Customer Personality Analysis

Yuechen Liu

12/21/2021

BIS 634

Supervisor: Robert McDougal

Introduction

Customer personality evaluation is an in-depth evaluation of a “perfect” consumer for a business enterprise. This facilitates organizations higher apprehend their clients and makes it simpler for them to promote their merchandise in keeping with the precise needs, behaviors, and issues of various kinds of clients.

Customer personality evaluation facilitates organizations to adjust their product primarily based on target clients from different kinds of consumer segments. For example, instead of spending money to promote a brand-new product to every consumer, the company can analyze which type of customers will most likely buy the product, and only market the product to a specific group of people, which can set the target market more precisely. Besides, the company can predict customers’ consumption motivation by collecting their basic information, in order to better adjust or improve the business models [1].

Data Resources and FAIRness

The original dataset is provided by Dr. Omar Romero-Hernandez from Kaggle public data repositories, named Customer Personality Analysis. The metadata of the dataset is available and licensed on Kaggle.

FAIRness

I. Findability: The dataset is public and can be searched through the Internet.
II. Accessibility: People can copy, modify, distribute and perform the work without asking permission. The dataset is accessible and downloaded by anyone via Kaggle API.
III. Interoperability: The dataset is stored in .csv format, and it uses a formal, accessible, shared, and broadly applicable language for information representation.
IV. Reusability: The dataset is published with a clear and accessible data usage license, with a clear and detailed description of the file content, columns, provenance, and license specifications.

Objective

After looking through the data set, there are several possible questions that are raised to analysis:

1. The correlation between expenses and other possible factors.
2. Who are the target customers?
3. What characteristics do they have?
4. What is the customer's consumption motivation?
5. How can we predict the customer consumption ability?

Exploratory Data Analysis

Data Overview

The dataset size is 507.6+ KB in .csv format, and it has 29 columns, 2,240 rows.
Data Dictionary

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, Teenhome: Number of children\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, MntFruits : Amount spent on wine\fruits in last 2 years
- MntMeatProducts, MntFishProducts: Amount spent on meat\fish in last 2 years
- MntSweetProducts, MntGoldProds: Amount spent on sweets\gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1-5: 1 if customer accepted the offer in the 1st-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

Data Processing

1. Drop missing value and unuseful columns

   There are 24 missing values in the "Income" column. Categories "Z_CostContact" and "Z_Revenue" have the same value in all the rows, as a result, they are not going to contribute anything to the model building. So we drop missing values and categories that are not useful for this assignment: "ID", "Z_CostContact", and "Z_Revenue".

2. Check outliers

   The outliers in "Income" are people who are extremely rich. I decided to remove them by filtering "Income" < 500,000.

   The outliers in "Year_Birth" are people who are extremely old. I decided to remove them by filtering "Year_Birth" > 1920.

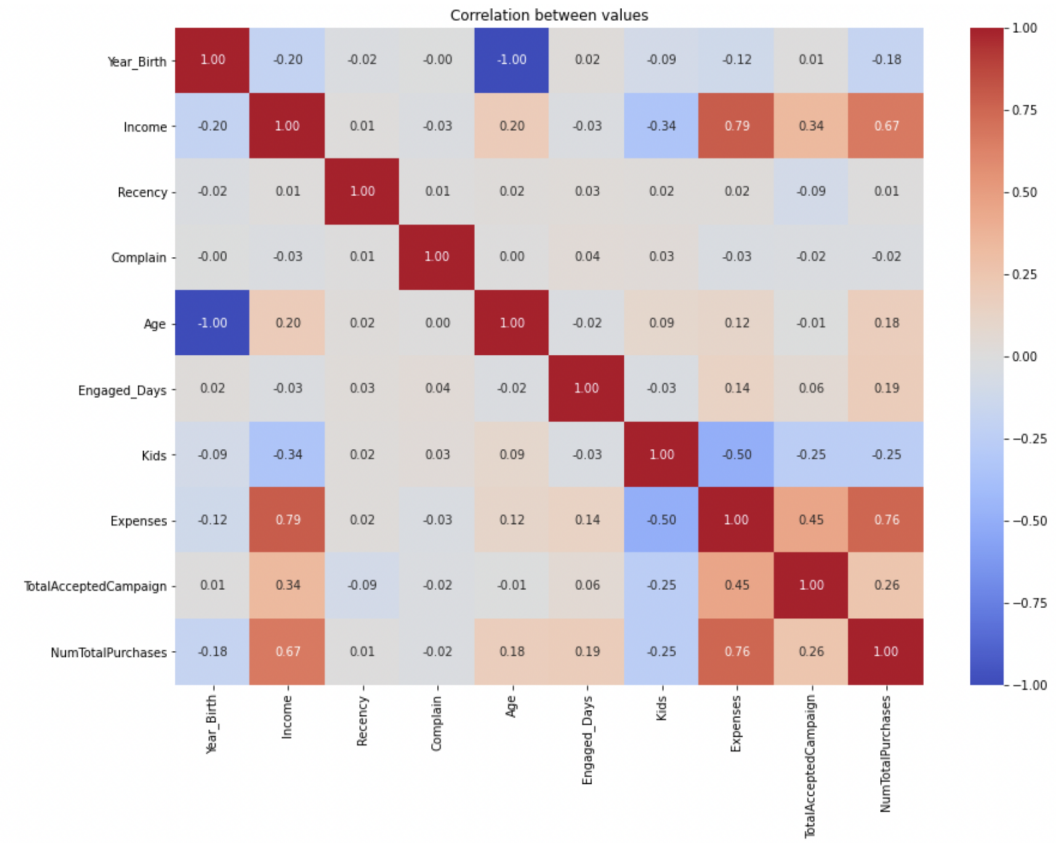
3. Combine different variables into one variable to reduce the number of dimensions

   There are some variables that can be grouped into one variable. I created 6 new columns, "Education", "Marital_Status", "Kids", "Expenses",

“TotalAcceptedCampaign”, and “NumTotalPurchases”. See details in the below table.

