**Predicting growth in children**

***Objective 1, Explore and wrangle the dataset***

STEP 1

Create smaller datasets with growth measurements (growth dataset), clinical (clinical dataset) and biochemical variables (biochemical dataset)

Growth measurements:

* "Length\_\_cm\_ time point 1,2,3,4,5 and 6 "
* "Weight\_\_cm\_ time point 1,2,3,4,5 and 6 "
* "MUAC\_\_cm\_ time point 1,2,3,4,5 and 6 "

Clinical variables:

* "Diarrhoea\_Exists",
* "Fever",
* "Night\_Sweats",
* "Weight\_Loss",
* "Breathlessness",
* "Weakness",
* "Loss\_of\_Apetite",
* "Cough",
* "Haemoptysis",
* "Vomiting",
* "Haematemesis",
* "Anorexia",
* "Rashes",
* "Odema"

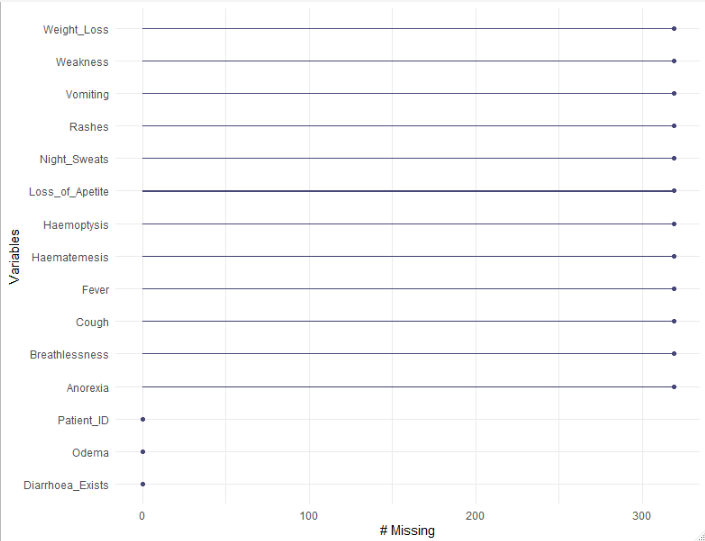
Biochemical variables:

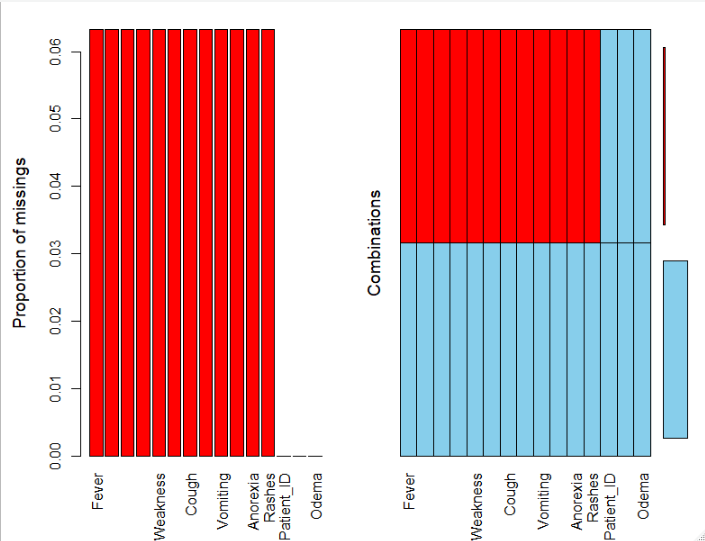
* "Patient\_ID",
* "Hematocrit\_PCV",
* # "Diff\_Leucocyte\_Count",
* "Total\_Leucocyte\_Count",
* "Neutrophil",
* "Anion\_Gap",
* "ESR",
* "Hemoglobin",
* "Potassium",
* "Sodium",
* "Chloride"

STEP 2

Explore the dataset: Calculate of rate of weight gain and missing data for the smaller datasets.

