Mall Case Study

Rana Samir

10/01/2021

Investigate the dataset

```
summary(mall)
```

```
##    CustomerID        Gender               Age        Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00
##  Spending.Score..1.100.
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00
```

```
str(mall)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : chr  "Male" "Male" "Female" "Female" ...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

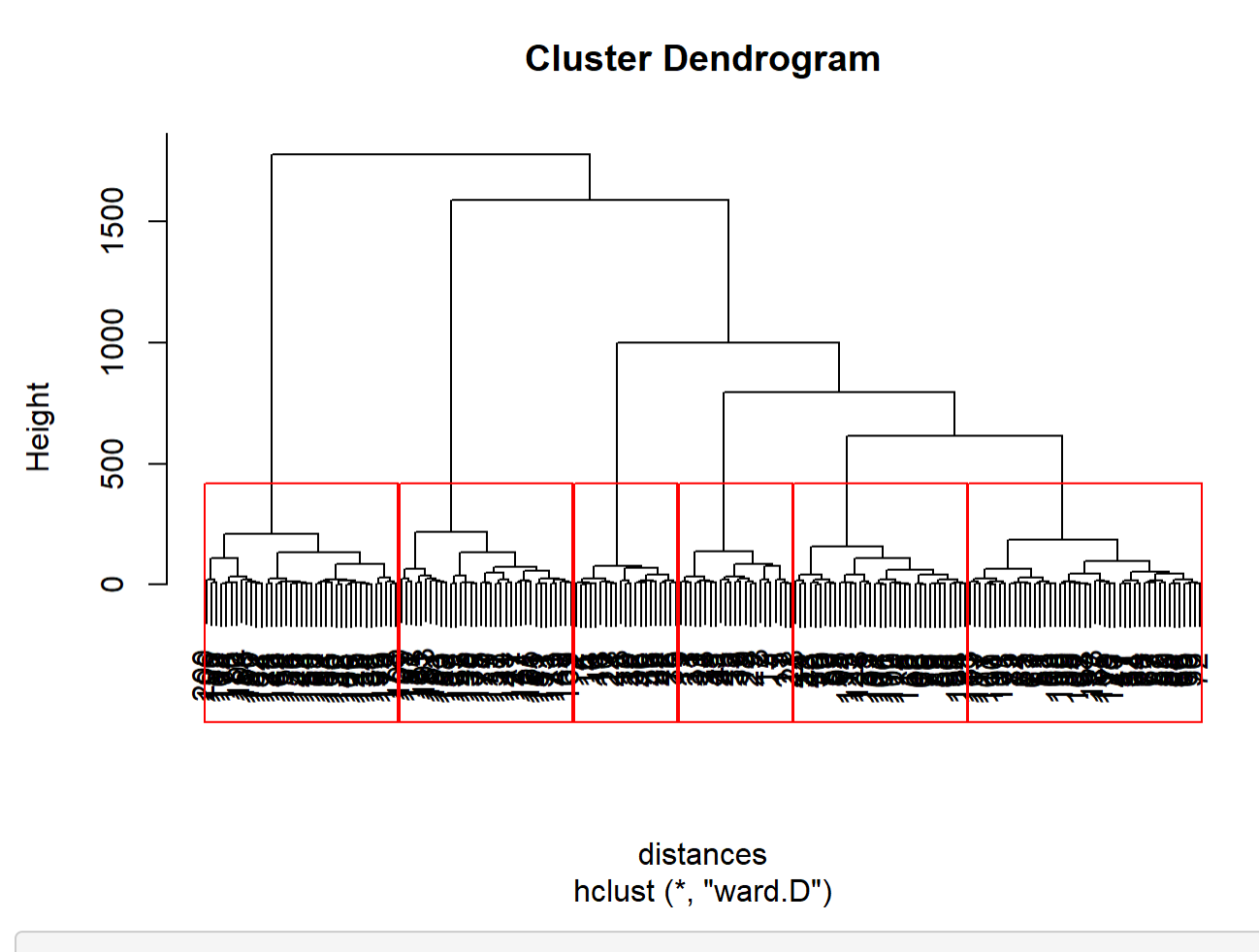
Prepare the data for analysis

