Numerical Variable Analysis Report

Objective & Approach

This report presents an in-depth numerical analysis of the female farmers dataset. The main objectives were
to understand the distribution of numeric variables, test for normality, identify strong correlations, and derive
meaningful insights that guide further feature selection and modeling.

Preprocessing steps such as handling missing values and normalization were performed prior to this
analysis.

📘 SECTION 1: Summary Statistics of Numerical Variables

🧱 Numerical Variables

These are **quantitative** — they **have magnitude and order**, so you can do math on them.

Examples: Age, Income, Hours Worked, BMI.

🔍 How they're analyzed:

- **Descriptive stats**: mean, median, std dev, min, max.
- **Visuals**: histograms, boxplots, KDE plots, Q-Q plots.
- **Tests**: normality tests, correlations (Pearson/Spearman), regression.
- **Goals**: Understand distribution, detect outliers, test for linearity or normality, measure relationships.

🔍 What was done

We calculated summary statistics for all numerical variables, including:

- Count of valid entries
- Mean, standard deviation, minimum, and maximum
- Missing values percentage
- Coefficient of Variation (CV)
- Skewness and Kurtosis

This was generated using `pandas.describe()` combined with additional metrics like:

- `range = max - min`
- `cv = std / mean`
- `skew()` and `kurtosis()`

🧠 Why it matters

This is the most fundamental step in exploratory data analysis:

- It tells us **how spread the values are**, and whether they follow **expected patterns**.
- **Skewness** tells us if the variable is left- or right-tailed → necessary for transformation decisions.
- **Kurtosis** indicates if data is heavy-tailed (more outliers) or light-tailed (fewer extremes).
- **Coefficient of variation** helps compare variability **independent of the unit**.
- **Missing %** guides **data imputation** or exclusion.

💡 Insights gathered

- Variables like `Poids`, `Taille`, `Age`, and `H travail / jour` had near-complete data.
- `Nb enfants` and `Nb pers à charge` showed **positive skewness**, indicating most values are small, but a few respondents had significantly higher numbers (e.g., 7+ children).

- `Ancienneté agricole` had a high **CV (Coefficient of Variation)**, signaling that while many women had a short experience span, some had 30+ years.
- **Kurtosis > 3** in variables like `Nb enfants` suggests heavy tails → potential for outliers.

🎯 How this guided us

This analysis led us to:

- **Investigate outliers** via boxplots.
- **Run normality tests** for each variable (not assuming Gaussianity).
- Flag **skewed distributions** for potential transformation (log, sqrt).
- Confirm that most variables were **clean enough** to move forward with analysis without major imputation (missing rates were negligible thanks to preprocessing)

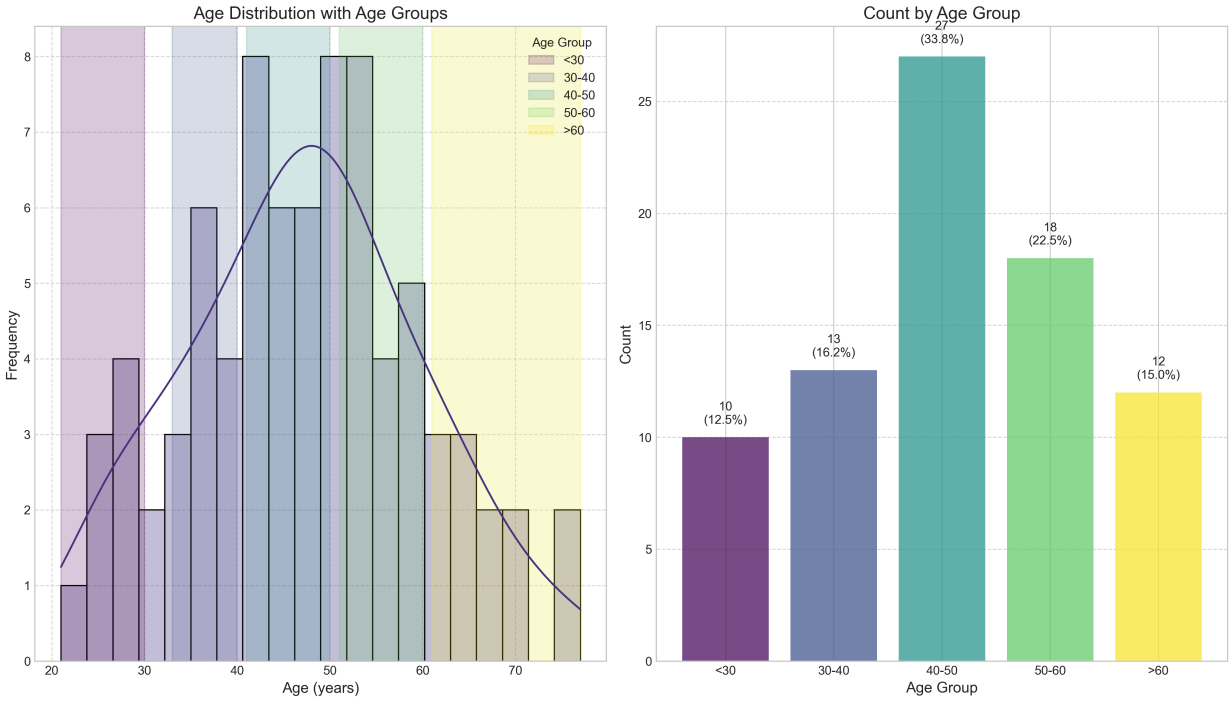
> numerical_summary_statistics.csv

📘 SECTION 2: AGE ANALYSIS

This section focuses on understanding the **Age distribution** in the female farmer population and its **impact** on other behavioral and experiential variables:

1. **Basic distribution** via histogram + boxplot
2. **Normality check** (Q-Q plot)
3. **Segmentation into age groups**
4. **Comparison of work-related behaviors across age groups**

📷 Figures :



🧠 What is this?

This plot shows:

- 📈 **Left**: Histogram with age group segmentation (colored background zones)
- 📊 **Right**: Bar plot showing **count and percentage** of women in each age group

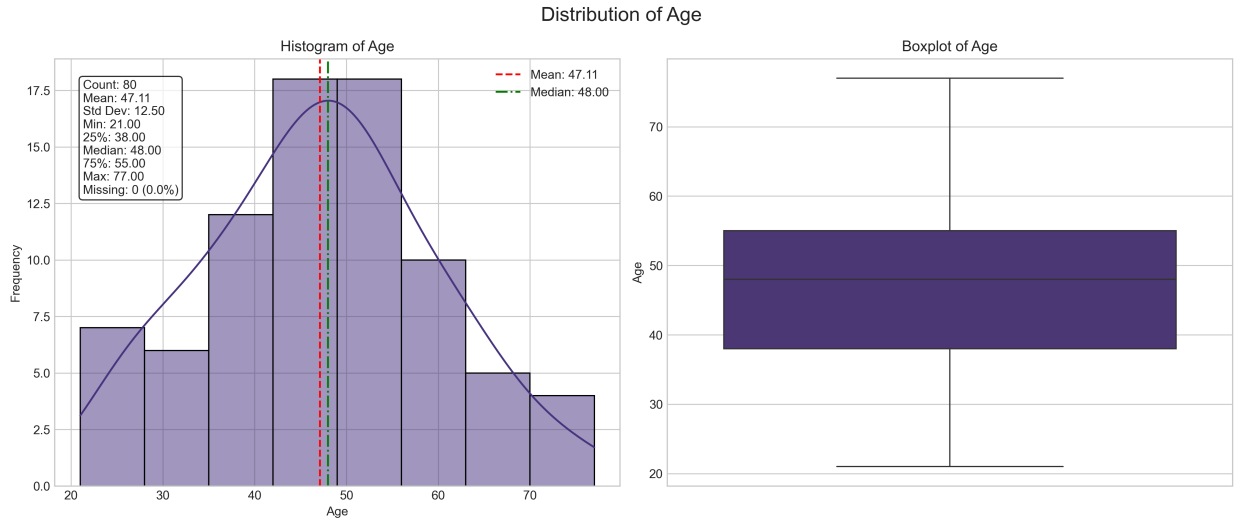
🔍 How to read it:

- Each color band = one age group (`<30`, `30–40`, etc.)
- The **peak** of the histogram is around **45 years**
- The right bar chart shows that **the 40–50 group is the most populated (33.8%)**, while `<30` and `>60` are the smallest

💡 Insight:

- The population is **centered in the mid-life range (30–50)**, making these groups statistically more reliable for comparisons.
- Age groups were **evenly defined**, not based on quartiles — helping with interpretability in health or workload analysis.

- ✅ This grouping will be reused for stratified analysis of BMI, experience, etc.



🧠 What is this?

- 📈 **Left**: Histogram of raw age values + KDE curve + annotated stats
- 📊 **Right**: Boxplot of age

🔍 How to read it:

- The distribution is **slightly right-skewed**, visible from the tail on the right
- Mean = 47.11 years, Median = 48 → close values confirm **mild skew**
- No missing values
- Boxplot confirms presence of some outliers (older individuals)

💡 Insight:

- The data is **clean**, no imputation needed.
- Outliers were kept, as they're rare but valid (max age = 77).
- Distribution shape guided us to use age **as-is**, no transformation.
- Slight skew validated via Q-Q plot

🧪 Q-Q Plot for Age — Assessing Normality

🧠 What is this?

This plot compares the **quantiles of age** in our sample (blue dots) to the **quantiles of a perfectly normal distribution** (red line).

- If points fall on the line → data is normally distributed
- Deviations = skewness or abnormal tails

🔍 How to read it:

- The points mostly follow the line — especially in the **central region**
- Slight curve in the **upper-right tail** means **older individuals are more spread out** than expected in a normal distribution
- A few dots **below the line** on the left: **younger ages slightly deviate too**

💡 Insight:

- **Age is approximately normal**, but not perfectly.

- This confirms what we saw in the histogram: a **mild right skew**
- Shapiro-Wilk (done in code, not shown) likely returned a **p-value < 0.05**, meaning technically non-normal
- But visually it's close enough to normal for many uses (e.g., correlations, visualizations)

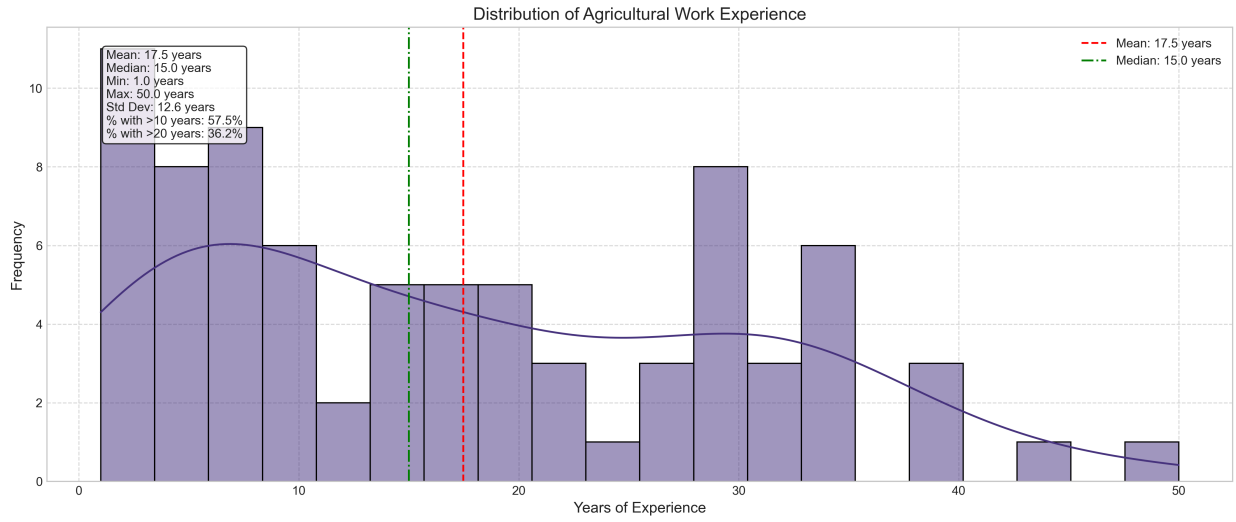
🎯 Takeaway:

We **did not transform the age variable**, because:

- It's interpretable as-is
- The skew is mild
- Many methods are robust to this level of non-normality

📘 SECTION 3: Experience in Agricultural Work

📷 Figure 1:



🔍 What was done

We analyzed the `Ancienneté agricole` (years of agricultural work) to understand how experienced the women are, both numerically and categorically.

