Cluster Analysis on Customer Behavior
Kyle Reedy

Introduction:

Every customer has unique characteristics and preferences, so there's no one-size-fits-all when businesses approach customers. That is why by segmenting customers based on their specific traits and behaviors, a business can gain deeper insights into each customer's needs and preferences. This enables the business to tailor its strategies and tactics to effectively meet the diverse needs of different customer segments, ultimately leading to increased profitability.
As a result, targeted campaigns and actions can enhance customer loyalty and engagement, leading to more meaningful conversations and ultimately, increased success for the business. The purpose of this project is to give meaningful recommendations for businesses that may need ideas after segmentations.

The dataset used for this project is a public dataset provided by Dr. Omar Romero-Hernandez located on Kaggle (https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis). The data set has 2240 observations with 29 different variables. Out of the 29 different variables, 25 are numerical and 4 are objects.

This paper explores the application of unsupervised learning algorithms in analyzing customer personalities and how businesses can leverage them to gain deeper insights into their customer base. It delves into various types of unsupervised learning algorithms and their functionality, shedding light on how they can be employed to enhance customer understanding and service.

Data Dictionary used for Project:

People

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

Products

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

Promotion

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

Place

- NumWebPurchases: Number of purchases made through the company’s web site
- NumCatalogPurchases: Number of purchases made using a catalog
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company’s web site in the last month

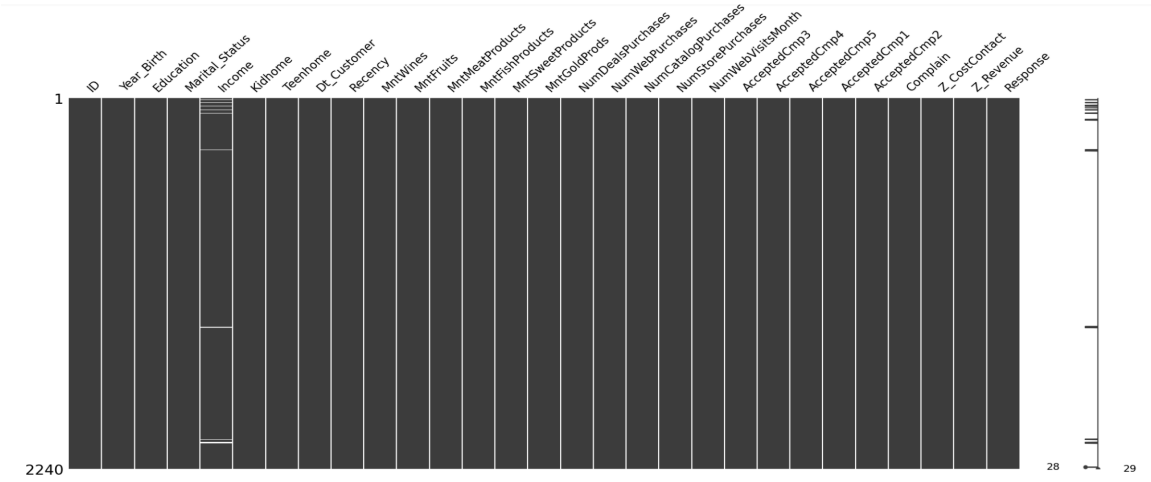
Methodology

Cleaning, Wrangling:

Data cleaning involves identifying and fixing incomplete, incorrect, inaccurate, or irrelevant parts of the data. This part of the project aims to detect and correct inaccurate records in a dataset. Data cleaning is crucial for customer segmentation as it ensures the accuracy and reliability of the data used for analysis. To clean data for segmentation, errors, inconsistencies, and missing values need to be identified and addressed through functions and visualizations.

There were only a couple of necessary steps when cleaning and wrangling the dataset.

1. The dataset had only 24 missing values in the ‘Income’ column
    a. Replaced null values with median values



2. The Date_Customer column is not in the right format
    a. Converting Date_Customer into the correct date format for further processes
3. Creating a ‘Customer_For’ Column utilizing the previously fixed Date_Customer Column
    a. After creating a variable for the oldest and newest registration dates, we can use those to subtract the newest registration with the newly created ‘Date_Customer’.

```
df2["Date_Customer"] = pd.to_datetime(df2["Dt_Customer"])
newest_date = df2["Date_Customer"].max().date()
oldest_date = df2["Date_Customer"].min().date()

print("The newest customer's enrollment date in the records:", newest_date)
print("The oldest customer's enrollment date in the records:", oldest_date)
```

```
The newest customer's enrollment date in the records: 2014-12-06
The oldest customer's enrollment date in the records: 2012-01-08
```

```
/var/folders/2r/qvwb_49162s3q75v3pr4d1x00000gn/T/ipykernel_97584/2632591379.py:1: U
M/YYYY format when dayfirst=False (the default) was specified. This may lead to inc
a format to ensure consistent parsing.
  df2["Date_Customer"] = pd.to_datetime(df2["Dt_Customer"])
```

Next, I'll create a new feature ("Customer_For") indicating the number of days since customers first started shop

```
days_since_registration = (newest_date - df2["Date_Customer"].dt.date).dt.days
df2["Customer_For"] = days_since_registration
```

The different customer segmentation demographics that will be looked into:

Age: This is important because a 25 year old who lives in Los Angeles will have significantly different behaviors and interests from someone who is from New York and is 55 years old.

Marital Status: Businesses can target customers based on whether they are single, married, engaged, divorced, etc. Their marital status can affect the household's way of living and spending.

Household with children: This is correlated with income, but a household with children will have unique needs and will obviously have more spending power than a household with no children.

