

EAI6010 Applications of Artificial Intelligence

Submitted By: Priyanka Balaji

Instructor: Abhijit Sanyal

**Introduction**

In this assignment, I have been given a customer dataset that contains 5000 rows and 59 columns that describe customer’s region, age, demographics, gender, lifestyle, telecommunication related information, card and loan debts related data. Our goal is to analyze the key features, do an exploratory analysis of the data, perform customer segmentation and profiling for the given data to provide market strategies and recommendations. I have performed my analysis in Python.

**Importing Libraries:**

We first would need to import all the required libraries needed for our analysis.



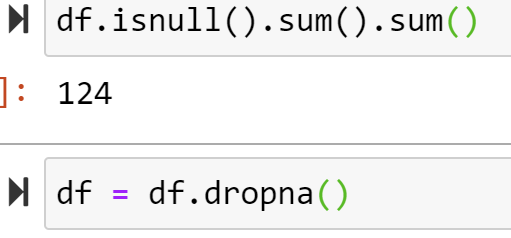
**Data Wrangling:**

Next, we would read our csv file and view the top 5 rows and columns of our dataset.



**Data pre-processing:**

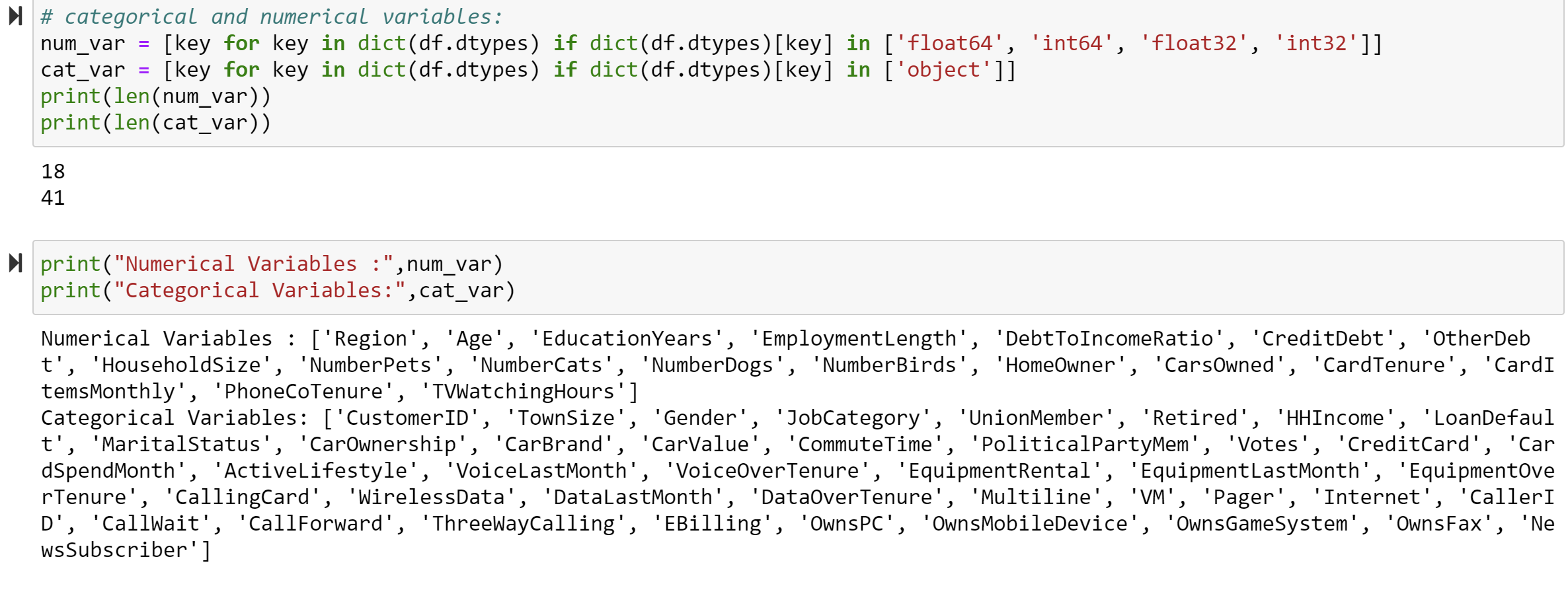
In this step, we are going to perform data imputation to check for missing values and to handle missing values. Also, we would be converting string and object data types to numeric values.



**Feature Engineering:**

To be followed by feature engineering. It is very usual to notice categorical features in a dataset. Nevertheless, the machine learning algorithm can only read numerical values. It is important to encode categorical features into numerical values.

