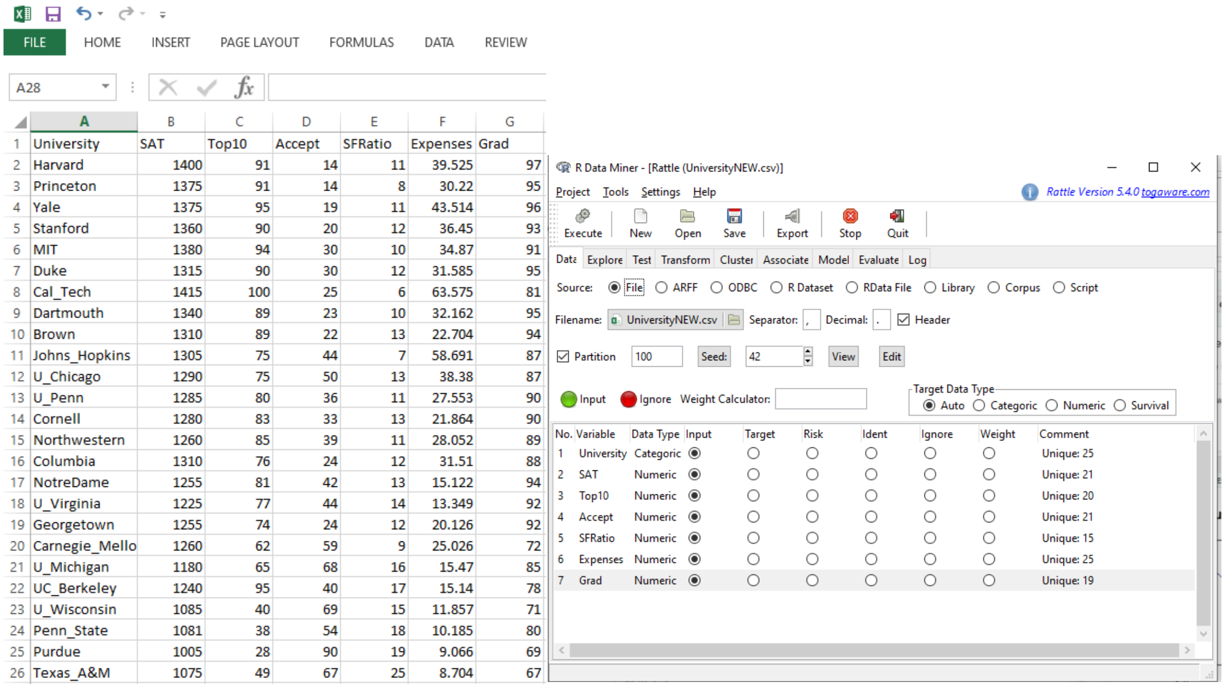
MODULE 1. PEER REVIEW ASSIGNMENT-PREDICTIVE ANALYTICS & DATA MINING
January 24th, 2021 - Aya Anisa Dwinidasari

Deliverable 1

From the University data set given, after we adjust the raw data into the processed data as mentioned on the instruction, I conducted two methods for clustering the data. After I input the csv file into Rattle, I adjust the data into 100 partition, so it will gain 42 Seed. Then I conduct the two methods of clustering. There are:



	A	B	C	D	E	F	G
1	University	SAT	Top10	Accept	SFRatio	Expenses	Grad
2	Harvard	1400	91	14	11	39.525	97
3	Princeton	1375	91	14	8	30.22	95
4	Yale	1375	95	19	11	43.514	96
5	Stanford	1360	90	20	12	36.45	93
6	MIT	1380	94	30	10	34.87	91
7	Duke	1315	90	30	12	31.585	95
8	Cal_Tech	1415	100	25	6	63.575	81
9	Dartmouth	1340	89	23	10	32.162	95
10	Brown	1310	89	22	13	22.704	94
11	Johns_Hopkins	1305	75	44	7	58.691	87
12	U_Chicago	1290	75	50	13	38.38	87
13	U_Penn	1285	80	36	11	27.553	90
14	Cornell	1280	83	33	13	21.864	90
15	Northwestern	1260	85	39	11	28.052	89
16	Columbia	1310	76	24	12	31.51	88
17	NotreDame	1255	81	42	13	15.122	94
18	U_Virginia	1225	77	44	14	13.349	92
19	Georgetown	1255	74	24	12	20.126	92
20	Carnegie_Mello	1260	62	59	9	25.026	72
21	U_Michigan	1180	65	68	16	15.47	85
22	UC_Berkeley	1240	95	40	17	15.14	78
23	U_Wisconsin	1085	40	69	15	11.857	71
24	Penn_State	1081	38	54	18	10.185	80
25	Purdue	1005	28	90	19	9.066	69
26	Texas_A&M	1075	49	67	25	8.704	67

