k

By Gaurav Taori(10525875)

Python Data Visualization for datasets- “Clients” and “Patches”

**Data Visualization for Dataset “Clients”:**

**Introduction**

This dataset refers to bank client information. Each row corresponds to a client.

**Number of Instances**: 43,193

**Number of Variables**: 9

1 - age (numeric)

2 - job : type of job (categorical: "admin.", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")

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

4 - education (categorical: "secondary", "primary", "tertiary")

5 - default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

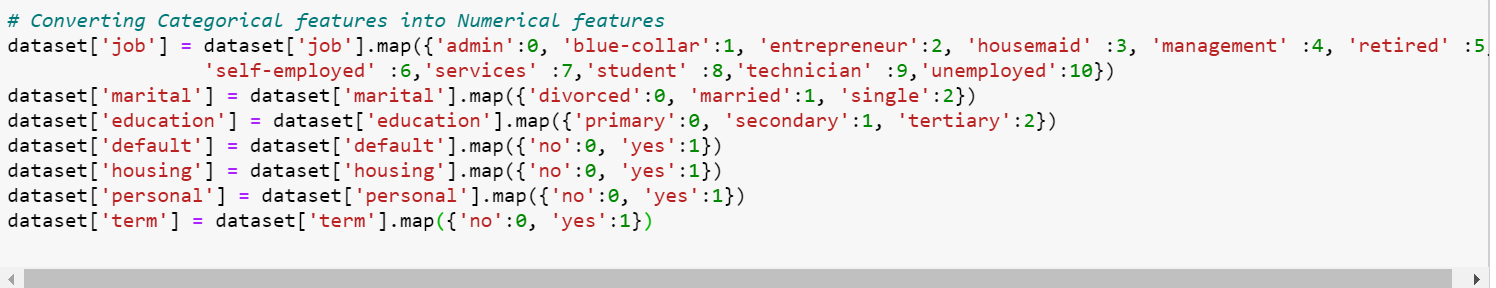
7 - housing: has housing loan? (binary: "yes", "no")

8 - personal: has personal loan? (binary: "yes", "no")

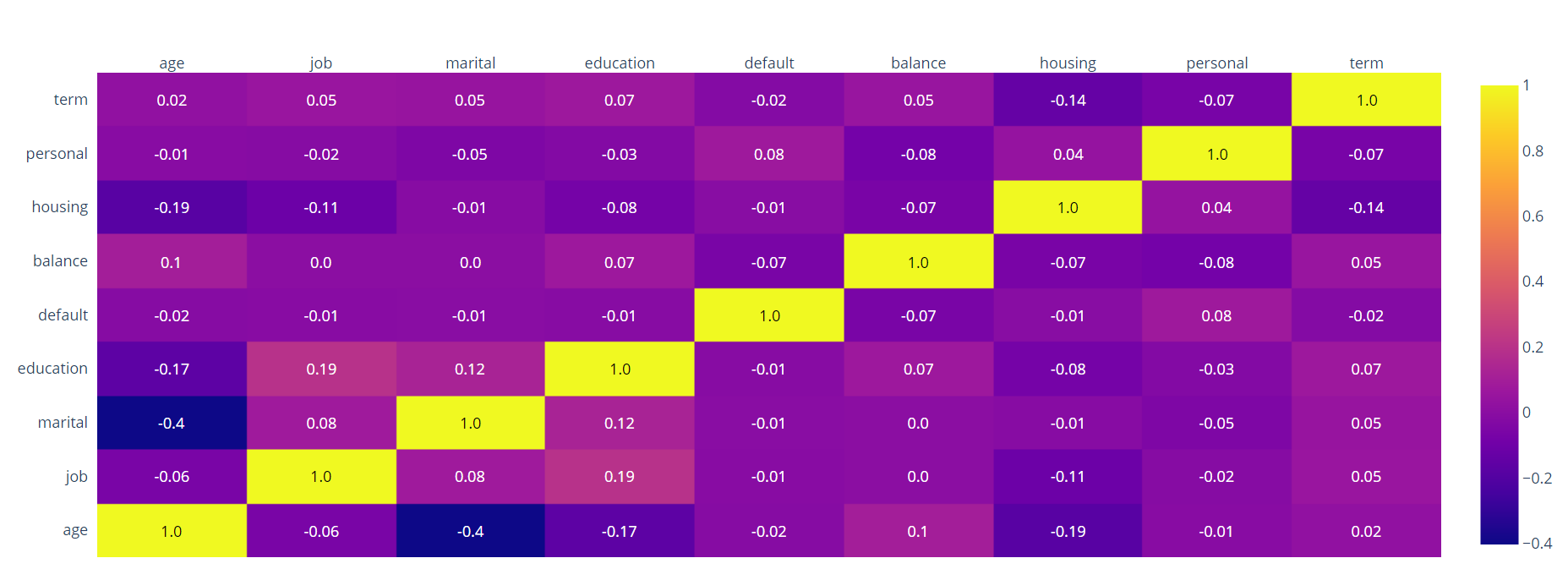
9 - term: has term deposit? (binary: "yes", "no")

