**CIDM 6355 Data Mining Methods** **LA4 Instruction & Template**

(40 points in total; Due 11:59 PM CDT, August 7, 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: Adrian Varela

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 08 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 pivot table has broken down the clusters from 2 and three cluster attributes. The 3 cluster gave more in depth analysis splitting attribute 0 and 1 further than combining them into just 0 as shown in image 1.

A screenshot of a computer

Description automatically generated

1. 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). I do not as the difference in actual means is probably significant.

**A screenshot of a computer

Description automatically generated**

**Part 2: Please submit your deliverables and answer questions required in Class 08 R Lab** (26 points).

1. 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). One of the main differences is the amount of clusters broken up for the R dendrogram versus the RM. The other difference is that RM let us view the numbers and easily read the clusters broken down in the dendrogram.

A screenshot of a computer

Description automatically generated

1. 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). In the chart, Clusters 2 is greatly higher for the average versus cluster 1 and 3 making cluster 2 have a higher grouping in MPG.

A screenshot of a computer

Description automatically generated

1. 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). The factor being 206.1 shows that there is high compatibility for the clusters.

A screenshot of a computer

Description automatically generated

1. 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 07 R Lab. (3 pts for your screenshot and 4 pts for your answer). They are the same as shown in the pivot table. Although the clusters are labeled differently they both show how each cluster gives the same results of totality in numbers.
2. A screenshot of a computer

   Description automatically generated