The Data Preparation (and Interpretation) Challenge

Research Question: "What can we learn about our data to make better analytical decisions, and how do we do it?"

---

1. Title and Research Question

1.1 Title

"The Data Preparation (and Interpretation) Problem"

1.2 Research Question

"What can we learn about our data to make better analytical decisions, and how do we do it?"

1.3 Relevance and Significance

Imagine this: Bill Gates walks into a bar. Suddenly, on average, everyone in the bar is a millionaire.

It's a classic joke, but it illustrates a serious point: a single outlier can completely distort our understanding of data, and the "average" can describe a reality that doesn't actually exist for anyone in the room. In data analysis, these problems aren't funny. They lead to wrong conclusions, failed predictions, and poor decisions. Before we can trust any analysis, we need to understand what our data actually looks like, not just what summary statistics tell us it looks like.

Real-world data is messy. Missing values, outliers, measurement errors, and unknown distributions are the norm rather than the exception. Data preparation (transforming raw data into something we can analyze confidently) encompasses many challenges: ensuring data integrity, handling duplicates, merging sources, managing formats, addressing temporal dependencies, and dealing with high dimensionality, among others.

This chapter focuses specifically on **data quality** and **distributional properties**. These two aspects are foundational: quality issues determine whether data accurately represents reality, while distributional properties determine which statistical methods we can validly apply. What makes them particularly challenging is how they interact. Missing values and outliers can distort distributions, making normal data appear skewed. Conversely, trying to assess distributions on dirty data leads to wrong conclusions about what the data actually looks like. When preparation goes wrong, the consequences go beyond technical invalidity. We may spot patterns that aren't real, miss relationships that matter, or base important decisions on faulty interpretations.

The following sections build the conceptual foundation and practical skills for tackling these challenges. We'll explore how quality issues manifest, how distributions behave, and how to navigate their interdependencies. The goal isn't perfection but developing the judgment to make sound preparation decisions.

---

2. Theory and Background

Understanding data preparation requires grounding in two foundational areas: data quality theory and statistical distribution theory. While each has been extensively studied independently, their interaction is often overlooked in practice. Critically, effective data preparation can't be purely mechanical. It requires understanding the **context** in which data was collected, the domain it represents, and the business or scientific questions it will answer.

2.1 Data Quality: Dimensions and Mechanisms

Data quality is multidimensional. Each dimension captures a different aspect of how well data represents reality, but assessing each dimension requires **contextual understanding** of what the data represents and how it was collected.

Completeness refers to whether data contains all expected values. However, understanding *why* data is missing requires context. Three fundamental mechanisms explain missing data:

- **MCAR (Missing Completely at Random)**: Missingness is independent of both observed and unobserved data. Example: randomly dropped test tubes. This is the least problematic mechanism and can be tested statistically.
- **MAR (Missing at Random)**: Missingness depends on observed data but not on the missing values themselves. Example: younger respondents skip income questions regardless of their actual income. Identifying MAR often requires understanding respondent behavior and survey context.
- **MNAR (Missing Not at Random)**: Missingness depends on the unobserved values. Example: high earners specifically avoid income questions. Detecting MNAR is nearly impossible without domain knowledge about why people might hide certain information.

Context matters profoundly for missing data. In a healthcare dataset, missing lab results might mean tests weren't ordered (patient too healthy), results were normal (not recorded), or data transfer failed (technical issue). Each scenario requires different handling. Understanding the data collection process, system architecture, and domain practices is essential.

Accuracy concerns whether data correctly represents real-world entities. Inaccuracy arises from systematic measurement errors (instrument calibration issues, human biases), random measurement noise, data entry mistakes (typographical errors, transposed digits), and outdated information. **Domain expertise is critical** for identifying accuracy issues: a body temperature of 110°F is medically impossible, while a customer purchasing 1,000 units might be an error in a retail context but normal in B2B sales.

