W2. Segmentation (Clustering)

December 18, 2018

0.1 Market Segmentation

Li Liu
12/17/2018

Case: Data of attribute importance on cars from 72 students (24 MBAs+49 undergrads)

```
In [33]: install.packages('gmodels', repos='http://cran.us.r-project.org')
         install.packages('mclust', repos='http://cran.us.r-project.org')
         install.packages('NbClust', repos='http://cran.us.r-project.org')
         install.packages('tidyverse', repos='http://cran.us.r-project.org')
         install.packages('factoextra', repos='http://cran.us.r-project.org')
         library(factoextra)
         library(tidyverse)
         library(NbClust)
         library(gmodels)
         library(mclust)
```

```
package 'gmodels' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\LI LIU\AppData\Local\Temp\RtmpWoPRGl\downloaded_packages
package 'mclust' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\LI LIU\AppData\Local\Temp\RtmpWoPRGl\downloaded_packages


Warning message:
"package 'NbClust' is in use and will not be installed"Package 'mclust' version 5.4.2
Type 'citation("mclust")' for citing this R package in publications.
```

0.1.1 Hierarchial Clustering

Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.

In [90]:
```
df<-read.csv("SegmentationData.csv",row.names=1)
head(df)
attach(df)
```

Trendy	Styling	Reliability	Sportiness	Performance	Comfort	MBA	Choice
10	20	35	5	20	10	MBA	Lexus
25	5	25	5	25	15	MBA	BMW
10	20	30	10	10	20	MBA	Lexus
10	15	30	10	20	15	MBA	BMW
20	10	40	1	14	15	MBA	Mercedes
20	30	10	20	10	10	MBA	Lexus

In [102]:
```
#Standarize raw data
stddf<-scale(df[,c("Trendy", "Styling",
                   "Reliability", "Sportiness", "Performance", "Comfort")])
#Calculate Euclidean Distance
dist<-dist(stddf,method="euclidean")
```

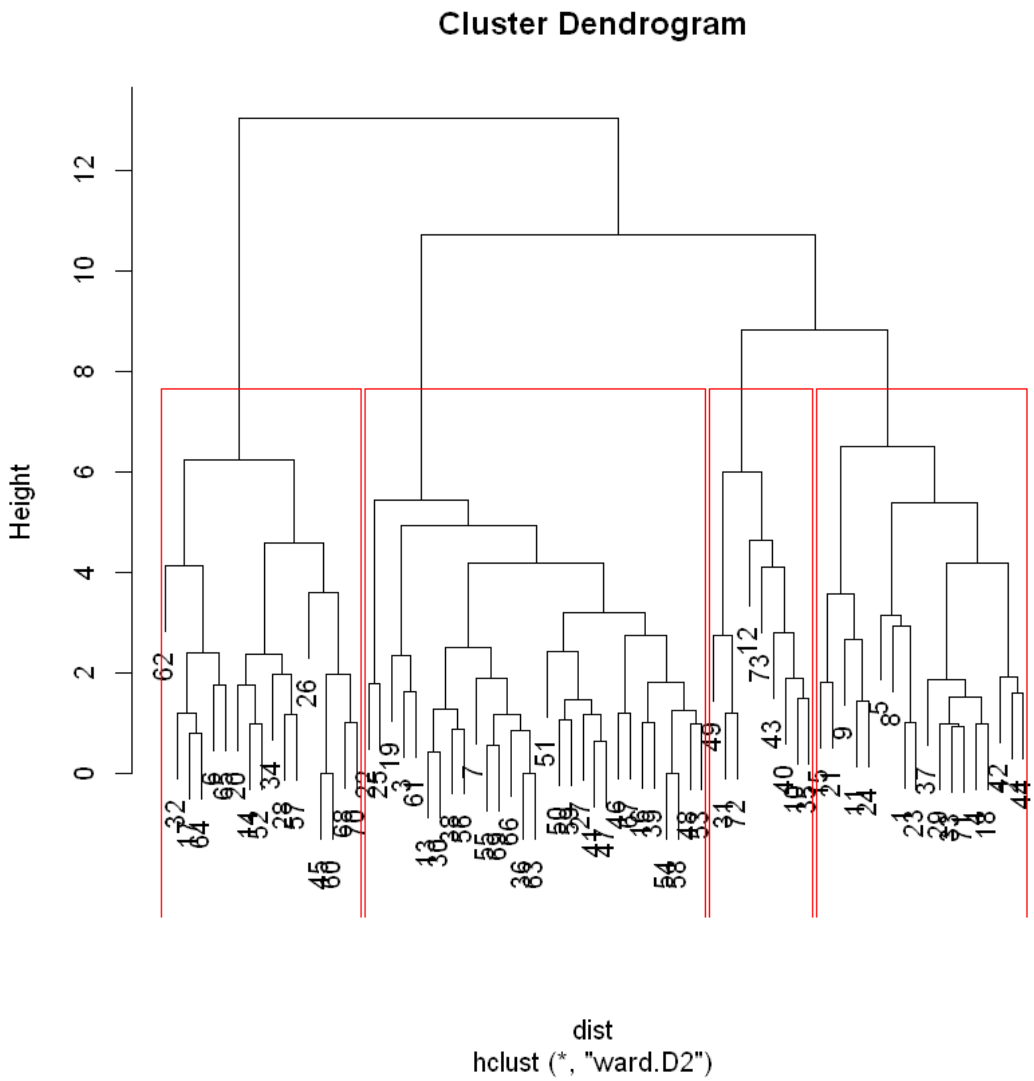
In [7]:
```
#Distance Matrix
as.matrix(dist)[1:10,1:5]
```

