National College of Ireland

A hybrid approach towards identifying optimal prices by segmenting customers using active and inactive criteria

MSc Research Project
Data Analytics

Tejali Gangane
Student ID: 21148872

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Tejali Gangane
Student ID:	21148872
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Qurrat Ul Ain
Submission Due Date:	14/08/2023
Project Title:	A hybrid approach towards identifying optimal prices by segmenting customers using active and inactive criteria
Word Count:	6,795
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	17th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A hybrid approach towards identifying optimal prices by segmenting customers using active and inactive criteria

Tejali Gangane
21148872

Abstract

Everchanging customer behaviour and meeting the demands of such dynamic customers have been the constant focus of the business to make themselves customer-centric. Along with that, analysing the factors affecting the purchase of a product and the price for such products is also important. The traditional methodology of setting optimal prices makes use of mathematical models. Few studies work towards the usage of customer segments for price prediction. Therefore, the study addressed in this report works towards further exploration of the usage of customer segmentation along with customer lifetime value (CLV) for generating optimal prices for each product and performing product segmentation as well. This study segments the customers based on the active or inactive state and generates CLV for each customer and uses the information to set up optimal prices. The study tweaks the formula for CLV with the addition of LoyaltyRate. The product segmentation is performed using a TF-IDF vectorizer and K-means clustering. The results show that, with the usage of such segmentation criteria, the optimal prices help to generate higher revenue for the businesses.

1 Introduction

Retail stores have existed for generations. Online retail stores came into existence around the 80s to 90s. The e-commerce sector witnessed exponential growth in the 2010s. During this time, the sales of online shopping businesses went over a billion dollars. Ever since then, the online shopping industry kept growing.

The circumstances surrounding the pandemic, restricted physical mobility and a heightened emphasis on remote transactions contributed to an amplified range of choices available to consumers in online shopping. In the retail industry, it is evident that as the e-commerce sector expanded and matured, a parallel progression led to the demands and behaviours of customers. Therefore, it is necessary to comprehend these patterns depicted by the consumers and their approach towards the purchases of products. Deciphering the factors affecting these will help businesses to make effective decisions concerning the products and their prices. Along with that, since the relationship between a retail business and a customer is not subscription-based, that is, there is no guarantee or information regarding the timeline until which the customer will stay associated with the business. If the concept of business is subscription based, one can know the period of time the

customers will remain with the business, ahead of time. As for subscription-based businesses, the customer is bound to the company for a certain period of time. Because, of this, it's difficult for businesses to analyse the information in a way which guarantees the timeline of a customer with the company. Therefore, in this scenario, it becomes important to work towards comprehending customer behaviour to understand the needs, demands and purchase patterns of the customers.

Retail businesses are responsible for offering products aligned with customer preferences and meeting their needs Yang et al. (2022). By actively identifying and comprehending evolving demands and preferences of their target market, retailers can tailor their product offerings to align with customer expectations. However, it is necessary to acknowledge that various factors about the products themselves play a significant role in their purchase. While retail businesses strive to offer products that align with customer preferences, certain inherent qualities, features, and product attributes can heavily influence consumer decision-making. Factors such as product quality, functionality, design, pricing, brand reputation, and perceived value all contribute to the attractiveness and desirability of a product in the eyes of customers. Therefore, the purchase intention of the customer has been important due to the ever-changing demands and needs of the customers Rakib et al. (2022b).

The need to prioritize the satisfaction of customers remains at the top, however, ensuring that businesses remain profitable at the same time Dehnavi et al. (2023). Balancing these equations will assist enterprises in being financially stable and also keep their customers happy.

Sun et al. (2023)'s study provided the idea of working towards understanding customers that are lost and gained and understanding the reason behind the same. Their study provided this as a scope of future work. This study works towards understanding the behaviour of customers and the factors that influence these customers to make purchases Rakib et al. (2022a). The customers addressed in this study are based on the extension of Sun et al. (2023)'s work. However, with a slight modification and addition of a few other techniques. The segmentation method used in this study is that of the customers being in an *Active* or *Inactive* states.

Based on these purchases the businesses earn money. Therefore, understanding ways in which the profits can be maximized is also part of this study. To ensure maximum profits for the business, optimal prices need to be set for the products. This study provides a hybrid approach to achieve price optimisation. It deals with the performance of customer and product segmentation along with the estimation of CLV. And uses the insights from these activities to generate optimal prices.

1.0.1 Research objectives

- To perform **customer segmentation** to understand factors affecting customer behaviour
- To perform **product segmentation** understand the similarity between products
- To compute **customer lifetime value** to understand customer's behaviour.
- To compute **optimal prices** by taking customer segments into consideration.

1.0.2 Research Question

How does the usage of segmentation techniques for customers and products along with the estimation of CLV (Customer Lifetime Value), help in setting optimal prices that could lead to higher revenues for retail industries?

2 Related Work

Understanding the demands of customers holds significant significance, especially in the retail sector, where businesses must give top priority to customer contentment in order to effectively fulfil their needs. Unlike subscription-based companies, which have contractual obligations with customers, retail enterprises need to focus on comprehending customer behaviour and preferences to create a customer-centric approach. It is essential for the business's development to identify the target customers and determine the ones that bring profitability and value to the company. Along with that identifying ways to retain customers is also an essential task for businesses to ensure customers remain loyal to them and the businesses execute smoothly.

A few critical tasks concerning a business's aim to retain customers and optimizes product prices are customer churn prediction and segmentation. Apart from that, this activity will also help companies to build marketing strategies. According to the recent literature, machine learning algorithms, explainable artificial intelligence (XAI) and graph neural networks would be considered a few beneficial approaches towards these tasks. This literature review will discuss different studies that use these techniques to address the problem of segmenting customers and their effect on pricing strategies.

