ALGORITHMIC TRANSPARENCY & TECHNICAL SPECIFICATION

PurifyData Pro architectural, statistical, privacy, and operational documentation

Document version	2.0.0
Application version reflected	PurifyData Pro Version 3.0
Classification	Technical specification / algorithmic transparency / user-facing method disclosure
Target audience	Researchers, postgraduate students, data analysts, statistical auditors, supervisors, and technical reviewers

1. Current Application Scope

PurifyData Pro is a browser-based research tool for preliminary data screening, outlier assessment, missing-data review, optional single imputation, exportable screening outputs, and initial assessment of univariate and multivariate normality.

The current application is a static web app protected by Supabase authentication. Users must log in before using the statistical workspace. Public access remains available for the landing page, sign-up, log-in, Privacy Notice, Terms of Use, and Contact page.

- Registration collects full name, email address, password, organisation, occupation, country, research interest, required Terms of Use and Privacy Notice agreement, and optional marketing consent.
- Profile information is saved in the Supabase *public.profiles* table using the authenticated user ID.
- Marketing consent, consent date-time, and withdrawal date-time are stored so that users can manage preferences through the profile or marketing-preferences page.
- Account deletion is designed to call a secure Supabase deletion function so the authenticated account can be removed automatically when the database RPC is installed.
- Raw uploaded CSV data are processed in the browser workspace and are not intentionally uploaded to Supabase.

Privacy distinction
PurifyData Pro uses Supabase for authentication and profile/consent management. Statistical datasets uploaded by users are held in the browser runtime state for analysis and are not stored as profile records. Closing or refreshing the browser tab clears the active in-memory dataset state.

2. Data Intake, Variable Selection, And Missing-Value Handling

CSV import is followed by automatic header detection, row parsing, numeric-variable detection, and a variable-selection pop-up. The user chooses the variables required for screening before results are recalculated.

- Blank cells, whitespace-only entries, null, undefined, and NA are treated as missing values.
- Other missing-value codes such as N/A, -99, . , or MISSING should be recoded before import.
- Missing-data deletion is controlled by an on/off checkbox. When active, cases with missing selected values are deleted listwise before screened exports are produced.
- When missing data deletion is inactive, missing selected values are reported and retained unless another active screening rule deletes the case.
- The app displays progress messages during CSV import, normality computation, imputation, and export operations.

3. Core Descriptive Statistics

For selected numeric variables, PurifyData Pro calculates sample size, missing-data frequency, mean, sample standard deviation, minimum, maximum, quartiles, median, skewness, excess kurtosis, standardized scores, IQR fences, Pearson correlations, and case-level diagnostic values.

- Mean:
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- Sample standard deviation:
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- Skewness: third standardized moment used as a distributional-shape diagnostic.
- Excess kurtosis: fourth standardized moment minus 3, used to assess tail weight relative to a normal distribution.
- Quantiles: calculated from sorted finite values using linear interpolation.

4. Univariate Outlier Detection and Deletion

Univariate outlier screening is one of the app's primary functions. The workflow has no variable-count cap and uses all selected numeric variables. Users may activate Z-score screening, IQR screening, both methods, or neither method.

Setting	Current options / default	Interpretation
Z-score cutoff	2.58, 3.00 default, 3.29, or 4.00	Flags value whose absolute standardized score exceeds the selected cutoff when the Z-score assessment is active.
IQR multiplier	1.50 default, 2.20, or 3.00 x IQR	Flags values outside Q_1 - multiplier x IQR or Q_3 + multiplier

		x IQR when the IQR assessment is active.
Missing deletion	On by default	Deletes rows with missing values on selected screening variables when active.
Delete flagged rows from screened exports	On by default	Controls whether flagged rows are removed from the exported screened datasets.

The univariate algorithm is iterative. After each pass, newly flagged cases are removed from the working screening set, descriptive statistics and cutoffs are recalculated, and the remaining cases are screened again. The loop stops only when no additional univariate outliers are detected.

Exported univariate screened CSV files include the screened data and diagnostic columns such as Z-score values for assessed variables where applicable.

5. Multivariate Outlier Detection and Deletion

Multivariate outlier screening uses squared Mahalanobis distance. It is also a primary screening function and does not impose a variable-count cap. It requires complete numerical observations for the selected variables in order to compute Mahalanobis distances.

- The user can turn Mahalanobis assessment on or off.
- The default Mahalanobis threshold is the 99.9th chi-square percentile, corresponding to $p < .001$.
- Available percentile options are 95th, 97.5th, 99th, and 99.9th.
- The chi-square degrees of freedom equal the number of selected variables used in the Mahalanobis calculation.
- The app also reports the empirical percentile of observed Mahalanobis distances in the final pass.
- If the ordinary covariance matrix is singular or unstable, the app attempts a ridge-regularized covariance inverse and reports this caution.
- Exported multivariate screened CSV files include Mahalanobis distance, Mahalanobis percentile rank, and distributional diagnostics including skewness and kurtosis columns where relevant.

