# MA Group Assignment

Sriganesh Balamurugan – 11915001

Raghu Punnamraju – 11915010

Anmol More – 11915043

Ref :

- https://www.r-bloggers.com/finding-optimal-number-of-clusters/

- https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/

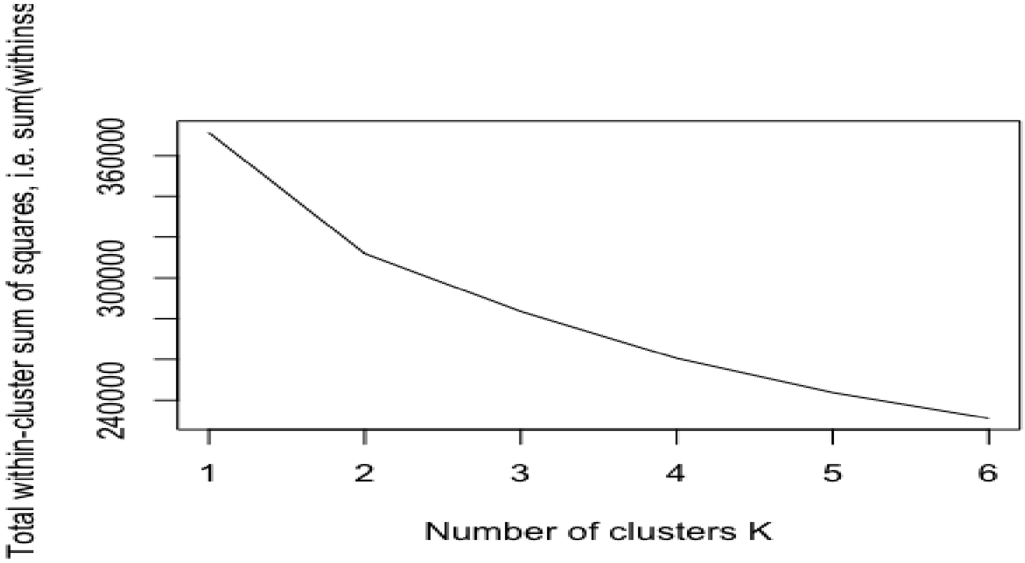
- https://medium.com/codesmart/r-series-k-means-clustering-silhouette-794774b46586

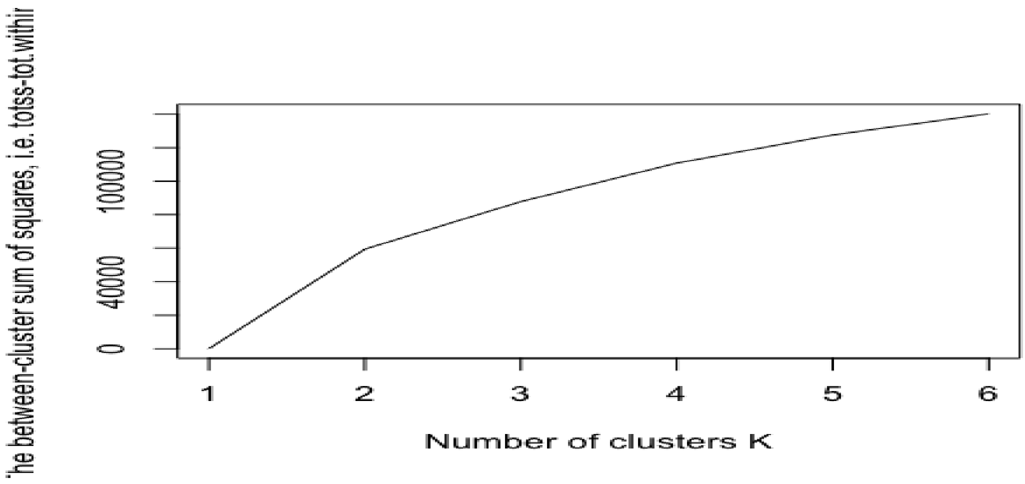
## Segmentation

### 1. Segment respondents based on the Partworth data (use any unsupervised learning technique).

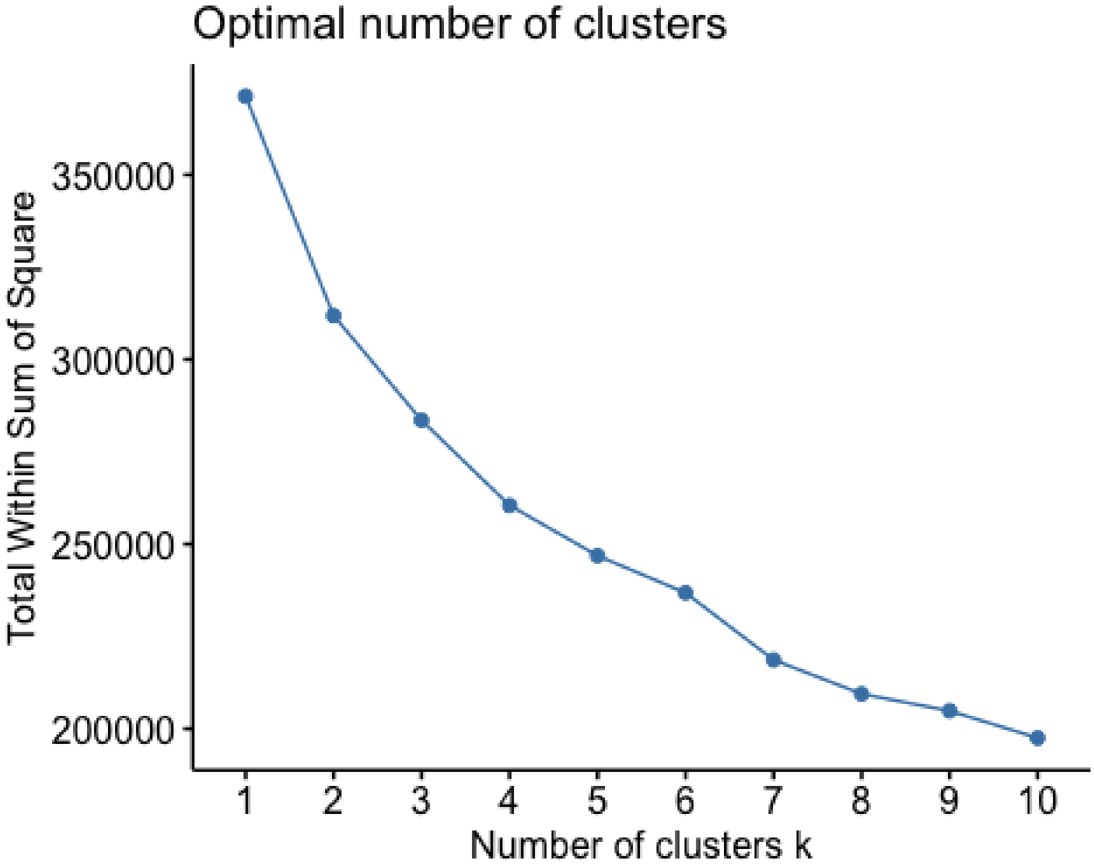
First we apply various clustering algorithms to come up with best K.

