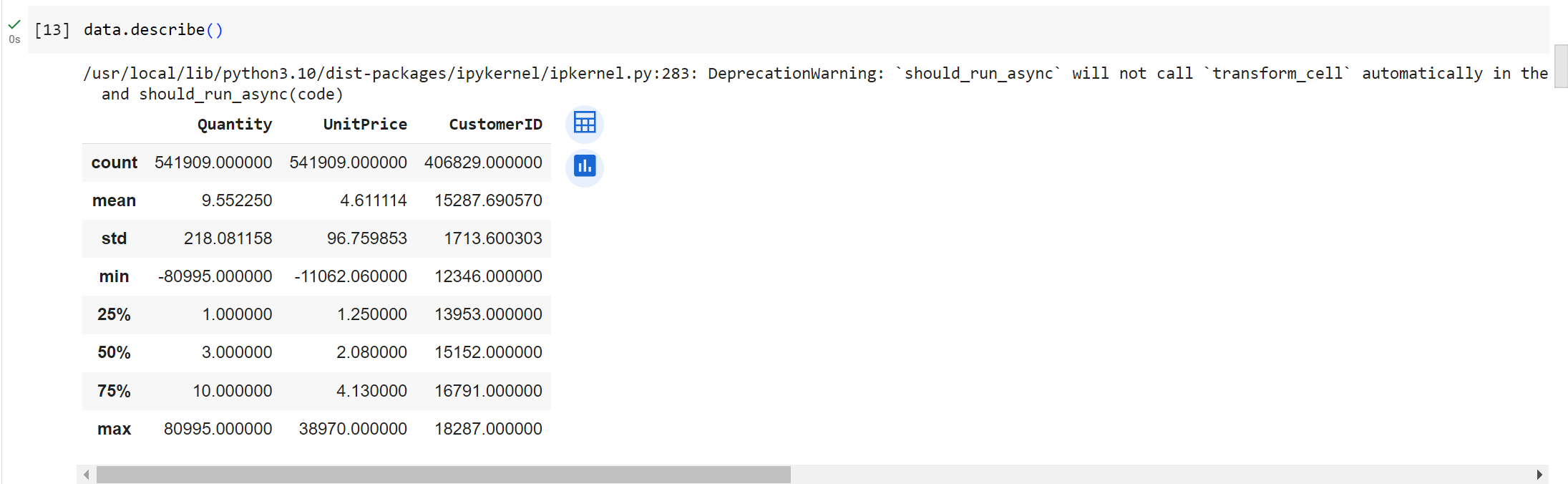


We’ll use the encoding as latin1 to read this file as there are some special characters in this file.)



