Ecosustainable Energy Production

STAT 430: Unsupervised Learning

Ilia Lomasov, Bota Kabiyeva, Vahe Mouradian

About the project

- **Gas turbine** in Turkey's northwestern region
- **2000 instances of 11 sensor measures** aggregated over one hour

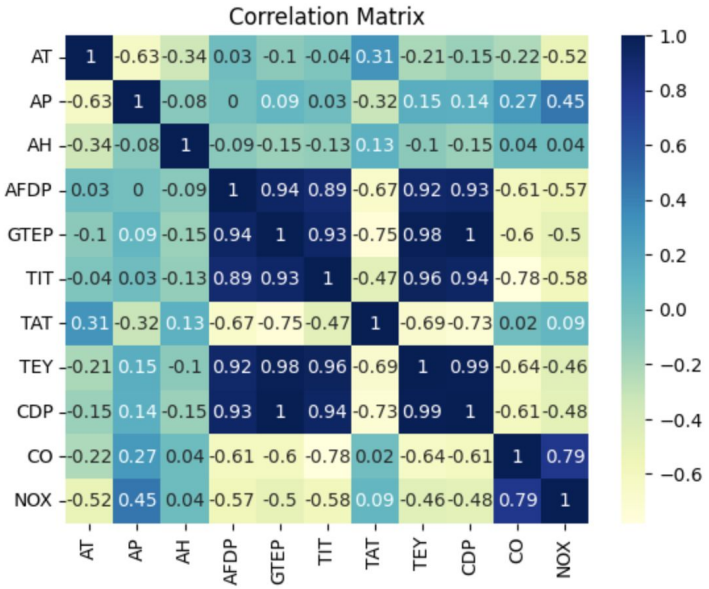
Variable	Abbr.	Unit
Ambient temperature	AT	°C
Ambient pressure	AP	mbar
Ambient humidity	AH	(%)
Air filter difference pressure	AFDP	mbar
Gas turbine exhaust pressure	GTEP	mbar
Turbine inlet temperature	TIT	°C
Turbine after temperature	TAT	°C
Compressor discharge pressure	CDP	mbar
Turbine energy yield	TEY	MWH
Carbon monoxide	CO	mg/m^3
Nitrogen oxides	NO_x	mg/m^3

- **GOAL:** identify general clustering **structure**, as well as clusters with:
  - High **TEY**, low **CO & NOx** (desirable)
  - Low **TEY**, high **CO & NOx** (undesirable)

- **IMPORTANT FOR**:
  - Emissions **regulation**
  - Reducing total measurement **costs**

