Lab 4 - Vizualisation

Karthikeyan Devarajan (Karde799)

9/23/2021

# Assignment 1

*1) For further analysis, import data to R and keep only the columns with the following numbers: 1,2,5,6,7,9,10,16,17,18,19. Use the first column as labels in further analysis.*

*2) Plot a heatmap of the data without doing any reordering. Is it possible to see clusters, outliers?*

It is very randomly grouped and not able to see clear clusters.  
*3) Compute distance matrices by a) using Euclidian distance and b) as one minus correlation. For both cases, compute orders that optimize Hamiltonian Path Length and use Hierarchical Clustering (HC) as the optimization algorithm. Plot two respective heatmaps and state which plot seems to be easier to analyse and why. Make a detailed analysis of the plot based on Euclidian distance. Use Euclidian Distance matrix in all coming steps.*

*3.1) Euclidean Distance*

The clustering can be easily viewed and Since euclidean distance is sum of squared difference. Euclidean distance works only for scaled data, the clustering is better for scaled data. The hours worked and the hours required for bread, big, Rice and Iphone 4s are inversely correlated in the last 10 cities and highly correlated in the top 10 cities.This is the opposite for the other variables such as Vacation days, clothing Index, Wage Net, and Food Cost.

The correlation distance matrix is basically average of two values.

*4) Compute a permutation that optimizes Hamiltonian Path Length but uses Traveling Salesman Problem (TSP) as solver. Compare the heatmap given by this reordering with the heatmap produced by the HC solver in the previous step – which one seems to be better? Compare also objective function values such as Hamiltonian Path length and Gradient measure achieved by row permutations of TSP and HC solvers (Hint: use criterion() function)*

The Hamiltonian Path Length optimization using Travelling Salesman Problem is better than the euclidean distance measurement.

*TSP and Gradient Measure criterion*

## The Criterion by using TSP is 27958 120.9386

## Registered S3 method overwritten by 'gclus':  
## method from   
## reorder.hclust seriation

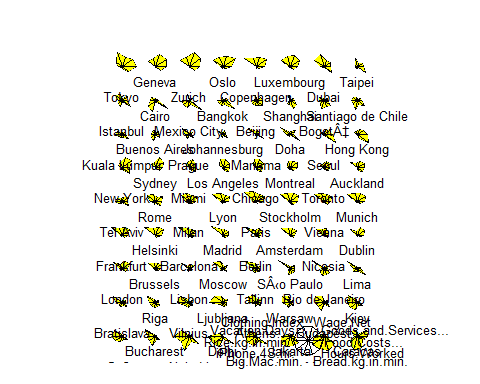
## The Criterion by using TSP is 41612 127.3152

*5) Use Ploty to create parallel coordinate plots from unsorted data and try to permute the variables in the plot manually to achieve a better clustering picture. After you are ready with this, brush clusters by different colors and comment about the properties of the clusters: which variables are important to define these clusters and what values of these variables are specific to each cluster. Can these clusters be interpreted? Find the most prominent outlier and interpret it.*

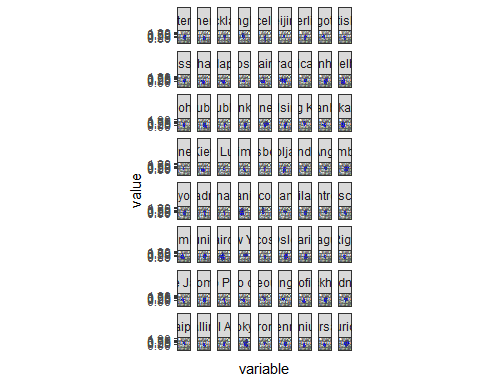
We could observe some clusters in each variables.  
i) There is cluster in vacation day between 20 and 30.  
ii) There is cluster in Clothing Index between 100 and 140.  
ii) There is a cluster in Minimum minutes for Rice between 10 and 20.

From the graph, we could deduct that total number of hours is significant variable.

*6) Use the data obtained by using the HC solver and create a radar chart diagram with juxtaposed radars. Identify two smaller clusters in your data (choose yourself which ones) and the most distinct outlier.*



## Warning: `funs()` was deprecated in dplyr 0.8.0.  
## Please use a list of either functions or lambdas:   
##   
## # Simple named list:   
## list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`:   
## tibble::lst(mean, median)  
##   
## # Using lambdas  
## list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.



## Warning: `mutate\_each\_()` was deprecated in dplyr 0.7.0.  
## Please use `across()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

## Warning: `add\_rownames()` was deprecated in dplyr 1.0.0.  
## Please use `tibble::rownames\_to\_column()` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.

## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode  
## No scatterpolar mode specifed:  
## Setting the mode to markers  
## Read more about this attribute -> https://plotly.com/r/reference/#scatter-mode

*7) Which of the tools you have used in this assignment (heatmaps, parallel coordinates or radar charts) was best in analyzing these data? From which perspective? (e.g. efficiency, simplicity, etc.)*

The heat maps is better in terms of simplicity. The other graph such as parallel coordinates and radar charts were difficult to understand. The parallel coordinate has lot of lines overlapping, thus making it difficult to distinguish between each other. In radar charts, the values are converted to radan value which is difficult to understand the correlation between parameters.