COSC 4337: Data Science II
Dr. Ricardo Vilalta - Spring 2023

TEAM MEMBERS

Luong Thuy Ngo - 2048659

Syaneth Khatt - 2114829

Karla Castello - 2138671

……………………...............

PHASE 1: DATA SELECTION REPORT

INTRODUCTION

Dataset: Customer Personality Analysis

The analysis helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers. The study helps modify its product based on its target customers from different types of customer segments.

DATA DESCRIPTION

- ➢ The selected dataset is the Customer Personality Analysis dataset, sourced from the Kaggle repository. With this data, we want to see common patterns of customers to see what type of customers is more likely to buy more from this company and determine certain strategies to target these customers in future campaign

Dataset link:

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?resource=download

- ➢ Column and data type

People

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

Products

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

Promotion

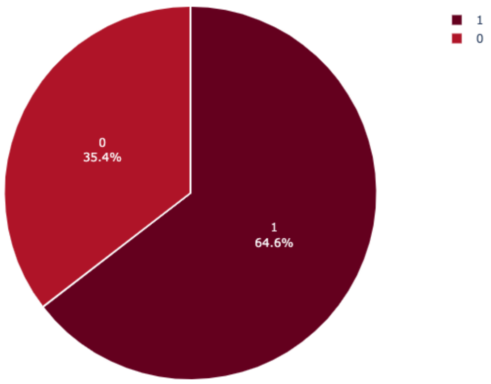
- NumDealsPurchases: Number of purchases made with a discount - integer
- AcceptedCmp1: 1 if the customer accepted the offer in the 1st campaign, 0 otherwise - integer
- AcceptedCmp2: 1 if the customer accepted the offer in the 2nd campaign, 0 otherwise - integer
- AcceptedCmp3: 1 if the customer accepted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

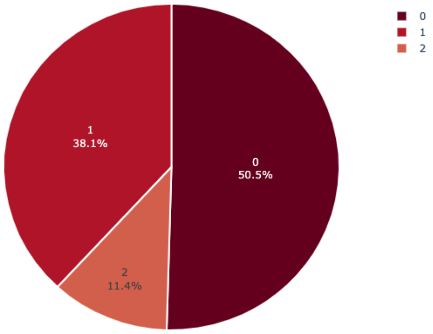
Place

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

EXPLORATORY DATA ANALYSIS

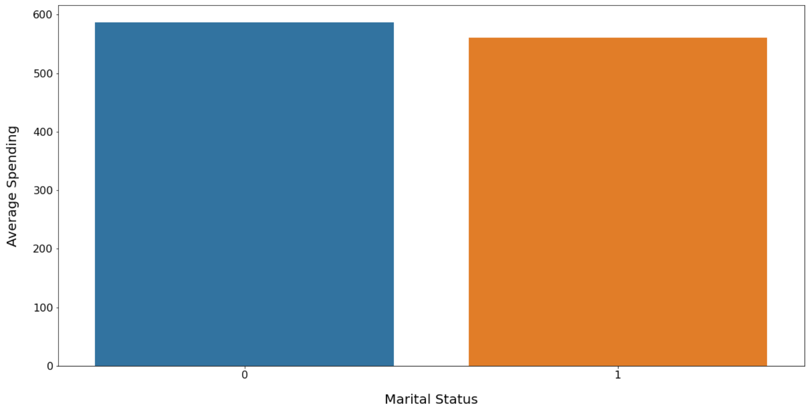
We made a few relationships between important features to easily compare data and obtain a surface clue from the dataset for further analysis.



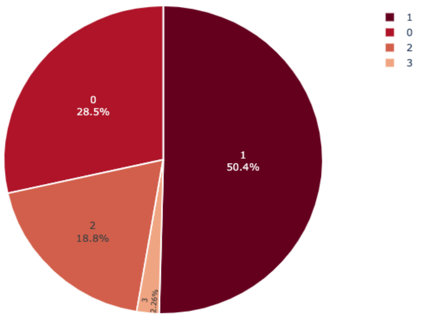


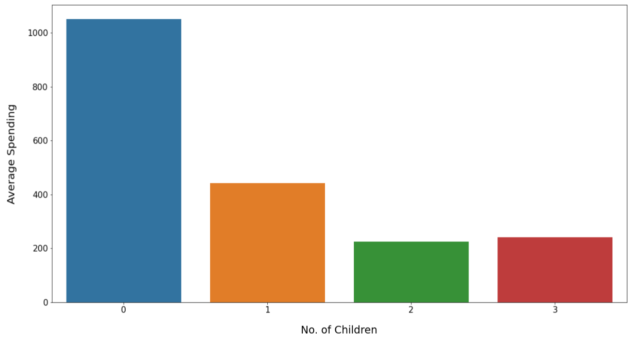
The first pie chart above shows that 64.6% are having partners(1) while 35.4% are single(0). In the Marital_Status feature, in particular, there are multi values that classify status but we categorized them into two main categories: partner(labeled as 1) and single(labeled as 0).

