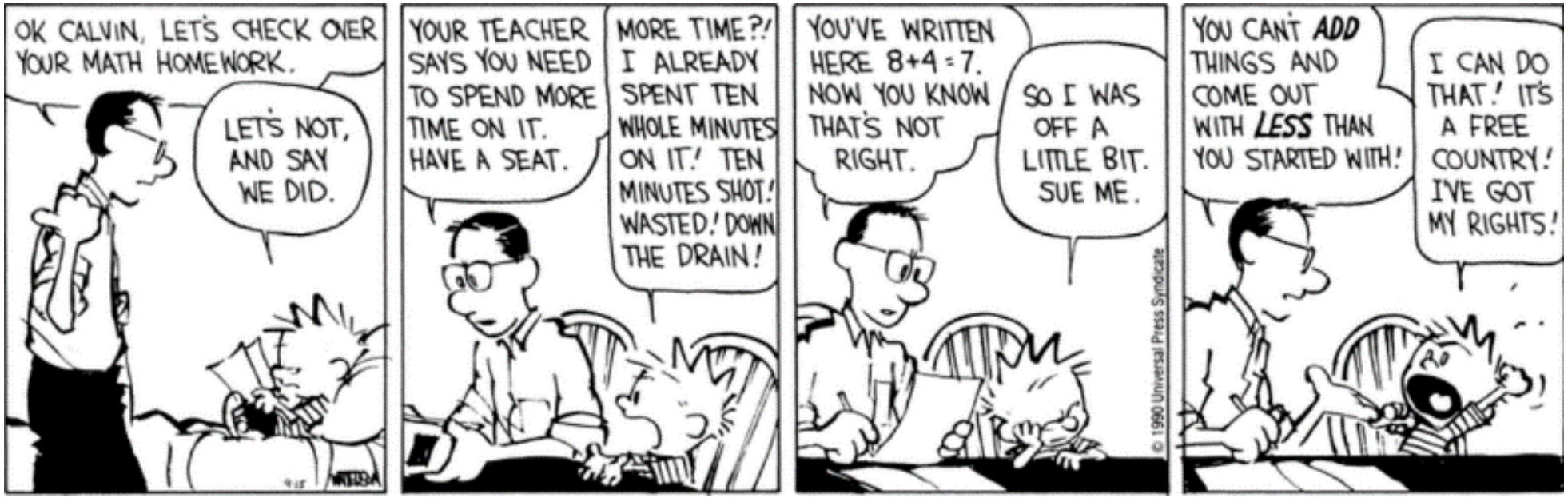
DATA ANALYSIS

INVALID ENTRIES

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



COMMON ERROR SOURCES

When dealing with **legacy**, **inherited** or **combined** datasets (i.e., 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

