SEGMENTATION OF CUSTOMER PROFILING VARIABLES FOR MARKET ANALYSIS USING EXPLORATORY DATA ANALYSIS AND K-MEANS CLUSTERING

BY

OKHAMENA AZEEZ KEHINDE

ENG1703969

DEPARTMENT OF COMPUTER ENGINEERING, UNIVERSITY OF BENIN

FACULTY OF ENGINEERING

UNIVERSITY OF BENIN

BENIN CITY

SEPTEMBER, 2023

SEGMENTATION OF CUSTOMER PROFILING VARIABLES FOR MARKET ANALYSIS USING EXPLORATORY DATA ANALYSIS AND K-MEANS CLUSTERING

BY

OKHAMENA AZEEZ KEHINDE
ENG1703969

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING, FACULTY OF ENGINEERING, UNIVERSITY OF BENIN, BENIN CITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF BACHELOR OF ENGINEERING(B.ENG) DEGREE IN COMPUTER ENGINEERING

SEPTEMBER, 2023

CERTIFICATION

This project was carried out by Okhamena Azeez Kehinde in the Department of Computer Engineering, Faculty of Engineering, University of Benin, Benin City, and is hereby certified

…….. ……..

Engr. Dr. U. Iruansi Date

(Project Supervisor)

……. ……..

Engr. Dr. (Mrs.) O. Okosun Date

(Ag. Head of Department)

DEDICATION

I humbly dedicate this work to God Almighty, acknowledging His immeasurable blessings and guidance that have sustained me throughout the process of this research. Firstly, I am profoundly grateful for the precious gift of life bestowed upon me. Through His grace and mercy, I have been allowed to embark on this journey and witness its completion.

I extend my gratitude to God for blessing me with a sound mind, allowing me to think critically, analyze, and comprehend the complexities of this research. His divine wisdom has illuminated my path and provided clarity amidst the challenges encountered along the way.

In addition, I thank God for granting me peace amidst the demanding nature of this work. His calming presence has alleviated anxiety and provided me with the strength to persevere during moments of uncertainty. His peace has been my constant anchor, keeping me grounded and focused on the task at hand.

Lastly, I sincerely appreciate His unwavering guidance throughout this research journey. His divine direction has steered me towards the right path, guiding my choices, and enabling me to make meaningful contributions. I recognize that my accomplishments are a testament to His loving guidance and presence in my life.

May this work be a testament to His goodness and a source of inspiration to others?

ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt gratitude to the almighty for granting me the strength and wisdom to undertake this project.

I extend my sincere appreciation to my dedicated supervisor, Engr. Dr. Usiholo Iruansi, for their unwavering support, guidance, and invaluable insights throughout this journey. Your mentorship has been instrumental in my success. I would also like to acknowledge the invaluable contributions of the entire staff members, whose collective efforts and expertise have impacted me, I say a big thank you to you all.

To my beloved parents, Mr. And Mrs. Okhamena, for your unwavering love, encouragement, and sacrifices have been my driving force. I am profoundly grateful for your endless support. a special thanks to my siblings for standing by me, offering their encouragement, and being a source of inspiration. I extend my heartfelt gratitude to my dear friends, Aziken Jefferson, Ismalia Jesse, Edoma Benjamin Oghosa, Eric Samuel, Adeyemo Ayoade, Akpomena Nero, and Imoru David, for their friendship, encouragement, and support. Your unwavering belief in me has been a constant source of strength

Thank you all for being an integral part of my journey, and for your unwavering support and encouragement.

ABSTRACT

Market analysis has evolved significantly over the years, with customer profiling and segmentation playing a pivotal role in understanding and catering to diverse consumer needs. One powerful approach to customer segmentation involves the combined use of Exploratory Data Analysis (EDA) and K-means clustering. EDA, encompassing univariate, bivariate, and multivariate analyses, allows for a comprehensive examination of customer data, while K-means clustering assists in identifying distinct customer segments based on similarities in various profiling variables.

This study employs Exploratory Data Analysis, encompassing univariate, bivariate, and multivariate analyses, to gain profound insights into customer demographics. The initial analysis revealed that a substantial portion of the customer base falls within the age range of 41 to 60 years, possesses first-degree qualifications, and is predominantly married, constituting approximately 65% of the sample. Additionally, income distribution exhibited a diverse pattern, with the majority earning between 0 to 100k$, but a noteworthy proportion having incomes exceeding 600k$. The bivariate analysis further unveiled intriguing insights, particularly in terms of spending patterns linked to educational backgrounds.

Employing K-means clustering on the customer profiling variables, this study successfully identified three distinctive customer clusters. These clusters were characterized as follows: the first cluster comprised low earners with corresponding low spending tendencies, the second cluster consisted of moderate earners exhibiting moderate spending habits, and the third cluster encompassed high earners known for their high spending behaviors. The integration of EDA and K-means clustering in this analysis provides valuable information for targeted marketing and sales strategies. By recognizing these distinct customer segments, businesses can tailor their approaches to cater to the specific needs and preferences of each group, thus enhancing their market competitiveness and overall success.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF PLATES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND STUDY

Market segmentation can be traced back to the early 1950s when Wendell R. Smith introduced the idea of dividing markets into homogeneous subsets. However, not until the 1970s did advancements in computing power and data analytics make segmentation more practical and widespread. Traditional methods, such as demographic and geographic segmentation, laid the foundation for more sophisticated approaches that considered customer behavior, psychographics, and purchase patterns (Michael John et al.,2022).

Exploratory data analysis is an essential preliminary step in data analysis and research. Its primary purpose is to gain a deep understanding of a dataset's properties and structure. During Exploratory data analysis, analysts use various statistical and graphical techniques to visually and quantitatively summarize the data's main characteristics. This process involves detecting outliers, exploring patterns, identifying trends, and assessing data distribution. it helps uncover relationships between variables, assess data quality, and make informed decisions about subsequent analytical approaches.

K-means clustering, a method of unsupervised machine learning, was proposed by Stuart Lloyd in 1957. The algorithm gained popularity with the seminal work of John MacQueen in 1967 and its formalization by James MacQueen in 1968. K-means seeks to partition data points into k clusters, where each cluster is represented by a centroid

that minimizes the sum of squared distances between data points and the centroid. The iterative nature of the algorithm ensures convergence to stable cluster assignments. (Tryon, Robert C. et al 1939).

Customer data segmentation using k-means clustering is an invaluable technique in modern market analysis, revolutionizing how businesses understand and interact with their customers. The versatility and effectiveness of k-means, an unsupervised machine learning algorithm, have made it the go-to method for dividing vast amounts of customer data into meaningful clusters based on similarity. (Lloyd, S et al, 2014)

The process of k-means clustering begins with selecting the appropriate number of clusters, known as k, which represents the number of segments the data will be divided into. This decision is crucial and can significantly impact the results. Businesses must strike a balance between having enough clusters to capture the nuances of customer behavior while avoiding excessive complexity that might hinder practical applications.
In dividing or segmenting markets, researchers typically look for common characteristics such as shared needs, common interests, similar lifestyles, or even similar demographic profiles.

Once the number of clusters is determined, the algorithm starts the iterative process of assigning data points to clusters. Initially, random data points are selected as cluster centroids. The algorithm then calculates the distance between each data point and each centroid and assigns the data point to the nearest cluster. After the initial assignment, the centroids are recalculated as the mean of all the data points in their respective clusters.

This process continues iteratively until the centroids stabilize, and data points stop changing their cluster assignments.

Another significant benefit is the ability to implement targeted marketing strategies. Once customer segments are identified, businesses can create personalized marketing campaigns that speak directly to the needs and desires of each group. This personal touch can lead to increased engagement, loyalty, and ultimately, higher conversion rates. Customer Segmentation is used to divide a company's customer base into distinct groups based on common characteristics. It helps companies better understand their customers ((n.d.). *What is Customer Segmentation*. The clever programmer. https://thecleverprogrammer.com/2023/06/08/what-is-customer-segmentation/)

Product customization is a crucial aspect of modern business success. With k-means clustering, companies can identify the specific needs and preferences of each customer segment and develop products or services that cater to those requirements. This tailored approach increases customer satisfaction and sets the foundation for long-lasting relationships.

Optimizing resource allocation is a perennial challenge for businesses. Limited financial or human resources must be utilized efficiently to maximize returns. K-means clustering aids in this aspect by revealing which customer segments hold the most significant potential for growth or profitability. By focusing resources on the most promising segments, companies can ensure that their efforts yield the best results.

1.2 PROBLEM STATEMENT

To effectively segment customer data using k-means clustering and exploratory data analysis for market analysis, enabling businesses to understand customer behavior, tailor marketing strategies, optimize resource allocation, and gain a competitive advantage in diverse and dynamic markets.

