PRATIQUES D'ENQUETES AVEC R

Nguemfouo Ngoumtsa Célina et Mamady I Berete

Vendredi, 18 Avril 2025

Introduction

L'épuration des données après enquête est indispensable pour garantir des résultats fiables. Avec R, on :

- Automatise et documente chaque étape,
- Traite rapidement les valeurs manquantes, aberrantes ou incohérentes,
- Standardise formats et unités,
- Crée et enrichit des variables selon la logique métier.

Cette démarche, à la fois **reproductible**, **flexible** et **efficace**, constitue la base de toute analyse rigoureuse.

Plan de l'exposé

Introduction

I. Partie théorique

1. Présentation d'une enquête
2. Traitement d'une base de données

II. Partie pratique

1. Cas d'une enquête ménage
2. Cas d'une enquête individu

III. Automatisation du traitement d'une base de données

I. Partie théorique

1. Présentation d'une enquête

Une **enquête** est une méthode de collecte d'information visant à analyser un phénomène, évaluer une situation ou vérifier des hypoyhèses. Elle s'appuie sur l'étude d'un échantillon representatif d'une population, via des questionnaires, des entretiens, des observation ou l'exploitation de données existantes. Elle a pour objectif de :

- **Décrire** : analyser les caractéristiques d'une population.
- **Expliquer** : identifier les relations entre différentes variables.
- **Prévoir** : anticiper les tendances à partir des données recueillies.
- **Evaluer** : Mesurer l'impact d'une politique ou d'un programme.

I. Partie théorique

1. Présentation d'une enquête

On distingue plusieurs types d'enquêtes qui peuvent être classés en plusieurs catégories:

- **Selon l'objectif de l'enquête** : on distingue les enquêtes descriptives, les enquêtes analytiques, les enquêtes évaluatives et les enquêtes expérimentales.
- **Selon la méthode de collecte** : on a les enquêtes par questionnaire, par entretien, par observation et par expérimentation.
- **Selon la Périodicité** : c'est ici que l'on retrouve les enquêtes ponctuelle, longitudinale et récurrente.
- **Selon l'unité statistique enquêtée** : là on distingue les enquêtes ménage, individuelle, sur l'entreprise, communautaire ou collective, etc.

I. Partie théorique

1. Présentation d'une enquête

Les principaux types de variables sont :

- **Les variables d'identification** : on retrouve parmi elles les identifiants de l'enquêteur et du répondant, la date et lieu de l'enquête ainsi que les coordonnées GPS (si collectées).
- **Les variables socio-démographiques** : il s'agit de l'âge, le sexe, l'état matrimonial, le niveau d'instruction, la profession, le revenu, la taille du ménage, etc.
- **Les variables de mode de vie** : c'est ici que l'on retroive le milieu de résidence, (urbain/rural), l'accès à l'eau, à l'électricité, à l'éducation et à la santé, activité principale, etc.

On distingue également les variables spécifiques au thème de l'enquête et les variables permettant d'évaluer la qualité des données (temps de réponse, . . .).

I. Partie théorique

2. Traitement d'une base de données

- **Traitement à chaud** : dès la phase de collecte, le superviseur peut intervenir pour limiter la propagation des erreurs. Il peut notament verifier le contrôle géographique, le chronométrage, le profil de réponse ainsi que le feedback quotidien.

Une fois l'enquête terminée, son appurage à froid peut commencer.

- **Visualisation de la base** : la première étape est d'importer et de visualiser la base dans l'environnement R. Pour ce faire, les librairies **haven**, **readr** ou **readxl** peuvent être utilisées. La visualisation de la base peut également necessiter la quelques graphiques ou tableaux. Pour ce faire, les libriries **ggplot2** et **gtsummary** peuvent être utilisées.

I. Partie théorique

2. Traitement d'une base de données

- **Traitement des doublons** : si des doublons sont détectés, ils doivent être supprimés de peur de biaiser les résultats. Toutefois, l'idéal reste toujours de contacter les ménages enquêtés pour avoir de vraies informations.
- **Traitement des valeurs manquantes** : on distingue généralement trois types de valeurs manquantes :
  - **MCAR (Missing Completely At Random)** : les données sont manquantes de façon complètement aléatoire. (Une ligne d'un questionnaire a été perdue à cause d'un problème technique.);
  - **MAR (Missing At Random)** : les données sont manquantes de manière conditionnelle. (Les personnes âgées répondent moins souvent à certaines questions sensibles.);
  - **MNAR (Missing Not At Random)** : les données sont manquantes de manière non aléatoire. (Une personne qui ne souhaite pas déclarer son revenu parce qu'il est très élevé ou très faible.)

I. Partie théorique

2. Traitement d'une base de données

- **Traitement des valeurs manquantes** : le traitement des valeurs manquantes dépend à la fois du **type de variable** concernée (numérique, catégorielle ou texte libre) et du **contexte de l'analyse**.
  - **Variables numériques** : elles peuvent être traitées par suppression, imputation simple, imputation conditionnelle ou régression. On distingue également des méthodes plus avancées comme la méthode des k-NN (k plus proches voisins), la méthode des arbres de décision ainsi que les méthodes bayésiennes ou multiple imputation.
  - **Variables catégorielles (facteurs)** : pour ce type de variables, on a comme méthode la suppression, l'imputation simple et l'imputation conditionnelle. Les modèles de classification (arbre de décision, régression logistique) peuvent également être utilisés, ainsi que d'autres méthodes comme l'utilisation de modèles d'apprentissage automatique adaptés aux variables catégorielles.

