tictactrip

Tictactrip - Test Technique

Introduction

Ce projet est réalisé dans le cadre d'un entretien technique avec la société Tictactrip. L'idée principale est de réaliser l'analyse des données de voyage fournies par la société.

Analyse des données

https://github.com/emnasouki/test_tictactrip/blob/main/Test_technique_Tictactrip.ipynb

L'objective de cette partie est de réaliser les analyses des données nécessaire affin d'extraire les informations intéressantes .

Discrption des données fournis par Tictactrip

1. cities :

https://github.com/emnasouki/test_tictactrip/blob/main/data/cities.csv

Contenant les villes desservies par tictactrip .

Numeric features:

- **Id :** Real number (ℝ), représente l'ID de ville, chaque ligne a son propre ID, unique.
- **latitude :** Real number (ℝ), représente la latitude de la ville .
- **longitude :** Real number (ℝ), représente la langitude de la ville .
- **population :** Real number (ℝ), représente la langitude de la ville .

Texte features:

- **local_name :** Texte, répresente l'ensemble des noms attribuées à la ville .
- **unique_name :** Texte ,répresente le nom unique de ville, unique .

2. stations :

https://github.com/emnasouki/test_tictactrip/blob/main/data/stations.csv

Contenant les stations desservies par tictactrip.

Numeric features:

- **Id :** Real number (ℝ), représente l'ID de station, chaque ligne a son propre ID, unique .
- **latitude :** Real number (ℝ), représente la latitude de la station .
- **longitude :** Real number (ℝ), représente la langitude de la station .

Texte features:

- **unique_name :** Texte ,répresente le nom unique de station, unique.

3. Providers :

https://github.com/emnasouki/test_tictactrip/blob/main/data/providers.csv

Contenant des information sur les différents providers.

Numeric features:

- **Id :** Real number (ℝ), représente l'ID , chaque ligne a son propre ID, unique .
- **company_id :** Real number (ℝ), représente l'ID des societés des voyages , chaque ligne a son propre ID, unique .
- **provider_id :** Real number (ℝ), représente l'ID de fournisseueur des voyages , unique, il y a des valeurs manquantes .

Categorical features:

- **transport_type :** Categorical ,répresente le type de transport (bus, train, carpooling,car).

Texte features:

- **name :** Texte ,répresente le nom de fournisseueur des voyages, unique.
- **fullname :** Texte ,répresente le nom complet de fournisseueur des voyages, unique.

Boolean features:

- **has_wifi :** Boolean ,la valeur est égale à True si le fournisseueur des voyages offre l'option de Wifi et la valeur est égale à False sinon , il y a des valeurs manquantes .
- **has_plug :** Boolean ,la valeur est égale à True si le fournisseueur des voyages offre l'option de plug et la valeur est égale à False sinon , il y a des valeurs manquantes .
- **has_adjustable_seats :** Boolean ,la valeur est égale à True si le fournisseueur des voyages offre l'option des sièges réglables et la valeur est égale à False sinon , il y a des valeurs manquantes .
- **has_bicycle :** Boolean ,la valeur est égale à True si le fournisseueur des voyages offre la possibilité d'apporter un vélo et la valeur est égale à False sinon , il y a des valeurs manquantes .

4. Tickets :

https://github.com/emnasouki/test_tictactrip/blob/main/data/ticket_data.csv

Contenant un historique de ticket (une ligne => une proposition de ticket sur tictactrip)

Numeric features:

- **Id :** Real number (ℝ), représente l'ID de ticket, chaque ligne a son propre ID, unique.
- **company :** Real number (ℝ), représente l'ID de societés des voyage de ticket .
- **o_station :** Real number (ℝ), représente l'ID de station de depart , il y a des valeurs manquantes.
- **d_station :** Real number (ℝ), représente l'ID de station d'arrivée , il y a des valeurs manquantes.
- **price_in_cents :** Real number (ℝ), représente la prix de ticket en cents .
- **middle_stations :** ensemble des ID des stations intermédiaires .
- **other_companies :** ensemble des ID des autres societés des voyage de ticket .
- **o_city :** Real number (ℝ), représente l'ID de ville de départ.
- **d_city :** Real number (ℝ), représente l'ID de ville d'arrivée.

