Master Informatique — SID

Big Data TP

**LABIAD Soukaina**

**Découverte et mise en œuvre de Spark**

## **Collection de données**

Après téléchargement des données à partir du lien : <https://s3.amazonaws.com/baywheels-data/201802-fordgobike-tripdata.csv.zip>, nous avons crée la table "***201802\_fordgobike\_tripdata***" dans notre notebook.

Pour bien commencer le traitement, nous avons changé le schéma de données pour que les colonnes ***duration\_sec***, ***start\_time*** et ***end\_time*** soient du type Timestamp.

## **Évaluation des données**

Dans la base de données choisie (février 2018), nous n'avons pas de valeurs nulles, donc pas besoin de faire une suppression.

Après évaluation des données, on remarque des problèmes de propreté:

* Des colonnes individuelles sont manquantes pour les valeurs d'heure, le nom du jour de la semaine, le nom du mois et la durée en minutes.
* Ensemble de données non classées par date.
* Des colonnes qui ne seront pas utilisées.

## **Nettoyage des données**

Nous avons commencé par un premier nettoyage, qui consiste à supprimer les lignes contenants des valeurs aberrantes pour l'attribut ***duration\_sec*** et les lignes qui dépassent 98%.

Après ce nettoyage des données, on remarque qu’on a toujours des problèmes de propreté.

Nous avons besoin de créer des nouvelles colonnes : ***start\_time\_hour***, ***end\_time\_hour***, ***time\_day\_of\_week***, ***time\_month*** et ***duration\_minut*** pour remédier au problème des colonnes manquantes pour les valeurs d'heure, le nom du jour de la semaine, le nom du mois et la durée en minutes.

Pour l’ensemble de données non classées par date, nous procéderons par un triage suivant la colonne ***start\_time***

Dans le même cadre de nettoyage de données, on trouve que les colonnes suivantes ne seront pas utilisées: ***start\_station\_latitude, start\_station\_longitude, end\_station\_latitude, end\_station\_longitude*** et ***bike\_id***. Elles sont donc à supprimer.

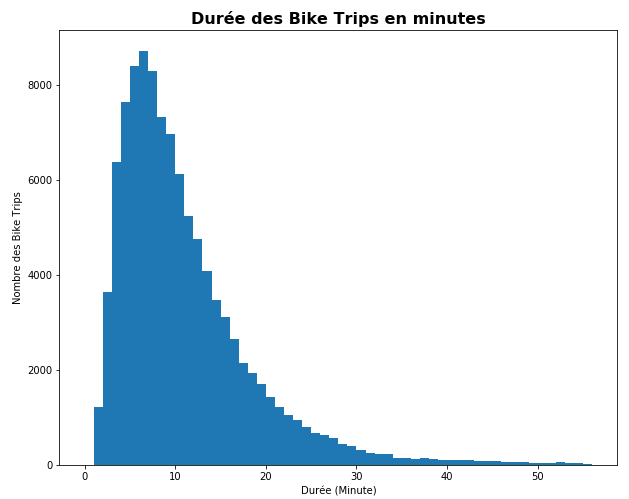
Après évaluation et nettoyage, on se trouve avec une base de données finale propre et valide.