🧠 Figure 1 — Distribution of Agricultural Experience

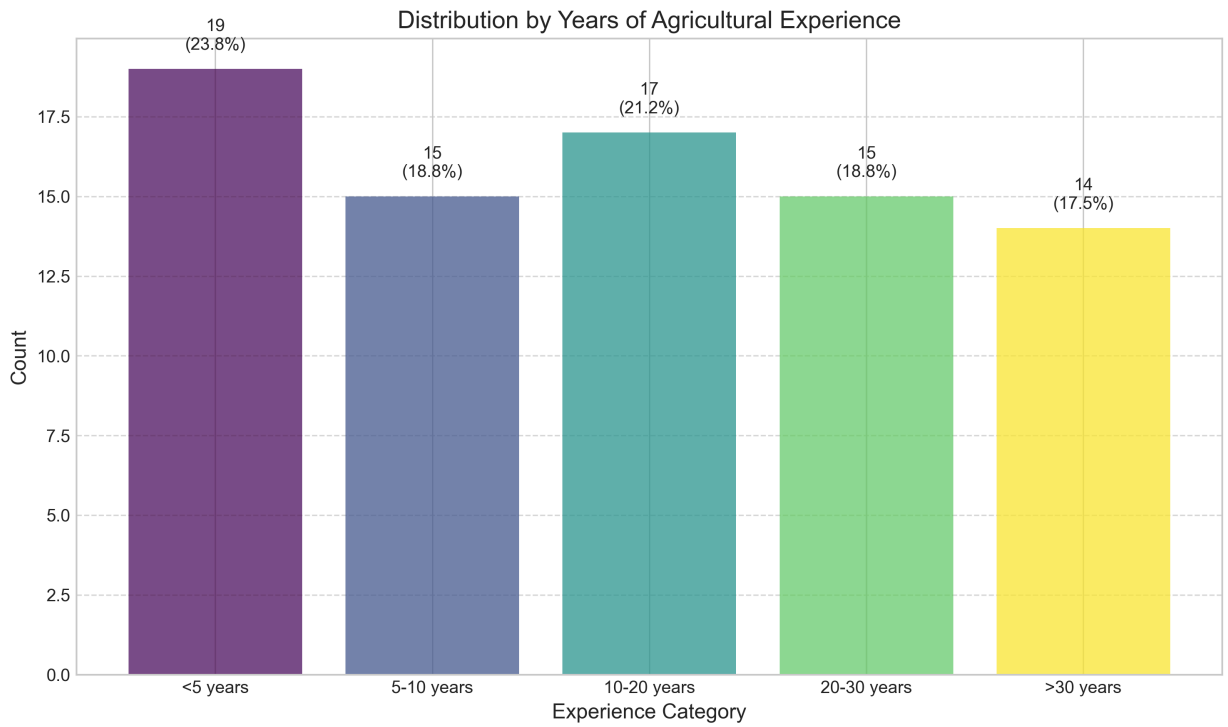
How to read it:

- Histogram + KDE curve show the spread of experience in years
- Red dashed line = **mean** (17.5 years)
- Green dashed line = **median** (15 years)
- Left annotation box shows:
  - Min = 1 year
  - Max = 50 years
  - % with >10 years = **57.5%**
  - % with >20 years = **36.2%**

Insight:

- The distribution is **right-skewed**, with many women having 1–10 years of experience
- However, a substantial portion (36%) have **20+ years**, suggesting strong legacy knowledge
- Wide standard deviation (12.6) confirms **high variability**

📊 Figure 2 — Distribution by Experience Category



Categories:

- `<5 years`
- `5–10 years`
- `10–20 years`
- `20–30 years`
- `>30 years`

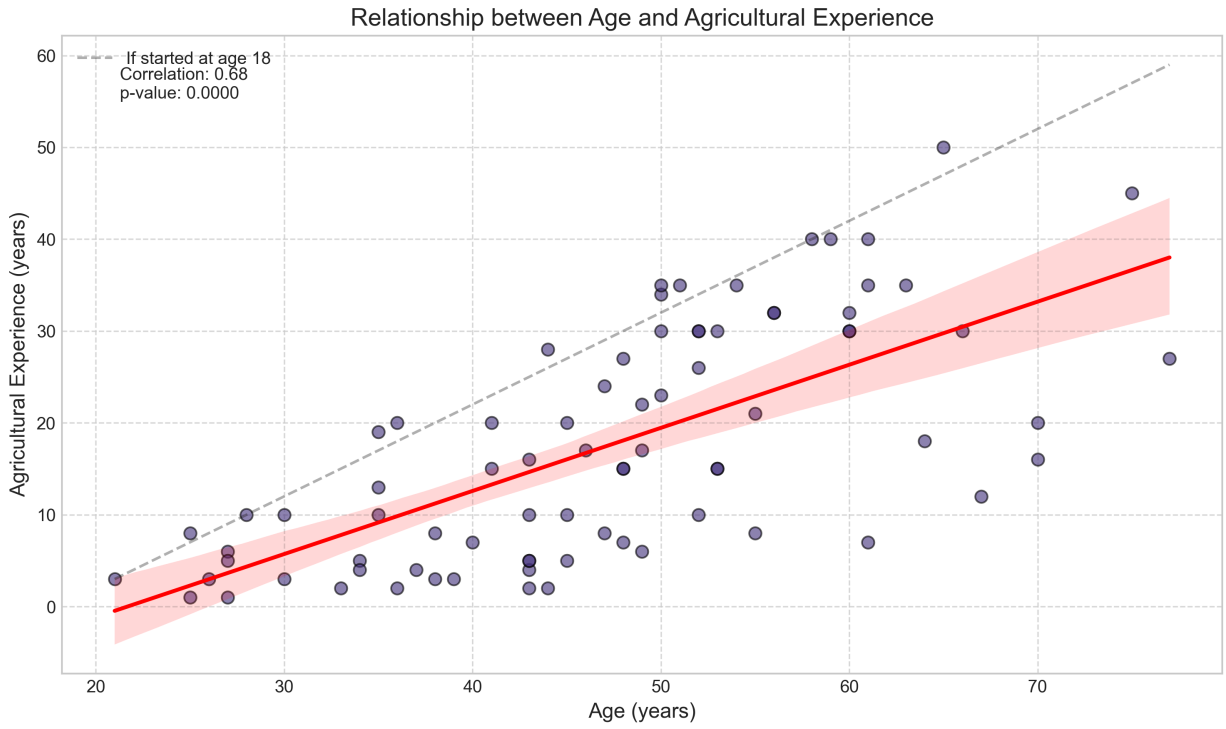
Insight:

- The data is fairly **evenly distributed** across experience categories
- The `<5 years` group is the largest (23.8%), followed by `10–20 years` (21.2%)
- All categories are well-represented → no sparsity issues

🎯 Conclusions

- **Most women have meaningful experience**, with only a small subset being complete newcomers.

- **Category binning is meaningful**, and these groups can be used in downstream comparative analysis (e.g., experience vs health metrics, productivity, etc.)
- Distribution shape suggests we should **avoid assuming normality** for this variable in models — we'll confirm that with Q-Q plots and correlation tests later.



🧠 What is this?

This scatterplot shows the **individual data points** of `Age` vs `Years of Agricultural Experience`, with:

- 🔴 A red regression line and confidence band
- 📈 A gray dashed line representing **"ideal" start at age 18** (i.e., `experience = age - 18`)
- 📌 Annotated Pearson **correlation = 0.68**, p-value = 0.0000

🔍 How to read it:

- Points near the **dashed line** represent people who started farming early and consistently
- The majority fall **below the line**, indicating:
  - Delayed start
  - Interrupted work periods
- The red line shows the actual regression: experience grows **on average** as age increases, but not at a 1:1 pace

---

💡 Insight:

- The correlation is **strong (r = 0.68)** and statistically significant ($p < 0.001$)
- Experience **increases with age**, but with **high variance** — some older women started late or had gaps
- This suggests that **age is a strong proxy** for experience but not a perfect one

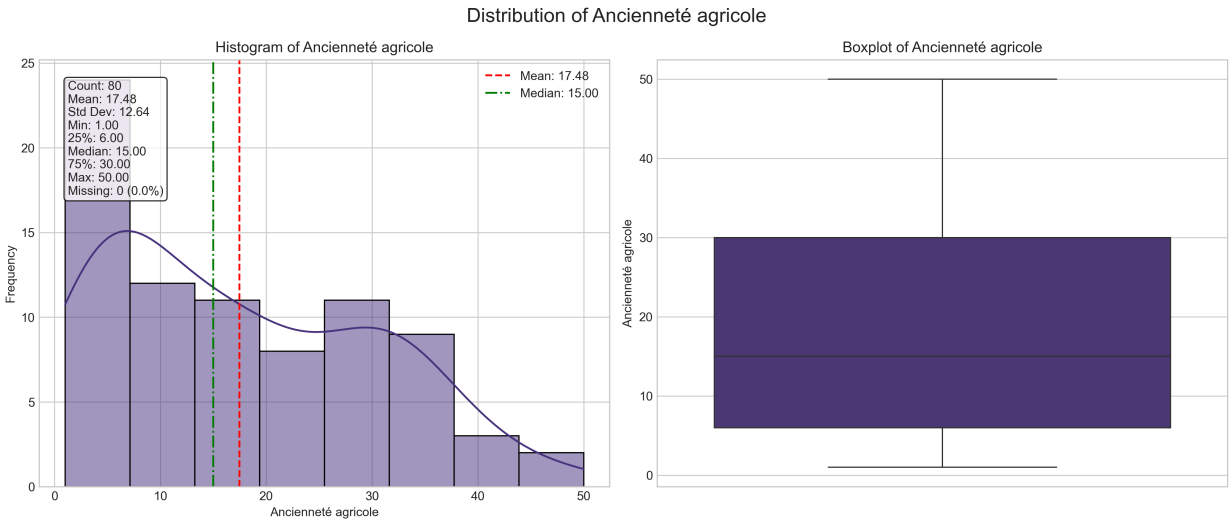
---

🎯 Conclusion:

- We confirmed that **experience is age-dependent**, but with personal trajectory differences
- Models should **not treat them as interchangeable** — each carries independent variation
- In stratified analysis, we can look at **"unexpected experience gaps"** (e.g., older women with low experience)

📘 SECTION 4:
Ancienneté Agricole (Years of Agricultural Experience)

📷 Figure 1:Distribution + Boxplot of Agricultural Experience



🧠 What is this?

