Course Project for Data Visualization

Summer,2020

Project Report

Dipkumar Patel

Noopa Jagadeesh

Prasanth Varma

Contents

Abstract

Visualization of any data is crucial as it helps to understand and analyze the information in a better way. The project deals with the visualization of the given dataset and understanding it by finding out the mapping between the decision space and objective space. Using the techniques like radviz, dendogram etc., we have tried to give the information related to the dataset. We have also achieved multi objective optimization by finding the objective function that could be minimized.

Introduction to the Visualization Techniques

Radviz

It is one of the best techniques as we can visualize more variables in 2-D projection. The variables are placed on the circle and are spaced equally. Observations are shown as points within the circle, and their positions are determined by a physics metaphor: each point is held in place with springs attached to the variable anchors at the other end. After the placing the points inside the circle, we can infer that the points which are closer to a variable have higher values than that of others.

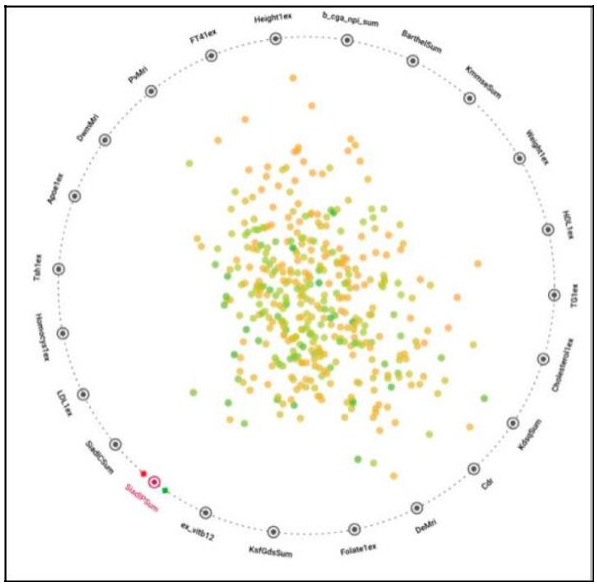


Fig 1: Sample Radviz Visualization

The main disadvantage with radviz is that if the number of variables is more, then the points tend to merge at the center making it difficult to interpret the data.

Dendogram

A dendrogram is a diagram showing the relation between the objects in the hierarchy. It is most commonly generated as a Hierarchical Clustering output. A dendrogram is mainly used to work out the best way of allocating objects to clusters. The height of the dendogram is very much essential to interpret. Also, clustering of the information can easily be understood with this technique.

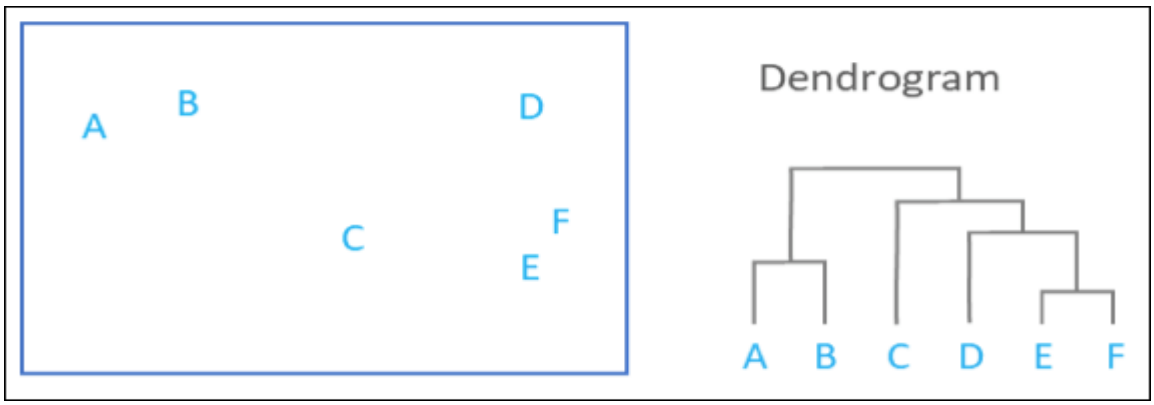


Figure 2: Sample Dendogram

While it provides the information regarding the clustering, it is not suitable for huge data; especially when it has mixed data types.

Heatmaps

A heat map (or heatmap) is a graphical representation of data showing values by colour. Heat maps make it easy to imagine and grasp complicated data in one glance. It can be visualized using python seaborn library. These are easy to build and can be used to visualize higher dimensional data.

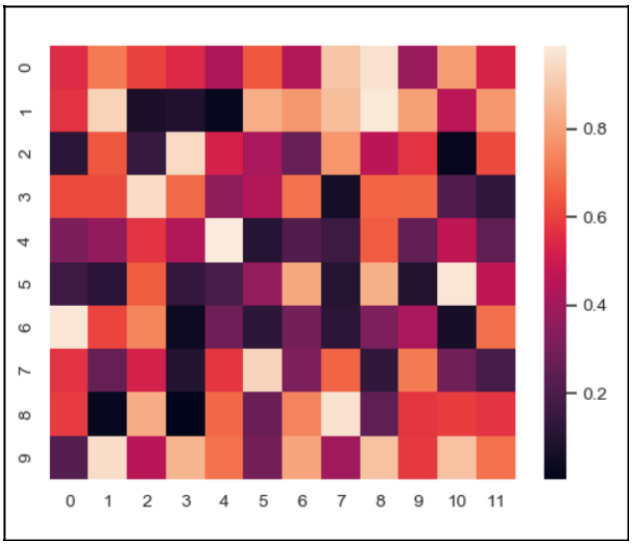


Figure 3: Sample Heatmap

Parallel Coordinates

It is mainly used as we can show multiple variables on a parallel coordinate and the values of the variable are plotted on the axis. Values are plotted as a series of lines connecting across every axis. That means each line is a collection of points placed on each axis, all connected together. The advantage of using this is that it can compare the datapoints easily and can help in discovering patterns or correlations across variables.

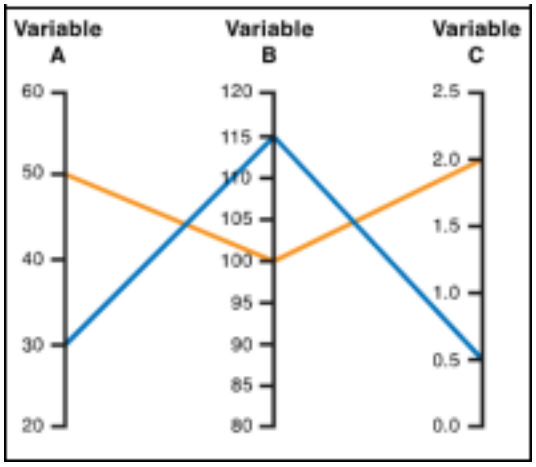


Figure 4: Sample Parallel Coordinate

The main behind the selecting these visualization techniques is that, they are very much helpful in visualizing high dimension data and are also helpful in clustering of the given dataset. In the next section, we will look into the given dataset description.

