Customer Classification Modelling in Complex Supply Chains

Akanksha
Bhargav Bachanahalli Shekarmurthy
Rufus Meera Gomathi Sankar
Shivang Ranjan

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

August 2024

Supervisors: Dr Paula Carroll

Head of School: Professor Anthony Brabazon

Dedication

This project is dedicated to all our professors and to our beloved friend Vignesh.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Preface

This capstone project focuses on customer classification within the pharmaceutical supply chain, specifically for Pfizer's operations in France. The pharmaceutical industry is characterized by complex supply chains, stringent regulations, and a diverse customer base. Effective customer classification is crucial for optimizing supply chain operations, improving service delivery, and enhancing customer satisfaction.

In this report, we have employed advanced machine-learning techniques to segment Pfizer's customer base into meaningful clusters. By analysing over 12,000 customers, including wholesalers, independent pharmacies, and hospitals, we have developed a robust clustering model that prioritizes value and volume metrics. The insights gained from this model are intended to aid Pfizer in making data-driven decisions regarding customer prioritization, contract negotiations, and operational efficiency.

Our findings not only highlight the importance of precise customer segmentation in the pharmaceutical industry but also provide actionable recommendations for Pfizer to optimize its supply chain strategies. This project represents a significant step toward leveraging data analytics to address the challenges of managing a complex and dynamic supply chain in the pharmaceutical sector.

Dublin,
August 2024

Akanksha
Bhargav Bachanahalli Shekarmurthy
Rufus Meera Gomathi Sankar
Shivang Ranjan

Acknowledgements

We would like to extend our heartfelt gratitude to all those who have contributed to the success of this project.

We express our sincere thanks to Pfizer, particularly Conor Riordan and Kunal Gupta for sponsoring this project and their invaluable support and guidance throughout the engagement. Their insights and expertise have been crucial in shaping our approach and achieving our objectives. We also wish to acknowledge the support from UCD Michael Smurfit Graduate Business School, our Head of School Prof. Anthony Brabazon, our Program Director Dr Michael MacDonnell and our supervisor Dr Paula Carroll. Their academic mentorship and encouragement have been crucial in navigating the complexities of this project.

We are deeply grateful to our module coordinators, including Dr Miguel Nicolau and Assoc Prof. Peter Keenan for their exceptional teaching on the technical aspects and for helping us develop our skills on Python and Data Visualization. We extend our thanks to Dr Debajyoti Biswas, Assoc Prof. Sean McGarraghy and Dr Elane Ruane for helping us build a solid foundation in statistics and Machine Learning and for showing us how to apply these concepts to real-world scenarios. Additionally, we appreciate Prof. Michael O'Neill and Dr Clare Branigan for honing our skills in stakeholder management, data storytelling and emotional intelligence while collaborating. Their teaching, feedback and encouragement have provided us with the foundation and knowledge needed to successfully complete this project.

Executive Summary

The report contains the analysis of 12,470 customers from France based on their booking data from January 2023 to May 2024. Fifteen (15) groups of customers are created identifying profitable customers and customers where Pfizer might not be profitable. Unsupervised machine learning models are used to arrive at these results. Furthermore, two problems and the cohort of customers with these problems are identified. The first problem has a potential of approximately €13M in sales and the second one, €441,000 in savings annually, which would be 5% of total revenues for this cohort of pharmacies.

There are 10,068 independent pharmacies followed by 1331 private hospitals, 998 public hospitals and 72 wholesalers. However, reinforcing the existing body of knowledge at Pfizer it is observed that 25% of customers make 86% of the total revenue of €2.5Billions. Of this, 80% of the revenue is contributed by 72 wholesalers followed by independent pharmacies contributing 10% (€252M), 7% by public hospitals and 2% from private hospitals.

The first problem identified is high average rejection rates for all three clusters of wholesalers with an overall rate of 8.5%. Wholesalers also account for €18M of the total €19.9M worth of rejected orders because customers do not accept the backorders. Hence this opens an opportunity of €8M of sales if Pfizer targets a 5% rejection rate and it increases up to €13M if the target is 2%, which would be industry best. Recommendations to address this problem are provided based on industry knowledge base and scholarly works. The functional recommendations are to improve demand forecasting, enhance supplier-customer collaboration, and optimise safety stock levels. This could be translated organisationally as a dedicated wholesaler operations manager who would be part of joint inventory and production plan sessions with the wholesalers, act as an advocate for the wholesalers within Pfizer, being the single point of contact post sales championing smooth fulfilment. Additionally, there are two broader supply chain-wide recommendations, one is to increase inventory visibility and improve the order fulfilment process.

The second problem identified is a higher average transportation cost of 5.41% for very small pharmacies which are 4, 626 in number. If Pfizer can reduce the cost to 1%, then there would be a saving of €441,000 annually. One recommendation is on the supplier side, shared-resources logistics, where Pfizer can partner with other pharma cos to consolidate their orders of common customers and zone-skipping and centralised fulfilment, where Pfizer can ship orders to a single large customer in a geo area and then distribute from there. Additionally,

Pfizer can have a newer business model where the smaller customers can be offered incentives to buy through one of Pfizer's wholesalers. This ensures that the revenue is not lost while transportation cost is reduced or eliminated and further all three parties gain from this model.

This work analysed the provided booking data using machine learning models and identified different cohorts of customers with defining characteristics which helps in prioritising them, servicing them. Then two major problems are identified and there are made into opportunities of high economic value. Lastly, industry best practices and personalised recommendations are provided to realise this potential.

Chapter 1 - Introduction

1.1 Overall Business Understanding

The pharmaceutical industry operates within one of the most complex supply chain environments, characterised by stringent regulatory requirements, a wide variety of stakeholders, and the critical nature of its products. Ensuring the timely and safe delivery of medications is paramount, necessitating sophisticated supply chain management practices. Customer classification within this context involves segmenting customers based on various attributes to better understand their needs and behaviors.

Pfizer, a leading pharmaceutical company, services a diverse customer base with widely varying sales volumes, product portfolio, geography, modes of shipping and ordering patterns. The company faces challenges in efficiently negotiating deals and prioritising customer orders due to these variations. Currently, customer service and order delivery processes do not fully account for the differences in customer size and behaviour, leading to misaligned contracts and potential delays.

1.2 Business Objective

The pharmaceutical industry is highly dynamic, with complex supply chains that involve multiple stakeholders, including API and drug manufacturers, distributors, and healthcare providers. Efficient supply chain management is critical for ensuring timely delivery of products, maintaining inventory levels, and meeting customer demands.

The primary objective is to develop a customer classification model to identify key customer segments among the existing wide customer base. Evaluate the segments produced and deliver this via a dashboard to the stakeholders. By classifying customers based on metrics such as value, volume, and frequency. Pfizer aims to prioritise customers, chart logistical strategies and aid contract negotiations from the insights generated from the model output., ensuring that high-value customers receive prioritised attention. However, the larger extended objective is to deliver this as a model or package where the input is the customer data, and the output is a classified segment of the customers via a dashboard along with intermediary data representation as part of EDA (Exploratory Data Analysis).

1.3 Data Description

Pfizer has access to comprehensive datasets on the Celonis platform, including customer details, orders, shipping details and product information.

Data attributes are crucial for segmentation and include:

- Customer Details: ID.
- Order Details: Items, Materials, Sales Document, expected delivery date etc.
- Value Details: Net Order Value.

Chapter 2 - Literature Review

We have reviewed the literature from the lens of exploring different clustering techniques to segment customers. We have also investigated model evaluation metrics used by researchers. Furthermore, we have investigated the latest research and learning in e-commerce for advancements in logistics and customer segmentation. The e-commerce domain has a customer delivery logistical resemblance to that of Pfizer's; hence it makes it a good fit. In addition, research in the pharmaceutical supply chain is also reviewed to understand case studies which fit the problem profile.

Li and Lee (2024) present a comprehensive study on enhancing traditional customer segmentation methods by integrating support vector machines (SVM) and K-means clustering algorithms. The primary goal is to improve the accuracy and efficiency of customer classification by leveraging the strengths of both SVM and K-means. Traditional segmentation approaches often struggle with large, complex datasets, affecting the formulation of effective marketing strategies. By utilising the nonlinear classification capability of SVM and the clustering power of K-means, the proposed model addresses these limitations, providing a robust framework for customer segmentation. The study constructs a customer segmentation model that first employs SVM to segment existing customer data. It then integrates SVM with K-means to refine these segments further, identifying distinct customer groups with specific characteristics. The model's performance is validated through simulation experiments, demonstrating significant improvements in segmentation accuracy and profitability predictions. Key metrics include an average error rate of 6.82%, a recall rate of 91.28%, and a profit prediction accuracy within 2.53% of the true value.

This research is particularly relevant to Pfizer's customer segmentation/classification as it offers a proven methodology to handle large-scale customer data effectively. The integration of SVM and K-means ensures precise segmentation, allowing for the development of tailored marketing strategies.

Hicham and Karim (2022) present an innovative approach to customer segmentation by employing a clustering ensemble combined with spectral clustering. This method integrates four fundamental clustering algorithms: DBSCAN, K-means, Mini Batch K-means, and Mean Shift, aiming to deliver a more consistent and high-quality segmentation outcome. The use of spectral clustering further enhances the integration of multiple clustering results, leading to improved clustering quality. This approach is particularly flexible in handling

diverse client data, which is a crucial aspect for developing tailored marketing strategies (Hicham and Karim, 2022). The paper emphasises the importance of understanding customer needs and behaviours in today's competitive business environment. By grouping customers with similar characteristics, companies can develop more effective marketing strategies. The proposed methodology involves feature engineering to transform raw data into meaningful features that can be used for clustering. The performance of the model is evaluated using metrics such as Adjusted Rand Index (ARI), Normalised Mutual Information (NMI), Dunn's Index (DI), and Silhouette Coefficient (SC), with the ensemble model outperforming individual clustering techniques (Hicham and Karim, 2022)

Applying this model to a dataset collected from Moroccan citizens, the study demonstrates its practical applicability and effectiveness in real-world scenarios. We can utilise a similar approach to segment customers based on various characteristics such as demographics, purchasing behaviour, and other relevant factors. The enhanced segmentation quality achieved through this ensemble approach can lead to more accurate targeting and personalised marketing efforts, ultimately improving customer engagement and satisfaction (Hicham and Karim, 2022).

