Cluster analysis

Homework Week 9

The assignment

Data preparation, identification

Load the new, improved `mtcars2` datset, available here: mtcars2 in Google Drive. Identify the structure of the data.

Analysis

Univariate relationships

Provide a scatterplot matrix of all continuous variables in the `mtcars2` dataset.

Distance matrix

Provide the code used to calculate an Euclidean distance matrix on the data plotted above. Do not provide the matrix in your submitted homework file.

Cluster analysis

Cluster diagram

Provide script for an hierarchical cluster diagram of the Euclidean distance matrix calculated above. Show the plot, with meaningful labels, and answer the following questions:

Text answers:

- Which three cars are the most similar, based on these data? How do you know, and what might account for their similarity?
- Which car or cars are the most unique? How do you know?
- Which car is more similar to the Toyota Corolla: the Porsche 914-2, or the Duster 360? How do you know?

Visualize clusters

Provide script for two cluster diagrams, one with two clusters, and another with three clusters.

Text answers:

- At which Euclidean distance do the two-group and three-group clustering scenarios cut the tree?
- How many clusters would be formed if one were to cut the tree at 150?

k-means clustering

Determine best number of clusters

Provide script that compares the residual Sum of Squares of various numbers of potential clusters. Plot the change in residual error for each cluster scenario, and answer the following questions:

Text answers:

- How many clusters do you think these data best sort into? On what do you base your answer?
- Which categorical variables in the `mtcars2` dataset have the same number of levels as the clustering scenario you selected above?

Test clusters

Text answers:

Pick one of the variables you identified above and:

- Provide code for a contingency table comparing the cluster scenario you selected above with the counts of each level of the cateogrical variable you selected above.
- Give code for and present a mosaic plot and χ^2 test based on the contingency table.
- Interpret these results.
- Identify an additional question of these data and the relationships in the cluster diagram you haven't been able to identify here.

Bonus round

There are a couple pretty odd associations in the cluster diagram—clearly `hclust` doesn't know much about style. Identify one or more of these odd pairs.