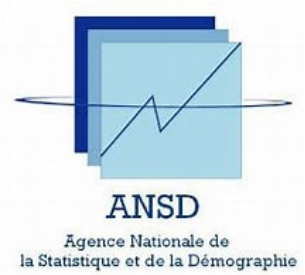
RÉPUBLIQUE DU SÉNÉGAL

Un Peuple - Un But - Une Foi



Agence Nationale de la Statistique et de la Démographie

École Nationale de la Statistique et de l'Analyse Économique

Exposé de R

PRATIQUE DES ENQUETES AVEC R - traitement à froid

Rédigé par :

NGUEMFOUO NGOUMTSA Célina

Mamady I BERETE

Élèves Ingénieurs Statisticiens Économistes

Sous la supervision de :

M. Aboubacar HEMA

Data analyst

Année scolaire : 2024/2025

Avant-propos

Les Ingénieurs Statisticiens Economistes (ISE) jouent un rôle essentiel dans la compréhension des dynamiques économiques, sociales et démographiques, et dans l'aide à la prise de décisions. Leur expertise contribue à l'élaboration et à l'évaluation des politiques publiques, ainsi qu'au développement de stratégies économiques efficaces. L'Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye (ENSAE) du Sénégal a été créée pour offrir une formation de qualité dans ce domaine. Elle propose plusieurs parcours spécialisés, notamment :

- **Analystes Statisticiens (AS)** : une formation de trois ans qui forme des analystes en statistique capables de traiter et d'analyser des données à des fins variées.
- **Ingénieurs Statisticiens Économistes (ISE)** : un cycle court de trois ans ou un cycle long de cinq ans, offrant une formation approfondie pour devenir des cadres spécialisés dans l'analyse statistique et économique. L'accès à l'ENSAE est soumis à un concours rigoureux, ouvert aux étudiants détenteurs du baccalauréat ou d'un diplôme universitaire selon le niveau d'admission. Membre du Réseau des Écoles Africaines de la Statistique (RESA), l'ENSAE collabore avec des institutions prestigieuses comme l'Institut Sous-Régional de Statistique et d'Économie appliquée (ISSEA) au Cameroun, l'École nationale Supérieure de Statistique et d'Économie appliquée (ENSEA) en Côte d'Ivoire, et l'École nationale d'Économie appliquée et de Management (ENEAM) au Bénin. Ce réseau permet d'harmoniser les programmes de formation et d'offrir une reconnaissance internationale des compétences des élèves. Les étudiants de la filière ISE apprennent durant leur cursus à se familiariser avec l'outil **R**, un outil essentiel pour tout analyste de données. C'est dans ce cardre que s'incrit cet exposé ayant pour thème : **Pratique d'enquête avec R**. À travers cet exposé, nous explorons le traitement des données provenant d'enquêtes.

Sommaire

Contents

Résumé

L'usage de **R** dans la pratique des enquêtes statistiques permet une gestion optimale des bases de données collectées lors des enquêtes. Ce document met en lumière les étapes clés du traitement des données à froid, notamment la visualisation, l'identification des doublons, le traitement des valeurs manquantes, des incohérences et des valeurs aberrantes. Grâce à **R**, il est possible d'automatiser et de rendre plus efficaces ces processus, ce qui facilite l'analyse de données massives et complexes. L'application de ces techniques sur des bases de données réelles, telles que celles issues des enquêtes menées au Sénégal, démontre l'importance du logiciel dans l'analyse des conditions de vie des ménages. En outre, l'intégration de **R Shiny** permet une gestion dynamique des données, renforçant ainsi la capacité à générer des rapports et visualisations directement exploitables.

Introduction

Dans le cadre de la formation à l'École Nationale de la Statistique et de l'Analyse Économique (ENSAE) de Dakar, l'utilisation d'outils informatiques modernes est primordiale pour le traitement et l'analyse des données issues des enquêtes statistiques. Parmi ces outils, **R** s'impose comme un logiciel incontournable en raison de sa flexibilité, de sa puissance et de ses capacités d'analyse avancées. En tant qu'outil de traitement statistique, **R** permet non seulement d'exécuter des analyses de données complexes, mais aussi d'assurer un nettoyage rigoureux des bases de données, un traitement efficace des valeurs manquantes et aberrantes, ainsi qu'une visualisation des résultats permettant une meilleure interprétation des données collectées.

Ce document s'inscrit dans une démarche pédagogique visant à approfondir la maîtrise de **R** pour le traitement des données d'enquêtes. Il met en lumière l'importance de ce logiciel dans le cadre de la formation des futurs statisticiens économistes, en leur offrant une plateforme robuste pour manipuler les grandes bases de données et appliquer des méthodes statistiques de manière optimale. À travers cet exposé, nous explorons les diverses étapes du traitement des données provenant des enquêtes, notamment la gestion des doublons, des valeurs manquantes, ainsi que l'identification et le traitement des incohérences, tout en mettant l'accent sur l'utilisation de **R Shiny** pour une gestion dynamique et interactive des données.

L'intégration de **R** dans notre parcours académique est un levier essentiel pour garantir une compréhension approfondie des processus statistiques et pour développer des compétences pratiques en analyse de données, essentielles dans notre futur rôle d'analystes.

Chapitre 1 : Partie théorique

I. Présentation d'une enquête

1. Définition

Une enquête est une méthode de collecte d'informations visant à analyser un phénomène, évaluer une situation ou vérifier des hypothèses. Elle s'appuie sur l'étude d'un échantillon représentatif d'une population, via des questionnaires, des entretiens, des observations ou l'exploitation de données existantes.

2. Objectifs

Les objectifs d'une enquête sont multiples et peuvent varier en fonction du domaine d'application :

- **Décrire** : Analyser les caractéristiques d'une population.
- **Expliquer** : Identifier les relations entre différentes variables.
- **Prévoir** : Anticiper les tendances à partir des données recueillies.
- **Évaluer** : Mesurer l'impact d'une politique ou d'un programme.

