Analyse statistique 2021/2022	**Fiche TP 2** Préparation des données	

Application 1 :

Pour déterminer les types des variables dans un tableau de données plusieurs fonctions sont utilisées tel que : str, describe, glimpse…

1. Commencer par installer et charger le package **questionr.**
2. Charger et visualiser la base de données **hdv2003**.
3. Stocker la base de données dans une variable appelée **« db ».**
4. Lister les noms de toutes les variables existantes dans le jeu de données précédents.
5. Classer les variables selon leurs types. (**str** et **describe)**
6. Changer le nom de la variable *bricol* et *occup* en *bricolage* et *occupation.*
7. On peut considérer que la variable numérique *freres.soeurs* est une « fausse » variable numérique et qu'une représentation sous forme d'une variable qualitative serait plus adéquate. Convertir alors la variable *freres.soeurs* en une variable qualitative.

Une opération courante consiste à modifier les valeurs d'une variable qualitative, que ce soit pour avoir des intitulés plus courts ou plus clairs ou pour regrouper des modalités entre elles. Il existe plusieurs possibilités pour effectuer ce type de recodage, mais ici on va utiliser la fonction **fact_recode (** ou **fact_collapse** ou **fact_other**) du package **forcats**. Celle-ci prend en argument une liste de recodages sous la forme *"Nouvelle valeur" = "Ancienne valeur"*.

8. Déterminer les modalités de la variable *qualif.*
9. Changer la modalité *« ouvrier specialise »* en *« ouvrier »*. Vérifier que tout s'est bien passé.
10. Regrouper maintenant les modalités *« ouvrier specialise »* et *« ouvrier qualifié »* dans la même modalité que vous l'appelez *« ouvrier »* et les modalités *« technicien »* et *« profession intermédiaire »* dans la même modalité *« interm »*

L'avantage des variables qualitatives est que leurs modalités peuvent être ordonnées.

11. Transformer la variable « trav.statisf » en une variable qualitative ordinale. **(ordered)**

Une autre opération relativement courante consiste à découper une variable numérique en classes. On utilise pour cela la fonction **cut** :

12. Regrouper les valeurs de la variable « age » dans 5 classes.

Remarque : vous pouvez générer un rapport automatique qui porte la description d'un tableau de données en utilisant la fonction ***create_report*** du package **DataExplore**.

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

	Gender	Manpower	Sales
1	M	25.00	343.00
2	F	NA	280.00
3	M	33.00	332.00
4	M	NA	272.00
5	F	25.00	NA
6	M	29.00	326.00
7	NA	26.00	259.00
8	M	32.00	297.00

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 4 : 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)

7. Importer le fichier
8. Décrire le jeu de données : dimension, descriptif des variables et résumé statistique
9. Justifier l'existence de données manquantes dans le fichier
10. Calculer le taux de données manquantes. Proposer alors un scénario de gestion.
11. Dans un premier lieu, on essayera d'imputer les données manquantes pour la variable NO2 (indication : Utiliser le package **Hmisc**)

11.1 Proposer une méthode de type « Generalized imputation »

11.2 Proposer une solution d'imputation en utilisant « Hot deck imputation »

12. On s'intéresse maintenant à la variable SO2.Utiliser l'algorithme KNN pour faire les imputations nécessaires (indication : utiliser le package **VIM**)

Application 5 : Calcul du Z-Score

Au cours de l'année 2019, les laboratoires agroalimentaires ont participé à un essai inter-laboratoires concernant des analyses d'un élément nutritif majeur (Nitrate). Le traitement statistique montre que l'ensemble des essais peut être considéré comme bon à très bon. La dispersion correspond à une dispersion "typique" de ces essais (préparation des échantillons, techniques analytiques utilisées,...). Le nombre de résultats suspects ou aberrants est faible. Pour détecter les résultats aberrants les laboratoires utilisent le Z-score qui est calculé comme suit :

$$Z - score = \frac{résultat\ laboratoire - valeur\ référence}{écart\ type\ référence}$$

Les références calculées ou fixées :

z-score ≤ 2 : bon

2 < z-score < 3 : suspect

z-score > 3 : insatisfaisant

Les résultats trouvés sont enregistré dans un fichier nommé **résultats analyses**.

1. Importer le fichier.
2. Calculer la moyenne et l'écart type des résultats obtenus.
3. Calculer les z-scores.
4. Donner un résumé sur les Z-scores en utilisant la commande (summary)
5. Interpréter le résultat obtenu.

Application 6 : transformation des données

Une transformation de variables peut résulter en une modification des variables dans un jeu de données ou en la création de nouvelles variables à partir de celles existantes.

Supposons ici que nous travaillons avec un jeu de données **cars** du package **dataset**. Les colonnes du data frame représentent des variables. Il s'agit d'observations prises sur des voitures. La distance parcourue (variable dist) par ces voitures entre le moment d'un freinage et l'atteinte d'une immobilisation complète a été mesurée. La vitesse de déplacement (variable speed) des voitures tout juste avant de freiner a aussi été mesurée.

1. Charger et visualiser la base de données cars.
2. Déterminer type des variable et faire un résumé statistique pour cette base.

3. Travailler sur une copie de ce jeu de données appelée « data1 » afin de ne pas modifier le jeu de données original.
4. Dans notre jeu de données, la distance est exprimée en pieds. Créer une nouvelle variable « dist2 » afin de transformer l'échelle de mesure de la variable « dist » pour des mètres sachant que un pied est l'équivalent de 0.3048 mètre.
5. Créer une nouvelle variable « speed2 » dans le jeu de données data1 contenant la vitesse en km/heure plutôt qu'en miles/heure, mais dans laquelle les valeurs de distance en miles à l'heure inférieures à 10 sont ramenées à des valeurs manquantes sachant que 1.60934 est le facteur de conversion entre un mile à l'heure et un kilomètre à l'heure.
6. Standardiser les valeurs de notre jeu de données data1.
7. Déterminer un résumé statistique pour la base data1.
8. Standardiser la base cars selon la formule suivante : **(x-min(x))/(max(x)-min(x))**