**Detect counterfeit notes using regression analysis**

This project was conducted as part of my OpenClassroom Data Analysis bootcamp. I made and then used a OLS regression model as well as classification analysis to predict weather certain banknotes are false using the six variables given of each bank note.

There is one file attached for the whole analysis. It is made in Python in a Jupyter notebook. There is also a presentation attached.

**Data Preparation**

I generated 4 dataframes out of the original data. One scaled version, one logarithmic version, and one PCA version. I eliminated variables that weakened the performance of a regression model based on their p value.

**Data Analysis**

I then selected the dataframe that worked best with the regression models. I later on trained both the regression model and k-means models on the selected dataframe, which was the scaled dataframe, and I used the model to then predict the legitimacy of each note.

Instructions:

1. Open folder: Analysis.ipynb
2. Run all cells.

**The project description:**

Your IT consulting company has been given a new mission by The Bureau of Engraving and Printing (BEP), a government agency within the U.S. Department of the Treasury as part of the fight against organized crime. Your mission, should you accept it, is to create a **counterfeit detection algorithm**.

**The data**

The BEP sent you a dataset containing the geometric characteristics of banknotes. For each of them, you'll find:

* The length of the note (in mm).
* The height of the note (measured on the left side, in mm).
* The height of the note (measured on the right side, in mm).
* The margin between the top edge of the note and the main image (in mm).
* The margin between the bottom edge of the note and the main image (in mm).
* The diagonal measurement of the note (in mm).

**Your tasks**

The BEP sent you the following **specific brief:**

**Task 1**

Briefly **describe** the data at hand with univariate and bivariate analyses.

**Task 2**

Carry out a **principal component analysis** of the sample :

* Analyzing eigenvalues scree plot.
* Representing variables using correlation circles.
* Representing individuals using factorial planes.
* Analyzing the representation quality and the contribution of individuals.

For each of these steps, comment on the results. The variable giving the genuine/counterfeit nature of the note will be used as an illustrative variable.

**Task 3**

**Apply a classification algorithm and analyze the result**. Visualize the resulting scores in the first factorial plane of the PCA and analyze them.

**Task 4**

**Model the data**. You will then create a program that can determine if a note is genuine or counterfeit.