Texte features:

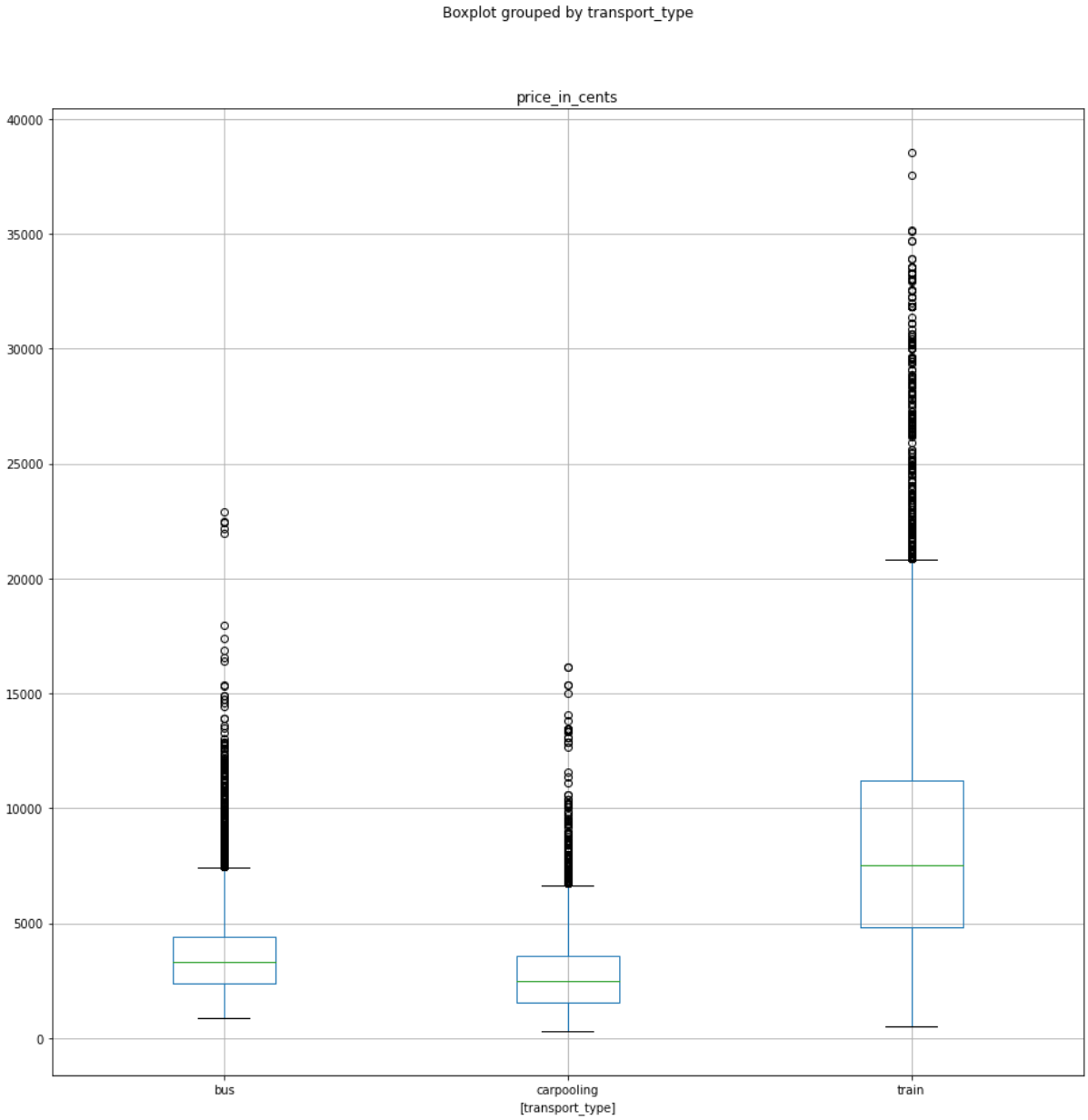
- **departure_ts :** Date, représente la date et l'heure de départ .
- **arrival_ts :** Date, représente la date et l'heure d'arrivée .
- **search_ts :** Date, représente la date de "search" .

