

Filter

feature vectors (3 min)

First steps (5 min)

Programming exercises (10 min)

Normalization (20 min)

Binning (15 min)

Scrubbing (5 min)

Qualities of good numerical features (5 min)

Polynomial transforms (5 min)

Test your knowledge (10 min)

Conclusion (2 min)

What's next

Working with categorical data (50 min)

Datasets, generalization, and overfitting (105 min)

Advanced ML models

<

Numerical data: Scrubbing

🔖

📄

Send feedback

Apple trees produce a mixture of great fruit and wormy messes. Yet the apples in high-end grocery stores display 100% perfect fruit. Between orchard and grocery, someone spends significant time removing the bad apples or spraying a little wax on the salvageable ones. As an ML engineer, you'll spend enormous amounts of your time tossing out bad examples and cleaning up the salvageable ones. Even a few bad apples can spoil a large dataset.

Many examples in datasets are unreliable due to one or more of the following problems:

Problem category	Example
Omitted values	A census taker fails to record a resident's age.
Duplicate examples	A server uploads the same logs twice.
Out-of-range feature values.	A human accidentally types an extra digit.
Bad labels	A human evaluator mislabels a picture of an oak tree as a maple.

You can write a program or script to detect any of the following problems:

- Omitted values
- Duplicate examples
- Out-of-range feature values

For example, the following dataset contains six repeated values:

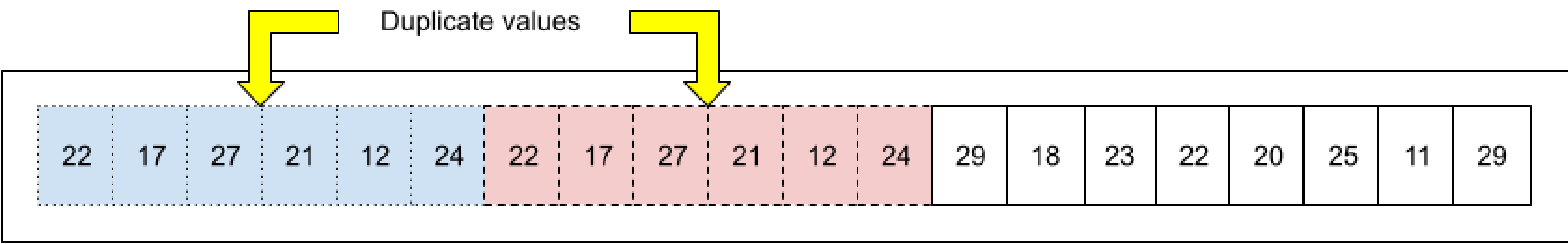


Figure 15. The first six values are repeated.

As another example, suppose the temperature range for a certain feature must be between 10 and 30 degrees, inclusive. But accidents happen—perhaps a thermometer is temporarily exposed to the sun which causes a bad outlier. Your program or script must identify temperature values less than 10 or greater than 30:

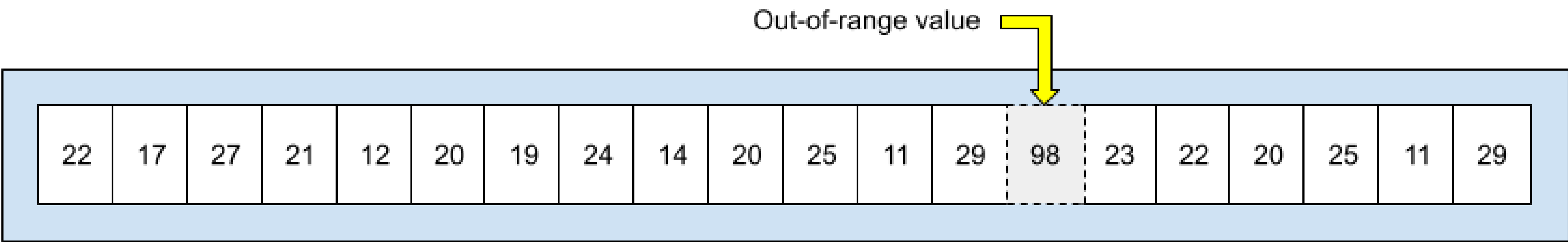


Figure 16. An out-of-range value.

When labels are generated by multiple people, we recommend statistically determining whether each rater generated equivalent sets of labels. Perhaps one rater was a harsher grader than the other raters or used a different set of grading criteria?

Once detected, you typically "fix" examples that contain bad features or bad labels by removing them from the dataset or imputing their values. For details, see the [Data characteristics](#) section of the [Datasets, generalization, and overfitting](#) module.

Help Center

Previous

← Binning (15 min)

Next

Qualities of good numerical features (5 min) →

Was this helpful?



Send feedback

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](#), and code samples are licensed under the [Apache 2.0 License](#). For details, see the [Google Developers Site Policies](#). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2024-10-09 UTC.

Connect

Blog

Instagram

LinkedIn

X (Twitter)

YouTube

Programs

Google Developer Groups

Google Developer Experts

Accelerators

Women Techmakers

Google Cloud & NVIDIA

Developer consoles

Google API Console

Google Cloud Platform Console

Google Play Console

Firebase Console

Actions on Google Console

Cast SDK Developer Console

Chrome Web Store Dashboard

Google Home Developer Console

Google for Developers

Android

Chrome

Firebase

Google Cloud Platform

Google AI

All products

Terms | Privacy

Sign up for the Google for Developers newsletter

Subscribe

🌐

English ▾