GENERAL ASSEMBLY

Feature Scaling

Sam Stack

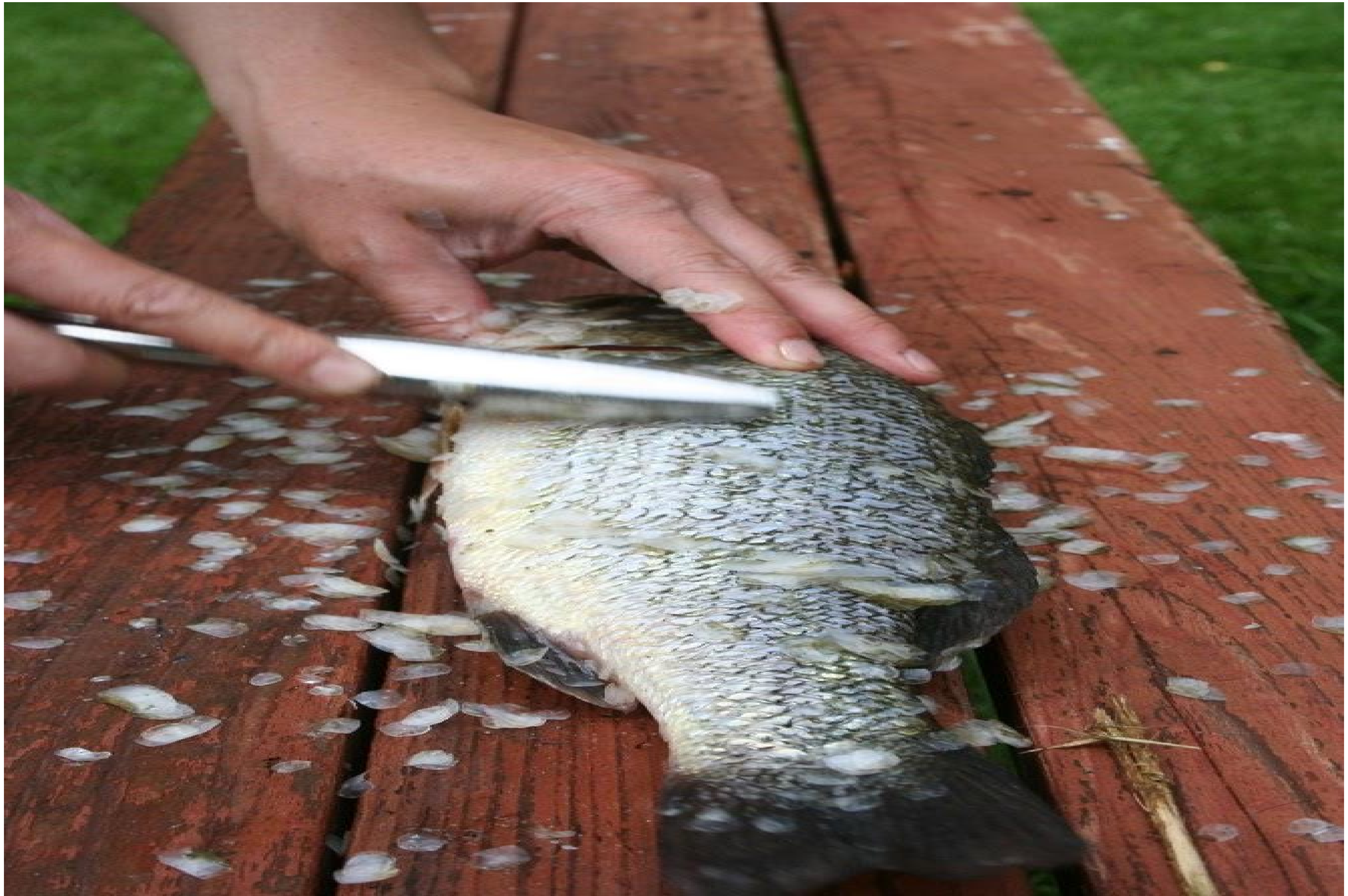
LEARNING OBJECTIVES

- Explain the benefits of scaling data.
- Identify situations where scaling data is beneficial.
- Scale data using Python and SKLearn.

