Assignment 11
Data Visualization: Advanced Topics – WS 20/21
Ali Asghar Marvi

Exploratory Data Analysis

Introduction:

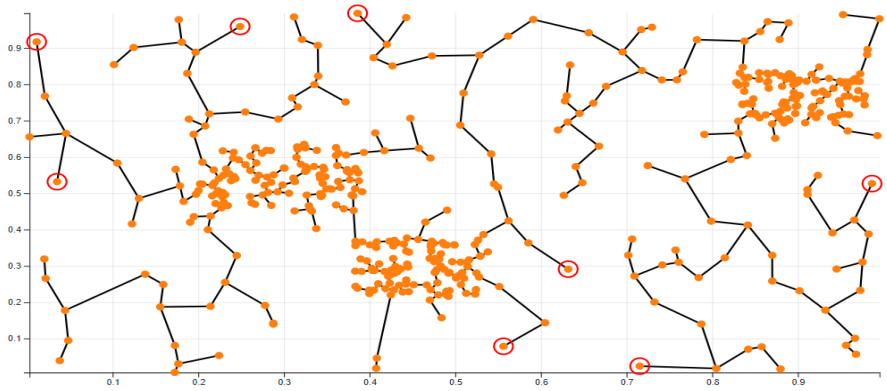
Exploratory.csv has 12 dimensions with numerical values. I added a "class" column with dummy string value to make it easy to be called in the framework. My framework composes of scatter plot implemented with MST and Tukey measure to detect outliers, Scatter Plot Matrix, Parallel Coordinates Plot with Edge Bundling, TSNE and MDS projections and Pixel-based plots. I have tweaked different pair of dimensions to detect possible patterns. I will be explaining how each view helped me and if it didn't it would be documented as well. Also all findings are validated using screenshots provided. Details are as follows:

Scatter Plot:

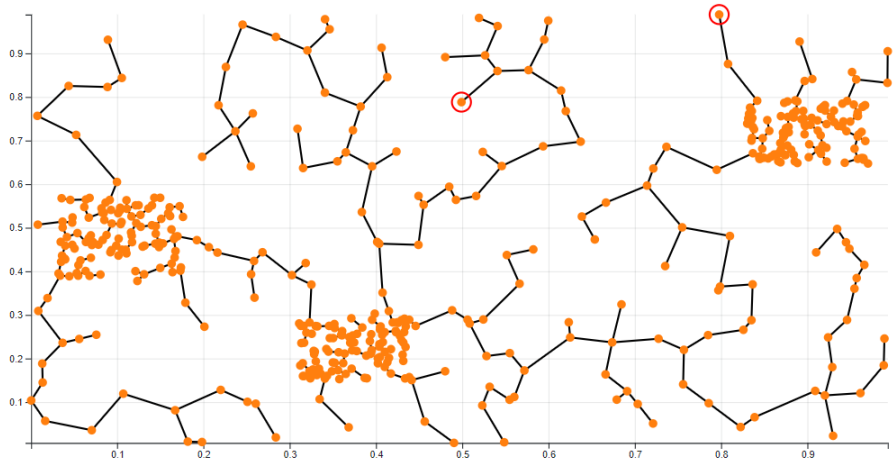
Scatter Plot view is hardcoded with first dimension and the second dimension. After alternating between subsequent dimensions it becomes visible that this data is normalized between 0 and 1.00. This is evident from the axes and its corresponding point on the plot. Now Scatter Plot Matrix would have helped me too without alternating between any but due to higher number of dimensions and better visibility Scatter Plot view was used only.

Each plot which rendered showed three groupings in the data apart from the noise (indicated by points lying far from clusters), this indicates that data has three distinct classes if analysis were to be furthered. Tukey measure implemented shows several outliers mainly the ones which values in range 0.98-1.00 but since it is prone to the dimensions of SVG used, some of the outliers detected did not make sense.

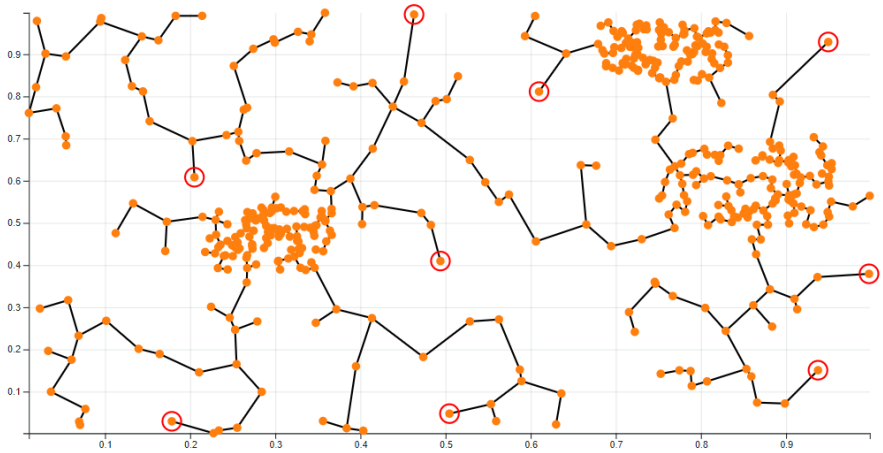
Plot between dim1 and dim2:



Plot between dim11 and dim12:

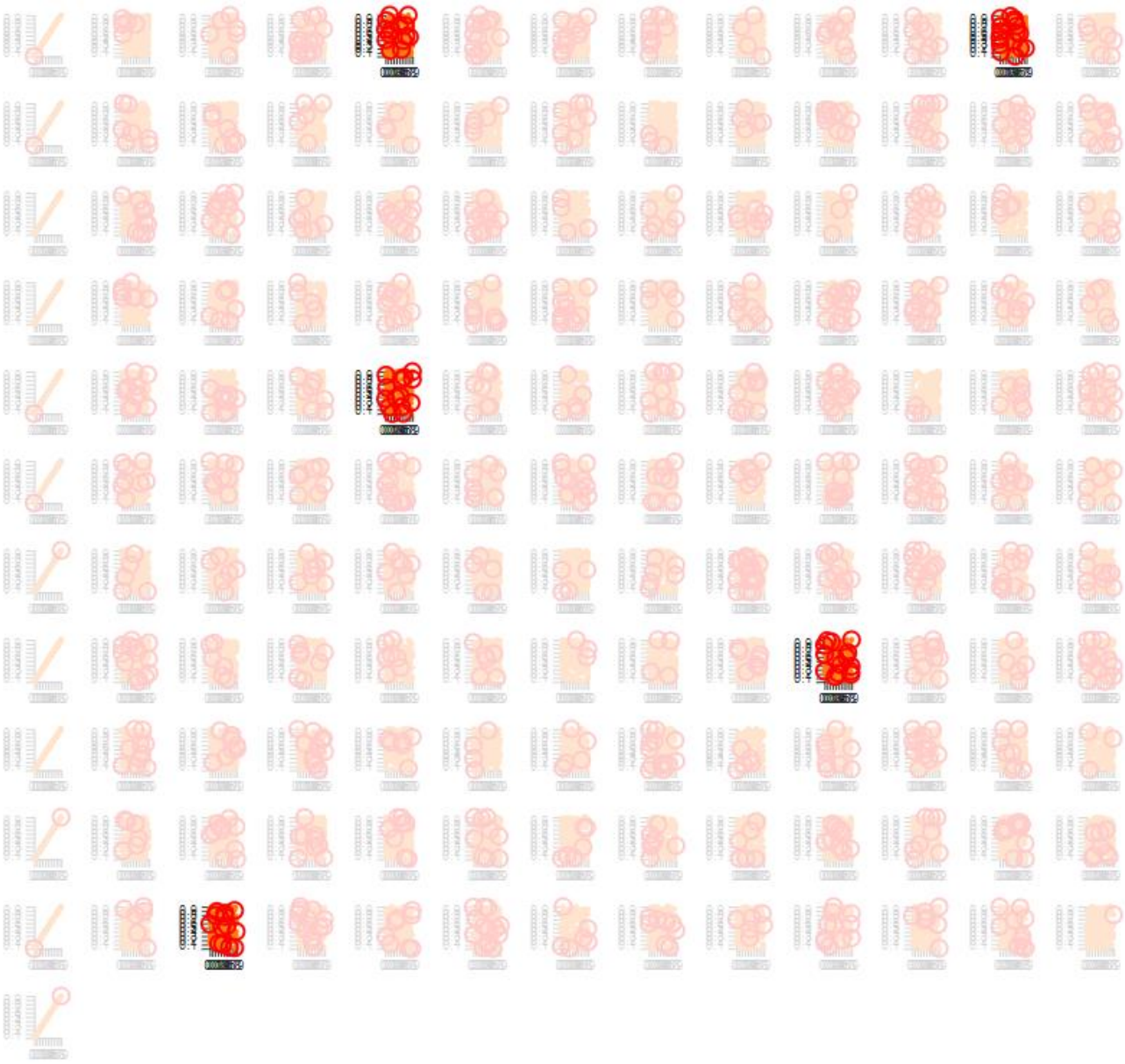


Plot between dim5 and dim6:



