Résumé des méthodes de l’analyse 2 Analyse Bivarié quanti-quanti :

Table des matières

1. Introduction

- Le datascientist, dans son travail doit réaliser plusieurs types d'étude sur les jeux de données qui l'intéresse. Parmi elles, les analyses bivariés permettent de mettre en lumière les potentielles corrélations entre les différentes variables du jeu de donnée.

Nous verrons donc tout d'abord la réalisation de l'import des données puis la manière dont un data analyste réalise des analyses bivariées et plus particulièrement, des analyses bivariées entres variables quantitatives.

Dans le cadre de ce TP, les packages « Pandas », « Scipy », « sklearn », « Numpy » et « matplotlib » ont été utilisés. Le jeu de donnée DepensesEduData.csv, représentant les dépenses de 26 pays en fonction de leur nombre d'étudiants, composera la base de notre étude.

2. Import des données

Afin d'apporter les données à étudiant dans python, on utilise le package Pandas qui simplifie la création et la manipulation de dataframe.

Suite à l'import des packages,

```
# import des packages
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import pearsonr
from sklearn.linear_model import LinearRegression
from scipy.stats import linregress
```

Il suffit d'utiliser read_csv, une fonction de pandas, pour lire un csv et le récupérer sous forme de dataframe exemple :

```
# Création du dataframe
df = pd.read_csv("data\DepensesEduData.csv", sep=";", header = 0, index_col=0)
```

Il en sort notre dataframe sur Python.

	nbEleves	Depenses
Allemagne	14065,4	106626,4
Autriche	1467,8	13722,8
Belgique	2427,7	18427,4
Bulgarie	1141,8	2968
Danemark	1151,6	12909,4

3. Exploration et travail sur les données.

Il est commun de devoir réaliser des modifications sur les jeux de données afin d'assurer la qualité et la pertinence de celles-ci. Dans notre cas certaines valeurs sont définies en tant qu'objet au lieu de floats. Nous devons donc modifier les datas pour les rendre utilisables dans le cadre d'une analyse bivariée quantit quantit.

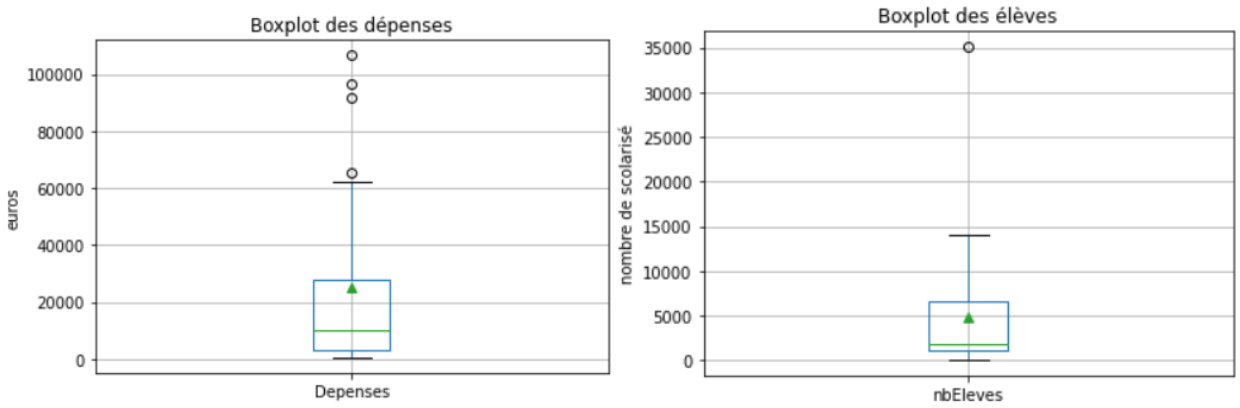
```
# On voit ici que les données sont de type objet, or on voudrait des float, on
remplace la virgule par un . pour le changement de type
def format(df, col):
    for i in range(len(df[col])):
        df[col][i] = df[col][i].replace(',', '.')
    df[col] = df[col].astype(float)
    return df
```

4. Etude bivariée

Il existe plusieurs moyens plus ou moins pertinent suivant les cas, de comparer les variables quantitatives entres elles.

