Data Preparation

DATA SCIENCE ESSENTIALS

Session 3

7. Data Quality and Data Wrangling

The Hot Mess

"Data is messy, you know."
"Even after it's been cleaned?"
"*Especially* after it's been cleaned."

Data **cleaning**, **processing**, **wrangling** are essential aspects of data science projects; analysts may spend **up to 80%** of their time on **data preparation**.

Data Wrangling and Tidy Data

Tidy data has a specific structure:

- each variable is in a single column
- each observation is in a single row
- each type of observational unit is in a single table

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

VS.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

Data Wrangling Functionality

Data wrangling functions should allow the analyst to:

- extract a subset of variables from the data frame
- extract a subset of observations from the data frame
- sort the data frame along any combination of variables in increasing or decreasing order
- to create new variables from existing variables
- to create (so-called) pivot tables, by observation groups
- database functionality (joins, etc.)
- etc.

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.

Approaches to Data Cleaning

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

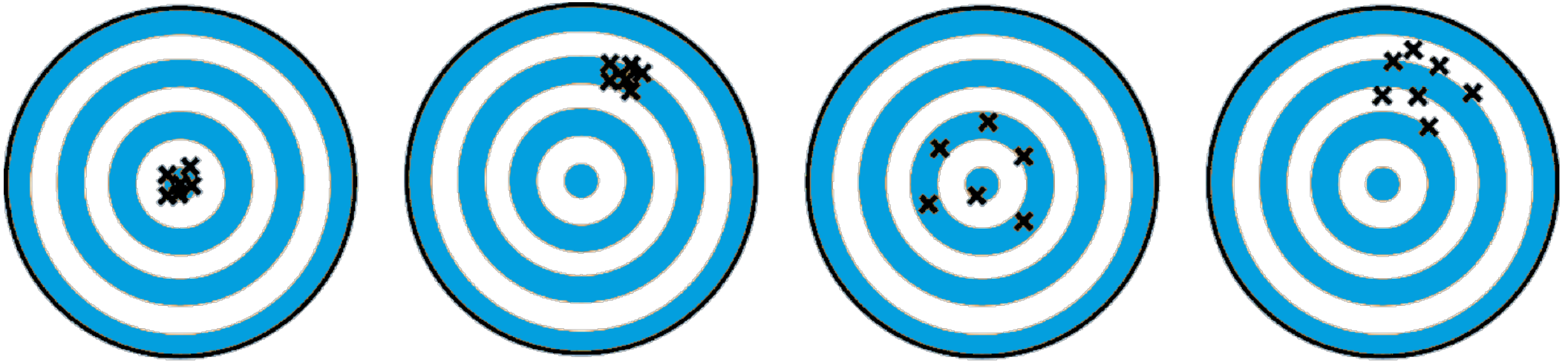
Data Soundness

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 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.

Data Soundness



accurate and precise

precise but not accurate

accurate but not precise

neither accurate nor very precise

Common Error Sources

When dealing with **legacy**, **inherited** or **combined** datasets (that is, datasets over which there is no collection and initial processing control):

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

[Bill Watterson, *Calvin and Hobbes*, 15 Sep. 1990]

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.

Detecting Invalid Entries

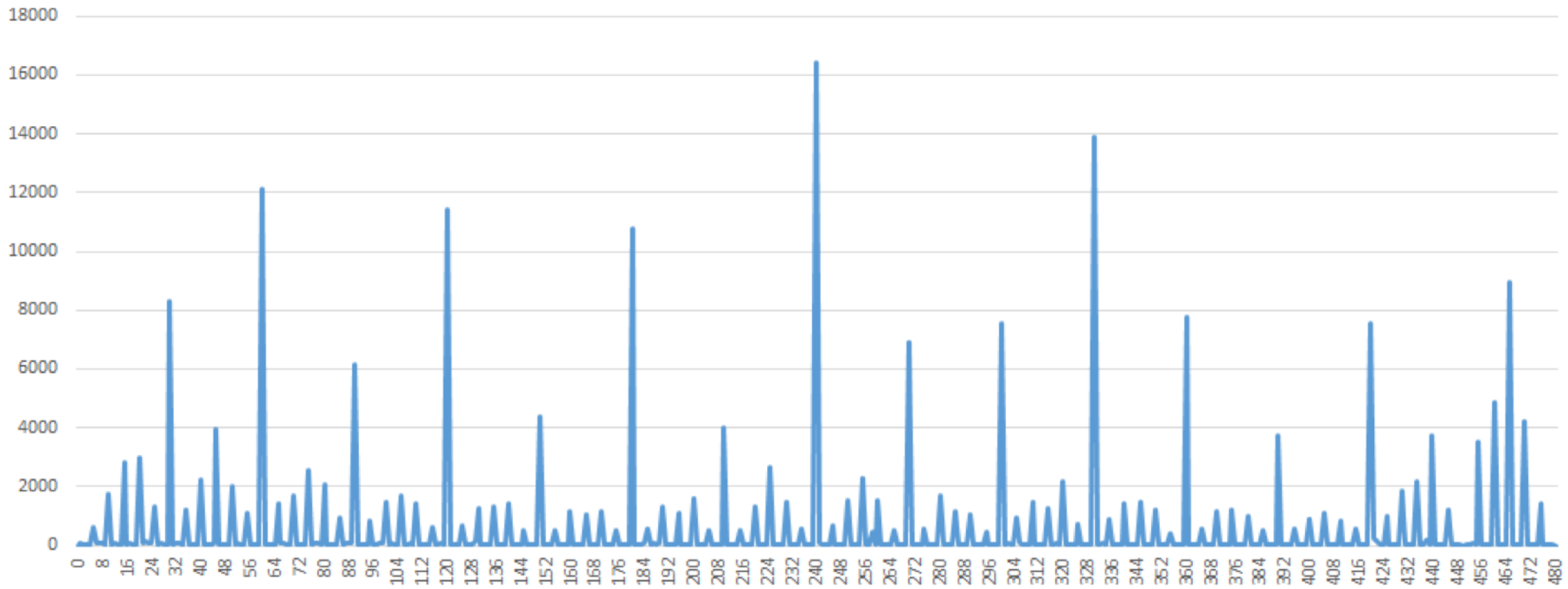
Univariate tests do not always tell the **whole** story.

This step might allow for the identification of potential outliers.

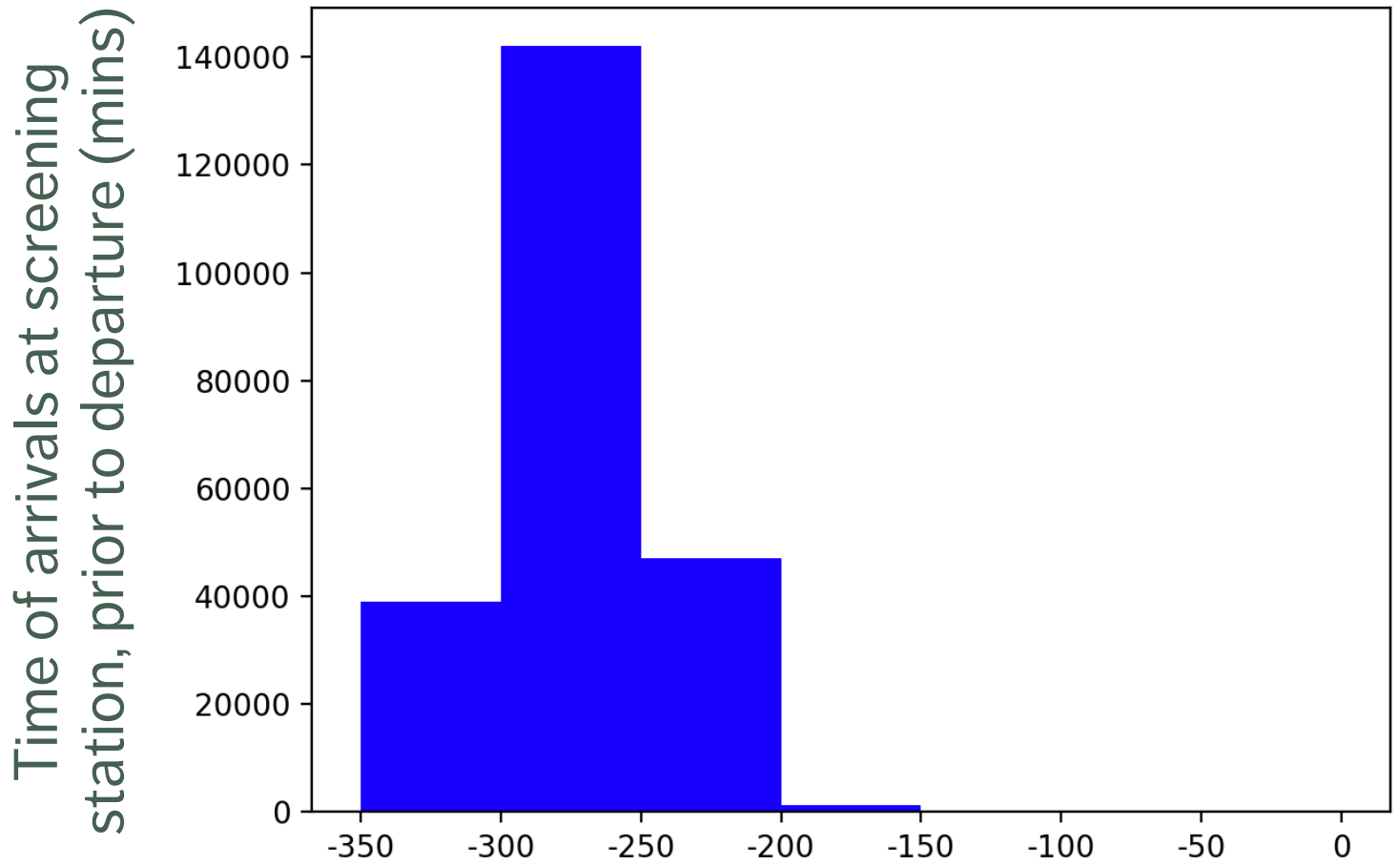
Failure to detect invalid entries ≠ all entries are valid.

Small numbers of invalid entries recoded as “missing.”

Detecting Invalid Entries



Detecting Invalid Entries



Detecting Invalid Entries

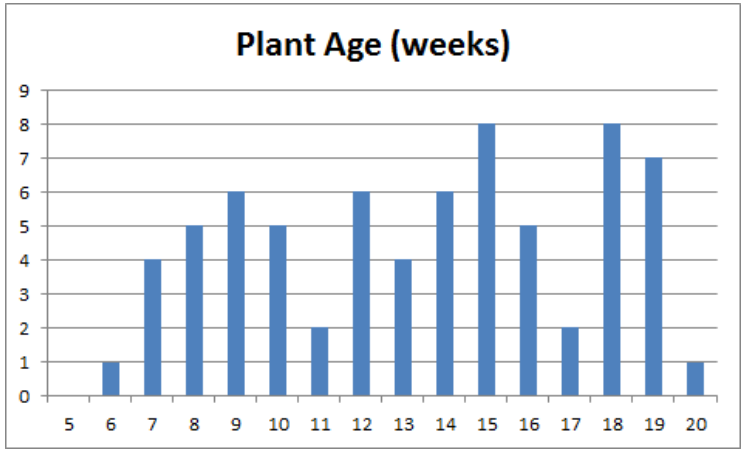
Sex	
Male	19
Female	17
(blank)	2
Total	**38**

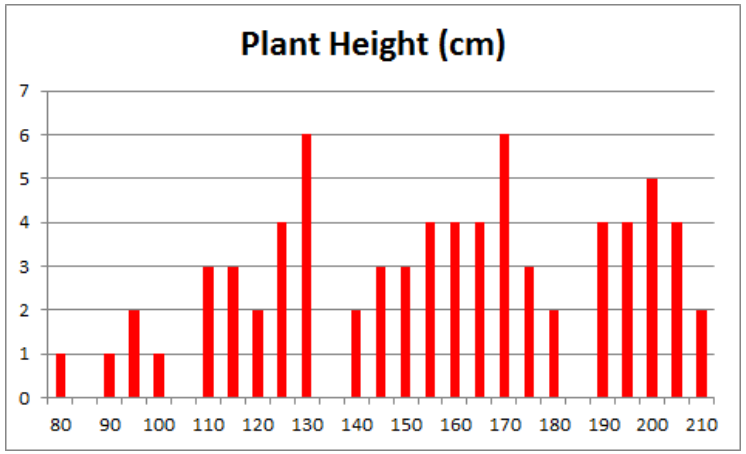
Pregnant	
Yes	7
No	27
99	1
(blank)	3
Total	**38**

vs.

