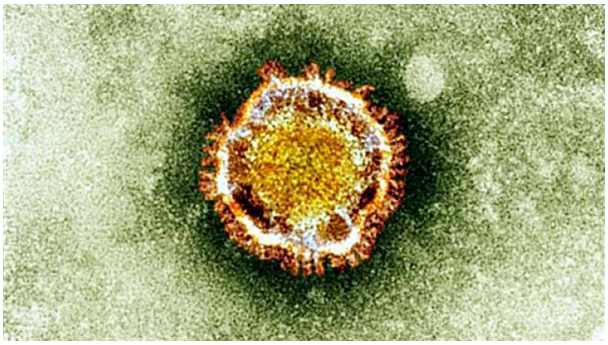
COVID-19

1. Présentation de l'activité



- Quotidiennement, des données sont récoltées dans le monde entier par les chercheurs des pays touchés pour pouvoir suivre et enrayer la propagation de la pandémie de Coronavirus.
- Sur la page suivante , on trouve des fichiers qui rassemblent ces données : https://www.data.gouv.fr/fr/datasets/coronavirus-covid19-evolution-par-pays-et-dans-le-monde-maj-quotidienne/ (https://www.data.gouv.fr/fr/datasets/coronavirus-covid19-evolution-par-pays-et-dans-le-monde-maj-quotidienne/)
- Les objectifs de cette activité sont :
  - De découvrir et d'utiliser une nouvelle bibliothèque.
  - D'utiliser un fichier au format **csv** depuis une URL, ce qui permet une consultation en temps réel.
  - Y extraire des données et trier des données selon des critères définis.
  - Réprésenter graphiquement l'évolution de la maladie par pays et par jour.

2. Pré-requis

Cette activité nécéssite l'utilisation des bibliothèques suivantes:

- `pandas` qui permet de travailler avec des bases de données.
- `matplotlib` et plus particulièrment le module `pyplot` pour les représentations graphiques.

Exercice 1 :

Exécuter la cellule ci-dessous.Si rien ne s'affiche, c'est que les modules sont correctement installées. Si une erreur apparaît, c'est que l'un des modules n'est pas installé.Il faut alors suivre la procédure suivante :

- Ouvrir une console de commande.
- taper `pip install pandas`
- taper `pip install matplotlib`
- Exécuter de nouveau la cellule pour vérifier que les modules sont installés

In [2]:

```
import pandas
from matplotlib import pyplot
```

Exercice 2 :

Sur le site proposé dans la présentation ci-dessous, télécharger le fichier **"Evolution par jour et par pays (CSV)"**.

1. Trouver une application pour ouvrir ce fichier.
2. Que signifie **csv** ?
3. Quel caractère est le séparateur dans ce fichier ?

Réponses :

1.
2.
3.

A retenir

- Le sigle CSV signifie "Comma-Separated Values" et désigne un fichier informatique de type tableur, dont les valeurs sont séparées par des virgules(ou un autre caractère de ce type).
- Le format CSV est un format de texte simple qui est utilisé dans de nombreux contextes lorsque de grandes quantités de données doivent être traitées.
- Ce type de fichier peut être lu dans un tableur.
- La plupart des langages de programmation comprennent ce format de fichier.

3. Exploiter un fichier `.csv` avec `pandas`

Exercice 3 :

1. Sur la page web précédente trouver l'URL stable de ce fichier et la recopier en réponse
2. Compléter le programme ci-après en remplaçant les `?` .L'éxécution de ce programme doit renvoyer un extrait du fichier.

Réponses :

1.

In [ ]:

```
#2. Programme à compléter

#récupérer le dataset à l'aide d'une URL
data = pandas.read_csv(?,
                       skiprows=3,
                       sep=?)

#formater la colonne date
data['Date'] = pandas.to_datetime(data['Date'], format='%Y-%m-%d')

#on ne garde que les données postérieures à une date donnée
data=data.loc[data['Date'] > '2020-02-20']

#Affiche les 5 premières lignes du fichier avec les noms de colonnes
print(data.head(5))
```

Exercice 4 :

A l'aide du programme précédent ainsi que de cette documentation : http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Data_Manipulation_Pandas.pdf (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Data_Manipulation_Pandas.pdf) , répondre au questions suivantes :

1. Afficher tous les noms de colonnes.
2. Combien de lignes possède aujourd'hui ce fichier ?
3. Afficher uniquement les colonnes `'Date'` , `'Pays'` et `'Infections'` .
4. Afficher uniquement les données de l'Espagne et les stocker dans une variable. Combien de personnes y sont pour le moment infectées ?
5. Trier les lignes de cette table par ordre alphabétique des pays.
6. Afficher les noms de tous les pays concernés par cette pandémie(sans doublons).
7. Afficher les lignes des pays dont le nombre de personnes infectées est aujourd'hui supérieur à 5000.

