Customer Segmentation

Project submitted to the

SRM University – AP, Andhra Pradesh

for the course project of

CSE338 Applied Data Science

Submitted by

MOHAN KRISHNA KOMATI(AP20110010433)

KAVYA BODAPATI(AP20110010164)

JASMITHA PUSULURI(AP20110010473)

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[Month, Year]

Table of Contents

1.Abstract

In these projects, I will be performing an unsupervised clustering data on the customer's records from a supermarket database. Customer segmentation is the practise of separating customers into groups that reflect similarities among customers in each cluster. This abstract focuses on customer segmentation using machine learning algorithms and highlights its significance in enabling businesses to deliver personalized experiences, improve customer targeting, and maximize marketing effectiveness. It explores various machine learning techniques commonly used for customer segmentation, including clustering algorithms, dimensionality reduction, and data pre-processing and data cleaning.

Customer segmentation in machine learning presents significant opportunities for businesses to gain a deeper understanding of their customer base and tailor their marketing strategies accordingly. By leveraging advanced algorithms and techniques, companies can unlock valuable insights from their data and deliver personalized experiences that drive customer satisfaction and loyalty. However, it is crucial to approach customer segmentation with careful consideration of ethical implications and continuous evaluation to ensure its effectiveness in a dynamic business environment. this abstract addresses the ethical considerations surrounding the use of machine learning in customer segmentation, such as data privacy, transparency, and potential biases. It highlights the importance of responsible and ethical implementation of machine learning techniques to ensure fairness and inclusivity in customer segmentation practices.

2.Methodology

2.1 **Data collection and preparation**

Identify the relevant data sources: Determine which data sources are available and can provide insights into customer behavior and characteristics. This may include sources such as customer relationship management (CRM) systems, transactional databases, website analytics, surveys, social media, or third-party data providers.

Gather comprehensive data: Collect a wide range of customer data points to capture various aspects of customer behavior and preferences. This may include demographic information (age, gender, location), transactional data (purchase history, order value), behavioral data (website visits, clicks, time spent), and any other relevant data that helps differentiate customer segments.

Feature engineering: Identify and extract relevant features that are likely to influence customer behavior and segmentation. This may involve creating new variables or combining existing ones to capture meaningful insights. For example, calculating average order value, frequency of purchases, or recency of interactions.

Normalize or standardize variables: Depending on the nature and scale of the variables, it may be necessary to normalize or standardize them to ensure they have a similar range or distribution. This step helps prevent certain variables from dominating the segmentation process due to their scale or magnitude.

2.2 **Feature Selection**:

Identify a set of features that are most relevant for customer segmentation.

Consider variables such as age, gender, location, purchase frequency, total spending, and any other variables that can differentiate customer behaviors or preferences.

Domain Knowledge: Start by leveraging your domain expertise and understanding of your business objectives to identify the key factors that may influence customer behavior and segmentation. Consider variables that are known to be relevant in your industry or have been found to be significant in previous research or studies.

Univariate Analysis: Conduct a univariate analysis by examining the relationship between each individual variable and the target variable (e.g., customer segment). This can be done using statistical tests such as t-tests, chi-square tests, or analysis of variance (ANOVA). Select variables that show a significant association with the target variable.

2.3 **Dimensionality Reduction:**

Apply dimensionality reduction techniques to reduce the number of features while retaining meaningful information.

Common techniques include Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE).

This step helps in eliminating redundant or less informative features, making subsequent clustering more effective.

Principal Component Analysis (PCA):

PCA is a widely used dimensionality reduction technique that identifies the directions, called principal components, along which the data varies the most.

It transforms the original variables into a new set of orthogonal variables, known as principal components, which are linear combinations of the original variables.

t-Distributed Stochastic Neighbor Embedding (t-SNE):

t-SNE is a nonlinear dimensionality reduction technique that focuses on preserving the local structure and relationships between data points.It is particularly effective for visualizing high-dimensional data in lower-dimensional spacest-SNE maps the original data points to a lower-dimensional space such that similar points in the original space are modeled as nearby points in the lower-dimensional space.

Linear Discriminant Analysis (LDA):

LDA is a dimensionality reduction technique that is specifically designed for supervised tasks such as classification or customer segmentation.

2.4 Clustering:

Apply clustering algorithms to group customers with similar characteristics into segments.

Popular clustering algorithms include k-means, hierarchical clustering, or density-based clustering (e.g., DBSCAN).

Experiment with different cluster numbers and evaluate the results using metrics such as silhouette score or within-cluster sum of squares (WCSS).

2.5 Interpretation and Profiling:

Analyze the resulting clusters to understand the characteristics and behaviors of each segment.

Examine the cluster centroids or representative samples to identify key features that distinguish one segment from another.

Create customer profiles for each segment, describing their characteristics, preferences, and behaviors.

2.6 Evaluation and Refinement:

Assess the quality and stability of the obtained clusters by conducting internal or external validation measures.

Repeat the clustering process with alternative algorithms or parameter settings to compare and refine the segmentation.

Consider conducting A/B testing or tracking business metrics to evaluate the effectiveness of the segmentation on marketing campaigns or customer engagement.

2.7 Application and Personalization:

Utilize the customer segments for targeted marketing campaigns, product recommendations, or tailored experiences.

Develop personalized strategies for each segment based on their unique needs and preferences.

Continuously monitor and adapt the segmentation as customer behavior evolves or new data becomes available.

2.8 Ethical Considerations:

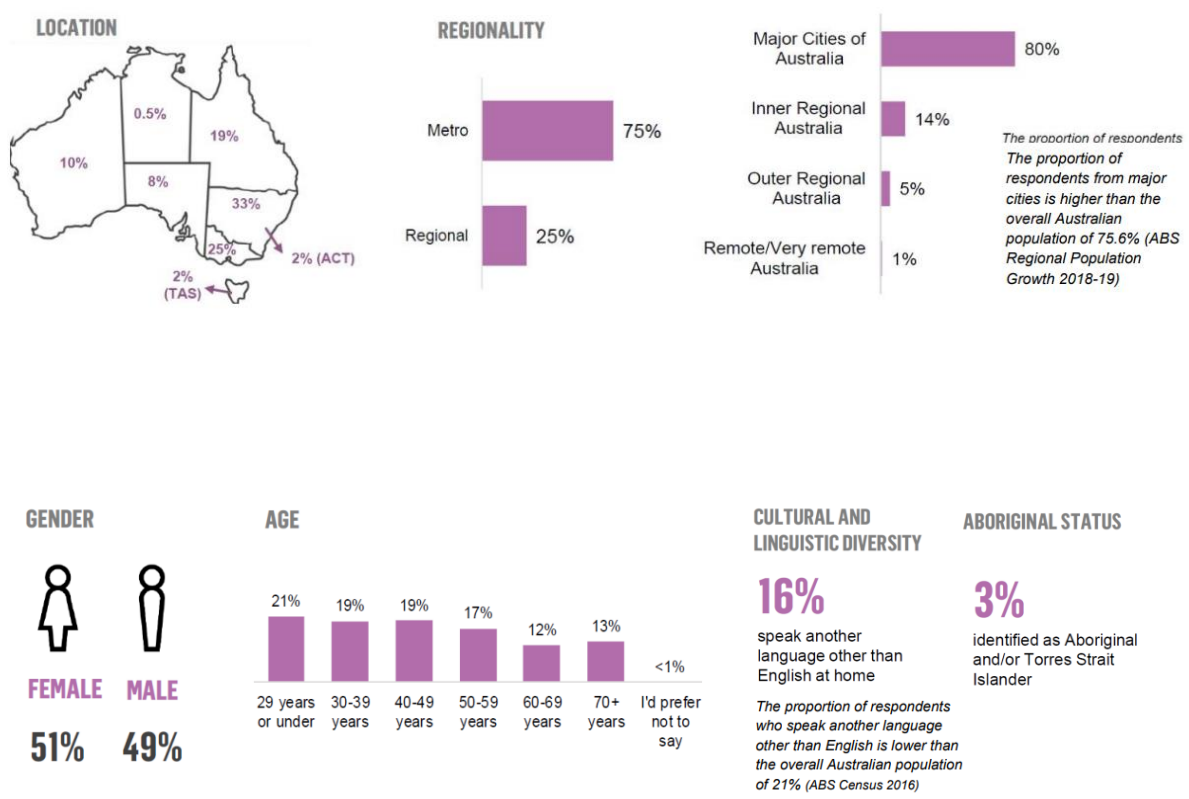Ensure compliance with data privacy regulations and protect customer information.

Address potential biases in the data or clustering algorithms to ensure fairness and inclusivity in the segmentation process.

Regularly audit and review the segmentation process to identify and mitigate any unintended consequences or ethical concerns

3. Survey:

Using customer segmentation process a detailed survey report on how the customers used to take medicines and shows their health problems and health services they done in health services we done in supermarket analysis Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster.

Demographic Profile of Respondents

LOCATION

0.5%
19%
10%
8%
33%
25%
2% (ACT)
2% (TAS)

REGIONALITY

Metro 75%
Regional 25%

Major Cities of Australia 80%
Inner Regional Australia 14%
Outer Regional Australia 5%
Remote/Very remote Australia 1%

The proportion of respondents The proportion of respondents from major cities is higher than the overall Australian population of 75.6% (ABS Regional Population Growth 2018-19)

GENDER

FEMALE 51%   MALE 49%

AGE

21% 29 years or under
19% 30-39 years
19% 40-49 years
17% 50-59 years
12% 60-69 years
13% 70+ years
<1% I'd prefer not to say

CULTURAL AND LINGUISTIC DIVERSITY

16% speak another language other than English at home

The proportion of respondents who speak another language other than English is lower than the overall Australian population of 21% (ABS Census 2016)

ABORIGINAL STATUS

3% identified as Aboriginal and/or Torres Strait Islander

Education

Primary education	1%
Secondary education	25%
Tertiary Cert	16%
Advanced Dip and Dip	14%
Bachelor's degree	25%
Graduate Dip/ Graduate Level Cert	5%
Postgraduate Degree	12%
Doctoral Degree	2%
I'd prefer not to say	<1%

The proportion of respondents with a bachelor degree qualification or above (45%), is higher than the overall Australian (15+) population of 22% (ABS