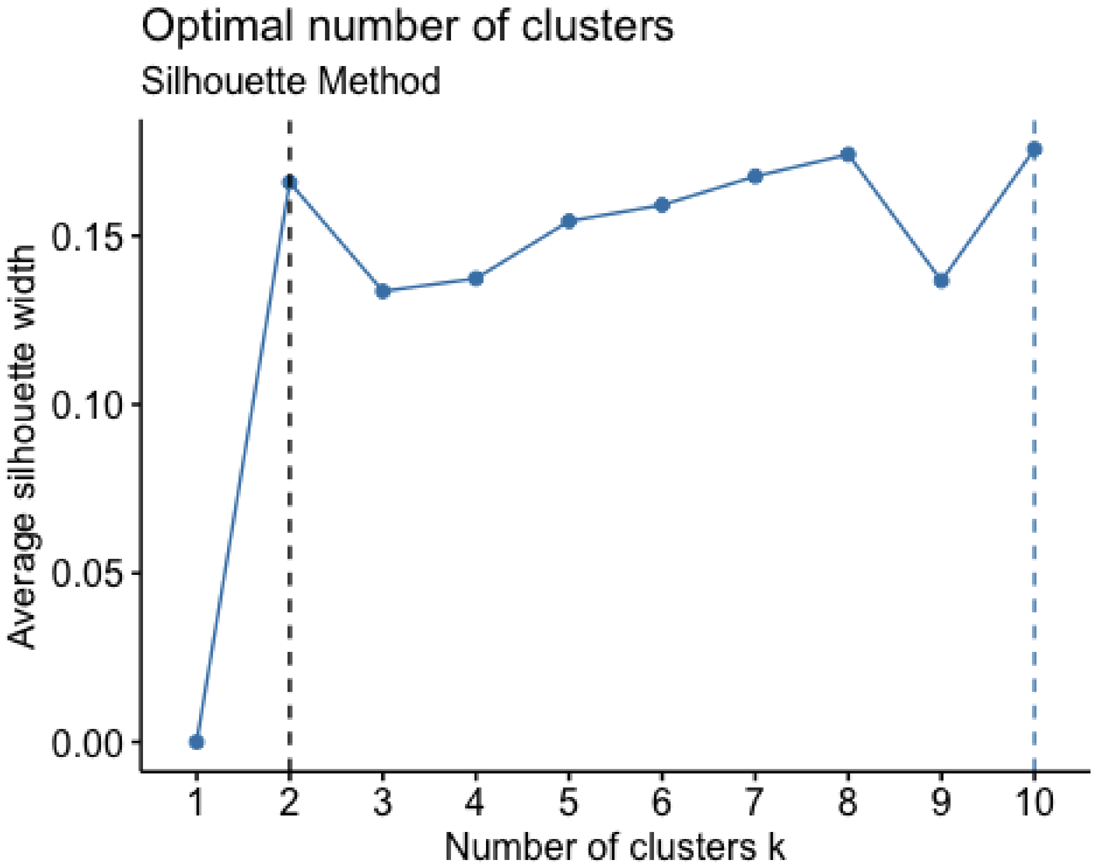
In order to find best K, we looked at within cluster and between cluster Sum of square differences for k.

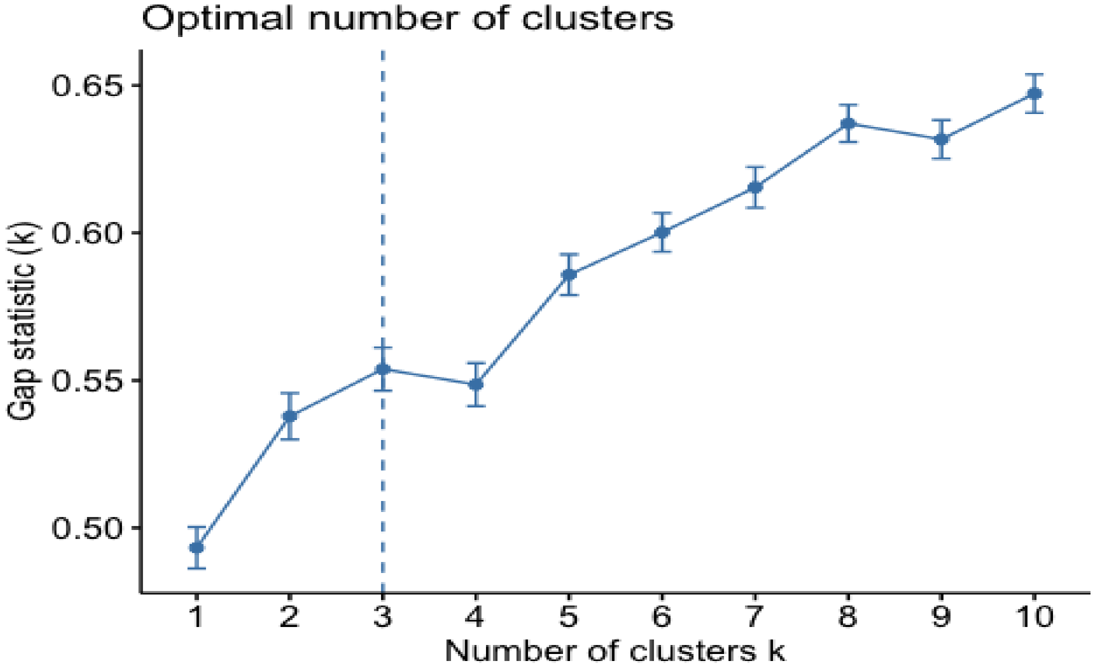




Looking at within cluster and between cluster differences, we can say that data can be divided in either 2 or 3 clusters. For further confirmation, we try various k means methods like Elbow, Silhouette and **find 2 to be most optimal clustering as shown in plot below**




Optimal number of clusters
Silhouette Method
Average silhouette width
0.00
0.05
0.10
0.15
1
2
3
4
5
6
7
8
9
10
Number of clusters k

