An-Najah National University

Department of Artificial Intelligence



Homework 1

Data Mining and Analysis

Submitted By: Aladdin Husni Odeh

Submitted To: Dr. Anas Toma

Contents

[1. Introduction 3](#_Toc151474200)

[1.1. Objectives 3](#_Toc151474201)

[1.2. Dataset overview 3](#_Toc151474202)

[2. Data Preprocessing and Cleaning 4](#_Toc151474203)

[2.1. Initial Assessment 4](#_Toc151474204)

[2.2. Cleaning Process 5](#_Toc151474205)

[2.2.1. Missing customer Ids handling 5](#_Toc151474206)

[2.2.2. Stock code and description handling 5](#_Toc151474207)

[2.2.3. Numeric values handling 6](#_Toc151474208)

[2.2.4. Outlier Detection and Treatment 7](#_Toc151474209)

[3. Data Analysis 8](#_Toc151474210)

[3.1. Statistical Analysis 8](#_Toc151474211)

[3.2. Trend Analysis 10](#_Toc151474212)

[6. References 12](#_Toc151474213)

# 1. Introduction

## 1.1. Objectives

In this report, we're going to analyze a dataset for an online store. The first step is preprocessing the data to make it ready for analysis and association rules generation.

The next step after cleaning the data is analyzing it. We want to see what the data tells us about customer trends and which products are popular in different countries.

Finally, we'll generate association rules to find out which products are often bought together. This is important because it can help the store know what products to focus on.

## 1.2. Dataset overview

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. [1]

The dataset has 541909 records with the following features:

| **Variable Name** | **Role** | **Type** | **Description** |
| --- | --- | --- | --- |
| InvoiceNo | ID | Categorical | A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation |
| StockCode | ID | Categorical | A 5-digit integral number uniquely assigned to each distinct product |
| Description | Feature | Categorical | product name |
| Quantity | Feature | Integer | the quantities of each product (item) per transaction |
| InvoiceDate | Feature | Date | the day and time when each transaction was generated |
| UnitPrice | Feature | Continuous | product price per unit |
| CustomerID | Feature | Categorical | A 5-digit integral number uniquely assigned to each customer |
| Country | Feature | Categorical | The name of the country where each customer resides |

Table : Dataset description from <https://archive.ics.uci.edu/dataset/352/online+retail>