1.3 AIM AND OBJECTIVES

The project aims to utilize exploratory data analysis and k-means clustering to segment customer profiling variables for market analysis

The objective of the project work includes:

- Acquisition of Dataset from Kaggle: Obtain the dataset from Kaggle to serve as the foundation for the analysis.
- Preprocessing of Dataset: Prepare the acquired dataset for analysis by cleaning, handling missing data, and formatting.
- Exploratory Data Analysis (EDA): Conduct EDA, which includes univariate, bivariate, and multivariate analyses to understand the dataset's characteristics and relationships.
- Application of K-Means Cluster Algorithm: Utilize the K-Means clustering algorithm to group data points with similar features.
- Determine the Optimal Number of Clusters: Employ both the Silhouette Score and the Elbow Point Method to identify the most suitable number of clusters for the data.
- Visualization, Observation, and Conclusion: Visualize the clustered data, observe patterns, and draw meaningful conclusions from the analysis, providing insights or recommendations as necessary.

1.4 SCOPE OF WORK

To use exploratory data analysis and implement customer data clustering using a k-means algorithm to create distinct customer segments. The analysis will provide valuable market insights, highlighting areas for improvement, enhancing customer experiences, and informing optimized business strategies. The project's focus is on data analysis and clustering techniques, without delving into the implementation of specific marketing strategies or business changes based on the findings.

1.5 RELEVANCE

Customer segmentation using k-means clustering and exploratory data analysis is highly relevant in various industries and business contexts. It offers several key advantages that make it a valuable tool for gaining insights and improving business strategies

1.6 OUTLINE OF THE THESIS

The subsequent sections of the report are structured as follows: Chapter Two provides an in-depth review of essential concepts and related works, encompassing details about the Dataset, methodology, and design employed in the study. In Chapter Three, the methodology and implementation processes are explained. Chapter Four presents the results and discussions, outlining the test procedures, data used, data analysis, and the resulting findings. Finally, Chapter Five concludes the report and offers valuable recommendations.

CHAPTER TWO

LITERATURE REVIEW

2.1 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an essential stage in data analysis. It entails thoroughly examining and using visualizations to comprehend the primary features, trends, and connections within datasets.

2.2 PURPOSE OF EXPLORATORY DATA ANALYSIS

2.2.1 DATA UNDERSTANDING AND FAMILIARITY

Exploratory data analysis is like diving into the dataset. It helps data experts get a full picture of the data, like its shape, how much of it there is, and what's inside. When they study things like the different kinds of data, how the numbers are spread out, and what's in the data, they really get to know what they're dealing with.

Exploratory data analysis helps analysts answer fundamental questions about the dataset, such as:

What are the key variables and their data types?

How many data points (observations) are there?

Are there missing values that need to be addressed?

What is the range and scale of the data?

2.2.2 PATTERN DISCOVERY AND INSIGHTS

Exploratory Data Analysis plays a crucial role in guiding decision-making processes. By gaining insights into the data's characteristics, analysts can make informed choices about subsequent data analysis steps.

Exploratory Data Analysis is a data analytics process to understand the data in depth and learn the different data characteristics, often with visual means. This allows you to get a better feel of your data and find useful patterns in it. (Avijeet Biswal et.al)

2.2.3 DECISION SUPPORT AND MODEL SELECTION

Exploratory Data Analysis plays a crucial role in guiding decision-making processes. By gaining insights into the data's characteristics, analysts can make informed choices about subsequent data analysis steps. This includes selecting appropriate modeling techniques and feature engineering strategies that align with the data's properties.

2.3 TYPES OF EXPLORATORY DATA ANALYSIS

2.3.1 UNIVARIATE ANALYSIS

Univariate analysis looks at one thing at a time. Its main aim is to figure out how that one thing is spread out and what it's like. Univariate analysis gives us the basics we need to explore and sum up individual things, and this is important for understanding what they are like and how they're connected to other things.

2.3.2 BIVARIATE ANALYSIS

Bivariate analysis involves the examination of the relationship between two variables simultaneously. Its primary goal is to understand how one variable may be related to or influenced by another variable. Bivariate analysis helps identify associations and

dependencies between pairs of variables, making it useful for hypothesis generation and understanding cause-and-effect relationship.

2.3.3 MULTIVARIATE ANALYSIS

Multivariate analysis extends the exploration to three or more variables simultaneously. It aims to uncover more complex relationships and patterns involving multiple variables

2.4 CLUSTERING ANALYSIS

Clustering being an unsupervised learning algorithm is trained on a dataset without explicit supervision or labeled output. In unsupervised learning, the goal is to find patterns, structures, or relationships within the data without having specific target values to guide the learning process. The algorithm explores the inherent structure of the data to identify similarities, groupings, or distributions.

2.5 WHAT IS UNSUPERVISED LEARNING AND HOW DOES IT RELATE TO CLUSTERING

unsupervised learning is a method we use to group data when no labels are present. Since no labels are present, unsupervised learning methods are typically applied to build a concise representation of the data so we can derive imaginative content from it (Pykes, K. (n.d.)). In this approach, the algorithm tries to identify patterns, structures, or relationships within the data on its own, without being provided with specific target labels or categories for the input data.

2.6 CUSTOMER SEGMENTATION AND IT'S TYPES

Segmentation means grouping entities based on similar properties. Entities could be customers, products, and so on (Ibrahim Abayomi Ogunbiyi et.al). Customer

segmentation in particular, means grouping customers based on similar features or properties as it can relate to their income, age, and purchasing ability, as such customers can be segmented based on:

a. Demographic segmentation

Demographic customer segmentation is a way to group people based on specific characteristics like age, gender, and income.

b. Behavioral segmentation

Behavioral segmentation is like sorting people based on their actions or behaviors. Instead of looking at their age or gender, we look at what they do or how they interact with products and services

c. Geographical segmentation

Geographic segmentation involves sorting people based on where they live or their location. They divide their customers into groups based on where they are located.

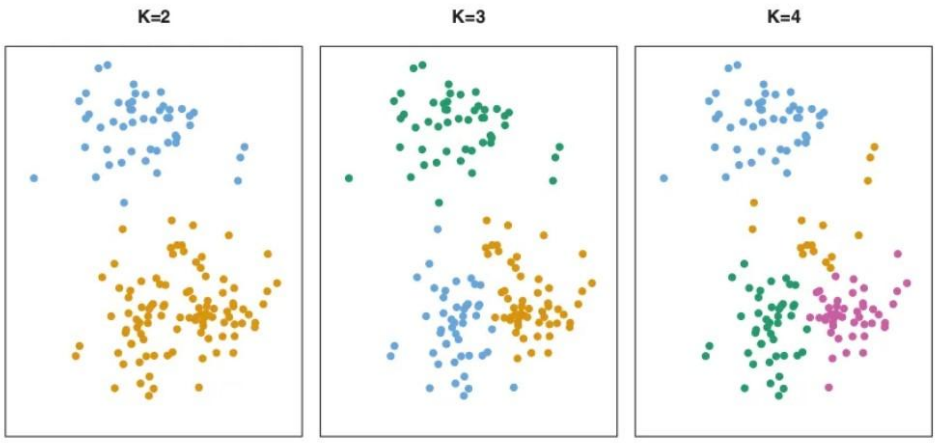
d. Psychographic segmentation

Psychographic segmentation involves sorting people based on what they think, feel, and believe. You can group people based on their interests, values, personalities, and beliefs.

2.7 K-MEANS CLUSTERING AS IT RELATES TO CUSTOMER SEGMENTATION

K-means clustering is a method for grouping n observations into K clusters. It uses vector quantization and aims to assign each observation to the cluster with the nearest mean or centroid, which serves as a prototype for the cluster. Originally developed for signal processing, K-means clustering is now widely used in machine learning to partition data points into K clusters based on their similarity. The goal is to minimize the sum of squared distances between the data points and their corresponding cluster centroids, resulting in clusters that are internally homogeneous and distinct from each other.

K-means clustering is an elegant and straightforward method for dividing data collection into K-separate, non-overlapping groups. To do K-means clustering, we must first select the required number of clusters K; then, each observation will be assigned exactly one of the K values by the K-means algorithm.



(Sharma, P. (n.d.). et.al 2023)

Figure 2.1

Let's try to understand this with a simple example. A bank wants to give credit card offers to its customers. Currently, they look at the details of each customer and, based on this information, decide which offer should be given to which customer.

