Strathmore UNIVERSITY

SU+ @ Strathmore University Library

Electronic Theses and Dissertations

2023

Examining Gaussian Mixture Models using clustering algorithms.

Oloo, John Mark
Strathmore Institute of Mathematical Sciences
Strathmore University

Recommended Citation

Oloo, J. M. (2023). *Examining Gaussian Mixture Models using clustering algorithms* [Strathmore University].

http://hdl.handle.net/11071/15389

Follow this and additional works at: http://hdl.handle.net/11071/15389

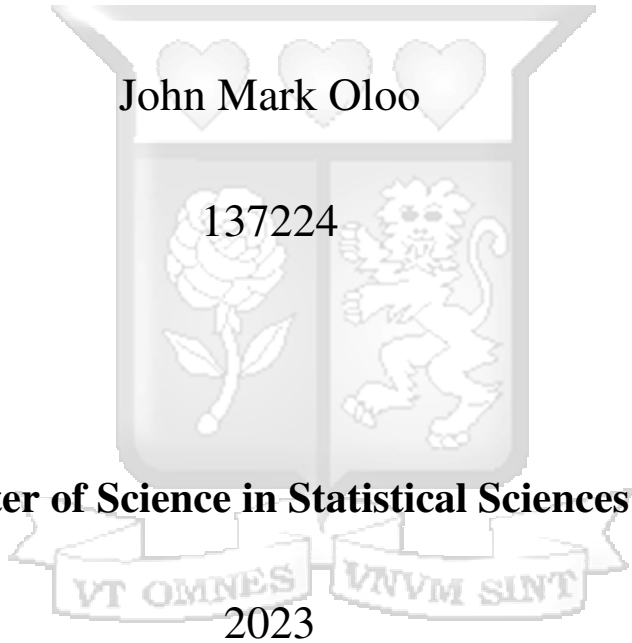
This work is availed for free and open access by Strathmore University Library.
It has been accepted for digital distribution by an authorized administrator of SU+ @Strathmore University.
For more information, please contact library@strathmore.edu

Examining Gaussian Mixture Models using Clustering Algorithms

By

John Mark Oloo

137224



Master of Science in Statistical Sciences

2023

Examining Gaussian Mixture Models using Clustering Algorithms

John Mark Oloo

137224

Submitted in total fulfilment of the requirements for the degree of Master of Science in Statistical Sciences of Strathmore University

Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

July 2023

This thesis is available for Library use through open access on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University.

Name: **John Mark Oloo**

Signature: 

Date: June 14, 2023

Approval

The thesis of John Mark Oloo was reviewed and approved by the following:

Dr. Collins Ojwang' Odhiambo

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Dr. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

Abstract

Clustering is an important data mining technique for finding homogeneous and heterogeneous groups in a data set. Identifying these groups from a sales data-set is important for estimating demand for a specific range of products. This research carried out a detailed analysis of Gaussian Mixture Models by using the expectation-maximization method to find optimal clusters on a sales data-set. The method combines expectation-maximization algorithm with the agglomerative hierarchical clustering, resulting in an effective, iterative process for estimating the model's parameters. In order to give accurate estimates for the ideal number of clusters, the expectation-maximization approach uses the hierarchical clustering to provide an initial guess for the algorithm. The goal is to boost sales performance of products sold by estimating demand and comparing sales over a particular period. The method segmented clients into groups with shared characteristics, such that customers within each subgroup could be offered products and promotions that are likely to interest them. Therefore, this study was interested in maximizing the distance between individual clusters and also minimizing the distance between items belonging to the same cluster.

The research experimented with sales data from a large liquor distribution company, examining how variables such as product, customer, sales region, and quantity sold affected overall sales volume and revenue. In order to identify deviation in product sales, the data-set was split into subsets. Also, before clustering and data pre-processing, exploratory data analysis was used to understand the features of the data. To correctly measure the performance of the clustering algorithm the study used the Bayesian Information Criterion as a goodness of fit metric. The results had two distinct clusters that represented analysis of 146 products and 223 customers from the dataset. These findings confirmed that Gaussian Mixture Models and EM algorithms are more effective at estimating the underlying key parameters and identifying subgroups of similar products and customers.

VT OMNES UNUM SINT

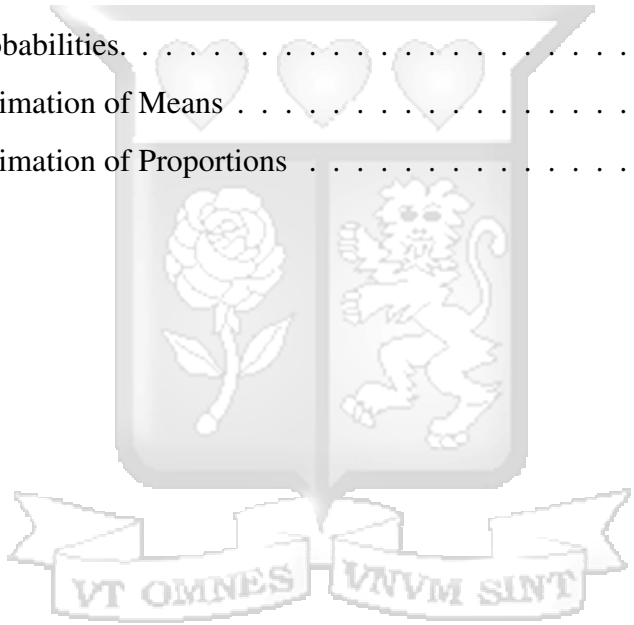
Table of contents

VT OMNES VNVM SINT

List of figures

VT OMNES VNVM SINT

List of tables



List of abbreviations

ERP	Enterprise Resource Planning
EM	Expectation-Maximization
KDV	Kensington Distillers and Vintners
GMM	Gaussian Mixture Models
SKU	Stock Keeping Unit
CL	Complete Linkage
SF	Sales Force
SL	Single Linkage
CRM	Customer Relationship Management
i.i.d	Independent and Identically Distributed
VAT	Value Added Tax
GA	Group Average
MDEA	Monument Distillers East Africa
BIC	Bayesian Information Criterion
MLE	Maximum Likelihood Estimation
MCMC	Markov Chain Monte Carlo
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
KPCA	Kernel Principal Component Analysis

Acknowledgement

My sincere gratitude goes to Dr. Collins Odhiambo, the supervisor of my dissertation, for his unwavering support and direction during the course of my research.

I would also like to express my appreciation to the instructors and staff at the Strathmore Institute for Mathematical Sciences for their guidance and instruction throughout this academic program.

In gratitude for their support along the way, I will always remember my friends Daniel Maangi and Dominic Lemusi, as well as my family.

I appreciate you everyone.

Dedication

This thesis is dedicated to my mother Elizabeth Mukhwana and my wife Sheilla Atieno Oloo.

Chapter 1

Introduction

1.1 Background to the Study

In a distributor business model, there exists a wide variety of methods for managing operations such as supply chain management, marketing plans, market research, and customer and product profiling. In either case, clustering this set of data elements is fundamental; otherwise, most of the data deposited will be of little benefit to anyone. Clustering is critical for knowledge discovery in such operations. It divides a data-set into various groups so that observations within each group are fairly similar, while observations in other groups are quite different (James et al., 2013). To understand whether observations are similar or different, this study took a few considerations. Assume we have a collection of n observations, each with q characteristics. These observations could represent various stock keeping units (SKUs), while the characteristics could include Sales Force (SF) measurements collected for each unit, such as the number of active customers or the quantity of products sold. There is a possibility that there are differences among the subgroups within the SKUs.

Clustering, which employs unsupervised learning techniques, is a widely used statistical data analysis method in this scenario (Seif, 2018). By employing clustering analysis, it is possible to extract meaningful information from the data by observing the groups that data points are categorized into when a clustering algorithm is applied. The selection of appropriate distance or dissimilarity measures for comparing two attributes is essential in all clustering techniques and various algorithms can be used to achieve this objective.

This study described two clustering algorithms and validated their performance through numerical examples;

1.1.1 Agglomerative Hierarchical Clustering

Hierarchical Clustering involves organizing data items into a tree structure with a predetermined number of clusters (HanumanthSastry and PrasadaBabu, 2013). This tree structure, known as a dendrogram, illustrates the grouping of units, where units within the same cluster are connected. The y-axis of the dendrogram represents the distances between the units within each cluster (Ledolter, 2013). Starting from the bottom, the individual units are depicted as leaves and are gradually combined to form smaller branches, which in turn merge to create larger branches. Eventually, the root of the tree represents a single cluster that includes all units.

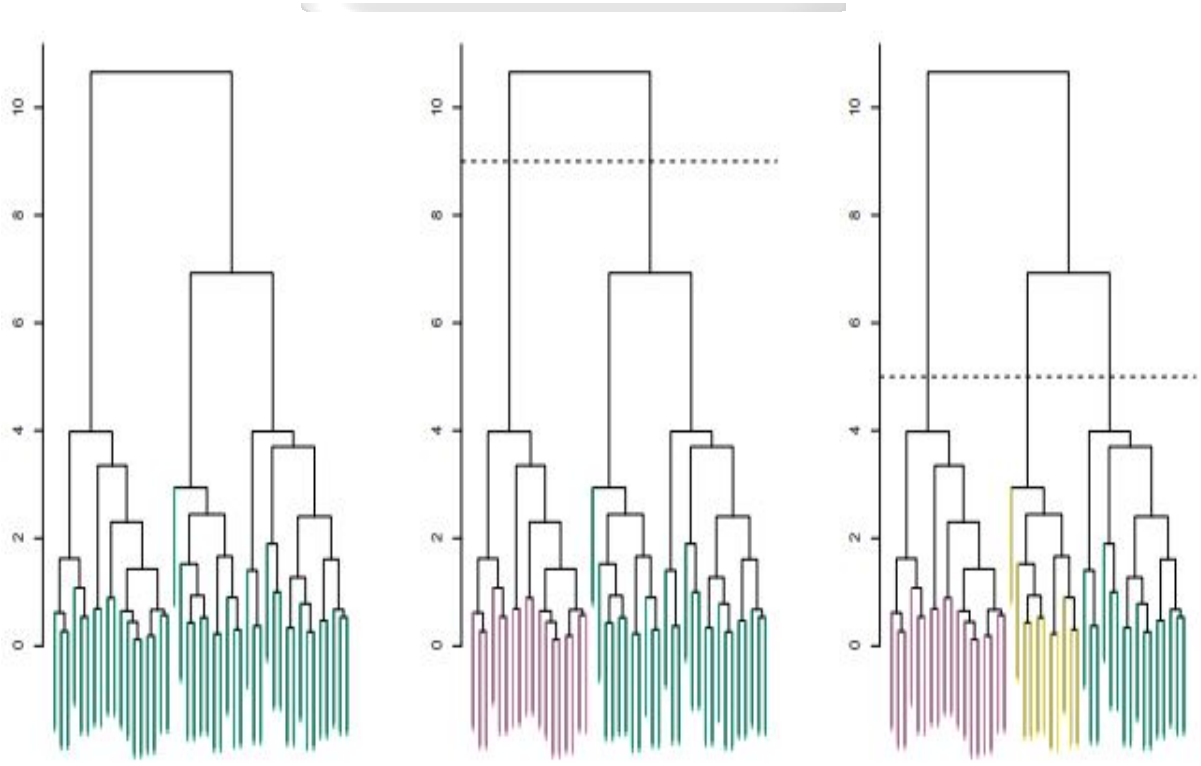


Figure 1.1: Dendogram

In this hierarchical approach, each object is initially considered as its own cluster (Sastry and Babu, 2013), and as the hierarchy is traversed, pairs of clusters are merged together (Maingi, 2015). The process starts with each observation being treated as a separate singleton cluster. In each of the $N-1$ steps, the two closest clusters are merged, resulting in one less cluster at the subsequent higher level. This merging is based on a dissimilarity measure that determines

the similarity between clusters.

Let A and B denote two groups. The dissimilarity distance, denoted as $d(A,B)$, between A and B is calculated using the pairwise observation dissimilarities, represented as d'_{ii}, where one member of the pair i is in A, and the other i' is in B. In agglomerative hierarchical clustering procedures, both a distance measure and a linkage criterion are required (Ledolter, 2013). The distance measure between two objects with feature vectors $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $x_j = (x_{j1}, x_{j2}, ..., x_{jp})$ is a non-negative and symmetric value, $d(x_i, x_j) = d(x_j, x_i) \geq 0$, where $d(x_i, x_j) = 0$ indicates $x_i = x_j$, and it satisfies the triangle inequality $(d(x_i, x_j) \leq d(x_i, x_k) + d(x_j, x_k))$. The Euclidean norm $d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(x_{ir} - x_{jr})^2}$ is commonly employed as a distance measure. The following three criteria are utilized to define the dissimilarity between groups in the context of agglomerative clustering.

Single linkage (SL) agglomerative clustering considers the inter-group dissimilarity to be the minimum dissimilarity between any pair of observations, defined as:

$$d_{SL}(A,B) = \min d_{ii'}, \text{ where } i \in A \text{ and } i' \in B.$$

Complete linkage (CL) agglomerative clustering, on the other hand, takes the inter-group dissimilarity to be the maximum dissimilarity between any pair of observations:

$$d_{CL}(A,B) = \max d_{ii'}, \text{ where } i \in A \text{ and } i' \in B.$$

Group average (GA) clustering calculates the average dissimilarity between groups based on the pairwise dissimilarities:

$$d_{GA}(A,B) = \frac{1}{N_A N_B} \sum_{i \in A} \sum_{i' \in B} d_{ii'},$$

where N_A and N_B are the respective number of observations in each group.

Unlike K-Means, Hierarchical Clustering eliminates the need for specifying the number of clusters in advance. Instead, we have the flexibility to explore and select the most suitable number of clusters based on the constructed tree (Seif, 2018). Hierarchical clustering is particularly advantageous when dealing with data that exhibits a hierarchical structure, aiming to uncover and capture the hierarchical relationships. However, in the context of the sales data discussed here, such hierarchical structure is not present. To address this, the study employed a Gaussian mixture approach that combines hierarchical clustering with the Expectation Maximization (EM) algorithm by using the initial estimation for the EM algorithm as the output of hierarchical clustering, thus enhancing the clustering process.

