**ABSTRACT**

This abstract introduces a novel method for identifying product bundles within large datasets. By leveraging advanced data analytics and machine learning techniques, our approach efficiently analyses product attributes, sales patterns, and customer behaviour to accurately identify and categorize product bundles. The proposed system significantly reduces manual effort and improves the accuracy of bundle identification, enabling businesses to optimize pricing strategies, enhance customer experience, and increase revenue.

Product Bundle Identification using Semi Supervised Learning:

Information systems

* World Wide Web
* Web applications
* Electronic commerce
* Online shopping

Theory of computation

* Theory and algorithms for application domains
* Machine learning theory

**Pictorial Identification of The Project** :



These are some kind of the representation of the project which deals about the bundling the frequent types of the materials required.

Some More Applications of The Project In Daily Life:





Online shopping also uses the process same as bundling the product criteria.

What is product bundling?

Product bundling is a technique in which several products are grouped together and sold as a single unit for one price. This strategy is used to encourage customers to buy more products. McDonald’s Happy Meals are an example of product bundles. Instead of selling a burger, soda, and french fries separately, they are sold as a combination, which leads to more sales than offering them separately.

## Advantages of product bundling

Bundling helps you do much more with your existing stock. Let’s take a look at the advantages of product bundling and how it can be beneficial for your business.

### 1.Increase your average order value

Product bundling can increase the profits and sales of individual items over time. By grouping your items together you can make your customers buy more than one product during a single purchase, which increases your average order value. For example: Instead of buying just one pencil during a single purchase, your customer can be given an option to buy a pencil, eraser and sharpener as a bundle, making them purchase more than one product thereby increasing your average order value.

### 2.Decreases marketing and distribution costs

Bundling enables you to sell more and decrease marketing and distribution costs. Instead of marketing every product you can group complementary products together and market them as a single product. By packaging different items together you only need one warehouse bin to store them instead of different bins. Also, bundling helps you ship fewer boxes of individual items and saves you money on postage.

### 3.Reduce inventory waste

Merchandise that doesn’t get sold remains in your inventory as dead stock, adding to your holding costs, and is eventually discarded as waste. You can use bundling to clear out this dead stock before it becomes a problem. If you bundle a slow-moving or stagnant item with a faster-selling product, customers will see the bundle as a bargain and be more inclined to buy it. This helps reduce your inventory waste, free up warehouse space, and decrease your inventory holding costs.

## Types and Examples of Product Bundles

There are several different bundling techniques which are used to group products:

1. Pure bundles
2. New product bundles
3. Mix-and-match bundles
4. Cross-sell bundles
5. Gifting bundles
6. Inventory clearance bundles
7. Buy-one-get-one bundles

### 1.Pure bundles

In pure bundling, the individual products that make up the bundle can be purchased only as a bundle and not as standalone products. This technique limits the choices offered to the consumer. For example, HelloFresh is a company which does pure bundling successfully. It bundles the ingredients that their customers need to cook a healthy meal. They offer meal options based on the number of people and recipes the customer requires each week, but they don’t allow you to choose the ingredients as individual items that can be bought separately.

### 2.New product bundling

In this technique, newly-launched products are grouped along with existing or popular products as a promotion to help customers discover your latest product. This method is used by ecommerce stores, which mix new products with their well-known merchandise to gain some exposure for the new product. The more well-received the existing product is in the market, the more it brings the buyer closer to the new product. For example: The Nintendo switch + the legend with Zelda product bundle, is one of the fast-selling Nintendo’s bundle, in this bundle Nintendo introduced their brand new games which is grouped together along with their existing best selling products. This bundle offers an unique Zelda carrying case which is available with this bundle exclusively and two brand new games (Breath of the wild and Super Mario Kart) along with the accessories for gaming.