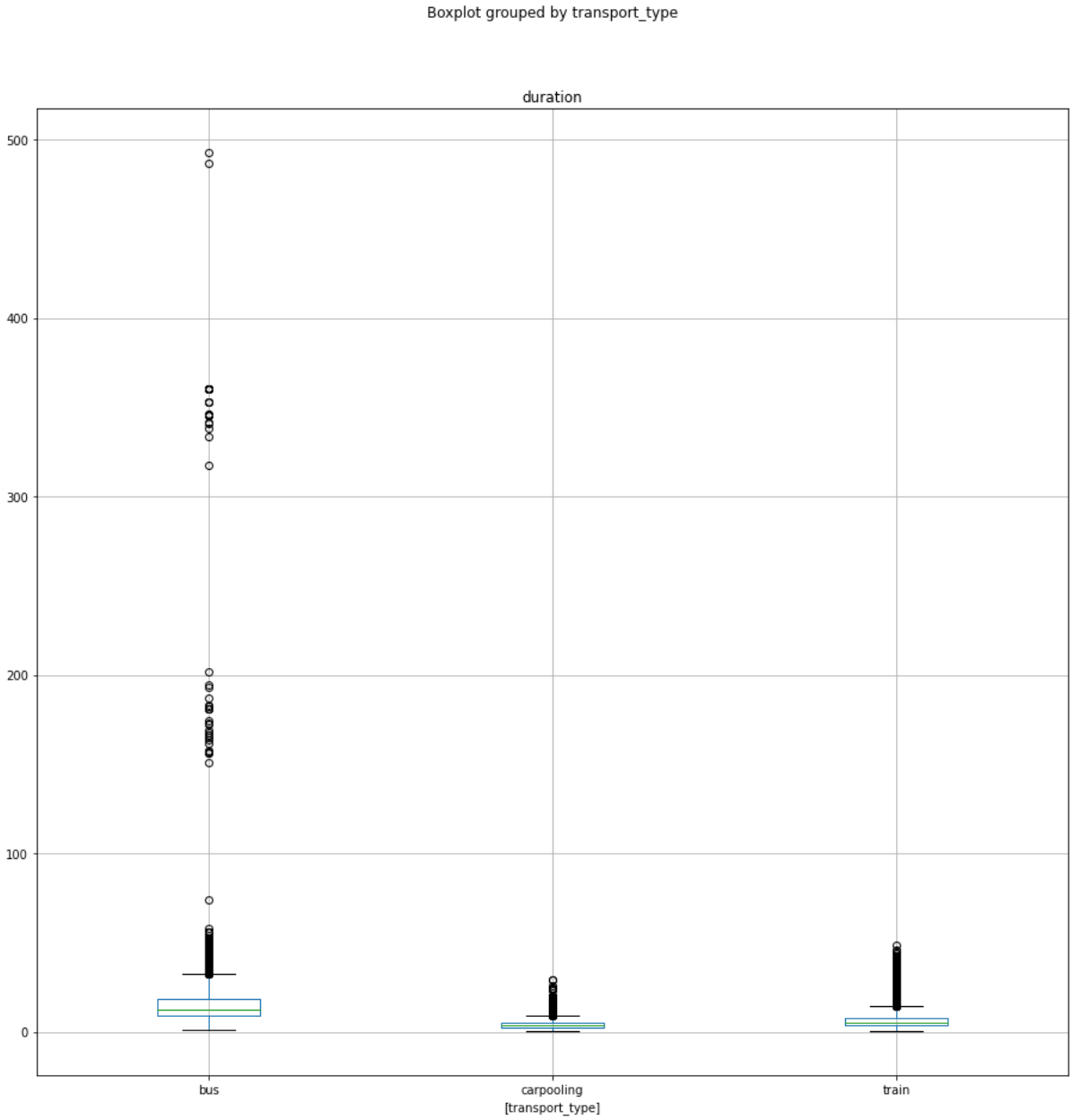
Extraction d'informations intéressantes et visualisation

1. Géneration des informations interessante distance,durée en heures et trajet de chaque voyage:

trajet	distance	duration
orleans to montpellier	503.136831	6.166667
orleans to montpellier	503.136831	17.833333
orleans to montpellier	503.136831	31.950000
orleans to montpellier	503.136831	21.583333
orleans to montpellier	503.136831	21.766667
...	...	...
paris to nantes	339.046766	13.500000
paris to nantes	339.046766	6.500000
paris to nantes	339.046766	6.750000
paris to nantes	339.046766	7.750000