Das (2015) presents a comprehensive study on the application of machine learning techniques to customer classification. Das (2015) evaluates three widely used algorithms: Naïve Bayes, K-Nearest Neighbours (KNN), and Support Vector Machines (SVM), to determine their effectiveness in classifying banking customers based on historical data. The choice of these algorithms is driven by their proven efficacy in classification tasks across various domains. The study’s methodology includes thorough data pre-processing steps such as data cleaning, feature selection, and transformation, which are critical for improving model accuracy and reliability. The performance of each algorithm is evaluated using metrics such as accuracy, precision, and recall, providing a detailed assessment of their strengths and limitations (Das, 2015).

The evaluation of different algorithms offers a comparative analysis that can guide the selection of the most suitable machine-learning techniques for customer segmentation tasks. The emphasis on data pre-processing aligns with the need to handle large and complex datasets in the pharmaceutical industry effectively. Moreover, the performance metrics used in the study can serve as benchmarks for assessing the effectiveness of the classification models, ensuring they meet the required standards for accuracy and reliability.

By applying the methodologies discussed in this paper, Pfizer can enhance its customer segmentation efforts, leading to more precise targeting and personalised marketing strategies.

The practical application of these models to classify banking customers demonstrates their utility in real-world scenarios, which can be translated to the pharmaceutical context. This approach can help Pfizer identify key customer segments, optimise marketing resources, and improve overall customer engagement and satisfaction.

Griva et al. (2024) presents a two-stage business analytics approach that combines behavioural and geographic segmentation using e-commerce delivery data. This study employs data mining and machine learning techniques, making it highly relevant to our project at Pfizer, which also aims to segment customers using order delivery data. The study integrates behavioural segmentation, focusing on customer purchase behaviours, with geographic segmentation, which considers the spatial characteristics of customers. This dual approach provides a comprehensive view of the customer base, crucial for Pfizer in understanding customer preferences and optimising delivery logistics. Griva et al. uses advanced techniques like Latent Dirichlet Allocation (LDA) for topic modelling and feature selection, offering detailed and actionable insights. Their analysis of real-world e-commerce data from a third-party logistics company demonstrates the practical applicability of their methods, providing a tested framework for our project.

Tabianan, Velu, and Ravi (2022) present a data-driven approach to customer segmentation aimed at optimising delivery strategies within the e-commerce sector. Their study employs K-means clustering and association rule mining to segment customers based on their delivery preferences and purchasing behaviour. By analysing these patterns, the study seeks to develop tailored delivery strategies that cater to the specific needs of different customer groups, thereby improving the efficiency and effectiveness of delivery operations. The methodology involves clustering customers using K-means to identify distinct segments, followed by the application of association rule mining to uncover frequent delivery patterns within each segment. This combination allows for a more granular understanding of customer behaviour, which is essential for optimising delivery routes and schedules. The study evaluates the performance of these strategies through simulation experiments, demonstrating significant improvements in delivery efficiency and customer satisfaction. Metrics such as delivery time reduction and increased on-time delivery rates are used to validate the approach (Tabianan, Velu, and Ravi, 2022).

This research is highly relevant to Pfizer's customer segmentation project. By leveraging similar data-driven techniques, Pfizer can enhance its delivery strategies, ensuring that pharmaceutical products are distributed more efficiently based on customer segmentation.

The ability to tailor delivery operations to specific customer groups can lead to improved service levels, reduced operational costs, and higher customer satisfaction.

For Pfizer, adopting a similar approach can lead to substantial improvements in their delivery operations. By segmenting customers based on delivery data and using predictive models to anticipate delivery needs, Pfizer can enhance the efficiency of their logistics network. This strategy can ensure timely delivery of pharmaceutical products, optimise resource allocation, and improve overall customer satisfaction.

Dang (2015) explore the application of machine learning algorithms for customer segmentation. Their study compares the effectiveness of several clustering algorithms, including K-means, hierarchical clustering, and DBSCAN. By employing these algorithms, the study aims to create more precise and actionable customer segments, which can be used to tailor marketing strategies and improve customer engagement. The methodology involves pre-processing the data through feature engineering to enhance the quality and relevance of the input features. For Pfizer, these clustering algorithms can be applied to identify distinct customer groups. Based on the results, Dang (2015) concluded that the simple K-means clustering algorithm is the fastest. The Make density-based clustering algorithm (MDBCA) is equally as fast as the simple K-means. However, the Agglomerative algorithm is more sensitive to noisy data and demonstrates greater variation in time complexity. In terms of time complexity and the dataset used, K-means produces better results compared to all other algorithms mentioned.

By adopting similar machine learning approaches, Pfizer can enhance its segmentation efforts, leading to more effective and personalised marketing strategies. The insights gained from these techniques can help Pfizer better understand its customer base, optimise marketing resources, and ultimately improve customer satisfaction and loyalty.

Mustafa et al. (2023) investigate the use of data mining techniques for customer classification in the e-commerce industry. Their study evaluates the performance of various algorithms, including Support Vector Classifier (SVC), K-nearest Neighbours, Decision Tree, Random Forest, AdaBoost Classifier and Gradient Boosting Classifier, to classify customers based on transaction details such as product name, quantity, price, and other IDs. The primary goal is to leverage advanced data mining techniques to manage the complexity and volume of data, ultimately enhancing customer relationship management and targeted marketing efforts. The study's methodology includes a thorough data pre-processing phase to clean and transform the raw data into suitable formats for analysis. Following this, the different data mining algorithms are applied to classify customers into distinct groups. The effectiveness of these

algorithms is assessed using precision. Combining the results from Gradient Boosting, k-NN and Random Forest algorithms provided better results. This finding underscores the importance of selecting the right data mining techniques to achieve optimal classification results in the e-commerce context (Mustafa et al., 2023).

For Pfizer, the insights from this research can be directly applied to enhance their customer classification efforts. By implementing the data mining techniques discussed in this paper, Pfizer can achieve more accurate customer classifications, enabling the development of highly targeted marketing strategies.

Abdulhafedh (2021) presents a clustering approach for customer segmentation in the banking industry. It analyses transaction data of credit card holders using K-means, Hierarchical clustering, and Principal Component Analysis (PCA). The study first applies clustering algorithms directly and then uses PCA for dimensionality reduction before reapplying clustering. Among the algorithms, K-means proved to be more effective than Hierarchical clustering for this dataset. The evaluation metrics used to compare the clustering performance included the Davies-Bouldin Index, Silhouette Score, and Dunn Index. These metrics indicated that K-means provided a better fit, especially after updating the number of clusters based on PCA results, which identified an additional segment overlooked by the other methods. The study underscores the importance of PCA in refining clustering outcomes and offers tailored marketing strategies for the identified customer segments to enhance customer engagement and profitability (Abdulhafedh, 2021).

Li and Lee (2024) discuss the application of big data analytics to enhance customer segmentation and improve supply chain efficiency in the pharmaceutical industry. The study highlights how advanced data analytics can be used to identify distinct customer segments based on purchasing patterns, demographic data, and prescription histories. These segments are then leveraged to optimise various supply chain processes, such as inventory management, distribution planning, and delivery scheduling, ensuring that the supply chain operations are aligned with the specific needs of each customer segment.

The research methodology involves collecting and pre-processing large datasets from various sources, followed by the application of machine learning techniques such as clustering algorithms and predictive analytics. The study uses these techniques to segment customers into meaningful groups and then integrates these segments into supply chain optimisation models. Key performance metrics, including delivery times, inventory levels, and customer satisfaction rates, are used to evaluate the effectiveness of the integrated approach. The

findings indicate that the use of big data analytics leads to more accurate customer segmentation and significant improvements in supply chain efficiency (Li and Lee 2024).

In addition to the above, various models have been proposed for customer classification. For instance, decision trees, neural networks, and logistic regression are commonly used. Decision trees, as discussed by Quinlan (1986) in "Induction of Decision Trees," provide an intuitive approach to customer classification by splitting the data into homogeneous groups based on decision rules. Neural networks, covered extensively by Goodfellow, Bengio, and Courville (2016) in Deep Learning, offer powerful techniques for handling complex and non-linear data patterns, making them suitable for customer classification in complex supply chains. Kumar and Reinartz (2018), in Customer Relationship Management: Concept, Strategy, and Tools explore the use of RFM analysis and other segmentation techniques to enhance customer relationship management. They highlight the importance of understanding customer value and behaviour to develop effective segmentation strategies.

Metric	Definition	Paper
Dunn's Index (DI)	A metric that evaluates the compactness and separation of clusters, with higher values indicating better clustering quality.	Hicham and Karim (2022)
Silhouette Coefficient (SC)	A measure of how similar an object is to its own cluster compared to other clusters, with higher values indicating better-defined clusters.	Hicham and Karim (2022)
Error Rate (ERA)	The error rate refers to the proportion of incorrect predictions made by the model relative to the total number of predictions.	Li and Lee (2024)
Recall Rate	The recall rate measures the ability of the model to correctly identify all relevant instances	Li and Lee (2024)

	(in this case, customers belonging to a specific segment).	
Profit Prediction Accuracy	Degree to which the predicted profit by the model aligns with the actual profit, reflecting the accuracy of profitability predictions.	Li and Lee (2024)
Adjusted Rand Index (ARI)	A metric used to measure the similarity between two data clusters, adjusted for chance.	Hicham and Karim (2022)
Normalized Mutual Information (NMI)	A measure of the amount of shared information between two clusters.	Hicham and Karim (2022)
Accuracy	A metric that measures how often the model correctly classifies instances out of the total instances.	Das (2015)
Precision	A measure of how many of the instances predicted as positive are actually positive.	Das (2015)

Table 1: Performance Metrics for Clustering

Model	Advantages	Disadvantages	Paper
Support Vector Machines (SVM) + K-means	Combines the nonlinear classification power of SVM with the clustering power of K-means for enhanced accuracy.	May be computationally intensive for large datasets.	Li and Lee (2024)
Agglomerative	Effective for	Can be slow and	

Clustering	hierarchical relationships and detecting nested clusters.	memory-intensive for large datasets.	
DBSCAN Clustering	Capable of finding arbitrarily shaped clusters and handling noise.	Struggles with varying cluster densities and high-dimensional data.	Dang (2015)
Clustering Ensemble with Spectral Clustering	Improves clustering quality by integrating results from multiple clustering algorithms.	Complex and may require careful tuning of parameters.	Hicham and Karim (2022)
Naive Bayes	Simple, fast, and effective for small datasets.	Assumes independence between features, which may not always be true.	Das (2015)
Gradient Boosting	High accuracy and ability to handle complex data patterns.	Can be prone to overfitting and requires careful tuning of parameters.	(Mustafa et al. 2023)

Table 2: Pros and Cons of different Clustering techniques

Chapter 3 - Methodology

This chapter is dedicated to outlining the methodology for developing a customer classification model within complex pharmaceutical supply chains, using Pfizer as a case study. The methodology is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, which comprises six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. This chapter delved into the first four phases, with a specific focus on business and data understanding, data exploration, scaling, and clustering models.

