# THE ESSENTIALS OF DATA PREPARATION

Patrick Boily[1,2,3]

**Abstract**

Once raw data has been collected and stored in a database or a dataset, the focus should shift to data cleaning and processing. This requires testing for **soundness** and fixing **errors**, designing and implementing strategies to deal with **missing values** and **outlying/influential observations**, as well as low-level **exploratory data analysis** and **visualisation** to determine what **data transformations** and **dimension reduction** approaches will be needed in the final analysis. In this report, we establish the essential elements of data cleaning and of data processing.

**Keywords**

Data cleaning, data preparation, data quality, missing values, anomalous observations, data transformations.

**Funding Acknowledgement**

Parts of this report were funded by Carleton University's Centre for Quantitative Analysis and Decision Support.

[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa
[2]Data Action Lab, Ottawa
[3]Idlewyld Analytics and Consulting Services, Wakefield, Canada
**Email**: pboily@uottawa.ca

## Contents



## 1. Introduction

**Data Validation**

**Martin K:** Data is messy, Alison.
**Alison M:** Even when it's been cleaned?
**Martin K:** Especially when it's been cleaned.

– P. Boily, J. Schellinck, *The Great Balancing Act*.

Data cleaning and data processing are essential aspects of quantitative analysis projects; analysts and consultants should be prepared to spend up to 80% of their time on data preparation, keeping in mind that:

- processing should **NEVER** be done on the original dataset – make copies along the way;
- **ALL** cleaning steps need to be documented;
- if **too much** of the data requires cleaning up, the data collection procedure might need to be **revisited**, and
- records should only be discarded as a **last resort**.

Another thing to keep in mind is that cleaning and processing may need to take place more than once depending on the type of data collection (one pass, batch, continuously), and that that it is essentially impossible to determine if all data issues have been found and fixed.[1]

[1]In this report, we are assuming that the datasets of interest contain only numerical and/or categorical observations. Additional steps must be taken when dealing with unstructured data, such as text or images.

## 2. General Principles

**Data Validation**

**Dilbert:** I didn't have any accurate numbers, so I just made up this one. Studies have shown that accurate numbers aren't any more useful that the ones you make up.
**Pointy-Haired Boss:** How many studies showed that?
**Dilbert:** [*beat*] Eighty-seven.

– Scott Adams, Dilbert, 8 May 2008

### 2.1 Approaches to Data Cleaning

There are two main **philosophical** approaches to data cleaning and validation:

- **methodical**, and
- **narrative**.

The **methodical** approach consists in running through a **check list** of potential issues and flagging those that apply to the data.

The **narrative** approach, on the other hand, consists in **exploring** the dataset while searching for unlikely or irregular patterns.

Which approach the consultant/analyst opts to follow depends on a number of factors, not least of which is the client's needs and views on the matter – it is important to discuss this point with the clients.

### 2.2 Pros and Cons

The methodical approach focuses on **syntax**; the checklist is typically **context-independent**, which means that it (or a subset) can be reused from one project to another, which makes data analysis pipelines **easy to implement** and **automate**. In the same vein, common errors are **easily identified**.

On the flip side, the check list may be quite extensive and the entire process may prove **time-consuming**.

The biggest disadvantage of this approach is that it makes it difficult to identify **new types of errors**.

The narrative approach focuses on **semantics**; even false starts may simultaneously produce **data understanding** prior to switching to a more mechanical approach.

It is easy, however, to miss important sources of errors and invalid observations when the datasets have a **large number of features**.

There is an additional downside: **domain expertise**, coupled with the narrative approach, may bias the process by neglecting "uninteresting" areas of the dataset.

### 2.3 Tools and Methods

A non-exhaustive list of common data issues can be found in the *Data Cleaning Bingo Card* (see Table 1); other possibilities obviously exist.

Other methods include

- **visualizations –** which may help easily identify observations that need to be further examined;
- **data summaries –** # of missing observations; 5-pt summary, mean, standard deviation, skew, kurtosis, for numerical variables; distribution tables for categorical variables;
- *n*-**way tables –** counts for joint distributions of categorical variables;
- **small multiples** – tables/visualizations indexed along categorical variables, and
- **preliminary data analyses** – which may provide "huh, that's odd..." realizations.

It is important to note that there is nothing wrong with running a number of analyses to flush out data issues, but remember to label your initial forays as **preliminary** analyses. From the client's perspective, repeated analyses may create a sense of unease and distrust, even if they form a crucial part of the analytical process (doing so will also facilitate invoicing, if that is part of your concern).

---

In our (admittedly biased and incomplete) experience,

- **computer scientists** and **programmers** tend to naturally favour the methodical approach, while
- **mathematicians** and **statisticians** tend to naturally favour the narrative approach,

although we have met plenty of individuals with unexpected backgrounds in both camps. This is not the place for identity politics: quantitative consultants and analysts need to be comfortable with **both** approaches.

As an example, the narrative approach is akin to working out a crossword puzzle with a pen and accepting to put down potentially erroneous answers once in a while to try to open up the grid (what artificial intelligence researchers call the "exploration" approach).

The mechanical approach, on the other hand, is similar to working out the puzzle with a pencil and a dictionary, only putting down answers when their correctness is guaranteed (the "exploitation" approach of artificial intelligence).

More puzzles get solved when using the first approach, but mistakes tend to be spectacular. Not as many puzzles get solved the second way, but the trade-off is that that it leads to fewer mistakes.

| | | | | |
|---|---|---|---|---|
| random missing values | outliers | values outside of expected range - numeric | factors incorrectly/iconsistently coded | date/time values in multiple formats |
| impossible numeric values | leading or trailing white space | badly formatted date/time values | non-random missing values | logical inconsistencies across fields |
| characters in numeric field | values outside of expected range - date/time | DCB! | inconsistent or no distinction between null, 0,not available, not applicable,missing | possible factors missing |
| multiple symbols used for missing values | ??? | fields incorrectly separated in row | blank fields | logical iconsistencies within field |
| entire blank rows | character encoding issues | duplicate value in unique field | non-factor values in factor | numeric values in character field |

**Table 1.** Data cleaning bingo card [personal communication, J.Shellinck].

## 3. Data Quality

**The Importance of Validation**

**Calvin's Dad:** OK Calvin. Let's check over your math homework.
**Calvin:** Let's not and say we did.
**Calvin's Dad:** Your teacher says you need to spend more time on it. Have a seat.
**Calvin:** More time?! I already spent 10 whole minutes on it! 10 minutes shot! Wasted! Down the drain!
**Calvin's Dad:** You've written here $8+4=7$. Now you know that's not right.
**Calvin:** So I was off a little bit. Sue me.
**Calvin's Dad:** You can't **add** things and come with **less** than you started with!
**Calvin:** I can do that! It's a free country! I've got my rights!

– Bill Watterson, *Calvin and Hobbes*, 15-09-1990.

The quality of the data has an important effect on the quality of the results: as the saying goes: "garbage in, garbage out."

Data is said to be **sound** when it has few issues with

- **validity** – are observations sensible, given data type, range, mandatory response, uniqueness, value, regular expressions, etc. (e.g. a value that is expected to be text value is a number, a value that is expected to be positive is negative, etc.)?;
- **completeness** – are there missing observations (more on this in a subsequent section)?;
- **accuracy and precision** – are there measurement and/or data entry errors (e.g. an individual has $-2$ children, etc., see Figure 1, linking accuracy to bias and precision to the standard error)?;
- **consistency** – are there conflicting observations (e.g. an individual has no children, but the age of one kid is recorded, etc.)?, and
- **uniformity** – are units used uniformly throughout (e.g. an individual is 6ft tall, whereas another one is 145cm tall)?