	1	2	3	4	5
1	0.000000	3.730216	2.802191	1.775616	2.746615
2	3.730216	0.000000	4.218662	3.017462	2.984534
3	2.802191	4.218662	0.000000	1.974683	3.331082
4	1.775616	3.017462	1.974683	0.000000	2.924141
5	2.746615	2.984534	3.331082	2.924141	0.000000
6	5.280741	5.984364	4.536636	4.783331	6.598513
7	3.589287	3.493220	3.339530	2.128324	4.765912
8	2.376538	3.128561	4.238938	2.526470	3.312364
9	4.458554	4.494805	4.619433	3.806746	5.084776
10	2.547435	3.213949	3.411743	2.175899	4.228049

In [103]:
```
#4-cluster

clust<-hclust(dist,method="ward.D2")
plot(clust)
#Cut trees
h_clust<-cutree(clust,4)
rect.hclust(clust,k=4,border='red')
table(h_clust)
hclust_summary <- aggregate(stddf[,c("Trendy", "Styling",
                                     "Reliability", "Sportiness",
                                     "Performance", "Comfort")],
                             by=list(h_clust),FUN=mean)
hclust_summary
```

```
h_clust
 1  2  3  4
18 29 17  9
```

Group.1	Trendy	Styling	Reliability	Sportiness	Performance	Comfort
1	-0.50357227	-0.6837159	1.09976574	-0.94569654	0.6548024	0.08642535
2	-0.01577854	-0.4249072	-0.28158545	0.50052114	-0.0989237	0.58621035
3	1.14725137	0.8552172	-0.65660558	0.16346240	-0.9192806	-0.69794311
4	-1.10904387	1.1211667	-0.05194561	-0.03015957	0.7455682	-0.74341374

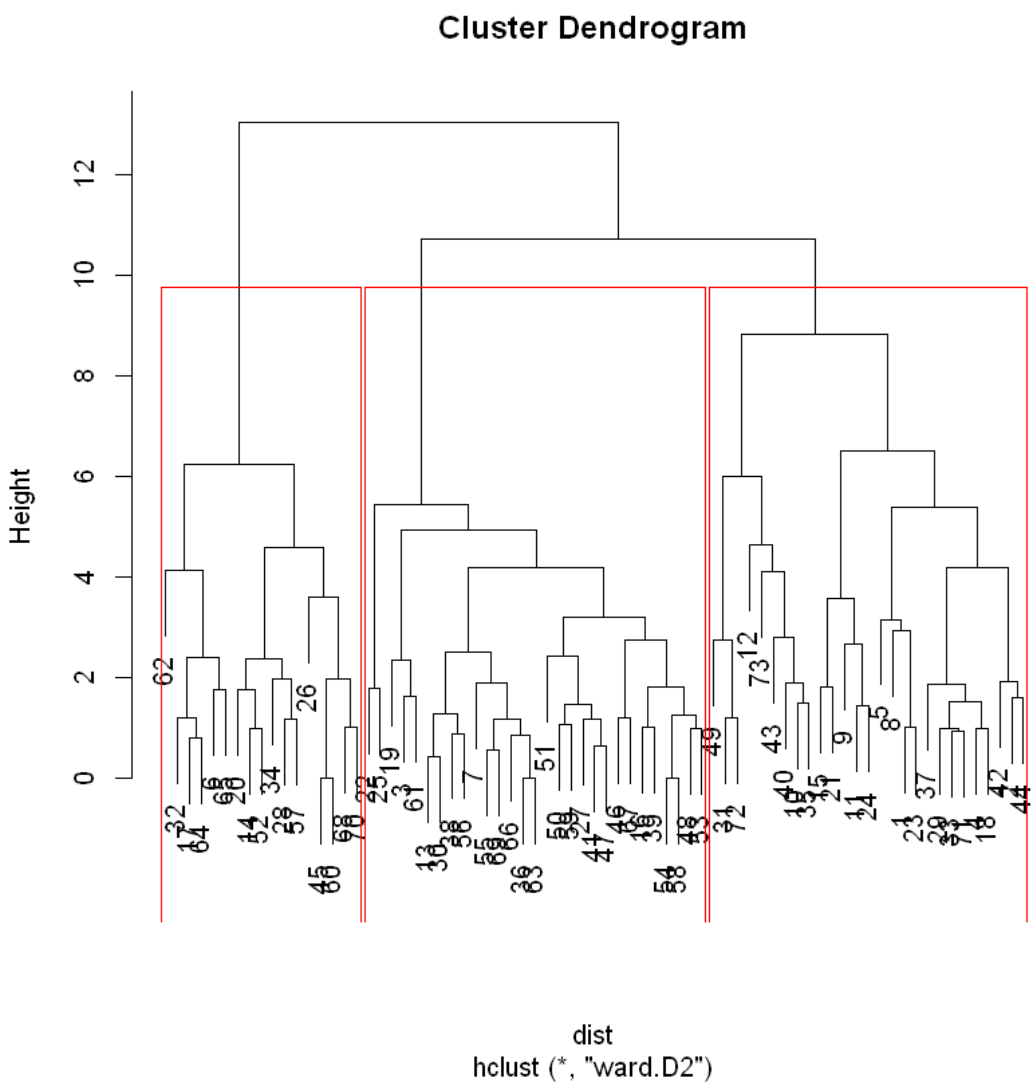


In [104]:
```
#3-cluster

clust<-hclust(dist,method="ward.D2")
plot(clust)
#Cut trees
h_clust<-cutree(clust,3)
rect.hclust(clust,k=3,border='red')
```

