DATA PROCESSING & DATA CLEANING

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
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/iconsistentl y 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

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

MISSING VALUE 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

etc.

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

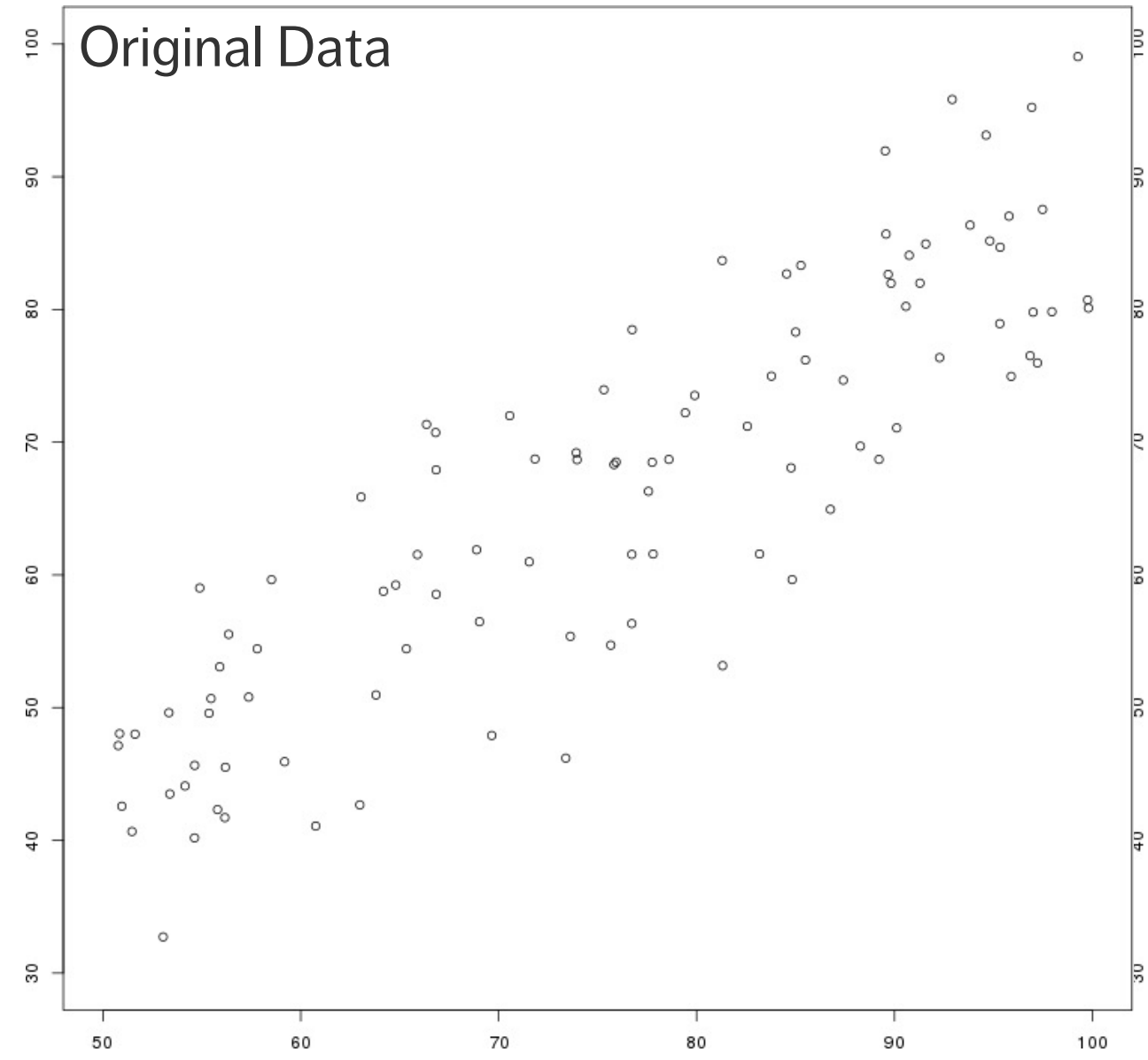
Last Observation Carried Forward (LOCF): substitute the missing values with previous values (in a longitudinal study)

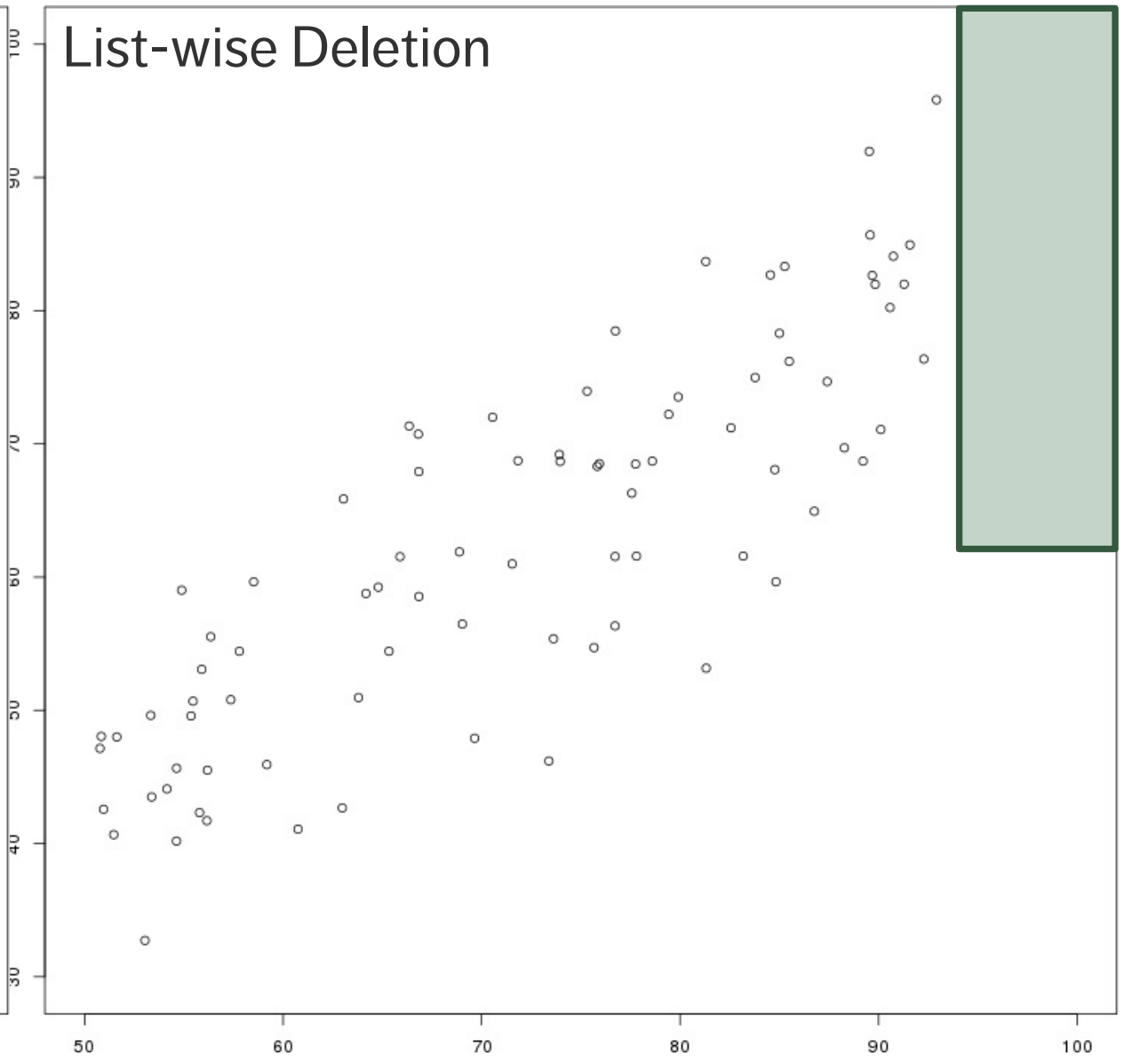
- Assumption: MCAR, values do not vary greatly over time
- Cons: may be too "generous", depending on the nature of study

