# Thomas Jones – CS-5530-0005 HW#4

## Data Analysis

The data being clustered is from the diamond dataset. Specifically, we are interested in the following columns: carat, depth, table, width, length, and height.

A black and white table with numbers

Description automatically generated

Since the data consists of a number of different scales we will need to use standard scaling to normalize it so the distances are uniform.

A black background with white text

Description automatically generatedSince we also have the categorical data we can use it as a possible check for how the clustering performs. If we separate out the categories and consider unique values per category. Naively multiplying these we get 280 possible combinations. That does not necessarily mean that there are 280 clusters nor that there is a correlation between say color and carat, but it does give some possible insight into how many clusters we might be looking for.

As there are only a few features we can also review the pair plot between them.

A graph of different sizes of data

Description automatically generated with medium confidence

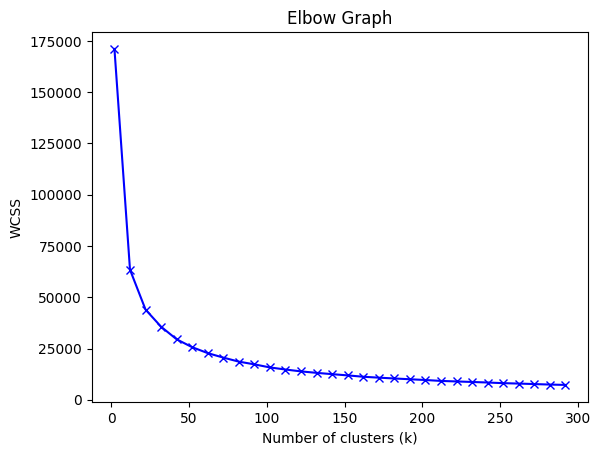
A graph of a graph

Description automatically generated with medium confidenceThe pair plot shows that while there are some trends in the data, mainly between carat and the dimensional features (length, height, width) there are not obvious clusters. The histograms do show some peaking. For example, width, shown to the right and might be a qualitative validation of the clustering.

## A graph with a blue line Description automatically generatedKMeans Clustering

The data was first clustered using k from 2 to 30 in steps of 2 using the default iteration count of 300.

There is not an obvious inflection point in the graph. Instead of a sharp turn the curve is gently sloping.



Based on our previous rough estimate of 280 unique values the clustering was extended to 300 in steps of 10. Given that there are nearly 54k samples in the dataset expecting clusters > 100 is not unreasonable, however analyzing a clustering this large might prove difficult.

## Hierarchical Clustering

Each of the linkage methods was explored. These are shown below. Note, we’re not particularly interested in the labels at the bottom since we are clustering over 100 samples. Instead, we can look at the overall structural elements of the linkages. The linkages were produced with optimal ordering on to get a better visualization of the dendrograms.

The initial dendrogram produced with default settings is shown below. However, the resulting labels at the bottom are difficult to interpret.

A diagram of a city

Description automatically generated with medium confidence

A graph with orange lines and numbers

Description automatically generated

A graph of a diagram

Description automatically generated with medium confidence

A graph with orange lines

Description automatically generated

A graph with orange lines

Description automatically generated

A graph with different colored squares

Description automatically generated

### A screenshot of a computer Description automatically generatedHierarchical Discussion

With optimal ordering we see that 91 and 97 are generally placed on opposite sides of the diagram, generally indicating they are farther apart than other leaves. Further analysis on why this is the case would need to be done. The raw data for these two data points is shown to the right. At least visually, the two diamonds appear to be similar and just based on this observation we might expect the two leaves to be close within the hierarchy. Instead they are joined only at the highest clade.

## PCA Analysis

Applying PCA to the scaled data then taking the best two components yields the following heatmap (as expected since PCA seeks to product linearly independent variables)

A black and white squares with white text

Description automatically generated

The explained variances for PC1 and PC2 are then [0.65534685 0.21406055] respectively.

### KMeans Clustering

We then clustered the PCA separated data and produced scatter plots for values of k from 2 to 18 in steps of 4 which yielded the following. The data does not clearly visually fall into defined clusters from the PCA separated data with the clustering instead producing simple distance based separation of the datapoints.

A yellow and purple dot diagram

Description automatically generated



A colorful dots with numbers

Description automatically generated with medium confidence

A colorful dots and numbers

Description automatically generated with medium confidence

A colorful dots with numbers

Description automatically generated with medium confidence

And the elbow graph which shows the same sloping as with the non-PCA version.

A graph with a blue line

Description automatically generated

### Hierarchical Clustering

The dendrograms for the hierarchical clustering of the PCA data is shown below. As with before, the leaves were truncated based on the method used in order to produce visually interpretable results.

A drawing of a line drawing

Description automatically generated with medium confidence

A graph of a diagram

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated with medium confidence

A diagram of a city

Description automatically generated

A diagram of a group of people

Description automatically generated with medium confidence

### Hierarchical Discussion

The clustering produced by the PCA data produces different hierarchies as would be expected from the data being transformed from 6 to 2 features, even when the features are now linearly separable.

### Overall Discussion

For the given 6 features, carat, depth, table, width, length, and height the data is not easily, if at all, separable into clearly defined clusters. The inclusion of price might provide for better separability as was seen in the previous exercise. Noting also that carat is a weight measure and diamonds, for the most part, are physically continuous across physical measurements, the overall results are not surprising for this domain.