# Student Alcohol Consumption

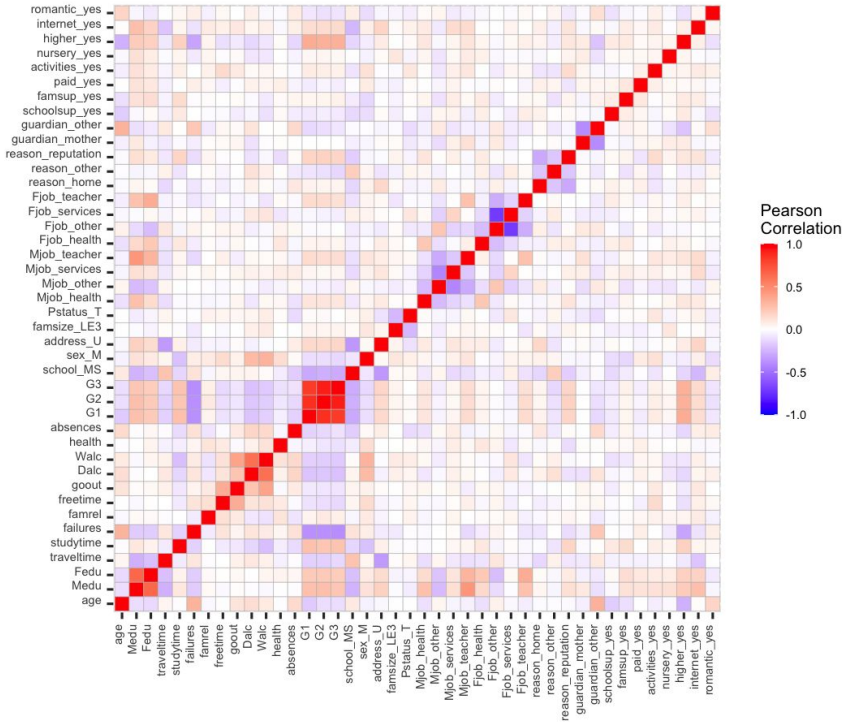
## Data Analysis Course, Mines ParisTech

Samuel Diebolt, 3rd year

# Dataset description

The dataset was obtained in a portuguese survey of student enrolled in a portuguese language course in secondary school. The selected features are described here.

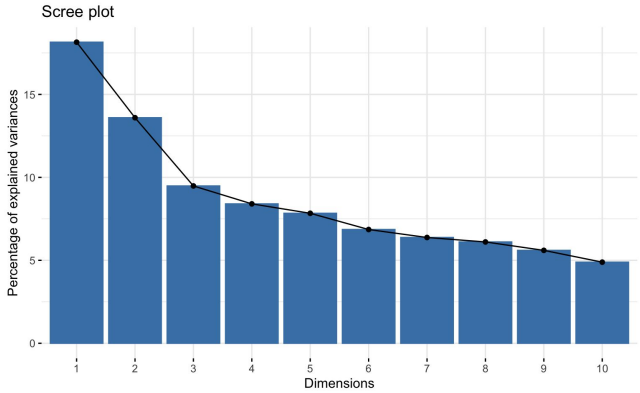
Looking at the correlation matrix on the right, grades are highly correlated. Since including nearly-redundant variables can cause methods like PCA to overemphasize their contribution, we choose to keep only the mean of these grades.