I. Partie théorique

2. Traitement d'une base de données

- **Traitement des valeurs manquantes** :
  - **Variables textuelles (libre)** : pour ce type de variable, on a la suppression ou le remplacement par une chaîne vide, la création d'une modalité "Manquant" ou "Non précisé", l'imputation par des règles logiques ou NLP (traitement automatique du langage) : comme une extraction de texte similaire à partir d'autres champs ou une classification sémantique si le texte a des structures régulières.

Il est important de noter que toutes les valeurs manquantes ne sont pas à imputer.

- **Traitement des valeurs aberrantes** : une valeur aberrante est une observation dont la distance à la distribution centrale dépasse significativement celle des autres points. Elle peut être légitime ou erronée. Elle peut être causée par des erreurs humaines, des problèmes techniques ou des caractéristiques intrinsèques de la population.

I. Partie théorique

2. Traitement d'une base de données

- **Traitement des valeurs aberrantes** : la détection de ces valeurs peut se faire de manière graphique par une boîte à moustaches (boxplot), un histogramme ou un scatterplot, mais également de manière statistique grace au Z-score, la méthode IQR, la distance de Mahalanobis et le score de robustesse. Parmi les méthodes de traitement des valeurs manquantes, on distingue la suppression des aberrations, l'imputation, la Winsorisation et la transformation. Des modèles robustes comme les arbres de decision peuvent également être utilisés dans le cadre du traitement des valeus aberrantes.
- **Traitement des incohérences** : une valeur incohérente est une donnée qui ne respecte pas les règles ou les attentes logiques d'un jeu de données.

2. Traitement d'une base de données

- **Traitement des incohérences** : par exemple, des valeurs numériques hors de portée attendue, des valeurs textuelles incorrectes ou mal formatées, des incohérence entre plusieurs variables et des valeurs manquantes ou vides là où elles sont attendues. Afin de détecter ces incohérences, on peut se servir de plage de valeurs, d'histogrammes et Boxplots, d'une détection par logique entre variables, de la vérification des formats de données, de la détection des valeurs manquantes ou vides, voire d'algorithmes de détection de valeurs aberrantes.

II. Partie pratique

1. Cas d'une enquête ménage

La base utilisée est la base ménage de l'EHCVM Sénégal de 2021.

- **Chargement et visualisation des données**

Afin de visualiser correcttement la base, il faut transformer les variables catégorielles en facteurs
On présente à présent quelques observations et quelles variables de notre base.

```
## # A tibble: 10 x 10
##    country  hhid  year grappe menage vague logem
##    <chr>   <dbl> <dbl>  <dbl>  <dbl> <dbl> <fct>
##  1 SEN       201  2021      2      1     2 Proprietaire tit
##  2 SEN       203  2021      2      3     2 Locataire
##  3 SEN       204  2021      2      4     2 Locataire
##  4 SEN       205  2021      2      5     2 Locataire
##  5 SEN       206  2021      2      6     2 Locataire
```

II. Partie pratique

1. Cas d'une enquête ménage

- **Traitement des doublons**

```
## La base contient 0 doublon.
```

On peut donc passer à l'étape suivante.

- **Traitement des valeurs manquantes**

Affichons tout d'abord les variables qui contiennent des valeurs manquantes :

II. Partie pratique

1. Cas d'une enquête ménage

Résumé des valeurs manquantes

Nom de la variable	Libellé	Nombre de NA	% de
superf	Superficie agricole (en ha)	4470	62.8

La seule variable qui comporte des valeurs manquantes est la variable superf. Mais son taux de NA est beaucoup trop grand pour imputer sans biaiser les données.

- **Traitement des valeurs aberrantes**

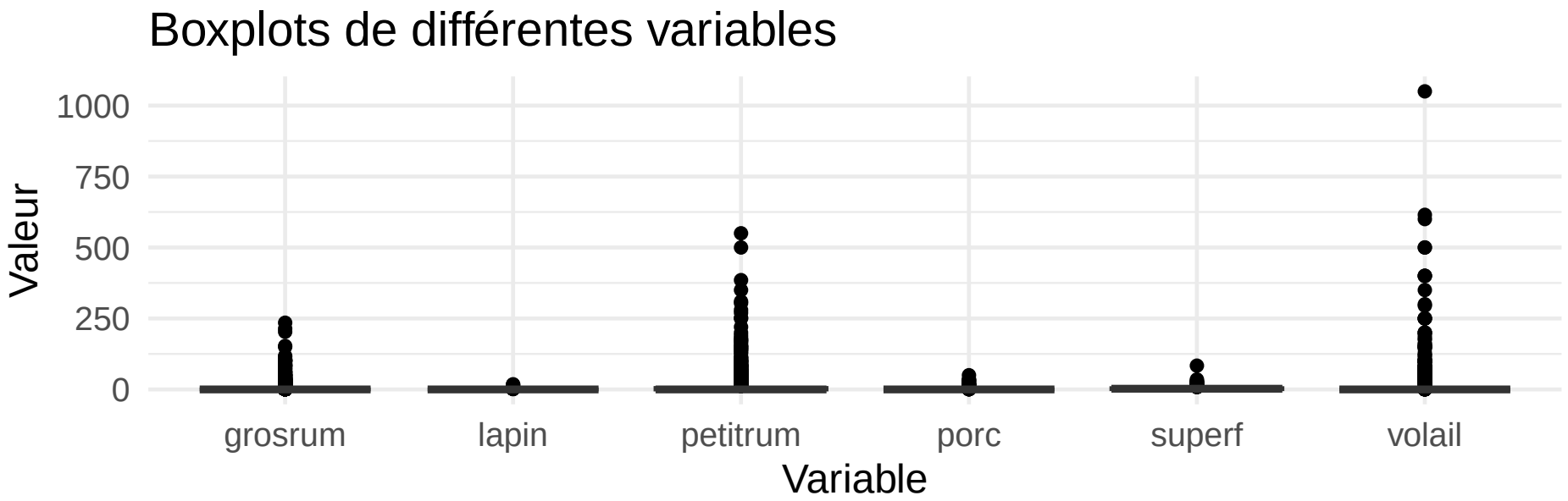
Commençons par lister toutes les variables numériques de la base :

```
##  [1] "hhid"     "year"     "grappe"   "menage"   "vague"
##  [7] "grosrum"  "petitrum" "porc"     "lapin"    "volail"
```

