Rapport TictacTrip

Exercice de crunching de data

Auteur : Mathieu Pierronne

Introduction :

Ce mini-rapport a pour but de faire un état des lieux du travail qui a été réalisé : des recherches et démarches entreprises ainsi que les résultats obtenus.

Ce travail s'est décomposé en quatre grande partie. Dans un premier temps, j'ai fait une analyse globale du dataset en étudiant notamment la répartition du prix et de la durée des différents trajets du jeu de données. Ensuite, j'ai étudié, comme demandé, la différence de prix moyen selon le mode de transport et la durée du trajet. Dans un troisième temps, j'ai voulu approfondir l'exploration des données en représentant notamment le prix en fonction de la compagnie de transport, des gadgets mis à disposition ou encore de l'heure de départ. Enfin, dans un dernier temps, j'ai essayé d'appliquer au jeu de données plusieurs algorithmes de prédiction pour prédire le prix des billets en fonction de différentes caractéristiques.

Avant de renter plus en détail dans mes résultats, je voulais vous remercier de la mission que vous m'avez donnée car j'ai pris beaucoup de plaisir à jouer avec ce jeu de données et d'en extraire de l'information. J'espère avoir l'occasion d'échanger avec vous de vive voix pour vous témoigner ma motivation et ma soif d'apprendre.

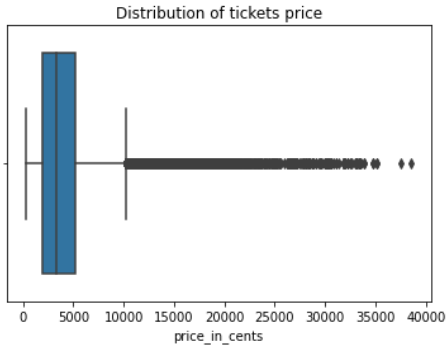
Sommaire

I/ Analyse globale

Cette première partie de l'exercice avait pour but de faire une analyse globale du jeu de données, en étudiant notamment la répartition du prix et de la durée des différents trajets sur ce jeu de données.

1) Répartition prix des billets

La répartition des prix des billets dans ce jeu de données est la suivante :




```
ticket_data['price_in_cents'].describe()

count    74168.000000
mean      4382.711061
std       3739.325367
min        300.000000
25%       1900.000000
50%       3350.000000
75%       5250.000000
max      38550.000000
Name: price_in_cents, dtype: float64
```

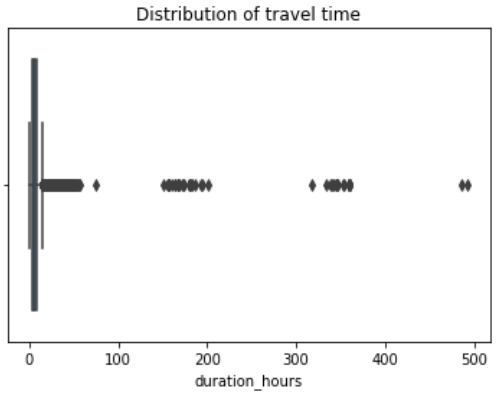
On peut voir que le prix moyen d'un ticket chez TicTactrip est de 4382 cents (soit 43,82€), le prix maximum est de 38 550 cents (soit 385,5€), et le prix minimum de 300 cents (soit 3€).

Ces valeurs sont basées sur un échantillon de 74 168 tickets.

En représentant la répartition des prix avec une boxplot, on observe que 75% des prix sont compris entre 3€ et 52,50€.

2) Répartition de la durée des trajets sur TicTacTrip

On veut maintenant déterminer la durée moyenne d'un trajet ainsi que son min et son max. Après traitement des données, on obtient la répartition suivante des données :



