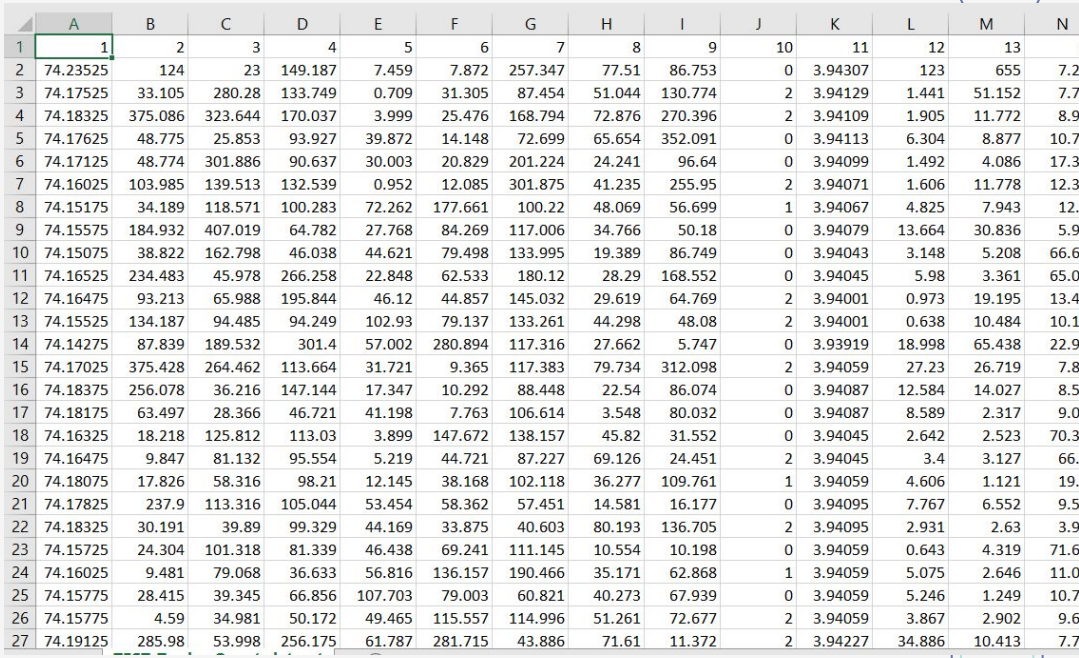
Analysing Unlabelled Data

Bending Spoons Task #3

Jaiveer Singh Madan

Problem Context and Dataset

- Unlabelled time-series dataset.
- 8,688 observations, 56 features
- No timestamps, labels, or business context



