Log-book of cyber-warriors

# 04.04, till 15:15

Computed the set of missing values:

A black screen with white text

Description automatically generated with low confidenceA black screen with white text

Description automatically generated with low confidence

Read about missing values [here](https://scikit-learn.org/stable/modules/impute.html), we can either do univariate (mean/mode/median) or multivariate imputation (reggresion).

Created a python script to clean dataframes, with two funcions:

1. Change\_to\_cat
2. Check\_almost\_duplicates (does not work yet)

# 13.04, till 23:00

One problematic thing : it might be hard to know if there are missing values in “high\_income” (which is the response variable), since the column is all 1s and 0s… (any 0 could be a missing value).

Classifying each row by missing column:

Chart, histogram

Description automatically generated

**We have to choose a threshold to remove the rows**