Continued:

- Understanding the overall "structure" of the data
  - PCA → Reducing Dimensions. / Visualization.
  - t-SNE → Visualization.

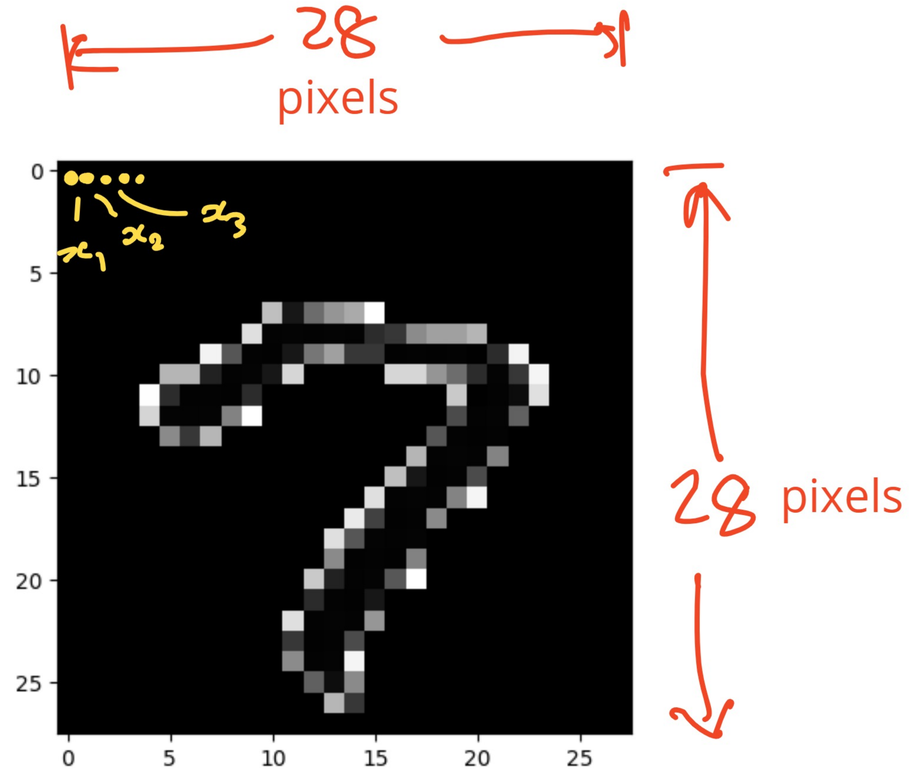
(Dimension reduction only for visualization and clustering))

- Feature Transformation
  - Encoding & Binning.

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# The MNIST Data Set

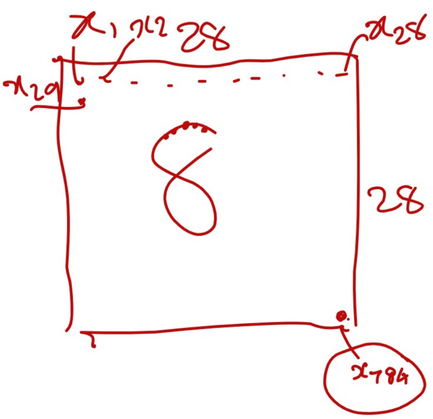
| | A | EX | EY | EZ | FA | FB | FC | FD | FE | FF | FG | FH | FI | FJ | FK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | label | 6x13 | 6x14 | 6x15 | 6x16 | 6x17 | 6x18 | 6x19 | 6x20 | 6x21 | 6x22 | 6x23 | 6x24 | 6x25 | 6x26 |
| 2 | 5 | 3 | 18 | 18 | 18 | 126 | 136 | 175 | 26 | 166 | 255 | 247 | 127 | 0 | 0 |
| 3 | 0 | 0 | 0 | 48 | 238 | 252 | 252 | 252 | 237 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 67 | 232 | 39 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 124 | 253 | 255 | 63 | 0 | 0 | 0 | 0 |
| 6 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 2 | 0 | 0 | 0 | 13 | 25 | 100 | 122 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1 | 237 | 253 | 252 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 3 | 43 | 105 | 255 | 253 | 253 | 253 | 253 | 253 | 174 | 6 | 0 | 0 | 0 | 0 |
| 10 | 1 | 5 | 63 | 197 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 143 | 247 | 153 | 0 | 0 |
| 12 | 3 | 254 | 254 | 254 | 254 | 254 | 66 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 3 | 155 | 155 | 131 | 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 6 | 38 | 178 | 252 | 253 | 117 | 65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 168 | 242 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 2 | 164 | 211 | 250 | 250 | 194 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 203 | 229 | 32 | 0 | 0 | 0 |
| 20 | 6 | 0 | 0 | 75 | 247 | 143 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 4 | 0 | 0 | 0 | 0 | 0 | 112 | 252 | 125 | 4 | 0 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 96 | 205 | 251 | 253 | 205 | 111 | 4 | 0 | 0 | 0 |
| 24 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 121 | 254 | 136 | 0 | 0 | 0 | 0 | 0 |
| 26 | 1 | 0 | 0 | 29 | 249 | 254 | 254 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 2 | 246 | 253 | 253 | 253 | 253 | 253 | 220 | 154 | 17 | 3 | 0 | 0 | 0 | 0 |
| 28 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | 3 | 207 | 255 | 254 | 254 | 254 | 97 | 80 | 80 | 44 | 0 | 0 | 0 | 0 | 0 |




MNIST stands for “Modified National Institute of Standards and Technology” database. It is a large database of small, square 28x28 pixel grayscale images of handwritten single digits between 0 and 9. The MNIST database contains 60,000 training images and 10,000 testing images, with each image labeled with the respective digit that it represents.

How to UNDERSTAND the structure of such a large data set?

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

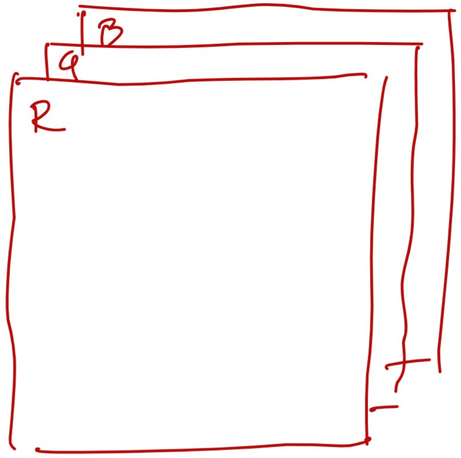



convert → (flattened)

[ 56 ... ]

[ 0 0 0 0 76 84 ... 200 ... x784 ]

