Result

To create a great dataset, changes need to be made to the downloaded dataset. The dataset has been gathered of 1484 yeast sequences from SWISS-PROT using the annotations from YPD. In this section, we are going to talk about the results we found from the dataset.

Codebook

As it is a large dataset, it is wise to look at the codebook first. However, there is no codebook. Nevertheless, there is a file with more information on all the attributes. This is where the information is extracted to make your own codebook.

Below a table with a description of all the attributes abbreviation, explanation and data types.

Name	Fullname	Datatypes
seq.name	Accession number for the SWISS-PROT database	str
mcg	McGeoch's method for signal sequence recognition	double
gvh	von Heijne's method for signal sequence recognition	double
alm	Score of the ALOM membrane spanning region prediction program	double
mit	Score of discriminant analysis of the amino acid content of the N-terminal region	double
erl	Presence of 'HDEL' substring	double
pox	Peroxisomal targeting signal in the C-terminus	logical
vac	Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins	double
nuc	Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins	double
loc.site	The class is the localization site	factor

Table 1: The codebook.

The last column is the sequence's localization site. There are 10 different possibilities for this. Yeast protein were classified into 10 classes: **cytoplasmic:** cytoskeletal (CYT); nuc]ear (NUC); vacuolar (VAC); mitochondrial (MIT); isomal (POX); **extracellular:** including those localized against the cell wall (EXC); proteins localized to the lumen of the endoplasmic reticulum (ERL); membrane proteins with a cleaved signal (ME1); membrane proteins with an uncleared signal (ME2); and membrane proteins with no N-terminal sign (ME3), where ME1, ME2,and ME3 proteins may be localized to the plasma membrane, the endoplasmic reticulum membrane, or the membrane of a golgi body.

Here are the 10 in question:

Abbreviation	Fullname	Amount
CYT	cytosolic or cytoskeletal	463
NUC	nuclear	429
MIT	mitochondrial	244
ME3	membrane protein, no N-terminal signal	163
ME2	membrane protein, uncleaved signal	51
ME1	membrane protein, cleaved signal	44
EXC	extracellular	37
VAC	vacuolar	30
POX	peroxisomal	20

Abbreviation	Fullname	Amount
ERL	endoplasmic reticulum lumen	5

Table 2: Sequence localization sites.

There are alot of CYT and NUC localization. ERL localization is the least. There are only 5 of these in the dataset.

Dataset

8 features were used in classification: the presence or absence of an HDEL pattern as a signal for retention in the endoplasmic reticulum lumen (erl); the result of discriminant analysis on the amino acid content of vacuolar and extracellular proteins (vac); the result discriminant analysis on the amino acid composition of the 20-residue N-terminal region of mitochondrial and non-mitochondrial proteins (mit); the presence absence of nuclear localization consensus patterns combined with a term reflecting the frequency of basic residues (nuc); and some combination of the presence of a short sequence motif and the result of discriminant analysis of the amino acid composition of the protein sequence (pox).

seq.name	mcg	gvh	alm	mit	erl	pox	vac	nuc	loc.site
ADT1_YEAST	0.58	0.61	0.47	0.13	0.5	0	0.48	0.22	MIT
ADT2_YEAST	0.43	0.67	0.48	0.27	0.5	0	0.53	0.22	MIT
ADT3_YEAST	0.64	0.62	0.49	0.15	0.5	0	0.53	0.22	MIT
AAR2_YEAST	0.58	0.44	0.57	0.13	0.5	0	0.54	0.22	NUC
AATM_YEAST	0.42	0.44	0.48	0.54	0.5	0	0.48	0.22	MIT
AATC_YEAST	0.51	0.4	0.56	0.17	0.5	0.5	0.49	0.22	CYT

Table 3: First 6 rows of the dataset.

As seen from the table the loaded dataset. Has a char datatype for the first and last column. And every other column has a double as datatype.

There are 1484 rows and 10 columns.