Dataset Description

The given dataset contains 7 subsets of data and few of the observation values are greater than 1 and had to be normalized to understand the information better. Totally, there are 573 observations in the given dataset. They contain 100 dimensions in each subset, while the objective function data varied. There are 3 subsets which had 3 objective functions, 2 subsets which had 5 objective functions and 2 subsets which had 10 objective functions.

Results

Majority of real-world problems are multi-objective problems which have multiple objectives. For example, we may want to optimize the design of the vehicle with the objectives speed, ride-comfort, handling, price, fuel-efficiency and safety. This becomes a 6 objective problem because there are 6 objectives that we wish to optimize at the same time. Each solution now has 6 objective values indicating how well it performs for each of our objectives. This introduces many difficulties, but the most significant difficulty is how do we decide if one solution is better than the another? All the 7 given dataset has decision space and objective space. Decision variables are used to model a given complex problem and objective variables reflect the objective functions used.

DTLZ2: No of Objective =3

Visualizing the Decision space

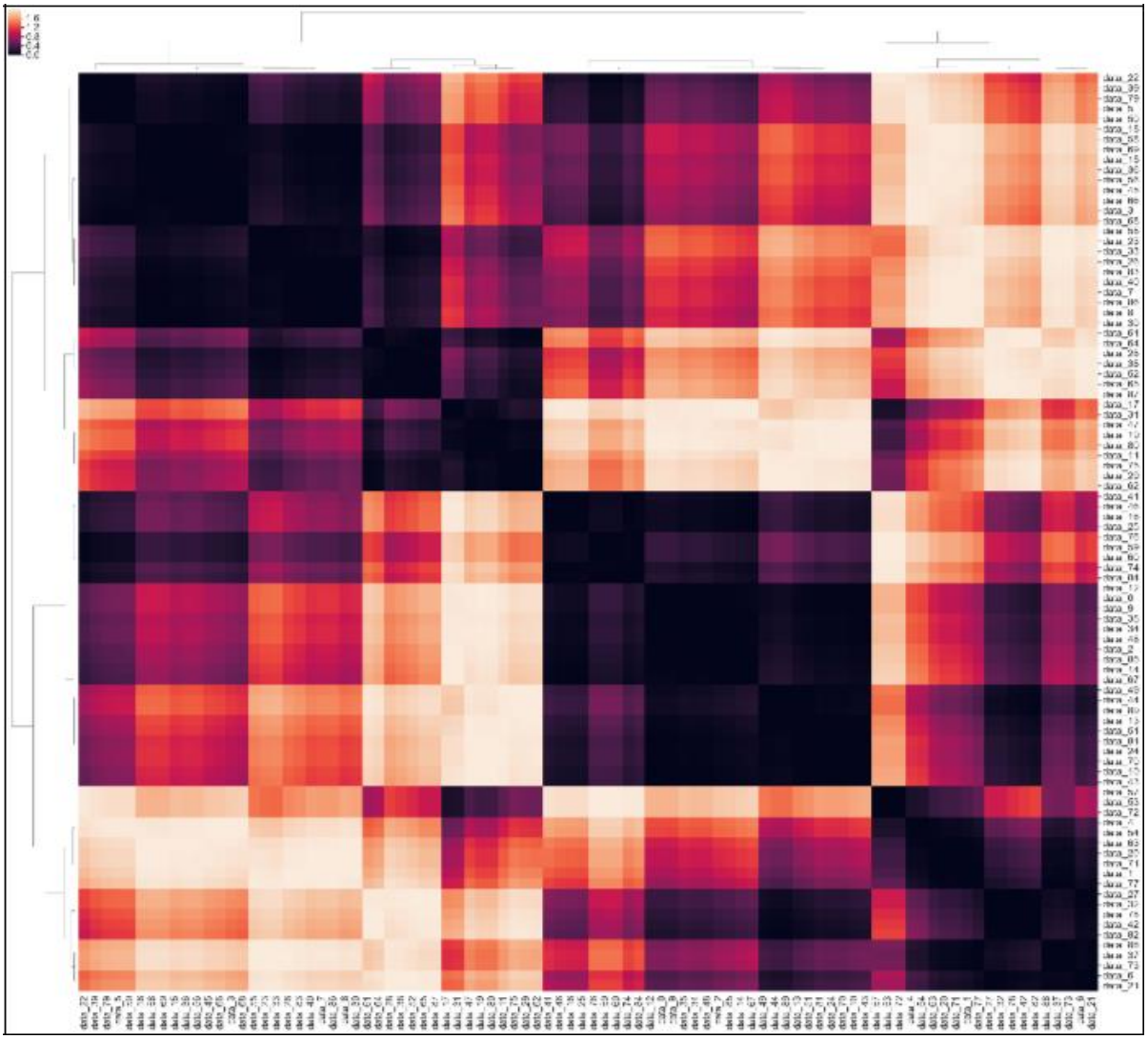


Figure 5: a. Clustermap

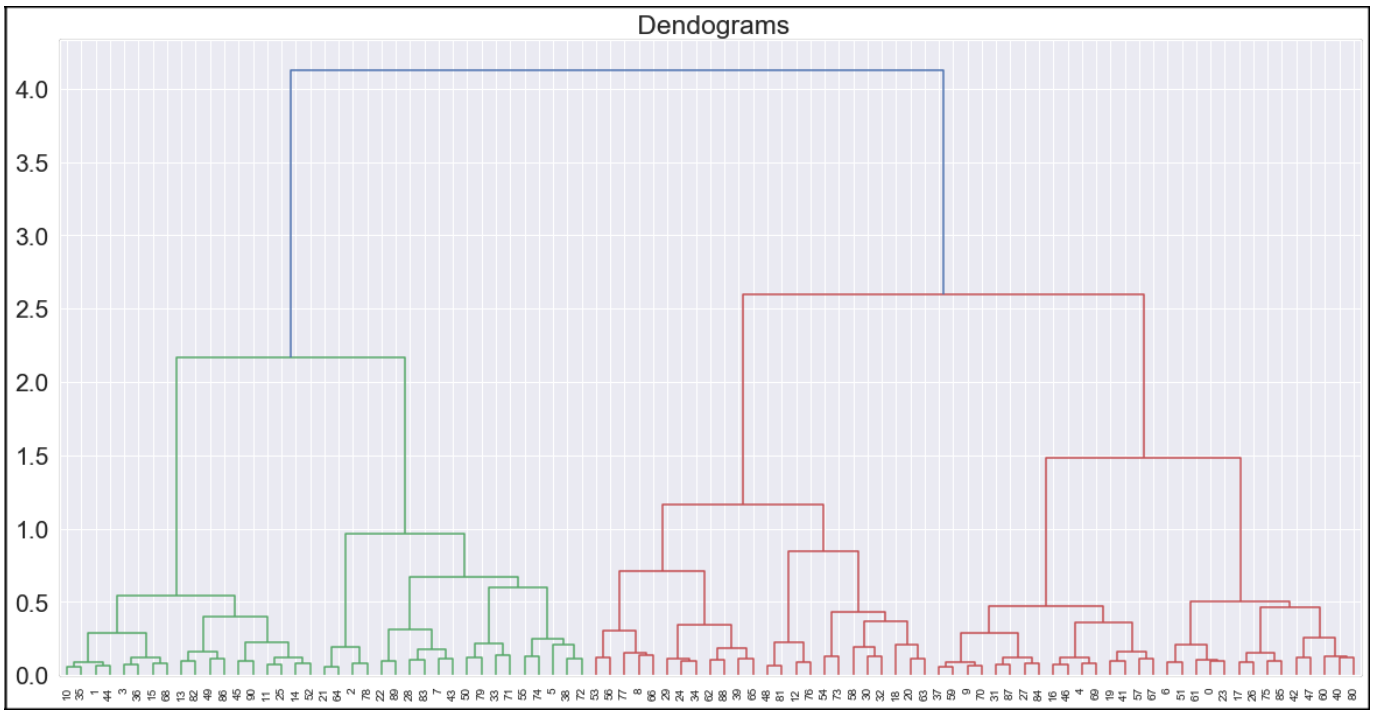
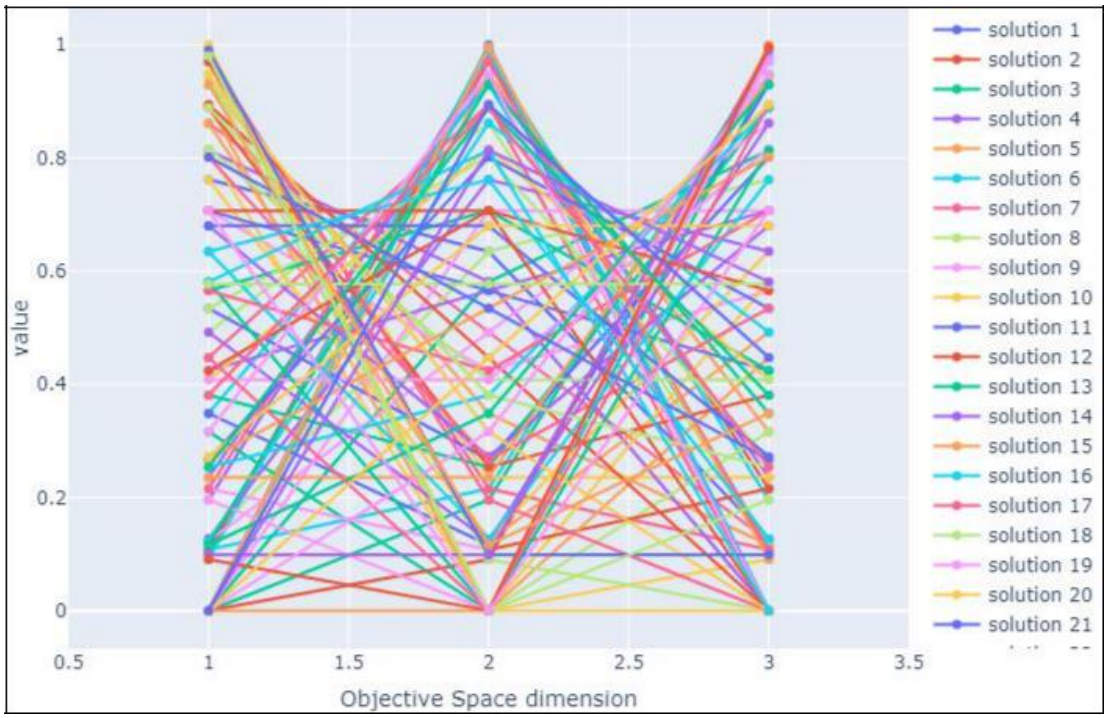
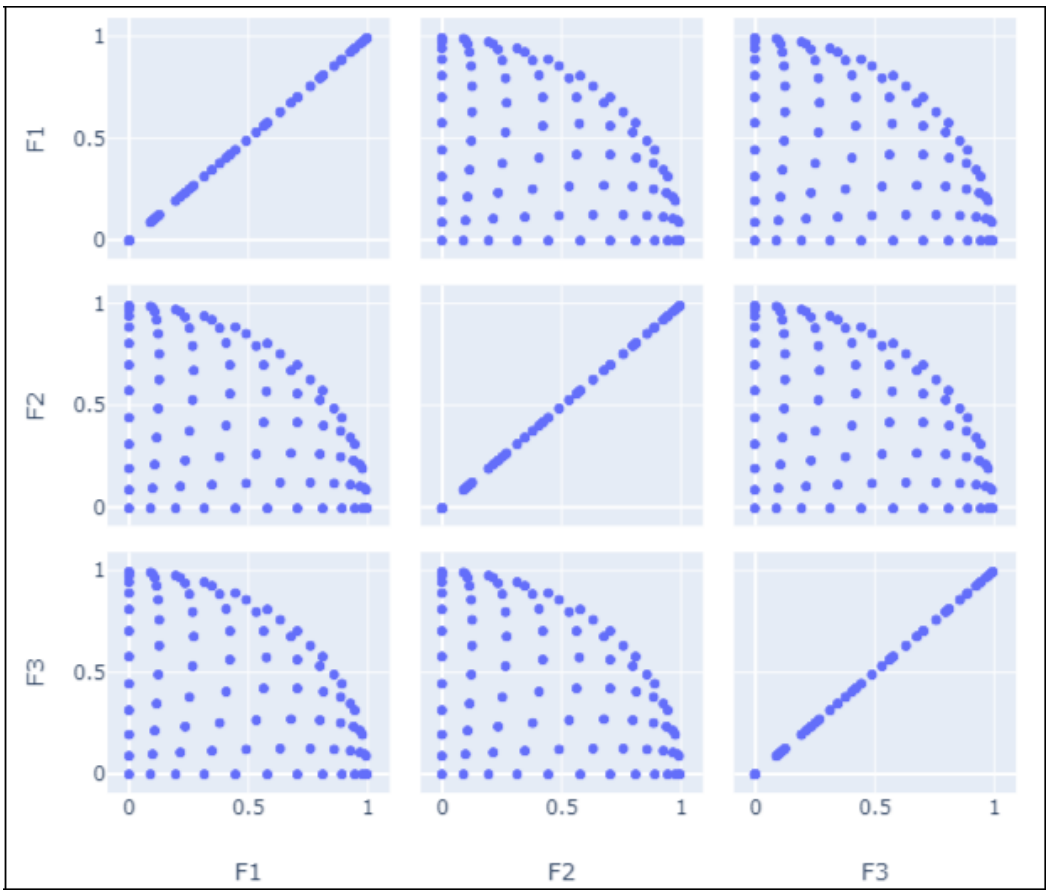


