**GROUP 320: CUSTOMER SEGMENTATION AND MARKET BASKET ANALYSIS FOR AN ONLINE RETAIL**

|  |  |  |
| --- | --- | --- |
| First Name | Last Name | Email address |
| Bhavya Sree | Bindela | bbindela@hawk.iit.edu |

Table of Contents

[**1. Introduction** 2](#_Toc57883753)

[**2. Data** 2](#_Toc57883754)

[**3. Problems and Solutions** 3](#_Toc57883755)

[**4. KDD** 3](#_Toc57883756)

[4.1. Data Processing 3](#_Toc57883757)

[4.1.1. Identify the Customers into different groups based on their behavior with the online retail. 3](#_Toc57883758)

[4.2. Data Mining Methods and Processes 13](#_Toc57883759)

[**5. Evaluations and Results** 13](#_Toc57883760)

[5.1. Evaluation Methods 13](#_Toc57883761)

[5.2. Results and Findings 13](#_Toc57883762)

[**6. Conclusions and Future Work** 13](#_Toc57883763)

[6.1. Conclusions 13](#_Toc57883764)

[6.2. Limitations 13](#_Toc57883765)

[6.3. Potential Improvements or Future Work 14](#_Toc57883766)

# **1. Introduction**

Customer Segmentation and Market analysis for an online retail non-store is to explore different types of customers and their granular purchase behavior.

**Customer segmentation** is the process of dividing the customers into different groups based on their behavior with the company. There will be different types of customers for any firm like some customers will be frequent buyers, some customers will be less frequent but buy high-value products, and so on. These are called different segments of customers. Using customer segments, the company can market their strategies to each group effectively.

**Market Basket Analysis** is a technique to identify the strength of association between pairs of products that are purchased together. This technique is based on a concept that if a person buys a product A, how likely the person will buy a product B.

Therefore, Market Basket Analysis is to uncover the product purchase behavior of customers by extracting the association rules of the products. These will be helpful to understand purchase decisions made by the customers and the company can use the association rules in their future marketing strategies.

# **2. Data**

Dataset belongs to a UK-based registered non-store online retail.

This dataset contains all the transactions that occurred between 01/12/2010 and 09/12/2011.

This dataset has been collected from UCI Machine Learning Repository.

The dataset can be found in the below link

<https://archive.ics.uci.edu/ml/datasets/online+retail>

**Format of the dataset**

Number of Instances: 541909

Number of Attributes: 8

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Description** | **Data Type** |
| InvoiceNo | Invoice number | Character |
| StockCode | Product Code | Character |
| Description | Product Name | Character |
| Quantity | Quantity of each product per transaction | Number |
| InvoiceDate | Invoice Date and Time | Number |
| UnitPrice | Price of the Product per unit | Number |
| CustomerID | Customer Number | Number |
| Country | Country Name | Character |

In this dataset, there is no target variable, as we will be doing unsupervised learning on the online retail dataset.

The whole project will be implemented in R.

# **3. Problems and Solutions**

Below are the research problems identified to solve in this project.

1. Identify the customers into different groups based on their behavior with the online retail. Identify the Loyal Customers.
2. Identify Top 10 Association rules for the products purchased in the online retail i.e., to identify the Product sets that are brought together in the online retail.

# **4. KDD**

## 4.1. Data Processing

## 4.1.1. Identify the Customers into different groups based on their behavior with the online retail.

* As a first step, loaded the data into R session.

Text

Description automatically generated

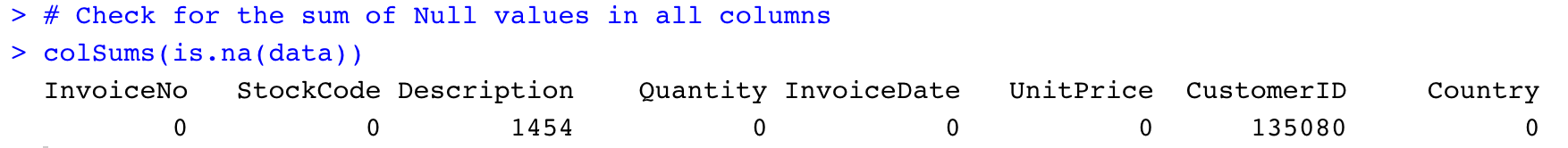
As the type of the InvoiceDate is Numerical, changed its type to date.

