Customer Segmentation

Link to github Repository : https://shorturl.at/cnqJ0

1. Data Description :

This data is taken from kaggle, here is the link to the original data : https://shorturl.at/sEG59

This dataset is composed of 29 columns :

- Customer related Information :
  - 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 the customer complained in the last 2 years, 0 otherwise
- Amount Spent of different 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
- Promotions :
  - 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
- Channel :
  - NumWebPurchases: Number of purchases made through the company’s website
  - NumCatalogPurchases: Number of purchases made using a catalogue
  - NumStorePurchases: Number of purchases made directly in stores
  - NumWebVisitsMonth: Number of visits to company’s website in the last month

2. Main objective :

Our approach involves utilizing Principal Component Analysis (PCA) for dimensionality reduction, clustering for customer segmentation, and customer profiling. PCA simplifies the dataset while retaining essential information. Clustering groups similar customers for tailored strategies. Profiling unveils specific traits and preferences within each cluster. These data-driven techniques inform marketing and engagement strategies.

3. Data Exploration and Transformation :

In this section, we will incorporate visual aids from my notebook to facilitate our data exploration and elucidate the relationships among the features themselves or between the features and the target variable. Prior to delving into this visual analysis, we will commence by acquainting ourselves with the structure of our dataset, scrutinizing the data types, and determining which columns merit removal due to their lack of relevance in constructing our forthcoming predictive machine learning model.

The complete notebook will be attached with this document.

- Null Values Handling :

While discovering the data types of different columns, we found that the income feature column contains some null values. In this case,

-We can either remove the rows with null values at income feature

- or we can fill the missing values through :
  - filling with the mean value
  - filling with the median value
  - searching for the feature(s) with which Income is most correlated and perform a linear regression
  - using any other regression model
  - **using K-Nearest Neighbors Imputation**

- Feature Engineering :

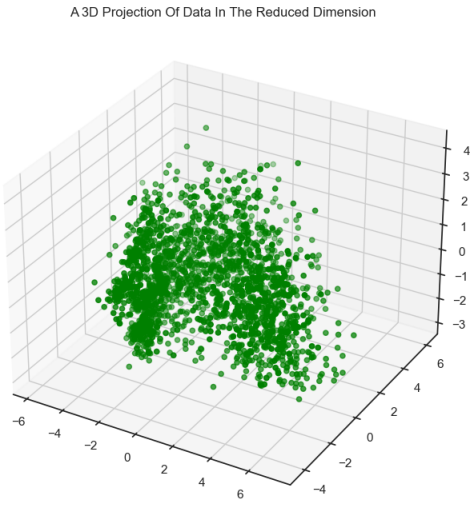
- We calculate a new feature that can definitely be helpful in our analysis, which is **Age**, calculating through subtracting the year_birth values from the current year (2023).
- We calculate the total spending of each customer through adding up all the amounts spent on different subtypes.
- Martial_Status column values can be grouped in just a binary variable of either living alone or with a partner to simplify calculations

- Calculate the number of children for a household by adding the number of kids and the number of teens.
- Calculate family size by adding the values of newly added columns of number of children and whether the household is living alone or with a partner to find the total family size.
- Education can be summarized in undergraduate, graduate or postgraduate.
- We can calculate for how long a household has been our customer using the Dt_Customer feature.

4. Principal Component Analysis :

- Before Starting our PCA, we should normalize our dataset, and one-hot encode our categorical data.
- Results show that :

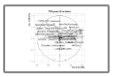
```
Principal Component 1: Explained Variance =
0.3058, Cumulative Variance = 0.3058, Feature
= Income
Principal Component 2: Explained Variance =
0.1137, Cumulative Variance = 0.4195, Feature
= Kidhome
Principal Component 3: Explained Variance =
0.0849, Cumulative Variance = 0.5044, Feature
= Teenhome
```

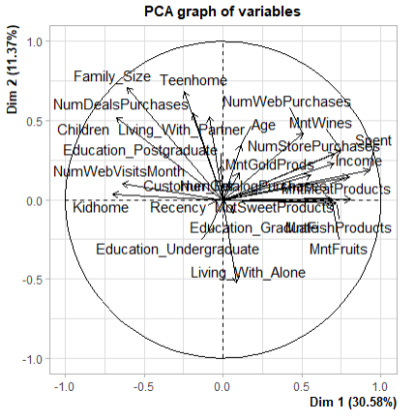


These 3 features explain 50.44% of the total variance of the data with the principal component Income explaining 30.58% of the total variance of the data variability.

```{R}
data <- read.csv(file.choose(),sep='\t')
```

```{R}
res.pca <- PCA(data, scale.unit = TRUE, ncp = 5, graph = T)
```

