1 Extraits du programme : contenus et capacités

Contenus	Capacités attendues
Données structurées	Utiliser un site de données ouvertes, pour sélectionner et récupérer des données.
Traitement de données structurées	Réaliser des opérations de recherche, filtre, tri ou calcul sur une ou plusieurs tables ; observer les différences de traitements possibles selon le logiciel choisi pour lire le fichier : tableur, programme Python.

Exemples d'activités
- Explorer les données d'un fichier CSV à l'aide d'opérations de tri et de filtre, effectuer des calculs sur ces données, réaliser une visualisation graphique des données. - À partir de deux tables de données ayant en commun un descripteur, montrer l'intérêt des deux tables pour éviter les redondances et les anomalies d'insertion et de suppression, réaliser un croisement des données permettant d'obtenir une nouvelle information.

(MINISTÈRE DE L'ÉDUCATION NATIONALE, 2019)

2 Activités avec le tableur

2.1 Obtenir des données

Les données ouvertes (Open Data en anglais) sont des informations accessibles librement et gratuitement, sous la forme de fichiers respectant des formats inter-opérables.
La finalité est de donner la possibilité à tout citoyen, toute entreprise ou association d'utiliser ces données numériques à ses propres fin d'analyse pour en extraire l'information désirée.
Une partie de ces données sont publiques, par exemple le site https://www.data.gouv.fr/fr/ contient un grand nombre de données publiques ou le site de l'insee https://www.insee.fr/fr/accueil. Ces données sont librement réutilisables.
En général, le site qui fournit le fichier des données donne aussi une description des données de la table (description du fichier avec définition des descripteurs)

2.2 Trier des données

1. Sur le site https://www.data.gouv.fr/fr/, faire une recherche pour trouver les données relatives aux **temps de parole des hommes et des femmes à la télévision et à la radio** et cliquer sur la partie relative aux **moyennes par années et par chaînes** : donner une description du contenu du fichier.
   Télécharger le fichier `20190308-years.csv` correspondant.
2. Ouvrir le fichier `20190308-years.csv` avec Libre Office en choisissant `Jeu de caractères : Unicode : UTF-8` et `Langue : Français(Suisse)`.

   Remarque : Lorsque vous ouvrez le fichier avec Libre Office, la boite de dialogue vous montre un aperçu du fichier : vous remarquerez que ici les nombres sont écrits avec le point comme séparateur décimal. Si vous choisissez `Langue : Français(France)`, les nombres ne seront pas reconnus comme des nombres mais comme des chaînes de caractère : impossible alors de faire des calculs.
3. Trier les données (*voir explications après les questions*) pour répondre aux questions suivantes :
   (a) En quelle année, le taux d'expression des femmes a-t-il été le plus important ? On précisera le nom du média et le taux.
   (b) En 2019, quel est le nom du média pour lequel le taux d'expression des femmes est le plus important ? On précisera le taux.
   (c) En quelle année, le taux d'expression des femmes a-t-il été le plus important sur un média public ? On précisera aussi le nom du média et le taux.

Comment trier des données dans Libre Office ?
Dans la barre de menu, choisir `Données` puis `Trier...`

Choisir ensuite les clefs de tri primaires et éventuellement secondaires en spécifiant si le tri est croissant ou décroissant.

2.3 Filtrer des données

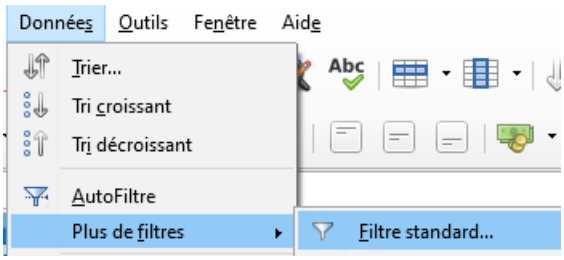
Une autre façon d'obtenir des informations est de filtrer les données pour n'afficher que celles qui nous intéressent.
Des outils de filtres permettent de ne faire afficher que certaines lignes d'une feuille de calcul suivant certains critères.

1. **Comment filtrer des données ?**
   La fonction AutoFiltre insère, au niveau d'une ou de plusieurs colonnes de données, une zone combinée permettant de sélectionner les enregistrements (lignes) à afficher.
   Pour se faire, sélectionnez (= mettre en bleu) les colonnes auxquelles vous désirez appliquer l'AutoFiltre.
   Ne pas en sélectionner revient à les sélectionner toutes (ce qui est préférable).