(Pulkit Sharma et.al ,2023)

Figure 2.2

2.7.1 PROPERTIES OF K-MEANS CLUSTERING

1. All the data points in a cluster should be similar to each other:

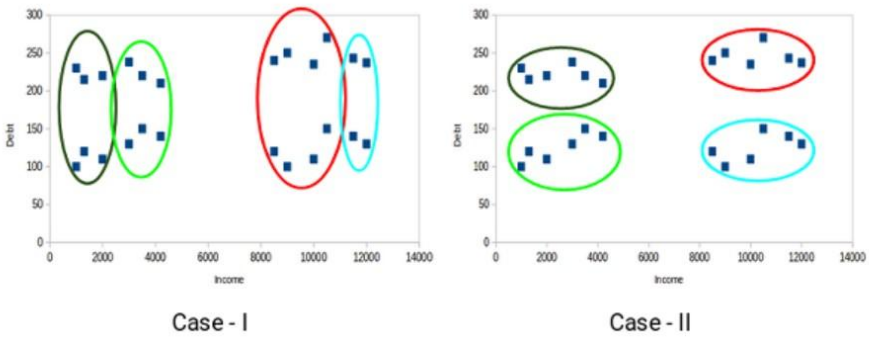
When you put things in a group, those things should be really similar to each other. In other words, all the stuff in the group should be a lot alike.

(Sharma , P. (n.d.). et.al ,2023)

figure 2.3

2. The data points from different clusters should be as different as possible:

When we put data into groups, we want the things in each group to be a lot like each other and different from things in other groups. This way, we can make sure that when we sort data into these groups, it makes sense and helps us understand and categorize the data better.



(Sharma , P. (n.d.). et.al ,2023)

Figure 2.4

2.8 METHODS TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS

Silhouette Score and Elbow Method are like tools that help us figure out how many groups we should make when using K-means clustering, which is a way to group data. These tools help us decide the best number of groups.

2.8.1 SILHOUETTE SCORE METHOD

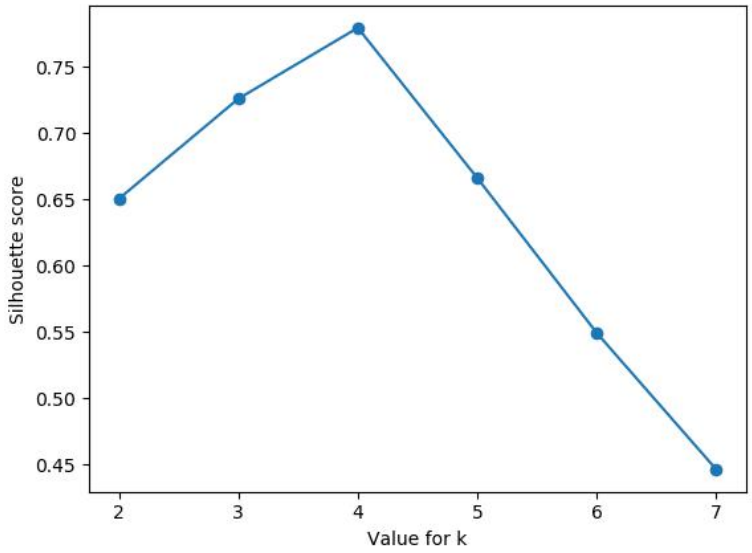
The Silhouette Score measures how similar each data point in one cluster is to the other data points in the same cluster compared to the nearest neighboring cluster.

The Silhouette Score ranges from -1 to 1

A high positive value (close to 1) indicates that the data point is well-clustered and far from neighboring clusters.

A score around 0 means that the data point is on or very close to the decision boundary between two neighboring clusters.

A negative score (close to -1) suggests that the data point may have been assigned to the wrong cluster.



(Sharma, P. (n.d.). et.al 2023)

Figure 2.5

To determine the optimal number of clusters using the Silhouette Score, you calculate the score for different numbers of clusters and choose the number that yields the highest average Silhouette Score across all data points.

2.8.2 ELBOW POINT METHOD

The Elbow Method is a visual technique for finding the optimal number of clusters by plotting the variance explained as a function of the number of clusters.

You perform K-means clustering for a range of cluster numbers, typically starting from 2 clusters and going up to some reasonable upper limit. For each cluster number, you calculate the total within-cluster sum of squares .

You then plot the number of clusters against the WCSS. The point at which the rate of decrease in WCSS starts to slow down (forming an "elbow" in the plot) indicates the optimal number of clusters.

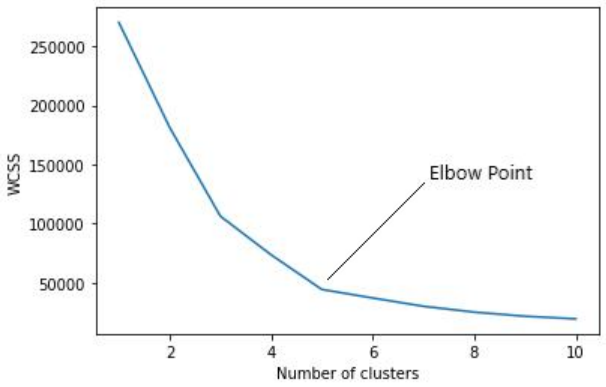


Figure 2.6 (Aditya et al.,2018)

2.9 REVIEW OF RELATED WORK

The works related to the study are reviewed and critiqued in this section, which includes the various approaches to segmenting customer data for market analysis.

In this regard**, Kaylivly Tabianan et al** proposed the "**K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data**". The research focuses on using K-Means clustering to perform intelligent customer segmentation based on customer purchase behavior data in the context of E-commerce systems. The goal is to divide customers into groups with similar characteristics to optimize services and products and increase business profits. E-commerce systems have become increasingly popular, especially during the COVID-19 pandemic, as people started using online shopping platforms to purchase essential items conveniently from home. Customer segmentation plays a crucial role in E-commerce to understand customers' needs and interests, which enables businesses to target the right customers with relevant products and services.

Jun Wu, Li Shi et al conducted a "**study on customer segmentation and value analysis in the e-business world"**, focusing on the dynamic customer purchase behaviors in China's online shopping environment. They emphasized the significance of data mining as the most suitable method for customer purchase behavior analysis, especially in the era of big data. The research aimed to develop a customer-oriented marketing strategy by predicting customer online behaviors based on data mining. To achieve this, they used real-world data from an enterprise in Beijing, China, and applied the RFM (Recency, Frequency, Monetary) model and K-means clustering algorithm for customer segmentation and value analysis.

E Ernawati et al conducted a "**study on customer segmentation using data mining methods and the RFM (Recency Frequency Monetary) model"**. The main objective of their research was to improve marketing strategies by dividing customers into smaller homogeneous groups and targeting each group individually. Customer segmentation is essential for businesses as not all marketing strategies are suitable for every customer. Data mining methods play a crucial role in identifying hidden trends and relationships within large datasets. The RFM model is widely used for customer behavior analysis, and it focuses on recency, frequency, and monetary factors to categorize customers into distinct groups. To conduct their research, the authors performed a literature search using three databases, namely Scopus, Web of Science (WoS), and Emerald. They used specific keywords related to customer segmentation, data mining, and the RFM model within the timeframe of 2015 to 2020. After eliminating duplicates and screening abstracts, they selected 44 articles that met their inclusion criteria. The study's results and discussions highlighted the significance of the RFM-based model in customer segmentation. commerce, finance, healthcare, and IT. It helps companies gain a deeper

Fahmida Afrin et al carried out research that focuses on **"customer segmentation using data mining techniques"**. They compare the performance of two clustering algorithms, K-means and Fuzzy C-means (FCM), after implementing Principal Component Analysis (PCA) for dimensionality reduction. The research conducted by Fahmida Afrin, Md. Al-Amin and Mehnaz Tabassum focus on customer segmentation using data mining techniques. They compare the performance of two clustering algorithms, K-means and Fuzzy C-means (FCM), after implementing Principal Component Analysis (PCA) for dimensionality reduction. The study begins with a brief introduction to data clustering, highlighting its importance in various fields, including

customer segmentation. It emphasizes that understanding customer behavior is crucial for businesses to develop effective marketing strategies. The authors then discuss the methodologies used in their research. They apply K-means and FCM algorithms for clustering and use PCA as a preprocessor to reduce the high-dimensional and noisy data. PCA is a technique that transforms correlated variables into a smaller set of uncorrelated variables called principal components. It helps in visualizing and analyzing the data more efficiently.