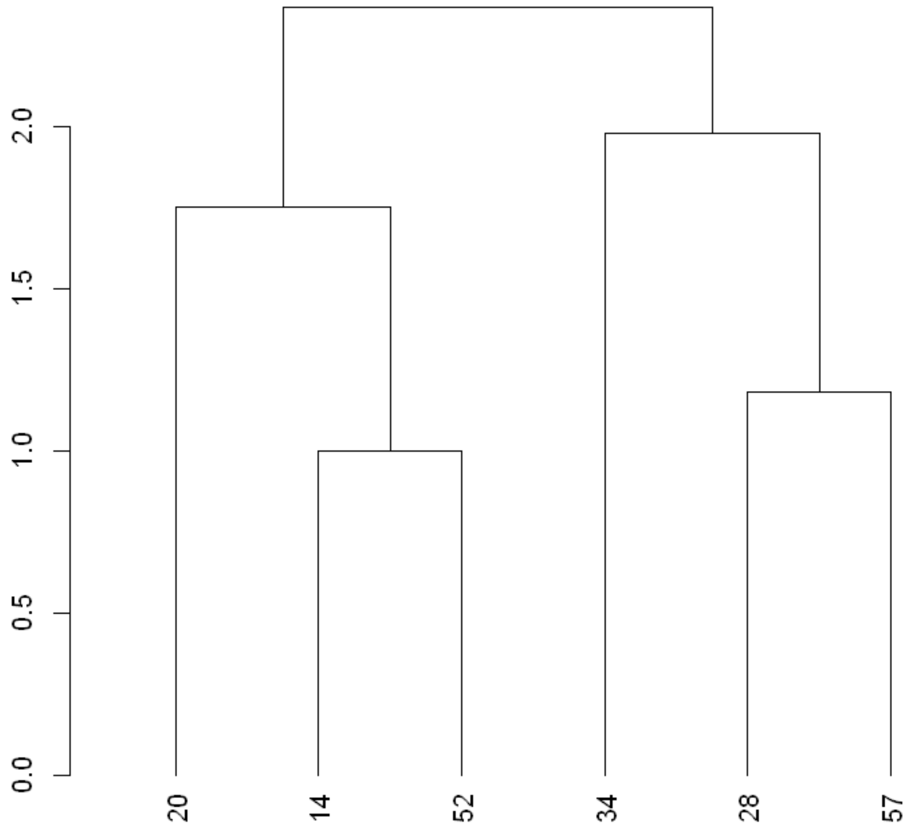
```r
table(h_clust)
hclust_summary <- aggregate(stddf[,c("Trendy", "Styling",
                                     "Reliability", "Sportiness",
                                     "Performance", "Comfort")],
                                      by=list(h_clust),FUN=mean)
hclust_summary

#Most significant factor in each cluster
#G1: Reliability; G2: Comfort; G3: Sportiness
```

```
h_clust
 1  2  3
27 29 17
```

Group.1	Trendy	Styling	Reliability	Sportiness	Performance	Comfort
1	-0.70539614	-0.08208834	0.7158620	-0.6405175	0.6850577	-0.1901877
2	-0.01577854	-0.42490717	-0.2815854	0.5005211	-0.0989237	0.5862104
3	1.14725137	0.85521724	-0.6566056	0.1634624	-0.9192806	-0.6979431



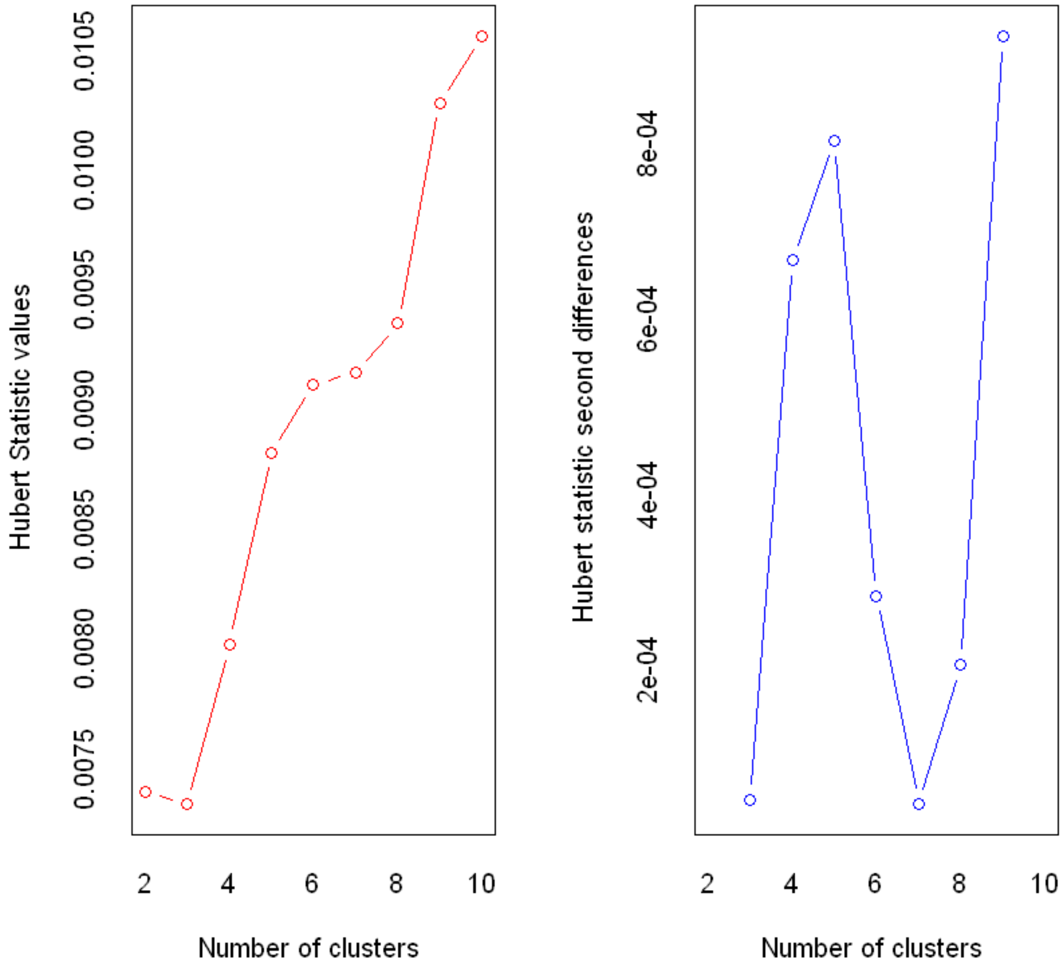
In [31]:
```
#Segment plot
plot(cut(as.dendrogram(clust),h=3)$lower[[3]])
```



NcCluster(): use 26 criteria to determine the number of clusters

```
In [32]: set.seed(1990)
         NbClust(data=stddf[,1:5],min.nc=2,max.nc=10,index='all', method="ward.D2")
```

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hu
                index second differences plot.
```




```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dir
                second differences plot) that corresponds to a significant increase of the valu
                the measure.

