Name: Quynh Diem Luong
Student Number: 808244

ASSIGNMENT 3

Exercise 1

a. *Make a trellis (small multiples, similar to Exercise 4 in Assignment 1) of 2-dimensional scatterplots of the point set in X such that you map the value of vector t to a suitable continuous colour scale (e.g., spectral scale)*

Answer:
Figure 1 below shows the trellis of two-dimensional scatterplots of the point set in X

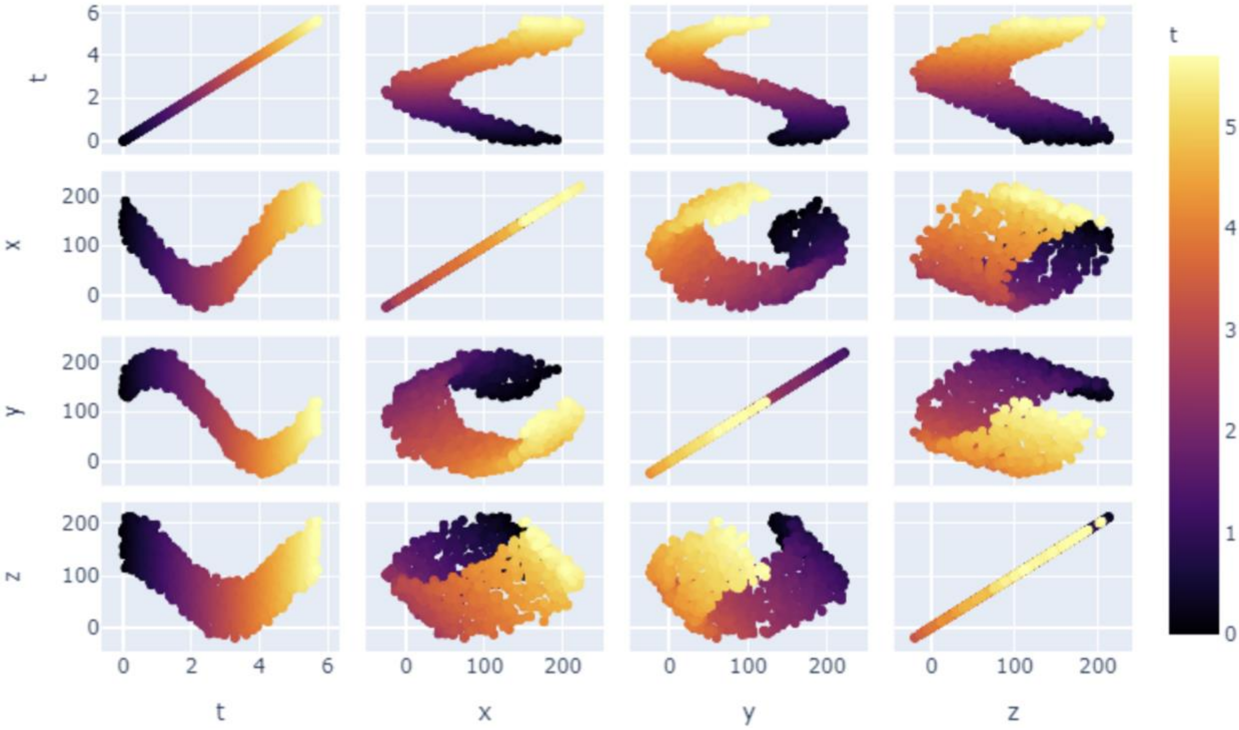

Figure 1. A trellis of two-dimensional scatterplots

As can be seen, the value of vector t is mapped to a suitable continuous color scale. The chosen color scale for the visualization is the Inferno color scale. The Inferno scale belongs to the viridisLite package, a color package designed to improve visualization readability for readers with common forms of color blindness. Thus, it is chosen in the visualization.

b. *Use pca, project the data to the first principal component, and make a plot of the data into one dimension using the same colour scale as in item (a) above. Also, make a histogram of the one-dimensional embedding. With pca, it is a good idea to center the data first. Why? What would happen if the data would be uncentered when looking for a maximum variance projection?*

Answer

Figure 2 below shows the plot where data is projected into one dimension using the same color scale as figure 1.