```
# data preparation

#Normalize data
#normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}
#mall$Age=normalize(mall$Age)
#mall$Annual.Income..k..=normalize(mall$Annual.Income..k..)
#mall$Spending.Score..1.100.= normalize(mall$Spending.Score..1.100.)
#Remove unnecesser columns
mall$CustomerID=NULL
#Remove duplicates
mall=unique(mall)
```

Hierarchical Clustering

```
#Start the Hirearichal clustring

#Compute distance
distances= dist(mall[2:4], method="euclidean")
clustermall= hclust(distances, method="ward.D")
```

Plot The Dendogram




```
#Assign clusters
clustegroups= cutree(clustermall, k=6)
#Invistigate the clusters
tapply(mall$Age, clustegroups, mean)
```

```
##        1        2        3        4        5        6
## 45.21739 24.80952 25.97143 54.59574 32.69231 41.68571
```

```
tapply(mall$Annual.Income..k.., clustegroups, mean)
```

```
##        1        2        3        4        5        6
## 26.30435 25.61905 55.60000 54.46809 86.53846 88.22857
```

```
tapply(mall$Spending.Score..1.100., clustegroups, mean)
```

```
##        1        2        3        4        5        6
## 20.91304 80.23810 48.77143 50.19149 82.12821 17.28571
```

Cluster #1 charachterstics

```
G1 = subset(mall, clustegroups==1)
G1
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
1	Male	19	15	39
3	Female	20	16	6
5	Female	31	17	40
7	Female	35	18	6
9	Male	64	19	3
11	Male	67	19	14
13	Female	58	20	15
15	Male	37	20	13
17	Female	35	21	35
19	Male	52	23	29

1-10 of 23 rows Previous 1 2 3 Next

Show G1 summary

```
summary(G1)
```

```
##     Gender               Age        Annual.Income..k.. Spending.Score..1.100.
##  Length:23          Min.   :19.00   Min.   :15.0       Min.   : 3.00
##  Class :character   1st Qu.:35.50   1st Qu.:19.5       1st Qu.: 9.50
##  Mode  :character   Median :46.00   Median :25.0       Median :17.00
##                     Mean   :45.22   Mean   :26.3       Mean   :20.91
##                     3rd Qu.:53.50   3rd Qu.:33.0       3rd Qu.:33.50
##                     Max.   :67.00   Max.   :39.0       Max.   :40.00
```

```
table(G1$Gender)
```

```
##
## Female   Male
##     14      9
```

Spending score is low

Annual Income is average

However, both spending score and annual income are small. The mean value of age in this cluster is 45, however this cluster have a various ages in it resulting in a high range of ages within the cluster.

Females is two times more than males

Cluster #2 charachterstics

```
G2 = subset(mall, clustegroups==2)
G2
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
2	Male	21	15	81
4	Female	23	16	77
6	Female	22	17	76
8	Female	23	18	94
10	Female	30	19	72
12	Female	35	19	99
14	Female	24	20	77
16	Male	22	20	79
18	Male	20	21	66
20	Female	35	23	98

1-10 of 21 rows Previous 1 2 3 Next

```
summary(G2)
```

```
##     Gender               Age        Annual.Income..k.. Spending.Score..1.100.
##  Length:21          Min.   :18.00   Min.   :15.00      Min.   :65.00
##  Class :character   1st Qu.:21.00   1st Qu.:19.00      1st Qu.:73.00
##  Mode  :character   Median :23.00   Median :24.00      Median :77.00
##                     Mean   :24.81   Mean   :25.62      Mean   :80.24
##                     3rd Qu.:29.00   3rd Qu.:33.00      3rd Qu.:87.00
##                     Max.   :35.00   Max.   :39.00      Max.   :99.00
```

```
table(G2$Gender)
```

```
##
## Female   Male
##     13      8
```

Spending score is High

Annual Income is average

Howeever, both spending score and annual income are small. The mean value of age in this cluster is between 25-35 age group. And females is more than 60% of the cluster

Cluster #3 charachterstics

```
G3 = subset(mall, clustegroups==3)
G3
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
28	Male	35	28	61
44	Female	31	39	61
48	Female	27	40	47
49	Female	29	40	42
50	Female	31	40	42
52	Male	33	42	60
53	Female	31	43	54
59	Female	27	46	51
62	Male	19	46	55
66	Male	18	48	59