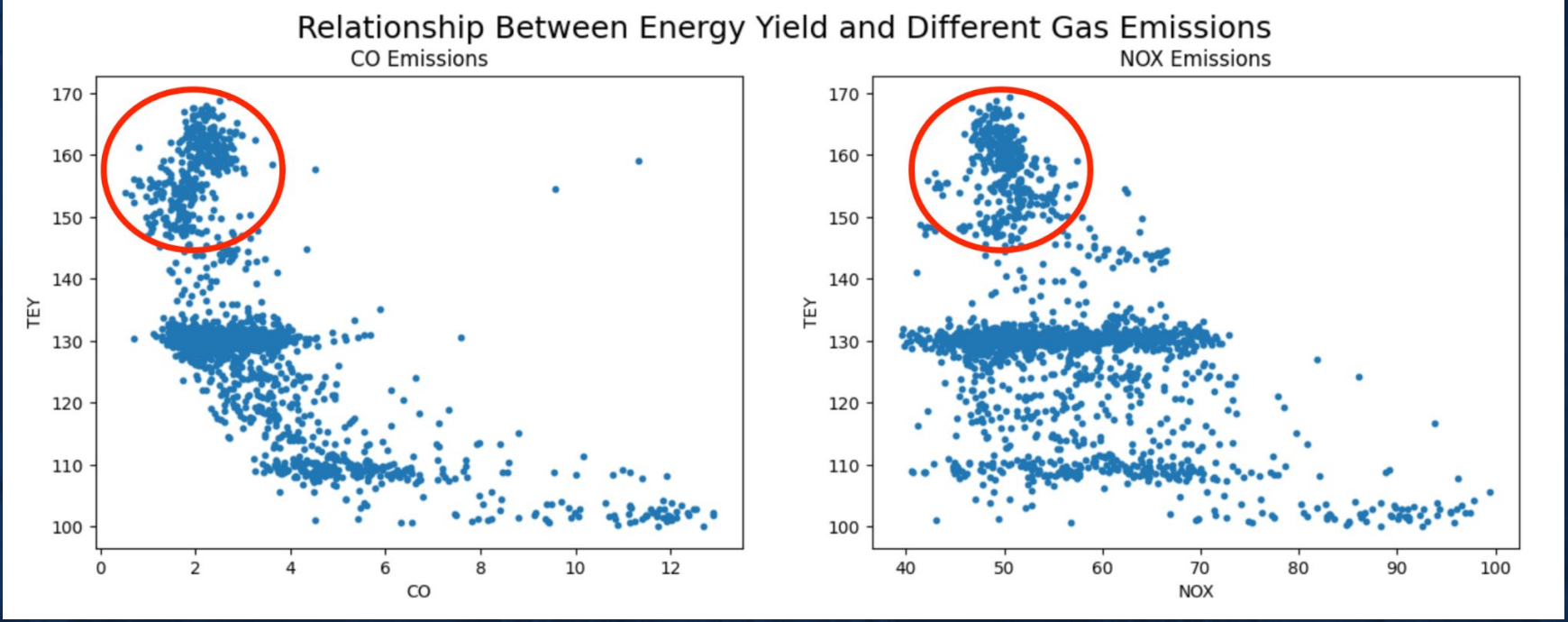
EDA of the data

- The **scale of variables** is very different - **mean-scaling** is needed
- Existence of **highly-correlated** variables
- Looked at pairwise scatter plots

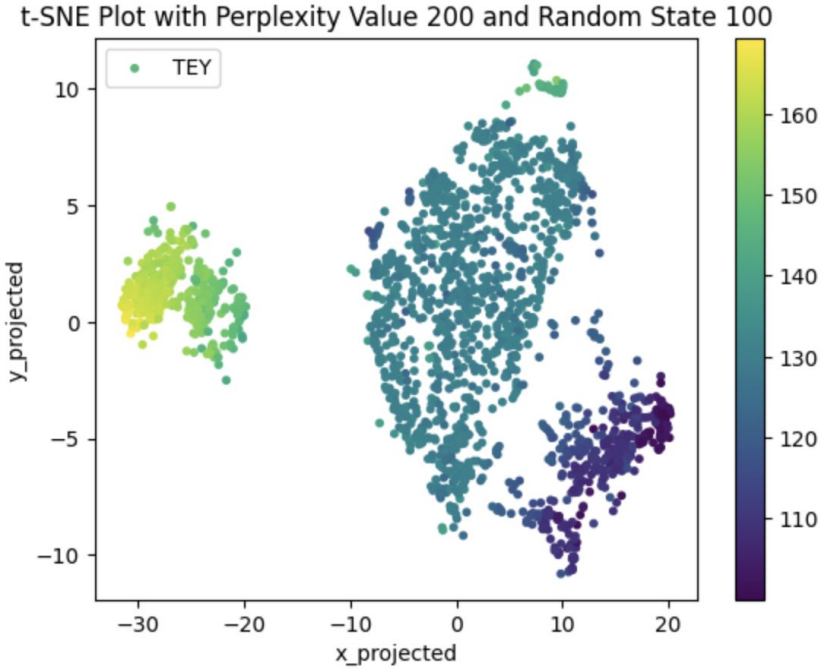


	AT	AP	AH	AFDP	GTEP	TIT	TAT	TEY	CDP	CO	NOX
count	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000	1965.000000
mean	14.480507	1018.535099	74.536957	3.520607	24.417182	1074.869160	546.917746	131.269863	11.883762	3.206167	56.271015
std	5.589764	7.693844	10.385808	0.560008	4.073701	17.560745	5.822548	15.612174	1.078619	1.955137	9.511217