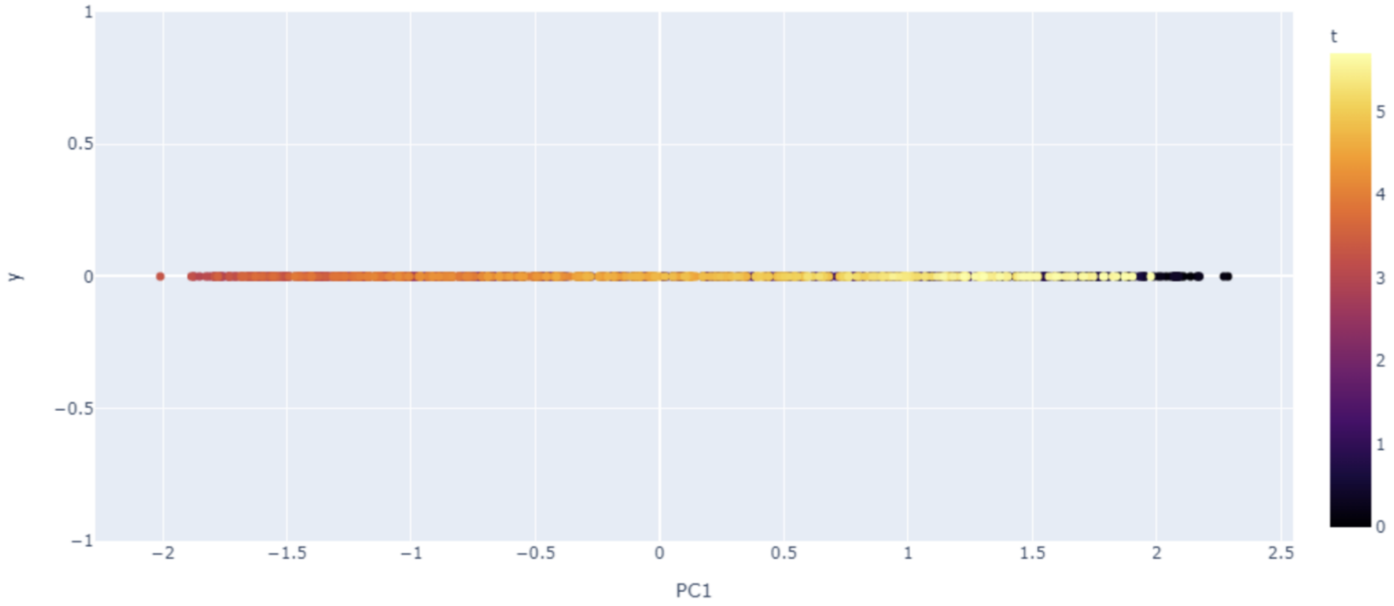

Figure 2. Data is projected into 1D

The histogram of one-dimensional embedding is shown in figure 3

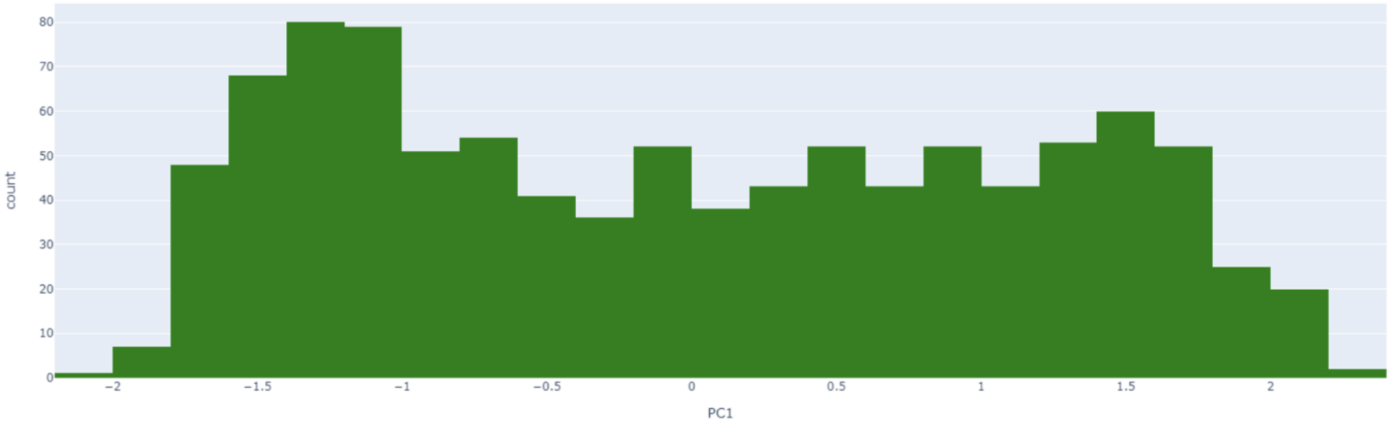

Figure 3. Histogram of one dimensional PCA

By centering a dataset, it means subtracting the mean value from each observation in the dataset. When the dataset has different units and scales, normalizing data before performing PCA could be important. Since high variance in the data could result in a high signal-to-noise ratio, dimensionality reduction results in a loss of some critical information. Thus, centering data could be used to make PCA less arbitrary.

c. *Make two-dimensional plots of the data projected to the first and second pca components, and to the second and third components. Based on these, what can you tell about the data set's shape?*

Answer:

Figure 5 and figure 6 below show how data is projected to the first and second PCA components as well as second and third components.