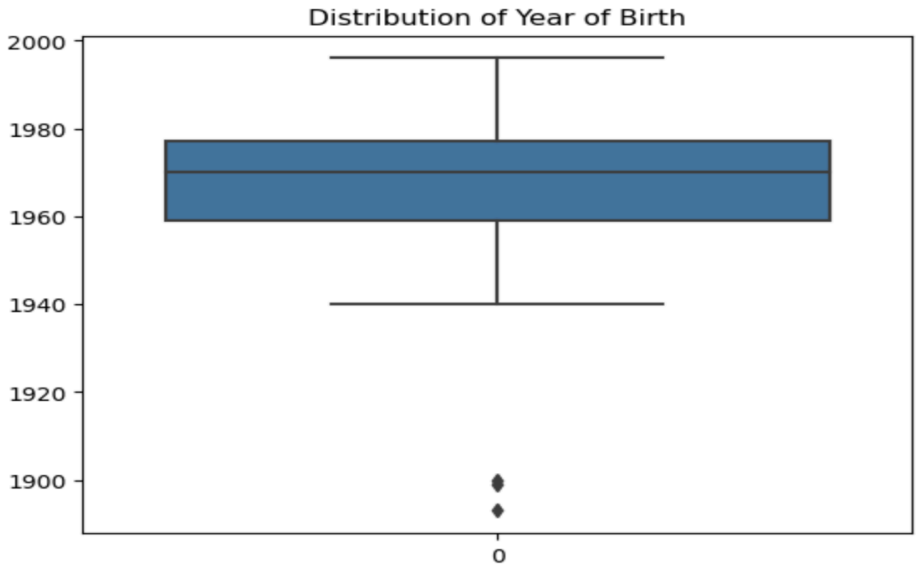
Income: Income is an obvious example of segmentation as it indicates how much a customer can spend. However, if this is combined with a household with children, many factors will change.

Education: Customers with higher levels of education have different thought and decision-making processes. Purchasing power will change as people with higher level degrees tend to earn more than those who have not received higher educational degrees.

Exploratory Data Analysis:

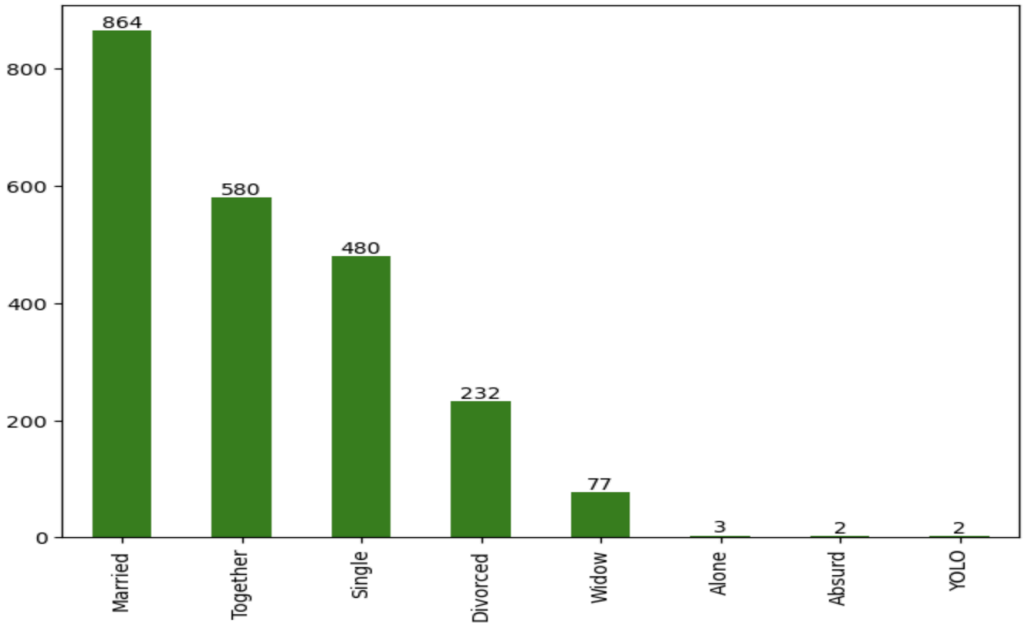
It's important to look at the customer demographics, especially through visualization to see what sort of dataset we are dealing with. Here are the features we are looking at in order to apply the customer segmentation.

‘Year of Birth’



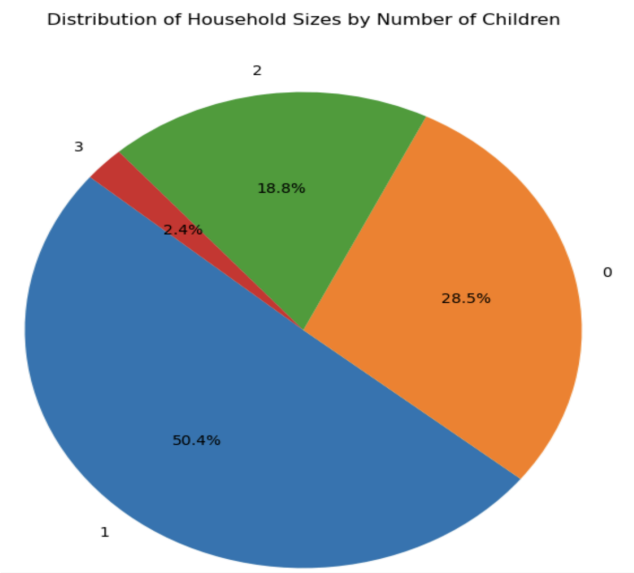
From the above graph, it shows that the year of birth starts from 1940 to the birth year of 1997.

