# **RAPPORT DU PROJET IA MASTER 1 BIG DATA ET MANAGEMENT**

***« Comprendre le marché de l’industries des taxis avant d’investir »***



Projet réalisé par :

Projet encadré par :

Pôle Académique IA de l’école de l’Intelligence Artificielle

**SOMMAIRE**

1. INTRODUCTION
2. RÉCUPÉRATION DES DONNÉES PAR SCRAPING
3. IMPORTATION ET VISUALISATION DES DONNÉES
4. ANALYSE ET NETTOYAGE DES DONNÉES
5. ETUDES ET ANALYSES STATISTIQUES
6. Visualisation statistique générale
7. Analyse du profit mensuel et annuel
8. CONCLUSION
9. **INTRODUCTION**

En raison de la croissance remarquable de l'industrie des taxis au cours des dernières années et de plusieurs acteurs clés du marché, d'énormes investissements sont prévus et, conformément à sa stratégie Go-to-Market (plan d'action d'entreprise, visant à procéder avec succès au lancement d'un produit ou service. Sa démarche consiste à mobiliser les ressources internes et externes d'une entreprise pour augmenter le taux de réussite de l'opération), ils souhaitent comprendre le marché avant de prendre décision finale. Les investisseurs souhaitent utiliser vos informations afin d'identifier la bonne entreprise pour réaliser leur investissement.

Base de données 4 ensembles de données individuelles ont été fournies. La période d'enregistrement va du 31/01/2016 au 31/12/2018. Vous trouverez ci-dessous la liste des ensembles de données fournies pour l'analyse :

Cab\_Data.csv - ce fichier contient les détails de la transaction pour 2 compagnies de taxi

Customer\_ID.csv - il s'agit d'une table de mappage qui contient un identifiant unique reliant les détails démographiques du client

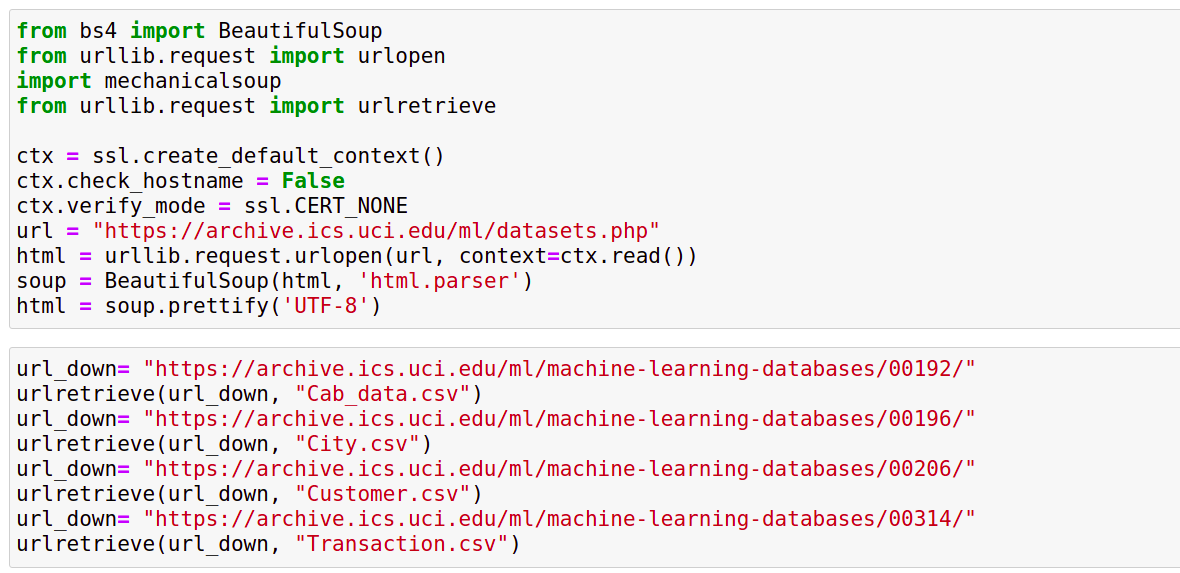
Transaction\_ID.csv - il s'agit d'une table de mappage qui contient le mappage de la transaction au client et le mode de paiement

City.csv - ce fichier contient la liste des villes américaines, leur population et le nombre d'utilisateurs de taxi

Pour ce faire, nous devons au cours de notre étude répondre à une série de questions :

* Y a-t-il une saisonnalité ?
* La demande de l'industrie du taxi a-t-elle tendance à augmenter avec le temps ?
* Le pourcentage de trajets rentables change selon la ville.
* Comment la demande varie selon l'âge ?
* Tarifs fidélité.
* Le mode de paiement fluctue-t-il d'une année à l'autre ? En fonction de l'âge ? Citywise?

1. **RÉCUPÉRATION DES DONNÉES PAR SCRAPING**

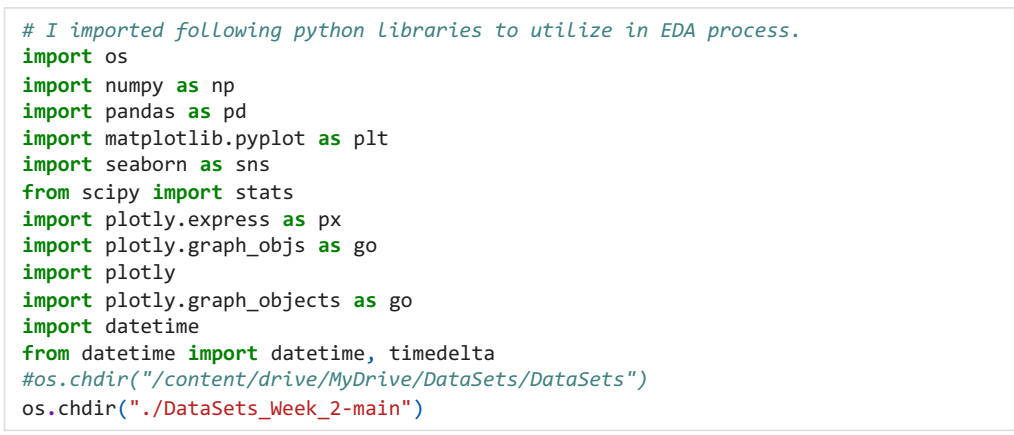
Le scraping est le processus automatisé de récupération (ou de grattage*)*  des données d'un site Web ou d’une plateforme. La première étape pour un data scientist est de récolter les données.