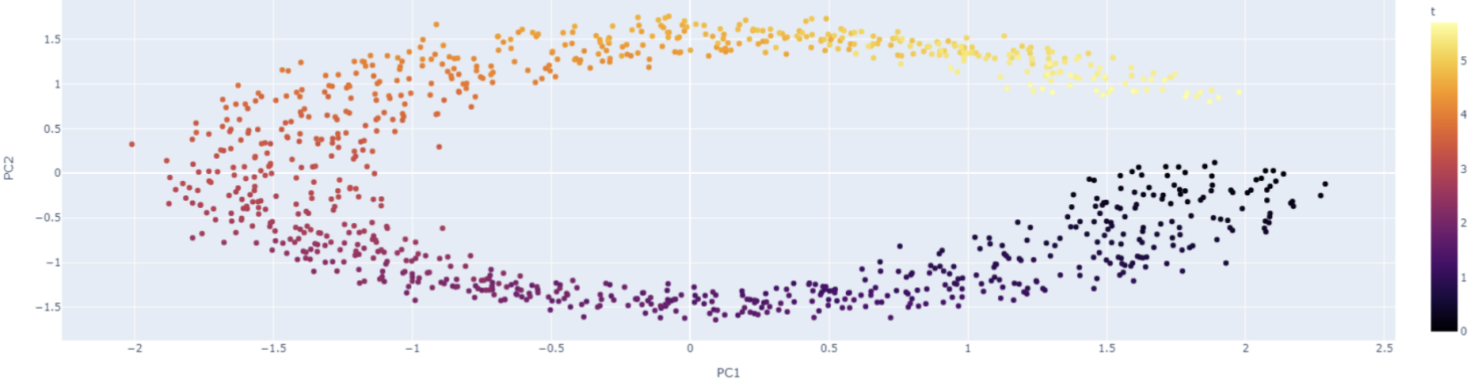

Figure 4. First and second PCA component plot

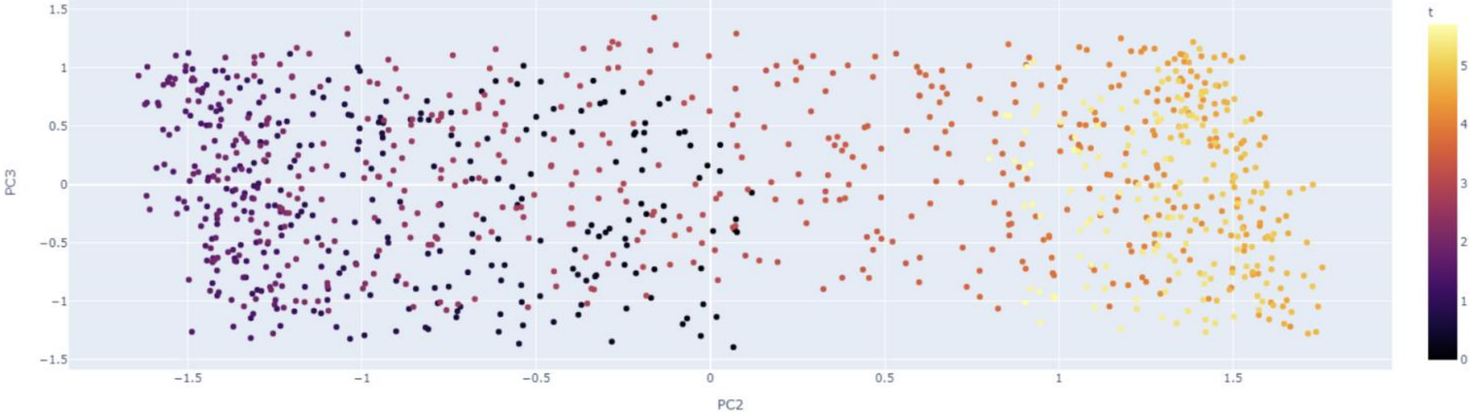

Figure 5. Second and third PCA component plot

In figure 4, we see that the shape of the data is formed into a curve. Data points are projected with similar t values and are clustered next to each other. Meanwhile, in figure 5, the data shape seems more distributed. We see that the data is proportional to the t value and increases regarding y-axis. In other words, the higher value in PCA2, the higher t-value.

d. *Use nonmetric MDS or Sammon mapping to embed the data into one or two dimensions and plot the data the same way you did in item (b) above*

Answer:

Figure 6 below shows one-dimensional plot using MDS mapping the same way in item b. In this visualization, the data has been centered and MDS mapping was applied to embed the data into one dimension.

We see that the data looks quite messy and compressed into one line. Data points with different t-value are widely scattered while data points with similar t-value are not adjacent. As a result, the projection performed by MDS is not as good as how it was performed in PCA.

