ML Clustering Analysis

CUNY SPS MSDS

Professor Samuel Gralnick

Team Sugar Cane: Euclid Zhang, Jie Zou, Zhenni Xie

Fall 2022

We previously used a method that searches for pairs within each sector for Pair Trading. This method is not able to find the pairs across sectors that may be good candidates for pairs trading. We may compare each pair of assets in the entire market but doing so is time consuming and it requires intense compute power. Hence, using Machine Learning to cluster assets into clusters and then search for pairs within each cluster would be a more efficient method. In this analysis, we will examine a few clustering methods to confirm if they be used to help identifying pairs for pairs trading.

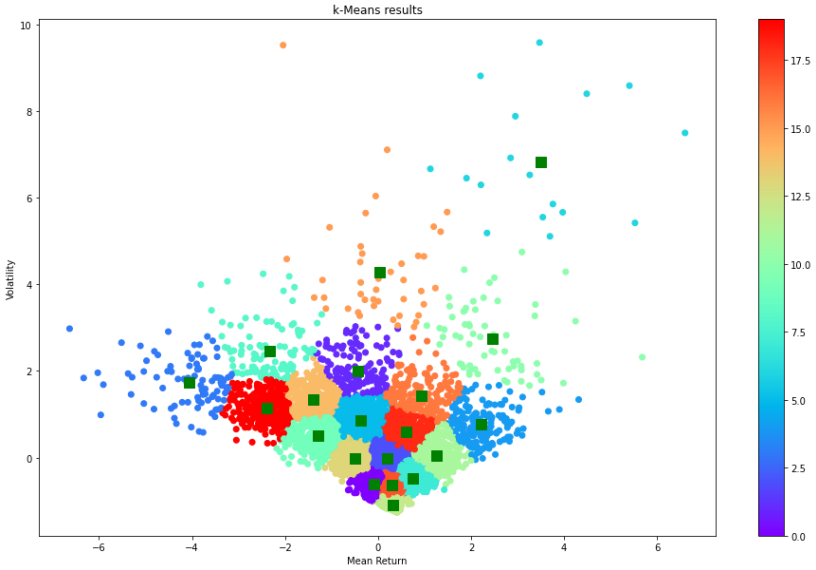
Our data is the daily adjusted close price of more than 8000 assets from 9/17/2021 to 9/16/2022. Downloading a large volume of data from Yahoo Finance is difficult. Instead of downloading the most recent data, we will use the data we previously downloaded.

The features that we use for clustering are the annualized return and volatility of the assets. We build the following models and compare their results:

1. KMeans
2. Hierarchical Clustering (Agglomerative Clustering)
3. Affinity Propagation

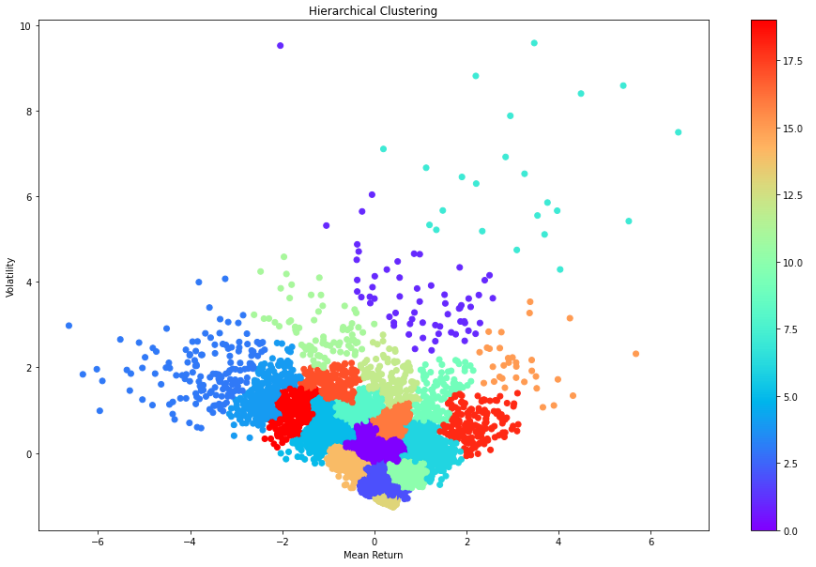
K-Means Clustering

Usually, we should calculate the Sum of square errors (SSE) within clusters or Silhouette score to find the optimal number of clusters. In this analysis, we have assets into 20 sectors. We would like to check if the majority of assets within the same sector will stay in the same few clusters. Therefore, we would set the number of clusters to be 20.