*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 10 proposed 3 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 3 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters
```

```
                 ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

$All.index

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedma
2	1.0799	19.4423	15.2666	-3.4128	63.8586	564035051	5038.8631	282.6111	3.1512
3	4.2356	19.1708	9.1557	-3.8247	125.3102	546888028	3415.0147	232.5975	5.4783
4	0.3189	17.2540	8.3534	-5.0953	159.1836	611300402	2593.8512	205.6938	7.0207
5	0.5079	16.3550	8.9849	-6.0343	202.7795	525669059	2134.0463	183.4809	8.8405
6	1.2239	16.3655	7.4554	-5.4291	240.3721	452300734	1641.4307	162.0669	10.4134
7	0.9987	16.1533	6.8924	-5.0795	272.0101	399115542	1309.7081	145.8389	11.7806
8	0.8637	16.0296	7.0554	-4.7518	305.5157	329419867	1063.7346	132.0490	13.4009
9	1.0446	16.1775	6.6456	-4.2941	344.2663	245198975	894.1088	119.1193	15.5956
10	0.9691	16.3520	6.6364	-3.8667	376.7922	193878168	720.2895	107.9138	17.7061

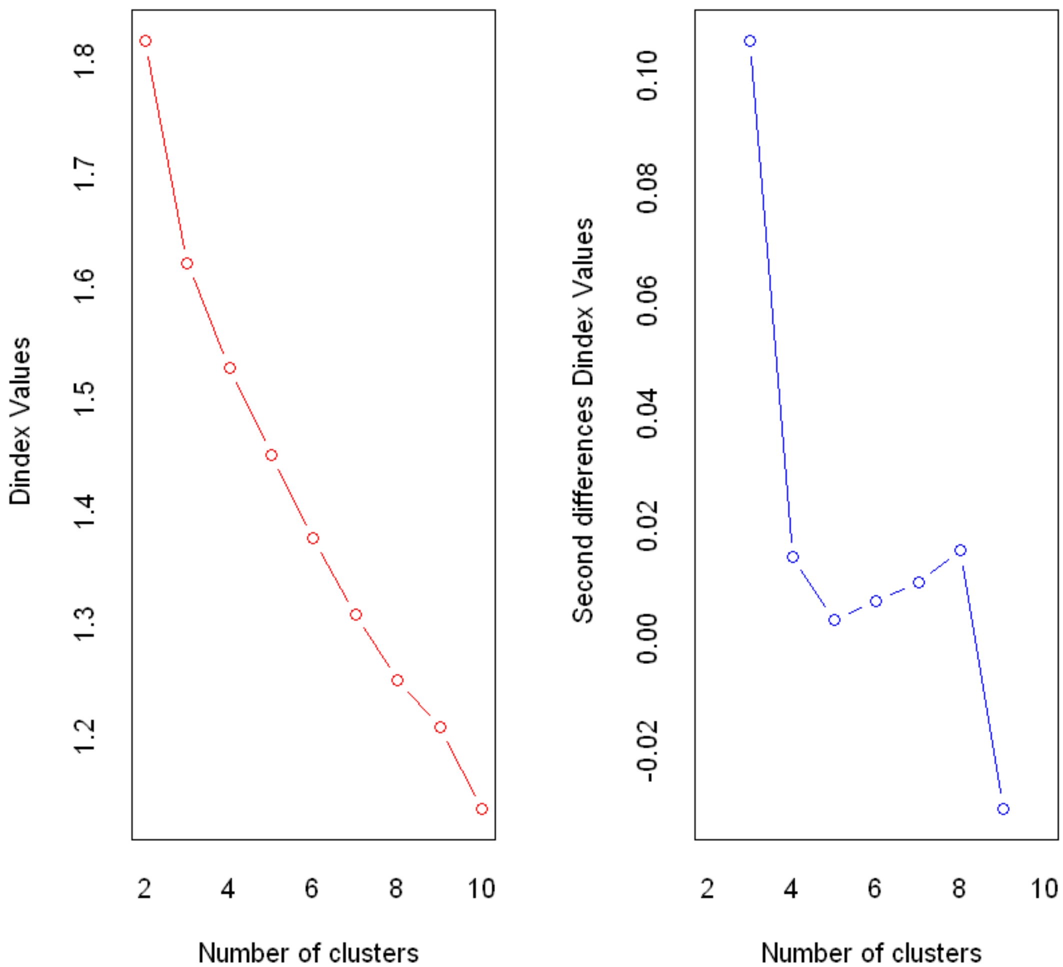
$All.CriticalValues

	CritValue_Duda	CritValue_PseudoT2	Fvalue_Beale
2	0.6251	32.3838	0.3822
3	0.4234	20.4305	0.2745
4	0.4864	22.1752	0.4062
5	0.2868	19.8896	0.0573
6	0.2552	20.4342	0.0034
7	0.3776	19.7832	0.1668
8	0.5502	25.3445	0.5019
9	0.5399	24.7090	0.3493
10	0.1164	30.3727	0.0422

$Best.nc

	KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW
Number_clusters	3.0000	2.0000	3.0000	2.0000	3.0000	3	3.000	3.0000
Value_Index	4.2356	19.4423	6.1109	-3.4128	61.4516	81559398	1623.848	23.1098

