CUSTOMER SEGMENTATION

TEAM - 3

SLIDES
Content

INTRODUCTION

- In today's data-driven world, understanding our customers' diverse characteristics and behaviors is essential for business success.
- The problem at hand is to cluster customers into distinct groups based on their demographics, behaviors, or purchase patterns.
- Leveraging a Kaggle dataset and utilizing powerful libraries like NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn, we aim to gain valuable insights from customer data.
- Our main focus is to implement the clustering algorithm, allowing us to segment customers effectively.
- Let's delve into the fascinating world of customer segmentation and discover how it can revolutionize our marketing strategies and improve customer engagement.

WHAT IS CUSTOMER SEGMENTATION?

- Dividing customers into distinct groups based on similar characteristics, behaviors, or preferences.
- Enables personalized marketing, improving customer satisfaction and loyalty.
- Optimizes marketing resources, leading to a higher return on investment.
- Enhances customer experience and retention by addressing specific needs.
- Uncovers new business opportunities and potential revenue streams.
- A data-driven approach for making informed decisions.
- Drives market differentiation and targeted marketing campaigns.
- Applicable throughout the customer lifecycle for continuous improvement.

Libraries

- **Numpy :** Used for numerical operations and array manipulations, providing efficient mathematical functions.
- **Pandas:** Utilized for data manipulation and analysis, offering powerful data structures like Data Frames.
- **Matplotlib:** A widely used plotting library for creating static, interactive, and publication quality visualizations.
- **Seaborn:** A data visualization library based on matplotlib, making it easy to create attractive statistical graphics.
- **Sklearn.cluster(kmeans):** Part of scikit learn library, used for implementing K-means clustering algorithm.
- **Sklearn.preprocessing(LabelEncoder):** Part of scikit-learn library, used for encoding categorical variables into numeric format.
- **Datetime:** Used for handling date and time related operations.
- **Plotly.express(px):** USed for interactive visualizations and creating expressive plots, graphs, and charts.

CUSTOMER SEGMENTATION PROCESS

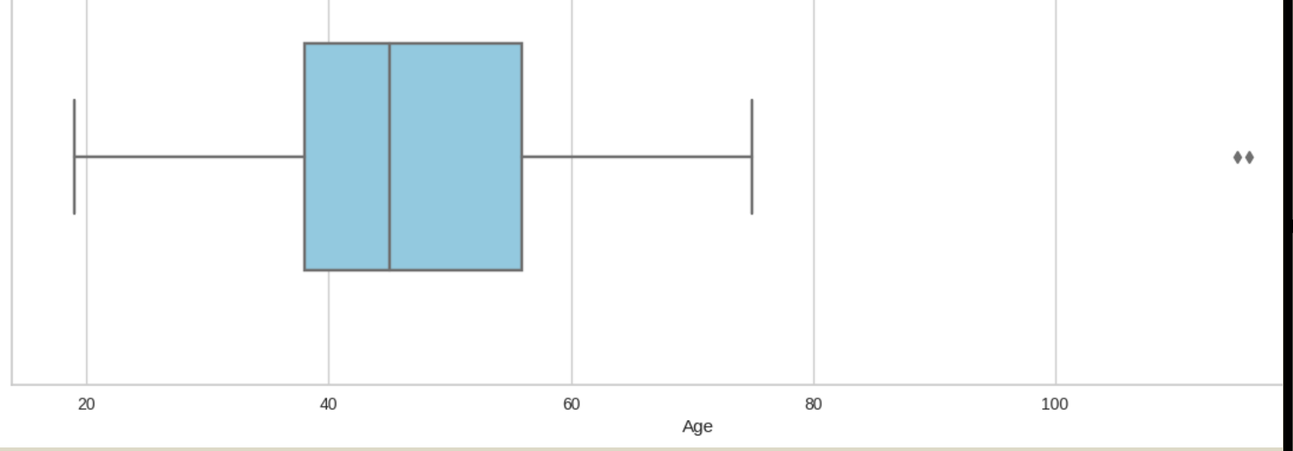
- **Data Collection:** Gather relevant customer data from various sources, such as demographics, purchase history, website interactions, and customer feedback.
- **Data Preprocessing:** Clean and prepare the data by handling missing values, removing duplicates, and converting categorical variables into a numerical format using techniques like Label Encoding.
- **Feature Engineering:** Extract and create meaningful features from the data that can be used for customer segmentation. For example, calculate customer lifetime value, frequency of purchases, recency of interactions, etc.
- **Exploratory Data Analysis (EDA):** Perform EDA to gain insights into the data and identify potential patterns or clusters that can guide the segmentation process.
- **Selecting Features:** Choose relevant features that are important for the segmentation task and can help in distinguishing different customer groups effectively.
- **Choosing the Clustering Algorithm:** Select an appropriate machine learning clustering algorithm.
- **Interpreting Clusters:** Analyze the resulting clusters to understand the characteristics, behaviors, and preferences of each customer group to understand their purchase pattern.

DATA CLEANING

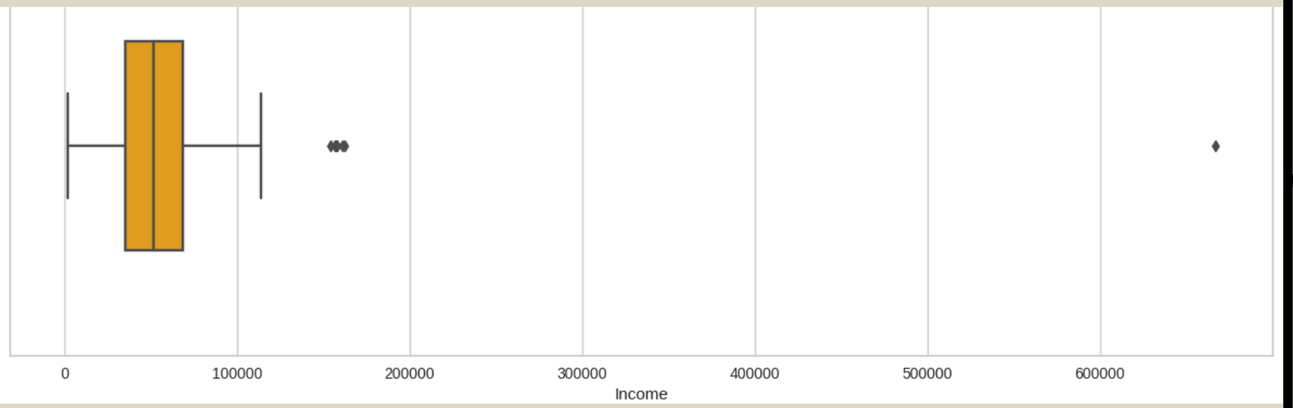
- **Handling Missing Values:** Identify and handle missing data through techniques like imputation (mean, median, or mode) or dropping rows/columns.
- **Removing Duplicates:** Detect and remove duplicate records to avoid bias in analysis and ensure data integrity.
- **Addressing Inconsistent Data:** Standardize data formats, units, and representations for consistency.
- **Dealing with Irrelevant Features:** Remove irrelevant or redundant features that do not contribute to the analysis.

REMOVING OUTLIER

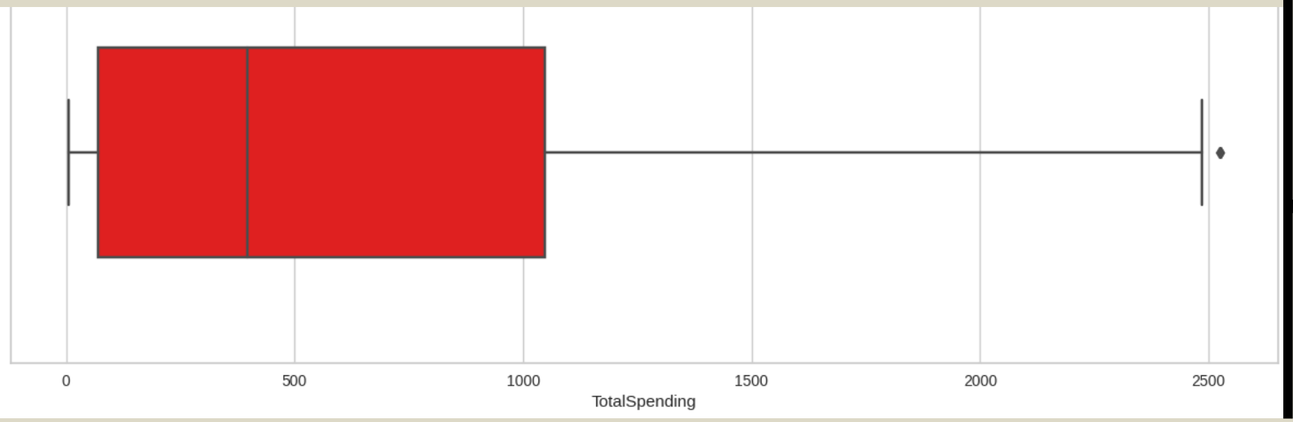
- **Outlier Detection:** Use statistical methods (Z-score, IQR) or visualizations (box plots) to identify outliers.
- **Outlier Handling Strategies:** Decide whether to remove outliers, transform them, or treat them separately based on domain knowledge and impact on the analysis.



