CUSTOMER SEGMENTATION
USING
MACHINE LEARNING

INDUSTRIAL ORIENTED MINI PROJECT REPORT

Submitted in partial fulfilment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

NAME	ROLL NUMBER
Kamtham Keerthi Reddy	22XW1A0525
Pinnam Sathwik Kumar	22XW1A0544
Mettu Sowjanya	22XW1A0537
Lodugu Ravi Kumar	23XW5A0507

Under the guidance of

Dr.CH. ASHA JYOTHI

Associate Professor
Department of Computer Science and Engineering

Department of Computer Science and Engineering

Jawaharlal Nehru Technological University Hyderabad

University College of Engineering Wanaparthy

Narsingaipalli, Gopalpet Road Wanaparthy Dist.- 509 103

2024 – 2025

ABSTRACT

Compelling choices are compulsory for any organization to produce good revenue. Nowadays contest is huge and all organizations are moving forward with their own different strategies. We ought to utilize information and take an appropriate choice.

Each individual is different from the other and we don't know what he/she purchases or what their likes are.

However, with the assistance of the machine learning method, one can sort out the information and can find the target group by applying a few algorithms to the dataset. Without this, It will be very troublesome and no better techniques are accessible to find the gathering of people with comparable person and interests in an enormous dataset.

Here, The customer segmentation utilizing K-Means clustering assists with gathering the information with the same ascribes which precisely helps to business the best.

We are going to use the elbow technique to track down the number of clusters and finally, we visualize the data.

TABLE OF CONTENTS

List of Figures

CHAPTER 1 - INTRODUCTION

1.1 Introduction

Fig 1.1.1 - Customer Segmentation

As more and more business being coming up everyday, it has become significantly important for the old businesses to apply marketing strategies to stay in the market as the competition has been cut to throat . Change or die have become the simple rule of marketing in today’ s world. As the customer base is increasing day by day it has become challenging for the companies to cater to the needs of each and every customer, this is where Data mining serves a very important role to unravel hidden patterns stored in the company’s database.

Customer segmentation is one of the application of data mining which helps to segment the customers with similar patterns into similar clusters hence, making easier for the business to handle the large customer base. This segmentation can directly or indirectly influence the marketing strategy as it opens many new paths to discover like for which segment the product will be good, customising the marketing plans according to the each segment, providing discounts for a specific segment, and decipher the customer and object relationship which has been previously unknown to the company.

Customer segmentation allows companies to visualise what actually the customers are buying which will prompt the companies to better serve their customers resulting in customer satisfaction, it also allows the companies to find who their target customers are and improvise their marketing tactics to generate more revenues from them. Clustering has been proven effective to implement customer segmentation. Clustering comes under unsupervised learning, having ability to find clusters over unlabelled dataset.

K-means Clustering:

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.

Customer Segmentation mostly is the course of division of the client base into a generally few gatherings called client sections to pretty such an extent that every client fragment comprises clients who mostly have fairly comparative qualities, showing how throughout the definitely long term, the opposition among organizations particularly is expanded and the enormous verifiable information that basically is accessible essentially has brought about the inescapable utilization of information mining methods in removing the significant and vital data from the data set of the association, actually contrary to popular belief.

The division depends on the similitude in various ways that really are pertinent to promoting like orientation, age, interests, and incidental ways of managing money, which actually is fairly significant. The client division generally has the significance as it incorporates, the capacity to basically alter the projects of the market with the kind of goal that it literally is kind of appropriate to every one of the client portion, support in business choice; ID of items related with every client portion and mostly manage the interest and supply of that item; distinguishing and focusing on the pretty potential client base, and foreseeing client surrender, giving headings in viewing as the arrangements, definitely contrary to popular belief.

The push of this paper kind of is to really recognize client sections utilizing the information mining approach, utilizing the dividing calculation called as K-means grouping calculation, which mostly shows that customer Segmentation particularly is the course of division of the client base into a actually few gatherings called client sections to definitely such an extent that every client fragment comprises clients who particularly have really comparative qualities, showing how throughout the really long term, the opposition among organizations actually is expanded and the enormous verifiable information that actually is accessible essentially has brought about the inescapable utilization of information mining methods in removing the significant and vital data from the data set of the association in a particularly big way.

The elbow strategy decides the fairly ideal groups, which specifically shows that as indicated, Bunching strategies particularly consider information tuples as items. They segment the information objects into gatherings or groups so that items inside a bunch essentially are like one another and unlike articles in different groups, which basically is quite significant.

1.2 Aim of the project

In today’s highly competitive market, businesses struggle to effectively target their customers due to a lack of understanding of diverse customer behaviors and preferences. A one-size-fits-all approach to marketing often leads to suboptimal engagement and poor resource allocation. To address this issue, there is a need for a data-driven solution that can segment customers into distinct groups based on their purchasing patterns, demographics, and behavioral data.

1.3 Existing system

In the traditional business environment, customer segmentation is often performed using manual or rule-based methods. These methods involve segmenting customers based on basic demographic attributes such as age, gender, location, or income level. While these attributes offer some insights, they are often too broad and fail to capture the nuanced behavioral patterns and preferences that exist within a customer base.

One of the most common existing approaches involves using predefined rules or static categories for segmentation. For instance, marketing teams might classify customers into "young adults," "middle-aged," or "seniors" based on age brackets. Alternatively, they may segment customers based on historical purchase volume (e.g., high spenders vs. low spenders) or product category preferences. These methods are often implemented through spreadsheet tools or basic database queries.

While simple and easy to implement, such systems have several limitations:

1. Lack of Personalization: Rule-based segmentation often leads to broad categories, which do not cater to individual needs or preferences. As a result, marketing campaigns lack personalization and may not resonate well with targeted audiences.

2. Static Groupings: Traditional methods use static rules that do not adapt to changing customer behaviour’s. For instance, a customer who recently shifted their purchasing habits will still be grouped according to outdated data.

3. Scalability Issues: As customer data grows, manual segmentation becomes increasingly time-consuming and error-prone. Businesses with thousands or

millions of customers cannot rely on manual methods without sacrificing accuracy and efficiency.

4. No Predictive Insight: Traditional systems provide descriptive insights based on historical data but lack the ability to predict future customer behavior, such as churn likelihood or potential product interests.

Due to these limitations, businesses often miss valuable insights that could drive revenue growth, customer loyalty, and marketing effectiveness. The absence of automation and adaptability in existing systems highlights the need for more intelligent, data-driven solutions.

1.4 Proposed system

The proposed system introduces a machine learning-based approach to customer segmentation that leverages advanced analytics and unsupervised learning techniques to identify meaningful patterns in customer data. By utilizing clustering algorithms such as K-Means, Hierarchical Clustering, or DBSCAN, this system can automatically group customers into distinct segments based on similarities in behavior, purchase history, demographics, and engagement levels.