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

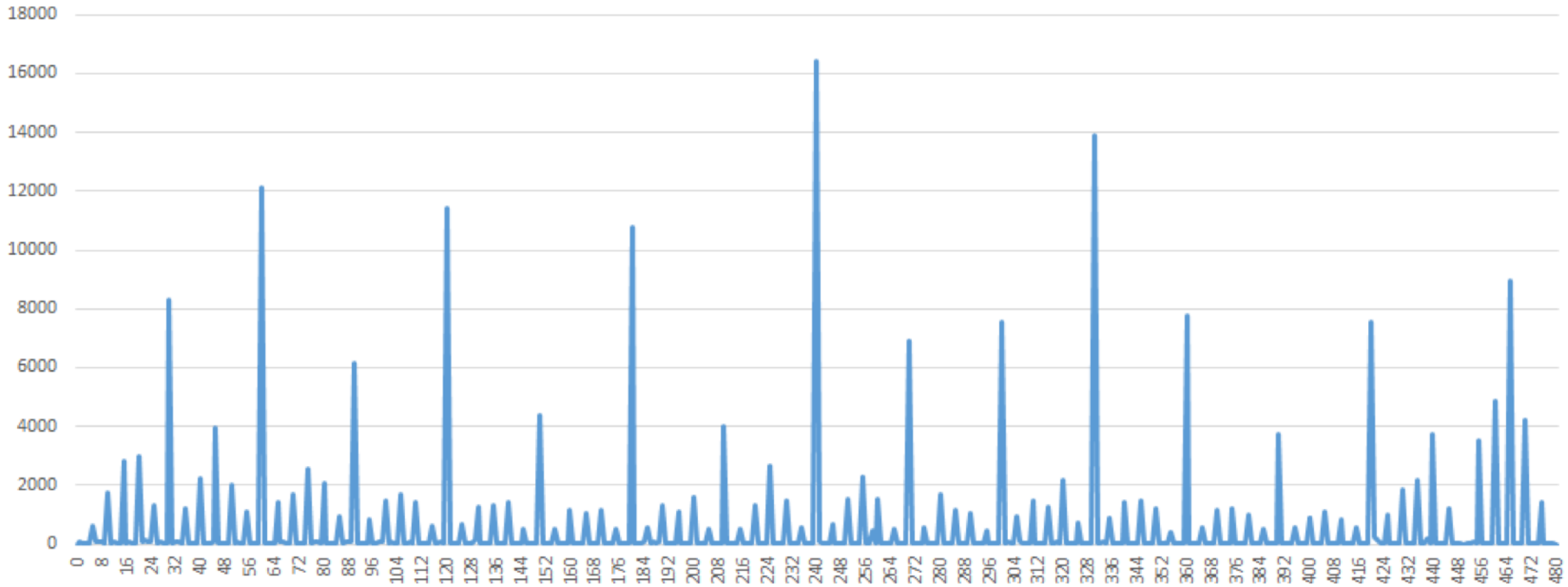
Univariate tests do not always tell the **whole** story.

