DATA VISUALIZATION

PYTHON DATA VISUALIZATION:

## VISUALIZE NATURAL GROUPINGS OR CLUSTERS.

## COURSE NO.

## STUDENT NAME

## INSTRUCTOR NAME

Table of Contents

[VISUALIZE NATURAL GROUPINGS OR CLUSTERS. 1](#_Toc47567490)

[COURSE NO. 1](#_Toc47567491)

[STUDENT NAME 1](#_Toc47567492)

[INSTRUCTOR NAME 1](#_Toc47567493)

[INTRODUCTION 3](#_Toc47567494)

[Cluster analysis 3](#_Toc47567495)

[Dimensional Reduction 4](#_Toc47567496)

[t-Distributed Stochastic Neighbor Embedding (t-SNE) 4](#_Toc47567497)

[UMAP Uniform Manifold Approximation and Projection 5](#_Toc47567498)

[MNIST\_SIGNS 5](#_Toc47567499)

[Dataset 5](#_Toc47567500)

[Data preprocessing 5](#_Toc47567501)

[T-SNE approach 6](#_Toc47567502)

[KMeans 6](#_Toc47567503)

[Mini Batch K Means 7](#_Toc47567504)

[Gaussian mixture 8](#_Toc47567505)

[UMAP approach 8](#_Toc47567506)

[K-means 8](#_Toc47567507)

[Gaussian Mixture 9](#_Toc47567508)

[Conclusion 9](#_Toc47567509)

# INTRODUCTION

Information perception is the realistic portrayal of information. It includes delivering pictures that convey connections among the spoke to information to watchers of the pictures. This correspondence is accomplished using a methodical planning between realistic imprints and information esteems in the formation of the representation. This planning sets up how information esteems will be spoken to outwardly, deciding how and how much a property of a realistic imprint, for example, size or shading, will change to reflect changes in the estimation of a datum.

To convey data unmistakably and effectively, information perception utilizes factual illustrations, plots, data designs and different instruments. Numerical information might be encoded utilizing specks, lines, or bars, to outwardly convey a quantitative message. Effective representation assists clients with breaking down and reason about information and proof. It makes complex information more open, justifiable and usable. Information representation is both a craftsmanship and a science.

# Cluster analysis

Cluster analysis or clustering is the errand of collection a lot of items so that objects in a similar gathering (called a group) are more comparative (in some sense) to one another than to those in different gatherings (groups). It is a fundamental assignment of exploratory information mining, and a typical strategy for measurable information investigation, utilized in numerous fields, including design acknowledgment, picture examination, data recovery, bioinformatics, information pressure, PC illustrations and AI.

Cluster analysis itself isn't one explicit calculation, yet the overall assignment to be settled. It very well may be accomplished by different calculations that contrast essentially in their comprehension of what establishes a bunch and how to effectively discover them.

# Dimensional Reduction

Dimensionality decrease, or feature reduction, is the change of information from a high-dimensional space into a low-dimensional space with the goal that the low-dimensional portrayal holds some significant properties of the first information, in a perfect world near its inborn measurement. Working in high-dimensional spaces can be bothersome for some reasons; crude information are regularly inadequate as a result of the scourge of dimensionality, and breaking down the information is generally computationally immovable. Dimensionality decrease is normal in fields that manage enormous quantities of perceptions as well as huge quantities of factors, for example, signal preparing, discourse acknowledgment, neuro informatics, and bioinformatics.

## t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-straight strategy for dimensionality decrease that is especially appropriate for the representation of high-dimensional datasets. It is widely applied in picture handling, NLP, genomic information and discourse preparing. To keep things straightforward, here's a concise review of working of t-SNE

* The calculations begins by ascertaining the likelihood of comparability of focuses in high-dimensional space and computing the likelihood of closeness of focuses in the relating low-dimensional space. The likeness of focuses is determined as the restrictive likelihood that a point A would pick point B as its neighbor if neighbors were picked in relation to their likelihood thickness under a Gaussian (typical dissemination) focused at A.
* It at that point attempts to limit the contrast between these restrictive probabilities (or likenesses) in higher-dimensional and lower-dimensional space for an ideal portrayal of information focuses in lower-dimensional space.
* To quantify the minimization of the whole of distinction of restrictive likelihood t-SNE limits the total of Kullback-Leibler difference of in general information focuses utilizing an angle plunge strategy.

## UMAP Uniform Manifold Approximation and Projection

UMAP represents Uniform Manifold Approximation and Projection. It's the new child on the dimensionality decrease obstruct (in 2018), and it is fundamentally the same as t-SNE. On the off chance that you contrast representations made and t-SNE and UMAP, you may make some hard memories disclosing to them separated.

