Explorative Data Analysis (EDA)

Import libraries

```
library(ggplot2)
library(ggcorrplot)
library(cowplot)
library(PerformanceAnalytics)
library(GGally)
library(knitr)
library(kableExtra)
```

Import final and final_individual tables

```
final <- read.csv("https://raw.githubusercontent.com/elypaolazz/dataset/main/final%20(1).csv")
final_individual <- read.csv("https://raw.githubusercontent.com/elypaolazz/dataset/main/final%20(1).csv
```

Tidy up

```
final <- na.omit(final)

final_individual <- na.omit(final_individual)
```

Compute variables correlation matrix

```
cor(final)
```

```
##             nquest       ncomp        nord          sex       anasc      staciv
## nquest  1.00000000 -0.02699042 -0.02168942  0.095148381 -0.09355298  0.05049659
## ncomp  -0.02699042  1.00000000  0.51613245 -0.313631734  0.32711351 -0.24284084
## nord   -0.02168942  0.51613245  1.00000000 -0.215243244  0.22669178 -0.15412186
## sex     0.09514838 -0.31363173 -0.21524324  1.000000000 -0.99970277  0.72181626
## anasc  -0.09355298  0.32711351  0.22669178 -0.999702768  1.00000000 -0.72197626
## staciv  0.05049659 -0.24284084 -0.15412186  0.721816257 -0.72197626  1.00000000
## ireg   -0.09432300  0.28968200  0.16432198 -0.603336868  0.60488621 -0.42872908
## y      -0.05072944  0.20106422  0.10407368 -0.008956435  0.01080913 -0.12671623
##              ireg            y
## nquest -0.0943230 -0.050729440
## ncomp   0.2896820  0.201064223
## nord    0.1643220  0.104073681
## sex    -0.6033369 -0.008956435
## anasc   0.6048862  0.010809133
## staciv -0.4287291 -0.126716225
## ireg    1.0000000 -0.136384548
## y      -0.1363845  1.000000000
```

```
cor(final_individual)
```

```
##             nquest       ncomp        nord          sex       anasc      staciv
## nquest  1.00000000 -0.02699042 -0.02168942  0.095148381 -0.09355298  0.05049659
## ncomp  -0.02699042  1.00000000  0.51613245 -0.313631734  0.32711351 -0.24284084
## nord   -0.02168942  0.51613245  1.00000000 -0.215243244  0.22669178 -0.15412186
## sex     0.09514838 -0.31363173 -0.21524324  1.000000000 -0.99970277  0.72181626
## anasc  -0.09355298  0.32711351  0.22669178 -0.999702768  1.00000000 -0.72197626
## staciv  0.05049659 -0.24284084 -0.15412186  0.721816257 -0.72197626  1.00000000
## ireg   -0.09432300  0.28968200  0.16432198 -0.603336868  0.60488621 -0.42872908
## y      -0.05072944  0.20106422  0.10407368 -0.008956435  0.01080913 -0.12671623
##              ireg            y
## nquest -0.0943230 -0.050729440
## ncomp   0.2896820  0.201064223
## nord    0.1643220  0.104073681
## sex    -0.6033369 -0.008956435
## anasc   0.6048862  0.010809133
## staciv -0.4287291 -0.126716225
## ireg    1.0000000 -0.136384548
## y      -0.1363845  1.000000000
```

