

GLO-4027 - Bike Sharing Demand

Planification

30 janvier 2019

Étudiants du 2ème cycle :

Bastien CHABAL

Corentin GIRAUD

[Projet Kaggle](https://www.kaggle.com/c/bike-sharing-demand#description) – [Repository GitHub](https://github.com/bastienchabal/GLO-4027)

# Les données

Le projet que nous avons choisi est intitulé **Bike Sharing Demand** et proposé par Kaggle.

Le système de partage de vélos au sein d’une ville est très simple : plusieurs kiosques sont répartis dans la ville et des personnes (abonnés ou non) peuvent louer un vélo et faire le trajet qu’elle souhaite jusqu’à un autre kiosque.

Les données sont réparties comme ceci :

* Le **set d’entrainement** contient deux ans de données, où sont relevés toutes les heures différentes informations, dont le nombre total de locations. Ce set ne contient que les données des **19 premiers jours** **de chaque mois**.
* Le **set de test** contient des données exactement similaires au set d’entraînement, mais pour **tous les jours après le 20 du mois inclus**.

# L’objectif

L’objectif ici est **de prédire le nombre de total de vélos loués à chaque heure** pour les périodes couvertes par le set de test, en utilisant seulement les informations du set d’entraînement.

La prédiction sera évaluée grâce au **Root Mean Squared Logarithmic Error** (RMSLE).

# Première approche

Après une première visualisation des données proposées par le set d’entraînement, on constate qu’il contient 10 886 points de données, répartis sur 12 attributs.

Le nombre d’attributs est conséquent, surtout que notre prédiction ne devra se faire que sur le nombre de location. La première analyse à faire serait donc de « trier » ses attributs par importance. Pour comprendre les relations qui les relient, on génère la **matrice de corrélation**. On s’intéresse pour l’instant principalement à la ligne « count » pour voir quels attributs corrèlent avec le nombre de location.



On constate d’abord que les attributs qui corrèlent le plus avec le *bike count* sont en fait des attributs sous-jacents à ce dernier : les locations **occasionnels** (0.69) et **enregistrés** (0.97), dont la somme donne le nombre de locations total.

Avec une corrélation de 0.97, l’attribut *registered* va se comporter de la même manière que l’attribut *bike count*. Ainsi, au lieu d’essayer de prédire directement le bike count, une première approche intéressante serait d’abord d’analyser les comportements de *registered* et *casual* séparément. Si ceux-ci s’avèrent différents, on découpe notre problème de base en deux sous-problèmes, amenant à deux prédictions plus précises dont on fera la somme.

# Observations

On repère déjà une différence entre *registered* et *casual* sur leur corrélation avec *workingday* (négative sur *casual*).

Les deux sont corrélés sur les températures et températures ressenties, on émet donc la conjecture que les journées chaudes sont propices à la location.



Les deux sont corrélés négativement avec l’humidité.

Enfin, il faudra s’intéresser l’évolution du nombre de location au sein d’une journée, pour voir si certaines heures sont plus propices à la location.