Datathon 4

Sushranth Hebbar

Introduction :- Use tabular datasets published by the World Health Organization to explore different matrix seriation methods.

Methods :- The approach to this task can be divided into 2 phases. The first phase is data preprocessing and the second is matrix seriation. The first step consists of computing a distance matrix from the given data. The second step focuses on seriating this matrix using libraries in R.

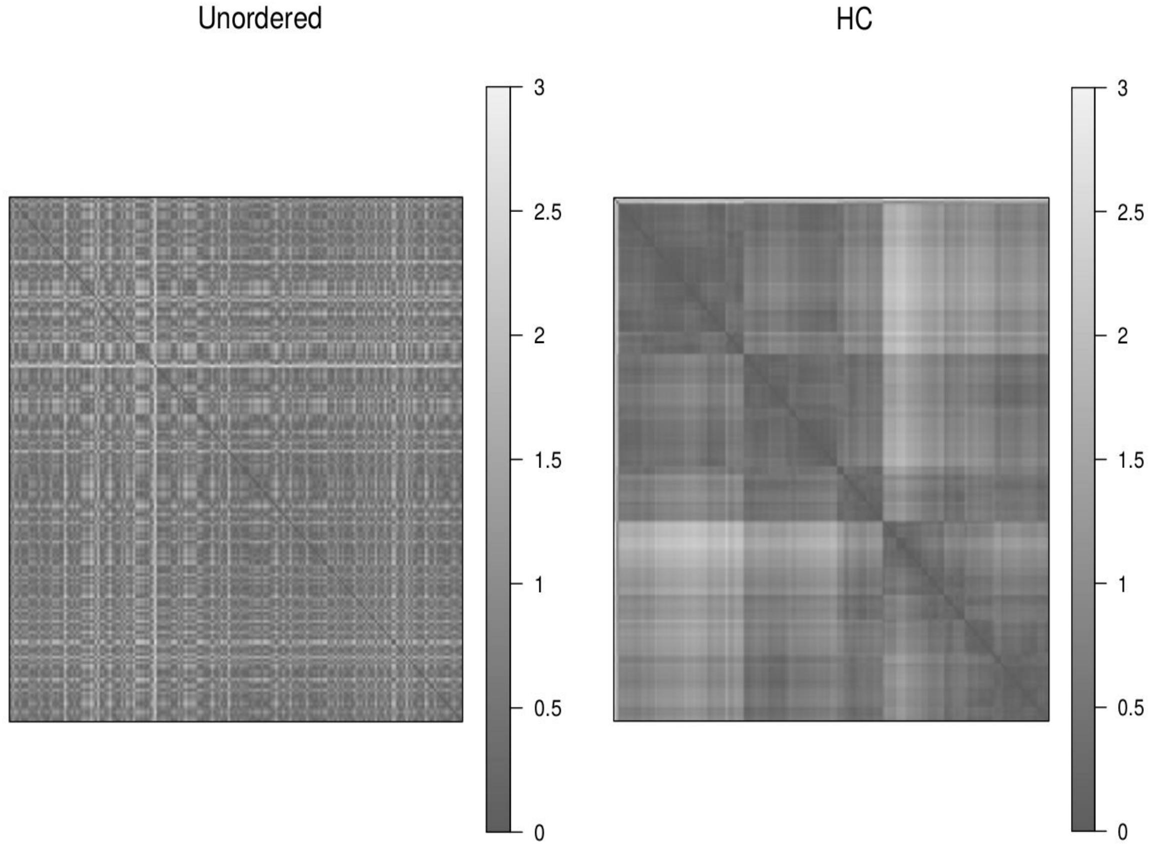
Preprocessing :- The covid_19_data.csv was used to generate the visuals. This is because this file contains the least amount of missing values and has time series data on the number of active cases and deaths country wise between January 22 to September 23. First, the dataset was loaded into a pandas dataframe. Then it was grouped based on the observed date. Then the confirmed cases for each province was summed up for every country for a particular time stamp was computed and stored in a dictionary. Then the square root transformation was applied on these values to ensure the stability of these values. Finally, they were converted into a dataframe and then the values were normalized.

Then the distance matrix was computed from the normalized matrix where each node in the matrix represents a country. The values in this matrix are positive. Two nodes which are similar to each other will have low distance value and vice-versa.This was later saved into a text file. It was then loaded and 5 different seriation methods were used on this matrix to observe patterns on this data. The scale in the image represents the distance value. Standard libraries from R were used as it generates results within a second as opposed to manually creating an implementation which would be slow and inefficient.

