Project: Pairs Selection

Milestone 2

Team Members: Kareema Batool, Nikhil Nayak, Nishtha Sardana, Saket Joshi

Introduction

Pairs trading is a market-neutral strategy where we use statistical techniques to identify two stocks that are historically highly correlated with each other. When there is a deviation in the price relationship of these stocks, we expect this to be mean reverting and buy the underperforming stock and simultaneously sell the outperforming one. If our mean-reversion assumption is valid then prices should converge to long term average and trade should benefit. However, if the price divergence is not temporary and it is due to structural reasons then there is a high risk of losing the money.

Problem Statement

Data - We have taken NIFTY-100 data for the past 3 years on a day level basis for our analysis. However, we have considered the companies who have complete data for our selected tenure and escaped any that were listed. For each of these companies, we have taken the 'closing value' for our further analysis. However, we have eliminated businesses that lack data for the chosen tenure, leaving only the top 88 stocks. For these 88 equities, we use the closing price on a daily basis.

In [4]:

```
A quick view of our cleaned dataset for 5 stocks
```

Out[4]:

	BHARTIARTL	TATASTEEL	BOSCHLTD	TITAN	INDUSTOWER
0	518.15	72.20	19791.90	852.45	369.95
1	507.05	72.95	19721.75	845.15	378.80
2	508.65	73.50	19692.95	856.30	378.65
3	513.35	76.00	19652.65	892.90	379.05
4	530.05	77.05	19693.15	909.70	372.60

Exploring the correlation and cointegration between the pair of stocks

As a first step, we start by exploring the highly correlated pairs in our dataset of 88 companies.

We have leveraged the Pearson Correlation Coefficient to gain a general understanding of the relationship between these companies before trying to explore and locate cointegrated stocks.

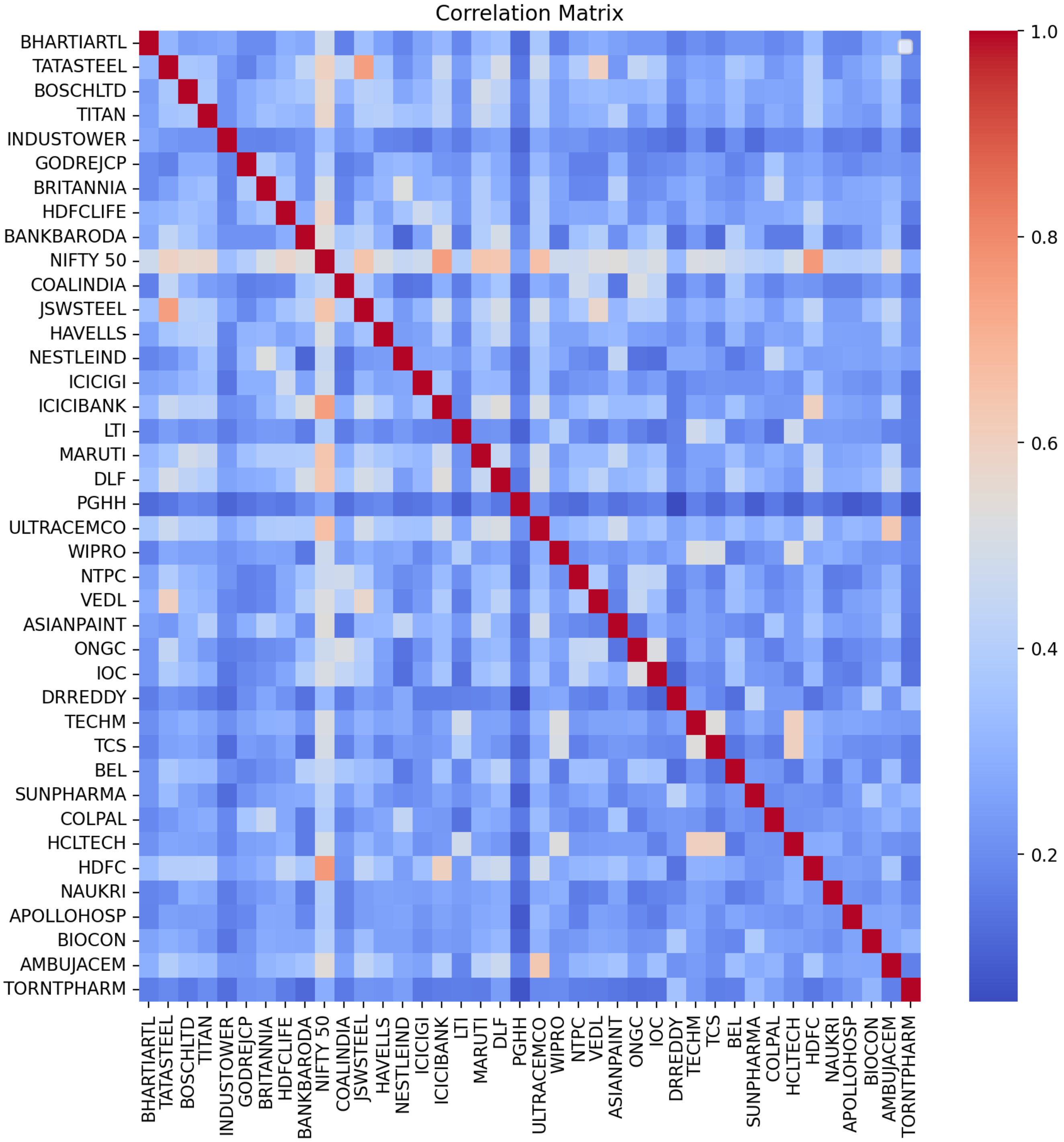
Pearson correlation coefficient varies between +1 to -1 and is a linear measure of the relationship between two variables. The value +1 indicates a strong positive correlation, zero indicates no relationship, and -1 indicates a strong negative relationship.

