Projet Analyse Statistique de données

Le projet se prépare en binôme.

Données

La recherche d'un jeu de données fait partie du projet. Trouver un jeu de données intéressant peut prendre du temps, ne négligez pas cette part du travail. Le jeu de données doit se prêter aux techniques d'analyses multidimensionnelles étudiées en cours et en tps (cours des semestres 1 et 2). Il doit donc comporter plusieurs variables et suffisamment d'individus. Pour appliquer les différentes techniques vues en cours, choisissez un jeu de données avec plusieurs variables quantitatives qui ne seront pas trop corrélées entre elles, et au moins une variable qualitative (si vous n'avez pas de variable qualitative, il est possible d'en créer une à partir d'une variable quantitative. A vous de voir si c'est pertinent). Attention aussi à choisir suffisamment d'individus, sans choisir un trop gros jeu de données.

- Vous pouvez chercher un jeu de données sur l'une des plateformes suivantes :

  `https://github.com/`

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

  `http://www.inside-r.org/howto/finding-data-internet`

  `http://archive.ics.uci.edu/ml/datasets.html`

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

  `https://www.kaggle.com/`

  `http://www.databasesports.com/`

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

  `http://www.umass.edu/statdata/statdata/data/index.html`

  `http://apcentral.collegeboard.com/apc/members/courses/teachers_corner/22027.html`

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

  `http://stats.oecd.org/`

  `http://data.worldbank.org/`

- Vous pouvez me proposer un jeu de données provenant d'une autre source.

Proposition de projet

Vous devez m'envoyer par mail votre proposition de projet sous la forme d'un pdf d'une dizaine de lignes. Votre proposition de projet présentera le jeu de données choisi, le lien du site où vous l'avez trouvé, et les méthodes que vous comptez mettre en oeuvre. Ce document me permettra de valider ou non votre choix. Vous me rendrez cette proposition de projet sur Moodle pour mi-avril.

Consignes générales pour la préparation du projet

Le projet consiste en la rédaction d'un rapport et une présentation orale. Le rapport doit être rédigé sous la forme d'un document .pdf comportant :

- Une synthèse de maximum 15 feuilles (discussion, extraits de sorties R, graphiques)
- Des annexes structurées comportant les codes et les extraits de sorties R que vous estimez utiles.

La synthèse comportera notamment :

1. Une présentation du jeu de données et des problématiques que celui-ci soulève.
2. Quelques lignes sur l'importation du jeu de données (difficultés éventuellement rencontrées, renvoyez au code en annexe s'il y a lieu).
3. Les analyses statistiques : pour chaque point étudié, vous motiverez celui-ci par quelques phrases, puis vous exposerez vos conclusions en illustrant celles-ci par des graphiques et éventuellement de courts extraits de sorties R. Vous donnerez en annexe les codes R et les sorties R discutées. Il vous est demandé de mettre en oeuvre un maximum de méthodes adéquates sur le jeu de données choisi.

   Les méthodes vues au premier semestre (statistique descriptive, regression linéaire, anova) sont des méthodes à essayer si le jeu de données le permet.

   N'hésitez pas à mentionner les méthodes que vous auriez pu essayer, même si celles-ci n'ont pas donné de résultats intéressants ou facilement interprétables. Vous pouvez aussi (si vous le souhaitez) apprendre de nouvelles méthodes d'analyse de données à cette occasion mais ceci ne devra constituer qu'une partie réduite du projet et ne sera compté que comme du bonus !
4. Vos principales conclusions à l'issue de l'analyse.

Le rapport .pdf sera à rendre à la fin du semestre et l'exposé aura lieu quelques jours plus tard. L'exposé doit être calibré pour durer 10 minutes + 5 minutes de questions. Les questions pendant l'exposé porteront sur votre étude mais pourront aussi porter sur les notions vues en cours et en TPs.

Bon projet !