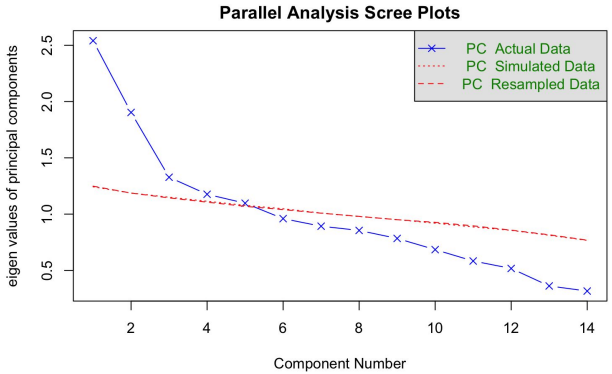


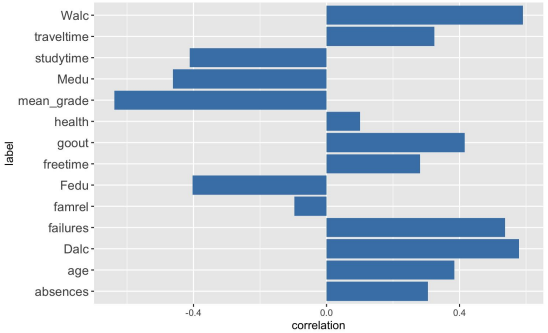
# PCA on quantitative variables

Without expert opinion on the dataset, we cannot determine the number of relevant factors in advance. Thus, it makes more sense to use PCA over FA to get a first look at our data.

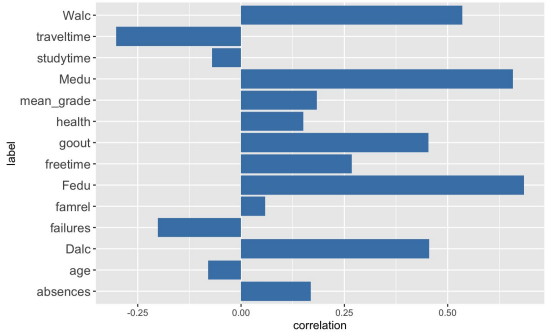
Looking at the scree plot on the right, a sharp break isn't obvious. Horn's parallel analysis was used to determine the number of components to keep. Therefore, interpretations for the first five principal components will be given in the next slide. Detailed explanations are given in the R notebook accompanying these slides.