OUTLIERS BY AGE



OUTLIERS BY INCOME

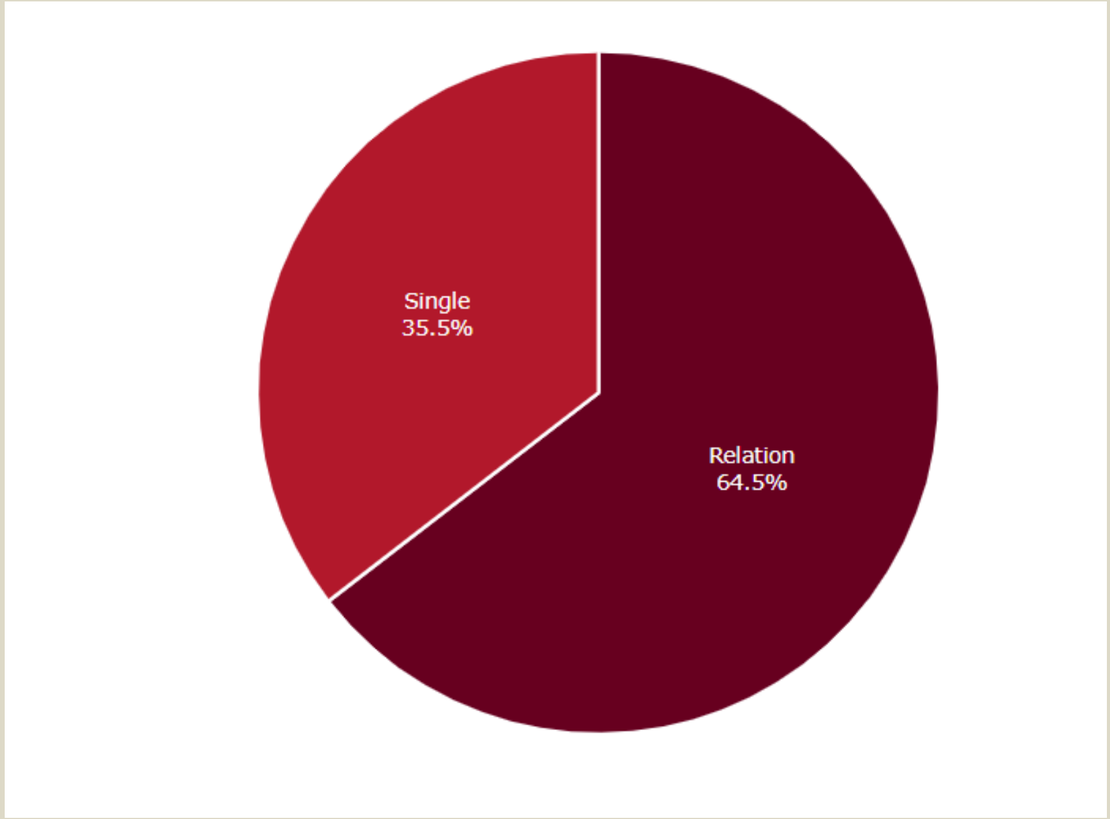


OUTLIERS BY TOTAL SPENDING

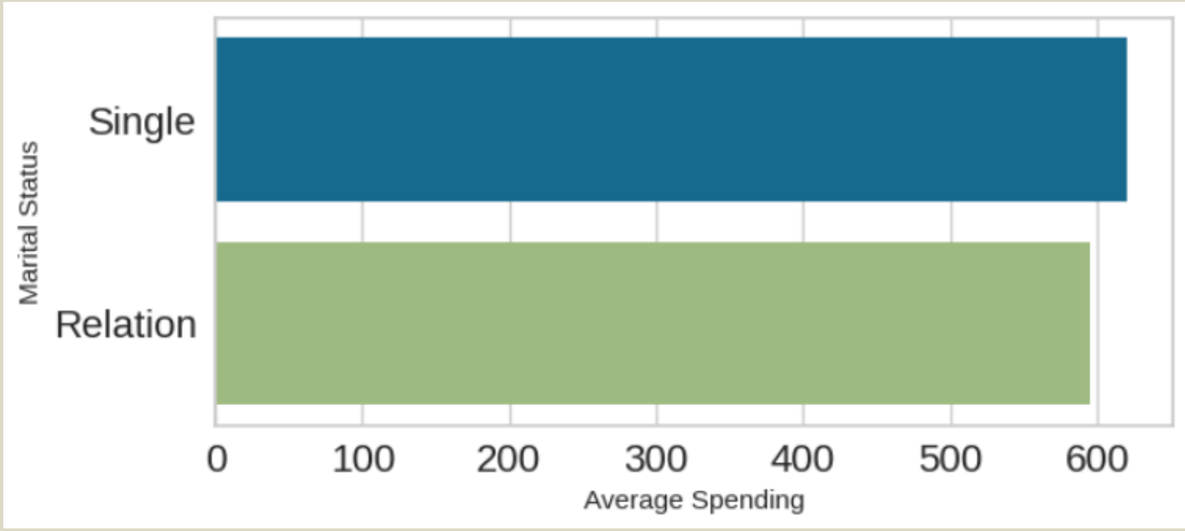
Exploratory Data Analysis (EDA)

- It is the initial and essential step in the data analysis process. It involves visually and statistically exploring a dataset to gain insights, discover patterns, and uncover underlying relationships within the data.
- The primary goal of EDA is to understand the data's structure, identify trends, outliers, and potential issues, which can guide subsequent data processing, modeling, and decision-making

