|  |
| --- |
| **REPUBLIQUE DU SENEGAL** |

|  |
| --- |
|  |

|  |
| --- |
| **\*\*\*\*\*\*\*\*\*\*** |
| ***Un Peuple - Un But - Une Foi*** |
| **\*\*\*\*\*\*\*\*\*\*** |
| **Agence nationale de la Statistique et de la démographie** |

|  |
| --- |
|  |

|  |
| --- |
| **\*\*\*\*\*\*\*\*\*\*** |
| **Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye** |

|  |
| --- |
|  |

|  |
| --- |
| **ÉVALUATION Projet statistique avec R** |

|  |  |
| --- | --- |
| **Rédigé par** | **Sous la supervision de** |
| Malick SENE | M. Aboubacar HEMA |
| *Élève Ingénieure Statisticienne Économiste* | *Data scientist* |

|  |
| --- |
| **Année académique 2024-2025** |

# Table des matières

# Liste des figures

# Liste des tables

# Partie 1 : Gestion et nettoyage des bases de données

## Vérification des doublons

Nous commencons par importer les library necessaires ainsi que les données.

Maintenant, nous allons vérifier s’il existe des doublons dans les bases suivantes : - food\_comp\_child\_baseline - food\_comp\_child\_endline - base\_menage

Le nombre de doublons dans la base Baseline est 0.

Le nombre de doublons dans la base Endline est 0.

Le nombre de doublons dans la base Ménage est 1.

Nous procédons donc maintenant à la suppression des doublons dans la base ménage

Après suppression de l’unique doublon dans la base Ménage , on retrouve alors que :

Le nombre de doublons dans la base Ménage est 0.

## Harmonisation des bases

Certaines variables dans la base food\_comp\_mother\_baseline n’avaient pas des noms explicites, bien que leurs labels soient identiques à ceux des autres bases.

Nous avons donc corrigé les noms pour assurer l’homogénéité entre les bases Baseline et Endline. Il sagit des variables energ\_kcal(nommé V1 dans la base baseline mother) et zinc\_mg(nommé V9 dans la base baseline mother)

Nous vérifions maintenant si les noms de variables dans les bases food\_comp\_mother\_baseline et food\_comp\_mother\_endline sont identiques.

Pour cela nous faisons un test logique : Les noms de variables sont-ils identiques entre les deux bases mother ? True.  
Les noms de variables sont-ils identiques entre les deux bases child ? True.

## Gestion des données manquantes

Pour la correction des valeurs manquantes, nous appliquons la règle suivante :

Si l’individu **n’a pas consommé de repas** (s1\_q1 == 0), alors toutes les valeurs de consommation nutritionnelle doivent être mises à 0 (et non laissées à NA).

La correction a été appliquée à toutes les bases. Désormais, lorsqu’un repas n’a pas été consommé (s1\_q1 == 0), toutes les variables nutritionnelles sont systématiquement mises à 0 si elles étaient manquantes.

## Consommation d’énergie moyenne par repas – Mères, enquête de base

Nous calculons la moyenne de la variable energ\_kcal pour les mères ayant effectivement consommé un repas lors de l’enquête de base (food\_comp\_mother\_baseline), c’est-à-dire les observations où s1\_q1 == 1.

La consommation d’énergie moyenne par repas pour les mères lors de l’enquête de base est de 714.72 kilocalories.

## Sauvegarde des bases corrigées

Nous sauvegardons maintenant toutes les bases de données ayant fait l’objet de corrections sous de nouveaux noms de fichiers .dta.

Toutes les bases ont été sauvegardées dans le dossier data/ avec des noms reflétant les versions corrigées.

# Partie 2 : Empilement et Fusion des données

## Baseline

### Empilage des bases

Nous créons une base unique en empilant les données de consommation journalière des mères et des enfants pendant l’enquête de base. Une variable groupe permet d’identifier la source de chaque observation.

La base conso\_baseline\_empilee contient désormais les consommations des mères et des enfants, avec 8512 observations au total.

### Renommage des variables nutritionnelles – Suffixe \_b (Baseline)

Pour identifier clairement les variables issues de l’enquête de **Baseline**, nous avons renommé toutes les variables de consommation nutritionnelle en leur ajoutant le suffixe **\_b**.

