DATA COLLECTION AND DATA PROCESSING

"People resist a census, but give them a profile page and they'll spend all day telling you who they are."

Max Berry, Lexicon

OUTLINE

1. Getting Ready for Analysis: Data Cleaning
2. Making Your Data (More) Manageable: Data Transformation
3. Ensuring Good Data: Data Quality and Data Validation

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

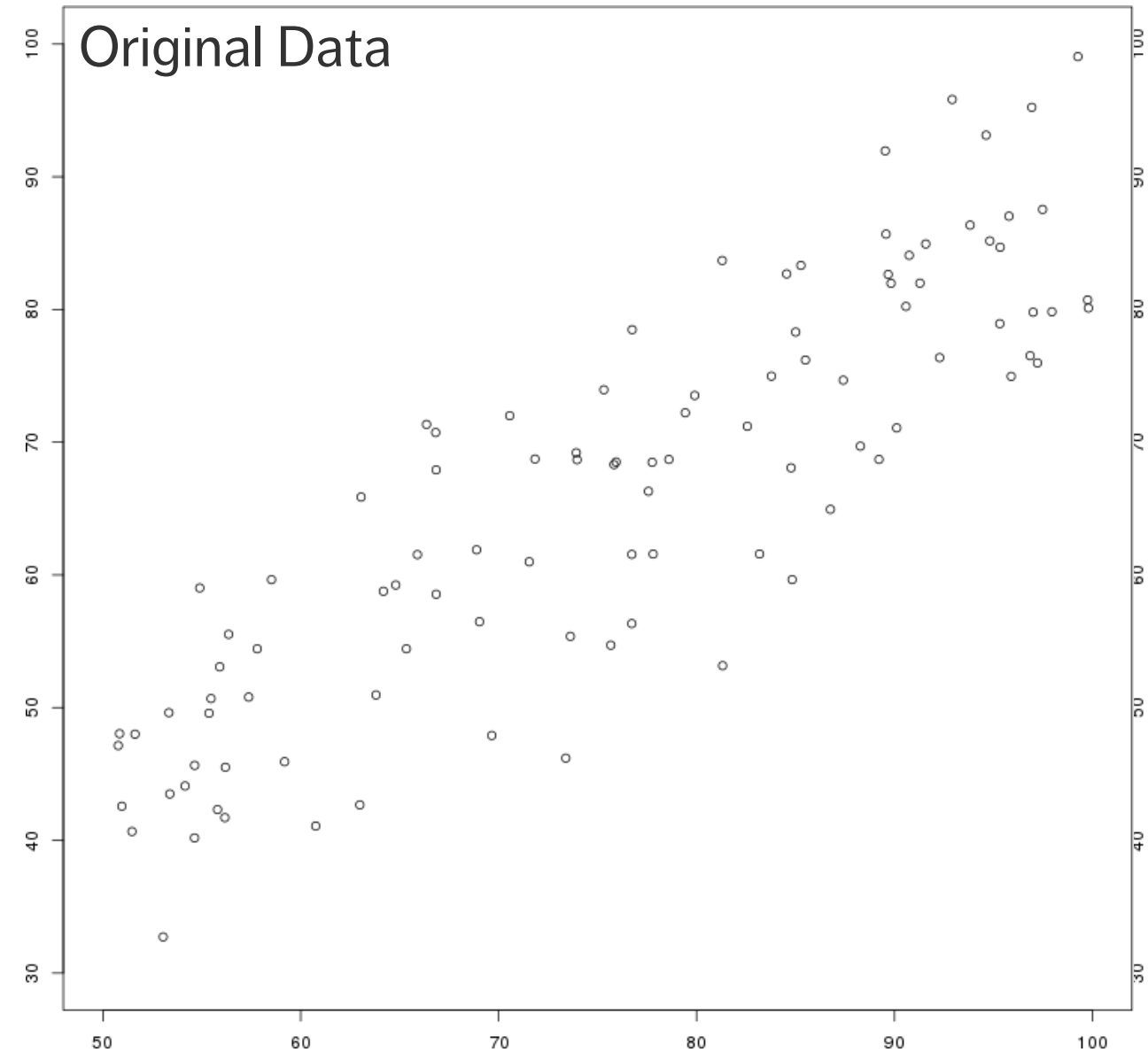
Stochastic regression imputation

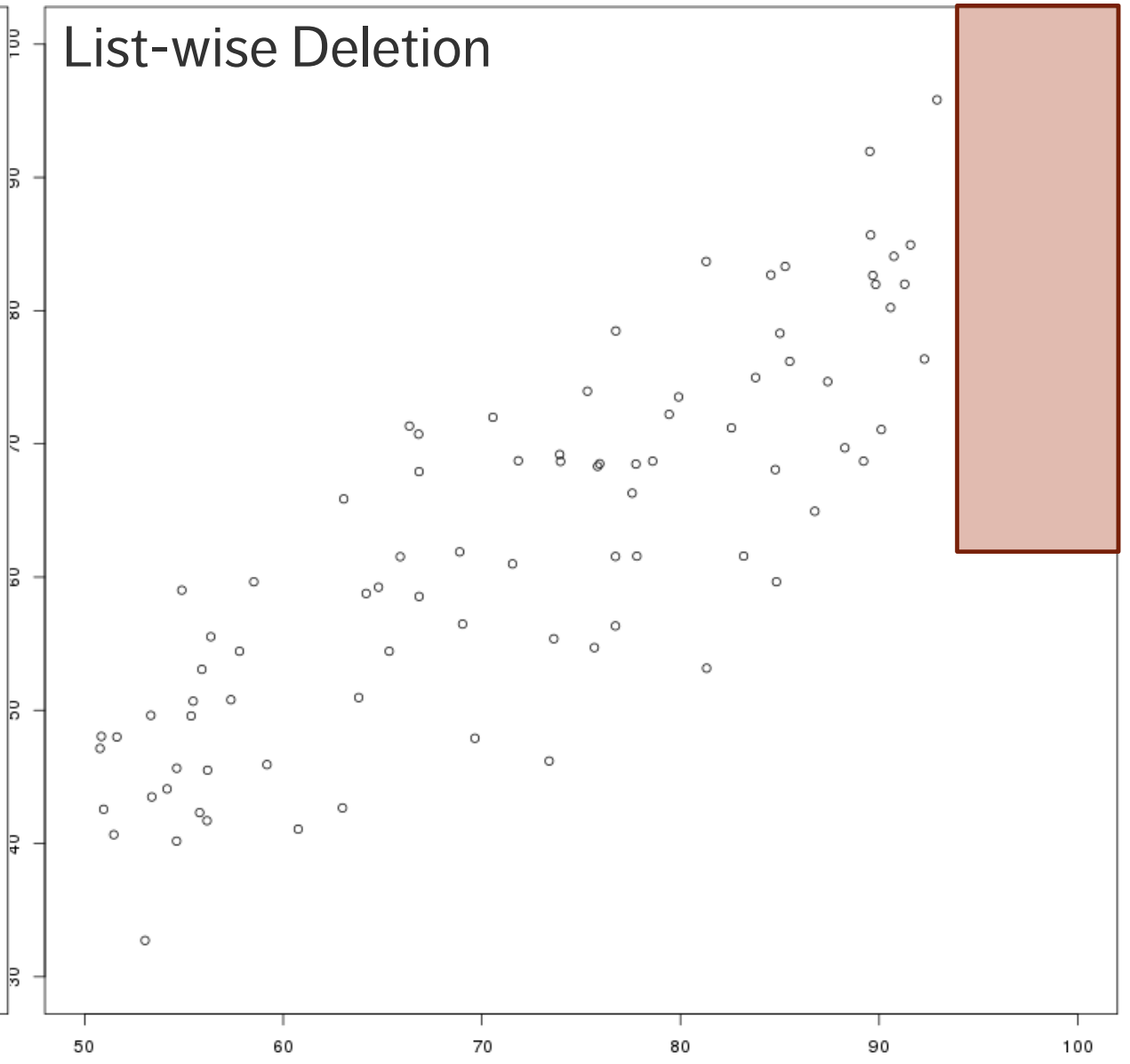
Last observation carried forward

k-nearest neighbours imputation

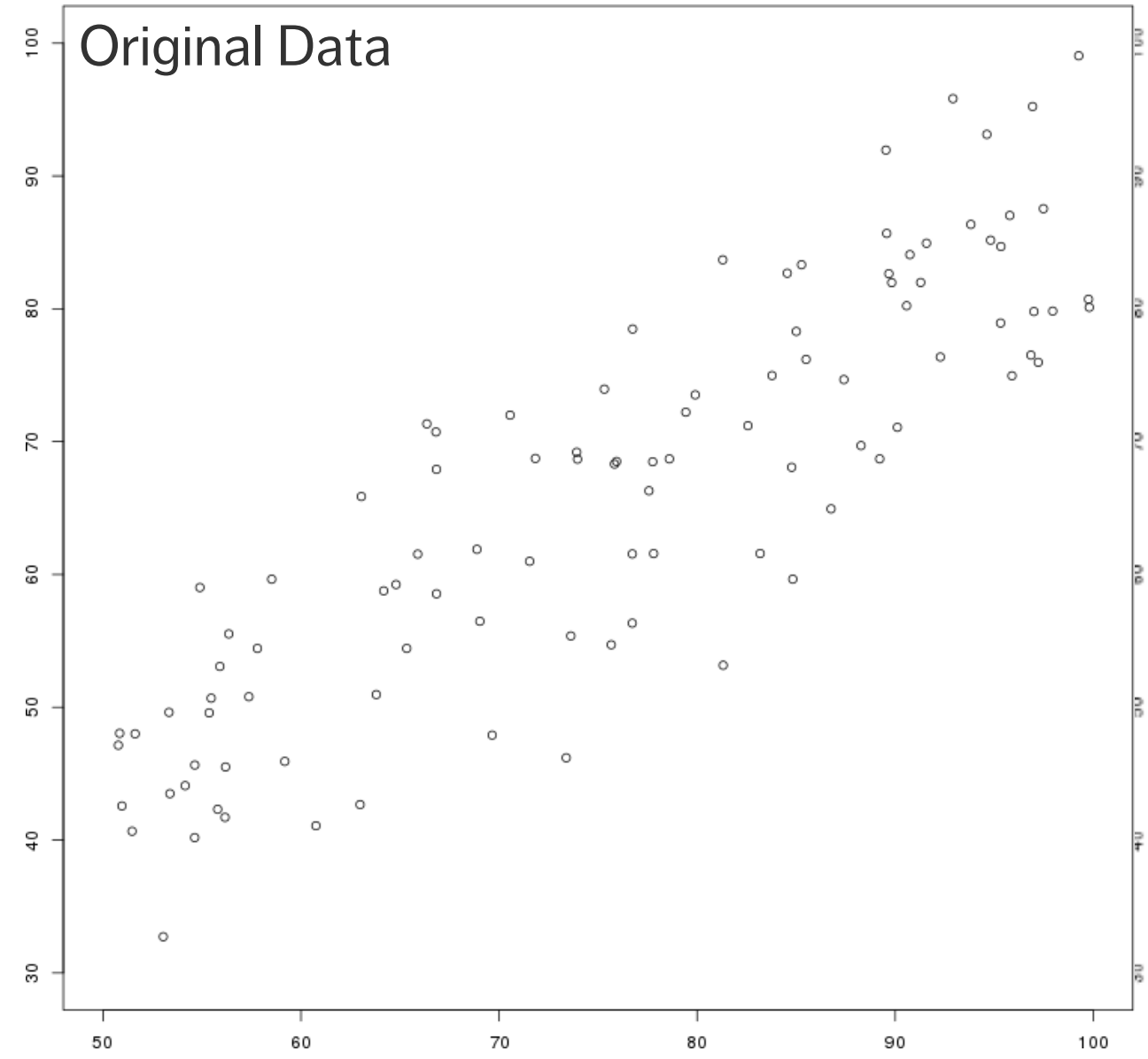
Multiple imputation

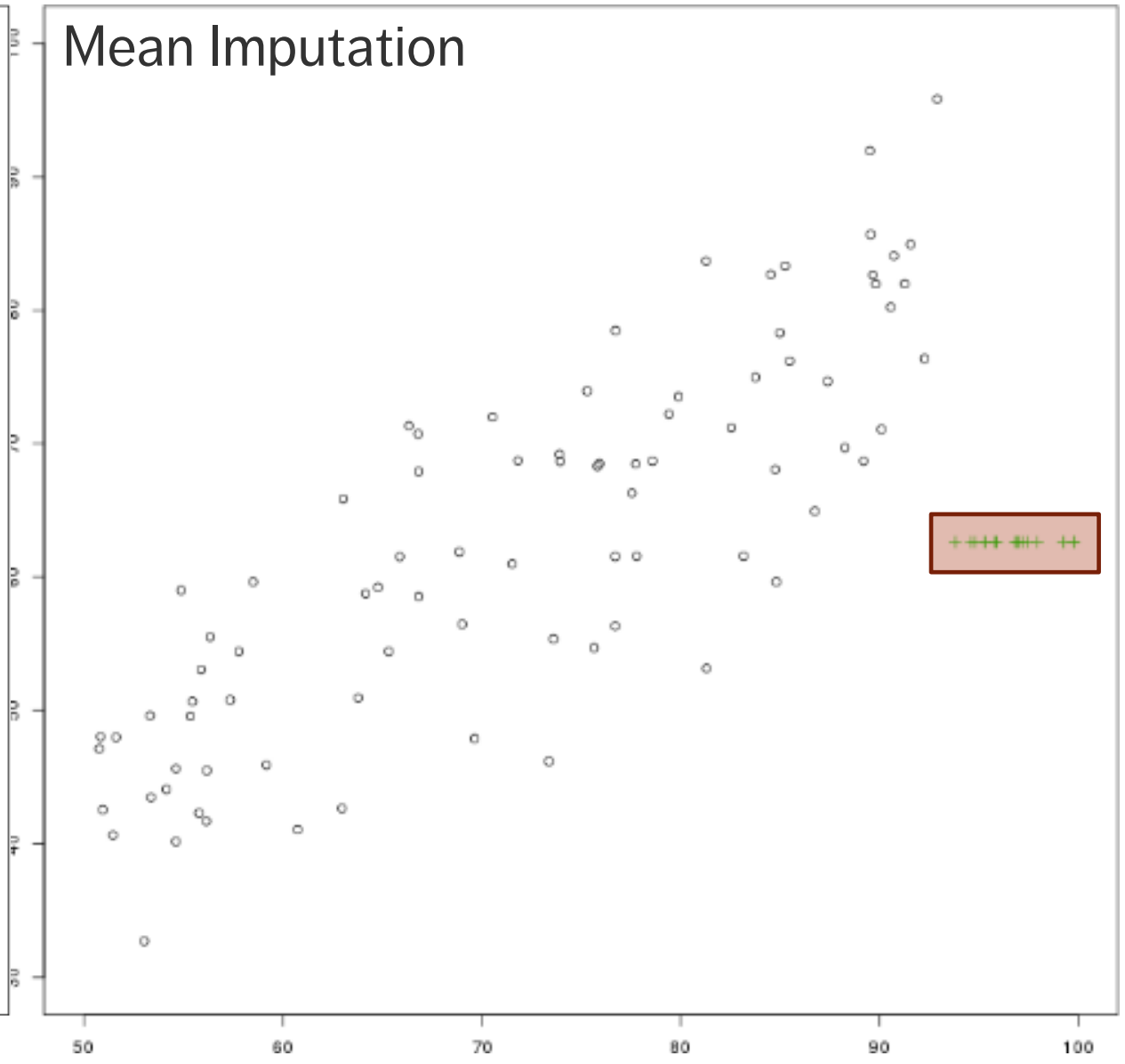
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