1-10 of 35 rows Previous 1 2 3 4 Next

```
summary(G3)
```

```
##     Gender               Age        Annual.Income..k.. Spending.Score..1.100.
##  Length:35          Min.   :18.00   Min.   :28.0       Min.   :29.00
##  Class :character   1st Qu.:20.50   1st Qu.:47.0       1st Qu.:42.00
##  Mode  :character   Median :26.00   Median :60.0       Median :50.00
##                     Mean   :25.97   Mean   :55.6       Mean   :48.77
##                     3rd Qu.:31.00   3rd Qu.:63.5       3rd Qu.:55.00
##                     Max.   :40.00   Max.   :76.0       Max.   :61.00
```

```
table(G3$Gender)
```

```
##
## Female   Male
##     23     12
```

Spending score is Average

Annual Income is relatively high

This cluster have a low spending score relative to its annual income. The mean value of age in this cluster is 25, however this cluster have a various ages in it resulting in a high range of ages within the cluster.

Females is two times more than males

```
G4 = subset(mall, clustegroups==4)
G4
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
47	Female	50	40	55
51	Female	49	42	52
54	Male	59	43	60
55	Female	50	43	45
56	Male	47	43	41
57	Female	51	44	50
58	Male	69	44	46
60	Male	53	46	46
61	Male	70	46	56
63	Female	67	47	52

1-10 of 47 rows Previous 1 2 3 4 5 Next

```
summary(G4)
```

```
##     Gender               Age       Annual.Income..k.. Spending.Score..1.100.
##  Length:47          Min.   :34.0   Min.   :40.00      Min.   :41.00
##  Class :character   1st Qu.:48.0   1st Qu.:48.00      1st Qu.:46.00
##  Mode  :character   Median :51.0   Median :54.00      Median :50.00
##                     Mean   :54.6   Mean   :54.47      Mean   :50.19
##                     3rd Qu.:64.0   3rd Qu.:62.00      3rd Qu.:55.00
##                     Max.   :70.0   Max.   :69.00      Max.   :60.00
```

```
table(G4$Gender)
```

```
##
## Female   Male
##     26     21
```

Spending score is Average

Annual Income is above average

This cluster have a low spending score relative to its annual income. The mean value of age in this cluster is 25, however this cluster have a various ages in it resulting in a high range of ages within the cluster.

Females is slightly higher than males

```
G5 = subset(mall, clustegroups==5)
G5
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
124	Male	39	69	91
126	Female	31	70	77
128	Male	40	71	95
130	Male	38	71	75
132	Male	39	71	75
134	Female	31	72	71
136	Female	29	73	88
138	Male	32	73	73
140	Female	35	74	72
142	Male	32	75	93

1-10 of 39 rows Previous 1 2 3 4 Next

```
summary(G5)
```

```
##     Gender               Age        Annual.Income..k.. Spending.Score..1.100.
##  Length:39          Min.   :27.00   Min.   : 69.00     Min.   :63.00
##  Class :character   1st Qu.:30.00   1st Qu.: 75.50     1st Qu.:74.50
##  Mode  :character   Median :32.00   Median : 79.00     Median :83.00
##                     Mean   :32.69   Mean   : 86.54     Mean   :82.13
##                     3rd Qu.:35.50   3rd Qu.: 95.00     3rd Qu.:90.00
##                     Max.   :40.00   Max.   :137.00     Max.   :97.00
```

```
table(G5$Gender)
```

```
##
## Female   Male
##     21     18
```

Spending score is high

Annual Income is high

This cluster have a high spending score as well as high annual income The mean value of age in this cluster is 32, with no high variability in the age range

Females to males ration is negligable

```
G6 = subset(mall, clustegroups==6)
G6
```

	Gender <chr>	Age <int>	Annual.Income..k.. <int>	Spending.Score..1.100. <int>
127	Male	43	71	35
129	Male	59	71	11
131	Male	47	71	9
135	Male	20	73	5
137	Female	44	73	7
139	Male	19	74	10
141	Female	57	75	5
145	Male	25	77	12
147	Male	48	77	36
149	Female	34	78	22

1-10 of 35 rows Previous 1 2 3 4 Next

```
summary(G6)
```

```
##     Gender               Age        Annual.Income..k.. Spending.Score..1.100.
##  Length:35          Min.   :19.00   Min.   : 71.00     Min.   : 1.00
##  Class :character   1st Qu.:35.00   1st Qu.: 77.50     1st Qu.:10.00
##  Mode  :character   Median :43.00   Median : 85.00     Median :16.00
##                     Mean   :41.69   Mean   : 88.23     Mean   :17.29
##                     3rd Qu.:47.50   3rd Qu.: 97.50     3rd Qu.:23.50
##                     Max.   :59.00   Max.   :137.00     Max.   :39.00
```

```
table(G6$Gender)
```

```
##
## Female   Male
##     15     20
```

Spending score is low

Annual Income is high

This cluster have a low spending score vs a annual income. The mean value of age in this cluster is 41, however this cluster have a various ages in it resulting in a high range of ages within the cluster.

This is the only cluster that has males higher than femals.

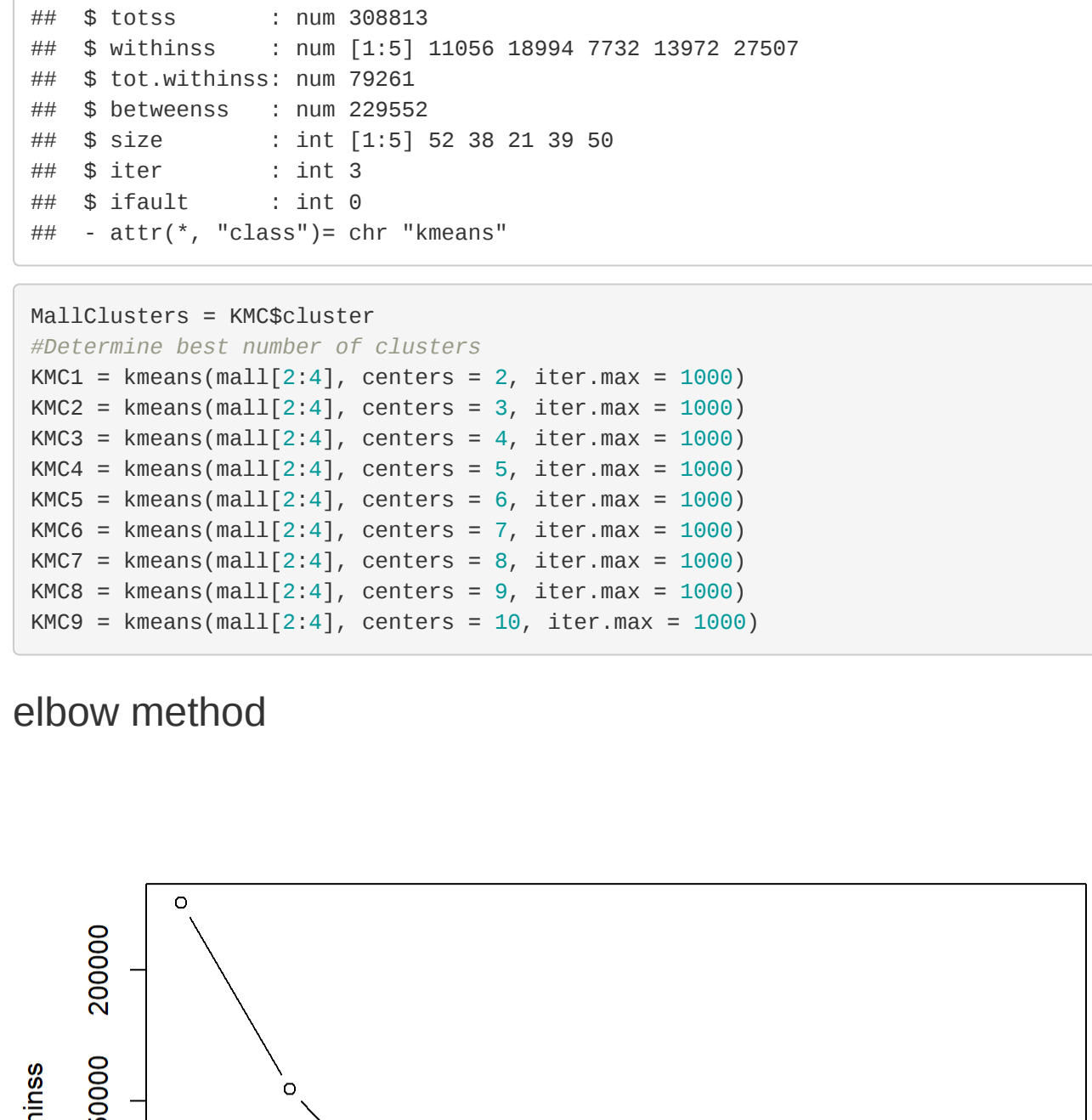
K-Means Clustring

```
# Run k-means
k=5
set.seed(1)
KMC = kmeans(mall[2:4], centers = k, iter.max = 1000)
str(KMC)
```

```
## List of 9
##  $ cluster     : Named int [1:200] 3 5 3 5 3 5 3 5 3 5 ...
##   ..- attr(*, "names")= chr [1:200] "1" "2" "3" "4" ...
##  $ centers     : num [1:5, 1:3] 53.7 40.4 44.1 32.7 24.8 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:5] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:3] "Age" "Annual.Income..k.." "Spending.Score..1.100."
##  $ totss       : num 308813
##  $ withinss    : num [1:5] 11056 18994 7732 13972 27507
##  $ tot.withinss: num 79261
##  $ betweenss   : num 229552
##  $ size        : int [1:5] 52 38 21 39 50
##  $ iter        : int 3
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