Observations we are interested in:



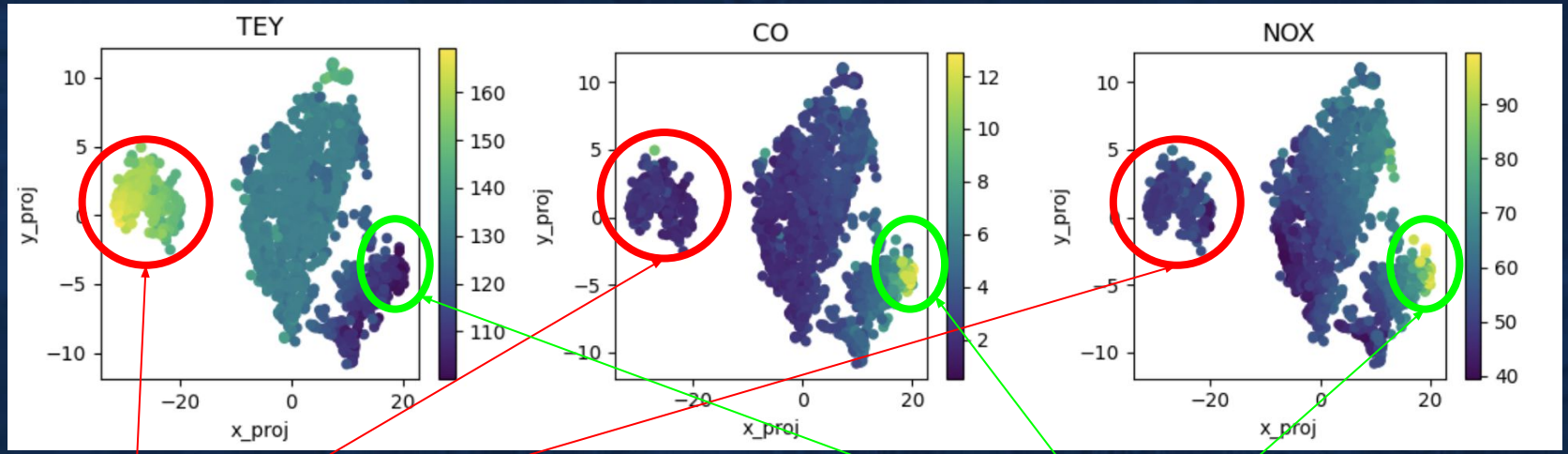
Clusterability Analysis



- Both Hopkins (~0.08) and t-SNE suggested that the dataset is **CLUSTERABLE**

- **STRUCTURE**:
  - **3 main clusters** + 2-3 subclusters in each
  - **Convex** shaped
  - Similar density, but **different size**
  - Main clusters - **well-separated**; subclusters - may be overlapping
  - According to t-SNE - not many noisy points

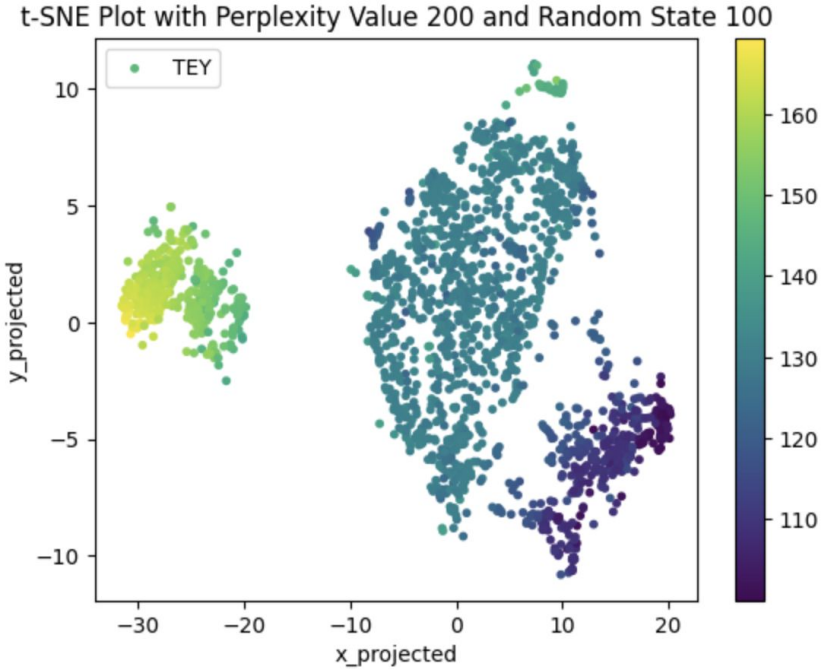
Observations we are interested in:



Max Energy - Min Emissions

Min Energy - Max Emissions

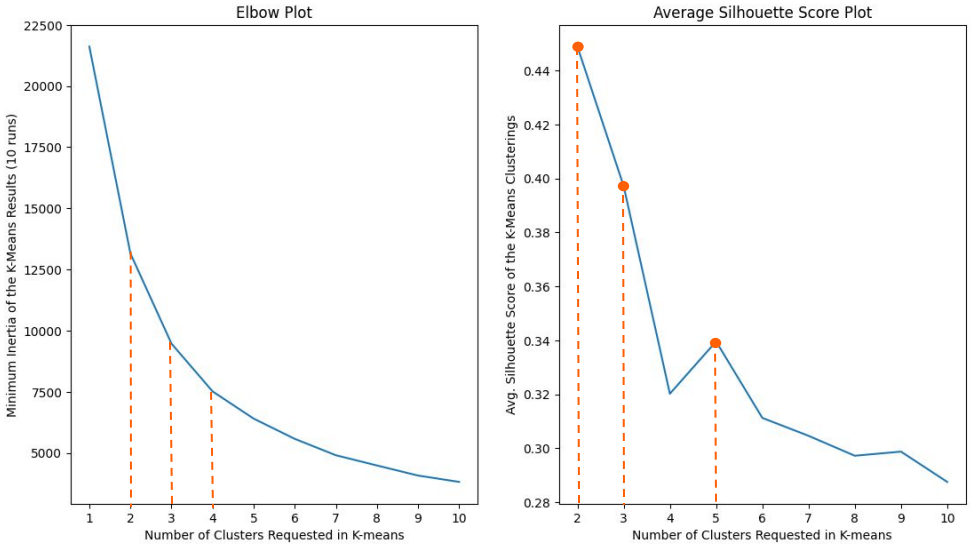
Algorithm selection motivation

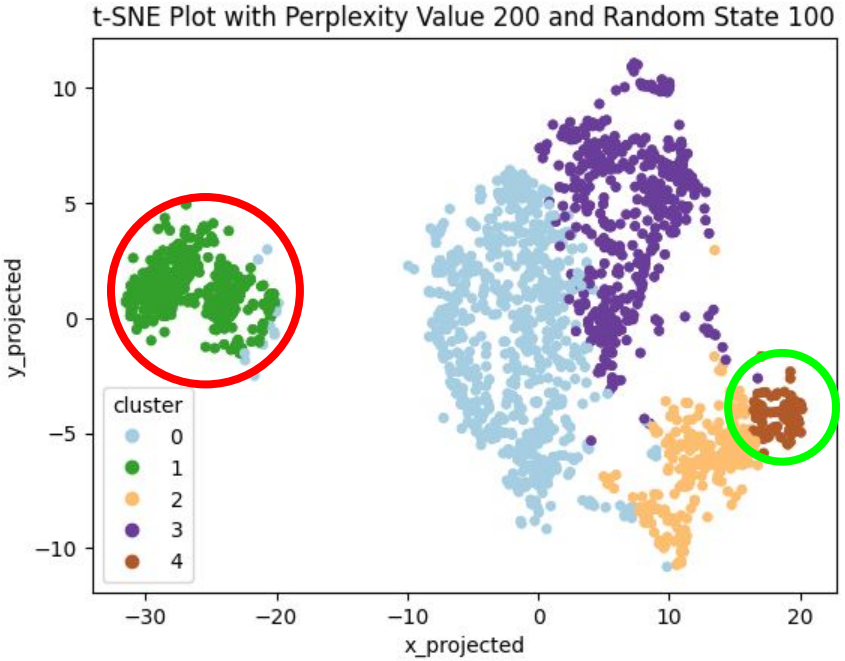


- K-Means:
  - **convex** shape of clusters
  - same **sparsity**
  - well-**separated**
  - little to **no fuzziness**
- HAC:
  - **nested** structure
  - clusters have **different sizes**
- DBSCAN:
  - doesn't assume **number of clusters**
  - clusters have **different sizes**
  - same **sparsity**
  - potential **noise** points

K-Means

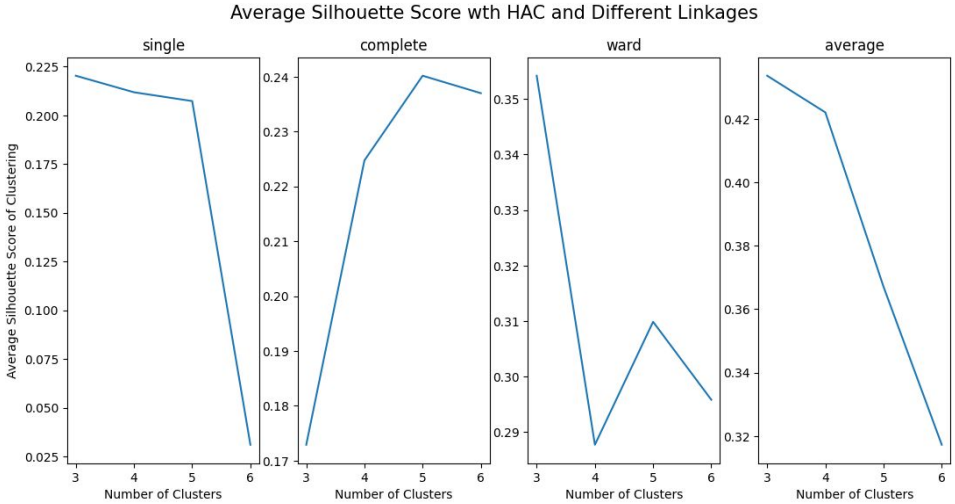
Elbow Method Results & Average Silhouette Scores for Turbine Data