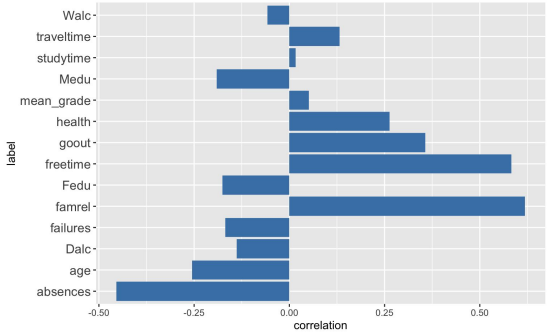





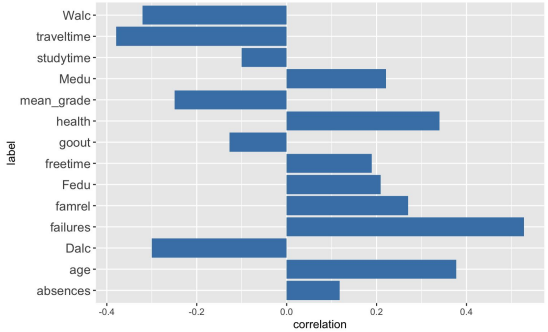

**PC1:** *school failures caused by alcohol consumption, lack of studying and/or parent's lack of education.*



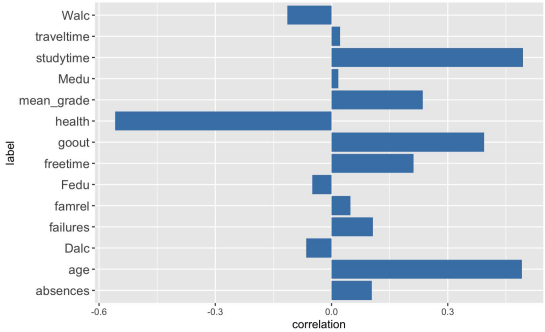

**PC2:** *social status.*




**PC3:** *family relationship.*




**PC4:** *school failures that are not caused by alcohol consumption or lack of studying.*



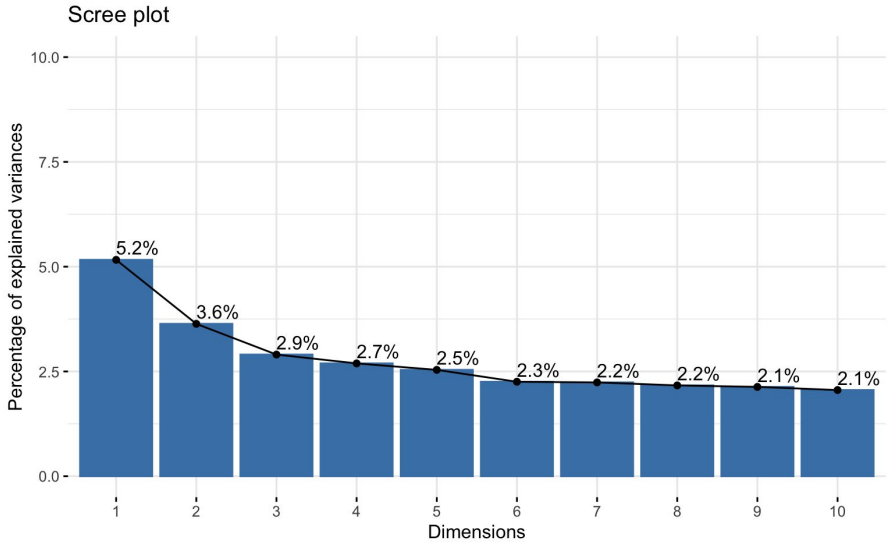

**PC5:** *older student who might be exhausted from studying, or might not be happy with their environment due to their older age.*

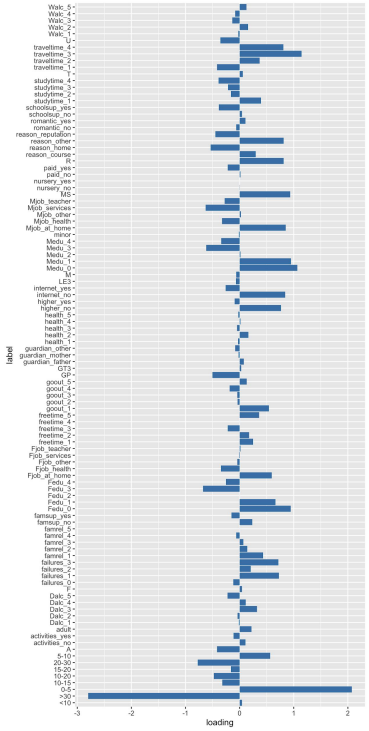
# Conclusions on PCA

- These five principal components only explain approximately 57% of the variance in the data and their interpretations were relatively vague.
- Factor analysis could have lead to better interpretations, in particular using methods of loading rotation such as varimax. However, the lack of expert opinion on the right number of factors lead us to use PCA, so as to give a first interpretation of the possible latent factors.
- PCA will now be compared to the next method of analysis: multiple correspondence analysis (MCA).