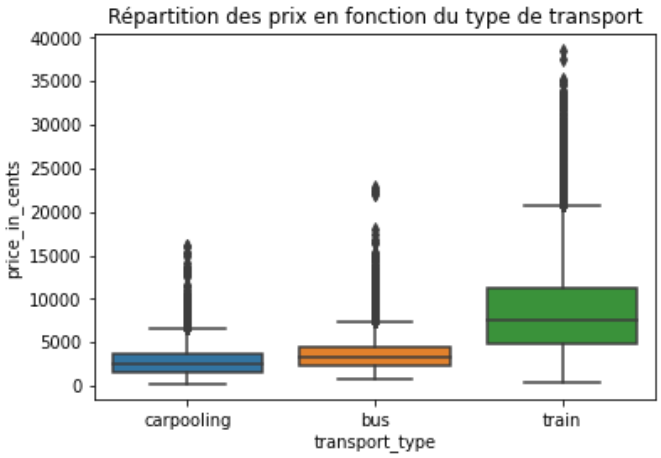

```
ticket_data['duration_hours'].describe()

count    74168.000000
mean         7.076948
std          9.916370
min          0.330000
25%          3.000000
50%          4.830000
75%          8.000000
max        492.850000
Name: duration_hours, dtype: float64
```

On peut voir que la durée moyenne d'un trajet sur cet échantillon de données est de 7 heures, la durée maximale est de 492 heures (soit 20,5 jours) et la durée minimale de 0.33 heures (soit 19 minutes). 75 % des trajets ont une durée inférieure à 8 heures.

II/ Répartition du prix selon la durée du trajet et le moyen de transport

1) Différence de prix selon le type de transport

On souhaite déterminer la différence de prix selon le type de transport. Après traitement des données, on obtient la répartition suivante :




```
====================================
la répartition des prix pour le carpooling
count    41441.000000
mean      2742.171907
std       1501.934054
min        300.000000
25%       1550.000000
50%       2500.000000
75%       3600.000000
max      16150.000000
Name: price_in_cents, dtype: float64
====================================
la répartition des prix pour le bus
count    13798.000000
mean      3652.448036
std       1913.197779
min        850.000000
25%       2390.000000
50%       3300.000000
75%       4400.000000
max      22900.000000
Name: price_in_cents, dtype: float64
====================================
la répartition des prix pour le train
count    18929.000000
mean      8506.634793
std       4888.064503
min        490.000000
25%       4800.000000
50%       7540.000000
75%      11200.000000
max      38550.000000
Name: price_in_cents, dtype: float64
```

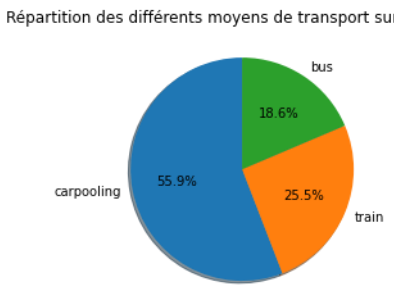
On peut observer que le mode de transport le moins cher est le covoiturage (carpooling), puis vient le bus et train est le mode de transport le plus cher.

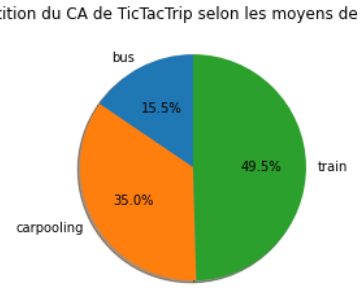
On constate également que plus le prix moyen est important, plus la répartition des prix est inégale : on retrouve ainsi dans les 3 cas un prix minimum similaire (de 3€ pour le covoiturage, 8,5€ pour le bus et 4,90€ pour le train) mais un prix maximum de 385€ pour le train, 229€ pour le bus et 361€ pour le covoiturage.

Les chiffres intéressants sont :

- Covoiturage :
  - moyenne : 25€
  - min : 3€
  - max : 161€
- Bus :
  - moyenne : 33€
  - min : 8,5€
  - max : 229€
- Train :
  - moyenne : 75€
  - min : 4,9€
  - max : 385€

Avant de passer à la suite, je trouvais intéressant de regarder la proportion de ces 3 différents modes de transport sur TicTacTrip et le pourcentage du Chiffre d'affaire (CA) pour chacun des types de transport. On obtient le graphe suivant :

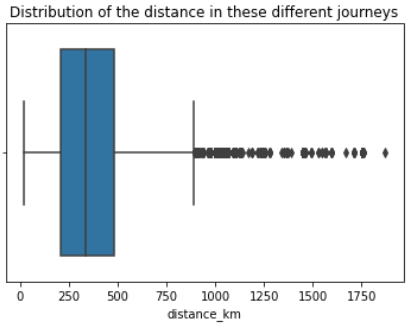




Il est intéressant de constater que le covoiturage est le moyen de transport le plus utilisé, vient ensuite le train puis le bus. Bien que le train ne représente que 25% des trajets, il représente près de 50% du chiffre d'affaire.