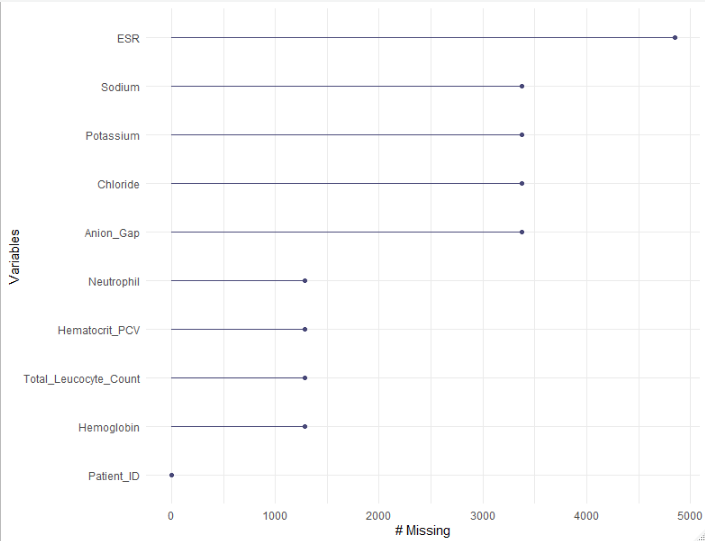
**Clinical dataset**

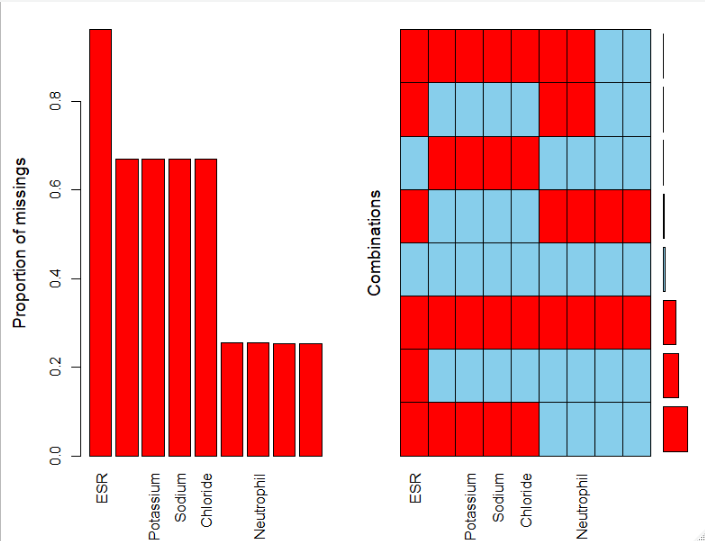




The plots show the frequency of missing data for each of the clinical variables in the clinical sub dataset. It is clearly shown that 319/5043 children miss all clinical variables at the same time.

**Biochemical dataset**

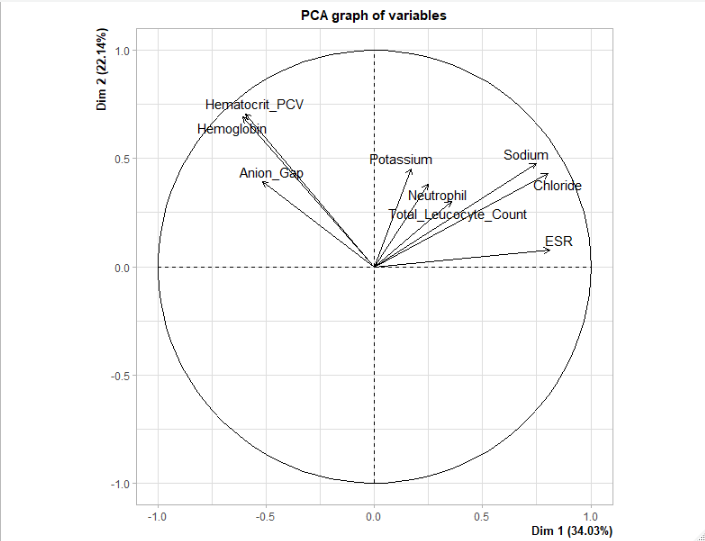
****



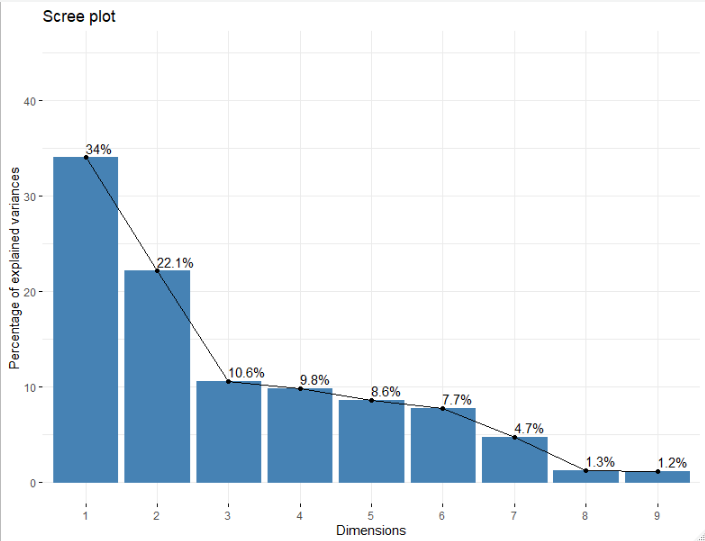
In the biochemical dataset the picture considering missing variables distribution is less clear. A bigger number of children is missing biochemical variables, while almost everyone is missing data on ESR and electrolytes.