This problem was addressed by Meng et al. (2023), Prabadevi et al. (2023), Sun et al. (2023), Hartoyo et al. (2023), Hu et al. (2023), Liu et al. (2022), Anitha and Patil (2022), Ljubičić et al. (2023), Vázquez et al. (2023) in their research.

Sun et al. (2023) performed the research concerning the segmentation of customers, and focused on grouping the customers specifically based on their life-cycle stage. Similarly, Anitha and Patil (2022)'s study aimed at identifying potential customers and maximizing sales for a specific area by making use of their historical purchase data and making the use of business intelligence. However, both of these studies made use of segmentation techniques called RFM (Recency, Frequency and Monetary Value). This involves the analysing of customers using the information for RFM and further categorizing them into segments.

The study conducted by Meng et al. (2023) makes use of adaptive clustering techniques to group the customers on the basis of usage patterns of electricity. The main objective of this research was to produce customized pricing strategies and to generate insights concerning ways in which electricity was utilized by the customers.

Looking into the study conducted by Hartoyo et al. (2023) based the segmentation activity on the appreciation of Indonesians for different automakers. The goal was to understand the influencers that result in the admiration by customers for different automakers and the way in which customers stay aware of the information related to brands. Their study also highlights the importance of the connection between the brand and the customer. Using these admiration-based segments, automobile firms can use this information for marketing and decision-making purposes.

The concerns for the business are not limited to customer segmentation. It further extends towards understanding the behaviour and influencing factors for leaving and

ensuring such customers are provided with personalized offers so they could be retained.

Prabadevi et al. (2023)'s study works towards the analysis of customer churn that is the customers that stopped making purchases from a business. Their study uses various machine learning algorithms to identify such customers. Their aim was to identify customers who are currently active in making purchases from the company and understand their behaviour to identify which one of them might leave in the future. Their study used various techniques to make such predictions. The techniques they used produced good results for accuracy which ranged from 78% to 83%.

Liu et al. (2022)'s study made the usage of pricing practice in an M/M/1 system to understand the patience of the customers. Their study provides insights concerning price adjustments performed by the service provider when the customers showed patience with respect to the waiting time. Their study worked towards the objective of finding a perfect combination of price hikes and discounts that can impact the patience of the customer. The results of their findings suggested the offering of customized services and consideration of patient threshold for clients to increase the income of the providers.

Another important factor that businesses need to consider is the independence of customers. Understanding whether a customer's decision is influenced by another customer can be an important factor towards customer segmentation.

Ljubičić et al. (2023)'s research talks about the prediction of a customer leaving using techniques that ensure the independence between customers with the help of GNNs (Graph neural networks). Their study focuses on the findings that customers that are within the same social circle as a churner have a higher chance of churning. This finding, therefore, challenges the assumption of customer independence. To deal with this issue, their study proposed a unique approach that makes use of GNNs to model the independence of customers with the help of a customer network. By making use of this methodology, the study helps with an enhanced way of assessing the probability of churning. This approach helps to capture the complex relationships between the customers and improve the accuracy for predicting churning customers.

Businesses need to understand the behaviour of the customers to ensure the evaluation of product performance, happiness of customers and overall profitability. This includes finding out the factors that influence the customers purchasing decisions. The informational insights derived from this will help the companies to make strategic decisions concerning the development of products and help them in effectively targeting suitable customers along with the necessary product information Holý et al. (2017). By understanding the behaviour of customers and the factors responsible for their product choices, companies will be able to make informed decisions to improve their product offerings and help customize their marketing efforts to reach the target customer.

To group the products together, Kondo and Okubo (2022) and Holý et al. (2017) used various different techniques. Depending on the behaviour of the customers across different channels, the items and the categories of products were grouped in the study conducted by Kondo and Okubo (2022). The segmentation strategy used in their study considers traits and the patterns of purchase performed by consumers for each product category. In comparison the study conducted by Holý et al. (2017), groups the product goods into clusters on the basis of the frequency of them appearing together in the shopping carts. The strategy used in the study works towards grouping the products as alternatives, where some of the clusters are formed of products which aren't most commonly found in the same baskets together.

Vázquez et al. (2023) clarifies with the confirmation that previously conducted per-

forms categorization of customers into different categories on the basis of the way they weigh with respect to certifications and the factors of sustainability while making the decisions for buying.

Guo et al. (2023)'s study talks about the increasing flexibility for energy systems and verifying the cost is reduced to ensure a balance is maintained between the demand and supply, their study built a price optimization model which was virtually concerning the electricity system. This model is based on the mechanism of credits. Their results depicted that, this virtual model for optimization of prices which was based on the mechanism of power credit was able to reduce the overall cost of power by around 7.8%.

Li et al. (2023)'s study highlights the cruciality of multi-product optimization of prices. It suggests that this is essential for decisions related to firms' operations. When multiple alternatives were faced, discrete choice models were applied widely for describing the choice behaviours of consumers.

Sridhar et al. (2021)'s study works towards the domain of inventory management specifically in the sector of retail. Their study built a model using a simulation software called Arena. They proposed an effective system with the help of experiments with the model along with the alteration of characteristics of the model. Their proposal for this system clarified that the model will be able to lower the level of inventory by 40% and the lost sales by 87%. They further optimized the system with the help of a module in the Arena called OptQuest.

After careful consideration and analysis of previous work, it was observed that most of the studies focus directly towards the generation of optimal prices. The literature lacks information about groupings of customers based on their states of being active or inactive. The research addressed in this report works towards filling these gaps and providing a hybrid approach for predicting prices.

3 Methodology

This report deals with the understanding of the implementation of price optimization strategies using customer behaviour, segmentation of products and the customer lifetime value. The research problem addressed in this study is assessing the effects of the behaviour of customers on the prices of different products. By understanding and identifying the factors that drive this behaviour, analyzing the relationship between customer behaviour and their product choices and the impact of customer lifetime value on the sales of products, this research will contribute by providing a blueprint of strategies that will help businesses use efficient methods for adjusting product prices

This research aims to assess the effects of customer behaviour for products based on their prices. This will intend to help in providing insights concerning ways to strategize pricing methods.