Figure 6: Dendogram

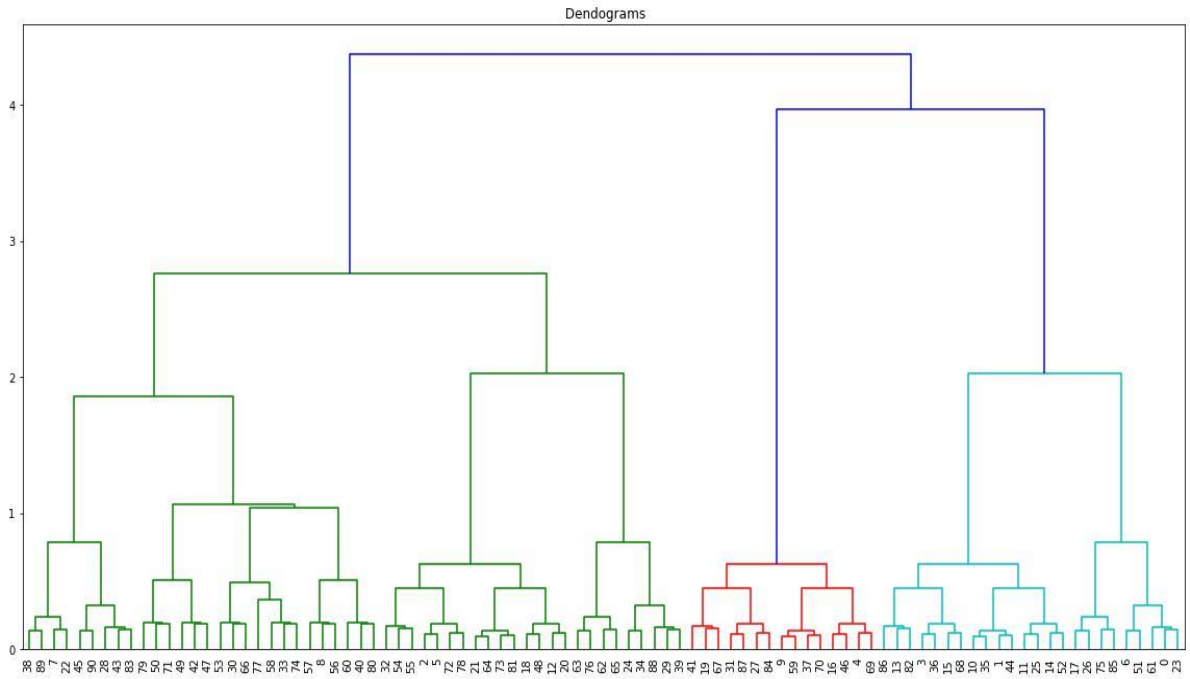
Visualizing the Objective space



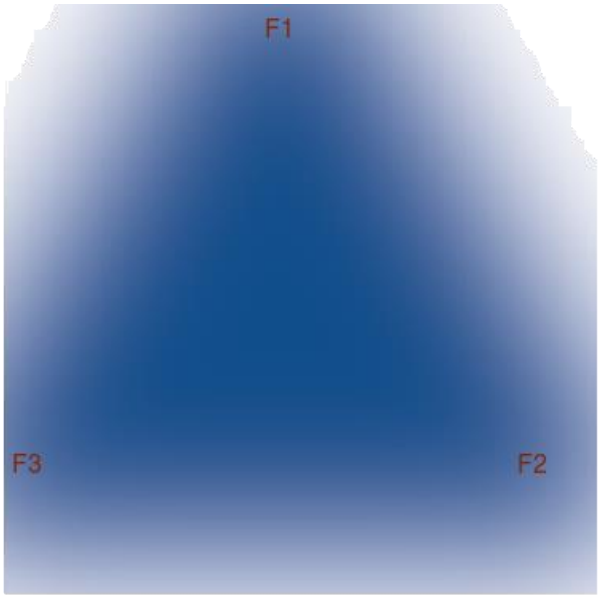
a



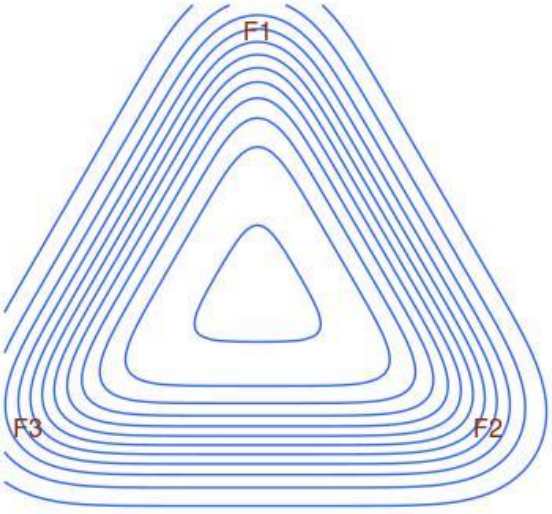
B



c



D



e

Figure 7: a. PCP b. PSP c. Truncated dendogram d. RadVis e. Contour plot

MaF3: No of Objective =3

