k-means Clustering
Author: Bharat Thakur

Background

In this assignment dataset containing TripAdvisor reviews of 249 different high volume reviewers in various categories was analyzed. The dataset comprises of anonymized user information and reviews in each of six categories i.e., sports, religious, nature, theatre, shopping, and picnic. After data transformation and performing descriptive analysis, K-means clustering algorithm was employed to segment the reviewers into distinct clusters.

Data Source

The reviews dataset used in this assignment was shared in the PROG8430 course.

Part 1

1. Data Transformation

Rename

The columns of the review dataset were renamed to meet assignment requirements.

```
names(Review_BT)

  [1] "User_Id_BT"   "Sports_BT"    "Religious_BT" "Nature_BT"    Theatre_BT"  
  [6] "Shopping_BT"  "Picnic_BT"    "Age_BT"       "Income_BT"    "Nbr_BT"
```

Standardize

In next step, the variables were scaled to have values in range [0,1]. Min-max scaling method was used over the normalization because dataset being relatively small and absence of significant outliers. All the variables were standardized but displaying only assigned variables i.e., “Religious” and “Shopping”.

```
#Create a standardization function
norm01 <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}


#Standardizing Variable Religious
Review_BT$Religious_MinMax_BT <- norm01(Review_BT$Religious_BT)


#Standardizing Variable Shopping
Review_BT$Shopping_MinMax_BT <- norm01(Review_BT$Shopping_BT)
```

2. Descriptive Data Analysis

Quantitative Summary

The various sample statistics and preview of data structure for the review dataset were generated using summary(), stat.desc() and str() function in R as shown below:

```
summary(Review_BT)

   User_Id_BT          Sports_BT         Religious_BT      Nature_BT    
  Length:249         Min.   :0.0051   Min.   :0.109   Min.   :0.088  
  Class :character   1st Qu.:0.0119   1st Qu.:0.156   1st Qu.:0.166  
  Mode  :character   Median :0.0192   Median :0.179   Median :0.208  
                     Mean   :0.0187   Mean   :0.184   Mean   :0.210  
                     3rd Qu.:0.0248   3rd Qu.:0.211   3rd Qu.:0.266  
                     Max.   :0.0323   Max.   :0.274   Max.   :0.377  
    Theatre_BT      Shopping_BT       Picnic_BT         Age_BT        Income_BT    
  Min.   :0.112   Min.   :0.106   Min.   :0.144   Min.   :18.0   Min.   :  963  
  1st Qu.:0.163   1st Qu.:0.146   1st Qu.:0.180   1st Qu.:27.0   1st Qu.:23790  