**Critical Visual Design decisions/approach:**

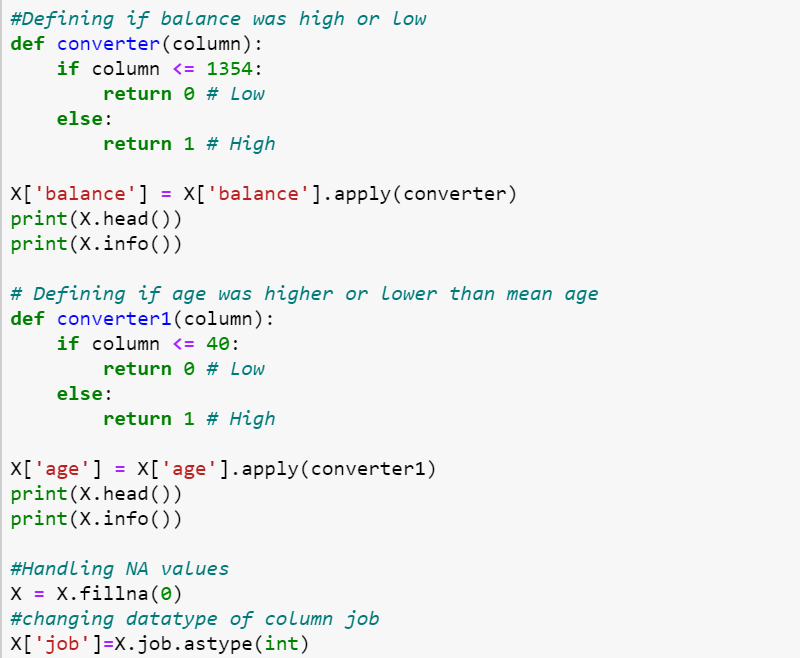
1. All categorical columns are first converted into Numerical Features.



1. In the correlation heat map, no high correlation could be found between the features.



1. Columns ‘balance’ and ‘age’ were converted to classified features as per their mean values. Also, since ‘job’ was showing as float column, we had to convert it into int.



1. Dataset is divided into four subsets as follows:

# Dividing data into subsets

#Personal + housing Data

subset1 = X[['housing','education','balance']]

#Loan related Data

subset2 = X[['default','personal','term']]

#Exclusive personal data

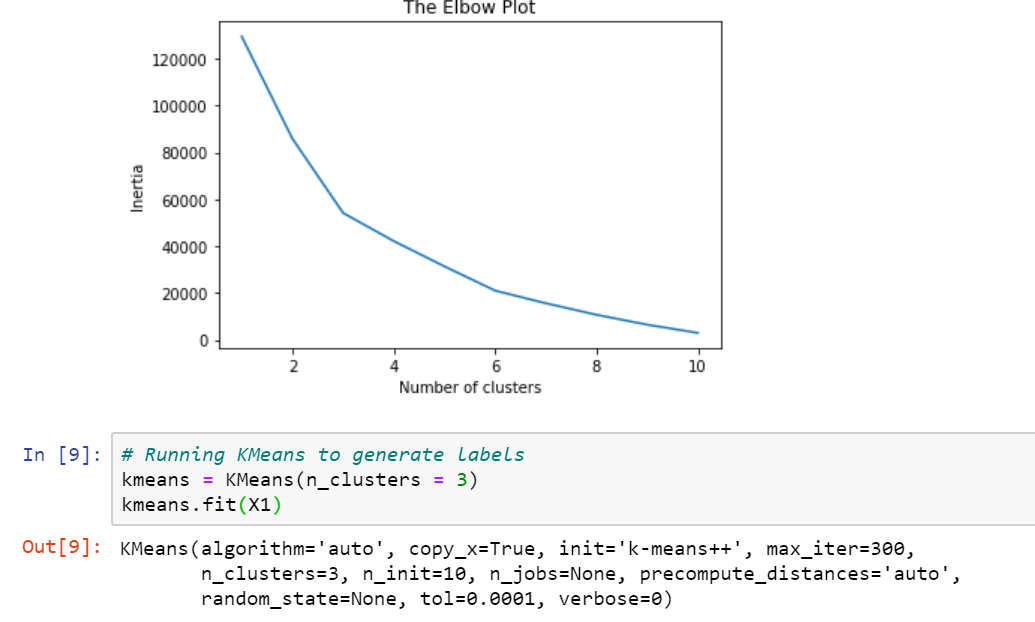
subset3= X[['age','job','marital','education']]

# Financial Data

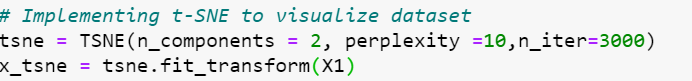
subset4=X[['default','balance','housing','personal','term']]

The subsets were made after much trials to get the best visualizations and then finally these 4 were selected. The first subset essentially consists of Personal data (education, bank balance) of an individual along with a portion of loan data (housing loan y/n). The second subset consists of purely finance related data like loan and credit data (default made y/n, personal loan y/n, term y/n). The 3rd subset has been tried to be created based on exclusive personal data like age, job, marital status and educational levels of a client. And 4th subset consist of Financial Data like default, balance, housing, personal, term.