2) Différence de prix selon la distance parcourue

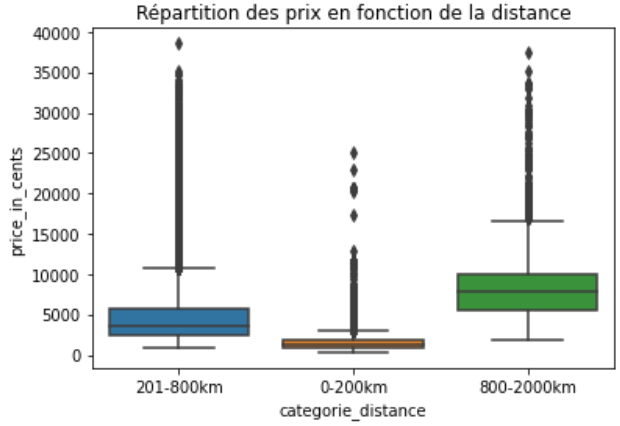
On veut ensuite déterminer le prix en fonction de la distance parcourue. Après avoir déterminé la distance de chaque trajet en utilisant la formule Haversine (on utilise les coordonnées lon-lat de la ville d'origine et de la ville d'arrivée), on obtient les résultats suivants :




```
ticket_data['distance_km'].describe()

count    74168.000000
mean       363.152147
std        194.991807
min         18.919179
25%        205.907209
50%        338.426341
75%        480.564632
max       1870.759110
Name: distance_km, dtype: float64
```

Il est interessant de constater que la distance moyenne parcourue est de 363km, la distance minimale est de 18km et la distance maximale de 1870km.

On a ensuite voulu déterminer la répartition de ces prix en fonctions de 4 catégories de distance du trajet : 0-200km, 201-800km, 800-2000km et 2000+km. J'obtiens les résultats suivants :

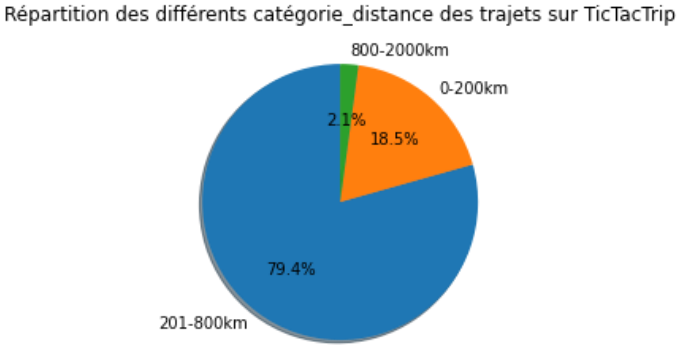



```
==========================================
la répartition des prix pour le categorie 0-200km  est:"
count    13724.000000
mean      1678.972311
std       1340.887966
min        300.000000
25%        950.000000
50%       1300.000000
75%       1800.000000
max      25100.000000
Name: price_in_cents, dtype: float64
==========================================
la répartition des prix pour le categorie 201-800km  est:"
count    58877.000000
mean      4893.836592
std       3745.577272
min        850.000000
25%       2500.000000
50%       3600.000000
75%       5800.000000
max      38550.000000
Name: price_in_cents, dtype: float64
==========================================
la répartition des prix pour le categorie 800-2000km  est:"
count     1567.000000
mean      8857.869177
std       5008.885867
min       1940.000000
25%       5500.000000
50%       7780.000000
75%      10050.000000
max      37550.000000
Name: price_in_cents, dtype: float64
```

Sans surprise, on constate que plus la distance est élevée, plus le prix est élevé.

On constate que le prix moyen pour la catégorie "0-200km" est de 16,78€, pour la catégorie "201-800km" de 48,93€ et pour la catégorie "800-2000km" est de 88,57€.