We can observe from the below generated heatmap that there are multiple pairs with a strong positive correlation.

In [38]:

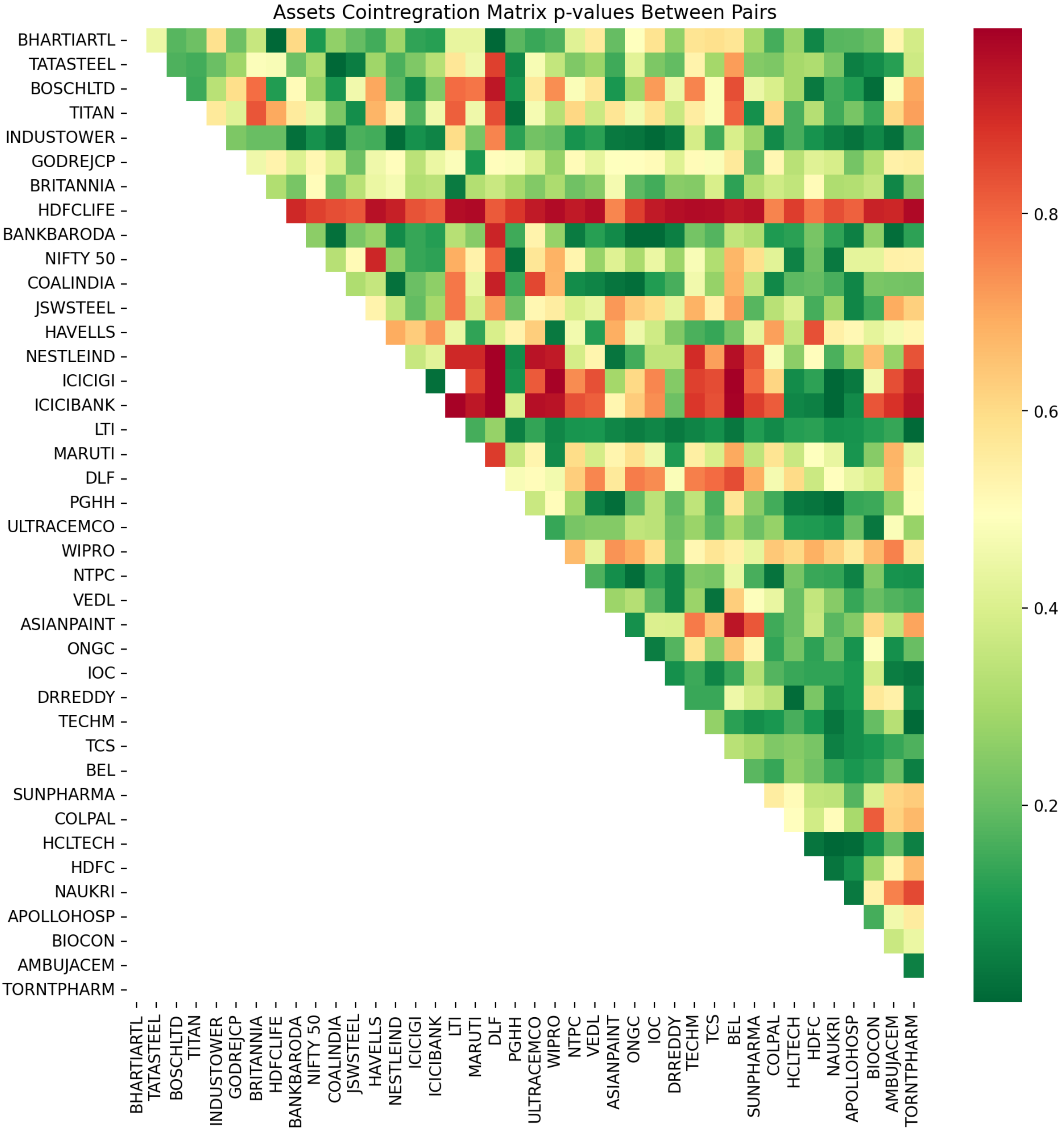
```
# We show correlation values for 40 sample stocks below
```

```
No artists with labels found to put in legend.  Note that artists whose label start w
ith an underscore are ignored when legend() is called with no argument.
```