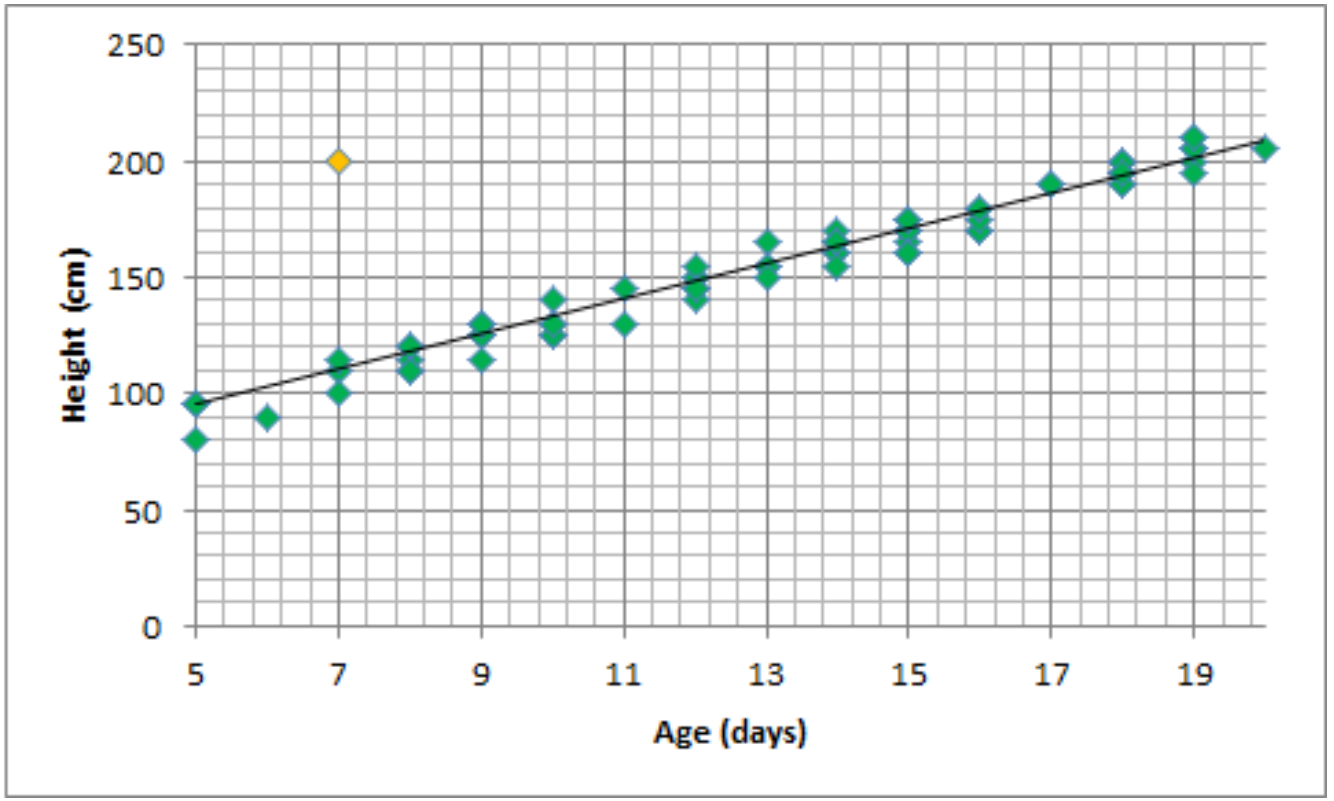
	Pregnant				Total
	Yes	No	99	(blank)	
Male	1	17	1	0	**19**
Female	6	9	0	2	**17**
(blank)	0	1	0	1	**2**
Total	**7**	**27**	**1**	**3**	**38**

Detecting Invalid Entries



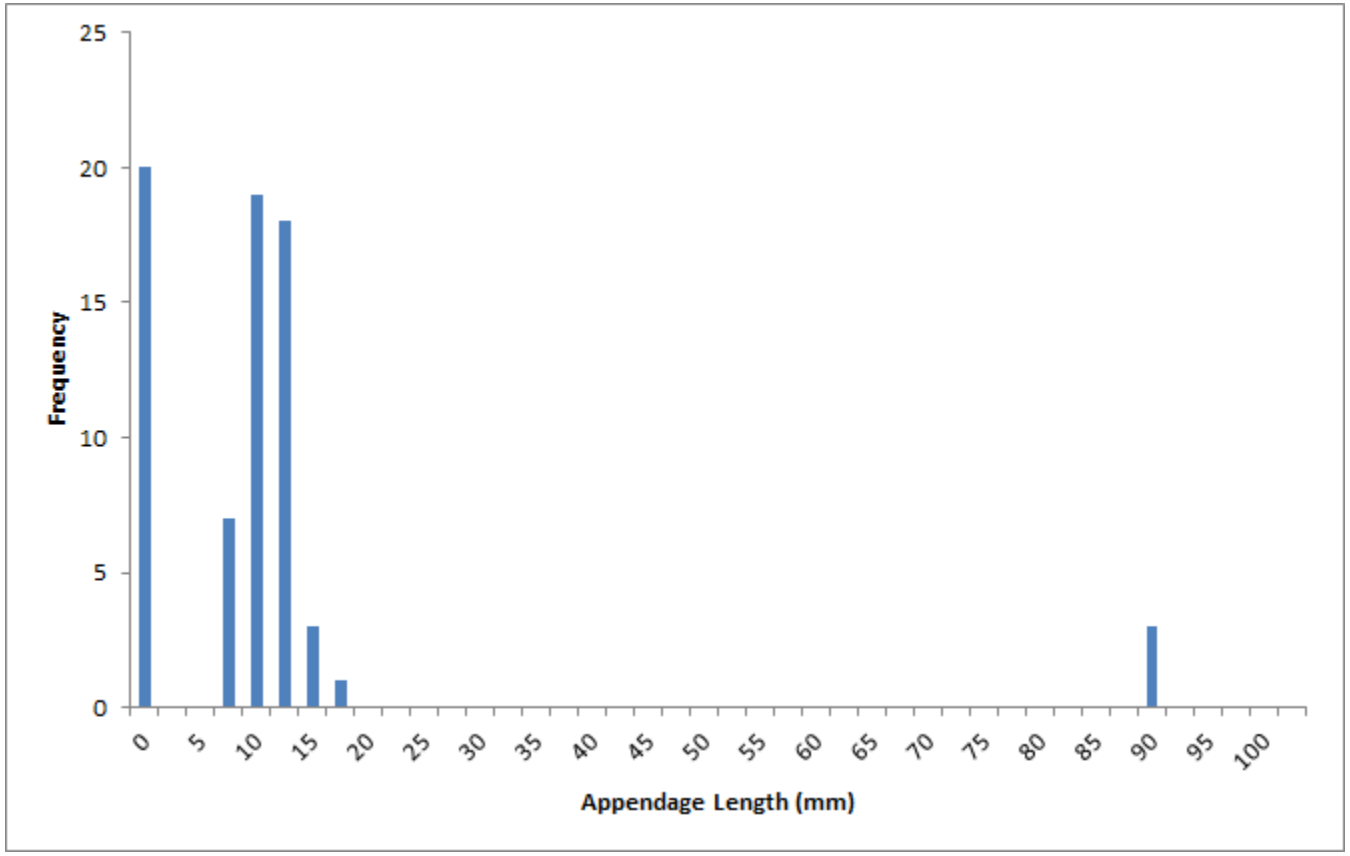


vs.



Detecting Invalid Entries

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



Suggested Reading

Data Quality

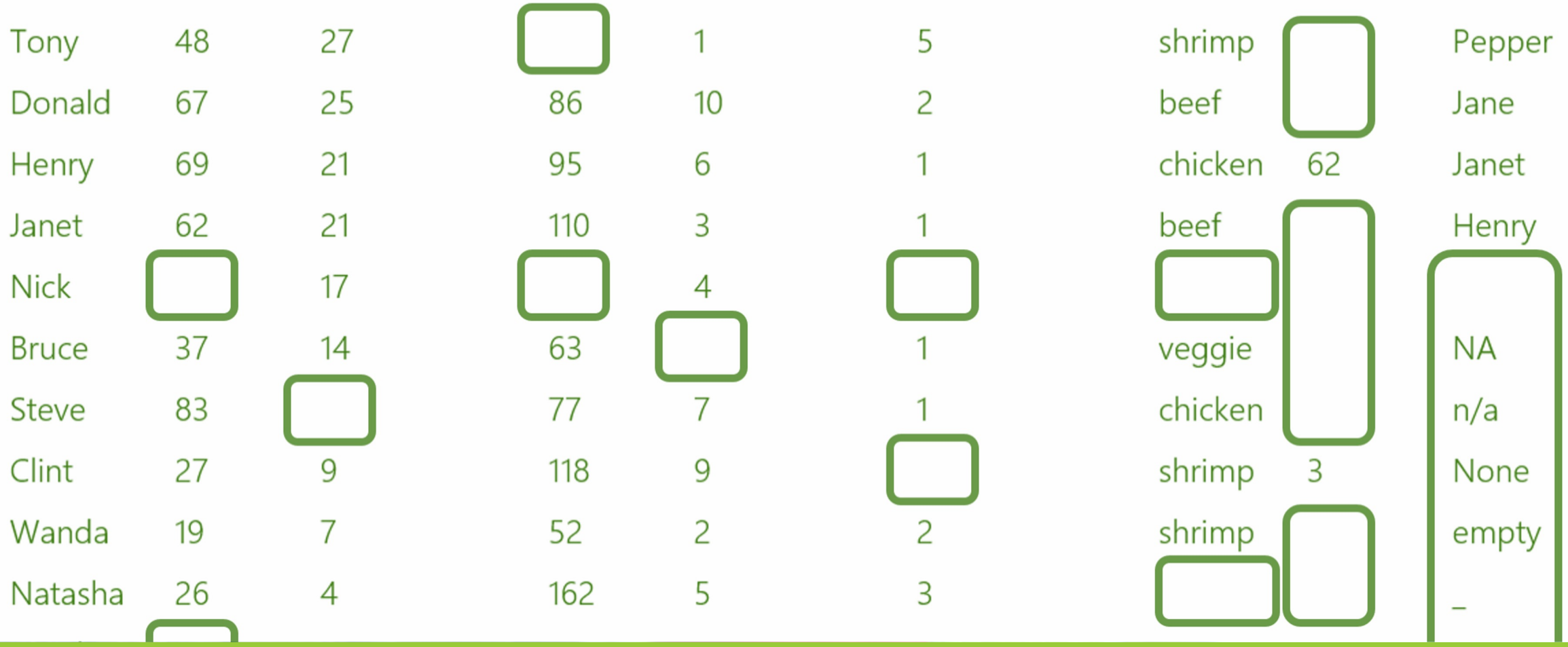
Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    - 15.1 Introduction
    - 15.2 General Principles
        - Approaches to Data Cleaning
        - Pros and Cons
        - Tools and Methods
    - 15.3 Data Quality
        - Common Error Sources
        - Detecting Invalid Entries

Exercises

Data Quality

1. Recreate the examples of The Tidyverse.
2. Turn the dataset found in the file cities.txt into a tidy dataset.
3. Does the dataset found in the file cities.txt appear to be of good quality (is it sound? does it have invalid entries?)
4. Create a list of items that could be used in a methodical data cleaning checklist. Use data that you have encountered in the past as inspiration (numerical, categorical, text data).



8. Missing Values

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**.

Finding missing values can help you deal with other data science problems.

The Case for Imputation

Not all analytical methods can easily accommodate missing observations:

- **discard** the missing observation
  - not recommended, unless the data is MCAR in the dataset as a whole
  - acceptable in certain situations (e.g., 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 Values Mechanism

Missing Completely at Random (MCAR)

- item absence is independent of its value or of auxiliary variables
- **example:** an electrical surge randomly deletes an observation in the dataset

Missing at Random (MAR)

- item absence is not completely random; can be accounted by auxiliary variables with complete info
- **example:** if women are less likely to tell you their age than men for societal reasons, but not because of the age values themselves)

Missing Values Mechanism

Not Missing at Random (NMAR)

- reason for nonresponse is related to item value (also called **non-ignorable non-response**)
- **example:** if illicit drug users are less likely to admit to drug use than teetotallers

In general, the missing mechanism **cannot be determined** with any certainty; we may need to make assumptions (domain expertise can help).

