Regression without regrets

M. Baillie, G. Heinze & M. Huebner

3/16/23

Table of contents

# Preface

The focus of this report is to provide guidance on conducting initial data analysis in a reproducible manner in the context of intended regression analyses.

# 1. Multivariate analyses

First load the required packages and data. Note:, from the univariate analyses the analysis data set for the lab parameters has been updated to include transformed variables. Therefore, we load the second iteration of the data data/ADLB\_02.rds.

## 1.1 V1: Association with structural variables

Attached the required structural variables to the lab data.

A scatterplot of each predictor with age, with different panels for males and females have been constructed. Associated Spearman correlation coefficients have been computed.

### 1.1.1 Key predictors

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |
| --- | --- |
|  |  |

### 1.1.2 Predictors of medium importance

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

### 1.1.3 Remaining predictors

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

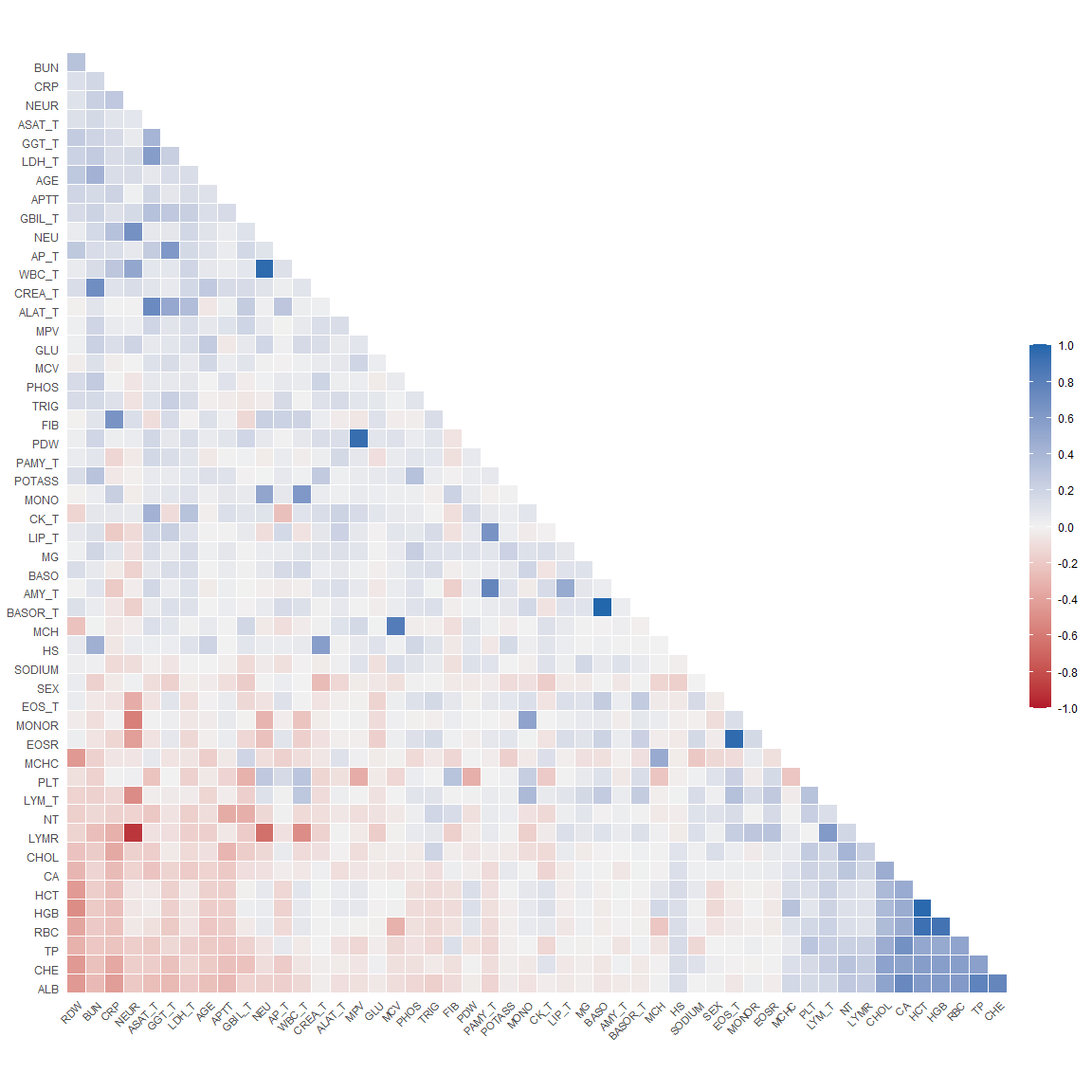
|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |
| --- | --- |
|  |  |