Avant de terminer cette partie, je trouvais intéressant de regarder la proportion de ces 3 catégories sur le jeu de données. On obtient le boxplot suivant :



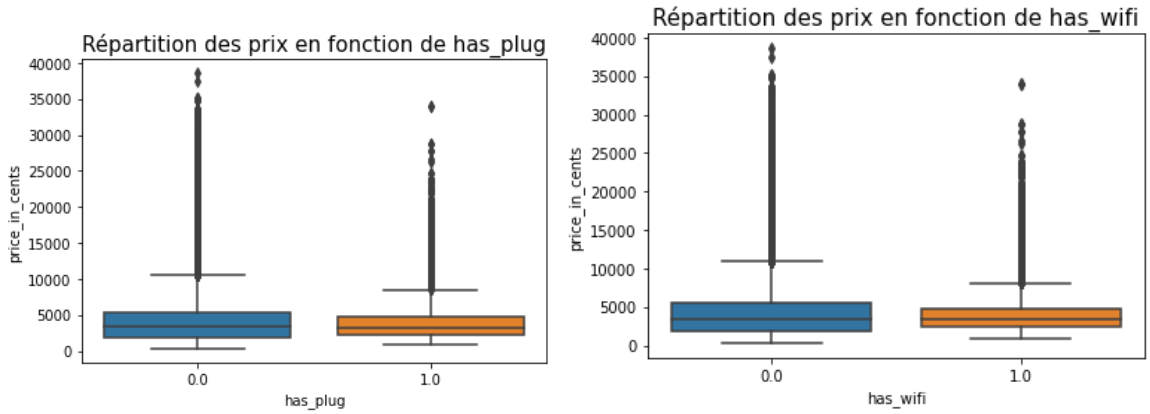
Il est intéressant de constater que les trajets de 200-800km représente la très grosse majorité des trajets sur la plateforme avec près de 80% des trajets totaux

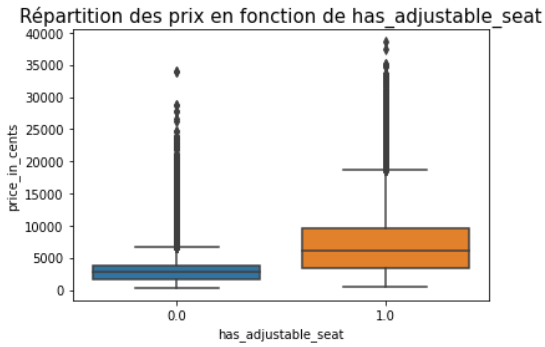
III/ Analyses supplémentaires

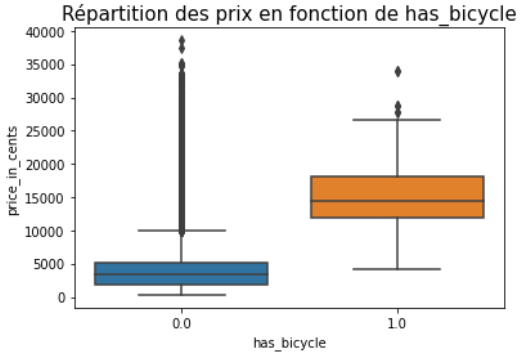
J'ai pu constater que de nombreuses données de la table providers.csv n'avaient pas été exploitées. Dans cette partie, je vais essayer de les exploiter pour extraire des informations supplémentaires.

1) Etude des gadgets sur le prix des billets

Dans cette partie, je me suis intéressé à la répartition du prix des billets en fonction des gadgets mise à disposition lors du trajet (wifi, plug, bicycle) . On obtient les graphes suivants :