Feature Scaling

What is scaling?

Fish Scaling



Mountain Scaling

Feature Scaling

What is mathematical* scaling?

Feature Scaling

What is scaling?

- Increase or decrease according to a fixed ratio.
- Change in size but not in shape (or distribution).

Feature Scaling

- Increase or decrease according to a fixed ratio.
- Change in size but not in shape (or distribution).

How can this be applied to data?

Feature Scaling

Why would we want to scale data?

Why do we scale data?

There are a number of good reasons why we scale our data:

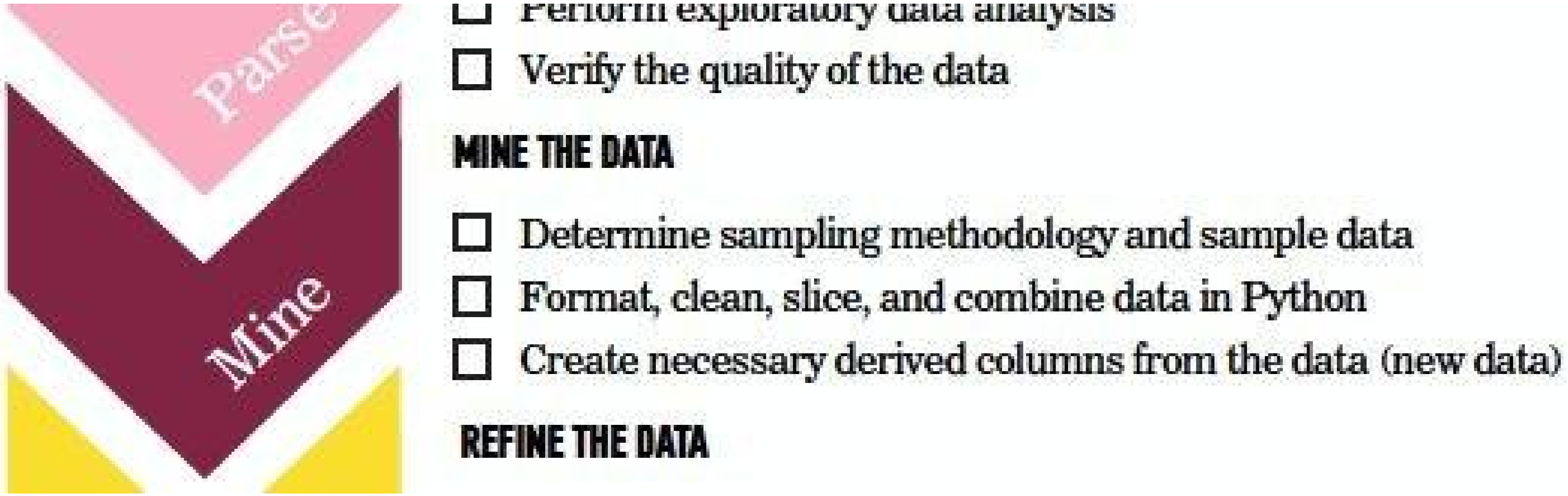
- To handle disparities in units.
- Cut computational expense
- Improve model performance (Especially Machine Learning)
- We scale for models to prevent the steps on different axis from varying widely.

It’s rarely a bad idea to scale your data!

Feature Scaling

Exception: Tree based Models.
Because they consider features independently

Feature Scaling



□ Perform exploratory data analysis

□ Verify the quality of the data

MINE THE DATA

□ Determine sampling methodology and sample data

□ Format, clean, slice, and combine data in Python

□ Create necessary derived columns from the data (new data)

REFINE THE DATA

Feature Scaling

How do we scale our data?

Feature Scaling

How do we scale our data?