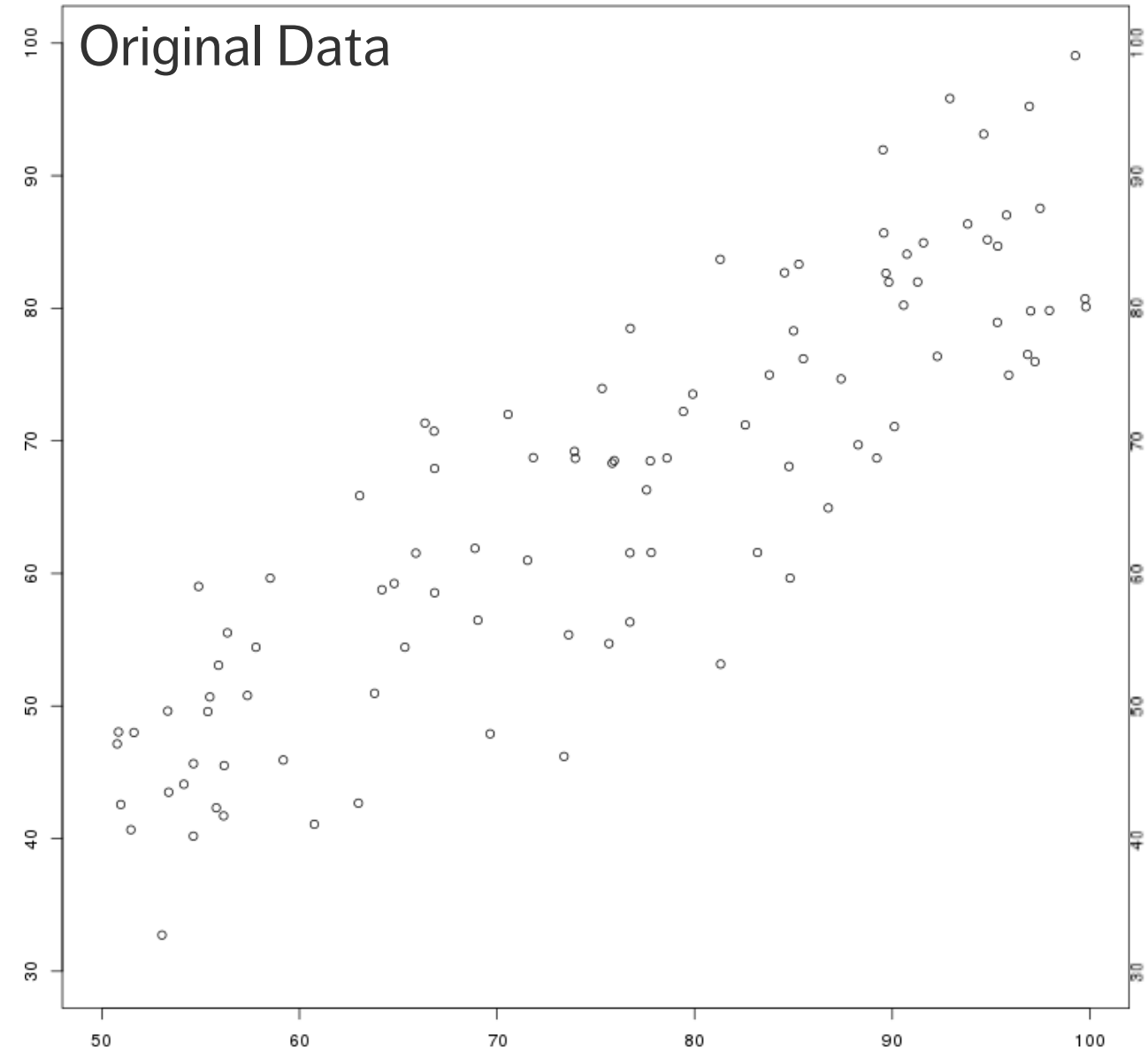


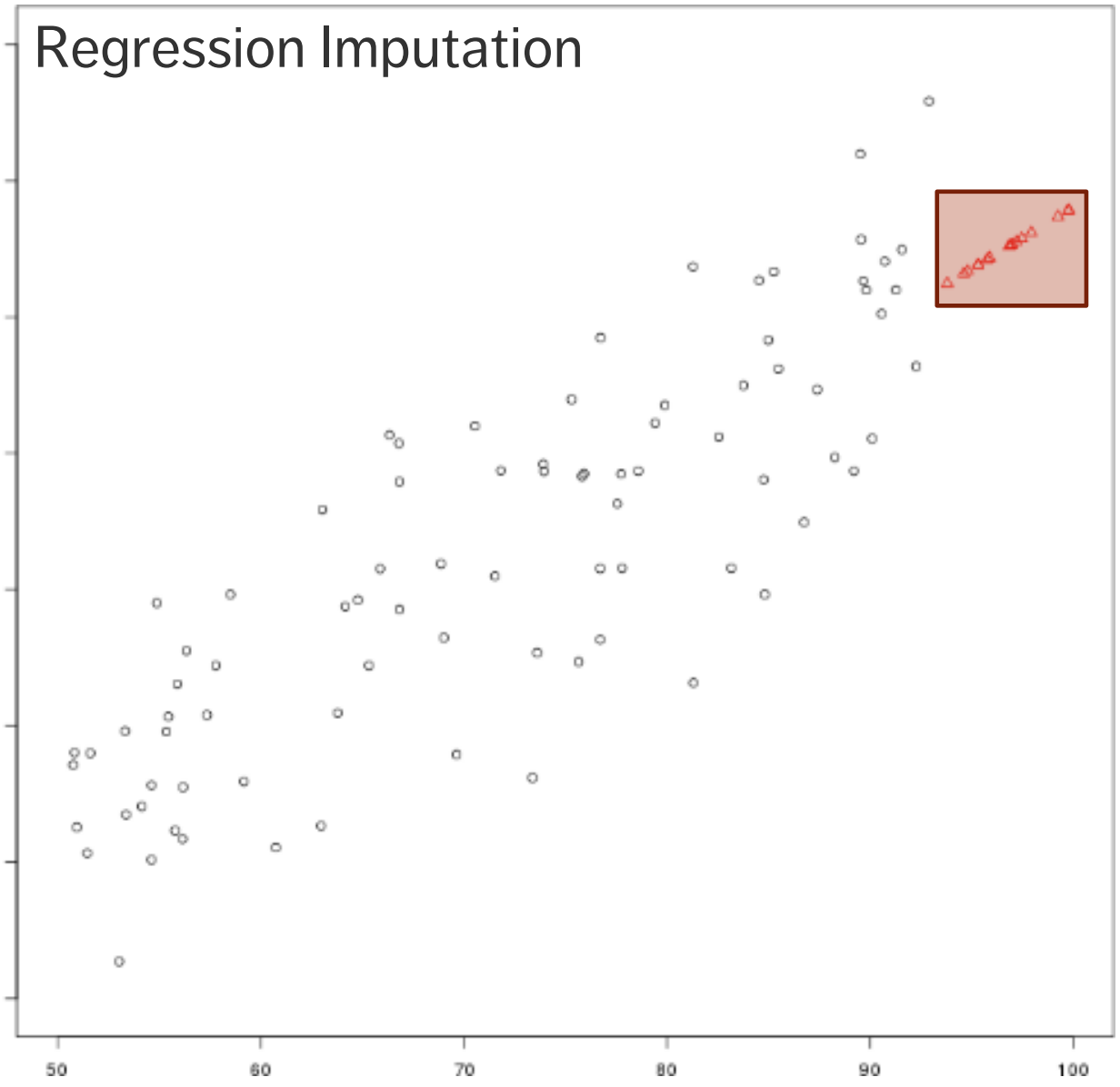
Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.



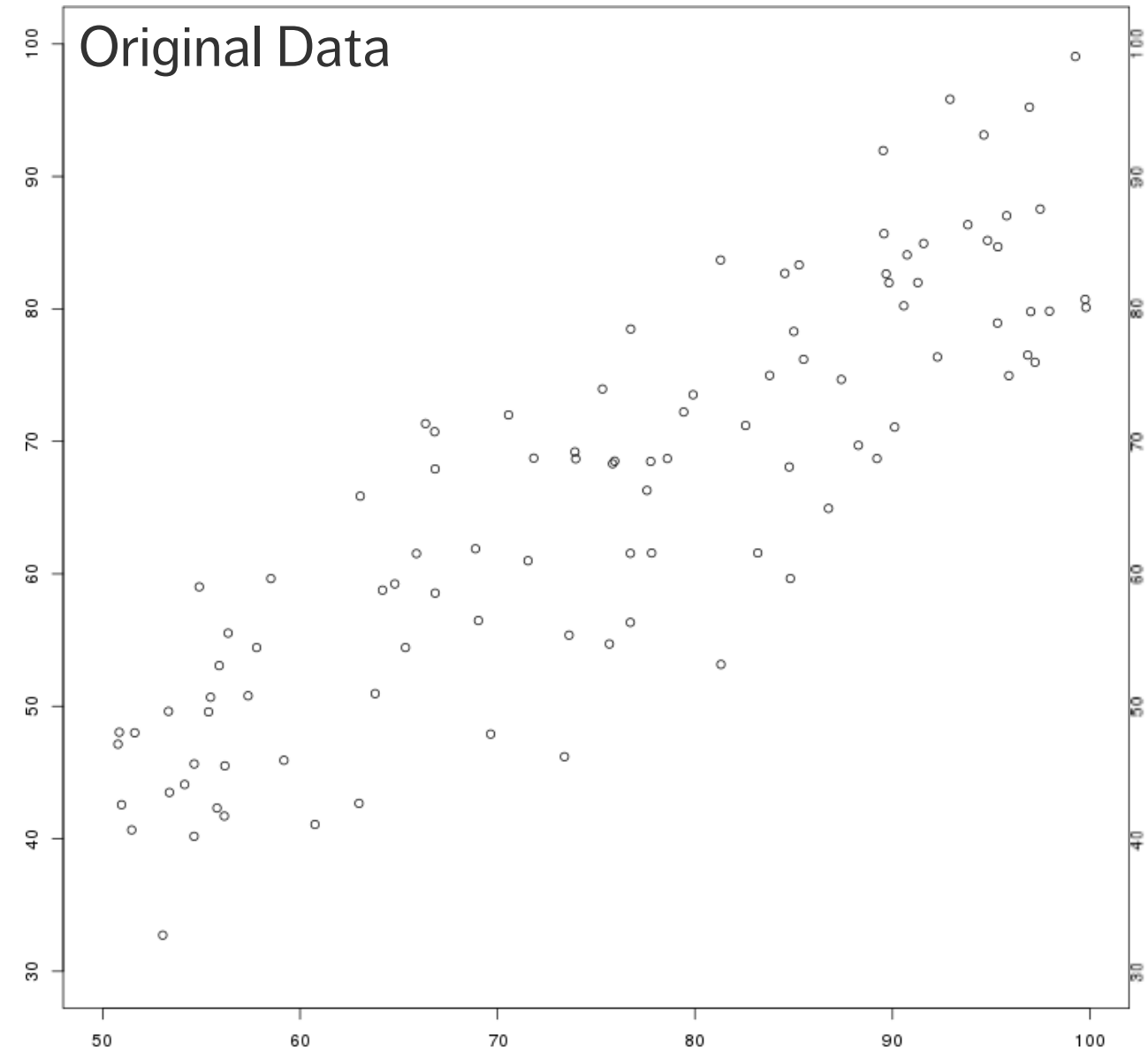


Artificial data: the y values of all points for which $x > 92$ have been erased by mistake.