# 2. Data Preprocessing and Cleaning

## 2.1. Initial Assessment

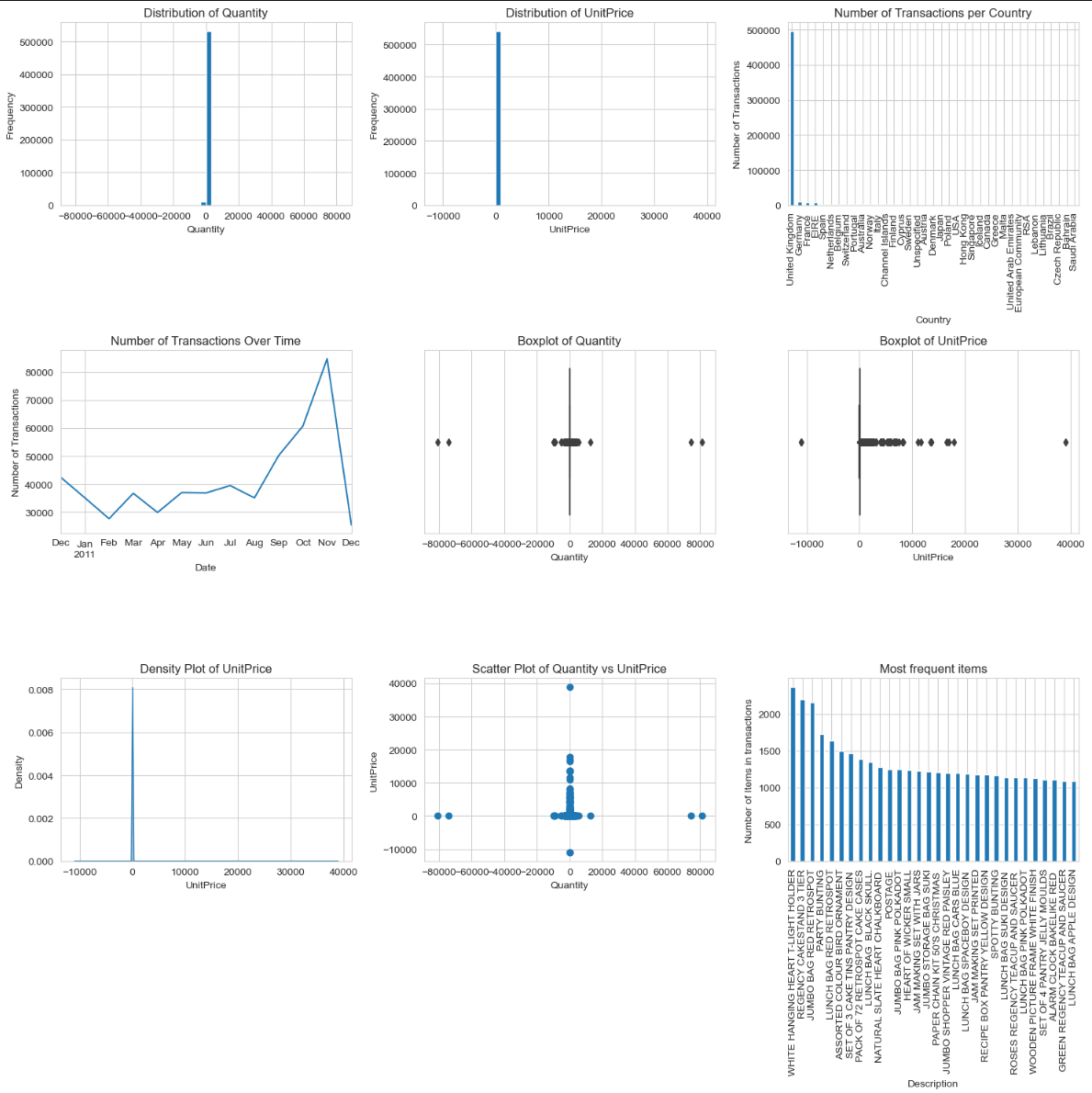


Figure : Data initial assessment

The initial step involved creating multiple histogram plots to observe the distribution of each feature. Additionally, box plots were utilized for numerical features to aid in understanding the data range and identifying any outliers.

From this preliminary examination, several key observations emerged:

* The data for both 'quantity' and 'unit price' showed widespread, indicating the likely presence of outliers. This was further supported by the insights gained from the box plot visualizations.
* A considerable amount of the data is from the United Kingdom, making it a significant region in this dataset.
* The dataset covers transactions from December 2010 to November 2011, offering a comprehensive view of retail activity over this period.
* Interestingly, 'POSTAGE' appears frequently as a product, despite not being an actual item for sale.

## 2.2. Cleaning Process

### 2.2.1. Missing customer Ids handling

The dataset contained approximately 135,000 records with missing customer IDs. An initial attempt to address this involved using forward filling, applied after grouping records by their invoice numbers. However, this approach proved ineffective, indicating that the customer ID was missing for all transactions under the same invoice.

Given that CustomerId is not a critical element for this study and considering there's no feasible method to generate its value, it was decided not to discard these records. Instead, a practical solution was adopted: filling in the missing customer IDs with a placeholder, labeled 'UnknownId'.

### 2.2.2. Stock code and description handling

|  |
| --- |
| *Number of unique descriptions 4211*  *Number of unique StockCode 4070*  *Description*  *check 146*  *? 47*  *damages 43*  *...*  *GARLAND WOODEN HAPPY EASTER 1*  *GARLAND, MAGIC GARDEN 1.8M 1* |

Output : Data Analysis - Cell #4

We began by examining the relationship between the stock code and product description. Ideally, each stock code should match only one product description. Our analysis found that this was true for most of the product descriptions (4051 of them). However, some descriptions were linked to more than one stock code. To resolve this, we decided to unify the stock codes for each description by selecting the first stock code listed.

