DATA CLEANING

DATA COLLECTION AND 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

LEARNING OBJECTIVES

Recognize the strengths and weaknesses of both major data cleaning approaches

Identify methods to handle missing observations

Increase familiarity with various anomaly detection or outlier tests

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.

TAKE-AWAYS

The narrative approach is similar to working out a crossword puzzle with a pen and putting down potentially wrong answers every once in a while to see where that takes you.

The mechanical approach is similar 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.

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

TYPES OF MISSING OBSERVATIONS

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.

There are two 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 value**
  - main drawback: we never know for a fact what the true value would have been
  - often the best available option

MISSING MECHANISMS

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)

IMPUTATION METHODS

List-wise deletion

Mean or most frequent imputation

Regression or correlation imputation

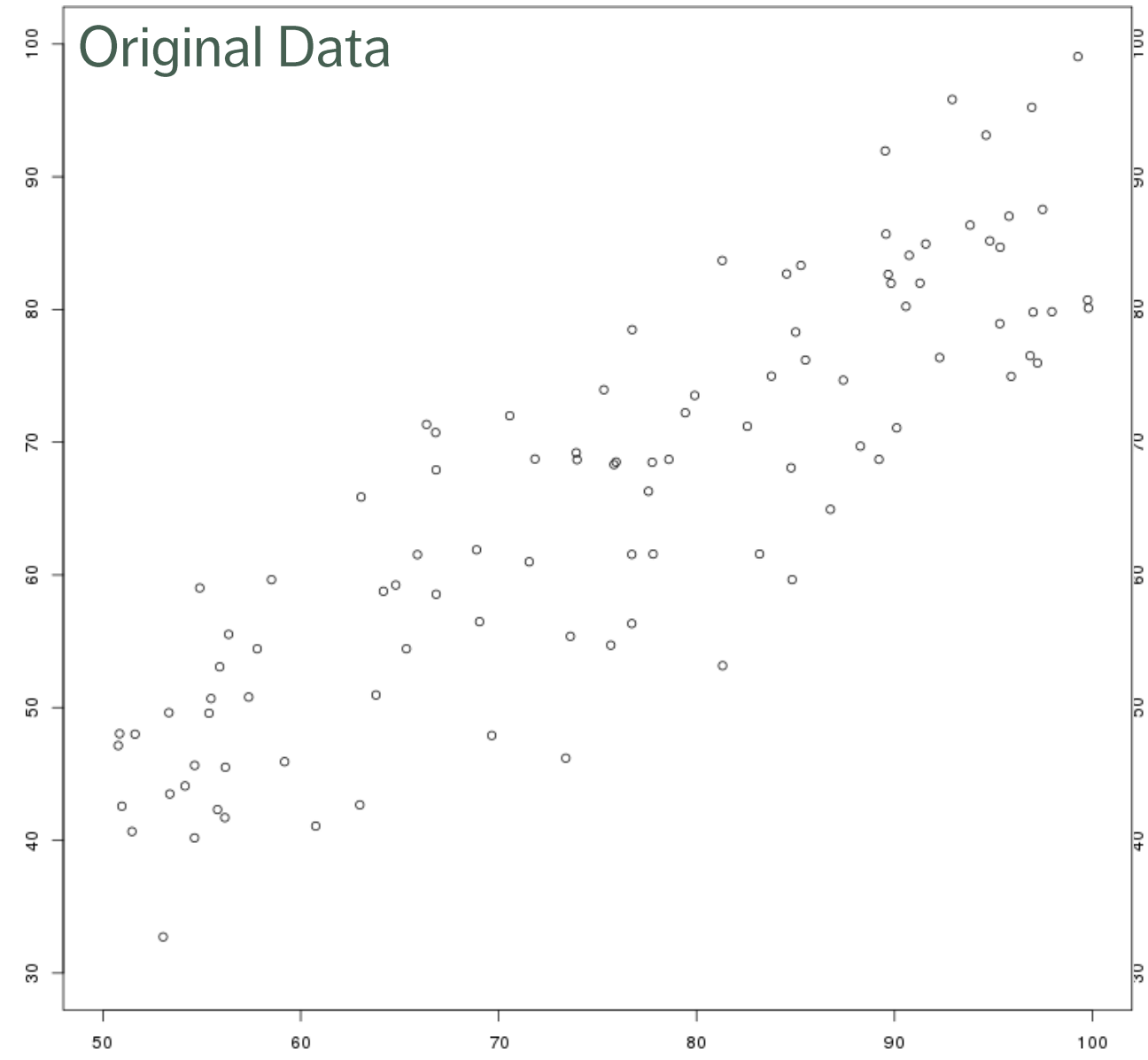