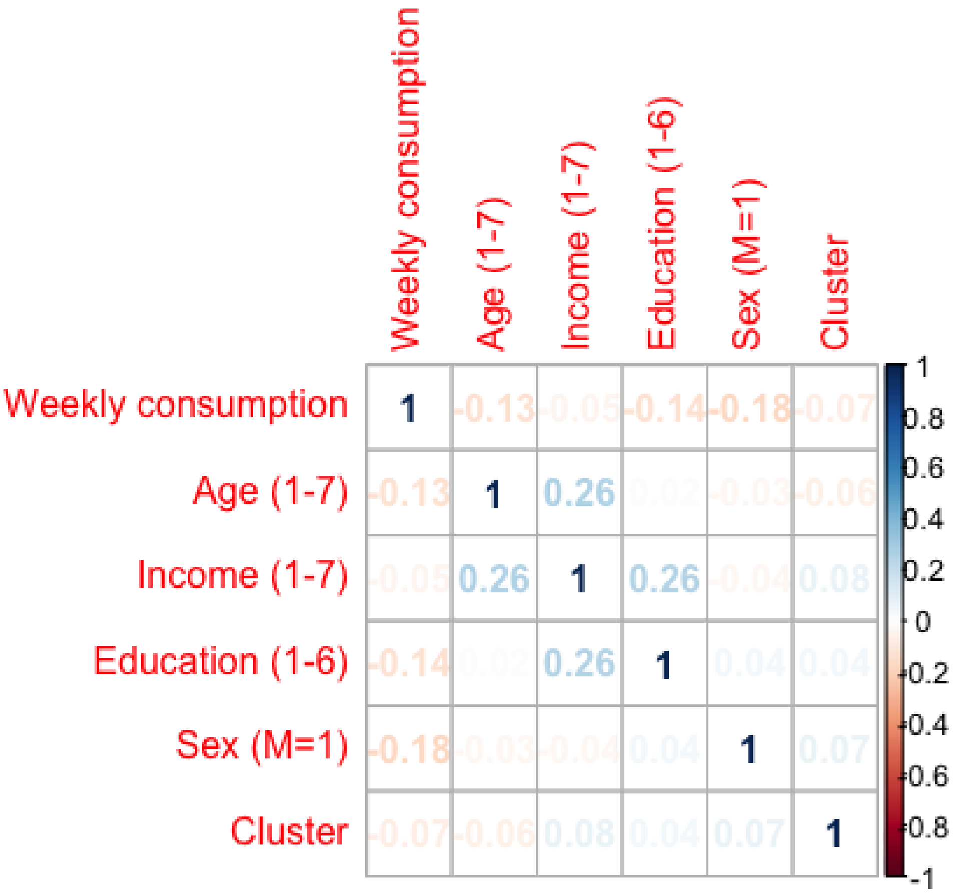

Optimal number of clusters
Gap statistic (k)
0.50
0.55
0.60
0.65
1
2
3
4
5
6
7
8
9
10
Number of clusters k

## 2. Use the Descriptors in the Demographic data sheet to perform classification (use any supervised learning technique) based on segments obtained in Step 1 and personify /describe each segment.

First we read demographics data and apply clusters obtained from conjoint data on Demographics data.

Before moving on to running the supervised learning techniques, we try to understand the distribution of each of the variables.

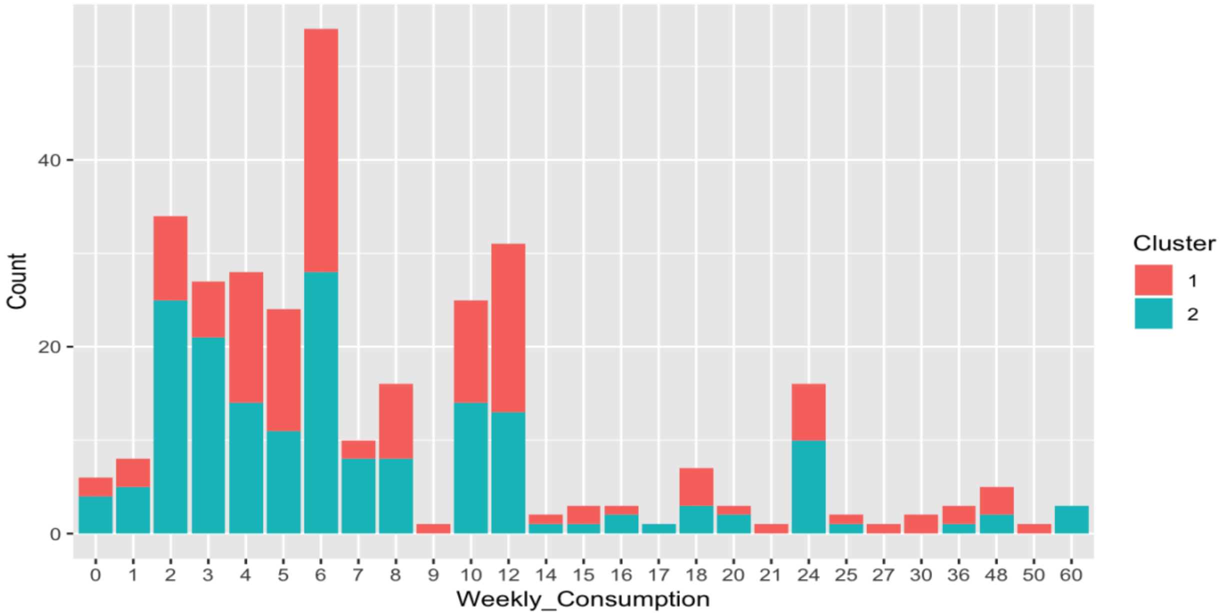
Correlation plot between variables and clusters looks like –

