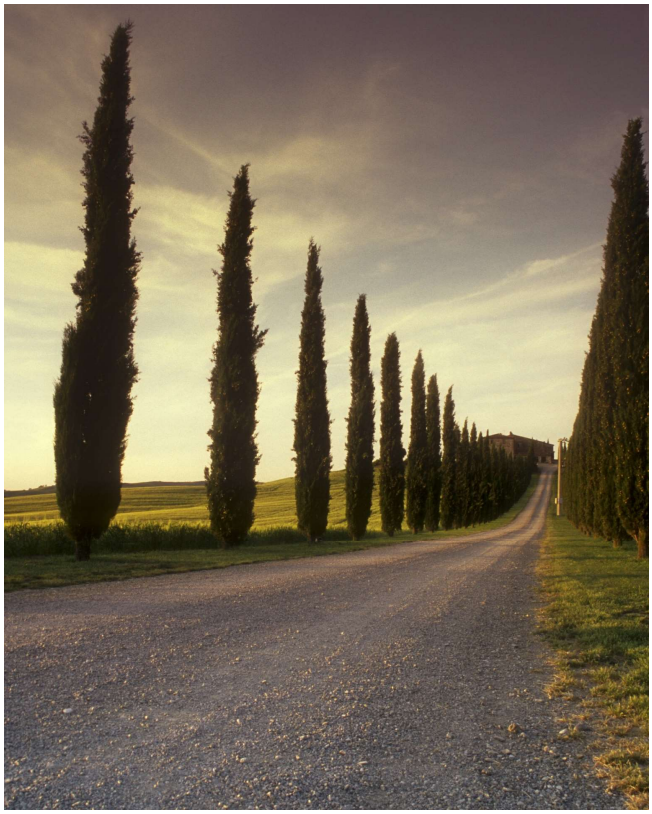


CRITICAL ANALYSIS REPORT

CA-02 DATA VISUALIZATION (B9DA106)

Authors: Bharat Jethwani(10519364), Kinjal Maru(10391312), Prasad Tambe(10515513)

Critical Analysis report

Contents

1. What is Dimensionality Reduction? 1
2. Why is Dimensionality Reduction required? 2
3. Techniques used in this study for dimension reduction: 3
Principal Component Analysis with k-means: 3
t-SNE with k-means: 3
UMAP with k-means: 3
Understanding Clients Data 4
Correlation Between Features: 6
Understanding Patches Data 7
Correlation Features: 9
5. Critical analysis of client dataset 11
Full-Client Dataset 11
6. Critical Analysis for Patches Dataset: 29
7. Conclusion 43
8. Team Contribution 43
9. Individual Contribution 43

1. What is Dimensionality Reduction?

We generate a large amount of data every day. In fact, over the past 3-4 years, 90% of the world's data has been generated. Below are just some of the examples of the kind of data being collected:

- Your smartphone apps collect a lot of personal information about you
- Facebook gathers data on what you like, share, post, places you visit, restaurants you like, etc.
- Casinos keep track of every move each customer makes
- Amazon collects data of what you buy, view, click, etc. on their site

With rising data generation and compilation, it becomes more and more difficult to imagine and draw conclusions; for such cases where we have a large number of variables, it is better to select a subset of the variables which capture as much information as the original set of variables. Reducing p dimensions of the data into a subset of k dimensions (k<<n). This is called dimensionality reduction. (Analytics Vidhya, 2018)

2. Why is Dimensionality Reduction required?

Below are the benefits of dimension reduction: (Analytics Vidhya, 2018)

• Dimensions are reduced by the volume of space used to store the data

• If we have large dimensions, certain algorithms don't work well. Therefore, the algorithm needs to be useful to reduce these dimensions

• Fewer dimensions contribute to lower time for computation/training

• It takes care of multicollinearity by removing redundant features. For example, you have two variables – 'time spent on a treadmill in minutes' and 'calories burnt.' These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require

• It helps in visualizing data. As discussed earlier, it is challenging to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly.

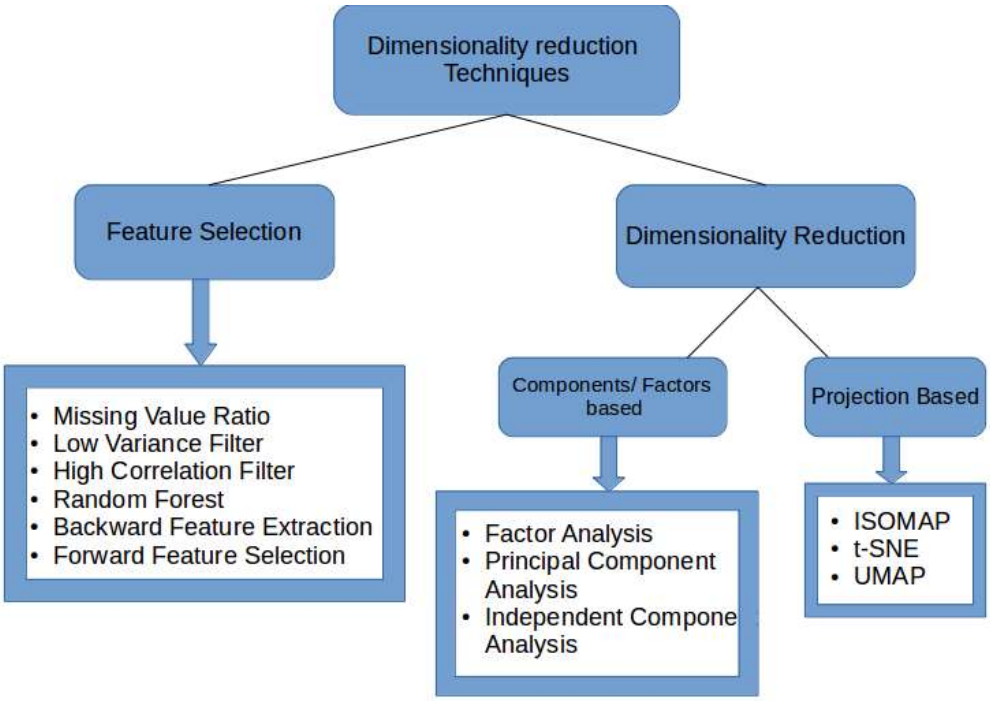
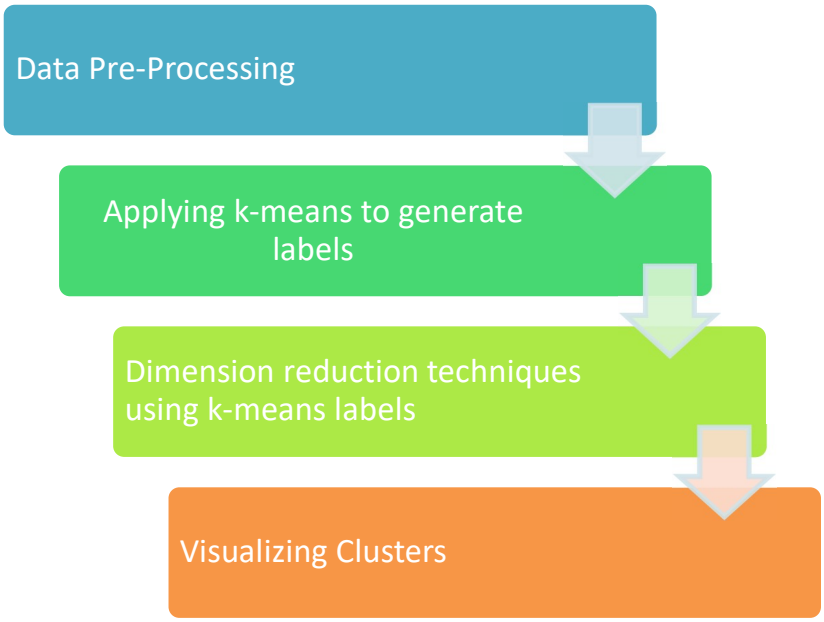


Fig 1. Common Dimension reduction techniques (Analytics Vidhya, 2018)

3. Techniques used in this study for dimension reduction:

Principal Component Analysis with k-means:



What is PCA?

This is one of the most widely used techniques for dealing with linear data. It divides the data into a set of components which try to explain as much variance as possible. (Analytics Vidhya, 2018)

We have reduced the dimensions using PCA and labelled it using k-means. K-means component were determined using elbow plot.

t-SNE with k-means:

What is t-SNE?

This technique also works well when the data is strongly non-linear. It works extremely well for visualizations as well (Analytics Vidhya, 2018)

We have reduced the dimensions using t-SNE and labelled it using k-means. K-means component were determined using elbow plot.

UMAP with k-means:

What is UMAP?

This technique works well for high dimensional data. Its run-time is shorter as compared to t-SNE (Analytics Vidhya, 2018)

We have reduced the dimensions using t- UMAP and labeled it using k-means. K-means component were determined using elbow plot.

Understanding Clients Data

The dataset contains client information for an unnamed bank, where each row corresponds to a new client. The dataset comprises 43,193 rows and nine columns. The list of columns being:

1. Age: This Feature is numeric in nature, with each row representing the client's age.

```
count    43193.000000
mean        40.764082
std         10.512640
min         18.000000
25%         33.000000
50%         39.000000
75%         48.000000
max         95.000000
Name: age, dtype: float64
```

Fig.2 Age Statistics

2. Job: This Feature describes the type of employment for a particular client, data being categorical in nature a specific client can only work as one of the listed jobs:
    - "admin."
    - "unemployed"
    - "management"
    - "housemaid"
    - "entrepreneur"
    - "student"
    - "blue-collar"
    - "self-employed"
    - "retired"
    - "technician"
    - "services"

```
count           43193
unique             11
top        blue-collar
freq             9278
Name: job, dtype: object
```

Fig 3. Job Statistics

Fig2 shows the statistics for Job, with “blue-collar” being the top value.

3. marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

This Feature is a definite type where the mar is classified into married, divorced, and none being divorced or widowed, which represents the specific client's marital status.

```
count       43193
unique          3
top       married
freq        25946
Name: marital, dtype: object
```

Fig 4 shows the statistics for marital, with "married" being the top value.

4. education (categorical: "secondary", "primary", "tertiary")
   This Feature is also a categorical type where it is divided into secondary, primary, tertiary defining the specific area of education.

```
count         43193
unique            3
top       secondary
freq          23131
Name: education, dtype: object
```

Fig 5 shows the statistics for education, with "secondary" being the top value.

5. default: has credit in default? (binary: "yes", "no")
   This Feature defines if the credit is applied to a particular client in default with the binary representation as "Yes" or "No".

```
count     43193
unique        2
top          no
freq      42411
Name: default, dtype: object
```

Fig 6 shows the statistics for default, with "no" being the top value.

6. balance: average yearly balance, in euros (numeric)
   This Feature represents the balance of a specific client, which is average yearly balance in euro's, which is in numeric type.

```
count     43193.000000
mean       1354.027342
std        3042.103625
min       -8019.000000
25%          71.000000
50%         442.000000
75%        1412.000000
max      102127.000000
Name: balance, dtype: float64
```

Fig 7 Shows the balance statistics.

7. housing: has a housing loan? (binary: "yes", "no")
   This Feature denotes whether the client is applicable for a housing loan or not, which is a binary type.

```
count     43193
unique        2
top         yes
freq      24292
Name: housing, dtype: object
```

Fig 8 shows the statistics for housing, with “yes” being the top value.

8. personal: has personal loan? (binary: "yes", "no")
   This Feature represents if the client has any personal loan, which is in binary type

```
count     43193
unique        2
top          no
freq      36086
Name: personal, dtype: object
```

Fig 9 shows the statistics for personal, with “no” being the top value.

9. term: has term deposit? (binary: "yes", "no")
   This Feature shows if the client has any term deposit, which is denoted in binary type.

```
count     43193
unique        2
top          no
freq      38172
Name: term, dtype: object
```

Fig 10 shows the statistics for the term, with ”no” being the top.

Correlation Between Features:

Some of the most rapid ways to improve the model are to recognize and reduce the strongly correlated data set features. Correlated features add to model noise and imprecision that makes it more challenging to achieve the desired result. However, in our dataset, we could not find any correlation between the two features, with the max being -0.4 between age and marital. No value is more significant than 0.5 or less than -0.5.

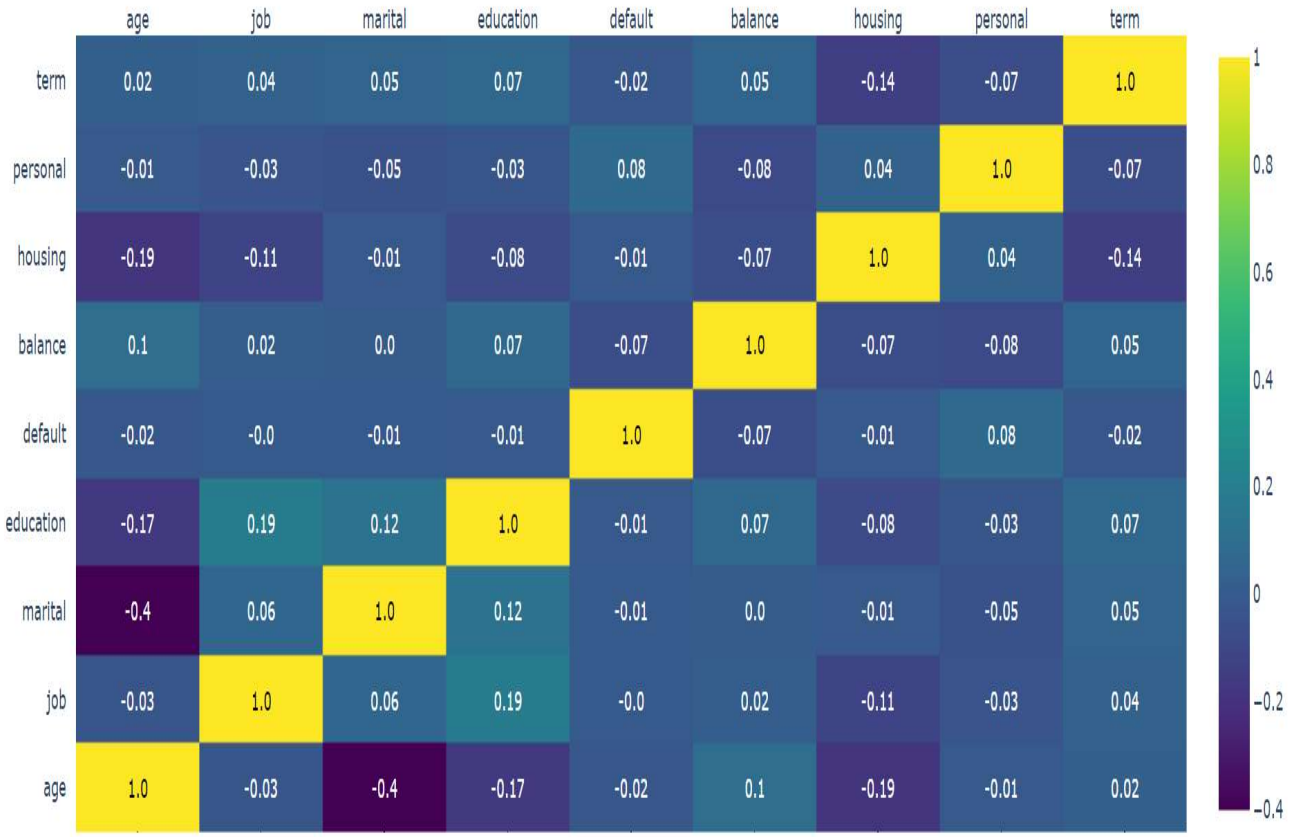


Fig 10. Correlation for Client Dataset

Understanding Patches Data

The dataset “Patches” contains cartographic data for observations made in the Alberta, Canada forests o ver various 30 m *30 m patches. The Dataset contains 15120 observations and 7 features.
The features included are: Elevation, Slope, Horizontal_Distance_To_Hydrology, Vertical_Distance_To_H ydrology,Horizontal_Distance_To_Roadways,Horizontal_Distance_To_Fire_Points,Tree

1. Elevation, Slope: This Feature is numeric in nature which shows the Elevation, slope of Alberta and Canada's forest.

```
count    15120.000000
mean      2749.322553
std        417.678187
min       1863.000000
25%       2376.000000
50%       2752.000000
75%       3104.000000
max       3849.000000
Name: Elevation, dtype: float64
```

Fig 11 shows the statistics of Elevation

Horizontal_Distance_To_Hydrology: This Feature denotes the distance between water level and the patc hes which are in horizontal in nature.

```
count    15120.000000
mean       227.195701
std        210.075296
min          0.000000
25%         67.000000
50%        180.000000
75%        330.000000
max       1343.000000
Name: Horizontal_Distance_To_Hydrology, dtype: float64
```

Fig 12 shows the statistics for Horizontal_Distance_To_Hydrology.

