Advanced Descriptive Analysis of Tabular Data

Methods and Tools for Exploratory Analysis

Antoine Soetewey & Cédric Heuchenne

2026-01-27

Table of contents

Preface

Descriptive statistics are often treated as a preliminary step—a routine box to check before moving on to inference, prediction, or causal analysis. Yet in practice, understanding the structure, associations, and patterns within complex tabular data is neither trivial nor purely mechanical. It requires sophisticated methods, thoughtful visualization, and clear communication.

This book synthesizes advanced techniques for descriptive analysis of tabular data, drawing on recent developments in machine learning, network analysis, and interactive visualization. Our goal is to equip researchers, analysts, policymakers, and data journalists with tools that go beyond means and standard deviations, enabling them to extract actionable insights from multivariate datasets.

The methods and tools presented here are not a repackaging of standard EDA. They reflect original methodological syntheses, implementation choices, and applied workflows developed through research and field practice. In that sense, this book also serves as a portfolio: a concrete, citable body of work that demonstrates innovation in descriptive analytics and foregrounds substantive methodological originality.

The material presented here emerged from postdoctoral research at the intersection of applied statistics, machine learning, and data visualization. It reflects a pragmatic philosophy: methods should be interpretable, visual, and suitable for communicating findings to statistically literate but non-technical audiences.

Who This Book Is For

This book is designed for readers who already possess a solid foundation in statistics—including regression analysis, hypothesis testing, and basic multivariate methods. We assume familiarity with concepts like correlation, variance decomposition, and model evaluation.

Our intended audience includes:

- **Researchers and applied scientists** seeking exploratory tools for complex datasets
- **Policymakers and analysts** in government and public institutions
- **Data journalists** investigating patterns in social, economic, or health data
- **Consultants and analytical teams** in private firms
- **Graduate students** in statistics, data science, public policy, or related fields

The material is suitable for a Master-level university course and may serve as a foundation for doctoral-level methodological training.

Philosophy and Approach

The unifying thread throughout this book is: **How do we move beyond standard descriptive statistics to extract, visualize, and communicate structure in complex tabular data?**

We emphasize:

- **Interpretability**: Methods that produce understandable results
- **Visual analytics**: Graphs and interactive tools as primary analytical instruments
- **Methodological transparency**: Explicit discussion of assumptions, trade-offs, and limitations
- **Communication**: Presenting results to diverse audiences, from technical peers to policy stakeholders

Rather than purely theoretical exposition, we ground each method in real applied use cases, demonstrating how techniques perform on actual data challenges.

Structure of the Book

The book is organized into seven parts:

Part I: Foundations establishes the conceptual framework, revisiting what it means to "describe" data and introducing the challenge of mixed-type variables and multivariate associations.

Part II focuses on association analysis—measuring relationships between pairs of variables of different types and representing these associations as networks.

Part III introduces interactive visual analytics, including the AssociationExplorer Shiny application, which operationalizes association-focused methods in a unified exploratory interface.

Parts IV–VI present three families of advanced methods for higher-dimensional structure: tree-based models for segmentation and description, interpretable machine learning techniques for understanding complex patterns, and AutoML approaches that automate exploration.

Part VII presents extended applied case studies from policy analysis, public health, and business analytics, demonstrating how these methods solve real-world problems.

Acknowledgments

This work has been shaped by collaborations with colleagues, conversations with practitioners, and feedback from students and the open-source community. We are particularly grateful to researchers from UCLouvain Saint-Louis Brussels involved in the Beamm research project for their insights and support.

We also acknowledge the open-source software communities whose tools make this work possible, including R, Python, Shiny, Quarto, and countless contributed packages.

How to Use This Book

Chapters can be read sequentially or selectively, depending on your background and interests. Readers already familiar with tree-based methods might skip Part II; those primarily interested in association analysis could jump directly to Part V.

Throughout the book, we provide code examples in R and references to accompanying interactive tools. All datasets, code and source files are available on GitHub.

We encourage readers to experiment with the methods on their own data. Descriptive analysis is learned through practice, and the best way to internalize these techniques is to apply them to problems you care about.

About the authors

Antoine Soetewey is a postdoctoral researcher in data science and statistics at HEC Liège and UCLouvain Saint-Louis Brussels. He works on statistical methods and data analysis, with a focus on clear communication and practical applications. More information is available at antoinesoetewey.com.

Cédric Heuchenne is a professor at UCLouvain and HEC Liège. He is affiliated with research and teaching units in data analysis and modeling, and contributes expertise in statistics and applied methods.