R Data Miner - [Rattle (UniversityNEW.csv)]

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: UniversityNEW.csv Separator: , Decimal: . ☑ Header

☑ Partition 100 Seed: 42 View Edit

● Input ● Ignore Weight Calculator:

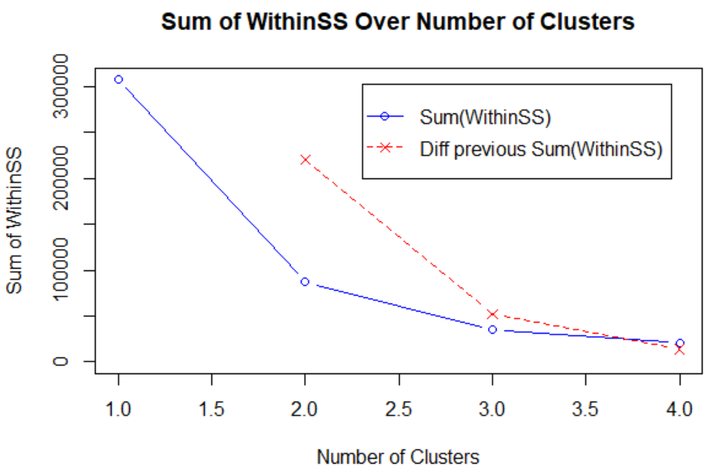
Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

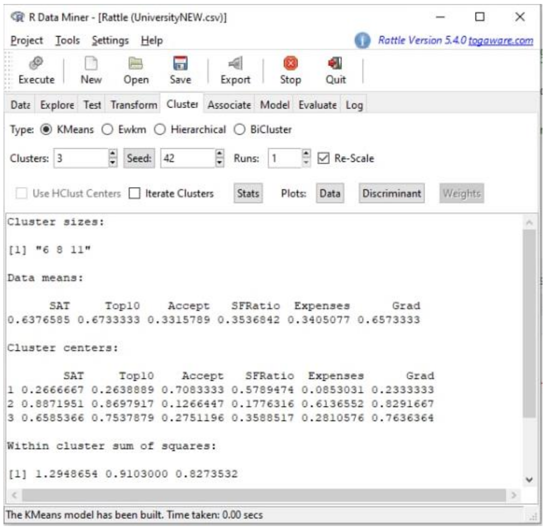
No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	University	Categoric	●	○	○	○	○	○	Unique: 25
2	SAT	Numeric	●	○	○	○	○	○	Unique: 21
3	Top10	Numeric	●	○	○	○	○	○	Unique: 20
4	Accept	Numeric	●	○	○	○	○	○	Unique: 21
5	SFRatio	Numeric	●	○	○	○	○	○	Unique: 15
6	Expenses	Numeric	●	○	○	○	○	○	Unique: 25
7	Grad	Numeric	●	○	○	○	○	○	Unique: 19

Deliverable 2

1. Partitional Clustering (K-means Cluster Method)

From this method, first I iterate the cluster into 10 and I got the elbow plot in the figure below. The intersection between red line and blue line is located on the 3.6 horizontal axis, so we conclude that the number of clusters is approximately 3.6. After that, I input the 3.6 cluster without iteration so we got three within cluster sum of squares: **1.2948654, 0.9103000, and 0.8273532.**