Income

Less than $10,000	3%
$10,000 - $24,999	7%
$25,000 - $49,999	17%
$50,000 - $74,999	19%
$75,000 - $99,999	16%
$100,000 - $149,999	19%
$150,000 - $174,999	5%
Over $175,000	5%
I'd prefer not to say	9%

Base n=1,503

Overall Health

Overall health rating

8% Excellent | 28% Very good | 42% Good | 18% Fair | 4% Poor | 0% I'd prefer not to say

Base: 1,503
Question: In general, my overall health is…

Health issues

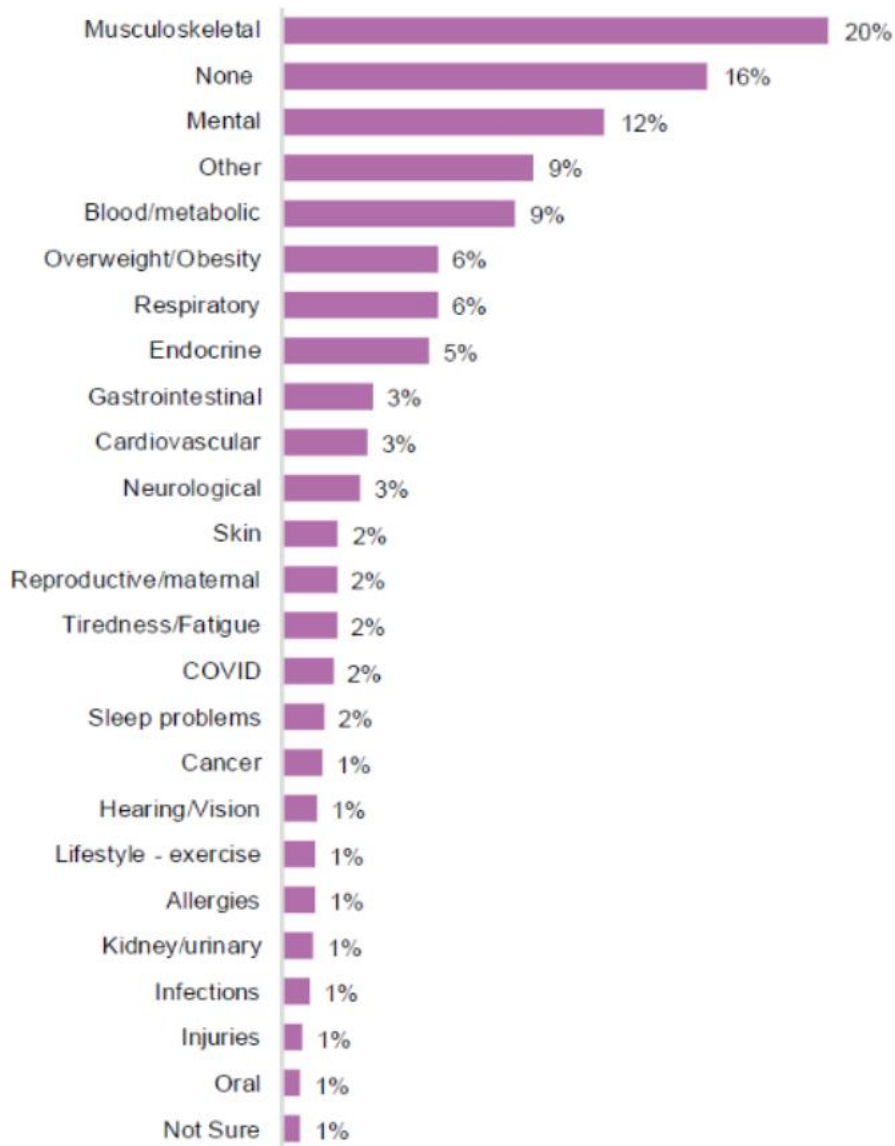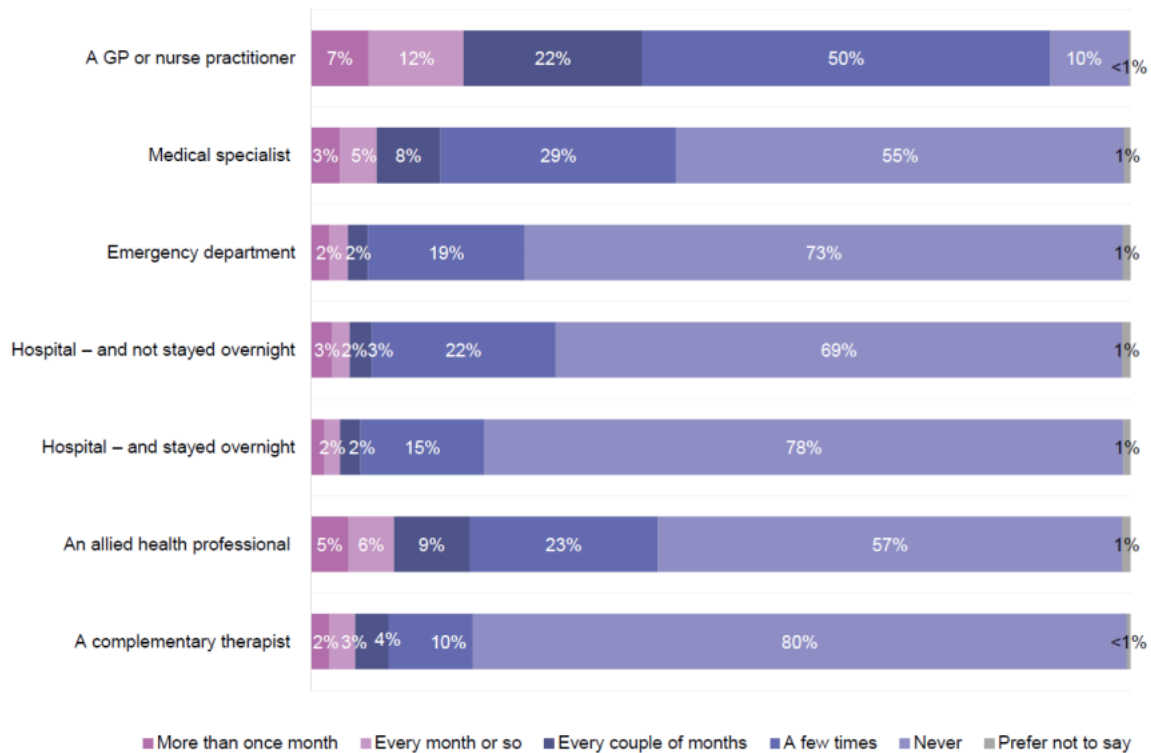|	
Arthritis	17%
Back pain	17%
