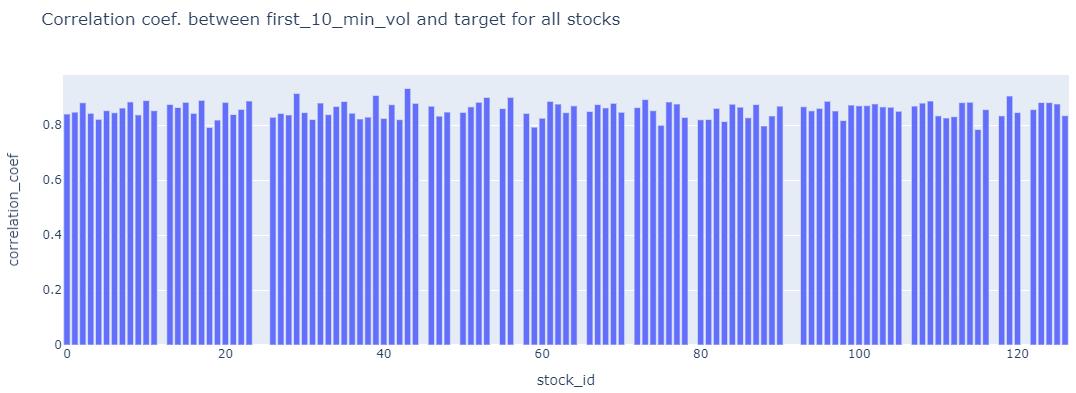
**IDEAS to gain insights**

# **Feature DEFINITION**

##### 

##### **1) Check correlation between target and book WAP realized\_vol in first 10 mins. Aggregate book WAP price using the log returns and realised volatility formulas for each time id. Plot scatterplot this against the corresponding target for all time id.**

yes confirmed!



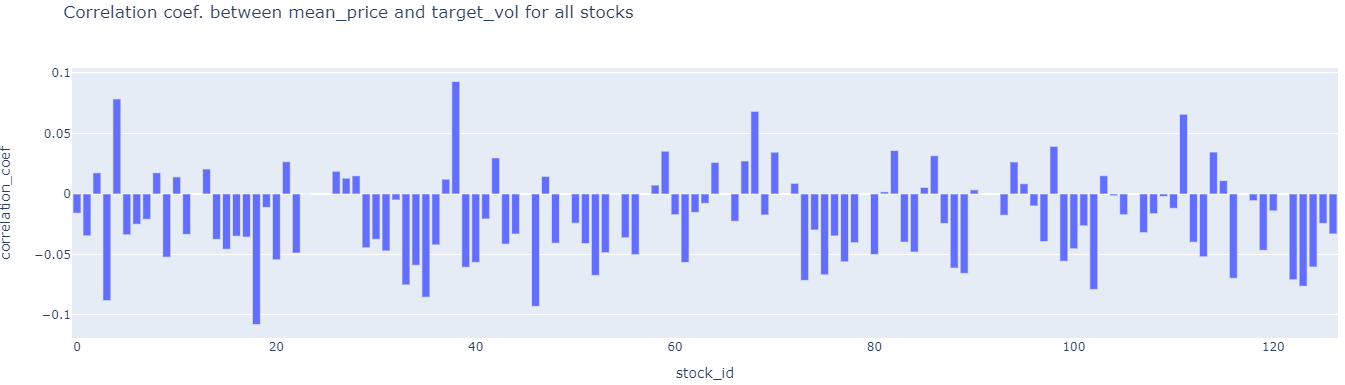
##### **2) Check correlation between**

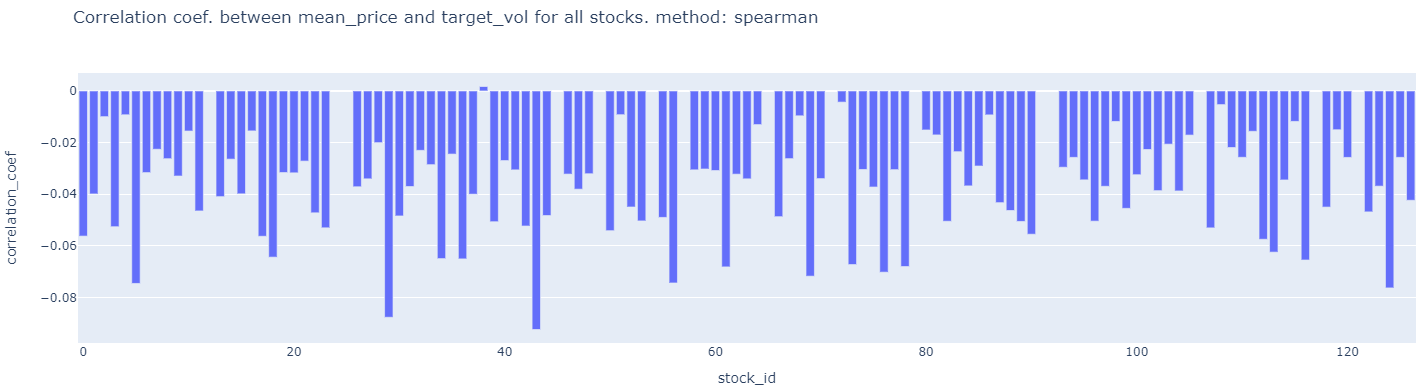
##### **i) target (2nd 10 mins vol.) and trade execution stock price at the available times.**

