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

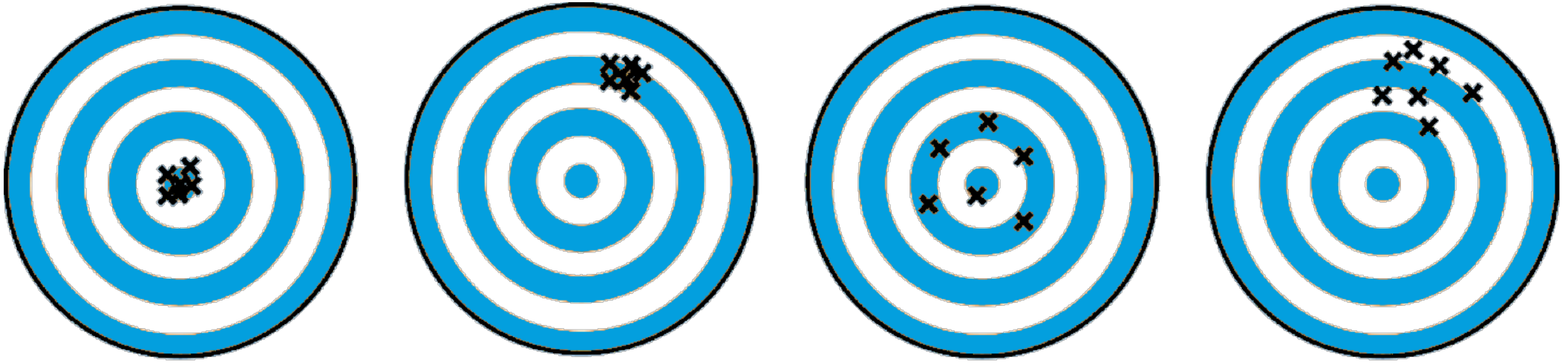
Data Soundness

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

Session 3

[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

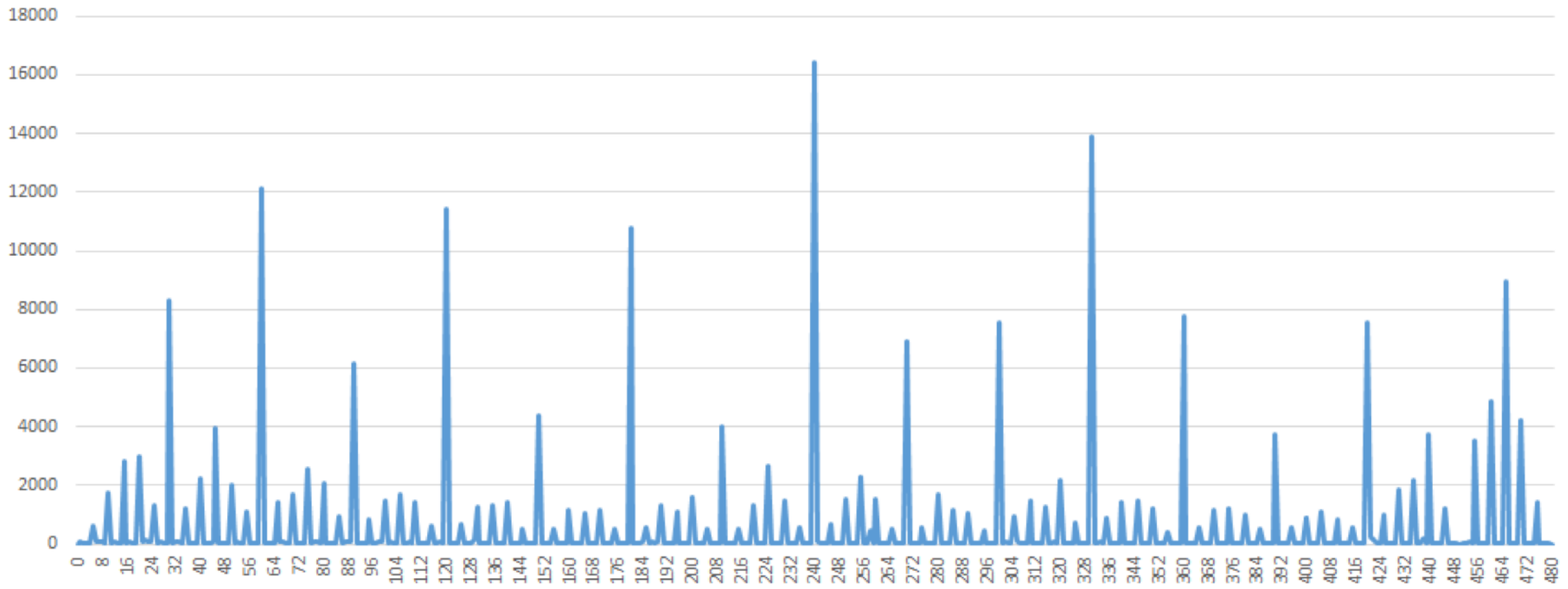
Univariate tests do not always tell the **whole** story.

