CDS 302 Fall 2023: Final Project Academic Journal
Vol. 1(X) (2022), 12-11, p-ISSN: CDS-302-DL1(Fall 2022)

Consumer Purchase Patterns

Avantika Prativadhi

Abstract

Consumer Sales Analysis consists of analyzing how a company is performing in terms of sales. It also provides an understanding of what the customers gravitate towards while purchasing products. This paper presents two main questions regarding the purchase pattern of consumers in a supermarket.These questions were answered through the use of various SQL queries. This analysis can help businesses maximize profits and minimize costs.

Keywords: *SQL ; Data Analysis; Consumer Purchase; Business ; Sales ; Profit*

1. Introduction

Analyzing data is a crucial step that companies take in order to grow their business. Regardless of the industry, every company has copious amount of data. This data allows analysts to find trends and patterns in the behavior of customers which then helps the company then make important decisions. In this paper, the focus is going to be on supermarkets as the data being used is of sales from a store. At the end of the day, the main goal for any company is to make great profit while keeping the losses as low as they can. For a company to improve it's sales, the understanding of consumer purchase behavior is very important. By knowing the audience and target demographic, the store can better market their products and strategize the kind of products they want to display more or display less. This will help make predictions and estimate future sales of the company. It allows a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. For example, the company can choose to market a new product by sending ads to everyone on their database. However, this will cost them a lot of money and also frustrate customers that are not interested in that product, which then drives business away. Instead,the company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment. According to the article by [1]cience,"data-driven businesses are 58% more likely to reach their revenue goals than those that do not use data."

The main questions that will be explored through the course of the paper are: Does the level of education of a customer determine how much they are willing to spend on household groceries like fruits, meat products, and wine? and Is there a difference between online and in store purchases based on the customers' marital status? These questions were analyzed and answered using various SQL queries. The next section will go in depth on the dataset that was used in this project.

2. Data

This dataset[2] was found on Kaggle and contains detailed information of a supermarket store's customers. It contains 29 total columns with information on various attributes of the customer like their education level, how many kids they have, what their income is, their marital status,

and their birth year. It also has information on the amount of money that the customers spend on products from the store like wine, meat products, fruits, and many more.

The table in below is an example of what the original columns in the dataset look like. Since the original has many columns that cannot fit here, only the first 4 are shown.

Table 1. First four columns of the Marketing Campaign Dataset

ID	YearBirth	Education	MaritalStatus
55224	1957	Graduation	Single
2174	1954	Graduation	Single
5324	1981	PhD	Married
7446	1967	Master	Together
387	1976	Basic	Married

Here, the ID refers to the individual customer; the yearBirth is the year that they were born, the Education is their highest degree of education; and the MaritalStatus is whether they are single, married, or with someone.

2.1. Database Schema

There are three tables that were joined to conduct the analysis for this project. The first table had all the information regarding the customer's personal attributes and the second table had information regarding the products. In the original database, there were also other tables pertaining other aspects of the business like their sales online and the use of specific coupons for the various products. However, since the first question of intent for the paper was specific to the consumer and the product. The other tables were ignored and the focus was on the customer's information and the particular products like meat, fruits, and wine. To build the SQL schema, the **CREATE TABLE** statement where both the tables were connected using the key ID which is a unique ID for each of the customer. The dataset does not have named of the customers so the ID helps differentiate each of them.

For the second question of intent, another table was created and used. This table contains information about the location of the purchase. Some customers preferred ordering their products online while some enjoying going in the actual store.

The figure below is an ER Diagram that gives a general overview of all the aspects of the database. All the different tables from the dataset are shown. Products, Customer, and Place were used for this analysis.

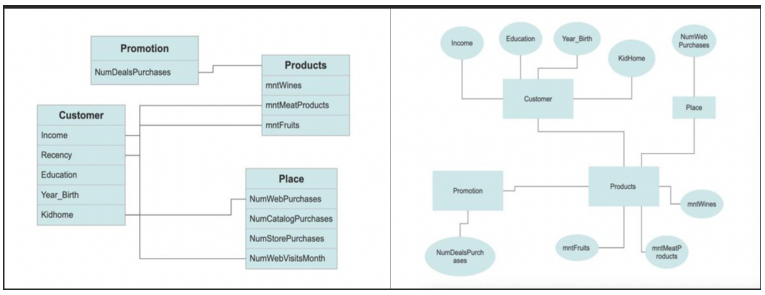

Figure 1. ER Diagram mapping the entities and relationships

3. Relevant operations

The volume of the dataset is very large as there are 2240 entries for each of the columns. To easily and quickly parse through all the values, the regular SQL queries that were used were **SELECT** with MIN, MAX, SUM, COUNT, AVG, GROUP BY, HAVING functions. The select function was used to select the specific column that was wanted, followed by **FROM** which specified which table the data should come from. This main statement was followed by the other functions which then helped answer the questions.

