Customer Personality
ANALYSIS

https://project-wzz2b3kmob3ruiuvkuebb4.streamlit.app

Meet
THE TEAM

MR. ABHISHEK B R

MR. MAYUR BALKRISHNA BARGE

MR. G. JEEVAN

MR VINAY SWAROOP

MS. SHREYAL UMREDKAR

MR. KRISHNENDU RAJ

MR. HARSHVARDHAN SURYAKANT PATIL

Table of CONTENTS

INTRODUCTION

In today's highly competitive market, understanding customers' preferences and behaviors is crucial for businesses seeking to tailor their products and services effectively. Customer segmentation through clustering analysis offers a powerful means to achieve this goal. By grouping customers with similar characteristics together, businesses can identify distinct market segments, personalize marketing strategies, and enhance customer satisfaction.

In this project, we aim to leverage clustering techniques to segment our customer base. By analyzing various customer attributes such as demographics, purchase history, browsing behavior, and product preferences, we seek to uncover meaningful patterns and segments within our customer data. Through this analysis, we aim to gain insights into the diverse needs and preferences of our customer base, enabling us to optimize marketing campaigns, improve product offerings, and ultimately drive business growth.

DATA OVERVIEW

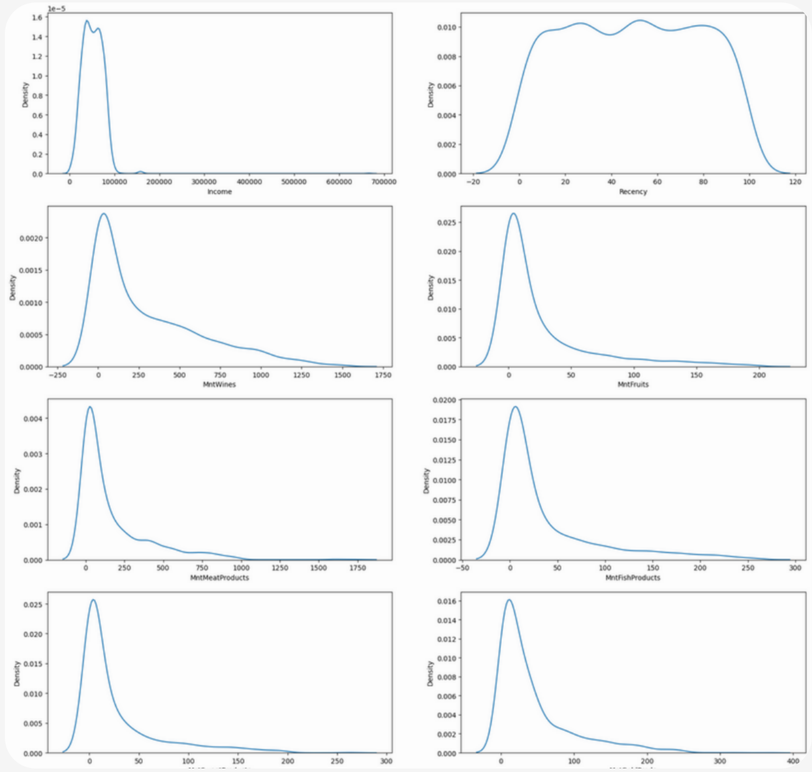
Count Plot

We Used it for exploring the distribution of categorical variables in a dataset. They help in understanding the relative proportions of different categories and identifying any patterns or anomalies.

Count plots make it easy to compare the frequency of categories side by side.