Visualizing the Decision space

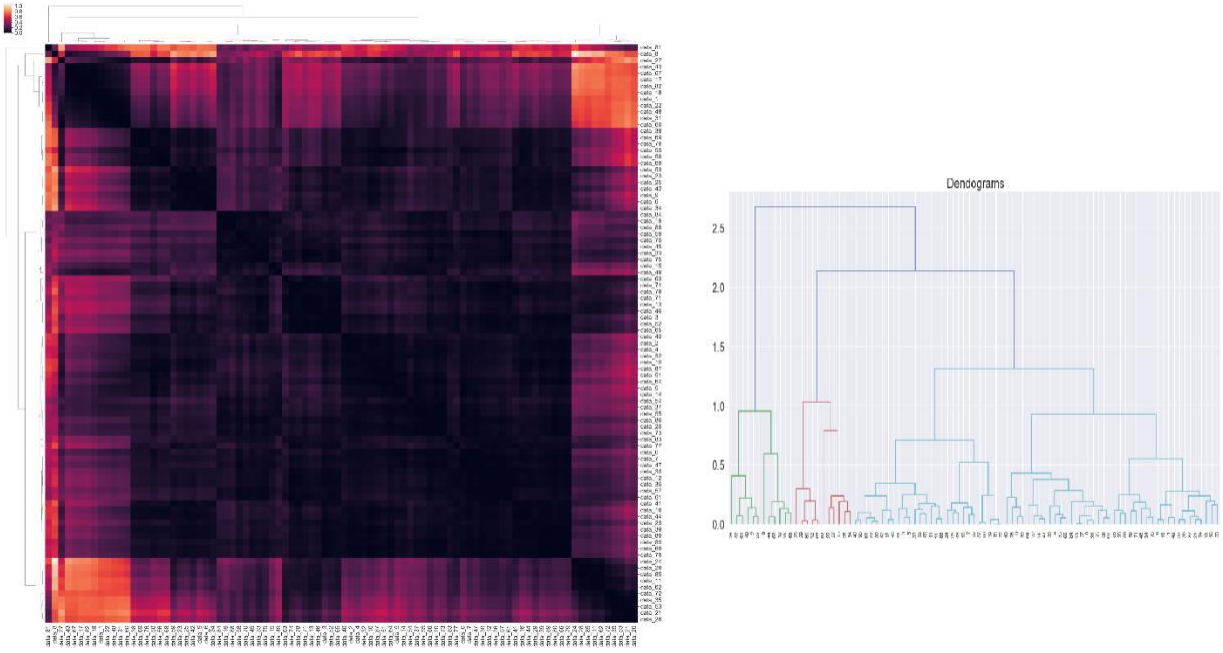


Figure 8: a. Clustermap b. Dendogram

Visualizing the Objective space

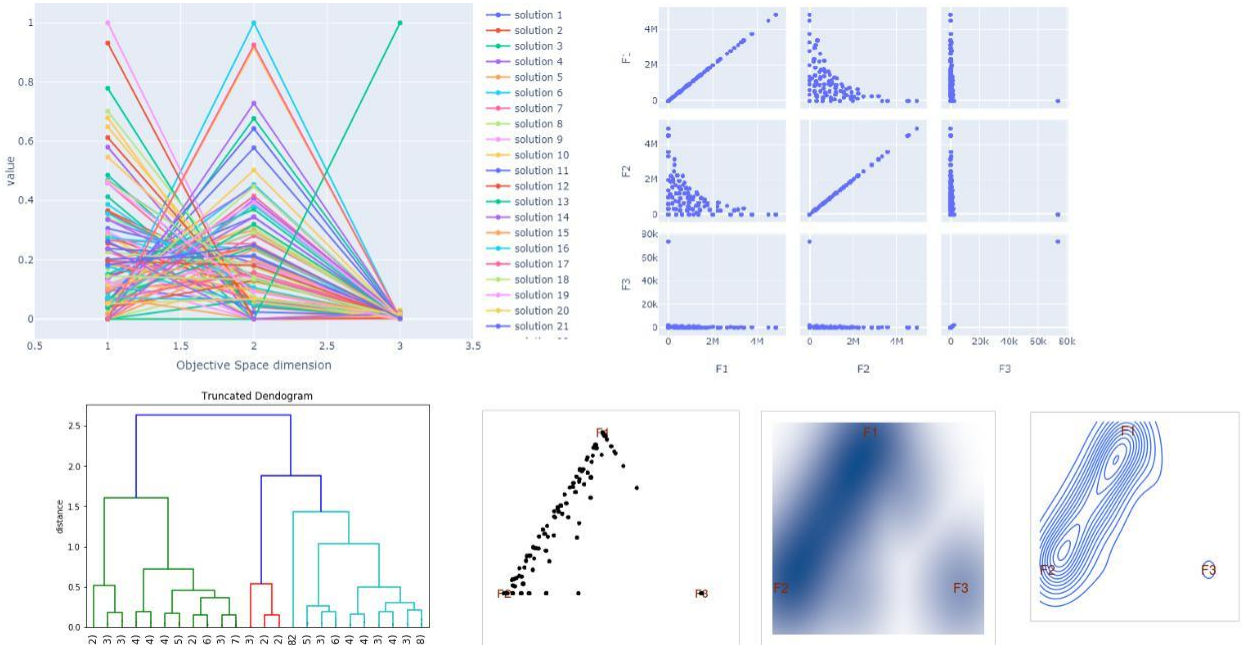


Figure 9: a. PCP b. PSP c. Truncated dendogram d. RadVis e. Density Plot f. Contour plot

