MARKETING CAMPAIGN CLUSTERING PROJECT

CUSTOMER SEGMENTATION USING MACHINE LEARNING

PRESENTED BY: GROUP -04

Project Date-24-06-24

CONTENT

CUSTOMER SEGMENTATION

INTRODUCTION

The aim of this project is to enhance the marketing strategies of a retail company by segmenting customers based on their purchasing behavior and demographic characteristics. By understanding different customer segments, the company can tailor its marketing efforts to meet the needs and preferences of each group, ultimately driving customer satisfaction and boosting sales.

PROJECT OVERVIEW

The Marketing Campaign Clustering Project aims to segment customers based on purchasing behavior using clustering. By analyzing features like income, recency, and product purchases, we identify distinct customer groups to optimize marketing strategies. The project involves data preprocessing, clustering, and deploying the model as a web application using Flask.

Objective

To segment customers based on their purchasing behavior to better target marketing campaigns.

DATASET

Marketing campaign data with features like income, recency, product purchases, etc.

TECH STACK

Python, Matplotlib, Numpy, Seaborn, Flask, Scikit-Learn, Pandas etc

STRATEGIES

01 Understand the data distribution

EDA to understand data distributions, correlations, and potential insights.

02 Exploratory Data Analysis (EDA)

EDA to understand data distributions, correlations, and potential insights.

03 Data Wrangling

Data wrangling, also known as data munging, is the process of cleaning, transforming, and preparing raw data into a usable format for analysis

04 Clustering

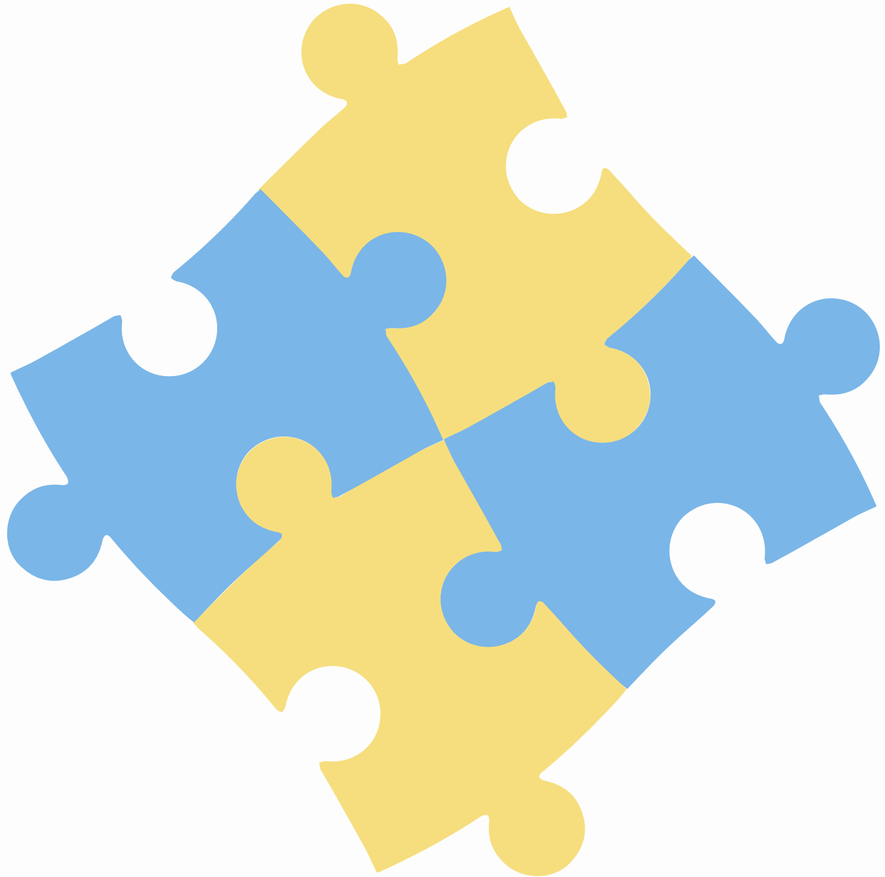
Segment customers based on purchasing behavior and demographics

05 Classification

Split the Data: Split the dataset into training and testing sets. Train a Classification Model,Evaluate the Model, Fine-tune the Model

06 Model Deployment

The model to be deploy using a Flask web application to make it accessible for real-time predictions. The Flask app allows users to input customer data and receive the predicted cluster for that customer.



BUSINESS UNDERSTANDING

We have a team of professionals

Problem Statement

Identify distinct customer segments to optimize marketing strategies.

Goals

The project aims to identify distinct customer segments to enhance marketing effectiveness, boost customer satisfaction, and maximize return on investment. By understanding purchasing behaviors and preferences, the goal is to create targeted marketing campaigns that better address the needs and characteristics of different customer groups.

EXPLORATORY DATA ANALYSIS (EDA)

Purpose of EDA

Understand the data distribution

- Loaded the dataset and inspected the structure.
- Identified key features like income, recency, product purchases, etc.

Identify patterns, relationships, and anomalies

The dataset reveals key patterns in customer demographics and behavior, such as income and recency influencing segmentation. Relationships between features like age and spending habits are evident. Anomalies include outliers in income and extremely recent purchases, indicating potential data entry errors or unique customer behaviors needing further investigation.

