Clustering of patients into different groups based on baseline and nadir phenotype blood count.

Following the PCA plot which implicated some cluster upon which the patient can be classified, the next step was to undego unsupervised clustering of patients into groups. Our prime motive was to find CNVs enriched in high toxicity groups. However, it should be taken in account that the phenotype data for the neutrophils are laden with a lot of NAs which complicates the issue of clustering of these methods. With few journal reviews, it seems that the area is active research in its own. So whatever imputation method I choose will be a contentious issue and supported with strong logic and results.

I used the package missMDA for the imputation of data

Few notes about missMDA

How does it handle the mixed dataset (that contains both quantitative and categorical data)?

How does it impute the values

PCA vs MCA(MCA 🡪 data analysis technique for nominal )

missMDA performs PCA on incomplete data sets aiming to estimate parameters and obtain graphical representations despite missing values.

Main idea is based on iterative PCA

Regularized iterative PCA

As it is based on a principal component method, imputation takes into account

both similarities between individuals and relationships between variables

the regularized iterative FAMD algorithm

FAMD 🡪 Factorial analysis of mixed data

MFA 🡪 multiple factor analysis

Unsupervised k-means clustering method a

Kmeans are ‘