Rapport sur l'exercice de crunching :

Introduction

Les données analysées contiennent les informations relatives à des recherches de voyages sur TicTaTrip. Nous disposons d'informations sur les voyages (entre Septembre 2017 et Mars 2018), sur les différentes localités qui constituent les points d'arrivée et d'origine de ces voyages, ainsi que sur les différents modes de transports utilisés.

Le script Jupyter ci-joint a pour vocation de donner une meilleure compréhension des données. Comme suggéré dans la consigne, je me suis concentré sur les différences de temps de trajet + prix en fonction du mode de transport et de la distance parcourue. J'ai également essayé d'explorer un peu le comportement des utilisateurs en regardant les dates privilégiées pour de voyage et d'achat de ticket. Nous distinguons trois types de transports : les bus, les trains et le co-voiturage ainsi que trois types de trajets : les courts trajets (<=200Km), les trajets moyens (200<trajet<=800Km) et les trajets longs (800<trajet<=2000Km). Le script contient les informations suivantes :

- Recherche de données manquantes dans les tables et du type de donnée de chaque attribut
- Prix minimum/maximum/moyen d'un ticket
- Temps de trajet minimum/maximum/moyen
- Date du premier et du dernier voyage
- Distance de trajet minimale/maximale/moyenne
- Graphique des prix moyens en fonction de la distance entre le point de départ et le point d'arrivée
- Graphique des temps de trajets moyens en fonction de la distance entre le point de départ et le point d'arrivée
- Graphique des prix moyens en fonction du type de transport
- Graphique des temps de trajets moyens en fonction du type de transport
- Providers les plus rentables (pour l'utilisateur) en fonction du type de transport (train ou bus) et de la distance entre le point de départ et le point d'arrivée
- Répartition des trajets en fonction du jour de la semaine
- Répartition des commandes de trajets en fonction du jour de la semaine

Problèmes rencontrés

1) Des **données manquantes** ont empêché un traitement uniforme des données. Tous les trajets faits en co-voiturage n'ont pas d'information sur les stations, ce qui est logique mais néanmoins ennuyeux pour le calcul de la distance du trajet. Plus anecdotiquement, la valeur du nom unique pour Hambourg était manquante et 3 providers n'avaient pas d'infos sur la présence ou non de WiFi, de sièges ajustables, de prises et d'emplacements pour vélo.

   Solution mise en œuvre : Les données de la ville d'Hambourg ont été rentrées manuellement vu qu'on pouvait aisément déduire la donnée manquante à partir de l'information à notre disposition. Je n'ai pas exploité les données sur les accommodations des moyens de transports. En ce qui concerne l'absence d'information sur les stations, je me suis servi des coordonnées GPS de la ville de départ et d'arrivée pour approximer la distance de trajet.

2) Le **format des timestamps** n'est pas régulier. En effet, la timezone était sous la forme +HH au lieu de +HHMM. De plus, la date de recherche de trajet était exprimée au millième de seconde près (au lieu d'être à la seconde près).

   Solution mise en œuvre : Après avoir vérifié que toutes les données de la base étaient au même fuseau horaire, j'ai choisi de supposer que la timezone était de +0000 pour tous les tickets, et j'ai décidé d'ignorer la partie décimale des secondes pour la date de recherche (j'aurai également pu la convertir en microsecondes).

3) Le **calcul de la distance parcourue** est très approximatif, en l'absence de données sur les lignes ferroviaires et les autoroutes.

   Solution mise en œuvre : Je me suis limité à un calcul de la distance géographique entre le point de départ et le point d'arrivée. Lorsque je ne connaissais pas ces derniers (dans le cas des voitures), je les ai approximés par les coordonnées géographiques de la ville de départ et de la ville d'arrivée. Si je suis assez confiant dans le fait que mon approximation est suffisamment bonne pour correctement classer les trajets dans les 3 catégories décrites précédemment, le fait de ne pas avoir la distance précise du trajet m'a empêché d'exploiter des données telles que la vitesse moyenne des trajets par exemple.