3.1 Approach

3.1.1 Study

The inclusion of the concept of customer segmentation for price optimization is underresearched in academic literature. The need for price optimization in businesses, specifically in the retail industry has been growing for the past couple of years Kirsch (n.d.). As there is a significant gap in grasping the advantages of this inclusion.

Even though, price optimization and customer segmentation are individually studied topics, the ways and benefits of their integration for developing a pricing strategy are yet to be explored. Most of the literature works towards individually studying these and ignores the added advantages of their usage together. Thereby, the availability of such opportunity to investigate or research more into this topic.

To achieve this, the study makes use of Quantitative data and performs analysis using different techniques like descriptive, correlation and experimental research. The data consists of various transactional records for a certain period of time. This includes the products brought in each transaction and the amount of products bought. The analysis will be performed to get an overview of the data.

- The descriptive approach helped in getting the summary of the data and variability present in the data.
- The correlation approach helped in understanding the effects of one variable over another.
- The experimental approach helped in understanding the cause-and-effect among the available variables.

This data is obtained from a secondary source. UCI ML repository provides similar datasets for the usage of different machine learning projects. The said dataset was used by different researchers and has various published literature where this data was used for various analyses.

3.1.2 Need

Customer segmentation is a key part of every retail business to ensure their business is customer-centric. This is done to ensure all the needs and expectations of customers are met. This segmentation can be performed on the basis of different characteristics like age, geography, spending pattern and so on. This way of grouping customers based on various factors will help in understanding customers that are similar to each other. This will further help businesses to target customers with specialised or customised offers based on their requirements and needs. Apart from that, understanding which group of customers are willing to spend more will help in devising high revenue-generating strategies.

3.2 Data Collection Methods

3.2.1 Data Source

The data has been taken from the UCI ML repository *Online Retail* (2015). This repository provides various datasets for budding machine learning engineers or data scientists to test out various algorithms. One such dataset was used for this study.

The dataset used here is **Online-Retail** dataset. The data is about an online retail store in the UK. It holds more than 500 thousand records of information related to transactions made for the year 2010-2011. The products available at this store are gift items. There are more than 4000 unique customers that made a purchase using this store.

This dataset provides enough information to perform the objectives of this study. The information regarding the invoices for each transaction and the products that were

purchased will help perform customer segmentation. This will further help in the identification of distinct groups among the customers. Furthermore, analysing this transactional data will help in exploring various patterns and similar characteristics among the customers.

The dataset also consists of information about the products like the stock code for each product and its description and unit price for each product. This will further be helpful for product segmentation. Understanding products demand patterns and their profitability and contribution towards revenue will also be essential.

The insights generated from the combination of customer and product segmentation will work as a base for achieving the objective of price optimization. Overall this dataset consists of all the attributes required to achieve the goals. This dataset is rich in information and will help in gaining insights into customer behaviour, product preferences and also be helpful in building pricing strategies.

Attributes	Description
InvoiceNo	Every transaction has unique identifier. If this attribute has suffix 'c', then the transaction represents a cancelled one
StockCode	Unique identifier for a product
Description	Name of the product
Quantity	Number of items purchased for a particular product
InvoiceDate	Date of purchase
UnitPrice	Price of one item
CustomerID	Unique identifier for a customer
Country	Place where the transaction was made

Figure 1: Data Dictionary

- InvoiceNo represents a unique transaction number for every order. The values prefixed with a 'C' are cancelled transactions.
- StockCode represents the unique identification value for the product. This is used to differentiate between products and can be used in place of the description of the products for easier identification.
- Description holds the name of the product.
- Quantity indicates the number of items that a particular product was ordered for
- InvoiceDate indicates the date the purchase was made.
- UnitPrice indicates the price of the product for a single item.

- CustomerID indicates the unique identification value for a customer.
- Country indicates the place where the transaction was made.

3.3 Software used

- PowerBI - This is a visualization tool. This tool was used to perform the descriptive analysis of the data. This tool allowed easier loading of the data. And had pre-existing charts, that allowed me to explore different variations in the data.
- Colab - This allows you to write Python code easily in the browser itself, without having to do any configurations beforehand. There, its usage allowed me to save time or configuration and focus my energy and resources on other parts of the project.

3.4 Proposed methodology

Figure 2, indicates the activities that will be performed as a part of the research. These activities are further elaborated throughout the report.

