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, *The Great Balancing Act*

IDLEWYLD Sysabee DAVHILL

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

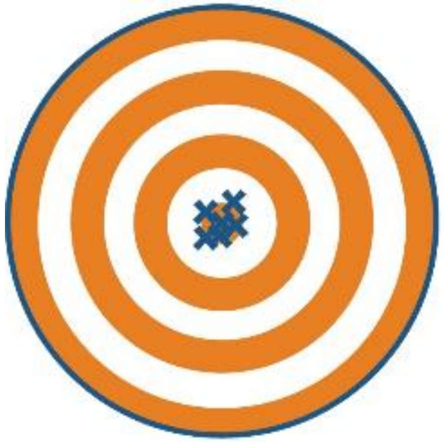
Familiarity with some examples of data validity issues

SOUND DATA

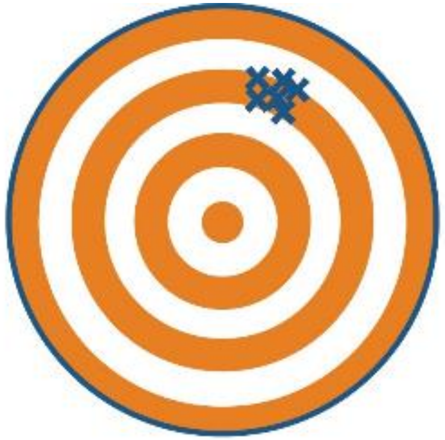
The ideal dataset will have as few issues as possible with:

- **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?

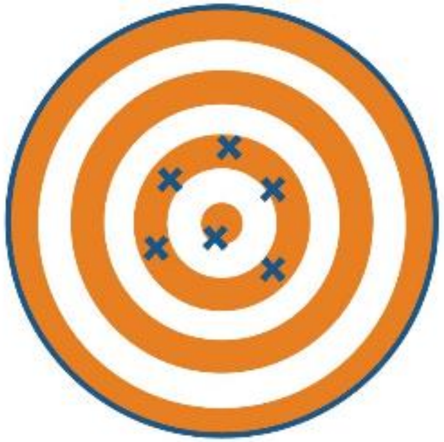
Checking for data quality issues at an early stage can save headaches later in the analysis.



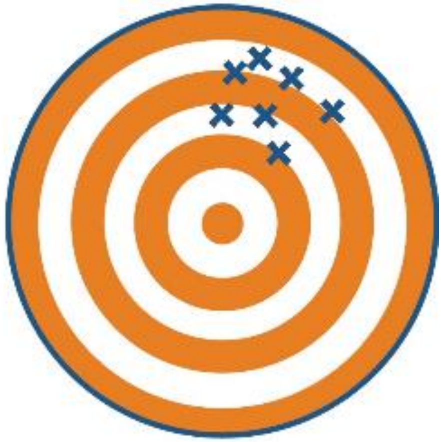
accurate and precise



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

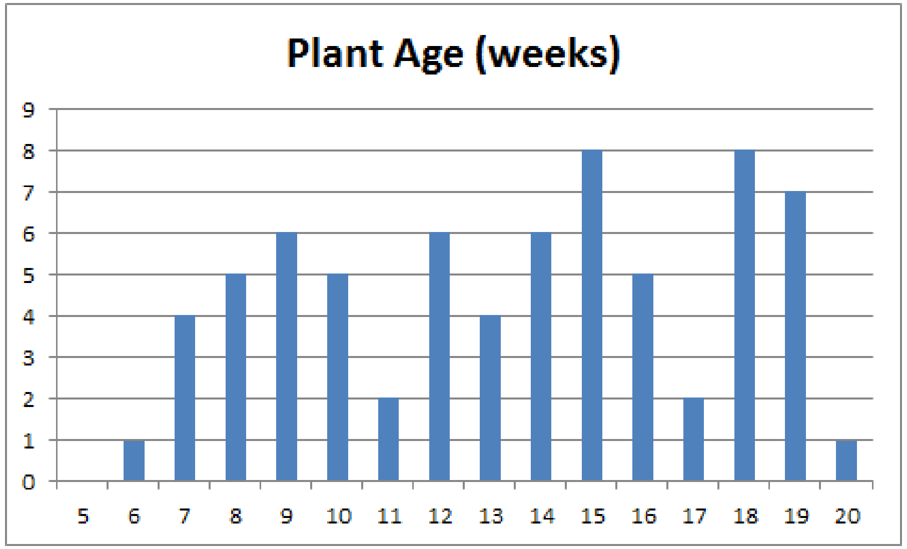
Potentially invalid entries can be detected with the help of:

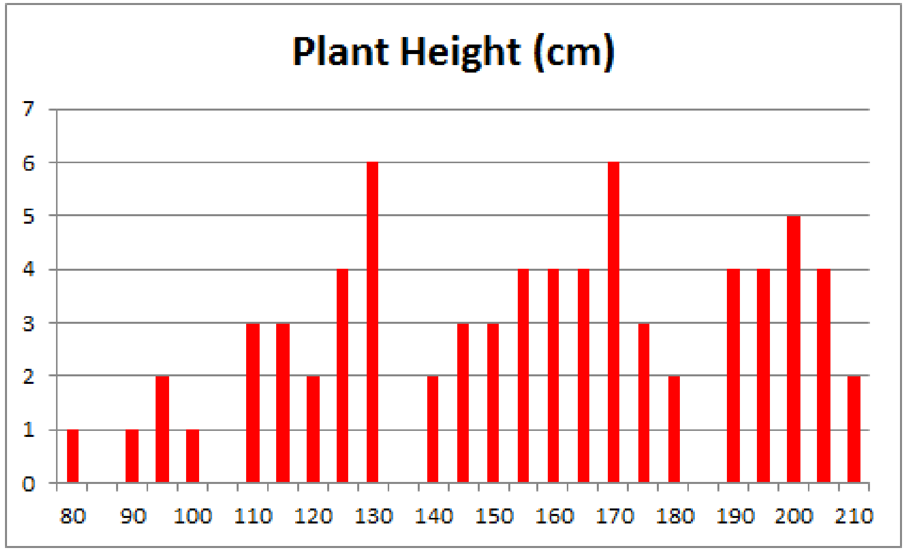
- **Univariate Descriptive Statistics**
  count, range, z-score, mean, median, standard deviation, logic check
- **Multivariate Descriptive Statistics**
  n-way table, logic check
- **Data Visualization**
  scatterplot, scatterplot matrix, histogram, joint histogram, etc.

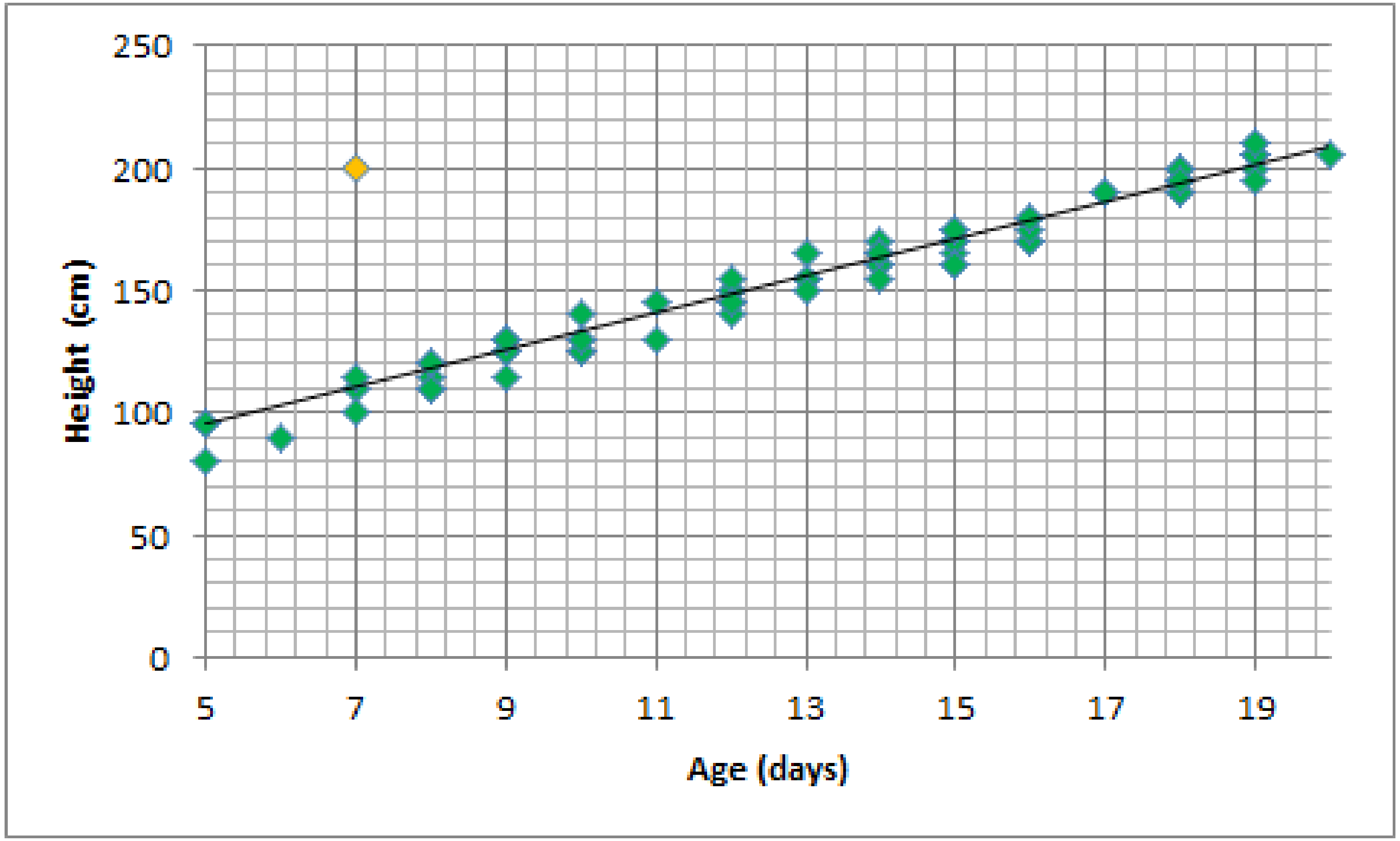
This step might allow for the identification of potential outliers.