MNIST → thousands of images · (labeled) ·



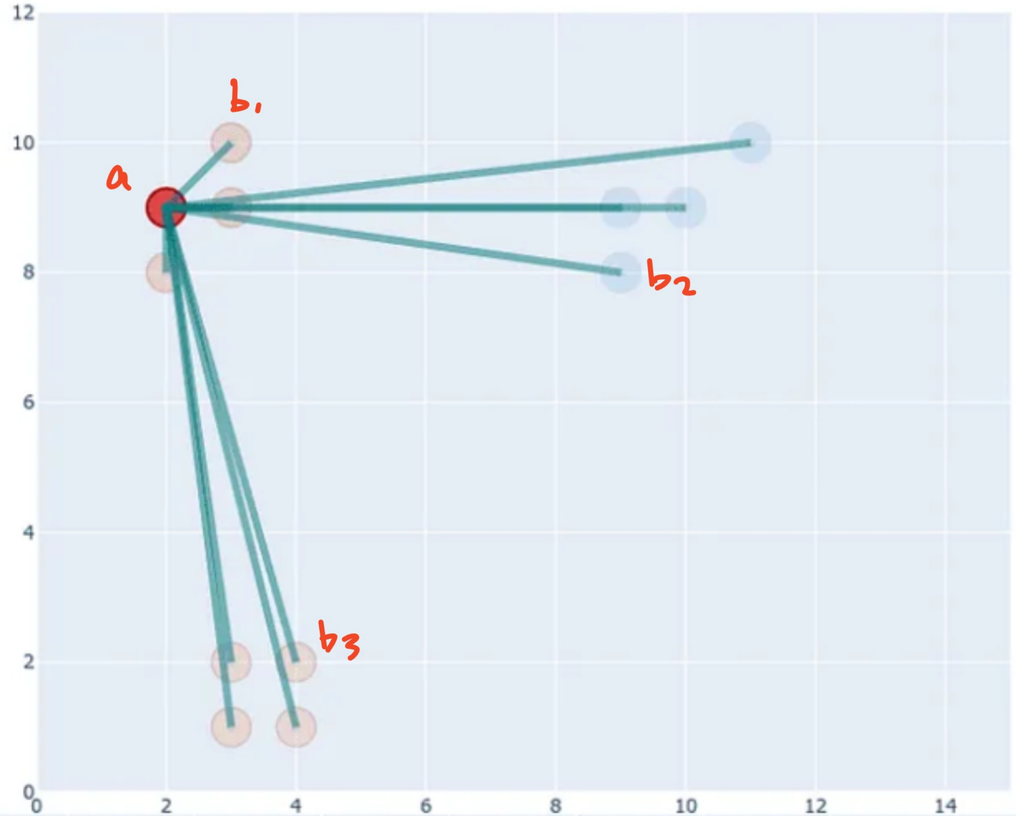

[

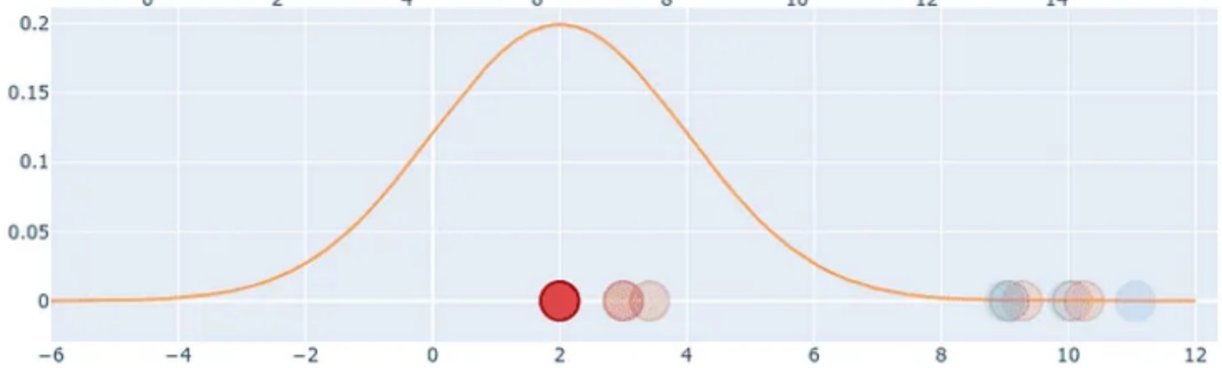
784 x 3 =

]

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

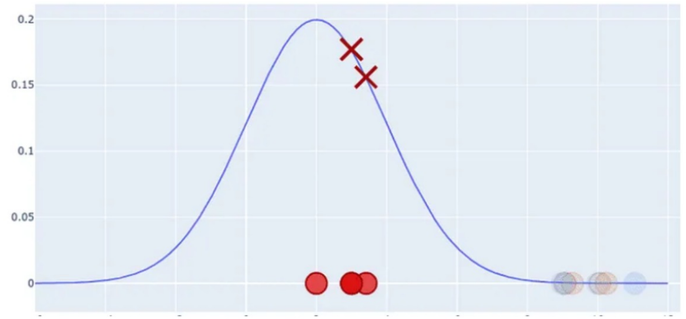
# t-SNE
## (t-distributed Stochastic Neighbour Encoding)