‘Marital Status’



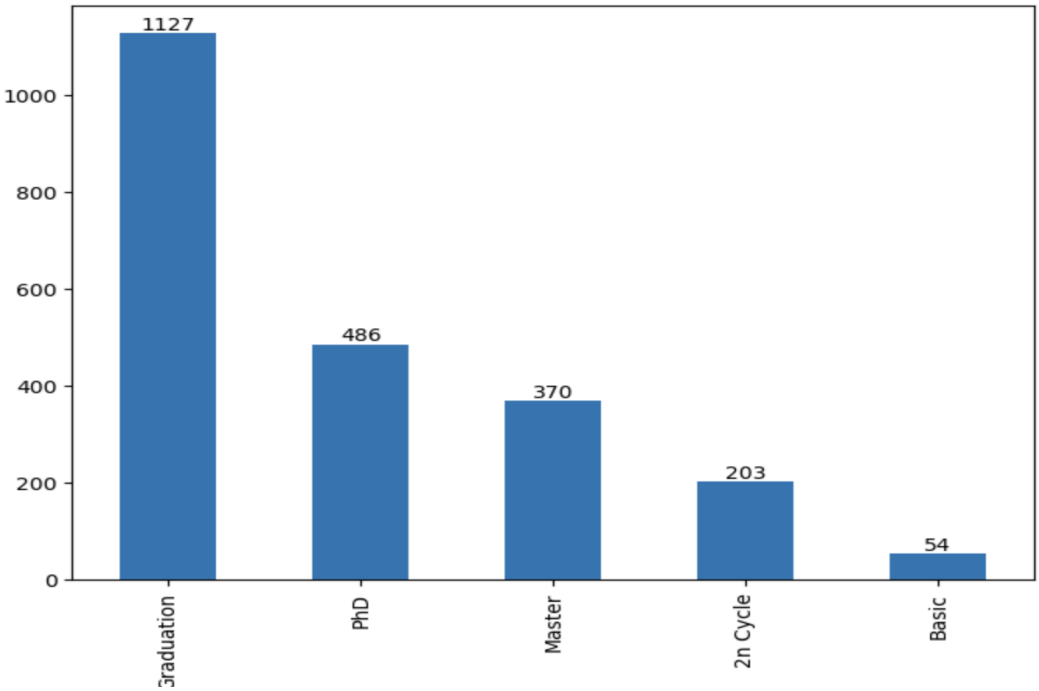
1,444 customers are either married or together while the other 796 are single, divorced, widow, alone, absurd and YOLO.

‘Household by Children Count’



Almost exactly half of the customers in the dataset have one child, whilst 28.5% of them have no children. Some have 2-3 total children.

‘Education’



Most customers have a bachelor's degree, while 856 have post-graduate degrees. The rest do not have degrees from higher education.

Preprocessing

The preprocessing part of the project involves several steps in order to fulfill the clustering process. This process involves handling missing variables, encoding categorical variables, featuring engineering, handling outliers, etc.

```
# Calculating customer age based on their birth year (2015 is the publish date of dataset)
df['Age'] = 2015 - df['Year_Birth']

# Calculating total amount spent by each customer
df['Spending'] = df[['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldPro

# Calculating the number of children in each household
df['Children'] = df['Kidhome'] + df['Teenhome']

# Identifying parental status based on the presence of children
df['Parental Status'] = np.where(df['Children'] > 0, 1, 0)

# Mapping marital status categories to simplified ones
df['Marital_Status'].replace({"Married": "Couple", "Together": "Couple", "Absurd": "Alone", "Widow": "Alone", "YOLO

# Mapping education categories to simplified ones
df['Education'].replace({"Basic": "Undergraduate", "2n Cycle": "Undergraduate", "Graduation": "Graduate", "Master":

# Calculating the total number of promotional campaigns accepted by each customer
df['Total Promo'] = df[['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5']].sum(axis=1

# Renaming columns for clarity
df.rename(columns={"Marital_Status": "Marital Status", "MntWines": "Wines", "MntFruits": "Fruits", "MntMeatProducts
```

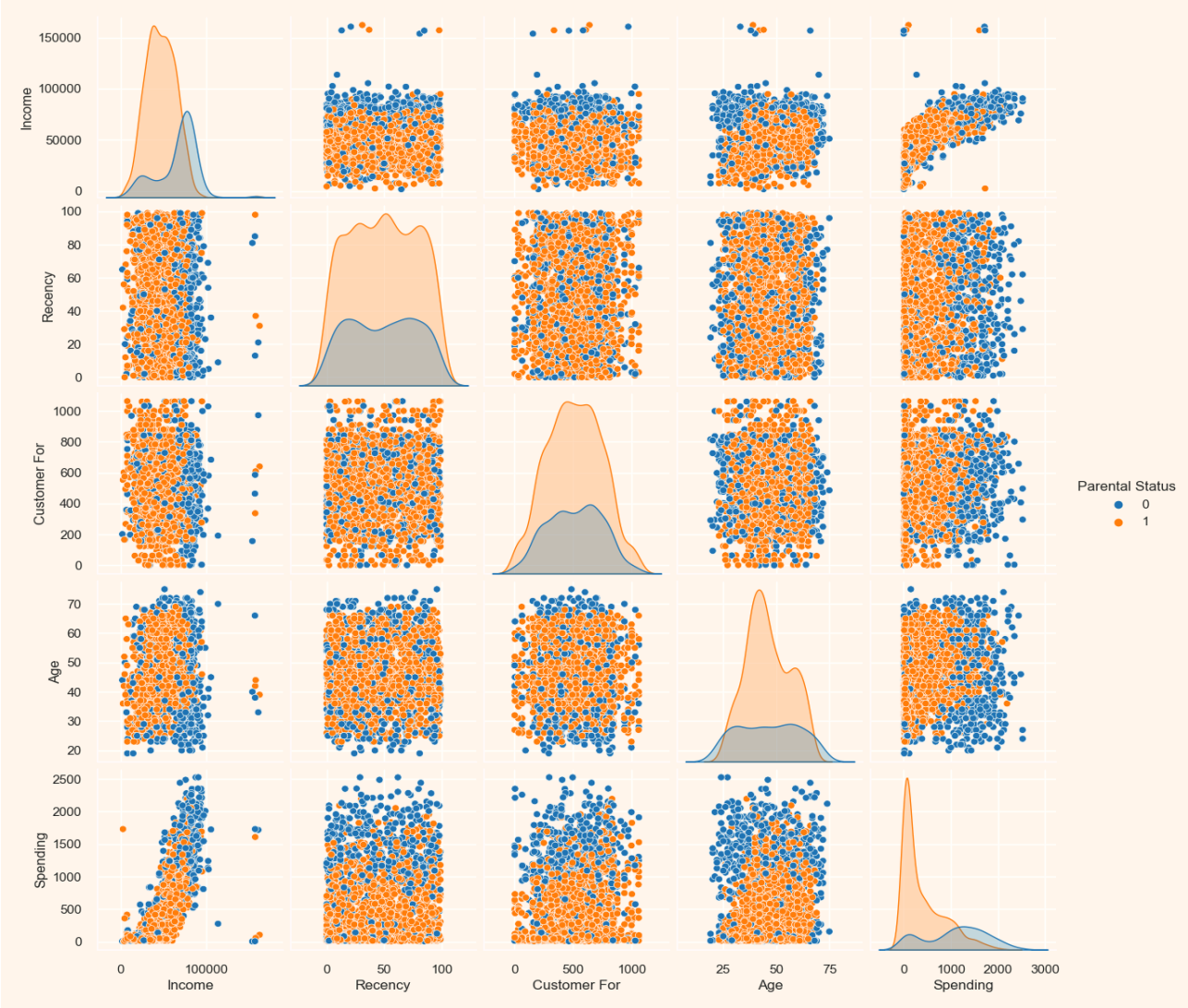
The figure above shows the new features being created in order to start the clustering process. Necessary features have been changed into numerical features and some stayed the same as categorical variables as we need them for the clustering process.
The following are the new columns created:

1. ‘Age’
2. ‘Spending’
3. ‘Children’
4. ‘Parental Status’
5. ‘Marital Status’
6. ‘Education’
7. ‘Total Promo’

In the next step, the following features have been dropped as they will not be used for the newly created dataframe.

```
to_drop = ["Dt_Customer", "Z_CostContact", "Z_Revenue", "Year_Birth", "Income_Group", "AcceptedCmp1", "AcceptedCmp2
df = df.drop(to_drop, axis=1)
```

1. ‘Dt_Customer’
2. ‘Z_CostContact’
3. ‘Z_Revenue’
4. ‘Year_Birth’
5. ‘Income_Group’
6. ‘AcceptedCmp’



The above figure displays plots amongst five different features with the hue of the newly created Parental Status. 0 meaning they have no children and 1 meaning the household does have children.

Removing Outliers

```
df = df[(df["Age"]<95)]
df[df["Age"]>95]
```

ID	Age	Education	Marital Status	Parental Status	Children	Kidhome	Teenhome	Income	Spending	...	Meat	Fish	Sweets	Gold	Web	Catalog	Store	Discount Purchases	T Pro

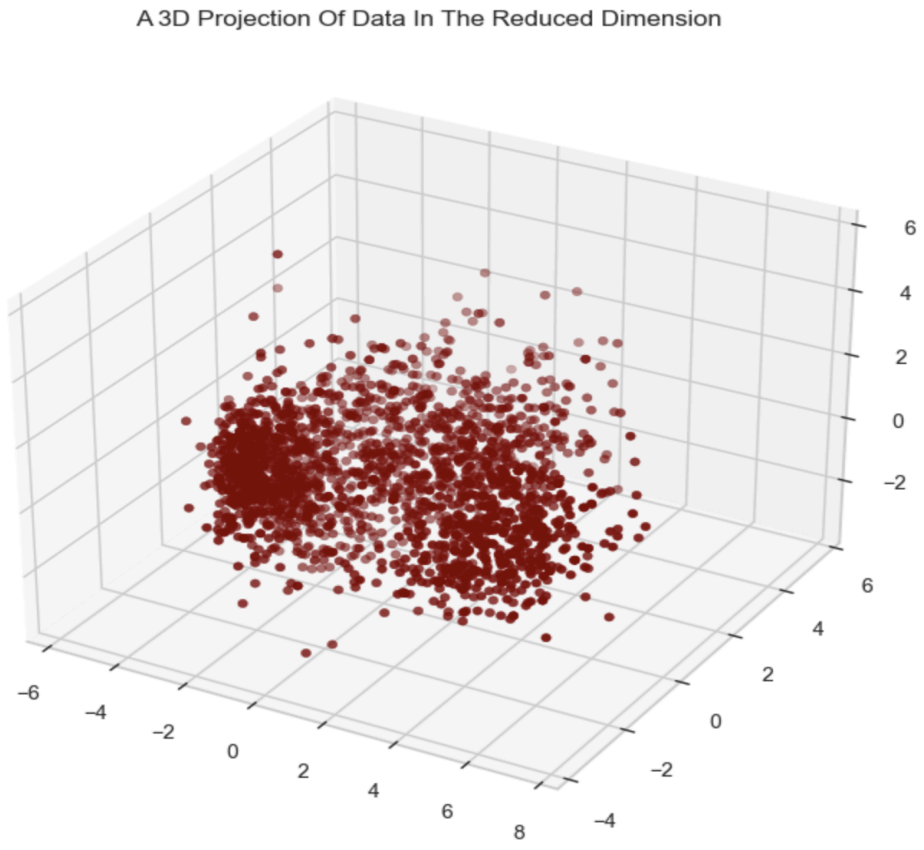
0 rows × 24 columns

From the above figure, there was one occurrence of someone that was over 150 years old which is not possible as the oldest person ever to live is 122 years old. That outlier has been removed.

Modeling

PCA:

Principal component analysis (PCA) is a method used to simplify complex datasets by reducing their dimensionality. This makes the data easier to interpret while minimizing the loss of information. In this part of this project, the final classification will be based on numerous variables, which can be thought of as different factors or characteristics. However, handling a large number of features can be challenging, especially when many of them are correlated and redundant. To address this issue, I reduced the dimensionality of the features before running them through a classifier. Specifically, I will reduce the dimensions to three and then visualize the resulting data frame.