Imputation Methods

- list-wise deletion
- mean or most frequent imputation
- regression or correlation imputation
- stochastic regression imputation
- last observation carried forward
- next observation carried backward
- 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/most frequent imputation: substitute missing values by average/most frequent value

- **assumption:** MCAR
- **cons:** distortions of distribution (spike at mean) and relationships among variables

Imputation Methods

Regression/correlation imputation: substitute missing values using fitted values based on other variables with complete information

- **assumption:** MAR
- **cons:** artificial reduction in variability, over-estimation of correlation

Stochastic regression imputation: regression/correlation imputation with a random error term added

- **assumption:** MAR
- **cons:** increased risk of type I error (false positives) due to small std error

Imputation Methods

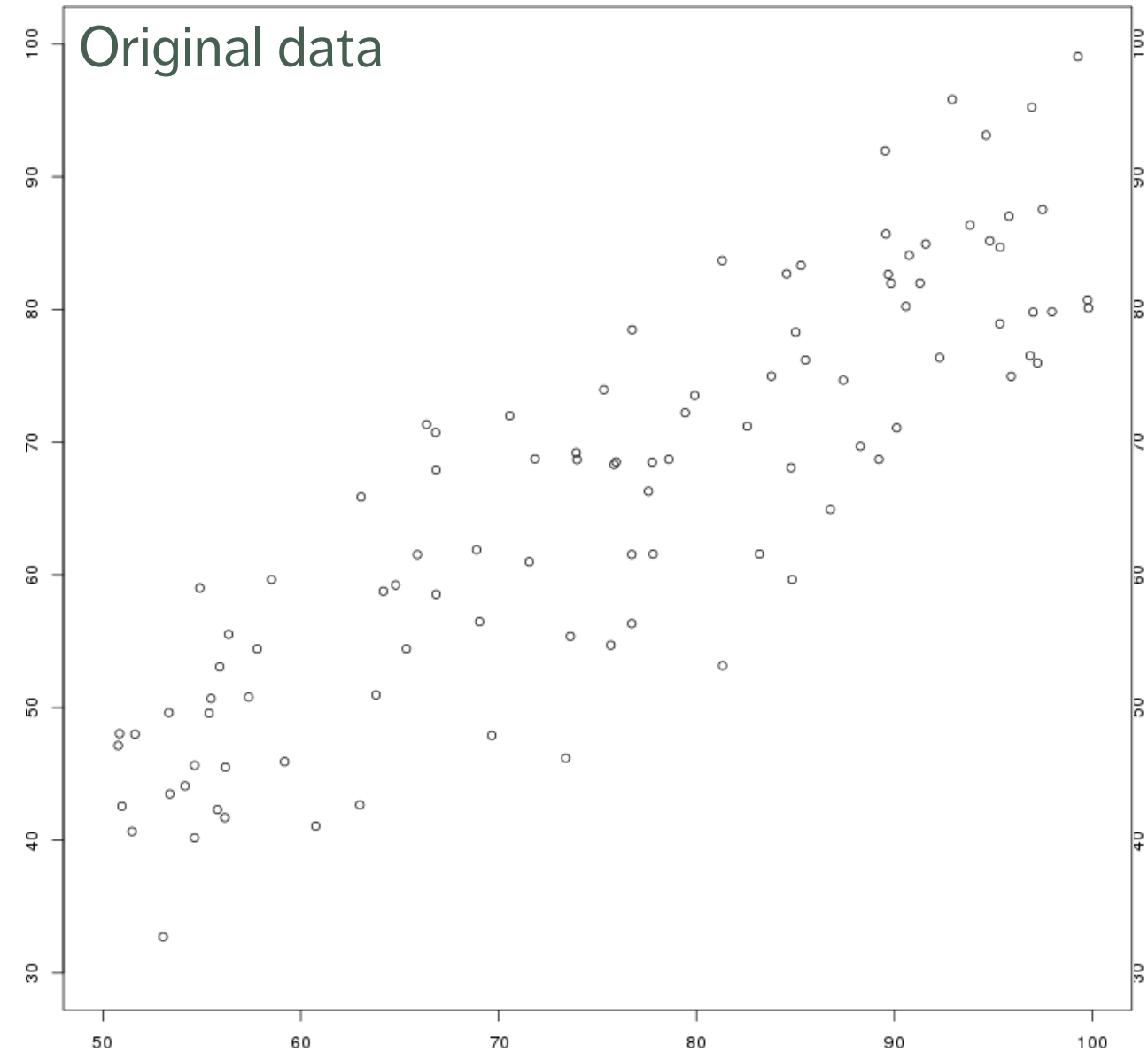
Last observation carried forward: substitute the missing values with latest previous values (in a longitudinal study)

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

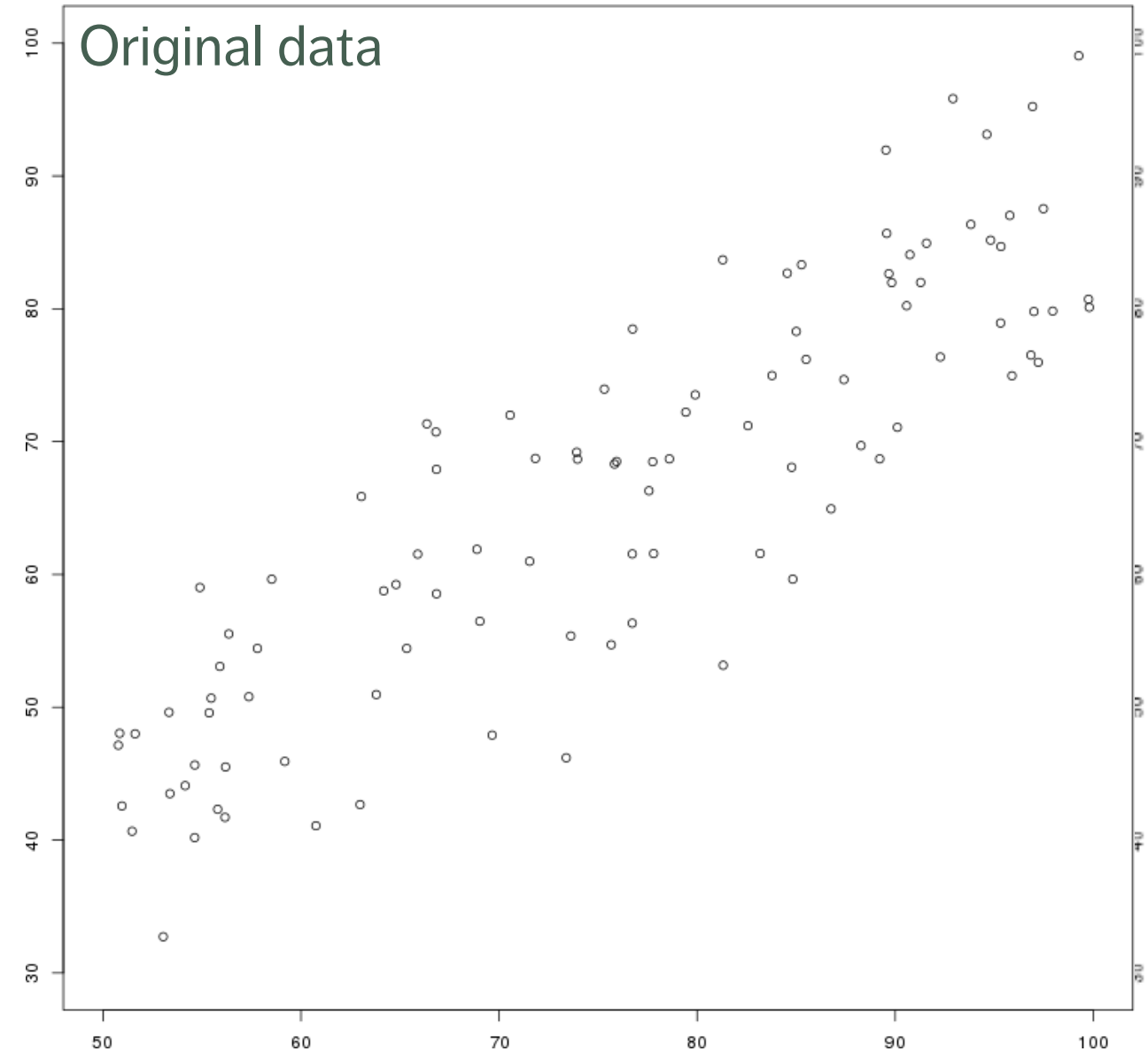
k nearest neighbour imputation (kNN): substitute the missing entry with the average from the group of the k most similar complete cases

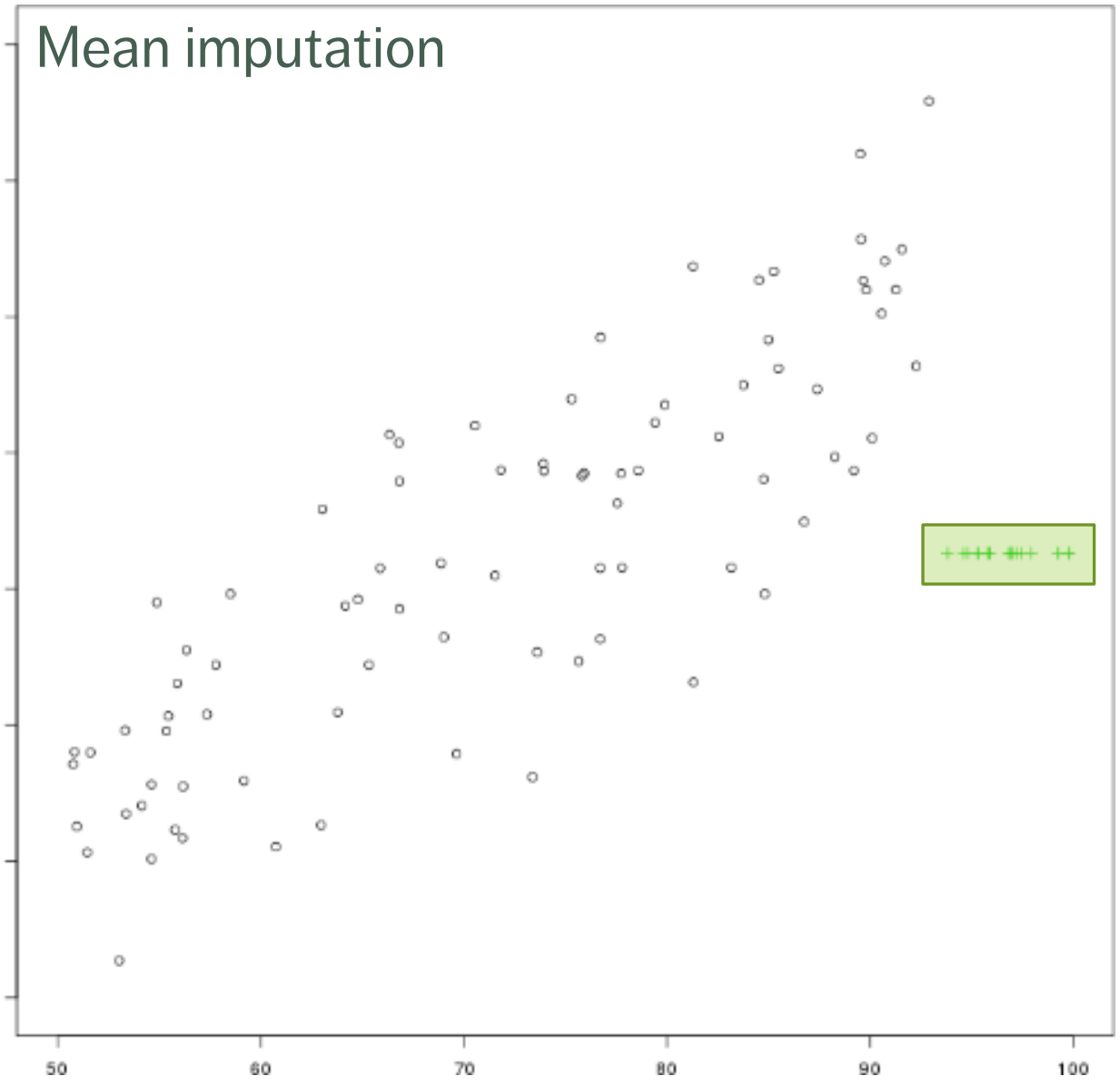
- **assumption:** MAR
- **cons:** difficult to choose appropriate value for k; possible distortion in data structure