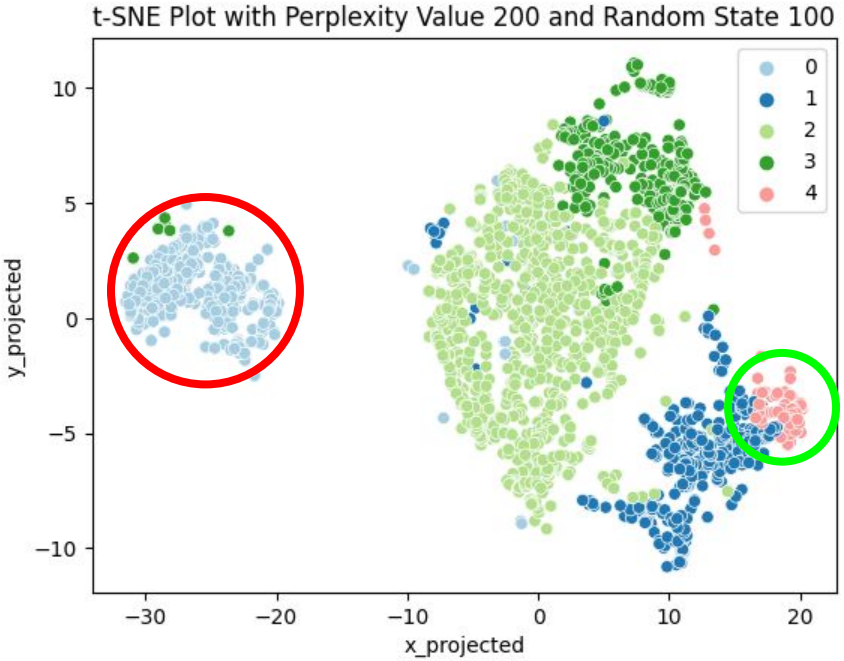




Elbow method suggested **k = 2, 3 or 4**
Silhouette plot suggested **k = 2, 3 or 5**
t-SNE plot comparison suggested **k = 5**