Failure to detect invalid entries $\neq$ all entries are valid.

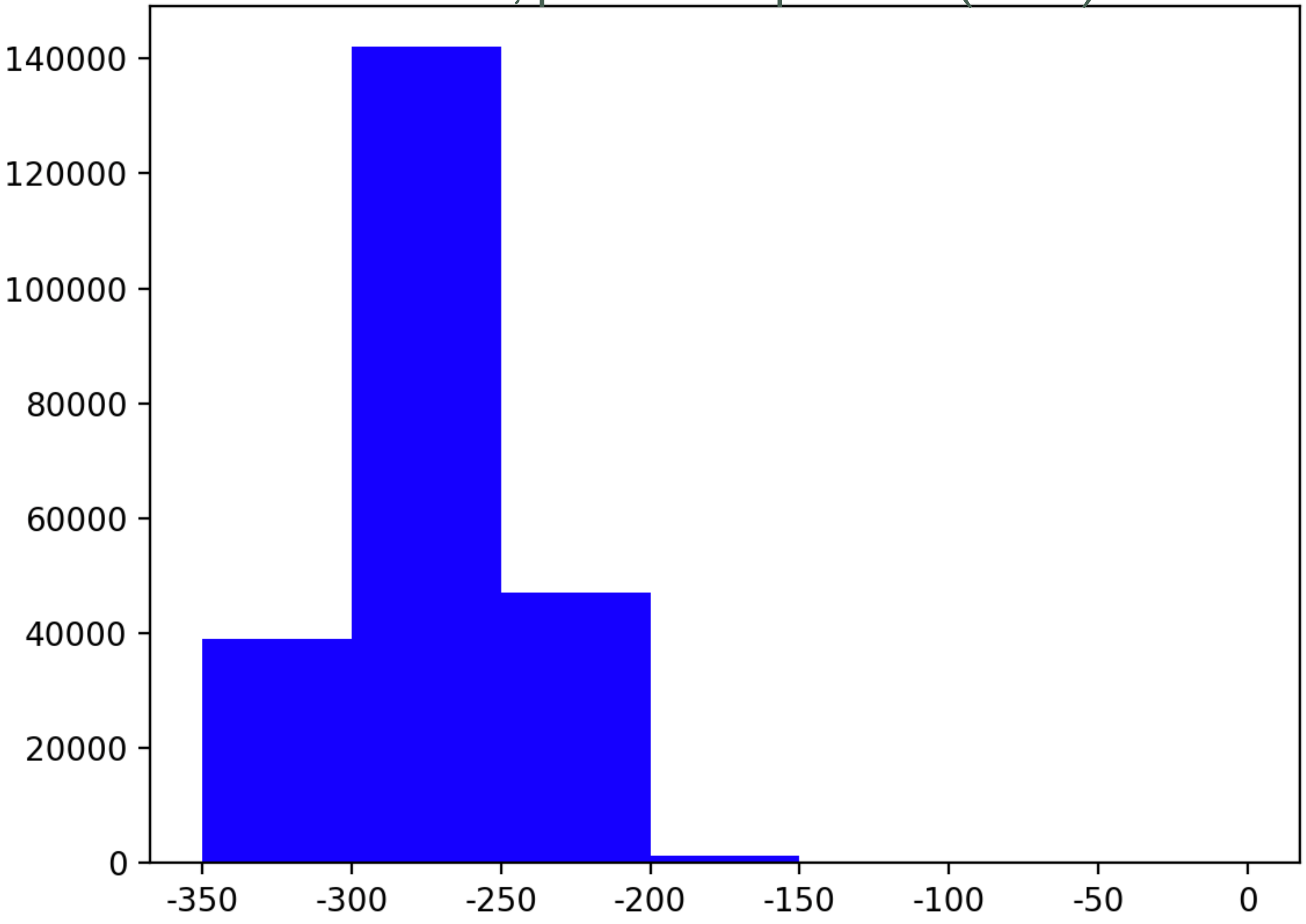
Small numbers of invalid entries recoded as "missing."