## **Analyse et visualisation des données**

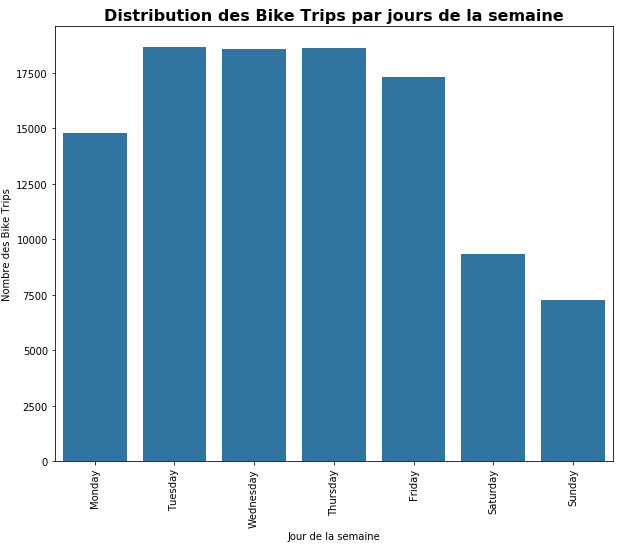
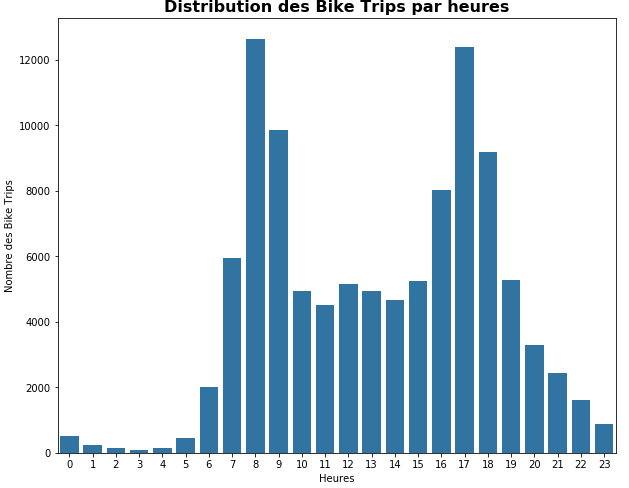
Dans cette partie, nous allons explorer la durée et la quantité des trajets à vélo par rapport aux caractéristiques de l'ensemble de données. Les fonctionnalités qui nous aideront à étayer cette enquête sont la ***duration\_minute***, ***start\_time***, ***user\_type***, ***start\_time\_hour***, ***time\_day\_of\_week*** et ***time\_month***.

1. **Exploration Univariante**

Commençons par l’affichage de la distribution de la variable principale ***duration\_minute***

******

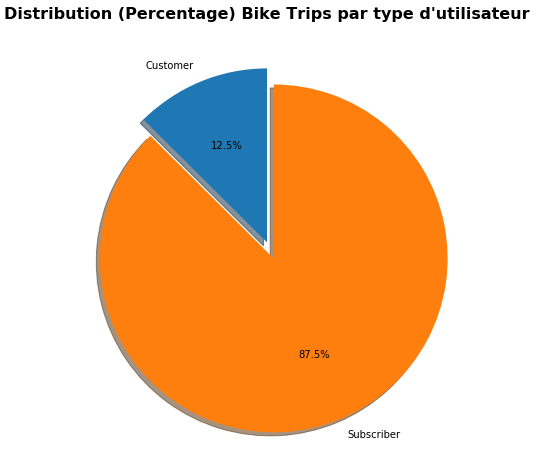
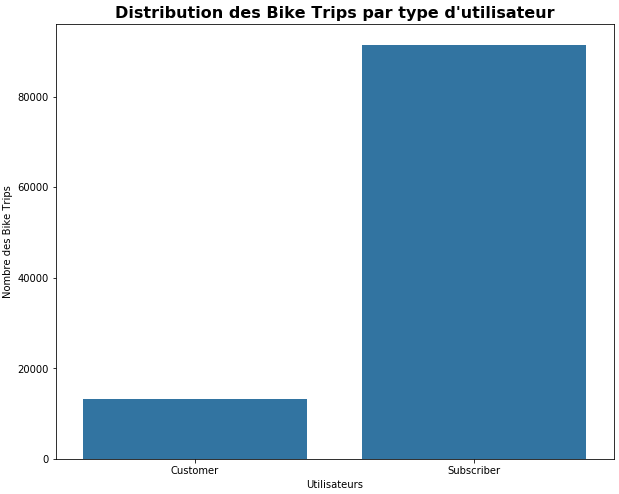
Nous observons qu'elle formait un graphique à asymétrie positive, où la majorité des trajets ont une durée comprise entre 5 et 10 minutes.

