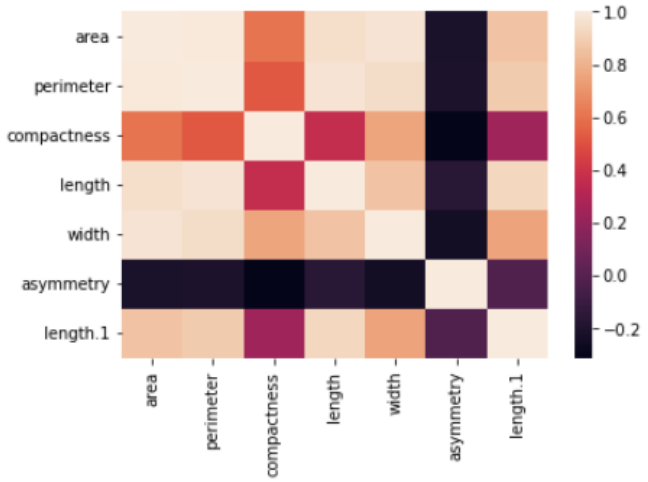
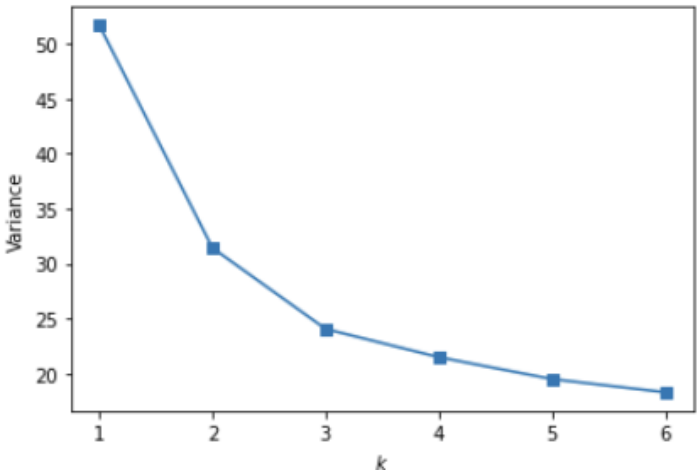
Intro

The dataset that I am working with has 7 columns of data on 210 types of seeds. My goal coming into this project is to determine the different general groups of seeds based on the variables available to me in this dataset. Using a clustering algorithm for this problem is the optimal way to get to a solution because clustering finds relationships and trends within large datasets among many dimensions. In this dataset, each seed has 7 attributes, and there are over 200 different seeds. Trying to group them by hand would be extremely inefficient and likely not very accurate. The seven seed attributes are length, width, area, perimeter, compactness, asymmetry, and "length.1". The measurements of the first five attributes make sense to me, but I am unsure what asymmetry and "length.1" might actually measure.

Pre-Processing

The heat map to the right shows that asymmetry doesn't correlate with any other variables, while compactness has weaker correlation with most everything than the four main variables that measure the size of the seed. This makes sense that variables measuring similar things would correlate together.

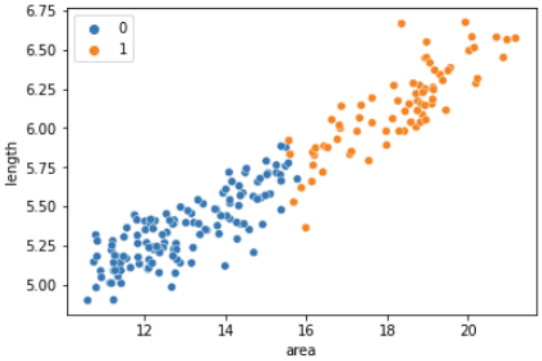


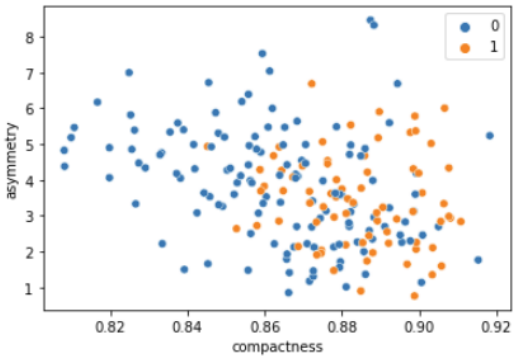


Before making my clustering model. I scaled the data and then made an elbow graph, which is shown to the left. Using two clusters was shown to be the best way to cluster this dataset, which is what I did.

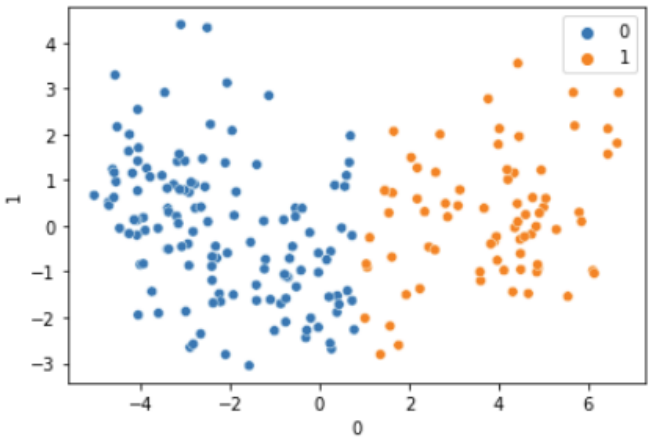
Modeling

When clustering, I used K-means clustering due to it being simpler and quicker. I also knew I was going to be visualizing it on a graph with only 2 dimensions. I made a scatter plot based on area and length, two variables that correlated highly together, which is shown below to the left. I also made a scatterplot between asymmetry and compactness, two variables that didn't correlate well with each other or other variables, shown below to the right.





We can see that the clusters have much sharper boundaries when they are shown on graphs with variables that have a stronger correlation. In addition to just a clustering model, I also used PCA analysis, shown to the right, to visualize how the two clusters look on different pairs of axes.



Impact

The real world impact of this analysis could help farmers, or anyone else that grows many different crops, find ways to maximize their yield. If a certain treatment helps certain types of seeds grow better, trying it on a seed in the same cluster could have a positive effect. However, solely looking at the size of a seed may be misleading. To completely understand how to maximize yield, more domain knowledge about agriculture and more information about the seeds is likely needed.