* Notwithstanding, UMAP seems to have some critical favorable circumstances over t-SNE:
* It's quicker than t-SNE.
* It catches worldwide structure better than t-SNE.
* The best part is that while t-SNE doesn't have a lot of utilization outside of representation, UMAP is a broadly useful dimensionality decrease method that can be utilized as preprocessing for AI.
* UMAP likewise has a strong hypothetical support as a complex estimation strategy, while t-SNE is principally a representation heuristic.

The fundamental hindrance of UMAP is its absence of development. It is another strategy, so the libraries and best practices are not yet solidly settled or powerful. Be that as it may, in case you're willing to be an early adopter, UMAP has a ton to offer.

# MNIST\_SIGNS

## Dataset

The dataset is labeled dataset with 24 classes. Hence we cluster it into 24 classes.

## 

## Design Approach

### Problem

Finding natural clusters in the dataset.

### Tools

Using python 3, libraries, Pandas, Numpy, Sklearn.

### Visual

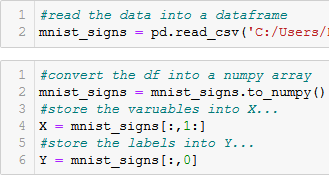
Plotly is charts plugin that creates interactive graphs which can be easily integrated into html, we display plotly scatter graph for each cluster algorithm

### Steps

* Load the data in a dataframe
* Convert it into array after preprocessing
* Using plolty to plot the results

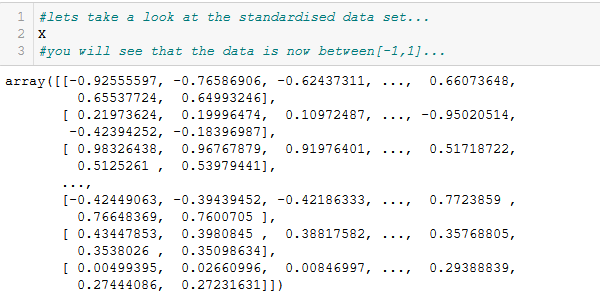
## Data preprocessing

The dataset contain 784 pixel features and 1 label feature. Hence we load the data using pandas dataframe and convert it into array using numpy.



We divide data into target Y and Features X.

The second step is data scaling here we use python sklearn library preprocessing package.



## T-SNE approach

We again use python library sklearn and manifold package,

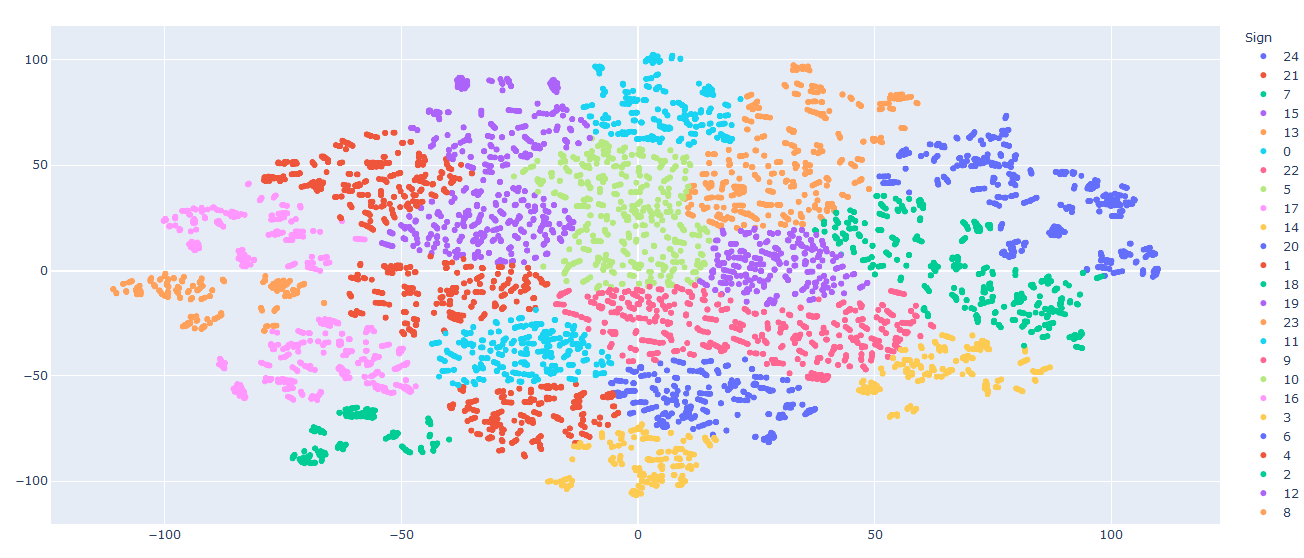


The X\_tsne variable stores the reduced array using the fit and transform function. This function fits over the data and then transforms it and saves it.

