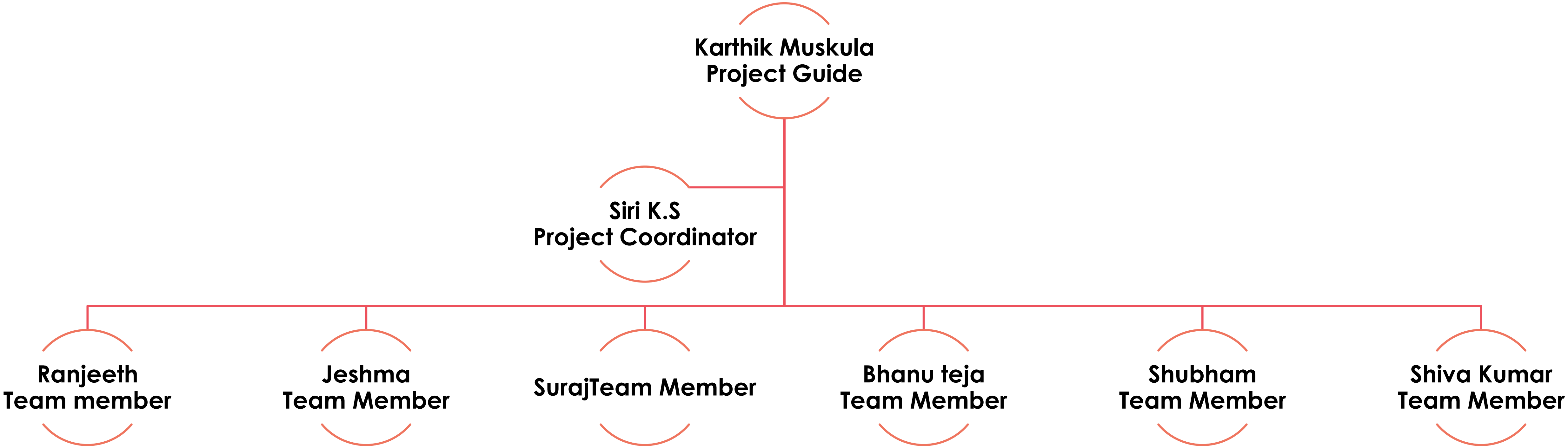
Customer Personality Analysis

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments.



Project Team Structure:

Milestones:

Milestones	Duration	Task Start – End Date
Kick off and Business Objective discussion	1 day	28^{th} April
Data set Details	1 Week – 1 ½ week	2^{nd} May
EDA	1 Week – 1 ½ week	6^{th} May
Model Building	1 Week – 1 ½ week	
Model Evaluation	1 Week	
Feedback		
Deployment	1 Week	
Final Presentation	1 day	

Introduction

Customer personality analysis goes beyond surface-level demographics to delve into the underlying motivations, desires, and pain points of individual customers. By segmenting customers based on their unique personality traits and preferences, businesses can create targeted marketing campaigns, personalize product offerings, and enhance the overall customer experience.

- Benefits of Customer Personality Analysis:

    1. Precision Targeting: By understanding the nuanced preferences of different customer segments, businesses can target their marketing efforts with precision, ensuring that messages resonate with the right audience at the right time.
    2. Product Customization: Armed with insights into customer preferences, businesses can tailor their products and services to meet the specific needs and desires of different customer segments, driving greater satisfaction and loyalty.
    3. Optimized Marketing Spend: Rather than adopting a one-size-fits-all approach, customer personality analysis allows businesses to allocate their marketing budget more efficiently, focusing resources on the channels and tactics that are most effective for each segment.
    4. Enhanced Customer Experience: By delivering personalized experiences that align with individual preferences, businesses can foster deeper connections with their customers, leading to increased engagement, loyalty, and advocacy.

Objective

- Deeper understanding of the customer base by categorizing customers into distinct groups or segments. These segments are identified based on various factors such as demographics (e.g., age, education, marital status), purchasing behavior (e.g., types of products purchased, frequency of purchases, amount spent), and response to promotional campaigns (e.g., acceptance of offers, engagement with marketing channels).
- By segmenting customers, businesses can uncover patterns and trends within their customer base, allowing for more targeted and personalized marketing strategies. This analysis enables businesses to tailor their products, services, and marketing efforts to meet the specific needs, preferences, and behaviors of different customer segments. Additionally, customer segmentation facilitates the identification of high-value customers, as well as opportunities for customer retention and acquisition.
- Overall, the goal of customer segmentation analysis is to enhance customer satisfaction, improve marketing effectiveness, and ultimately drive business growth by better understanding and catering to the diverse needs of the customer base.

Importance of Customer Personality Analysis

Understanding Customer Needs:

Personalization

Customers today expect personalized experiences tailored to their unique preferences and requirements. By understanding their needs, businesses can create customized offerings that resonate with individual customers, leading to increased satisfaction and loyalty.

Anticipating Demand

By analyzing past purchasing patterns and gathering feedback, businesses can anticipate future demand and proactively address customer needs before they arise. This not only enhances the customer experience but also drives revenue growth through repeat business and referrals.

Analyzing Customer Behaviors:

Purchase Journey

Every customer follows a distinct journey from awareness to purchase and beyond. By mapping these journeys and identifying key touchpoints, businesses can optimize their marketing efforts to guide customers seamlessly through the sales funnel.

Channel Preferences

Different customers prefer to engage with brands through different channels, whether it's social media, email, or in-person interactions. By analyzing customer behaviors across these channels, businesses can tailor their marketing strategies to meet customers where they are most active.

Addressing Customer Concerns:

Feedback Loop

Customer feedback is a valuable source of insights into their concerns, pain points, and areas for improvement. By actively soliciting and responding to feedback, businesses can demonstrate their commitment to customer satisfaction and foster trust and loyalty.

Resolving Issues

Inevitably, customers will encounter issues or challenges along their journey. By promptly addressing these concerns and providing solutions, businesses can turn potentially negative experiences into opportunities to delight customers and strengthen relationships.