New Columns	Replaced Variables	New Variables
Education	"Graduation", "PhD", "Master", "2n Cycle"	“Graduate”
Education	"Basic"	"Undergraduate"
Marital_Status	'Married', 'Together'	'Relationship'
Marital_Status	'Divorced', 'Widow', 'Alone', 'YOLO', 'Absurd'	'Single'
Kids	'Kidhome' + ‘'Teenhome'	
Expenses	'MntWines' + 'MntFruits'+ 'MntMeatProducts' + 'MntFishProducts' + 'MntSweetProducts' +'MntGoldProds'	
TotalAcceptedCampaign	'AcceptedCmp1' + 'AcceptedCmp2' + 'AcceptedCmp3' + 'AcceptedCmp4' + 'AcceptedCmp5'+'Response'	
NumTotalPurchases	'NumWebPurchases'+ 'NumCatalogPurchases'+ 'NumStorePurchases' + 'NumDealsPurchases'	

4. Correlation between values



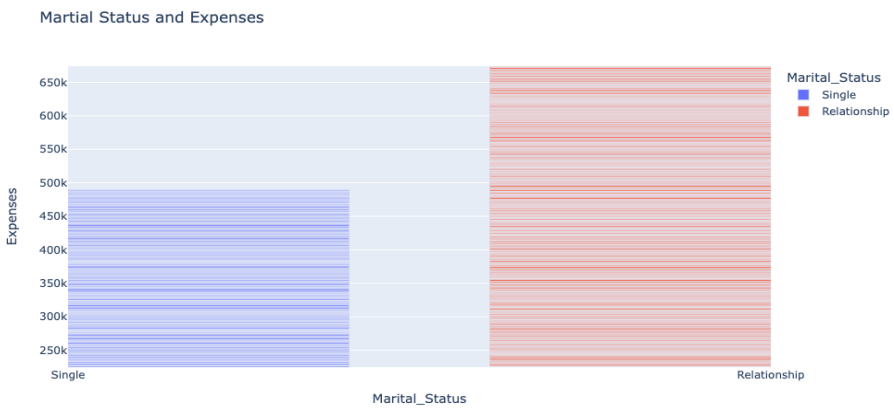
When we look at the correlation table, all the data looks clean. Assuming a strong relationship above 0.70, it is obvious that there are some strong positive relationships between income and expenses, expenses, and the number of total purchases. The higher expenses are associated with higher income.

Besides, there are some moderate relationships, such as kids' number and expenses having a moderate negative correlation.

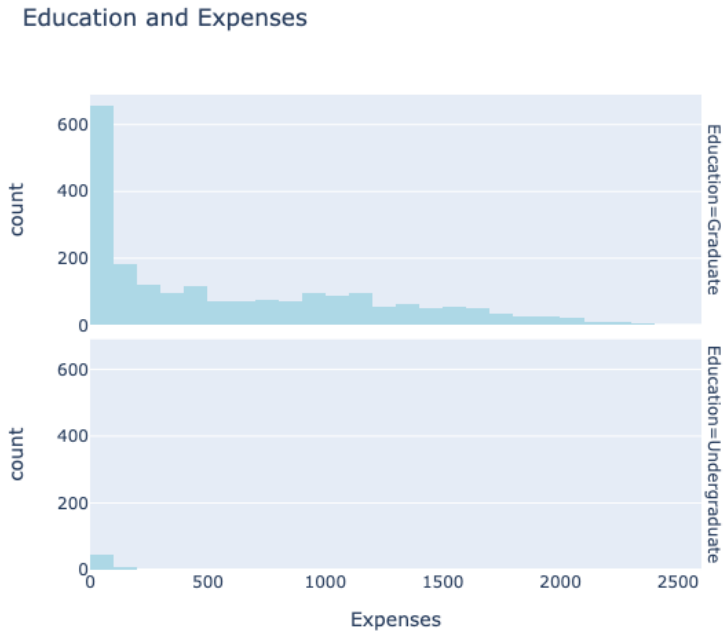
Data visualization

To better understand the relationship between each factor and "expenses", I draw graphs to visualize the possible associations.

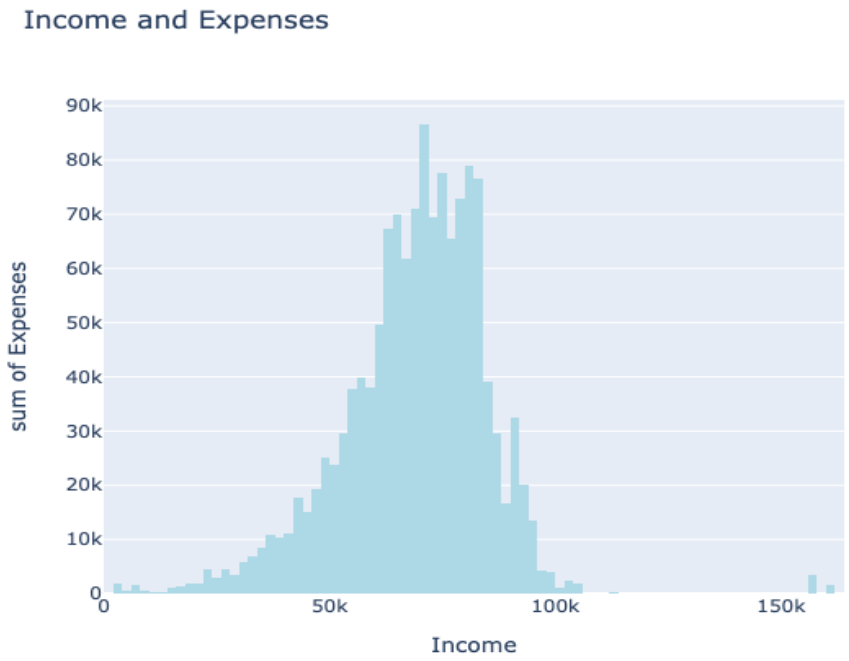
a) Relationship between marital status and expenses: customers who are in relationship tend to have higher expenses.



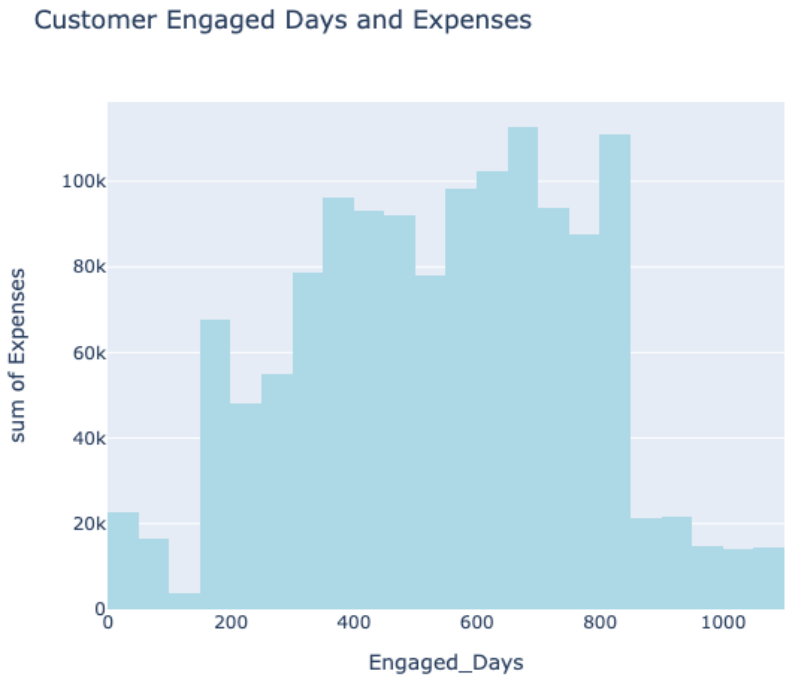
b) Relationship between education and expenses: customers who are graduates tend to have higher expenses.