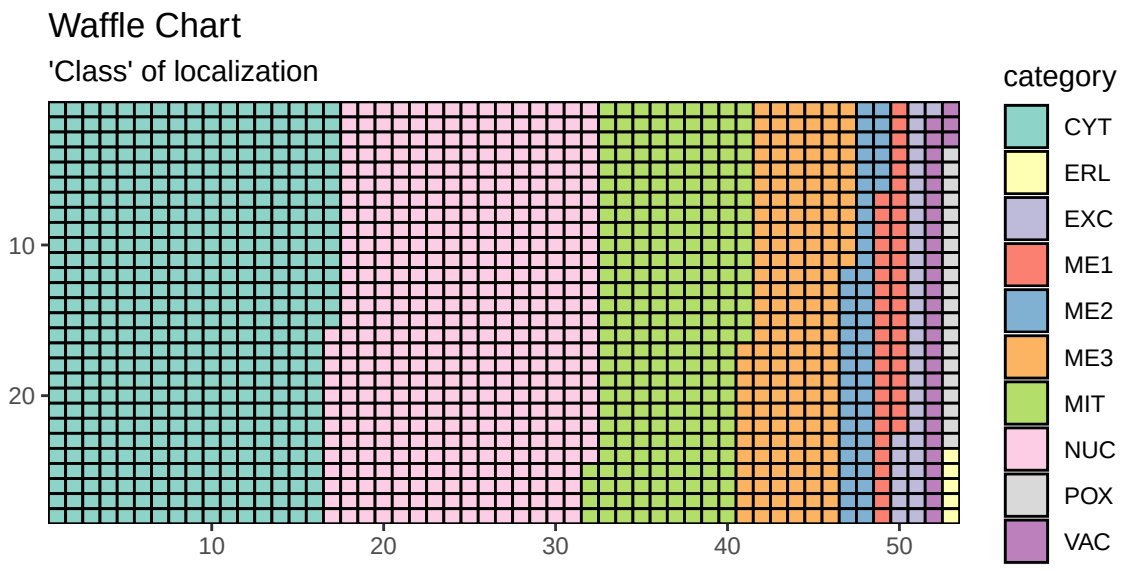

Figure 1: A waffle chart of the categorical composition

There are alot of CYT and NUC localization. And there are only 5 of ERL.

Clean the data

To ensure the quality of the data. The data had to be transformed before it could be worked with.

The first column has been dropped since it is not necessary. Since the sequence names contribute nothing to create a prediction model.

mcg	gvh	alm	mit	erl	pox	vac	nuc	loc.site
0.58	0.61	0.47	0.13	0.5	0	0.48	0.22	MIT
0.43	0.67	0.48	0.27	0.5	0	0.53	0.22	MIT
0.64	0.62	0.49	0.15	0.5	0	0.53	0.22	MIT
0.58	0.44	0.57	0.13	0.5	0	0.54	0.22	NUC
0.42	0.44	0.48	0.54	0.5	0	0.48	0.22	MIT
0.51	0.4	0.56	0.17	0.5	0.5	0.49	0.22	CYT

Table 4: First 6 rows of the dataset with the first column dropped.

As stated on the website there are 0 missing values.

```
##  Factor w/ 10 levels "CYT","ERL","EXC",..: 7 7 7 8 7 1 7 8 7 1 ...
```

The data type of the column loc.site has been succesfully changed to factors.

The ERL column has changed. From an double to a logical datatype. As seen in figure 2 below, only numbers 1 and 0.5 appear, but for a logical datatype you need 1 and 0. All 0.5 have been changed to 0.