In [1]:

```
#1.
```

In [2]:

```
#2.
```

In [3]:

```
#3.
```

In [5]:

```
#4.
```

In [6]:

```
#5.
```

In [7]:

```
#6.
```

In [8]:

```
#7.
```

Synthèse :

- `pandas` crée un objet appelé "dataFrame"(dans l'exercice précédent, il s'agit de la variable `data`), une sorte de tableau à double entrée composé de lignes et de colonnes.
- On peut ensuite extraire une partie des données en agissant sur les colonnes et les valeurs contenues dans celle-ci

4. Représenter les données avec `matplotlib`

Exercice 5 : Reprenons le dataFrame créé avec `pandas` :

In [33]:

```
#récupérer le dataset à l'aide d'une URL"
data = pandas.read_csv('https://www.data.gouv.fr/fr/datasets/r/f4935ed4-7a88-44e4-8f8a-33910a151d42',
                       skiprows=3,
                       sep=';')

#formater la colonne date
data['Date'] = pandas.to_datetime(data['Date'], format='%Y-%m-%d')

#on ne garde que les données postérieures à une date donnée
data=data.loc[data['Date'] > '2020-02-20']
```

1. Extraire les données pour la France

In [ ]:

```
#1.
datafr='à compléter'
print('datafr')
```

1. Compléter la liste `Xfr` pour qu'elle contienne les valeurs de la colonne `'Date'` et la liste `Yfr` pour qu'elle contienne les valeurs de la colonne `'Infections'` .Si les instructions sont correctes, une courbe doit apparaître après l'exécution de la cellule.

In [ ]:

```
#2.
Xfr='à compléter'
Yfr='à compléter'

#Pour éviter un message d'avertissement à l'affichage des courbes
from pandas.plotting import register_matplotlib_converters
register_matplotlib_converters()

#taille du graphique et légendes des axes
pyplot.figure(figsize=(20, 6))
pyplot.xlabel('Date')
pyplot.ylabel('Nombres de personnes')

#représentation des données
pyplot.plot(Xfr,Yfr,label='Infections-France')

#légende des courbes et grille
pyplot.legend(loc=0)
pyplot.grid()
```

1. Modifier le code de la question précédente pour ajouter la courbe des décès.

Exercice 6 :

Recopier ci-dessous le code précédent et ajouter les courbes des infections et des décès pour l'Espagne.

In [10]:

```
#Réponse
```

Exercice 7 :

Le script ci-dessous doit permettre de représenter les valeurs des colonnes `'Infections'` et `'Deces'` en fonction des valeurs de la colonne `'Date'`, et ce, pour n'importe quel pays choisi par l'utilisateur. Pour cela on crée deux fonctions :

- La fonction `donnees(pays)` qui prend en paramètre le nom du pays(chaîne de caractères) et qui renvoie un tuple de trois éléments contenant les valeurs des colonnes respectives Date, Infections, Deces.
- La fonction `graph(pays)` qui prend en paramètre le nom du pays(chaîne de caractères) et qui affiche les courbes dans le repère.
- A l'éxécution de ce script, les données pour les pays 'France', 'Italie' et 'Espagne' doivent s'afficher.

1. Compléter ces deux fonctions en remplaçant les caractères `?` jusqu'à ce qu'il n'y ait plus de message d'erreur

In [ ]:

```
#Réponse
import pandas
from matplotlib import pyplot

#Pour éviter un message d'avertissement à l'affichage des courbes
from pandas.plotting import register_matplotlib_converters
register_matplotlib_converters()

#récupérer le dataset à l'aide d'une URL"
data = pandas.read_csv('https://www.data.gouv.fr/fr/datasets/r/f4935ed4-7a88-44e4-8f8a-
33910a151d42',
                       skiprows=3,
                       sep=';')

#formater la colonne date
data['Date'] = pandas.to_datetime(data['Date'], format='%Y-%m-%d')

#on ne garde que les données postérieures à une date donnée
data=data.loc[data['Date'] > '2020-02-20']


def donnees(pays):
    D=data.loc[?]
    X=?
    Y=?
    Z=?
    return X,Y,Z

def graph(pays):
    X,Y,Z=?
    pyplot.plot(?,?,label='Infections-'+pays)
    pyplot.plot(?,?,label='Décès-'+pays)

pyplot.figure(figsize=(20, 6))
pyplot.xlabel('Date')
pyplot.ylabel('Nombres de personnes')

graph('France')
graph('Italie')
graph('Espagne')

pyplot.legend(loc=0)
pyplot.grid()
```

1. En modifiant le programme ci-dessus, comparer les données de la France et du Japon.

Exercice 8 : Un peu d'espoir

Par la méthode de votre choix, programmer l'affichage des courbes des infections et des guerisons pour la Chine depuis le début de l'épidémie.

In [ ]:

```
#Réponse
```

FIN (de la pandémie)