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:

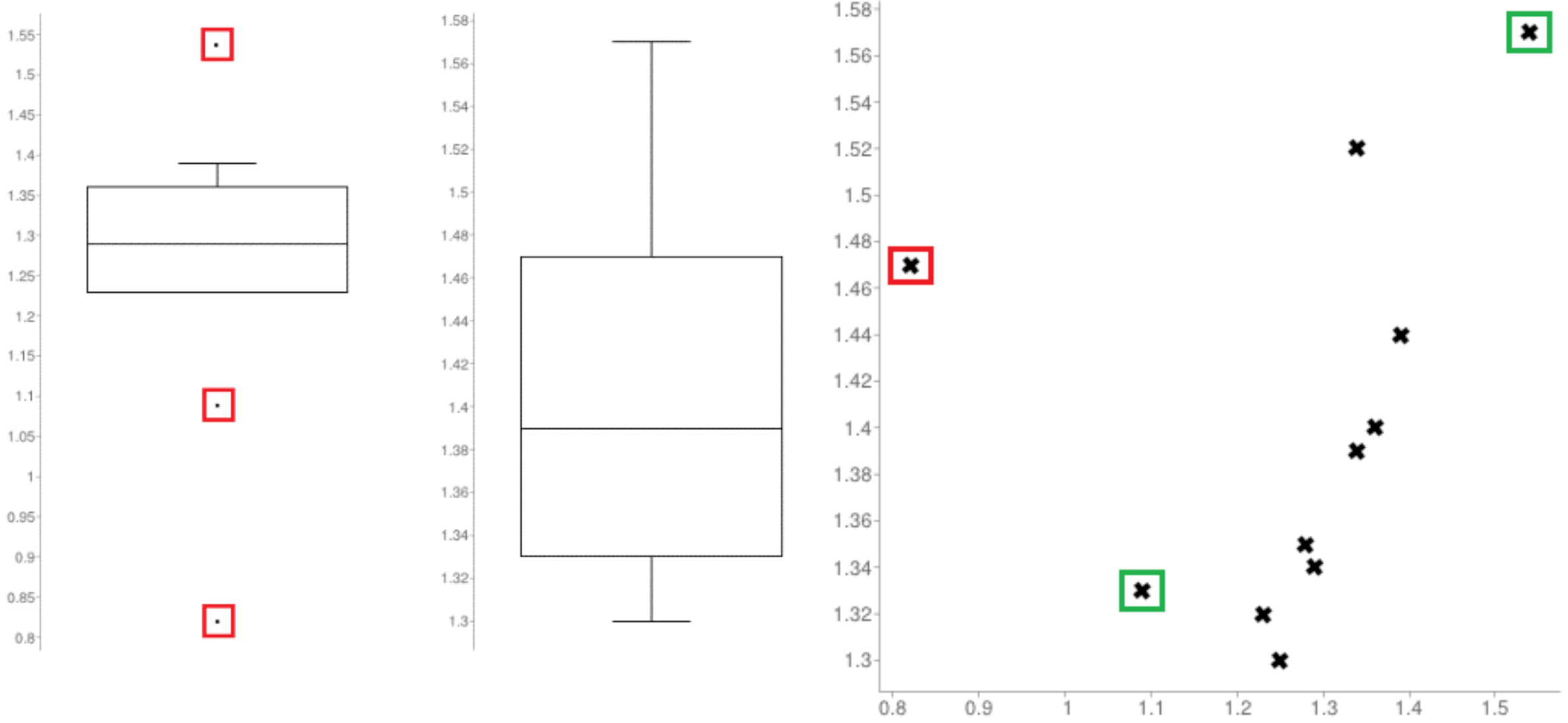
- domain expertise required to tag the data
- classification or regression task (probabilities and inspection rankings)
- rare occurrence problem (more on this later)

Unsupervised methods don't use external information:

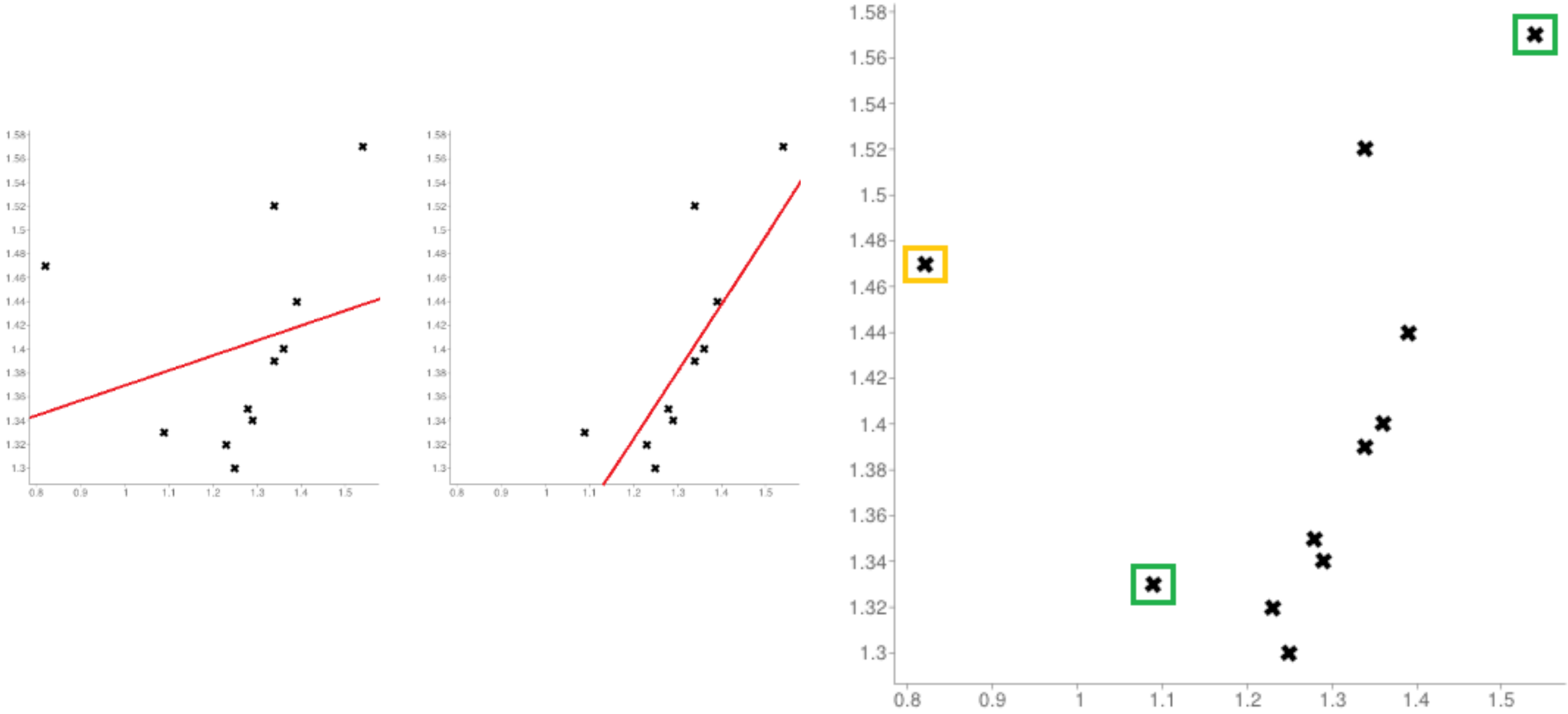
- traditional methods and tests
- can also be seen as a clustering or association rules problem

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.

DATA REDUCTION AND TRANSFORMATIONS

DATA COLLECTION AND DATA PROCESSING

LEARNING OBJECTIVES

Familiarity with the following concepts:

- Dimensionality of data
- Curse of Dimensionality
- Feature selection
- Principal Component Analysis (PCA)
- Data transformation
- Scaling
- Discretization

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.

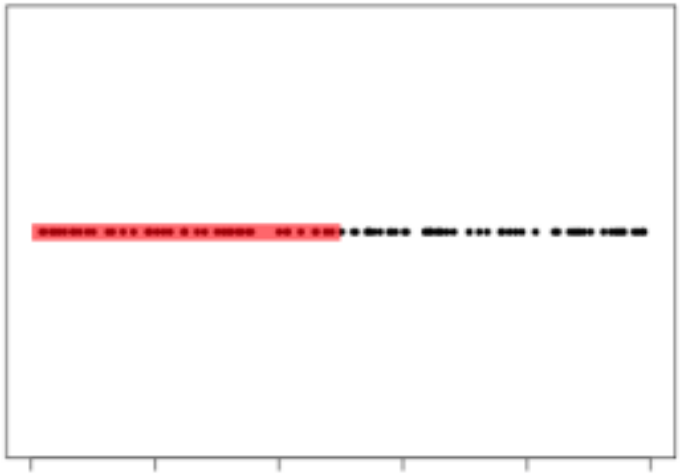
Here the term dimension is an extension of the use of the term to refer to the size of a vector.

