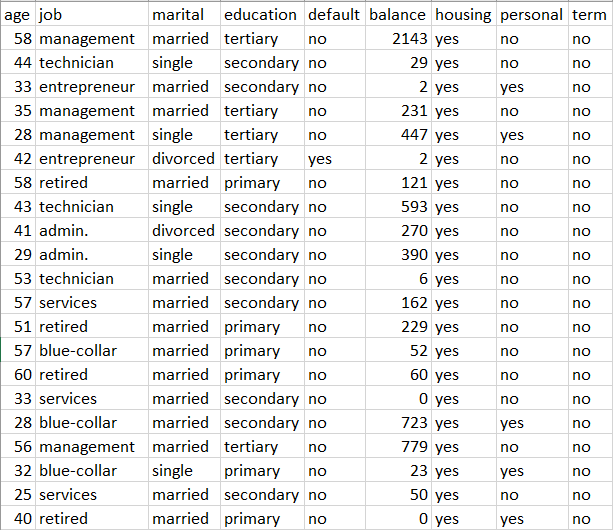
|  |  |
| --- | --- |
| Data Visualisation  Report - CA02 | Aswin Ramdas (10534162) Karan Perla (10539155) Usman Hameed (10534421) Akshay Sharma (10536542)  M.Sc. Data Analytics |

**Analysis and Visualisation of “Clients” Data Set**

**Data Set:**

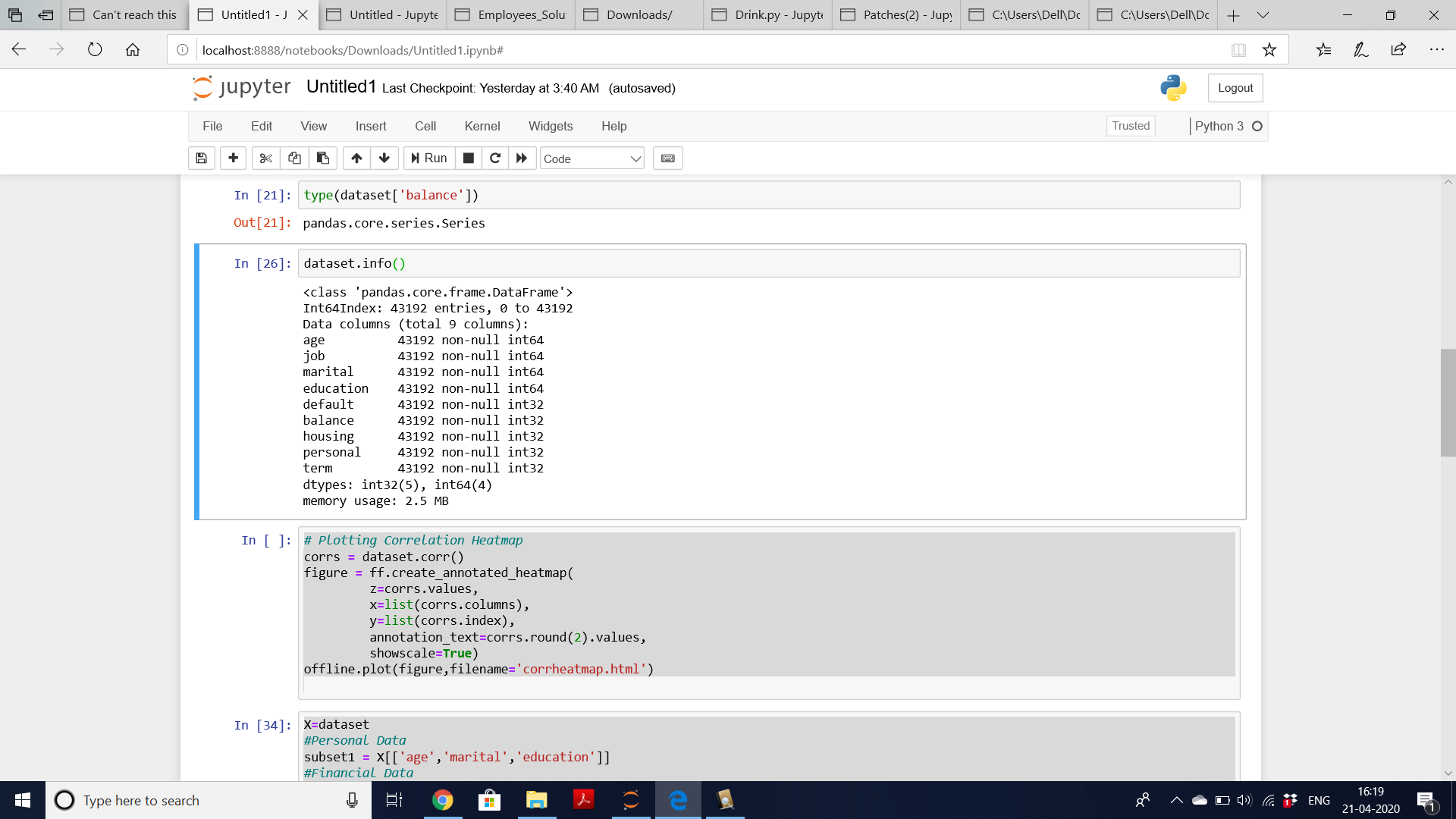
Clients data set has 43193 instances and 9 variables. This data set is all about client’s information and each row defines the client. The data set contains nine columns each of this giving us the information about the client’s different status which includes age, job, marital, education, default, balance, housing, personal and term.



