

# 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).**

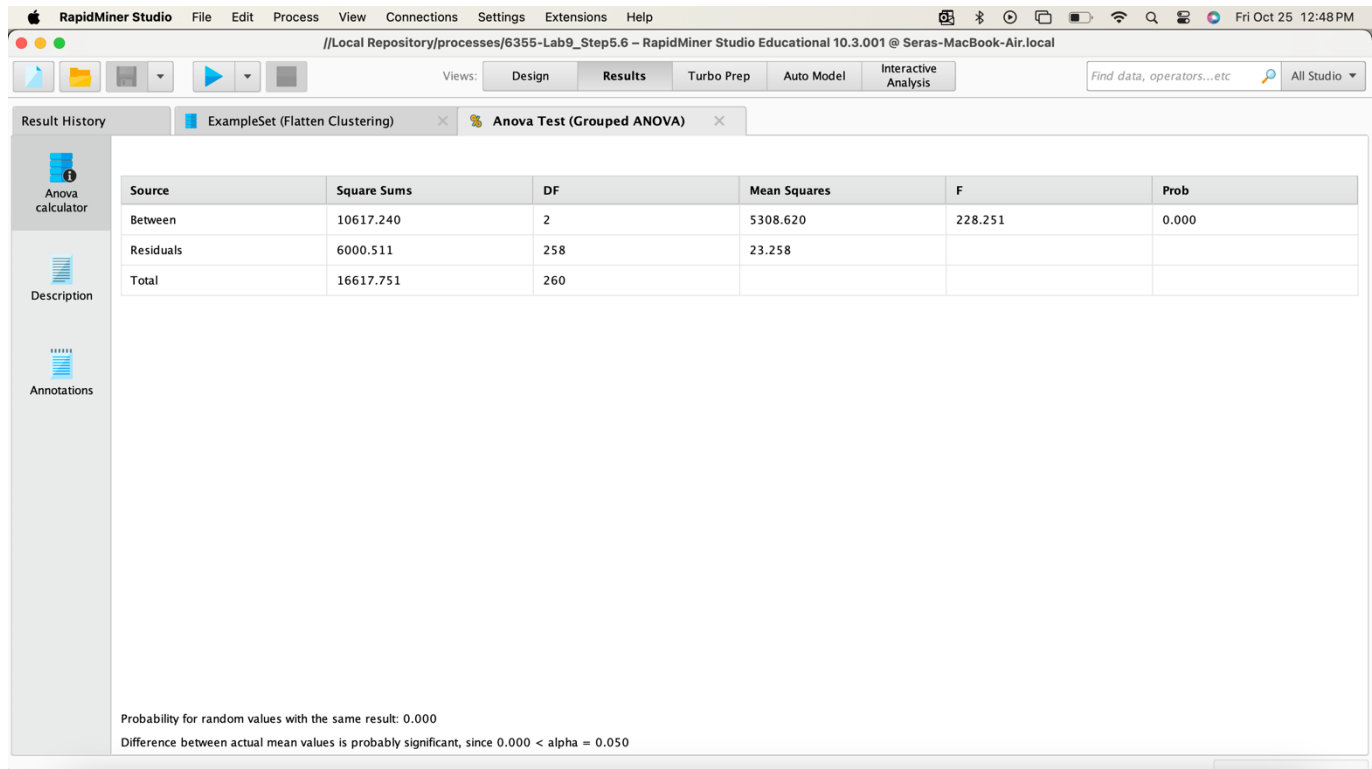
- [1] Step 4.12. Empirical Examination: Take a screenshot of your PivotTable for the empirical examination with date and time (Screenshot 1). What conclusion can you make based on the PivotTable? (3 pts for your screenshot and 4 pts for your answer).

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable is set to show 'Count of id' for 'Row Labels' and 'Column Labels'. The data is summarized as follows:

Row Labels	cluster_0	cluster_1	Grand Total
cluster_0	76	76	152
cluster_1	129	56	185
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).



The screenshot shows the RapidMiner Studio interface with the 'Results' tab selected. The main window displays the 'Anova Test (Grouped ANOVA)' results table. The table has six columns: Source, Square Sums, DF, Mean Squares, F, and Prob. The rows are: Between (10617.240, 2, 5308.620, 228.251, 0.000), Residuals (6000.511, 258, 23.258), and Total (16617.751, 260). The p-value is 0.000, which is less than the alpha level of 0.05, indicating a significant difference between the groups.

