Step 1: Data Standardization and Missing Data Analysis

Before any imputation could begin, we needed to understand the extent and patterns of missing data in the dataset.

Why It Matters

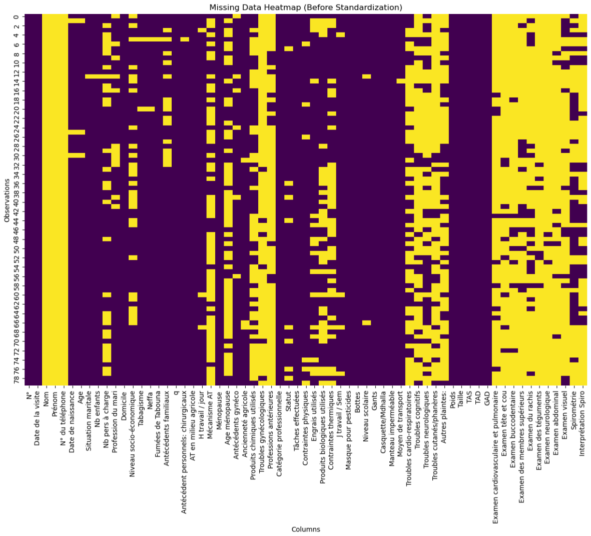
Inconsistent missing value representations obscured the true extent of data gaps, complicating analysis and imputation. Understanding missingness patterns and mechanisms was critical to choosing appropriate imputation methods and avoiding biased results.

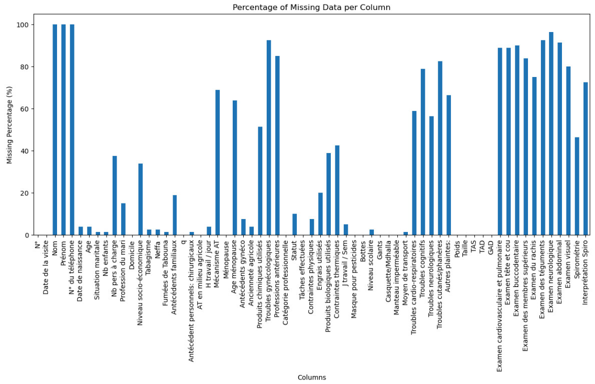
Methodology

We implemented a three-part approach:

1. **Standardizing Missing Values**
    - Used pandas to replace all variants ('non spécifié', '#VALUE!', '-', etc.) with NaN.
    - This standardization enabled accurate quantification and visualization of missing data patterns.
2. **Visualizing Missing Patterns**
3. **Analyzing Missing Mechanisms**

Analysis of Missing Data Visualizations(Before removing the descriptive text columns)