Stochastic regression imputation

Last observation carried forward

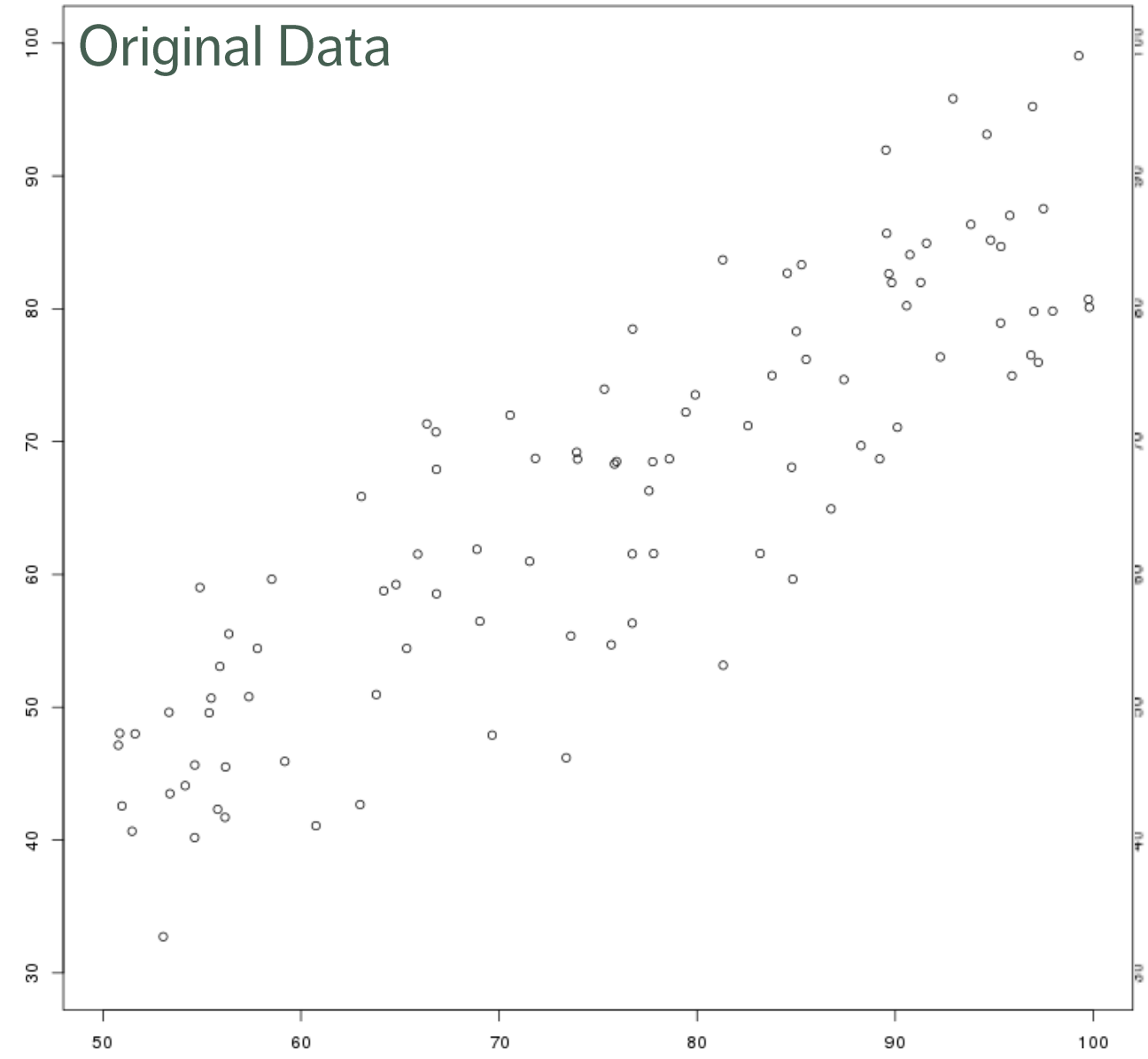
k-nearest neighbours imputation

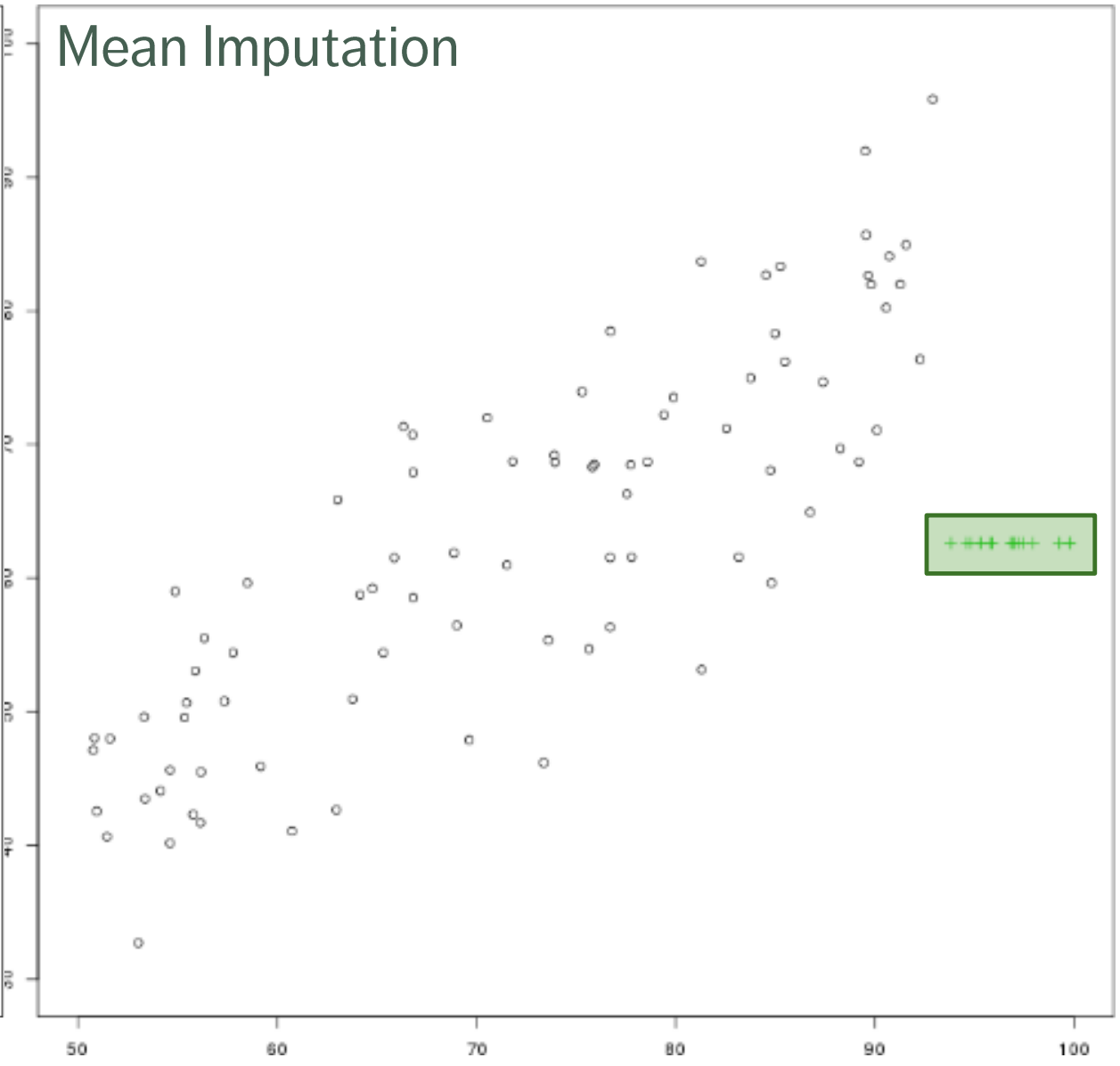
Multiple imputation

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

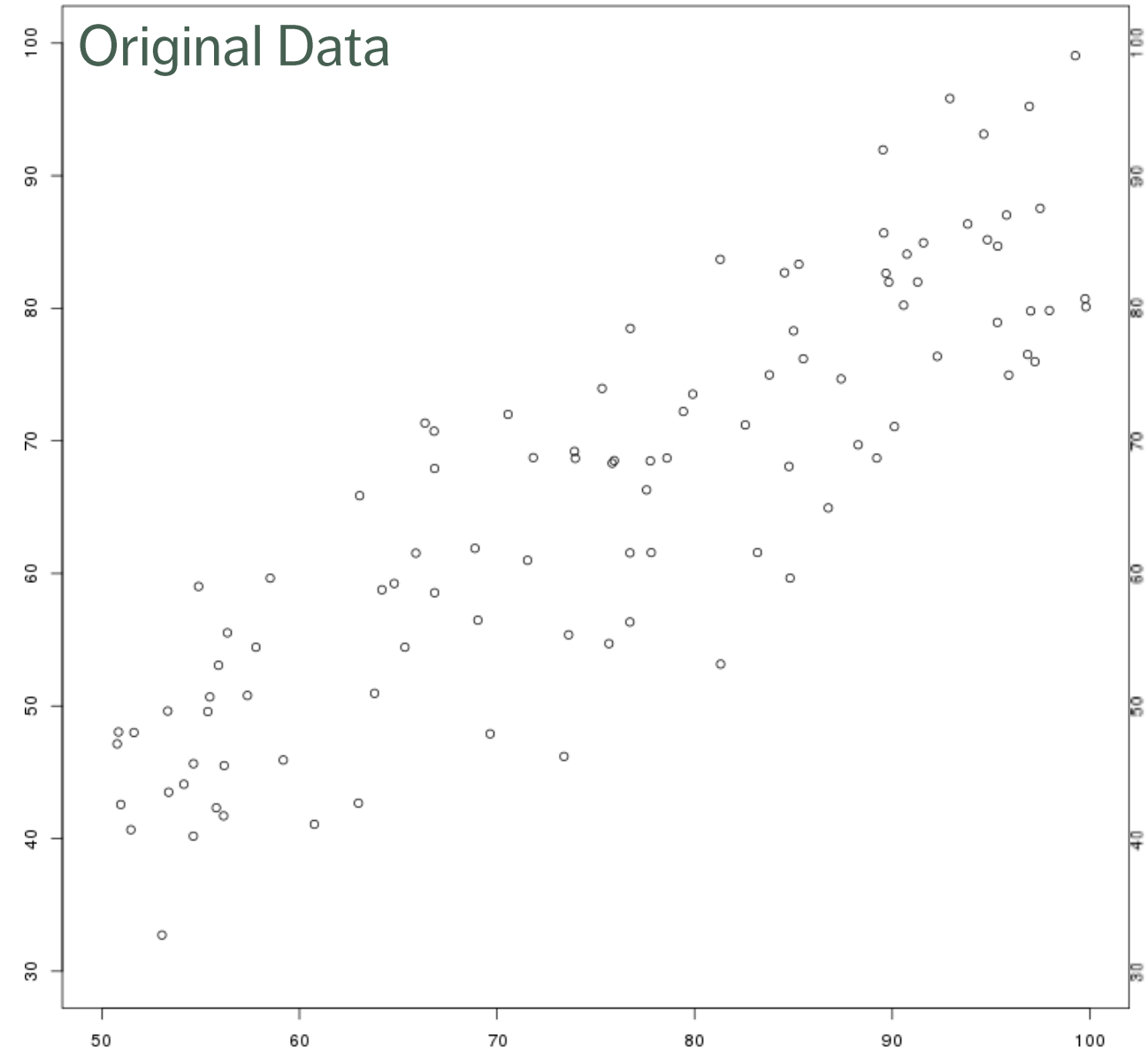


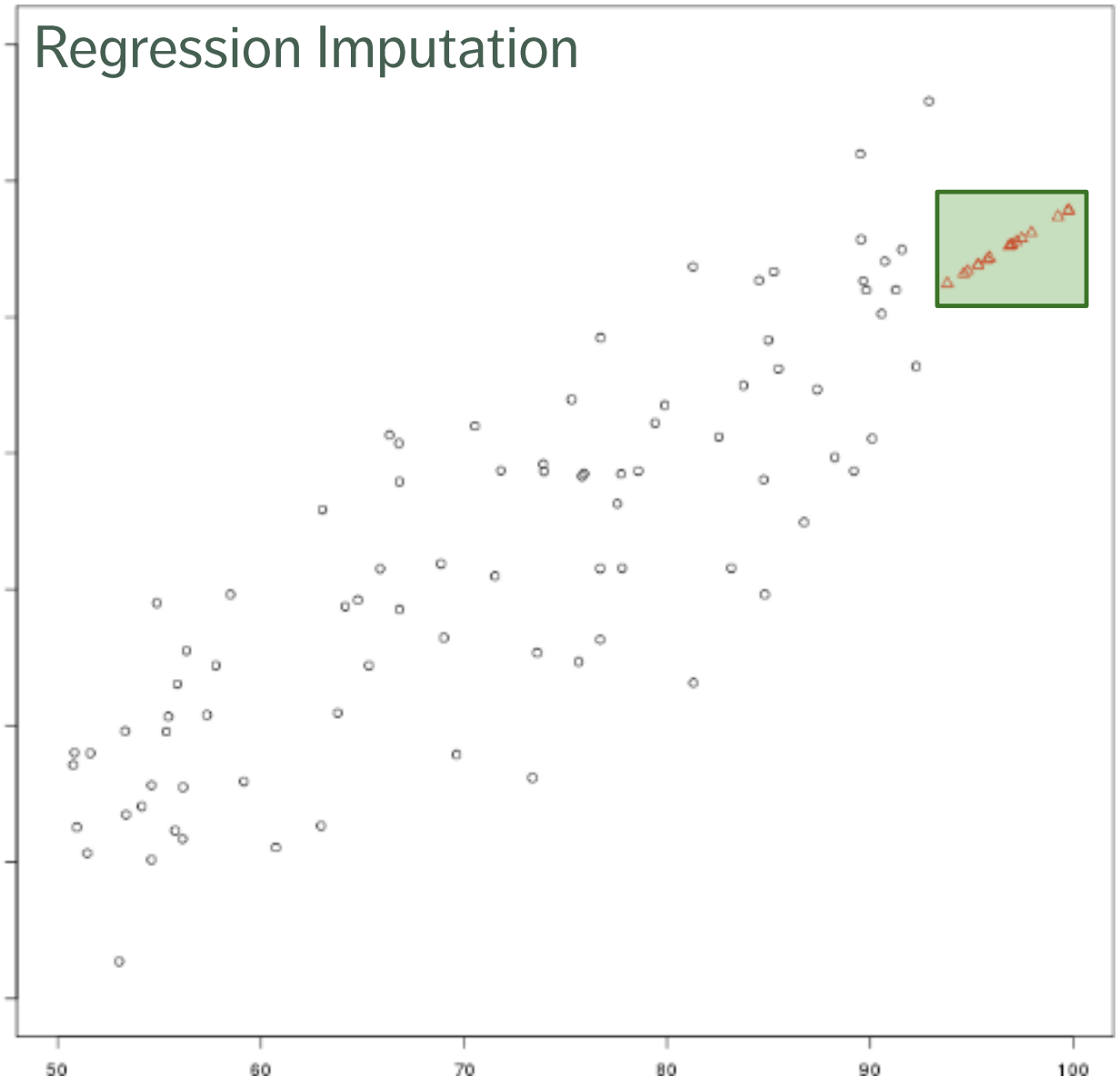
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