The next pie chart displays the Education level of customers. Half of the customers are graduate students(0). There are 38% of students are post-graduate (1) includes Master’s Degree and Phd. Only about 11% are without college degree.

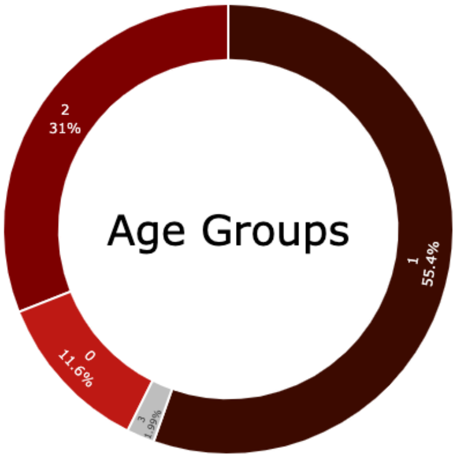


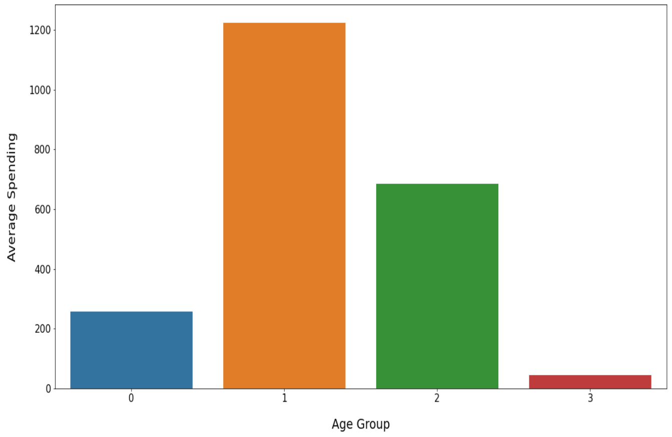
We observe that the average spending of single people (0) is slightly higher than the average spending of people with partners(1). Both of them spent approximately greater than 540.



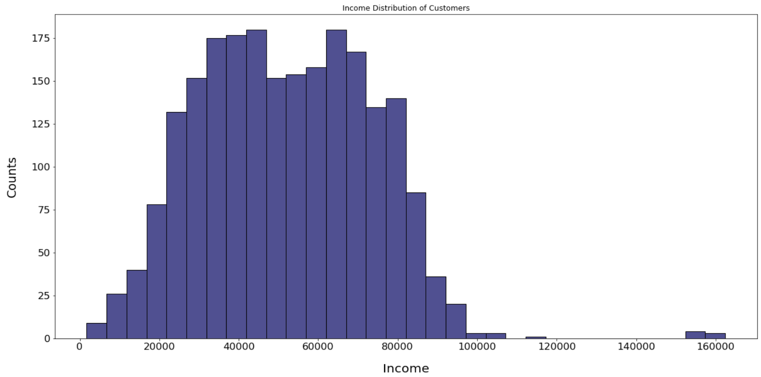


Based on the pie chart, about 50% of customers have only one child. 28% of customers do not have any children at home while about 18% of them have 2 children and followed by 2% with 3 children. We conducted a relationship between the number of children and average spending to find which house spends the most on the products in this company. It indicates that customers with no children are more likely to spend money than customers with 1, 2, and 3 children.

