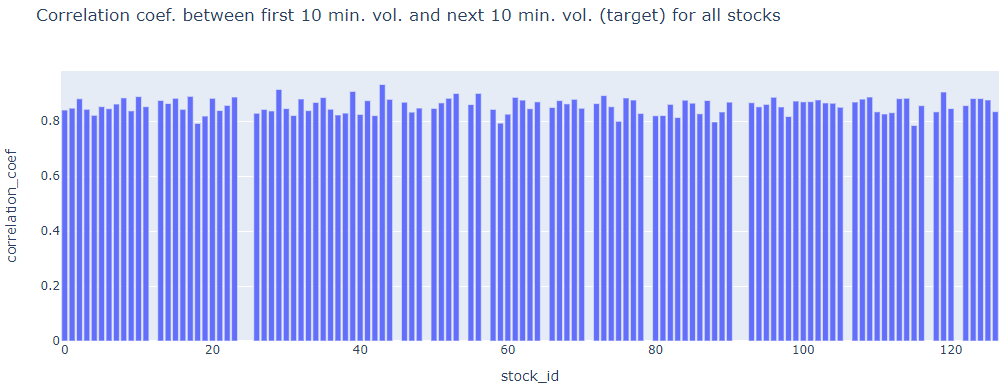
# **Key Insights that are important**

Record Key Insights that cannot be missed!

## 1) Correlation between real. Vol in first 10 mins and next 10 mins (target) for all stocks

****

## past 10 min real vol is a good predictor of future 10 min real vol.

##### **Check correlation between**

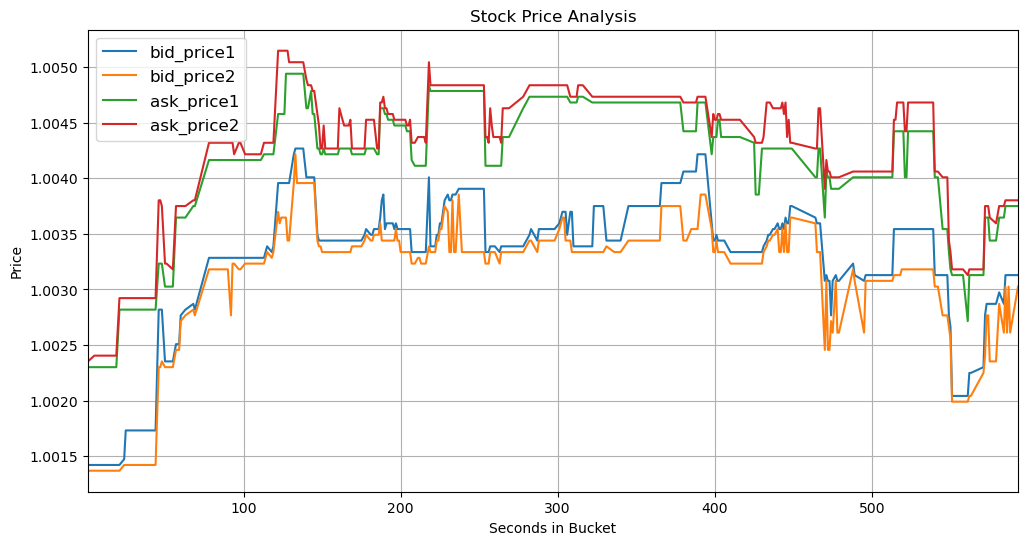
##### **i) target (2nd 10 mins vol.) and trade execution stock price at the available times.**

Aggregate **trade execution price using the log returns and realised volatility formulas for each time id. Plot scatterplot this against the corresponding target for all time id.**