Exploratory Data Analysis (EDA)

Understanding the Dataset

Attributes - People

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	Complain
Customer's unique identifier	Customer's birth year	Customer's education level	Customer's Marital Status	Customer's yearly household income	Number of children in customer's household	Number of teenagers in customer's household	Date of customer's enrollment with the company	Number of days since customer's last purchase	1 if the customer complained in the last 2 years, 0 otherwise

Attributes - Products

MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
Amount spent on wine in last 2 years	Amount spent on fruits in last 2 years	Amount spent on meat in last 2 years	Amount spent on fish in last 2 years	Amount spent on sweets in last 2 years	Amount spent on gold in last 2 years

Understanding the Dataset

Attributes - Promotion

NumDealsPurchases	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	Response
Number of purchases made with a discount	1 if customer accepted the offer in the 1st campaign, 0 otherwise	1 if customer accepted the offer in the 2nd campaign, 0 otherwise	1 if customer accepted the offer in the 3rd campaign, 0 otherwise	1 if customer accepted the offer in the 4th campaign, 0 otherwise	1 if customer accepted the offer in the 5th campaign, 0 otherwise	1 if customer accepted the offer in the last campaign, 0 otherwise

Attributes - Place

NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth
Number of purchases made through the company's website	Number of purchases made using a catalogue	Number of purchases made directly in stores	Number of visits to company's website in the last month

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in customer segmentation analysis as it helps to understand the underlying patterns and relationships within the dataset. Here's a brief overview of the EDA process for customer segment analysis:

- **Understanding the Dataset:**
  - The dataset contains information about customers including demographic details, purchasing behavior, and marketing campaign responses.
- **Data Cleaning and Preprocessing:**
  - Customer segment analysis involves breaking down your customer base into distinct groups based on various characteristics such as demographics, purchasing behavior, and preferences.
  - The process of data gathering for customer segment analysis entails collecting comprehensive information about your customers from diverse sources. This includes both internal sources, such as CRM (Customer Relationship Management) systems and transactional databases, as well as external sources like surveys, social media, and market research reports.
  - It's crucial to ensure the quality and accuracy of the data collected includes Data cleaning involves with Remove or impute missing values, Check for duplicates and remove them if necessary, Convert categorical variables into a suitable format for analysis and Scale numerical variables if needed.

Descriptive Statistics:

- Calculate basic statistics for numerical columns like mean, median, minimum, maximum, and standard deviation to understand the distribution and range of values.
- For categorical columns like Education and Marital_Status, calculate frequency counts to understand the distribution of categories.

Exploratory Analysis(Visualization of Data):

- Visualize the distribution of numerical variables using histograms, boxplots, or density plots.
- Explore the relationship between numerical variables using scatter plots or correlation matrices.
- Analyze the distribution of categorical variables using bar plots or pie charts.
- Investigate any potential trends or patterns in the data.

Feature Engineering:

- Create new features, such as calculating total spending by summing up the spending columns.
- Encode categorical variables for modeling purposes.

Correlation Analysis:

- Calculate correlations between numerical variables to identify any significant relationships.
- Visualize correlations using a heatmap.

Target Variable Analysis:

- Explore the distribution of the target variable (Response) to understand class imbalance.
- Visualize the relationship between the target variable and other features.

- **Outlier Detection:**
  - Identify outliers in numerical variables using statistical methods like z-score or visualization techniques like boxplots.
  - Decide whether to remove outliers or transform the data.
- **Feature Importance:**
  - Determine which features are most important for predicting the target variable using techniques like feature importance scores or model-based feature selection.
- **Modeling:**
  - Based on the insights gained from EDA, select appropriate modeling techniques such as regression, classification, or clustering.
  - Split the dataset into training and testing sets.
  - Train and evaluate the model performance using appropriate metrics.

Data Cleaning and Preprocessing:

```
customerdata.columns
```

```
Index(['ID', 'Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome',
       'Teenhome', 'Dt_Customer', 'Recency', 'MntWines', 'MntFruits',
       'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Complain', 'Response'],
      dtype='object')
```

```
#Import Libraries
import pandas as pd
import numpy as np

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier

#ignore warnings
import warnings
warnings.filterwarnings('ignore')
```

```
customerdata = pd.read_excel("../DS Project1/marketing_campaign.xlsx")
```

```
customerdata
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3
0	5524	1957	Graduation	Single	58138.0	0	0	2012-09-04	58	635	...	7	0
1	2174	1954	Graduation	Single	46344.0	1	1	2014-03-08	38	11	...	5	0
2	4141	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	426	...	4	0
3	6182	1984	Graduation	Together	26646.0	1	0	2014-02-10	26	11	...	6	0
4	5324	1981	PhD	Married	58293.0	1	0	2014-01-19	94	173	...	5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	10870	1967	Graduation	Married	61223.0	0	1	2013-06-13	46	709	...	5	0
2236	4001	1946	PhD	Together	64014.0	2	1	2014-06-10	56	406	...	7	0
2237	7270	1981	Graduation	Divorced	56981.0	0	0	2014-01-25	91	908	...	6	0
2238	8235	1956	Master	Together	69245.0	0	1	2014-01-24	8	428	...	3	0
2239	9405	1954	PhD	Married	52869.0	1	1	2012-10-15	40	84	...	7	0

2240 rows × 29 columns

```
# 1. Data Cleaning
# Check for missing values
missing_values = customerdata.isnull().sum()
print("Missing Values:\n", missing_values)
```

```
Missing Values:
 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
```

Descriptive Statistics:

Descriptive statistics provide a snapshot of the central tendency, variability, and distribution of the variables in the dataset, which is essential for understanding the characteristics of the customer base and informing further analysis, such as customer segmentation

```
customerdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   datetime64[ns]
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: datetime64[ns](1), float64(1), int64(25), object(2)
memory usage: 507.6+ KB
```