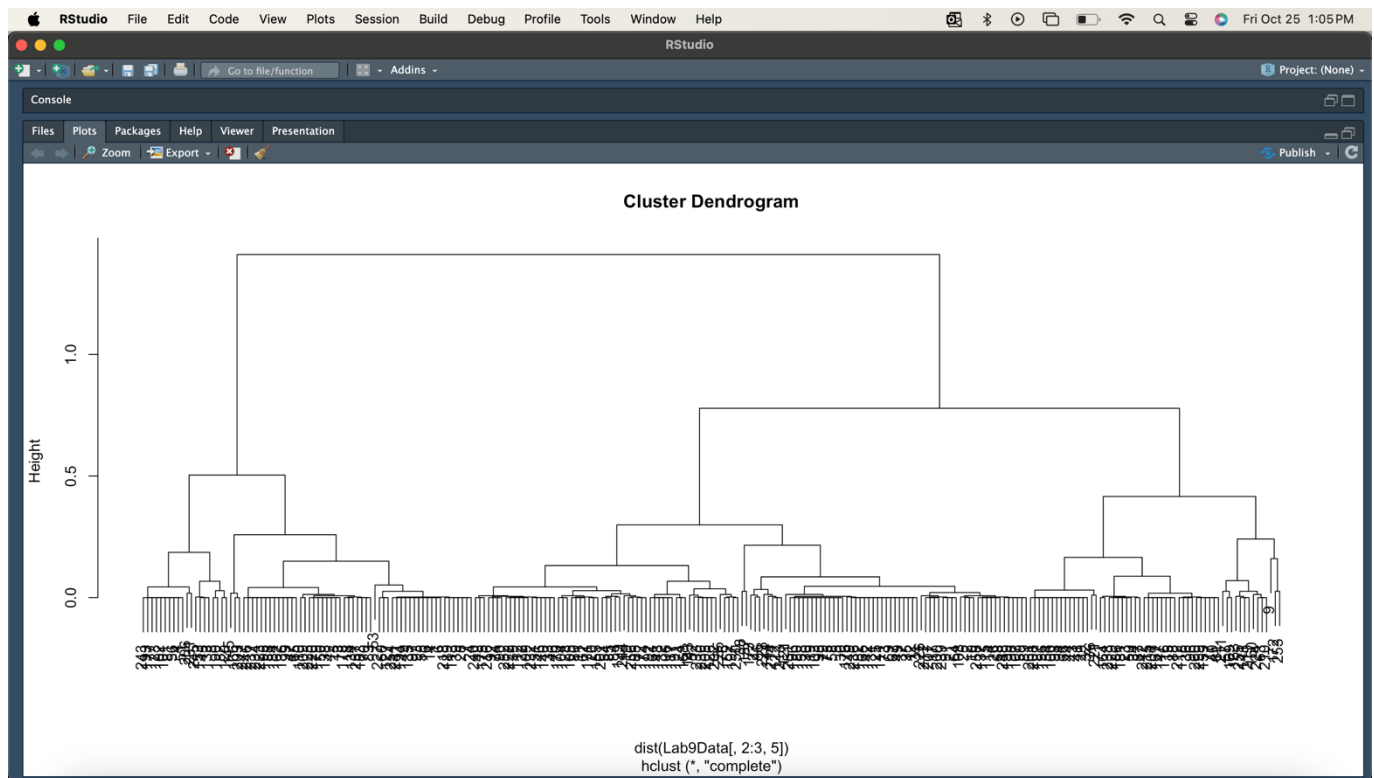
Source	Square Sums	DF	Mean Squares	F	Prob
Between	10617.240	2	5308.620	228.251	0.000
Residuals	6000.511	258	23.258		
Total	16617.751	260			

Probability for random values with the same result: 0.000  
Difference between actual mean values is probably significant, since  $0.000 < \alpha = 0.050$

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.