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

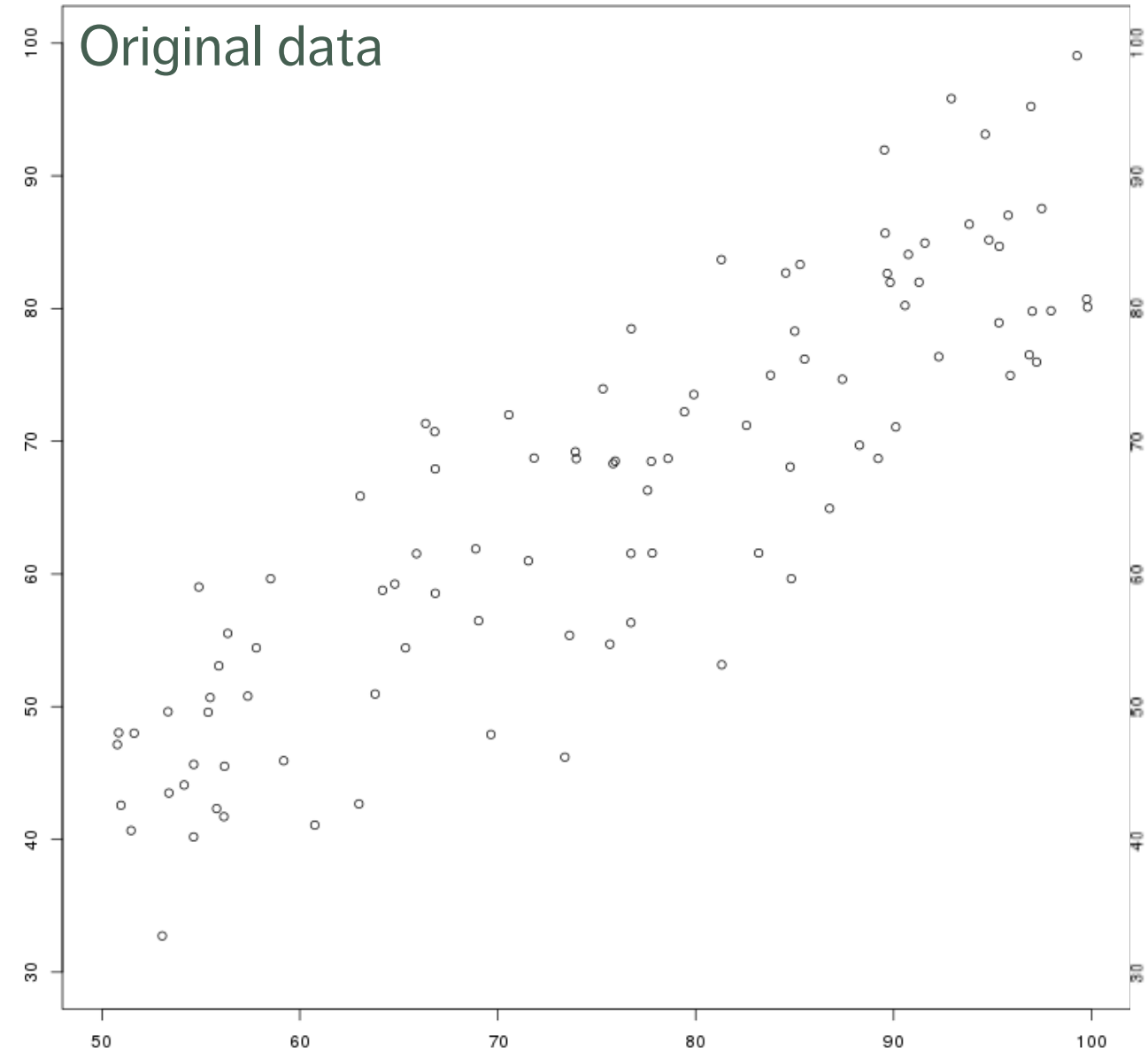


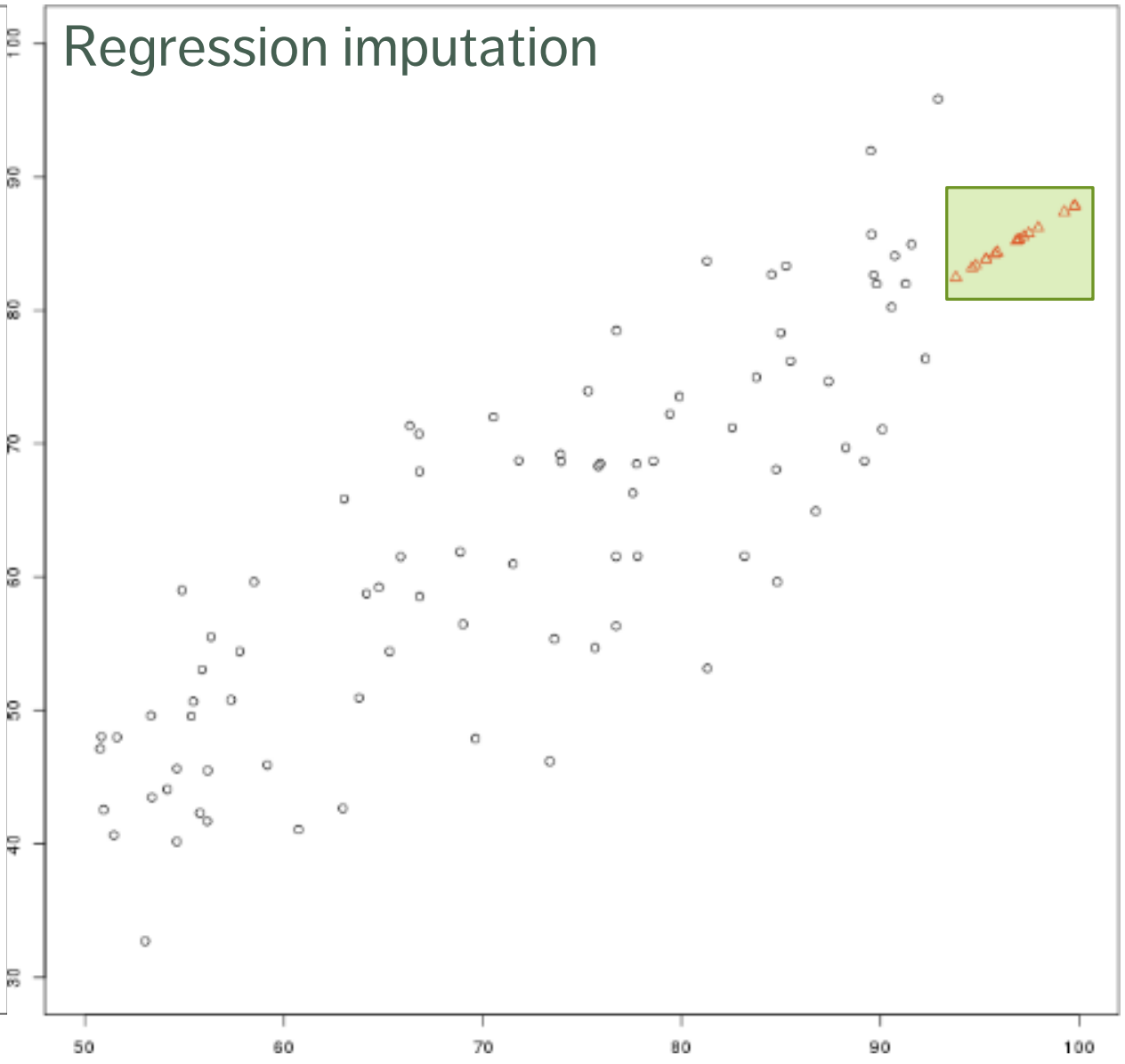
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



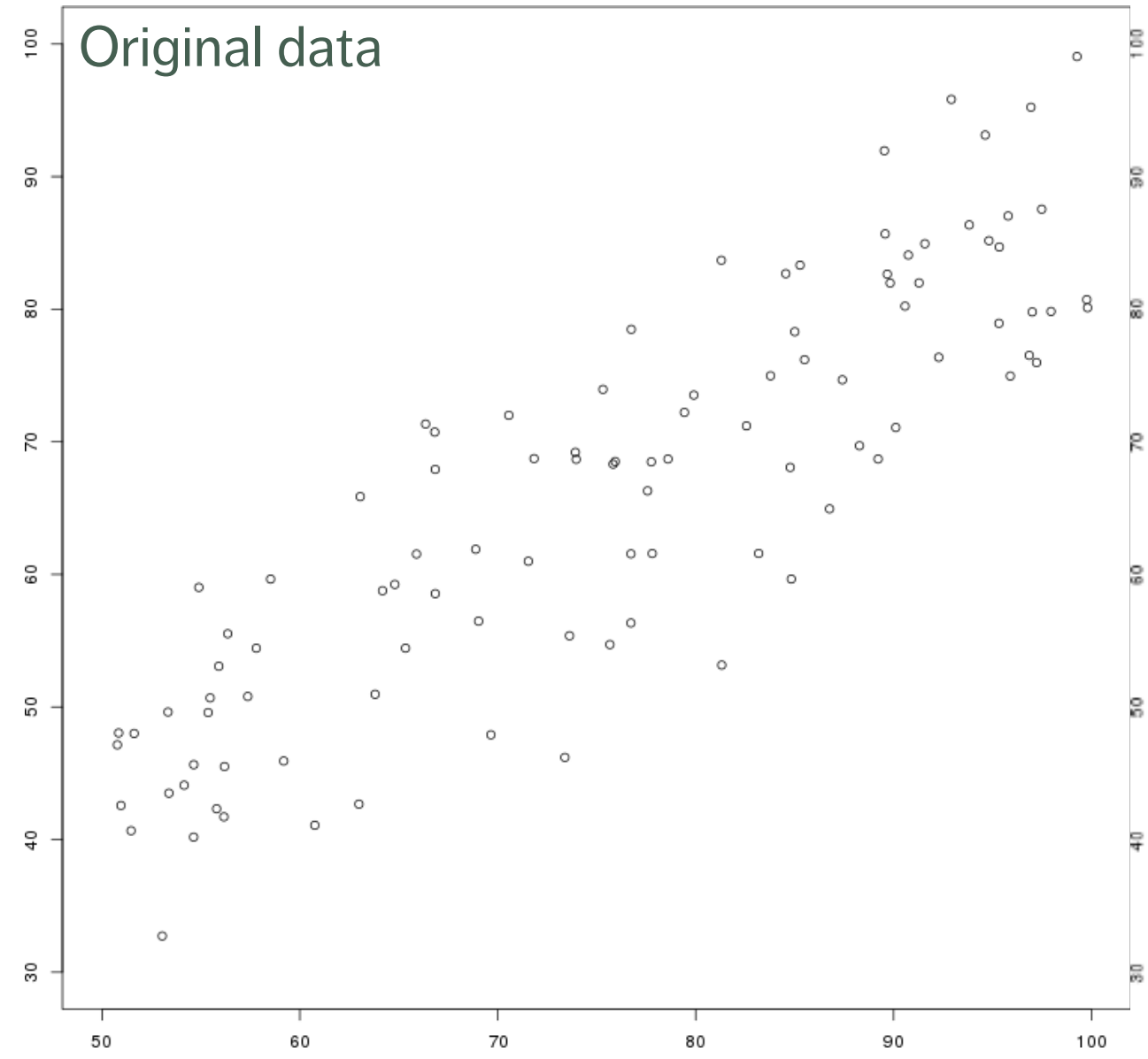


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





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
- if the analysis output is not a single value but some complicated mathematical object, this approach is unlikely to be useful

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 **MCAR** or **MAR**, 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 … **No-Free Lunch theorem**!

Suggested Reading

Missing Values

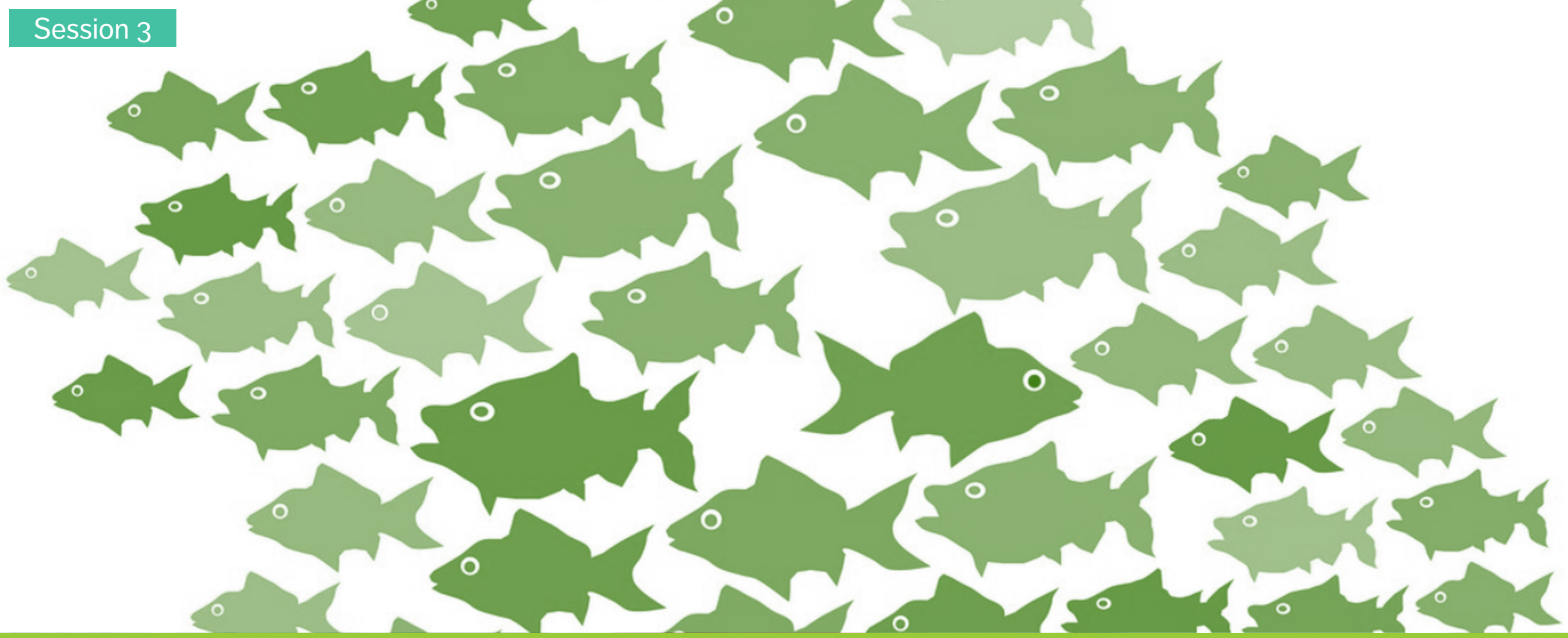
Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.4 Missing Values
    - Missing Value Mechanisms
    - Imputation Methods
    - Multiple Imputation

Exercises

Missing Values

1. Recreate the examples of Imputation Methods.
2. Recreate the missing value imputation process (data cleaning) used in Example: Algae Bloom.
3. Conduct kNN imputation on the grades dataset with various values of k.
4. Conduct multiple imputation on the grades dataset using stochastic regression in order to estimate the slope and intercept for the line of best fit.



9. Anomalous Observations

Anomalous Observations

In practice, an **anomalous observation** may arise as

- a **"bad" object/measurement:** data artifacts, spelling mistakes, poorly imputed values, etc.
- a **misclassified observation:** according to the existing data patterns, the observation should have been labeled differently;
- an observation whose measurements are found in the **distribution tails** of a large enough number of features;
- an **unknown unknown**: a completely new type of observations whose existence was heretofore unsuspected.

Anomalous Observations

Observations could be anomalous in one context, but not in another:

- A 6-foot tall adult male is in the 86th percentile for Canadian males (tall, but not unusual)
- in Bolivia, the same man would be in the 99.9th percentile (very tall and unusual)

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?

Outliers

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*)

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.

Detecting Anomalies

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

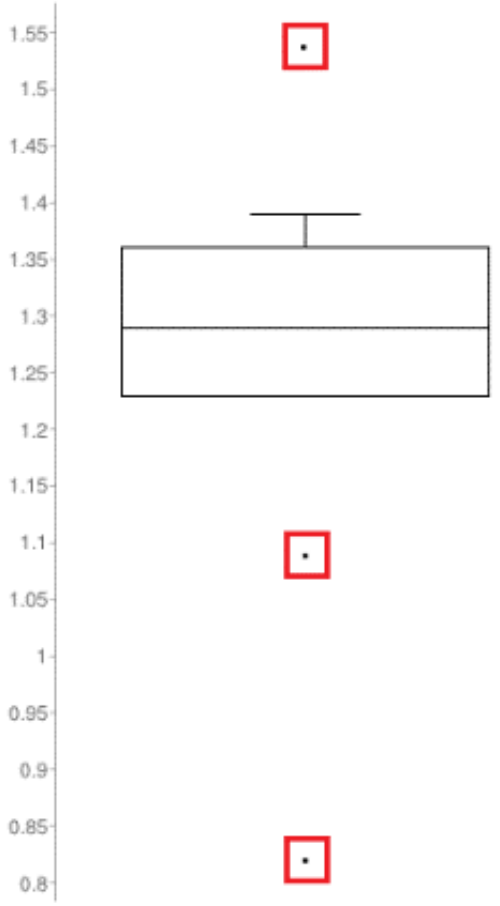
Anomalies are by definition **infrequent**, and typically shrouded in **uncertainty** due to small sample sizes.

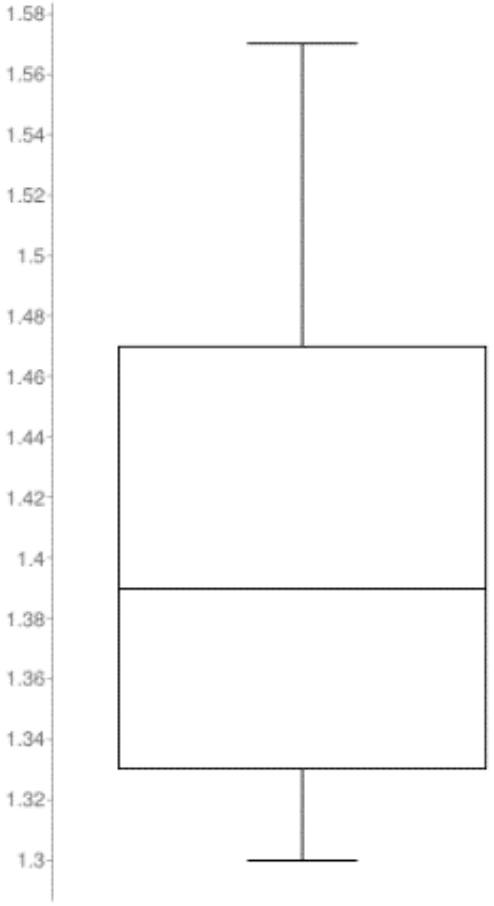
Differentiating anomalies from noise or data entry errors is **hard**.

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

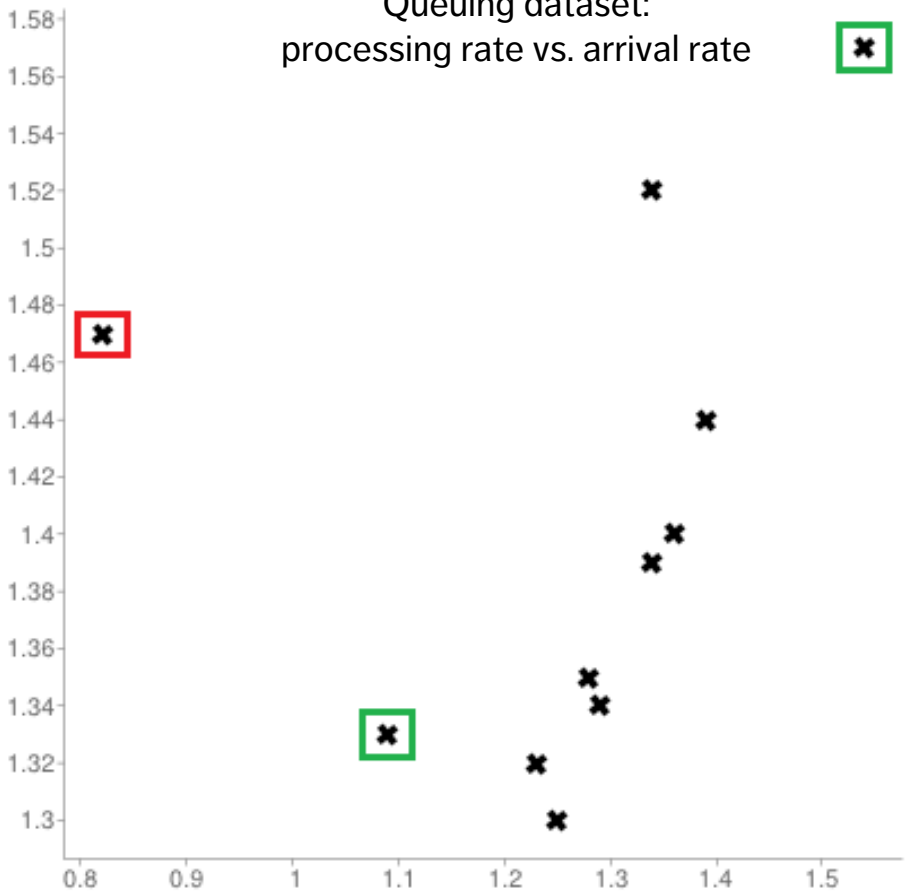
Anomalies associated with malicious activities are typically **disguised**.

Visual Outlier Detection





Queuing dataset:
processing rate vs. arrival rate



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)

Careful: once anomalous observations have been removed from the dataset, previously “regular" units may become anomalous.

Anomaly Detection Algorithms

Supervised methods use a historical record of labeled anomalous observations:

- domain expertise is required to tag the data
- classification or regression task
- rare occurrence problem

		Predicted Class: Normal	Predicted Class: Anomaly
Actual Class	Normal	TN	FP
	Anomaly	FN	TP

Unsupervised methods don't use external information:

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

Anomaly Detection Algorithms

The mis-classification cost is often assumed to be symmetrical, which can lead to **technically correct but useless** outputs.

For instance, most (99.999+%) air passengers do not bring weapons with them on flights; a model that predicts that no passenger is smuggling a weapon would be 99.999+% accurate, but it would miss the point completely.

For the **security agency**, the cost of wrongly thinking that a passenger is:

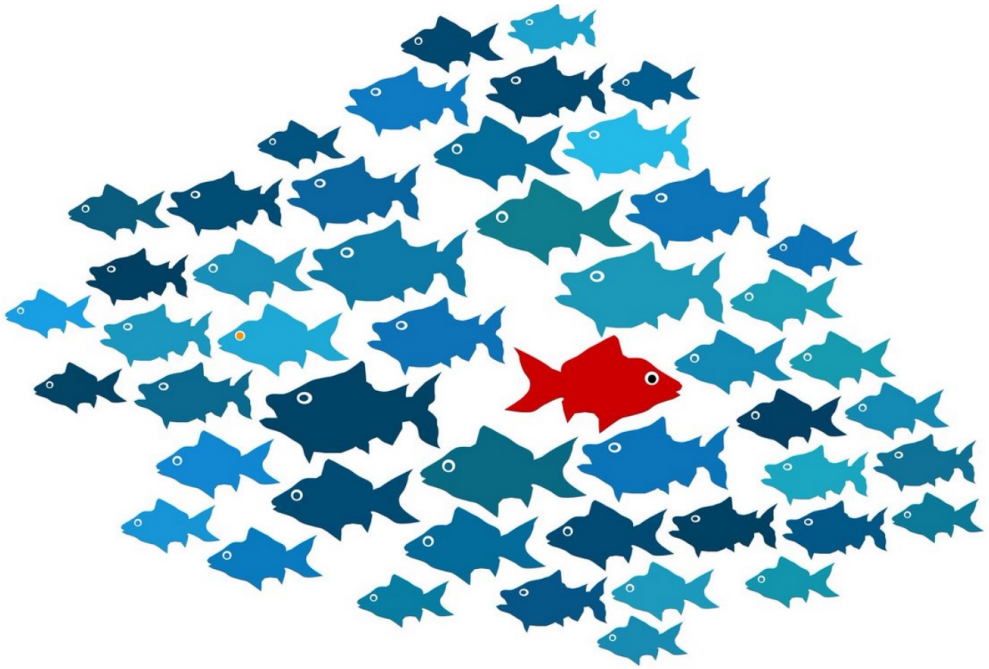
- smuggling a weapon ⟹ cost of a single search
- NOT smuggling a weapon ⟹ catastrophe (potentially)

The wrongly targeted individuals may have a different take on this!

Anomaly Detection Algorithms

If all participants in a workshop except for one can view the video conference lectures, then the one individual/internet connection/computer is **anomalous** – it behaves in a manner which is different from the others.

But this **DOES NOT MEAN** that the different behaviour is necessarily the one we are interested in...



Simple Outlier Tests

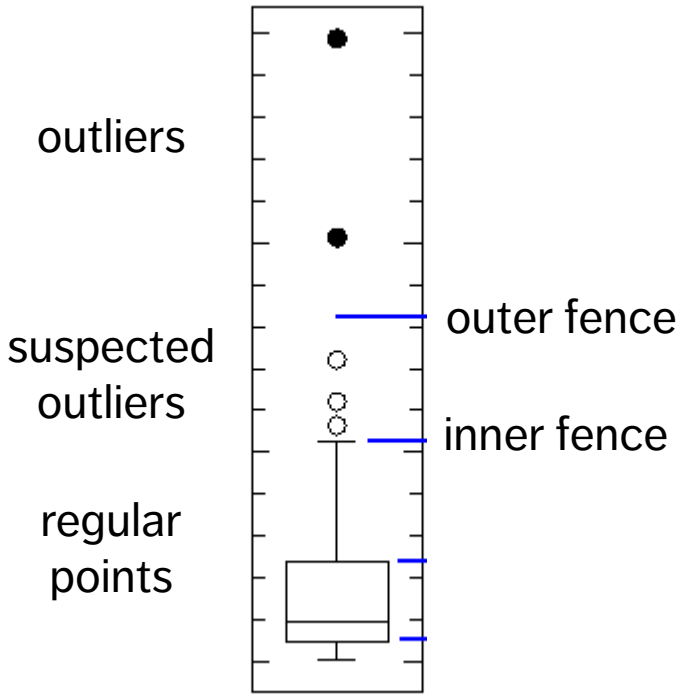
Tukey's Boxplot test: for normally distributed data, regular observations typically lie between the **inner fences**

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

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**.



Simple 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 simple tests:

- **Grubbs** (univariate)
- **Tietjen-Moore** (for a specific # of outliers)
- **generalized extreme studentized deviate** (for unknown # of outliers)
- **chi-square** (outliers affecting goodness-of-fit)

Sophisticated Anomaly Detection

- **DBSCAN**, **OR$_h$**, and **LOF** (unsupervised outlier detection)
- **rank-power** method (supervised outlier detection)
- **distance** or **density-based** methods (with exotic distance measures)
- **autoencoders and reconstruction error** (deep learning method)
- **rare-occurrence** methods (oversampling, undersampling, CREDOS, PN, SHRINK, SMOTE, DRAMOTE, SMOTEBoost, RareBoost, MetaCost, AdaCost, CSB, SSTBoost, etc.)
- **AVF**, **Greedy** algorithms (categorical data)
- **PCA**, **DOBIN**, and other **projection** methods (for high-dimensional data)
- **subspace** methods and **ensemble** methods

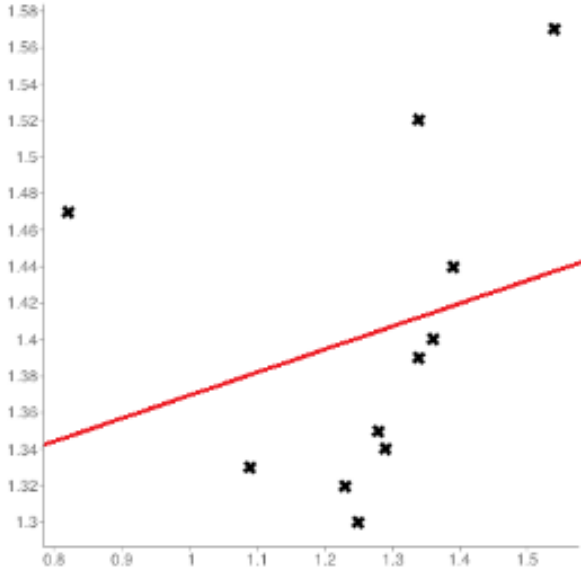
Influential Observations

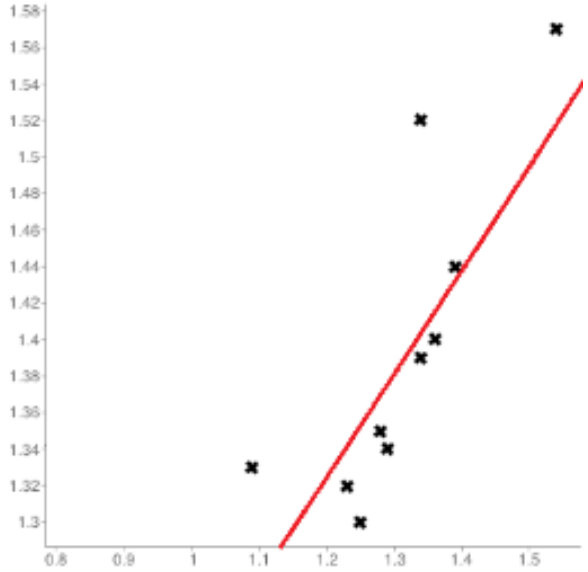
Influential data points are observations whose absence leads to **markedly different** analysis results.

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

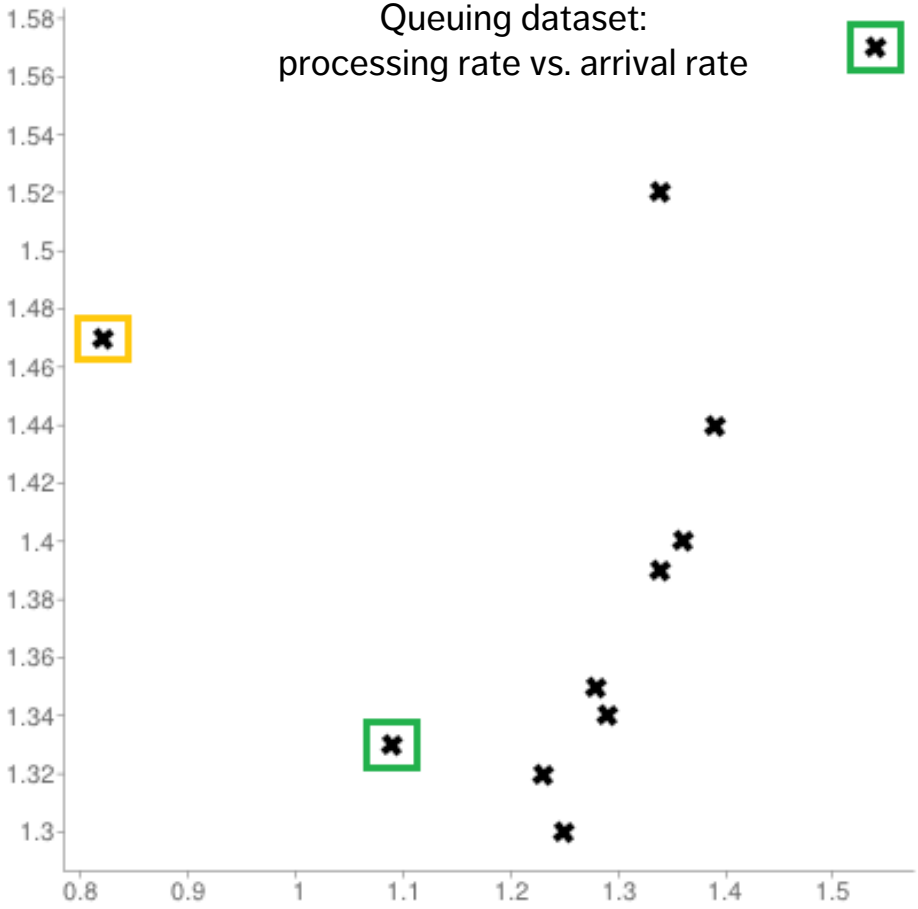
Outliers may be influential data points; influential data points need not be outliers (and *vice-versa*).

Influential Observations





Queuing dataset:
processing rate vs. arrival rate



Anomaly Detection Remarks

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 data 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**.

Suggested Reading

Anomalous Observations

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.5 Anomalous Observations
    - Anomaly Detection
    - Outlier Tests
    - Visual Outlier Detection

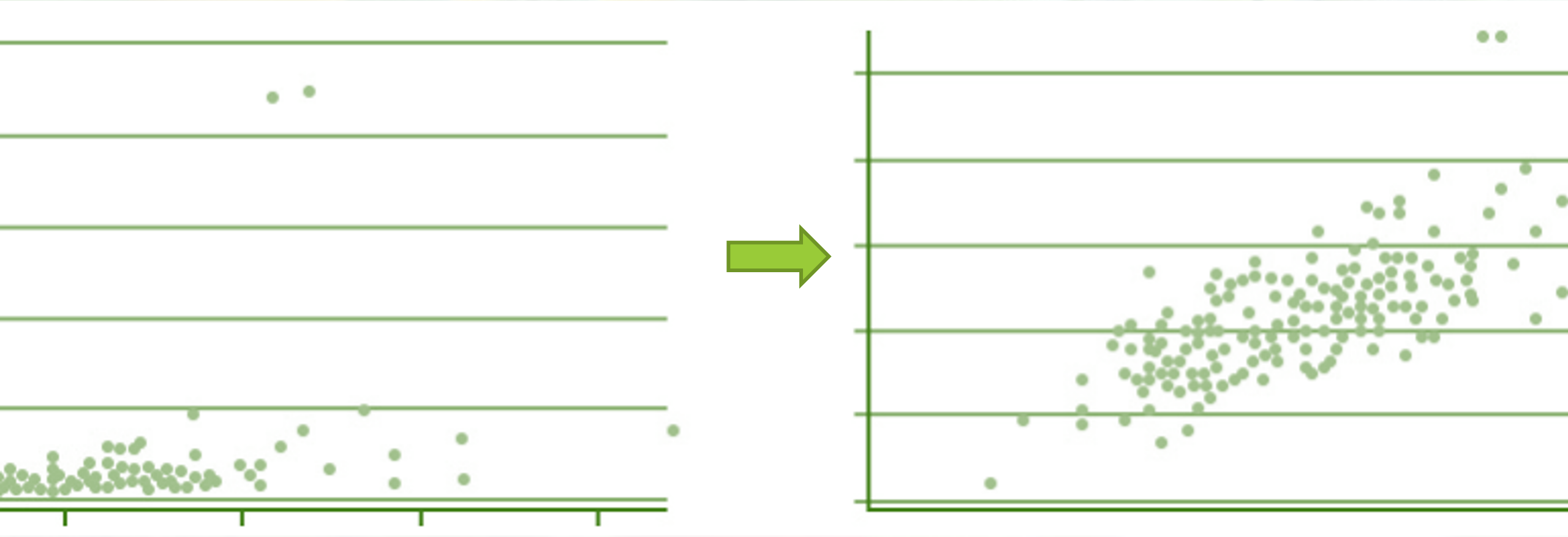
Volume 4: Techniques of Data Analysis

26. Anomaly Detection and Outlier Analysis

Exercises

Anomalous Observations

1. Recreate the anomaly detection process used in Example: Algae Bloom.
2. Find anomalous observations in the cities.txt and grades datasets (if applicable).
3. Find anomalous observations in a dataset of your choice.



10. Dimensionality and Data Transformations

Dimensionality of Data

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

We can think of the number of variables used to describe each object (row) as a vector describing that object: the dimension is simply the **size** of that vector.

(**Note:** "dimension" is used differently in business intelligence contexts)

High Dimensionality and Big Data

Datasets can be "big" in a variety of ways:

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

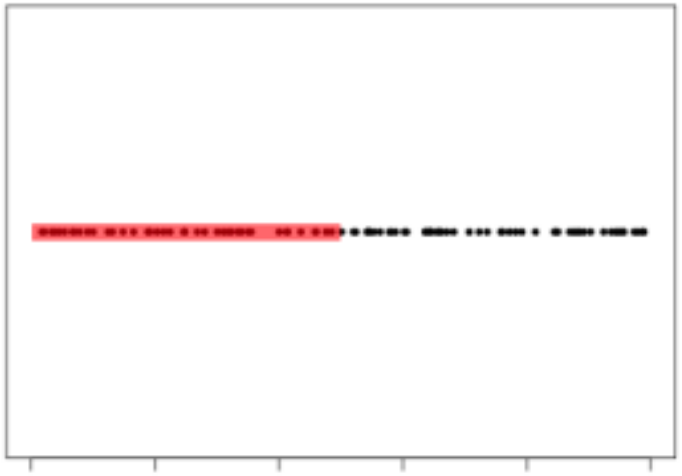
Examples:

- 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 **sparse data**

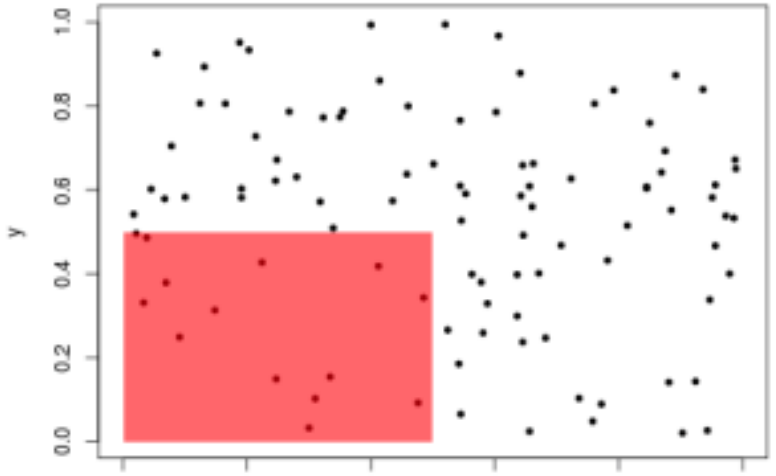
[http://simplystatistics.org]

Curse of Dimensionality

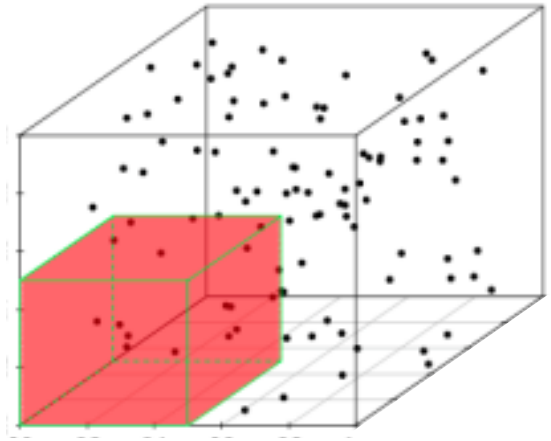
42% of data is captured



14% of data is captured



7% of data is captured



$N = 100$ observations, uniformly distributed on $[0,1]^d, d = 1, 2, 3.$
% of observations captured by $[0,1/2]^d, d = 1,2,3.$

Sampling Observations

Question: does every row of the dataset need to be used?