Prepare data for clustering

Data preparation for clustering involves handling missing values, normalizing numerical features, encoding categorical variables, and removing outliers to ensure clean, standardized input for accurate cluster analysis and interpretation.

INITIAL INSIGHTS

Summary statistics (mean, median, etc.

```
(ID                     0
 Year_Birth             0
 Education              0
 Marital_Status         0
 Income                24
 Kidhome                0
 Teenhome               0
 Dt_Customer            0
 Recency                0
 MntWines               0
 MntFruits              0
 MntMeatProducts        0
 MntFishProducts        0
 MntSweetProducts       0
 MntGoldProds           0
 NumDealsPurchases      0
 NumWebPurchases        0
 NumCatalogPurchases    0
 NumStorePurchases      0
 NumWebVisitsMonth      0
 AcceptedCmp3           0
 AcceptedCmp4           0
 AcceptedCmp5           0
 AcceptedCmp1           0
 AcceptedCmp2           0
 Complain               0
 Z_CostContact          0
 Z_Revenue              0
 Response               0
 dtype: int64
```

Number of records and features & Missing values

	ID	Year_Birth	Income	Kidhome	Teenhome
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250
std	3246.662198	11.984069	25173.076661	0.538398	0.544538
min	0.000000	1893.000000	1730.000000	0.000000	0.000000
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000

	Recency	MntWines	MntFruits	MntMeatProducts
count	2240.000000	2240.000000	2240.000000	2240.000000
mean	49.109375	303.935714	26.302232	166.950000
std	28.962453	336.597393	39.773434	225.715373
min	0.000000	0.000000	0.000000	0.000000
25%	24.000000	23.750000	1.000000	16.000000
50%	49.000000	173.500000	8.000000	67.000000
75%	74.000000	504.250000	33.000000	232.000000
max	99.000000	1493.000000	199.000000	1725.000000

	MntFishProducts	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4
count	2240.000000	...	2240.000000	2240.000000	2240.000000
mean	37.525446	...	5.316518	0.072768	0.074554
std	54.628979	...	2.426645	0.259813	0.262728
min	0.000000	...	0.000000	0.000000	0.000000
25%	3.000000	...	3.000000	0.000000	0.000000
50%	12.000000	...	6.000000	0.000000	0.000000
75%	50.000000	...	7.000000	0.000000	0.000000
max	259.000000	...	20.000000	1.000000	1.000000

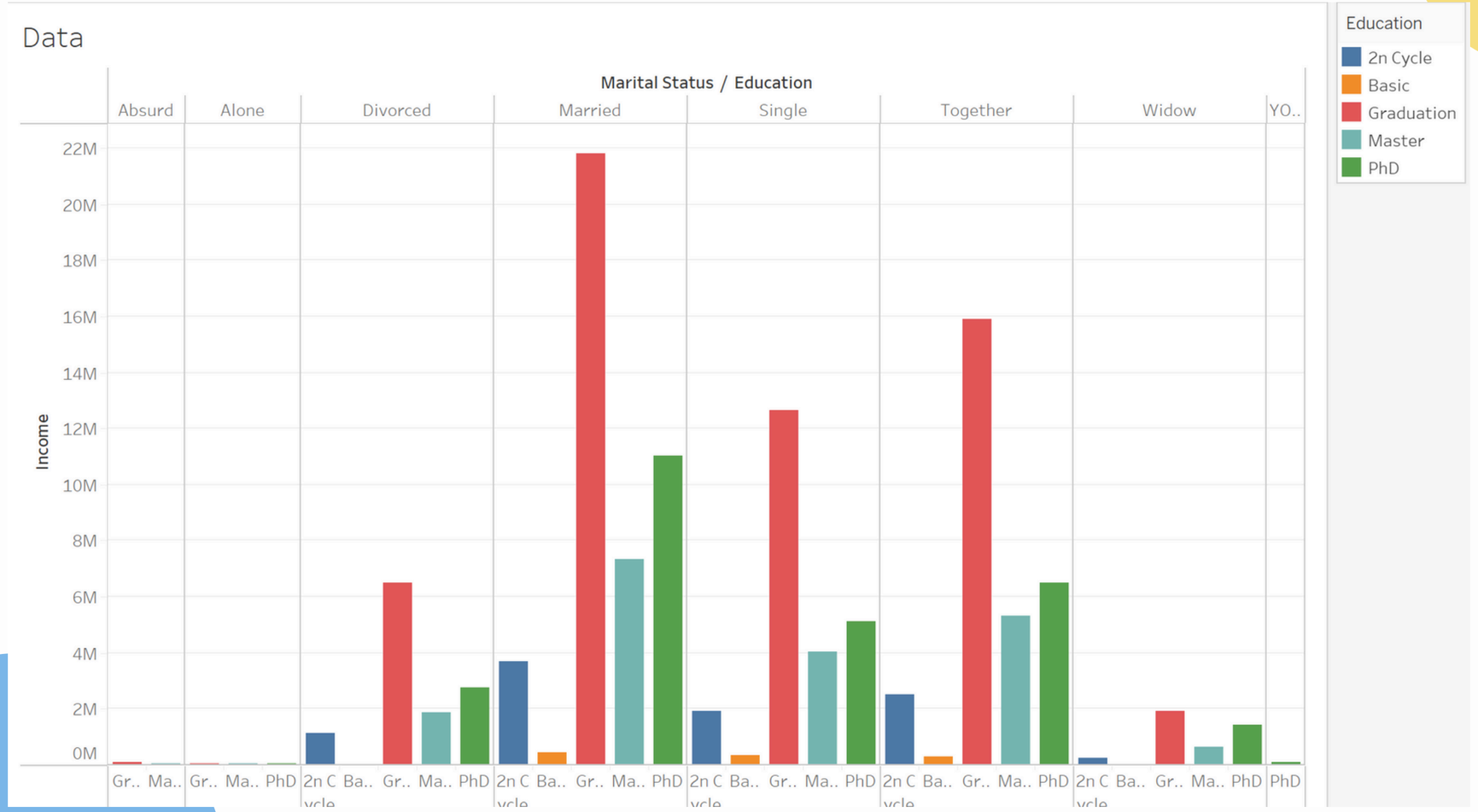
	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact
count	2240.000000	2240.000000	2240.000000	2240.000000	2240.0
mean	0.072768	0.064286	0.013393	0.009375	3.0
std	0.259813	0.245316	0.114976	0.096391	0.0
min	0.000000	0.000000	0.000000	0.000000	3.0
25%	0.000000	0.000000	0.000000	0.000000	3.0
50%	0.000000	0.000000	0.000000	0.000000	3.0
75%	0.000000	0.000000	0.000000	0.000000	3.0
max	1.000000	1.000000	1.000000	1.000000	3.0