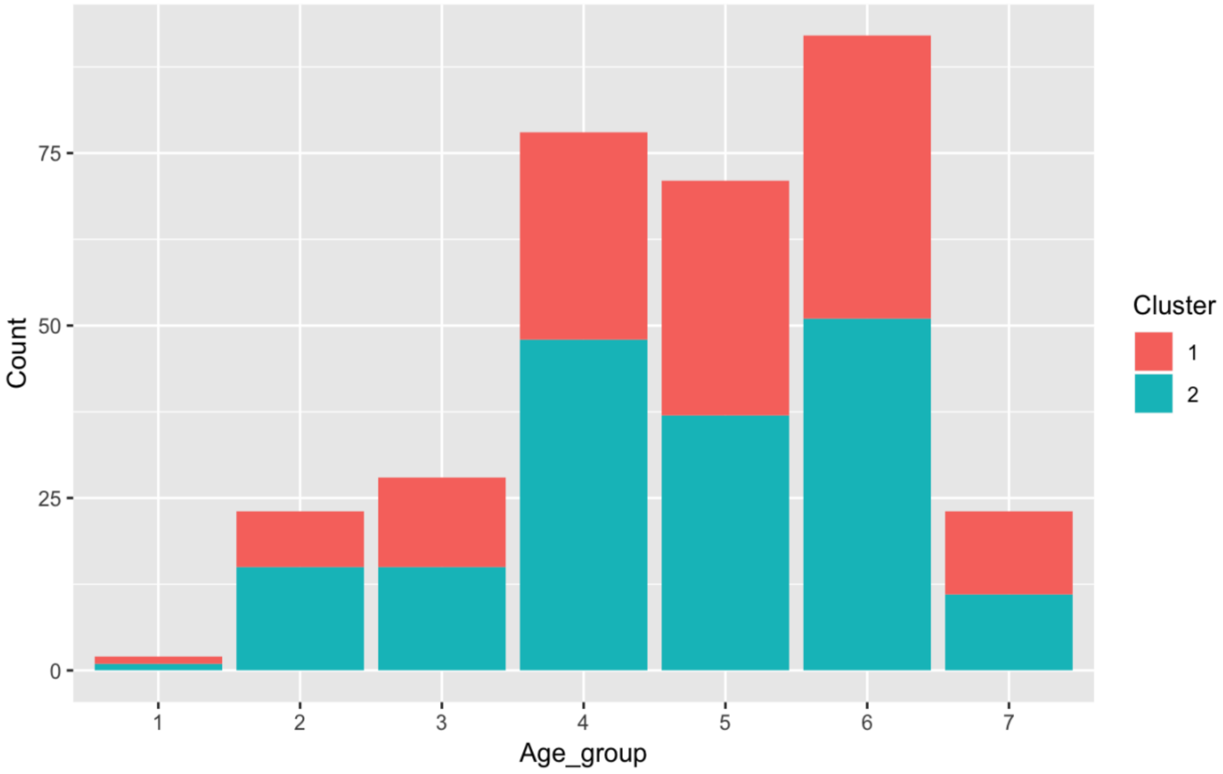


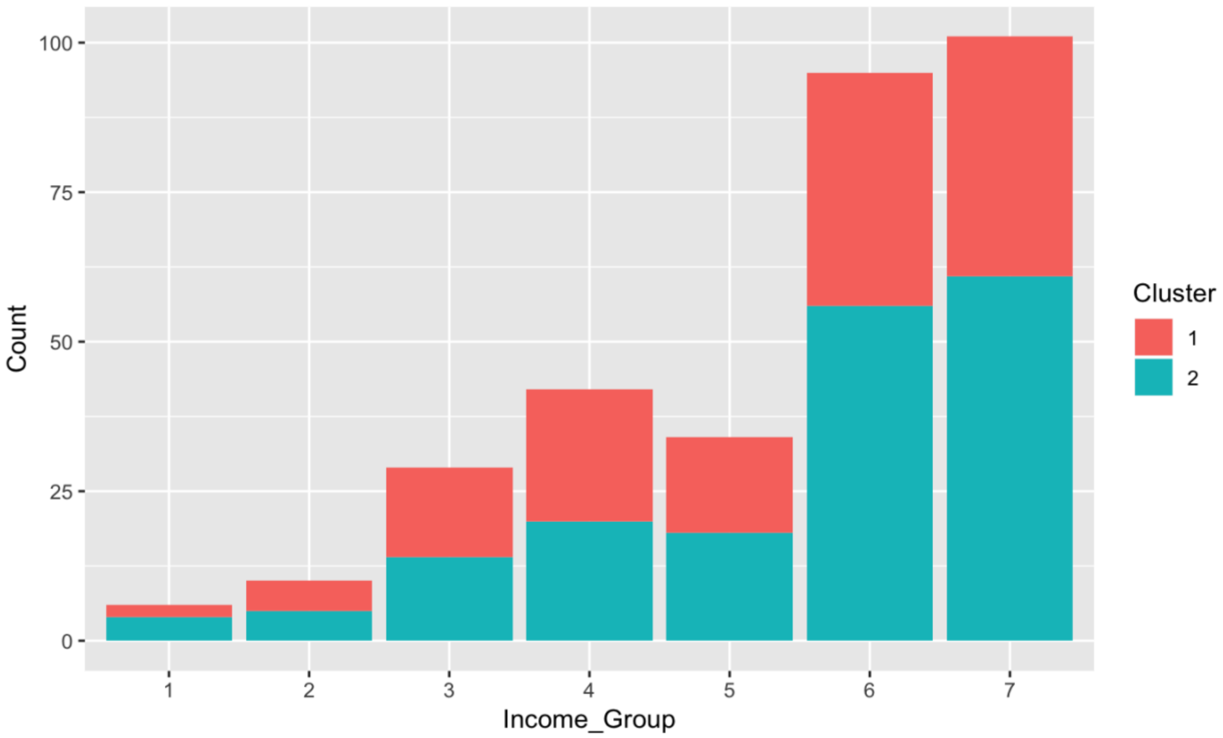
| | Weekly consumption | Age (1-7) | Income (1-7) | Education (1-6) | Sex (M=1) | Cluster |
|---|---|---|---|---|---|---|
| Weekly consumption | 1 | -0.13 | -0.05 | -0.14 | -0.18 | -0.07 |
| Age (1-7) | -0.13 | 1 | 0.26 | 0.02 | -0.03 | -0.06 |
| Income (1-7) | -0.05 | 0.26 | 1 | 0.26 | -0.04 | 0.08 |
| Education (1-6) | -0.14 | 0.02 | 0.26 | 1 | 0.04 | 0.04 |
| Sex (M=1) | -0.18 | -0.03 | -0.04 | 0.04 | 1 | 0.07 |
| Cluster | -0.07 | -0.06 | 0.08 | 0.04 | 0.07 | 1 |

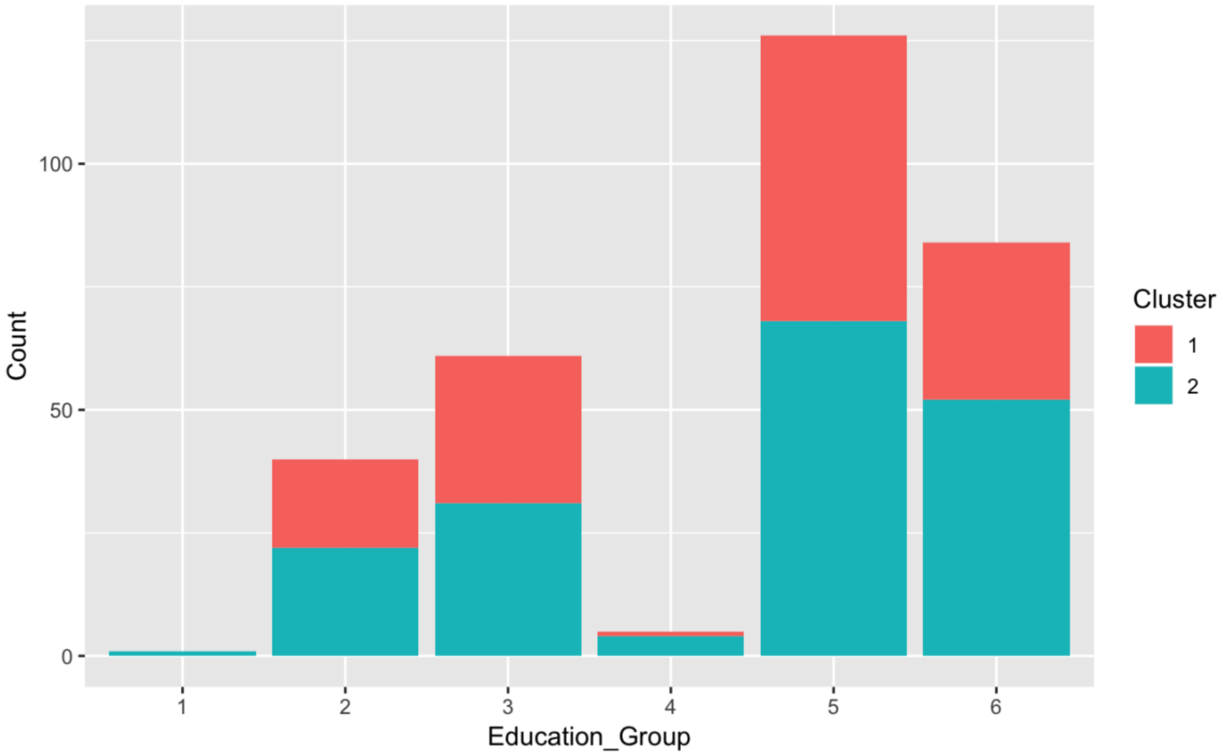
we see from above correlation plot that, our clusters are not correlated to any variables, to some extent age, income and education are weakly correlated

