Insights from Data

Individual project 2 Report

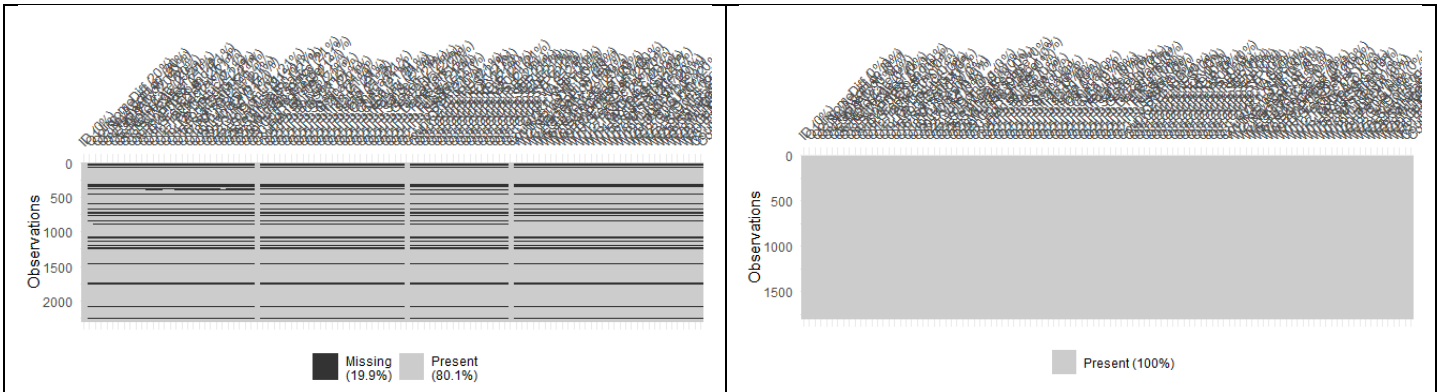
Maxwell Fundi Njiru

0. Project code

All project code for this task may be found on my Github repository

1. Data description and preprocessing

The questions ***vaalikone_questions_all.csv*** was a data frame with 108 variables(columns) and 2306 observations(rows). From the initial assessment using ***vis_miss*** function from ***Visdat*** package, I noticed that the data had 19.9% missing data as shown in in the graph below. Since the missing data was largely across the rows, not mostly one variable missing lots of data, I dropped all the rows with missing values and ended up with clean data. Additionally, I carried out conversion of columns data types to factors and ordered factors.



Graph showing percentage of missing data	Graph after handling missing vaues

Fig 1 – Graphs of missing values

2. Selection of distance measure

The Gower method was used with all columns except the ID column of the data. This was used because it is a good with datasets variables with mixed datatypes. There seems to be about 6 clusters from the data from the distance matrix visualization below

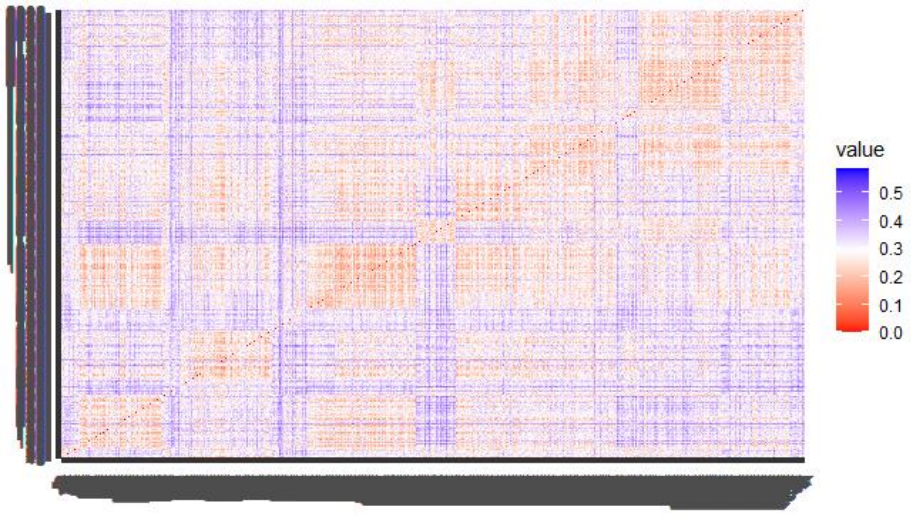


Fig 2 - distance matrix visualization using Gower

3. Clustering

Following the calculation of distance measures using gower method, and based on the distance matrix visualization, I conducted a hierarchical clustering using 6 clusters. However, on evaluating the clusters, I found that the clusters were of poorly separated and of poor quality with a silhouette width of 0.03. I therefore redid clustering with 3 clusters which seemed better but still not great. While the clustering improved, it was still quite poor with 3 clusters.