After ensuring a one-to-one match between stock code and description, we encountered another issue: 1454 records had a stock code but no corresponding description. To address this, we created a universal map linking each stock code to its description and filled in the missing descriptions accordingly.

In the end, there were 122 records that still had missing descriptions. We chose to remove these records from our analysis, as they are essential for the accuracy of our study.

Our initial findings revealed incorrect values appearing as single-word descriptions in the data. To investigate this further, we displayed all product descriptions that were just one word long. This revealed that all single-word descriptions, except for one product labeled 'SOMBRERO', represented incorrect entries, and needed cleaning.

|  |
| --- |
| *['POSTAGE' 'CARRIAGE' 'Manual' 'amazon' '?' 'SOMBRERO' 'check' 'damages' 'DAMAGED' 'faulty' 'Found' 'found' 'counted' 'Dotcom' 'samples/damages' 'Amazon' 'showroom' 'MIA' 'Adjustment' 'damages/display' 'broken' '?lost'*  *'damages?' 'cracked' 'Damaged' 'SAMPLES' 'returned' 'damaged' 'Display' 'Missing' 'adjustment' 'adjust' 'crushed' 'samples' 'mailout' 'wet/rusty' 'damages/dotcom?' 'smashed' 'missing' 'FOUND' 'dotcom' 'FBA' 'ebay' 'Damages/samples' '?display?' '?missing' 'Crushed' 'test' '??' 'Dagamed' 'WET/MOULDY' 'mouldy' 're-adjustment' …]* |

Output : Data analysis - Cell #7

Following this step, we achieved a clean and accurate mapping of stock codes to product descriptions.

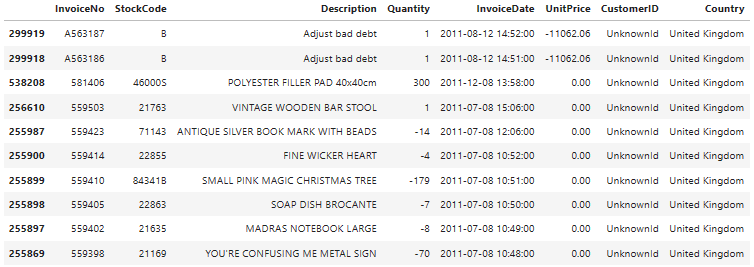
### 2.2.3. Numeric values handling

#### 2.2.3.1. Unit price preprocessing

When we first looked at the data, we noticed some extreme outliers. To understand these better, we examined the descriptions of these items. After several attempts, we discovered that all transactions with a unit price over 650 were for invalid items and could be removed from the dataset.

|  |
| --- |
| *array(['DOTCOM POSTAGE', 'AMAZON FEE', 'Bank Charges', 'Adjust bad debt',*  *'CRUK Commission'], dtype=object)* |

Output : Data analysis - Cell #9

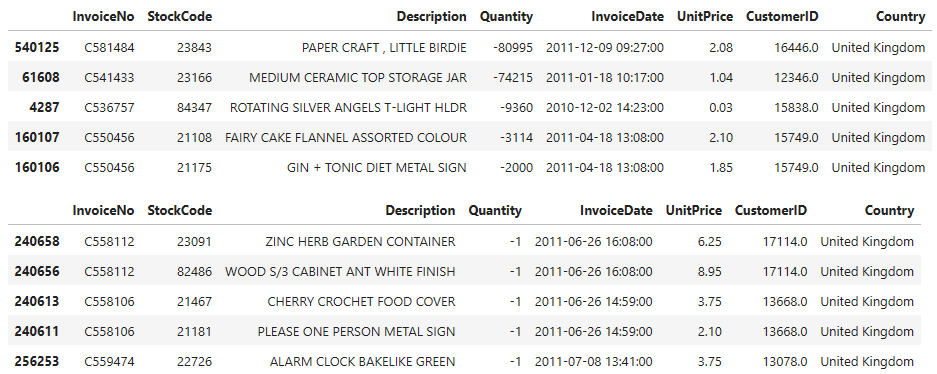
The initial assessment also highlighted some issues with prices, including negative and zero values:

Output : Data analysis - Cell #10

* Transactions with very high unit prices turned out to be for non-product entries, so we decided they should be removed.
* There were a couple of entries with high negative prices labeled as "Adjust bad debt." We chose to remove these as well.
* We also found several instances where zero prices were associated with negative quantities. These were removed for data accuracy.