### KMeans

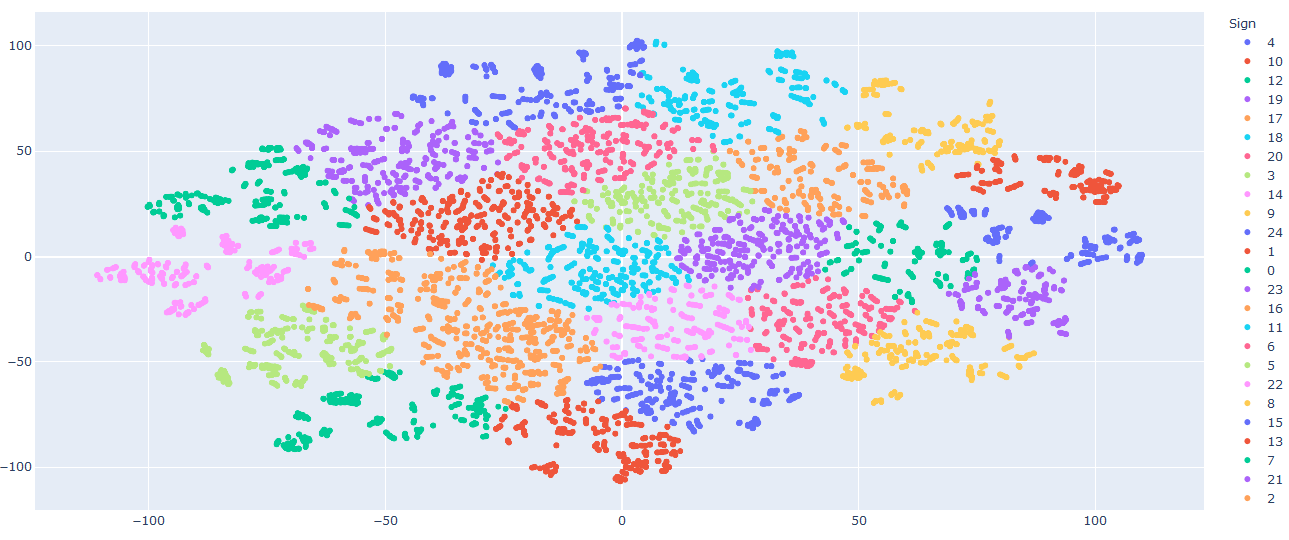
k-means clustering is a strategy for vector quantization, initially from signal handling, that intends to parcel n perceptions into k bunches in which every perception has a place with the group with the closest mean (group focuses or bunch centroid), filling in as a model of the group. This outcomes in a dividing of the information space into Voronoi cells. It is well known for bunch investigation in information mining. k-implies bunching limits inside group changes (squared Euclidean separations), yet not customary Euclidean separations, which would be the more troublesome Weber issue: the mean upgrades squared blunders, while just the geometric middle limits Euclidean separations.

We make 25 clusters,the results are as follows.



