Stock Market Prediction

Amod Sahasrabudhe
Northeastern University
Boston, MA
sahasrabudhe.a@northeastern.edu

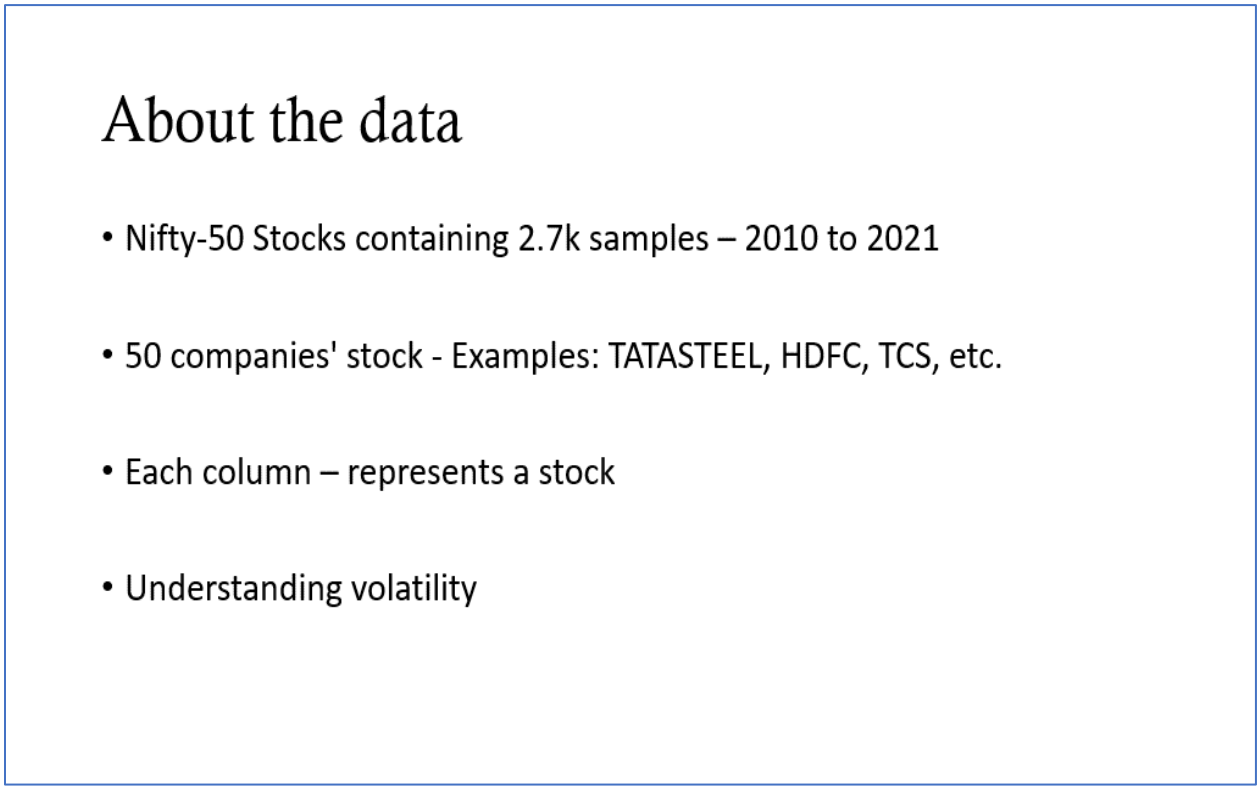
Grania Machado
Northeastern University
Boston, MA
machado.g@northeastern.edu

Rebecca Quay Dragon
Northeastern University
Boston, MA
dragon.r@northeastern.edu

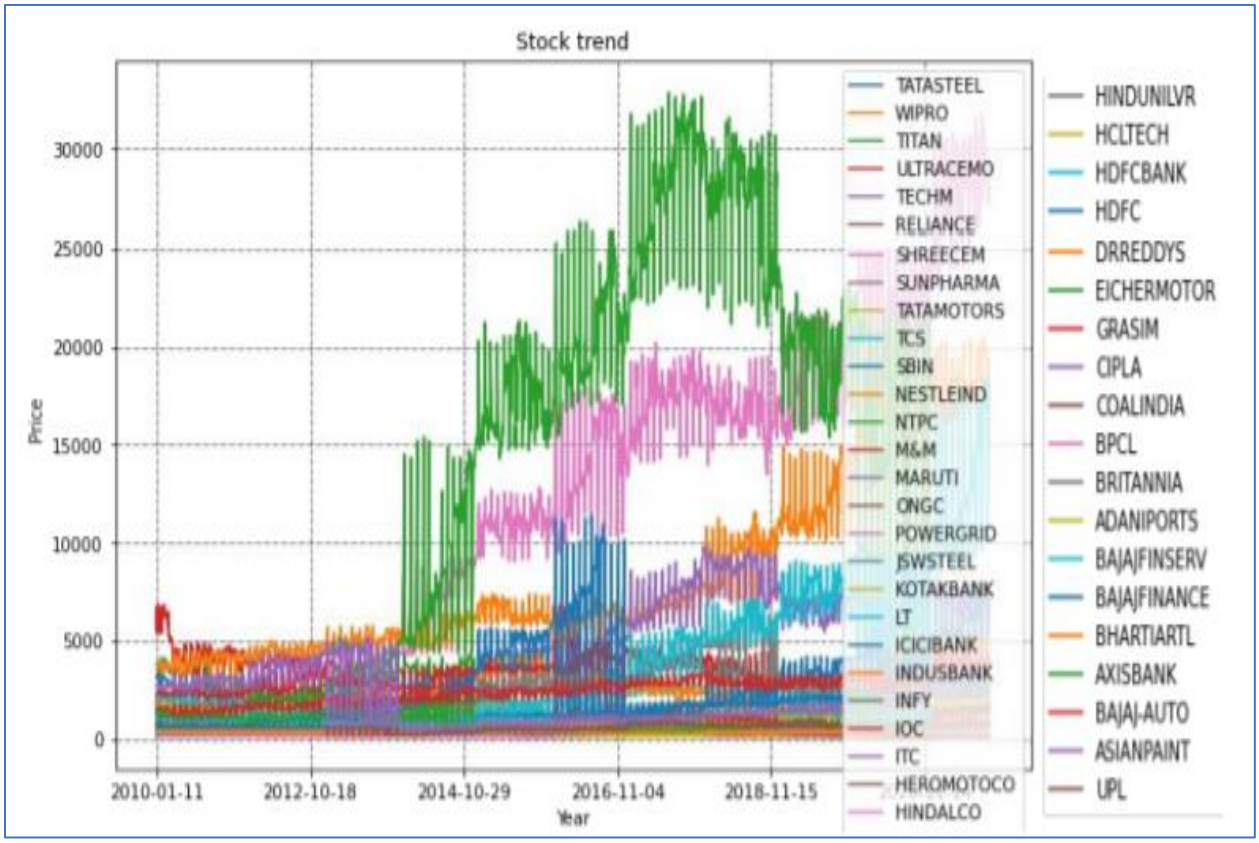
Introduction

- Stocks - Indicator of a nation's economy
- 80% of the trading through algorithms
- Machine learning algorithm
- Clustering - better classification and prediction