$Best.partition **1** 1 **2** 1 **3** 1 **4** 2 **5** 1 **6** 3 **7** 2 **8** 2 **9** 1 **10** 2 **11** 1 **12** 2 **13** 2 **14** 3 **15** 1 **16** 2 **17** 3 **18** 2 **19** 3 **20** 3 **21** 1 **22** 2 **23** 1 **24** 1 **25** 3 **26** 3 **27** 2 **28** 3 **29** 2 **30** 2 **31** 3 **32** 3 **33** 2 **34** 3 **35** 3 **36** 2 **37** 1 **38** 2 **39** 2 **40** 1 **41** 2 **42** 1 **43** 2 **44** 1 **45** 3 **46** 2 **47** 2 **48** 2 **49** 1 **50** 2 **51** 1 **52** 3 **53** 2 **54** 2 **55** 2 **56** 2 **57** 3 **58** 2 **59** 2 **60** 3 **61** 3 **62** 3 **63** 2 **64** 3 **65** 3 **66** 2 **67** 2 **68** 3 **69** 2 **70** 3 **71** 2 **72** 3 **73** 1



Cross Tabulation With Tests For Factor Independence link: https://www.rdocumentation.org/packages/gmodels/versions/2.18.1/topics/CrossTable CrossTable(x, y, digits=3, max.width = 5, expected=FALSE, prop.r=TRUE, prop.c=TRUE, prop.t=TRUE, prop.chisq=TRUE, chisq = FALSE, fisher=FALSE, mcnemar=FALSE, resid=FALSE, sresid=FALSE, asresid=FALSE, missing.include=FALSE, format=c("SAS","SPSS"), dnn = NULL, ...)

```
In [92]: CrossTable(df$MBA,h_clust,prop.chisq = FALSE,
                    prop.r = T, prop.c = T,prop.t = F,chisq = T)
```

```
   Cell Contents
```

```
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  73


             | h_clust
      df$MBA |         1 |         2 |         3 | Row Total |
-------------|-----------|-----------|-----------|-----------|
         MBA |        14 |         6 |         4 |        24 |
             |     0.583 |     0.250 |     0.167 |     0.329 |
             |     0.519 |     0.207 |     0.235 |           |
-------------|-----------|-----------|-----------|-----------|
   Undergrad |        13 |        23 |        13 |        49 |
             |     0.265 |     0.469 |     0.265 |     0.671 |
             |     0.481 |     0.793 |     0.765 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |        27 |        29 |        17 |        73 |
             |     0.370 |     0.397 |     0.233 |           |
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  7.03013     d.f. =  2     p =  0.02974588
```

0.1.2 K-means clustering

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"), trace=FALSE)

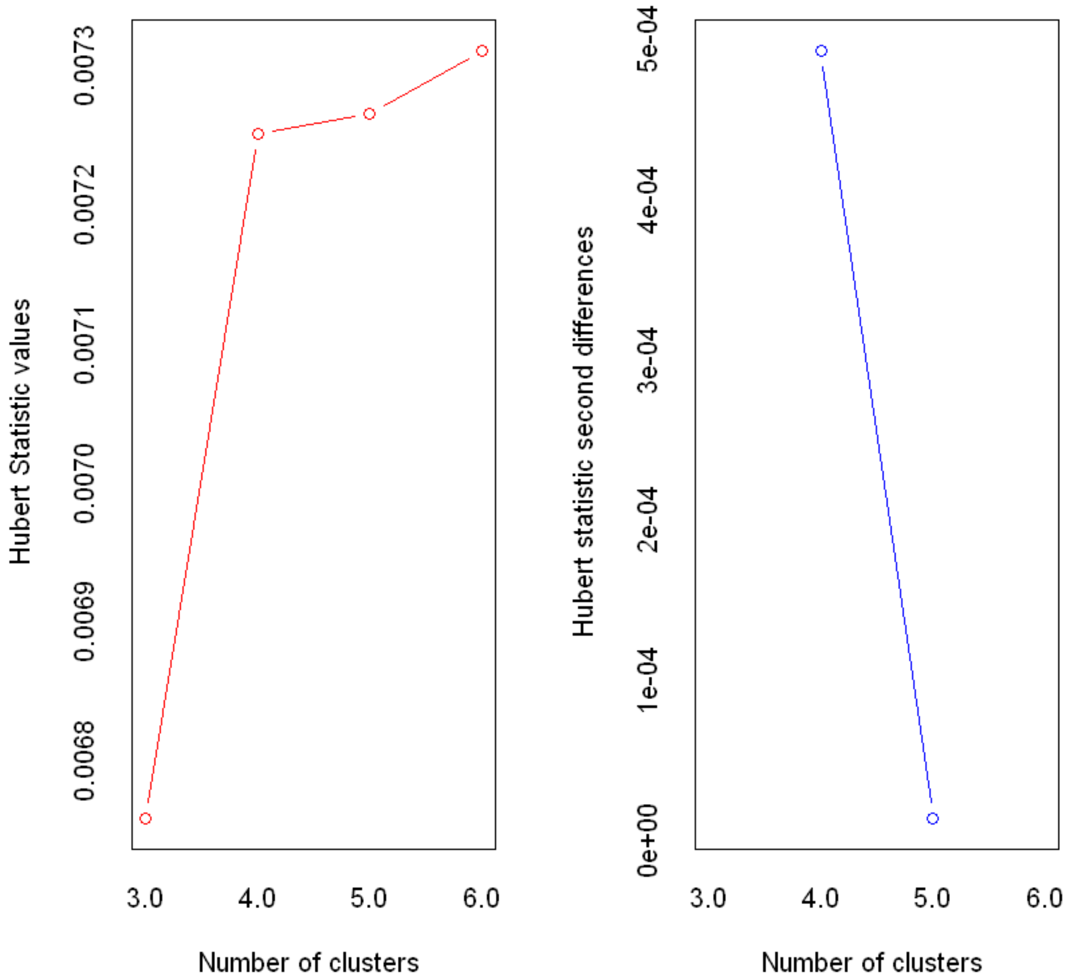
K-Means algorithm (from slides): 1. Start by randomly assigning each subject to a cluster, s=1,...,S 2. Compute the centroid of each cluster and the distance of each subject to each of the clusters centroids 3. Reassign each subject to the cluster with closest centroid 4. Repeat steps 2 and 3 until no further reassignment is possible (i.e., when the within-cluster variance is minimized)