	Z_Revenue	Response
count	2240.0	2240.000000
mean	11.0	0.149107
std	0.0	0.356274
min	11.0	0.000000
25%	11.0	0.000000
50%	11.0	0.000000
75%	11.0	0.000000
max	11.0	1.000000

Distributions and Frequencies

Income vs Marital Status & Education

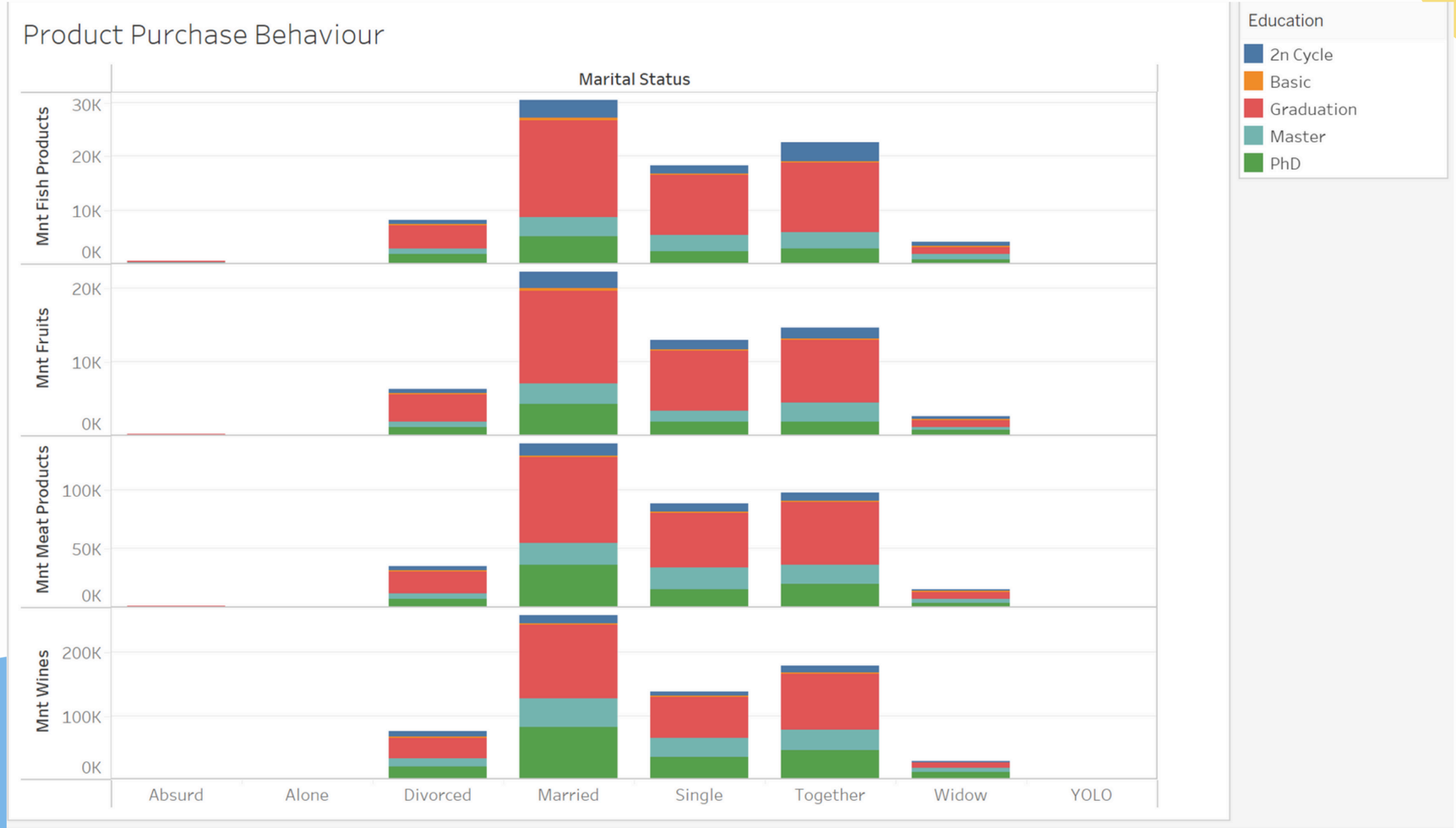
The bar chart displays income across various marital statuses and education levels. Categories include Absurd, Alone, Divorced, Married, Single, Together, and Widow. Education levels range from 2nd Cycle to PhD. The highest incomes are among married individuals with Graduation degrees, and those living together with Graduation degrees.