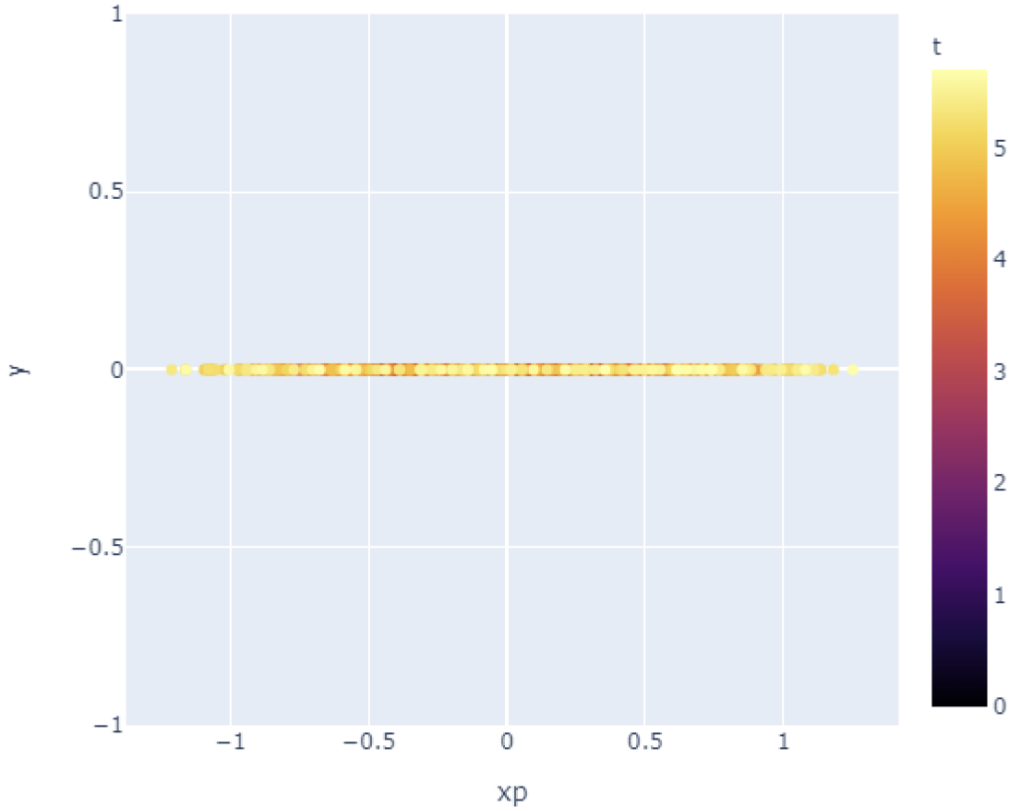

Figure 6. MDS Mapping Plot in 1D

The histogram of one-dimensional MDS mapping is shown in figure 7.

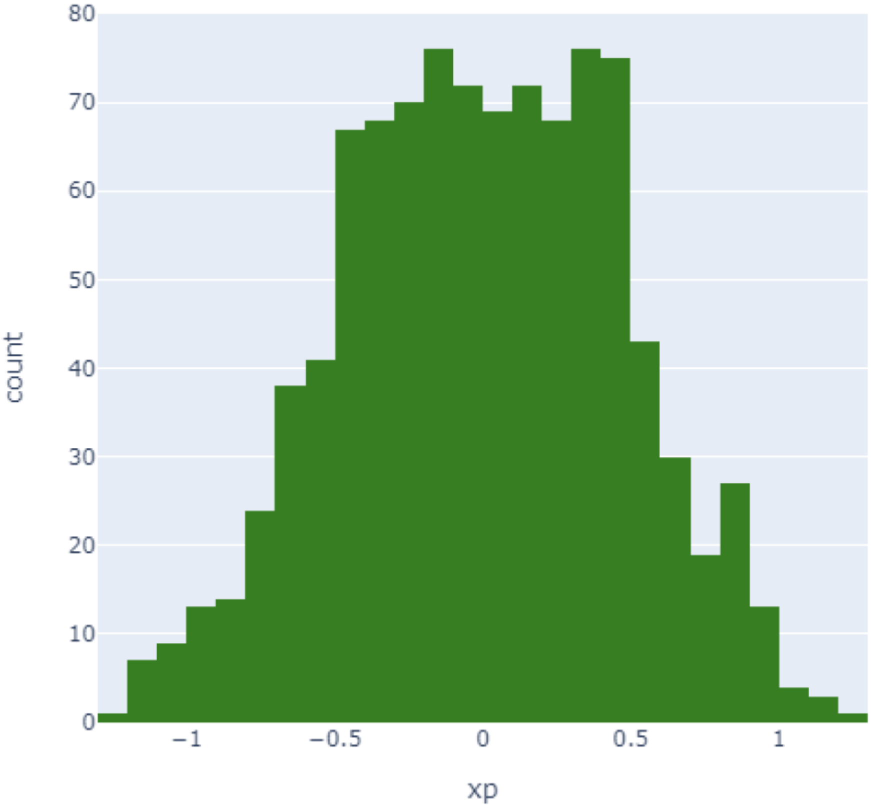

Figure 7 Histogram of one-dimensional MDS mapping

e. *Use Isomap, discussed in the lectures, to embed the data into one or two dimensions and plot the data the same way you did in item (b) above.*

Answer:

Figure 8 shows how data is embedded into one dimension using Isomap. In this visualization, the data is projected well on the x-axis. The t-values are clustered together and are well distributed along the x-axis from low value to high value. Compared to PCA and MDS mapping, Isomap visualization shows better visualization regarding the projection of data.