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	2	3	4	5	6	7	8	9	10	11	12	13	
2	74.23525	124	23	149.187	7.459	7.872	257.347	77.51	86.753	0	3.94307	123	655	7.2
3	74.17525	33.105	280.28	133.749	0.709	31.305	87.454	51.044	130.774	2	3.94129	1.441	51.152	7.7
4	74.18325	375.086	323.644	170.037	3.999	25.476	168.794	72.876	270.396	2	3.94109	1.905	11.772	8.9
5	74.17625	48.775	25.853	93.927	39.872	14.148	72.699	65.654	352.091	0	3.94113	6.304	8.877	10.7
6	74.17125	48.774	301.886	90.637	30.003	20.829	201.224	24.241	96.64	0	3.94099	1.492	4.086	17.3
7	74.16025	103.985	139.513	132.539	0.952	12.085	301.875	41.235	255.95	2	3.94071	1.606	11.778	12.3
8	74.15175	34.189	118.571	100.283	72.262	177.661	100.22	48.069	56.699	1	3.94067	4.825	7.943	12.
9	74.15575	184.932	407.019	64.782	27.768	84.269	117.006	34.766	50.18	0	3.94079	13.664	30.836	5.9
10	74.15075	38.822	162.798	46.038	44.621	79.498	133.995	19.389	86.749	0	3.94043	3.148	5.208	66.6
11	74.16525	234.483	45.978	266.258	22.848	62.533	180.12	28.29	168.552	0	3.94045	5.98	3.361	65.0
12	74.16475	93.213	65.988	195.844	46.12	44.857	145.032	29.619	64.769	2	3.94001	0.973	19.195	13.4
13	74.15525	134.187	94.485	94.249	102.93	79.137	133.261	44.298	48.08	2	3.94001	0.638	10.484	10.1
14	74.14275	87.839	189.532	301.4	57.002	280.894	117.316	27.662	5.747	0	3.93919	18.998	65.438	22.9
15	74.17025	375.428	264.462	113.664	31.721	9.365	117.383	79.734	312.098	2	3.94059	27.23	26.719	7.8
16	74.18375	256.078	36.216	147.144	17.347	10.292	88.448	22.54	86.074	0	3.94087	12.584	14.027	8.5
17	74.18175	63.497	28.366	46.721	41.198	7.763	106.614	3.548	80.032	0	3.94087	8.589	2.317	9.0
18	74.16325	18.218	125.812	113.03	3.899	147.672	138.157	45.82	31.552	0	3.94045	2.642	2.523	70.3
19	74.16475	9.847	81.132	95.554	5.219	44.721	87.227	69.126	24.451	2	3.94045	3.4	3.127	66.
20	74.18075	17.826	58.316	98.21	12.145	38.168	102.118	36.277	109.761	1	3.94059	4.606	1.121	19.
21	74.17825	237.9	113.316	105.044	53.454	58.362	57.451	14.581	16.177	0	3.94095	7.767	6.552	9.5
22	74.18325	30.191	39.89	99.329	44.169	33.875	40.603	80.193	136.705	2	3.94095	2.931	2.63	3.9
23	74.15725	24.304	101.318	81.339	46.438	69.241	111.145	10.554	10.198	0	3.94059	0.643	4.319	71.6
24	74.16025	9.481	79.068	36.633	56.816	136.157	190.466	35.171	62.868	1	3.94059	5.075	2.646	11.0
25	74.15775	28.415	39.345	66.856	107.703	79.003	60.821	40.273	67.939	0	3.94059	5.246	1.249	10.7
26	74.15775	4.59	34.981	50.172	49.465	115.557	114.996	51.261	72.677	2	3.94059	3.867	2.902	9.6
27	74.19125	285.98	53.998	256.175	61.787	281.715	43.886	71.61	11.372	2	3.94227	34.886	10.413	7.7

