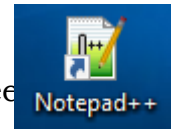
LE TRAITEMENT DE DONNEES EN TABLES

Partie 1 : Le fichier CSV

Vous regrouperez vos réponses dans un document word que vous nommerez « Traitement_Donnees_Tables_Nom_Prénom »

On trouve énormément de données sur internet. Une partie de ces données sont publiques, par exemple le site data.gouv recense un grand nombre de données publiques.

Explorez pendant quelques minutes le site data.gouv.



Sur ce site, on peut par exemple télécharger un fichier listant les lycée Normands. Vous le trouverez dans votre dossier sous le nom « **lycees25** »

- Dans quel format est enregistré ce fichier ?
- Ouvrez ce fichier avec un éditeur de texte, Notepad++ par exemple. Que remarquez-vous ?
- Enregistrez ce fichier dans le format « **lycees25.txt** ».

Voici ce que nous dit Wikipédia sur le format CSV :

***C**omma-**S**eparated **V**alues, connu sous le sigle **CSV**, est un format informatique ouvert représentant des données tabulaires sous forme de valeurs séparées par des virgules.*

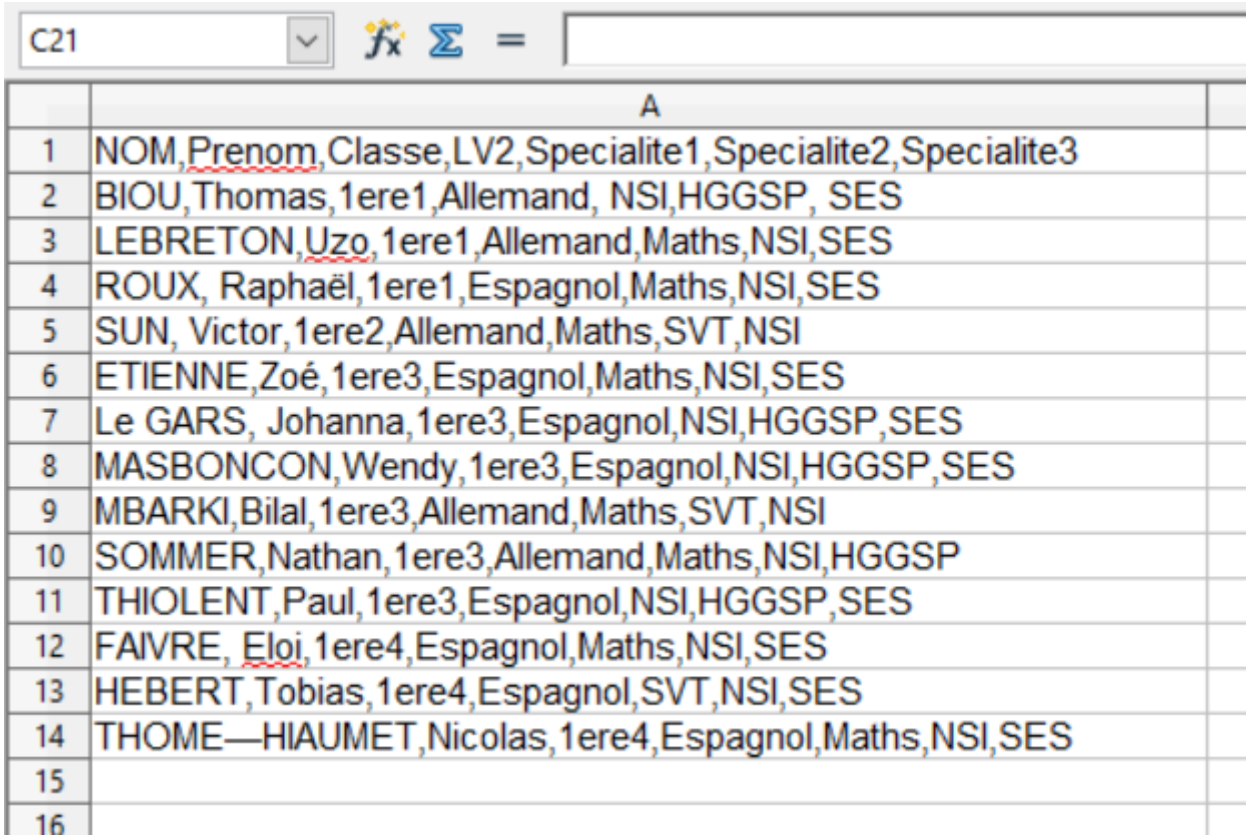
Un fichier CSV est un fichier texte. Chaque ligne du texte correspond à une ligne du tableau et les virgules correspondent aux séparations entre les colonnes. Les portions de texte séparées par une virgule correspondent ainsi aux contenus des cellules du tableau.