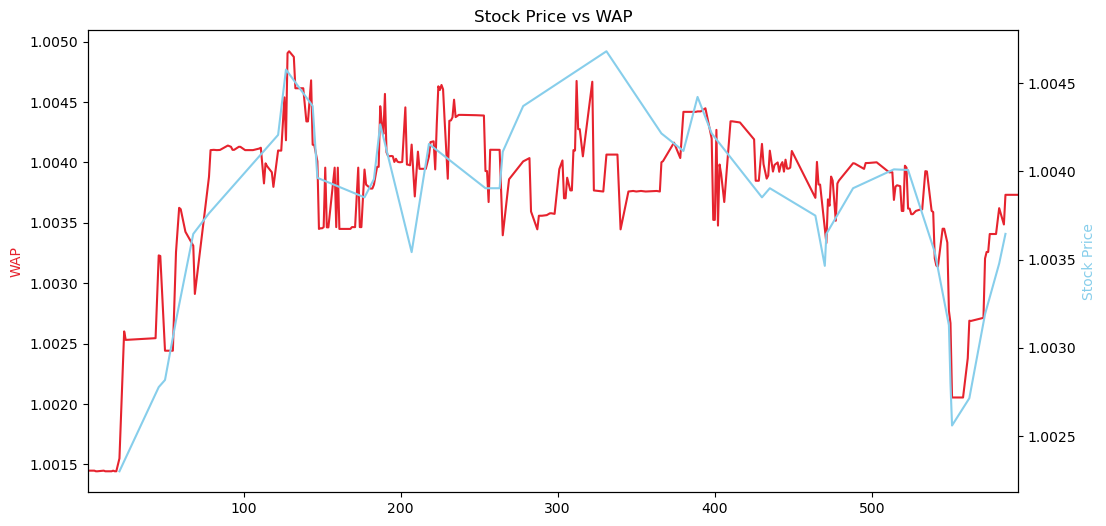
****

****

## 2) Bid Price VS Ask Price (For Stock 0 at time id 5):

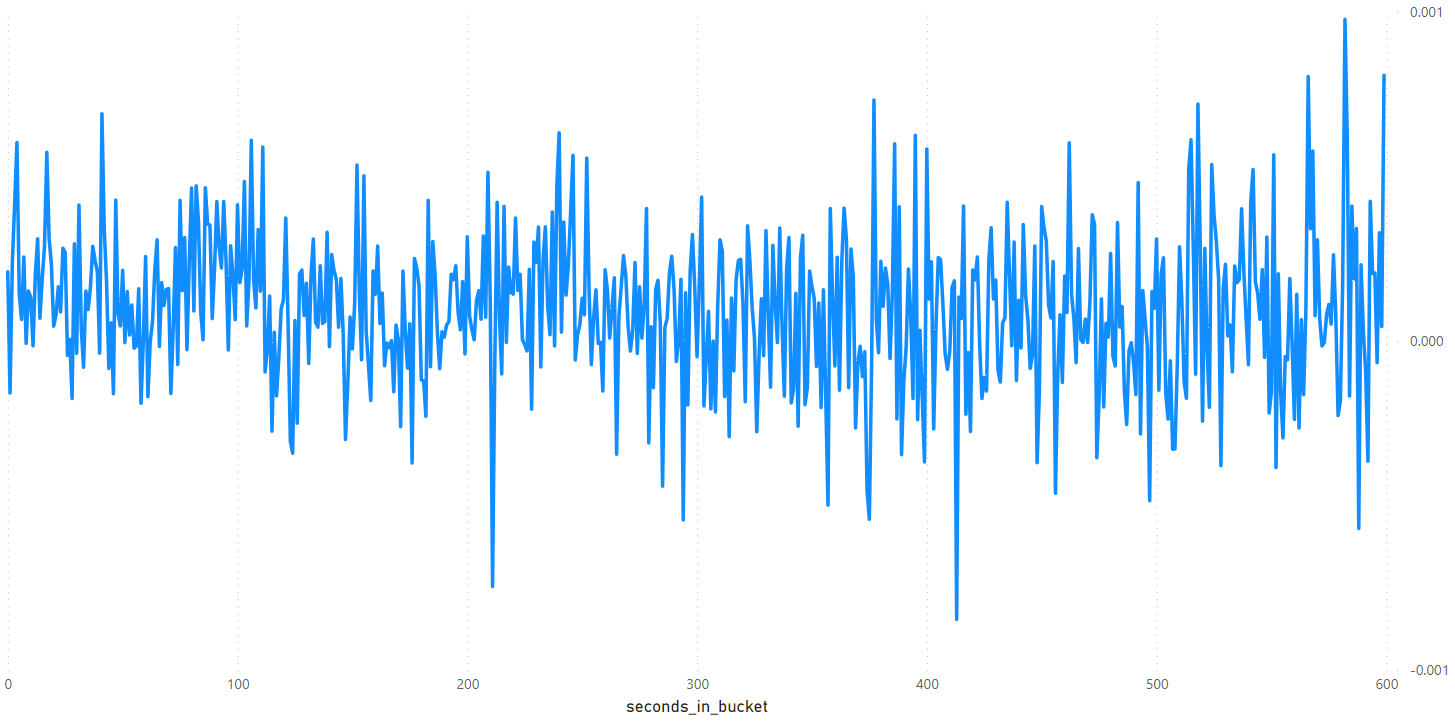
****

## 3) Stock Price VS WAP (For Stock 0 at time id 5):



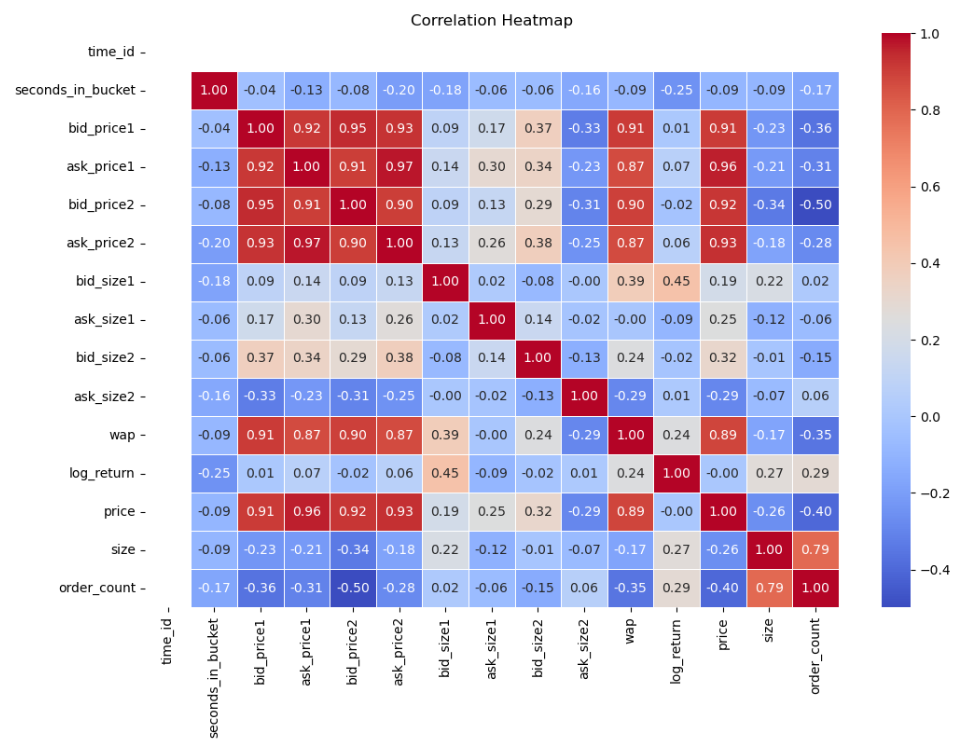
* Bid price and ask price form the lower and upper limit for WAP. WAP is correlated with the actual trade execution price (stock price) as shown above. Could we say that high volatilities in calculated WAP could indicate a high volatility in stock price?

## 4) Difference between Stock Price change and WAP for stock id 0 for aggregated all time id (Please verify this):



This represents the mean difference between the aggregated average of WAP across all time IDs and