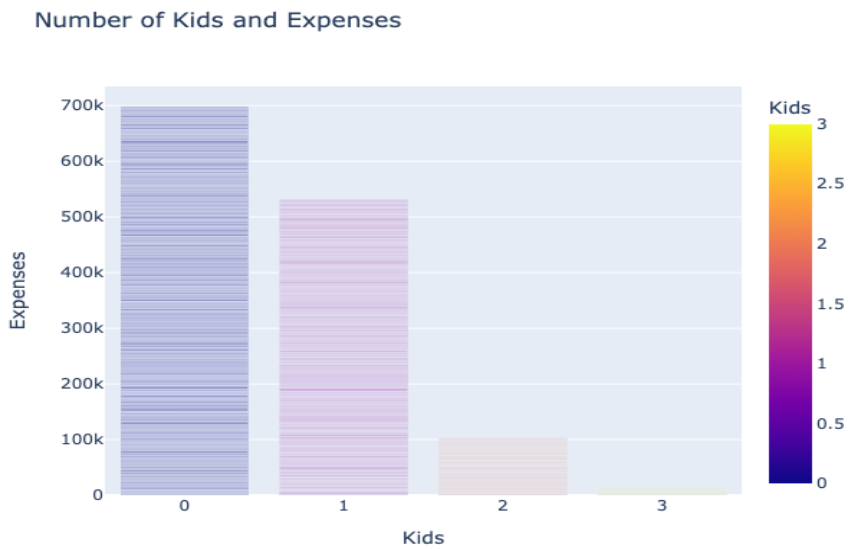
c) Relationship between income and expenses: customers who have higher income (around 80k) tend to have higher expenses.



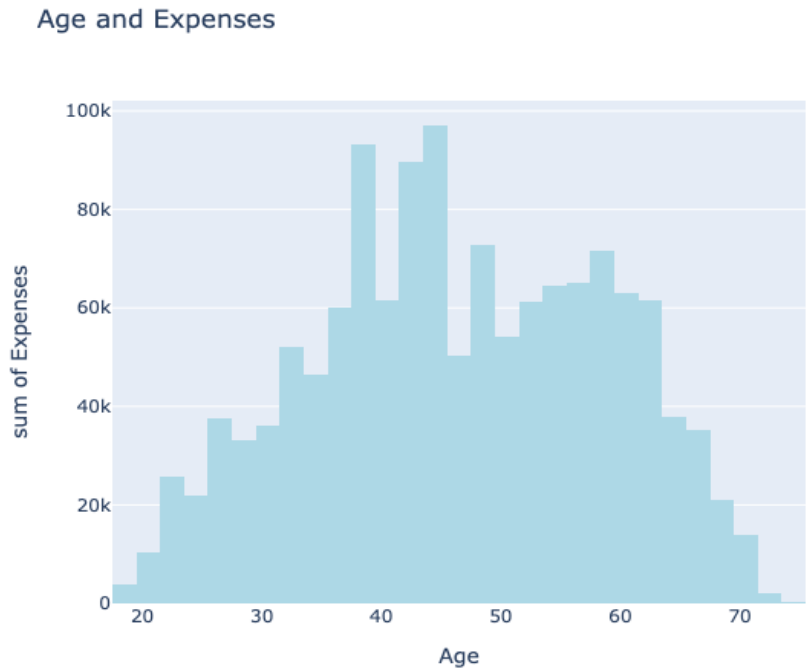
d) Relationship between engaged days and expenses: customers who engage in the company longer (650-700 days) tend to have higher expenses.



e) Relationship between number of kids and expenses: customers who have fewer than 1 kid tend to have higher expenses.



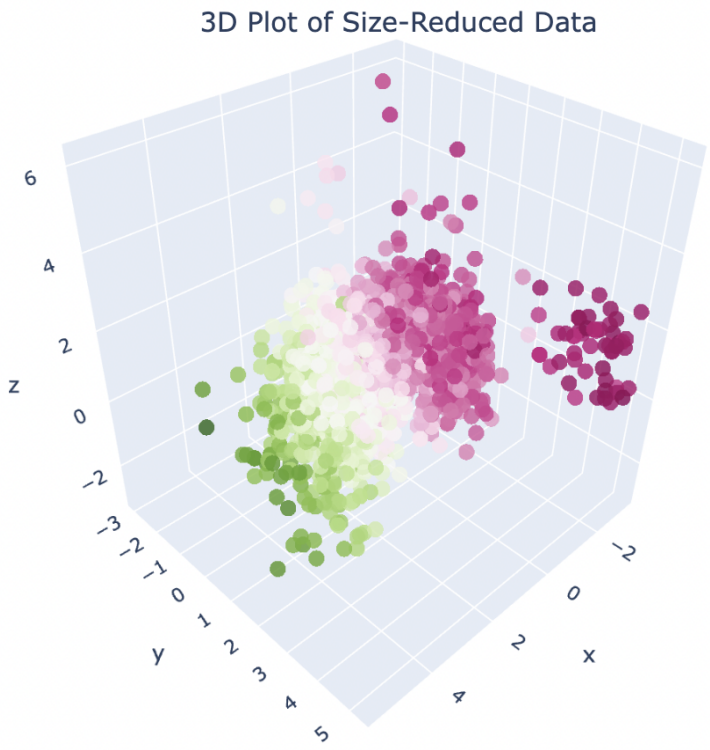
f) Relationship between age and expenses: customers who are around 40 years old tends to have higher expenses.



PCA

Process high-dimensional data will cause high processing power and cost. The higher the number of features in this dataset, the more difficult to deal with. Thus, it is important to do dimension reduction before we do the K-Means Clustering.
I use PCA to preprocess the data to perform K-Means Clustering.

1. LabelEncoder(): encode categorical columns with value between 0 and n_classes-1,
2. StandardScaler(): standardization process by removing the mean and scaling to unit variance.
3. PCA () to reduce feature dimensions to 3

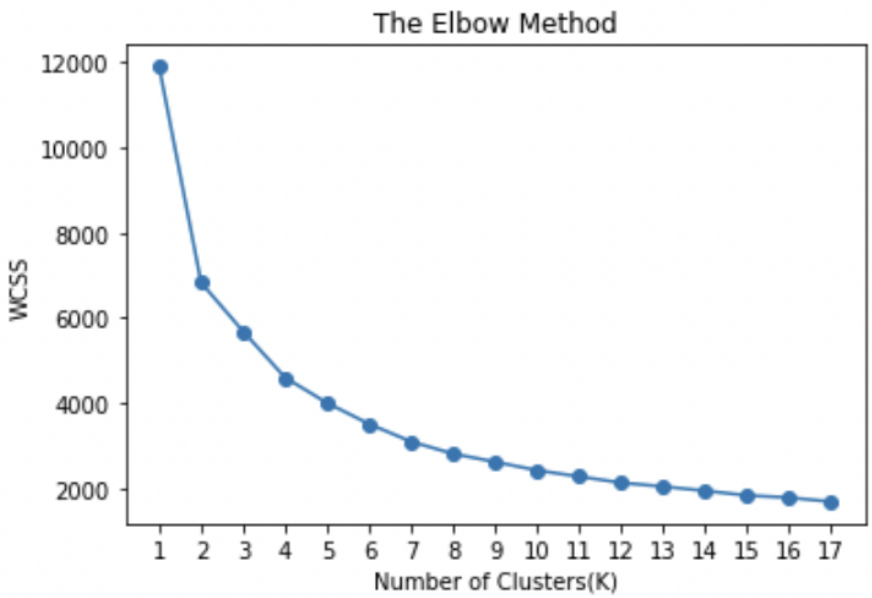


Based on the above PCA plot, we can clearly see that there are three clusters.

Use Elbow Method to find the optimal number of clusters

We cannot randomly select the number of clusters from a dataset. Each cluster is

formed by calculating and comparing the distances of data points within a cluster to its center. As a result, we calculate the Within-Cluster-Sum-of-Squares (WCSS) to find the right number of clusters. WCSS is the sum of squares of the distances of each data point in all clusters to their respective centers, and the goal is to minimize the sum. Assume we have n observations in a dataset, and we specify n number of clusters, which means k = n; so WCSS turns to 0 since data points themselves become centers and the distance will be 0, in turn this will perform a perfect cluster; but this is almost impossible as we have many clusters as the observations. Thus, we use Elbow Method to find the optimum value for k by fitting the model in a range of values of k. We randomly initialize the K-Means algorithm for a range of k values and plot it against the WCSS for each k value.

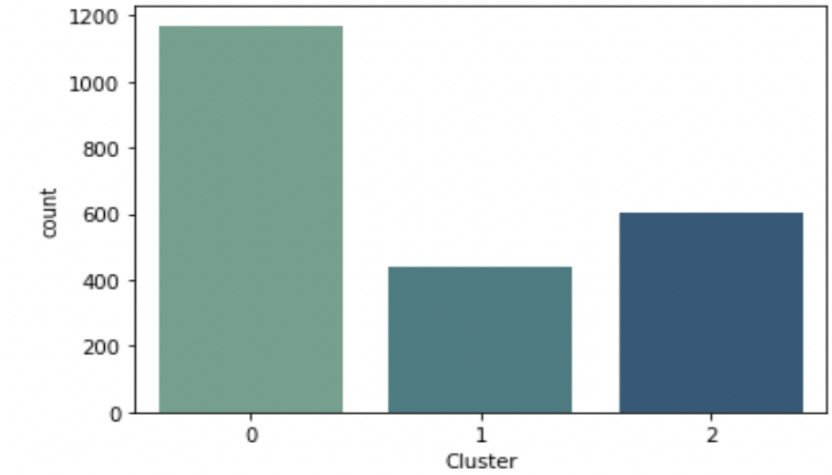


For the above given graph, the optimum value for k would be 3. As we can see that with an increase in the number of clusters, the WCSS value decreases. We select the value for k, the "elbow", on the basis of the rate of decrease, to indicate the model fits best at that point. In the graph, from cluster 1 to 2 to 3 there is a huge drop in WCSS. After 3 the drop is minimal, thus we chose 3 to be the optimal value for k. Based on the Elbow Method, we can find the optimal number of clusters is 3.

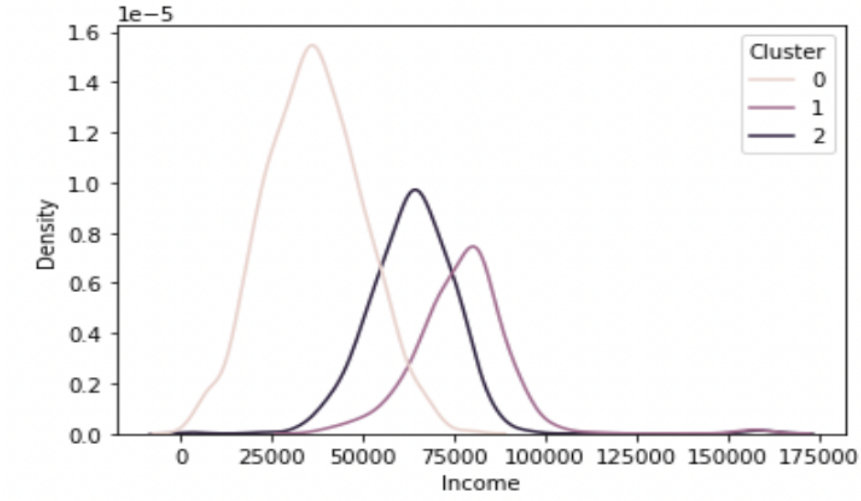
K-Means Clustering

I use K-Means Clustering to find the most optimal customers. I set k as 3, which means the result will be 3 clusters, and each cluster will share similar characteristics, and I use income level to group customers into three types: “Bronze Customer”, “Silver Customer”, and “Gold Customer”, and compare each groups’ characteristics from 5 factors: Education, Marital Status, Kids, Age, and Engaged Days.