The multivariate procedure is iterative. In each pass, cases exceeding the selected Mahalanobis chi-square cutoff are removed, the covariance structure is recalculated, and the remaining complete cases are retested until no further multivariate outliers are identified.

6. Normality Assessment

Normality assessment is an additional diagnostic function rather than the main deletion engine. Variable-count and display caps apply to normality outputs for browser responsiveness; these caps do not limit univariate outlier deletion or Mahalanobis outlier deletion.

6.1 Univariate Normality

- Moment-based assessment uses user-entered skewness and excess-kurtosis cutoffs; the default cutoffs are absolute skewness ≤ 2 and absolute excess kurtosis ≤ 7.
- Shapiro-Wilk diagnostics use a Royston-style p-value approximation. In Fast, Balanced, and Detailed modes, the app may test an evenly spaced subset for browser responsiveness. When Full available cases is selected, PurifyData Pro applies SPSS-style Shapiro-Wilk to all finite tested values.
- Kolmogorov-Smirnov normal-fit diagnostics estimate the mean and standard deviation from the tested values. In Fast, Balanced, and Detailed modes, the app may test an evenly spaced subset and does not apply Lilliefors correction. When Full available cases is selected, PurifyData Pro applies SPSS-style Kolmogorov-Smirnov with Lilliefors correction to all finite tested values.
- Histograms with fitted normal curves, univariate Q-Q plots, and boxplots are generated for visual review.
- A variable is classified as approximately normal when at least one active univariate numerical method supports normality. This is a permissive preliminary-screening rule, not a universal definition of normality.
- Corrective-method suggestions are shown only when violations are detected.

Formal-test mode	Shapiro-Wilk values	Kolmogorov-Smirnov values	Use case
Fast default	Up to 250	Up to 500	Routine screening and large datasets.
Balanced	Up to 500	Up to 1,000	More detail when the browser remains responsive.
Detailed	Up to 1,000	Up to 2,000	Smaller or moderate datasets require more formal test detail.
Full available cases	All finite values	All finite values	Best for comparison with SPSS/R but may be slower.

6.2 Multivariate Normality

- Mardia's multivariate skewness assesses asymmetric joint departure from multivariate normality.
- Mardia's multivariate kurtosis assesses joint tail-weight departure from normal-theory expectations.
- The Henze-Zirkler-type diagnostic uses Monte Carlo simulation with limited runs to remain responsive in the browser.
- The multivariate chi-square Q-Q plot compares squared Mahalanobis distances with expected chi-square quantiles.
- The visual Q-Q criterion uses a correlation cutoff of $r \geq 0.97$ as supportive visual evidence.
- The app combines the multivariate results into graded interpretations such as approximately reasonable, mixed evidence, asymmetry, tail-weight departure, mild statistical departure, or clear evidence of non-normality.
- Corrective-method suggestions are displayed only when violations or cautionary departures are detected.

7. Graphical Diagnostic Outputs

- Histograms include a fitted normal curve for univariate normality review.
- Univariate Q-Q plots compare observed quantiles with expected normal quantiles.
- Boxplots support outlier review before and after univariate deletion.
- The multivariate chi-square Q-Q plot evaluates whether squared Mahalanobis distances resemble the chi-square pattern expected under multivariate normality.
- Plot output is limited to the first 24 selected variables for responsiveness.

8. Imputation

PurifyData Pro includes optional single-imputation methods for exploratory and preliminary use. These methods do not represent uncertainty caused by missing data and should not be described as multiple imputations.

Method	Mechanics	Main caution
Mean	Fills missing values with the arithmetic mean of observed values in the same variable.	Can reduce variance and is sensitive to skewness/outliers.
Median	Fills missing values with the 50th percentile of observed values.	More robust for skewed variables, but still a single-imputation method.
Random hot-deck	Use a user-specified seed and randomly select observed donor values from the same variable.	Results depend on the seed and donor pool.
Regression-style	Uses available correlated variables and absolute Pearson correlations to form a weighted prediction.	Correlation-weighted single imputation; not a fitted multiple-regression or multiple-imputation model.

9. Exports, Notes, and Reporting Support

- Univariate screened CSV export.
- Multivariate screened CSV export.
- Normality assessment CSV export.
- Imputed dataset CSV export.
- Detailed univariate and multivariate screening notes with copy buttons.
- Consolidated decisions tab summarising activated assessments and suggested analysis paths.
- Overview cards showing rows, variables, numeric variables, Uni. retained, Multi. retained, assessment coverage, cutoff values, and quality flags.