This step might allow for the identification of potential outliers.

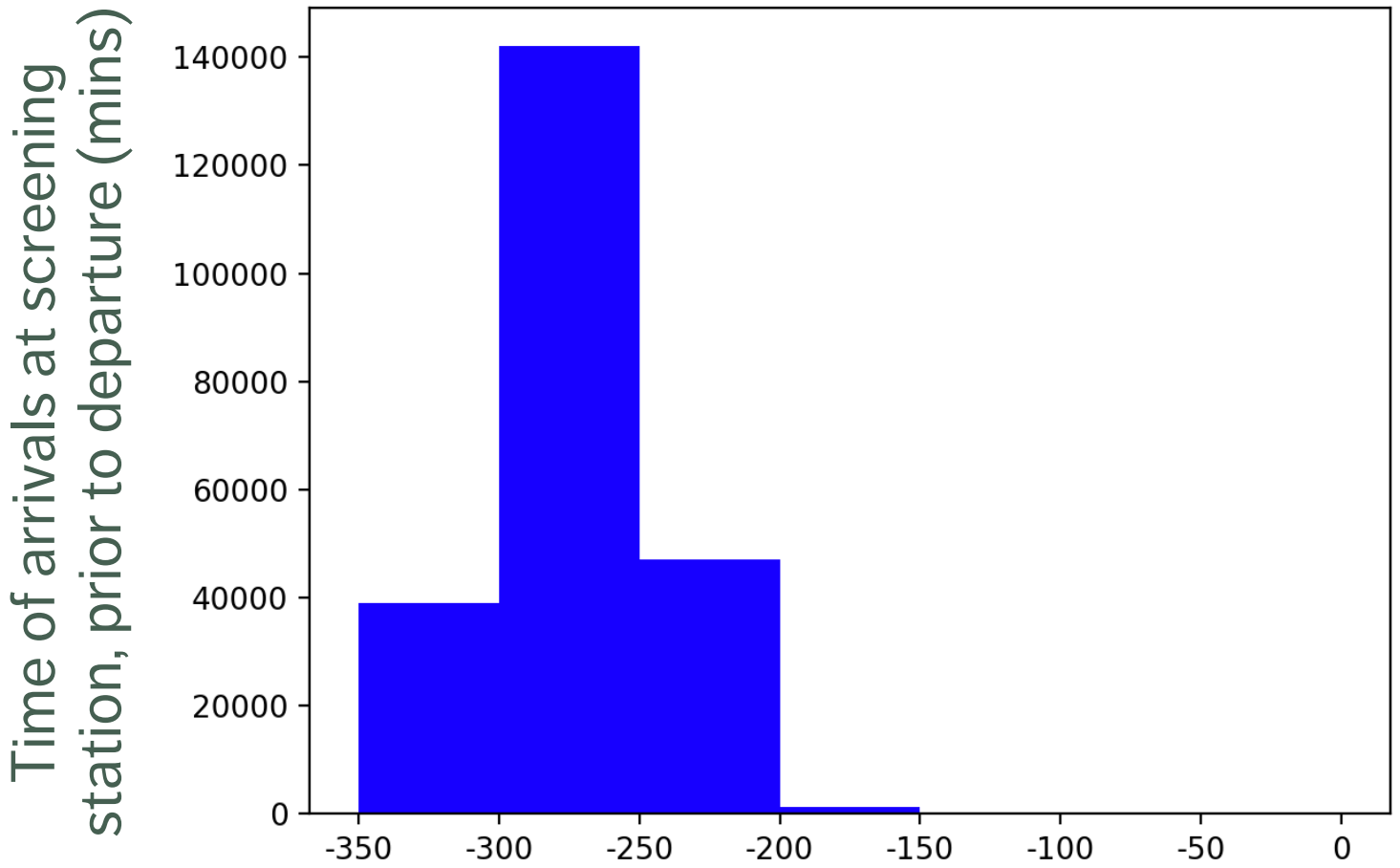
Failure to detect invalid entries ≠ all entries are valid.