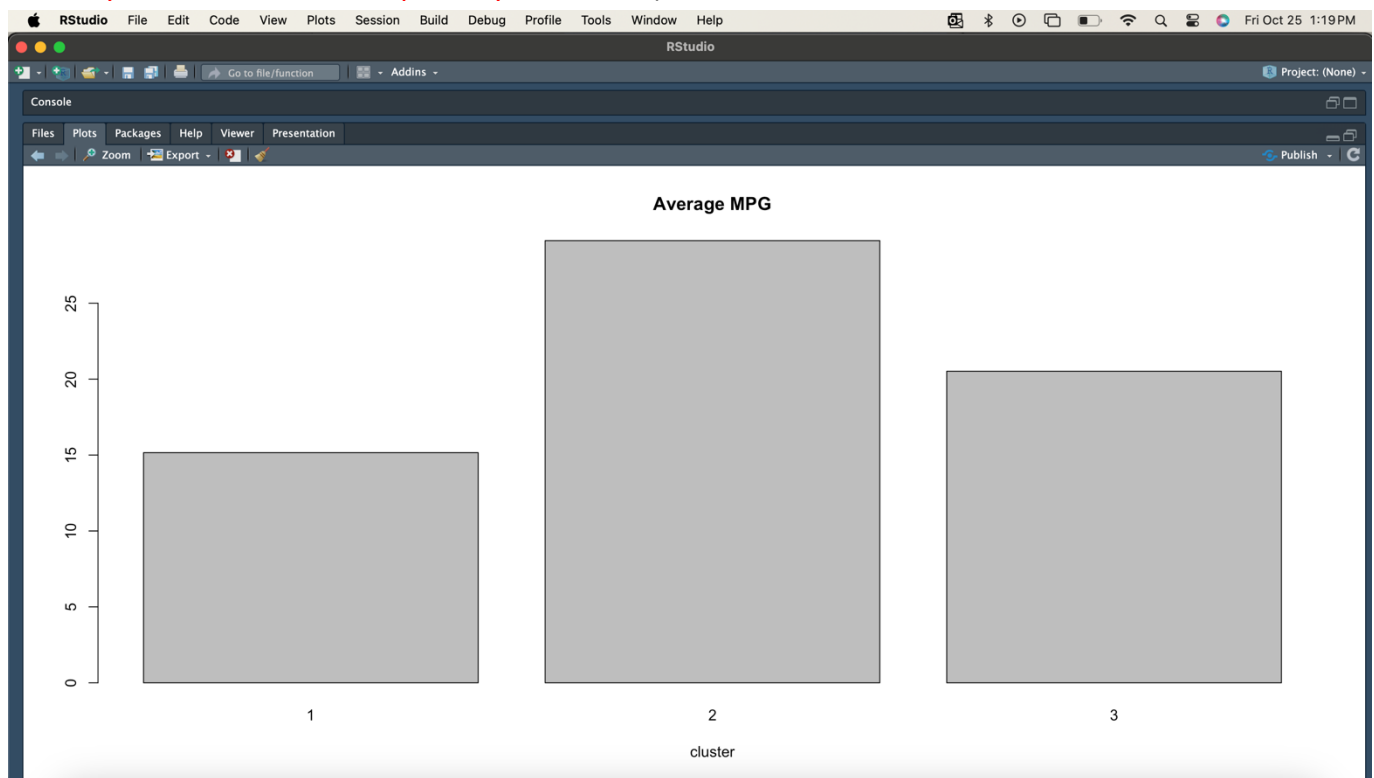
**Part 2: Please submit your deliverables and answer questions required in Class 09 R Lab (26 points).**