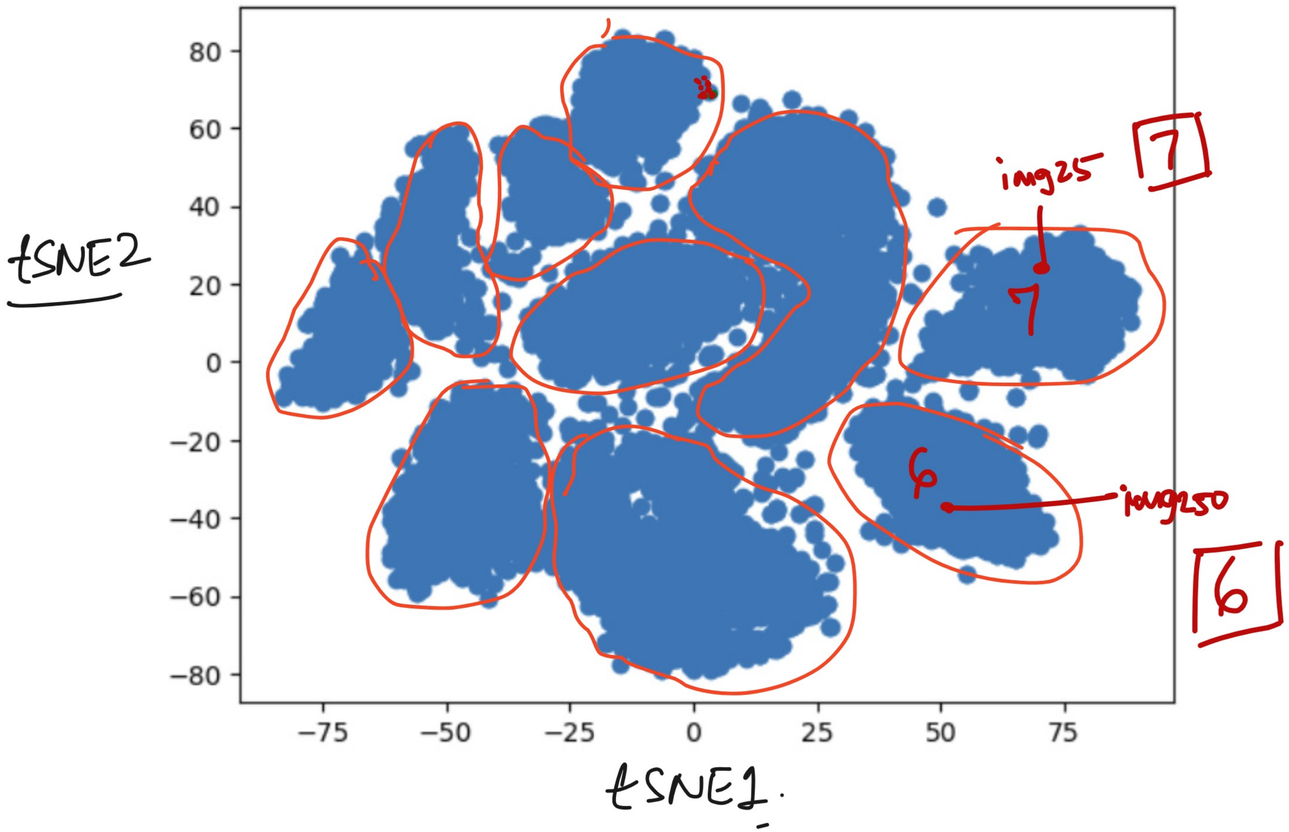
- t-SNE is a machine learning algorithm that is **used for dimensionality reduction and data visualization**.
- It works by finding the similarity measure between pairs of instances in higher and lower dimensional spaces, and tries to maintain the probability distribution for data samples in lower dimensions the same as the probability distribution of data samples in higher dimensions.
- The main advantage of t-SNE is the **ability to preserve local structure**, meaning that points which are close to one another in the high-dimensional data set will tend to be close to one another in the chart - which aspect is advantageously used for visualization.

- (A 'simulation' of the t-SNE algorithm is provided later in this slide deck)




This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# MNIST dataset visualized using t-SNE



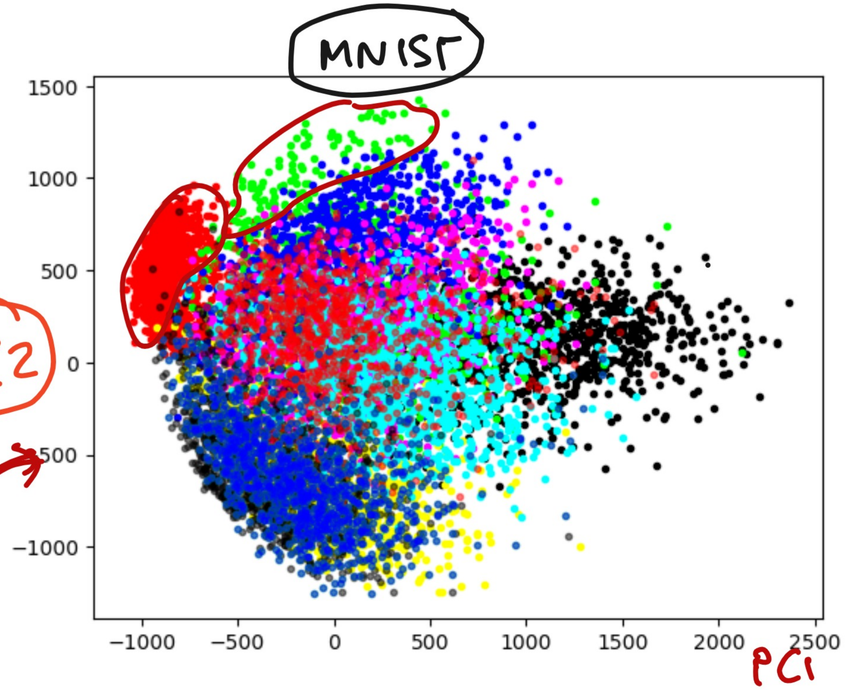

- t-SNE based visualization tells us that there are about 10 clusters in the MNIST dataset.
- Once this is known, we can pass the t-SNE components of the dataset through a clustering algorithm like KMeans (with K = 10) in order to group the observations into clusters.
- Once the clustering is done, detailed analysis of the cluster (in this case review of images within the cluster) can help us assign meaningful labels to the clusters (eg: '0', '1', '2', etc.)
- t-SNE can thus help us understand the 'structure' of the data, help us create labels if none existed earlier, and decide on the further course of action.

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

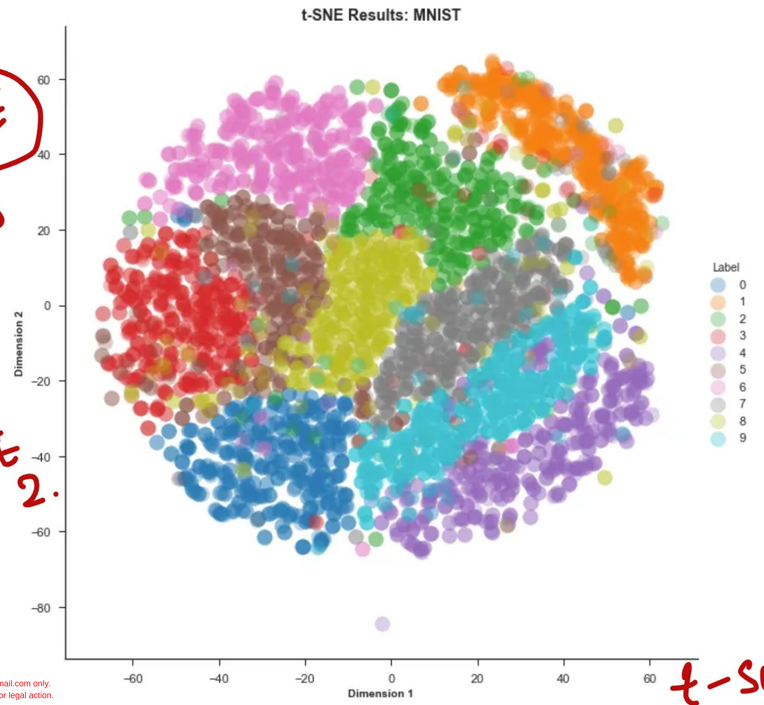
Visualization of MNIST data using PCA and t-SNE