**Table

Description automatically generated**

(I) Pre-Processing the data

* Check for the Missing values



Data has missing values in Description and CustomerID columns.

As CustomerID is the unique ID given to every customer, we cannot replace the missing values with any valid value. Therefore, removing the records with Null values in CustomerID.

A picture containing diagram

Description automatically generated

After removing the records with missing values in CustomerID, there are no more missing values in the data.

* Check the Data Structure (Summary of the data)

Summary of the data is as follows

Table

Description automatically generated

The minimum value of Quantity is identified as -80995.00. The value of Quantity can be negative only in the return transactions. As we won’t be considering return transactions for our analysis, we will be filtering the return transactions out and consider only sales transactions.

The minimum unitPrice is identified as 0.00, which won’t be the case in the sales transactions. Therefore, we will be considering the records, whose quantity is greater than 0.

A picture containing table

Description automatically generated

(II) Transforming the data

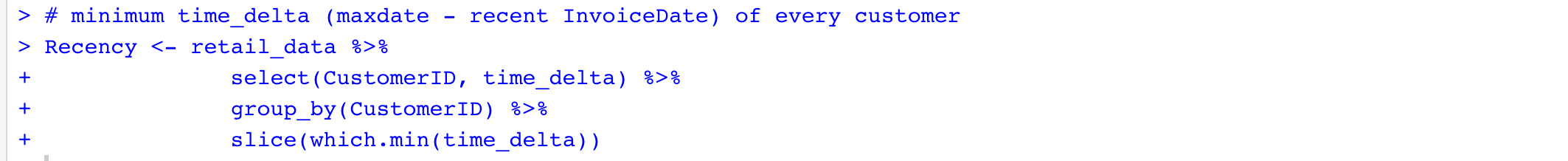
Recency, Frequency, Monetary (RFM) is a data-driven technique that is mostly used for customer segmentation. In this method, we use time from the last purchase (Recency), a total number of purchases in a period of time (Frequency) and total amount spent in a period of time (Monetary) attributes.

As a part of data Transformation, calculate Recency, Frequency, Monetary for every customer and form a new dataset.

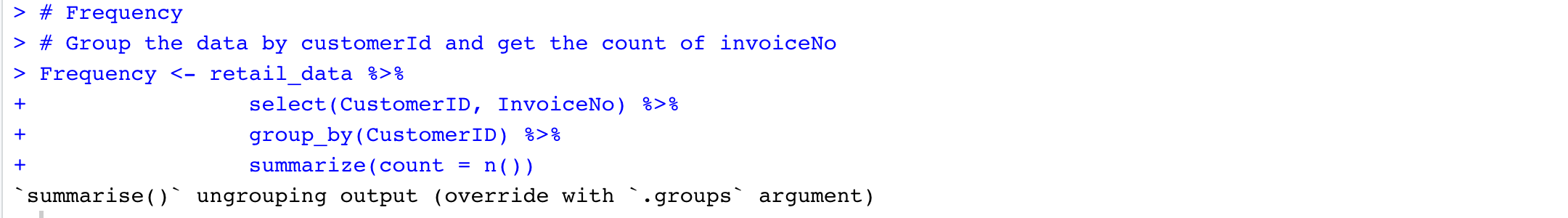
Recency is the Number of days from the last purchase. For this, first we will calculate the time difference of every Invoice i.e., difference between max InvoiceDate and InvoiceDate. The minimum time difference of every customer is taken as the Recency.