Après la récolte des données, passons au traitement et analyse.

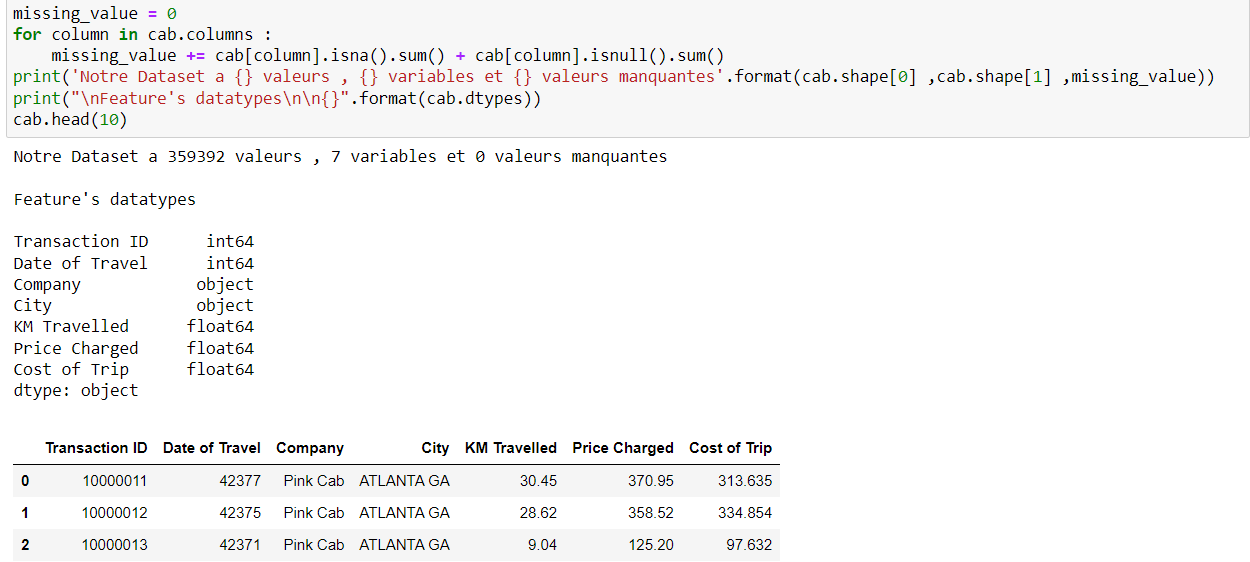
1. **IMPORTATION ET VISUALISATION DES DONNÉES**

La première tâche consiste à importer les bibliothèques qui nous serviront durant notre projet. Pour ce faire, il est strictement nécessaire d’avoir des compétences en data science et analyse afin de pouvoir anticiper les étapes à travers lesquelles nous effectuerons le traitement de nos données.

Les librairies sont en fait importées dans l’ordre croissant de l'exécution des tâches.

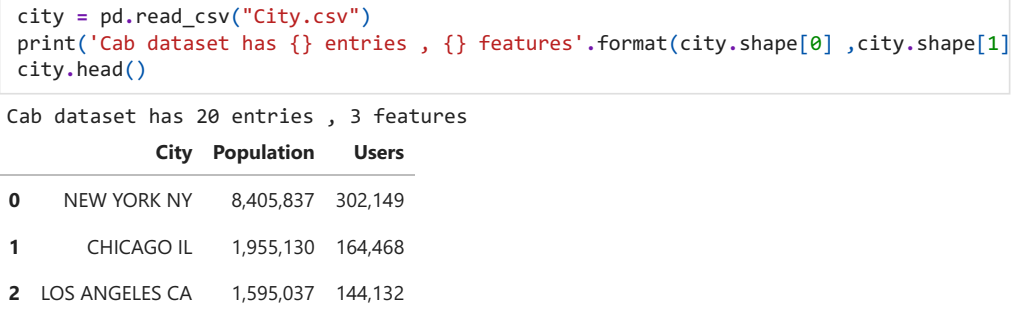


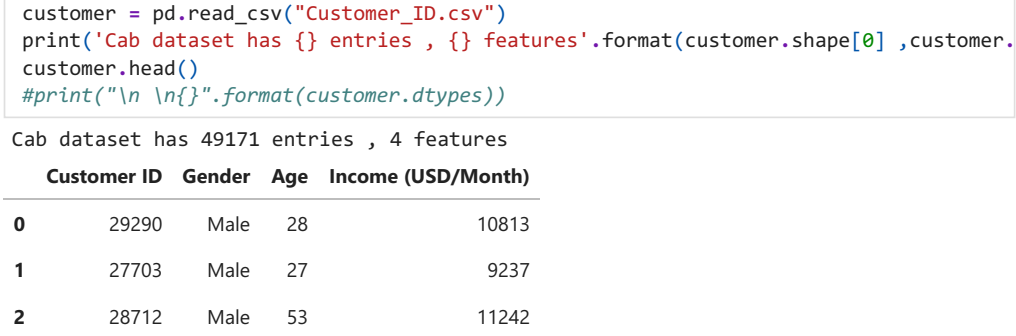
Une fois nos librairies importées, nous procédons à la visualisation de chaque fichier csv (comma-separated values) afin de prendre note premièrement de la taille de tout un chacun, mais aussi des informations primaires importantes et capitales telles que le type de valeurs. Pour cela, nous avons procédé comme le montrent les images suivantes :