3. Types d'enquêtes

Dans le domaine des études statistiques et des sciences sociales, il existe une grande variété de types d'enquêtes. Ces dernières peuvent être classées selon plusieurs critères, notamment leur objectif, leur méthode de collecte, leur périodicité ou encore l'unité statistique étudiée. Cette typologie permet d'adapter l'outil d'enquête aux besoins spécifiques d'analyse.

a. Typologie selon l'objectif de l'enquête

Selon l'objectif visé, on distingue principalement :

- **Les enquêtes descriptives** : elles visent à dresser un état des lieux ou à observer des tendances. Par exemple, les recensements de population permettent de décrire la structure démographique d'un pays à un moment donné.
- **Les enquêtes analytiques** : elles cherchent à mettre en évidence des relations de cause à effet entre différentes variables. Par exemple, une enquête peut analyser le lien entre le niveau d'instruction et l'accès à l'emploi.
- **Les enquêtes évaluatives** : elles sont utilisées pour mesurer l'impact d'un programme ou d'une intervention. Par exemple, une évaluation des effets d'un programme de transfert monétaire sur la scolarisation des enfants.
- **Les enquêtes expérimentales** : elles reposent sur des expériences contrôlées, comme le test d'une politique publique sur un groupe cible, afin d'en évaluer l'efficacité avant une généralisation.

b. Typologie selon la méthode de collecte

La méthode de recueil de l'information influence grandement la qualité et la nature des données :

- **Par questionnaire** : les répondants remplissent un formulaire, soit en face à face, par téléphone ou en ligne.
- **Par entretien** : il peut s'agir d'entretiens individuels ou de groupes (focus groups), souvent plus qualitatifs.
- **Par observation** : l'enquêteur observe directement le comportement ou les conditions, avec ou sans interaction avec les sujets.
- **Par expérimentation** : les données sont collectées dans le cadre d'un test, par exemple avant la mise sur le marché d'un produit.

c. Typologie selon la Périodicité

La fréquence de l'enquête détermine la temporalité de l'analyse :

- **Enquête ponctuelle** : réalisée une seule fois, elle fournit un instantané de la réalité.

- **Enquête longitudinale** : elle suit un même échantillon sur une période donnée afin d'observer l'évolution dans le temps.
- **Enquête récurrente** : conduite à intervalles réguliers (par exemple, chaque année), elle permet d'effectuer des comparaisons temporelles.

d. Typologie selon l'unité statistique enquêtée

L'unité statistique désigne l'entité sur laquelle les observations sont effectuées. Voici les principaux types d'enquêtes en fonction de cette unité :

- **Enquête ménage** : l'unité est le ménage, défini comme un groupe de personnes vivant sous le même toit et partageant les ressources. C'est par exemple le cas des enquêtes sur les conditions de vie, les revenus et dépenses, ou encore le logement. Ces enquêtes permettent d'analyser les conditions économiques, sociales et démographiques des ménages.
- **Enquête individuelle** : ici, chaque individu constitue une unité statistique distincte. C'est par exemple le cas d'enquêtes démographiques et de santé (comme les EDS), enquêtes d'opinion, ou sur les comportements de consommation. Elles permettent de recueillir des données personnelles (santé, opinions, comportements, etc.).
- **Enquête sur l'entreprise** : l'unité statistique est l'entreprise ou la structure économique. C'est le cas des enquêtes sur la production industrielle, les pratiques de gestion, ou les conditions de travail. Elles servent à comprendre la structure, les performances et les caractéristiques économiques des entreprises.
- **Enquête communautaire ou collective** : elle porte sur une communauté (village, groupe ethnique, zone rurale). C'est ici que l'on retrouve les enquêtes sur la santé communautaire ou sur l'accès à l'éducation dans une région spécifique. Ces enquêtes permettent d'identifier les besoins et spécificités d'un groupe ou d'un territoire.

Nous pouvons également citer les enquêtes *institutionnelles*, les enquêtes *sur les biens ou produits*, les enquêtes *territoriales* ou *géographiques* ainsi que les enquêtes *longitudinales*.

4. Catégorie courante de variable dans les enquêtes

L'analyse des enquêtes repose sur un ensemble de variables récurrentes, classées en grandes familles. Leur qualité et leur fiabilité influencent directement la pertinence des résultats. Voici les principaux types de variables :

a. Variables d'identification

Ces variables sont essentielles pour assurer la traçabilité et l'organisation des données collectées.

- **Identifiants de l'enquêteur et du répondant** : ils permettent d'attribuer chaque réponse à un enquêteur spécifique et à un répondant unique. Il est important de s'assurer de l'unicité des identifiants, et d'une bonne correspondance avec les questionnaires.
- **Date et lieu de l'enquête** : Ces informations permettent de s'assurer que les entretiens ont été réalisés conformément au plan d'échantillonnage.
- **Coordonnées GPS (si collectées)** : utile pour géolocaliser précisément les points d'enquête.

b. Variables socio-démographiques

Ces variables fournissent un portrait de base des répondants et sont indispensables pour toute analyse segmentée. Il s'agit de **l'âge, le sexe, l'état matrimonial, le niveau d'instruction, la profession, le revenu, la taille du ménage, etc.**

c. Variables de mode de vie

Elles décrivent l'environnement socio-économique et les conditions de vie des répondants. 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, ...).

II. Traitement d'une base de données

Avant d'engager toute analyse statistique, il est primordial d'« appurer » la base de données issue de l'enquête : il s'agit de détecter, corriger ou documenter toutes les anomalies et de structurer les informations de manière cohérente

1. Traitement à chaud

Dès la phase de collecte, le superviseur peut intervenir pour limiter la propagation des erreurs. Voici quelques vérifications qu'il peut effectuer:

- **Contrôle géographique** : en temps réel, vérifier que chaque point GPS appartient bien à la zone d'échantillonnage prévue ; les relevés hors périmètre peuvent être automatiquement signalés à l'enquêteur pour correction immédiate.
- **Chronométrage** : un formulaire de 30 à 40 questions doit normalement prendre 10 à 15 minutes ; si un questionnaire est bouclé en moins de 2 minutes, on suspecte un remplissage automatique ou bâclé. À l'inverse, un temps supérieur à 30 minutes peut indiquer des difficultés de compréhension.
- **Profil de réponse** : comparer les schémas de réponses entre enquêtés successifs ; une similitude trop forte (tous les mêmes choix) suggère un copié-collé ou une fraude.
- **Feedback quotidien** : grâce aux logs de synchronisation, le superviseur identifie les enquêteurs dont les saisies montrent des anomalies répétées et organise des rappels à l'ordre ou des sessions de recadrage sur le terrain.

