**CIDM 6355 Data Mining Methods HW4**

(60 points in total; Due 11:59 PM Central Time, October 23, 2023)

Requirements: This homework 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 homework should be addressed to the instructor. You are required to follow the instruction to complete all the questions and deliverables. This is an individual homework assignment, 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: Jordan Unfred

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 question required in Week 9 Lab- RapidMiner (36 points in total: 6 points for each).**

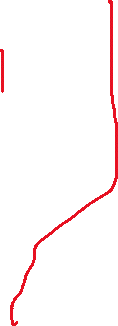
1. Step 3.2. Observe the Scatter 3D plot from any perspective by rotating it using your cursor. What problem do you observe? Hint questions: How are clusters are distributed at each dimension? Does each attribute contribute to the clustering model equally? Is there any attribute dominating the clustering model?

In your 3D plot, the fact that most data points cluster in a specific area suggests that three particular attributes hold significant influence in shaping your data's patterns. These attributes are more important than others in your analysis and may benefit from further exploration and dimensionality reduction for improved model efficiency.

1. Step 4.4. If you are asked to generate five clusters from this dendrogram manually, how you are going to do? Please show it in your dendrogram and then take a screenshot of it with date and time (Screenshot 1). Which cluster has the largest number of records? Please label it in your dendrogram.

A screenshot of a computer

Description automatically generated



I tried to complete the task, but I could not get the dendrogram to cut in the user interface of RM. Cluster 2 would have the largest number of records

1. Step 4.11. Compare the two 3D Scatter Plots (Steps 4.9 and 4.10) and then think about how many clusters are better, 2 or 3? Why?

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

To determine whether two or three clusters are better in the 3D scatter plots from Steps 4.9 and 4.10, we need to scrutinize the distribution and separation of data points. If distinct, well-separated groupings are visible, this suggests the presence of clusters. The choice between two or three clusters should align with the practical interpretation of the data, cluster validation metrics, and the specific goals of your analysis. Be mindful of overfitting and consider cross-validation to make an informed decision. Ultimately, the optimal number of clusters depends on a comprehensive assessment of your data's characteristics and the context in which you are working.

1. Step 4.12. Empirical Examination: Take a screenshot of your PivotTable for the empirical examination with date and time (Screenshot 2). What conclusion can you make based on the PivotTable?

A screenshot of a computer

Description automatically generated

The PivotTable presents a breakdown of data points across three clusters: cluster\_0, cluster\_1, and cluster\_2. Cluster\_1 has the largest number of data points, with 16,852, followed by cluster\_0 with 9,872, while cluster\_2 contains 7,466 data points. This tabulation helps in understanding the distribution of data within each cluster, providing insights into the composition of the clusters.

1. Step 5.8. Take a screenshot of your column and scatter charts with date and time (Screenshots 3 and 4). What conclusion can you make from each of the two charts?

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

We can conclude that cluster 2 has the largest population

1. Step 5.9. Take a screenshot of the ANOVA Test table with date and time (Screenshot 5). Based on the ANOVA table, do you think the mean mpg of the three clusters differ at the 95% confidence level? Why?

**A screenshot of a computer

Description automatically generated**

I do not think the confidence level is different between the three clusters as the data is highly similar between all of them on every table

**Part 2: Please submit your deliverables required in Week 9 Lab- R (24 points in total and 6 points for each: 3 pts for a screenshot and 3 pts for your answer).**

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.

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.

A screenshot of a computer

Description automatically generated

I couldn’t get the barplot to display

1. Deliverable R3: take a screenshot of the ANOVA result with date and time and make your conclusion.

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 Week 8 R Lab.

A screenshot of a computer

Description automatically generated

I couldn’t get the PivotTable to work because my R is not outputting the clusters into the csv. The only data that I could get the same was the sum of ID