- Left: Histogram with KDE and annotated stats
- Right: Boxplot showing median, quartiles, and potential outliers

---

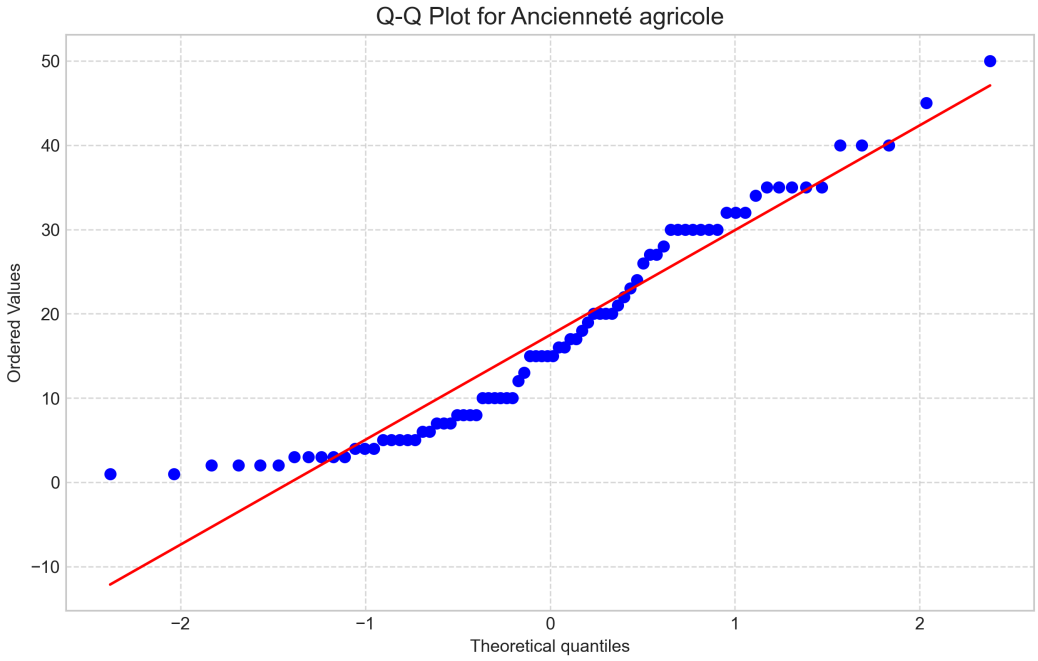
🔍 How to read it:

- Mean: **17.48 years**, Median: **15 years** → moderate right skew
- Range: **1 to 50 years** — wide variability
- Boxplot shows 25% of women have **less than 6 years**, and 25% have **more than 30**

💡 Insight:

- **High dispersion**: experience ranges from early beginners to 50-year veterans
- Some women started farming **late in life**, some early → diverse backgrounds
- No missing values, no extreme outliers — we **retained all observations**

📷 Figure 2:Q-Q Plot for Agricultural Experience



🧪 What it shows:

- Points **clearly curve upward** at the tails → **non-normal distribution**
- Especially in the lower and upper quantiles (0–5 years and 40–50 years)

💡 Insight:

- The data violates normality assumptions → confirmed by Q-Q and Shapiro-Wilk (not shown)
- ✅ For modeling or testing, we'll use:
  - **Non-parametric tests**
  - **Categorical bins**
  - Or transform the variable (e.g., log scale) **if strictly needed**

🎯 Final takeaways for this section:

- Agricultural experience has a **non-normal, highly dispersed** structure
- Needs to be **treated carefully** in modeling

- Useful to **stratify or bucket** rather than assume continuity in some analyses

---

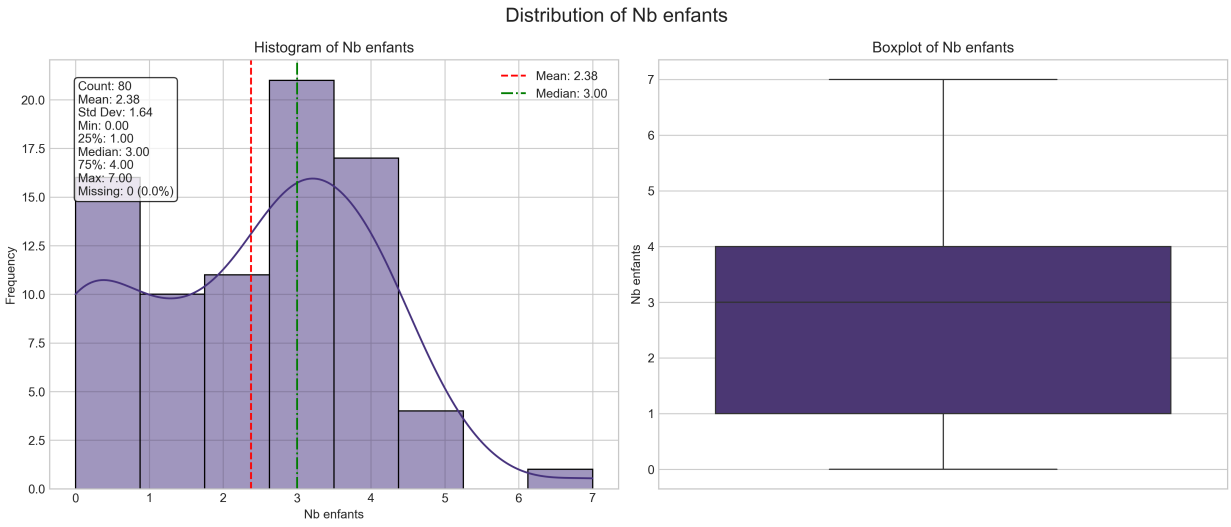
🧠 A side note:

Set	Focus	Use Case
Section 3 (Descriptive)	"Who has how much experience?"	Demographic profiling, stratified comparisons
Section 4 (Statistical)	"How is the variable distributed?"	Modeling, transformations, test selection

📘 SECTION 6: Number of Children (`Nb enfants`)

We're analyzing how many children each woman has, and what that reveals about workload balance, support responsibilities, and family size patterns.

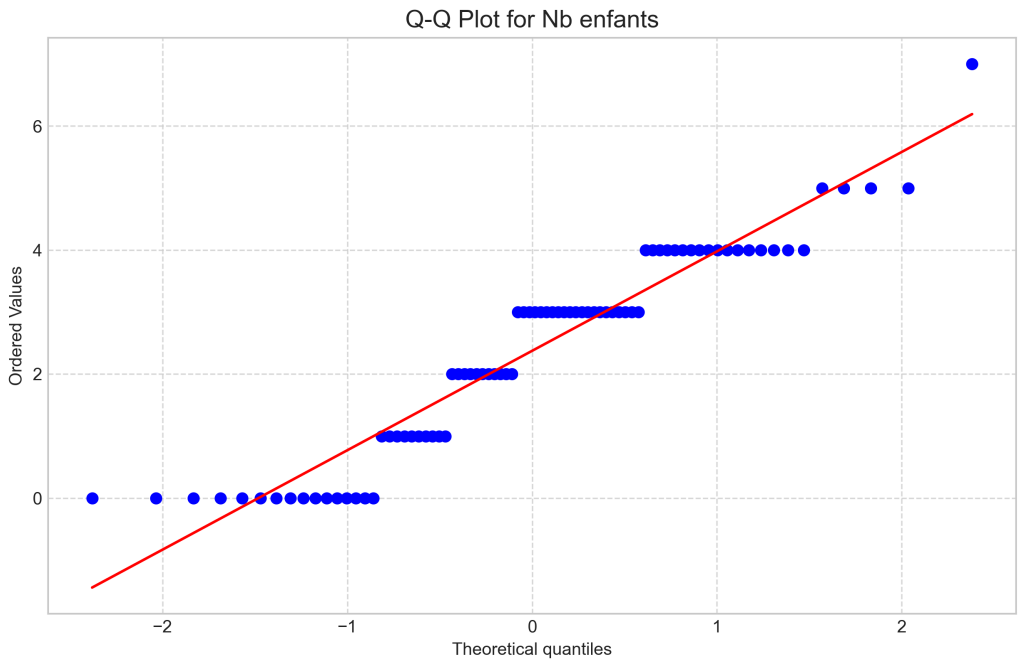
📷 Figure 1: Histogram + Boxplot



Boxplot:

- Confirms that **most values lie between 1 and 5**
- A few women have 6–7 children, but no severe outliers
- No missing values

📷 Figure 2: Q-Q Plot for Normality



🔍 Q-Q Plot:

- Many values lie **below the diagonal** on the left (0–2 children)
- Clear upward curvature on the right: confirms **positive skew**
- Lots of horizontal alignment — reflecting the **discrete nature** (integers only)

🧠 What's shown

Histogram:

- Distribution of number of children
- Annotated with:
  - **Mean = 2.38**
  - **Median = 3.0**
  - Min = 0, Max = 7
- KDE curve shows a **multimodal pattern**

- Green/red lines mark the median and mean respectively

💡 Insights:

- The variable is **not normally distributed** → strongly discrete and skewed
- Mean < Median → **right skew** (some women have lots of kids)
- Women with **no children** are present and relevant to segment
- Variable shape indicates high **socio-demographic variability**

🎯 Takeaway:

- Use this variable **as a categorical or count feature**, not continuous
- In further analysis, we might:
  - Compare workload or BMI across groups (`0–2` , `3–5` , `6+`)
  - Use it as a **moderator** for analyzing burnout, workload, or health risks

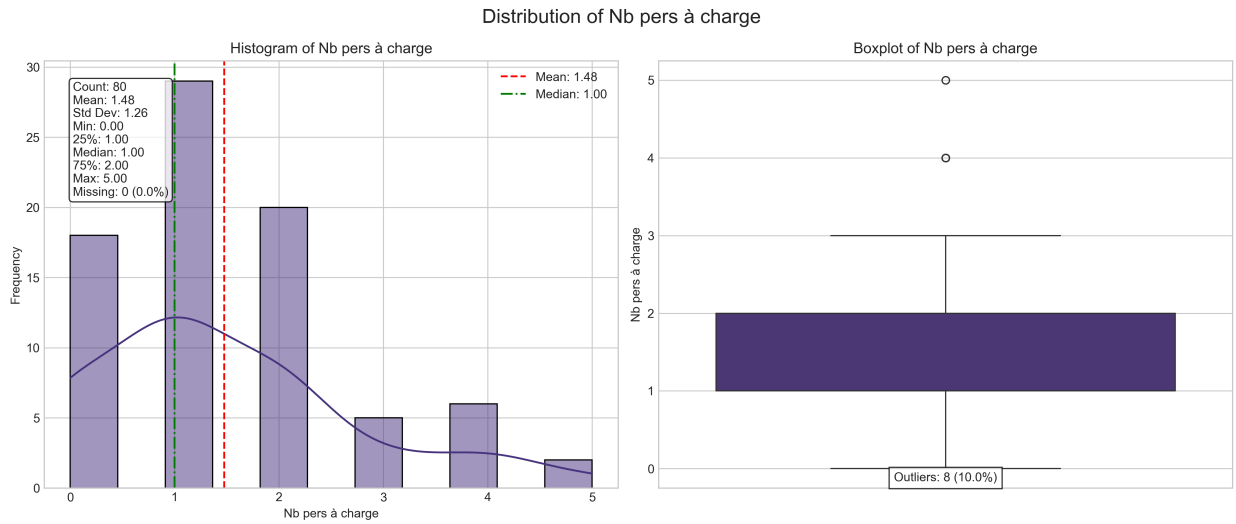
---

📘 SECTION 7: Number of Dependents (`Nb pers à charge`)

This section captures **how many people each woman supports financially or personally**, such as children, elderly, or other dependents.

