# 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 and tried to think about how the data could be used to solve some of the above problems.

## 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.

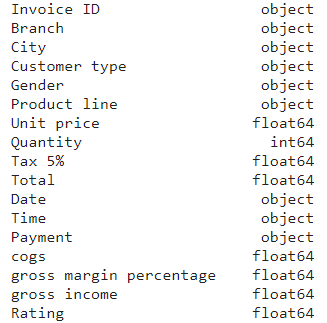
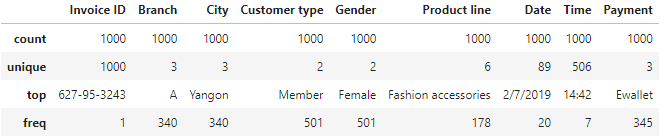


Figure 1: Data type of all columns in the dataset

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.



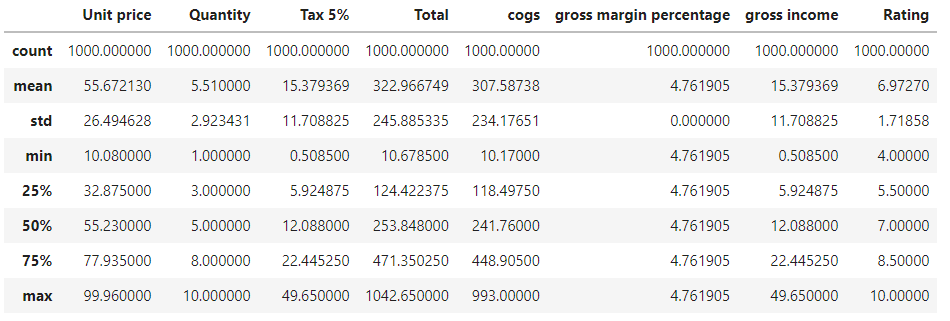
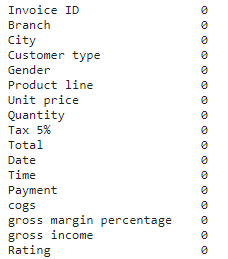


Figure 2: Summary statistics for all data columns

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.



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

(Include the charts and describe them)

## Insights

(Point out at least 5 insights in bullet points)