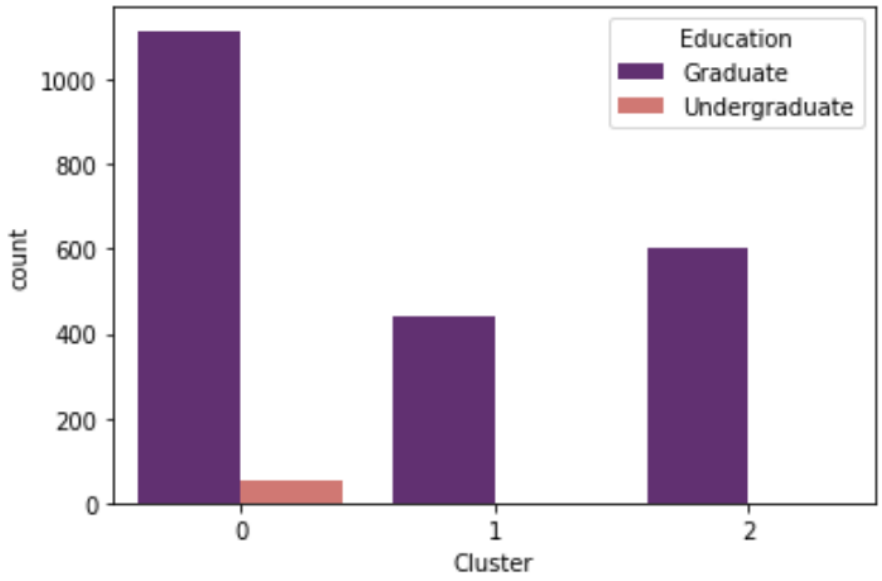
The number of people in each cluster.



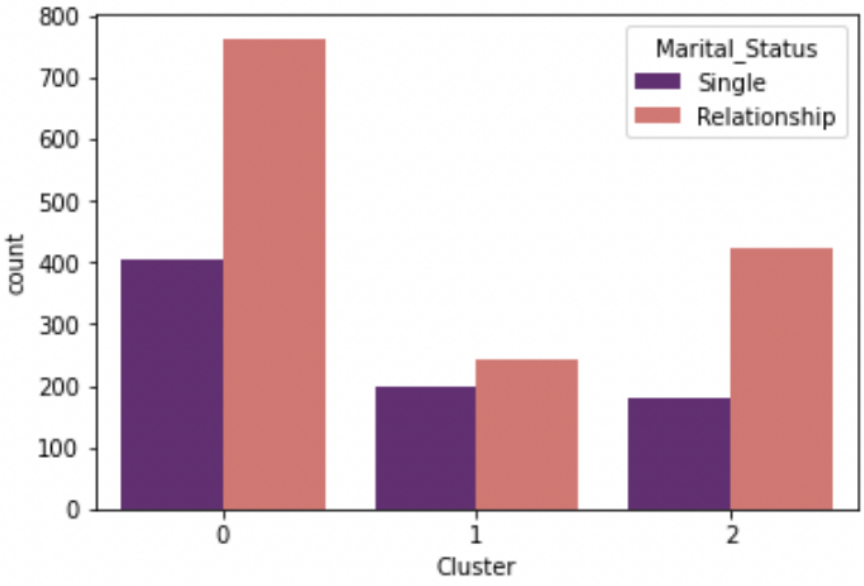
Income situation in each cluster.



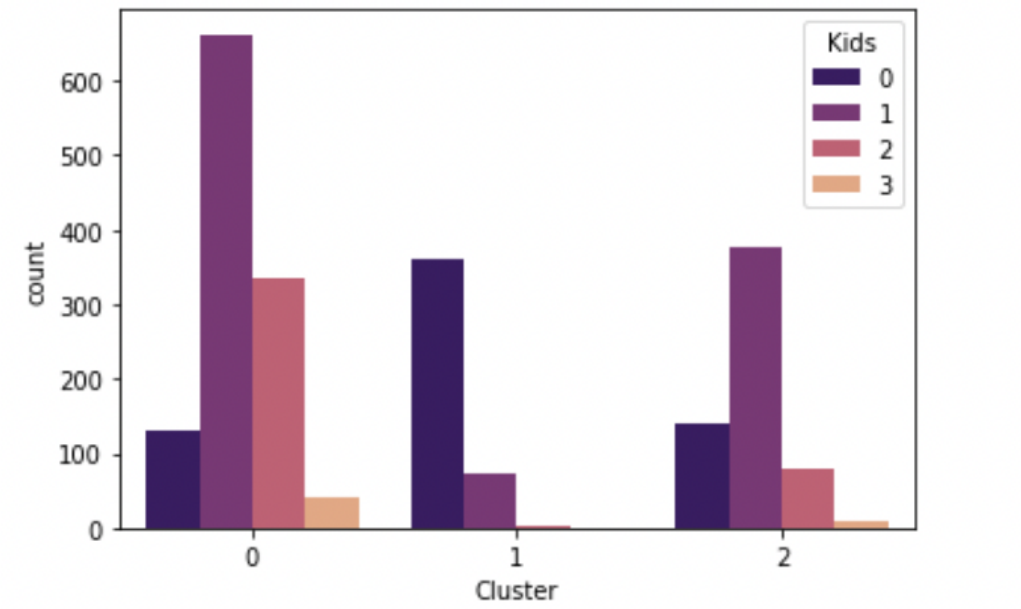
Education situation in each cluster.



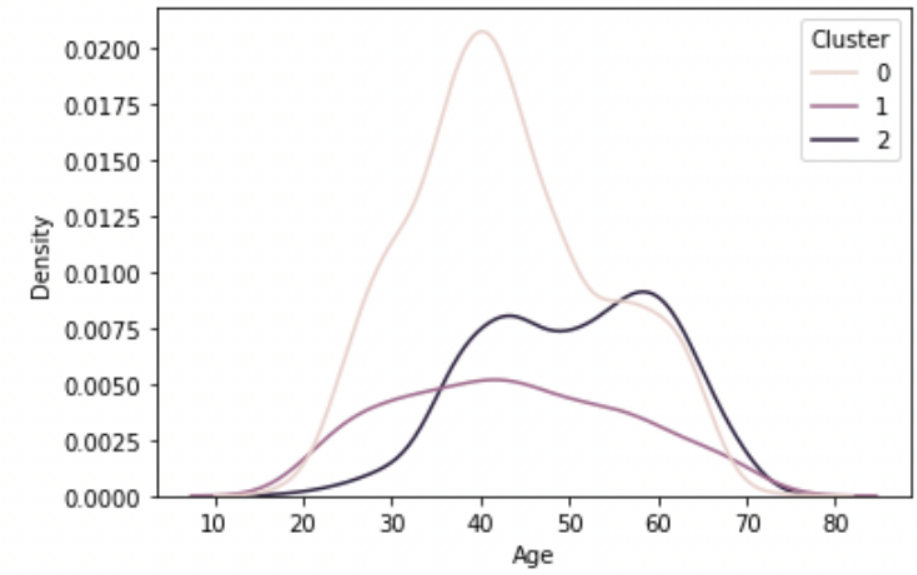
Marital status in each cluster.