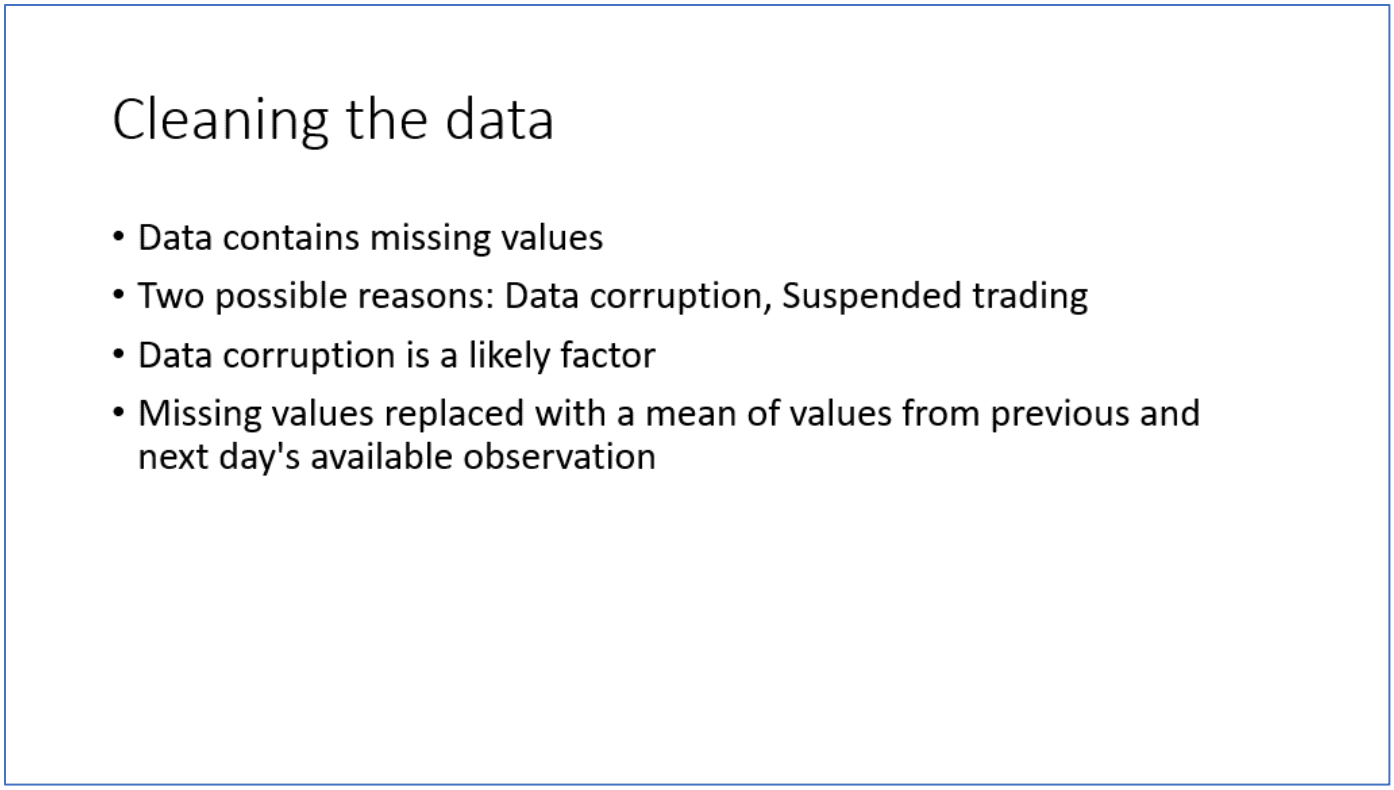
The stock market, specifically the top 50 stocks, represents not only safe, proven investments but also indicates how the economy is performing. There are countless investment firms which use algorithmic trading to bolster their portfolios. Today, 80 percent of trades are done by algorithms. Many individuals and firms attempt to implement machine learning algorithms to better classify and predict stock movements and pricing. These machine learning algorithms seek better investment outcomes compared to that of competing investors. A large problem machine learning engineers face when building algorithms for stock trading is overfitting. We hope that clustering our data will help us differentiate the diverse types of equities and their movements to develop algorithms which correctly match the clustered groups. This will reduce the overfitting problem if new equities are evaluated into the correct cluster prior to using a predictive algorithm such as an unsupervised naïve bayes. This approach, along with attention to parameter tuning, will allow our team to make competitive, well-informed algorithms.



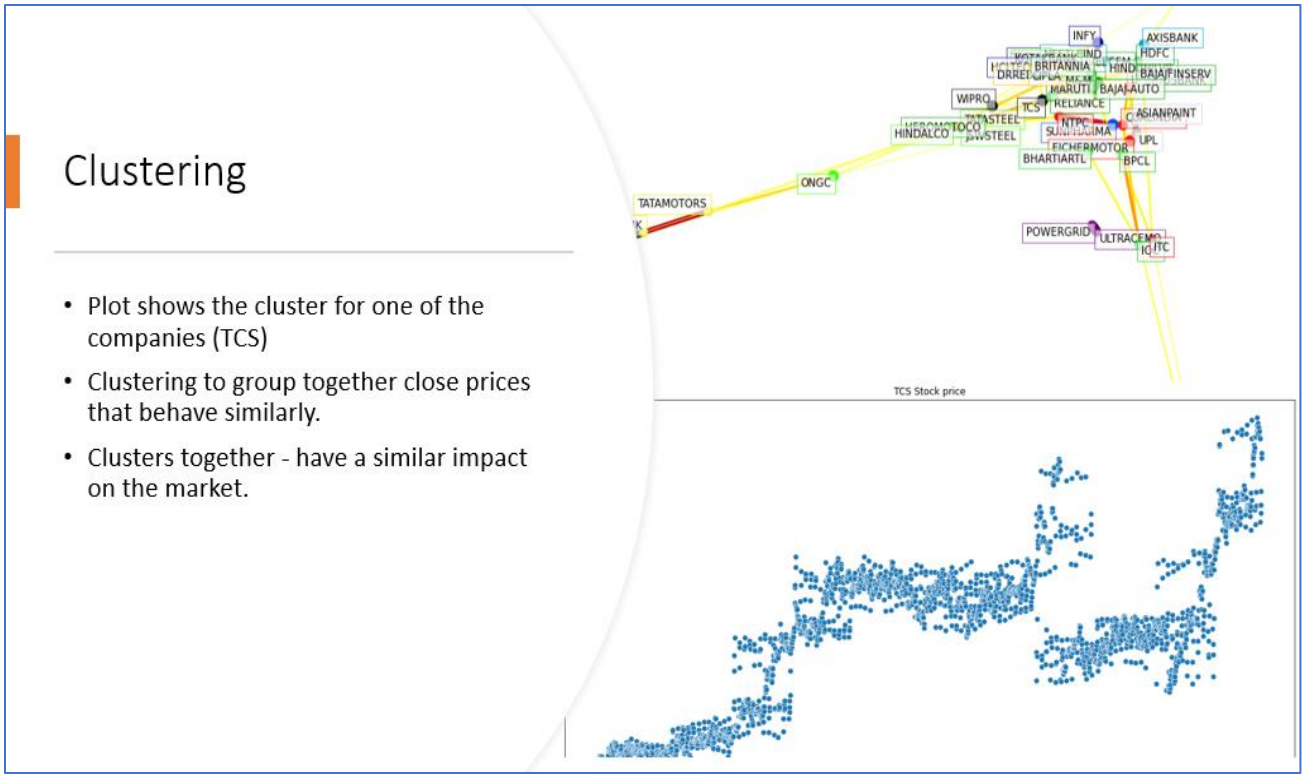
The dataset contains about 2.7k samples for each NIFTY stock from the year 2010 to 2021. Each column represents a NIFTY-50 stock. Examples of such stocks are: HDFC, TCS, TATASTEEL, etc. The data represents the stock's closing price as the market fluctuates during the day. We will be using EDA to understand the overall trends of the top 50 stocks to ensure when we compare these to our initial clustering algorithms, we are finding comparable results regarding upwards or downwards trends between stocks over time. We have also evaluated the mean and standard deviations of these top 50 stocks to understand the volatility. Volatility is a key factor when evaluating potential gain from quick upward shifts and dips in the market.