```
# Visualize distributions of numerical features
numerical_features = customerdata.select_dtypes(include=['int64', 'float64']).columns.tolist()
print("\nVisualizing distributions of numerical features:")
for feature in numerical_features:
    plt.figure(figsize=(8, 6))
    sns.histplot(customerdata[feature], kde=True)
    plt.title(f'Distribution of {feature}')
    plt.xlabel(feature)
    plt.ylabel('Frequency')
    plt.show()
```

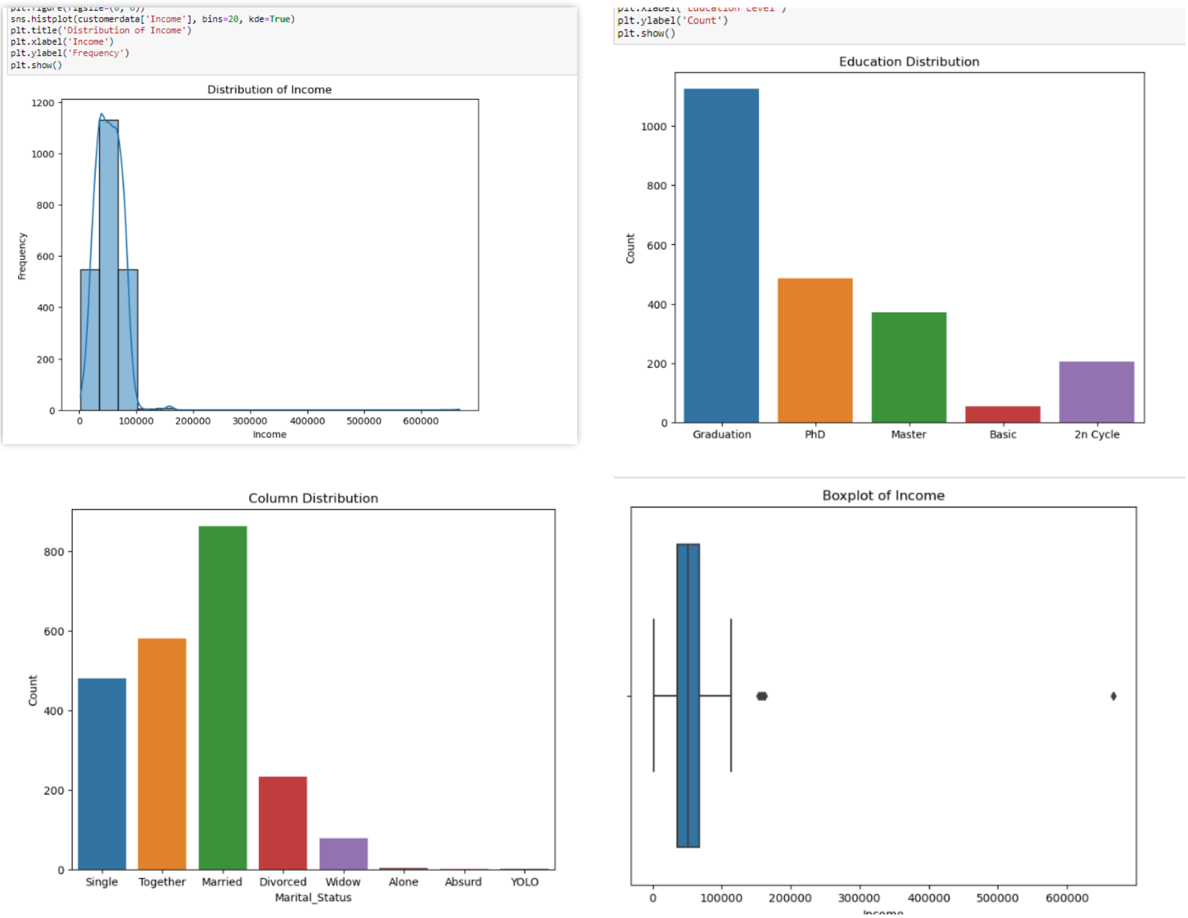
```
# Visualize distributions of categorical features
categorical_features = customerdata.select_dtypes(include=['object']).columns.tolist()
print("\nVisualizing distributions of categorical features:")
for feature in categorical_features:
    plt.figure(figsize=(8, 6))
    sns.countplot(customerdata, x=feature)
    plt.title(f'Distribution of {feature}')
    plt.xlabel(feature)
    plt.ylabel('Count')
    plt.xticks(rotation=45)
    plt.show()
```

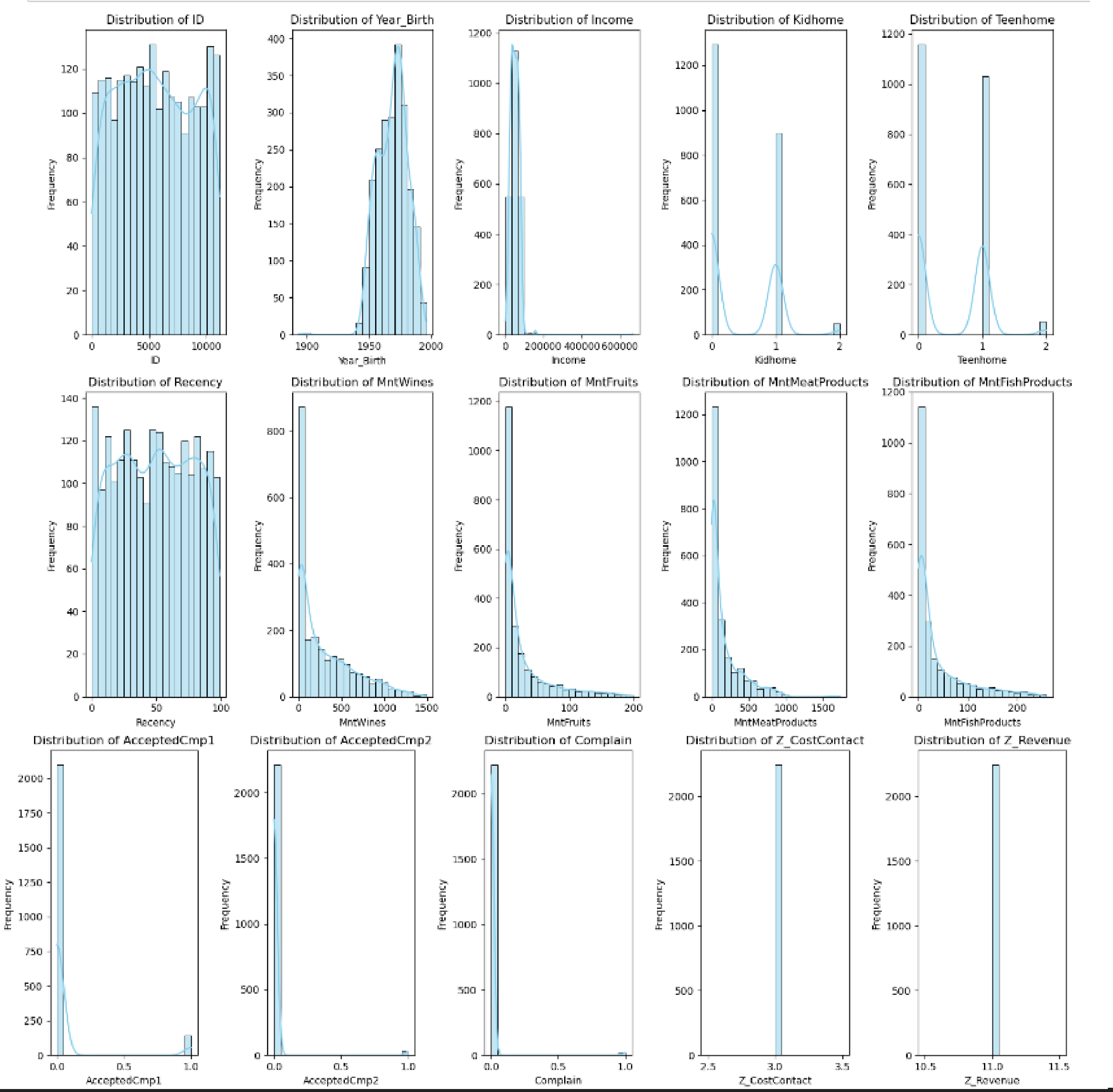
```
# 2. Descriptive Statistics
descriptive_stats = customerdata.describe().T
print("Descriptive Statistics:\n", descriptive_stats)
```

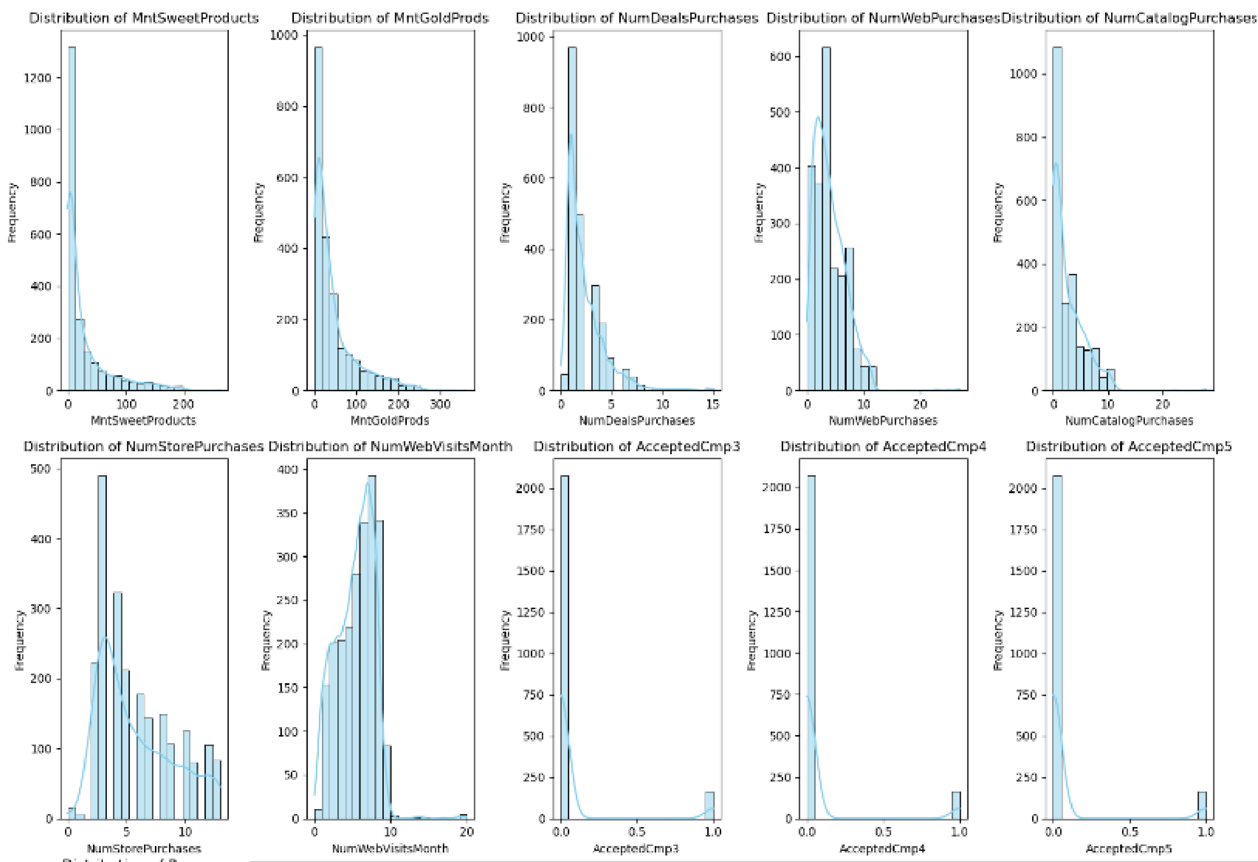
```
Descriptive Statistics:
                       count          mean           std     min       25%  \
ID                   2240.0   5592.159821   3246.662198     0.0   2828.25
Year_Birth           2240.0   1968.805804     11.984069  1893.0   1959.00
Income               2240.0  52247.251354  25037.797168  1730.0  35538.75
