TD : Utilisation de Pandas

Analyse du jeu de données : "Parcours et réussite des bachelières et bacheliers inscrits pour la première fois en licence"

Ce jeu de données a été obtenu à la page suivante :
https://www.data.gouv.fr/datasets/parcours-et-reussite-des-bachelieres-et-bacheliers-inscrits-pour-la-premiere-fois-en-licence-donnees-consolidees/

Travail préliminaire

Importation des bibliothèques

On commence par importer les bibliothèques que l'on va utiliser.

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Importation des données

On peut maintenant importer le jeu de données sur lequel on va travailler.

Importez le fichier "fr-esr-parcours-et-reussite-des-bacheliers-en-licence.csv".

Attention, il faut préciser que le séparateur est le point-virgule : « ; », et non la virgule « , ».

In [ ]:
```

```

Découverte des données

Affichez les dimensions du dataframe.

In [ ]:

Affichez les 5 premières lignes.

In [ ]:

Affichez les informations générales concernant les colonnes.

In [ ]:

Nettoyage des données

Suppression des données inutiles pour l'analyse

D'après les titres des colonnes, le jeu de données contient deux types d'informations :

- colonnes 14 à 22 : des statistiques sur le passage de la L1 à la L2
- colonnes 23 à 28 : des statistiques sur l'obtention de la licence (i.e. la validation de la L3).

Ici, on ne va s'intéresser qu'aux statistiques de passage de la L1 à la L2.

Supprimez les 5 dernières colonnes qui ne sont pas utiles pour notre analyse.

In [ ]:

Affichez de nouveau les informations générales concernant les colonnes pour vérifier que la suppression a bien eu lieu.

In [ ]:

Suppression des données redondantes

Les titres des colonnes suggèrent que :

- Les six premières colonnes indiquent la filière de chaque ligne.
- La filière est divisée en "Grande discipline", puis en "Discipline", puis en "Secteur disciplinaire".
- Chacun de ces trois niveaux est représenté par un identifiant (une chaîne de caractères ou un entier) et un intitulé.

Pour vérifier qu'on comprend bien l'organisation des données, nous allons vérifier que chaque identifiant est associé à une seule catégorie.

Remarque : d'après les informations affichées précédemment, doit-on se méfier des données manquantes pour ces six colonnes ?

Réponse :

1. analyse des colonnes "Id Grande discipline" et "Grande discipline" :
Pour vérifier que chaque "Id Grande discipline" est associé à une seule "Grande discipline" (et vice-versa), affichez le nombre de valeurs différentes présentes pour les colonnes "Id Grande discipline" et "Grande discipline", ainsi que pour les couples ("Id Grande discipline", "Grande discipline").

In [ ]:

2. analyse des colonnes "Id Discipline" et "Discipline" :
Procédez comme précedemment.

In [ ]:

3. analyse des colonnes "Id Secteur disciplinaire" et "Secteur disciplinaire" :
Procédez comme précedemment.

In [ ]:

Affichez les 6 premières colonnes en les triant d'abord par "Id Grande discipline", puis par "Id Discipline" et enfin par "Id Secteur disciplinaire".

In [ ]:

Chaque niveau de filière est représenté par un identifiant et un intitulé. Cela génère de la redondance d'informations que nous allons éliminer. Cependant, afin de garder toute l'information, nous allons créer des dictionnaires qui associent les identifiants aux intitulés.

Écrivez une fonction "make_dict" qui prend en paramètre le nom de deux colonnes d'un dataframe, et renvoie un dictionnaire des éléments de la première vers ceux de la seconde.

Utilisez cette fonction pour créer un dictionnaire associant chacune des colonnes "Id Grande discipline", "Id Discipline" et "Id Secteur disciplinaire" respectivement aux colonnes "Grande discipline", "Discipline" et "Secteur disciplinaire". Affichez-les afin de les vérifier.

Aide : consultez la documentation des fonctions "drop_duplicates", "set_index" et "to_dict".