Affichons maintenant la distribution des fonctionnalités ***datetime***

Lorsque nous regardons la distribution des balades à vélo par heure, nous voyons un pic de déplacement entre 7 et 10 et un autre pic entre 16 et 20;

Quand on regarde la même variable par jour de la semaine, on remarque que du lundi au vendredi nous avons la plupart des trajets avec des pics entre le mardi et le jeudi et un nombre minimal le week-end.

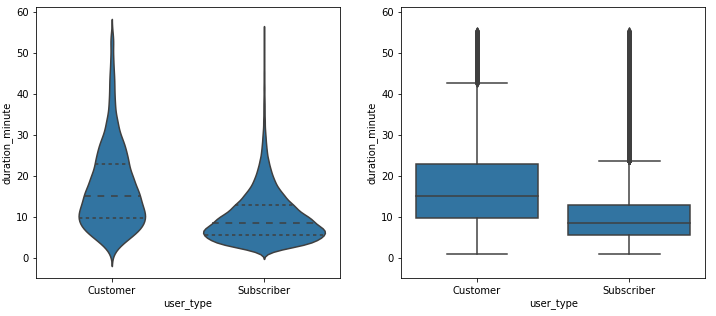
Notons que ces trajets sont faits principalement dans les heures de déplacement pour partir ou revenir du travail.

Passons à la distribution de la fonctionnalité ***utilisateur***

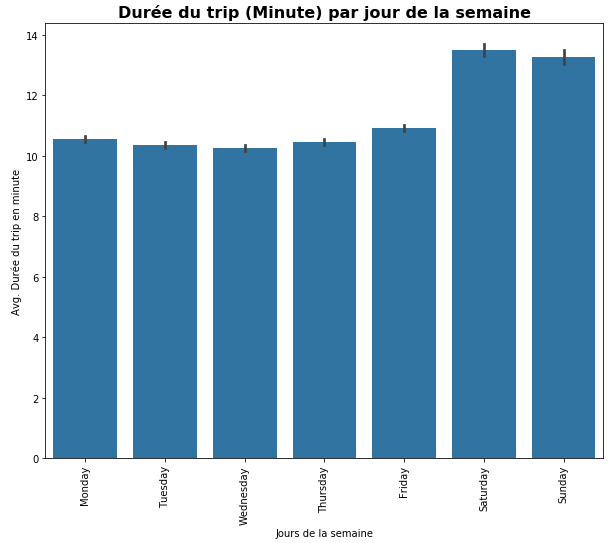
On voit ici que plus de 87,5% des utilisateurs qui utilisent les services sont des abonnés.

1. **Exploration Bivariante**

Commençons par comparer ***la durée des trajets*** par les autres fonctionnalités.

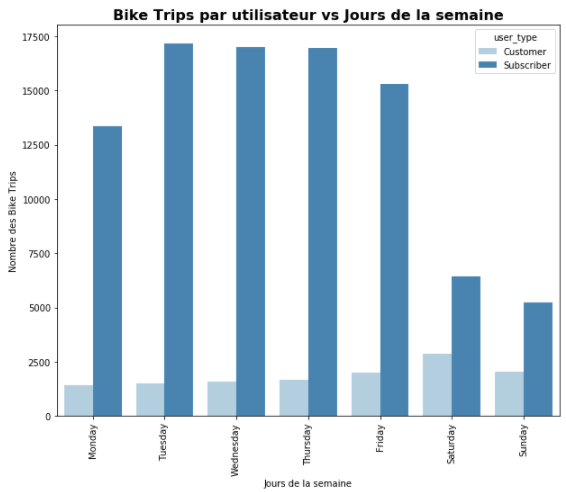


Nous voyons que les utilisateurs "***Customer***" ont tendance à effectuer des trajets plus longs que les utilisateurs "***Subscriber***".

