Studies of Independent Variables

05.04.2022

Bibliotheken laden, Hilfsfunktion

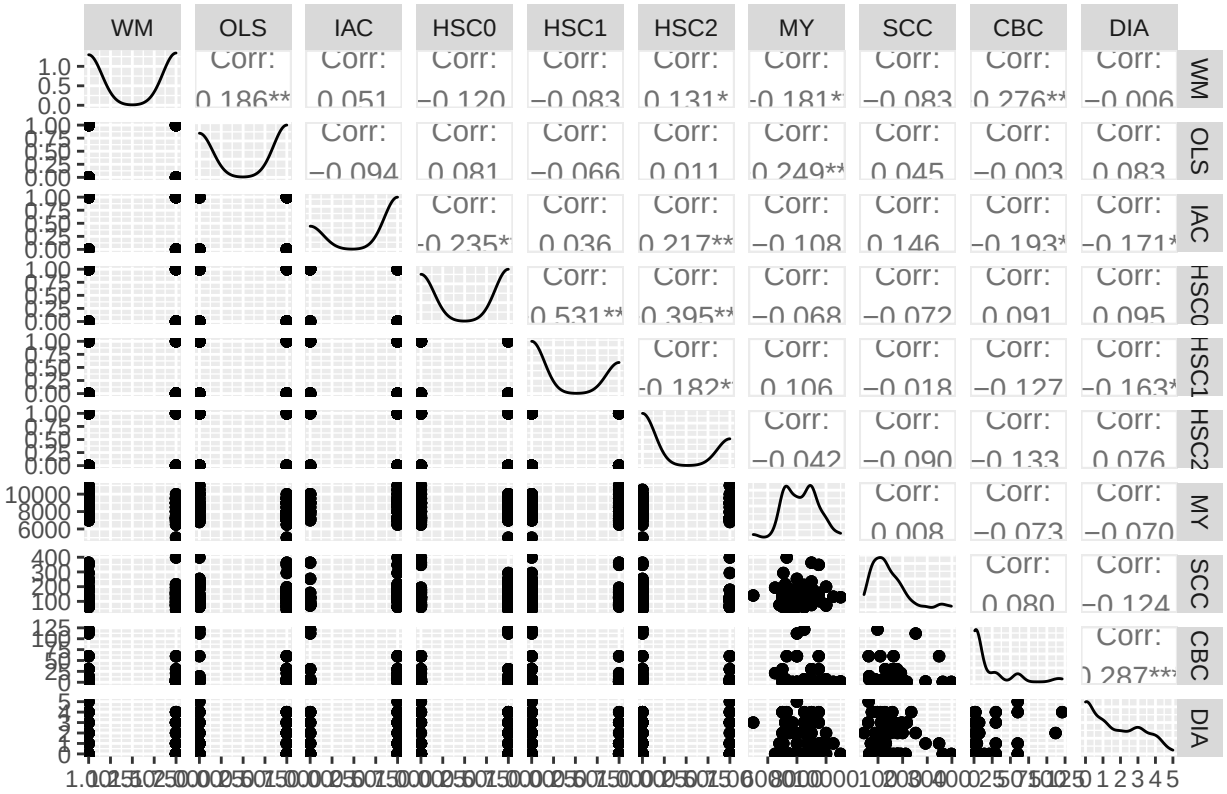
```
#library(ggplot2)     # moderne plots
library(GGally)

debug <- T           # debug printout
debug <- F           # kein debug printout
Log <- function(string) {
  if(debug){print(string)}
}
```

For all MY Groups :

- Resistenzen.Rmd generated Resistenzen[Schicht].csv, read it in
- plot variables and calculate correlations