Graphical user interface, text, application, email

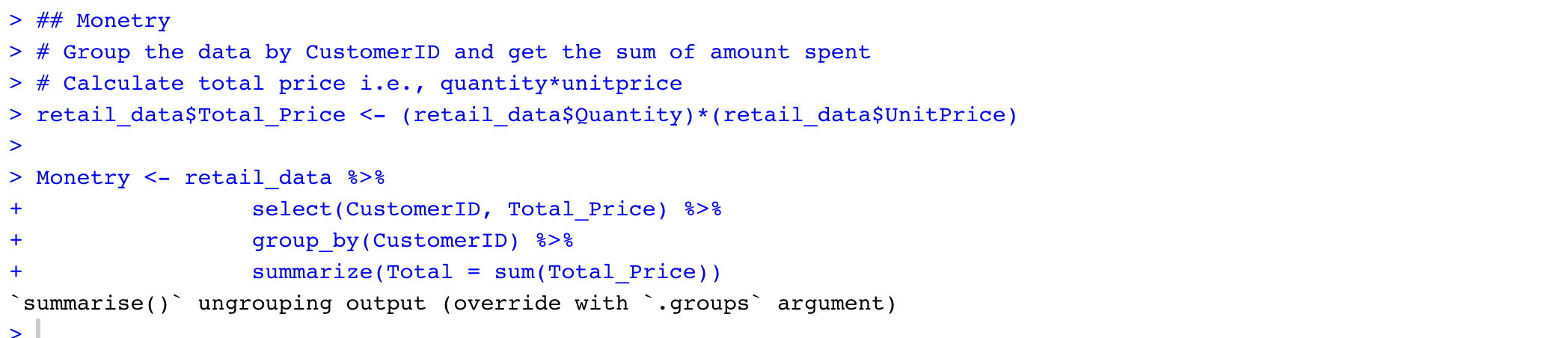
Description automatically generated



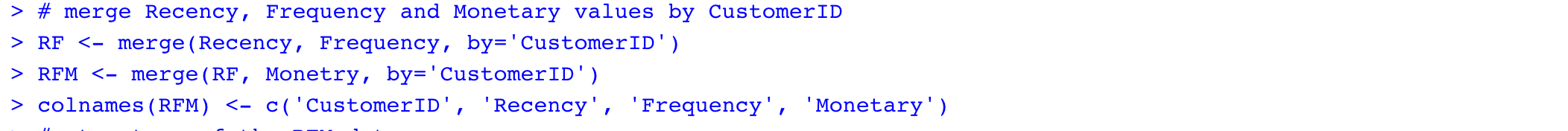
Frequency is the number of visits by the customer. To calculate this, we will group the data by CustomerId and get the count of Invoice number.



Monetary is the total amount spent by the customer. To calculate this, we will group the data by CustomerId and get the sum of amount spent (UnitPrice \* Quantity).



Now merge Recency, Frequency and Monetary values by CustomerID and form new dataset.



Structure of Summary of the RFM data is as below

A picture containing table

Description automatically generated

* Pre-processing the RFM data – Check for the outliers



A picture containing diagram

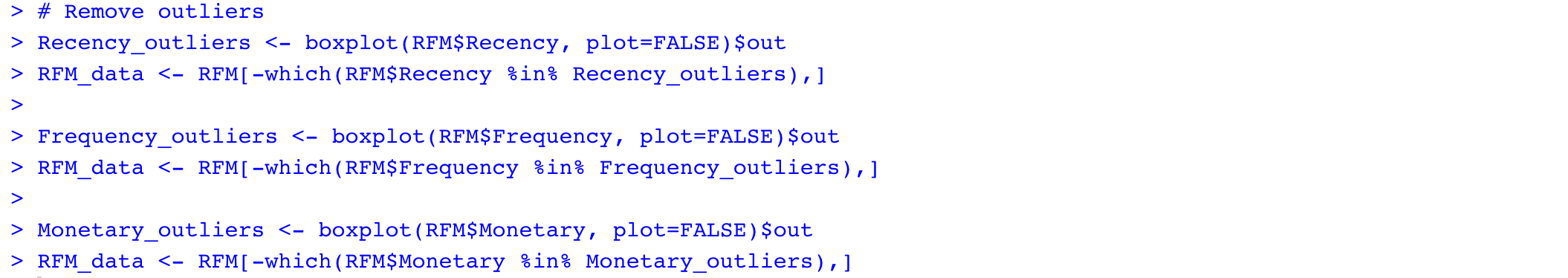
Description automatically generated

From the boxplot, it is clear that Recency, Frequency and Monetary values have outliers.

For customer segmentation, we will be using clustering techniques.

Outliers will have more effect on clustering as they can change the centroids of the clusters. Therefore, we will be removing outliers as the pre-processing for clustering.

Remove Outliers

****

Structure of the RFM data after removing the outliers is as below

Chart, scatter chart

Description automatically generated

(II) Standardizing the data

In order to compare the data and draw conclusions, we need to scale the numerical data.