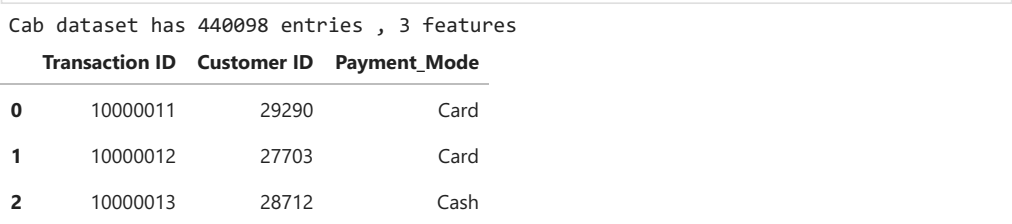
On note bien que le premier dataset n’a aucune valeur manquante et à partir de la boucle on peut obtenir quelques informations basiques.

Nous procédons donc ainsi pour les 3 autres datasets restants.





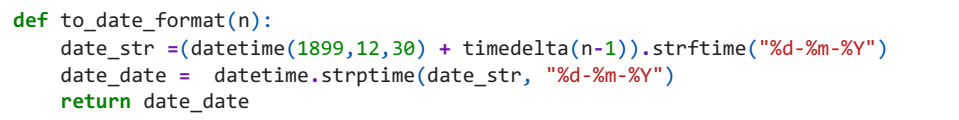


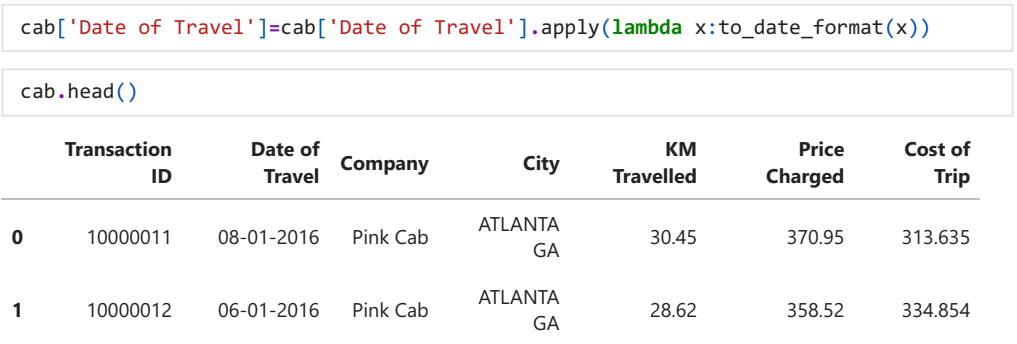


1. **ANALYSE ET NETTOYAGE DES DONNÉES**

Le volet Analyse et nettoyage des données occupe un très grand pourcentage dans la répartition du temps de travail d’un Data Scientist. Cela consiste en quelque sorte à vérifier la présence des valeurs manquantes et se décider sur leur devenir (les supprimer ou les remplacer par la médiane par exemple en fonction du type de travail qu’il y aura à faire), renommer les variables s’il le faut, affecter à chaque variable le type qui lui sera opportun pour l’analyse statistique.

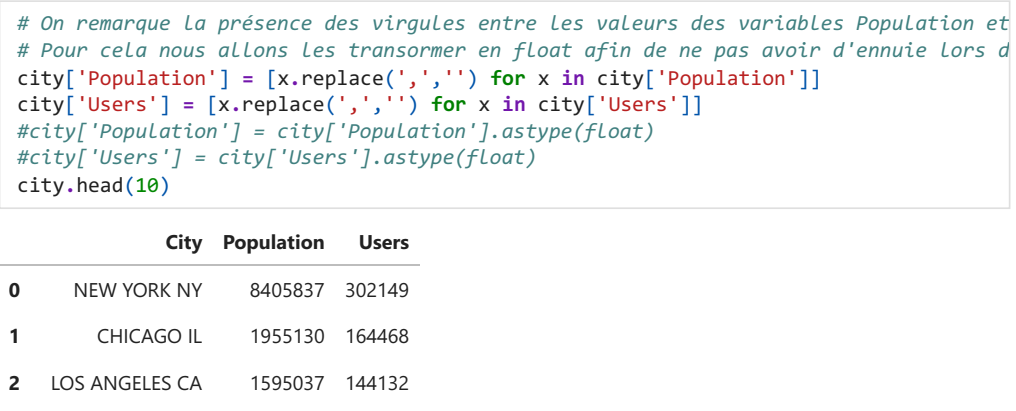
En nous rappelant des observations faites plus haut, nous avons noté l’absence des valeurs manquantes, ce qui est déjà positif ; mais nous avons été intrigués par la variable “Date of Travel” qui était en format <int64>. Alors nous commencerons par modifier ce format en <datetime>.