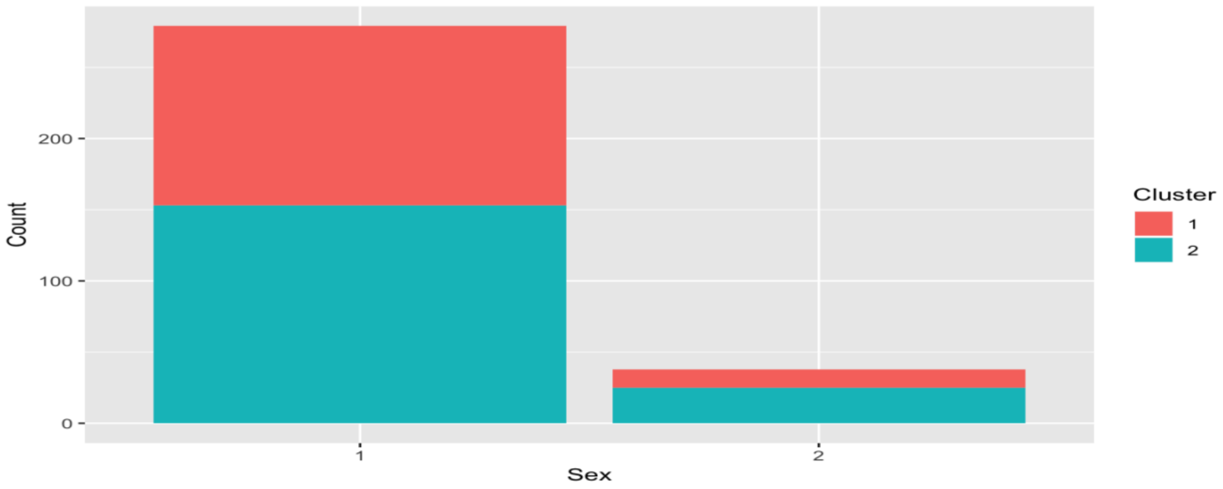
Plot of each variable against no. of clusters defined to identify relationships -


Count
0
20
40
0
1
2
3
4
5
6
7
8
9
10
12
14
15
16
17
18
20
21
24
25
27
30
36
48
50
60
Weekly_Consumption
Cluster
1
2


Count
0
25
50
75
1
2
3
4
5
6
7
Age_group
Cluster
1
2


Count
100
75
50
25
0
1
2
3
4
5
6
7
Income_Group
Cluster
1
2


Count
100
50
0
1
2
3
4
5
6
Education_Group
Cluster
1
2

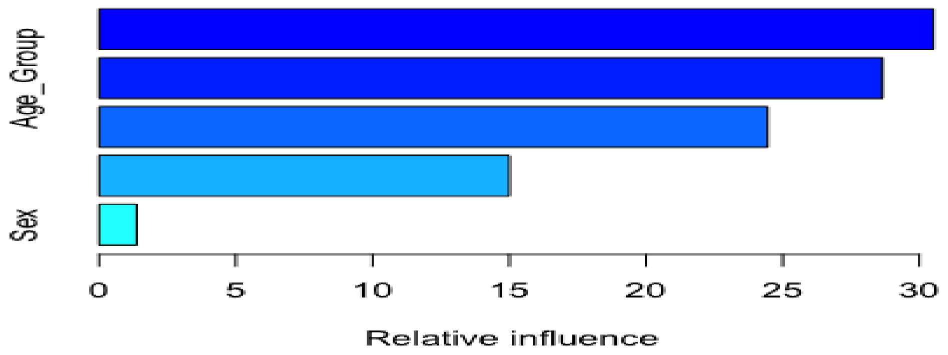


check summary of data to understand distribution of each of the variables

```
## data
##
##  6  Variables      317  Observations
## ---------------------------------------------------------------------------
-
## Weekly_Consumption
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      317        0       27     0.99     9.42    8.941        2        2
##      .25      .50      .75      .90      .95
##        4        6       12       24       25
##
## lowest :  0  1  2  3  4, highest: 30 36 48 50 60
## ---------------------------------------------------------------------------
-
## Age_Group
##        n  missing distinct     Info     Mean      Gmd
##      317        0        7    0.948     4.77    1.523
##
## Value          1     2     3     4     5     6     7
## Frequency      2    23    28    78    71    92    23
## Proportion 0.006 0.073 0.088 0.246 0.224 0.290 0.073
## ---------------------------------------------------------------------------
-
## Income_Group
##        n  missing distinct     Info     Mean      Gmd
##      317        0        7    0.936    5.451    1.686
##
## Value          1     2     3     4     5     6     7
## Frequency      6    10    29    42    34    95   101
## Proportion 0.019 0.032 0.091 0.132 0.107 0.300 0.319
## ---------------------------------------------------------------------------
```

