



CANADIAN  
FOREIGN  
SERVICE  
INSTITUTE

L'INSTITUT  
CANADIEN  
DU SERVICE  
EXTÉRIEUR



## Introduction to Data Analysis

# DATA PROCESSING

Patrick Boily

Data Action Lab | uOttawa | Idlewyld Analytics

[pboily@uottawa.ca](mailto:pboily@uottawa.ca)

# DATA CLEANING

DATA PROCESSING

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

T. Orchard, M. Woodbury

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

I. Asimov



# FOUR VERY IMPORTANT REMARKS

**NEVER** work on the original dataset. Make copies along the way.

Document **ALL** your cleaning steps and procedures.

If you find yourself cleaning too much of your data, **STOP**. Something might be off with the data collection procedure.

Think **TWICE** before discarding an entire record.

# APPROACHES TO DATA CLEANING

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

- methodical
- narrative

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

The **narrative** approach consists of **exploring** the dataset and trying to spot unlikely and irregular patterns.

# PROS AND CONS

## Methodical (syntax)

- Pros: checklist is **context-independent**; pipelines **easy to implement**; common errors and invalid observations **easily identified**
- Cons: may prove **time-consuming**; cannot identify new types of errors

## Narrative (semantics)

- Pros: process may simultaneously yield **data understanding**; false starts are (at most) as costly as switching to mechanical approach
- Cons: may miss important sources of errors and invalid observations for datasets with **high number of features**; domain knowledge may bias the process by neglecting uninteresting areas of the dataset

# TOOLS AND METHODS

## Methodical

- list of potential problems (Data Cleaning Bingo)
- code which can be re-used in different contexts

## Narrative

- visualization
- data summary
- distribution tables
- small multiples
- data analysis

## Data Cleaning Bingo

random missing values	outliers	values outside of expected range - numeric	factors incorrectly/inconsistently 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 inconsistencies within field
entire blank rows	character encoding issues	duplicate value in unique field	non-factor values in factor	numeric values in character field



# APPROACHES TO DATA CLEANING

The narrative approach is akin to working out a crossword puzzle with a pen and putting down potentially wrong answers **occasionally**, to see where that takes you.

The mechanical approach is akin to working it out with a pencil, a dictionary, and never jotting down an answer unless you are certain it is correct.

You'll solve more puzzles (and it will be flashier) the first way, but you'll rarely be wrong the second way.

It's the same thing with data: analysts must be comfortable with both approaches.

# TYPES OF MISSING OBSERVATIONS

Blank fields come in 4 flavours:

- **Nonresponse**  
an observation was expected but none had been entered
- **Data Entry Issue**  
an observation was recorded but was not entered in the dataset
- **Invalid Entry**  
an observation was recorded but was considered invalid and has been removed
- **Expected Blank**  
a field has been left blank, but expectedly so

Too many missing values (of the first three type) can be indicative of **issues with the data collection process** (more on this later); too many missing values (of the fourth type) can be indicative of **poor questionnaire design**.

# THE CASE FOR IMPUTATION

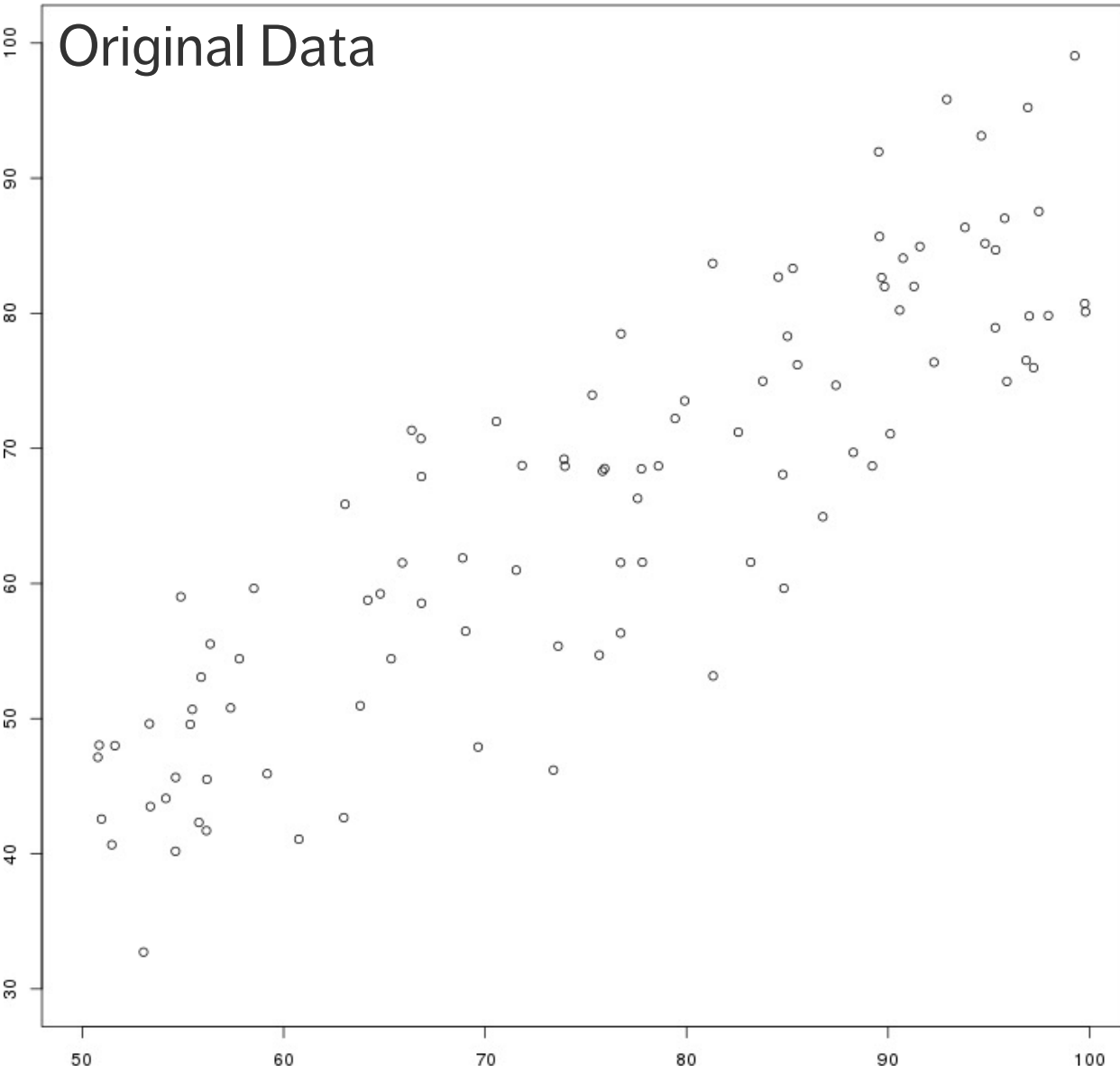
Not all analytical methods can easily accommodate missing observations – 2 options:

- **Discard** the missing observation
  - not recommended, unless the data is missing completely randomly in the dataset as a whole
  - acceptable in certain situations (such as a small number of missing values in a large dataset)
- Come up with a **replacement (imputation) value**
  - main drawback: we never know what the true value would have been
  - often the best available option

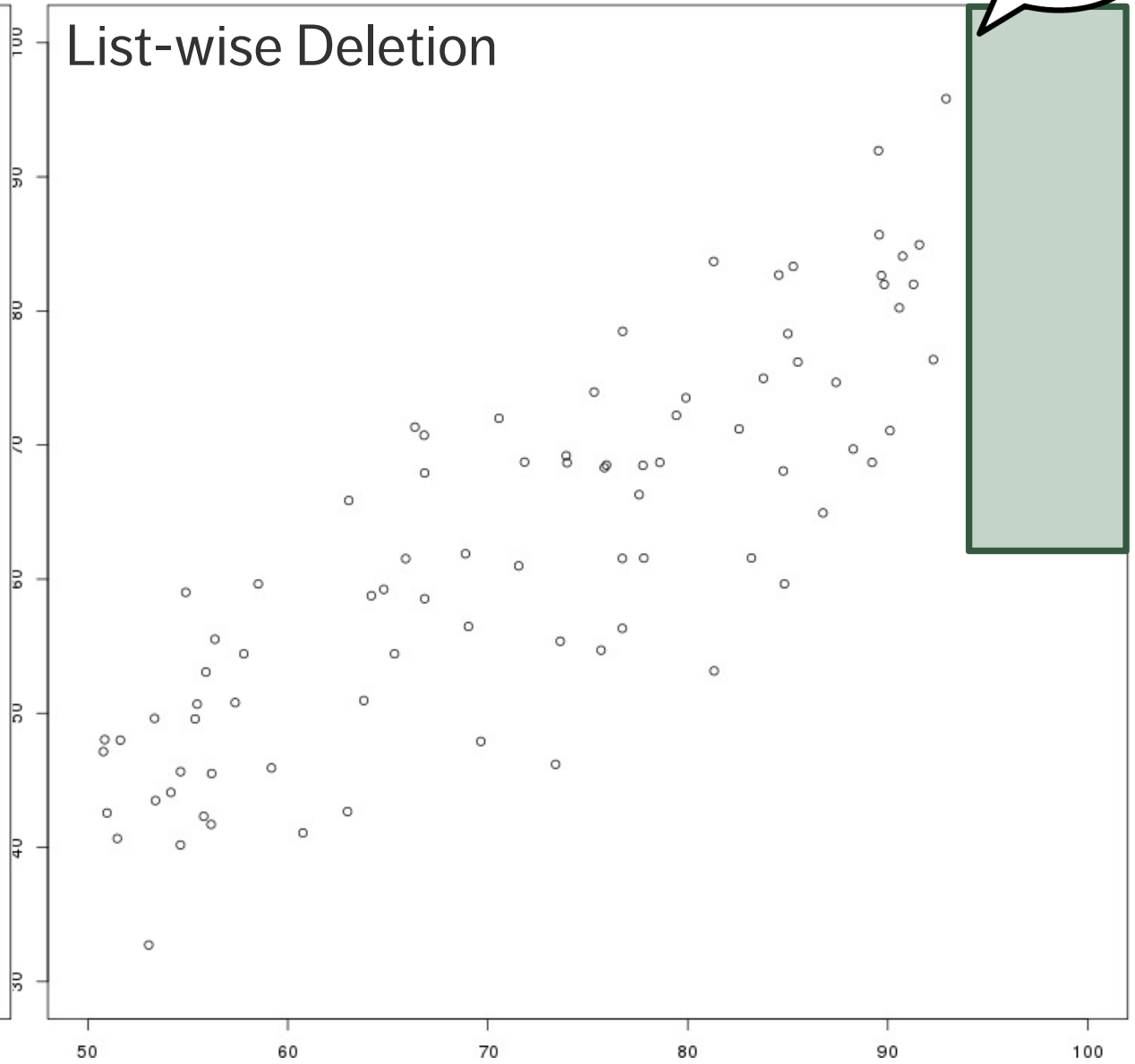
**Artificial data:** the  $y$  values of all points for which  $x > 92$  have been erased by mistake.



Original Data



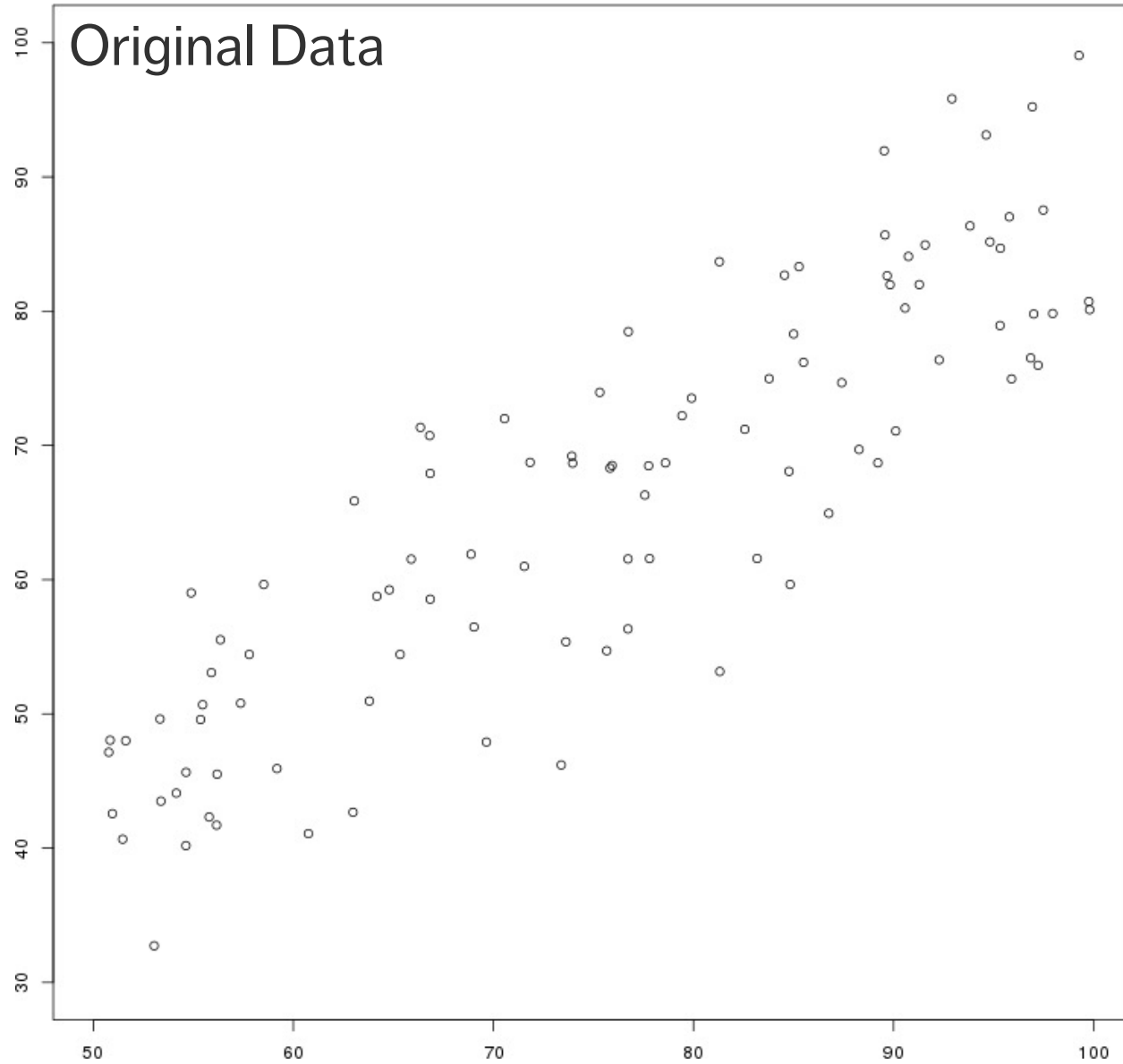
List-wise Deletion



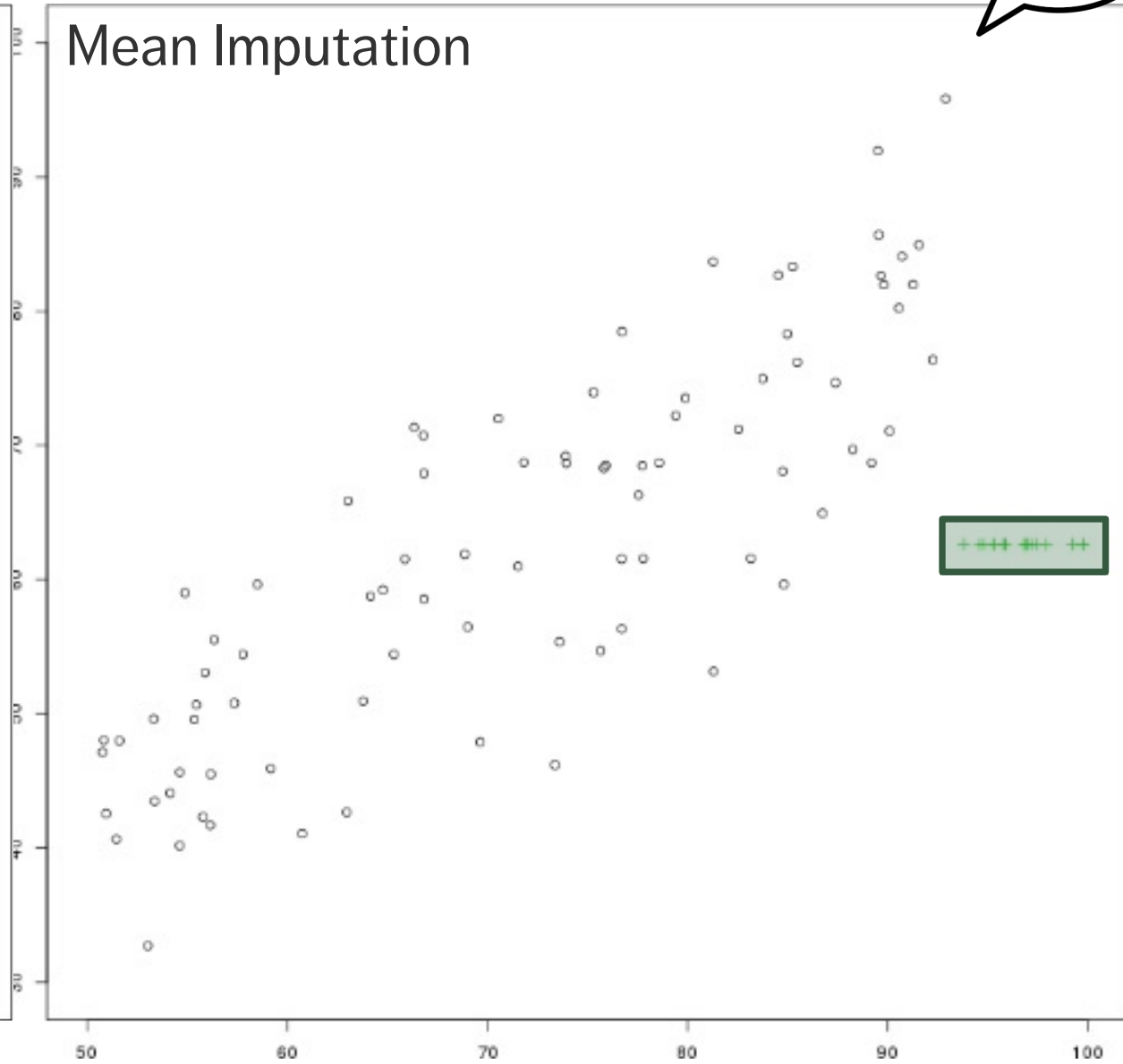
**Artificial data:** the  $y$  values of all points for which  $x > 92$  have been erased by mistake.



Original Data



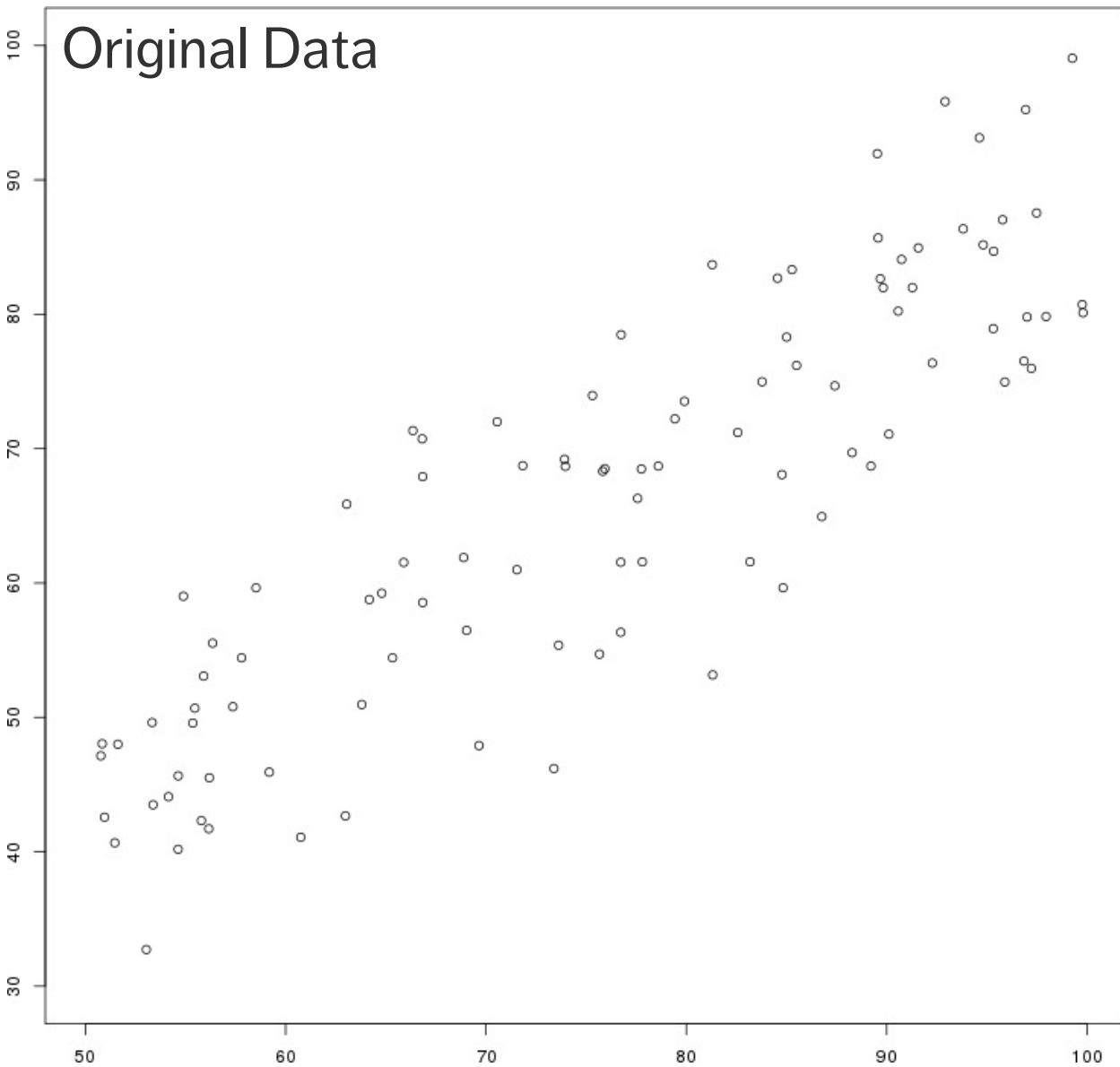
Mean Imputation



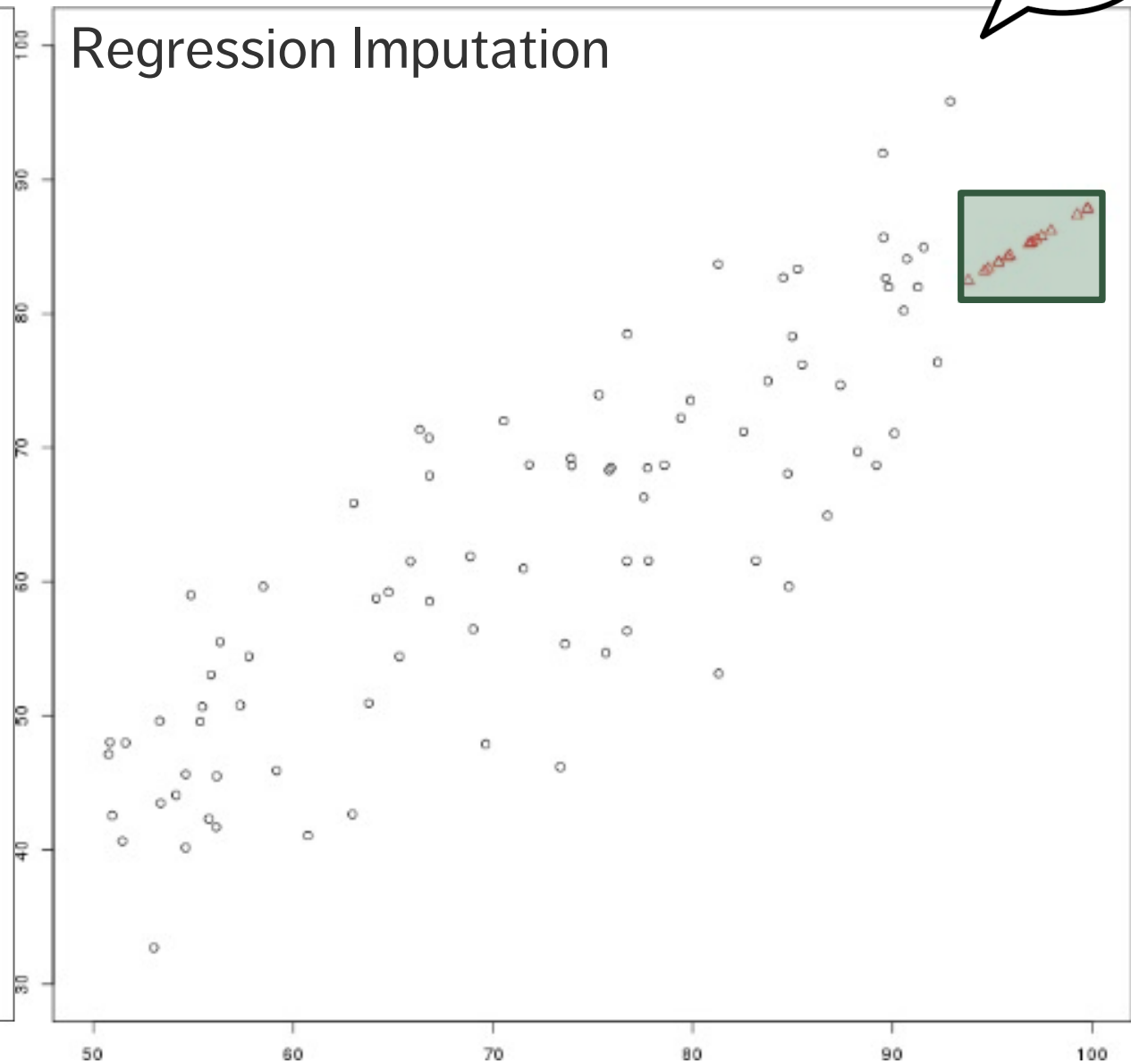
**Artificial data:** the  $y$  values of all points for which  $x > 92$  have been erased by mistake.



Original Data

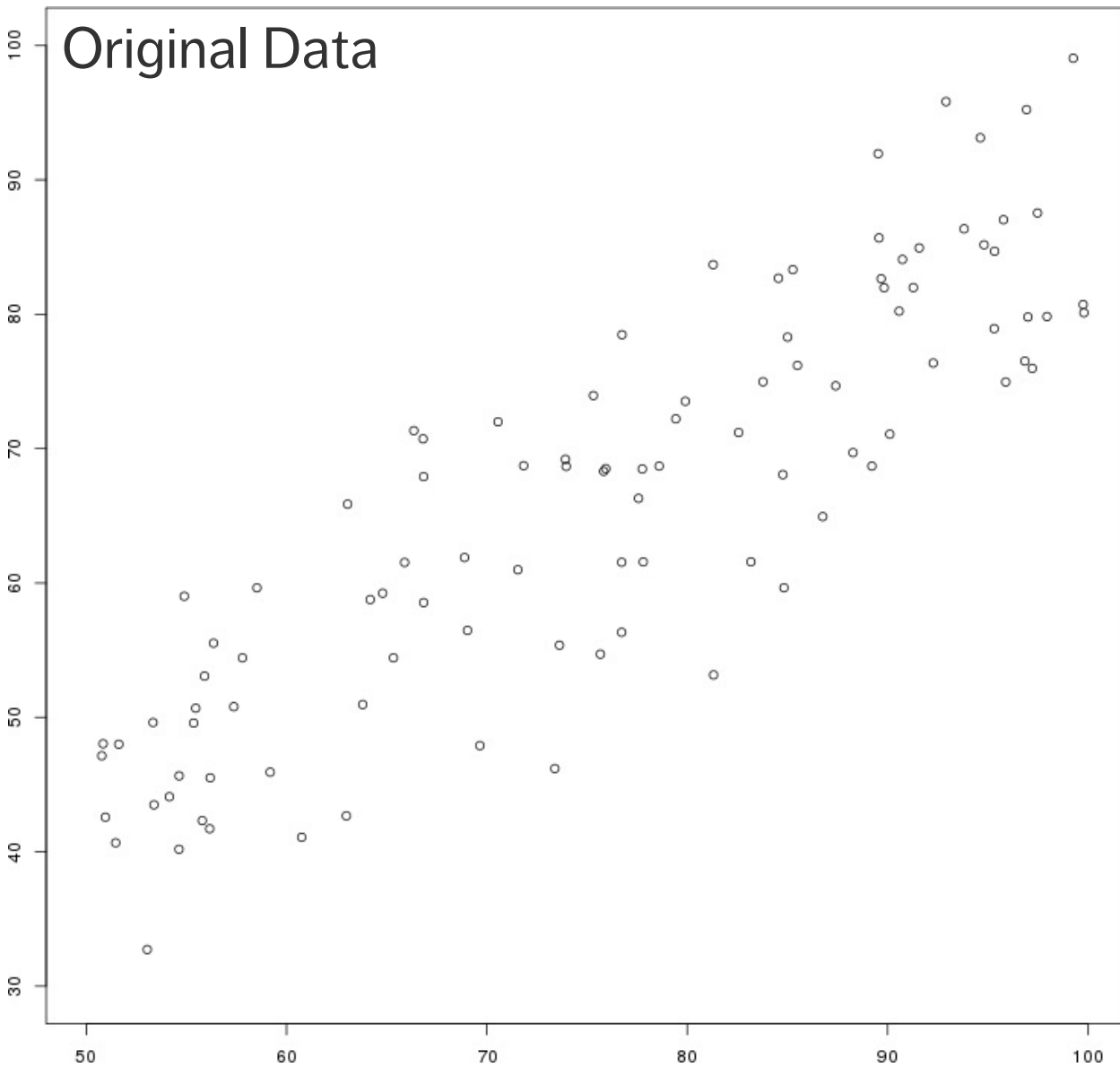


Regression Imputation

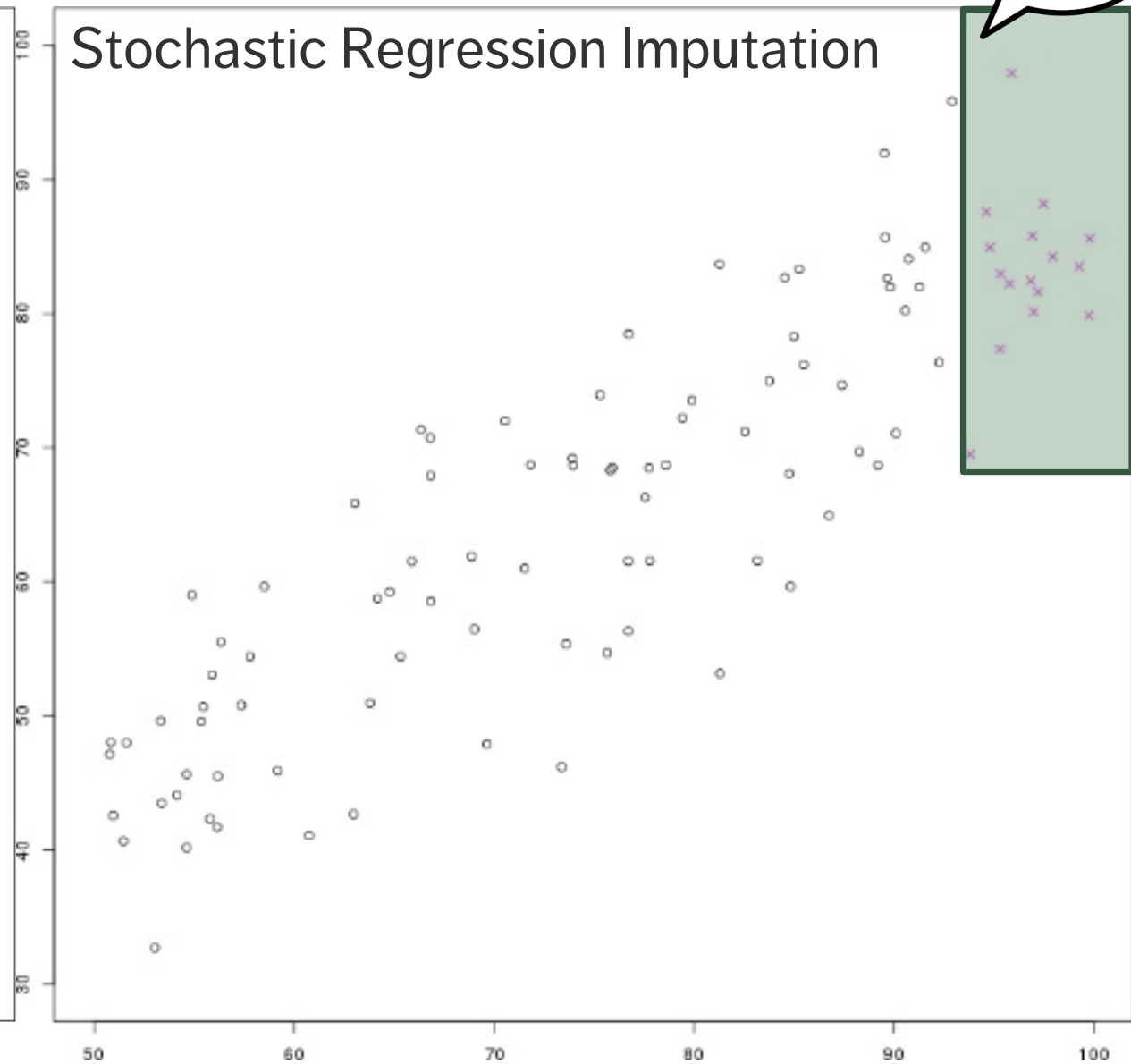


**Artificial data:** the  $y$  values of all points for which  $x > 92$  have been erased by mistake.

Original Data



Stochastic Regression Imputation



# TAKE-AWAYS

Missing values cannot simply be ignored.

The missing mechanism cannot typically be determined with any certainty.

Imputation methods work best when values are missing completely at random or missing at random, but imputation methods tend to produce biased estimates.

In single imputation, imputed data is treated as the actual data; multiple imputation can help reduce the noise.

Is stochastic imputation best? In our example, yes – but beware the *No-Free Lunch* theorem!



# SPECIAL DATA POINTS

**Outlying observations** are data points which are **atypical** in comparison to

- the unit's remaining features (**within-unit**),
- the field measurements for other units (**between-units**),

or as part of a collective subset of observations.

Outliers are observations which are **dissimilar to other cases** or which **contradict known dependencies** or rules.

Careful study is needed to determine whether outliers should be retained or removed from the dataset.

# SPECIAL DATA POINTS

**Influential data points** are observations whose absence leads to **markedly different** analysis results.

When influential observations are identified, **remedial measures** (data transformations) may be required to minimize their undue effects.

Outliers **may** be influential data points, influential points **need not be** outliers.

# DETECTING ANOMALIES

Outliers may be anomalous along any of the unit's variables, or in combination.

Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

Differentiating anomalies from noise or data entry errors is **hard**.

Boundaries between normal and deviating units may be **fuzzy**.

When anomalies are associated with malicious activities, they are typically **disguised**.

# DETECTING ANOMALIES

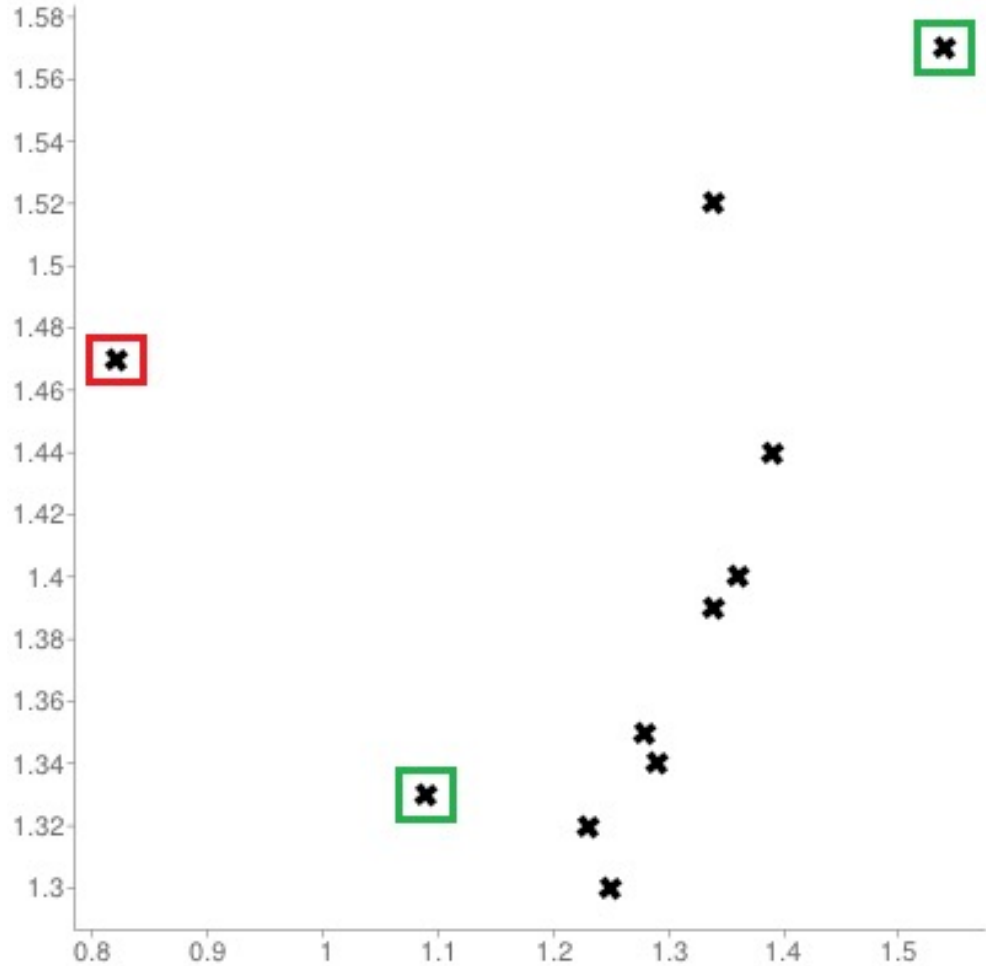
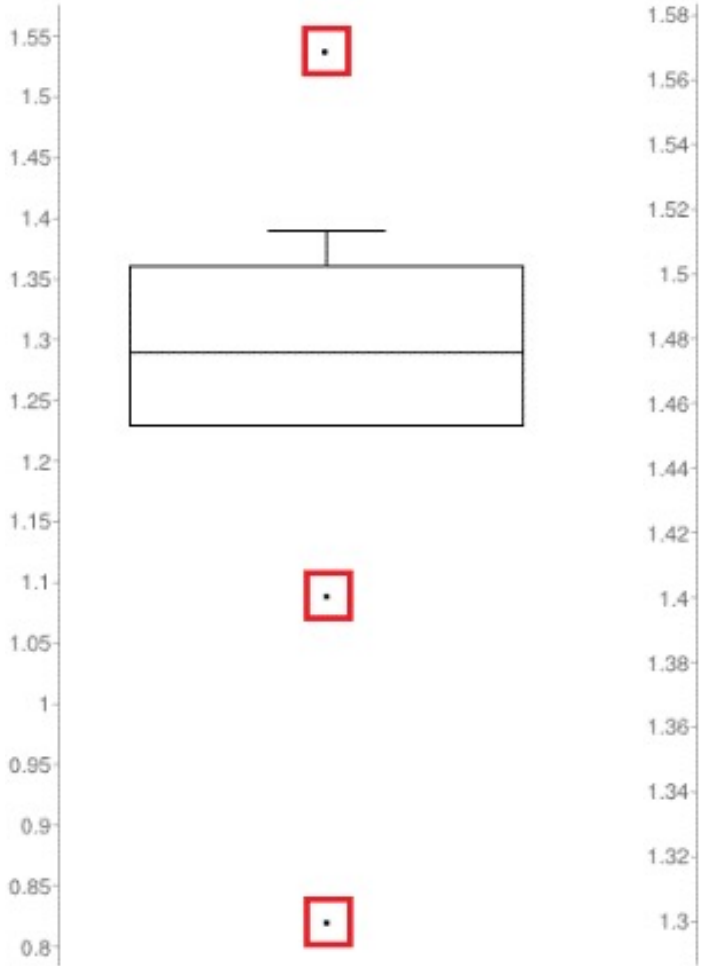
Numerous methods exist to identify anomalous observations; **none of them are foolproof** and judgement must be used.

Graphical methods are easy to implement and interpret.

- **Outlying Observations**  
box-plots, scatterplots, scatterplot matrices, Cooke's distance, normal qq plots
- **Influential Data**  
some level of analysis must be performed (leverage)

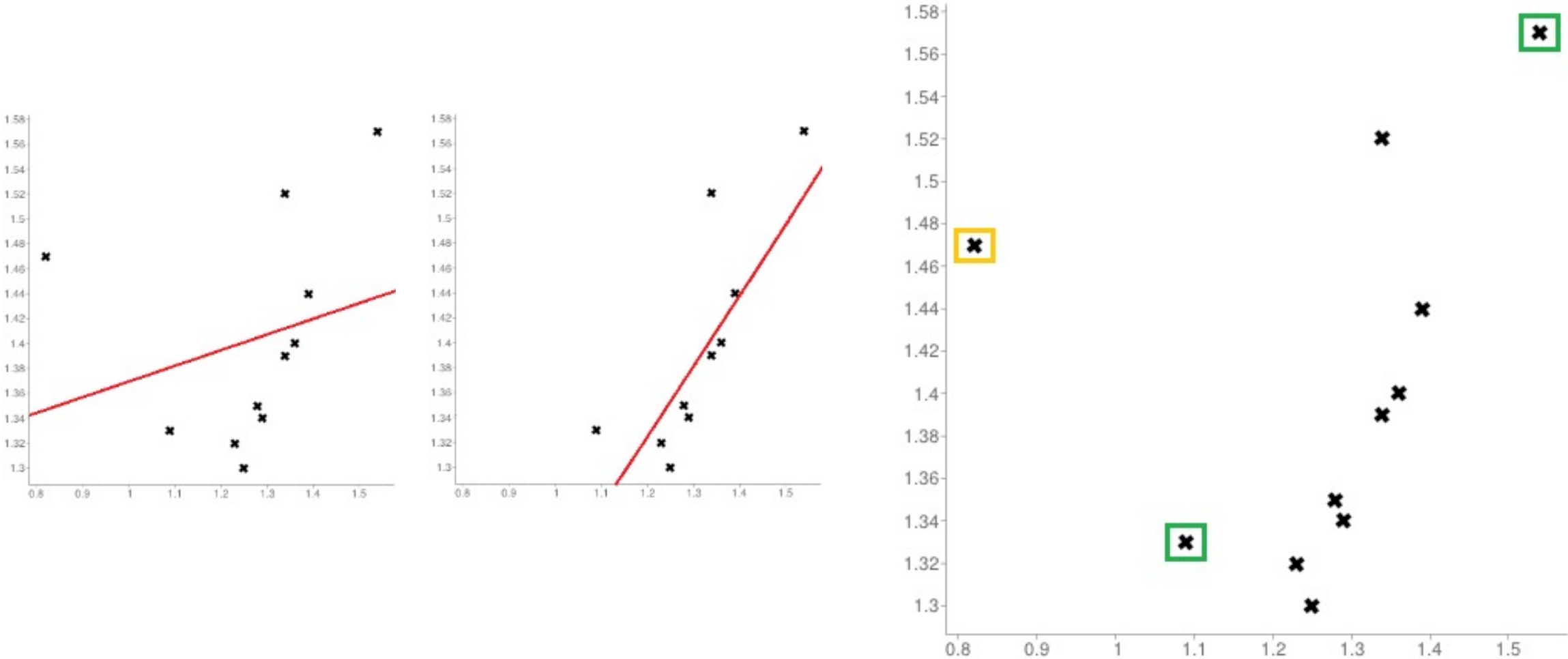
Once anomalous observations have been removed from the dataset, previously “regular” units may become anomalous.

# OUTLIERS



Queuing dataset: processing rate vs. arrival rate

# INFLUENTIAL OBSERVATIONS



Queuing dataset: processing rate vs. arrival rate



# TAKE-AWAYS

Identifying influential points is an iterative process as the various analyses have to be run numerous times.

Fully automated identification and removal of anomalous observations is **NOT recommended**.

Use transformations if the data is **NOT** normally distributed.

Whether an observation is an outlier or not depends on various factors; what observations end up being influential data points depends on the specific analysis to be performed.

# DATA REDUCTION AND TRANSFORMATIONS

DATA PROCESSING





# DIMENSIONALITY OF DATA

In data analysis, the **dimension** of the data is the number of variables (or attributes) that are collected in a dataset, represented by the number of columns.

The term dimension is an extension of the use of the term to refer to the size of a vector.

We can think of the variables used to describe each object (row) as a **vector** describing that object.

**Note:** the term dimension is used differently in business intelligence contexts.



But more data is always better, right?

It depends.

# HIGH DIMENSIONALITY & BIG DATA

Datasets can be “big” in a variety of ways:

- too large for the **hardware** to handle (cannot be stored or accessed properly due to # of observations, # of features, or the overall size)
- size can go against **modeling assumptions** (# of features  $\gg$  # observations)

## Examples:

- Multiple sensors recording 100+ observations per second in a large geographical area over a long time period = **very big dataset**.
- In a corpus' Term Document Matrix (cols = terms, rows = documents), the number of terms is usually substantially higher than the number of documents, leading to **excessively sparse data**.

# CURSE OF DIMENSIONALITY

Unless the dataset size grows exponentially with its dimension, the performance of any model we build is likely to suffer due to the **Curse of Dimensionality**.

Possible solutions:

- **sampling observations**
- **feature selection** (easy-ish) and/or dimension reduction (hard).

We look for ways to preserve the signal while shrinking the dimension: it's easier to find needles in small haystacks!

(This is actually a thorny problem... but we'll avoid the technical details in this course).

# SAMPLING OBSERVATIONS

**Question:** does every observation (row of the dataset) need to be used?

If rows are selected randomly, the resulting sample might be **representative** of the entire dataset.

## **Drawbacks:**

- if the signal of interest is rare, sampling might drown it altogether
- if aggregation is happening down the road, sampling will necessarily affect the numbers (passengers vs. flights)
- even simple operations on a large file (finding the # of lines, say) can be taxing on the memory and in terms of computation time – **prior information on the dataset structure can help**

# FEATURE SELECTION

Removing **irrelevant** or **redundant** variables is a common data processing task.

## Motivations:

- modeling tools do not handle these well (variance inflation due to multicollinearity, etc.)
- dimension reduction (# variables  $\gg$  # observations)

## Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

# FEATURE SELECTION METHODS

**Filter methods** inspect each variable individually and score them according to some **importance metric**.

The less relevant features (i.e. importance score below some set threshold) are then removed.

**Wrapper methods** seek feature subsets for which the evaluation criterion used by the eventual analytical method is “optimized”.

The process is **iterative**, and typically computationally intensive: candidate subsets are used in the analysis until one produces an acceptable evaluation metric for the analysis.

# FEATURE SELECTION METHODS

**Unsupervised methods** determine the importance of a feature based only on its values.

**Supervised methods** evaluate each feature's importance by studying the relationship with a **target feature** (correlation, etc.)

Wrapper methods are usually supervised.

**Unsupervised filter methods:** removing constant variables, ID-like variables (different on all observations), features with low variability, etc.



# DISCRETIZING

To reduce computational complexity, a numeric variable may need to be replaced by an **ordinal** variable (from *height* value to “*short*”, “*average*”, “*tall*”, for instance).

**Domain expertise** can be used to determine the bins’ limits (although that 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
- the bins each have the same width
- the performance of some modeling tool is maximized

# DATA QUALITY AND DATA VALIDATION

DATA PROCESSING

**Martin:** Data is messy.

**Allison:** Even when it's been cleaned?

**Martin:** Especially when it's been cleaned.

P. Boily, *Introduction to Quantitative Consulting*

# SOUND DATA

The ideal dataset will have as few issues as possible with:

- **Validity:** data type, range, mandatory response, uniqueness, value, regular expressions
- **Completeness:** missing observations
- **Accuracy and Precision:** related to measurement and/or data entry errors; target diagrams (accuracy as bias, precision as standard error)
- **Consistency:** conflicting observations
- **Uniformity:** are units used uniformly throughout?

Checking for data quality issues at an early stage can save headaches later in the analysis.

# SOUND DATA



accurate and  
precise



precise but  
not accurate



accurate but  
not precise



neither accurate  
nor very precise

# COMMON SOURCES OF ERROR

When dealing with **legacy**, **inherited** or **combined** datasets (that is, datasets over which you have little control):

- missing data given a code
- 'NA'/'blank' given a code
- data entry error
- coding error
- measurement error
- duplicate entries
- heaping

# DETECTING INVALID ENTRIES

Potentially invalid entries can be detected with the help of:

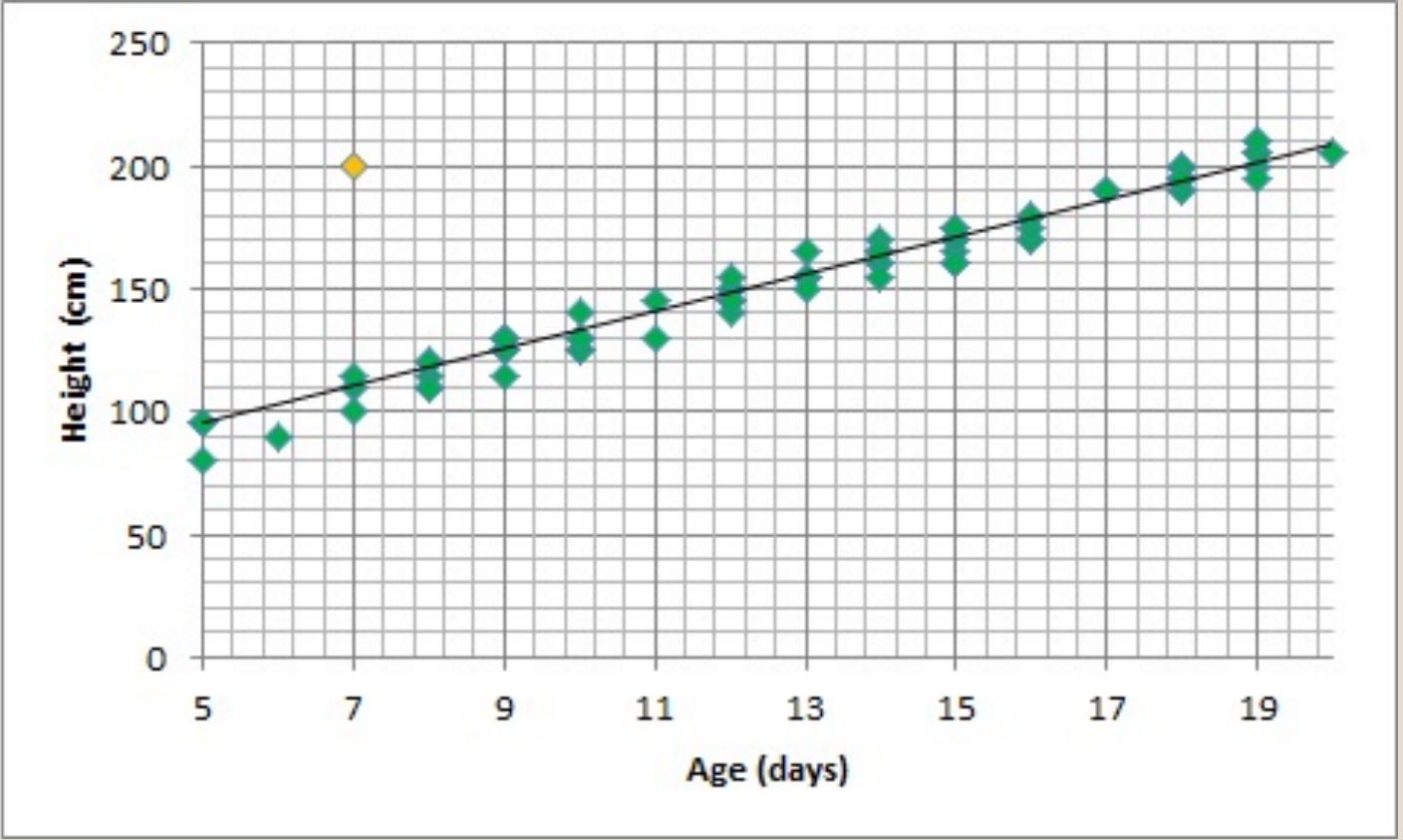
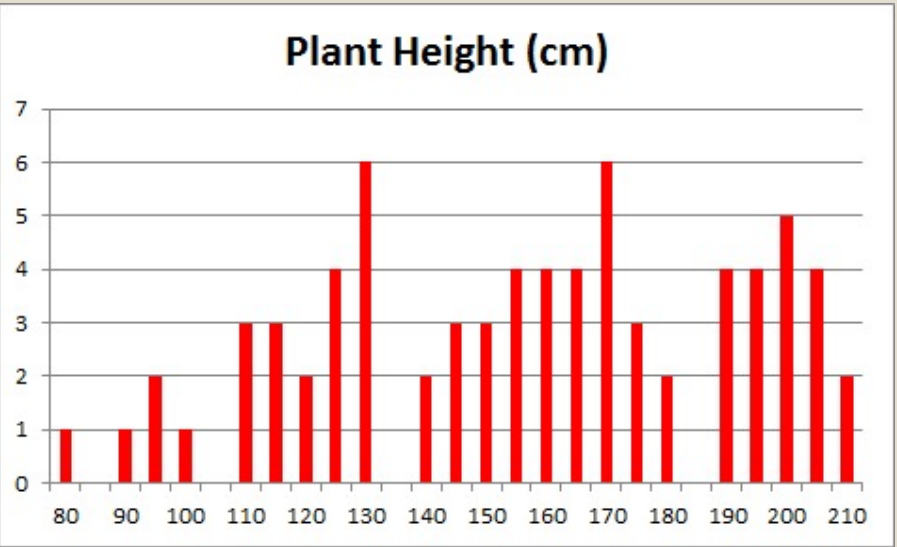
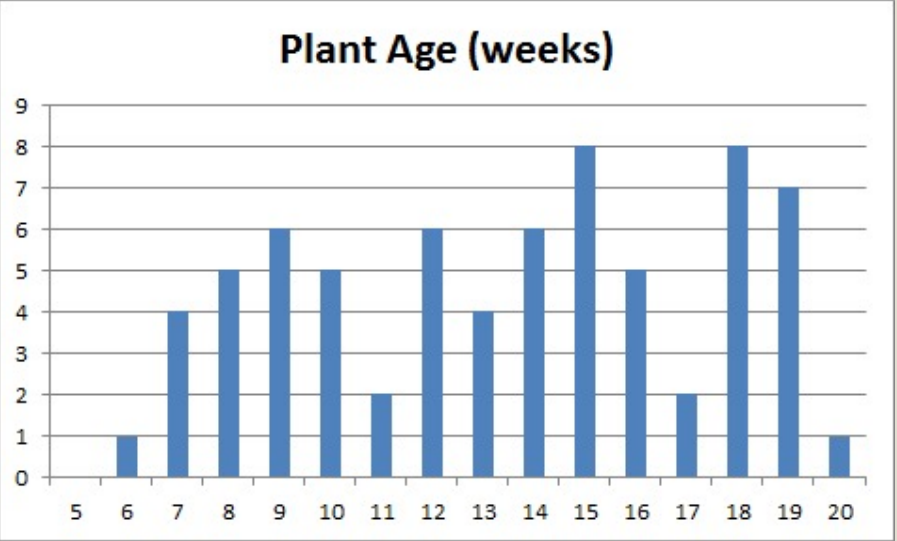
- **Univariate Descriptive Statistics**  
count, range, *z*-score, mean, median, standard deviation, logic check
- **Multivariate Descriptive Statistics**  
*n*-way table, logic check
- **Data Visualization**  
scatterplot, scatterplot matrix, histogram, joint histogram, etc.

This step might allow for the identification of potential outliers.

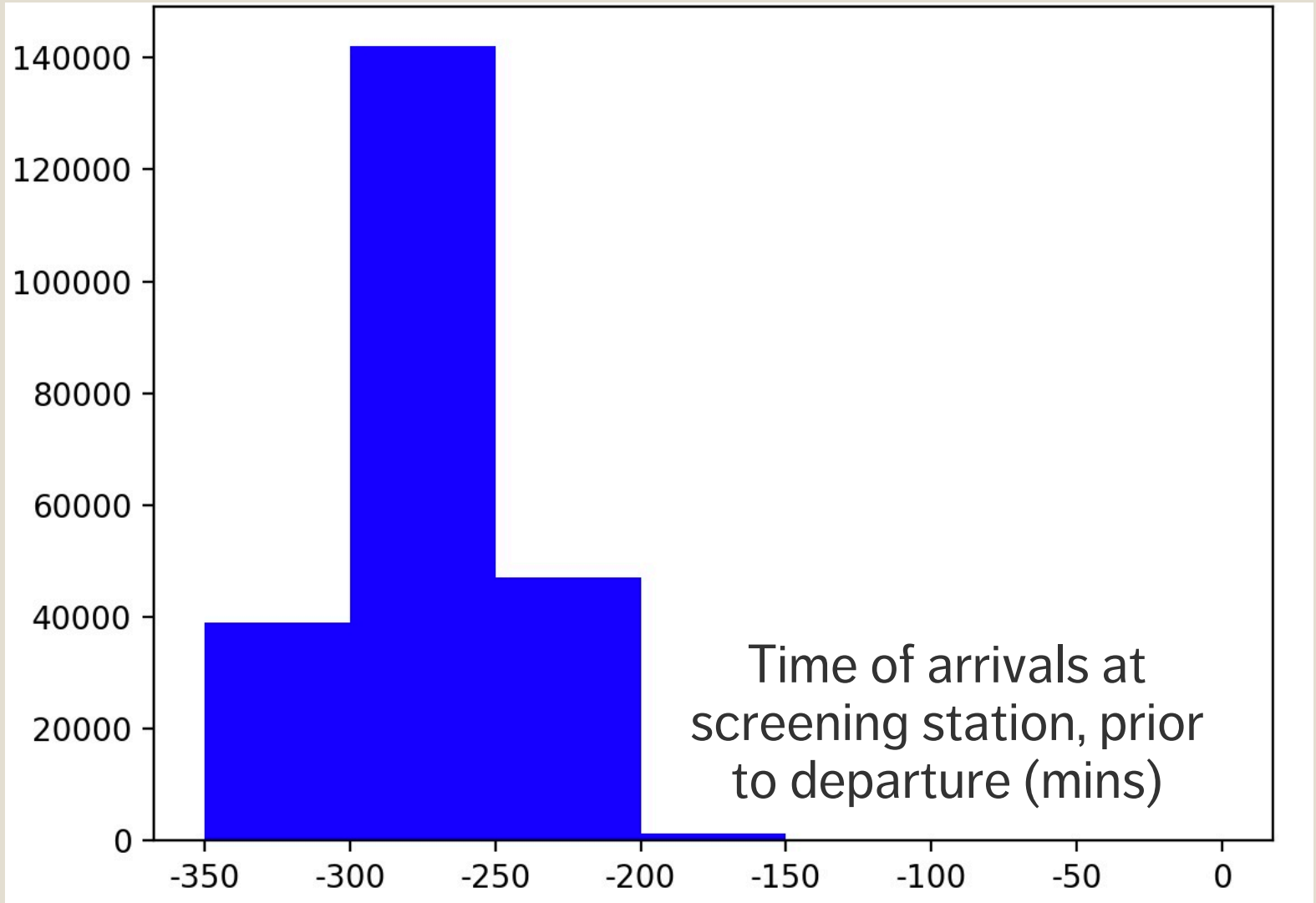
Failure to detect invalid entries **does not mean** that all entries are valid.

Small numbers of invalid entries recoded as “missing.”

# ILLUSTRATION



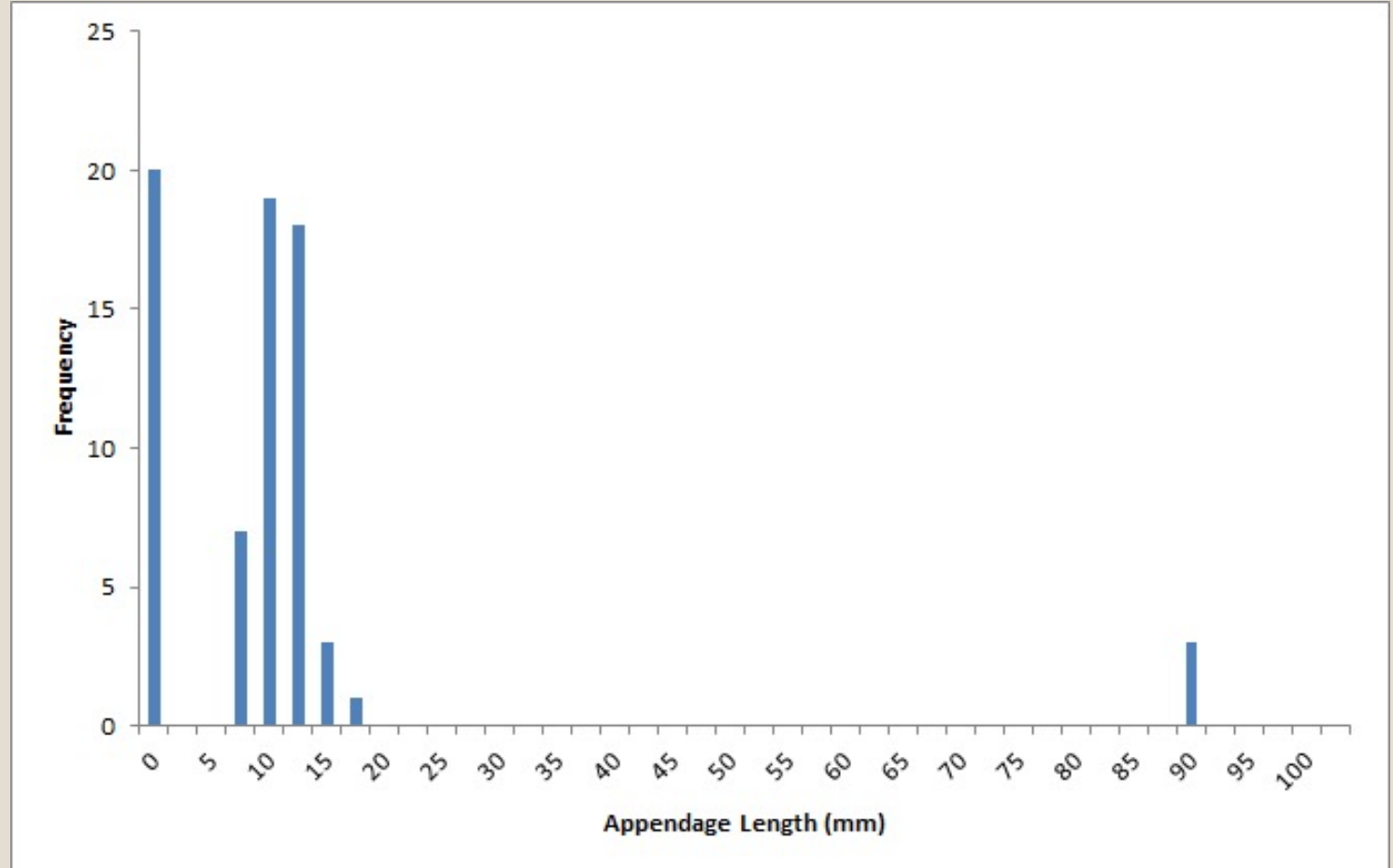
# ILLUSTRATION





# ILLUSTRATION

<i>Appendage length (mm)</i>	
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





# TAKE-AWAYS

Don't wait until after the analysis to find out there was a problem with data quality.

Univariate tests don't always tell the whole story.

Visualizations can help.

Context is crucial – you may need more context about the data in order to make sense of what you see... but whatever the situation, you need to understand the dataset quality.