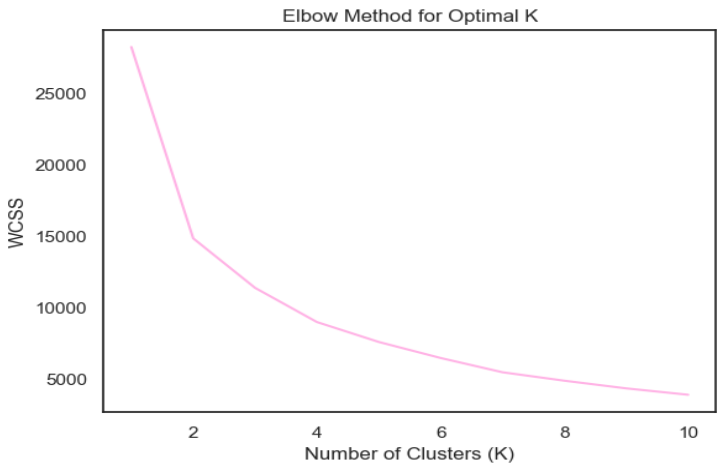


The factor map provide us with some useful information about our features as well :

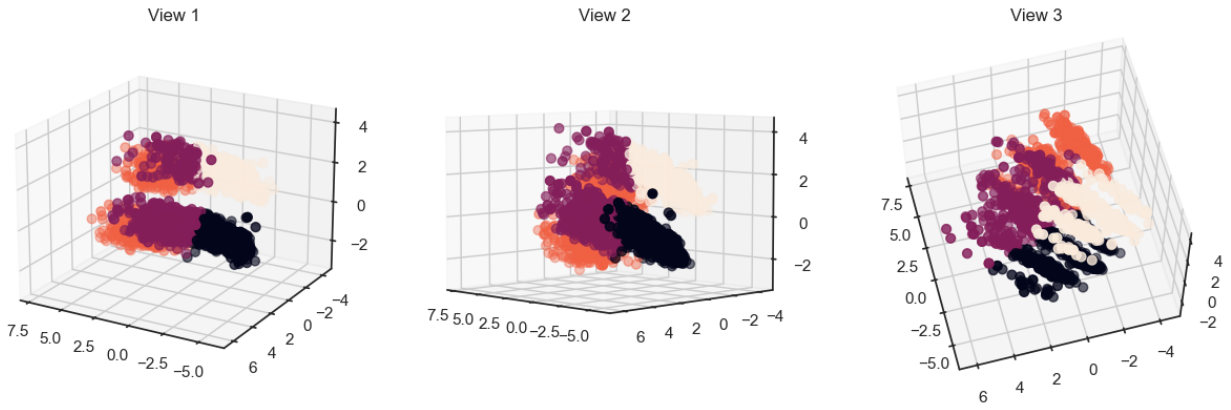
- Income and Spending are highly correlated features.
- Family Size is highly correlated with NumDealsPurchases.
- There is no correlation between the number of web visits per month and the number of web puchases.
- There is a strong correlation between income and {amount spent on wine, amount spent on meat}
- Income and Undergraduate studies are negatively correlated
- Amount spent on fruits is highly correlated with amount spent on Fish
- There is a strong correlation between age and Number of Store Purchases.