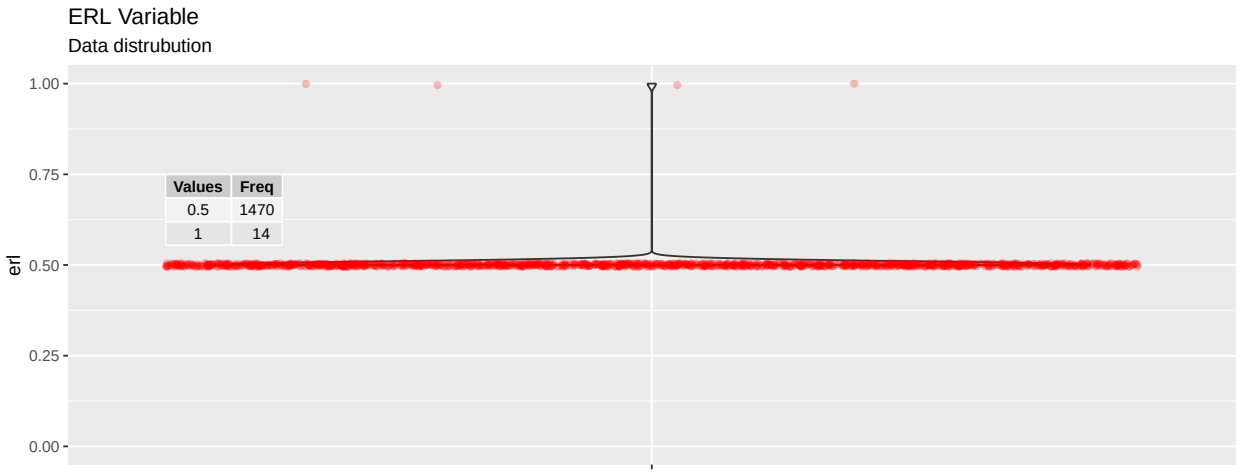

Figure 2: ERL column data distrubution

mcg	gvh	alm	mit	erl	pox	vac	nuc	loc.site
double	double	double	double	logical	double	double	double	factor

Table 5: Each column name and the typeof the datatype.

Each column has been changed so that it has the right data type.

Exploratory data analysis

mcg	gvh	alm	mit
Min. :0.1100	Min. :0.1300	Min. :0.21	Min. :0.0000
1st Qu.:0.4100	1st Qu.:0.4200	1st Qu.:0.46	1st Qu.:0.1700
Median :0.4900	Median :0.4900	Median :0.51	Median :0.2200
Mean :0.5001	Mean :0.4999	Mean :0.50	Mean :0.2612
3rd Qu.:0.5800	3rd Qu.:0.5700	3rd Qu.:0.55	3rd Qu.:0.3200
Max. :1.0000	Max. :1.0000	Max. :1.00	Max. :1.0000

pox	vac	nuc
Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.4800	1st Qu.:0.2200
Median :0.0000	Median :0.5100	Median :0.2200
Mean :0.0075	Mean :0.4999	Mean :0.2762
3rd Qu.:0.0000	3rd Qu.:0.5300	3rd Qu.:0.3000
Max. :0.8300	Max. :0.7300	Max. :1.0000

Table 6: Summary of the dataset.

As observed in table 6 the data has already been transformed with a min-max normalisation. All the datapoints are between 0 and 1.Mcg and gvh are normally distributed. Alm and mit are skewed but nothing to crazy to be worried about. Vac and nuc are really skewed however, not yet bad enough to worry about it. Pox is very oddly distributed.