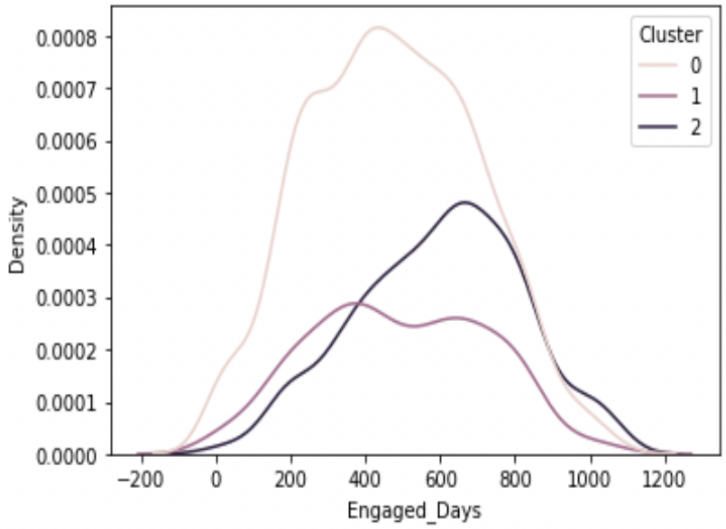
Kids number in each cluster.



Age situation in each cluster.



Engaged days with the company in each cluster.



Based on the K-Means clustering, I have the following results.

Cluster 0: Lowest to moderate income: Lowest expenses—> “Bronze Customer”

- Education: the proportion of undergraduates is higher than other clusters.
- Marital Status: more are in a relationship.
- Kids: most of them have more than 1 kid.
- Age: most of them are around 40 years old.
- Engaged Days: most of them have enrolled in the company for around 400 days.

Cluster 1: Highest income: Highest expenses—> “Gold Customer”

- Education: graduates.
- Marital Status: approximately half of them is in a relationship; half of them are not.
- Kids: most of them do not have kids.
- Age: most of them are around 40 years old.
- Engaged Days: most of them have enrolled in the company for around 400 days and 700 days.

Cluster 2: Moderate income: Moderate expenses—> “Silver Customer”

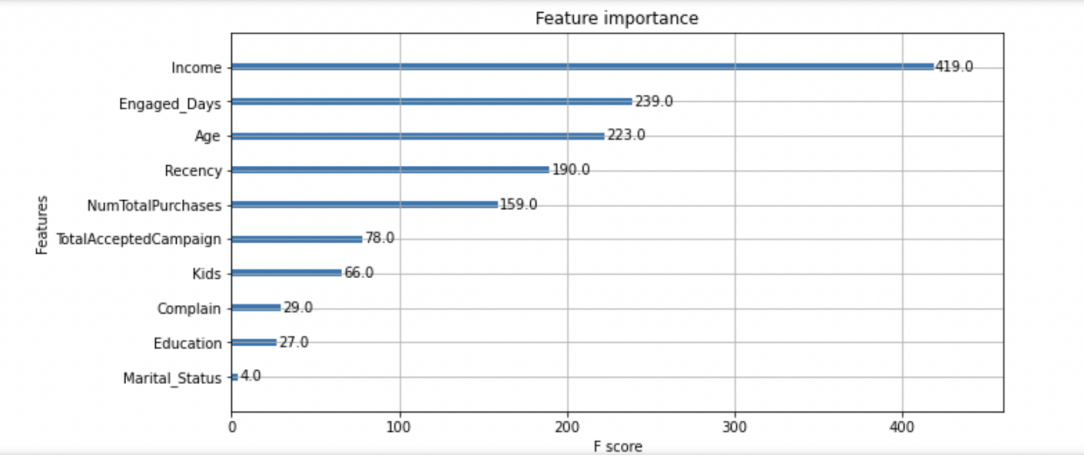
- Education: graduates.
- Marital Status: more are in a relationship.
- Kids: most of them have 1 kid.
- Age: most of them are around 40 years old and 60 years old.
- Engaged Days: most of them have enrolled in the company for around 650 days.

Regression Model

To understand which characteristics are the most related to expenses, I use XGBoost for this linear regression predictive modeling. The dataset I used is after cleaning and is stored into a new .csv file called “data.csv.” I scale the dataset and use 80% of the data to train the model; 20% of the data for testing to see if there exists significant loss.

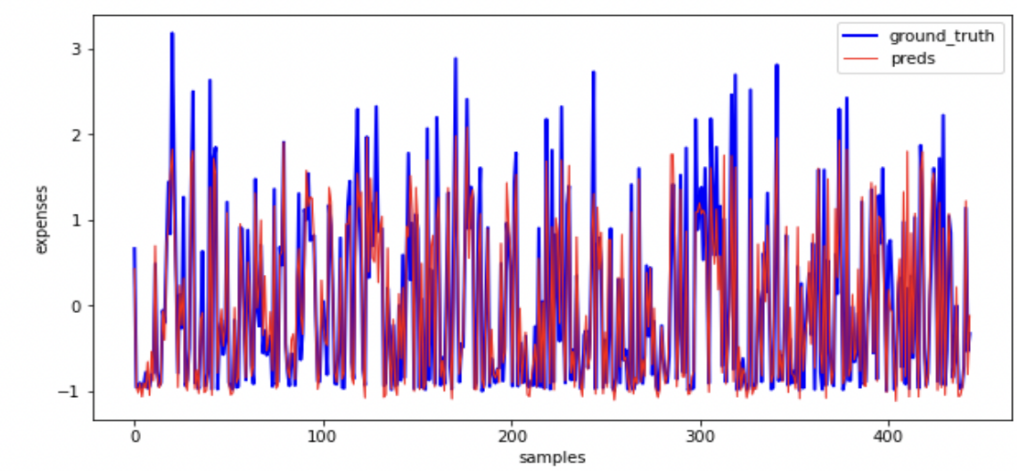
Y = ax+b

I choose "Education","Marital_Status","Income","Recency","Complain","Age", "Engaged_Days","Kids","TotalAcceptedCampaign","NumTotalPurchases" as X; “expenses” as Y.

