KMean Clustering

Kumar Gaurav

15/09/2020

```
#####################################################################
#----------- Loading Library-----------------------------------#
#####################################################################
library (cluster)
```

```
#####################################################################
#----------- Loading Dataset-----------------------------------#
#####################################################################
cust_spend=read.csv('/home/kumar/Documents/Projects and Practices/practice/R/cust_spe
nd_analysis/Cust_Spend_Data_New.csv')
head(cust_spend)
```

```
##   Cust_ID          Name Avg_Mthly_Spend No_Of_Visits Apparel_Items FnV_Items
## 1       1  Abraham \xca            1123           28             1        16
## 2       2    Adela \xca            9818           13             5         2
## 3       3  Adelina \xca            9824           10            10         2
## 4       4   Adrian \xca            3097           23             2         8
## 5       5 Adrianna \xca             817           28             1        17
## 6       6     Aide \xca            3039           21             1         8
##   Staples_Items
## 1            14
## 2             5
## 3             2
## 4             9
## 5            17
## 6            12
```

```
#####################################################################
#----------- Checking Null Value-----------------------------------#
#####################################################################
colSums(is.na(cust_spend))
```

```
##         Cust_ID            Name Avg_Mthly_Spend    No_Of_Visits   Apparel_Items
##               0               0               0               0               0
##       FnV_Items   Staples_Items
##               0               0
```

```
summary(cust_spend[,-c(1,2)])
```

```