2. Visualisation des prix et des durées selon les différents types de transport :

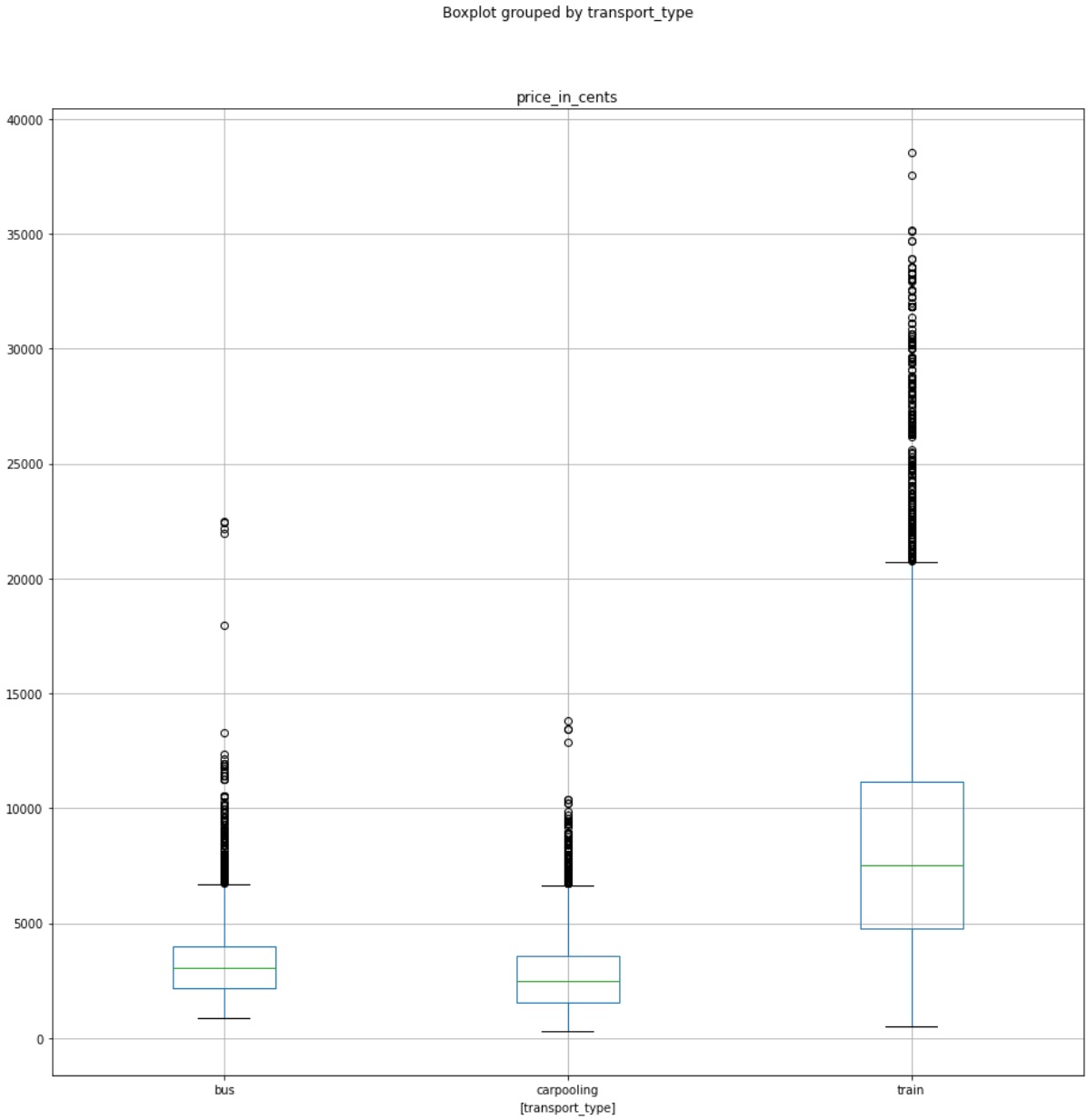

Boxplot grouped by transport_type
price_in_cents
40000
35000
30000
25000
20000
15000
10000
5000
0
bus
carpooling
train
[transport_type]

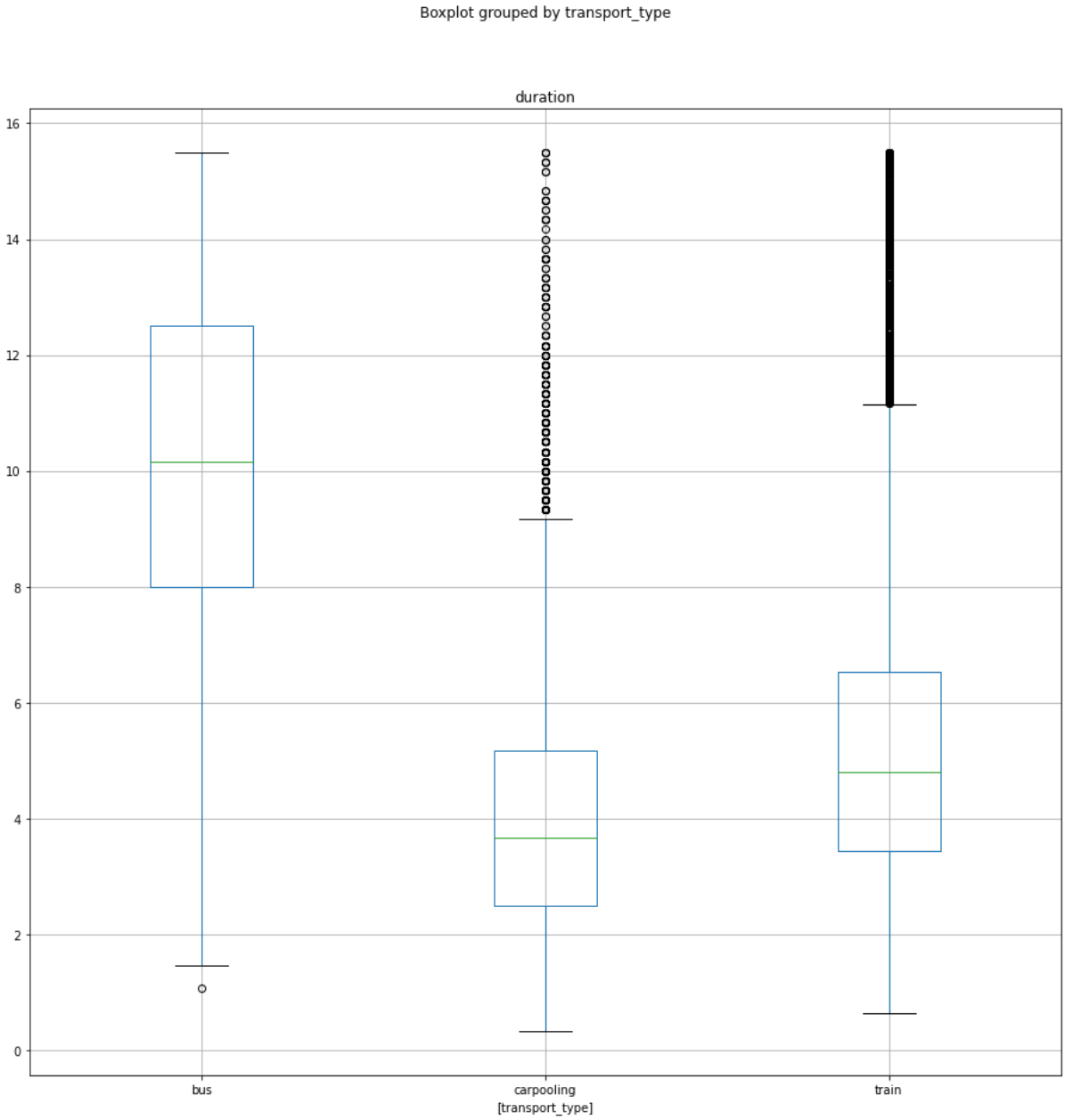


Selon cette visualisation, nous pouvons constater l'existence de plusieurs durées aberrantes, ou la nécessité de les éliminer.
Pour atteindre ce but j'ai utilisé la méthode **IQR** pour les détecter et puis les supprimer.

Chaque exemple de données qui se situe à une valeur supérieure à (Q1 + 1.5 (Q1 +Q3)) est considéré comme une anomalie.

3. Visualisation des prix et des durées selon les différents types de transport après élimination des données aberrantes :


Boxplot grouped by transport_type
price_in_cents
40000
35000
30000
25000
20000
15000
10000
5000
0
bus
carpooling
train
[transport_type]



Présentation des résultats

1. Presentation des prix min, moyen ,max et median par trajet et durée min, moyenne ,max et mediane par trajet :

	duration				price_in_cents			
	min	max	mean	median	min	max	mean	median
trajet								
agde to amsterdam	8.883333	8.950000	8.916667	8.916667	19300	22000	20750.000000	20850.0
agen to dijon	12.400000	15.016667	13.708333	13.708333	9860	13650	11755.000000	11755.0
agen to marseille	5.600000	12.333333	8.290000	7.958333	2000	8920	4042.666667	3400.0
agen to marseille-aeroport	5.000000	8.000000	5.833333	5.333333	3700	4050	3833.333333	3800.0
agen to paris	9.666667	14.500000	12.166667	12.250000	2600	3190	2797.500000	2700.0
...	...	...	...	...	...	...	...	...
villefranche-sur-cher to bordeaux	3.166667	14.750000	7.458333	5.841667	3060	7670	5988.750000	6465.0
vitre to nice	9.883333	11.083333	10.483333	10.483333	15070	21920	18495.000000	18495.0
zurich to dijon	8.166667	8.250000	8.208333	8.208333	1400	2400	1900.000000	1900.0
zurich to liege	5.833333	11.500000	7.722222	5.833333	4050	4190	4096.666667	4050.0
zurich to strasbourg	4.916667	10.583333	7.000000	6.583333	1400	1780	1590.000000	1590.0

1294 rows × 8 columns

2. Présentation des différents prix moyens, min, max et médians et des temps moyens, min, max et médians selon les différents types de transport et selon la distance du trajet :

- les types de transport : train, bus et covoiturage
- la distance du trajet : (0-206km),(207-339km),(340-481km) et (481+km)

  ces distances sont choisies en fonction de la distribution des quantiles.

		duration				price_in_cents			
		min	max	mean	median	min	max	mean	median
RangeDistance	transport_type								
0-206km	bus	1.083333	15.500000	7.093757	6.750000	850	9800	2088.541373	1984.0
	carpooling	0.333333	14.333333	2.242571	2.333333	300	13450	1377.688039	1450.0
	train	0.650000	15.500000	3.532682	2.900000	490	25100	3712.950000	3350.0
207-339km	bus	2.483333	15.500000	9.081734	8.500000	1000	11810	2807.467053	2400.0
	carpooling	1.333333	14.333333	3.818259	3.666667	950	13800	2587.597020	2650.0
	train	1.483333	15.433333	4.876881	4.083333	1400	31850	7359.501493	6295.0
340-481km	bus	4.483333	15.500000	9.886426	9.666667	1180	22480	3490.065702	3200.0
	carpooling	2.666667	13.500000	5.308721	5.166667	1550	9200	3640.212252	3600.0
	train	2.150000	15.500000	5.956007	5.466667	2400	33900	9339.042849	8380.0
481+km	bus	3.416667	15.500000	12.346061	12.500000	1400	17980	3975.000000	3780.0
	carpooling	3.833333	15.500000	7.797021	7.666667	2300	10400	5318.825556	5250.0
	train	2.516667	15.483333	6.690941	5.933333	1400	38550	11305.764356	11000.0

Price regressor

https://github.com/emnasouki/test_tictactrip/blob/main/Price_regressor.ipynb

Pour la prédiction des prix, j'ai utilisé l'algorithme Random Forest et l'ai entraîné sur des données collectées précédemment, en choisissant comme features:

- Duration
- Prix_en_cent
- Id du ville de de départ
- Id du ville de d'arrivée
- Company Id
- Type de transport
- Les offres du tranport (has_bicycle, has_wifi, has_adjustable_seats , has plug)

https://github.com/emnasouki/test_tictactrip/blob/main/data/All_data.csv

Après avoir examiné les résultats de ce modèle (erreur quadratique moyenne : 957), il est peut-être préférable d'ajouter davantage de données/caractéristiques et d'essayer des autre modèles tel que la régression linéaire.
Il est également utile de se concentrer davantage sur l'ingénierie des features afin d'inclure des caractéristiques plus fiables pour notre modèle de régression des prix.

Quelques hypothèses à vérifier

Après avoir analysé les données et observé les résultats, une variété d'hypothèses se sont imposées:

- Le prix des billets dépend-il des dates des jours feriés?
- Pour une même distance, les prix changent-ils en fonction de la ville ?
- Est-il utile de tenir compte des distances entre les stations intermédiaires ?