### Mini Batch K Means

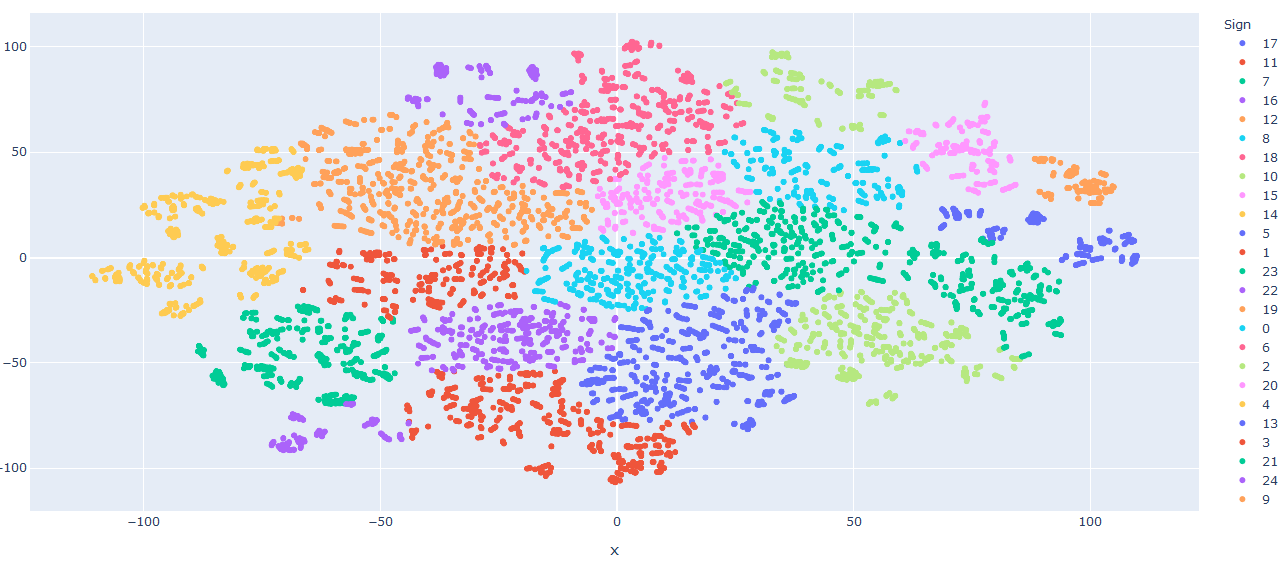
fundamental thought is to utilize little arbitrary batches of information of a fixed size, so they can be put away in memory. Every iteration another irregular example from the dataset is acquired and used to refresh the groups and this is rehashed until union. Every smaller than usual batch refreshes the clusters utilizing a curved blend of the estimations of the models and the information, applying a learning rate that diminishes with the quantity of cycles. This learning rate is the converse of the quantity of information appointed to a bunch during the procedure. As the quantity of cycles builds, the impact of new information is diminished, so combination can be identified when no adjustments in the clusters happen in a few back to back iterations.



### Gaussian mixture

Gaussian Mixture Models use the soft clustering technique for assigning data points to Gaussian distributions. Hence, for a dataset with d features, we would have a mixture of k Gaussian distributions (where k is equivalent to the number of clusters), each having a certain mean vector and variance matrix.

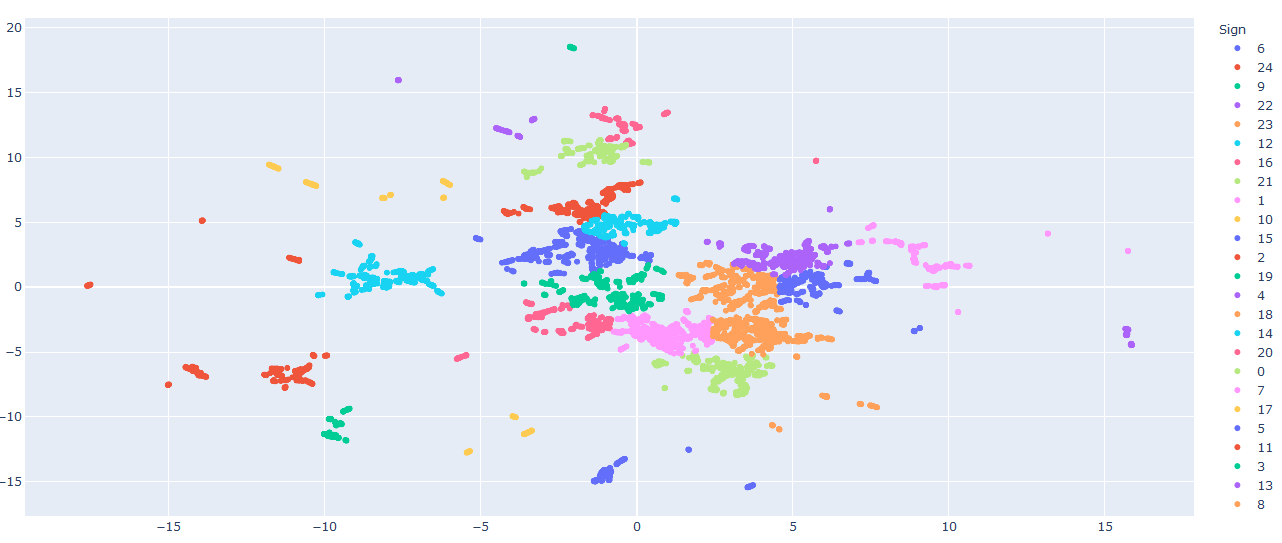
The results of this method for 25 clusters, are.



