Walmart Project: Use market basket analysis to classify shopping trips

Introduction

With the advancement of ecommerce, it is challenging for retail stores like Walmart to maintain or increase their sales, as these days people prefer to get the products with just one touch on their smart phones. For retail sales to increase and to provide better services to their customers, it becomes necessary for the retail stores to find and understand the pattern of the goods being purchased by the customers based on the past customer experience and the goods being bought in the past. For an instance, Milk, eggs, bread etc. are the products which are consumed on daily basis, so, whenever we enter any retail store like Walmart, we find that these products are kept far away from the entrance, so that whenever a customer wants to purchase these goods, they must walk through all the sections to get these products. The strategy behind this idea is that, until the customer reaches the end section, on their way he/she can put the items they see and feel important in their carts and eventually end up buying more products than what they had originally intended to buy. To come up with such strategies, we perform the market basket analysis on the historical data which contained the details about the transactions like, different products in an itemset belonging to their specific departments being purchased on a day of the week. Based on this analysis conclusions are made, which help in decision-making for the retailers to stock the products, discontinue or reduce the number of products which consumers do not prefer to buy or increase the quantity of those products which people often buy. Market basket analysis helps in associating the goods that go together by finding the support, confidence and lift of these respective products. With the help of these parameters, the association rules can be formed, which helps the retail stores like Walmart to make the shopping experience better for the customers thereby gaining profit by increasing their sales. In this project we use this analysis to classify the trips of the customers. These trips are classified into 38 distinct categories. The trips range from a last-minute trip, a weekend trip to a grocery shopping trip.

Dataset

We have used the dataset from Kaggle: <https://www.kaggle.com/c/walmart-recruiting-trip-type-classification>. There were two datasets available for the analysis, training set and test dataset. Training data had 647054 observations and 7 attributes, in which one of the attributes was the labelled output data, ‘TripType’, which was used to classify the shopping trips. Test data had 653646 observations and 6 attributes.

The 7 attributes are,

* Visit Number- An Id corresponding to a single trip by a single customer, basically a transaction Id.
* Weekday-The day of the week on which the trip was made to the store.
* UPC- The UPC number of the product purchased, which is a unique Id of the product.
* ScanCount - The number of the given item that was purchased. A negative value indicated that the product was returned.
* DepartmentDescription- A high-level description of the department to which the item belonged. There are total 69 distinct Departments.
* FinelineNumber - A more refined category for each of the products, created by Walmart.
* TripType - A categorical id representing the type of shopping trip the customer made.

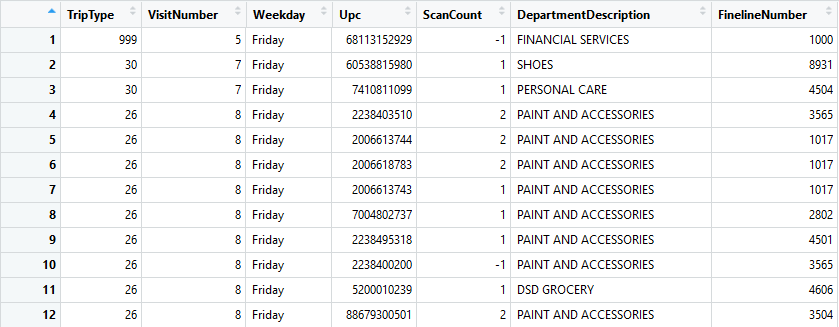


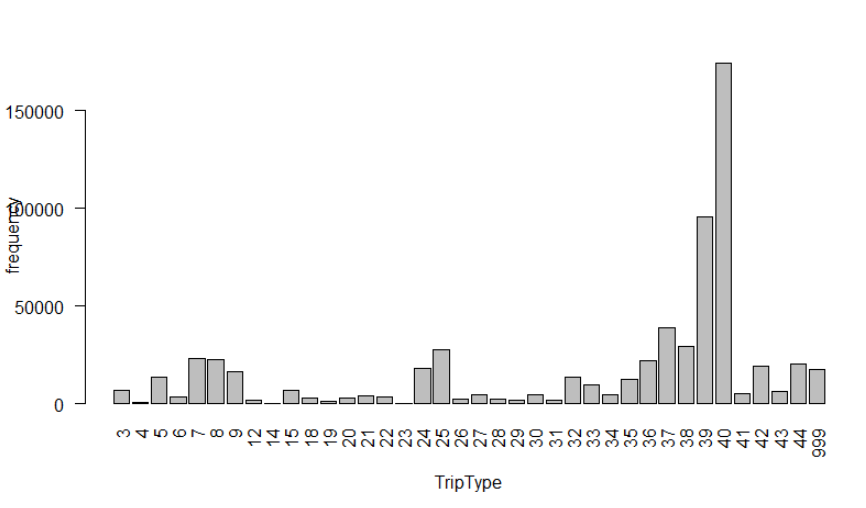
Figure 1: This figure shows the first 12 sample observations of our dataset

Data Visualization

We have plotted a histogram of the frequency/count of each trip type.