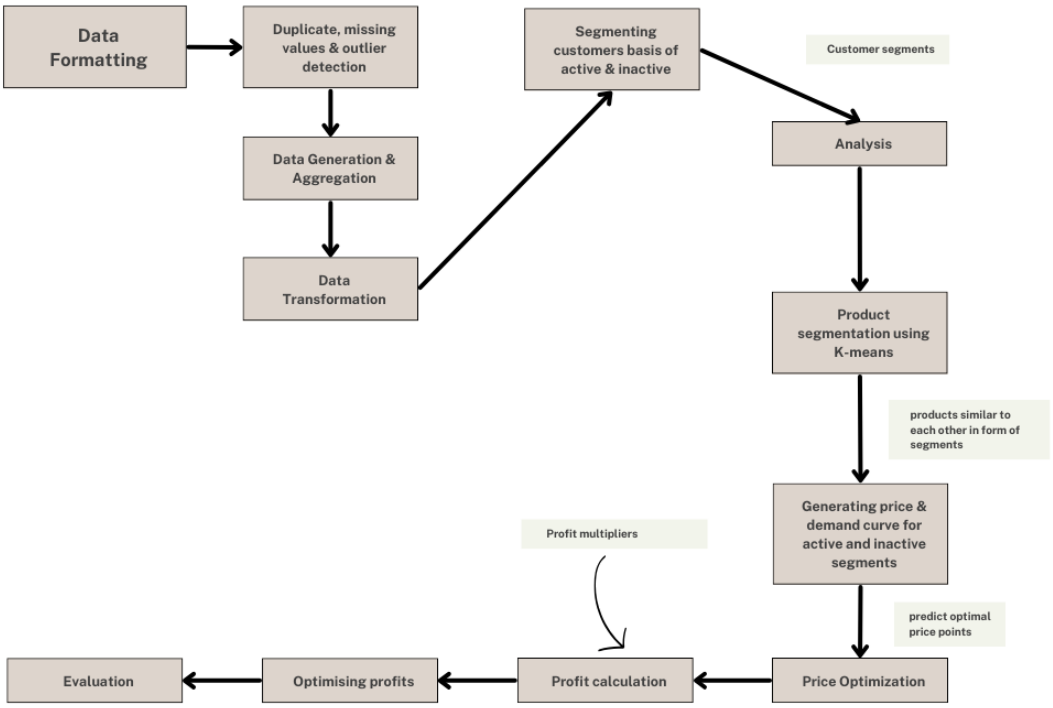

Figure 2: Proposed flow of the research

3.5 Processing and Analysis of Data

3.5.1 Data Preparation

Data preparation is an essential phase in the process of analysing data. Making sure that the raw data is transformed and cleaned in a consistent and structured way will help generate meaningful and accurate insights. For this dataset, several steps were taken to ensure the data was righteously processed for further usage.

3.5.2 Data Generation

There were several steps that were taken into consideration for the generation of data. This was done to ensure the data is polished and made available in a way useful for the study.

3.5.3 Data Seggregation

The original data had records for transactions that were cancelled and transactions that were successful. These cancelled transactions were denoted by prefixing C in as a part of InvoiceNo. Therefore, the cancelled and successful transactions were separated.

3.5.4 Data Aggregation

The data available was aggregated to get an overview of the data. It helps to get a broader view of the data and helps generate insights in a way which will help the businesses to get a deeper understanding. The process of data aggregation using the concept of RFM technique will help to perform segmentation based on the monetary involvement of the customer and will help to get an overview to understand the spending behaviour of the customer. This will further be helpful in understanding the behaviour of customers and generate insights which can help in making decisions.

3.6 Techniques used

3.6.1 Customer Segmentation

The data generated from the previous stage of CLV estimation was used for the segmentation of customers. The segmentation was done on the basis of active and inactive customers(as previously stated). This type of segmentation was done to differentiate and understand the purchase behaviour of customers and their transition between active and inactive states.

3.6.2 Product Segmentation

The data generated to hold unique product information from previous stages was used as input to segment products. This process will help in tailoring any marketing efforts related to product offerings to better meet the needs and demands of the customers. Anitha and Patil (2022)'s study suggests that the usage of the k-means algorithm is beneficial for grouping as it only takes a singular parameter when compared to other algorithms.

3.6.3 Price Demand Curve

The price demand curve will be built to show the relationship between the price of the product and the quantity associated with it. Along with that, these data points will be visualized. The slope of the demand curve can also help reflect the elasticity of the price of demand. Which indicates how the changes in price causes changes in quantity. Kedia et al. (2020)'s study suggests that the demand curve helps to understand product demands for certain price points. This will further help in understanding the elasticity of the product. Their study further explained the relationship between price and demand using an equation -

$$Ed = \frac{\Delta Q}{Q} . \frac{P}{\Delta P}$$

Figure 3: Formula for price demand elasticity

3.6.4 Price Optimization

Predicting optimal prices for the products is an important part of this study. The setting up of optimal prices for the products was performed by taking into consideration previously created segments called *Active* and *inactive*. To achieve the inclusion of customer segmentation in the decision of optimal pricing, below are the steps that were performed -

- Calculation of profit multipliers
  - The profit multipliers are the factors which are applied to the base profit value for a product which can help in calculating the potential profit for various customer segments. It is used based on the characteristics and behaviours concerning various customer segments.
  - It helps businesses to tailor the prices for different products by taking into consideration the segments specific to their audiences.
  - Here, they were used to understand how less or more profitable a particular product will be concerning a specific segment.
  - The profit multipliers are calculated on the basis of past data. The purchase amounts were analysed for all the products and taken into consideration the insights from customer segments.
- Profit calculation
  - This represents the objective which needs to be optimized. It defines the maximum profits that need to be achieved by adjusting the prices for various customer segments.
  - The calculation of profit will be based on the price that will be set and the characteristics of the segments.

4 Design & Implementation

4.1 Overview of the data

Below is the overview of the data through visualization. Here are some insights regarding the same -

- The business generated more than eight million in revenue.
- The business provides its products for ordering in up to 37 countries.
- There are more than eighteen thousand customer-specific transactions that had taken place.
- Growth in revenue can be observed throughout the year, where there seemed to be a higher jump than usual from Quarter 3 to Quarter 4.
- Number of customers grew by a few hundred over every quarter. And the business also witnessed a jump of more than seven-hundred customers from Quarter 3 to Quarter 4.

