Building Features from Numeric Data

USING NUMERIC DATA IN MACHINE LEARNING ALGORITHMS

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Pre-processing data for ML models

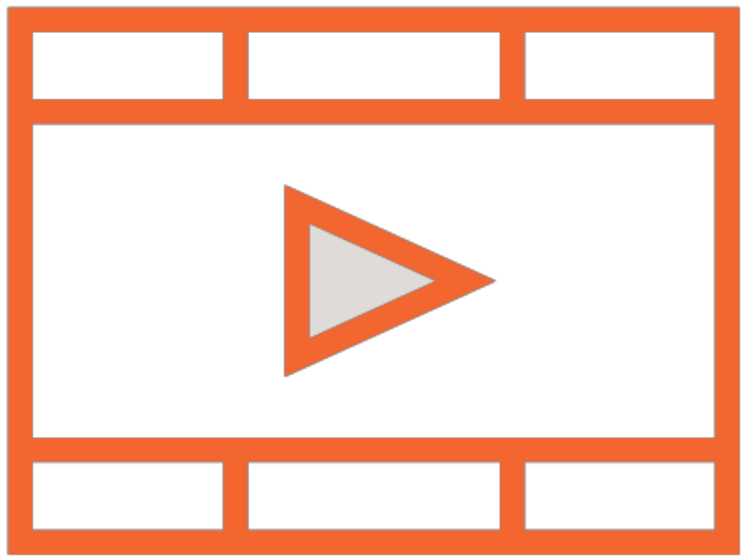
Using mean and variance to standardize and scale data

Box plots for outlier detection and data exploration

Outlier removal using quartile range selection

Prerequisites and Course Outline

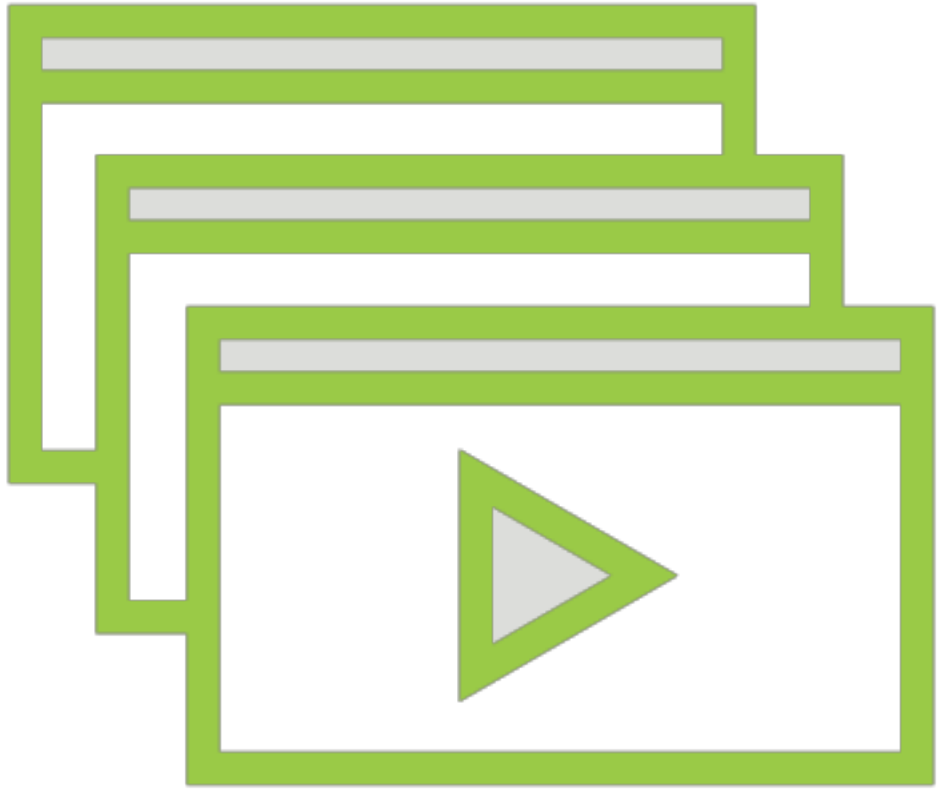
Prerequisites



Working with Python and Python libraries

Basic understanding of machine learning algorithms

Prerequisites



Understanding Machine Learning by David Chappell

Building Machine Learning Models in Python with scikit-learn by Janani Ravi

Understanding Machine Learning with Python by Jerry Kurata

Course Outline

Using numeric data in ML models

- Mean, variance and standard deviation
- Standardization and scaling numeric data

Normalization to unit norm

- Normalization and cosine similarity
- L1, L2 and max normalization

Scaling and advanced transformations

- Continuous data to categorical form
- Working with polynomial features
- Transforming data to normal distribution

Numeric Features in Training Data

Numeric Features

Can represent any kind of information

The range of each feature will be different

The average and dispersion of features will also be different

Comparing different features is hard

Machine learning algorithms typically do not work well with numeric data with different scales

Feature Scaling

Scaling

Standardization

Feature Scaling

Scaling

Standardization

Numeric values are **shifted and rescaled** so all features have the same scale i.e. within the same minimum and maximum values

Feature Scaling

Scaling

Standardization

Often data scaled to be in the range of 0 to 1, many people call this normalization

Feature Scaling

Scaling

Standardization

The feature range of data is something that you can specify

Feature Scaling

Scaling

Standardization

Does not bind values to a specific range

Feature Scaling

Scaling

Standardization

Centers data round the mean and divides each value by the variance so all features have 0 mean and unit variance