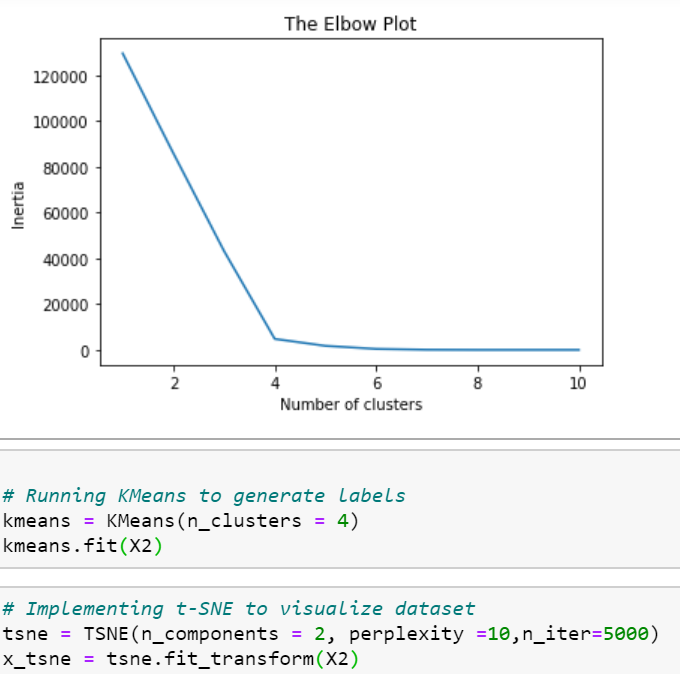
1. The elbow plot for the first subset showed a clear kink at k=3 and we got best visualization for the same, hence k=3 was finalized for the first subset.

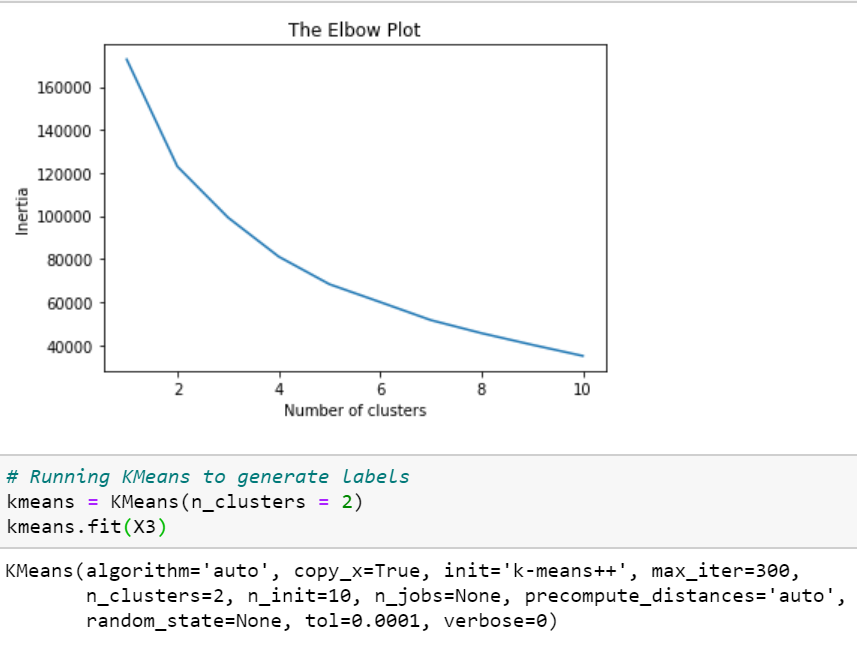
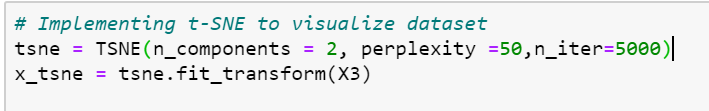


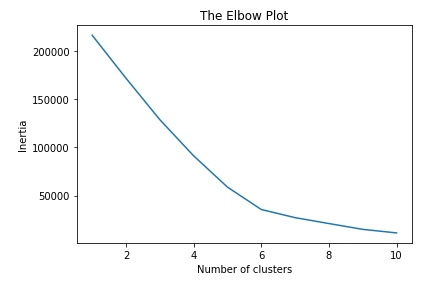
1. Best visualization for subset 1 was obtained for a relatively low perplexity value (since the number of dimensions are not that high) and a high iteration. Perplexity values were tried starting with perplexity=50 and then taking it down to perplexity=10. Iterations were tried starting from 1000 and trying up to 5000 and we got the best results for perplexity=10 and iterations=3000.

