Rapport Data TicTacTrip

I - Démarches :

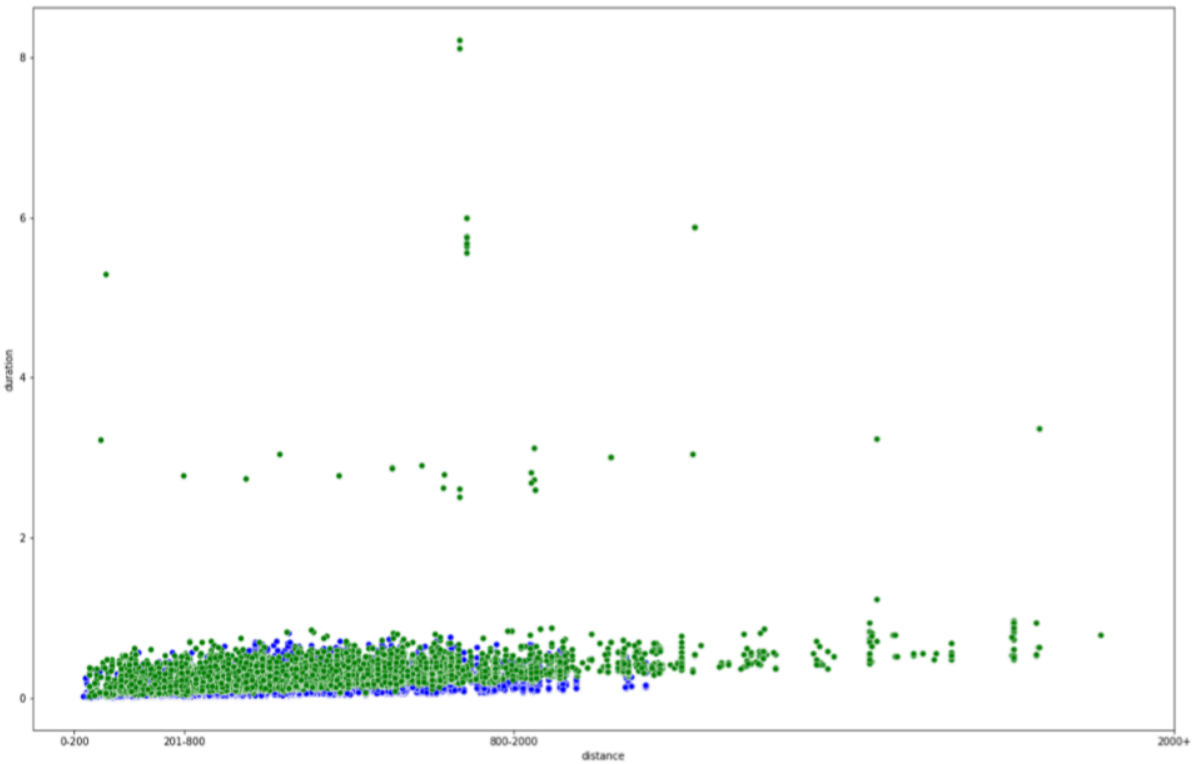
J'ai commencé tout d'abord par introduire les fichiers .CSV dans des DataFrames pour voir quelle type de Data je vais étudier.
Dès le début, j'ai remarqué des valeurs "Null" et ca peut poser des problèmes en terme de calcul. donc pour quelques parties du code, j'ai enlevé les lignes qui contiennent des valeurs "Null"

Pour le prix moyen, min, et max, les fonctions built-in de Python et Numpy ont pu faire le job. J'ai suivi deux manières pour faire cela.
La première j'ai prix tout les tickets du fichiers "ticket_data.csv" et voir le max, le min et le moyen. du prix. La deuxième j'ai fait un GroupBy par les trajets (o_station, d_station) pour voir les mêmes statistiques pour les trajets uniquement.

Pour la durée, fallait faire la différence entre la data sous les colonnes (arrival_ts,departure_ts), j'ai passé une fonction de parsing pour faciliter la différence des dates et puis avoir le résultat.