- [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).

```

1 # define and choose dataset
2 Lab9Data <- read.csv(file.choose(), header = T)
3 # show summary statistics of the data and check the units of each attribute
4 summary(Lab9Data)
5 # normalize the three attributes to the range 0-1
6 Lab9Data$cylinders <- (Lab9Data$cylinders - min(Lab9Data$cylinders)) / (max(Lab9
7 Lab9Data$cubicinches <- (Lab9Data$cubicinches - min(Lab9Data$cubicinches)) / (ma
8 Lab9Data$weightlbs <- (Lab9Data$weightlbs - min(Lab9Data$weightlbs)) / (max(Lab9
9 # show summary statistics of normalized data
10 summary(Lab9Data)
11 # use hclust for hierarchical clustering; hclust requires the data in the fo
12 # by default, the complete linkage method is used for hclust; use the 2nd, 3
13 clusters <- hclust(dist(Lab9Data[, 2:3, 5]))
14 # generate a cluster dendrogram
15 plot(clusters)
16 # cut off the dendrogram tree at the desired number of clusters using cutre
17 clusterCut <- cutree(clusters, 3)
18 # generate a new column called label to save the cluster in the dataset
19 Lab9Data$label <- clusterCut
20 # use the library ggplot2
21 library(ggplot2)
22 # draw a chart to show the distribution of mpg for each cluster
23 gplot(X = label, y = mpg, data = Lab9Data)
24 # compute the average mpg for each cluster and assign a new name for the mea
25 # two lines are used, but one is okay
26 meanmpg <- aggregate(Lab9Data$mpg, list(Lab9Data$label), mean)
27 colnames(meanmpg) <- c("cluster", "averagempg")
28 meanmpg
29 # generate a bar chart to show the mean mpg for each cluster
30 barplot(meanmpg$averagempg, main = "Average MPG", names.arg = meanmpg$cluster,
31 # conduct ANOVA test
32 summary(aov(mpg ~ factor(label), data = Lab9Data))
33
34

```

```

R 4.3.1 ~ /
Max. :46.60 Max. :1.0000 Max. :1.000000 Max. :230.0 Max. :1.00000 Max. :25.00
year brand
Min. :2005 Length:261
1st Qu.:2008 Class :character
Median :2011 Mode :character
Mean :2011
3rd Qu.:2014
Max. :2017
> clusters <- hclust(dist(Lab9Data[, 2:3, 5]))
> plot(clusters)
> clusterCut <- cutree(clusters, 3)
> Lab9Data$label <- clusterCut
> library(ggplot2)
> gplot(X = label, y = mpg, data = Lab9Data)
Error in gplot(X = label, y = mpg, data = Lab9Data) :
could not find function "gplot"
> gplot(X = label, y = mpg, data = Lab9Data)
Warning message:
'gplot()' was deprecated in ggplot2 3.4.0.
This warning is displayed once every 8 hours.
Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
generated.
> meanmpg <- aggregate(Lab9Data$mpg, list(Lab9Data$label), mean)
> colnames(meanmpg) <- c("cluster", "averagempg")
> meanmpg
 cluster averagempg
1 1 15.15789
2 2 29.12362
3 3 20.51897
> barplot(meanmpg$averagempg, main = "Average MPG", names.arg = meanmpg$cluster,
> summary(aov(mpg ~ factor(label), data = Lab9Data))
Df Sum Sq Mean Sq F value Pr(>F)
factor(label) 2 9788 4894 206.1 <2e-16 ***
Residuals 258 6126 24
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

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	cluster	R 3 cluster	Count of id	Column Labels	Grand Total
1	cluster_0	1	76	1	76
2	cluster_1	2	127	2	129
3	cluster_2	3	56	3	56
4	cluster_0	1	76	1	76
5	cluster_1	2	127	2	129
6	cluster_2	3	58	3	58
7	cluster_0	1	76	1	76
8	cluster_1	2	127	2	129
9	cluster_2	3	56	3	56
10	cluster_0	1	76	1	76
11	cluster_1	2	127	2	129
12	cluster_2	3	58	3	58
13	cluster_0	1	76	1	76
14	cluster_1	2	127	2	129
15	cluster_2	3	56	3	56
16	cluster_0	1	76	1	76
17	cluster_1	2	127	2	129
18	cluster_2	3	58	3	58
19	cluster_0	1	76	1	76
20	cluster_1	2	127	2	129
21	cluster_2	3	56	3	56
22	cluster_0	1	76	1	76
23	cluster_1	2	127	2	129
24	cluster_2	3	58	3	58
25	cluster_0	1	76	1	76
26	cluster_1	2	127	2	129
27	cluster_2	3	56	3	56
28	cluster_0	1	76	1	76
29	cluster_1	2	127	2	129
30	cluster_2	3	58	3	58
31	cluster_0	1	76	1	76
32	cluster_1	2	127	2	129
33	cluster_2	3	56	3	56
34	cluster_0	1	76	1	76
35	cluster_1	2	127	2	129
36	cluster_2	3	58	3	58

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.