### 3.Mix-and-match bundles

The mix-and-match bundling technique allows the customer to choose among multiple similar products. This is mostly done by brick-and-mortar stores for fast-moving consumer products such as perishables or bulk items. Here, you specify a few products for your customers to choose from and they can create their own custom bundle from the options available. This method helps the customer feel that they’re in direct control of what they want to buy, thereby increasing the perceived value of the item. It’s the perfect method for encouraging your customers to buy products in bulk without forcing them to buy items which don’t interest them. For example, some retail stores offer a deal where you can match complementary pieces of clothing from an array of choices for a fixed price, such as any shirt along with any pair of trousers for $50.

### 4.Cross-sell bundles

In this bundling technique, retailers sell a complementary product as an add-on to a main product. This type of bundling works well with lower-priced items, or accessories or parts that go with a more expensive item. For example, if you buy an iPhone, you would probably like to buy a case along with it. So the iPhone and case can be sold together as a bundle.

### 5.Gifting bundles

Gift bundles are aimed at shoppers who want to give a bundle of complementary products together to a loved one. This type of bundle is mostly sold during holiday seasons. For example, beauty brand Estee Lauder offers a popular protect-and-hydrate gift set containing four skincare products that work together.

### 6.Inventory clearance bundling

In this bundling technique, you pair a faster-moving item in the inventory with a stagnant or slower-moving item to clear inventory space and decrease your inventory holding costs. This method includes discounts on your bundles so that shoppers who are interested in a top-selling item will see the whole bundle as a bargain and will be more inclined to buy it. For example, the popular specialty tea retailer T-WE found out that their tea accessories were selling faster than their teas (which was unfortunate, because the teas offered a higher profit margin).

### 7.Buy-one-get-one bundles

This bundling is used when you buy one main item, you can avail a discount for another complimentary product or get another product free. This is a best used technique for one time purchase products For example, electronics, if a customer buys a hair dryer they wouldn’t be coming in to buy the same product again. Hence, offering a complimentary product, discount or gift card will encourage your customers to add more items to their carts at a lesser price.

Bundling adds value to your products by adding extra features or products to your existing purchase. You can tailor your product offerings according to the preference of your customers to align with their wants.

## Introduction to R

R is a language and environment for statistical computing and graphics. It is a [GNU project](http://www.gnu.org/) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R’s strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the [Free Software Foundation](http://www.gnu.org/)’s [GNU General Public License](https://www.r-project.org/COPYING) in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

## The R environment

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

* an effective data handling and storage facility,
* a suite of operators for calculations on arrays, in particular matrices,
* a large, coherent, integrated collection of intermediate tools for data analysis,
* graphical facilities for data analysis and display either on-screen or on hardcopy, and
* a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R, like S, is designed around a true computer language, and it allows users to add additional functionality by defining new functions. Much of the system is itself written in the R dialect of S, which makes it easy for users to follow the algorithmic choices made. For computationally-intensive tasks, C, C++ and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

Many users think of R as a statistics system. We prefer to think of it as an environment within which statistical techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

### Advantages of R Programming

**Open Source:** R is an open-source language and is free to download and use. One can also contribute by optimizing its source code.

**Platform independent:** R is platform-independent and can work on all the operating systems like UNIX, Windows, and Mac.

**Data Wrangling:** Through its packages like readr and dplyr, R has the capability of converting a messy code into a structured one.

**Plots and Graphs:** Through ggplot and plotly, R creates attractive graphs with notations and formulas.

**Package Availability:** R has numerous packages dedicated to the development of machine learning, data analysis, and statistical projects.

### Disadvantages of R

**Memory:** R consumes more memory as all the objects get stored in the physical memory. Over time, as the program has bigger data, the process slows down.

**Security:** R lacks basic security that makes it practically difficult to embed in web applications.

**Difficult to learn:** Unlike Python, R is a complicated language and is difficult for a beginner to learn.

**Slow Runtime:** R is a slow processing language. In comparison to other languages such as MATLAB and Python, it takes more time to give an output.

**Data Handling:** Data handling in R is tedious as it requires all the data to be in one place. It is not ideal for Big Data. However, it does have an integration that makes handling slightly easier.

**ABOUT APRIORI ALGORITHM:**

Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyses that people who bought product A also bought product B.

The primary objective of the apriori algorithm is to create the association rule between different objects. The association rule describes how two or more objects are related to one another. Apriori algorithm is also called frequent pattern mining. Generally, you operate the Apriori algorithm on a database that consists of a huge number of transactions. Let's understand the apriori algorithm with the help of an example; suppose you go to Big Bazar and buy different products. It helps the customers buy their products with ease and increases the sales performance of the Big Bazar. In this tutorial, we will discuss the apriori algorithm with examples.

## What is Apriori Algorithm?

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

Components of Apriori algorithm

The given three components comprise the apriori algorithm.

1. Support
2. Confidence
3. Lift

Let's take an example to understand this concept.

We have already discussed above; you need a huge database containing a large no of transactions. Suppose you have 4000 customers transactions in a Big Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together.

Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

### **Support**

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= 400/4000 = 10 percent.

### **Confidence**

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

Hence,

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

### **Lift**

Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.

Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

= 50/10 = 5

It means that the probability of people buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone. If the lift value is below one, it requires that the people are unlikely to buy both the items together. Larger the value, the better is the combination.

How does the Apriori Algorithm work in Data Mining?

We will understand this algorithm with the help of an example

Consider a Big Bazar scenario where the product set is P = {Rice, Pulse, Oil, Milk, Apple}. The database comprises six transactions where 1 represents the presence of the product and 0 represents the absence of the product.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Transaction ID** | **Rice** | **Pulse** | **Oil** | **Apple** |
| t1 | 1 | 1 | 1 | 0 |
| t2 | 0 | 1 | 1 | 1 |
| t3 | 0 | 0 | 0 | 1 |
| t4 | 1 | 1 | 0 | 1 |
| t5 | 1 | 1 | 1 | 0 |
| t6 | 1 | 1 | 1 | 1 |

The Apriori Algorithm makes the given assumptions

* All subsets of a frequent itemset must be frequent.
* The subsets of an infrequent item set must be infrequent.
* Fix a threshold support level. In our case, we have fixed it at 50 percent.

**Step 1**

Make a frequency table of all the products that appear in all the transactions. Now, short the frequency table to add only those products with a threshold support level of over 50 percent.

We find the given frequency table.

|  |  |
| --- | --- |
| **Product** | **Frequency (Number of transactions)** |
| Rice (R) | 4 |
| Pulse(P) | 5 |
| Oil(O) | 4 |
| Milk(M) | 4 |

The above table indicated the products frequently bought by the customers.

**Step 2**

Create pairs of products such as RP, RO, RM, PO, PM, OM. You will get the given frequency table.

|  |  |
| --- | --- |
| **Itemset** | **Frequency (Number of transactions)** |
| RP | 4 |
| RO | 3 |
| RM | 2 |
| PO | 4 |
| PM | 3 |
| OM | 2 |

**Step 3**

Implementing the same threshold support of 50 percent and consider the products that are more than 50 percent. In our case, it is more than 3

Thus, we get RP, RO, PO, and PM

**Step 4**

Now, look for a set of three products that the customers buy together. We get the given combination.

1. RP and RO give RPO
2. PO and PM give POM

**Step 5**

Calculate the frequency of the two item sets, and you will get the given frequency table.

|  |  |
| --- | --- |
| **Itemset** | **Frequency (Number of transactions)** |
| RPO | 4 |
| POM | 3 |

If you implement the threshold assumption, you can figure out that the customers' set of three products is RPO.

We have considered an easy example to discuss the apriori algorithm in data mining. In reality, you find thousands of such combinations.

Note: Small Project On Product Bundle Identification Steps To Get The Details Of The Bundle That Too With A Jitter Plot Analysis Of The Data Using Apriori Algorithm.

The "arules" package is an R library that provides functionality for mining frequent itemsets and association rules from transaction data. It is widely used in market basket analysis and other data mining tasks.

To get started with the "arules" package, you'll need to install it first. You can do this by running the following command in R:

Once the package is installed, you can load it into your R session using the **library()** function:

The "arules" package provides several important classes and functions for association rule mining. Here are some of the key ones:

1. **read.transactions()**: This function is used to read transaction data from various formats, such as a matrix or a CSV file, and convert it into the transaction format required by the package.
2. **apriori()**: This function is used to mine frequent itemsets and association rules from transaction data using the Apriori algorithm. It allows you to specify parameters such as minimum support and confidence thresholds.
3. **inspect()**: This function is used to visualize and explore the results of association rule mining. It allows you to view the generated rules and associated statistics, such as support, confidence, and lift.
4. **subset()**: This function allows you to filter association rules based on specific criteria, such as minimum support or minimum confidence.

These are just a few examples of the functions provided by the "arules" package. The package also includes functions for manipulating and manipulating transaction data, measuring rule interestingness, and performing rule-based predictions.

**Product Bundle Identification Project in R Step By Step:**

To identify product bundles in R, you can use association rule mining algorithms such as Apriori or FP-growth. Here is a step-by-step guide to implementing the Apriori algorithm:

*step 1:*

Load the necessary libraries:

$$

library(arules)

library(arulesViz)

$$

*step 2:*

Import your data into R. This should be in the form of a transaction dataset, where each row represents a unique transaction and each column represents an item in that transaction.

Convert your data into a binary format using the transactions() function:

$$

basket <- read.transactions("yourdata.csv", sep=",")

$$

*step 3:*

Use the apriori() function to run the Apriori algorithm on your data:

$$

rules <- apriori(basket, parameter=list(support=0.005, confidence=0.2, minlen=2))

$$

Here, support is the minimum support threshold (i.e., the minimum percentage of transactions that contain a particular itemset), confidence is the minimum confidence threshold (i.e., the minimum percentage of times the consequent occurs when the antecedent occurs), and minlen is the minimum length of an itemset.

*step 4:*

Use the summary() function to view the results of the Apriori algorithm:

$$

summary(rules)

$$

This will show you the number of rules that were generated, as well as statistics such as support, confidence, and lift for each rule.

*step 5:*

Visualize the rules using the plot() function:

$$

plot(rules)

$$

This will generate a scatterplot of the rules, where each point represents a rule and the x and y axes represent the support and confidence, respectively.

*step 6:*

Optionally, you can filter the rules based on certain criteria using the subset() function:

$$

filtered\_rules <- subset(rules, subset = lift > 1.2 & confidence > 0.5)