Chaque contrôle à chaud permet de rectifier immédiatement les cas manifestes et de réduire l'effort de nettoyage a posteriori. Cette vérification est importante et ne doit pas être négligée.

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

2. Visualisation de la base

La première étape 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.

3. Traitement des doublons

La première chose à vérifier lors du traitement à froid des données est la présence de 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.

4. Traitement des valeurs manquantes

Les valeurs manquantes dans une base de données peuvent apparaître pour différentes raisons (non-réponse, erreur de saisie, données non disponibles, etc.). On distingue généralement trois types de valeurs manquantes :

- **MCAR (Missing Completely At Random)** : Les données sont manquantes de façon complètement aléatoire. L'absence de données n'est liée ni aux variables observées ni aux variables non observées.
  Exemple : 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, c'est-à-dire que l'absence de réponse dépend d'autres variables observées dans le jeu de données.
  Exemple : les personnes âgées répondent moins souvent à certaines questions sensibles.
- **MNAR (Missing Not At Random)** : Les données sont manquantes de manière non aléatoire, c'est-à-dire que l'absence dépend de la variable elle-même.
  Exemple : une personne qui ne souhaite pas déclarer son revenu parce qu'il est très élevé ou très faible.

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**.

a. Variables numériques

Voici quelques méthodes de traitement des valeurs manquantes :

- **Suppression** : supprimer les lignes ou colonnes contenant trop de valeurs manquantes (si la proportion est faible). Il faut faire attention à ne pas biaiser l'analyse si les données ne sont pas MCAR.
- **Imputation simple** : cette méthode consiste à remplacer les valeurs manquantes par la moyenne, la médianne (plus robuste aux valeurs extrêmes ou une valeur constante.
- **Imputation conditionnelle** : c'est une sorte d'amélioration de l'imputation simple. On impute toujours la moyenne ou la médianne, mais en faisant des regroupement par variable catégorielle (sexe, région).
- **Régression** : cette méthode permet de prédire la variable manquante à partir d'autres variables (modèle linéaire).

On distingue également des méthodes plus avancées comme la méthode des **k-NN (k plus proches voisins)**, imputation basée sur les observations similaires; la méthode des **arbres de décision** avec des modèles comme *random forest* pour estimer les valeurs manquantes; ainsique les méthodes **bayésiennes** ou **multiple imputation** pour refléter l'incertitude liée au manque.

b. Variables catégorielles (facteurs)

les méthodes d'imputayion les plus utilisées pour les variables catégorielles sont :

- **Suppression** : comme précédemment, on peut proceder à la suppression si la proportion est faible ou si la catégorie n'est pas cruciale.
- **Imputation simple** : cette méthode consiste à remplacer par les valeurs manquantes par la **modalité la plus fréquente** (mode) ou une valeur spéciale (comme "Inconnu", "Non répondu") pour garder la trace du manquant.
- **Imputation conditionnelle** : Il s'agit de regrouper les observations par d'autres variables catégorielles comme sexe ou zone géographique avant de calculer le mode et de proceder à l'imputation.

Les modèles de classification (arbre de décision, régression logistique) peuvent également être utiisés pour prédire la modalité manquante, ainsi que d'autres méthodes comme l'utilisation de modèles d'apprentissage automatique adaptés aux variables catégorielles.

c. Variables textuelles (libre)

Voici quelques méthodes d'imputation pour ce type de variable :

- **Suppression ou remplacement par une chaîne vide** (**""**) si la variable est peu informative.
- **Création d'une modalité "Manquant"** ou "Non précisé".
- **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 (ex : intitulé de métier, nom de commune).

Le choix de la méthode d'imputation dépend non seulement du type de la variable, mais également du **taux de valeurs manquantes**, du **type de données** et de leur importance dans l'analyse, du **modèle statistique ou machine learning** utilisé ensuite et des **ressources disponibles** (temps, capacité de calcul, expertise).

Il est important d'**analyser les motifs de manquants** avant toute imputation pour ne pas introduire de biais.

Une autre chose à noter est que toutes les valeurs manquantes ne sont pas à imputer, par exemple, un enfant de 6 ans qui présente NA à la variable emplois n'est pas une valeur manquante. Imputer cette valeur reviendrait à fausser complètement la base de données. Enfin, le meilleur de traiter une valeur est autant que possible chercher à avoir accès aux réelles données.

5. Traitement des valeurs aberrantes

Dans toute étude quantitative, les valeurs aberrantes (outliers) sont des observations extrêmes qui divergent sensiblement de la tendance générale des données. Leur présence peut biaiser les estimations statistiques, fausser les modèles et conduire à des conclusions erronées. Le traitement approprié des valeurs aberrantes est donc essentiel pour garantir la fiabilité des analyses.

Une valeur aberrante se définit comme une observation dont la distance à la distribution centrale dépasse significativement celle des autres points. Elle peut être légitime (phénomène rare) ou erronée (erreur de saisie, de mesure, etc.). Elle peut être causée par

des erreurs humaines (saisie, codage), des problèmes techniques (capteurs défaillants) ou des caractéristiques intrinsèques de la population (phénomènes extrêmes).