## MNIST Data Set

| | A | EX | EY | EZ | FA | FB | FC | FD | FE | FF | FG | FH | FI | FJ | FK |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | label | 6x13 | 6x14 | 6x15 | 6x16 | 6x17 | 6x18 | 6x19 | 6x20 | 6x21 | 6x22 | 6x23 | 6x24 | 6x25 | 6x26 |
| 2 | 5 | 3 | 18 | 18 | 18 | 126 | 136 | 175 | 26 | 166 | 255 | 247 | 127 | 0 | 0 |
| 3 | 0 | 0 | 0 | 48 | 238 | 252 | 252 | 252 | 237 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 67 | 232 | 39 | 0 | 0 | 0 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 124 | 253 | 255 | 63 | 0 | 0 | 0 | 0 |
| 6 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 2 | 0 | 0 | 0 | 13 | 25 | 100 | 122 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1 | 237 | 253 | 252 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 3 | 43 | 105 | 255 | 253 | 253 | 253 | 253 | 253 | 174 | 6 | 0 | 0 | 0 | 0 |
| 10 | 1 | 5 | 63 | 197 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 143 | 247 | 153 | 0 | 0 |
| 12 | 3 | 254 | 254 | 254 | 254 | 254 | 66 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 3 | 155 | 155 | 131 | 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 6 | 38 | 178 | 252 | 253 | 117 | 65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 1 | 168 | 242 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 2 | 164 | 211 | 250 | 250 | 194 | 15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 203 | 229 | 32 | 0 | 0 | 0 |
| 20 | 6 | 0 | 0 | 75 | 247 | 143 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 4 | 0 | 0 | 0 | 0 | 0 | 112 | 252 | 125 | 4 | 0 | 0 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 96 | 205 | 251 | 253 | 205 | 111 | 4 | 0 | 0 | 0 |
| 24 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 121 | 254 | 136 | 0 | 0 | 0 | 0 | 0 |
| 26 | 1 | 0 | 0 | 29 | 249 | 254 | 254 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 2 | 246 | 253 | 253 | 253 | 253 | 253 | 220 | 154 | 17 | 3 | 0 | 0 | 0 | 0 |
| 28 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | 3 | 207 | 255 | 254 | 254 | 254 | 97 | 80 | 80 | 44 | 0 | 0 | 0 | 0 | 0 |