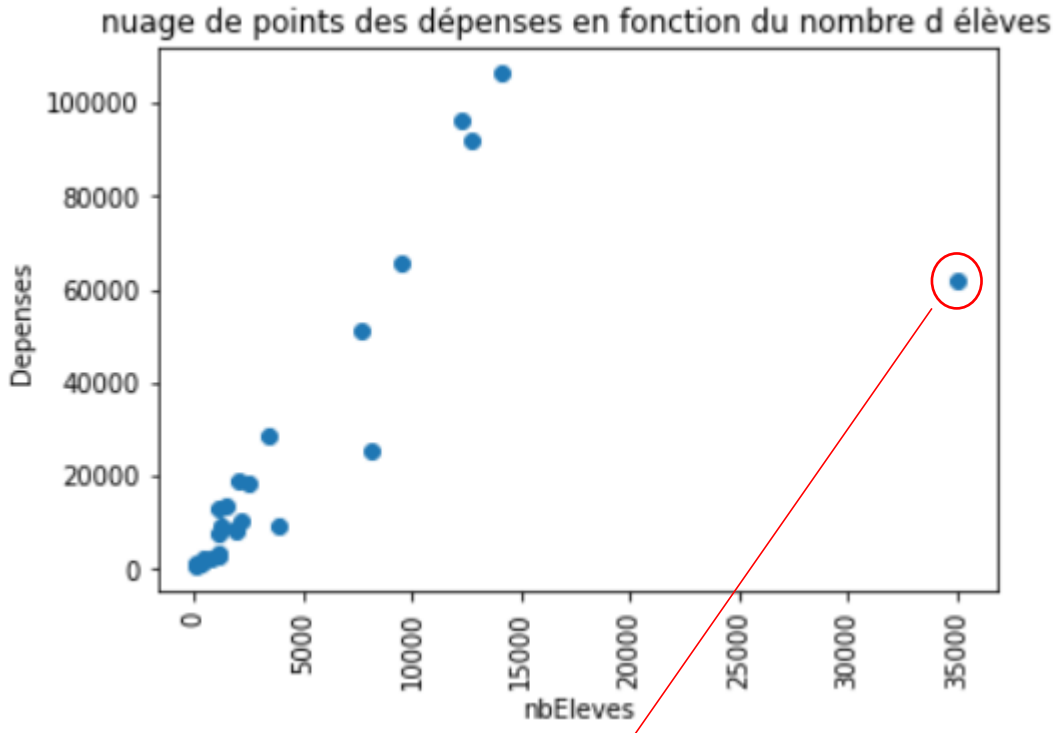
4.1. Graphiques

On peut ainsi réaliser des graphiques pour rendre l'analyse de données plus visuel. L'utilisation de la boîte de Tukey (ou boite à moustache) est la plus répandue du fait de sa clarté et son efficacité.

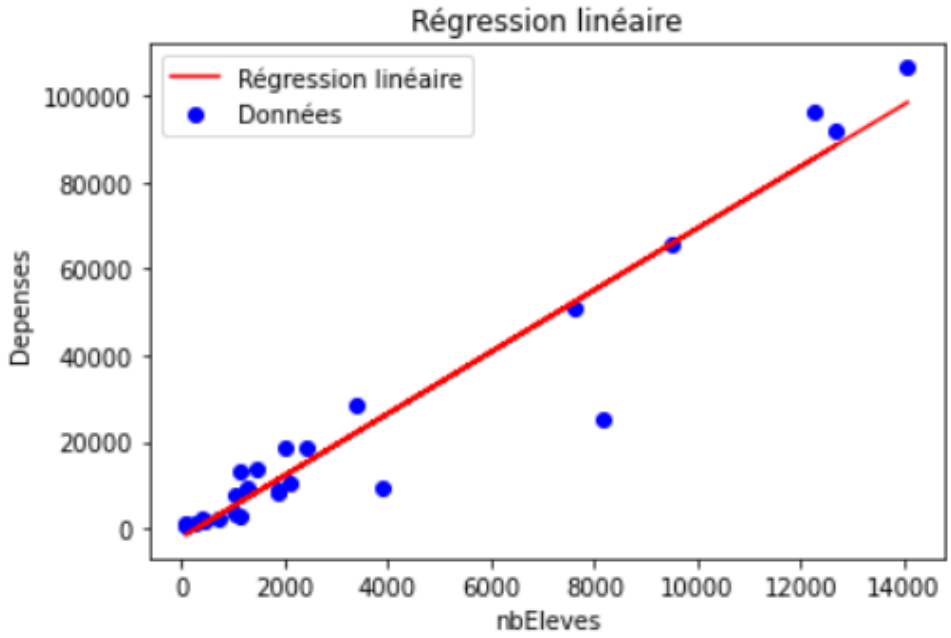


Ces deux graphes nous donnent déjà beaucoup d'information. Il existe une valeur aberrante. Un pays possède énormément d'étudiant scolarisés. On remarque également que la moyenne est assez éloignée de la médiane, impliquant un indice de dispersion élevé. (Grande différence entre les deux extremums).

On peut également réaliser un nuage de point pour avoir un rendu visuel de la position de chaque pays en fonctions des deux variables (nbEtudiant et dépenses).



Une valeur semble perturber le graph puisqu'un pays semble avoir bien trop d'étudiant. On « nettoie » ce dataset de cette valeur aberrante et réalisons une régression linéaire pour vérifier visuellement l'existence d'une potentielle corrélation entre les variables :



Les points semblent alignés de manière plutôt logique sur la droite de régression, il semble qu'il existe une corrélation entre la variable représentant le nombre d'étudiant et les dépenses du pays dans l'établissement.

4.2. Régression Linéaire calculatoire

On réalise une régression linéaire pour obtenir des résultats calculatoires :

:

```
# seconde méthode de régression linéaire ; scipy stats
lr = linregress(df_propre['nbEleves'], df_propre['Depenses'])
lr
```

Et on obtient :

```
LinregressResult(slope=7.153620702871415, intercept=-2093.810191533619, rvalue=0.966056406103996,
pvalue=5.13084626980696e-15, stderr=0.39887462128097034, intercept_stderr=2230.1956987465246)
```

La « rvalue » vaut 0.966 révélateur d'une très grande corrélation entre les variables représentant le nombre d'étudiants et celle des dépenses en euro des pays dans l'éducation.

(+ c'est proche de 1 plus c'est corrélé).