**CIDM 6355 Data Mining Methods HW4**

(60 points in total; Due 11:59 PM Central Time, March 25, 2024)

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: Mehnaz Afrose

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 (32 points in total).**

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 (3 pts for your screenshot and 4 pts for your answer).

A screenshot of a computer

Description automatically generated

In this dendrogram, the red line I drew cuts through five intersection points across the graphs, generating five clusters. The blue circles indicate the cutting points between the clusters and the horizontal line. From this visual representation, I identified the cluster that includes the largest number of records. Under this cluster, we can see that there is the highest data density compared to the other clusters.

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? (4 pts for your answer).

Answer: By comparing the two 3D Scatter Plots, I think three clusters are better.

Because all the data seems naturally to fall into three distinct groups. While we select 3 numbers of clusters, it shows three distinct clusters with three different colors or from three different categories. But when we select the number of clusters 2, it still shows three distinct clusters, but two of them are in same color, which implies that they are from same category.

There is no significant overlap among the three clusters. Adding the third cluster helps to better capture the nuances and variations within the data.

Also, the purpose of this clustering is grouping similar items together. There is a clear visualization of data falls for three distinct categories according to the similarity.

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? (3 pts for your screenshot and 3 pts for your answer).

A screenshot of a computer

Description automatically generated

We can observe from this pivot table that cluster 1 in the 2-cluster model has the exact same number of data objects (76) as cluster 0 in the 3-cluster model.

The two clusters include the same data objects. I checked this in the Excel file. Using a filter, I kept all the data labeled as cluster\_1 in the 2-cluster column, and it had exactly 76 entities. Cluster 0 from 3-cluster model and cluster\_1 from 2-cluster model has a one-to-one relationship.

Cluster\_0 from 2-cluster model contains 129 elements of cluster\_1 from 3-cluster model, and 56 elements of cluster\_2 from 3-cluster model.

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? (4 pts for your screenshots and 4 pts for your answer).

A screenshot of a computer

Description automatically generated

From the column chart, we can observe that most of the average MPG cars fall into cluster 1. Cluster 1 has the highest average MPG number, which is around 28. Cluster 2 has the highest average MPG of 20, and cluster 0 has 15. On the other hand, the average cubic inches, where the range of cubic inches is considered from 0 to 1 (where 1 being the biggest vehicle and 0 being the smallest) are only divided between cluster 0 and cluster 2. Comparing to cluster 2, cluster 0 has bigger cars, as its value is around 1. There are no elements of average cubic inches in cluster 1. Cluster 0 has more elements than cluster 2 of average cubic inches.

A screenshot of a computer

Description automatically generated

According to this scatter chart, we can conclude that most Japanese and European cars are clustered into cluster 1, with very few of them in cluster 2, and none of these brands in cluster 0. Only cars from the US are in clusters 0, 1, and 2. When we choose the “color” parameter as “mpg”, we can observe that US brand cars with mpg less than 30 are clustered in cluster 0. Some Japanese and European cars, along with some US cars with similar mpg, are clustered in cluster 2. However, for cluster 1, we cannot assume, as it is very scattered in color.

But if we select the color parameter as cylinders, we can find that all heavy-weight cars, which require more cylinders, are made in the US, and they are in cluster 0 (as all the dots are red and only in the US portion). Comparing this, Japan and Europe make lightweight cars that require fewer cylinders to operate (the color of dots is light blue), and these cars fall into cluster 1. Cluster 2 includes some medium-weight cars (the color of dots is orange) from all brands, with US brands having more scatterplots than others.

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? (3 pts for your screenshot and 4 pts for your answer).

**A screenshot of a computer

Description automatically generated**

Based on the ANOVA table provided:

The F-statistic is 228.251.

The degrees of freedom (DF) for between groups is 2 and for residuals is 258.

The p-value associated with the F-statistic is 0.

With a p-value of 0.00, which is less than the typical significance level of 0.05, we reject the null hypothesis (H0) in favor of the alternative hypothesis (H1). Therefore, we conclude that at least two clusters of cars have different average mpg.

In other words, there is sufficient evidence at the 95% confidence level to suggest that the mean mpg of the three clusters differs.

**Part 2: Please submit your deliverables required in Week 9 Lab- R (**28 points in total and 7 points for each question: 3 pts for your screenshot and 4 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. In the dendrogram produced in R, the height of the clusters is shown. However, in RapidMiner, the dendrogram doesn’t display any parameters.
2. In the R dendrogram, there are scale measurements present along the x-axis, representing the individual data point values. However, in the RapidMiner dendrogram, no data points are labeled.
3. Using the scale measurements from the R dendrogram, we can calculate the height of any cluster or the distance between two clusters. On the other hand, in the RapidMiner dendrogram, we cannot calculate such measurements because it lacks scale measurements.
4. Deliverable R2: take a screenshot of the chart with date and time and describe it briefly.

A screenshot of a computer

Description automatically generated

In this chart, we observe that the average mpg is highest in the second cluster, labeled as cluster\_2. The highest bar in this cluster exceeds 25. Thus, cars with a higher average mpg predominantly belong to cluster\_2. Cars with an average mpg ranging around 20 are predominantly found in cluster\_3, while those with average mpg around 15 are primarily assigned to cluster\_1.

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

The ANOVA results indicate a significant difference in mean values among the clusters labeled in the dataset. With a high F-value of 206.1 and an extremely low p-value of less than 2e-16, we can reject the null hypothesis and conclude that there are significant differences in the average values of the clusters. This suggests that at least two clusters have different average values.

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

From the pivot table, we observe that the two clustering results in R and RM are the same because we obtained a rank zero diagonal matrix in the pivot table. This implies that all the elements of one cluster in RM belong to a single cluster in R. For example, all 76 elements from cluster\_1, obtained from R, belong to only a single cluster of RM: cluster\_0. The same pattern applies to the remaining clusters as well. Specifically, 129 elements of cluster\_2 in R belong to cluster\_1 in RM, and 56 elements from cluster\_3 belong to cluster\_2 in RM. Given the rank zero diagonal matrix, we can conclude that the two clustering results in R and RM are identical.