DSC 498 Project Report

Shopping Behavior Analysis with Data Analytics Tools

Nicholas Cen

Introduction:

We all buy things from stores. Sometimes we go to stores near our homes, and sometimes we go to stores farther away. There can be many factors that influence which stores we go to, such as the selection of merchandises, the prices, or whether there is something else going on at or near to the store that we are also interested in, such as restaurants, barbershops, shopping malls, movie theaters, to name just a few.

Looking from the store perspective, it's important to identify, attract, and retain the most profitable customers. Customer analytics is powered by consumer behavioral data that is examined by a company running a store or a chain of stores. Collecting this data involves an activity whereby customer's buying behavior is analyzed through data collections, with the results being used by marketing professionals looking to increase revenues.

Shopping behavior analysis refers to the process of gathering data on the actions of buyers in a retail environment, and then using that data to identify their buying patterns and preferences. Some of the factors that are considered during the analysis can include (Sightcorp, 2019):

- How do shoppers navigate the store?
- Which products draw the shoppers' attention?
- Is there a noticeable demographic trend? E.g. Do working-age people prefer to shop in the evenings? Or in the mornings on the way to work?

This type of analysis enables marketing, sales, and logistics staff to predict market trends, which is useful when making setting up promotions, designing store layout, and buying decisions, for example (Sightcorp, 2020).

The goal of customer analytics is to create a single, accurate view of the customer for better decision making on how to acquire and retain customers.

Accurate targeting of a customer is very important for a retail business. This is the case of a chain store, for example, Shaws/Star Market, Stop & Shop, Market Basket, which has locations throughout the region, offering comparable selection of goods and at the same prices. Customers shopping at one shop instead of others not because of price differences.

Apart from tailoring the shopping experience of loyal customers, customer loyalty programs can be used to monitor the purchase behavior after the sale is made. For example, if the customer has made an additional purchase (complementing the previous purchase) or has made a product return.

Whether it is through free shipping or freebies, customer loyalty is an excellent yardstick for observing customer behavior. Customer loyalty metrics can help a company track the buying patterns of each customer and understand the merchandise preferred by each demographic group (Countants, 2020).

Many retailers nowadays have accumulated huge quantities of data to focus analytics on to understand customer choice. For example, Tesco, Co-op, and Metro, along with other big European retailers, have as part of their loyalty/reward programs built internal databases that are extremely comprehensive and which combine demographics with transactional data. Every time a customer swipes their loyalty card at the checkout, information is captured about the products they bought, the frequency of buying, and their responsiveness to price and promotions, among others (Uncles, 2010; Sigurdsson, 2015). In fact, loyalty programs generate a ton of data based on shopping patterns at the stores. Retailers that gather that information can then act upon the insight it provides after some data crunching. Such actions include increasing in-store promotions, changing the store layout, adjusting pricing, targeting promotions etc. (Sightcorp, 2019).

Furthermore, the possession of the purchase history of each loyalty card member opens up the potential for more one-to-one marketing and targeted promotions. Customers showing no purchasing behavior in a particular category may be stimulated to make a purchase through a targeted coupon with a considerable discount. The British retailer Tesco, for instance, has for a long time focused on increasing sales to regular customers and enhancing loyalty with targeted coupon offers delivered through its Clubcard program. According to Davenport et al (Davenport, 2011), Tesco’s analysis of purchase information on their Clubcard members has provided insight for more sophisticated targeted coupons. As a result, Clubcard members buying diapers for the first time not only get coupons for baby wipes and toys, but also for beer.

“Data analysis revealed that new fathers tend to buy more beers, because they are spending less time at the pub” (Davenport et al., 2011, p. 4). Tesco’s aim has not only been to expand the range of customer purchases through targeted coupons, but also to target regular customers with deals on products they usually buy. Similarly, the German retailer group Metro has its Payback loyalty program (Gedenk, 2010) and the Norwegian retailer group Co-op has its Co-op

Member Program – both with a potential of delivering valuable data for promotion analysis and planning, as well as consumer insight on other areas of interest for the retailers.

There are many things we can analyze. However, to set the scope of our project here, we'll study the shopping behavior of customers buying at stores of the same chain that offer the same prices on the same products but are at varying distances from where they live.

Data

The dataset we used is the retail market data of Coop, one of the largest Italian retail distribution companies (Coscia, 2013).

The whole dataset contains retail market data in a time window that goes from January 1st, 2007 to December, 31st 2011. The active and recognizable customers in that interval are 1,066,020. A customer is active if they have purchased something during the data time window, while they are recognizable if the purchase has been made using a membership card. The 138 stores of the company cover an extensive part of Italy, selling 345,208 different items.

This is the dataset released as a companion for the paper “Explaining the Product Range Effect in Purchase Data“, presented at the BigData 2013 conference (Pennacchioli, 2013).

