Filter

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
Neural networks (75 min)
Embeddings (45 min)



# Numerical data: Conclusion

On this page
Additional Information
What's next

A machine learning (ML) model's health is determined by its data. Feed your model healthy data and it will thrive; feed your model junk and its predictions will be worthless.

Best practices for working with numerical data:

- Remember that your ML model interacts with the data in the **feature vector**, not the data in the **dataset**.
- **Normalize** most numerical **features**.
- If your first normalization strategy doesn't succeed, consider a different way to normalize your data.
- **Binning**, also referred to as **bucketing**, is sometimes better than normalizing.
- Considering what your data *should* look like, write verification tests to validate those expectations. For example:
  - The absolute value of latitude should never exceed 90. You can write a test to check if a latitude value greater than 90 appears in your data.
  - If your data is restricted to the state of Florida, you can write tests to check that the latitudes fall between 24 through 31, inclusive.
- Visualize your data with scatter plots and histograms. Look for anomalies.
- Gather statistics not only on the entire dataset but also on smaller subsets of the dataset. That's because aggregate statistics sometimes obscure problems in smaller sections of a dataset.
- Document all your data transformations.

Data is your most valuable resource, so treat it with care.

## Additional Information

- The *Rules of Machine Learning* guide contains a valuable Feature Engineering section.

## What's next

Congratulations on finishing this module!

We encourage you to explore the various MLCC modules at your own pace and interest. If you'd like to follow a recommended order, we suggest that you move to the following module next: **Representing categorical data**.

**Key terms:**

- Binning
- Bucketing
- Dataset
- Feature
- Feature vector
- Normalization

Help Center

Previous
Test your knowledge (10 min)

Next
Introduction (5 min)

Was this helpful?

Send feedback