Zhao Xian et al explore the use of **"historical sales and behavioral data analytics to construct a recommendation model for traditional offline stores looking to boost their sales by transitioning to online B2C business models"**. The sudden COVID-19 pandemic has forced many offline shops to face challenges, leading them to seek new ways of doing business. The researchers emphasize the potential of online shopping, especially with the advanced algorithms used in recommender systems and customer value management. These technologies enable customer-oriented marketing strategies, which can enhance customer satisfaction and increase corporate profits simultaneously. To achieve this transformation, the study highlights the importance of data mining as the most effective way to analyze customer purchase behavior and discover hidden useful information from massive online transaction data. Start-up e-commerce companies, like cosmetics e-commerce companies discussed in the study, can utilize data mining techniques to find customer favorites, optimize inventory, predict sales, and make product recommendations. This helps maximize sales, avoid stock shortages, and reduce sales stagnation due to outstanding stocks. Traditional offline shops often face limited markets, long sale cycles, and complex and time-consuming sale processes. However, with the right data analytics and recommendation models, they can adapt to

the current market situation and overcome these challenges. The study proposes a combination of recency, frequency, and monetary (RFM) analysis methods and the k-means clustering algorithm to segment customer levels in the company. The association rule theory and the apriori algorithm are then utilized for shopping basket analysis and product recommendations based on the results.

Ardvin Kester S et al aim to analyze the preferences of consumers for Samgyeopsal, a popular Korean grilled dish, in the Philippines. The study focuses on the attributes of Samgyeopsal, which include the main entrée, cheese inclusion, cooking style, price, brand, and drinks. The researchers employed Conjoint Analysis and k-means clustering to understand consumer preferences and identify different market segments. The popularity of Samgyeopsal in the Philippines has grown due to the influence of Hallyu, the Korean Wave, which has introduced Korean culture, including its cuisine, to a global audience. To better understand the preferences of consumers, the researchers collected 1018 responses through an online survey conducted on social media platforms, using a convenience sampling approach. The respondents were asked to rate the importance of each attribute in their decision-making process when choosing Samgyeopsal. The results of the Conjoint Analysis revealed the relative importance of each attribute. The main entrée was found to be the most crucial attribute, with a weightage of 46.314%. Following that, cheese inclusion was rated second with 33.087%, price ranked third at 9.361%, drinks were fourth with 6.603%, and cooking style was rated the least important at 3.349%.

CHAPTER THREE

THEORETICAL FRAMEWORK

3.1 WORKFLOW PROCESS

The procedures and methods adopted to achieve the objectives of this research work are outlined in Fig. 3.1.

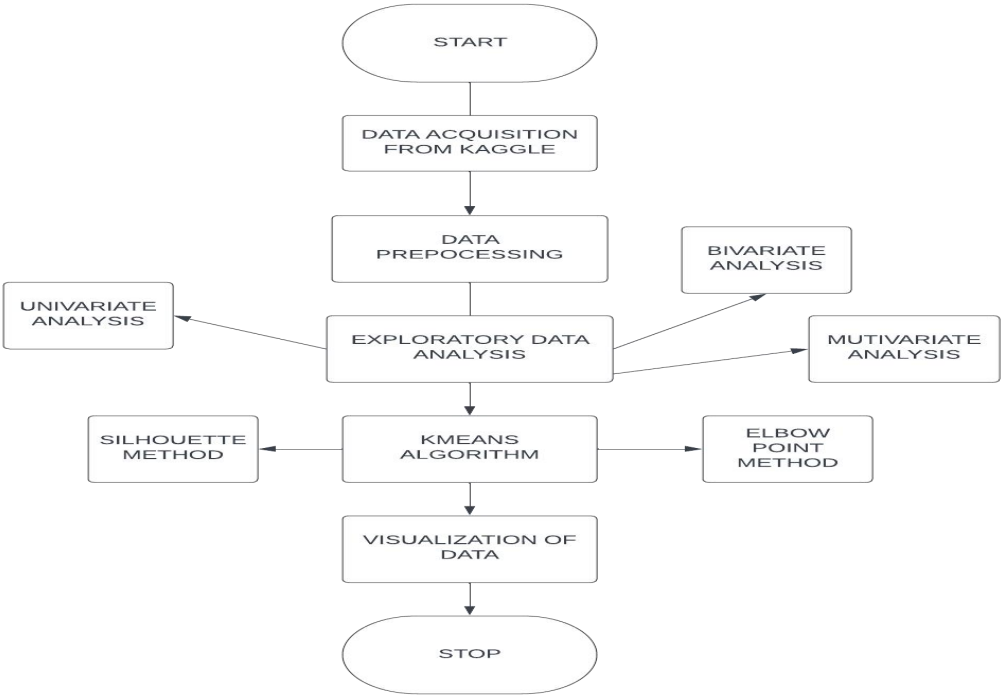


Figure 3.1

From the workflow diagram above it can be seen, that the process begins with obtaining a comprehensive dataset that includes diverse customer attributes, such as demographics, purchase history, behavioral patterns, and other relevant information. The quality and completeness of the dataset are crucial as they directly impact the accuracy and effectiveness of the clustering results. Data cleaning and preprocessing techniques may be employed to handle missing values and ensure data integrity.

Following data acquisition, exploratory data analysis is conducted to gain deeper insights into the dataset's characteristics. Exploratory Data Analysis involves data summarization, visualization, and outlier detection to identify patterns and understand the data's underlying structure. This step is instrumental in making informed decisions during the clustering process and helps ensure that data quality issues are addressed early on.

The next critical step in the flow is to determine the optimal number of clusters for the K-Means algorithm. Two widely used techniques, the elbow method, and the silhouette score are utilized for this purpose. The elbow method involves plotting the sum of squared distances (inertia) of data points to their assigned cluster centroids for different numbers of clusters. The optimal number of clusters is determined at the "elbow" point, where the inertia starts to level off, indicating that adding more clusters does not significantly improve the clustering. Simultaneously, the silhouette score measures the similarity or dissimilarity between data points within clusters and across different clusters. A higher silhouette score indicates well-defined and distinct clusters and the optimal number of clusters corresponds to the maximum silhouette score.

With the optimal number of clusters established, the K-Means algorithm is implemented. The algorithm aims to partition the data into K clusters, where each data point is assigned to the cluster whose centroid is closest to it.

Visualizing the K-Means clustering results is a crucial step to gaining a comprehensive understanding of the customer segments.. Visualizations offer insights into the characteristics of each cluster, enabling businesses to comprehend the preferences and behaviors of different customer groups.

Once the clustering process is completed, the results are obtained for each customer. Clustering outcomes include cluster assignments for each customer, enabling businesses to effectively segment their customer base. Additionally, the reconstruction error, measuring the similarity or dissimilarity between customers within each cluster, is utilized to assess the clustering quality. By setting a threshold on the reconstruction error, unique or anomalous customer segments can be identified, allowing businesses to focus on these distinctive groups for targeted marketing strategies.

3.2 ACQUISITION OF DATASET FROM KAGGLE

The dataset for this project is obtained from a public repository (Dr. Omar Romero-Hernandez et al) and was published on Kaggle.

The dataset focuses on customer personality and how it affects the market, and includes the following profiling variables amongst other columns in the dataset

3.3 PREPROCESSING OF DATASET

From the dataset obtained from Kaggle, we can process and analyze it using Python. With Pandas, we read and handle missing data, making the data ready for visualization. Once we finish these steps, we can create charts and graphs to find patterns and insights in the data. Since the dataset is in CSV format, we use Pandas to open it easily.