3.1 CRISP-DM

3.1.1 Business Understanding

The pharmaceutical industry, characterised by its complex and dynamic supply chains, demands highly efficient and adaptable customer management strategies. For Pfizer, this involves developing a refined understanding of its diverse customer base to ensure optimal service delivery and efficient supply chain management. The primary business objectives are:

1. To develop a robust customer clustering model focusing on value and volume metrics that categorises Pfizer's customer base on existing customer segments - Wholesale, Retail, and Hospitals (public/private).
2. To deploy this model in a dashboard for visualisation to access and interpret customer segments effortlessly and to rank customers within each cluster.
3. To ensure that the model improves customer prioritisation to support effective contract negotiations.

Based on our initial business understanding, we focused on addressing three key questions. However, after completing the evaluation phase and revisiting the business understanding phase, we identified additional business problem from our clusters. Hence, our final question is as below:

4. To analyse the developed clusters, identify underlying issues and provide strategic recommendations to Pfizer aimed at optimizing operations and achieving cost savings.

3.1.2 Data Understanding

Data Collection and Description:

The data required for this project encompasses comprehensive details on customer orders, including key information such as sales order identifiers, line-item numbers, company codes

representing corporate entity and types of sales orders. The dataset further includes data on existing Business Unit splits and material.

Additionally, customer-specific information is provided including customer identifiers, customer groupings (Wholesalers, Independent Pharmacies, Public Hospitals, Private Hospitals) and the origin of the sales order, specifying the channel or source of the order placement. The dataset also captures the timeline of the order process, from the creation date of the sales order and individual line items to the dates of purchase orders, requested and actual delivery dates, and various service level agreement (SLA) metrics.

Critical logistical information is included, such as picking dates, delivery document identifiers, and the first goods issue dates, along with metrics on the workdays required to complete each stage of the order process. The dataset also contains flags and indicators for the full completion of deliveries, reasons for order or line-item rejections and any reductions in order quantities.

Performance metrics such as On-Time-In-Full (OTIF) indicators, order cycle time, delays, and specific KPIs for picking, packing, goods issue, and proof of delivery are tracked to assess the efficiency and accuracy of the order fulfilment process. The financial aspect is covered by including the net order value, delivered quantity, and SAP-confirmed quantities.

The data has been anonymized for use in this project. Based on our understanding from Pfizer, we built the data dictionary, and the following table contains the columns and a brief description of what each column means.

Column Name	**Description**
Sales Document	Identifier for the sales order
Item	Line item number within the sales order (so line number)
Company Code	Code representing the company within the corporate structure
Sales Document Type	Type of the sales order
BU Split	Business Unit Split, indicating division or department within the company
Material	Actual Item (Medicines)
Customer ID	Identifier for the customer placing the order

Customer group 2	Category or segment of the customer
Order Origin	Source or channel through which a sales order was placed
SO Creation	Date and possibly time when the sales order was created
SO Item Creation date	Date when the specific line item was created in the system
SO Item Creation Time	Time when the specific line item was created in the system
Purchase order date	Date when the customer placed the purchase order
SAP RDD	Requested Delivery Date by the customer in the SAP system
Calculated Req. PGI. Date	Calculated required date for picking, packing and goods issue
Calculated Req. Del. Date POD	Calculated required delivery date for proof of delivery
Actual Delivery Date (POD)	Actual date when delivery was confirmed
SLA Days PGI	Service Level Agreement days for picking, packing and goods issue
SLA Days POD	Service Level Agreement days for proof of delivery
Create Picking Date	Date when the order was picked from the inventory
Delivery Document	Identifier for the delivery document associated with the order
Delivery Creation	Date and possibly time when the delivery order was created
First Goods Issue Date	Date when the goods were first issued from the inventory
Workdays until PGI	Number of working days taken until picking, packing and goods issue
Workdays to POD	Number of working days taken until proof of delivery
Latest Order Entry	Latest date when order entry was modified or updated

In Full Delivery FLAG	Indicator flag showing if the delivery was completed in full
Reason for Rejection	Reason code for order or line item rejection, if applicable
Supply Policy Amount	Amount or quantity as per the supply policy
Reason for Reduction	Reason code for any reduction in order quantity or amount
OBD Date	Outbound delivery date
OTIF Month	On-Time In-Full performance metric for the month
Order Cycle time	Total time taken from order placement to delivery
Delay	Delay in order process, in any
On Time	Indicator if the order was processed on time
On Time PGI	Indicator if picking, packing, and goods issue were on time
Delivery On Time	Indicator if the delivery was on time
In Full	Indicator if the order was delivered in full
OTIF KPI	On-Time-In-Full Key Performance Indicator
OTIF KPI PGI	On-Time-In-Full Key Performance Indicator for picking, packing, and goods issue
OTIF KPI POD	On-Time-In-Full Key Performance Indicator for proof of delivery
SLA calculation method	Method used to calculate service level agreements
Net Order Value	Total value of the sales order
SAP Del Qty	Quantity delivered as per the SAP system
Delivered Quantity	Actual quantity delivered to the customer
SAP Conf Qty	Quantity confirmed in the SAP system
PET 1st Qty	Possibly the first quantity of a specific product or batch in the PET system.

Table 3: Data Dictionary

Exploratory Data Analysis (EDA)

EDA was conducted to identify patterns, detect outliers, and test underlying assumptions through visual and statistical methods. EDA is performed on different metrics based on value, volume, rejection, OTIF and frequency.

Insights from the initial EDA

Orders: The initial EDA reveals that out of 12,470 total customers, wholesalers contribute to 80% of the net order value (excluding returns) across 115,279 orders indicating a significant concentration of high-value orders among a very small customer base.

Returns: Out of 271,181 orders, 1,389 were returned, with a total value of €296k. Notably, 6 out of 543 customers who made returns, account for €264k, representing 90% of the return value.

Orders by value: Among the orders by value, three customers stand out with exceptionally high order values:

- Customer ID 2000004623: placed 10 orders, 9 of which are each worth a maximum of €185,644.
- Customer ID 2000003010: placed 9 orders, 8 of which are each worth a maximum of €180,933.
- Customer ID 2000003886: placed 2 orders, each worth €96,533.

Per order value: The per order value analysis reveals that there are 10-15 orders exceeding €2 million, with the highest single order valued at an impressive €7 million.

Volume (Number of orders per customer): The volume analysis of orders per customer reveals that 10 customers have placed more than 1,000 orders, with the range extending from 1,000 to 30,000 orders. The 90th percentile of customers placed 25 orders each, while the median number of orders per customer is 4. This distribution highlights a small group of highly active customers and a broader base of customers with fewer orders.

On-time In Full Delivery (OTIFKPI): The On-Time In Full (OTIF) delivery analysis shows that 93% (256k) of the orders met the OTIF criteria. Additionally, 94% of delivery documents, totalling 302k, were also OTIF. However, 7% of orders were non-OTIF. Notably,

50% of these non-OTIF orders were attributed to just 5 customers and 93% of this 7% are from pharmaceutical wholesalers.

Rejection: Regarding order cancellations, 1.25% of orders (26k) were rejected. A significant 98% of these rejections occurred because customers do not accept backorders. Five customers are responsible for 60% of these rejected orders. Specifically, customer 3000234587 has a rejection rate of 22% (5,849 orders), and customer 3000223730 has a rejection rate of 14% (3,882 orders) indicating that a small number of customers are heavily contributing to the overall rejection rate (ratio of number of item lines rejected to the total number of item lines), primarily due to the reason that "customer do not accept backorders".

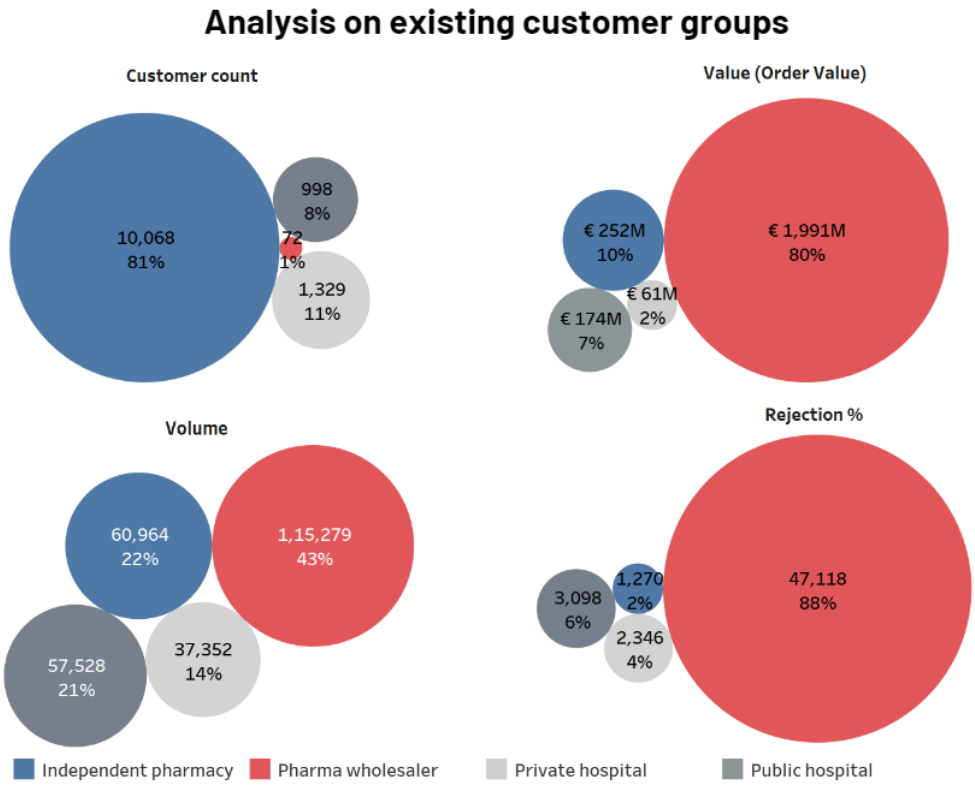

Figure 1: Analysis of existing customer group

3.1.3 Data Preparation

Data Modelling:

- Since the goal is to cluster customers, the data is modelled in such a way that each row represents one customer and the customer's associated features like Customer ID, BU split, customer group, sales order count, items quantity, delivery document count, material count, net order value, rejection, OTIF, weeks with orders.

