University of Pisa

Department of Computer Science

Data Mining
Group 18

Supermarket Analysis

Authors:

Donato Meoli
Enrico D'Arco
Luigi Quarantiello

November 5, 2020

Contents

1 Introduction

This report shows the analysis we made over the shopping sessions registered by a supermarket; the goal is to understand the customers' behavior and to define different customer profiles.

Don't know if we want to keep this introduction

2 Data Understanding

The dataset contains informations about the purchases of each customer of the supermarket. The dataset contains **471910** rows, each of one divided into **8** columns; each entry represents the purchase of a product by a customer, and it comprehends also informations about the date of the transaction, the price of the product and the quantity purchased.
A key information we extracted from the dataset, based on the products descriptions and the frequency of purchases per customer, is that the supermarket in question is not a food store, but instead a grocery store, that sells in particular fast-moving consumer goods.

2.1 Data Semantics

In this section, we describe the semantic of each attribute, providing also some statistics about them.

BasketID
A code that identifies a shopping session for a customer. It remains the same for all the rows that indicate different products inside a single session.
We have **22190** distinct Basket IDs.
We define *good* **BasketID** as a six digit code, instead a *bad* **BasketID** is composed by a character, *C* or *A*, followed by a six digit code.

BasketDate
The date and time of a transaction. Basket Dates that correspond to the same Basket ID have the same value.
The dates range from 2010 to 2011, while the times range from 6 to 21.

Sale
A numeric attribute that indicates the price of a product. Since there is no currency associated with the sale price, we assume that it is always the same.

CustomerID
A code that identifies a customer. It is a five-digit code.
We have **4372** unique customer IDs, with some *NULL* values.

CustomerCountry
A categorical attribute that indicates the country of origin of each customer.
We have **37** different countries.

ProdID
A code that identifies a product.
It most cases, it is a five-digit string; sometimes, it has a letter at the end, while other times it is a string of characters.

We have **3953** distinct Product IDs. We define *good* **ProdID** as a five digit code, possibly followed by a variable number of characters; everything else, which are strings followed by some numbers, is labelled as a *bad* **ProdID**

ProdDescr
A string with a description of a product.
We have **4097** unique Product Descriptions, with some *NULL* values.

Qta
A numeric attribute that represents the purchased quantity for each product.

2.2 Assessing Data Quality

Now, we describe a deeper analysis to assess the quality of data.
First, we checked for duplicate rows, and we found a total of **5232** duplicate records. Since each row represents a single purchase from a customer in a specific date, we have considered these duplicates as errors, and so we decided to drop them. After this correction, we have a dataset consisting of **466678** records.

Then, we focused on the missing values; in the dataset, we have only two columns with some *NULL* values; in particular, we have **CustomerID**, with **65073** null objects, and **ProdDescr**, with **753**.

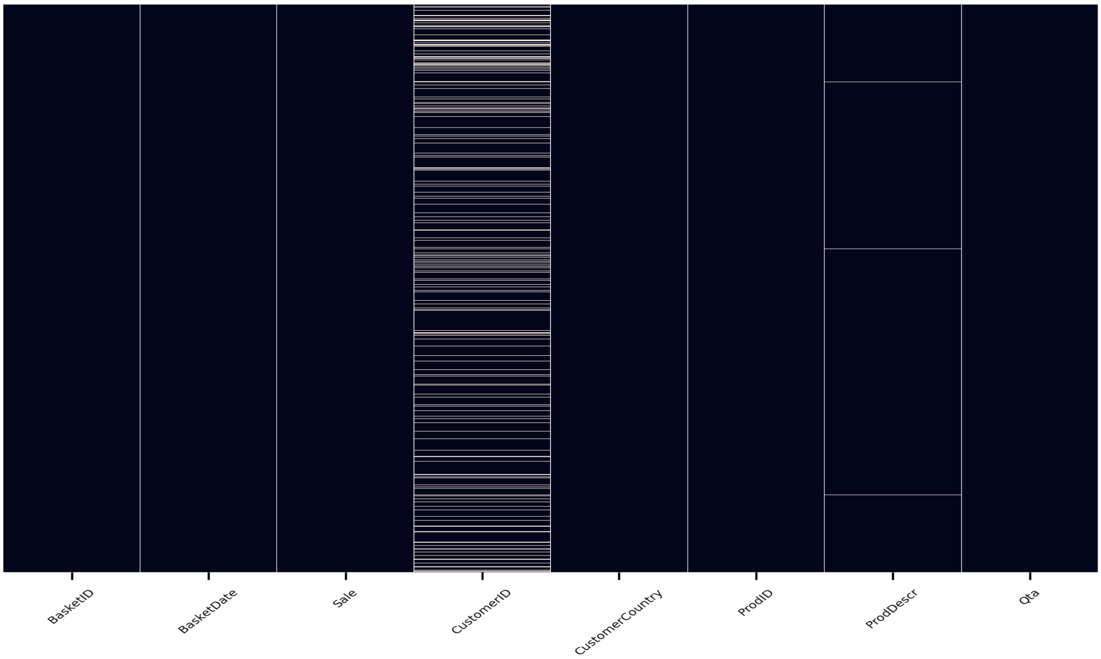


Figure 1: Missing values

We checked attributes like **BasketDate**, **Sale** and **Qta** to see if there were some syntactic errors, but we did not find any of them.