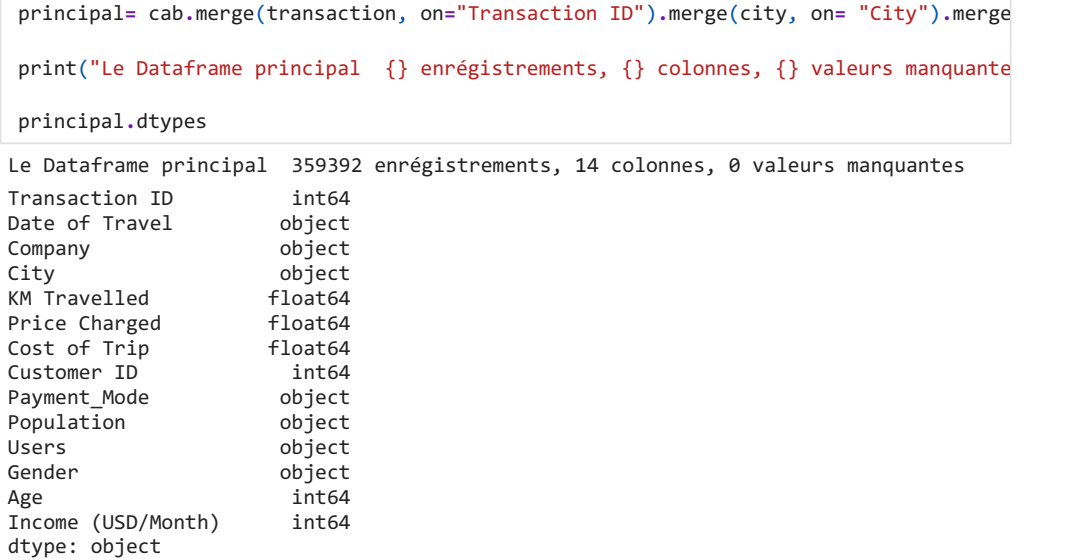




En observant aussi le dataset City.csv, on a remarqué la présence des virgules entre les chiffres. Donc on va procéder à la modification des variables concernées.

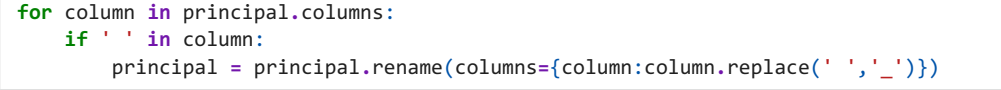


Nous pouvons donc procéder à la fusion des 4 datasets afin d’en avoir un seul compact. Pour ce faire, la fonction <merge> est très efficace et nous a accompagnés dans cette tâche.

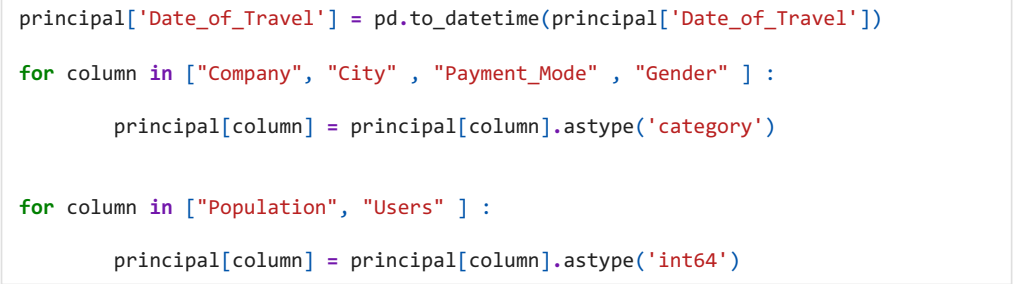


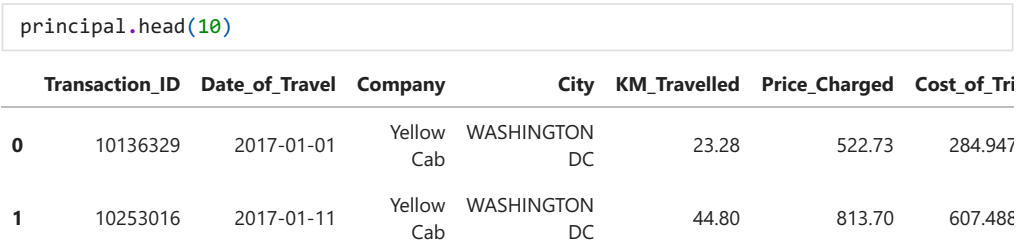
Une dernière étape de nettoyage est donc nécessaire afin de s'assurer que toutes les variables de notre grand dataset soient écrites de façon plus simple (sans les espaces) et ensuite que les types de toutes les variables soient conformes à l’étape suivante qui sera l’étude statistique.

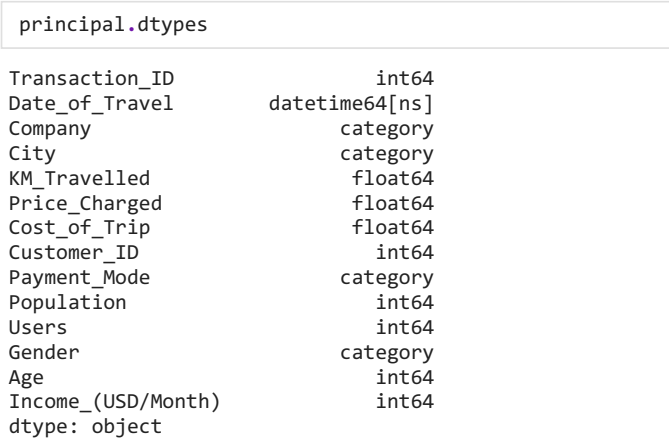
On commence donc par remplacer les espaces dans les noms des variables par des underscores,



Ensuite, convertir les types de variables.