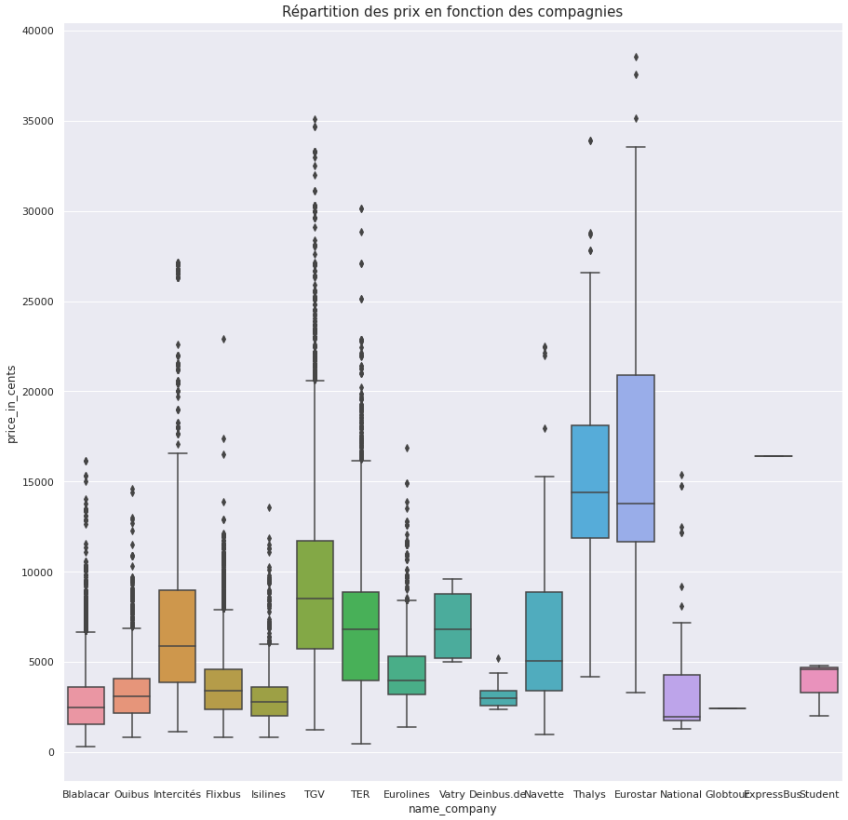


On constate que si la wifi et le plug n'ont pas trop d'influence sur le prix, le fait d'avoir des sièges ajustables ou la possibilité de prendre un vélo augmente considérablement le prix.

2) <u>Répartition des prix en fonction des compagnies</u>

On représente d'abord le prix en fonction des compagnies. On constate que les compagnies les plus chères sont les compagnies de train vers l'étranger (thalys et Eurostar) et les compagnies les moins chères sont les compagnies de covoiturage avec notamment Blablacar.

On remarque également que TGV et Eurostar sont les compagnies avec l'écart type le plus élevé (autrement dit avec les prix les plus éparpillé).



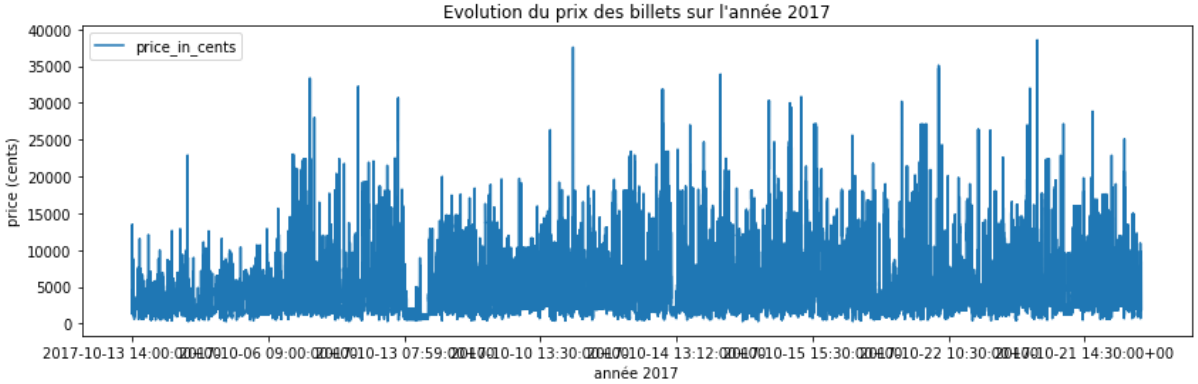
3) Etude de l'évolution des prix en fonction des mois, années et heures.

Je considérais qu'il était intéressant de faire une analyse temporelle des prix et d'étudier les prix en fonction du temps.

Tout d'abord, il aurait été intéressant d'étudier l'évolution des prix au cours des dernières années pour voir s'il y a une évolution significative des prix ainsi que l'évolution des prix sur une année (évolution sur au cours des mois) pour voir si on peut observer des saisonnalités.

