Title: Wholesale customer dataset analysis

Abstract

A wholesale distributor possesses data on the yearly expenditure of various items in their stores across diverse regions and channels. Utilizing this information, an analysis report was generated through a series of tasks encompassing data cleaning, exploration, analysis, and visualization of variable relationships, alongside classification employing K-Nearest Neighbor and Decision Tree models. Each task enhances comprehension of the dataset and primes it for modeling and further analysis.

Introduction

This report was written to present the results of data analysis from the Wholesale Customers Dataset, which describes customers of a wholesale distributor and shows how much they spend each year on different types of products, measured in monetary units. The goal of the project is to gain insights from the dataset by analyzing relationships between variables, predicting the trends of the market. Thereby, reporting to stakeholders to answer the business questions that have been raised.

Task 1

In selecting the "Wholesale Customers Dataset" as the foundation for the project, some important factors were considered like the data integrity, information diversity, and applicability to the project's objectives. This data set includes numerical features such as annual spending on different products, also contains categorical indicators such as sale channels and sale regions, makes it a great fit for the report's purpose.

Before exploring the dataset, some of the following important steps have been taken:

- **Data Import**: Import the dataset to Jupiter Notebook, using Python
- **Data Cleaning**: Examine the dataset to identify any missing or inaccurate information and clean the dataset accordingly. This may involve removing or imputing missing data or correcting any obvious errors.
- **Data Description**: Generate summary statistics such as mean, median, and standard deviation for each column of the dataset. This will help in understanding the distribution of data in each column.

Task 2: Data Exploration

2.1 Exploring the columns in the dataset:

After understanding the basic information about the data set, a deeper investigation is conducted by examining the specific data within each column of the table. The data consists of 440 large retailers' annual spending on 6 different varieties of products in 3 different regions (1 for Lisnon, 2 for Oporto and 3 for Other) and across different sale channels (1 for Retail, 2 for Horeca).

Channel: The bulk of the data is centered around the value 1, indicating that the majority of customers are associated with the Horeca channel, account for 67.7%. Conversely, the Retail channel (indicated by the value 2) significantly less common compared to the Horeca wholesale channel, only 32.3%. HORECA stands for Hotels, Restaurants, and Catering, representing the business-to-business (B2B) clientele of wholesalers, who supply consumer goods to the hospitality and food and beverage sectors, as opposed to individual retail customers. Specific in the chart below:

Figure 2.1. Sale chanels propotion

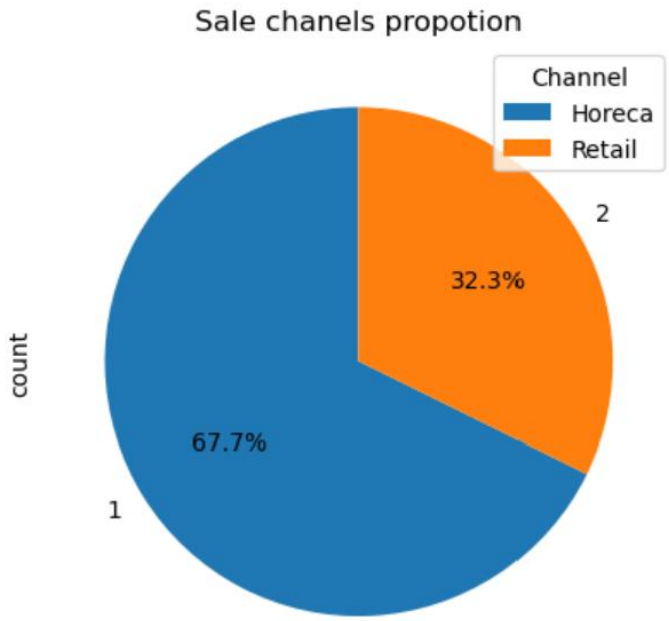
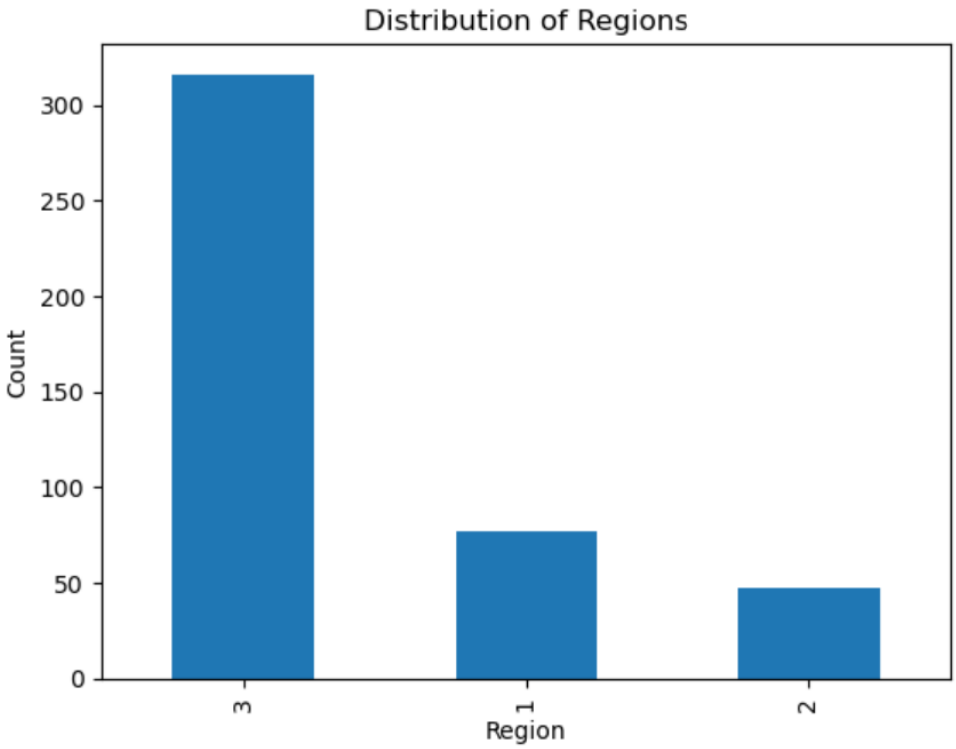


