Andrew Stambaugh

MechE 329

Professor Liu

10/25/2022

Homework 5: K Means Clustering Steel Property Data

In this assignment, I utilized sklearn in Python to cluster the data given to us on different types of steel and their mechanical, thermal, and electrical properties.

Data Preprocessing

For both mechanical and thermal properties, I standardized the data using z-scores. In both datasets, many of the features possess vastly different scales, such as Poisson Ratio being 0.25-0.29 and Tensile Strength Ultimate being 350-1700. If the data were not standardized, the k-means algorithm would place a heavy bias on the much larger variables as its cost function is calculated from the total squared distance between each cluster's center and the total squared distance between all points and their cluster centers. In turn, variables with larger scales will have a more profound influence on the algorithm's results than those with smaller scales. On the contrary, the opposite can occur in variables with low-variance scales. In this case, because the variance is low, when standardized, even the smallest deviations will be seen as a major difference in the data. For example, Annealed Stainless and 316 and 25% Hardened Steel are very similar mechanically, and thus, should likely be clustered together. However, their poisson ratios are different by 0.02, which is two standard deviations in difference. This large difference alone will make the algorithm cluster the two apart, despite every other mechanical property being very similar. Because of this, I decided to halve the z-scores of the Poisson ratios to reduce their influence.

Optimal Clusters

Mechanical

A preliminary indicator of the optimal number of k-means clusters is a graph of the objective cost vs the number of k means clusters. In this graph, the optimal number of clusters is where the line appears to form an elbow.

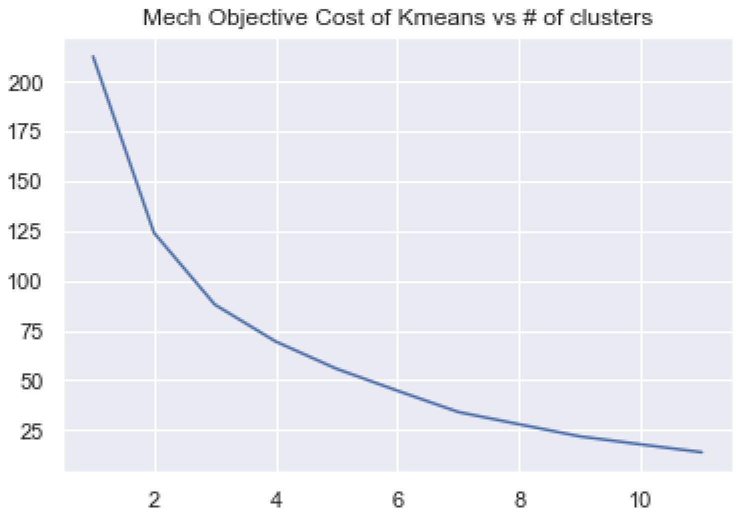


Figure 1: Objective Cost for Mechanical Clusters

In the graph above, we see that the elbow for the mechanical clusters is not sharp, indicating that the optimal number of clusters is anywhere from 3 to 5. Thus, we need to test this range of clusters and decide the optimum based on each scenario's interpretability. When interpreting clusters, the most important factor is that when the clusters are visualized, we can see clear differences between them with little overlap. Taking a look at the clusters below:

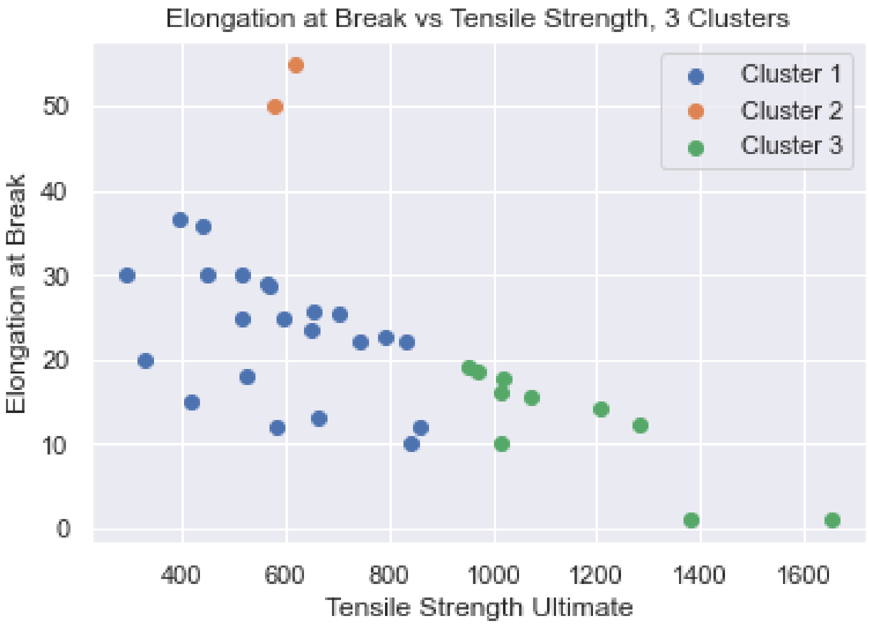


Figure 2: Mechanical Clusters Visualized (k = 3)

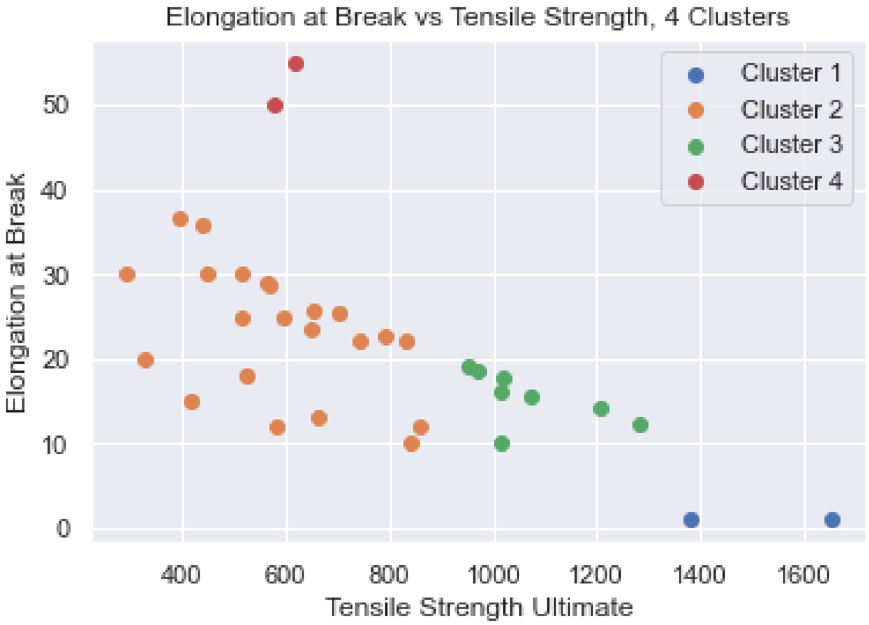


Figure 3: Mechanical Clusters Visualized (k = 4)

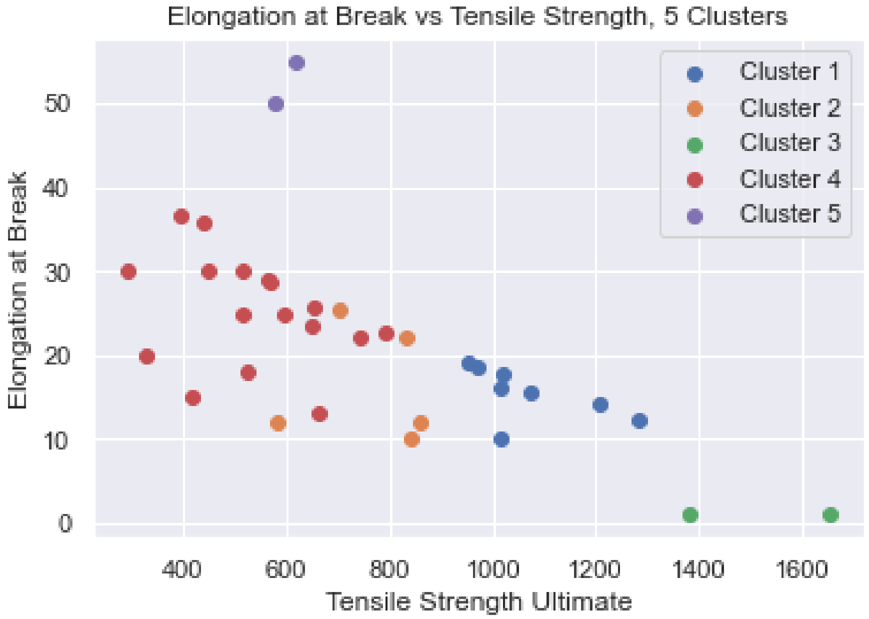


Figure 4: Mechanical Clusters Visualized (k = 5)

We can immediately see that when k = 3, the clusters are clearly defined, with no overlap. When k = 4, the clusters are almost exactly the same, except the points in the bottom right are now in their own cluster. However, when k = 5, clusters 2 and 5 overlap, making them harder to interpret. We can conclude that using 5 or more clusters will only parse the data further and make more, less meaningful clusters. While similar, k = 4 is better than k = 3 in this case because the extra cluster in k = 4 clearly defines a set of outliers that are vastly different from the other data points in regard to elongation at breaking point. Thus, k = 4 is the optimal number of clusters.

Thermal

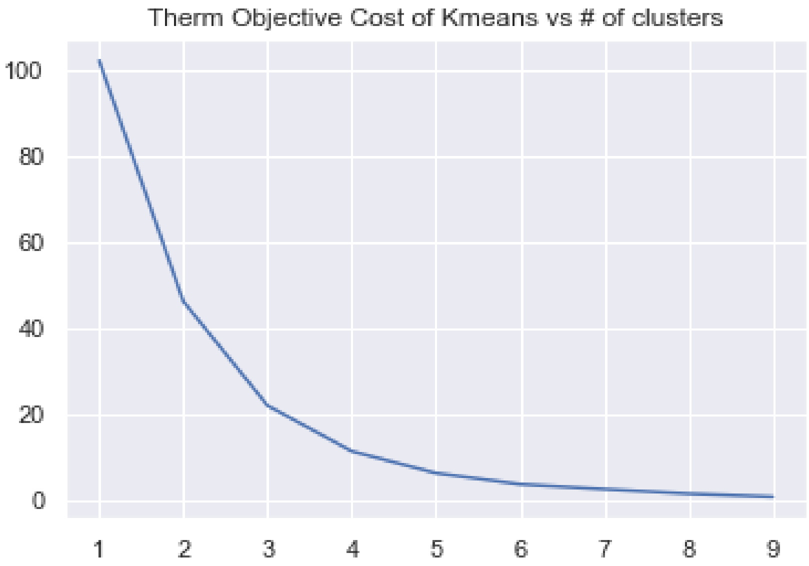


Figure 5: Objective Cost for Thermal Structures

In this elbow, we see that 3-4 clusters are the range for the optimum. Taking a look at 3D plots of the clusters:

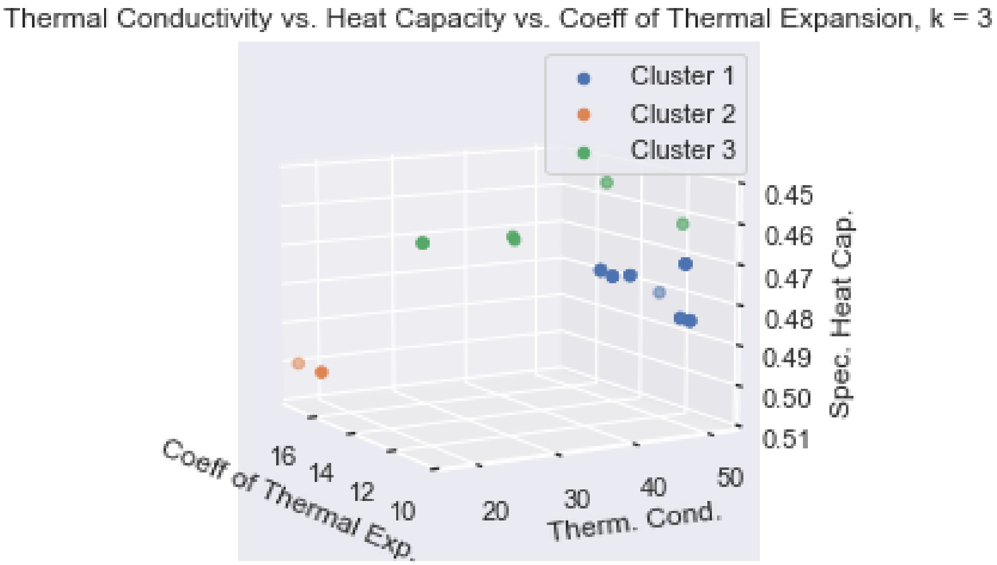


Figure 6: Thermal Clusters Visualized (k = 3)

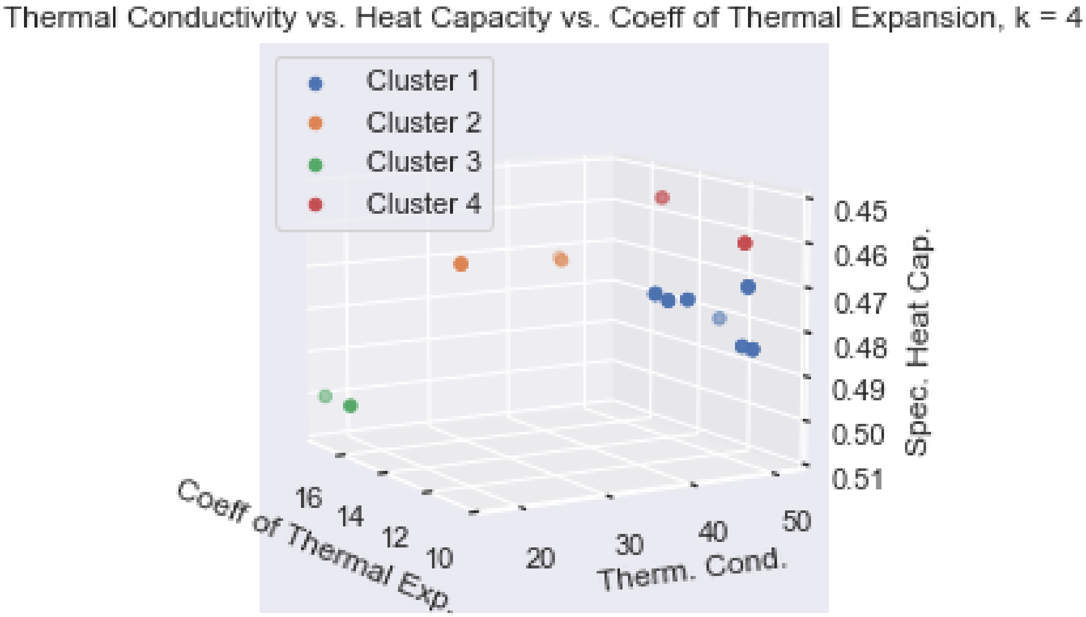


Figure 6: Thermal Clusters Visualized (k = 4)

We see that both scenarios have similarly defined clusters. However, when k = 3, cluster 3 is fairly dichotomous, with one half having very high thermal conductivity and a low coefficient of thermal expansion and the other having a low thermal conductivity and a high coefficient of thermal expansion. This dichotomy is not something that should exist within a cluster, and k = 4 properly separates these two halves. Thus, coupling this with the elbow chart, we can conclude that 4 clusters is the optimal solution.

Electrical

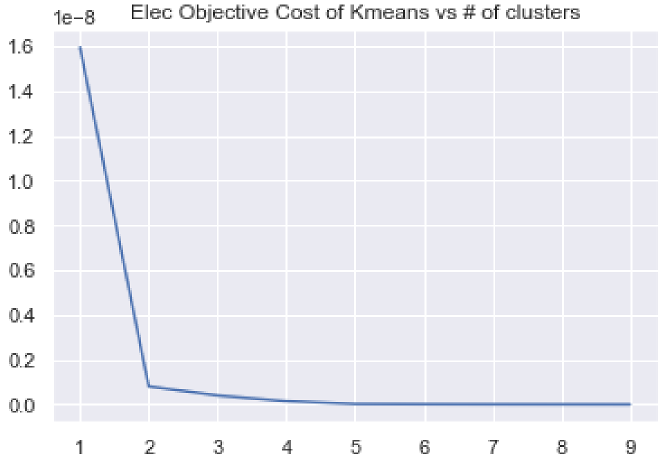


Figure 7: Objective Cost of Electrical Clusters

The elbow chart clearly shows that 2 is the optimal number of clusters.

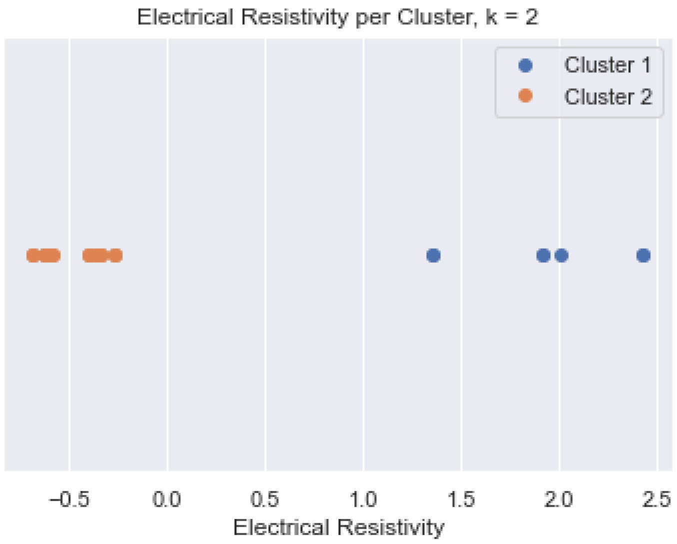
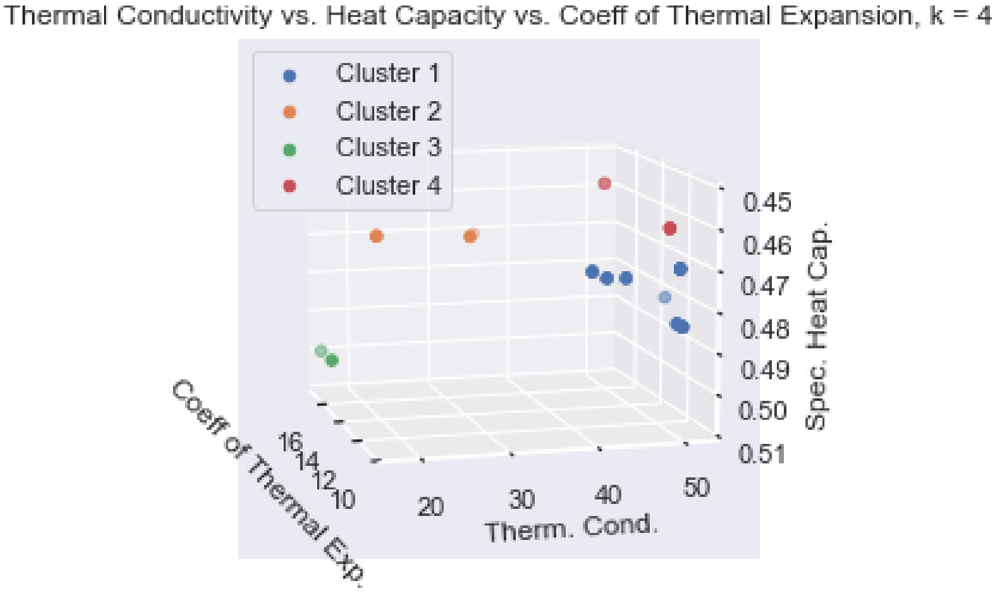


Figure 7: Electrical Clusters Visualized (k = 2)

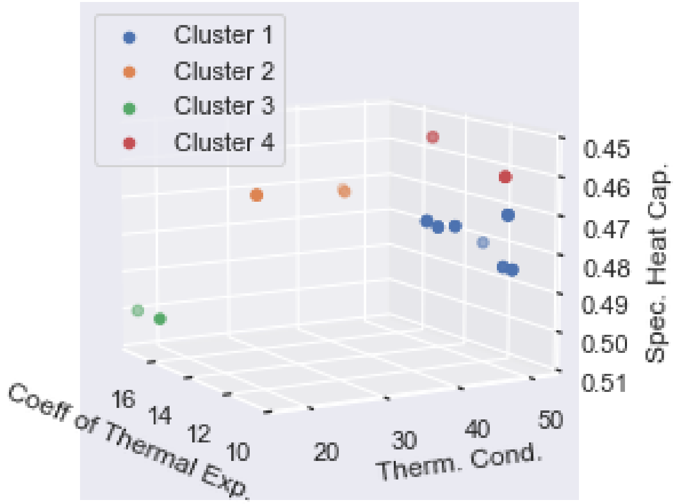
Fortunately, the clusters show a clear difference in the data, and visually, we can see that adding more clusters would not produce any more meaningful results.

Thermal Clusters Visualized in 3D

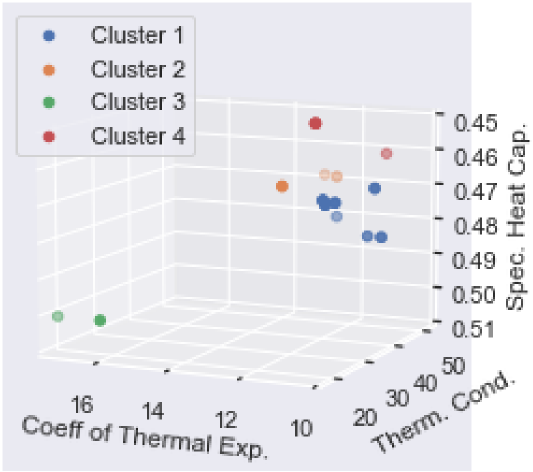
The optimal number of thermal clusters was 4. Here is the graph below with multiple angles for better depth perception.



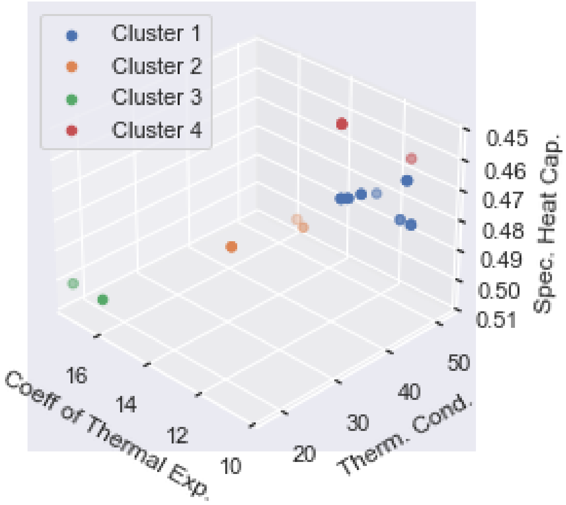
Thermal Conductivity vs. Heat Capacity vs. Coeff of Thermal Expansion, k = 4



Thermal Conductivity vs. Heat Capacity vs. Coeff of Thermal Expansion, k = 4



Thermal Conductivity vs. Heat Capacity vs. Coeff of Thermal Expansion, k = 4


● Cluster 1
● Cluster 2
● Cluster 3
● Cluster 4
Coeff of Thermal Exp.
Therm. Cond.
Spec. Heat Cap.