Cela concerne les variables suivantes : energ\_kcal, protein\_g, lipid\_tot\_g, calcium\_mg, iron\_mg, zinc\_mg, vit\_b6\_mg, vit\_b12\_mcg, vit\_c\_mg

Elles ont été renommées respectivement en :  
energ\_kcal\_b, protein\_g\_b, lipid\_tot\_g\_b, …, vit\_c\_mg\_b.

Nous avons utilisé la fonction rename\_with() du package **dplyr**, qui permet de modifier plusieurs noms de variables à la fois.

La fonction paste0() a été utilisée pour **concaténer automatiquement** le nom original de chaque variable avec le suffixe \_b, sans ajouter d’espace.

Exemple : paste0("energ\_kcal", "\_b") retourne "energ\_kcal\_b".

### Consommation journalière

Dans la base conso\_baseline\_empilee, chaque individu (mère ou enfant) est observé jusqu’à **quatre fois par jour**, chaque observation représentant un **repas consommé** dans la journée.

Les données de consommation nutritionnelle sont donc **réparties sur plusieurs lignes** pour un même individu.

L’objectif ici est de :  
- **Calculer la consommation totale journalière** en additionnant les apports en énergie et en nutriments **sur l’ensemble des repas**,  
- Et de **rejoindre** cette information à la base initiale, de façon à ce que chaque ligne de repas contienne **à la fois l’information propre à ce repas** et **les totaux journaliers**.

Nous procédons ainsi:

* Nous utilisons group\_by(hhid, s1\_q2) pour **regrouper les lignes correspondant à un même individu** (même ménage, même groupe mère/enfant).
* Ensuite, nous utilisons summarise() + across() pour calculer la **somme** de chaque nutriment (les colonnes terminant par \_b), et créer de nouvelles colonnes comme energ\_kcal\_b\_jour.
* À l’aide de la fonction merge(), nous joignons les totaux journaliers à la base conso\_baseline\_empilee en utilisant les clés communes : hhid, s1\_q2.

### Export de la base finale – baseline\_final.dta

Nous devons créer une base baseline\_final.dta qui contient, pour chaque individu (mère ou enfant), **les apports nutritionnels journaliers totaux**.  
Elle doit inclure uniquement les variables suivantes :  
hhid, s1\_q2, energ\_kcal, protein\_g, lipid\_tot\_g, calcium\_mg, iron\_mg, zinc\_mg

Un ménage (hhid) doit avoir :  
- **une seule ligne mère**, identifiée par s1\_q2,  
- **plusieurs lignes enfants**, chaque enfant ayant son propre s1\_q2.

Nous avons donc **choisi de conserver une ligne par individu**, ce qui implique :  
- une seule ligne pour la mère,  
- une ou plusieurs lignes pour les enfants d’un même ménage.

Méthode utilisée :

1. Sélectionner les colonnes contenant les **totaux journaliers** (\*\_b\_jour)
2. Supprimer les doublons (une ligne par hhid + s1\_q2)
3. Renommer les colonnes comme demandé
4. Exporter la base au format .dta

## Endline

Nous rassemblons les données issues de l’enquête **Endline** en une seule base regroupant : - les consommations journalières des **mères**, - et celles des **enfants**.

Une variable groupe est ajoutée pour distinguer les deux, et les variables nutritionnelles sont renommées avec le suffixe \_e pour indiquer qu’elles proviennent de l’enquête Endline.

1. On ajoute une colonne groupe dans chaque base.
2. On empile les deux bases avec bind\_rows().
3. On renomme toutes les variables nutritionnelles de energ\_kcal à vit\_c\_mg avec rename\_with() en ajoutant le suffixe \_e.

### Résumé des consommations journalières par individu – Endline

Nous souhaitons créer une base résumant, pour chaque individu, la **consommation journalière totale** (somme des 4 repas) pour tous les nutriments.  
Cette étape correspond à la commande **collapse** dans Stata.

1. On utilise group\_by() pour regrouper les lignes correspondant au même individu (hhid, s1\_q2).
2. On applique summarise() avec sum(..., na.rm = TRUE) pour calculer les apports journaliers.
3. Le nom de chaque variable nutritionnelle sera suffixé avec \_e\_jour pour indiquer qu’il s’agit de la somme journalière Endline.