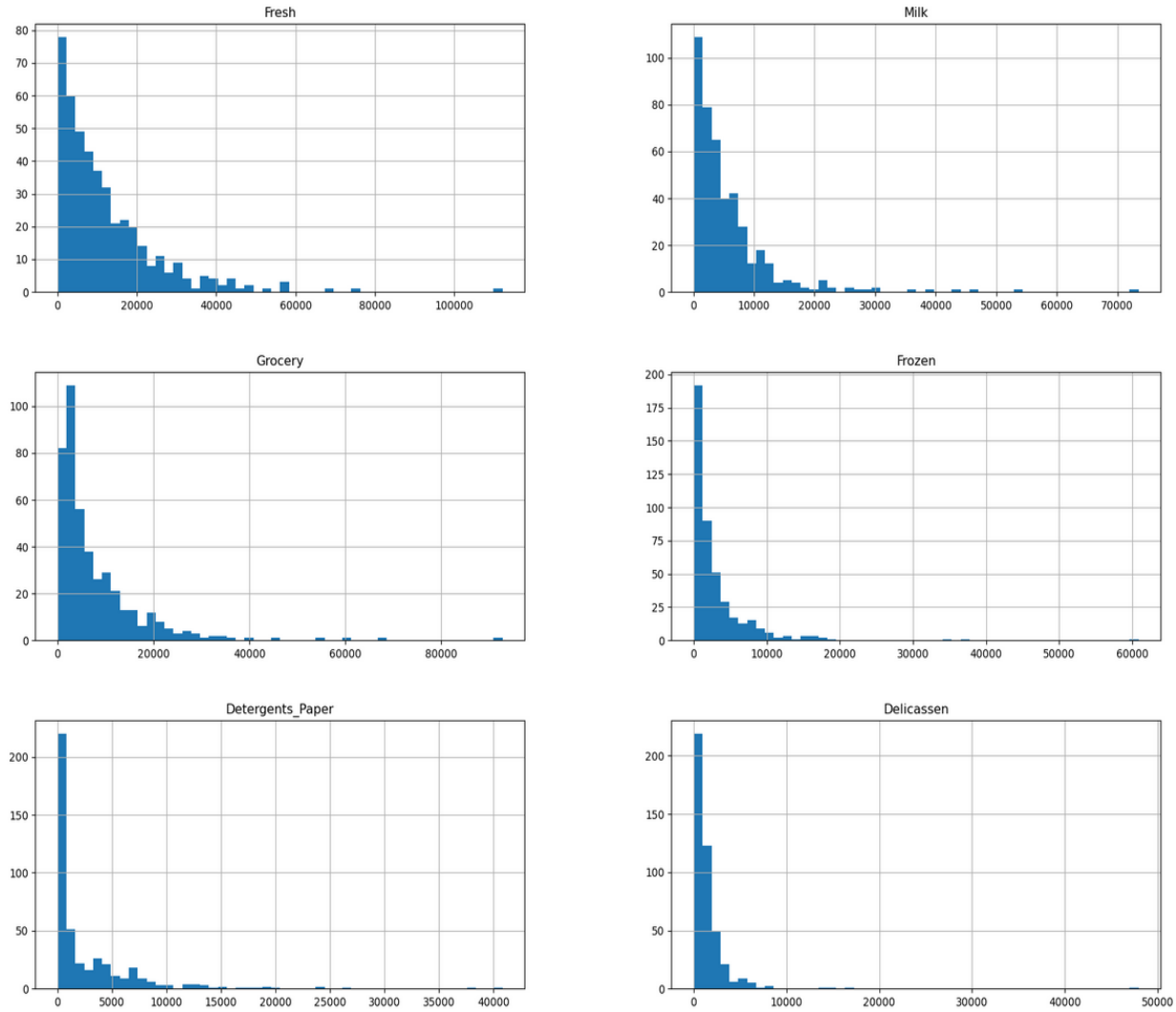
Figure 2.2. Distribution of Regions



Region: The majority of the data is centered around the value of 3, which is an aggregation of "Other Region" of that country. It doesn't provide further details about which cities/locations contribute to such a density of wholesale customers in whole. There are

smaller concentrations around the values of 1 (Lisnon) and 2 (Oporto), but they have fewer customers compared to the rest of the regions combined.

Figure 2.3. Distribution of numerical columns:



Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen: The numerical columns depicting annual spending across various product categories exhibit right-skewed distributions, suggesting that most customers in the dataset have low to moderate spending habits. Additionally, a handful of outliers in these columns indicate a small number of customers who spend significantly higher amounts.

2.2 Exploring the relationship between some pairs of column in the dataset:

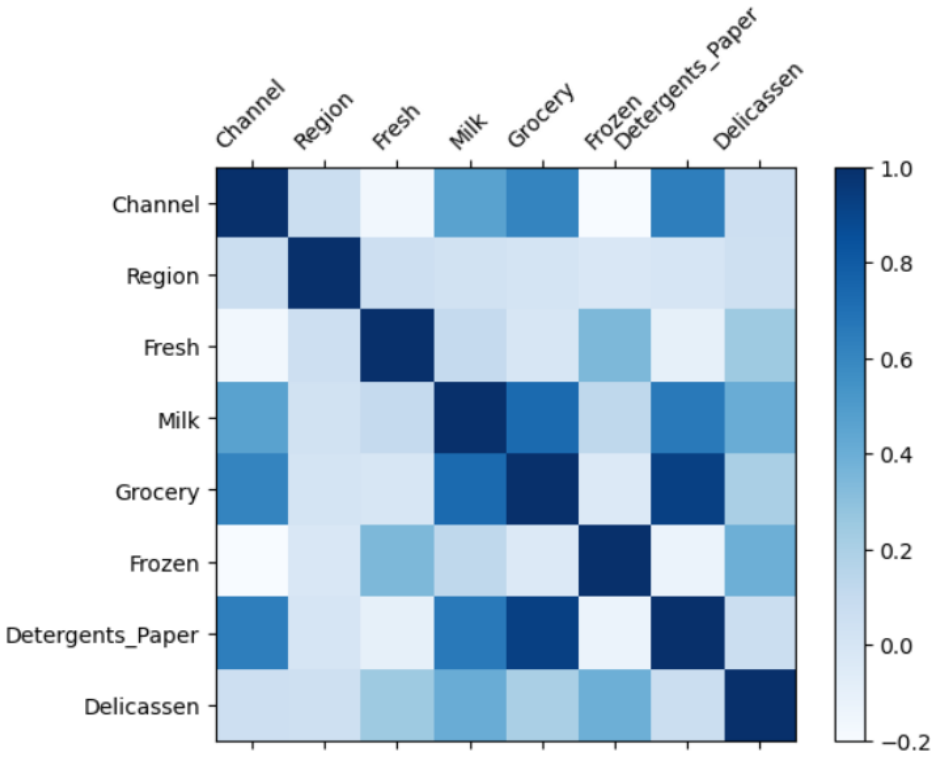
The pairwise correlation of all columns was calculated in order to investigate the relationship between pairs of column. The correlation is a number between -1 and 1 that indicates how closely the two variables are related. A positive correlation indicates that

the variables tend to increase together, a negative correlation indicates that when one variable increases, the other tends to decrease, and a correlation close to 0 indicates that there's no linear relationship between the variables.