##  Avg_Mthly_Spend  No_Of_Visits   Apparel_Items      FnV_Items
##  Min.   :  549   Min.   : 2.00   Min.   : 0.000   Min.   : 1.000
##  1st Qu.: 4156   1st Qu.:15.00   1st Qu.: 3.000   1st Qu.: 6.000
##  Median : 4516   Median :18.00   Median : 4.000   Median : 7.000
##  Mean   : 4801   Mean   :17.86   Mean   : 3.961   Mean   : 7.624
##  3rd Qu.: 4910   3rd Qu.:20.00   3rd Qu.: 5.000   3rd Qu.: 8.000
##  Max.   :10000   Max.   :29.00   Max.   :10.000   Max.   :19.000
##  Staples_Items
##  Min.   : 0.000
##  1st Qu.: 5.000
##  Median : 8.000
##  Mean   : 8.339
##  3rd Qu.:10.000
##  Max.   :20.000
```

```
#########################################################################
#---------- scaling Features-----------------------------------#
#########################################################################
cust_spend_scaled=scale(cust_spend[,3:7],center = T,scale = T)
head(cust_spend_scaled)
```

```
##      Avg_Mthly_Spend No_Of_Visits Apparel_Items  FnV_Items Staples_Items
## [1,]      -1.7886149    2.3226702    -1.4402261  2.2850177     1.3319197
## [2,]       2.4394544   -1.1120312     0.5051058 -1.5340945    -0.7855857
## [3,]       2.4423720   -1.7989715     2.9367706 -1.5340945    -1.4914208
## [4,]      -0.8287289    1.1777697    -0.9538931  0.1026678     0.1555278
## [5,]      -1.9374119    2.3226702    -1.4402261  2.5578114     2.0377548
## [6,]      -0.8569323    0.7198096    -1.4402261  0.1026678     0.8613629
```

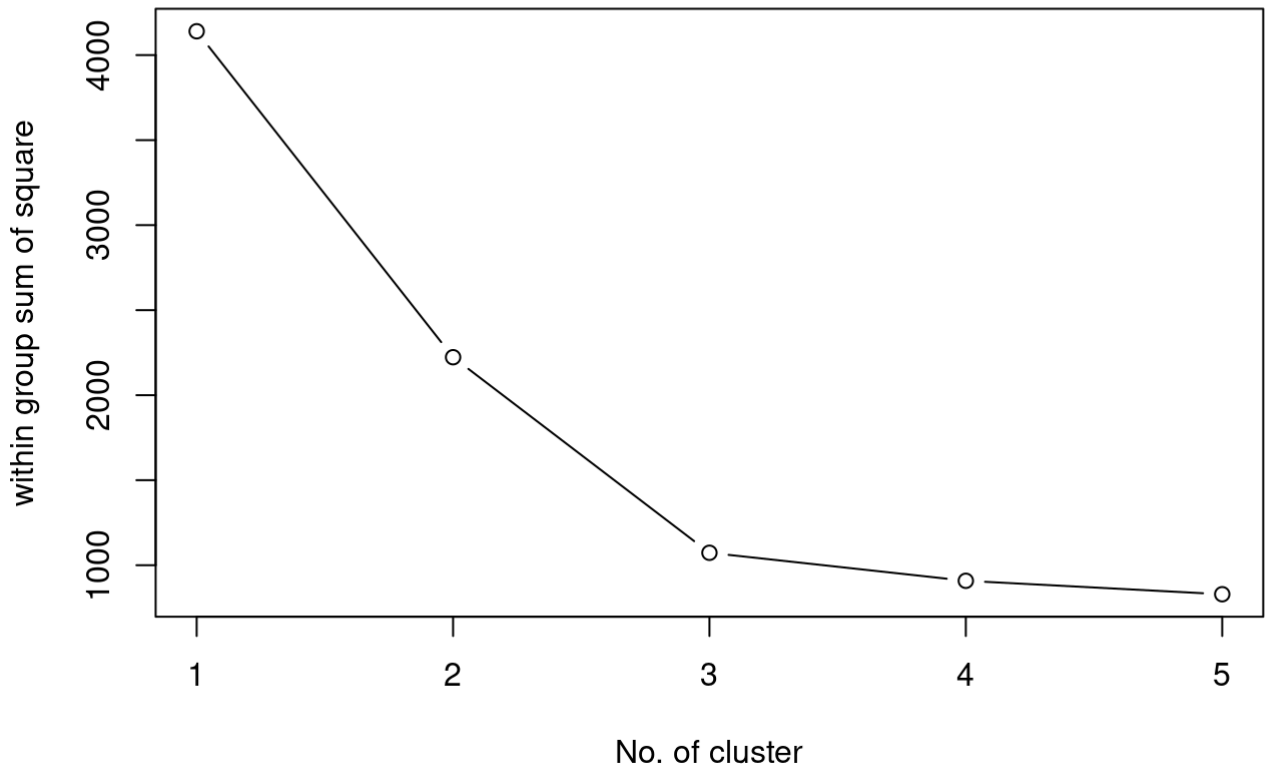
```
#########################################################################
#---------- Finding silhouette score----------------------------#
#########################################################################
k=2:10
silhouette_score=function(k){
  km<-kmeans(cust_spend_scaled,centers = k,nstart = 25)
  ss<-silhouette(km$cluster,dist(cust_spend_scaled))
  mean(ss[,3])
}


avg_sil=sapply(k,silhouette_score)

plot(k,type='b',avg_sil,frame=FALSE,xlab = 'No. of clusters',
     ylab = 'Avg Silhouette Score',main = 'Silhouette')
```


```
#####################################################################
#---------- Finding WSS score----------------------------------#