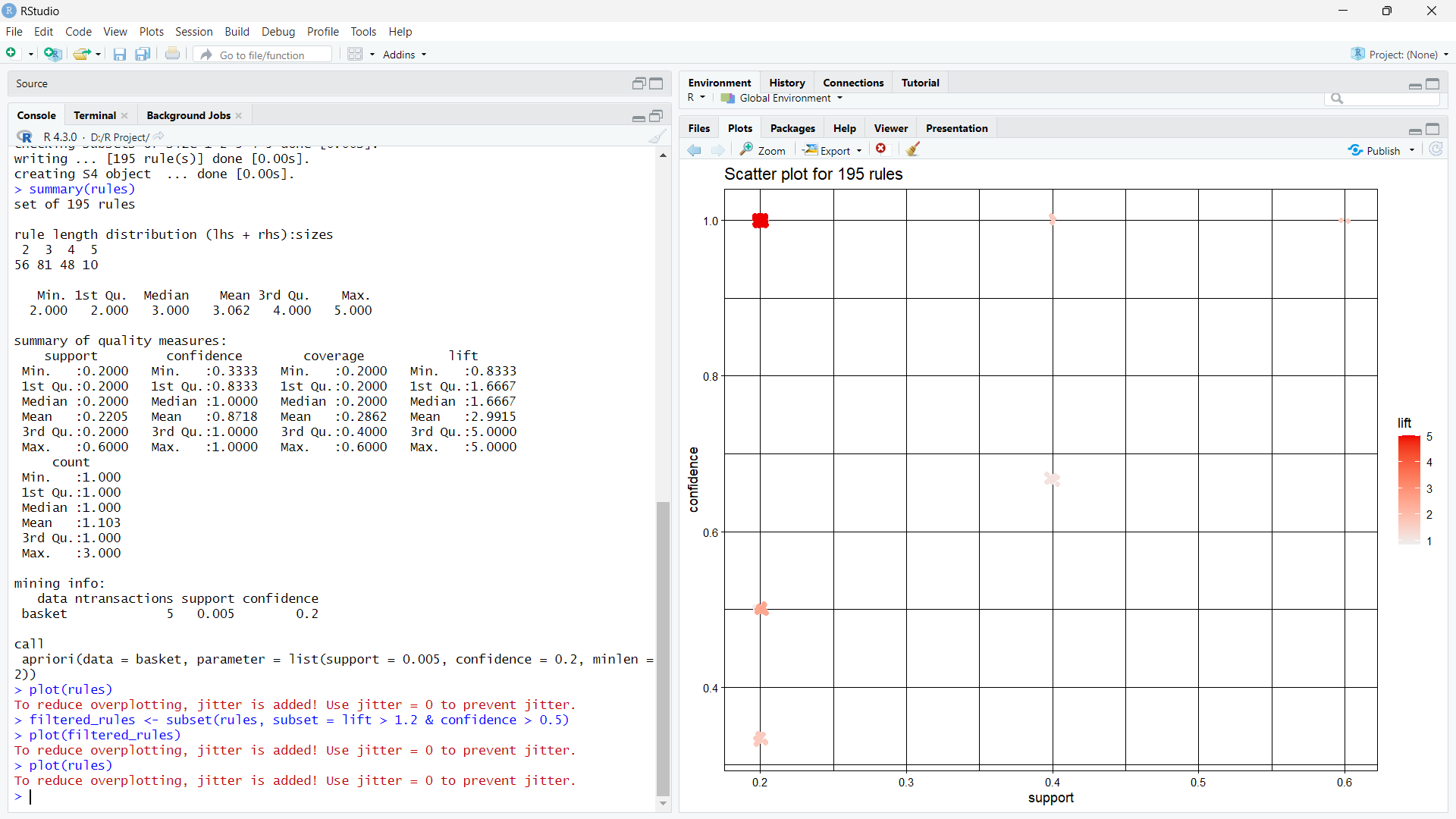
$$

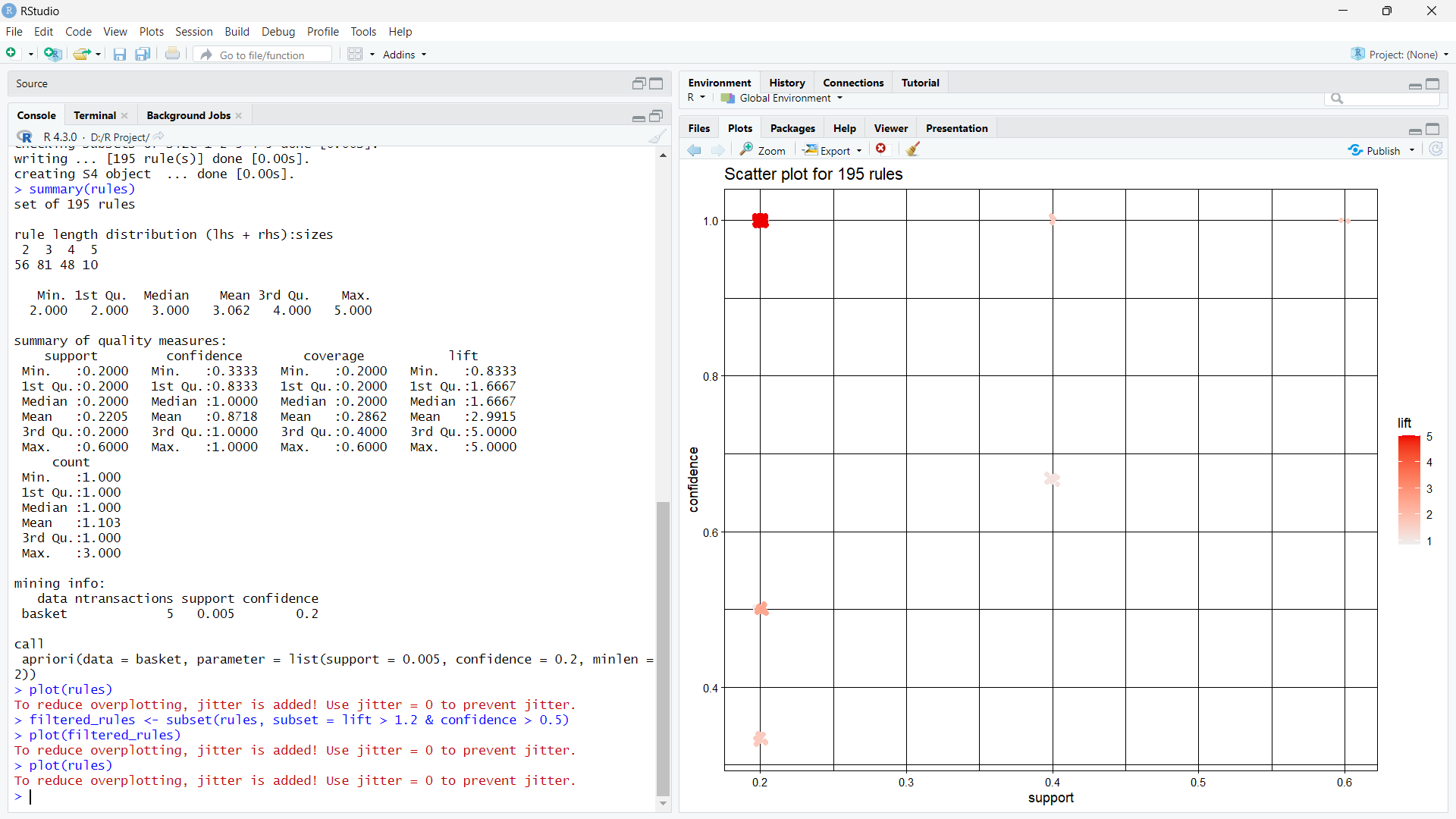
This will filter the rules based on a minimum lift and confidence threshold.

Overall, the Apriori algorithm is a powerful technique for identifying product bundles in large datasets, and it can be easily implemented in R using the arules and arulesViz packages.