ATTENTION : La virgule est un standard pour les données anglo-saxonnes, mais pas pour les données aux normes françaises. En effet, en français, la virgule est le séparateur des chiffres décimaux. Il serait impossible de différencier les virgules des décimaux et les virgules de séparation des informations. C'est pourquoi on utilise un autre séparateur : le point-virgule (;). Dans certains cas cela peut engendrer quelques problèmes, vous devrez donc rester vigilants sur le type de séparateur utilisé.

- Ouvrez le fichier téléchargé qui liste les lycées Bas-Normands avec « Open Office », et enregistrez-le au format « .ods » sous le nom : « **lycees25_OO** ».
- Ouvrez alors le fichier nouvellement créé avec un éditeur de texte. Que remarquez-vous ?

Comment est structuré un fichier CSV ?

Prenons comme exemple le fichier « **Eleves_1eres_G2_NSI_2021-2022.csv** ».



Dans une même colonne, on regroupe l'ensemble des informations concernant un élève. Ces informations sont séparées par une virgule « , ».

"NOM", "Prenom" , "Classe" , "LV2", "Spécialité1" , "Spécialité2" et "Spécialité3", sont appelés « **descripteurs** » alors que, par exemple, "BIOU", "MBARKI" et "SOMMER" sont les « **valeurs** » du descripteur "NOM".

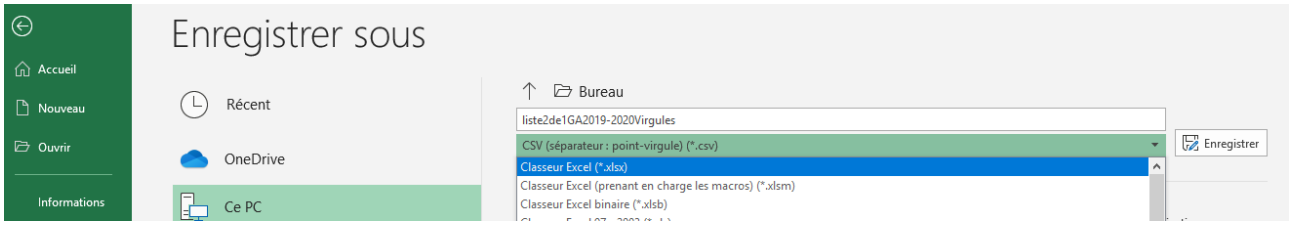
Donnez les différentes valeurs du descripteur "Spécialité2"

Ouvrez le fichier « **Eleves_1eres_Groupe_NSI_2021-2022.csv** » de votre Groupe.

Dans cet exemple qui ne comporte que 7 descripteurs, on peut encore lire assez facilement le contenu du fichier. Lorsque le fichier de données regroupe un très grand nombre de données, cela devient fastidieux et il est préférable de retrouver la mise en forme sous colonne comme le donne le fichier type excel.

Comment convertir ce fichier ?

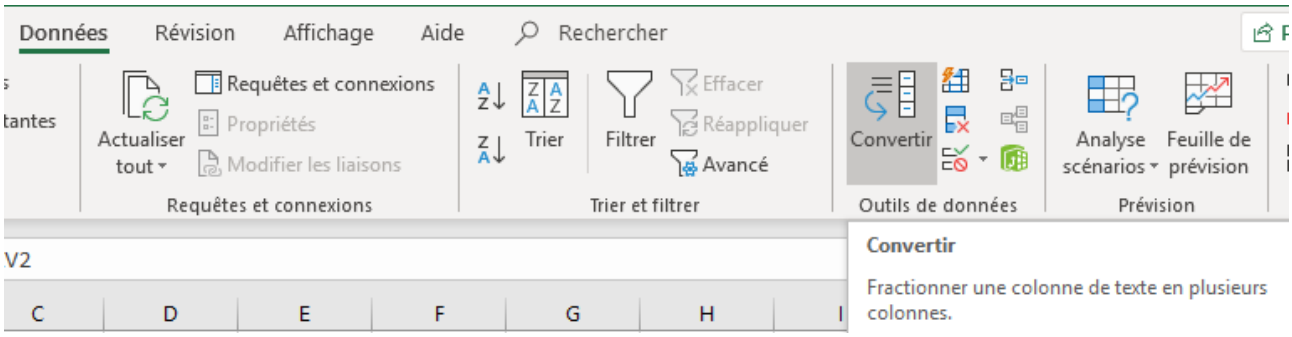
Pour convertir le fichier, il ne suffit pas de modifier le format d'enregistrement en procédant comme ci-dessous.