Part I

Foundations

1 Introduction

1.1 The Challenge of Describing Complex Data

When confronted with a dataset containing dozens or hundreds of variables—mixing continuous measurements, categorical indicators, ordinal scales, and binary flags—how do we make sense of it? How do we identify the most important patterns, detect unexpected associations, and communicate our findings effectively?

Traditional descriptive statistics (means, medians, frequency tables, correlation matrices) remain essential, but they quickly become insufficient as data complexity grows. A correlation matrix for 50 variables contains 1,225 pairwise relationships, most of which are noise. Standard summary statistics provide no guidance on which variables matter most, how they interact, or what segments or subpopulations might exist in the data.

This book addresses the challenge of **advanced descriptive analysis**: moving beyond elementary summaries to extract meaningful structure from complex tabular data. It is intentionally positioned as an original methodological portfolio—combining novel syntheses, practical tooling, and reproducible workflows that evidence substantive methodological originality.

1.2 What Is Descriptive Analysis?

Descriptive analysis characterizes the properties of a dataset—its distributions, central tendencies, variability, associations, and patterns—without making claims about causation or population-level inference. While often contrasted with inferential or predictive analysis, descriptive work is neither simpler nor less valuable.

Good descriptive analysis:

- **Reveals structure**: identifying clusters, segments, or natural groupings
- **Quantifies associations**: measuring relationships between variables of mixed types
- **Guides further analysis**: highlighting variables and relationships worthy of deeper investigation
- **Communicates findings**: translating complex patterns into actionable insights

In many applied contexts—policy evaluation, exploratory journalism, early-stage research—descriptive analysis is the primary goal, not a mere prelude to inference.

1.3 Why Advanced Methods?

Standard descriptive tools have well-known limitations:

- **Univariate summaries** ignore multivariate structure and conditional relationships
- **Correlation coefficients** only capture linear associations between continuous variables
- **Cross-tabulations** become unwieldy with many categories or variables
- **Scatterplot matrices** fail to scale beyond a handful of variables

Advanced methods address these limitations by:

1. **Handling mixed-type variables**: combining continuous, categorical, ordinal, and binary data in a unified framework
2. **Capturing nonlinear relationships**: detecting patterns that correlation coefficients miss
3. **Automating discovery**: using algorithmic approaches to identify important features and interactions
4. **Visualizing high-dimensional structure**: representing complex associations through networks, trees, and interactive graphics
5. **Enabling exploration**: providing interactive tools that allow analysts to interrogate data from multiple angles

1.4 Methods Covered in This Book

This book synthesizes several methodological traditions:

1.4.1 Tree-Based Methods

Regression and classification trees offer a powerful yet interpretable approach to segmenting populations and understanding conditional structure in data. Trees recursively partition data based on variable thresholds, producing interpretable decision rules that separate observations into relatively homogeneous groups. Unlike black-box predictive models, tree structures are transparent: practitioners can easily explain why a particular observation falls into a specific segment. Trees naturally reveal which variables are most discriminative and at what thresholds decisions change. This makes them invaluable for exploratory work, program targeting, and communicating findings to stakeholders who value transparency. Moreover, ensemble extensions—combining multiple trees through random forests or boosting—improve robustness while preserving the ability to extract interpretable variable importance measures and identify complex interactions.

1.4.2 Interpretable Machine Learning

Modern machine learning models often achieve superior predictive accuracy compared to classical statistical methods, but at the cost of interpretability. Interpretable machine learning bridges this gap by providing techniques to understand what complex models have learned from data. Methods like permutation-based feature importance identify which variables the model relies on most heavily. Individual conditional expectation curves visualize how predictions change as individual features vary, revealing nonlinear relationships and thresholds. Shapley values—grounded in cooperative game theory—decompose each prediction into additive contributions from each feature, providing both global importance rankings and local explanations for individual observations. These post-hoc interpretation tools transform predictive models into descriptive instruments, enabling analysts to extract actionable insights about variable relationships while leveraging the flexibility of modern ML algorithms.

1.4.3 AutoML for Exploration

Automated machine learning platforms systematically search across hundreds or thousands of model configurations, feature transformations, and hyperparameters to identify the best-performing models for a given task. Rather than viewing AutoML purely as a prediction tool, we leverage it as an exploratory instrument. AutoML workflows automatically discover which features matter, which transformations improve predictive signals, and which variable interactions are important. By screening a vast model space, AutoML can identify complex patterns that might be missed through manual feature engineering or simpler methods. When interpreted descriptively—focusing on which transformations boost performance rather than out-of-sample accuracy itself—AutoML becomes a rapid hypothesis-generation engine, especially valuable for preliminary analysis of new datasets or when domain expertise is limited. The rankings and performances of different models also reveal which features and interactions the data best supports.