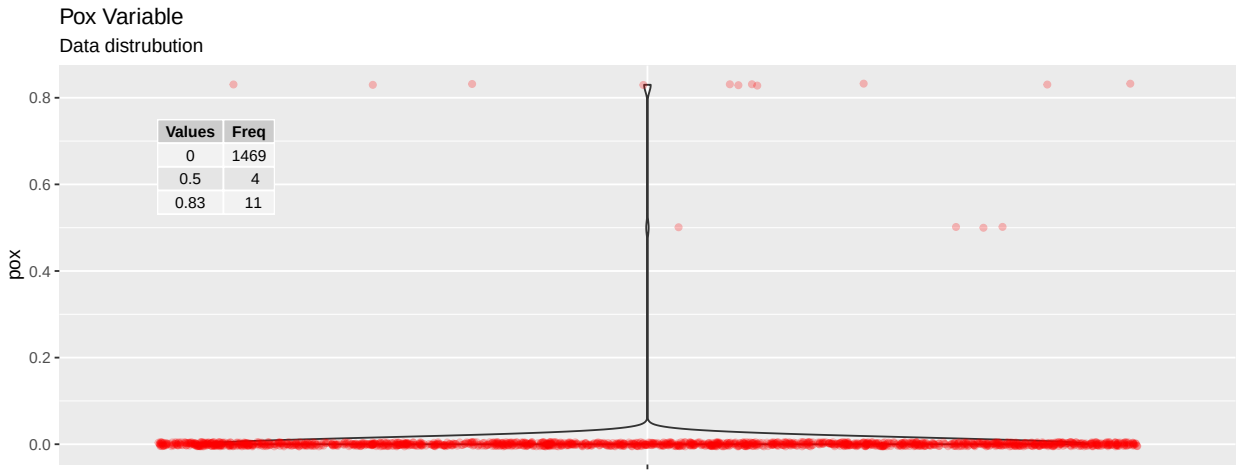

Figure 3: Closer look at the pox variable

Almost all the datapoint are 0. This could cause problems later on. Since more than 95% of all the datapoints are 0.

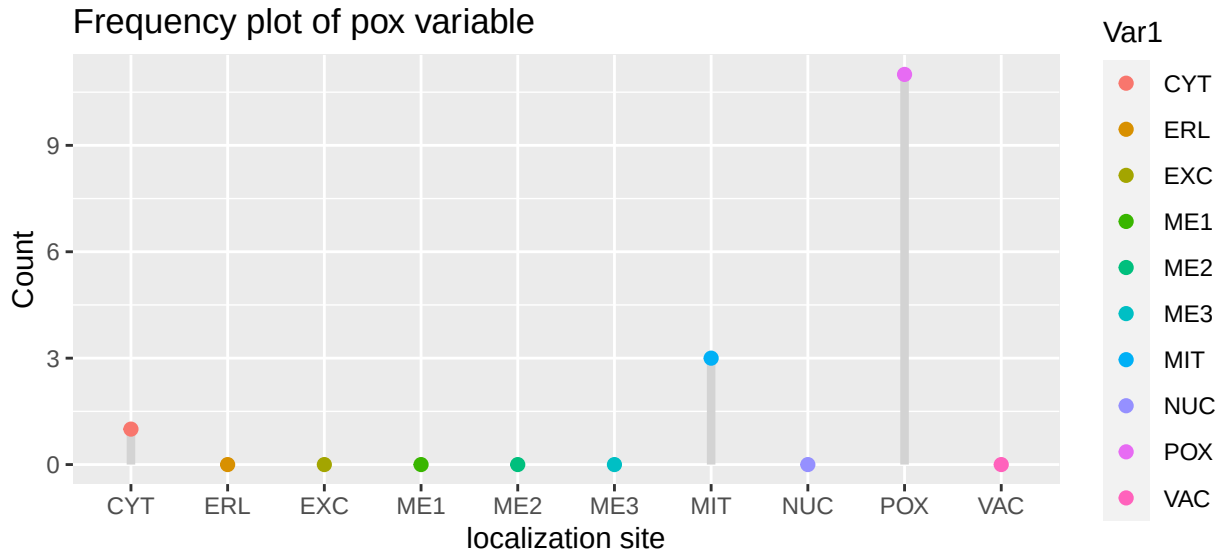

Figure 4: Frequency plot of pox variablee

If the pox value is not 0 then the classification is Pox, Mit or Cyt. This means that Pox has a high probability when the Pox value is not 0.