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

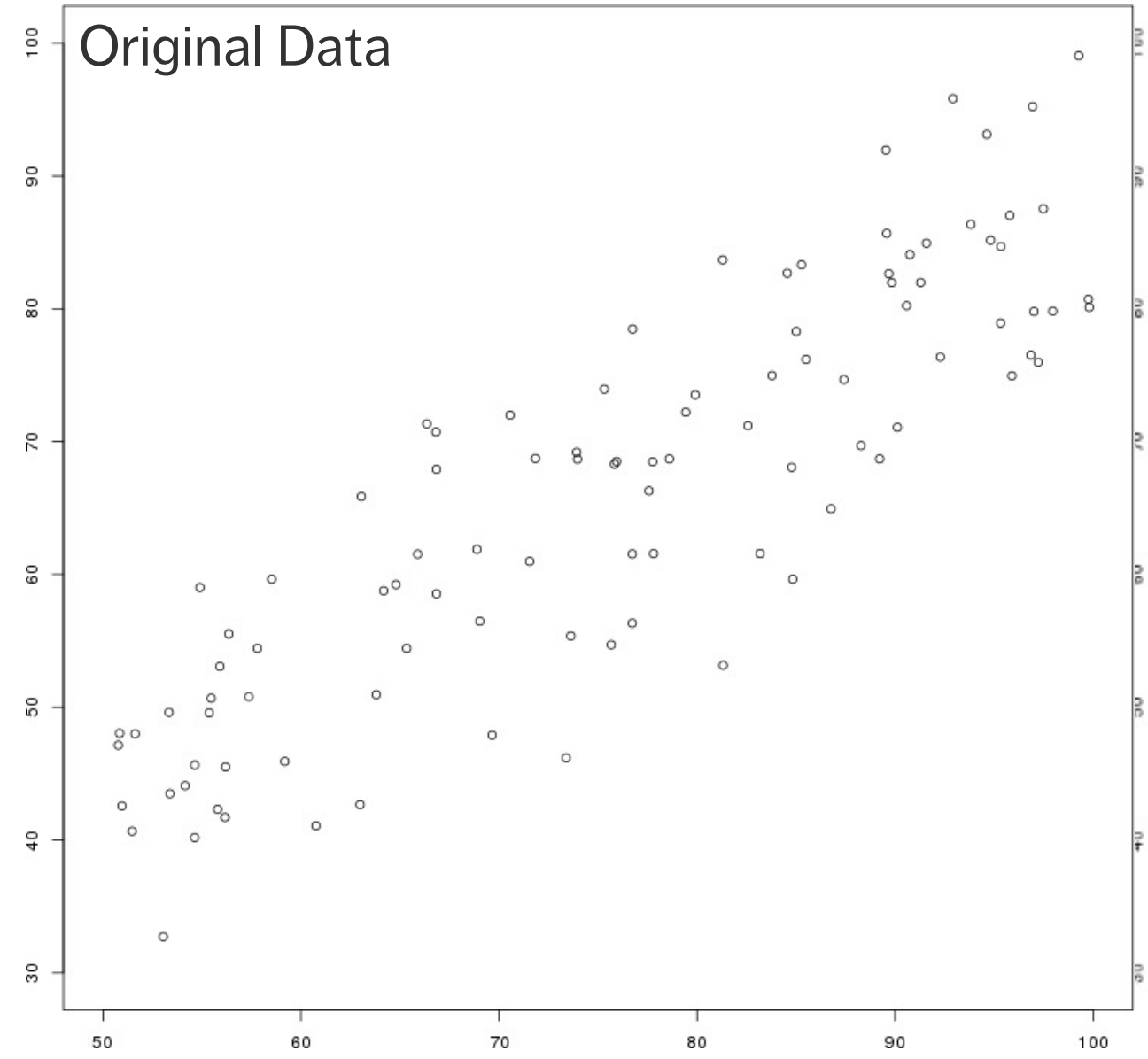
- Assumption: MAR
- Cons: difficult to choose appropriate value for k. Possible distortion in data structure ($\mathrm{k} > 1$)

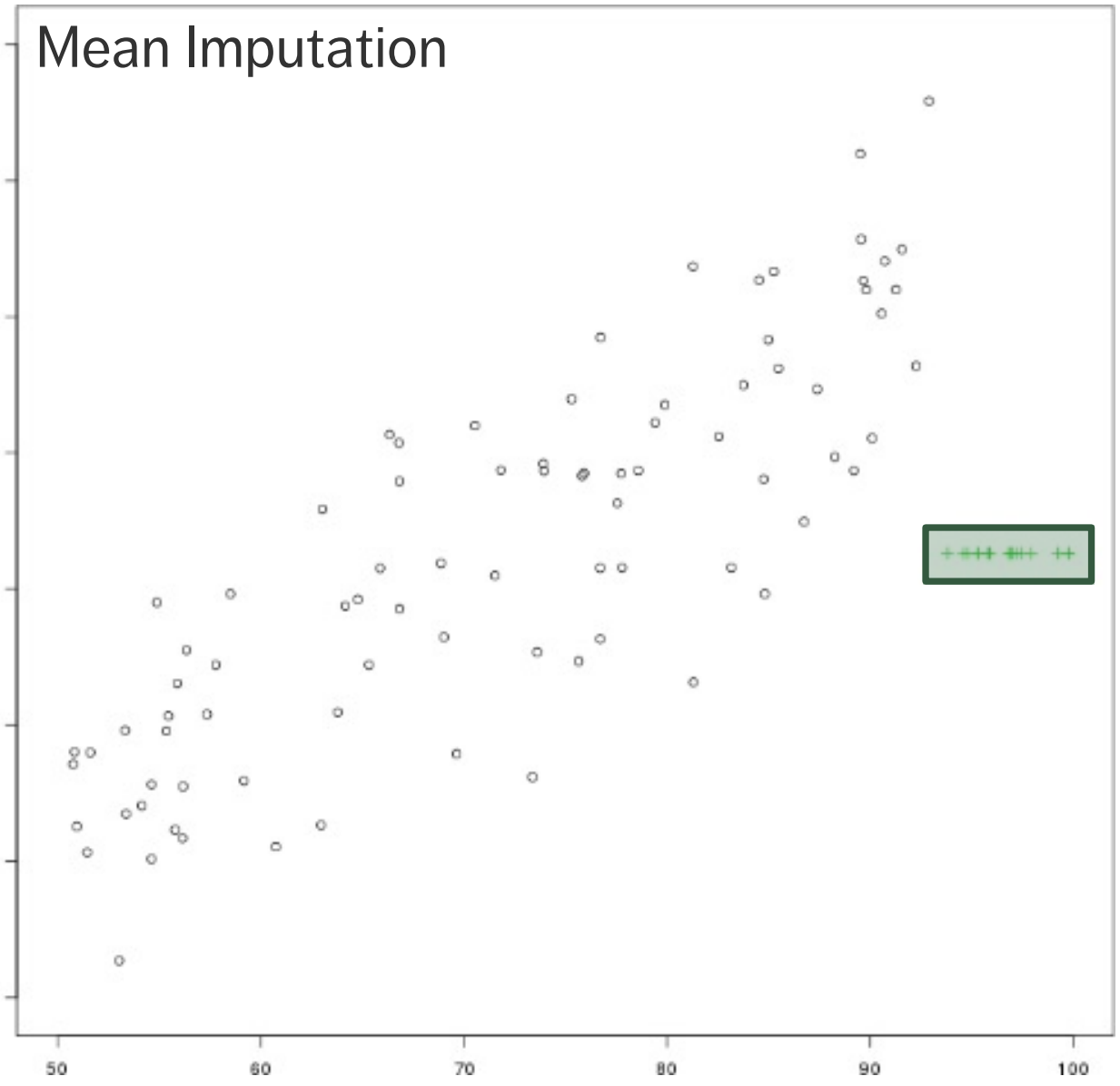
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