La détection de ces valeurs peut se faire de manière graphique ou statistique :

- **Approche graphique** :
  - **Boîte à moustaches (boxplot)** : points au-delà des moustaches (1,5 × IQR) sont suspects.
  - **Histogramme** : pic isolé dans les extrémités.
  - **Scatterplot** (pour variables continues) : nuages de points mettant en évidence les points isolés.
- **Approche statistique** :
  - **Z-score** : $z_i = (x_i - \bar{x})/s$. Valeurs |z| > 3 souvent considérées comme aberrantes.
  - **Méthode IQR** : observations < Q1 - 1,5 × IQR ou > Q3 + 1,5 × IQR
  - **Distance de Mahalanobis** (multidimensionnel)
  - **Score de robustesse** : basés sur la médiane et l'écart absolu médian (MAD).

    $$\text{MAD} = \text{median}(|x_i - \text{median}(x)|)$$

    $$\text{robust\_z}_i = (x_i - m)/(1.4826 \times \text{MAD})$$

Voici quelques méthodes de traitement des valeurs aberrantes :

- **Suppression des aberrations** : C'est la méthode d'imputation la plus simple, mais elle est très dangereuse, car elle conduit à une perte d'information, et peut biaiser les données si les outliers sont légitimes.
- **Imputation** : la valeur de remplacement peut être la moyenne, la médiane, ou une méthode plus sophistiquée (k-NN, régression multiple).
- **Winsorisation** : il s'agit de rapprocher les valeurs extrêmes aux percentiles (p.ex. 5e et 95e).
- **Transformation** : elle consiste à faire une transformation logarithmique, racine carrée, ou Box–Cox des données afin de réduire l'effet des valeurs extrêmes.

Des modèles robustes comme les arbres de decision peuvent également être utilisés dans le cadre du traitement des valeus aberrantes.

6. Traitement des incohérences

Une **valeur incohérente** fait référence à une donnée qui ne respecte pas les règles ou les attentes logiques d'un jeu de données, ce qui peut entraîner des erreurs dans l'analyse. Les incohérences dans les données peuvent être dues à plusieurs facteurs, comme une mauvaise collecte, des erreurs de saisie, ou des anomalies dans la logique des valeurs par rapport à d'autres variables.

Voici quelques exemples de valeurs incohérentes :

- **Valeurs numériques hors de portée attendue** : par exemple, une variable "âge" dans une enquête pourrait avoir une valeur de 200, ce qui est incohérent (car l'âge ne peut pas être supérieur à 120 dans la plupart des cas). Ou encore un salaire de 0 FCFA pour une personne travaillant dans une entreprise.
- **Valeurs textuelles incorrectes ou mal formatées** : une variable "sexe" pourrait avoir une valeur "Homme" ou "Femme", mais une valeur "Autre" ou "Indéterminé" dans un cas où cette information ne devrait pas exister. Ou une colonne "date de naissance" peut contenir des valeurs comme "1999-02-30" ou des formats incorrects (par exemple "31/02/2022").
- **Incohérence entre plusieurs variables** : dans une base de données sur les ventes, une "quantité vendue" pourrait être indiquée comme étant de 1000, mais si le "prix unitaire" est 0, cela créerait une incohérence par rapport au montant total de la vente. Si une variable "sexe" est indiquée comme "Femme" et une variable "État civil" indique "Marié(e)", mais la colonne "Âge" est indiquée comme étant "10 ans", il y a une incohérence dans l'ensemble des données.
- **Valeurs manquantes ou vides là où elles sont attendues** : par exemple, une variable "numéro de téléphone" peut avoir une valeur vide pour un client qui a fourni ses informations.

Il existe plusieurs approches pour détecter ces valeurs incohérentes dans un jeu de données. Ces approches peuvent être réalisées de manière manuelle (en inspectant les données) ou automatisée à l'aide de techniques de *contrôle de la qualité des données*.

Voici quelques méthodes pour détecter des valeurs incohérentes :

- **Détection par validation des valeurs numériques** :

- *Plage de valeurs* : Vérifiez que les valeurs numériques se trouvent dans une plage raisonnable. Par exemple, pour l'âge, les valeurs devraient être comprises entre 0 et 120 ans.
- *Histogrammes et Boxplots* : Utilisez des graphiques comme les histogrammes ou boxplots pour visualiser les distributions des données et identifier les valeurs extrêmes ou aberrantes.

- **Détection par logique entre variables** :
  - Utilisez des règles logiques pour vérifier les relations entre les variables. Par exemple, si "sexe" = "Homme", alors "grossesse" ne peut pas être une variable valide.
  - Vérifiez que les relations entre les variables sont cohérentes. Si une variable "quantité" a une valeur supérieure à une certaine limite et une variable "prix" est égale à zéro, l'analyse doit déclencher une alerte.

- **Vérification des formats de données** :
  - Assurez-vous que les formats des dates, des numéros de téléphone, des adresses email, etc., sont conformes aux attentes. Par exemple, une date de naissance doit être dans un format valide (yyyy-mm-dd), sinon elle peut être signalée comme incohérente.
  - En R, vous pouvez utiliser des expressions régulières (regex) pour vérifier les formats de chaînes.

- **Détection des valeurs manquantes ou vides** :
  - Vérifiez les valeurs manquantes dans les colonnes essentielles à l'analyse. Une valeur vide dans une variable qui doit absolument être renseignée (comme le "numéro de téléphone" d'un client) est une incohérence.
  - Utilisez des fonctions comme is.na() pour détecter les valeurs manquantes.

- **Utilisation d'algorithmes de détection de valeurs aberrantes** :
  - Utilisez des algorithmes statistiques comme *IQR (interquartile range)* ou *Z-scores* pour identifier des valeurs anormales. Ces valeurs peuvent être considérées comme incohérentes si elles sont extrêmement éloignées de la moyenne.

Chapitre 2 : Partie pratique

Tout au long de ce chapitre, nous utiliserons deux bases issues de l'**EHCVM** (Enquête Harmonisée sur les Conditions de Vie des Ménages) menée au Sénégal en 2021. Cette enquête s'inscrit dans un cadre d'harmonisation des statistiques de la pauvreté et des conditions de vie afin d'améliorer l'élaboration et l'évaluation des politiques publiques.

I. Cas d'une enquête ménage

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

1. Chargement et visualisation des données

La première étape pour traiter une base est de la charger.