1.1.2 Expectation Maximization (EM) Clustering using Mixtures of Normal Distributions

Expectation-maximization algorithm presents an alternative approach to clustering sales data, offering greater flexibility in terms of cluster covariance. This algorithm serves as a method for conducting maximum likelihood estimation in the presence of latent variables. It achieves this by initially estimating the values of the latent variables and subsequently optimizing the model (Brownlee, 2019). The expectation-maximization algorithm is a powerful and versatile technique commonly employed for density estimation when dealing with hidden data. In the context of this study, it was applied to the Gaussian Mixture Model. The mixture modeling assumption proposes that the data is an independent and identically distributed (i.i.d) sample from a population characterized by a probability density function (Hastie et al., 2009). This density function is defined as a combination of component density functions, with each component density representing one of the clusters. The model is then fitted to the data using maximum likelihood estimation or corresponding Bayesian approaches. The EM algorithm has found widespread application in econometric and sociological studies where there are unknown factors influencing the outcomes (Schmee and Hahn, 1979).

The algorithm comprises two key steps: an expectation step and a maximization step. During the expectation step, the algorithm calculates the expected values of the unknown underlying variables, taking into account the current parameter estimate and conditioned on the observed

data (Moon, 1996). Subsequently, the maximization step generates a new parameter estimate based on the computed expectations. These two steps are iteratively repeated until convergence is achieved. This method is particularly valuable in estimating substitute and lost demand when only sales and product availability data are observable. This is because not all products are consistently available in every period, owing to stock-outs and availability controls. Additionally, when performing the estimation, the study also considered the aggregate market share to account for broader market dynamics.

1.2 Monument Distillers in Kenya

Due to rapid population growth, a growing middle class, and a dynamic private sector, the liquor industry in Kenya has experienced impressive growth. Although East African Breweries Limited (EABL) continues to dominate the market, it has faced strong competition from local brewers and importers of international brands in recent years. EABL currently holds approximately 90% share of the Kenyan beer market and is expanding its presence in other parts of East Africa (Laura Wood, 2019). It is noteworthy that consumers in East Africa are increasingly purchasing entry-level branded products and emerging brands due to their purchasing power. Kensington Distillers and Vintners (KDV), a London-based premium beverages company, expanded into the East African market under the brand name Monument Distillers East Africa (MDEA). As part of its expansion strategy, and in line with other multinational companies, MDEA established offices in Kenya in 2019 with the primary goal of targeting the emerging middle class consumers with its premium drinks offerings. The company has already hired more than 30 employees, highlighting KDV's commitment to generating employment in the countries where they invest.

Monument Distillers, operating in East Africa, holds the distribution rights for a wide range of global brands from Sazerac, a leading American spirits company. These brands include Popov Vodka, Southern Comfort, Myers Rum, Firewater, Paddy's Irish Whisky, and Buffalo Trace. Additionally, their portfolio features international brands such as Martini, Bacardi,

Four Cousins, Montego Rum, Squadron Rum, Lupini Sambuca, Bertrams Brandy, Cape Velvet, Bannerman's Finest Scotch Whisky, and Grace du Roi Fine Wine. This extensive range of brands gives Monument Distillers a competitive advantage in the market, as the target market seeks association with iconic, high-quality, and premium international brands.

In Kenya, the per capita alcohol consumption is 3.4 liters, which is a third less than what individuals in Uganda and Tanzania consume on average per year. Uganda, Tanzania, and Rwanda all have above-average alcohol consumption rates of more than 9 liters per person per year. Among these countries, 56% of alcohol consumption comes from formal sources, including recorded beer, liquor, and wine sales, while the remaining 44% is from informal sources such as chang'aa and other illicit alcohol. Despite its size, the liquor industry in Kenya has demonstrated resilience during economic downturns, making it a profitable and stable sector. Monument Distillers East Africa has experienced steady sales growth in Kenya, driven by increased marketing efforts and engagement in the market facilitated by its enthusiastic staff.

1.3 Statement of the Problem

The process of sales forecasting can be complex (Jung and Jeong, 2012), as it involves uncertainties stemming from consumer behavior, posing risks to the liquor distribution industry. The manufacturing of products for sale relies on innovation, driven by the anticipation of consumer needs and the decision to fulfill those needs. However, this anticipation alone is not sufficient. There must be a demand for these products, which requires customers with the ability, desire, and willingness to make purchases.

In retail demand forecasting, two significant challenges arise: estimating lost demand when items are sold out and accurately accounting for substitution effects among related stock keeping units (SKUs). Many retail demand forecasting models are based on observed sales data, assuming that each SKU receives an independent flow of customer requests (Vulcano et al., 2012). However, neglecting the demand that is lost when a customer's preferred choice is unavailable leads to a negatively biased demand forecast. This bias can be especially sig-

nificant if products remain unavailable for extended periods. Simultaneously, when products are out of stock, customers may turn to available substitute products, leading to increased sales. Neglecting to account for the sales of these substitute products in demand forecasting introduces an overestimation bias among the set of available SKUs. It is crucial to address these issues to obtain an accurate estimate of the true underlying demand for products. The uncertainty of sales, coupled with the complexities and ambiguities involved, significantly impacts supply chain performance (Zhang, 2004). Various factors, including promotions, market trends, and special events, further contribute to the complexity of sales modeling. Promotions, for instance, are a common practice at Monument Distillers, aimed at boosting sales and influencing demand dynamics. The combination of different promotional strategies, frequency, price adjustments, and product displays in outlets can yield sales enhancements (Blattberg and Neslin, 1993), resulting from increased purchasing rates or higher consumption.

A common challenge in modeling involves estimating a joint probability distribution for a dataset of this nature (Brownlee, 2019). Density estimation requires selecting a probability distribution function and its corresponding parameters that best describe the joint probability distribution of the observed data. Maximum likelihood estimation is a popular approach for solving this problem, treating it as an optimization task to find the parameter values that provide the best fit to the joint probability of the data sample. However, this approach assumes that the dataset is complete or fully observed, meaning that all relevant variables are present. However, this assumption does not always hold true, as there may be hidden or latent variables that cannot be directly observed but still influence other variables in the dataset. In the case of the sales data mentioned, some data is missing, and the number of underlying data points is unknown. The EM algorithm is well-suited for addressing such problems as it employs iterative methods to estimate the underlying parameters. In this case, the goal is to estimate the demand by SKU using a set of historical data. Unlike conventional maximum likelihood estimation, the presence of latent variables makes the EM algorithm more effective, as computing the maximum likelihood estimate becomes challenging when latent variables are involved (Murphy, 2018). The EM algorithm operates by alternating between computing conditional expected values of the parameter estimates to obtain an

expected log-likelihood function (E-step) and maximizing this function to obtain improved estimates (M-step) (Vulcano et al., 2012). In this study, the EM algorithm is applied to historical sales data from a one-year period to measure demand and correct for bias among the available SKUs.

1.4 Objectives of the Study

1.4.1 General Objective

The primary goal of this research was to discover inherent cluster structure within sales data by utilizing a Gaussian Mixture Model.

1.4.2 Specific Objectives

1. To apply the Gaussian Mixture distribution by combining hierarchical clustering with the EM algorithm
2. To estimate the likelihood that a data point belongs to the Gaussian distribution
3. To partition the sales data to detect deviation in product sales thus comparing sales over a particular period

1.4.3 Research Questions

1. Does grouping similar data points together into clusters accurately predict customer purchasing behavior ?
2. Does using hierarchical clustering to provide an initial guess for the EM algorithm improve the accuracy of the model ?
3. How do the latent variables in the data influence sales performance ?

1.5 Justification

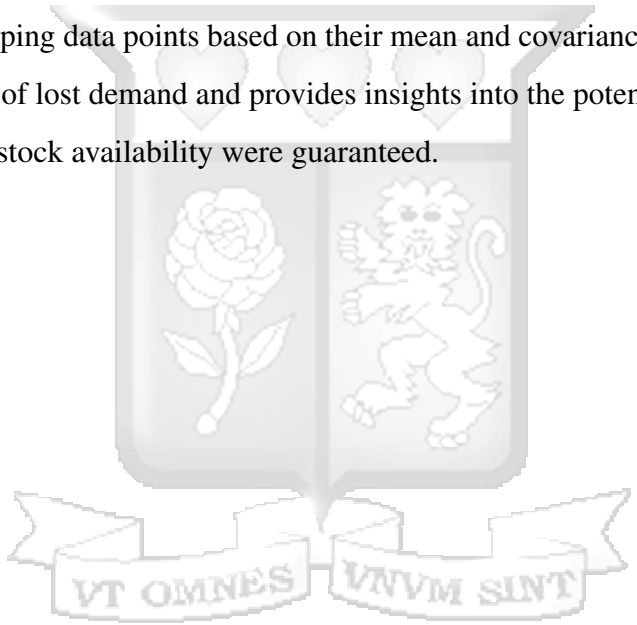
Statistical methods are commonly employed to estimate the parameters of probability distributions in order to accurately represent the density of a given training dataset. The Gaussian mixture model (GMM) is a specific type of mixture model that combines multiple normal probability distributions, requiring the estimation of mean and standard deviation parameters for each component. Various techniques exist for parameter estimation in GMMs, with maximum likelihood estimation (MLE) being widely used. In the dataset examined in this study, the data points originate from two distinct processes. Each process follows a normal distribution, but due to their similarities, it is difficult to assign a point to a specific distribution. These processes correspond to latent variables that affect the data but are not directly observable. In such cases, the EM algorithm is a suitable approach for estimating the distribution parameters by assigning probabilistic class membership rather than categorical assignment. Unlike algorithms that categorically assign instances, the EM algorithm guarantees finding model parameters with equal or greater likelihood at each iteration. It addresses the presence of latent variables by producing maximum likelihood estimates when there is a complex mapping from an underlying distribution to the observed distribution. The sales data analyzed in this study exhibits a many-to-one relationship, where a customer may purchase multiple products, and vice versa. The resulting estimates inherit the desirable statistical properties of MLE, including consistency, asymptotic normality, and asymptotic efficiency.

VT OMNES UNVM SINT

1.6 Significance of the Study

This study aimed to provide valuable insights for individuals interested in using clustering procedures for predictive modeling. The analysis of sales data series plays a crucial role in informing managerial decisions, such as demand forecasting, inventory management, and production planning. Additionally, this study contributes to the understanding of the key factors determining success in the liquor industry in Kenya and East Africa as a whole, benefiting stakeholders including liquor manufacturers, marketers, and sellers.

The utilization of clustering with GMM-EM, based on the normal data distribution, proves to be a suitable tool for analyzing sales data. It offers the advantage of incorporating clusters and effectively grouping data points based on their mean and covariance. This model enables accurate estimation of lost demand and provides insights into the potential customer base for various products if stock availability were guaranteed.



Chapter 2

Literature Review

2.1 Introduction

In this chapter, we present a comprehensive review of previous research on cluster analysis using the EM algorithm for Gaussian Mixtures. We delve into the underlying concepts of these methods and explore their diverse applications in the literature.

2.2 Clustering Algorithms

2.2.1 Hierarchical-based Clustering

In a study by Jung et al. (2009), a methodology for business process analysis was proposed, which employed hierarchical clustering based on process similarity. The aim was to analyze a collection of accumulated process models and provide assistance for designing new processes. The researchers demonstrated the methodology using example processes and showcased how hierarchical merged models could be induced from the identified process clusters. The proposed method proved to be beneficial in leveraging structural similarity metrics for business processes.

In another study conducted by Argüelles et al. (2014), hierarchical clustering techniques were utilized to identify regional business clusters as an initial step in developing cluster-based strategies. This approach offered two key advantages. Firstly, it involved applying objective clustering methods to the results of principal components analysis, resulting in a more refined cluster solution. Secondly, a mixed algorithm was employed during the clustering process, enhancing the reliability and stability of the final outcomes.

Kettani et al. (2014) discussed the popularity of agglomerative hierarchical clustering algorithms due to their versatility and conceptual straightforwardness. They presented a simplified algorithm with a reduced computational complexity specifically designed for clustering large datasets. Experimental evaluations conducted on several standard datasets confirmed the effectiveness of the proposed approach.

In a study conducted by Kalsoom and Halim (2013), hierarchical clustering techniques were used for the purpose of classifying individual drivers based on their driving behavior. The driving state of each driver was carefully examined, and the results were compared to determine the most suitable and efficient clustering approach for the given data. The clustering procedures were then applied to the dataset to group the drivers according to their behavior. The findings demonstrated that the employed methods effectively grouped the data accurately, resulting in a satisfactory overall performance.

In their study, Kettani et al. (2014) presented a method for agglomerative hierarchical clustering that offers simplicity and ease of implementation. The proposed approach does not require any tuning parameters, except for the number of clusters k, making it particularly suitable for clustering large datasets. The effectiveness of the proposed approach was verified through experiments conducted on various widely-used datasets. However, it should be noted that the study did not include an appropriate data pre-processing method, which resulted in a decreased accuracy of the clustering. Consequently, the findings revealed a reduction in clustering accuracy for certain datasets..

2.2.2 Distribution-based Clustering

Preheim et al. (2013) presented a straightforward clustering algorithm based on distributions for operation calls. This algorithm considers both the genetic distance and the distribution of sequences across samples. The method is designed to detect errors that are systematically generated across samples. Through their study, the authors demonstrated that this approach achieves higher accuracy in grouping reads into call units, as evidenced by its performance in a mock community and environmental samples.

Steiner et al. (2015) conducted a study where they focused on clustering music characteristic texts that are shared and posted on various social networks. Their objective was to provide personalized music recommendation services within specific activity contexts. To achieve this, they employed a Gaussian distribution-based approach. They proposed an algorithm that addresses the task of identifying and grouping precise and near-duplicate media items originating from multiple social networks. Similarly, Yao et al. (2018) proposed a text clustering method based on Gaussian Mixture Models (GMM). Their method integrated both text similarity and commodity object similarity to normalize brand entities of e-commerce products. The findings of their study demonstrated that this approach exhibited strong overall performance in the normalization of brand entities.