Scatter Plot Matrix:

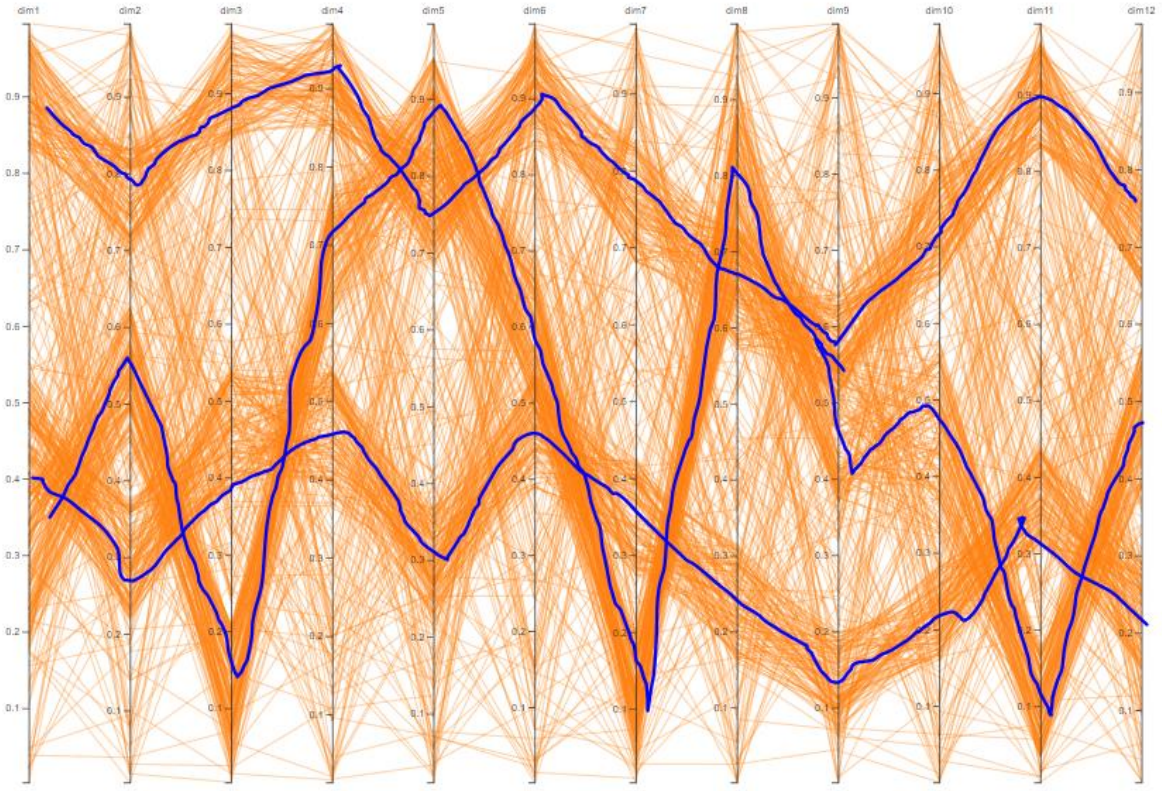
Even though SPLOM is not evident enough to display relation between each dimension but with implementation of Tukey's measure on a slider, several dimensions can be filtered using average outlying measure. So lets say we assign a threshold of 0.1, we can see that "dim11", "dim4", "dim9" and "dim2" can be filtered for having outliers but this information is not enough to completely discard a dimension. Techniques like Correlation Plot or Histograms would have proven it to be better in further scrutinizing dimensions. Find screenshot below:



Parallel Coordinate Plot:

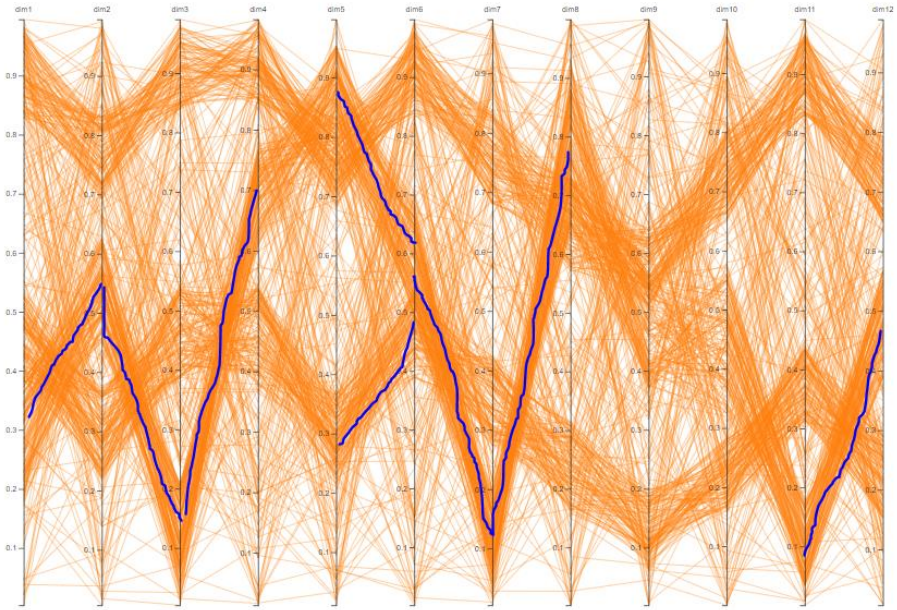
Upon plotting the PCP, three shaded regions are evident across the axes. This again denotes three clusters present in the data. Upon bundling the edges with Beta set to 1, we can see how edges are bundled at the centroids. Since centroids are computed using the median of dimensions wrt class variable, it can be seen that median of dimensions are generally between 0.4-0.7 unlike in dim4 and dim5 where the centroid is formed above 0.7. As far as detection of outliers is concerned, data seems to be scaled in the manner such that there are no points located unreasonably on the axes. PCPs are helpful in identifying any visible inverse collinearity. One can identify it using dim7 and dim8, dim3 and dim4, dim11 and dim12, etc. Please find screen shots of clusters identified and marked inverse relationship:

Clusters:



Alpha: 0 Beta: 0

Marked region for inverse relationship:



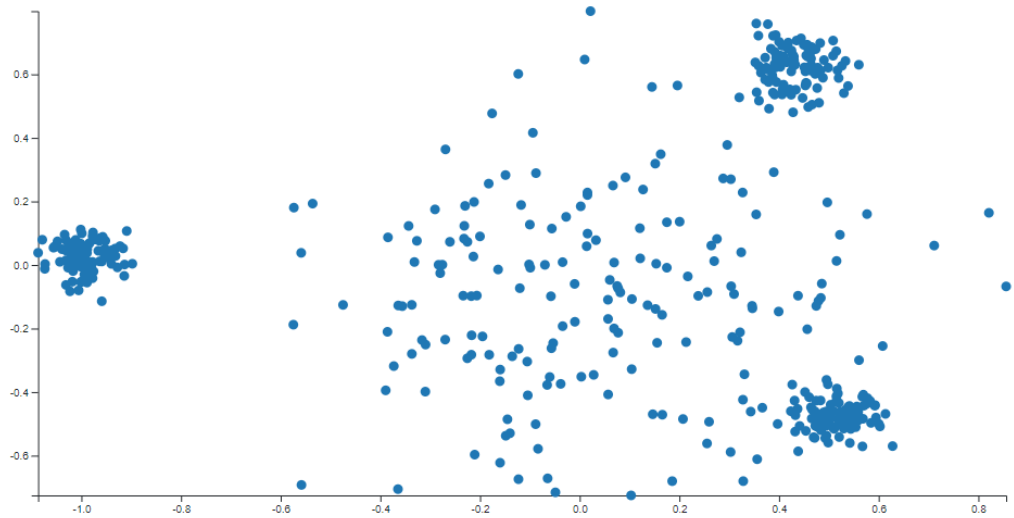
Alpha: 0 Beta: 0

Inverse relationship can be seen between majorities of data lines which are not part of visible clutters. This proves potential of inverse relationship in the data.

