**Electronic Assignment Cover sheet**

**Student (s) Number as per your student card:**
10505324, 105323237, 10532445

**Course Title:** Master of Science in Data Analytics

**Lecturer Name:** Kunwar Madan

**Module/Subject Title:** Data Visualization

**Assignment Title:** CA 2

**No of Words:** 1032

# Dimensionality Reduction using UMAP

There were two datasets given to us for dimensionality reduction. First dataset was of images of hand gestures of English letters from A to Z. Second dataset was of telecom company customers.

**Dimensionality Reduction for Dataset 1:**

Dataset 1 that is image data of English letters from A to Z (except letters J and Z) was in the form of 28 X 28 pixels. The dataset had 784 dimension and there was one label column. The main task was to identify the natural clusters and represent entire data in 2 dimensions.

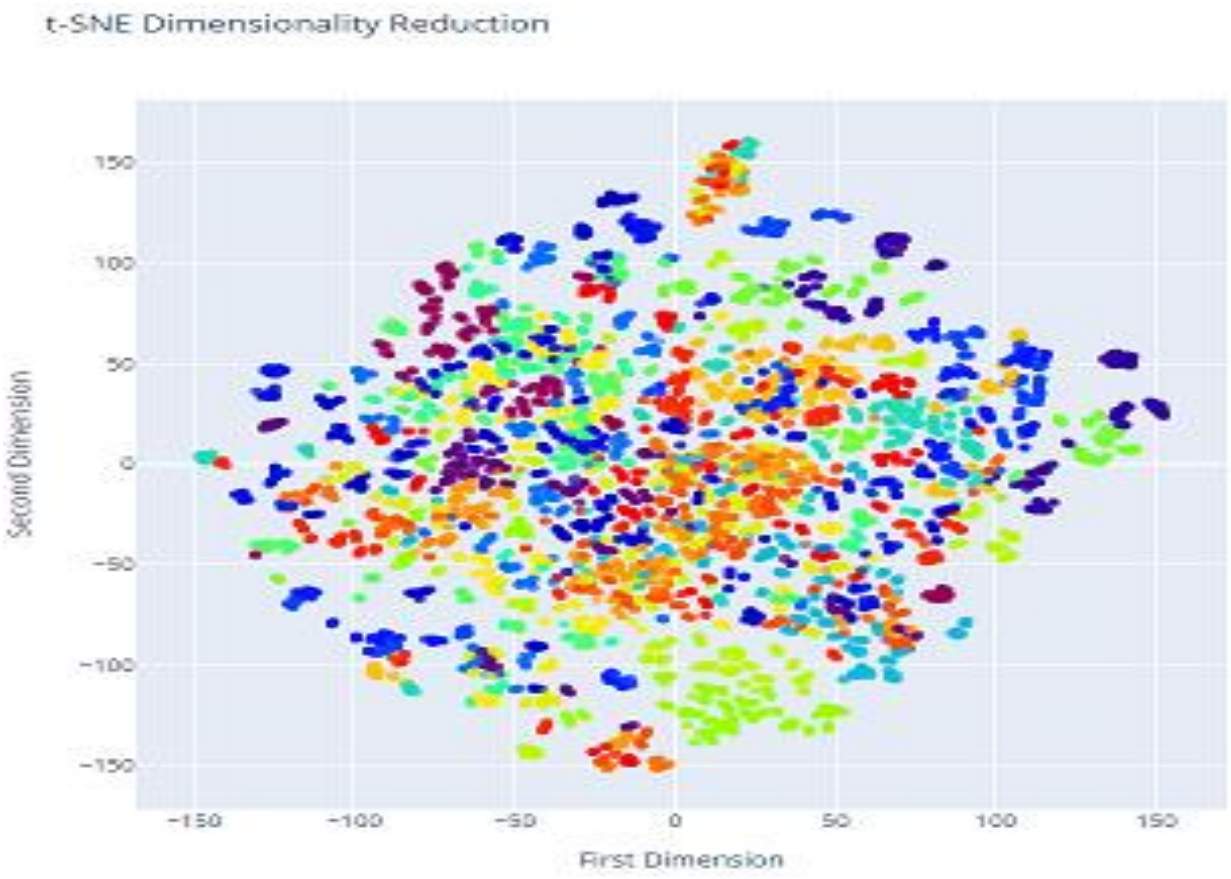
**Data Preparation:**

There were no missing values in the dataset, but the data was needed to be converted into normal form. Before converting data to normal form, the features and labels were split into two different variables. Further, the features were converted into normal for so that each pixel had mean of zero and standard deviation of one. Once the data was prepared it was good to go with next step of dimensionality reduction.

**Dimensionality Reductions:**

There are two important techniques for dimensionality reduction t-SNE and UMAP when there are many dimensions involved in the dataset. t-SNE or t-distributed Stochastic Neighbour Embedding randomly maps the higher dimensionality data point to 2 dimensions and then by

taking the gaussian distribution it adjusts the neighbouring points. t-SNE was introduced in 2008 and performs well better than PCA. But, the only drawback of t-SNE is that it takes more time to process the dimensions are increased. We had used t-SNE for analysis and found the following result:

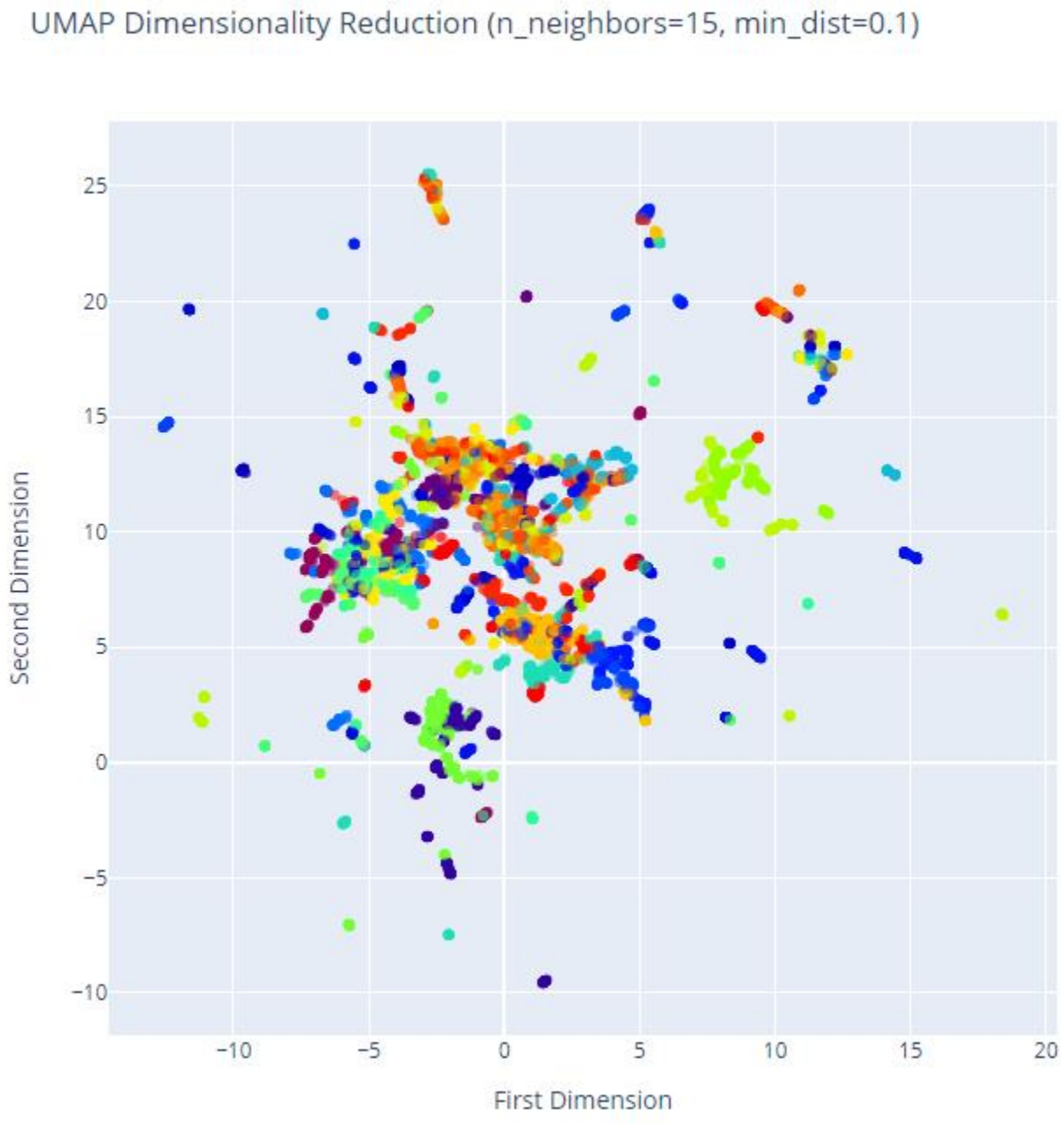



t-SNE Dimensionality Reduction (perplexity = 20, n_iter=2000)

As we can observe that many of the clusters are jumbled within each other and the visualization looks a mess. Also, the processing time took by t-SNE was between 5-10 mins.

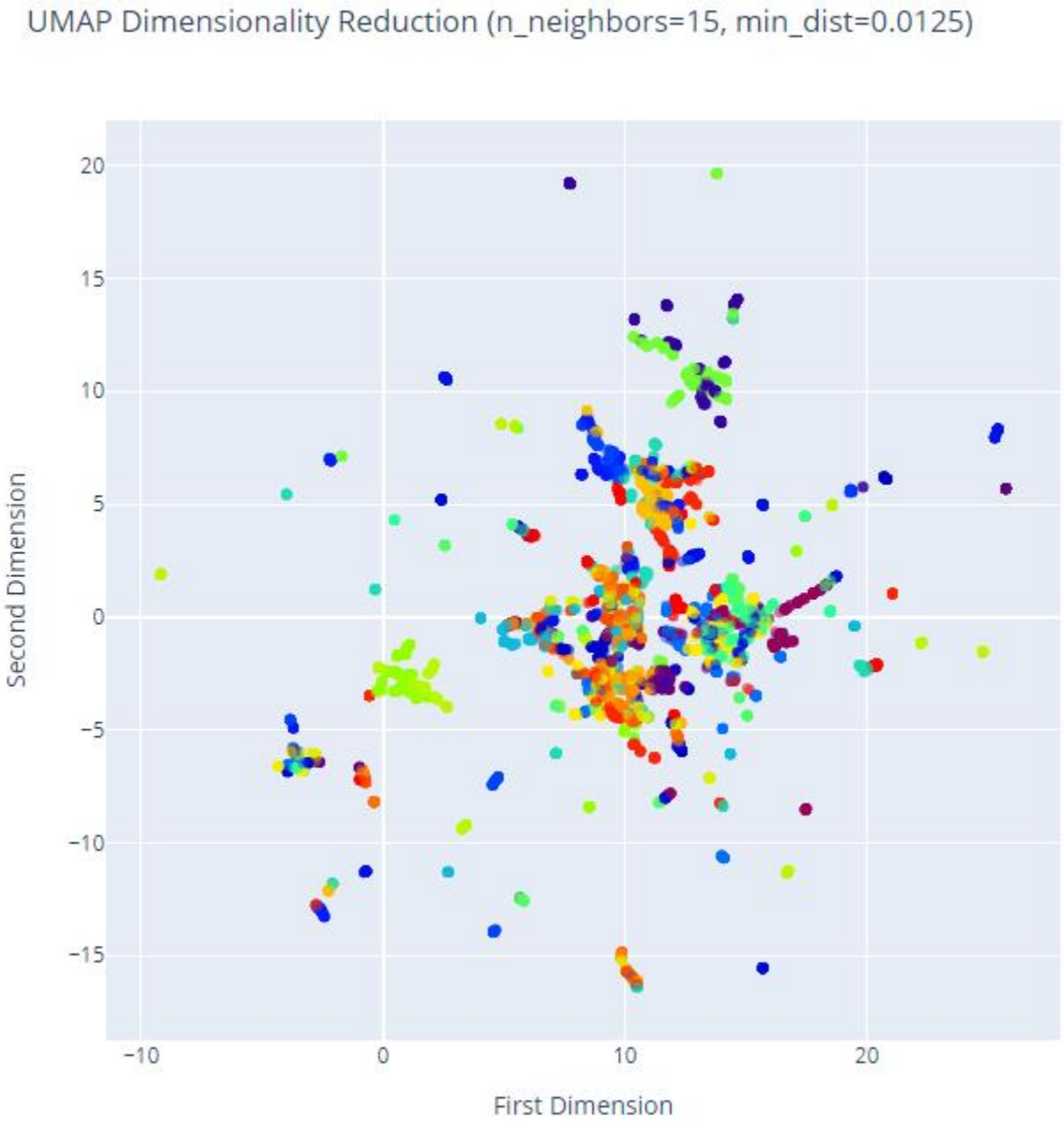
Whereas there is another technique UMAP or Uniform Manifold Approximation Projection which was introduced recently in 2018. UMAP forms much better clusters than t-SNE also the processing time

for UMAP is much less. However, by using the standard initial parameter values that is (n_neighbors=15, min_dist=0.1) clusters were not formed but using UMAP it was easy to fine tune the parameters as it executed easily.




(UMAP1.html)

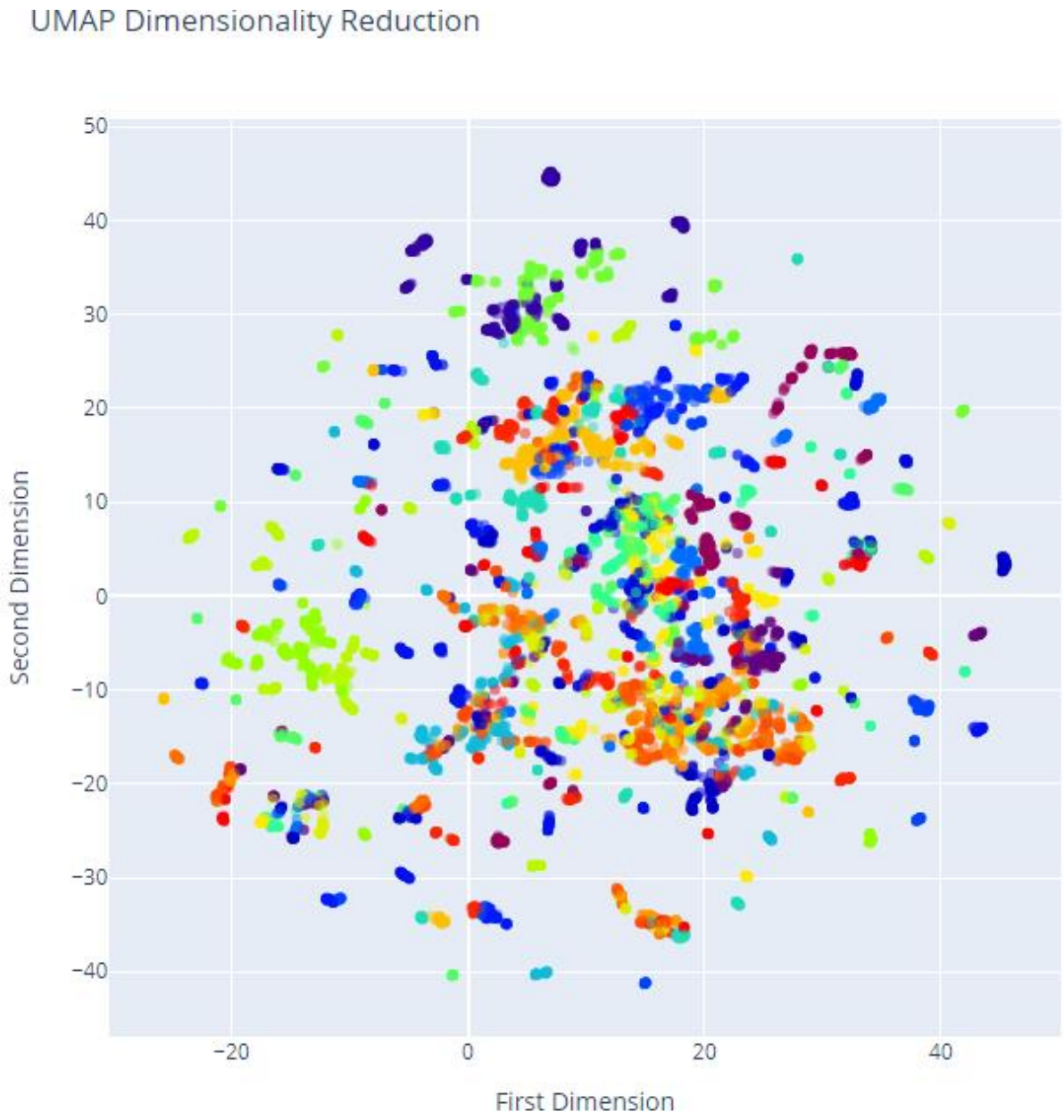
As the data points were already closely packed, we decided to decrease the min_dist to 0.0125. Keeping the n_neighbour same.