Vertical_Distance_To_Hydrology: This feature indicates the vertical distance between the water level an d patches.

```
count    15120.000000
mean        51.076521
std         61.239406
min       -146.000000
25%          5.000000
50%         32.000000
75%         79.000000
max        554.000000
Name: Vertical_Distance_To_Hydrology, dtype: float64
```

Fig 13 shows the statistics for Vertical_Distance_To_Hydrology.

Horizontal_Distance_To_Roadways: This Feature represents the Horizontal Distance between the patche s and the roadways which is numeric in nature.

```
count    15120.000000
mean      1714.023214
std       1325.066358
min          0.000000
25%        764.000000
50%       1316.000000
75%       2270.000000
max       6890.000000
Name: Horizontal_Distance_To_Roadways, dtype: float64
```

Fig 14 shows the statistics for Horizontal_Distance_To_Roadways.

Horizontal_Distance_To_Fire_Points: This Feature denotes that Horizontal_Distance between the patche s and fire points which is in Numeric.

```
count    15120.000000
mean      1511.147288
std       1099.936493
min          0.000000
25%        730.000000
50%       1256.000000
75%       1988.250000
max       6993.000000
Name: Horizontal_Distance_To_Fire_Points, dtype: float64
```

Fig 15 shows the statistics for Horizontal_Distance_To_Fire_Points.

Tree: This is a categorical feature “Tree” whose distinct values are : “Spruce” which means Spruce tree w as found predominant or “Other” indicates the trees other than Spruce were found predominant.

```
count     15120
unique        2
top       Other
freq      12960
Name: Tree, dtype: object
```

Fig 16 shows the statistics for Tree, with “other” being the top value.

Correlation Features:

In order to visualize the correlations between all features, we are going to analyse a Correlations Heat M ap (Figure 1).

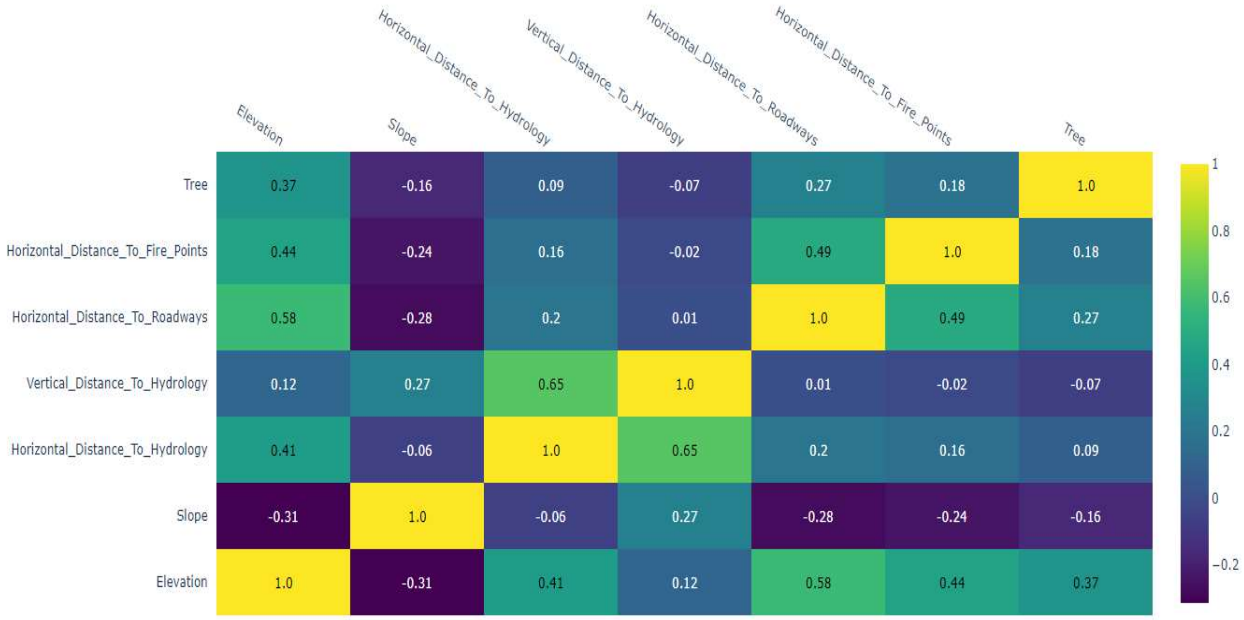


Figure 17(Correlation for Patches)

From figure 17 ,We can see positive relation of 0.58 between Elevation and Horizontal_Distance_To_Roadways .In general ,roadways are constructed at a particular distance from the forest patches ,That means It is safer to construct roadways from large elevated forest patches.

We can also see 0.65 correlation between Horizontal_Distance_To_Hydrology and Vertical_Distance_To_Hydrology ,Both are linear because larger the patch wider the water surface.

Note: We can remove one of the two correlated features while processing the subsets.

5. Critical analysis of client dataset

Full-Client Dataset

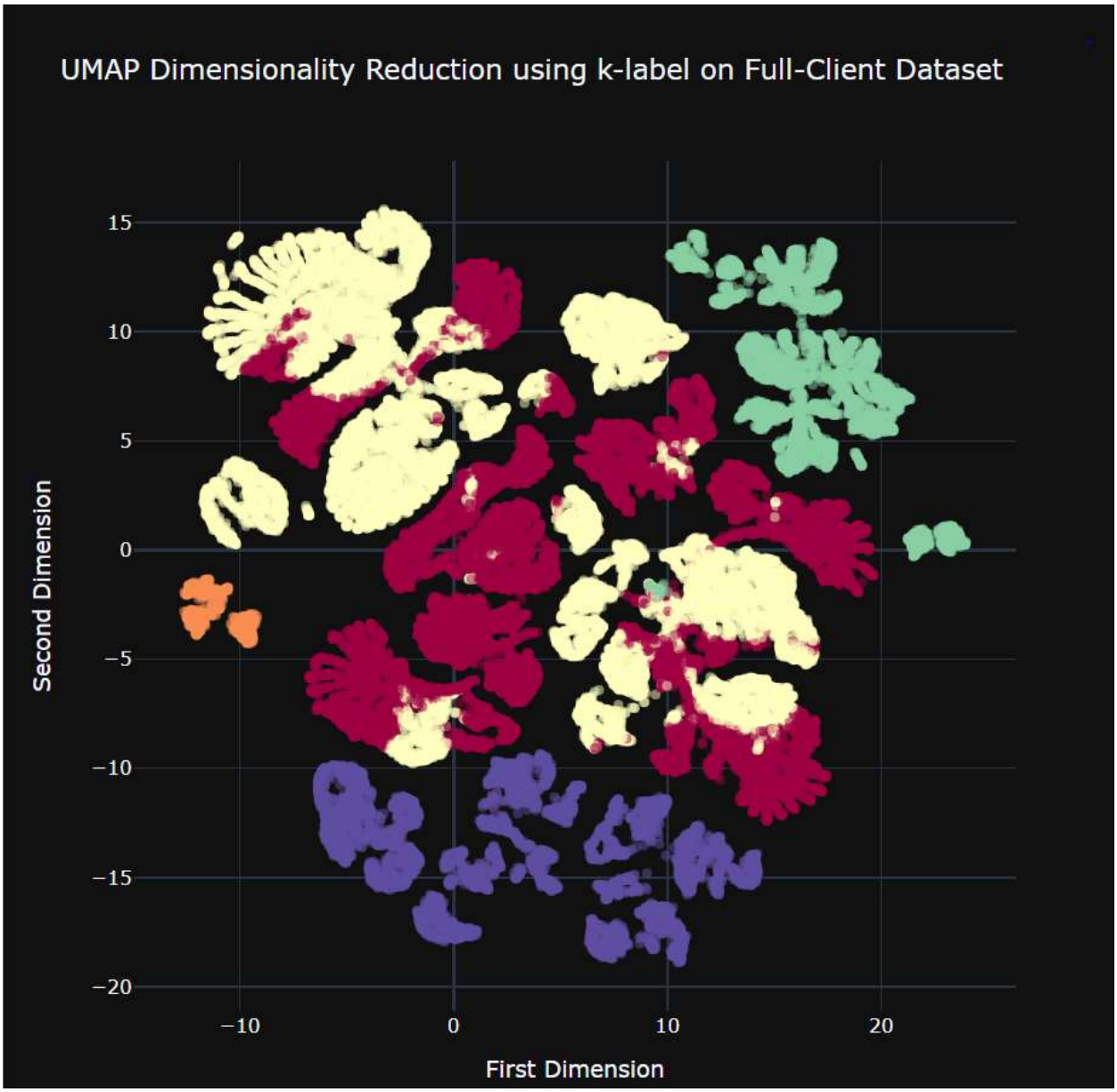


Fig 18 UMAP Result on Client Dataset

We tried to analyse how are clusters grouped and below are for the same:

1. **Cluster 1(Green):**
    - Situated on the upper side of first dimension
    - Person's age is ranged between 24 to 73
    - Job criteria is mostly admin and blue-collar
    - Person has mostly not taken personal loan

2. **Cluster 2(Red):**
    - Age is between range 20 to 50
    - Marital status is married and single with dominant married
    - Mostly all the jobs except housemaid or retired or unemployed
3. **Cluster 3(Purple):**
    - Marital status is mostly married
    - Person has taken personal loan
    - Dominant on the lower side of second dimension below -10
4. **Cluster 4(Yellow):**
    - Person has not taken personal Loan
    - All the job except unemployed ,technician and student
    - Mixed marital status with dominant Married
5. **Cluster 5(Orange):**
    - Smallest cluster on the dataset
    - Average age is between 25 to 60
    - Mixed job status

Insights:

Full Datastet:

• Usually clients with marital status as "married" have housing loans but no personal loans.

• Clients in the age of 30 and job status 7 usually have balance less than 400.

• Clients above 40 having marital status "married" tend to have more balance usually greater than 4000.

• Clients having age above 35 have descent balance with marital status being "Yes" and education status as "Secondary".

• Clients having highly side of balance having housing "no" and personal "no" i.e. Having no loan.

• Clients having education status as "primary" usually have either housing status "yes" or personal status "yes" or both.

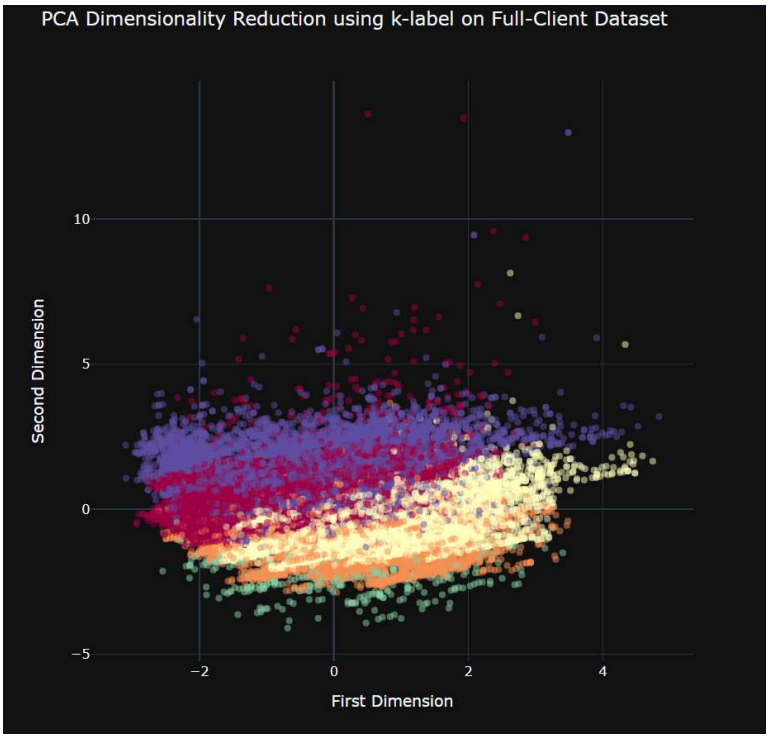


Fig 19. PCA Result on Full Client Dataset

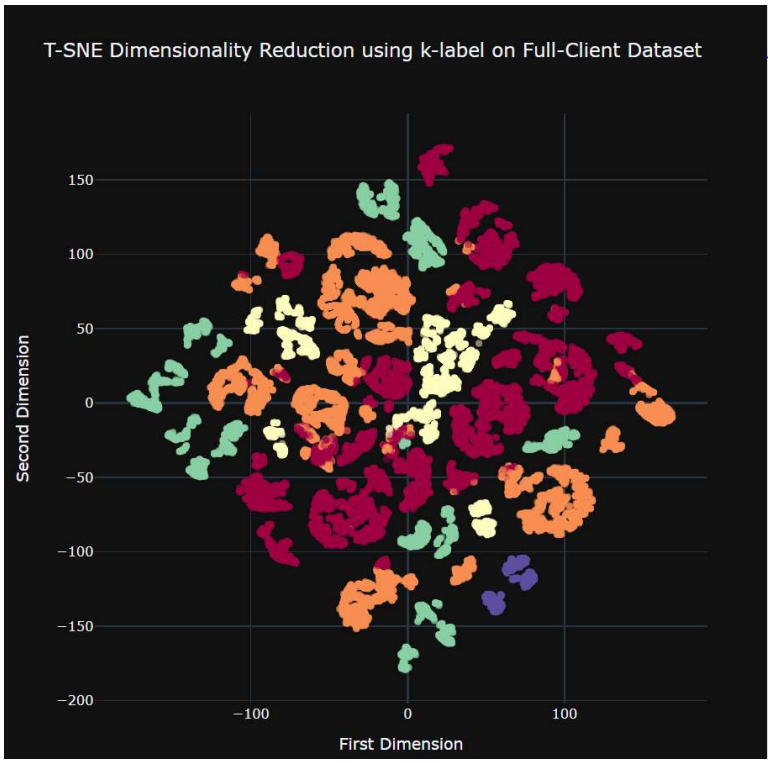


Fig 20 T-SNE Result on Full Dataset

Subset 1:

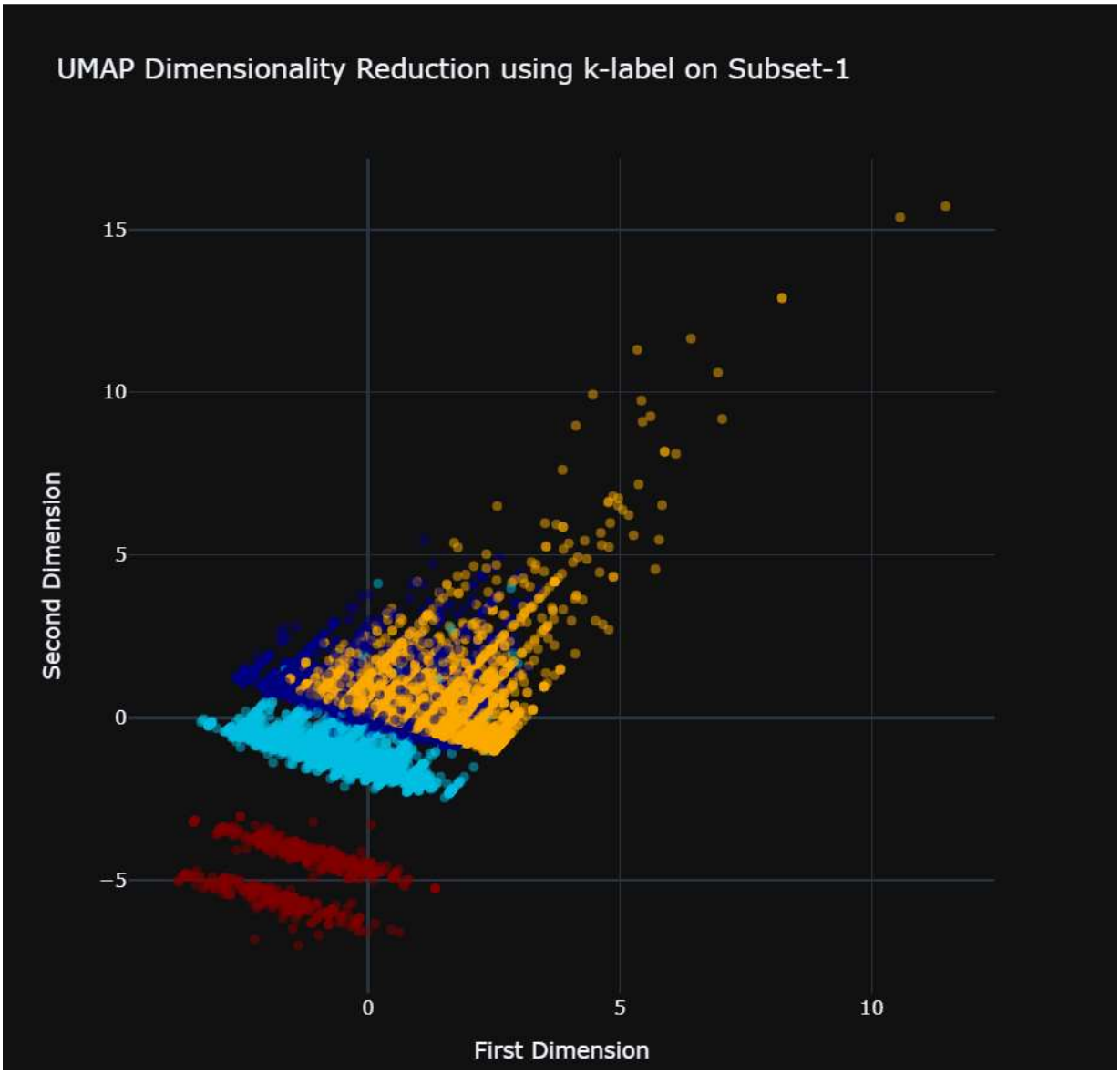


Fig 21 PCA Result on Subset 1

- **We did not get expected results, so we did not waste our time finding insights and understanding how clustering works**

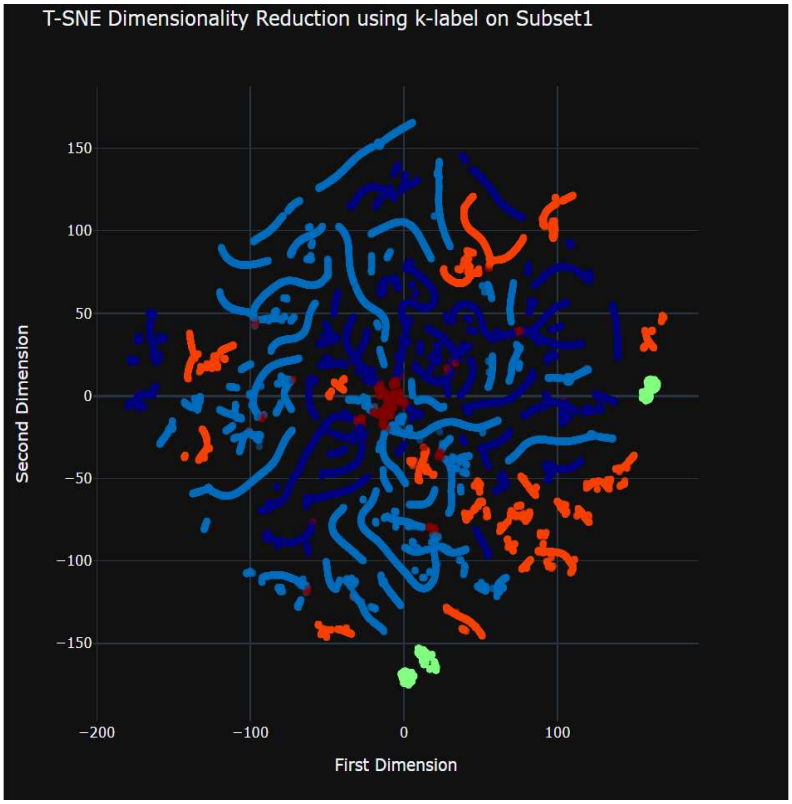


Fig 22 T-SNE Result on Subset 1

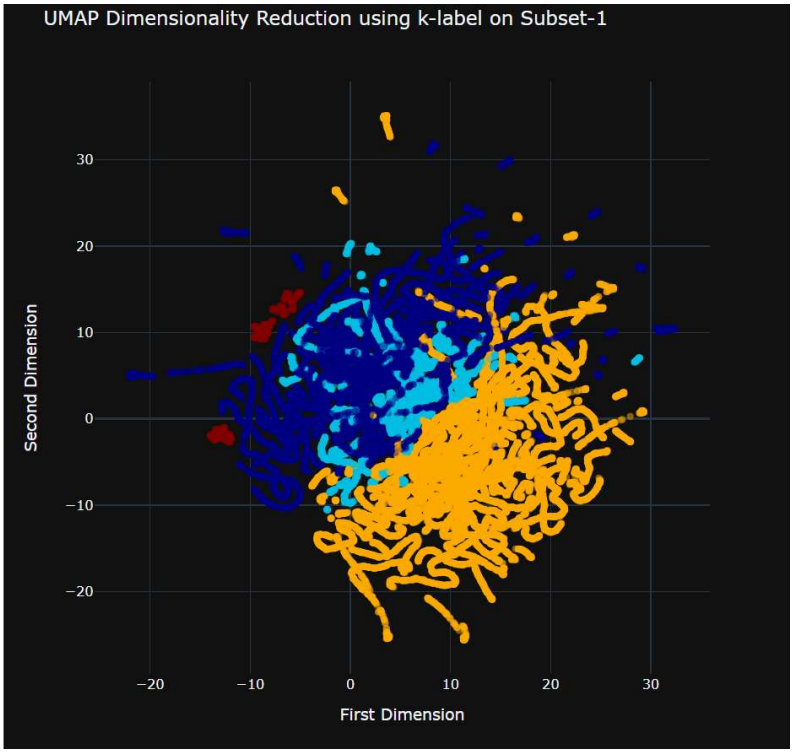


Fig.23 PCA Result on Subset 1

Subset 2:

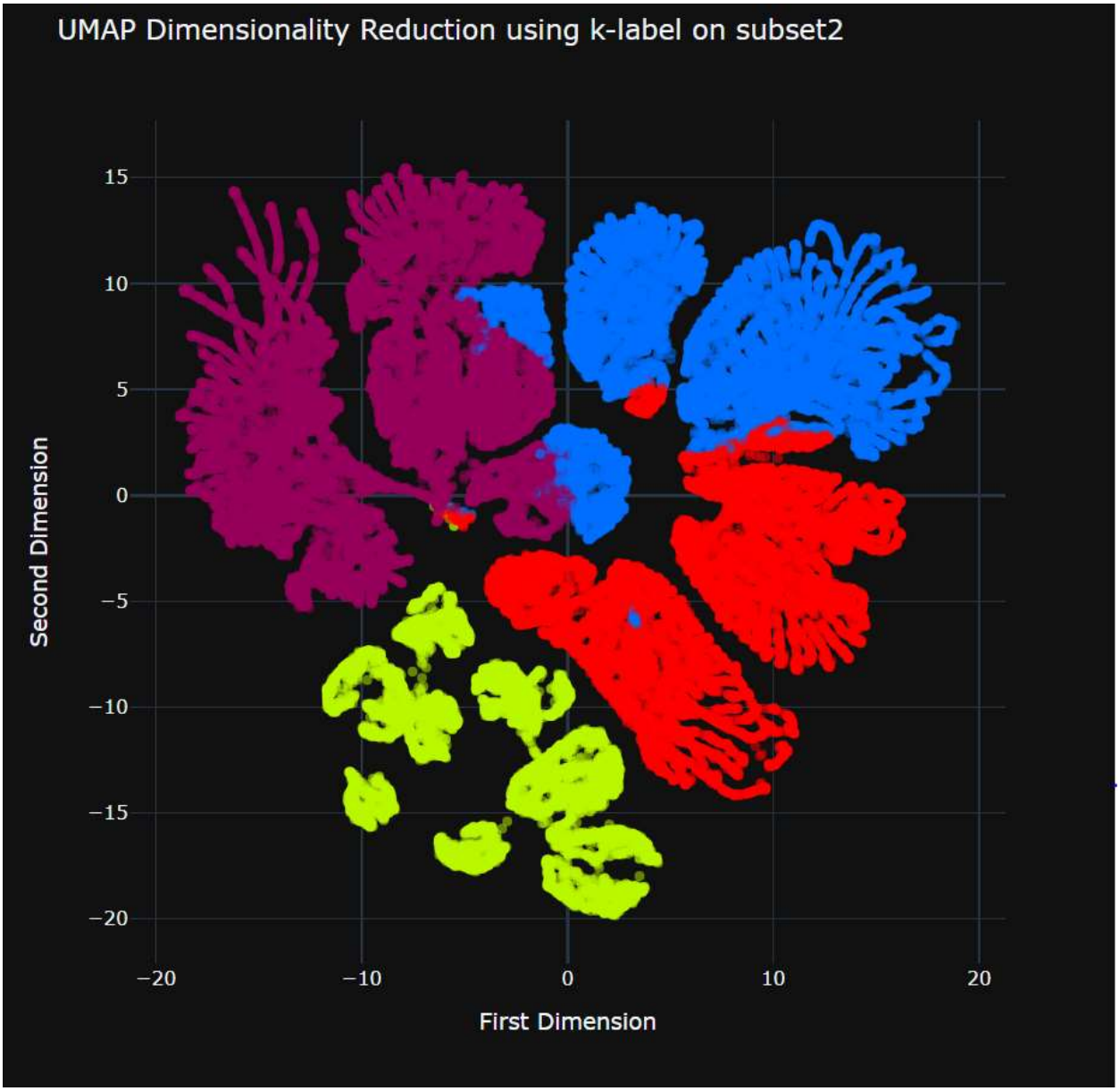


Fig 24 UMAP Result on Subset 2

We tried to analyse how are clusters grouped and below are for the same:

1. Cluster 1(Purple):

- Lower age is dominant range mostly 20's and 30's
- Education is mostly primary and secondary

- Do not have personal Loan
- Person has not taken housing Loan

2. Cluster 2(Blue):

- Dominant on the upper right side of the first dimension
- Person has taken housing loan
- Education is mostly primary and secondary
- Jobs are mostly unemployed, blue-collar, admin or housemaid

3. Cluster 3(Red):

- Dominant lower right side of the first dimension
- Person has taken housing loan
- Age of the person is on the higher side mostly above 30
- Education is mostly secondary and primary
-

4. Cluster 4(Green):

- Dominant on the lower side of second dimension below -5
- Person has taken Personal Loan
- Person has not taken housing loan
- Person's Education is mostly secondary

Insight for subset 2:

• Clients having educational status "secondary" usually have high balance with personal "no" and housing "no.

• Clients having job status 0 are usually in their 30's having low balance with personal "no "and housing "no".

• Clients in their 40's with job "Blue-collar" have descent balance usually in the range of 500-5000 with housing "yes" and personal "no".

• Clients in their 30's with educational status as "secondary" have housing "yes".

• Clients in their late 40's and 50's having job have very high balance with housing "yes".

• Clients with education status "tertiary "and job level either "management" or "self-employed" usually have housing "no" and personal "no".

• Clients in their 30's usually have both personal and housing as "yes".