#### 2.2.3.2. Quantity preprocessing

Our initial review of the data revealed many negative quantities, some of which were extremely high. To address this, we started by looking into these negative quantities:



Output : Data analysis - Cell #11

We found several instances of very high negative quantities. These appeared to be errors in the data. There were also numerous instances of quantities listed as -1. These are likely customer returns. Since our study focuses on items sold, these returns are not relevant and can be considered for removal.

### 2.2.4. Outlier Detection and Treatment

A screenshot of a graph

Description automatically generated

Figure : Outliers for numeric features after processing

|  |
| --- |
| *UnitPrice Quantiles (98th to 99.9th percentile):*  *0.980 13.29000*  *0.990 16.95000*  *0.995 21.23000*  *0.999 125.94848*  *Name: UnitPrice, dtype: float64*  *Quantity Quantiles (98th to 99.9th percentile):*  *0.980 72.0*  *0.990 100.0*  *0.995 160.0*  *0.999 432.0*  *Name: Quantity, dtype: float64* |

Output : Data analysis - Cell #15

We also noticed that both the quantity and price features in our dataset had outliers, some of which were quite significant. Here's how we addressed these:

1. **Price Outliers**: There were many outliers in price, including two extremely high values. It was important to be cautious here, as not all high prices are unreasonable or errors. By examining different percentiles, specifically the 99.9th percentile, we found that most high prices were valid. However, upon closer inspection, we identified 'DOTCOM POSTAGE' and 'AMAZON FEE' as invalid products and removed these entries.
2. **Quantity Outliers**: In the case of quantity, the high numbers mostly seemed like valid transactions for a retail setting. There were, however, two extremely high outliers in quantity. After careful consideration, we decided these were not realistic and removed them.

Ultimately, we chose to retain the outliers that seemed valid and only remove those that were clearly invalid.

# 3. Data Analysis

## 3.1. Statistical Analysis

To provide clearer visualization and aid in the decision-making process, we examined the data at the 99th percentile. This analysis was solely for a better understanding of the dataset's characteristics, and no data was removed based on this percentile. The insights gained from this perspective were instrumental in identifying valid outliers and ensuring that only irrelevant or incorrect data points were excluded.