Analysis Structure and Methodology

01

Investigation/EDA

02

Hypothesis Formation

03

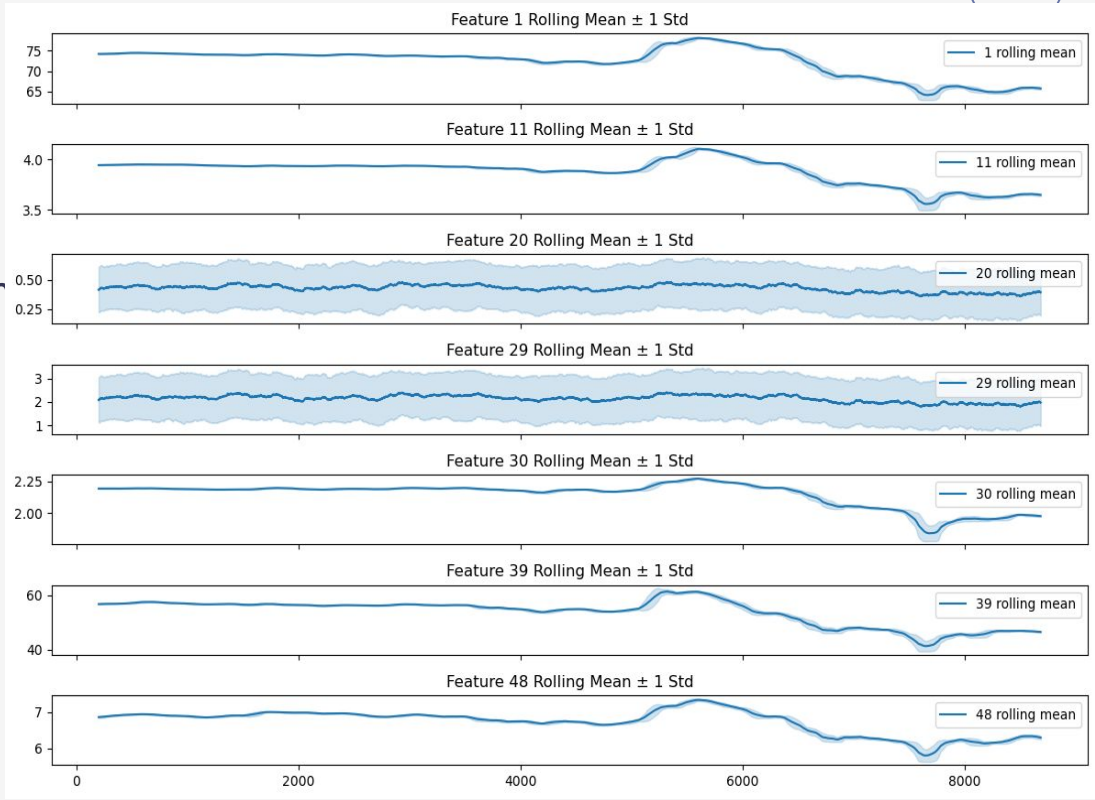
Hypothesis Testing

04

Prediction Model Fitting

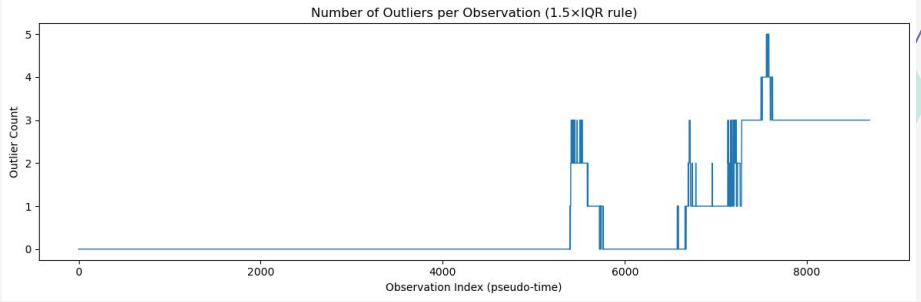
Investigations #1

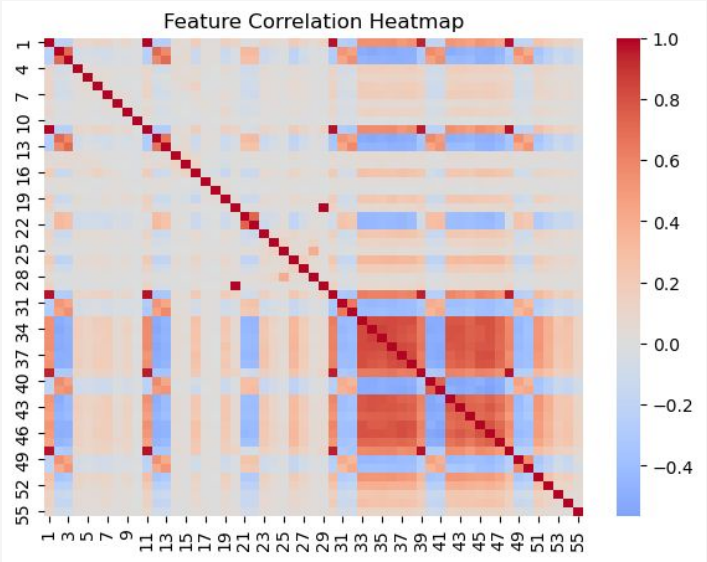
- Checked for monotonic features to infer time
- Introduced a sequential tim index as a working assumption
- Verified that row order reflects real system evolution rather than random observations



Investigations #2

- Outlier-heavy features identified
- Outliers cluster in time, not randomly
- Heavy-tailed distributions observed
- Strong correlations reveal shared drivers





Hypothesis Testing

Hypotheses

There is a regime shift in data

✅ **VERIFIED**(Binary Segmentation)