MaF3: No of Objective =5

Visualizing the Decision space

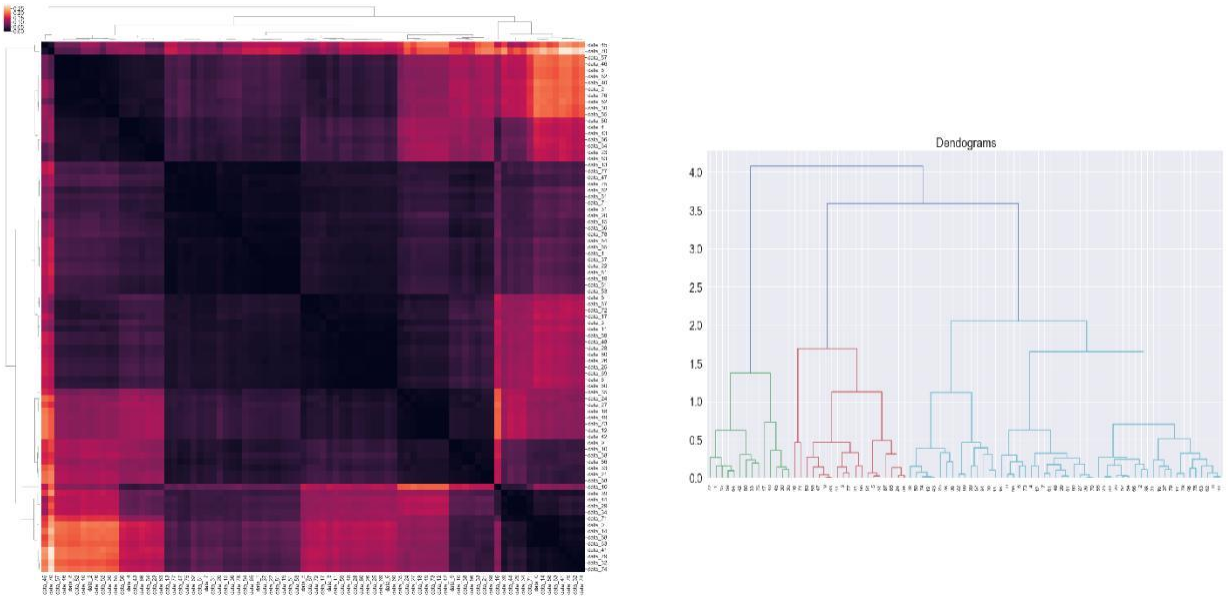


Figure 10: a. Clustermap b. Dendogram

Visualizing the Objective space

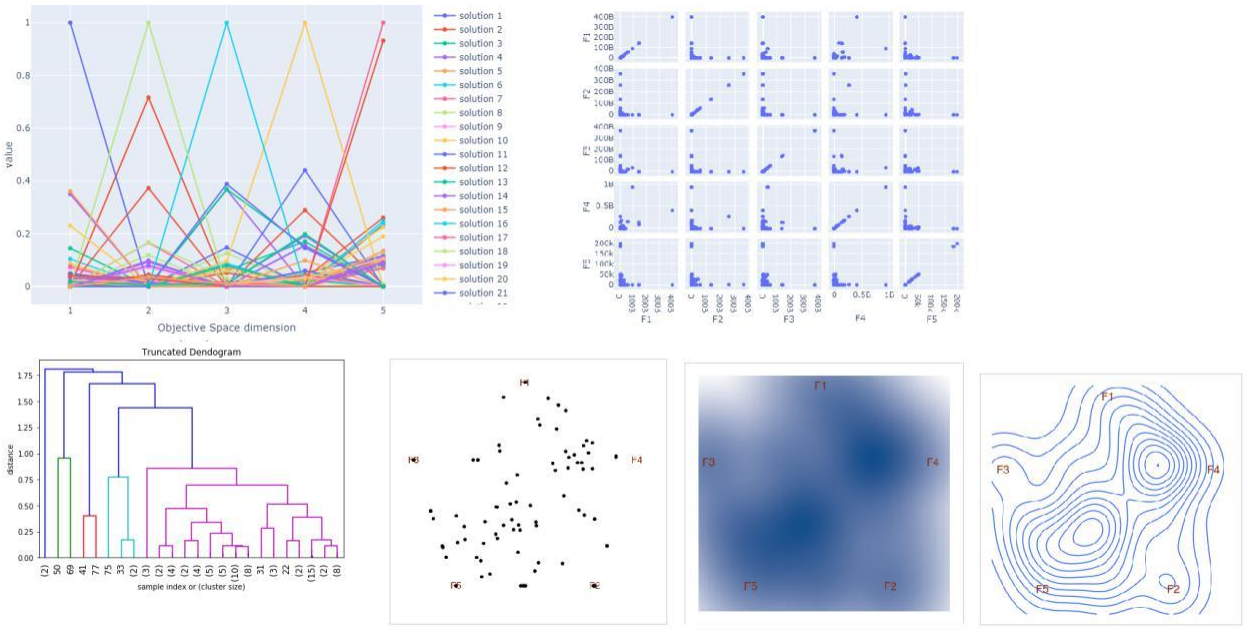


Figure 11: a. PCP b. PSP c. Truncated dendogram d. RadVis e. Density plot f. Contour plot

MaF5: No of Objective =10

Visualizing the Decision space

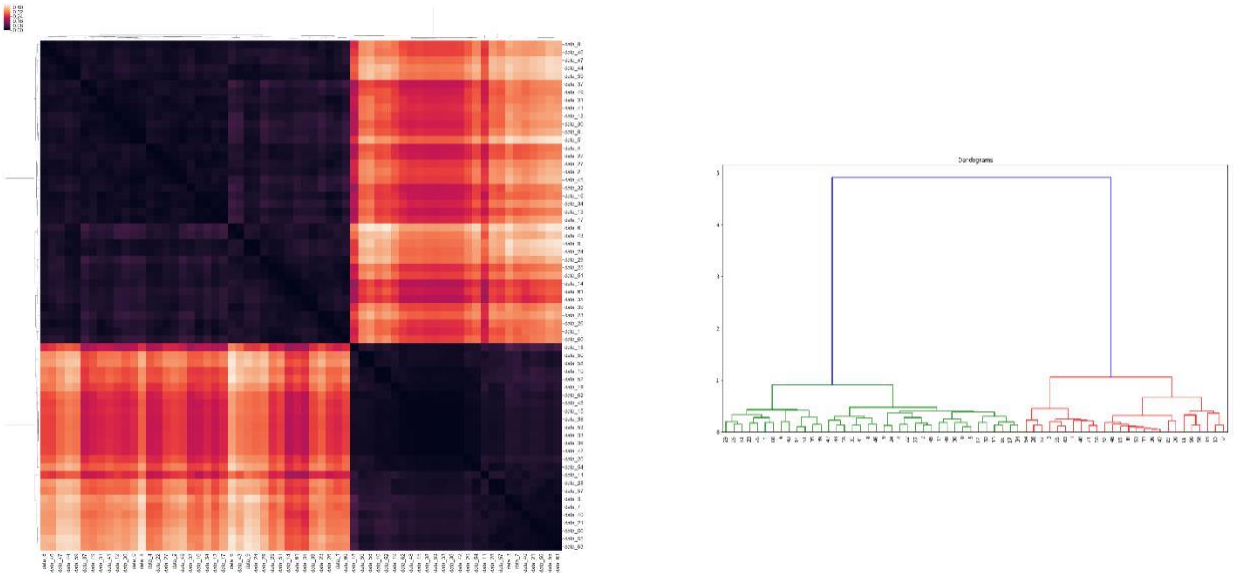


Figure 12: a. Clustermap b. Dendogram

Visualizing the Objective space

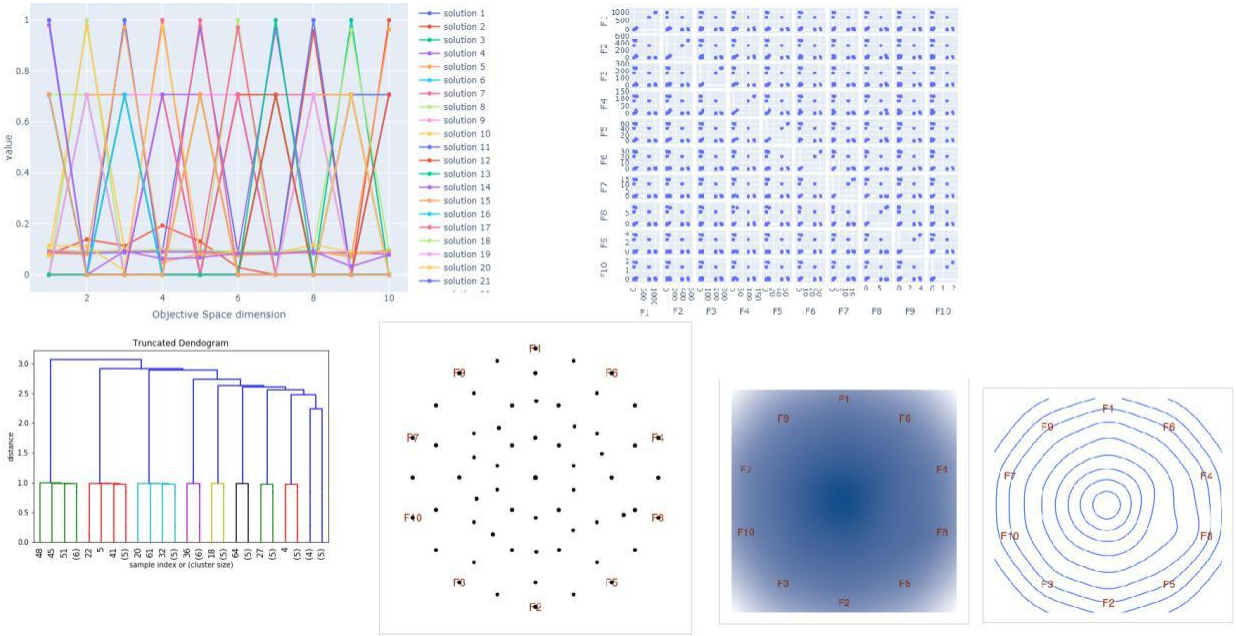


Figure 13: a. PCP b. PSP c. Truncated dendogram d. RadVis e. Density plot f. Contour plot