With a heatmap, you can easily see where there are correlations between variables.

mcg	gvh	alm	mit	pox	vac	nuc	varnames
1.0000000	0.5816314	-0.1639513	0.1581755	0.0055970	0.0750427	-0.1245404	mcg
0.5816314	1.0000000	-0.2718000	0.1403139	0.0003918	0.0887594	-0.1029840	gvh
-0.1639513	-0.2718000	1.0000000	0.0596683	0.0093779	-0.1858054	-0.0220428	alm
0.1581755	0.1403139	0.0596683	1.0000000	-0.0090398	-0.1035914	-0.0547965	mit
0.0055970	0.0003918	0.0093779	-0.0090398	1.0000000	0.0208997	-0.0356586	pox
0.0750427	0.0887594	-0.1858054	-0.1035914	0.0208997	1.0000000	0.0896904	vac
-0.1245404	-0.1029840	-0.0220428	-0.0547965	-0.0356586	0.0896904	1.0000000	nuc

Table 7: The calculated correlation matrix.

This has been tranformed to a long matrix.

varnames	variable	cor
mcg	mcg	1
mcg	gvh	0.5816
mcg	alm	-0.164
mcg	mit	0.1582
mcg	pox	0.005597
mcg	vac	0.07504

Table 8: The long calculated correlation matrix.

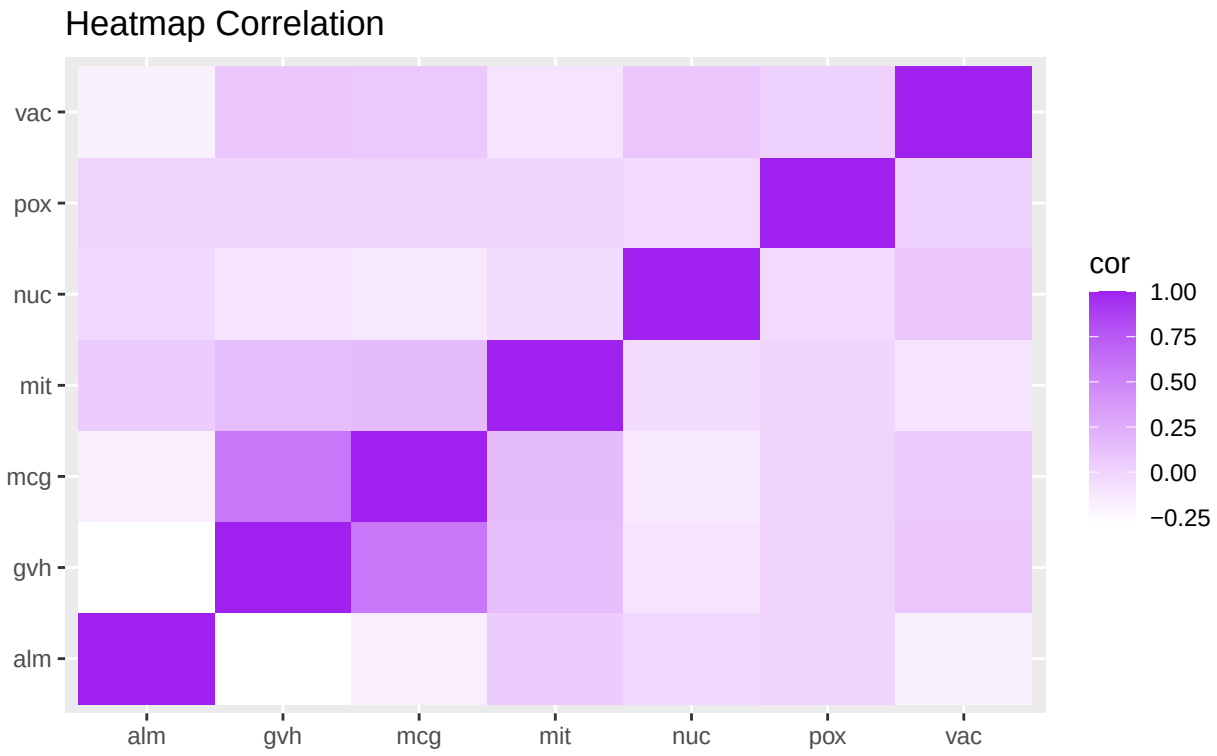

Figure 5: A heatmap pairwise correlation of selected numeric variables

As seen from the heatmap there is some correlation between mcg and gvh. While on the contrary Pox barely has any correlation with a other variable.