This step might allow for the identification of potential outliers.

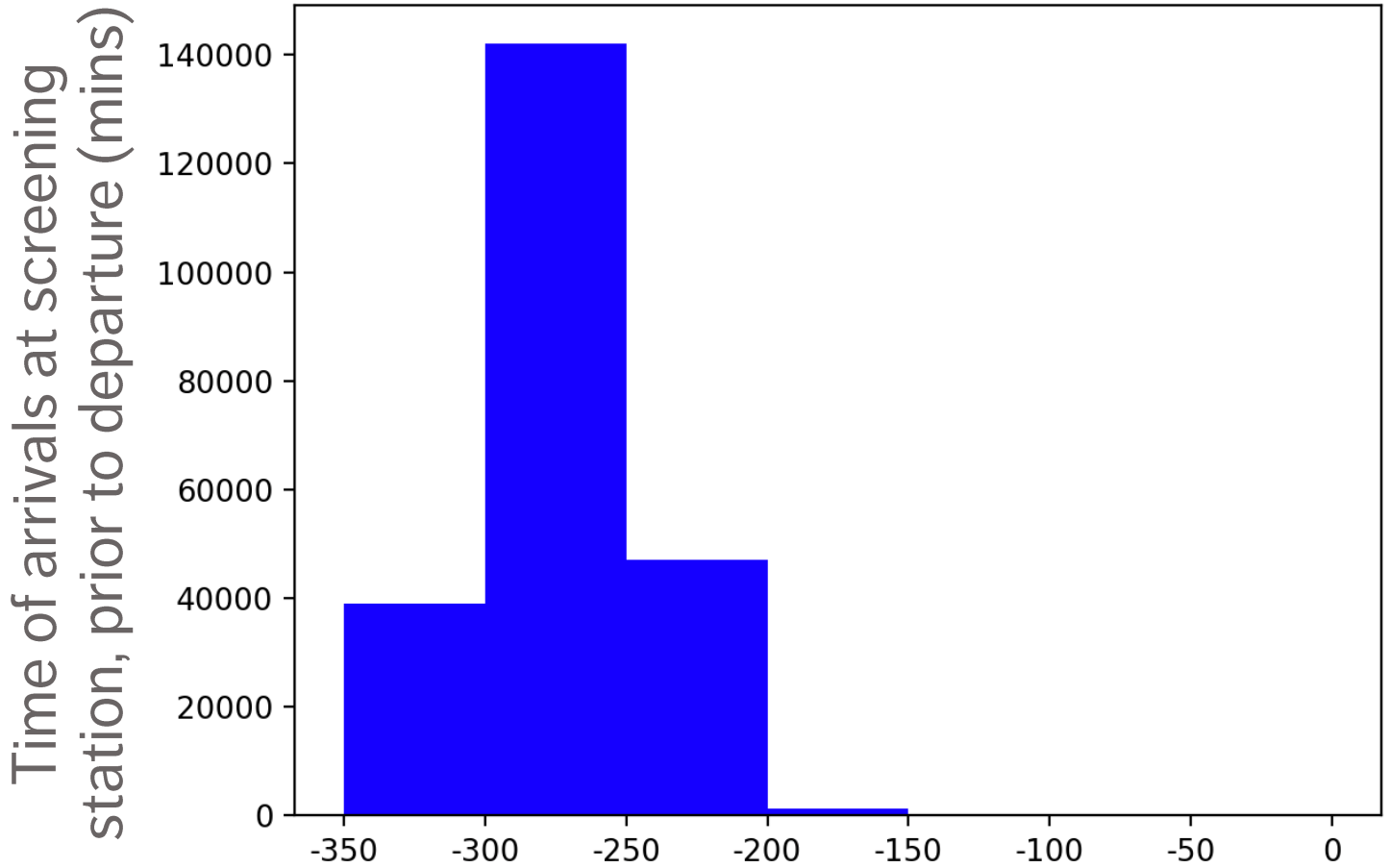
Failure to detect invalid entries ≠ all entries are valid.