#####################################################################
wssplot=function(data,nc=15,seed=123){
  wss=c()
  for (i in 1:nc) {
    set.seed(seed)
    wss[i]=sum(kmeans(data,centers = i)$withinss)}
  plot(1:nc,wss,type = 'b',xlab = 'No. of cluster',
       ylab = 'within group sum of square')
}

wssplot(cust_spend_scaled,nc=5)
```




```
#########################################################################
#----------- K-mean Clustering-----------------------------------#
#########################################################################
kmean.clus=kmeans(x=cust_spend_scaled,centers = 3)
kmean.clus
```

```
## K-means clustering with 3 clusters of sizes 584, 139, 106
##
## Cluster means:
##   Avg_Mthly_Spend No_Of_Visits Apparel_Items   FnV_Items Staples_Items
## 1      -0.1529762  -0.01771023   -0.08448967 -0.07436781   -0.07249884
## 2       1.8495352  -1.23887628    1.44628250 -1.24167537   -1.19859234
## 3      -1.5825214   1.72213752   -1.43104998  2.03795923    1.97116658
##
## Clustering vector:
##   [1] 3 2 2 1 3 1 1 2 1 1 2 1 2 1 3 3 3 1 1 3 1 2 2 3 3 2 1 3 3 3 3 1 1 1 3 3 2
##  [38] 1 2 3 2 3 3 3 2 3 2 3 2 3 3 2 1 3 2 2 3 3 2 2 3 3 3 3 2 1 1 3 3 3 3 2 1 2
##  [75] 3 1 2 1 2 1 2 2 1 2 3 2 2 2 1 3 3 1 2 1 3 2 3 2 3 2 1 3 3 2 1 3 2 2 2 3 2
## [112] 2 2 1 2 1 2 2 1 2 2 2 3 3 2 3 2 2 3 2 1 2 3 3 1 1 1 2 1 1 3 3 3 2 2 3 2 2
## [149] 2 2 3 2 2 2 1 3 2 3 3 1 1 3 1 2 1 2 1 3 3 1 1 2 1 2 1 2 2 2 2 3 2 1 2 1 1
## [186] 3 3 2 2 3 3 1 1 1 1 2 1 1 1 2 3 2 1 3 1 2 2 1 3 3 1 3 3 3 1 2 3 1 3 2 2 2
## [223] 1 1 3 3 2 2 1 2 3 2 2 1 2 2 2 1 3 3 3 3 1 3 2 3 2 3 2 3 2 1 1 2 2 1 2 3 1
## [260] 3 2 3 3 2 3 3 1 2 2 3 2 2 2 2 1 2 1 2 3 3 2 1 2 2 2 3 3 2 2 3 2 2 2 2 3 2
## [297] 2 1 1 2 1 2 2 1 2 2 2 1 3 3 3 2 1 2 1 3 1 2 2 1 1 3 1 3 1 2 2 2 2 2 2 1 1
## [334] 2 2 