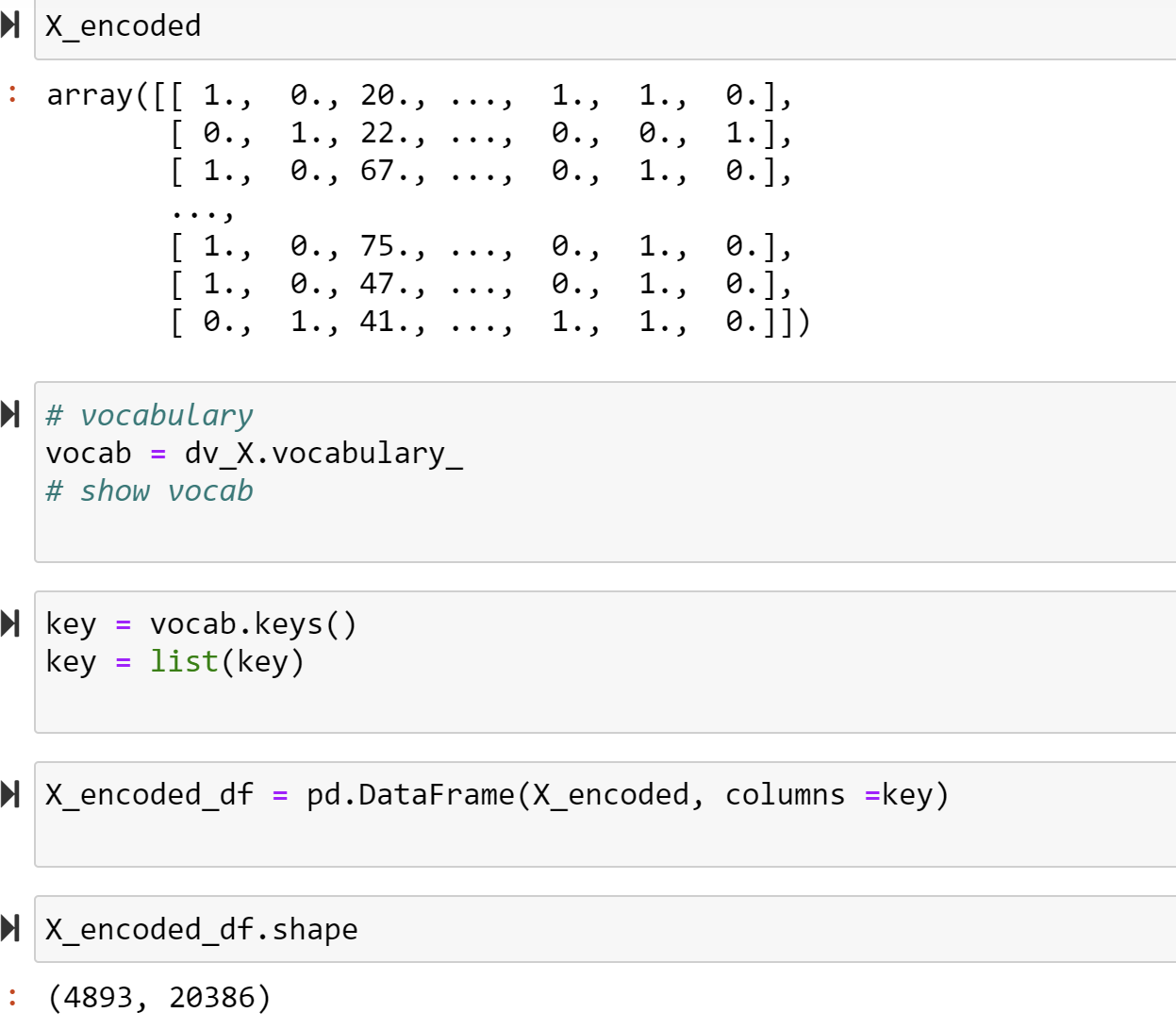
In our dataset, we observe many columns containing categorical variables and numerical variables as shown below:



**One-Hot-Encoding:**

For this situation, a one-hot encoding can be applied to the integer portrayal. This is the place where the integer encoded variable is taken out and another binary variable is added for every unique integer value.



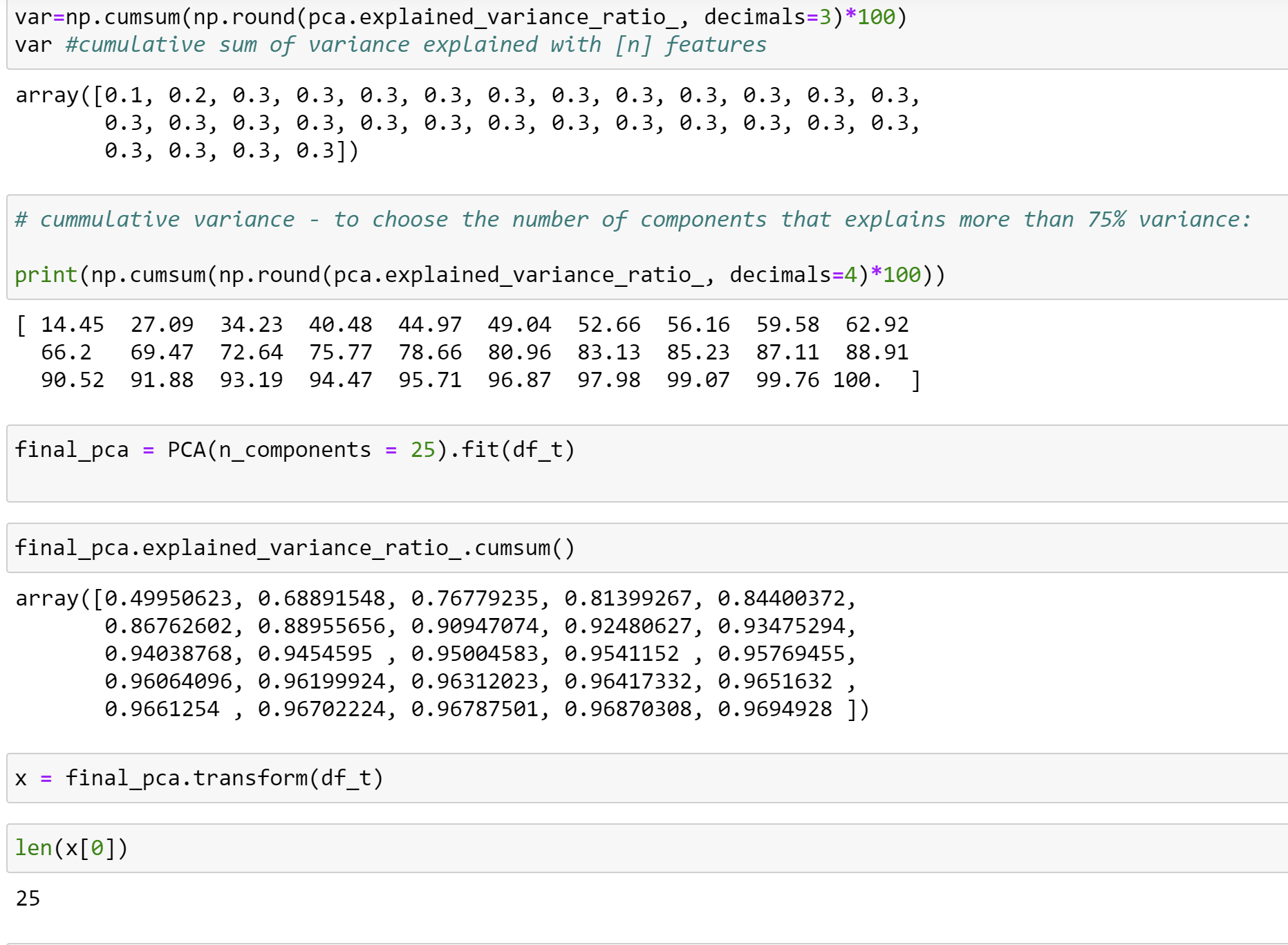




**Principal Component Analysis:**

We are going to do a dimensionality reduction to extract important features in our dataset. For our scenario, n\_components would estimate the number of principal components in the transformed data. Later we would visualize as to how much variance has occurred described utilizing these components By using explained\_variance\_ratio\_ .