- Various aggregation methods such as median, count, sum, average and variance are calculated for each measure at the customer level based on all the sales order data available for the customer.

Data Scaling:

Given the distance-sensitive nature of clustering algorithms, normalising the data is crucial. Scaling techniques such as Z-Scaling, Min-Max Scaling and Standardisation have been considered to ensure that the clustering algorithms function optimally without bias towards variables with higher magnitude.

3.1.4 Modelling

Pfizer has existing customer groups. When initial clustering was done with all available customers, the clusters turned out to match Pfizer's existing customer groups (Private and public hospital, retailers and wholesalers). Hence, we proceed with intra customer group clustering to provide specific value-volume features-based customer clusters.

Clustering algorithms explored based on Literature review:

- K-means Clustering
- DBSCAN
- Hierarchical/Agglomerative Clustering
- Gaussian Mixture Model

3.1.5 Evaluation

- Elbow Method – To select the optimal number of clusters.
- Silhouette score – To analyse quality of the clusters formed.
- Dunn's Index (DI) - To analyse quality of the clusters formed. Dunn's index is a ratio of the smallest distance between any two clusters to the largest intra-cluster distance within any cluster.
- Feature Importance – To understand the contribution of each feature towards the clustering.

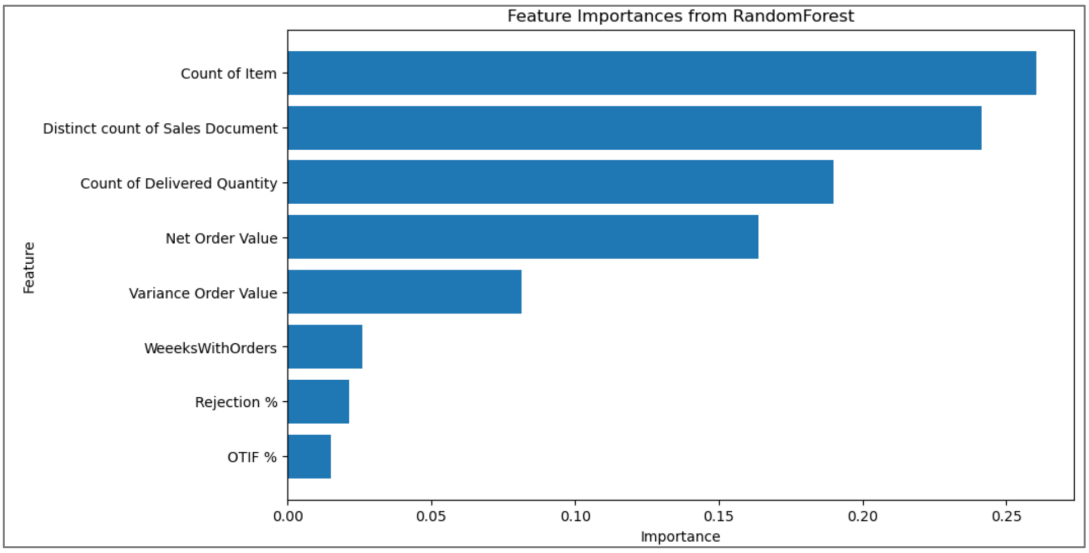

Figure 2: Feature Importance in Clustering Analysis

- Characteristics of each cluster – To understand the characteristics of each cluster and how each cluster varies from others.

Wholesalers:

Algorithm	Silhouette Score	Dunn's Index	Top 4 Features
Agglomerative	0.30	0.01	Rejection%, OTIF%, Count of delivery quantity, net order value
K means	0.30	0.02	Rejection%, OTIF%, Count of delivery quantity, net order value
DBSCAN	0.32	-	Rejection%, Variance order value, Count of delivery quantity, net order value
Gaussian Mixture	0.63	0.95	Count of Item, Count of Sales Document, Count of delivery quantity, net order value

Table 4: Comparison of different Clustering Algorithms for Wholesalers

Based on the evaluation of various clustering algorithms, the Gaussian Mixture Model (GMM) demonstrated superior performance in both key metrics: Silhouette Score and Dunn's Index. GMM achieved a significantly higher Silhouette Score of 0.63, compared to the other algorithms which ranged from 0.30 to 0.32. This indicates that GMM formed clusters that are

well-separated and internally cohesive. Additionally, the Dunn's Index for GMM was 0.95, surpassing the other algorithms, further validating the quality of clustering. Moreover, GMM's top features align closely with Pfizer's value-volume based clustering requirement. The count of items, sales documents, delivery quantities, and net order value were the top features used for GMM clustering.

Independent Pharmacies:

Algorithm	Silhouette Score	Dunn's Index	Top 4 Features
Agglomerative	0.46	0.26	Count of Item , Distinct count of Delivery Document, Distinct count of Sales Document, Count of Delivered Quantity
Agglomerative2 (3 clusters)	0.35	0.039	Count of Item , Distinct count of Material, Count of Delivered Quantity , WeeeksWithOrders.
K-means	0.41	0.042	Count of Item, Distinct count of Material, Count of Delivered Quantity, Distinct count of Delivery Document
DBSCAN	-0.18	0.024	Median order Value , Count of Delivered Quantity, Count of Item , Net Order Value
GMM- 4	0.26	0.051	Count of Item , Median order Value, Net Order Value, Distinct count of Sales Document.
GMM - 3	0.21	2.3	Net Order Value , Median order Value , Variance Order Value , Count of Delivered Quantity

Table 5: Comparison of different Clustering Algorithms for Independent pharmacies

There are 10,068 independent pharmacies which Pfizer service across France. There were 6 best models after implementing GridSearchCV. Based on the evaluation metrics of silhouette score, agglomerative clustering with four clusters has the best score of 0.46, however Dunn's

index is 0.26. The top four important features are predominantly volume based metrics. The results of DBSCAN clustering algorithm had poor scores where silhouette score is -0.018 which is indicative of very poor clustering where points are added to a wrong cluster than the nearest cluster.

Gaussian mixture model with three clusters has a silhouette score of 0.21, while Dunn's index is 2.3 which is highest which is the ratio of minimum inter cluster distance to maximum intra cluster distance. This clustering also had the top three features which are value centric and the fourth feature is volume based. These three clusters have a fairly even sizes 2348, 3094, 4626. Hence, this clustering is chosen as our proposal for Pfizer.

Public and Private Hospitals (Combined):

Algorithm	Silhouette Score	Dunn's Index	Top 4 Features
Agglomerative	0.9	0.006	Distinct count of Delivery Document, Count of Item, Weeks with Orders, Distinct count of Sales Document
K means	0.6	0.005	Distinct count of Delivery Document, Count of Item, Distinct count of Sales Document, Weeks with Orders
DBSCAN	0.7	0.01	Count of Delivered Quantity, Net Order Value, Distinct count of Sales Document', Rejection
Gaussian Mixture	0.5	0.00014	Net Order Value, Distinct count of Delivery Document, Count of Delivered Quantity, Variance order value

Table 6: Comparison of different Clustering Algorithms for Combined Hospitals (Public and Private)

Agglomerative Clustering shows the highest Silhouette Score (0.9), indicating that the clusters are well-defined and cohesive. However, its Dunn's Index is relatively low (0.006), suggesting that the separation between clusters might not be optimal. DBSCAN has a moderately high Silhouette Score (0.7) and the highest Dunn's Index (0.01) among the four, suggesting good cluster cohesion and separation. The use of the k-distance graph further enhances DBSCAN's capability to handle noise and density variations, making it particularly

suitable for datasets with outliers or non-uniform cluster shapes. Also, DBSCAN prioritises major volume and value metrics for the clustering process. The need to balance both cohesion within clusters and separation between them, DBSCAN appears to be the best approach, especially with the additional insights provided by the k-distance graph that clear inflexion points allowed for precise parameter tuning, making DBSCAN effective at identifying clusters with varying densities and handling noise, which is crucial in real-world datasets where data points might not fit neatly into predefined clusters.