Key features and benefits of the proposed system include:

1. **Data-Driven Segmentation**: Instead of relying on manual rules, the system analyzes a broad set of features—such as Recency, Frequency, and Monetary (RFM) values, product categories, website interaction data, and customer demographics—to discover hidden patterns and form more accurate clusters.

2. **Unsupervised Learning**: Clustering algorithms require no labelled data. This makes them ideal for exploratory tasks like segmentation, where the optimal number and nature of segments are not known in advance.

3. **Scalability and Automation**: The system can handle large datasets efficiently and adaptively, allowing businesses to scale their segmentation efforts as their customer base grows. The segmentation process can be automated and scheduled periodically, ensuring that customer groups reflect the latest data.

4. **Enhanced Marketing Precision**: With more refined and behaviorally driven segments, businesses can create targeted marketing strategies, personalize communication, and improve conversion rates. For example, one segment may respond better to email promotions, while another prefers social media offers.

5. **Dynamic Updates**: Unlike static rule-based systems, the proposed system supports periodic re-training and re-segmentation, allowing customer groups to evolve with changing behavior patterns.

6. **Visualization and Interpretability**: Tools like Principal Component Analysis (PCA) can be used to visualize high-dimensional clusters in 2D or 3D space, aiding business users in interpreting the segmentation results. Dashboards and reports can be integrated to support marketing decisions.

7. **Extensibility**: The system is designed to be modular, allowing the integration of supervised learning models in the future for use cases such as predicting customer churn, lifetime value, or upsell potential based on segment membership.

In summary, the proposed machine learning-based customer segmentation system provides a robust, scalable, and intelligent approach to understanding customer diversity. It enables businesses to go beyond basic demographics and unlock deeper behavioral insights, ultimately driving smarter decision-making and improved customer experiences.

1.5 Objectives

The main objective of this project is to implement an unsupervised machine learning approach to segment customers based on their purchasing behavior and demographic data. These segments can help businesses tailor marketing strategies, enhance product recommendations, and improve overall customer experience.

- ➢ Develop a low-cost product for analyzing customer trends
- ➢ Design an optimal distribution strategy.
- ➢ Choose specific product features for deployment.

1.6 Methodology

The customer segmentation project employs a structured and systematic approach to analyze customer data using machine learning techniques, particularly unsupervised learning. The methodology followed in this project is divided into several key stages: data collection, data preprocessing, exploratory data analysis, feature selection, model selection and implementation, cluster evaluation, and result interpretation.

1. Data Collection

The first step involves gathering a dataset containing customer-related information. This may include demographic data (age, gender, location), behavioral data (purchase frequency, recency, monetary value), and engagement data (web activity, response to campaigns). For this project, a publicly available customer dataset or organizational customer data (if available) is used as the basis for analysis.

2. Data Preprocessing

Raw data often contains inconsistencies, missing values, and irrelevant features. The following preprocessing steps are applied:

- **Handling Missing Values:** Missing entries are either removed or imputed using mean/median/mode techniques.
- **Data Cleaning:** Duplicate records and irrelevant columns are eliminated.
- **Data Transformation:** Categorical variables are encoded using techniques like one-hot encoding or label encoding.

- **Normalization:** Features are scaled using Min-Max normalization or StandardScaler to bring them to a common scale, which is essential for clustering algorithms.

3. Exploratory Data Analysis (EDA)

EDA is conducted to gain insights into the dataset and understand the relationships between different variables. Visualizations such as histograms, box plots, scatter plots, and heatmaps are used to identify patterns, correlations, outliers, and distributions. This step helps in selecting meaningful features for clustering.

4. Feature Engineering and Selection

Based on the understanding from EDA, key features are selected or derived. A common technique used in customer segmentation is the **RFM (Recency, Frequency, Monetary)** model, where:

- **Recency** refers to how recently a customer made a purchase.
- **Frequency** refers to how often they purchase.
- **Monetary** refers to how much money they spend.

These features are computed from transactional data and provide a strong foundation for clustering.

5. Model Selection and Clustering

Unsupervised learning is used to perform customer segmentation. The chosen algorithm is **K-Means Clustering**, due to its simplicity and effectiveness. The following steps are followed:

- **Elbow Method** is applied to determine the optimal number of clusters (k) by plotting the Within-Cluster-Sum-of-Squares (WCSS) and identifying the point of inflection.
- **K-Means Algorithm** is then applied with the selected value of k to segment the customers.

6. Evaluation of Clustering

Unlike supervised models, clustering does not have ground truth labels. Therefore, internal validation metrics are used to evaluate the clustering performance:

- **Cluster Visualization**: 2D or 3D plots (using PCA or t-SNE) are used to visually inspect the separation between clusters.

Tools and Technologies Used

- **Programming Language**: Python
- **Libraries**: pandas, NumPy, scikit-learn, matplotlib, seaborn
- **Clustering Algorithms**: K-Means, Hierarchical Clustering
- **Dimensionality Reduction**: PCA
- **Environment**: Jupyter Notebook / Google Colab

CHAPTER 2 - REQUIREMENT AND ANALYSIS

The requirement and analysis phase is a critical component of the project, This phase ensures a clear understanding of project software and hardware requirements, and constraints, providing a solid foundation for successful project delivery.

2.1 Software Requirements

- Anaconda Navigator
- Jupyter
- Python 3.10 +
- Language: Python
- Operating System: Windows 10 and above

2.2 Hardware Requirements

- Processor - Processor core i3 and above
- Speed - 1.1 GHZ
- RAM – 4GB
- Hard disk - 500 GB
- Keyboard - Standard Keyboard
- Monitor - LED Monitor

2.3 Machine Learning

Machine Learning for customer segmentation

Machine learning methodologies are a great tool for analyzing customer data and finding insights and patterns. Artificially intelligent models are powerful tools for decision-makers. They can precisely identify customer segments, which is much harder to do manually or with conventional analytical methods.

There are many machine learning algorithms, each suitable for a specific type of problem. One very common machine learning algorithm that’s suitable for customer segmentation problems is the k-means clustering algorithm. There are other clustering algorithms as well such as DBSCAN, Agglomerative Clustering, and BIRCH, etc.

Why would you implement machine learning for customer segmentation?

Manual customer segmentation is time-consuming. It takes months, even years to analyze piles of data and find patterns manually. Also if done heuristically, it may not have the accuracy to be useful as expected.

Customer segmentation used to be done manually and wasn’t too precise. You’d manually create and populating different data tables, and analyze the data like a detective with a looking glass. Now, it’s much better (and relatively easy thanks to rapid progress in ML) to just use machine learning, which can free up your time to focus on more demanding problems that require creativity to solve.

Ease of retraining

Customer Segmentation is not a “develop once and use forever” type of project. Data is ever-changing, trends oscillate, everything keeps changing after your model is deployed. Usually, more labeled data becomes available after development, and it’s a great resource for improving the overall performance of your model.

There are many ways to update customer segmentation models, but here are the two main approaches:

- Use the old model as the starting point and retrain it.
- Keep the existing model and combine its output with a new model.

Better scaling

Machine learning models deployed in production support scalability, thanks to cloud infrastructure. These models are quite flexible for future changes and feedback. For example, consider a company that has 10000 customers today, and they’ve implemented a customer segmentation model. After a year, if the company has 1 million customers, then ideally we don’t need to create a separate project to handle this increased data. Machine Learning models have the inherent capability to handle more data and scale in production.

Higher accuracy

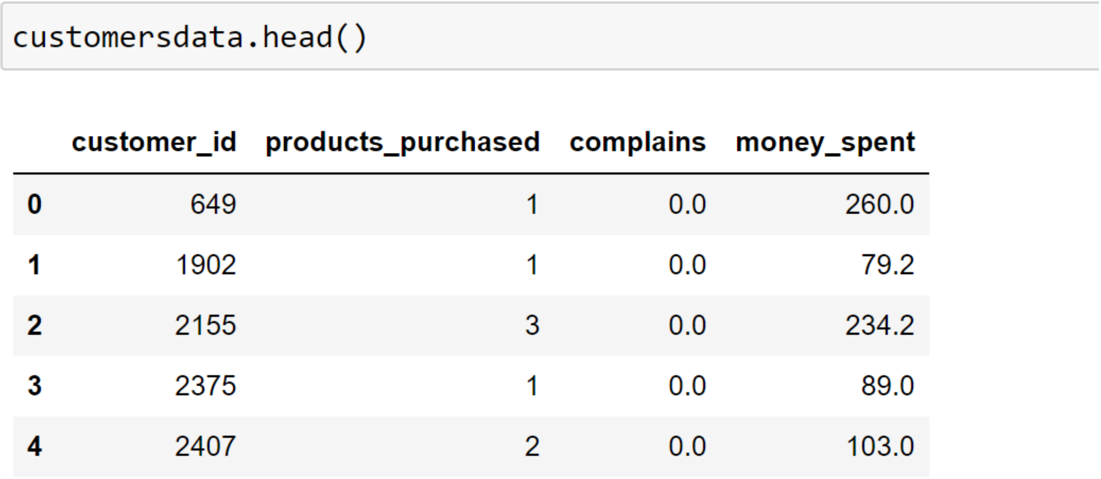
The value of an optimal number of clusters for given customer data is easy to find using machine learning methods like the elbow method. Not only the optimal number of clusters but also the performance of the model is far better when we use machine learning.

2.4 Dataset

Exploring customer dataset and its features

Let’s analyze a customer dataset. Our dataset has 24,000 data points and four features. The features are:

- Customer ID – This is the id of a customer for a particular business.
- Products Purchased – This feature represents the number of products purchased by a customer in a year.
- Complaints – This column value indicates the number of complaints made by the customer in the last year
- Money Spent – This column value indicates the amount of money paid by the customer over the last year.



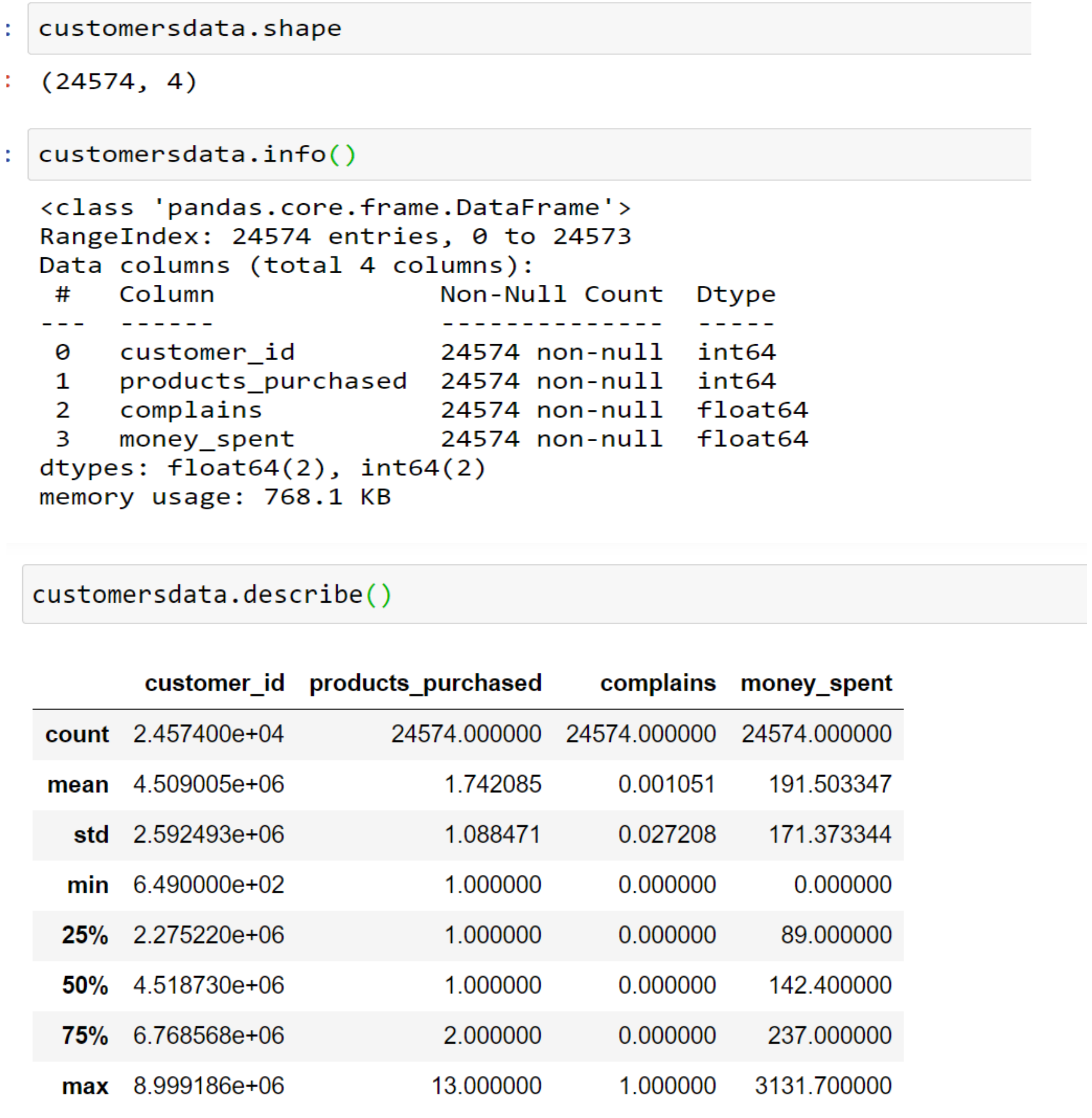
```
customersdata.head()
```

	customer_id	products_purchased	complains	money_spent
0	649	1	0.0	260.0
1	1902	1	0.0	79.2
2	2155	3	0.0	234.2
3	2375	1	0.0	89.0
4	2407	2	0.0	103.0