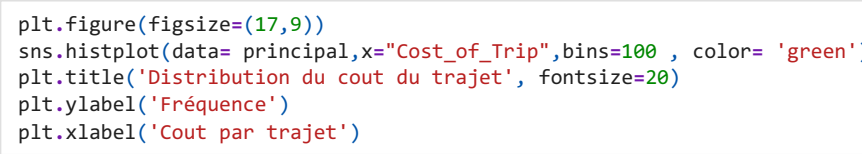


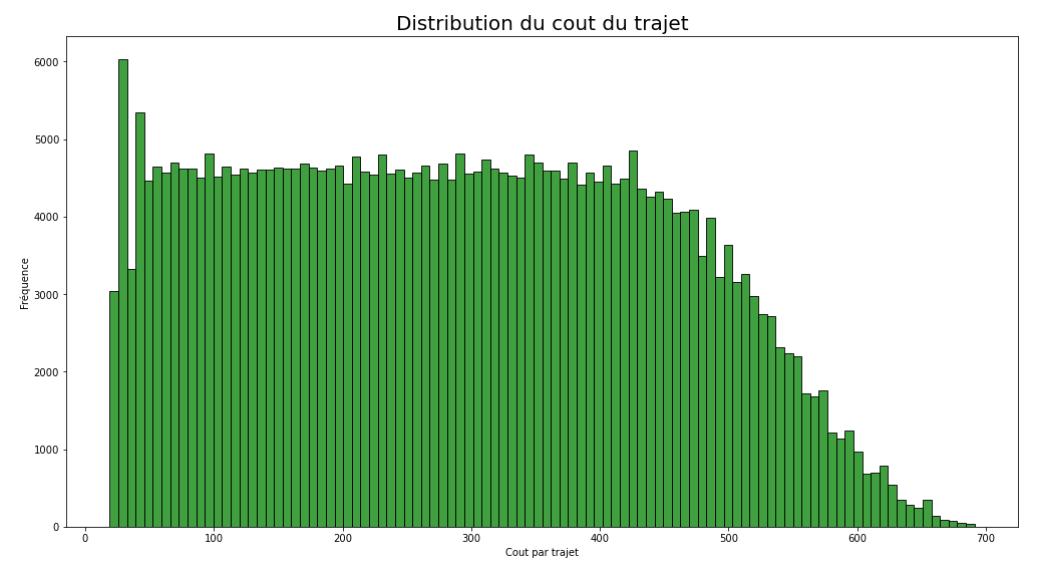
Une fois tout fait, on peut passer à l’étape suivante qui consistera aux analyses statistiques et les visualisations.

1. **ETUDES ET ANALYSES STATISTIQUES**
2. **VISUALISATION STATISTIQUE GENERALE**

Nous nous concentrerons plus sur les différentes thématiques : Distribution, Valeurs aberrantes, Corrélation, Covariance…

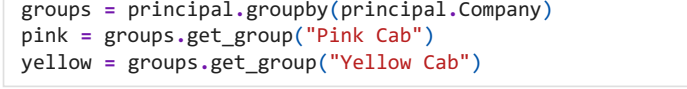
Commençons par une visualisation graphique de la variation globale du coût du trajet.



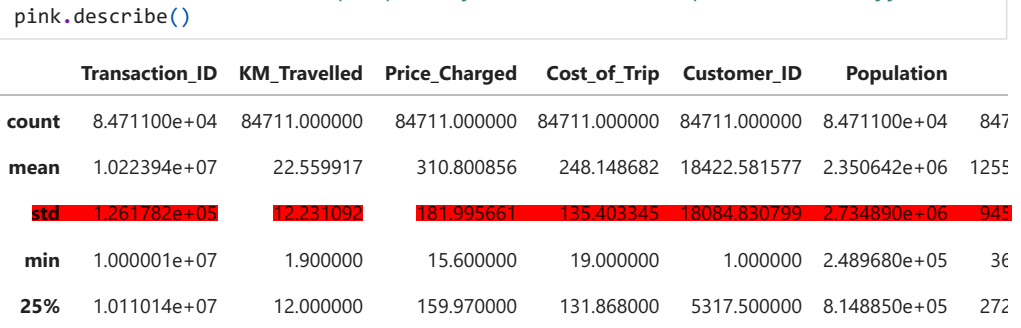


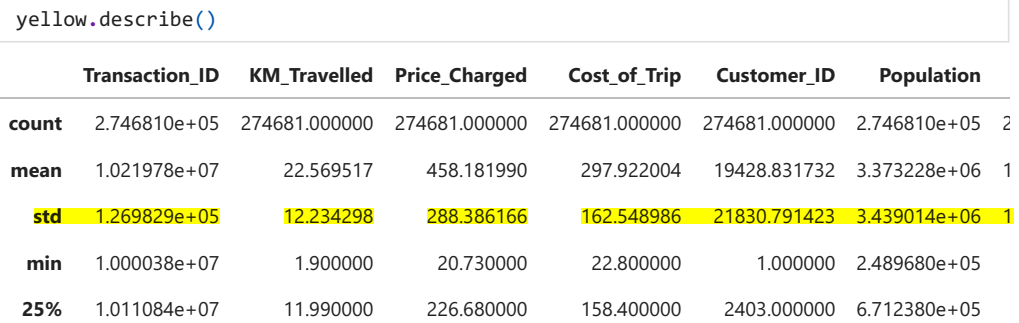
En observant ce graphique, on ne peut ressortir que les trajets à coût très élevés (supérieur à 600) sont de très faible fréquence.

Afin de mener à bon port notre étude, nous devons aller plus en profondeur. Nous commençons donc par regrouper le dataset en fonction du type d’entreprise ("Company").

