Data Visualization

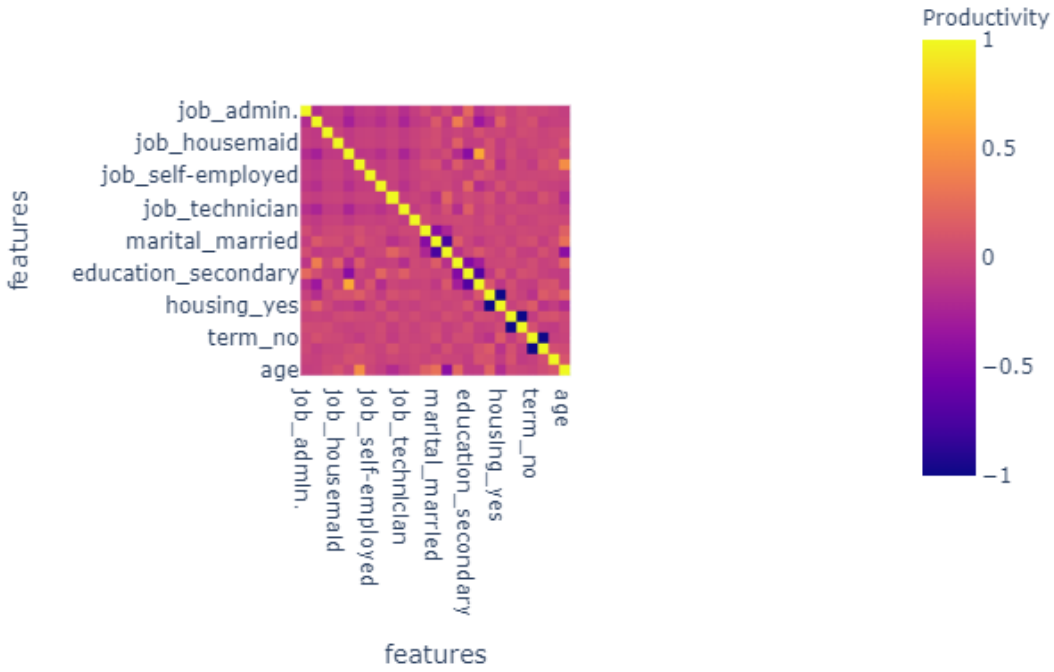
Professor: Kunwar Madan

Presented by:

Sidharth N Koparde (10521114)

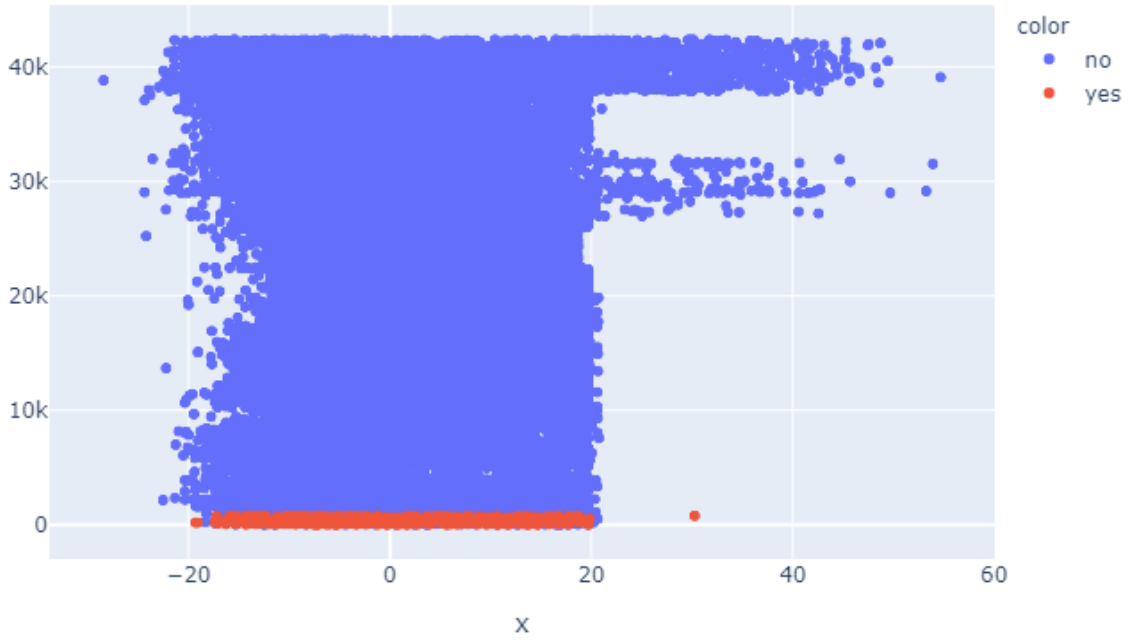
This is a report of the visualization task, this task, not only discusses two datasets, but also optimizes the number of clusters. This makes the clusters different, with each cluster highlighting on a unique aspect. Moreover, it can be said that different set of clusters are there and all of these tend to lead to some variation. The two datasets that are in process, indicate the potential variation of the variables. The first set of data is about clients, which studies different features of the clients, this dataset is derived to predict which clients would tend to default, and where the probability of default is less. Moreover, the second set of data indicates that the patches in Canada. These patches tend to predict the variation and also illustrate the potential of a dataset that predicts whether a patch is spruce or is it some other. Based on altitude, slope, and different distances from hydrology, roadways and fire points. This illustrates the potential of each variable, or feature. In order to identify each segment, it can be seen that the potential of those points tend to lead to a specific segment.

The first dataset is based on clients, and as stated it identifies the defaulters from the ones who did not default. Below is the correlation matrix in form of a heatmap:



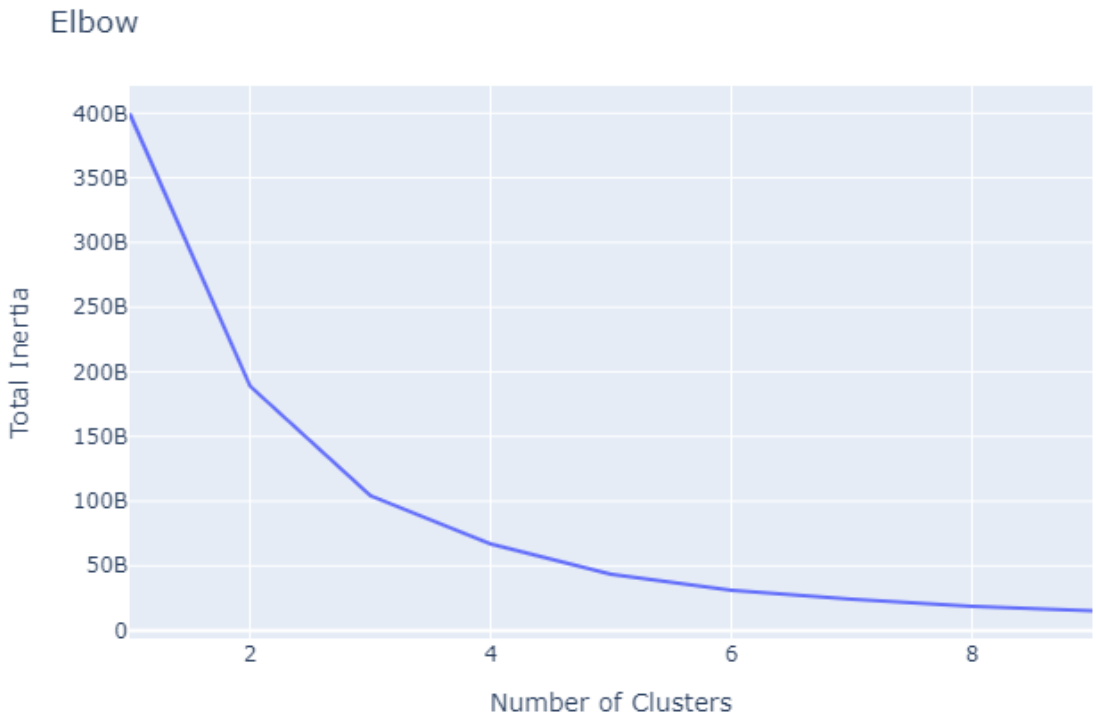
It can be seen that this correlation matrix has more than 30 variables, and the potential of these variables indicate that the relationship between job factors. This also indicates the overall potential of the correlation factor; the relationship dampens as it moves towards age and housing.

The next part of the analysis is the dimensionality reduction and the chart for that is as per PCA, this is the dimensionality reduction algorithm that is to be used here, so the potential of a PCA on the dataset of clients is stated as follows:



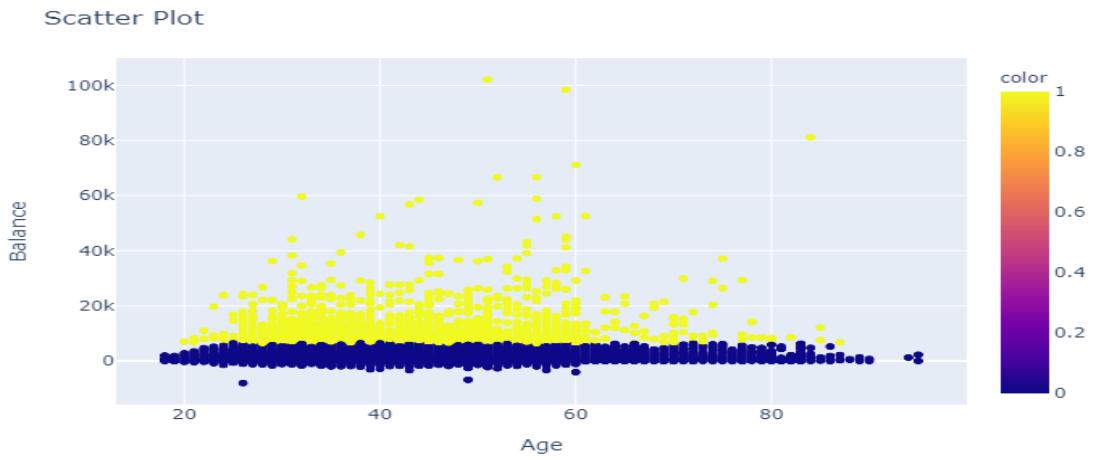
This is a plot, the color patterns indicate the potential expansion on the said domain, and here, for the sake of simplicity, default is colored, the red part indicates that the part defaulters, the rest of the blue is non-defaulters. It can be observed that the potential defaulters are below the point and not above it, therefore, it can be used as a classifier, and tends to be an indicator of the difference between the two classes.

In the first dataset, the clusters are identified, and the number of optimal clusters in the dataset are illustrated. These clusters tend to be optimized in form of an elbow chart. This chart has cluster on x axis, and inertia on y axis. As the clusters increases, the inertia decrease, the optimal decrease is identified and concluded, this is done as follows for the client dataset.