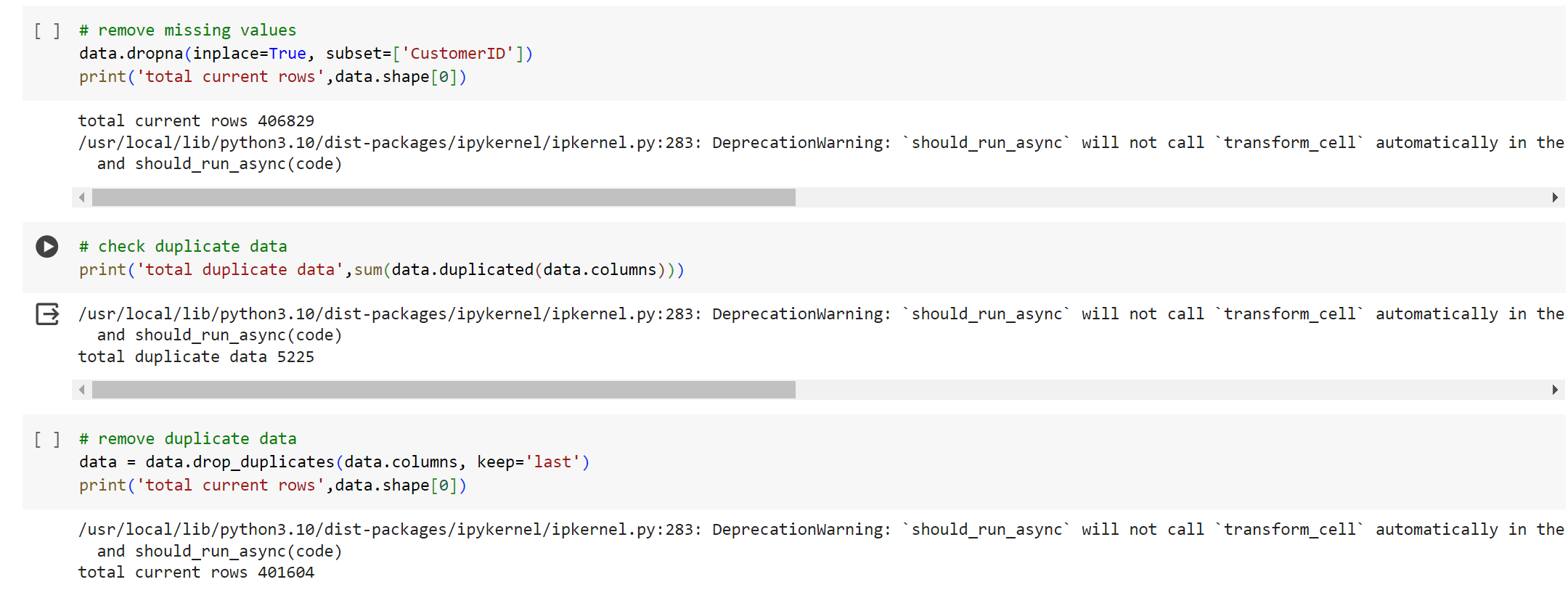


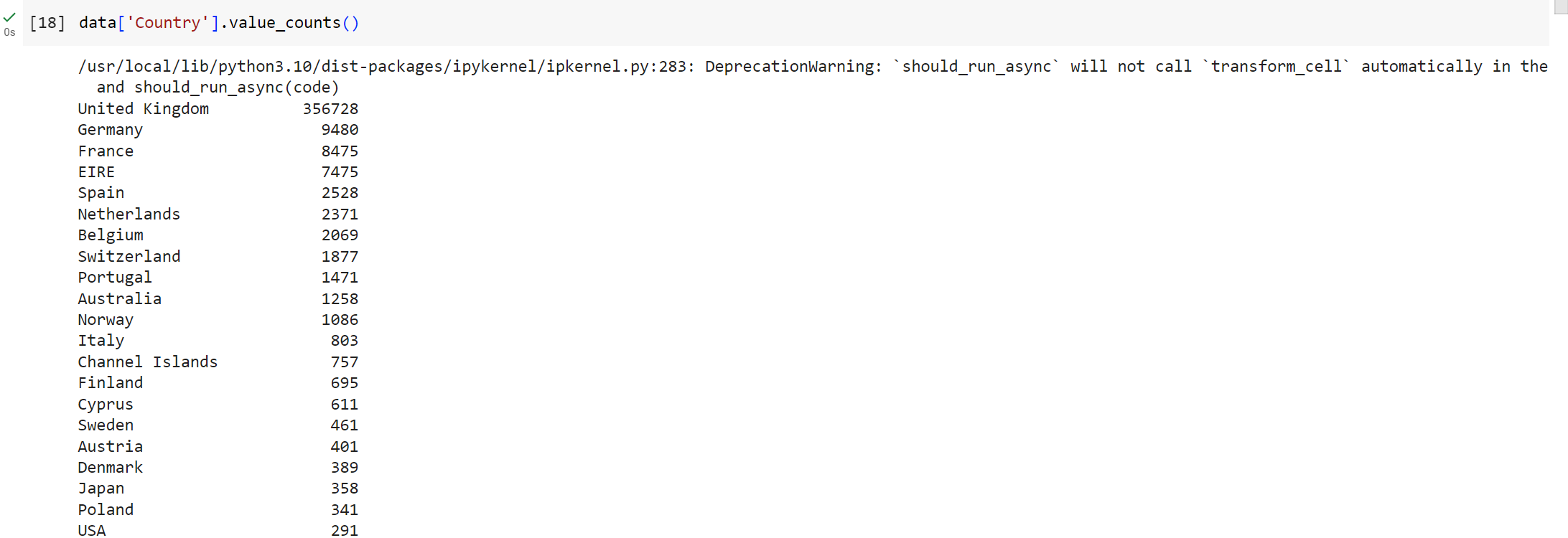


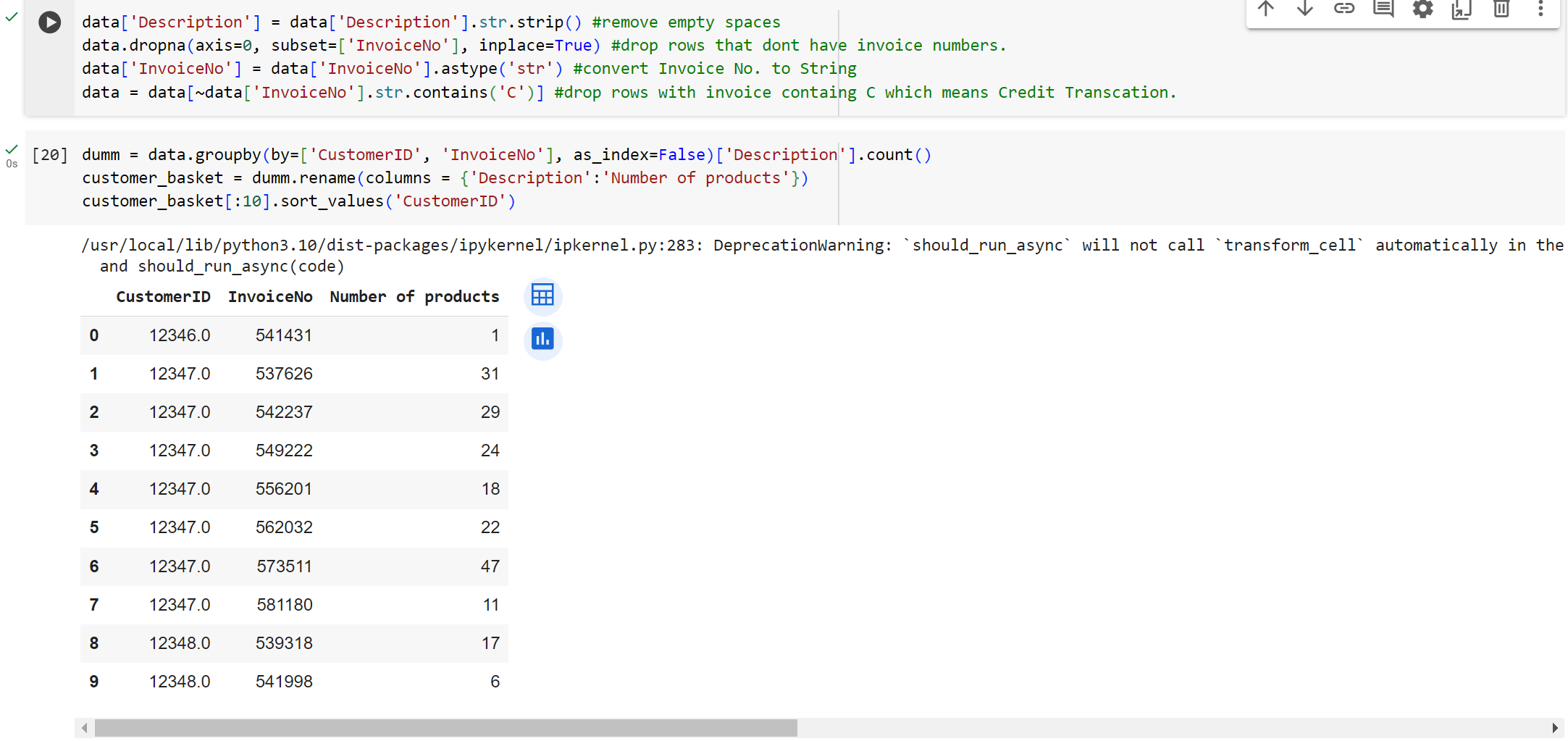


Data Cleaning

* Check missing values
* Remove missing values
* Check duplicate data
* Remove duplicate data
* First, we will remove spaces in descriptions.
* Drop the rows that don’t have invoice numbers
* Remove the credit transactions (those with invoice numbers containing C)