the price for stock ID 0. As we, can observe there is not much difference between both trends so we can safely say that the **correlation between WAP and Stock price (from trade data) is High.**

## 5) Correlation Heat Map (For Stock 0 at time id 5):



The correlation heatmap presented above reveals that our stock demonstrates a robust positive correlation with [bid\_price1, ask\_price1, bid\_price2, ask\_price2, wap], with an approximate 90% correlation coefficient. Conversely, there is a notable negative correlation of around 40% with the [order\_count] variable.

## **6) Target realized volatility clustering analysis**

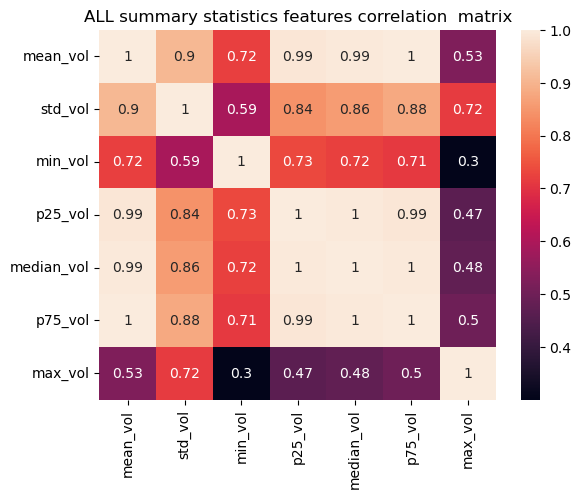
**Analysis file :** target\_eda\_across\_stocks.ipynb