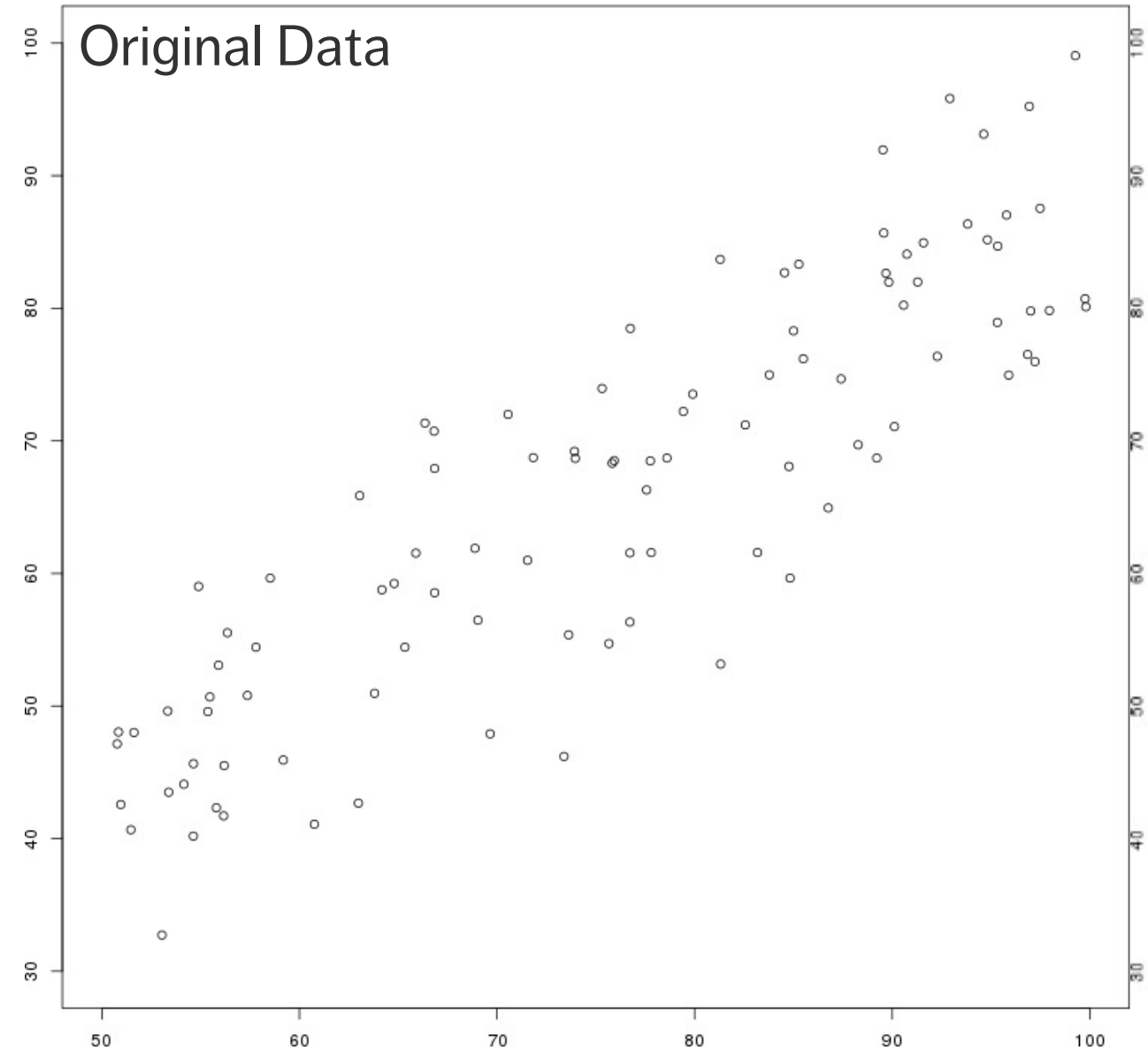


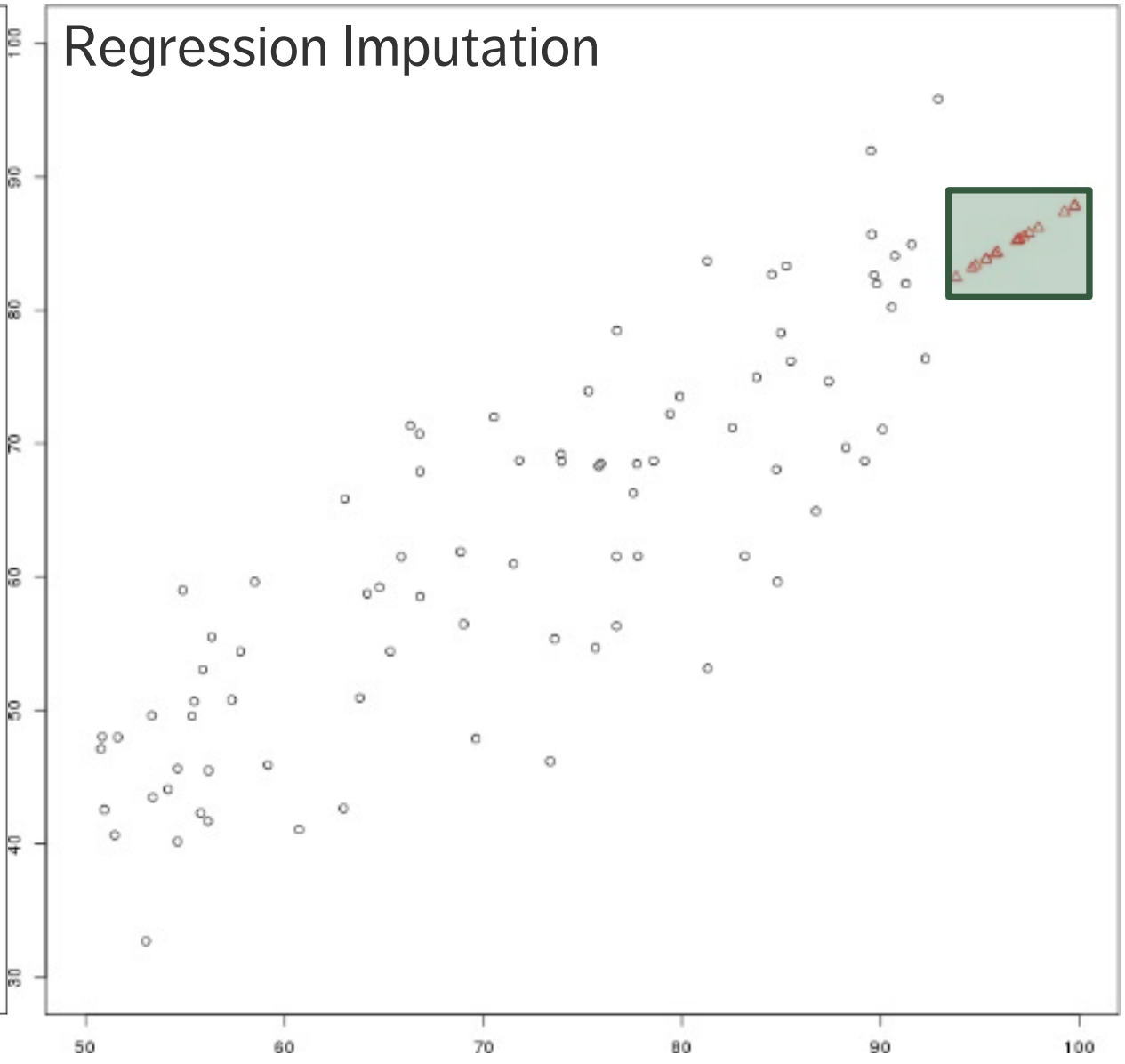
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