Small numbers of invalid entries recoded as “missing.”

Detecting Invalid Entries



Detecting Invalid Entries



Detecting Invalid Entries

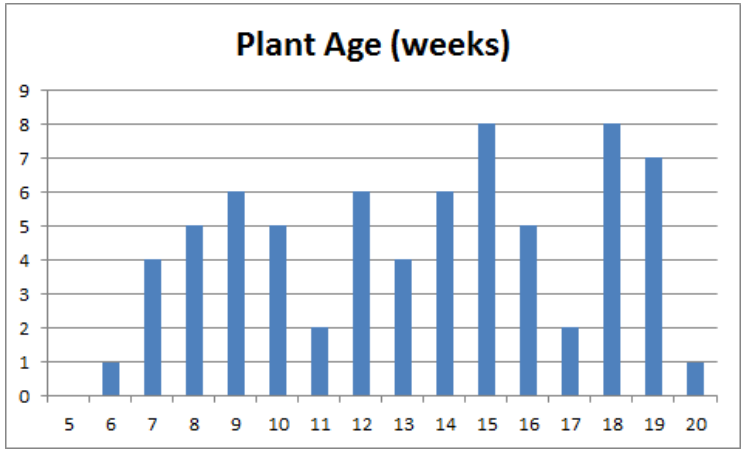
Sex	
Male	19
Female	17
(blank)	2
Total	**38**

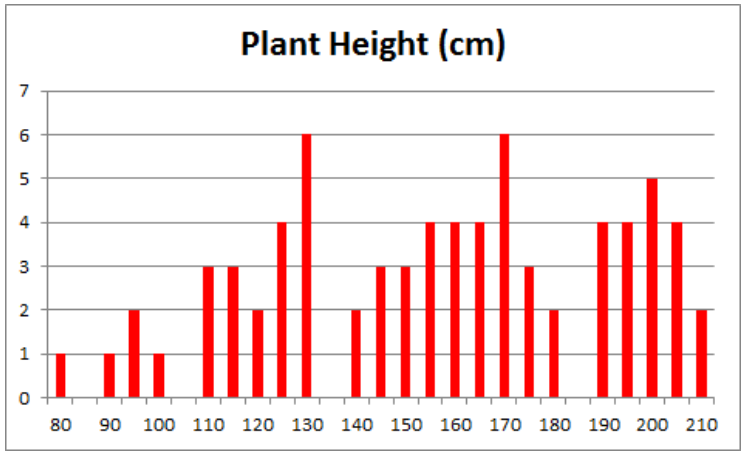
Pregnant	
Yes	7
No	27
99	1
(blank)	3
Total	**38**

vs.