Marital Status

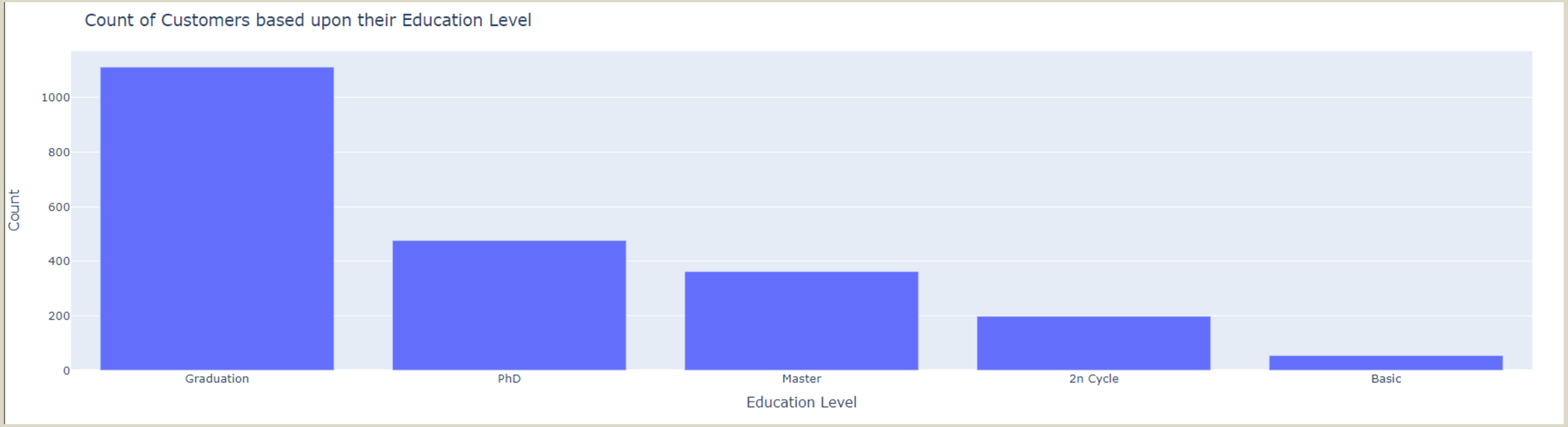


Count

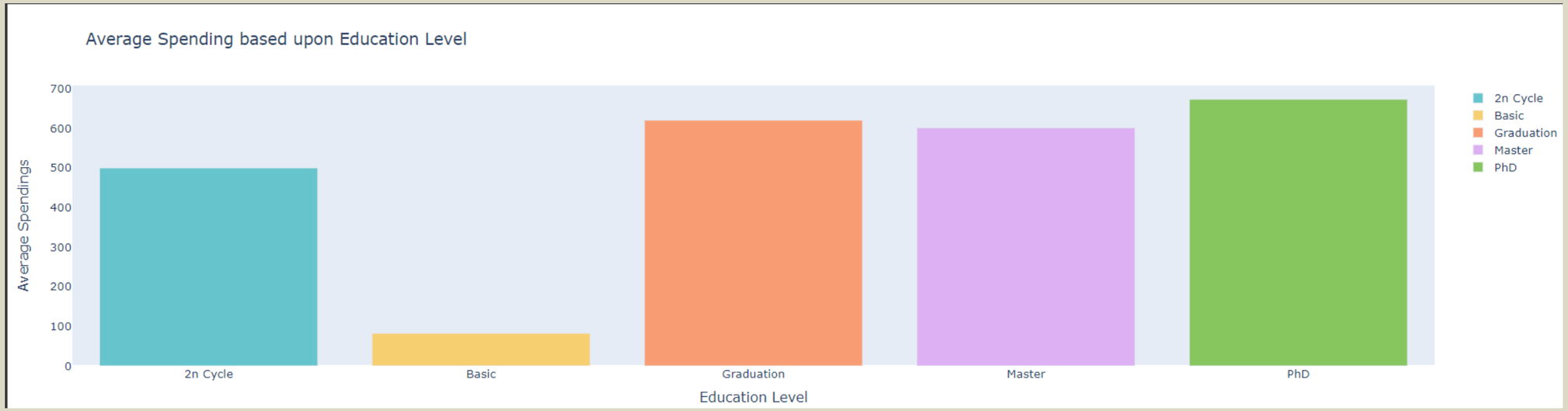


Average Spending

Education Level

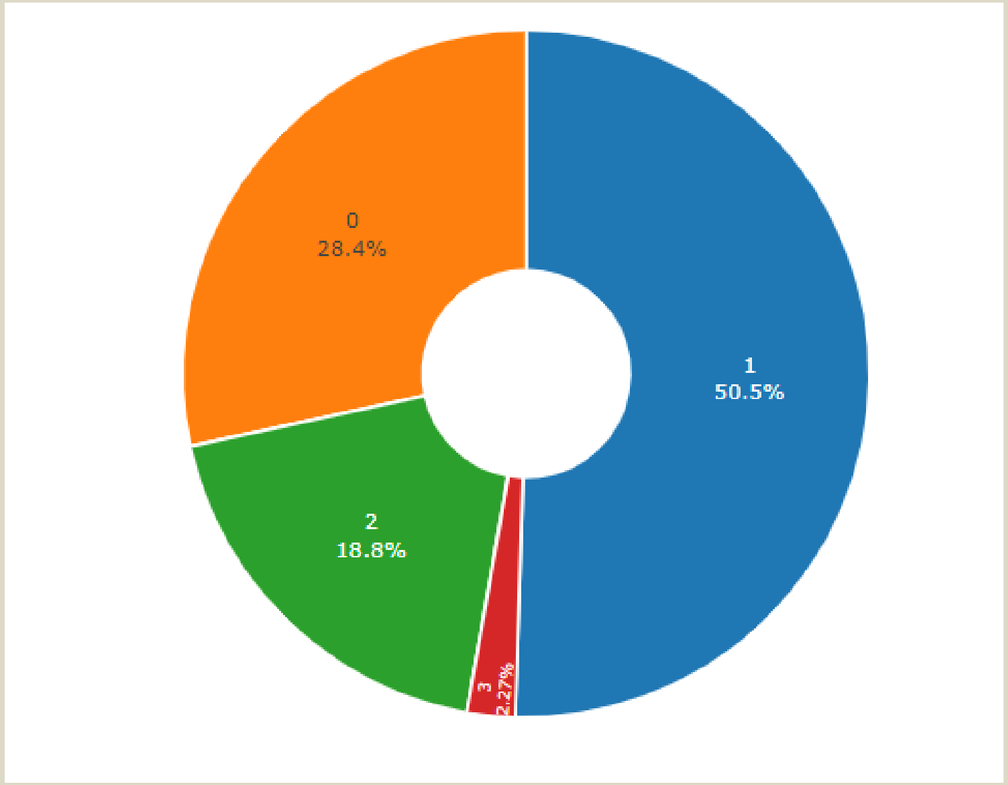


Count

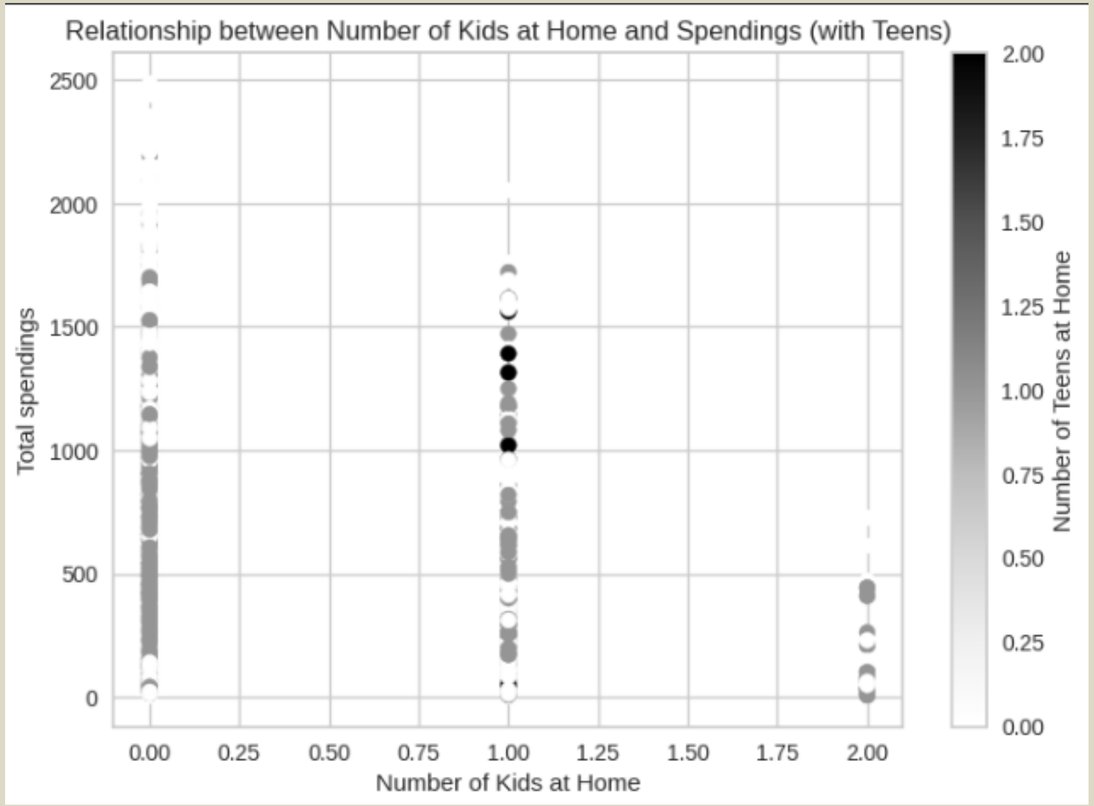


Average Spending

Number of Children

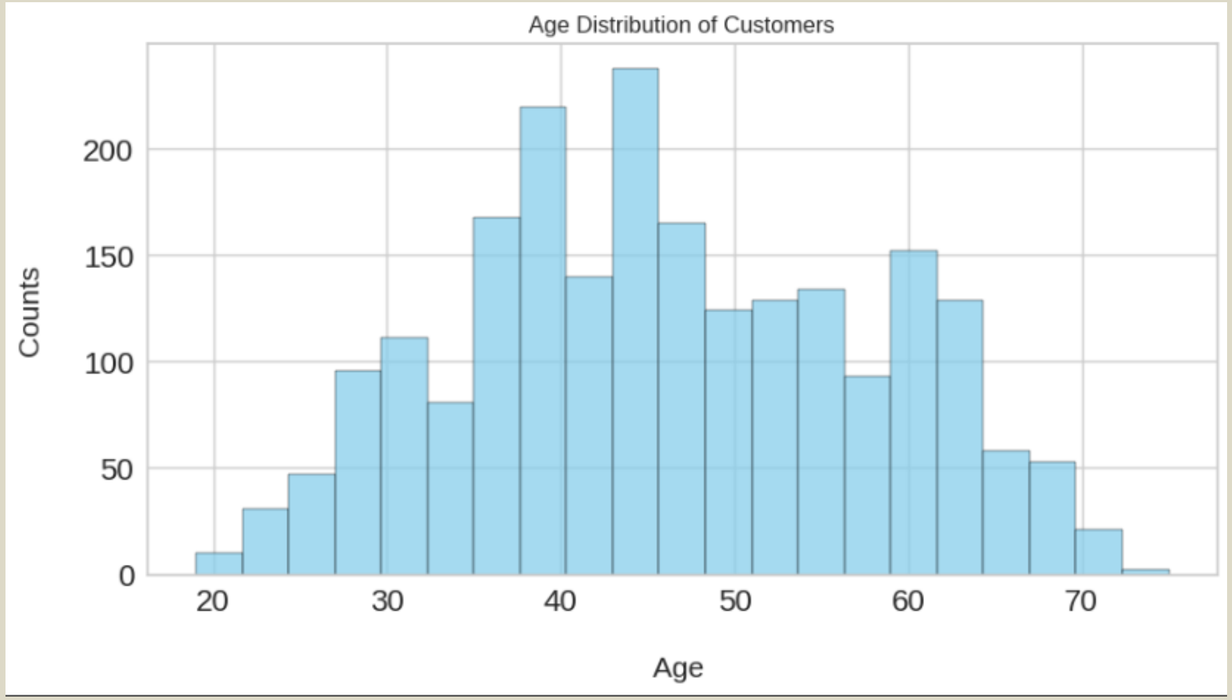


Count

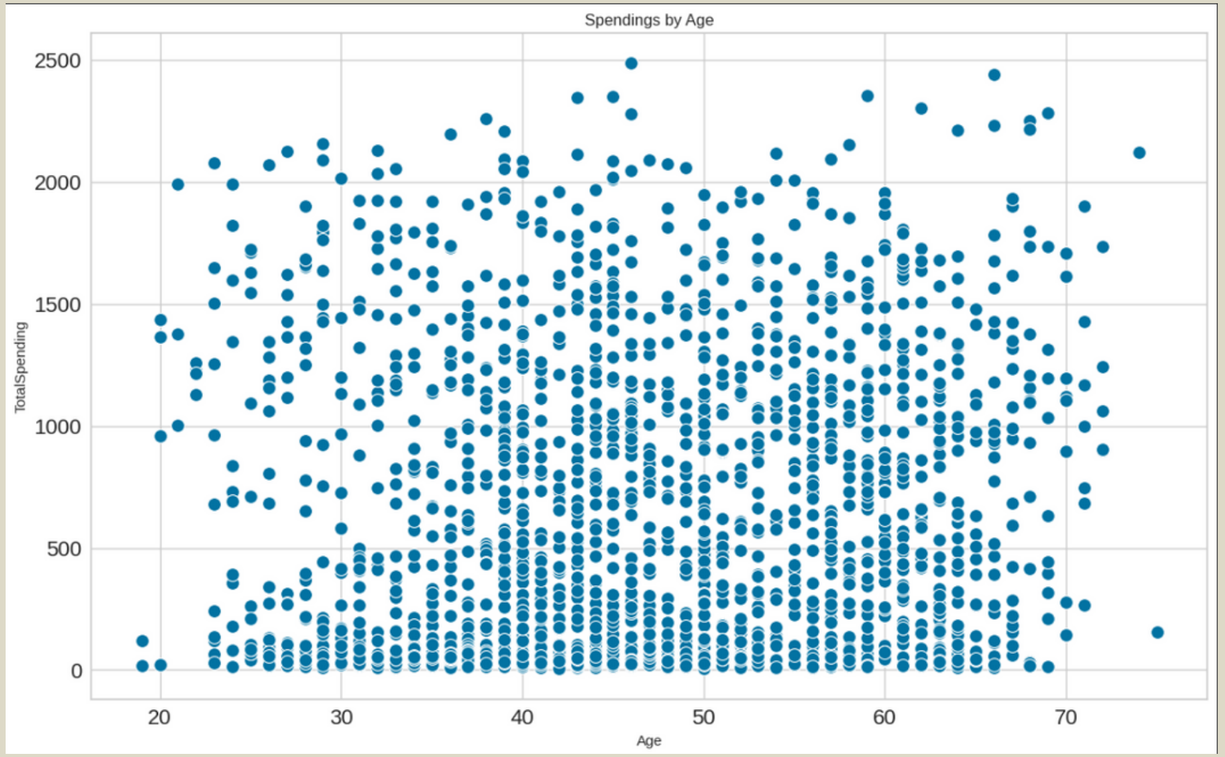


Average Spending

Age Distribution of Customer

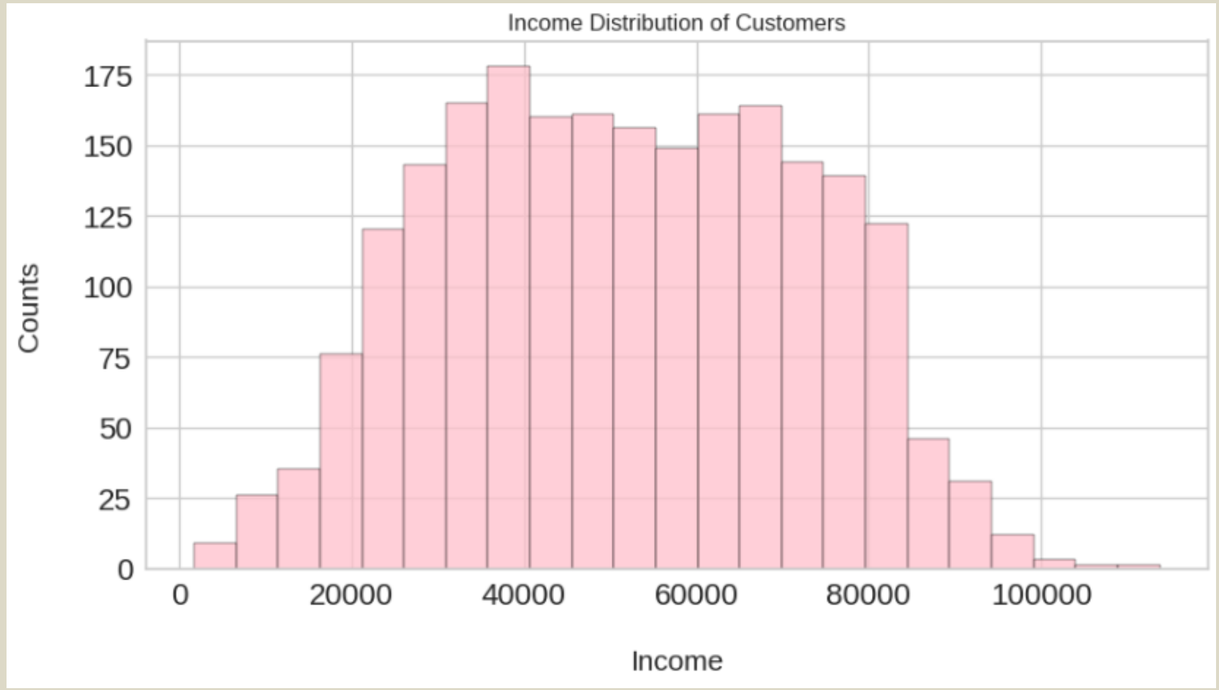


Count

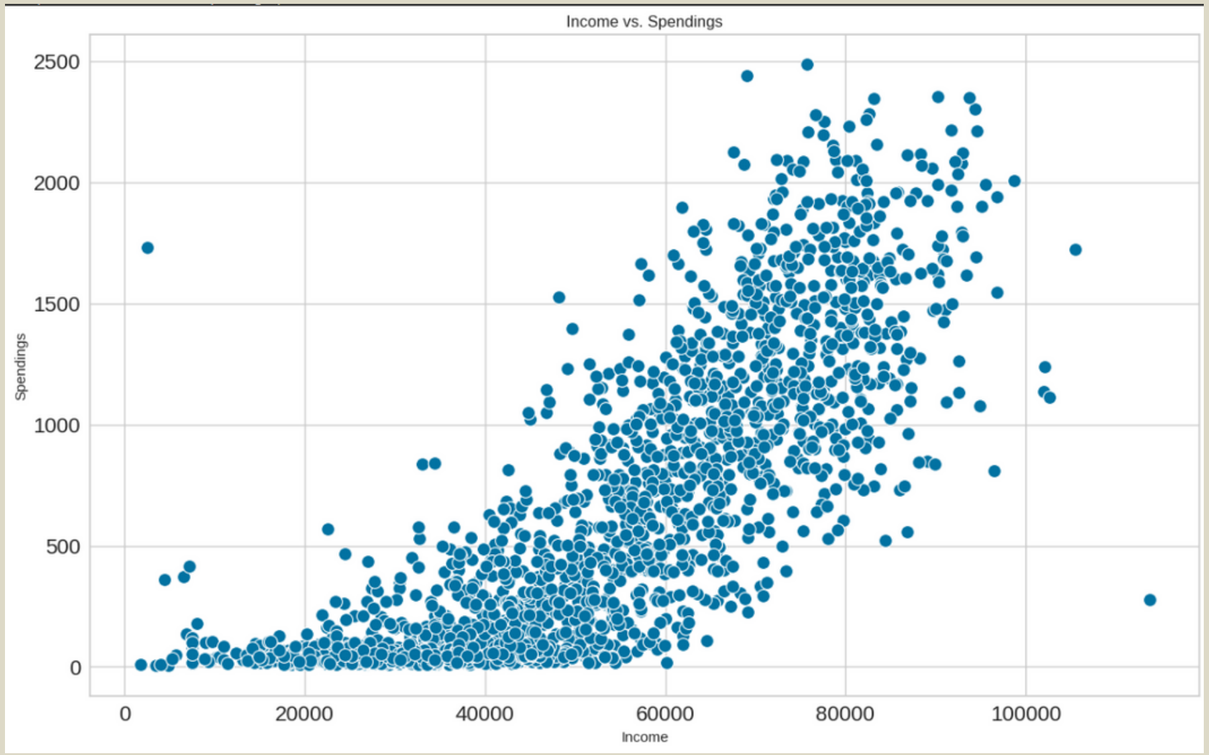


Average Spending

Income Distribution of Customer

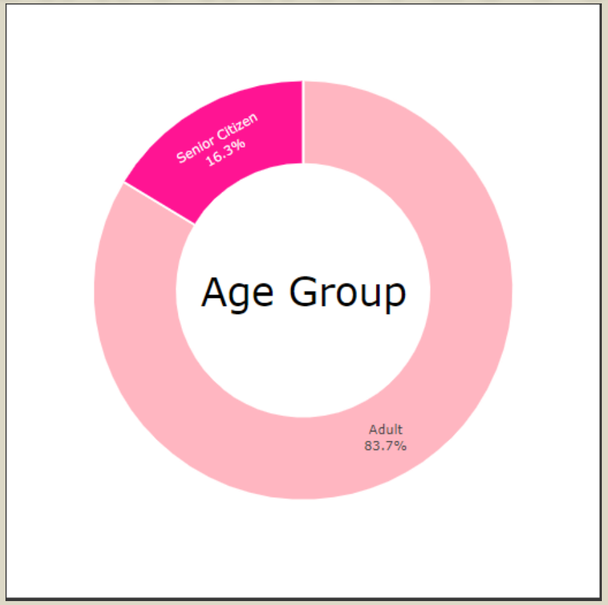