🧠 What's shown

📷 Figure 1: Histogram + Boxplot



Histogram:

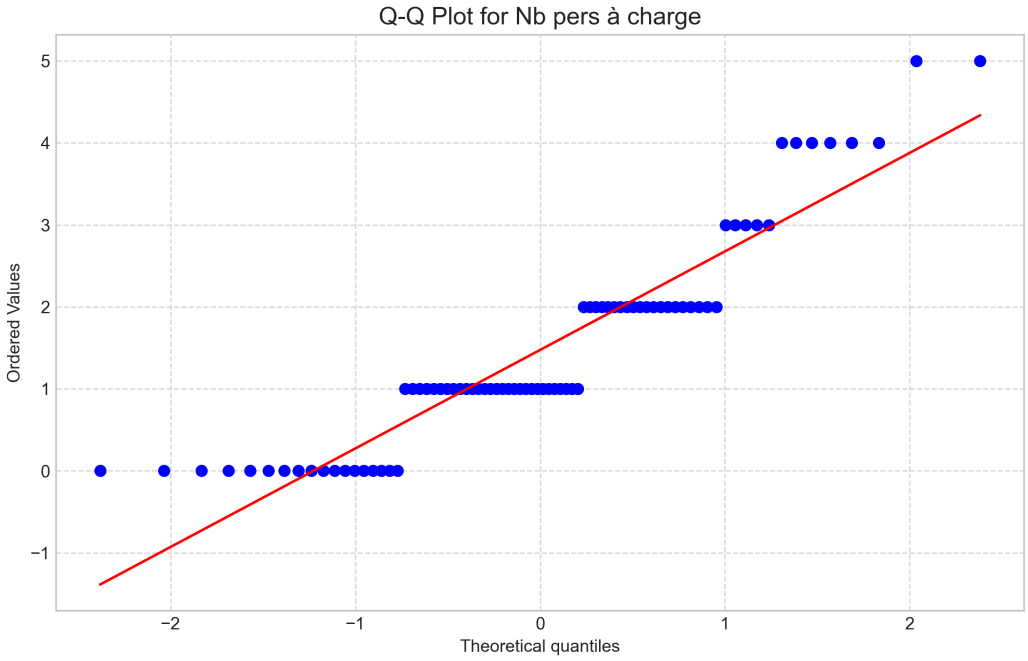
- Distribution of dependent counts (0 to 5)
- Mean: **1.48**, Median: **1**
- The most common value is **1**
- Min = 0, Max = 5

📷 Figure 2: Q-Q Plot for Normality

Boxplot:

- Clean distribution with a few mild outliers (not extreme)
- **10% of women are flagged as outliers** for having 4–5 dependents
- Still retained — not abnormal in extended family systems

🔍 Q-Q Plot:



- The discrete nature of the data is evident — **step-like jumps** in values
- Slight curvature shows **right skewness**, but less severe than with number of children
- Deviation from normality is expected and confirmed

💡 Insights:

- Most women support **1–2 people**
- A small group handles **4–5 dependents**, creating extra personal load
- This variable is discrete and skewed, so:
  - Avoid using it as-is in linear models
  - Treat it as **categorical or ordinal** (e.g., `0`, `1–2`, `3+`) for grouped analysis

---

🎯 Takeaway:

- This variable can be very informative in relation to:
  - **Workload balancing**

- **Time pressure**
- **Physical and emotional stress**

- Will be used later in modeling (possibly as a moderating or control variable)

📘 SECTION 5 : Weekly Work Hours

🧠 What this shows:

- A histogram with **hours worked per week**
- Blue solid line = **mean = 40.6h**
- Red dashed line = **standard reference (40h/week)**
- KDE curve illustrates density (smoothed frequency)
- Annotated box shows:
  - Range: **6h to 77h**

- Median = 42h
- **53.8% work > 40h/week**
- **27.5% work > 50h/week**

🔍 How to read this:

- The distribution is **wide and nearly flat** — no dominant peak
- Significant spread in working hours:
  - Some women work **very few hours**
  - Others exceed 70 hours — a clear **overwork segment**
- The histogram suggests **multimodal behavior** — different roles or schedules

---

💡 Key insights:

- **More than half** the women exceed the standard work week → strong indicator of labor intensity
- Almost **1 in 3 work 50+ hours/week** → important for health and time allocation analysis
- High standard deviation (16h) reflects **broad lifestyle/work differences**
- Median > mean confirms **slightly right-skewed** distribution (a few very high workers pulling the mean)

🎯 What we concluded:

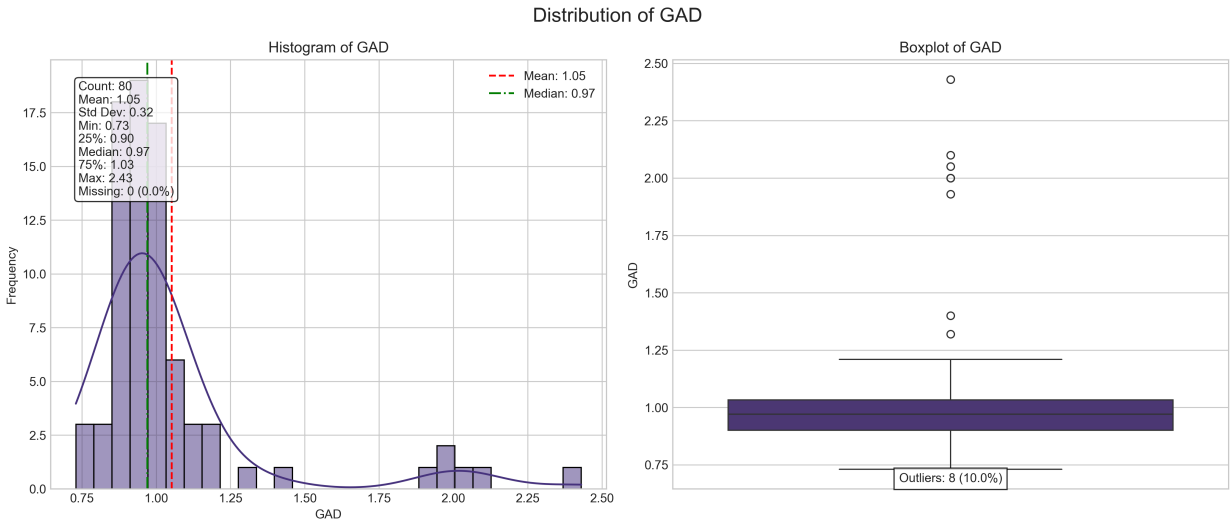
- This variable is **not suitable for normality-based tests**
- We should analyze it:
  - As a **continuous variable** in regressions
  - And/or segment it into ranges: `<30`, `30–40`, `40–50`, `>50`
- Critical for:
  - Linking to BMI, fatigue, and productivity

- Comparing with age and dependents

📘 SECTION 8: GAD (General Anxiety Disorder Score)

“A GAD score of 2.0+ is commonly used in clinical studies to flag potential anxiety disorders, though thresholds vary.”

📷 Plots:



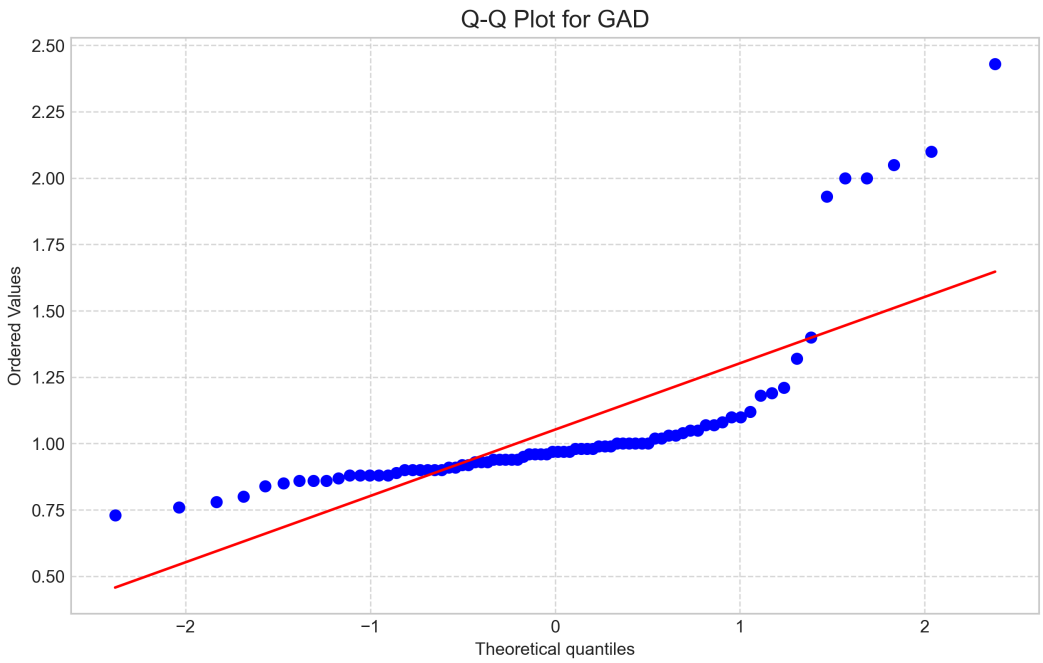
🧠 Key Stats:

- **Mean = 1.05, Median = 0.97**
- **Min = 0.73, Max = 2.43**
- 8 outliers identified (10%)

🔍 What it shows:

- The distribution is **positively skewed** (long right tail)
- Majority of scores are tightly packed between **0.9 and 1.1**
- **Outliers are high-scoring cases**, indicating possible severe anxiety

📉 Q-Q Plot:



- Points deviate strongly above the diagonal at the end → confirms **right skew**
- Not normally distributed — consistent with histogram and outliers

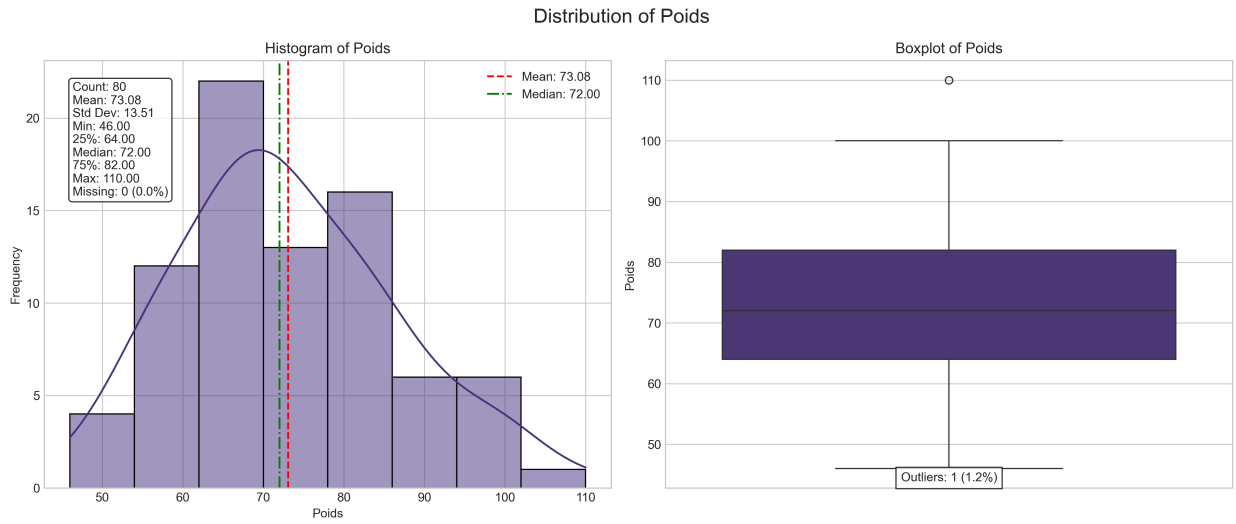
---

🎯 Conclusion:

- GAD is **non-normal**, skewed, with **clear outliers**
- For modeling: use non-parametric methods or transform
- High GAD cases may warrant **individual investigation** or flagging

📘 SECTION 9: Poids (Weight)

📷 Plots:



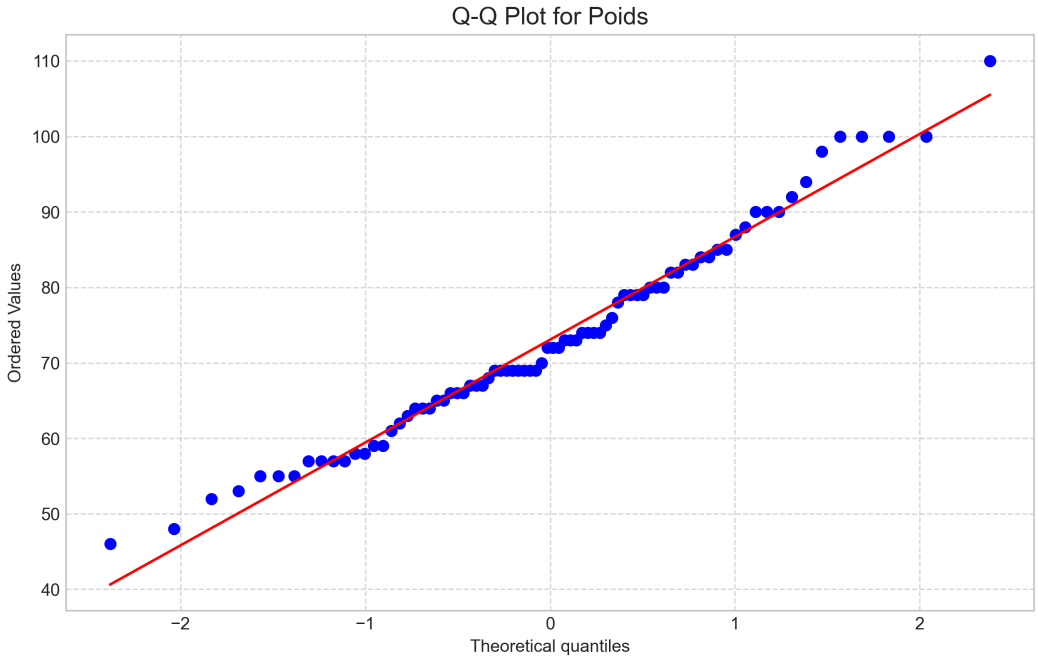
🧠 Key Stats:

- **Mean = 73.1 kg**, **Median = 72.0 kg**
- Range: **46–110 kg**
- 1 outlier (1.2%) > 100 kg

🔍 What we see:

- Distribution is **slightly right-skewed** — more values on the higher end
- Majority fall between **60–85 kg**
- Boxplot shows **tight IQR**, with just one mild outlier

📉 Q-Q Plot:



- Nearly all points hug the diagonal → **close to normal**
- A few slight deviations in the tails

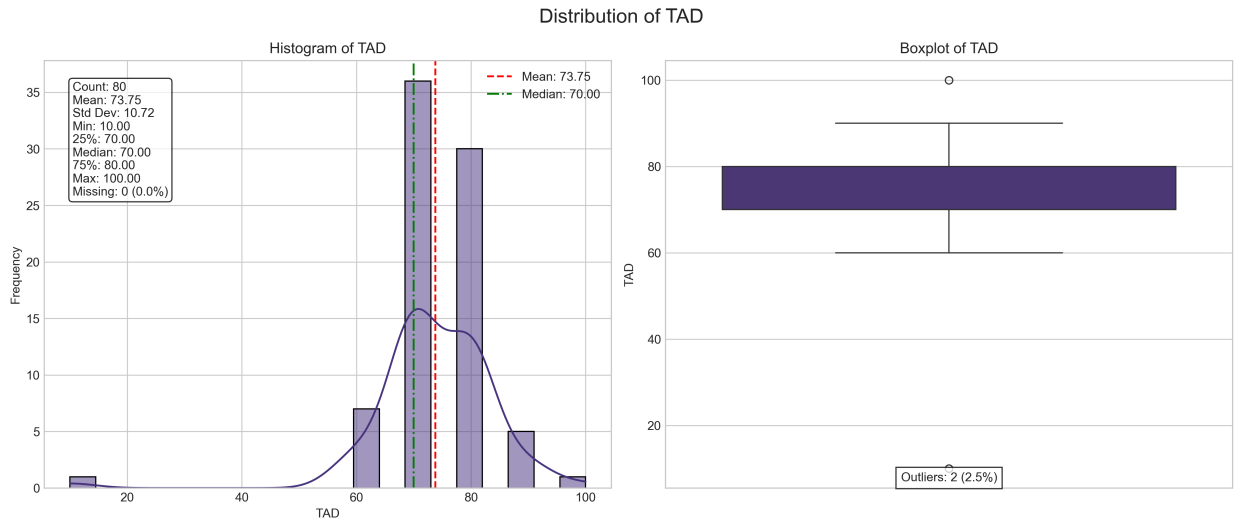
💡 Interpretation:

- This is a **clean, well-behaved variable**
- Slight skew but **acceptable for modeling**
- Used later for **BMI calculation** and correlations with work, age, and health variables

---

📘 SECTION 10: TAD (Diastolic Blood Pressure)

📷 Plots:



🧠 Key Stats:

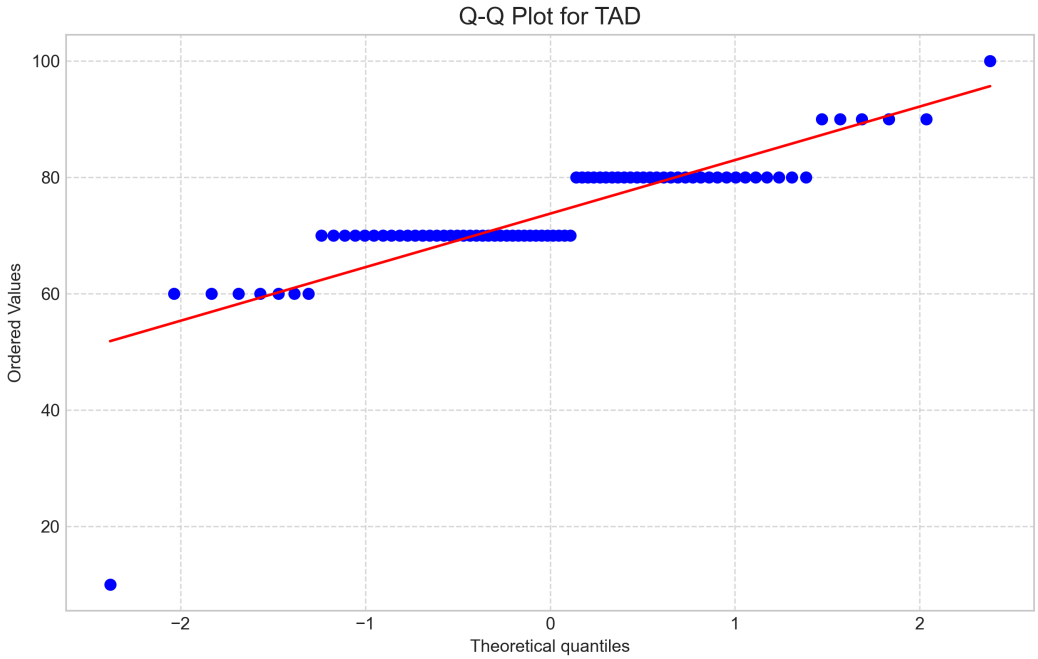
- **Mean = 73.75 mmHg**, **Median = 70 mmHg**
- Range: **10–100 mmHg**
- 2 outliers (2.5%), including one extreme low (10)

🔍 What we see:

Histogram + Boxplot:

- Mostly centered around **70–80 mmHg**, which is within **normal clinical range**
- A few **extreme outliers**, possibly due to **entry error** or rare hypotension
- Slight right skew, but mostly symmetrical

Q-Q Plot:



- Heavy **clustering around specific values** (e.g., 70, 80)
- Strong deviations in lower tail → confirms **non-normality** and **heaping**

💡 Interpretation:

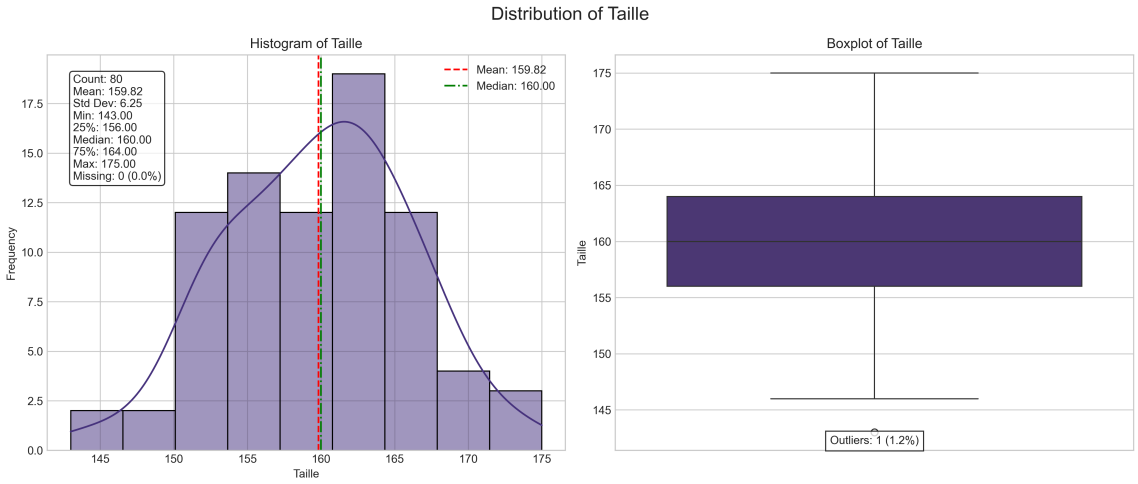
- Most women fall in a **healthy TAD range**
- However, data shows signs of **digit preference** (lots of 70s and 80s)
- Outliers must be examined — e.g., the 10 mmHg point may be a **data error**

🎯 Conclusion:

- TAD is **usable for analysis**, but not normal → prefer **non-parametric tests**
- Consider **winsorizing or flagging** outliers during modeling
- May correlate with age, BMI, and stress-related variables

📘 SECTION 11: Taille (Height)

📷 Plots:



🧠 Key Stats:

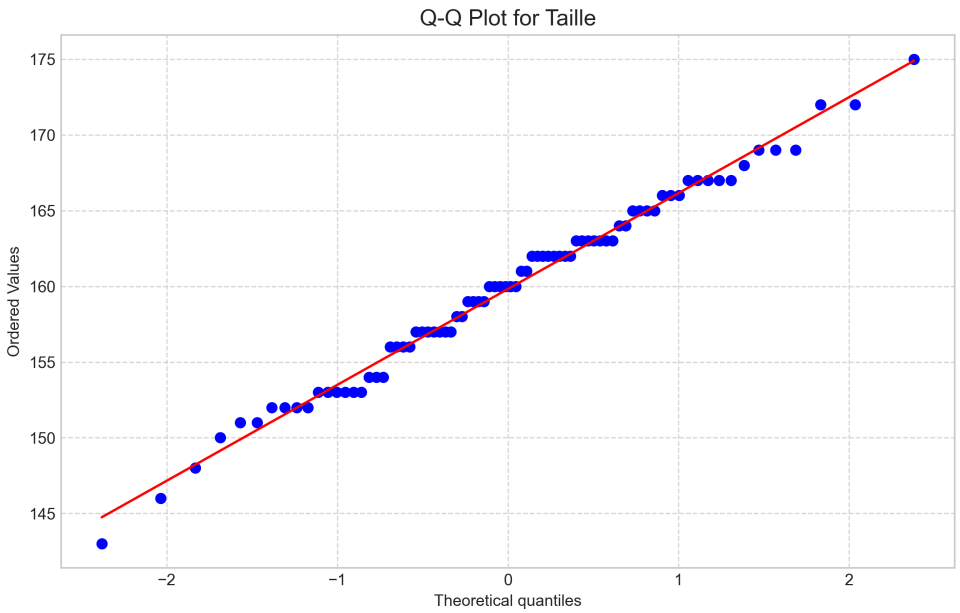
- **Mean = 159.8 cm**, **Median = 160.0 cm**
- Range: **143–175 cm**
- One mild outlier (1.2%)