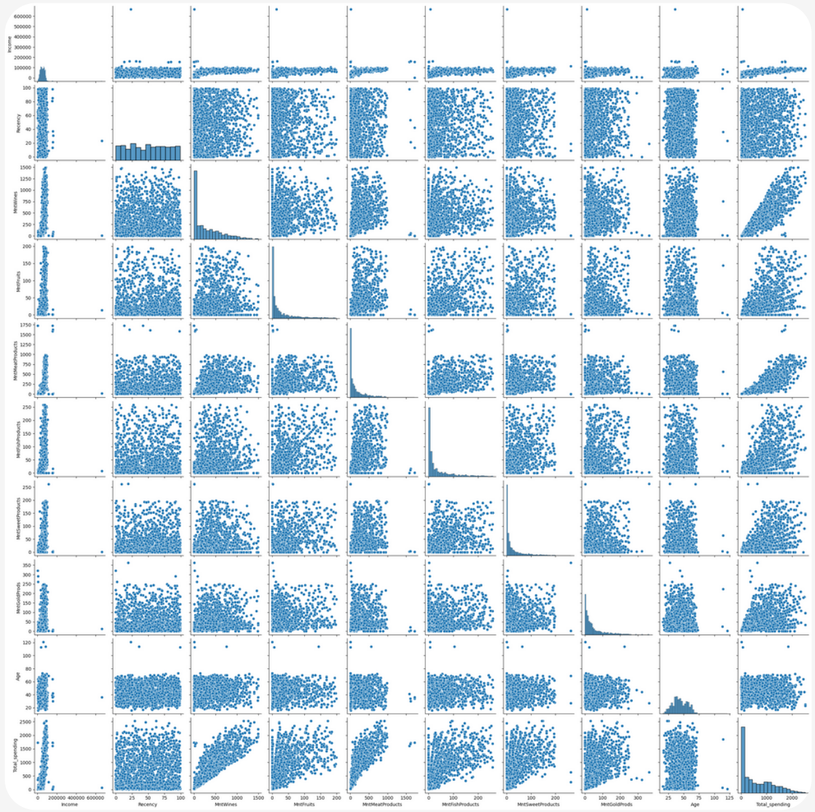
This has allowed us for a quick understanding of the distribution of data across different categories.

DATA OVERVIEW



KDE plot

By estimating the probability density function of Few found Skew in few variables which can be considered as outliers in the data.
It helped us to visualize the shape, spread, and central tendency of the data.



DATA OVERVIEW

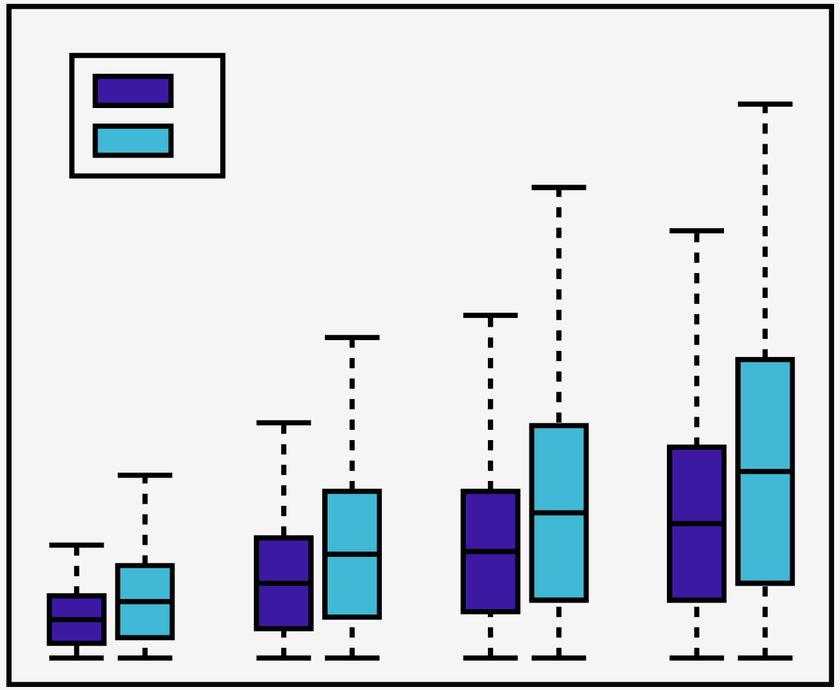
Pair plot

It displays scatterplots for each pair of variables along with histograms of each variable's distribution on the diagonal.