Yang et al. (2017) presented a novel approach for a knowledge recommendation system using Gaussian Mixture Models (GMM). Their aim was to predict users' ratings of knowledge items and provide personalized knowledge guidelines. To accomplish this, they trained a GMM-based user model using a Vector Space Model with Rating (VSMR). The trained user model was then employed to predict ratings for knowledge items, with higher scores indicating user interest and the highly scored items were recommended to the users. The effectiveness of their approach was evaluated through two experiments, which demonstrated its successful performance. The experiments utilized open-source data, and the results showcased the method's superior prediction accuracy.

Rahim et al. (2021) investigated the application of Gaussian Mixture Models (GMM) on vibration data obtained from an aircraft wing box structure for Structural Health Monitoring (SHM) purposes. To process the high-dimensional data, they employed Kernel Principal Component Analysis (KPCA) to reduce its dimensionality. KPCA transformed the continuous signal into discrete data, enabling the identification of clusters within the data. The study demonstrated that the GMM-EM based data clustering model effectively fitted the data density, considering operational variations. It emphasized the clustering of the reduced vibration data using KPCA, with a focus on SHM and the baseline's initial parameters.

An et al. (2015) conducted a study on predicting user interests in social networks using the GMM approach. They utilized real experimental datasets for verification purposes. The model operated through a series of steps: gathering user statistics and messages from Sina

Weibo, obtaining each user's interest eigenvalue sequence, and establishing a GMM model for user clustering. The GMM-based method took into account prediction accuracy and computational time. The findings demonstrated that GMM achieved a high prediction accuracy of 93.9%, while also managing computational complexity effectively.

Vulcano et al. (2012) presented a method that utilizes a distribution-based approach to estimate replacement and lost demand when certain products are unavailable during specific time periods. Their approach combined a multinomial logit demand model with a non-homogeneous Poisson model, considering practical data such as observed sales, product availability, and an aggregate market share estimate. The authors employed the EM procedure and treated observed demand as an incomplete observation of the primary demand. The EM algorithm demonstrated simplicity, efficiency, and provided an estimation of lost sales. The researchers confirmed that the EM algorithm possesses significant practical utility due to its ease of implementation, reliance on realistic input data, and the quality of its results.

Moon (1996) presented the EM algorithm for a signal processing task that estimated parameters of a probability distribution function. The author provided a detailed implementation of the algorithm to demonstrate its computational steps and highlight its versatility across various applications. The study concluded that the EM algorithm is a favorable choice for estimation tasks due to its adaptability and guaranteed convergence, making it a valuable option for addressing a wide range of estimation challenges.

Conlon and Mortimer (2013) devised an EM algorithm to address missing data in a wireless inventory system operating on a continuous time model of demand. They installed this system on 54 vending machines, monitoring product availability at four-hour intervals. The aim of their study was to illustrate how data reflecting short-term variations in the choice set can be utilized to identify substitution patterns, even when complete information about changes to the choice set is unavailable. The results revealed significant disparities in demand estimates, with the revised model projecting substantially greater impacts of stock-outs on profitability. However, implementing the E-step becomes challenging when multiple products experience stock-outs simultaneously, as it necessitates estimating an exponential number of parameters. This drawback is acknowledged in the study.

Talluri and Van Ryzin (2004) applied the EM method to estimate arrival rates and parameters

of a multinomial logit choice model, leveraging consumer-level panel data where no-purchase outcomes were unobservable. Their analysis focused on a single-leg yield management problem, where the choice behavior of buyers was explicitly modeled. The effectiveness of this approach was supported by empirical evidence provided by Vulcano et al. (2010). Ratliff et al. (2008) conducted a comprehensive review of demand uncensoring literature in the context of sales control scenarios. They also proposed a heuristic method to jointly estimate spill and recapture across multiple flight classes, employing balance equations that built upon the framework proposed by Andersson (1998). A comparable approach was previously presented by Ja et al. (2001).

In their study, Anupindi et al. (1998) introduced an approach to estimate consumer demand in cases where the preferred choice is unavailable. This model was specifically developed for an industry that significantly influences retail sales of various consumer products. The authors employed a continuous-time demand model and devised an EM method to address instances of stock-outs in a periodic review coverage. To handle the potential variables, they imposed a constraint that allowed at most two SKUs to be out of stock. By deriving maximum likelihood estimates (MLEs) of the demand parameters and examining substitution probabilities, they discovered that naively predicting demand rates based on observed sales rates led to biased results, even for products with minimal instances of low inventory.

2.2.3 Density-based Clustering

Campello et al. (2013) proposed an enhanced approach that combines density-based methods with hierarchical clustering to create a clustering hierarchy. From this hierarchy, they constructed a simplified tree of significant clusters. To obtain a "flat" partition comprising only the most relevant clusters, which may correspond to different density thresholds, the authors proposed a new measure called cluster stability. This measure formalizes the goal of maximizing the overall stability of selected clusters. They developed an algorithm that computes an optimal solution to this problem. Through their experiments, they demonstrated that this hierarchical clustering approach outperformed existing density-based clustering

methods across a wide range of real-world datasets.

Bohm et al. (2004) proposed a density-based algorithm to identify clusters of preference-weighted connected points in subspaces. This algorithm focuses on finding local subgroups with low variance along one or more attributes. In high-dimensional feature spaces, clusters tend to exist only in specific subspaces, and the algorithm aims to capture these patterns. However, the algorithm had limitations in terms of accuracy.

Khandare and Alvi (2018) investigated the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and K-means clustering algorithms, aiming to enhance the quality of clusters and improve the efficiency of the K-means algorithm. The study was conducted using a standard dataset of online retail stores, consisting of eight attributes and approximately 500,000 instances of online retailing. Initially, the standard K-means algorithm was applied to the retail dataset, followed by clustering aggregation and spectral analysis to examine dataset properties and select preliminary centroids. To improve the algorithm, additional steps were incorporated into the standard K-means algorithm. However, when comparing the proposed algorithm with the standard clustering algorithm, it was found that the proposed algorithm did not significantly improve efficiency due to its focus solely on cluster quality. Consequently, productivity was adversely affected by the algorithm's low efficiency.

2.2.4 Centroid-based Clustering

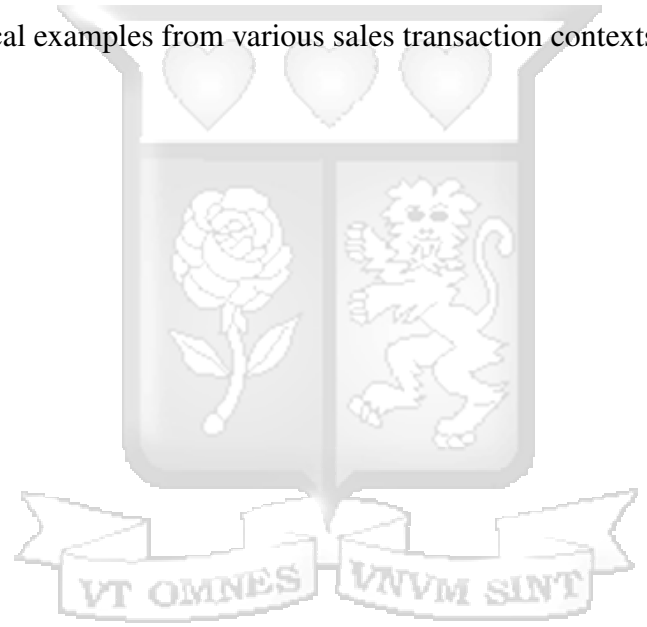
Anitha and Patil (2022) proposed the K-Means clustering method to identify prospective customers and provide timely and pertinent data to businesses in the retail industry. The analysis was conducted by examining the sales history and purchasing behavior of consumers in a systematic manner. The application of the K-Means algorithm involved analyzing real-time transactional and retail datasets to forecast consumer purchasing behavior and identify relevant patterns. However, it should be noted that the method employed a random selection of cluster centers, leading to varying clustering results across different runs of the algorithm.

Sastry et al. (2013) applied centroid based clustering methods to segment sales data into meaningful and valuable clusters. By utilizing K-Means and EM clustering algorithms,

they examined the data to identify variations in product sales within a specific time frame, unveiling interesting patterns that could enhance sales revenue and increase sales volume. The study utilized annual sales data from a metal-based business enterprise to analyze both sales volume and value. The results provided evidence that partitioning techniques such as K-Means and EM algorithms are more suitable for analyzing sales data compared to density-based methods like DBSCAN.

2.3 Our Research

The existing literature demonstrates a growing interest in estimating demand and customer behavior within the retailing context, with some studies employing the EM method. However, to the best of the researcher's knowledge, no prior study has explored the application of the EM-GMM clustering methodology to a liquor retail dataset, incorporating hierarchical steps. Thus, the primary contribution of this research is to utilize Gaussian Mixture Models that combine hierarchical clustering with the Expectation-Maximization algorithm for demand estimation. The initial guess for the EM algorithm is provided by the agglomerative hierarchical clustering. This study offers a comprehensive analysis of the EM-GMM approach, presenting numerical examples from various sales transaction contexts involving SKUs.



Chapter 3

Methodology

3.1 Introduction

In this section, we present the essential background information on the Gaussian Mixture Model (GMM) and its implementation using the Expectation-Maximization (EM) algorithm. The GMM is employed to combine multiple probability distributions and derive parameter estimates. We apply this model to a sales dataset to evaluate its effectiveness and gain valuable insights and recommendations for companies utilizing a distribution sales model.

3.2 Retail Data

The clustering algorithm was employed on a retail dataset comprising 19,000 transactions, covering a period of one year from April 2021 to March 2022. The dataset encompassed 34 features, including information such as sales date, sales volume, sale price, SKU descriptions, customer details, regional data, and salesperson information.

Different locations listed in the data-set gave a full overview of the customer base and purchasing behaviour in Kenya. That is, Nairobi East, Nairobi West, North Rift, South Rift, Lake Region, Coast, Mt. Kenya, and Key Accounts (supermarkets).

To ensure a comprehensive analysis and achieve satisfactory results for both algorithms, the clustering process was conducted in two stages. Firstly, the data pertaining to product analysis was consolidated into 10 categories representing the primary brands of MDEA. Secondly, the data related to customer analysis was categorized into 223 distinct customer segments. This division facilitated the use of the entire dataset as a reference for a relatively straightforward clustering task.

Table 3.1: Definition of key variables.

Variable	Description
InvoiceNo	a unique number generated while issuing an invoice to a customer
InvoiceDate	The date which the invoice was prepared and products provided
CustomerName	Individuals or wholesalers that purchase the products
SKU Code	a unique identifier for various stock keeping units
Qty	The volume of products sold in pieces
Amount	The revenue generated excluding VAT
Region	An area that represents where sales have taken place
Inventory Description	The definition of different SKUs listed together with their capacity
Cases	The volume of products sold in cases
Category	The class or division of products

3.3 Measures of Clusterability

After performing data preprocessing, an assessment of clusterability was conducted to determine if the data exhibited inherent cluster structures. If the data was found to be clusterable, a clustering algorithm was developed and its effectiveness validated by applying clustering quality measures Adolfsson et al. (2019). This section introduces key properties that guided the selection of appropriate clusterability measures for the analysis:

- **Effectiveness:** The chosen clusterability measure needed to accurately identify the presence of clusters in the data, minimizing any ambiguity.
- **Algorithm Independence:** The clusterability measure should not be reliant on a specific clustering algorithm. The objective was to assess whether the data can be clustered using various approaches.
- **Efficiency:** The clusterability measure should be computationally efficient for practical

utility. It should be capable of being computed in low polynomial time and run efficiently on large datasets without being overly restrictive.

In this study, the data exhibited clear characteristics, allowing the researcher to determine effective clusterability measures. Each cluster was described using Gaussian density parameters, including mean and covariance matrix. The evaluation of differences between the data and other clustering datasets provided valuable insights for selecting suitable clusterability notions in future research. To facilitate a more concise comparison, the component Gaussians were constrained to have scalar covariance matrices (Hastie et al. (2009)). This combination of extensive simulations and analysis on real data enriched the understanding of clusterability evaluation and enabled the comparison of different approaches in terms of their effectiveness.

3.4 Research Design

The study employs quantitative approaches to examine and depict the relationships between variables, adopting an experimental research design. According to Degu and Yigzaw (2006), a research design provides guidance to researchers on the collection, analysis, and interpretation of observations. The dataset utilized in this study is sourced from Monument Distillers East Africa Limited, a company specializing in the sale and distribution of premium alcoholic beverages and emerging brands.

The analysis phase involves comprehensive data pre-processing and model optimization procedures. Outliers are identified and removed, ensuring that incomplete or irrelevant data is rectified and aligned with the required format. This prepares the dataset for subsequent model deployments and optimizations, which help reveal information that is independent of the model. Particular attention is given to the quality of the data, specifically focusing on the quality of each attribute. Additionally, less valuable data attributes are considered for elimination. The datasets are downloaded from an ERP system in Microsoft Excel format and imported into R for analysis. The statistical computations and analyses are conducted using R version 4.2.1 program obtained from CRAN, serving as the primary software for the study.

3.5 Expectation Maximization Algorithm

The EM method belongs to a category of algorithms designed for maximum likelihood estimation. This approach allows for the estimation of various parameters in situations where latent variables are present. Within this section, the focus is on maximizing the likelihood function while assuming that all the data, including any hidden or latent variables, originates from an exponential family distribution.

Considering a sample of independent observations denoted as $X = (x_1, x_2, ..., x_n)$, we can infer that the random variables $\mathbf{X}$ and $\mathbf{Z}$ follow the distribution P_θ, where $\theta \in \Theta$ represents an unknown parameter. Typically P_θ is an exponential family, and $\mathbf{X}$ is a vector containing the observed data, while $\mathbf{Z}$ represents the latent variable. The primary objective is to maximize the marginal probability of $\mathbf{X}$. The challenge lies in maximizing $\mathsf{P}_\theta(X) = \sum_{\mathbf{z}} \mathsf{P}_\theta(\mathbf{X}, \mathbf{Z})$, as the summation over $\mathbf{z}$ introduces multiple modes and multiple maxima, making the task difficult. However, the EM algorithm provides a solution to this problem. The EM algorithm iteratively refines the estimation of θ through the following steps:

1. Initialization: Beginning with initial parameter guesses $\theta_0 \in \Theta$.

2. Expectation Step: Computing the expectation $Q(\theta, \theta_t) = E_{\theta_t}(\log \mathsf{P}_\theta(\mathbf{X}, \mathbf{Z}) \,|\, \mathbf{X} = \mathbf{x})$ for $t = 0, 1, 2, \ldots$ until convergence. Here, $Q(\theta, \theta_t)$ represents the conditional expectation of the logarithm of the joint distribution under θ.