Count

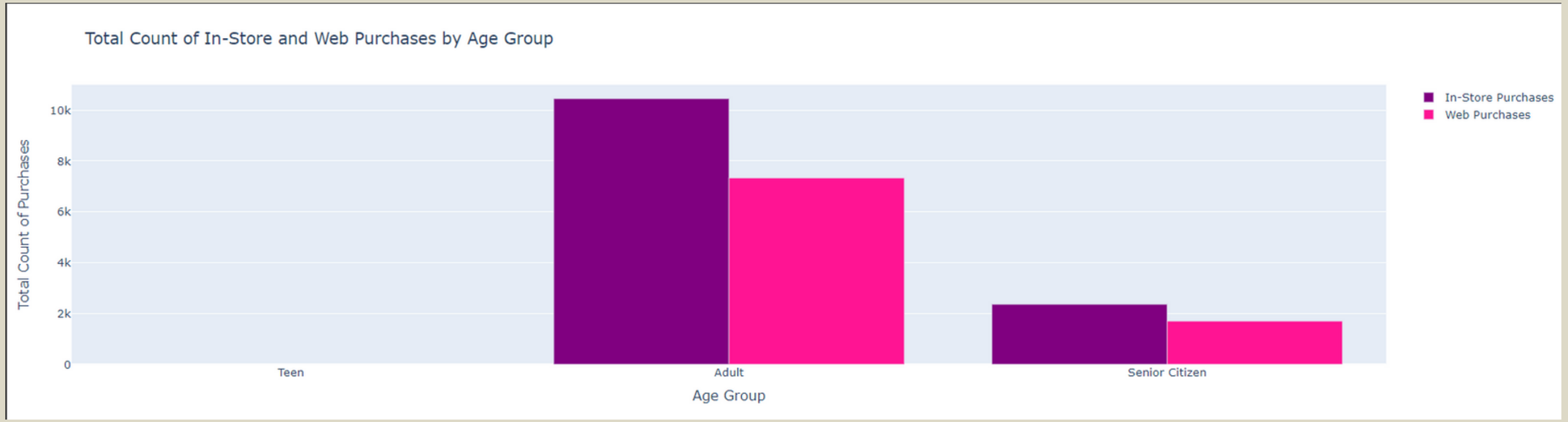


Average Spending

Education Level



Count

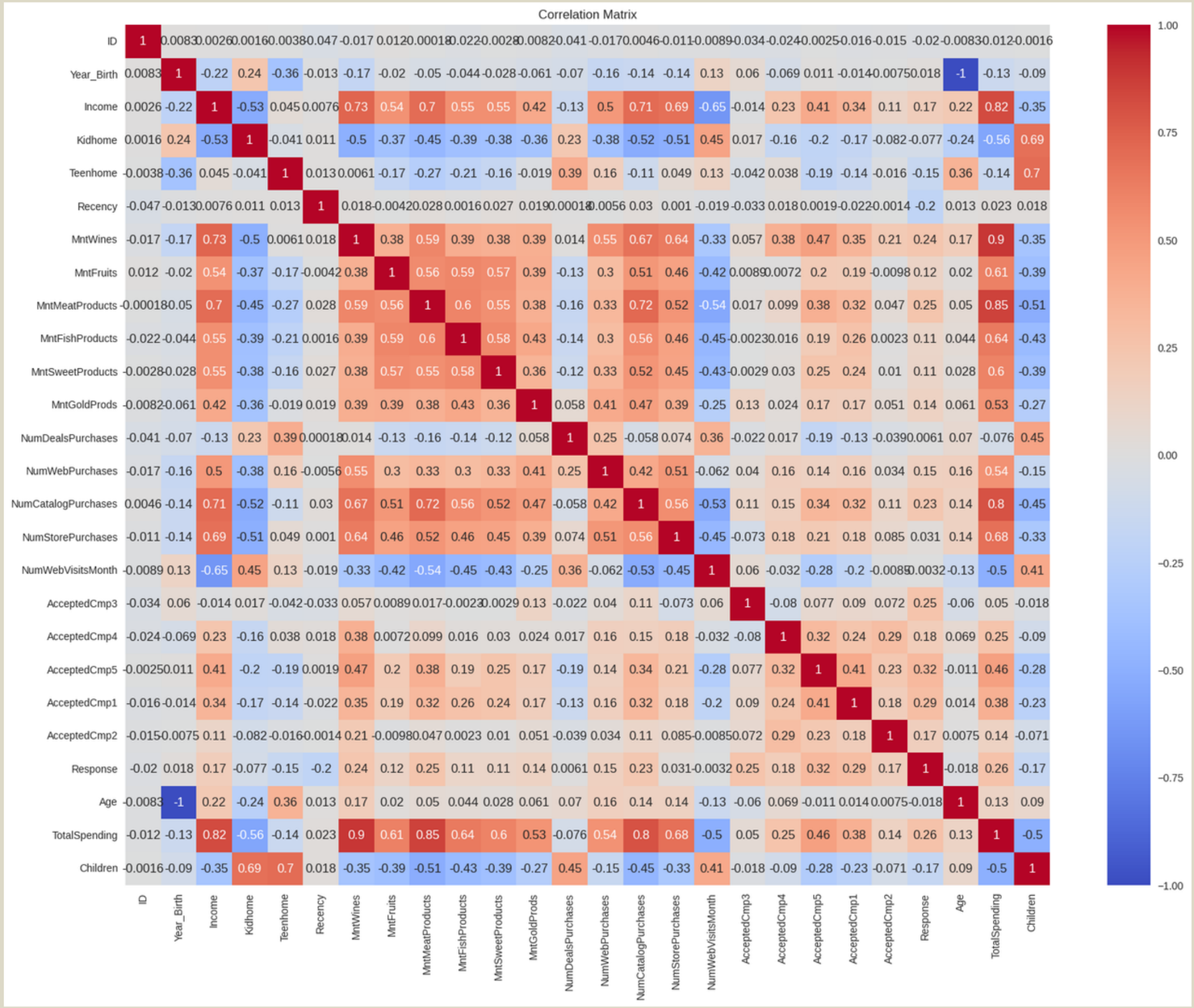


Average Spending

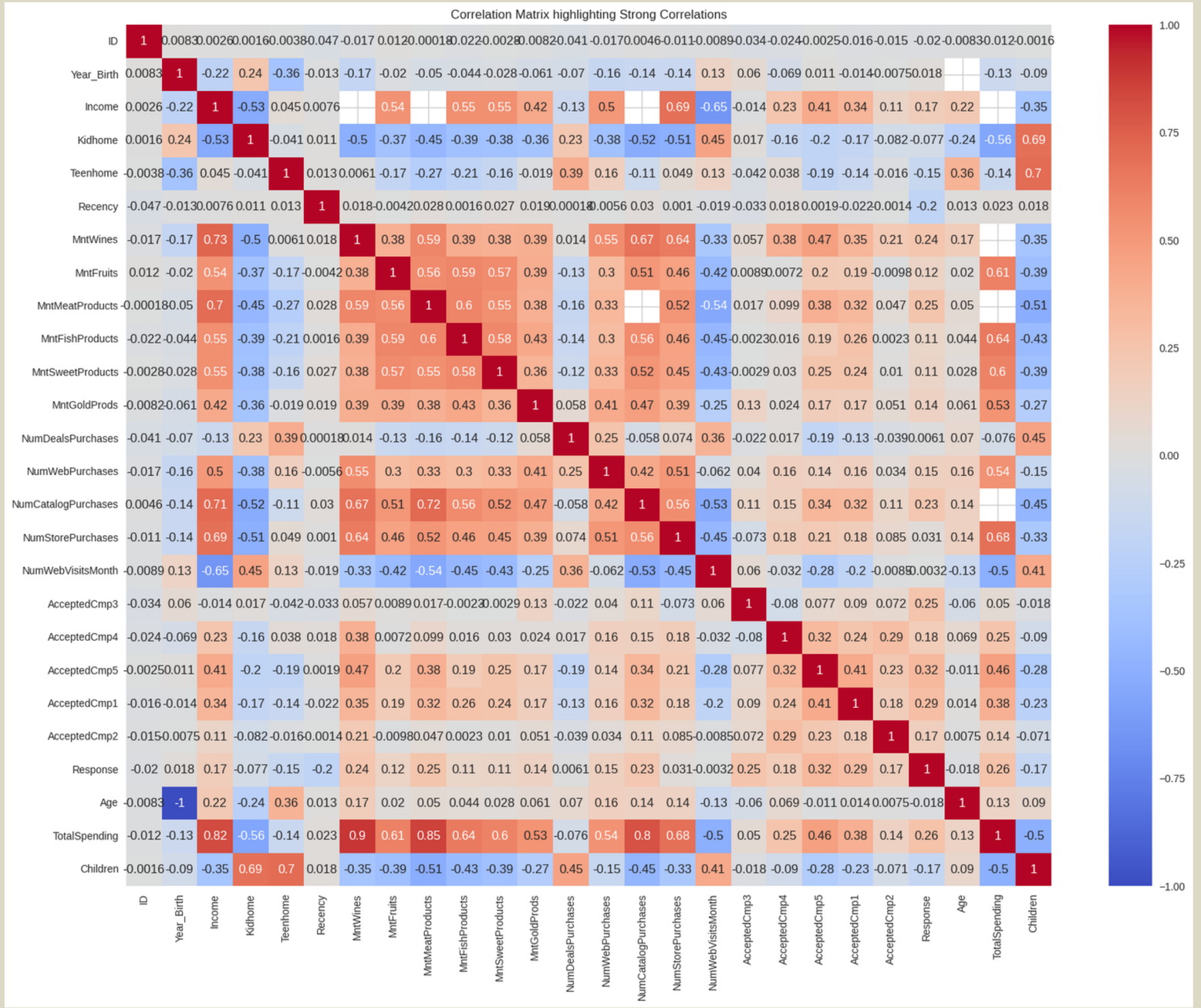
CORRELATION MATRIX

- A correlation matrix is a table that shows the correlation coefficients between multiple variables in a dataset.
- Correlation coefficients quantify the strength and direction of the linear relationship between two variables.
- The matrix provides a comprehensive view of how each variable is related to all other variables in the dataset.

Correlation Matrix