Fig 2.4.1 – Customers data head

Pre-processing the dataset

Before feeding the data to the k-means clustering algorithm, we need to pre-process the dataset. Let's implement the necessary pre-processing for the customer dataset.



```
:  customersdata.shape
:  (24574, 4)
```

```
:  customersdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24574 entries, 0 to 24573
Data columns (total 4 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   customer_id        24574 non-null  int64  
 1   products_purchased 24574 non-null  int64  
 2   complains          24574 non-null  float64
 3   money_spent        24574 non-null  float64
dtypes: float64(2), int64(2)
memory usage: 768.1 KB
```

```
customersdata.describe()
```

	customer_id	products_purchased	complains	money_spent
count	2.457400e+04	24574.000000	24574.000000	24574.000000
mean	4.509005e+06	1.742085	0.001051	191.503347
std	2.592493e+06	1.088471	0.027208	171.373344
min	6.490000e+02	1.000000	0.000000	0.000000
25%	2.275220e+06	1.000000	0.000000	89.000000
50%	4.518730e+06	1.000000	0.000000	142.400000
75%	6.768568e+06	2.000000	0.000000	237.000000
max	8.999186e+06	13.000000	1.000000	3131.700000

Fig 2.4.2 – Pre-processing dataset

2.5 Libraries

1. Pandas

- **Purpose**: A powerful data analysis and manipulation library.
- **Key Features**:
    - Provides data structures like **DataFrame** and **Series**.
    - Useful for reading/writing data (CSV, Excel, SQL, etc). Supports data cleaning, filtering, aggregation, and reshaping.
- **Use Case**: Organizing and processing structured data in tabular form.

2. NumPy

- **Purpose**: Fundamental library for numerical and scientific computing in Python.
- **Key Features**:
    - Provides efficient array operations with ndarray.
    - Supports linear algebra, random number generation, and basic math functions.
- **Use Case**: Performing vectorized operations and mathematical computations on large datasets.

3. Scikit-learn (sklearn)

- **Purpose**: A machine learning library offering simple and efficient tools.
- **Key Features**:
    - Implements machine learning algorithms for classification, regression, clustering, etc.
    - Includes data preprocessing and model evaluation tools.
- **Use Case**: Here, KMeans is used for unsupervised clustering of data points into groups.

4. Plotly

- **Purpose**: Interactive graphing and visualization library.
- **Key Features**:
    - Allows creation of interactive, web-based plots.
    - plotly.express offers a quick way to create simple charts.
    - plotly.graph_objects allows fine-grained customization of complex visualizations.
- **Use Case**: Visualizing clusters, data distribution, and trends in an interactive format.

5. Matplotlib

- **Purpose**: A 2D plotting library for creating static, animated, and interactive visualizations.
- **Key Features**:
  - Compatible with NumPy and Pandas data structures.
  - Provides extensive control over plot appearance.
- **Use Case**: Often used for basic or publication-quality plots and charts.

Chapter 3 - System Design

The system design phase focuses on defining the architecture, components, and data flow of the system, ensuring a robust and scalable solution that meets the requirements.