Kidhome              2240.0      0.444196      0.538398     0.0      0.00
Teenhome             2240.0      0.506250      0.544538     0.0      0.00
Recency              2240.0     49.109375     28.962453     0.0     24.00
MntWines             2240.0    303.935714    336.597393     0.0     23.75
MntFruits            2240.0     26.302232     39.773434     0.0      1.00
MntMeatProducts      2240.0    166.950000    225.715373     0.0     16.00
MntFishProducts      2240.0     37.525446     54.628979     0.0      3.00
MntSweetProducts     2240.0     27.062946     41.280498     0.0      1.00
MntGoldProds         2240.0     44.021875     52.167439     0.0      9.00
NumDealsPurchases    2240.0      2.325000      1.932238     0.0      1.00
NumWebPurchases      2240.0      4.084821      2.778714     0.0      2.00
NumCatalogPurchases  2240.0      2.662054      2.923101     0.0      0.00
NumStorePurchases    2240.0      5.790179      3.250958     0.0      3.00
NumWebVisitsMonth    2240.0      5.316518      2.426645     0.0      3.00
AcceptedCmp3         2240.0      0.072768      0.259813     0.0      0.00
AcceptedCmp4         2240.0      0.074554      0.262728     0.0      0.00
AcceptedCmp5         2240.0      0.072768      0.259813     0.0      0.00
AcceptedCmp1         2240.0      0.064286      0.245316     0.0      0.00
AcceptedCmp2         2240.0      0.013393      0.114976     0.0      0.00
Complain             2240.0      0.009375      0.096391     0.0      0.00
Response             2240.0      0.149107      0.356274     0.0      0.00