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [445] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [482] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [519] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [556] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [593] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [630] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [704] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [815] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##
## Within cluster sum of squares by cluster:
## [1] 681.4936 237.4893 154.1035
##  (between_SS / total_SS =  74.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
cust_spend$Cluster=kmean.clus$cluster
#View(cust_spend)

aggr=aggregate(cust_spend[,-c(1,2,8)],list(cust_spend$Cluster),mean)
aggr
```

```
##   Group.1 Avg_Mthly_Spend No_Of_Visits Apparel_Items FnV_Items Staples_Items
## 1       1        4486.682     17.77911      3.787671  7.351027      8.030822
## 2       2        8604.835     12.44604      6.935252  3.071942      3.244604
## 3       3        1546.830     25.37736      1.018868 15.094340     16.716981
```

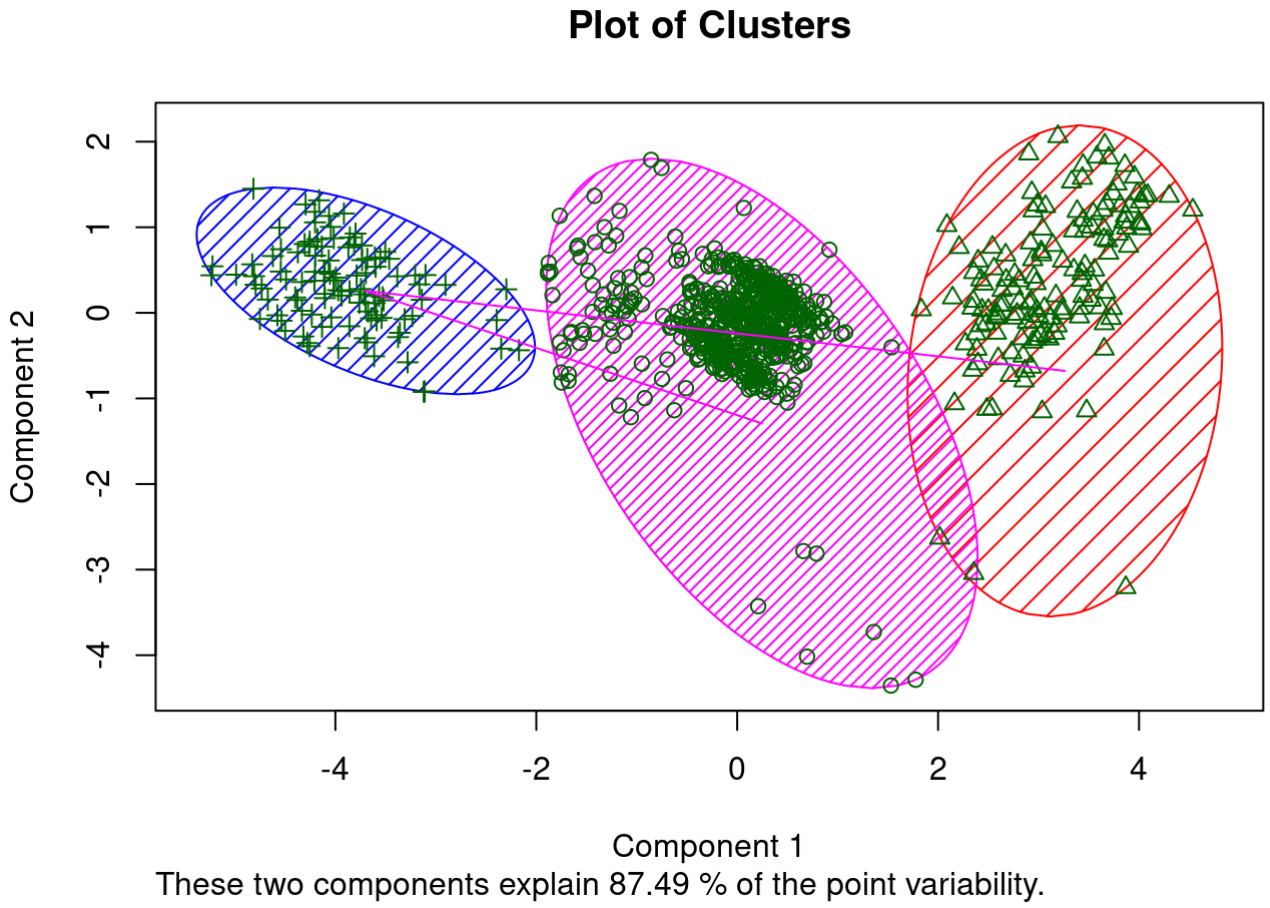
```
sil=silhouette(kmean.clus$cluster,dist(cust_spend_scaled))
head(sil[,1:3])
```

```
##      cluster neighbor sil_width
## [1,]       3        1 0.6560546
## [2,]       2        1 0.5124245
## [3,]       2        1 0.5823802
## [4,]       1        3 0.3751968
## [5,]       3        1 0.7031666
## [6,]       1        3 0.2153388
```

```
summary(sil)
```

```
## Silhouette of 829 units in 3 clusters from silhouette.default(x = kmean.clus$clust
er, dist = dist(cust_spend_scaled)) :
##  Cluster sizes and average silhouette widths:
##       584       139       106
## 0.6006331 0.5160516 0.5997410
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04131 0.55732 0.62901 0.58634 0.66882 0.72848
```

```
########################################################################
#----------- Ploting cluster plot-----------------------------#
########################################################################
clusplot(cust_spend[,3:7],kmean.clus$cluster,lines = 1,color = T,shade = T,main = 'Pl
ot of Clusters')
```



These two components explain 87.49 % of the point variability.