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          mur   toit  so
##    <chr>   <dbl> <dbl>  <dbl>  <dbl> <dbl> <fct>          <fct> <fct> <
##  1 SEN     201  2021     2     1     2 Proprietaire titre Oui   Oui   Oui
##  2 SEN     203  2021     2     3     2 Locataire          Oui   Oui   Oui
##  3 SEN     204  2021     2     4     2 Locataire          Oui   Oui   Oui
##  4 SEN     205  2021     2     5     2 Locataire          Oui   Oui   Oui
##  5 SEN     206  2021     2     6     2 Locataire          Oui   Oui   Oui
##  6 SEN     207  2021     2     7     2 Locataire          Oui   Oui   Oui
##  7 SEN     208  2021     2     8     2 Locataire          Oui   Oui   Oui
##  8 SEN     212  2021     2    12     2 Locataire          Oui   Oui   Oui
##  9 SEN     213  2021     2    13     2 Proprietaire titre Oui   Oui   Oui
## 10 SEN     214  2021     2    14     2 Locataire          Oui   Oui   Oui
```

2. Traitement des doublons

Determinons le nombre de doublons que contient notre base :

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

On peut donc passer à l'étape suivante.

3. Traitement des valeurs manquantes

Pour traiter les valeurs manquantes, affichons tout d'abord les variables qui contiennent des valeurs manquantes avec le nombre et le taux de valeurs manquantes :

Résumé des valeurs manquantes

Nom de la variable	Libellé	Nombre de NA	% de NA
superf	Superficie agricole (en ha)	4470	62.800

La seule variable qui comporte des valeurs manquantes est la vaiable superf. Mais son taux de NA est beaucoup trop grand pour imputer sans biaiser les données. Il s'agit d'une variable où il doit y avoir des NA, du fait que tous les ménages ne possèdent pas de superficie agricole.

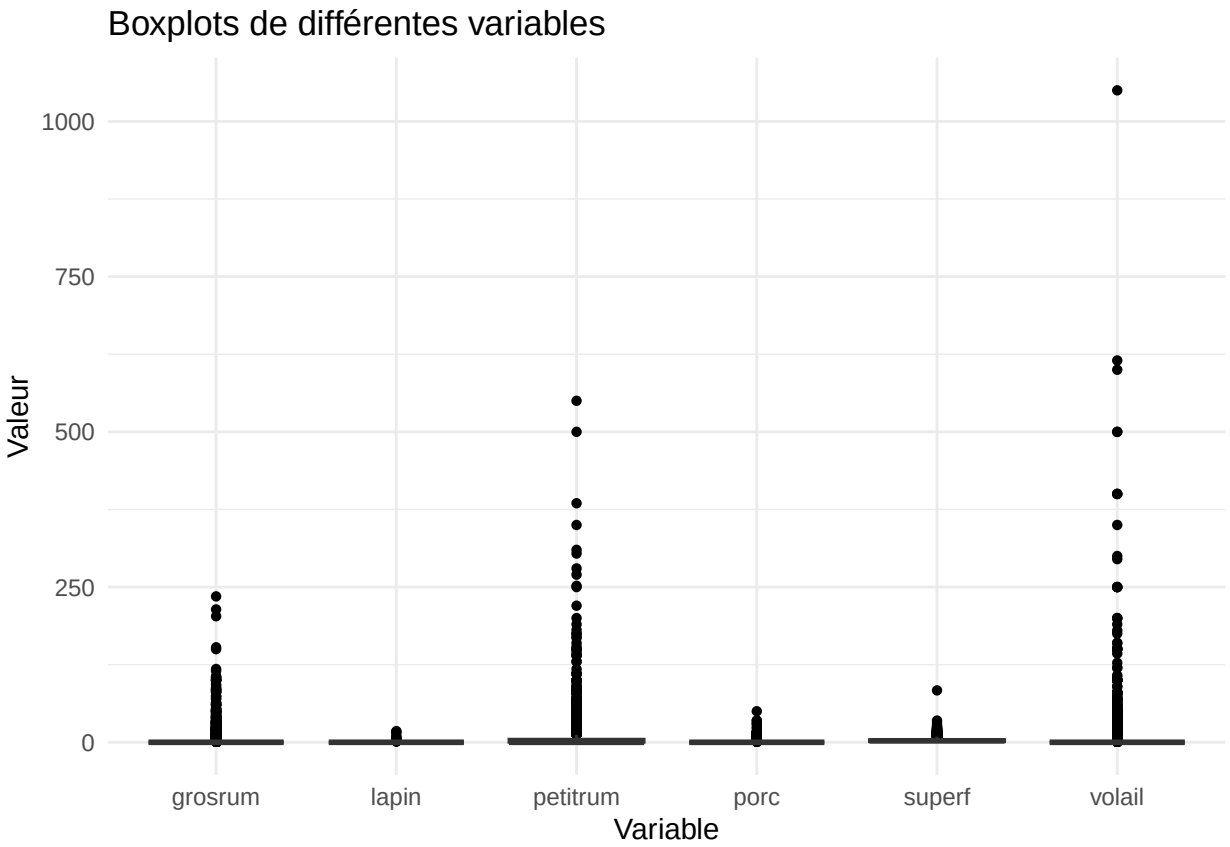
4. Traitment des valeurs aberrantes

Une fois les valeurs manquantes traitées, traitons à présent les valeurs abérrantes.

Commençons par lister toutes les variables numériques de la base :

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

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. Cas d'une enquête individu

1. Chargement et visualisation des données

On commence par charger la base de 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   zaemil region
##    <chr>   <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <fct>  <dbl> <fct>
##  1 SEN     2021     2   201      2      1      1 Dakar     11 dakar
##  2 SEN     2021     2   201      2      1      2 Dakar     11 dakar
##  3 SEN     2021     2   201      2      1      3 Dakar     11 dakar
##  4 SEN     2021     2   201      2      1      6 Dakar     11 dakar
##  5 SEN     2021     2   201      2      1      7 Dakar     11 dakar
##  6 SEN     2021     2   201      2      1      8 Dakar     11 dakar
##  7 SEN     2021     2   201      2      1     13 Dakar     11 dakar
##  8 SEN     2021     2   201      2      1     14 Dakar     11 dakar
##  9 SEN     2021     2   201      2      1     16 Dakar     11 dakar
## 10 SEN     2021     2   203      2      3      2 Dakar     11 dakar
```

2. Traitement des doublons

Determinons le nombre de doublons que contient notre base :

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

On peut donc passer à l'étape suivante.

3. Traitement des valeurs manquantes

Pour traiter les valeurs manquantes, affichons tout d'abord le svariables qui contiennent des valeurs manquantes avec le nombre et le taux de valeurs manquantes :

Résumé des valeurs manquantes

Nom de la variable	Libellé	Nombre de NA	% de NA
mstat	Situation de famille	16	0.000
religion	Religion	742	1.200
ethnie	Ethnie	1141	1.800
nation	Nationalité	742	1.200
agemar	Age premier marriage	39532	62.200
aff30j	probleme sante	50518	79.500
durarr	Duree arret activite pour maladie	55700	87.700
con30j	Consulte 30 dern. jours	50518	79.500
handit	Handicap tout niveau	7795	12.300
handig	Handicap majeur seul	7795	12.300
educ_scol	Niv. educ. actuel	47255	74.400
branch	Branche activite	43411	68.300
sectins	Sect. institutionnel empl. prin.	41479	65.300
csp	CSP empl. prin.	41376	65.100
volhor	Horaire an. travail empl. prin.	47456	74.700
salaire	Salaire an. empl. prin.	58232	91.700
sectins_sec	Secteur instit. emploi sec.	58905	92.700
csp_sec	CSP emploi sec.	58905	92.700
volhor_sec	Horaire an. travail emploi sec.	60863	95.800
salaire_sec	Salaire an. emploi sec.	63246	99.600
serviceconsult	Service de santé consulté	50518	79.500
persconsult	Personnel de santé consulté	50518	79.500

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   zaemil region de
##  <chr>   <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl> <fct>  <dbl> <fct>  <fc
## 1 SEN     2021     2  8208     82      8      5 Dakar     11 dakar  guediawaye
## 2 SEN     2021     2  8614     86     14      6 Dakar     11 dakar  guediawaye
## 3 SEN     2021     2 14008    140      8     10 Thie~      5 diour~ diourbel
## 4 SEN     2021     2 14210    142     10      4 Thie~      5 diour~ diourbel
## 5 SEN     2021     1 21310    213     10      8 Sain~      4 SAINT~ SAINT LOU
## 6 SEN     2021     1 22003    220      3     18 Zigu~     10 tamba~ bakel
## # i 45 more variables: commune <chr>, milieu <fct>, hhweight <dbl>,
## #   resid <fct>, sexe <fct>, age <dbl>, lien <fct>, mstat <fct>,
## #   religion <fct>, ethnie <fct>, nation <fct>, agemar <dbl>, mal30j <
## #   aff30j <fct>, arrmal <fct>, durarr <fct>, con30j <fct>, hos12m <fc
## #   couvmal <fct>, moustiq <fct>, handit <fct>, handig <fct>, alfa <fc
## #   alfa2 <fct>, scol <fct>, educ_scol <fct>, educ_hi <fct>, diplome <
## #   telpor <fct>, internet <fct>, activ7j <fct>, activ12m <fct>, ...
```

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 moisn de 15 ans l'observation "Célibataire" pour la variable mstat, et NA pour la variable agemar.

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
```