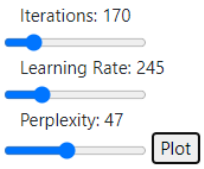
TSNE and MDS:

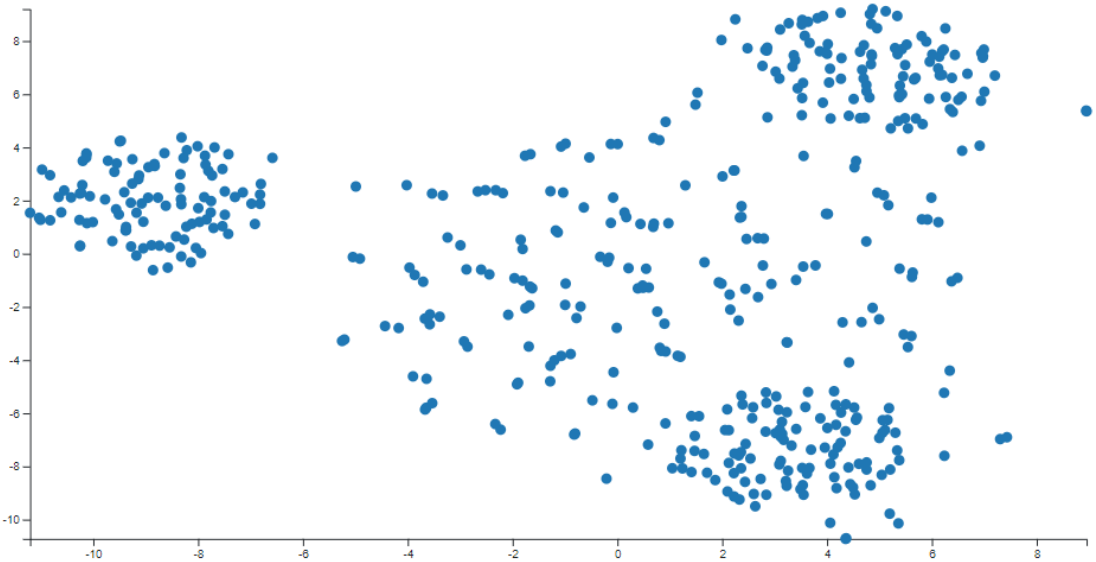
I will be discussing TSNE and MDS together since these are both projection techniques and are responsible for identifying clusters in the dataset. One does with probabilities and another does it by identifying coll inearity between dimensions using Euclidean distance. Both visualizations are sufficient enough to identify three clusters in the dataset:

MDS:



TSNE:

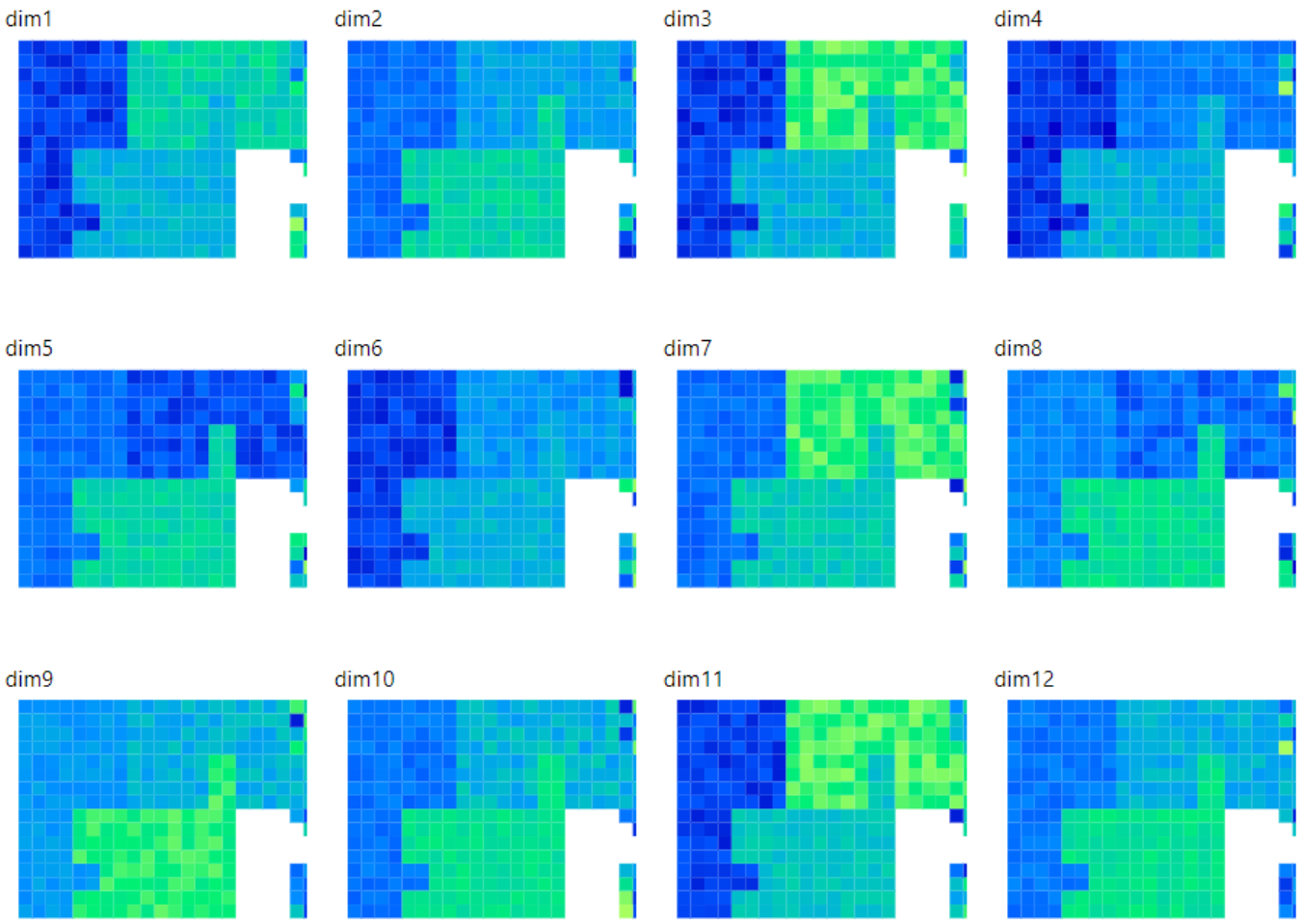




Based on the screenshots above three groupings can be identified in both MDS and TSNE plots.

Pixel Plot:

Pixel plot consist of each dimension plotted individually where each pixel is represented by small colored squares of each Hilbert curve in dimension. Each bar shows three different color coordination denoting three different clusters. Find screenshot below:



Note that no sorting is needed anymore since clusters are evident.

Conclusion:

Plots like Correlational Plots, Histograms could have been fruitful in doing analysis w.r.t ranges of values and identify in which range values are cluttered more. Histogram bars would have been helpful in identifying the frequency of values lying within a certain range. Such an analysis could have been further validated using a Pie Chart or Donut Chart.