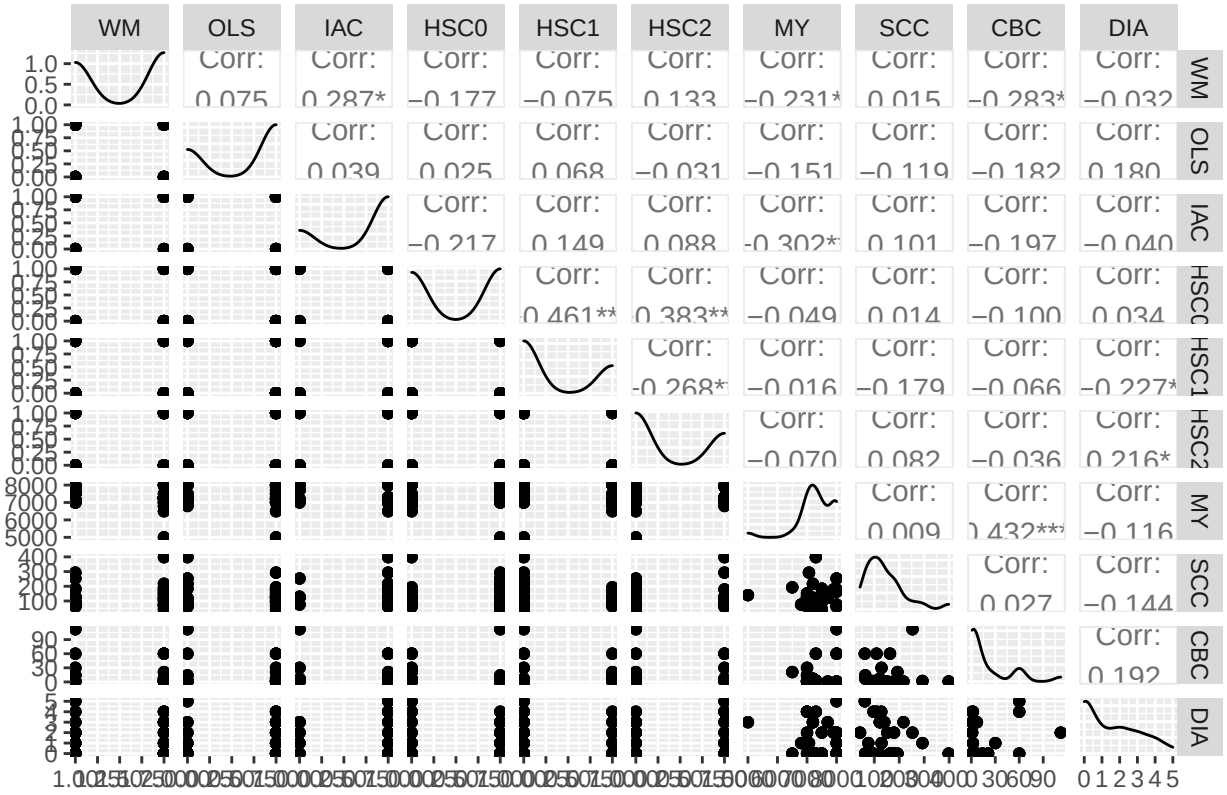
```
for( Schicht in c("U", "LE8000","GT8000") ) {     # Un-stratisfied /  Less than or Equal to 8000 / Greater Than 80
   FileIn <- paste( "Resistenzen",Schicht,".csv" , sep="" )
   Resistenzen <- read.csv(FileIn)

  # csv schreiben fügt vorne Index-Spalte an; diese entfernen :
  Resistenzen[,1] <- NULL

  if(debug){View(Resistenzen)}

  df <- data.frame(WM     = Resistenzen$WM.group,    # unabhängige Variablen extrahieren
                   OLS    = Resistenzen$OLS.group,   # incl. Titel kürzen, sonst Platzprobleme ...
                   IAC    = Resistenzen$IAC.group,
                   HSC0   = Resistenzen$HSC0,
                   HSC1   = Resistenzen$HSC1,
                   HSC2   = Resistenzen$HSC2,
                   #HSC3   = Resistenzen$HSC3,
                   #HSC4   = Resistenzen$HSC4,
                   #HSC5   = Resistenzen$HSC5,
                   MY     = Resistenzen$MY,
                   SCC    = Resistenzen$SCC,
                   CBC    = Resistenzen$CBC,
                   DIA    = Resistenzen$DIA)
  #View(df)
  print(ggpairs(df, title = paste("group:",Schicht), upper=list(continuos=wrap("cor",size=6))))
  print("")
}
```