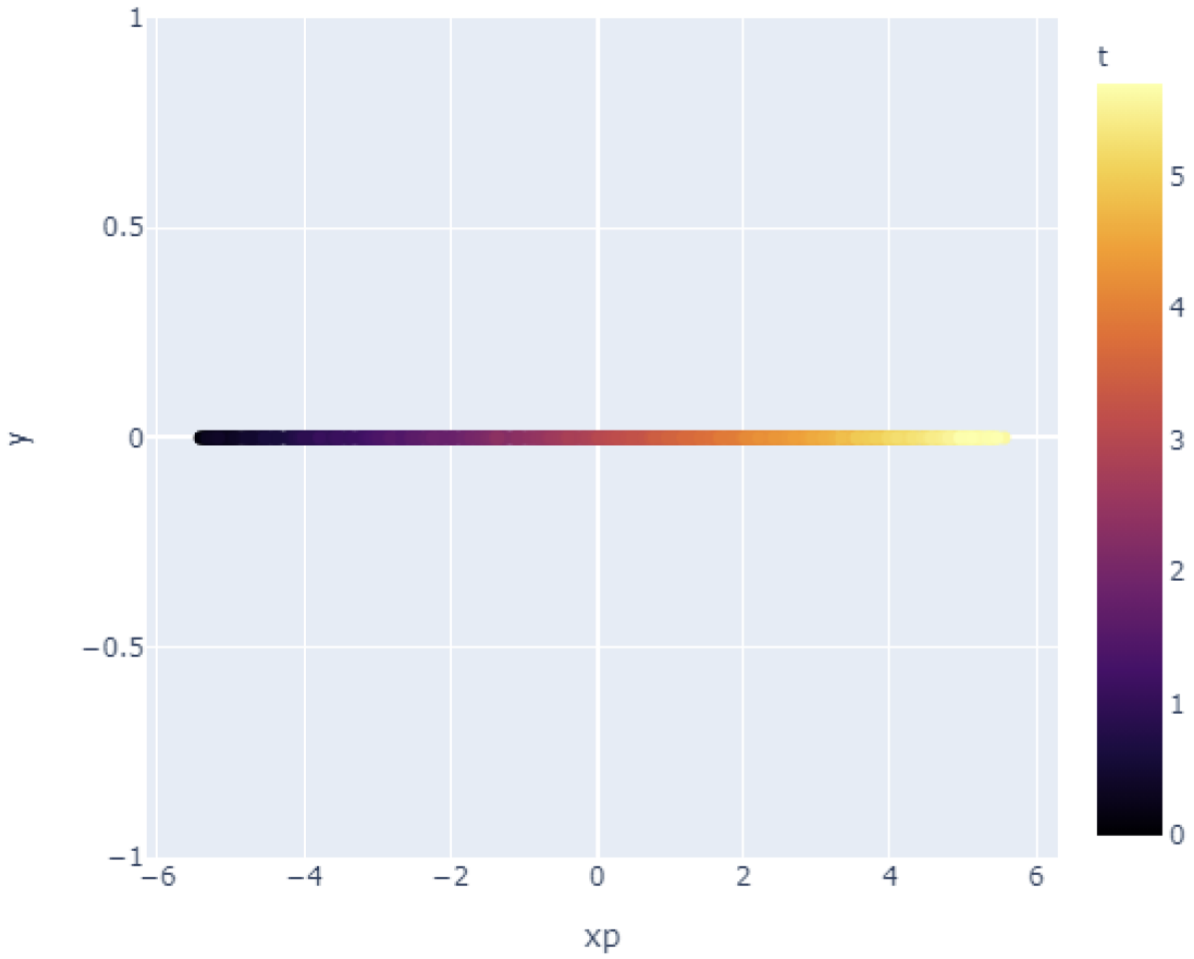

Figure 8. ISO Mapping in 1D

Exercise 2

Download from MyCourses the dataset population_data.csv which contains statistics on population age structure in Finnish municipalities. Compute and plot Metric mds (mmds) and Sammon mapping. Annotate some selected (or all) places, for example, main cities, provinces, places where you have been/born, etc. Compute Shepard plot (a scatter plot of output distances as a function of input distances) and compare the plots of mmds and Sammon mapping. Which method predicts which distances better?

Answer

Figure 9 and figure 10 show the MDS and Sammon mapping in the big cities of Finland such as Espoo, Helsinki, Oulu, Tampere, Vantaa, Turku, Lahti, and Kuopio.

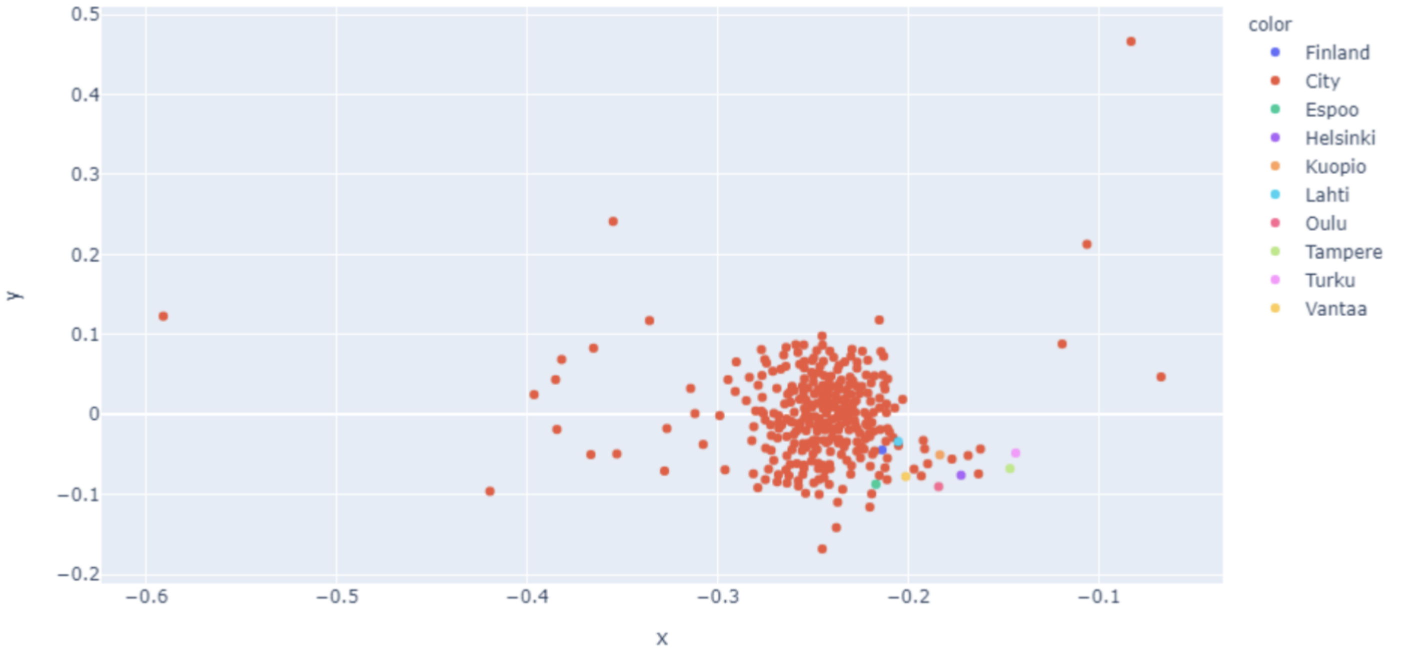

Figure 9. The plot of MMDS mapping in Finnish big cities

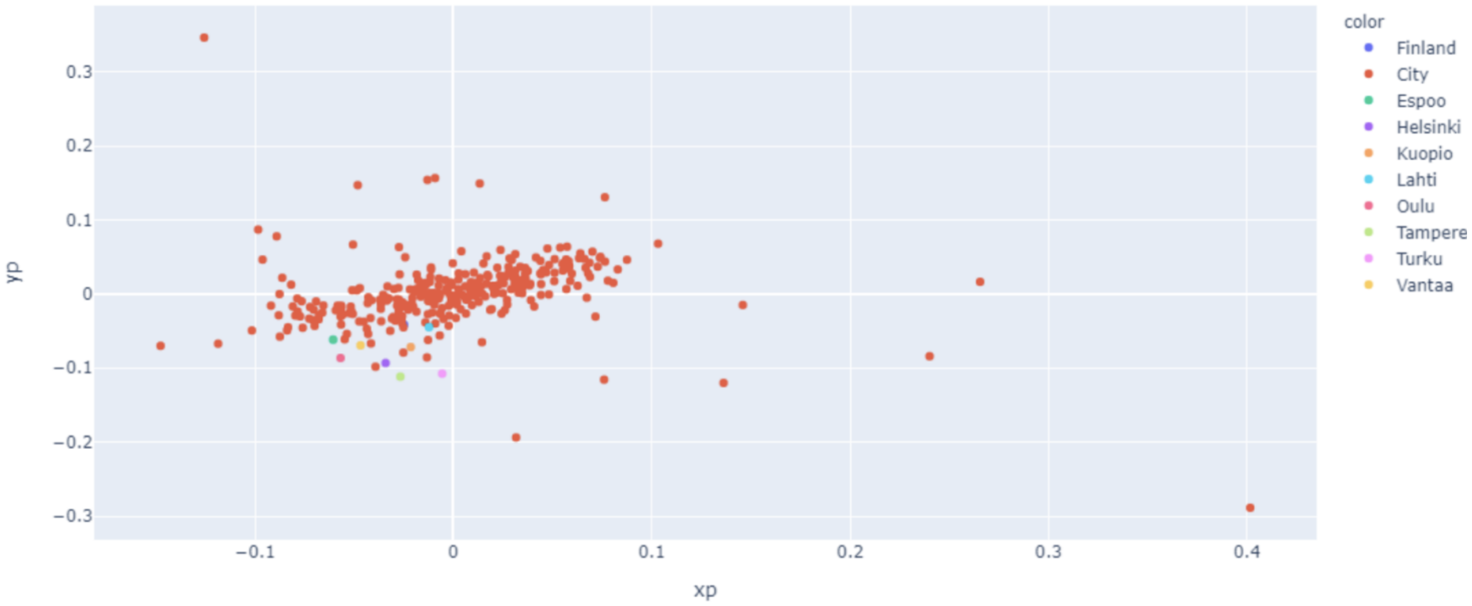

Figure 10. Sammon Mapping in Finnish big cities

The Shepard plots for MMDS mapping and Sammon mapping are shown in figure 11 and figure 12 below. As can be seen, both of the Shepard plots have a similar structure to each other.