In [42]:
```
Kclu<-kmeans(stddf,3,iter.max=100,nstart=100)
Kclu
```

```
K-means clustering with 3 clusters of sizes 18, 32, 23

Cluster means:
        Trendy    Styling Reliability Sportiness Performance     Comfort
1 -0.637247817 -0.6837159   1.1781135 -1.0328905   0.7785740  0.08642535
2 -0.003271873 -0.3788069  -0.3496669  0.4977728  -0.0445069  0.53615835
3  0.503267855  1.0621176  -0.4355087  0.1157956  -0.5473961 -0.81359668

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
 1  1  2  1  1  3  2  1  1  3  1  2  2  3  1  2  3  2  2  3  1  2  1  1  2  3
27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
 2  3  1  2  3  3  1  3  3  2  1  2  2  3  2  1  3  1  2  1  2  2  3  2  2  3
53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
 2  2  2  2  3  2  2  2  2  3  2  3  3  2  2  3  3  3  2  3  1

Within cluster sum of squares by cluster:
[1]  81.39207  83.90060 111.49649
 (between_SS / total_SS =  35.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```
In [27]: NbClust(data=stddf[,1:5],min.nc=3,max.nc=6,index='all', method="kmeans")
```

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hu
                index second differences plot.
```




```
*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Di
                second differences plot) that corresponds to a significant increase of the valu
                the measure.

*******************************************************************
* Among all indices:
* 9 proposed 3 as the best number of clusters
* 8 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 5 proposed 6 as the best number of clusters

                   ***** Conclusion *****
```

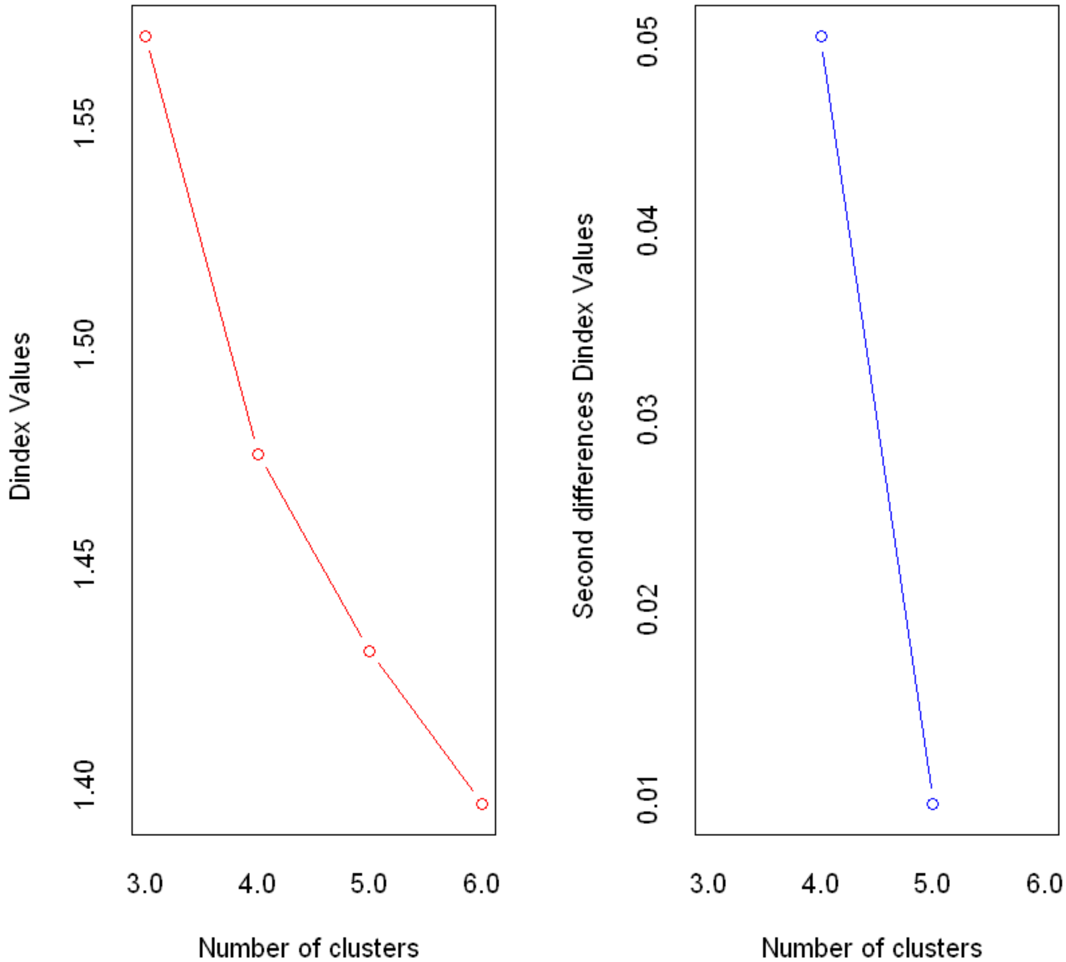
```
* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

		KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW	Friedman
	3	2.6615	23.0027	9.9985	-2.3384	161.6503	332431565	3142.612	217.2313	6.9183
$All.index	4	2.6211	20.5604	5.9101	-3.4050	206.2171	320952678	2494.735	190.0809	8.7864
	5	0.3090	17.9546	3.5125	-5.0208	211.9276	463754305	1884.086	175.0844	9.3054
	6	0.1667	15.5757	15.2717	-5.9661	250.0332	396233308	1841.278	166.4846	11.9435

		CritValue_Duda	CritValue_PseudoT2	Fvalue_Beale
	3	0.4864	23.2312	0.4623
$All.CriticalValues	4	0.2552	43.7876	1.0000
	5	0.1725	47.9798	1.0000
	6	0.4234	20.4305	1.0000