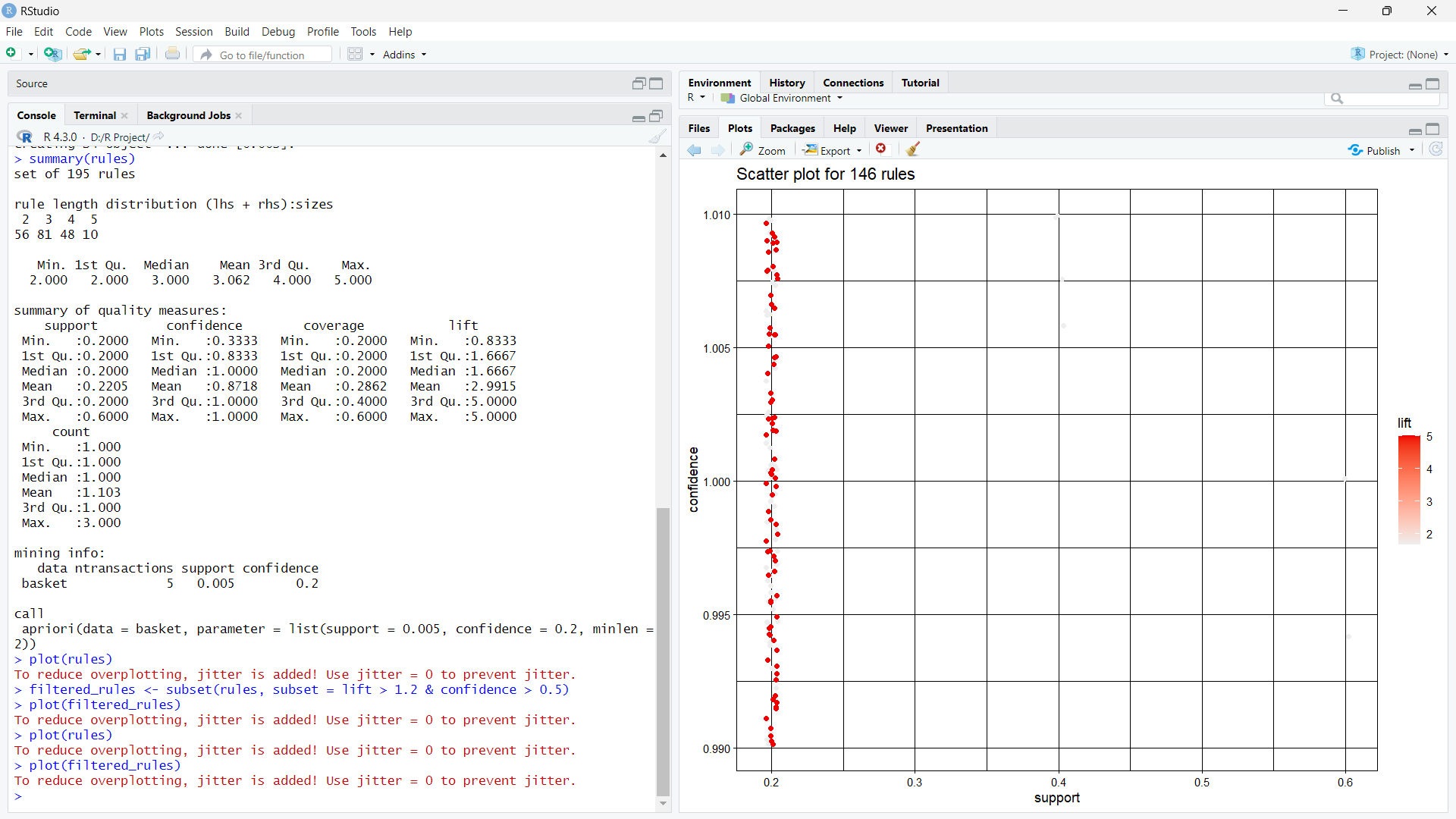


Output For the Above Code using R:

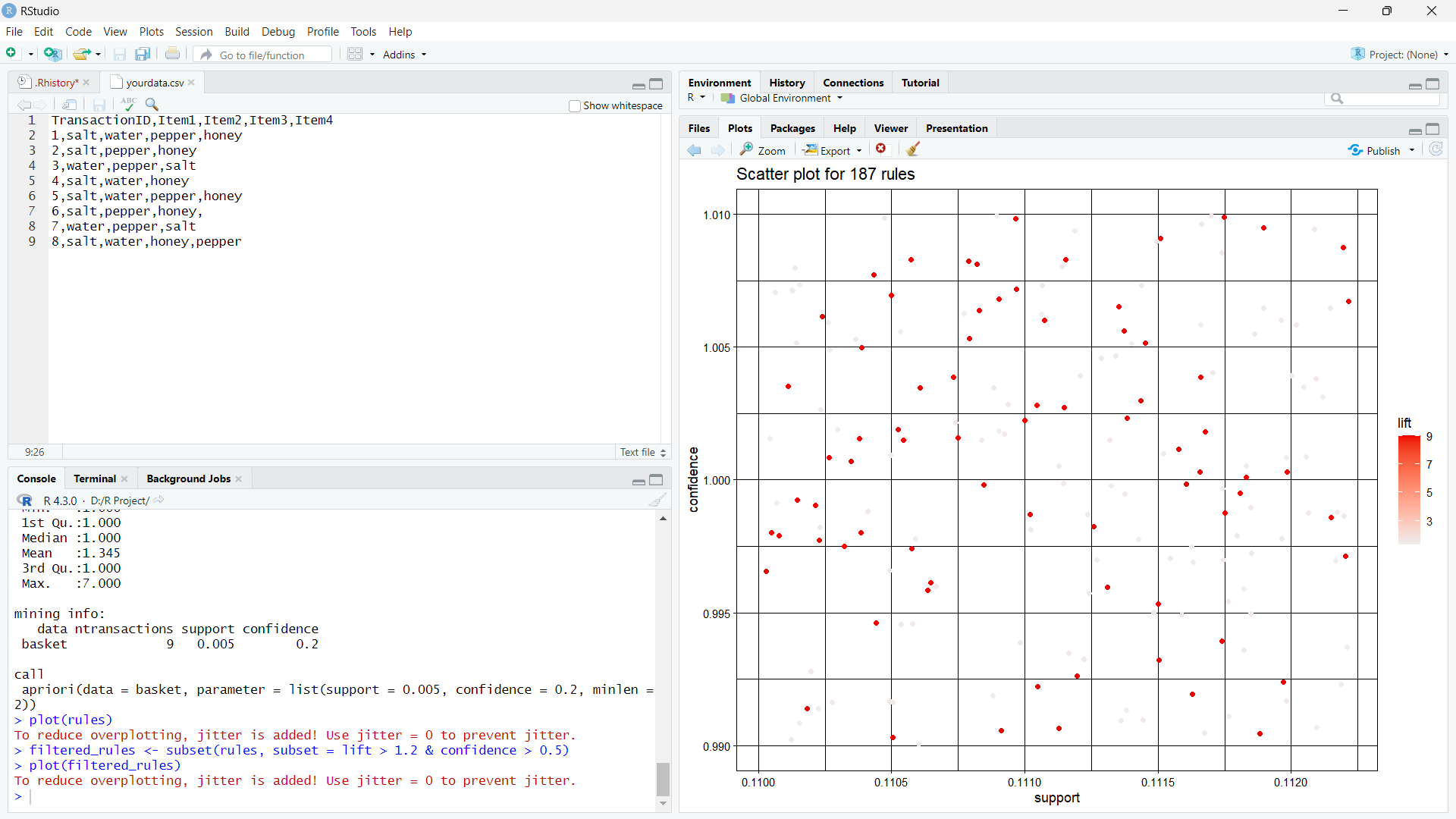




Here The Graph Represents The 195 Rules to Be Followed.



Here The Scatter Plot Represents the Plotting Of filtered\_Rules and To Follow The 146 Rules.



here we can observe that the modification in the data may result in the number of the rules to be followed by the apriori algorithm also.

intensionally we increased some data to the previous data and now, the rules of the apriori algorithm increased to 187 rules according to the data.

APPLICATIONS OF PRODUCT BUNDLE IDENTIFICATION:

1. Mix low selling items with fast-moving ones

2. Show a way out of the ‘multiple choice’ complication

3. Make the most of buy-more-pay-less bundles

4. Reward your customers with a discount

5. Give them a "reason" to buy more

6. Harness the power of personalization

7. Make it easy for them to keep buying from you

8. Get them into the DIY mode

9. Make seasonal shopping more attractive

10. Help customers discover new products

11. Make it cheaper to buy favourite.