CIDM 6355 Data Mining Methods LA5 Instruction & Template

(40 points in total; Due 11:59 PM CDT, October 27, 2024)

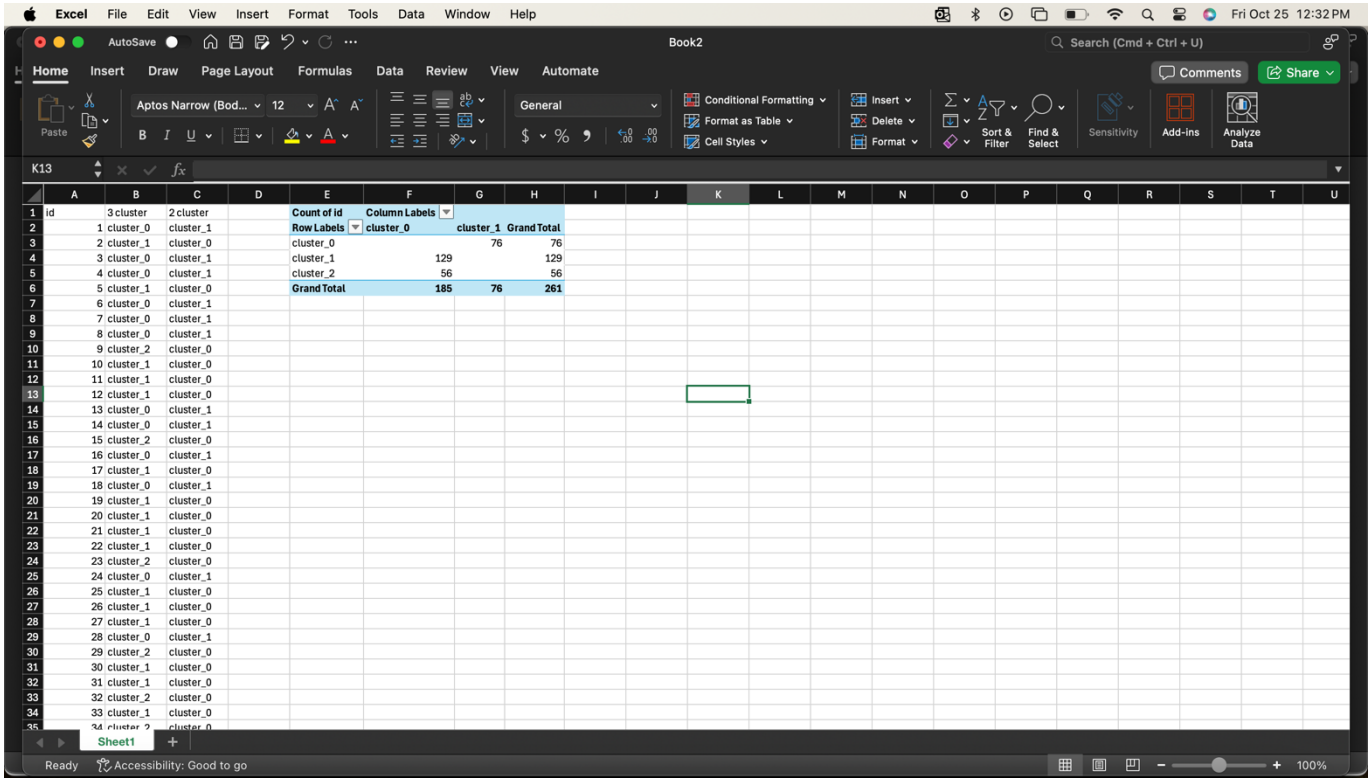
Requirements: This learning activity is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else before the due time. Any question about the learning activity should be addressed to the instructor. You are required to follow the instruction to complete all the questions and deliverables. This is an individual learning activity, so sharing your RM processes, R scripts, screenshots, or answers with other students or parties is considered as cheating, which will be reported to the university authority. In addition, it is your responsibility to make your answers meet the required format; otherwise, you might lose points because of wrong format. Screenshots without date and time can only receive up to 50% of points. Please read, understand, and comply with these requirements in this homework assignment by typing your name as below.

Your name: Sera Hill

Please go over the Lab Instruction before you answer the following questions. **Please DONOT change the question number**.