Figure 2.4. The correlation of all columns:

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Channel	1.000000	0.062028	-0.169172	0.460720	0.608792	-0.202046	0.636026	0.056011
Region	0.062028	1.000000	0.055287	0.032288	0.007696	-0.021044	-0.001483	0.045212
Fresh	-0.169172	0.055287	1.000000	0.100510	-0.011854	0.345881	-0.101953	0.244690
Milk	0.460720	0.032288	0.100510	1.000000	0.728335	0.123994	0.661816	0.406368
Grocery	0.608792	0.007696	-0.011854	0.728335	1.000000	-0.040193	0.924641	0.205497
Frozen	-0.202046	-0.021044	0.345881	0.123994	-0.040193	1.000000	-0.131525	0.390947
Detergents_Paper	0.636026	-0.001483	-0.101953	0.661816	0.924641	-0.131525	1.000000	0.069291
Delicassen	0.056011	0.045212	0.244690	0.406368	0.205497	0.390947	0.069291	1.000000

Figure 2.5. The correlation matrix:



Through Figure 2.4 and Figure 2.5, it can be easily observed that dataset has some significant correlations. Detergents_paper and Grocery, Detergents_paper and Milk or Grocery and Milk have a very strong positive linear relationship while Detergents_paper and Frozen, Fresh and Detergents_paper have weak linear relationship.

The following three tables will help you see the relationship between these pairs of products more clearly.

Figure 2.6.

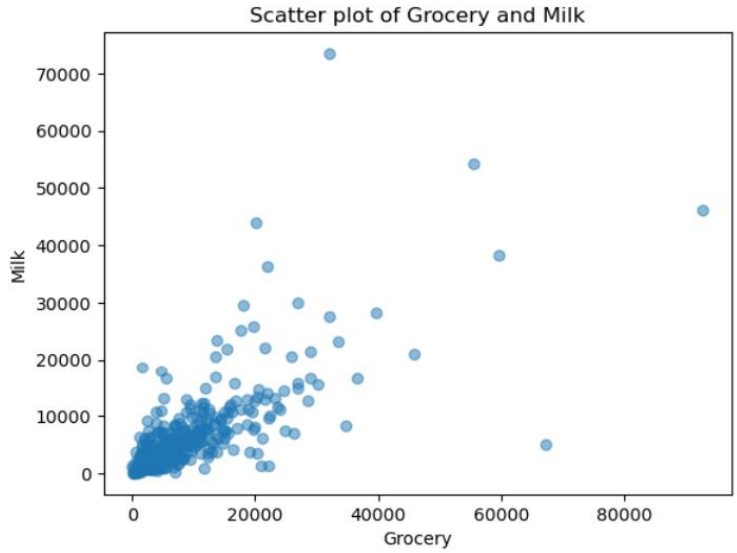


Figure 2.7.

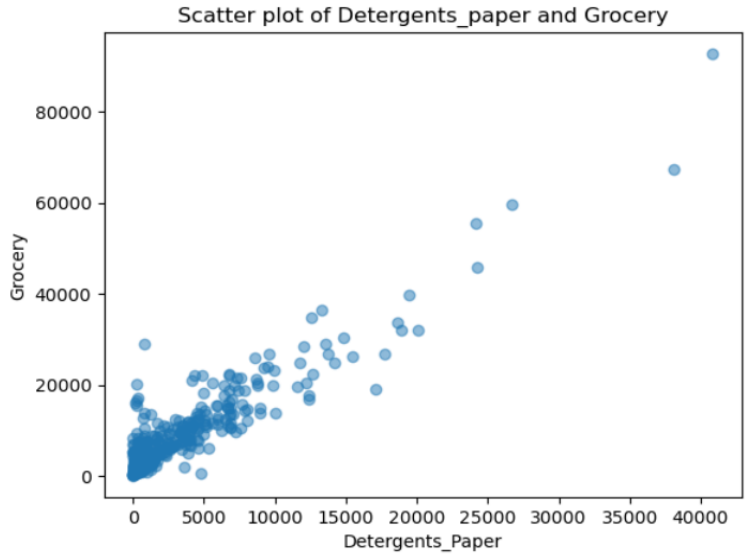
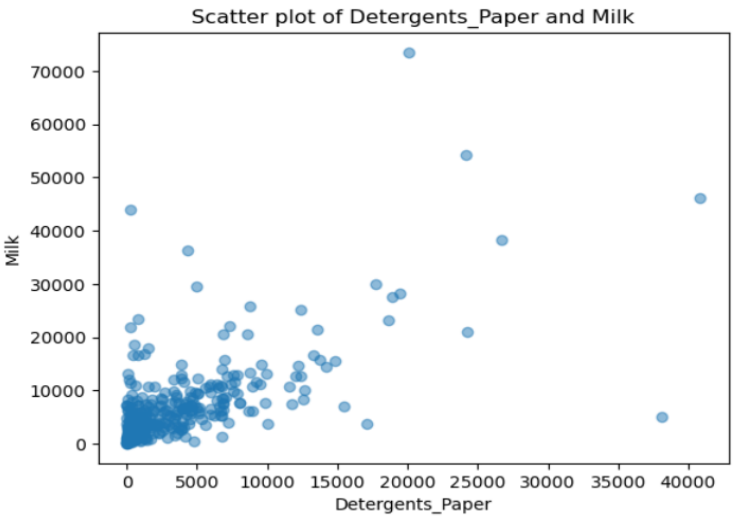


Figure 2.8.



