2020/2021

**TP 2 Préparation des données**

**Application 1 : Jeu de données réel (data\_baby)**

Les données concernent le poids à la naissance de bébés américains de sexe masculin. Pour expliquer les variations de cette variable, d'autres ont été enregistrées, concernant la mère de l’enfant : taille, poids, âge, etc... .

1. Importer le fichier.

2. Décrire le jeu de données : contenu du tableau de données : nom des variables, dimension.

3. Justifier graphiquement l’existence de données aberrantes dans le fichier, en utilisant la variable âge.

4. Identifier les points aberrants (on s’intéresse à la variable âge) et remplacer la valeur aberrante par NA. r

5. Vérifier si toutes les variables ont bien été filtrées.

6. Tracer le poids de la mère en fonction de son nombre de grossesses antérieures (indication : utiliser la fonction **boxplot**)

****

**Application 2 : Utilisation de la méthode « Similar case imputation »**

1. Construire le data frame suivant sous le nom DB

2. Calculer la moyenne de la variable Manpower pour chaque genre (Gender) 3. Imputer les valeurs trouvées dans les cases convenables de la variable Manpower en utilisant la méthode « similar case imputation »

**Application 3 : Jeu de données réel Gazela.xls**

Il s’agit d’un ensemble de données horaires collectées par l’ANPE (Agence nationale de protection de l’environnement) autour de plusieurs polluants dans la station de surveillance de la qualité de l’air située à Cité La Gazelle Ariana (2008-2009)

1. Importer le fichier

2. Décrire le jeu de données : dimension, descriptif des variables et résumé statistique

3. Justifier l’existence de données manquantes dans le fichier

4. Calculer le taux de données manquantes. Proposer alors un scénario de gestion.

5. Dans un premier lieu, on essayera d’imputer les données manquantes pour la variable NO2 (indication : Utiliser le package **Hmisc**)

4.1 Proposer une méthode de type « Generalized imputation »

4.2 Proposer une solution d’imputation en utilisant « Hot deck imputation » 5. On s’intéresse maintenant à la variable SO2.Utiliser l’algorithme KNN pour faire les imputations nécessaires (indication : utiliser le package **VIM**)