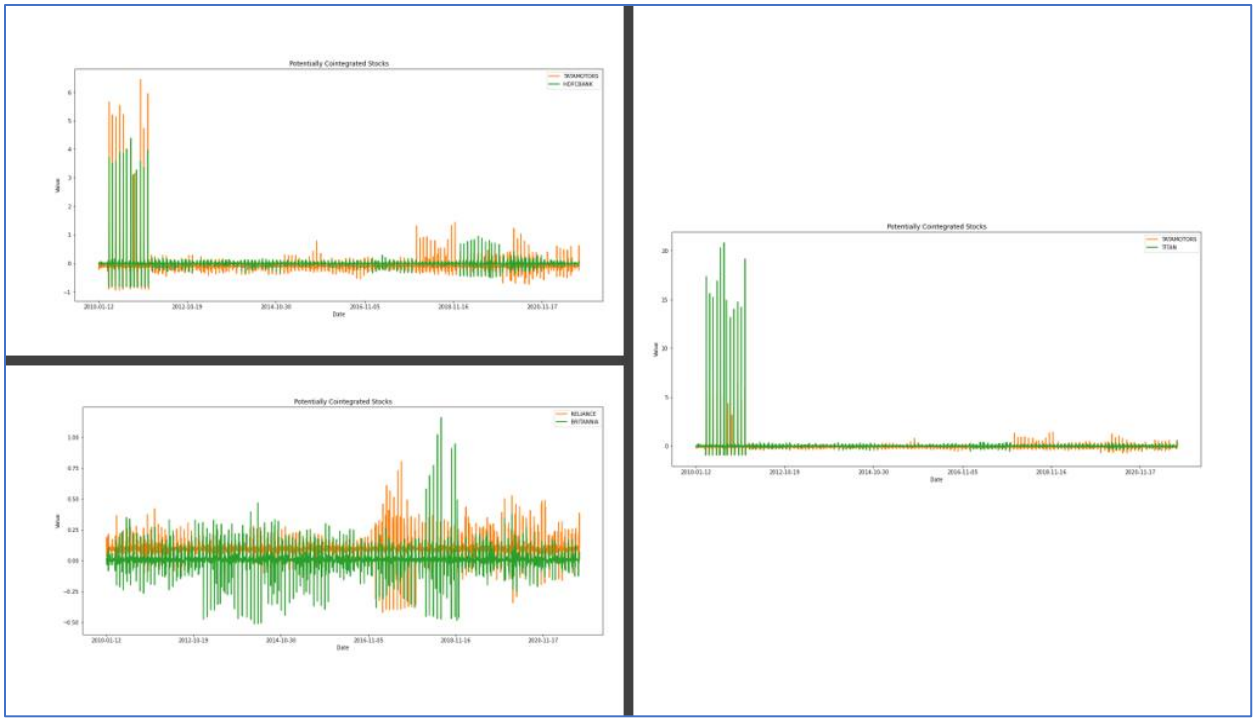
We are currently testing k-means with dynamic time warping to model the price fluctuations of the 50 stocks. Dynamic time warping is measuring the similarity between two sequences of time. This allows us to model the stock fluctuations based upon stock within the same predetermined cluster.



During the initial look of the data, we realized that our data contained missing values, and further inspection of the stocks led us to suspect that this is possible on two accounts: data corruption and suspended trading activity. Data corruption is a likeliness during data extraction and documentation while company acquisition, massively fluctuating stocks might be certain reasons for stock trading suspension.



The plots below are part of our initial EDA (Exploratory Data Analysis) as well as our first attempt using a clustering algorithm to identify trends within the time series. The first plot shows a cluster of the stock prices for one of the companies: TCS is plotted. We can see in this plot the individual volatility of a stock. This is important to our analysis because similarly volatile stocks are likely to behave similarly to external market stocks.

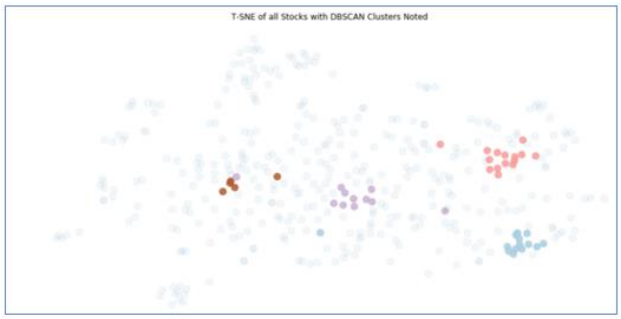


We also observe SHREECEM, NESTLEIND, EICHERMOTOR behave similarly while maintaining high asset prices. This trend may be substantiated by further clustering. This interaction AXISBANK and BHARTIARTL (similar pricing from 2016-2020) is interesting because of the lack of correlation earlier on in the time series. Hopefully, the dynamic changes within the market can be captured by adding data about external market conditions so that we can better understand how various stages within each company's existence are influenced and therefore correlated through a currently unknown variable

DBSCAN

Density-based spatial clustering of applications with noise

- Clusters stocks and exclude the stocks that do not fit into a cluster
- Gives sensible clusters of common stocks for candidate pairing which can be used for validating relationships



Density-based spatial clustering of applications with noise (DBSCAN) can be used to cluster stocks and exclude the stocks that do not fit into a cluster. The basic difference of density-based technique with that of partitioning technique is that it is not based on distance, but on density. The two parameters for DBSCAN are eps (how close points should be to each other to be considered part of a cluster) and minPoints (the minimum number of points to form a dense region). From these parameters, the DBSCAN algorithm then creates clusters from the set of points we feed it. Points that are in low density regions are classified as outliers. We have found 4 clusters. As an attempt to visualize what has happened in 2D, we tried with T-SNE, which is an algorithm for visualizing very high dimension data in 2D. Visualizing the discovered pairs helped us gain confidence that the DBSCAN output is sensible, and this can be used in validating relationships. To understand the stocks pairing relationship, we visualize the cluster to observe how these pairs of stocks have performed in a similar pattern.