The strong linear relationship between these three factors can be explained by the following reasons:

- Market supply and demand: If the supply or demand for one product increases, the supply or demand for the other product may also increase. For example, if consumers buy more milk, they may also purchase more fruits and vegetables to maintain a healthy diet.
- Consumer habits: Consumers may have a habit of buying milk and fruits and vegetables together, creating a linear relationship between the two products.

The following two charts will depict the weak linear relationship between two pairs Detergents_paper and Frozen, Fresh and Detergents_paper:

Figure 2.9.

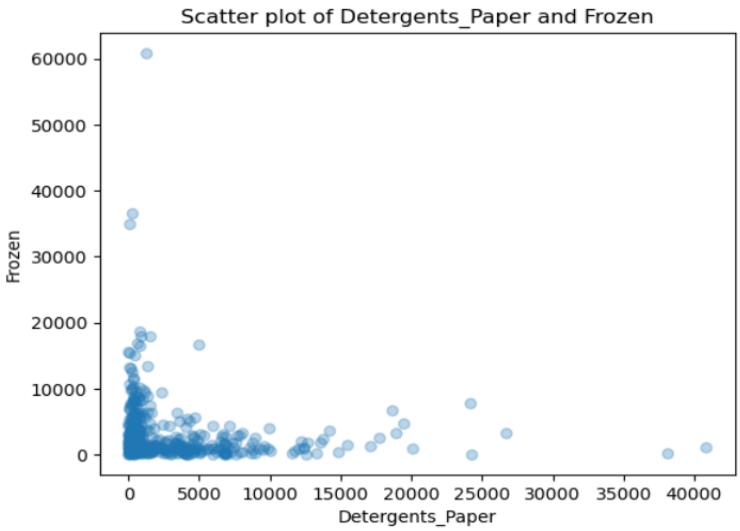
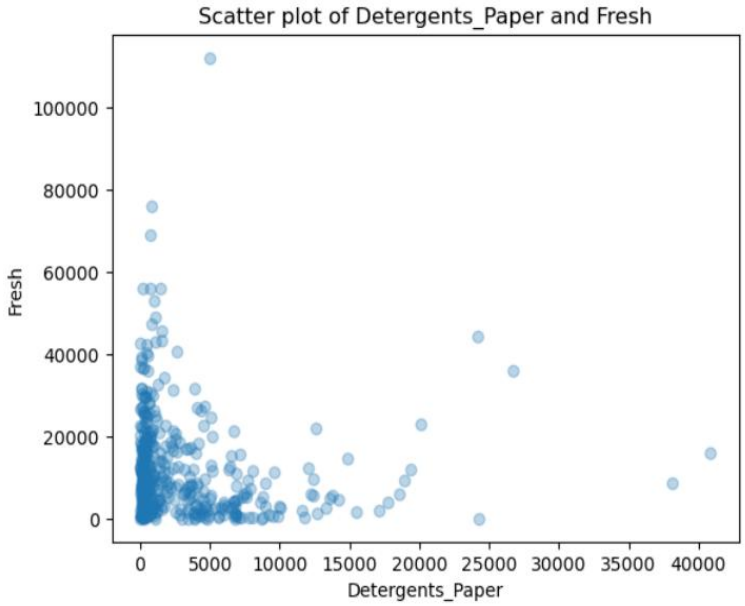


Figure 2.10.

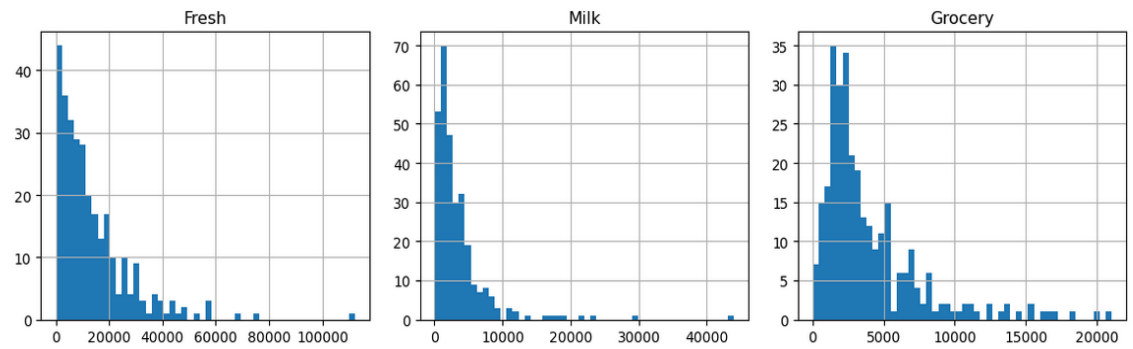


The weak linear relationships may stem from the fact that these categories serve distinct purposes and are not typically purchased together. Fresh produce is primarily consumed as food, while Detergents_paper products are used for cleaning and household purposes. Consequently, consumers' purchasing behavior in one category may not strongly influence their purchases in the other, resulting in a weak correlation between the two variables.

2.3 Answer the question: "Do spending patterns differ between Horeca and Retail customers across product categories?"

To investigate this issue, we will divide the dataset into two parts: one containing product groups sold through retail channels and the other consisting of products purchased by the Horeca group. The dataset after division will be represented by charts as shown below:

Figure 2.11 Spending amount on Horeca



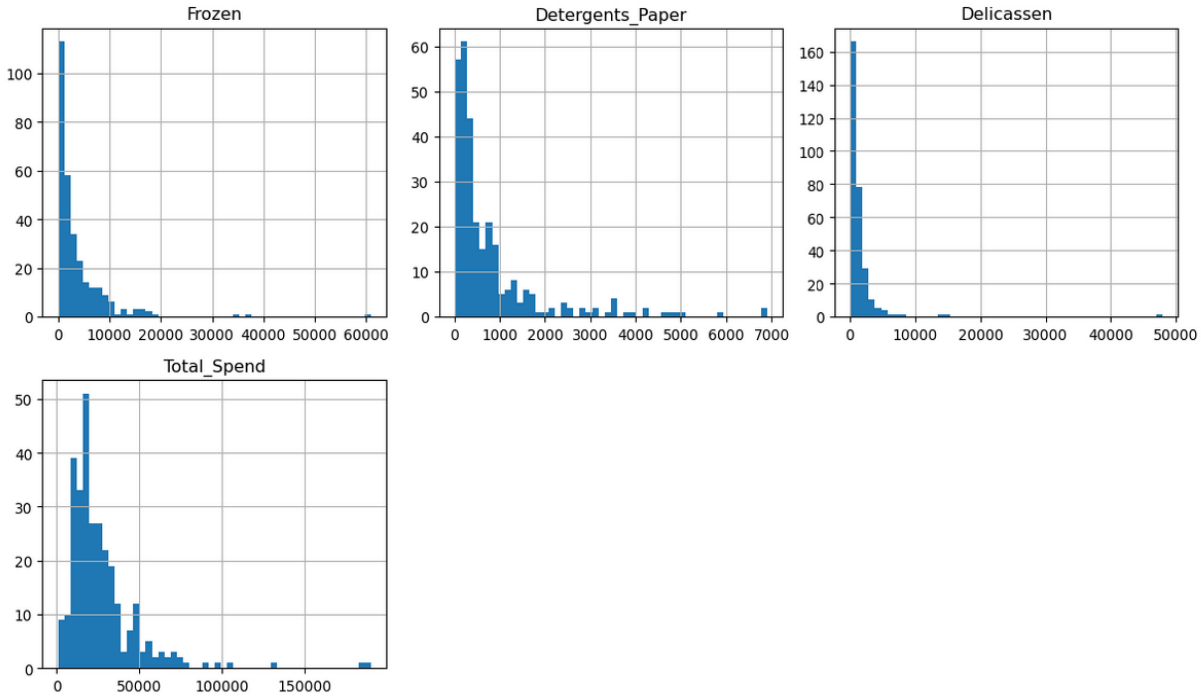
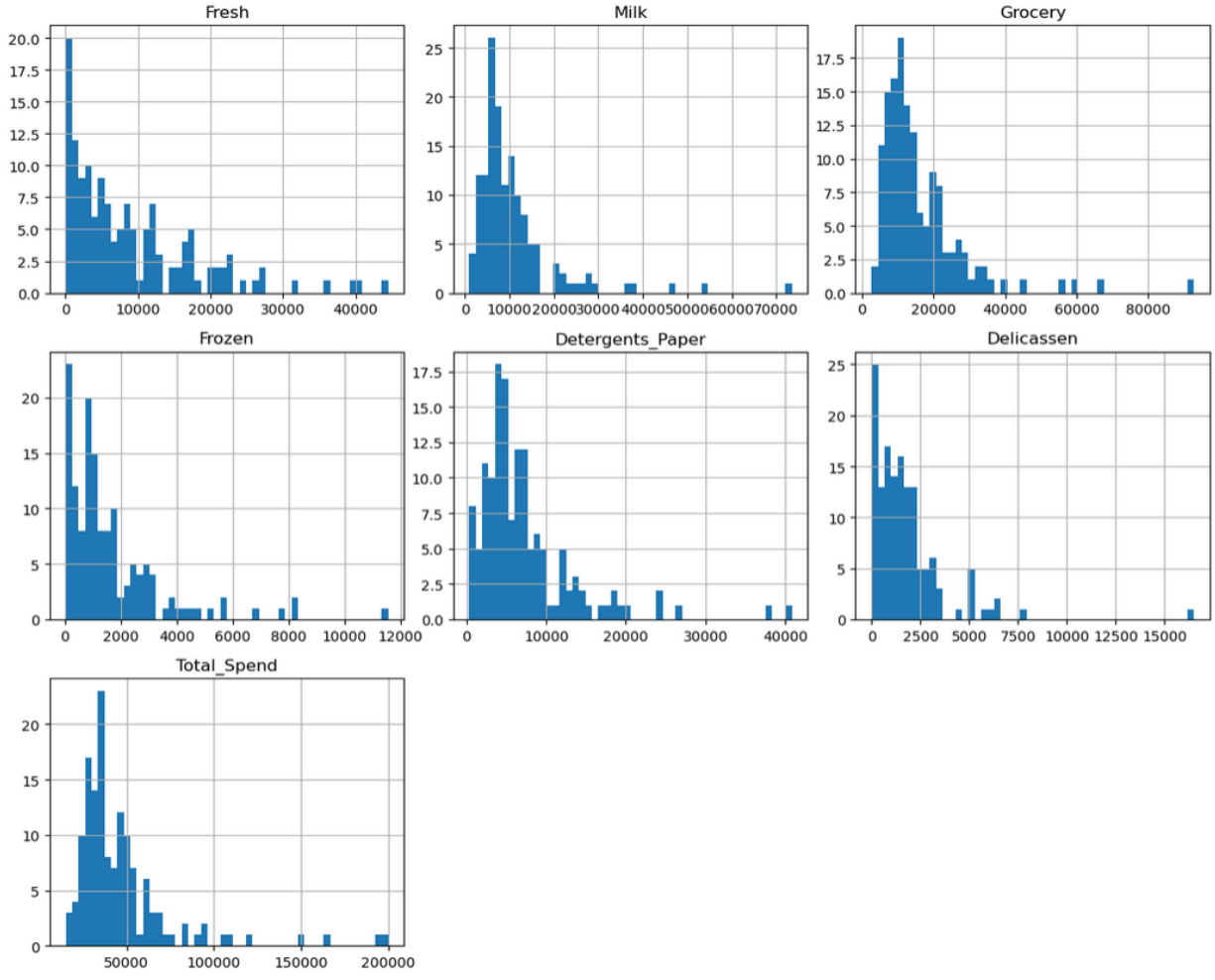


Figure 2.11 Spending amount on Retail channel:



Upon examining the histograms from both the Horeca and Retail channels, we made several observations:

- Despite Horeca customers making up nearly 68% of the total, their average annual spending across all categories is only 26.844 while Retail customers account for 46.619
- Retail customers spend more on Milk, Groceries, and Detergents_paper than Horeca customers.
- Both channels see similar overall spending patterns for the rest of the categories, despite Horeca customers slightly spend more on food-related categories such as: Fresh produce and Frozen products.

