Forecasting Task

We provide an artificially modified data sample about a set of contracts.

The **total gross** is the final gross salary that we will pay to the worker at the end of his/her working period. This information varies at the end of the period, since the worker often perform a different number of hours than originally planned, affecting the final payment.

The current method to estimate the total gross of a worker (`est_total_gross`) is not always very effective. The goal of this task is to build a **model that forecasts the estimated total gross** with the available data.
However, first a a **data analysis** phase need to be provided, to show which are the tools and the steps required to study and understand the variables and their relations. Then, **data insights** about the features used, such as their predictive power, should be presented and commented.

As a preliminary analysis, it's important to understand the data and as proof-of-concept we require to answer to the following points:

- Always highlight any data inconsistency or data mistake you find (if there are).
- Any particular pattern among the salary values? (for `job_function` , `company`, period of the year, *etc.*)
- Any interesting relation between the `fee_percentage` and the `job_function` or `company`?
- Which are the `job_function` and `companies` that required to work the most on the weekend?
- Could you find patterns for the cancelled contracts (when `cancellation_date.notnull()`)?
- Any personal *insight* provided is a plus...

Data

The dataset is available here.

Load the dataset into a Python dataframe and make the necessary pre-processing and cleansing steps.

There are various columns that can be used for the task, but the most important are the following:

- `total_gross`: it's the target variable, the value we want to predict.

- `est_total_gross`: it's the current "prediction", you can consider it as a baseline.
- `hiring_*`: this prefix refer to the contract
- `wp_*`: this prefix refer to the "work period" that is a subset of the contract, in general it depends about the country's law (*e.g.*, in Spain the majority of monthly contracts will have a `work_period` for each month, it's basically the "nomina". In UK it usually has a weekly frequency.)

All the columns have been artificially generated or carefully anonymized in order to maintain their predictive power.

Evaluation

The result should be presented in a **Jupyter Notebook in Python**.

The evaluation will account for:

- Cleaning and documentation of the various steps.
- Data analysis and feature insights.
- Motivation about the decision made (normalization, ML model, metrics chosen, *etc.*).
- Accuracy of the forecasting.
- Any other personal comment and initiative.

Note: in case of doubts about the data we require the candidate to make her/his assumptions without being blocked.