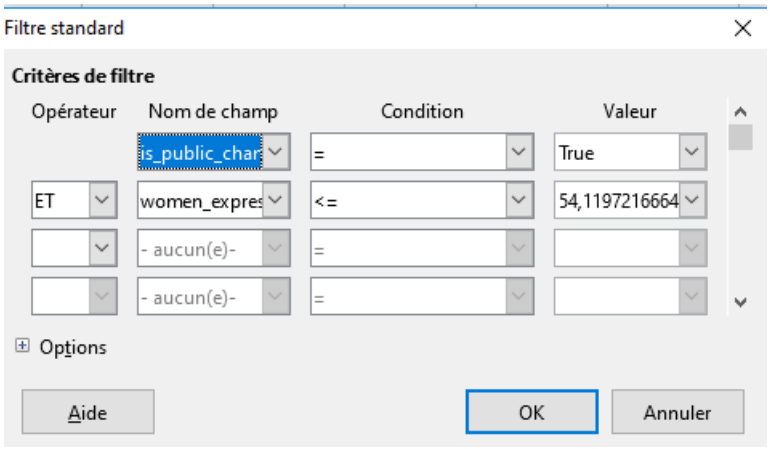
   Choisir `Données` puis `Autofiltre` ou cliquer sur l'icône : 
   Les flèches des boîtes combinées sont visibles dans la première ligne de la plage sélectionnée.
   Pour lancer le filtre, cliquer sur la flèche de déroulement située dans l'en-tête de la colonne et choisissez un élément.

   Seules les lignes dont le contenu correspond aux critères de filtre sont affichées. Les autres lignes sont filtrées. Si les numéros des lignes ne se suivent pas, cela indique que les lignes ont été filtrées. La colonne utilisée pour le filtre est identifiée par une flèche de couleur différente.
   Lorsque vous appliquez un Autofiltre supplémentaire sur une autre colonne de plage de données filtrées, alors les autres zones combinées listent seulement les données filtrées.
   Pour afficher à nouveau tous les enregistrements, sélectionnez l'entrée "tout" dans la zone combinée de l'AutoFiltre.
   Pour cesser d'utiliser l'AutoFiltre, sélectionnez toutes les cellules sélectionnées à l'étape 1 et choisissez de nouveau Données - Filtre - AutoFiltre.

2. Reprendre le fichier `20190308-years.csv` et répondre aux trois questions posées précédemment en utilisant les filtres.
3. Vous pouvez aussi choisir la fonction Filtre Standard. Choisir `Données` puis `Plus de filtres` puis `Filtre standard...`



La boîte de dialogue Filtre standard s'affiche et vous permet de définir un filtre standard selon des critères que vous pouvez choisir dans les menus déroulants :



Répondre à la question suivante en utilisant un filtre standard :
En 2018, quel est le nom du média pour lequel le taux d'expression des femmes est le plus important ?

2.4 Croisement des données

Lorsqu'on dispose de deux tables qui ont au moins un descripteur en commun, on peut obtenir une nouvelle information, on peut créer une nouvelle table en rajoutant un ou plusieurs descripteurs absents dans une des deux tables.

1. Prenons les deux tables suivantes :

Nom	Ville	Entrées
Musée des Beaux-Arts	Lyon	334 459
Musée d'Art Roger Quillot	Clermont-Ferrand	78 386
Musée Dauphinois	Grenoble	76 413
Musée Alpin	Chamonix	35 747
Musée d'art contemporain	Lyon	135 000

Ville	Région	Habitants
Grenoble	Isère	160 649
Lyon	Rhône	513 275
Chamonix	Haute Savoie	8 906
Clermont-Ferrand	Puy-de-Dôme	141 398

d'après Françoise Tort.

(a) Quel est le nombre d'habitants de la ville du musée Alpin ?
(b) Dans quelle région se trouve le musée qui a le plus petit nombre d'entrées ?
(c) Combien y a-t-il de musées dans la ville qui a le plus grand nombre d'habitants ?

Des que les tables contiennent un nombre important de données, il est impossible de faire des recherche manuellement.

2. Dans Libre Office, on aura besoin d'une fonction qui vérifie si une valeur spécifique est contenue dans la première colonne d'une `matrice`. Elle renvoie alors la valeur dans la même ligne de la colonne désignée par `index`.

La formule correspondante à écrire dans une cellule est `=RECHERCHEV(critère_de_recherche;matrice;index;trié)`

- ⋆ `critère_de_recherche` est la valeur recherchée dans la première colonne de la matrice.
- ⋆ `matrice` est la référence qui doit comprendre au moins deux colonnes.
- ⋆ `index` est le numéro de la colonne dans la matrice qui contient les valeurs devant être renvoyées. La première colonne a le numéro 1.
- ⋆ `trié` est un paramètre facultatif qui indique si la première colonne de la matrice est triée en ordre croissant. Saisissez la valeur logique FAUX ou 0 si la première colonne n'est pas triée en ordre croissant. Les colonnes triées peuvent être recherchées plus rapidement et la fonction renvoie toujours une valeur, même si la valeur de recherche ne correspond pas exactement. Dans les listes non triées, la valeur de recherche doit correspondre exactement. Sinon la fonction renvoie ce message `Erreur: valeur non disponible`.

3. Sur le site https://sql.sh/736-base-donnees-villes-francaises, télécharger la liste des villes françaises sous le format `csv` : le nom du fichier est `villes_france.csv`. Attention ce fichier ne contient pas de descripteurs à la première ligne mais les données du fichier sont décrites sur la page du site.
   Sur le site de l'INSEE, dans la partie `Définitions, méthodes et qualité, Géographie administrative et d'étude, Téléchargement, Code officiel Géographique, Millésime 2019 : Téléchargement des fichiers`, télécharger les fichiers sur les régions et les départements avec l'extension `.csv`.
   Les fichiers après décompression ont pour noms `region2019.csv` et `departement2019.csv`.
4. Ouvrir `region2019.csv` avec Libre Office puis sauvegarder ce fichier sous le nom `ActivitesRegions.csv` (attention à l'encodage).
   Ensuite, insérer les fichiers `departement2019.csv` et `villes_ france.csv` chacun dans une nouvelle feuille.
   Pour cela, choisir dans la barre de menu `Feuille` puis `Inserer une feuille` puis cocher la case `A partir d'un fichier` ....
5. Dans la feuille `departement2019`, on souhaite rajouter le nom des régions, ce qui possible puisque les deux collections départements et régions ont en commun le descripteur donnant le numéro des régions.
   (a) Dans la cellule H1 de la feuille `departement2019`, tapez le nom du descripteur à rajouter par exemple `NomRegion`.
   (b) Dans la cellule H2, taper la formule `=RECHERCHEV(B2;region2019.A$2:F$19;4;0)`
   Explication de la formule :
   Cette instruction effectue une recherche du contenu de la cellule B2 (c'est-à-dire ici 84) dans les cellules A2 à F19 de la feuille "region2019". Elle trouve cette valeur 84 à la cellule A17. Elle affecte alors à la cellule H2 le contenu de la cellule D17, c'est-à-dire AUVERGNE-RHONE-ALPES, situé dans la 4ème colonne de la zone de recherche qui est la plage A$2 :F$19 (d'où le "4" dans la formule). Le nom AUVERGNE-RHONE-ALPES devrait donc apparaître dans la cellule H2 si vous avez tapé correctement cette formule. Le dernier paramètre, 0, spécifie que la colonne A dans laquelle la recherche se fait n'est pas triée par ordre croissant. Les $ sont ici pour que par recopie vers le bas la plage de recherche reste inchangée.
   (c) Recopier la formule de la cellule H2 vers le bas jusqu'à la dernière ligne remplie.
   Pour cela, copier la cellule H2 (ctrl+c) , sélectionner la cellule H3 et appuyer en même temps sur ctrl+Maj+fin (ce qui sélectionne la plage H3 :H976) puis coller (ctrl+v)
   (d) Rajouter le nom du département dans la feuille `villes_france` en utilisant la feuille `departement2019` puis y rajouter le nom de la région.
   (e) Utiliser un filtre pour ne voir que les villes de votre département dans la feuille "villes_france".
   (f) Insérer une nouvelle feuille que vous appellerez "Calvados" ou "Manche" ou "Orne" et y copier-coller les lignes sélectionnées par le filtre.
   (g) Dans cette dernière feuille, calculer le nombre moyenne d'habitants en 2010 et 2012. Donner l'altitude minimale et l'altitude maximale .....