Part 1: Please submit your deliverables and answer questions required in Class 09 RM Lab (14 points).

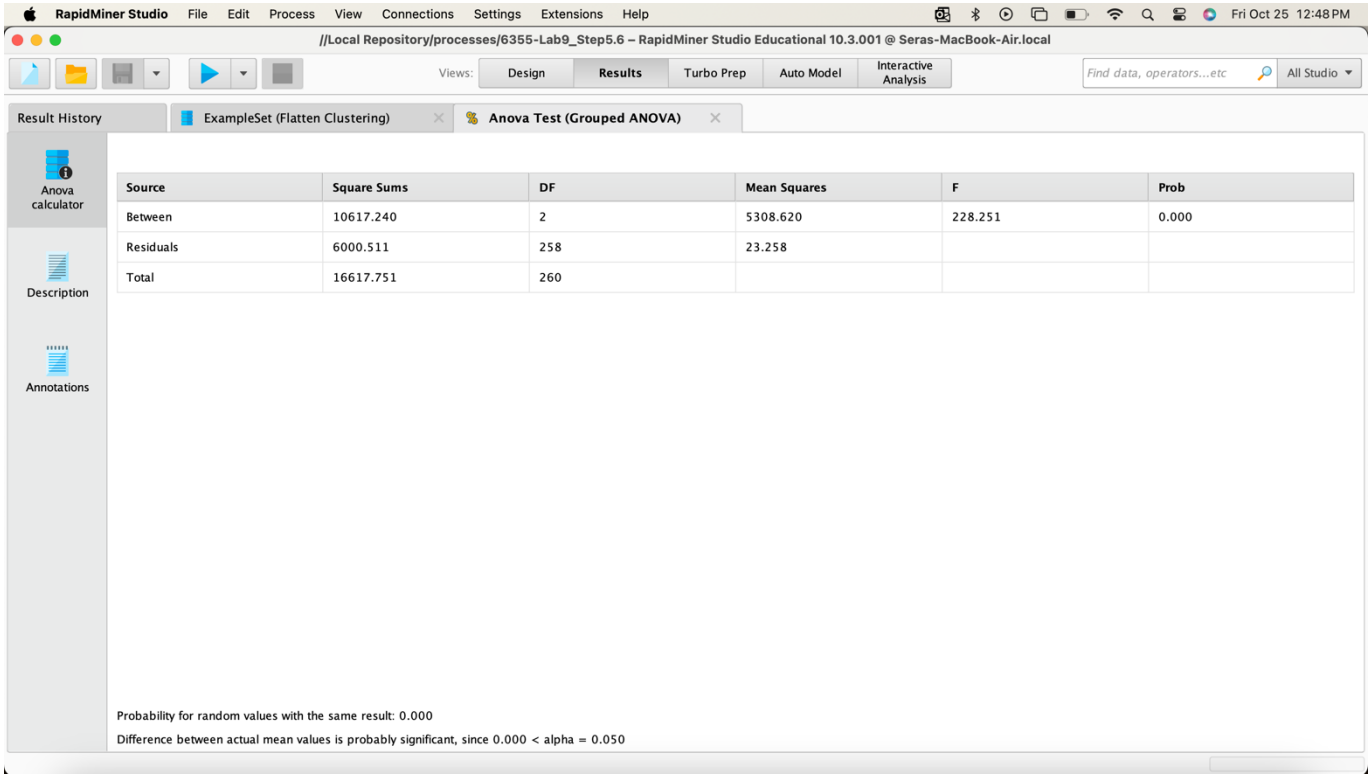
[1] Step 4.12. Empirical Examination: Take a screenshot of your PivotTable for the empirical examination with date and time (Screenshot 1). What conclusion can you make based on the PivotTable? (3 pts for your screenshot and 4 pts for your answer).



Count of id	Column Labels		
Row Labels	cluster_0	cluster_1	Grand Total
cluster_0		76	76
cluster_1	129		129
cluster_2	56		56
Grand Total	185	76	261

There is a clear segmentation within the clusters, that have an imbalance between them.

[2] Step 5.9. Take a screenshot of the ANOVA Test table with date and time (Screenshot 2). Based on the ANOVA table, do you think the mean mpg of the three clusters differ at the 95% confidence level? Why? (3 pts for your screenshot and 4 pts for your answer).