Mean, Variance and Standard Deviation

Data in One Dimension

Pop quiz: Your thoughtful, fact-based point-of-view on these numbers, please

Mean as Headline

The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

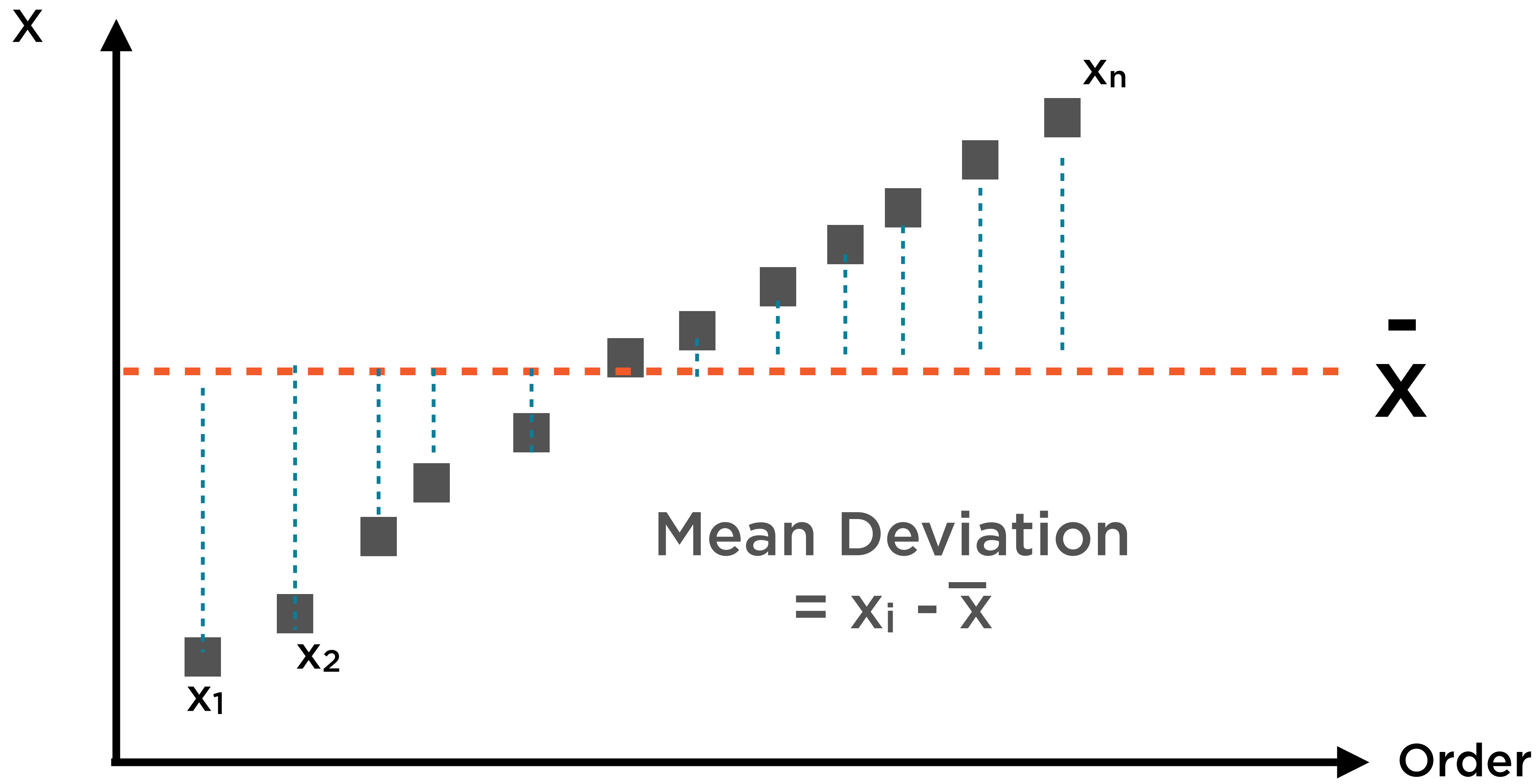
Variation Is Important Too

"Do the numbers jump around?"