II. Partie pratique

1. Cas d'une enquête ménage

Les variables qui nous interressent sont **superf, grosrum, petitrum, porc, lapin** et **volail**.



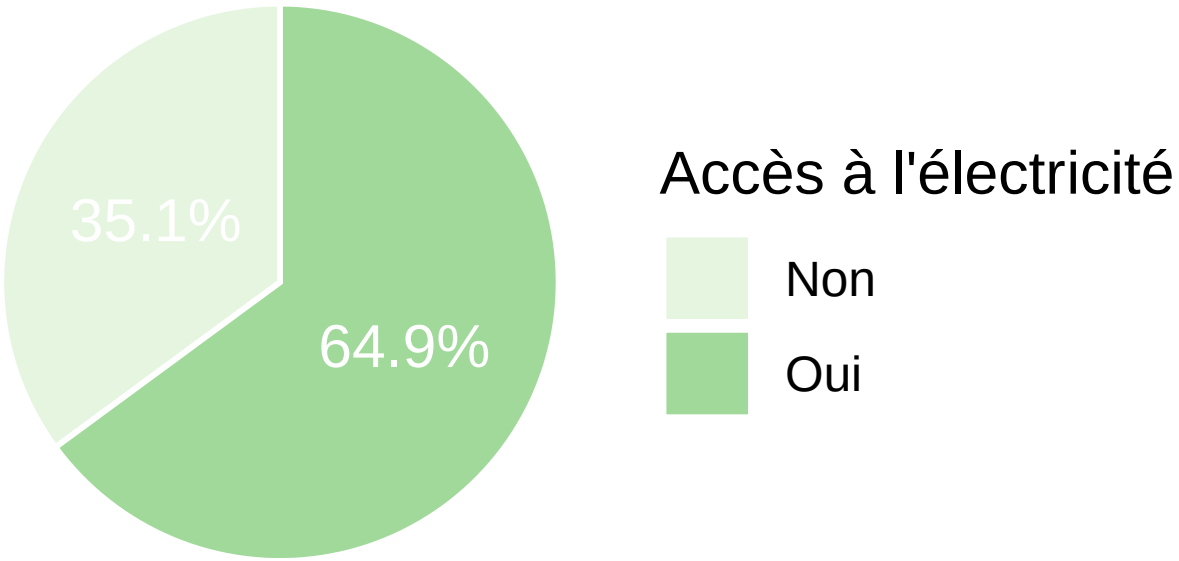
Comme on le voit sur les boxplots, ces variables sont toutes spécifiques au ménage. Les traiter pourrait biaiser les donner et faire perdre les informations.

II. Partie pratique

1. Cas d'une enquête ménage

- **Statistiques descriptives**

Répartition de la variable elec_ac



II. Partie pratique

1. Cas d'une enquête ménage

Table 1: Résumé des accès & équipements du ménage

Variable	Effectif (%)[1]
Acces reseau electrique	
Non	2,498 (35%)
Oui	4,622 (65%)
Toilettes saines	
Non	3,027 (43%)
Oui	4,093 (57%)
Menage a TV	
Non	3,121 (44%)
Oui	3,999 (56%)
Menage a fer electrique	
Non	6,853 (96%)

II. Partie pratique

2. Cas d'une enquête individu

- **Chargement et visualisation des données** Il s'agit de la base de données EHCVM Sénégal de 2021.

Afin de visualiser correcttement la base, il faut transformer les variables catégorielles en facteurs
On présente à présent quelques observations et quelles variables de notre base.

```
## # A tibble: 10 x 10
##    country  year vague  hhid grappe menage numind zae   zae
##    <chr>   <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <fct>  <c
##  1 SEN      2021     2   201      2      1      1 Dakar
##  2 SEN      2021     2   201      2      1      2 Dakar
##  3 SEN      2021     2   201      2      1      3 Dakar
##  4 SEN      2021     2   201      2      1      6 Dakar
##  5 SEN      2021     2   201      2      1      7 Dakar
```

2. Cas d'une enquête individu

- **Traitement des doublons**

```
## La base contient 0 doublon.
```

On peut donc passer à l'étape suivante.