There needs to be data correlation with the classes. The heigth of the peak doesn't matter much, shifted peaks do.

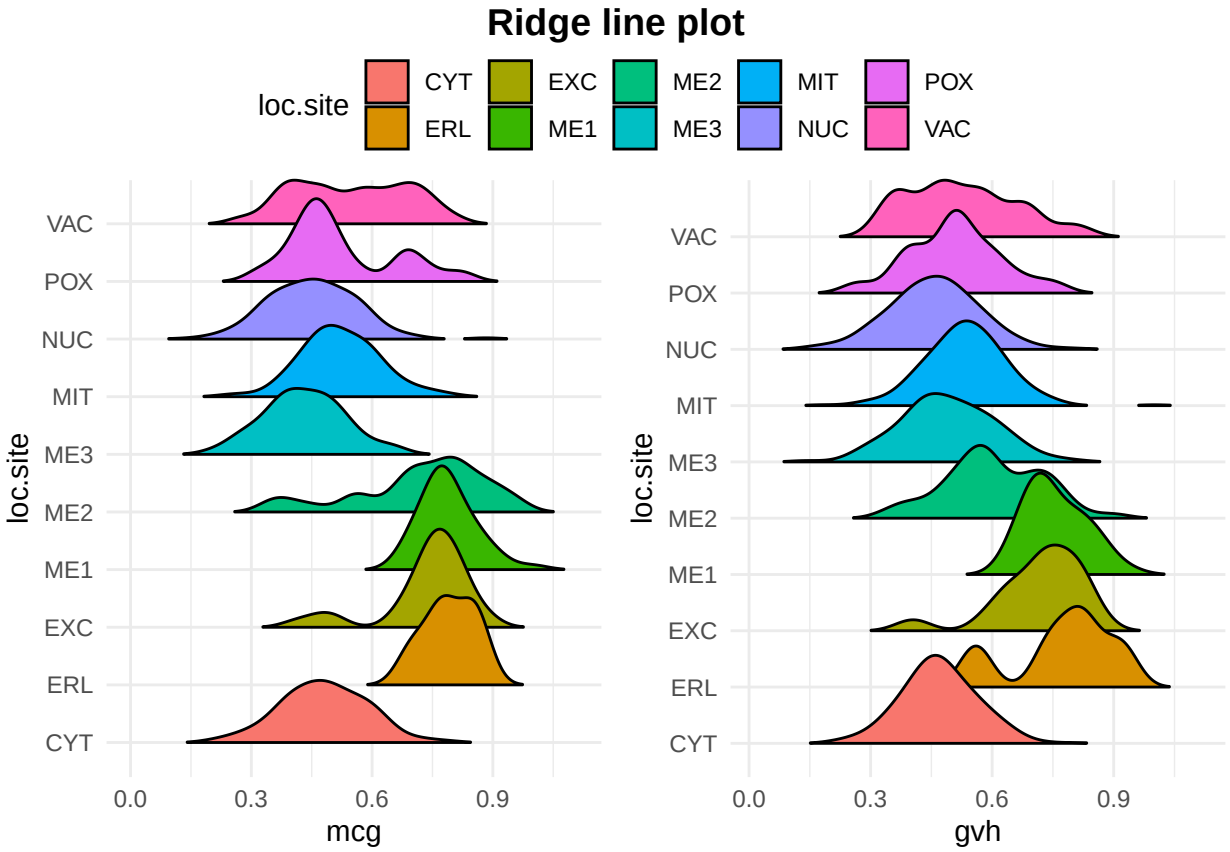

Figure 6: Ridge line plot between mcg and gvh

There are shifted peaks. this shows a distribution of the different classes.