## 1.2 V2: Correlation coefficients between all predictors

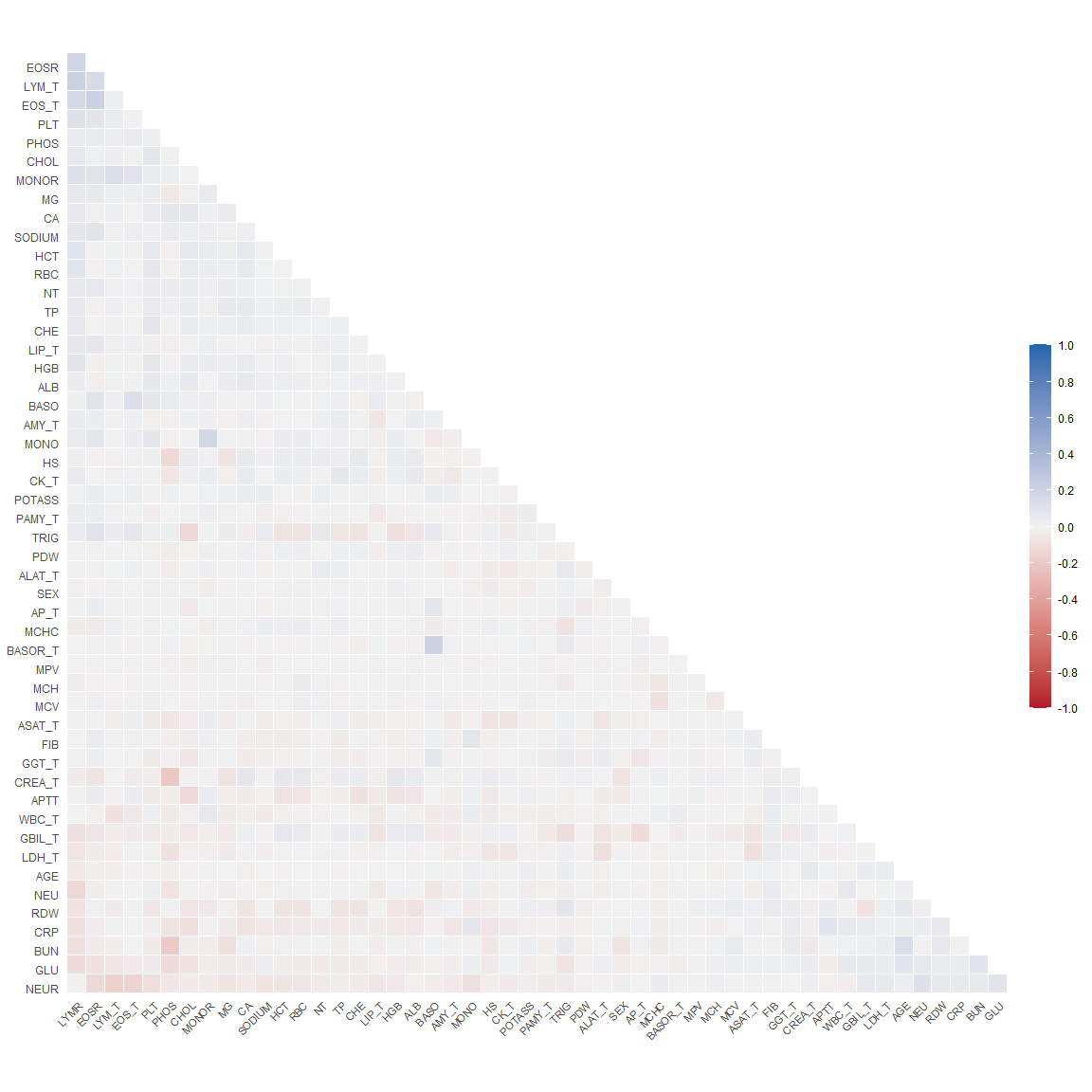
Calaulcate correlation matrix using Spearman correlation coefficient.

The Spearman correlation coefficients are depicted in a quadratic heat map:

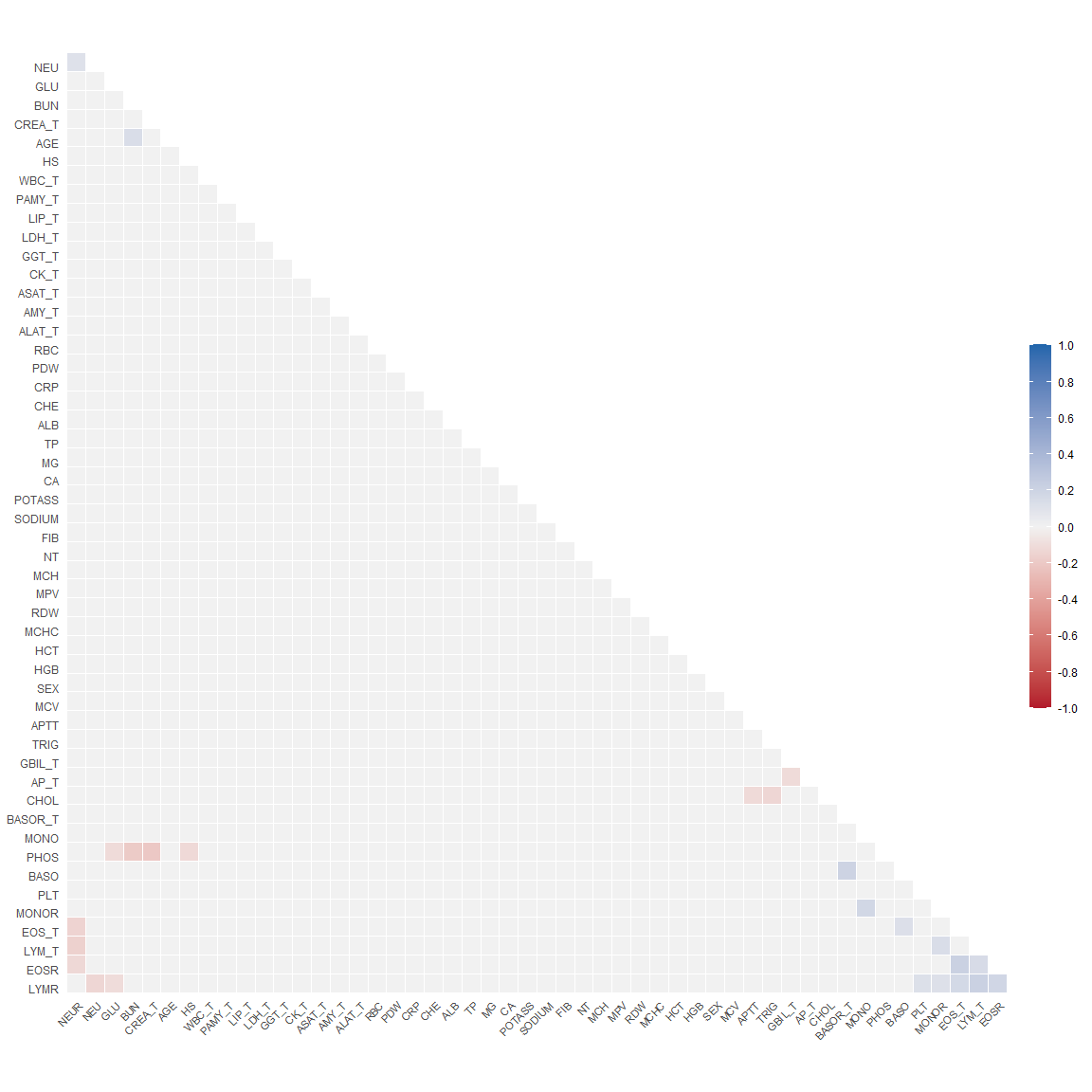


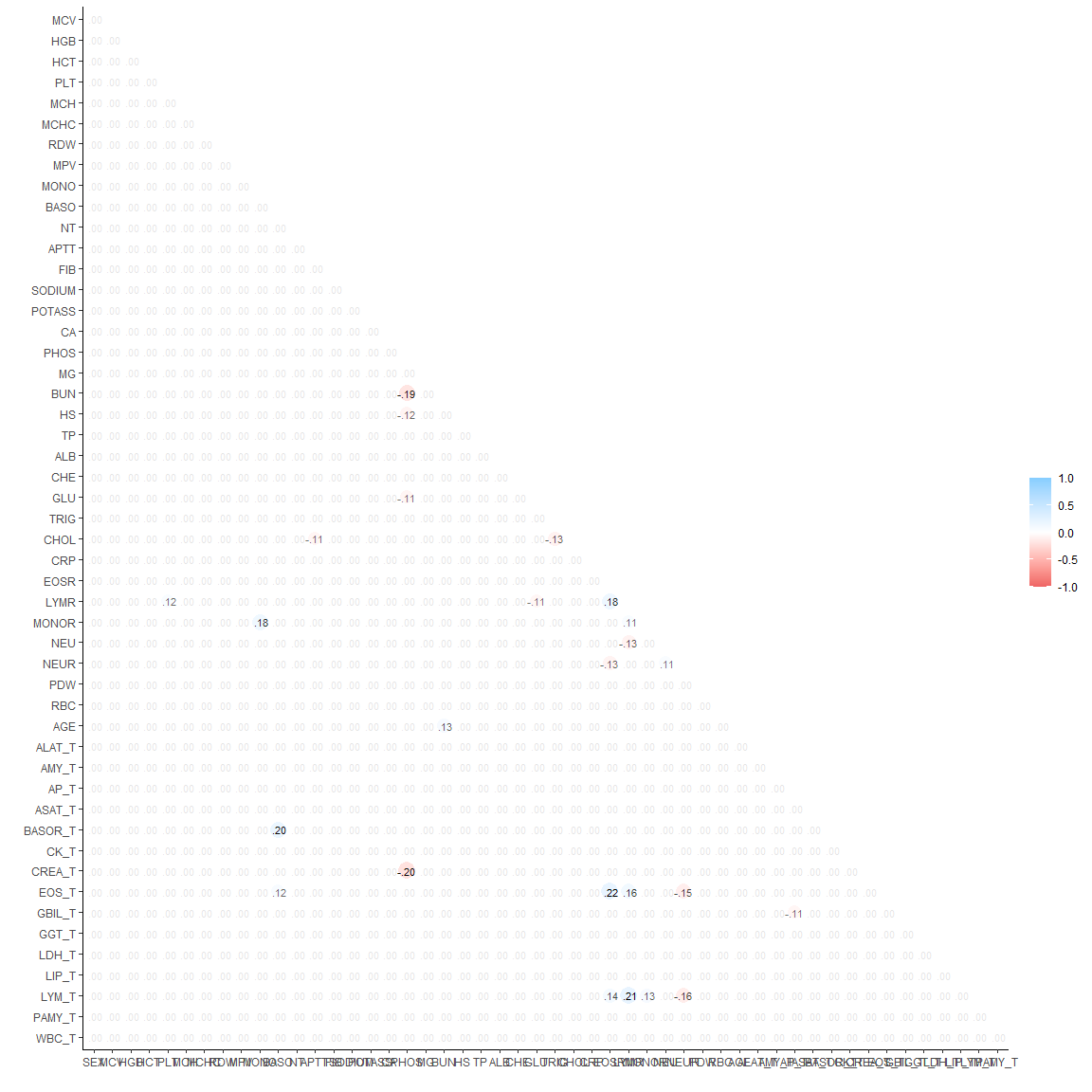
### 1.2.1 VE1: Comparing nonparametric and parametric predictor correlation