The top rows of the new data (Clust\_data) is as below

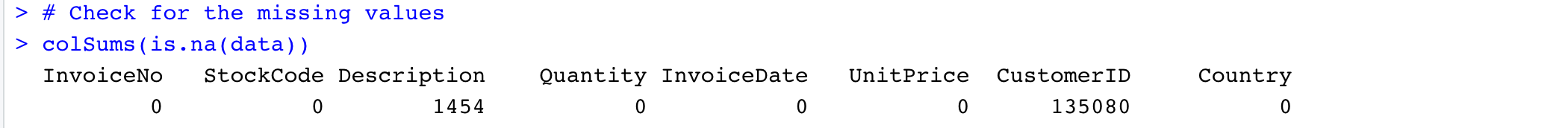
A picture containing background pattern

Description automatically generated

**4.1.2. Identify Top 10 Association rules for the products in the online retail i.e., to identify the Product sets that are frequently bought together in the online retail.**

(I) Pre-processing the data

* Check for the missing values



Data has missing values in Description and CustomerID columns.

As description (ProductName) is the main and only one attribute, we use for association rules and the number of missing values is small compared to the total size of the dataset i.e., 541909. We will be removing the records with Null values in Description.

As we won’t be using customerID for this analysis i.e., while building association rules for the products that are brought together.



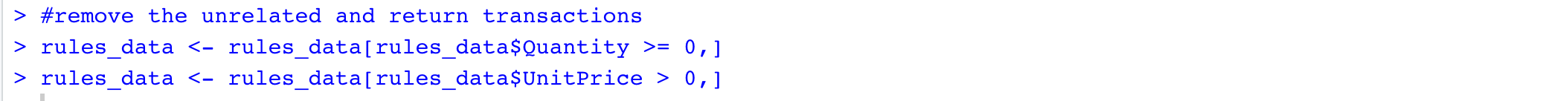
Summary of the data is as below

A picture containing table

Description automatically generated

The minimum value of Quantity is identified as -80995.00. The value of Quantity can be negative only in the return transactions. As we won’t be considering return transactions for our analysis, we will be filtering the return transactions out and consider only sales transactions.

The minimum UnitPrice is identified as 0.00, which won’t be the case in the sales transactions. Therefore, we will be considering the records, whose quantity is greater than 0.



* Check for the presence of irrelevant data in Description (ProductName) attribute.

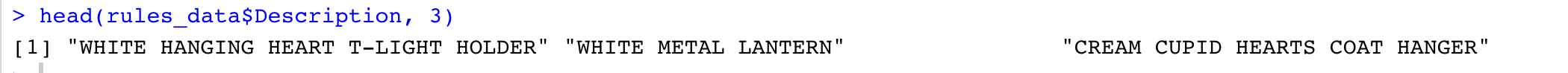
For this, as a first step, check whether the unique number of StockCodes is same as the unique number of products (Descriptions).



Generally, each stockCode represents a unique Product and a Product can have more than one StockCode as the StockCode number can be reviewed whenever the product is restoked (after the completion of all products of that StockCode in the shop).

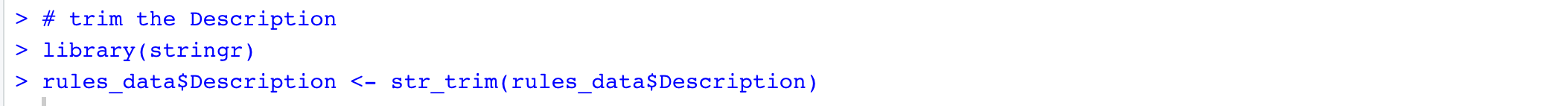
In this case, there are a greater number of products than the stockCode. Therefore, we need to check whether all the Descriptions (ProductName) are valid or not.

Some of the descriptions in the data is as below

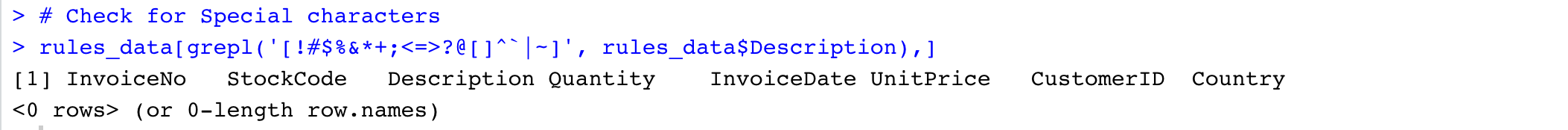


* Trim the Description (Product Name)

As Description is the type Character, we trim the data to remove any leading and trailing spaces.