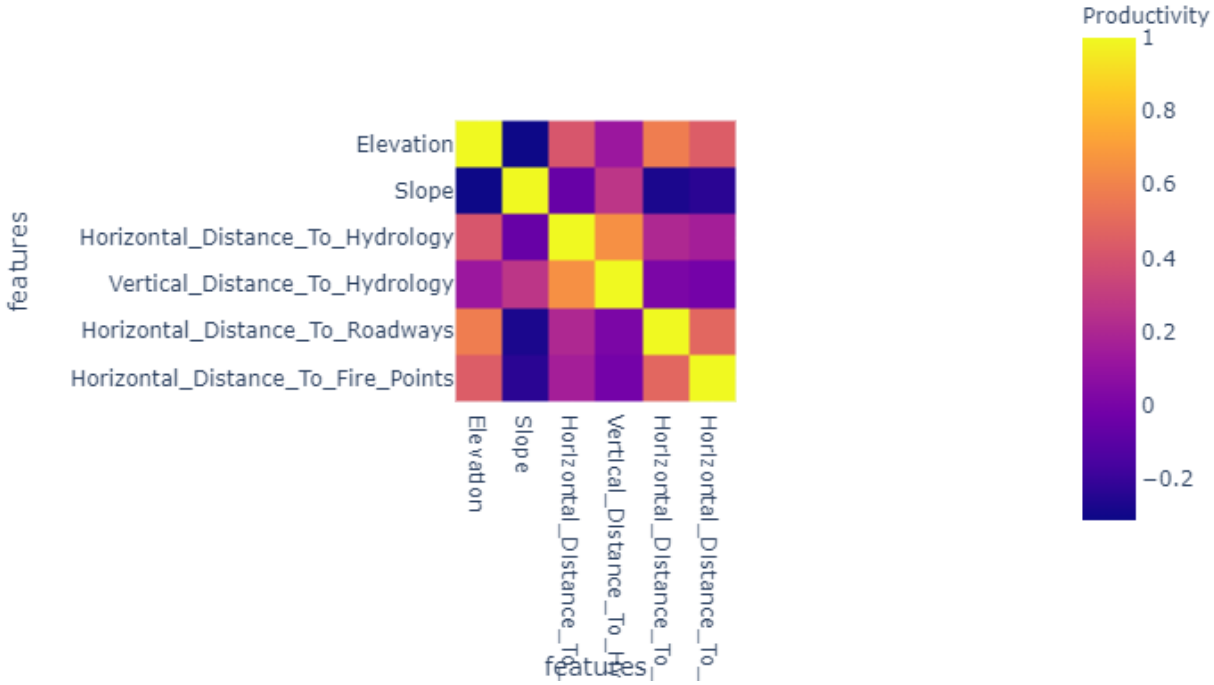
This illustrates that the optimal number of clusters are 2, and this is due to the fact that after 2 clusters. The decrease in inertia is continuous in terms of the number of clusters, therefore, the optimal number of clusters are the point where the elbow chart has most significant turn. Moving on, as stated above, that the point of optimality is not the point of least inertia, as inertia constantly decreases, instead, it is the point where the marginal inertia, or the change in inertia is the most, and the point in that study here is the distribution with 2 clusters.

The change in inertia decreases, in order to move forward, it can be illustrated that in the client's database, the number of clusters should be 2, this can be depicted in the following diagram:



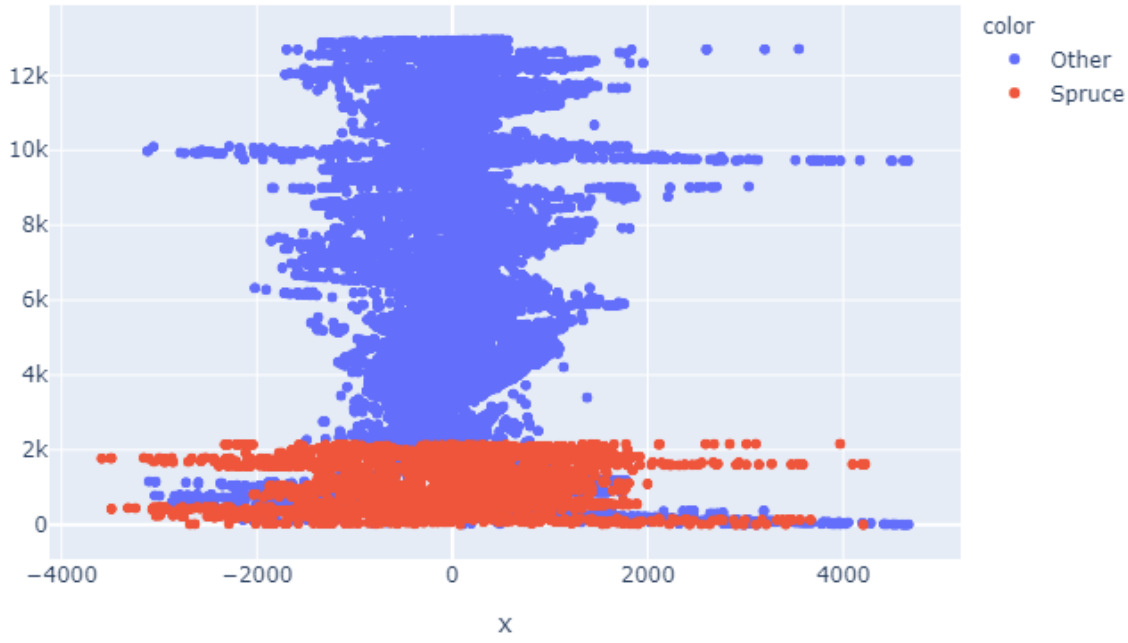
In this diagram, there are two segments, the first segment is that of low balances and the ones with high balances. This is illustrated in terms of the two colors. Though the age does not cause much difference, but balance seems like a key divider. In terms of the variation in clustering, it can be said that the low balance cluster can indicate the cluster that majorly comprises of defaulters. In addition, it can be seen that in the PCA for client's dataset, the lower dataset indicated that the lower variation and low balance are the key features in terms of defaulters. So, in moving that forward, it can be said that the cluster for low balance in all age groups is the cluster that represents the low cluster. Additionally, it can be added the other cluster comprises of the population with more variation and higher balances, so it can be assumed similarly that the higher cluster represents the non-defaulters.