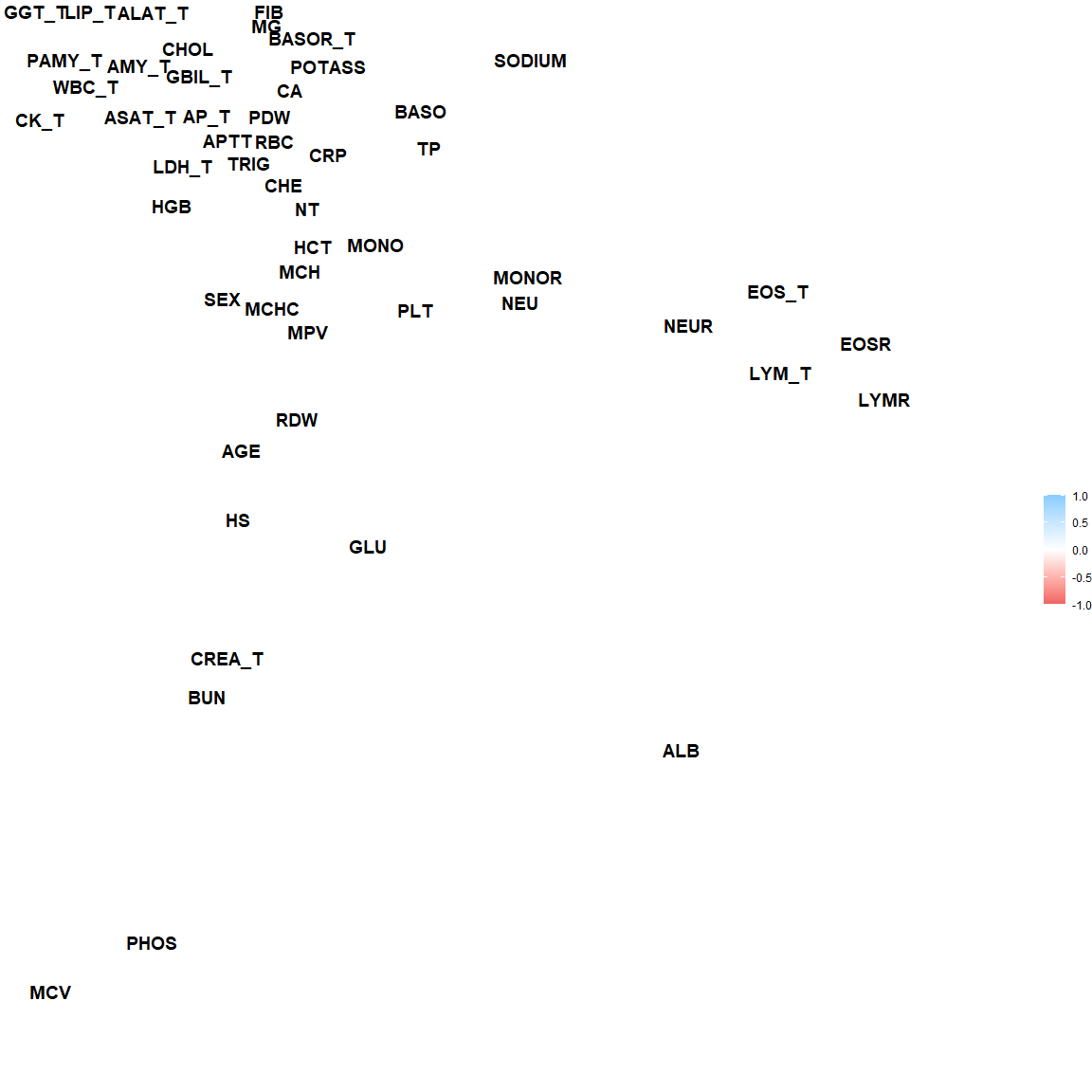
Plot the difference matrix

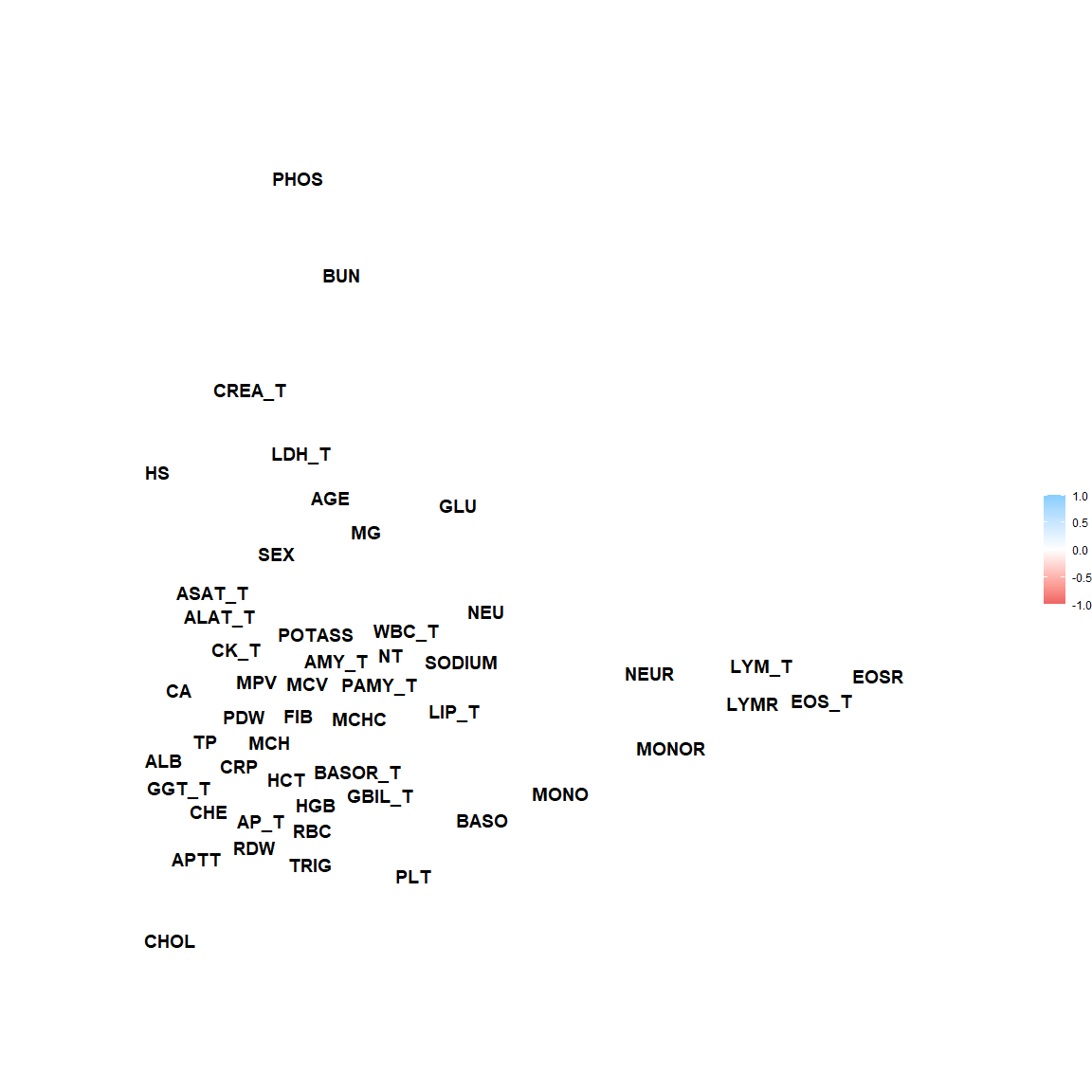


Plot the matrix but suppress differences less then 0.1.









