Projet d'étude statistique
rapport à remettre
avant mardi 3 mai 2024 minuit

1 But du projet

Le but de projet est d'analyser un jeu de données et de construire un modèle de régression. La démarche sera donc la suivante :

1. analyse descriptive du jeu de données : taille du jeu de données, type des variables à disposition, observations manquantes/aberrantes...
2. analyse des variables du jeu de données : traitements univariés (répartition, adéquation à une loi théorique...)
3. analyse de la liaison entre la variable que l'on veut expliquer et les autres variables du jeu de données
4. construction de modèles de régressions simples
5. construction d'un modèle de régression multiple, avec prise en compte éventuelle des croisements entre variables,
6. si possible, une ACP suivie d'une CAH.

Les parties 4 et 5 sont à faire en `SAS`, la partie 6 en `R`, il n'y a pas de contrainte pour les parties 1 à 3.

2 Instructions

2.1 Pour le rapport

Le rapport sera envoyé par mail aux adresses suivantes : `jeromepcollet@gmail.com` et `renaud.mozet@ensae.org` avant la date limite (les retards seront pénalisés). Le rapport sera au format .pdf.
Le rapport sera construit selon un plan et contiendra une introduction et une conclusion ; sa longueur sera d'environ 15 pages.

2.2 Pour la soutenance

Des soutenances pour présenter les travaux seront organisées. La date sera fixée ultérieurement et chaque groupe devra ensuite s'inscrire pour l'horaire de sa soutenance.
La durée d'une soutenance est 30 minutes : 15 à 20 minutes d'exposé par tous les étudiants du groupe et quelques questions. Un support de présentation de type diaporama est impératif.

3 Les projets

3.1 Évaluation de risque automobile

On dispose des caractéristiques de 205 types de voitures, et des pertes dues aux accidents avec ce type de voiture. On souhaite prévoir les pertes. La difficulté de ce sujet est que la variable à prévoir est binaire, il faut donc effectuer une régression logistique.

3.2 Consommation électrique d'une famille

On dispose de 829 mesures journalières de la consommation d'électricité d'une famille francilienne. Les variables explicatives (en dehors du jour de la semaine) sont pour la plupart à construire : période de vacances, position dans l'année, etc.

3.3 Voitures d'occasion

Le problème consiste à évaluer le prix de voitures d'occasion. Il faudra faire attention à la variable `Model`, car un modèle utilisant cette variable s'ajuste très bien aux données, mais a peu de pouvoir prédictif.

3.4 Impact de la pollution sur la mortalité

Ce sujet est d'actualité : on estime la mortalité en fonction de la pollution, de la structure sociale de la population, du climat.

3.5 Prévision du prix des logements (5 sujets différents)

On prévoit le prix d'un logement, à Munich, Taipei ou Sacramento. Dans le jeu de données de Sacramento, on dispose de plus de données de longitude et latitude, ce qui permet (en `R`) d'estimer un modèle semi-paramétrique les prenant en compte.

3.6 Velib de Londres et Washington DC

On dispose de données sur des vélos partagés. Il faut prévoir le nombre de vélos loués à chaque instant. On pourra par exemple se demander si la sensibilité de la demande à la météo est la même pendant les jours et heures ouvrés et et les autres.

4 Procédure de choix

Vous m'enverrez, avant le 16 février minuit, un mail contenant : la liste des membres du groupe de travail (2 ou 3 personnes, groupes plus grands interdits, groupe de 1 déconseillé), 3 sujets préférés par ordre de préférence (le préféré en premier, ...). Les sujets sont nommés et numérotés ci-dessous.

1. ConsoFamille
2. Munich
3. PollutionMortalite
4. RisqueAutomobile
5. Sacramento
6. Taipei
7. VelibLondres
8. VelibWashington
9. VoituresOccasion