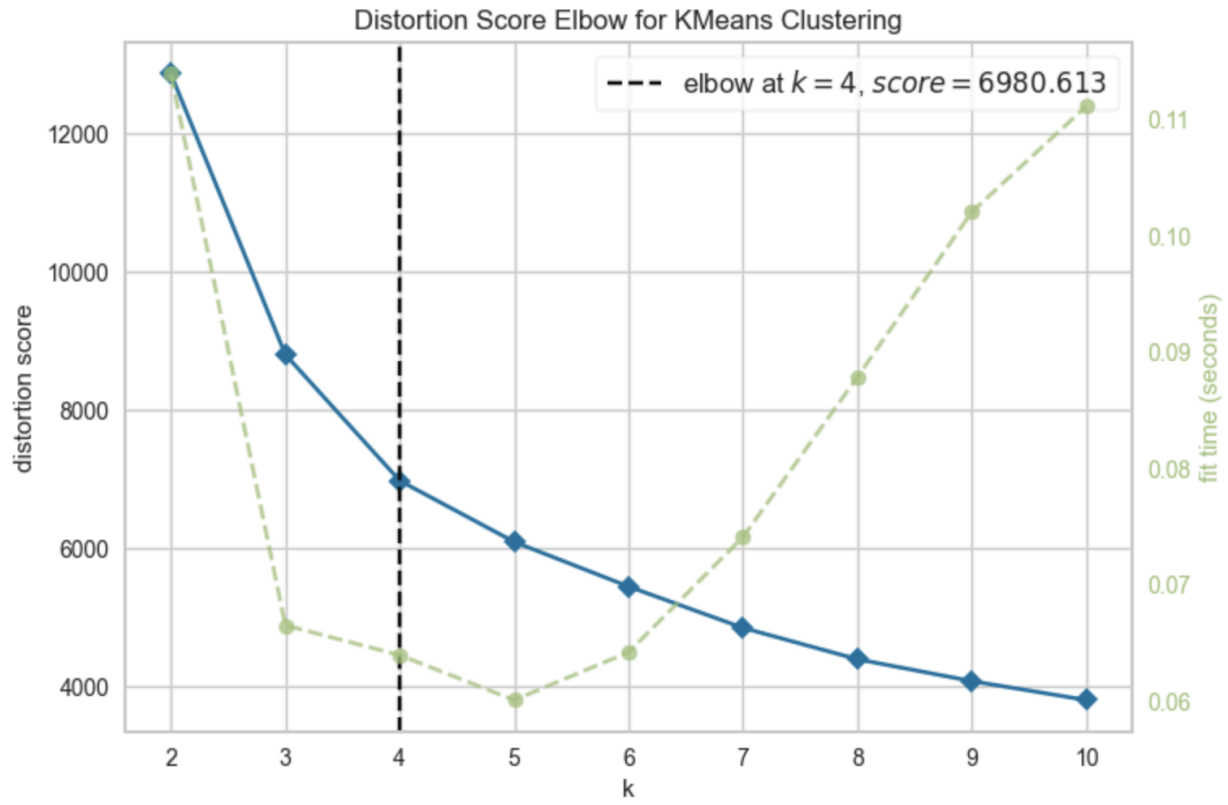


The above plot shows that after the principal component analysis, the data has been processed into three dimensions.

K-Means Clustering:
K-means is a clustering algorithm that assigns each data point to a cluster based on its proximity to a centroid. The main objective of K-means is to identify the optimal number of clusters (K) in a dataset. This algorithm originated in signal processing as a vector quantization technique.

The goal of K-means clustering is to partition a set of observations into K clusters, where each observation is assigned to the cluster with the closest mean, serving as the prototype of the cluster. By using K-means, we aim to discover meaningful clusters within the data. In this specific case, we intend to identify and select four clusters that best represent the underlying structure of the dataset.

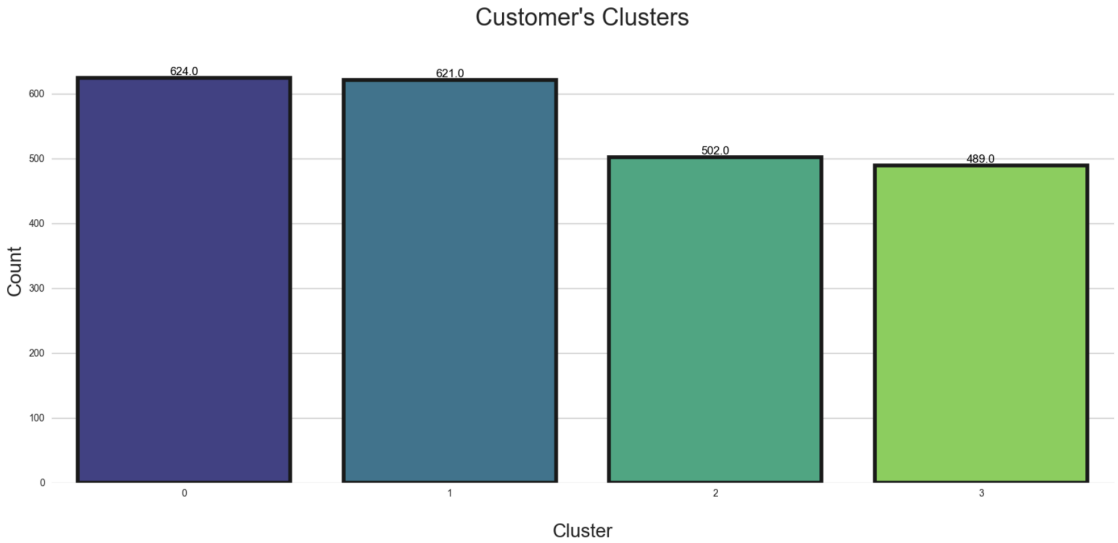
K-means clustering is an important method for clustering customers because it allows businesses to group customers with similar characteristics or behaviors together. By identifying these clusters, businesses can gain insights into customer segments, such as their preferences, purchasing patterns, or response to marketing campaigns. This information enables businesses to tailor their products, services, and marketing strategies to better meet the needs of different customer groups, ultimately improving customer satisfaction and driving business growth.



Hierarchical Clustering:

From the above plot, it demonstrates the results of applying hierarchical clustering to the dataset. The data has been transformed into four different clusters.

Customer Profiling

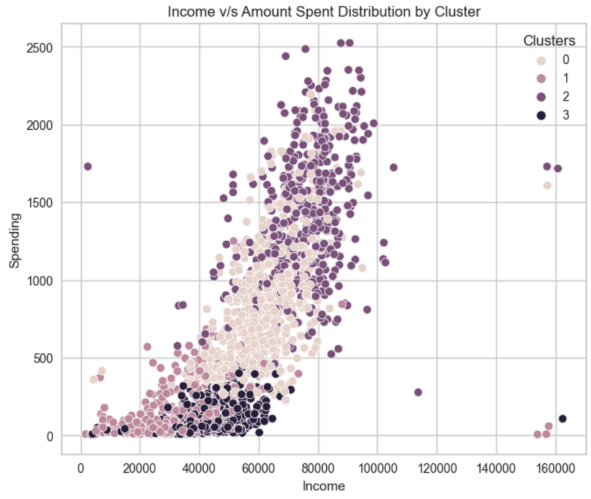


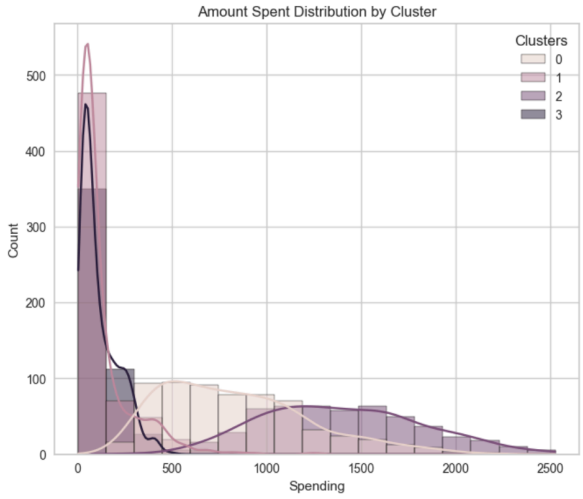
The above plot shows the count of customers for each cluster after utilizing K-Means Clustering with the elbow method.

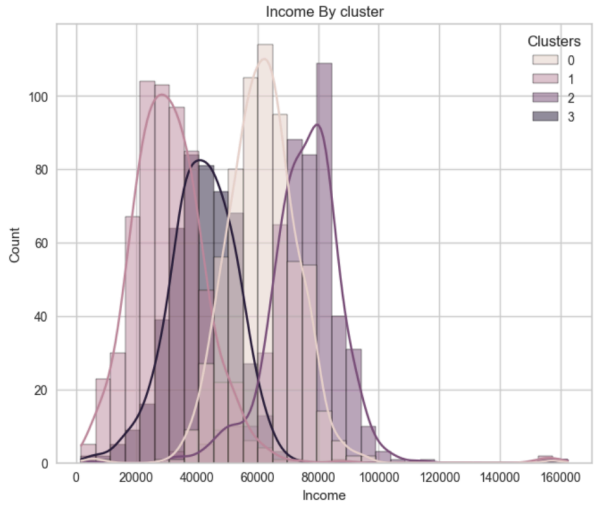
Customer Profiling EDA

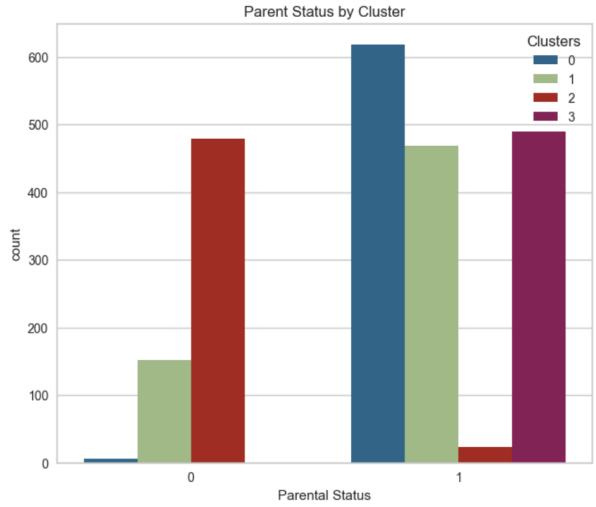
Here are the following features we will be looking into through the customer profiles based on clustering:

- Income vs Spending
- Count of Spending Amount
- Count of Different Incomes
- Count of Parental Statuses





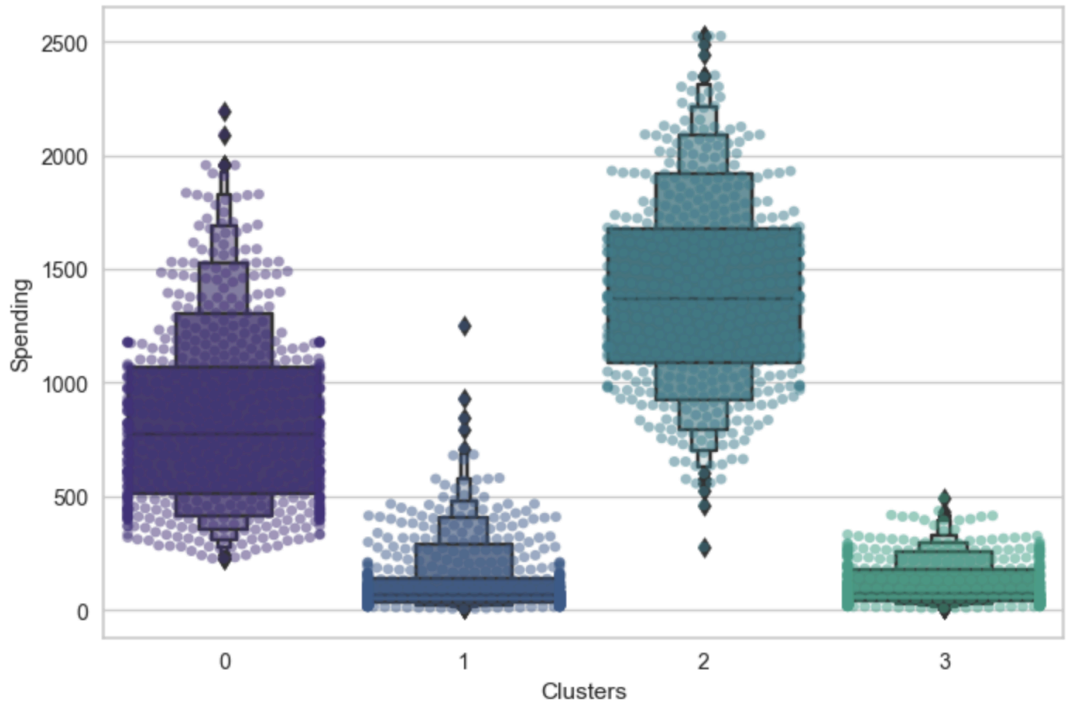




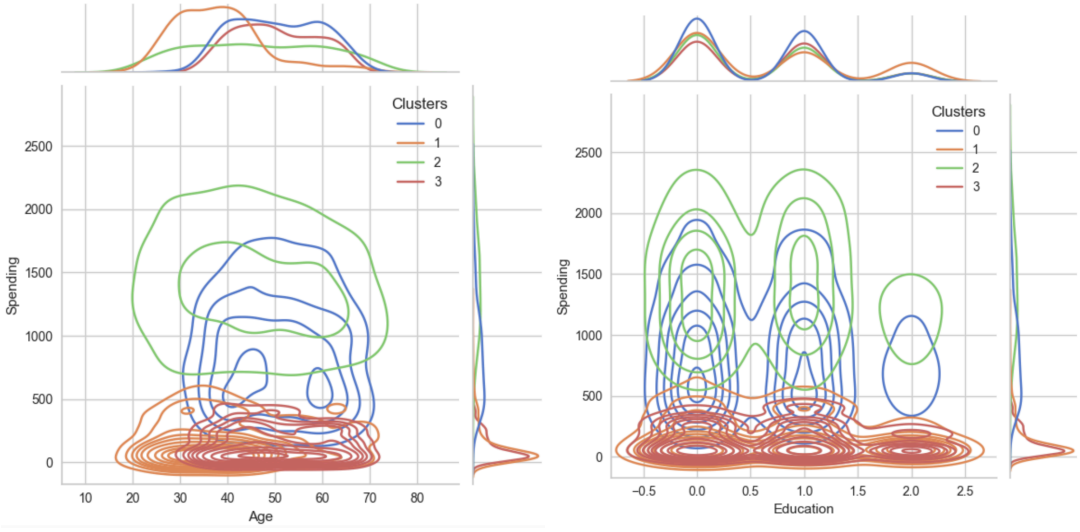
Insights:

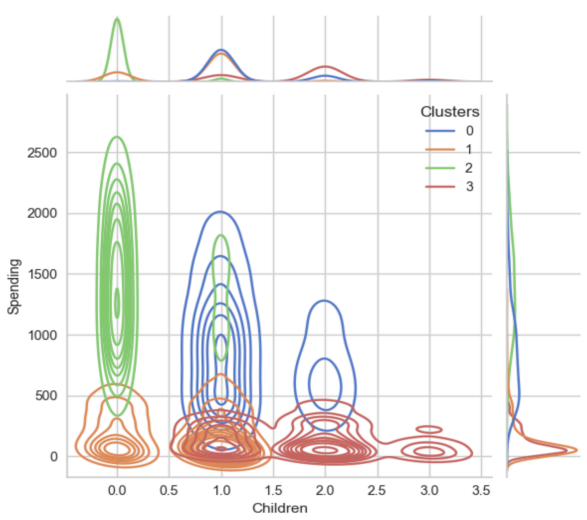
- Cluster 0: average income | high spending
- Cluster 1: low income | low spending
- Cluster 2: high income | high spending
- Cluster 3: average income | low spending