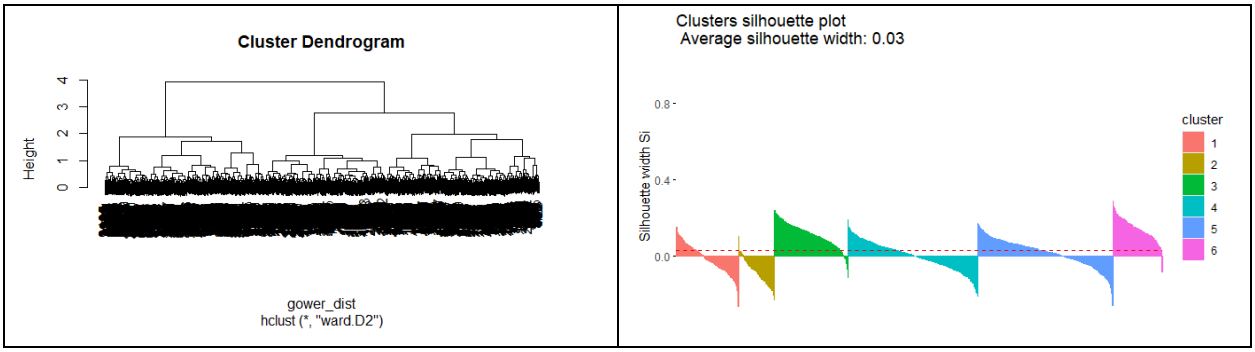


Fig 3 – Dendrogram and silhouette plot for 6 clusters

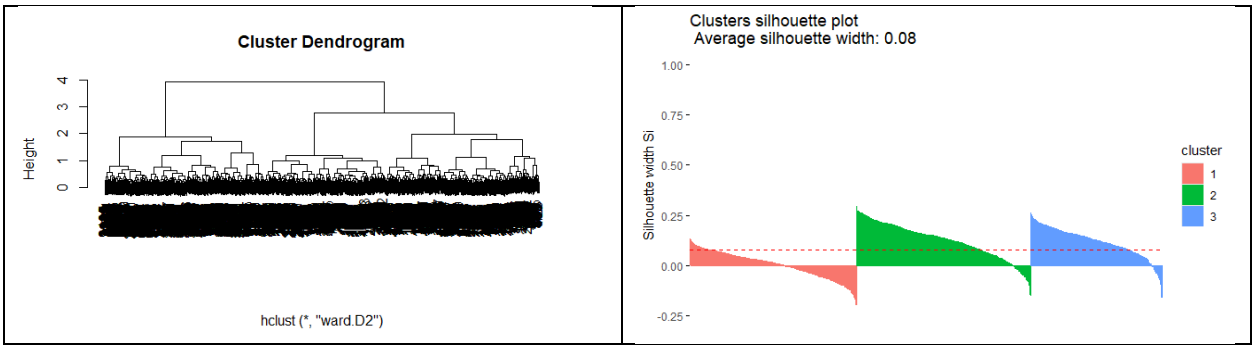


Fig 4 – Dendrogram and silhouette plot for 3 clusters

Finally, I plotted a bar graph to assess the degree to which clusters correspond to political parties. The bar graph showed that cluster 1 had very diverse range of political parties which mean they might have some similar features. Cluster 2 has fewer parties with high number of counts which shows that they might strong relationships. Cluster 3 had fewer political parties represented with many having very low counts

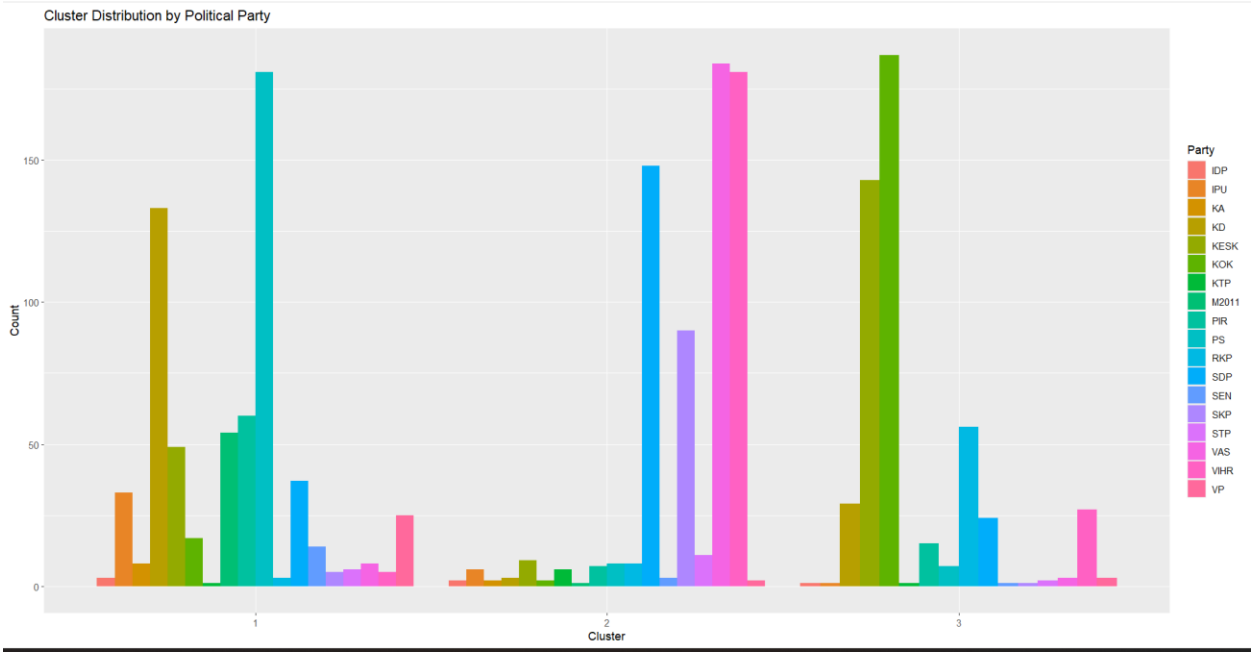


Fig 5 – Bargraph of clusters with parties