**Working with Credit One Data**

This project is being executed with the aim of developing a model capable of predicting how much credit to allow someone or, at least, whether they should be approved or not. Since Credit One is an external client and stores all their data in a MySQL database, we had to query it through Structured Query Language and retrieve the data into a Pandas dataframe.

As potential business values, I can highlight the ability to create a programming environment with the right libraries. Downloading, using, adapting. While it is true that there are many types of solutions to a problem when it comes to programming. Between learning new coding/applications, as well as adapting ready-made solutions but according to the problem to be solved, the objective of the study can be achieved.

In addition, the dedication to understanding the business problem and the interactions of all the variables extracted from the Credit One data is noteworthy, as it facilitates the exploration of the data to find the required quality insights. Some of the most important lessons learned are shown below:

* Previously we worked with a csv file where the table it contained was easier to understand. In the case of Credit One I did learn the fact that the time to study a database is variable, so we cannot expect to have a recipe when it comes to understanding the different types of databases.
* When I started the project, I started running different queries to the database without having used Pandas-Profiling from the beginning, which quickly creates an EDA report, and in much less time than a more manual study. From this report, we identified variable definition errors, repeated rows, inadequate data types, etc...
* When working with a database of so many variables, a very good level of order is required to discard and segregate all the really valuable data.
* Given that from the beginning we have the categorical dependent variable that we want to predict with our model, it is easy to start experimenting with all types of analyses and contrasts of this variable with all the others. This in order to identify which features have the most influence on the behavior of the dependent variable.

Regarding my findings, after plotting the correlation chart for all variables, we found that LIMIT\_BAL, SEX, EDU, AGE and MARRIAGE are not highly correlated to each. PAY\_0, PAY\_2, PAY\_3, PAY-4, PAY\_5, PAY\_6 are correlated much higher degree than other variables. In this way, we can have a brief thought is that those people who can pay the bill on time will highly be possible to pay next bill one time. The probability of their late payment is mainly based on their previous behaviors instead of their characteristics.