There exists long-term equilibrium relationships

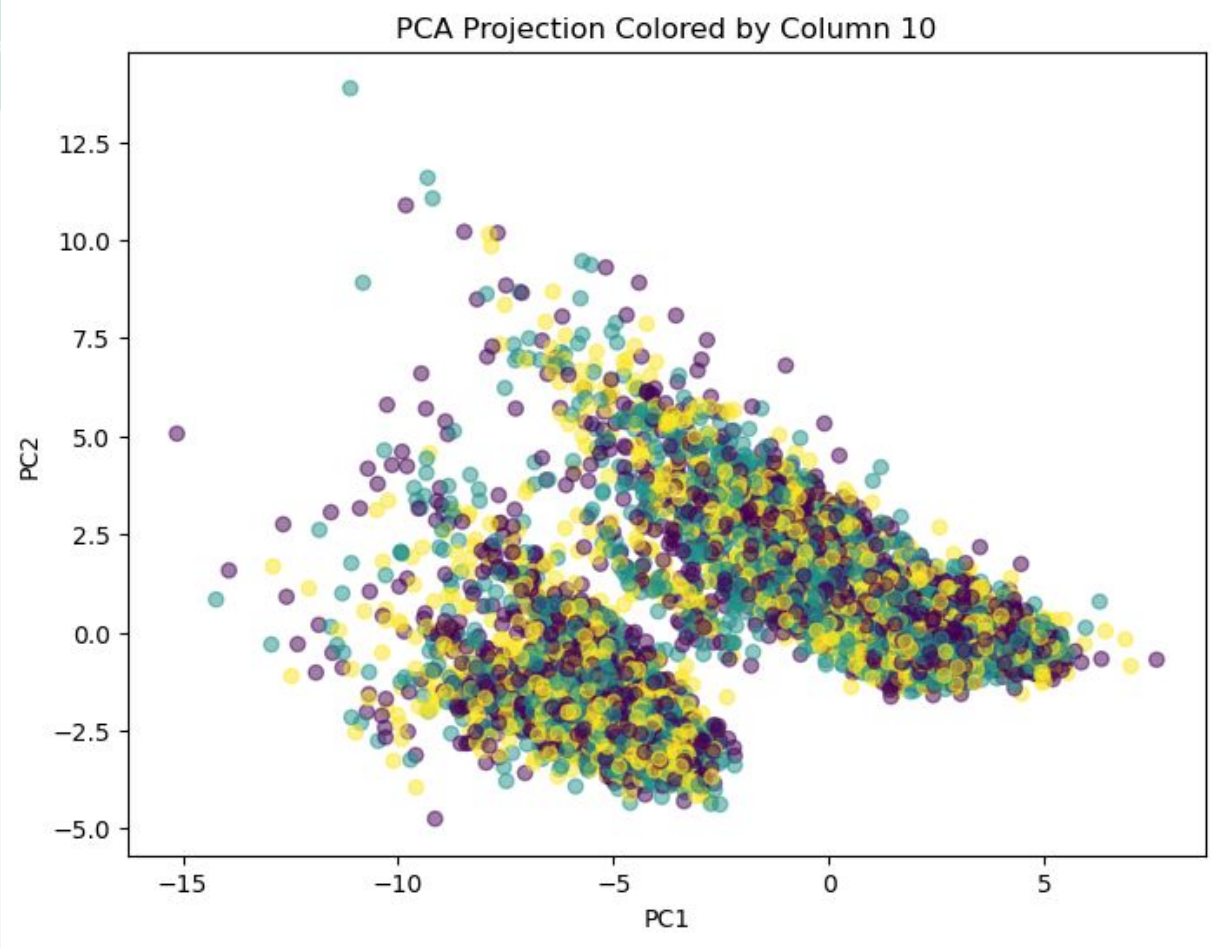
✅ **VERIFIED** (Johansen Test)

There exists time-series like structure

✅ **VERIFIED** (Ljung Box Test)