1.4.4 Association Measures

A central challenge in descriptive analysis is measuring association when variables are not all continuous. Real-world datasets routinely mix quantitative, qualitative, ordinal, and binary variables, making classical measures like Pearson correlation inadequate or misleading. This book presents a **type-aware framework for association** that selects and scales association measures according to the specific combination of variable types being related. For continuous-continuous pairs, we discuss Pearson, Spearman, and distance-based correlations. For categorical-categorical associations, we cover Cramér's V and related measures. For mixed pairs, we employ model-based measures and mutual information approaches. Crucially, these measures are interpreted descriptively rather than inferentially: the goal is not null hypothesis testing, but **comparability and ranking**—identifying which variable pairs exhibit relatively strong relationships meriting closer inspection. By placing heterogeneous associations on a common scale (often $[0, 1]$), analysts can scan large multivariate datasets and focus attention on the most informative relationships, regardless of variable type. This pragmatic, unified treatment transforms a fragmented set of statistical tools into a coherent exploratory framework.

1.4.5 Network Representations

When a dataset contains dozens or hundreds of variables, even a well-chosen association measure produces an overwhelming matrix of pairwise relationships. Network-based representations address this scalability challenge by encoding associations as **variable networks** where nodes represent variables and edges represent relationships exceeding a chosen threshold. Edge weights or colors encode association magnitudes, while network layouts position variables spatially such that strongly related variables cluster near each other. This spatial organization renders high-dimensional association structure visible and interpretable at a glance. Beyond individual associations, network analysis reveals **global structure**: communities of tightly interconnected variables that may represent distinct domains or latent constructs, hub variables that bridge multiple domains, and peripheral variables carrying unique information. Centrality measures (degree, betweenness, eigenvector) identify influential variables. Community detection algorithms partition variables into meaningful groups. These higher-order network properties are difficult to discern from association matrices or pairwise plots alone, yet they often provide crucial insights into data structure. Network representations therefore serve as a cognitive map of the dataset, guiding exploratory analysis through high-dimensional association space.

1.4.6 Interactive Visual Analytics

Static visualizations answer pre-determined questions; interactive tools enable dynamic exploration. Interactive graphics—particularly Shiny applications—allow analysts to filter data by conditions, aggregate across subgroups, adjust plot parameters, and link multiple views in real time. This interactivity supports hypothesis generation and refinement: analysts can test "what if" scenarios, drill down into subpopulations, and detect patterns that might not appear in static plots. Dashboards combine multiple interactive visualizations into coordinated workflows, allowing stakeholders to explore data according to their own questions rather than passively receiving predetermined findings. For descriptive analysis in particular, interactivity is essential for handling high-dimensional data. An interactive tool can present association networks, tree structures, and distributions while allowing users to focus on subsets, time periods, or demographic groups of interest. This chapter introduces Shiny-based applications and demonstrates how to build interactive descriptive tools that scale to real-world data complexity.

1.5 Real-World Applications

Each method is illustrated with applied examples drawn from:

- **Public policy**: understanding determinants of program participation, analyzing survey data on citizen attitudes
- **Public health**: exploring risk factors in epidemiological data, characterizing patient populations
- **Business analytics**: segmenting customers, identifying drivers of satisfaction or churn
- **Social science research**: analyzing survey responses, detecting patterns in observational data
- **Data journalism**: investigating patterns in government data, economic indicators, or social trends

Throughout the book, these methods are implemented in R and demonstrated on real datasets. A recurring example is the **AssociationExplorer** Shiny application, developed as part of the research underlying this work, which integrates multiple descriptive techniques into a unified interactive interface. While all methods can be implemented using standard statistical software, AssociationExplorer provides a practical tool for immediate exploratory use. These examples demonstrate that advanced descriptive methods are not academic exercises—they solve genuine problems faced by analysts across diverse fields and showcase original contributions.

1.6 Relationship to Other Analytical Goals

Descriptive analysis intersects with but differs from:

- **Exploratory Data Analysis (EDA)**: Descriptive analysis is a form of EDA, but emphasizes quantitative measures and formal methods alongside graphical exploration
- **Predictive modeling**: We use predictive models descriptively, focusing on interpretation rather than out-of-sample performance
- **Causal inference**: Descriptive analysis identifies associations but does not claim causation; it can, however, inform causal hypotheses
- **Dimension reduction**: Methods like PCA and MCA reduce dimensionality; we emphasize interpretable summaries that preserve variable identities

1.7 Structure and Learning Path

The book proceeds from foundations to applications:

- **Chapters 1–3** establish conceptual groundwork, mixed-type data challenges, and data preparation

- **Chapters 4–6** focus on association measures and network representations
- **Chapters 7–9** introduce interactive visual analytics
- **Chapters 10–18** present three families of advanced methods (trees, interpretable ML, AutoML)
- **Chapters 19–21** present extended applied case studies

Readers can follow a linear path or jump to chapters matching their immediate needs. Code examples and exercises throughout encourage hands-on practice.