If rows are selected randomly (with or without replacement), 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 – **prior information on the dataset structure can help**

Feature Selection

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

Motivations:

- modeling tools do not handle these well (variance inflation due to multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

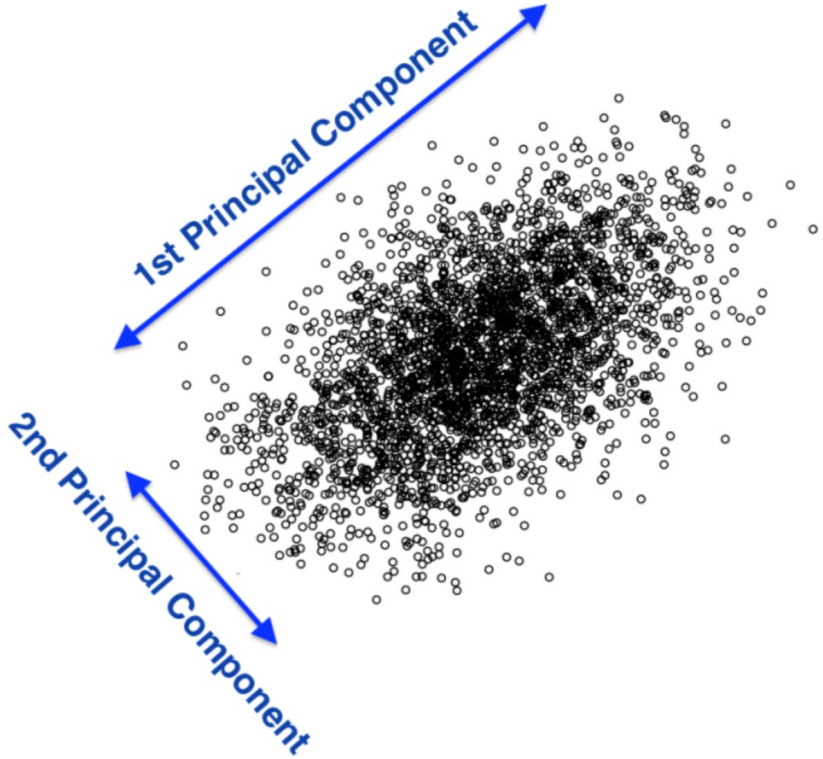
[A. Ng, K. Soo, *Numsense!*]

Dimension Reduction: PCA

Motivational Example: Nutritional Content of Food

What is the best way to differentiate food items? Vitamin content, fat, or protein level? A bit of each?

Principal Component Analysis (PCA) can be used to find the combinations of variables along which the data points are **most spread out** (dimension reduction).



[A. Ng, K. Soo, *Numsense!*]

Dimension Reduction: PCA

Presence of nutrients appears to be **correlated** among food items.

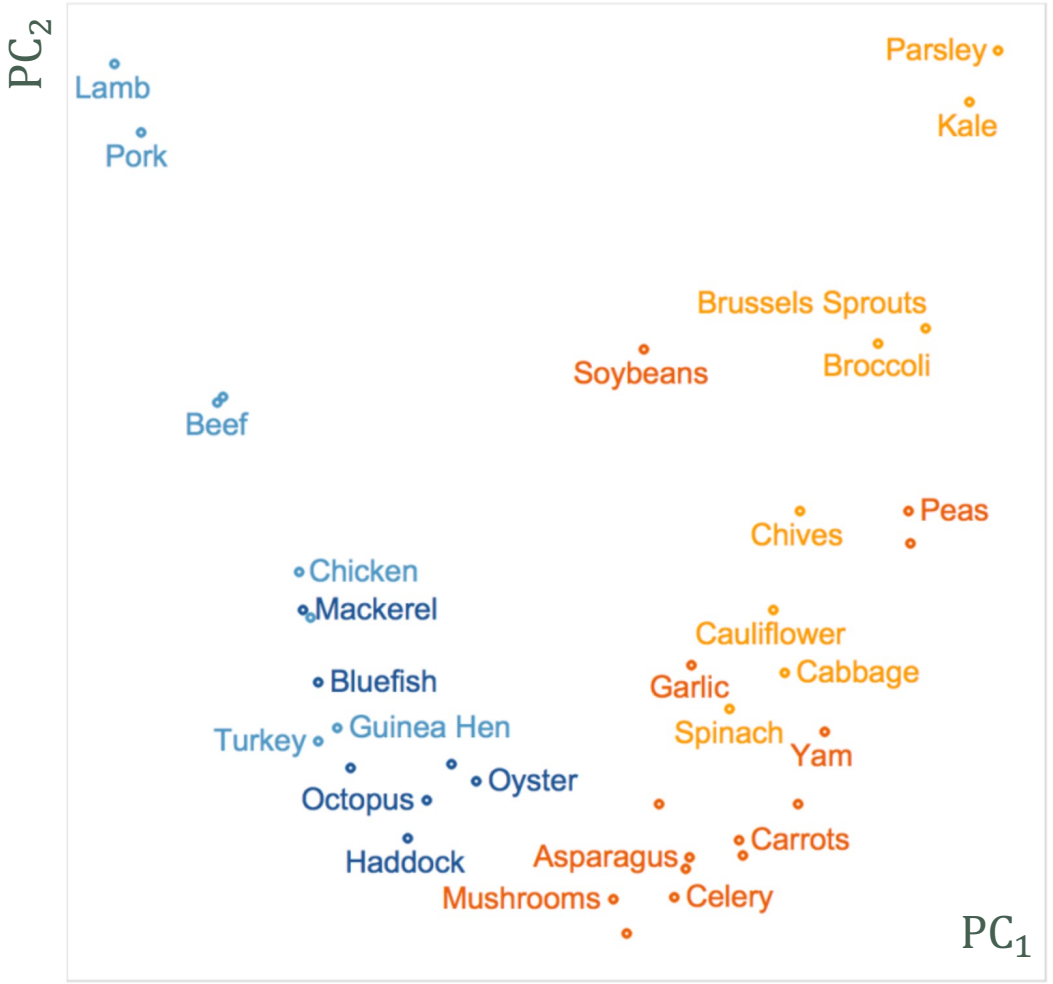
In the (small) sample consisting of Lamb, Pork, Kale, and Parsley, *Fat* and *Protein* levels seem in step, as do *Fiber* and *Vitamin C*.

In a larger dataset, the correlations are $r = 0.56$ and $r = 0.57$.

How much could 2 **derived** variables explain?

$$PC_1 = -0.45 \times \text{Fat} - 0.55 \times \text{Protein} + 0.55 \times \text{Fiber} + 0.44 \times \text{Vitamin C}$$
$$PC_2 = 0.66 \times \text{Fat} + 0.21 \times \text{Protein} + 0.19 \times \text{Fiber} + 0.70 \times \text{Vitamin C}$$

PCA Differentiation



PC_1 differentiates vegetables from meats; PC_2 differentiates 2 **sub-categories** within these:

- **meats** are concentrated on the left (low PC_1 values)
- **vegetables** are concentrated on the right (high PC_1 values)
- **seafood** have lower *Fat* content (low PC_2 values) and are concentrated at the bottom
- **non-leafy veggies** have lower *Vitamin C* content (low PC_2 values) and are also bunched at the bottom

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

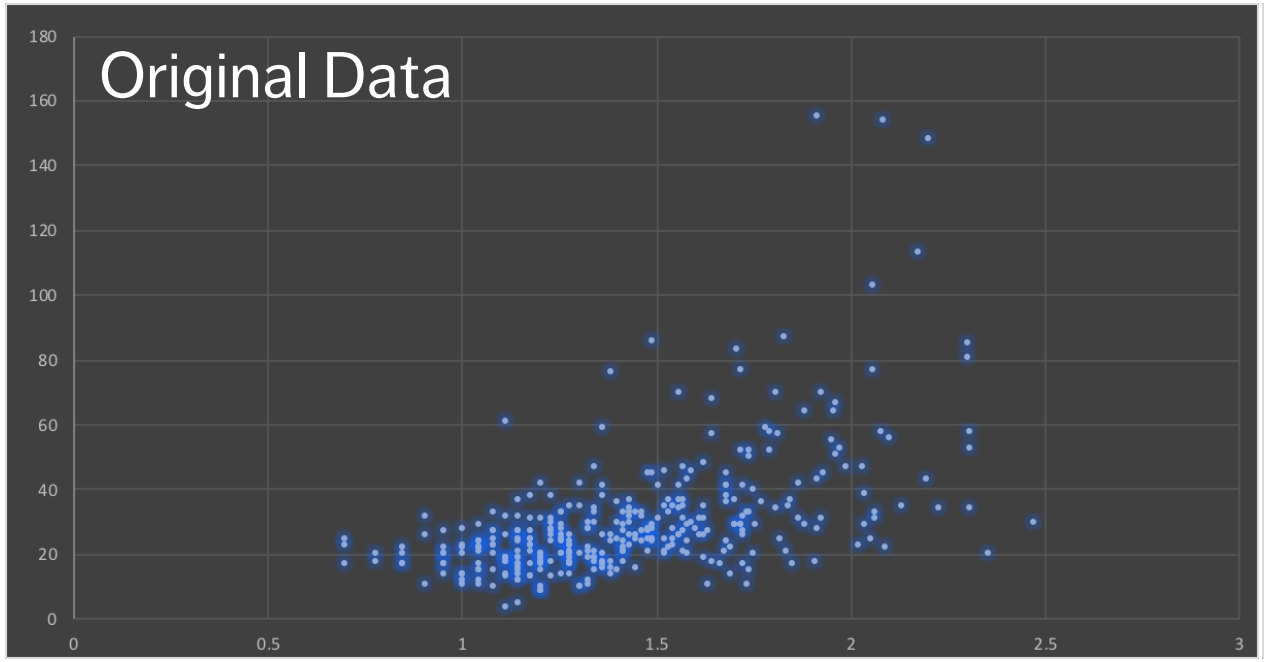
The second approach requires an **inverse transformation** to be able to draw conclusions about the **original data**.