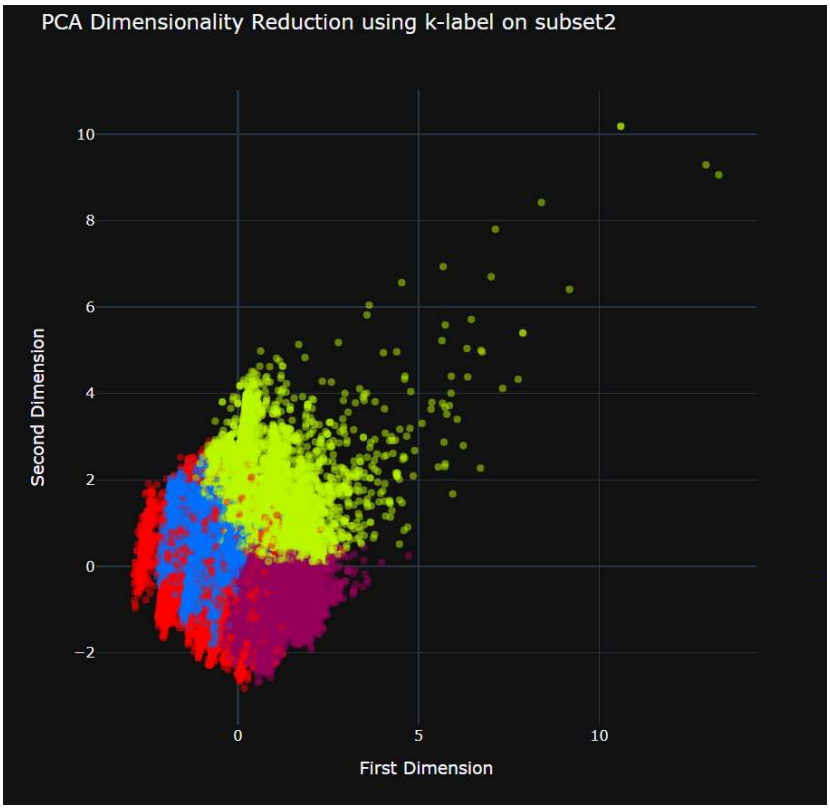


Fig 25 PCA Result on Subset 2

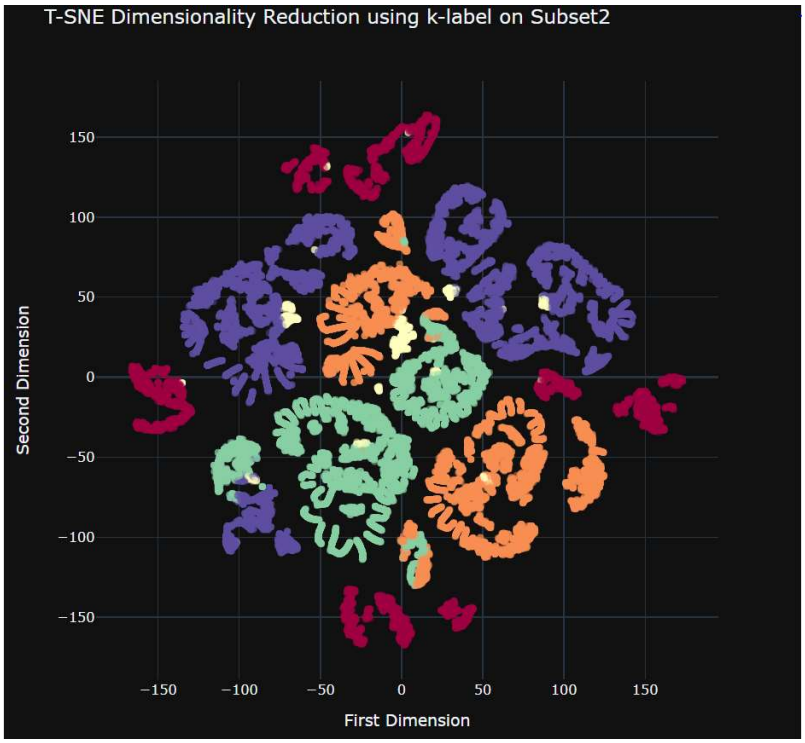


Fig 26 T-SNE Result on Subset 2

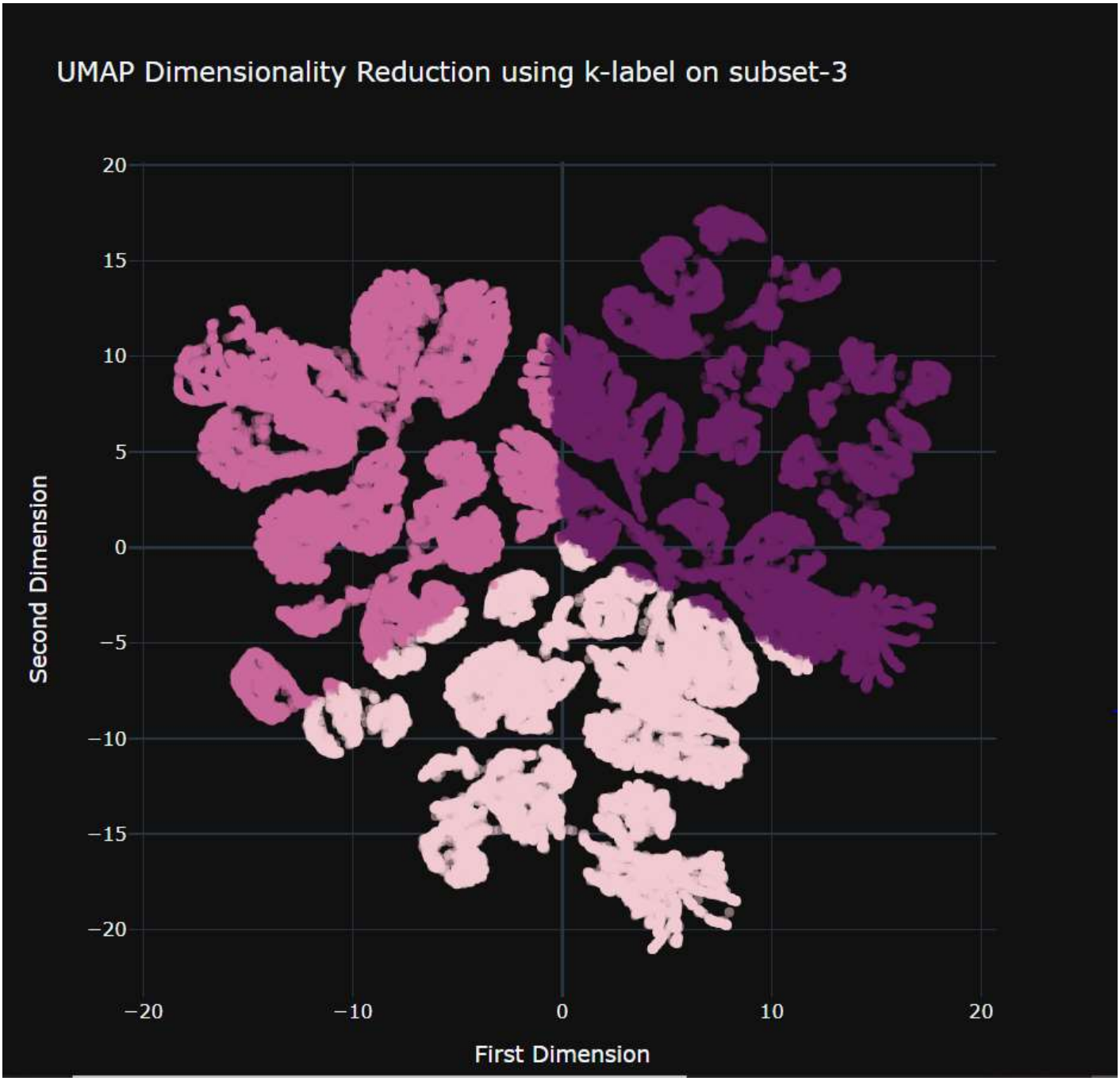


Figure 27 UMAP Result on Subset 3

We tried to analyse how are clusters grouped and below are for the same:

1. **Cluster 1(pink):**
    - Dominant on mostly left side of the first dimension
    - Education is primary and secondary with secondary being the most
    - Person has taken house loan
    - Person is mostly married however singles are also there
    - Balance is on lower side
2. **Cluster 2(Purple):**
    - Dominant on mostly right side of first dimension
    - Person is mostly divorced or married
    - Cluster has both personal and housing loan
3. **Cluster 3(Light Pink):**
    - Dominant on mostly right side or the first dimension
    - Person has not taken housing loan
    - Person has not taken house loan

Insight:

• Clients with marital status “married” and education “secondary” usually have balance less than 2000 with housing “yes”.

• Clients with education “primary” have personal “no” and housing “no”.

• Clients in their 30’s with job “management” marital status as “married” and education as “secondary” have balance less than 1000.

• Clients belonging to “management” are usually above 40’s having balance as less as 0.

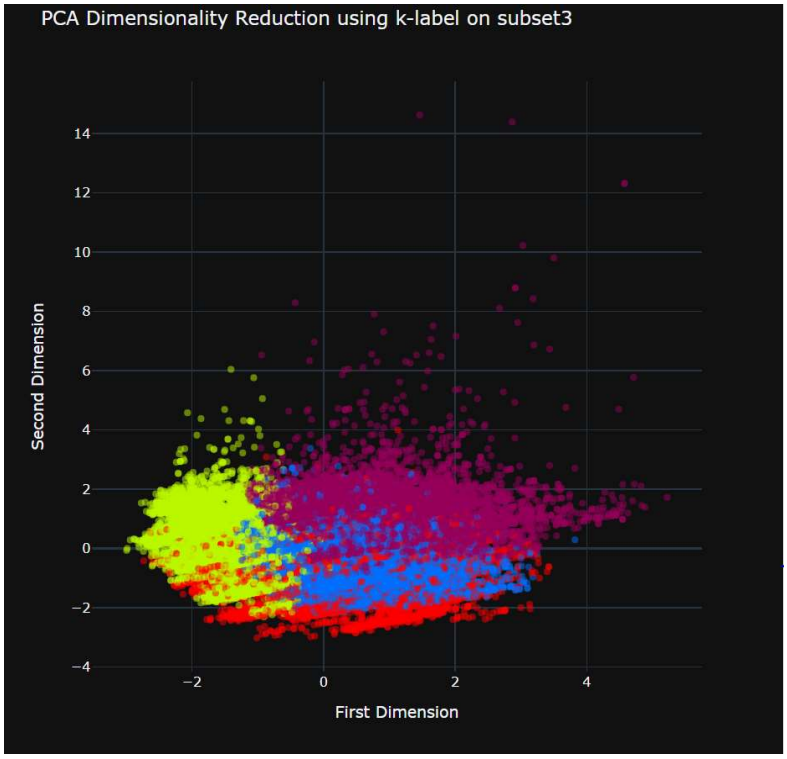


Fig 28 PCA Results on Subset 3

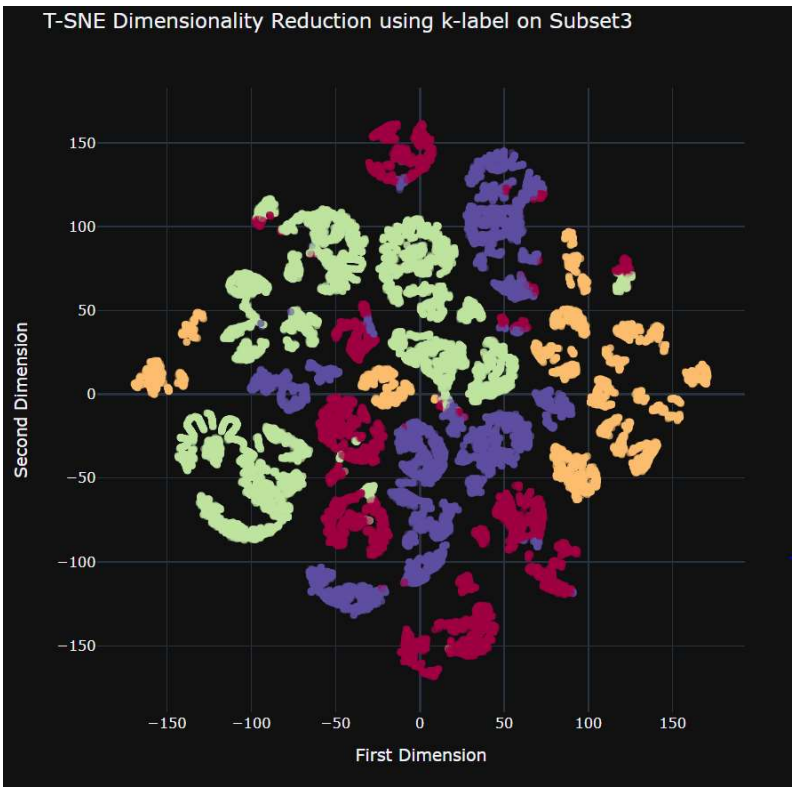


Fig 29 T-SNE Results on Subset 3

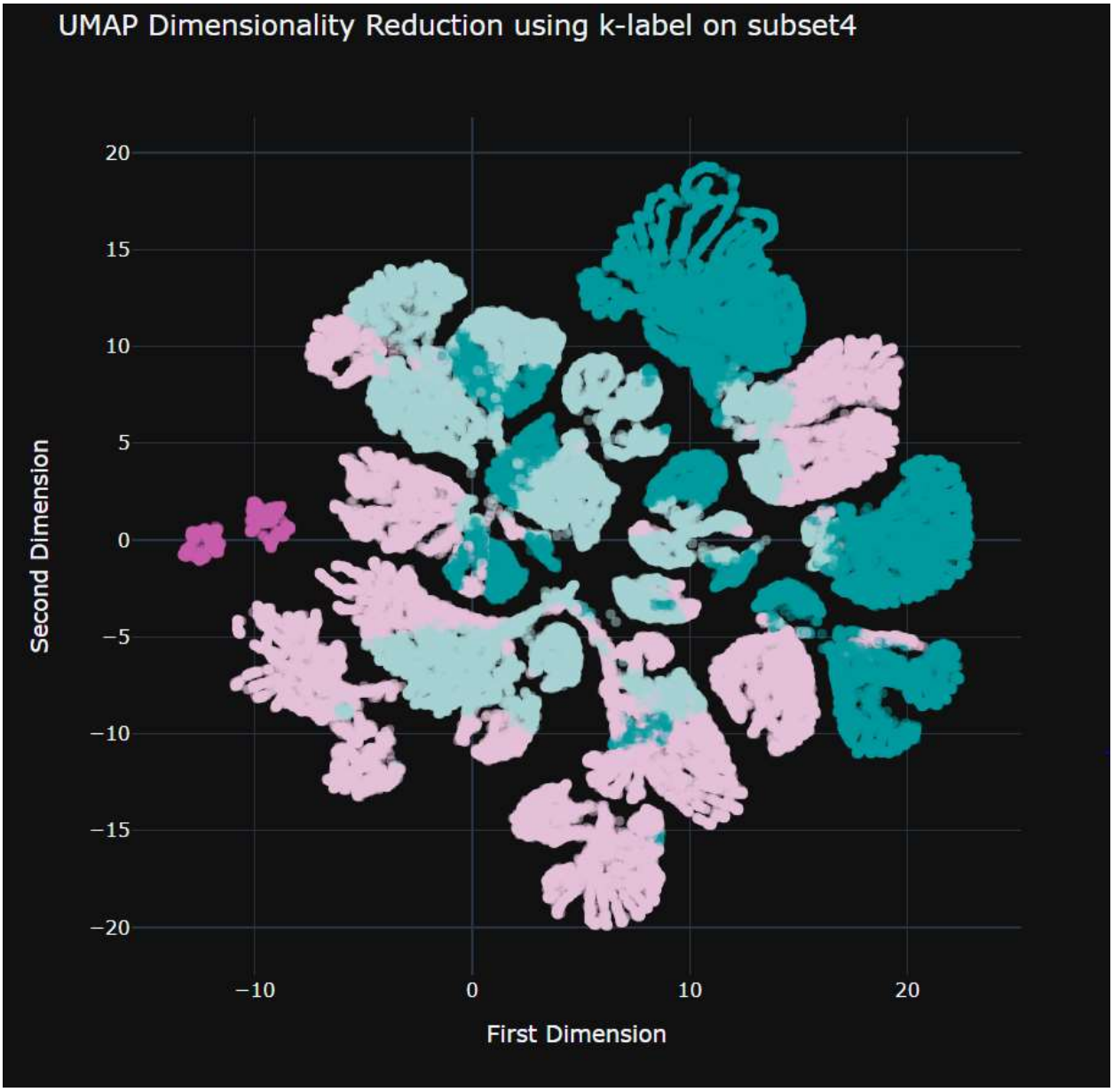


Fig 30 UMAP Results on Subset 4

- **We did not get expected results, so we did not waste our time finding insights and understanding how clustering works**