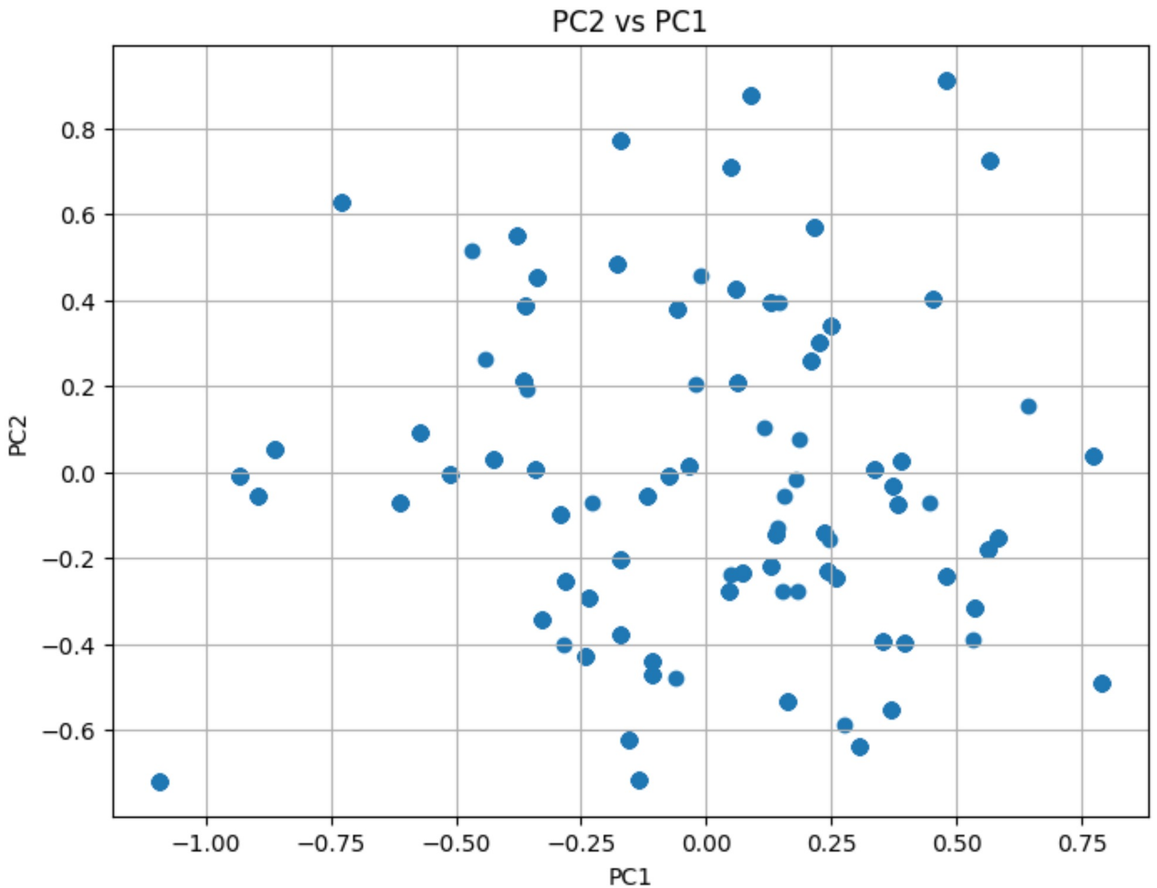

Scatter plot of PC1 and PC2

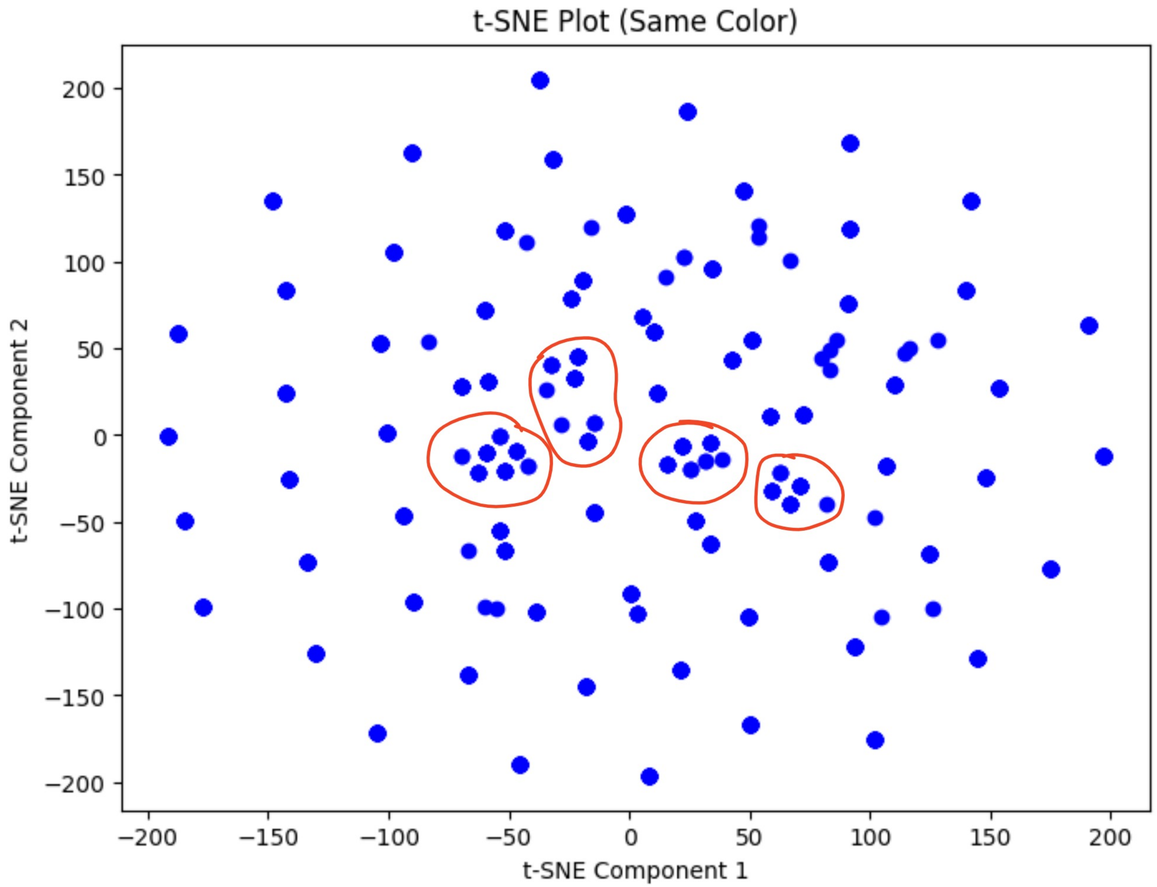



Note: Usually, PCA precedes t-SNE; Principle Components can be used as input to t-SNE