  Median :0.187   Median :0.183   Median :0.197   Median :38.0   Median :47986  
  Mean   :0.197   Mean   :0.188   Mean   :0.202   Mean   :37.4   Mean   :47433  
  3rd Qu.:0.234   3rd Qu.:0.216   3rd Qu.:0.225   3rd Qu.:48.0   3rd Qu.:67165  
  Max.   :0.303   Max.   :0.319   Max.   :0.269   Max.   :55.0   Max.   :99949  
      Nbr_BT    Sports_MinMax_BT Religious_MinMax_BT Nature_MinMax_BT
  Min.   :353   Min.   :0.000    Min.   :0.000       Min.   :0.000   
  1st Qu.:494   1st Qu.:0.250    1st Qu.:0.285       1st Qu.:0.268   
  Median :595   Median :0.518    Median :0.425       Median :0.416   
  Mean   :596   Mean   :0.498    Mean   :0.460       Mean   :0.421   
  3rd Qu.:710   3rd Qu.:0.724    3rd Qu.:0.618       3rd Qu.:0.614   
  Max.   :843   Max.   :1.000    Max.   :1.000       Max.   :1.000   
  Theatre_MinMax_BT Shopping_MinMax_BT Picnic_MinMax_BT Age_MinMax_BT
  Min.   :0.000     Min.   :0.000      Min.   :0.000    Min.   :0.000  
  1st Qu.:0.269     1st Qu.:0.187      1st Qu.:0.291    1st Qu.:0.243  
  Median :0.394     Median :0.365      Median :0.425    Median :0.541  
  Mean   :0.450     Mean   :0.385      Mean   :0.464    Mean   :0.523  
  3rd Qu.:0.639     3rd Qu.:0.519      3rd Qu.:0.654    3rd Qu.:0.811  
  Max.   :1.000     Max.   :1.000      Max.   :1.000    Max.   :1.000  
  Income_MinMax_BT NBR_MinMax_BT
  Min.   :0.000    Min.   :0.000  
  1st Qu.:0.231    1st Qu.:0.288  
  Median :0.475    Median :0.494  
  Mean   :0.469    Mean   :0.495  
  3rd Qu.:0.669    3rd Qu.:0.729  
  Max.   :1.000    Max.   :1.000  
```

```
stat.desc(Review_BT)

          User_Id_BT   Sports_BT Religious_BT Nature_BT Theatre_BT Shopping_BT
 nbr.val          NA 249.0000000    249.00000 249.00000  249.00000   249.00000
 nbr.null         NA   0.0000000      0.00000   0.00000    0.00000     0.00000
 nbr.na           NA   0.0000000      0.00000   0.00000    0.00000     0.00000
 min              NA   0.0050761      0.10867   0.08828    0.11151     0.10558
 max              NA   0.0323415      0.27363   0.37722    0.30266     0.31902
 range            NA   0.0272654      0.16496   0.28895    0.19115     0.21344
 sum              NA   4.6470097     45.93361  52.27343   49.16132    46.76731
 median           NA   0.0192000      0.17880   0.20849    0.18689     0.18343
 mean             NA   0.0186627      0.18447   0.20993    0.19744     0.18782
 SE.mean          NA   0.0004562      0.00233   0.00389    0.00275     0.00312
 CI.mean          NA   0.0008985      0.00459   0.00766    0.00542     0.00615
```