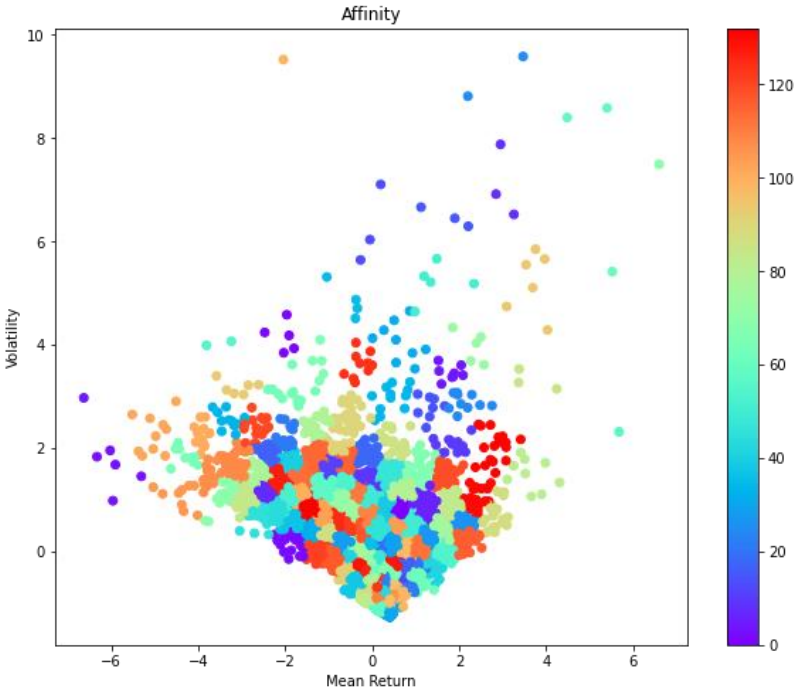
As showed in the graph, assets with similar Mean Return and Volatility are clustered into the same group and each group is enclosed by a clear boundary. The KMean method is doing well in clustering the assets.

Hierarchical Clustering (Agglomerative Clustering)



Similar to the plot of k-means clustering, assets with similar Mean Return and Volatility are clustered into the same group. The boundaries are clear but the shapes are not as regular as the ones in k-means clustering.