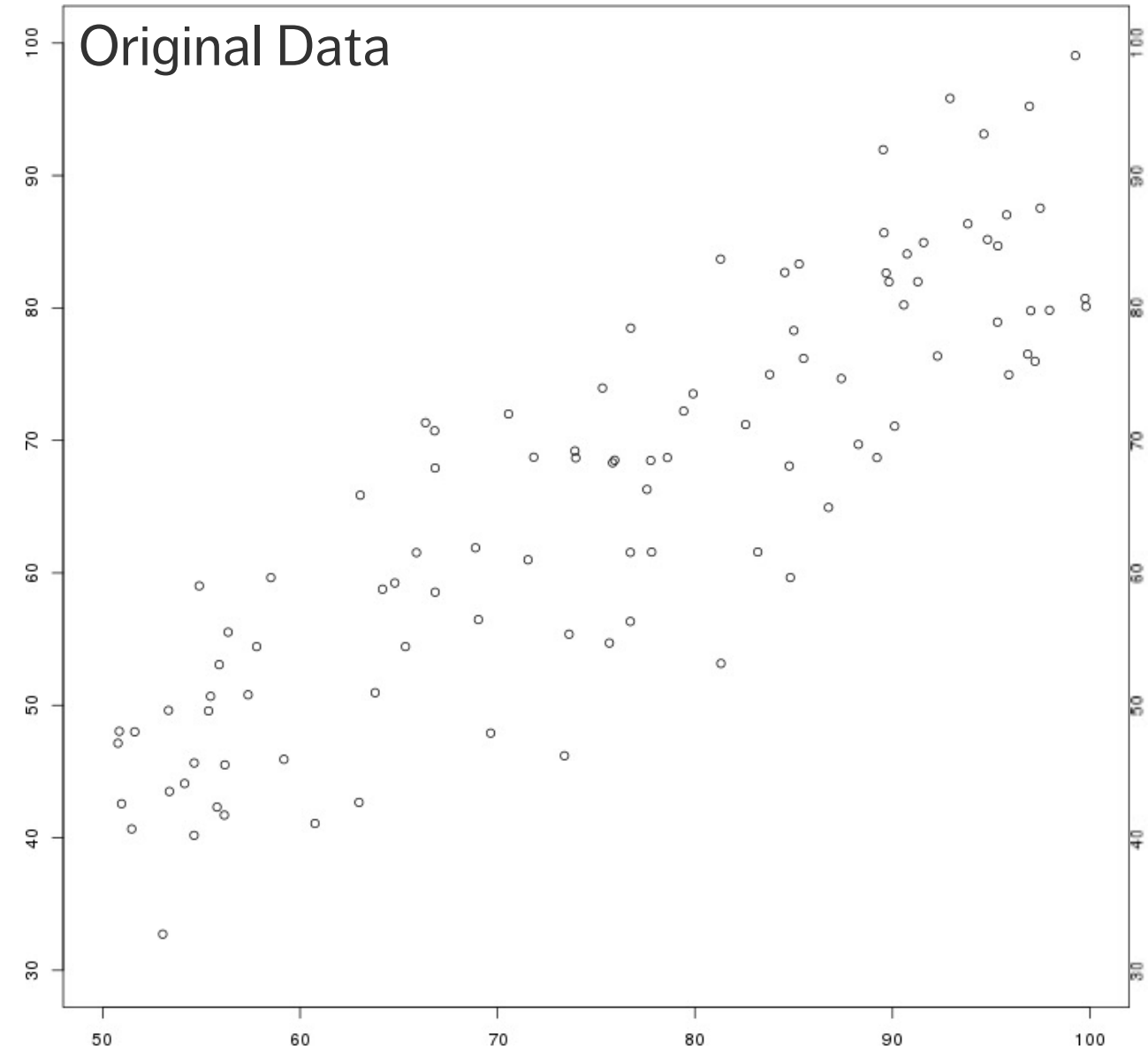


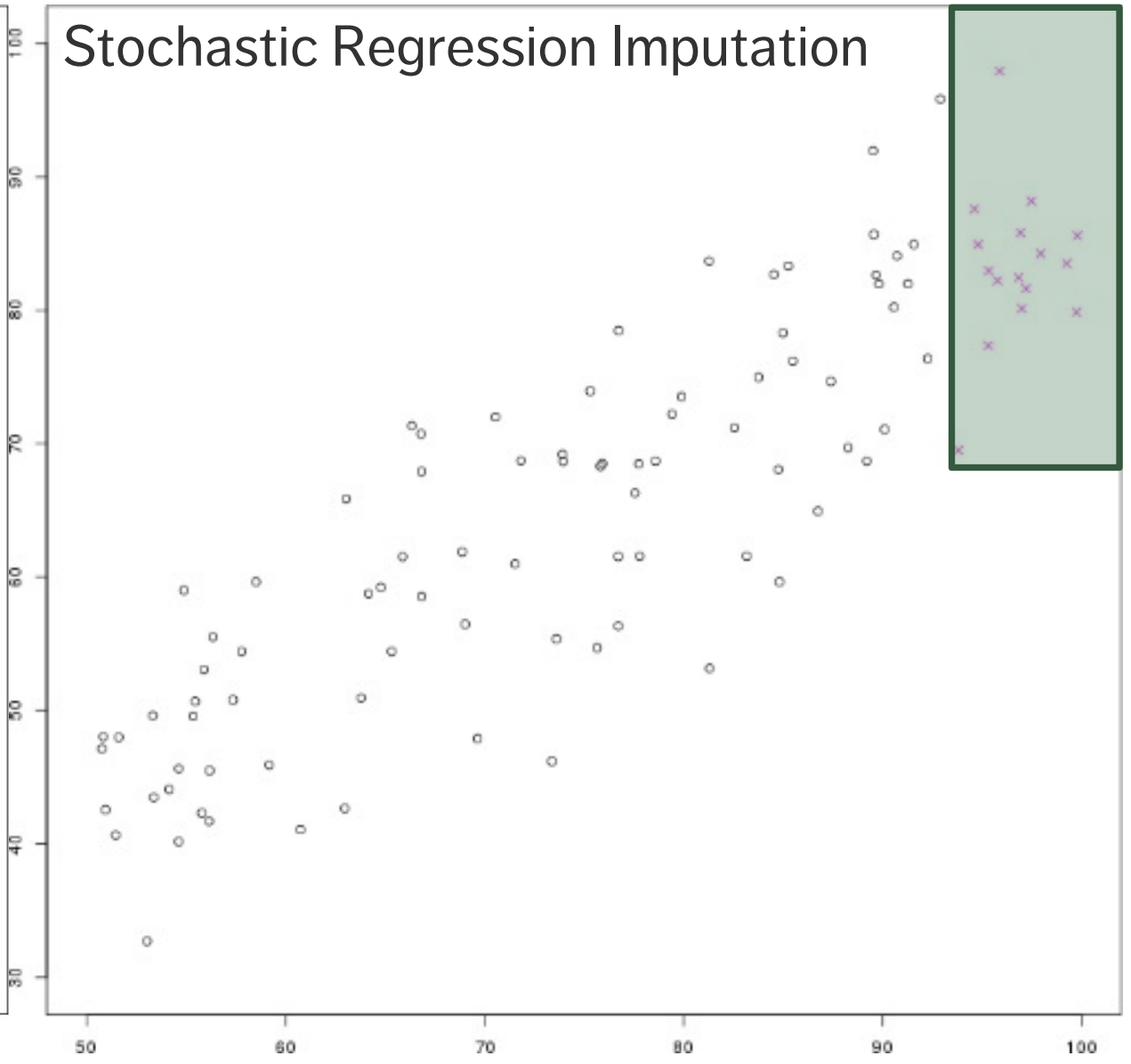
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





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** (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.

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.

OUTLIER TESTS

Normality is an assumption for most tests.

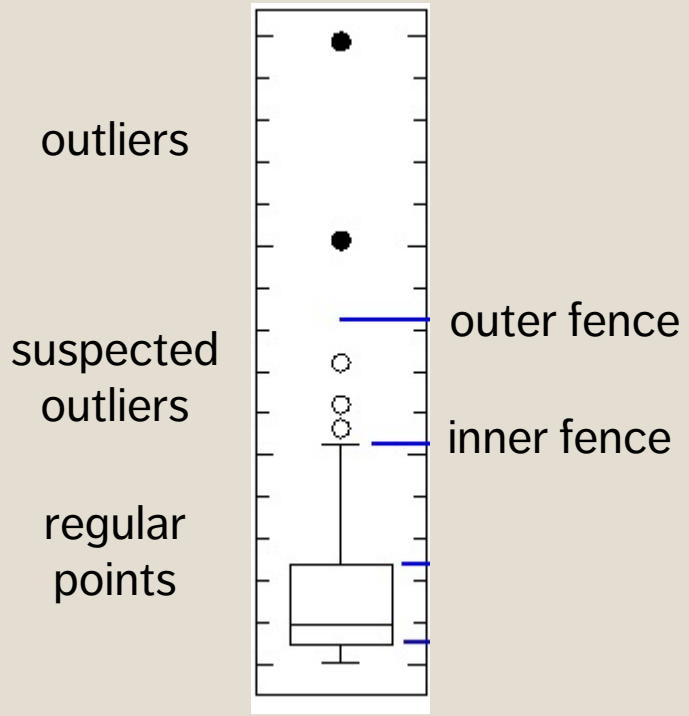
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

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

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

Grubbs test used to detect a single outlier

Dixon Q test used to find outliers in (extremely) small datasets (**dubious validity**)

Mahalanobis distance can be used to find multi-dimensional outliers

Tietjen-Moore test for a specific # of outliers

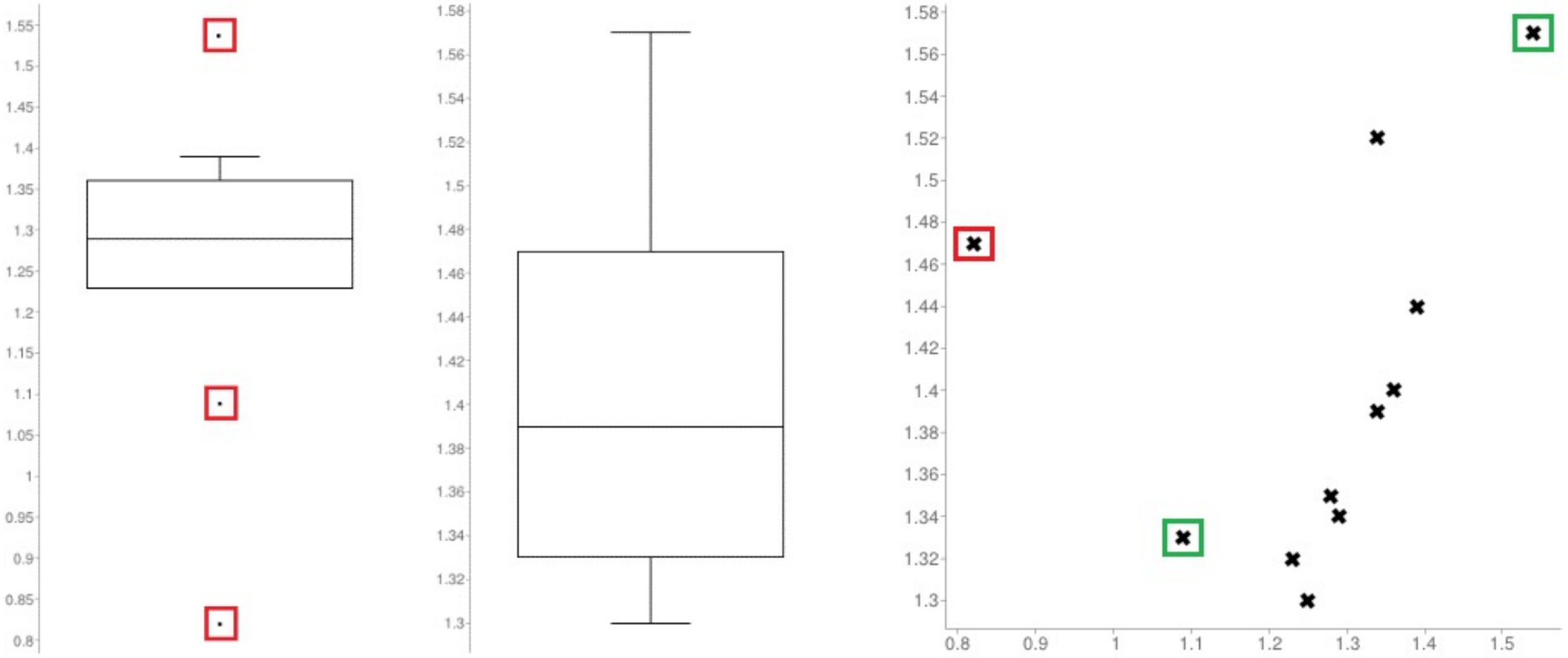
Generalized extreme studentized deviate test for unknown # of outliers