UMAP2.html

There was not much difference in the clusters, but clusters were centred at one point, so we decided to increase the ‘spread’ another parameter of UMAP.

Now we can see that the clusters are pretty much separated from each other. Also, there are some interesting clusters formed we can observe from this setting

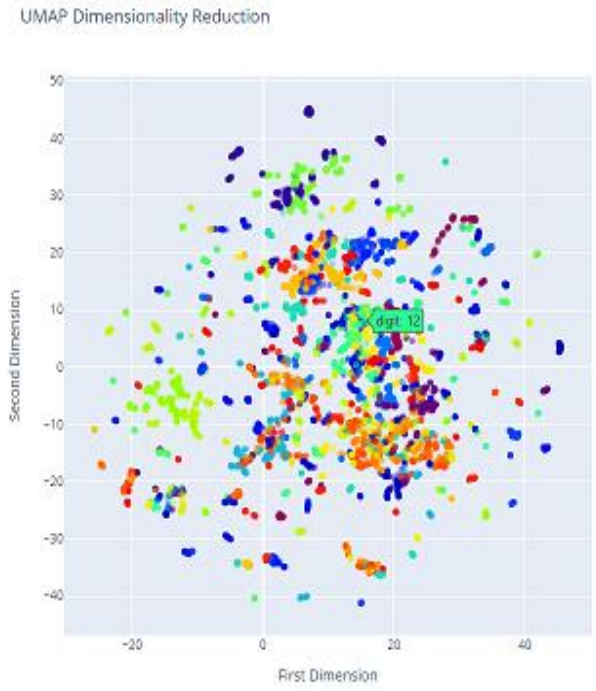



UMAP.html

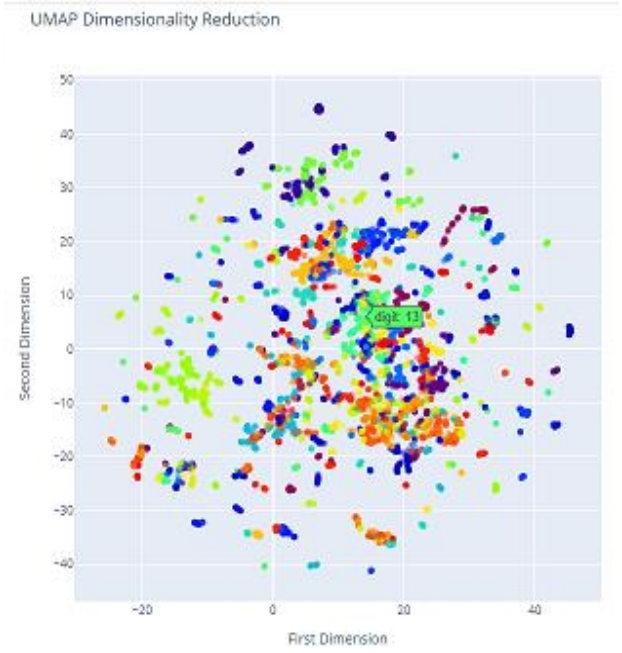
In the below image alphabets like ‘A’, ‘M’, ‘M’, ‘N’, ‘S’, ‘U’, ‘E’ look very much similar to each other.


*Figure 1: Hand gestures in the American Sign Language representing 24 English alphabets*

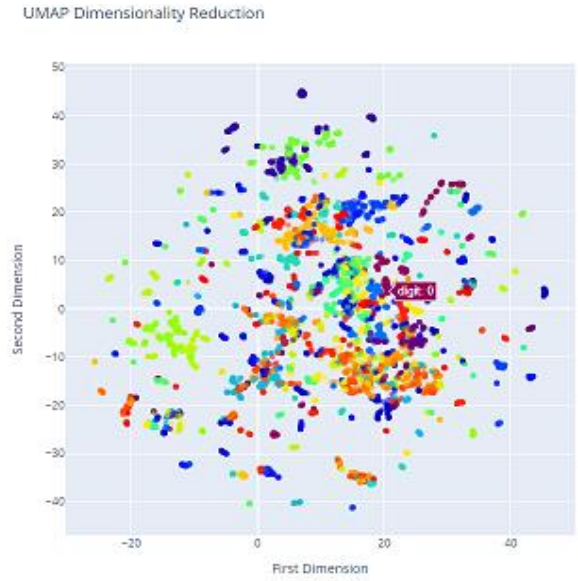
Thus, it is observed that the data point representing these alphabets are clustered together.


UMAP Dimensionality Reduction
Second Dimension
First Dimension
digit: 12

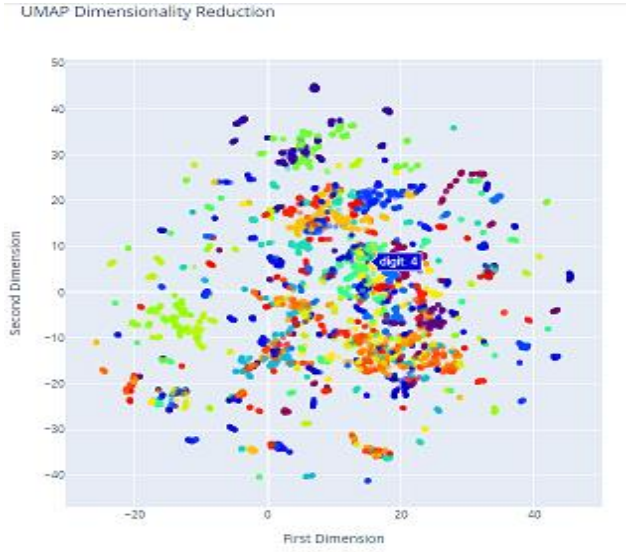
12 = M


UMAP Dimensionality Reduction
Second Dimension
First Dimension
digit: 13

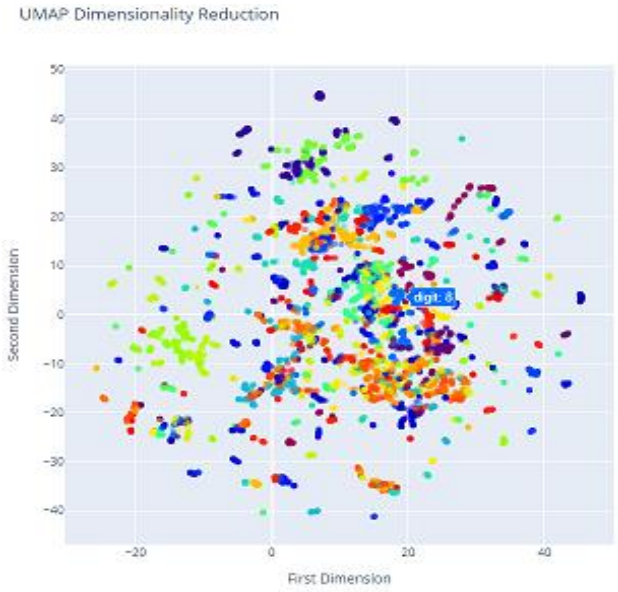
13 = N

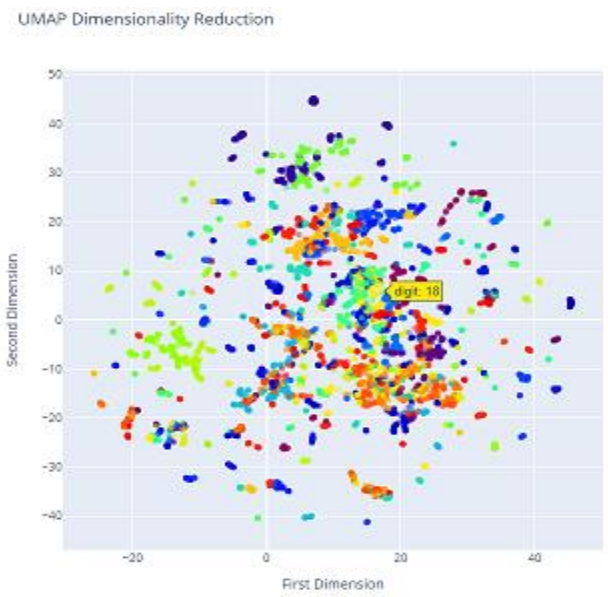

UMAP Dimensionality Reduction
Second Dimension
First Dimension
digit: 0

0 = A


UMAP Dimensionality Reduction
Second Dimension
First Dimension
digit: 4

4 = E



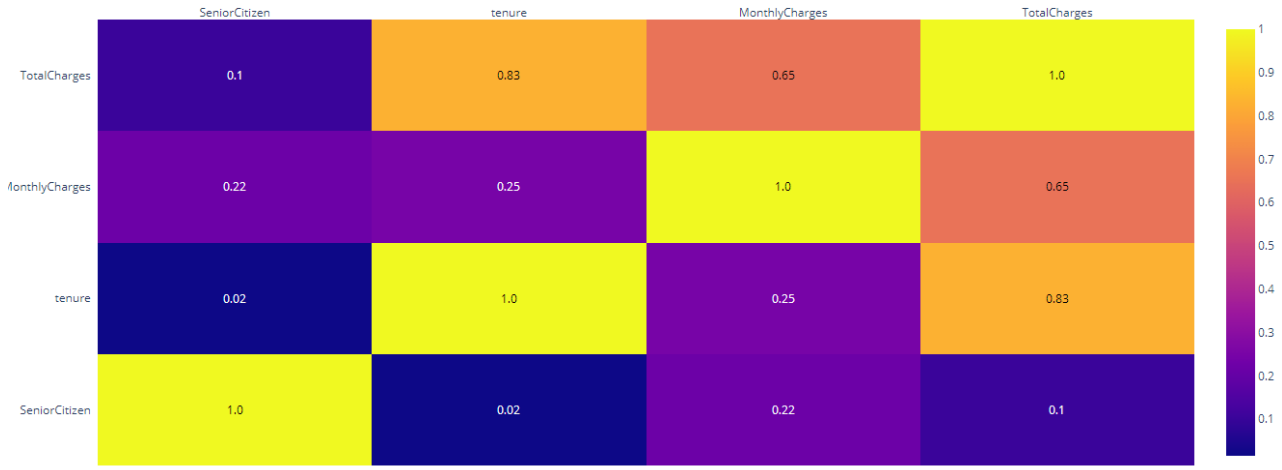




8 = I

18 = S

**Dimensionality Reduction for Dataset 2:**

Second dataset was of customers of telecom company. After importing data and initial analysis it was found that there were too many categorical variables in the dataset which are needed to be converted to numerical before applying dimensionality reduction technique. Before converting the categorical variables to numerical variables, the correlation of existing numerical variables was calculated and then the most highly correlated values with causation were removed.

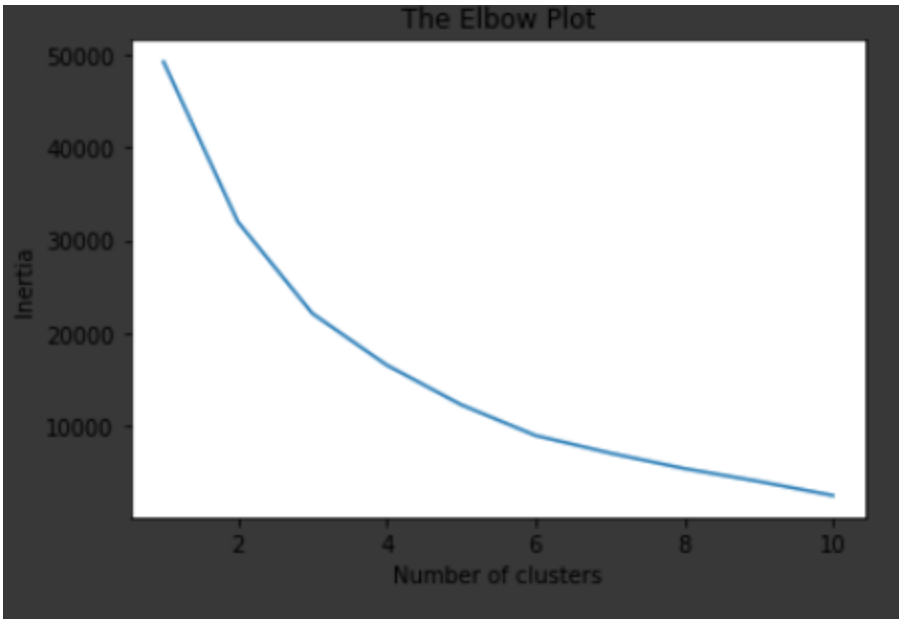



The above correlation heatmap tells us that 'TotalCharges' is highly related to fields, 'MonthlyCharges' and 'tenure' with the increase of 'tenure' and 'MonthlyCharges' can get affected hence there is causation too. Therefore, we decided to remove common column 'TotalCharges'.

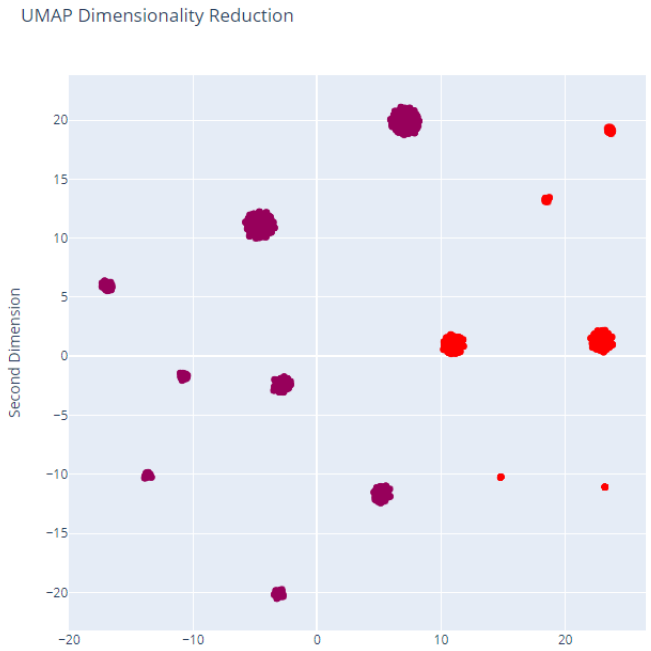
Then the categorical variables were converted to numerical using dummy columns. This led to increase in the dimensions of out dataset to 44 columns. Further, these 44 columns were distributed to subsets with some centre subject. Personal data was taken into one subset, another subset was formed with the information related to services, and final dataset was related to contract information. All the subset then was converted to normalized form so that each column has mean of 0 and standard deviation of one.

**Analysis on Subset 1 (Personal Data):**

As there was no label column it was really difficult to decide on what basis the clusters should be formed. There to identify the number of clusters in subset 1 first, elbow plot was used.




Thus the above elbow plot clearly states that there 2 clusters in the subset. Then, kmeans clustering algorithm was used to identify those 2 clusters. Finally, UMAP was used to represent the entire subset into 2 dimensions and color the clusters with the labels of kmeans. Following was the obtained result.