We have got insight into the relationships between variables in a dataset.

As income increases spending also increases

DATA OVERVIEW

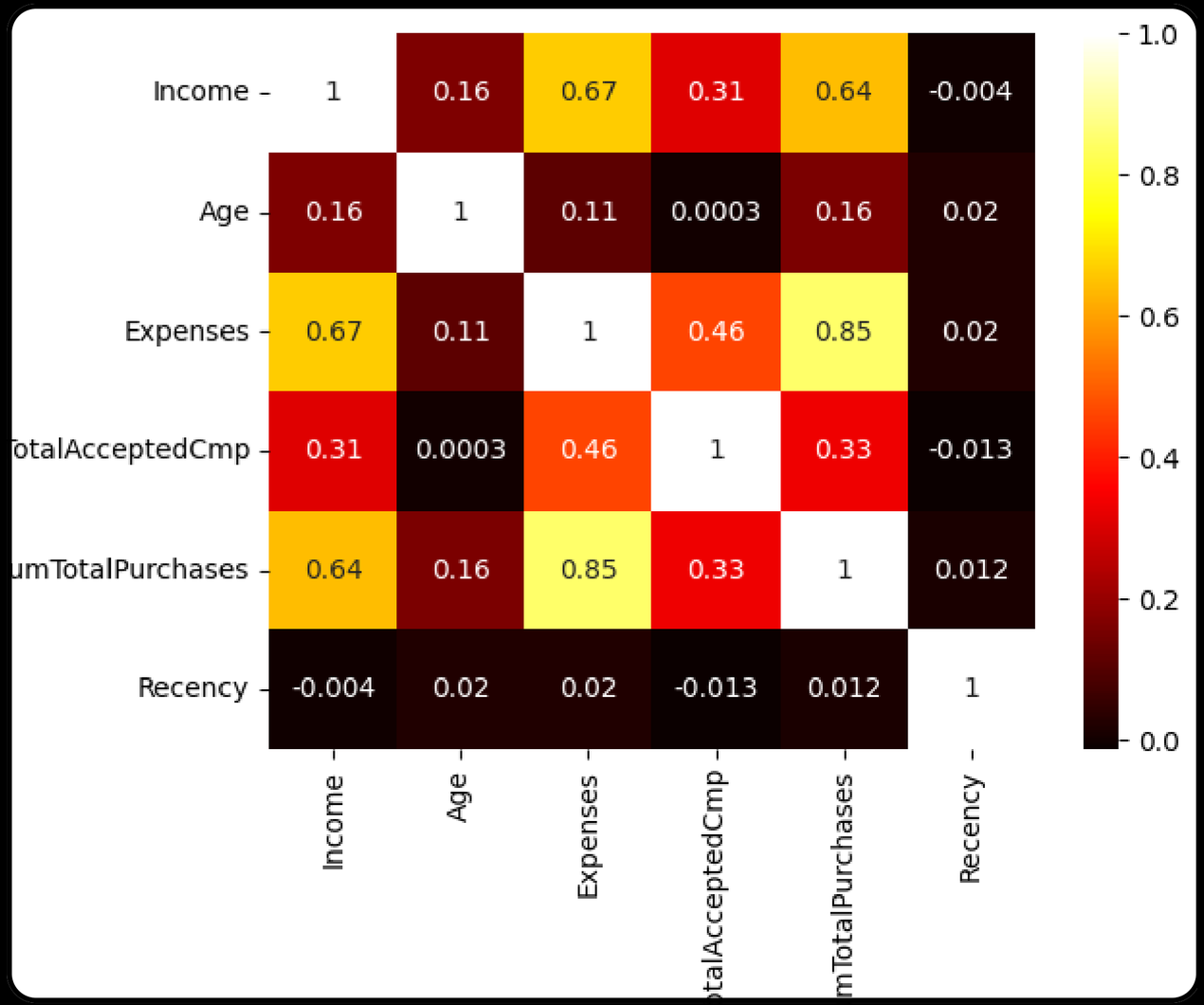


Box plot

we can see the skewness and some outliers in age, income, the amount spent on gold, meat, etc. features
where Recency has a normal distribution and doesn't have any outliers
we found outliers in age and income.
We found outliers in a few others as well like products and purchases.