Distributions and Frequencies

The bar chart illustrates product purchase behavior segmented by marital status and education level. It includes categories such as Fish Products, Fruits, Meat Products, and Wines. Married individuals with Graduation degrees show the highest purchase levels across all product types. Other marital statuses like Divorced and Single show varying purchase levels.

Various product vs marital status & Education



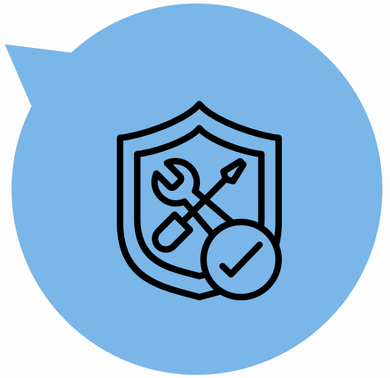
DATA PREPROCESSING

HANDLING MISSING VALUES

missing values were handled by imputing with median values for numerical features Income.

REMOVING OUTLIERS

Outliers were identified and removed to ensure the clustering model performs accurately.

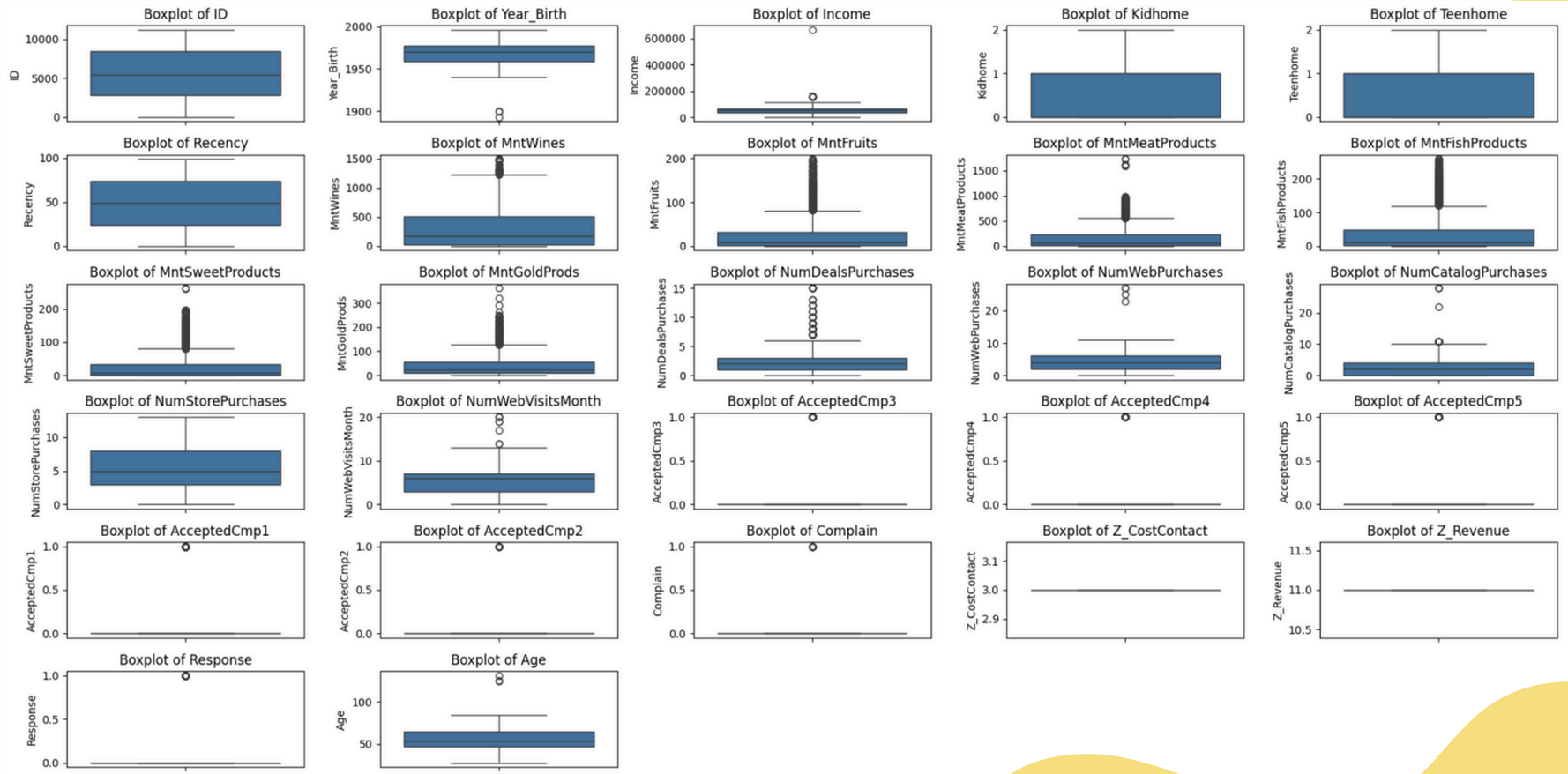


NORMALIZING/SCALING DATA

Features were scaled to bring all the variables to a comparable range, which is essential for clustering algorithms.