Another health condition	16%
A mental health condition	16%
Asthma	12%
Diabetes	9%
Disability	6%
Cardiovascular disease	4%
Chronic obstructive pulmonary disease (COPD)	3%
Cancer	2%
Carer of someone with chronic condition	5%
None of the above	43%
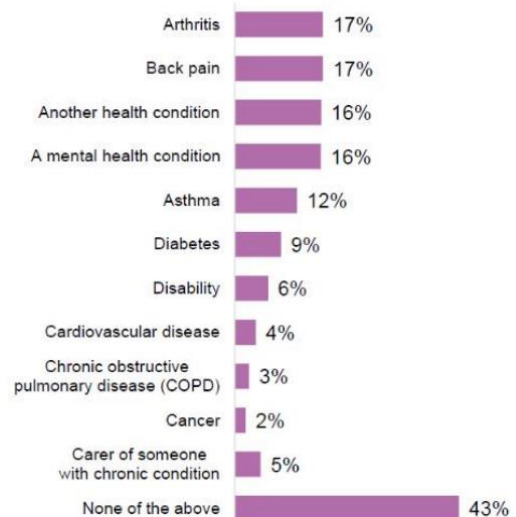
Base: 1,503
Question: Below are some more statements about your health. Please select all that apply to you. Just give your best guess if you are unsure.
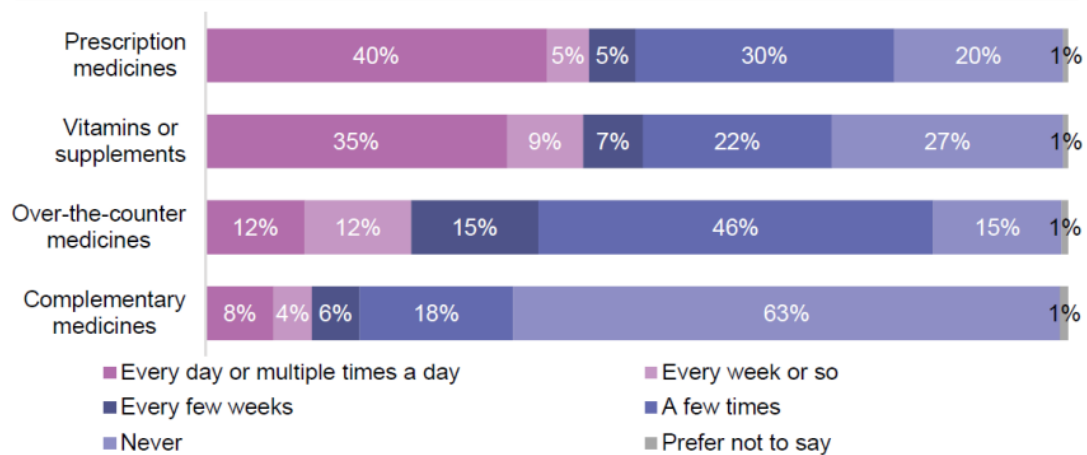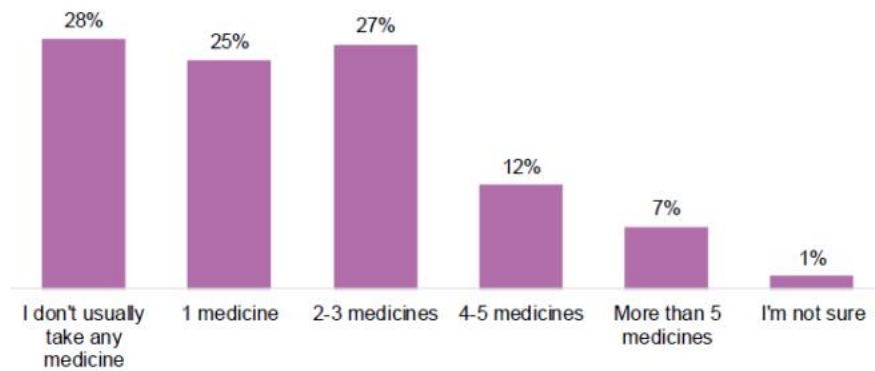
Health issues