- **Traitement des valeurs manquantes**

Affichons tout d'abord le svariables qui contiennent des valeurs manquantes avec le nombre et le taux de valeurs manquantes :

II. Partie pratique

2. Cas d'une enquête individu

Résumé des valeurs manquantes

Nom de la variable	Libellé	Nombre de NA
mstat	Situation de famille	16
religion	Religion	742
ethnie	Ethnie	1141
nation	Nationalité	742
agemar	Age premier marriage	39532
aff30j	probleme sante	50518
durarr	Duree arret activite pour maladie	55700
con30j	Consulte 30 dern. jours	50518
handit	Handicap tout niveau	7795
handig	Handicap majeur seul	7795
educ_scol	Niv. educ. actuel	47255

II. Partie pratique

2. Cas d'une enquête individu

- **Traitement des valeurs manquantes**: on se rend compte que les variables qui affichent plus de 5% de NA sont des variables qui ont des auts. L'impuation des valeurs manquantes ne va donc considerer que les variables qui ont moins de 5% des NA. Il s'agit des variables **mstat**, **religion**, **ethnie** et **nation**.
  - **Variable mstat (situation de famille)** : affichons les 16 observations qui présentent des valeurs manquantes à la variable mstat :

```
## # A tibble: 6 x 56
##   country  year vague  hhid grappe menage numind zae   zaen
##   <chr>   <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <fct>  <db
## 1 SEN      2021     2  8208     82      8      5 Dakar
## 2 SEN      2021     2  8614     86     14      6 Dakar
## 3 SEN      2021     2 14008    140      8     10 Thie~
## 4 SEN      2021     2 14210    142     10      4 Thie~
```

II. Partie pratique

2. Cas d'une enquête individu

- **Traitement des valeurs manquantes**
  - **Variable mstat (situation de famille)** : en observant ces 16 observations, on se rend compte qu'il ya des incohérences dans la base : on remarque que ces variables ont presque tous les mêmes modalités aux variables age et agemar. La conclusion qui a été faite est que les deux variables ont certainement été confondues. Pour imputer ces observations, nous avons donc attribué à tous ceux qui avaient moins de 15 ans l'observation "Célibataire" pour la variable mstat, et NA pour la variable agemar.

2. Cas d'une enquête individu

- **Traitement des valeurs manquantes**
  - **Variable mstat (situation de famille)**

Pour les autres valeurs manquantes, nous allons imputer en utilisant une méthode d'imputation conditionnelle. Avec les variables classe d'age que nous allons créer et lien de parenté. Le mode sera imputé aux valeurs manquantes de la variable mstat. On vérfie ensuite qu'il n'y a plus de valeurs manquantes pour la variable mstat :

```
## valeurs manquantes : 0
```

II. Partie pratique

2. Cas d'une enquête individu

- **Traitement des valeurs manquantes**
  - **Variable religion** : on a 742 obsevations qui présentent une valeur manquante pour la variable religion. Pour traiter ces valeurs manquantes, on va faire une imputation conditionnelle en utilisant la variable ménage (numéro du ménage), et en attribuant à chaque valeur manquante le mode de la variable region du groupe dans lequel il se trouve avec la variable ménage. On s'assure qu'il n'ya plus de valeurs manquante pour cette variable.

```
## valeurs manquantes : 0
```

La même procédure a été appliquée pour les variables **ethnies** et **nation**.

```
## valeurs manquantes : 0
```

```
## valeurs manquantes : 0
```

2. Cas d'une enquête individu

- **Traitment des valeurs aberrantes** : commençons par lister toutes les variables numériques de la base :

```
##  [1] "year"        "vague"       "hhid"        "grappe"
##  [6] "numind"      "zaemil"      "hhweight"    "age"
## [11] "volhor"      "salaire"     "volhor_sec"  "salaire_sec
```

Les variables qui nous interressent sont: **age, agemar, volhor, salaire, volhor_sec** et **salaire_sec**.

II. Partie pratique

2. Cas d'une enquête individu

Table 2: Résumé statistique de la variable age

Statistique	**N = 63,530**
Age en annees	
Moyenne = Mean	Moyenne = 24.1
Écart-type = SD	Écart-type = 19.6
Médiane = Median	Médiane = 18.0
Min = Min	Min = 0.0
Q1 = Q1	Q1 = 9.0
Q3 = Q3	Q3 = 36.0
Max = Max	Max = 108.0

II. Partie pratique

2. Cas d'une enquête individu

- **Traitment des valeurs aberrantes**
  - **Variables age et agemar** : la plage d'âge est de 0 à 108 ans, ce qui semble normal On conclu que la variable âge ne comporte aucune variable abérrante. Le même procédé a été appliqué pour la variable agemar.