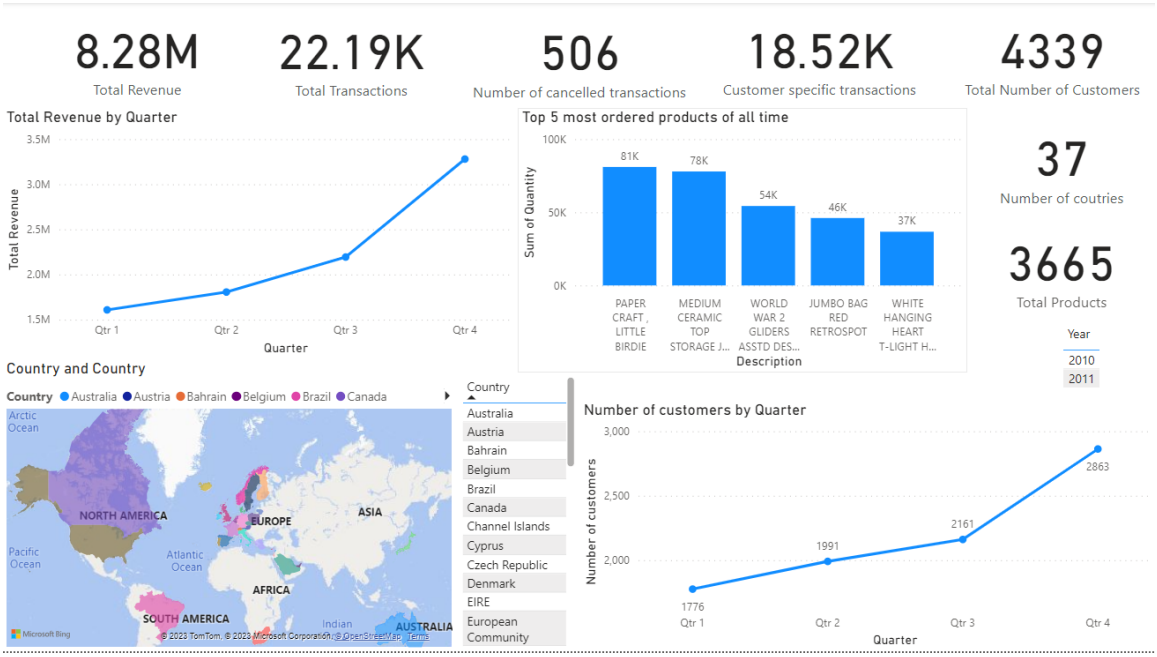

Figure 4: Overview of the data

4.2 Processing and Analysis of Data

4.2.1 Data Preparation

First of all, all of the rows where the *CustomerID* was not present, were removed. Since the data consists of information concerning the transactions that occurred within a certain timeline and assuming that these transactions involved a purchase made by the customer; having no CustomerID associated with the transaction made no sense for the research

that is being conducted. Hence, removing such rows will help clear out the unnecessary clutter.

Further, the *InvoiceDate* column which contained date and time information in string format, was converted to a Date datatype. This was done to ensure correct processing of the date information when further analysis is done. The date components were extracted from this column to only store the date information. This step helped in further performing analysis of understanding how sales take place throughout the dates.

Ensuring accurate geographic-related insights, a comprehensive check for values concerning the *Country* attribute was made. Specific rows containing the value "Unspecified" were removed from the dataset. This was done to ensure, the analysis related to the customer's living arrangements was accurately addressed.

To prevent the redundant values from skewing the analysis presented from the data, the duplicate values were removed. Along with that, a check for missing values was performed. The results indicated, that after previous activities, there were no missing values remained as such.

4.2.2 Data Generation, Segregation & Aggregation

- **Data Generation**
  - First of all, an attribute called *AmountSpent* was generated which resulted as a product of *Quantity* and *UnitPrice*. This value represents a monetary expenditure associated with an order placed for a product based on the quantity. Thereby, helping in understanding the amount of money earned in a transaction for a product.
  - Multiple sub-datasets were created to hold customized specific data pertaining to different information. A dataframe was generated that held information about the unique products concerning data about the *StockCode* and *Description*. This was done to get a better idea about the individual products.
  - Furthermore, another dataframe was created to store the history of prices for each product. The attributes present in this record are *SotckCode*, *Description*, *UnitPrice*, *InvoiceDate*. This data provides an overview of the changes in price that took place for each product. It helps in understanding the price fluctuations that take over time.
- **Data Segregation**
  - Segregation cancelled and valid transactions

    The original data consists of transactions that were cancelled by customers. There the resulting values in columns like Quantity or UnitPrice were denoted by a negative value. Performing aggregation and modelling on such data can provide improper results. Therefore, it is necessary to filter out this data. To achieve this, transactions having the InvoiceNo beginning with a 'C' were filtered to be cancelled transactions and the rest of them were deemed to be valid transactions.
- **Data Aggregation**
  - RFM

* RFM data was created by aggregating the original data. RFM indicates *Recency, Frequency and Monetary Value*. The frequency indicates the number of transactions made by the customer in a particular timeline. Recency indicates the number of days since the customer last made a purchase. Monetary Value indicates the amount of money spent by the customer over the lifetime associated with the company.
* Recency will be curated with the help of the attribute *InvoiceDate* from the original data. It consists of the calculation of the difference in days from a day ahead of the latest transaction date and the date since the customer last transacted. The newly created attribute *LastPurchaseDays* represents the Recency value.
* Frequency will be the result of the aggregation of the number of transactions that occurred for each user. With the help of a unique count of *InvoiceNo* for each customer, this attribute was generated.
* Monetary value will represent the financial aspect concerning each customer. It depicts the aggregation of the money spent by the customer for every transaction that they made.

– Customer Purchases

To get further clarity on the customer transactions, a data table was created to store information about the existing transactions made by customers based on monthly intervals. The attributes of this data were depicted by month and year as a single attribute name. These months and years ranged from the first transaction date present in the dataset to the last transaction date. The resulting values were binary indicating whether a customer made a transaction in a particular month of the respective year.

– Customer Activity Attendance

Another data table was created to indicate the purchase activity of the customer. This was derived from the *Customer Purchases* table. It showed the unique number of times a customer showed up throughout the timeline of the dataset. These values are indicated by the column *AttendanceThroughYear*

– CLV estimation

In the context of this study, another additional factor was included called *LoyaltyRate*.

Below are some of the steps taken to compute the CLV.