Plot correlation matrix

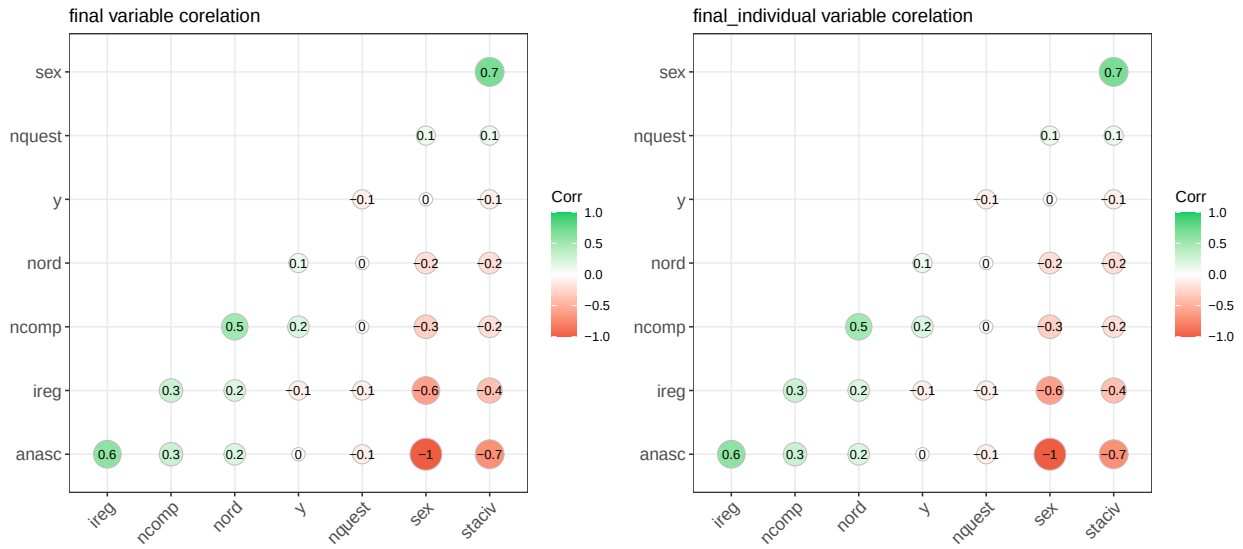
```
# final
# Correlation matrix
corr_final = round(cor(final), 1)

# Plot
final_corplot <- ggcorrplot(corr_final, hc.order = TRUE,
                            type = "lower",
                            lab = TRUE,
                            lab_size = 3,
                            method="circle",
                            colors = c("tomato2", "white", "springgreen3"),
                            title="final variable corelation",
                            ggtheme=theme_bw)

# final_individual
# Correlation matrix
corr_final_indiv = round(cor(final_individual), 1)

# Plot
final_indv_corplot <- ggcorrplot(corr_final_indiv, hc.order = TRUE,
                                 type = "lower",
                                 lab = TRUE,
                                 lab_size = 3,
                                 method="circle",
                                 colors = c("tomato2", "white", "springgreen3"),
                                 title="final_individual variable corelation",
                                 ggtheme=theme_bw)

plot_grid(final_corplot, final_indv_corplot)
```