3 Traitement des données avec Python pour comprendre comment ça marche

3.1 Lire un fichier

Pour ouvrir un fichier, il faut commencer par créer un objet python `f` (par exemple) représentant le fichier créé et le mettant en état de lecture avec l'option `'r'` et comme c'est l'option par défaut, il suffit d'écrire : `f= open('nomFichier')` où `'nomFichier'` désigne le nom avec l'extension.
Pour le lire on utilise la méthode `read` : `f.read()`

Écrire l'exemple ci-dessous dans un script puis l'exécuter.

```
f=open('baselog.csv',encoding='UTF8')  # ou open('baselog.csv', 'r',,encoding='UTF8')
contenu=f.read()
f.close() # un fichier ouvert doit être toujours fermé
```

Dans la console, taper les instructions suivantes :

```
>>>contenu
>>>type(contenu)
>>>print(contenu)
```

On a ainsi récupérer le texte sous forme d'une chaîne de caractères, ce qui n'est pas très pratique pour traiter le fichier.

3.2 Lire un fichier CSV avec le module csv

Pour traiter des données numériques qui sont enregistrées dans un fichier `.csv`, il est fortement conseillé d'avoir enregistré le fichier au format csv avec l'encodage UTF-8 et un point pour le séparateur décimal si on souhaite faire des calculs avec les données.

Le module `csv` est un des modules qui permet de lire un fichier CSV et de le "transformer" en liste que l'on peut ensuite manipuler.

Écrire l'exemple ci-dessous dans un script puis l'exécuter.

```
import csv
f=open('baselog.csv',encoding='UTF8')
r=csv.reader(f)
Log=list(r)
f.close()
```

Dans la console, taper les instructions suivantes :

```
>>>print(Log) # ou Log selon l'environnement utilisé
>>>type(Log)
```

On récupère ainsi une liste de listes : la liste de toutes les lignes de la table, chaque ligne étant une liste et chaque élément de cette dernière est de type chaîne.
Plus précisément, un élément de `Log` obtenu par l'instruction `Log[i]` (où `i` est l'indice de l'élément) correspond à une ligne de la table.
L'instruction `len(Log)` qui renvoie la longueur de la liste `Log` correspond donc au nombre de lignes de la table. L'instruction `len(Log[0])` qui renvoie la longueur du premier élément de la liste `Log`, correspond donc au nombre de colonnes de la table. Enfin, une cellule de la table sera accessible par `Log[i][j]` où `i` est l'indice de la ligne et `j` celui de la colonne.

Dans la console, taper les instructions suivantes :

```
>>>nblignes=len(Log)
>>>nbcolonnes=len(Log[0])
>>>Ligne1=Log[0] ; print(Ligne1)
>>>Ligne3=Log[2] ; print(Ligne3)
```

A quoi correspond les éléments de la liste `Ligne1` ?

Pour récupérer, par exemple, la première colonne autrement dit la liste des éléments d'indice 0 de chaque ligne qui dans notre exemple correspond à l'identifiant `'nom'`, on peut utiliser la fonction suivante :

```
def Colonne(L,j):
    #renvoie la liste des éléments de la colonne d'indice j de la table L
    C=[]
    nblignes=len(L)
    for i in range(nblignes):
        C.append(L[i][j])
    return(C)
```

Expliquer "en langage naturel" comment opère cette fonction.
L'exécuter et dans la console taper l'instruction qui permet d'obtenir la liste des éléments de la première colonne de la table (Log).

3.3 Croisement des données

1. A partir du fichier `'basemail.csv'`, obtenir la liste `Mail` correspondant à cette table.
2. Vérifier que ces deux tables `Log` et `Mail` ont un descripteur en commun en affichant dans la console les deux listes contenant les descripteurs de celles-ci.
3. On souhaite écrire une fonction `loginPass(adressemail)` qui prend en paramètre la chaine `adressemail` et renvoie le login et le mot de passe correspondant. Compléter :
   On rappelle que la méthode `index` pour les listes permet d'obtenir l'indice d'un élément d'une liste.

   ```
   def loginPass(adressemail):
       CmailMail= Colonne(Mail,0) #on récupère la liste des éléments de la colonne
       #qui contient les mails dans la liste Mail
       CnomMail= ..................#on récupère la liste des éléments de la colonne
       #qui contient les noms dans la liste Mail
       indiceLmail=CmailMail.index(adressemail) #on obtient l'indice de la ligne
       #où se trouve adressemail dans CmailMail
       lenom=CnomMail[indiceLmail] #on obtient le nom correspondant à  adressemail
       lemotdepasse=Mail[indiceLmail][2] #.....................................................
       CnomLog=.....................#on récupère la liste des éléments de la colonne
       #qui contient les noms dans la liste Log
       indiceLlog=..................... #on obtient l'indice de la ligne
       #où se trouve lenom dans CnomLog
       lelogin=.................. #on obtient le mot de passe correspondant à lenom
       return(lelogin,lemotdepasse)
   ```

4. Quel est le login et le mot de passe du détenteur de l'adresse `'mp@chez.moi'` ?

4 Représenter des données

Exemples d'activités
- Réaliser une visualisation graphique des données.