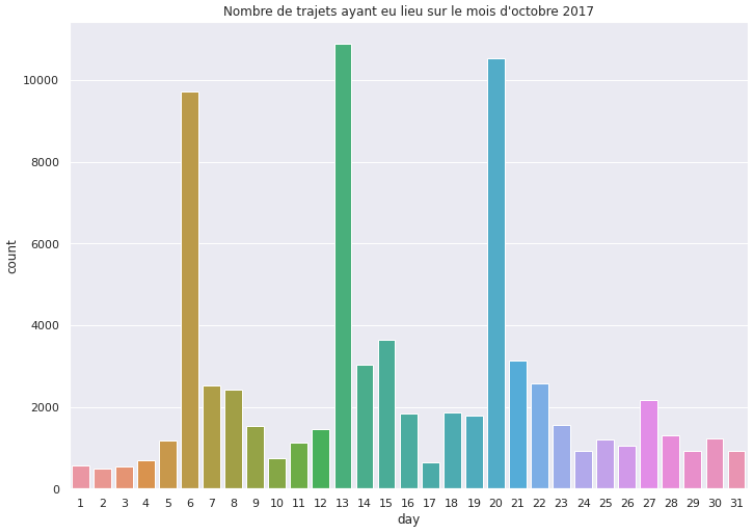
Cependant, nous avons des données seulement pour l'année 2017 et principalement pour le mois d'octobre (mois 10), comme le montre le graphe ci-dessous. De telles analyses ne peuvent alors être effectuées...

On peut toutefois faire une représentation temporelle des données qu'on possède. On obtient le résultat suivant :



On remarque une légère augmentation du prix entre le début et la fin de l'année 2017.

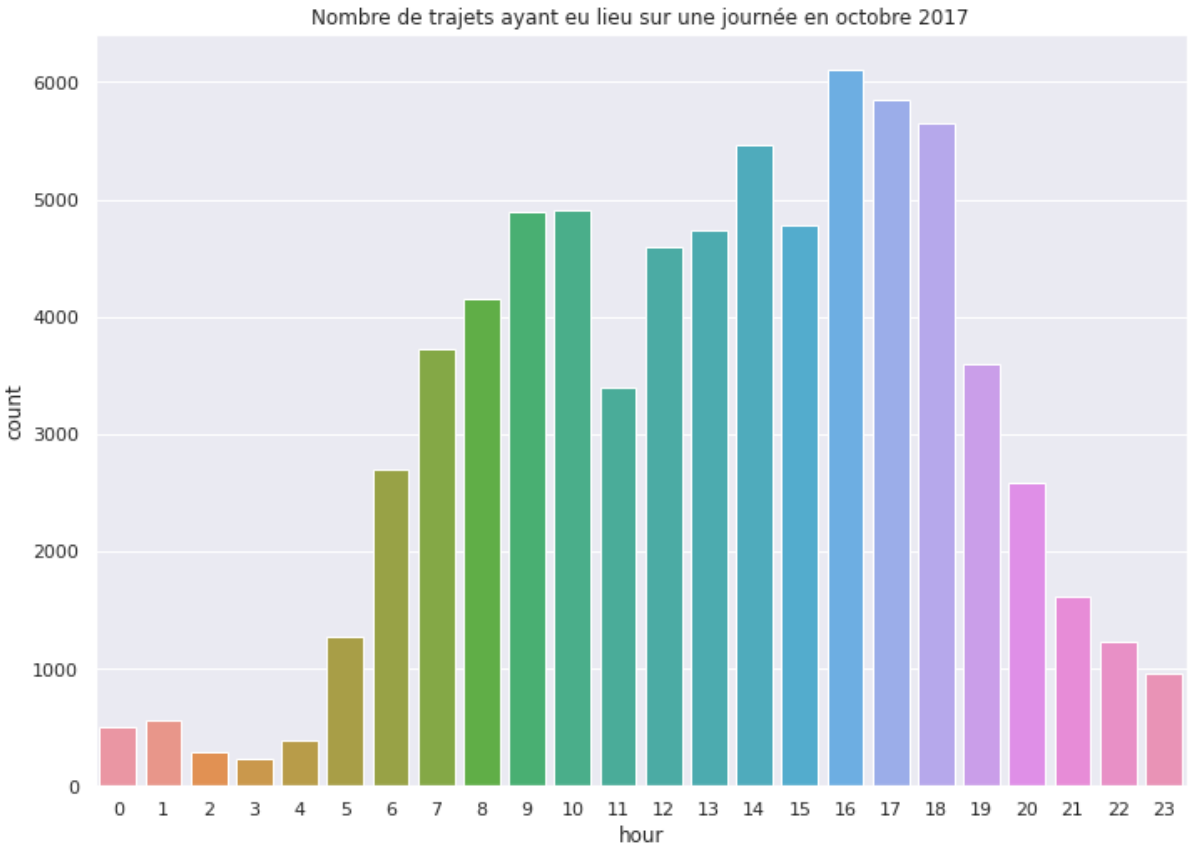
On représente ensuite le nombre de trajets par heure et par jour. On obtient les résultats suivants :