Correlation Martix setting a Threshlod value



CLUSTERING MODELS CONSIDERED

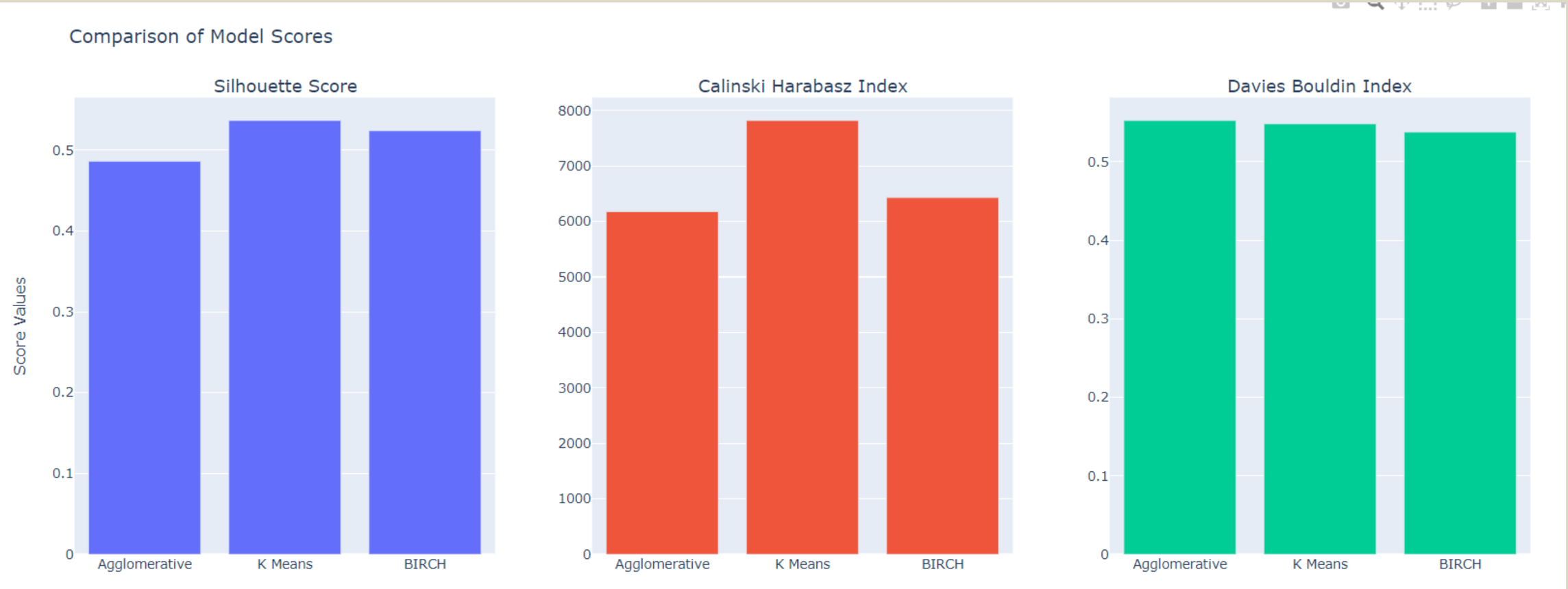
AGGLOMERATIVE	K-MEANS	BIRCH
The hierarchical method merges the closest data points or clusters iteratively until forming a dendrogram. The number of clusters can be determined after the clustering process.	The partition-based method assigns data points to K clusters based on the distance to centroids. Requires the user to specify the number of clusters (K) beforehand.	Hierarchical method that uses a tree-like structure (CF tree) to efficiently cluster large datasets. Allows controlling the granularity of clustering and can handle new data incrementally.

MODEL COMPARISION

Based on

- **Silhouette Score:** Measures cluster quality based on data point distances within and between clusters. A higher score indicates better clustering.
- **Calinski Harabasz Index:** Assesses clustering quality by comparing inter-cluster and intra-cluster dispersion. A higher index implies improved clustering.
- **Davies Bouldin Index:** Evaluates clustering by measuring the average similarity between each cluster and its most similar cluster. A lower index indicates better clustering.