Table 3: Résumé statistique de la variable `horaire travail emplois principal`

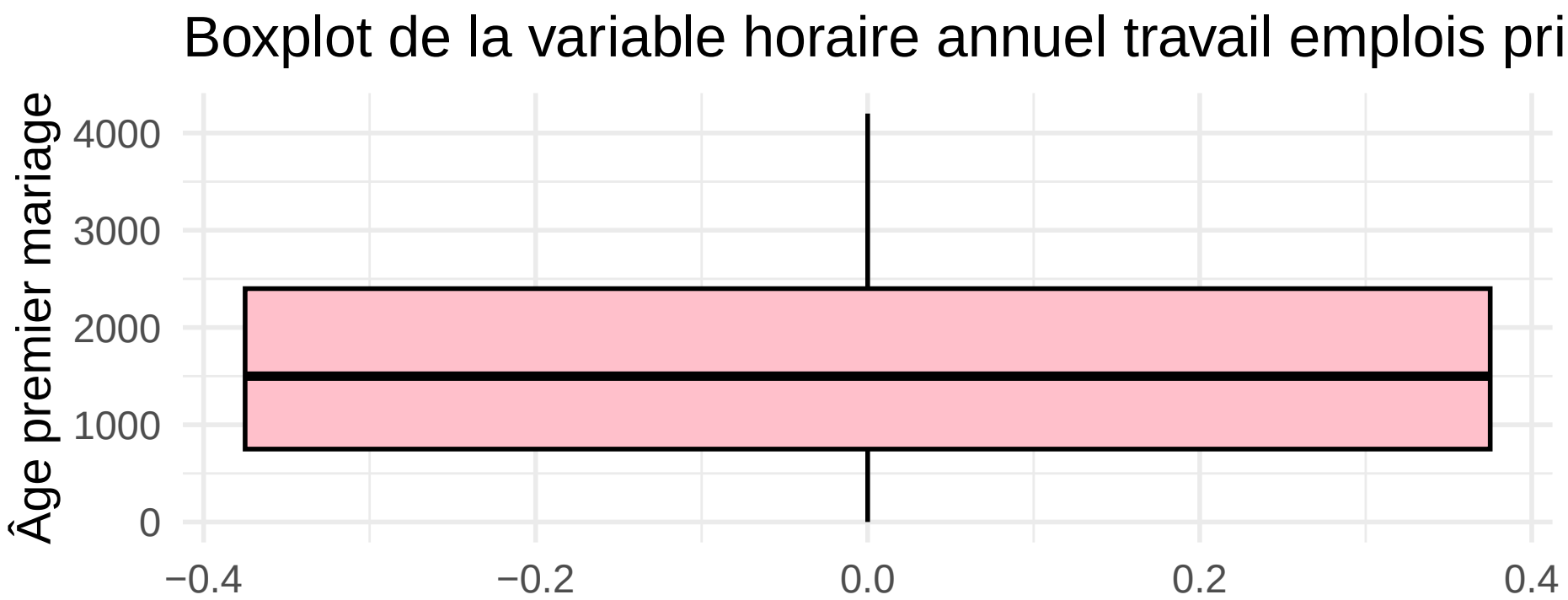
Statistique	**N = 63,530**
Horaire an. travail empl. prin.	
Moyenne = Mean	Moyenne = 1,613.2
Écart-type = SD	Écart-type = 997.9
Médiane = Median	Médiane = 1,500.0
Min = Min	Min = 0.5
Q1 = Q1	Q1 = 750.0
Q3 = Q3	Q3 = 2,400.0
Max = Max	Max = 4,200.0
Unknown	47,456

2. Cas d'une enquête individu

- **Traitment des valeurs aberrantes**
  - **Variable volhor (horaire annuel travail emplois principal)** : le maximum d'heures de travail est 4200 heures par an, soit environ 11 heures et demi par jour de travail, ce qui semble plutôt logique. Poursuivons l'analyse avec un boxplot :

II. Partie pratique

2. Cas d'une enquête individu



Le boxplot ne détecte aucune valeur aberrante. On peut donc passer à la variable suivante.

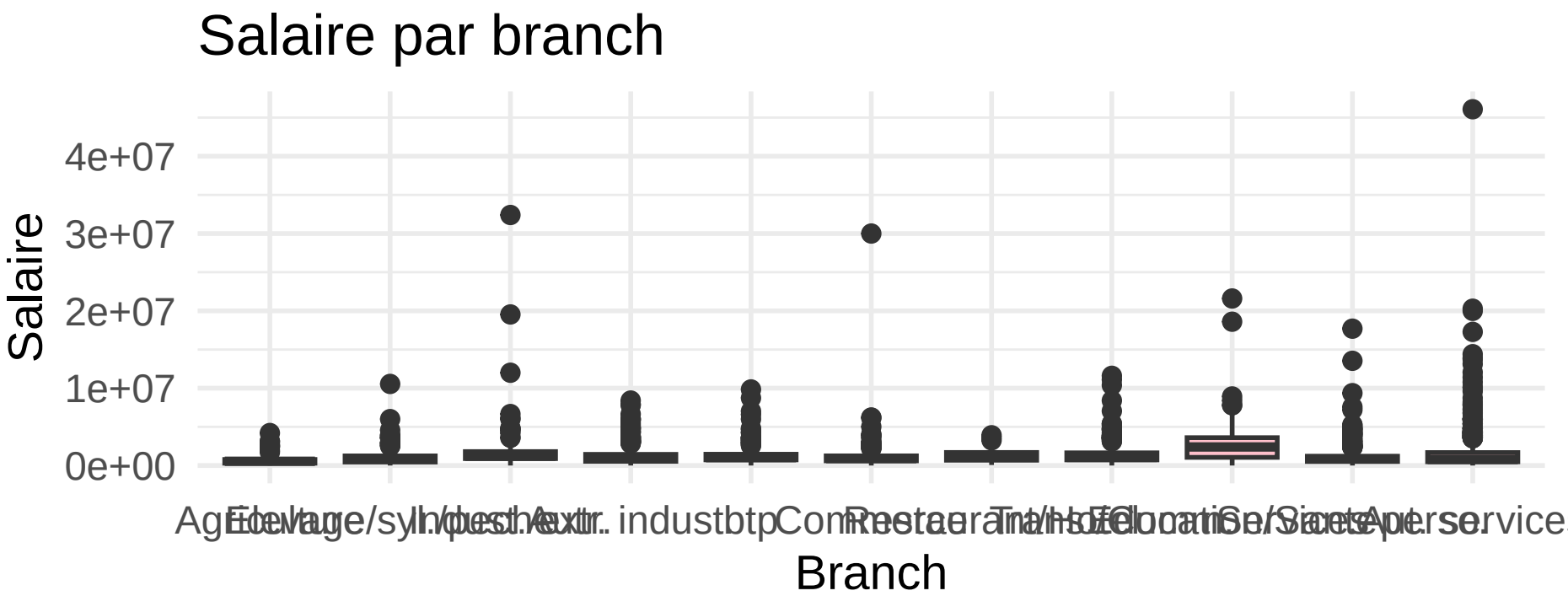
Table 4: Résumé statistique de la variable `salaire annuel emplois principal`

Statistique	**N = 63,530**
Salaire an. empl. prin.	
Moyenne = Mean	Moyenne = 1,366,140.0
Écart-type = SD	Écart-type = 1,705,367.9
Médiane = Median	Médiane = 936,000.0
Min = Min	Min = 0.0
Q1 = Q1	Q1 = 576,000.0
Q3 = Q3	Q3 = 1,631,152.6
Max = Max	Max = 46,080,000.0
Unknown	58,232

II. Partie pratique

2. Cas d'une enquête individu

- **Traitment des valeurs aberrantes**
  - **Variable salaire** : pour visualiser les éventuelles valeurs aberrantes, nous allons afficher les boxplots de la variable salaire par branche d'activité.