Consistency addresses contradictions within or across datasets. Internal inconsistencies (age = 25 but account_age = 30 years) are often detectable statistically, but understanding whether they represent errors or legitimate scenarios (business accounts vs. individual accounts) requires business context. Cross-dataset inconsistencies may reflect different measurement timepoints, system definitions, or valid business processes rather than errors.

Validity ensures data conforms to defined formats, types, and constraints. However, what constitutes "valid" depends entirely on context. A transaction amount of $0 might be invalid in a sales system but valid in a returns/refunds context. Age over 100 might seem invalid but could be legitimate in certain healthcare datasets. **Business rules** define validity, not just statistical patterns.

Outliers and Their Interpretation: Outliers are tricky because they might be errors or valid extreme values. They can be univariate (unusual in one dimension) or multivariate (unusual in combination). Detection methods include Z-score methods (assume normality), IQR methods (more robust), and modern approaches like Isolation Forest.

However, statistical detection must be followed by contextual investigation. **Domain knowledge is your primary tool** for figuring out what to do. Measurement errors should be corrected or removed, while valid extremes should generally be kept. For example, an extremely dissatisfied customer with many support tickets isn't an outlier to delete but potentially the most important case to study. A support ticket taking 200 hours might be a decimal point error (20.0 hours) or a legitimate complex case. Context determines the answer.

2.2 Statistical Distributions: Theory and Testing

Most parametric statistical methods (t-tests, ANOVA, regression) assume data or residuals follow specific distributions, typically normal. These assumptions aren't arbitrary. They're mathematically necessary for deriving valid test statistics, p-values, and confidence intervals. When violated, Type I and Type II error rates increase, parameter estimates may be biased, and confidence intervals may have incorrect coverage. However, the Central Limit Theorem provides some safety net: with large samples (typically $n > 30$), many methods stay reasonably valid even with moderate departures from normality.

Common Distributions in Real Data: Different processes produce different shapes, and **understanding the domain** helps predict what to expect. Normal distributions arise from additive processes and many small independent effects (human height in similar populations, measurement errors). Log-normal distributions describe multiplicative processes (income, house prices grow through percentage changes, not fixed amounts). Exponential distributions model time between events (customer arrivals, equipment failures). Poisson distributions describe count data (website visits per hour, defects per product). Heavy-tailed distributions appear in power-law phenomena (wealth distribution follows the "80-20 rule").

Importantly, bimodal distributions often indicate mixture populations (satisfied vs. dissatisfied customers) and reflect meaningful structure rather than problems. **Domain context** reveals whether bimodality is a data artifact or represents real subgroups that should be analyzed separately.

Assessment Methods combine visual and statistical approaches. Visual methods include histograms (overall shape), Q-Q plots (compare your data against theoretical distributions), and box plots (symmetry and outliers). Statistical tests include Shapiro-Wilk (most powerful for normality, especially with samples under 5,000), Kolmogorov-Smirnov (works for any distribution comparison), and Anderson-Darling (puts more weight on the tails). Descriptive statistics like skewness (measures asymmetry) and kurtosis (measures tail heaviness) complement these approaches. Sample size matters: large samples flag tiny deviations, small samples might miss real problems.

Context guides interpretation: Statistical tests tell you if data deviates from theoretical distributions, but domain knowledge tells you whether that deviation matters. For customer satisfaction scores (scale 1-10), slight skewness might be unimportant, but extreme bimodality might reveal critical insights about two distinct customer segments.

Transformations: Power transformations can help when data violates distributional assumptions. Common ones include logarithmic (for right-skewed data from multiplicative processes), square root (for count data), and Box-Cox (which automatically finds the best transformation parameter). The Yeo-Johnson transformation extends this to handle negative values.

However, transformations aren't always the right answer, and **context determines when to transform**. Bimodal data can't be made unimodal and shouldn't be forced. Transformed scales (log dollars vs. dollars) may be hard for stakeholders to understand even if statistically better. When non-normality reflects real, meaningful structure (like those bimodal satisfaction scores), keeping it provides more insight than transforming it away. Sometimes robust non-parametric methods are better, especially when keeping the original scale helps interpretation.