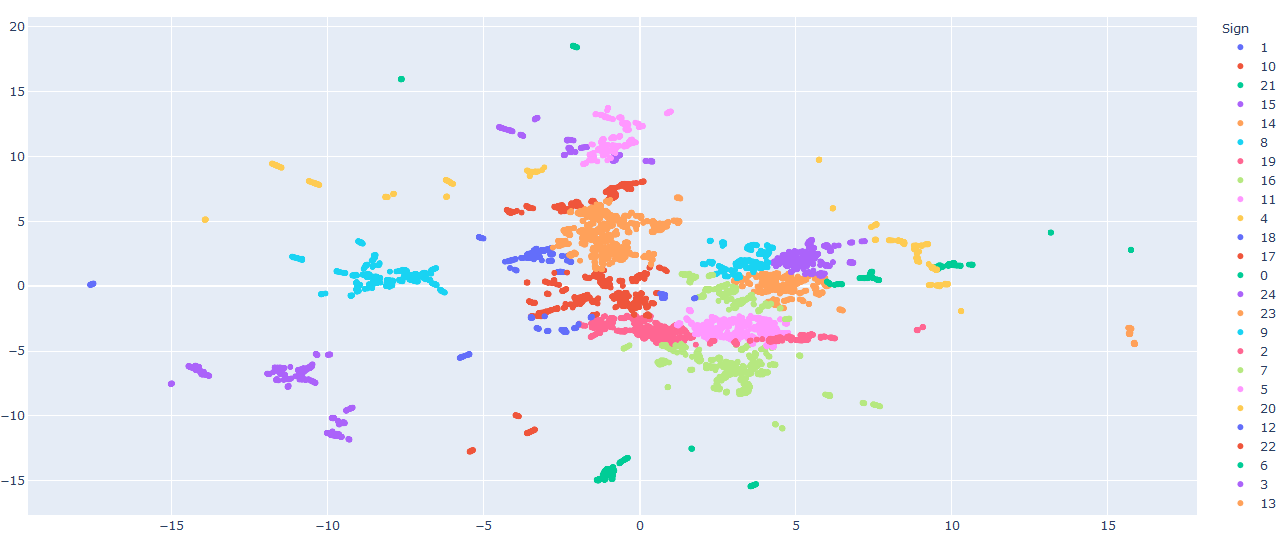
## UMAP approach

In order to compare the effectiveness we also applied the UMAP to the data.

### K-means



### Gaussian Mixture

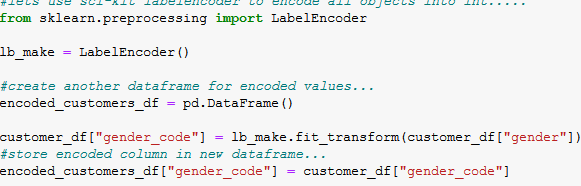


## Conclusion

The results of UMAP are much better for some clusters the distance between many clusters is greater than T-SNE reduction.

# Customers

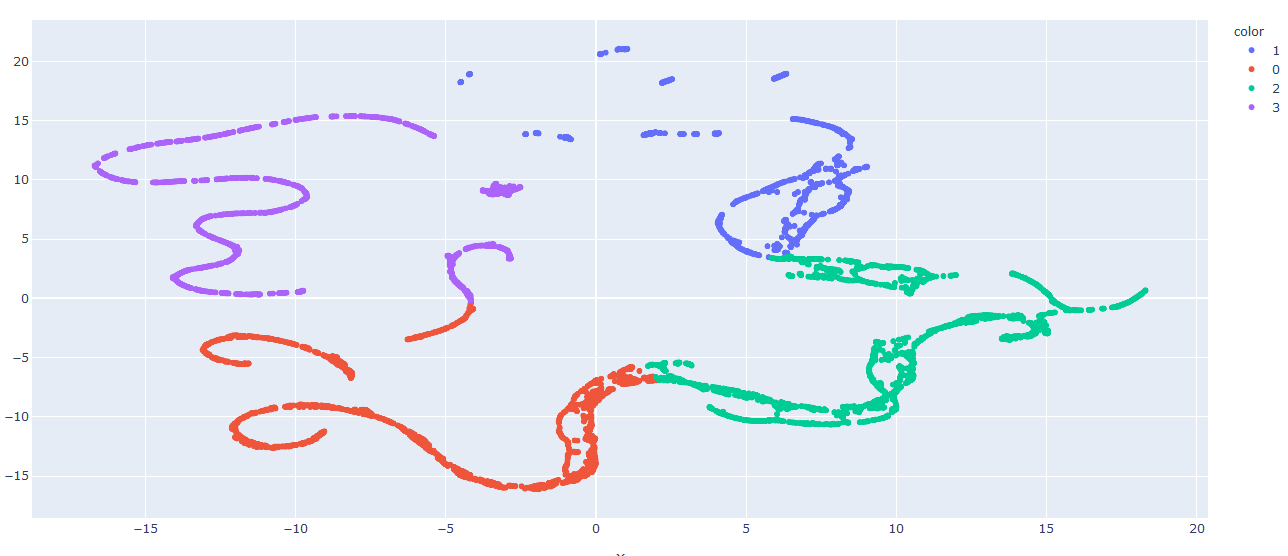
This dataset is unlabeled hence the approach will be different. Also there is a lot of string type data which we encode using sklean preprocessing package.