DTLZ6: No of Objective =10

Visualizing the Decision space

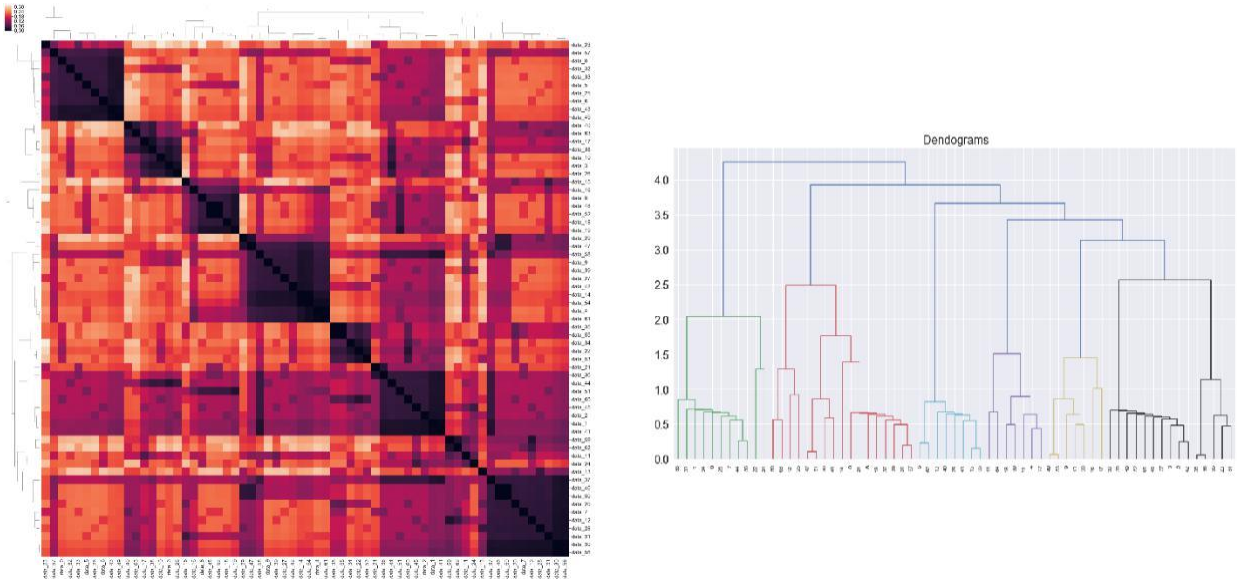


Figure 14: a. Clustermap b. Dendogram

Visualizing the Objective space

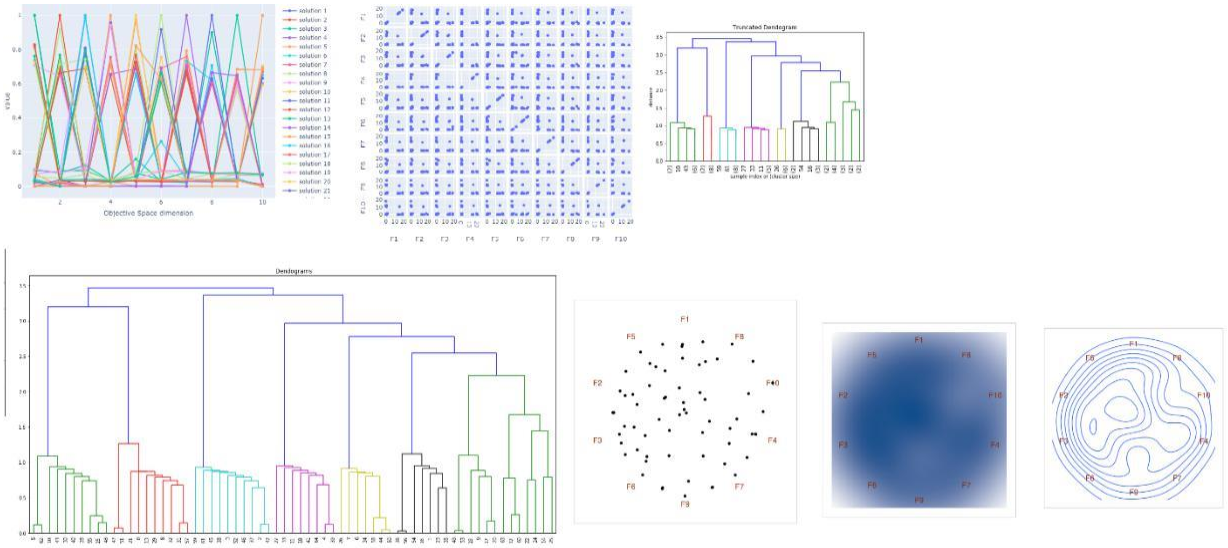


Figure 15: a. PCP b. PSP c. Dendogram d. Truncated dendogram e. RadVis f. Density plot g. Contour plot

MaF2: No of Objective =3

Visualizing the Decision space

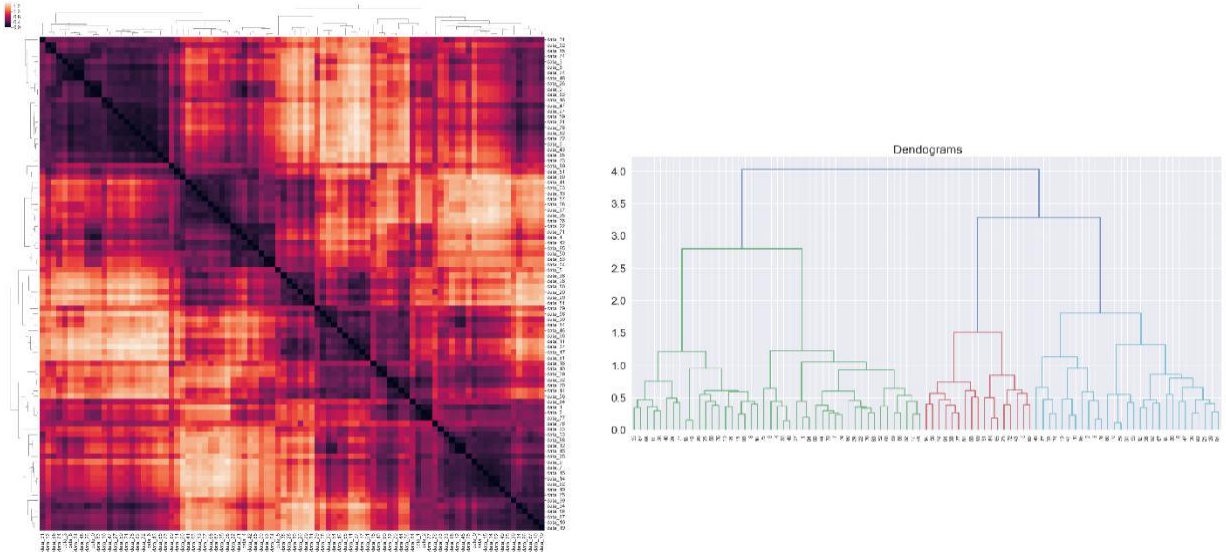


Figure 16: a. Clustermap b. Dendogram

Visualizing the Objective space

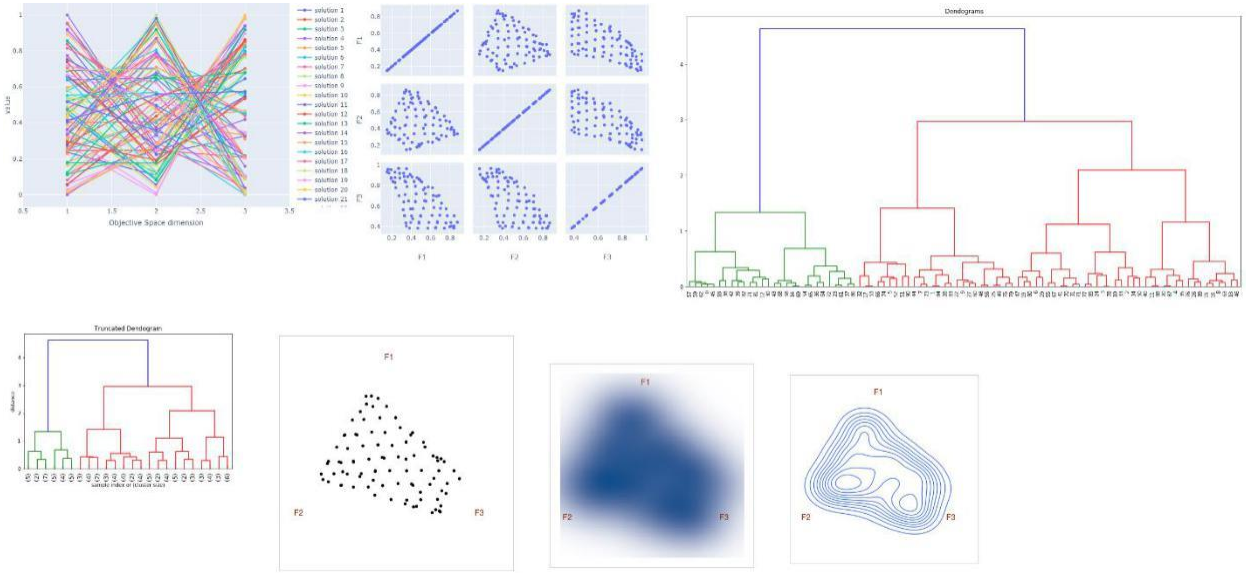


Figure 17: a. PCP b. PSP c. Dendogram d. Truncated dendogram e. RadVis f. Density plot g. Contour plot

MaF2: No of Objective =5

Visualizing the Decision space

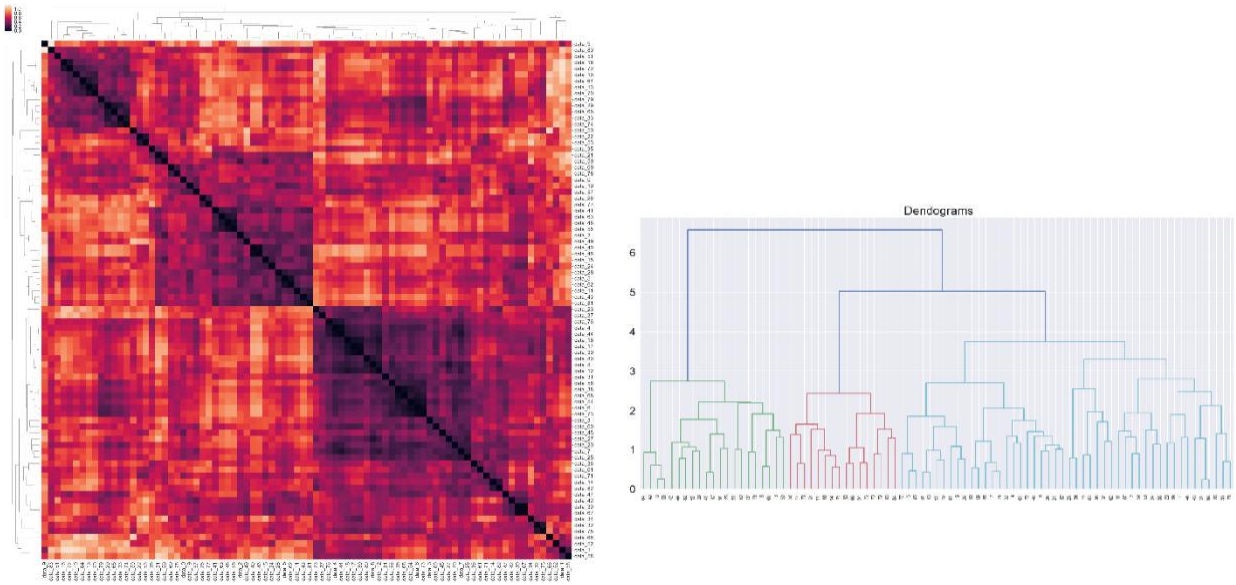


Figure 18: a. Clustermap b. Dendogram

Visualizing the Objective space

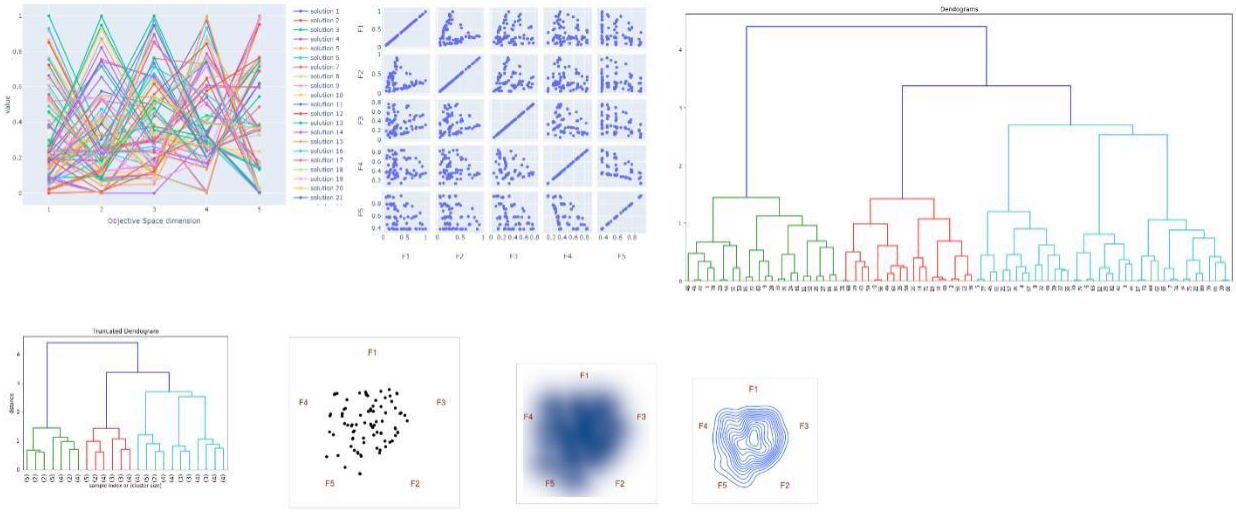


Figure 19: a. PCP b. PSP c. Dendogram d. Truncated dendogram e. RadVis f. Density plot g. Contour plot

Limitations

1. In the radviz visualization (for 10 variables subset), we cannot see the points tending towards a variable. So, it will be difficult for us to interpret the higher values of a variable in the given dataset.
2. Though using parallel coordinates, we can visualize the data without any loss, it is very difficult to interpret for datasets with large number of objective functions.

Conclusion

We have proposed the use of two visualization techniques to understand the clustering in decision space using clustermaps and dendogram. The clustering in objective space is done using the techniques parallel coordinates, pairwise scatter plot, dendogram, radviz, density plot and contour maps. The mapping between decision and objective space has been achieved. Clustering were also visible in all decision space and objective space. We have achieved multi object optimization for datasets.