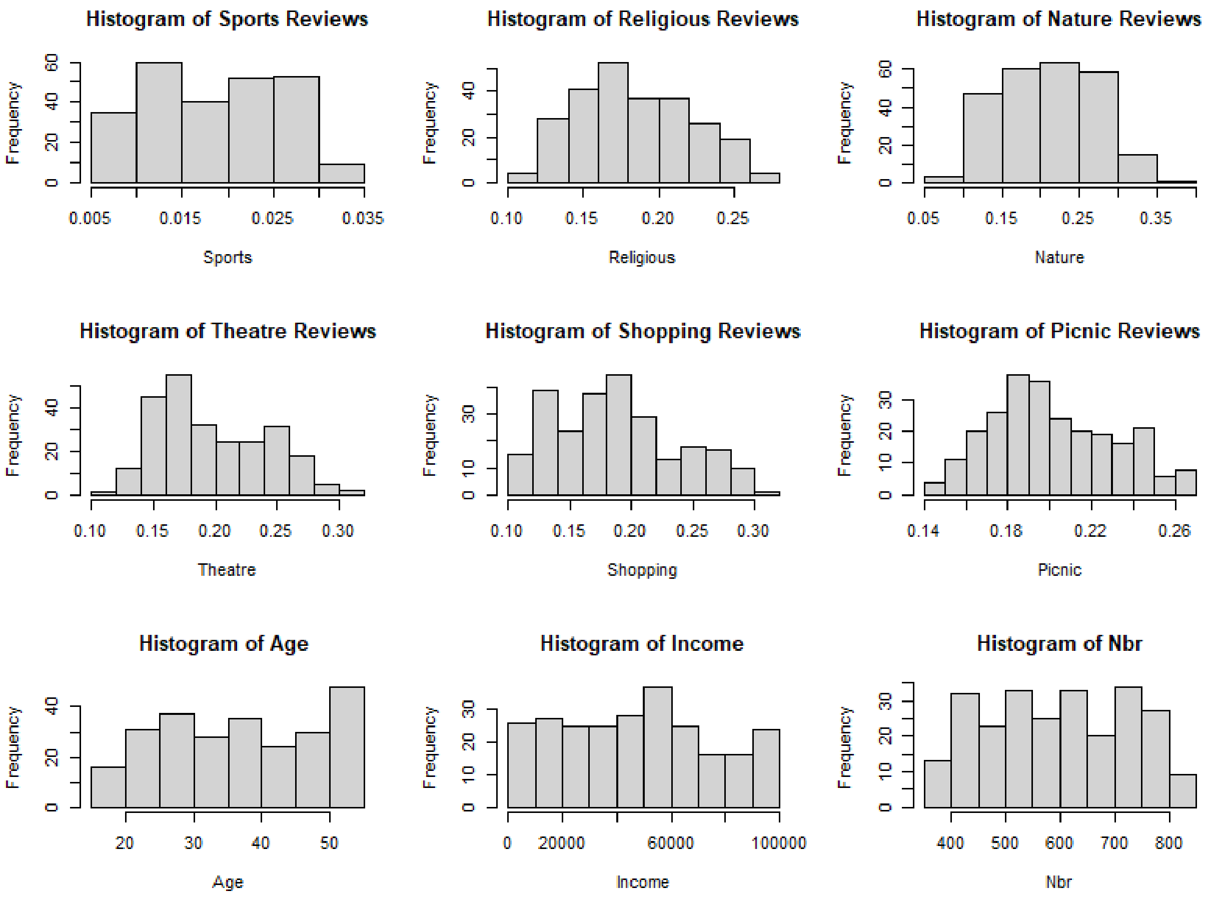
```
var              NA   0.0000518      0.00135   0.00376    0.00189     0.00243
std.dev          NA   0.0071989      0.03680   0.06135    0.04342     0.04929
coef.var         NA   0.3857383      0.19950   0.29222    0.21994     0.26245
          Picnic_BT   Age_BT     Income_BT     Nbr_BT Sports_MinMax_BT
nbr.val  249.000000  249.000       249.000    249.000         249.0000
nbr.null   0.000000    0.000         0.000      0.000           1.0000
nbr.na     0.000000    0.000         0.000      0.000           0.0000
min        0.143780   18.000       962.900    353.000           0.0000
max        0.268603   55.000     99949.100    843.000           1.0000
range      0.124822   37.000     98986.200    490.000           1.0000
sum       50.217320 9303.000  11810928.500 148330.000         124.0786
median     0.196891   38.000     47985.800    595.000           0.5180
mean       0.201676   37.361     47433.448    595.703           0.4983
SE.mean    0.001870    0.716      1757.358      8.125           0.0167
CI.mean    0.003684    1.410      3461.249     16.002           0.0330
var        0.000871  127.651 768988463.254  16436.911           0.0697
std.dev    0.029513   11.298     27730.641    128.207           0.2640
coef.var   0.146339    0.302         0.585      0.215           0.5299
         Religious_MinMax_BT Nature_MinMax_BT Theatre_MinMax_BT
nbr.val             249.0000         249.0000          249.0000
nbr.null              1.0000           1.0000            1.0000
nbr.na                0.0000           0.0000            0.0000
min                   0.0000           0.0000            0.0000
max                   1.0000           1.0000            1.0000
range                 1.0000           1.0000            1.0000
sum                 114.4211         104.8379          111.9329
median                0.4251           0.4161            0.3943
mean                  0.4595           0.4210            0.4495
SE.mean               0.0141           0.0135            0.0144
CI.mean               0.0278           0.0265            0.0284
var                   0.0498           0.0451            0.0516
std.dev               0.2231           0.2123            0.2272
coef.var              0.4855           0.5043            0.5054
         Shopping_MinMax_BT Picnic_MinMax_BT Age_MinMax_BT Income_MinMax_BT
nbr.val            249.0000         249.0000      249.0000         249.0000
nbr.null             1.0000           1.0000        9.0000           1.0000