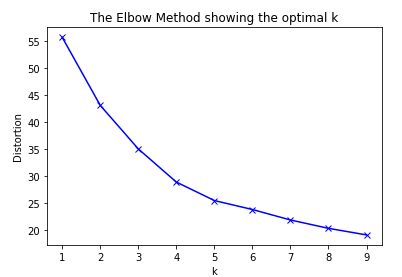
## UMAP Approach

We reduce the dimensions into 2, and then apply clustering algoithms.

### K-means

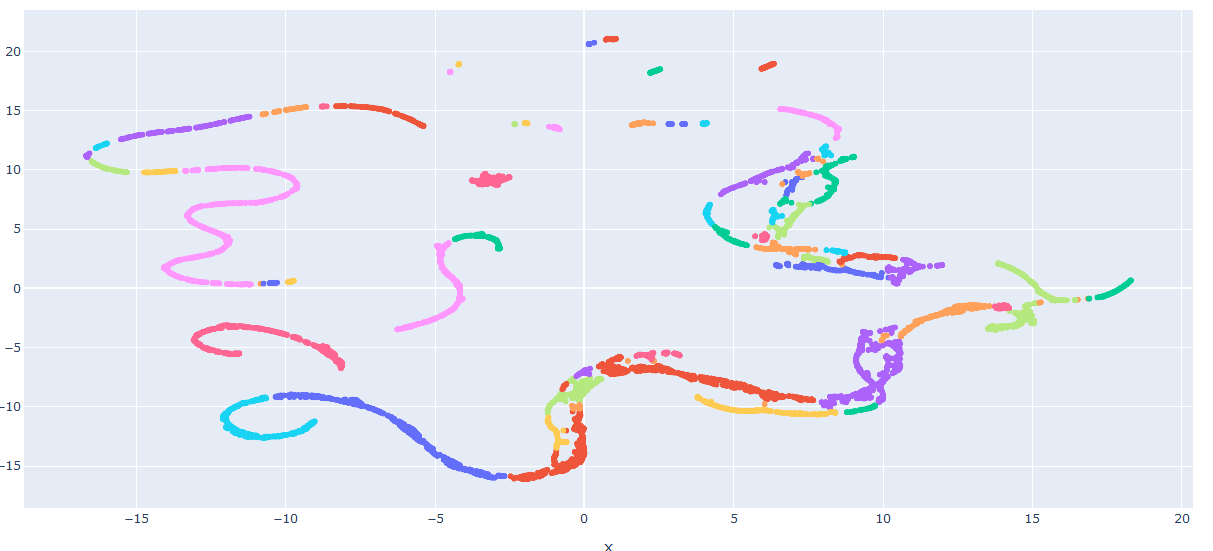


The data is very well distributed, however one problem exists how do we decide the optimal number of cluster for this we use the elbow method.

The elbow is pretty smooth hence we decide to select any number between 3 to 5, in our case we take 4 as optimal number. As shown in the above image [K-Means cluster].