                         50%       75%       max
ID                    5458.5   8427.75   11191.0
Year_Birth            1970.0   1977.00    1996.0
Income               51741.5  68289.75  666666.0
Kidhome                  0.0      1.00       2.0
Teenhome                 0.0      1.00       2.0
Recency                 49.0     74.00      99.0
MntWines               173.5    504.25    1493.0
MntFruits                8.0     33.00     199.0
MntMeatProducts         67.0    232.00    1725.0
MntFishProducts         12.0     50.00     259.0
MntSweetProducts         8.0     33.00     263.0
MntGoldProds            24.0     56.00     362.0
NumDealsPurchases        2.0      3.00      15.0
NumWebPurchases          4.0      6.00      27.0
NumCatalogPurchases      2.0      4.00      28.0
NumStorePurchases        5.0      8.00      13.0
NumWebVisitsMonth        6.0      7.00      20.0
AcceptedCmp3             0.0      0.00       1.0
AcceptedCmp4             0.0      0.00       1.0
AcceptedCmp5             0.0      0.00       1.0
AcceptedCmp1             0.0      0.00       1.0
AcceptedCmp2             0.0      0.00       1.0
Complain                 0.0      0.00       1.0
Response                 0.0      0.00       1.0
```

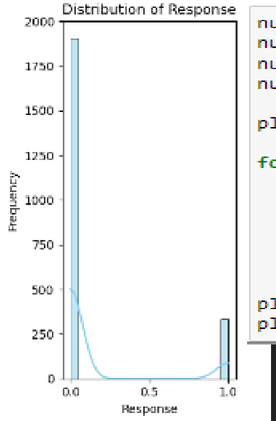
Exploratory Analysis: Visualization of data

- Data visualization plays a crucial role in EDA by providing intuitive insights into the dataset. Visualizations such as heatmaps, scatter plots, and bar charts help identify trends, patterns, and anomalies in the data, making it easier to communicate findings to stakeholders.