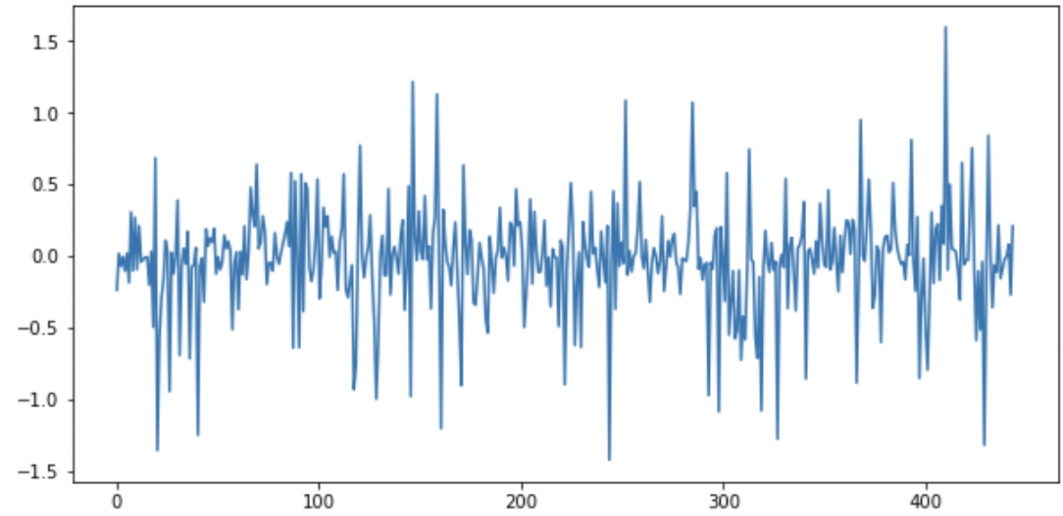


From the above graph, we can see that “Income” has the highest F score, which means that “Income” is most related to “Expenses”: with the higher “Income”, the “Expenses” also increases; other influential factors are “Engaged_Days”, “Age”, “Recency”, NumTotalPurchases”, “TotalAcceptedCampaign”, “Kids”, “Complain”, “Education”, “Marital_Status”.

The model can use different factors to predict expenses. To test the model performance, I first use the model to predict the result, and then use the real data to test if the predicted data is matching the real data.



From the above graph, we can see that the result is highly matching. Next, we will have a more direct look at the loss in the model.

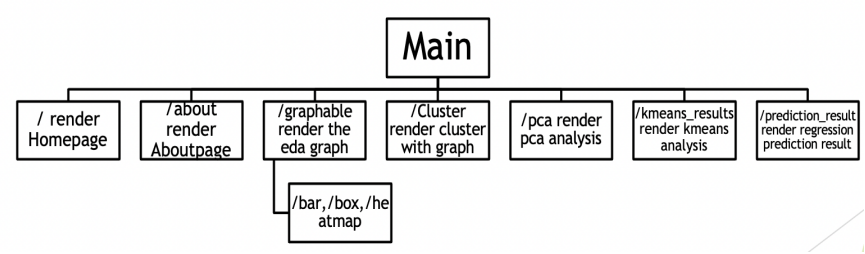


As we can see, the loss is varied around 0.0, which means that the model is relatively accurate.

API and Web Front-end

API Server

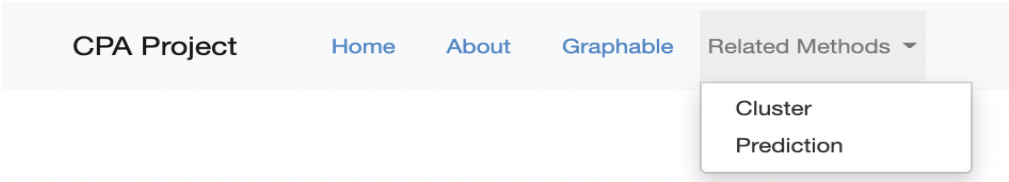
I create a Flask app and set configurations. The API server framework is as below.



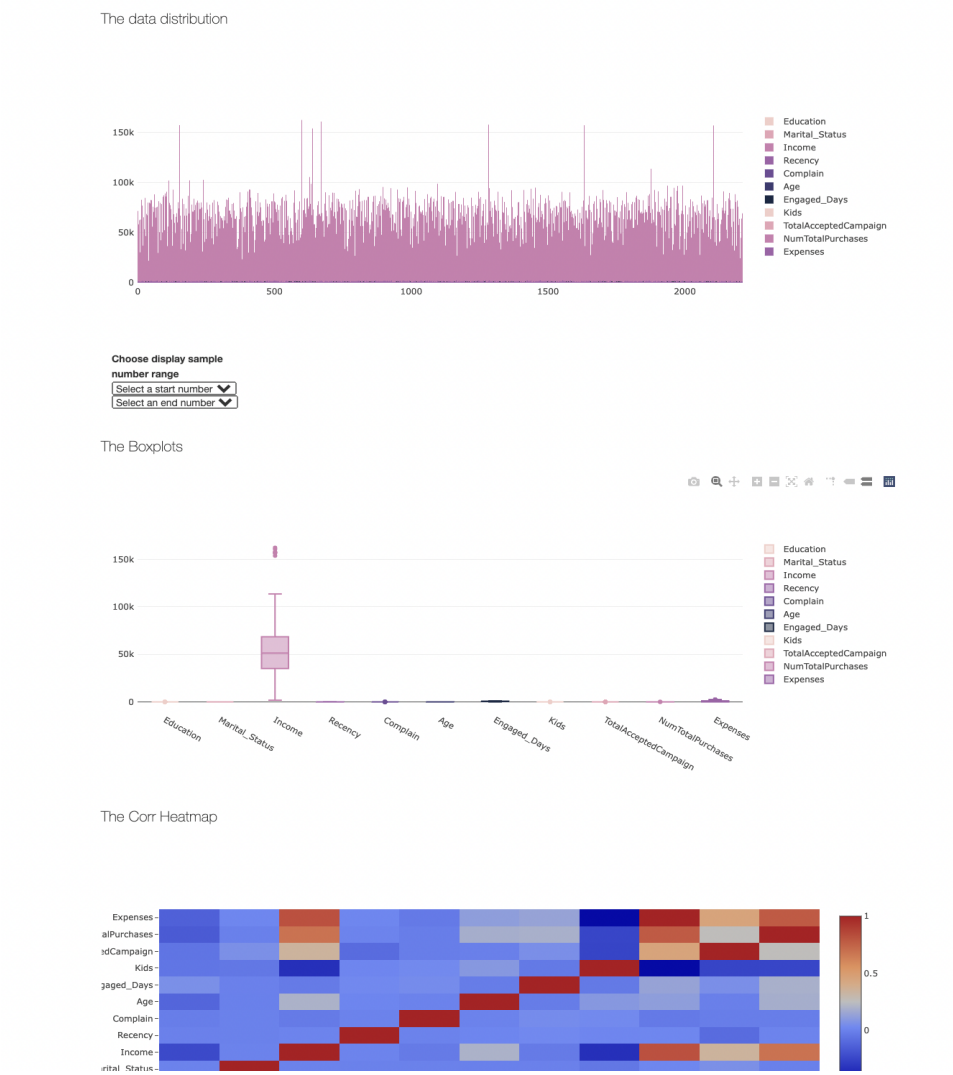
Web Frontend

The web frontend has 6 pages.