🔍 Observations:

Histogram + Boxplot:

- Distribution is **very symmetrical** and tightly grouped
- Most heights fall in the **155–165 cm** range
- Only one slight low-end outlier (~143 cm), likely valid

Q-Q Plot:



- Points align closely with the diagonal line → **strong normality**
- Ideal shape for statistical modeling

💡 Interpretation:

- **Height is normally distributed** in the sample
- Clean, reliable variable for modeling (especially for BMI computation)
- No indication of data entry error

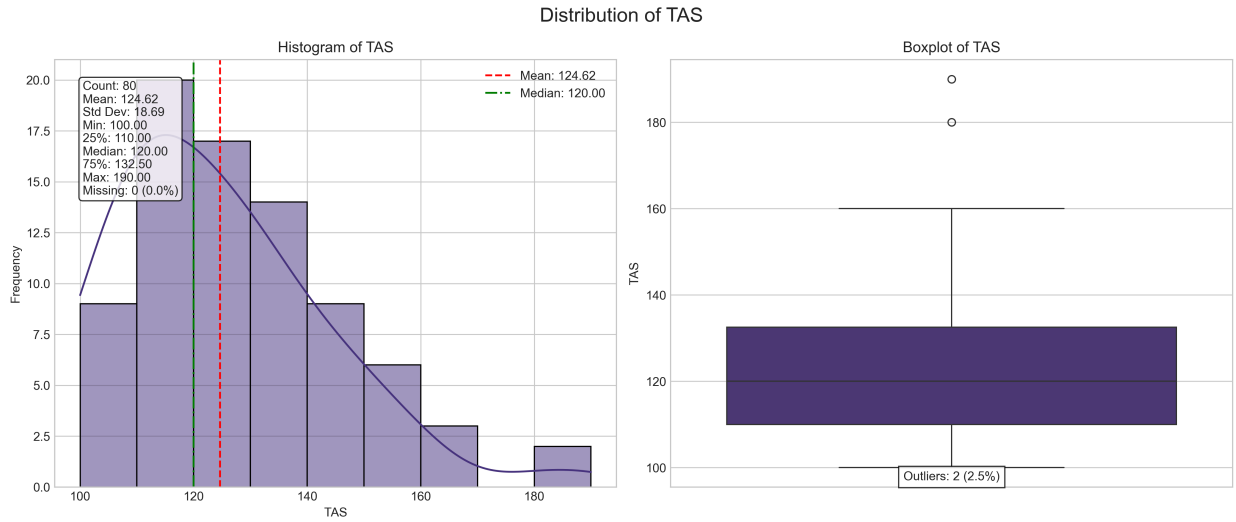
🎯 Conclusion:

- Use as-is or for BMI (`Poids / (Taille/100)^2`)
- May correlate with other biological metrics like weight, blood pressure, workload capacity

---

📘 SECTION 12: TAS (Tension Artérielle Systolique / Systolic Blood Pressure)

📷 Plots:



🧠 Key Stats:

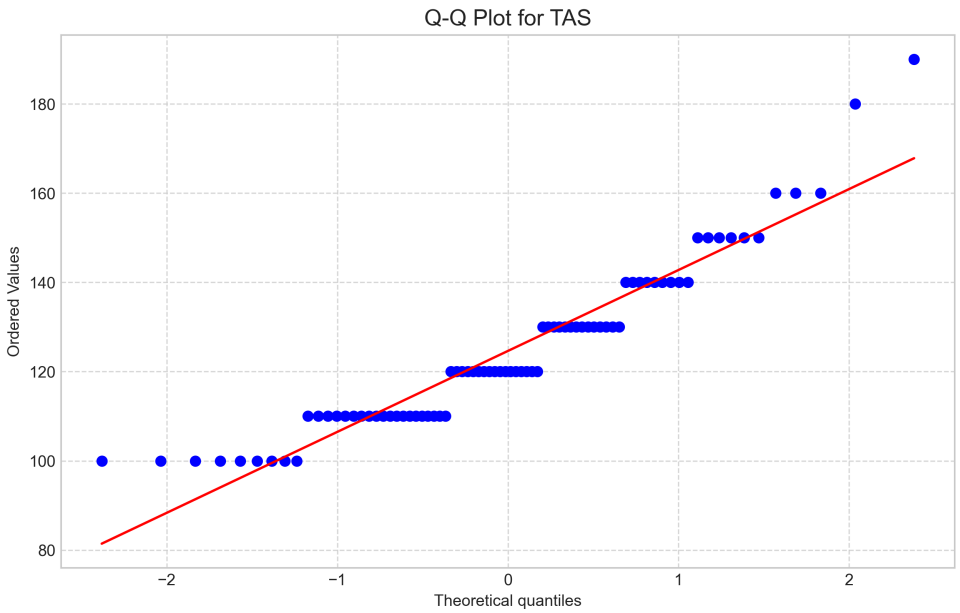
- **Mean = 124.6 mmHg**, **Median = 120 mmHg**
- Range: **100–190 mmHg**
- Outliers: 2 values above **180 mmHg** (2.5%)

🔍 Observations:

Histogram + Boxplot:

- Mild **right-skewed** distribution
- Most values fall in the **110–135 mmHg** range (normal to pre-hypertension)
- A few **extreme systolic values**, indicating possible **hypertension risks**
- Slight asymmetry — mean > median

Q-Q Plot:



- Deviations from the diagonal, especially at the **tails**
- Indicates **non-normality** — heavier upper tail

💡 Interpretation:

- A good portion of the women have **normal TAS**
- But the **upper outliers (e.g. 180–190 mmHg)** are critical to monitor — potentially **undiagnosed hypertensives**
- Combined with TAD and age, this can indicate cardiovascular risks

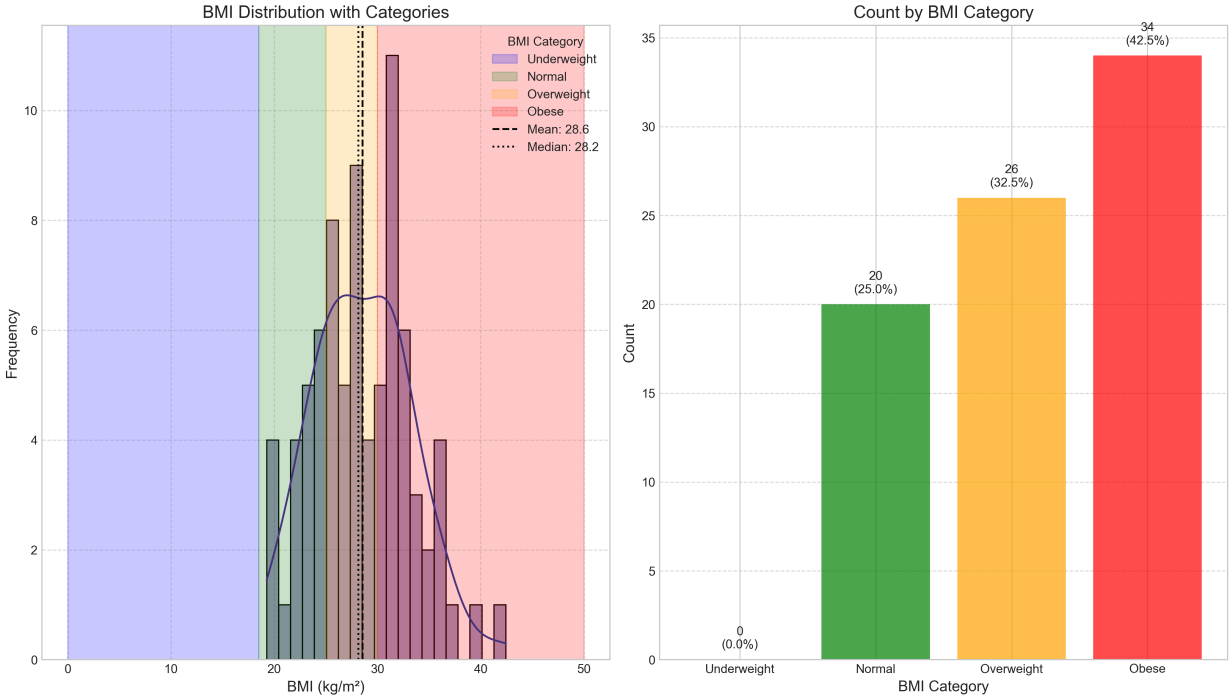
---

🎯 Conclusion:

- TAS is **slightly skewed**, and some values require attention
- Useful for deriving risk categories or investigating links with **BMI, age, GAD**
- May inform future health screening or targeted support

📘 SECTION: BMI ANALYSIS

BMI (Body Mass Index) is a numerical value that helps assess whether a person's **weight is appropriate for their height**.



📊 1. BMI Distribution with Categories (Left Plot)

What the plot shows:

- A histogram overlaid with a KDE curve represents the **distribution of BMI** values in the dataset.
- Background colors split the BMI range into standard **WHO categories**:
  - **Underweight**: BMI < 18.5
  - **Normal**: 18.5 ≤ BMI < 25

📈 2. Count by BMI Category (Right Plot)

What it shows:

- This bar chart quantifies how many individuals fall into each BMI category:
  - **Obese**: 34 individuals (42.5%)
  - **Overweight**: 26 individuals (32.5%)
  - **Normal**: 20 individuals (25.0%)

- **Overweight**: 25 ≤ BMI < 30
- **Obese**: BMI ≥ 30

Statistical indicators:

- **Mean BMI = 28.6**
- **Median BMI = 28.2**
- The mean and median being very close indicates a **roughly symmetric distribution** (though slightly skewed right).

Insight:

- The peak of the distribution is in the **Overweight to Obese** range.
- There are no individuals categorized as **Underweight**.
- The concentration around BMI 25–35 indicates a **health risk cluster** (above-normal weight range), which may have implications for designing health interventions in this population.

- **Underweight**: 0 individuals (0.0%)

Key Finding:

- A striking **75% of the population is either overweight or obese**.
- **Obesity alone affects nearly half (42.5%)** of the sample—this highlights a **public health concern**.