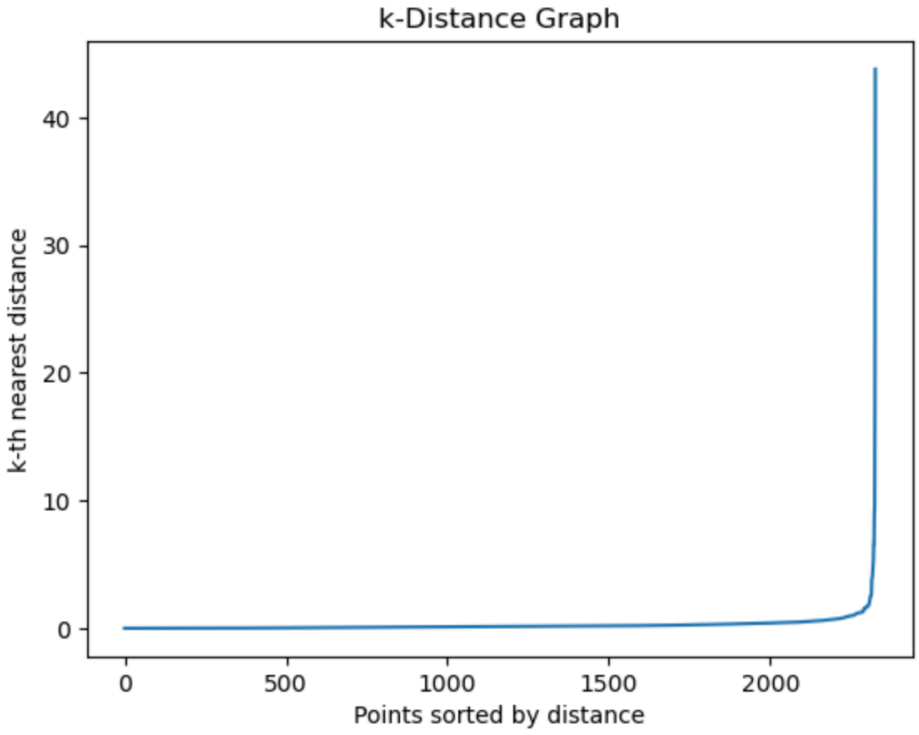

Figure 3: K-distance Graph for evaluating DB Scan parameters

Private Hospitals:

Algorithm	Silhouette Score	Dunn's Index	Top 4 Features
Agglomerative	0.625	0.018	Distinct Count of Delivery Document, Distinct Count of Sales Document, Weeks with Orders, Count of Item
K means	0.60	0.011	Distinct Count of Delivery Document, Count of Item, Distinct Count of Sales Document, Weeks with Orders

DBSCAN	0.295	0.023	Count of Item, Net Order Value, Distinct Count of Delivery Document, Count of Delivered Quantity
Gaussian Mixture	0.188	-	Net Order Value, Distinct Count of Delivery Document, Variance Order Value, Distinct Count of Sales Document

Table 7: Comparison of different Clustering Algorithms for Private Hospital

There are 1331 private hospitals which Pfizer services across France. Agglomerative clustering stands out as the most effective algorithm, achieving the highest Silhouette Score (0.625), indicating that the clusters are well-defined and exhibit strong cohesion. It also maintains a competitive Dunn's Index (0.018), suggesting a decent balance between cluster separation and cohesion. K-means clustering, while slightly trailing with a Silhouette Score of 0.60, shows a lower Dunn's Index (0.011), implying that the separation between clusters is less optimal compared to Agglomerative clustering. However, K-means prioritizes critical Volume metrics such as the Distinct Count of Delivery Document, Count of Item, and Distinct Count of Sales Document, making it effective for analysing specific metrics. DBSCAN, despite having a lower Silhouette Score (0.295), exhibits a relatively higher Dunn's Index (0.019), indicating its strength in handling clusters with noise and density variations, particularly useful for datasets with outliers. The Gaussian Mixture model, with the lowest Silhouette Score (0.188) and absence of Dunn's Index, is the least effective among the algorithms. Given the need to maximize both cluster cohesion and separation, Agglomerative clustering emerges as the most suitable approach.

Public Hospitals:

Algorithm	Silhouette Score	Dunn's Index	Top 4 Features
Agglomerative	0.41	-	Count of Item, Distinct Count of Delivery Document, Distinct Count of Material, Weeks with Orders
K means	0.53	0.011	Distinct Count of Delivery Document, Count of Item,

			Weeks with Orders, Distinct Count of Sales Document
DBSCAN	0.476	0.015	Distinct Count of Delivery Document, Distinct Count of Sales Document, Count of Delivered Quantity, Distinct Count of Material
Gaussian Mixture	0.368	-	Distinct Count of Delivery Document, Count of Item, Distinct Count of Sales Document, Weeks with Orders

Table 8: Comparison of different Clustering Algorithms for Public Hospital

There are 998 public hospitals. Based on the evaluation of various clustering algorithms, K-means clustering demonstrated superior performance in one of the metrics. It achieved a higher Silhouette Score of 0.53, compared to the other algorithms which ranged from 0.35 to 0.47. This indicates that K-means formed clusters that are well-separated and internally cohesive. Top features align closely with Pfizer's value-volume based clustering requirement. Count of items, sales document, delivered quantities and weeks with orders were the top features used for K-means clustering.

Chapter 4 - Results & Analysis

4.1 Overall Clustering Results

Pfizer's existing customer groups (Private and public hospitals, retailers and wholesalers) match the initial clustering results. The initial round of clustering gave us two clusters, broadly one that represented the pharma wholesalers and the other represented other non-wholesalers. On further clustering, the non-wholesalers, the 2 broad clusters obtained represented independent pharmacies and hospitals (public and private). To bring out more insights, further intra-customer-group clustering was performed and within each existing customer group clusters were created.

As discussed in the methodology section, the unsupervised clustering machine learning algorithms tried include Agglomerative clustering, K means, DBSCAN and Gaussian Mixture Models.

This intra-group clustering addresses the first business question from our business understanding, which was to develop a robust customer clustering model focusing on value and volume metrics that categorises Pfizer's customer base on existing customer segments - Wholesale, Retail, and Hospitals (public/private). We answer this question in the following sections with our final clusters, the rationale behind choosing these clusters and their analysis.

4.2 Intra Clustering Results

4.2.1 Wholesalers

Pfizer has a total of 72 wholesale customers out of the total 12k customers. But these wholesalers contribute over 80% of the total revenue. Hence, performing analysis by clustering these customers will Pfizer make specific decisions. The customer features used for developing the clustering model are 'Customer ID', 'C-Group', 'Count of Delivered Quantity', 'Count of Item', 'Distinct count of Sales Document', 'Net Order Value', 'WeeeksWithOrders', 'Variance Order Value', 'OTIF %' and 'Rejection %'. Once the features are loaded, they are scaled using the StandardScaler module available in the sklearn library in Python. Once the features are scaled, the Elbow method is used to identify the range for the optimal number of clusters. GridSearchCV was used for hyper-parameter tuning to study the results for different parameters for various algorithms. Based on the elbow method, the

number of customers in each cluster and characteristics of the clusters, we concluded that 3 clusters would be the ideal option.

Wholesalers Clustering:

Wholesalers

	Large Wholesalers	Moderate Volume Wholesalers	Small Volume - High Performance Wholesalers
Distinct count of Customer ID	7	14	51
Net Order Value	€ 1,835M	€ 84M	€ 72M
Median Net Order Value	€ 211,709K	€ 5,662K	€ 924K
Avg. Std Dev Net Order Value	€ 59,479	€ 40,208	€ 13,281
Avg. DeliveryCostPercentofDeliveryNetvalue	0.16%	0.35%	0.49%
Median Distinct count of Sales Document	12,743	210	131
Count of Item	482,991	26,977	44,503
Count of Delivered Quantity	24,874,593	1,203,308	1,080,075
Avg. Distinct count of Material	216	113	113
Median Weeeks With Orders	74	69	61
Avg. Rejection	8.64%	9.79%	5.12%
Avg. Return %	0.08%	1.31%	0.70%
Avg. Delay %	0.86%	7.22%	2.61%
Avg. Avg. Lead time	3	6	3
Avg. InFull rate %	80.26%	80.13%	88.57%

Figure 4: Clustering Characteristics for Wholesale

Cluster Characteristics:

1) Large Wholesalers:

Large Wholesalers have the highest net order value at €1,835M, with a median net order value of €211,709K. These 7 wholesalers contribute to 92% of revenue and 98% of the order quantity indicating their significant market impact. They have high order frequency and their delivery cost as a per cent of net-order value is very low (0.1%). These wholesalers also have a high rejection rate of 8.64%.

2) Moderate Volume Wholesalers:

They deal with smaller quantities, as seen in their lower count of delivered quantities and items, yet they have the highest rejection (9.79%) and delay rates (7.22%) among all the clusters. Their average lead times are twice as the other clusters. These wholesalers also have a high rejection rate of 9.79%. Delay rate also seems to be high at 7.22%.

3) Small Volume – High-Performance Wholesalers:

Even though they contribute less in terms of revenue and ordered quantity, their rejection % is almost half of moderate wholesalers and they have a 10% better in-full rate than other clusters.

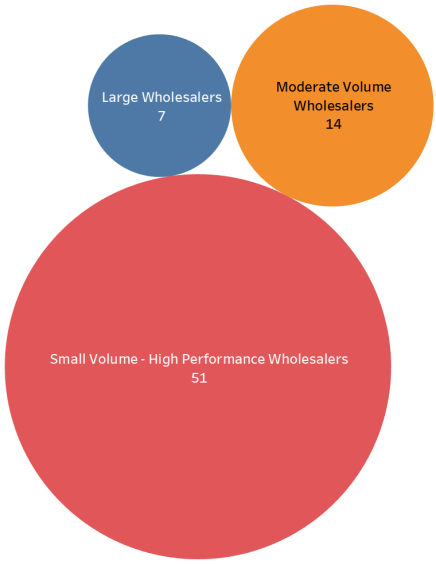

Figure 5: Wholesale customers' size

4.2.2 Independent Pharmacies

10,068 independent pharmacies are distributed across three clusters. As in the below table, the median net order value varies across clusters with € 1,000 for very small pharmacies which order 3 times (median) a year, high ratio of delivery cost to the net order value and an average distinct count of materials, higher than that of the medium-sized pharmacies. Another cluster is high-worth pharmacies, with relatively least number of customers, but the highest median net order value of €28,000 compared to €15,000 and €2,000 for the other two clusters.

Independent pharmacies

	High Worth Pharmacies	Medium Sized Pharmacies	Very Small Pharmacies
Distinct count of Customer ID	3,094	2,348	4,626
Net Order Value	€ 177M	€ 65M	€ 10M
Median Net Order Value	€ 28K	€ 15K	€ 1K
Avg. Std Dev Net Order Value	€ 2,358	€ 1,056	€ 181
Avg. DeliveryCostPercentofDeliveryNetvalue	1.75%	0.65%	5.41%
Median Distinct count of Sales Document	9	3	3
Count of Item	139,877	11,932	95,902
Count of Delivered Quantity	2,666,637	67,800	1,910,976
Avg. Distinct count of Material	15	2	10
Median Weeeks With Orders	9	3	3
Avg. Rejection	0.66%	2.03%	0.27%
Avg. Return %	0.27%	0.27%	0.11%
Avg. Delay %	4.99%	4.51%	5.33%
Avg. Avg. Lead time	3	3	3
Avg. InFull rate %	96.91%	96.15%	97.82%

Figure 6: Clustering Characteristics for Independent Pharmacies

Cluster Characteristics:

The broader assessment is that 45% of the pharmacies account for only 4% of the revenue from pharmacies. 30% of all pharmacies which form high-worth pharmacy clusters account for 70% of the revenue. Medium-sized pharmacies have a unique feature where they order less variety of materials (average distinct count of materials is two), with the highest rejection rate and least delivery cost to net order value ratio.