HeatMap

- The highest correlation between features is 0.85.
- There are no continuous features that correlates with Expenses below 0.02 and above 0.85
- Day_engaged and Age have the lowest correlation with expenses, almost reaching 0.14 while NumTotalPurchases have the highest correlation with the expenses.



DATA CLEANING

We have found null values in income that are dropped.
We don't have any duplicates.
Removed outliers from the data (Income, age, products, and purchases)
Also dropped variables that don't add any value to the analysis(Z_CostContact, Z_Revenue, and few others)

Feature ENGINEERING

We have generated new features by combining or transforming existing ones.
Standardization: Converting categorical variables into numerical representations
Extracting useful information from date and time variables, such as day of the week, month, or time elapsed since a specific event.
Scaling numerical features and important features to reduce dimensionality **(PCA)**

Clustering

We need to form clusters on Unsupervised data.
We plotted an L-bow graph for best clustering or k value
Predicting the higher silhouette score for better results.
when we formed 2-3 clusters, we got the best results.

Hierarchical

We formed 2 clusters according to the dendrogram but we are dealing with a big data set that doesn't give great results so we avoided it

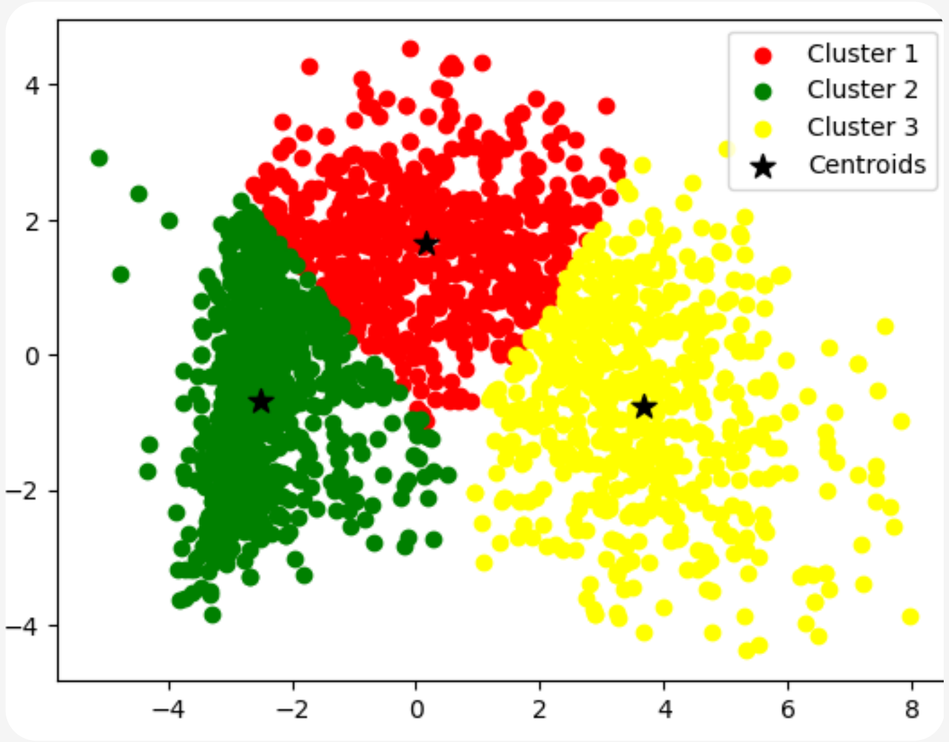
K-mean

We formed 3 clusters according to the L-bow graph the avg silhouette score was around 40.04 which was the highest

DBSCAN

Just like l-bow for k-mean, we use K Distance Graph for DBSCAN
we found the highest silhouette score when eps=6, and min_samples=60

K-Mean



- We chose K-mean for the better results
- We plotted an L-bow graph for best clustering or k value
- Predicting the higher silhouette score for better results.
- Assigned each data point to the nearest cluster centroid.
- Recalculate the cluster centroids by taking the mean of all data points assigned to each cluster.
- Repeated steps 2 and 3 until the cluster assignments no longer change significantly or a maximum number of iterations is reached.
- The algorithm produced 3 cluster centroids and assigned each data point to one of the clusters.