$$\text{Range} = X_{max} - X_{min}$$

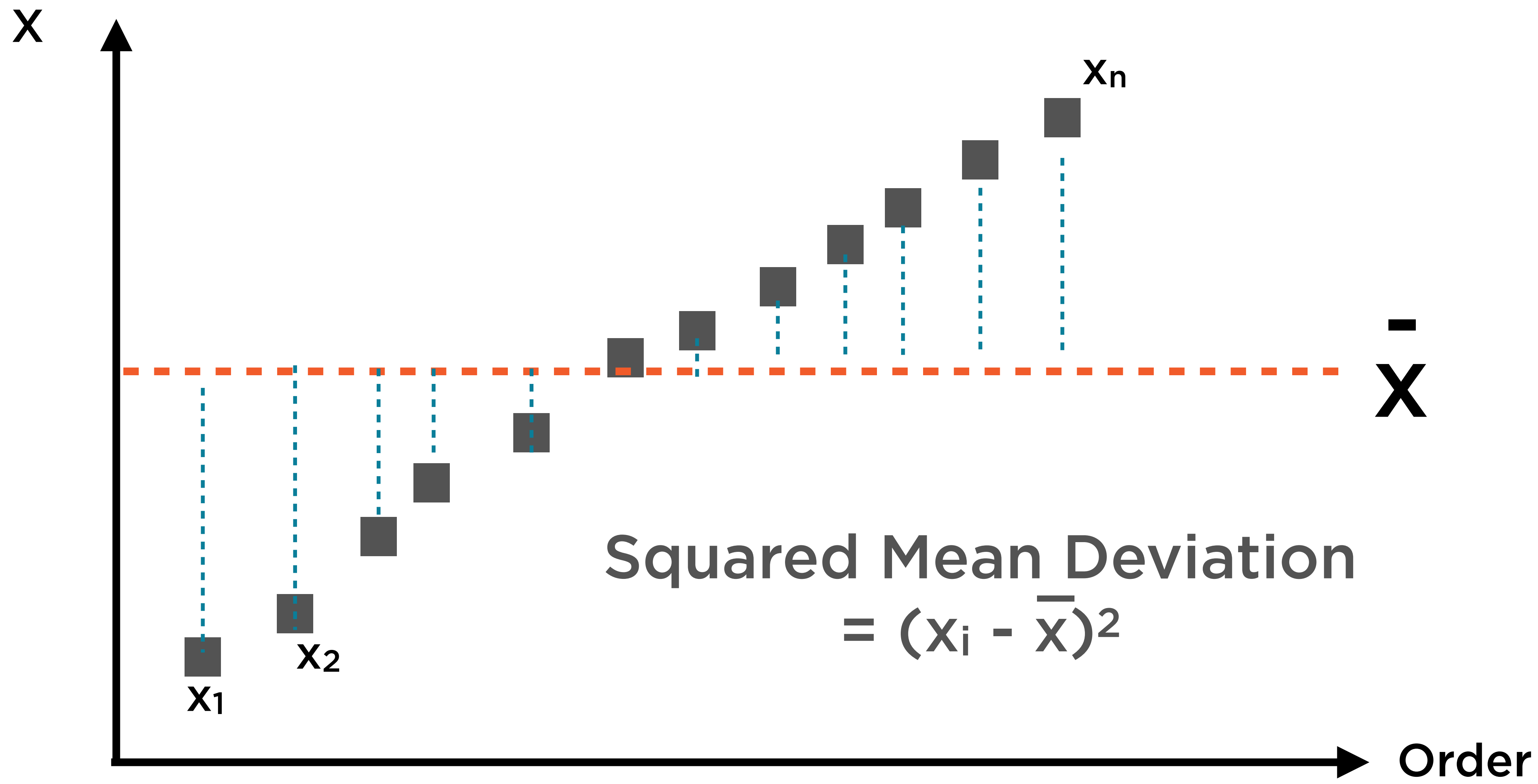
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



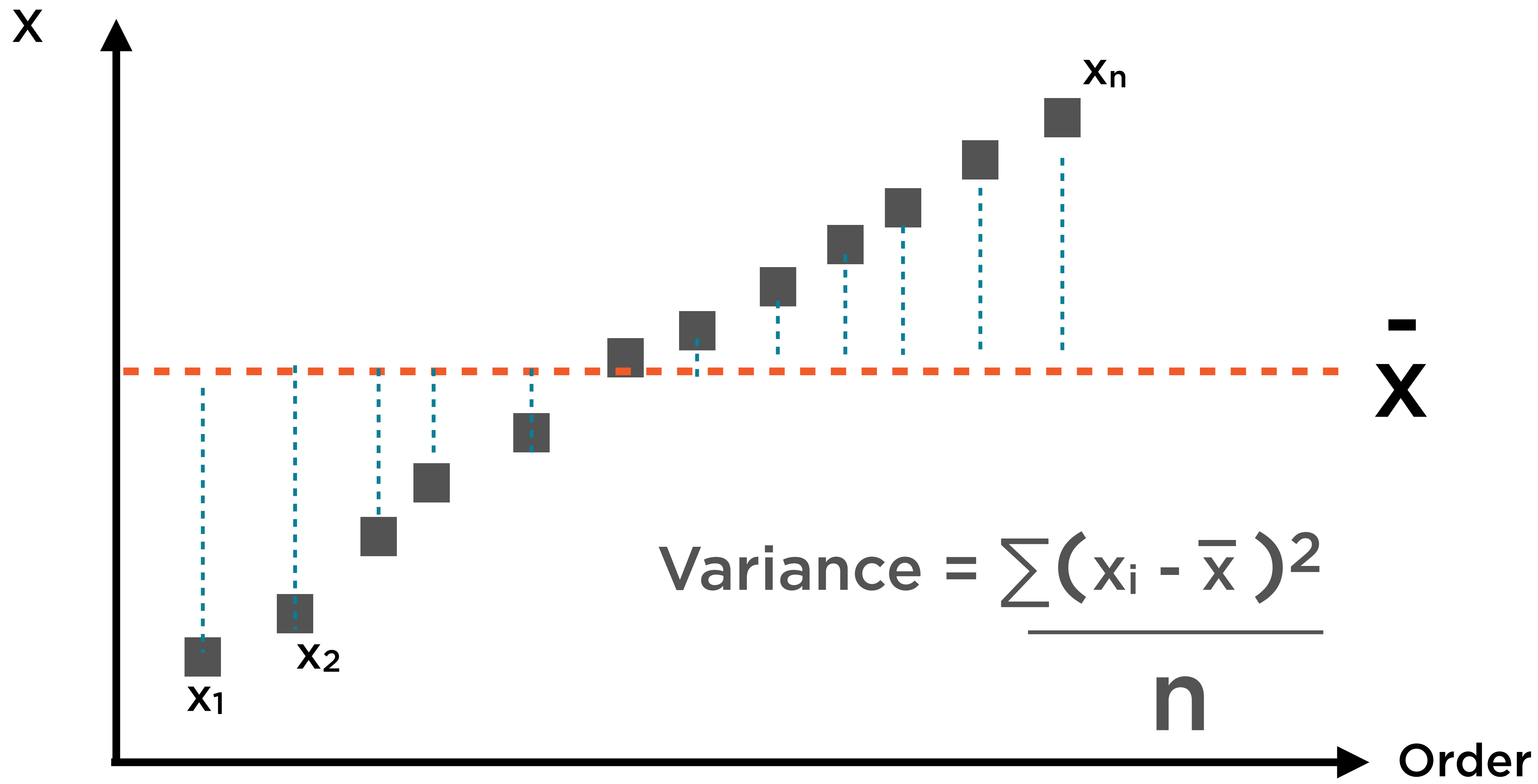
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