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

The 'blood test' dataset

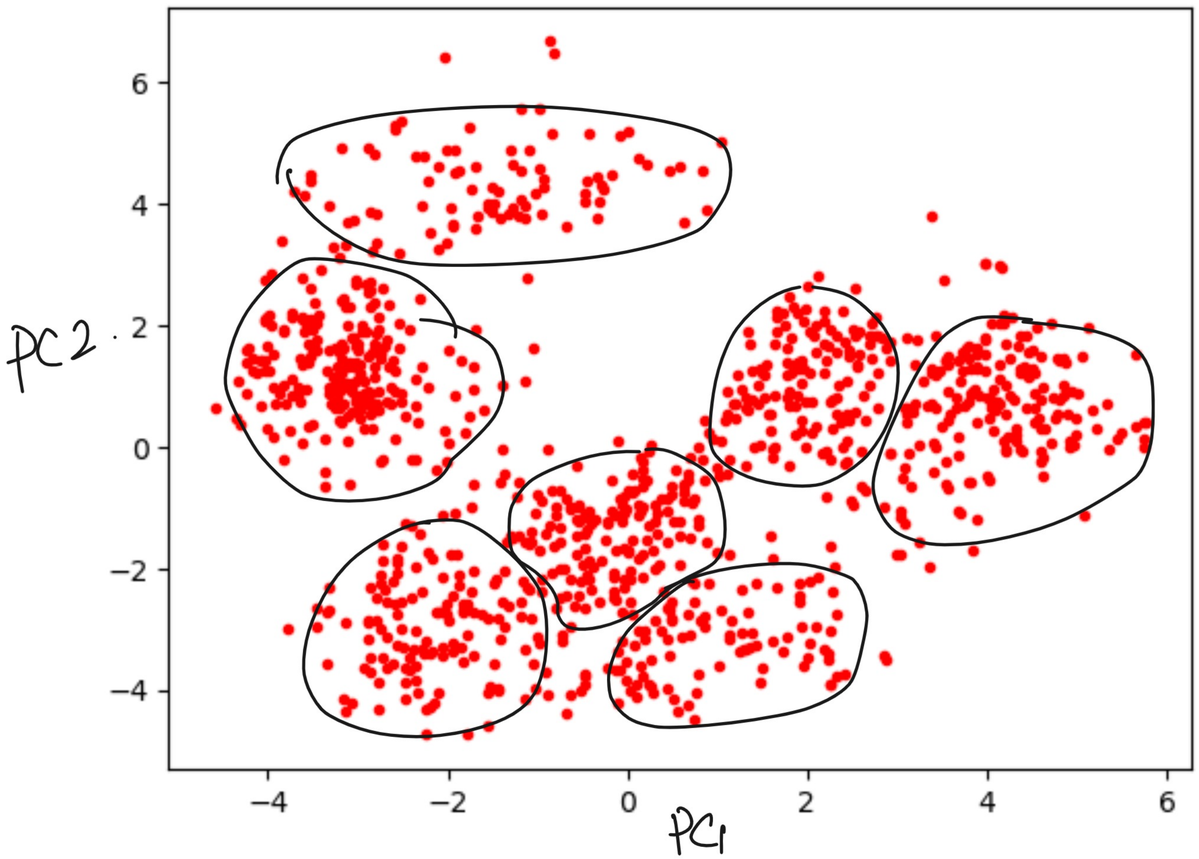


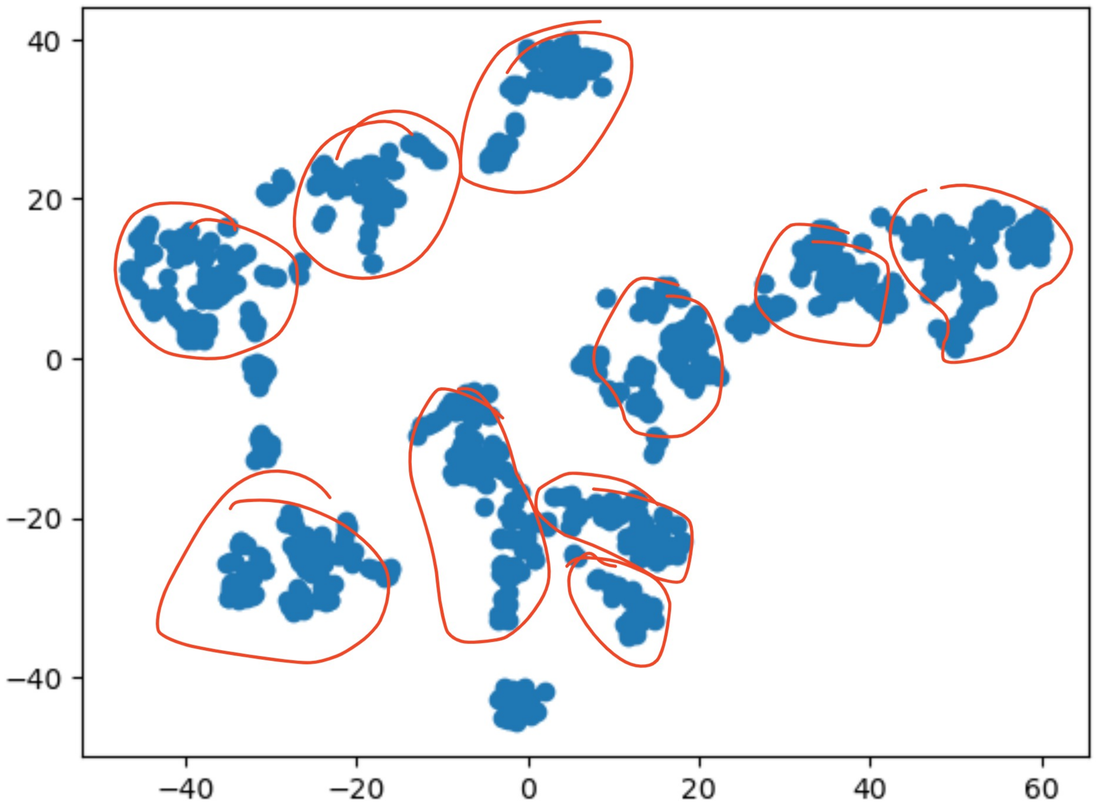




No major clusters are seen in either PC-based or t-SNE based visualization
(though t-SNE does throw up some minor clusters. What could they be?)

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# Chemical process plant dataset



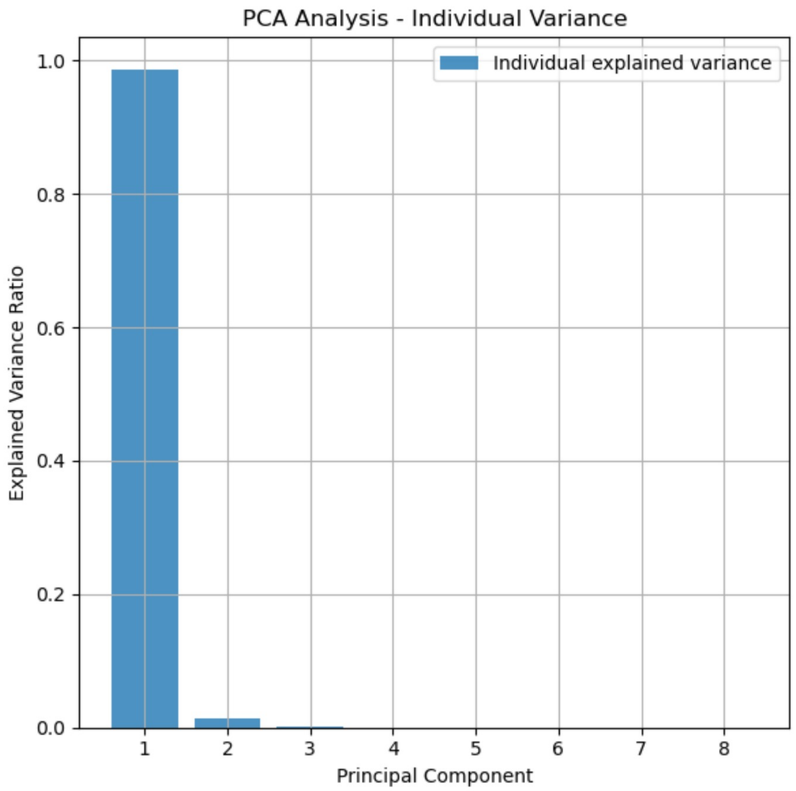



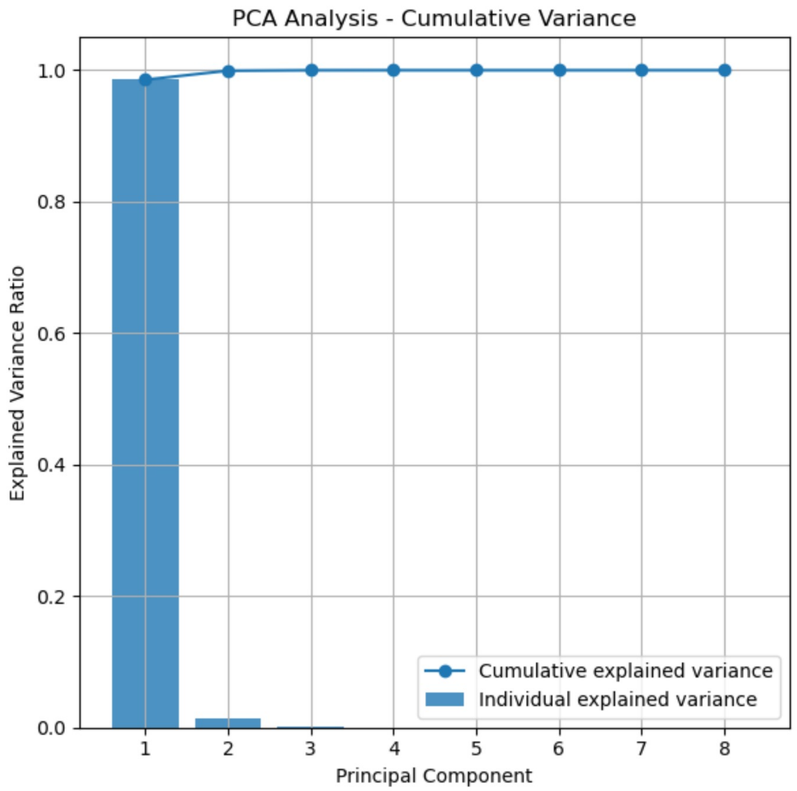

We can observe many distinct clusters in both, PC2 v/s PC1 and t-SNE visualizations