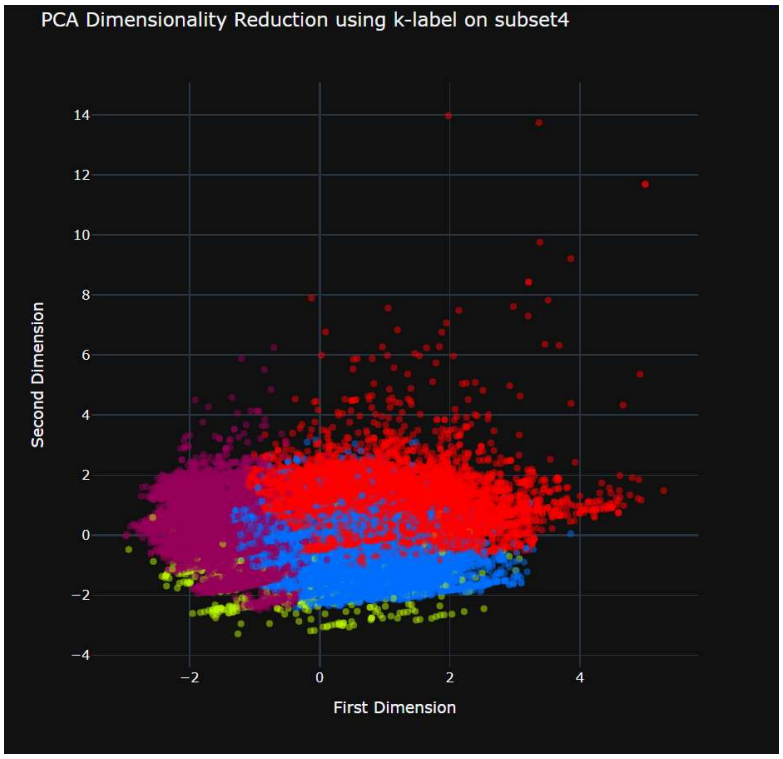


Fig 31 PCA Results on Subset 4

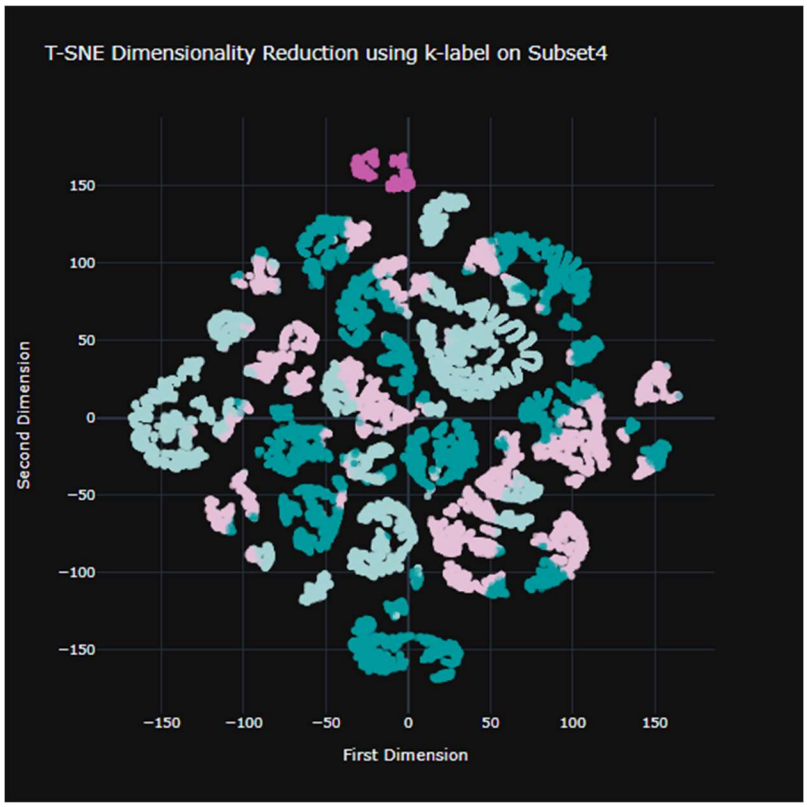


Fig 32 T-SNE Results on Subset 4

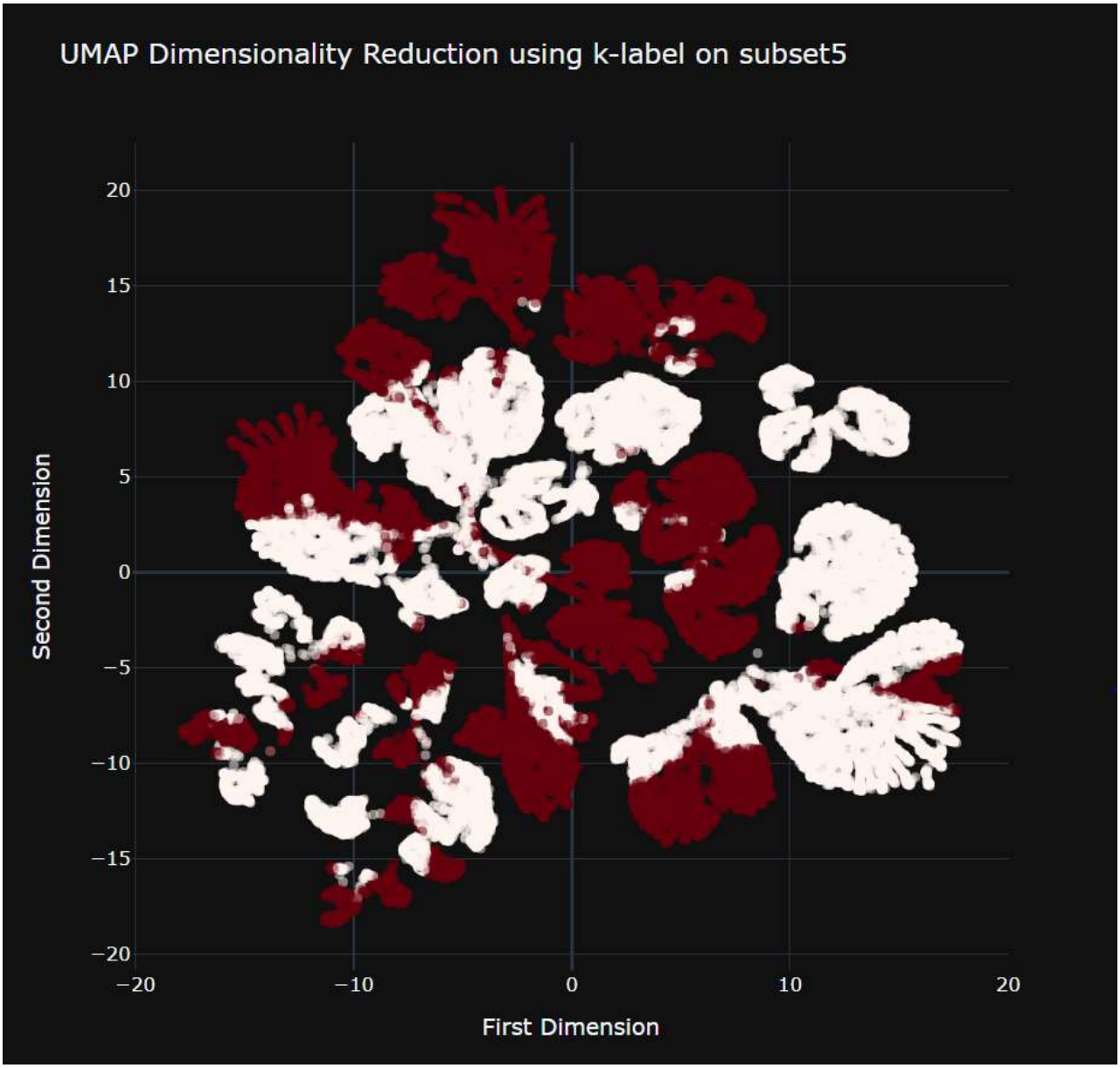


Fig 33 UMAP Results on Subset 5

We tried to analyse how are clusters grouped and below are for the same:

1. Cluster 1(Red):

- Dominant with education Secondary and tertiary
- Person has not taken personal loan
- Average yearly balance is on the lower side with value less than 3000
- Person has taken house loan

2. Cluster 2(Yellow):

- Age of the person is on the larger side above 40
- Education of a person is mostly
- Person has not taken house loan

Insights on Subset 5:

• Clients with marital status “married” and in their 60’s have education “primary” and personal “no”.

• Clients having job level “student” or “technician” with marital status as “single” and education “secondary” usually have housing loan as “no”.

• Clients in their 30’s with job status as “student “ marital status as “ married” education “secondary” have personal “no” and housing “yes”.

• Clients marital status as “married” and education “tertiary” have personal “yes” housing “yes” with balance usually greater than 2000.

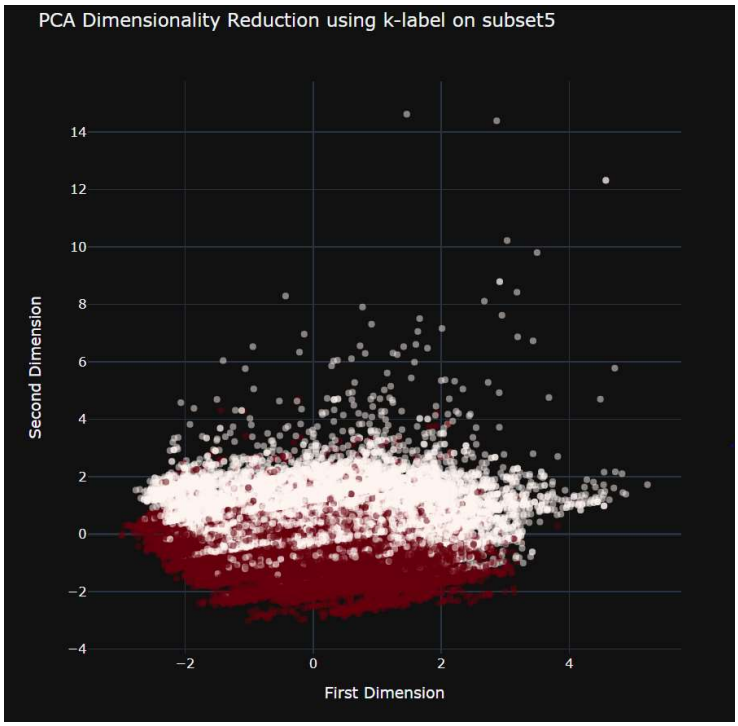


Fig 34 PCA Results on Subset 5

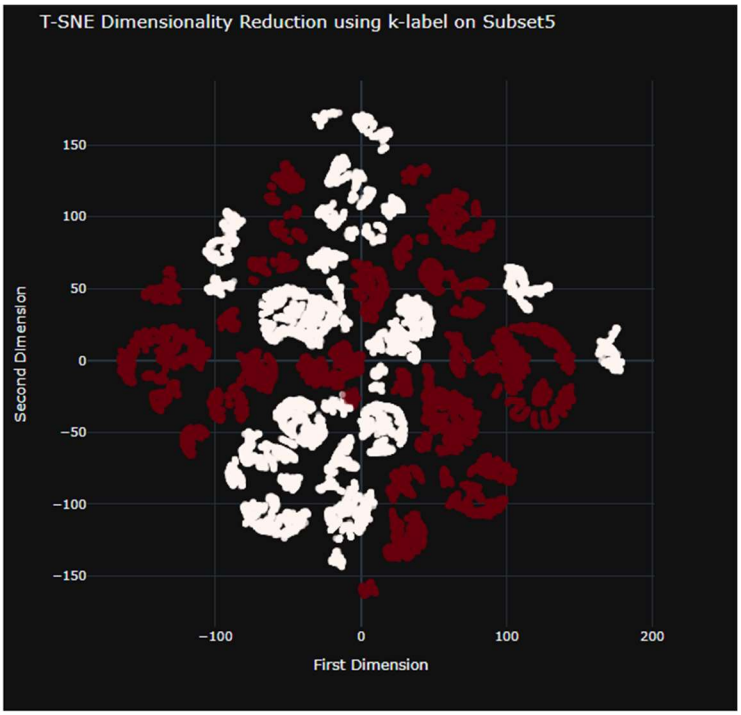


Fig 35 T-SNE Results on Subset 5

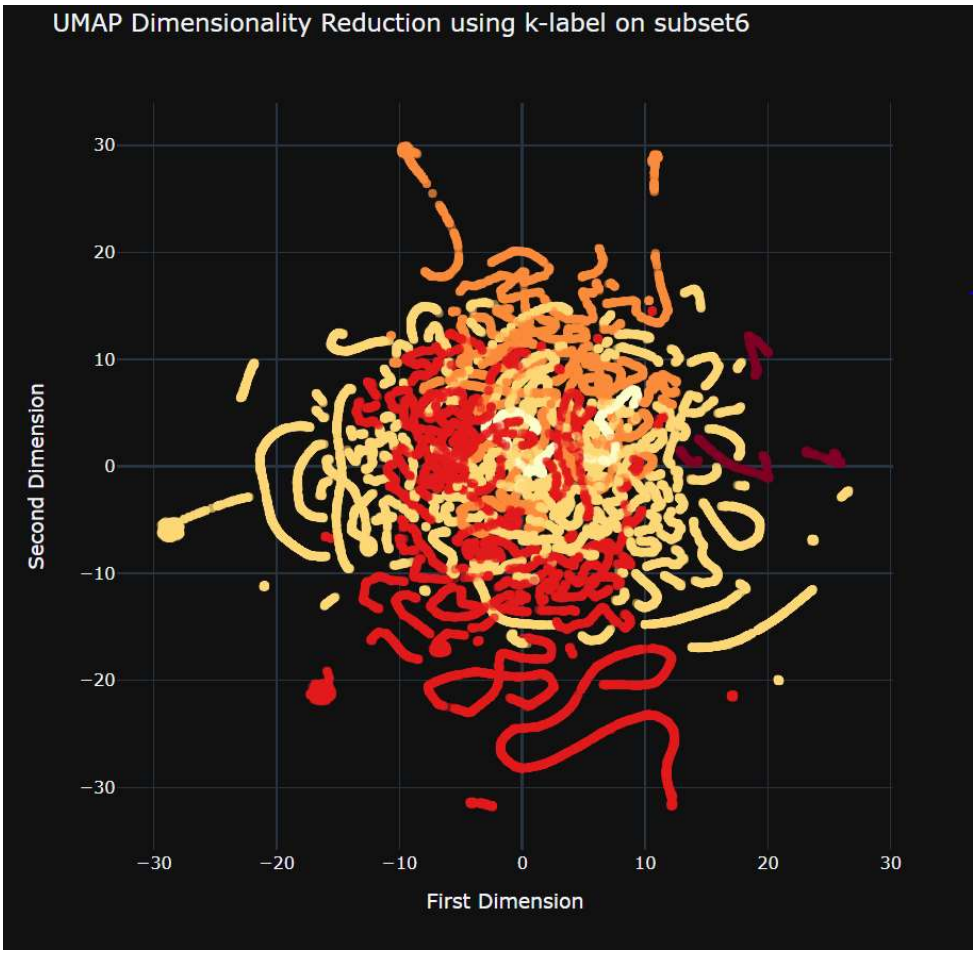


Fig 36 UMAP Results on Subset 6

- **We did not get expected results, so we did not waste our time finding insights and understanding how clustering works**