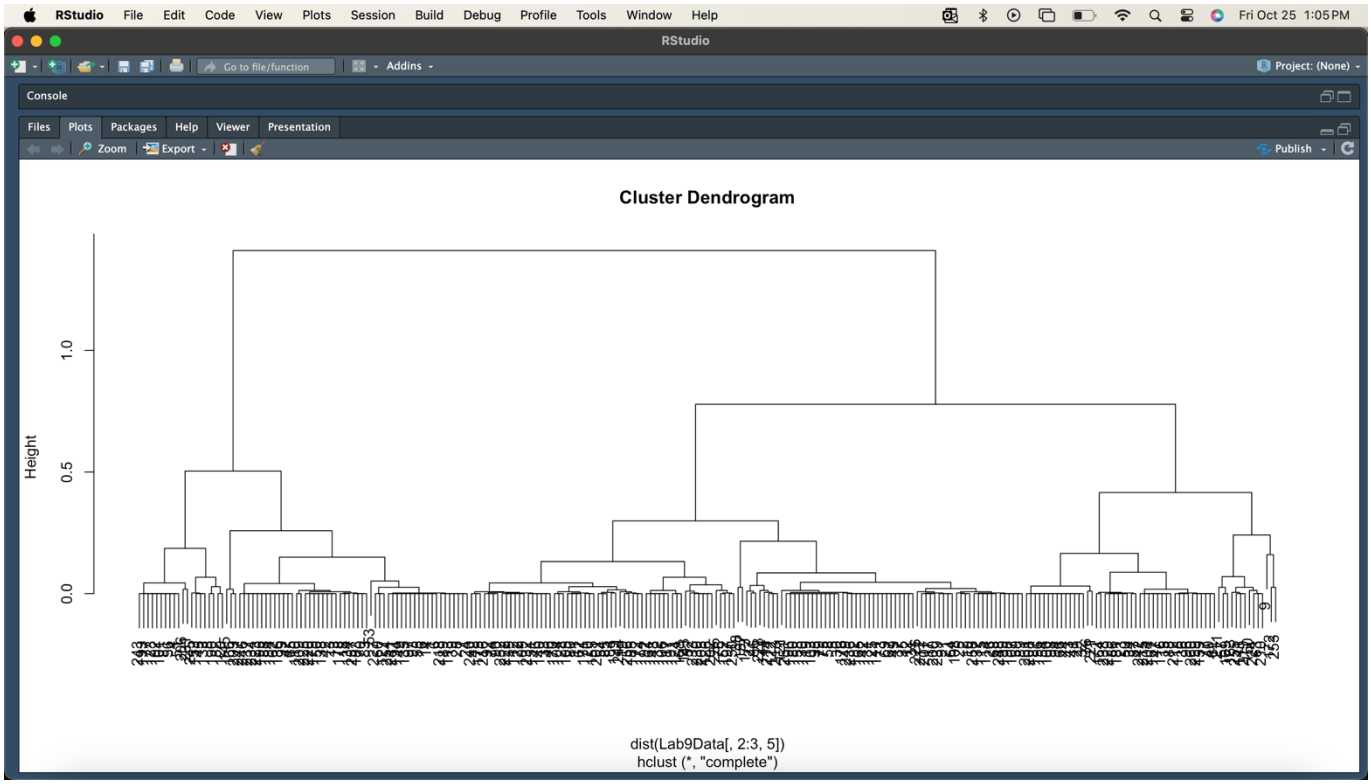


Source	Square Sums	DF	Mean Squares	F	Prob
Between	10617.240	2	5308.620	228.251	0.000
Residuals	6000.511	258	23.258		
Total	16617.751	260			

Because the p-value is lower than our alpha of .05, we can reject the null hypothesis. There is a significant difference in the average MPG between the three clusters.

