Evidence of Comparisons

Data Description:

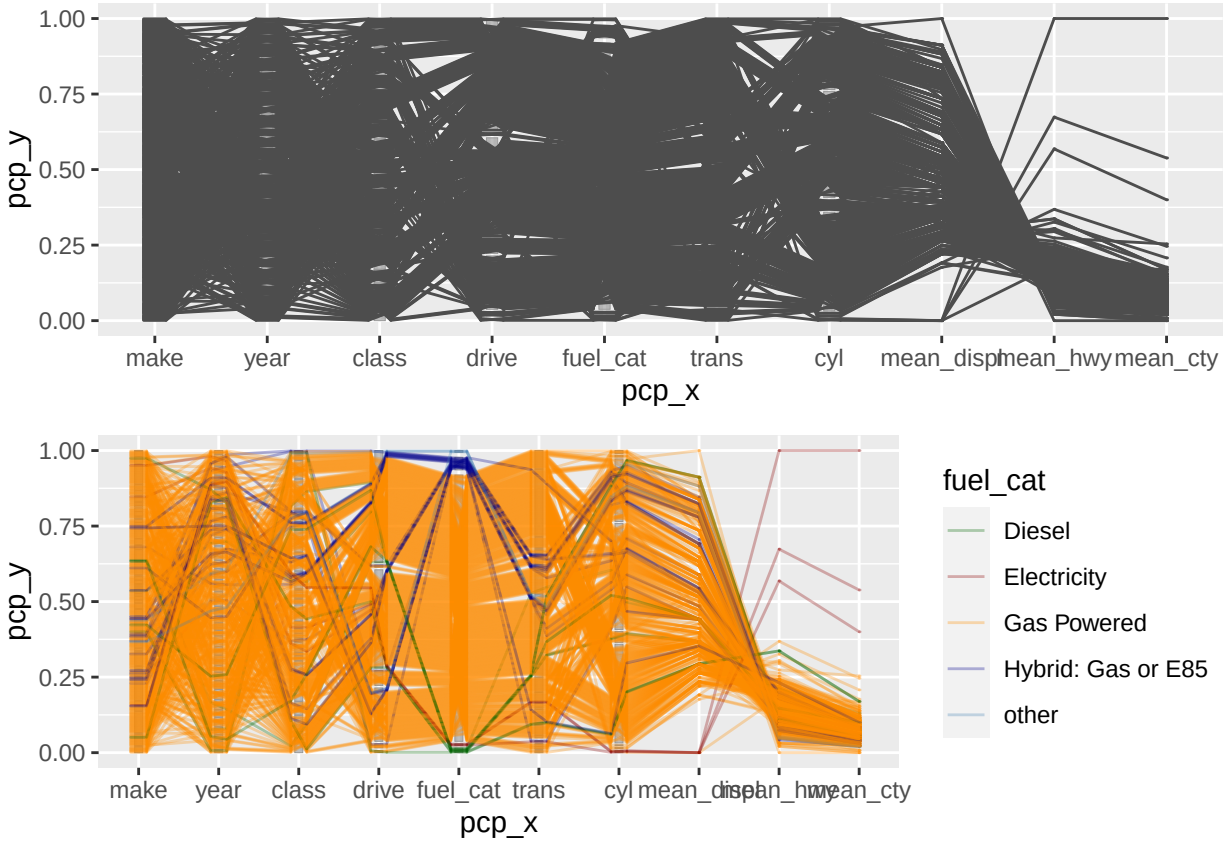
Fuel economy data for 1984-2015 from the US EPA, conveniently packaged for R users. FuelEconomy Github written by Dr. Hadley Wickham.

Detection of Outliers Within Clusters

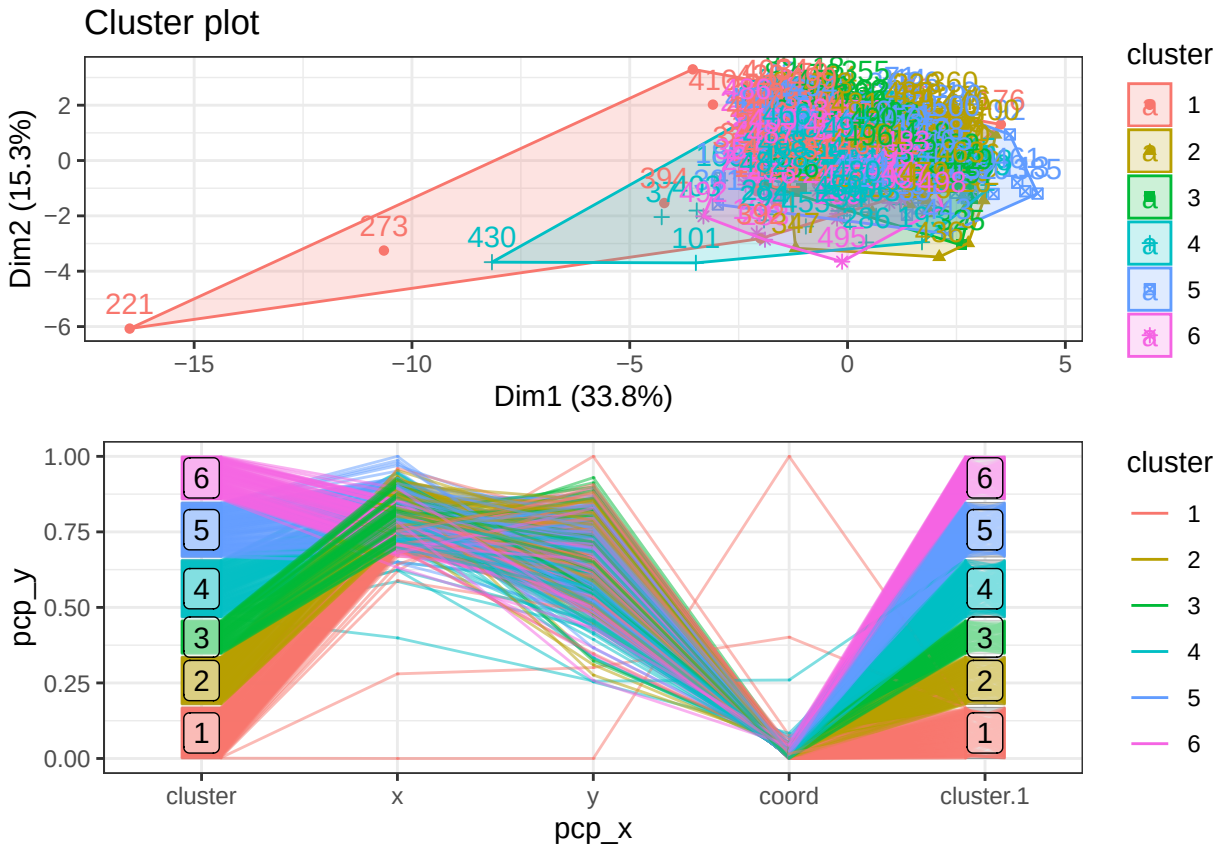
Finding outliers within groups is an integral part of data analysis, and scatter plots and parallel coordinate plots (PCPs) can be very different in how well they do this. It can be hard to see outliers in scatter plots, especially if they are close to the edges of groups or if the data is very complex. Scatter plots can only show two or three dimensions at a time, which means that outliers in higher dimensions might not be seen because they are not displayed simultaneously (Feng et al., 2010). Also, when groups in a scatter plot overlap, outliers may blend in with the clusters, making them harder to find.

PCPs, on the other hand, are a more reliable way to find outliers within groups. A vertical scale shows every parameter in the data, and lines are drawn between the variables to show the data points. Lines distinct from the general pattern that other lines in a cluster follow are usually called outliers in PCPs. This difference can be seen, even in data with many dimensions, because each dimension is shown on its own plane, making any oddity in any dimension easy to spot (Holten & Van Wijk, 2010). As a result, PCPs can better show outliers that may be part of the same cluster in some aspects but not in others. This gives them a more thorough and detailed strategy for finding outliers in large datasets.

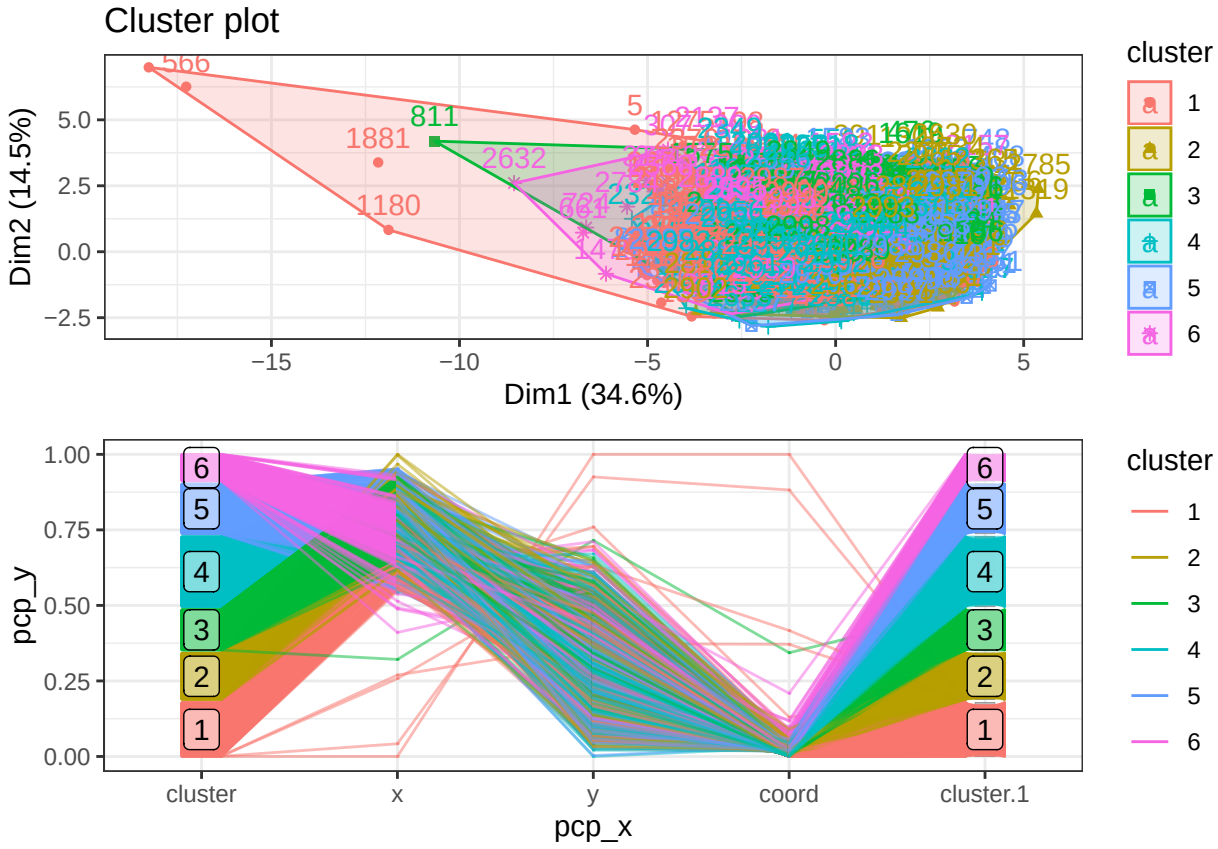
Without Clustering



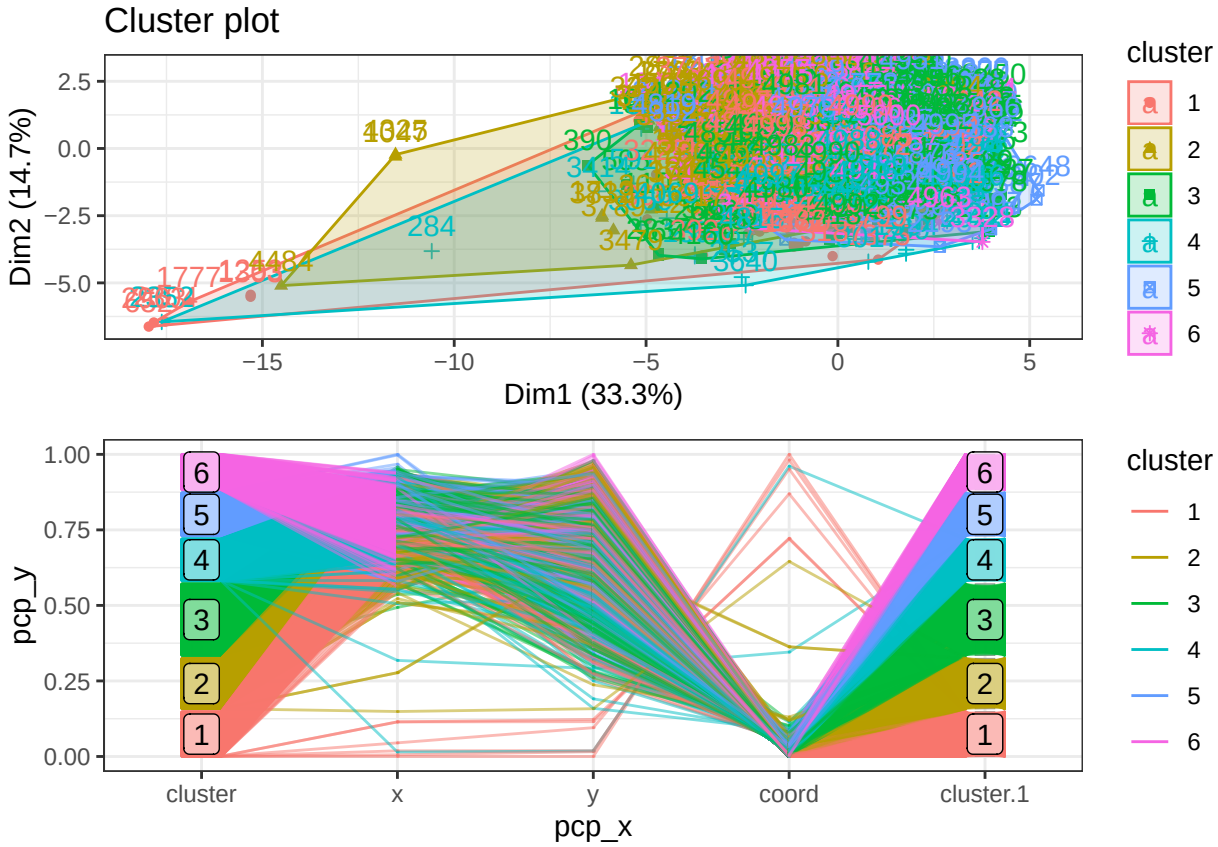
PAM Clustering Method 1 – Using Random Selection of 500 rows



PAM Clustering Method 2 – Using Random Selection of 3000 rows



PAM Clustering Method 3 – Using Random Selection of 5000 rows



References

1. Holten, D., & Van Wijk, J. J. (2010). Evaluation of cluster identification performance for different PCP variants. Computer Graphics Forum. Link to paper
2. Feng, D., Kwock, L., Lee, Y., & Taylor, R. (2010). Matching visual saliency to confidence in plots of uncertain data. IEEE Transactions on Visualization and Computer Graphics. Link to paper