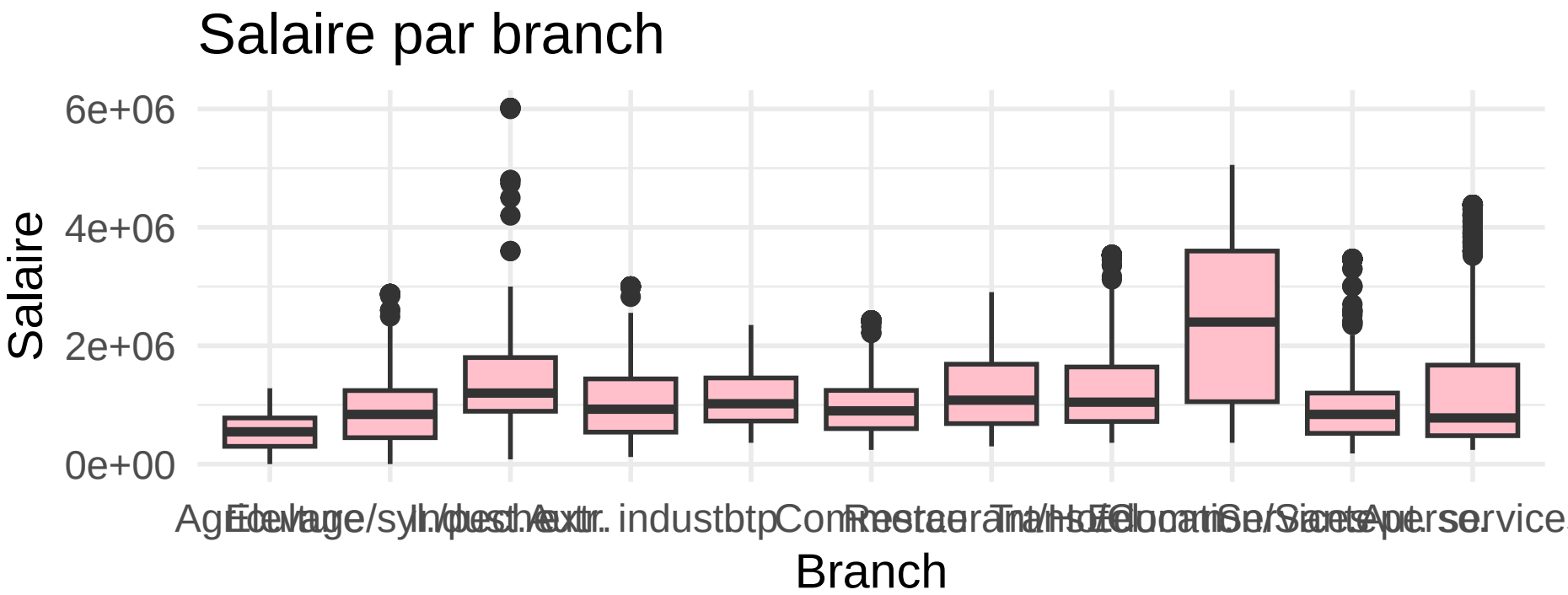


II. Partie pratique

2. Cas d'une enquête individu

- **Traitment des valeurs aberrantes**
  - **Variable salaire** : nous allons imputer les valeurs manquantes par la méthode de la **winsorisation**, et par catégorie de barnche d'activité.

Voici les nouveaux boxplots après traitement des valeurs aberrantes :



II. Partie pratique

2. Cas d'une enquête individu

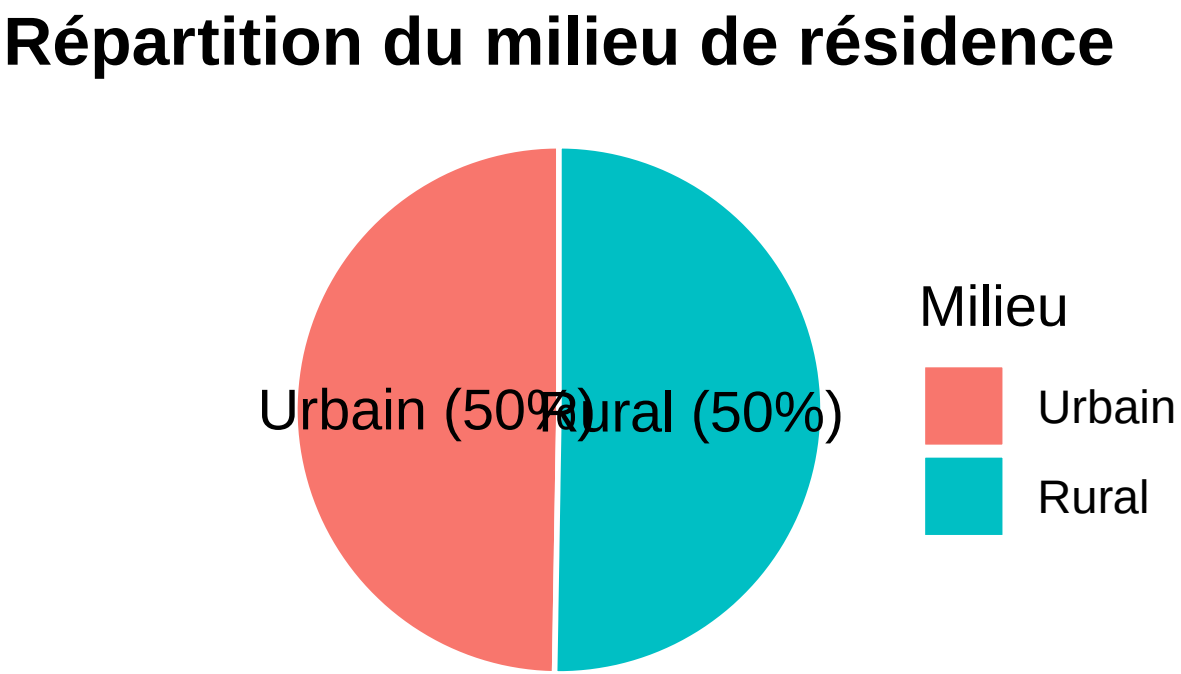
- **Traitment des valeurs aberrantes**
  - **Variable salaire** : On se rend compte que les valeurs extrêmes ont disparues. La même procédure sera suivie pour les variables **volhor_sec** et **salaire_sec**.