- **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
```

- **Variable ethnie**

Là également, il sera question d'imputer les valeurs manquantes de la variable ethnie en la regroupant avec la variable ménage.

Une fois l'imputation terminée, déterminons s'il n'y a plus de valeurs manquante pour cette variable.

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

- **Variable nation**

La même procédure que précédement sera appliquée pour traiter les valeurs manquantes de la variable nation.

Une fois l'implémentation terminée, on se rassure qu'il n'y a plus de valeurs manquantes.

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

4. Traitment des valeurs aberrantes

Une fois les valeurs manquantes traitées, traitons à présent les valeurs abérrantes.

Commençons par lister toutes les variables numériques de la base :

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

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

- **Variable age**

Une valeur abberrante serait une variable trop forte pour être un âge. Affichons donc un résumé statistique de la variable :

Table 1: **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

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.

- **Variable agemar (âge premier mariage)**

Le même processus sera appliqué pour cette variable. On a alors le tableau suivant :

Table 2: **Résumé statistique de la variable `age premier mariage`**

Statistique	**N = 63,530**
agemar	
Moyenne = Mean	Moyenne = 21.7
Écart-type = SD	Écart-type = 5.6
Médiane = Median	Médiane = 20.0
Min = Min	Min = 12.0
Q1 = Q1	Q1 = 18.0
Q3 = Q3	Q3 = 25.0
Max = Max	Max = 85.0
Unknown	39,576

On remarque que la plage d'âge de premier mariage est de 10 à 80 ans. De plus, 25% des répondants à cette question se sont mariés pour la première fois avant 18 ans. Ces informations laissent quelques doutes quant à la véracité de ces informations, d'autant plus que plus haut, le constat que les variables age et agemar semblaient avoir été cofondues.

Toutefois, tenter de d'imputer ces valeurs aberrantes avec des valeurs qui sont sans doutes fausses contribuerait d'avantage à biaiser les données.

- **Variable volhor (horaire annuel travail emplois principal)**

Ici également, nous allons commencer par afficher un résumé statique de la variable :

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

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 :

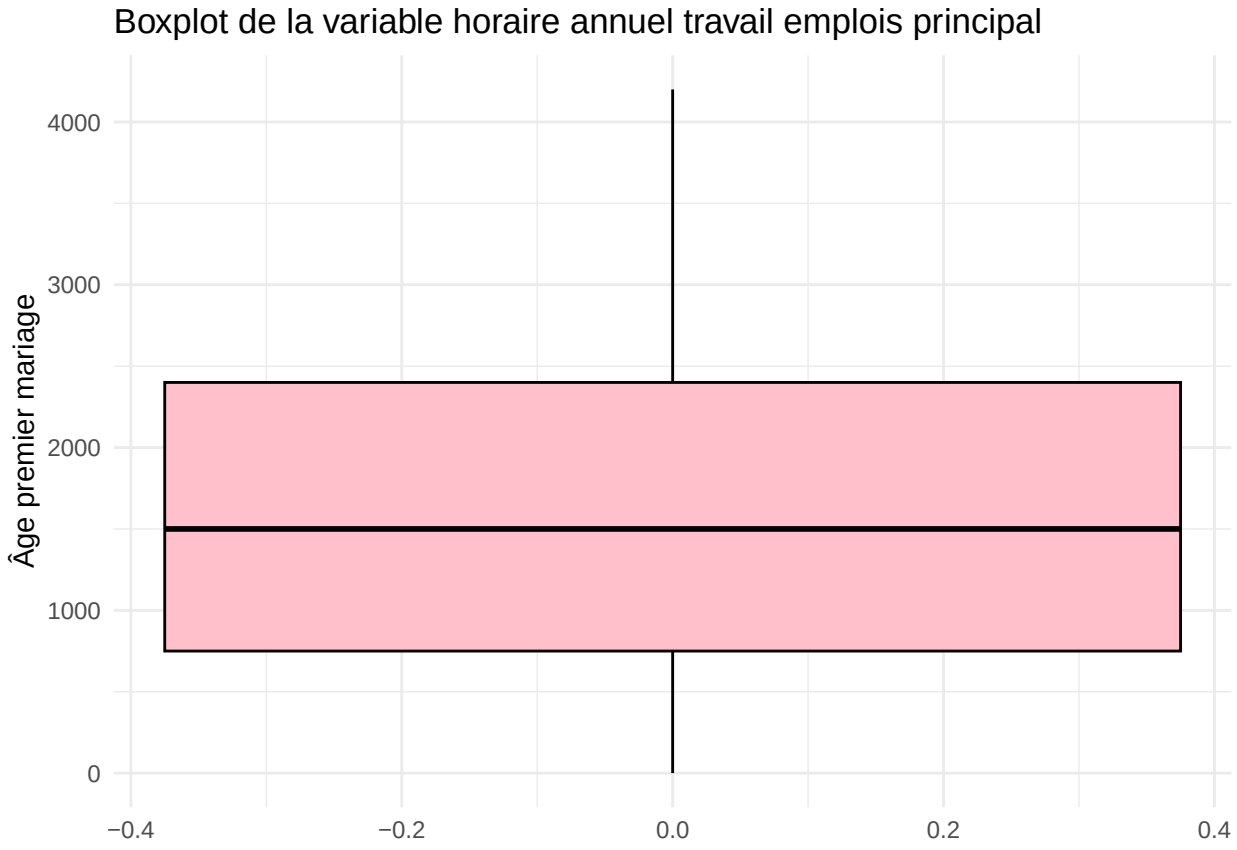


Table 4: **Résumé statistique de la variable `salaire annuel emplois principal`**

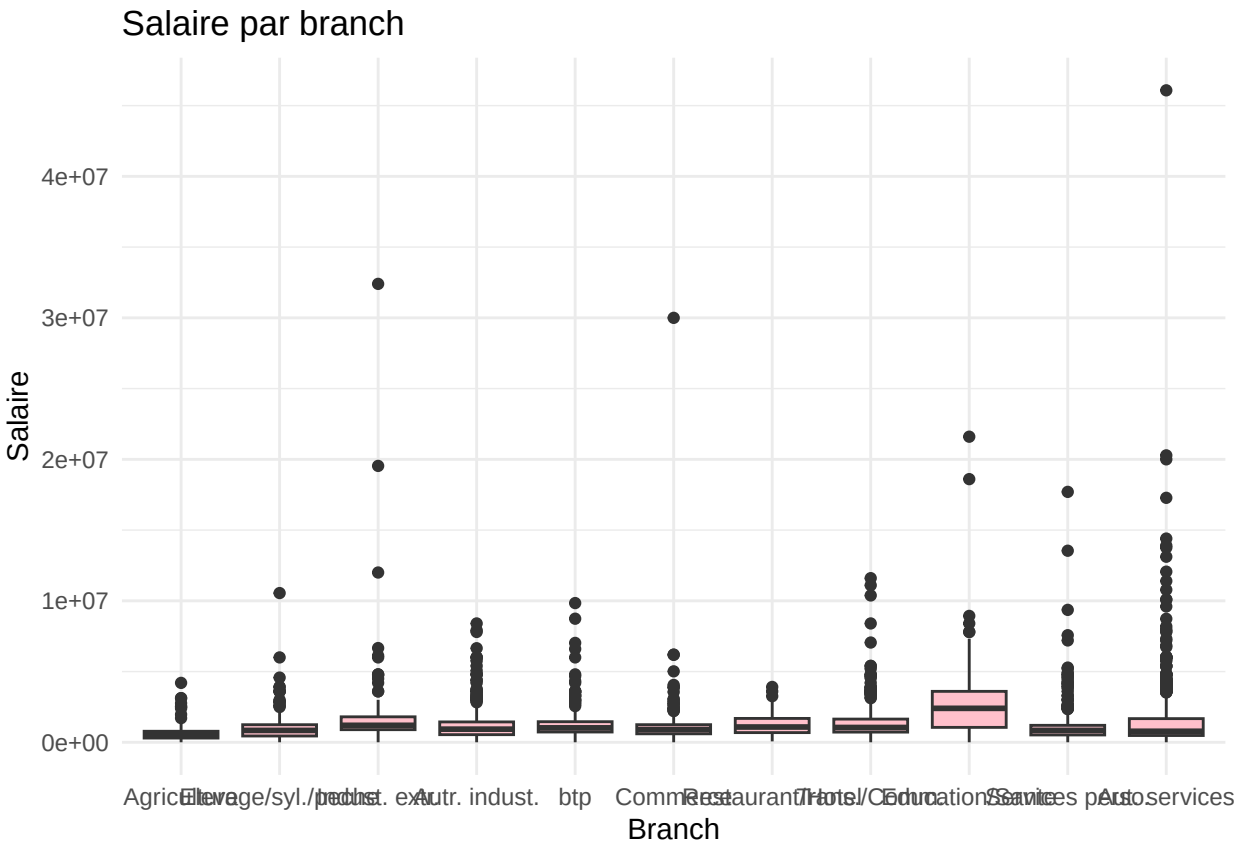
Statistique	**N = 63,530**
Salaire an. empl. prin.	
Moyenne = Mean	Moyenne = 1,366,140.0
Écart-type = SD	Écart-type = 1,705,367.9
Médiane = Median	Médiane = 936,000.0
Min = Min	Min = 0.0
Q1 = Q1	Q1 = 576,000.0
Q3 = Q3	Q3 = 1,631,152.6
Max = Max	Max = 46,080,000.0
Unknown	58,232

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

- **Variable salaire** :

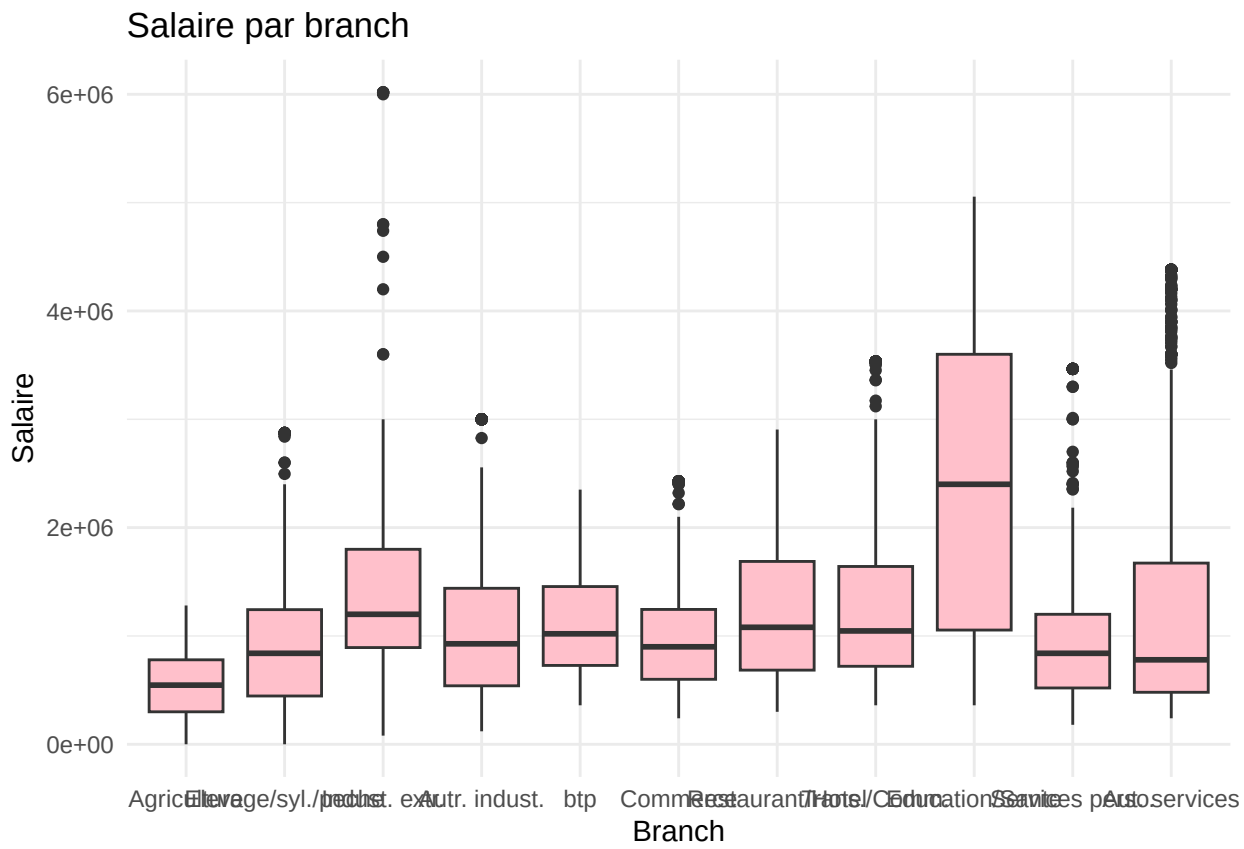
Le même procédé sera suivi pour la variable salaire. Voici un résumé statistique de la variable :

Pour visualiser les éventuelles valeurs aberrantes, nous allons afficher les boxplots de la variable salaire par branche d'activité.



On se rend compte qu'il y a des effectivement des valeurs aberrantes. Bien que celapuisse s'epliquer par la spécifité de la région, pour ne pas que ces valeurs influencent les données et perdre des informations, 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 :