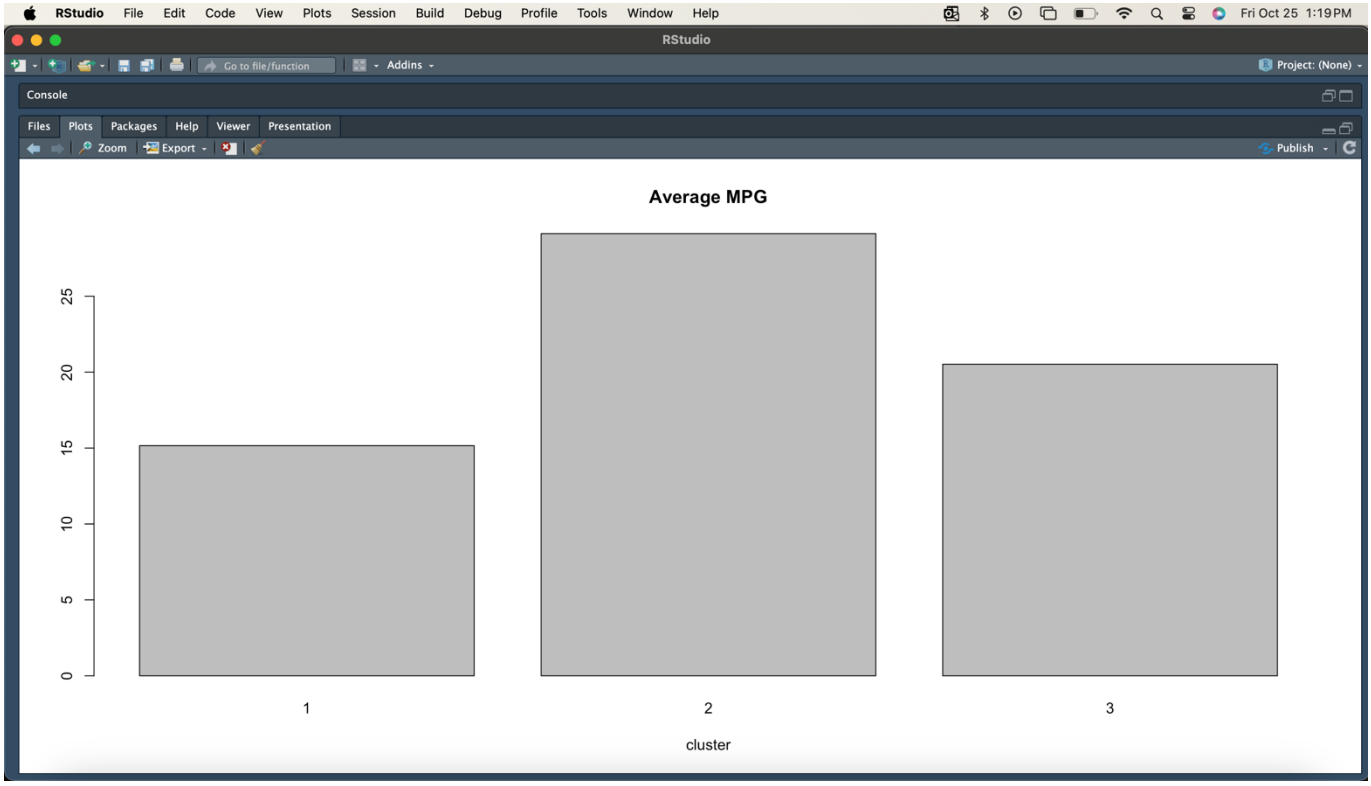
Part 2: Please submit your deliverables and answer questions required in Class 09 R Lab (26 points).

[3] Deliverable R1: take a screenshot of the dendrogram with date and time. Compare it with the one generated in RM and find at least two differences (3 pts for your screenshot and 4 pts for your answer).