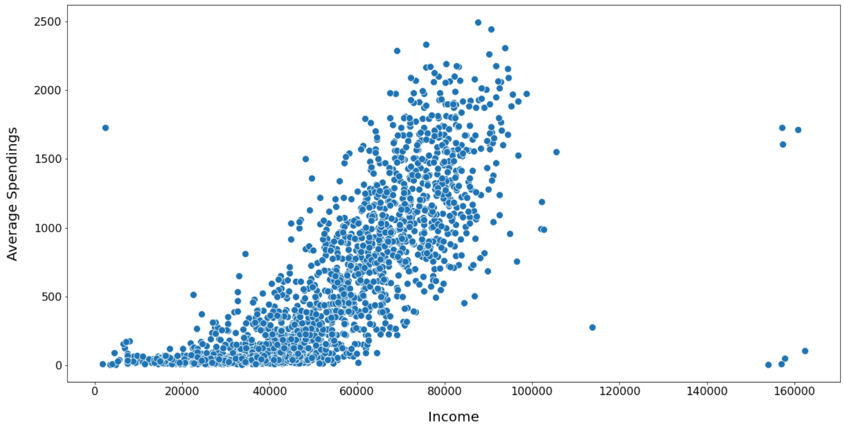




From the pie chart, it can be seen that the majority of customers consists of middle age to senior adults. More than 50% of customers are middle age adults. Likewise, middle age group of people spent much more than any other age groups including adults, senior adults, and teens respectively by looking at the relationship between age group and average spending.



There are outliers but the salaries of customers have normal distribution with most of the customers earning between 25000 and 85000.



The relationship between income and average spendings is linear which indicates that customers having higher income spent more money.

FEATURE ENGINEERING

1. Data Cleaning

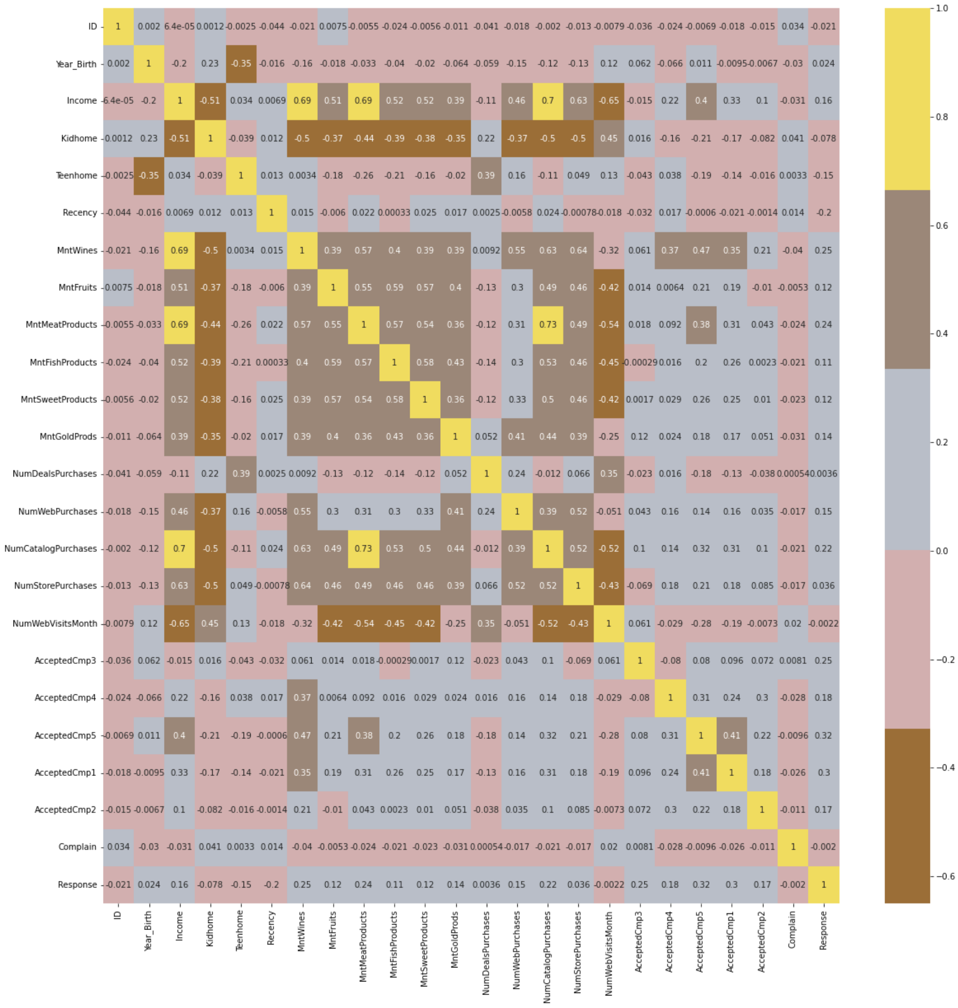
When data was imported, the first step we took was to check for missing values of the data. Out of the 2240 observations, only 24 had null values in the 'Income' column. We decided to drop these values since it was a very small percentage of the data and it would also keep our model accurate. Besides that, we also checked and removed all duplicate values in data.