```python
numerical_columns = customerdata.select_dtypes(include=['int64', 'float64']).columns
num_cols_count = len(numerical_columns)
num_cols_per_row = 5  # Number of numerical columns per row
num_rows = (num_cols_count - 1) // num_cols_per_row + 1

plt.figure(figsize=(15, 5*num_rows))

for i, column in enumerate(numerical_columns):
    plt.subplot(num_rows, num_cols_per_row, i + 1)
    sns.histplot(customerdata[column], kde=True, bins=20, color='skyblue')
    plt.title(f'Distribution of {column}')
    plt.xlabel(column)
    plt.ylabel('Frequency')

plt.tight_layout()
plt.show()
```

Feature Engineering

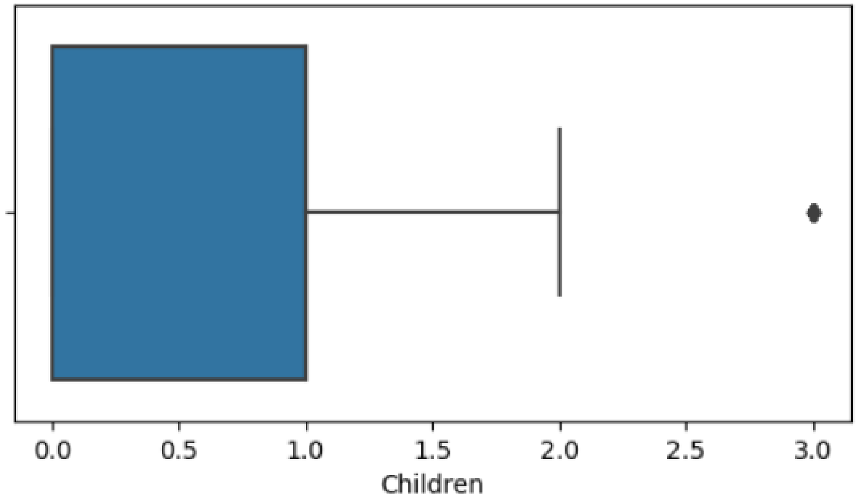
Feature engineering plays a critical role in the success of machine learning models by improving their interpretability, generalization, and performance. It requires a deep understanding of the data and domain expertise to effectively create, transform, and select features that capture the underlying patterns and relationships in the data

Feature Creation: This involves generating new features from the existing dataset or domain knowledge.

```
# families with kids
customerdata['Children'] = customerdata['Kidhome'] + customerdata['Teenhome']
```

```
plt.figure(figsize=(6,3))
sns.boxplot(data=customerdata, x='Children')
```

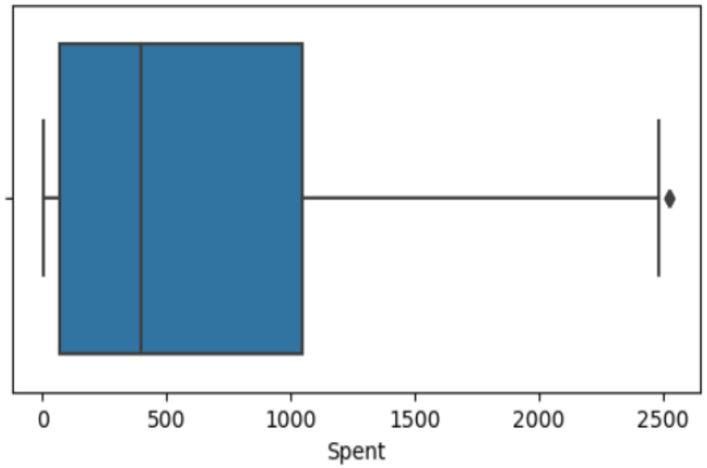
<Axes: xlabel='Children'>




```
# Feature engineering
# Example: Creating a new column 'Spent' by summing up all spending columns
# amount spent on all types of purchases in last 2 years
customerdata['Spent'] = customerdata['MntWines'] + customerdata['MntFruits'] + customerdata['MntMeatProducts'] + \
    customerdata['MntFishProducts'] + customerdata['MntSweetProducts'] + customerdata['MntGoldProds']
```