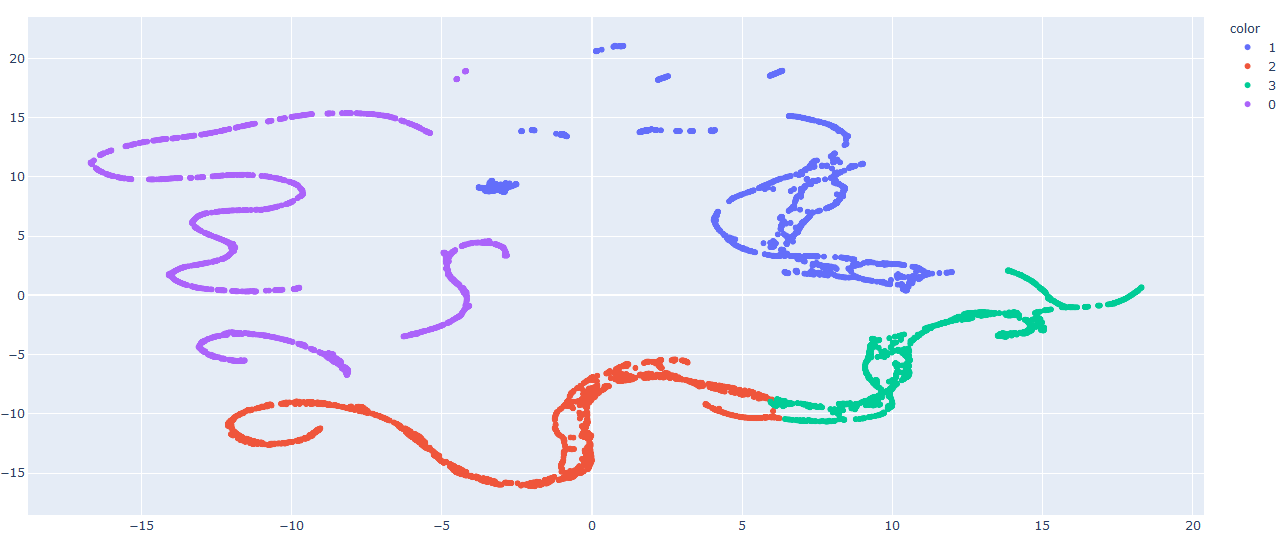
### DB Scan

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.



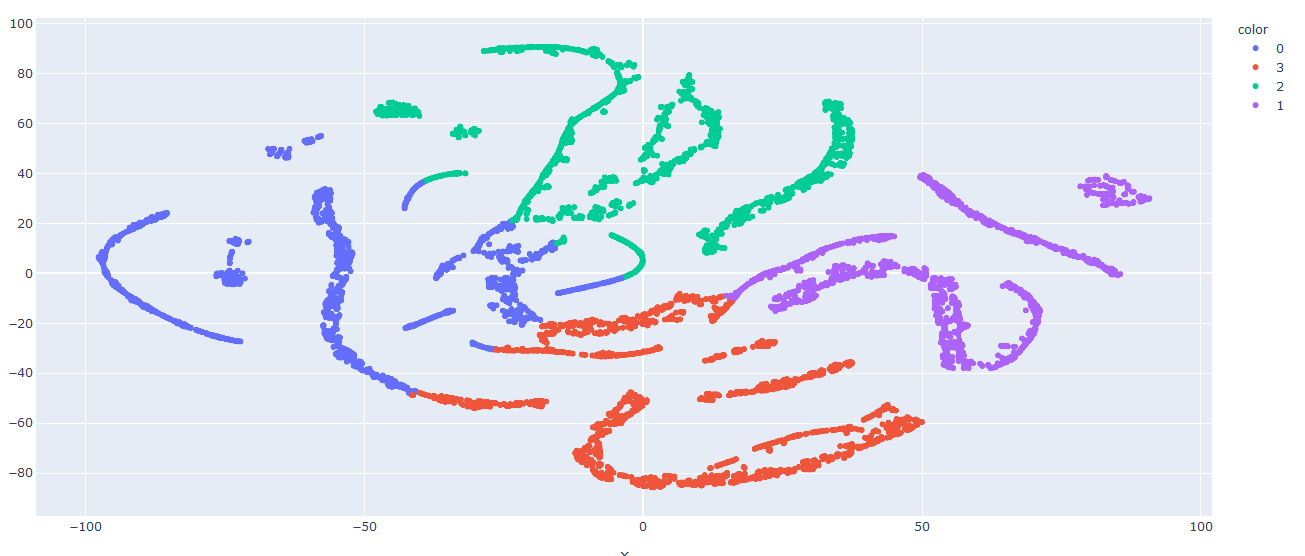
### Agglomorative Clustering

The **agglomerative clustering** is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting). The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects. The result is a tree-based representation of the objects, named dendrogram*.*