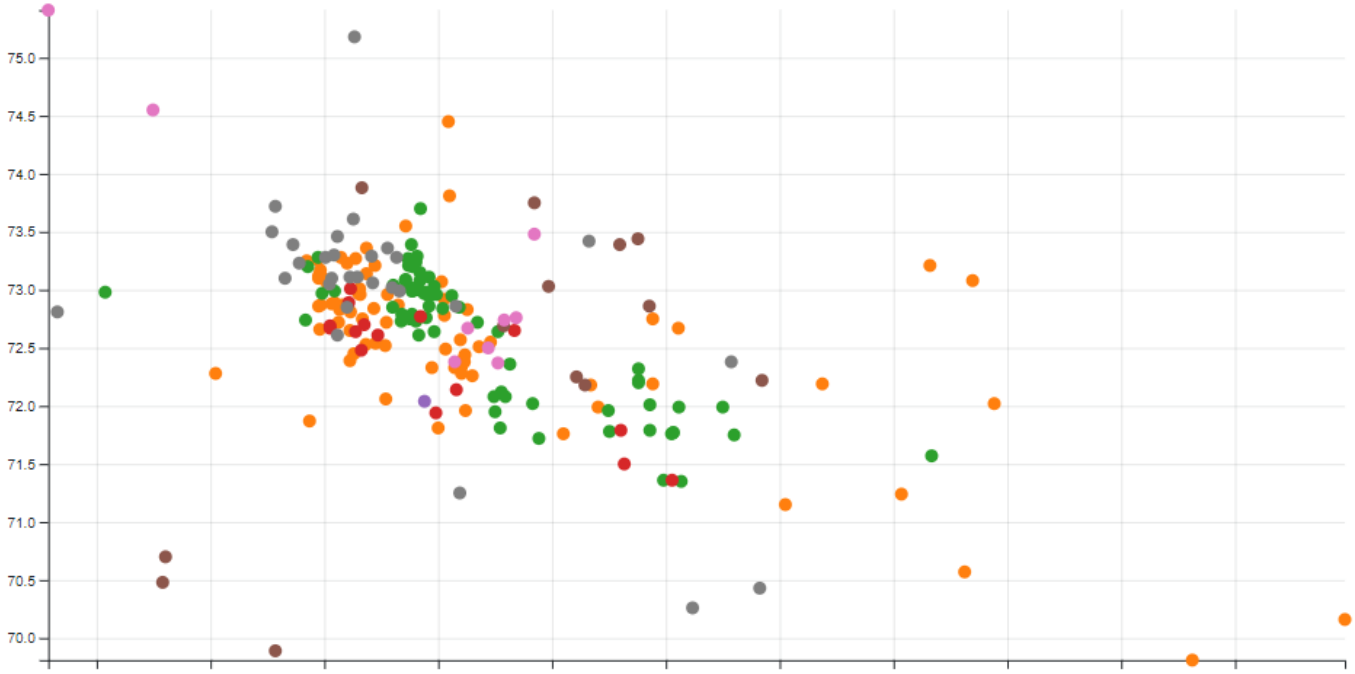
(P.T.O)

Confirmatory Analysis

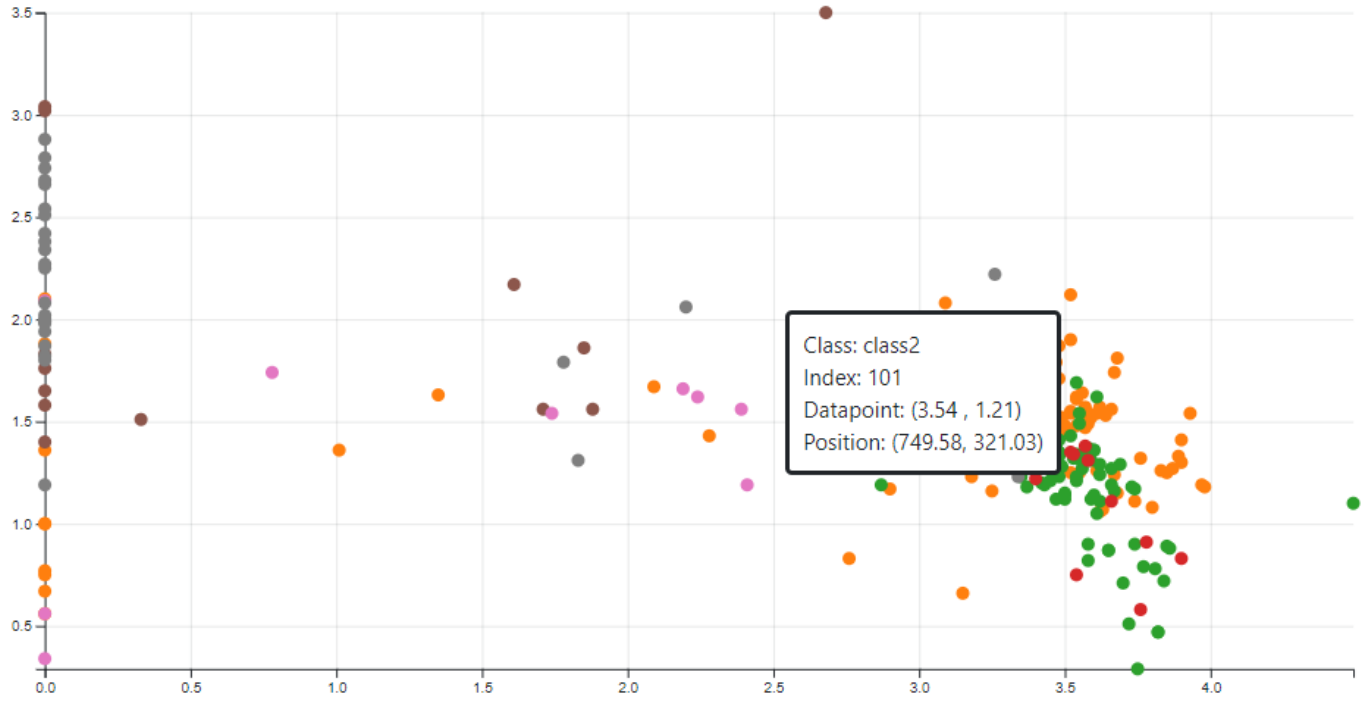
1 ➔

Scatterplot is use to identify discriminating dimension. I plotted by tweaking between different pair of dimensions and among mostly all of them class2 turns out to be the evident represented by green color. As far as the dimension is concerned which is solely responsible for different clusters for same class variable is dimension6. For further validation please check screenshots below:

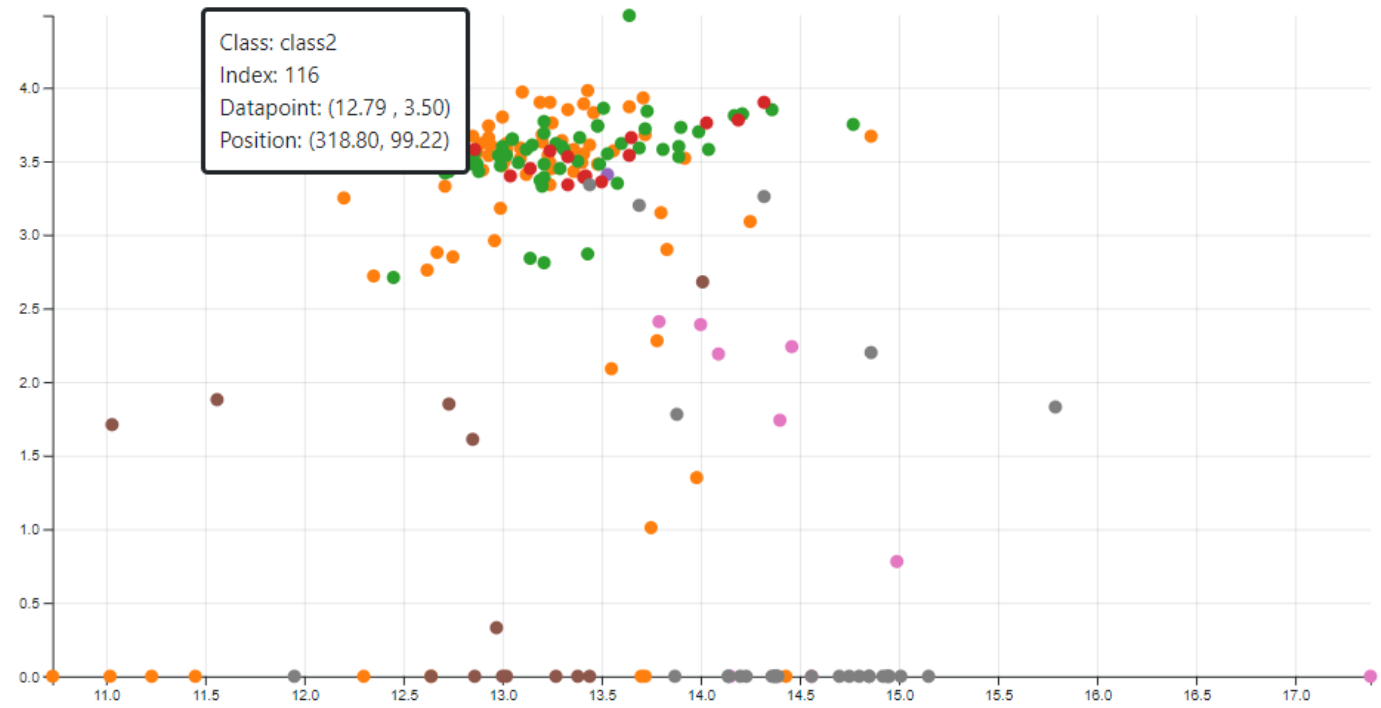
Dim1 vs Dim2:



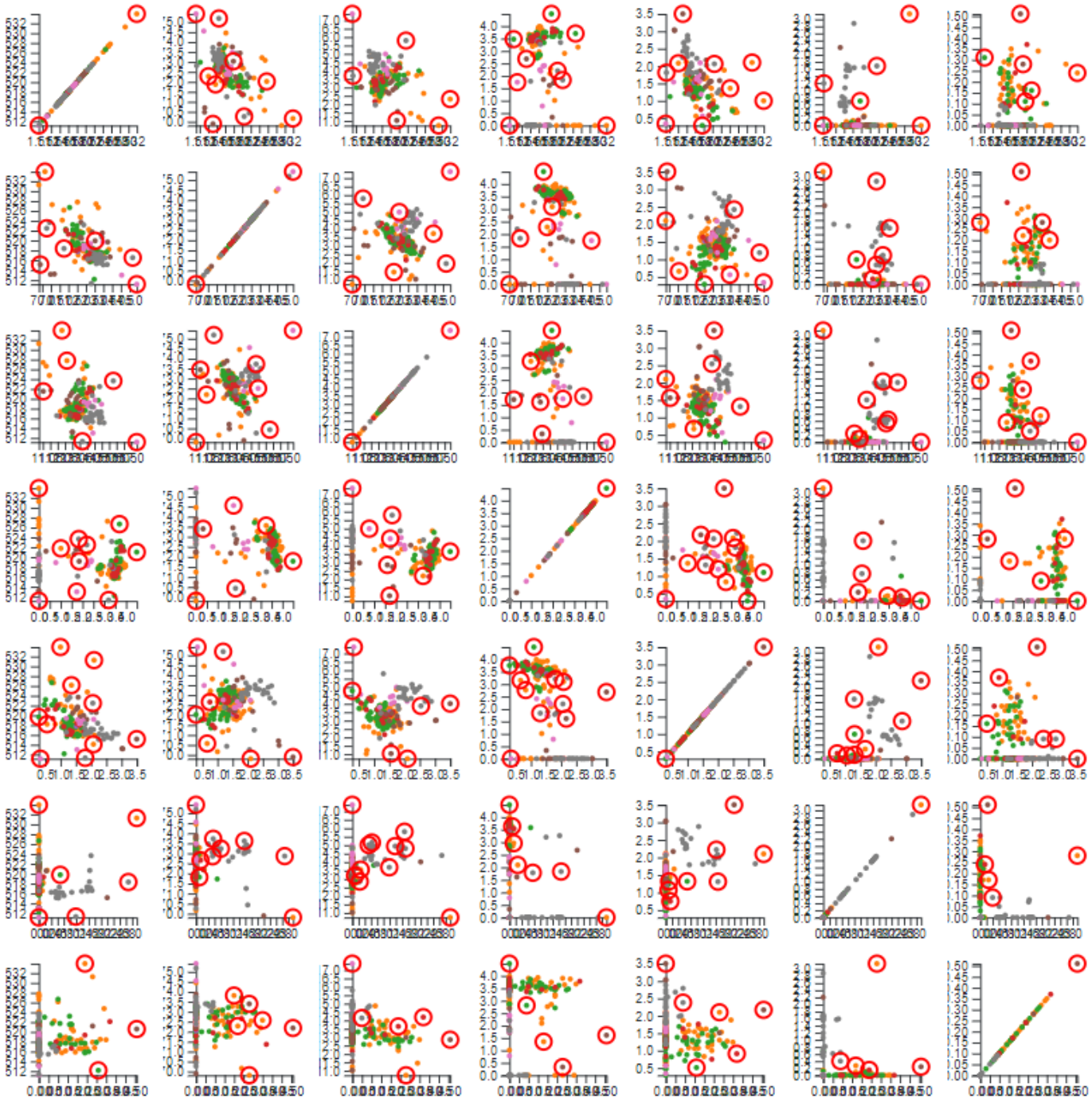
Dim4 vs Dim5:



Dim3 vs Dim4:



SPLOM further validates this finding by identifying grouping in the purple for dimension 6:

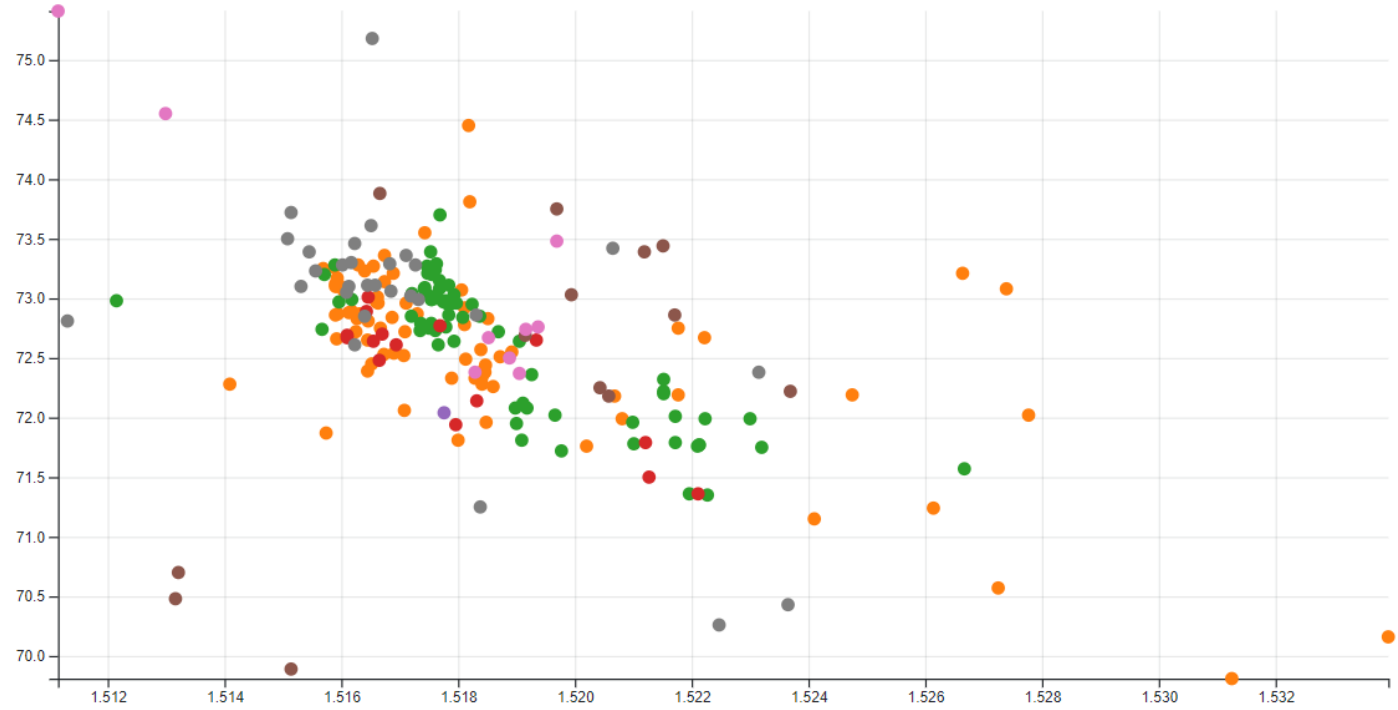


2 ➔

Again using Scatter Plot we can see overlapping points with different colors mainly Green, Orange and Red, denoting that class1, class2 and class3 is similar. Again this is tested using different pair of dimensions and is further validated by bundling and smoothening edges on Parallel Coordinate Plot. MDS and TSNE return the same overlapping clutter. Screenshots of which are given below:

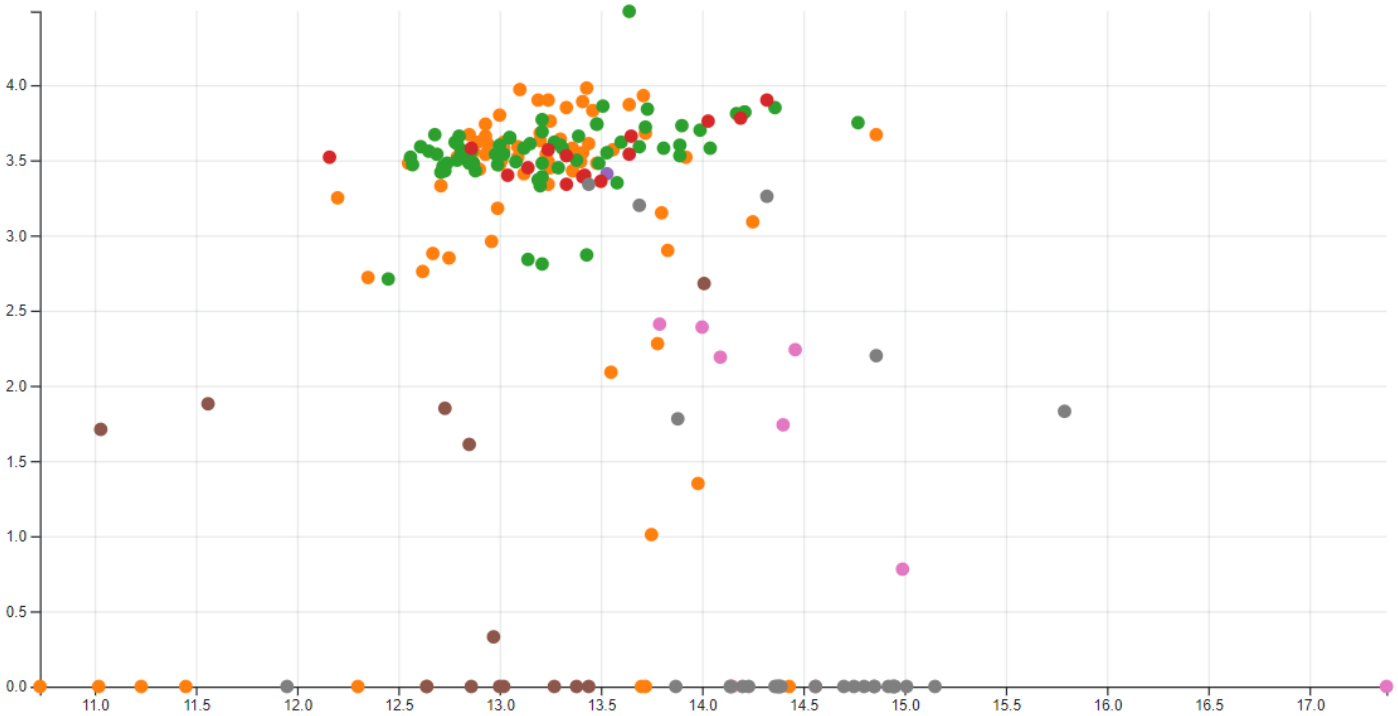
Scatter Plot Dim1 vs Dim2:

Class6 in grey color seems to be the similar one too however its not the case in other dimension pairs.



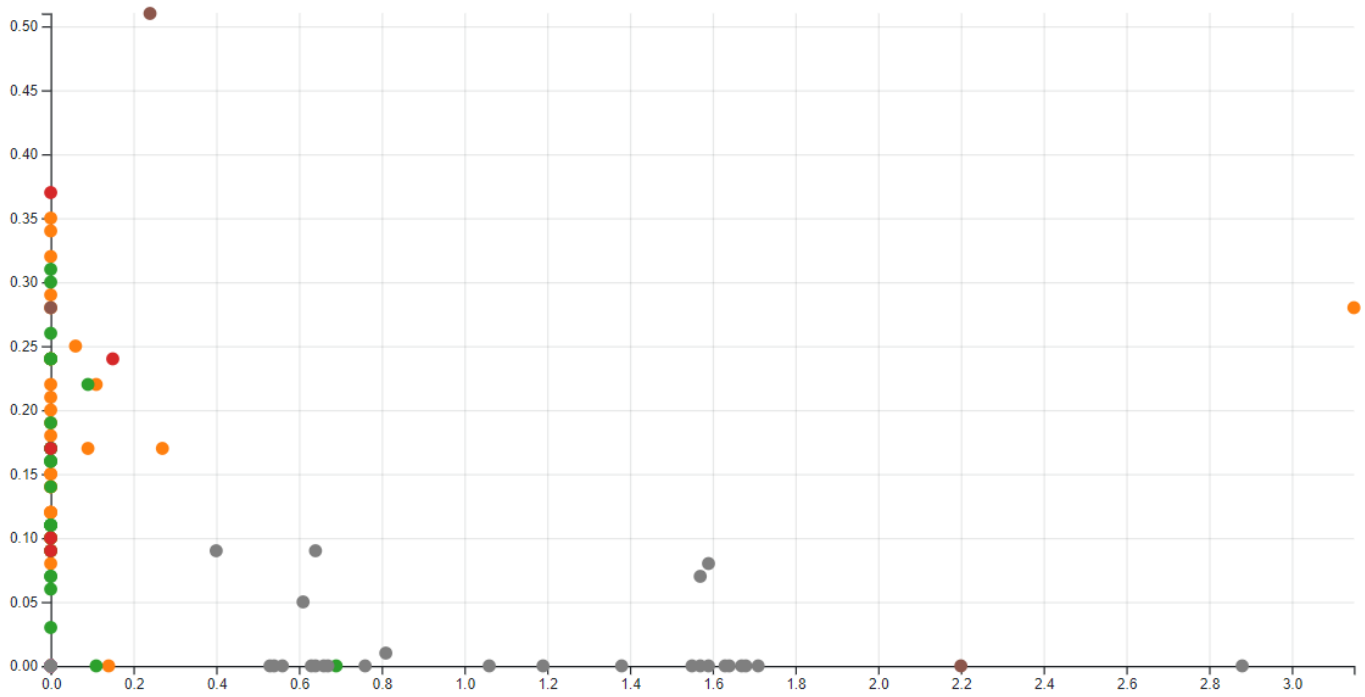
Dim3 vs Dim4 Scatter Plot:

As seen below class6 (grey points) are not part of clutter anymore



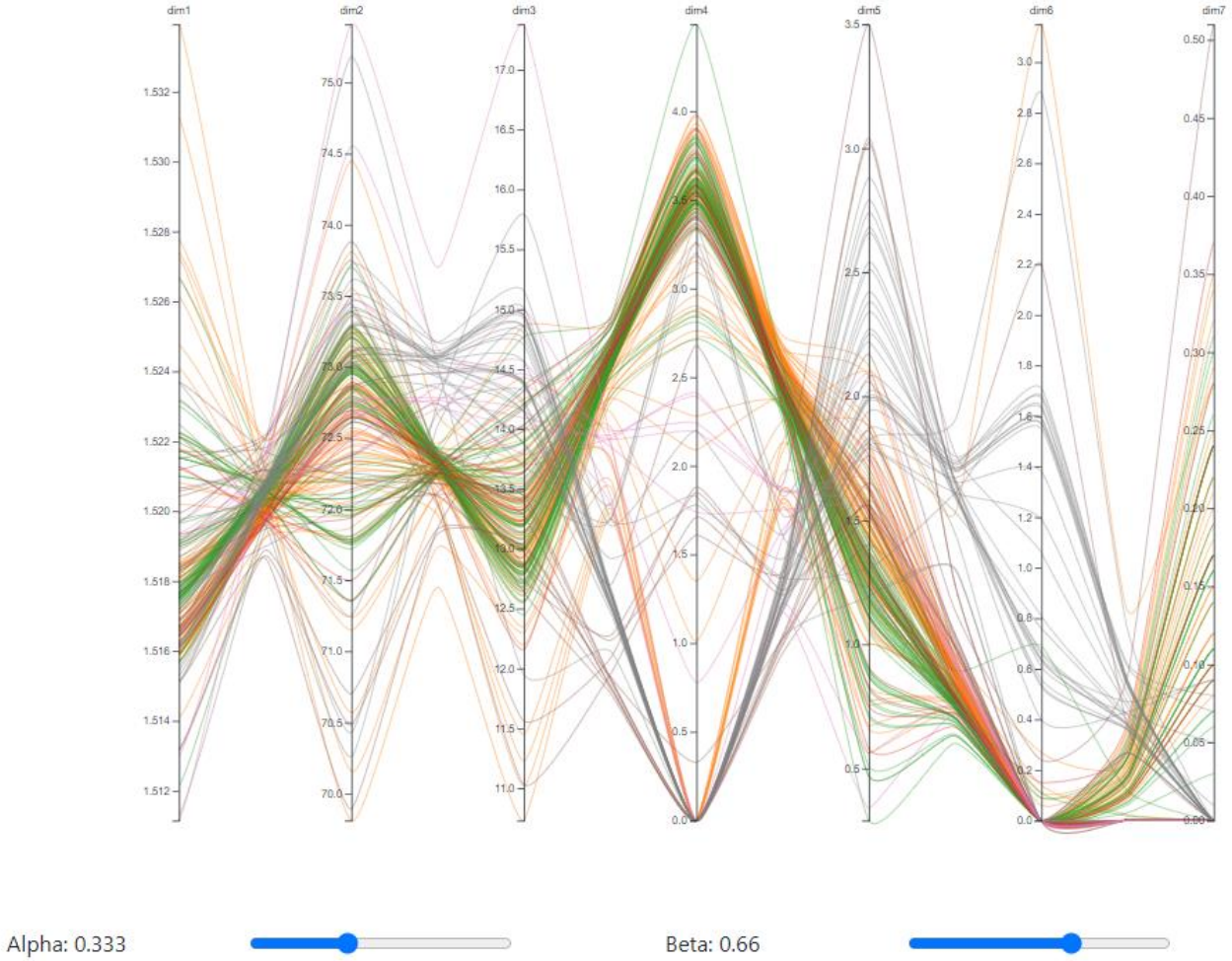
Dim6 vs Dim7 Scatter Plot:

Again is the case between dimension 5 and dimension 7:



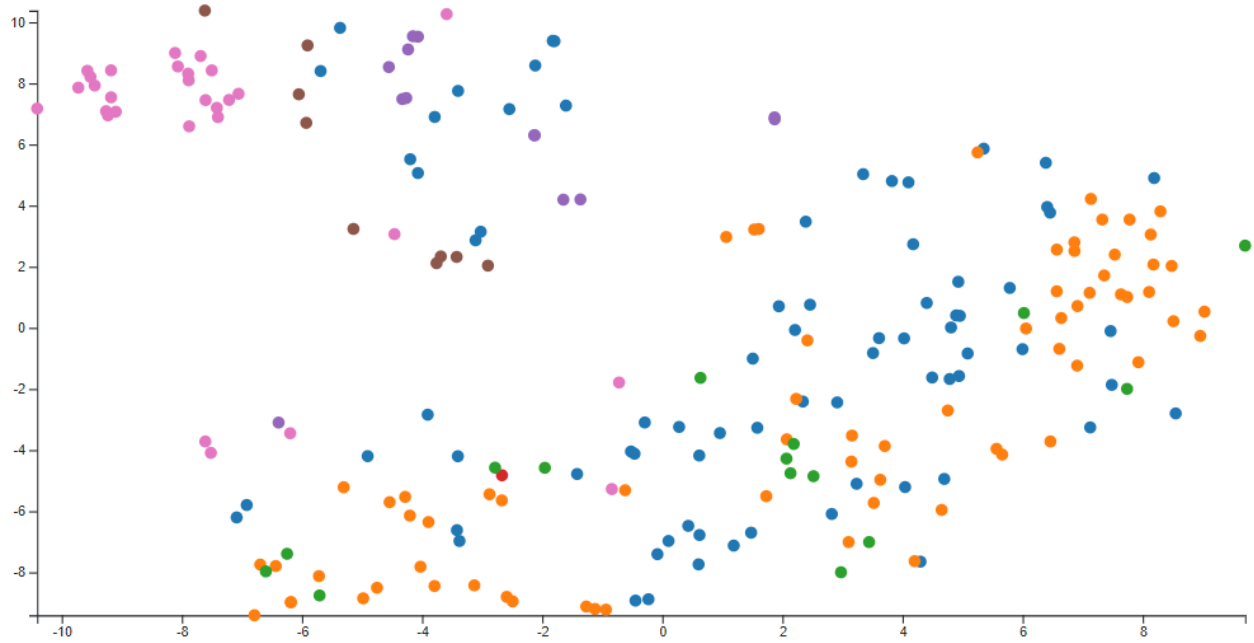
Parallel Coordinate Plot with bundled edges:

Findings are further validated with overlapping green, red and orange (referring to same class1, class2, class3) paths.



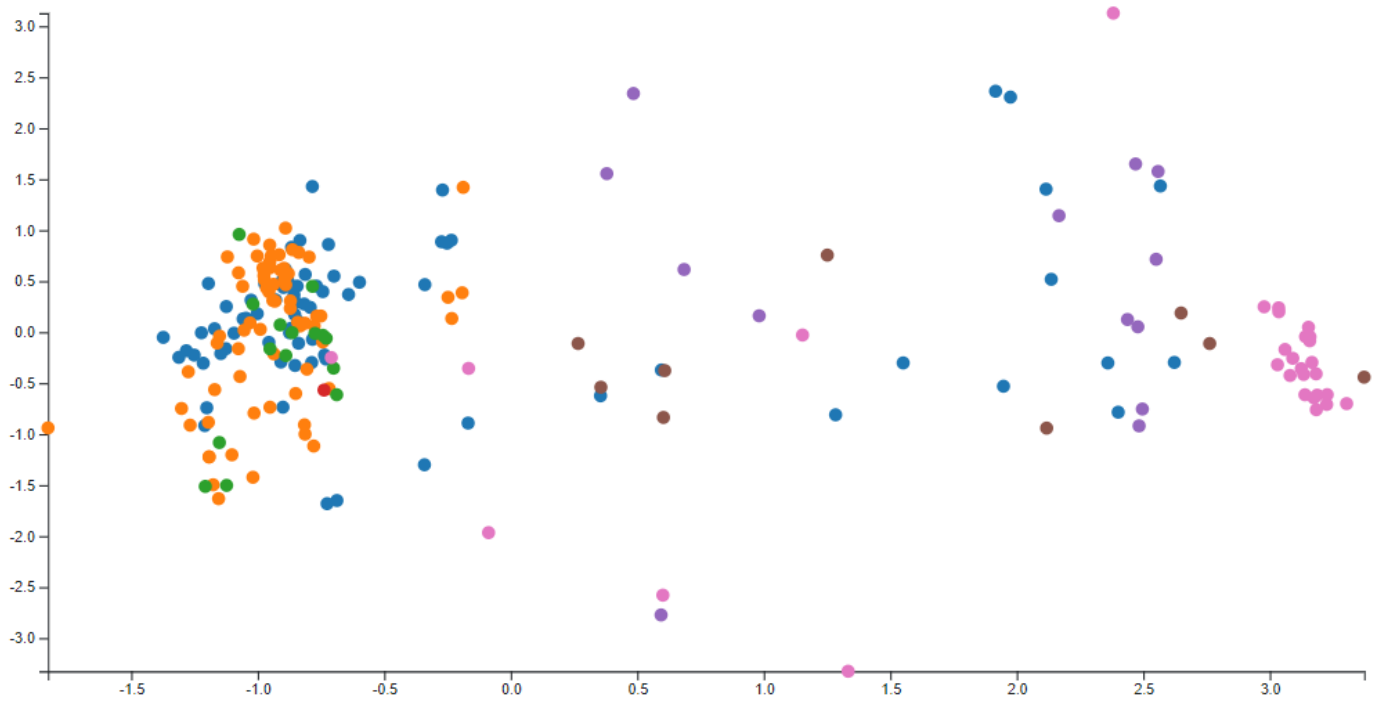
TSNE Plot:

Same is done by TSNE where class1, class2 and class3 are similar (colors orange, green and blue) :



MDS Plot:

Same is returned by MDS where class1, class2 and class3 are similar (colors orange, green and blue):



Pixel plot is not so much helpful since it is not meant to find similar classes but rather the range of distribution in values.