Health issue	Percentage
Musculoskeletal	20%
None	16%
Mental	12%
Other	9%
Blood/metabolic	9%
Overweight/Obesity	6%
Respiratory	6%
Endocrine	5%
Gastrointestinal	3%
Cardiovascular	3%
Neurological	3%
Skin	2%
Reproductive/maternal	2%
Tiredness/Fatigue	2%
COVID	2%
Sleep problems	2%
Cancer	1%
Hearing/Vision	1%
Lifestyle - exercise	1%
Allergies	1%
Kidney/urinary	1%
Infections	1%
Injuries	1%
Oral	1%
Not Sure	1%

Use of Health Services

	More than once month	Every month or so	Every couple of months	A few times	Never	Prefer not to say
A GP or nurse practitioner	7%	12%	22%	50%	10%	<1%
Medical specialist	3%	5%	8%	29%	55%	1%
Emergency department	2%	2%		19%	73%	1%
Hospital – and not stayed overnight	3%	2%	3%	22%	69%	1%
Hospital – and stayed overnight	2%	2%		15%	78%	1%
An allied health professional	5%	6%	9%	23%	57%	1%
A complementary therapist	2%	3%	4%	10%	80%	<1%

Legend: ■ More than once month ■ Every month or so ■ Every couple of months ■ A few times ■ Never ■ Prefer not to say

Use of Medicines

Usages of medicines

	Every day or multiple times a day	Every week or so	Every few weeks	A few times	Never	Prefer not to say
Prescription medicines	40%	5%	5%	30%	20%	1%
Vitamins or supplements	35%	9%	7%	22%	27%	1%
Over-the-counter medicines	12%	12%	15%	46%	15%	1%
Complementary medicines	8%	4%	6%	18%	63%	1%

Legend:
■ Every day or multiple times a day   ■ Every week or so
■ Every few weeks   ■ A few times
■ Never   ■ Prefer not to say

Base: 1,503
Question: In the past year, how often have you taken…?
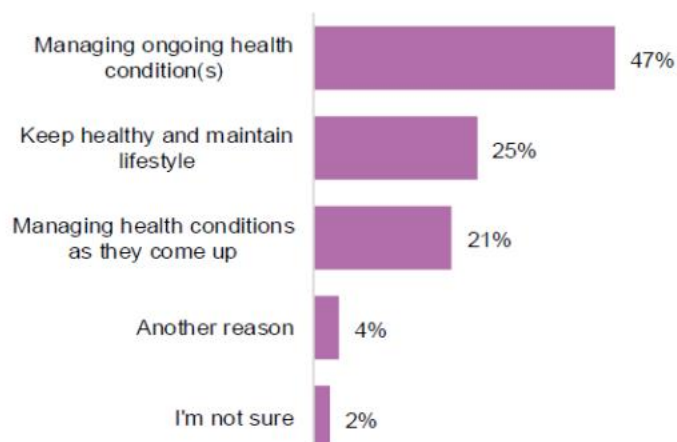
Number of medicines taken