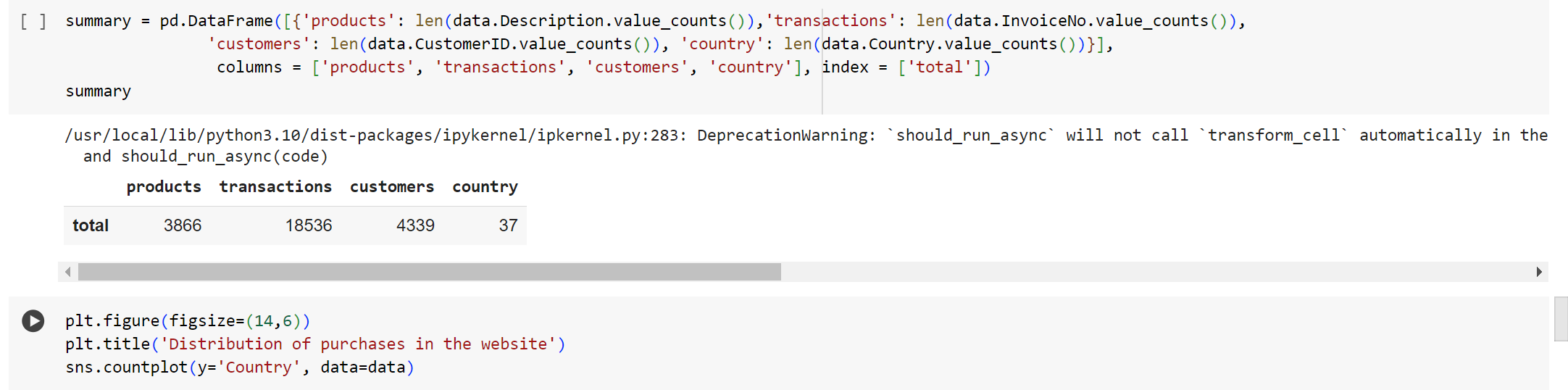


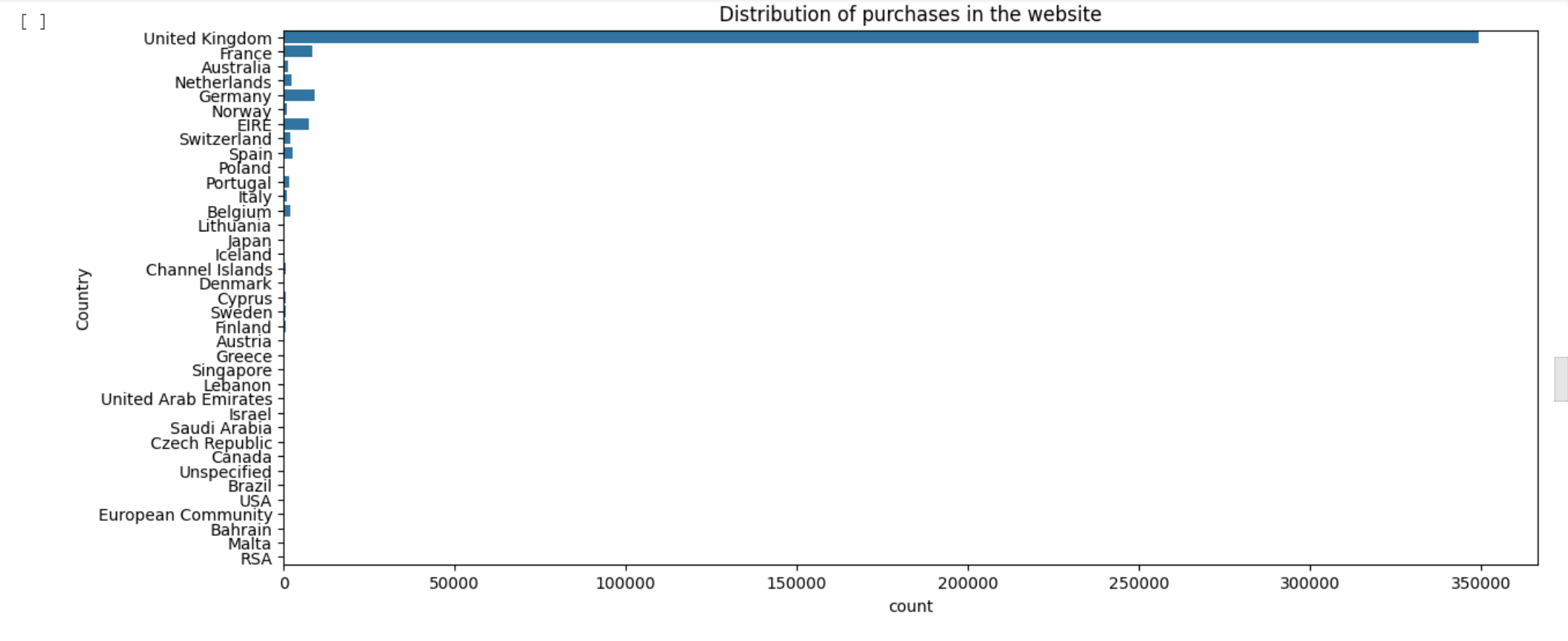


Insight from this data

* It was possible to determine that certain customer had only visited once and made a single purchase (CustomerID 12346).
* regular customers who purchase a lot of goods with each order (CustomerID 573511)