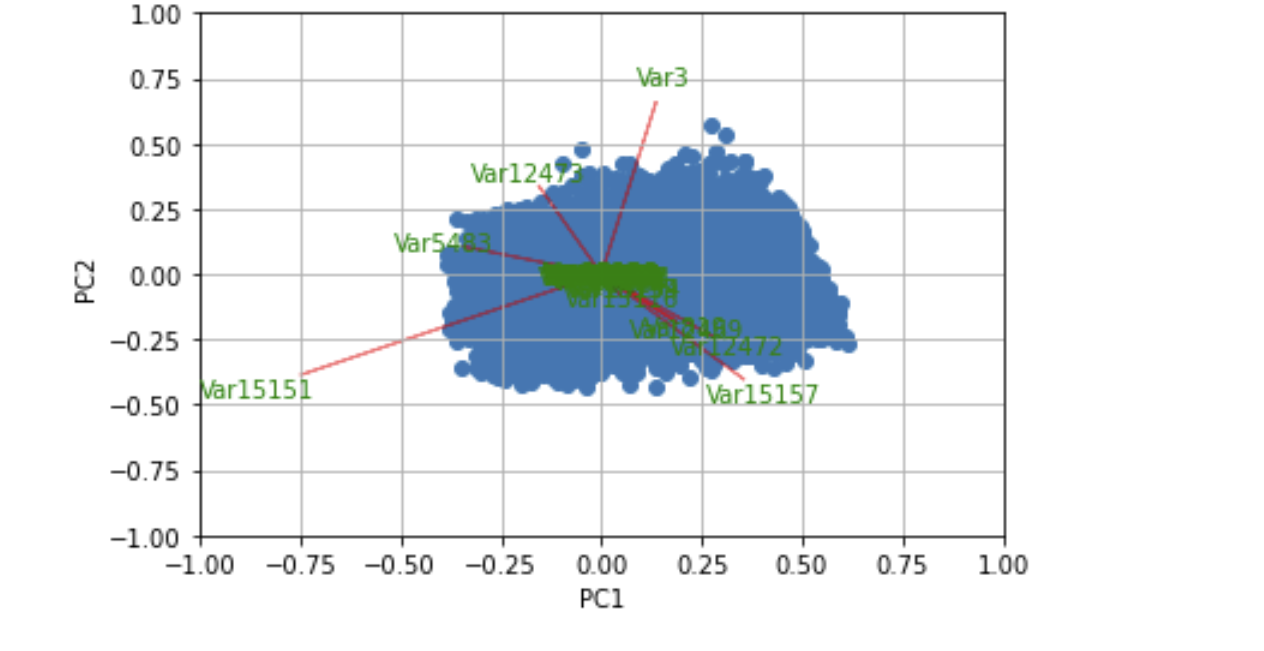
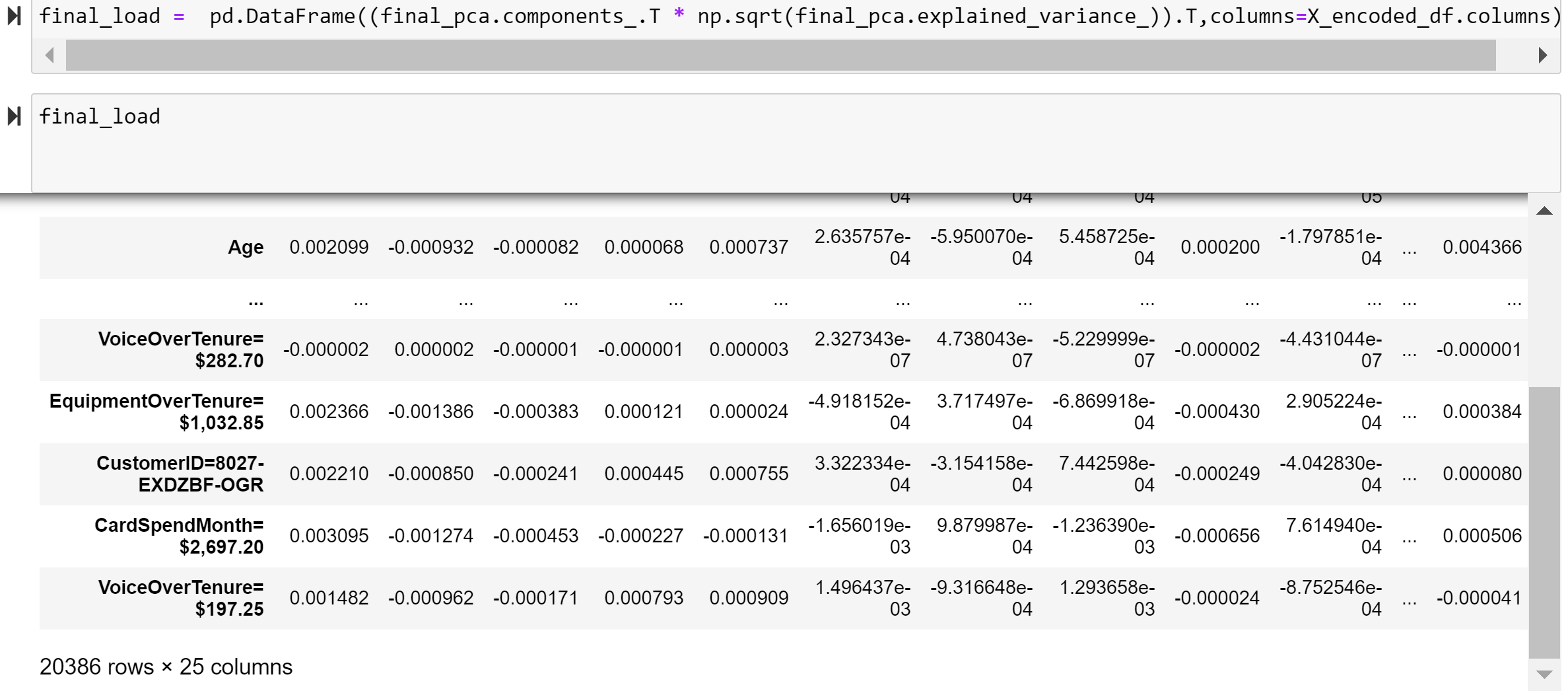