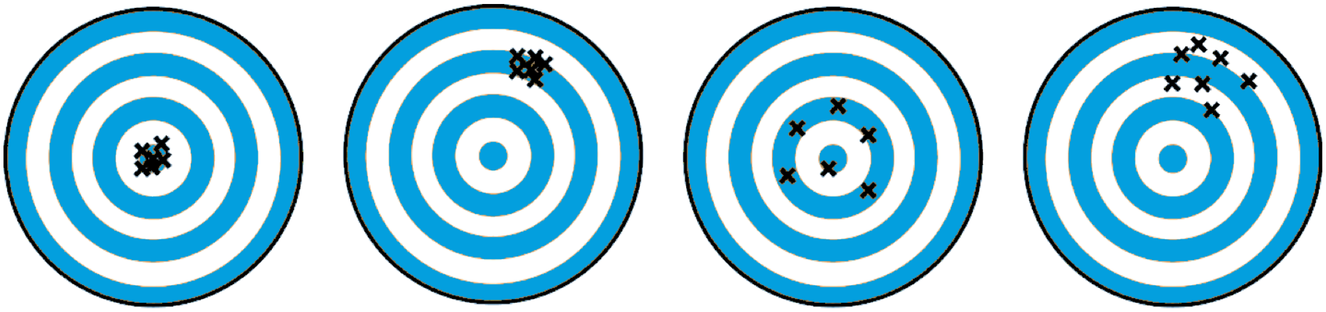


**Figure 1.** Accuracy as bias, precision as standard error [author unknown].

Finding an issue with data quality after the analyses are completed is a surefire way of losing the client's trust – check early and often!

### 3.1 Common Sources of Error

If the analysts have some control over the data collection and initial processing, regular data validation tests are easier to set-up.

When the analysts are dealing with **legacy**, **inherited**, or **combined** datasets, it can be difficult to recognise errors arising (among others) from

- missing data being given a code;
- 'NA'/'blank' entries being given a code;
- data entry errors;
- coding errors;
- measurement errors;
- duplicate entries, and
- heaping (see Figure 2 for an example).

### 3.2 Detecting Invalid Entries

Potentially invalid entries can be detected with the help of a number of methods:

- **univariate descriptive statistics** – $z$—score, count, range, mean, median, standard deviation, etc.;
- **multivariate descriptive statistics** – $n$—way tables and logic checks, and
- **data visualization** – scatterplot, histogram, joint histogram, etc.

We will not be discussing these methods, but we will point out that univariate tests do not always tell the whole story.

---

Consider, for instance, a medical dataset consisting of 38 patients' records, containing, among others, fields for the **sex** and the **pregnancy status** of the patients. A summary of the data of interest is afforded by the frequency counts (1-way tables) shown in Table 2. The analyst can quickly notice that some values are missing (in green) and that an entry has been miscoded as 99 (in yellow). Using only these univariate summaries, however, it is impossible to decide what to do with these invalid entries.

The 2-way frequency counts shed some light on the situation, and uncover other potential issues with the data.

One of the green entries is actually blank along the two variables; depending on the other information, this entry could be a candidate for **imputation** or outright **deletion** (more on these concepts in the next section).

Three other observations are missing a value along exactly one variable, but the information provided by the other variables may be complete enough to warrant imputation. Of course, if more information is available about the patients, the analyst may be able to determine why the values were missing in the first place, although privacy concerns at the collection stage might muddy the waters.

The mis-coded information on the pregnancy status (99, in yellow) is linked to a male client, and as such re-coding it as 'No' is likely to be a reasonable decision (although not necessarily the correct one).

A similar reasoning process should make the analyst question the validity of the entry shaded in red – the entry might very well be correct, but it is important to at least inquire about this data point, as the answer could lead to an eventual re-framing of the definitions and questions used at the collection stage.

---

In general, there is no universal or one-size-fits-all approach – a lot depends on the **nature of the data**. As always, domain expertise can help.

Remember that a failure to detect invalid entries is **not a guarantee** that there are in fact no invalid entries in the dataset. It is important not to oversell this step to the client. When only a small number of invalid entries are detected, the general recommendation is to treat these values as **missing**, which we discuss presently.

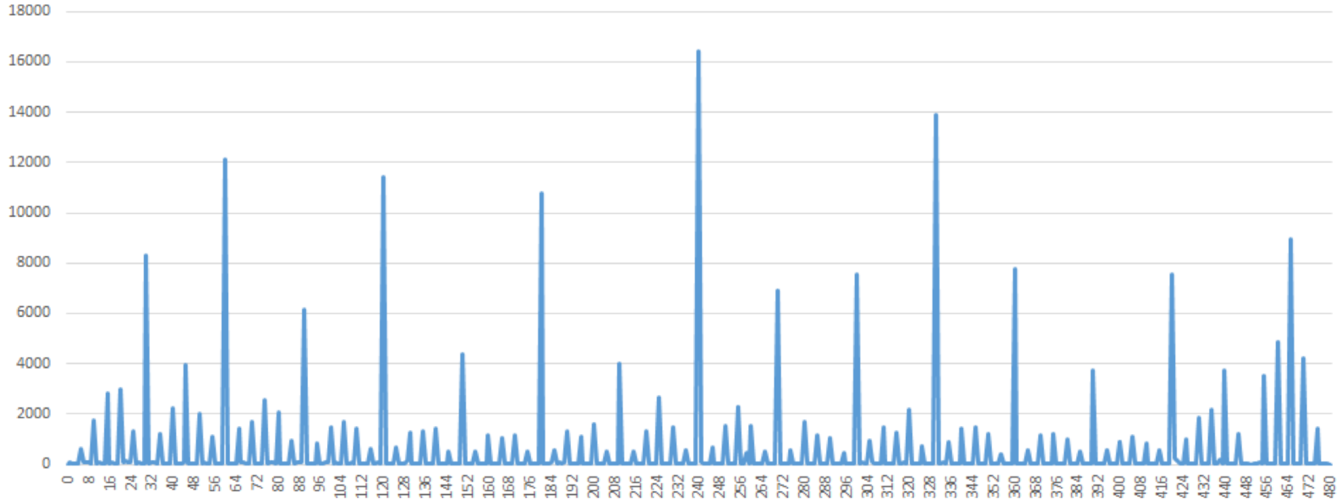


**Figure 2.** An illustration of heaping: self-reported time spent working in a day [personal file]. The entries for 7, 7.5, and 8 hours are omitted. Note the rounding off at various multiples of 5 minutes.

| | | |
|---|---|---|
| Sex | Male | 19 |
| | Female | 17 |
| | (blank) | 2 |
| Total | | 38 |

| | | |
|---|---|---|
| Pregnant | Yes | 7 |
| | No | 27 |
| | 99 | 1 |
| | (blank) | 3 |
| Total | | 38 |

| | | Pregnant | | | | Total |
|---|---|---|---|---|---|---|
| | | Yes | No | 99 | (blank) | |
| Sex | Male | 1 | 17 | 1 | 0 | 19 |
| | Female | 6 | 9 | 0 | 2 | 17 |
| | (blank) | 0 | 1 | 0 | 1 | 2 |
| Total | | 7 | 27 | 1 | 3 | 38 |

**Table 2.** Summary data for an (artificial) medical dataset: 1—way tables (left), 2—way table (right).

## 4. Missing Values

**Easier Said Than Done**

Obviously, the best way to treat missing data is not to have any.