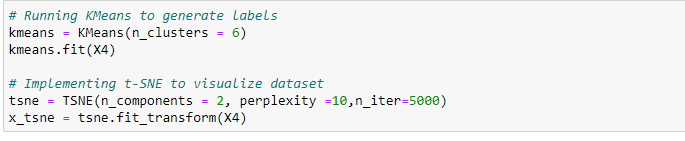


1. For subset 2, the elbow plot showed kink at 4 hence K=4 is taken. Also, perplexity=10 and iteration=5000 provided the best results.



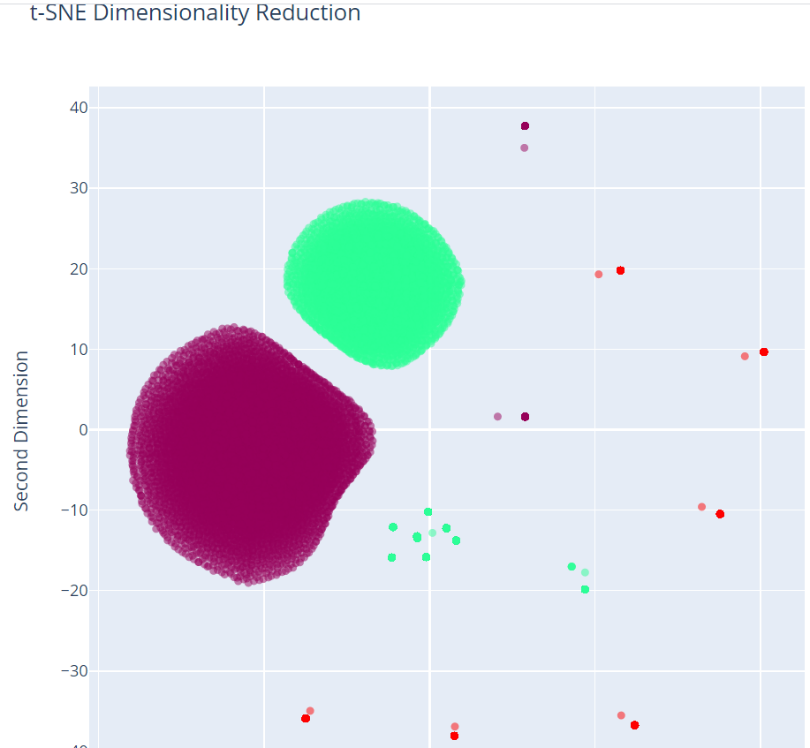
1. Similarly, for subset 3, K=2 was taken and perplexity =50, n\_iter=5000 provided the best visuals. 
2. The elbow plot for the fourth subset showed a clear kink at k=6 and we got best visualization for the same, hence k=6 was finalized for the fourth subset.



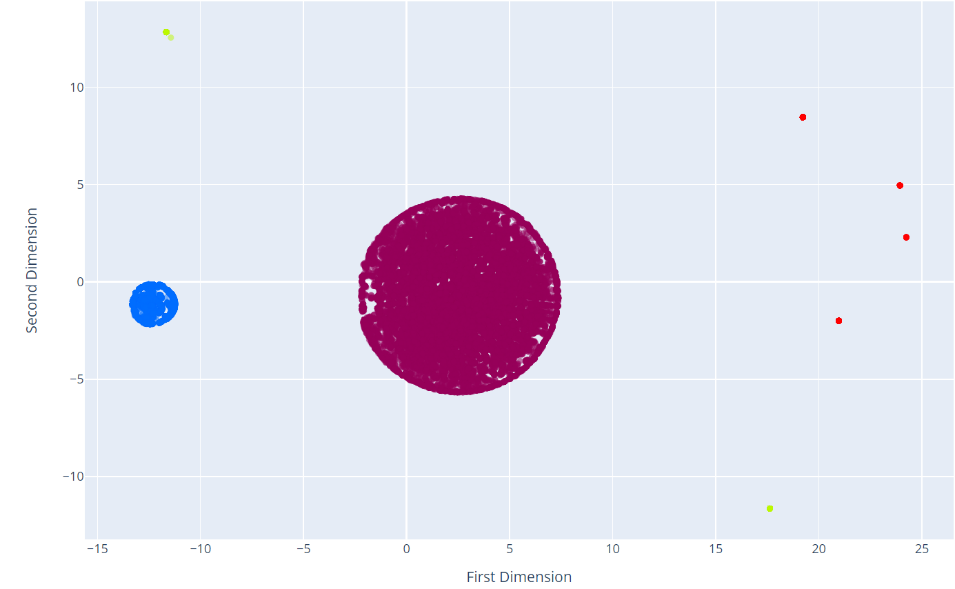
****

1. The screen clippings for the visuals obtained from the 4 subsets are attached herewith and are further explained in details in the ppt’s .