3. Maximization Step: Determining the new estimate θ_{t+1} by maximizing the Q function $Q(\theta, \theta_t)$:

$$\theta_{t+1} \in \arg\max_{\theta} Q(\theta, \theta_t).$$

4. Iteration: Iterating between the E-step and M-step until convergence is achieved.

By following this iterative process, the EM algorithm progressively improves the estimate of θ, thereby addressing the challenge of maximizing the marginal probability.

3.6 Gaussian Mixture Model

In this model, each cluster is represented by a Gaussian distribution characterized by its mean and covariance James et al. (2013). We have a variable $\mathbf{Z}$ which is a vector-valued random variable $(e_1, ..., e_m)$, with the standard bases;

$$e_1 = (1,0,0,...,0),$$
$$e_2 = (0,1,0,...,0),$$
$$\dots \text{ and,}$$
$$e_m = (0,0,0,...,1).$$

Therefore, the generative process for this is some vector valued random variable $\mathbf{Z} \in (e_1, ..., e_m)$, and $\mathsf{P}(\mathbf{Z} = e_k) = \propto_k$, where $\propto_k$ is a p.m.f of the finite set $\mathbf{Z} \in (e_1, ..., e_m)$. Given $\mathbf{Z} = e_k$, the distribution of $\mathbf{X}$ is normal with covariance C_k, $\mathbf{X} \sim \mathbb{N}(\mu_k, c_k)$, where $\mu_1, \dots, \mu_m \in \mathbb{R}^d$ (d-dimensional real-valued vectors) and $C_1, \dots, C_m$ are $d \times d$ covariance matrices. This indicates that for every $\mathbf{Z}$, there corresponds a mean and a covariance for one of the Gaussians.

To generate the p.d.f for this model, we first get the marginal distribution of $\mathbf{X}$,

$$\mathsf{P}(X) = \sum_{\mathbf{z}} \mathsf{P}(\mathbf{X}|\mathbf{Z})\mathsf{P}(\mathbf{Z})$$
$$= \sum_{k=1}^{m} \mathsf{P}(\mathbf{Z} = e_k)\mathsf{P}(\mathbf{X}|\mathbf{Z} = e_k),$$

$$\text{where } \mathsf{P}(\mathbf{Z} = e_k) = \propto_k,$$

$$= \sum_{k=1}^{m} \propto_k \mathbb{N}(\mathbf{X}|$$

This equation represents the density function of a mixture of Gaussians, when we have m different gaussian p.d.fs and the coefficient $\propto_k$ is a combination of these p.d.fs. These $\propto_{k_s}$

are the mixing coefficients. A mixture of gaussians is used to model various clusters, as it gives a probabilistic model for a clustering type of problem like the one in this study.
In general, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$, is going to be useful when estimating parameters using the EM algorithm, with each gaussian density having a mean μ_k and covariance C_k. And this is the product of all k_s multiplied by the normal density at x. The $\mathsf{z_k}$ exponents in this product make the factors that do not belong there just once when $\mathsf{z_k} = 0$ and they may get just the normal thing when $\mathsf{z_k} = 1$.

$$\mathsf{P}(\mathbf{X},\mathbf{Z}) = \prod_{k=1}^{m} \propto_k^{\mathsf{z_k}} \frac{1}{\sqrt{2\pi}c_k} \exp\left(-\frac{(\mathsf{X}-\mu_k)^2}{2c_k^2}\right)^{z_k}$$
$$\text{where, } \mathbf{Z} = (z_1, ..., z_m).$$

By applying Bayes' rule, the conditional probability is determined through the following calculation:

$$\mathsf{P}(\mathbf{Z} = e_k|X) = \frac{\mathsf{P}(\mathbf{X}|e_k)\,\mathsf{P}(e_k)}{\mathsf{P}(x)} = \frac{\propto_k \frac{1}{\sqrt{2\pi}c_k} \exp\left(-\frac{(\mathsf{X}-\mu_k)^2}{2c_k^2}\right)}{\sum_{\iota=1}^{m} \propto_\iota \frac{1}{\sqrt{2\pi}c_t} \exp\left(-\frac{(\mathsf{X}-\mu_t)^2}{2c_t^2}\right)}$$

$$\text{and } \mathsf{P}(\mathbf{Z} = e_k|X) = \mathsf{E}(\mathbf{I}(\mathbf{Z} = e_k) \mid X = x).$$

In the next section, the EM algorithm is applied to derive estimations for the parameters of the Gaussian mixture model.

3.6.1 EM for the Gaussian Mixture Model

In this scenario, the parameters are represented by a vector comprising of $\propto$ (p.m.f for $\mathbf{Z}$), the set of μ_s and the set of covariance matrices C.

$$\theta = (\propto, \mu_k, c_k),$$

Starting by formulating EM for exponential families, we have the expectation with respect to θ_0 of the i^{th} sufficient statistic as a function of $(\mathbf{X}, \mathbf{Z})$,

$$\mathsf{E}_{\theta_0}(\mathsf{S}_\mathsf{j}(\mathbf{X}, \mathbf{Z}) \mid X = x) \;=\; \mathsf{E}_\theta \mathsf{S}_\mathsf{j}(\mathbf{X}, \mathbf{Z}) \; for\; each\; \mathsf{j}.$$

We aim to establish if the joint distribution on $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family, to figure out what the sufficient statistics are;

$$\mathsf{P}(\mathbf{X}, \mathbf{Z}) \;=\; \prod_{k=1}^{m} \propto_k^{\mathsf{z_k}} \; \frac{1}{\sqrt{2\pi} c_k} \exp\left(-\frac{(\mathsf{X} - \mu_k)^2}{2c_k^2} \right)^{z_k}$$

The variables are $\mathbf{X}$ and $\mathbf{Z}$, while the parameters are $\propto_k$, μ_k and c_k. As an exponential family, we have;

$$\prod_{k=1}^{m} e^{\mathsf{z_k} \; \log(\propto_k)} \; \frac{1}{\sqrt{|2\pi c_k|}} \; \exp(-\frac{1}{2}(x \; - \mu_k)^T \; C_k^{-1} \; (x - \mu_k)),$$

but we can use the precision matrix which is defined as the inverse of the covariance matrix

$$\Lambda_k \;=\; C_k^{-1}$$

$$\prod_{k=1}^{m} e^{\mathsf{z_k} \; \log \propto_k} \; \sqrt{\frac{|\Lambda_k|}{2\pi}}^{\mathsf{z_k}} \; \exp(-\frac{\mathsf{z_k}}{2}(x \; - \; \mu_k)^T \; \Lambda_k^{-1} \; (x - \mu_k))$$

For proportionality, 2π is dropped and we get;

$$\propto_{\mathsf{x,z}} \; \exp(\sum_{k=1}^{m} \; (\mathbf{Z_k} \; \log \; \propto_k \; + \; \frac{\mathbf{Z_k}}{2} \; \log \; |\Lambda_k| \; - \; \frac{\mathbf{Z_k}}{2}(\mathbf{X}^T \Lambda_k \mathbf{X} \; - \; 2\mu_k^T \Lambda_k \mathbf{X} \; + \; \mu_k^T \Lambda_k \mu_k))).$$

Where $\mathbf{X}^T \Lambda_k \mathbf{X}$ is a scalar and it is equal to its trace.

Simplifying;

$$\exp(\sum_{k=1}^{m} \mathbf{Z_k}\beta_k \; - \; \frac{1}{2}\sum_{k=1}^{m} tr(\mathbf{Z_k}\mathbf{X}\mathbf{X}^T\Lambda_k) \; + \; \sum_{k=1}^{m} \mathbf{Z_k}\mathbf{X}^T\Lambda_k\mu_k)$$

$$where\ \beta_k \; = \; \log \propto_k \; + \frac{1}{2} \; + \; \log\,|\Lambda_k| \; - \; \frac{1}{2}\,\mu_k^T\,\Lambda_k\mu_k \; and \; tr(\mathbf{X}\mathbf{X}^T\Lambda_k) \; = \; tr(\mathbf{X}^T\Lambda_k\mathbf{X})$$

The first set of sufficient statistics for this distribution is; $\mathbf{Z}$, $\mathbf{Z_k}\mathbf{X}\mathbf{X}^T$ $(k = 1,...,m)$, and $\mathbf{Z_k}\mathbf{X}^T$ $(k = 1,...,m)$.

Assuming that $\mathbf{X}$ and $\mathbf{Z}$ are i.i.d from this distribution, we get;

$$\mathsf{P}_\theta(X_1,...,X_n,\, Z_1,...,Z_n) \; = \; \prod_{i=1}^{n} \mathsf{P}_\theta(\mathbf{X_i},\, \mathbf{Z_i})$$

$$= \; \exp(\textstyle\sum_k \,(\sum_i \mathbf{Z}_{ik})\beta_k \; - \; \frac{1}{2}\,\sum_k tr(\,(\sum_i \mathbf{Z}_{ik}\,\mathbf{X}_i\,\mathbf{X}_i^T)\,\Lambda_k) \; + \; \sum_k(\sum_i\,\mathbf{Z}_{ik}\mathbf{X}_i)\,\Lambda_k\mu_k)$$

where $\mathrm{X}_i \; = \; (X_{i1},...,X_{id} \;\; and \;\; \mathbf{Z_i} \; = \; (Z_{i1},...,Z_{im})$.

This is an exponential family over all the random variables and the sufficient statistics in this case are $(\sum_i \mathbf{Z}_{ik})\beta_k$, $(\sum_i \mathbf{Z}_{ik}\,\mathbf{X}_i\,\mathbf{X}_i^T)$, and $(\sum_i\,\mathbf{Z}_{ik}\mathbf{X}_i)$. Using the expression $\mathsf{E}_{\theta_0}(\mathsf{S_j}(\mathbf{X},\mathbf{Z}) \mid X = x) \; = \; \mathsf{E}_\theta \mathsf{S_j}(\mathbf{X},\mathbf{Z})$ *for each* j, we generate these sufficient statistics for the gaussian model where $\mathbf{X} \; = \; (x_1,...,x_n)$ and $\mathbf{Z} \; = \; (z_1,...,z_n)$;

1. First sufficient statistic $(\sum_i \mathbf{Z}_{ik})\beta_k$,

$$\mathsf{E}_{\theta_0}(\sum_i \mathbf{Z}_{ik} \mid X = x) \; = \; \mathsf{E}_\theta(\sum_i \mathbf{Z}_{ik}),$$

$$but,\; \mathsf{P}(\mathbf{Z}_i = e_k) \; = \; \mathsf{P}(\mathbf{Z}_{ik} \; = \; 1) \; = \propto_k .$$

$$\mathsf{E}_\theta(\mathbf{Z}_{ik}) \; = \propto_k,$$

$$hence, \sum_i^n \mathsf{E}_{\theta_0}(\mathbf{Z}_{ik} \mid \mathbf{X}_i) \;=\; \sum_i^n \propto_k = \; n_k$$

$$\therefore \propto_k = \; \tfrac{n_k}{n}.$$

2. Second sufficient statistic $(\sum_i \mathbf{Z}_{ik}\, \mathbf{X}_i\, \mathbf{X}_i^T)$,

$$\mathsf{E}_{\theta_0}\,(\sum_i \mathbf{Z}_{ik}\, \mathbf{X}_i \mid x) \;=\; \mathsf{E}_\theta(\sum_i \mathbf{Z}_{ik}\, \mathbf{X}_i) \;=\; \sum_i \mathsf{E}_\theta(\mathsf{E}_\theta(\mathbf{Z}_{ik}\mathsf{X}_i \mid \mathbf{Z}_{ik})).$$

$$\text{Let } \; r_{ik} = r_{ik}(\theta_0, X_i, k),$$

$$\mathsf{E}_{\theta_0}\,(\sum_i \mathbf{Z}_{ik}\, \mathbf{X}_i \mid x) \;=\; \sum_i \mathsf{E}_{\theta_0}(\mathbf{Z}_{ik} \mid x_i) x_i \;=\; \sum_i r_{ik} x_i,$$

$$\mathsf{E}_\theta(\mathbf{Z}_{ik}\mathbf{X_i} \mid \mathbf{Z}_{ik} \;=\; 1) \;=\; \mathsf{E}_\theta(\mathbf{X_i} | \mathbf{Z}_{ik} \;=\; 1) \;=\; \mu_k,$$

$$\sum_i \mathsf{P}_\theta(\mathbf{Z}_{ik} \;=\; 1)\mu_k \;=\; \sum_i \propto_k \mu_k$$

$$\text{But } \; \propto_k = \; \tfrac{n_k}{n}$$

$$\text{Multiplying both sides by } \mu_k \;,\; n_k \mu_k \;=\; n \propto_k \mu_k$$

$$\therefore \;\; \mu_k \;=\; \tfrac{1}{n_k} \sum_i \; r_{ik} x_i.$$

3. Third sufficient statistic $(\sum_i \; \mathbf{Z}_{ik}\mathbf{X}_i)$,

$$\mathsf{E}_{\theta_0}\,(\sum_i \; \mathbf{Z}_{ik}\, \mathbf{X}_i\, \mathbf{X}_i^T \mid \mathbf{X} = x) \;=\; \mathsf{E}_\theta(\sum_i \; \mathbf{Z}_{ik} \mathbf{X}_i \mathbf{X}_i^T) \;=\; \sum_i \mathsf{E}_\theta(\mathsf{E}_\theta(\mathbf{Z}_{ik}\; \mathbf{X}_i\; \mathbf{X}_i^T \mid \mathbf{Z}_{ik})),$$

$$\mathsf{E}_{\theta_0}\,(\sum_i \; \mathbf{Z}_{ik}\, \mathbf{X}_i\, \mathbf{X}_i^T \mid \mathbf{X} = x) \;=\; \sum_i \; r_{ik} x_i x_i^T,$$

$$\text{Also } \; \mathsf{E}_\theta(\mathbf{Z}_{ik}\mathbf{X}_i\mathbf{X}_i^T \mid \mathbf{Z}_{ik} = 1) \;=\; c_k + \mu_k \mu_k^T.$$

$$\sum_i \mathsf{E}_\theta(\mathsf{E}_\theta(\mathbf{Z}_{ik}\, \mathbf{X}_i\, \mathbf{X}_i^T \mid \mathbf{Z}_{ik})) = \sum_i \mathsf{P}_\theta(\mathbf{Z}_{ik} \;=\; 1)\,(c_k + \mu_k \mu_k^T) = n \propto_k \; (c_k + \mu_k \mu_k^T) = n_k (c_k + \mu_k \mu_k^T),$$

$$\text{where } \; \mathsf{P}_\theta(\mathbf{Z}_{ik} = 1) \;= \propto_k$$

$$\therefore \;\; C_k \;=\; (\tfrac{1}{n_k} \sum_i r_{ik} x_i x_i^T) - \mu_k \mu_k^T.$$

C_k is the covariance matrix and μ_k is the mean.

3.7 Model Validation

Model validation comes once the model development steps are complete. This involves assessing whether the model accurately captures the system's behavior (Aumann (2007)). In this study, effective validation processes were identified, highlighting significant variations among them. These differences provide valuable insights for clustering users to make informed decisions when choosing a clustering technique. To strike a balance between data fitting accuracy and model complexity, the Bayesian Information Criterion (BIC) is utilized.

3.7.1 Bayesian Information Criterion (BIC)

As the models are fitted, the aim is to minimise the loss of information in the process. The BIC shows how much information is lost in the model. A lower BIC value indicates a better-performing model. The BIC incorporates both the likelihood and a penalty term, which penalizes models with a higher number of parameters and a larger sample size to prevent overfitting. By considering these factors, the BIC helps guard against the risk of overfitting and ensures a more reliable model. It is calculated using the following formula:

$$\mathrm{BIC} = -2\ln\bar{L} + k \cdot \ln(n)$$

where ln represents the natural logarithm, $\bar{L}$ is the maximized value of the likelihood function, k is the number of parameters in the model, and n is the sample size.

In this study, it was computed as:

$$-2\ln\bar{L} + 10 * \ln(195)\ for\ products$$
$$\text{and } -2\ln\bar{L} + 223 * \ln(1429)\ for\ customers,$$

3.8 Data Analysis

The initial stage of the data analysis involved preprocessing the retail data to ensure its cleanliness and reliability. Several steps were taken, including the removal of outliers with missing or negative values, as well as duplicate instances. Redundant attributes were also eliminated. Exploratory data analysis techniques were then employed to gain a deeper understanding of the variables. The data was also summarized by using descriptive statistics for two distinct sets of numerical samples i.e, the product data set and the customer data set. The analysis proceeded in two main steps. Firstly, simulated data was used to assess the methodology's ability to identify a known sales demand system and determine the optimal amount of data required for accurate estimates. Secondly, the analysis was conducted on two industry-specific datasets, one focusing on products and the other on customers. Throughout the analysis, a stopping criterion was established based on the difference between consecutive iterations of the EM method's $\bar{X}$ matrices. The implementation of the algorithm utilized the *mclust* package, a widely used R package for model-based clustering, classification, and density estimation employing Gaussian mixture modeling Scrucca et al. (2016). Prior to analyzing the dataset, the *mclust* package was installed and loaded, providing functions for parameter estimation using the EM algorithm for normal mixture models with various covariance structures, as well as simulation functions.

In addition to *mclust*, other libraries such as *ggplot*2, *mixtools*, and *flexmix* were loaded to facilitate the analysis process. *ggplot*2 was used to create visually appealing graphics and map variables to aesthetics. The final model was evaluated, and the findings were documented in **chapter 4**.

3.8.1 Getting Initial Parameters

The assumption of the Gaussian mixture model is that each latent class is characterized by a distinct set of means and covariances. The parameters are intialized and iteratively updated for consistency and accuracy of the results. The **R** package *mclust* employs an EM algorithm to calculate the parameters of the GMM. By using the Bayesian Information Criterion to

select the optimal number of clusters and estimate the corresponding model parameters, it automatically determines the number of clusters. The initialization of means and covariances in *mclust* follows a specific procedure. Initially, the package estimates a one-component Gaussian distribution based on the data. It then performs an EM procedure to estimate the parameters for each subsequent number of components, from two to a specified maximum number. The initial values for means and covariances are obtained from the results of the previous component, by splitting the largest component into two subcomponents. The EM algorithm in *mclust* maximizes the likelihood of the data by iteratively updating the means, covariances, and mixing proportions. The procedure continues until convergence is reached, where the likelihood improvement between iterations falls below a specified threshold. After classifying each data point into a cluster, the mixture, mean, and covariance parameters can be initialized within those groups Melnykov and Maitra (2010).

3.9 Preliminary Data Analysis and Results

This section presents the data generated for the study, along with a comparison of the results obtained using four benchmark procedures. The product data-set consists of 195 observations, while the customers data-set contains 1429 observations. Both data-sets consist of three variables. Scatter plots were used to visualize the initial relationship between two variables, namely cases and revenue, providing insights into the data's behavior. Additionally, the mean, standard deviation, and Gaussian distributions were estimated for the variables.

3.9.1 Scatter Plot of Products and Customers

A plot was created to display two variables as shown in **figure 3.1**, namely cases and revenue. In this plot, the independent variable is represented on the x-axis, while the dependent variable is represented on the y-axis.

The scatter plots show a positive linear correlation between the two variables as observed from the trend of the plotted points, which generally form a linear relationship. **Figure 3.1**

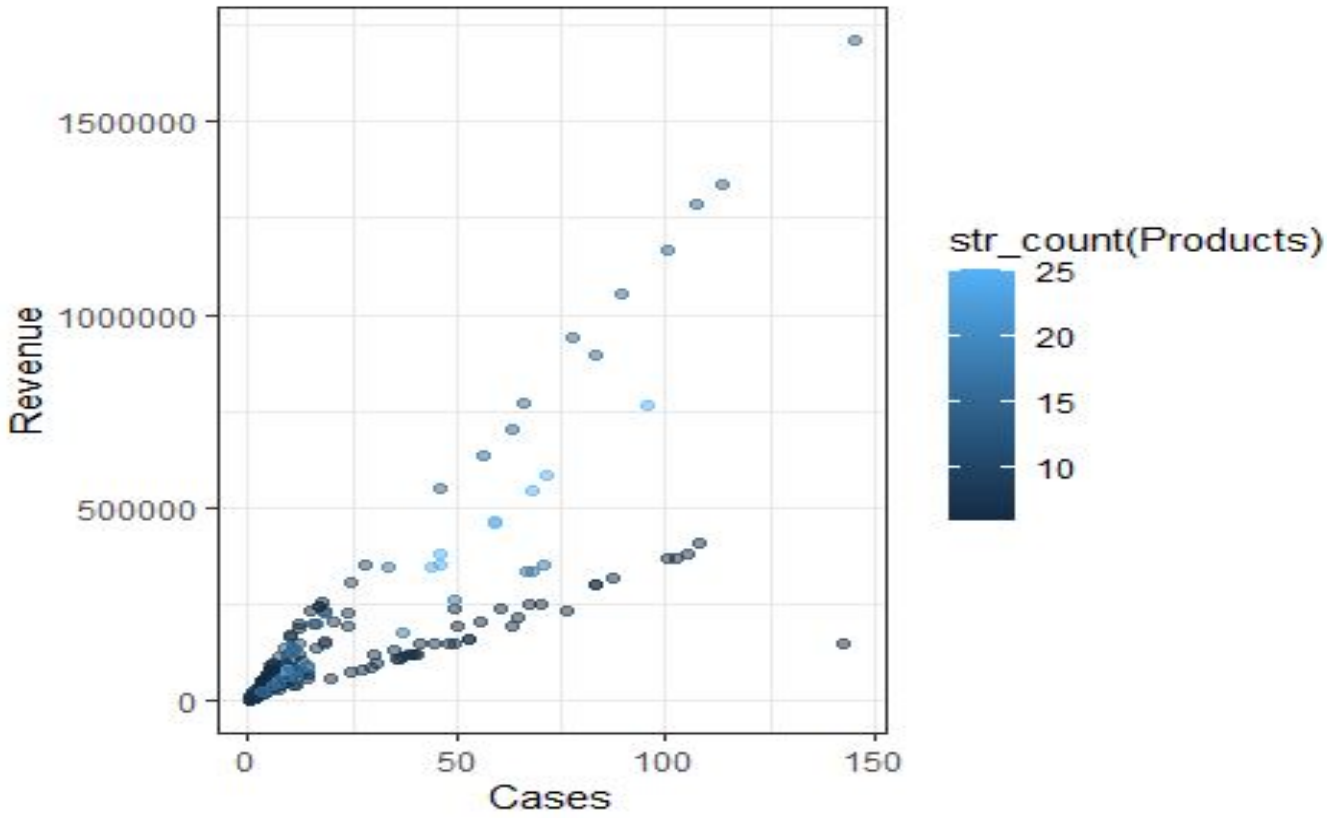


Figure 3.1: Products data-set

illustrates the distribution of the 10 products based on the number of cases sold and the corresponding revenue obtained, analysed within the fiscal year of 2021-2022 in the company. Down in the left there is a high concentration of products with low revenue and cases and up in the right are the long dispersed groups with high volume and revenue. **figure 3.2** shows the distribution of cases sold and revenue obtained for all the customers who purchased within the same fiscal year.

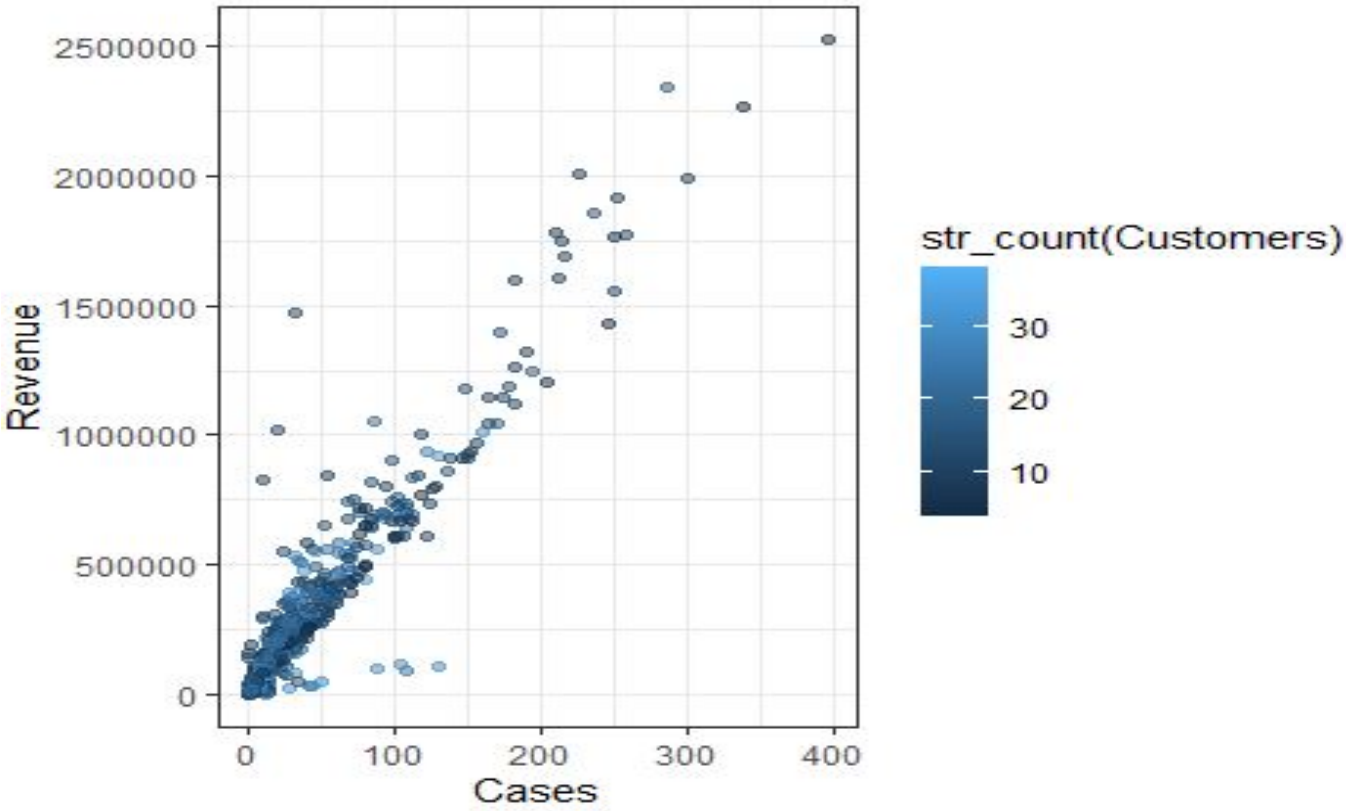


Figure 3.2: Customers data-set

3.9.2 Gaussian Distribution

figure 3.3 shows the estimated mean, standard deviation, and the gaussian distribution plot for the products data-set, considering the cases sold in one fiscal year.

This gave the researcher a good starting point in the study, to understand the relationship of the two variables all at a quick glance.

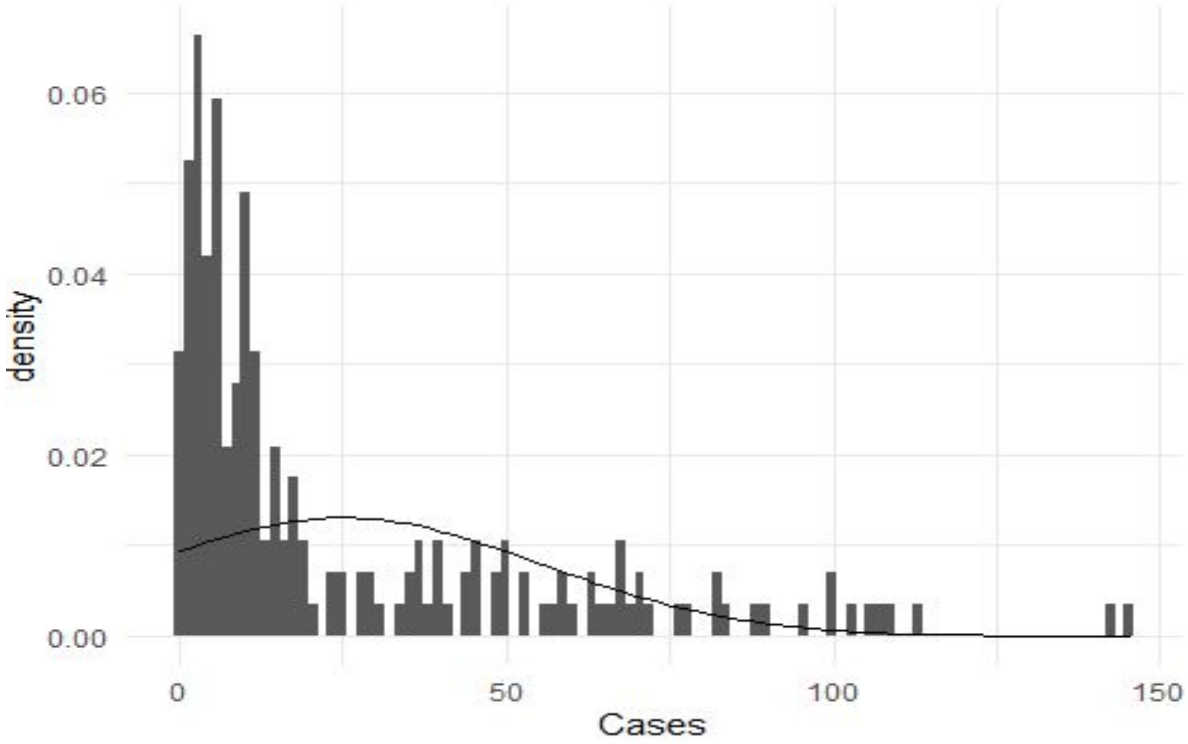


Figure 3.3: Products data-set

Chapter 4

Presentation of Research Findings

4.1 Introduction

The probabilistic model used in this study models the data from a generative process whereby each component corresponded to a distinct cluster. To address various challenges such as determining the number of clusters, initializing parameters, and modeling component densities, a specific form of the GMM was chosen. The Expectation Maximization algorithm was then utilized to estimate the parameters. This iterative algorithm involved an initialization step followed by alternating E-step and M-step computations to address computational circularity. By fitting the available data-sets to the generative model, the parameters were estimated to maximize the likelihood of the observed data. This estimation process also provided probabilities for each data point, indicating the goodness-of-fit to the underlying distribution. High fit probabilities indicated a close match to the distribution, while low fit probabilities suggested anomalies. In this section, the author focused on analyzing the sales data-set to gain insights into retail analytics, sales trends, inventory levels, and consumer demand. These insights were valuable for making informed business decisions, such as maximizing profit margins, minimizing costs associated with overstocking, and identifying areas for improvement within the business processes. Variables such as Invoice number, Invoice dates, customer name, SKU code, Quantity sold, Amount, Region, Inventory description, and Product category were directly mentioned in **Table 3.1**. However other variables such as currency, outlet status, unit of measurement, and unit price were not mentioned in the table. The presented results and findings, covered two main variables in the study (volume of products sold, customers, and revenue) which was a reflection of the contribution of all variables in the raw data. Customer segmentation was analyzed to

determine marketing targets based on the purchasing behavior of customers' while product segmentation was analyzed to determine the volume that needs to be at the warehouse and also to forecast the availability of SKUs that were highly fast-moving.
The assumption made in the GMM model was that the classes exhibit characteristics of a normally distributed density function. This assumption proves effective in estimating class-conditional probabilities.

4.2 Exploratory Data Analysis

The data was summarized across all the numeric variables to provide a description of; measures of central tendency, measures of variability, and the 95 percent confidence interval of the mean.

Table 4.1: Products

Variable	**min**	**max**	**median**	**iqr**	**mean**	**sd**	**se**	**ci**
Cases	0.17	145.5	11	34.9	25.1	30.5	2.2	4.3
Revenue	1250	1711749.6	84568.9	156136.5	172854.9	250230.7	17919.4	35341.8

From the products data-set, there is a maximum of 145.5 and minimum of 0.17 cases sold in the entire period. The quantity sold had a mean and standard deviation of 25.1 and 30.5 respectively with a mean revenue of 172,854.9. The 95% confidence interval of the mean is 4.3 and 35,341.8 for cases and revenue respectively.

Table 4.2: Customers

Variable	**min**	**max**	**median**	**iqr**	**mean**	**sd**	**se**	**ci**
Cases	0.08	396.2	8	18.3	21.1	38.5	1	1.9
Revenue	0	2528710.8	60689.7	143067.9	160955.4	276780.7	7321.8	14362.7

From the customers data-set, there is a maximum of 396.17 cases and minimum of 0.08

purchased in the entire period. Cases sold have a mean and standard deviation of 21.106 and 38.459 respectively with a mean revenue of 160,955.358. The 95% confidence interval of the mean is 1.996 and 14,362.7 for cases and revenue respectively.

4.3 Results

The results highlighted in this section were set into three categories. The first section gave the output of product data analysis, the second gave the output of customers data analysis, and the third section went ahead to fit the model and give the visual output of the analysis. All the procedures and analyses were carried out using **R** functions, as explained in **section 3.9**. The obtained results aligned with the objectives outlined in **chapter 1**.

4.3.1 Products Data

Table 4.3 presents the comparative analysis of the mixture model formed on the probabilities of belonging to each cluster. Based on the performance, there is a high probability of products being in the first cluster. Additionally, only Grace Du Roi, Cape Velvet Cream, Kensington London Dry Gin, LSD Cap, and Squadron Dark Rum could be traced in the first three clusters. This could be due to the consistency of sales among the highlighted SKUs compared to the rest.

Table 4.3: Probabilities.

Product	prob-cluster1	prob-cluster2	prob-cluster3
Grace Du Roi	0.9793814	0.02061856	0.04123711
Squadron Dark Rum	0.9793811	0.02015756	0.04236711
LSD Cap	0.9893814	0.02216846	0.04330712
Kensington London Dry Gin	0.9693814	0.02071656	0.04723711
Cape Velvet Cream	0.9536082	0.04639175	0.09278351

Hence, the probabilities facilitated the estimation of the means and the proportions of the clusters.

Table 4.4: Estimation of Means

mean-cluster1	**mean-cluster2**	**mean-cluster3**
10.36788	10.61171	10.61142

Table 4.5: Estimation of Proportions

props-cluster1	**props-cluster2**	**props-cluster3**
0.4896907	0.5103093	0.0110121

VT OMNES VNVM SINT

Considering the bi-variate nature of the data-set, the means were estimated in two dimensions, the standard deviation and the covariance between the two variables (cases and revenue), as well as the proportions. The researcher started by first initializing two clusters. The ellipse curve in **Figure 4.1** below showed how the *x* and *y* data points corresponded to the gaussian mixture model.

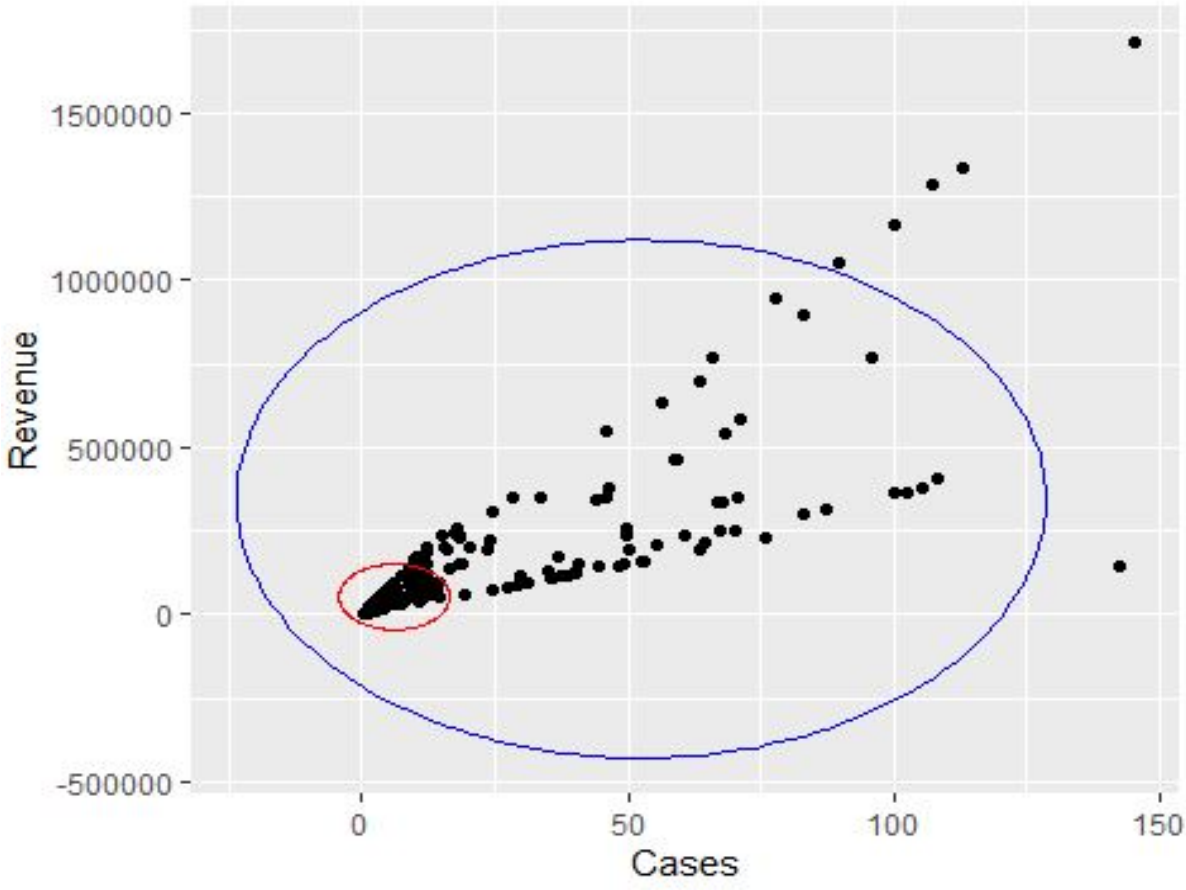


Figure 4.1: Distribution of product data-points 1

The data can be observed to exhibit a tendency to cluster into two distinct groups. The overall pattern of the data reveals variations not only in the means but also in the covariance. The blue cluster and the red cluster exhibit distinct differences in their respective means and covariance. The blue cluster has a lot of spread in both axes whereas the red cluster has a less spread. Most cases sold are concentrated between 0 - 50. An equivalent distribution is observed for revenue generated which is concentrated between 0 - 500,000. During the

E-step, the points outside the clusters are more probable under the blue cluster than the red. The points inside the clusters in the middle are equally explained to be in either clusters. There are few points exceeding the boundary of the blue cluster and this could be linked to products that brought a higher revenue to the company due to better marketing strategies, stock availability and pricing.

Upon updating the model parameters to calculate the weighted mean and covariances, it was observed that the red cluster model undergoes a slight shift to better align with the data located near its boundary. Similarly, the blue cluster model adjusts to better accommodate the data with a higher probability under its component.

After iterating the process, we see that the clusters shift and tilt slightly as in **Figure 4.2**.

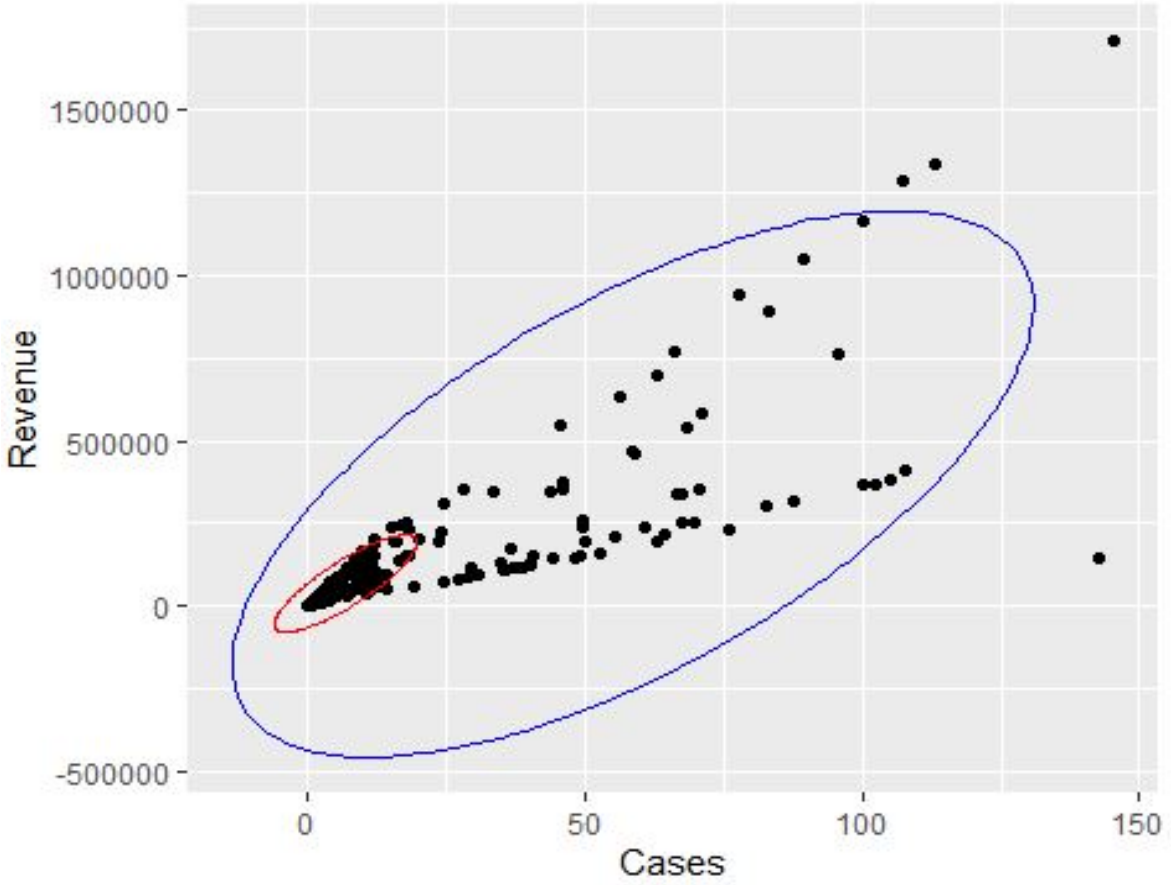


Figure 4.2: Distribution of product data-points 2

By several iterations, the red and blue clusters adjusted further to be entirely responsible for

the dispersed data near their boundaries. Eventually, this model converges and then it ceases to change much.