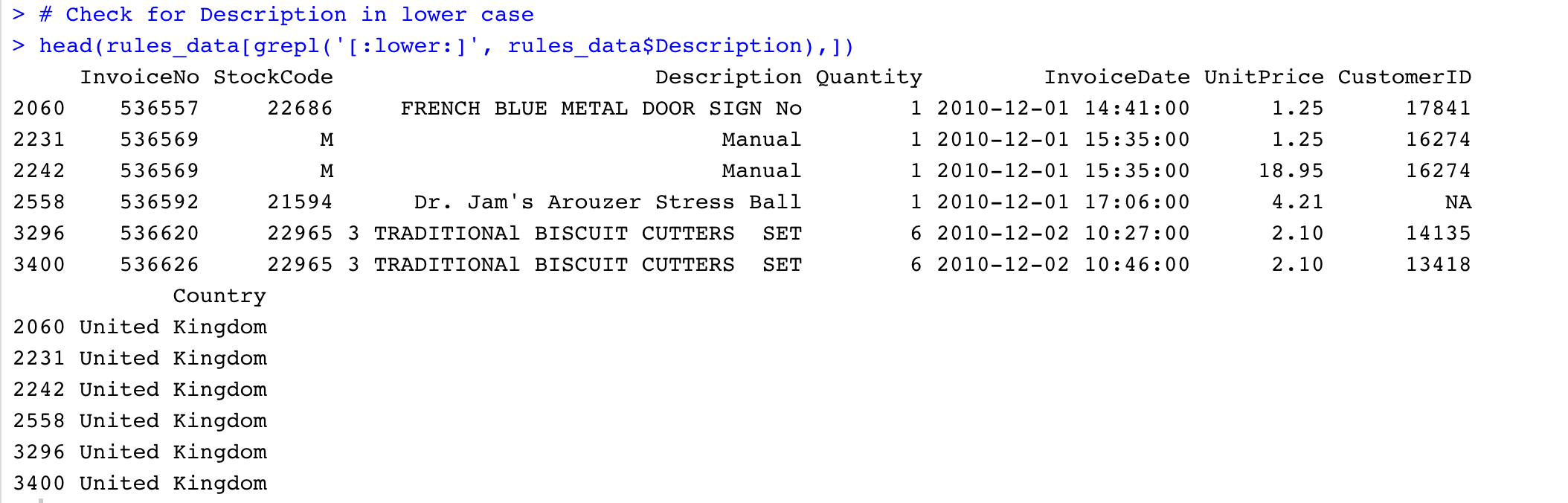


* Check for the special characters in the Description (Product Name)

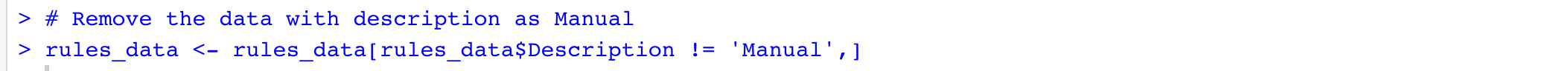


* Check for the lowercase letters in Description

It seems Descriptions are in uppercase letters. Therefore, checking if there are any lowercase letters too in the Description.

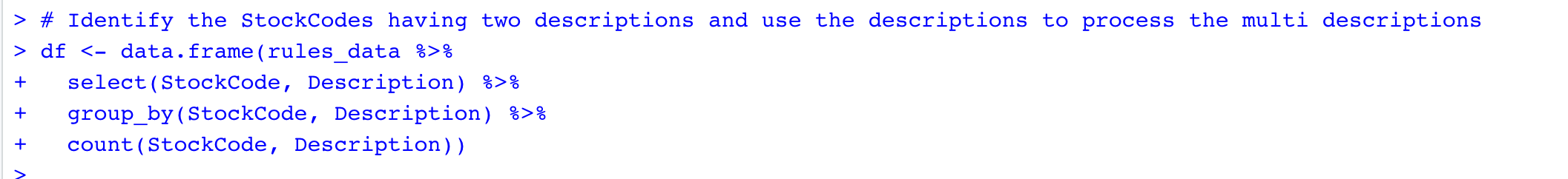


There are many rows with manual as Descriptions (just shown few for reference). As product Name manual makes no sense, we will be removing the records with Manual in Description.



* Check for the StockCodes with multiple products i.e, Descriptions

To achieve this, first we need to group the data by StockCode and Description and get the unique StockCode and Description. Then check for the duplicate StockCodes in the list of Unique StockCode and Description.



List of the Descriptions of the duplicated StockCodes is as below.

A picture containing text, newspaper

Description automatically generated

Here, we got a list of 453 Descriptions (given few for reference). When we observer this list, there are many typo errors in the product names.