Task 3: Data Modelling

One of the key aspects of supervised machine learning is model evaluation and validation. When assessing the predictive performance of a model, it is essential that the prediction process is not biased. Using the train_test_split() function from the scikit-learn library, the dataset can be divided into subsets to minimize the risk of bias during evaluation. The data set will be splitted to form 3 different suites of trainning and test set as following ratio:

- Suite1: 50% for training and 50% for testing
- Suite2: 60% for training and 40% for testing
- Suite3: 80% for training and 20% for testing

K-Nearest Neighbor and Decision Tree classification was applied to each Suite to make a prediction for 'Channel' feature in the dataset. Performances of the model on the training and test sets was evaluated in terms of:

- Confusion matrix
- Accuracy score
- Precision score
- Recall score
- F1 score

On Suite 1: Performance of two models was calculated and showed in following graph:

Figure 3.1. Confusion matrix of 2 model:

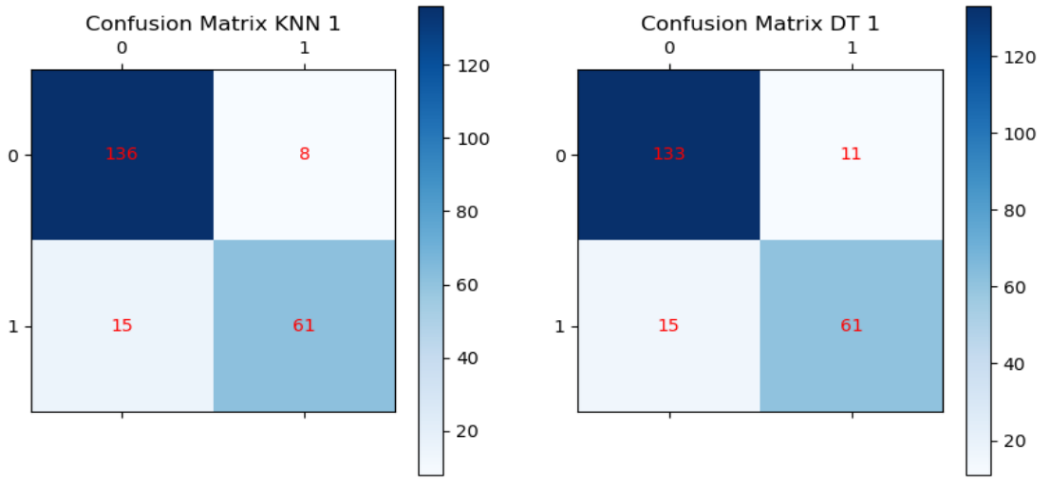
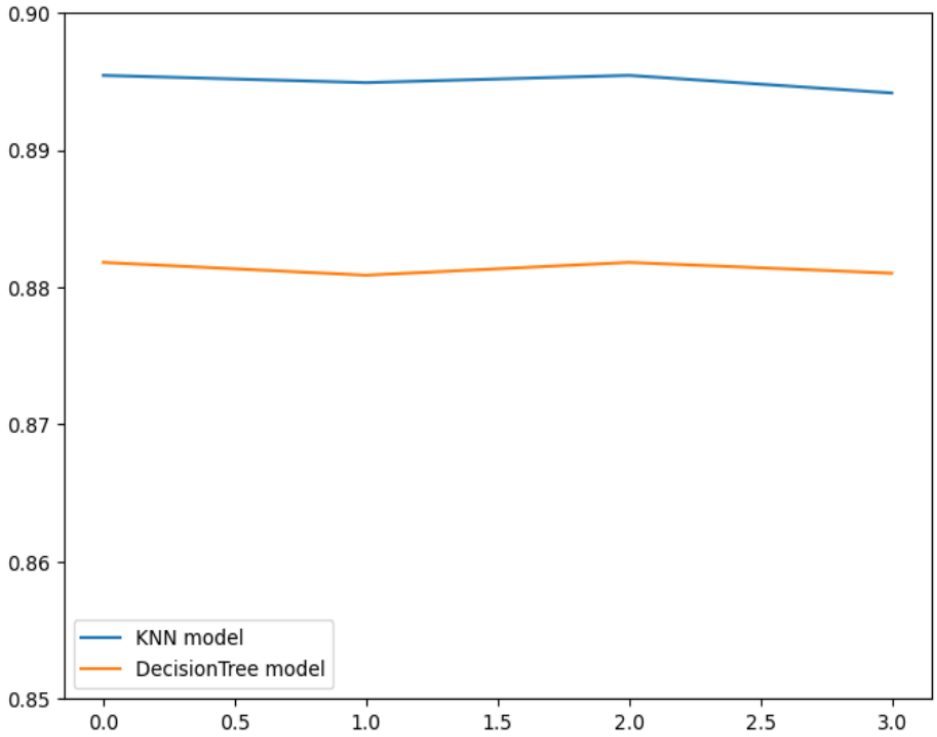


Figure 3.2. Different between the other evaluation value:



From the two figures, we can observe that the predictions of the two models are quite similar . Both confusion matrices have low FP and FN values, with other metrics ranging from 0.88 to 0.89.

<u>On Suite 2:</u>

Figure 3.3. Confusion matrix of 2 model:

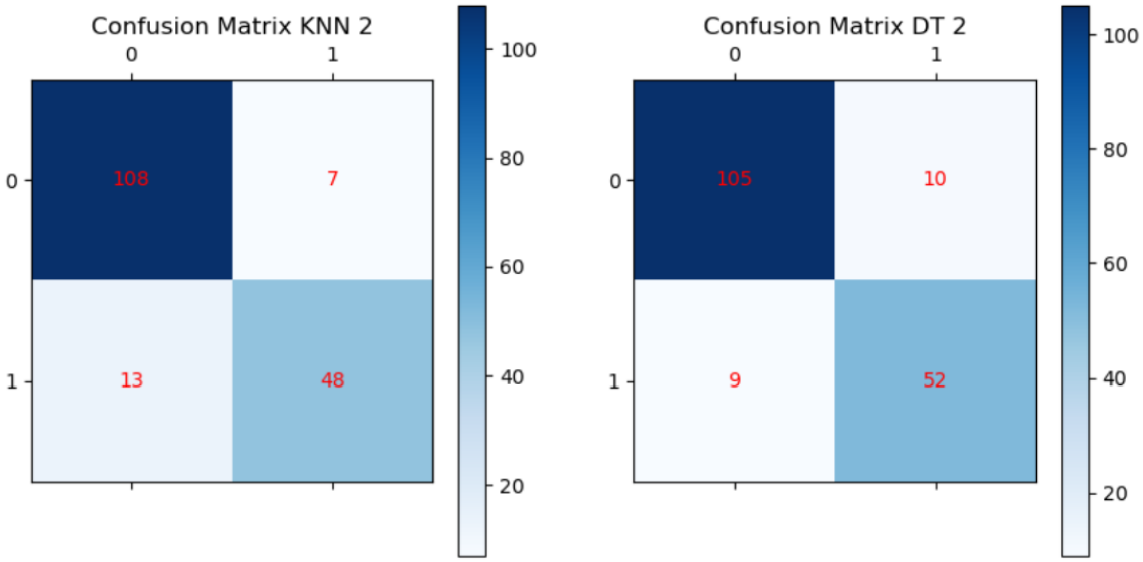
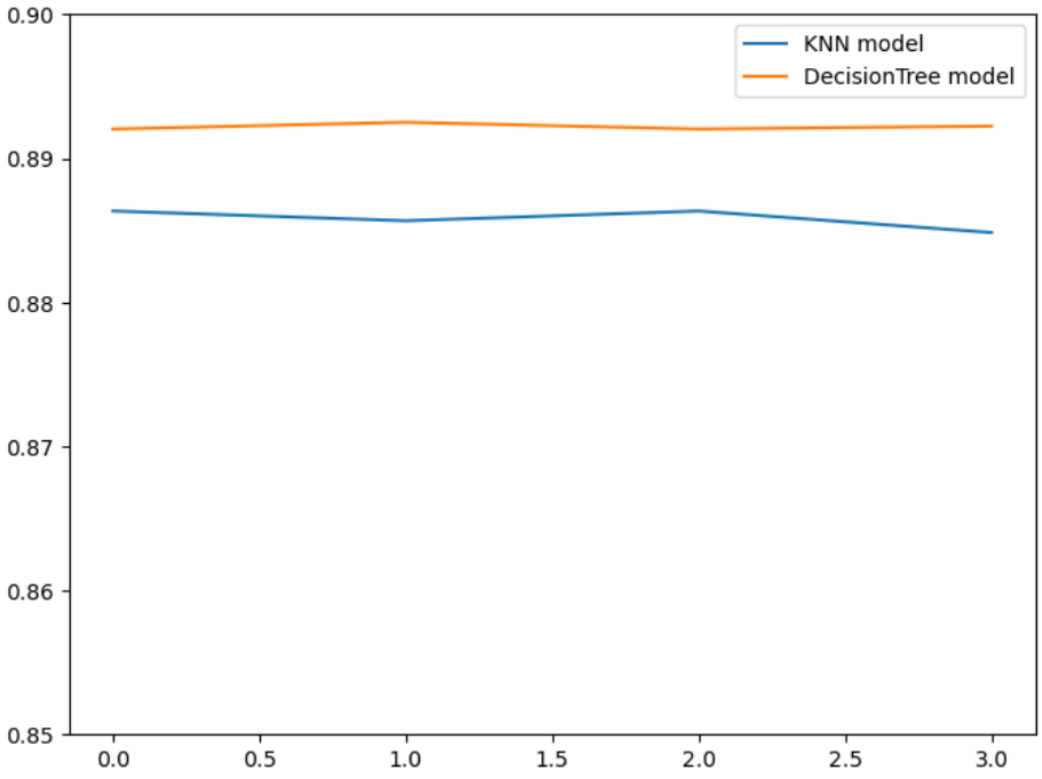


Figure 3.4. Different between the other evaluation value:



In the second suite, both models also provide fairly similar prediction results.

<u>On Suite 3:</u>

Figure 3.5. Confusion matrix of 2 model:

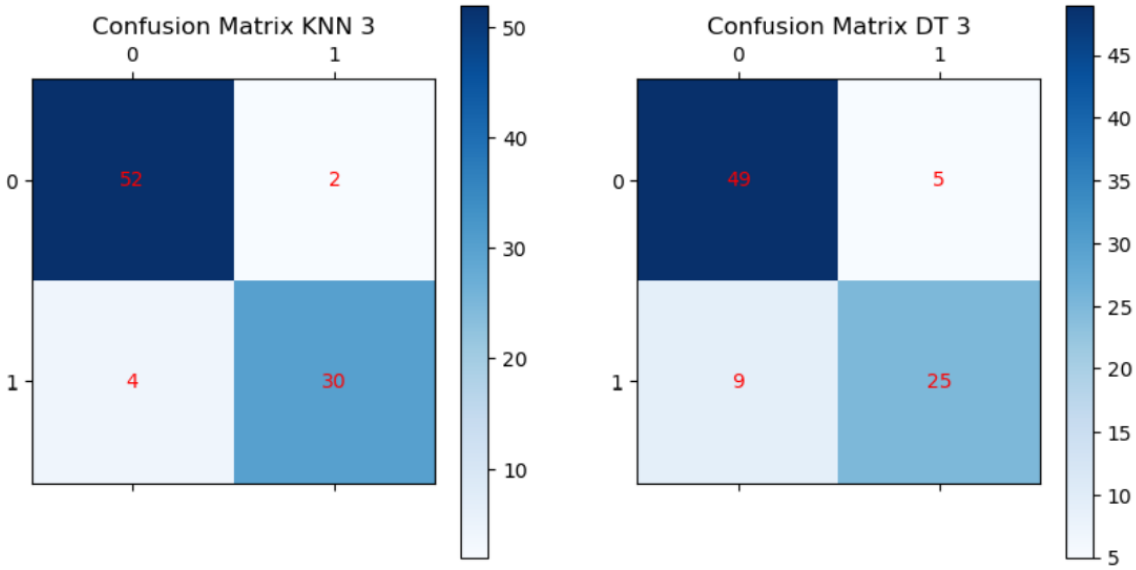
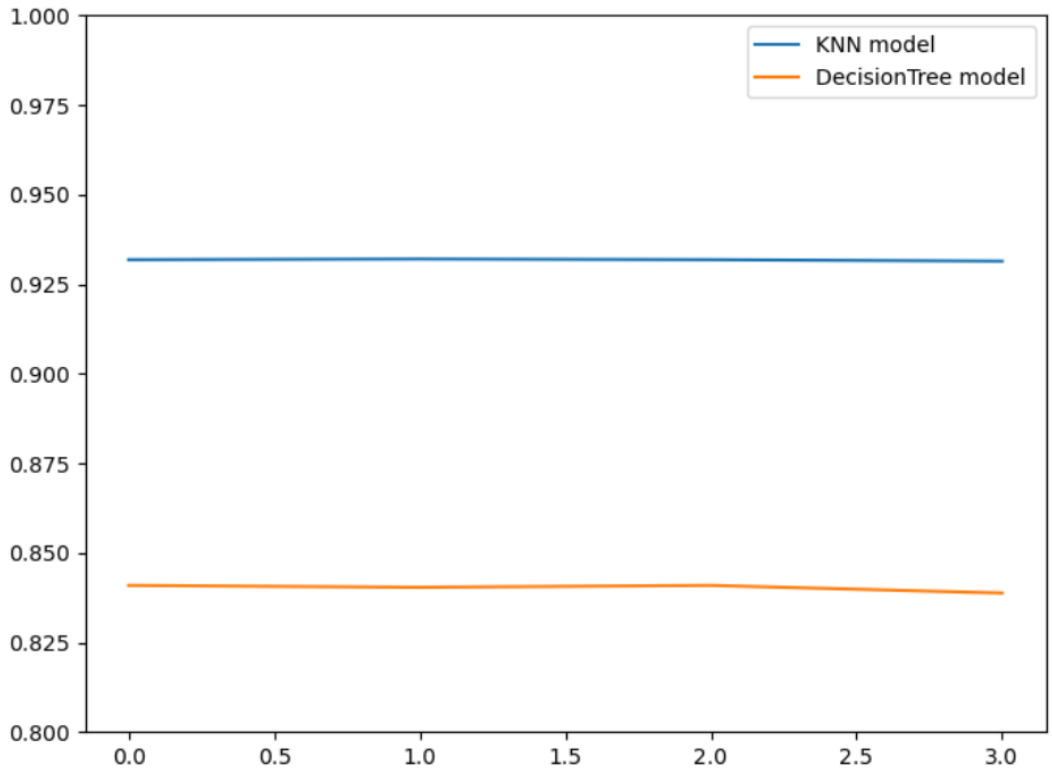