– T. Orchard, M. Woodbury, *A Missing Information Principle: Theory and Applications*, 1972

Why does it matter that some values may be **missing**? Well, they can potentially introduce bias into the analysis, which is rarely (if at all) a good thing, but, more pragmatically, they may interfere with the functioning of most analytical methods, which can not easily accommodate missing observations without breaking down.[2]

Consequently, when faced with missing observations, analysts have two options: they can either **discard** the missing observation (which is not typically recommended, unless the data is missing completely randomly), or they can **create a replacement value** for the missing observation (the **imputation** strategy has drawbacks since we can never be certain that the replacement value is the true value, but is often the best available option; information in this section is taken partly from [2–5]).

[2]For instance, The canonical equation $\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}^\top\mathbf{Y}$ of linear regression cannot be solved as $\mathbf{X}^\top\mathbf{X}$ is not defined if some observations are missing.

Blank fields come in 4 flavours:

- **nonresponse** – an observation was expected but none was entered;
- **data entry issues** – an observation was recorded but was not entered in the dataset;
- **invalid entries** – an observation was recorded but was considered invalid and has been removed, and
- **expected blanks** – a field has been left blank, but not unexpectedly so.

Too many missing values of the first three types can be indicative of **issues with the data collection process**, while too many missing values of the fourth type can be indicative of **poor questionnaire design** (see [29] for a brief discussion on these topics).

Either way, missing values cannot simply be **ignored**: either the

- corresponding record is removed from the dataset (not recommended without justification, as doing so may cause a loss of auxiliary information and may bias the analysis results), or
- missing value must be **imputed** (that is to say, a reasonable replacement value must be found).

### 4.1 Missing Value Mechanisms

The relevance of an imputation method is dependent on the underlying **missing value mechanism**; values may be

- **missing completely at random** (MCAR) – the item absence is independent of its value or of the unit's auxiliary variables (e.g., an electrical surge randomly deletes an observation in the dataset);
- **missing at random** (MAR) – the item absence is not completely random, and could, in theory, be accounted by the unit's complete auxiliary information, if available (e.g., if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves), and
- **not missing at random** (NMAR) – the reason for nonresponse is related to the item value itself (e.g., if illicit drug users are less likely to admit to drug use than teetotalers).

The consultant's main challenge in that regard is that the missing mechanism cannot typically be determined with any degree of certainty.

### 4.2 Imputation Methods

There are numerous statistical **imputation** methods. They each have their strengths and weaknesses; consequently, consultants and analysts should take care to select a method which is appropriate for the situation at hand.[3]

- In **list-wise deletion**, all units with at least one missing value are removed from the dataset. This straightforward imputation strategy assumes MCAR, but it can introduce bias if MCAR does not hold, and it leads to a reduction in the sample size and an increase in standard errors.
- In **mean** or **most frequent imputation**, the missing values are substituted by the average or most frequent value in the unit's subpopulation group (stratum). This commonly-used approach also assumes MCAR, but it can creates distortions in the underlying distributions (such as a spike at the mean) and create spurious relationships among variables.
- In **regression** or **correlation imputation**, the missing values are substituted using a regression on the other variables. This model assumes MAR and trains the regression on units with complete information, in order to take full advantage of the auxiliary information when it is available. However, it artificially reduces data variability and produces over-estimates of correlations.
- In **stochastic regression imputation**, the regression estimates are augmented with random error terms added. Just as in regression estimation, the model assumes MAR; an added benefit is that it tends to produce estimates that "look" more realistic than regression imputation, but it comes with an increased risk of type I error (false positives) due to small standard errors.
- **Last observation carried forward** (LOCF) and its cousin **next observation carried backward** (NOCB) are useful for longitudinal data; a missing value can simply be substituted by the previous or next value. LOCF and NOCB can be used when the values do not vary greatly from one observation to the next, and when values are MCAR. Their main drawback is that they may be too "generous" for studies that are trying to determine the effect of a treatment over time, say.
- Finally, in ***k*-nearest-neighbour imputation**, a missing entry in a MAR scenario is substituted by the average (or median, or mode) value from the subgroup of the $k$ most similar complete respondents. This requires a notion of **similarity** between units (which is not always easy to define reasonably). The choice of $k$ is somewhat arbitrary and can affect the imputation, potentially distorting the data structure when it is too large.

---

What does imputation look like in practice?

Consider the following scenario (which is, somewhat embarrassingly, based on a true story): after marking the final exams of the 100 students who did not drop her course in *Advanced Retroencabulation* at State University, Dr. Helga Vanderwhede plots the final exam grades ($y$) against the mid-term exam grades ($x$) as in Figure 3.

She takes a quick look at the data and sees that final exam grades are **correlated** with mid-term exam grades: students who perform well on the mid-term tend to perform well on the final, and students who perform poorly on the mid-term tend to perform poorly on the final, as is usually the case.

She also sees that there is a fair amount of variability in the data: the noise is not very tight around the (eye-balled) line of best fit.

Furthermore, she realizes that the final exam was harder than the students expected[4] – she suspects that they simply did not prepare for the exam seriously (and not that she made the exam too difficult, no matter what her ratings on `RateMyProfessor.com` suggest), as most of them could not match their mid-term exam performance.

As Dr. Vanderwhede comes to term with her disappointment, she takes a deeper look at the numbers, at some point sorting the dataset according to the mid-term exam grades. It looks like good old Mary Sue performed better on the final than on the mid-term (where her performance was already superlative), scoring the only perfect score.

[3] Imputation methods work best under MCAR or MAR, but keep in mind that they all tend to produce **biased estimates**.

[4] The slope of the line of best fit is smaller than 1.

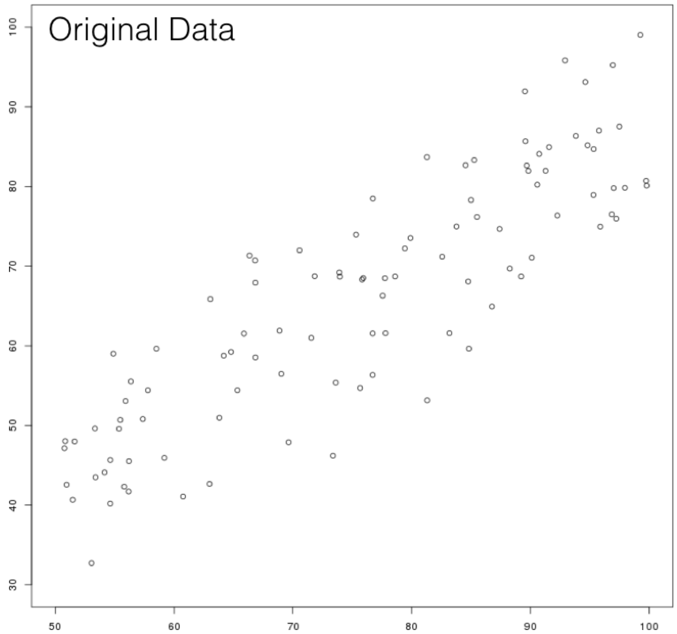


**Figure 3.** Dr. Vanderwhede's original *Advanced Retroencabulation* dataset; mid-term grades ($x$—axis), final exam grades ($y$—axis).

What a great student Mary Sue is! And such a fantastic person – in spite of her superior intellect, she is adored by all of her classmates, thanks to her sunny disposition and willingness to help at all times. If only all students were like Mary Sue...

She continues to toy with the spreadsheet until the phone rings. After a long and exhausting conversation with Dean Bitterman about teaching loads and State University's reputation, Dr. Vanderwhede returns to the spreadsheet and notices in horror that she has accidentally deleted the final exam grades of all students with a mid-term grade greater than 92.

What is she to do?

A technically-savvy consultant would advise her to either undo her changes or to close the file without saving the changes,[5] but let's assume for the time being that, in full panic mode, the only solution that comes to her mind is to impute the missing values.