There are a few

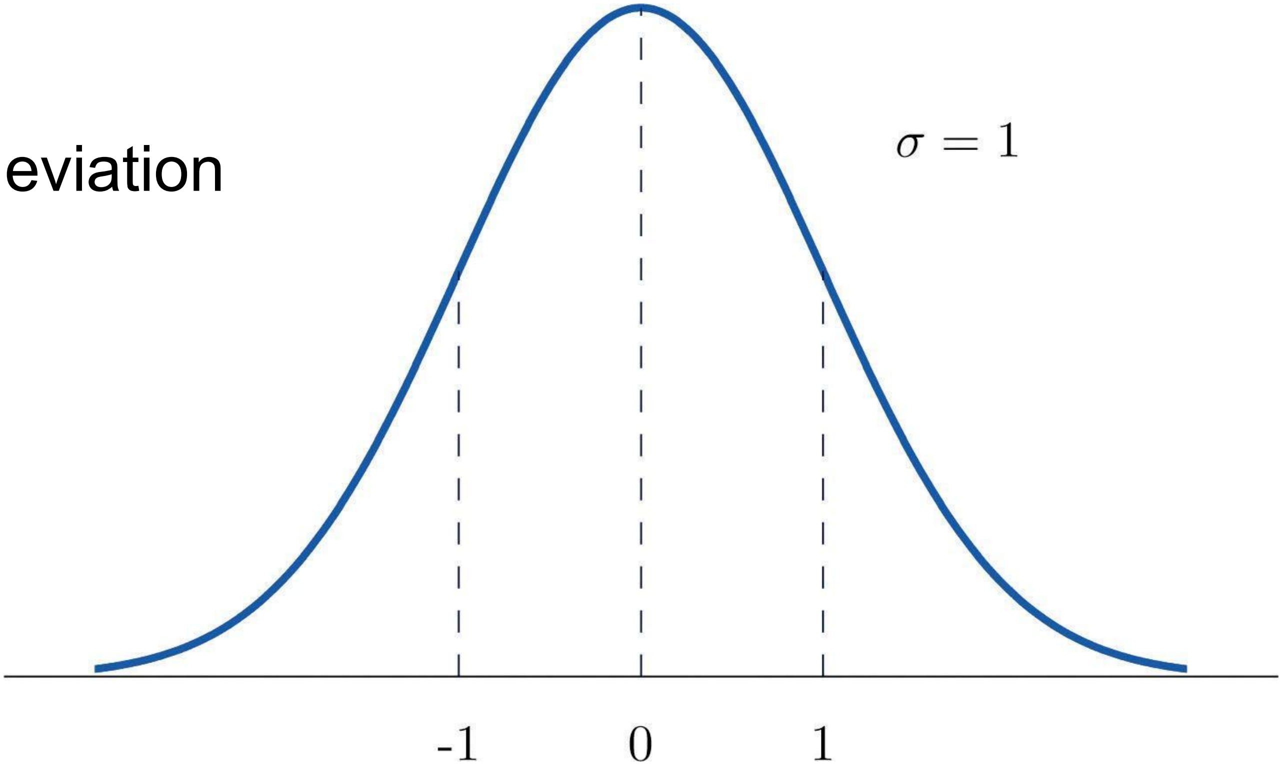
1. **Standardization:** (Z/T-Score Normalization) Scales a feature to have a mean of 0 and standard deviation/variance equals 1.
2. **Normalization:** Reduce Data to a neutral (standard) scale typically 0 to 1

Feature Scaling

1. Standardization: Takes a feature and rescales it to have a mean of zero and a variance of 1.

- Compute the mean and standard deviation

x’ = (x - mean) / std_dev



Feature Scaling

When would we want to standardize?

- KNN: Scaling is necessary if you want all the features to contribute equally!
    - KNN Models take into consideration Euclidean distances.
- K-Means: Since the algorithm works off of computing means, unscaled data will severely affect the model
- Logistic Regression, Neural Networks, and SVMs - unscaled data will disproportionately weigh some data points

Feature Scaling

How do we scale our data?

There are a few

1. ~~**Standardization:** (Z/T-Score Normalization) Scales a feature to have a mean of 0 and standard deviation/variance equals 1.~~

2. **Normalization:** Reduce Data to a neutral (standard) scale typically 0 to 1

Feature Scaling

Normalization: Scales data to exist within a neutral (standard) range. Typically 0-to-1

Probably the most common application of normalization is Min-Max Scaling.

$$x' = \frac{(x - min)}{(max/min)}$$

Feature Scaling

Demo: Scaling in Python

Conclusion