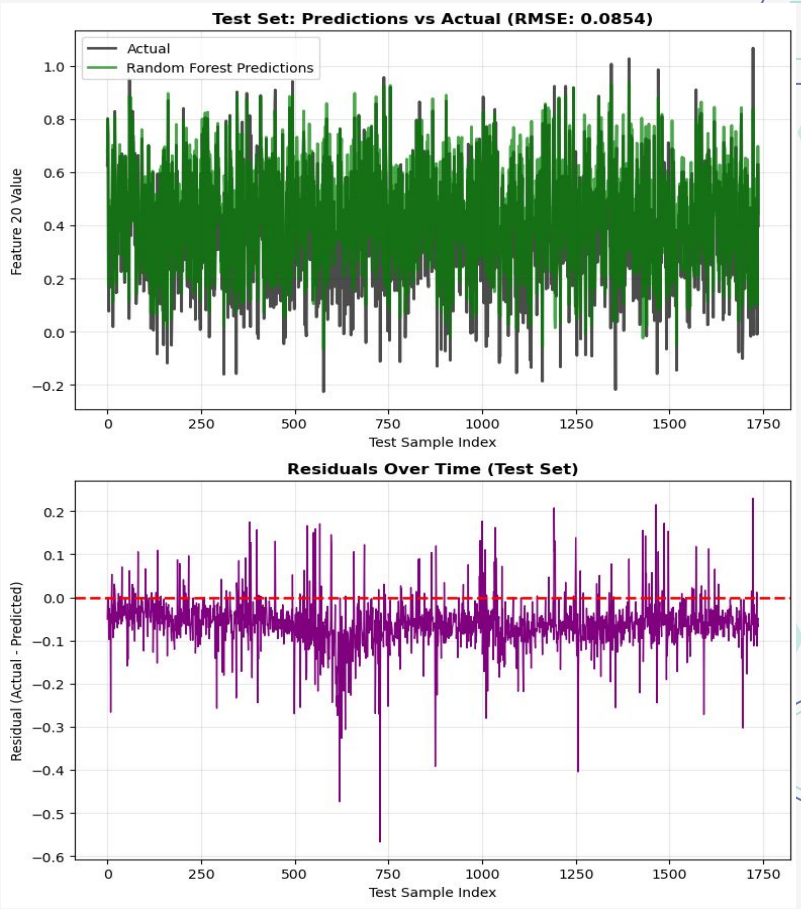
Col 10 forms clusters

❌ **NOT VERIFIED** (ANOVA + Visual Plot)


PCA Projection Colored by Column 10
PC2
PC1
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−15
−10
−5
0
5

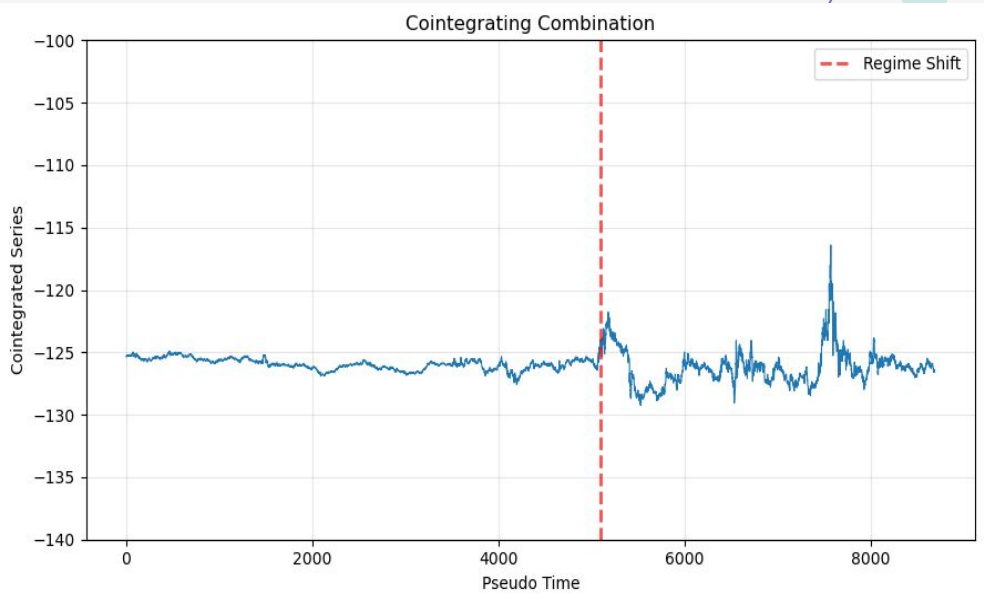
Machine Learning Prediction Modelling

- Predictions closely track actual behavior over time
- Model captures short-term fluctuations without lag
- Strong explanatory power ($R^2 \approx 0.83$)
- Enables reliable short-term monitoring and early detection of behavioral shifts



Key Takeaways and Business Implications

- Coordinated metric changes and rising outliers signal real regime changes, not noise- can be some external event.
- Strong correlation and cointegration show many signals reflect the same underlying driver.
- Stationary features provide clean baselines for monitoring and alerts.
- ML captures near-term behavior well, useful for early warnings rather than long-term forecasts.



Conclusion and Next Steps

This project shows how starting with no context, careful investigation can reveal structure, detect change, and extract insights that are useful for real decision-making.

Potential next steps in this project include:

1) Expand feature coverage
   Looking at more features and relationships between those.

2) Metric consolidation and dimensionality reduction
   Use the discovered correlations and cointegration structure to collapse redundant metrics into a smaller, more interpretable set.