Réalisez cette procédure en enregistrant le fichier dans le format « excel » sous le nom « **Eleves_1eres_NSI_xlsx** » et commentez le résultat obtenu.

Pour obtenir une colonne par descripteur, il faut utiliser la fonction « **convertir** » accessible depuis le menu « **données** ».

Pour cela , vous devez sélectionner la colonne qui comporte les données.



Une page de dialogue (Etape 1) s'ouvre alors. Choisir « délimité » puis faire « suivant ». Dans l'étape 2, il faut renseigner le séparateur utilisé (ici, la virgule). L'étape 3 permet d'affecter le bon format de données pour chaque colonne. Les valeurs de chaque descripteur sont maintenant regroupées par colonnes. Enregistrer votre fichier sous le nom « **Eleves_1eres_NSI_Colonnes** »

Vous pouvez constater que les données sont bien "rangées" dans un tableau avec des lignes et des colonnes ; voilà pourquoi on parle de **données tabulaires**.

Il est possible de trouver sur le web des données beaucoup plus complexes à traiter comme par exemple la liste des 36700 communes de France.

Connectez-vous sur le site « **sql.sh** ». Téléchargez la liste des villes françaises au format « .csv », et enregistrez-le sous le nom « **Villes_France**».

Comment exploiter ce fichier ?

Vous pouvez vous rendre compte que le nombre de données est difficilement exploitable ! En augmentant la largeur de la colonne A, vous pouvez voir que toutes les données sont bien dans la même colonne !

Modifiez le fichier pour que les valeurs d'un même descripteur soient regroupées dans une même colonne. Enregistrez le fichier sous le nom « **Villes_France_Colonnes** ».

Donnez le nom, le code postal, l'altitude maximale, l'altitude minimale et le code de la commune où vous habitez.

Donner le code postal de la commune qui a la plus grande altitude. Expliquez la démarche utilisée.

Pour répondre à ces questions, vous avez dû lire une à une les 36700 lignes du fichier… ! Ou alors utiliser les fonctions de tri du tableur, entrainant parfois une modification de ce même fichier.

L'intérêt du fichier csv est que l'on peut l'utiliser dans un programme, pour effectuer un traitement automatique.

Nous allons utiliser le langage python pour lire un fichier csv réduit nommé « **Villes_Manches_csv**»

Ouvrez le fichier python « **Lecture_Villes_Manche_csv** » avec le logiciel PySripter et exécutez-le.

Quel résultat obtient-on ?

Le programme «**Ville_Altitude-Max** » permet de récupérer la ville dont l'altitude maximale est la plus grande. (On se limite au fichier réduit des villes de la Manche : « **Echantillon_Villes_Manche_csv** »).

Ouvrez le fichier « **Ville_Altitude-Max** » dans PyScripter, exécutez-le.

Que réalise la première partie du fichier ?

Ajoutez des commentaires sur cette première partie du fichier et enregistrez-le sous le nom : « **Ville_Altitude-Max _NOM_Prenom** »

Exécutez ensuite la fonction « **maxi(VilleAltitude)** » qui apporte la réponse à la question « Quelle est la ville dont l'altitude maximale est la plus grande ? »

Applications :

1- Répondre à ces questions en utilisant le fichier réduit des villes de la Manche : « **Echantillon_Villes_Manche_csv** »

Créer une fonction « **mini(VilleAltitude)** » qui renvoie la ville dont l'altitude maximale est la plus petite.

Modifier le programme précédent afin de faire afficher la ville dont l'altitude minimale est la plus grande.

Puis un programme qui permet d'afficher la ville dont l'altitude minimale est la plus petite.

2- Reprendre le fichier « **Eleves_1eres_NSI** »

Créer un programme qui renvoie la liste des noms tous les élèves du groupe et la liste de tous les prénoms.

Puis compléter le programme pour récupérer un dictionnaire dont les clés sont les noms des élèves et les valeurs les prénoms.

3- Reprendre le fichier « **lycees25** » des lycées Bas-Normands.

Créer un programme qui renvoie la liste de tous les lycées Bas-Normands.

En utilisant le fichier « **lycees25_Manche** » :

- Créez un programme qui renvoie le lycée de la Manche le plus au sud.
- Créez un programme qui renvoie le nombre de lycées avec internat.