3.1. Does the level of education of a customer determine how much they are willing to spend on household groceries like fruits, meat products, and wine?

The process of answering this question started with first finding the max amount spent for each type of product. To do this, the SELECT command was used to select the Education column, along with the three different columns. Using the function max() on each of the products, the maximum money spent on each of them was shown. The FROM command was used to specify which table the information should be collected from and finally, the Group by command was used to group all the results based on the degree type. The table below shows the output result of the query.

Table 2. Maximum Amount Spent On Products Grouped By Degree Type

Education	Wines	Meat	Fruits
Graduation	1,492	1,725	199
PhD	1,493	1,622	197
Master	1,486	925	194
Basic	228	122	122

Based on the results of the query, there is not that much difference among the total amount being spent on wine between people with Grad, PhD, and Master students. But there is a significant different between those groups and the people with the Basic (Bachelors) degree. Similarly, with meat products people with Grad degrees and PhD's tend to spend more than people with Masters and Bachelors. In terms of fruits, there does not seem to be a major difference. This makes sense as Wine can be seen as a more luxury item and there are higher end brands that sell it for a larger price. Hence, people are spending a lot of money for that product and people making more money are able to afford it. Whereas, with fruits, there is a certain price that they are sold at. Usually, that price is pretty low so the income of the people does not affect that as much.

To further expand on this query, the **WHERE** clause can be added to find certain attributes. For example, to find the same information as above but only for customers that are married, the where clause can be used along with the comparison operator "=" which checks if the statement on the right is equal to the statement to the left and gives the result of all the true values.

3.2. Is there a difference between online and in store purchases based on the customers' marital status

To answer this question of intent, a similar query as the last one was done. However, in this one, rather than finding the maximum value of the product, the count was used. The count function returns the number of rows that match the criteria. Since count was used on the "NumWebPurchases" column, the count of each of the web purchases was returned. The group by clause was used on the marital status column so that the number of purchases for each of the different status' was shown. Finally, the order by clause was used to order the returned result in descending order(desc) so that it was easier to analyze the different groups. The table below shows the result of this query.

Table 3. Number of Web Purchases done Grouped by Marital Status

Marital Status	Online Purchases
Married	857
Together	573
Single	471
Divorced	232

The results show that people in a relationship are more likely to shop online as more online purchases were done by groups that are married or together while the single and divorced seem to purchase less online.

4. Significance of the operations

The various SQL functions made it possible to quickly find the results of the questions. What would have taken ages to look through each of the rows and find the answers manually, only took a few minutes using the operations. The major operations which made the queries possible were comparison operations like equal to and greater than or less than. When trying to find the results of a particular group, the equal to operator helped as it made it possible to select a certain category. When the desired output was only the group of customers with Masters degrees, adding "Education ="Master" was all that was needed to get the result. The greater than or less was useful when trying to filter the data and find when the customers spent greater than or less than a price. This operator especially came in handy when trying to find the minimum amount of money the customers were spending. Since the default was set to 0 where there was no data, the result for the minimum kept showing as 0. So, by adding the greater than operator (MntWines $>$ 0), the results were the lowest numbers that were greater than 0. This can also be used to find the number of wines bought at a certain price or lower price.

Along with these operators, the functions like avg helped find the average of the amount that the customers were willing to spend; max and min gave the most amount and the least amount that they were willing to spend. Then, the clauses like Select, Group by, having, From, and order by made it easy to find the right column of data needed to execute all the sets of queries.

5. Conclusion

Large amount of data can be quickly accesses and analyzed using used a few commands and operations. This helps the company make major progress with their sales every year. Nevertheless, this is just a small portion of the vast number of questions this dataset and the SQL queries can answer. This study can be further progressed by taking the other columns into consideration and using other operators to look further into what the customers gravitate towards for their shopping.

References

[1] Pagotto, D. (2022, December 6). How Sales Data Analysis Can Help Grow Your Business. 5 Reasons Why Sales Data Analysis Will Boost Your ROI. Retrieved December 10, 2022, from https://www.cience.com/blog/sales-data-analysis.

[2] Romero-Hernandez, O. (2021, February 8). Customer Personality Analysis. Retrieved October 29, 2022, from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.