Figure 3.6. Different between the other evaluation value:



In the third suite, there is a difference in the prediction results between the two models. The KNN model provides very high prediction results, which may indicate overfitting. Conversely, the DT model yields lower prediction results compared to its performance on the other two datasets, suggesting potential underfitting.

In addition to comparing the results between 2 models in the same data suite, the model evaluation results are also visualized among 3 suites.

Figure 3.7.

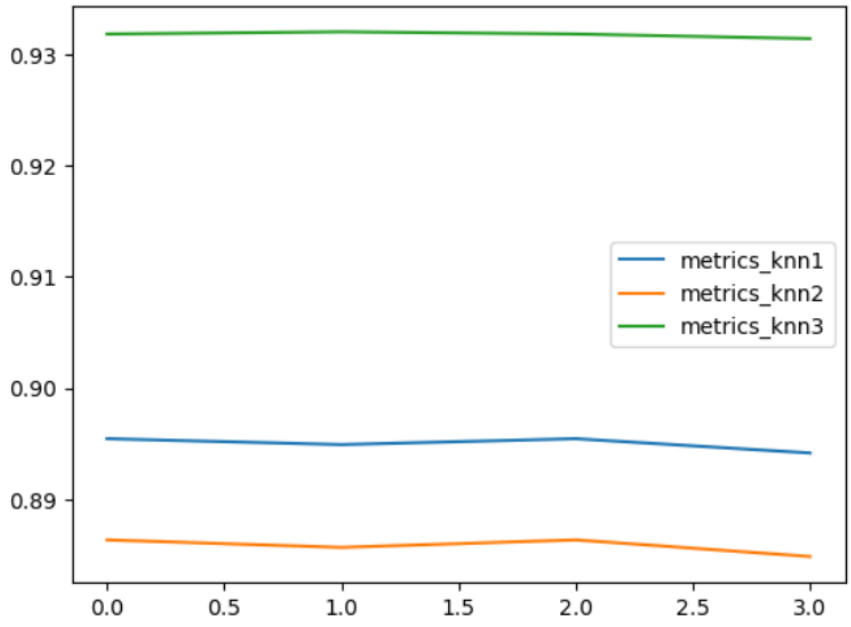
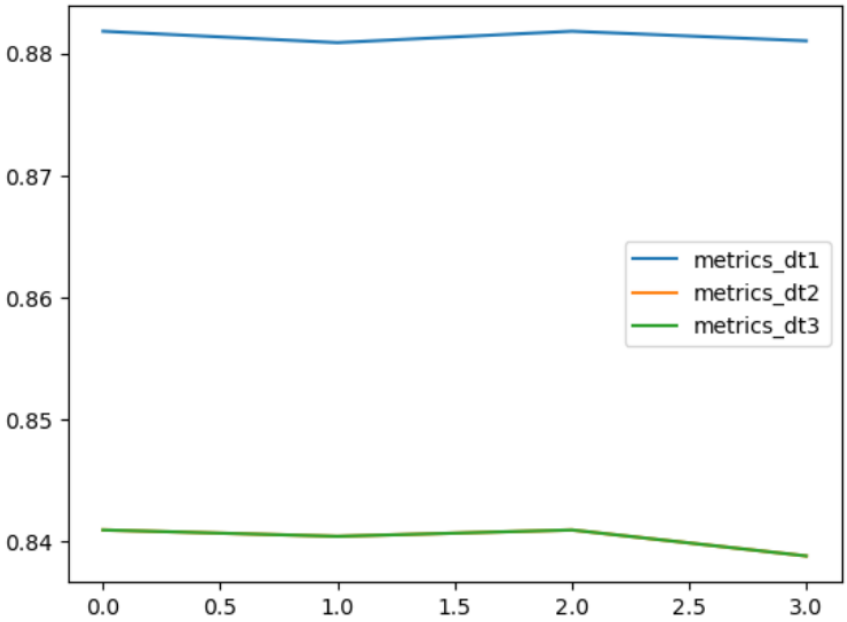


Figure 3.8.



Conclution

The report has provided thorough analysis of the data in the wholesale customer dataset, along with predictions based on data analysis models. These findings have addressed stakeholders' business questions, aiding in future strategy development and decision-making.