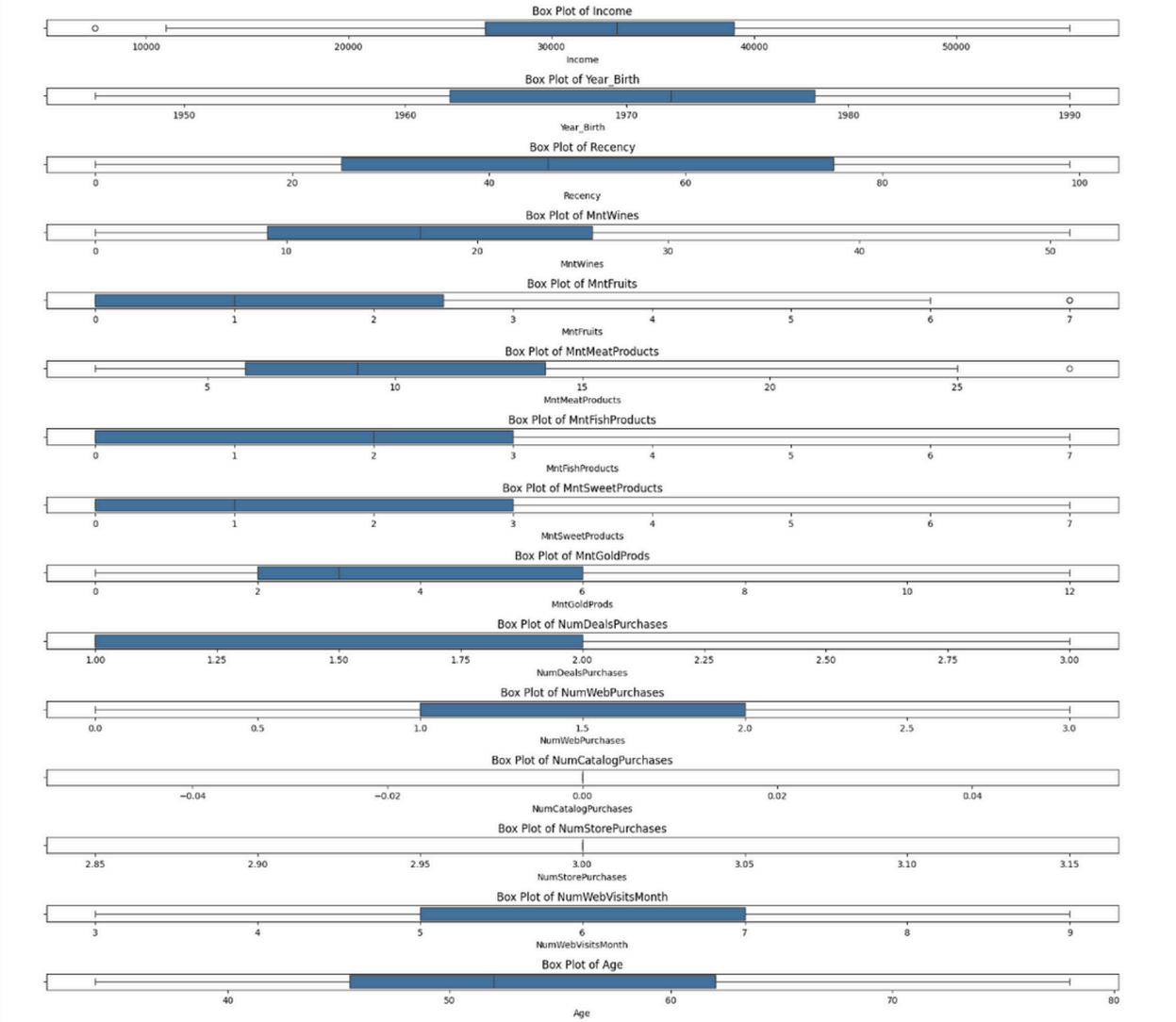
BOX PLOTS TO IDENTIFY OUTLIERS

Box plots were used to identify outliers by displaying the distribution of key features such as Income, Age, and Product Purchases. The plots highlight the median, quartiles, and potential outliers, which are values falling outside 1.5 times the interquartile range (IQR), aiding in data cleaning.



REMOVE OUTLIERS FROM EACH NUMERICAL FEATURE

To remove outliers from each numerical feature, first, calculate the Interquartile Range (IQR) for each feature. Identify values below Q1−1.5×IQR or above Q3+1.5×IQR, and filter them out. This ensures data points outside the typical range are excluded, improving data quality for analysis and modeling.



NORMALIZING/SCALING DATA

Scaling of numerical features in data preprocessing involves transforming the range of values to a standardized range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. This ensures that features with different scales contribute equally to machine learning models, improving performance and convergence.