3.1 Data Flow Diagram (DFD)

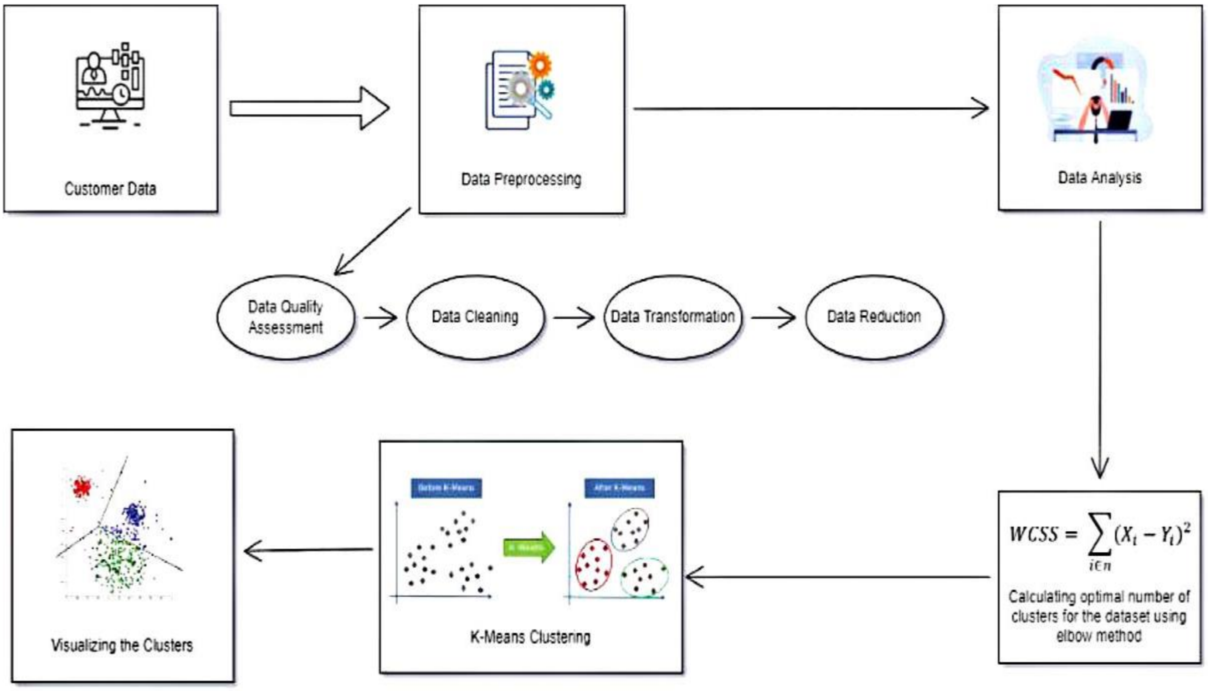


Fig 3.1.1 – Data flow diagram

3.2 Entity Relationship Diagram (ER – Diagram)

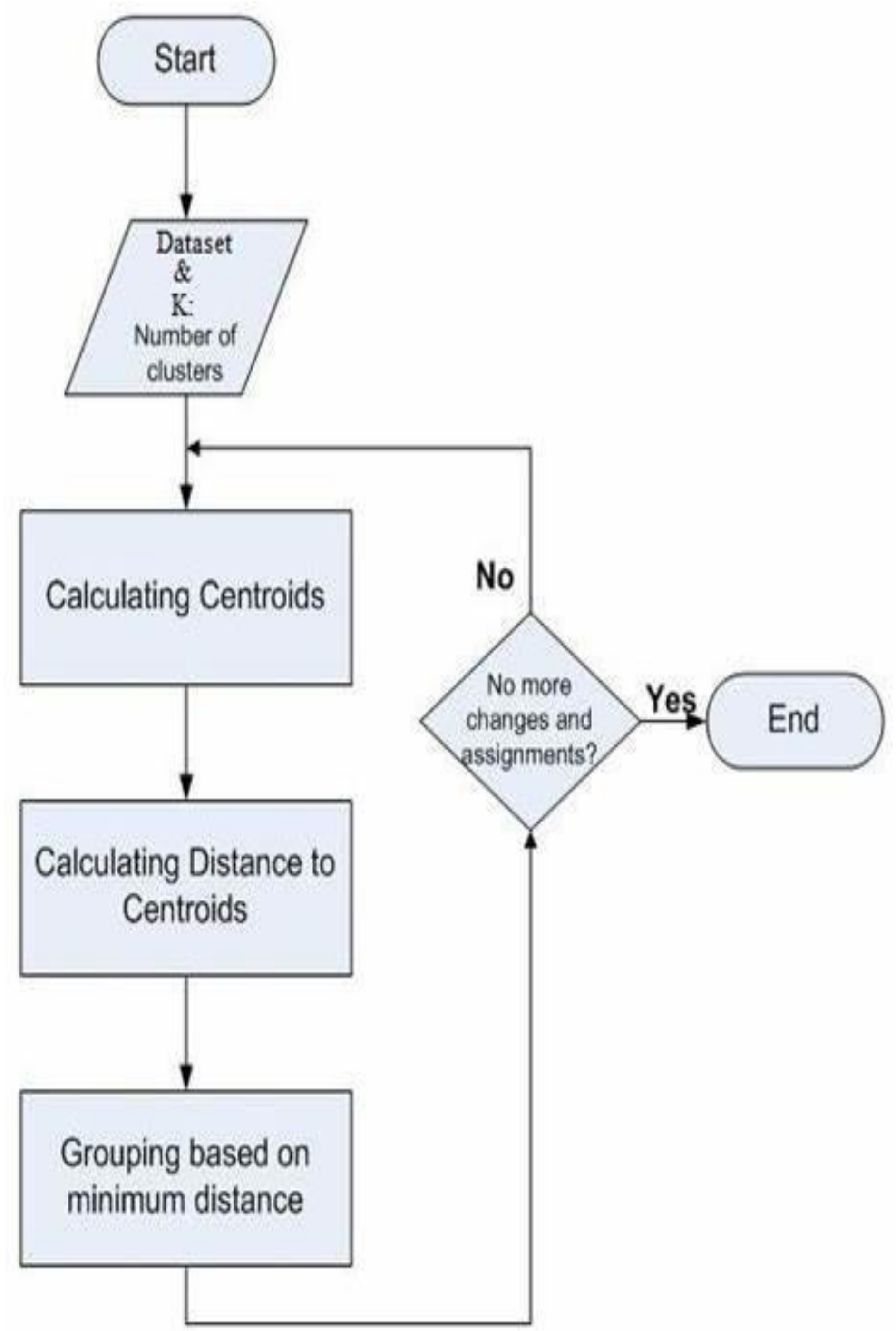

Start
Dataset
&
K:
Number of
clusters
Calculating Centroids
Calculating Distance to Centroids
Grouping based on minimum distance
No
No more changes and assignments?
Yes
End

Fig 3.1.2 – Entity Relationship Diagram

3.3 Algorithms Used

K-means Clustering:

It is the simplest algorithm of clustering based on partitioning principle. The algorithm is sensitive to the initialization of the centroids position, the number of K (centroids) is calculated by elbow method (discussed in later section), after calculation of K centroids by the terms of Euclidean distance data points are assigned to the closest centroid forming the cluster, after the cluster formation the barycentre's are once again calculated by the means of the cluster and this process is repeated until there is no change in centroid position.

Choosing the optimal number of clusters:

Elbow method is applied to calculate value of K for the dataset.

Step-1: Run the algorithm for various values of k i.e making the k vary from 1 to 10.

Step-2: Calculate the within cluster squared error.

Step-3: Plot the calculated error, where a bent elbow like structure will form, will give the optimal value of clusters.

SSE is calculated by -:

$$\sum_{i=1}^{k} \sum_{Xj \in Si} ||Xj - \mu i||^2$$

Xj = data point in Si cluster

μ i = centroid of the cluster

Algorithm:

Step-1: Initialize the K (= 5) clusters.

Step-2: Assign the data point that is closest to any particular cluster.

Step-3: Recalculate the centroid position based on the mean of the cluster formed

Step-4: Repeat step 2 and 3 until the centroid position remains unchanged in the previous and current iteration.

Elbow Method :

Elbow method is used for finding optimal value of K for K-means clustering algorithm. This method works by finding the SSE of each data point with its nearest centroid with different values of K. As value of K increases the SSE will decrease and at a particular value of K where there is most decline in the SSE is the elbow, the point at which we should stop dividing data further.

The Elbow method is a widely used technique in clustering analysis, particularly in customer segmentation projects. It helps determine the optimal number of clusters (K) for grouping customers based on their characteristics, behaviors, or preferences.

How the Elbow Method Works

1. Calculate Sum of Squared Errors (SSE): For each value of K, calculate the SSE or distortion score, which measures the difference between each data point and its assigned cluster center.

2. Plot SSE against K: Visualize the SSE values against different K values, typically ranging from 1 to 10.

3. Identify the Elbow Point: Look for the point on the plot where the rate of decrease in SSE becomes less steep, forming an "elbow" shape. This point indicates the optimal number of clusters.

Importance in Customer Segmentation

The Elbow method is essential in customer segmentation because it helps:

1. Identify Distinct Customer Groups: By determining the optimal number of clusters, businesses can identify distinct customer segments with unique characteristics, needs, and preferences.

2. Improve Targeting and Personalization: With well-defined customer segments, businesses can tailor their marketing strategies, product offerings, and customer experiences to meet the specific needs of each segment.

3. Enhance Customer Satisfaction and Loyalty: By understanding the unique needs and preferences of each customer segment, businesses can deliver more effective and personalized experiences, leading to increased customer satisfaction and loyalty.

Best Practices

To get the most out of the Elbow method in customer segmentation:

1. Use Relevant Data: Ensure that the data used for clustering is relevant to the business goals and customer characteristics.

2. Experiment with Different K Values: Try different K values and evaluate the resulting clusters to ensure the optimal number of clusters is chosen.

3. Combine with Other Methods: Consider combining the Elbow method with other clustering evaluation techniques, such as silhouette analysis or gap statistic, to validate the results.