5. Clustering :

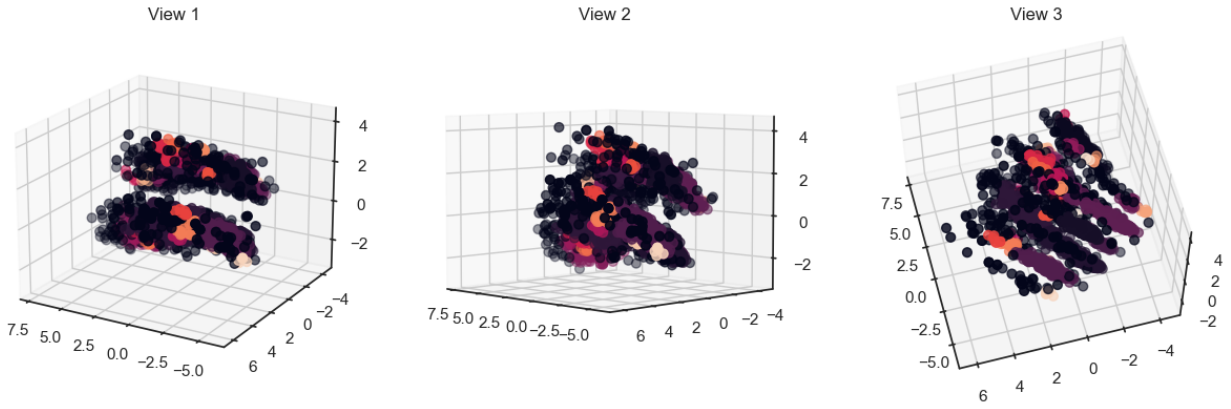
5.1 K-Means Clustering :



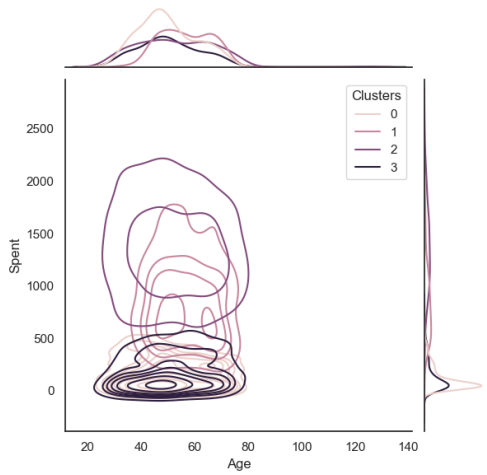
Using the Elbow method, we can see that the optimal number of clusters is equal to 4. Below is the visualization of clusters with a 3D representation.

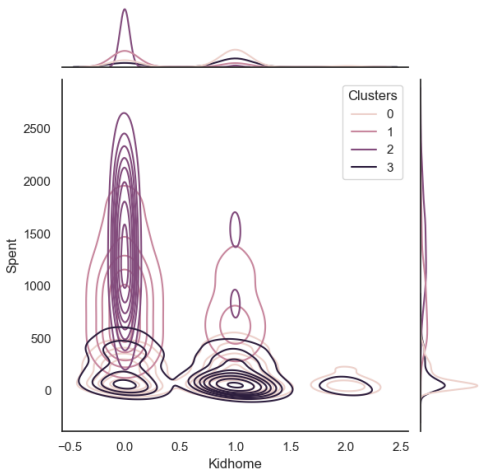


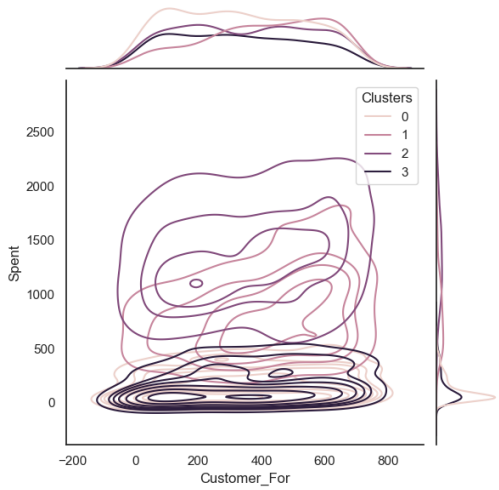
5.1 Density-Based Spatial Clustering of Applications with Noise :