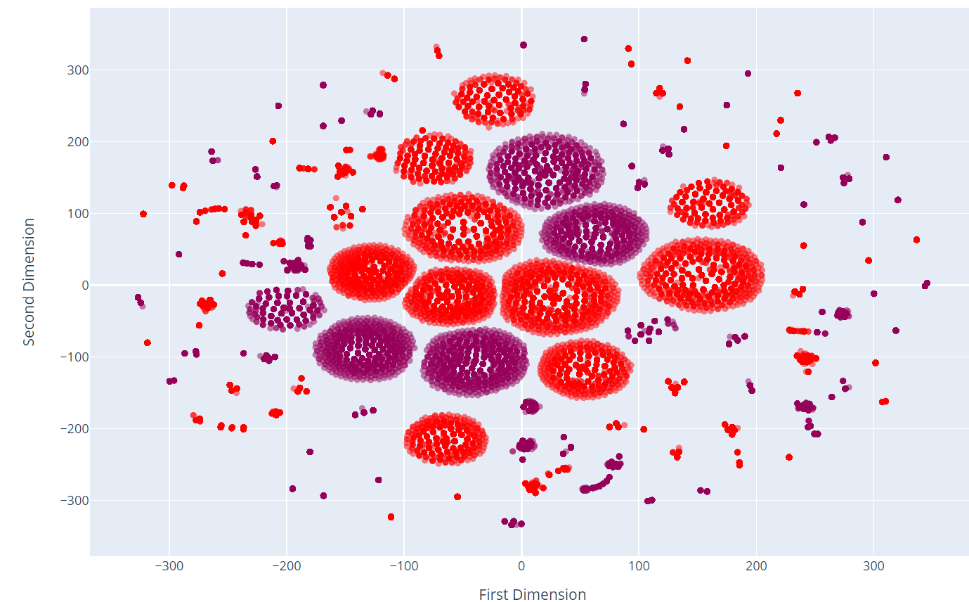
Subset 1:



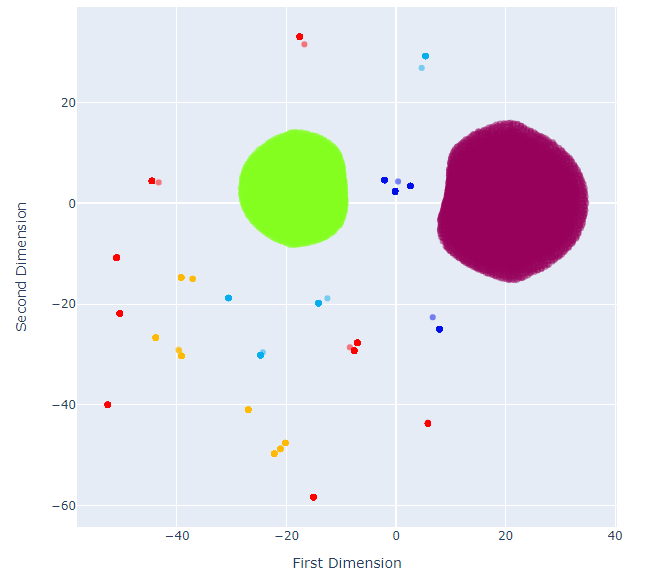
Subset 2:



Subset 3:



Subset 4:



**Visualisation on Patches Dataset**

**Description**

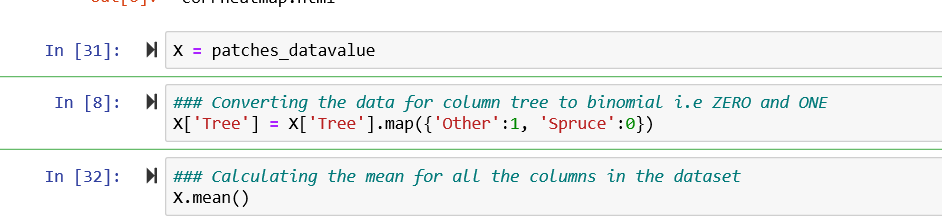
Information for forest of Canada and Alberto describing the categorial data for observations made over short patches of 30 X 30m.

Total Shape of the dataset is with 7 attributes and 15,120 samples.