### Export de la base finale – endline\_final.dta

Nous extrayons une base finale contenant :  
- une ligne par **individu** (mère ou enfant),  
- avec les **consommations journalières totales** pour les principaux nutriments,  
- et uniquement les variables demandées : hhid, s1\_q2, energ\_kcal, protein\_g, lipid\_tot\_g, calcium\_mg, iron\_mg, zinc\_mg.

Méthode utilisée :

1. Nous sélectionnons les variables utiles de resume\_journalier\_endline.
2. Nous les **renommons** pour enlever le suffixe \_e\_jour.
3. Nous exportons la base finale au format .dta.

Il n’est pas nécessaire d’utiliser distinct() ici car la base resume\_journalier\_endline a déjà été construite à partir d’un group\_by(hhid, s1\_q2) suivi d’un summarise().  
Cela garantit qu’il y a **exactement une seule ligne par individu**, ce qui respecte pleinement la structure attendue de la base finale.

## Fusion des bases Baseline et Endline – Par individu

Nous fusionnons les bases baseline\_final.dta et endline\_final.dta de sorte que chaque individu soit représenté par **une seule ligne**, contenant :  
- ses **informations Baseline** (suffixe \_b)  
- ses **informations Endline** (suffixe \_e)

La fusion est réalisée sur les variables hhid et s1\_q2, qui identifient chaque individu de manière unique.

Étapes de la fusion :

1. Charger les deux bases depuis le dossier data/.
2. Renommer les colonnes nutritionnelles pour éviter l’écrasement.
3. Fusionner les deux bases avec merge() en conservant tous les individus.

Nous vérifions que la base fusionnée contient exactement le même nombre d’observations que les bases baseline\_final.dta et endline\_final.dta.

Nombre d’observations dans la base Baseline : 2128

Nombre d’observations dans la base Endline : 2128

Nombre d’observations dans la base fusionnée : 2128

## Fusion avec les caractéristiques des ménages

Nous completons la base fusion\_finale en y ajoutant les **caractéristiques socio-économiques du ménage**, extraites de la base base\_menage\_final.

Les variables intégrées sont :  
- Région, commune, village (regionid, communeid, villageid)  
- Statut matrimonial (poly)  
- Taille du ménage (hhsize)  
- Niveau d’éducation du chef de ménage (hh\_primary)  
- Ratio de dépendance (dependencyratio)  
- Score HFIAS (hfias\_score)  
- Variable de traitement (T1)

Etape de la fusion :

1. Sélection des variables pertinentes dans base\_menage\_final
2. Fusion par la variable hhid avec fusion\_finale

## Sauvegarde de la base finale complète

La base fusion\_complete, contenant à la fois : - les apports nutritionnels issus des enquêtes **Baseline** et **Endline**, - et les **caractéristiques socio-économiques** du ménage, est maintenant sauvegardée sous le nom **base\_finale.dta**.

# Partie 3 : Statistiques descriptives avec le package gtsummary

## Tableau 1 : Apports nutritionnels moyens à Baseline et Endline

Ce tableau présente les **moyennes et écarts-types** des principaux nutriments consommés quotidiennement par les individus,  
au moment de l’enquête **Baseline** (début de l’étude) et **Endline** (fin de l’étude).  
Les indicateurs analysés incluent : l’énergie (en kcal), les protéines, les lipides, le calcium, le fer et le zinc.

Les variables de la Baseline sont préfixées par \*\*\_b**, celles de l’Endline par** \_e**.**  
**Les résultats sont exprimés sous la forme :** moyenne ± écart-type\*\*

Table : Apports nutritionnels moyens à Baseline et Endline

| Nutriment | **N = 2,128**1 |
| --- | --- |
| Énergie (kcal) - Baseline | 1,954 ± 1,174 |
| Protéines (g) - Baseline | 61 ± 60 |
| Lipides (g) - Baseline | 34 ± 41 |
| Calcium (mg) - Baseline | 338 ± 346 |
| Fer (mg) - Baseline | 25 ± 23 |
| Zinc (mg) - Baseline | 9.9 ± 7.0 |
| Énergie (kcal) - Endline | 1,967 ± 1,194 |
| Protéines (g) - Endline | 61 ± 58 |
| Lipides (g) - Endline | 35 ± 41 |
| Calcium (mg) - Endline | 352 ± 392 |
| Fer (mg) - Endline | 23 ± 21 |
| Zinc (mg) - Endline | 9.6 ± 6.8 |
| 1Mean ± SD | |

## Tableau 2 : Apports nutritionnels moyens à Endline selon la variable de traitement (T1)

Ce tableau compare les **apports nutritionnels moyens à Endline** selon la variable T1,  
qui indique si un individu appartient au **groupe de traitement** ou au **groupe témoin** dans l’étude.