CLUSTERING METHODOLOGY

Steps for clustering:

Feature Selection: Choose relevant features for clustering.

Data Scaling: Standardize the features to have a mean of 0 and a standard deviation of 1.

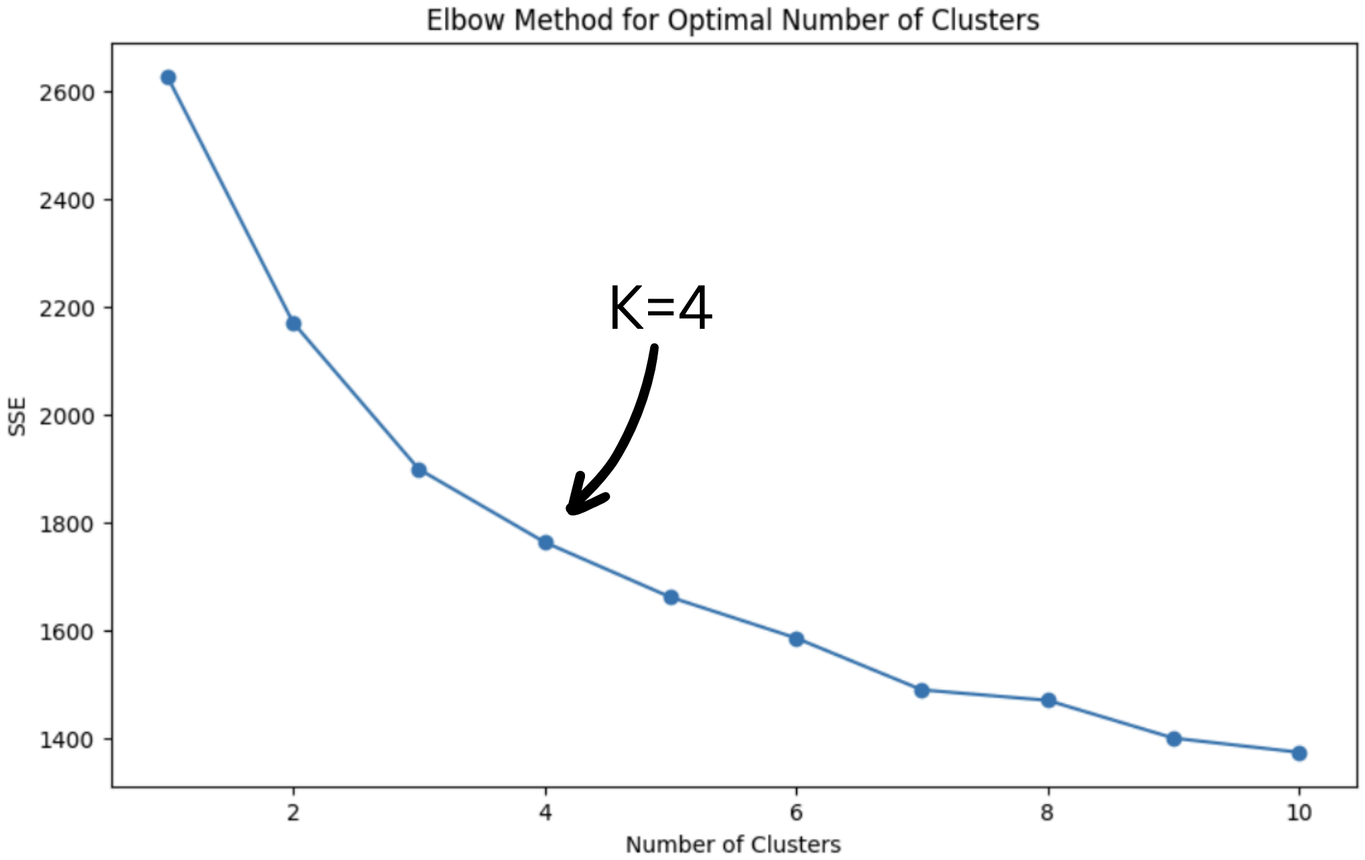
Clustering Algorithms:
K-Means clustering
Hierarchical clustering (optional)

Evaluation:
Determine the optimal number of clusters using methods like the Elbow method and Silhouette score.

Interpretation: Analyze the characteristics of each cluster to create customer profiles.