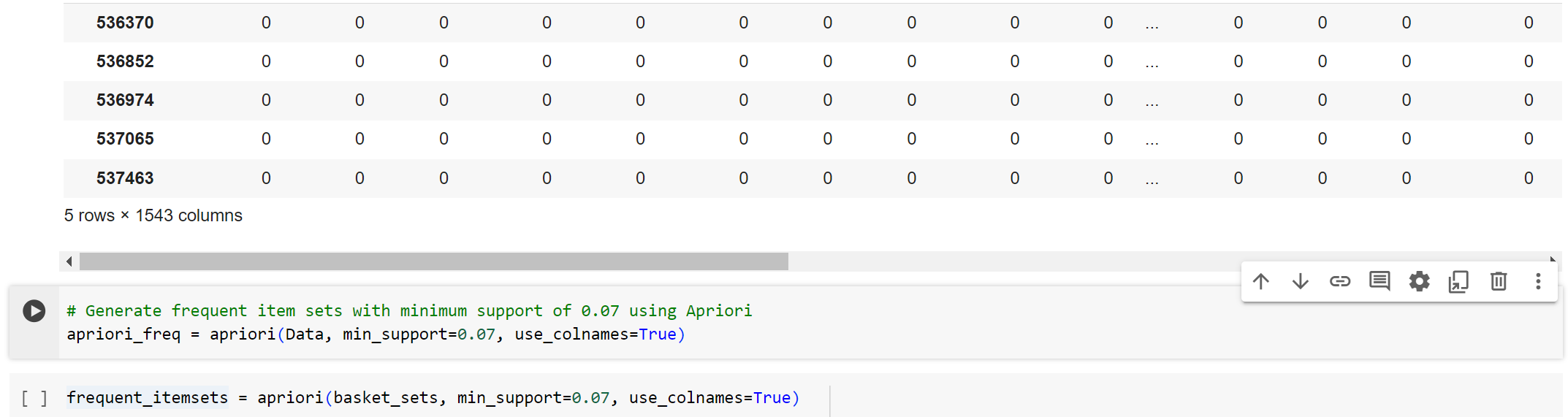
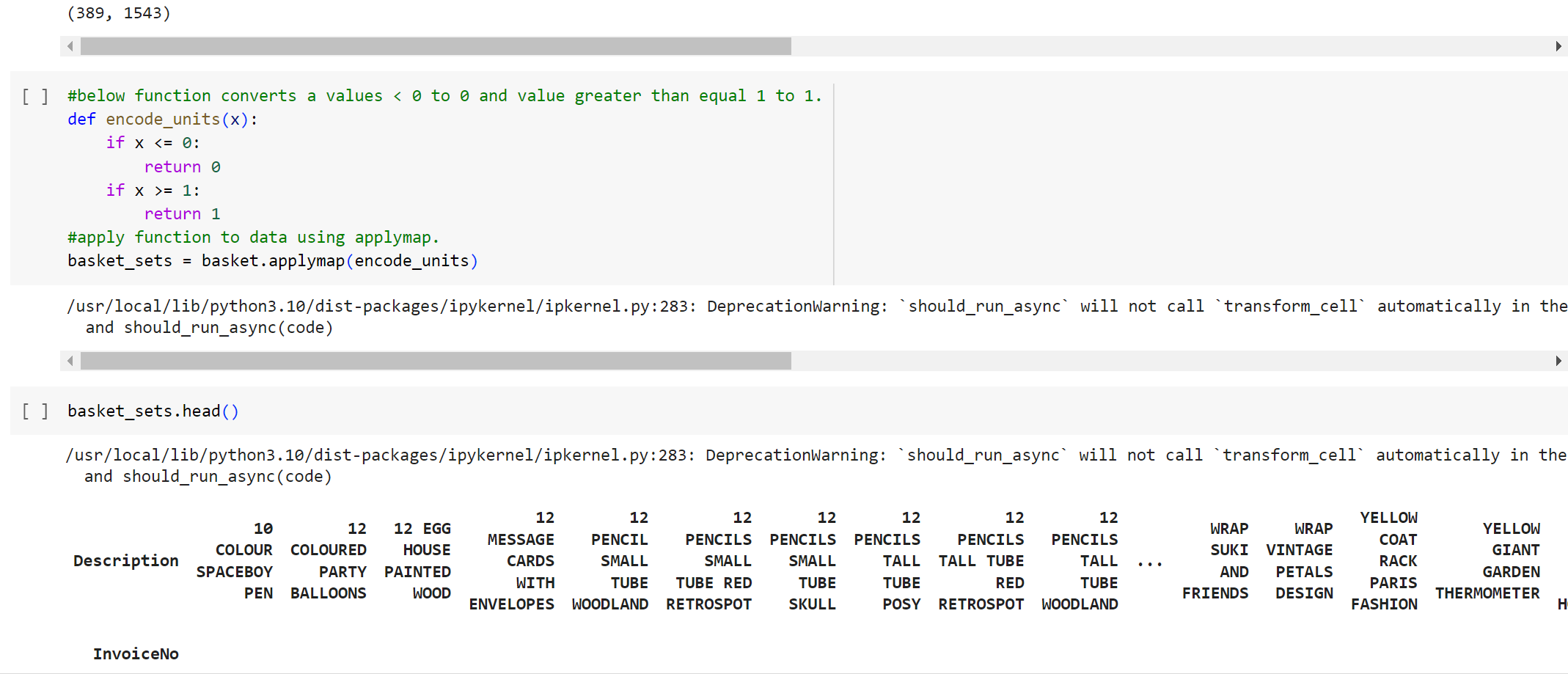
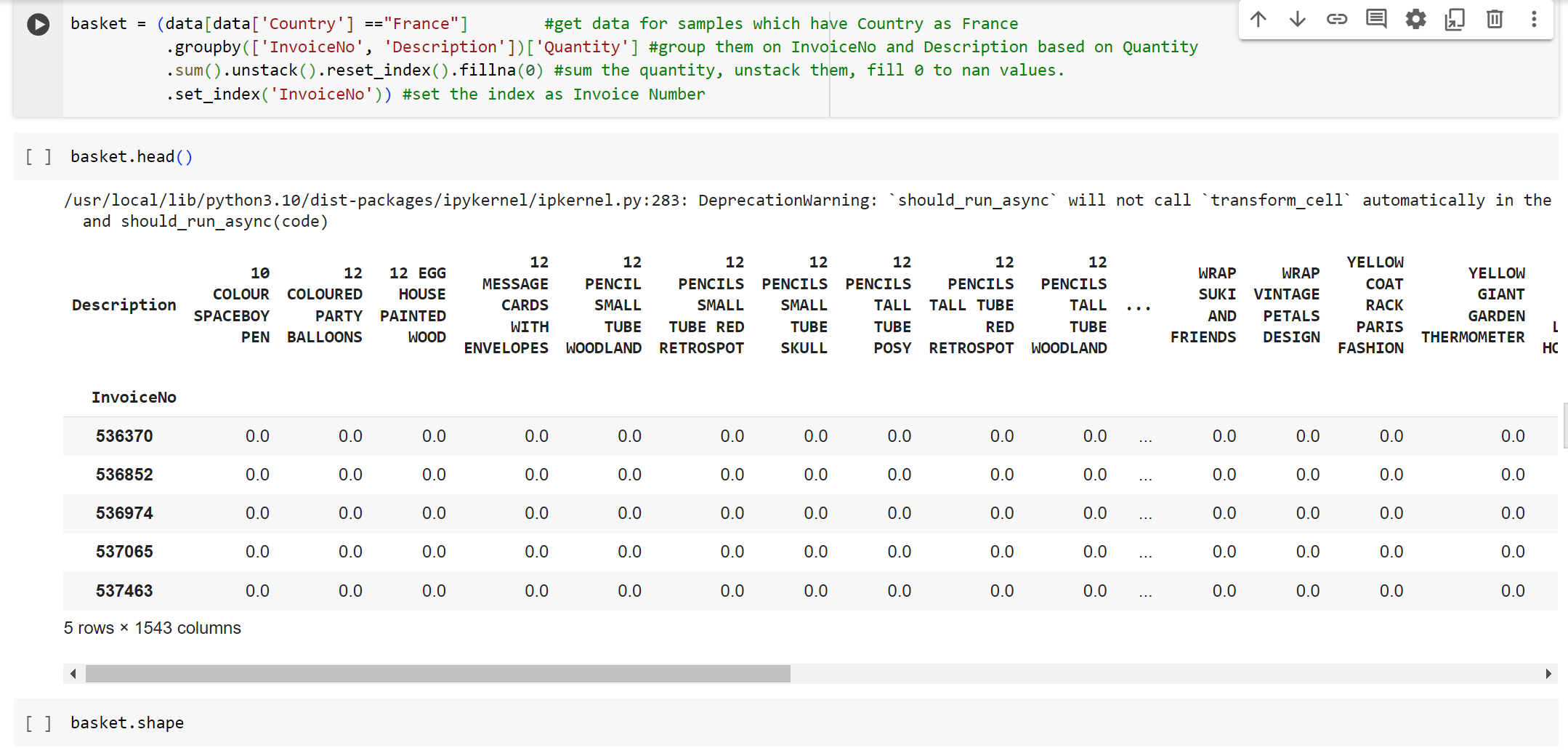
The retailer appears to have 3866 distinct items. They service 4339 clients and are present in 37 countries with over 18536 total order transactions.