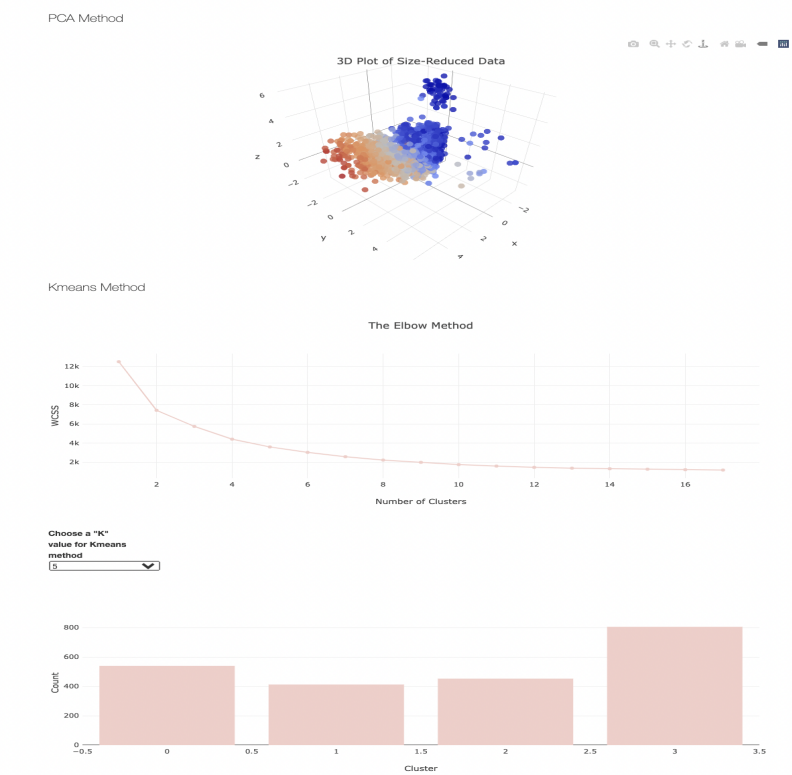
1. The Homepage and About page are the project overview, including the project background, metadata, content overview.



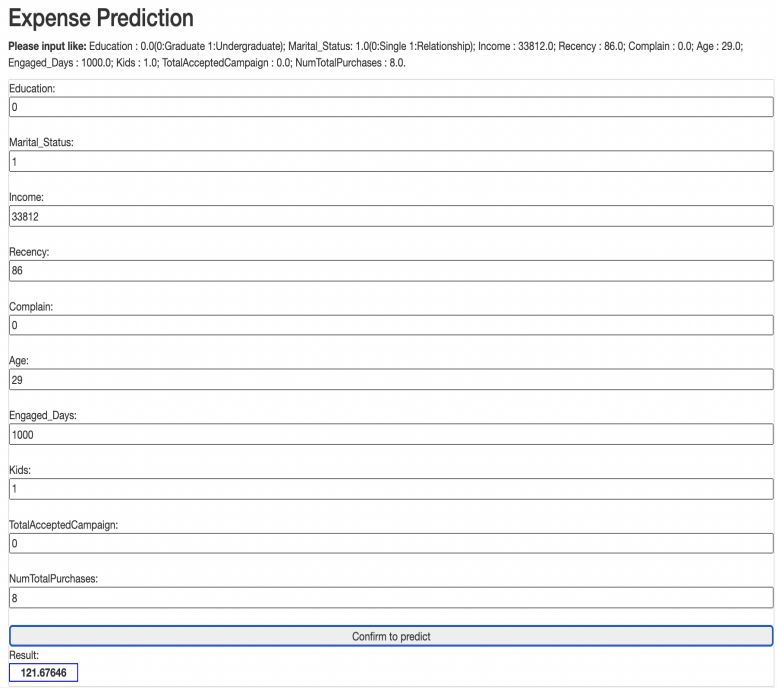
2. The Graphable page contains the summary statistics of different data distributions. You can click on the specific variable and see the data distribution; also, you can select a sample number range to display. Besides, you can choose the Boxplots and Correlation Heatmap.



3. The Related Methods page has two pages. The first one is the Cluster page, including PCA Method and Kmeans Method. You can click on those two methods to see the results and graphs. For Kmeans Methods, you can select different “K” numbers to get different senes of clusters distributions.



4. The second page is Xgboost Prediction. You can see the result of the project. More importantly, you can have expense prediction by simply typing in the number into different factors, and clicking on “Confirm to predict”, to get the predicted expense of customers.



Discussion

It is critical for companies to locate the target market and promote their products to the customers who have the most possibility to buy. Based on the above customer personality analysis, I have several findings and recommendations.

1. Interestingly, the correlation between kids' numbers and expenses is a negative correlation. After I did the research, one journal gives me the answer: "The relationship between income and family size, which is hypothesized to be positive, often is negative in empirical studies. This perverse result is thought to occur because of the many correlations between income and other factors that affect fertility. In this research, these other factors--such as the net price of a child, the opportunity cost of the wife's time, and supply factors--are statistically controlled, and the income effect is positive and significant. When the net price of a child is not controlled, however, the income effect becomes negative and significant. " [2]
2. Direct marketing. Different types of customers have different needs and consumption standards, the company marketing department can be based on that to sell the product. For example, the most expensive and highest quality product can be marked to the "Gold Customer", who has the highest expenses. This strategy can be realized by email, online adverts, promotion letters, etc.
3. Predict consumer motivation. The company can use the model to improve sales forecast by predicting the expense based on costumer's motivation (income, age, etc.); create product promoters; target more active buying customers.

Difficulties

1. This is my first time independently investigating a dataset without being instructed on how to analysis step by step. Thus, I had a hard time selecting an appropriate analysis way.
2. It is not easy to find a "perfect" dataset. I have changed one dataset because the older one has many missing values, and the association between different variables is not significant, therefore I felt difficult to find the research topic based on that dataset.

Future Work

There are many variables can be analyzed in the future work.

1. Investigate marketing places. I did not investigate variables that are related to the place. This can be further researched to help companies to find the appropriate marketing place for the specific type of customers.
2. Classify expenses. Expenses can be classified into different types of products, such as food, commodity, luxury, etc. With more detailed categories and information, the customer can be targeted more precisely.
3. Find more associated factors. Other factors may be related to expenses, such as occupation, health condition, location, etc.

Reference

1. https://www.kaggle.com/imakash3011/customer-personality-analysis
2. https://www.jstor.org/stable/2061527