In [36]:

```
## We show cointegration values for 40 stocks below
```



We have further defined a function to calculate p-values of the cointegration test for each pair, to generate a heatmap. If the p-value is less than 0.05, the null hypothesis can be rejected and the cointegration between the two time series with distinct symbols.

We can see in the below heatmap that there are many pairs with a p-value of less than 0.05. This means that for these pairs we can reject the null hypothesis and they can be cointegrated.

Baseline Model

Baseline model - Pair trading is one of the most popular trading strategies since 1980 for finding statistical arbitrage opportunities in the stock market. The logic behind pairs trading is to trade pairs of stocks belonging to the same industry or having similar characteristics, such that their historical returns move together and are expected to continue to do so in the future. In this project, we will use machine learning/deep learning methods to find statistical arbitrage opportunities in the stock market using pair trading strategy. The baseline model recognizes correlated pairs using clustering methods.

Approach in the baseline model for pairs trading is to apply PCA for dimension reduction on a large set of features (different stocks) in order to ease computation of unsupervised clustering algorithms like DBSCAN/t-SNE for clustering of similar stocks. Following which, we use cointegration tests to extract all possible combinations of stocks in each cluster that are within 5% significance level (p-value is 0.05). These pairs of stocks are the final predictions of the model which are correlated enough for pair trading.

Specifications of the baseline model:

1. PCA is a mathematical procedure that transforms a large number of variables into a smaller number of uncorrelated variables called principal components. In the baseline model, PCA is used to reduce daily stock prices of 88 companies to 50 variables while trying to keep as much variance as possible. Each of the resulting principal components can be seen as representing a risk factor, and the stocks will be clustered based on these components to find the stock pairs.
2. Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering non-parametric algorithm. We preferred DBSCAN over K-Means because it has a couple of important advantages. Specifically, DBSCAN does not cluster all stocks, i.e. it leaves out stocks which do not neatly fit into a cluster, and the number of clusters does not need to be specified.
3. t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique. Once the data is clustered using DBSCAN, we use t-SNE to visualize the high dimensional data and its clusters in a two-dimensional graph.