nbr.na               0.0000           0.0000        0.0000           0.0000
min                  0.0000           0.0000        0.0000           0.0000
max                  1.0000           1.0000        1.0000           1.0000
range                1.0000           1.0000        1.0000           1.0000
sum                 95.9462         115.4924      130.2973         116.8968
median               0.3648           0.4255        0.5405           0.4750
mean                 0.3853           0.4638        0.5233           0.4695
SE.mean              0.0146           0.0150        0.0194           0.0178
CI.mean              0.0288           0.0295        0.0381           0.0350
var                  0.0533           0.0559        0.0932           0.0785
std.dev              0.2309           0.2364        0.3054           0.2801
coef.var             0.5993           0.5098        0.5835           0.5967
         NBR_MinMax_BT
nbr.val       249.0000
nbr.null        1.0000
nbr.na          0.0000
min             0.0000
max             1.0000
range           1.0000
```

```
 sum           123.3327
 median          0.4939
 mean            0.4953
```

```
str(Review_BT)
```

```
 'data.frame':    249 obs. of  19 variables:
  $ User_Id_BT         : chr  "User 1" "User 2" "User 3" "User 4" ...
  $ Sports_BT          : num  0.00513 0.00567 0.00554 0.00528 0.00508 ...
  $ Religious_BT       : num  0.197 0.176 0.139 0.179 0.249 ...
  $ Nature_BT          : num  0.203 0.215 0.269 0.203 0.137 ...
  $ Theatre_BT         : num  0.177 0.215 0.241 0.251 0.15 ...
  $ Shopping_BT        : num  0.174 0.195 0.139 0.201 0.241 ...
  $ Picnic_BT          : num  0.244 0.193 0.208 0.161 0.218 ...
  $ Age_BT             : int  44 35 18 24 52 51 32 43 33 48 ...
  $ Income_BT          : num  53441 61412 66359 61344 53164 ...
  $ Nbr_BT             : int  390 353 361 379 394 385 376 386 386 416 ...
  $ Sports_MinMax_BT   : num  0.00191 0.02162 0.01702 0.00737 0 ...
  $ Religious_MinMax_BT: num  0.538 0.406 0.181 0.429 0.849 ...
  $ Nature_MinMax_BT   : num  0.396 0.44 0.624 0.398 0.169 ...
  $ Theatre_MinMax_BT  : num  0.342 0.543 0.677 0.728 0.2 ...
  $ Shopping_MinMax_BT : num  0.322 0.421 0.154 0.445 0.635 ...
  $ Picnic_MinMax_BT   : num  0.8 0.391 0.513 0.138 0.597 ...
  $ Age_MinMax_BT      : num  0.703 0.459 0 0.162 0.919 ...
  $ Income_MinMax_BT   : num  0.53 0.611 0.661 0.61 0.527 ...
  $ NBR_MinMax_BT      : num  0.0755 0 0.0163 0.0531 0.0837 ...
```