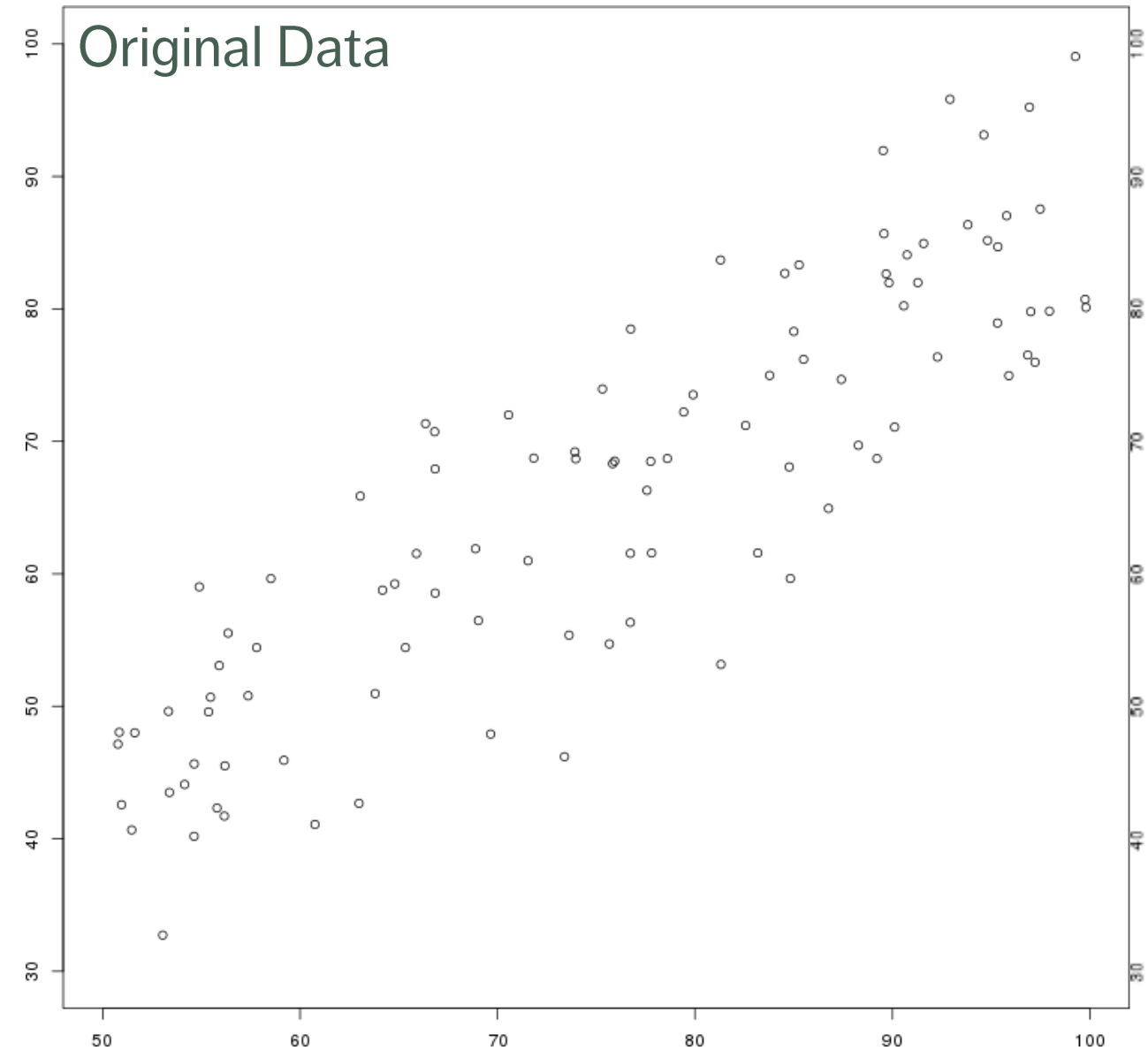


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





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



MULTIPLE IMPUTATION

Imputations increase the noise in the data.

In **multiple imputation**, the effect of that noise can be measured by consolidating the analysis outcome from multiple imputed datasets.

Steps:

1. Repeated imputation creates m versions of the dataset.
2. Each of these datasets is analyzed, yielding m outcomes.
3. The m outcomes are pooled into a single result for which the mean, variance, and confidence intervals are known.

MULTIPLE IMPUTATION

Advantages

- **flexible**; can be used in a various situations (MCAR, MAR, even NMAR in certain cases).
- accounts for **uncertainty** in imputed values
- fairly easy to implement

Disadvantages

- m may need to be fairly **large** when there are many missing values in numerous features, which slows down the analyses
- what happens if the analysis output is not a single value but some more complicated mathematical object?

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 (such as data transformations) may be required to minimize their undue effects.

Outliers may be influential data points, yet influential data points need not be outliers (weighted data).

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, 2D tour, 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.

OUTLIER TESTS

Supervised methods use a historical record of labeled anomalous observations:

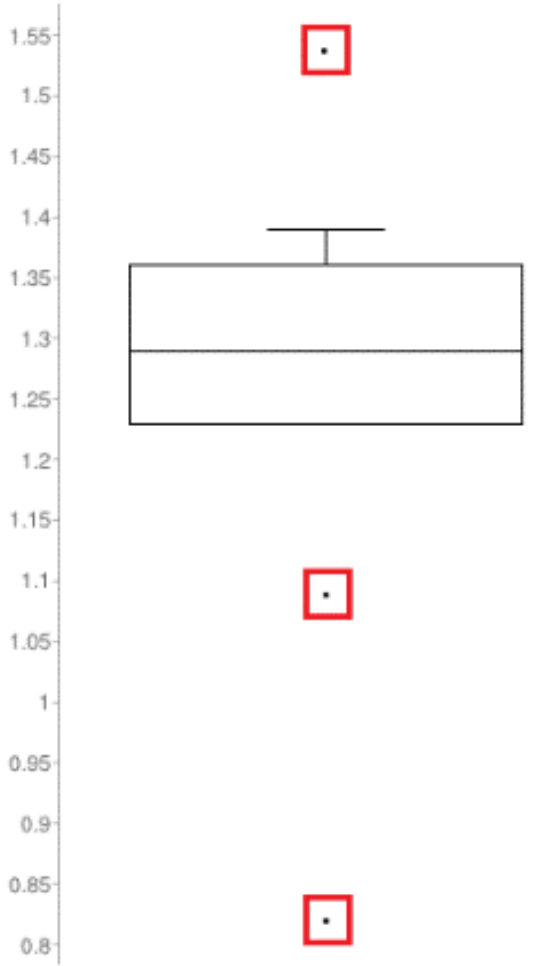
- domain expertise required to tag the data
- classification or regression task (probabilities and inspection rankings)
- rare occurrence problem (more on this later)

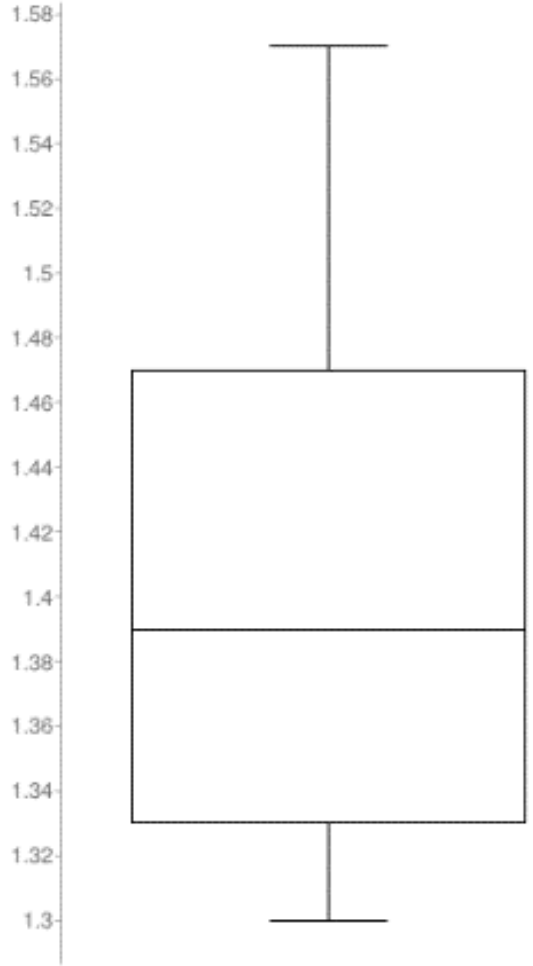
Unsupervised methods don't use external information:

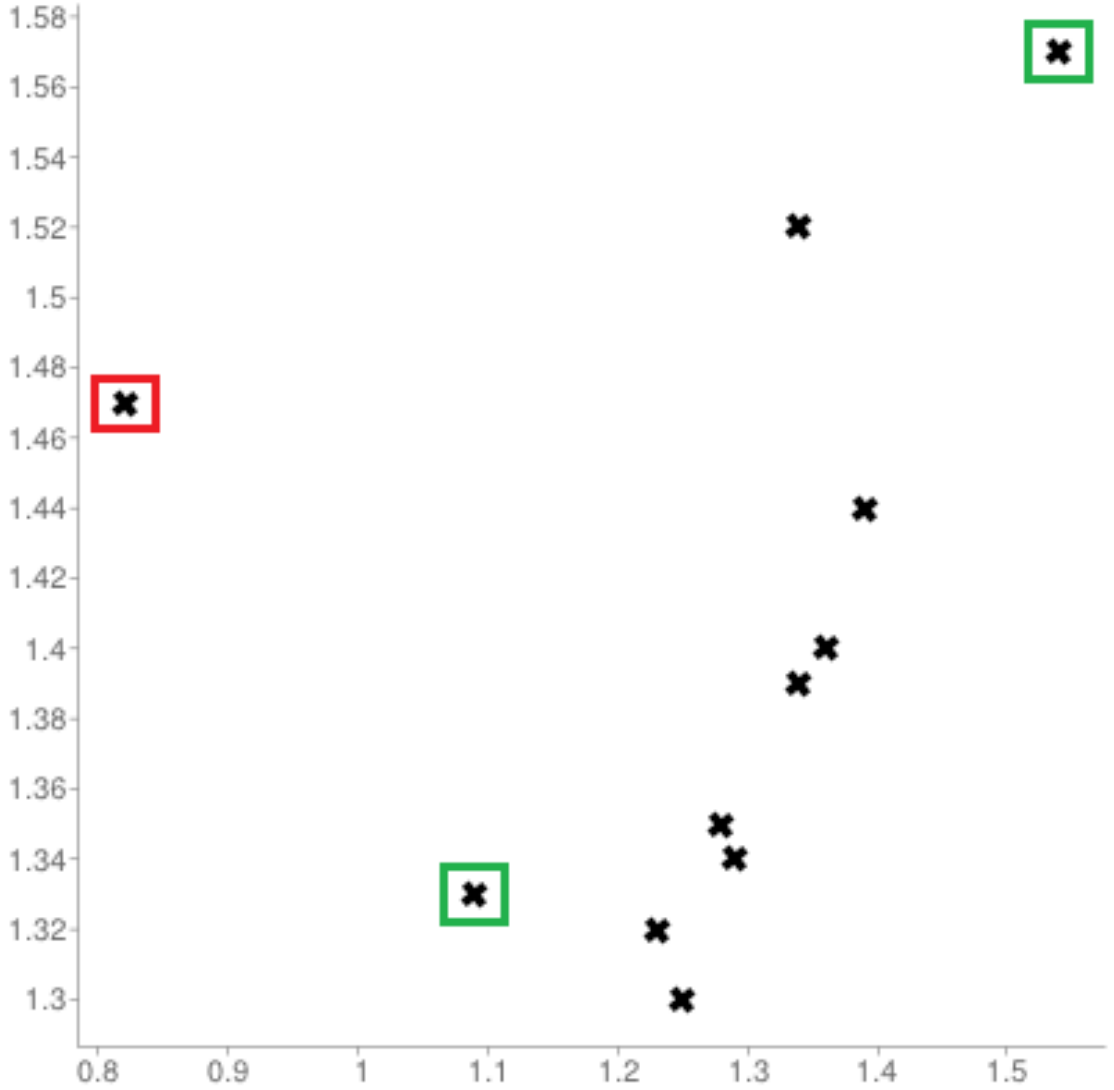
- traditional methods and tests
- can also be seen as a clustering or association rules problem

Semi-supervised methods also exist.