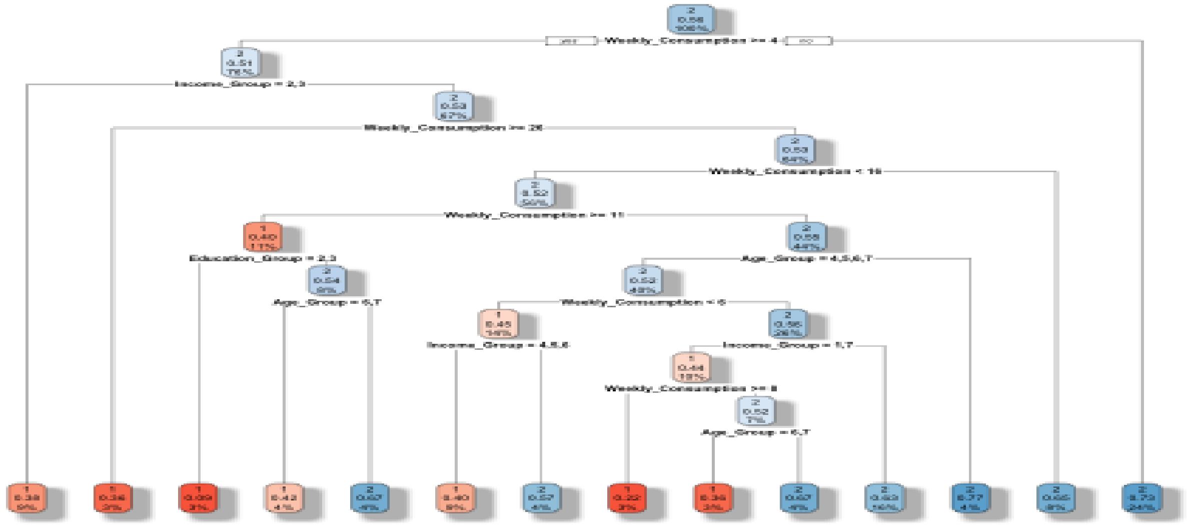
```
-
## Education_Group
##        n  missing distinct     Info     Mean      Gmd
##      317        0        6    0.909    4.473    1.509
##
## Value          1     2     3     4     5     6
## Frequency      1    40    61     5   126    84
## Proportion 0.003 0.126 0.192 0.016 0.397 0.265
## ---------------------------------------------------------------------------
-
## Sex
##        n  missing distinct     Info     Mean      Gmd
##      317        0        2    0.317     1.12   0.2117
##
## Value         1    2
## Frequency   279   38
## Proportion 0.88 0.12
## ---------------------------------------------------------------------------
-
## Cluster
##        n  missing distinct     Info     Mean      Gmd
##      317        0        2    0.739    1.562    0.494
##
## Value          1     2
## Frequency    139   178
## Proportion 0.438 0.562
## ---------------------------------------------------------------------------

##                 Length Class  Mode
## call               3   -none- call
## type               1   -none- character
## predicted        317   factor numeric
## err.rate        1500   -none- numeric
## confusion          6   -none- numeric
## votes            634   matrix numeric
## oob.times        317   -none- numeric
## classes            2   -none- character
## importance         5   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            14   -none- list
## y                317   factor numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

**We ran regression based algorithms, as well as Decision Tree (as seen in R Code) and Random Forest. Since Random forest gives us clear picture of our segments we further analyse and classify our data based on trees produced from random forest**

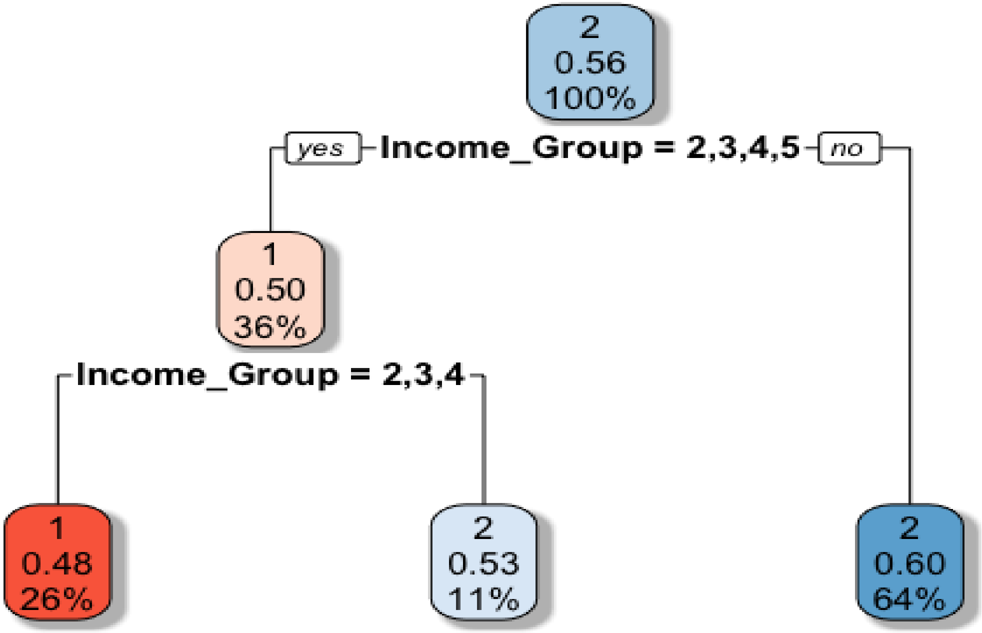