Chi-square test for outliers affecting goodness-of-fit

DBSCAN, OR_h and **LOF** for unsupervised outlier detection

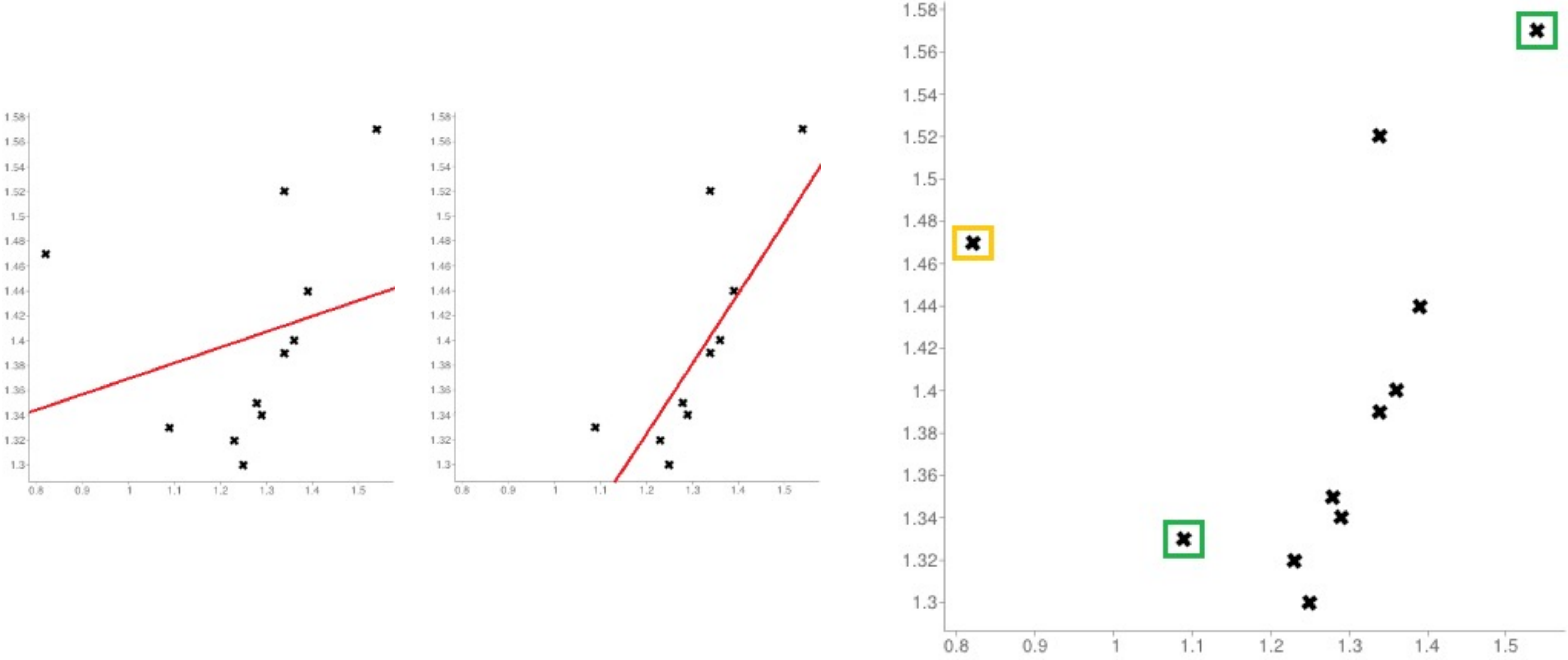
[Personal file]

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.

EXERCISE

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?

EXERCISE

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

Another answer: 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 algae.

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

EXERCISE

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?

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 ≫ # 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 multi-colinearity, 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.

SUPERVISED FILTER METHODS

Correlation between a feature X and a target variable Y (features which are highly correlated with the target variable are retained, but this approach is limited if the relationship to the target variable is **non-linear)**.

Mutual Information of nominal target Y from nominal feature X (same approach).

Classification Tasks

- Gain Ratio, Inf Gain, Gini, MDL, etc.

Regression Tasks

- MSE of Mean, MAE of Mean, Relief (evaluates features simultaneously), etc.

COMMON TRANSFORMATIONS

Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

If the raw data does not meet the requirements, we can either

- abandon the model
- attempt to **transform** the data

The second approach requires an inverse transformation to be able to draw conclusions about the original data.

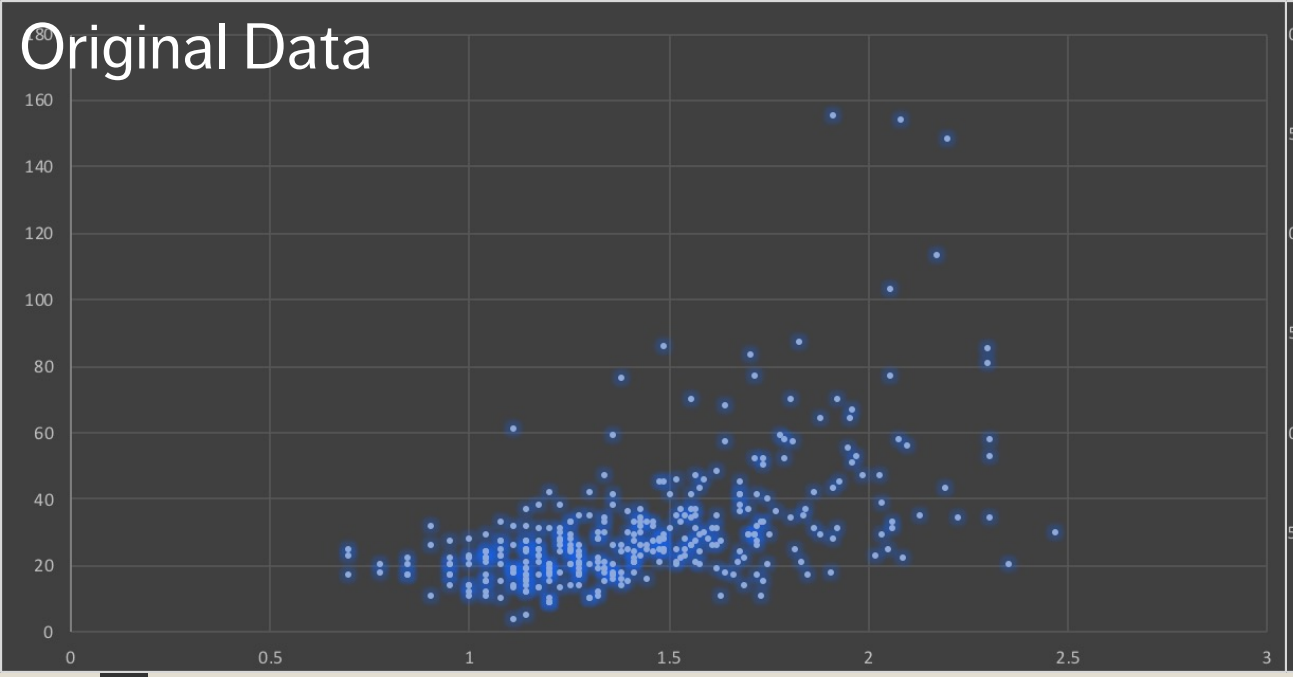
COMMON TRANSFORMATIONS

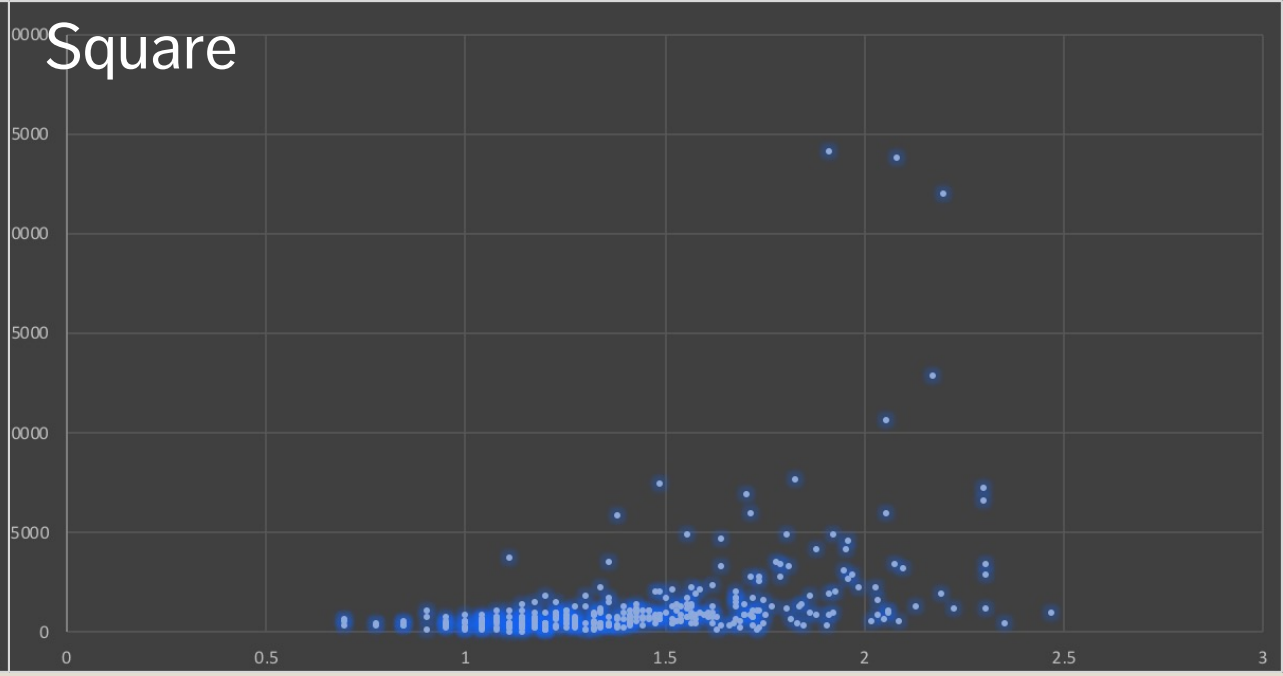
In the regression context, transformations are **monotonic:**

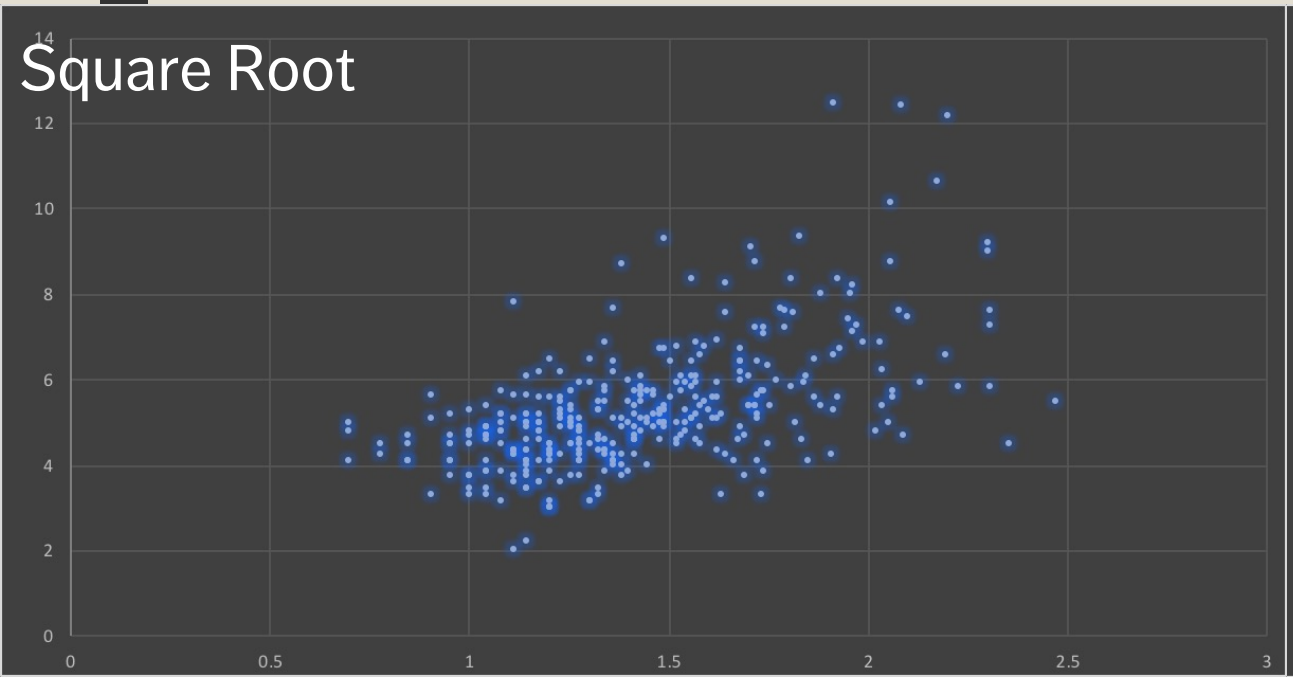
- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

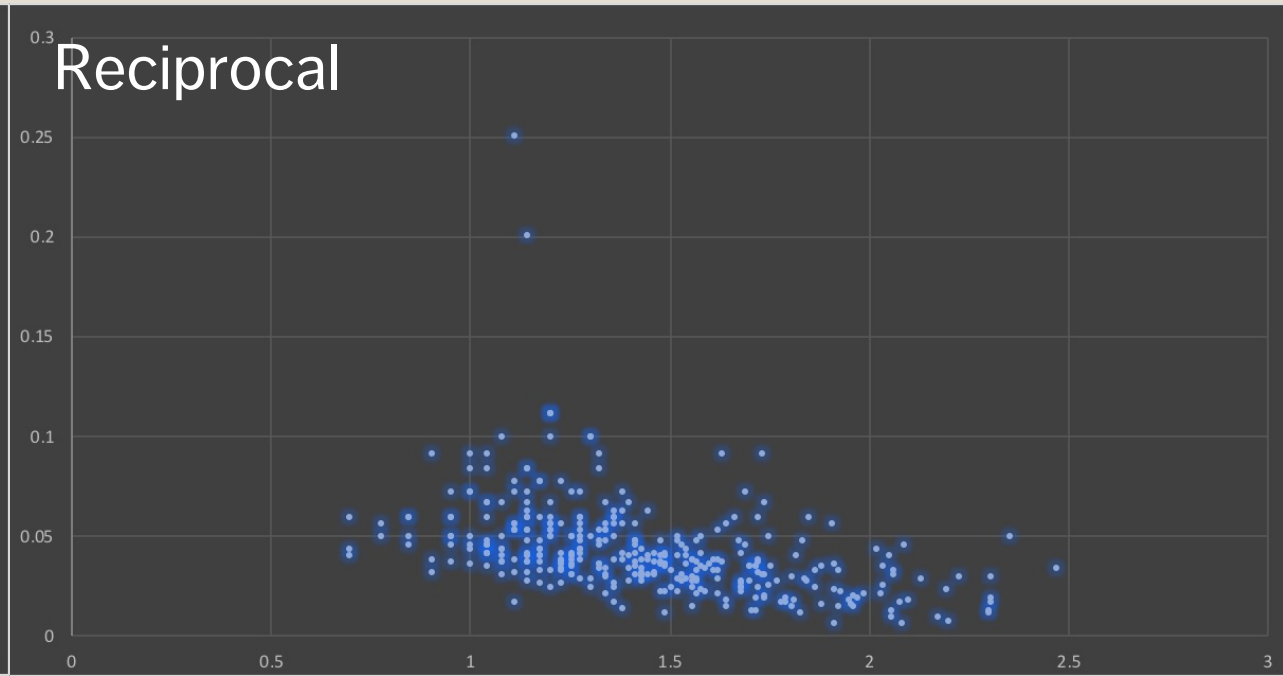
Transformations on X may achieve linearity, but usually at some price (correlations are not preserved, for instance).

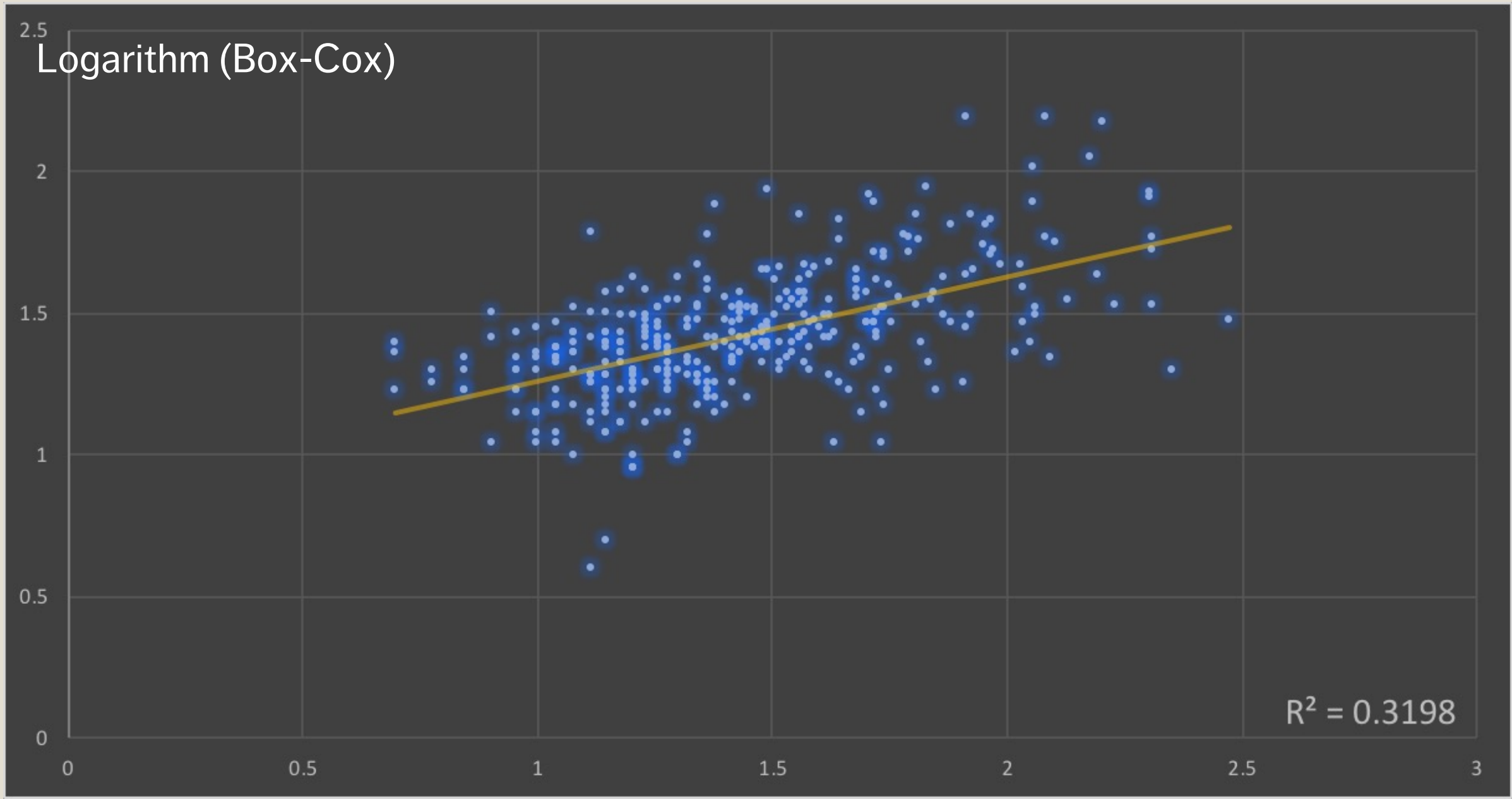
Transformations on Y can help with non-normality and unequal variance of error terms.


Original Data
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Square
5000
0000
5000
0000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3


Logarithm (Box-Cox)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
R² = 0.3198

SCALING

Numeric variables may have different **scales** (weights and heights, for instance).

The variance of a large-range variable is typically greater than that of a small-range variable, introducing a bias (for instance).

Standardization creates a variable with mean 0 and std. dev. 1:

$$Y_i = \frac{X_i - \bar{X}}{s_X}$$

Normalization creates a new variable in the range $[0,1]$: $Y_i = \frac{X_i - \min X}{\max X - \min X}$

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

CREATING VARIABLES

New variables may need to be introduced:

- as **functional relationships** of some subset of available features
- because modeling tool may require **independence of observations**
- because modeling tool may require **independence of features**
- to simplify the analysis by looking at **aggregated summaries** (often used in text analysis)

Time dependencies ⟶ time series analysis

Spatial dependencies ⟶ spatial analysis

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.

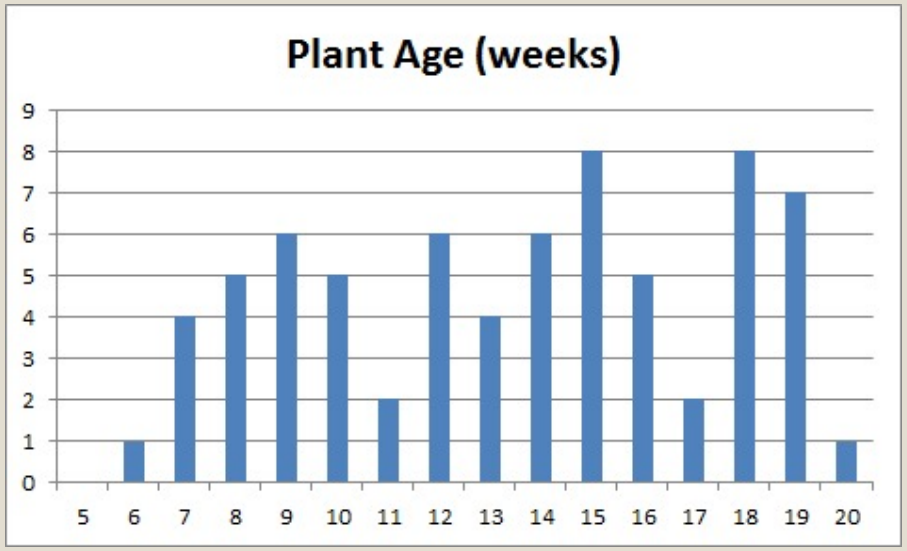
This step might allow for the identification of potential outliers.

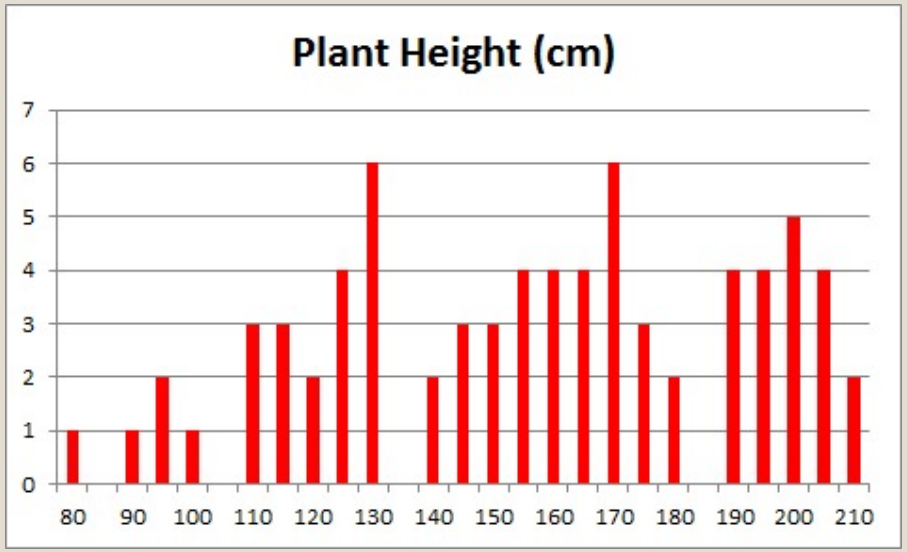
Failure to detect invalid entries **does not mean** that all entries are valid.

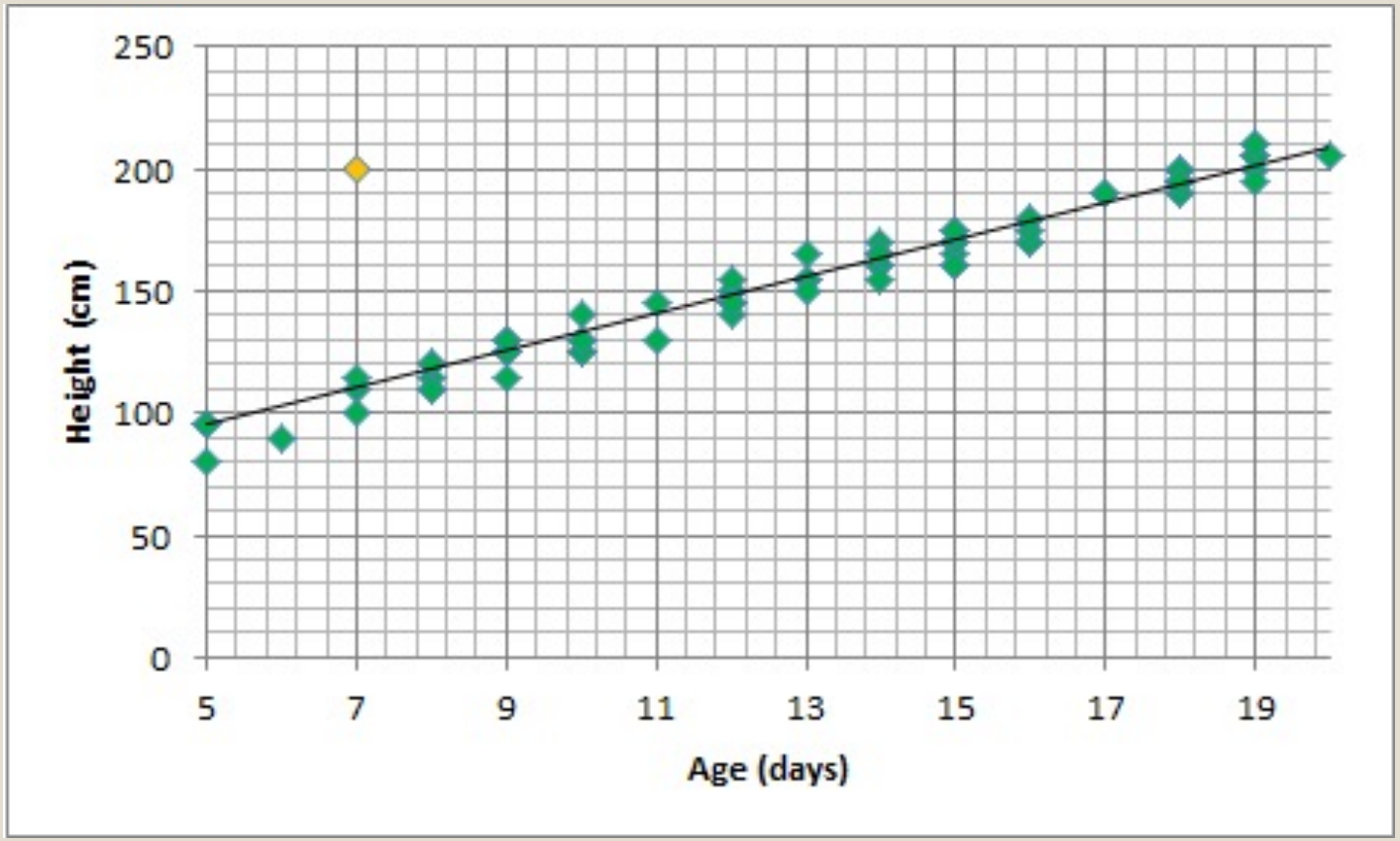
Small numbers of invalid entries recoded as “missing.”

[Personal file]

ILLUSTRATION

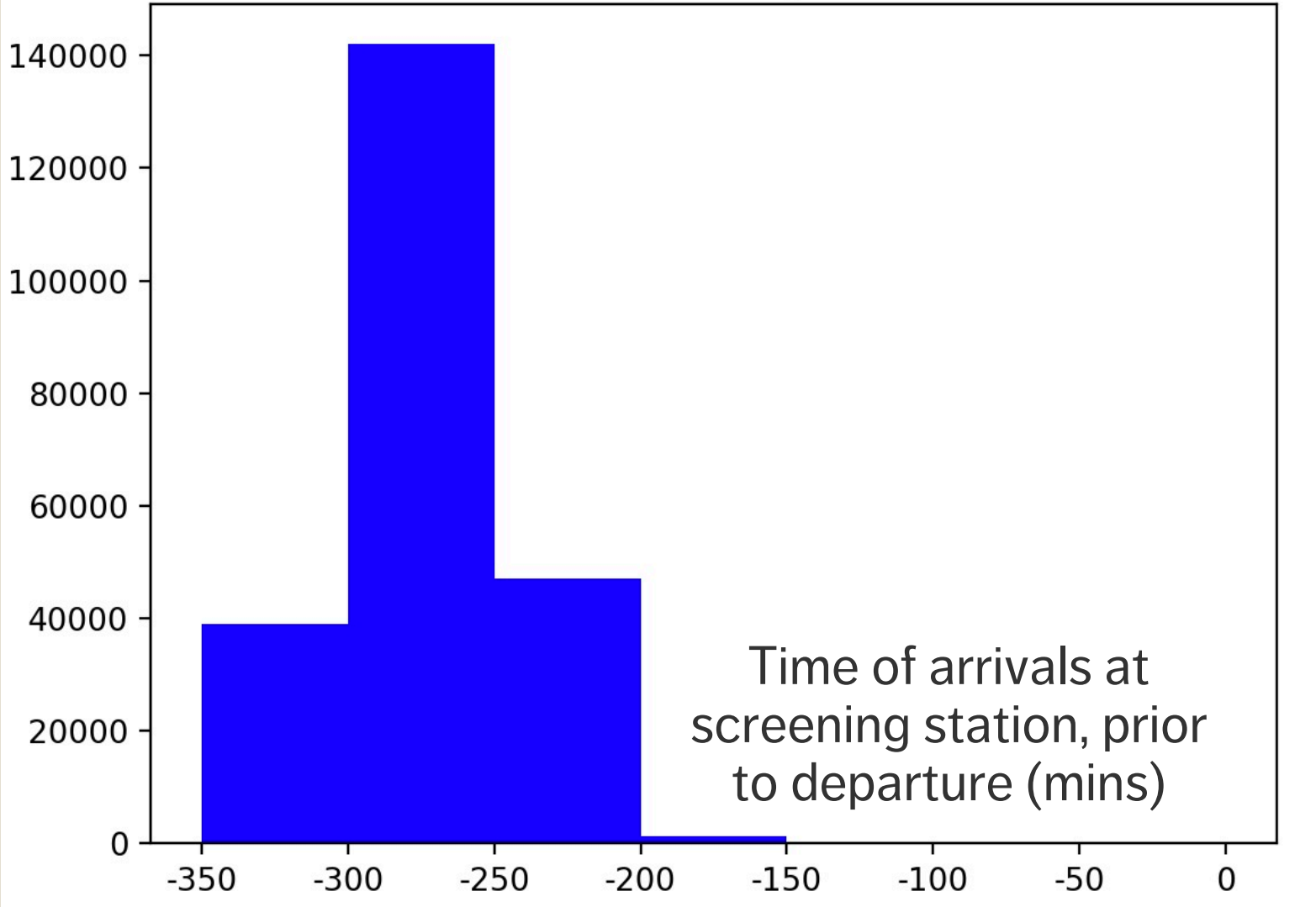






[Personal file]

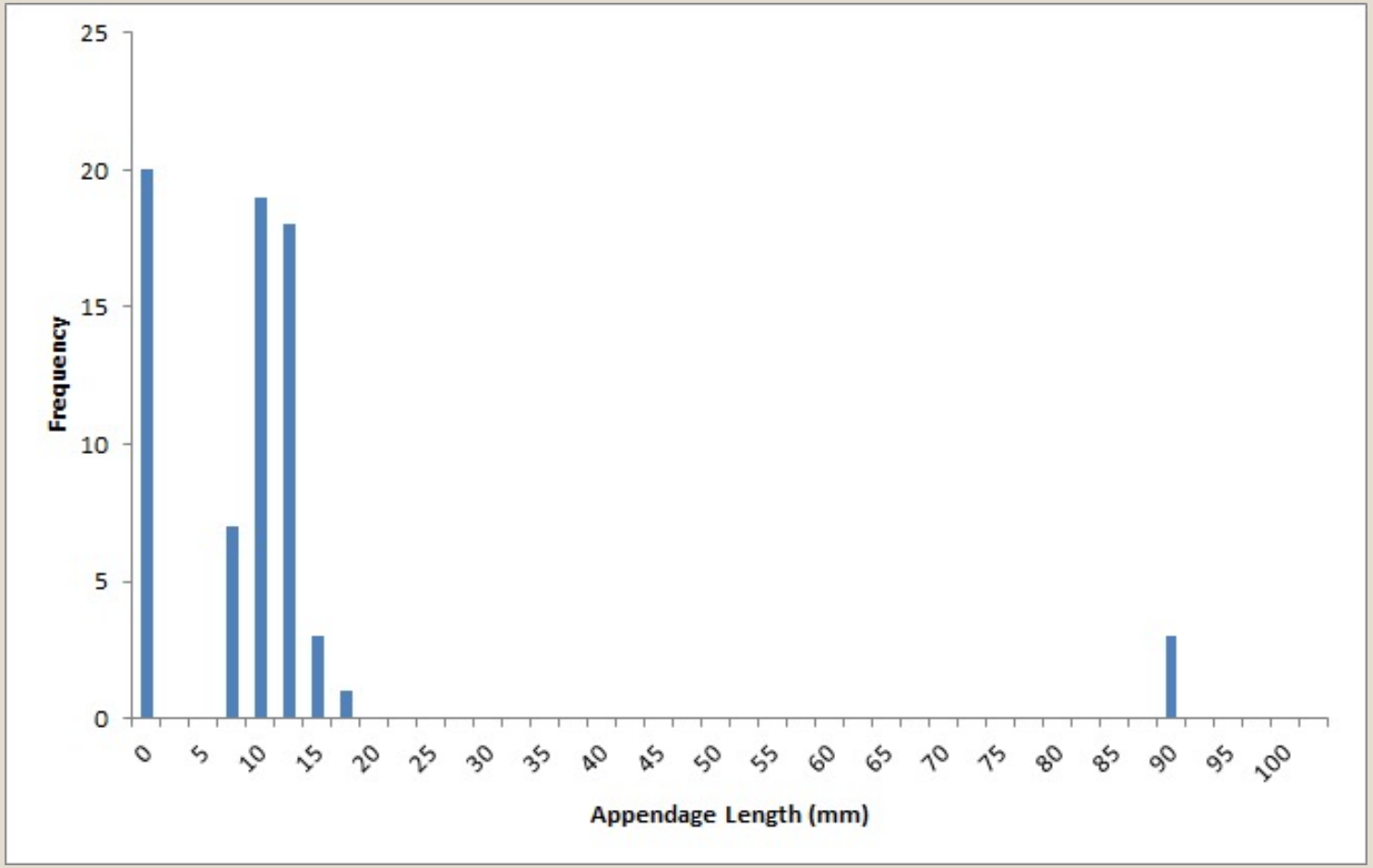
ILLUSTRATION



[Personal file]

ILLUSTRATION

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



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.