**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: Mariam Adegbindin

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

Amongst the five clusters, Cluster 5 has the largest number of records. But if we look at the lowest level, we can see that there are more clusters with larger numbers.

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).

The three-cluster model outperforms the two-cluster one in our dataset as each cluster is proportional to one of the three features in the three-cluster 3D scatter plot. When compared, the 2-cluster model is imbalanced since it produces two clusters consisting of the same groups and one cluster consisting of separate groups.

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

An unstructured cluster with several differences will be produced by the model with two clusters when cluster\_1 and cluster\_2 is joined. Because there is a fair distribution of clusters in the three-cluster model, it will perform better.

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

The bar chart represents the “Average MPG (Miles Per Gallon**)”** for each cluster, the X-axis shows the three clusters (labeled as Cluster 0, Cluster 1, and Cluster 2), the Y-axis represents the **average** MPG values. Cluster1 (29 mpg) has the highest average MPG while Cluster 0 (15 mpg) and Cluster 2 (20 mpg) have lower average MPG values. Cars in Cluster 1 are more fuel-efficient, achieving higher MPG. Cluster 0 and Cluster 2 may represent different types of vehicles or driving conditions.

A screenshot of a computer

Description automatically generated

We can see distinct groupings of car brands based on their MPG values. Some clusters have higher MPG values, while others have lower MPG values. Certain car brands stand out as having exceptionally high or low fuel efficiency. By examining the scatter plot, we can identify which car brands are associated with better fuel efficiency (higher MPG) and which ones are less fuel-efficient. Cluster\_0 which has only the U.S has the lowest mpg, and cluster\_1 which includes the U.S, Europe, and Japan has the highest mpg, cluster\_2 has the second to highest mpg which included all the three brands with more of U.S.

If you look at the average mpg for each cluster group in both the bar chart and the scatter plot, you can see that clusters\_0, cluster\_1, and cluster\_2 are similar.

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

At the 95% confidence level (α = 0.05), since the p-value is less than 0.05, we reject the null hypothesis, concluding that there is a significant difference in the mean mpg of the three clusters. We have evidence to suggest that the mean MPG of the three clusters differs significantly.

**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**).**

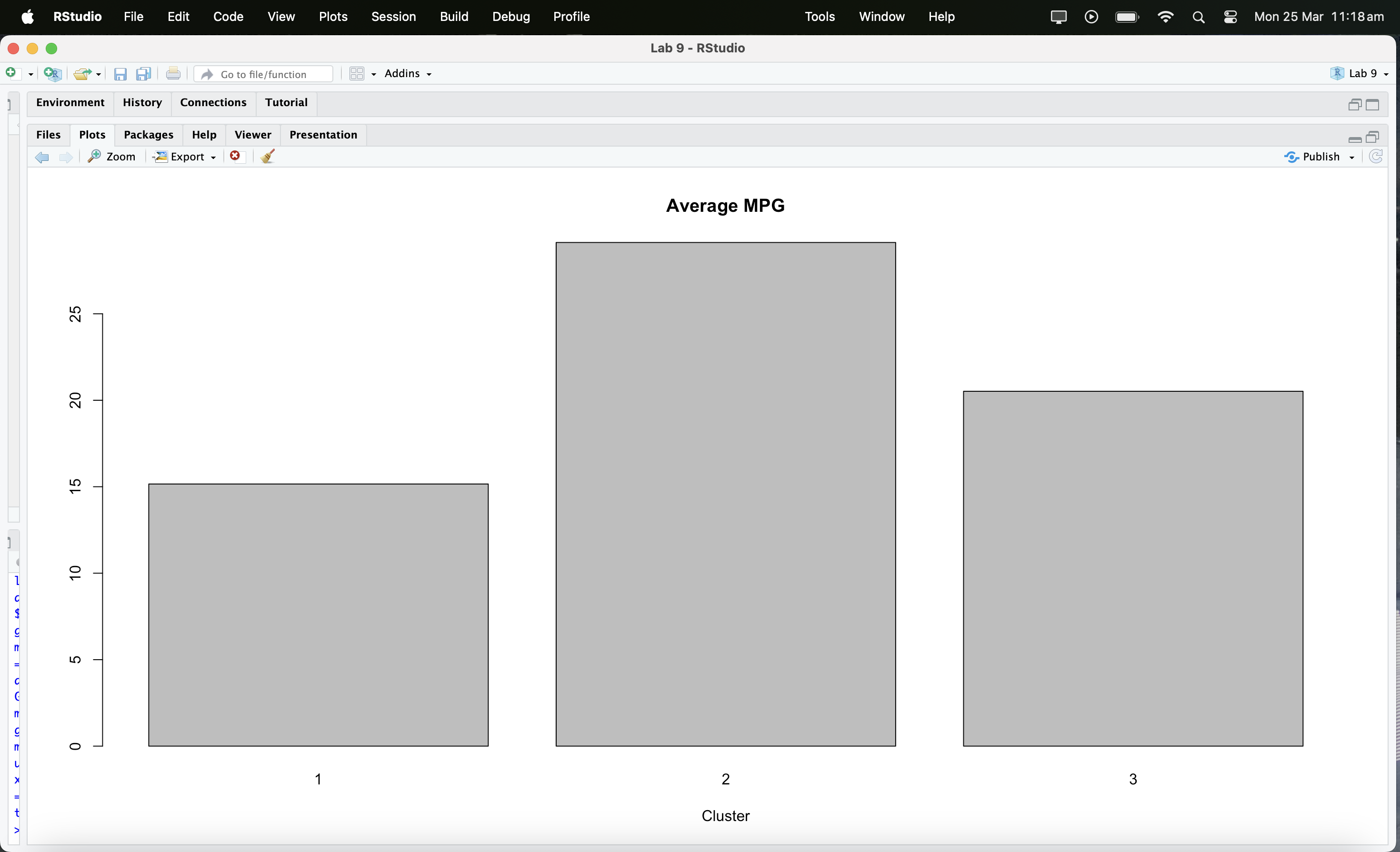
1. Deliverable R1: take a screenshot of the dendrogram with date and time. Compare it with the one generated in RM and find at least two differences.

A screenshot of a computer

Description automatically generated

In R there is a ruler on the left to show what points the clusters start from so its easier to quantify. In RM a ruler measurement is not included in the dendrogram. In RM you can see how the clusters are grouped at the lowest level of data while in R the clusters at the lowest level aren’t as clear.

1. Deliverable R2: take a screenshot of the chart with date and time and describe it briefly.



There are three groups in the bar chart, and their average values are displayed in each of the clusters. Out of all the clusters, Cluster 1 has the worst average fuel economy (15 mpg), Cluster 2 has the best (30 mpg), and Cluster 3 has the average of both (15 mpg) and (30 mpg).

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

Based on the extremely small p-value (less than 0.05) we reject the null hypothesis and conclude that there is a significant difference in mean 'mpg' among the different levels of the 'label' factor. In other words, the 'label' factor significantly affects the 'mpg'.

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

There was a total of 261 results from both models, which show all three cluster groups. A small number separates each RapidMiner cluster group from its R counterpart. There are differences between RapidMiner and R in the values of cluster\_0 (76) and cluster\_1 (127), cluster\_2 (56) and cluster\_3 (129).