The five seriation methods used were :-

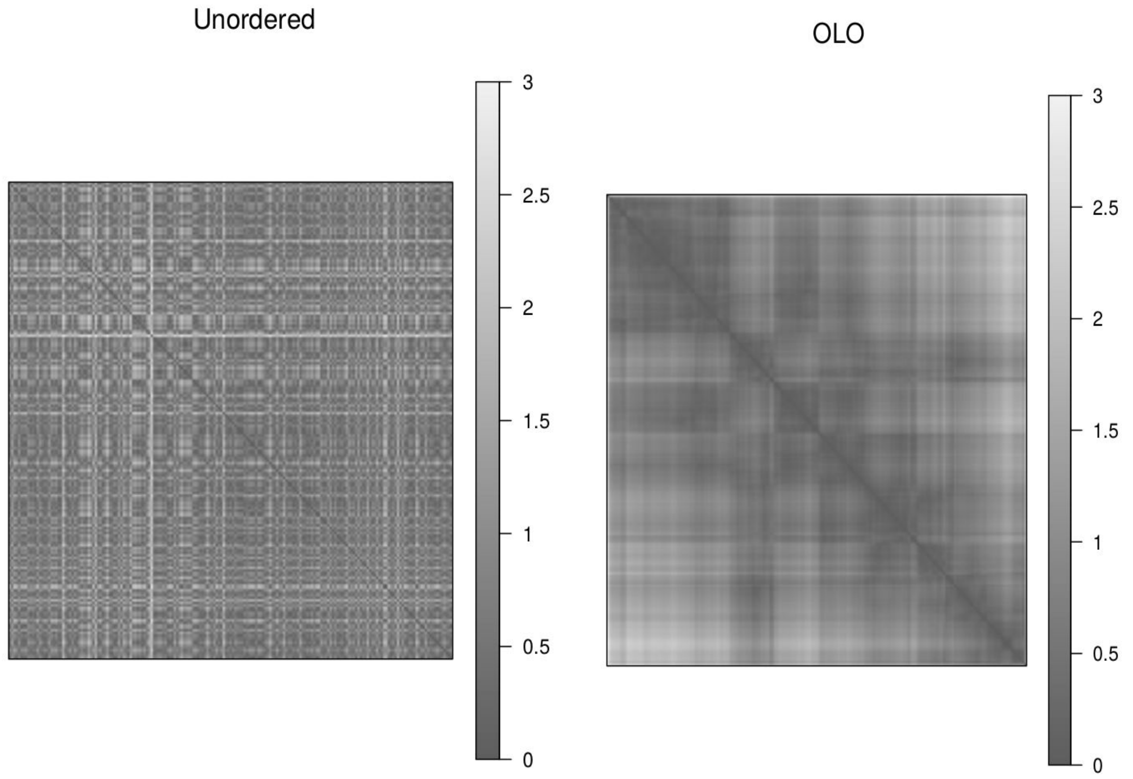
- Hierarchical Clustering
- Optimal Leaf Ordering
- Travelling Salesman Problem
- Rank Two Ellipse Seriation
- Bond Energy Algorithm

Hierarchical Clustering:-



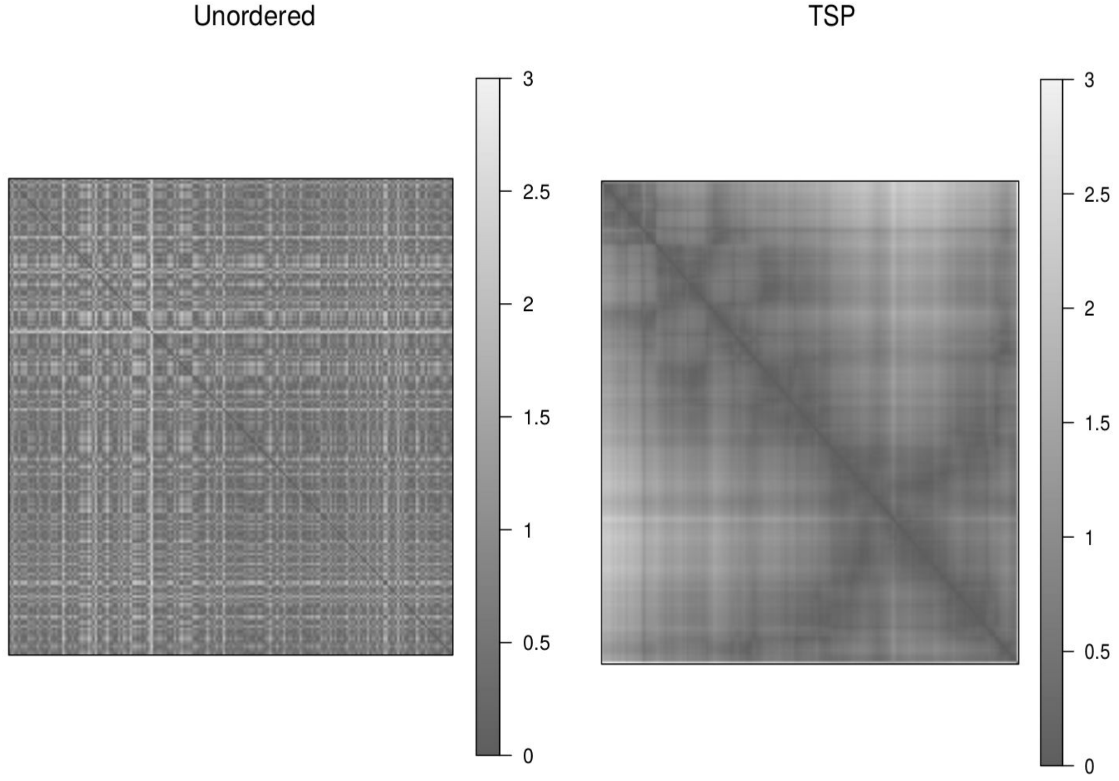
This algorithm is useful for deriving clusters of similar data elements. This algorithm aims at producing grouping patterns from the data. In the above figure on the right, there are dark boxes along the main diagonal. Dark boxes indicate nodes which are similar to each other and which are likely to form connected components in the graph. From the figure we can notice that every dark box can be subdivided into smaller sub boxes indicating smaller communities inside larger ones.