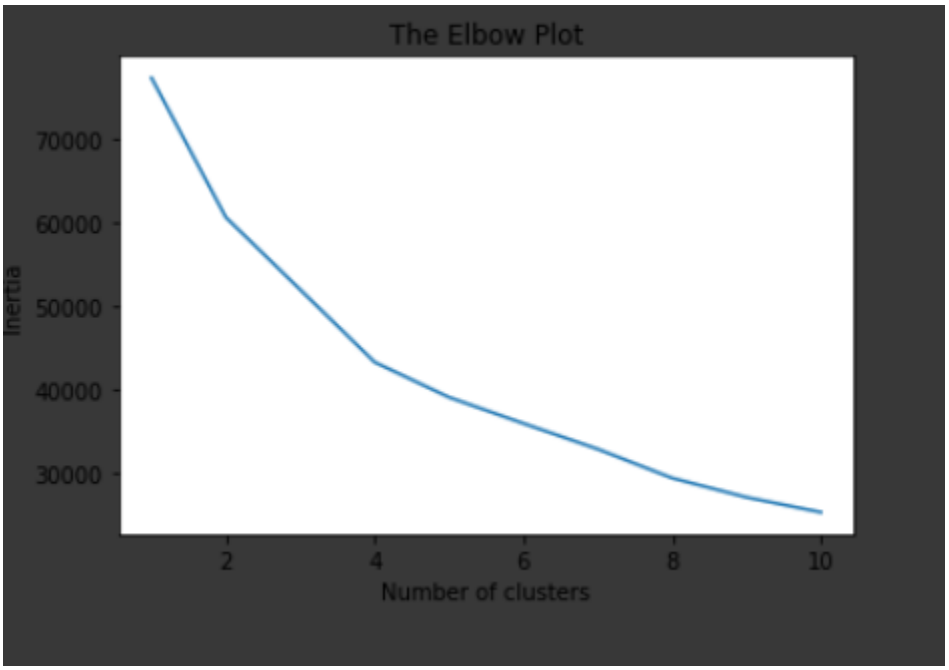



Custumers_UMAP1.html

There were two main clusters found based on Dependent ‘Yes’ or ‘No’ and then further there were subclusters based on other categorical variables.

**Analysis on Subset 2:**

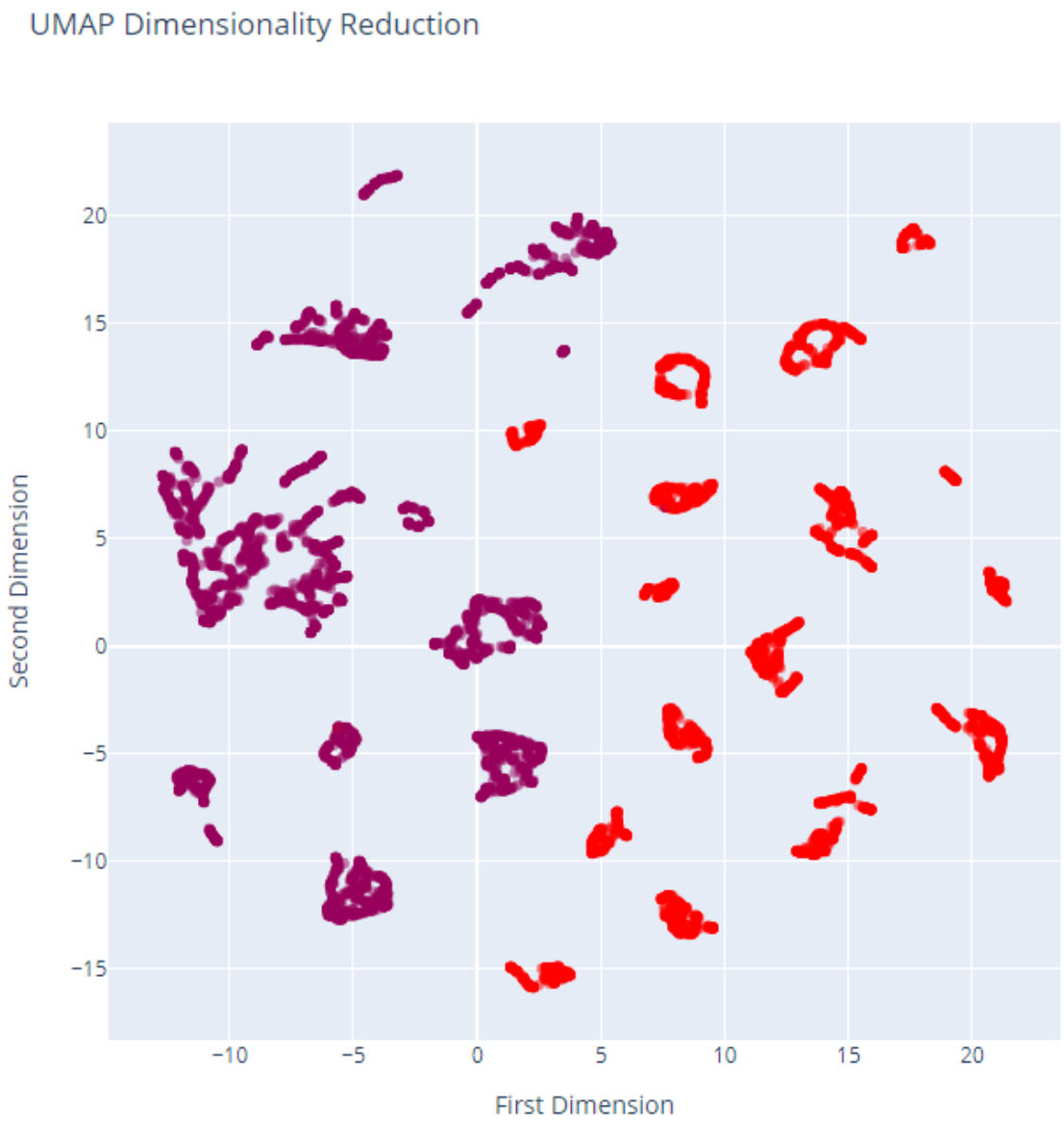
Subset 2 consisted information about the contract which customers has with the company. Like subset 1, for subset 2 to identify the number of clusters elbow plot was constructed.




In this elbow plot too, there are 2 clusters found. Then by running kmeans clustering those clusters were found and labelled in the UMAP dimensionality reduction technique.