Figure PCA component analysis



**Exploratory Data Analysis:**

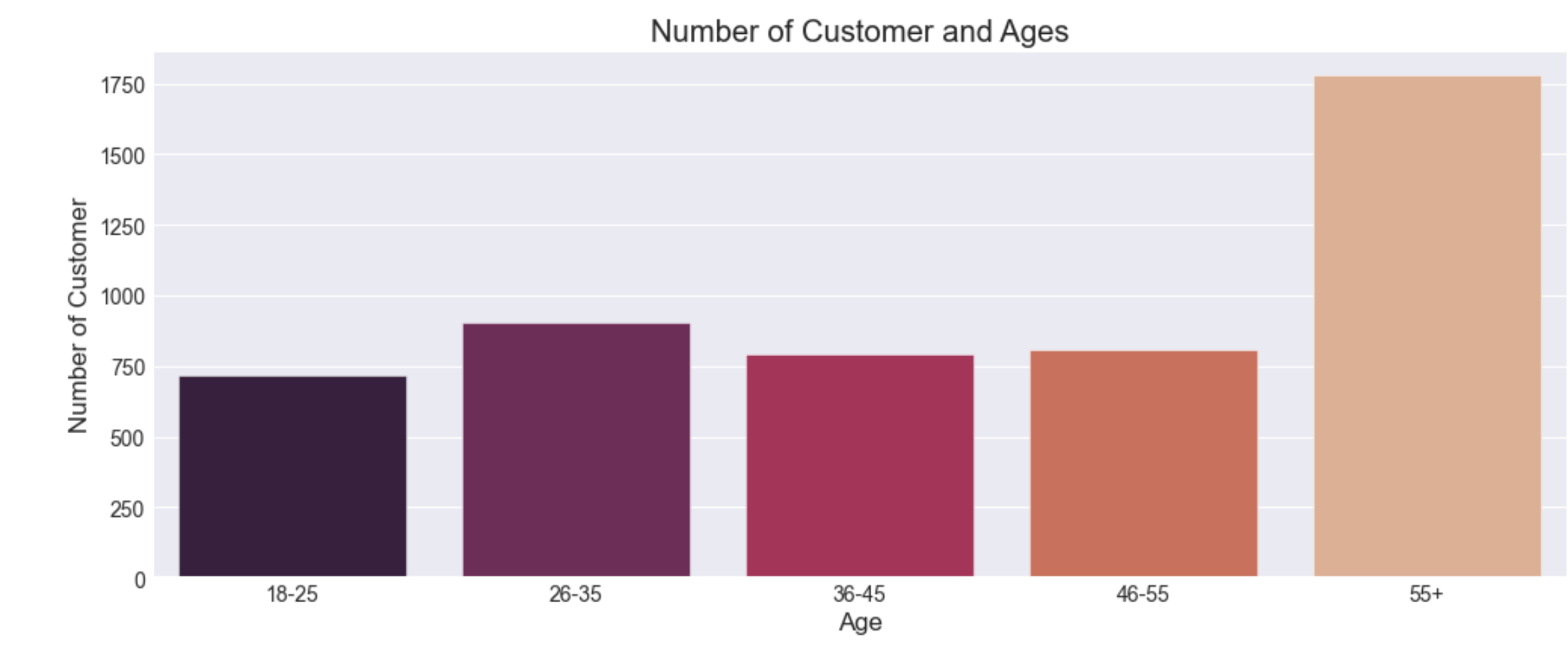
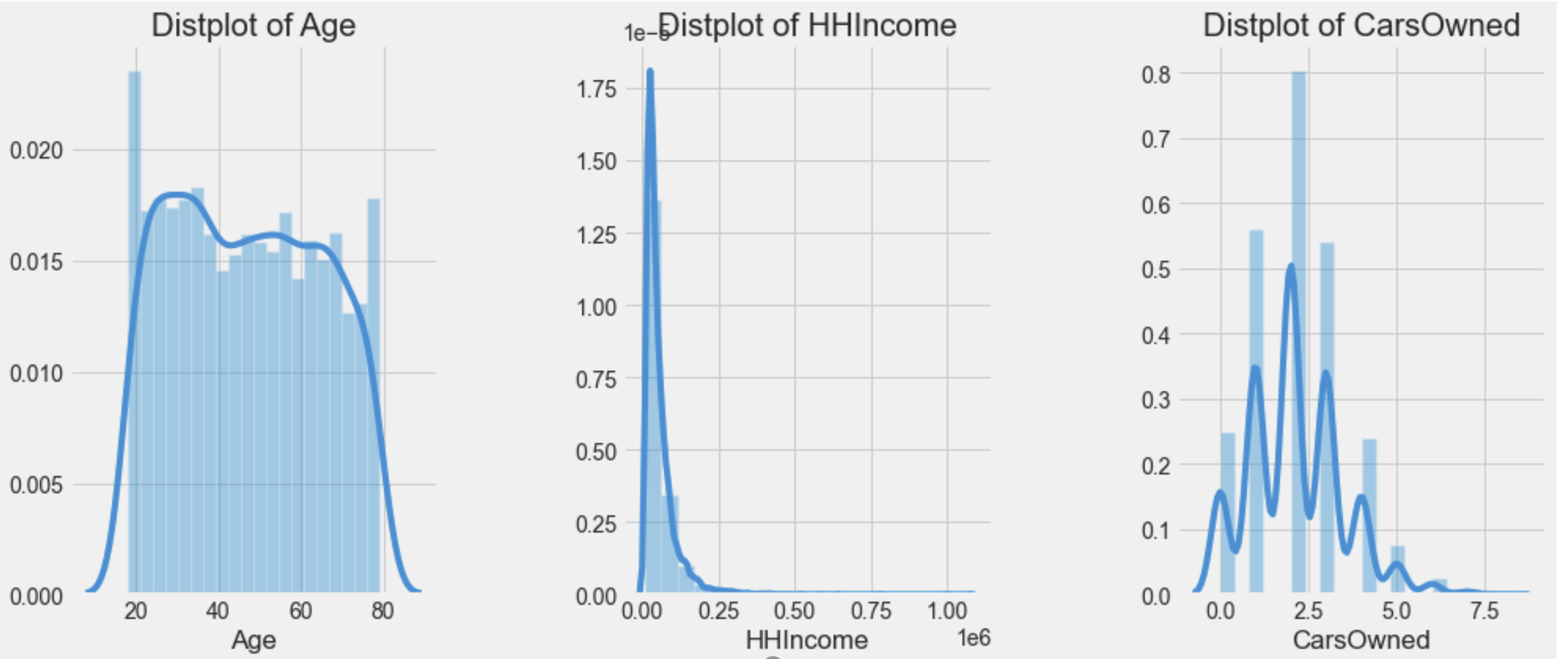


Figure Bar chart Customer vs Age

From this chart, we observe that the number of customers are greater in the Age group above 55 and least between Age group 18-25.



