M2Info- 2017 - Traitement de masses de données

1 Données

Dans le cadre de la mouvance opendata, de nombreuses bases de données sont librement consultables en ligne. Sélectionnez un ensemble de données et une problématique associée, et présentez vos conclusions dans un court rapport. Vous veillerez à questionner la précision des données et leur capacité à supporter votre investigation (nécessité de certaines hypothèses additionnelles, taille de sampling,...). Vous pouvez travailler en groupes (3 max).

Les données doivent être disponibles en ligne ou fournies, idéalement au format csv. Elles doivent comporter au minimum 1000 observations. Voici quelques exemples :

Egalité des sexes Etudiez l'impact du genre sur le salaire des employés de la ville de San-Francisco.
Données : `http://transparentcalifornia.com/salaries/san-francisco/`

Le retour du prénom à la mode On dit que les modes des prénoms sont cycliques. Est-ce vrai ? Exemple en utilisant les prénoms américains
Données : `https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-natior`

Succès Kickstarter Que peut-on dire sur les projets Kickstarter qui ont été financés avec succès ?
Données : `http://datapolymath.paperplane.io/`

Quelques sources d'inspiration :

Reddit/dataIsBeautiful compilation d'analyses exploratoires très vaste.

Kaggle competitions Compétitions de machine learning

2 Rendu

Vous rendrez **une archive** tar compressée avec gzip (.tgz) contenant
— un document **pdf** de **4** pages maximum, contenant au minimum 5 graphiques **commentés**.
— le(les) scripts R **commentés** réalisant les graphiques du rapport.
— les données ou un lien vers les données (accessible directement au correcteur).

L'envoi sera effectué par mail a `gtredan@laas.fr` au plus tard le 8 Décembre 2017 à minuit. Le sujet du mail sera "[M2INFO] rapport NOM1 NOM2..." ou NOMi est le nom de famille du $i$ème membre du binôme.

3 Barème

— Respect du format de rendu : 3 points
— Qualité du rapport : 7 points
— Qualité du code R : 5 points
— Profondeur du traitement : 5 points

Voici le tarif des manquements suivants :
— Pénalité de retard : 1 point par tranche de 8 heures après la deadline
— Utilisation de format non-ouverts : Note/2
— Plagiat, sources non citées : tarif variable selon gravité de l'infraction
— Orthographe et grammaire : tarif variable selon gravité de l'infraction