**Attributes summary:**

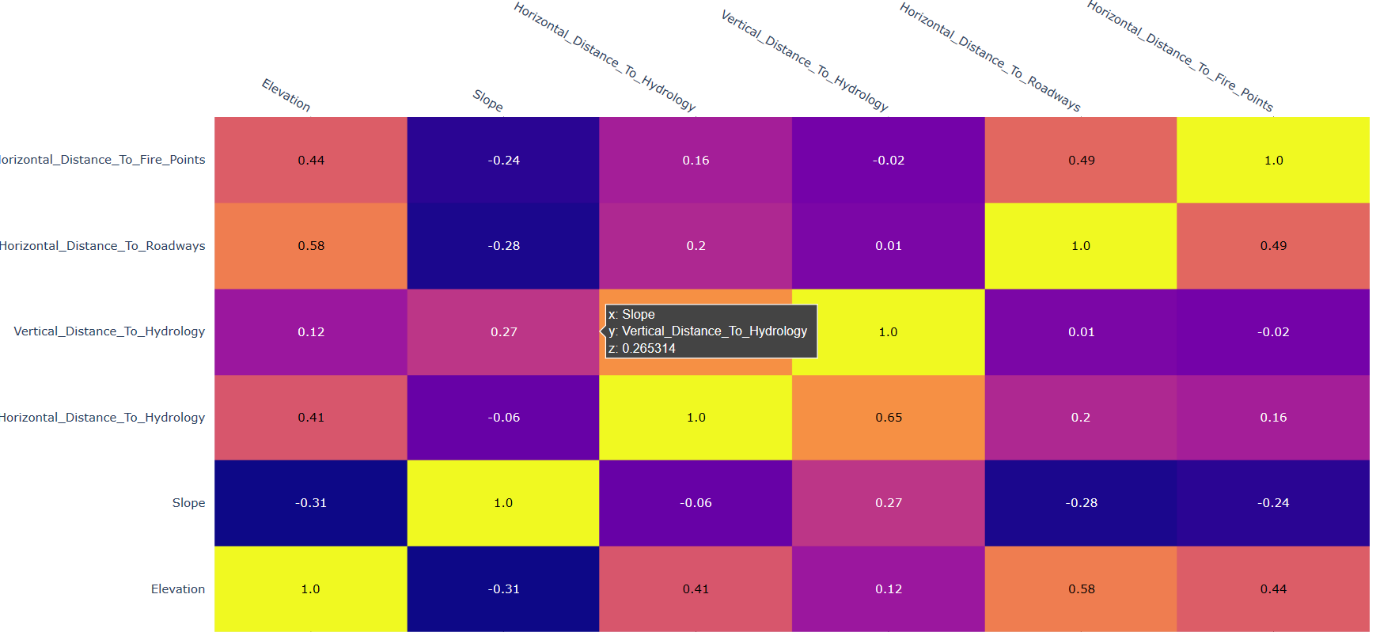
1. Elevation – Elevation calculated in meters.
2. Slope – Measured in degree
3. Horizontal\_Distance\_To\_Hydrology – Distance to nearest surface water.
4. Vertical\_Distance\_To\_Hydrology – Vertical distance to nearest surface water.
5. Horizontal\_Distance\_To\_Roadways- Horizontal distance to nearest roadway.
6. Horizontal\_Distance\_To\_Fire\_Points – Nearest fire station
7. Tree – Kind of tree observed in the particular patch. Either “Spruce” or “Other”
8. All the categorial variable were converted into Numerical.

Here only Column: Tree was categorial which we mapped to 0 and 1 for the two categories.



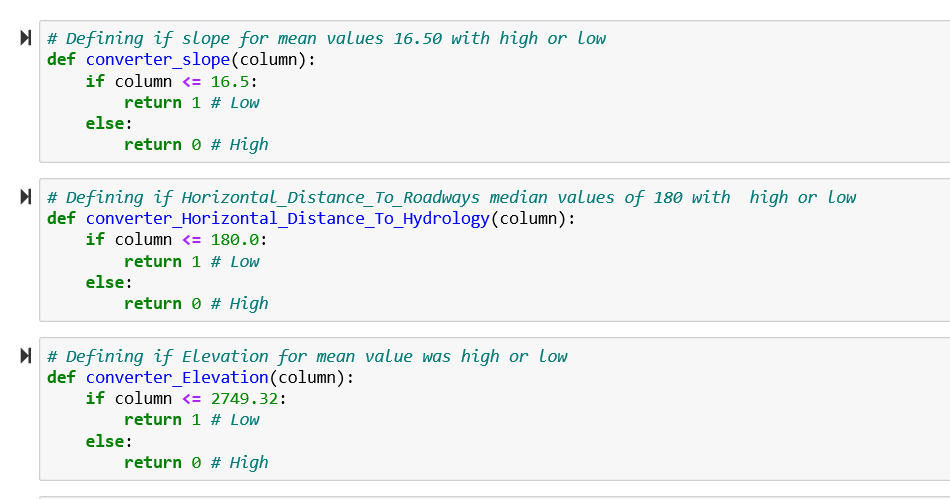
1. Co-relation heat map for Patches database.

Few strong correlations were seen between certain columns as per the heat map, but since the causations were not clear , no column was decided to be dropped.



1. Columns were converted in accordance with the mean and median depending on

the data and this were converted using the functions.