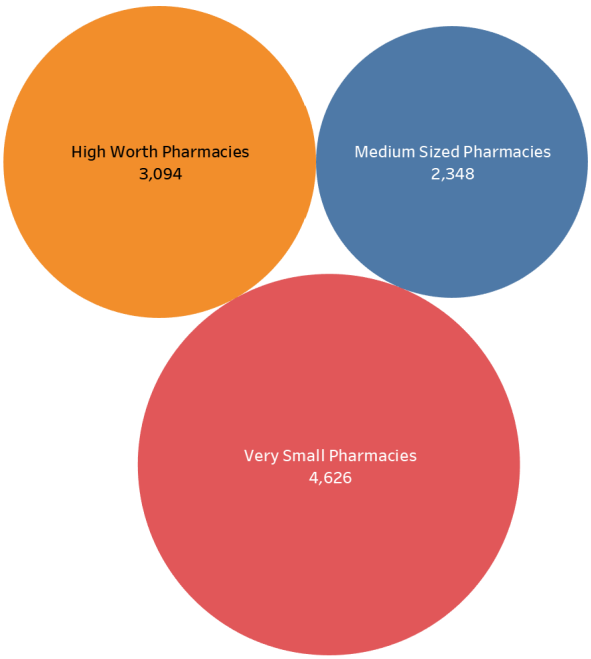

Figure 7: Independent Pharmacies customers' size

4.2.3 Combined Hospitals (Private and Public Hospitals)

Pfizer has a total of 2,327 hospital customers (combining public and private hospitals), which constitute around 19% of the total customer base of 12k. These hospital customers contribute approximately 9% to the total revenue, with private hospitals contributing 2% and public hospitals contributing 7%. Given their significant, yet smaller share of revenue compared to wholesalers, analysing and clustering these hospital customers will allow Pfizer to make targeted decisions to enhance engagement and optimize strategies.

The customer features used for developing the clustering model are 'Customer ID', 'C-Group', 'Count of Delivered Quantity', 'Count of Item', 'Distinct count of Sales Document', 'Net Order Value', 'Weeks with Orders', 'Variance Order Value', and 'Rejection %'. After scaling the features using the StandardScaler and utilising the Elbow method to determine the optimal number of clusters, the k-distance graph was utilized for the DBSCAN algorithm to identify the appropriate epsilon value which is crucial for clustering. GridSearchCV was also employed for hyper-parameter tuning to explore a variety of parameters for different algorithms. After applying the Elbow method, analysing customer distribution and considering cluster characteristics, we found that three clusters were the optimal solution for segmenting hospital customers. By incorporating the k-distance graph, we were able to further refine the clustering approach, especially for DBSCAN, ensuring distinct and well-defined clusters that accurately represent the nuances of the hospital customer base.

Combined Hospitals DB Scan Clustering Characteristics:

Combined Hospitals

	High Worth Hospitals	Low Worth Hospitals	Medium Worth Hospitals
Distinct count of Customer ID	105	15	2,205
Net Order Value	€ 137,435K	€ 6K	€ 98,011K
Median Net Order Value	€ 842,159	€ 119	€ 6,177
Avg. Std Dev Net Order Value	€ 11,561	€ 86	€ 1,101
Avg. DeliveryCostPercentofDeliveryNetvalue	0.12%	0.27%	0.93%
Median Distinct count of Sales Document	206	2	19
Count of Item	55,974	47	152,183
Count of Delivered Quantity	1,612,671	363	1,844,539
Avg. Distinct count of Material	47	2	12
Median Weeeks With Orders	69	2	18
Avg. Rejection	2.92%	49.63%	3.45%
Avg. Return %	0.60%	0.00%	0.51%
Avg. Delay %	5.12%	0.00%	2.89%
Avg. Avg. Lead time	2	2	2
Avg. InFull rate %	93.44%	50.37%	93.53%

Figure 8: Clustering Characteristics for Combined Hospitals (Public and Private)

Cluster Characteristics:

1) High Worth Hospitals:

High Worth Hospitals have a significant net order value of €137,435K, with a median net order value of €842,159. These 105 hospitals are crucial contributors, with a high volume of delivered quantities (1,612,671) and items (55,974). Their average delivery cost as a per cent

of net-order value is extremely low at 0.12%, indicating efficient operations. Additionally, they maintain a low average rejection rate of 2.92% and a strong average in-full rate of 93.44%, reflecting reliable and consistent performance.

2) Low Worth Hospitals:

Low Worth Hospitals have the smallest net order value of €6K, and with only 15 hospitals in this cluster, they contribute minimally in terms of revenue and order quantities. However, this cluster shows significant challenges, with the highest rejection rate of almost 50%, indicating possible issues with either stringent acceptance criteria or process inefficiencies. Their average lead times and in-full rates are on par with the other clusters, but their high rejection rate indicates a need for closer scrutiny and potential improvements.

3) Medium Worth Hospitals:

Medium Worth Hospitals, though contributing less than High Worth Hospitals, have a significant net order value of €98,011K. With 2,205 hospitals in this cluster, they exhibit a balanced approach with moderate volumes of delivered quantities (1,844,539) and items (152,183). They maintain a relatively low rejection rate of 3.45% and an impressive in-full rate of 93.53%, similar to High Worth Hospitals.

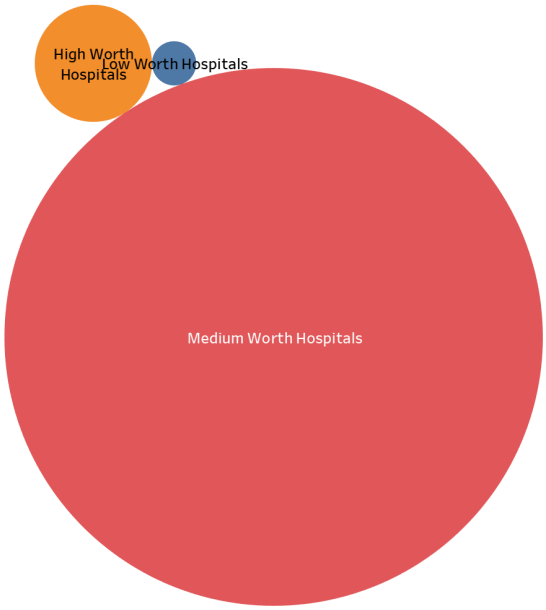

Figure 9: Combined Hospitals customers' size

4.2.4 Private Hospitals

Pfizer's 1331 private hospital customers, split into three clusters, account for about 2.46% of total revenue. The average rejection rate for high-activity hospitals and low-activity hospitals is greater than 2% which is considered high.

Private Hospital

	High Activity Hospitals	Low Activity Hospitals	Moderate Activity Hospitals
Distinct count of Customer ID	163	1,157	9
Net Order Value	€ 34,805K	€ 14,212K	€ 12,126K
Median Net Order Value	€ 145,443	€ 2,014	€ 1,309,072
Avg. Std Dev Net Order Value	€ 2,861	€ 645	€ 15,631
Avg. DeliveryCostPercentofDeliveryNetvalue	0.19%	1.15%	0.11%
Median Distinct count of Sales Document	90	13	131
Count of Item	33,972	36,665	3,527
Count of Delivered Quantity	589,002	387,849	132,511
Avg. Distinct count of Material	24	8	39
Median Weeeks With Orders	58	13	64
Avg. Rejection	2.24%	4.41%	1.00%
Avg. Return %	0.20%	0.44%	0.25%
Avg. Delay %	3.93%	3.09%	4.82%
Avg. Avg. Lead time	2	2	2
Avg. InFull rate %	95.15%	92.74%	96.87%

Figure 10: Clustering Characteristics for Private Hospitals

Cluster Characteristics:

1) High Activity Hospitals:

Representing 12% of customers, these hospitals lead in delivered quantity and maintain other metrics close to the best.

2) Moderate Activity Hospitals:

Though only 9 in number, they have the highest median net order value, lowest delivery cost ratio, highest order consistency in terms of weeks with orders, and the best in full rate, but face the most delays

3) Low Activity Hospitals:

Despite engaging 86% of customers, they contribute a similar net order value to the 9 moderate activity hospitals, with the highest average rejection rate and average delivery cost per cent while having the lowest median weeks with orders and net order value.

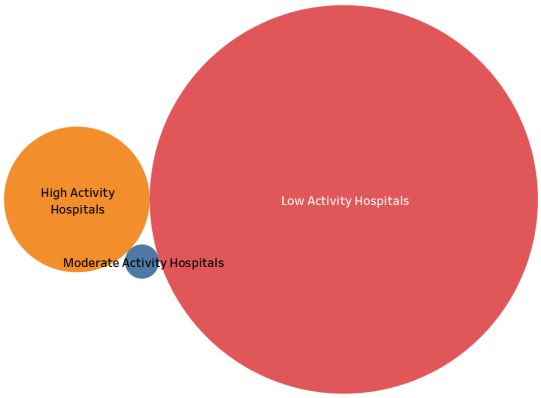

Figure 11: Private Hospitals customers' size

4.2.5 Public Hospitals

Pfizer's 998 public hospital customers, split into three clusters, account for about 7% of total revenue. The average rejection rate for high-activity hospitals and low-activity hospitals is greater than 2% which is considered high. Similarly, the average delay rate for very high-activity hospitals is greater than 7% which is considered high.