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

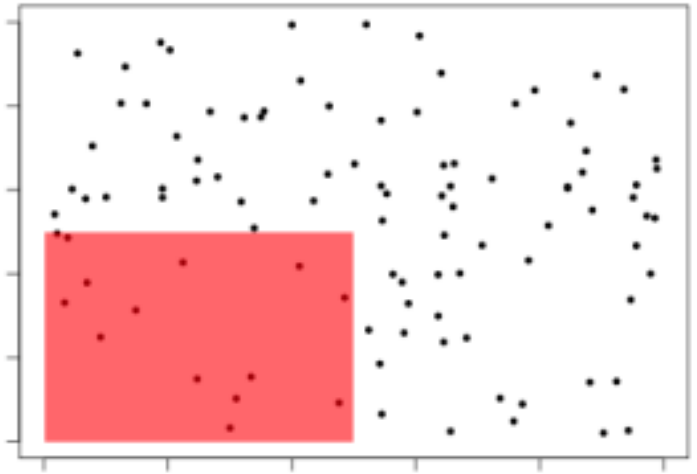
(Note – the term dimension is used differently in business intelligence contexts)

CURSE OF DIMENSIONALITY

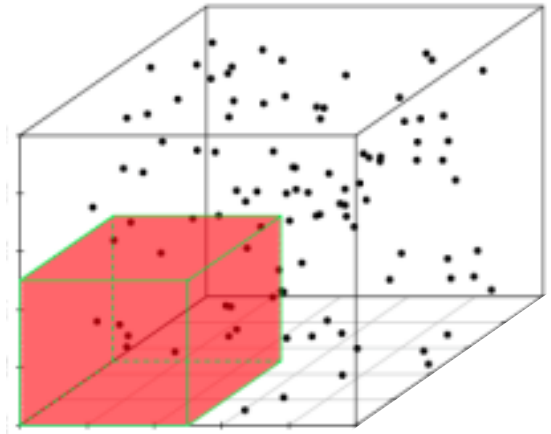
42% of data is captured



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 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 multicolinearity, etc.)
- dimension reduction (# variables ≫ # observations)

Approaches:

- filter vs. wrapper
- unsupervised vs. supervised

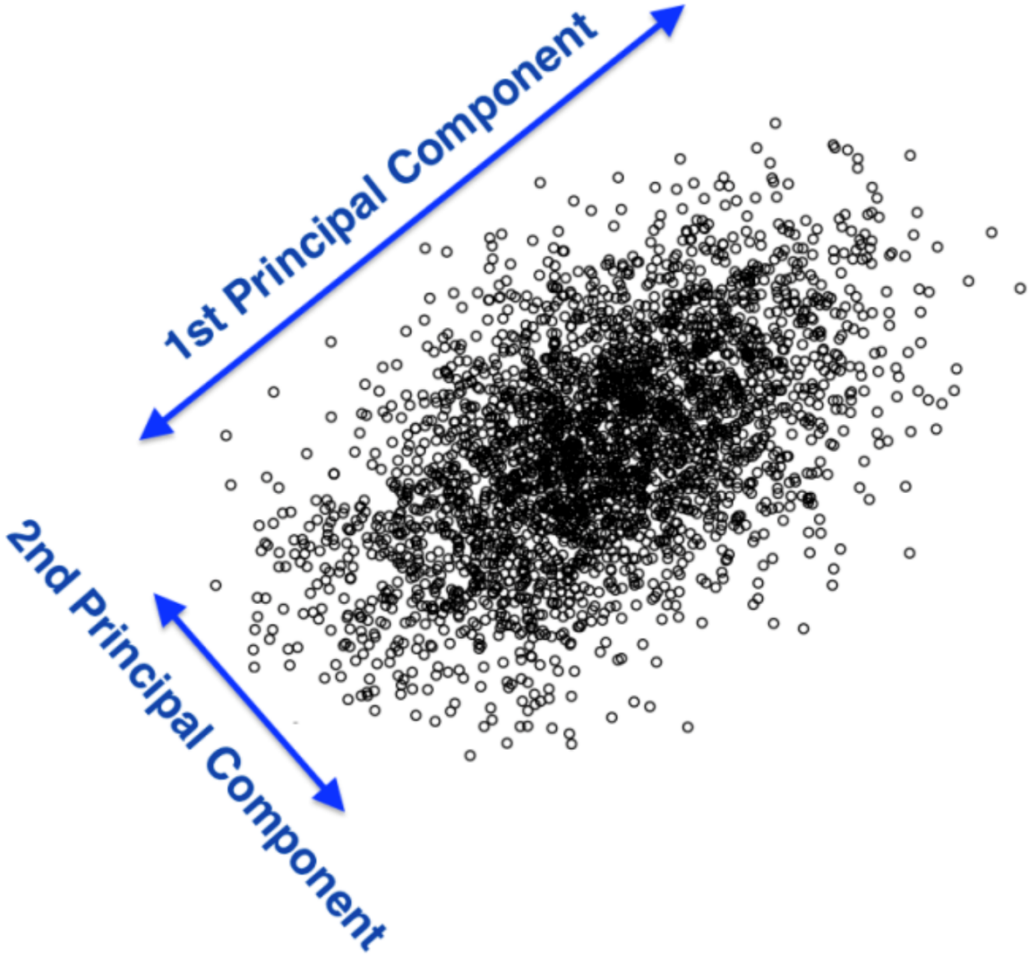
[A. Ng, K. Soo, *Numsense!*]

PRINCIPAL COMPONENT ANALYSIS

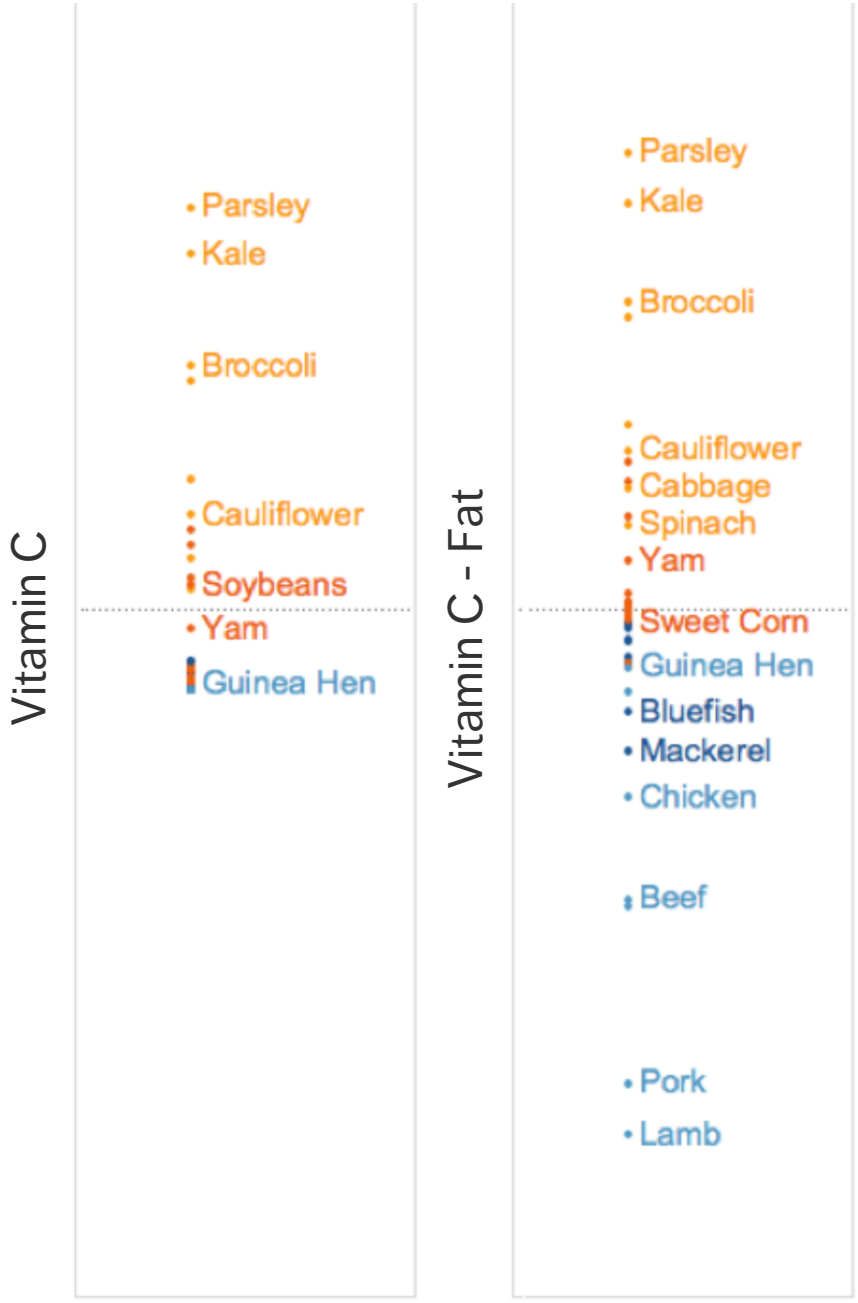
Motivational Example:
Nutritional Content of Food

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.**



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



DIFFERENTIATION

Vitamin C is present in various levels in fruit and vegetables, but not in meats. It **separates** vegetables from meats, and specific vegetables from one another (to some extent), but the meats are **clumped together** (left).

The situation is reversed for *Fat* levels, so the **combination** of vitamin C and fat **separates** vegetables from meats, and **spreads** vegetables and meats (right).

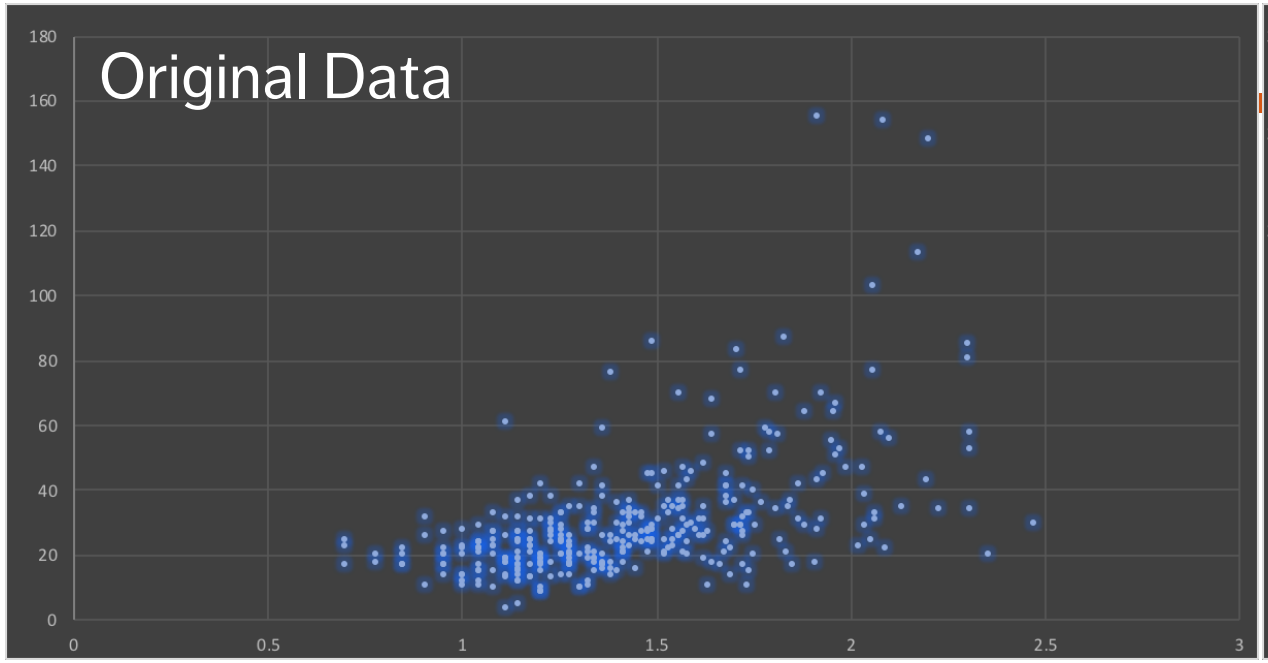
COMMON TRANSFORMATIONS

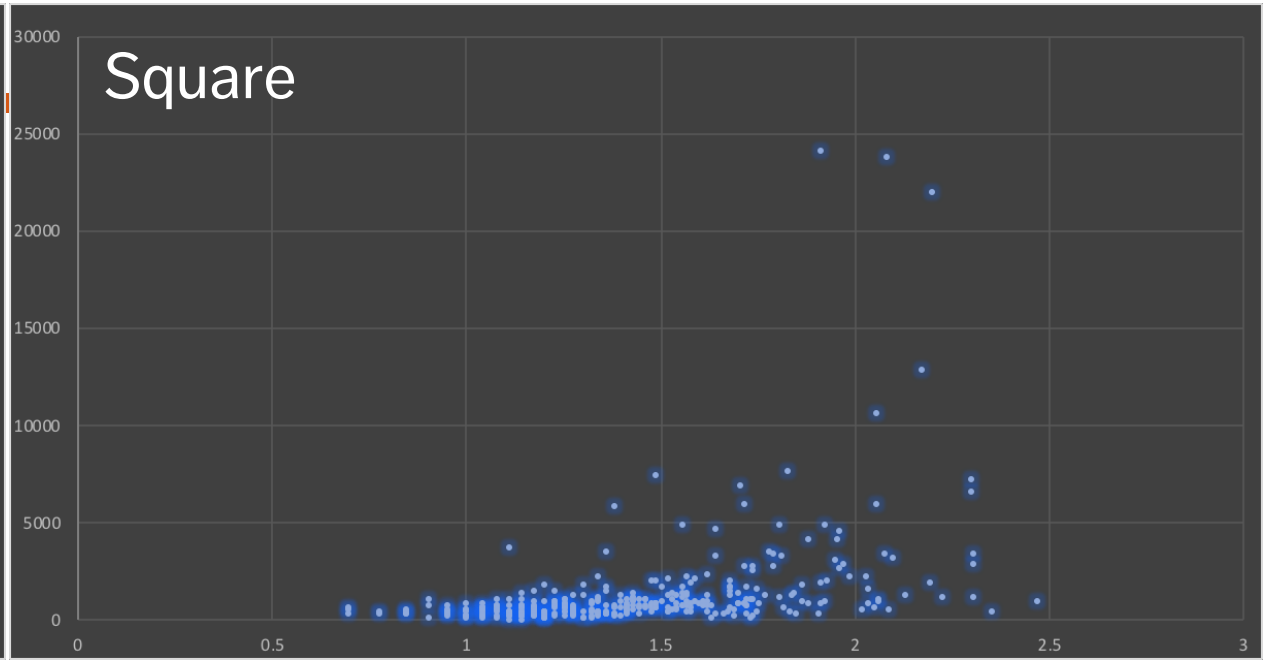
Models sometimes require that certain data assumptions be met (normality of residuals, linearity, etc.).

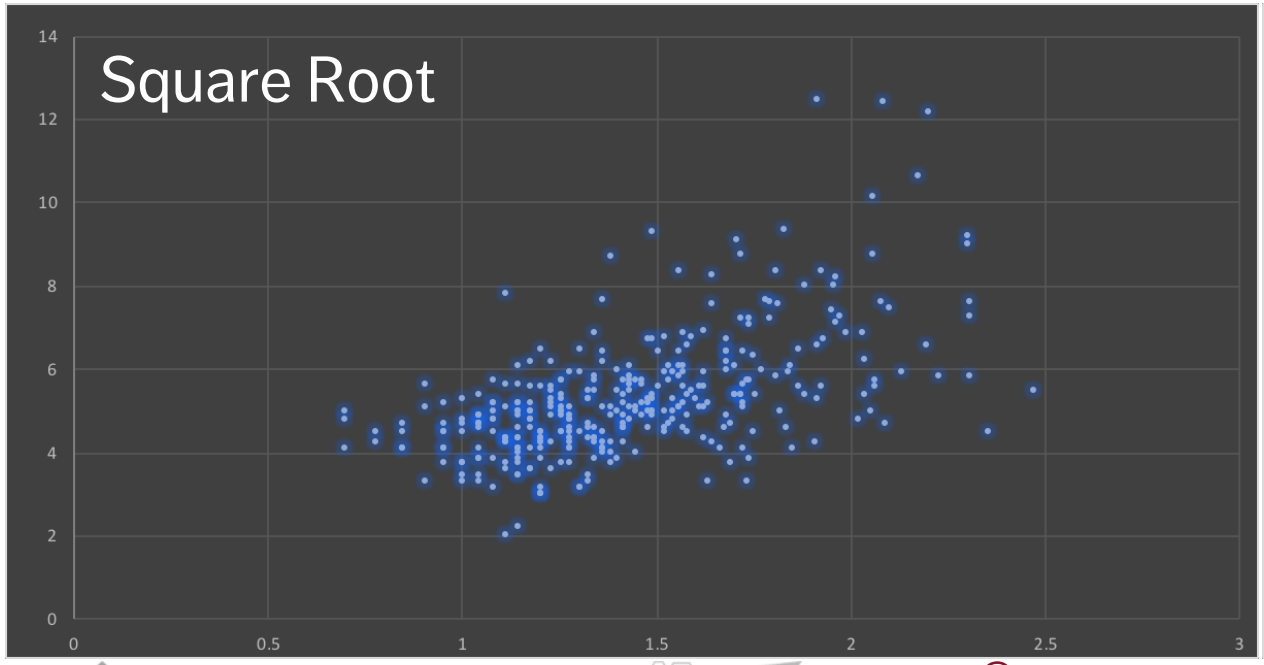
If the raw data does not meet the requirements, we can either

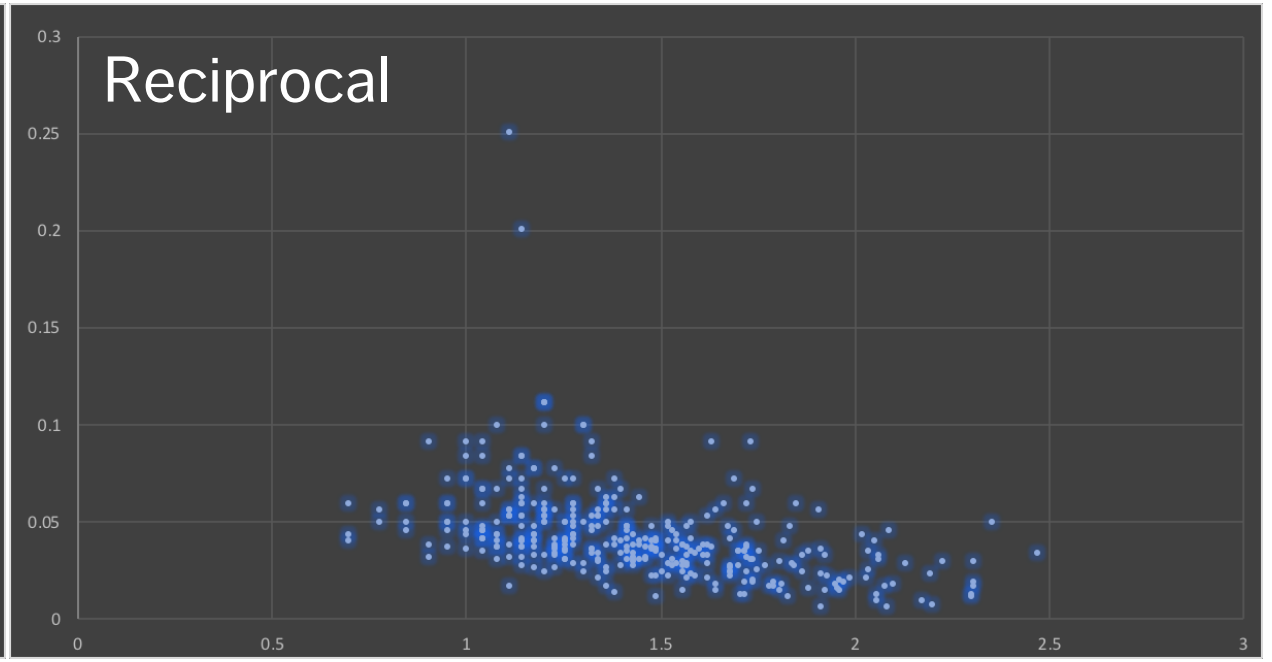
- abandon the model
- attempt to **transform** the data

