UCD PA: Data Analytics for Business – Final Project

Charlotte McCarthy
10/02/2022

GitHub URL

https://github.com/CharlottePPB/UDCPA_CharlotteMcCarthy

Abstract

Introduction

Supermarket data is widely collected around the world and has a variety of possible use cases. Advertising and marketing companies can use the data to predict which customers are most likely to buy their products and then target offers, promotions and adverts directly at those customers. Supply chain managers can use the data to predict which products will be most in demand at certain periods of time e.g., ice cream sales will go up in the summer, while soup sales go up in winter. They use this information to plan which products to order and in what quantities to ensure that supermarkets are always fully stocked.

When completing this project, I kept those real-world use cases in mind. I chose to focus on the customer data and place myself in the role of marketing / promotion manager deciding which customers to target for which marketing campaigns and at what time.

Dataset

The dataset chosen for this project came from Kaggle. I chose this dataset because it was opensource and came from a real supermarket company. The one disadvantage of this dataset is the limited number of rows, just 1000. A full dataset capturing the same information would likely have millions of rows.

The dataset captures purchases across three branches of a large supermarket chain. It contains a mix of customer data, like which branch the customer visited and whether they're a member of the supermarket loyalty scheme, and transaction data, like what the customer purchased and how much profit the supermarket made.

Column	Description
Invoice ID	The auto generated invoice ID
Branch	The supermarket branch
City	The city where the branch is located
Customer Type	Is the customer a member of the loyalty scheme?
Gender	Gender of the customer
Product Line	Product line of item purchased
Unit Price	Unit price of item purchased
Quantity	Quantity of items purchased
Tax 5%	Total tax paid by customer
Total	Total amount paid by customer
Date	Date the purchase was made
Time	Time the purchase was made
Payment	Payment method used

COGS	Cost of the goods sold
Gross Margin %	Margin percentage for the supermarket
Gross Income	Gross profit for the supermarket
Rating	Customer satisfaction rating for their visit

Table 1: Description all columns in the dataset

Implementation Process

The data has nine object columns, seven float columns and one integer column. The pandas dtypes function was used to easily view this information.

```
Invoice ID                 object
Branch                     object
City                       object
Customer type              object
Gender                     object
Product line               object
Unit price                float64
Quantity                    int64
Tax 5%                    float64
Total                     float64
Date                       object
Time                       object
Payment                    object
cogs                      float64
gross margin percentage   float64
gross income              float64
Rating                    float64
```

Figure 1: Data type of all columns in the dataset

Based on the data types, two new dataframes were created. One dataframe held all the object data, while the other held all the numeric data. This was done so that the describe function could be used to look at the summary statistics for all the data in the two dataframes.

	Invoice ID	Branch	City	Customer type	Gender	Product line	Date	Time	Payment
count	1000	1000	1000	1000	1000	1000	1000	1000	1000
unique	1000	3	3	2	2	6	89	506	3
top	627-95-3243	A	Yangon	Member	Female	Fashion accessories	2/7/2019	14:42	Ewallet
freq	1	340	340	501	501	178	20	7	345

	Unit price	Quantity	Tax 5%	Total	cogs	gross margin percentage	gross income	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	307.58738	4.761905	15.379369	6.97270
std	26.494628	2.923431	11.708825	245.885335	234.17651	0.000000	11.708825	1.71858
min	10.080000	1.000000	0.508500	10.678500	10.17000	4.761905	0.508500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	118.49750	4.761905	5.924875	5.50000
50%	55.230000	5.000000	12.088000	253.848000	241.76000	4.761905	12.088000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	448.90500	4.761905	22.445250	8.50000

Figure 2: Summary statistics for all data columns

Once the summary statistics were viewed, the time and date functions were transformed from objects to datetimes using the pandas to_datetime function. This was done to make them easier to manipulate later.

The next step was to look for null values within the dataset and replace them if necessary. The initial search for null values was done by combining the isnull() and sum() functions. This adds up the count of null values in each column. By using this method, we can get an indication of how many null values exist and where they exist within the dataset.

```
Invoice ID                0
Branch                    0
City                      0
Customer type             0
Gender                    0
Product line              0
Unit price                0
Quantity                  0
Tax 5%                    0
Total                     0
Date                      0
Time                      0
Payment                   0
cogs                      0
gross margin percentage   0
gross income              0
Rating                    0
```

Figure 3: Count of Null Values within the dataset

This dataset has no missing values; therefore, no replacement is required in this case. If there were missing values, the options for dealing with them would include deleting all rows with missing values or replacing the missing values with the mean value for that column. The method for dealing with the missing values would depend on what percentage of data was missing.

