

Présentation projet NF26

Loïc Adam, Xiang Li
TD2 Groupe 19-20

Introduction

- Etude d'un jeu de données volumineux.
- Conception de bases de données.
- Fonctions python pour répondre à des questions.

Jeu de données

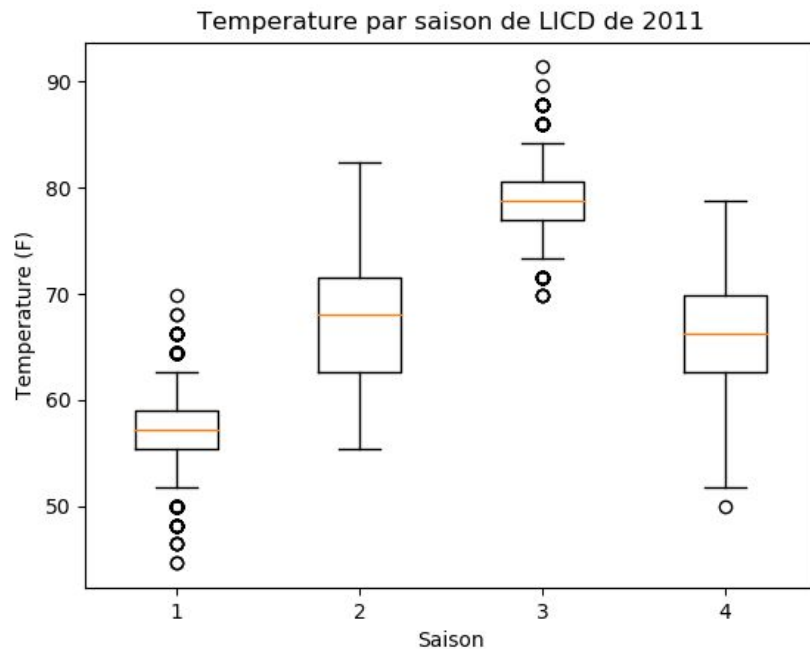
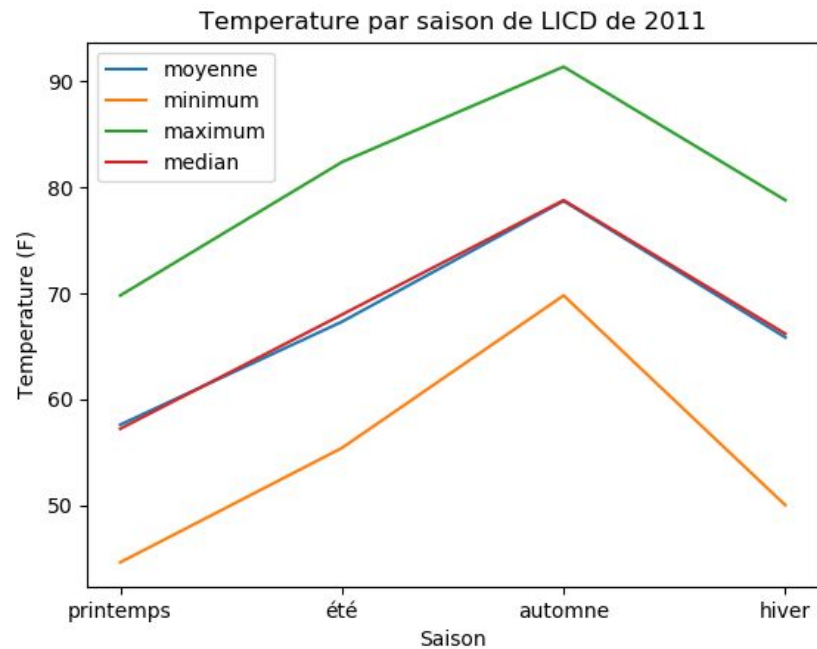
- Données ASOS (météorologie) en Italie de 2005 à 2014. Pas de données avant 2011.
- 3 175 698 enregistrements. 31 variables. Une grande partie des variables souvent vides (null ou trois espaces).
- 104 stations + 1 en Irak !

Bases de données

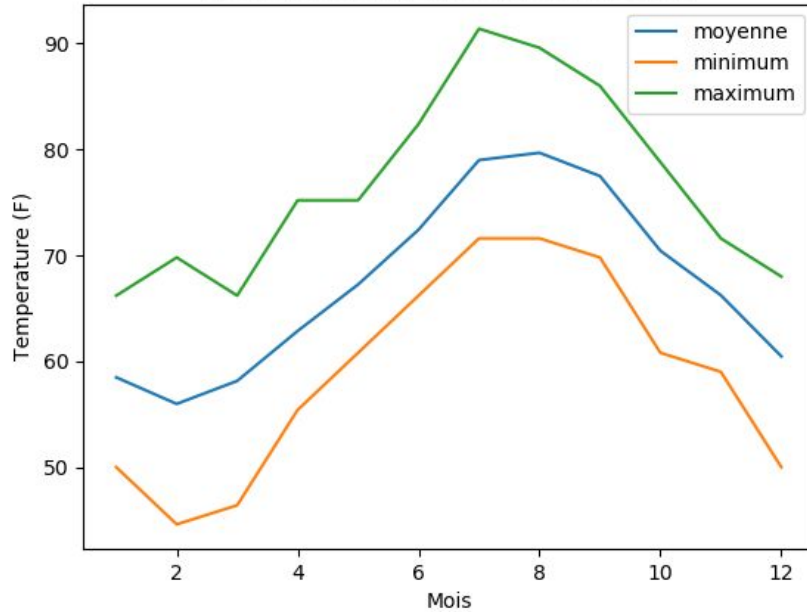
- Deux bases de données Cassandra pour répondre à nos questions.
- AsosItalyStation pour la question 1.
 - Partition : station, année.
 - Tri : mois, jour, heure, minute.
- AsosItalyTime pour la question 2 et 3.
 - Partition : année, mois.
 - Tri : date d'observation, longitude, latitude.

Question 1

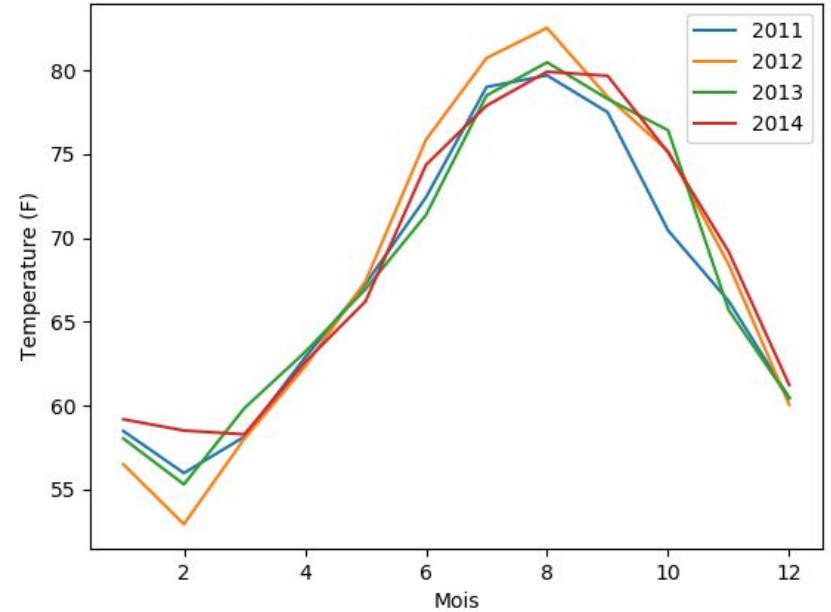
- Plusieurs types de schéma pour présenter l'historique :
 - En fonction du mois et de la saison pour une année donnée.
 - En fonction du mois pour plusieurs années.
- Plusieurs critères :
 - Moyenne, médiane, minimum, maximum.
- Trouver la plus proche station (distance euclidienne).



Temperature par mois de LICD de 2011



Historique de temperature par mois de LICD



Question 2

- Récupérer toutes les données d'un instant t en fonction d'une variable.
- Sélection des données : aucune valeur null acceptée.
- Création de la carte.



Question 3

- Sélection de données.
- Choix de variables :
 - Température de l'air.
 - Température de rosée.
 - Humidité relative.
 - Pression.
 - Vitesse du vent.
- Prétraitement : calculer minimum, maximum, moyenne, variance.



Conclusion

- Jeu de données volumineux, avec beaucoup de variables et de manque.
- Calcul en parallèle et gestion de la mémoire (générateurs).
- Limites sur chaque réponse :
 - Interface trop minimaliste.
 - Exploitation trop limitée des données (pas assez de carte, k-moyennes).
 - Correction des données (null et Irak).

Merci de votre écoute !
Nous écoutons vos
questions !