After preprocessing, it can be found that null values are not in the dataset

```
df.isnull()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	AcceptedCmp1	AcceptedCmp2	Complain
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2236	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2237	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2238	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2239	False	False	False	False	False	False	False	False	False	False	...	False	False	False

Figure 3.2

3.4 EXPLORATORY DATA ANALYSIS

The types of exploratory data analysis that was carried out include:

3.4.1 UNIVARIATE ANALYSIS

3.4.2 STEPS IN CARRYING OUT UNIVARIATE ANALYSIS

Step 1: Obtain the column in the dataset for analysis

Select a specific column from the dataset for detailed examination.

Step 2: Perform this analysis for all relevant columns

Apply the chosen analysis method to the column in the dataset

Step 3: visualization and understand the pattern, relationships

Create visual representations and explore patterns and connections within the data

Step 4: Repeat for all columns needed

3.4.3 BIVARIATE ANALYSIS

3.4.4 STEPS IN CARRYING OUT BIVARIATE ANALYSIS

Step 1: Select Two Columns

Choose the two columns (variables) you want to compare. For example, you might compare the "age" and "Marital status" columns to investigate if there's a connection between these two factors.

Step 2: Group and analyze

Group the column and analyze the data and this can be done using the describe method on the dataset variable, which pulls out the information needed to gain insight into the data

Step 3: Visualization

Plot the results to visualize the results of the distribution of the columns among the customers. You can use a histogram, box plot, or other appropriate visualization techniques to display the distribution of ages.

Step 4: Repeat for all two-column comparisons in the dataset when needed

This allows you to understand each column's individual properties and compare them to other columns or datasets. Repeat Step 4 as necessary for each column or variable of interest in your dataset.

3.4.5 MULTIVARIATE ANALYSIS

3.4.6 STEPS IN CARRYING OUT BIVARIATE ANALYSIS

Step 1: Group the customer features to be analyzed i.e. in our case we used 3 column comparison from the dataset

Step 2: Visualize the result for the comparison

Step3: Repeat steps 1 and 2 for all 3 or more column comparisons for as much as needed

3.5 K-MEANS CLUSTER ALGORITHM

3.5.1 EXPRESSION FOR THE K-MEANS ALGORITHM

The expression for the algorithm is as follows (Kmeans et al):

$J=\sum i=1n\|xi-centroid\_j\|\ ^2$ ----------(3.1)

J = This represents the cost function or objective that K-means aims to minimize. It's often called "inertia" or "within-cluster sum of squares." The goal is to minimize this value

n: The total number of data points in the dataset.

xi: A specific data point in the dataset. K-means assigns each data point to one of the K clusters.

Centroid_j: The centroid of a cluster. This is a point that represents the center of the cluster and is calculated as the mean (average) of all the data points assigned to that cluster.

3.6 HOW TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS

3.6.1 BY USING THE WITHIN-CLUSTER SUM OF SQUARES AND THE ELBOW GRAPH METHOD

3.6.2 EXPRESSION FOR THE ELBOW POINT METHOD

The expression for the equation is as follows (David et al):

SSD(k) = Σ(distance(xi, Cj)^2) ----------(3.2)

Xi = data point.

Cj = centroid of the cluster to which xi belongs

3.6.3 STEPS IN DETERMINING THE NUMBERS OF CLUSTERS USING WCSS AND THE ELBOW GRAPH METHOD

a) **Calculate K-Means for Various K Values**: Run the K-Means method first for a range of K values (for instance, from 1 to a respectably high number). Calculate the WCSS for every K value.

b) **Calculate the WCSS for each K:** Calculate the sum of squared distances between data points and their cluster centers (cluster centroids) for each value of K.

c) **Plot WCSS Against K:** Create a line plot or a scatter plot with K on the x-axis and WCSS on the y-axis. The plot will typically show a decreasing trend in WCSS as the number of clusters (K) increases.

d) **Identify the Elbow Point:** Look for the "elbow point" on the plot. The elbow point is the value of K at which the rate of decrease in WCSS starts to slow down significantly.

e) **Select the Number of Clusters (K):** Based on the Elbow Method, choose the value of K at the elbow point as the number of clusters for your K-Means algorithm.

3.7 BY USING THE SILHOUETTE SCORE METHOD TO DETERMINE THE OPTIMAL NUMBER OF CLUSTERS

3.7.1 EXPRESSION FOR SILHOUETTE EQUATION

The expression for the silhouette equation is given as :

S(i)= b(i)−a(i) / max(an(i),b(i)) ----------(3.3)

n = total number of data points.

a(i) = average distance from data point i to all other data points in the same cluster as i.

b(i) = smallest average distance from data point I to all data points in any other cluster (i.e., the nearest neighboring cluster to which I do not belong).

S(i)= silhouette score for data point I, which measures how similar it is to its cluster compared to other clusters.

3.7.2 STEPS IN DETERMINING THE OPTIMAL NUMBER OF CLUSTERS USING THE SILHOUETTE SCORE

a) **Run K-Means for Different Values of K:** As with the Elbow Method, start by running the K-Means algorithm for a range of K values (e.g., from 2 to a reasonably high number).
b) **Calculate Silhouette Score for Each K:** For each value of K, calculate the Silhouette Score for the entire dataset.
c) **To calculate the Silhouette Score for an individual data point:** Calculate the average distance (a) of the data point to all other data points within the same cluster.
d) **Plot Silhouette Score Against K:** Create a line plot or a bar plot with K on the x-axis and the corresponding Silhouette Score on the y-axis. The plot will show the Silhouette Score for each value of K.
e) **Identify the Optimal K:** Look for the value of K that gives the highest Silhouette Score. The value of K that maximizes the Silhouette Score is considered the optimal number of clusters.
f) **Confirm the Number of Clusters:** Based on the Silhouette Score and the visual inspection, confirm the number of clusters you want to use in your K-Means algorithm.

3.8 VISUALIZATION, OBSERVATION, AND CONCLUSION DRAWN FROM THE DATA

Performing visualization using Matplotlib and Seaborn for customer data segmentation using k-means can be divided into several steps:

a) **VISUALIZATION OF EXPLORATORY DATA ANALYSIS CLASSES AND CUSTOMER CLUSTERS**

Visualize data from the univariate analysis, bivariate analysis, and multivariate analysis to gain insights into the data. After that carry out the visualization for clustered data by creating scatter plots by using the K-means clustering algorithm.

OBSERVATIONS AND CONCLUSIONS

Make observations from the visualizations. Are there any clear separations between clusters? Are certain customer groups more closely related? What characteristics define each cluster? Based on these insights, conclude customer segmentation and identify potential marketing strategies for each cluster.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 RESULT PRESENTATION OVERVIEW

This section presents the results of the methods carried out in Chapter 3. Section 4.2 presents results from obtaining the dataset and preprocessing. Section 4.3 presents the results from the univariate, Bivariate, and multivariate analysis. Section 4.4 presents the results from evaluating the datasets and grouping them into clusters so easy identification.

4.2 RESULT PRESENTATION FROM OBTAINING THE DATASET AND PREPROCESSING

The column that outputs a True is null, while any column that outputs a False has no null value

```
df.isnull()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	AcceptedCmp1	AcceptedCmp2	Complain
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False
4	False	False	False	False	False	False	False	False	False	False	...	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2236	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2237	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2238	False	False	False	False	False	False	False	False	False	False	...	False	False	False
2239	False	False	False	False	False	False	False	False	False	False	...	False	False	False

Plate 4.1

4.3 RESULTS OBTAINED FROM THE EXPLORATORY DATA ANALYSIS

4.4 UNIVARIATE ANALYSIS

4.4.1 AGE DISTRIBUTION

```
sns.histplot(data=df, x="Age", bins = list(range(10, 150, 10)))
plt.title("Distribution of Customer's Age")
plt.savefig("Age.png");
```

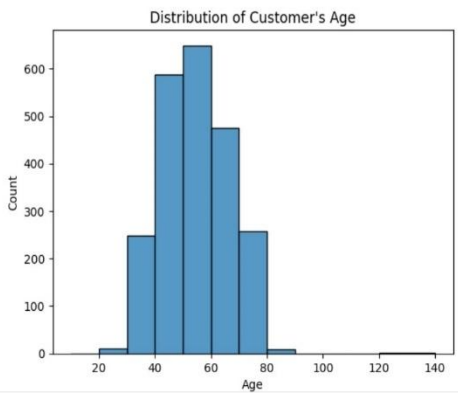


Plate 4.2

Looking at the chart above, we can tell that the majority of our customers fall in the age range of 41 to 60.

4.4.2 EDUCATION

```
df["Education"].value_counts(normalize=True).plot.bar(figsize=(8, 6))
plt.xticks(rotation=45)
plt.title("Frequency of Customer's Education [proportion]");
```

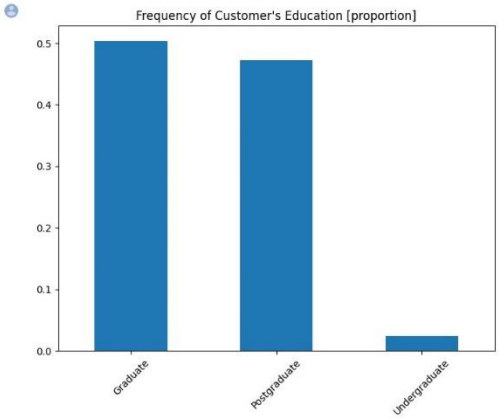


Plate 4.3

The summary above shows that 50% of our customers have completed their first-degree graduation. After that, the next largest group of customers has a bachelor's degree, followed by those with a postgraduate education.