4.3.2 Customer Data

The same procedure was applied to the customers data-set. Two clusters are initialized and the means, standard deviation, and covariance between the two variables (cases and revenue) estimated.

The ellipse curve in **Figure 4.3** below showed how the data points corresponded to the GMM.

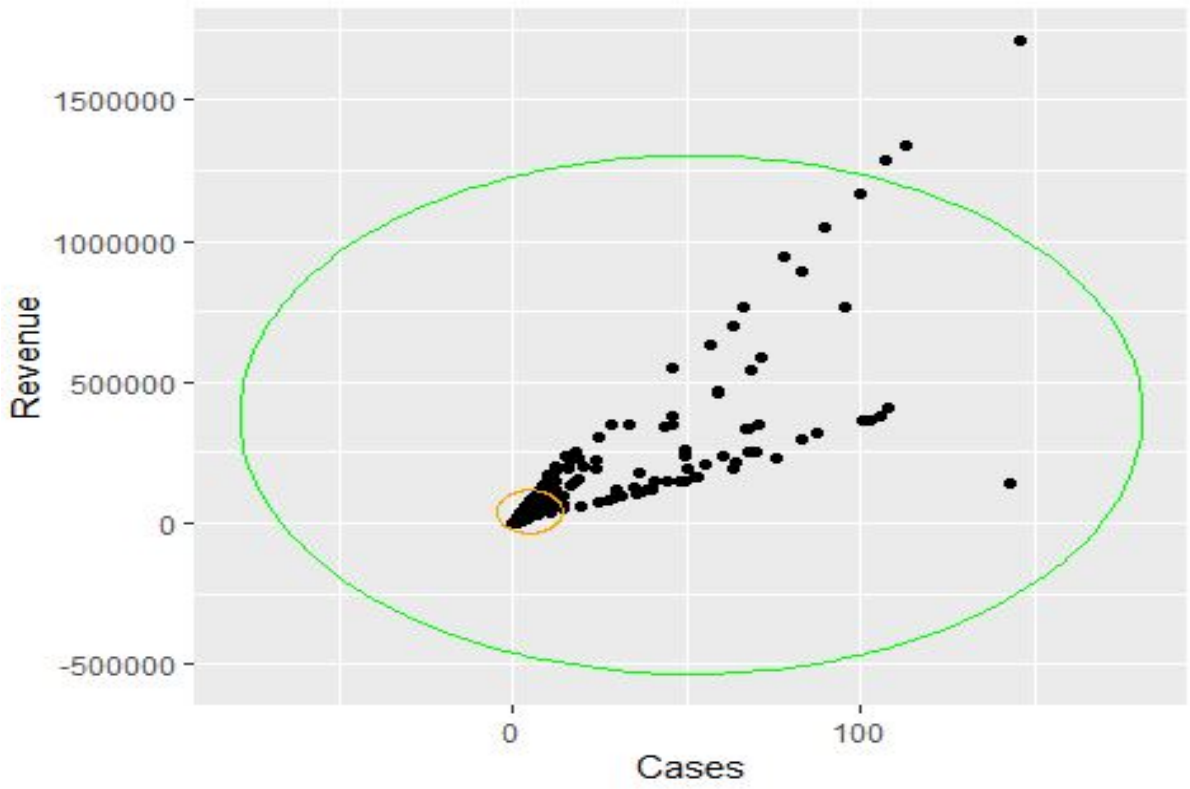


Figure 4.3: Distribution of customer data-points 1

The green and orange clusters exhibit distinct means and covariances. The green cluster displays a wide spread in both axes, while the orange cluster has a less spread. It is observed during the E-step that the points outside the clusters have a higher probability of belonging

to either the green or orange cluster. The points inside the clusters in the mid-point can either fall in the green cluster or the orange cluster. The few points that exceed the boundary of the green cluster could be associated with customers who were exceptional in purchase of many cases hence generating a higher revenue.

Figure 4.4 shows the clusters shifting and adjusting further trying to accommodate the dispersed data near the boundaries. This is after several iterations.

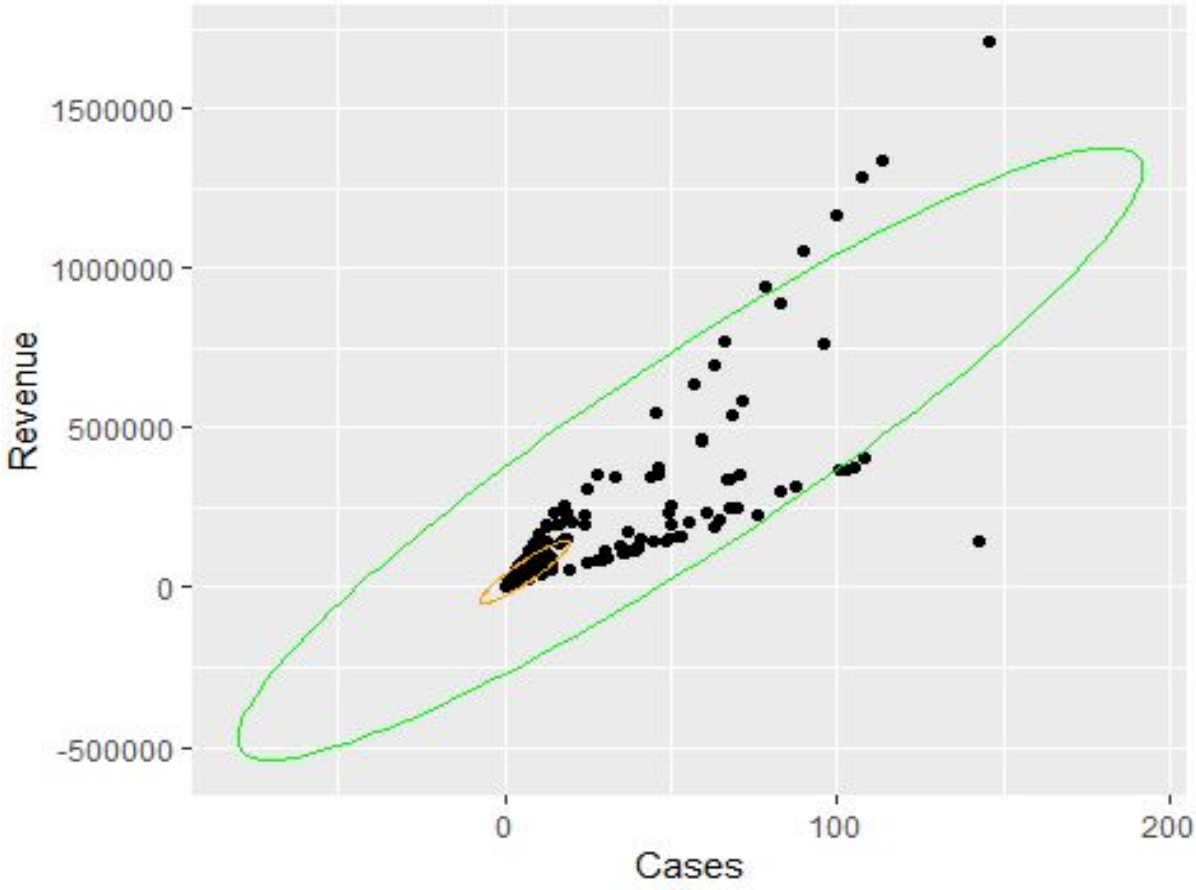


Figure 4.4: Distribution of customer data-points 2

The researcher stopped after the model converged and there was less change.

4.3.3 Model Performance

The models were evaluated using Bayesian Information Criterion. The bigger the number of clusters, the better the model became. It was determined that the data could be divided into

two distinct clusters, each with its own characteristics. During the calculation of the BIC, it was observed that as the number of iterations increased, the log-likelihood was maximized.

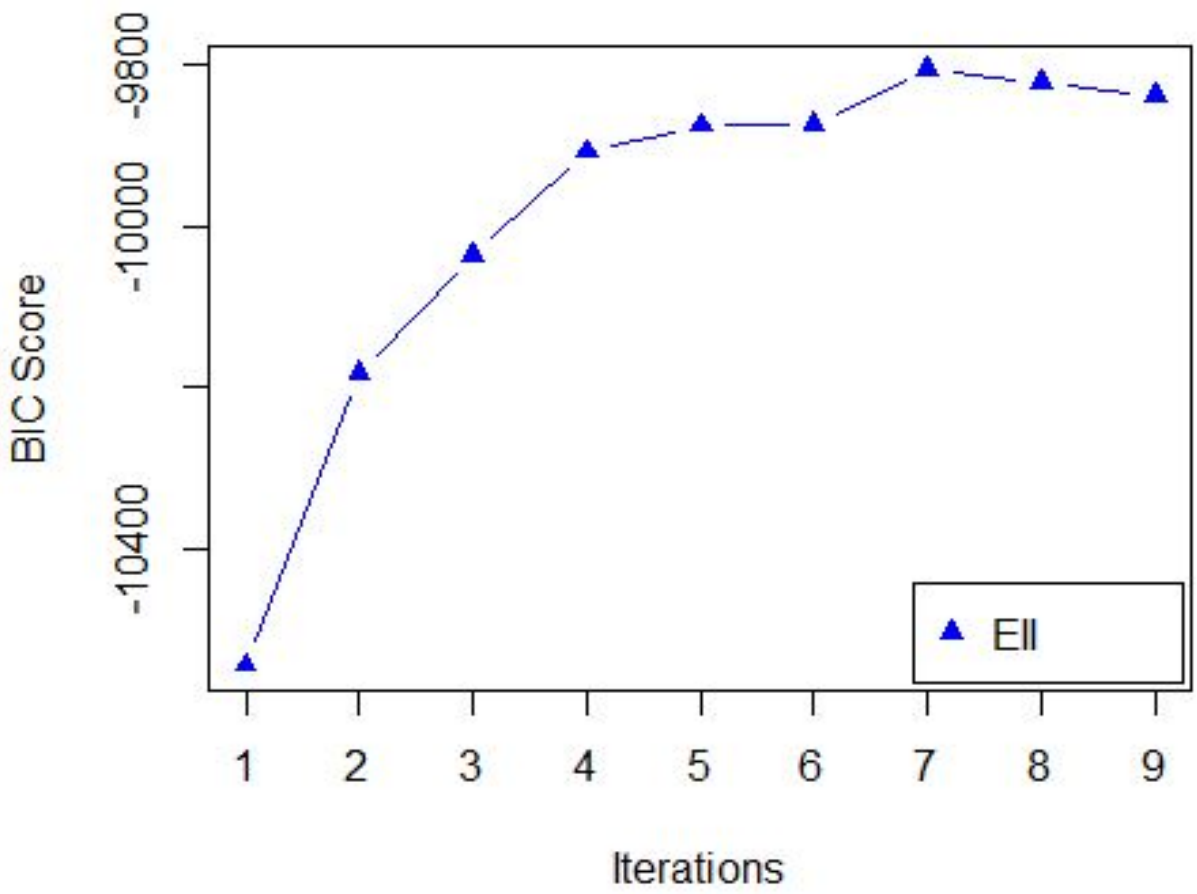


Figure 4.5: Calculating BIC

The curve is fairly smooth and monotone. As the iterations progress, the EM algorithm reaches a point of diminishing returns, where the change between each subsequent iteration becomes negligible. It is important to note that all the BIC scores were negative. Additionally, starting from the sixth iteration, the gradient of the curve remains nearly constant, indicating that the BIC score reaches a plateau and does not significantly increase beyond that point. This observation suggests that the model has converged, and further increasing the number of iterations does not yield substantial improvements.

Chapter 5

Discussion of Research Findings

5.1 Introduction

This chapter provides a comprehensive overview of the development and validation components of the model, and also the study outcomes in relation to the research objectives. This chapter concludes by describing the application of the model in identifying products and customers that would improve sales performance. The development of the model was carried out using **R**, an open-source programming language widely recognized for its data analysis capabilities and its statistical software. The key packages utilized for analysis and data manipulation were *mclust* and *mixtools*, which offer essential functionalities for clustering analysis and data manipulation.

Several quantitative criteria were summarized from the dataset to select a suitable clusterability measure. The findings indicated that GMM method using EM is effective in classifying datasets based on their level of clusterability. Different methods exhibited varying performance based on factors such as dimensionality, outlier treatment, and cluster shape and separability (Adolfsson et al. (2019)). Within the EM algorithm, the estimation step aimed to determine a latent variable value for each data point, while the maximization step focused on optimizing the parameters of the probability distributions. This iterative process aimed to accurately capture the density of the data. The repetitions continued until an optimal set of latent values and a maximum likelihood estimation were attained, effectively fitting the data.

5.2 Our case study

The proposed approach was applied to a retail dataset from MDEA company, which operates in multiple locations in Kenya. The dataset consisted of diverse products distributed to customers. The goal of this study was to uncover potential relationships between customer groups and their purchasing patterns. Customers with similar preferences were grouped together in clusters, and individual customer attributes were dispersed within these larger clusters. Each cluster exhibited distinct characteristics, including a set of base values and associated variation ranges, which could be used to suggest standard settings for new sales regions.