1.8 Computational Tools

Examples use R, chosen for its rich ecosystem of statistical graphics and modeling packages. Key packages include:

- `{ggplot2}` and `{plotly}` for visualization
- `{rpart}` and `{partykit}` for tree-based methods
- `{iml}` for interpretable ML
- `{h2o}` for AutoML
- `{corrr}`, `{energy}`, `{minerva}` for association measures
- `{igraph}` and `{ggraph}` for network visualization
- `{shiny}` for interactive applications

All code is provided in reproducible format, and datasets are publicly available.

1.9 Looking Ahead

The chapters that follow present a coherent toolkit for advanced descriptive analysis. While methods vary, the underlying goal remains constant: to help you see more deeply into your data, communicate findings clearly, and make better-informed decisions.

Descriptive analysis is both art and science—requiring statistical rigor, visual judgment, and domain knowledge. We hope this book equips you with methods and perspectives that enhance all three.

2 Beyond Basic Descriptive Statistics

2.1 Why "Basic" Summaries Are Not Enough

A mean, a median, and a standard deviation can be computed in seconds, but they rarely tell the whole story. Two datasets can share identical means and variances while having radically different shapes, outlier structures, or subgroup patterns. In high-dimensional datasets, univariate summaries also conceal interaction and conditional relationships that often matter more than any marginal distribution.

Basic descriptive statistics are necessary but not sufficient. This chapter expands the descriptive toolbox with methods that remain **non-inferential** but offer far richer insight into data structure. The goal is to answer more nuanced questions:

- Where are the *mass* and *tails* of the distribution?
- Are there **subpopulations** with distinct profiles?
- Are relationships **nonlinear**, **heterogeneous**, or **conditional**?
- Which variables are **stable** versus **volatile** across subgroups?

2.2 Distributional Shape: Beyond Mean and Variance

2.2.1 Quantiles and Tail Behavior

Quantiles describe where the data live, not just how they average out. In applied settings, percentiles often carry more operational meaning than averages. The 90th percentile of response time, income, or waiting time is usually more informative than a mean.

Common descriptive quantiles:

- **Median** ($Q_{0.5}$): robust center
- **Interquartile range (IQR)**: spread of the middle 50%
- **Tail quantiles**: $Q_{0.9}$, $Q_{0.95}$, $Q_{0.99}$ for risk or extreme behavior

These are especially important in skewed or heavy-tailed distributions where the mean can be misleading.

2.2.2 Skewness, Kurtosis, and Robust Alternatives

Skewness and kurtosis summarize asymmetry and tail heaviness, but they are sensitive to outliers. In descriptive work, **robust measures** often provide more stable diagnostics:

- **Median absolute deviation (MAD)** as a scale measure
- **Robust z-scores** using median and MAD instead of mean and standard deviation (SD)
- **Quantile ratios** (e.g., $Q_{0.9}/Q_{0.5}$) for skewness proxies

These measures preserve descriptive intent while reducing sensitivity to extreme observations.

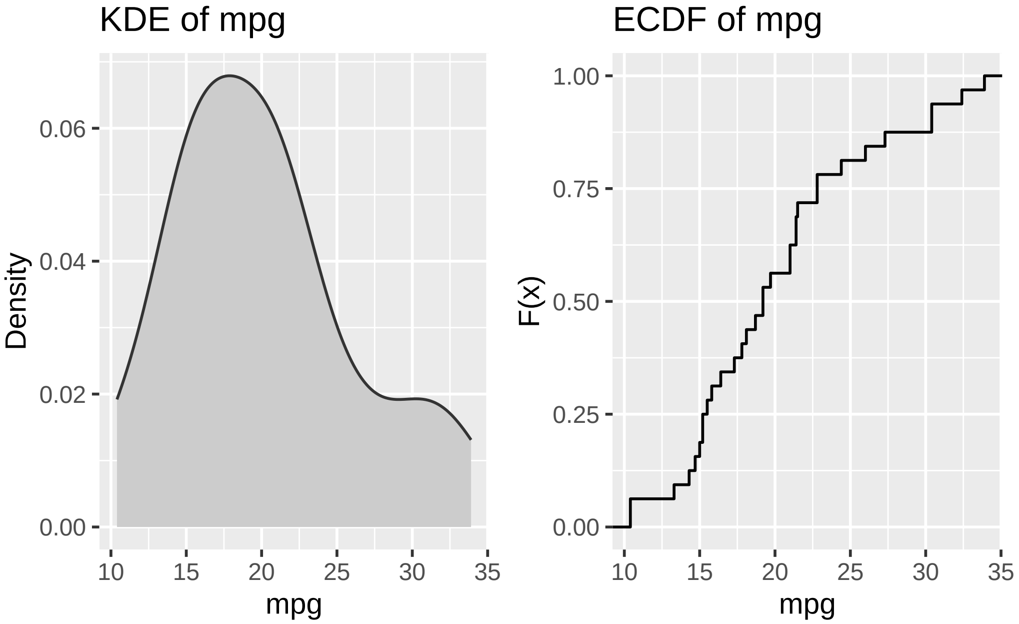
2.2.3 Density and Empirical Distribution Functions