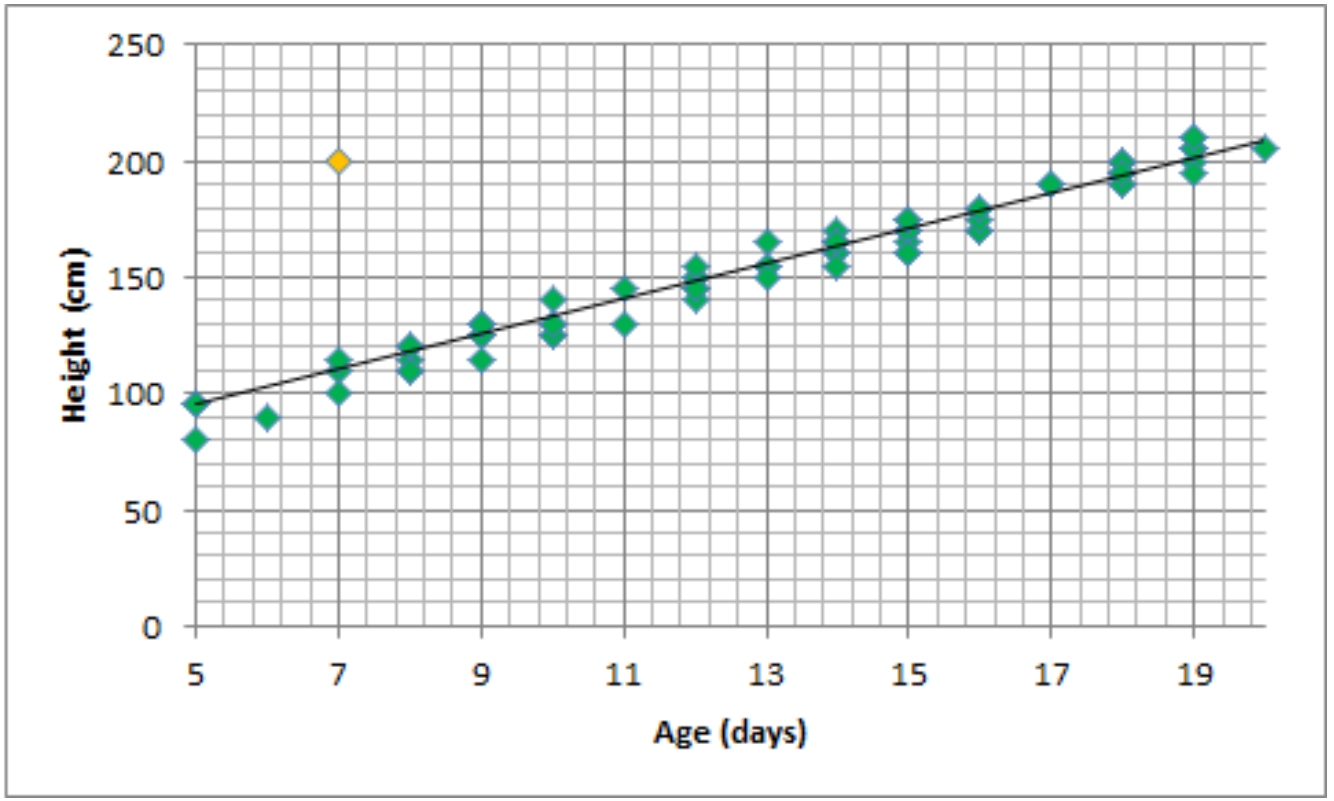
	Pregnant				Total
	Yes	No	99	(blank)	
Male	1	17	1	0	**19**
Female	6	9	0	2	**17**
(blank)	0	1	0	1	**2**
Total	**7**	**27**	**1**	**3**	**38**

Detecting Invalid Entries



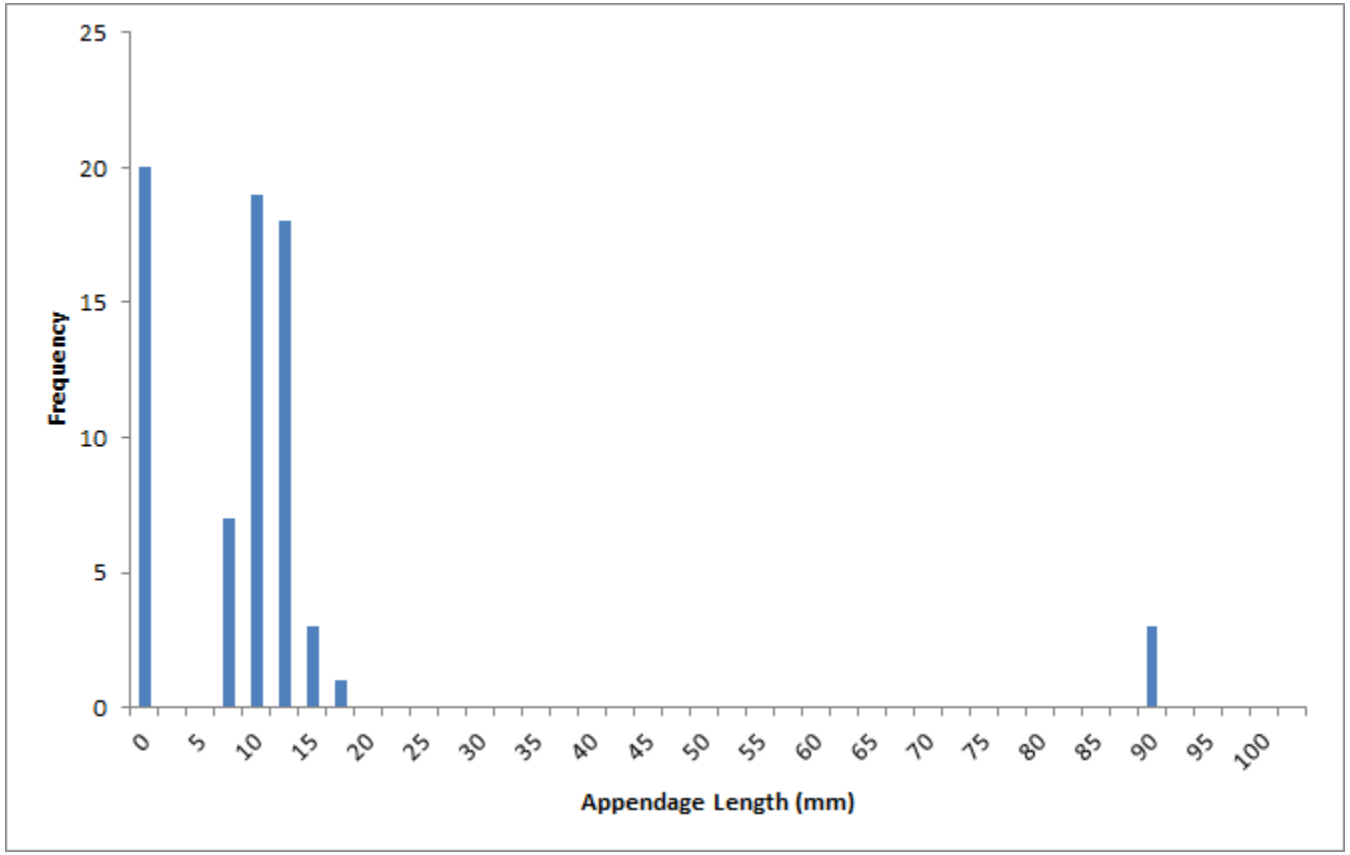


vs.



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
Data Preparation

Introduction

General Principles

- Approaches to Data Cleaning
- Pros and Cons
- Tools and Methods

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).