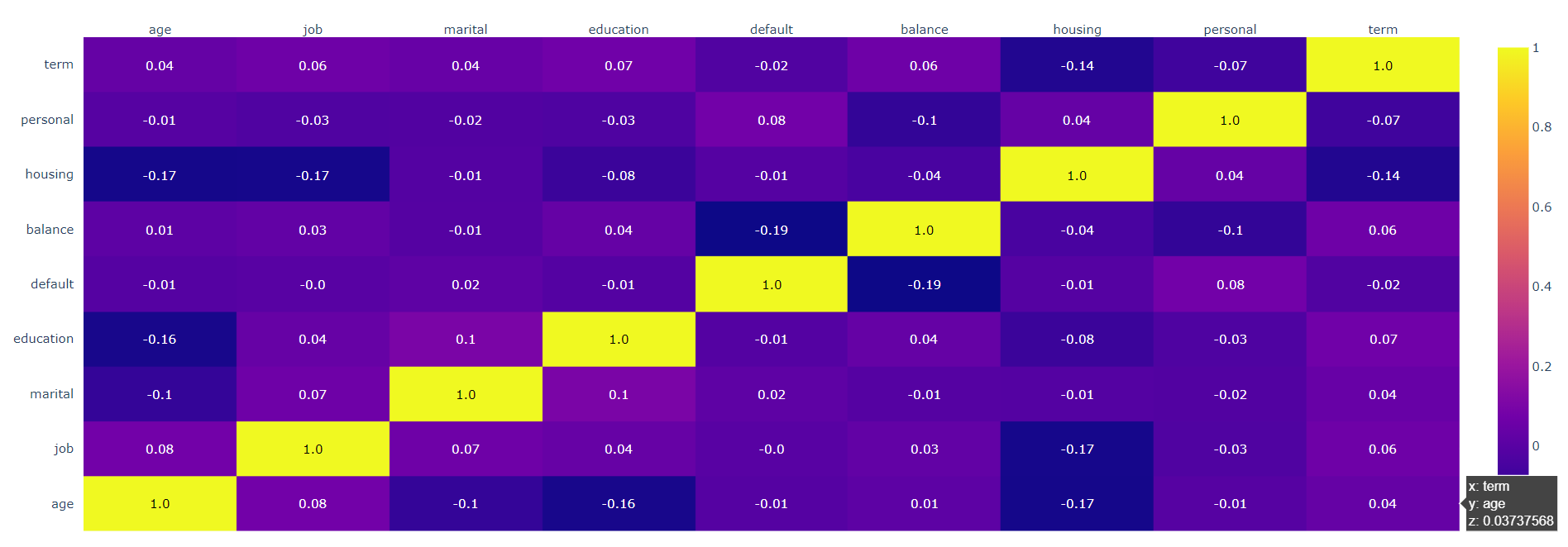
The analysis was done on this dataset and visualisations are created on python.

Relationship between variables

We had binary values and categorical values, we converted them into numerical values. We have more two columns there that were continues values in age and balance which we converted into numbers for the sake of clustering algorithm.



We used Label Encoder for converting categorical data into numeric data. We can plot the correlation matrix using this data. Correlation Plot is shown below



We now find relationship between some variables using correlation plot.

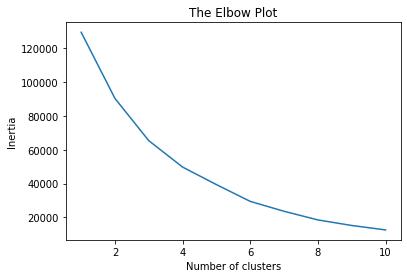
* The variable “personal” is correlated with “default”.
* The variable “housing” is negatively correlated with “education”.

Since we can see that none of the variables are highly correlated, we are taking all the variables in consideration.