The results obtained from clustering both products and customers revealed that transactions involving similar customer preferences were grouped together, as were products with similar characteristics and descriptions. Consequently, if a new potential customer emerges, they can be assigned to the corresponding customer group based on the mean value of the clusters obtained.
Similarly, commercial manager or investor can utilize these findings to achieve the following objectives for their business enterprise;

- **Customer Segmentation:** From the findings, distinct customer groups have been identified based on their preferences and purchasing behavior. By understanding these customer segments, a manager can tailor marketing strategies and promotions to each group's specific needs and preferences. This can include targeted advertising campaigns, personalized offers, and product recommendations, leading to improved customer satisfaction and increased sales.

- **Stockout Prevention:** By knowing the customer preferences and purchasing behavior, a manager can anticipate demand for specific products and take proactive measures to prevent stockouts. They can use the results to set appropriate stock levels for different products and customer segments. Additionally, they can monitor inventory levels and implement replenishment strategies to ensure timely availability of popular products, reducing the likelihood of stockouts and potential revenue loss.

- **Demand Forecasting:** By analyzing the clusters of products and customers, a business can gain insights into the demand patterns and potential sales opportunities. They can identify which products are popular among specific customer groups and adjust inventory levels accordingly to minimize stockouts or overstocking. Therefore this research can help optimize inventory management, reduce costs, and ensure a more efficient supply chain.

- **Marketing Campaign Optimization:** The results can provide information in designing effective marketing campaigns that resonate with each group. By tailoring marketing messages, channels, and offers to specific customer segments, a manager can optimize the marketing budget, increase campaign effectiveness, and drive higher conversion rates.

- **Expansion and New Market Entry:** The identified customer clusters and associated product preferences can also guide decisions when expanding into new sales regions or targeting new customer segments. An investor can use the cluster profiles as a reference to identify potential similarities with existing customer groups. This information can assist in setting standard settings and product assortments for new markets, reducing the risk of offering irrelevant products and optimizing the chances of success.

By leveraging the results from the study, a business can enhance operational efficiency, improve customer satisfaction, and drive sales growth for the company.

5.3 Model Performance

The BIC criterion provided an assessment of the effectiveness of the GMM in predicting the data. A lower BIC value indicated a better model fit and a more accurate representation of the underlying, unknown distribution. To prevent overfitting, the BIC penalized models with a larger number of clusters. In this study, the BIC was employed to evaluate the efficiency of the parameterized model. The model with the lowest BIC was selected, indicating an optimal number of clusters (in this case, **2**). Furthermore, the fitted model was utilized to estimate the latent parameters for both existing and new data points, offering valuable insights for analysis.

5.4 Limitations of the Study

The study encountered several limitations that should be acknowledged. Firstly, it did not account for the potential sales in a new region, thus neglecting the analysis of predicting sales in entirely new markets. Consequently, the methodologies employed may not be practical when applied to new regions. Secondly, the analysis focused only on volume and revenue, overlooking the influence of other variables such as tax rates, invoice dates, outlet status, and units of measurement on overall sales. Moreover, data consistency was an issue for the first 5 months of the entire period, which resulted in the exclusion of customers who made single or minimal purchases during that time. The researcher made efforts to streamline the dataset while ensuring it still aligned with the study objectives.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

The findings of this study demonstrate the utility and adaptability of Gaussian Mixture Models in analyzing product demand and customer purchasing behavior. The application of GMMs, along with the EM algorithm, enabled the clustering of sales data and estimation of model parameters. This iterative process improved the likelihood of the model fitting the data and ensured convergence. Additionally, the use of the BIC facilitated model selection by penalizing more complex models and assessing their predictive accuracy. The outcomes of the study revealed valuable insights, particularly in identifying high-value customers and products. Notably, the green cluster exhibited a concentration of high-revenue customers, while the blue cluster consisted of high-revenue products.

6.2 Recommendation

The suggested methodology offers a practical solution for businesses dealing with large datasets, specifically for conducting accurate sales predictions. The machine learning approach described in this research paper presents an efficient framework for data preprocessing and decision-making processes. By implementing this approach, organizations can effectively handle substantial data volumes and benefit from intelligent sales forecasting.

References

Adolfsson, A., Ackerman, M., and Brownstein, N. C. (2019). To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26.

An, D., Zheng, X., Rong, C., Kechadi, T., and Chen, C. (2015). Gaussian mixture model based interest prediction in social networks. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*, pages 196–201. IEEE.

Andersson, S.-E. (1998). Passenger choice analysis for seat capacity control: A pilot project in scandinavian airlines. *International Transactions in Operational Research*, 5(6):471–486.

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.

Anupindi, R., Dada, M., and Gupta, S. (1998). Estimation of consumer demand with stock-out based substitution: An application to vending machine products. *Marketing Science*, 17(4):406–423.

Argüelles, M., Benavides, C., and Fernández, I. (2014). A new approach to the identification of regional clusters: hierarchical clustering on principal components. *Applied Economics*, 46(21):2511–2519.

Aumann, C. A. (2007). A methodology for developing simulation models of complex systems. *Ecological Modelling*, 202(3-4):385–396.

Blattberg, R. C. and Neslin, S. A. (1993). Sales promotion models. *Handbooks in operations research and management science*, 5:553–609.

Bohm, C., Railing, K., Kriegel, H.-P., and Kroger, P. (2004). Density connected clustering with local subspace preferences. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 27–34. IEEE.

Brownlee, J. (2019). A gentle introduction to expectation-maximization (em algorithm). *Machine Learning Mastery, Oct*, 31.

Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*, pages 160–172. Springer.

Conlon, C. T. and Mortimer, J. H. (2013). Demand estimation under incomplete product availability. *American Economic Journal: Microeconomics*, 5(4):1–30.

Degu, G. and Yigzaw, T. (2006). Research methodology: lecture notes for health science students. *Addis Ababa: The Carter Center (Ethiopian Public Health Training Initiative)*, pages 45–50.

HanumanthSastry, S. and PrasadaBabu, M. (2013). Analysis & prediction of sales data in sap-erp system using clustering algorithms. *arXiv e-prints*, pages arXiv–1312.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Ja, S., Rao, B., and Chandler, S. (2001). Passenger recapture estimation in airline rm. In *AGIFORS 41st Annual Symposium*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jung, H. and Jeong, S.-J. (2012). Managing demand uncertainty through fuzzy inference in supply chain planning. *International Journal of Production Research*, 50(19):5415–5429.

Jung, J.-Y., Bae, J., and Liu, L. (2009). Hierarchical clustering of business process models. *International Journal of Innovative Computing, Information and Control*, 5(12):1349–4198.

Kalsoom, R. and Halim, Z. (2013). Clustering the driving features based on data streams. In *Inmic*, pages 89–94. IEEE.

Kettani, O., Ramdani, F., and Tadili, B. (2014). An agglomerative clustering method for large data sets. *International Journal of Computer Applications*, 92(14).

Khandare, A. and Alvi, A. (2018). Efficient clustering algorithm with enhanced cohesive quality clusters. *International Journal of Intelligent Systems and Applications*, 11(7):48.

Laura Wood, S. P. M. (2019). Understanding alcoholic drinks market dynamics in kenya, 2019 - eabl controls about 90% of the kenyan beer market and continues to expand into the rest of east africa. *ResearchAndMarkets.com*.

Ledolter, J. (2013). *Data mining and business analytics with R*. John Wiley & Sons.

Maingi, M. N. (2015). A survey on the clustering algorithms in sales data mining.

Melnykov, V. and Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series).

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., and Alm, E. J. (2013). Distribution-based clustering: using ecology to refine the operational taxonomic unit. *Applied and environmental microbiology*, 79(21):6593–6603.

Rahim, S. A., Manson, G., and Aziz, M. (2021). Data clustering based on gaussian mixture model and expectation-maximization algorithm for data-driven structural health monitoring system. *International Journal of Integrated Engineering*, 13(7):167–175.

Ratliff, R. M., Venkateshwara Rao, B., Narayan, C. P., and Yellepeddi, K. (2008). A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings. *Journal of Revenue and Pricing Management*, 7(2):153–171.

Sastry, H. S. and Babu, M. P. (2013). Cluster analysis of material stock data of enterprises. *International Journal of Computer Information Systems*, 6(6):8–19.

Sastry, S. H., Babu, P., and Prasada, M. (2013). Analysis & prediction of sales data in sap-erp system using clustering algorithms. *arXiv preprint arXiv:1312.2678*.

Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data. *Technometrics*, 21(4):417–432.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289.

Seif, G. (2018). The 5 clustering algorithms data scientists need to know. *Towards Data Science*.

Steiner, T., Verborgh, R., Gabarro, J., Mannens, E., and Van de Walle, R. (2015). Clustering media items stemming from multiple social networks. *The Computer Journal*, 58(9):1861–1875.

Talluri, K. and Van Ryzin, G. (2004). Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33.

Vulcano, G., Van Ryzin, G., and Chaar, W. (2010). Om practice—choice-based revenue management: An empirical study of estimation and optimization. *Manufacturing & Service Operations Management*, 12(3):371–392.

Vulcano, G., Van Ryzin, G., and Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research*, 60(2):313–334.

Yang, N., Wang, G., Hao, J., Yan, Y., and Han, H. (2017). A gmm-based user model for knowledge recommendation. In *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, pages 1–5. IEEE.

Yao, L., Li, Z., Nie, T., and Zhao, Z. (2018). Research and implementation of a brand normalization method across e-commerce platforms. In *2018 4th International Conference on Big Data Computing and Communications (BIGCOM)*, pages 236–242. IEEE.

Zhang, X. (2004). The impact of forecasting methods on the bullwhip effect. *International journal of production economics*, 88(1):15–27.

Appendix A

The R codes used for model fitting and estimation of parameters in Chapter 3 and Chapter 4 are shown below;

A.1 Libraries and data summary

```
library(mclust)
library(mixtools)
library(flexmix)
library(ellipse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(rstatix)
library(DataExplorer)
library(reshape)
theme_set(theme_bw())

##Summarizing the data across all numeric variables
setwd("C:\\Users\\JohnOloo\\Desktop\\Project")
Product <- read.csv("Products.csv", header = T)
Products_data <- Product %>%
  drop_na()
Customer <- read.csv("Customers.csv", header = T)
Customers_Data <- Customer %>%
  drop_na()
get_summary_stats(Customers_Data)
```

```
# Scatter plot - Products
Products_data %>%
  ggplot(aes(x = Cases, y = Revenue, col = str_count(Products)))+
  geom_point(alpha = 0.5)

# Scatter plot - Customers
Customers_Data %>%
  ggplot(aes(x = Cases, y = Revenue, col = str_count(Customers)))+
  geom_point(alpha = 0.5)
```

A.2 Calculating Expectatation and model fitting

```
# Calculating PDF with class means and covariances.
q <- cbind(mvpdf(x = Products_data[, 1:2], mu = mu[1, ], sigma = cov[[1]]),
mvpdf(x = Products_data[, 1:2], mu = mu[2, ], sigma = cov[[2]]),
mvpdf(x = Products_data[, 1:2], mu = mu[3, ], sigma = cov[[3]]))

# Expectation Step for each class.
r <- cbind((a[1] * q[, 1])/rowSums(t((t(q) * a))), (a[2] * q[, 2])
/rowSums(t((t(q)* a))), (a[3] * q[, 3])/rowSums(t((t(q) * a))))

# Choosing the highest row-wise probability
eK <- factor(apply(r, 1, which.max))

# Total Responsibility
mc <- colSums(r)

# Updating Mixing Components.
a <- mc/NROW(Products_data)
```

```
# Updating the Means
mu <- rbind(colSums(Products_data[, 1:2] * r[, 1]) * 1/mc[1],
colSums(Products_data[, 1:2] * r[, 2]) * 1/mc[2], colSums(Products_data[, 1:2]
* r[, 3]) * 1/mc[3])

# Updating Covariance matrix.
cov[[1]] <- t(r[, 1] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[1, ]))) %*% (r[, 1] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[1, ]))) * 1/mc[1]
cov[[2]] <- t(r[, 2] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[2, ]))) %*% (r[, 2] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[2, ]))) * 1/mc[2]
cov[[3]] <- t(r[, 3] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[3, ]))) %*% (r[, 3] * t(apply(Products_data[, 1:2], 1,
function(x) x - mu[3, ]))) * 1/mc[3]

## Calculating BIC
```{r eval=TRUE, echo=TRUE, warning=FALSE}
# Computing the sum of the mixture densities by taking the log and adding the
column vector.
loglik <- sum(log(apply(t(t(q) * a), 1, sum)))
bic <- -2 * loglik + 14 * log(NROW(Products_data)) # BIC equation
# After every iteration we can plot.
par(mfrow = c(1, 2))
plot(mclustBIC(Products_data[, -3]), modelNames = "EII", type = "l", lwd = 2,
col = "blue", main = "Log-Likelihood", xlab = "Iterations", ylab = "BIC Score")
```
```

Appendix B

SU Ethical Clearance Release Letter

Strathmore UNIVERSITY

13th October 2022

John Oloo
Student number: 137224
john.oloo@strathmore.edu

Dear John,

RE: Examining Gaussion Mixture Models using Clustering Algorithms

This is to inform you that the Office of Graduate Studies on 12th October 2022 received your acknowledgement of breach in ethical processes given that you have already collected data and written the Thesis prior to obtaining Ethical clearance. The ethics approval process is ONLY done before any collection of primary or secondary data.

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: This is not in any way an ethical approval letter.

Yours sincerely,

Dr. Bernard Shibwabo
Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C

Originality Report

JohnOloo Dissertation

ORIGINALITY REPORT

19%	17%	8%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.r-bloggers.com Internet Source	2%
2	su-plus.strathmore.edu Internet Source	2%
3	zdoc.site Internet Source	1%
4	machinelearningmastery.com Internet Source	1%
5	Johannes Ledolter. "Clustering", Wiley, 2013 Publication	1%
6	rs.itum.mrt.ac.lk Internet Source	1%
7	hdl.handle.net Internet Source	1%