Démarches de traitement des données:

Introduction :

Ici notre but sera de trouver un moyen de décrire et présenter nos données sur trois axes majeurs que sont :

- Les différentes valeurs de temps que peut prendre un trajet
- La relation entre la distance et le prix d'un ticket, en comptant l'influence du type de transport. Et la prédiction du prix de ceci.
- Les différents coût des ticket de différents pays, rapportés à leurs distance

Procédure:

On travaillera tout au long de ce livrable avec la bibliothèque **Pandas** de python. De plus, nous utiliserons aussi **sklearn** pour élaborer nos modèles plus loin.

Après avoir fait le tour des différents DataFrames proposés par l'exercice, et ce en affichant les premières lignes de ceci. On peut commencer à avoir une idée sur leur contenu et la manière dont leurs colonnes sont agencées.

Ticket frame et durée d'un trajet:

Si on se concentre maintenant sur la table "ticket_frame", on peut ainsi extraire des informations générales sur elle grâce à la méthode "describe" qui nous renvoie par exemple la moyenne, l'écart-type, valeur min-max …etc

Plus loin, on peut aussi travailler à extraire les valeurs relatives à la durée d'un trajet. Pour cela on soustrait simplement les valeurs des colonnes **depart_ts** et **arrival_ts**. Par contre, il faut avant cela transformer les chaînes de caractères représentant les dates au format ISO en objets "date". Pour ce faire on utilisera la bibliothèque **datetime** qui contient la fonction isoparser, qui convertira les Strings en objets acceptant les opérations arithmétiques.

On peut par la suite afficher les valeurs précédentes sous forme de graphe. On voit qu'il y a certaines valeurs aberrantes avec une durée de 20 jours de trajet pour 700km de distance.

On peut aussi afficher l'intervalle de temps qu'il y a entre la date de la recherche d'un billet et la date de départ. On y voit que la majorité des utilisateurs recherche un billet au minimum un mois à l'avance.

Relation Prix/Distance:

Ici le principe est de déterminer la distance d'un trajet avec les informations **o_stations**, **d_stations** et **middle_stations**. Cependant pour de nombreux trajets, ces valeurs sont notées null, dans ce cas on prendra alors les valeurs **o_city** et **d_city** pour calculer la distance.

Pour ce faire on utilisera la bibliothèque **geopy** qui comporte un outil permettant de calculer une distance suivant une géodésique en fonction de deux coordonnées (latitude, longitude). Cette manière de faire a tout de même sa faille car celle-ci calcule une distance en "vol d'oiseau" qui est une vague approximation du chemin que prennent les véhicules.

Une manière d'y remédier aurait été de recueillir la distance à partir de l'API de Google Maps, cependant cette approximation ne nous dérangera pas pour continuer notre analyse.

En premier lieu, on affichera les valeurs du prix en fonction de la distance sous forme d'un nuage de points. Après cela, on refera la même manipulation mais en prenant cette fois-ci **le type de transport** en compte, on colora alors les points **selon leurs types**.

On pourra par la suite utiliser **sklearn** pour entraîner un modèle de régression linéaire avec les données précédentes. Le modèle devra déterminer le prix d'un ticket en fonction de la distance en kilomère pour chacun des types de transport. Le but est de pouvoir entrer manuellement les valeurs (distance, type_transport) pour obtenir en retour le prix estimé.

Relation Prix/Destinations:

Dans cette partie, on commencera tout d'abord à déterminer les villes d'origine et de destinations les plus prisées (avant même de faire les tests, on s'attend à ce qu'elle soit toute en France). On fera par la suite la même analyse sur les pays d'origine et de destination.

Cela nous amène au prochain point où on déterminera la relation entre le pays de destination et le prix du voyage. Ici on s'attendait certainement à ce que plus le pays de destination est loin plus le prix du trajet est élevé; ce qui est le cas d'ailleurs. Pour contrer cela, on normalise le prix par la distance en kilomètre à parcourir jusqu'à la ville de destination, pour calculer par la suite le prix moyen au kilomètre.