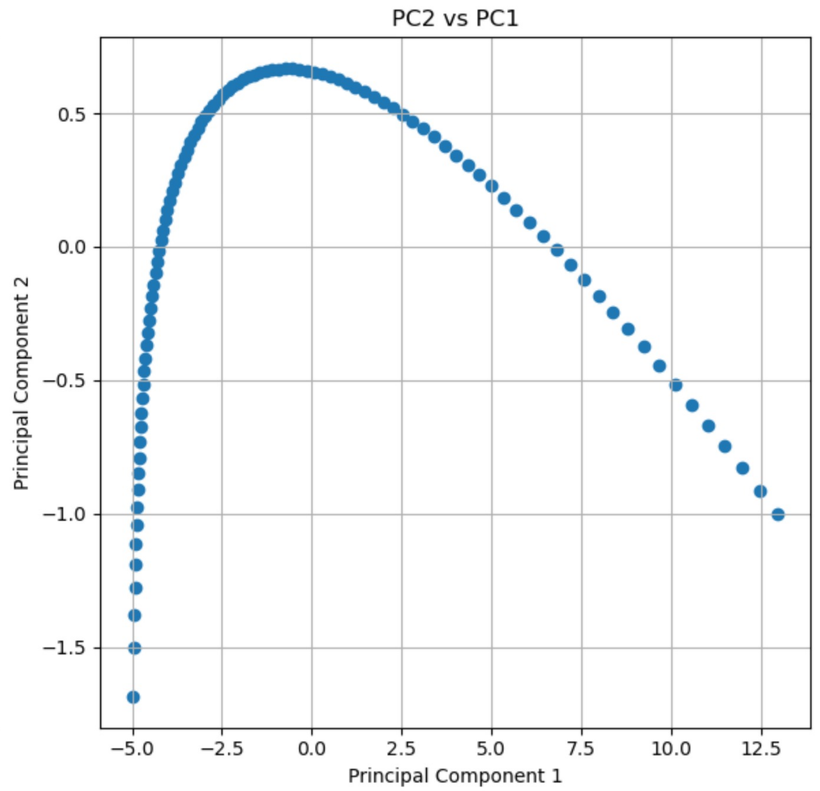
(The next steps involve investigating the clusters and 'labelling' the observations based on the investigation. Once labelled, the data set can be further used to create prediction models)

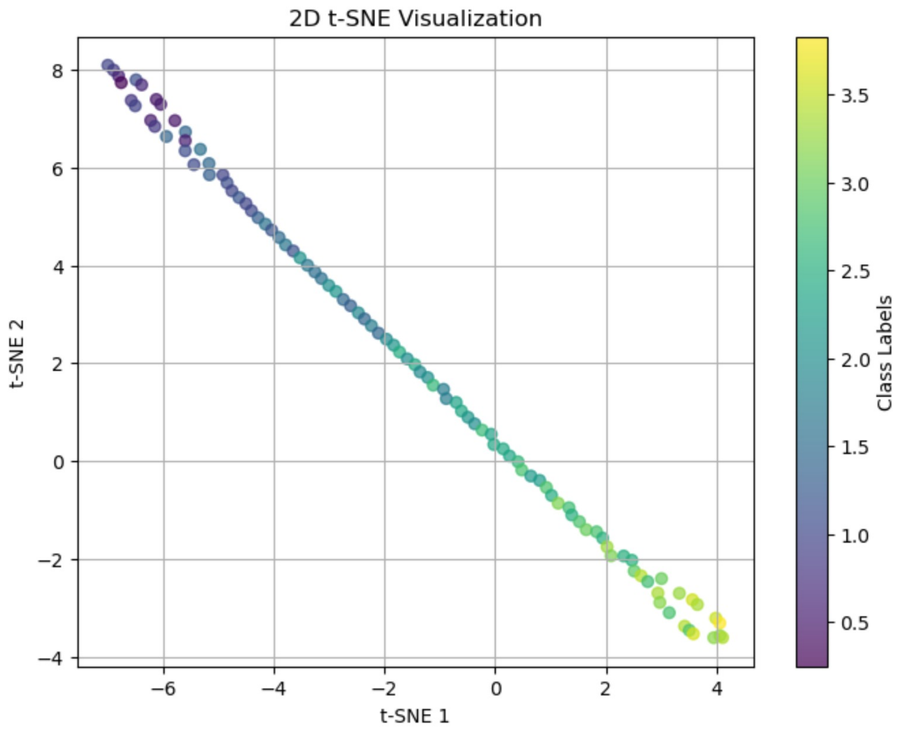
This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# The 'curse-of-dimensionality' data set - with 8 features




Observations:

- Even though the data set contains 8 features, PCA tells us that only one principle component is sufficient to capture the variance in the data, hinting at drastic dimensionality reduction.
- PC2 v/s PC1 also hints at non-linearity in the data set.
- The t-SNE plot also hints at the almost uni-dimensional nature of the data.
- It would be interesting to get the t-SNE plot of the Principle Components !!

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# PCA v/s t-SNE

| Feature | PCA (Principal Component Analysis) | t-SNE (t-Distributed Stochastic Neighbor Embedding) |
|---|---|---|
| **Primary Purpose** | Dimensionality reduction, feature extraction | Visualization of high-dimensional data in lower dimensions |
| **Transformation Type** | Linear | Non-linear |
| **Information Preservation** | Preserves global variance | Preserves local similarities (can distort global relationships) |
| **Interpretability** | Relatively easy to interpret components | Difficult to interpret components |
| **Deterministic/Stochastic** | Deterministic (same results on repeated runs) | Stochastic (different results on repeated runs) |
| **Suitability for Prediction Models** | Suitable as a feature extraction method | **Generally unsuitable for direct use in prediction models** |
| **Distance Preservation** | Aims to preserve overall distances | Distorts global distances, focuses on local distances |
| **Use Cases** | Feature reduction, noise reduction, data compression | Visualizing clusters, exploring data structure, detecting patterns |

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

# 'Simulation' of the t-SNE algorithm

Assume 3 points in 'n dimensional' space (n = 3 in this example)

We would like to 'visualize' these points in 2D

We have three points in 3D space:

| Point | Coordinates |
|---|---|
| A | (1,1,1) |
| B | (2,1,1) |
| C | (5,5,5) |

We will calculate the Gaussian probability distribution for each point by:

1. Computing pairwise Euclidean distances
2. Applying the Gaussian similarity formula
3. Normalizing the probabilities so they sum to 1

For each point $x_i$, the probability of another point $x_j$ being its neighbor is given by:

$$p_{j|i} = \frac{\exp(-d_{ij}^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-d_{ik}^2/2\sigma_i^2)}$$

where:

- $d_{ij}$ is the Euclidean distance between $x_i$ and $x_j$.
- $\sigma_i$ is a local scaling factor (we will assume a fixed value for simplicity).
- The denominator ensures probabilities sum to 1 for each row.

First, let's calculate the pairwise Euclidean distances:

$$d(A,B) = \sqrt{(2-1)^2 + (1-1)^2 + (1-1)^2} = \sqrt{1} = 1$$

$$d(A,C) = \sqrt{(5-1)^2 + (5-1)^2 + (5-1)^2} = \sqrt{16+16+16} = \sqrt{48} \approx 6.93$$

$$d(B,C) = \sqrt{(5-2)^2 + (5-1)^2 + (5-1)^2} = \sqrt{9+16+16} = \sqrt{41} \approx 6.40$$

The distance matrix:

| | A | B | C |
|---|---|---|---|
| A | 0 | 1 | 6.93 |
| B | 1 | 0 | 6.40 |
| C | 6.93 | 6.40 | 0 |

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

## 4. Compute Gaussian Probabilities

Let's assume σ = 2 for all points (for simplicity).

