690V – Visual Analytics
Homework 3
Suhas Keshavamurthy

Question:

You will use data set 1 (with no missing values) for simple clustering. Select 2 clustering algorithms (or more) from a library. Run your algorithms on the data set. Explore and compare using interactive visualizations. Discuss what's different on each, what their parameters are impacting, what's missing, etc... .

Data

https://github.com/SuhasKMurthy/Visual-Analytics/tree/master/HW3

Wholesale customers data.csv

Visualization:

Github link –
https://github.com/SuhasKMurthy/Visual-Analytics/blob/master/HW3/vis.py

How to run the file:

In a command terminal change to the appropriate directory
Then enter ‘bokeh serve –show vis.py’ in the terminal
A page in the default browser should open with the visualization

There are 2 graphs shown in the visualization.
One of them is used to demonstrate K-Means clustering. The other demonstrates Ward agglomerative method.

We are comparing the annual spending of two products against each other. The data can be further drilled down by ‘Channel’ and ‘Region’.

We observe in the data that most of the data is clustered around smaller units of spending in both the products. However there are outliers to this for both the products under comparison. In the K-Means method, the outliers tend to have an outsized effect on the classification. As the number of clusters (input) increases, the number of classifiers for the outliers tend to remain constant whereas the number of classifiers closer to the higher density of data increases.

In the Agglomerative Clustering method with Ward linkage criterion, the effect of outliers on data is less prevalent. Also as required by the Scikit learn library the criterion for choosing connectivity matrix for structured Ward, with the parameter of choosing n-neighbours has neglible effect on the clustering (The appropriate default range of value needs to be investigated). This method is better able to classify the outliers. As the number of clusters (input) increases, the number of classifiers for the outliers tend to represent more accurately. When we select 2 clusters, this method is better able to provide the appropriate clusters. This method provides more weightages to the linkages between the data points than K-Means.

Because it appears that the distribution of data is random, both the clustering methods provide similar results though the case with outliers is a little different.