Université de Bordeaux
M2 MAS - Parcours MSS
UE : Projet Données Massives
Année 2017-2018

Liste des projets Open Data

Le but de ce projet informatique est de traiter des données libres en utilisant les méthodes de statistique descriptive, de représentations graphiques, et d'inférence vues tout au long de votre formation. Ces données libres (Open data) sont accessibles directement sur Internet, éventuellement au travers de packages *R*.

Dans ce document, différents projets sont proposés qui visent à analyser des données en libre accès. Ce projet informatique est très libre, en particulier toutes les initiatives personnelles et les approches innovantes seront fortement appréciées. Toutefois, pour réaliser les traitements statistiques de vos données, il est demandé de travailler à partir du logiciel *R*.

Dans la liste de projets ci-dessous les questions proposées ne sont que des suggestions et n'ont pas vocation à être exhaustives. De ce fait, vous êtes encouragés à croiser les données d'un projet avec d'autres données libres accessibles sur Internet et qui vous sembleraient pertinentes. Pour enrichir votre projet, vous pourriez juger intéressant d'utiliser les données des sites décrits sur le site web du projet : `https://www.math.u-bordeaux.fr/~arichou/site/references/index.html` ou tout autre jeu de données ouvertes disponible sur internet.

1 Description des projets

1.1 Twitter

Le package *twitteR* est une interface *R* qui permet d'accéder à des données sur le site de microblogage Twitter : `https://twitter.com`. Ces données sont constituées de mots et de phrases déposées quotidiennement par des millions d'utilisateurs dans le monde, ainsi que de méta-données sur les utilisateurs de ce réseau social. Le package est disponible sur le CRAN. Quelques exemples d'analyse statistique de données sur Twitter sont disponibles aux adresses :

- analyse textuel de mots dans des tweets :
  `https://sites.google.com/site/miningtwitter/home`
- analyse de l'état émotionnel d'un utilisateur de Twitter :
  `http://thinktostart.com/create-twitter-sentiment-word-cloud-in-r/`

Traitement des données : il pourra être intéressant dans un premier temps de reprendre certains exemples d'analyse de données proposées dans les sites web listés ci-dessus puis de les enrichir et/ou de s'intéresser à des problématiques proches.

1.2 Taux d'abstention aux élections en France

Les fichiers disponibles aux adresses suivantes :

https://www.data.gouv.fr/fr/datasets/elections-legislatives-1958-2012/
https://www.data.gouv.fr/fr/datasets/elections-regionales-1986-2010/
https://www.data.gouv.fr/fr/datasets/elections-cantonales-1988-2011/

contiennent des données sur les résultats des élections législatives, régionales et municipales en France entre 1958 et 2012. D'autre part, on dispose d'indicateurs socio-économiques et démographiques sur les communes françaises sur le site de l'INSEE à cette adresse :

https://www.insee.fr/fr/statistiques

Dans ce projet, on cherche à comprendre quelles sont les données socio-économiques et démographiques à l'échelle des communes françaises qui permettent d'expliquer l'abstention aux élections en France.

Traitement des données : il pourra être intéressant dans un premier temps de reprendre certains exemples d'analyse de données sur ce thème proposés par les étudiants des années précédentes, puis de les enrichir.

1.3 Tribunaux de commerce

Les tribunaux de commerces mettent à disposition de nombreuses données sur le site :
https://data.infogreffe.fr/explore/?sort=modified
On y trouve en particulier des données concernant les créations et les radiations d'entreprises en France.

Traitement des données : il pourra être intéressant dans un premier temps de réaliser une présentation des données de créations et de radiations d'entreprises à l'aide d'outils de statistique descriptive. Puis dans un second temps, vous pourriez essayer de croiser ces données avec d'autres données socio-économiques, étudier les données d'un point de vu géographique, vous intéresser à l'évolution temporelle ou bien encore réaliser une étude statistique sur les données textuelles que représentent les noms d'entreprises.

1.4 Bordeaux

A l'instar de nombreuses villes françaises, Bordeaux met à disposition de nombreuses données ouvertes concernant ses équipements, ses habitants et ses transports :
http://opendata.bordeaux.fr/

Traitement des données : il pourra être intéressant dans un premier temps de se focaliser sur une catégorie de données et de les présenter à l'aide d'outils de statistique descriptive. Puis dans un second temps, vous pourriez essayer de croiser ces données avec des données socio-économiques des habitants de Bordeaux, ou bien encore confronter ces informations aux données d'autres grandes villes françaises. Vous pouvez étendre les travaux déjà effectués en 2016-2017 par des étudiants du Master MAS-MSS.

1.5 Open Data de l'Assurance Maladie

L'assurance maladie a mis en place un site de données en ligne sur différents thèmes tels que les dépenses d'assurance maladie ou les prescriptions de médicaments délivrés en officine de ville. Ces données sont disponibles à cette adresse :
`http://open-data-assurance-maladie.ameli.fr`

et des exemples d'études statistiques à partir de ces données sont présentés ici :
`http://www.lemonde.fr/les-decodeurs/article/2017/11/28/medicaments-et-remboursements-la-base-de-donnees-open-medic-en-6-points_5221378_4355770.html`

Traitement des données : il pourra être intéressant de reproduire dans un premier temps les exemples d'analyse statistique présentés dans l'article ci-dessus (identification des médicaments les plus coûteux, ou bien des niveaux de remboursement par habitant et par région). Dans un second temps, il serait également pertinent de coupler ces données avec des variables explicatives de type socio-économiques disponibles sur le site de l'INSEE.

1.6 Open Data Réseaux Energies

Le site `https://opendata.reseaux-energies.fr/explore/?sort=modified` permet la mise à disposition de données autour des thématiques telles que la production, la consommation et le stockage de différentes sources d'énergie dans les territoires et régions de France.

Traitement des données : dans un premier temps il pourra être intéressant de faire une analyse descriptive de ces données, et de croiser des informations telles que les courbes de charge consommation brute régionale avec la température régionale pour analyser d'éventuelles corrélations. Dans un second temps, une étude dynamique de l'évolution temporelle de la consommation d'énergie par région en France au cours de l'année pourrait être intéressant et une recherche de variables explicatives (météo régionale) de la variation de la consommation serait pertinent.

2 Travail à effectuer

Après avoir choisi l'un des projets décrits ci-dessus, le travail à effectuer se décompose en deux parties :

- Vous devrez réaliser un compte-rendu de vos travaux sous la forme d'un site web sur Internet qui expliquera votre démarche dans l'analyse de ces données. Vous devrez fournir votre code *R* correctement commenté, et proposer des outils de visualisation adéquats pour présenter vos résultats (via Rmarkdown, Shiny).
- Vous ferez une présentation de 20 minutes présentant une partie de vos travaux suivie de 10 minutes de questions