**IMPLEMENTATION OF T-SNE FOR CLUSTER FORMATION**

1. Formation of subsets:

Depending on the attributes in the dataset, creating subsets was crucial for getting proper clusters by find the accurate combinations. Around 7 subsets were created using the attributes from the dataset. However clear and proper clusters were not formed. The final set for the subset combination formed and described below with the detailed analysis.

**SUBSET 1:**

Subset = [Slope, Horizontal\_Distance\_To\_Hydrology, Vertical\_Distance\_To\_Hydrology, Horizontal\_Distance\_To\_Roadways]

Slope column from the dataset was used for bifurcating the data into two sets using the mean of the column. Based on the mean value i.e. 16.50 data was divided mean value greater than 16.50 as 1 and less then mean values as 0. Using this set of values subset was created with the remaining columns. The results obtained from this is better and separated the data into two clusters vary cleanly. It is showing approximately 2 percent of disturbance. Fig. 1 shows the cluster formation using slope as mean. Using K-means total number of clusters used was 2.

For TSNE, Perplexity used was 50 and iteration carried out was around 2000, Random state = 1.

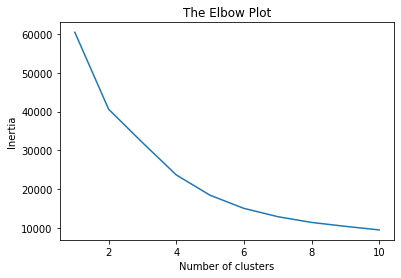


Fig.1.1 Number of clusters using k-means for subset 1

Result for Subset 1:

Fig.1.2 describes the result for scatter plot for the subset using two clusters.

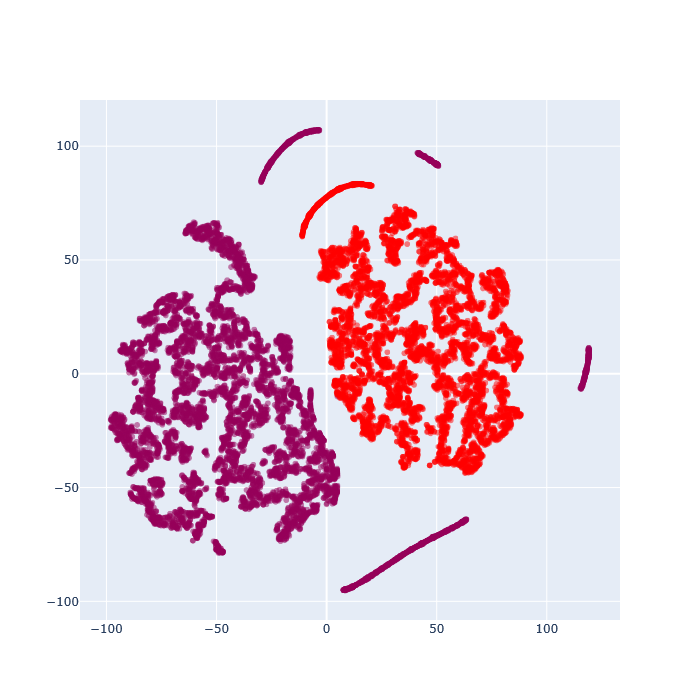


Fig.1.2 Scatter plot for Subset 1

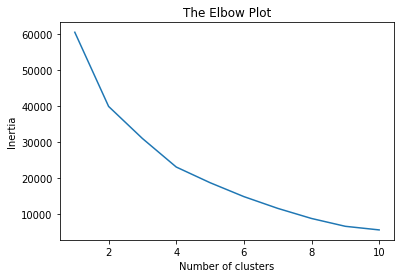
**SUBSET 2**

Using the median of the column **Horizontal\_Distance\_To\_Hydrology** from the dataset median was calculated around 180 and converted the data above 180 as median to 1 and less then median value to 0. The columns used for subset were

**SUBSET** = X[['Slope’, ‘Elevation', 'Horizontal\_Distance\_To\_Hydrology', 'Vertical\_Distance\_To\_Hydrology']].

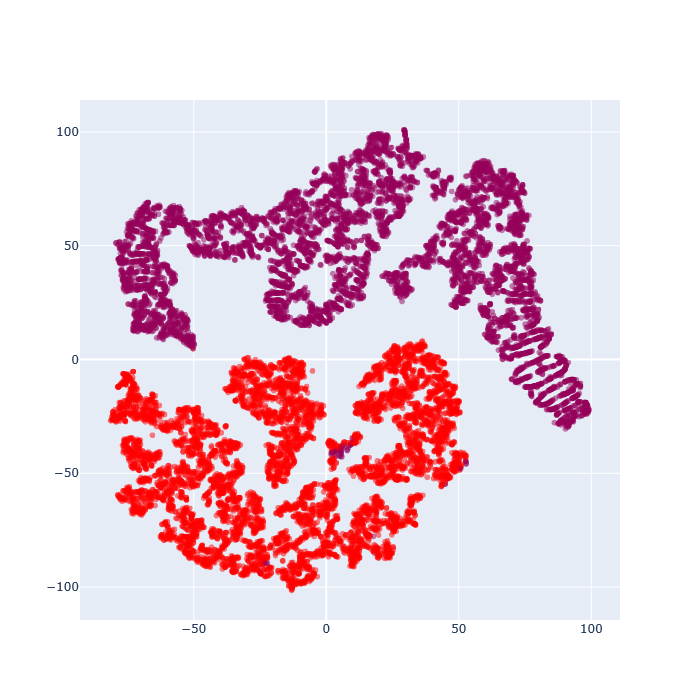
Using the standard scalar, the subset features were standardized. For the transformed data was used for finding the number of clusters to be used. Using the K-means we calculated the no of clusters for this subset. And it turned around to be 2. Using the TSNE, with the below parameter scatter plot was plotted with fine clear distribution. Fig 1.2 show the plot with the 2 number of clusters.