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.**

**Yes confirmed**

****

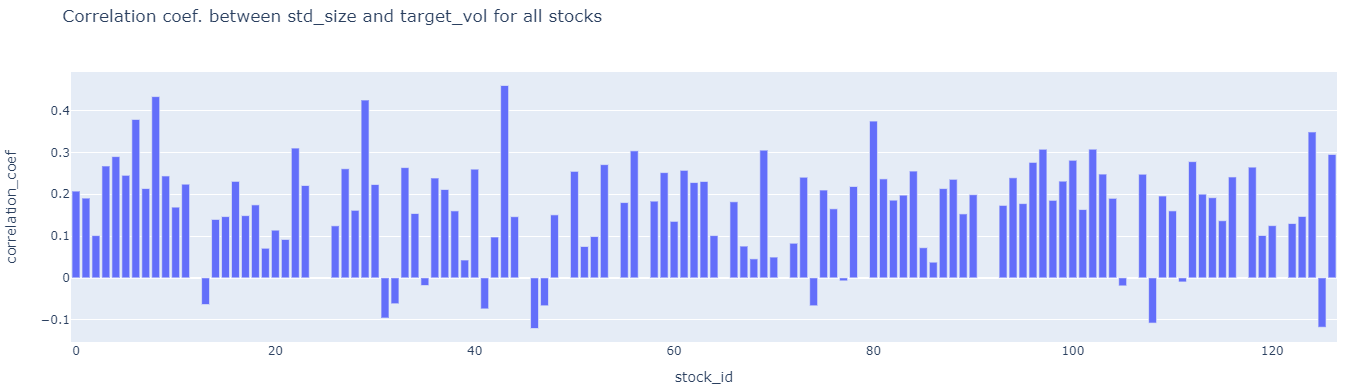
****

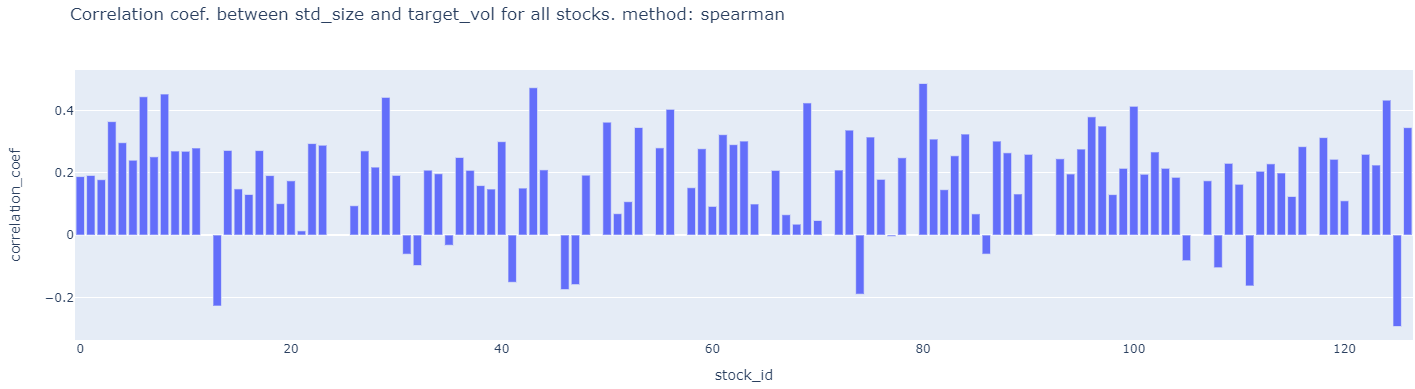
****

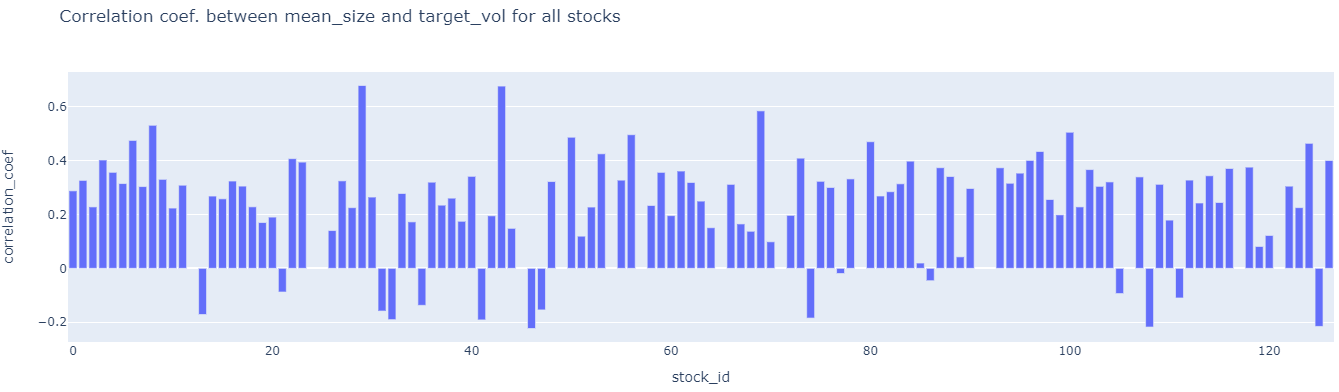
****

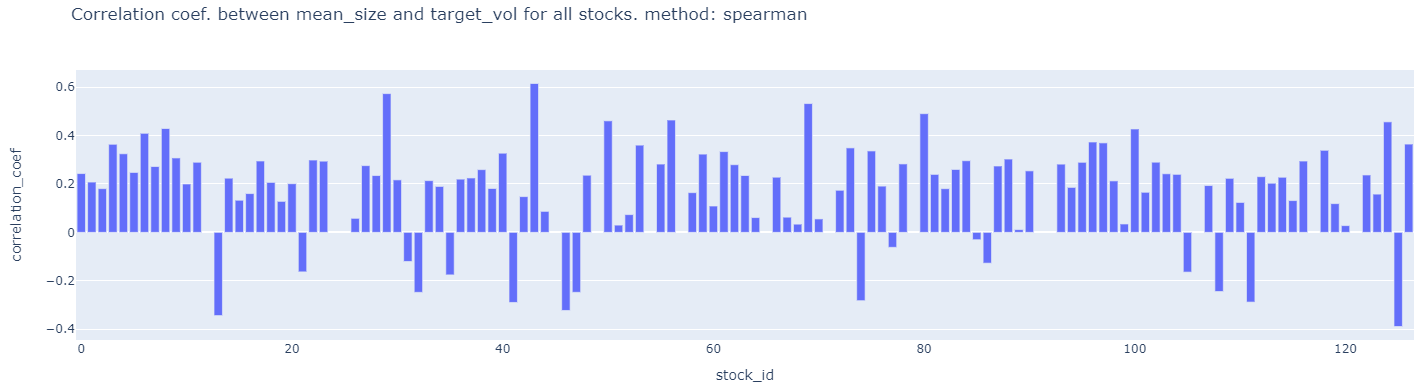
##### **ii) target (2nd 10 mins vol.) and size at the available times.**

Aggregate size by taking standard deviation or mean at each time id. **Plot scatterplot of this against the corresponding target for each time id.**

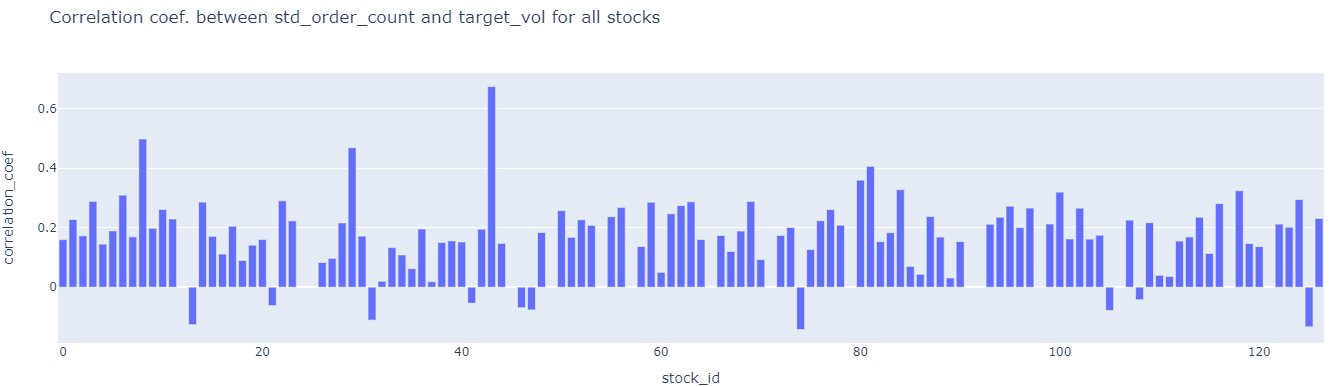
****

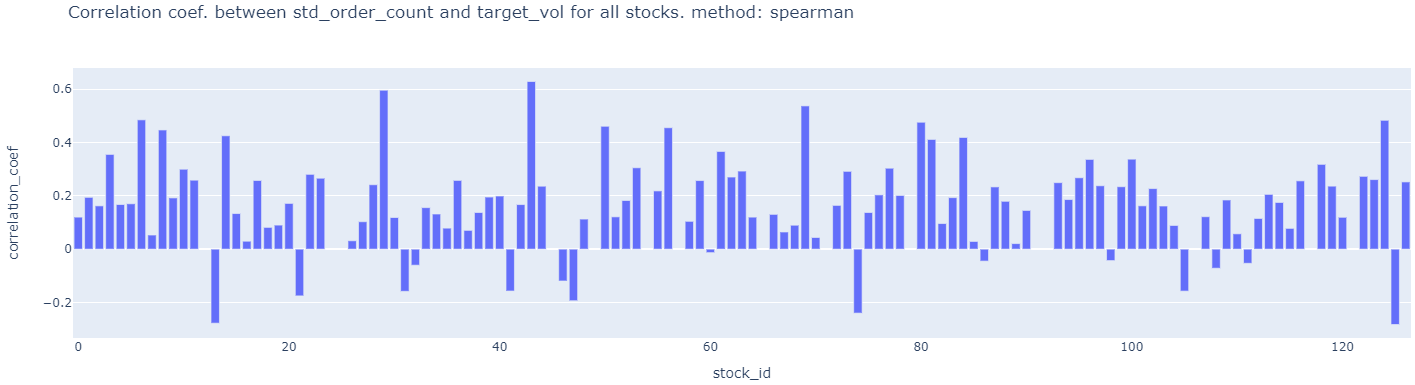
****

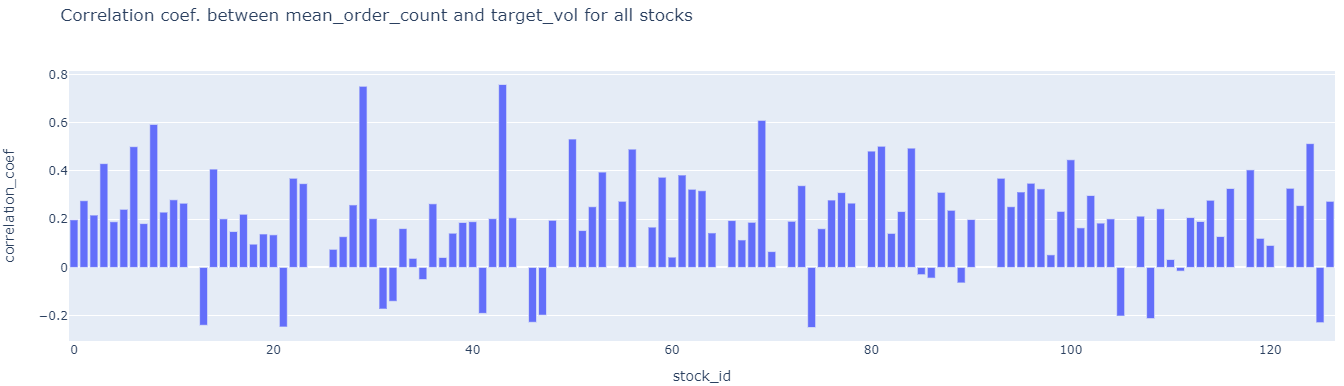
****

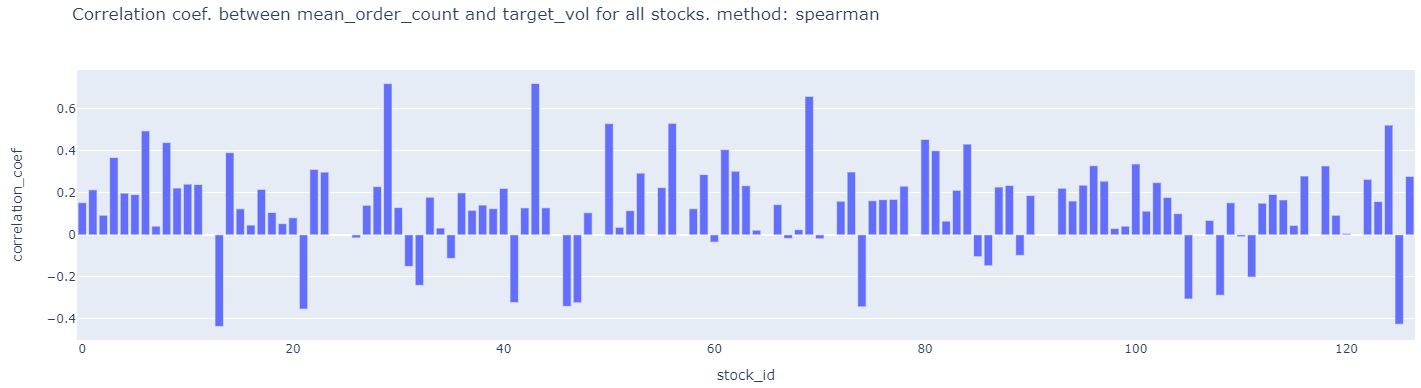
****

##### **ii) target (2nd 10 mins vol.) and order\_count at the available times.** Aggregate **order\_count** by taking standard deviation or mean at each time id. **Plot scatterplot of this against the corresponding target for each time id.**







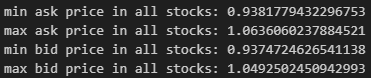


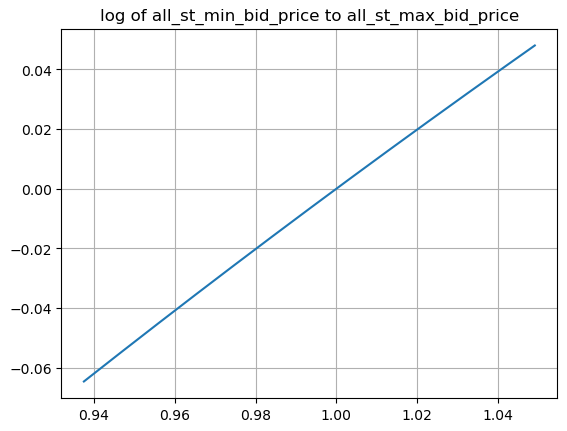
##### **2a)** Try to calculate spearman’s rank correlation for the cases of low linear pearson’s correlation. It might be higher due to non-linear correlation??? **DONE ABOVE. VERY SIMILAR.**

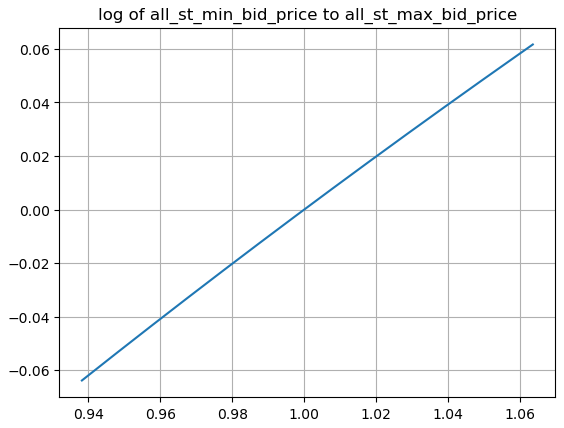
##### **3) check correlation in the first 10 mins. Target volatility between any two stocks and then identify which stocks have highest correlation. This is possible because even though time\_id is shuffled, they are the same for each stock so the scatter plot does not care about time. 112 choose 2 = 6216. compute this numerically. DO NOT PLOT. Rank the correlations.**

**DONE and updated in Key Insights file.**

4) ## check how prices are distributed in the book data so that we can see how log(s\_t2/s\_t1) transformation of prices below and above 0 affects volatility. ## due to log nature check if non-linearity is visible. A point in graph below is for a single stock.



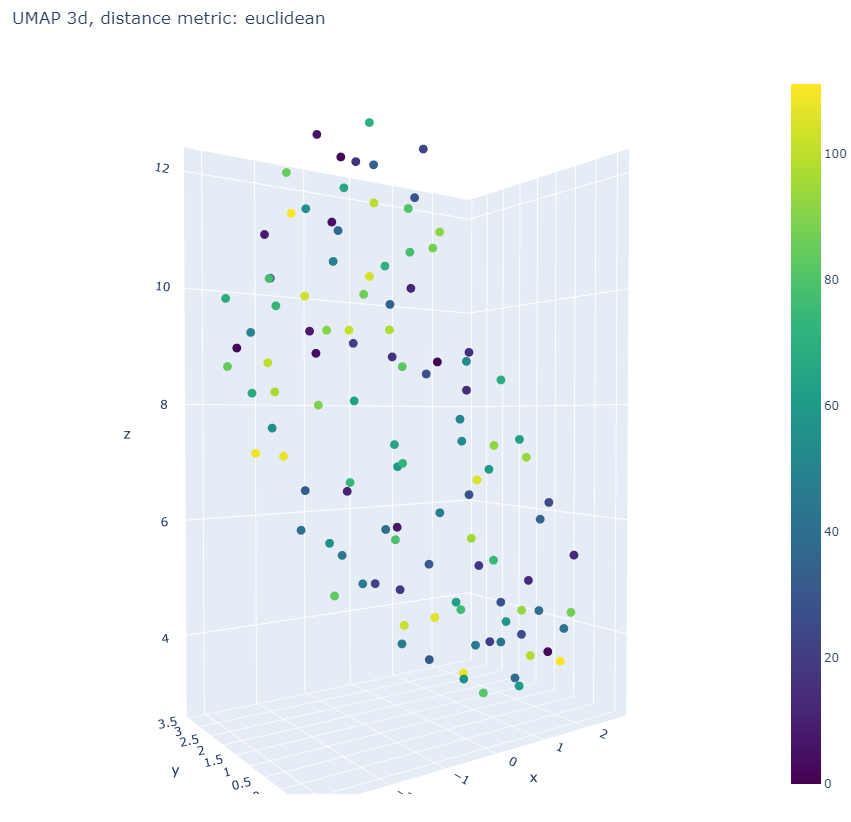




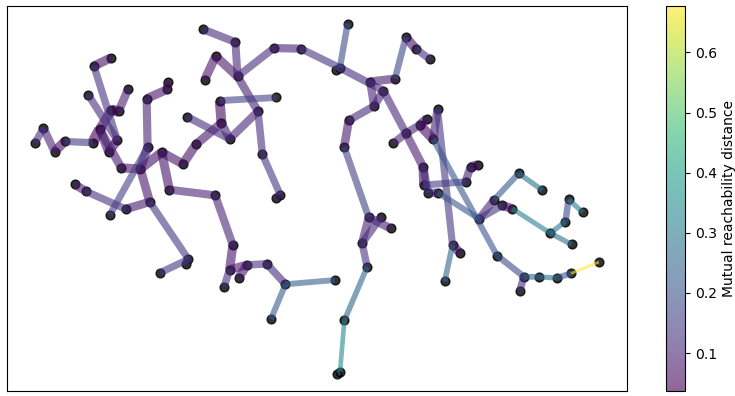
## Non-linearity is NOT visible as variance in price is low. so NO need to treat prices < 1 differently from prices > 1.

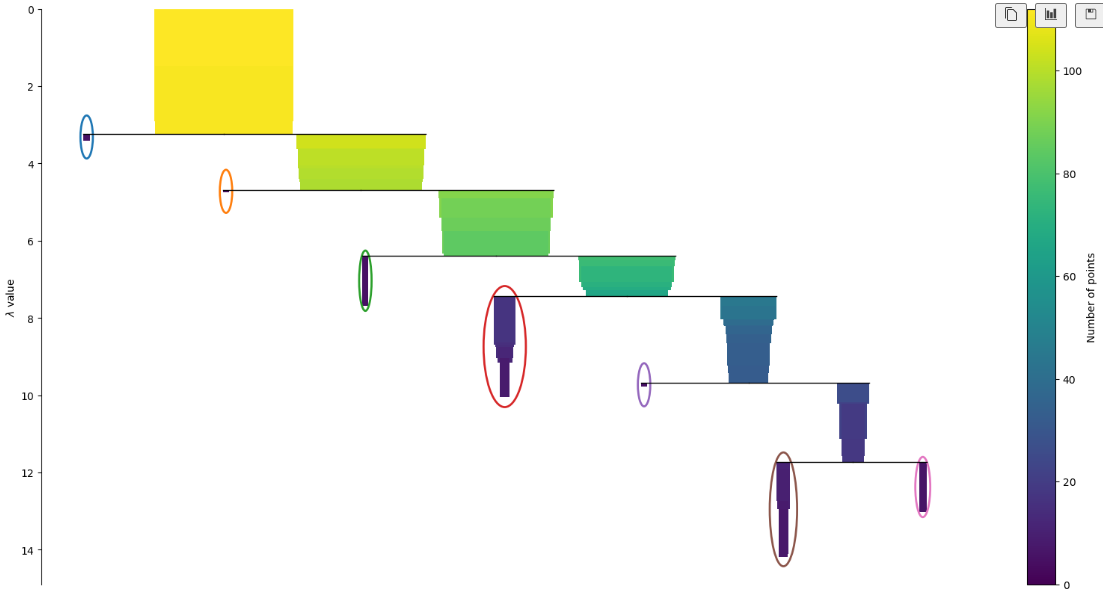
5) perform target volatility clustering across stocks using summary statistics features like mean, median, min, max etc..

Ans: The stocks volatilities are so homogeneous that it's difficult to separate/cluster them using summary stats features and hdbscan clustering algorithm. Maybe better features and/or different clustering algo. Might work. This is done in target\_eda\_across\_stocks.ipynb file

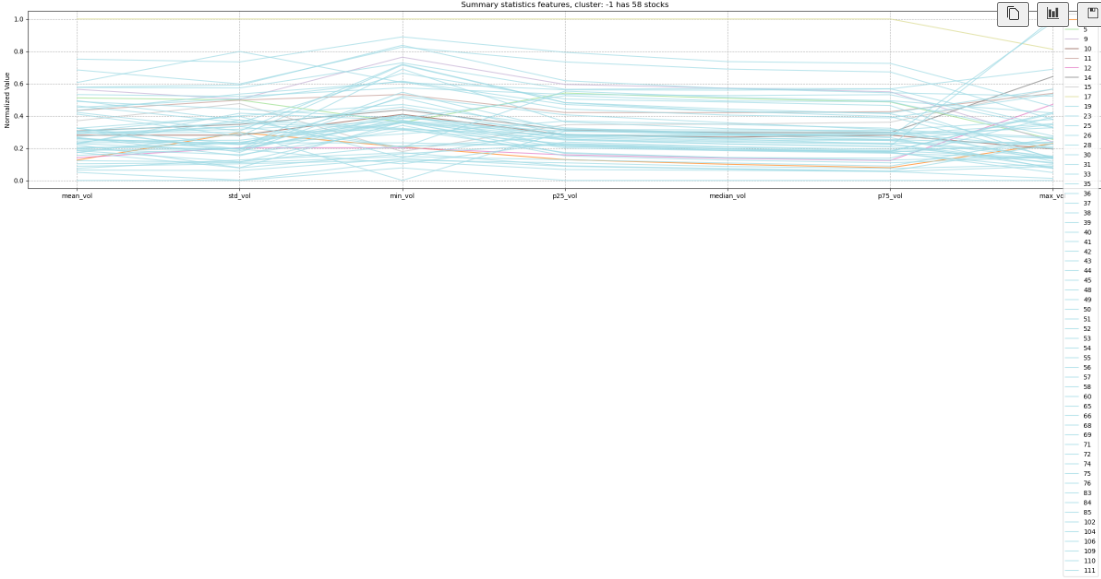
Colors represent different stocks. 

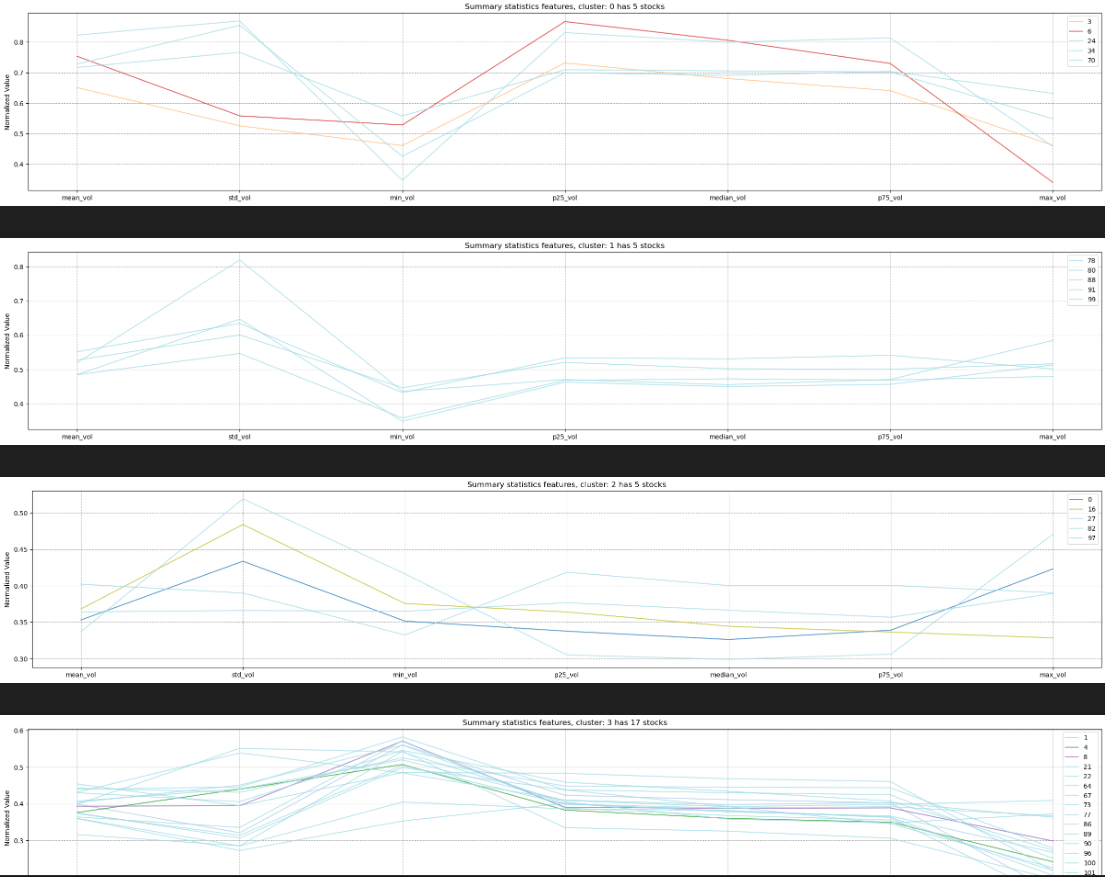
HDBSCAN analysis below was NOT fruitful!! .

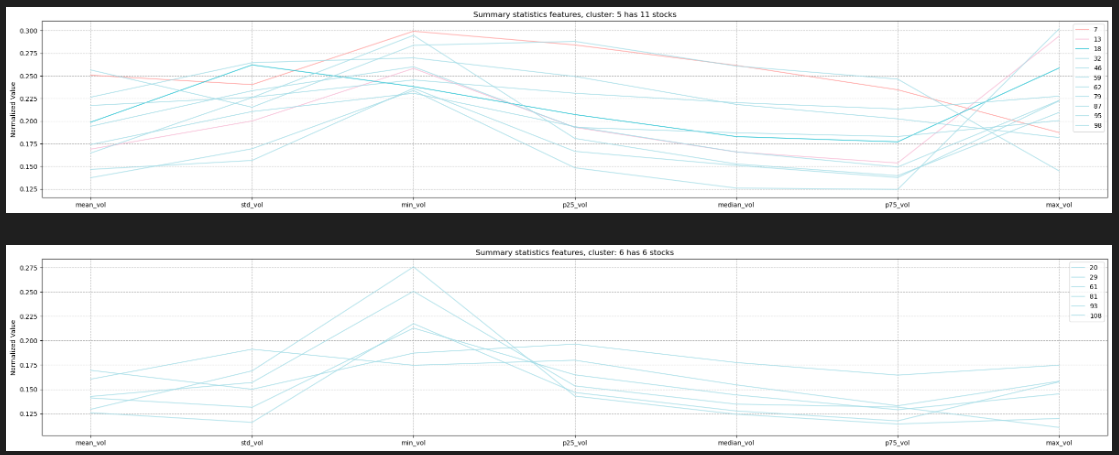




There are too many outliers in this clustering method.

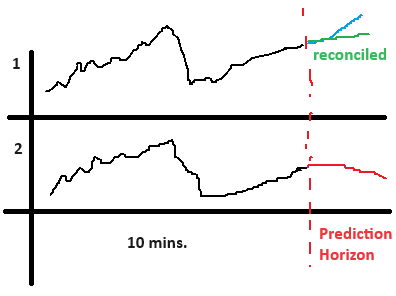






Hierarchical Agglomerative clustering using WARD linkage is better.

6) If two stocks (i.e. time series within a time\_id i.e. 10 min interval) are highly correlated for many time\_ids then we can improve the independent forecast of one stock by using the independent prediction of the correlated other stock. Reconcile the two predictions for the first stock to get more accurate forecast for 1st stock. Correlation captures shared patterns between the two stocks while independent stock prediction is based solely on within stock patterns. Do similarly for stock 2 using stock 1’s prediction. The other stock is acting like a covariate.



This analysis is SAME as in suggestion 3) which is **DONE and updated in Key Insights file.** Although data analysed is from TARGET and NOT from input features i.e. book\_train and trade\_train data. It’s better still.

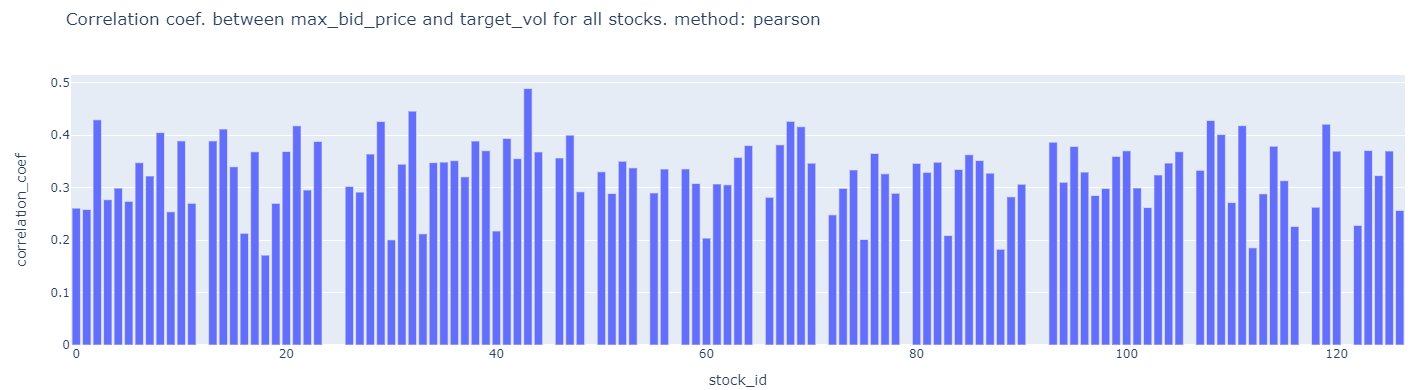
7a) Check if minimum/maximum of bidsize1 and asksize1 in a time\_id correlated with target realized volatitlity for the same time\_id?

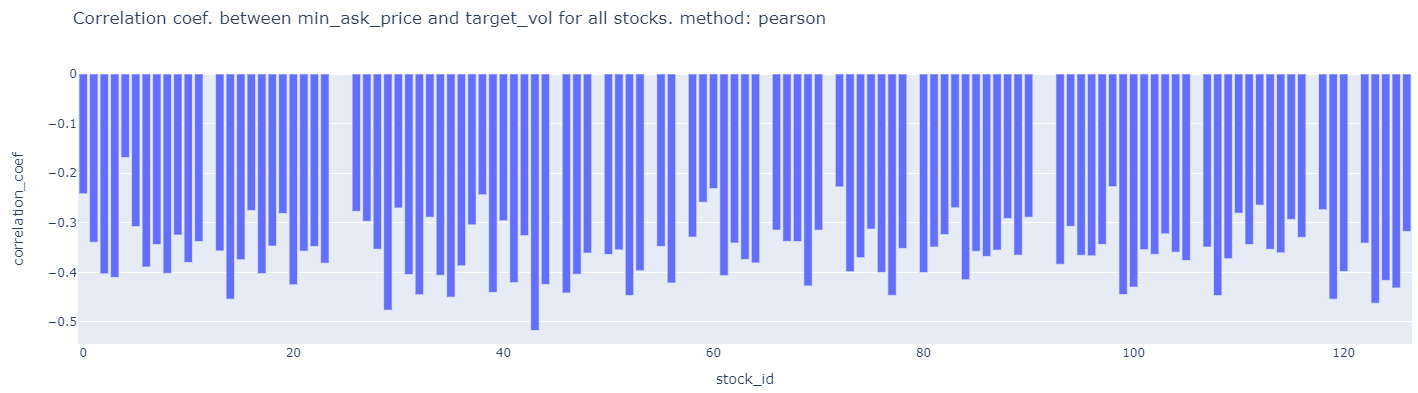


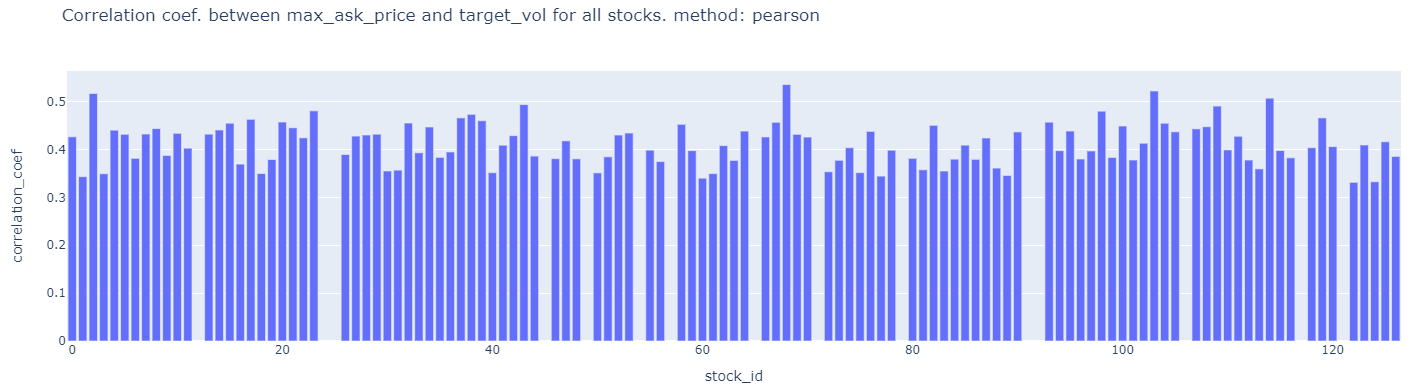


7b) Check if minimum/maximum of bidprice1 and askprice1 is correlated with target realized volatitlity for the same time\_id?





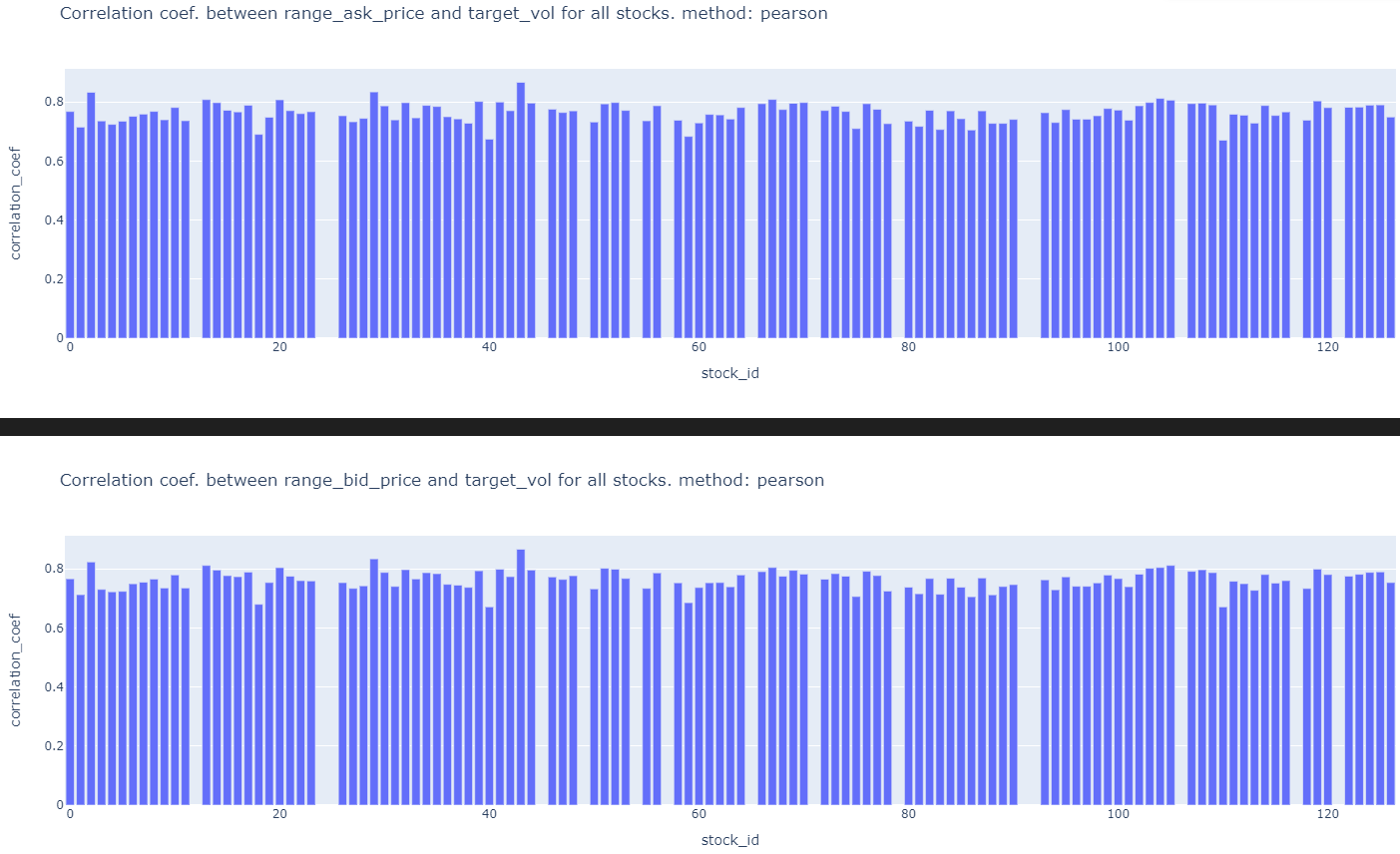




7c) check if the difference between minimum and maximum (i.e. range) of bidsize1 and range of asksize1 is correlated with target realized volatitlity for the same time\_id?

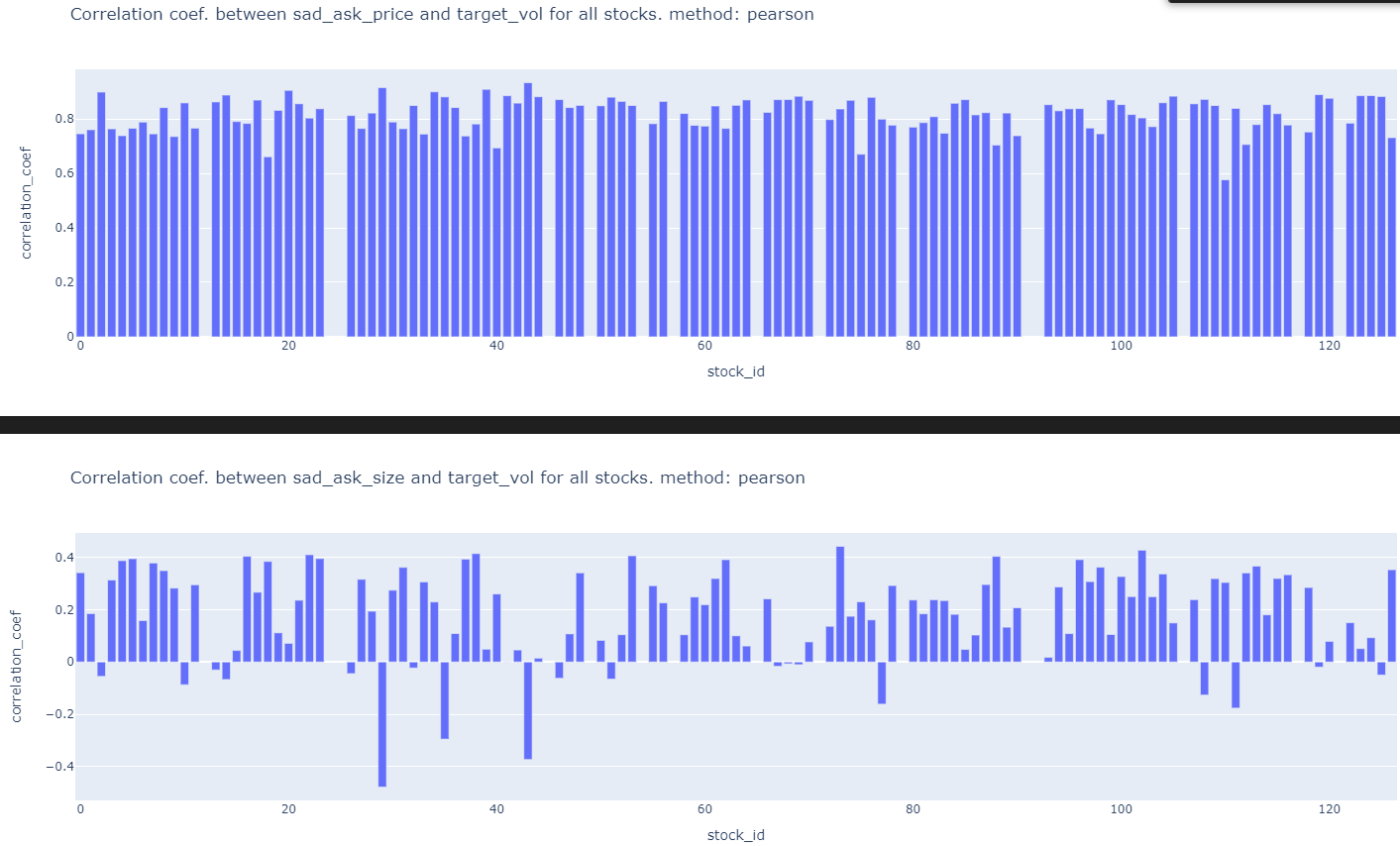


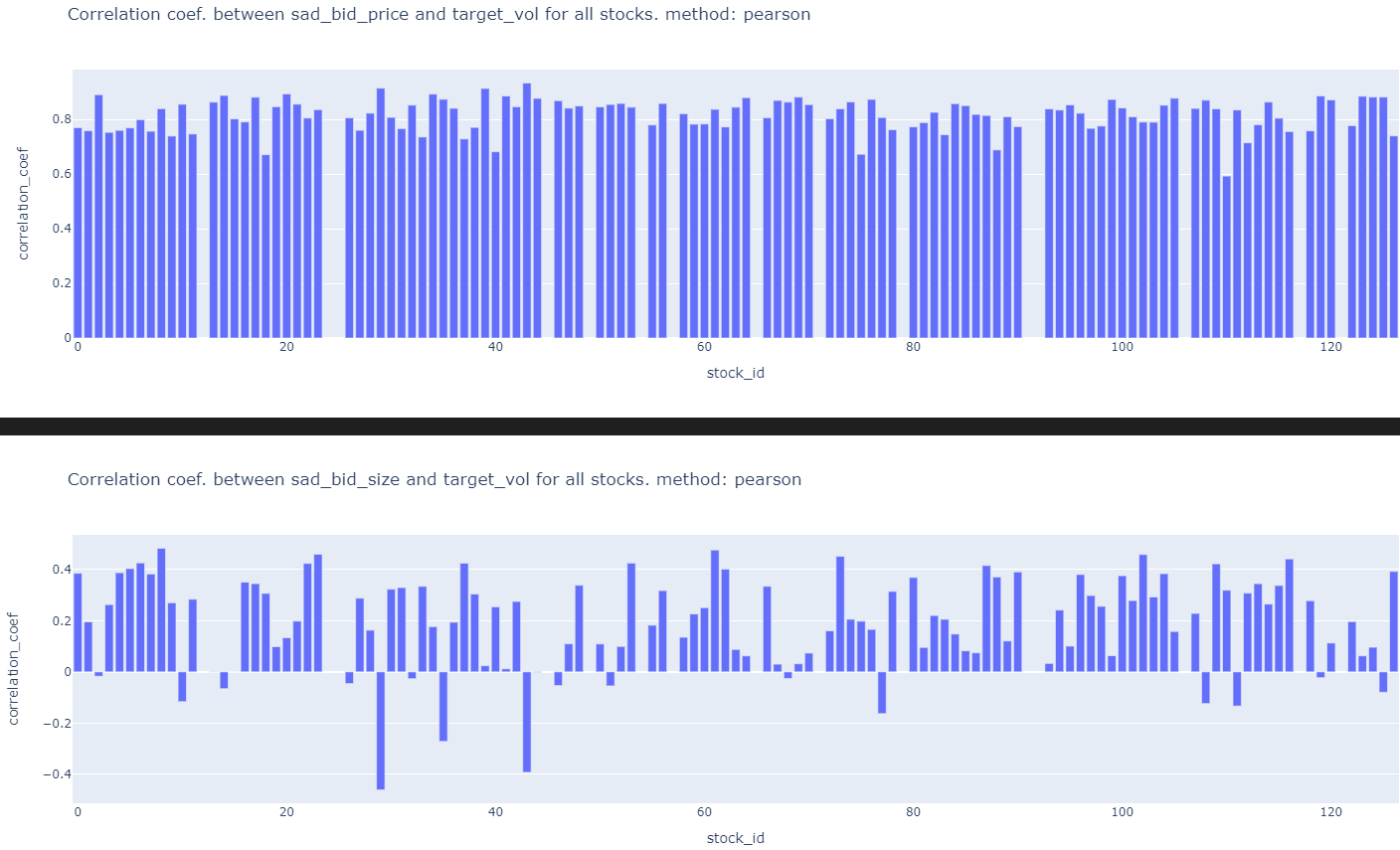
7d) check if the difference between minimum and maximum (i.e. range) of bidprice1 and range of askprice1 is correlated with target realized volatitlity for the same time\_id?



7e)Also check the correlation between (minimum – maximum) (i.e. range) vs. realized volatitliy of target. WHAT??

7f) Also check if the ~~average~~ sum of absolute bid\_price1 in t2 minus bid\_ price1 in t1 within a time\_id is correlated with target. Check the same for ask\_price1. Sum of absolute differences is better than average because price changes are too small and there are 600 seconds in a bucket so average values will be too small.



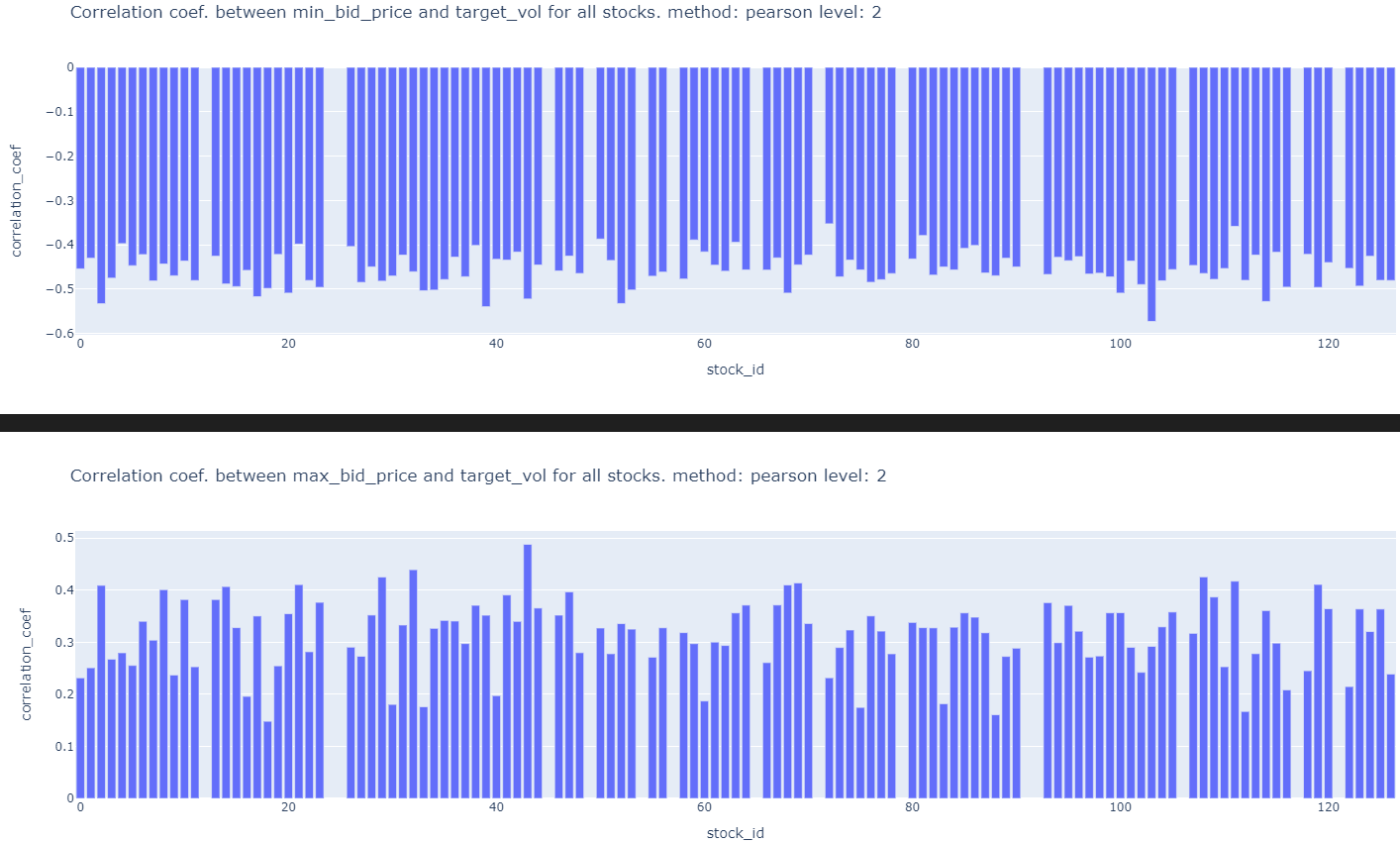


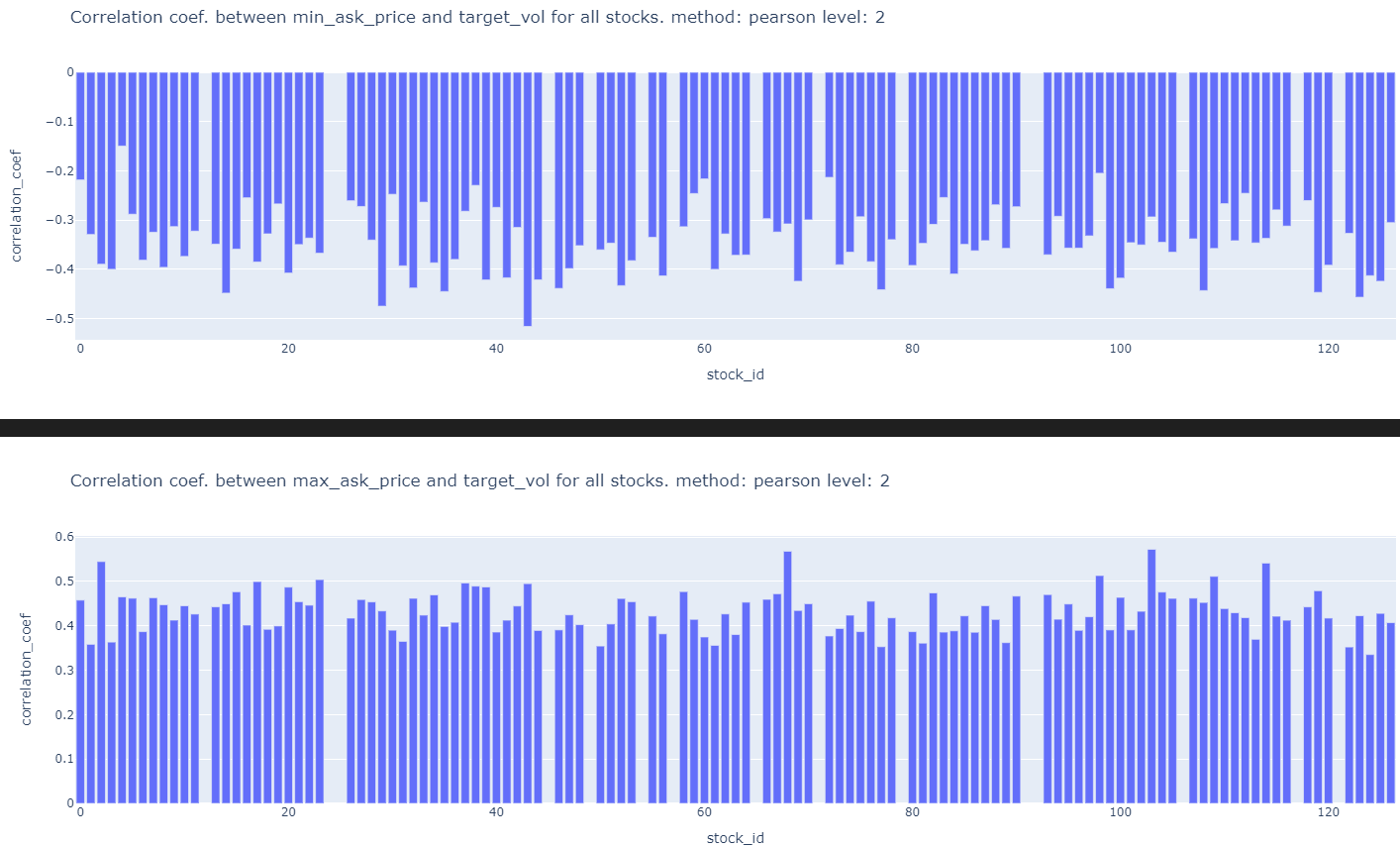
7g) calculate correlation between bid\_price1 and ask\_price1 in a time\_id (positive correlation leads to a larger/smaller wap’s numerator) then check correlation with target realized voaltitlity. This code is done by Kartik already??

7h) calculate correlation between bid\_size1 and ask\_size1 in a time\_id (positive correlation leads to a larger/smaller wap’s denominator) This code is done by Kartik already??

7i) Calculate pairs of correlation between a time series in this list [bid\_price1, ask\_price1] and this list [bid\_size1, ask\_size1] in a time\_id (negative correlation leads to a larger/smaller wap) i.e. corr(bid\_price1, ask\_size1), corr(bid\_price1, bid\_size1), corr(ask\_price1, bid\_size1), corr(ask\_price1, bid\_size1) This code is done by Kartik already??

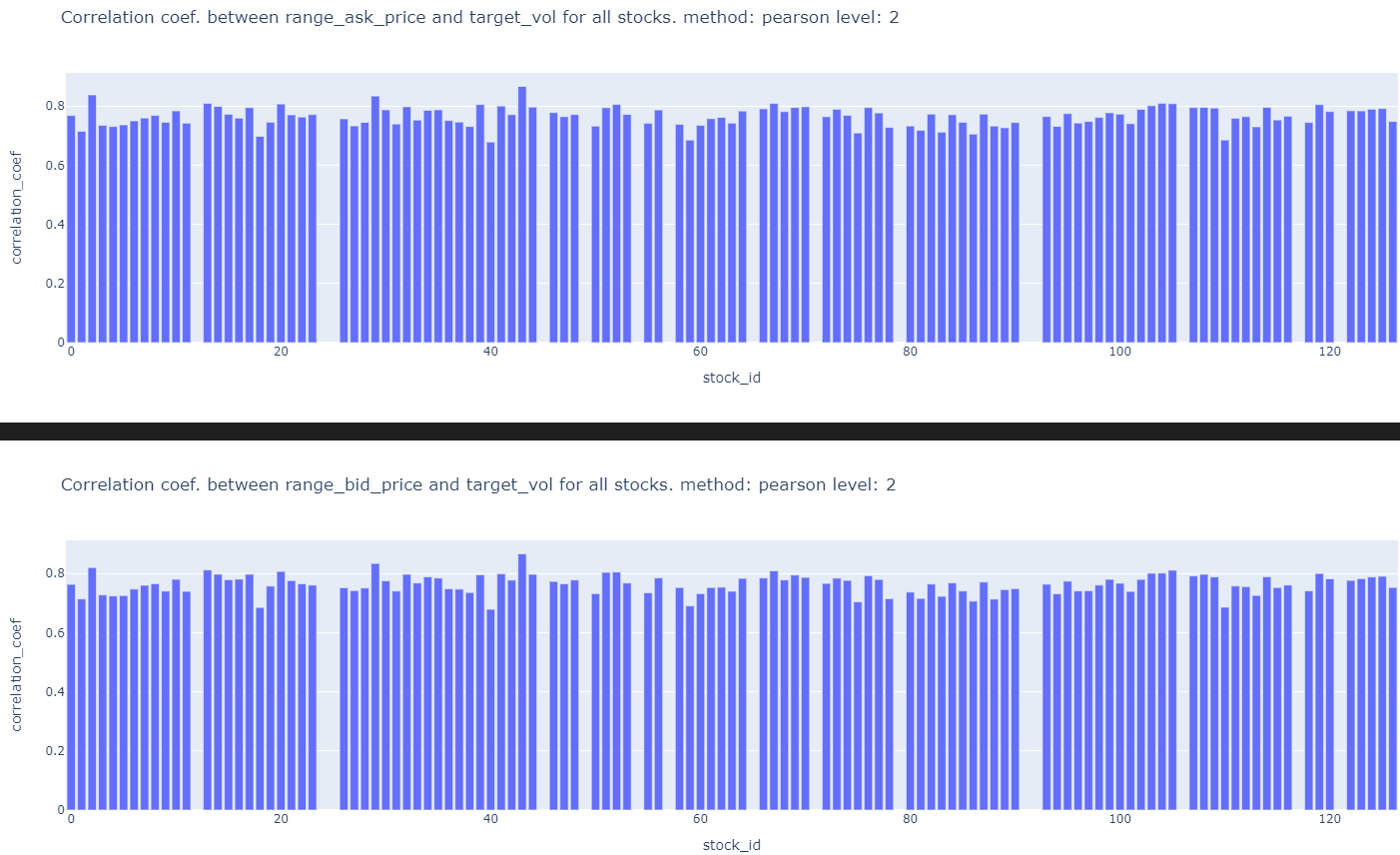
7j) Do the above for bid/ask\_price2, bid/ask\_size2 as well.

















ALL of suggestion 7 is done in features\_eda\_within\_stocks.ipynb

This is because from the WAP formula we can see that there is large variation in wap time series caused by min and max values of bidprice1 and askprice1 bidsize1 and asksize1 respectively. This affects returns and volatility.

7k) Do the above for trade\_train.parquet

8) What features in the first 10 min that affect volatility in the next 10 mins?

Feature that has Delayed effect on volatility?

Ans: Rolling average or moving average volatility can identify a lasting trend of high or low volatility. (peeterson)

9) Use a GARCH model on the time series available in the first 10 mins. to predict volatility into X (X<10) minutes of the second 10 minutes. Check correlation for different average x minutes realized volatility against the target realized volatility and choose the best x minutes of average volatility for each stock. (peeterson)

10) Plot ACF of squared returns time series to leverage the Volatility clustering (persistence) phenomenon to predict/extend into second 10 minutes. Remember volatility is square root of sum of squared returns.(jx)

11) WAP price trend up/down in the second 10 mins is negatively correlated with realized volatility in the 2nd 1o mins. This is called the leverage effect. Can we forecast trend in the 2nd 10 mins? Try using moving average of WAP of different window sizes and check if current price is above or below average. to predict short term trend in 2nd 10 mins.

12) When calculating realized volatility did you first ffill for the missing book order seconds?

13) Compute Wap2 using ask/bid\_price/size 2.

14) Just use skew, kurtosis, min, max, std,and all other statistics for each bid ask wap etc… in each time id.

15) Check if Standard Deviations of Price Variables is POSITIVELY Highly Correlated with Target

16)Check if Minimums of Prices Variables is NEGATIVELY Highly Correlated with Target;

17) the std of bid price was often more highly correlated with the target than the wap calculation.

18) Use bid ask spread . ask\_price1 / bid\_price1 - 1

# **Feature SELECTION**

19) always check correlation between features to ensure no multicollinearity issues.

18 features target real vol

7 pc target real vol