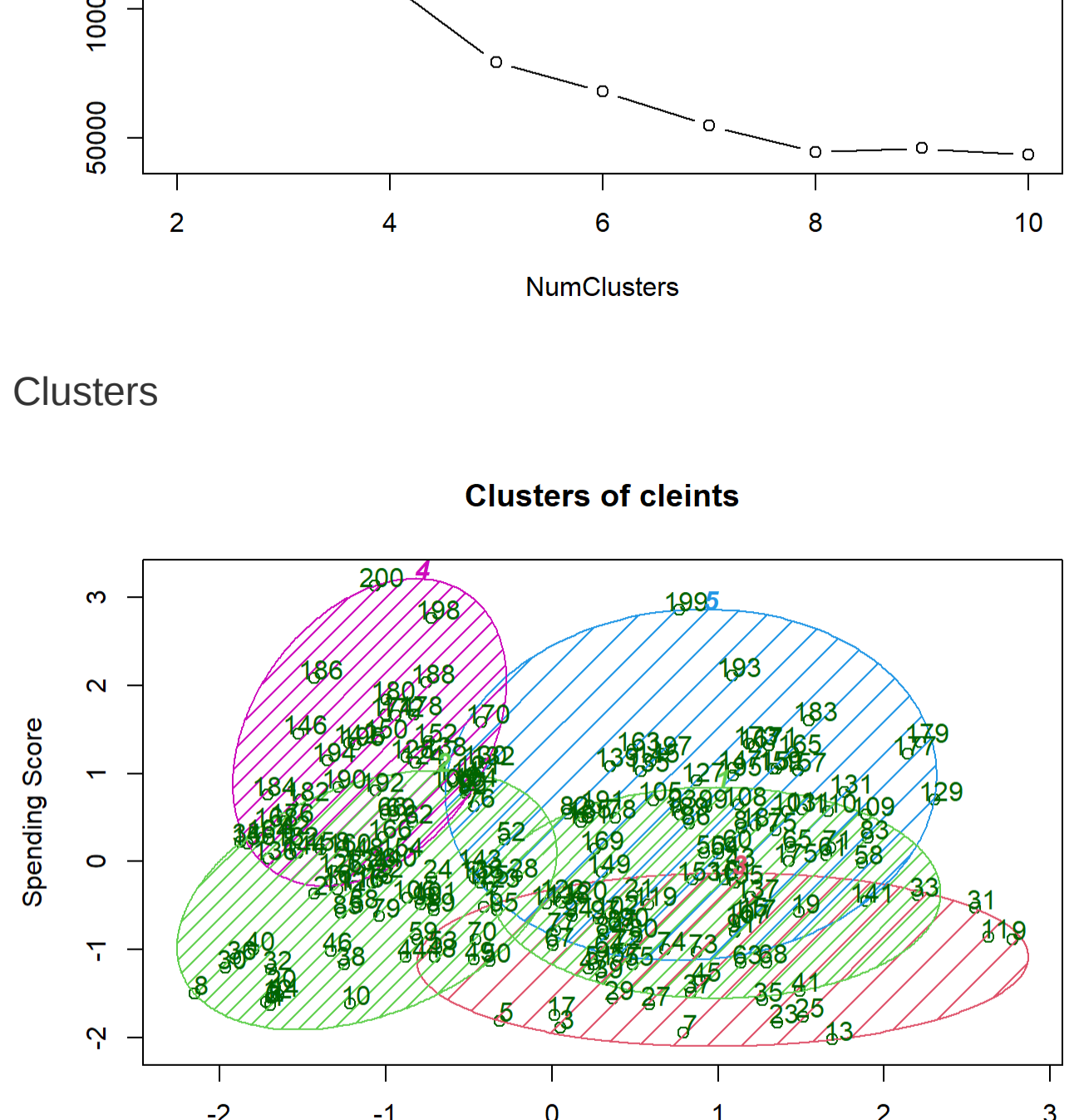
```
MallClusters = KMC$cluster
#Determine best number of clusters
KMC1 = kmeans(mall[2:4], centers = 2, iter.max = 1000)
KMC2 = kmeans(mall[2:4], centers = 3, iter.max = 1000)
KMC3 = kmeans(mall[2:4], centers = 4, iter.max = 1000)
KMC4 = kmeans(mall[2:4], centers = 5, iter.max = 1000)
KMC5 = kmeans(mall[2:4], centers = 6, iter.max = 1000)
KMC6 = kmeans(mall[2:4], centers = 7, iter.max = 1000)
KMC7 = kmeans(mall[2:4], centers = 8, iter.max = 1000)
KMC8 = kmeans(mall[2:4], centers = 9, iter.max = 1000)
KMC9 = kmeans(mall[2:4], centers = 10, iter.max = 1000)
```

elbow method



5 is the best number of cluster

Clusters




```
aggregate(mall, by=list(cluster=KMC4$cluster), mean)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or logical:
## returning NA
```

cluster <int>	Gender <dbl>	Age <dbl>	Annual.Income..k.. <dbl>	Spending.Score..1.100. <dbl>
1	NA	53.71154	54.42308	48.73077
2	NA	24.80000	41.46000	63.70000
3	NA	44.14286	25.14286	19.52381
4	NA	32.69231	86.53846	82.12821
5	NA	40.39474	87.00000	18.63158

5 rows

```
KMC4$size
```

```
## [1] 52 50 21 39 38
```

Conclusion

For the Hierarchical clustring

Clusters 5,6 have the highest income. Clusters, 1,2 have the lowest income, and the other two clusters, 3,4 have an average income.

The spending score of these clusters is not proportional to their income. As the highest spending score clusters are 2,5, and the lowest spending score was given to people in cluster 6.

The notable trend in the data was that the high spending score was in the clusters of ages between 25-35 with a low variability within the cluster, like 2,5. The 40s age groups are all have a low spending score.

The gender of the customers didn't affect the spending score in general So the recommendation for a marketing team, in order to have more customer acquisitions, to make the main targeting to the clusters of low ages and average income and spending score in order to make them a high spending customers. So focus on clusters 3,4. Then Secondarly target clusters with high income but low spending score and a various age groups within such as cluster 6. Cluser 1 was neglected from the targetting as it have a low income and also a high mean age within the cluster. Clusters 2,5 are already a loyal high spending customers that doesn't need an advertising ad especially for them and the rest of the targetting will work as a nice brand awerness for them.

For the K-Means clustring

The results of the K-Means clustring wasn't very different from the Hierarchical clustring. The most high spending cluster is 4, the least spending clusters are 3,5, and the last cluaters,1,2, have an average spending score. The trend deduced from the data are the same as the last one. the high spending score is associated with the younger age groups. and the 40s clusters have a low spending score. So the recomendation will be to target the clusters of the average spending and other than the 40s age, and target clusters with a high dispertion in age means. As mentioned the averge spending score clusers are 1,2 which have a mean age values of 25 and 54.

There were no big effect on the conclusion due to the change in the number of clusters between Hierarchical and k-means methods. Even in the number of recommended clusters to be targtted, there are no significant difference in the number of people in both clusters. However, I think the 6 clusters will be more accurate in the targetting campigns. So the following recommendations will be based on the Hierarchical.

Summary and Recommendation

- Clusters 3,4 with average income and average spending score can be easily converge, so they should be targetted by advertising campaigns.
- Cluster 2 with low income but high spending score should be given a speical offers to make sure they stay loyal to the brand.
- Cluster 5 with the highest spending score and high income, should be targetted periodically, as they have more tendency to spend more in the brand.
- Cluster 1,6 have a low spending score although their distinct income values, they should be given a offers for free stuff to try the brands in order to make an acquisition.