

≡

Filter

Data

▾ Working with numerical data (85 min)

≡ Introduction (3 min)

≡ How a model ingests data with feature vectors (5 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)

<

Numerical data: First steps



Send feedback

- On this page
- Visualize your data
- Statistically evaluate your data
- Find outliers

Before creating feature vectors, we recommend studying numerical data in two ways:

- Visualize your data in plots or graphs.
- Get statistics about your data.

Visualize your data

Graphs can help you find anomalies or patterns hiding in the data. Therefore, before getting too far into analysis, look at your data graphically, either as scatter plots or histograms. View graphs not only at the beginning of the data pipeline, but also throughout data transformations. Visualizations help you continually check your assumptions.

We recommend working with pandas for visualization:

- [Working with Missing Data \(pandas Documentation\)](#)
- [Visualizations \(pandas Documentation\)](#)

Note that certain visualization tools are optimized for certain data formats. A visualization tool that helps you evaluate protocol buffers may or may not be able to help you evaluate CSV data.

Statistically evaluate your data

Beyond visual analysis, we also recommend evaluating potential features and labels mathematically, gathering basic statistics such as:

- mean and median
- standard deviation
- the values at the quartile divisions: the 0th, 25th, 50th, 75th, and 100th percentiles. The 0th percentile is the minimum value of this column; the 100th percentile is the maximum value of this column. (The 50th percentile is the median.)

Find outliers

An **outlier** is a value *distant* from most other values in a feature or label. Outliers often cause problems in model training, so finding outliers is important.

When the delta between the 0th and 25th percentiles differs significantly from the delta between the 75th and 100th percentiles, the dataset probably contains outliers.

★ **Note:** Don't over-rely on basic statistics. Anomalies can also hide in seemingly well-balanced data.

Outliers can fall into any of the following categories:

- The outlier is due to a *mistake*. For example, perhaps an experimenter mistakenly entered an extra zero, or perhaps an instrument that gathered data malfunctioned. You'll generally delete examples containing mistake outliers.
- The outlier is a legitimate data point, *not a mistake*. In this case, will your trained model ultimately need to infer good predictions on these outliers?
 - If yes, keep these outliers in your training set. After all, outliers in certain features sometimes mirror outliers in the label, so the outliers could actually *help* your model make better predictions. Be careful, extreme outliers can still hurt your model.
 - If no, delete the outliers or apply more invasive feature engineering techniques, such as **clipping**.

- A **Key terms:**
- [Clipping](#)
 - [Outliers](#)

Help Center

Previous

←

How a model ingests data with feature vectors (5 min)

Next

→

Programming exercises (10 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 2025-02-26 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 ▾