Affinity Propagation



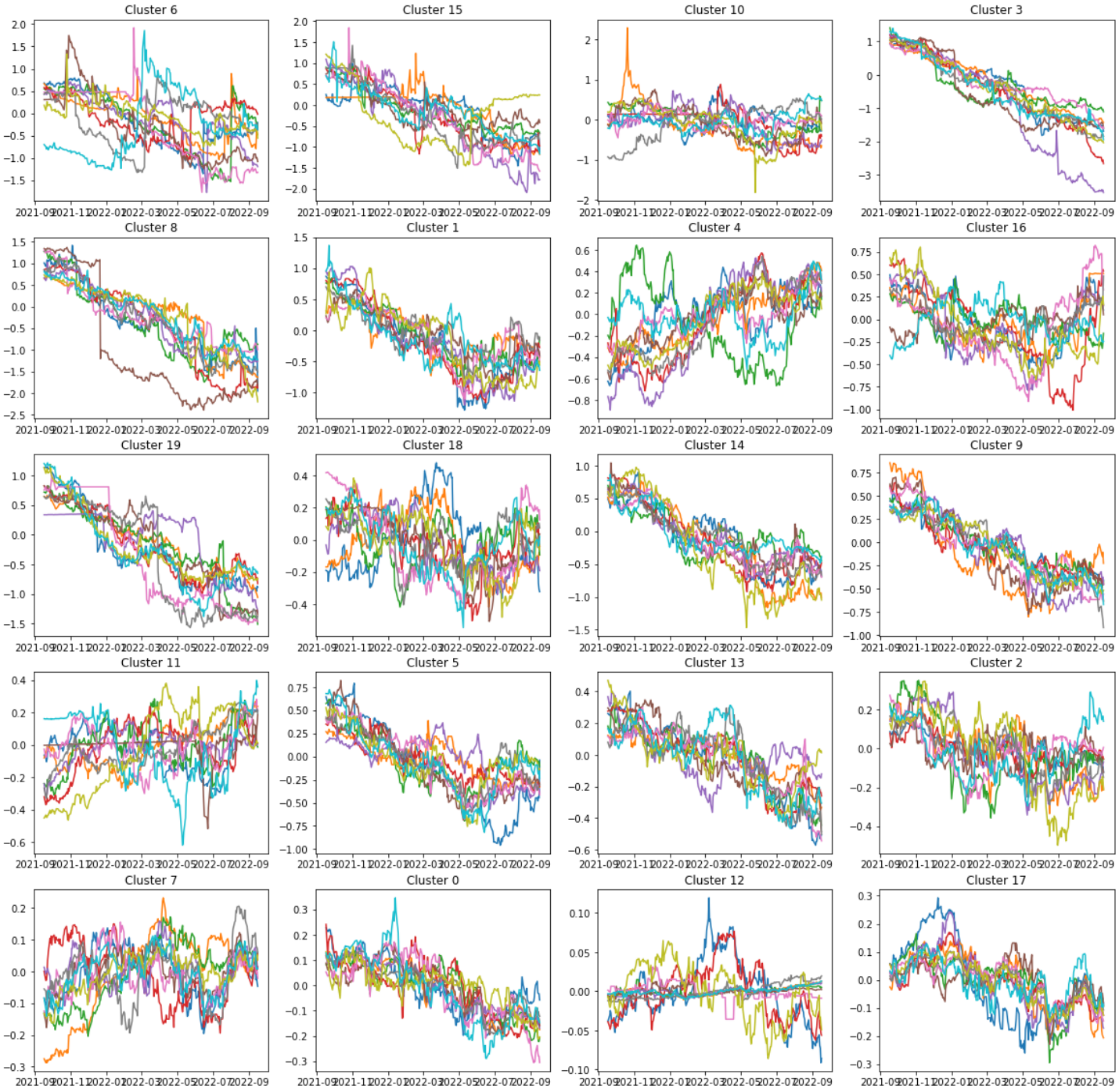
The number of clusters is determined by the Affinity Propagation model. From the graph, this method creates too many clusters, which is not suitable for our analysis. we will focus on either the KMean or Hierarchical Clustering model.

We can compare the Silhouette Score of the KMean and Hierarchical Clustering models to select our optimal model. The following are the Silhouette Scores of the two models.

km 0.3517323801100116
hc 0.30516820009814355

Given the KMean model has the highest Silhouette Score, it is selected as our optimal model.

Now, let us visualize the results of the clusters. We would like to check if the assets in the same cluster have similar price movements. Since the number of assets in some clusters is too large, we only plot 10 assets in each cluster.



Most of the assets within the same cluster have similar patterns, with a few exceptions. The exceptions tell us that assets with similar mean return and volatility may not have the same pattern. We will perform cointegration test to identify pairs within each cluster.

The following plot shows the distribution of clusters within each sector.

Some of the Sectors such as Utilities and Energy Minerals, have majority of the assets staying the same clusters. Other Sectors such as Transportation and Technology Service, have assets distributed into many different clusters. This shows that finding pairs within sectors may work for some sectors, but not for the others.

From more than 8000 assets, we identify more than 440,000 pairs. 42.5% of the pairs that have both assets in the same sector. We can see that the strategy of finding pairs within each sector can still find a considerable number of pairs. However, the method fails to find the pairs across sections. The method of using clusters is doing good in finding inter-sector pairs and cross-sector pairs.

We also find the strength of the relationship between two sectors by checking the number of cross-section pairs. For example, HEALTH SERVICES has the highest number of pairs with HEALTH TECHNOLOGY. The MISCELLANEOUS sector, which contains mainly ETFs, has a strong relation with the FINANCE sector. The

FINANCE and MISCELLANEOUS sectors also have strong relations across all other sectors, which is reasonable since all industries need to be funded.

The ML Clustering method does a great job in identifying inter-sector and cross-sector pairs. It also help us finding some relations between sectors.