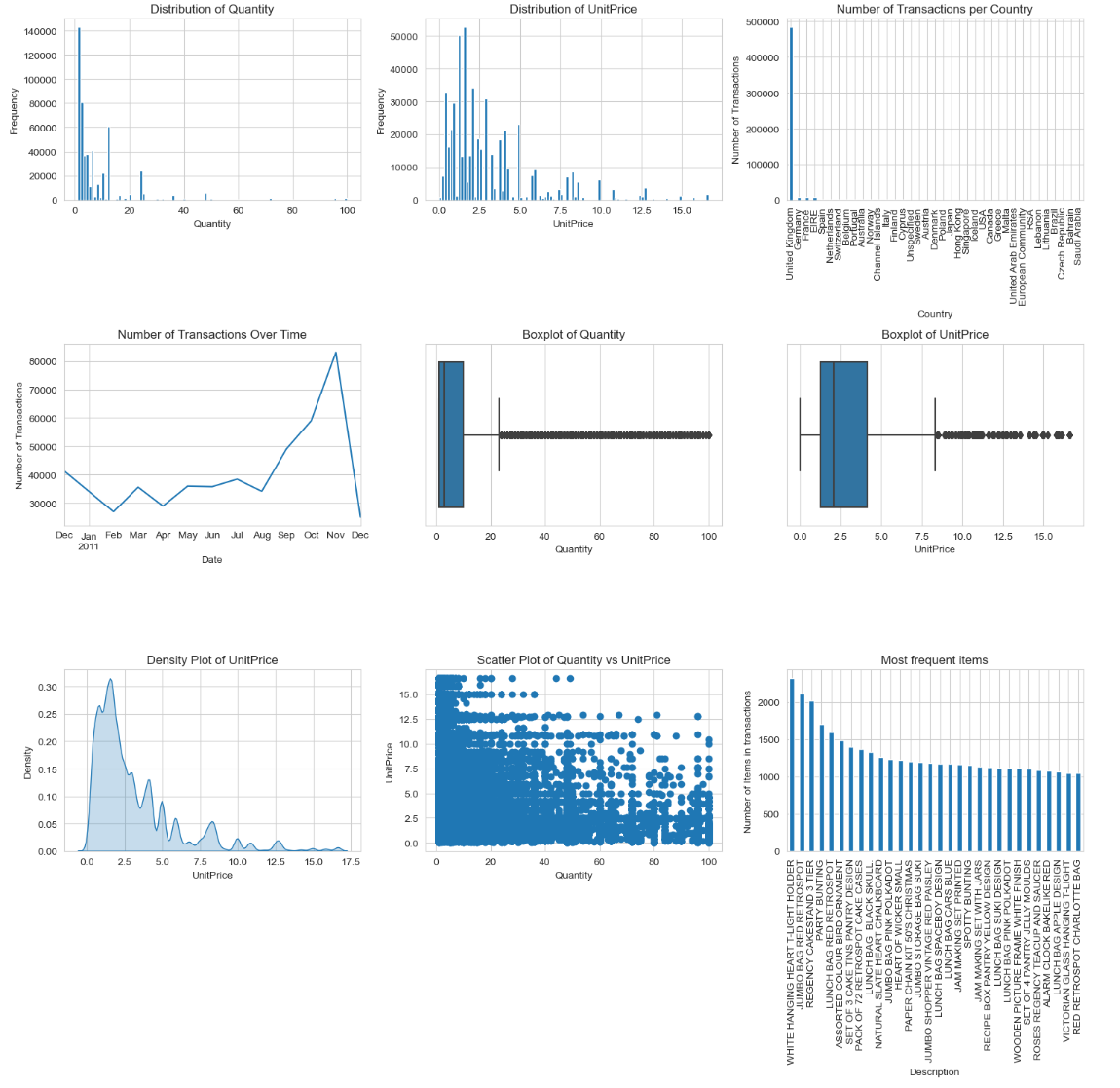


Figure : Data after preprocessing (99% percentile)

The basic plots of our dataset revealed some key initial insights:

1. The majority of transactions involved quantities ranging from 1 to 15 units per item.
2. Most transactions were for items priced between 0 and 5 units.
3. While there are some outliers in the data, these appear to be valid instances, representing either expensive products or bulk purchases by retailers.
4. A noticeable increase in the number of transactions was observed starting from August 2011. This suggests that the online retailer may have expanded into new markets during this period.