Public Hospital

	High Activity Hospitals	Low Activity Hospitals	Very High Activity Hospitals
Distinct count of Customer ID	211	786	1
Net Order Value	€ 120,214K	€ 26,979K	€ 27,116K
Median Net Order Value	€ 272K	€ 8K	€ 27,116K
Avg. Std Dev Net Order Value	€ 5,043	€ 1,508	€ 41,576
Avg. DeliveryCostPercentofDeliveryNetvalue	0.19%	0.84%	0.01%
Median Distinct count of Sales Document	122	20	1,253
Count of Item	87,623	43,845	2,572
Count of Delivered Quantity	1,576,965	512,346	258,900
Avg. Distinct count of Material	45	11	102
Median Weeeks With Orders	63	18	74
Avg. Rejection	2.11%	3.49%	0.66%
Avg. Return %	0.38%	0.70%	0.54%
Avg. Delay %	2.82%	2.62%	7.39%
Avg. Avg. Lead time	2	2	2
Avg. InFull rate %	94.40%	93.24%	96.93%

Figure 12: Clustering Characteristics for Public Hospitals

Cluster Characteristics:

1) **High Activity Hospitals** :

A single hospital contributes 15% of the net order value, matching the total of all low-activity hospitals. It boasts the highest median net order value, and lowest delivery cost ratio, but faces high delay rates. These hospitals generate 68% of public hospital revenue, with the highest item count, delivered quantity, and the lowest return percentage.

2) **Low Activity Hospitals:**

Making up 78% of public hospitals, they have the lowest median net order value, high delivery cost ratio, low order frequency, and the highest average rejection rate.

3) **High-Activity Hospitals:**

They generate 68% of public hospital revenue, leading in item count, delivered quantity, and maintaining the lowest average return percentage.

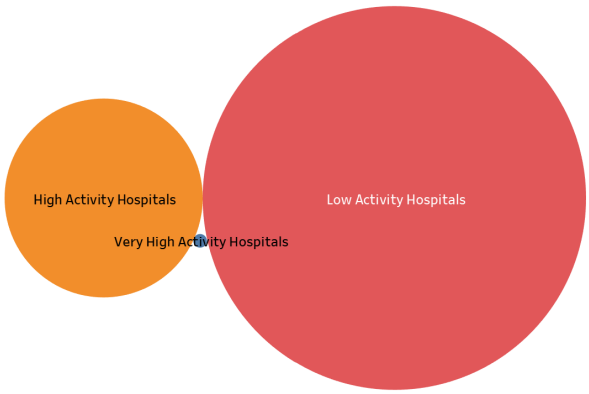

Figure 13: Private Hospitals customers' size

Chapter 5 - Conclusion and Future Research

5.1 Conclusion

We have 12 clusters, three each of four customer groups and three clusters with all hospitals put together. Hence we have arrived at 15 clusters for 12, 470 customers of Pfizer from France. Amongst three clusters for all customer groups, there exists a cluster with the lower number of customers in it (Large wholesalers, high-worth pharmacies, high-worth hospitals) which contributes disproportionately to the revenue. These clusters cover 25% of customers while accounting for 86% of the total revenue, however, they account for 88% of all the rejections.

On the contrary, there exists another cluster with more customers from the group but contributes significantly less in terms of revenue, these clusters are Low-value pharmacies: (4k - 40% customers, 4% of pharmacies revenue), Low Activity Private Hospital cluster (1K customers, small size orders), Low Activity Public Hospital cluster.

5.2 Answering Business Questions

The first requirement was to develop a robust customer clustering model focusing on value and volume metrics that categorises Pfizer's customer base on existing customer segments - Wholesale, Retail, and Hospitals (public/private). In response to this, we developed 15 customer clusters from 20 different models focusing on value and volume metrics, effectively categorising Pfizer's customer base into Wholesale, Retail, and Hospitals (public/private) segments. The intra-clustering results further refined these categories, revealing distinct customer characteristics and behaviour patterns discussed in section 4.2. As detailed in Chapter 4, this clustering approach optimises customer prioritisation and supports Pfizer's strategic goals.

The second objective was to deploy the customer clustering model in a dashboard for visualisation, allowing easy access and interpretation of customer segments, and enabling users to rank customers within each cluster. We have implemented this in a Tableau dashboard, which provides a clear and intuitive interface for exploring the clustered data and cluster characteristics. Pfizer's can effortlessly analyse and rank customers within each cluster, aiding in strategic decision-making. However, this is not deployed but we are sharing tableau desktop files.

The third goal was to ensure the model enhances customer prioritization to support effective contract negotiations. The final output of the project achieves this by providing rankings for each cluster, which is indicated by its name and within clusters, ranking to each customer is assigned based on value and volume metrics. By analysing broader characteristics like Net Order Value, transportation costs, and delay rates, the output equips Pfizer with data-driven insights to negotiate contracts more effectively and realistically.

Lastly, we have addressed the business question of analysing the developed clusters to identify underlying issues and providing strategic recommendations to Pfizer aimed at optimising operations and achieving cost savings. We have formulated recommendations grounded in industry knowledge and scholarly works. The first set of recommendations has been consolidated into an organizational proposal, specifically a job description for a role called Wholesaler Operations Manager (WOM). The second issue is mitigating higher transportation costs. Based on industry best practices, we suggest a shared-resources logistics approach on the supplier side and zone skipping and centralized fulfilment as a customer-side strategy. Further, we propose a different business model where smaller customers are consolidated to purchase through Pfizer's wholesalers. The detailed discussion of these recommendations can be found in Section 5.3 below.

5.3 Recommendations

We have identified the main issues, first one the major one is the very high rejection rate with large wholesalers and high-worth private hospitals. The acceptable rejection rate (industry standard) for pharmaceutical supply chains is 2% (Alicke et al,2014), but for large and moderate-volume wholesalers it is 8.64% and 9.79%. High-worth private hospitals also have a high rejection rate of 9.86%. Furthermore, the reason for rejection in 93% of all rejections is that the customer wouldn't accept backorders (partial orders). Pfizer is unable to fulfil the order, i.e. Pfizer is short of these materials and can fulfil the order in multiple orders.

The rejection for overall wholesalers is 8.5% (ratio of rejected item lines to total item lines) which accounts for €18M of net order value. This €18M also accounts for 90% of overall rejections, which is €19.9M. Hence, if Pfizer can achieve 2% of rejection rate, that would be a sale of €13M assuming that each item line costs the same amount (€18M/47114 = €382 per item line). If one can reduce rejection rate to 5%, then it's a sale of €7.4M.

Hence, owing to the magnitude of cost savings for Pfizer our first set of recommendations include solutions to reduce rejections. Here we have arrived at three major recommendations and two generic recommendations.

1. Improve Demand Forecasting

Accurate demand forecasting is critical to avoid stockouts and reduce backorders. Investing in advanced demand forecasting tools and collaborating with customers on forecasts can help minimize backorders. George and Elrashid (2023) found that demand forecasting significantly enhances drug supply chain performance, which can lead to fewer stockouts and rejections. Specifically, they emphasize that effective demand forecasting is crucial for managing inventory levels and ensuring that the right products are available when needed, thereby minimizing the likelihood of order rejections from customers.

2. Enhance Supplier- Customer Collaboration

Collaborating closely with suppliers (here it is Pfizer) is important to ensure adequate supply to consumers (Praveen and Suraj, 2021). Partnering with suppliers on forecasts, inventory, and production plans can reduce the risk of backorders (FasterCapital, 2024). Pfizer should partner with wholesalers and other customers in joint inventory and production plans and forecasts.

3. Optimize Safety Stock Levels

Finding the right balance of safety stock is critical. One study noted "Finding the right balance to optimize inventory based on the product portfolio and on historic demand and shipment patterns will be a continuing challenge." (Supply Chain Brain, 2021) Leveraging demand forecasts and supply chain analytics can help determine optimal safety stock levels to minimize backorders.

Next, we have the two generalist recommendations based on the literature survey.

4. Increase Inventory Visibility

Real-time inventory visibility across the supply chain is key to proactively managing stockouts. A survey found that "more than a quarter of firms had experienced damaged, spoiled or lost inventory. Roughly 90% reported said they didn't have full visibility into their supply chains and didn't trust the in-transit data they were receiving." (Rockeman, 2022). Deploying active tracking devices and analytics can provide the visibility needed to avoid backorders.

5. Improve Order Fulfilment Processes

Streamlining order fulfilment processes can help avoid stockouts. A study proposed "A unique supply chain disruption shortage management roadmap...which could help guide practice and drug delivery by providing strategic insights on a system-wide view of drug delivery performance." (Chadist, 2021). Mapping and optimizing order management processes can minimize backorders.

Based on the above recommendations from various sources, we observed that these are the responsibilities of a role known as POMs (partner operation managers) at Cisco Systems which is the second best on Gartner's global supply chain rankings in 2023 (Gartner, 2023). Two of the author's experience at Cisco Systems helped recognise this and hence we are proposing a role (if it doesn't exist at Pfizer already) Wholesaler operations manager with the following role and responsibilities:

Role

Be a wholesaler's advocate within the organisation's supply chain and help streamline and smoothen order fulfilment at each stage of the process from quote to order closure.

1. Manage stakeholders

a. Engage in joint inventory and production planning sessions with wholesalers.

b. Escalation management concerning issues raised by wholesalers and co-ordinate with different teams within Pfizer to alleviate problems.

2. Own the forecasting process and result quality for the wholesaler while partnering with engineering teams within Pfizer. Define the process to sense the wholesaler's inventory levels.

3. Come up with a unique supply chain disruption shortage management roadmap.

The second issue we have identified is the higher transportation cost. Alicke et al (2014) suggest that the best pharma companies have 12% of COGS (cost of goods sold) as supply chain cost. Supply chain cost would include warehousing, transportation, supply chain overhead staff, inventory holding, and obsolescence. However, based on Pfizer's 2023 annual income statement (Wall Street Journal, 2024), COGS was 40% of the revenue/sales. We have used net order value as a proxy to revenue/sales and hence we arrive at 4.7% (12% supply chain cost of 40% COGS). While 5% of the revenue should be supply chain cost, transportation cost should be much less than 5%, however, we are considering 5% as the benchmark to show an opportunity to save €441,000 annually on transportation costs.

We have low-value pharmacies accounting for 4,646 customers, €10 million in revenue and 5.41% average transportation cost of net order value. Hence, 5.41% of €10M, which is €541,000 is the transportation cost. If we are able to reduce the cost of transportation to 1%

we can save up to €441,000 annually with respect to France alone. We have chosen 1% of the revenue as a realistic transportation cost that would be upto 20% of the overall supply chain cost. Within our assessment, most of the clusters (except 4 out of 15 clusters) have this under 1%.

