**Critical visual design**

**decisions report for**

**Data Visualization**

**Submitted by:**

**Koushik Gowda**

**Santosh Vinayaka**

**Goutham Bitra**

**Amith Murali**

**1.Introduction**

Our main goal is to extract meaningful insights from the data by visualizing natural groupings or clusters in the given datasets. The two data sets used are Clients.csv and Patches.csv. We have used PCA and t-SNE algorithms in python to visualize the data. The different relationships among the variables have been explored to get better insights from the data sets using correlation heatmap. We have also created subsets of the data to get better interpretations from the visualizations.

**2.Dataset**

The first dataset used is Clients.csv. The data set consists of information about the clients of a bank. It has 43,193 instances and 9 variables. 2 variables are of the data type integer and 7 variables are of the data type object. The second data set used is Patches.csv. The dataset refers to a cartographic data that consists of observations made over different 30m X 30m patches in the forests of Alberta, Canada. It has 15,120 instances and 7 variables. 6 variables are of the data type integer and one variable ‘Tree’ is of the data type object. We have used LabelEncoder to convert the categorical features into numerical features. The data is also checked for null values and blanks to make sure that there is no error in the visualization. We have not removed any column from both the datasets as there is no correlation and causation between the columns.

**3.Approach**

Once the data was prepared for visualization, we plotted a correlation heat map to explore the relationship among the variables. We created two subsets for Clients.csv dataset. One subset with the columns default, housing, balance and age. The second subset has the columns education, marital, housing and personal. We created three subsets for Patches.csv dataset. The first one for the hydrology data, the second one for Roadways data and the third one for Fire points. This is done to get better visualization results. The numerical features are normalized such that each feature has a mean of 0 and variance of 1. All the subsets are analyzed to find the number of clusters (K) using the elbow plot method, then we run. The k-means algorithm to generate labels. The datasets used have multiple dimensions, in order to visualize this type of data we have used Dimensionality reduction technique. The two dimensionality reduction techniques that we have used are Principal component analysis and t-SNE. PCA is an unsupervised dimensionality reduction technique, which is used to cluster similar data points based on the feature correlation between them without any labels (Supervision). The major challenge we face in today’s world is the volume of data and the variables that define the data. Principal components are the key to PCA, when the data is projected into a lower dimension from a higher dimension. Th lower dimensions are the principal components that captures most of the variance. T-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear dimensionality reduction technique used for dimensionality reduction. The algorithm calculates the probability of similar points in high dimensional and low dimensional space. The difference between these similar points is minimized for a perfect representation of data points in lower dimensional space.

**4.Critical visual design decisions**

We have used PCA and t-SNE algorithms in both the data sets and selected the best one from which we could extract meaningful insights. Firstly, we used PCA as the dimensionality reduction technique on both the datasets to visualize the data using a scatterplot. PCA does not work on both the datasets as it is non-linear and the variance is less than 75%. We also observed a lot of overlapping in the scatterplot which makes it difficult for us to visualize and extract insights from the data. Hence, We have used t-SNE to create a 2-D plot and we have also used k-means algorithm. When we used t-SNE to find natural clusters in our data, we observed that the clusters are very close to one another. To overcome this, we use the k-means algorithm. The k-means algorithm generates labels for the clusters, such that we can differentiate the clusters and visualize the data. We have used elbow plot to find the value of K, that is to find out how many clusters can exist in the data. In Clients.csv dataset we have used two subsets. In subset1 we get a good plot when we set the perplexity as 30 and the number of iterations as 2500. in subset2 the perplexity is set to 45 and the number of iterations is set to 5000 to get a good plot. In Patches.csv dataset we have used 3 subsets. Subset1 has the perplexity set to 30, subset 2 has the perplexity set to 35 and the subset 3 has the perplexity set to 50. The number of iterations is set to 2000 for all the subsets. The above-mentioned values of perplexity and the number of iterations gives us the best plots.

**6. Contribution**

Data visualization as an assignment was a challenging and interesting one for us. We made sure that each one of our group members contribute equally in the assignments, so that the learning outcomes from the assignments benefit all of us. Since we have two data sets, we split up into two groups and took up one dataset each. Goutham and koushik worked on subsets of Clients.csv dataset and provided meaningful insights from the visualizations. Goutham worked on the first subset and Koushik worked on the second subset. As the first dataset has a lot of data when compared to the second one, they needed more time to come up with good visualizations for better insights. Amith and Santosh worked on the second dataset Patches.csv. We made three subsets of it and two of them where worked on by Santosh and one of them was worked on by Amith. Once we had the visualizations and insights ready for both the datasets, we divided the report into 6 sections and split it among us. The recordings in the power point is done by Amith by using the insights provided by all of us for each subset.