* Initially, the first and last purchase date of the customer was identified and later this information was merged together. Thereby, the result was a dataframe consisting of unique customer IDs along with the information regarding the date of the first and last purchase they made.
* Further, the difference in months between the first date and the last date was calculated. Later, the resultant dataframe was merged with the data generated from customer activity.
* Then the *LoyaltyRate* was calculated based on the division between *AttendanceThroughYear* and the difference in months. This was done to ensure that the loyalty rate was calculated based on the customer's lifetime of existence and not based on the company's existence. This factor indicates the frequency rate of the customer's interaction with the company. Every

customer begins their interaction with the company at a different point in the company's lifetime. Hence, the judgement criteria for understanding how often a customer purchases with the company varies. Therefore, defining a formula for LoyaltyRate that takes into consideration the first and last transactions made by the customer for a company is beneficial for providing a more granular analysis.

* Finally the calculation of CLV was performed as the product between the values of *AmountSpent*, *LastPurchaseDays*, *AvgBasketSize*, *LoyaltyRate*.

4.3 Modelling Approach

4.3.1 Customer Segmentation

Segmentation of customers was performed to group the customers in a particular way for further analysis. The goal of the study is to incorporate the segmentation of customers in the decision-making of optimal prices for the products.

- The cut-off date was set as the 1st of September 2011.
- The filtration was done based on all the customers whose *FirstPurchaseDate* was before the cut-off date.
- Furthermore, segments were defined based on the *LastPurchaseDate*. If the transaction date was later than the cut-off date, the customer was set to be inactive and if the date was earlier than the cut-off date, the customer was said to be active.

4.3.2 Product Segmentation

Product segmentation was done to understand the similarities between the products. Descriptions of the products were used to group the products. This was done using a TF-IDF vectorizer and calculation of cosine similarities. Based on these cosine similarities and providing it as an input to the k-means clustering algorithm, the clusters were formed.

To perform product segmentation -

- Stored unique product descriptions in a variable.
- Initialized a TfidfVectorizer
- The product descriptions were transformed into TF-IDF vectors to represent the text data in the form of numerical information. This was useful in the measurement of the importance of every word.
- Cosine similarities between the descriptions were computed. This was performed as each description of the product consisted of information about the product's characteristics and features. This allowed us to understand the similarity between the descriptions of each product. The result was a similarity matrix.
- Further, clustering was performed on this similarity matrix using K-Means clustering. The number of clusters was chosen as specified in the design specification. Anitha & Patil (2022) described the usage of the K-means algorithm for the grouping of customers. Their research underlined the key importance of this algorithm for the process of segmentation. The reason is just a single parameter as input for the algorithm, which represents the clusters needed for the grouping criteria.

- The dataframe used to store unique product prices was merged with this segmented product data.

4.3.3 Price Demand Curve

The price-demand curve was generated for the data to understand the relationship between demand and price. The computation of the price and demand curve was performed as follows -

- Grouping the valid transactions by *UnitPrice* and aggregating the data based on the sum of Quantity.
- Furthermore, a line chart is built based on the data.

4.3.4 Price Optimization

Prediction of optimal prices by making use of insights from segmentation was performed by the calculation of profit multipliers and calculation of profits. Identification of reference segment was performed to calculate these profit multipliers

- Calculation of profit multipliers

  To compute the profit multipliers, below are the steps.

  - Calculating the average spent for each segment.
  - Identifying the segment that has the highest average amount spent. This segment was used as the reference segment. This is done to ensure relative profitability and use it as a comparison base for another segment.
  - The profit multipliers are computed as the division of the average amount spent by the average amount spent for the reference segment.
  - In concern to the study addressed in this report, there exist just two segments. These are active and inactive. Therefore, one of the segments which have the highest average amount spent had a profit multiplier of 1.

- Profit calculation

  - To perform the calculation of profits
    * Two things are required - the price and the customer segment
    * Initially, the profit base is calculated which is a product of the price provided and the constant pre-defined value for the base profit multiplier.
    * Calculating the profits as a product of the base profit and the segment multiplier value which is dependent on the segment value.

- Optimizing profits The bounds were set and provided as inputs for the optimisation function. These values for bounds were based on the minimum and maximum values available for the products. In scenarios where there existed only one UnitPrice for the product, the full price was considered to be an addition of 1 unit to the minimum price.

5 Evaluations & Results

5.1 Evaluation

5.1.1 Customer Segmentation

Fig. 4 indicates the segmentation that was performed on the data. These segments were "Active" and "Inactive". It can be observed that there were more than three thousand customers that fall under the "Active" segment, whereas there are around 300 customers that fall under the "Inactive" segment.

Segment	CustomerID
Active	3958
Inactive	341

Figure 5: Count of customers in each segment

5.1.2 Product Segmentation

Fig. 5 indicates the results of the Silhouette-score analysis that was performed for product segmentation using K-means clustering. The computation of this score was done to assess whether the formation of clusters was performed accurately. If the score was closer to 1, it indicates that the clusters are far away from each other and if the score is closer to -1, it indicates that the products are assigned to the wrong clusters. The score generated here is closer to 0, indicating the products are assigned within the right clusters.

```
Silhouette Score: 0.073031188527108 31
```

Figure 6: Silhouette scores for product segmentation using K-Means clustering

5.1.3 Price Demand Curve

Fig. 6 indicates the price-demand curve which was generated for the online-retail dataset. It was observed that the shape of the curve was similar to the shape of the axis. This indicates that the demand is price elastic. This means a change in price can affect the demand as well.