Results of the baseline model & updated problem statement We used 1216 rows of data i.e. 1216 consecutive days of closing price for 88 companies. We used PCA with 50 components. Following which 3 clusters were discovered using DBSCAN algorithm. t-SNE is mainly used for visualization in 2 dimensions. Following 3 clusters were formed:

1. Cluster 0 - BRITANNIA, NESTLEIND, COLPAL, HINDUNILVR
2. Cluster 1 - ASIANPAINT, BERGEPAINT, PIDILITIND
3. Cluster 2 - TCS, HCLTECH, INFY

Pairs within each cluster were investigated for correlation using cointegration. Cointegration tests identify scenarios where two or more non-stationary time series are integrated together in a way that they cannot deviate from equilibrium in the long term. The tests are used to identify the degree of sensitivity of two variables to the same average price over a specified period of time. Thus, it can be used for finding stock pairs that are correlated in the long term. Among the 88 companies 3 pairs were found to be significant using a p-value of 0.05. Pairs found were - ('TCS', 'HCLTECH'), ('TCS', 'INFY'), ('HCLTECH', 'INFY').

***Updated problem Statement**

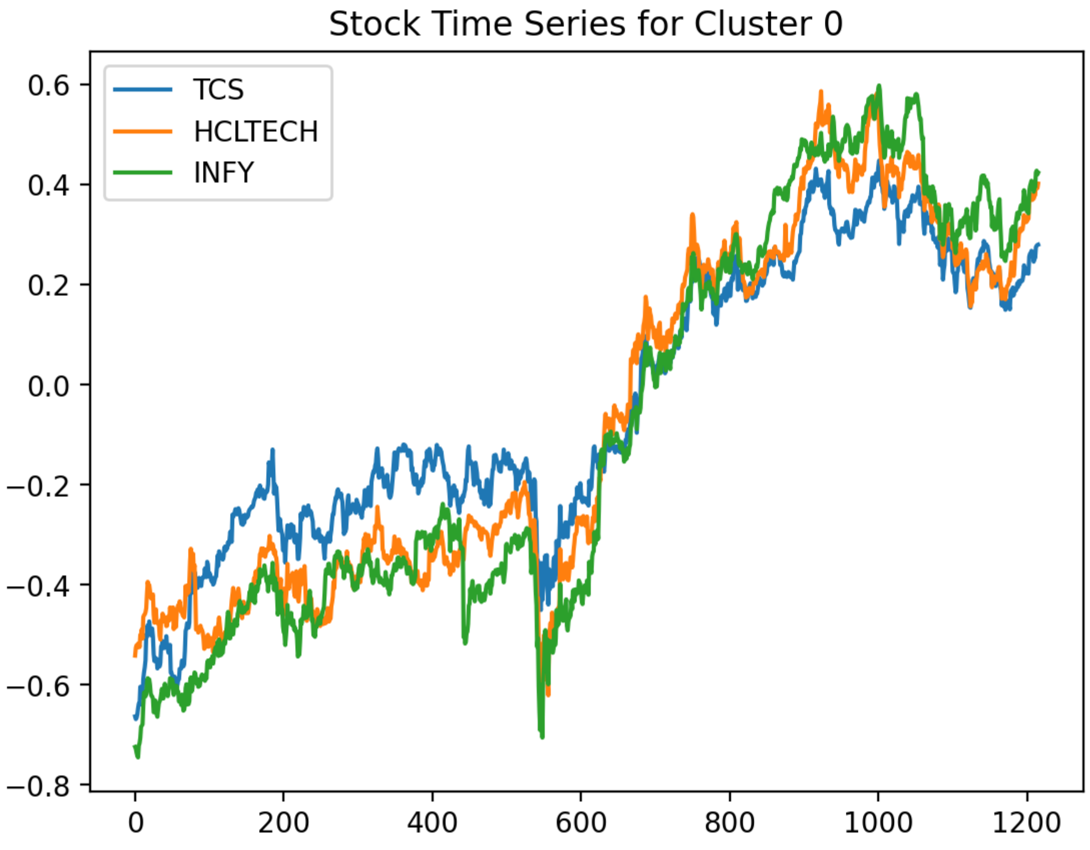
1. The baseline model tells if two stock prices are closely related, but in the next milestone we intend on predicting the value/gains of investing in these pair stocks in a specified long term interval.
2. We will investigate different base estimators and clustering based architecture to minimize hypothesis testing problem, which occurs when several tests are done due to a large number of stock pairs.
3. We will define and evaluate strategies to use the correlation / cointegration pairs identified for trading. One of such strategies will be mean-reversion. These will be baselined and compared against different pairs identification architectures.

