Modélisation sur des données « Open Data »

On vous demande de réaliser une étude sur des données « Open Data » accessibles en ligne (ex. https://www.data.gouv.fr/fr/ ; http://open-data-assurance-maladie.ameli.fr/ ; etc. – Les bases qui servent pour les challenges de type Kaggle sont proscrites).

1. Vous avez le choix des données à exploiter. Qu'importe le thème et le sujet, vous devez avant tout développer une problématique de modélisation linéaire en précisant la variable cible à expliquer, les variables explicatives potentielles, les attentes en termes d'interprétation que vous pouvez attendre dans cette étude.
2. Les données doivent être parfaitement sourcées c.-à-d. je dois pouvoir les retrouver facilement en ligne à partir de vos indications.
3. Vous avez le choix des outils. Vous devez simplement m'envoyer tous les éléments qui me permettront de reproduire vos calculs (code R ou Python, fichiers Excel, etc.).
4. Il y aura certainement une régression linéaire multiple à réaliser (j'imagine, j'espère en tous les cas). Après, il vous appartient de déterminer les analyses supplémentaires à utiliser pour corroborer votre discours et valider votre travail (ex. étude des résidus, analyse des points atypiques, etc.). Le plus important est de justifier l'usage de telle ou telle technique, la lecture des résultats que vous en faites, et les préconisations que vous en tirez (ex. vous décidez de supprimer tel point parce qu'il exagérément influent dans la régression ; sélection de variables ; etc.).
5. Vous rédigerez un rapport d'une douzaine de pages présentant votre étude, votre démarche et vos résultats. Adoptez une démarche standardisée en vous calant sur le modèle CRISP-

DM par exemple (cf. https://www.the-modeling-agency.com/crisp-dm.pdf ; voir en particulier la Section IV « The CRISP-DM outputs »)[1]. Les chapitres types seraient :

(a) Présentation du problème et des objectifs de l'étude ;

(b) Présentation des données, comment et de quelle manière pourront-elles répondre au cahier des charges défini en (a), description rapide des données (nombre d'observations, de variables, présence ou non de données manquantes, statistiques descriptives, etc.) ;

(c) Préparation des données, les éventuels recodages ou normalisations et autres ;

(d) Modélisation c.-à-d. définition et estimation des paramètres du modèle explicatif / prédictif à partir des données, présentation du modèle ;

(e) Evaluation, diagnostic et vérification de la qualité de votre modèle, affinage de votre modèle ;

(f) Interprétation c.-à-d. quels phénomènes de causalité montrent vos résultats, quelles sont les variables explicatives les plus déterminantes, comment pèsent-elles sur la variable à expliquer.

(g) Bilan critique de votre étude. Points forts, points faibles, portée de votre travail, conclusion.

Vous pouvez adapter la structure de votre document au regard des spécificités de votre étude bien sûr. Essayez d'adopter un discours varié mêlant texte, tableaux et graphiques.

Le travail est à réaliser en équipes de 3 étudiants.

[1] Ce document du site « WIKISTAT » devrait vous aider également : « Pour rédiger un rapport de statistique ».

A RENDRE

A. Un rapport au format PDF décrivant l'étude (**une douzaine de pages**). Indiquez clairement sur la page de garde les noms des étudiants qui y ont contribué.

B. Le matériel ayant permis de le réaliser (code source, fichiers de travail, etc.). Je dois pouvoir reproduire votre travail et obtenir exactement les résultats indiqués dans votre rapport !

C. Zippez le tout dans un fichier archive portant les noms des étudiants et le numéro de groupe de TD (**Groupe_TD_numero_GRIEZMAN_MBAPPE_POGBA.zip** par exemple)[2].

D. Envoyez-le-moi par e-mail (adresse mail de l'enseignant) pour (telle date).

E. Vous devez présenter votre travail durant une soutenance d'une douzaine de minutes. Vous mettrez l'accent sur les objectifs de votre étude, les principaux résultats obtenus, l'interprétation de ces résultats, les difficultés rencontrées, et le bilan critique de votre propre travail. Tous les membres du groupe doivent participer à la présentation.

[2] Respectez scrupuleusement ces spécifications pour que je puisse traiter facilement vos retours.