esprit Se former autrement HONORIS UNITED UNIVERSITIES	**Projet Statistiques- 4 DS**	**2023/2024**

En utilisant la base de données et sa fiche descriptive, vous êtes invités à réaliser les tâches suivantes :

<u>Tâche 1 : Importation des données</u>

1. Importer les bases de données de votre projet
2. On s'intéresse à l'étude des trois premiers mois uniquement de l'année 2014. Supprimer alors les données inutiles

Tâche 2: Pré-traitement des données

1. **Création de nouvelles variables**

- En se basant sur la colonne Date,créer une nouvelle variable qualitative "Mois" Dans chaque base de données

3. **Valeurs aberrantes**

- Analyser toutes les variables de la base de données et détecter les valeurs aberrantes.
- Gérer les valeurs aberrantes

4. **Valeurs manquantes**

- Étudier le taux des valeurs manquantes et préciser leur structure
- Proposer une méthode pour l'imputation des valeurs manquantes (justifier votre choix).

5. **Normalisation des données**

- Effectuer la mise à l'échelle des variables si nécessaire.Justifier vos choix

<u>Tâche 2 :Analyse Univariée</u>

Réaliser une analyse univariée de chaque variable de nos données en faisant un résumé statistique et en détectant la distribution des données tout en utilisant des méthodes numériques et graphiques.

Tâche 3 :Analyse bivariée

Étudier la corrélation entre les variables (deux à deux) de la base de données et valider les résultats avec les tests d'hypothèse appropriés tout en mettant l'accent sur la relation entre la variable "O3" et les autres variables du jeu de données.

Tâche 4 :Régression linéaire

1. Effectuer des analyses de régression linéaire simple pour examiner la relation entre "O3" (variable quantitative) et une autre variable.
2. Effectuer des analyses de régression linéaire multiple pour examiner la relation entre "O3" (variable quantitative) et d'autres variables.

3. Proposer une stratégie détaillée pour améliorer la performance du modèle de régression
4. Justifier le choix des métriques et des tests pour comparer les différents modèles.
5. Réduire la dimension de la base de données en utilisant l'analyse en composantes
6. principales et régresser la variable cible quantitative en fonction des nouvelles variables synthétiques

Tâche 5 :Modèles additifs généralisés (GAM) (Réf1)

Les modèles additifs généralisés (**Modèles GAM**) ont montré des performances assez importantes dans la modélisation des données liées à la pollution de l'air

1. Effectuer une étude bibliographique sur ces modèles
2. Utiliser ces modèles afin de comprendre la relation entre la variable cible "O3" et les autres variables explicatives.

Livrables :

Vous êtes invités à préparer :

1. **Un script R : le script doit être fonctionnel et exécutable, contenant toutes les tâches demandées.**
2. **Une présentation : Une présentation descriptive du travail demandé. Un maximum de 30 diapositives.**

Références:

Réf 1: Atelier 8: Modèles additifs généralisés (qcbs.ca)