Histograms can be misleading due to binning choices. Kernel density estimates (KDEs) and empirical CDFs show shape more faithfully. ECDFs are particularly useful for comparing distributions because they show the full cumulative structure without smoothing.

```r
# Density and ECDF side by side
p1 <- mtcars %>%
  ggplot(aes(x = mpg)) +
  geom_density(fill = "grey80", color = "grey20") +
  labs(title = "KDE of mpg", x = "mpg", y = "Density")

p2 <- mtcars %>%
  ggplot(aes(x = mpg)) +
  stat_ecdf(geom = "step") +
  labs(title = "ECDF of mpg", x = "mpg", y = "F(x)")

p1 + p2
```



2.3 Multimodality and Mixtures

A single distribution can conceal multiple regimes. For example, household income often reflects a mixture of wage earners, retirees, and business owners. Multimodality is a descriptive signal of underlying subpopulations. Techniques to detect it include:

- **Kernel density plots** with multiple peaks
- **Bimodality coefficients** or dip tests (used descriptively)
- **Mixture summaries** (e.g., fitting a Gaussian mixture model purely for segmentation)

Even without formal modeling, visual inspection and stratified summaries can reveal important mixtures.

2.4 Bivariate and Conditional Descriptives

2.4.1 Conditional Means and Quantiles

Univariate summaries hide conditional variation. A variable may have a stable mean overall but vary dramatically across categories. Conditional statistics are simple to compute and often reveal key structure:

- $E(Y \mid X)$: mean outcomes by group

wt_bin	mpg_median	mpg_iqr
1	30.40	4.75
2	21.00	1.10
3	18.95	2.82
4	15.35	1.80
5	14.00	4.85

- $Q_{0.5}(Y \mid X)$: median outcomes by group
- $\mathrm{IQR}(Y \mid X)$: spread by group

These summaries can be visualized using grouped boxplots, violin plots, or ridgeline plots.