Base: 1,503
Question: On average, how many medicines would you take each day?
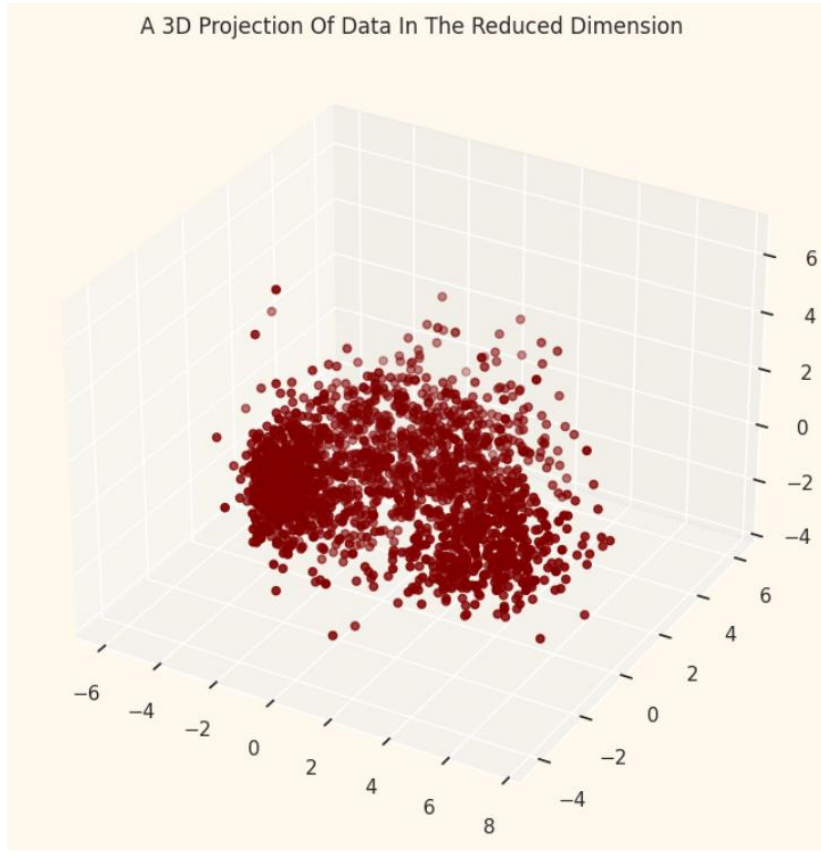This includes over-the-counter and complementary medicines and vitamins or supplements.
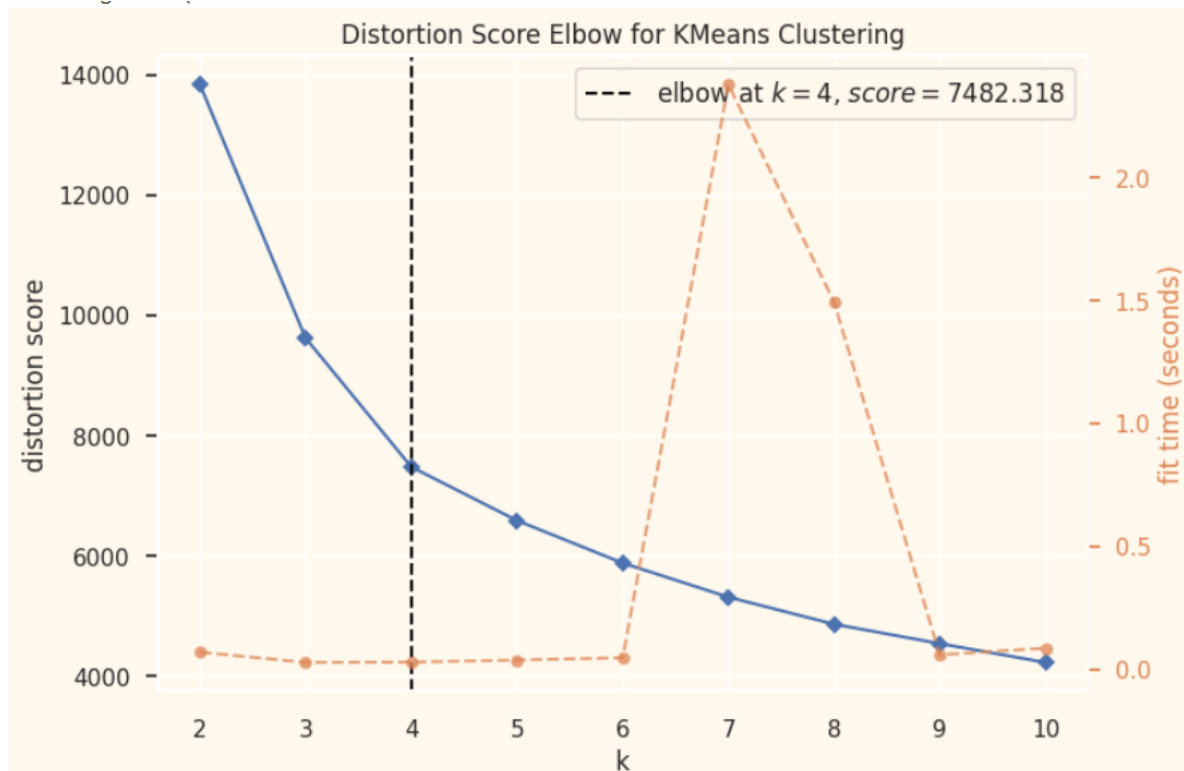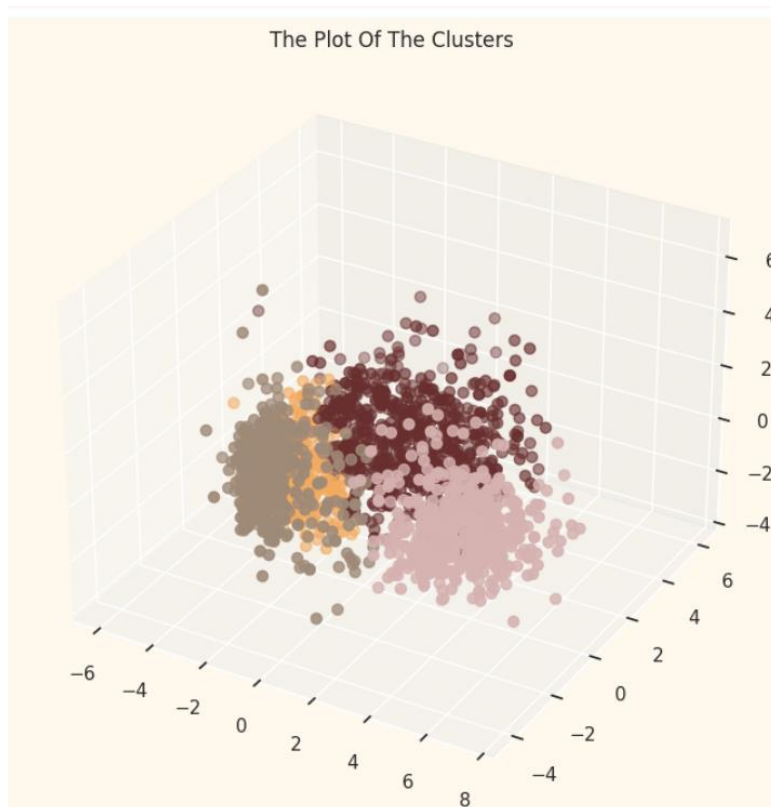
Main reason for taking medicines
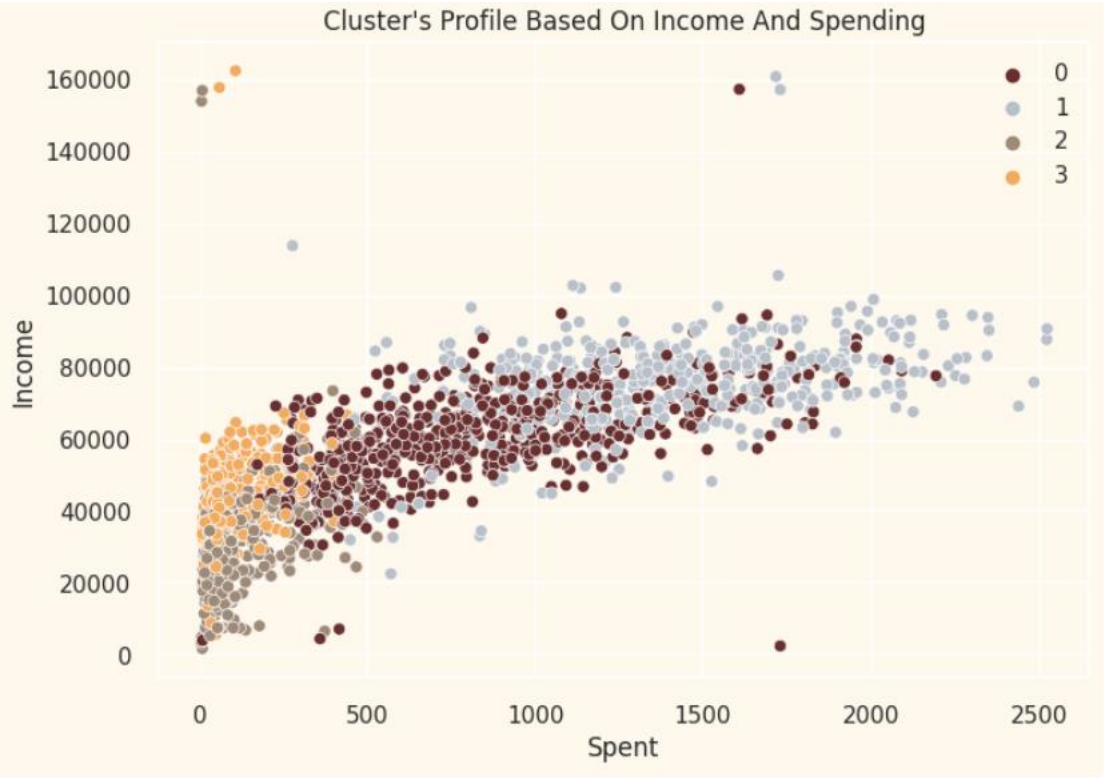
4. Results:

4.1 Dimensionality Reduction:

A 3D Projection Of Data In The Reduced Dimension

4.2 Clustering-K Means

Distortion Score Elbow for KMeans Clustering

4.3 Plot of Clusters:

The Plot Of The Clusters

4.4 Cluster's Profile Based on Income And Spending

Cluster's Profile Based On Income And Spending

4.5 Clusters for count of promotion Accepeted:

Count Of Promotion Accepted
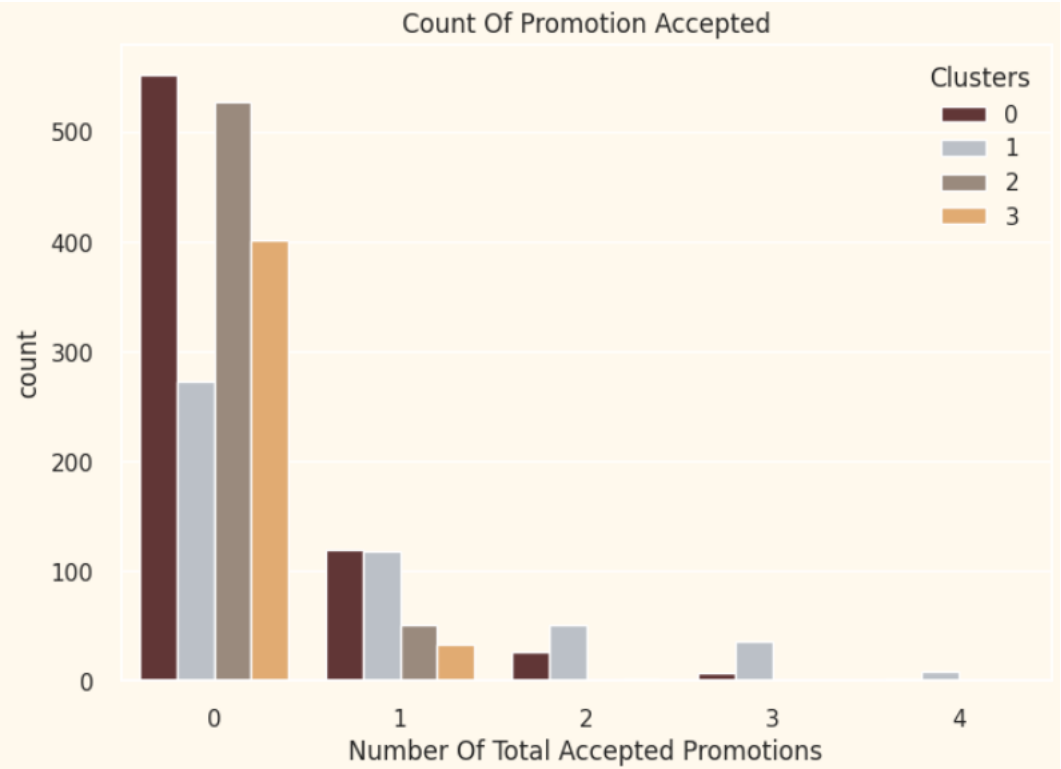
4.6 Total no of deals purchased in form of clusters:

Number of Deals Purchased
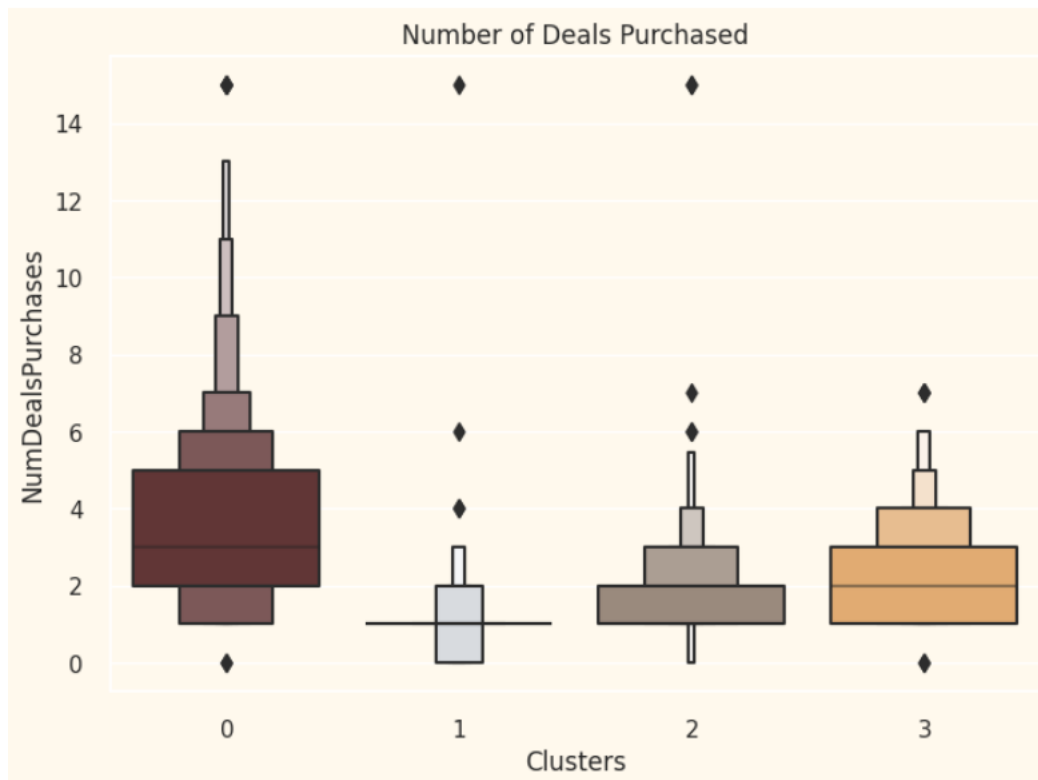
5. Conclusion:

In this project, I performed unsupervised clustering. I did use dimensionality reduction followed by agglomerative clustering. I came up with 4 clusters and further used them in profiling customers in clusters according to their family structures and income/spending. This can be used in planning better marketing strategies. And finding reports of deals in medicines as by customers health conditions of their age and their gender.

Customer segmentation provides a powerful tool for dividing into groups on based on shared characteristics and behaviors. Through the application of clustering, supermarkets can discover hidden patterns, preferences, and trends within their customer data that might not be immediately apparent

Machine learning algorithms for clustering, such as k-means, hierarchical clustering, or density-based clustering, enable supermarkets to automate the segmentation process and handle large and complex datasets. These algorithms can handle a wide range of customer attributes and variables, including demographic information, purchasing behavior, frequency of visits, product preferences, and more. By incorporating advanced clustering techniques, supermarkets can gain a comprehensive understanding of their customer base and make data-driven decisions.

6. References:

For dataset:

https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering/input

Dimensionality Reduction:

https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/

Clustering

https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Survey reports:

https://www.nps.org.au/assets/Consumer-Segmentation-Survey_Final-Report_AT.pdf

https://www.sciencegate.app/app/document#/10.32628/cseit217654/related-documents