```
##                                   var   rel.inf
## Income_Group             Income_Group 30.515456
## Age_Group                   Age_Group 28.651370
## Weekly_Consumption Weekly_Consumption 24.454245
## Education_Group       Education_Group 15.001676
## Sex                               Sex  1.377253
```

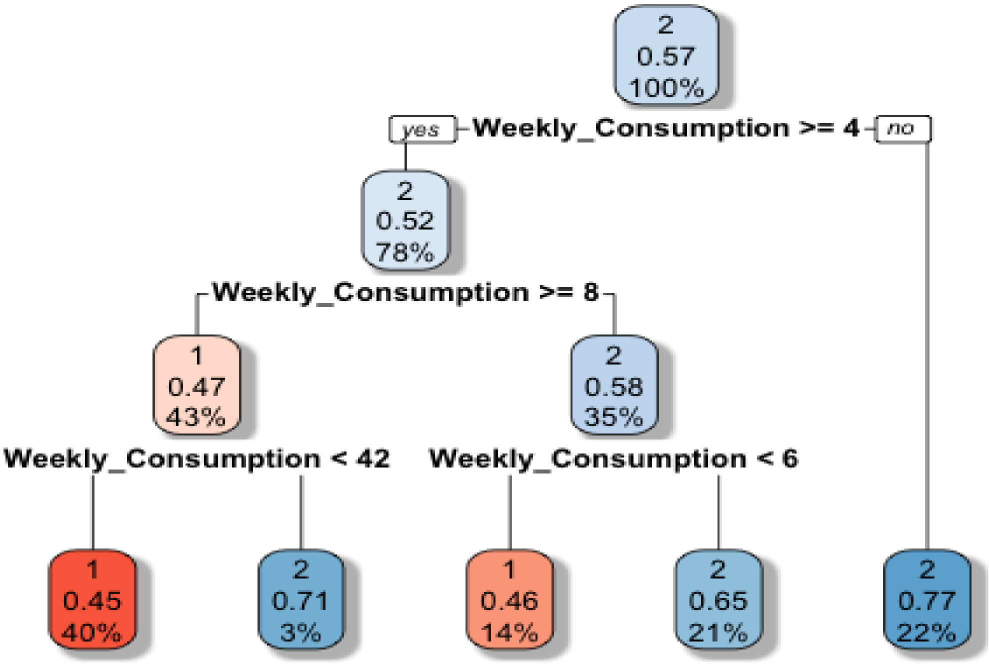
**As we can see from plot above Income, Age and Weekly Consumption classifies the clusters best. We plotted various different trees to personify our clusters.**



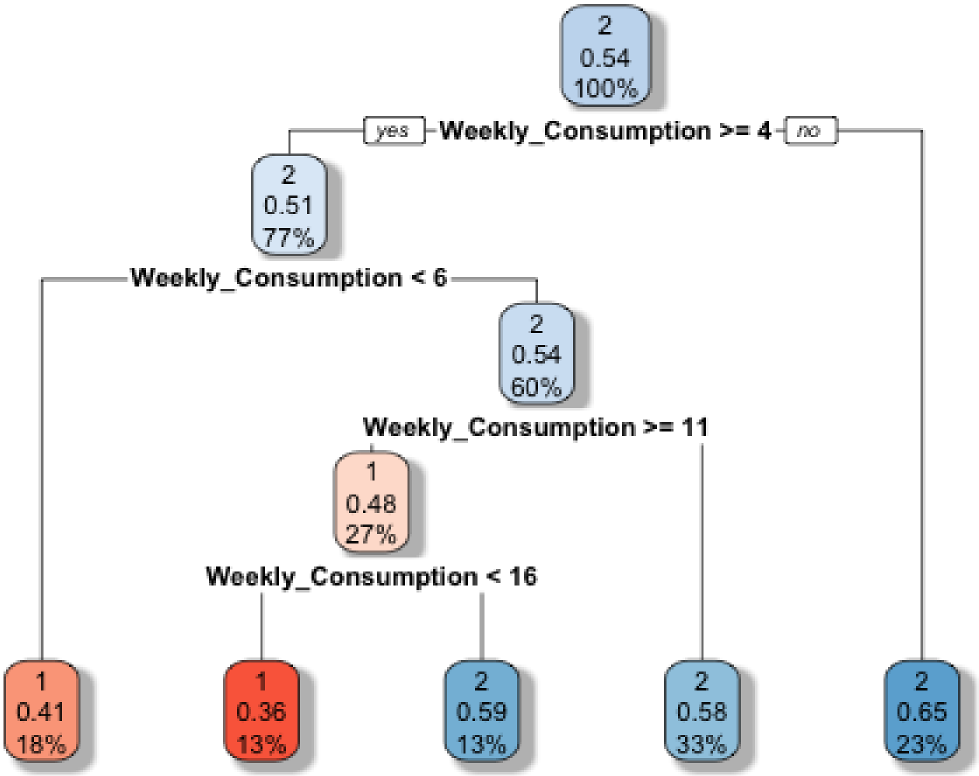
**It is evident in below plot that high income group people (75%) fall in cluster 2**



**Digging further, we can see that, younger population within age group of below 40 have higher weekly consumption than average consumption of 9.4. Around 83% are falling in falling in segment 1**



Some further analysis on higher age group 40+ age, shows 50%+ fall in cluster 2



Also in histogram for cluster vs demographics data, we see similar pattern. We see that cluster 2 has lower weekly consumption and fall in higher age group and income group

**Finally we can divide the whole data in two groups –**

1. **Young and heavy drinkers** - These are people in age group of less than 40, and have average weekly consumption of 10.24 bottles/cans

2. **Old and occasional drinkers** - These are older people in age group of above 40, and have average weekly consumption of 8.8. Also, we observed that 75% of people in this group earn more than $50k.

So, we can say that older and higher income group people are occasional drinkers and prefer costly brands over younger lot who are heavy drinkers are prefer cheaper priced brands.

## Targeting

### 3. Calculate the market Shares of existing brands in each segment.

Calculations in excel sheet

In order to calculate the market share, we followed following steps –

i) Create a matrix of properties of brands and their attribute profiles

ii) Calculated utility of each product and each respondent

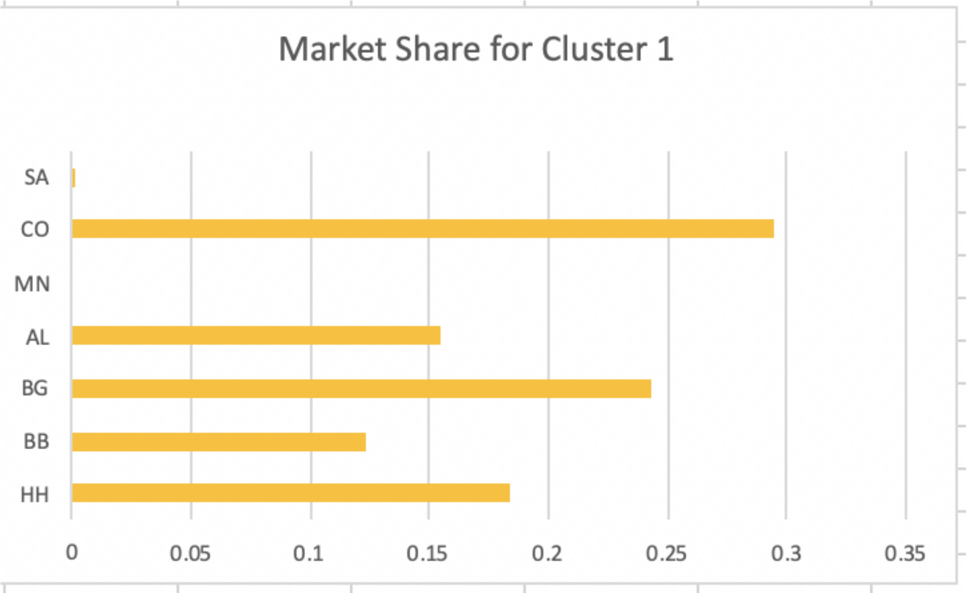
iii) Calculate the choice probability of each respondent for each product