The available data is downloaded and organized as three CSV data sets, for customer purchase history data, customer-store distances and product prices. We can load them as tables into a relational database that supports analytic functions in data queries, such as Oracle, or SQL Server.

For this project, the three tables are loaded into an Oracle 18c database as three relational tables described below, as purchase data, price data and distance data, for some preliminary data processing before being loaded into R Studio for further analysis.

Purchase Data

The first column is the customer id, the second is the product id, the third is the shop id and the fourth is the total number of items that the customer bought the product at that particular shop. There are 1,048,576 records in this table without duplicates.

Column	**Data Type**

CUSTOMER_ID	Integer
PRODUCT_ID	Integer
SHOP_ID	Integer
QUANTITY	Integer

Price Data

The first column is the product id and the second column is its unit price. The price is in Euro and it is calculated as the average unit price for the time span of the dataset. There are 4,567 records in this table without duplicates

Column	**Data Type**
PRODUCT_ID	Integer
PRICE	Double

Distance Data

The first column is the customer id, the second is the shop id and the third is the distance between the customer's house and the shop location. The distance is calculated in meters as a straight line so it does not take into account the road graph. There are 301,830 records in this table without duplicates.

Column	**Data Type**
CUSTOMER_ID	Integer
SHOP_ID	Integer
DISTANCE	Double

Methodology

Since the purchase data is for customers who did shopping at stores belonging to the same chain, we need to first figure out if they have been buying different products at different stores. If that were the case, we can assume that the reason they are going to different stores is because they need different products. From the purchase data, we can find out if a customer bought the same product from multiple shops. Since the price for the same product is the same at the different stores belonging in the same chain, we can explore the data to tease out other things that may affect their decisions to purchase the same product at any of the stores at varying distances from where they live, such as other available products they also buy, or the total amount of purchase from each of the shopping trips.

We first enhance the data set by creating the following view SM_PURCHASE_ALL by joining the three tables so that all available customer-shop distance and product price are in the same data set.

SM_PURCHASE_ALL

Column	Data Type	Description
CUSTOMER_ID	Integer	Customer Id
PRODUCT_ID	Integer	Product Id
SHOP_ID	Integer	Shop Id
QUANTITY	Integer	Quantity of product bought by customer at shop
DISTANCE	Double	Distance between customer and shop
PRICE	Double	Price of the product

We then find out the purchase records as a subset of SM_PURCHASE_ALL for customers who bought the same product at more than one store by using a subquery finding out such customer and product pairs.

SM_PURCHASE_MULTI_SHOP

Column	Data Type	Description
CUSTOMER_ID	Integer	Customer Id
PRODUCT_ID	Integer	Product Id
SHOP_ID	Integer	Shop Id
QUANTITY	Integer	Quantity of product bought by customer at shop
DISTANCE	Double	Distance between customer and shop
PRICE	Double	Price of the product

Next, we use SQL analytic functions to find the distance rankings of the stores in both ascending and descending orders, number of products bought, and total amount spent for customers who bought the same products from multiple shops at varying distances from where they live.

SM_PURCHASE_ANALYTICS1

Column	**Data Type**	**Description**
CUSTOMER_ID	Integer	Customer Id
PRODUCT_ID	Integer	Product_Id
SHOP_ID	Integer	Shop Id
NUM_PRODUCS_BOUGHT	Integer	Total number of products bought by customer at shop
AMOUNT_SPENT	Double	Total amount spent by customer at shop
DISTANCE_RANKING_DESC	Integer	Distance ranking in descending order of shop for the customer who bought the product. For a given product, the store farthest away from the customer has a ranking of 1
DISTANCE_RANKING_ASC	Integer	Distance ranking in ascending order of shop for the customer who bought the product. For a given product, the store nearest the customer has a ranking of 1
DISTANCE	Double	Distance between customer and shop

We next find the distinct records without the individual products, so that they are only about the customers, how many products they bought, how much money they spent at each of the stores, and the distance ranking of each of the stores from where they live.