Chapter 4 - Performance Analysis

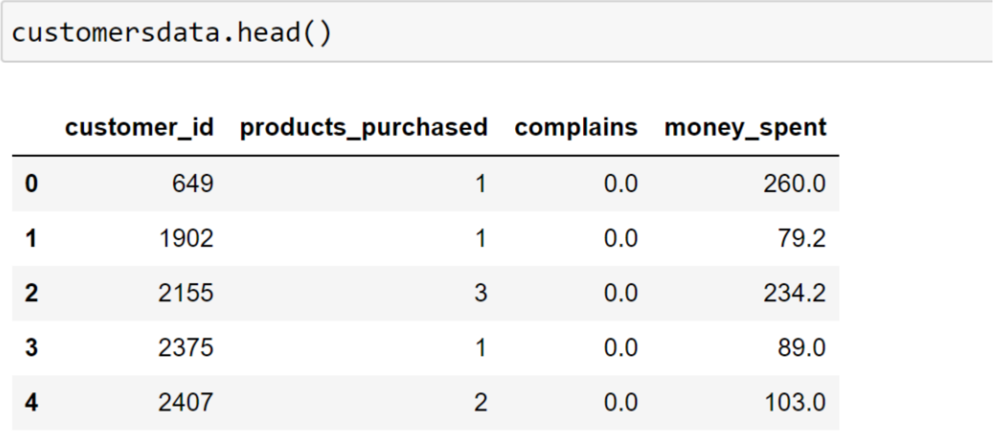
This section evaluates the system's performance, identifying bottlenecks and optimizing code for efficiency, scalability, and reliability to ensure a seamless user experience.

4.1 Source Code

```
# Import required libraries

import pandas as pd

import numpy as np

from sklearn.cluster import KMeans

import plotly.express as px

import plotly.graph_objects as go

import matplotlib.pyplot as plt

#Load customers data

customersdata = pd.read_csv("customers-data.csv")

# Define K-means model

kmeans_model = KMeans(init='k-means++',  max_iter=400, random_state=42)

# Train the model

kmeans_model.fit(customersdata[['products_purchased','complains','money_spent']])

# Create the K means model for different values of K

def try_different_clusters(K, data):

    cluster_values = list(range(1, K+1))

    inertias=[]

    for c in cluster_values:

        model= KMeans(n_clusters = c,init='k-means++',max_iter=400,random_state=42)

        model.fit(data)
        inertias.append(model.inertia_)
    return inertias
# Create the K means model for different values of K
def try_different_clusters(K, data):
    cluster_values = list(range(1, K+1))
    inertias=[]
    for c in cluster_values:
        model= KMeans(n_clusters = c,init='k-means++',max_iter=400,random_state=42)
        model.fit(data)
        inertias.append(model.inertia_)
    return inertias
# Find output for k values between 1 to 12
outputs= try_different_clusters(12,
customersdata[['products_purchased','complains','money_spent']])
distances = pd.DataFrame({"clusters": list(range(1, 13)),"sum of squared distances":
outputs})
# Finding optimal number of clusters k
figure = go.Figure()
figure.add_trace(go.Scatter(x=distances["clusters"],y=distances["sum of squared
distances"]))
figure.update_layout(xaxis = dict(tick0 = 1,dtick = 1,tickmode = 'linear'),
                  xaxis_title="Number of clusters",
                  yaxis_title="Sum of squared distances",
                  title_text="Finding optimal number of clusters using elbow method")
figure.show()
```

Chapter 5 - Result

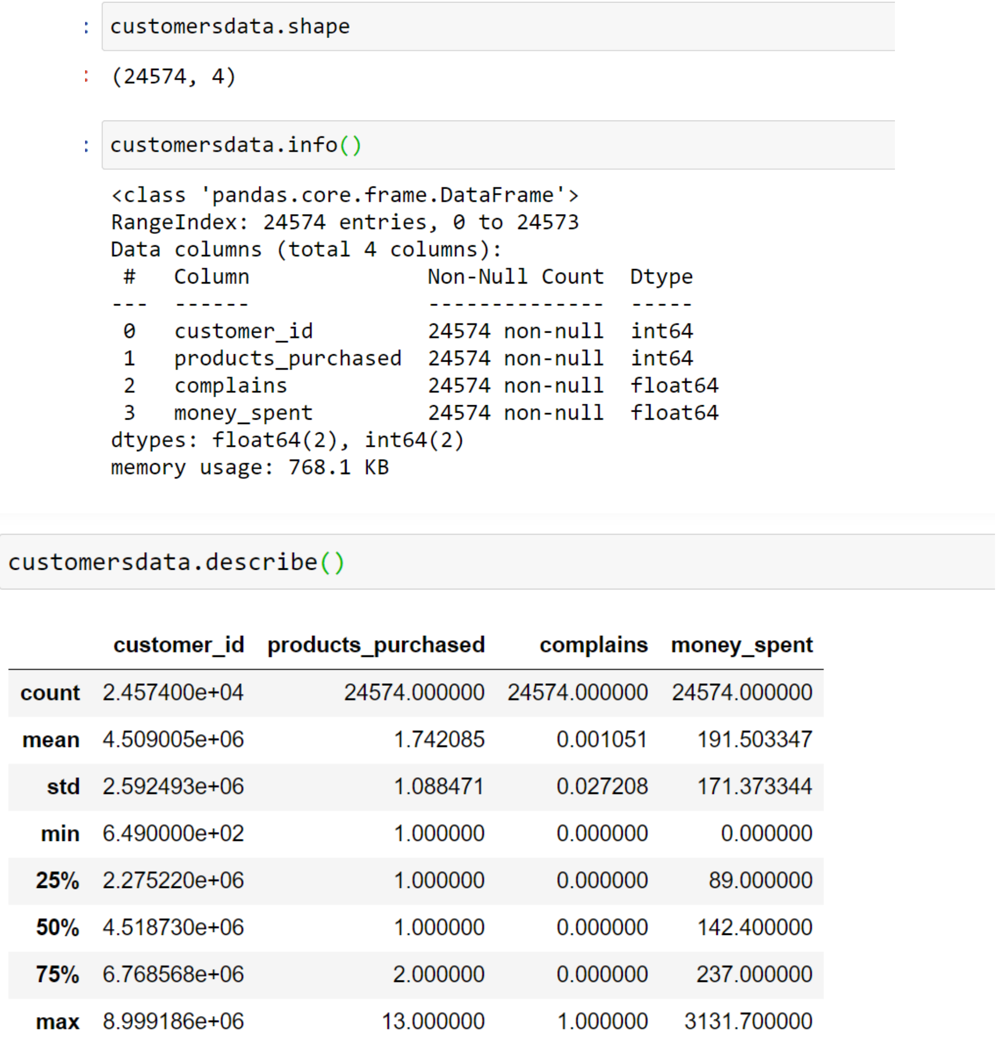
5.1 Output Screens



```
customersdata.head()
```

	customer_id	products_purchased	complains	money_spent
0	649	1	0.0	260.0
1	1902	1	0.0	79.2
2	2155	3	0.0	234.2
3	2375	1	0.0	89.0
4	2407	2	0.0	103.0

Fig 5.1.1 – Customers data



```
: customersdata.shape
```

```
: (24574, 4)
```

```
: customersdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 24574 entries, 0 to 24573
Data columns (total 4 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   customer_id         24574 non-null  int64  
 1   products_purchased  24574 non-null  int64  
 2   complains           24574 non-null  float64
 3   money_spent         24574 non-null  float64
dtypes: float64(2), int64(2)
memory usage: 768.1 KB
```

```
customersdata.describe()
```

	customer_id	products_purchased	complains	money_spent
count	2.457400e+04	24574.000000	24574.000000	24574.000000
mean	4.509005e+06	1.742085	0.001051	191.503347
std	2.592493e+06	1.088471	0.027208	171.373344
min	6.490000e+02	1.000000	0.000000	0.000000
25%	2.275220e+06	1.000000	0.000000	89.000000
50%	4.518730e+06	1.000000	0.000000	142.400000
75%	6.768568e+06	2.000000	0.000000	237.000000
max	8.999186e+06	13.000000	1.000000	3131.700000