group: U

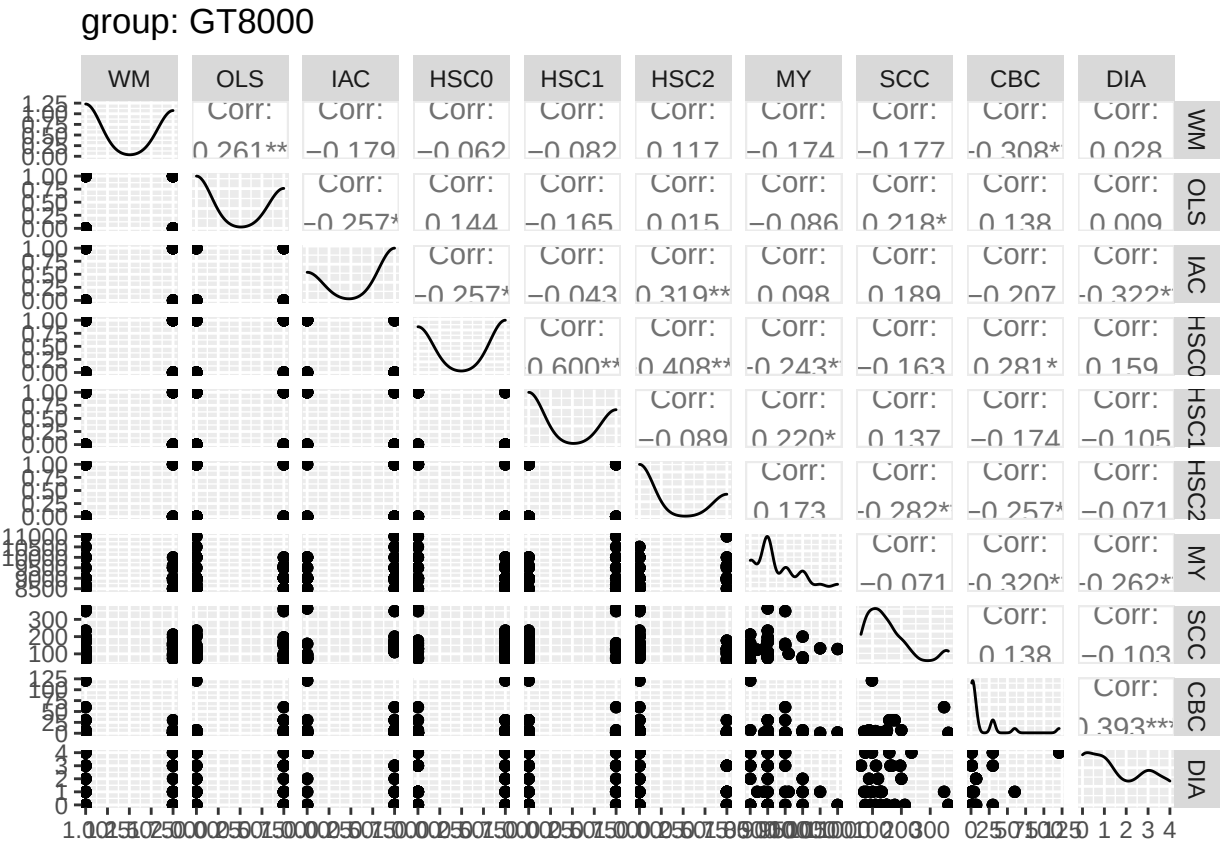



```
## [1] ""
```

group: LE8000




```
## [1] ""
```




```
## [1] ""
```

Linear Depence?

Linear Dependence implies multicollinearity, so in its presence the logistic regression would be unreliable.

The maximum correlation magnitude amounts to

- 53.2% (HSC0 with HSC1) in the unstratisfied analysis
- 60.0% (HSC0 with HSC1) in the stratisfied analysis MY > 8000

It might be better to not include HSC0 and HSC1 in one multivariate logistic regression to avoid collinearity problems.

Outliers?

Im wesentlichen sind nur die plots ohne diskrete Variablen gut zu interpretieren (für die anderen könnte man bessere Grafiken machen, hätte dann aber immer noch das Problem der beliebigen Kodierung).

In Histogrammen und Streuplots sehe ich einen Ausreisser mit MY = 5000, das ist Farm 32.

- ist sie als problematisch bekannt?
- das ist aber *kein* Problem für die Regression: MY wird zur Schichtung verwendet, nicht als unabhängige Variable