Optimal Leaf Ordering :-



In addition to the above mentioned clustering methods, smoothing the clusters by ordering the vertices according to the neighboring distance values reveals the structures more clearly and smoothens the transitions from one connected component to another. This algorithm reveals the connection between different clusters. An optimal ordering is computed globally, so as to minimize the sum of distances between successive rows while traversing the clustering tree in depth-first order.

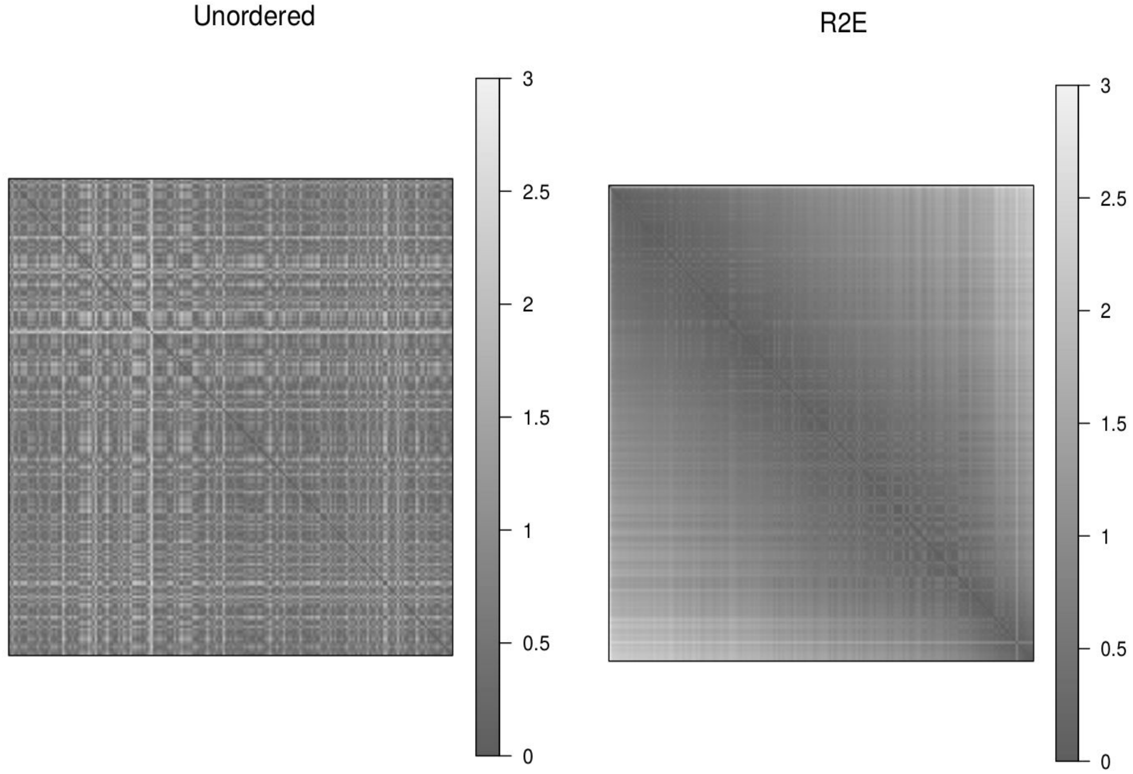
Travelling Salesman Problem:-


Unordered

TSP

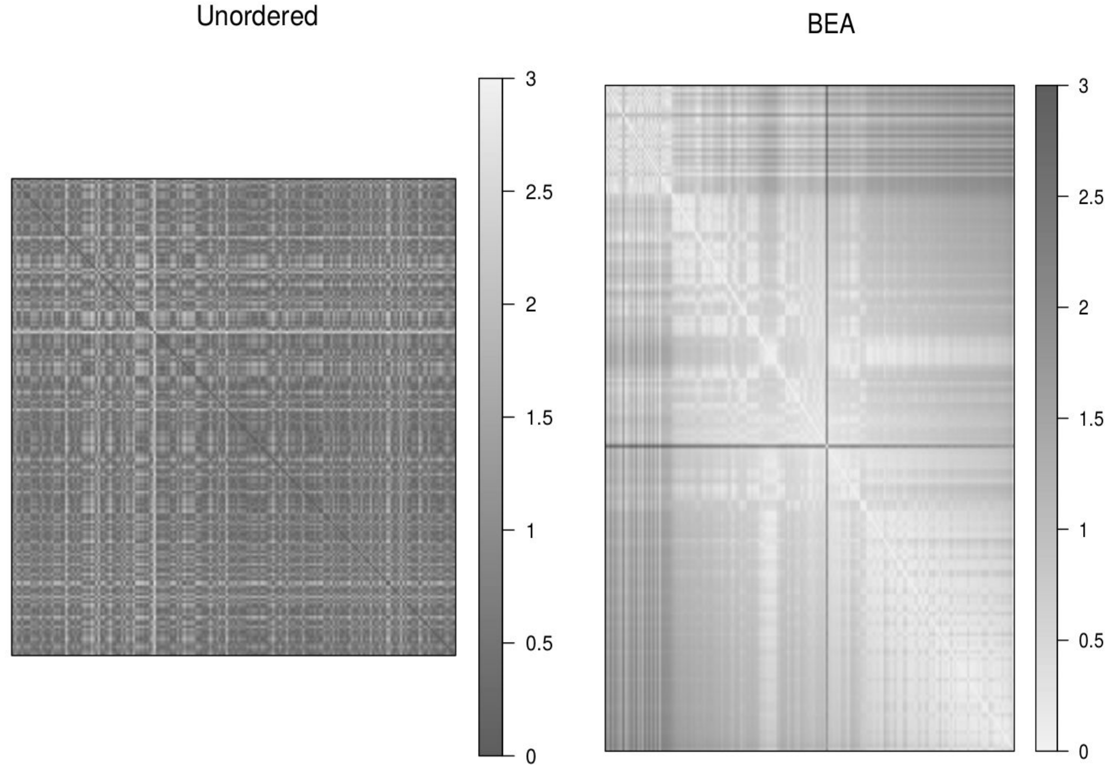
This falls under the graph theoretic approach where the permutation problem is transformed into the space of graph enumeration. The central idea of graph-theoretic approaches is to exploit the graph structure for computing a linear order that optimizes a graph-theoretic layout cost function. TSP matrix reordering approaches model each row, respectively column, as a city and translate the row/column-wise similarity into virtual distances. From the above diagram, we can see that the number of clusters is less compared to the previous diagrams. So, TSP tends to reveal local patterns but may fail to optimize for the whole matrix.

Rank Two Ellipse Seriation:-



This is a spectral method. The eigenvectors and eigenvalues are computed from the matrix. The first few or the last few vectors are used to construct a permutation. These types of approaches are sensitive towards the noise in the data. Outliers, missing values, distributions of the data can affect the quality of the visual generated. The assumption is that the core matrix structure can be extracted from only a few dominant dimensions. Similarly connected nodes tend to be closer in the eigenspace. So, clusters can be detected using spectral approaches.

Bond Energy Algorithm:-



This algorithm falls under the heuristic approach. Heuristics are methods that transform the problem of finding an ordering into another problem space that abstracts the problem appropriately and allows for computationally efficient problem solving. Heuristic approaches transform the matrix reordering problem,such that specific assumptions are met. While problem simplification algorithms are usually fast, they suffer inherently from this restriction. If a dataset is not of the expected form, the results will be inappropriate for an analysis. From the above image it is apparent that the current dataset is not appropriate for this algorithm. It is hard to spot well defined clusters in the above image. So, this dataset does not simplify well which is why the results are poor.

References:-

https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12935
https://cran.r-project.org/web/packages/seriation/vignettes/seriation.pdf
https://cran.r-project.org/web/packages/seriation/seriation.pdf
https://www.r-project.org/other-docs.html