Fig 5.1.2 - Pre-processing the dataset

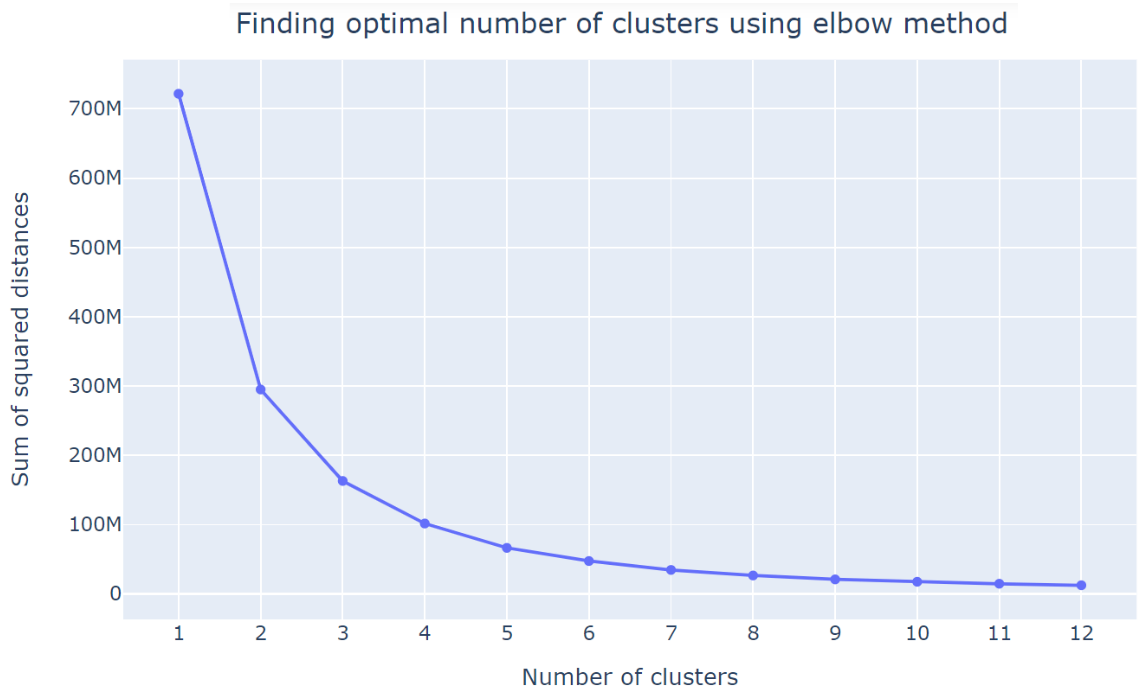
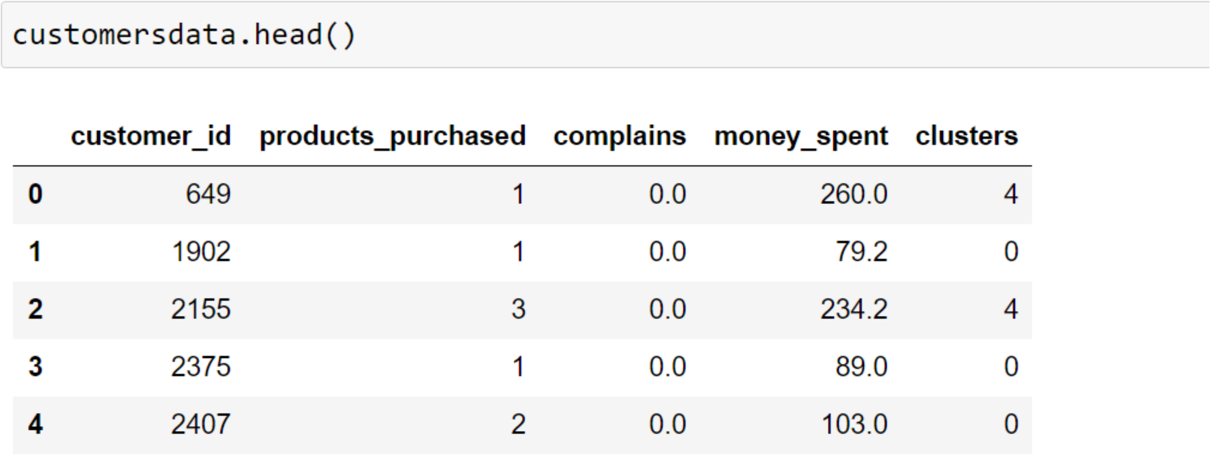


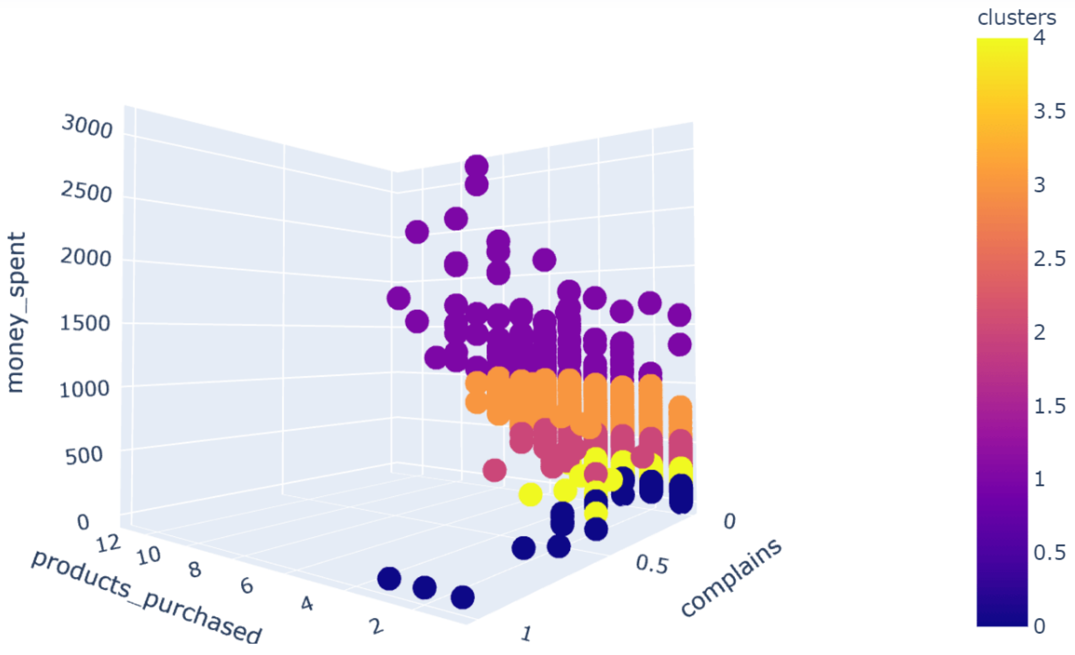
Fig 5.1.3 - Finding optimal number of clusters k



```
customersdata.head()
```

	customer_id	products_purchased	complains	money_spent	clusters
0	649	1	0.0	260.0	4
1	1902	1	0.0	79.2	0
2	2155	3	0.0	234.2	4
3	2375	1	0.0	89.0	0
4	2407	2	0.0	103.0	0