```
plt.figure(figsize=(6,3))
sns.boxplot(data=customerdata, x='Spent')
```

<Axes: xlabel='Spent'>



Feature Engineering:

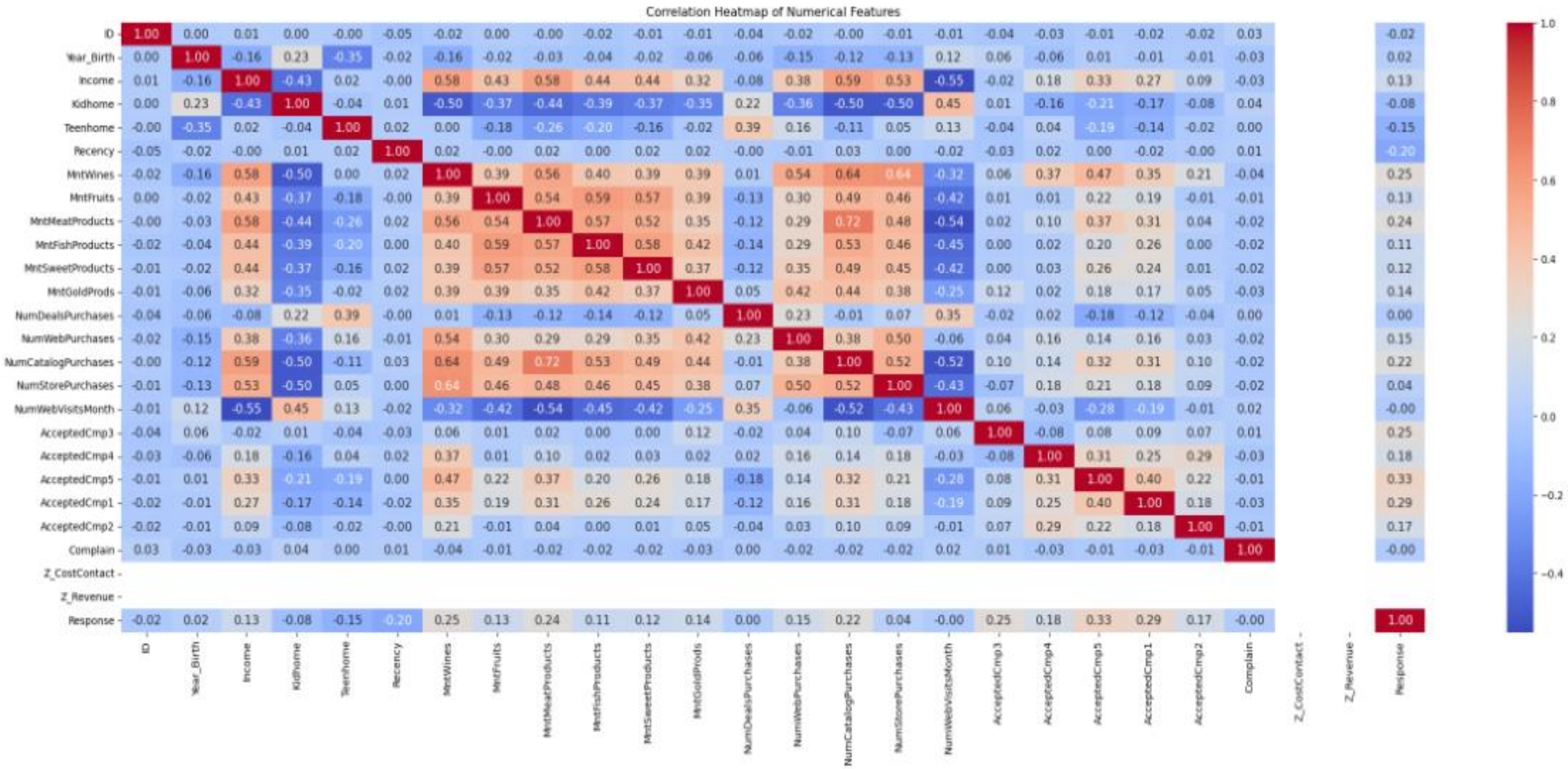
Removing	Encoding	Encoding
Removing Unwanted features	Label Encoding	Label Encoding

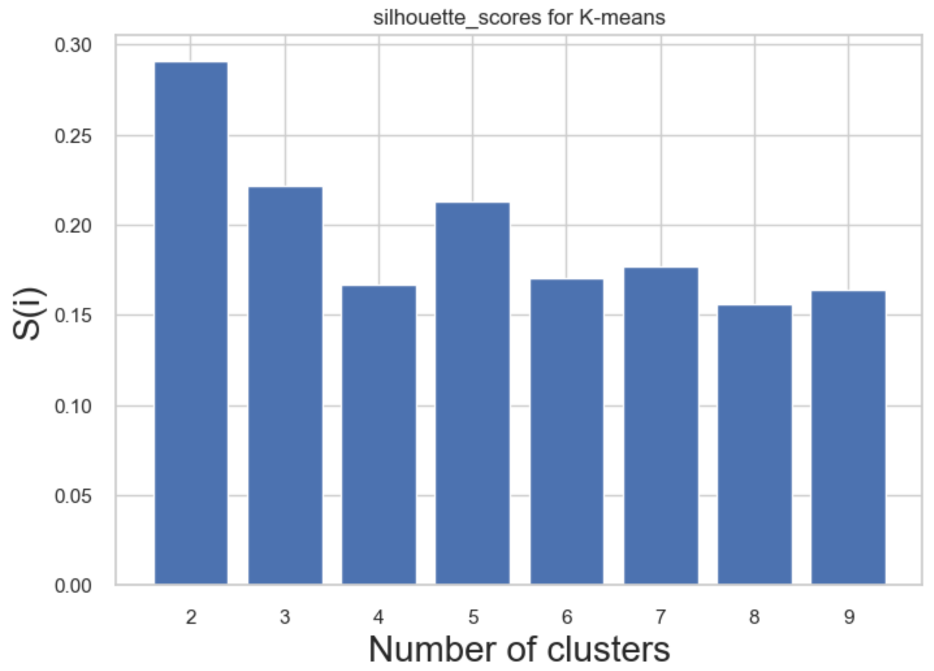
Correlation Analysis:

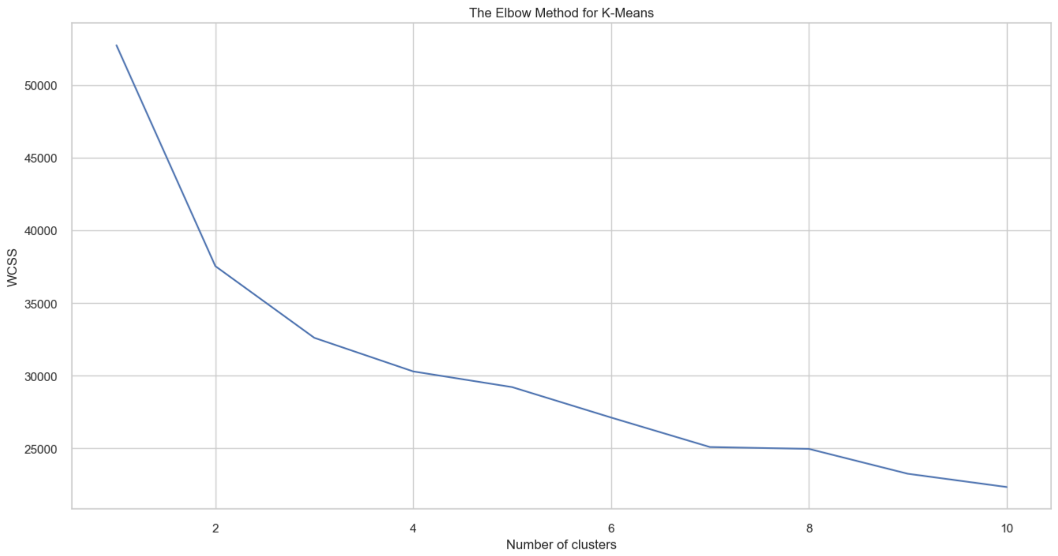
Correlation analysis is a powerful tool for exploring the associations between variables in a dataset, providing valuable insights into the underlying structure and informing subsequent analysis and modeling decisions.