$\exp(\mathrm{Ui})/\sum_{j=1}^{n} \exp\ (Uj)$

iv) Sum up the choice probabilities to get market share

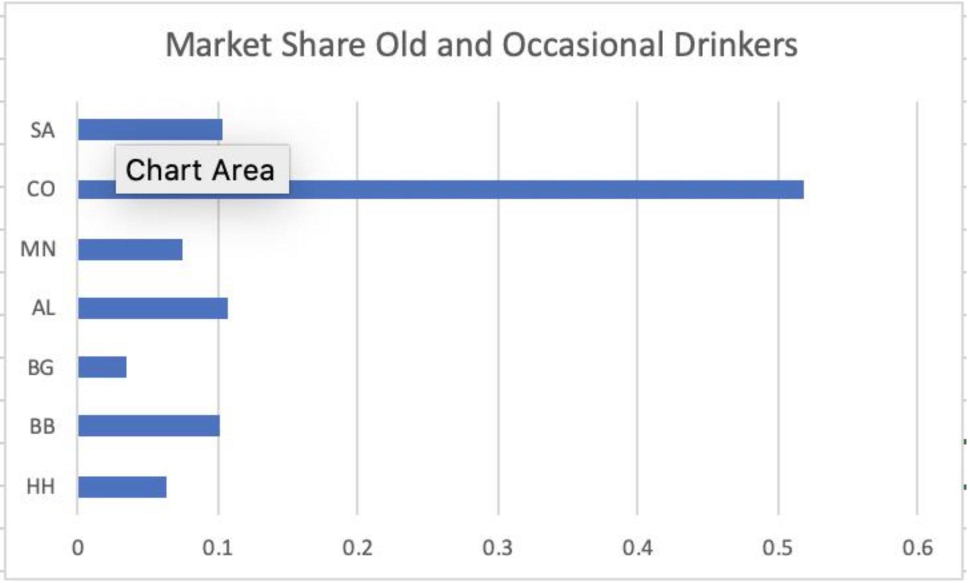
Market Share of cluster 1 (**Young and heavy drinkers**) –

| HH | BB | BG | AL | MN | CO | SA |
|---|---|---|---|---|---|---|
| 0.184 | 0.123 | 0.243 | 0.155 | 0 | 0.295 | 0.001 |



Market share of cluster 2 (**Old and occasional drinkers**) –

| HH | BB | BG | AL | MN | CO | SA |
|---|---|---|---|---|---|---|
| 0.063 | 0.101 | 0.034 | 0.107 | 0.074 | 0.518 | 0.103 |



## 4. If a new brand NB is to be introduced in the market, which segment (question 1&2) will you suggest NB should target, based on your analysis (make appropriate assumptions if required).

Target for NB segment should be cluster 2 (Old and Occasional drinkers). Reasons being –

i) It can command a greater market share in its own cluster 13% compared to only 3% targeted customers in cluster 1 (Young and Mature)

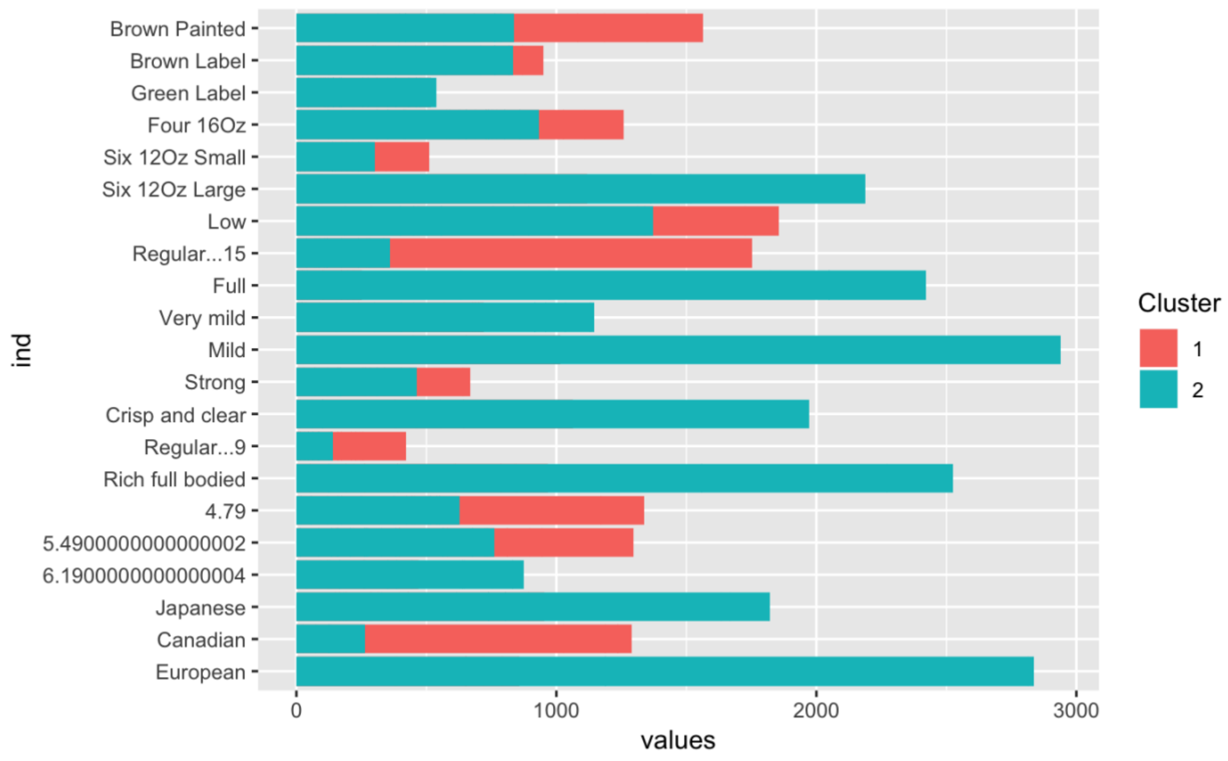
ii) It can capture the market for BB and AL which are coming and small packaging and preferred by old and occasional drinkers

iii) Coming with small packaging and still priced at 5.49, NB is comparatively costlier than market leader CO. So, target young and heavy drinkers won't fetch it much market, as they are price sensitive group with lower income < $50K

Other reasons to target cluster 2 (Old and Occasional drinkers) for new brand is –

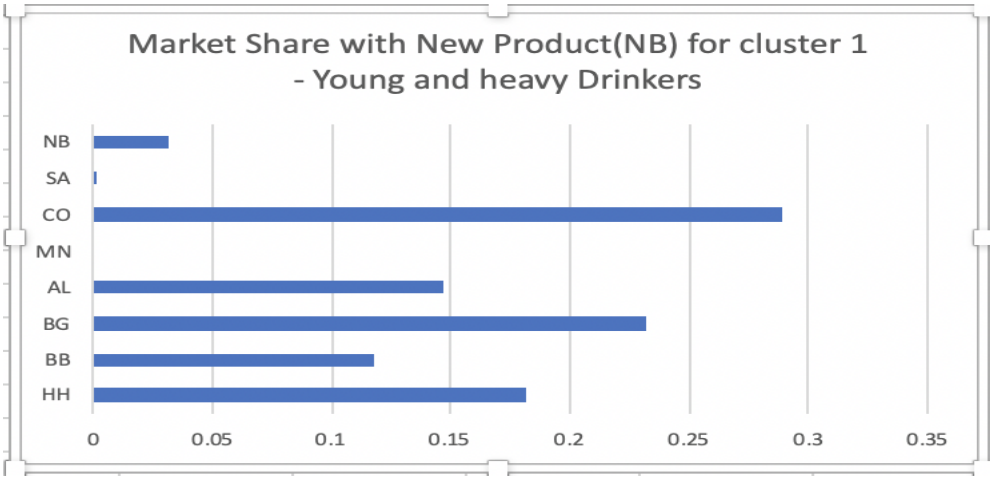
People in cluster 2 prefer - Japanese & European. Brown label, mild and very mild beers

People in cluster 1 prefer - Brown painted, Canadian beers



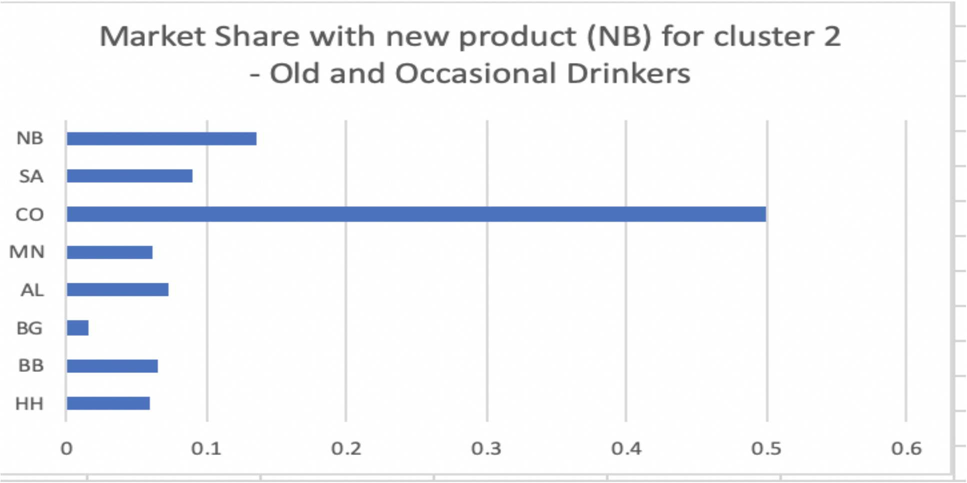
Market Share of cluster 1 with new brand –

| HH | BB | BG | AL | MN | CO | SA | NB |
|---|---|---|---|---|---|---|---|
| 0.181 | 0.118 | 0.232 | 0.147 | 0 | 0.289 | 0.001 | 0.032 |



Market Share of cluster 2 with new brand –

| HH | BB | BG | AL | MN | CO | SA | NB |
|---|---|---|---|---|---|---|---|
| 0.059 | 0.066 | 0.016 | 0.073 | 0.062 | 0.5 | 0.09 | 0.135 |


Market Share with new product (NB) for cluster 2
- Old and Occasional Drinkers
NB
SA
CO
MN
AL
BG
BB
HH