#R code for plotting frequency of each Trip Type

barplot(table(train\_data$TripType), las=2,xlab='TripType', ylab='frequency')

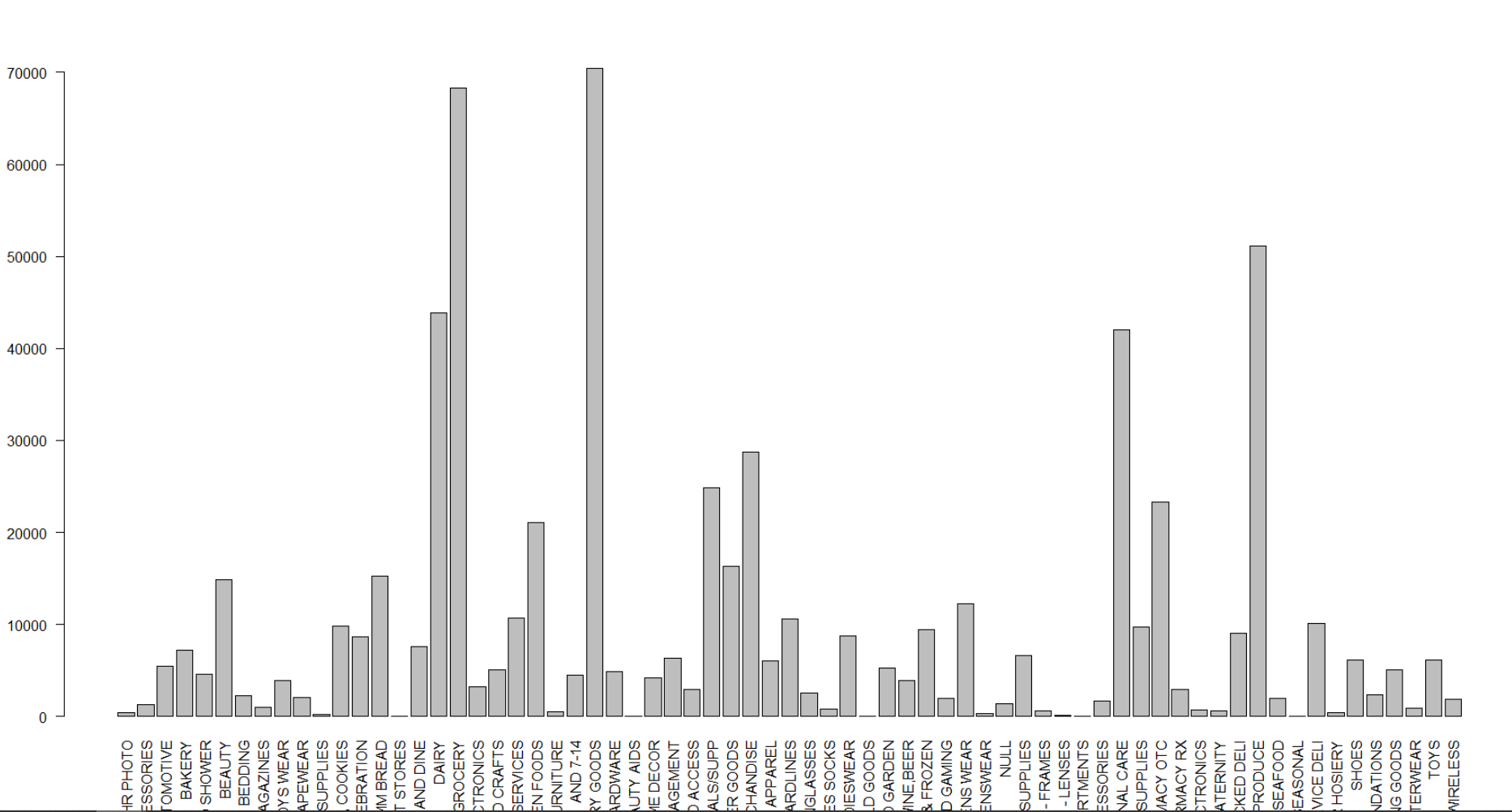


We see from the above plot that maximum number of people made a trip of type 40 and very few people made a trip of type 4 or 14 or 23.

We have plotted a histogram of the frequency/count of each DepartmentDescription.

# R code for plotting frequency of each Department Description

barplot(table(train\_data$DepartmentDescription), las=2,xlab='DepartmentDescription', ylab='frequency')



We observe that, "GROCERY" and "GROCERY AND DRY GOODS" have the highest frequency counts.