Variance is the second-most important number to summarize this set of data points

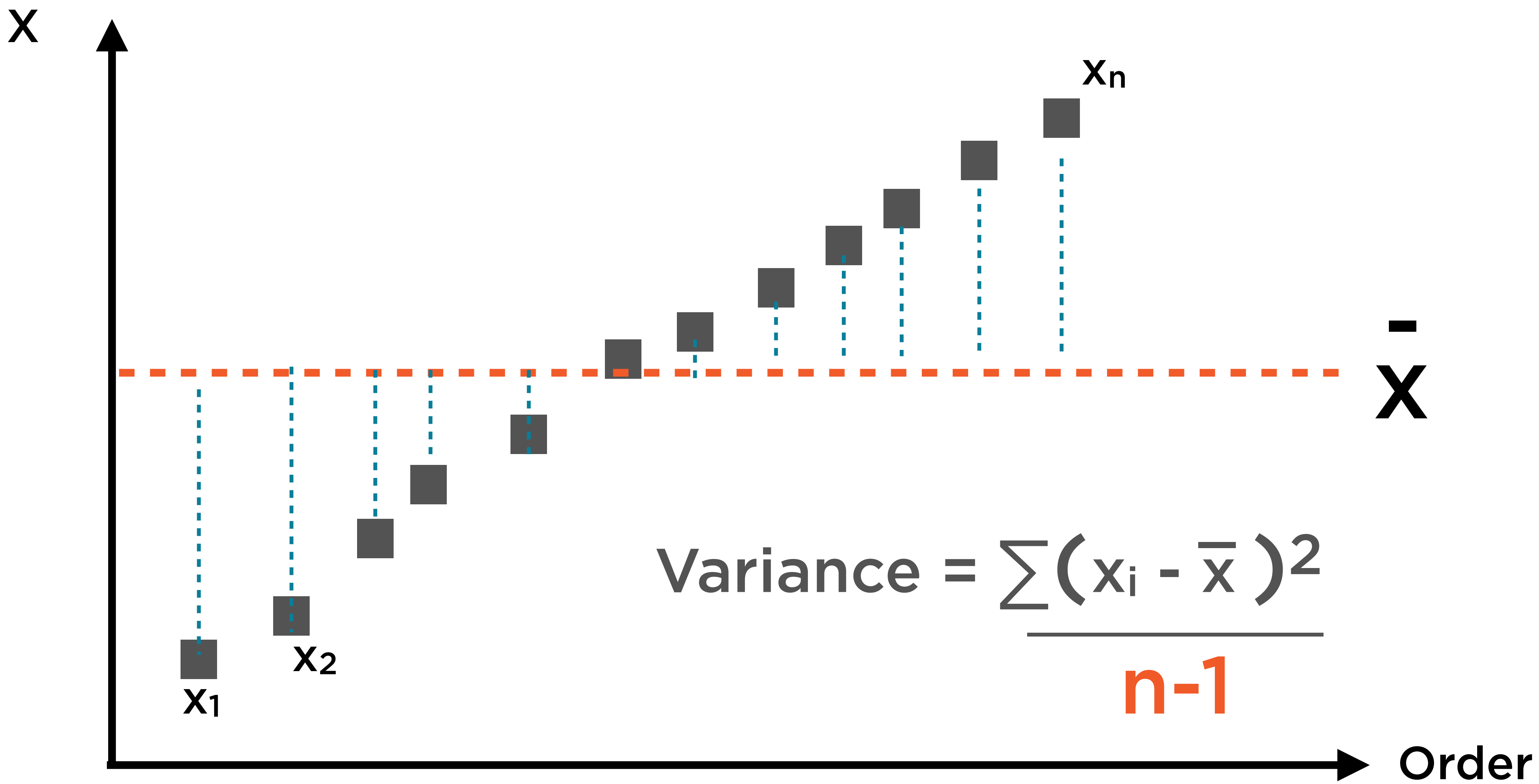
Variance as Asterisk



$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

We can improve our estimate of the variance by tweaking the denominator - this is called Bessel's Correction

Mean and Variance

Mean and variance succinctly summarize a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Variance and Standard Deviation

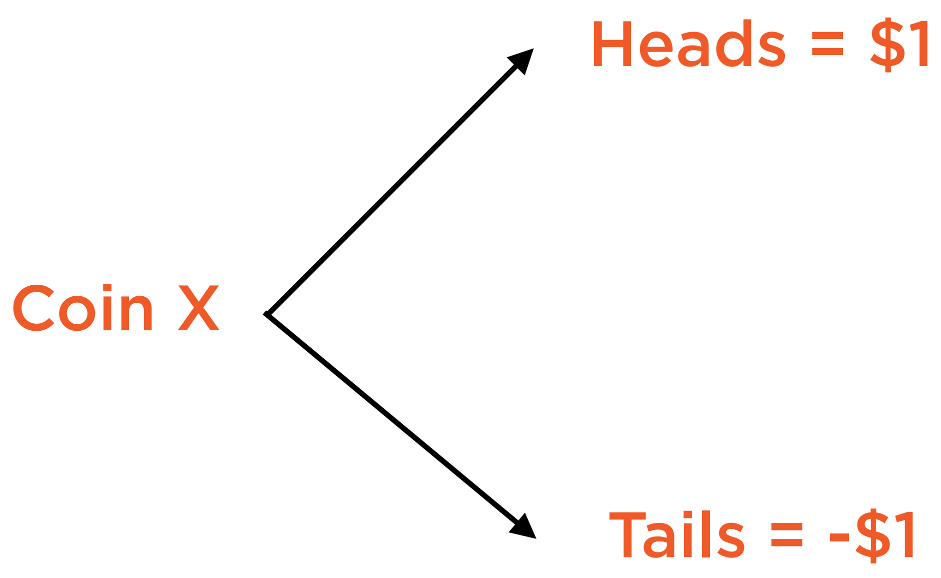
Standard deviation is the square root of variance