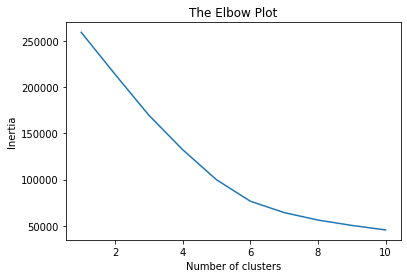
# K-means Clustering:

K-means clustering technique is used to visualise groupings of clients in our data set. As we all know, K-means is an algorithm of unsupervised machine learning. It is used for creating clusters to observe the relationships. Features in the data set have been normalised for implementing K-means clustering. We used Standard Scalar function for the same. Standard Scalar function was used to transform all the columns in the data set so that the distribution will have standard deviation as “1” and mean value as “0”. Elbow method was used to find the clusters in the dataset. We have created three subsets in the Clients Data set. Elbow plots have been created for all the 3 subsets.

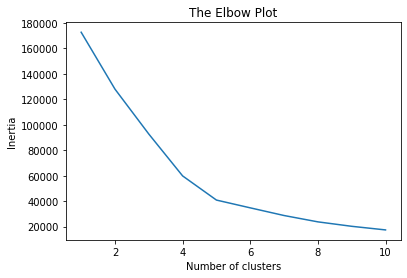
Elbow plot of 1st Subset:



Elbow plot of 2nd Subset:



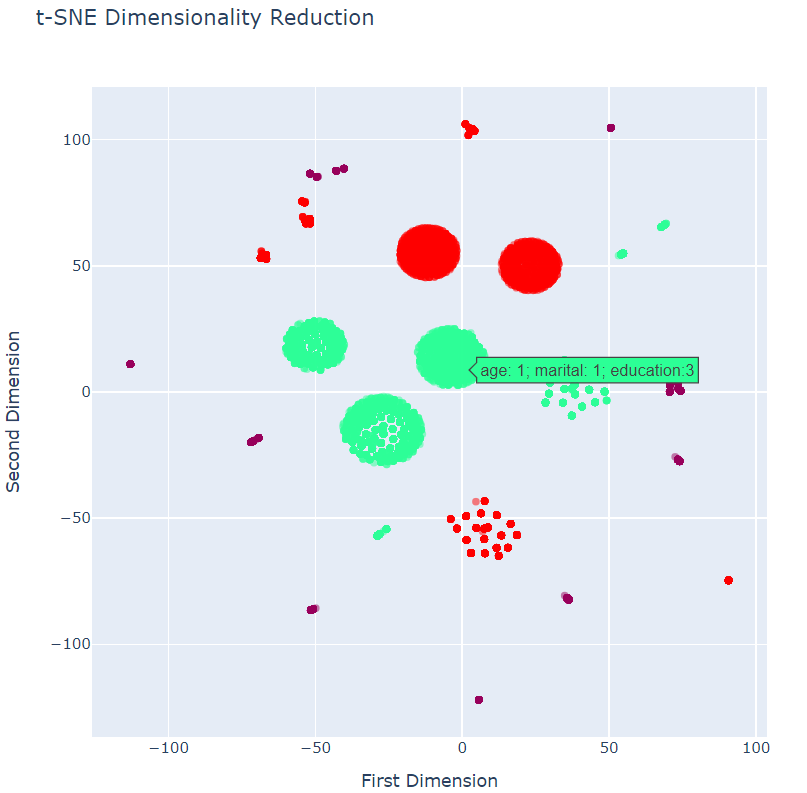
Elbow plot of 3rd Subset:



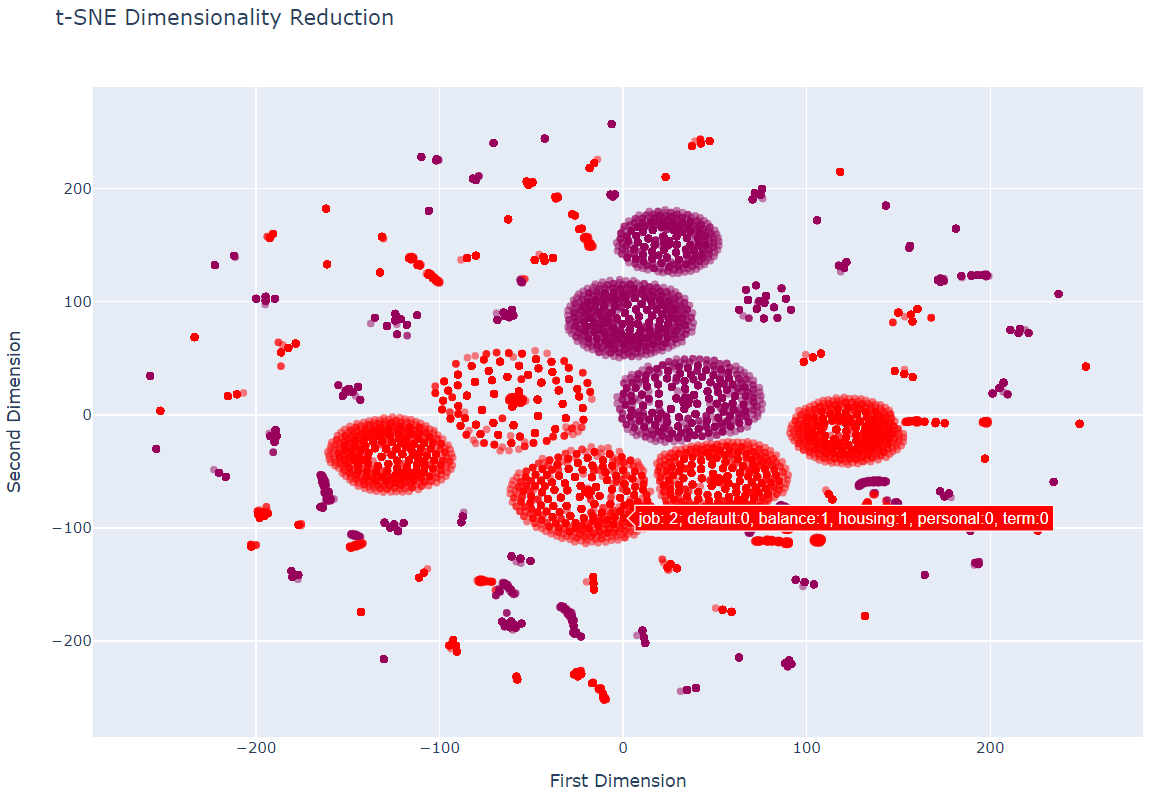
The number of clusters in our data can be determined by the location of the elbow bend. As we can see from the figure, the bend is at value 6. Hence, our ideal clusters should be 6.

# t-SNE (t-Distributed Stochastic Neighbour Embedding):

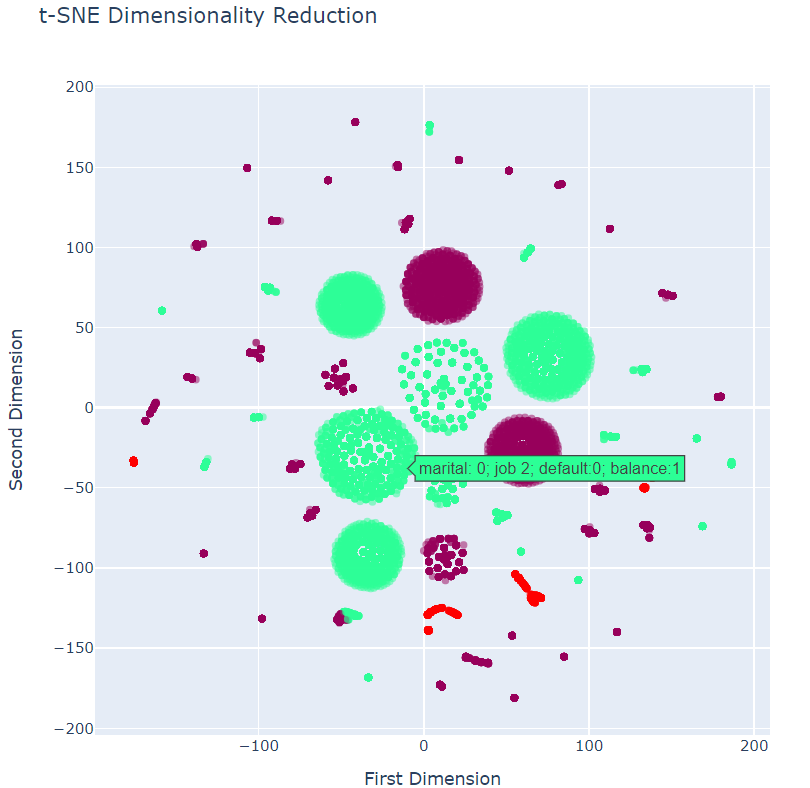
For visualising high dimensional data sets, we use t-SNE. The plot has been attached below:



This clustering analysis shows us the cluster formed on basis age, marital status and education level of the client, which helps us to understand that most of the big clusters of this bank are highly educated. Most of the clients of this bank are in the age group of 20-40. Most of the clients of this bank are not married.



