Projet de Gestion des données : « Passe ton bac d'abord ! » (Sujet3)

Lucille CARADEC et Tom-Hadrian SY

Introduction

Dans le cadre de ce projet, nous nous sommes demandé si la météo pouvait avoir un effet sur les résultats au baccalauréat. Pour des raisons de simplicité, nous nous sommes limités à l'étude du facteur température sur le taux de réussite au bac.

Cette problématique générale peut être déclinée sous deux angles :

(1) Sur une même année, existe-t-il un lien significatif entre température et résultats du bac pour les établissements français de différentes régions ?
(2) Au sein d'une même région, des différences de températures interannuelles pourraient-elles avoir un effet sur les résultats du bac ?

Démarche de travail

Collecte des données

Pour répondre aux questions posées, les 3 jeux de données ont été téléchargé sous forme de fichiers csv, depuis la plateforme data.gouv.fr. Nous avons donc obtenu :

- un grand jeu de données « Bac » portant sur les performances au bac de tous les établissements français, sur plusieurs années

- un jeu de données « Dpt » reliant les départements aux régions françaises

- un jeu de données « Climat » décrivant les caractéristiques météorologiques des régions sur une période donnée

Construction d'une base de données

La base de données est construite à partir des fichiers csv sur Python (package pymysql), dans le module MySQLManager.py. Ce module contient les classes et méthodes nécessaires à l'ouverture de la base de données avec pymysql, la construction, suppression et le remplissage des tables.

Trois tables, correspondant aux trois jeux de données (voir ci-dessus) sont donc créées et remplies (main.py, partie 1). La base est décrite par le schéma relationnel par la figure 1.

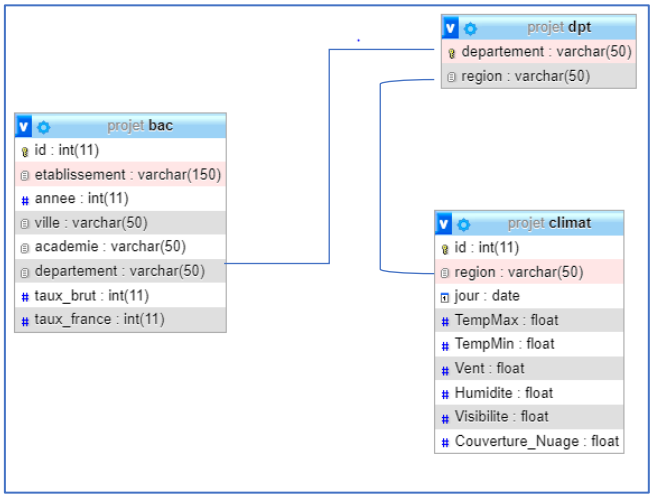

Figure 1- Schéma relationnel de la base de données construite

Traitements de la base et analyses des résultats

Question 1 : Sur une même année, existe-t-il un lien significatif entre température et résultats du bac pour les établissements français de différentes régions ?

Pour la première question, la requête suivante a été émise : (main.py, partie 2)

```
SELECT b.taux_brut, AVG(c.TempMax) TempMax, AVG(c.TempMin) TempMin FROM Bac b
    INNER JOIN Dpt d ON b.departement = d.departement
INNER JOIN Climat c ON c.region = d.region
WHERE b.annee = 2019 AND YEAR(c.jour) = 2019
AND MONTH(c.jour) = mois GROUP BY b.id
```

Il a s'agit de réaliser deux jointures internes entre les trois tables. Pour chaque région, nous avons sélectionné le taux de réussite au bac ainsi que les températures minimales et maximales. Cette opération a été répétée pour différents mois (variable « mois », soulignée dans la requête), à but de comparaison.

Analyse et résultat : Une régression linéaire simple, du taux de réussite au bac en fonction de la température moyenne de la région a été réalisée. Nous obtenons des coefficients de corrélation R^2 très faibles, pour chaque modèle créé, ce qui révèle l'absence de lien entre les deux variables.

Question 2 : Au sein d'une même région, des différences de températures interannuelles pourraient-elles avoir un effet sur les résultats du bac ?

```
SELECT AVG((TempMax + TempMin)/2) TempMoy, YEAR(jour)
    FROM Climat
WHERE MONTH(jour) IN (5,6) AND region = "HAUTS-DE-FRANCE"
GROUP BY YEAR(jour)
```

La région de référence choisie a été Hauts-de-France (Lille). Cette première requête calcule, pour 2018 et 2019, la température moyenne des mois de mai et juin, dans la région concernée.

```
SELECT b.taux_brut taux_2018, bb.taux_brut taux_2019
    FROM Bac b
    INNER JOIN Bac bb
    ON b.etablissement = bb.etablissement
    WHERE b.ville = 'LILLE' AND b.annee = 2018 AND bb.annee = 2019
```

La deuxième requête sélectionne, dans la table bac, les taux de réussite de la région de Lille, aux années 2018 et 2019.

Analyse et résultat : Le résultat de la première requête montre que la température moyenne dans la région des Hauts-De-France a diminué entre 2018 et 2019 (mois de mai et juin). En comparaison, le résultat de la deuxième requête montre que, parmi les établissements de la région, de 2018 à 2019, 34 établissements ont amélioré leur taux de réussite, 22 l'ont vu diminuer, et 6 ont obtenu le même taux que l'an passé. Cette observation aurait tendance à nous faire dire qu'une baisse de la température engendrerait une meilleure réussite, mais cette affirmation est très incertaine et l'étude nécessiterait des tests supplémentaires.