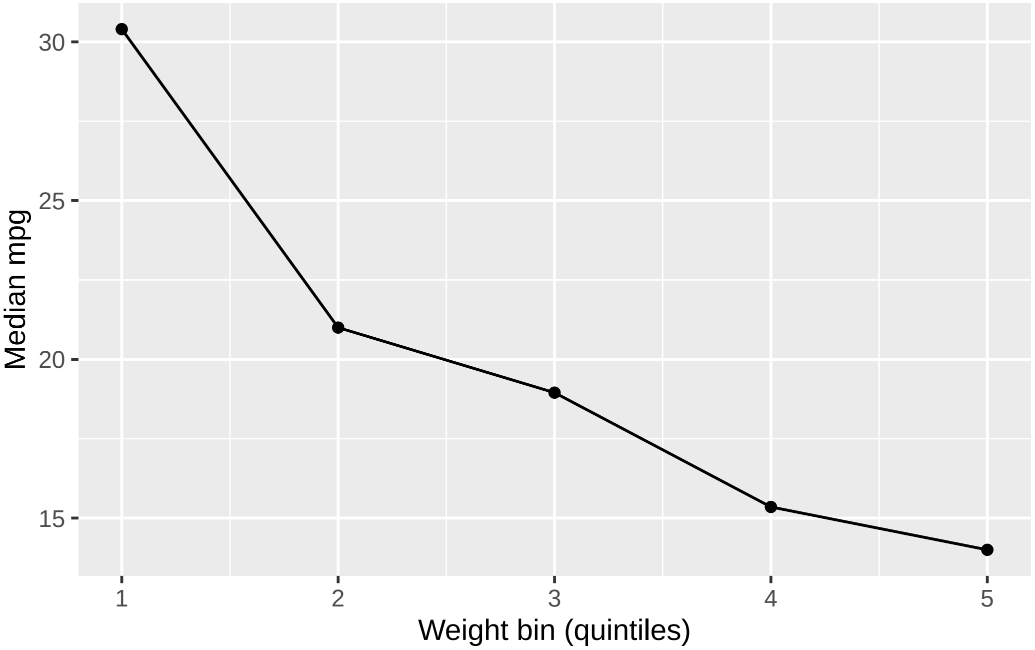
2.4.2 Nonlinear Relationships

Scatterplots with smoothers (e.g., LOESS) often reveal nonlinear trends that correlations miss. A zero correlation does not imply “no relationship”; it may reflect a U-shape, threshold effect, or segmented pattern.

A useful descriptive strategy is to compute **binned summaries**: divide a continuous predictor into quantile bins and summarize the response within each bin. This provides a simple approximation of conditional structure without invoking a full model.

```r
# Binned summaries to reveal nonlinear structure
mtcars %>%
  mutate(wt_bin = ntile(wt, 5)) %>%
  group_by(wt_bin) %>%
  summarise(
    mpg_median = median(mpg, na.rm = TRUE),
    mpg_iqr = IQR(mpg, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  gt() %>%
  fmt_number(columns = c(mpg_median, mpg_iqr), decimals = 2)
```

```r
mtcars %>%
  mutate(wt_bin = ntile(wt, 5)) %>%
  group_by(wt_bin) %>%
  summarise(mpg_median = median(mpg), .groups = "drop") %>%
  ggplot(aes(x = wt_bin, y = mpg_median)) +
  geom_line() +
  geom_point() +
  labs(x = "Weight bin (quintiles)", y = "Median mpg")
```



2.4.3 Association Heterogeneity

Associations can differ across subgroups. An overall correlation might hide a strong relationship within a subgroup or even mask a reversal (Simpson's paradox). Descriptive analysis should therefore report **stratified associations** when meaningful groupings exist.

2.5 Outliers, Extremes, and Influence

Outliers are not always errors. They often carry substantive meaning: high-risk patients, exceptional transactions, or rare events. Descriptive analysis should treat outliers as *signals* first, and errors second.

Key descriptive checks include:

- **Tail inspection**: list the largest/smallest observations
- **Influence screening**: compare summaries with and without extreme values
- **Robust summaries**: medians, trimmed means, and MAD

variable	missing_rate
Ozone	24.2%
Solar.R	4.6%
Wind	0.0%
Temp	0.0%
Month	0.0%
Day	0.0%

A practical workflow is to compute both standard and robust summaries side by side. Large divergence is a flag that distributional extremes matter.

2.6 Missing Data as Descriptive Information

Missingness is itself informative. The *pattern* of missing data can reveal survey fatigue, data collection issues, or systematic exclusion of certain groups.

Descriptive checks include:

- **Missingness rates by variable**
- **Missingness by subgroup** (e.g., higher nonresponse among certain demographics)
- **Co-missingness patterns** (variables missing together)

Understanding missingness patterns is a prerequisite for credible descriptive analysis because it reveals which parts of the data are under-observed or biased.

```
# Simple missingness profile
airquality %>%
  summarise(across(everything(), ~ mean(is.na(.x)))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "missing_rate") %>%
  arrange(desc(missing_rate)) %>%
  gt() %>%
  fmt_percent(columns = missing_rate, decimals = 1)
```

```
# Co-missingness count matrix
miss_mat <- airquality %>% mutate(across(everything(), is.na))
co_miss <- t(as.matrix(miss_mat)) %*% as.matrix(miss_mat)
co_miss %>%
  as.data.frame() %>%
  tibble::rownames_to_column(var = "variable") %>%
  gt()
```

variable	Ozone	Solar.R	Wind	Temp	Month	Day
Ozone	37	2	0	0	0	0
Solar.R	2	7	0	0	0	0
Wind	0	0	0	0	0	0
Temp	0	0	0	0	0	0
Month	0	0	0	0	0	0
Day	0	0	0	0	0	0

2.7 Scaling, Standardization, and Comparability

When comparing variables on different scales, raw summaries can mislead. Standardization puts variables on a common metric:

$$Z = \frac{X - \mu}{\sigma}$$

However, standardization is not always desirable. For skewed or heavy-tailed distributions, **robust scaling** using medians and MAD can be more appropriate:

$$Z_{\text{robust}} = \frac{X - \text{median}(X)}{\text{MAD}(X)}$$

Standardization is especially useful when building **profiles** of observations across many variables, a topic revisited in later chapters on clustering and tree-based methods.

