Web scraping

Récupération des données météorologiques
site : infoclimat.fr

Installation de Python :

- Suivre les étapes : https://www.python.org/downloads/

Installation de la bibliothèque Pandas :

- Se rendre dans le terminal et exécuter cette commande :

```
> pip install pandas
```

Installation de la bibliothèque Beautiful Soup :

- Se rendre dans le terminal et exécuter cette commande :

```
> pip install beautifulsoup4
```

Installation de la bibliothèque Requests :

- Se rendre dans le terminal et exécuter cette commande :

```
> pip install requests
```

Comment exécuter le fichier Python ?

- Se rendre dans le terminal à l'endroit où se trouve le fichier `data_meteo.py`
- Exécuter cette commande :

```
> python data_meteo.py
```

Comment modifier les dates de récupération des données ?

- Ouvrir le fichier `data_meteo.py` et se rendre dans le main : `if __name__ == "__main__"`
- Dans la fonction `get_data(2018, 2018, 1, 2)`, saisir les dates souhaitées

Description des fonctions :

- `def get_url(mois, année):`

Cette fonction récupère l'adresse URL des données météorologiques à la date donnée en paramètre.

- `def sup_le(string):`

Lors de la récupération de la température maximale extrême du mois, la température est suivie du jour où elle a été mesurée (exemple : 19.4 le 23). Cette fonction permettra donc de récupérer uniquement la température.

- `def remplir_dict(temp, station, dic, département):`

  Cette fonction associe la température au numéro du département en l'ajoutant dans le dictionnaire. Elle vérifie préalablement si le numéro de département récupéré est bien dans la liste des départements de la France métropolitaine.

- `def cal_moy(dic):`

  Cette fonction calcule la moyenne des températures des stations appartenant aux mêmes départements.

- `def creer_dft(département):`

  Cette fonction crée la transposition d'un data frame.

- `def get_data(deb année, fin annee, deb mois, fin mois):`

  Cette fonction récupère les données météorologiques (les températures moyennes, maximales extrêmes et les moyennes des températures maximales du mois) ainsi que les départements et exporte toutes ces données dans des fichiers csv.

Sorties :

- `data_tmm.csv` : feuille de calcul Excel contenant les **températures moyennes du mois** pour chaque département

- `data_txm.csv` : feuille de calcul Excel contenant les **moyennes des températures maximales du mois** pour chaque département

- `data_txx.csv` : feuille de calcul Excel contenant les **températures maximales extrêmes du mois** pour chaque département

Adaptation de la base de données des sirops

Fichier Excel fourni par le client

Installation de Python :

- Suivre les étapes : https://www.python.org/downloads/

Installation de la bibliothèque Pandas :

- Se rendre dans le terminal et exécuter cette commande :

```
>    pip install pandas
```

Installation de la bibliothèque XlsxWriter :

- Se rendre dans le terminal et exécuter cette commande :

```
>    pip install XlsxWriter
```

Installation de la bibliothèque Openpyxl :

- Se rendre dans le terminal et exécuter cette commande :
  > `pip install openpyxl`

Comment exécuter les fichiers Python ?

- Se rendre dans le terminal à l'endroit où se trouve les fichiers `tri_bdd.py`, `separation_bdd.py` ainsi que la base de données
- Exécuter consécutivement les commandes :
  > `python tri_bdd.py`
  > `python separation_bdd.py`

Sorties :

- `Base_donnees_triee.xlsx` : feuille de calcul Excel contenant la nouvelle base de données triée en extension xlsx

- `data_sirops.xlsx` et `data_sirops.csv` : feuille de calcul contenant les données des sirops associées à leur identifiant en extensions xlsx et en csv

- `identifiants_sirops.xlsx` et `identifiants_sirops.csv` : feuille de calcul contenant les informations des sirops associés à leur identifiant en extensions xlsx et en csv