Afterwards, we checked the *semantic accuracy* of the entries in the dataset, to see if some value was not coherent with the logic of its attribute.
In particular, we found:

- 2 records with a negative value for **Sale**;
- Some rows with "*strange*" **ProdDescr**, not referring to a particular product
  - Some of them describe the product's conditions, *e.g.* '*Damaged*', '*wet rusty*', '*Unsaleable destroyed*'. We noticed that these records have all *NULL* **CustomerID**, so we consider these objects not as actual purchases, but instead as internal operations of the supermarket staff;
  - Other ones are referring to some online buying, *e.g.* '*amazon*', '*ebay*', '*dotcom sales*', but we couldn't be able to understand the real meaning of these objects, since they are very few in the dataset;
  - Others, like '*Manual*' and '*Next Day Carriage*', are related to some particular situations, that are not interesting in relation to our analysis;
- Some rows have negative values for **Qta**; here, we discover an interesting pattern: in fact, all the rows with a **BasketID** starting with **C**, have a negative **Qta**. We interpret this records as refunds.

For what regards the outliers, we used a box plot to visualize the two continuous attributes that we have, that are **Sale** and **Qta**.
From these plots, we found that, for both attributes, the box is very flat, meaning that the vast majority of the values fall in a small range. Nevertheless, there are several outliers, someone with a value really far away from the median; these values could represent an issue for the analysis we will perform. Plus, here we can notice the negative values we already found during the semantic analysis.

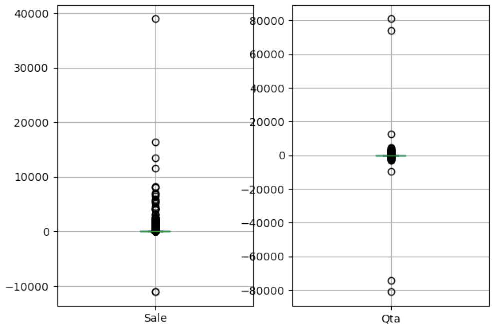


Figure 2: Box Plot for outliers detection

2.3 Variables Transformations

We chosen to categorize the attribute **BasketDate**, by first splitting the date information and the time information, and then

- for the time, we decided to divide the day hours into 5 categories:
  - *Early morning*, from 6 to 9;
  - *Morning*, from 9 to 12;
  - *Lunch time*, from 12 to 15;
  - *Evening*, from 15 to 18;
  - *Late evening*, from 18 to 21.
- for the date, we chosen to keep the week of the transaction.

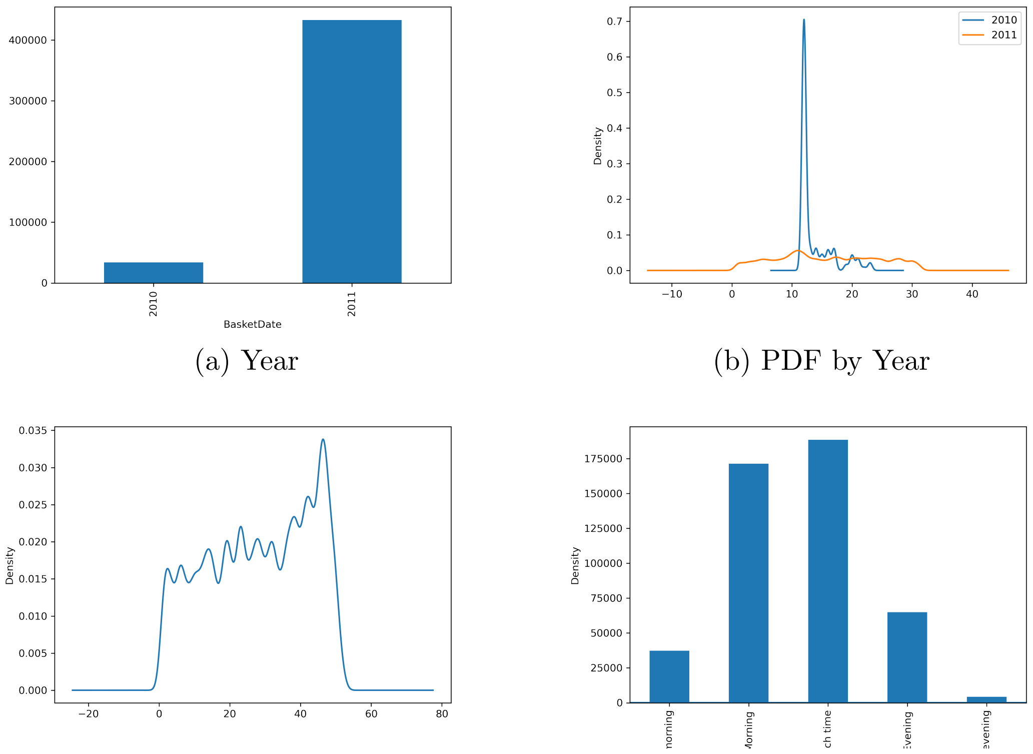
The rationale behind these choices is that we are more interested in the period in which a transaction occurs, rather that its precise date and time, to be able to cluster customers with a similar shopping behavior; furthermore, since the dataset is not from a grocery store, the customers with several shopping sessions per day or per week are really rare, and so we decided to use a broader partitioning.
After these categorization, we ended introducing **DayTime** and **Week**.