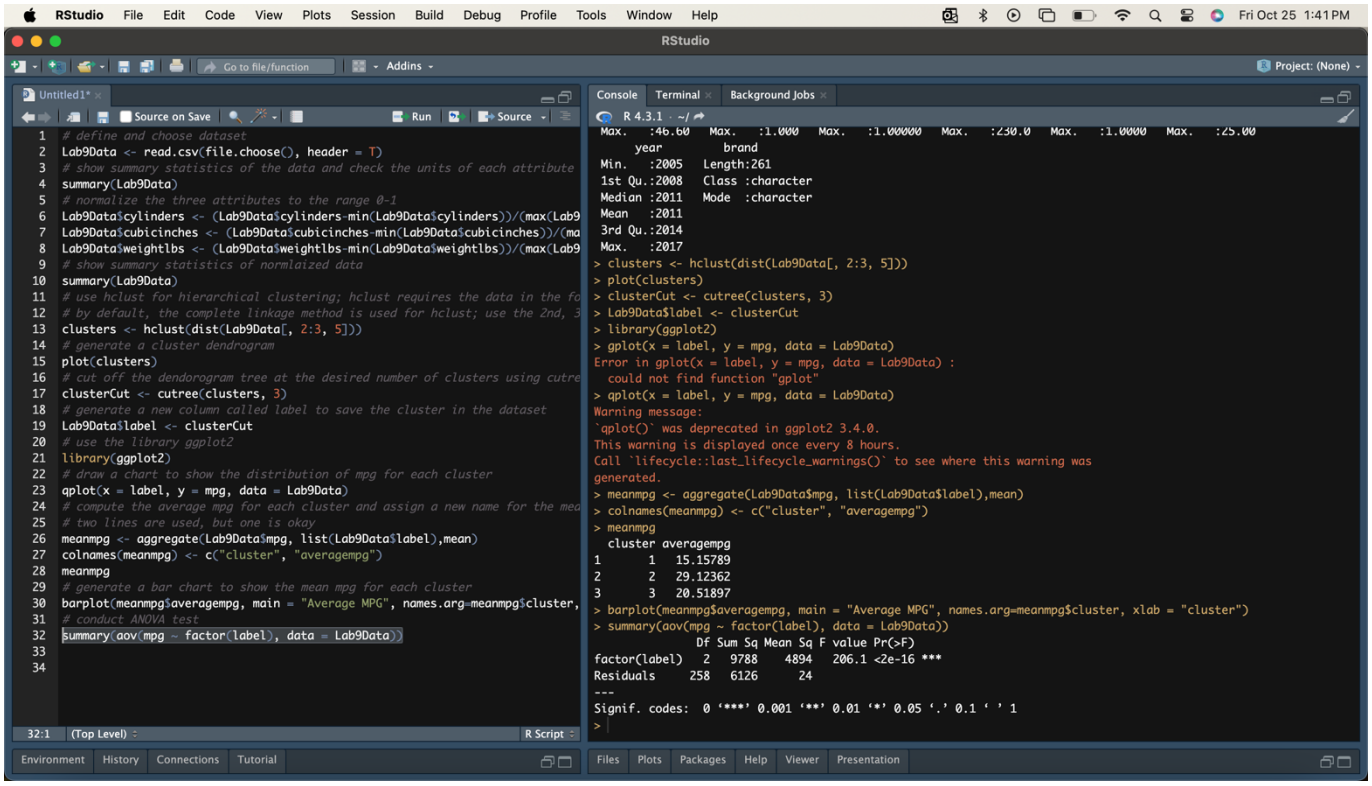
The dendrogram that was generated in RM has more branches than the one from R. The distances between the clusters vary between the dendrograms generated from RM and R.

[4] Deliverable R2: take a screenshot of the chart with date and time and describe it briefly (3 pts for your screenshot and 3 pts for your answer).