```
# Visualize correlations between numerical features
print("\nVisualizing correlations between numerical features:")
plt.figure(figsize=(30, 10))
sns.heatmap(customerdata[numerical_features].corr(), annot=True, cmap='coolwarm', fmt=".2f", annot_kws={"size": 12})
plt.title("Correlation Heatmap of Numerical Features")
plt.show()
```

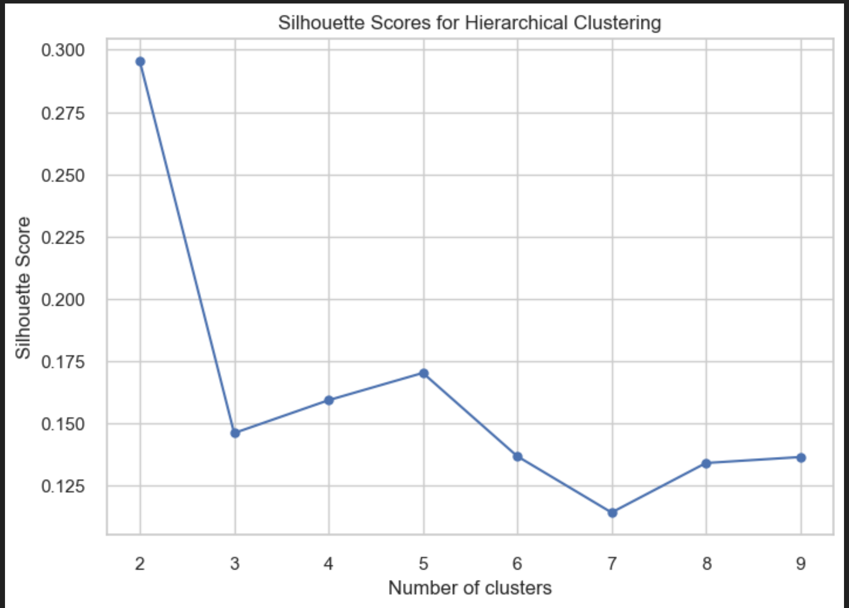
```
Visualizing correlations between numerical features:
```

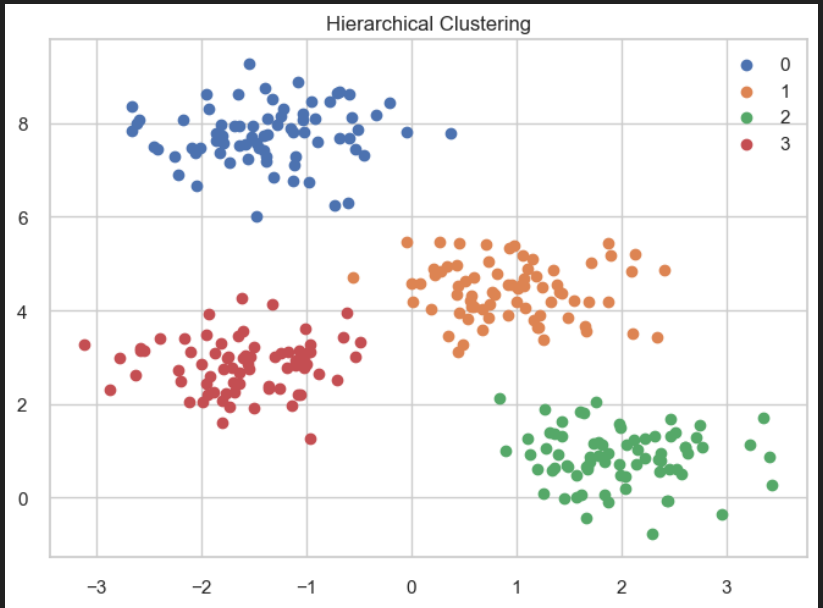






K- Means Clustering





Hierarchical Clustering

Choosing number of clusters

Although score of 3_clusters to 5_clusters is decent silhouette score however we are choosing 3_clusters of K-means .

3

because observations seems to be more evenly distributed among the clusters and making strategy for 3 cluster is much more convenient and silhouette scores are dropping after cluster 3 Scores

Defining Clusters

Cluster 0:
Moderate Spenders with Minimal Family Size

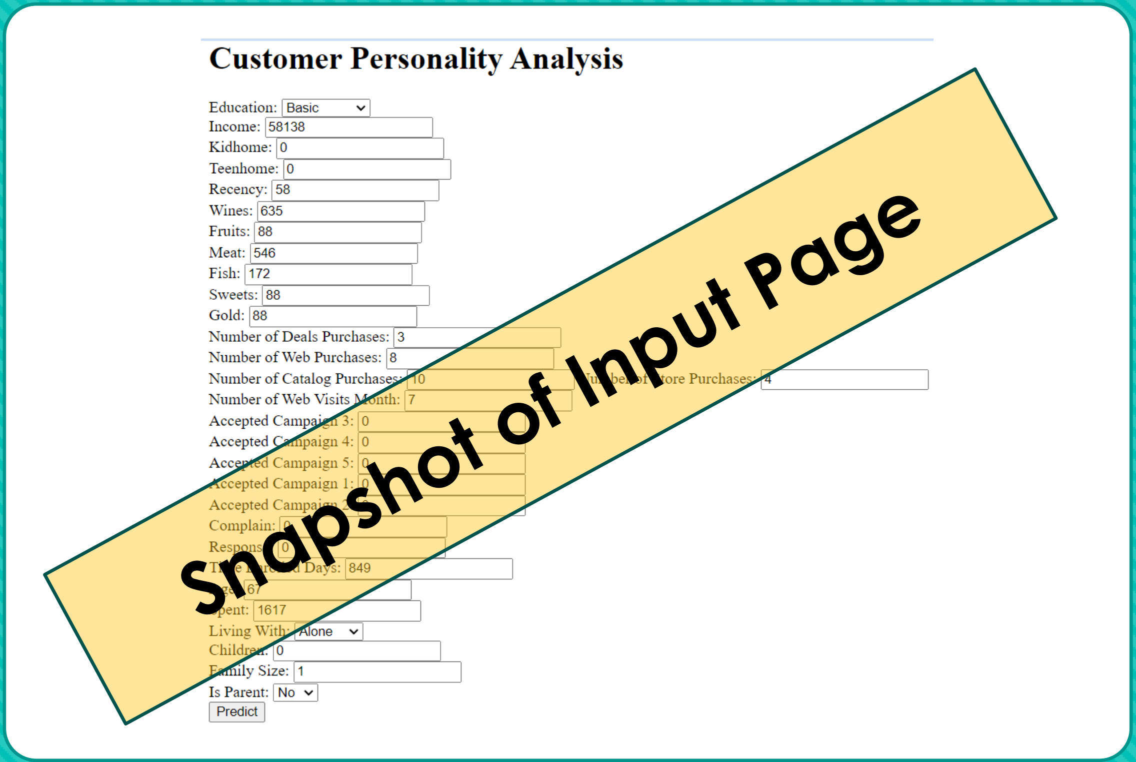
Cluster 1:
Low Income, Low Spenders with Children

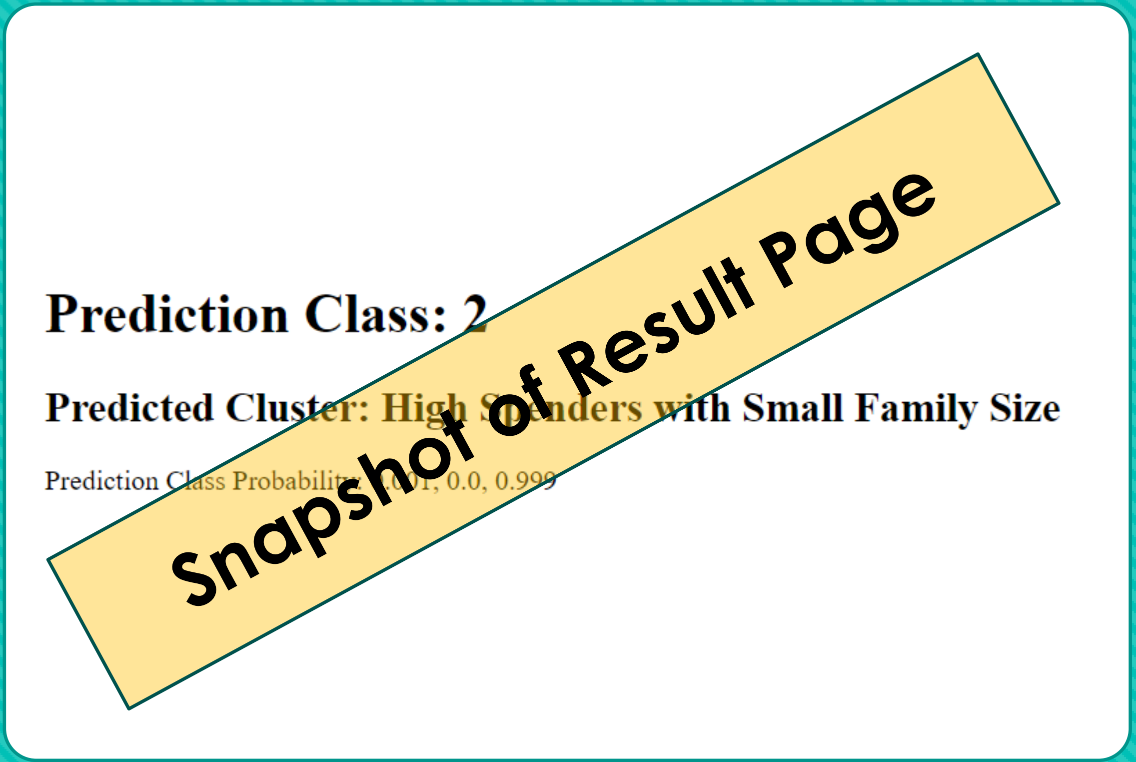
Cluster 2:
High Spenders with Small Family Size

MODEL BUILDING

Model	Accuracy	Mean Cross-Validation Scor
Gradient Boosting	97.99%	96.81%
Random Forest	97.99%	96.75%
Decision Tree	96.20%	95.74%
Naive Bayes	94.41%	94.29%
Logistic Regression	87.70%	87.84%
K-Nearest Neighbors	85.23%	85.21%
Support Vector Machine	78.75%	78.10%

Hence, we will deploy model using Gradient Boosting classifier as Three clusters technique giving 97% accuracy.
Dump the model files for Deployement





Model Deployment

Challenges Faced

Defining clusters using features poses a significant challenge.

Choosing the right number of clusters is proving to be difficult.

Handling high-dimensional data adds another layer of complexity to the clustering process.

Thank you