We observe that, as Age and HHIncome are show a gradual increase and decrease as respective to each other and Number of cars owned are increasing and rising at a peak value of 2.5 frequency.

Figure Distribution plot between 3 variables

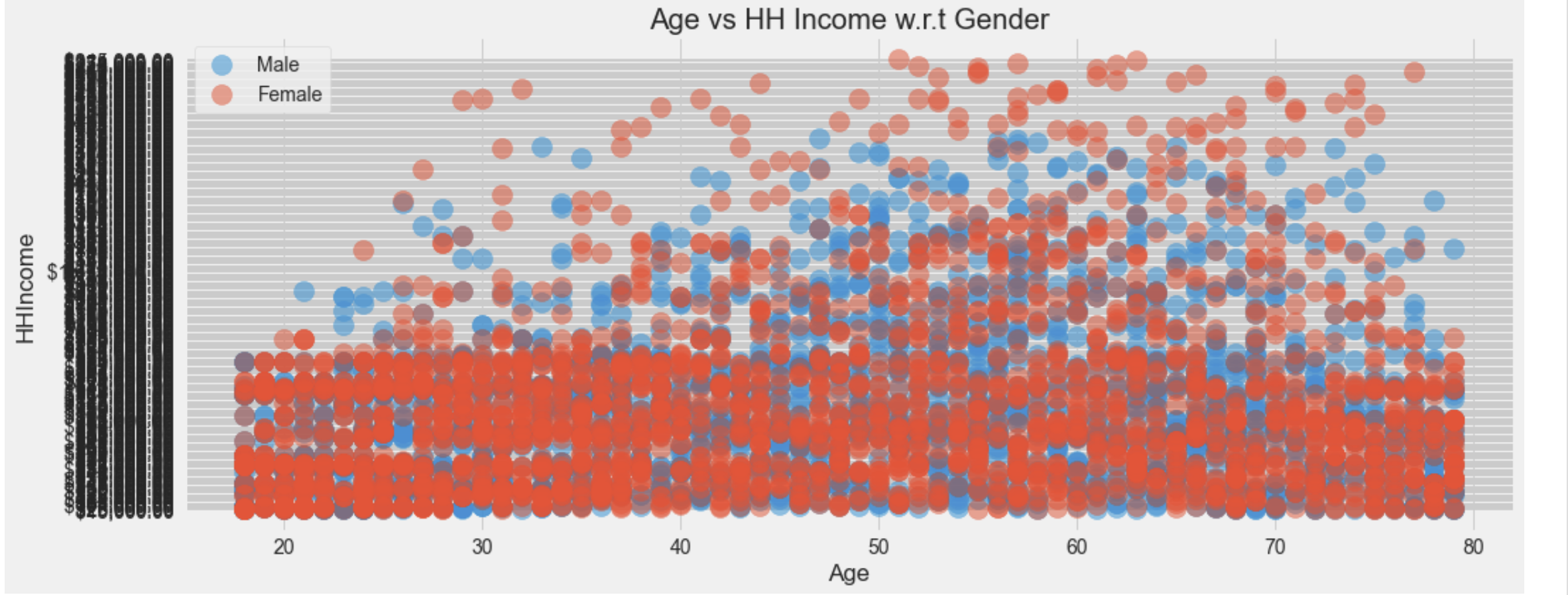


Figure Scatter plot Age vs HH Income

From the scatter plot, we observe that HH Income increases over Age and is maximum observed in Female than in Male.

**Clustering (K-Means):**

In our analysis, we utilized the K-Means clustering method to do our customer segmentation. We have created cluster labels and the by utilizing the Elbow method, we have computed and figured out the optimum number of clusters. Utilizing these number of clusters , we have interpreted the clusters and segmented our features to find the similarities and differences that our cluster segments generated.



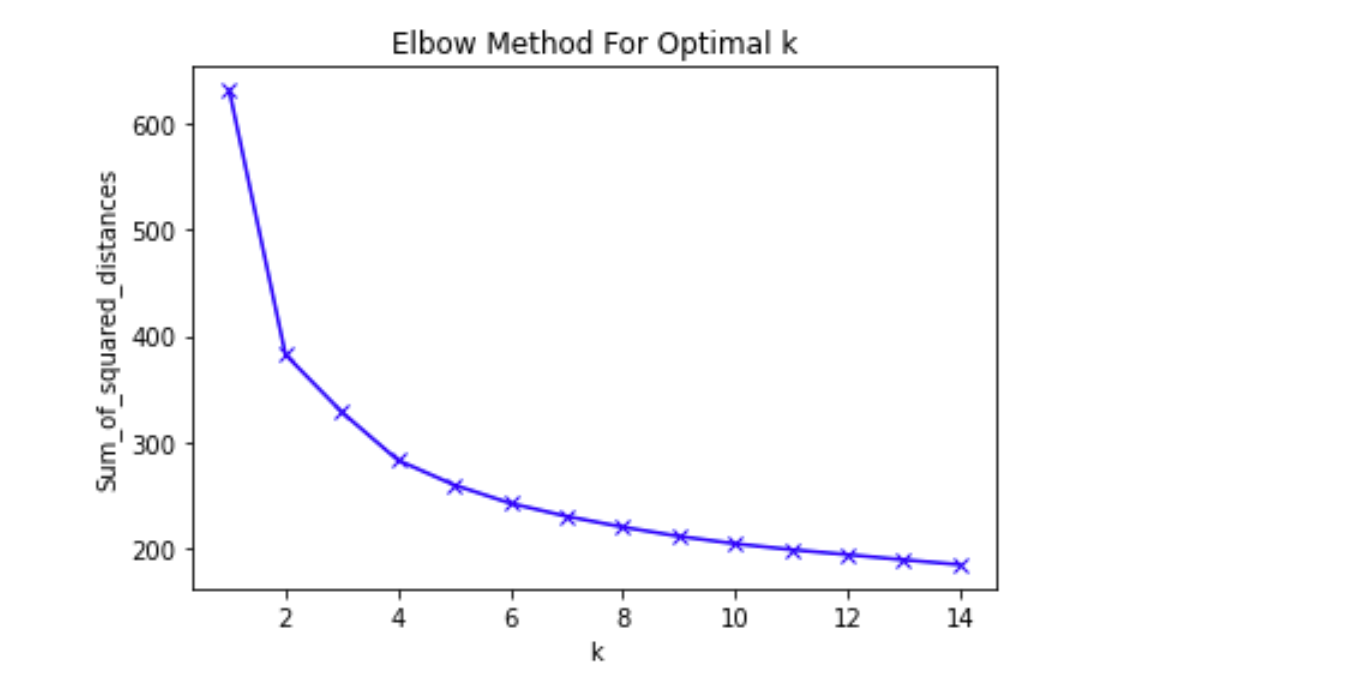


Figure Elbow method K means

From the plot above, we can observe that the elbow bend occurs at k level of 4 and hence we have decided the optimum number of clusters to be 4.