- **Statistiques descriptives**

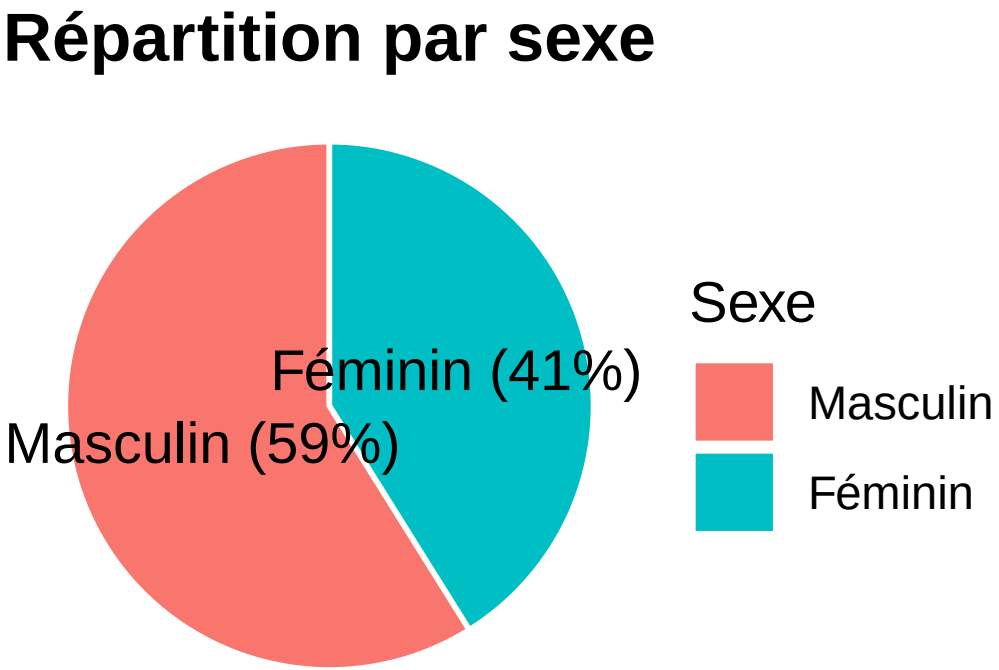
Table 5: Résumé des variables catégorielles : religion, handit, alfa

Variable	**Effectif (%)**[1]
religion	
Musulman	19,207 (95%)
Chrétien	879 (4.4%)
Animiste	26 (0.1%)
Autre Réligion	3 (<0.1%)
Sans Réligion	4 (<0.1%)
Handicap tout niveau	
Non	18,280 (91%)
Oui	1,839 (9.1%)

