'data.frame': 600 obs. of 46 variables:

$ Member.id : int 1010010 1010020 1014020 1014030 1014190 1017020 1017110 1017160 1017360 1017460 ...

$ SEC : int 4 3 2 4 4 4 4 4 4 1 ...

$ FEH : int 3 2 3 0 1 3 2 3 3 3 ...

$ MT : int 10 10 10 0 10 10 10 10 10 5 ...

$ SEX : int 1 2 2 0 2 2 2 2 2 1 ...

$ AGE : int 4 2 4 4 3 3 4 2 4 4 ...

$ EDU : int 4 4 5 0 4 4 1 4 4 7 ...

$ HS : int 2 4 6 0 4 5 3 5 6 3 ...

$ CHILD : int 4 2 4 5 3 2 2 3 4 4 ...

$ CS : int 1 1 1 0 1 1 1 0 1 1 ...

$ Affluence.Index : int 2 19 23 0 10 13 11 0 17 6 ...

$ No..of.Brands : int 3 5 5 2 3 3 4 3 2 4 ...

$ Brand.Runs : int 17 25 37 4 6 26 17 8 12 13 ...

$ Total.Volume : int 8025 13975 23100 1500 8300 18175 9950 9300 26490 7455 ...

$ No..of..Trans : int 24 40 63 4 13 41 26 25 27 18 ...

$ Value : num 818 1682 1950 114 591 ...

$ Trans...Brand.Runs : num 1.41 1.6 1.7 1 2.17 1.58 1.53 3.13 2.25 1.38 ...

$ Vol.Tran : num 334 349 367 375 638 ...

$ Avg..Price : num 10.19 12.03 8.44 7.6 7.12 ...

$ Pur.Vol.No.Promo.... : Factor w/ 53 levels "0%","100%","33%",..: 2 43 48 2 15 2 52 48 44 2 ...

$ Pur.Vol.Promo.6.. : Factor w/ 42 levels "0%","1%","10%",..: 1 3 13 1 7 1 13 1 3 1 ...

$ Pur.Vol.Other.Promo..: Factor w/ 33 levels "0%","1%","10%",..: 1 14 26 1 19 1 1 30 1 1 ...

$ Br..Cd..57..144 : Factor w/ 87 levels "0%","1%","10%",..: 34 14 25 37 47 77 42 36 35 67 ...

$ Br..Cd..55 : Factor w/ 81 levels "0%","1%","10%",..: 7 65 47 52 8 59 43 64 1 6 ...

$ Br..Cd..272 : Factor w/ 40 levels "0%","1%","10%",..: 1 1 1 1 1 1 2 1 1 1 ...

$ Br..Cd..286 : Factor w/ 43 levels "0%","1%","10%",..: 1 1 23 1 1 1 1 1 1 1 ...

$ Br..Cd..24 : Factor w/ 29 levels "0%","1%","10%",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Br..Cd..481 : Factor w/ 37 levels "0%","1%","10%",..: 1 30 1 1 1 1 1 1 1 1 ...

$ Br..Cd..352 : Factor w/ 42 levels "0%","1%","10%",..: 1 1 1 1 1 1 1 1 1 1 ...

$ Br..Cd..5 : Factor w/ 25 levels "0%","1%","10%",..: 1 7 9 1 1 1 1 1 1 16 ...

$ Others.999 : Factor w/ 433 levels "0.0%","0.7%",..: 191 290 138 1 348 375 193 41 251 153 ...

$ Pr.Cat.1 : Factor w/ 97 levels "0%","1%","10%",..: 18 24 6 1 1 17 66 36 5 59 ...

$ Pr.Cat.2 : Factor w/ 101 levels "0%","1%","10%",..: 54 53 28 37 47 42 65 36 90 3 ...

$ Pr.Cat.3 : Factor w/ 78 levels "0%","1%","10%",..: 7 70 47 51 8 56 41 71 1 6 ...

$ Pr.Cat.4 : Factor w/ 69 levels "0%","1%","10%",..: 51 46 1 1 59 22 18 14 1 11 ...

$ PropCat.5 : Factor w/ 101 levels "0%","1%","10%",..: 48 43 19 37 82 46 83 58 70 19 ...

$ PropCat.6 : Factor w/ 62 levels "0%","1%","10%",..: 1 28 5 1 1 3 1 1 21 39 ...

$ PropCat.7 : Factor w/ 65 levels "0%","1%","10%",..: 1 24 24 1 1 1 14 1 1 9 ...

$ PropCat.8 : Factor w/ 60 levels "0%","1%","10%",..: 1 13 2 1 39 2 2 1 1 1 ...

$ PropCat.9 : Factor w/ 34 levels "0%","1%","10%",..: 1 2 2 1 1 32 1 1 13 1 ...

$ PropCat.10 : Factor w/ 30 levels "0%","1%","10%",..: 1 1 1 1 1 1 1 1 1 1 ...

$ PropCat.11 : Factor w/ 39 levels "0%","1%","10%",..: 1 31 1 1 1 1 1 1 1 1 ...

$ PropCat.12 : Factor w/ 15 levels "0%","1%","12%",..: 7 1 5 1 1 1 1 2 1 1 ...

$ PropCat.13 : Factor w/ 35 levels "0%","1%","10%",..: 1 1 1 1 1 1 1 1 1 1 ...

$ PropCat.14 : Factor w/ 77 levels "0%","1%","10%",..: 7 63 47 51 8 56 41 70 1 6 ...

$ PropCat.15 : Factor w/ 38 levels "0%","1%","10%",..: 22 1 1 1 1 17 3 19 1 19 ...

Step 1: data preparation

1. Variable selection
2. The variables that describe purchase behavior (including brand loyalty)

b. The variables that describe the basis for purchase

c. The variables that describe both purchase behavior and basis of purchase

1. Transform factors into integers
2. From the percentage of purchases made for each brand, pick the highest one and assign it to a new variable named “Max.Br.Cd”.
3. Remove the individual brand purchase percentage variables and the demographic variables.
4. Normalize integer variables
5. Any NA’s? (no)
6. This is a case of unsupervised learning, so no need for validation/test data

Step 2: unsupervised learning

1. Using the elbow method and silhouette method, determine the number of clusters (k=2).
2. Apply k-mean algorithm to the data.
3. Compare the 2 clusters.
   1. Cluster 1 is lower in SEC, higher in EDU, higher affluence index, have purchased more brands, way higher brand runs, purchase volume, number of transactions varies more, average number of transactions and total value of purchases are higher, less average transactions per brand run, lower average volume per transaction

Average price of purchase is higher

|  |  |  |  |
| --- | --- | --- | --- |
|  | Demographic | Purchase behavior | Basis of purchase |
| Cluster 1 (N = 530) | Slightly higher in socioeconomic class (mean of SE= 2.4), with higher education level and higher affluence index | Have purchased more brands with a higher brand run and a lower average transaction/volume per brand run; made more overall transactions with higher volume higher total value, and price of purchase is higher. On average, the maximum percentage of purchase made within one brand is 69%.  **In summary, cluster 1 display less brand loyalty but a higher purchasing power.** | Purchases were made at a higher average price. |
| Cluster 2 (N = 70) | Slightly lower in socio economic class (mean of SE = 3.4), with lower education level and lower affluence index | **On the contrary, cluster 2 seem to have stronger brand loyalty but is weaker in purchasing power.** | Purchases were made at a lower average price. |

1. To develop a model, I define my target audience as customers who have lower brand loyalty and higher purchasing power (Cluster 1). I will use the k-NN algorithm to develop this model.