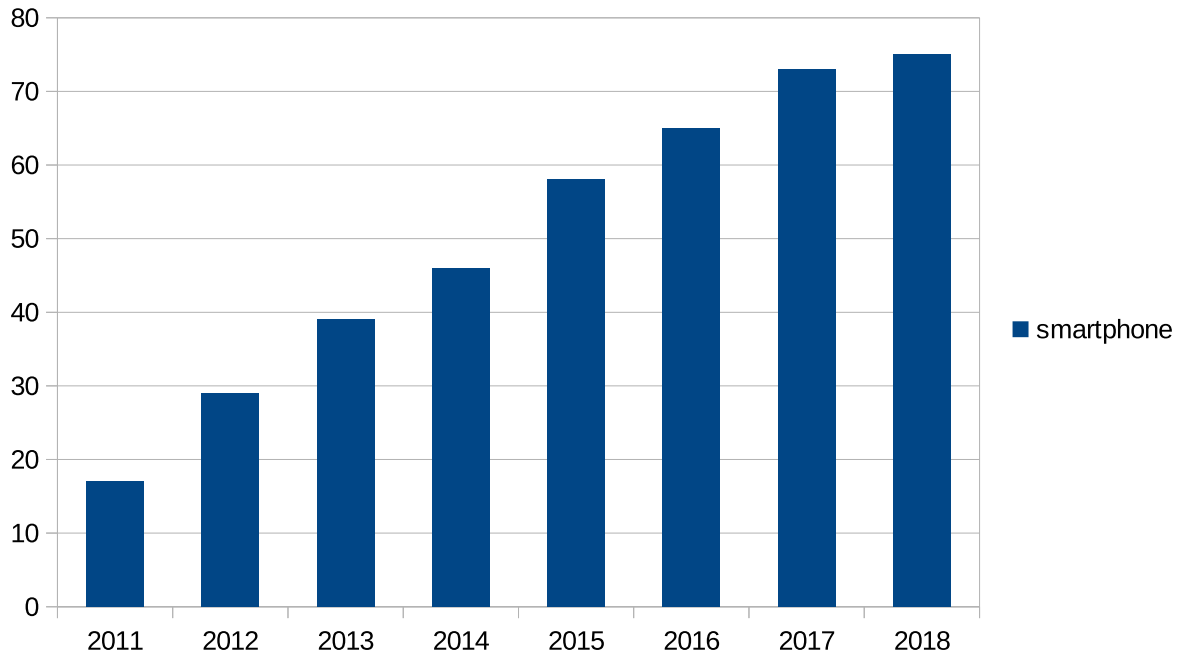
On dispose d'un fichier donnant la répartition en % de la population en France entre 2011 et 2018, par type de téléphone mobile utilisé.
Source : CREDOC, Enquêtes sur les "Conditions de vie et les Aspirations"

4.1 Avec Libre Office

Les données ont été stockées dans le fichier `typemobile.csv`.

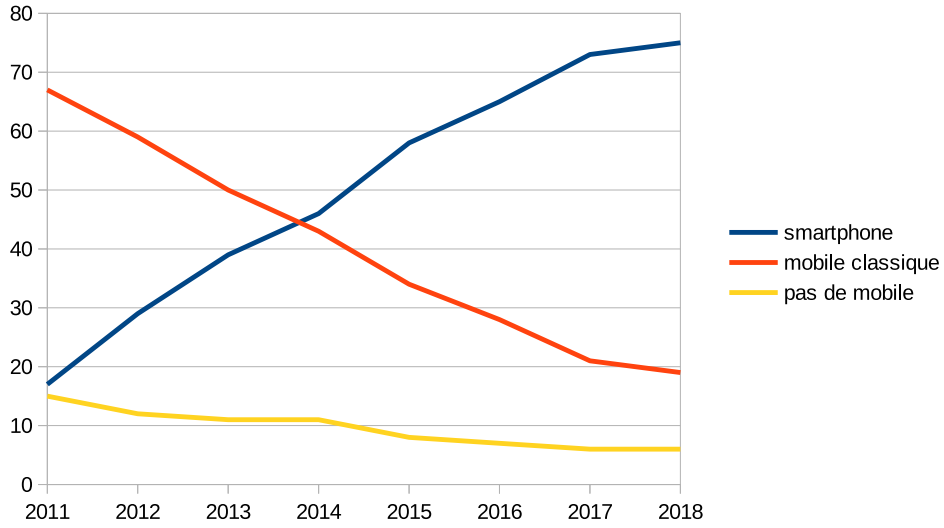
1. On souhaite visualiser l'évolution de l'équipement en smarthphone.
   (a) Ouvrir ce fichier (en UTF8)
   (b) Sélectionner les lignes 1 et 2
   (c) Cliquer sur l'icône (ou dans le menu `Insertion/Diagramme`)
   Choisir :
   - Type de diagramme : `Colonnes`
   - Plage de données (à sélectionner si non sélectionnées avant) : choisir `Séries de données en lignes` et cocher les cases `première ligne comme étiquette` et `première colonne comme étiquette`
   - Cliquer sur `Terminer`