Screening notes report the number of cases before and after screening, missing-data handling, active outlier criteria, thresholds used, iterative deletion logic, and deleted or retained cases. Notes omit inactive assessment methods so that the reporting text matches the user's selected settings.

10. Computational Boundaries and Safeguards

Limit	Current value	Purpose
Normality plot display	24 variables	Prevents excessive SVG/histogram/Q-Q/boxplot rendering.
Univariate normality computation	50 selected variables	Limits additional normality computation only; does not cap

		univariate outlier detection/deletion.
Multivariate normality variables	80 variables	Limits additional multivariate normality diagnostics only; does not cap Mahalanobis outlier detection/deletion.
Multivariate normality rows	Responsive subset up to 120 complete cases	Avoids freezing during Mardia, Henze-Zirkler, and chi-square Q-Q diagnostics.
Formal test case modes	Fast, Balanced, Detailed, Full	Allows the user to balance speed against SPSS/R comparability.
Henze-Zirkler simulation	Up to 19 runs, sometimes fewer for larger problems	Keeps browser-based diagnostics responsive.
Correlation matrix display	45 variables	Hides the displayed matrix above this limit while preserving underlying screening calculations.
Case table display	1,000 cases	Limits on-screen case table length; CSV exports include eligible rows.
Data preview	80 columns	Prevents preview rendering from overwhelming the browser.

11. Authentication, Public Pages, and Account Management

- Supabase authentication supports sign-up, login, logout, email verification, resend verification, forgot password, reset password, user profile, marketing preferences, and account deletion.
- Only the Supabase Project URL and publishable/anonymous key are present in the front end. The service-role key and database password are not placed in the static app.
- Privacy Notice, Terms of Use, and Contact are presented as public floating pages with close controls.
- Users must agree to the Terms of Use and Privacy Notice before registration.
- Marketing communication consent is optional and can be withdrawn.

12. Intended Use and Statistical Cautions

- PurifyData Pro is intended for preliminary screening, education, and transparent reporting support.
- A flagged observation is not automatically erroneous and should not be deleted without substantive or methodological justification.
- Normality decisions should be interpreted alongside plots, sample size, variable measurement level, outliers, model assumptions, and discipline-specific expectations.
- For theses, dissertations, journal articles, CFA, SEM, clinical research, or other high-stakes work, important findings should be verified using validated statistical software and the full dataset.
- The app's Kolmogorov-Smirnov normal-fit diagnostic may still differ slightly from SPSS because statistical packages can use different numerical approximations, but Full available cases mode applies the Lilliefors correction for closer SPSS-style comparison.
- When comparing Shapiro-Wilk or Kolmogorov-Smirnov p-values with SPSS/R, use the Full available cases option and test the same screened CSV.

13. Reproducibility Checklist

- Application version: PurifyData Pro Version 3.0.
- Selected variables.
- Missing-data deletion setting.
- Z-score cutoff and whether Z-score assessment was active.
- IQR multiplier and whether IQR assessment was active.
- Skewness and kurtosis cutoffs and whether moment assessment was active.
- Univariate formal-test case limit.
- Mahalanobis percentile and whether Mahalanobis assessment was active.
- Whether flagged rows were deleted from screened exports.
- Imputation method and random seed, if imputation was used.
- Any responsive-subset caution reported for multivariate normality.
- Exported screened/imputed CSV files and copied screening notes.

14. Citation

Recommended citation:

Citation
Ady Hameme, B. N. A. (2026). PurifyData Pro: Data screening and initial normality assessment (Version 3.0) [Web application]. ADV Research and Consultancy. https://purifydatapro.myadvrc.workers.dev/

Example methods statement: Initial data screening was conducted using PurifyData Pro, version 3.0 (Ady Hameme, 2026). The assessment included missing-value review, iterative univariate outlier screening, squared Mahalanobis-distance analysis, descriptive distributional statistics, graphical diagnostics, optional imputation, and preliminary univariate and multivariate normality assessment.

15. Technical Appendix: Mathematical Formulas Used in the App

This appendix records the principal formulas and computational rules used by PurifyData Pro. The formulas support algorithmic transparency and should be read together with the cautions in the main document. Some p-values and quantiles use numerical approximations so that the application can run inside a browser.

15.1 Descriptive Statistics and Distribution Shape

Procedure	Formula / Calculation	How the app uses it
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	Summarises the centre of each selected numeric variable.
Sample standard deviation	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$	Measures sample dispersion using Bessel's correction.
Z-score	$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$	Supports univariate standardized-score outlier screening.
Skewness	$g_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$	Assesses distributional asymmetry against the user-selected skewness cutoff.
Excess kurtosis	$g_2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$	Assesses tail weight against the user-selected excess-kurtosis cutoff.
Linear-interpolated quantile	$Q_p = x_k + r(x_{k+1} - x_k)$	Calculates quartiles, medians, and other percentiles from sorted finite values.

15.2 Univariate Outlier Screening

Procedure	Formula / Calculation	How the app uses it
IQR	$IQR = Q_3 - Q_1$	Forms the basis of boxplot-style outlier fences.
Lower IQR fence	$Lower\ fence = Q_1 - m(IQR)$	Flags low-tail univariate outliers when IQR assessment is active.
Upper IQR fence	$Upper\ fence = Q_3 + m(IQR)$	Flags high-tail univariate outliers when IQR assessment is active.
Iterative deletion	$Screen \rightarrow flag \rightarrow delete\ flagged\ cases \rightarrow recalculate \rightarrow repeat\ until\ no\ new\ flags$	Reduces masking effects by recalculating limits after each deletion pass.

15.3 Correlation, Covariance, and Mahalanobis Distance

Procedure	Formula / Calculation	How the app uses it
Pearson correlation	$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$	Used in the correlation matrix and correlation-weighted imputation.
Covariance	$cov(X_j, X_k) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k})$	Builds the covariance matrix for multivariate calculations.
Ridge-regularised covariance	$\Sigma_{reg} = \Sigma + \lambda I$	Used when the ordinary covariance matrix is singular or unstable.
Squared Mahalanobis distance	$D_i^2 = (x_i - \mu)' \Sigma^{-1} (x_i - \mu)$	Measures each complete case's multivariate distance from the selected-variable centre.
Mahalanobis cutoff	$Cutoff = \chi^2_{percentile}(df = p)$	Flags multivariate outliers using the selected chi-square percentile.
Empirical percentile rank	$Percentile_i = \frac{count(D^2 \leq D_i^2)}{N}$	Reports how extreme an observed Mahalanobis distance is within the analysed dataset.

15.4 Normality Assessment

Procedure	Formula / Calculation	How the app uses it
Shapiro-Wilk W	$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	Used as a univariate formal normality diagnostic with a Royston-style p-value approximation.
Kolmogorov-Smirnov D	$D = max_x \lvert F_n(x) - \Phi\left(\frac{x - \bar{x}}{s}\right) \rvert$	Measures the largest distance between the empirical distribution and fitted normal distribution.
Normal Q-Q plot correlation	$r_{QQ} = corr\begin{pmatrix} observed\ ordered\ values, \\ expected\ normal\ quantiles \end{pmatrix}$	Supports visual normality when the Q-Q pattern is sufficiently linear.
Mardia multivariate skewness	$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(x_i - \mu)' \Sigma^{-1}(x_j - \mu)\right]^3$	Assesses asymmetric joint departure from multivariate normality.
Mardia multivariate kurtosis	$b_{2,p} = \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu)' \Sigma^{-1}(x_i - \mu)]^2$	Assesses multivariate tail-weight departure from normal theory.
Mardia kurtosis z	$z = \frac{b_{2,p} - p(p+2)}{\sqrt{\frac{8p(p+2)}{n}}}$	Provides a standardized kurtosis departure statistic.
Henze-Zirkler diagnostic	$HZ = distance\begin{pmatrix} empirical\ characteristic\ function, \\ theoretical\ MVN\ characteristic\ function \end{pmatrix}$	Implemented as a simulation-based diagnostic with limited runs for browser responsiveness.
Chi-square Q-Q plot	$sorted(D_i^2)\ compared\ with\ \chi^2\ quantiles,$	Shows whether squared

	$df = p$	Mahalanobis distances follow the expected multivariate-normal distance pattern.

15.5 Imputation

Procedure	Formula / Calculation	How the app uses it
Mean imputation	$x_{missing} = \bar{x}_{observed}$	Replaces missing values with the observed variable mean.
Median imputation	$x_{missing} = median(x_{observed})$	Replaces missing values with the observed variable median.
Random hot-deck	$x_{missing} = random\ observed\ donor\ value\ selected\ using\ seed$	Uses a reproducible pseudo-random donor process.
Correlation-weighted regression-style imputation	$\hat{x} = \frac{\sum_{j=1}^{m} \lvert r_j \rvert \hat{y}_j}{\sum_{j=1}^{m} \lvert r_j \rvert}$	Combines predictions from available correlated variables.
Bivariate prediction component	$\hat{y}_j = \mu_{target} + r_j \frac{x_j - \mu_j}{s_j} s_{target}$	Predicts a target value from one available driver variable before weighted combination.

Transparency notes: the formulas above describe the main computational logic. Browser-based implementation details such as floating-point precision, case limits, responsive subsets, and numerical approximations may create small differences from SPSS, R, Python, or other statistical packages.