In [ ]:

Maintenant que l'on peut retrouver les intitulés des différents niveaux de filière à partir de leurs identifiants, on peut enlever les colonnes correspondantes dans le dataframe, pour limiter la taille de données à traiter.

Supprimez les colonnes contenant les intitulés des grandes disciplines, disciplines, et sections disciplinaires. Afficher la liste des colonnes pour vérifier la suppression.

In [ ]:

De manière similaire, créez les dictionnaires suivants :

- "Id Série ou type de Bac" -> "Série ou type de Bac"
- "Id Âge au bac" -> "Âge au bac"
- "Id Sexe" -> "Sexe"
- "Id Mention au Bac" -> "Mention au Bac"

Une fois ces dictionnaires créés, supprimez les colonnes avec les intitulés.

In [ ]:

Traitement des données manquantes

Affichez les 10 premières lignes du dataframe.

In [ ]:

On observe qu'il y a une ligne pour laquelle il n'y a pas de données. Cette ligne correspond aux hommes ayant passé un bac professionnel avec mention bien qui se sont inscrit en chimie : sans doute il n'y a eu aucun élève dans ce cas en 2014, d'où l'absence de données.

Remplacez tous les NaN par des 0.0. Affichez de nouveau les 10 premières lignes du dataframe pour vérifier que le remplacement a bin été fait.

In [ ]:

Suppression des données inutiles

Que contient la colonne "Année de cohorte des données sur le passage entre L1 et L2" ? À votre avis, pourquoi seule la valeur "2014" apparaît ?

In [ ]:

Réponse :

Comme cette colonne n'est pas informative, supprimez-la.

In [ ]:

Donnez le nombre d'étudiants qui suivent des études de santé. (Id Secteur disciplinaire = 7).

In [ ]:

On observe qu'il y en a très peu. On peut imaginer, sans certitude, que cela est dû au changement de la filière de médecine, qui fait l'objet de statistiques séparées. Les quelques lignes restantes doivent être un reliquat, ou des erreurs d'inscription.

Supprimez toutes les lignes correspondantes à ces étudiants.

In [ ]:

Nous en avons fini avec le nettoyage du dataframe, nous allons pouvoir commencer à travailler avec les données !

Analyse du dataframe

Taille de la cohorte

Combien d'élèves sont présent·e·s dans l'échantillon que nous avons ?

In [ ]:

Influence de la mention sur la validation de L1

Trouvez le nombre de passage en L2 en 1 an, le nombre de redoublement en L1 et le nombre de passage en L2 en 2 ans pour chaque mention.

Gardez le résultat dans une variable *statistiques_par_mention*.

In [ ]:

On va maintenant regarder si la filière d'origine (= série du bac) des élèves a une influence sur la validation de la L1 (en 1 an et en 2 ans confondus), en fonction de la filière universitaire (= la grande discipline) suivie à l'université. Ces données vont être représentées dans un tableau à double entrée.

Dans un premier temps, réfléchissez *à la sortie que vous voulez obtenir* :

- quelles données souhaitez-vous représenter en colonne et en ligne ?
- comment pouvez-vous calculer les données présentes dans le tableau ?

On veut obtenir un tableau à double entrée avec, en ligne, la filière universitaire et, en colonne, la filière d'origine. Les données présentes seront le taux de passage en L2 pou chaque couple (filière universitaire, filère d'origine).

Dans un second temps, réfléchissez à comment y parvenir et codez votre solution.

Gardez le résultat dans une variable *statistiques_par_filiere*.

In [ ]:

Modifiez le résultat pour que, dans le tableau, les idenfitiants des grandes disciplines et celui des séries du bac soient remplacés par les intitulés correspondants (utilisez les dictionnaires précédemment créés).

In [ ]:
```
df_statistiques_par_filiere.rename(index = gd_disc_dict, inplace = True)
df_statistiques_par_filiere.rename(columns = serie_dict, inplace = True)
df_statistiques_par_filiere
```

In [ ]: