Intermediate Report
Aaron Goh & Ritesh Sharma
Spring 2019
CS 5140

1. What progress you have made towards your proposed goal? (just data collection is not an option)

The data has been collected through CapIQ and cleaned to remove missing data values. To begin gaining insight into the dataset, clustering algorithms were used to do the initial analysis. Among the clustering algorithms, k-means and DBSCAN were primarily used with Euclidean distance as its measurement pre-set. In plotting the results of both the DBSCAN and the k-means algorithm, the graphical representations seem to demonstrate that there is one cluster existent with scattered outliers outside of the cluster.

While in observation, most of the data would be considered as one cluster, there may be trends occurring in the more dense part of the cluster. In order to simplify analysis and observe these trends, the next proposition is to split the dataset into subsets by "Total Revenue Value" or "Ebitda" into some pre-set bin ranges. These fields are chosen in particular as they are traditional measurements to see a company's relative size and success. In expectation, this should be correlated with the scale of the data and allow us to make the data less noisy to analyze.

2. If you tried some basic approaches: what worked well and what did not?

The initial approach of running k-means algorithm returned poor results. The data was not properly clustered and increasing k came back with visualizations that made little sense where some data points were in more sparse locations, data was split in a non-uniform manner between two centers.

Have observed a singularly dense cluster where the rest of the data was scattered, the next algorithm attempted was the DBSCAN. Differing epsilons were attempted. Using smaller epsilons ranged from ~ 5000 - 15,000, approximately two clusters were observed within each dataset's dense cluster location.

This strongly suggests that there may be multiple clusters given smaller subsets of the data and in excluding "outliers" of the data, this may also improve the performance of the k-means algorithm.

3. What could be done to improve the basic approaches?

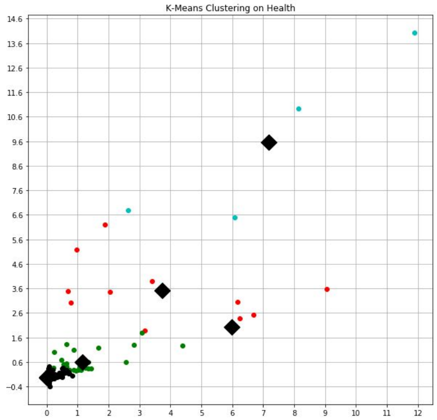
By binning into subsets of the data, it is expected that the outliers will be removed in the k-means algorithm and thus perform better in clustering for the subset containing the the dense cluster identified in the initial k-means run.

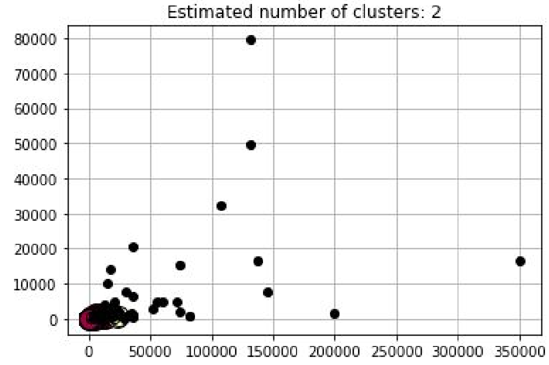
4. What experiments have you run and are you planning to run to demonstrate the effectiveness?

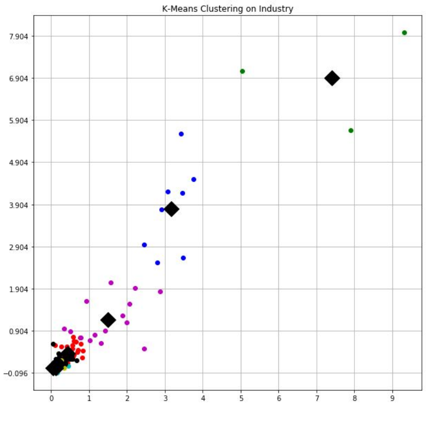
The K-means algorithm and the DBSCAN were run in order observe if there were clear differences in the businesses scraped. The initial observation has demonstrated that there are a large number of outliers and due to the scale of distances observed, it is uncertain if subsets of the data may have key differences to better improve performance.

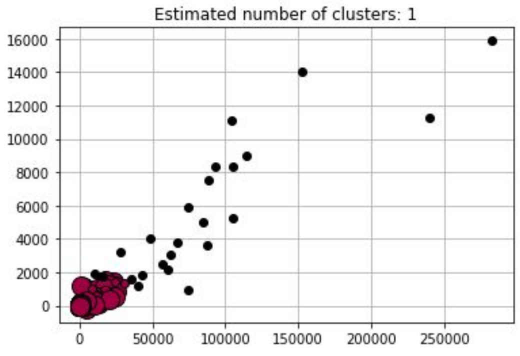
Once the data has been binned into subsets, the "elbow" graph can be used to determine the most effective k and whether the current observation is strongly lacking in performance.

The eventual goal is to run regression by cluster of the dataset to better gain insight and create meaningful comparisons in the datasets. The goal is to observe trends of the EBITDA and Total Enterprise Value by the other data fields.


K-Means Clustering on Health


Estimated number of clusters: 2


K-Means Clustering on Industry


Estimated number of clusters: 1