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


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


Square
30000
25000
20000
15000
10000
5000
0
0
0.5
1
1.5
2
2.5
3


Square Root
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3


Reciprocal
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.5
1
1.5
2
2.5
3

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 COLLECTION AND DATA PROCESSING

Martin: Data is messy.
Allison: Even when it's been cleaned?
Martin: Especially when it's been cleaned.

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

LEARNING OBJECTIVES

Understand common sources of data error and types of potential issues

Understand difference between accuracy and precision

Understand, at a high level, some techniques for detecting data issues

Familiarity with some examples of data validity issues

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.

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

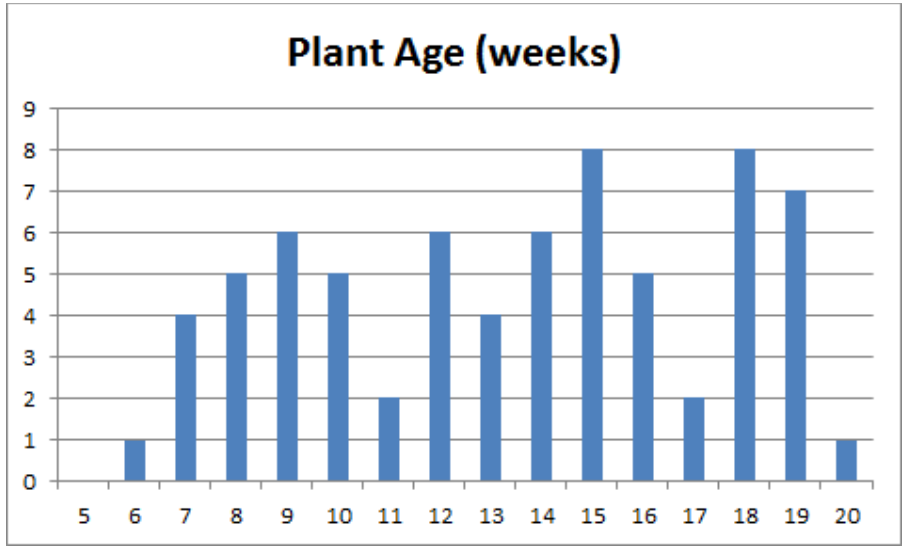
Potentially invalid entries can be detected with the help of:

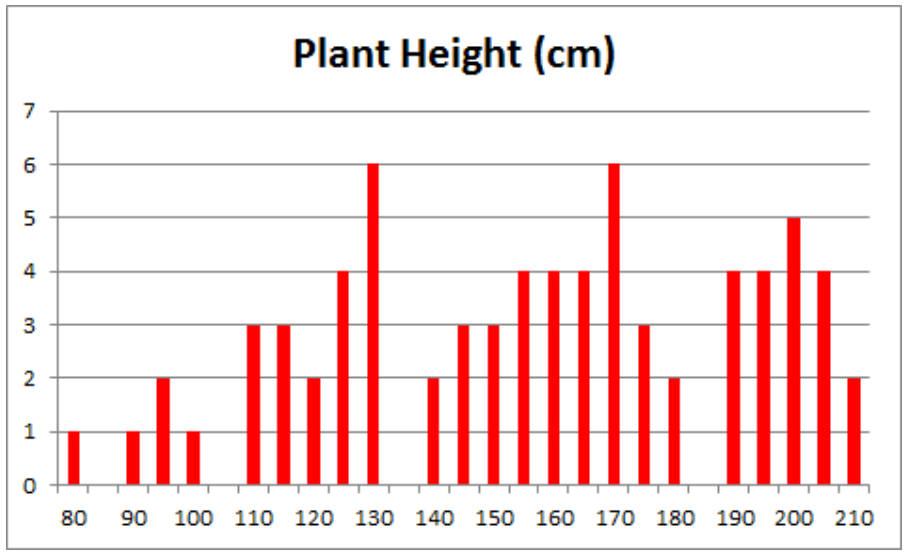
- **Univariate Descriptive Statistics**
  count, range, z-score, mean, median, standard deviation, logic check
- **Multivariate Descriptive Statistics**
  n-way table, logic check
- **Data Visualization**
  scatterplot, scatterplot matrix, histogram, joint histogram, etc.

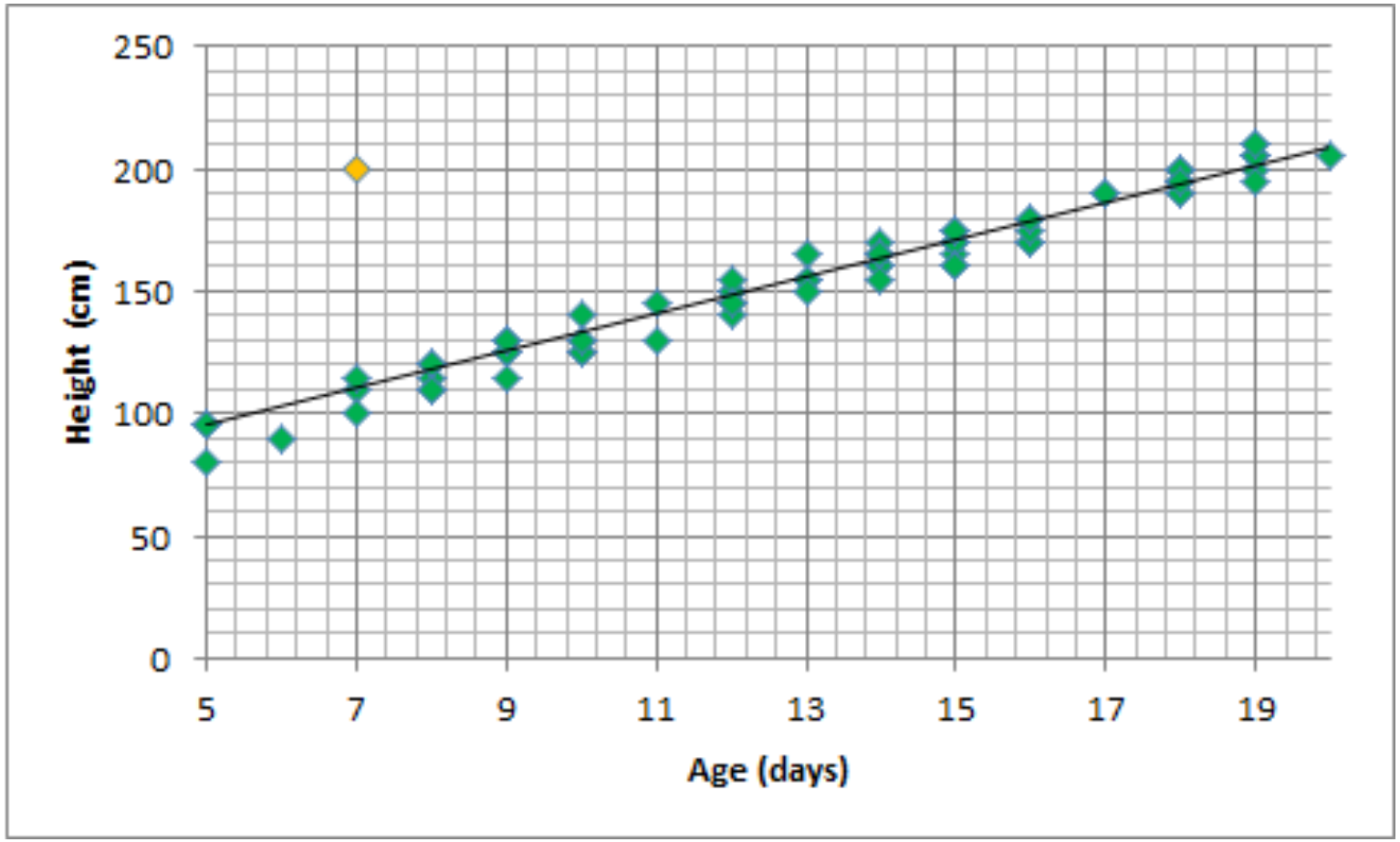
This step might allow for the identification of potential outliers.

Failure to detect invalid entries ≠ all entries are valid.