SM_PURCHASE_ANALYTICS2

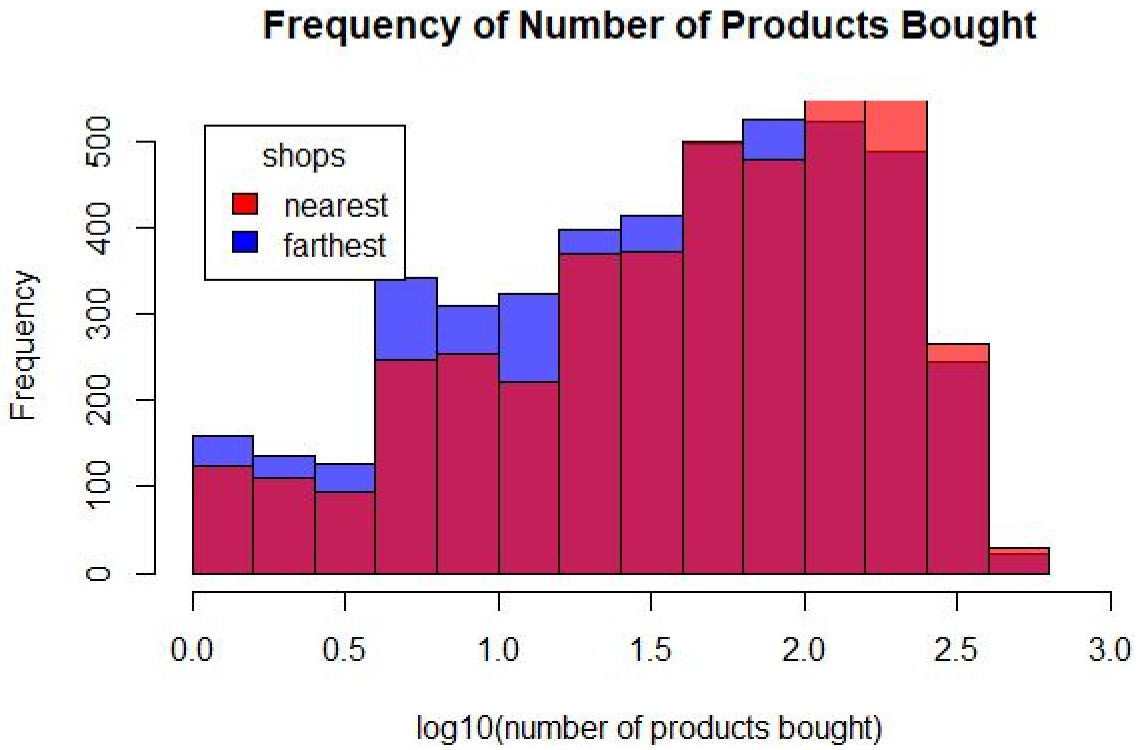
Column	**Data Type**	**Description**
CUSTOMER_ID	Integer	Customer Id
SHOP_ID	Integer	Shop Id

NUM_PRODUCS_BOUGHT	Integer	Total number of products bought by the customer at the shop
AMOUNT_SPENT	Double	Total amount spent by the customer at the shop
DISTANCE_RANKING_DESC	Integer	Distance ranking in descending order of shop for the customer. The store farthest away from the customer has a ranking of 1
DISTANCE_RANKING_ASC	Integer	Distance ranking in ascending order of shop for the customer. The store nearest the customer has a ranking of 1
DISTANCE	Double	Distance between customer and shop

Next we load up the data set into R Studio to do further analysis of customer behavior. Based on the data we have from them buying the same product at multiple shops, we can compare their shopping patterns at the store farthest from where they live and at the store closest to them.

Results

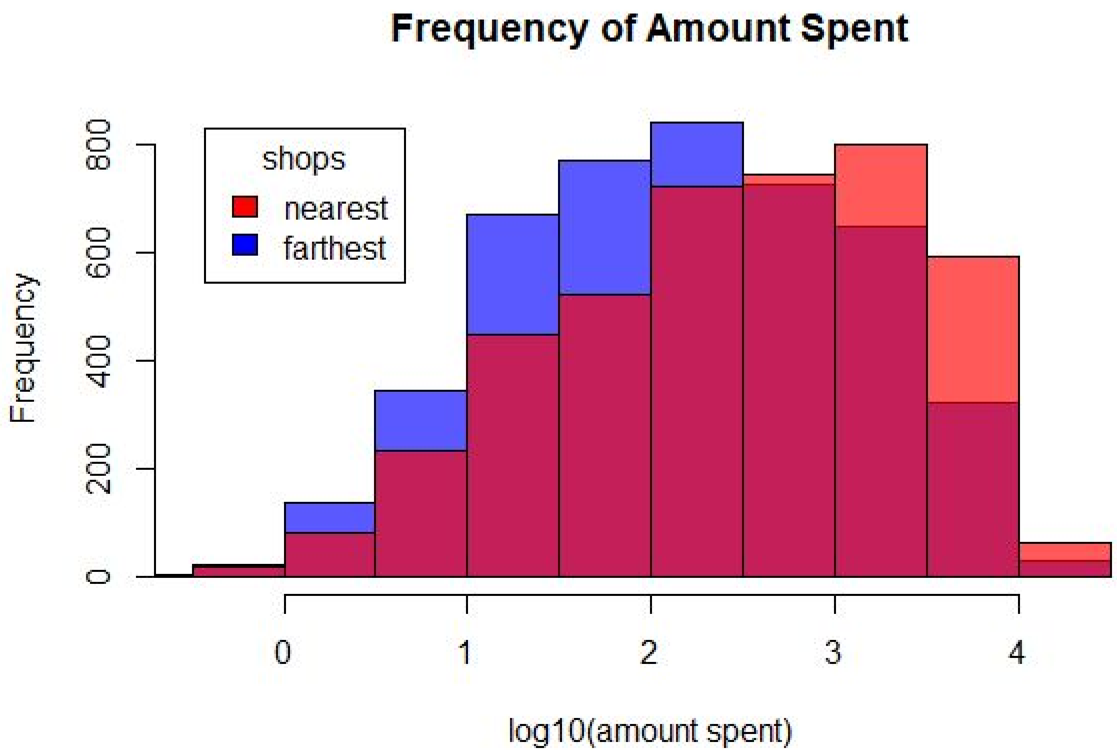
By plotting the data in histograms to compare the occurrences of total number of products bought by customers at stores closest to them vs at stores farthest from them,