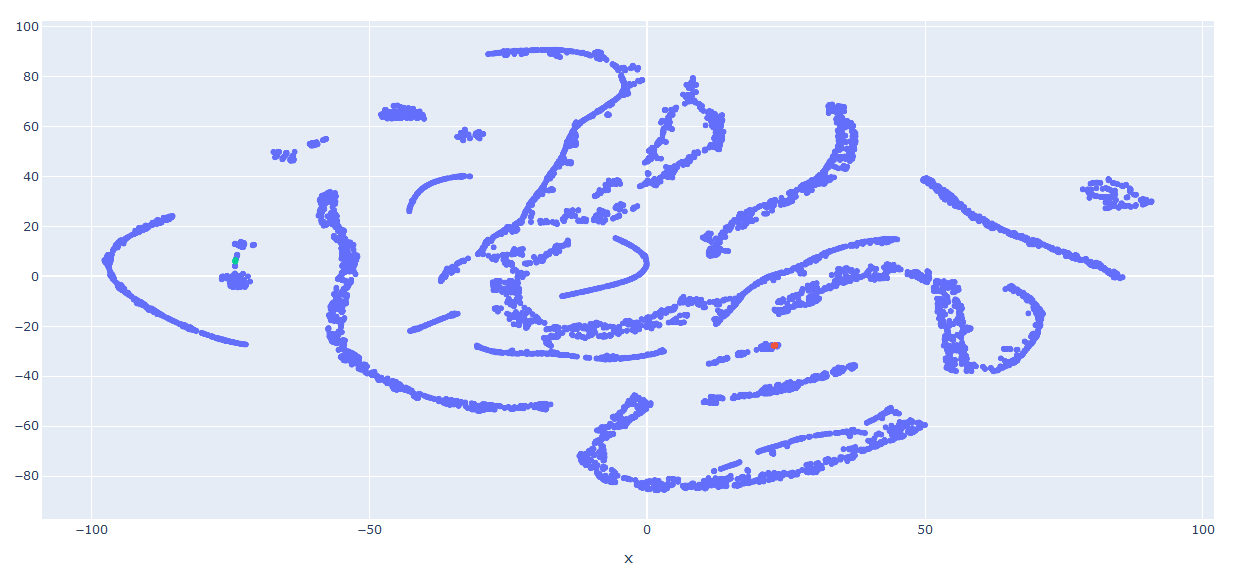


## TSNE Approach

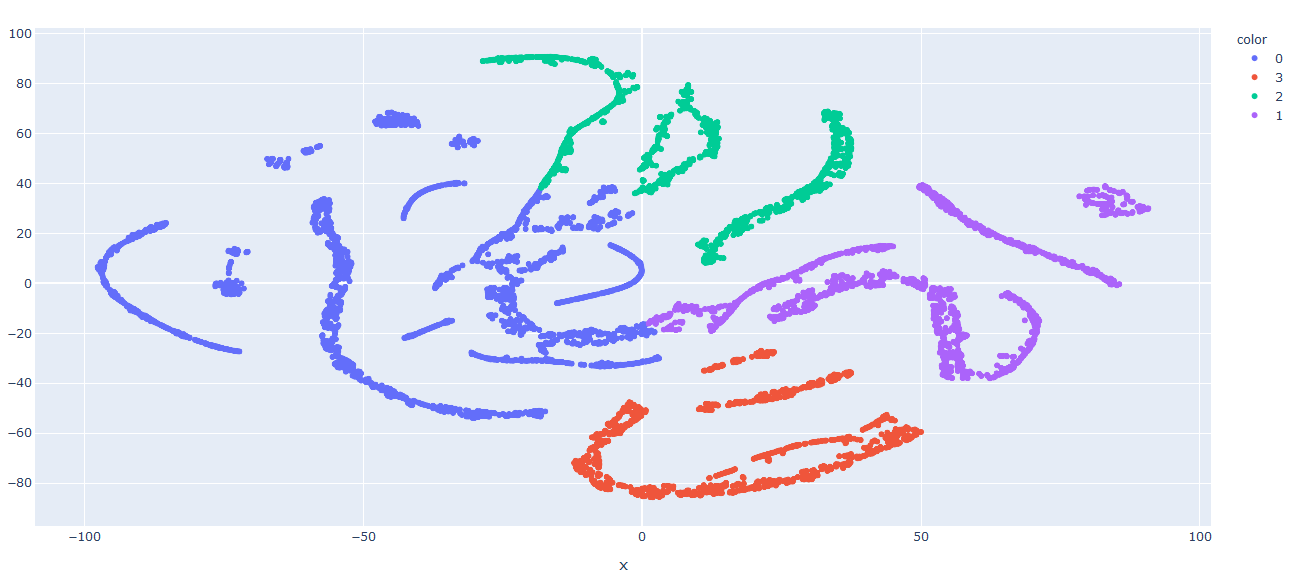
### K-Means



### DBSCAN



### Agglomorative



## Conclusion

The above visuals display that the results of UMAP are better than T-SNE. The cluster distance and boundaries are better observed in UMAP.