Small numbers of invalid entries recoded as “missing.”

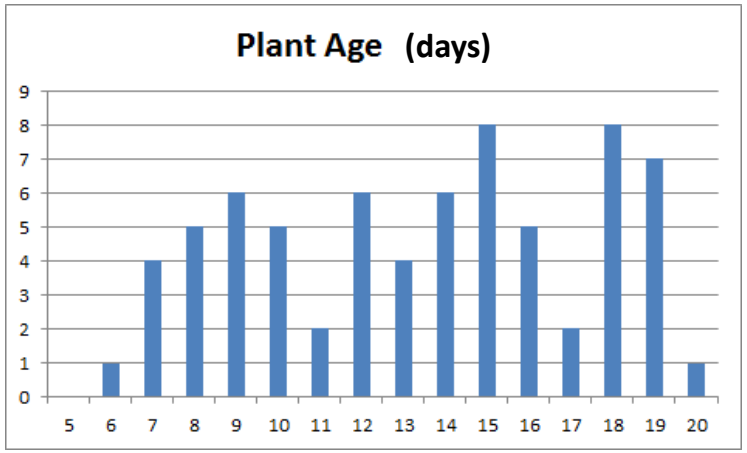
DETECTING INVALID ENTRIES

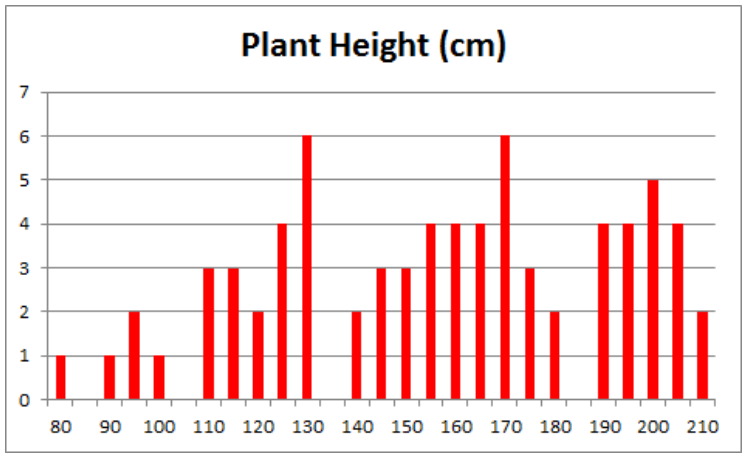


DETECTING INVALID ENTRIES