This analysis shows the clusters formed according to financial data of the clients. The cluster are made on the basis of job, default, balance, housing, personal and term. The biggest clusters that were formed shows that most of the clients of this bank have taken only housing loan and are relatively financially secure. Most form of debt to the bank awed by the clients is only in terms of housing loans. Those who are in the middle class of the income range are the most frequent clients of this bank, namely, having 0-20k of average yearly income in euros. Those who hold stable jobs are most likely to have a clear debt record. Those who hold high paying jobs are most likely to have housing loans. Those who have blue-collar jobs are most likely to have credit in default.

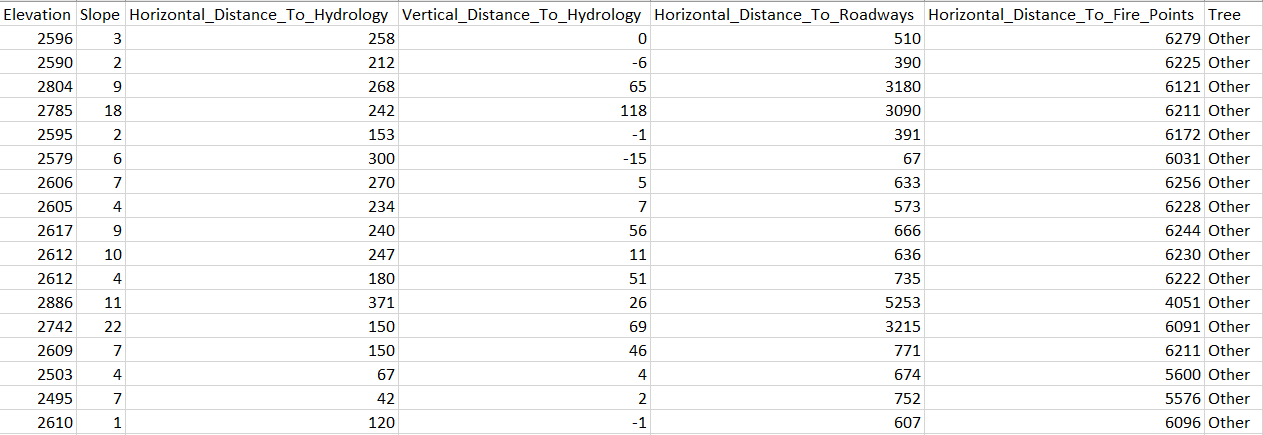


This analysis shows the churn factors used in this dataset namely, marital status, job, default, balance. The observations here are that most of the clients of this bank tend to not have credit in default. Very few clients have more than 50k average yearly balance

**Analysis and Visualisation of Patches Data Set**

**Data Set:**

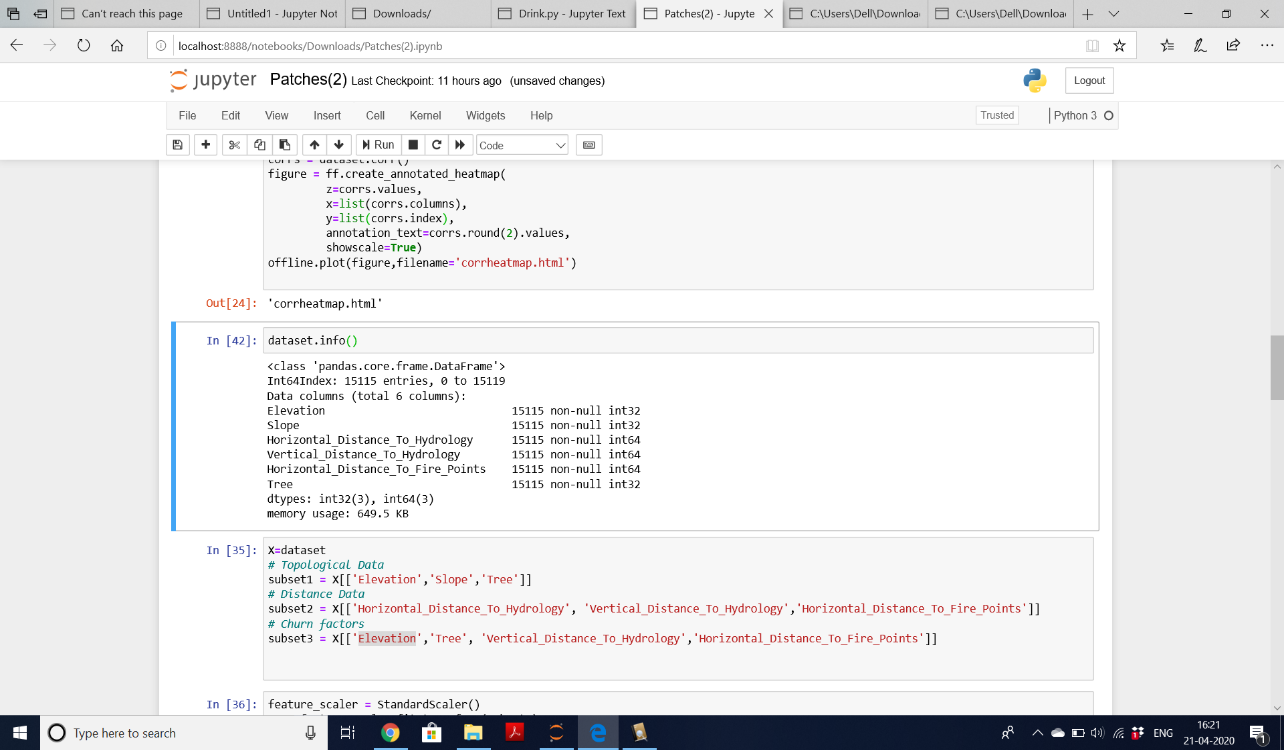
Patches data set consists of 15120 instances and 7 variables. This dataset is all about cartographic information that was made on different patches in the forest of Alberta, Canada. It contains 7 columns each one exploring different dimension categories that are Elevation, Slope, Horizontal Distance to Hydrology, Vertical distance to Hydrology, Horizontal Distance to Roadways, Horizontal distance to Fire points and Tree.