A 1-way ANOVA test was carried out on the data.

The P-value: $1.2036763 \times 10^{-146}$ is < 0.05. So there is a significant difference.

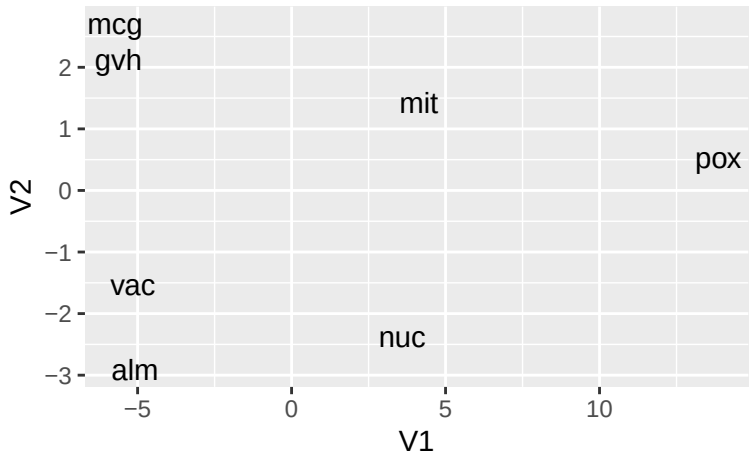

Figure 7: Classical (Metric) Multidimensional Scaling

As found in the heatmap Mcg and Gvh are clustered. Vac and Alm are somewhat clustered they are close to eachother. Nuc, Mit and Pox are all on their own.

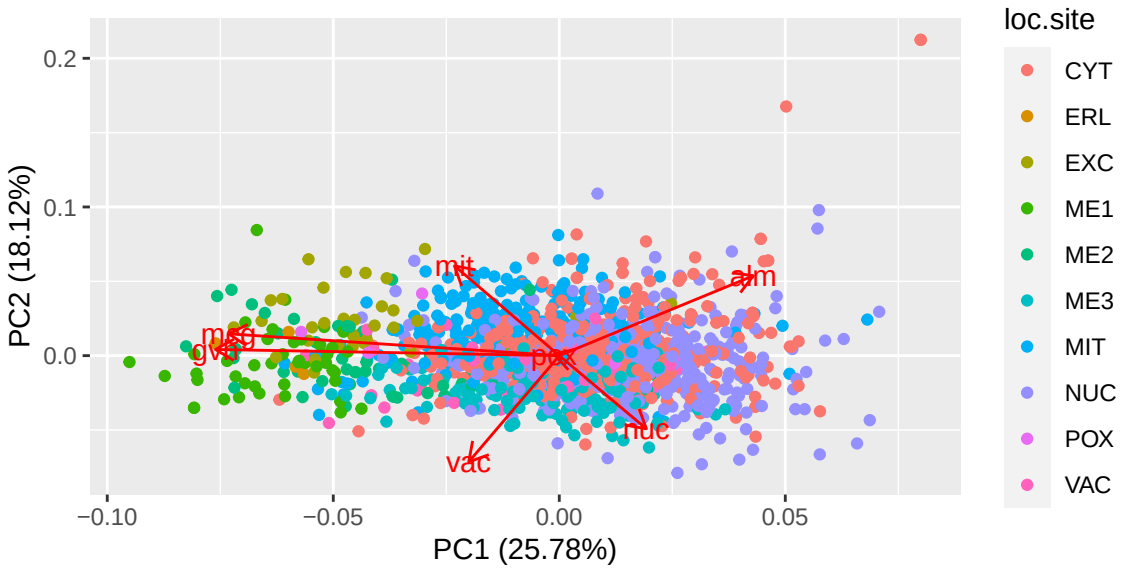

Figure 8: PCA plot

Mcg and Gvh are also clustered in this PCA plot. As in contrast with the MDS plot Vac and Alm are not clustered at all. Pox is in the centre.

Discussion

Conclusion