# Exercice 1 - EDA and Data Cleansing

Using the data named: **hrm-employee-attrition\_unclean.csv**, Please meke following steps:

1. Use SweetViz to make the initial EDA.
2. Check each variable using graphs grouping those graphics by the data types (categorical-nominal, ordinal, and continuous).
3. Use boxplots to detect which variables have extreme values (outliers).
4. By using graphs (scatterplots) and statistical analysis, check the correlation between the numeric variables.
5. By using graphs (histograms/distribution plots) and statistical analysis, check the difference of the numeric variables within categories of different categorical variables.
6. Using PCA, check if there are subgroup of individuals that can be identified. You can also use Cluster analysis for this task.
7. Check the missing values. Which variables have to be deleted? Which variables may need to be transformed into categories? Which variables need to be evaluated for their missing mechanism?. Explain why you decided to use each method.
8. Clean the data from outliers. Describe which methods are the correct for each variable.
9. Clean the data for missingness. Which variable will need to be deleted? Which variables will be converted to categories, and which variables will be imputed? How do you determine which variables can be imputed and which can’t be imputed?
10. Check again the data with the automatic EDA (SweetViz)