Predictor pairs for which Spearman and Pearson correlation coefficients differ by more than 0.1 correlation units will be depicted in scatterplots:

| x | y | r |
| --- | --- | --- |
| PLT | LYMR | 0.1161529 |
| MONO | MONOR | 0.1761864 |
| BASO | BASOR\_T | 0.2003626 |
| BASO | EOS\_T | 0.1172575 |
| BUN | AGE | 0.1329799 |
| EOSR | LYMR | 0.1818832 |
| EOSR | EOS\_T | 0.2171330 |
| EOSR | LYM\_T | 0.1449661 |
| LYMR | PLT | 0.1161529 |
| LYMR | EOSR | 0.1818832 |
| LYMR | MONOR | 0.1108623 |
| LYMR | EOS\_T | 0.1643959 |
| LYMR | LYM\_T | 0.2145499 |
| MONOR | MONO | 0.1761864 |
| MONOR | LYMR | 0.1108623 |
| MONOR | LYM\_T | 0.1319019 |
| NEU | NEUR | 0.1062152 |
| NEUR | NEU | 0.1062152 |
| AGE | BUN | 0.1329799 |
| BASOR\_T | BASO | 0.2003626 |
| EOS\_T | BASO | 0.1172575 |
| EOS\_T | EOSR | 0.2171330 |
| EOS\_T | LYMR | 0.1643959 |
| LYM\_T | EOSR | 0.1449661 |
| LYM\_T | LYMR | 0.2145499 |
| LYM\_T | MONOR | 0.1319019 |

Data prep

Plot the data.

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

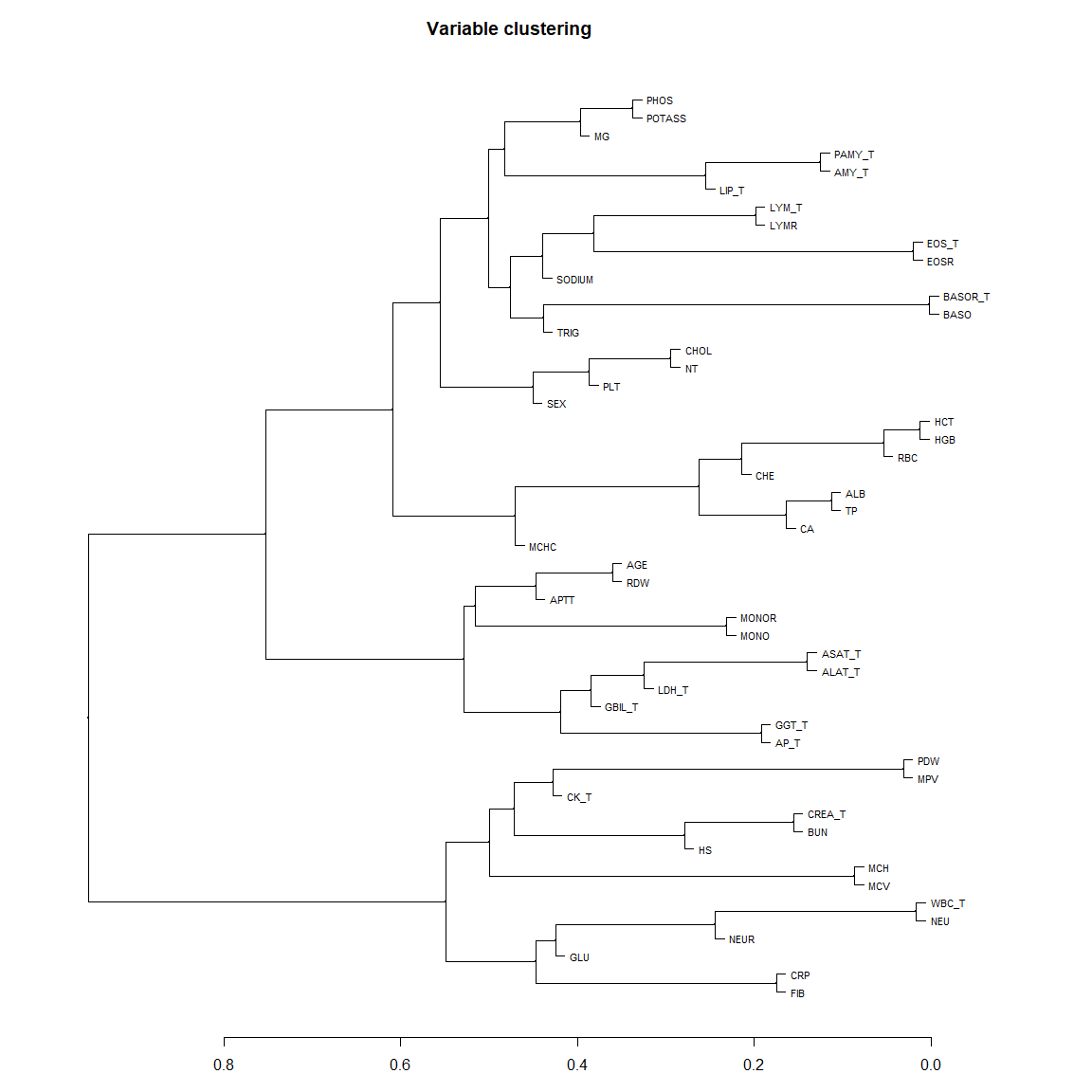
|  |  |  |
| --- | --- | --- |
|  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