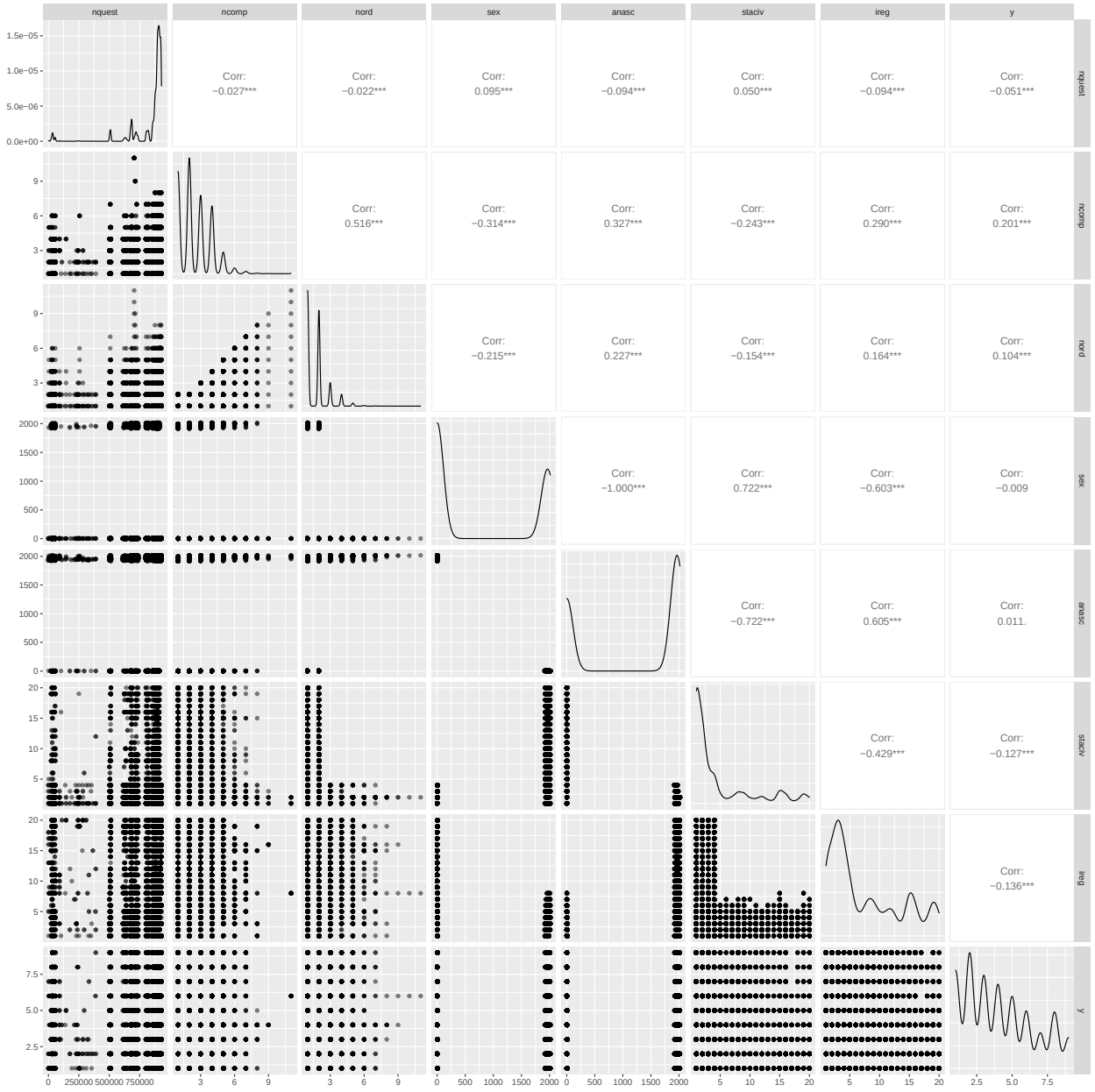


As the correlographs show, the Exploratory Data Analysis (EDA) detected several high Pearson correlation coefficients. In particular, a $\rho > |\pm 0.6|$ is detected between: **staciv-anasc**, **staciv-sex**, **anasc-ireg**, and **ireg-sex** in both datasets. A strong negative correlation ($\rho = -1$) characterises the couple **sex-anasc**. These results can already cut out some classification methods having the assumption of independence between predictors.