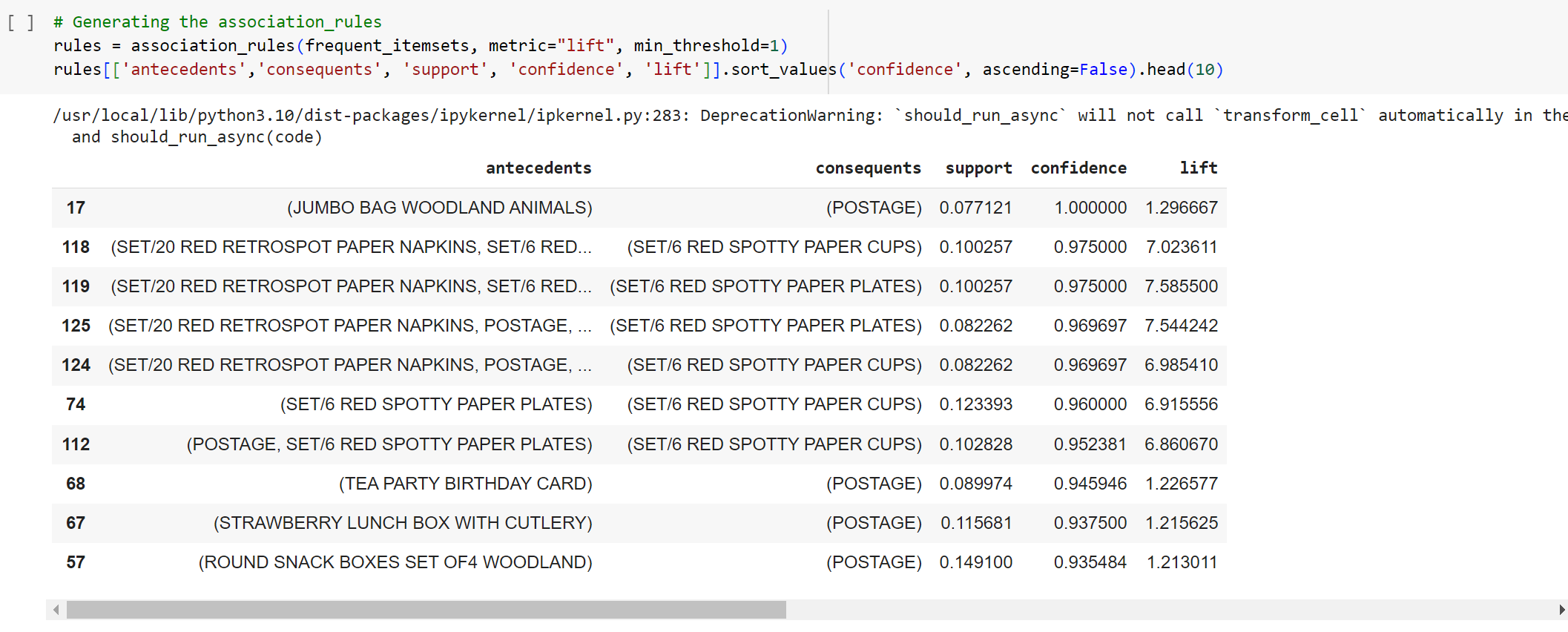
Following the cleanup, we must combine the items into a single row of transactions, with a single hot encoded product for each item. We are only looking at sales for France in order to keep the data set short.

The data contains a lot of zeros, but we also need to make sure that any positive values are converted to ones and that any values that are less than zeros are set to zeros.