Model BUILDING

- We have gone with the Random Forrest model for the dataset.
- As they are less prone to overfitting
- This model is built on n_estimators=100, random_state=42
- The data is split into train and test
- It finding the accuracy we got 96.6%

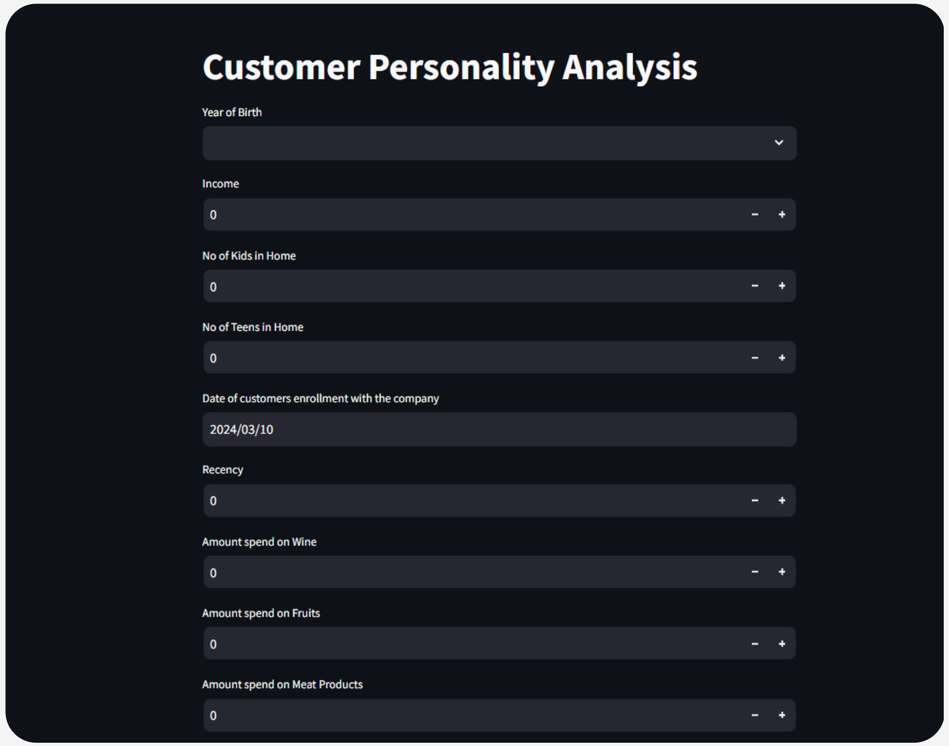
Model EVALUATION

- The model is validated on KFold and Cross_val_score.
- We have done on random forest model and X, Y which was initialized in the model, with cv=5.
- we got an Average Accuracy of 0.95 for it.
- In the process, we found out that Xgboost and a few other models give close to 0.96.

CLASSIFICATION_REPORT

PROJECT	precision	recall	f1 scoreses	support
0	0.95	0.93	0.94	120
1	0.97	0.98	0.98	130
2	0.97	0.97	0.97	194
Accuracy			0.97	444
Macro Avg	0.96	0.96	0.96	444
Weighed Avg	0.97	0.97	0.97	444

Deployment



Integration Steps:

Preprocessing and Model Training:

- Outline the preprocessing steps for data preparation.
- Describe how the Random Forest model is trained on the prepared data.

