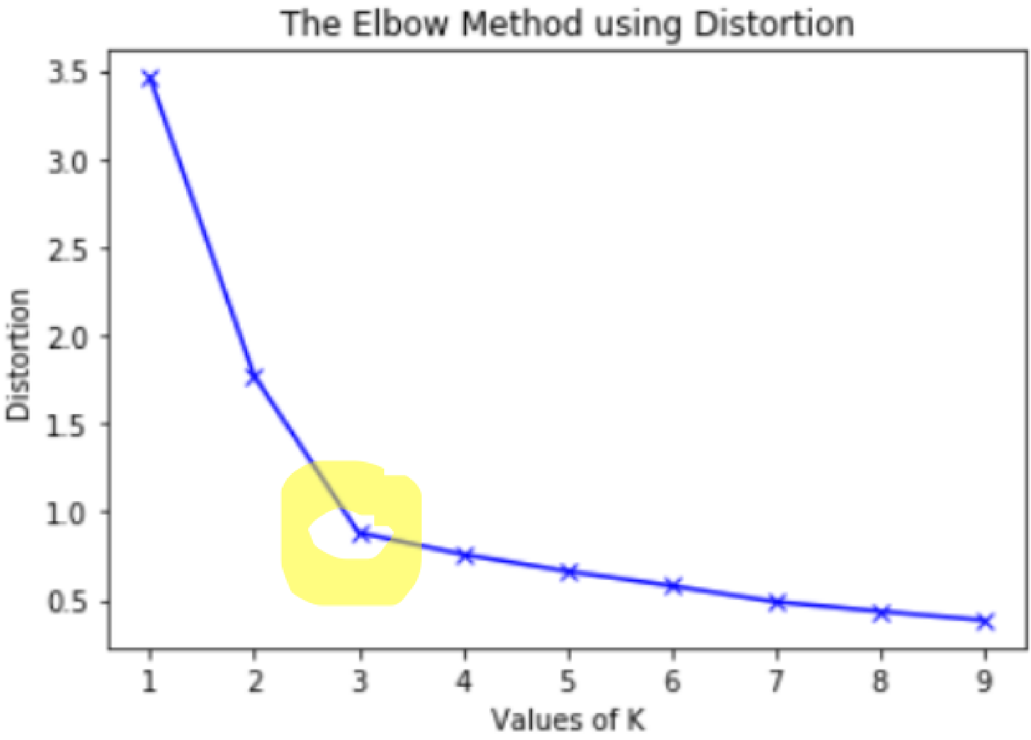
19-04-2022

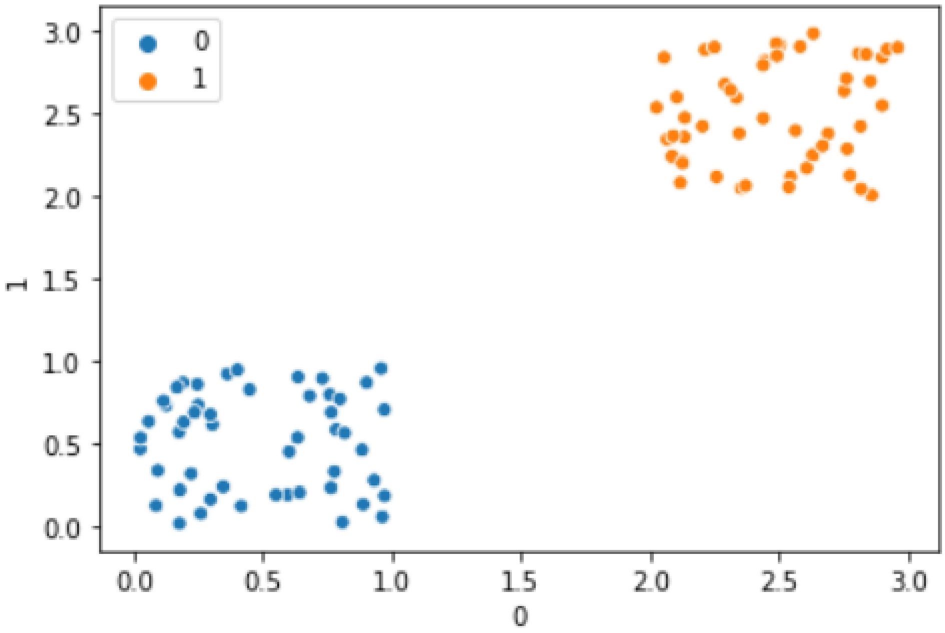
Assignment 8

by Mathias Frank Parking

1. My walkthrough of the book examples can be found at: https://github.com/mp525/DS-Assignment-8/blob/master/Assignment%208.ipynb
2. Most common use cases of *K-means* involve unsupervised learning and clustering of data. Some of the use cases of *K-means* are; Market Segmentation, Data Analysis, Anomaly Detection, Dimensionality Reduction, Segmentation of images and search engines involving image searches. DMSCAN is very good at identifying outliers, whilst also being very good at forming clusters based on the varying densities of the data. So if we have data that has densities in the data which can't be seen as somewhat circular and regular, we need DMSCAN to make accurate clusters. DBSCAN is also commonly used in recommendation engines that can accurately predict what users may do/like in the future and therefore make recommendations for them.
3. Clustering is the use of unsupervised learning to organize data points into groups or clusters related by either distance or density.
4. One of the ways we can make a plot to decide which K-value will give the lowest inertia value is the "elbow" plot. The point on the graph where the inertia has the highest decrease, while keeping the K-value as low as possible, is the point where we see an "elbow-curve" in the graph.

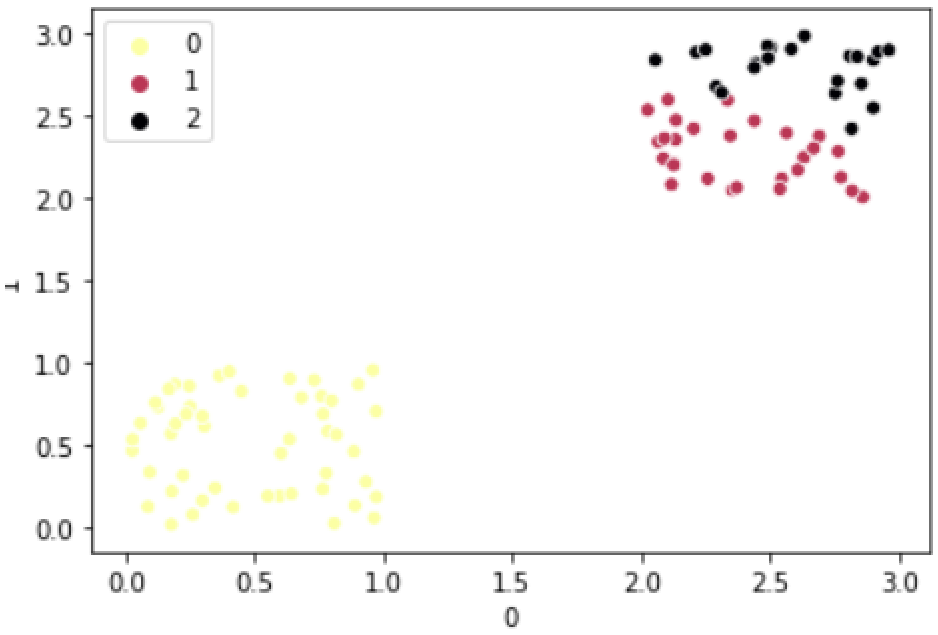


Another method of determining the best value of K, is the Silhouette Score. The silhouette score is a metric used for calculating the goodness of a clustering technique.

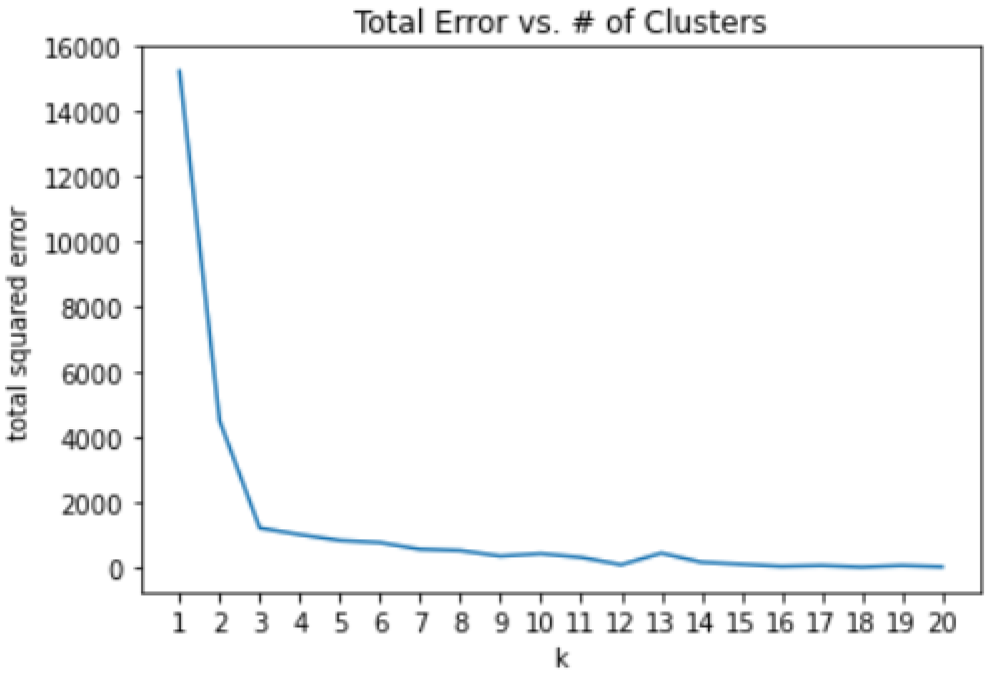


The silhouette score goes from -1 to 1, and the closer to 1 it is, the more separated the clusters are. The number of clusters, K, can be chosen based on the results of these graphs. The first graph has k=2 and a higher silhouette score, and the second graph has k=3 and a lower silhouette score.

5. The best algorithm for large datasets is the K-means algorithm while the DBSCAN is the preferred one for data with high density regions.
6. "High Density Mean" in DBSCAN means the average distance from a given datapoint to that datapoint's centroid.
7. Depending on what k value we give the K-means, we are then already deciding the quality of our model. On the following graphs we can see that if we use the algorithm



with k=3 we get the lowest value of K whilst keeping the inertia low as well.



8. The model can be found at https://github.com/mp525/DS-Assignment-8/blob/master/Olvietti.ipynb