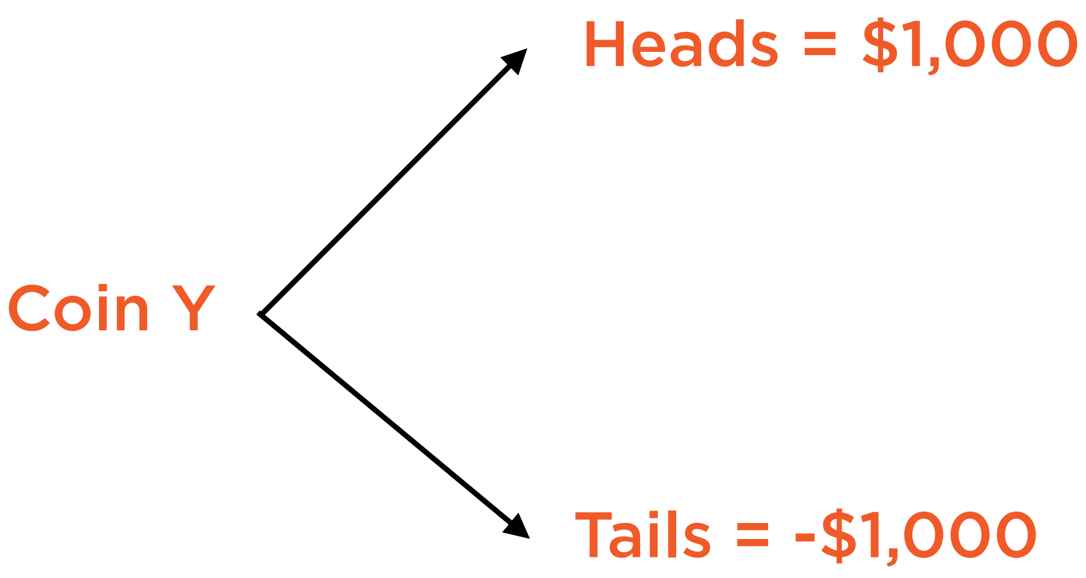
$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Understanding How Variances Work

Tossing Two Coins





Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Tabulate the possible outcomes (assume each coin is a fair one)

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0$ $\bar{y} = 0$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	\$1	\$1,000
Heads	Tails	\$1	-\$1,000
Tails	Heads	-\$1	\$1,000
Tails	Tails	-\$1	-\$1,000

$\bar{x} = 0 \quad \bar{y} = 0$

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n}$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
$1	1
$1	1
-$1	1
-$1	1

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n} = 1$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$y_i - \bar{y}$	$(y_i - \bar{y})^2$
$1,000	10,00,000
-$1,000	10,00,000
$1,000	10,00,000
-$1,000	10,00,000

$$\text{Variance} = \frac{\sum(y_i - \bar{y})^2}{n} = 1{,}000{,}000$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	\$1	\$1,000
Heads	Tails	\$1	-\$1,000
Tails	Heads	-\$1	\$1,000
Tails	Tails	-\$1	-\$1,000

$\bar{x} = 0$ $\bar{y} = 0$

$\mathrm{Var}(x) = 1$ $\mathrm{Var}(y) = 1{,}000{,}000$

As stakes grow, variance gets big faster than the mean

Tossing Two Coins

Coin X
- Heads = $1
- Tails = -$1

Coin Y
- Heads = $1,000
- Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

As stakes grow 1000x, variance grows 1,000,000x

Demo

Calculating mean, variance, and standard deviation

StandardScaler

Feature Scaling

Scaling

Standardization

Feature Scaling

Scaling

Standardization

Scaling Data

$$\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \ldots & x_{3k} \\ \ldots & & \ldots \\ x_{n1} & & x_{nk} \end{bmatrix}$$

Maximum and minimum values of X_1, X_2, ... X_k can be very different

Scaling Data

$$\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \dots & x_{3k} \\ \dots & & \dots \\ x_{n1} & & x_{nk} \end{bmatrix}$$

Scaling refers to having all data in the same range i.e. same maximum and minimum values

Scaling Data

$$
\begin{bmatrix}
x_{11} & & x_{1k} \\
x_{21} & & x_{2k} \\
x_{31} & \cdots & x_{3k} \\
\cdots & & \cdots \\
x_{n1} & & x_{nk}
\end{bmatrix}
$$

Scaling operations are applied to features i.e.
to all data in a single column

Scaling Data

$$\begin{bmatrix} x_{11} & & x_{1k} \\ x_{21} & & x_{2k} \\ x_{31} & \cdots & x_{3k} \\ \cdots & & \cdots \\ x_{n1} & & x_{nk} \end{bmatrix}$$