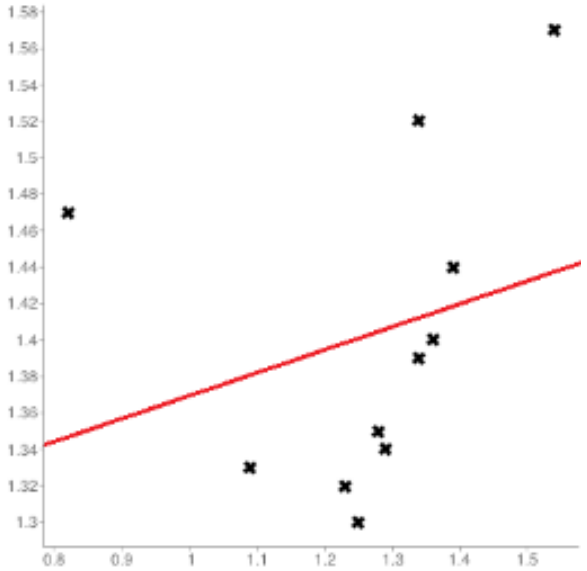
Queuing dataset: processing rate vs. arrival rate

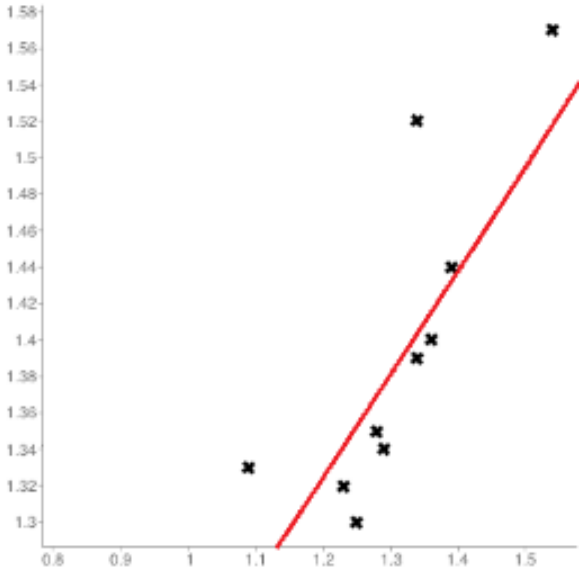


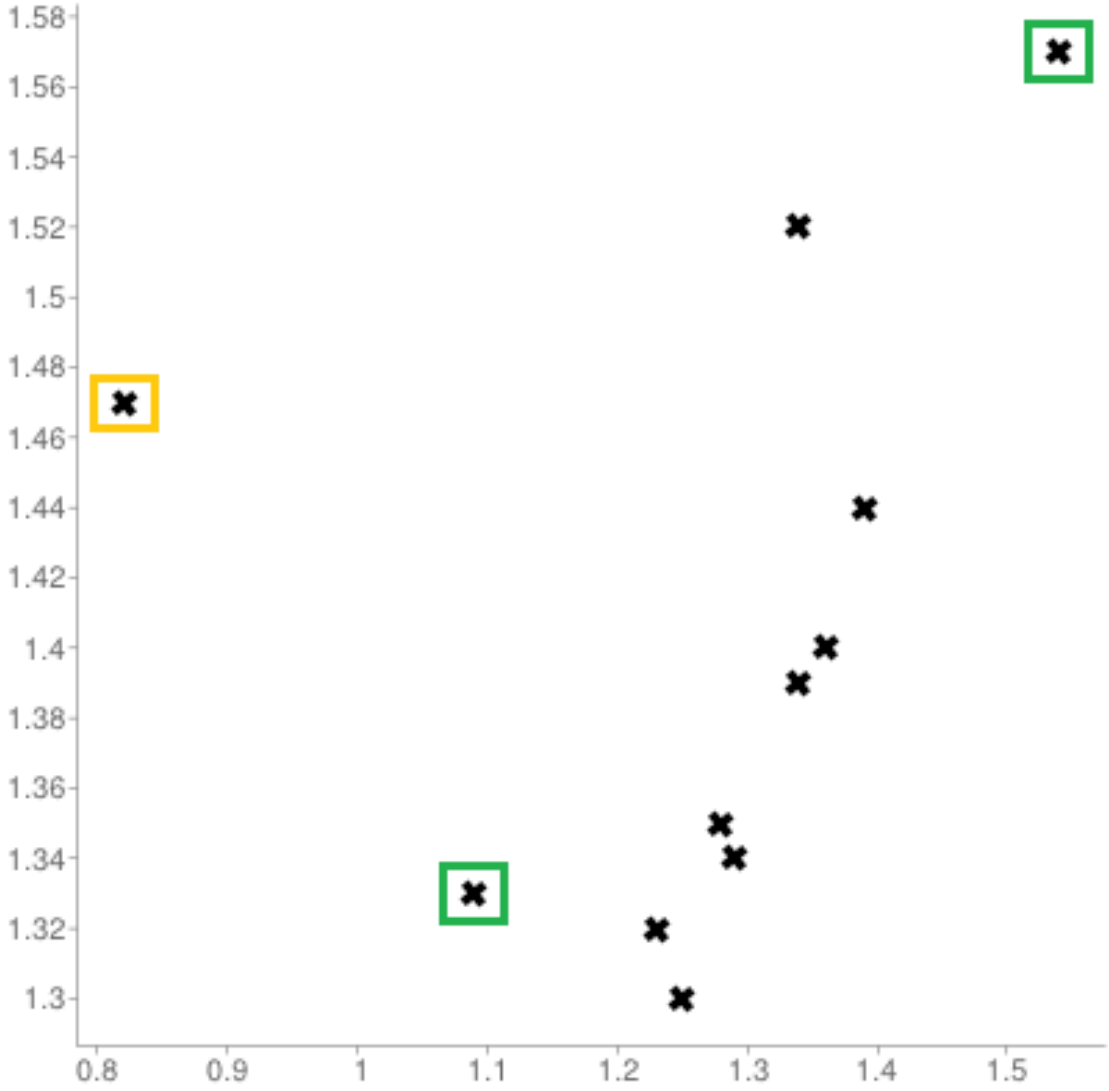




Queuing dataset: processing rate vs. arrival rate







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.

EXERCISES

The ability to monitor and perform early forecasts of various river algae blooms is crucial to control any ecological harm they can cause.

The `algae_bloom.csv` dataset is used to train a learning model consists of:

- **chemical properties** of various water samples of European rivers
- the **quantity of seven algae** in each of the samples, and
- the **characteristics of the collection process** for each sample.

What is the data science motivation for such a model, given that we **can** actually analyze water samples to determine if various harmful algae are present or absent?

EXERCISES

The answer is simple: chemical monitoring is **cheap** and **easy to automate**, whereas biological analysis of samples is **expensive** and **slow**.

Another answer is that analyzing the samples for harmful content does not provide a better understanding of algae bloom **drivers**: it just tells us which samples contain the harmful algaes.

Can our model provide a more thorough understanding of the algae situation?

EXERCISES

Locate and determine the structure of the algae bloom dataset, and provide a summary of its features.

Compute the number of missing values for each record.

Identify some potential anomalous observations in the same dataset.

What strategies could you use to deal with such observations / records?

Supplemental Material

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 for 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 for 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

OUTLIER TESTS

Normality is an assumption for most tests.

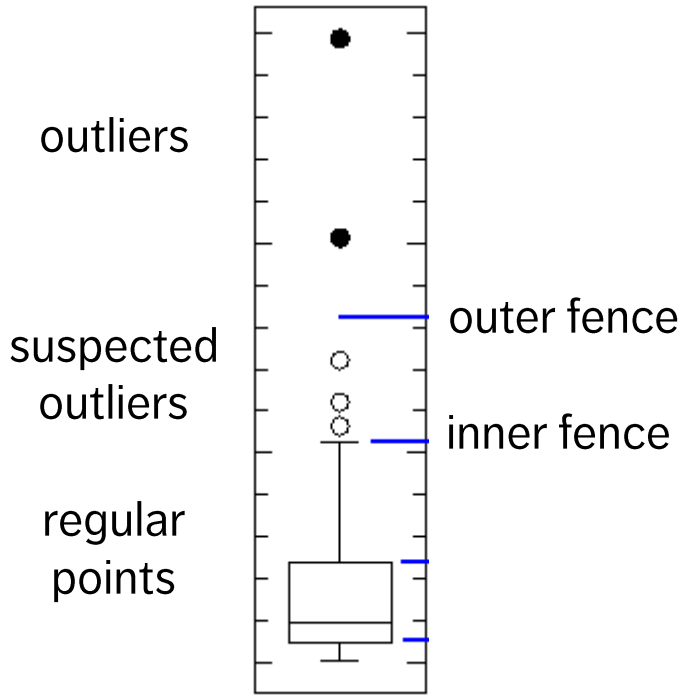
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

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

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

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

Outliers lie beyond the outer fences.



OUTLIER TESTS

The **Grubbs Test** is a univariate test. Consider

- x_i: value of feature X for i^{th} unit, $1 \leq i \leq N$
- $\bar{x}$: mean value of feature X
- s_x: standard deviation of feature X
- α: significance level
- $T(\alpha, N)$: value of the t-distribution at significance $\alpha/2N$

The i^{th} unit is an **outlier along feature** X if

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

OUTLIER TESTS

The **Dixon Q Test** is used in experimental sciences to find outliers in (extremely) small datasets (dubious validity).

The **Mahalanobis Distance** (linked to the leverage) can be used to find multi-dimensional outliers (when relationships are linear).

Other tests:

- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)
- **DBSCAN, OR_h** and **LOF** (unsupervised outlier detection)

IMPUTATION METHODS

List-wise deletion: remove units with at least one missing values.

- Assumption: MCAR
- Cons: can introduce bias (if not MCAR), reduction in sample size, increase in standard error

Mean or **Most Frequent Imputation:** substitute missing values by average value or most frequent value

- Assumption: MCAR
- Cons: distortions of distribution (spike at mean) and relationships among variables

IMPUTATION METHODS

Regression or **Correlation Imputation:** substitute missing values by using regression based on other variables (with complete information)

- Assumption: MAR
- Cons: artificial reduction in variability, over-estimation of correlation

Stochastic Regression Imputation: regression imputation with random error terms added

- Assumption: MAR
- Cons: increased risk of type I error (false positives) due to small std error

IMPUTATION METHODS

Last Observation Carried Forward (LOCF): substitute the missing values with previous values (in a longitudinal study)

- Assumption: MCAR, values do not vary greatly over time
- Cons: may be too "generous", depending on the nature of study

k-Nearest-Neighbour Imputation (kNN): substitute the missing entry with the average from the group of the k most **similar** complete respondents

- Assumption: MAR
- Cons: difficult to choose appropriate value for k. Possible distortion in data structure ($k > 1$)