We also decided to create the attribute **TotalPrice**, which is simply the product of **Sale** and **Qta**, and represents the total amount spent by a customer for each record. We made this choice mainly to have a clearer look to the dataset, emphasizing an important information that was not explicit in the original table, and also to simplify the extraction of some features and statistics.

2.4 Variables Distribution

In this section, we will study the distribution of various attributes, by plotting some interesting properties about them.

First, we analyse the **BasketDate**.
From the Figure 3a, we can see that the attribute is highly unbalanced; in fact, we have that almost all the records are related to transaction of the 2011, while the objects from 2010 are very few. Indeed, the rows of 2011 represent about the 93% of the whole dataset.
Furthermore, from the Figure 3b, that represents an estimation of the probability density function of **BasketDate** divided by year, we can appreciate the different distributions for the two years.
In fact, for the 2010, we have a very uneven plot, which indicates that the records are not uniformly distributed with respect to the days in a month. That is justified



(a) Year

(b) PDF by Year

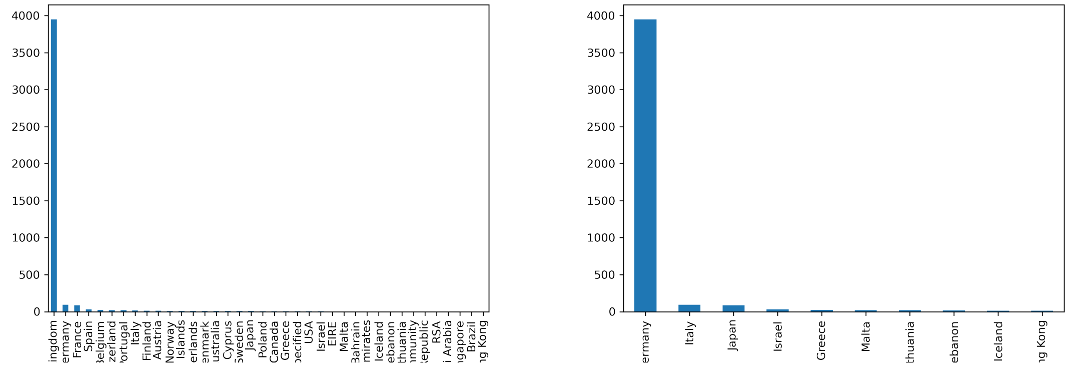
(c) PDF by Week

(d) Day Time

Figure 3: BasketDate Distributions

by the fact that, for the majority of the months in 2010, there were registered only transactions from a single day; this day, that is the 12th, is the value for which the plot shows the peak.
On the other hand, the distribution for the 2011 is much more homogeneous, meaning that the transactions were registered for most days in the months of that year.
Another interesting distributions are plotted in Figure 3c and 3d.
In the first one, we can see that the last weeks of the year are the one with more purchases; that is consistent with our knowledge, since those are the weeks closest to Christmas time, that typically represents a great period of shopping.
In the second one, the focus is instead on the hours in a day; we found that, unsurprisingly, the most popular hourly range goes from 9 to 18.

Some others statistics can be visualized for the attribute **CustomerCountry**.
In Figure 4a, we can see the distribution of the CustomerID with respect to the country; from the plot, it is clear that the most frequent country is the United Kingdom, that is present in about the 90% of the rows.
To see also some properties of the other countries, in Figure 4b, we plotted the



(a)

(b)

Figure 4: CustomerCountry Distributions

most frequent countries, excluding the UK. Here, we can see that the second most frequent country is Germany, while the other ones are almost irrelevant, since they are in very few objects.

Finally, we see some informations about the correlation of the attributes, to see if some of them are redundant.
From the Figure 5, we can see that almost all the attributes are uncorrelated, except for **TotalPrice**, that shows ah high correlation with **Qta**; that follows what we expected, since **TotalPrice** is, by construction, dependent on **Qta**. So, we conclude that all the original columns are independent, and so we don't need to perform any further manipulation.

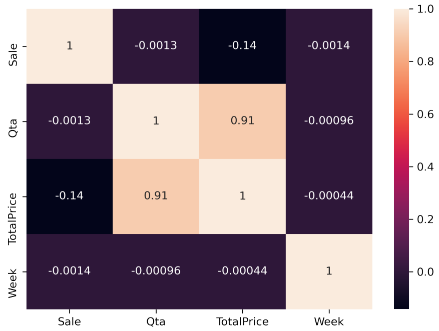


Figure 5: Correlation Matrix