Compute more informative matrices

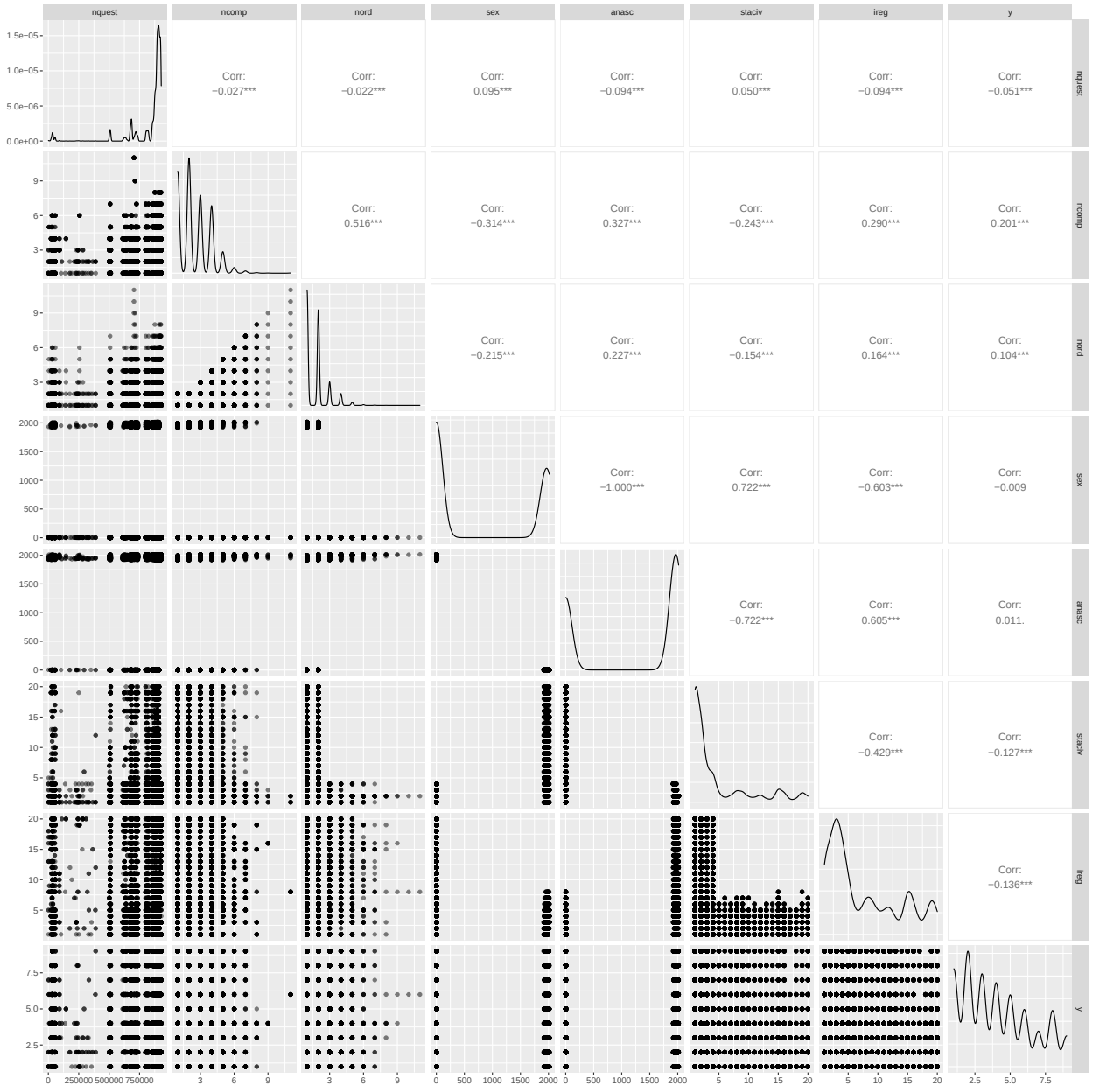
```
final_chart <- ggpairs(final,
                       lower = list(continuous = wrap("points", alpha = 0.5)),
                       diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha=0.5 ))
                       )

final_chart
```




```
final_indiv_chart <- ggpairs(final_individual,
                     lower = list(continuous = wrap("points", alpha = 0.5)),
                     diag = list(discrete="barDiag", continuous = wrap("densityDiag", alpha=0.5 ))
                     )

final_indiv_chart
```



The variables contained in those datasets are: - *nquest* and *nord*: both integer numbers. nquest (household ID) represents the primary key to merge household level. In order to merge individual level information it must be considered together with nord (ID of each household member). - *sex*: define males as 1 and females as 2. - *anasc*: year of birth of the respondent - *staciv*: marital status which is inserted as integer number ranging from 1 to 6. Specifically, it identifies: 1 = celibe/nubile, 2 = convivente, 3 = Sposato/a, 4 = Vedovo/a, 5 Separato/a, and 6 Divorziato/a. - *ireg*: integer number (1-20) reporting the NUTS2 codes, meaning the numbers associated with the Italian regions. In particular:

The variables' density functions highlight an absence of normality. Therefore, they do not satisfy the assumptions of statistical learning methods such as LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis). Indeed, the assumptions that must be satisfied in LDA and QDA:

- **Multivariate normality**: the predictors taken into consideration should follow a gaussian distribution for each grouping variable level.

Table 1: Regions and corresponding NUTS2 code

Region	NUTS2
Piemonte	1
Valle d'Aosta	2
Lombardia	3
Trentino Alto Adige	4
Veneto	5
Friuli Venezia Giulia	6
Liguria	7
Emilia Romagna	8
Toscana	9
Umbria	10
Marche	11
Lazio	12
Abruzzo	13
Molise	14
Campania	15
Puglia	16
Basilicata	17
Calabria	18
Sicilia	19
Sardegna	20

- **Homoscedasticity**: LDA assumes an equal variance among the different predictor variables, while in QDA the assumption of a common variance/covariance matrix across classes does not hold.
- **Multicollinearity**: the predictor variables should not be significantly correlated

As seen above, these assumptions are not met and they are expected to perform poorly or at least worse than other classification algorithms, such as k-NN and Random Forest.