HAC with Ward Linkage

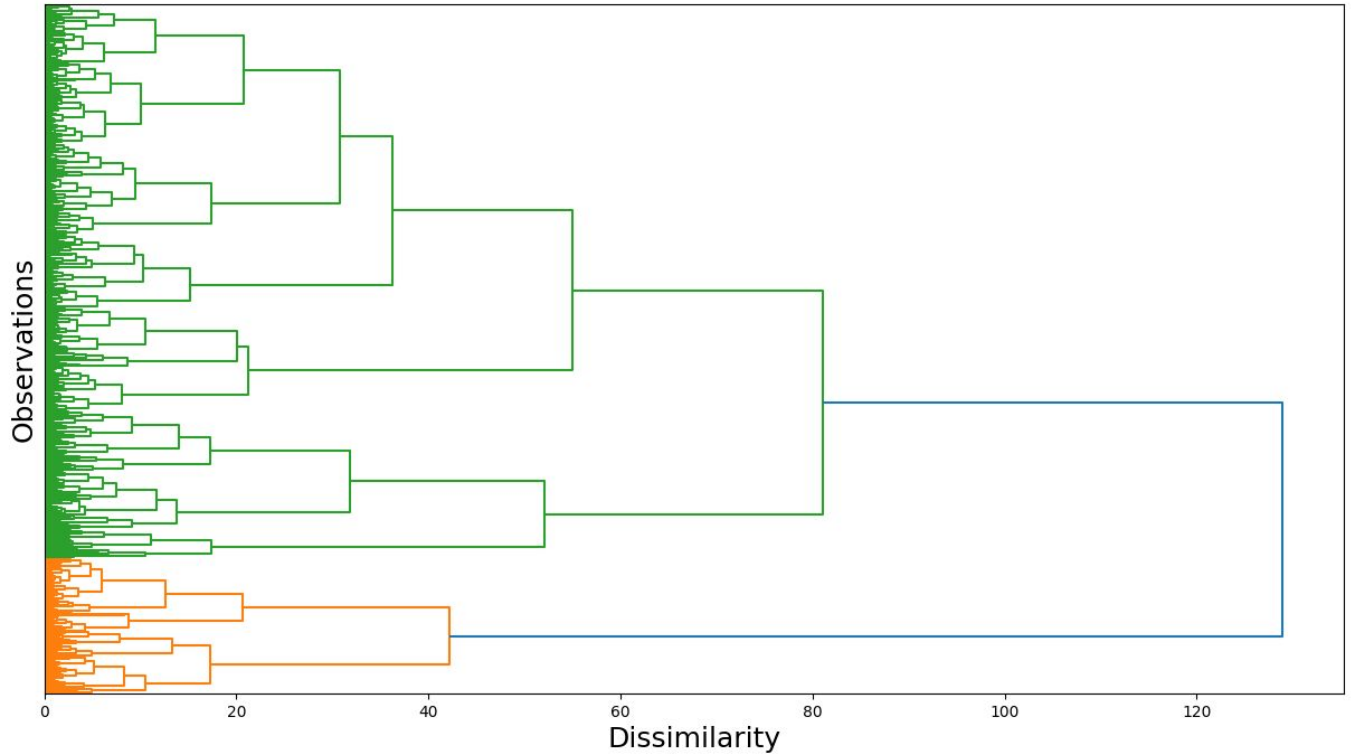




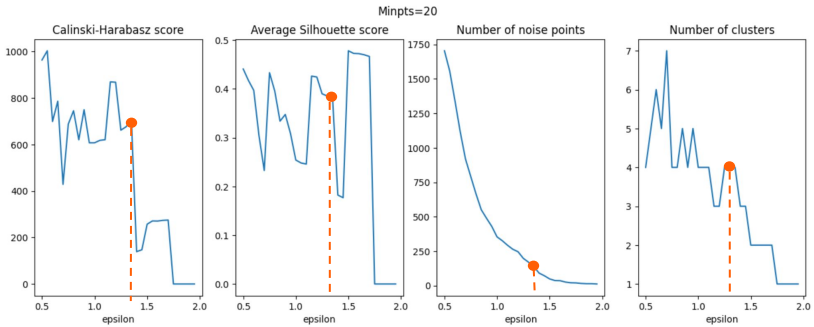
Average linkage had highest average Silhouette scores