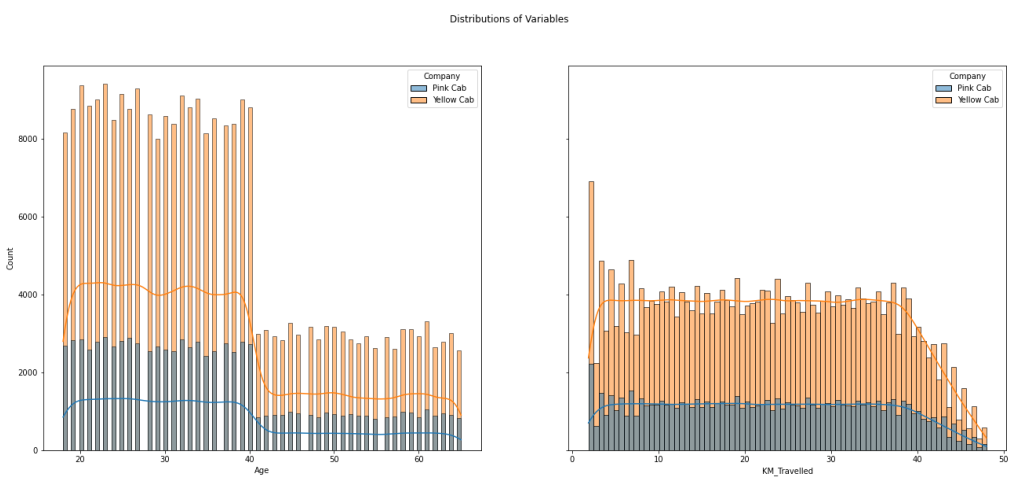


On a ainsi le dataset pink d’un côté, et yellow de l’autre côté. En effectuant simultanément les visualisations primaires de chacun des deux datasets, on note une similitude générale, c'est-à-dire, les valeurs des variables qui sont communes aux deux types d’entreprises sont égales. Mais en étant un peu plus curieux, on peut noter déjà des différences, par exemple en observant les déviations standard, il est très facile de se rendre compte que les valeurs sont différentes. C’est en fait en se basant sur ces différences que nous saurons répondre à la problématique principale et ainsi orienter dans le bon sens les investissements.



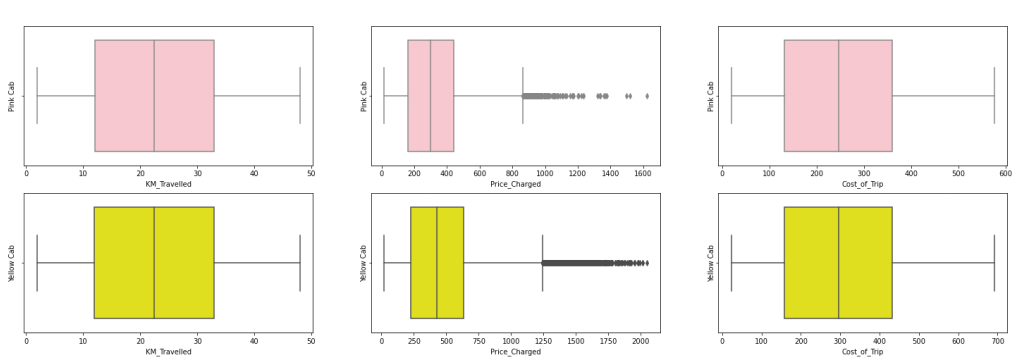


Une méthodologie performante de visualisation et d’estimation des densités de probabilité des variables aléatoires est le KDE (Kernel Density Estimation). Nous allons l’utiliser ici afin d’apprécier les différences de variations de densité de chaque entreprise.



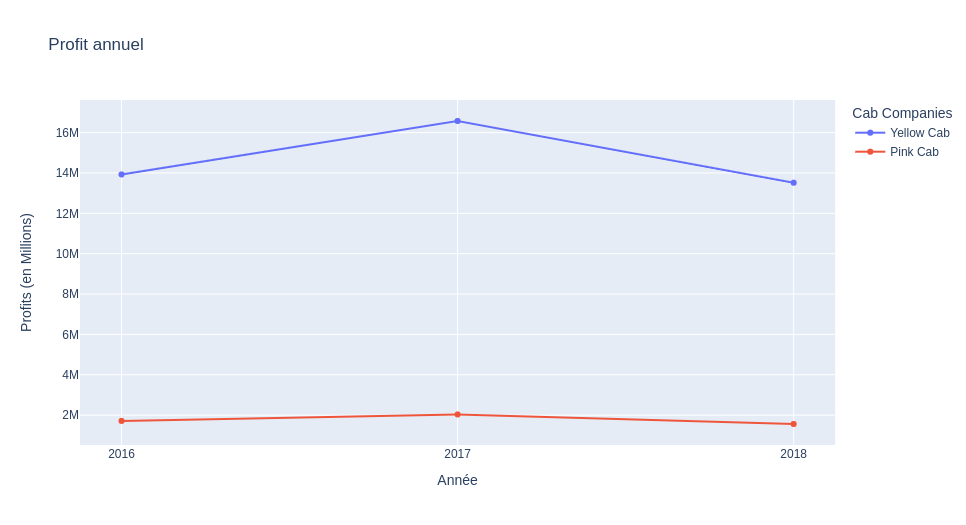
En observant ces deux graphiques représentatifs de la variation de densité de l'âge et des distances pour chacune des compagnies, on se rend compte que la yellow a une distribution beaucoup plus statique et moins oscillante que Pink, ce qui traduit donc un certain équilibre.

L’analyse des boxplots relève des informations pertinentes.