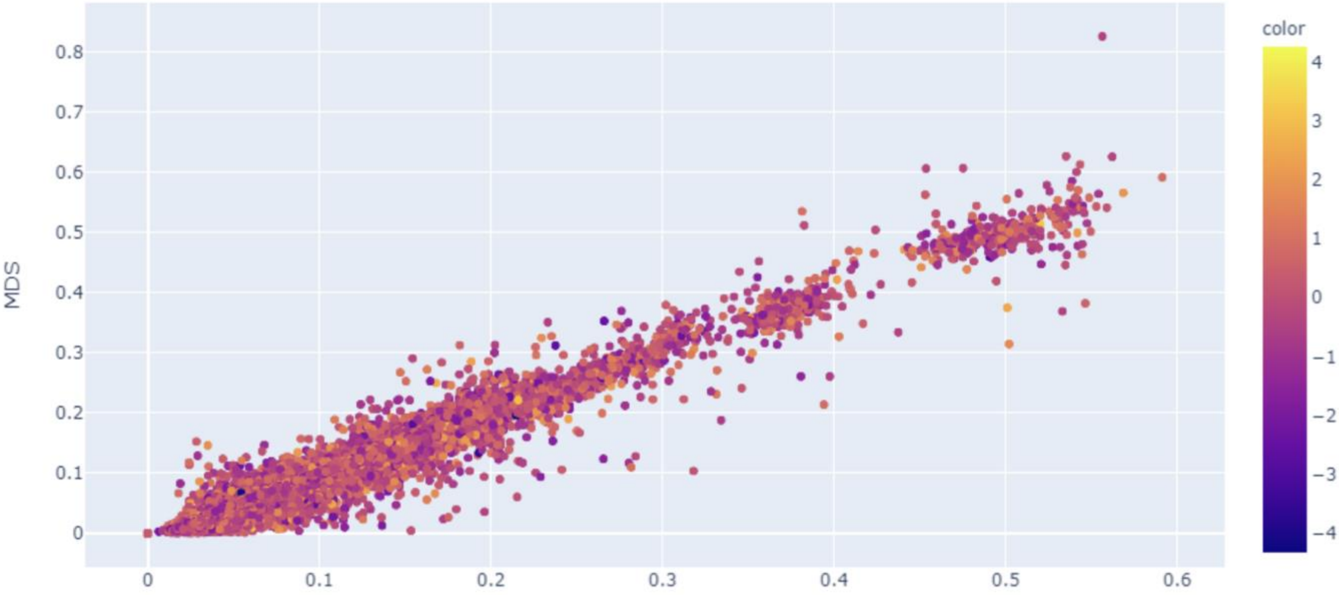

Figure 11. MMDS Shepard Plot

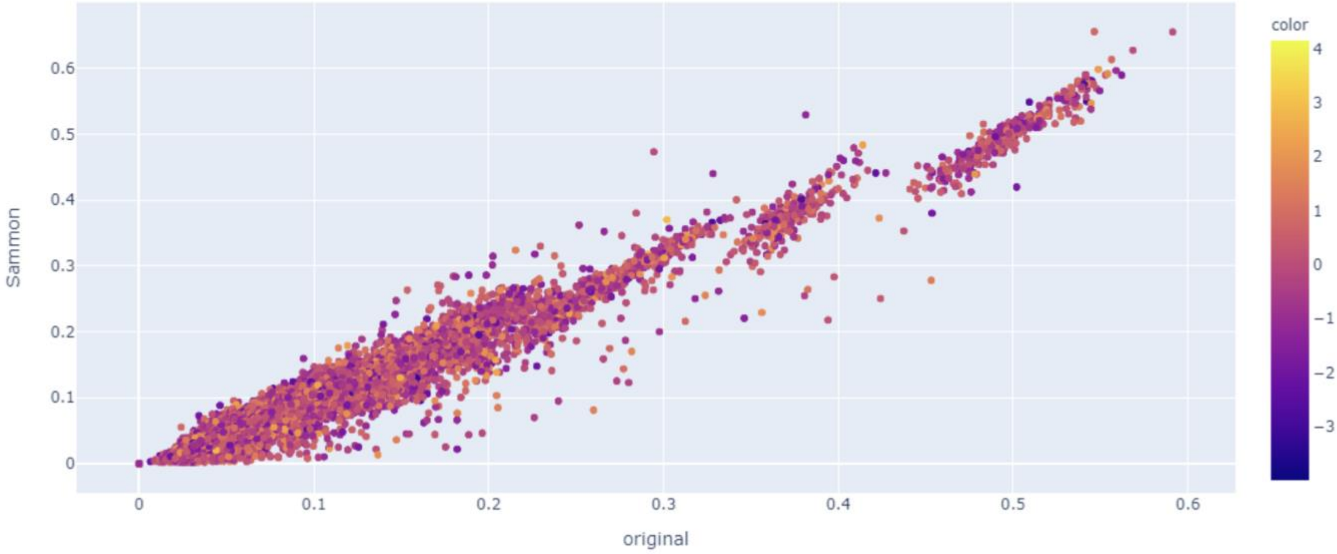

Figure 12. Sammon Shepard Plot

The MSE calculation was conducted using Python scikit-learn library. Scikit learn has returned the result as indicated in the table below

Method	MSE
MMDS Mapping	0.00018
Sammon Mapping	0.00014

Even though the plots seem similar to each other by visualization, MSE calculation reveals that Sammon Mapping is slightly better than MMDS Mapping. The MSE calculation of Sammon Mapping is lesser, which mean that the Sammon Mapping predicts the distance better.

Exercise 3

From My courses, download the dataset network data.tgf, which contains a network defined as an adjacency list (explained below). Visualize it using the principles introduced in the last lecture (and the general Tufte's principles taught earlier). Explain why your visualization is appropriate for this network and how you produced it. Also, visually indicate each node's given attribute labels and try to make different network substructures visible.

Answer

Figure 13 below shows the visualization of network data defined as an adjacency list. The visualization was made using Microsoft Visio software.