Moving on to the second dataset, it can be said that this too has to be solved with clustering. This has to get started with correlation, and in order to get it started, this needs to increase the potential, so first of all, the correlation is going to be studied, since, all of these variables are quantitative in nature, in terms of the dataset, therefore, it can be said that there is no requirement of dummy variables. Moving forward, the correlation heatmap is as follows:



It can be seen that this has a total of six variables, and the highest correlation is between the horizontal distance to roadways and horizontal distance to fire points. This indicates that these two, when combined can illustrate the clusters that could make clustering possible in this distribution. Moreover, it can be said that this correlation division also illustrates the potential negative correlation the strongest of that indicates that the slope and elevation takes place, but in an overall context, it can be highlighted that the correlation is very weak as compared to the positive correlation.

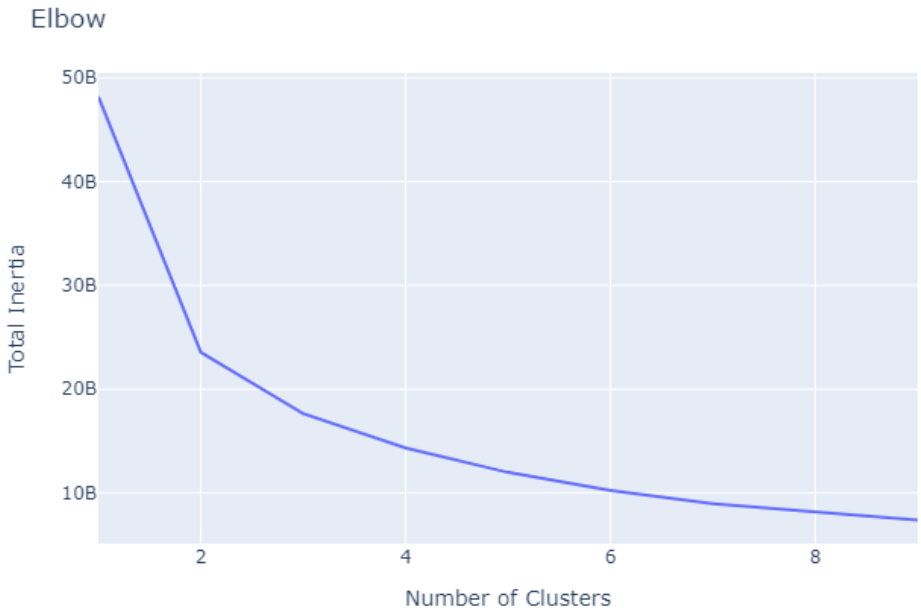
In the second part of the second dataset, there is another algorithm, that not only minimizes the dimensions but also points out the primary difference between these factors. Furthermore, this also points to the visualization that is indicative of the 2 dimensions theory. For this PCA is used and the visualization is done as follows:



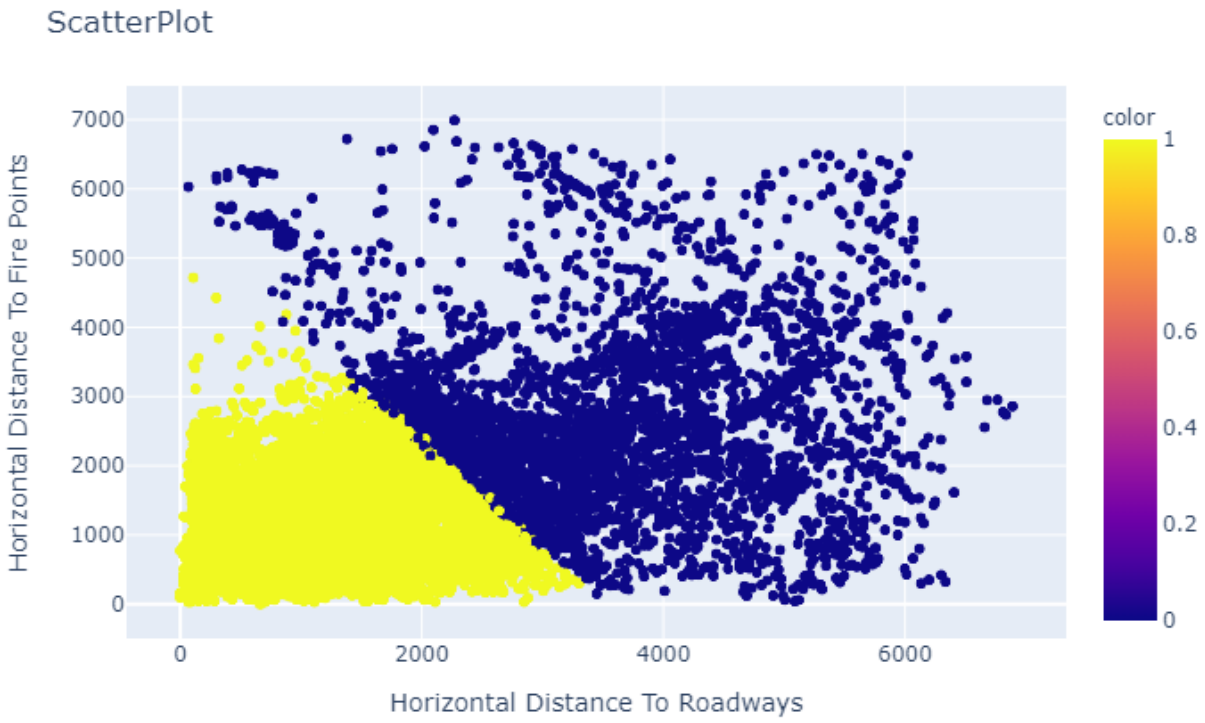
There are two tree types in this dataset, and if one wants to identify each tree type in terms of the overall results, it can be seen that the spruce has the dimensions below 2000, and above 2k, the dimensions indicate that spruce has a higher division, and the other type of trees are closely related. Moreover, it can be said that the variation is also more in other tree types and not in case of the spruce tree.

In order to move of, and on observing the dataset, and its features there were various features observed, as this had features like altitude and slope, along with distances, after numerous and

vigorous testing the clusters were developed in form of the following elbow diagram.



It can be seen that the pivot point is 2, here as well, so it can be said that the optimal number of clusters are two, before 2 the inertia fell drastically, and after 2, the case was reversed. So, for this study the clusters should be 2. Keeping in mind the above results, the following can be the cluster chart for patches, and it would indicate 2 clusters. The decrease in inertia is continuous in terms of the number of clusters, therefore, the optimal number of clusters are the point where the elbow chart has most significant turn. Moving on, as stated above, that the point of optimality is not the point of least inertia, as inertia constantly decreases, instead, it is the point where the marginal inertia, or the change in inertia is the most, and the point in that study here is the distribution with 2 clusters.



In this model, the distance to roadways and the distance to fire points are taken respectively on x and y axis, respectively. It can be seen that there are two segments, one is the segment that is close to both the roads and fire points. It can further be seen that the potential variation also indicates that these choke points can result in a very fine division in the two groups. In order to further analyze these variables, it can be said that the variation is not based on high and low, but, it is based on having lower values in both distances collectively, and in both distances observed in conjunction. This is an indicator of the various results that are studied and indicates a potential for variation that leads to lower values and higher values.

It can be concluded that these segments are identified and in terms of both the datasets, there were two clusters, in the clients, there were those who maintained low balances, and the ones who had considerable balance. In the patch's dataset, there were two clusters, the one that are

near the roads, and the ones that are far. Additionally, it can be said that the potential dataset illustrates a medium tendency and also reflects upon the dominant side of the tree. This further illustrate that the dimensions in the dataset have strong relationship. Moreover, in both the datasets Principal Component analysis is used for dimensionality reduction.