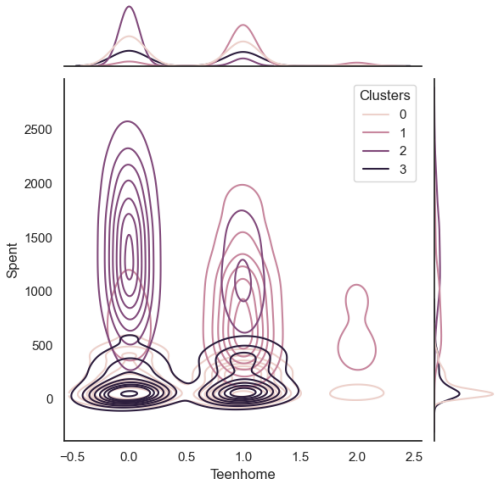


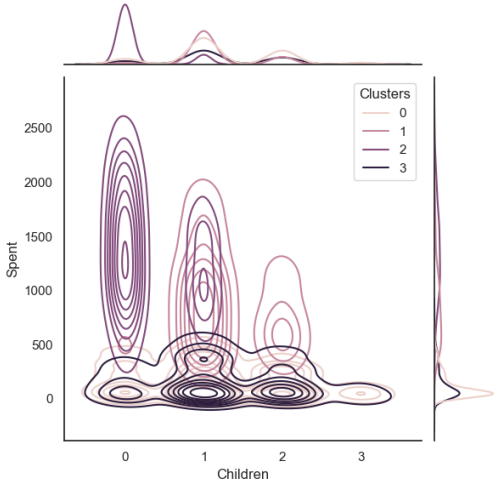
6. Profiling :

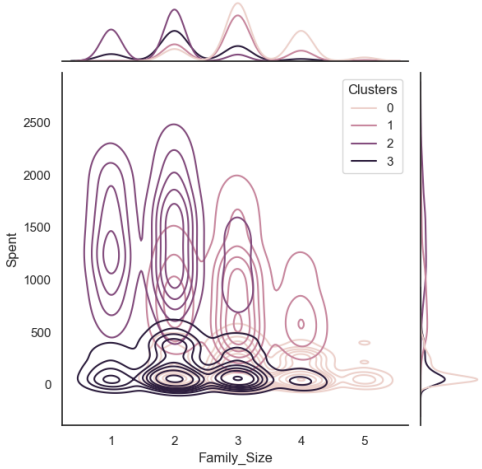

Clusters
0
1
2
3
Spent
Age

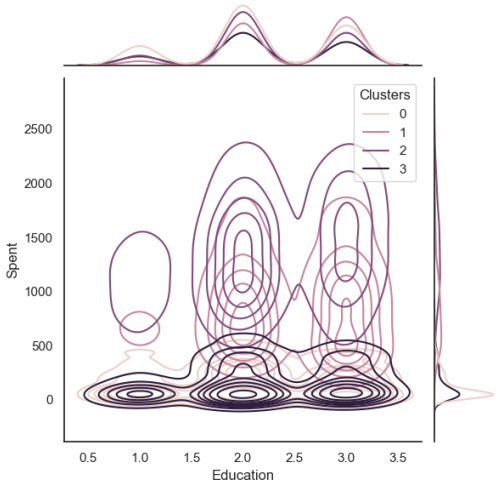

Clusters
0
1
2
3
Spent
Kidhome

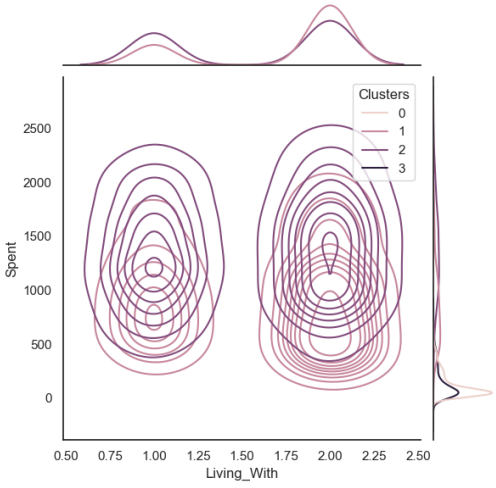

Clusters
0
1
2
3
Spent
Customer_For


Clusters
0
1
2
3
Spent
Teenhome


Clusters
0
1
2
3
Spent
Children


Clusters
0
1
2
3
Spent
Family_Size


Clusters
0
1
2
3
Spent
Education


Clusters
0
1
2
3
Spent
Living_With

Cluster 0 :
- Younger people
- More undergraduates compared to other clusters
- They live with families, not alone and not with a partner
- The least spending cluster

Cluster 1 :
- older households
- They live with a partner
- The second most spending cluster
- Maximum of 1 kid at home

Cluster 2 :
- Span all ages
- Most of them are not parents
- More likely to be living with a partner
- The tend to have achieved their graduate studies
- Most Spending Cluster

Cluster 3 :
- They are parents
- They have relatively lower income
- They have families of maximum 4 individuals
- They don't spend too much

7. **Possible Enhancements :**

- **Predictive Modeling:** Use customer segments as a target variable and build predictive models to understand which segments are more likely to respond to specific marketing campaigns. This can improve campaign targeting.

- **Longitudinal Data**: Incorporate time-series data to analyze how customer behavior changes over time and how segmentation evolves with seasonality or external factors.