K-Means DTW

- K-Means Dynamic Time Warp was used to represent clusters with greater than 2 stocks in a time series.
- DTW compares elements in a series allowing us to view K-Means as a time series.

- Dynamic Time Warp:

$$DTW(x, y) = \min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(x_i, y_j)^2}$$

K-means paired with the dynamic time warping algorithm allows us to use k-means as a time series modeling algorithm which applied to stock prices to demonstrate and capture their behavior over a period. K-means works as usual by clustering each of the data points to a mean value which best suits its assigned cluster. The dynamic time warping changes the k-means algorithm by finding the optimal match between the sequences in a series. The common Euclidean distance used by k-means alone is not suitable for a time series because it does not evaluate the similarity between two temporal sequences. By comparing the two temporal sequences, such as comparing the points within two features at the same index the algorithm builds a “warped” path that both features take. The number of K clusters defined in the algorithm results in the number of paths which the series will separate into.

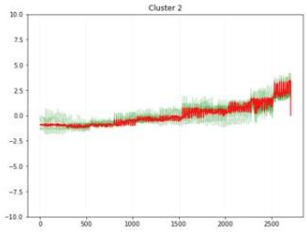
Results of K-Means DTW- Highlights

Red: Predictive
Green: Actual
Movements

K-Means DTW Cluster 2 for:

RELIANCE, SHREECEM, M&M, MARUTI, INDUSBANK, HEROMOTOCO, HDFC, BPCL, BRITANNIA, BAJAJFINSERV, BHARTIARTL, BAJAJ-AUTO.

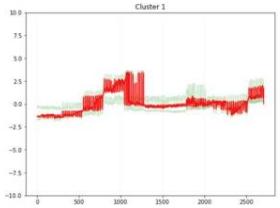
DTW < 50



K-Means DTW Cluster 1 for:

WIPRO, TECHM, TCS, INFY, HCLTECH

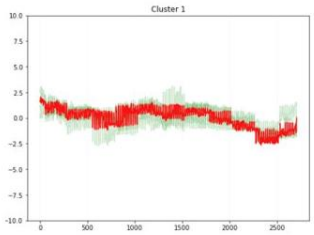
DTW < 50



K-Means DTW Cluster 1 for:

NTPC, LT, IOC, EICHERMOTOR, COALINDIA, ASIANPAINT

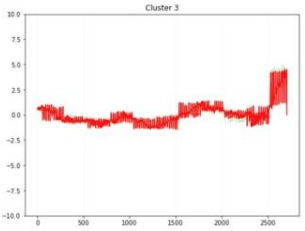
DTW < 50



K-Means DTW Cluster 3 for:

TATASTEEL, ONGC, JSWSTEEL, KOTAKBANK, HINDALCO

DTW < 50

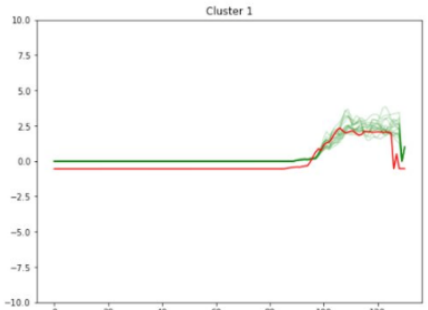


In the above slide, we show the K-Means DTW applied to clusters determined by our DBSCAN. Having a DTW measure of 0 means that all the series being measured are the same. Our DTW measures are below 50. Although this distance seems high this measurement is bound only by the distances between each series. Therefore, for price movements we can assume that a DTW below 50 is accomplishing the pattern modeling we were striving for. This is further substantiated by the predicted and actual movements of the stocks in the four graphs. It is important to note that this specific algorithm can only be applied to clusters with greater than two sequential members.