5.1.4 Price Optimization

Fig. 7 indicates the optimal prices that were generated for *Active* and *Inactive* segments. It can be observed that certain products have similar prices for both these segments in comparison to the original price. This is the expected behaviour of the model, as

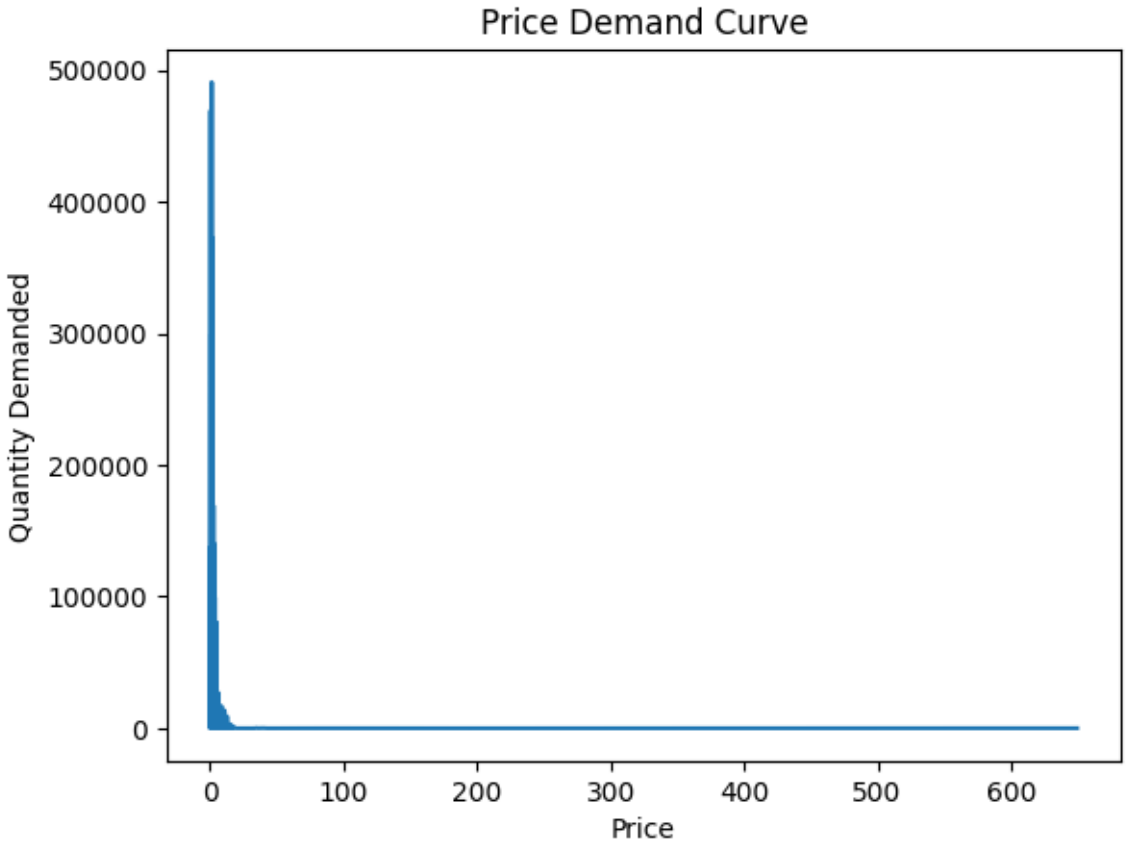

Figure 7: Price-Demand Curve

the chances are such products are already optimally priced. Hence, such results were generated.

5.2 Results

Based on the generated optimal prices for the two individual segments, the overall revenue generated from each product was computed. Looking at Fig. 8, column *OriginalRev* shows the revenue generated based on the original *UnitPrice*. The *PriceChangeActiveRev* depicts the revenue based on the Active segment. It can be observed that there is a significant difference between these two revenues.

Fig. 9 indicates the total revenue calculated based on the original unit price and the total revenue calculated based on the revenue generated for the Active segment.

6 Conclusion and Future Work

Generating optimal prices by taking customer segments into consideration has resulted in promising outcomes. It was observed based on the results that the optimal prices generated and when simulated for existing quantities, resulted in significantly higher revenues. It was also observed that certain products' optimal prices generated were similar to the original product price, this indicates these products were already at the optimal price level.

	StockCode	Active	Inactive	UnitPrice
0	10002	1.85	1.85	0.850000
1	10080	0.85	1.39	0.390000
2	10120	1.21	1.21	0.210000
3	10123C	1.65	1.65	0.650000
4	10124A	1.42	0.42	0.420000
5	10135	2.46	1.25	1.255000
6	11001	3.29	2.69	2.083333
7	15030	1.29	1.29	0.290000
8	15034	0.14	1.14	0.105000
9	15036	0.83	0.83	0.707500
10	15044A	5.79	3.95	3.763333
11	15056BL	12.46	5.95	5.912000
12	15056P	12.46	6.95	4.600000
13	15058A	7.95	7.95	7.350000
14	15058C	7.95	8.95	3.950000
15	15060B	8.29	4.75	6.020000
16	16008	0.25	0.25	0.100000
17	16014	0.42	1.42	0.320000
18	16015	0.50	1.50	0.430000
19	16045	0.12	0.12	0.040000

Figure 8: Price-Demand Curve

Therefore, these insights also provide useful guidance on understanding and identifying the products that are at the optimal level, so any further product improvements shall not be the main focus and can be considered as a background activity. Grouping the customers based on the category of being active or inactive will help the businesses tailor their marketing strategies uniquely based on the behaviour of each segment. This approach promotes the devising of personalized pricing strategies. This is not limited to pricing through segments but also takes into consideration grouping similar products together and later devising a pricing strategy.