		KL	CH	Hartigan	CCC	Scott	Marriot	TrCovW	TraceW
$Best.nc	Number_clusters	3.0000	3.0000	6.0000	3.0000	4.0000	4	4.0000	4.0000
	Value_Index	2.6615	23.0027	11.7591	-2.3384	44.5668	154280514	647.8765	12.1538

$Best.partition **1** 2 **2** 2 **3** 1 **4** 2 **5** 2 **6** 1 **7** 3 **8** 2 **9** 2 **10** 3 **11** 2 **12** 3 **13** 3 **14** 1 **15** 2 **16** 3 **17** 1 **18** 3 **19** 3 **20** 1 **21** 2 **22** 3 **23** 2 **24** 2 **25** 3 **26** 1 **27** 3 **28** 1 **29** 2 **30** 3 **31** 1 **32** 1 **33** 2 **34** 1 **35** 1 **36** 3 **37** 2 **38** 3 **39** 3 **40** 3 **41** 3 **42** 2 **43** 3 **44** 2 **45** 1 **46** 2 **47** 3 **48** 3 **49** 1 **50** 3 **51** 3 **52** 1 **53** 3 **54** 3 **55** 3 **56** 3 **57** 1 **58** 3 **59** 3 **60** 1 **61** 1 **62** 1 **63** 3 **64** 1 **65** 1 **66** 3 **67** 3 **68** 1 **69** 3 **70** 1 **71** 3 **72** 1 **73** 2



Concordance between kmeans() and hclust() cluster memberships

```
In [105]: CrossTable(h_clust,Kclu$cluster,prop.chisq=FALSE,
                      prop.r=T,prop.c=T,prop.t=T,chisq=T)
```

```
Warning message in chisq.test(t, correct = FALSE, ...):
"Chi-squared approximation may be incorrect"
```

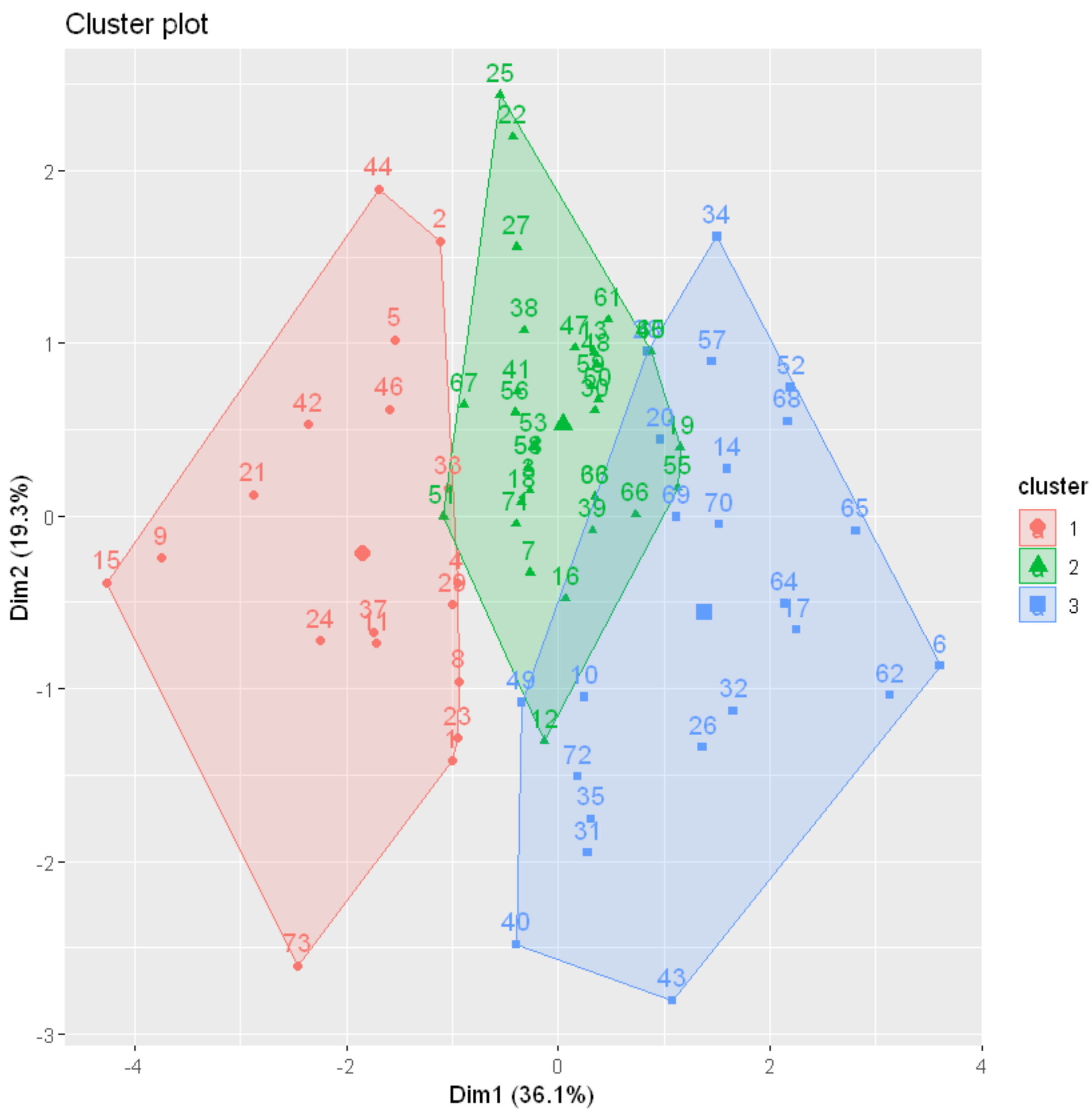
```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
```

```
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  73


             | Kclu$cluster 
     h_clust |         1 |         2 |         3 | Row Total | 
-------------|-----------|-----------|-----------|-----------|
           1 |        17 |         3 |         7 |        27 | 
             |     0.630 |     0.111 |     0.259 |     0.370 | 
             |     0.944 |     0.094 |     0.304 |           | 
             |     0.233 |     0.041 |     0.096 |           | 
-------------|-----------|-----------|-----------|-----------|
           2 |         1 |        27 |         1 |        29 | 
             |     0.034 |     0.931 |     0.034 |     0.397 | 
             |     0.056 |     0.844 |     0.043 |           | 
             |     0.014 |     0.370 |     0.014 |           | 
-------------|-----------|-----------|-----------|-----------|
           3 |         0 |         2 |        15 |        17 | 
             |     0.000 |     0.118 |     0.882 |     0.233 | 
             |     0.000 |     0.062 |     0.652 |           | 
             |     0.000 |     0.027 |     0.205 |           | 
-------------|-----------|-----------|-----------|-----------|
Column Total |        18 |        32 |        23 |        73 | 
             |     0.247 |     0.438 |     0.315 |           | 
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test 
------------------------------------------------------------
Chi^2 =  77.06958     d.f. =  4     p =  7.26997e-16
```

