**PROJET DATA SCIENCE – Divvy Bikes (Chicago)**

**Données :**

Fichiers de trajets et de stations des «vélib » pour la période 01/03/2018 – 31/07/2018 extraits avec le code « 05 »

**Objectif du projet :**

Analyser l’utilisation des Divvy Bikes et illustrer les résultats obtenus

**Exigences du CEPE :**

1. Démontrer que nous sommes capables d’appliquer les algorithmes vus en cours
2. Faire une visualisation des résultats (shiny)

**Etudes à mener :**

*Partie à modifier ou enrichir si vous avez des meilleures idées !*

1. **Classification des stations en fonction de leur utilisation**

Fichier à utiliser : Stations

Description : Identifier les clusters des stations et voir s’ils correspondent p.ex. aux quartiers résidentiels, centres des affaires, zones touristiques etc.

Méthodes à utiliser : K-means, CAH, autres ?

Calculs à réaliser :

*Etape 1 :* Pour chaque ligne du fichier des stations calculer le taux d’utilisation de la station

Taux d’utilisation = Nombre de vélos disponibles / Nombre de bornes en service

*Etape 2* : Construire un fichier avec 1 ligne par station et les taux d’utilisation à des différents instants en colonnes

*Prérequis pour cette étape*: s’assurer que les relevés sont réalisés au même temps pour l’ensemble de stations et à un intervalle de temps constant.

*Etape 3* : Tester !!!

Pour commencer :

* Choisir un jour ouvré, p.ex. le mardi 04/07/2017 et comme variable explicative (en colonne) conserver le taux d’utilisation de la station toutes les 30 minutes (48 colonnes),
* Tester la méthode K-means avec les différents nombres de clusters et choisir un nombre de clusters qui donne des résultats satisfaisants,
* Tester la méthode CAH.

Une fois ce premier exercice est fait :

* Voir ce qui se passe si on garde dans les données sources les taux d’utilisation des stations toutes les 10 minutes ou toutes les heures => le résultat change-t-il ?
* Essayer de réaliser le clustering pour un dimanche (p.ex. 09/07/2017). Les clusters sont-ils très différents par rapport à ceux obtenus pour le jour ouvré ?
* Voir ce qui se passe si au lieu de travailler sur une seule journée, on met comme variables explicatives les taux d’utilisation moyens de l’ensemble de jours ouvrés (p.ex. taux d’utilisation moyen à 10h, à 10h30 etc).
* …

1. **Etude de « migrations » de la population cyclistes via clustering de trajets**

Fichier à utiliser : Trajets

Description : Regrouper des trajets « similaires » en termes de localisation des stations de départ et d’arrivée pour pouvoir ensuite visualiser les principales directions de déplacements de la population

Méthode à utiliser : méthode basée sur la densité (dbscan), car cette méthode permet d’exclure les trajets atypiques

Calculs à réaliser : **Magda : Plus de détails jeudi, je suis dessus et j’ai des idées. Il faut jouer avec des coordonnées géographiques ;)**

1. **Prévoir l’utilisation du réseau Divvy Bike en fonction des variables explicatives**

Fichiers à utiliser : Trajets, Météo de plages, autres ?

Description : Prévoir le nombre de trajets réalisés dans la journée en fonction du jour de la semaine, des informations météorologiques et éventuellement d’autres variables explicatives

Méthodes à utiliser : Régression multiple avec choix de variables, Ridge, Lasso, Elastic net, Forêts aléatoires, autres ? + choix du modèle avec la validation croisée

Calculs à réaliser :

*Etape 1 :* A partir du fichier des Trajets, pour chaque jour de la période étudiée calculer le nombre de trajets

et récupérer le jour de la semaine, par exemple :

|  |  |  |
| --- | --- | --- |
| **Jour** | **Jour de la semaine** | **Nombre de trajets** |
| 01/03/2017 | Wednesday | 13 456 |
| 02/03/2017 | Thursday | 15 222 |
| … |  |  |

*Etape 2 :* A partir du fichier de Météo de plages récupérer les variables explicatives, p.ex.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Jour** | **Température minimale** | **Température maximale** | **Indicateurs liés au vent** | **Indicateurs liés à la pluie** | **…** |
| 01/03/2017 | 3°C | 6°C |  |  |  |
| 02/03/2017 | 7°C | 10°C |  |  |  |
| … |  |  |  |  |  |

*Etape 3 :* Croiser les deux fichiers en vérifiant si nous disposons des informations météo pour l’ensemble de jours

*Etape 4 :* Appliquer les modèles, voir lequel marche le mieux et quelle est la qualité de l’ajustement. Si nécessaire, réfléchir sur d’autres variables explicatives.