**Flourishing in the Workplace – An exploration**

A company wishes to study the work-life balance of its employees. To do so, they have collected data from 248 employees. You are in charge of analyzing those data and to make a written report on your conclusions, as well as of making recommendations if appropriate.

1. The data are available in an Excel file named « Base\_248\_HD.xlsx »

The variables of this dataset are the following:

* 1. ID : identification number of the respondent (a number)
  2. Sex : the gender of the respondent
     1. The value 1 indicates a male
     2. The value 2 indicates a female
  3. famstatus : the family status of the respondent
     1. 1 : Single
     2. 2 : Separate / Divorced
     3. 3 : Couple
  4. education : The educational level of the respondent
     1. 1 : Has not completed high school degree
     2. 2 : Has completed only a high school degree
     3. 3 : Has completed up to 2 years after high school degree
     4. 4 : Has completed up to 3 years after high school degree
     5. 5 : Has completed up to 4 years after high school degree
     6. 6 : Has completed 5 years after high school degree or more
  5. age : age of the respondent
  6. prof\_cat : the flourishing state of the respondent in an occupational context
     1. Languishing (value = 1)
     2. Moderately Mentally Healthy (value = 2)
     3. Flourishing (value = 3)
  7. priv\_cat : the flourishing state of the respondent in a private context
     1. Languishing (value = 1)
     2. Moderately Mentally Healthy (value = 2)
     3. Flourishing (value = 3)
  8. positivity : the positivity ratio value, i.e., the ratio between the average positive emotions, over the average negative emotions
     1. If the ratio is lesser than 1, the individual is diagnosed as depressed
     2. If the ratio is between 1 and 2, the individual is languishing
     3. If the ratio is between 2 and 3, the individual is moderately emotionally healthy
     4. If the ratio is greater than 3, the individual is emotionally flourishing
  9. prof\_quant : the total score of the respondent’s mental flourishing in a professional context
  10. priv\_quant : the total score of the respondent’s mental flourishing in a private context
  11. flow : the Flow score of the respondent, measured as performance, mastery and focus

**What is Statistical Learning?**

In Data Science before you do anything you first start by exploring your dataset. The basic steps in data exploration are following:

**Step 1 - Univariate Data Analysis**

1. If the variables are categorical.
   1. Generate the summary table of each variable.
   2. Plot their Pie Chart
   3. Plot their Bar Chart
2. If the variables are quantitative.
   1. Plot their histogram
   2. Plot their boxplot

**Step 2 – Bivariate Data Analysis**

1. If the variables are both categorical.
   1. Generate the contingency table
   2. Check the significance of their relationship with the chi-square test & provide Cramer’s V
   3. Plot their side-by-side bar charts
   4. Plot their stacked bar charts
2. If the variables are both quantitative.
   1. Compute the correlation (table)
   2. Check the significance of their relationship with the correlation test & provide the r value
   3. Plot their scatter plot (matrix)
3. If the variables are mixed categorical & quantitative.
   1. Compute the anova table
   2. Check the significance of the difference in values between groups
   3. Plot the grouped boxplots

**Question:**

Which variables could be selected as outcome (target, response, dependent variable) or variable to be explained from the others?

Justify your suggestions with statistical & theoretical arguments.

**Step 3 – Modeling**

Once you have selected an outcome among the variables of your dataset, the other variables become the predictors. Now the goal is to find a function that captures in the best possible way the relationships between the outcome and the predictors. This process is called “modeling” and statistical learning is one way of addressing it.

One goal of modeling is thus to explain the variability (or variance) in the outcome from the predictors. This approach to modeling is associated with “supervised learning” in statistical learning.

However, this variance may not be explained by the predictors only, i.e., by the relationship between the predictors and the outcome only. It may also be explained by the existence of subgroups of individual instances or subgroup heterogeneity. The process to account for these subgroups is associated with “unsupervised learning” in statistical learning.

On other occasions, the number of variables is too numerous, many of which are associated (correlated or linked) in various ways. In such cases, the relationships between the outcome and the predictors can become confusing, opaque or too complex. To reduce this complexity and to simplify interpretability, dimension reduction is recommended. The set of tools to reduce dimensions is also associated with “unsupervised learning” in statistical learning.

**Practice**

**Goal**

Apply the three steps of data analysis on the <Flourishing> dataset.

**Exercise 1**

Explore univariate descriptive analysis.

**Exercise 2**

Explore bivariate descriptive analysis.

**Exercise 3**

Apply supervised learning on a quantitative outcome variable (recommended: Flow). With a linear regression model and a K nearest neighbors regressor (KNN) model

**Exercise 4**

Apply supervised learning on a categorical outcome variable (recommended: Flow). With a logistic regression model and a K nearest neighbors classifier (KNN) model

**Exercise 5**

Apply unsupervised learning on the observations (rows).

Explore links between the emerging classes and the flourishing states and use it to draw a final interpretation.

**Exercise 6**

Apply unsupervised learning on the variables (columns).

See If you can draw some interpretation…