Overall Interpretation:

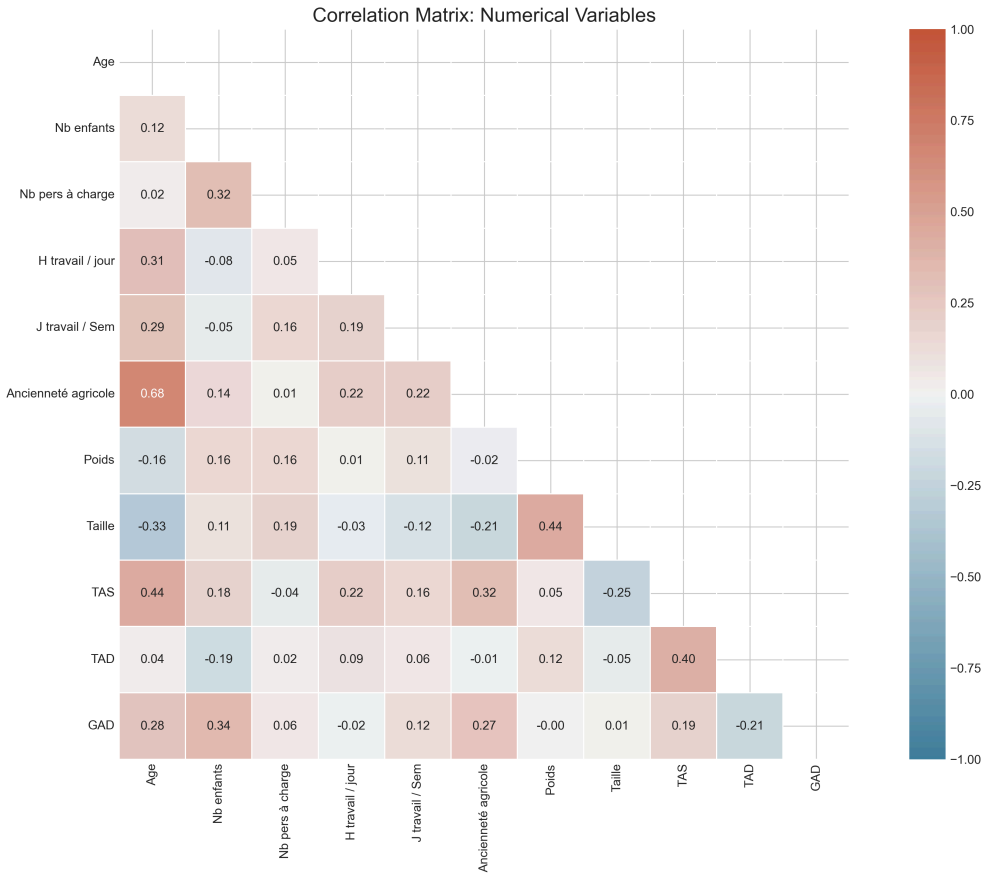
- The BMI profile of this dataset suggests a **high prevalence of excess body weight**.
- These findings could be cross-analyzed with variables such as age, work hours, and experience to determine possible contributing factors (e.g., sedentary patterns, age-related weight gain, etc.).

✅ **Conclusion:**

- BMI analysis reveals a **dominant overweight/obese pattern**, which aligns with the slightly right-skewed distribution.
- Public health implications are critical: this could indicate increased risk of metabolic diseases, especially in a farming population that is traditionally expected to be more physically active.

📊 Correlation Analysis

This heatmap visualizes
Pearson correlation coefficients between numerical variables. It helps identify how strongly two variables move together (positively or negatively).



A correlation matrix was generated to explore linear relationships among the numerical variables. The strongest positive correlation was found between **Age** and **Ancienneté agricole** (r = 0.68), confirming that older individuals tend to have more agricultural experience. Additionally, **Poids** and **Taille** (r = 0.44) showed a moderate positive relationship, as expected. Blood pressure indicators were moderately correlated: **TAS** and **TAD** (r = 0.40), aligning with physiological norms. Other noteworthy observations include a mild positive link between **GAD** (anxiety score) and **Number of children** (r = 0.34), suggesting potential stress factors.

Most work-related variables (e.g., hours worked per day) showed negligible correlations with other metrics, indicating limited linear influence

📌 Key takeaways from this matrix:

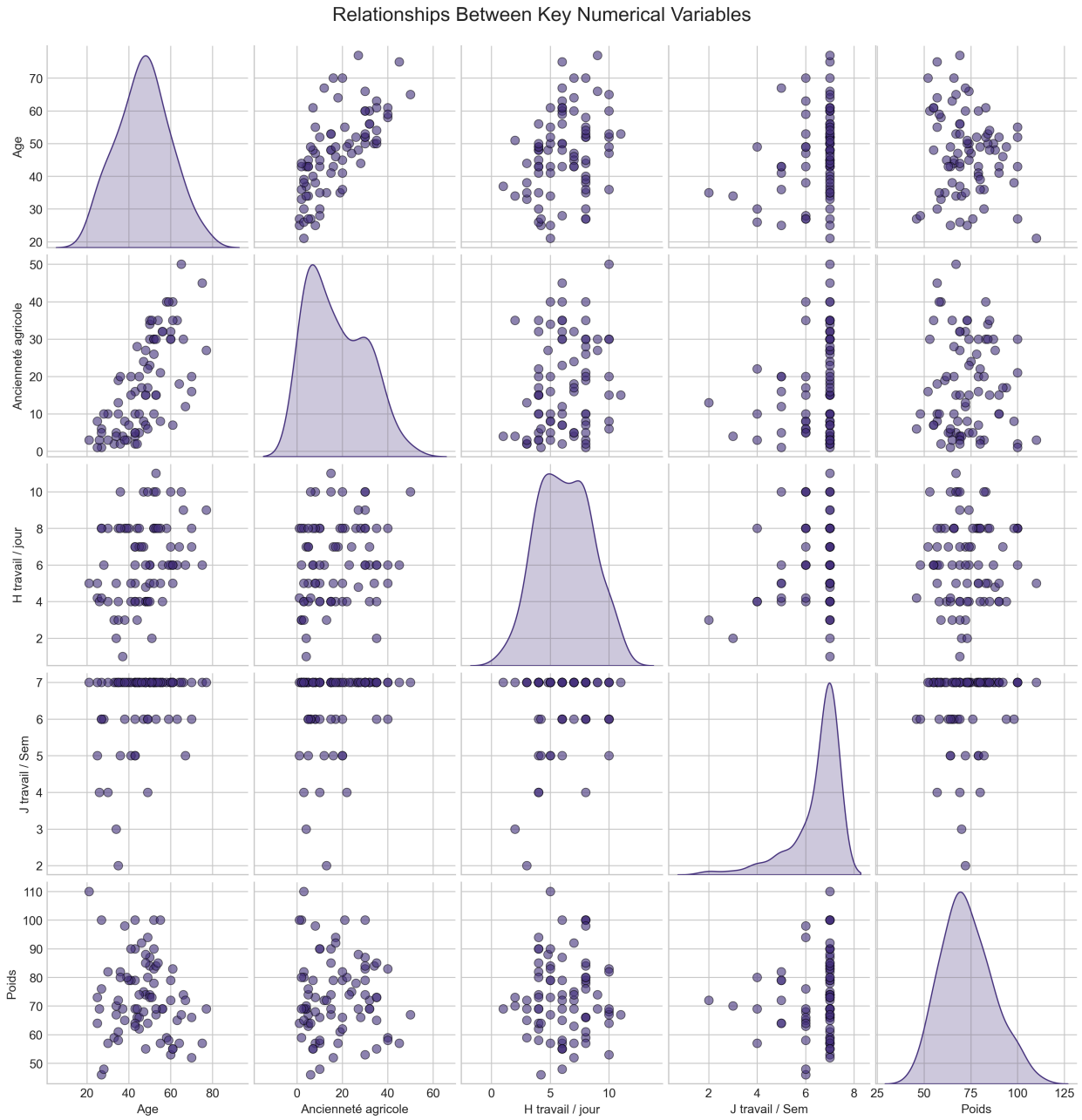
Variable Pair	Correlation	Interpretation
`Age` & `Ancienneté agricole`	**0.68**	Strong positive — older individuals tend to have more agricultural experience
`TAS` (systolic pressure) & `TAD` (diastolic)	**0.40**	Moderate positive — consistent with medical expectations
`Taille` & `Poids`	**0.44**	Positive — taller people tend to weigh more
`GAD` (General Anxiety Disorder score) & `Nb enfants`	**0.34**	Slight correlation — more children might be associated with more stress/anxiety
`Age` & `TAS`	**0.44**	Older individuals tend to have higher systolic pressure

📘 Section X: Summary of Numerical Analysis

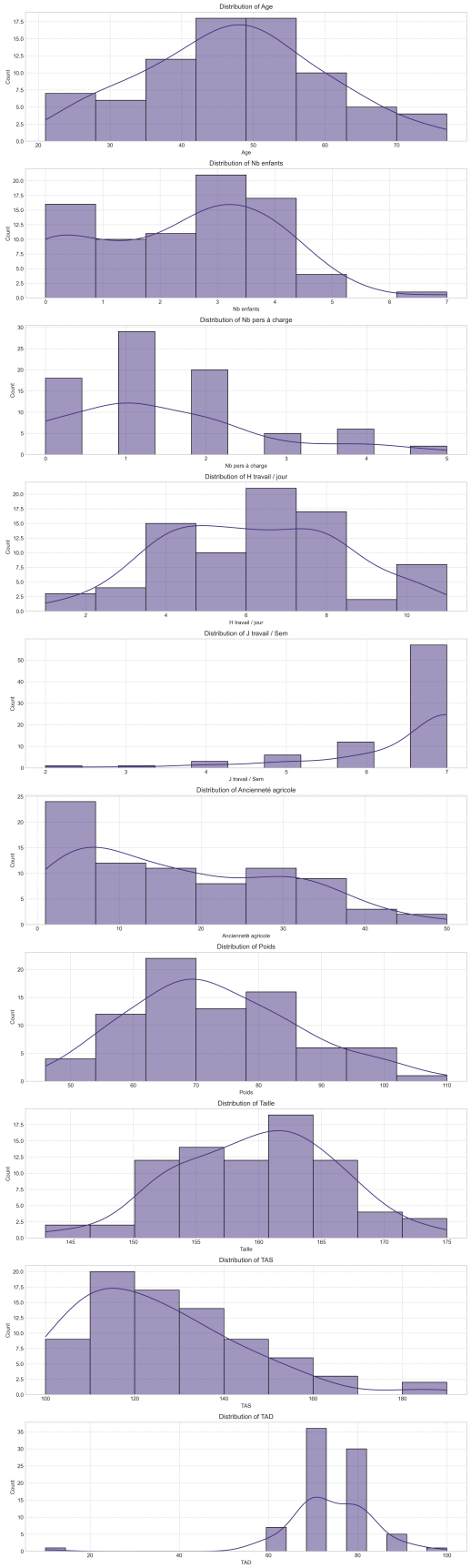
The exploratory analysis of numerical variables revealed diverse distribution patterns, relationships, and insights relevant to the population under study.

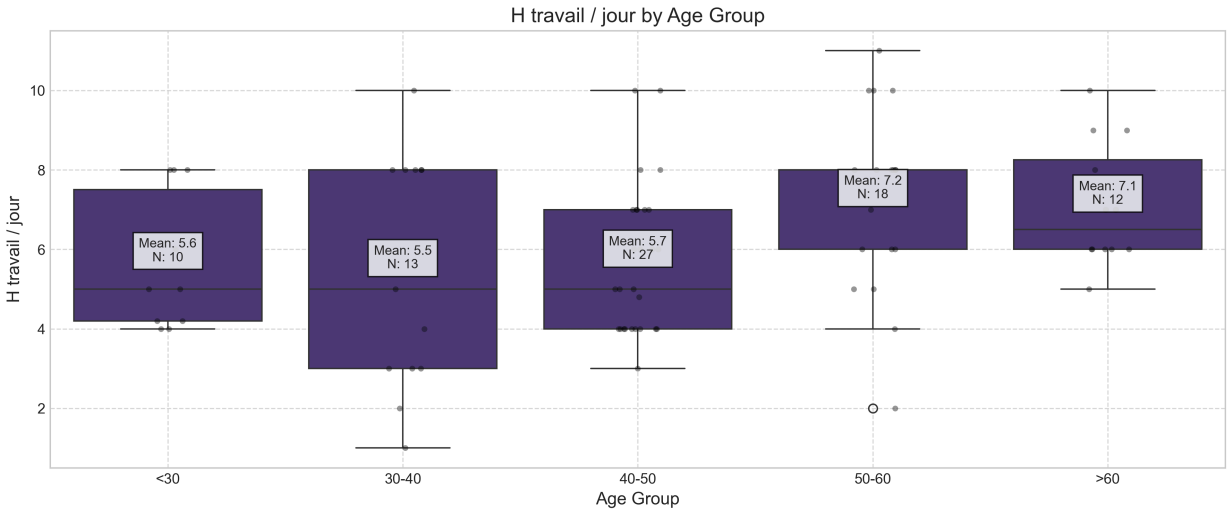
- **Distributions** were examined via histograms, boxplots, and Q-Q plots, highlighting non-normality in most variables (as confirmed by normality tests), along with key outliers (e.g., GAD, TAS, Poids).
- The **correlation matrix** emphasized significant relationships:
  - A **strong positive correlation** between *Age* and *Ancienneté agricole* (r = 0.68), as expected.
  - Moderate correlations between *TAS* and *TAD* (r = 0.40), and *TAS* and *Age* (r = 0.44).
- The **pairplot visualization** reinforced these associations and highlighted variable clusters with meaningful interactions (especially among age, experience, and blood pressure metrics).