4.4.3 MARITAL STATUS

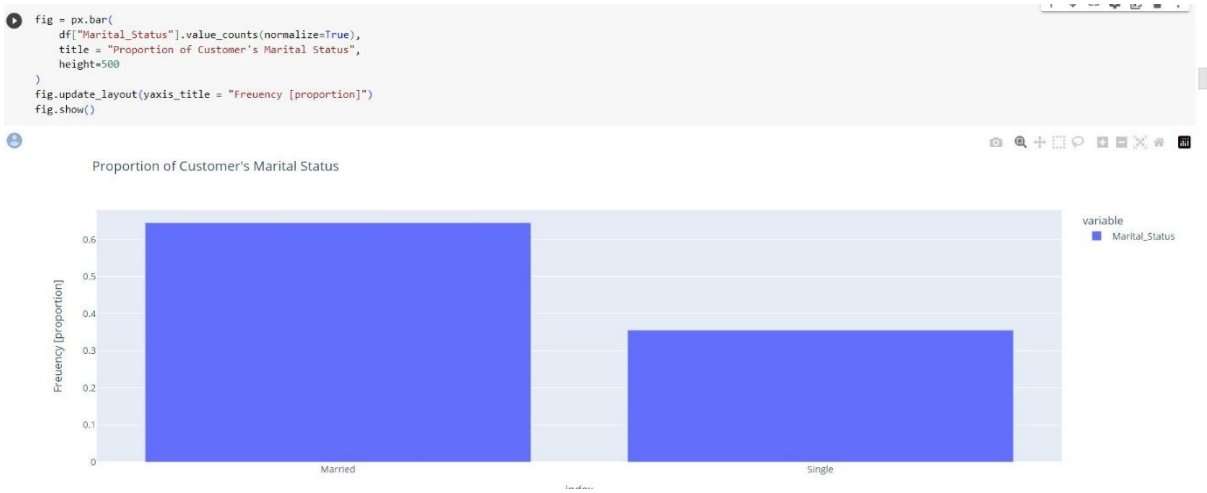


Plate 4.4

The summary above tells us that about 65% of our customers are married, and the rest, nearly 35%, are single.

4.4.4 INCOME DISTRIBUTION

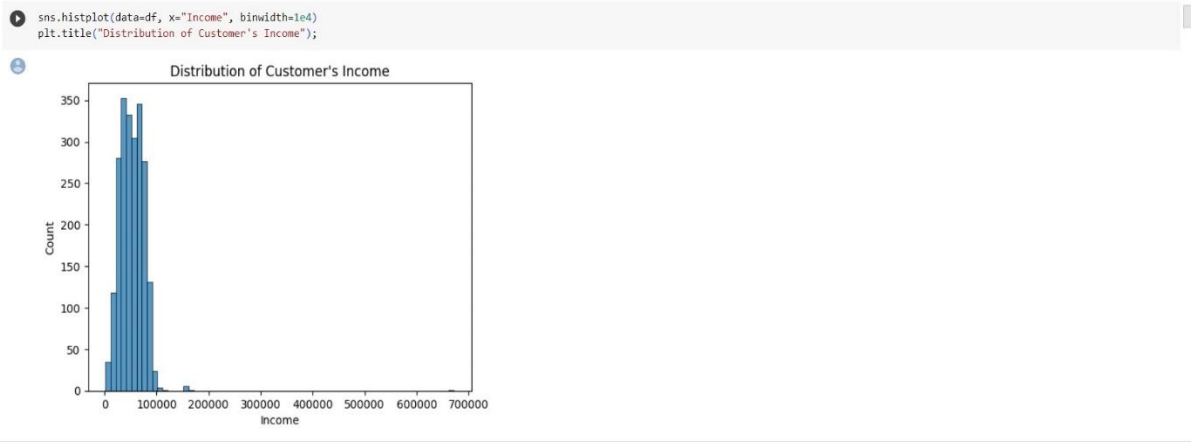


Plate 4.5

It is seen that the majority of customer's income is within 0-100k\$. However, we have other customers that earn way more than that (above 600k\$)

4.5 BIVARIATE ANALYSIS

4.5 RELATIONSHIP BETWEEN AGE AND TOTAL AMOUNT SPENT

4.5.1 USING SCATTERPLOT FROM THE SEABORNE LIBRARY IN PYTHON

Looking at the result below, it's clear that age doesn't have a strong connection with how much money customers spend. In simple terms, knowing a person's age doesn't really tell us how much they'll spend.

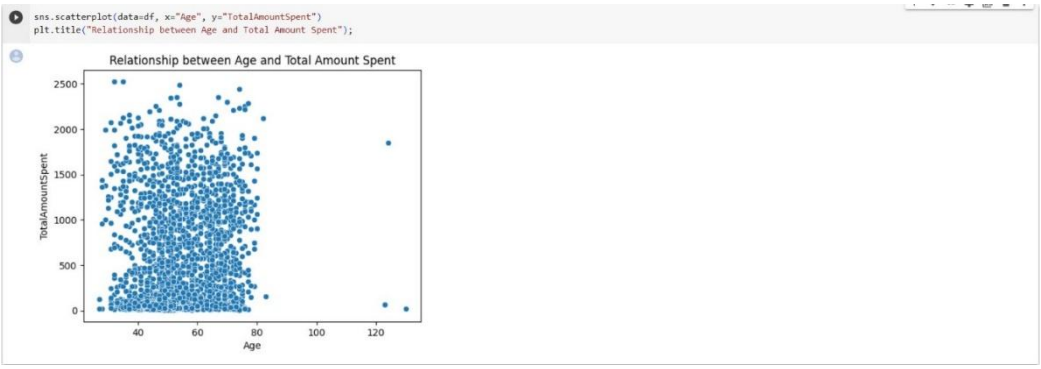


Plate 4.6

4.5.2 USING THE BAR CHAT FROM THE MATPLOTLIB LIBRARY

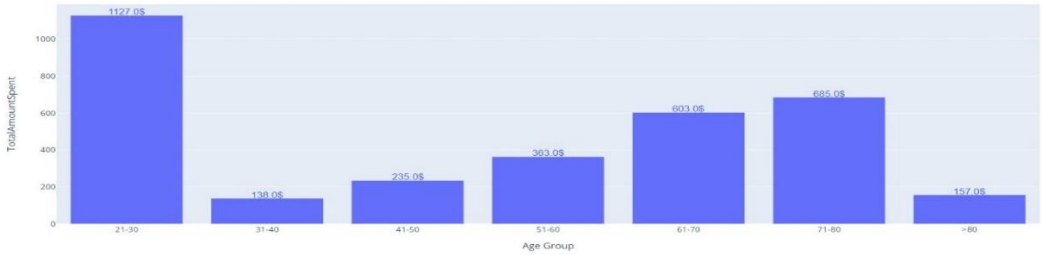


Plate 4.7

Looking at the summary above, we find the average spending for different age groups. It appears that the age group spending the most on average is those between 21 and 30 years old. Next in line are customers between 71 and 80 years old. Now, let's check how these two groups compare.

4.5.3 RELATIONSHIP BETWEEN EDUCATION AND TOTAL AMOUNT SPENT

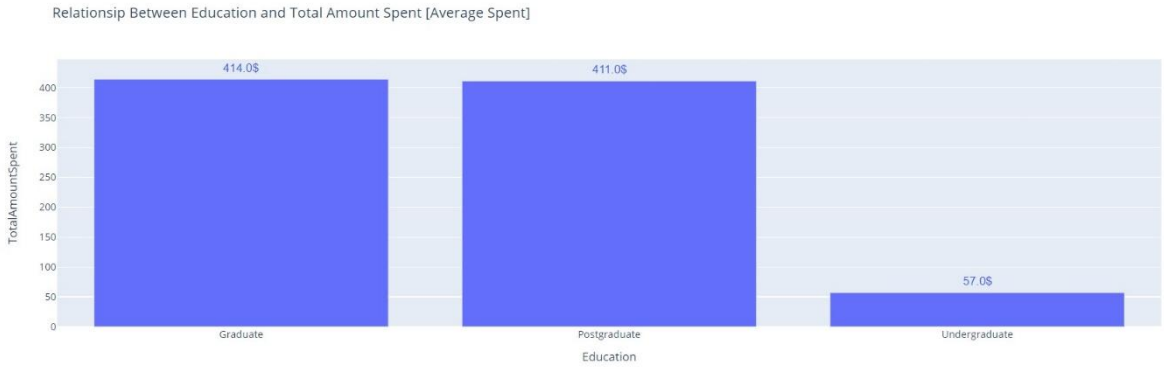


Plate 4.8

The visualized data shows little spending difference between graduate and postgraduate customers, but a significant drop for undergraduates. Graduate+ spends about 7 times more

4.5.4 RELATIONSHIP BETWEEN MARITAL STATUS AND INCOME

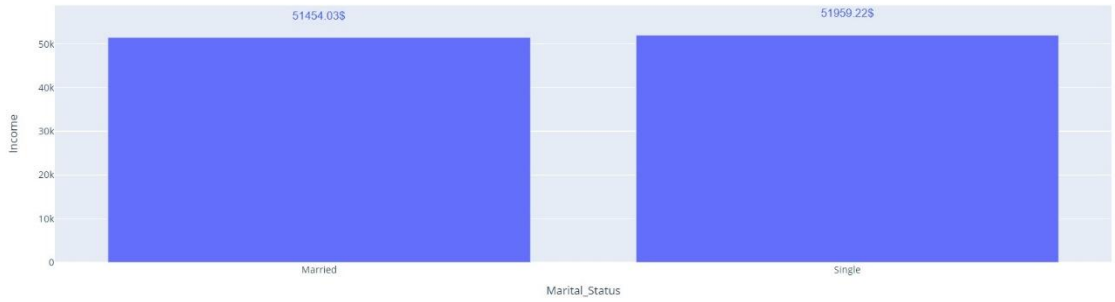


Plate 4.9

Based on the results shown above, it's clear that there's no connection between whether someone is married and how much money they make. Customers generally earn about the same amount, regardless of their marital status.

4.5.4 RELATIONSHIP BETWEEN AGE GROUP AND INCOME OF THE CUSTOMER

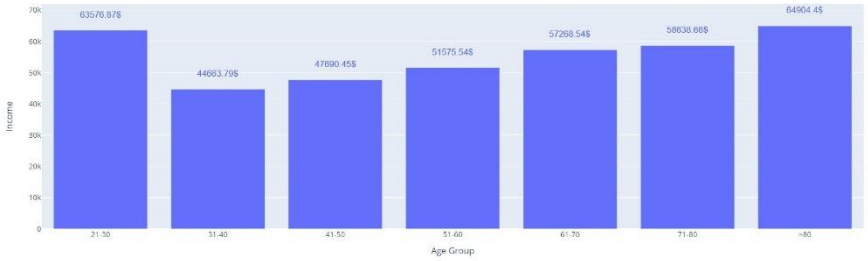


Plate 4.10

Looking at the chart, it's clear that the people who make the most money, on average, are those above 80 years old, followed by those between 71 and 80. The interesting

thing is, except for the folks aged 21 to 30, there's a general pattern where income tends to go up as people get older.

4.5.5 RELATIONSHIP BETWEEN EDUCATION AND INCOME OF THE CUSTOMERS

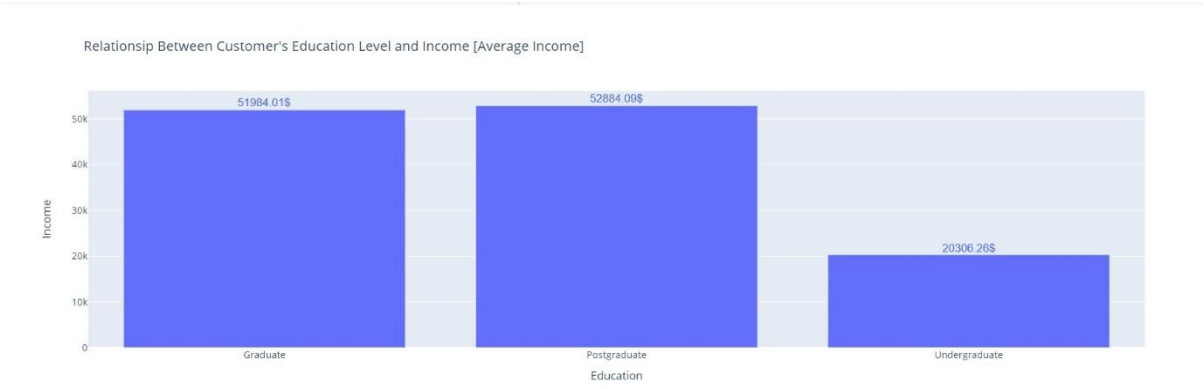


Plate 4.11

Looking at the chart, it's clear that people who have completed graduate and postgraduate degrees make twice as much money as those with just an undergraduate degree

4.5.6 RELATIONSHIP BETWEEN MARITAL STATUS AND TOTAL AMOUNT SPENT BY CUSTOMER

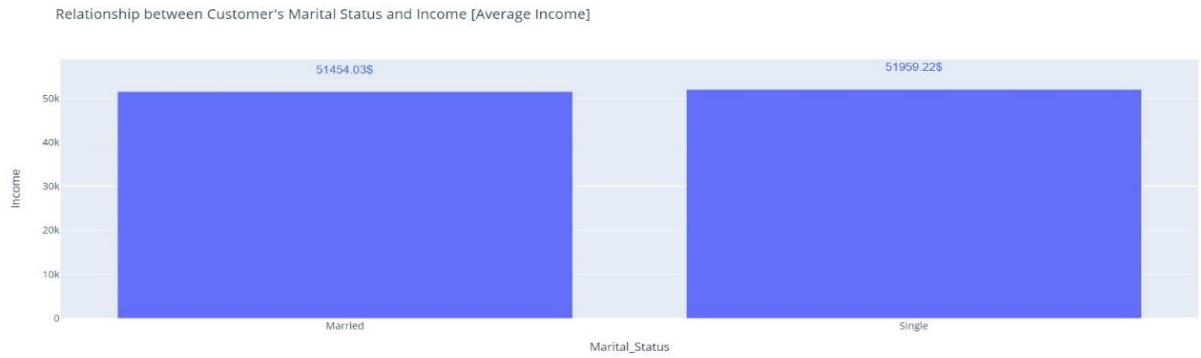


Plate 4.12

From the above diagram, we can see from the above summary that there is no relationship between the customer's marital status and the average amount spent.

4.5.7 RELATIONSHIP BETWEEN INCOME AND TOTAL AMOUNT SPENT BY THE CUSTOMER

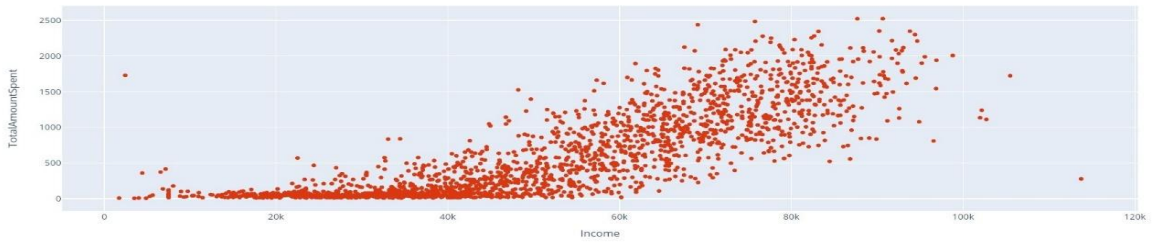


Plate 4.13

We can see from the above summary that the Income of a customer determines the total amount to be spent on the product. As a customer's income increases so does what they buy increases.

4.5 MULTIVARIATE ANALYSIS

4.6.1 RELATIONSHIP BETWEEN INCOME, EDUCATION, AND TOTAL AMOUNT

Plate 4.14

Firstly from the exploratory Data Analysis, we saw that Income was the key indicator that determined the amount a customer would spend.

Also In terms of Education, we noticed customers with a graduate education level and above tend to spend 12 times more than those customers with an undergraduate education level.

4.7 RESULT OBTAINED FROM THE K-MEANS CLUSTERS

4.7.1 RESULTS OBTAINED FROM WITHIN THE CLUSTER SUM OF SQUARES AND ELBOW POINT METHOD

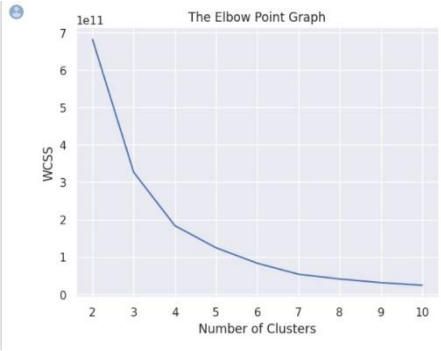


Plate 4.15

From the above, we can see from the diagram the maximum number of clusters at the elbow point is 3

4.7.2 RESULT OBTAINED FROM THE SILHOUETTE SCORE METHOD

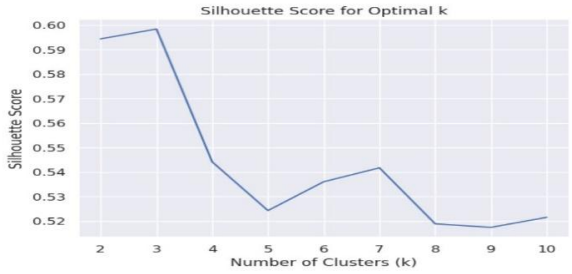


Plate 4.16

From the above diagram, we can see that the maximum point of clusters was obtained at the highest point of the 3

4.7.3 RESULTS OBTAINED FROM THE CLUSTERS OF INCOME AND TOTAL AMOUNT SPENT

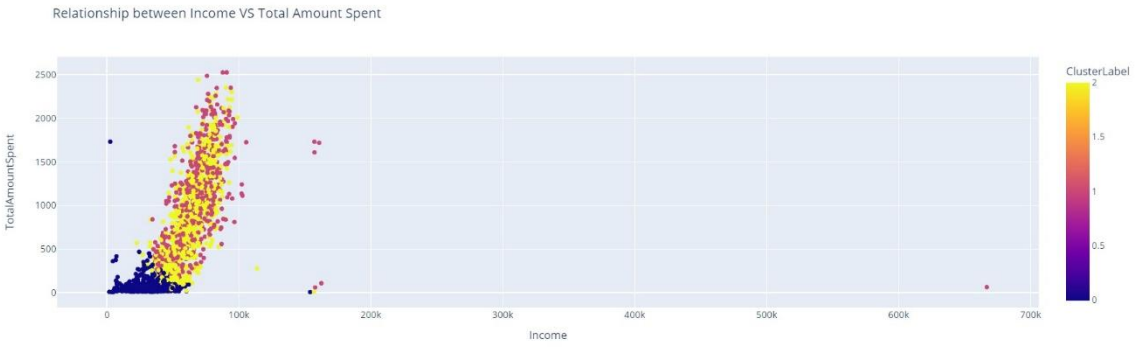


Plate 4.17

From the above diagram, we can inference:

- Cluster 0(blue cluster): Customers with low Income and Low spending.
- Cluster 1(orange cluster): Customer with moderate Income and Moderate spending.
- Cluster 3(yellow cluster): Customers who earn much and spend much.

4.7.4 RESULTS OBTAINED FROM THE CLUSTERS OF INCOME, TOTAL AMOUNT SPENT, AND AGE

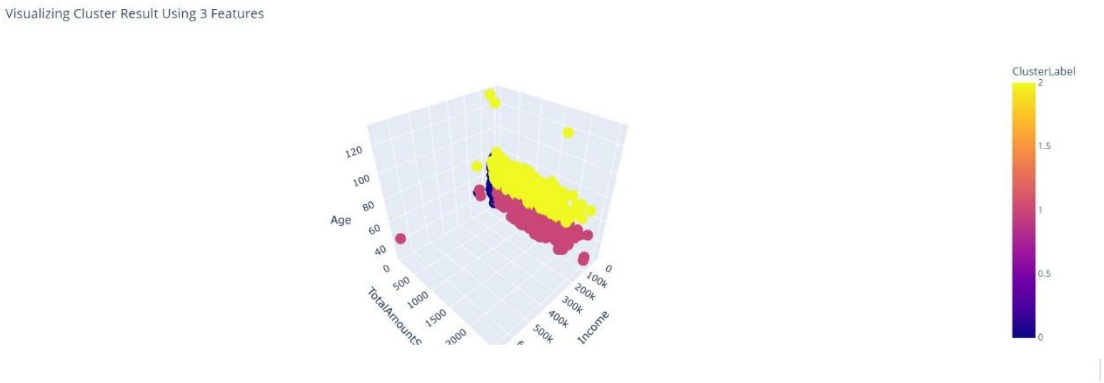


Plate 4.18

- Cluster 1 depicts young customers who earn a lot and also spend a lot.
- Cluster 2 translates to old customers that earn a lot and also spend high.
- Cluster 3 depicts young customers who earn low and spend low.

4.8 PERFORMANCE ANALYSIS

In comparison to the works carried out by the author in my literature review, all the authors employed only the elbow point method and this work employed both the elbow point method and the silhouette method to determine the optimal number of clusters, resulting in a higher accuracy in determining the numbers of cluster. Moreover, our comprehensive exploratory data analysis added depth to our approach, showcasing its superiority over the referenced study, which relied solely on the elbow point method for cluster determination

4.9 DISCUSSION

The exploratory data analysis revealed insights into customer demographics. Most are aged 41-60, hold first-degree graduations, and are married (65%). Income distribution is mainly within 0-100k$, with some earning above 600k$. The bivariate analysis highlighted spending differences based on education. K-means clustering identified three clusters: low earners with low spending, moderate earners with moderate spending, and high earners with high spending. This analysis aids in targeted marketing and sales strategies.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The project's data analysis has given us important information that can help us make better business decisions. We now know more about our customers and how they spend their money. Many of our customers are between 41 and 60 years old, which is an important group for us to focus on. Also, a lot of our customers have completed a college degree, showing that education is important in how they shop. Plus, since 65% of our customers are married, we should consider family-related factors in our marketing.

Income distribution within our customer base presents a fascinating picture. While the majority of our customers earn between 0-100k$, it's noteworthy that a significant proportion boasts incomes exceeding 600k$. This finding suggests the existence of diverse economic backgrounds among our customers, presenting both challenges and opportunities in tailoring our products and marketing strategies.

The relationship between education levels and spending behavior cannot be overstated. Our exploratory Data Analysis elucidated that customers with graduate and postgraduate degrees exhibit substantially higher spending tendencies compared to their undergraduate counterparts. This distinction opens up avenues for targeted promotions, product offerings, and loyalty programs to incentivize higher education customers.

5.2 RECOMMENDATION

a) **Granular Marketing Segmentation**: To fully leverage our insights, we recommend implementing granular marketing segmentation. This approach involves dividing our customer base into smaller, more specific groups based on age, education, and income, among other variables.

b) **Personalization**: With a wealth of data at our disposal, personalized marketing should be at the forefront of our strategy. Utilize customer data to create individualized offers, content, and recommendations that cater to their preferences and spending habits.

c) **Customer Journey Mapping**: Understanding the customer journey is crucial. Map out the various touchpoints and interactions customers have with our brand, and optimize each stage to enhance customer satisfaction and loyalty.

d) **Feedback Mechanisms**: Establish robust feedback mechanisms to continually gauge customer satisfaction and adapt strategies accordingly. This includes soliciting feedback through surveys, monitoring social media, and analyzing customer reviews.

REFERENCES

David, R., et al. (2016). Dickson, Peter R.; Ginter, James L., "Market Segmentation, Product Differentiation, and Marketing Strategy, " *Journal of Marketing,* Vol. 51, No. 2, 1987, p. 1

Driver, H. E., & Kroeber, A. L. (1932). Quantitative Expression of Cultural Relationships. University of California Publications in American Archaeology and Ethnology, 211–256.

Estivill-Castro, V. (2002). Why So Many Clustering Algorithms – A Position Paper. ACM SIGKDD Explorations Newsletter, 4(1), 65–75. doi:10.1145/568574.568575.

Everitt, B. (2011). Cluster Analysis. Chichester, West Sussex, U.K: Wiley. ISBN 9780470749913.

Pantea Keikhosrokiani, Z. X. (n.d.). *An RFM Model Using K-Means Clustering to Improve Customer Segmentation and Product Recommendation*. Https://www.Igi-Global.com/. https://www.igi-global.com/chapter/an-rfm-model-using-k-means-clustering-to-improve-customer-segmentation-and-product-recommendation/305699

Lloyd, S. (1982). Least Squares Quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137. doi:10.1109/TIT.1982.1056489.

Microsoft Academic Search: Most Cited Data Mining Articles. (2010, April 18). (Note: DBSCAN is ranked 24th in the list)

(n.d.). *How to Perform Customer Segmentation in Python – Machine Learning Tutorial*. Freecodecamp. https://www.freecodecamp.org/news/customer-segmentation-python-machine-learning/

(n.d.). *Implementing Customer Segmentation Using Machine Learning*. Neptune ai. https://neptune.ai/blog/customer-segmentation-using-machine-learning

(n.d.). *Understanding K – Means Clustering WIth Customer Segmentation Usecase*. Www.Analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2021/07/understanding-k-means-clustering-using-customer-segmentation/

(n.d.). *What is Customer Segmentation*. The cleverprogrammer. https://thecleverprogrammer.com/2023/06/08/what-is-customer-segmentation/

Sibson, R. (1973). SLINK: An Optimally Efficient Algorithm for the Single-link Cluster Method. The Computer Journal, 16(1), 30–34. doi:10.1093/comjnl/16.1.30.

Tabianan, K., Velu, S., Ravi, V., Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Xu, G. (2016).

Tryon, R. C. (1939). Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.

Zubin, J. (1938). A Technique for Measuring Like-mindedness.The Journal of Abnormal and Social Psychology, 33(4), 508–516. doi:10.1037/h0055441.

Biswal, A. (n.d.). What is Exploratory Data Analysis? Steps and Market Analysis. Https://www.Simplilearn.com/. https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

Sharma , P. (n.d.). *The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications*. Https://www.Analyticsvidhya.com/. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#What_Is_K-Means_Clustering?

Pykes, K. (n.d.). *Introduction to Unsupervised Learning*. Https://www.Datacamp.com/. https://www.datacamp.com/blog/introduction-to-unsupervised-learning