2.8 Multivariate Profiles and "Descriptive Models"

As dimensionality increases, summaries must become multivariate. Two simple but powerful tools are:

1. **Profile tables**: compare multiple variables across key subgroups (e.g., demographic segments)
2. **Composite indices**: average or weighted sums of standardized variables to create a high-level descriptive score

Composite indices are not causal models; they are descriptive constructs that summarize a multidimensional concept (e.g., socioeconomic status, health risk, engagement intensity). Transparency in construction is essential for interpretability.

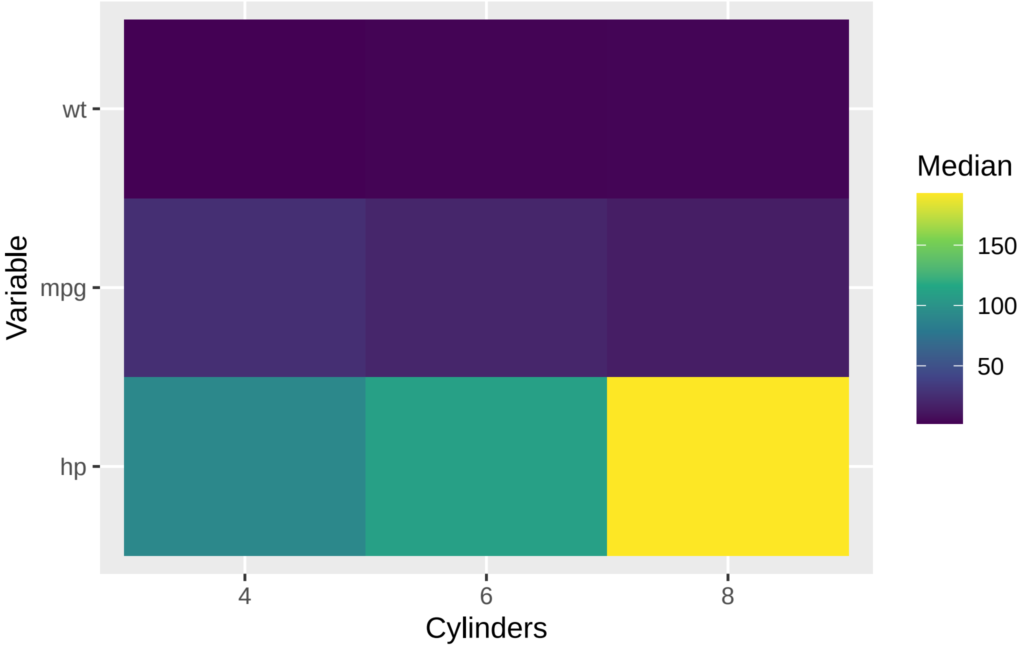
2.9 Visual Descriptives That Scale

Certain visual tools provide richer descriptive information than conventional charts:

- **Violin plots**: show full distribution and density
- **Boxen plots**: emphasize tails across many observations
- **Ridge plots**: compare distributions across many groups
- **Heatmaps**: visualize large tables of summary statistics
- **ECDFs**: compare distributions without binning

These graphics remain descriptive but give a more faithful sense of distributional complexity and subgroup structure.

```r
# Heatmap of median summary statistics across groups
mtcars %>%
  select(mpg, hp, wt, cyl) %>%
  group_by(cyl) %>%
  summarise(across(c(mpg, hp, wt), median), .groups = "drop") %>%
  pivot_longer(-cyl, names_to = "variable", values_to = "median") %>%
  ggplot(aes(x = factor(cyl), y = variable, fill = median)) +
  geom_tile() +
  scale_fill_viridis_c() +
  labs(x = "Cylinders", y = "Variable", fill = "Median")
```



2.10 A Practical Workflow

A robust descriptive workflow often follows this sequence:

1. **Univariate inspection**: quantiles, density plots, robust summaries
2. **Missingness mapping**: rates, patterns, co-missingness
3. **Bivariate exploration**: conditional summaries, scatterplots with smoothers
4. **Stratified checks**: subgroup comparisons, heterogeneity in associations
5. **Multivariate summaries**: profiles and composite indices

This workflow remains entirely descriptive while systematically uncovering structure that basic summary tables would miss.