Analyse statistique

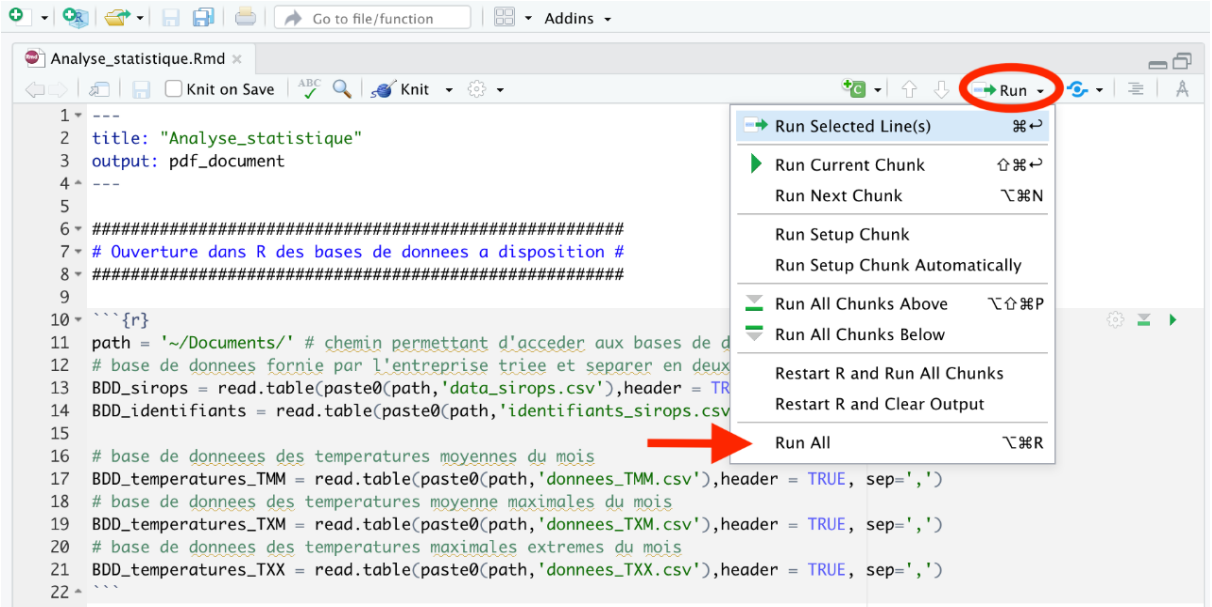
Etude de la corrélation entre les ventes de sirops et la météo et estimation de l'impact d'un été chaud sur ces ventes

Installation de Rstudio et des bibliothèques utilisées:

- Se rendre sur : https://www.rstudio.com/products/rstudio/download/#download
- Si la bibliothèque "RColorBrewer" n'est pas installée sur votre PC, tapez la commande `install.packages("RColorBrewer")` dans la console de Rstudio

Comment exécuter le fichier R ?

- Mettre toutes les bases de données fournies dans le même répertoire
- Ouvrir le fichier `Analyse_statistique.Rmd` avec Rstudio
- Dans la variable `path`, au tout début du fichier R, mettre le chemin d'accès à ces bases de données
- Cliquer sur la petite flèche à côté du bouton Run puis sur "Run All" tout en bas.



Description des fonctions :

- `moyennes_temperatures_France(BDD)` :

Cette fonction récupère une base de données des températures par département et par mois et renvoie une matrice colonne des moyennes des températures en France par mois.

- `dates(BDD)` :

Cette fonction récupère une base de données des ventes de sirops par sirop et par mois et renvoie un tableau contenant les dates de début et de fin de vente de chacun des sirops

- `correlation_tri(sirops, temperatures, choix, identifiants)` :

Cette fonction récupère une base de données des ventes de sirops par sirop et par mois, une base de données des températures par département et par mois, une chaîne de caractère précisant si l'on travaille sur quel type de sirop on travaille et une base de données contenant les descriptifs des sirops associés à leur identifiant. Elle permet de calculer, pour chaque type de sirop, sa corrélation avec la température puis renvoie un tableau des corrélations trié par ordre croissant.

- `correlation_graphes(sirops, temperature, choix, identifiants)` :

Cette fonction récupère une base de données des ventes de sirops par sirop et par mois, une base de données des températures par département et par mois, une chaîne de caractère précisant sur quel type de sirop on travaille et une base de données contenant les descriptifs des sirops associés à leur identifiant. Elle permet de calculer, pour chaque type de sirop, sa corrélation avec la température puis de tracer 2 diagrammes en barres. Le premier représentant les types de sirops les plus corrélés et le second représentant ceux qui le sont le moins.

- `correlation_par_type(sirops, BDD_identifiants, temperatures, type):`

Cette fonction récupère une base de données des ventes de sirops par sirop et par mois, une base de données contenant les descriptifs des sirops associés à leur identifiant, une base de données des températures par département et par mois et une chaîne de caractère précisant sur quel type de sirop on travaille. Elle permet de calculer, pour chaque type de sirop, sa corrélation avec la température puis renvoie un tableau des corrélations trié par ordre croissant et trace 2 diagrammes en barres. Le premier représentant les types de sirops les plus corrélés et le second représentant ceux qui le sont le moins.

- `regression_lineaire_simple(sirop):`

Cette fonction récupère un entier correspondant au numéro du sirop à évaluer. Elle permet de calculer et de tracer :

- une régression linéaire des ventes de ce sirop en fonction de la température
- un intervalle de confiance
- un intervalle de prédiction

Sorties :

- `Correlations_sirops.csv` : un fichier Excel contenant un tableau trié par ordre croissant contenant les **corrélations de chacun des sirops**.

- `Correlations_parfums.csv` : un fichier Excel contenant un tableau trié par ordre croissant contenant les **corrélations de chacun des parfums de sirop**.

- `Correlations_gammes.csv` : un fichier Excel contenant un tableau trié par ordre croissant contenant les **corrélations de chacune des gammes de sirop**.

- `Correlations_marques.csv` : un fichier Excel contenant un tableau trié par ordre croissant contenant les **corrélations de chacune des marques de sirop**.