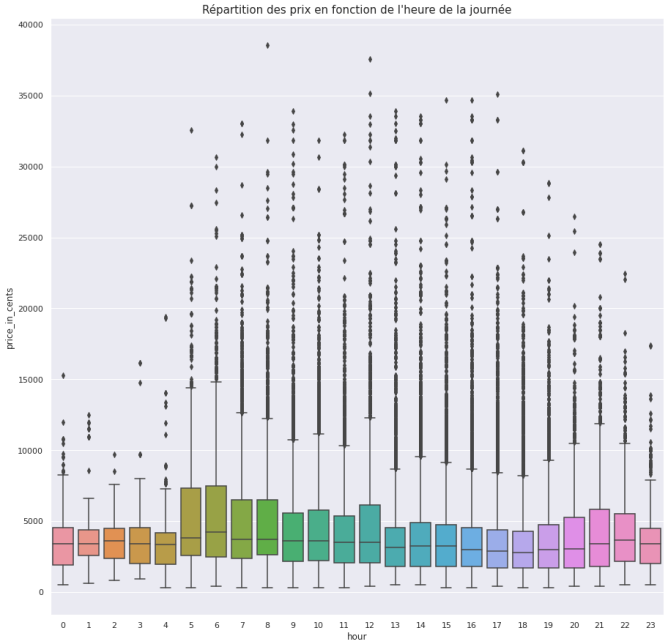
Il est interessant de constater qu'on observe 3 pics le 6, 13 et 20 octobre, qui correspondent tous 3 à des vendredi. Cela est cohérent car lors des départs en weekend, les gens ont plus tendance à faire des trajets longs.

On remarque également une saisonnalité par semaine : il y a un pic le vendredi et plus de départs le samedi, dimanche et lundi qu'en semaine.

On termine en faisant la représentation par heure :



Il y a bien évidement plus de trajet la journée plutôt que de nuit. On fait enfin, la repéartion des prix sur ces différentes heures :



Si le prix moyen à l'heure ne varie pas trop sur une journée, on peut constater que ce prix est beaucoup plus variable (écart type beaucoup plus important) pour des trajets de journée que de nuit.

IV/ Machine Learning : prédictions

Dans cette dernière partie, j'ai voulu appliquer des algorithmes de Machine Learning au jeu de données pour prédire le prix des billets en fonction de différents critères comme la compagnie, l'heure, le jour, les gadgets, la distance ou la durée.

Ayant très peu de données sur les années et les mois, je n'ai pas pris en compte cette caractéristique dans la prédiction même si elle a évidemment une influence.

Par conséquent, nos features (qui aux paramètres influcent le prix) sont : duration_hours, distance_km, has_wifi, has_plug, has_adjustable_seat, has_bicycle, day, hour, transport_type et name_company et le label (qui correspond à la valeur à prédire) est price_in_cents.

Après avoir encodé les string en float avec la fonction One Hot Encoding, j'ai divisé le jeu de données en 10 parties avec la fonction Kfold. Cette technique consiste à diviser l'échantillon original en k échantillons, à sélectionner un de ces k échantillons comme ensemble de validation pendant que les k-1 autres échantillons constituent l'ensemble d'apprentissage.

Cette méthode est bien plus précise qu'une division binaire du dataset en une partie d'apprentissage et une seconde de prédiction.

Une fois cette division du dataset effectué, j'ai appliqué différents algorithmes de Machine Learning pour prédire le price_in_cents à partir des features.

J'obtiens les résultats suivants :

Algorithme	Hyperparamètre	R2_score
DecisionTree	Max_depth = 10	0.8479813291734217
KNeighbors	Max_k = 1	0.8697737612664579
GradientBoosting	N_estim = 300	0.8478964
SVM	/	/
AdaBoost	/	/

Tous ces algorithmes de prédiction semblent relativement similaires et fonctionnent plutôt bien (il y a juste un problème avec AdaBoost).

Merci pour votre temps accordé et pour cette mission.

Mathieu Pierronne