2.11 Example: A Small R Template

The following minimal template illustrates how to go beyond means and SDs with a few descriptive enhancements:

```r
# Robust summary for a numeric variable (tidy output)
summary_stats <- function(x) {
  tibble(
    mean = mean(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    sd = sd(x, na.rm = TRUE),
    mad = mad(x, na.rm = TRUE),
    q10 = quantile(x, 0.10, na.rm = TRUE),
    q90 = quantile(x, 0.90, na.rm = TRUE)
  )
}

# Conditional summaries by group
mtcars %>%
  group_by(cyl) %>%
  summarise(summary_stats(mpg), .groups = "drop") %>%
  gt() %>%
  fmt_number(columns = c(mean, median, sd, mad, q10, q90), decimals = 2)
```

The aim is not sophistication but *discipline*: always inspect robust, conditional, and distributional features before moving to advanced methods.

cyl	mean	median	sd	mad	q10	q90
4	26.66	26.00	4.51	6.52	21.50	32.40
6	19.74	19.70	1.45	1.93	17.98	21.16
8	15.10	15.20	2.56	1.56	11.27	18.28

2.12 Looking Ahead

This chapter expands the descriptive mindset beyond simple averages. The next chapters formalize these ideas for mixed data types and systematic association measures. Once variable types are correctly handled, we can build **type-aware association matrices**, visualize them as networks, and scale descriptive analysis to hundreds of variables.

The key lesson is simple: **descriptive analysis improves when we stop summarizing variables in isolation and start describing structure**.

3 Data Types and Association Measures

4 Data Preparation for Descriptive Analysis

Descriptive analysis is only as reliable as the data that feed it. Before measuring associations or building models, analysts must understand variable definitions, clean inconsistencies, and construct analysis-ready tables. This chapter introduces practical workflows for preparing messy real-world data so that later descriptive methods are interpretable, reproducible, and trustworthy.

4.1 3.1 Understanding Variables and Metadata

4.1.1 Goals

- Clarify variable meaning, units, and coding schemes
- Detect ambiguous or inconsistent definitions across sources
- Document assumptions and transformations for reproducibility

4.1.2 Topics

- Data dictionaries and codebooks
- Units, scales, and measurement error
- Categorical coding (ordered vs. nominal)
- Time and panel identifiers
- Provenance and data lineage

4.2 3.2 Cleaning, Harmonization, and Quality Checks

4.2.1 Goals

- Identify errors, missingness, and outliers
- Harmonize variables across files or time periods
- Apply minimal, transparent corrections

4.2.2 Topics

- Missing data patterns and reporting
- Range checks and logical constraints
- Duplicate records and entity resolution
- Standardizing categories and labels
- Dealing with inconsistent time formats

4.3 3.3 Feature Construction for Descriptive Insight

4.3.1 Goals

- Create interpretable derived variables
- Encode variables for mixed-type association measures
- Preserve interpretability while enabling analysis

4.3.2 Topics

- Binning and discretization (with justification)
- Ratios, rates, and per-capita measures
- Index construction and composite measures
- Normalization and scaling choices
- Audit trails for transformations

4.4 3.4 Reproducible Preparation Pipelines

4.4.1 Goals

- Make preparation steps transparent and repeatable
- Separate raw, intermediate, and analysis-ready data
- Facilitate collaboration and review

4.4.2 Topics

- Scripted pipelines vs. manual edits
- Versioning datasets and metadata
- Validation checks as code
- Summary reports for prepared datasets

4.5 3.5 Applied Example: Preparing a Mixed-Type Dataset

A brief walkthrough demonstrates how a raw survey dataset is transformed into an analysis-ready table with clear variable definitions, cleaned categories, and documented transformations. The focus is on traceability: every transformation is explained, and the resulting dataset is ready for association analysis and visual exploration.

Part II

Association Analysis

5 Unified Association Measures for Mixed-Type Variables

6 Extensions of Correlation—Nonlinear and Conditional

7 Network-Based Representations of Associations

Part III

Interactive Visual Analytics

8 Principles of Interactive Exploration

9 The AssociationExplorer Application

10 Communicating Findings Through Visualization

Part IV

Tree-Based Methods for Description

11 Regression Trees for Exploratory Segmentation

12 Classification Trees and Confusion Matrix Insights

13 Ensemble Methods as Descriptive Instruments

Part V

Interpretable Machine Learning

14 Interpretable ML—An Overview

15 Feature Importance and Variable Selection

16 Partial Dependence and Individual Conditional Expectation

17 Shapley Values and Additive Explanations

Part VI

AutoML for Exploration

18 AutoML as a Descriptive Tool

19 Automated Feature Engineering and Interaction Discovery

Part VII

Applied Case Studies

20 Case Study—Public Policy and Program Evaluation

21 Case Study—Public Health and Epidemiological Data

22 Case Study—Business Analytics and Customer Insights

23 Conclusion

References