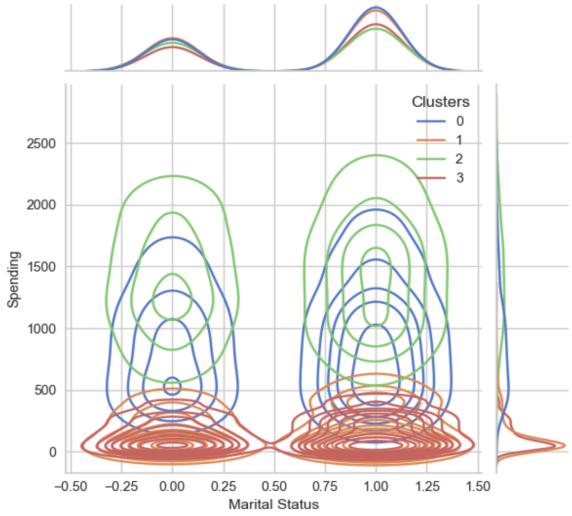
Who are the Biggest Customers

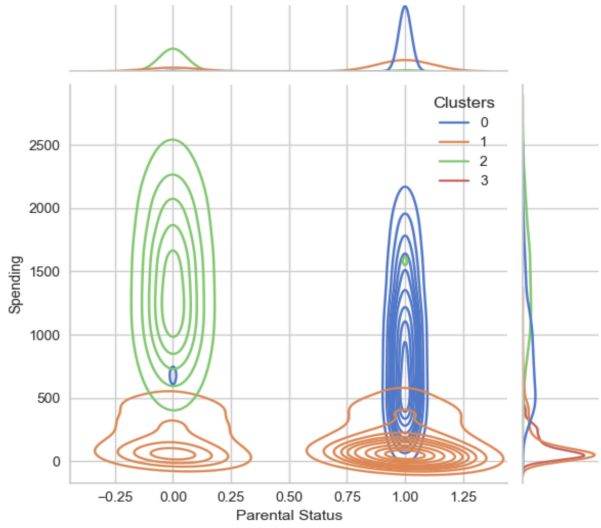


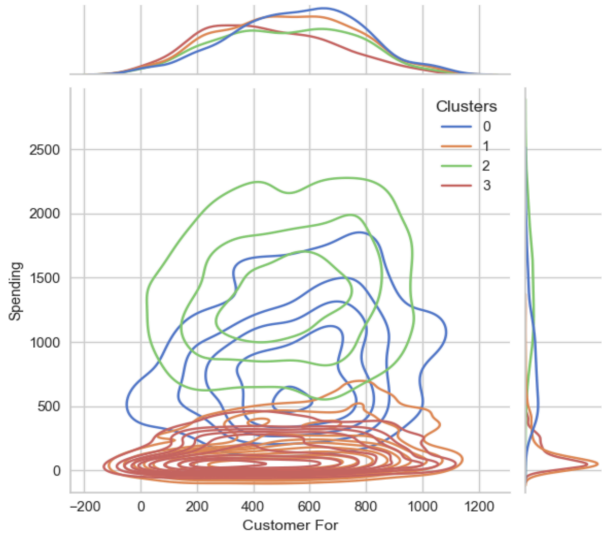
Clusters 0 and 2 are the largest customers, looking at this swarmplot. The only difference from the two clusters is that 2 spends more on average than 0.











Insights From Cluster EDA

Cluster 0:

- Almost all are parents
- A third utilized promos before purchasing products
- Most customers have an age range between 35 and 65
- Their spending range is in between 100 and 1750
- Customers from this cluster tend to spend at least 500 but majority less than 1000
- The customers of cluster 0 are graduate or in postgraduate level or in undergraduate level
- Most of the customers of cluster 2 are married
- Most of the customers of have one child, some have two
- The majority of customers have made 2 to 4 discounted purchases

Cluster 1:

- Some are parents
- Less than 10% utilized promos before purchasing products
- Most customers have an age range between 30 and 50
- Their spending range is in between 0 and 600
- Most of the customers of cluster 1 are graduate or in postgraduate level
- Some of the customers of cluster 1 are married, some not married
- Majority of customers do not have a child or have one child
- Cluster 1: The majority of customers have made 1 discounted purchases

Cluster 2:

- Almost all are not parents
- Most utilized promos before purchasing products
- Cluster 2: Most customers have an age range between 30 and 65
- Their spending range is in between 750 and 2300
- Customers from this cluster tend to spend at least more than 1000
- Most of the customers of cluster 2 are graduate and very few in postgraduate level
- Most of the customers of cluster 2 are married
- Majority of customers do not have a child
- Cluster 2: The majority of customers have made 1 discounted purchases

Cluster 3:

- Almost all are parents
- Less than 10% utilized promos before purchasing products
- Most customers have an age range between 40 and 65
- Their spending range is in between 0 and 500
- All the customers of cluster 3 are graduate or in postgraduate level and some are in undergraduate level
- Some of the customers of cluster 1 are married, some not married
- Most customers have one or two children, some have three
- Cluster 3: The majority of customers have made 1 to 3 discounted purchases

Conclusion and Recommendations

It is extremely difficult and unfavorable for companies to treat each customer similarly as each customer has their own personality and demographics. That is why customer segmentation is a necessary process for any business (in this case retail stores) to implement strategies by customizing relationships with their customers. Unsupervised learning offers a powerful approach to analyzing customer personality traits. By leveraging clustering algorithms and dimensionality reduction techniques, businesses can uncover meaningful patterns and correlations within customer data. These insights enable businesses to anticipate customer behaviors and preferences more accurately. By integrating these findings into marketing strategies and customer service initiatives, companies can deliver more tailored experiences that resonate with their target audience.

From this project, we were able to identify four clusters from the dataset of households with various demographics. Based on their demographics, their behaviors are contrasting, leading to volatile consumer behavior. With this in mind, below are three recommendations for businesses to implement after a refined cluster analysis.

Targeted Marketing Strategies: Marketing teams can tailor marketing campaigns to the specific needs and preferences of each customer cluster. Focus on promoting products and offers that are most relevant to each segment, such as family-oriented bundles for Cluster 0, value-driven promotions for Cluster 1, and exclusive discounts for frequent shoppers in Cluster 2.

Product Assortment Optimization: We can optimize product offerings to cater to the preferences of each cluster. Ensure that product assortments align with the spending habits and purchasing behaviors of customers in each segment. For example, offer budget-friendly options and loyalty rewards for Cluster 1, while providing premium products and higher-end offerings for Cluster 2.

Promotion Strategy Refinement: Stores can refine promotion strategies to maximize engagement and conversion rates for each customer segment. This can be done by tailoring promotions to match the promo utilization and spending patterns of customers in each cluster. We can also Implement targeted promotional campaigns, personalized discounts, and loyalty programs to incentivize purchases and enhance customer loyalty across all segments. If a household has a certain number of children, the promotional campaigns can be targeted towards lets say, a household with two children with, married, but no graduate degrees.

References:

Ackerman, Margareta, and Shai Ben-David. "A Characterization of Linkage-Based Hierarchical Clustering." Journal of Machine Learning Research, vol. 17, 2016, pp. 1–17. URL: https://www.jmlr.org/papers/volume17/11-198/11-198.pdf.

Piech, Chris. "K-Means." 2013. Stanford University, https://stanford.edu/~cpiech/cs221/handouts/kmeans.html. Written by Chris Piech. Based on a handout by Andrew Ng.

Kaggle Data Source: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis