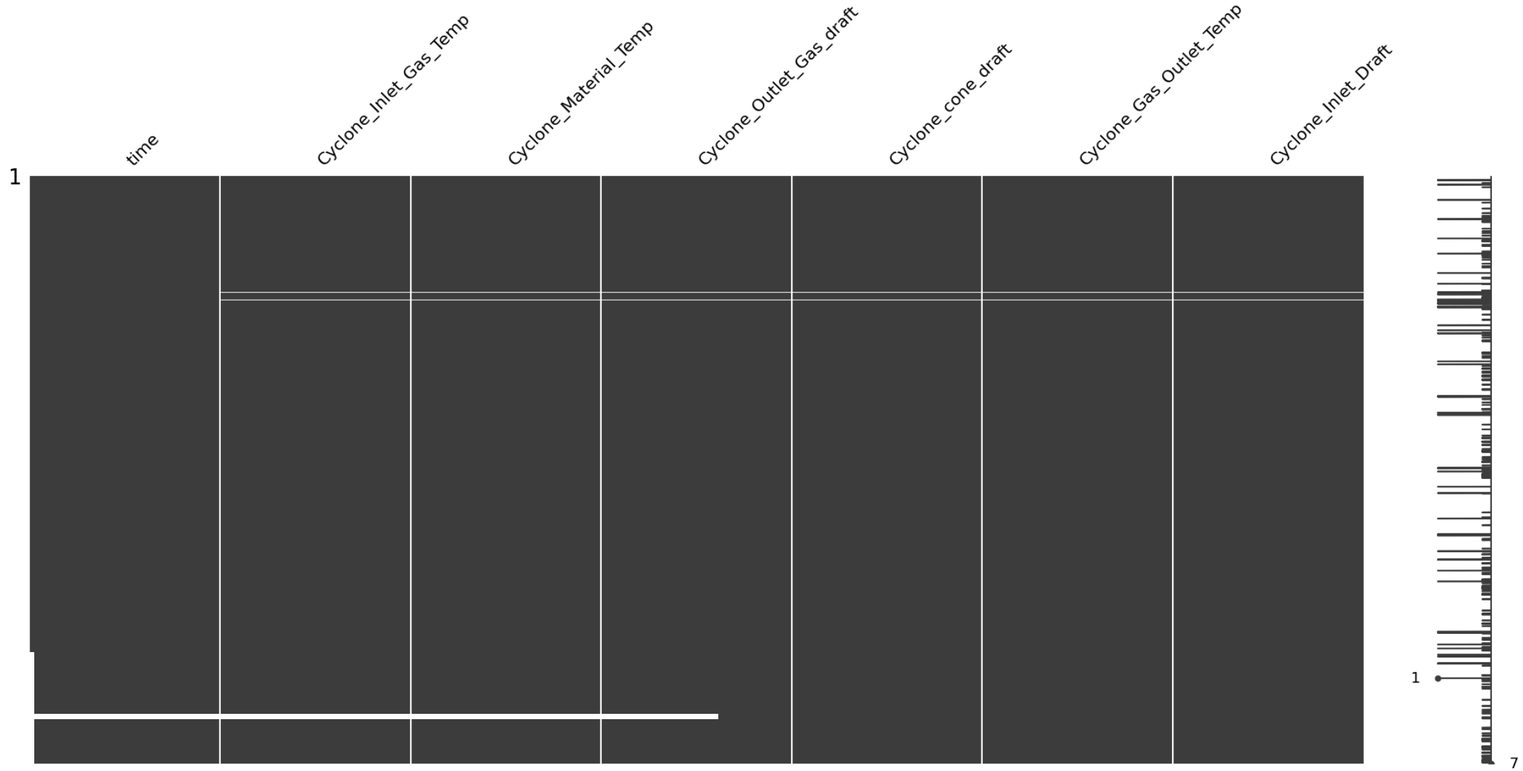
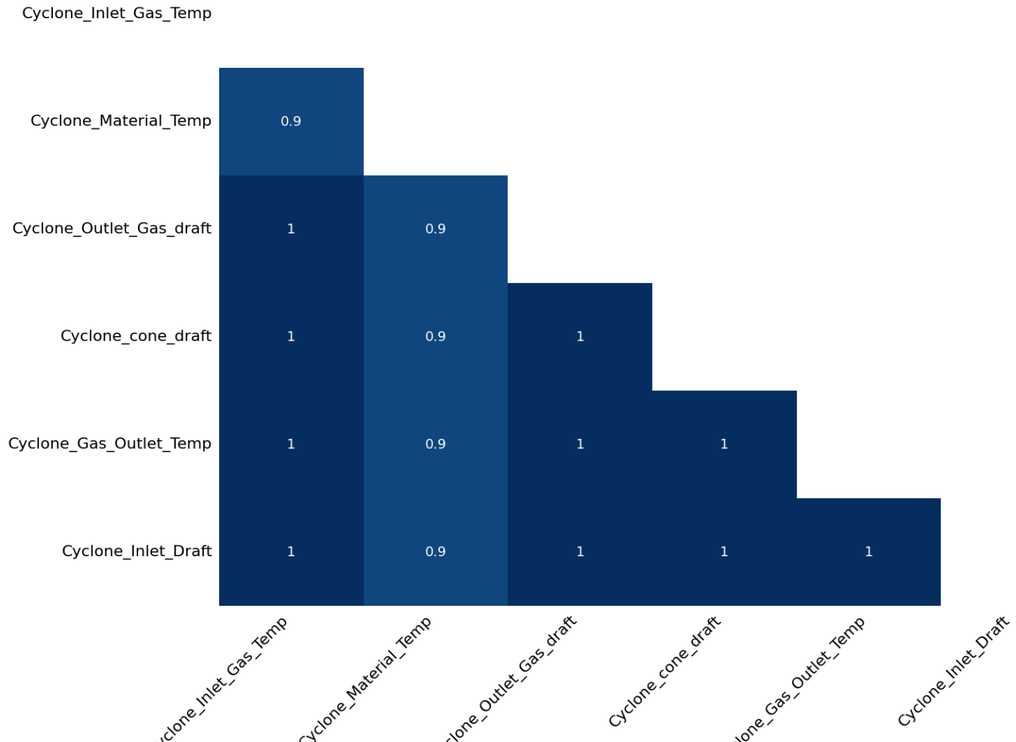
DATA SCIENCE ASSIGNMENT

Arshad Jamal

MISSING VALUES

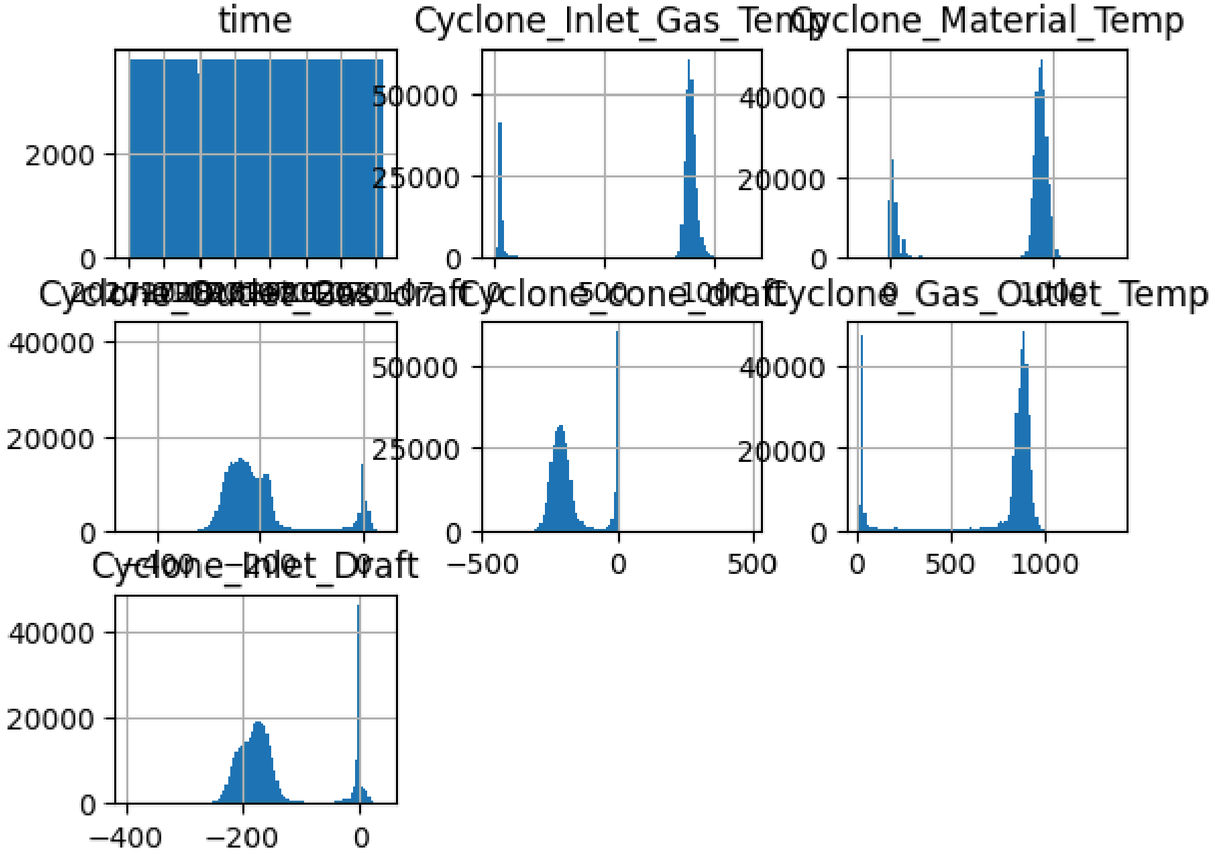
since we had very less percentage of vals missing, we cd remove them but before used missingno lib to see some patterns and identify missing data as Mnar or mar or mcar.

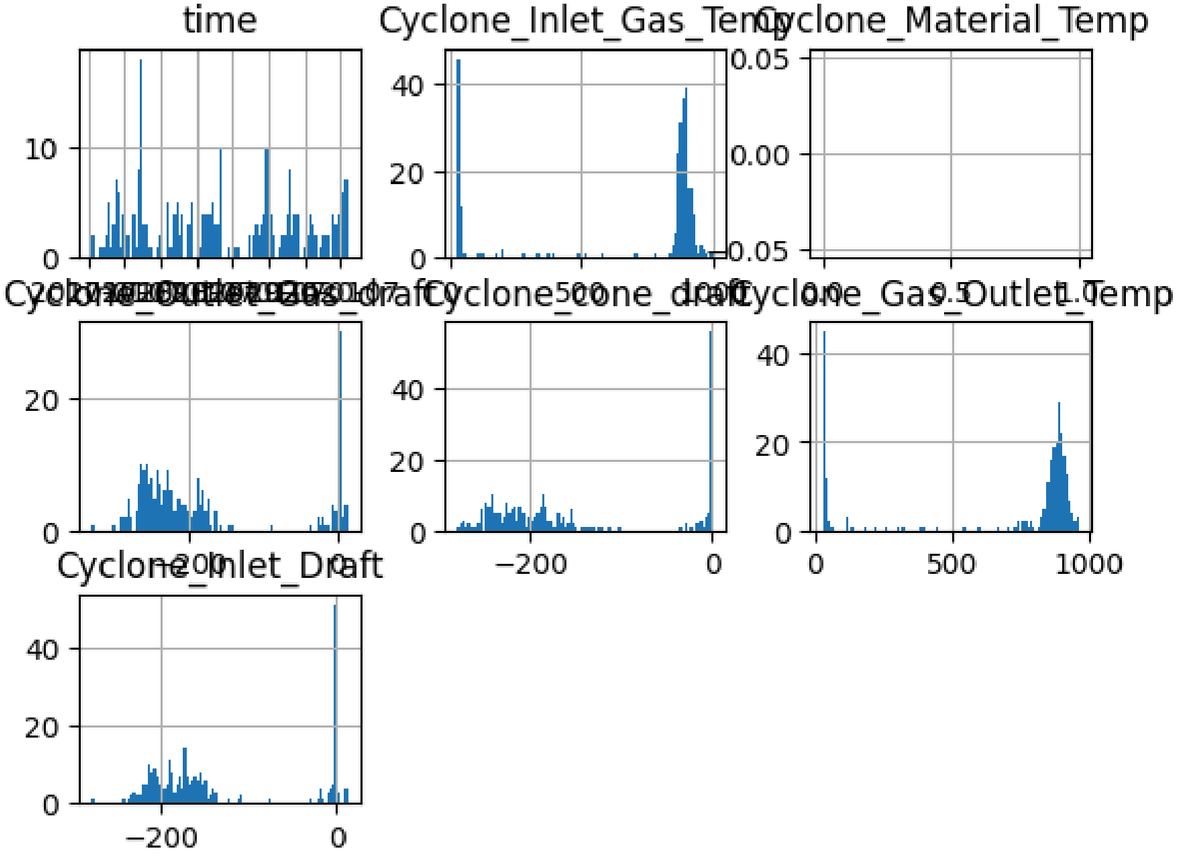




min cols =1 nd max = 7 that means where one col is missing all other are as well, high similarity indicates the same thus removal best option.

still, used dropna with ‘all’ first nd then checked again for some info.





left shows the overall distribution of the data nd the right shows the data which was left after doing drop na with all param, so now we see that those points werent outliers but just usual points of the distribution, removing which wdnt cause any hinderance or loss of info.
Hence, removal with ‘any’ approved!

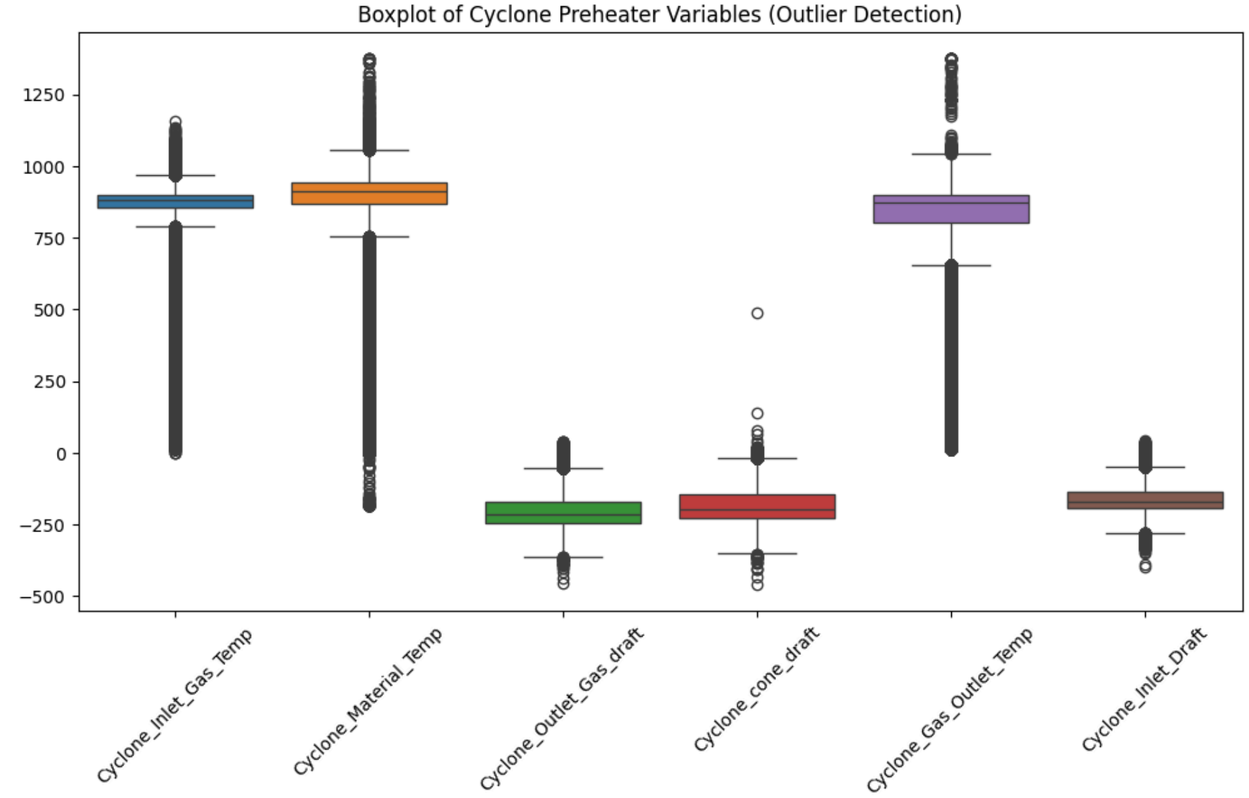
BOXPLOTS

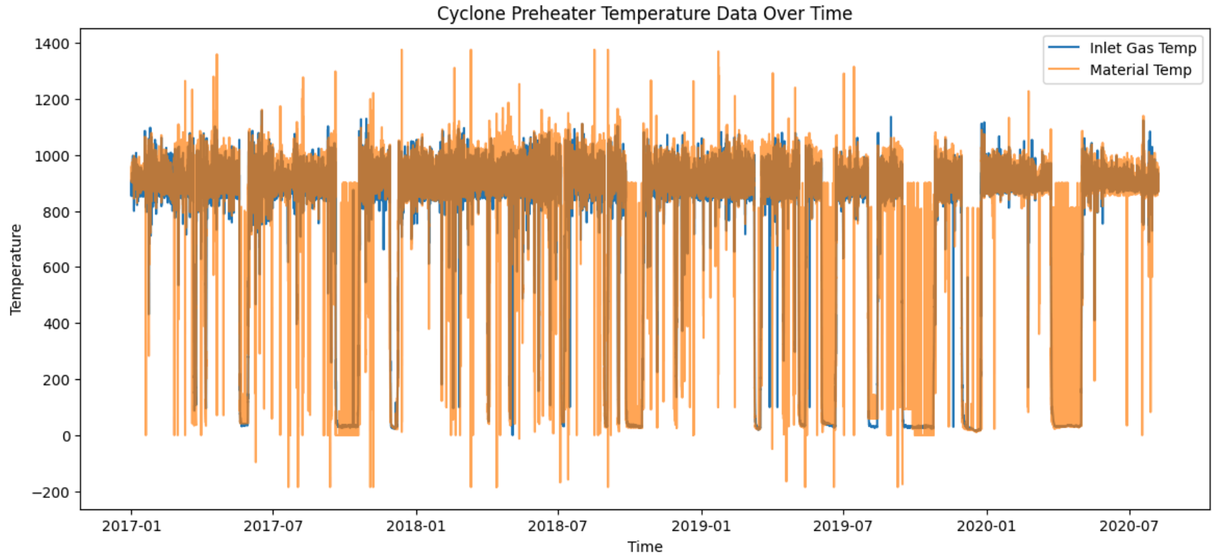
Indicates the need for multivariate outlier detection since

1. all variables show extreme outliers
2. univariate wd fail due to feature interaction
3. correlated cols
4. singularity may not be the cause of outlier, but combinations must be.

the graph below shows:

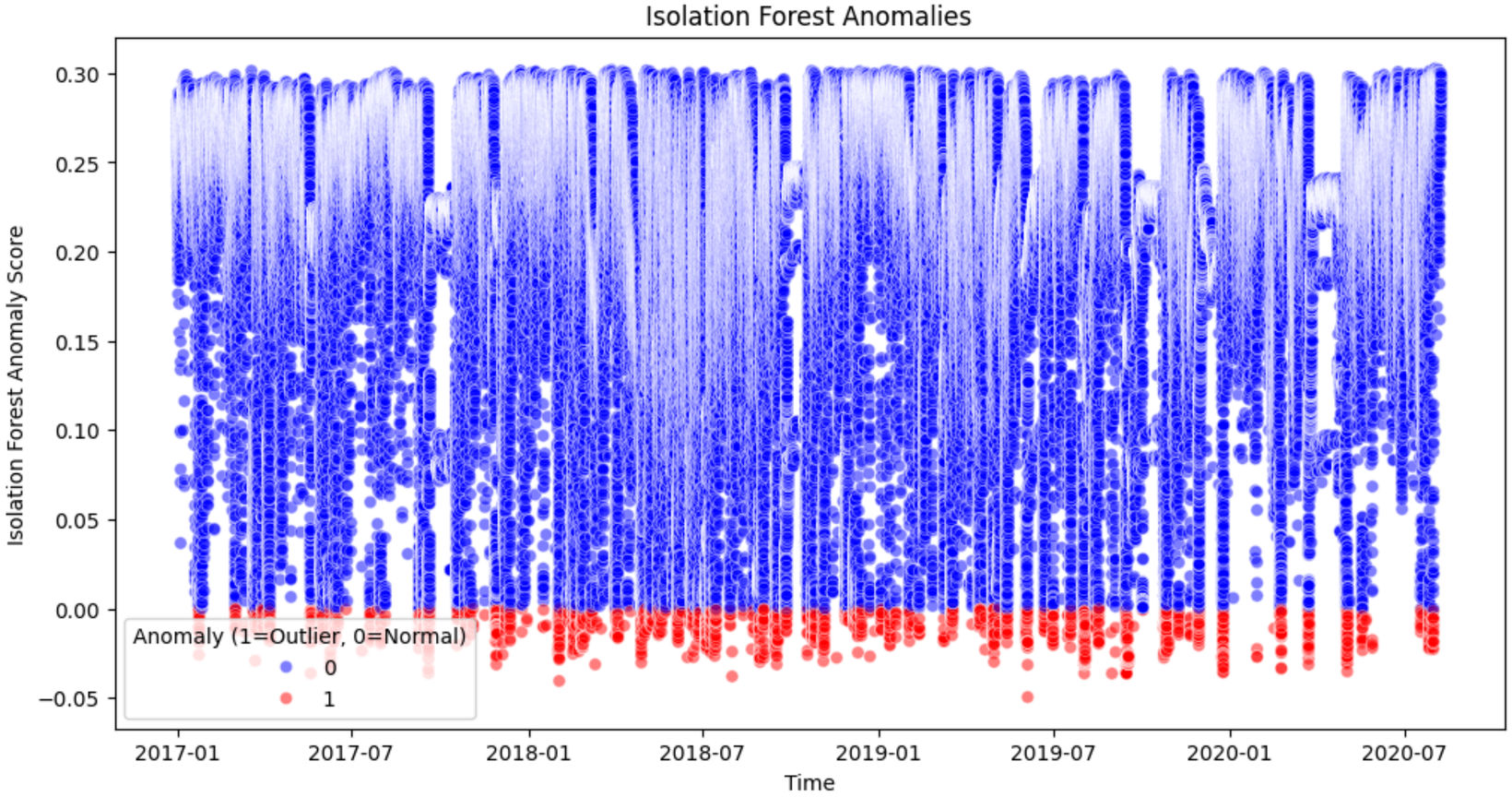
1. the follow of stable range
2. spikes nd drops
3. sudden temp drops near 0
4. maybe shutdowns , failures
5. two closely follow each other





Now lets figure out the best approach for our dats since our dimensionality is good but the rows are almost 350k

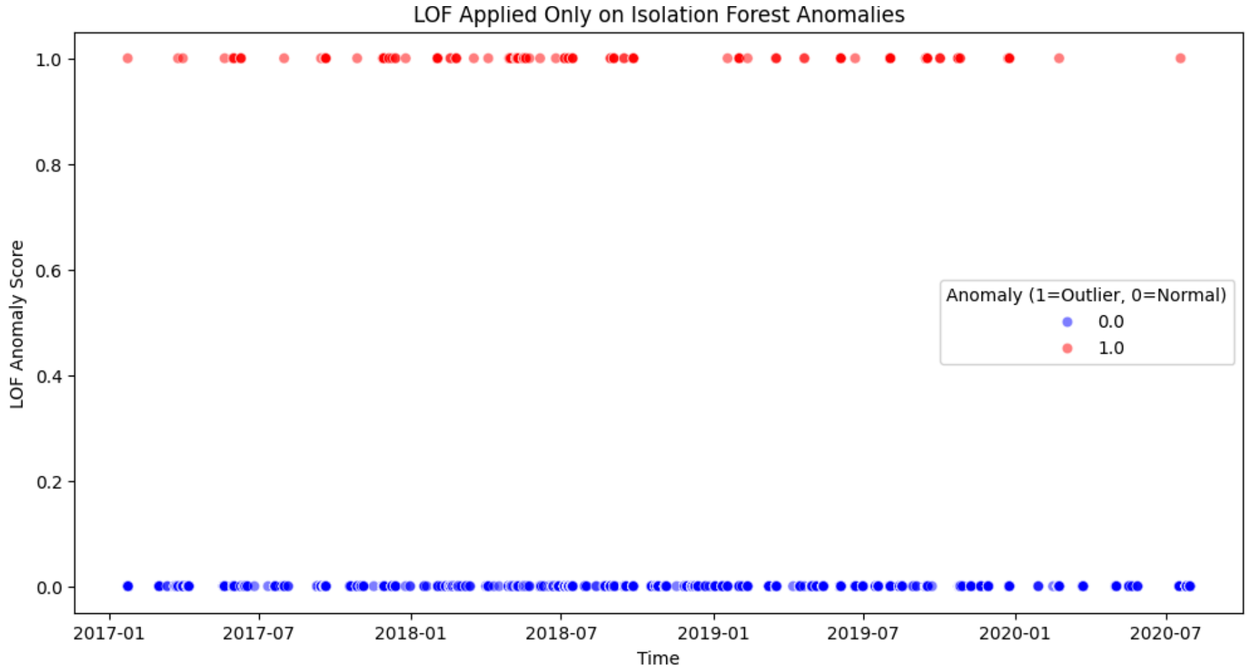
1. DB scan is poorly scalable with slow speed so we cant use it
2. HDB Scan slightly more efficient
3. Downsample wdnt help
4. LOF compares pairwsie distances so slow as well
5. Isolation Forest- excellent nlogn, very fast and thus the best choice for our use case

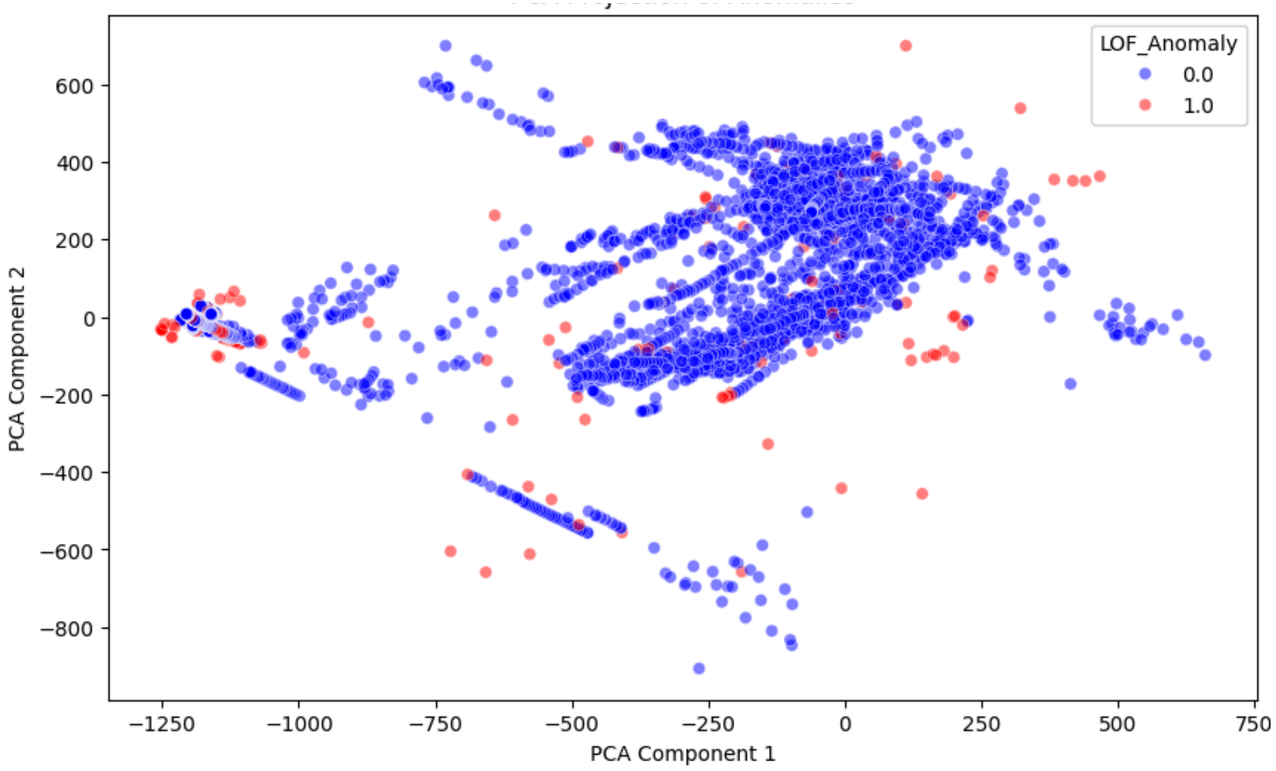


Observations:

1. Most anamolies have low anomaly score
2. detected anomalies are primarily concentrated at the bottom of dist, reinforcing consitency
3. frequency remains steady- might be following a recurring pattern rather than being random.
4. few normal pts show high scores- natural variability rather thatn outliers.

The red points represent anomalies (outliers) detected by Isolation Forest and then evaluated by (LOF).
The fact that these anomalies have high LOF scores (~1.0) suggests that LOF agrees with Isolation Forest's anomaly classification.



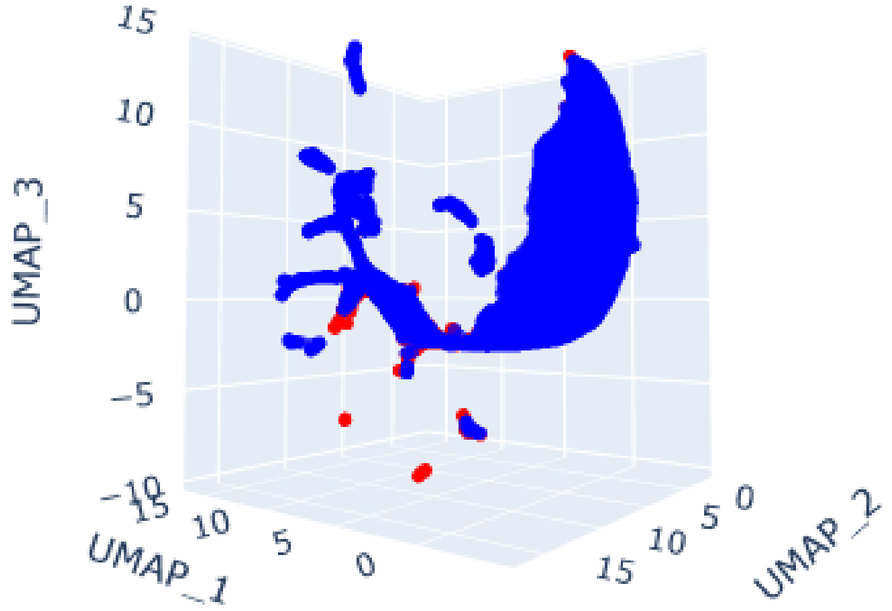


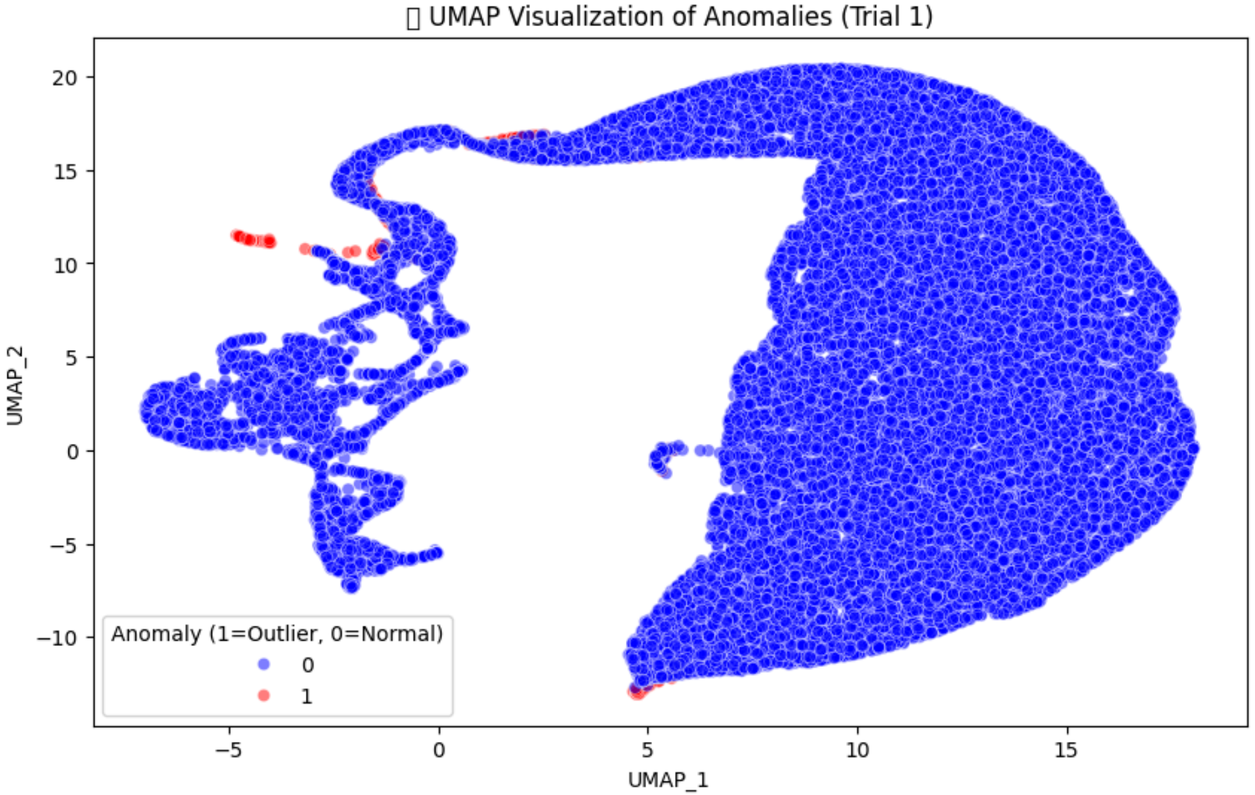
effectively detects outliers and shows both borderline and extreme anomalies,

using pca helped the above

Plotted UMAP post Pca for faster results:

1. Outliers are mostly on the edges
2. main dense cluster rep normal df, dominant pattern
3. Need for refining the model more
4. Parameter Tuning





1. some red points are embedded within blue region similar to above
2. cd hint towards borderline anomalies, or maybe potential misclassifications.
3. 3D obviously provides better inference