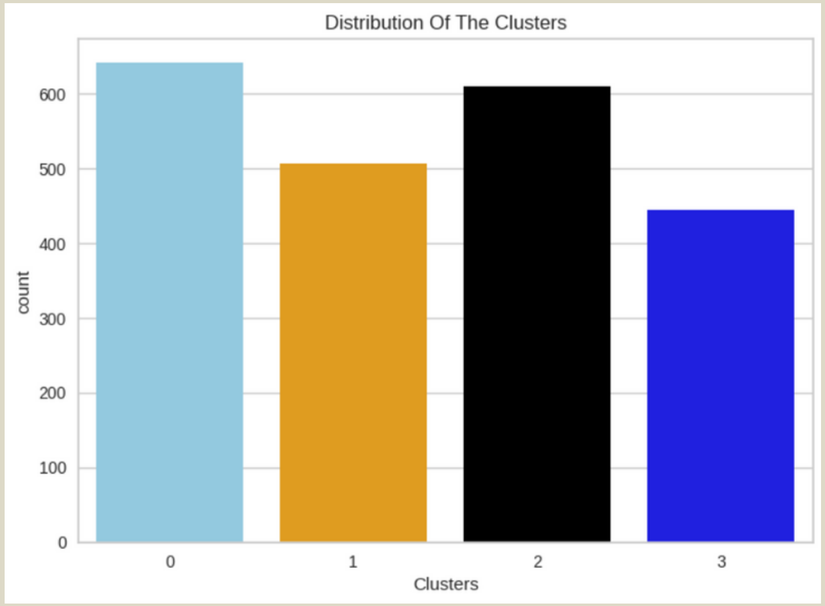
Score/Index	K Means	Agglomerative	BIRCH
Silhouette	0.5366	0.4862	0.5239
Calinski Harabasz	7821.2135	6177.4623	6431.4202
Davies Bouldin	0.5483	0.5527	0.5379

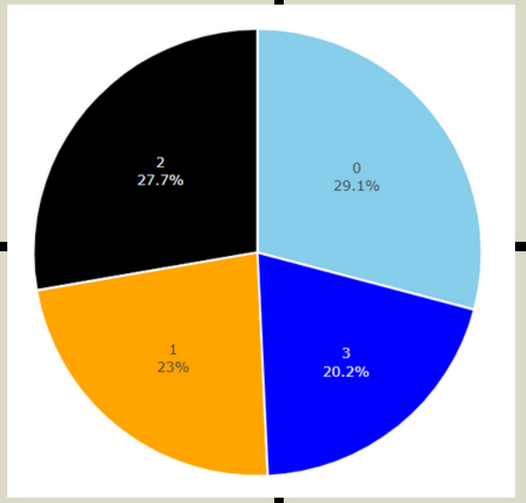


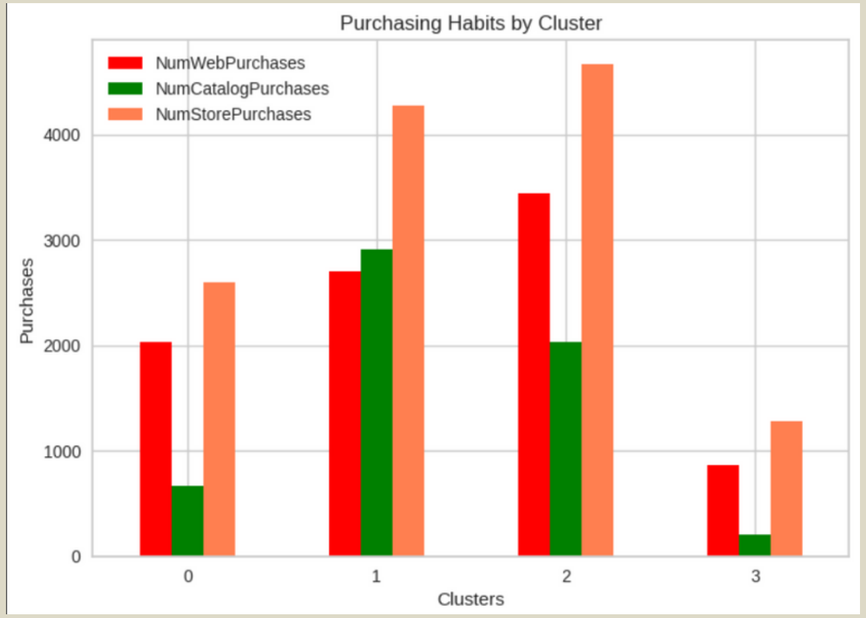
Best Performer: Among the three algorithms, K-Means demonstrates the highest Silhouette Score, Calinski Harabasz Index, and low Davies Bouldin Index.

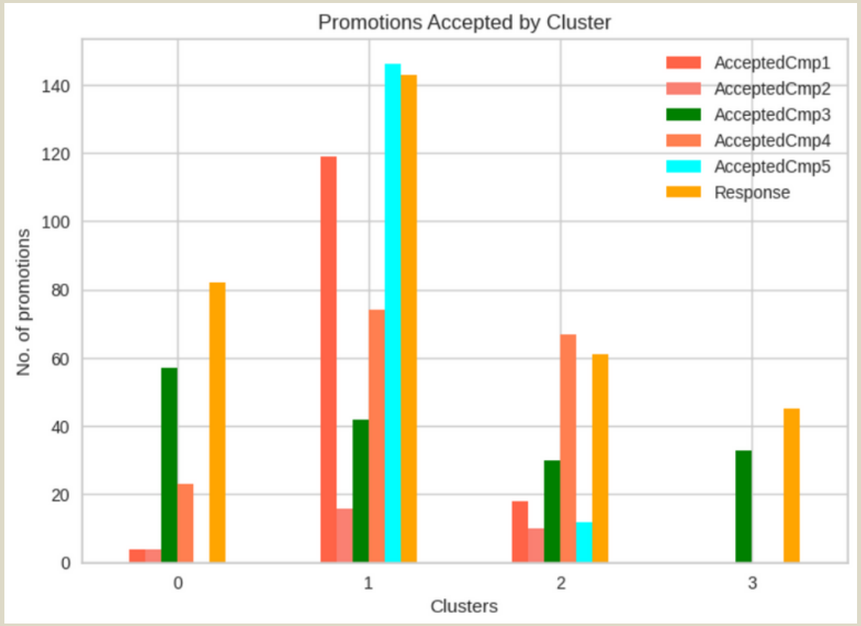
Conclusion: Based on the comparison, K-Means outperforms Agglomerative and BIRCH, making it the most suitable algorithm for this clustering task.