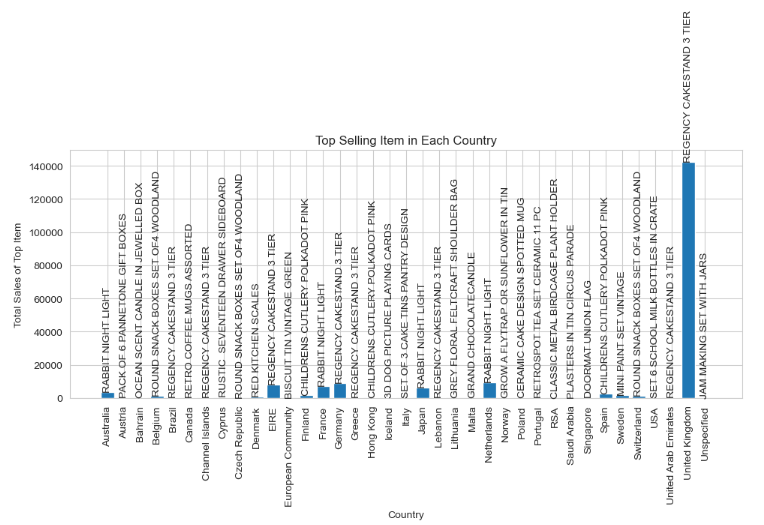


Figure : Top sellers per country

* **Product Popularity by Country**: The table shows a diverse range of top-selling products across different countries. For example, the 'REGENCY CAKESTAND 3 TIER' is the top seller in multiple countries like EIRE, Germany, and the United Kingdom, indicating its widespread appeal. However, each country has unique top sellers, such as the 'RABBIT NIGHT LIGHT' in Australia, Japan, France, and the Netherlands, suggesting varying consumer preferences.
* **Geographical Trends**: Products like 'REGENCY CAKESTAND 3 TIER' and 'RABBIT NIGHT LIGHT' appear multiple times across different countries, indicating their universal appeal. In contrast, some items are unique to specific regions, like 'PACK OF 6 PANNETONE GIFT BOXES' in Austria, showing localized preferences.
* **Market Opportunities**: Identifying top-selling items in each country can help the retailer tailor its inventory and marketing strategies to specific markets. For example, focusing on 'CHILDRENS CUTLERY POLKADOT PINK' in markets like Hong Kong, Finland, and Spain, where it's a top seller.

## 3.2. Trend Analysis

The figure presented illustrates the sales trends across the top 10 countries. A logarithmic scale is employed to enhance visualization, given the substantial differences in sales volumes between the countries. This scale choice enables a clearer focus on the trend patterns.

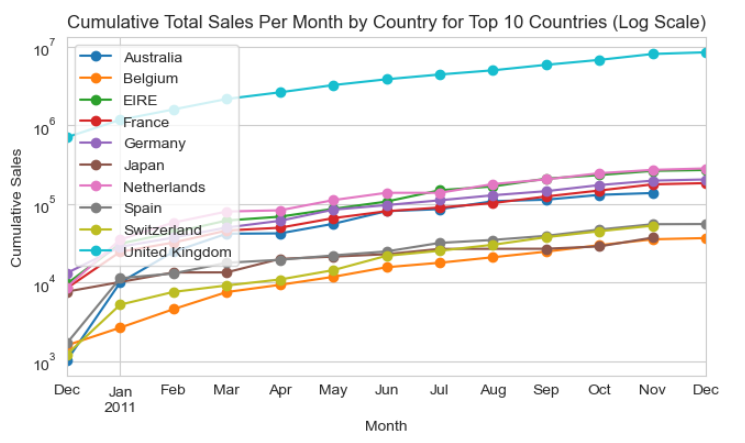


Figure : Sales trend per country

* **Growth Over Time**: All countries show an upward trend in sales throughout the year. This suggests that the store's business is growing over time.
* **United Kingdom Dominance**: The United Kingdom has much higher sales than any other country, showing it's the biggest market for the store.
* **Seasonal Patterns**: There seems to be a general increase in sales for all countries starting around August and continuing through the end of the year, which could be due to seasonal shopping increases, like back-to-school or holiday sales.
* **Market Differences**: While all countries show growth, the rate is different. For example, Japan and Australia have a steeper increase compared to countries like Belgium and Switzerland, indicating faster growth.

A graph of sales and sales

Description automatically generated with medium confidence

Figure : Customer trends per month

Taking into account that the data extends only until December 9th, the chart suggests the following trends for customer behavior:

* **Retained Customers**: There is a consistent growth in retained customers, indicating good customer loyalty and suggesting that existing customers are continuing to engage with the business throughout the year.
* **New Customers**: The number of new customers shows a significant increase in November, which might be related to holiday promotions attracting first-time buyers. However, the data does not cover the whole of December, so the expected further increase in new customers due to the holiday season is not fully captured.
* **Total Customers**: The total number of customers increases in November, likely influenced by both returning and new customers. The data suggests a peak in customer activity during this period, which may continue to rise if the data for the entire month of December were available.

Overall, the chart indicates a healthy customer base with a notable rise in customer engagement towards the end of the year, which could have potentially been even higher if the data included the entire holiday shopping season.

# 6. References

[1] Online Retail. (2015). UCI Machine Learning Repository. https://doi.org/10.24432/C5BW33.