For example, “16 PC CUTLERY SET PANTRY DESIGN” and “16 PIECE CUTLERY SET PANTRY DESIGN” are same with same StockCode, but these will be treated as different by the data mining techniques and these will effect the support, confidence and lift values of our association rules.

Therefore, we need to treat these typo errors if the description has the same StockCode.

Text

Description automatically generated

Treated almost 200 products which are having type errors (just a few are shown for reference, whole code will be available in the code file).

Now, the structure of the rules\_data is as below

Graphical user interface, text, application

Description automatically generated

Check for the unique number of StockCodes and unique number of Product Names.



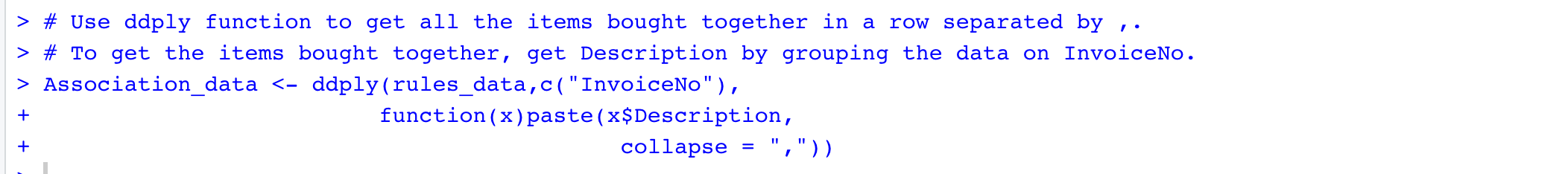
Now, the counts look valid as a product can have multiple StockCodes and as we are using product name to get the frequent item sets. This difference won’t be affecting our association rules.

(II) Transforming the data

The granularity of our dataset is one line per item.

In order to know the products that are frequently bought together, we need to have all the items bought together in a single row i.e., the granularity of the dataset should be one row per Invoice.

Therefore, to get the items that are bought in a single invoice, we need to group the data by Invoice No and concatenate the products using ‘,’.



Structure of the association data is as below

A picture containing text

Description automatically generated

Remove the column InvoiceNo, as we won’t be needing it in our process and change the column name from V1 to items.



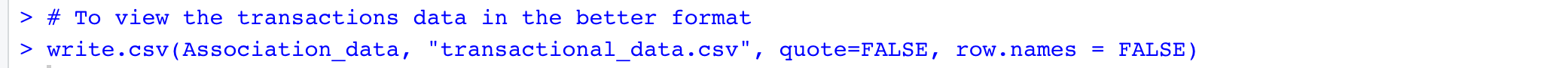
In order to build the association rules, the attributes need to be the type of factor.

A picture containing text

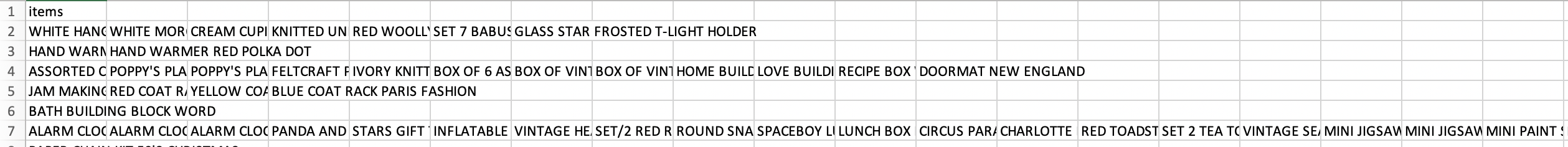
Description automatically generated

Association data contains the products that are bought together (i.e., in a single Invoice).

To view this data in better format, writing the association data to a csv file.



The csv file looks like below



Each row contains all the products that are bought in a single Invoice.

## 4.2. Data Mining Methods and Processes

**4.1.2. Identify Top 10 Product Sets that are frequently bought together Association rules for the products in the online retail i.e., to identify the Product sets that are frequently bought together in the online retail.**

We will be using Association rules mining technique in order to know the products that are frequently bought togther.

# **5. Evaluations and Results**

## 5.1. Evaluation Methods

xxxx

## 5.2. Results and Findings

xxx

# **6. Conclusions and Future Work**

## 6.1. Conclusions

xxx

## 6.2. Limitations

xxx

## 6.3. Potential Improvements or Future Work

Xxx