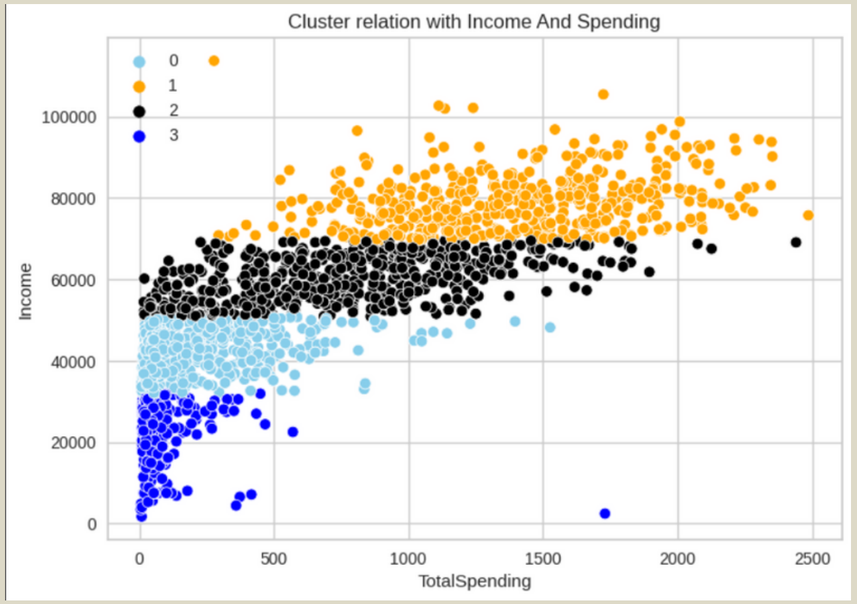
Evaluating the clusters

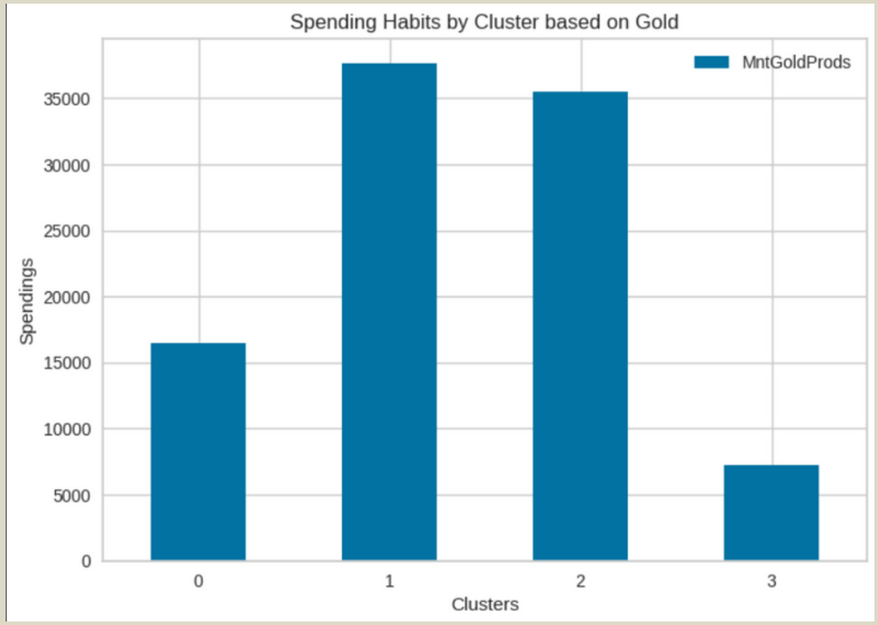

Distribution Of The Clusters
count
Clusters
0
1
2
3
0
100
200
300
400
500
600

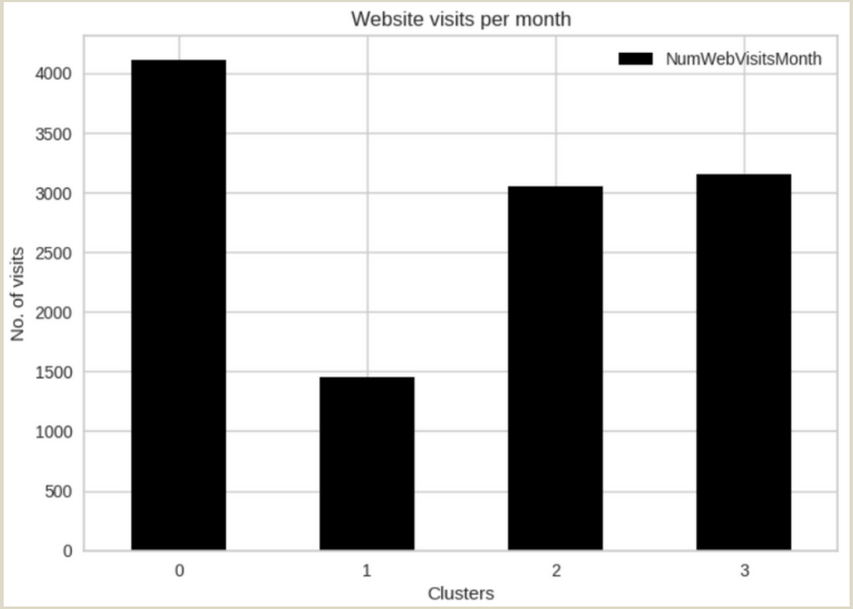

0
29.1%
1
23%
2
27.7%
3
20.2%

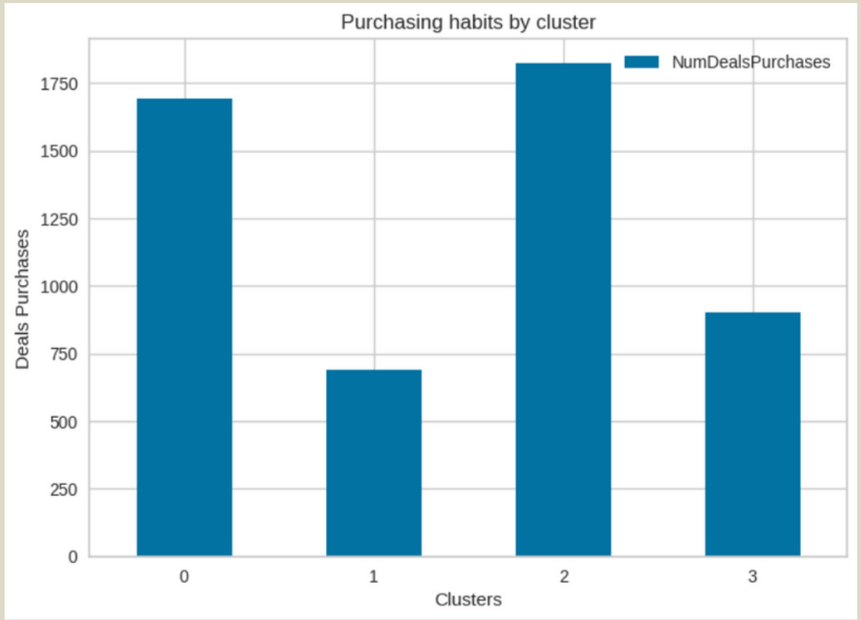

Purchasing Habits by Cluster
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
Purchases
Clusters
0
1
2
3
0
1000
2000
3000
4000

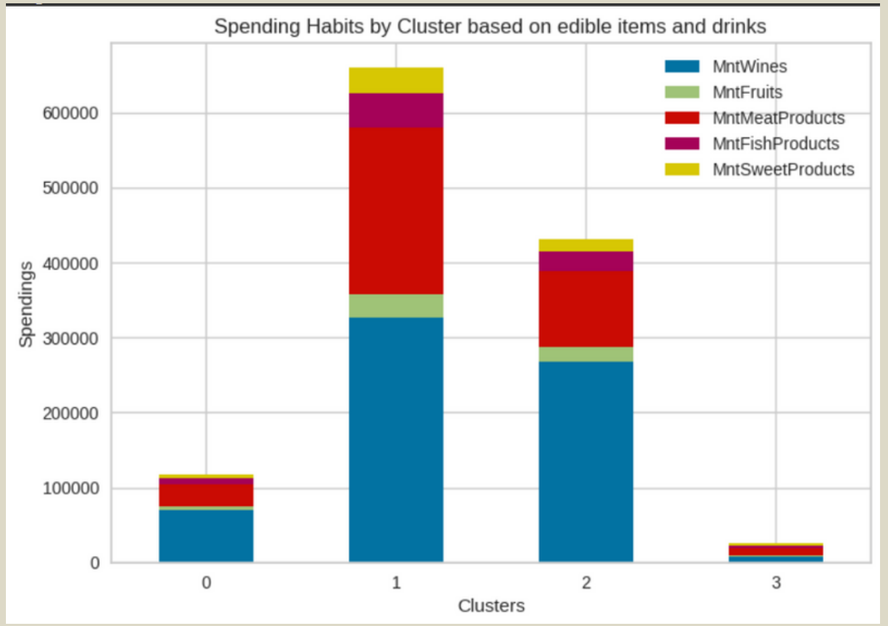

Promotions Accepted by Cluster
AcceptedCmp1
AcceptedCmp2
AcceptedCmp3
AcceptedCmp4
AcceptedCmp5
Response
No. of promotions
Clusters
0
1
2
3
0
20
40
60
80
100
120
140


Cluster relation with Income And Spending
0
1
2
3
Income
TotalSpending
0
500
1000
1500
2000
2500
0
20000
40000
60000
80000
100000


Spending Habits by Cluster based on Gold
MntGoldProds
Spendings
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
Clusters


Website visits per month
NumWebVisitsMonth
No. of visits
0
500
1000
1500
2000
2500
3000
3500
4000
0
1
2
3
Clusters


Purchasing habits by cluster
NumDealsPurchases
Deals Purchases
0
250
500
750
1000
1250
1500
1750
0
1
2
3
Clusters


Spending Habits by Cluster based on edible items and drinks
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
Spendings
0
100000
200000
300000
400000
500000
600000
0
1
2
3
Clusters

Conclusion

- Campaign 5 was the most successful, with the highest acceptance across all customer clusters.
- Customers in the 1st cluster exhibited the greatest acceptance of offers, outperforming other clusters.
- Cluster 3 customers showed the least response to campaigns and spending, requiring targeted marketing efforts.
- Customers from middle-income clusters 0 and 2 were most attracted to deals, indicating a preference for discounted offers.
- Offline stores were the preferred shopping channel for all clusters, suggesting a need to focus marketing efforts on this channel to maintain customer loyalty.

Thank You!

Aman Raj
2005362@kiit.ac.in

Simaran Bhardhaj
simranbhardwaj2607@gmail.com

VardaanKhosla
khoslavardaan1@gmail.com

Rishabh Kumar
rishabhquasar23@gmail.com

Dipti Verma
2006173@kiit.ac.in

Sharad Kumar Agarwal
skagarwal485@gmail.com