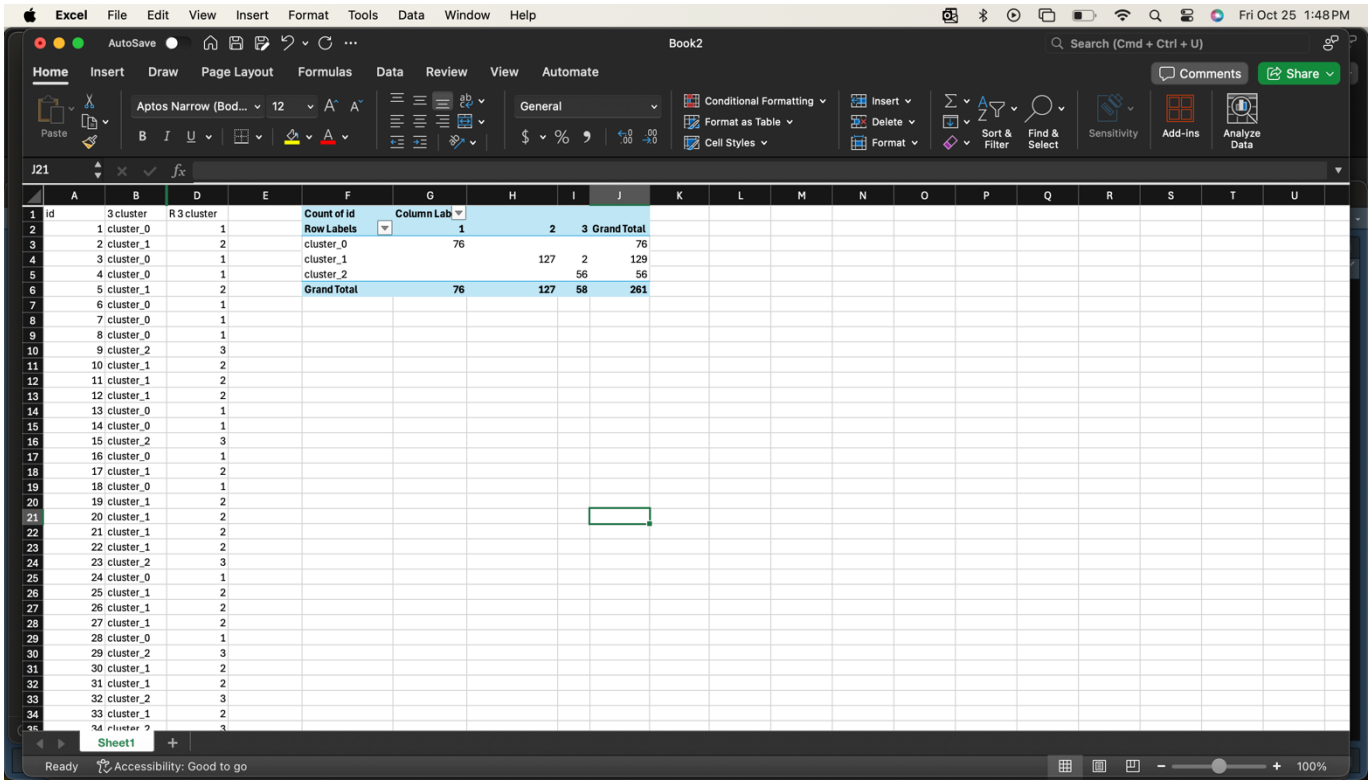
Based on the bar chart above, we can see that each cluster has a different average MPG with cluster 2 having the highest average MPG and cluster 1 having the lowest.

[5] Deliverable R3: take a screenshot of the ANOVA result with date and time and make your conclusion (3 pts for your screenshot and 3 pts for your answer).



As with the ANOVA chart generated in RM, this one generated in R shows a significant p-value and we can conclude that the mean MPG are significantly different between each cluster.

[6] Deliverable R4: save the cluster result in a csv file and then compare it with the cluster result (3-cluster model) generated at Step 4.8 in the RapidMiner lab. Are they the same? Include the screenshot of your PivotTable with date and time. Follow the same procedure we used for deliverable R4 in Class 08 R Lab. (3 pts for your screenshot and 4 pts for your answer).



id	3 cluster	R 3 cluster
1	cluster_0	1
2	cluster_1	2
3	cluster_0	1
4	cluster_0	1
5	cluster_1	2
6	cluster_0	1
7	cluster_0	1
8	cluster_0	1
9	cluster_2	3
10	cluster_1	2
11	cluster_1	2
12	cluster_1	2
13	cluster_0	1
14	cluster_0	1
15	cluster_2	3
16	cluster_0	1
17	cluster_1	2
18	cluster_0	1
19	cluster_1	2
20	cluster_1	2
21	cluster_1	2
22	cluster_1	2
23	cluster_2	3
24	cluster_0	1
25	cluster_1	2
26	cluster_1	2
27	cluster_1	2
28	cluster_0	1
29	cluster_2	3
30	cluster_1	2
31	cluster_1	2
32	cluster_2	3
33	cluster_1	2
34	cluster_1	2
35	cluster_2	3

Count of id	Column Labi			
Row Labels	1	2	3	Grand Total
cluster_0	76			76
cluster_1		127	2	129
cluster_2			56	56
Grand Total	76	127	58	261

There are slight differences between the 3 clusters generated between RM and R. Cluster 2 in RM has 129 data points while the second cluster in R has 127. Cluster 2 has 56 data points while the third cluster in R has 58.