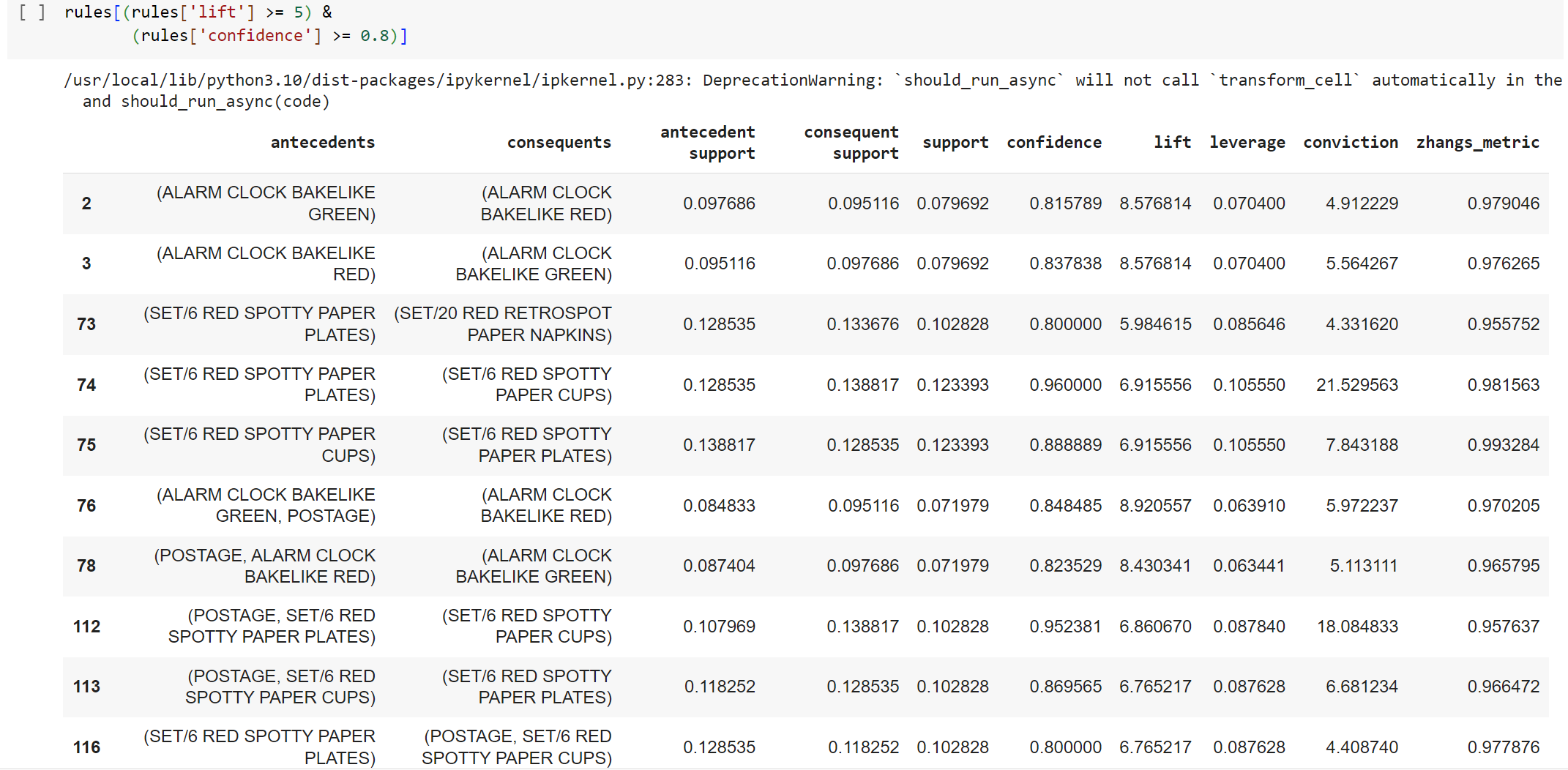


We can create frequent item sets with a support of at least 7% now that the data is correctly formatted (we chose this threshold to ensure we can find enough relevant examples).

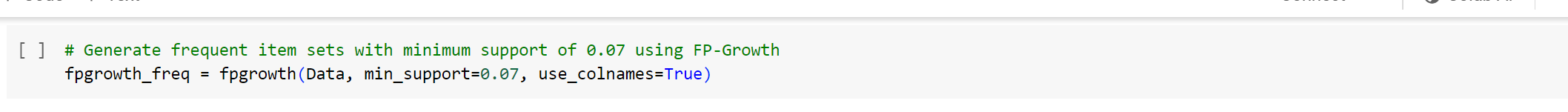
Generating the rules with the appropriate lift, confidence, and support is the last stage.

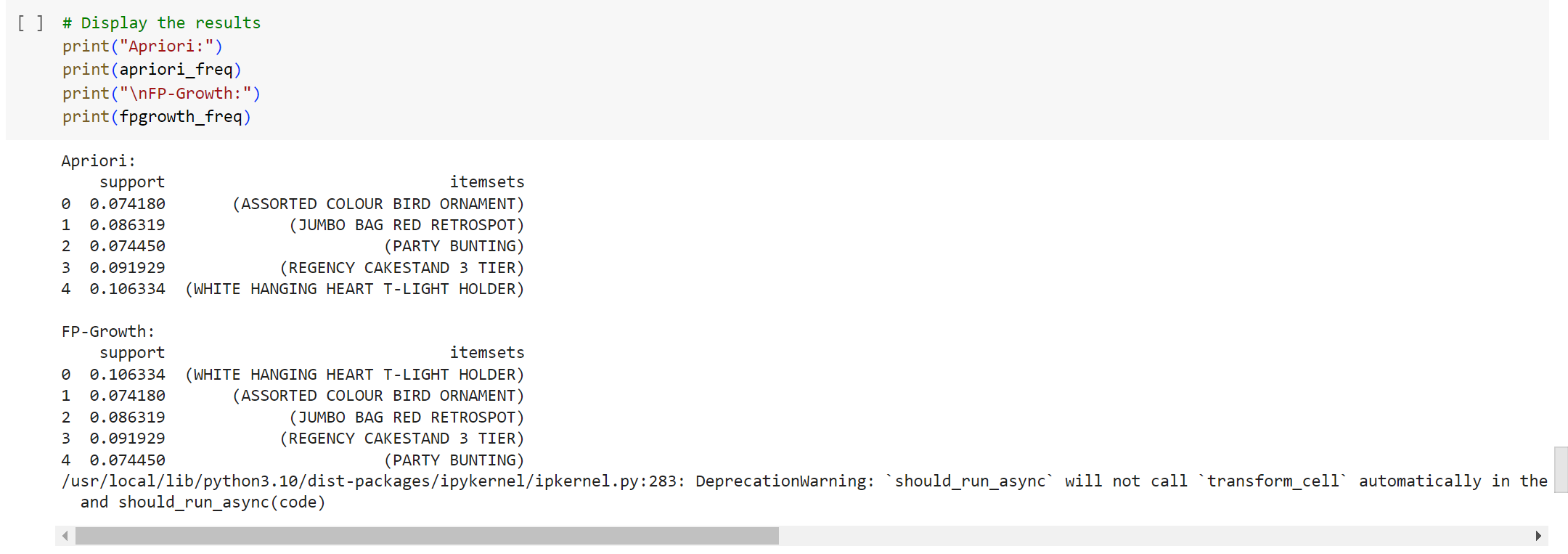


To check lift and confidence, we filter rules. The domain knowledge will be useful in this section of the analysis. As we don't have that, we'll only search for a few instances to help us along.Using the normal pandas code, we can filter the dataframe. Seek for a significant lift (5) and strong confidence (.8) in this instance.