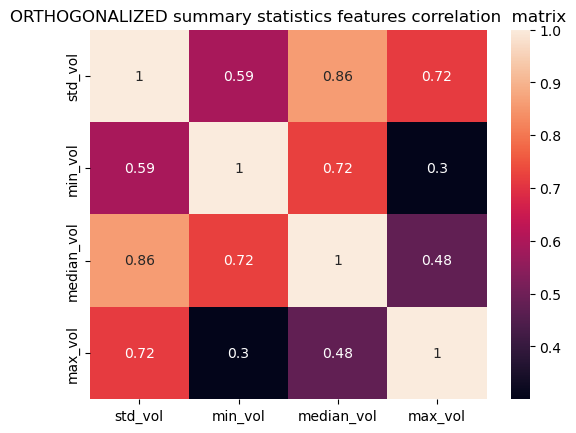
**Objective:** Group stocks that are similar to each other. The stocks that are similar can help to predict each other when used as covariates.

**Clustering based on :**

1. Target realized volatility Summary statistical features (i.e. Distributions) : [mean\_vol,std\_vol,min\_vol,p25\_vol,median\_vol,p75\_vol,max\_vol]
2. Temporal target realized volatility correlation : pearson correlation between stocks for all time\_ids.

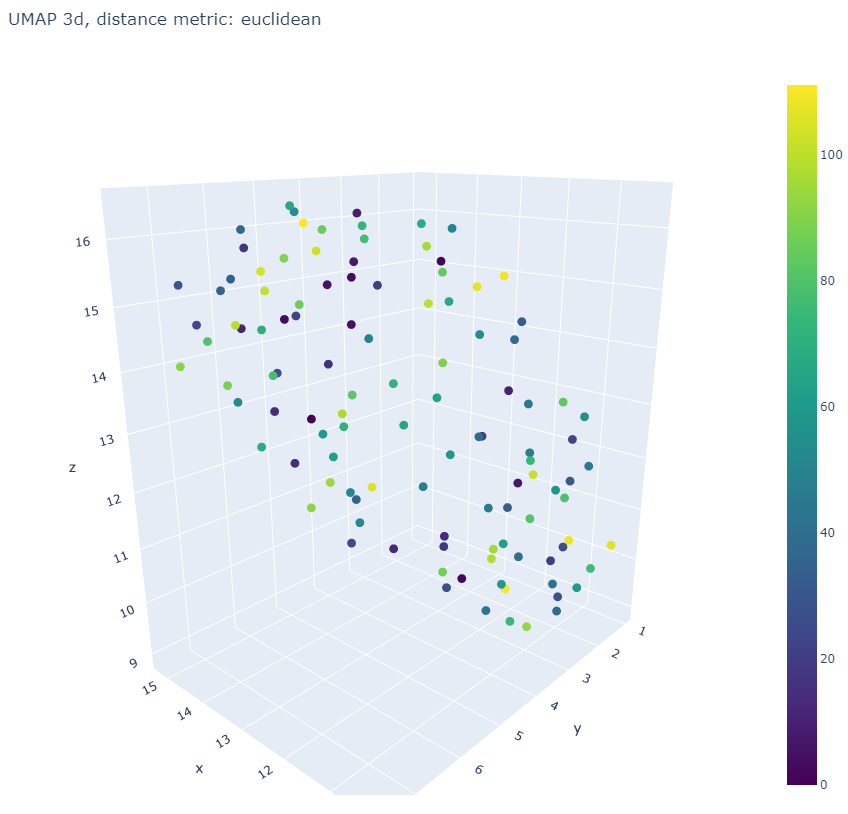
**Clustering on Summary statistical features**

1. Perform feature orthogonalization by removing correlated features 

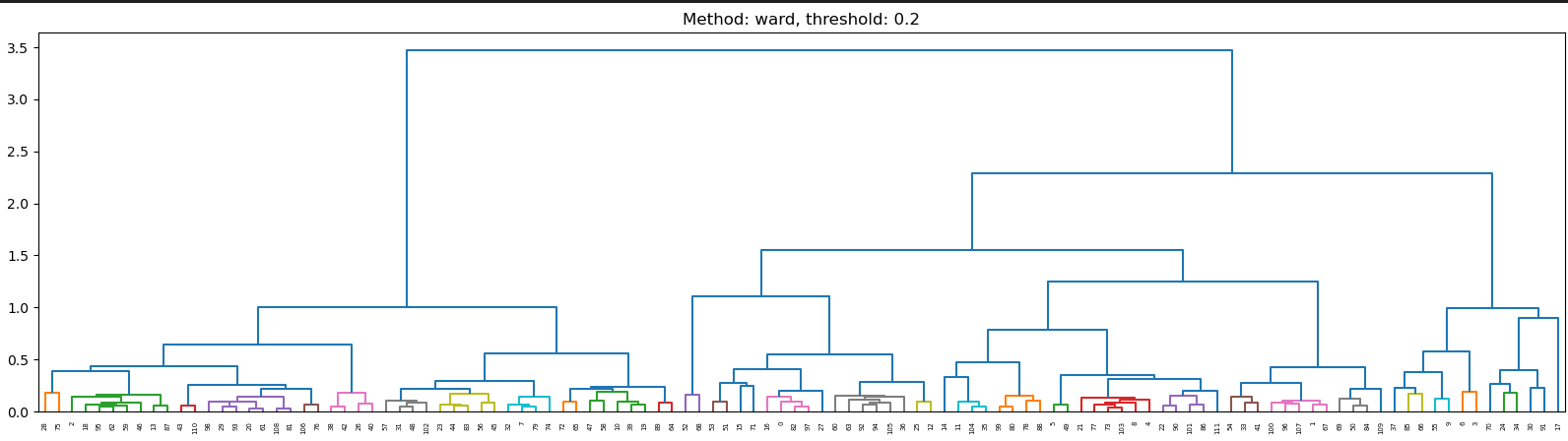


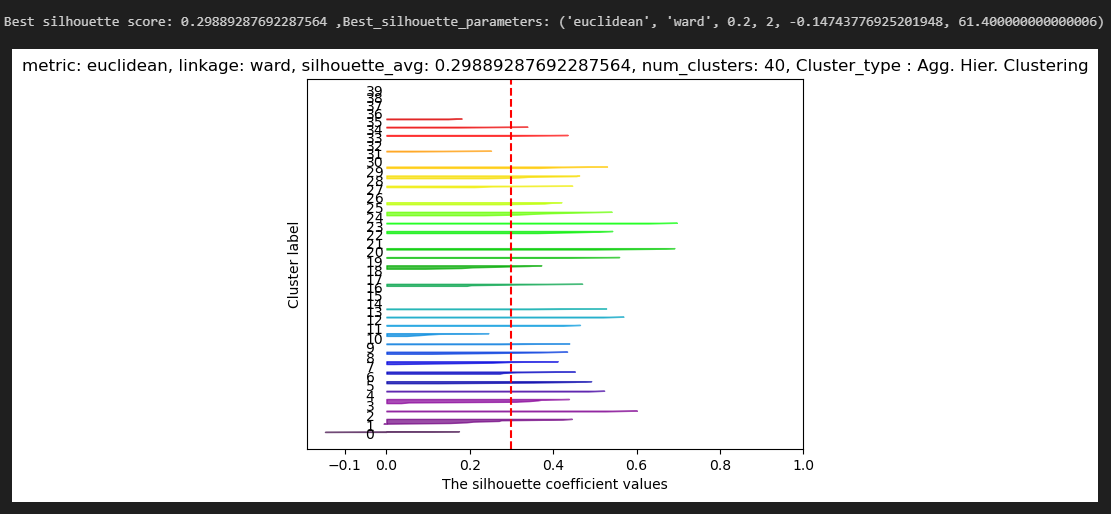
For large matrices PCA can also be used for feature orthogonolizaiton at the cost of losing intrepretatvility of principal components

1. Perform Dimensionality reduction to visualise how data is distributed in order to choose the appropriate clustering algorithm



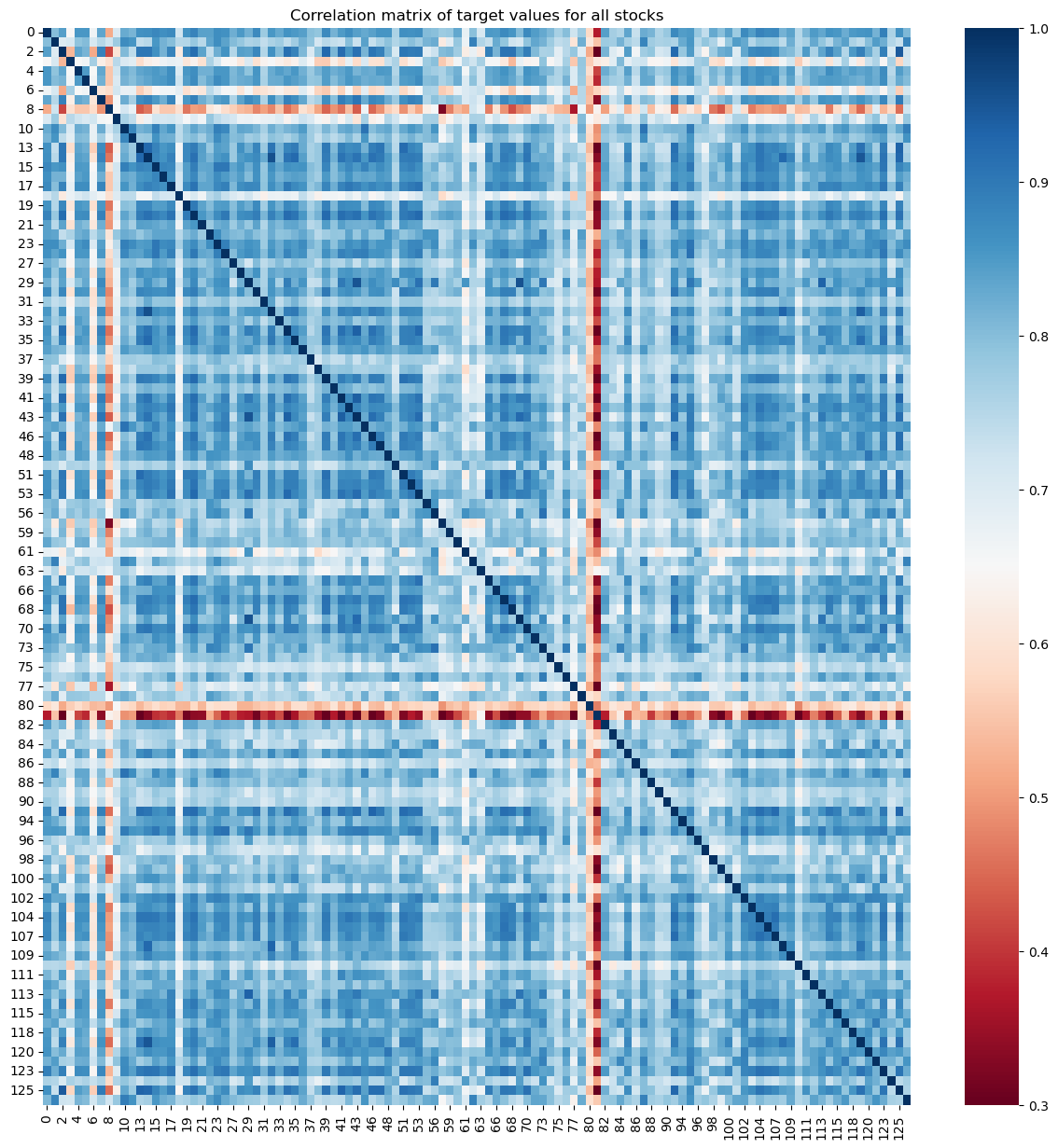
1. Perform clustering using various clustering algorithms and hyperparameters and check which one maximises the silhouette average score.





**Clustering on Temporal target realized volatility correlation**

1. Find ALL the unique time\_ids from all stocks. Create a new target dataframe (train\_common\_time\_ids\_df ) with all these time\_ids for all the stocks. Use forward and backward fills to populate missing values.
2. Find correlation matrix of this train\_common\_time\_ids\_df
3. Plot the (unclustered /ordered stocks) correlation matrix



No negative or zero correlation. Correlation starts at 0.3.

1. For each stock sort the most correlated stocks. This can help to validate clusterings formed later on.
2. Perform agglomerative hierarchical clustering to cluster similar stocks based on pearson correlation. Stocks that realised volatility that move together are clustered together.

The linkage function takes the pairwise correlations between stocks. I.e. (N\*(N-1))/2 entries in the upper triangular part of the correlation matrix.