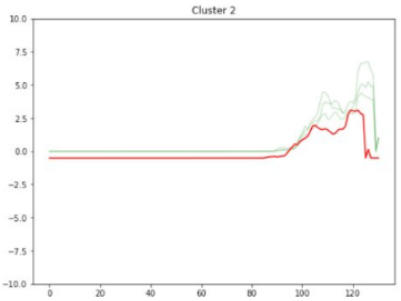
Expanded Stock Market Research 8k+ Stocks

- 8,000+ Stocks were pulled from the Yahoo API.
- First Clustered Using K-Means without DTW into 2 Clusters.
- DTW measurements range from good <20 to <80.

Red: Predictive
Green: Actual Movements

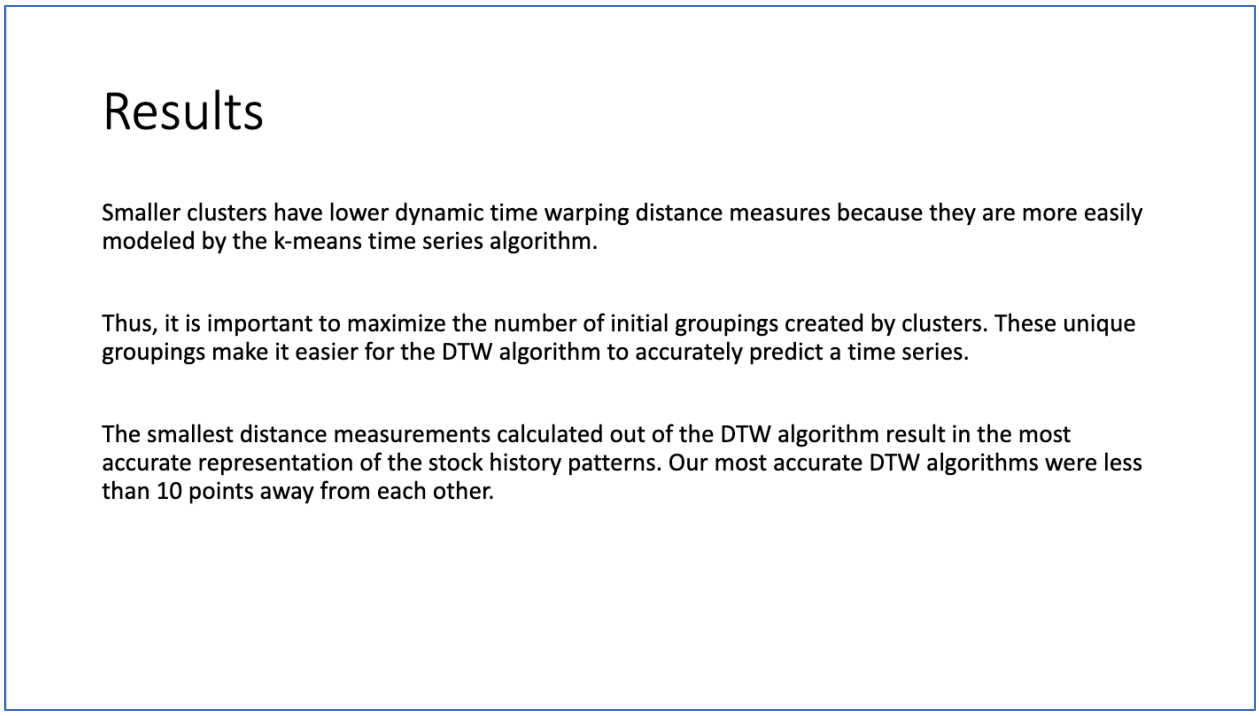


Group 14



Group 8

Our expanded research sought to apply the same clustering and techniques to 8,000 stocks pulled from the yahoo finance API. The Silhouette scores were higher for these clustering's, and we were able to accurately model the stock movements of 20 predefined stock groupings each into 5 different time series clusters created by the DTW algorithm. Interestingly, it was better to get lower clustering scores when first segmenting the 8,000 stocks into 20 distinct groups than to maximize the silhouette scores of these pairings. This is a tradeoff that was unforeseen in our initial analysis. Since the first k-means algorithm feeds into the second k-means DTW algorithm we have a tradeoff created between accurately predicting the clusters and creating small enough groupings that the DTW algorithm can accurately find a unique paths for each series.



We found our most accurate results modeling the stock history in our group 8. Our distance measurements were less than 15 points for each of the DTW clusters group 8 was allotted to. Our highest distances using this method were in the low 70s. For stock prices, this is still an accurate representation of those assets in the market.