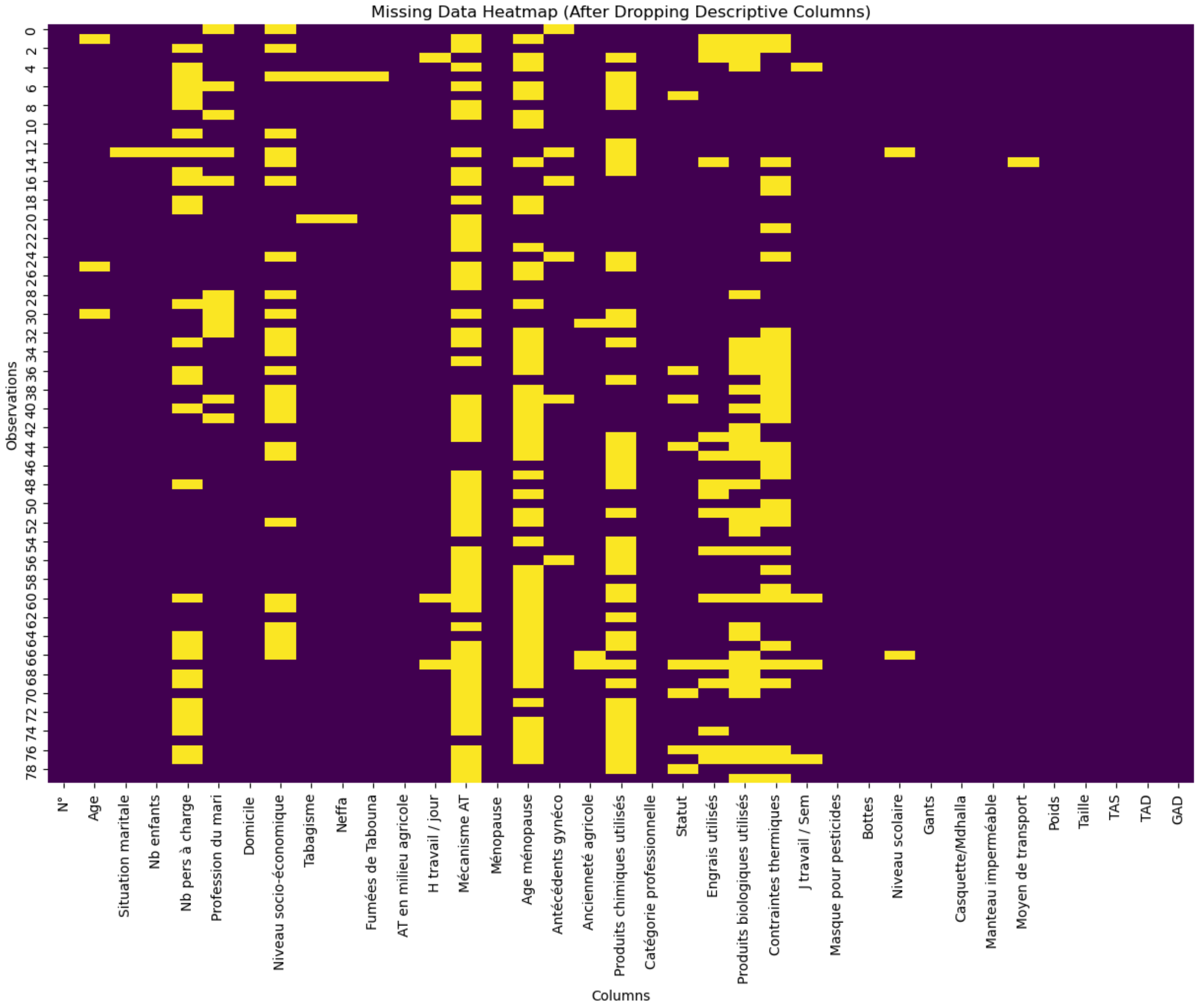




Analysis of Missing Data Visualizations(After removing the descriptive text columns)

The visualizations provide clear insight into the missing data patterns in the female farmers dataset after the removal of descriptive text columns:

Heatmap Analysis

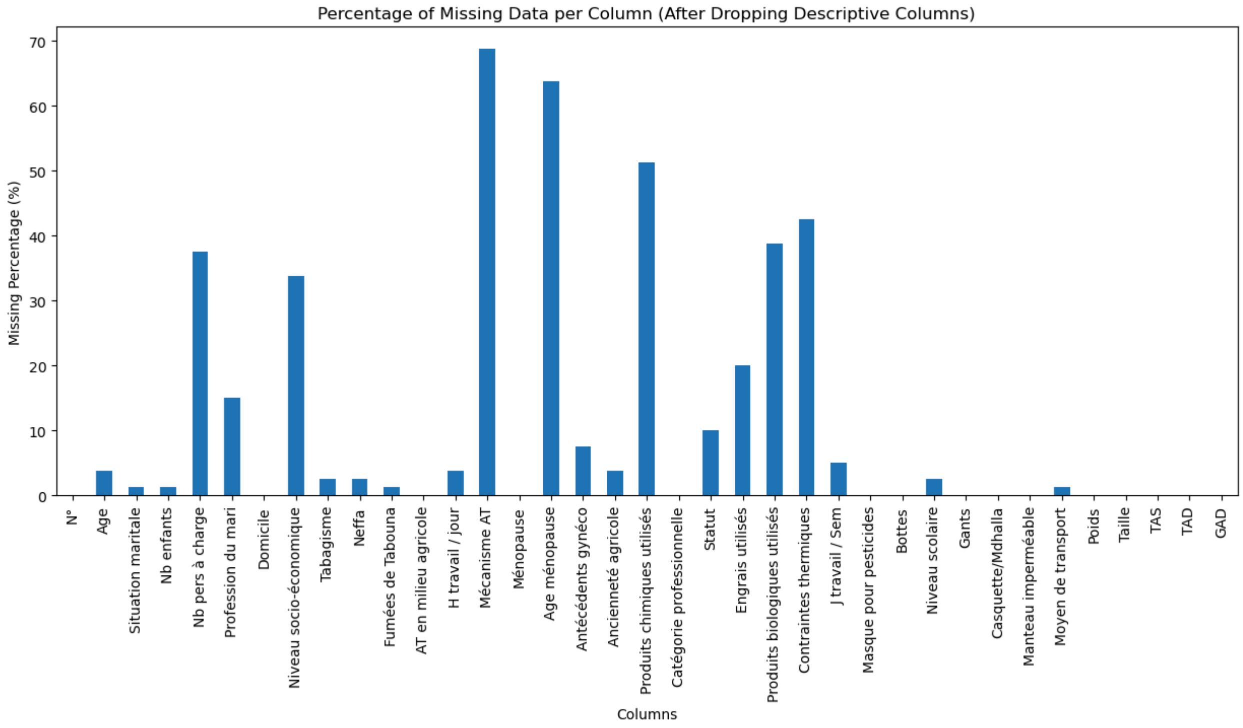


The missing data heatmap reveals structured patterns rather than random distribution:

- **Right side completeness**: Essential health metrics like weight (`Poids`), height (`Taille`), blood pressure measurements (`TAS`, `TAD`, `GAD`) have almost no missing values (appear as solid purple columns)
- **Structured missingness**: Certain variables show clear patterns where missing values (yellow) cluster together, suggesting systematic rather than random missingness
- **Variable-specific patterns**: Some variables like `Age ménopause` and `Antécédents gynéco` show high levels of missingness that appear related to specific participant characteristics

- **Observation patterns**: Some participants (rows) have more missing data across multiple fields, suggesting potential issues with specific data collection sessions

Bar Chart Analysis



The percentage bar chart quantifies the missingness by variable:

- **High missingness variables**:
  - `Age ménopause` (~69%)
  - `Antécédents gynéco` (~64%)
  - `Produits biologiques utilisés` (~51%)
- **Moderate missingness**:
  - `Nb pers à charge` (~38%)
  - `Niveau socio-économique` (~34%)
  - `Produits chimiques utilisés` (~39%)

    - `Contraintes thermiques` (~43%)

- **Low missingness**:
  Most demographic and clinical measurement variables show less than 5% missing values

Key Insights

1. The missing data follows clear patterns that suggest certain variables were collected only under specific conditions (e.g., `Age ménopause` was likely only recorded for menopausal women)
2. Essential health metrics and demographic information are nearly complete, providing a solid foundation for analysis
3. The moderate-to-high missingness in work exposure and environmental factors suggests these may have been optional components of data collection
4. The relatively complete nature of the protective equipment usage variables (`Masque pour pesticides`, `Bottes`, `Gants`) will allow for reliable analysis of protective behaviors

These visualizations confirm that removing the descriptive text columns was appropriate, as they allow for a clearer focus on variables with sufficient data for meaningful analysis.

Implications for Further Analysis

The standardization process successfully converted inconsistent missing representations to a unified format, enabling us to see that:

1. The missingness isn't random but follows meaningful patterns related to the data collection process
2. Essential variables for analyzing worker profiles and basic health metrics are largely complete
3. Variables with high missingness appear to follow a Missing Not At Random (MNAR) pattern, particularly health conditions and specific exposures