Type: ● KMeans ○ Ewkm ○ Hierarchical ○ BiCluster

Clusters: 3 Seed: 42 Runs: 1 ☑ Re-Scale

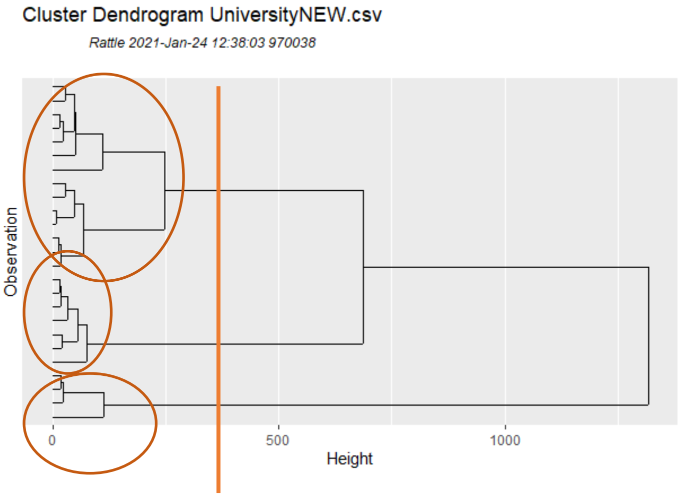
□ Use HClust Centers □ Iterate Clusters Plots:

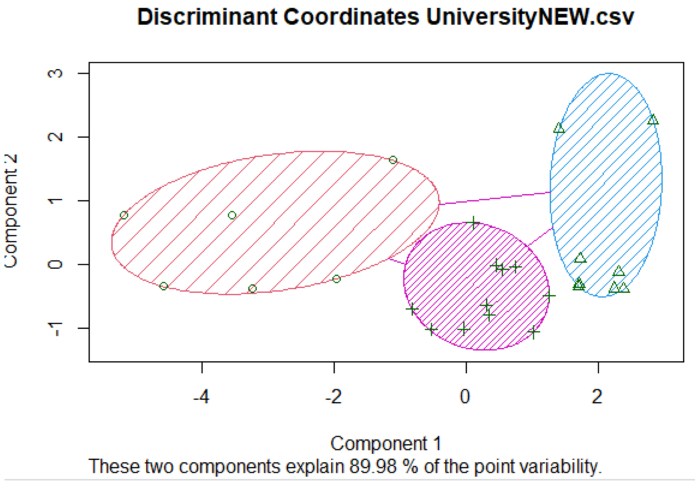
```
Cluster sizes:

[1] "6 8 11"

Data means:

      SAT     Top10    Accept   SFRatio  Expenses      Grad
0.6376585 0.6733333 0.3315789 0.3536842 0.3405077 0.6573333

Cluster centers:

        SAT     Top10    Accept   SFRatio  Expenses      Grad
1 0.2666667 0.2638889 0.7083333 0.5789474 0.0853031 0.2333333
2 0.8871951 0.8697917 0.1266447 0.1776316 0.6136552 0.8291667
3 0.6585366 0.7537879 0.2751196 0.3588517 0.2810576 0.7636364

Within cluster sum of squares:

[1] 1.2948654 0.9103000 0.8273532
```

The KMeans model has been built. Time taken: 0.00 secs

2. Hierarchical Method (dendogram method).

   The second method I used is hierarchical method from discriminant and dendogram. From this method, it's shown that there are 3 clusters defined from the dendogram and from the K-means I got in the partitional methods on the step one. As we see from the discriminant coordinates with 89.98% point variability. The comparison between K-means clustering is there is an ambiguity in the dendogram, but after we compare the discriminant plot we can see clearly the cluster size so I can cut the dendogram cluster with vertical line to make the clustering clearer.





Deliverable 3

The challenge that I might find is the comparison between the two methods mightly give slight different result. As the matter of fact the K-means clustering in partitional methods need an insight of the data overall, but the dendogram methods can give a more adjusted result in clustering as well it can be helped by the discriminant plot. My finding is, from the University data set above, there are 3 cluster of “class” University from the distribution of the all variable (SAT, top 10, Accept, SF Ratio, Expenses)