**Makayla McKibben**

**DSC500**

**Date: July 09, 2024**

**Exercise 5.2**

K-Means clustering is a statistical operation that breaks data sets into discreet parts where the data points in each part are related. The number of parts to split a data set into is determined based on some trial and error and visual evaluation. In order to determine the appropriate number of parts, one should plot the value of K (number of groups) against the respective distortion value (Bento, 2018). A distortion value is "the sum of squared difference between each data point and the centroid," the centroid is the cluster's center (Bento, 2018). The plot of K vs. distortion value is then analyzed, and the elbow of the plot will determine how many groups or K-values to use.

A potential situation where using K-Means clustering would be beneficial has to do with real estate. Analysis of the cost of a house, number of bedrooms, square footage of the interior, and square footage of the lot seem like a lot of data to cross-reference. However, to see how these categories interact, we need to know their relationships. Using K-Means clustering can show that seemingly disparate data points are interrelated.

To plot the cost of the home vs. square footage of the interior, number of bedrooms, and square footage of the lot gives us three plots. We should look for the elbow on our K-Means cluster vs. distortion plot to find the optimal number of groups (K-values). Once determining the optimal number of groups has occurred, apply that number of groups to the original three plots.

In this example, while looking at housing data, the optimal number of clusters for this data slice is five. Applying those five clusters to the original three plots changes the data visualization. See Figures 1-3 for the original plots, Figure 4 for the elbow plot, and five-cluster data in Figures 5-7. The cluster grouping greatly enhances the readability of the data. Additionally, the data is now set up for any further data analysis we may want to conduct.A graph with blue lines and numbers

Description automatically generated

**Figure 1.** Sale price of houses vs. number of bedrooms

A graph showing a blue line

Description automatically generated with medium confidence

**Figure 2.** Sale Price vs. Interior Square Footage

A graph with many dots

Description automatically generated with medium confidence

**Figure 3.** Sale Price vs. Lot Square Footage

A graph of a number of clusters

Description automatically generated

**Figure 4.** Elbow Plot

A graph with colorful lines

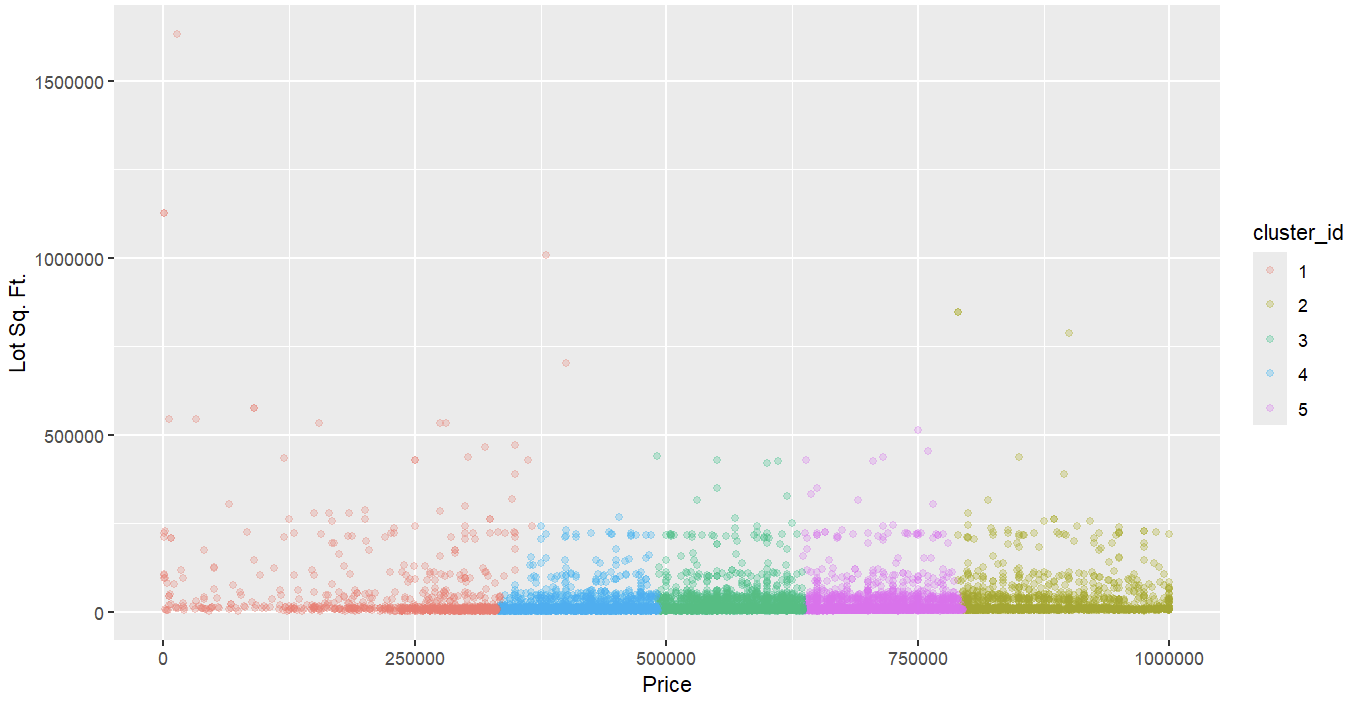
Description automatically generated with medium confidence

**Figure 5.** Sale Price vs. Bedrooms with Five Clusters

A graph showing different colored lines

Description automatically generated

**Figure 6.** Price vs. Interior Square Footage



**Figure 7.** Sale Price vs. Lot Square Footage

References

Bento, C. (2018, December 3). K-means in real life: Clustering workout sessions. Medium. https://towardsdatascience.com/k-means-in-real-life-clustering-workout-sessions-119946f9e8dd