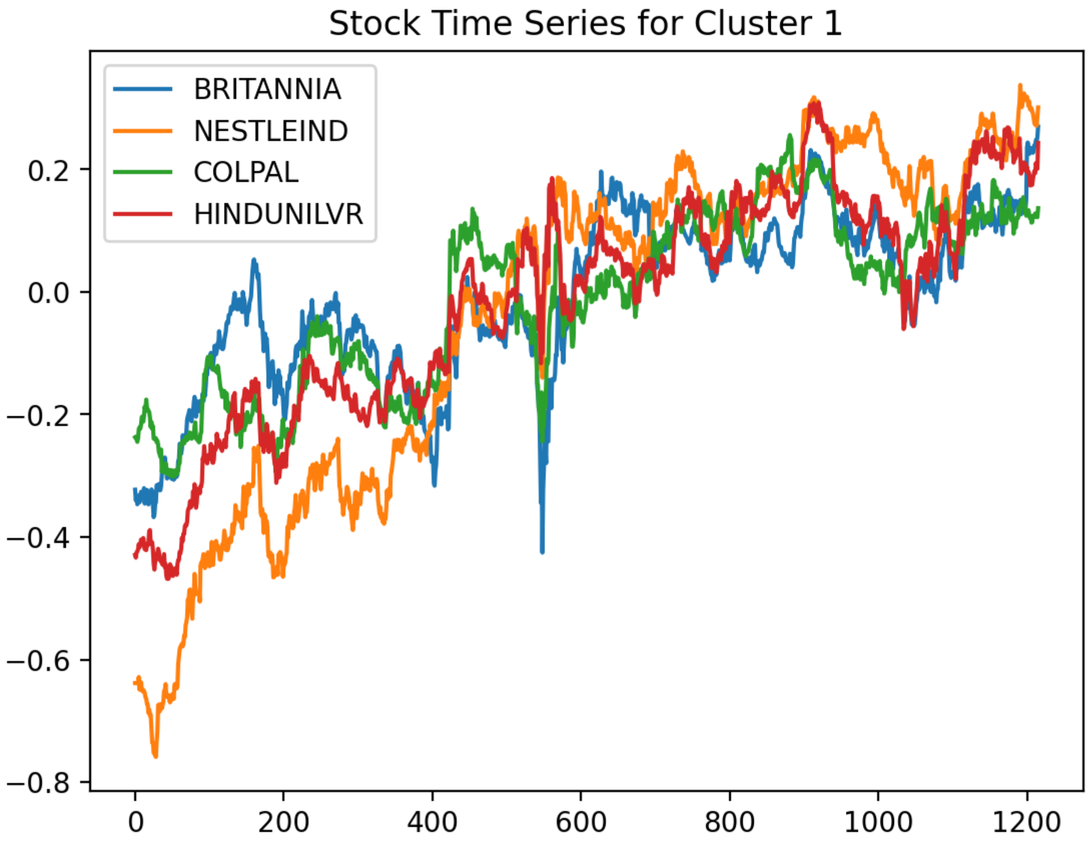
`Code for Baseline model submitted separately`

In [26]:

```

```





References

We have used the following resources as reference for our project:

- http://stat.wharton.upenn.edu/~steele/Courses/434/434Context/PairsTrading/PairsTradingGGR.pdf
- https://hudsonthames.org/employing-machine-learning-for-trading-pairs-selection/

- https://cs230.stanford.edu/projects_fall_2018/reports/12446738.pdf