   On obtient :



2. On souhaite visualiser l'évolution de l'équipement en téléphonie mobile
   (a) Sélectionner les 4 lignes
   (b) Cliquer sur l'icône (ou dans le menu `Insertion/Diagramme`)
   Choisir :
   - Type de diagramme : `ligne` puis `lignes seules`
   - Plage de données (à sélectionner si non sélectionnées avant) : choisir `Séries de données en lignes` et cocher les cases `première ligne comme étiquette` et `première colonne comme étiquette`
   - Cliquer sur `Terminer`

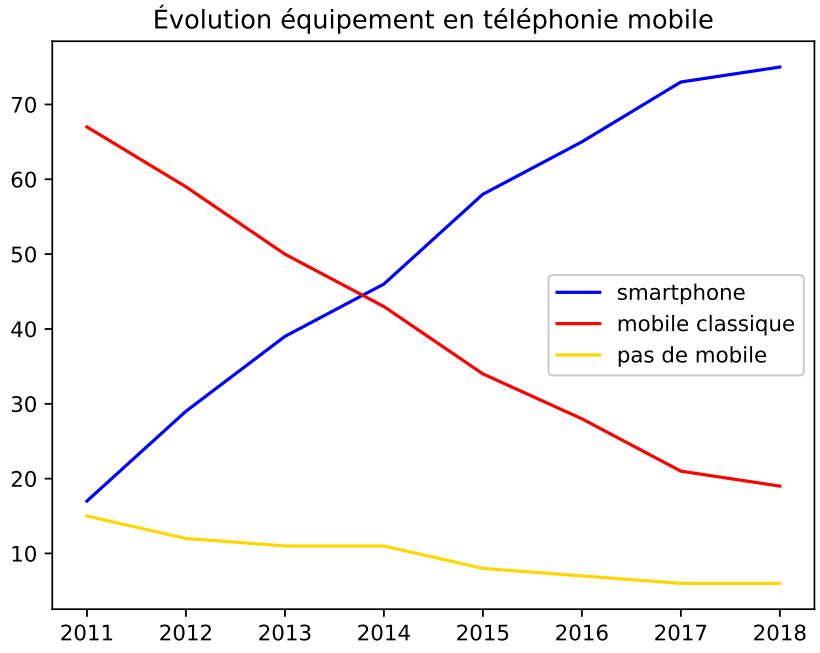
   On obtient :



4.2 Avec Python pour comprendre comment opère le tableur

Les données ont été stockées dans trois listes : `Smart`, `Classique` et `Aucun`.

```
from matplotlib.pyplot import *
Annee=[2011,2012,2013,2014,2015,2016,2017,2018]
Smart=[17,29,39,46,58,65,73,75]
Classique=[67,59,50,43,34,28,21,19]
Aucun=[15,12,11,11,8,7,6,6]
title('Évolution équipement en téléphonie mobile')
plot(Annee,Smart,'b' ,label='smartphone' )
plot(Annee,Classique,'r',label='mobile classique')
plot(Annee,Aucun,'gold',label='pas de mobile')
legend()
show()
```



5 PIX : évaluation nationale

- 1.1 Mener une une recherche d'information
- 1.2 Gérer des données
- 1.3 Traiter des données
- 2.2 Partager et publier
- 3.4 Programmer
- 4.2 Protéger les données personnelles et la vie privée
- 4.3 Protéger l'environnement
- 5.3 Construire un environnement numérique