# Insights from Data

# Individual project 2 Report

# Maxwell Fundi Njiru

**Documentation and Reporting:**

o Document the entire process, including data preprocessing, clustering techniques, and

results (code and the output).

o Prepare a concise yet complete report highlighting key findings and insights

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.

|  |  |
| --- | --- |
| A graph showing a number of objects  Description automatically generated | A grey rectangular object with text  Description automatically generated with medium confidence |
| Graph showing percentage of missing data | Graph after handling missing vaues |

Selection of distance measure

* Gover was used with all colums except the ID column of the data as it is a good good with datasets with mixed variables. There seems to be about 6 clusters

A close up of a screen

Description automatically generated

\conducted hieracical clustering using 6 clusters

On evaluation

A graph of a plot

Description automatically generated with medium confidence

The average silhouette width is **0.03**, indicating **poor clustering quality**. The clusters are not well-separated or cohesive.

Tried this with 3 clusters

A graph of a plot

Description automatically generated with medium confidence

The clustering improved but still seemed to be quite poor…

* 1. Characteristics of identified clusters

A graph of different colored bars

Description automatically generated

 **Clusters 1 and 6 (Red and Pink)**:

* Many points have negative or near-zero silhouette widths, meaning they are poorly assigned or lie near cluster boundaries.

 **Clusters 2 and 3 (Yellow and Green)**:

* These have slightly better silhouette widths, but many points are still near zero or negative.

 **Clusters 4 and 5 (Cyan and Blue)**:

* Most points have negative silhouette widths, suggesting they may overlap significantly with other clusters or are misassigned.
* Preprosessing steps
  + Handling missing values
  + Conversion of colunns - (Q columns to factors and W columns to ordered factors).

1. Clustering methodology
   1. Selection of distance measure
   2. Clustering algorithm
2. Clusters visualisarion
   1. Characteristics of identified clusters