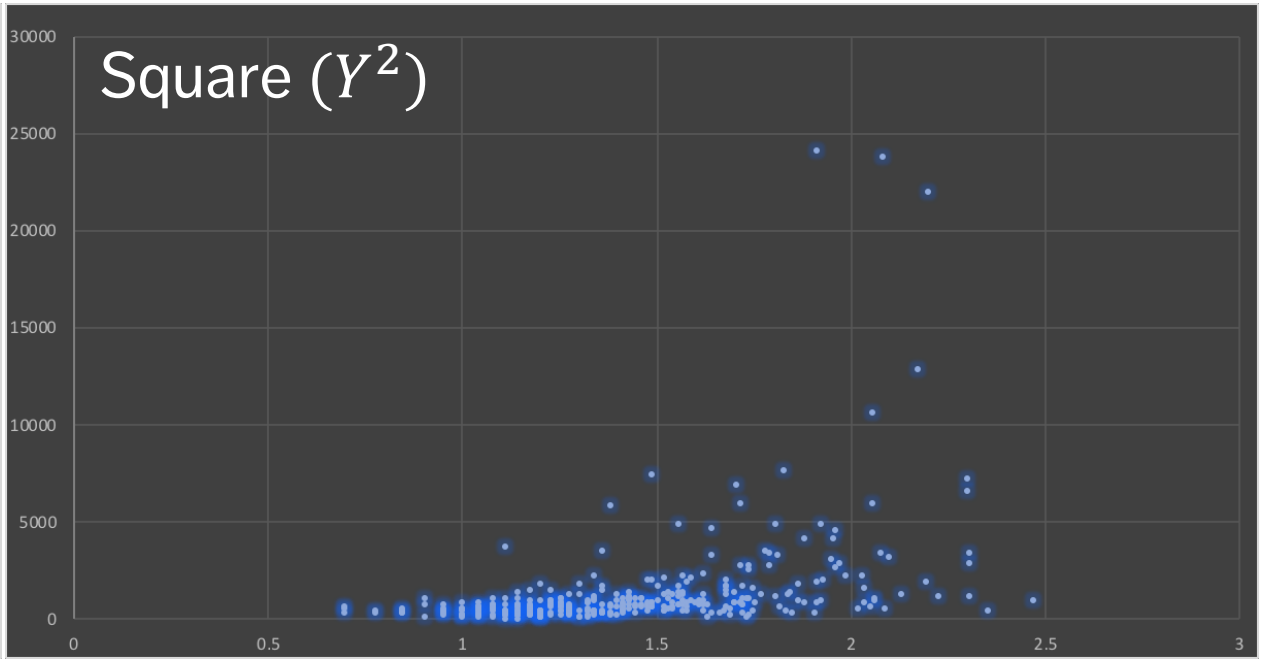
Common Transformations

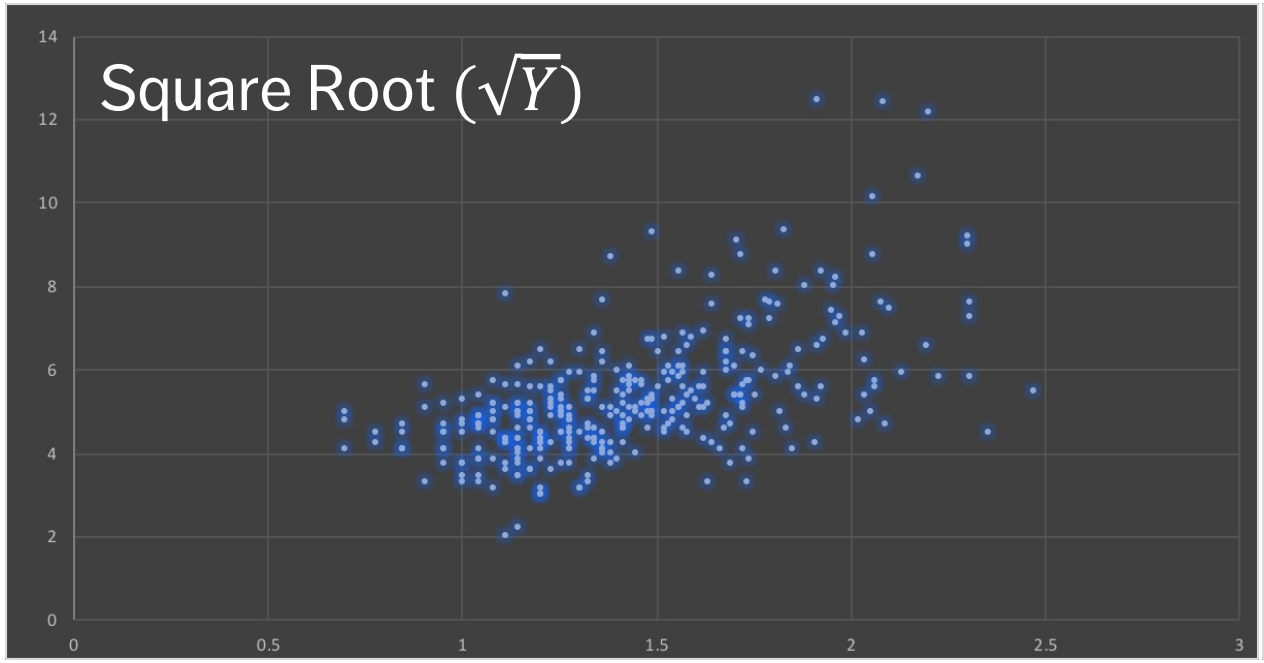
In the data analysis context, transformations are **monotonic:**

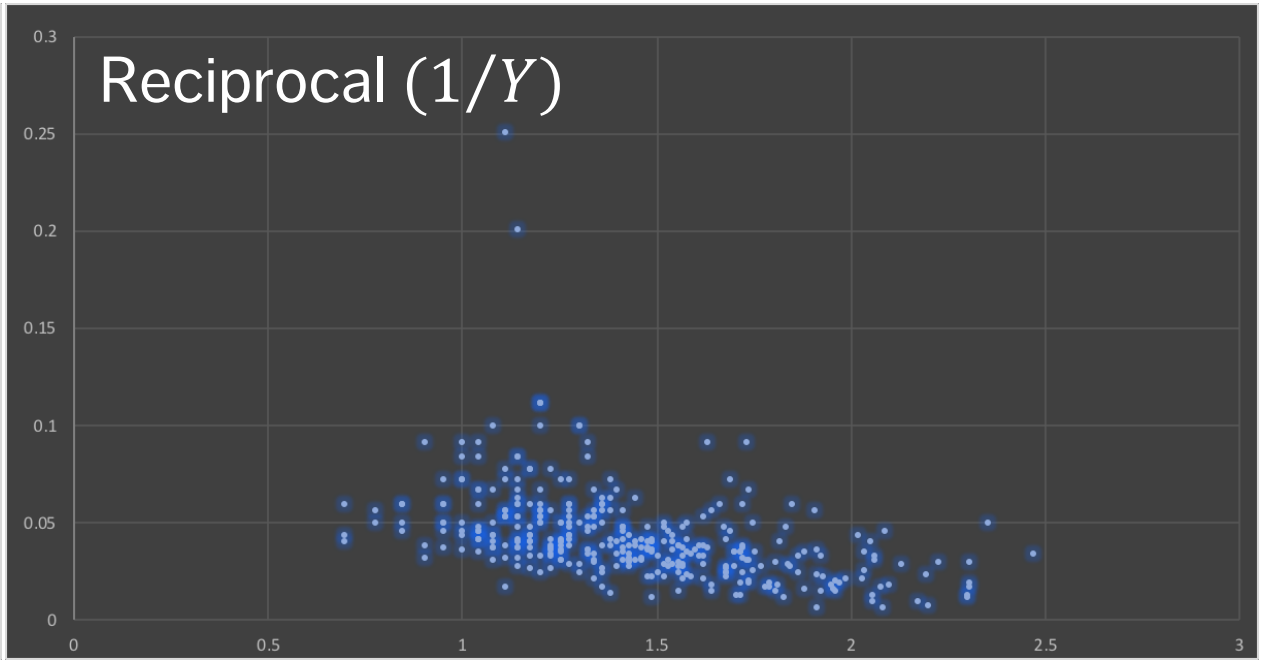
- logarithmic
- square root, inverse, power: W^k
- exponential
- Box-Cox, etc.

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
180
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3


Square (Y^2)
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root ($\sqrt{Y}$)
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal ($1/Y$)
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3

gm = geometric mean

Box-Cox Transformation

Assume the usual model $Y_j = \sum_i \beta_i X_{j,i} + \varepsilon_j$ with either

- skewed residuals
- not-constant variance
- non-linear trend

The **Box-Cox transformation** $Y_j \mapsto Y_j'(\lambda)$ suggests a choice: select λ which maximizes the corresponding log-likelihood

$$Y_j'(\lambda) = \begin{cases} \mathrm{gm}(\boldsymbol{Y}) \times \ln(Y_j), & \lambda = 0 \\ \lambda^{-1} \mathrm{gm}(\boldsymbol{Y})^{1-\lambda} \times (Y_j{}^{\lambda} - 1), & \lambda \neq 0 \end{cases}$$

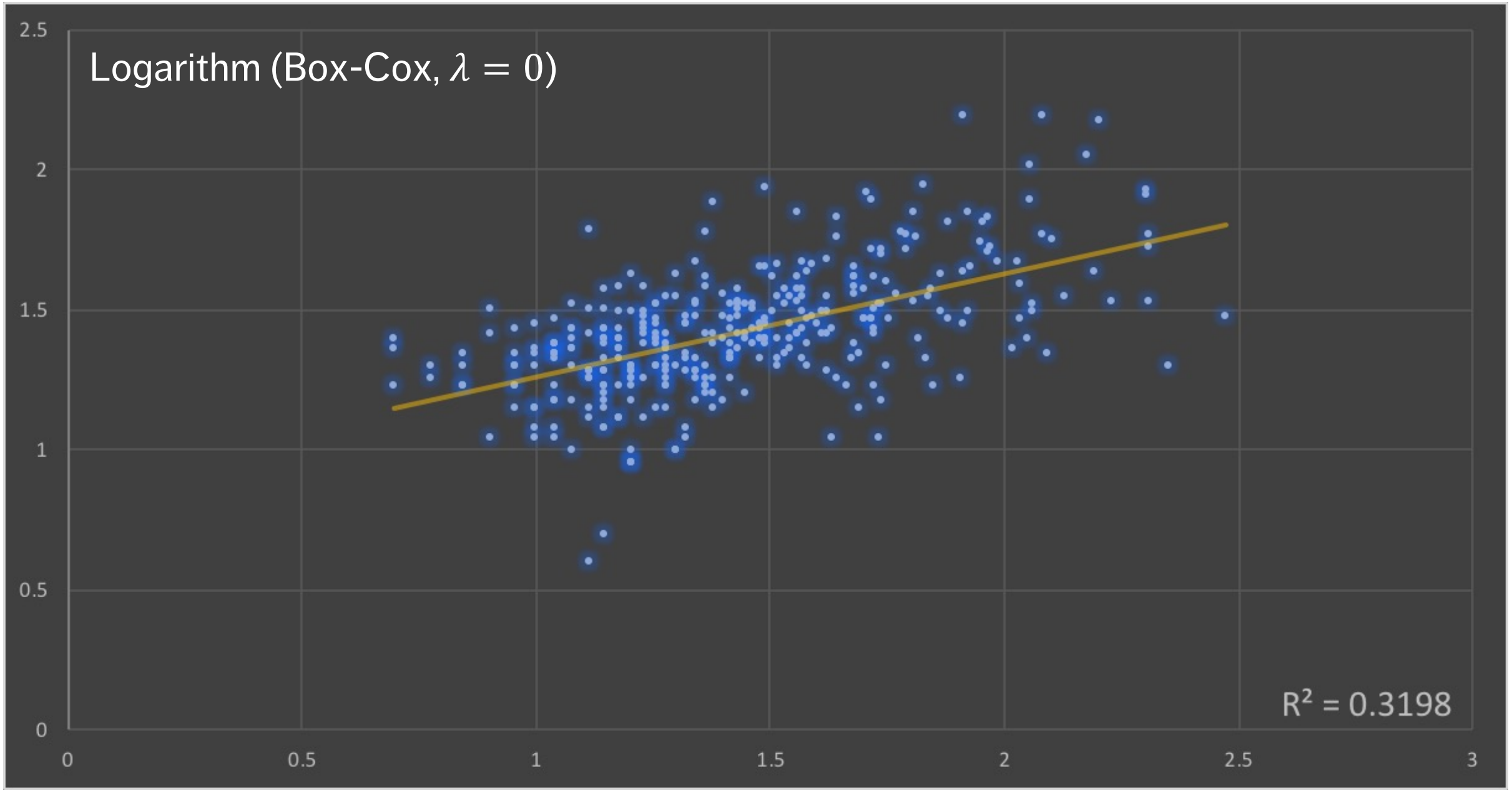
Box-Cox Transformation

The procedure provides a **guide** to select a transformation.

Theoretical/practical **rationales** may exist for a particular choice of λ.

Residual analysis is still required to ensure that the choice was appropriate.

Better to work with (interpret) the transformed data.


Logarithm (Box-Cox, $\lambda = 0$)
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
2.5
3
$R^2 = 0.3198$

Scaling

Numeric variables may have different **scales** (i.e., weights and heights).

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 may 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 (lags?)

Spatial dependencies ⟶ spatial analysis (neighbours?)

Suggested Reading

Dimensionality and
Data Transformations

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

15. Data Preparation
    15.6 Data Transformations
    - Common Transformations
    - Box-Cox Transformation
    - Scaling
    - Discretizing
    - Creating Variables

Volume 3: Spotlight on Machine Learning

23. Feature Selection and Dimension Reduction

Exercises

Dimensionality and
Data Transformations

1. Using Example: Algae Bloom as a basis, scale, discretize, and create new variables out of the algae blooms dataset.
2. Scale, discretize, and create new variables out of the grades and cities.txt datasets.
3. Scale, discretize, and create new variables out of a dataset of your choice.