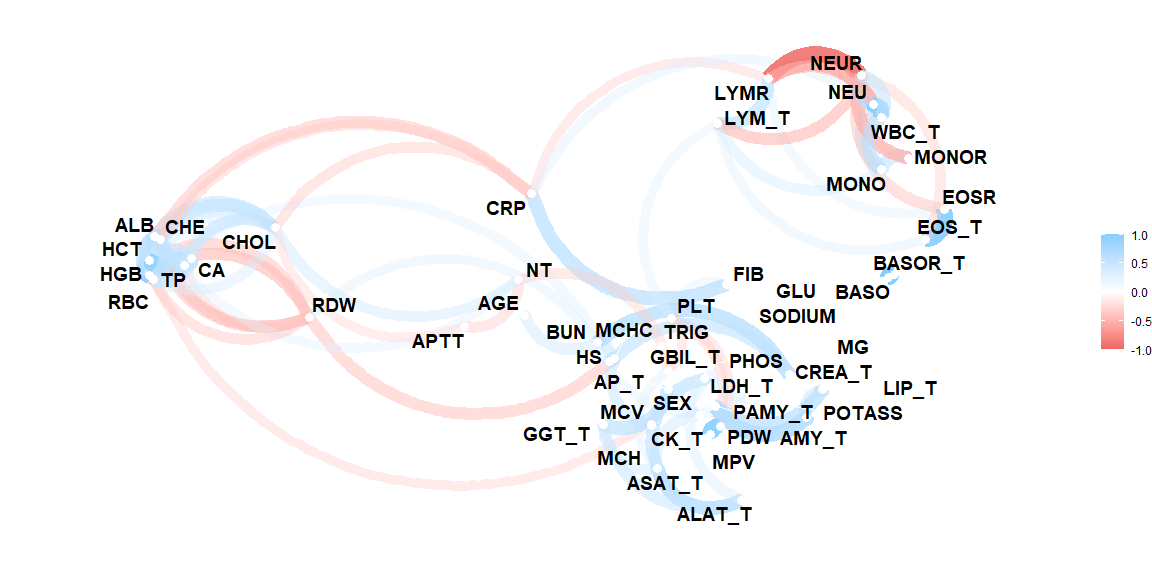
|  |  |
| --- | --- |
|  |  |

### 1.2.2 VE2: Variable clustering

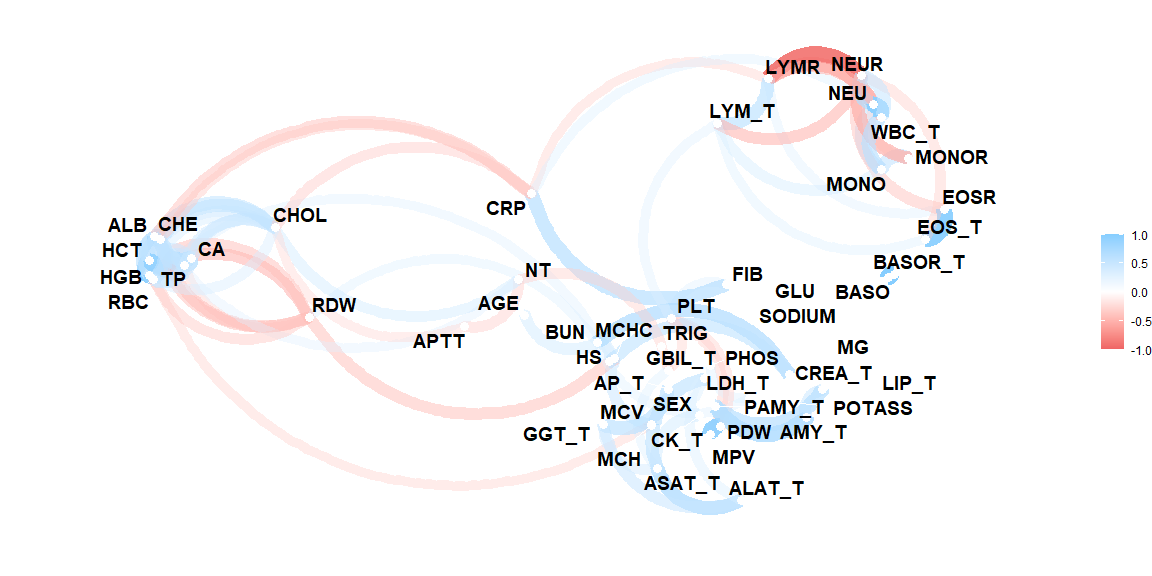
A variable clustering analysis has been performed to evaluate which predictors are closely associated. The dendrogram groups predictors by their correlation.



plot network plot



plot network with min cor of 0.3.