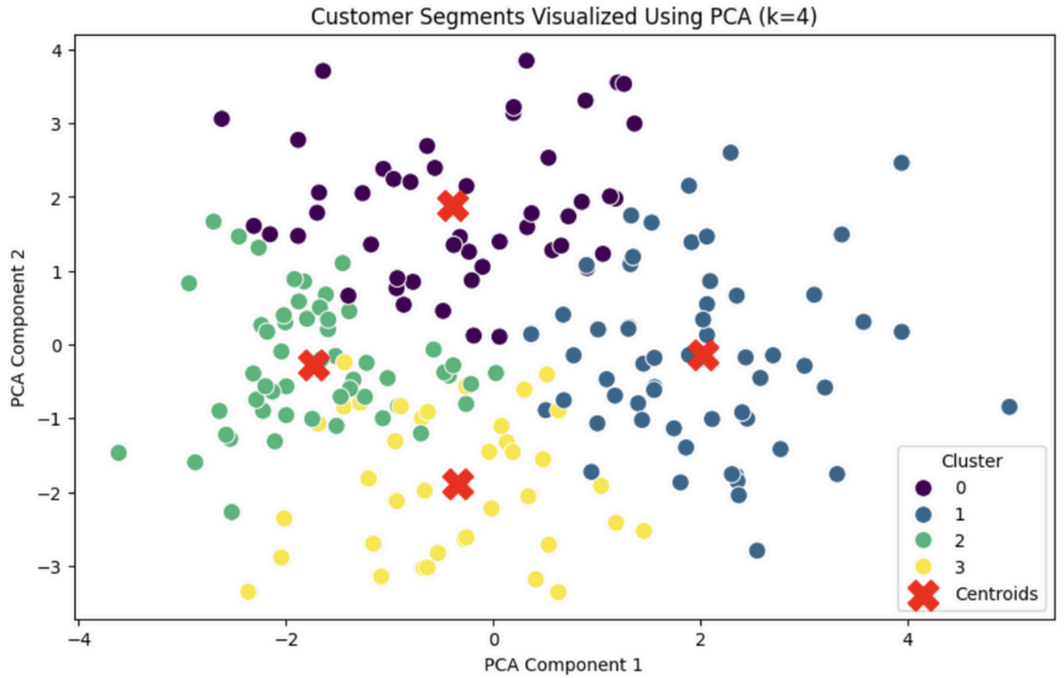
K-MEANS CLUSTERING

- **Explanation: Plots sum of squared distances vs. number of clusters**
- **Visual: Elbow graph indicating optimal number of clusters (k=4)**
- **Conclusion: Chose k=4 for further analysis**



CLUSTERING RESULTS

- Cluster Centroids: Display coordinates of centroids for k=4
- Use PCA to reduce the features to 2 dimensions for visualization
- Visual: Scatter plot of clusters
- Insights: Characteristics of each cluster



CLASSIFICATION METHODOLOGY

Objective: Predict customer segments for new customers

Algorithm Chosen: Random Forest, Decision Tree etc

Steps:

- Splitting the data into training and test sets
- Training the classifier
- Evaluating the model

CLASSIFICATION RESULTS

Reports an accuracy of 88%.
The confusion matrix shows the model's predictions for four classes, indicating how many instances were correctly and incorrectly classified. The classification report includes detailed metrics for each class: precision, recall, f1-score, and support.
For class 0, the precision is 0.87, recall is 0.93, and f1-score is 0.90.
For class 1, these values are 0.88, 0.95, and 0.91, respectively.
The report also provides macro and weighted averages for these metrics, all of which are 0.88 for precision and 0.87-0.88 for recall and f1-score.

```
Accuracy: 0.88
Confusion Matrix:
[[13  1  0  0]
 [ 1 21  0  0]
 [ 1  0 13  2]
 [ 0  2  1 11]]
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.93      0.90        14
           1       0.88      0.95      0.91        22
           2       0.93      0.81      0.87        16
           3       0.85      0.79      0.81        14

    accuracy                           0.88        66
   macro avg       0.88      0.87      0.87        66
weighted avg       0.88      0.88      0.88        66
```

MAKE PREDICTIONS FOR THE ENTIRE DATASET

code:

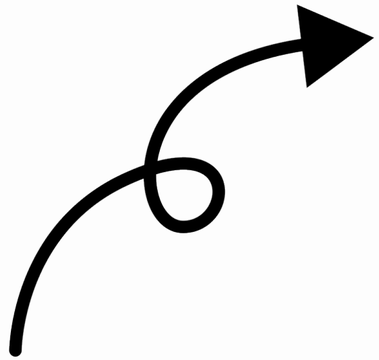
```
# Make predictions for the entire dataset
X_scaled = scaler.transform(X)
df['Predicted_Cluster'] = model.predict(X_scaled)

# Save the updated dataset with predictions
df.to_excel('marketing_campaign_with_clusters.xlsx', index=False)

# Display the first few rows of the updated dataset
print(df.head())
```