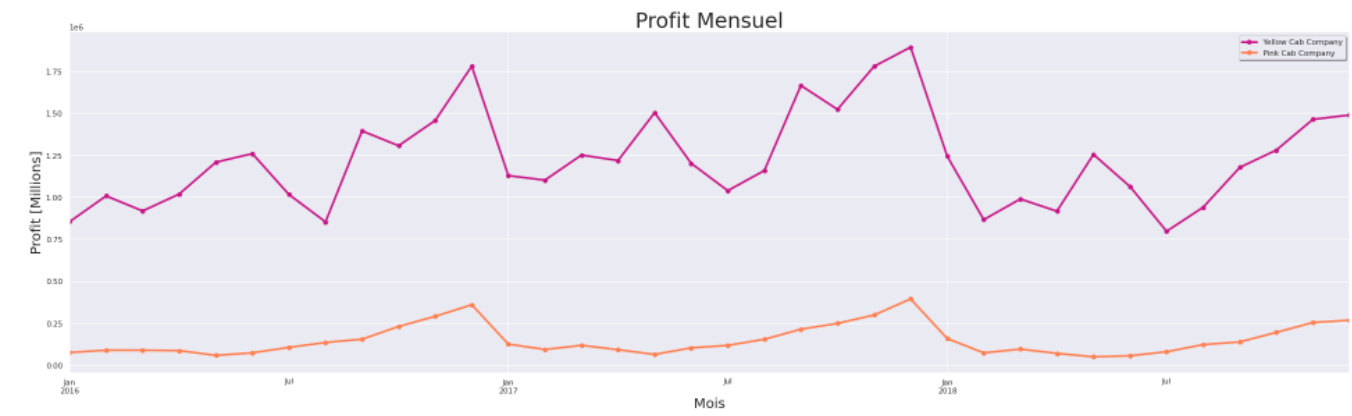
* La distribution des kilomètres parcourus est identique pour les deux entreprises
* On note la présence des outliers dans la variable “Price\_Charged”, on ne tiendra pas compte de cela car ce n’est pas une variable déterminante pour notre analyse.
* Au niveau de “Cost\_of\_Trip”, on se rend compte que Yellow a des coûts unitaires allant au-delà de 400$ tandis que Pink a un maximum d’environ 360$, les deux ayant donc une médiane assez équilibrée

1. **PROFIT ANNUEL ET MENSUEL**

Le graphique suivant nous permet de visualiser la variation du profit annuel des deux entreprises.

En l’observant, il nous vient presque à l’idée de se dire que les deux ont une variation très similaire, mais Yellow, qui est tout en dessus, subit une forte décroissance de 2017 à 2018. Mais nous ne pouvons pas tirer cette conclusion sans avoir calculé le coefficient directeur de chacune des droites afin de déterminer celui qui a une chute plus brutale.

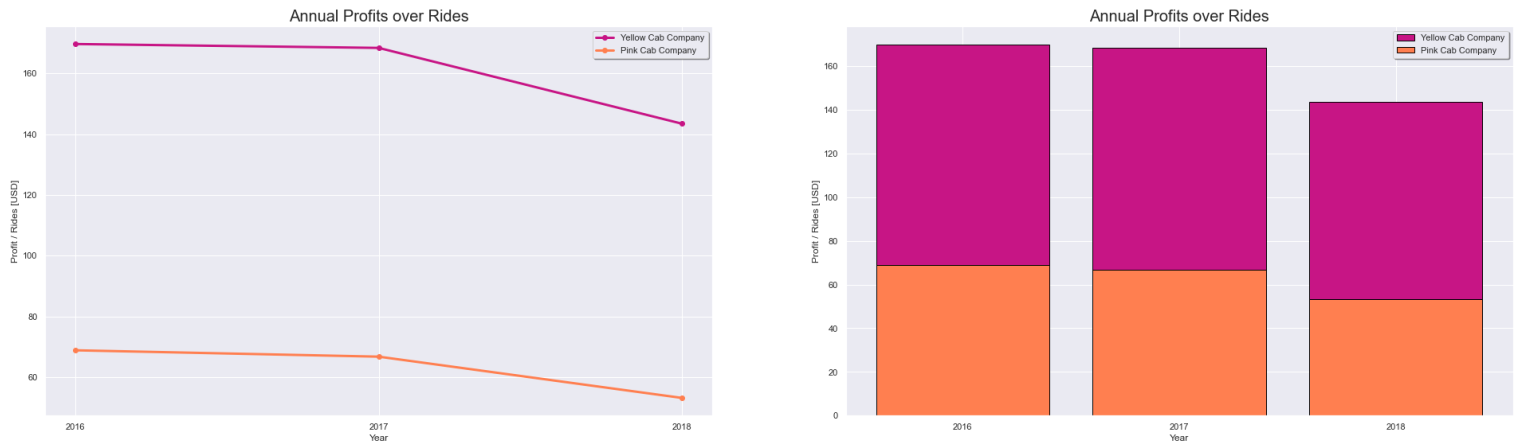
Si on décide de voir la variation mensuelle du bénéfice net, on obtiendra un graphique tel que :



On remarque que Yellow a des fluctuations importantes ce qui peut se traduire par un marché très mouvementé et donc la demande pourrait être aussi importante.