Scaling operations are applied to features i.e.
to all data in a single column

Feature Scaling

Scaling

Standardization

Standardization centers features to have a mean of 0 and a variance of 1

Standardizing Data

$$\begin{bmatrix} X_{11} & & X_{1k} \\ X_{21} & & X_{2k} \\ X_{31} & \cdots & X_{3k} \\ \cdots & & \cdots \\ X_{n1} & & X_{nk} \end{bmatrix}$$

$$\begin{array}{ccc} avg(X_1) & \cdots & avg(X_k) \\ stdev(X_1) & \cdots & stdev(X_k) \end{array}$$

Standardizing Data

$$\begin{bmatrix} \dfrac{x_{11} - avg(X_1)}{stdev(X_1)} & \cdots & \dfrac{x_{1k} - avg(X_k)}{stdev(X_k)} \\ \cdots & & \cdots \\ \dfrac{x_{n1} - avg(X_1)}{stdev(X_1)} & & \dfrac{x_{nk} - avg(X_k)}{stdev(X_k)} \end{bmatrix}$$

Each column of the standardized data has mean 0 and variance 1

Standardizing Data

$$\begin{bmatrix} \boxed{\begin{matrix} \dfrac{x_{11} - avg(X_1)}{stdev(X_1)} \\ \cdots \\ \dfrac{x_{n1} - avg(X_1)}{stdev(X_1)} \end{matrix}} & \cdots & \begin{matrix} \dfrac{x_{1k} - avg(X_k)}{stdev(X_k)} \\ \cdots \\ \dfrac{x_{nk} - avg(X_k)}{stdev(X_k)} \end{matrix} \end{bmatrix}$$

Standardization is applied to features i.e. to all data in a single column

Standardizing Data

$$
\begin{bmatrix}
\dfrac{x_{11} - avg(X_1)}{stdev(X_1)} & \cdots & \dfrac{x_{1k} - avg(X_k)}{stdev(X_k)} \\
\cdots & & \cdots \\
\dfrac{x_{n1} - avg(X_1)}{stdev(X_1)} & & \dfrac{x_{nk} - avg(X_k)}{stdev(X_k)}
\end{bmatrix}
$$

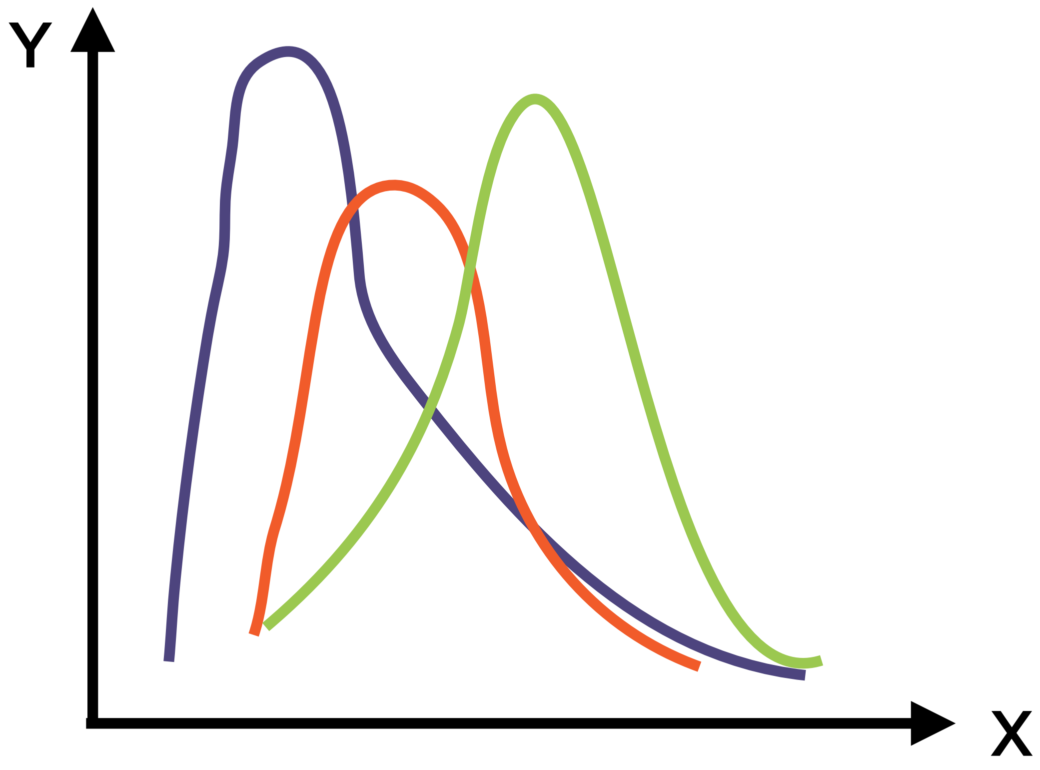
Standardization is applied to features i.e. to all data in a single column

StandardScaler

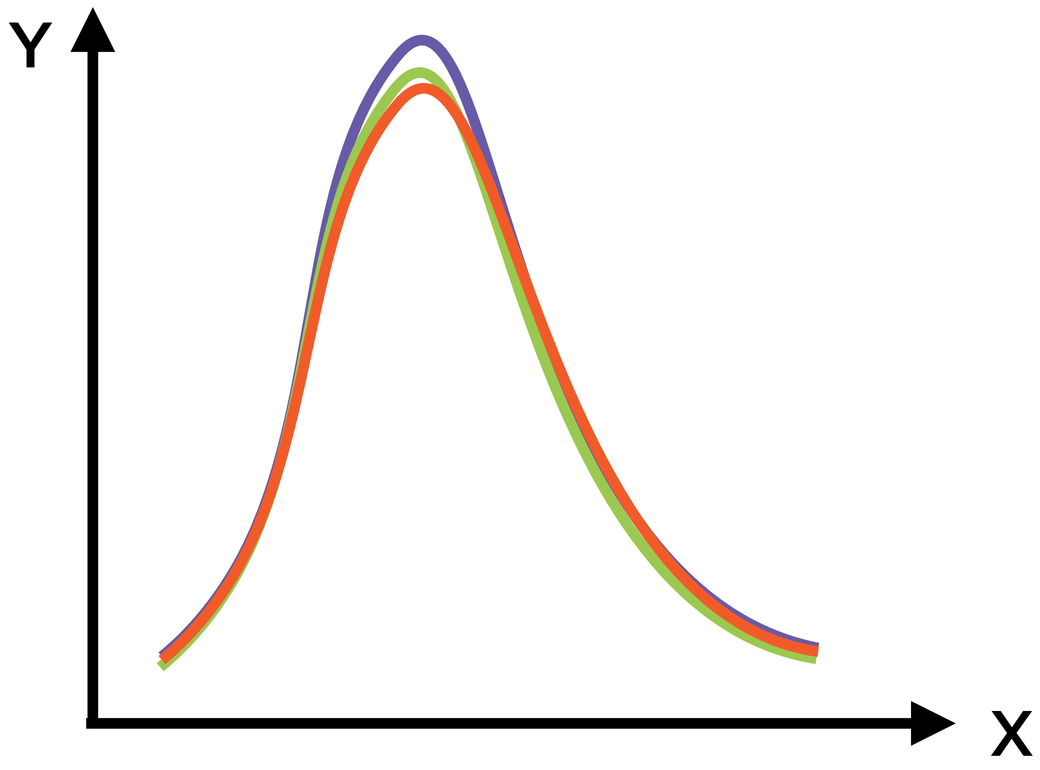
$$z = \frac{x_i - \mathrm{mean}(x)}{\mathrm{stdev}(x)}$$

StandardScaler operates column-by-column and yields features with zero mean and unit variance

StandardScaler



Before



After

Demo

Scaling numeric features using the StandardScaler

RobustScaler

The StandardScaler is very sensitive to the presence of outliers in the data

StandardScaler

$$z = \frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$$

Mean is a measure of central tendency and standard deviation is a measure of dispersion

RobustScaler

$$z = \frac{x_i - \text{median}(x)}{\text{Inter-quartile Range}(x)}$$

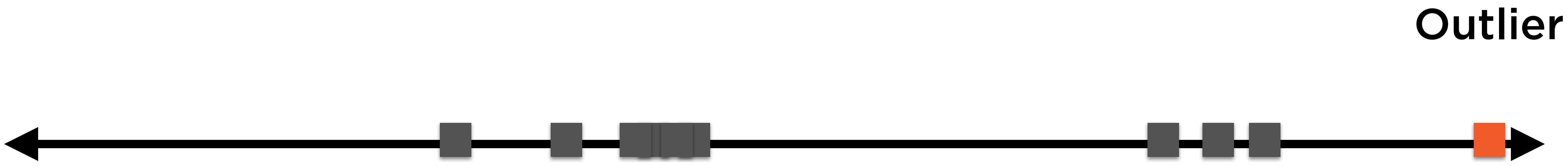
Median is also a measure of central tendency and inter-quartile range is also measure of dispersion

RobustScaler

$$z = \frac{x_i - \text{median}(x)}{\text{Inter-quartile Range}(x)}$$

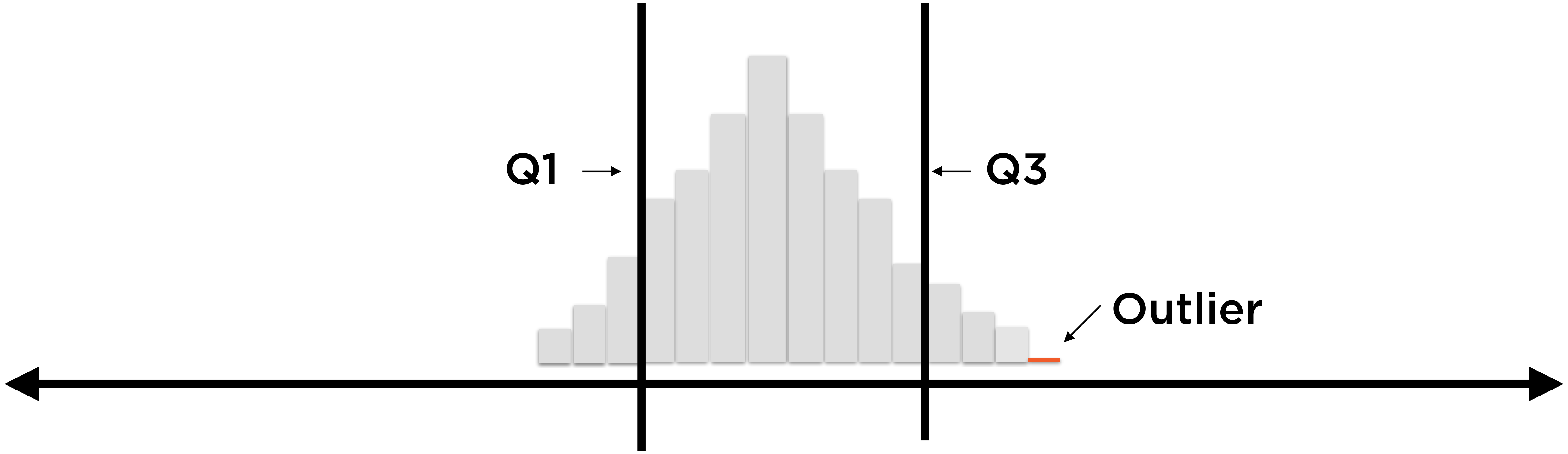
RobustScaler is a scaler whose output does not change much due to outliers

Outliers



Outliers might represent data errors, or genuinely rare points legitimately in dataset

Inter-quartile Range

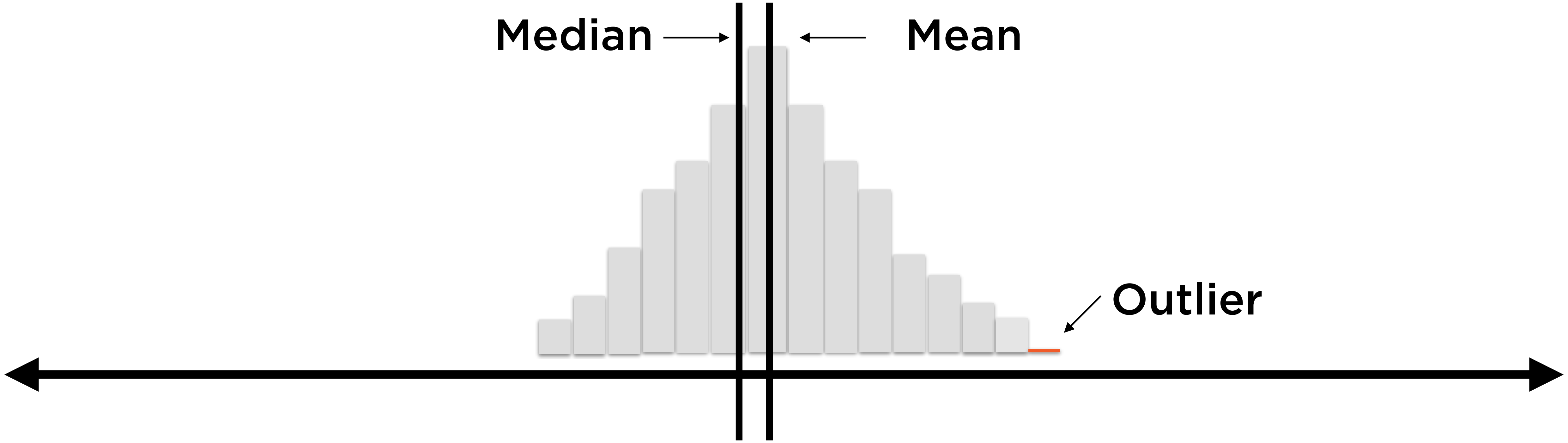


Q3 = 75th percentile: 75% of points smaller than this

Q1 = 25th percentile: 25% of points smaller than this

Inter-quartile Range (IQR) = 75th percentile - 25th percentile

Median



Median = 50th percentile: 50% of points on either side

Unlike mean, median changes little due to outliers

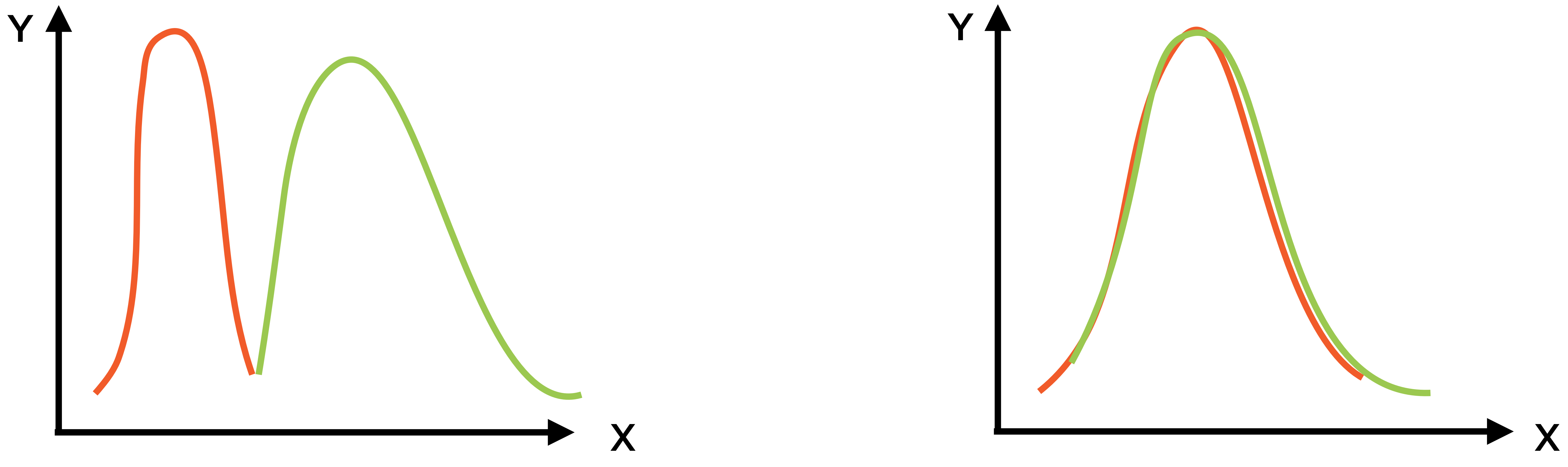
Median is used in numerator of RobustScaler

RobustScaler

$$z = \frac{x_i - \text{median}(x)}{\text{Inter-quartile Range}(x)}$$

RobustScaler is a scaler whose output does not change much due to outliers

RobustScaler



Before

After

Demo

Scaling data using the RobustScaler

Summary

Pre-processing data for ML models

Using mean and variance to standardize and scale data

Box plots for outlier detection and data exploration

Outlier removal using quartile range selection