Les indicateurs nutritionnels analysés sont : - l’énergie en kilocalories (kcal), - les protéines en grammes (g), - et les lipides en grammes (g).

Les résultats sont exprimés sous la forme **moyenne ± écart-type**, pour chaque groupe.

Table : Apports nutritionnels à Endline selon le traitement (T1)

| Nutriment | **Controle**  N = 7121 | **Intervention**  N = 1,4161 |
| --- | --- | --- |
| Énergie (kcal) – Endline | 1,874 ± 1,141 | 2,014 ± 1,217 |
| Protéines (g) – Endline | 58 ± 41 | 62 ± 66 |
| Lipides (g) – Endline | 33 ± 37 | 37 ± 43 |
| 1Mean ± SD | | |

## Tableau 3 : Apports nutritionnels à Endline selon le statut matrimonial du ménage

Ce tableau présente les **apports nutritionnels moyens à Endline** selon le **statut matrimonial** du ménage,  
indiqué par la variable poly, qui distingue les **ménages polygames** des **ménages monogames**.

Les apports sont exprimés en **moyenne ± écart-type** pour chaque groupe.

Table : Apports nutritionnels à Endline selon le statut matrimonial (polygamie)

| Nutriment | **Non**  N = 1,1081 | **Oui**  N = 1,0201 |
| --- | --- | --- |
| Énergie (kcal) – Endline | 1,948 ± 1,207 | 1,988 ± 1,180 |
| Protéines (g) – Endline | 61 ± 67 | 61 ± 48 |
| Lipides (g) – Endline | 35 ± 43 | 36 ± 39 |
| 1Mean ± SD | | |

## Tableau 4 : Caractéristiques des ménages

Ce tableau présente des **statistiques descriptives** sur les principales caractéristiques des ménages de l’échantillon.  
Les indicateurs incluent :

* La **taille du ménage** (hhsize)
* Le **niveau d’éducation primaire complété** par le chef de ménage (hh\_primary)
* Le **ratio de dépendance** (dependencyratio)
* Le **score HFIAS** (hfias\_score), mesurant la sécurité alimentaire du ménage.

Les statistiques sont exprimées en **moyenne ± écart-type**.

Table : Caractéristiques des ménages

| Caractéristique du ménage | **N = 2,128**1 |
| --- | --- |
| Taille du ménage | 8.9 ± 4.6 |
| Chef de ménage a complété le primaire |  |
| Non | 1,990 (94%) |
| Oui | 138 (6.5%) |
| Ratio de dépendance | 1.37 ± 0.65 |
| Score HFIAS (0 à 27) | 3.2 ± 4.4 |
| 1Mean ± SD; n (%) | |

## Tableau 5 : Apports nutritionnels moyens selon le niveau d’insécurité alimentaire (HFIAS)

Ce tableau présente les **apports nutritionnels moyens à Endline** en fonction du **niveau d’insécurité alimentaire**  
des ménages, mesuré par le **score HFIAS** (Household Food Insecurity Access Scale).

Les ménages ont été classés en trois groupes : - **Faible insécurité alimentaire** : score HFIAS ≤ 9 - **Modérée** : entre 10 et 18 - **Élevée** : > 18

Les apports en énergie, protéines et lipides sont présentés en **moyenne ± écart-type** pour chaque groupe.

Table : Apports nutritionnels selon le niveau d’insécurité alimentaire

| Nutriment | **Faible**  N = 1,9161 | **Modérée**  N = 1981 | **Élevée**  N = 141 |
| --- | --- | --- | --- |
| Énergie (kcal) – Endline | 1,975 ± 1,193 | 1,909 ± 1,212 | 1,658 ± 1,035 |
| Protéines (g) – Endline | 61 ± 59 | 63 ± 50 | 46 ± 31 |
| Lipides (g) – Endline | 36 ± 41 | 35 ± 40 | 24 ± 22 |
| 1Mean ± SD | | | |