2. Cas d'une enquête individu

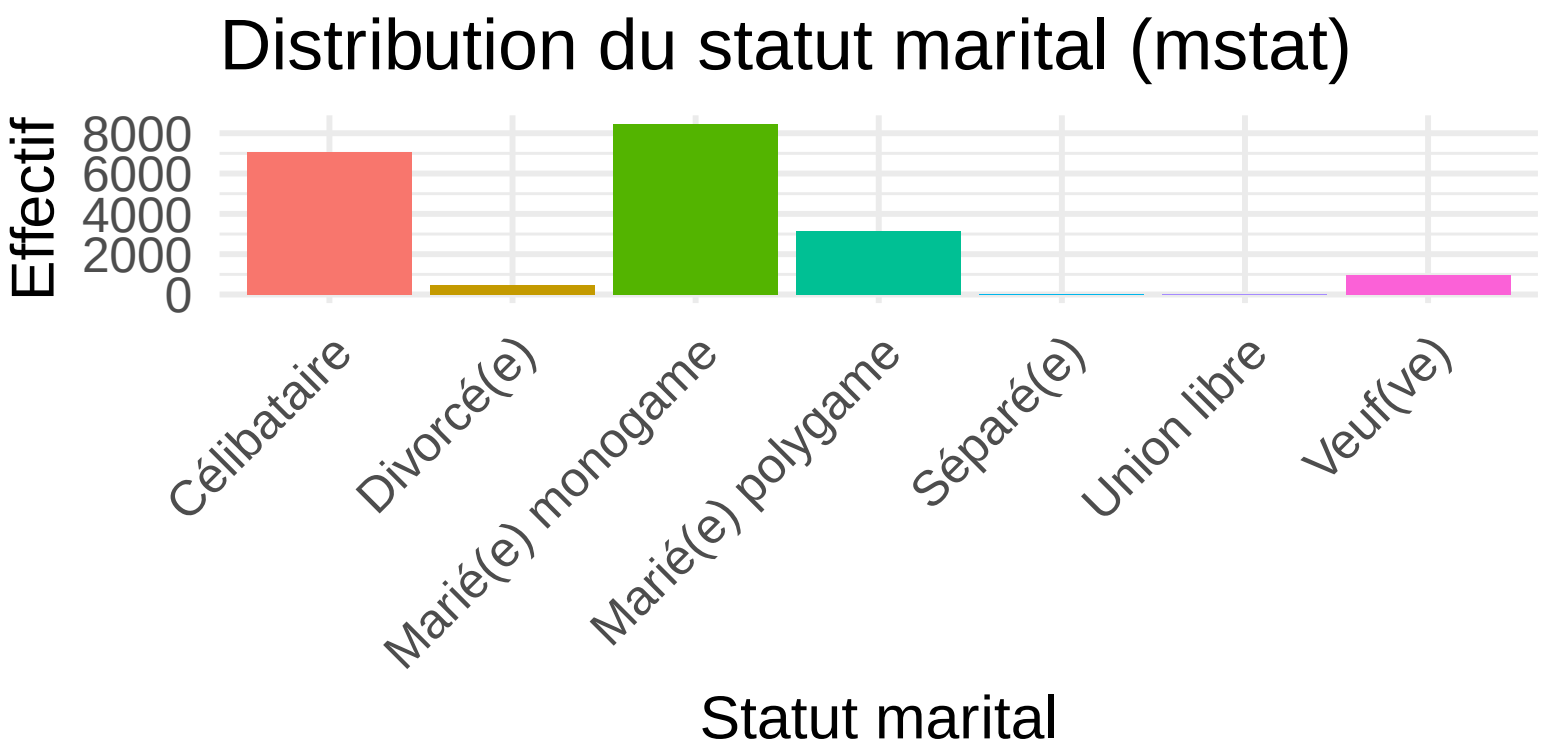


2. Cas d'une enquête individu



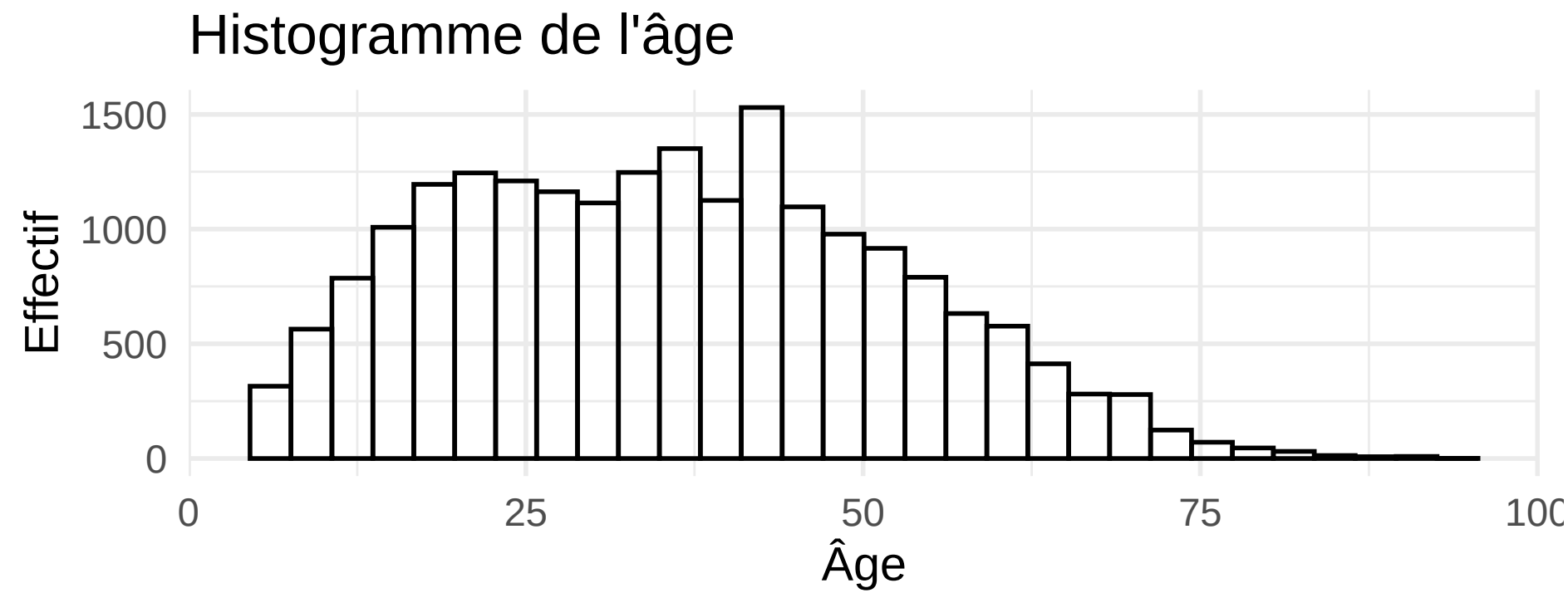
II. Partie pratique

2. Cas d'une enquête individu



II. Partie pratique

2. Cas d'une enquête individu



III. Automatisation du traitement d'une base de données

Une application a été conçue sur R shiny pour automatiser le traitement des bases de données.

Ouvrir la Plateforme de traitements d'enquêtes