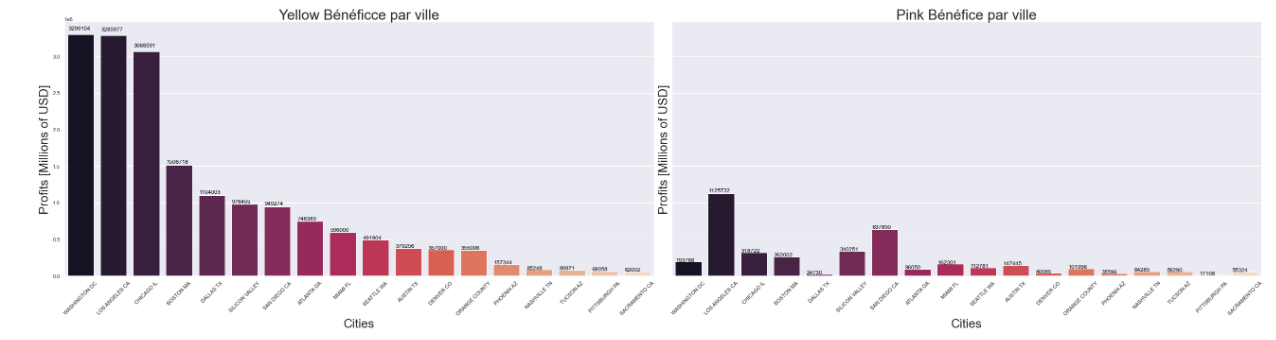
Allons maintenant vérifier l’efficacité de chacune des entreprises. Le profit par trajet est un indicateur qui mesure l'efficacité de l'entreprise en termes de coûts opérationnels.

Bénéfice par trajet = (Bénéfices totaux sur une certaine période de temps) / (Nº de trajets sur cette période de temps). Graphiquement on a obtenu un résultat suivant:



On remarque donc que les deux sociétés présentent le même comportement, leur efficacité tend à diminuer avec le temps.

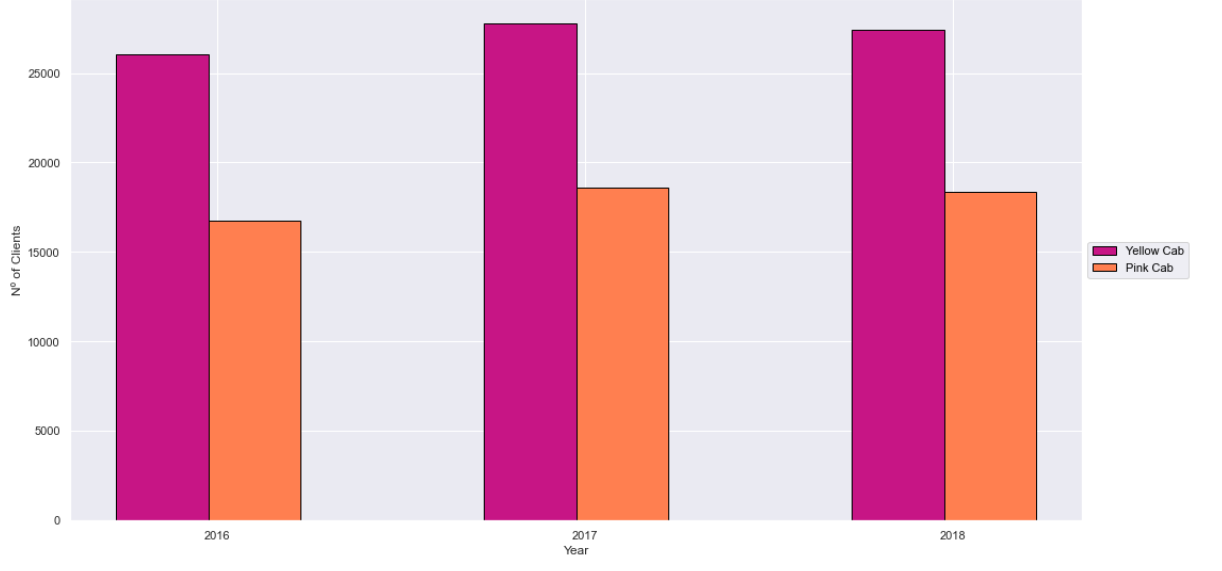
Un autre indicateur de mesure serait le bénéfice net réalisé par chacune des entreprises dans chaque ville. Le graphique ci-dessous illustre les résultats obtenus.



On remarque que dans les mêmes villes, Yellow a un bénéfice beaucoup plus élevé respectivement à son concurrent direct.

Allons maintenant regarder la fidélisation des clients de chaque entreprise par année. Pour ce faire, on peut utiliser un code qui regroupe l’ID de chaque client et additionne son nombre de voyages sur une année. En faisant ainsi on aura une idée de la compagnie la plus attractive et donc à qui les clients font le plus recours.

Clients ayant plus de 10 voyages par ans/ entreprise



**VI. CONCLUSION**

Nous avons évalué les deux compagnies de taxis sur les points suivants :

1. Analyse des bénéfices

Bénéfices : des bénéfices plus élevés au fil du temps et moins de fluctuations mensuelles

Bénéfices sur les trajets : meilleur taux de profit / trajet dans le temps. Yellow Cab a 2,5 bénéfices supplémentaires par rapport au trajet

Bénéfices Citywise : Yellow Cab a une plus grande part de marché dans chaque ville.

Manèges rentables : en supposant que sur un taux de 80 %, l'entreprise fonctionne bien. Yellow Cab a une haute performance parmi toutes les villes. En revanche,

Pink Cab n'est pas performant dans 8 villes.

2. Analyse de la demande

Demande : Yellow Cab a plus que triplé la demande de Pink Cab Company

3. Analyse des clients

Taux de fidélité : En supposant que les deux classes, Yellow Cab a un taux de fidélité élevé et un taux de fidélité moyen plus élevé.

Répartition des modes de paiement : les deux sociétés présentent la même répartition des modes de paiement dans le temps, par ville et par âge.

Donc sur la base des points ci-dessus, nous recommandons formellement **Yellow** pour un investissement qui serait rentable.