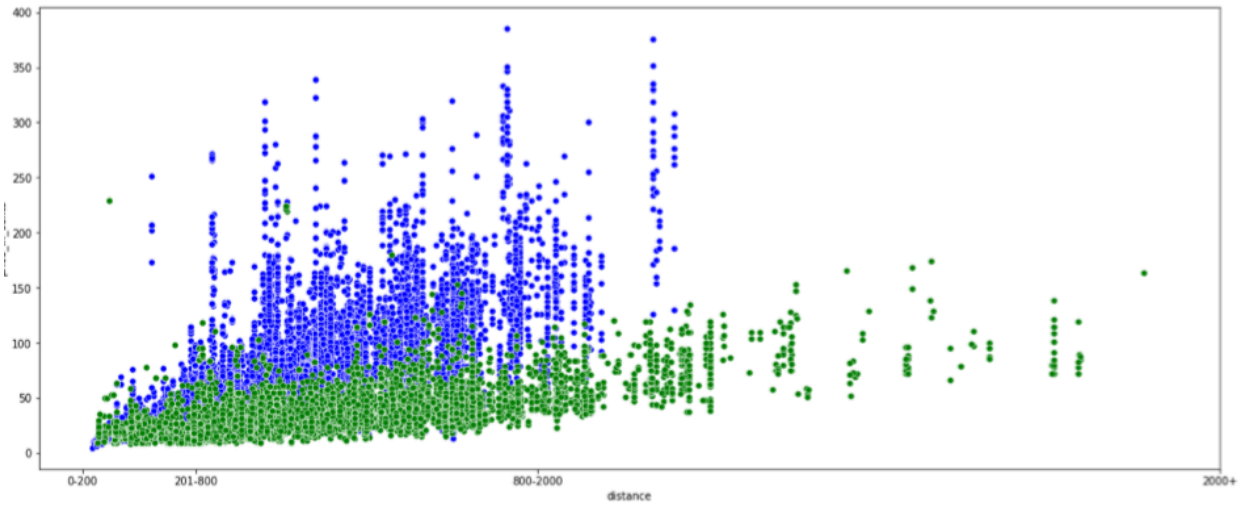
Pour la deuxième question, je voulais faire la visualisation pour afficher le résultat, donc fallait faire un lien entre le fichier "tickets_data.csv" et "providers.csv", le dernier c'est où on trouve le "transport_type".
Vous pouvez voir toute la démarche sur le fichier du code.

- **Constatation de Visualisation des trajets par rapport au nombre d'heure et la distance** :



on peut voir que pour le train, les voyages ne dépassent pas les 2 heures quelle que soit la distance, et aussi que les voyages en train ne vont pas plus que les ~2000km.
aussi, si on compare la longueur des tableaux **“df_train”** et **“df_bus”,** on peut dire que les gens aime voyager en train qu'en bus.

- **Constatation de Visualisation des trajets par rapport au prix et la distance :**



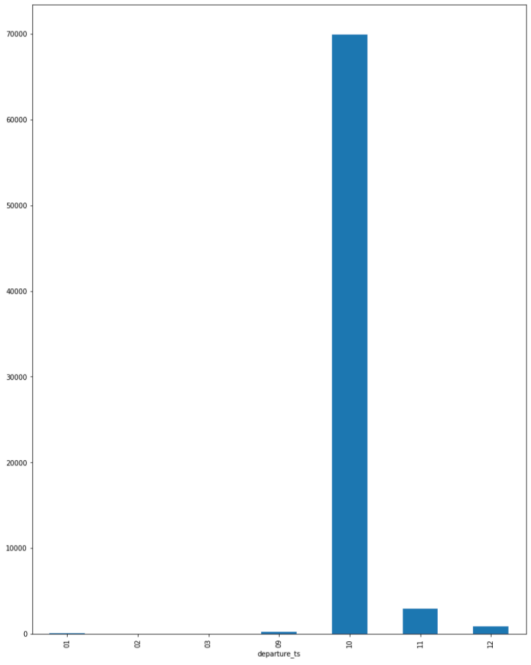
La première remarque c'est que les prix en train sont bien plus élevés qu'en bus.
et pour le train les prix s'augmentent exponentiellement par rapport au distance, alors que sur le bus les prix sont un peu stable

Tâches ajoutées :

La dataset est tellement riches de données qu’on peut l’utiliser pour beaucoup de chose, j’ai ajouté deux tâches; une visualisation des tickets achetés par mois, et un model SVM pour la prédiction des prix.
J’aimerais bien ajouté d’autres tâches mais vu le temps et la période de préparation d’examens après les vacances et aussi les projets scolaires, j’ai pas tout le temps que je veux.

- Constatation de Visualisation pour voir les mois où les gens achètent plus de tickets ou bien où les gens voyagent.

on constate que plus que 90% des tickets ont été achetés le mois d'octobre.



Model SVM pour la prédiction des prix :

J'ai prix en considération 6 features que j'ai vu qu'elles sont pertinentes, normalement on peut faire une "cross-validation" ou du "feature-selection" pour prendre les features les plus fits (ANOVA-f Value,Test Khi-deux…).

le model a un score de 0.81, qui n'est pas mal pour un model qui prend seulement 6 features en input.

II - Soucis dans la Data :

Les dataframes contiennent des valeurs "NULL" spécialement en relation avec le type de transport "carpooling" et "car", et par conséquent, les études amenées n'ont pas été appliqués sur ces deux types de transport. On peut les remplacer par des valeurs "0" ou aléatoire mais ce n'est pas efficace.

les types de transport "carpooling" et "car" n'ont pas une station d'origine et station destination (valeur null dans le tableau tickets) et donc la distance ne peut pas être prise en compte pour ces deux types de transports.