The methods used like single/complete/ward etc. are different ways to combine the indiv. stocks and then clusters of stocks until there is only 1 cluster left. <https://www.youtube.com/watch?v=8QCBl-xdeZI>

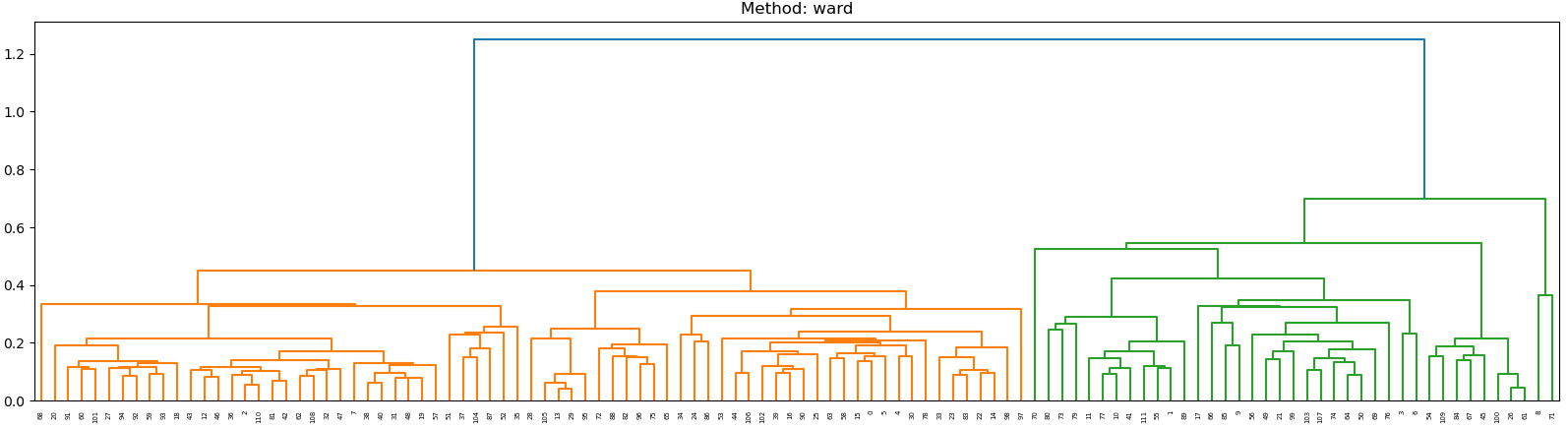
<https://www.youtube.com/watch?v=vg1w5ZUF5lA>,

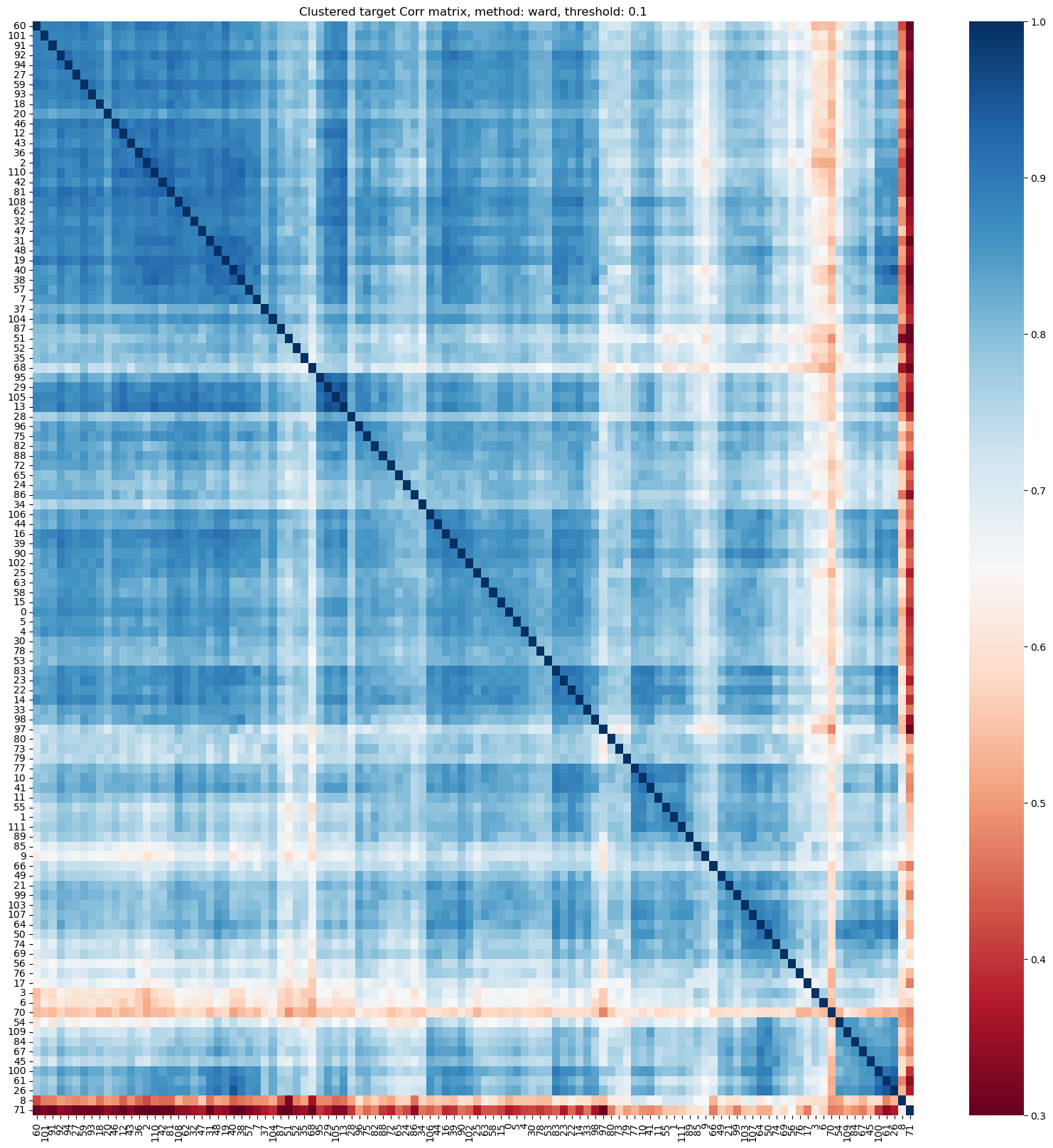
This is shown on dendrogram. The y-axis is the distance we provided (i.e. 1 - pearson\_correlation ) for methods like single and complete BUT not for all methods e.g. average.

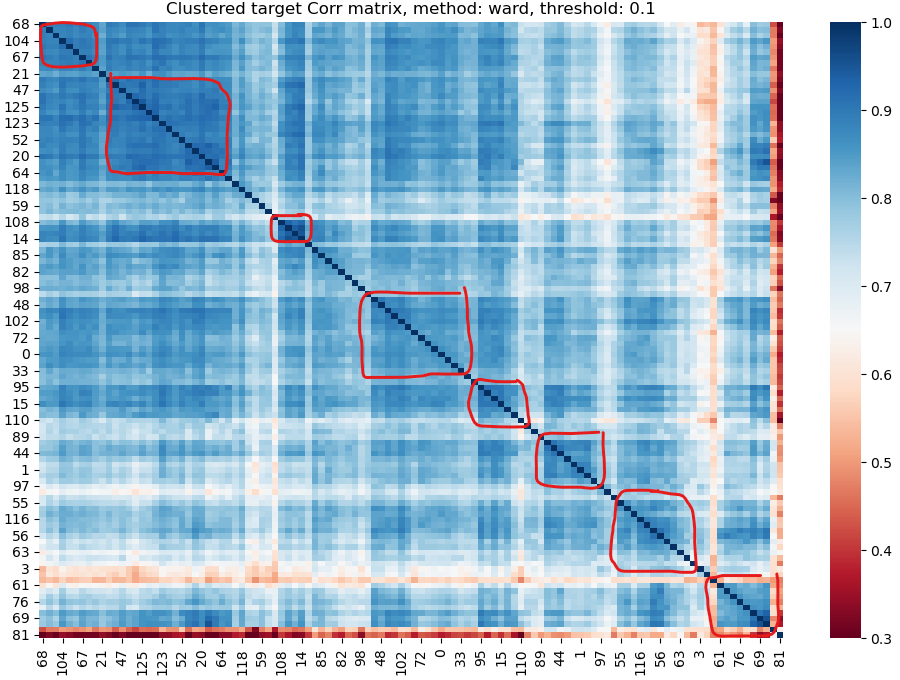
**Best methods: 1)** Ward, 2)Complete

Ward method is the best because it creates more compact clusters due to centroid formation and clusters combine based on new centroid formations. Other methods combining may note be good because they are combining based on edge point of the cluster alone whereas ward combines based on centroid which is better representative of the cluster than a single point i..e edge point.

WARD method: Try this first and see if prediction accuracy is good.







Look at the diagonal line. The square shaped clusters along the diagonal represent the groups/clusters of correlated stocks.

COMPLETE method : Try this if the first method’s prediction accuracy is NOT good.