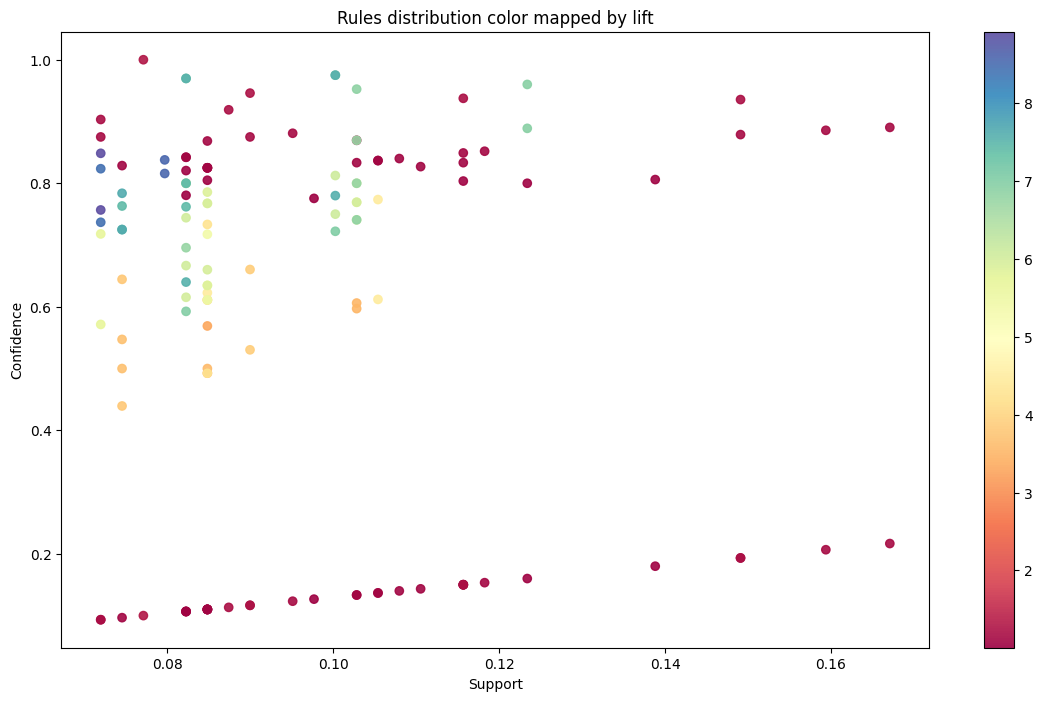
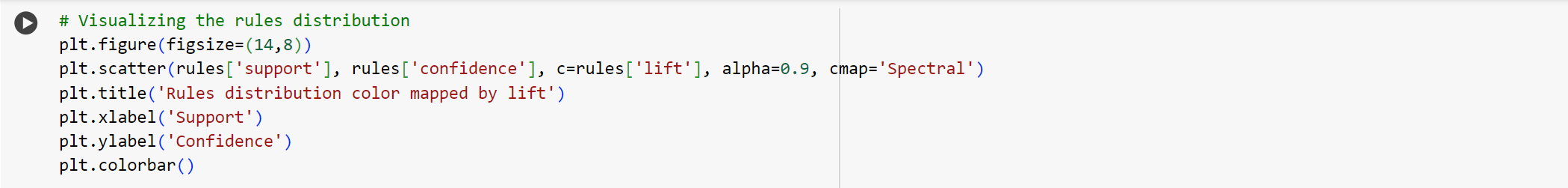


Same with FP-growth we’ll generate frequent item sets with the min\_support of 0.07.Same with FP-growth we’ll generate frequent item sets with the min\_support of 0.07.





Visualizing the rules distribution



# Result

* The retailer appears to have 3871 distinct items. They service 4334 clients and are present in 37 countries with over 18500 total order transactions.
* We can learn that the e-commerce serves 4339 clients in 37 countries worldwide, has 18536 total order transactions, and offers 3866 distinct products.

* The UK leads the world in e-commerce with 16579 invoice transactions, followed by Germany and France with 1694 and 1541 transactions, respectively.

* Support threshold of France is (0.06).
* According to the rules, it appears that the red paper cups, napkins, and plates are bought in a way that is more than what the overall likelihood would imply, as do the green and red alarm clocks.
* Grouping products that co-occur in the design of a store’s layout to increase the chance of cross-selling

ANALYSIS

Conclusion & Future work

# Conclusion

Market basket analysis, facilitated by unsupervised learning methods such as association rule mining, remains a powerful tool for retailers seeking to understand customer purchasing behaviour and optimize their marketing strategies. This study has demonstrated the effectiveness of algorithms such as Apriori and FP-growth in extracting meaningful associations from retail transaction data, providing valuable insights for decision-making.

Through the analysis of association rules, retailers can identify product affinities and customer preferences, enabling them to create targeted promotions, improve product recommendations, and enhance overall customer satisfaction. By leveraging these insights, retailers can also optimize inventory management and store layout to better meet customer needs and increase profitability.

# Future work

Future research in the area of market basket analysis in retail markets has a number of promising directions. To increase the precision and applicability of attribution rules, one possible study topic is the integration of new data sources, such as consumer demographics and web browsing activity. Retailers may get a deeper understanding of their customers and modify their marketing tactics by fusing transaction data with customer segmentation and predictive modelling approaches.

Moreover, there is still a problem with the scalability of association rule mining algorithms, particularly when merchants gather a lot of transaction data. Subsequent studies should concentrate on creating scalable and more effective algorithms for handling massive volumes of data and enabling real-time transaction data analysis.

Furthermore, the scalability of association rule mining algorithms remains a challenge, especially when retailers accumulate large amounts of transaction data. Future research can focus on developing more efficient and scalable algorithms to process large amounts of data and enable real-time analysis of transaction data.

Additionally, there is a need for more research in the areas of association rule interpretability and market basket analysis insight usability. The development of decision support systems and visualization tools that make it easier to convert insights into workable plans could increase market basket analysis's usefulness for merchants.

All things considered, market basket analysis is still a useful technique for merchants looking to learn from transaction data. The effectiveness of market basket analysis in retail decision-making can be further improved by researchers and practitioners by tackling the problems mentioned earlier and looking into new research directions.

REFRENCES