The Gaussian similarity formula:

$$p_{j|i} = \frac{\exp(-d_{ij}^2/2\sigma^2)}{\sum_{k \neq i} \exp(-d_{ik}^2/2\sigma^2)}$$

Now, let's calculate for each point:

**For Point A:**

$$\exp(-1^2/2(2)^2) = \exp(-1/8) \approx 0.882$$

$$\exp(-6.93^2/2(2)^2) = \exp(-48.02/8) = \exp(-6.00) \approx 0.0025$$

$$\text{Sum} = 0.882 + 0.0025 = 0.8845$$

$$p(B|A) = 0.882/0.8845 \approx 0.997$$

$$p(C|A) = 0.0025/0.8845 \approx 0.003$$

**For Point B:**

$$\exp(-1^2/2(2)^2) = \exp(-1/8) \approx 0.882$$

$$\exp(-6.40^2/2(2)^2) = \exp(-40.96/8) = \exp(-5.12) \approx 0.0059$$

$$\text{Sum} = 0.882 + 0.0059 = 0.8879$$

$$p(A|B) = 0.882/0.8879 \approx 0.993$$

$$p(C|B) = 0.0059/0.8879 \approx 0.007$$

**For Point C:**

$$\exp(-6.93^2/2(2)^2) = \exp(-6.00) \approx 0.0025$$

$$\exp(-6.40^2/2(2)^2) = \exp(-5.12) \approx 0.0059$$

$$\text{Sum} = 0.0025 + 0.0059 = 0.0084$$

$$p(A|C) = 0.0025/0.0084 \approx 0.298$$

$$p(B|C) = 0.0059/0.0084 \approx 0.702$$

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

## 5. Final Probability Matrix

| | A | B | C |
|---|---|---|---|
| A | - | **0.997** | 0.003 |
| B | 0.993 | - | 0.007 |
| C | 0.298 | 0.702 | - |

In the low-dimensional space (2D), we **randomly initialize** the positions of A, B, and C. Let's say:

- $y_A = (0.5, 0.5)$
- $y_B = (0.6, 0.5)$
- $y_C = (2.0, 2.0)$

Now, we need to **measure the similarities in 2D.**

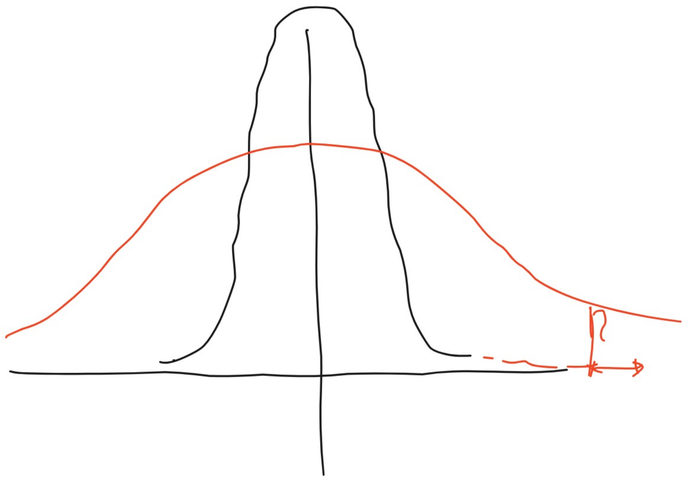
Instead of using a **Gaussian distribution**, t-SNE uses a **t-distribution** with one degree of freedom:

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

This ensures that:

- **Nearby points have high probability.**
- **Far away points still have some probability** (thanks to long tails in t-distribution).

Normal v/s t-distribution



This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

We now have:

1. High-Dimensional Probabilities $p_{ij}$ (from the Gaussian)
2. Low-Dimensional Probabilities $q_{ij}$ (from the t-distribution)

t-SNE wants $q_{ij}$ to match $p_{ij}$, meaning similarities in 2D should reflect those in high dimensions.

We measure the difference using KL Divergence:

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

KL = Kullback Leibler

- If $p_{ij} > q_{ij}$ → The points are too far apart in 2D → Move them closer.
- If $p_{ij} < q_{ij}$ → The points are too close together in 2D → Push them apart.



This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

1. Compute Similarities in High-Dimensional Space (Gaussian-based $p_{ij}$)
2. Initialize Points Randomly in 2D
3. Compute Similarities in Low-Dimensional Space (t-distribution-based $q_{ij}$)
4. Minimize KL Divergence Between $p_{ij}$ and $q_{ij}$
5. Use Gradient Descent to Move Points
6. Iterate Until Convergence
7. Visualize the Final 2D Map

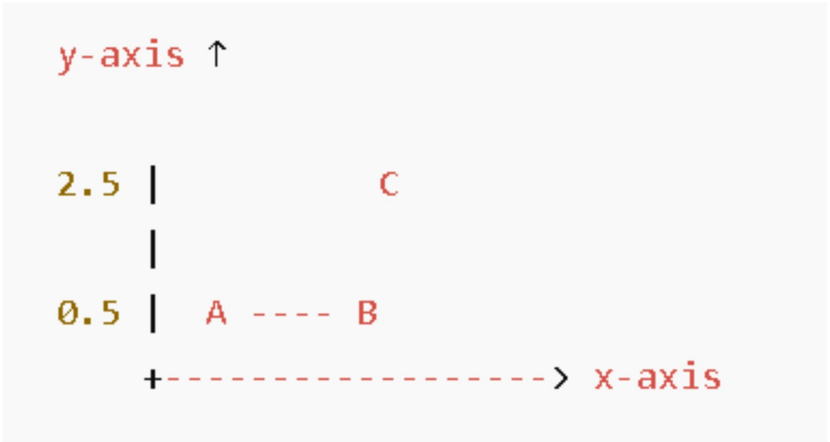
To reduce KL divergence, t-SNE moves the points gradually by computing the gradient:

$$\Delta y_i = \eta \sum_j (p_{ij} - q_{ij})(y_i - y_j)$$

where:

- $\eta$ is the learning rate (controls how much we move each step).
- The movement is in the direction that reduces KL divergence.

1. Start with random 2D positions.
2. Compute similarities in 2D (q-values using t-distribution).
3. Compare with high-dimensional similarities (p-values using Gaussian).
4. Move points using gradient descent to reduce KL divergence.
5. A and B stay close, C moves far.
6. Iteration continues until stable positioning is found.




This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.