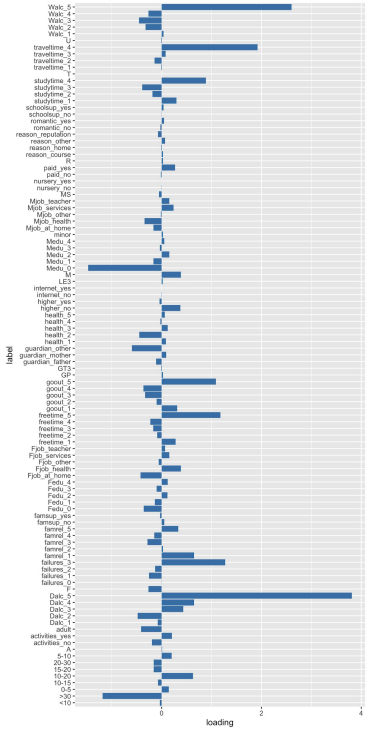
# MCA on categorical variables

- Since the dataset contains mixed data, i.e. nominal, ordinal and numerical features, it was preprocessed by converting quantitative variables to ordinal.
- The following variables were converted:
  - *age*: "minor", "adult";
  - *absences*: "0-10", "10-20", "20-30", ">30";
  - *mean grade*: "0-5", "5-10", "10-15", "15-20".
- MCA was then performed using 5 factors, so as to compared results with PCA.

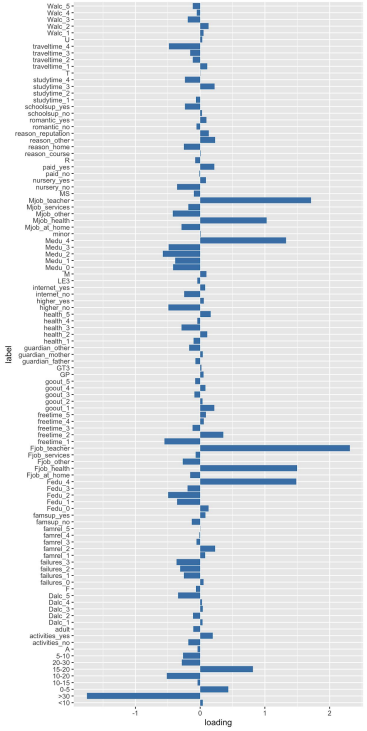




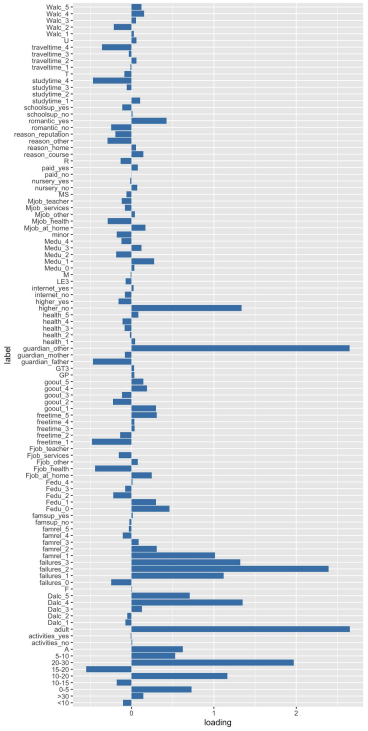

**Factor 1:** *low number of absences (0-5)*



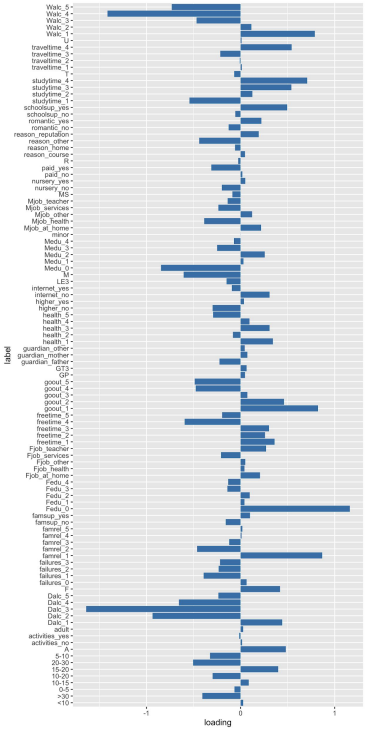

**Factor 2:** *high consumption of alcohol, in particular during the work days.*




**Factor 3:** *parents education.*




**Factor 4:** *adult students, possibly living by themselves, having relatively poor results in other classes.*




**Factor 5:** *families with low social status.*