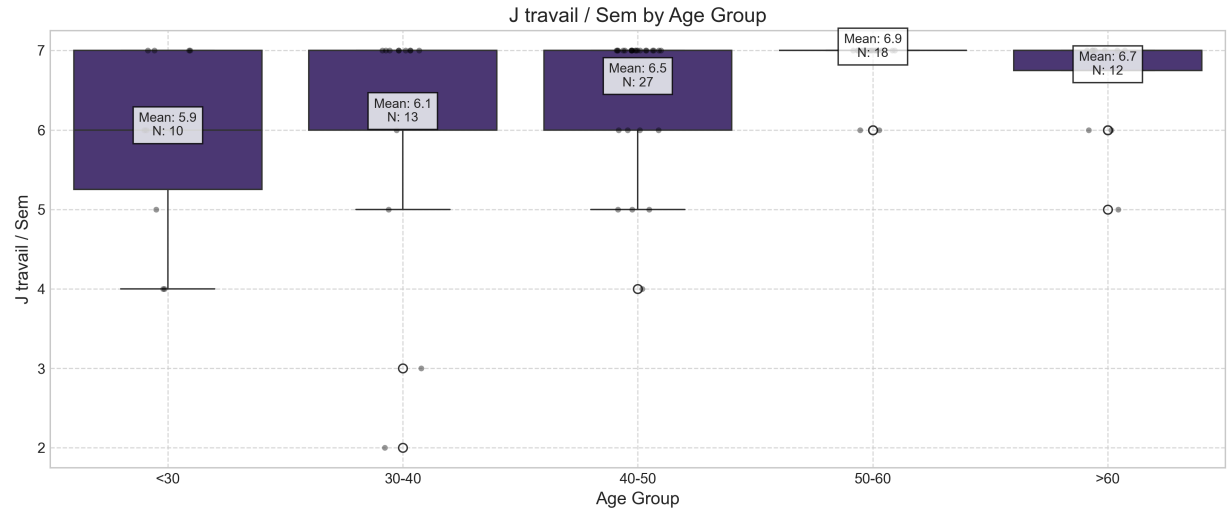
- The **BMI analysis** showed that **42.5% of the population falls into the obese category**, with a mean BMI of 28.6. This suggests a potential public health concern regarding weight-related risks.
- Work-related variables (hours/day, days/week) exhibited wide variability, reflecting diverse labor intensities within the sample.

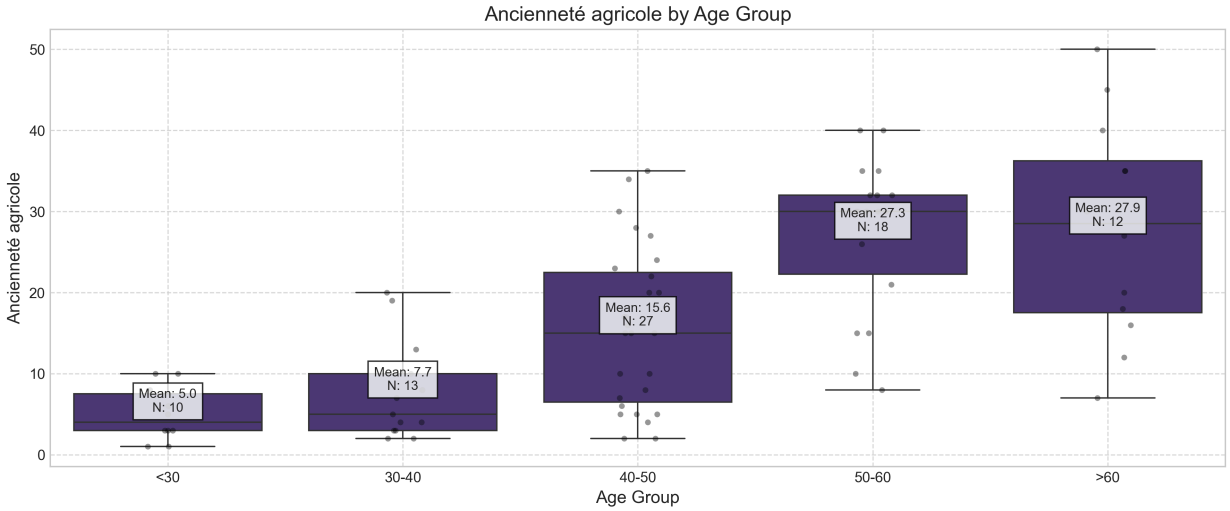


This combined evidence provides a strong foundation for deeper modeling, risk stratification, or hypothesis testing in subsequent stages of the project.


Distribution of Age
Count
Age
Distribution of Nb enfants
Count
Nb enfants
Distribution of Nb pers à charge
Count
Nb pers à charge
Distribution of H travail / jour
Count
H travail / jour
Distribution of J travail / Sem
Count
J travail / Sem
Distribution of Ancienneté agricole
Count
Ancienneté agricole
Distribution of Poids
Count
Poids
Distribution of Taille
Count
Taille
Distribution of TAS
Count
TAS
Distribution of TAD
Count
TAD







📘 SECTION X: SUMMARY AND IMPLICATIONS

This section synthesizes the key findings from the numerical analysis and discusses their implications for understanding and addressing the occupational health issues faced by female farmers.

🔍 Key Numerical Patterns

1. **Age and Experience Distribution**
    - Mean age of 47.11 years (range: 20-77 years)
    - Slightly right-skewed distribution with most farmers in 40-50 age group (33.8%)
    - Agricultural experience averaging 17.48 years (range: 1-50 years)
    - Strong correlation between age and experience (r = 0.68), but with significant individual variation
2. **Work Intensity Metrics**
    - Average weekly work of 40.6 hours (range: 6-77 hours)
    - 53.8% work more than the standard 40-hour week
    - 27.5% work more than 50 hours per week
    - Significant spread in working hours suggesting diverse work roles and demands
3. **Physical Health Indicators**
    - Mean weight of 73.1 kg (range: 46-110 kg)
    - Mean height of 159.8 cm (range: 143-175 cm)
    - Mean BMI of 28.6, indicating predominant overweight/obese status (75% above normal BMI)
    - 42.5% of the population classified as obese
4. **Cardiovascular Health Metrics**
    - Mean systolic blood pressure (TAS) of 124.6 mmHg (range: 100-190 mmHg)
    - Mean diastolic blood pressure (TAD) of 73.8 mmHg (range: 10-100 mmHg)

- Evidence of undiagnosed hypertension in a subset of the population
- General Anxiety Disorder (GAD) scores with a mean of 1.05 (range: 0.73-2.43)

5. **Family Structure Variables**
   - Average of 2.38 children per woman (range: 0-7)
   - Mean of 1.48 dependents (range: 0-5)
   - Positive skew in both variables, indicating some women with significantly higher family responsibilities

💡 Integrated Insights

1. **Age-Experience-Health Nexus**
   The analysis reveals a complex relationship between age, agricultural experience, and health indicators. Older, more experienced farmers show higher blood pressure values and other physiological stress markers, suggesting cumulative occupational health impacts. The correlation between age and systolic blood pressure (r = 0.44) highlights the interaction between natural aging processes and occupational stress factors.
2. **Work Intensity and Well-being**
   The wide distribution of weekly work hours (6-77 hours) reveals significant variability in work intensity. This pattern suggests different agricultural roles and responsibilities, with potential implications for physical strain and recovery time. The multimodal distribution of work hours indicates distinct subgroups with different work patterns that merit targeted investigation.
3. **BMI-Cardiovascular Risk Pattern**
   The high prevalence of overweight and obesity (75% of the population) represents a significant health concern. This pattern, combined with elevated blood pressure readings in some individuals, signals cardiovascular risk that may compound occupational hazards. The finding contradicts expectations of a physically active farming population, suggesting potential nutritional, cultural, or socioeconomic factors at play.
4. **Family Burden-Health Relationship**
   The analysis identified a potential relationship between family structure

(number of children, dependents) and anxiety indicators (GAD). The mild positive correlation (r = 0.34) between number of children and GAD scores suggests that family responsibilities may contribute to psychological strain, which could influence occupational health behaviors and outcomes.

5. **Workload-Age Distribution**
   The integration of age and work hours data reveals that older women tend to work slightly fewer hours than middle-aged women, but with high individual variation. This suggests a complex relationship between age, work capacity, and socioeconomic necessity that requires further investigation through multivariate analysis.

🔄 Relationship to Categorical Analysis

The numerical analysis complements the categorical findings in several ways:

- Provides quantifiable measures of health status that contextualize the categorical health complaints
- Quantifies work intensity patterns that help explain categorical occupational exposure differences
- Offers precise age and experience metrics that underpin demographic categories
- Allows for correlation analysis of continuous variables that can validate categorical associations
- Enables identification of outliers and subgroups that may require special attention in categorical analyses

🎯 Implications for Further Analysis

These findings provide direction for subsequent analytical steps:

1. **Principal Component Analysis (PCA)**
   - Will help identify key dimensions of numerical variability
   - Can reveal hidden patterns across multiple physiological metrics
   - Will enable reduction of dimensionality while preserving important variance

2. **Multiple Correspondence Analysis (MCA)**
    - Will integrate categorical variables with numerical indicators
    - Can reveal clusters of farmers with similar multidimensional profiles
    - Enables visualization of complex relationships between different variable types
3. **Combined Approach**
    - Integration of numerical findings with categorical analyses
    - Development of composite risk indices incorporating both data types
    - Creation of multidimensional farmer profiles for targeted intervention

📘 CONCLUSION

The numerical analysis of the female farmers dataset has revealed detailed patterns across demographic, workload, physical, and cardiovascular health domains. These patterns highlight important relationships between age, experience, work intensity, and health indicators that provide a quantitative foundation for understanding occupational health in this population.

Several critical insights emerge that can guide interventions:

1. **Cardiovascular health monitoring** should be prioritized, given the high prevalence of elevated BMI and blood pressure readings in a subset of the population.
2. **Age-appropriate approaches** are needed that address the different physiological profiles and cumulative exposures across age groups.
3. **Work intensity considerations** should be incorporated into occupational health guidelines, particularly for the significant portion working more than 50 hours weekly.
4. **Family burden factors** should be considered when designing support systems and interventions, as family responsibilities appear to correlate with stress indicators.
5. **Individualized assessment** is necessary given the high variability observed in most numerical variables, cautioning against one-size-fits-all approaches.

The numerical analysis has established a strong quantitative foundation for the multivariate analyses (PCA and MCA) that will follow, enabling a more comprehensive understanding of the complex factors influencing the health and safety of female farmers in this population. By integrating these numerical insights with categorical findings, a more holistic picture emerges that can inform targeted, effective interventions to improve occupational health outcomes.