Certain aspects of the study can be examined further in the near future. These are: (1) Although the study identifies products that are optimally priced, instead of keeping such products out of the spotlight completely, they can be monitored on the sidelines for any changes in demands or customer preferences. This will help the businesses stay prepared for any sudden changes that can occur concerning demands for any products. (2) The segments "Active" and "Inactive" can be further assessed on a granular level and more sub-segments can be formed. (3) These segments can be used to monitor the way the customers keep moving from active to inactive state and vice-versa. This will be useful as well for churn prediction, as every customer who stops making a purchase after a while does not indicate that the customer has left for good. Instead, its previous

	StockCode	Active	UnitPrice	OriginalRev	PriceChangeActiveRev
0	10002	1.85	0.850000	10.200000	22.200
1	10080	0.85	0.390000	4.680000	10.200
2	10120	1.21	0.210000	1.050000	6.050
3	10123C	1.65	0.650000	0.650000	1.650
4	10124A	1.42	0.420000	1.680000	5.680
5	10135	2.46	1.255000	12.550000	24.600
6	11001	3.29	2.083333	33.333333	52.640
7	15030	1.29	0.290000	1.450000	6.450
8	15034	0.14	0.105000	2.520000	3.360
9	15036	0.83	0.707500	8.490000	9.960
10	15044A	5.79	3.763333	22.580000	34.740
11	15056BL	12.46	5.912000	17.736000	37.380
12	15056P	12.46	4.600000	13.800000	37.380
13	15058A	7.95	7.350000	14.700000	15.900
14	15058C	7.95	3.950000	9.875000	19.875
15	15060B	8.29	6.020000	24.080000	33.160
16	16008	0.25	0.100000	2.400000	6.000
17	16014	0.42	0.320000	6.400000	8.400
18	16015	0.50	0.430000	4.300000	5.000
19	16045	0.12	0.040000	4.000000	12.000

Figure 9: Reveune comparison for predicted optimal prices concerning ACTIVE segment

```
Original Revenue based on the original UnitPrice =  7733.47975
Revenue after optimal prices were depicted for active segment =  10281.605
```

Figure 10: Total revenue before and after optimal price was generated

pattern of purchases can be used to identify whether the customer has become inactive or is about to leave.

References

Anitha, P. and Patil, M. M. (2022). Rfm model for customer purchase behavior using k-means algorithm, *Journal of King Saud University - Computer and Information Sciences* **34**(5): 1785–1792.

Dehnavi, M. N., Yazdian, S. A. and Sadjadi, S. J. (2023). Evaluating effective criteria on customer satisfaction using the best-worst method and optimizing resource allocation, case study iran aseman airlines, *Journal of Air Transport Management* **109**: 102375.

Guo, Z., Xu, W., Yan, Y. and Sun, M. (2023). How to realize the power demand side actively matching the supply side? ——a virtual real-time electricity prices optimization model based on credit mechanism, *Applied Energy* **343**: 121223.
URL: *https://www.sciencedirect.com/science/article/pii/S0306261923005871*

Hartoyo, H., Manalu, E., Sumarwan, U. and Nurhayati, P. (2023). Driving success: A segmentation of customer admiration in automotive industry, *Journal of Open Innovation: Technology, Market, and Complexity* **9**: 100031.
URL: *https://www.sciencedirect.com/science/article/pii/S2199853123001336*

Holý, V., Sokol, O. and Černý, M. (2017). Clustering retail products based on customer behaviour, *Applied Soft Computing* **60**: 752–762.

Hu, X., Liu, A., Li, X., Dai, Y. and Nakao, M. (2023). Explainable ai for customer segmentation in product development, *CIRP Annals* .

Kedia, S., Jain, S. and Sharma, A. (2020). Price optimization in fashion e-commerce, *arXiv:2007.05216 [cs, stat]* .
URL: *https://arxiv.org/abs/2007.05216*

Kirsch, K. (n.d.). The ultimate guide to price optimization.
URL: *https://blog.hubspot.com/sales/price-optimization*

Kondo, F. N. and Okubo, T. (2022). Understanding multi-channel consumer behavior: A comparison between segmentations of multi-channel purchases by product category and overall products, *Journal of Retailing and Consumer Services* **64**: 102792.

Li, L., Li, M., Zhang, H. and Zhang, L. (2023). Price optimization under the extended nested logit model, *Operations Research Letters* **51**: 54–59.
URL: *https://www.sciencedirect.com/science/article/pii/S0167637722001535*

Liu, J., Chen, J., Bo, R., Meng, F., Xu, Y. and Li, P. (2022). Increases or discounts: Price strategies based on customers' patience times, *European Journal of Operational Research* **305**.

Ljubičić, K., Merćep, A. and Kostanjčar, Z. (2023). Churn prediction methods based on mutual customer interdependence, *Journal of Computational Science* **67**: 101940.

Meng, F., Ma, Q., Liu, Z. and Zeng, X.-J. (2023). Multiple dynamic pricing for demand response with adaptive clustering-based customer segmentation in smart grids, *Applied Energy* **333**: 120626.

Online Retail (2015). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5BW33.

Prabadevi, B., Shalini, R. and Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms, *International Journal of Intelligent Networks* **4**. **URL:** *https://www.sciencedirect.com/science/article/pii/S2666603023000143*

Rakib, M., Pramanik, S., Amran, M., Islam, M. and Sarker, M. (2022a). Factors affecting young customers' smartphone purchase intention during covid-19 pandemic, *Heliyon* **8**(9): e10599.

Rakib, M. R. H. K., Pramanik, S. A. K., Amran, M. A., Islam, M. N. and Sarker, M. O. F. (2022b). Factors affecting young customers' smartphone purchase intention during covid-19 pandemic, *Heliyon* **8**: e10599.

Sridhar, P., Vishnu, C. and Sridharan, R. (2021). Simulation of inventory management systems in retail stores: A case study, *Materials Today: Proceedings* **47**.

Sun, Y., Liu, H. and Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model, *Heliyon* p. e13384.

Vázquez, J.-L., Lanero, A., García, J. A. and Moraño, X. (2023). Segmentation of consumers based on awareness, attitudes and use of sustainability labels in the purchase of commonly used products, *Sustainable Production and Consumption* **38**: 115–129. **URL:** *https://www.sciencedirect.com/science/article/pii/S2352550923000696*

Yang, Y., Chu, W.-L. and Wu, C.-H. (2022). Learning customer preferences and dynamic pricing for perishable products, *Computers & Industrial Engineering* **171**: 108440.