Projet d'Analyse de Données
FISA 2025-2026

Document de travail

1 Présentation

Le projet d'analyse de données est l'occasion de mettre en oeuvre les méthodes d'analyse exploratoire et de modélisation sur des données réelles choisies par les étudiants. Le projet se déroule en binome.

Le travail attendu consiste en un rapport d'étude en pdf et un fichier de code (.r ou .rmd) correctement rédigé qui présente les résultats des analyses effectuées. Aucune soutenance n'est prévue.
Le rendu est tout mis dans un fichier `.zip` et le nom du document contiendra les noms et prénoms de binom : `NOM1_Prenom1_NOM2_prenom2.zip`, par exemple `NGWALANGWALA_Arnaud_MAHIEDDINE_Abed.zip`.

Il contient

— un fichier des codes utilisés avec des explications, en format `.r` ou `.rmd`
— un rapport en `.pdf` décrit les résultats de l'annalyses des données, les commentaires...

Attention : L'absence de rendu de l'un des deux fichers sera synonyme d'un zéro.

2 Déroulement du projet

2.1 Préparation des données

Les données sont dans le fichier `housing.csv` avec la description trouvée ici.

Il s'agit de s'assurer de la mise en forme des données sous la forme d'un tableau $n \times p$, avec les individus en ligne et les variables en colonne. La préparation se fait par les taches suivantes :

1. détecter la présence de données manquantes ; de valeurs éventuellement aberrantes ;
2. s'assurer du type des différentes variables (quantitatives, qualitatives) ;
3. définir une problématique, ou un ensemble de questions au vu de ces données ou selon les objectifs pour lesquels la base de données a été constituée ;
4. proposer les méthodes d'analyse selon le problème défini.

L'ensemble de cette pré-analyse sera reporté dans la première partie du rapport, qui présentera donc les données (leur source notamment), et les objectifs. Celui-ci servira à l'encadrant à évaluer la faisabilité du projet.

2.2 Le rapport

Le rapport contiendra une présentation détaillée des données, et il est attendu d'y trouver plusieurs analyses statistiques :

1. des analyses descriptives univariées et bivariées : les distributions, les tests statistiques des des hypothèses, les corrélations,... avec les représentations graphiques.
2. faire les transformations de données si nécessaire.
3. des analyses factorielles multiples : l'ACP, l'ACM, ... avec l'interprêtation.
4. de la modélisation (régression linénaire simple)

5. de la classification (si possible).

La date limite de remise des projets est **le 02/01/2026** sur `exam.ensiie.fr` sous le dépôt `ando2025_fisa2a_projet`.

L'évaluation du projet portera sur les points suivants :

- Présentation : qualité de la rédaction, et de la présentation des résultats (utilisation pertinente des tableaux et des graphiques en nombre contrôlé, équilibre avec des annexes).
- Problématique et modélisation : Définition d'une bonne problématique adaptée aux données, bon choix (motivé) des méthodes et modèles proposées.
- Méthodes : évaluation de la bonne compréhension des méthodes utilisées basée sur une bonne description des procédures utilisées et la qualité (justesse) des commentaires des sorties statistiques obtenues par les logiciels.

L'utilisation pertinente et juste de méthodes non-vues directement en cours (par exemple telles que le modèle linéaire généralisée ou classification) sera valorisée et donnera lieu à des bonus.