Results

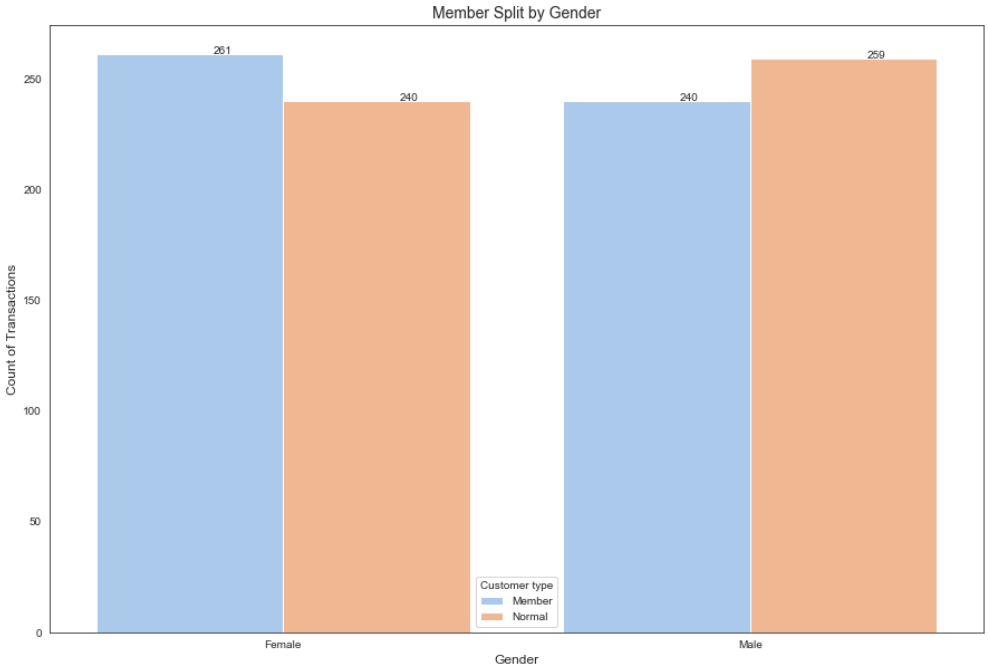

Figure 4: Breakdown of transactions by Gender and Member Type

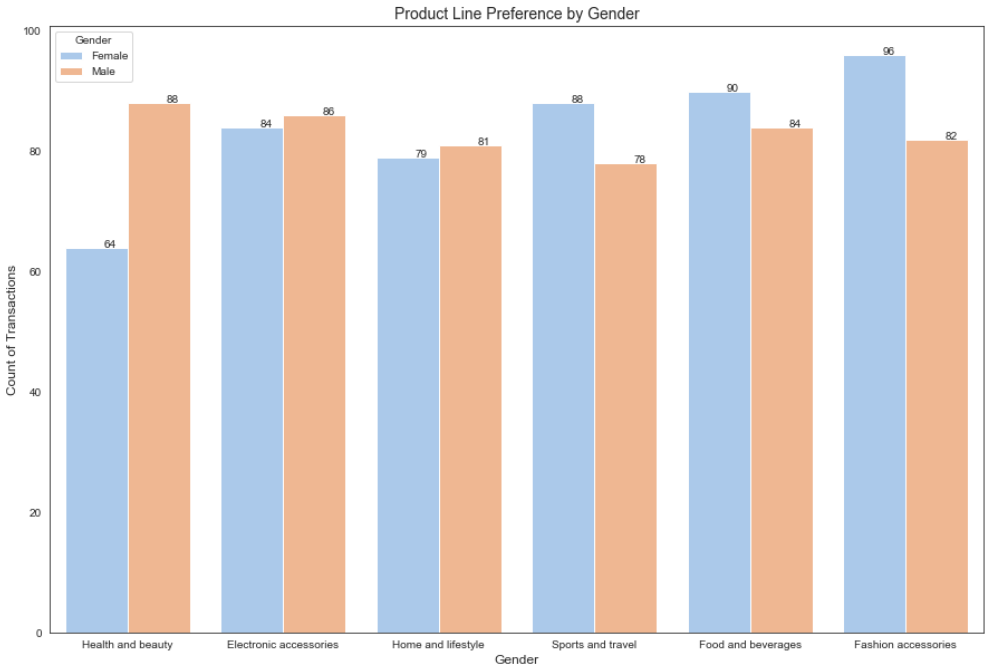

Figure 5: Seaborn count plot that shows a breakdown of purchases by Gender and Product Line