From the UMAP visualization it can be observed that there are two clusters formed based on the Contract type that is on the month basis

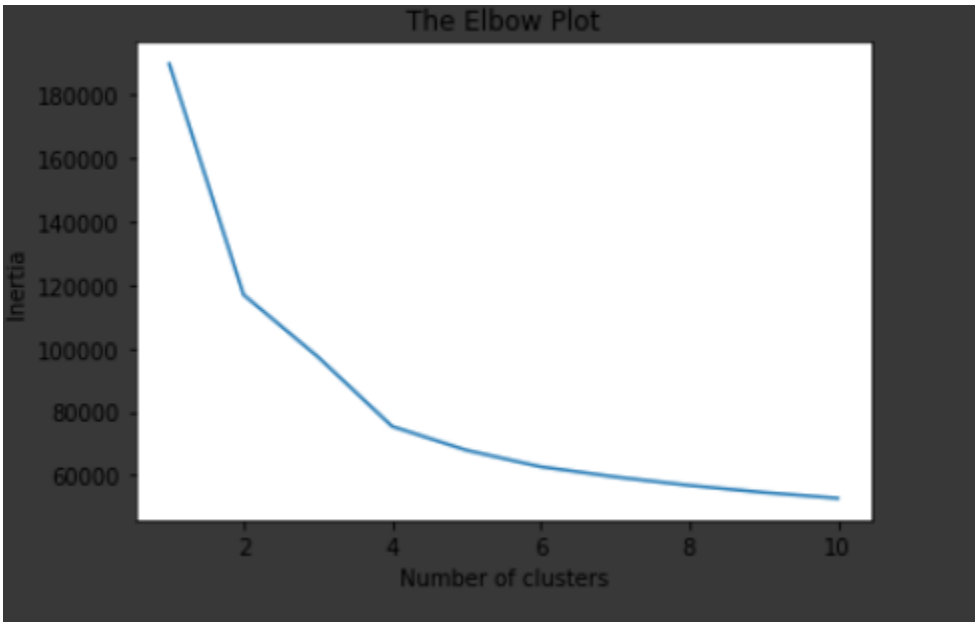
or year basis. Further, more cluster can be found based on the payment methods too. It is observed in the monthly contract-based cluster with the increase in tenure monthly charges also increases. But, in the year-based clusters there is a decrease in monthly charge with increase in tenure.



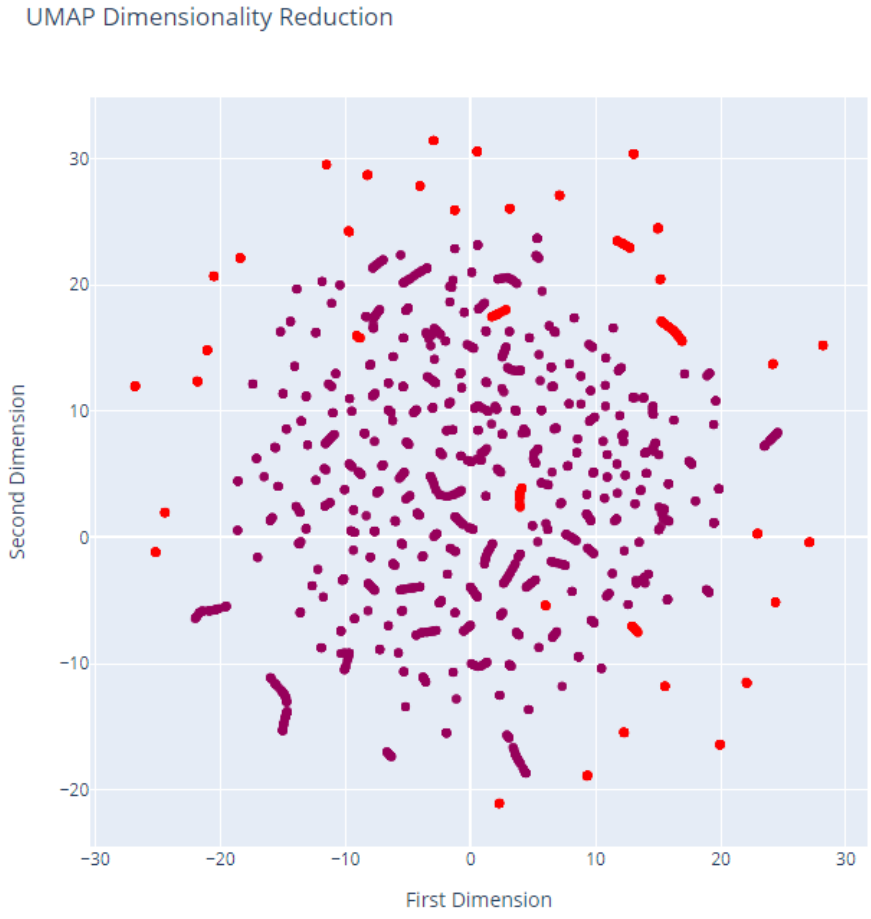

Customers_UMAP2.html

**Analysis on Subset 3:**

Subset 3 is made of all the services used by the customers. After taking elbow plot 2 clusters were found and those clusters were then used to colour the UMAP visualizations.



In this subset there are clusters formed based on internet service used or not, but no further information can be gained from these clusters.




Clusters_UMAP3.html

## Conclusion

In this report we have used UMAP to represent the higher dimensional data in two dimensions. Also, KMeans clustering methods is used to find the clusters in the dataset.

## Individual Contribution:

The approach and the results of the CA was planned and discussed together still there were some parts which where distributed among group members.

Aliasgar Electricwala – I have contributed towards the implementation of Dataset 1.

Adit Deshpande – I have implemented towards the implementation of Dataset 2.

Monika Tambe – I have Contributed towards the documentation of this CA.