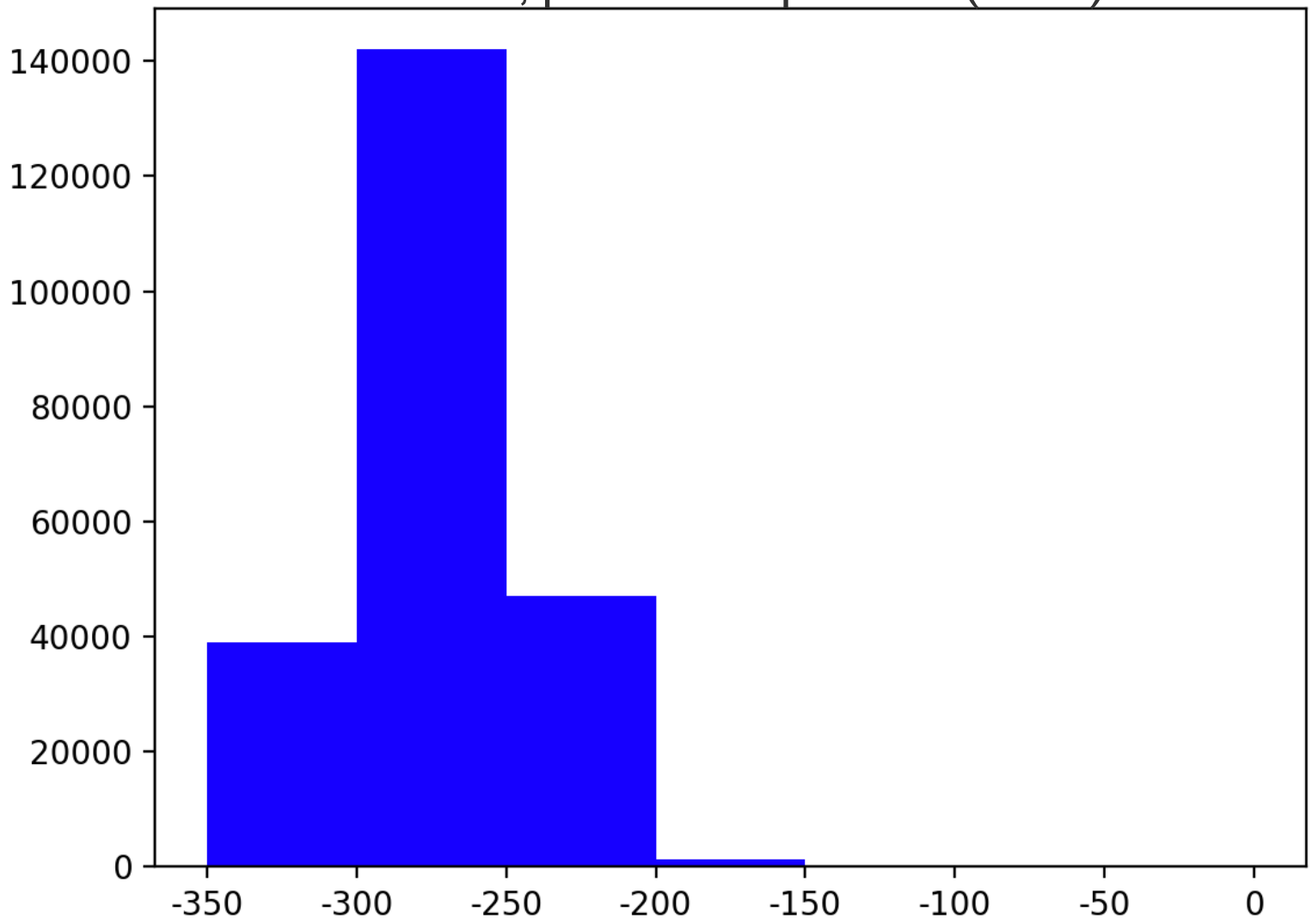
Small numbers of invalid entries recoded as “missing.”


Plant Age (weeks)
0
1
2
3
4
5
6
7
8
9
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

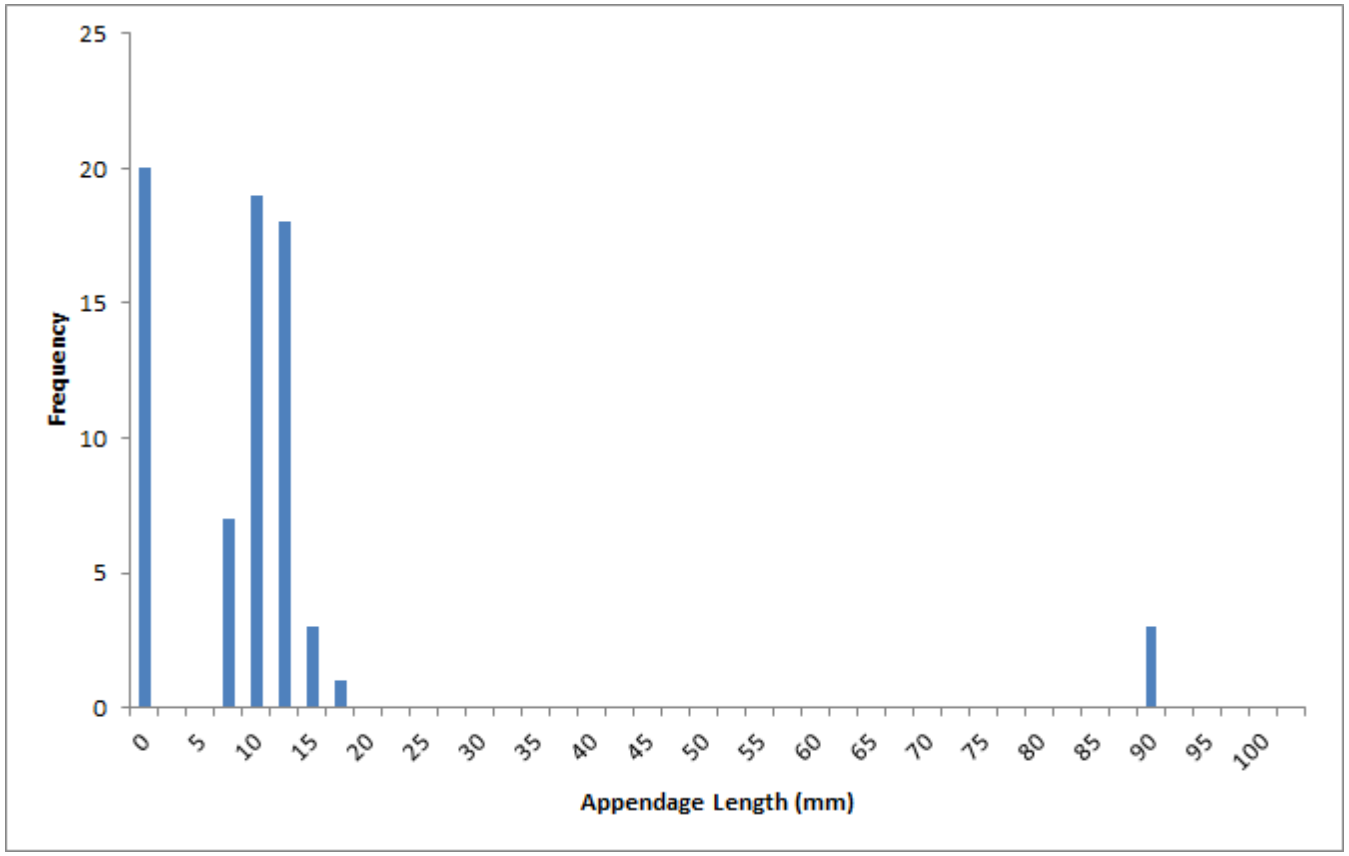

Plant Height (cm)
0
1
2
3
4
5
6
7
80 90 100 110 120 130 140 150 160 170 180 190 200 210


Height (cm)
0
50
100
150
200
250
5 7 9 11 13 15 17 19
Age (days)

Time of arrivals at screening station, prior to departure (mins)



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.

REFERENCES

DATA COLLECTION AND DATA PROCESSING

REFERENCES

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

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

Orchard, T. and Woodbury, M. [1972], *A Missing Information Principle: Theory and Applications,* Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Berkeley.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. and Solenberger, P. [2001], *A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models*, Survey Methodology, v.27, n.1, pp.85-95, Statistics Canada, Catalogue no. 12-001.

REFERENCES

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

Wikipedia entry for Data Cleansing

Wikipedia entry for Imputation

Wikipedia entry for Outliers

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

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

REFERENCES

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

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

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

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

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

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

REFERENCES

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

Steiger, J.H. , Transformations to Linearity, lecture notes.

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

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

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

REFERENCES

Robnik-Sikonja, M., Savicky, P., CORElearn package documentation, v1.51.2, CRAN.

Ng, A., Soo, K., Principal Component Analysis Tutorial, June 15, 2016.

Principal component analysis, on Wikipedia

Hastie, T., Tibshirani, R., Friedman, J. [2009], *The Elements of Statistical Learning (2nd ed.)*, ch.2, Springer.

Smith, L.I. [2002], A Tutorial on Principal Component Analysis

Shlens, J. [2014], A Tutorial on Principal Component Analysis, arXiv.org

Nonlinear dimensionality reduction, on Wikipedia

J. Leskovec, A. Rajaraman, J. Ullman [2015] *Mining of Massive Datasets*, Cambridge University Press.