Plant Age (weeks)

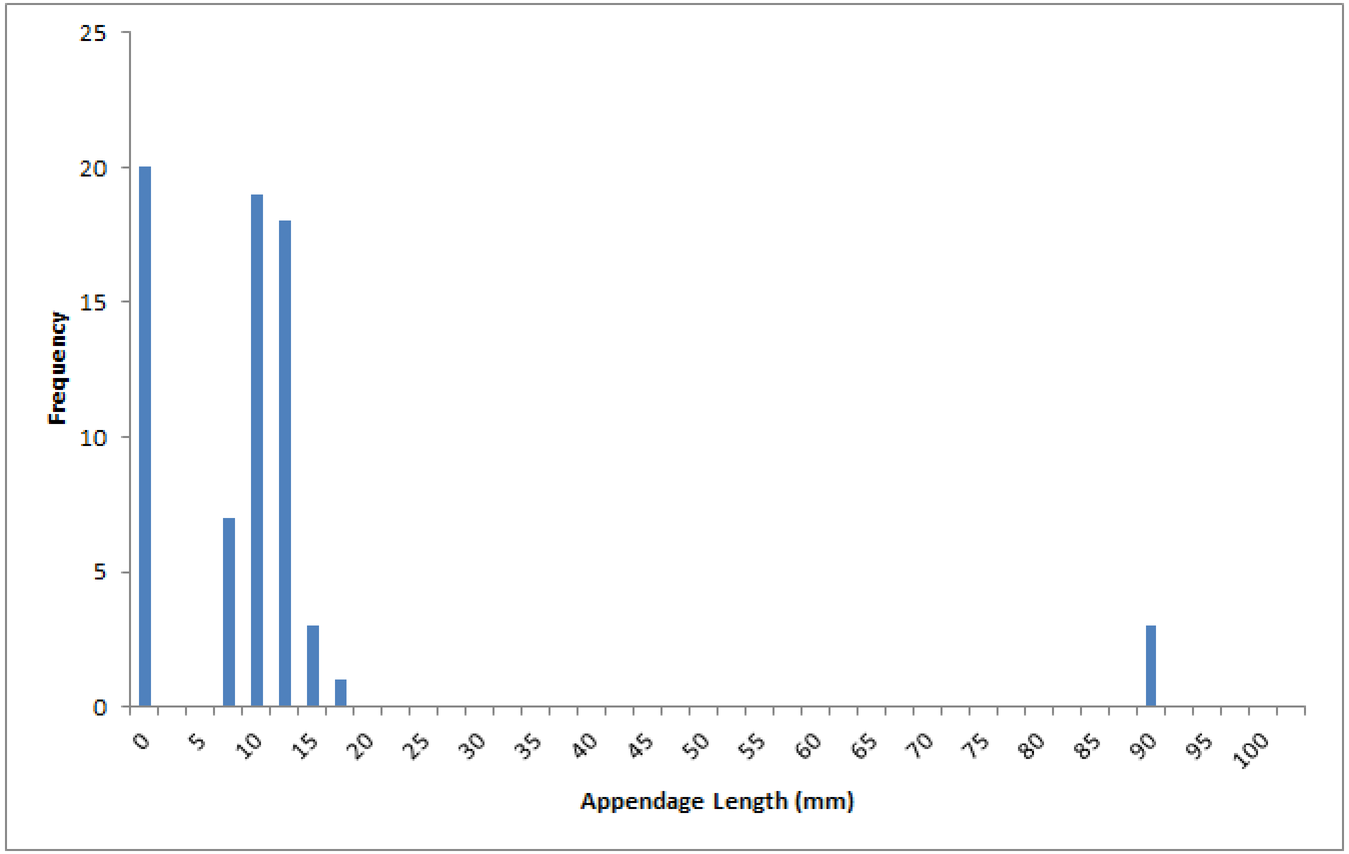

Plant Height (cm)



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.