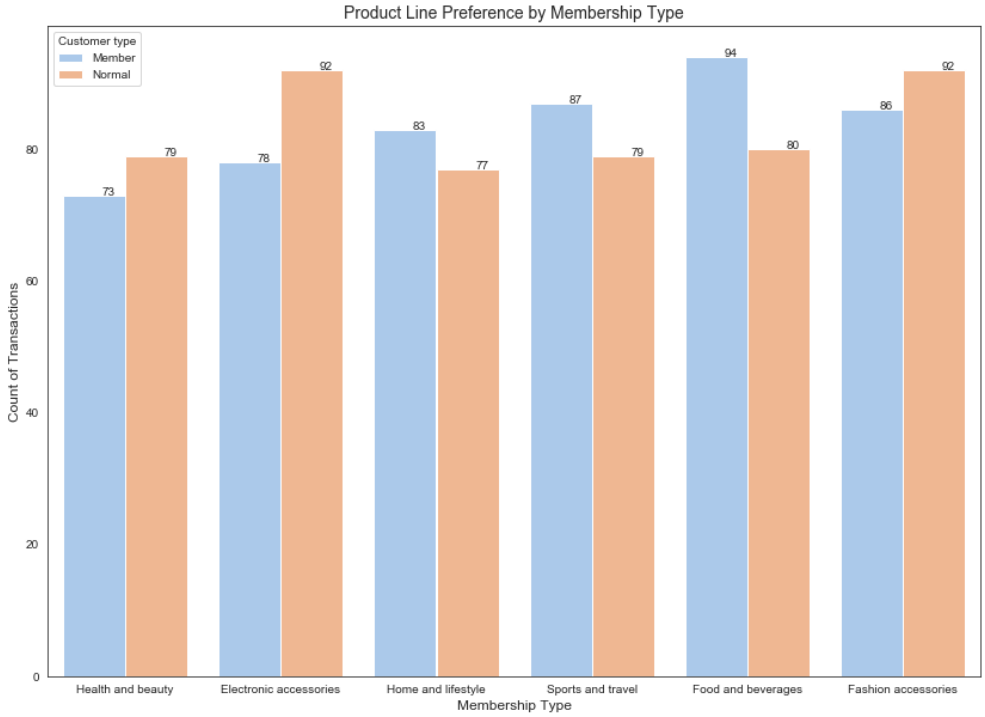

Figure 5: Seaborn count chart breaking down transaction count by Membership Type and Product Line

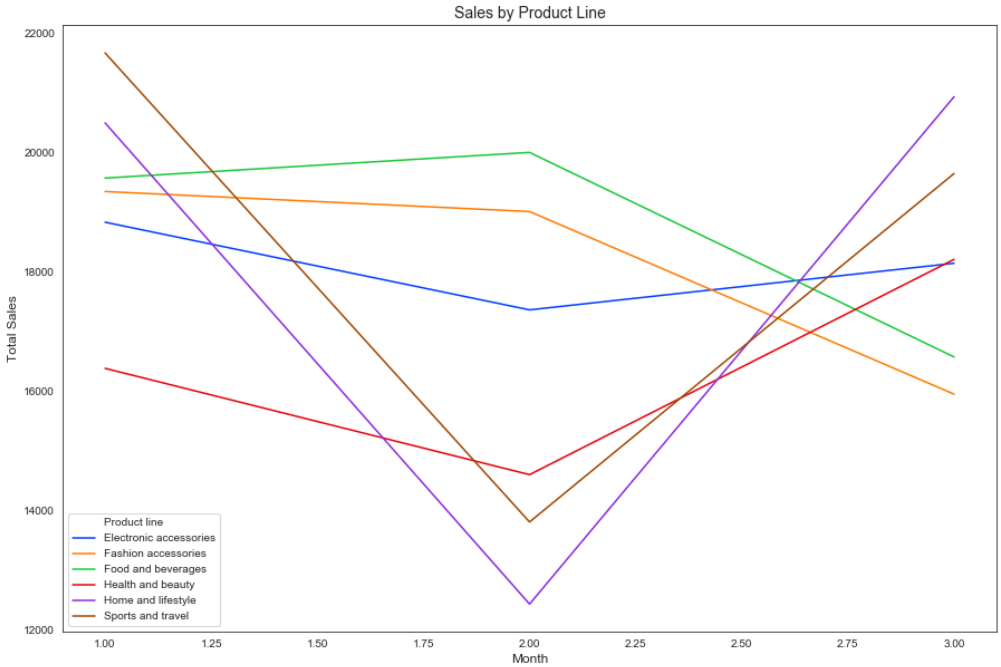

Figure 6: Seaborn line plot that tracks total sales by Month and Product Line

Insights

1) More women than men are member of the supermarket discount scheme
2) Men buy more health and beauty product than women, which is surprising given that these products are often advertised more heavily towards women
3) Women buy more fashion accessories, food and beverage products and sports and travel products than men. Since sports products are more heavily advertised towards men, it would be interesting to see a more detailed breakdown of the products within the sports and product category to see what products within that category women are most interested in
4) People who are members buy more food and beverage products than non-edible goods
5) There was a large dip in sales in February across four product categories, perhaps sending additional promotions at this time of year would boost sales and prevent that dip in future years

Potential Follow Up Work

If I was looking to perform Machine Learning on this dataset, I would look to build perform supervised learning and build a classifier model to identify customers that are most likely to sign up to the membership scheme in the future so I can target these customers with advertisements showing the benefits of the scheme.

I would use the Customer Type column as my label and the other columns as my features. I would try using a decision tree or random forest algorithm for my prediction.

References

Kaggle.com