2.3 The Critical Interaction

Here's the key insight often missed: quality and distribution interact bidirectionally, and **context and domain expertise** are essential for navigating this interaction.

Quality Issues Distort Distributions: Missing data creates selection bias. Imagine a truly normal variable with mean of 100 and standard deviation of 15. If values above 110 are systematically missing (MNAR), what you observe will look left-skewed with a lower mean and smaller variance. **Without understanding why data is missing**, you can't tell if the skewness is real or just an artifact.

Outliers from data entry errors inflate variance, cause normality tests to fail, and make distributions appear heavy-tailed when they're not. A single bad entry (age = 800 instead of 80) can completely change what the distribution looks like. Interestingly, measurement errors can have a paradoxical effect: they add noise that sometimes makes non-normal data appear *more* normal through the Central Limit Theorem.

The Logical Dependency: This creates a tricky circular problem. You need clean data to reliably assess distributions, but assessing quality issues is easier when you know what distribution to expect. Outlier detection methods assume certain distributions (Z-score assumes normality). Imputation methods use distributional assumptions. Even consistency checks often involve expectations about how data should be distributed. Understanding your data's structure helps you choose better imputation strategies, which then affects what the final distribution looks like.

Breaking the circle requires domain knowledge: Understanding how the data was collected helps you figure out why things might be missing. Knowing the business process tells you whether an outlier is probably an error or a valid extreme case. Subject matter expertise suggests what distribution shape you should expect based on what the data actually represents.

Think of it this way: what you observe in your data is the true underlying pattern mixed with distortions from quality problems. Missing data introduces selection bias (you're only seeing part of the picture). Outliers from errors inflate variance. Measurement noise adds static. Inconsistencies create artificial patterns. The goal of data preparation is to remove these distortions so you can see the real pattern, but you can't do that effectively with statistics alone. You need to understand both the statistical properties **and** the real-world context that generated the data.

Practical Implications: Don't trust what distributions look like in dirty data. Your initial checks are just exploratory, not definitive. Some things that look like distribution problems are actually quality problems in disguise. After you clean the data, check the distributions again because they might look completely different. Document what things looked like before and after so you have an audit trail.

Most importantly, **bring in domain expertise throughout**. Statistical tests can't tell you whether an extreme value is a typo or a real outlier. They can't tell you whether a pattern is meaningful or an artifact. They can't explain why data might be systematically missing. You need people who understand the context to make these calls. The best statistical methods in the world won't save you if you don't understand what the data represents. On the flip side, even simple statistical approaches work well when combined with good contextual understanding.

---

3. Problem Statement

The central problem in data preparation is this: **you cannot determine if your data supports valid statistical conclusions without first assessing its quality and distributional properties, yet assessing these properties reliably is difficult when the data has quality issues.** This creates a circular dependency that requires systematic investigation and informed judgment to resolve.

Specifically, when analyzing a dataset, you need to know:

- Whether missing data introduces bias that invalidates your conclusions
- Whether outliers represent errors that distort results or valid extremes that contain important information
- Whether variables meet the distributional assumptions of your intended statistical methods
- Whether apparent distributional violations are real properties of the data or artifacts of quality problems

The challenge is that these questions are interconnected. Missing data can make normal distributions appear skewed. Outliers from data entry errors can cause normality tests to fail. You can't reliably test distributions without addressing quality, but handling quality issues often requires understanding expected distributions.

3.1 Application Context 1: Pizza Sales Analysis

This analysis uses Plato's Pizza transaction data containing 12 variables across one year of orders (order details, pricing, timing, pizza characteristics). The restaurant manager needs insights about peak periods, sales patterns, and operational efficiency. Before answering business questions like "What's our average order value?" or "What are best-selling pizzas?", I must ensure the data quality and distributional properties support valid statistical analysis. Transaction data often contains entry errors (incorrect prices, impossible quantities), missing values (incomplete orders), timing anomalies, and variables with different distributional shapes (order counts, prices, times). Preparing

this data requires identifying which issues are data problems versus legitimate business patterns.