For TSNE, Perplexity used was 50 and iteration carried out was around 2000, Random state = 1.

****

**Fig.1.3 Elbow plot for subset 2**

**Scatter plot for Subset 2**



**Fig.1.4 Scatter plot for Subset 2**

**SUBSET 3**

Using the column Elevation for taking the mean and using the converter function to replace the values with mean greater than 2749.32 as 1 and less the mean value as 0. Furthermore, based on this dataset, below subset was created for plotting the clusters.

**SUBSET:**

Subset=X['Slope','Elevation','Tree','Horizontal\_Distance\_To\_Hydrology', 'Vertical\_Distance\_To\_Hydrology','Horizontal\_Distance\_To\_Roadways']].

For the above subset the k-mean calculated with total 3 clusters. Using the clusters value as 3 the TSNE was performed. As generated results with 3 different and clear dataset clusters. Since the elbow plot with k-means has been 4 but using number of clusters as 4 was not giving clear scatter plot so with total 3 number cluster more precise scatter plot was created.

For TSNE, Perplexity used was 50 and iteration carried out was around 2000, Random state = 1. Fig.1.6 displays the result of the scatter plot.

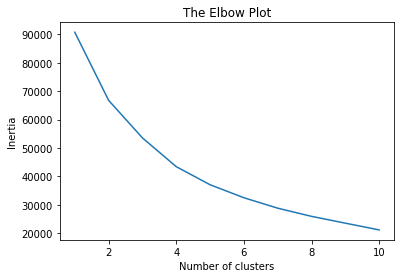


Fig.1.5 Elbow plot for subset 2

**Results for Subset 3:**

Fig.1.6 describes the scatter plot for subset 3 with total 3 clusters which are better and can classify the data properly.

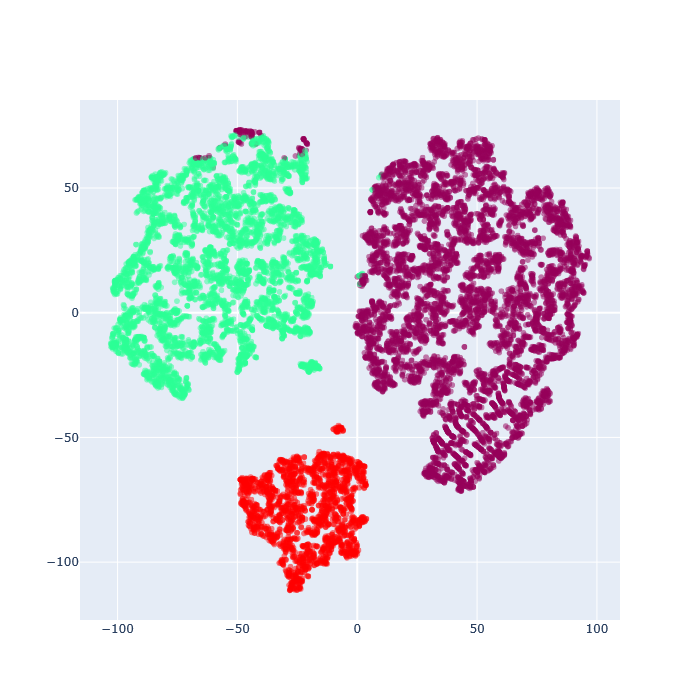


Fig.1.6 Scatter plot for subset 3 with total no of cluster 3

**Individual Contributions:**

**Himanshu Kumar: Student Number 10525006:**

**Akash Shivaji Kadam: Student Number 10523055**

Worked on the Patches dataset, Implemented the coding part Patches this has been attached to the assignment copy. Details analysis for dataset and decided the number of subset and tried other various combination for getting the result. Total final three subset where generated and scatter plot was plotted. Also, the report section for patches dataset was created.

**Gaurav Pradeep Taori: Student Number 10525875**

Worked on the Clients dataset, contributed in the pre-processing part along with Project members. Implemented the subset 4 in the clients dataset and tried various combination for getting results in the form of plots attached in the report. Also contributed as team member in the project.