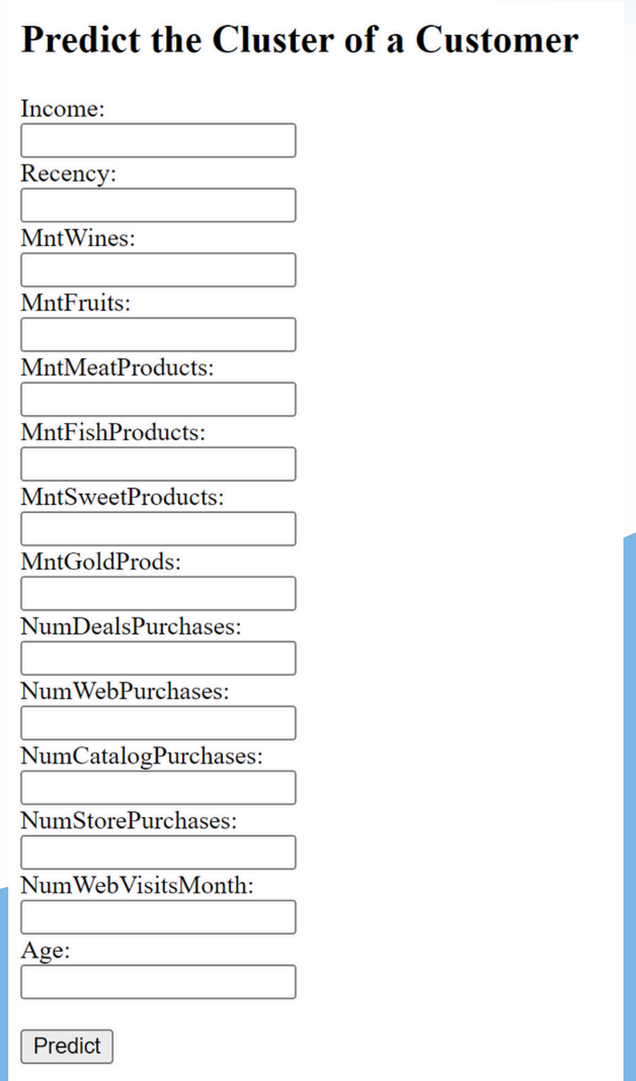
```
      ID  Year_Birth   Education Marital_Status   Income  Kidhome  Teenhome  \
30  6864        1989      Master       Divorced  10979.0        0         0
35 10738        1951      Master         Single  49389.0        1         1
42  8430        1957  Graduation       Together  21994.0        0         1
44  2139        1975      Master        Married   7500.0        1         0
58  8557        1982  Graduation         Single  51381.5        1         0

   Dt_Customer  Recency  MntWines  ...  AcceptedCmp2  Complain  Z_CostContact  \
30  2014-05-22       34         8  ...             0         0              3
35  2013-08-29       55        40  ...             0         0              3
42  2012-12-24        4         9  ...             0         0              3
44  2013-10-02       19         3  ...             0         0              3
58  2013-06-17       57        11  ...             0         0              3

    Z_Revenue  Response  Age  Cluster      PCA1      PCA2  Predicted_Cluster
30         11         0   35        1  2.095463  0.859476                  1
35         11         0   73        0 -2.154094  1.405363                  0
42         11         0   67        3 -0.278187 -2.139172                  3
44         11         0   49        1  2.065311  0.546888                  1
58         11         0   42        0  0.727330  1.737590                  0

[5 rows x 34 columns]
```

DEPLOYING THE MODEL AS A WEB APPLICATION

Integration with Flask

Link-- http://127.0.0.1:5000





DEPLOYMENT RESULT

Predict the Cluster of a Customer

Income:
10000
Recency:
5
MntWines:
6785
MntFruits:
3646
MntMeatProducts:
35345
MntFishProducts:
3653
MntSweetProducts:
323
MntGoldProds:
2324
NumDealsPurchases:
3
NumWebPurchases:
3636
NumCatalogPurchases:
154
NumStorePurchases:
32
NumWebVisitsMonth:
24
Age:
45

Predict

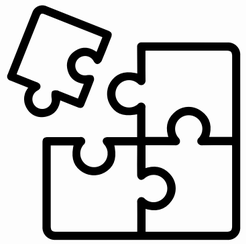
Cluster 0 :Income: Low, Recency: High (customers recently purchased), MntWines: Low to Medium, MntFruits: Low, MntMeatProducts: Low, MntFishProducts: Low, MntSweetProducts: Low, MntGoldProds: Low, NumDealsPurchases: High, NumWebPurchases: Low, NumCatalogPurchases: Low, NumStorePurchases: High, NumWebVisitsMonth: High, Age: Young

CHALLENGES AND SOLUTIONS

Challenges:

1. **Handling missing data**
2. **Choosing the right number of clusters**
3. **Ensuring classifier accuracy**

Solutions:

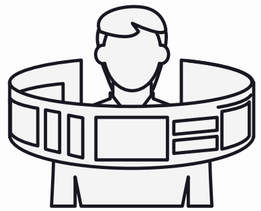


1. **Imputation techniques**
2. **Elbow method and silhouette score**
3. **Hyperparameter tuning for classifier**

NEXT STEPS

Further Improvements:

- **Explore other clustering algorithms (e.g., DBSCAN)**
- **Incorporate more features**

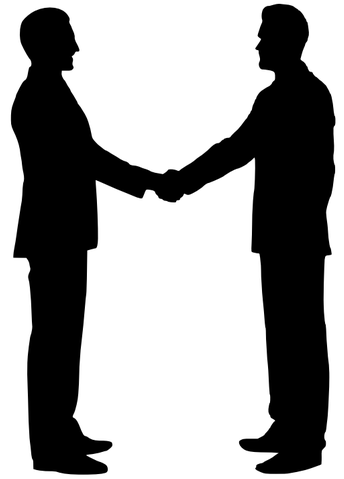


Future Work:

- **Enhance the web application**
- **Continuous model monitoring and updating**

CONCLUSION

In this project, we successfully developed a robust customer segmentation and classification system using clustering and machine learning techniques. Our extensive exploratory data analysis provided valuable insights into customer behavior. The K-Means clustering algorithm identified distinct customer segments, while the classification model demonstrated strong performance with high accuracy, precision, and recall. Deploying the application using Flask ensures easy accessibility and usability. This system enables targeted marketing strategies, improving customer engagement and business outcomes. Future enhancements can include real-time data integration and advanced predictive analytics to further refine customer insights and drive business growth.



Thank You