Out of the raw data, the 'Income' feature had an extreme outlier. The observation was removed to reduce the standard deviation and represent real population characteristics.

From the data, we also perform the calculation of ‘Age’ based on the ‘Year_birth’ column. Then, we remove outliers in the Age columns (those with age greater than 80).

Lastly, we removed any unnecessary features that would not tell us anything about the data. We did this by checking if any features had over 99.9% of similar data in the columns. There were two features, ‘Z_CostContact’ and ‘Z_Revenue’ that had 100% of similar data, therefore those two columns were removed.

2. Feature selection



To make the data simpler and easy to read, we changed the year of birth of each customer to their age as well as found out how long each customer has been enrolled with the company based on the ‘Dt_Customer’ attribute.

In this company, there were several campaigns to which customers accepted offers, therefore we added these observations into one column that shows the total number of accepted offers from campaigns for each observation.

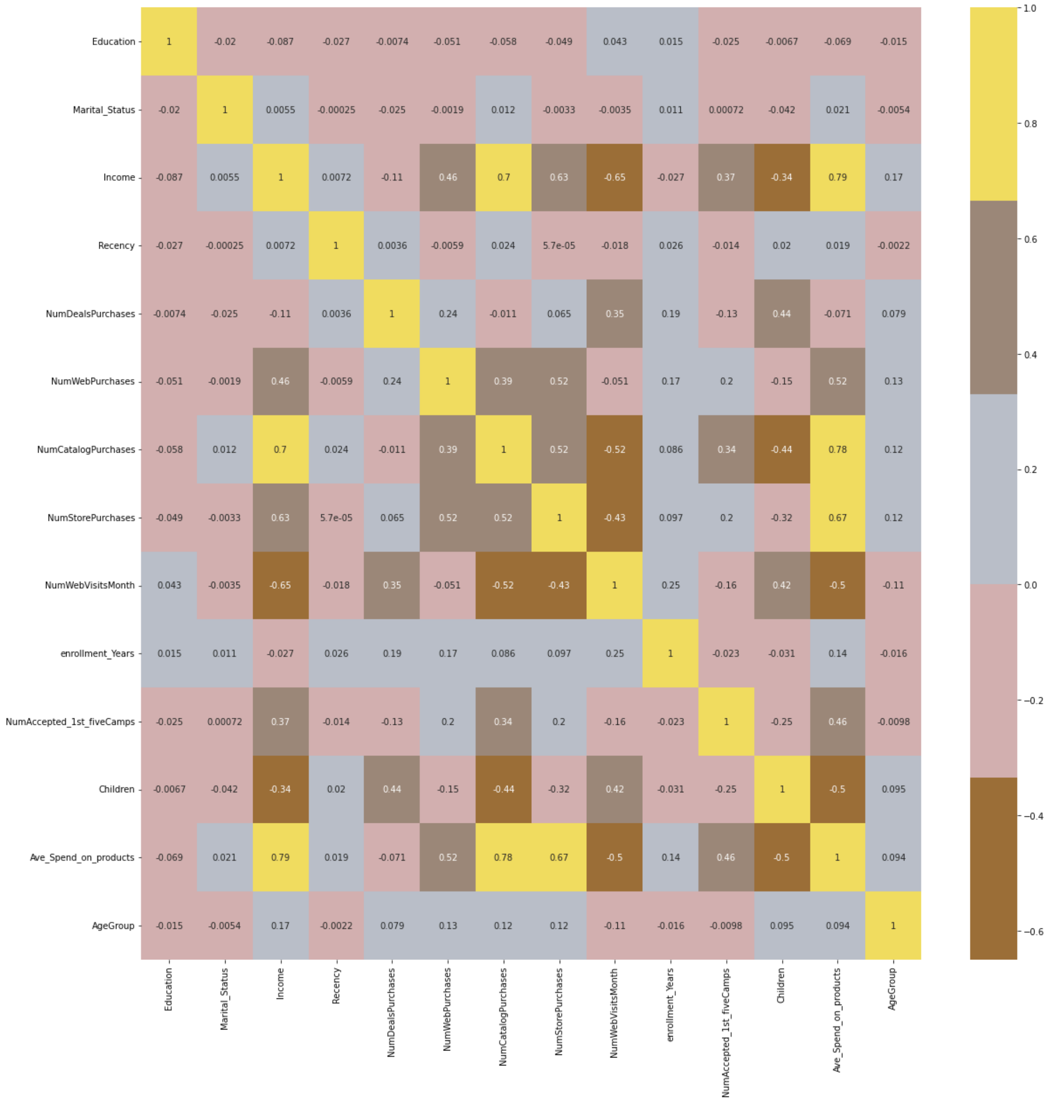
We combined the ‘KidHome’ and ‘TeenHome’ columns into ‘Children’ columns to represent the total number of children for each observation.

There were many unique values in ‘Education’ and ‘Marital Status’ features, therefore we classified marital status into two categories: single and partner, and we classified education into three categories: graduated, postgraduate, and pregraduated.

In order to better understand each customer segment, we decided to use categorical variables to group the ages of the observations. In the ‘Age Group’ attribute, we made 4 categories: teen, adult, middle age adult, and senior adult.

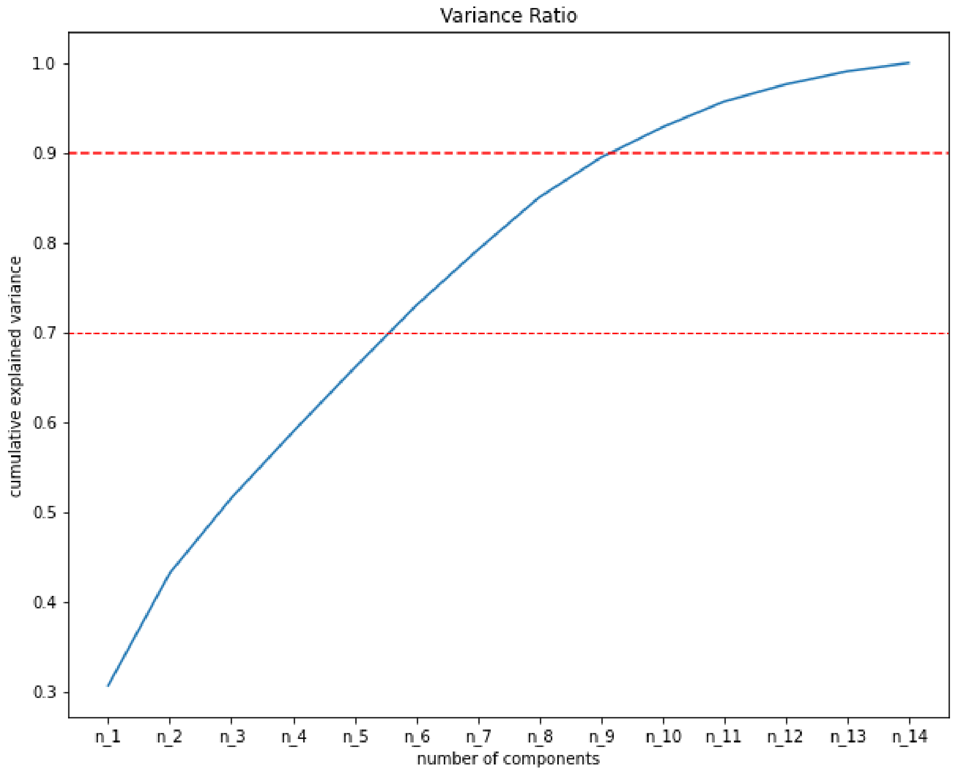
Lastly, we created a new feature by adding all the different products bought from the company and averaging the amount spent. This feature is the most important because it shows how much money these customers are puting into the company.

After adding new features and dropping irrelevant and redundant features, we have clean data with fewer features.

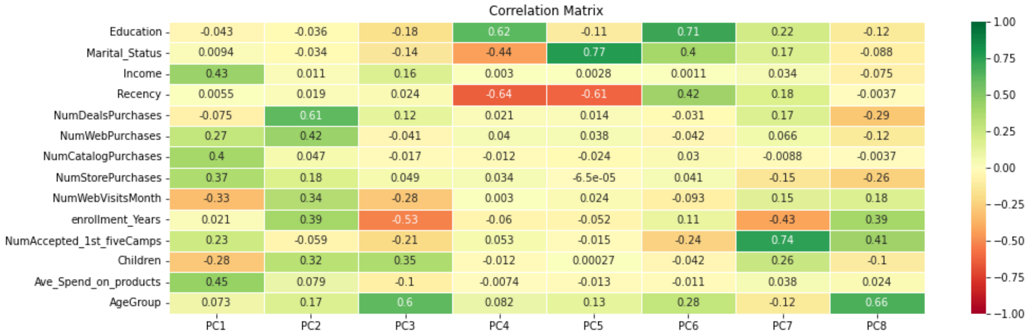


3. Dimensional reduction (PCA)

To perform PCA, we first changed the categorical variables to numerical and then standardized the data. Using the PCA library, we call fit to compute vectors that we can project our data onto. We then calculated the cumulative explained variance and plot it to decide how many principal components we should have in order to maximize variance and reduce dimensions.



The graph shows the amount of variance depending on the number of components. Here, we want to preserve around 80% of the variance, therefore we decided to keep 8 components.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Education	-0.043	-0.036	-0.18	0.62	-0.11	0.71	0.22	-0.12
Marital_Status	0.0094	-0.034	-0.14	-0.44	0.77	0.4	0.17	-0.088
Income	0.43	0.011	0.16	0.003	0.0028	0.0011	0.034	-0.075
Recency	0.0055	0.019	0.024	-0.64	-0.61	0.42	0.18	-0.0037
NumDealsPurchases	-0.075	0.61	0.12	0.021	0.014	-0.031	0.17	-0.29
NumWebPurchases	0.27	0.42	-0.041	0.04	0.038	-0.042	0.066	-0.12
NumCatalogPurchases	0.4	0.047	-0.017	-0.012	-0.024	0.03	-0.0088	-0.0037
NumStorePurchases	0.37	0.18	0.049	0.034	-6.5e-05	0.041	-0.15	-0.26
NumWebVisitsMonth	-0.33	0.34	-0.28	0.003	0.024	-0.093	0.15	0.18
enrollment_Years	0.021	0.39	-0.53	-0.06	-0.052	0.11	-0.43	0.39
NumAccepted_1st_fiveCamps	0.23	-0.059	-0.21	0.053	-0.015	-0.24	0.74	0.41
Children	-0.28	0.32	0.35	-0.012	0.00027	-0.042	0.26	-0.1
Ave_Spend_on_products	0.45	0.079	-0.1	-0.0074	-0.013	-0.011	0.038	0.024
AgeGroup	0.073	0.17	0.6	0.082	0.13	0.28	-0.12	0.66

We use heatmap to easily identify the contribution and value of eigenvector to its principal component. We determined the level of correlation above 0.4 is deemed important. For example, in PC1, we observe that Income and Average spent on products are positively related which can be interpreted as customers with higher income purchased more products from this company on average. In PC2, when there were good deals, we are able to see the increase in number of purchase on company's website. PC3 is a measure of number of years enrolled in company and where the customer lies in age. PC3 distinguished older age groups and newer customers. PC4 is a measure of the education level, marital status of the customer and the number of days since the customers last purchase. PC4 distinguishes customers with higher education, customers who are single and less days since their last purchase. PC5 distinguishes customers who have partners and have a more recent purchase with the company. PC6 distinguishes customers with higher education that have partners and have more days since their last purchase. PC7 distinguishes customers who are more recently enrolled into the company and have accepted more offers in the campaigns. Lastly, PC8 distinguishes customers in higher age groups (older people) and that have accepted more offers in the campaigns.

REFERENCES

Conor O'Sullivan, Towards Data Science, "A Step-By-Step Introduction to PCA

https://towardsdatascience.com/a-step-by-step-introduction-to-pca-c0d78e26a0dd

Patel, Akash. , Kaggle, 2022, "Customer Personality Analysis"

www.kaggle.com/datasets/imakash3011/customer-personality-analysis.

PennState: Statistics Online Courses, "11.6 - Example: Places Rated after Standardization: STAT 505"

https://online.stat.psu.edu/stat505/lesson/11/11.6

Seungbum Lim, Kaggle, 2022, "How to select the optimal number for PCA, Kmeas"

https://www.kaggle.com/code/seungbumlim/how-to-select-the-optimal-number-for-pca-kmeans#5.-%EC%A3%BC%EC%84%B1%EB%B6%84%EB%B6%84%EC%84%9D(PCA)%EC%9D%84-%EC%9D%B4%EC%9A%A9%ED%95%9C-%EB%B3%80%EC%88%98-%EC%B6%94%EC%B6%9C

INSTRUCTIONS TO RUN THE CODE

1. Download the dataset from this link:

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/download?datasetVersionNumber=1

2. Extract the zip folder to get the file named ‘marketing_campaign.csv’
3. Place the dataset into the same folder as the file named ‘notebook_firstDeliverable.ipynb’
4. Open the file ‘notebook_firstDeliverable.ipynb’ with Jupyter notebook
5. Run the code (will need to download the necessary libraries if it is not pre-installed on your computer)