She knows that the missing final grades are MAR (and not MCAR since she remembers sorting the data along the $x$ values); she produces the imputations shown in Figure 4. She remembers what the data looked like originally, and concludes that the best imputation method is the stochastic regression model.

But this only applies to this specific example. In general, that might not be the case, however, due to various *No Free Lunch* results.[6]

[5]Or to re-enter the final grades by comparing with the physical papers

[6]"There ain't no such thing as a free lunch" – there is no guarantee that a method that works best for a dataset even works well for another.

The principal take-away from this example is that various imputation strategies lead to different outcomes, and perhaps more importantly, that even though the imputed data might "look" like the true data, we have no way to measure its **departure from reality –** any single imputed value is likely to be completely off.

Mathematically, this might not be problematic, as the average departure is likely to be relatively small, but in a business or personal context, this might create gigantic problems – how is Mary Sue likely to feel about Dr. Vanderwhede's solution in the previous example? How is Dean Bitterman likely to react, if he finds out about the imputation scenario from irrate students?

Even though such questions are not quantitative in nature, their answer will impact any actionable solution.

### 4.3 Multiple Imputation

Another drawback of imputation is that it tends to increase the noise in the data, because the imputed data is treated as the *actual* data.

In **multiple imputation**, the impact of that noise can be reduced by consolidating the analysis outcome from multiple imputed datasets. Once an imputation strategy has been selected on the basis of the (assumed) missing value mechanism,

1. the imputation process is repeated $m$ times to produce $m$ versions of the dataset;
2. each of these datasets is analyzed, yielding $m$ outcomes, and
3. the $m$ outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known.

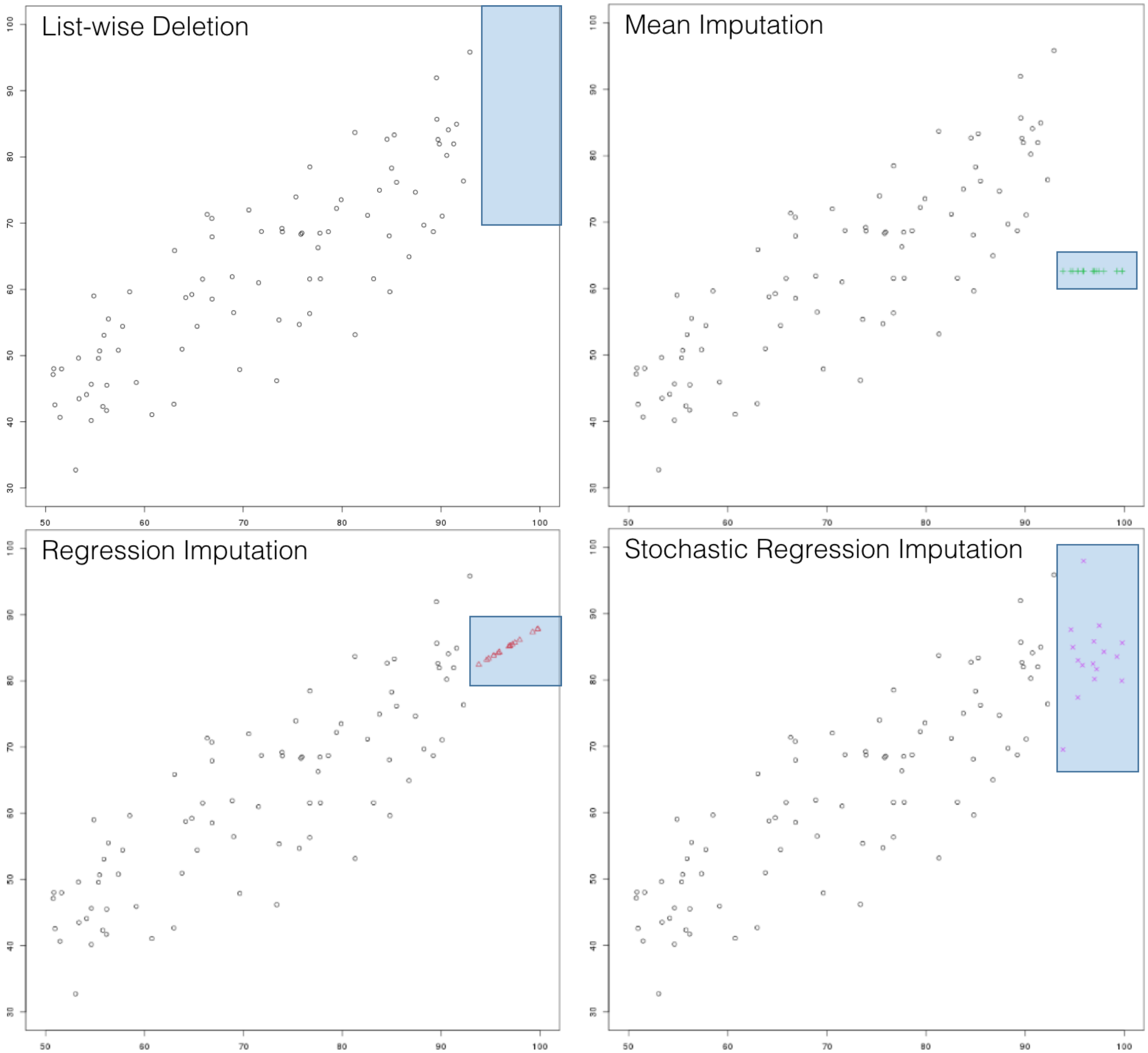


**Figure 4.** Imputed values for Dr. Vanderwhede's dataset.

On the plus side, multiple imputation is **easy to implement**, **flexible**, as it can be used in a most situations (MCAR, MAR, even NMAR in certain cases), and it accounts for **uncertainty** in the imputed values.

However, *m* may need to be quite **large** when the values are missing in large quantities from many of the dataset's features, which can substantially slow down the analyses.

There may also be additional technical challenges when the output of the analyses is not a single value but some more complicated object.

A generalization of multiple imputation was used by Transport Canada to predict the Blood Alcohol Level (BAC) content level in fatal traffic collisions that involved pedestrians [31].

## 5. Anomalous Observations

(The contents of this section are taken from [30]).

**The Good Doctor's Take**

The most exciting phrase to hear [...], the one that heralds the most discoveries, is not "Eureka!" but "That's funny...".

– Isaac Asimov (attributed)

**Outlying observations** are data points which are **atypical** in comparison to the unit's remaining features (*within-unit*), or in comparison to the measurements for other units (*between-units*), or as part of a collective subset of observations.

Outliers are thus observations which are **dissimilar to other cases** or which contradict **known dependencies** or rules.[7]

Observations could be anomalous in one context, but not in another. Consider, for instance, an adult male who is 6-foot tall. Such a man would fall in the 86th percentile among Canadian males [26], which, while on the tall side, is not unusual; in Bolivia, however, the same man would land in the 99.9th percentile [26], which would mark him as extremely tall and quite dissimilar to the rest of the population.[8]

A common mistake that analysts make when dealing with outlying observations is to remove them from the dataset without carefully studying whether they are **influential data points**, that is, observations whose absence leads to **markedly different** analysis results.

When influential observations are identified, remedial measures (such as data transformation strategies) may need to be applied to minimize any undue effect. Outliers may be influential, and influential data points may be outliers, but the conditions are neither necessary nor sufficient.

### 5.1 Anomaly Detection

By definition, anomalies are **infrequent** and typically surrounded by **uncertainty** due to their relatively low numbers, which makes it difficult to differentiate them from banal **noise** or **data collection errors**.

Furthermore, the boundary between normal and deviant observations is usually **fuzzy**; with the advent of e-shops, for instance, a purchase which is recorded at 3AM local time does not necessarily raise a red flag anymore.

When anomalies are actually associated to **malicious activities**, they are more than often **disguised** in order to blend in with normal observations, which obviously complicates the detection process.

Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used. Methods that employ graphical aids (such as box-plots, scatterplots, scatterplot matrices, and 2D tours) to identify outliers are particularly easy to implement, but a low-dimensional setting is usually required for ease of interpretability.

Analytical methods also exist (using Cooke's or Mahalanobis' distances, say), but in general some additional level of analysis must be performed, especially when trying to identify influential points (*cf.* **leverage**).

With small datasets, anomaly detection can be conducted on a case-by-case basis, but with large datasets, the temptation to use **automated detection/removal** is strong – care must be exercised before the analyst decides to go down that route.[9]

In the early stages of anomaly detection, **simple data analyses** (such as descriptive statistics, 1- and 2-way tables, and traditional visualisations) may be performed to help identify anomalous observations, or to obtain insights about the data, which could eventually lead to modifications of the analysis plan.

### 5.2 Outlier Tests

How are outliers *actually* detected? Most methods come in one of two flavours: **supervised** and **unsupervised** (we will discuss those in detail in later sections).

Supervised methods use a historical record of **labeled** (that is to say, previously identified) anomalous observations to build a **predictive classification or regression model** which estimates the probability that a unit is anomalous; domain expertise is required to tag the data. Since anomalies are typically **infrequent**, these models often also have to accommodate the **rare occurrence problem**.[10]

Unsupervised methods, on the other hand, use no previously labeled information or data, and try to determine if an observation is an outlying one solely by comparing its behaviour to that of the other observations.

The following traditional methods and tests of outlier detection fall into this category:[11]

- Perhaps the most commonly-used test is **Tukey's boxplot test**; for normally distributed data, regular observations typically lie between the **inner fences**

  $$Q_1 - 1.5(Q_3 - Q_1) \quad \text{and} \quad Q_3 + 1.5(Q_3 - Q_1).$$

  **Suspected outliers** lie between the inner fences and their respective **outer fences**

  $$Q_1 - 3(Q_3 - Q_1) \quad \text{and} \quad Q_3 + 3(Q_3 - Q_1).$$

  Points beyond the outer fences are identified as **outliers** ($Q_1$ and $Q_3$ represent the data's 1st and 3rd quartile, respectively; see Figure 5).
- The **Grubbs test** is another univariate test, which takes into consideration the number of observations in the dataset. Let $x_i$ be the value of feature $X$ for

[7] Outlying observations may be anomalous along any of the individual variables, or in combination.

[8] Anomaly detection points towards interesting questions for analysts and subject matter experts: in this case, why is there such a large discrepancy in the two populations?

[9] This stems partly from the fact that once the "anomalous" observations have been removed from the dataset, previously "regular" observations can become anomalous in turn in the smaller dataset; it is not clear when that runaway train will stop.

[10] Supervised models are built to minimize a cost function; in default settings, it is often the case that the mis-classification cost is assumed to be symmetrical, which can lead to technically correct but useless solutions. For instance, the vast majority (99.999+%) of air passengers emphatically do not bring weapons with them on flights; a model that predicts that no passenger is attempting to smuggle a weapon on board a flight would be 99.999+% accurate, but it would miss the point completely.

[11] Note that **normality** of the underlying data is an assumption for most tests; how robust these tests are against departures from this assumption depends on the situation.

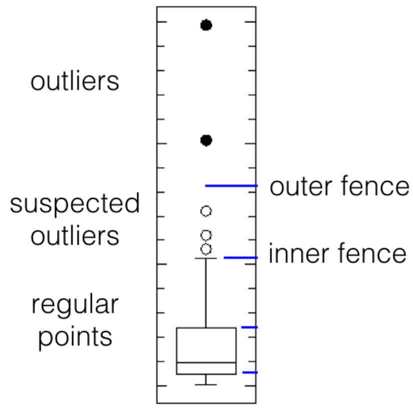


**Figure 5.** Tukey's boxplot test; suspected outliers are marked by white disks, outliers by black disks.

the $i^{\text{th}}$ unit, $1 \leq i \leq N$, let $(\overline{x}, s_x)$ be the mean and standard deviation of feature $X$, let $\alpha$ be the desired significance level, and let $T(\alpha, N)$ be the critical value of the Student $t$-distribution at significance $\alpha/2N$. Then, the $i^{\text{th}}$ unit is an **outlier along feature** $X$ if

$$|x_i - \overline{x}| \geq \frac{s_x(N-1)}{\sqrt{N}} \sqrt{\frac{T^2(\alpha, N)}{N - 2 + T^2(\alpha, N)}}.$$

- Other common tests include:
  - the **Mahalanobis distance**, which is linked to the leverage of an observation (a measure of influence), can also be used to find multi-dimensional outliers, when all relationships are linear (or nearly linear);
  - the **Tietjen-Moore** test, which is used to find a specific number of outliers;
  - the **generalized extreme studentized deviate** test, if the number of outliers is unknown;
  - the **chi-square** test, when outliers affect the goodness-of-fit, as well as
  - DBSCAN and other clustering-based outlier detection methods.

### 5.3 Visual Outlier Detection

The following three (simple) examples illustrate the principles underlying visual outlier and anomaly detection.

**Example 1.** On a specific day, the height of several plants are measured. The records also show each plant's age (the number of weeks since the seed has been planted).

Histograms of the data are shown in Figure 6 (age on the left, height on the middle).

Very little can be said about the data at that stage: the age of the plants (controlled by the nursery staff) seems to be somewhat haphazard, as does the response variable (height). A scatter plot of the data (rightmost chart in Figure 6), however, reveals that growth is strongly correlated with age during the early period of a plant's life for the observations in the dataset; points clutter around a linear trend. One point (in yellow) is easily identified as an **outlier**. There are (at least) two possibilities: either that measurement was botched or mis-entered in the database (representing an invalid entry), or that one specimen has experienced unusual growth (outlier). Either way, the analyst has to investigate further.

**Example 2.** A government department has 11 service points in a jurisdiction. Service statistics are recorded: the monthly average arrival rates per teller and monthly average service rates per teller for each service point are available.

A scatter plot of the service rate per teller ($y$ axis) against the arrival rate per teller ($x$ axis), with linear regression trend, is shown in the leftmost chart in Figure 7. The trend is seen to inch upwards with increasing $x$ values.

A similar chart, but with the left-most point removed from consideration, is shown in the middle chart of Figure 7. The trend still slopes upward, but the fit is significantly improved, suggesting that the removed observation is unduly **influential** (or anomalous) – a better understanding of the relationship between arrivals and services is afforded if it is set aside.

Any attempt to fit that data point into the model must take this information into consideration. Note, however, that influential observations depend on the analysis that is ultimately being conducted – a point may be influential for one analysis, but not for another.

**Example 3.** Measurements of the length of the appendage of a certain species of insect have been made on 71 individuals. Descriptive statistics have been computed; the results are shown in Figure 8.

Analysts who are well-versed in statistical methods might recognize the tell-tale signs that the distribution of appendage lengths is likely to be asymmetrical (since the skewness is non-negligible) and to have a "fat" tail (due to the kurtosis being commensurate with the mean and the standard deviation, the range being so much larger than the interquartile range, and the maximum value being so much larger than the third quartile).

The mode, minimum, and first quartile values belong to individuals without appendages, so there appears to be at least two sub-groups in the population (perhaps split along the lines of juveniles/adults, or males/females). The maximum value has already been seen to be quite large compared to the rest of the observations, which at first suggests that it might belong to an **outlier**.

The histogram of the measurements, however, shows that there are 3 individuals with very long appendages (see right-most chart in Figure 8): it now becomes plausible for these anomalous entries to belong to individuals from a different species altogether who were **erroneously added** to the dataset. This does not, of course, constitute a proof of such an error, but it raises the possibility, which is often the best that an analyst can do in the absence of subject matter expertise.

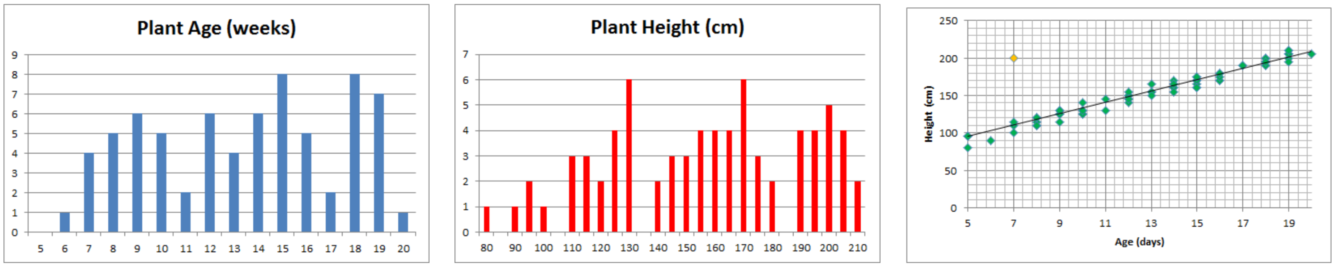


**Figure 6.** Summary visualisations for an (artificial) plant dataset: age distribution (left), height distribution (middle), height vs. age, with linear trend (right).

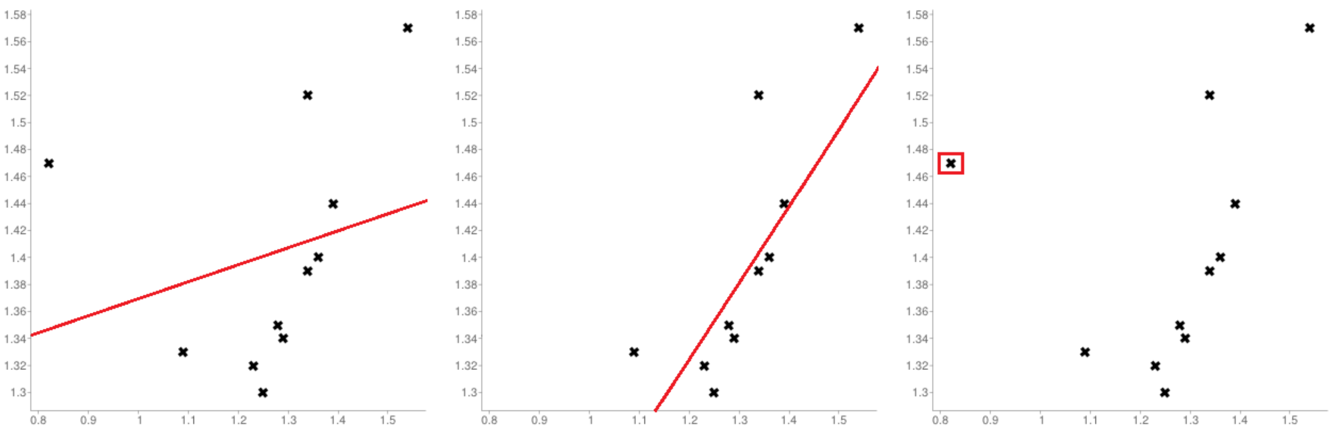

**Figure 7.** Visualisations for an (artificial) service point dataset: trend for 11 service points (left), trend for 10 service points (middle), influential observations (right).

| *Appendage length (mm)* | |
|---|---|
| Mean | 10.35 |
| Standard Deviation | 16.98 |
| Kurtosis | 16.78 |
| Skewness | 4.07 |
| Minimum | 0 |
| First Quartile | 0 |
| Median | 8.77 |
| Third Quartile | 10.58 |
| Maximum | 88 |
| Range | 88 |
| Interquartile Range | 10.58 |
| Mode | 0 |
| Count | 71 |

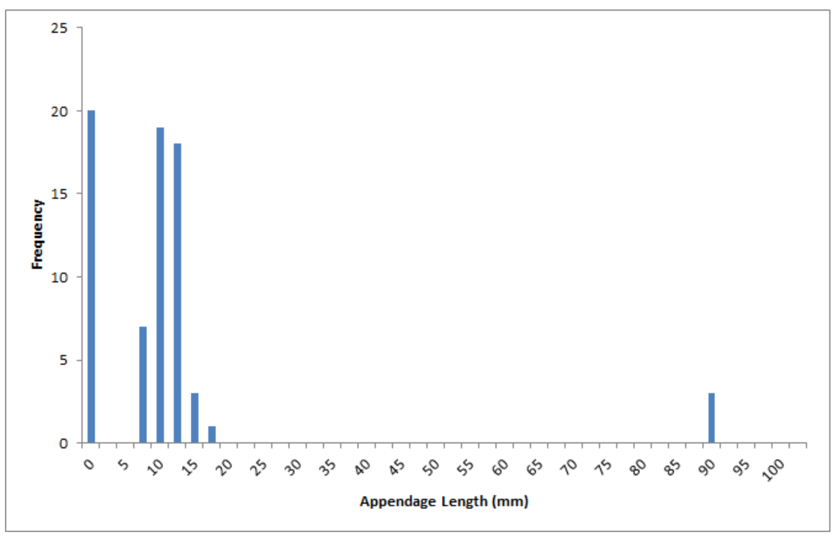


**Figure 8.** Summary and visualisation for an (artificial) appendage length dataset: descriptive statistics (left), appendage length distribution (right).

## 6. Data Transformation

**It's Also True of Data**

History is the transformation of tumultuous conquerors into silent footnotes.

– Paul Eldridge, American educator

This **crucial** last step is often neglected or omitted altogether. Various transformation methods are available, depending on the analysts' needs and data types, including:

- **standardization** and **unit conversion**, which put the dataset's variables on an equal footing – a requirement for basic comparison tasks and more complicated problems of clustering and similarity matching;

- **normalization**, which attempts to force a variable into a normal distribution – an assumption which must be met in order to use a number of traditional analysis methods, such as ANOVA or regression analysis, and
- **smoothing methods**, which help remove unwanted noise from the data, but at a price – perhaps removing natural variance in the data.

Another type of data transformation is pre-occupied with the concept of **dimensionality reduction**. There are many advantages to working with low-dimensional data [32]:

- **visualization methods** of all kinds are available to extract and present insights out of such data;
- high-dimensional datasets are subject to the so-called **curse of dimensionality**, which asserts (among other things) that multi-dimensional spaces are vast, and when the number of features in a model increases, the number of observations required to maintain predictive power also increases, but at a **substantially higher rate** (see Figure 9);
- another consequence of the curse is that in high-dimension sets, all observations are roughly **dissimilar** to one another – observations tend to be nearer the dataset's boundaries than they are to one another.

Dimension reduction techniques such as **principal component analysis**, **independent component analysis**, and **factor analysis** (for numerical data), or **multiple correspondence analysis** (for categorical data) project multi-dimensional datasets onto low-dimensional but high information spaces (the so-called **Manifold Hypothesis**). Some information is necessarily lost in the process, but in many instances the drain can be kept under control and the gains made by working with smaller datasets can offset the losses of completeness [32].

### 6.1 Common Transformations

Models often require that certain data assumptions be met. For instance, ordinary least square regression assumes:

- that the response variable is a **linear combination** of the predictors;
- **constant** error variance;
- **uncorrelated residuals**, which may or may not be statistically independent;
- etc.

In reality, it is rare that raw data meets all these requirements, but that does not necessarily mean that we need to abandon the model – an **invertible** sequence of data transformations may produce a derived data set which *does* meet the requirements, allowing the consultant to draw conclusions about the original data.

In the regression context, invertibility is guaranteed by **monotonic** transformations: identity, logarithmic, square root, inverse (all members of the power transformations), exponential, etc. (illustrations are provided in Figure 10).

There are rules of thumb and best practices to transform data, but analysts and consultants should not discount the importance of exploring the data visually before making a choice.

Transformations on the predictors $X$ may be used to achieve the **linearity assumption**, but they usually come at a price – correlations are not preserved by such transformations, for instance. Transformations on the target $Y$ can help with **non-normality** of residuals and **non-constant variance** of error terms.

Note that transformations can be applied **both** to the target variable or the predictors: as an example, if the linear relationship between two variables $X$ and $Y$ is expressed as $Y = a + bX$, then a unit increase in $X$ is associated with an average of $b$ units in $Y$.

But a better fit might be afforded by either of

$$\log Y = a+bX, \quad Y = a+b\log X, \quad \text{or} \quad \log Y = a+b\log X,$$

for which:

- a unit increase in $X$ is associated with an average $b\%$ increase in $Y$;
- a 1% increase in $X$ is associated with an average $0.01b$ unit increase in $Y$, and
- a 1% increase in $X$ is associated with a $b\%$ increase in $Y$, respectively.

### 6.2 Box-Cox Transformation

The choice of transformation is often as much of an art as it is a science. There is a common framework, however, that provides the optimal transformation, in a sense. Consider the task of predicting the target $Y$ with the help of the predictors $X_j$, $j = 1, \ldots, p$. The usual model takes the form

$$y_i = \sum_{j=1}^{p} \beta_j X_{x,i} + \varepsilon_i, \quad i = 1, \ldots, n.$$

If the residuals are skewed, or their variance is not constant, or the trend itself does not appear to be linear, a power transformation might be preferable, but if so, which one?

The **Box-Cox transformation** $y_i \mapsto y_i'(\lambda)$, $y_i > 0$ is defined by

$$y_i'(\lambda) = \begin{cases} (y_1 \ldots y_n)^{1/n} \ln y_i, \text{if } \lambda = 0 \\ \frac{y_i^\lambda - 1}{\lambda}(y_1 \ldots y_n)^{\frac{1-\lambda}{n}}, \text{if } \lambda \neq 0 \end{cases};$$

variants allow for the inclusion of a shit parameter $\alpha > 0$, which extends the transformation to $y_i > -\alpha$. The **suggested** choice of $\lambda$ is the value that maximises the log-likelihood

$$\mathscr{L} = -\frac{n}{2} \log\left( \frac{2\pi\hat{\sigma}^2}{(y_1 \ldots y_n)^{2(\lambda-1)/n}} + 1 \right).$$

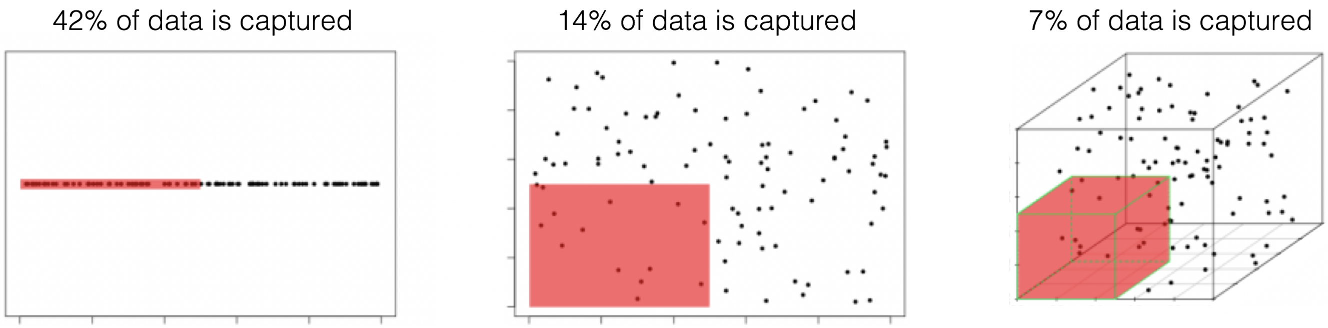


**Figure 9.** Illustration of the curse of dimensionality; $N = 100$ observations are uniformly distributed on the unit hypercube $[0,1]^d$, $d = 1,2,3$. The red regions represent the smaller hypercubes $[0,0.5]^d$, $d = 1,2,3$. The percentage of captured datapoints is seen to decrease with an increase in $d$ [28].

There might be theoretical rationales which favour a particular choice of $\lambda$ – these are not to be ignored. It is also important to produce a residual analysis, as the best Box-Cox choice does not necessarily meet all the least squares assumptions.

Finally, it is important to remember that the resulting parameters have the least squares property **only with respect to the transformed data points**.

### 6.3 Scaling

Numeric variables may have different scales (weights and heights, for instance). Since the variance of a large-range variable is typically greater than that of a small-range variable, leaving the data **unscaled** may introduce biases, especially when using unsupervised methods.

It could also be the case that it is the relative positions (or rankings) which is of importance, in which case it could become important to look at relative distances between levels:

- **standardisation** creates a variable with mean 0 and standard deviations 1:
$$Y_i = \frac{X_i - \overline{X}}{s_X};$$
- **normalization** creates a variable in the range [0,1]:
$$Y_i = \frac{X_i - \min\{X_k\}}{\max\{X_k\} - \min\{X_k\}}.$$

These are not the only options. Different schemes can lead to different outputs.

### 6.4 Discretizing

To reduce computational complexity, a numeric variable may need to be replaced with an **ordinal** variable (*height* values could be replaced by the qualitative "*short*", "*average*", and "*tall*", for instance.

Of course, what these terms represent depend on the context: Canadian short and Bolivian tall may be fairly commensurate, to revisit the example at the start of the preceding section.

It is far from obvious how to determine the bins' limits – **domain expertise** can help, but it could introduce unconscious bias to the analyses. In the absence of such expertise, limits can be set so that either the bins each:

- contain the same **number of observations**;
- have the same **width**, or
- the performance of some modeling tool is maximised.

Again, various choices may lead to different outputs.

### 6.5 Creating Variables

Finally, it is possible that new variables may need to be introduced (in contrast with dimensionality reduction). These new variables may arise

- as **functional relationships** of some subset of available features (introducing powers of a feature, or principal components, say);
- because modeling tool may require **independence of observations** or **independence of features** (in order to remove multicolinearity, for instance), or
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis).

There is no limit to the number of new variables that can be added to a dataset – but consultants should strive for **relevant additions**.

### References

[1] Chapman, A. [2005], Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data, Report for the Global Biodiversity Information Facility, Copenhagen.

[2] van Buuren, S. [2012], Flexible Imputation of Missing Data, CRC Press, Boca Raton.

[3] Hagiwara, S. [2012], Nonresponse Error in Survey Sampling - Comparison of Different Imputation Methods, Honours Thesis, Carleton University, Ottawa.

[4] Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. [2001], A Multivariate Technique for

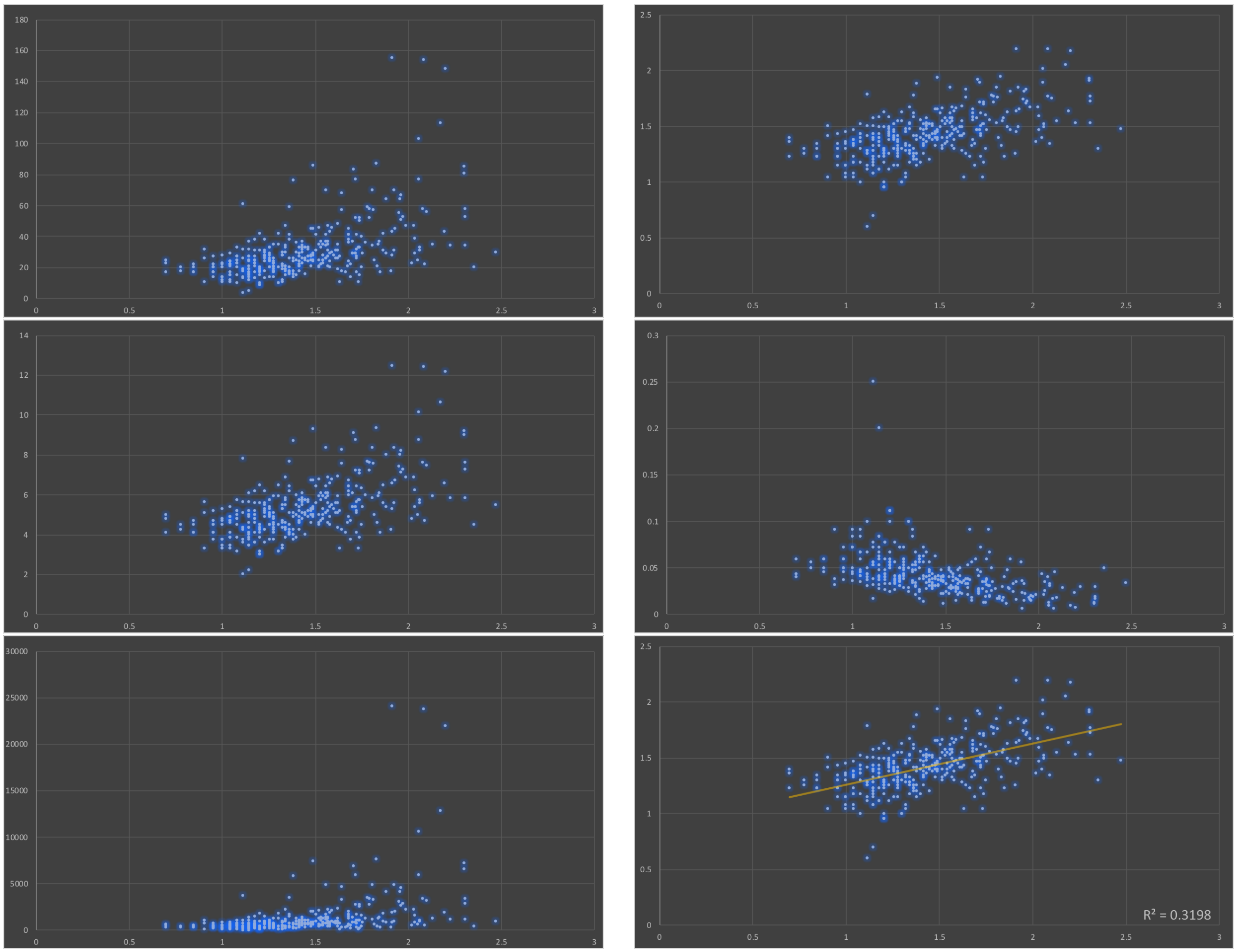


**Figure 10.** Examples of data transformations, for a subset of the BUPA liver dataset [27]. From left to right, top to bottom: original data, $Y' = \log Y$, $Y' = \sqrt{Y}$, $Y' = \frac{1}{Y}$, $Y' = Y^2$, and Box-Cox best choice ($\approx$ log).

Multiply Imputing Missing Values Using a Sequence of Regression Models, Survey Methodology, v.27, n.1, pp.85-95, Statistics Canada, Catalogue no. 12-001.

[5] Rubin, D.B. [1987], Multiple Imputation for Nonresponse in Surveys, Wiley, New York.

[6] Kutner, M., Nachtsheim, C., Neter, J. and Li, W. [2004], Applied Linear Statistical Models, 5th ed., McGraw-Hill/Irwin, New York.

[7] Green, S. and Salkind, N. [2011], Using SPSS for Windows and Macintosh - Analyzing and Understanding Data, 6th ed., Prentice Hall, Upper Saddle River.

[8] Wikipedia entry for Data Cleansing

[9] Wikipedia entry for Imputation

[10] Wikipedia entry for Outliers

[11] Torgo, L. [2017], Data Mining with R (2nd edition), CRC Press.

[12] McCallum, Q.E. [2013], Bad Data Handbook, O'Reilly.

[13] Kazil, J., Jarmul, K. [2016], Data Wrangling with Python, O'Reilly

[14] de Jonge, E., van der Loo, M. [2013], An Introduction to Data Cleaning with R, Statistics Netherlands.

[15] Pyle, D. [1999], Data Preparation for Data Mining, Morgan Kaufmann Publishers.

[16] Weiss, S.M., Indurkhya, I. [1999], Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers.

[17] Buttrey, S.E. [2017], A Data Scientist's Guide to Acquiring, Cleaning, and Managing Data in R, Wiley.

[18] Aggarwal, C.C. [2013], Outlier Analysis, Springer.

[19] Chandola, V., Banerjee, A., Kumar, V. [2007], Outlier detection: a survey, Technical Report TR 07-017, Department of Computer Science and Engineering, University of Minnesota.

[20] Hodge, V., Austin, J. [2004], A survey of outlier detection methodologies, Artif.Intell.Rev., 22(2):85-126.

[21] Feng, L., Nowak, G., Welsh, A.H., O'Neill, T. [2014], imputeR: a general imputation framework in R.

[22] Steiger, J.H., Transformations to Linearity, lecture notes.

[23] Wood, F., Remedial Measures Wrap-Up and Transformations, lecture notes.

[24] Dougherty, J., Kohavi, R., Sahami, M. [1995], Supervised and unsupervised discretization of continuous features, in Machine Learning: Proceedings of the Twelfth International Conference, Prieditis, A., Russell, S. (eds), Morgan Kaufmann Publishers.

[25] Orchard, T., Woodbury, M. [1972], A Missing Information Principle: Theory and Applications, Berkeley Symposium on Mathematical Statistics and Probability, University of California Press.

[26] Height Percentile Calculator, by Age and Country, https://tall.life/height-percentile-calculator-age-country/

[27] Dua, D., Karra Taniskidou, E. [2017], Liver Disorders dataset, UCI Machine Learning Repository.

[28] https://simplystatistics.org/2014/10/24/an-interactive-visualization-to-teach-about-the-curse-of-dimensionality/

[29] Boily, P. [2020], *Principles of Data Collection*, Data Science Report Series, data-action-lab.com.

[30] Cissokho, Y., Fadel, S., Millson, R., Pourhasan, R., Boily, P. [2020], *Anomaly Detection and Outlier Analysis*, Data Science Report Series, data-action-lab.com.

[31] Boily, P. [2007], *An Imputation Algorithm of Blood Alcohol Content Levels for Drivers and Pedestrians in Fatal Collisions*, Transport Canada.

[32] Leduc, O., Macfie, A., Maheshwari, A., Pelletier, M., Boily, P. [2020], *Feature Selection and Dimension Reduction*, Data Science Report Series, data-action-lab.com.