3 ➔

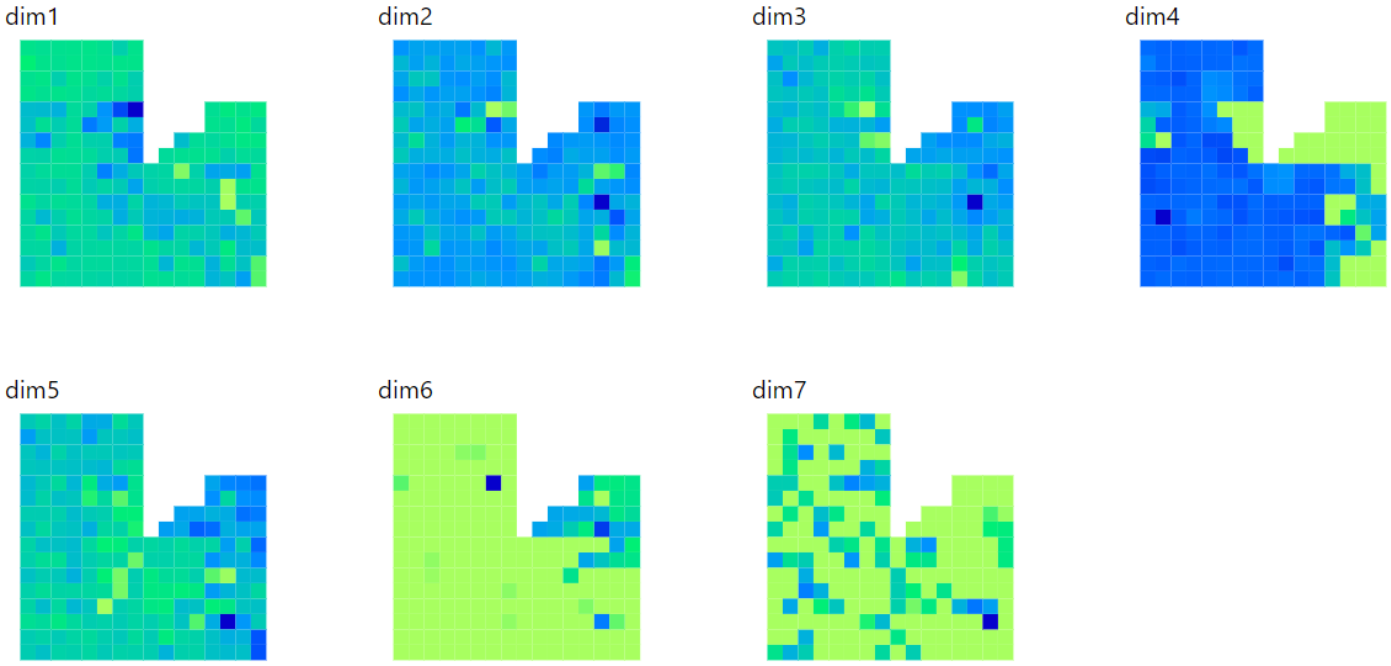
For this hypothesis we will use Pixel Plot to find variation in dimensions only. However, it is not helpful enough to figure which class does the variation applies to. To further dig in we will be using Scatter Plot Matrix to identify variations in corresponding classes. If we look at the Matrix carefully we would see class1 and class4 points spread out in dimension1 and dimension2. Similarly, between dimension1 and dimension6 the variation is high for class6 (grey point) but is not the case for other classes. However, it is not the case between dimension2 and dimension4 for class6 especially. Same can be seen between dimension5 and dimension4

Between dimension3 and dimension6 the class6 is scattered more than the rest which are mainly on the horizontal axis only. Between dimension7 and dimension6 the plot has values closer to the axes where grey is mainly towards the dim6 axes, this denotes that there are a lot of similar or same values on dimension 7 with some outliers. This distribution of dimension 7 is less varied as compared to the rest of classes.

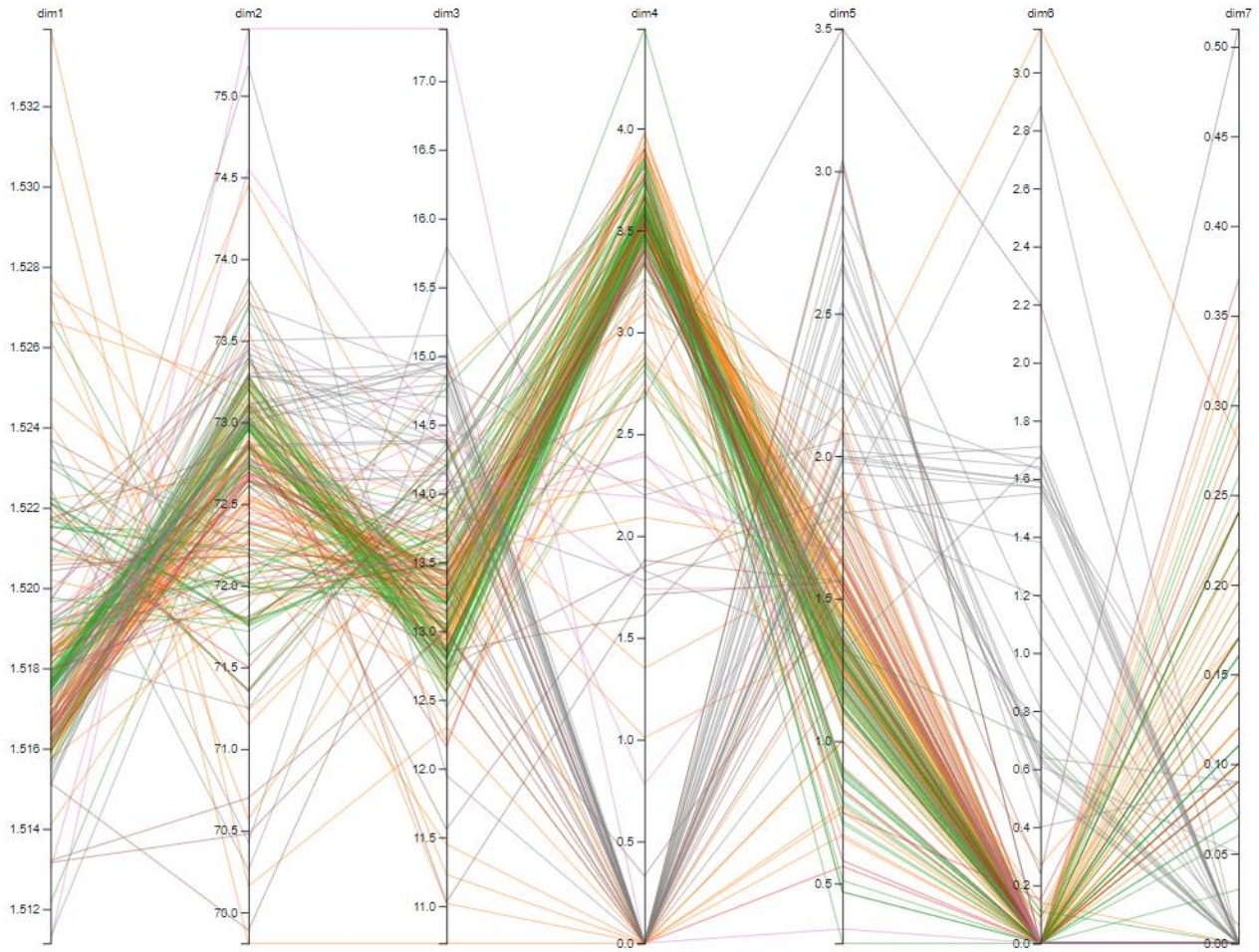
Dim7 and dim4 has class1 more scattered but is not the case in dimension2 and dimension6 plot which can be seen from the matrix that dim6 has similar values for class1. Dimension 5 and dimension 4 also has scattered value for class1 too.

We can also see an outlier in dimension7 for class1 and class 6 as seen from the dim7 vs dim7 plot. We can also see outliers for class1 in dimension6 and dimension1. The bit about outliers is further verified from Pixel Plot where in dim6 the color coordination with other pixels is completely off for one data point. Same can be said about each dimension where there is at least one outlier. This is further identified using Parallel Coordinate Plot where there is an outlier in every dimension on the top. This outlier is the lowest bound of range of values in each dimension

Pixel Plot for confirmatory analysis:



Parallel Coordinate Plot:



With presence of noise and several outliers it can be seen that classes are varying across all dimensions and paths follow a more “zig-zag” structure.

Conclusion:

This prototype depends a lot on labelling and color mapping which is not the case when comparing TSNE or MDS plot with the rest. This could have been better if there were tooltips on hovering any point or line. This signifies how much role a tooltip plays in further drilling of information being visualize d. As far as distribution of classes is considered then Pie Chart, Donut Chart or a Bar Chart would have been way convenient in identifying the distribution of class variables. As far as identifying variation is considered ensuring reordering of axes in Parallel Coordinate Plot would have assisted further in identifying corresponding classes. Such a control at user end along with the visualization would have been really helpful.