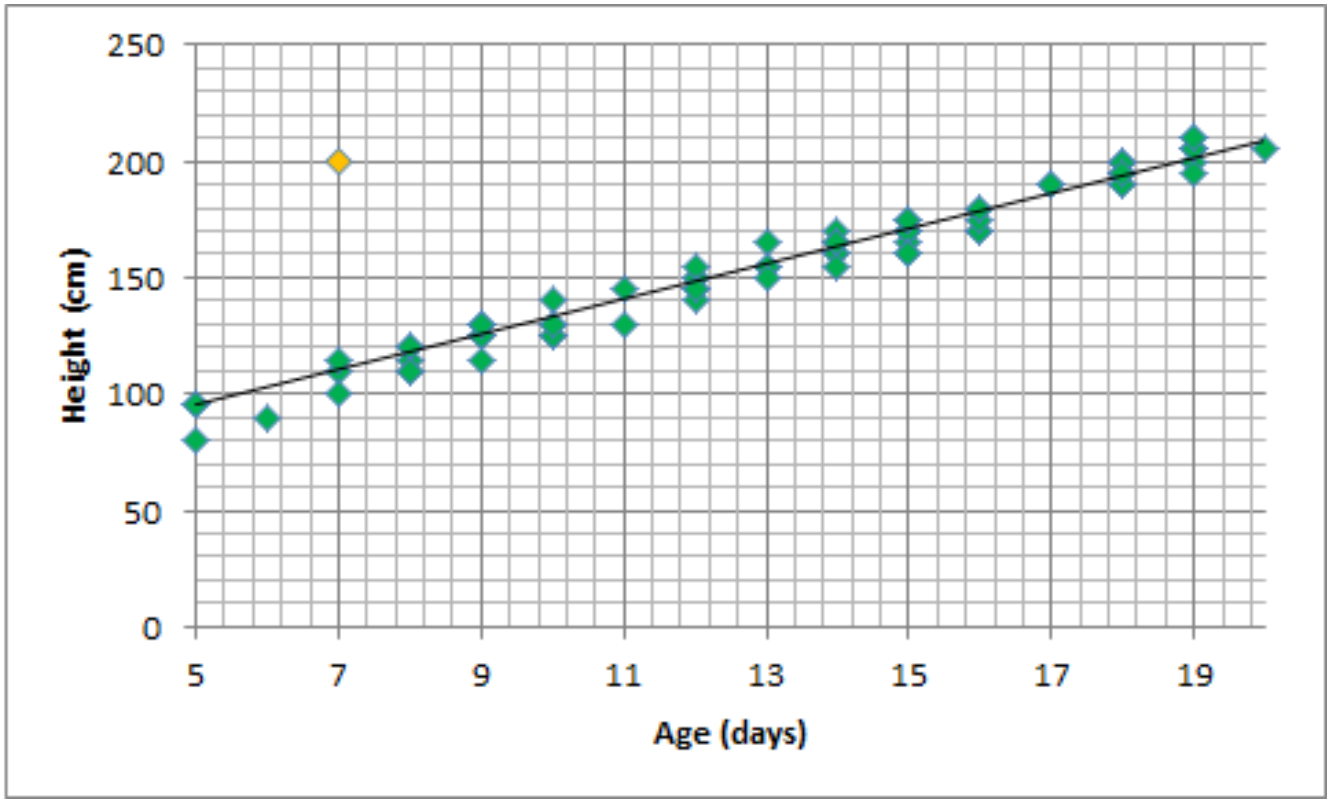


DETECTING INVALID ENTRIES



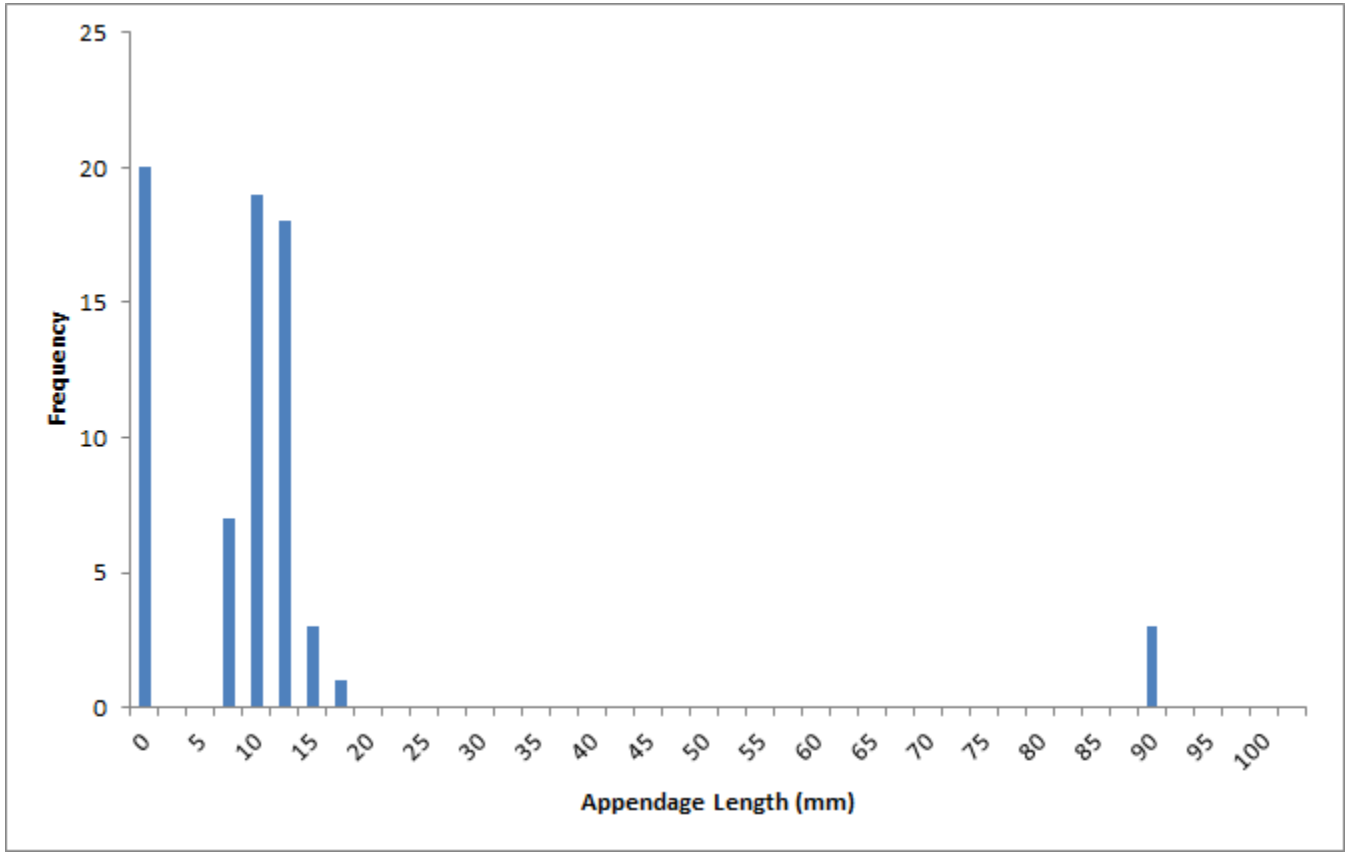


vs.



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



DATA ANALYSIS

INVALID ENTRIES