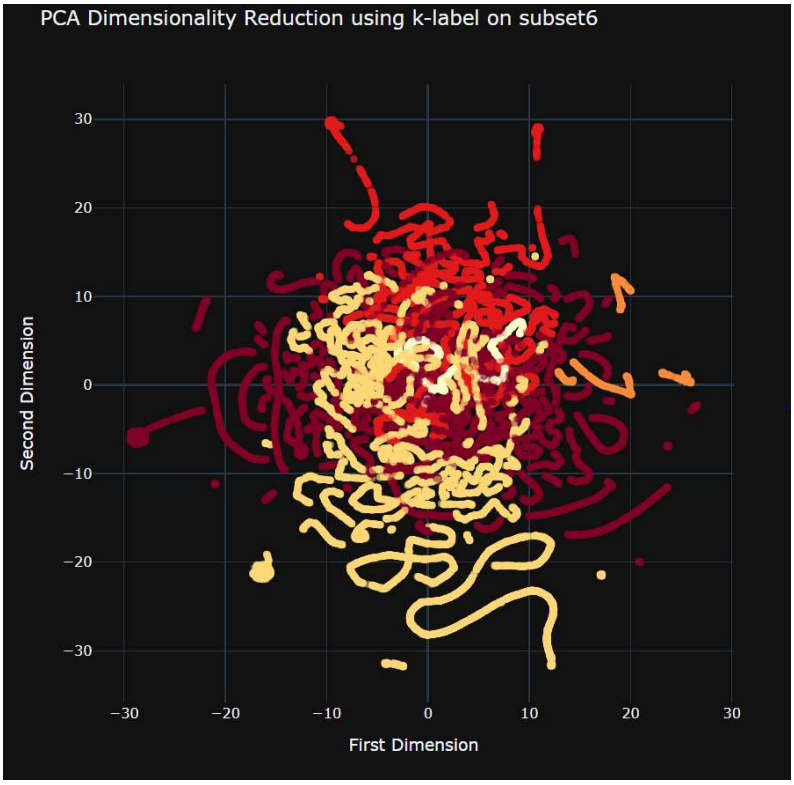


Fig 37 PCA Results on Subset 6

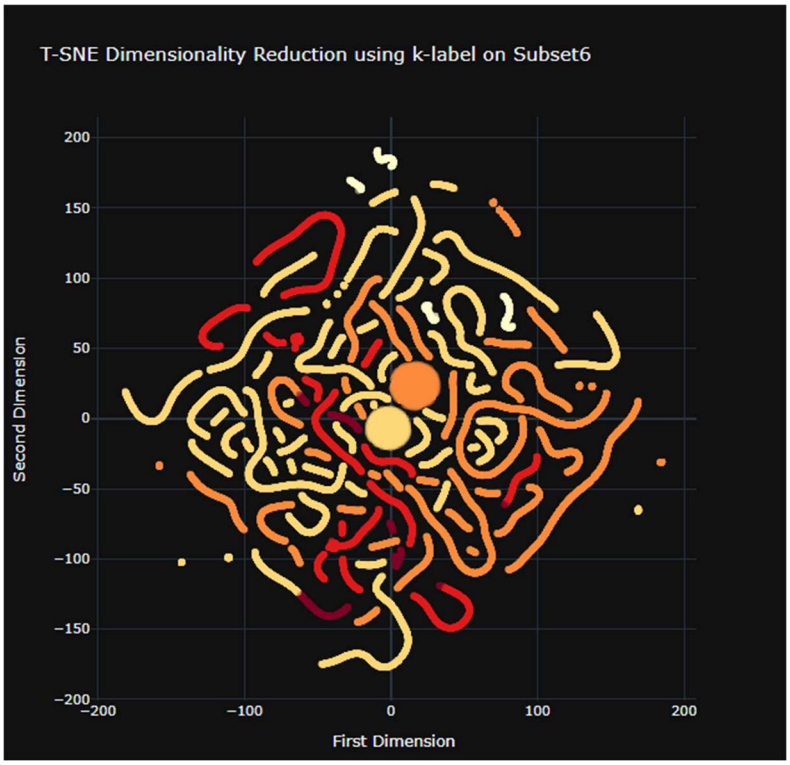


Fig 38 T-SNE Results on Subset 6

6. Critical Analysis for Patches Dataset:

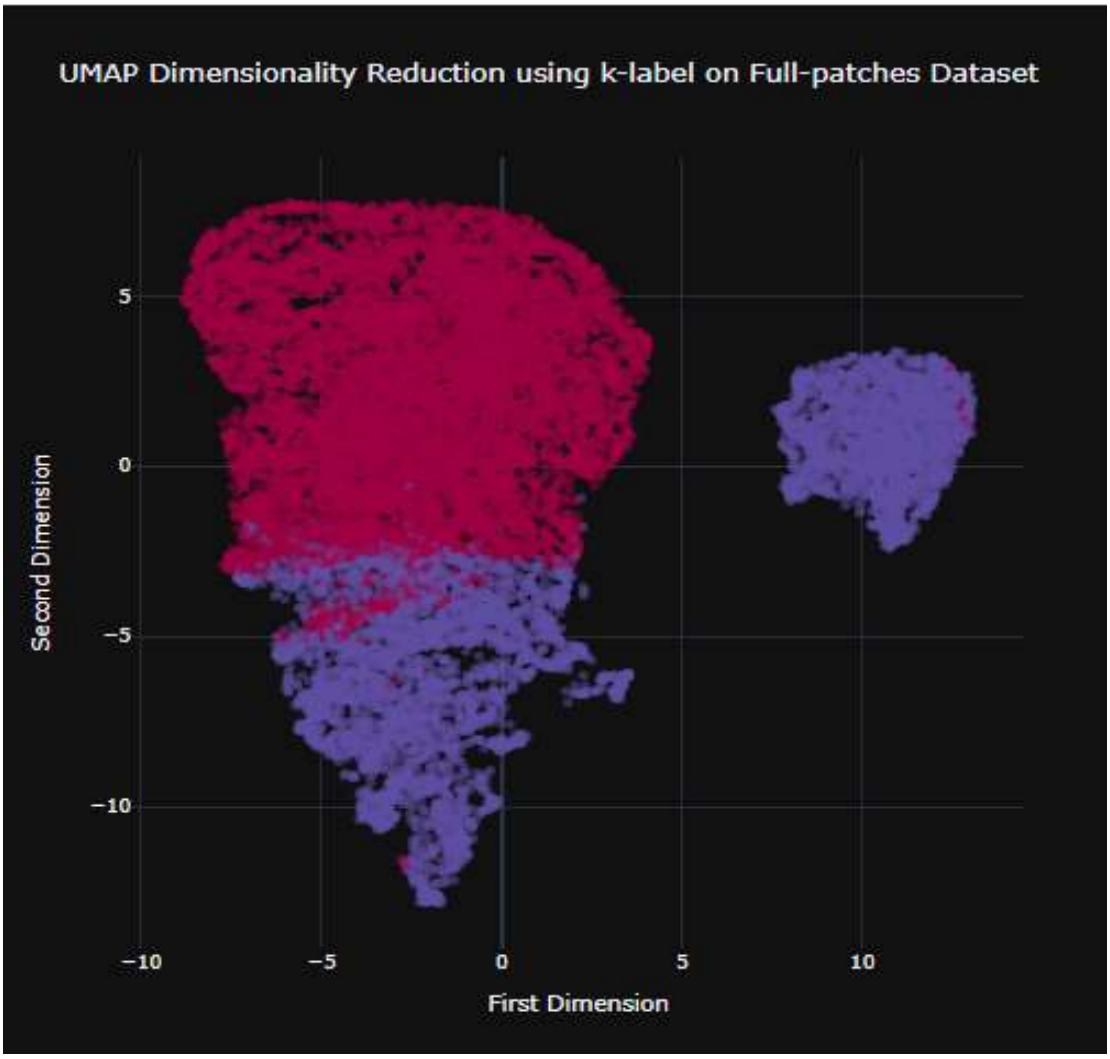


Fig 39 UMAP Results on Full Patches Dataset

We tried to analyse how are clusters grouped and below are for the same:

1. Cluster 1-Purple

- Dominant from -10 to -3 on second dimension
- Elevation is mostly above 3400
- Horizontal road is mostly on lower side below 2000

2. Cluster 2- Red

- Domninant from -3 and above

- Elevation mostly on lower side
- Horizontal road is mostly on higher side

Insights and Analysis:

Full Dataset:

- If the elevation range is between 2600 to 3000 usually have vertical hydro less than 80.
- If the vertical hydro is on the lower side with the slope less than 20 usually have horizontal fire value between 400 to 2000.
- If the slope is on higher side that is above 20, elevation is mostly on the lower side.
- If the elevation is on lower side the horizontal road is also on the lower side i.e elevation is directly proportional to horizontal road.
- If the horizontal road is on the higher side that is with value above 3000 in the range of 110-240.

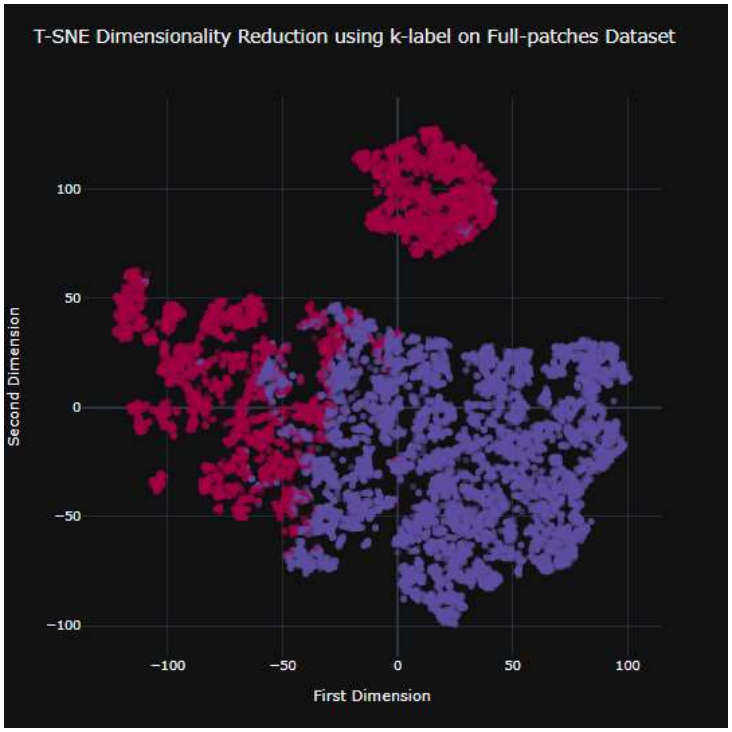


Fig 38 T-SNE Results on Full Patches Dataset

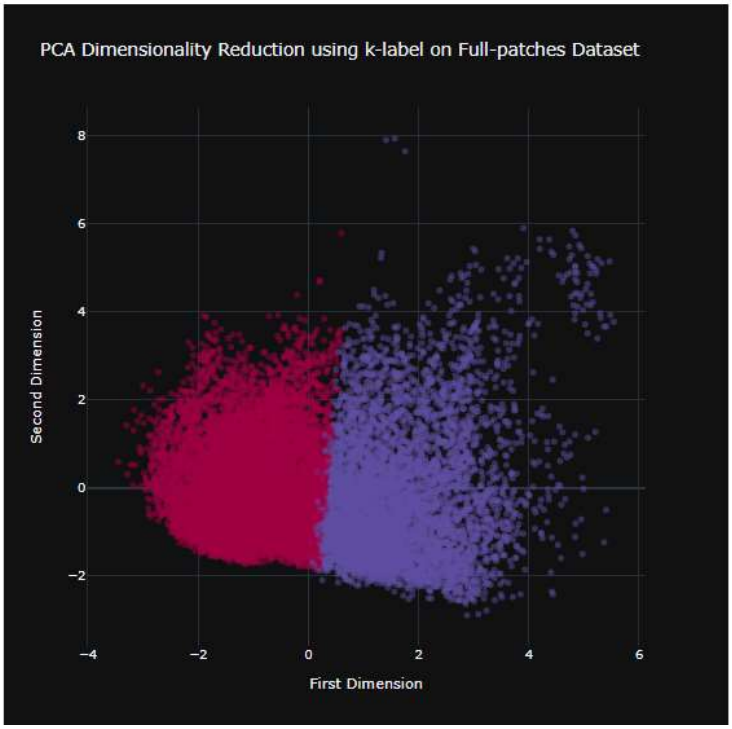


Fig 38 PCA Results on Full Patches Dataset

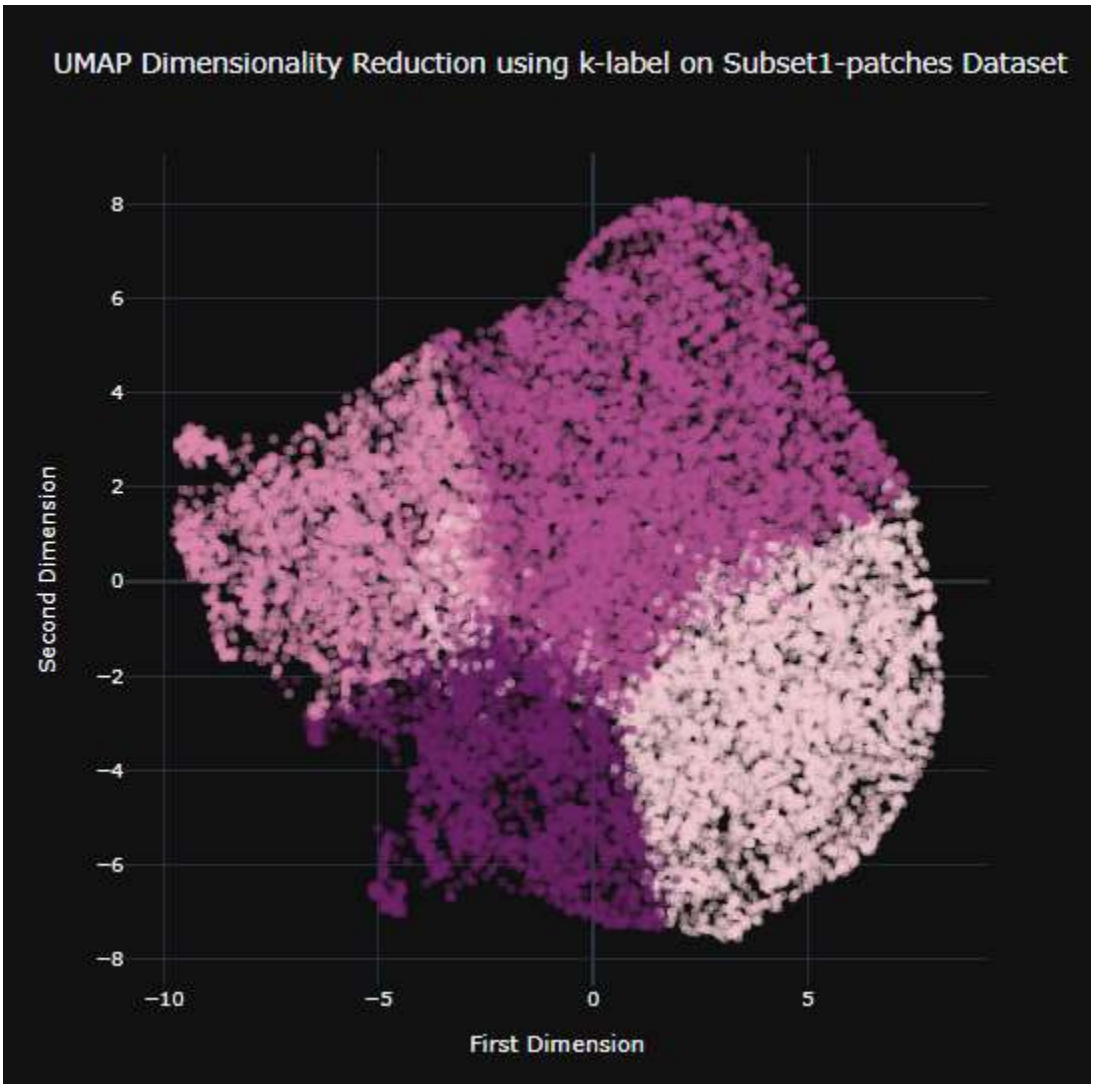


Fig 40 UMAP Results on Subset 1

We tried to analyse how are clusters grouped and below are for the same:

1. Cluster 1 (pink):

- Dominant from -10 to -5 on first dimension
- Vertical Hydro is on lower side mostly less than 80

- Elevation range is between 2600 to 3000

2. **Cluster 2 (purple):**
    - Dominant from 0 to above on first dimension
    - Slope is on lower side mostly less than 20
    - Horizontal fire value is between 400 to 2000
    - Vertical hydro is on lower side

3. **Cluster 3 (Light purple) :**
    - Dominant from 0 to above on first dimension
    - Elevation is mostly on lower side
    - Slope is on higher side mostly above 20
    - Horizontal road is on lower side mostly 200 to 1600 range
4. **Cluster 4 (Dark purple):**
    - Dominant from -2 to -8 on second dimension
    - Horizontal road is on higher side with value above 3000
    - Vertical hydro range between 110 and 240

Insights for Subset 1:

- If the Tree belongs to others vertical hydro is usually on the smaller side.
- If the Tree belongs to Spruce the horizontal hydro is usually on the lower side.
- If the Tree belongs to others vertical hydro is less than 70 with horizontal hydro being on the lower side.

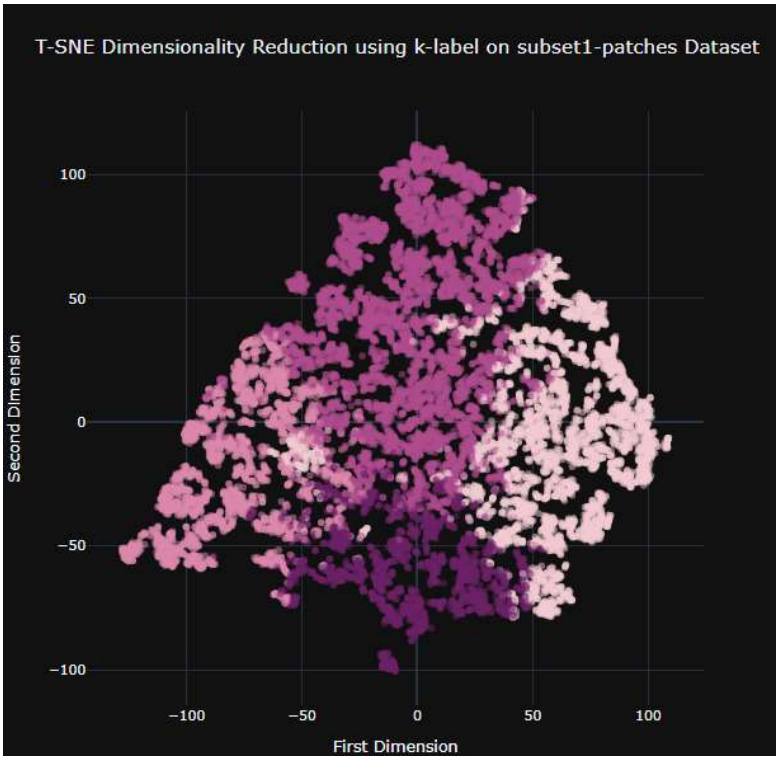


Fig 41 T-SNE Results on Subset 1

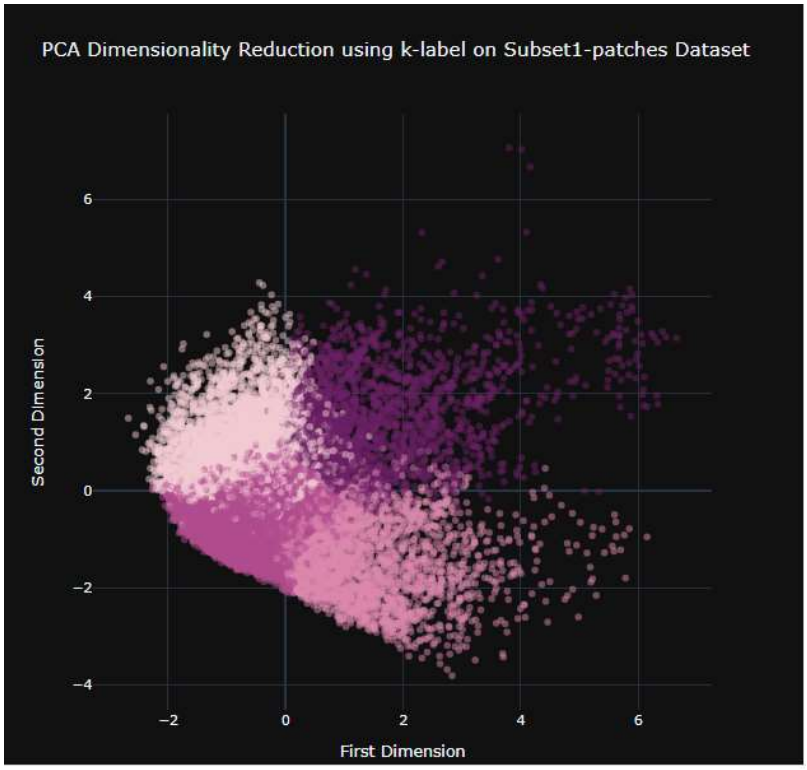


Fig 42 PCA Results on Subset 1

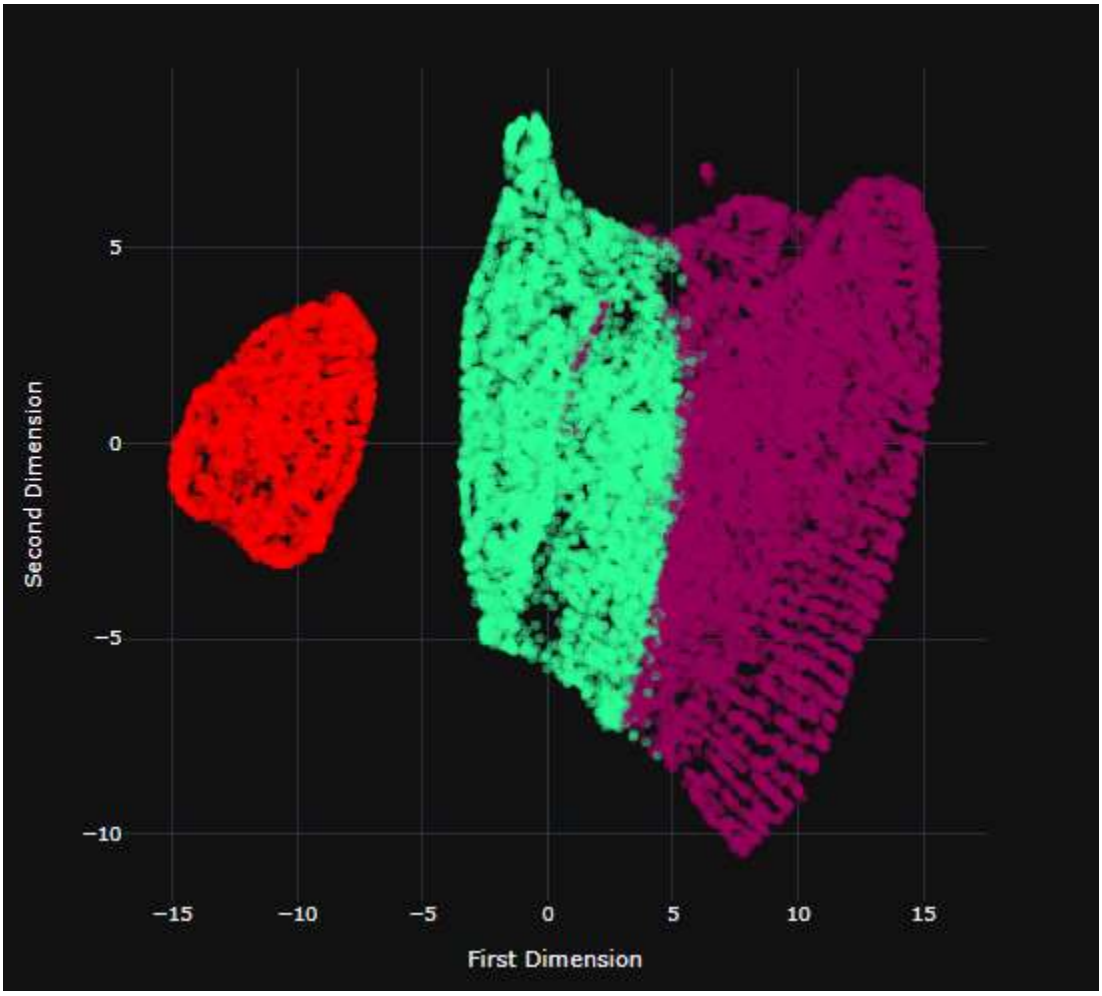


Fig 43 UMAP Results on Subset 3

We tried to analyse how are clusters grouped and below are for the same:

1. Cluster 1 (Red):

- Dominant on the left side of the first dimension
- Tree is 0

- Vertical hydro is on smaller side

2. Cluster 2 (Green):

- Dominant from -5 to 2 on first dimension
- Tree is 1
- Horizontal hydro is between range 100 to 600

3. Cluster 3 (Purple):

- Dominant on the right side of the first dimension
- Horizontal hydro is on lower side
- Tree is 1

Insights for Subset 3:

- With vertical hydro on the higher side and slope less than 20. The Tree usually belongs to others.
- Vertical hydro less than 75 and slope on the lower side. The Tree usually belongs to others.
- With Vertical hydro on the lower side with elevation mostly above 2600.The Tree belong to Spruce.