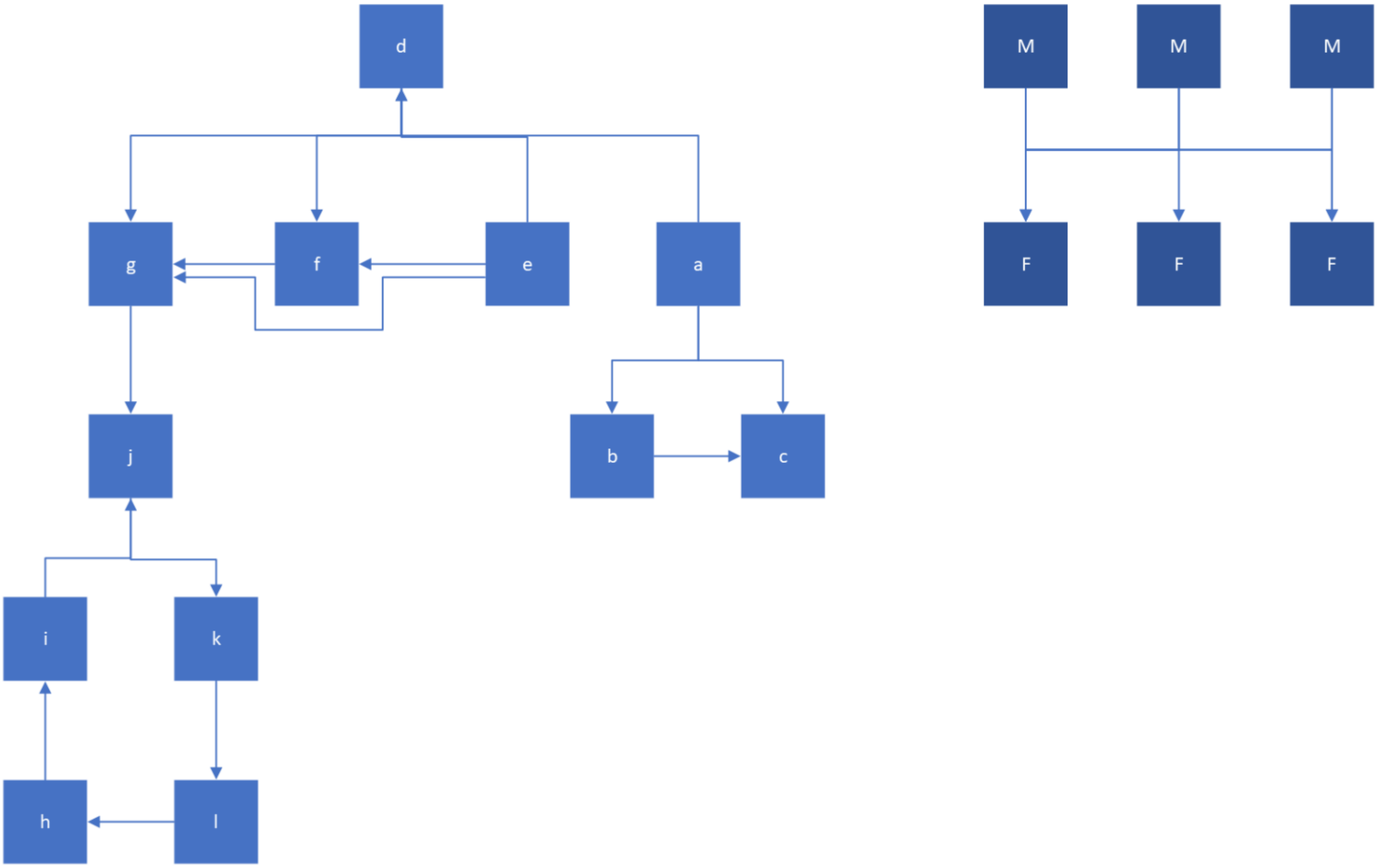

Figure 13. Visualization of the network defined as an adjacency list

As can be seen from the visualization, the network has two distinguished subgraphs. Since each node has a relationship with other nodes, a tree layout is used to visualize the network's adjacency.

The process of creating the tree layout is firstly to find the most frequent node which appears in the network. In our first subnetwork, the most frequent node is d. Thus, it is chosen as the root node. From the root node, we start establishing its relationship with other root's neighbor nodes. Afterward, we continue expanding the network further by including neighbor nodes of all existed nodes in the network. In the end, we got the visualization of the first subnetwork.

Meanwhile, in the second subnetwork, the appearance frequency is all similar in all the nodes. We also notice that all the M nodes are connecting to F nodes. Thus, all M nodes are chosen as root nodes in the second subnetwork.

The graph also follows Tufte's principles such as data-ink, chartjunk, aesthetics and techniques. Data-ink contains empty space and ink inside the visualization. It holds a non-erasable and non-redundant core inside the graphics. We see that if any part of the graph is removed, the visualization will miss some critical information. Thus, the information delivery is affected. In other words, the graph obeys the data-ink rule in Tufte's principle.

The graph also does not contain any interior decoration of the graphic that does not tell the viewer anything new. All pieces are meant to deliver new information. Thus, chartjunk principles are also followed.

By using the tree layout in this visualization, nodes and edges are evenly distributed and the edge bending ratio is minimized. Also, the edge crossing and edge length are minimized in both visualizations to help readers detect the relations among different nodes faster. Thus, the aesthetic and techniques principles are also followed at this point.