```
## # A tibble: 11 x 5
##    branch           min_salaire   p05     p95 max_salaire
##    <fct>                  <dbl> <dbl>   <dbl>       <dbl>
##  1 Agriculture                0     0 1269937.    1281210.
##  2 Elevage/syl./peche         0     0 2869800     2874000
##  3 Indust. extr.          80000 80000 6002700     6018000
##  4 Autr. indust.         120000 120000 3000000    3000000
##  5 btp                   360000 360000 2349720    2350800
##  6 Commerce              240000 240000 2410908.   2428098.
##  7 Restaurant/Hotel      300000 300000 2794125    2905500
##  8 Trans./Comm.          360000 360000 3527892.   3534018.
##  9 Education/Sante       360000 360000 5043750    5055000
## 10 Services perso.       180000 180000 3390750.   3465000.
## 11 Aut. services         240000 240000 4380000    4380000
```

On se rend compte que les valeurs extrêmes ont disparues.

La même procédure sera suivie pour les variables **volhor_sec** et **salaire_sec**.

Chapitre 3 : Automatisation du traitement d'une base de données

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

Conclusion

En conclusion, l'utilisation de *R* dans le traitement des données d'enquêtes revêt une importance capitale pour les ingénieurs statisticiens économistes formés à l'ENSAE de Dakar. Le logiciel permet de traiter efficacement de grandes quantités de données, de garantir la qualité des résultats et d'assurer une analyse approfondie et rigoureuse. Grâce à ses capacités d'automatisation via *R Shiny, il facilite la gestion des bases de données et réduit les erreurs humaines. Ce travail souligne donc l'impact crucial de **R* dans la statistique appliquée, tout en mettant en évidence son rôle central dans la formation des futurs professionnels du domaine. Ce logiciel devient ainsi un allié indispensable dans le traitement des enquêtes et l'analyse des données en général, apportant efficacité et précision aux analyses statistiques.