Exercices

Chapitre 5 - Prétraitement des données

Hiver 2025

L'objectif de ces exercices est de réaliser un prétraitement sur un ensemble de données.

Soit l'ensemble de données `carsPreprocessing.xlsx` disponible sur Léa. Ce fichier contient deux feuilles `cars.xls`, le fichier original des données et `carsMod.xls`, un fichier modifié des données que nous allons utilisé pour les excercices du cours.

Excercice 1 : Lecture des données

1. Téléchargez les données du fichier 'carsPreprocessing.xlsx' et stockez le contenu de la feuilles `cars` dans `df` et le contenu de la feuille carsMod dans df1
2. Déterminez le nombre d'individus et de variables
3. Identifiez le type de chacune des variables

Excercice 2 : Traitement des données manquantes

1. Comptez les individus ayant des données manquantes
2. Affichez les individus ayant des données manquantes
3. Supprimez les individus ayant des données manquantes.
4. Stockez le résultat dans df1Del
5. Vérifiez le nombre d'individus et de variables de df1Del
6. Remplacez les individus ayant des données manquantes quantitatives par la moyenne (df1ImputeMean) et par la mediane (df1ImputeMediane)

7. Affichez les résultats d'imputation par les deux stratégies
8. Comparez les résultats aux données réelles (`df`)
9. Commentez les résultats

Excercice 3 : Traitement des données aberrantes

1. Représentez le diagramme de dispersion de la variable `prix` en fonction de `puissance` de l'ensemble de données `df`
2. Représentez le diagramme en boite de la variable `prix` et de la variable `puissance`
3. Commentez les résultats
4. Détectez les valeurs aberrantes en utilisant un seuil η de 3 et un seuil η de 2
5. Représentez les diagrammes en boite correspondants
6. Commentez les résultats

Excercice 4 : Normalisation et standardisation

1. Créer un tableau `x` avec la variable `cylindree`
2. Normalisez la variable `x` en utilisant une mise à l'échelle min-max et stocker le résultat dans `x_minmax`
3. Standardisez la variable `x` et stocker le résultat dans `x_standard`
4. Repésentez l'histogramme des variables `x`, `x_minmax` et `x_standard`