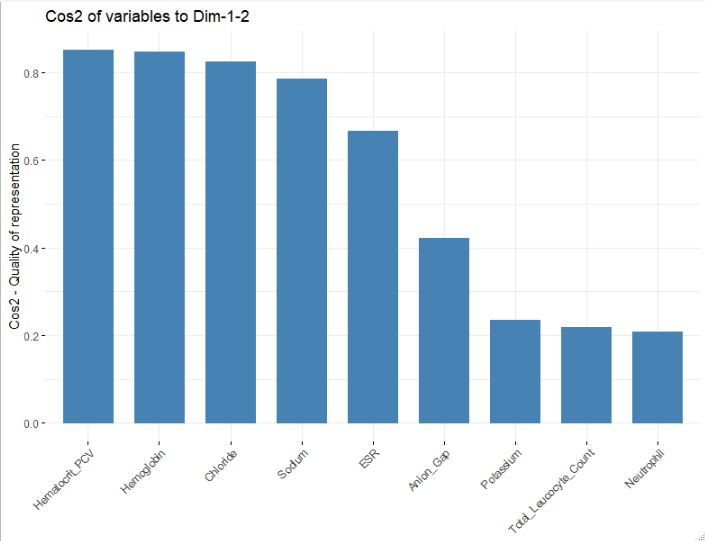
STEP 3

Since missingness in the biochemical sub-dataset is not consistent, we will investigate hidden patterns in the dataset, and whether some variables are correlated introducing redundancy in the dataset, by conducting PCA. Since this dataset is incomplete, any missing data were imputed.



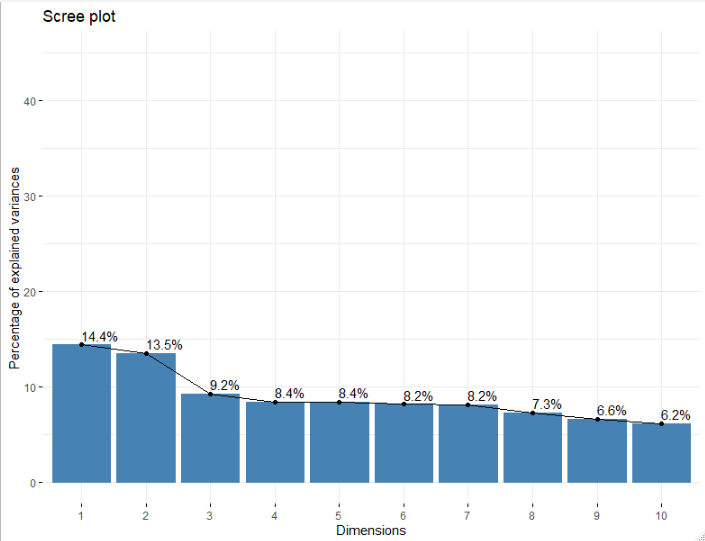
Two principal components were identified explaining 56.1% of the total variance. The plot above is a variable correlation plot, showing the relationships between all variables. Positively correlated variables are grouped together. Variables that are away from the origin are well represented on the factor map.