Ward linkage was the most **meaningful**

Dendrogram for HAC with Ward Linkage



DBSCAN

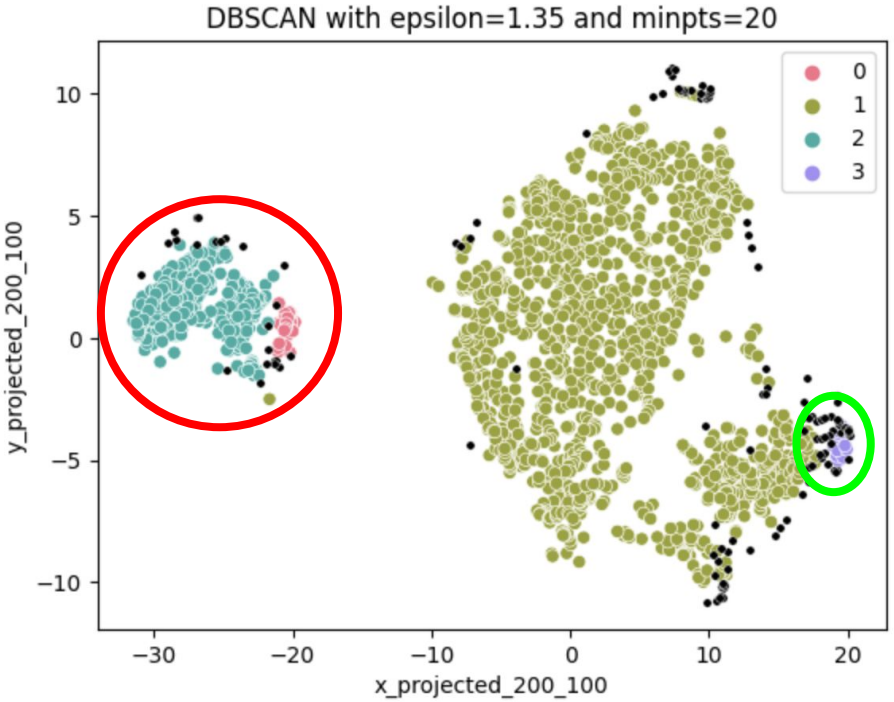


Parameters: **minpts=20, eps=1.35**

Output: 4 clusters, **139 noise** points

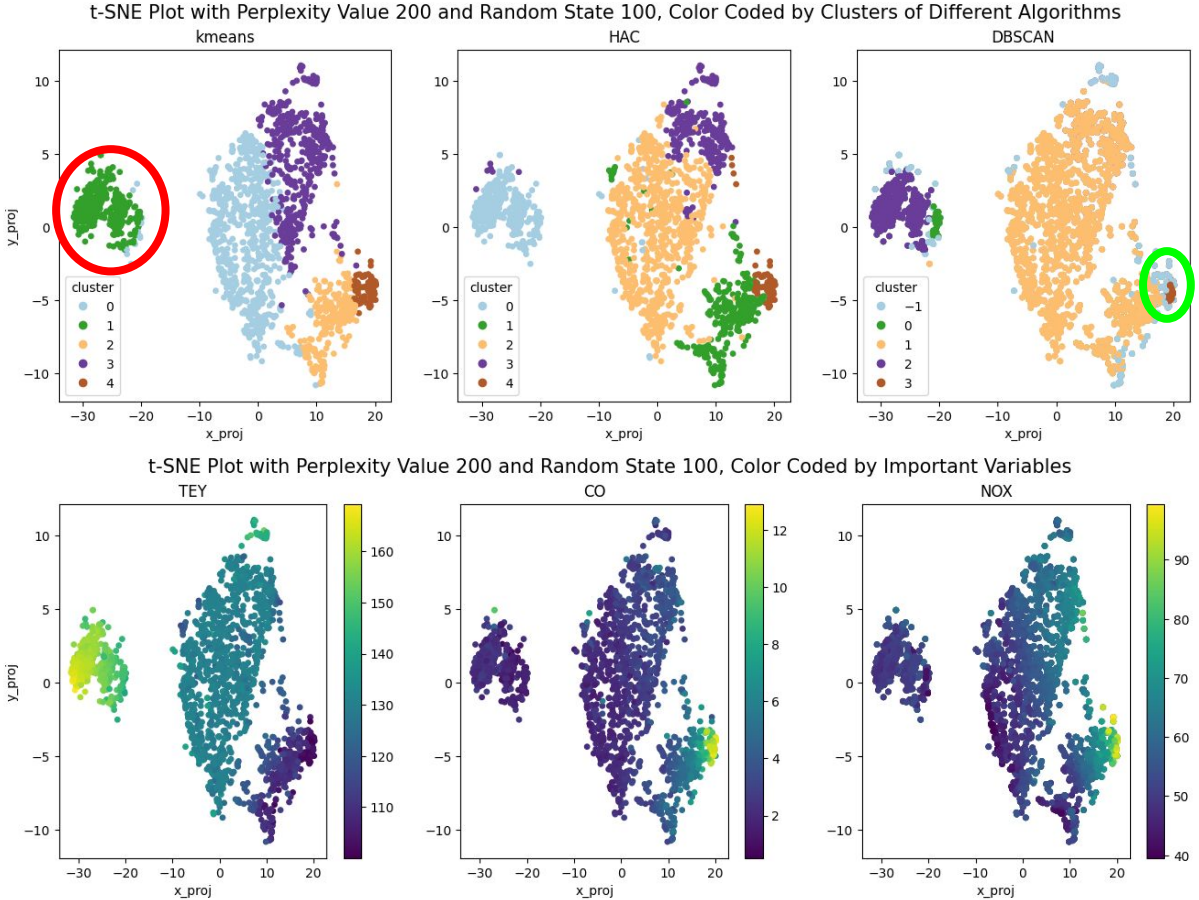
Summary:

- Separated the data into **4 clusters** - aligns with t-SNE
- Identified "**extreme**" clusters



Comparison

- **General Clustering Structure:**
  - K-Means did the best
  - Difference in separation of the middle cloud
- **“Extreme” observations**:
  - Leftmost: K-Means and HAC
  - Rightmost: DBSCAN
- **Average Silhouette Score:**
  - 0.34 for K-Means
  - 0.31 for HAC
  - 0.30 for DBSCAN



Summary

Clustering Structure

- All algorithms found 4-5 cluster - the data has underlying clustering structure
- Compared the similarities and differences of each cluster
- Helpful to estimate the energy output and emissions for the turbine.

Observing extreme subsets

- All the clustering algorithms - labeled extreme sets
- Variables like "AT" or "AP" don’t affect TEY and emissions values
- "AFDP" or "TAT" are highly associated with energy yield
- Maintaining those parameters at the level specified in this subsets, the factory could regulate the emissions as well as associated costs.

Thanks!