Fig 5.1.4 - After adding the new column, named clusters



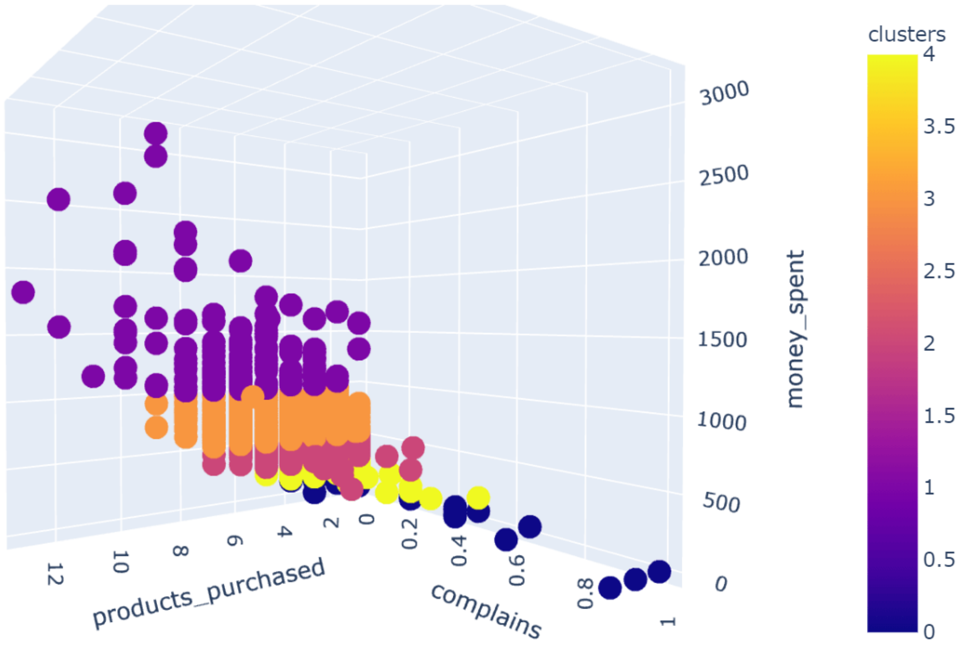


Fig 5.1.5 – Final visualizations

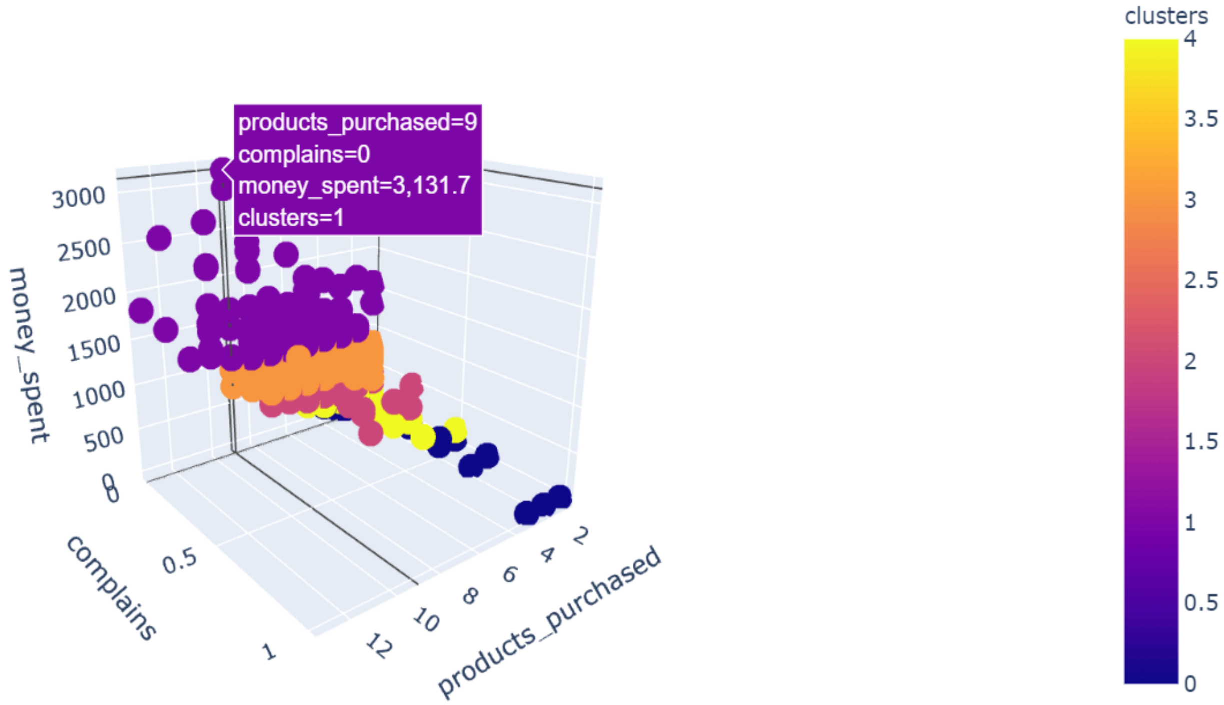


Fig 5.1.6 – Final visualizations

Chapter 6 – Conclusion

It’s not wise to serve all customers with the same product model, email, text message campaign, or ad. Customers have different needs. A one-size-for-all approach to business will generally result in less engagement, lower-click through rates, and ultimately fewer sales. Customer segmentation is the cure for this problem.

Finding an optimal number of unique customer groups will help you understand how your customers differ, and help you give them exactly what they want. Customer segmentation improves customer experience and boosts company revenue.

That’s why segmentation is a must if you want to surpass your competitors and get more customers. Doing it with machine learning is definitely the right way to go.

References

[1] Implementing Customer Segmentation Using Machine Learning [Beginners Guide]

[2] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu Electrical/Electronics & Computer Engineering Department, University of Uyo, Uyo, Akwa Ibom State, Nigeria “Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services”, IJARAI,Year: 2015.

[3] T.NelsonGnanarajDr.K.Ramesh Kumar N.Monica“Survey on mining clusters using new k-mean algorithm from structured and unstructured data”, IJACST,Year: 2014.

[4] Implementing Customer Segmentation Using Machine Learning

[5] https://ijcrt.org/papers/IJCRT_196650.pdf