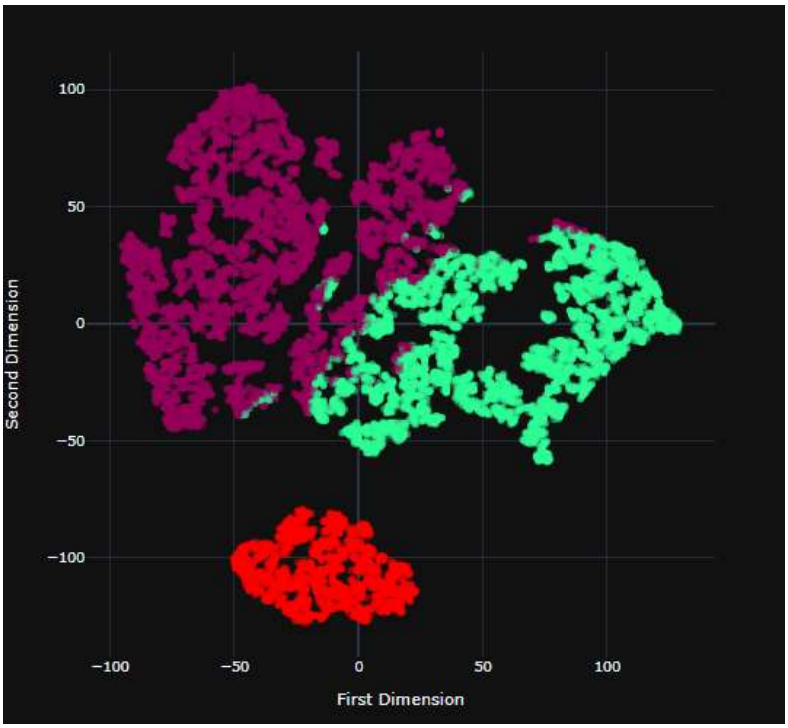


Fig 44 T-SNE Results on Subset 3

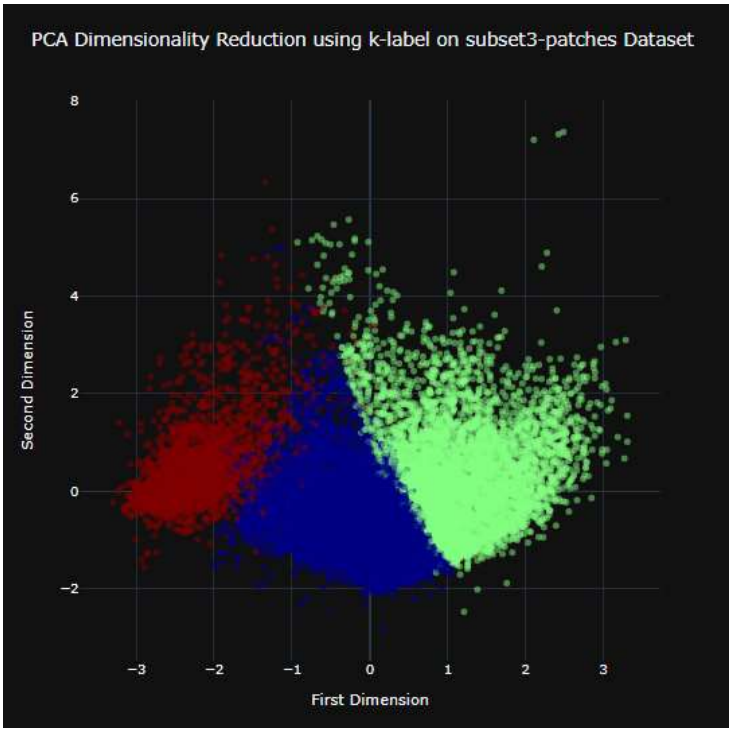


Fig 45 PCA Results on Subset 3

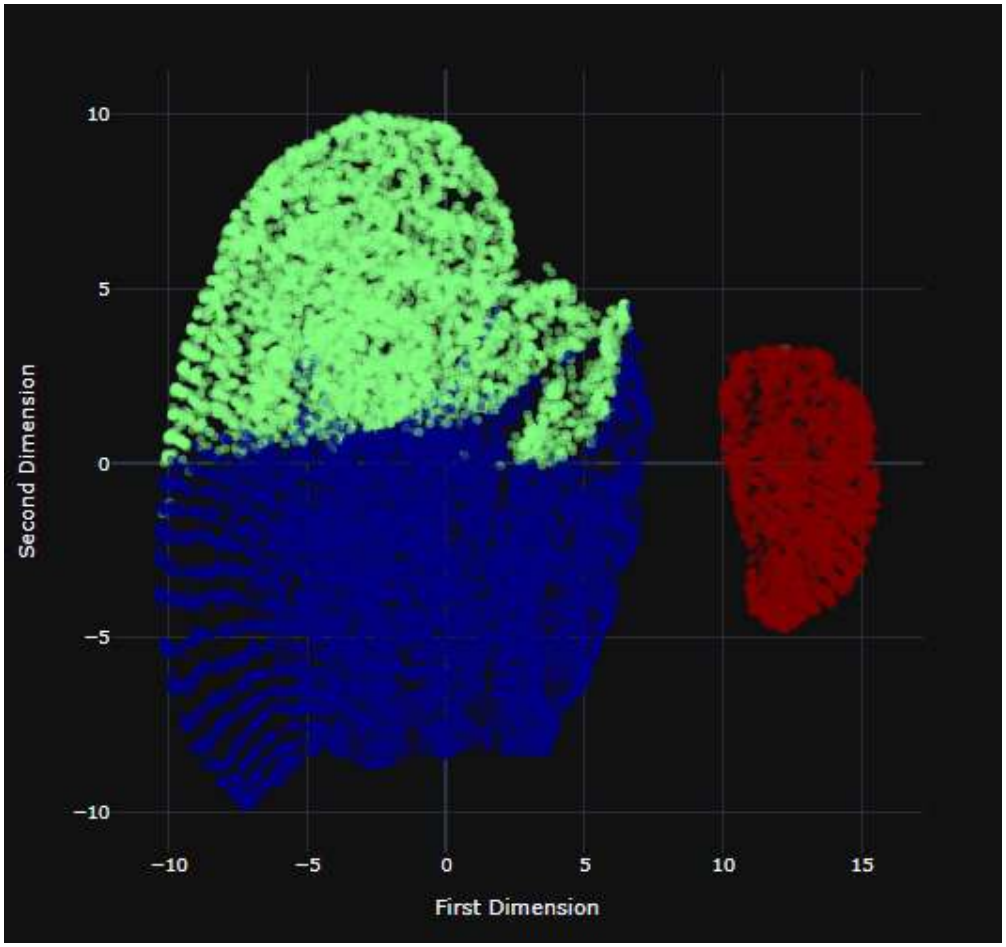


Fig 45 UMAP Results on Subset 3

We tried to analyse how are clusters grouped and below are for the same:

1. **Cluster 1(Green):**
    - Dominant on the upper side of the second dimension
    - Vertical hydro is mostly on the higher side
    - Value of Tree 0
2. **Cluster 2 (Blue):**
    - Dominant on the lower side of the second dimension
    - Vertical hydro is mostly on lower side with value less than 75
    - Slope is mostly on lower side
    - Value of Tree 0
3. **Cluster 3 (Red):**

- Dominant on the right side of the first dimension
- Vertical hydro is on lower side
- Elevation is on higher side mostly above 2600
- Value of Tree is 1\

Insights for Subset 3

- With vertical hydro on the higher side and slope less than 20. The Tree usually belongs to others.
- Vertical hydro less than 75 and slope on the lower side. The Tree usually belongs to others.
- With Vertical hydro on the lower side with elevation mostly above 2600.The Tree belong to Spruce.

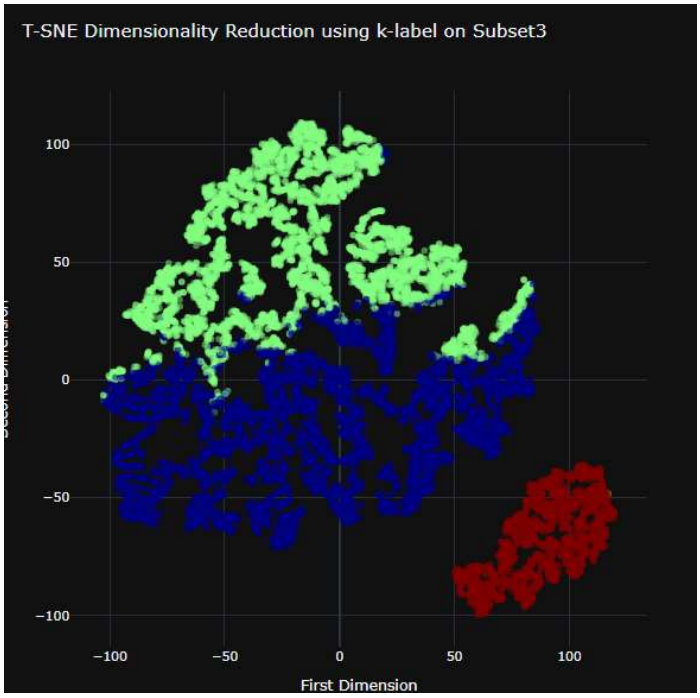


Fig 46 T-SNE Results on Subset 3

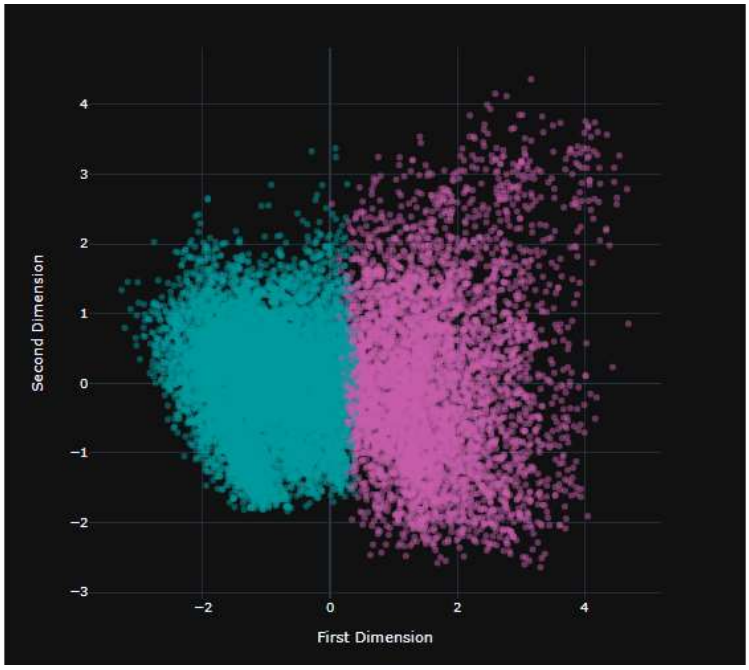


Fig 45 PCA Results on Subset 3

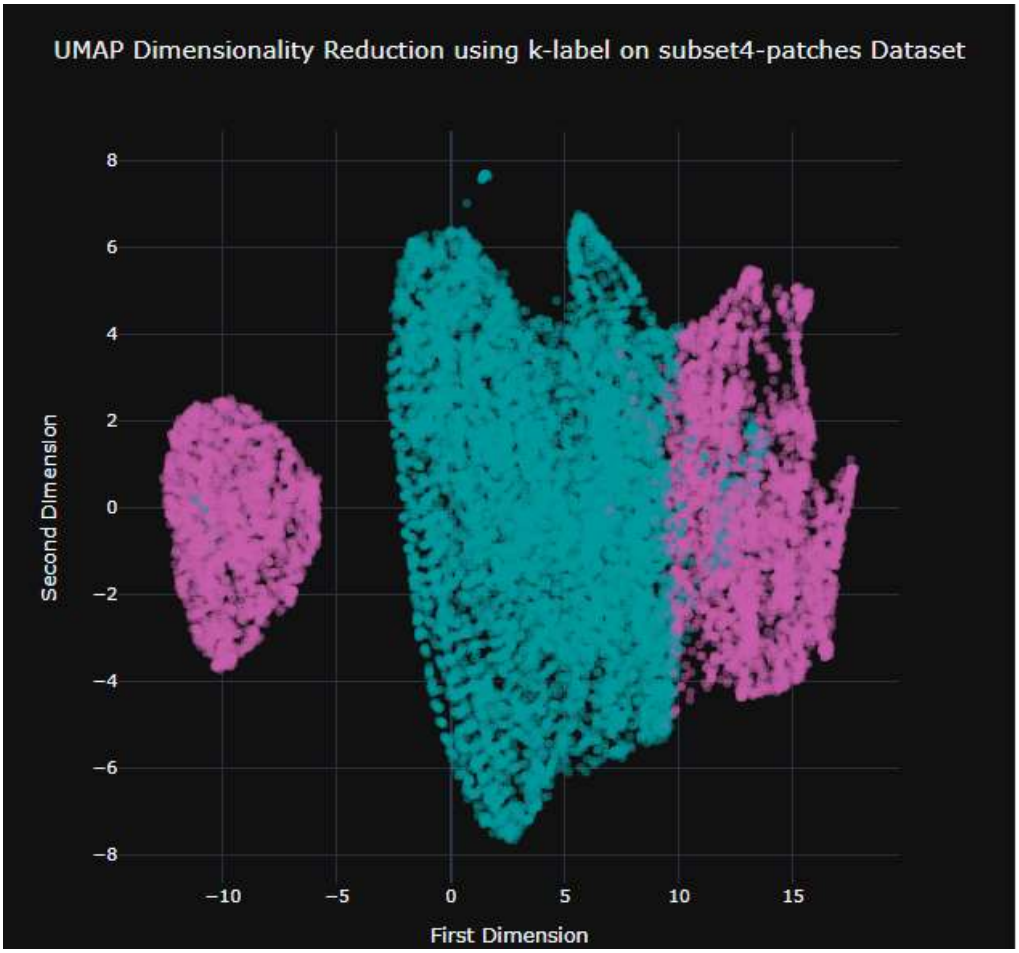


Fig 46 UMAP Results on Subset 4

We tried to analyse how are clusters grouped and below are for the same:

1. **Cluster 1(Pink):**
    - Dominant on the left side of the left side of first dimension
    - Elevation is mostly on lower side
    - Horizontal hydro is on Higher side
2. **Cluster 2(Blue):**
    - Dominant on the mid of the graph .i.e in the range between -5 to 10 on first dimension
    - Elevation is mostly on higher side

Insights for Subset 4:

- If the elevation is on the lower side, horizontal hydro is usually on the higher side.
- With both elevation and the slope on the higher side, the tree belongs to other's category.
- With both horizontal road and elevation on the higher side. The Tree always belongs to others.

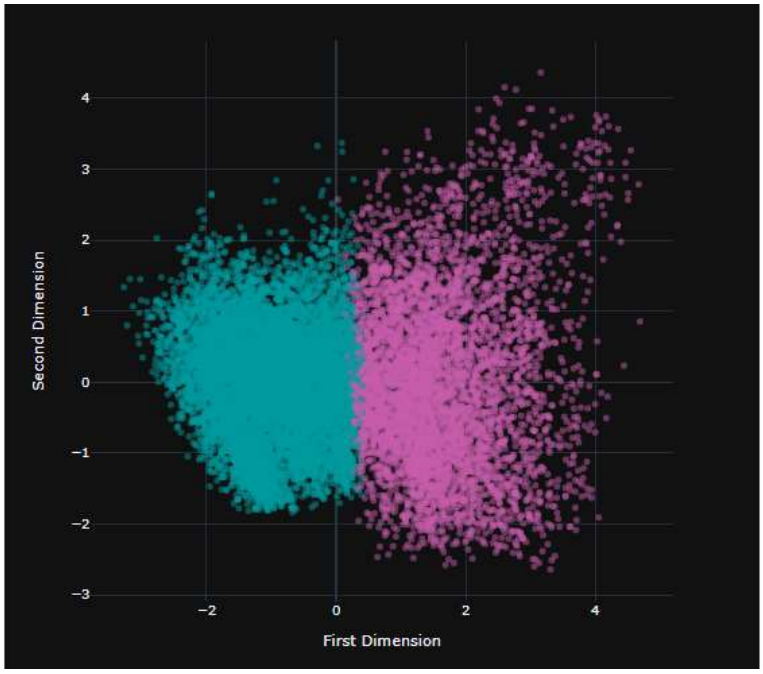


Fig 47 PCA Results on Subset 4

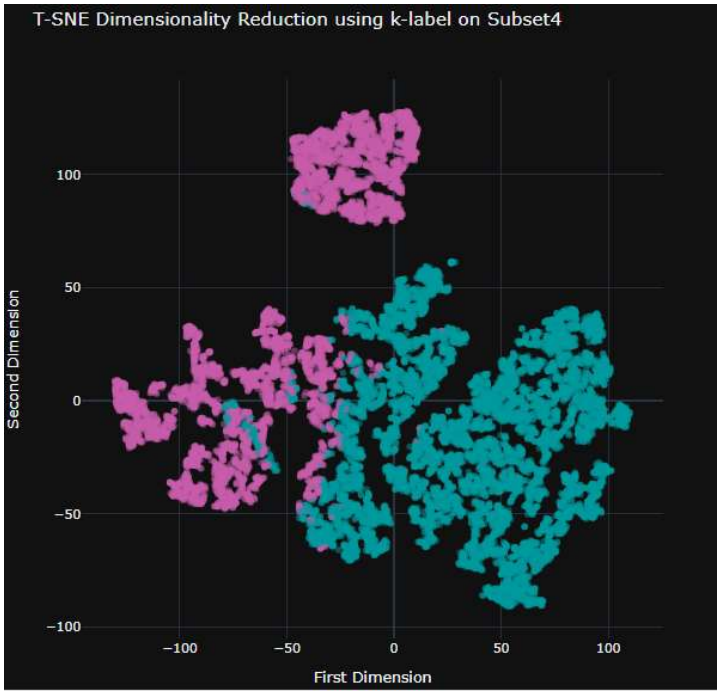


Fig 48 T-SNE Results on Subset 4

7. Conclusion

We would like to conclude that, UMAP and t-SNE give almost the same results, but UMAP clustering was a bit better. UMAP, t-SNE, PCA didn't work well for clients-subset1 and clients-subset6. T-SNE can take life long to run. PCA is only suitable for basic cluster understanding.

8. Team Contribution

Bharat started with understanding the study and forming objectives and passing on the information to the team. With the given information to approach Prasad started with data exploration and data understanding, followed by data pre-processing(i.e., Scaling, Normalization, and discretization). Then he started applying PCA with k-means algorithm, where he used k-means to extract the number of components using elbow plot. While kinjal researched on t-SNE and gave all of her time waiting for t-SNE results and still finding the best adjustments for perplexity and n_iter. Also, she kept her patience to understand how clusters are formed. Parallelly Bharat was studying UMAP and how does it work while creating interestingly fast and colorful visualizations. Once all the results came Bharat did an in-depth analysis of all of the findings comparing the dimensional reduction techniques and feeling lucky with the existence of UMAP. Once done, Bharat extracted insights and kinjal recorded the same in the presentation.

9. Individual Contribution

Bharat Jethwani :

He researched on UMAP, and how does it work. Comparing different methodology and forming the best algorithm that would work for our study. Along with Comparing the dimensional reduction techniques and extracting the insights. He contributed his work to words in the report.

Kinjal Maru:

She researched on t-SNE, and how does it work i.e. how would changing perplexity, n_iter affect the visualization and understanding how clusters are forming groups, not just contributed her work into words but also voice.

Prasad Tambe:

He started with exploring the data, converting categorical data, Scaling the data, creating sub-sets using different feature selection techniques and applying PCA with k-means, And contributed his work into words in the report.