Graphical Summary (Before Transformation)

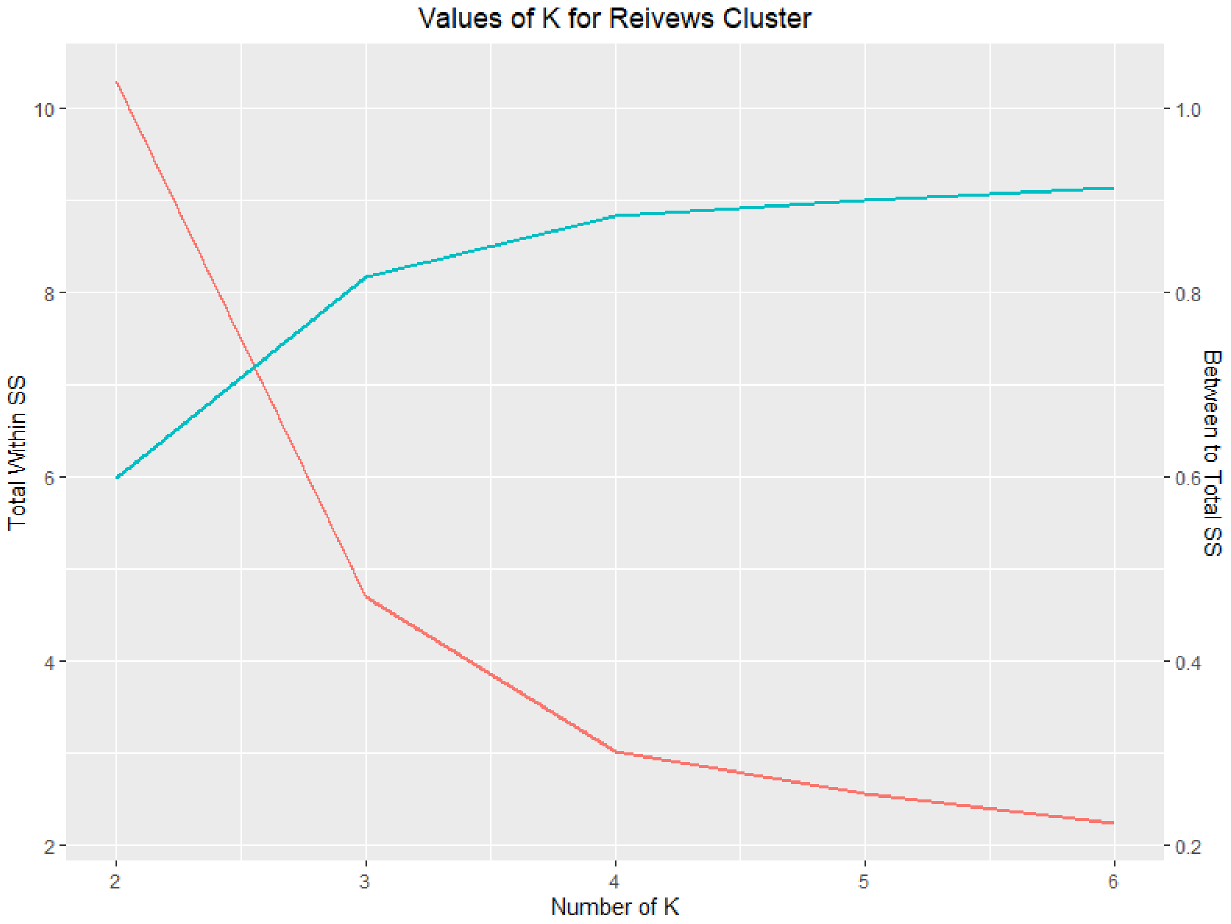


3. Clustering

3.1 Create segmentation/cluster schemes for k=2,3,4,5,6.

```
####################################################
## Create Clusters for K = 2:6                    ##
####################################################

for(i in 2:6){

  tmp_clstr <- paste("Clstr", "cnt", toString(i), "Rev", "BT", sep = "_")
  ClstrRev_BT <- kmeans(ReviewClstrData_BT, iter.max=10, centers=i, nstart=10)
  assign(tmp_clstr,ClstrRev_BT)

  Review_BT[paste("cluster_", toString(i), sep = "")]<- factor(ClstrRev_BT$cluster)
# Adding Cluster tags to variables
  Review_BT$cluster <- factor(ClstrRev_BT$cluster)
}

Review_BT$cluster <- NULL
```

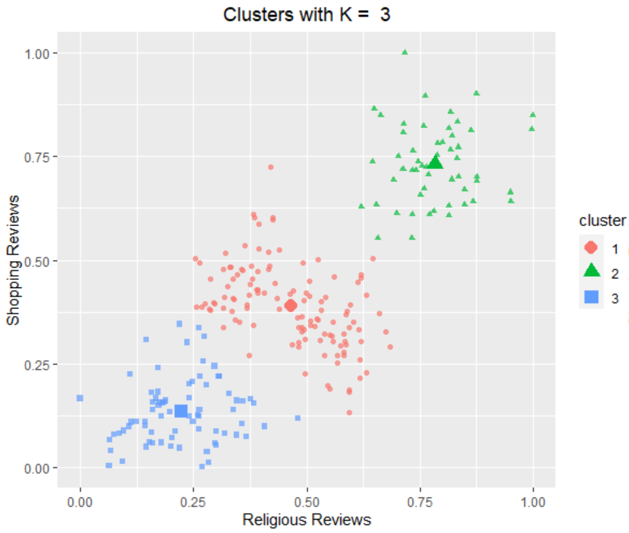
3.2 Create the WSS plots as demonstrated in class and select a suitable k value based on the "elbow".