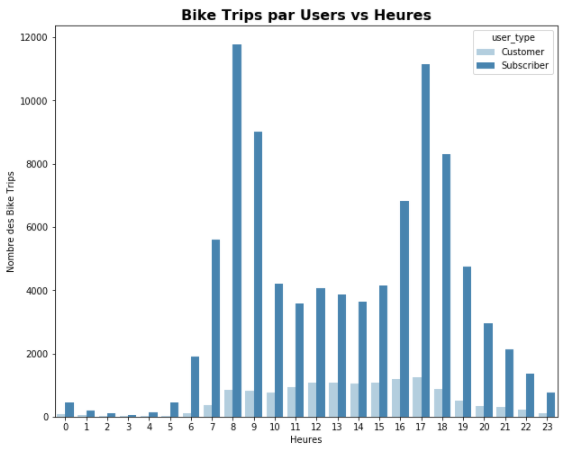


Nous observons que pendant les week-ends, les gens ont tendance à faire des trajets plus longs.

Faisons maintenant une comparaison de ***user\_type*** par les autres fonctionnalités



Ici, nous constatons une forte baisse du nombre d'abonnés le week-end, tandis que l’utilisateur client a moins d'impact.



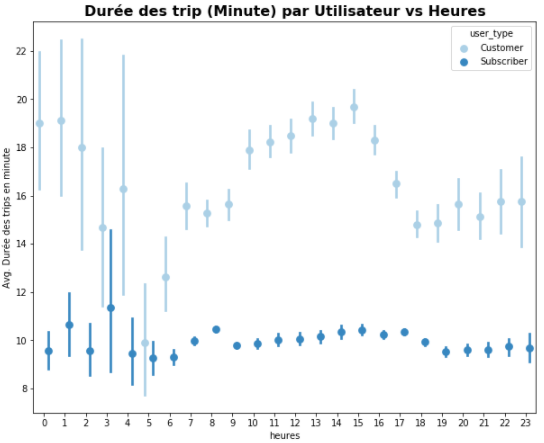
Là encore, on remarque la tendance à l'augmentation des déplacements aux heures de 7 à 10 et de 16 à 20, principalement pour les utilisateurs « Subscriber ».

Les utilisateurs « Customer » suivent une tendance plus discrète avec plus de déplacements dans l'après-midi.

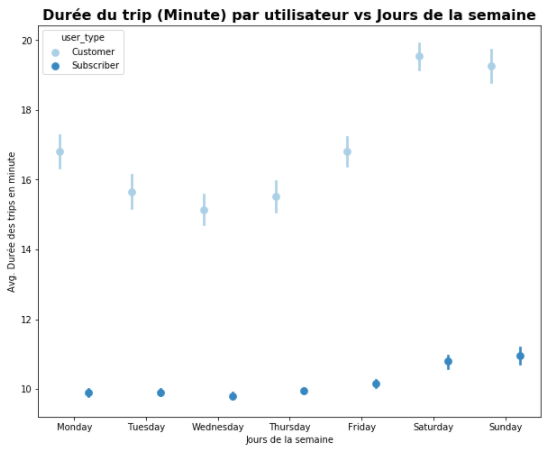
Pour conclure, l’habitude des déplacements varie beaucoup entre "Subscriber" et "Customer". Le “Subscriber” utilise le service de vélos pendant les heures de travail, de sorte que la plupart des déplacements ont eu lieu en semaine (du lundi au vendredi) et surtout aux heures de pointe (lorsqu’on travaille le matin et part l'après-midi), tandis que les “Customer” s'amusent généralement dans l'après-midi ou le week-end.

1. **Exploration Multivariante**

Dans cette partie nous comparons ***la durée moyenne du trajet*** en minute ***le user\_type*** par les autres fonctionnalités



Nous avons remarqué que la durée moyenne du trajet horaire, met le “Customer” avec une augmentation du temps pendant les heures de l'après-midi, tandis que le “Subscriber” a peu de variation pendant la journée des heures ouvrables.



Nous avons également observé que le “Customer” passait plus de temps en déplacements durant le week-end, tandis que le “Subscriber” a peu de volatilité pour le week-end.