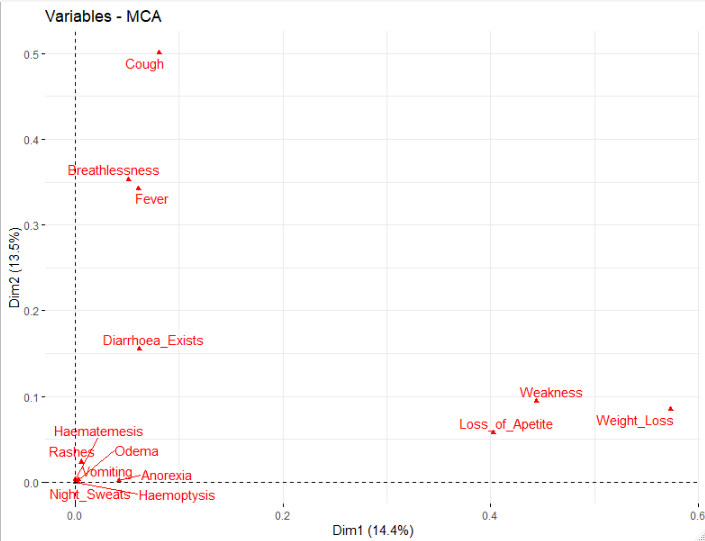


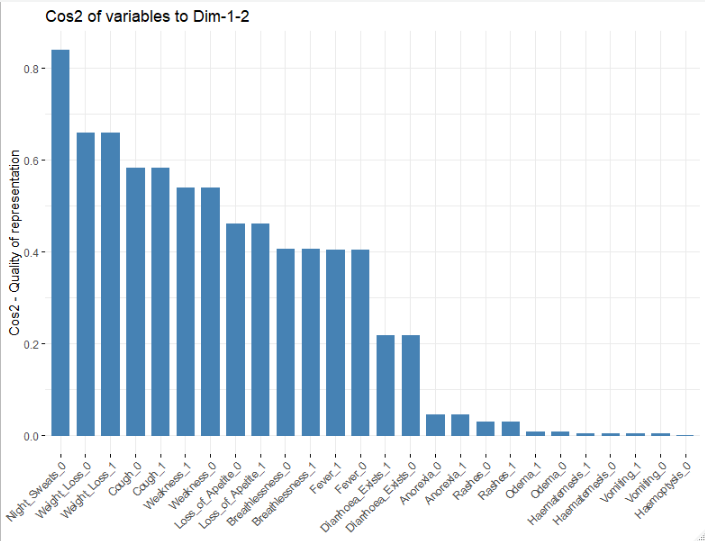
The quality of the variables on a factor map is called cos2. A high cos2 indicated a good representation of the variable on the principal component.

The same data mining process is repeated for the clinical variables. Since the clinical variables are categorical MCA (a generalization of PCA) analysis was conducted to identify groups with similar clinical profile after excluding the 319 children with missing data.



The clinical dataset shows less clear separation (different clinical profiles) of the clinical dataset. Some variables are correlated (weakens, loss of appetite, loss of weight) as expected.





STEP 3 (Topic of discussion)

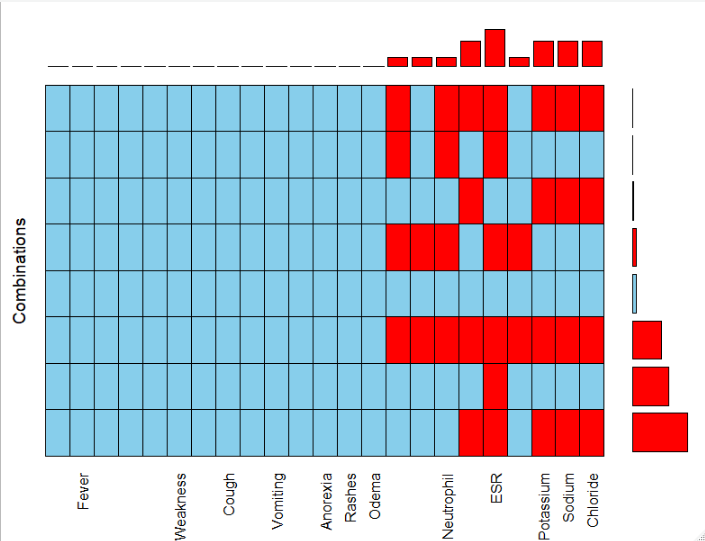
Create an index of missingness for each child and make groups.

Clinical data excluding patients with more than 10 NAs (n=319)

|  |  |
| --- | --- |
| **#NAs** | **#Patients** |
| 0 | 147 |
| 1 | 1317 |
| 3 | 3 |
| 4 | 35 |
| 5 | 2164 |
| 9 | 1057 |
| 12 | 8 |
| 13 | 66 |
| 15 | 1 |
| 17 | 145 |
| 21 | 99 |

From the total dataset containing all 21 variables (clinical and biochemical) we identified the “most problematic patients”, meaning those that had NAs in the majority of the variables (>10). The total number of these cases was 319 patients. After excluding these patients, we repeated this study from the other dimension, that of variables. We studied what combinations of variables had the most NAs between them. As it can be seen from the table below, the worst combination was that of {Anion\_gap, ESR, Potassium, Sodium, Chloride}, for which 43% of the patients (n=2044) had only NAs. The second worst combination included only ESR, for which about 28% of the patients (n=1317) had only NAs. Finally, the third worst combination included all biochemical markers, for which about 22% of the patients (n=1057) had only NAs.

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Combinations** | **Count** | **Percent** |
| **1** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0 | 147 | 3.1117697 |
| **2** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:0:0:0:0 | 1317 | 27.8789162 |
| **3** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:0:0:1:1:1 | 35 | 0.7408975 |
| **4** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:1:0:1:1:1 | 2044 | 43.2684166 |
| **5** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:0:1:0:1:0:0:0:0 | 3 | 0.0635055 |
| **6** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:0:1:1:1:0:1:1:1 | 1 | 0.0211685 |
| **7** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:1:1:0:1:1:0:0:0 | 120 | 2.5402202 |
| **8** | 0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:1:1:1:1:1:1:1:1 | 1057 | 22.3751058 |



Topics of discussion

* Do the presented results make clinical sense?
* Which should be the NAs cut-off?
* Do we need all biochemical variables? Since some of them are correlated, are they redundant?
* Assuming non randomness, Can NAs (and their combination) be a clinical feature the predicts growth?
* Next steps: Data cleaning, prediction