On vous demande de produire un rapport. Soignez la rigueur, la présentation, la clarté de votre document. On préférera la qualité à la quantité.

1 Partie commune à tout le groupe

Vous devez d'abord choisir un jeu de données. Vous trouverez au dos de ce document une liste d'idées de sites où le chercher mais vous êtes tout à fait libre de le chercher ailleurs. Ce jeu de données doit d'une part vous intéresser au sens où vous êtes curieux de pouvoir essayer d'en tirer de l'information, et d'autre part être suffisamment riche pour qu'il soit intéressant de lui appliquer des méthodes statistiques d'analyse de données. Cela signifie notamment qu'il doit comporter suffisamment de variables (au moins 6 ou 7 mais ça peut éventuellement être beaucoup plus), éventuellement de natures différentes, et suffisamment d'observations (au moins quelques dizaines). Si vous souhaitez quand même étudier un jeu de données qui ne respecte pas ces conditions, vous devez absolument me le soumettre d'abord en motivant votre choix.

L'objet du projet est d'étudier ce jeu de données, ce qui inclut nécessairement au moins les étapes suivantes (pas nécessairement dans cet ordre exact) :

- Prise de contact avec le jeu de données (contexte de la collecte des données, modalités de collecte, signification des variables...)
- Importation du jeu de données et éventuellement « nettoyage » de ces données. Notamment, il conviendra de vérifier si le jeu de données comporte des valeurs manquantes. On préférera à ce sujet un jeu de données avec suffisamment d'individus pour lesquels toutes les variables ont été observées pour pouvoir tout simplement supprimer les individus pour lesquels il y a des observations manquantes (il existe certes des approches plus subtiles pour traiter les données manquantes).
- Étude descriptive du jeu de données d'un point de vue statistique (nombre et nature des variables, quelques indicateurs statistiques sur ces données : position, dispersion...).
- Définition d'un enjeu pour l'étude statistique (explorer le jeu de données et chercher à mettre en évidence des informations intéressantes ; chercher quelles variables sont liées ou semblent jouer un rôle notable...).
- Choix des méthodes et mise en œuvre de l'analyse statistique.
- Analyse des résultats.

Vous préparerez un compte-rendu de ce travail. Celui-ci ne devra pas dépasser 10 pages (donc 5 recto-verso maximum) : une des difficultés de ce travail est bien sûr de choisir de manière pertinente les résultats et graphiques que vous présenterez, qui seront nécessairement une petite partie des résultats obtenus.

A priori, ce compte-rendu ne devrait pas contenir de code. En revanche, vous rendrez aussi un fichier R (ou Rmarkdown) contenant l'ensemble du code utilisé, commenté proprement, qui permettra au correcteur de reproduire l'ensemble des résultats que vous aurez obtenus. Par exemple, dans la troisième étape décrite ci-dessus, vous devez bien sûr vous pencher sur chacune des variables étudiées. Tous les résultats obtenus à ce sujet (surtout s'il y a beaucoup de variables) n'auront probablement pas leur place dans le compte-rendu : à vous de choisir d'éventuellement rendre compte d'une partie d'entre eux. En revanche, le code utilisé pour chacune des variables doit être disponible dans le fichier R.

2 Partie individuelle du travail

La partie individuelle du travail sera constituée de votre compte-rendu du TP 1. Comme annoncé en TP, celui-ci doit être préparé individuellement.

3 Rapport

Votre rapport devra donc consister en un fichier pdf incluant le compte-rendu sur le projet d'une part, et le compte-rendu de TP d'autre part. La date limite pour rendre ce rapport est le 19 janvier 2020 à minuit (sauf information contraire avant le 21 décembre 2019). Les modalités de rendu du rapport et du fichier R seront précisées d'ici là.

4 Idées de sources de données...

https://archive.ics.uci.edu/ml/index.php

https://challengedata.ens.fr

http://lib.stat.cmu.edu/datasets/

http://www.amstat.org/publications/jse/jse_data_archive.htm

http://www.rdatamining.com/resources/data

http://pbil.univ-lyon1.fr/R/enseignement.html (menu données)

https://www.kaggle.com/

http://www.kdnuggets.com/datasets/

http://www.data.gouv.fr/

http://stats.oecd.org/

http://data.worldbank.org/

https://mran.microsoft.com/applications/

The End.