Visualize K-means cluster “If there are more than two dimensions (variables) fviz_cluster will perform principal component analysis (PCA) and plot the data points according to the first two principal components that explain the majority of the variance.”

link: https://uc-r.github.io/kmeans_clustering

```
In [63]: fviz_cluster(Kclu, data = stddf)
```



Disadv of K-means: 1. require to specify the clusters number 2. sensitive to outliers 3. data ordering change results

In [76]:
```
#Add clustering results as new column

df<-data.frame(stddf)
df$cluster <- Kclu$cluster
head(df)
```

Trendy	Styling	Reliability	Sportiness	Performance	Comfort	cluster
-0.6844274	0.4957124	1.7657213	-1.1636815	0.08545324	-1.10717882	1
1.4386548	-1.9167545	0.3554625	-1.1636815	0.82808260	-0.08408953	1
-0.6844274	0.4957124	1.0605919	-0.1827491	-1.39980547	0.93899976	2
-0.6844274	-0.3084433	1.0605919	-0.1827491	0.08545324	-0.08408953	1
0.7309607	-1.1125989	2.4708507	-1.9484275	-0.80570199	-0.08408953	1
0.7309607	2.1040236	-1.7599257	1.7791159	-1.39980547	-1.10717882	3

0.1.3 Latent Class Analysis

From slides:
Uses a statistical model (vs. numerical algorithm) to form clusters
Assumes that data follow a finite mixture of normal distributions
Estimates a family of models and selects the best based on the Bayesian Information Criterion(BIC)
Outputs cluster means and cluster membership for each subject
Requires a large sample size

```
In [78]: lca<-Mclust(stddf[,1:5],verbose=FALSE,modelNames = "VEE")
         summary(lca)
```

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEE (ellipsoidal, equal shape and orientation) model with 2 components:

 log.likelihood  n df       BIC       ICL
      -431.8862 73 27 -979.6149 -990.9626

Clustering table:
 1  2
47 26
```

```
In [80]: lcaclust_summary <- aggregate(stddf[,c("Trendy", "Styling", "Reliability", "Sportines
         lcaclust_summary

         #G1: Styling/Reliability
         #G2: Spoortiness/Comfort
```

Group.1	Trendy	Styling	Reliability	Sportiness	Performance	Comfort
1	0.02627812	0.2424889	0.1364223	-0.2787552	-0.04411188	-0.1929288
2	-0.04750275	-0.4383453	-0.2466095	0.5039037	0.07974071	0.3487559

```
In [101]: #The segments are identifiable: Most MBAs in cluster 1; Undergrads are clustered int
          CrossTable(df$MBA,lca$classification,prop.chisq = FALSE,
                     prop.r = T, prop.c = T,prop.t = F,chisq = T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  73


             | lca$classification 
      df$MBA |         1 |         2 | Row Total | 
-------------|-----------|-----------|-----------|
         MBA |        20 |         4 |        24 | 
             |     0.833 |     0.167 |     0.329 | 
             |     0.426 |     0.154 |           | 
-------------|-----------|-----------|-----------|
   Undergrad |        27 |        22 |        49 | 
             |     0.551 |     0.449 |     0.671 | 
             |     0.574 |     0.846 |           | 
-------------|-----------|-----------|-----------|
Column Total |        47 |        26 |        73 | 
             |     0.644 |     0.356 |           | 
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test 
------------------------------------------------------------
Chi^2 =  5.599129     d.f. =  1     p =  0.01796941 

Pearson's Chi-squared test with Yates' continuity correction 
------------------------------------------------------------
Chi^2 =  4.435671     d.f. =  1     p =  0.03519539 
```

```
In [100]: #The segments are meaningful:
          #BMV and Lexus are associated with cluster 1; Mercedes half-half.
          CrossTable(lca$classification,df$Choice,prop.chisq = FALSE,
                     prop.r = T, prop.c = T,prop.t = F,chisq = T)
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|-------------------------|


Total Observations in Table:  73


                   | df$Choice
lca$classification |       BMW |     Lexus |  Mercedes | Row Total |
-------------------|-----------|-----------|-----------|-----------|
                 1 |        21 |        17 |         9 |        47 |
                   |     0.447 |     0.362 |     0.191 |     0.644 |
                   |     0.656 |     0.773 |     0.474 |           |
-------------------|-----------|-----------|-----------|-----------|
                 2 |        11 |         5 |        10 |        26 |
                   |     0.423 |     0.192 |     0.385 |     0.356 |
                   |     0.344 |     0.227 |     0.526 |           |
-------------------|-----------|-----------|-----------|-----------|
      Column Total |        32 |        22 |        19 |        73 |
                   |     0.438 |     0.301 |     0.260 |           |
-------------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  4.014183     d.f. =  2     p =  0.134379
```