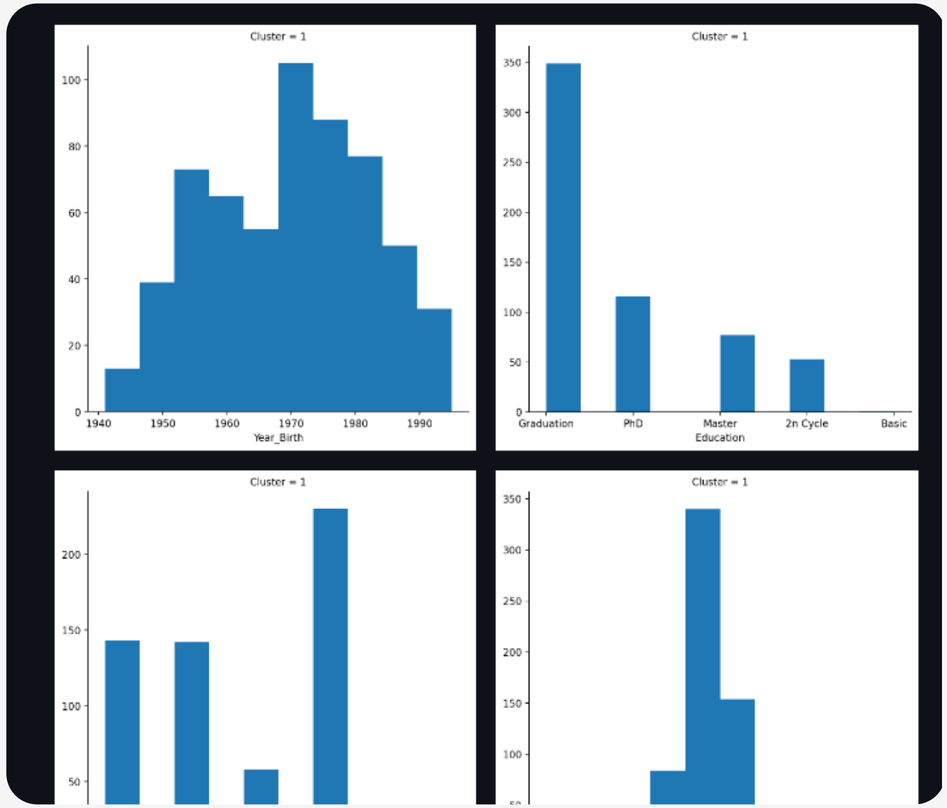
Exporting the Model:

- Explain the process of saving the trained Random Forest model using serialization techniques like Pickle or Joblib.

Building the Streamlit App:

- Provide an overview of setting up a Streamlit application.
- Mention the installation process and basic structure of a Streamlit script.

Deployment



Loading the Model in Streamlit:

- Demonstrate how to load the pre-trained Random Forest model within the Streamlit app.

Creating the User Interface:

- Showcase how to design an interactive user interface using Streamlit's intuitive widgets.
- Mention the incorporation of features like sliders, text inputs, and buttons for user interaction.

- Summarize the key points discussed, emphasizing the ease and effectiveness of deploying a Random Forest model with Streamlit.

Conclusion

0 cluster:

- Middle-income people (average income equals 50000)
- The average age is 52 years
- Have an education (Graduation, 2n Cycle, Master, PhD)
- People without families, people with families with and without children
- Quite often buy wines, but they also often buy meat
- Most often make purchases on the web
- The average number of purchases is 13

1 cluster:

- High-income people (average income equals 70000)
- The average age is 55 years
- Have an education (Graduation, 2n Cycle, Master, PhD)
- Have a family with children (Teenhome)
- Quite often buy wines, but they also often buy meat
- Most often make purchases in the stores themselves
- Most often make purchases (compared to other clusters)

2 cluster:

- Average income equals 38000
- The average age is 49 years
- Have an education (Graduation, 2n Cycle, Master, PhD)
- People with families with and without children
- A low number of purchases and, accordingly, spend little money on purchases