In the following scatterplots we show predictor pairs with Spearman correlation coefficients greater than 0.8:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |
| --- | --- |
|  |  |

### 1.2.3 VE3: Redundancy

Variance inflation factors (VIF) will be computed between the candidate predictors. This will be done for the three possible candidate models, and using all complete cases in the respective candidate predictor sets. Since , we also report the multiple R-squared values. Redundancy was further explored by computing parametric additive models for each predictor in the key predictor model and the extended predictor model. VIFs and multiple are reported from those models, again for the three predictor sets.

#### 1.2.3.1 VIF for key predictor model

~SEX + MCV + HGB + HCT + PLT + MCH + MCHC + RDW + MPV + MONO +   
 BASO + NT + APTT + FIB + SODIUM + POTASS + CA + PHOS + MG +   
 BUN + HS + TP + ALB + CHE + GLU + TRIG + CHOL + CRP + EOSR +   
 LYMR + MONOR + NEU + NEUR + PDW + RBC + AGE + ALAT\_T + AMY\_T +   
 AP\_T + ASAT\_T + BASOR\_T + CK\_T + CREA\_T + EOS\_T + GBIL\_T +   
 GGT\_T + LDH\_T + LIP\_T + LYM\_T + PAMY\_T + WBC\_T

Available sample size:  
 3979 ( 27.08 %)

Variance inflation factors:

SEX MCV HGB HCT PLT MCH MCHC RDW MPV MONO   
 1.34 129.57 254.36 249.42 1.92 179.46 47.48 1.89 9.17 4.30   
 BASO NT APTT FIB SODIUM POTASS CA PHOS MG BUN   
 3.48 1.46 1.22 2.91 1.44 1.41 2.08 1.56 1.30 3.23   
 HS TP ALB CHE GLU TRIG CHOL CRP EOSR LYMR   
 1.76 4.51 6.34 2.83 1.25 1.40 1.87 2.84 77.58 2279.22   
 MONOR NEU NEUR PDW RBC AGE ALAT\_T AMY\_T AP\_T ASAT\_T   
 548.43 9.41 3752.11 9.14 40.08 1.45 3.90 2.72 2.34 6.56   
BASOR\_T CK\_T CREA\_T EOS\_T GBIL\_T GGT\_T LDH\_T LIP\_T LYM\_T PAMY\_T   
 5.65 2.22 2.87 3.40 1.61 2.60 2.25 2.18 4.59 3.40   
 WBC\_T   
 16.31

Multiple R-squared:

SEX MCV HGB HCT PLT MCH MCHC RDW MPV MONO   
 0.2521 0.9923 0.9961 0.9960 0.4791 0.9944 0.9789 0.4705 0.8909 0.7673   
 BASO NT APTT FIB SODIUM POTASS CA PHOS MG BUN   
 0.7123 0.3166 0.1799 0.6566 0.3035 0.2927 0.5191 0.3608 0.2322 0.6908   
 HS TP ALB CHE GLU TRIG CHOL CRP EOSR LYMR   
 0.4334 0.7785 0.8423 0.6469 0.2025 0.2874 0.4659 0.6473 0.9871 0.9996   
 MONOR NEU NEUR PDW RBC AGE ALAT\_T AMY\_T AP\_T ASAT\_T   
 0.9982 0.8938 0.9997 0.8906 0.9751 0.3126 0.7433 0.6322 0.5719 0.8475   
BASOR\_T CK\_T CREA\_T EOS\_T GBIL\_T GGT\_T LDH\_T LIP\_T LYM\_T PAMY\_T   
 0.8230 0.5495 0.6516 0.7057 0.3786 0.6157 0.5546 0.5416 0.7821 0.7057   
 WBC\_T   
 0.9387

#### 1.2.3.2 VIF for model with key predictors and predictors of medium importance

#### 1.2.3.3 VIF for all predictor model

#### 1.2.3.4 Redundancy by parametric additive model: key predictor model

#### 1.2.3.5 Redundancy by parametric additive model: key predictors and predictors of medium importance

#### 1.2.3.6 Redundancy by parametric additive model: all predictors