

Exercice de crunching de data

Il y a 4 fichiers CSV dans le dossier data,

- ***ticket_data.csv*** : Contenant un historique de ticket (une ligne => une proposition de ticket sur tictactrip)

- ***cities.csv*** les villes desservies par tictactrip (lien grâce aux colonnes o_city (origin_city), d_city (destination_city) de ticket_data)

- ***stations.csv*** les stations desservies par tictactrip (lien via o_station, d_station de ticket_data)

- ***providers.csv*** infos sur les différents providers (lien via company de ticket_data)

Un provider est une "sous-compagnie". Par exemple TGV et TER sont deux providers de VSC (voyages-sncf).

La mission :

- extraire les infos intéressantes type prix min, moyen et max, durée min/max/moyenne par trajet
 - différence de prix moyen et durée selon le train, le bus et le covoit selon la distance du trajet (0-200km, 201-800km, 800-2000km, 2000+km)
 - le plus d'infos bonus
-

Le résultat :

- Tes scripts
- et (pas nécessairement) un mini-rapport expliquant tes recherches et ta démarche.

Cela peut se présenter sous la forme d'un **jupyter notebook** publié sur Github / Gitlab ou une autre plateforme en ligne.

Langage requis : **Python**, fortement conseillé de l'accompagner du package **Pandas**

NB : Attention à la qualité et à la rigueur du code.

Comme par exemple :

Graphes, prédictions de prix, rapport des soucis relevés dans les données, visualisation interactive, sourcing & utilisation de données externes pertinentes