**K=4**

**Customer Segmentation:**

The below segmentation is obtained by clustering using K-Means on 2 PCA components. Each color represents a particular cluster with a total of 4 cluster segments In which the customers are segmented overall.

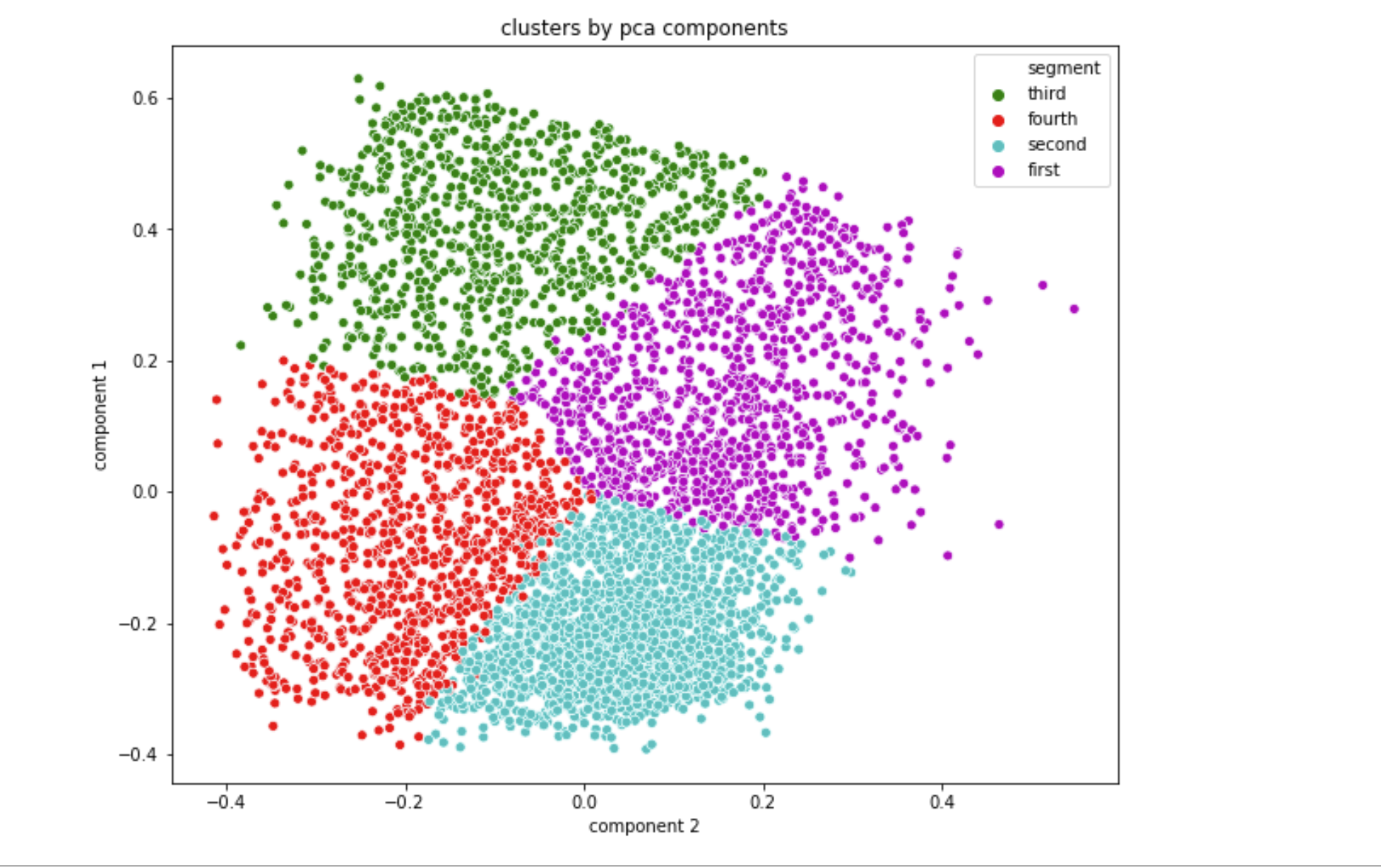


Figure Customer segmentation clusters

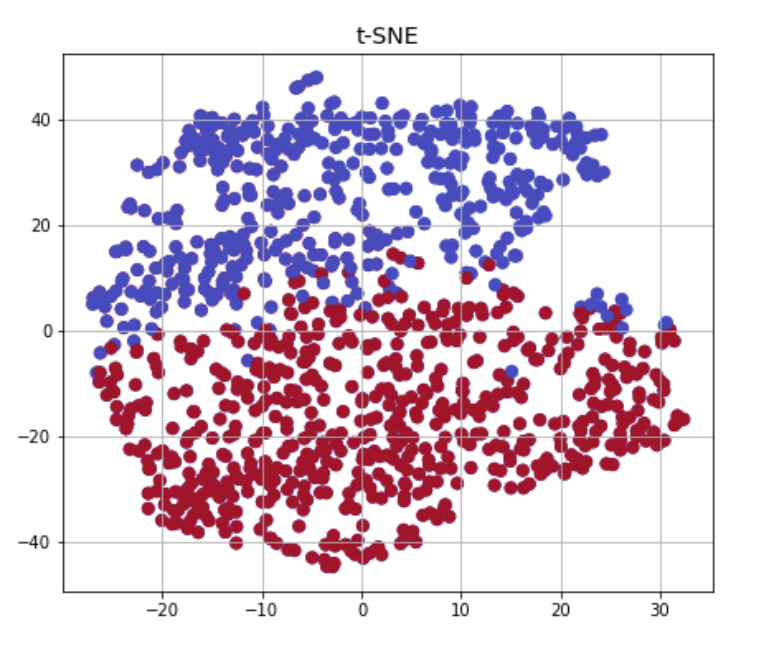


Figure Cluster wise customer segmentation t-SNE

The above figure displays two dimensional visualization of high dimensional dataset. T-SNE is aa technique to display the high dimensional data in two dimensions. From the above plot, we can observe it cearly separates it into two clusters.

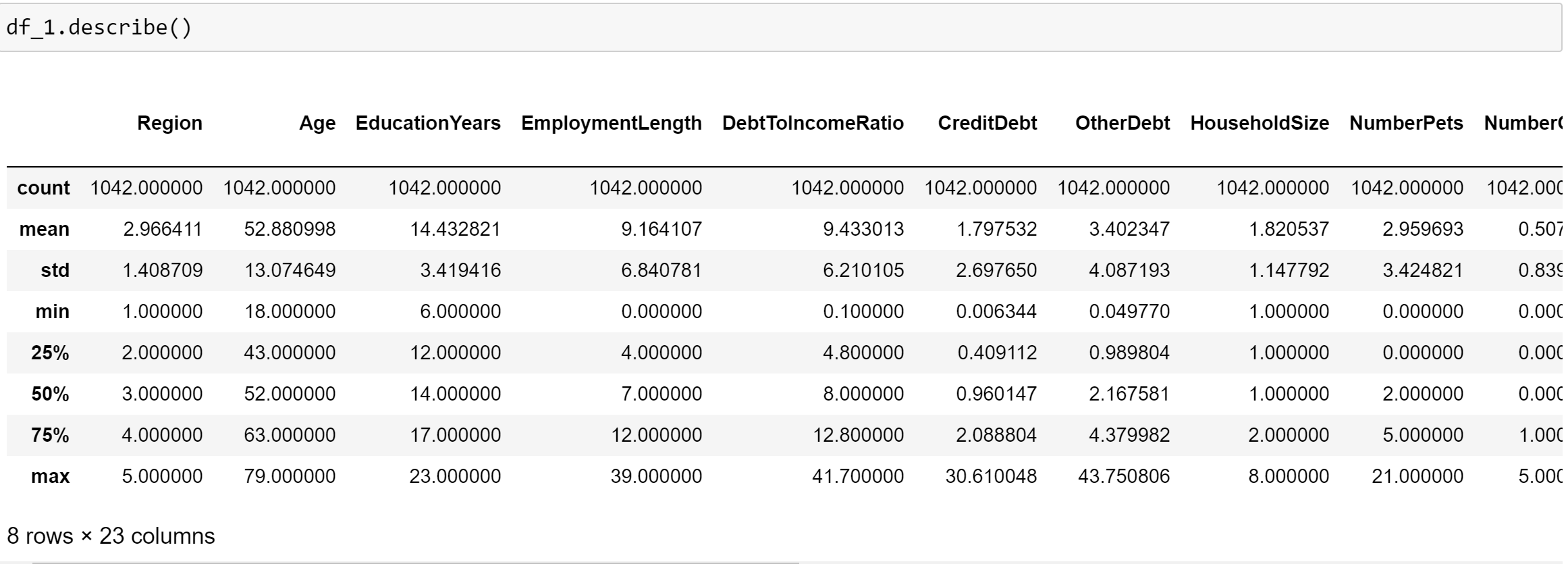
Then we assign these clusters to individual dataframes and get the generated table output as follows:

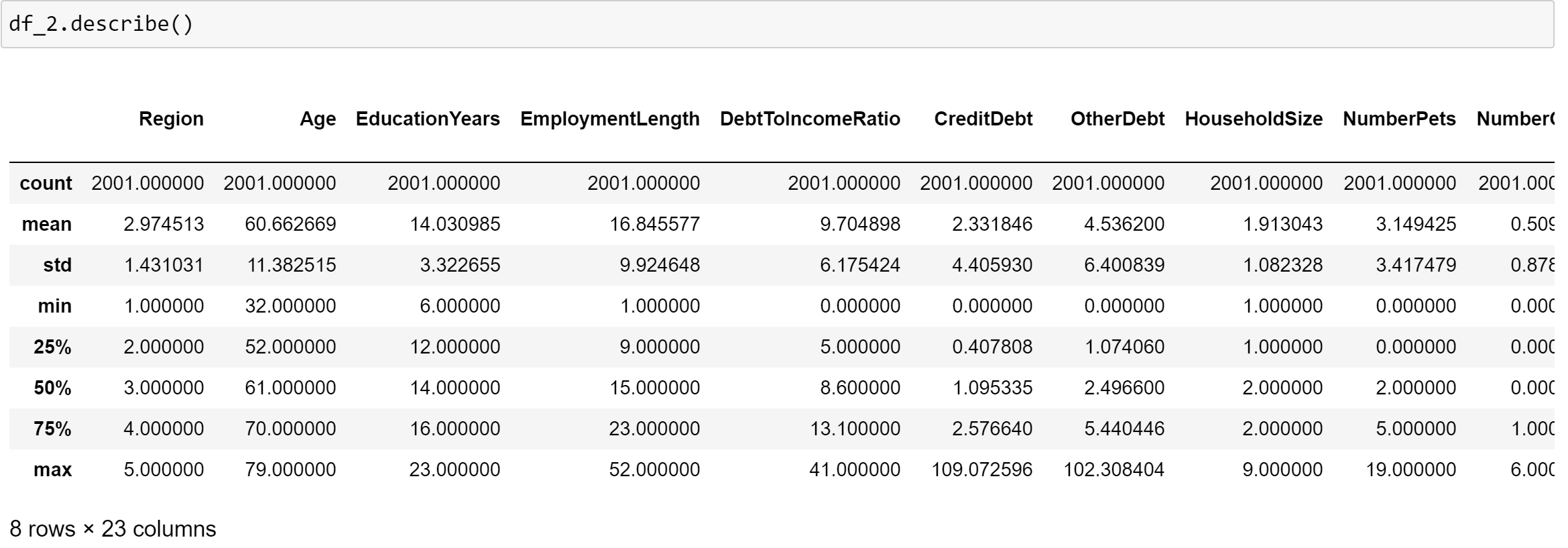
Customer Value Segmentation

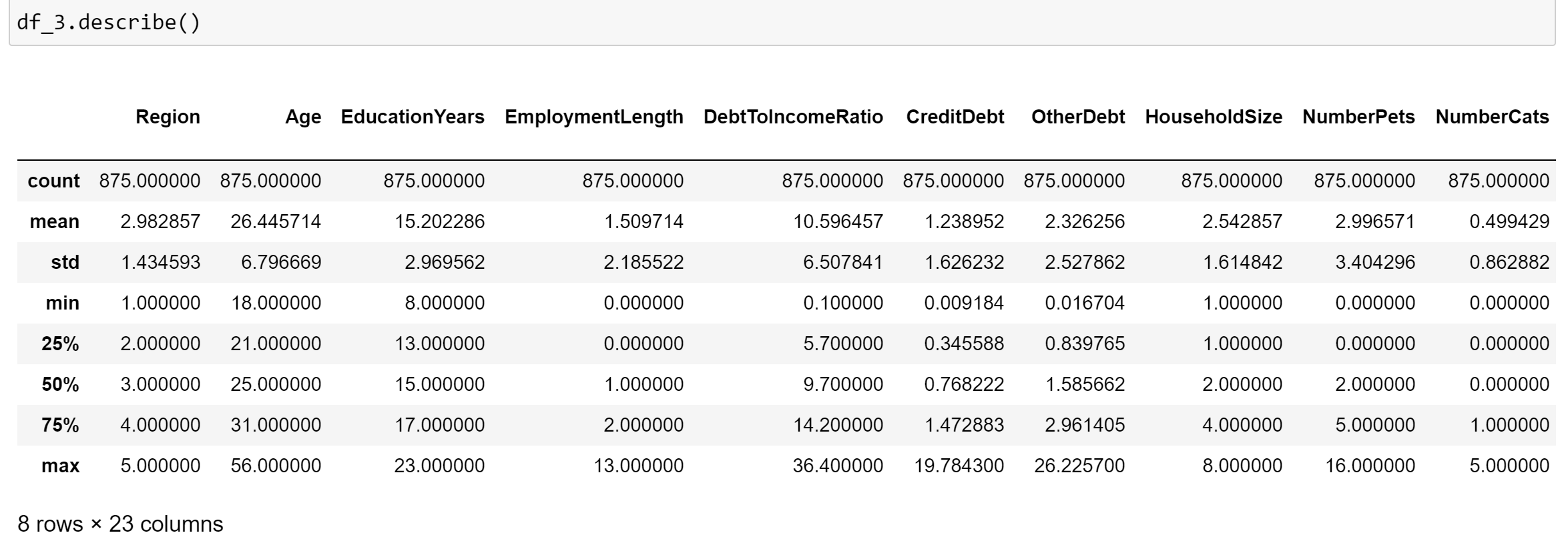
Customer Behaviour Segmentation

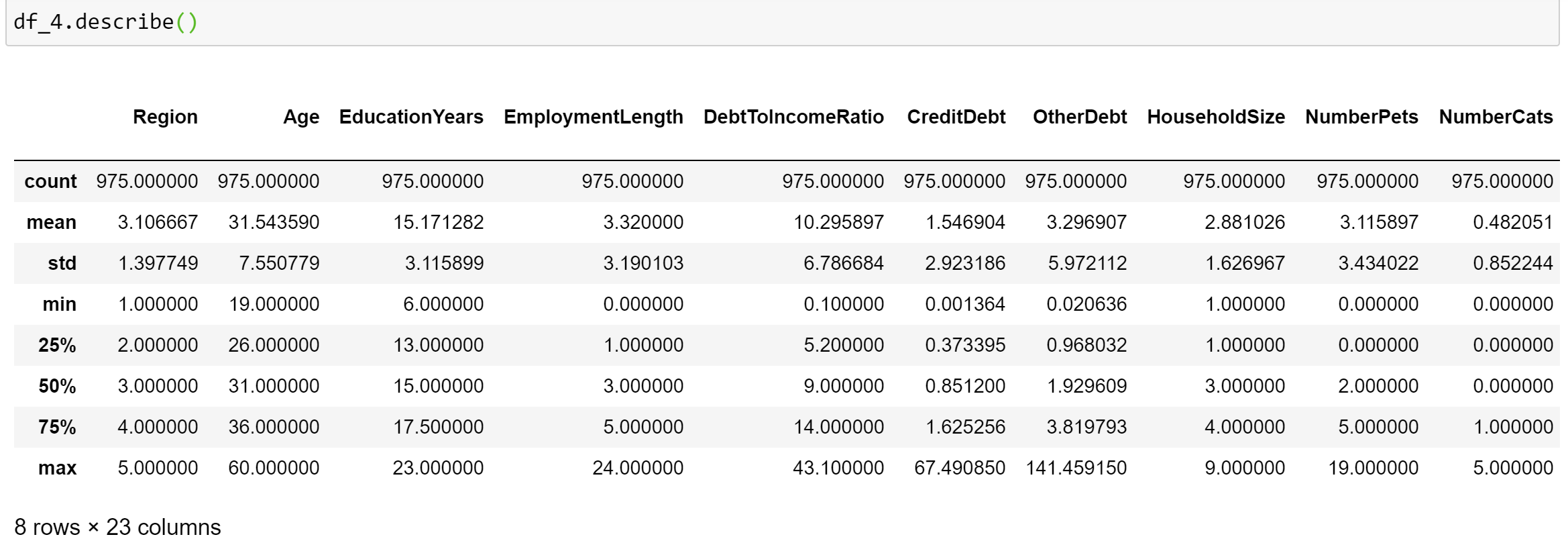
Customer Life style Segmentation

Customer Teleco Segmentation









The above data frames are created for each cluster. The cluster1, cluster 2, cluster 3 and cluster 4 is put into df\_1 , df\_2, df\_3 and df\_4 data frames respectively.

In df1\_1 and df\_2 shows that mean age of 52 and 60. df\_3 and df\_4 has the mean age of 26 and 31 respectively.

We believe that cluster 3 and cluster 4 are high value customers because they will have more phone conversations compared to cluster 1 and cluster who are aged people from 52 to 60.

Current Value Segmentation focuses on identifying the contribution that a customer makes to overall organisational profitability based on current relationships with the organisation.

Lifestyle Value Segmentation identifies the expected (predicted) contribution to overall organisational profitability based on expected ‘lifestyle relationships with the organisation.

In implementing these solutions, organisations need to be clear about their definitions of profit, contribution, revenue and so on. The closer a segmentation scheme moves toward measuring the precise contribution made by a customer, the more useful it may be, from a bottom-line perspective, when it comes to managing the customer base. However, accuracy always involves a trade-off: in this case, the proportion of information derived from factors that marketing can directly influence (spend, retention, price of calls and so on) becomes increasingly diluted by other factors that marketing cannot influence, such as credit-worthiness, competitor prices, and internal cost allocation.