Interestingly, the data shows that when the number of products bought is fewer than 100, customers who buy the same products at multiple stores are more likely to buy at the farthest stores than at the nearest stores, but the trend reverses when the number of products bought is more than 100, i.e., they are less likely to buy at the farthest store. Both seem to support a narrative that customers tend to buy fewer products at stores farther away from where they live.

From this analysis reported in (Pennacchioli, 2013) who did a more sophisticated study with more available data elements, they conclude that customers are willing to go to stores that are farther away to buy items of higher utilities to them, such as DVD players, bigger screen TV-sets etc., and as such, they tend to buy fewer things during such shopping trips, as they can get items of more common utilities such as Pizza, Frozen Side Dishes, School Notebooks etc. at nearby stores. Our finding is in agreement with their finding.

Next we look at the comparison of occurrences of total amount spent at the nearest stores and at the farthest stores



Similarly, the data shows that when the total amount spent is smaller than 300 euros, customers who buy the same products at multiple stores are more likely to buy at the farthest stores than at the nearest stores, but the trend reverses when the total amount spent is more than 300 euros, i.e., they are less likely to buy at the farthest store. Both seem to support a narrative that customers tend to spend less amounts at stores farther away from where they live, but are willing to travel further to buy products of higher utilities that are small in quantity and of higher prices.

Discussion/Conclusion

Based on the results, it seems like the least amount of products bought, customers are more likely to go to the supermarket from a farther away distance. But the more items bought or the larger amount of money spent, it seems to be the nearby stores that are attracting the customers more.

Customer behavior is influenced by many different factors. We can draw from our own experience that we as customers don't buy large-screen TV sets that often. But when we do, we can travel further to stores that offer more and better choices, which can be some regional superstores. On the other hand, for the things we need more often and of more common utility, such as toilet papers, bread, pasta, etc. we tend to stick to the stores that are closer by.

Analytic Tool Set

We used the following tools for our analysis work

Tool Type	**Tool Name**
Relational Database	Oracle 18c
SQL Command Line Interface	SQL Developer 19.2
Data Analysis and Visualization	R Studio

References

Coscia, M. (2013, February). *Supermarket Data*. From http://michelecoscia.com/wp-content/uploads/2013/02/supermarket_data.zip

Countants. (2020, January 2). *Why Consumer Behavior Analysis Is So Relevant to the eCommerce business?* From https://medium.com/datadriveninvestor/why-consumer-behavior-analysis-is-so-relevant-to-the-ecommerce-business-8f49c250ca9c

Davenport, T. H. (2011, December). Know what your customers want before they do. *Harvard Business Review*, pp. 84–92.

Farnworth, R. (2020, July 27). *How Data Science and AI Are Changing Supermarket Shopping.* From towardsdatascience.com/how-data-science-and-ai-are-changing-supermarket-shopping-e47f63f4b53f

Gedenk, K. N. (2010). Sales promotion. In M. &. Krafft, *Retailing in the 21st Century – Current and Future Trends* (pp. 345-359). Heidelberg: Springer Berlin.

Pennacchioli, D. C. (2013, February). *Explaining the Product Range Effect in Purchase Data*. From https://www.michelecoscia.com/wp-content/uploads/2013/09/geocoop.pdf

Sightcorp. (2019, November 8). *Customer Analytics - What is Customer Analytics*. From University of Amsterdam: https://sightcorp.com/knowledge-base/customer-analytics

Sightcorp. (2020, January 10). *Shopping Behavior Analysis*. From https://sightcorp.com/knowledge-base/shopping-behavior-analysis/

Sigurdsson, V. L. (2015, December). *Behavior Analysis of In-Store Consumer Behavior*. From https://www.researchgate.net/publication/282815374_Behavior_Analysis_of_In-Store_Consumer_Behavior

Uncles, M. (2010). Understanding retail customers. In M. &. Krafft, *Retailing in the 21st Century – Current and Future Trends* (pp. 159-173). Heidelberg: Springer Berlin.