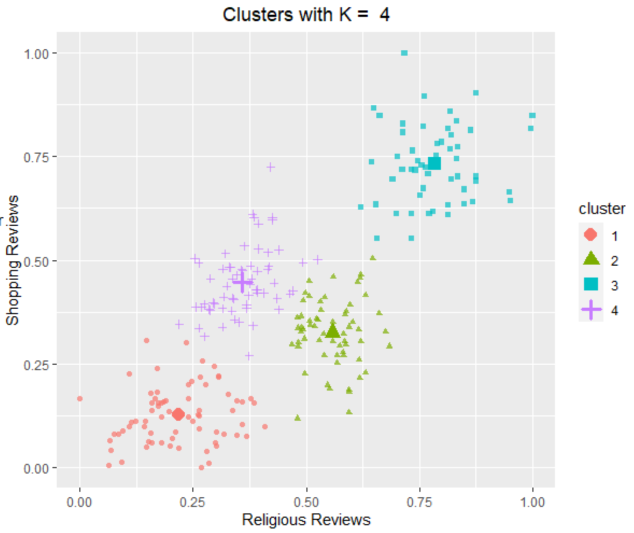


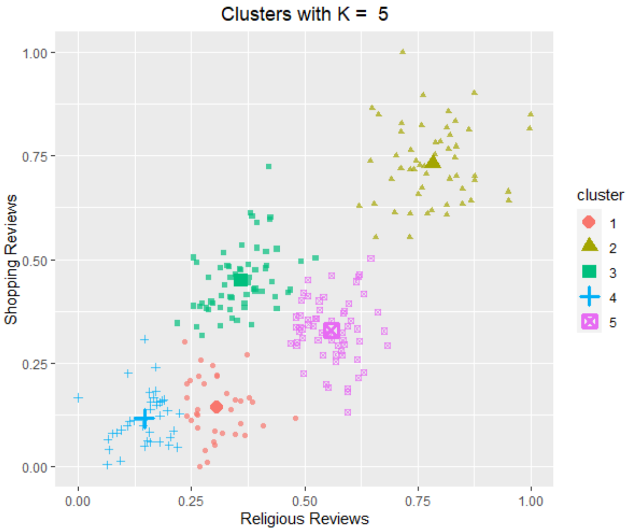
Comment: Based on elbow plot it appears from K = 4 onwards, increasing cluster count offers diminishing returns in terms of reducing sum of squares within clusters and increasing percentage of variance explained.

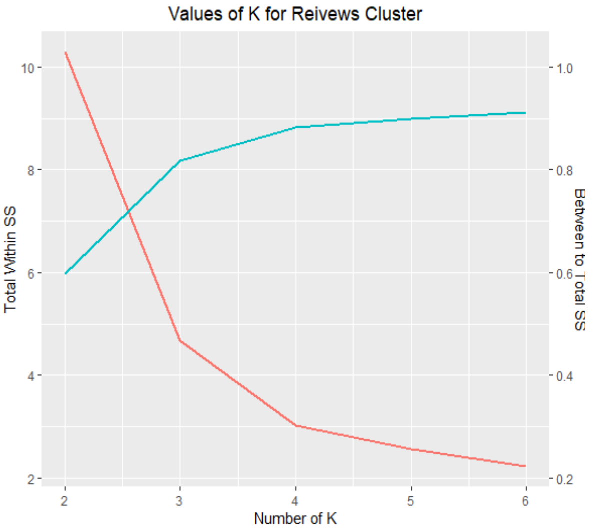
4. Evaluation of Clusters

4.1 Scatter Plot and WSS Plot









4.2 Choose best set of clusters

Visually it appears the clusters corresponding to K = 3, lead to more natural segmentation of the data. Further increasing cluster count from three to four does not lead to drastic reduction in sum of squares within clusters or increase in variance explained by additional cluster. Additional domain knowledge regarding the application of the resulting segmentation may influence the answer.

4.3 Summary Tables for Clusters

Cluster	Sports	Religious	Nature	Theatre	Shopping	Picnic	Age	Nbr	N
1	0.0206	0.145	**0.283**	0.203	0.134	0.214	37.4	600	75
2	0.0175	0.185	0.203	**0.209**	0.189	0.197	37.2	581	121
3	0.0185	**0.238**	0.123	0.162	**0.262**	0.197	37.7	624	53

4.4 Create suitable descriptive names for each cluster.

Cluster 1 – Hikers

Cluster 2 – Bardolator

Cluster 3 – Shopaholic Monks

4.5 Suggest possible uses for this clustering scheme

One suitable use of this clustering scheme would be making product recommendations. For example, reviewers belonging to cluster 3 are more inclined towards religious and shopping related activities. So, it would be a good idea to recommend shopping related activities to people who demonstrate religious proclivity or vice versa.