3.2 Application Context 2: Housing Price Prediction

The second analysis examines the Ames Housing dataset with 79 variables describing residential properties in Ames, Iowa. The goal is predicting final sale prices, but with so many features, outlier detection becomes critical. A house priced unusually high might be a data entry error, a luxury property with legitimate premium features, or genuinely mispriced in the market. Similarly, missing data in features like basement ceiling height or garage area might indicate the feature doesn't exist (no basement, no garage) or represent incomplete data collection. Statistical methods can flag unusual cases, but determining whether they're errors requiring correction or valid extremes requiring retention depends on understanding real estate markets and housing characteristics. This example particularly emphasizes the challenge of distinguishing data quality issues from legitimate variation in high-dimensional data.

---

4. Problem Analysis

4.1 Constraints and Assumptions

Constraints:

- **Sample size:** Cannot remove significant data (48,620 records provide strong statistical power)
- **Business requirements:** Need to answer operational questions (peak times, best sellers, average order value)
- **Computational:** Standard analysis tools (Python, pandas, scipy)
- **Interpretability:** Results must be actionable for restaurant management

Assumptions:

- **Data collection:** Transaction data assumed accurate from point-of-sale system
- **Completeness:** Initial inspection shows no missing values, but need verification
- **Outliers:** Unusual prices or quantities may be legitimate (bulk orders, promotions) or errors
- **Expected distributions:**
    - Revenue/prices: Likely right-skewed (log-normal)
    - Order counts: Temporal patterns expected (day of week, seasonal)

- Quantities: Discrete distribution (Poisson-like)

4.2 Approach and Logic

Chosen Sequence: Assessment → Outlier Detection → Distribution Analysis

Why this order?

1. **Initial Assessment:** No missing data found, so focus shifts to outliers and distributions
2. **Outlier Detection:** Identify extreme values in pricing and quantities before distribution testing
3. **Distribution Testing:** Validate assumptions on clean data

Key Principles:

- Domain knowledge essential (reasonable price ranges, typical order sizes)
- Visual + statistical validation
- Document all findings

Decision Framework:

For each numerical variable:

1. Check summary statistics (mean, median, range)
2. Visualize distribution (histogram, boxplot)
3. Identify outliers (IQR method, domain rules)
4. Investigate outliers (error vs. legitimate extreme)
5. Test normality (Shapiro-Wilk, Q-Q plot)
6. Transform if needed (log for revenue/prices)

---

5. Solution Explanation

5.1 Pipeline Overview

Three-Stage Systematic Approach:

Stage 1: Initial Assessment → Stage 2: Outlier Investigation → Stage 3: Distribution Validation

Each stage builds on the previous, with validation checkpoints to ensure data quality improvements don't introduce new artifacts.

5.2 Implementation Framework

Stage 1: Initial Data Assessment

Objective: Understand the data landscape before making any modifications.

Key Activities:

- Check for missing values across all variables
- Calculate summary statistics (mean, median, range, standard deviation)
- Identify data types and verify logical consistency
- Create initial visualizations (histograms for continuous variables)

Output: Complete picture of data quality issues and initial distributional characteristics.

Stage 2: Outlier Investigation

Objective: Identify and classify extreme values as errors or legitimate observations.

Detection Methods:

- Statistical: IQR method (values beyond Q1-1.5×IQR or Q3+1.5×IQR)
- Domain-based: Business rules (valid price ranges, reasonable quantities)
- Visual: Boxplots to identify extreme observations

Classification Framework:

For each flagged outlier:

IF violates business rules (e.g., negative price):

→ Data error, requires correction/removal

ELIF within possible range but unusual:

→ Investigate context (bulk order? promotion?)

ELSE:

→ Valid extreme, retain for analysis

Stage 3: Distribution Analysis and Validation

Objective: Determine if variables meet statistical assumptions and identify appropriate transformations.

Assessment Process:

- Visual inspection: Histograms, Q-Q plots, boxplots

- Statistical tests: Shapiro-Wilk for normality (with awareness of large sample sensitivity)
- Descriptive statistics: Skewness and kurtosis calculations

Transformation Strategy:

For right-skewed continuous variables (revenue, prices):

→ Apply log transformation

→ Re-test for normality

For count data (quantities, orders):

→ Check Poisson/negative binomial fit

→ No transformation typically needed

For categorical data:

→ Frequency distributions

→ Chi-square tests for independence

Validation at Each Stage:

- Document findings before and after each step
- Verify transformations improve distributional properties
- Confirm business interpretability of results

---

6. Results and Discussion

6.1 Results

Data Quality Assessment:

- **Complete dataset:** 48,620 records with 0% missing values
- **No duplicates:** All transactions unique
- **Valid ranges:** Prices $9.75-$83.00, quantities 1-4 pizzas
- **Conclusion:** High-quality dataset suitable for analysis

Table 1: Key Metrics

Metric	Value
Total Records	48,620
Missing Values	0 (0%)
Total Revenue	$817,860
Average Order Value	$38.31
Total Pizzas Sold	49,574

Distribution Findings:

- **Revenue:** Right-skewed (typical for monetary data)
- **Category Performance:** Classic leads (30% of orders, $220,053 revenue)
- **Size Preference:** Large pizzas dominant (38% of orders)
- **Temporal Patterns:** Friday peak (8,106 orders), Spring highest revenue ($210,537)

Figure 1: Revenue by Category (Classic dominates)
Figure 2: Pizza Size Distribution (Large 38%)
Figure 3: Category Distribution Pie Chart (relatively balanced)

6.2 Discussion

Key Findings:

1. Exceptional Data Quality
Perfect completeness unusual for real-world data. Verified with business to confirm no missing transactions. Suggests well-designed POS system with mandatory fields.

2. Revenue Patterns Match Theory
Right-skewed distribution confirms log-normal expectation for monetary data. High-volume/low-price (Classic) vs. premium items (Chicken) create expected distribution shape.

3. Business Insights from Clean Data

- Friday staffing critical (17% of weekly orders)
- Classic pizzas drive volume, Chicken drives margins
- Spring/Summer opportunities for promotional campaigns

Practical Impact:
Clean data enables:

- Valid hypothesis testing (comparing days/seasons)
- Regression modeling for sales forecasting
- Reliable operational decisions

Limitations:

- Single year of data (can't validate long-term trends)
- No customer demographics (can't segment behavior)
- Aggregated transactions (individual patterns unknown)

Conclusion:
This dataset exemplifies how quality data reveals clear distributional patterns and enables confident statistical analysis. The theoretical framework (quality → distribution → valid inference) was validated: complete data allowed immediate distribution assessment, which confirmed expected patterns (right-skewed revenue, discrete quantities), enabling actionable business recommendations.

---

7. References

Foundational Statistical Texts:

1. Batini, C., & Scannapieco, M. (2016). *Data and Information Quality: Dimensions, Principles and Techniques*. Springer International Publishing.
2. Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198-1202.
3. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
4. Van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). CRC Press.

Distribution Testing and Transformation:

5. Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2), 211-252.
6. D'Agostino, R. B., & Stephens, M. A. (Eds.). (1986). *Goodness-of-Fit Techniques*. Marcel Dekker.
7. Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591-611.

Outlier Detection:

8. Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 73-79.

Data Quality and Preparation:

9. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Cengage Learning.
10. Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(6), 1-12.

Statistical Software and Implementation:

11. McKinney, W. (2012). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.
12. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
13. Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 57, 61.

Dataset Source:

14. Yeole, D. (2024). Cheesy Conclusions: Analyzing Pizza Sales. *Kaggle Dataset*. Retrieved from https://www.kaggle.com/datasets/