Based on the literature survey we have the following recommendations:

1. Shared-Resources Logistics

This is a method where different suppliers consolidate multiple orders with them to ship bulk orders to the same customers. Implementing shared-resources logistics can significantly optimize transportation operations. By collaborating with other manufacturers to consolidate shipments, companies can reduce costs associated with less-than-truckload (LTL) shipments. Cardinal Health's Exclusive Pharmaceutical Transportation Network (EPTN) is a prime example, where manufacturers share truck capacity and streamline delivery routes (Pai, 2014). Pfizer will have to partner with other pharmaceutical suppliers locally to consolidate these orders.

2. Zone Skipping and Centralized Fulfilment

Employing zone-skipping techniques can consolidate shipments and reduce costs. By grouping orders from multiple pharmacies or hospitals in the same geographic area, companies can take advantage of bulk shipping rates. Additionally, centralizing fulfilment from a main pharmacy rather than distributing from multiple locations can streamline operations and reduce transportation expenses (Jameela, 2024 and Intelligent audit, 2024). Pfizer can here work with customers (mainly hospitals and pharmacies) located in the same geographical area to consolidate their orders and then fulfil these orders.

Additionally, Pfizer can have a newer business model where the smaller customers can be offered incentives to buy through one of Pfizer’s wholesalers. Incentives which could include discounts which would offset the transportation cost or other credits. This model ensures that the revenue is not lost for Pfizer while transportation cost is reduced or eliminated and further all three parties gain from this model. Wholesaler gains a customer and revenue, pharmacy gains a supplier who could deliver quicker along with incentives and Pfizer could build a stronger relationship with its customers while saving on the transportation cost and maintaining the revenue.

5.4 Future Work

The current customer classification model provides insights into Pfizer’s diverse customer base in France, offering a solid foundation for strategic decision-making. However, several

avenues for future work can further enhance the model's accuracy and applicability across different regions and business contexts.

A significant opportunity lies in the integration of additional supply chain data to refine and expand the model's capabilities. Incorporating logistical costs, such as freight and carbon footprint, as well as operational costs like support case expenses and inventory holding costs, can provide a more holistic view of the supply chain. Moreover, integrating regulatory attributes of materials, particularly those related to highly regulated products such as Botox or radioactive materials, can improve the accuracy of the model and its relevance to diverse customer segments.

While the above work of supplementing with more data would improve the cluster (model) quality, recommendations give rise to one regression problem and two optimisation problems. The first regression problem is demand forecasting whose accuracy should be 75%, which is considered average and more than 86% is considered to be best in the pharmaceutical industry (Alicke,2014). The other two optimisation problems are, first is to optimize safety stock levels and second is to map very small pharmacies to wholesalers who are Pfizer’s customers. Data points like geographical regions and material ids could help in this mapping.

Appendices

-

Cluster results - combined hospitals

C-Group (group)	Combined Hospital Cluster	Distinct count of Customer ID	Net Order Value	Avg. Rejection	Avg. Delay %	Avg. Distinct count of Delivery Document	Avg. InFull rate %
Independent pharmacy	High Worth Pharmacies	3,094	€ 177.127M	0.66%	4.99%	14	96.91%
	Medium Sized Pharmacies	2,348	€ 64.983M	2.03%	4.51%	4	96.15%
	Very Small Pharmacies	4,626	€ 9.887M	0.27%	5.33%	4	97.82%
Pharma Wholesaler	Large Wholesalers	7	€ 1,834.801M	8.64%	0.86%	14,625	80.26%
	Moderate Volume Wholesalers	14	€ 83.619M	9.79%	7.22%	275	80.13%
	Small Volume - High Performance Wholesalers	51	€ 72.477M	5.12%	2.61%	199	88.57%
Combined Hospitals	High Worth Hospitals	105	€ 137.435M	2.92%	5.12%	280	93.44%
	Low Worth Hospitals	15	€ 0.006M	49.63%	0.00%	2	50.37%
	Medium Worth Hospitals	2,205	€ 98.011M	3.45%	2.89%	37	93.53%

Table 9: Final Cluster (Combined Hospitals)

Cluster results - separate hospitals

Independent pharmacy	High Worth Pharmacies	3,094	€ 177.13M	0.66%	4.99%	14	96.91%
	Medium Sized Pharmacies	2,348	€ 64.98M	2.03%	4.51%	4	96.15%
	Very Small Pharmacies	4,626	€ 9.89M	0.27%	5.33%	4	97.82%
Pharma Wholesaler	Large Wholesalers	7	€ 1,834.80M	8.64%	0.86%	14,625	80.26%
	Moderate Volume Wholesalers	14	€ 83.62M	9.79%	7.22%	275	80.13%
	Small Volume - High Performance Wholesalers	51	€ 72.48M	5.12%	2.61%	199	88.57%
Private Hospital	High Activity Hospitals	163	€ 34.80M	2.24%	3.93%	125	95.15%
	Low Activity Hospitals	1,157	€ 14.21M	4.41%	3.09%	18	92.74%
	Moderate Activity Hospitals	9	€ 12.13M	1.00%	4.82%	181	96.87%
Public Hospital	High Activity Hospitals	211	€ 120.21M	2.11%	2.82%	207	94.40%
	Low Activity Hospitals	786	€ 26.98M	3.49%	2.62%	30	93.24%
	Very High Activity Hospitals	1	€ 27.12M	0.66%	7.39%	1,324	96.93%

Table 10: Final Cluster (Separate Hospitals)

References

Abdulhafedh, A., 2021. Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. Journal of City and Development 3, 12–30. https://doi.org/10.12691/jcd-3-1-3

Alicke, K., Ebel, T., Schrader, U. and Shah, K., 2014. Finding opportunity in uncertainty: A new paradigm for pharmaceutical supply chains. [pdf] McKinsey & Company. Available at: https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/pharma%20and%20medical%20products/pmp%20new/pdfs/finding_opportunity_in_uncertainty-introductory_chapter.ashx [Accessed 3 July 2024].

Chadist, P. (2021) Responding to disruptions in the pharmaceutical supply chain. The Pharmaceutical Journal. Available at: https://pharmaceutical-journal.com/article/research/responding-to-disruptions-in-the-pharmaceutical-supply-chain [Accessed 6 July. 2024].

Collaborating With Suppliers And Customers [WWW Document], n.d. . FasterCapital. URL https://fastercapital.com/keyword/collaborating-with-suppliers-and-customers.html (accessed 6.17.24).

Dang, S., 2015. Performance Evaluation of Clustering Algorithm Using Different Datasets. IJARCSMS 3, 167–173.

Das, T.K., 2015. A customer classification prediction model based on machine learning techniques, in: 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT). Presented at the 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), pp. 321–326. https://doi.org/10.1109/ICATCCT.2015.7456903

Gartner, 2023. These are the best supply chains of 2023. [online] Available at: https://www.gartner.com/en/articles/these-are-the-best-supply-chains-of-2023 [Accessed 13 July 2024].

George, S., Elrashid, S., 2023. Inventory Management and Pharmaceutical Supply Chain Performance of Hospital Pharmacies in Bahrain: A Structural Equation Modeling Approach. Sage Open 13, 21582440221149717. https://doi.org/10.1177/21582440221149717

Griva, A., Zampou, E., Stavrou, V., Papakiriakopoulos, D., Doukidis, G., 2024. A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data. Journal of Decision Systems 33, 1–29. https://doi.org/10.1080/12460125.2022.2151071

Hicham, N., Karim, S., 2022. Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. International Journal of Advanced Computer Science and Applications (IJACSA) 13. https://doi.org/10.14569/IJACSA.2022.0131016

Intelligent Audit, 2024. Pharmaceutical Shipping: Tips to Lower Transport Spend [WWW Document], n.d. URL https://www.intelligentaudit.com/blog/pharmaceutical-shipping-tips-to-lower-transport-spend (accessed 7.12.24).

Jameela, M., 2024. Last-Mile Costs through Regional Placement: The Power of Zone Skipping. URL https://wareiq.com/resources/blogs/what-is-zone-skipping/ (accessed 7.12.24).

Li, X., Lee, Y.S., 2024. Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm. Journal of Cases on Information Technology (JCIT) 26, 1–16. https://doi.org/10.4018/JCIT.336916

Nabeel Mustafa, S.M., Akhtar, A., Peter Noronha, J.T., Salman, M., Baig, M.A., 2023. Customer Segmentation using Machine learning Techniques, in: 2023 International Multi-Disciplinary Conference in Emerging Research Trends (IMCERT). Presented at the 2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT), pp. 1–7. https://doi.org/10.1109/IMCERT57083.2023.10075194

Pai, S., 2014. Creating the first, pharma-only 3PL network. Pharmaceutical Commerce, September/October, 9(5), pp. 31-32. Available at: https://coldchain-tech.com/images/documents/pharmcomm_20140910.pdf [Accessed 16 July 2024].

Praveen, T. and Suraj, V. V. (2021). A study on supplier and buyer coordination on pharma supply chain. Prayukti - Journal of Management Applications, 1(1), pp. 09-18. Available at: https://bschool.dpu.edu.in/download/journal/Volume1-Issue1/PJMA-M-02.pdf [Accessed 15 July 2024].

Praveen, T., V. Suraj, V., 2021. A study on supplier and buyer coordination on pharma supply chain. PJMA 01. https://doi.org/10.52814/PJMA.2021.1102

Rockeman, O. (2022) Oversupply, Damage Sees €163 Billion in Inventory Tossed Annually. Bloomberg, 10 November. Available at: https://www.bloomberg.com/news/newsletters/2022-11-10/supply-chain-latest-inventory-waste-totals-163-billion-a-year [Accessed 14 July. 2024].

Supply Chain Brain. (2022). Five critical challenges facing pharma supply chains. Available at: https://www.supplychainbrain.com/articles/34798-five-critical-challenges-facing-pharma-supply-chains [Accessed 17 July. 2024].

Tabianan, K., Velu, S., Ravi, V., 2022. K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. Sustainability 14, 7243. https://doi.org/10.3390/su14127243

Wall Street Journal, 2024. Pfizer Inc. Annual Income Statement. [online] Available at: https://www.wsj.com/market-data/quotes/PFE/financials/annual/income-statement [Accessed 3 July 2024].