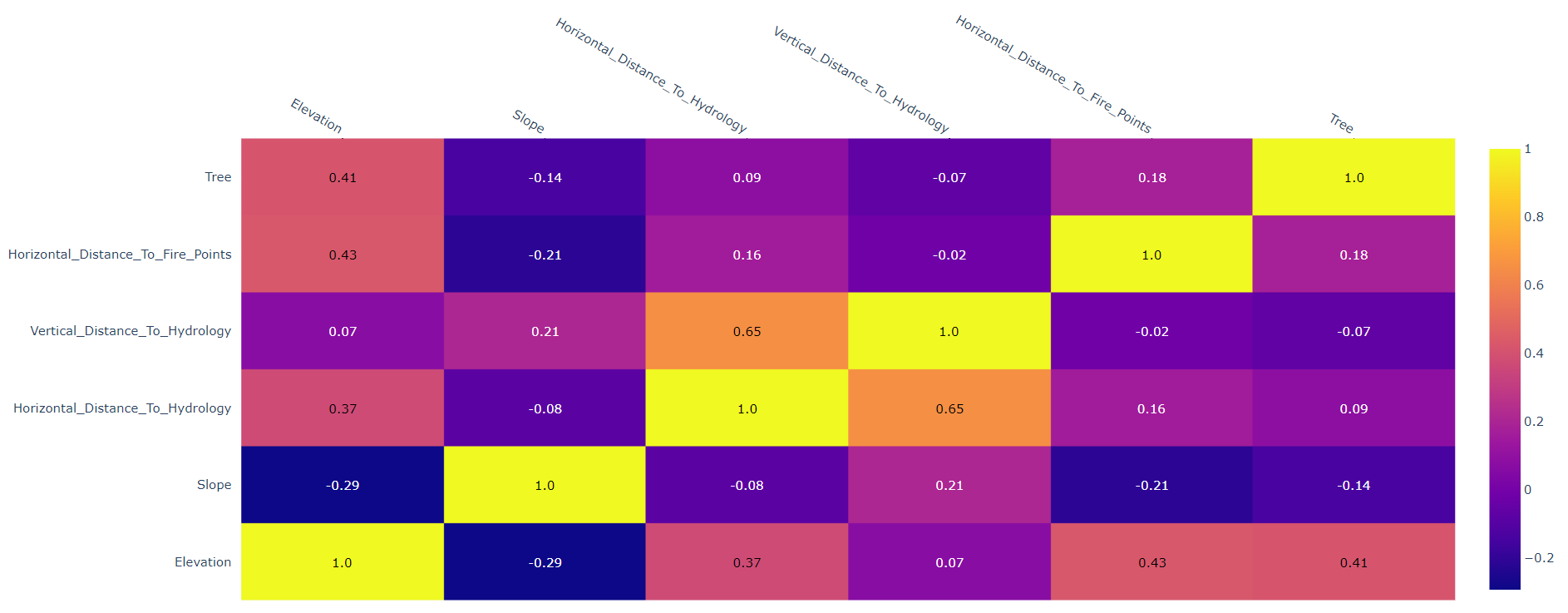
The analysis was done on this dataset and visualisations are created on python.

Relationship between Variables

We had binary values and categorical values, we converted them into numerical values. We have more two columns there that were continues values in age and balance which we converted into numbers for the sake of clustering algorithm.



We used Label Encoder for converting categorical data into numeric data. We can plot the correlation matrix using this data. Correlation Plot is shown below



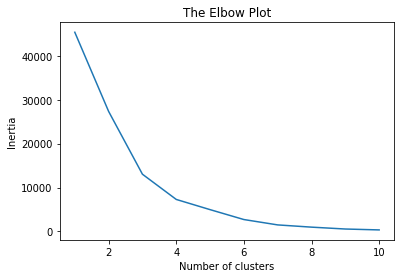
We now find relationship between some variables using correlation plot.

* The variable “vertical\_distance\_to\_hydrology” is highly correlated with “horizontal\_distance\_to\_hydrology”.
* The variable “horizontal\_distance\_to\_hydrology” is highly correlated with “vertical\_distance\_to\_hydrology”.
* There is negative correlation between “slope” and “elevation”.

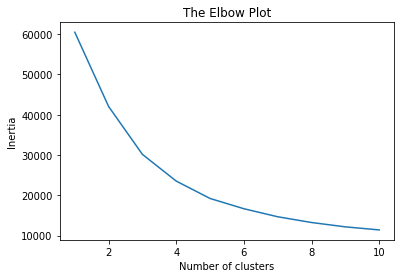
# K-means Clustering:

K-means clustering technique is used to visualise groupings of clients in our data set. As we all know, K-means is an algorithm of unsupervised machine learning. It is used for creating clusters to observe the relationships. Features in the data set have been normalised for implementing K-means clustering. We used Standard Scalar function for the same. Standard Scalar function was used to transform all the columns in the data set so that the distribution will have standard deviation as “1” and mean value as “0”. Elbow method was used to find the clusters in the dataset. We have created two subsets in the Patches Data set. Elbow plots have been created for both the subsets.

Elbow Plot of 1st Subset:



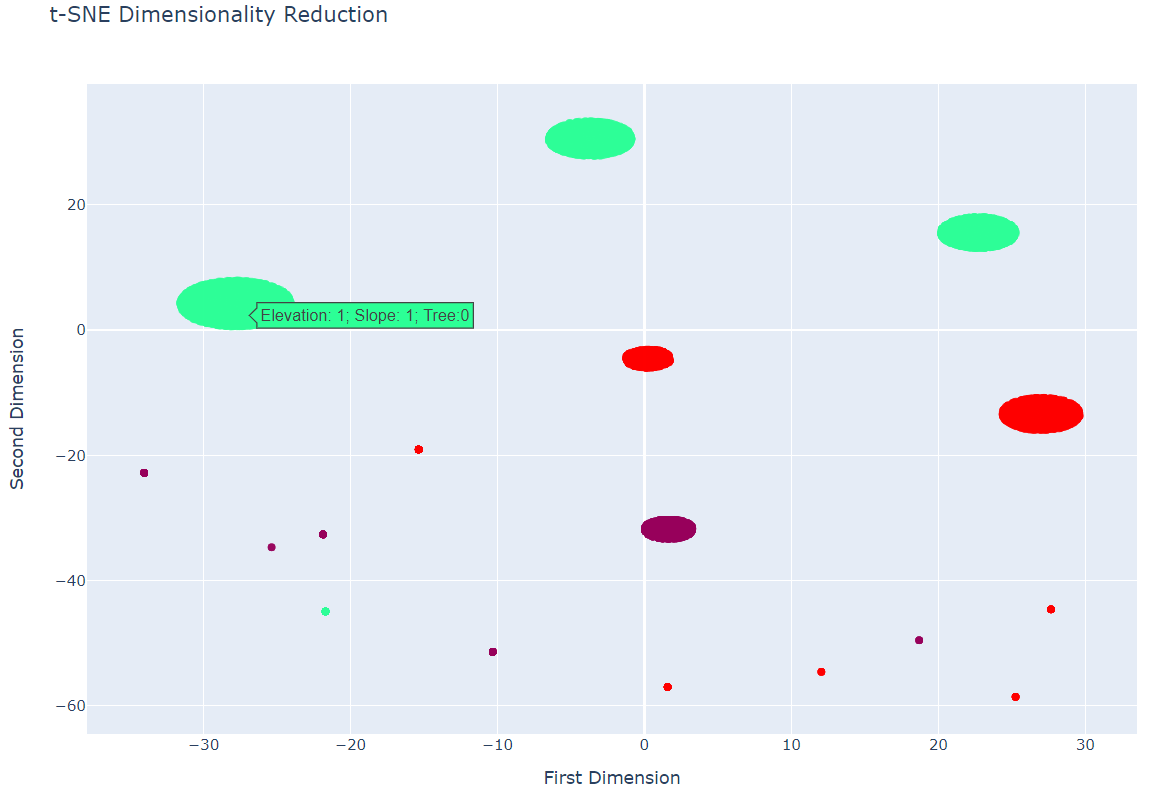
Elbow Plot of 2nd Subset:



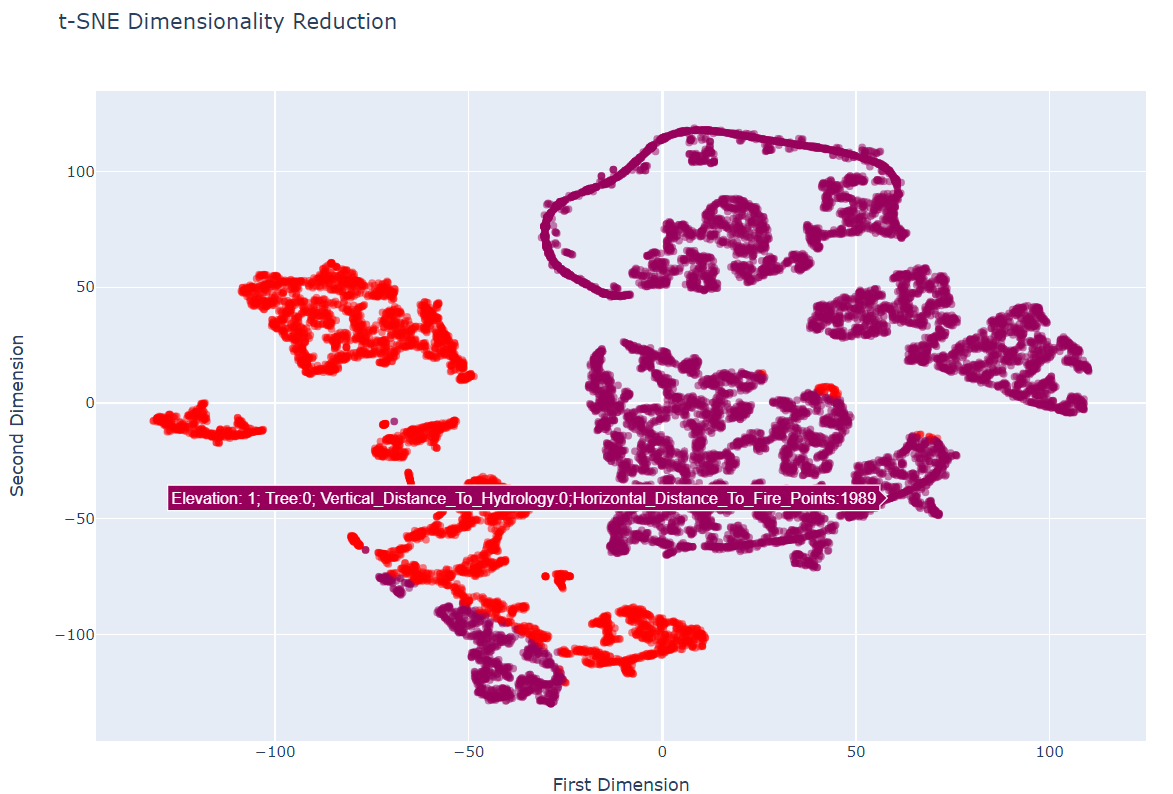
The number of clusters in our data can be determined by the location of the elbow bend. As we can see from the figure, the bend starts at 2, then there is a bend at 4. Hence, our ideal clusters should be 4.

# t-SNE (t-Distributed Stochastic Neighbour Embedding):

For visualising high dimensional data sets, we use t-SNE. The plot has been attached below:



This analysis shows us the clusters formed on the basis of Elevation, slope and Tree. We can infer from this analysis that the spruce trees are present on elevated land. The slope tends to be less in areas of high density of Spruce trees.Slope seen to be steepest when elevation is at one and tends to be more regular when at higher elevation



Observations from this t-SNE analysis:

Spruce trees only occur at high elevation.

Spruce trees seems to be far away from ignition points, suggesting that it may not burn easily.

wildfire generally did not happen at high elevation in those Canadian forest in Canada. This might be probably due to cold temperatures.

**Contribution for this assignment**:

All the group members actively participated in all the phases with each part being dedicatedly allocated to each person. The specific contributions are mentioned below:

1. Akshay Sharma (10536542) – Preparing the data for clients dataset and creating the presentation with audio

2. Karan Perla (10539155) – K-means clustering and t-SNE for clients dataset and creating the report

3. Usman Hameed (10534421) – Plotting the correlation matrix for both the datasets

4. Ashwin Ramdas (10534162) – Preparing the data for patches dataset and Implementing t-SNE plots for Patches dataset.