Preparing Data for Machine Learning

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Problems working with data

Dealing with outliers and missing values

Reading, visualizing and exploring a dataset

Building a simple predictive model

Garbage In, Garbage Out

If data fed into an ML model is of poor quality, the model will be of poor quality

Problems with Data

Insufficient data	Too much data	Non-representative data
Missing data	Duplicate data	Outliers

Problems with Data

Problems with Data

Insufficient data	Too much data	Non-representative data
Missing data	Duplicate data	Outliers

Problems with Data

Insufficient data

Too much data

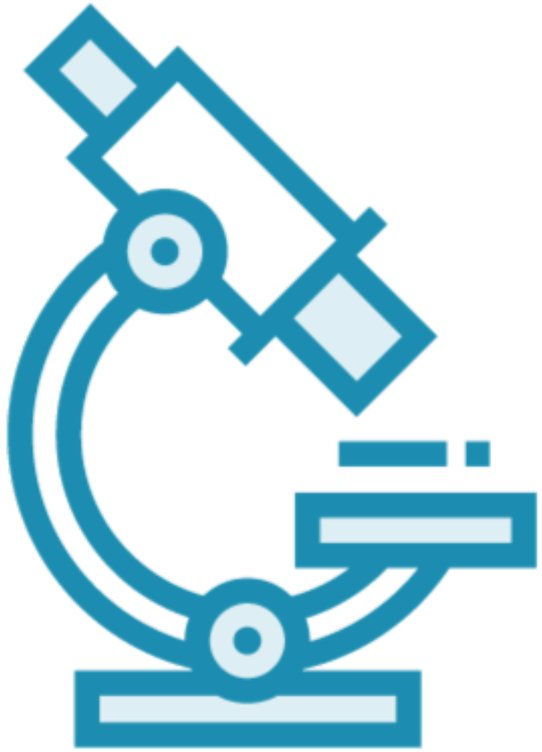
Non-representative data

Missing data

Duplicate data

Outliers

Insufficient Data



Models trained with insufficient data perform poorly in prediction

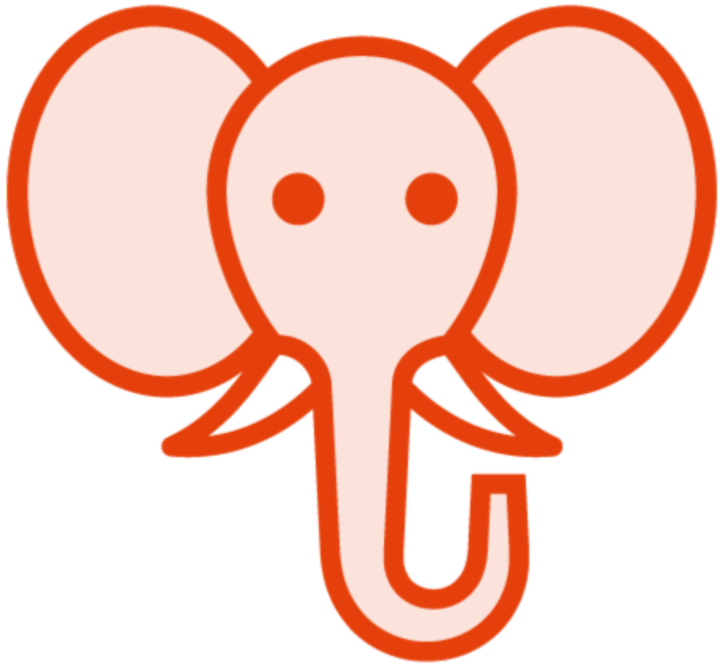
Paradoxically leads to either

- Overfitting: Read too much for too little data
- Underfitting: Build overly simplistic model from available data

Problems with Data

Insufficient data	**Too much data**	Non-representative data
Missing data	Duplicate data	Outliers

Too Much Data



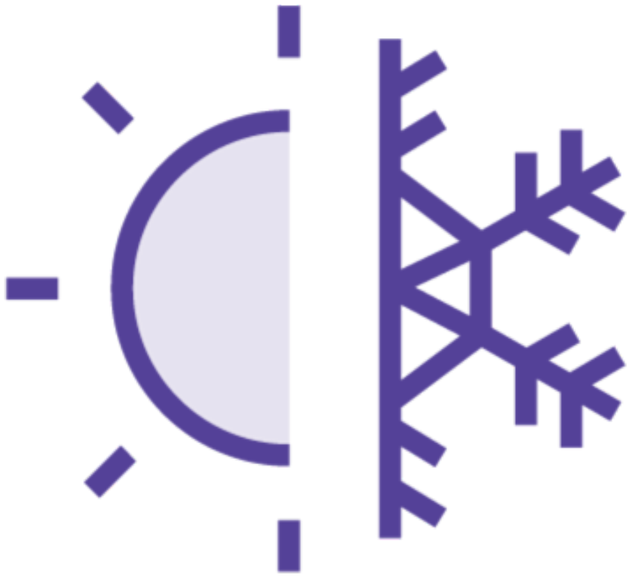
Data might be excessive in two ways

- Curse of dimensionality: Too many columns
- Outdated historical data: Too many rows

Problems with Data

Insufficient data	Too much data	**Non-representative data**
Missing data	Duplicate data	Outliers

Non-representative Data



Data not representative of the real world i.e. biased

Leads to biased models that perform poorly in practice

Mitigate using oversampling and undersampling

Problems with Data

Insufficient data

Too much data

Non-representative data

Missing data

Duplicate data

Outliers

Cleaning Data

Data cleaning procedures can help significantly mitigate effect of

- Missing data
- Outliers

Problems with Data

Insufficient data	Too much data	Non-representative data
Missing data	**Duplicate data**	Outliers

Duplicate Data

If data can flagged as duplicate, problem relatively easy to solve

- Simply de-duplicate

Can be hard to identify in some applications

- Real-time streaming

Missing Values and Outliers

Data Cleaning and Preparation

Missing Data

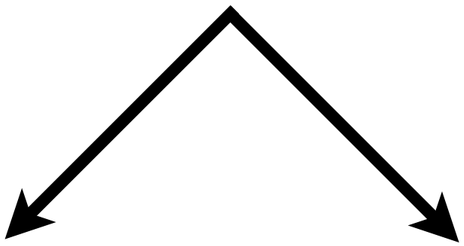
Outlier Data

Data Cleaning and Preparation

Missing Data

Outlier Data

Missing Data



Deletion

Imputation

Deletion a.k.a. Listwise Deletion

Delete an entire record (row) if a single value (column) is missing. Simple but can lead to bias.

Listwise Deletion

Most common method in practice

Can reduce sample size significantly

If values are not missing at random, can introduce significant bias

Imputation

Fill in missing column values, rather than deleting records with missing values. Missing values are inferred from known data.

Imputation

Methods range from very simple to very complex

Simplest method: Use column average

Can interpolate from nearby values

Can even build model to predict missing values

Multivariate Imputation

Univariate imputation: Rely only on known values in same feature

Multivariate imputation: Use all known data to infer missing data

- Construct regression models from other columns to predict this column
- Iterative repeat for all columns

Hot-deck Imputation

Sort records based on any criteria

For each missing value, use immediately prior available value

“Last Observation Carried Forward”

For time series, equivalent to assuming no change since last measurement

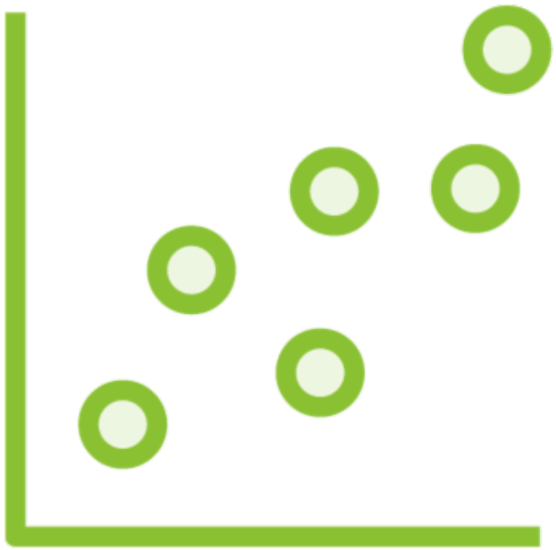
Mean Substitution

For each missing value, substitute mean of all available values

Has effect of weakening correlations between columns

Can be problematic when bivariate analysis required

Regression



Fit model to predict missing column based on other column values

Tends to strengthen correlations

Regression and mean substitution have complementary strengths

Data Cleaning and Preparation

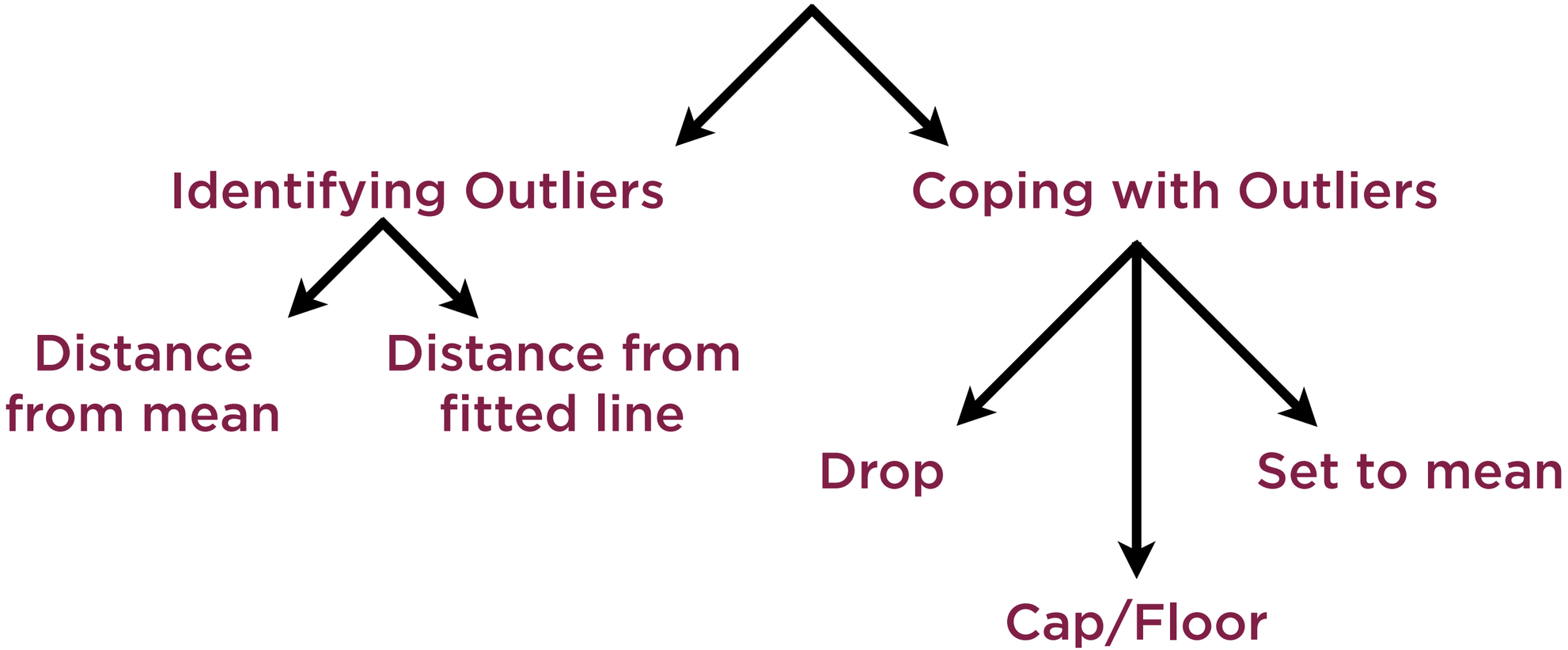
Missing Data

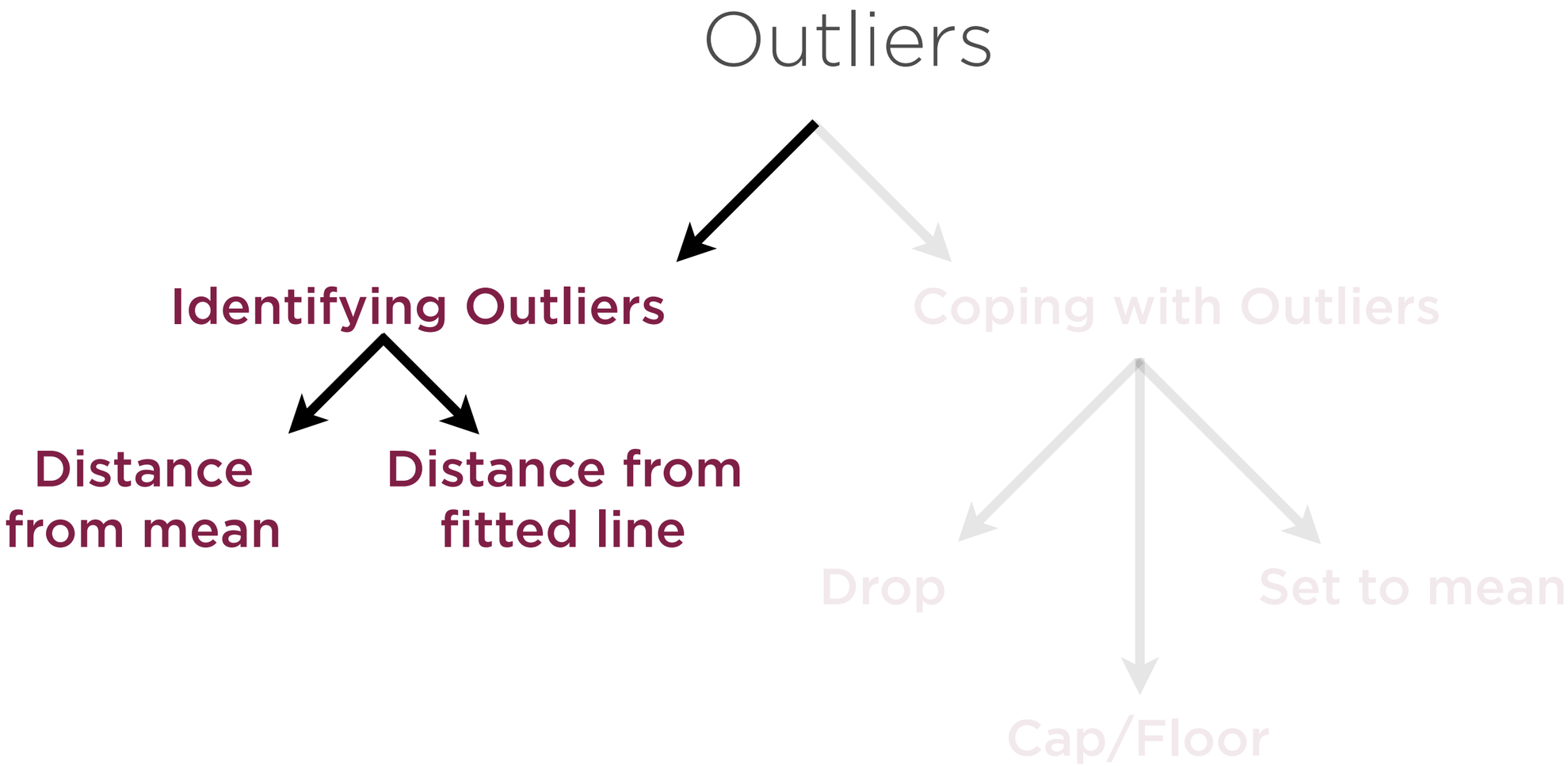
Outlier Data

Outlier

A data point that differs significantly from other data points in the same data set.

Outliers


Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Cap/Floor
Set to mean


Outliers
Identifying Outliers
Coping with Outliers
Distance from mean
Distance from fitted line
Drop
Set to mean
Cap/Floor

Identifying Outliers

Distance from mean

Distance from fitted line

Identifying Outliers

Distance from mean

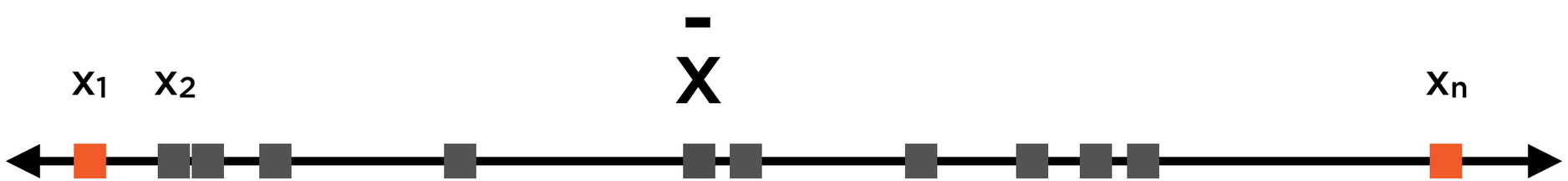
Distance from fitted line

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

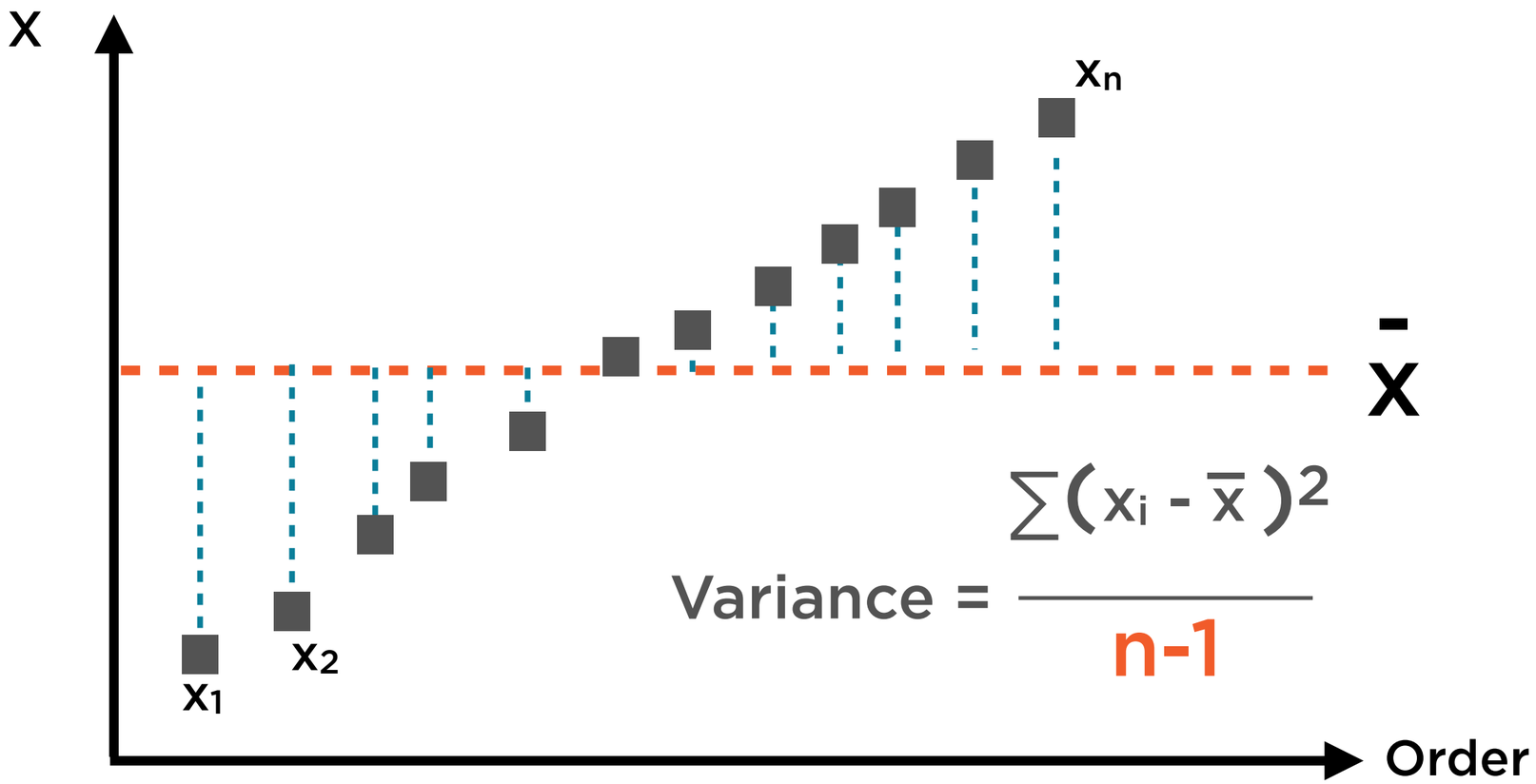


"Do the numbers jump around?"

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

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

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

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

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

Outliers

Points that lie more than 3 standard deviations from the mean are often considered outliers

Outliers

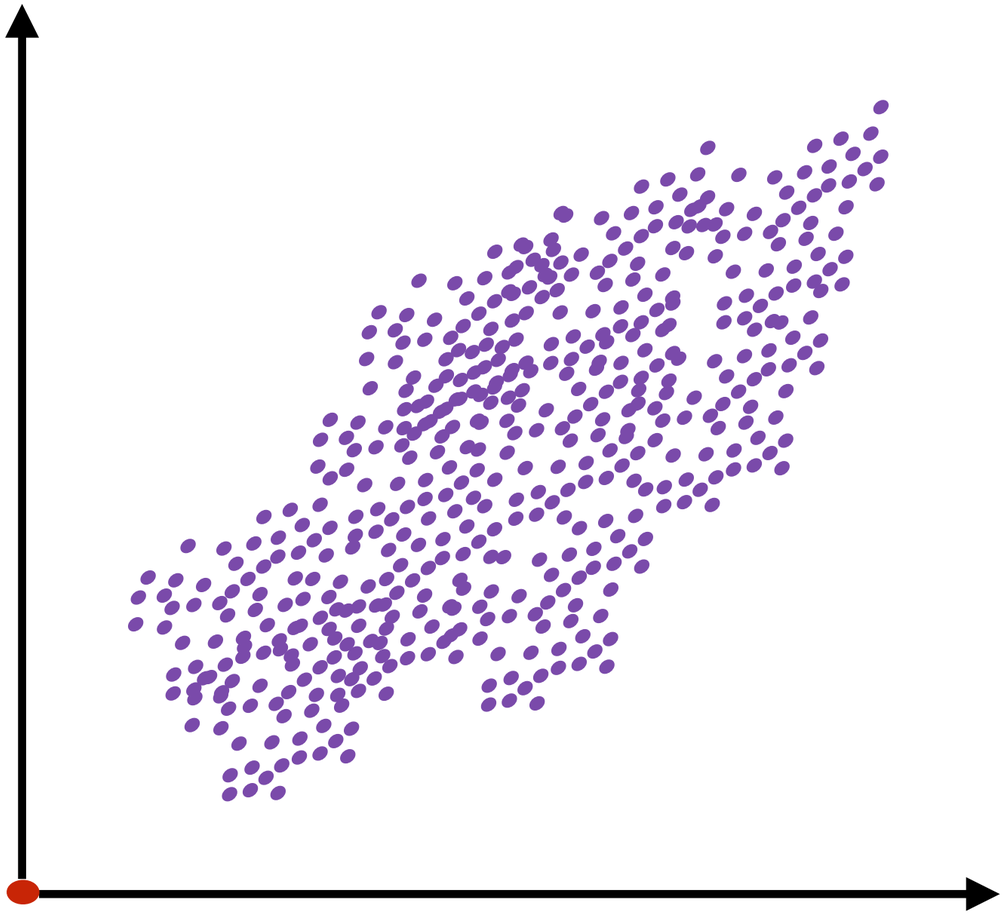
Points that lie more than 3 standard deviations from the mean are often considered outliers

Identifying Outliers

Distance from mean

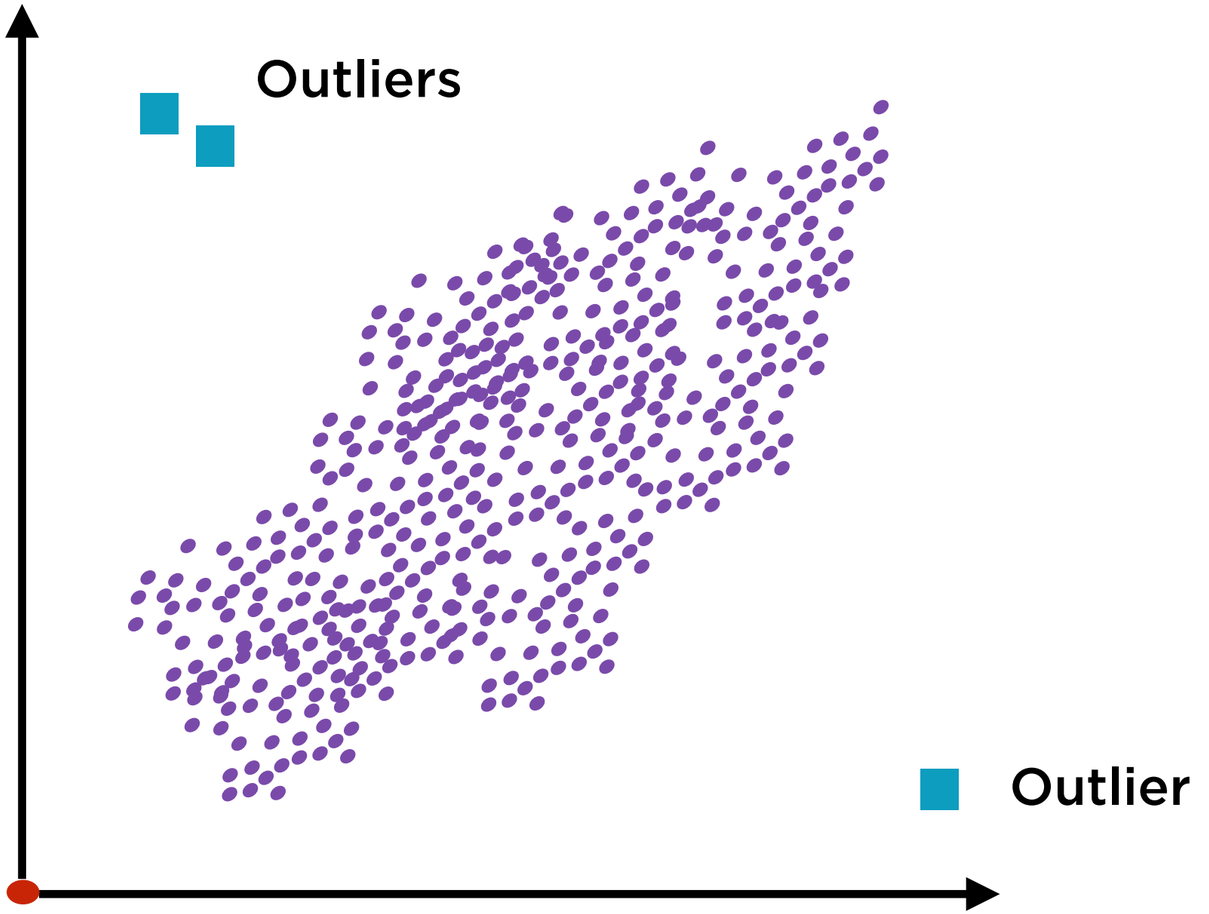
Distance from fitted line

Outliers



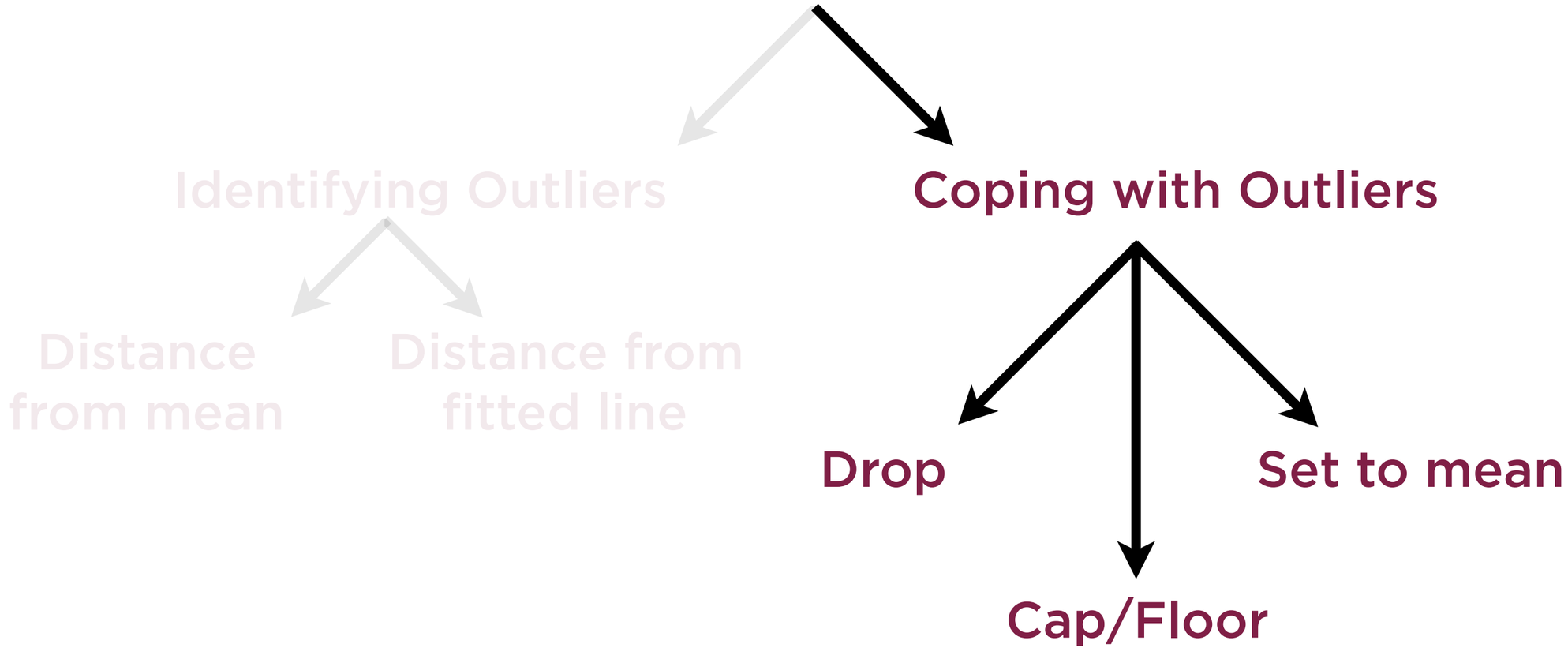
Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers

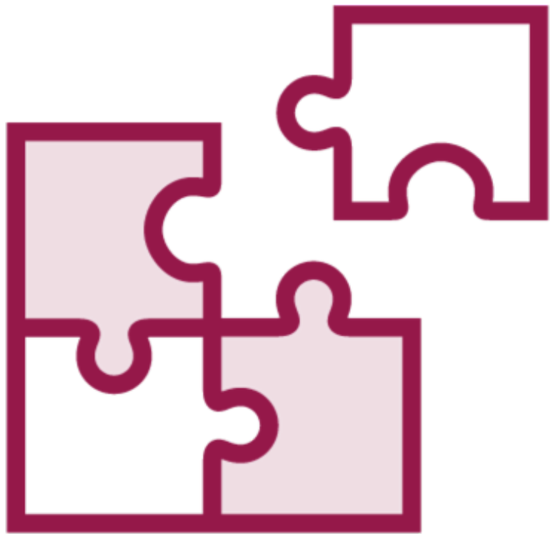


Outliers might also be data points that do not fit into the same relationship as the rest of the data

Outliers


Identifying Outliers
Distance from mean
Distance from fitted line
Coping with Outliers
Drop
Set to mean
Cap/Floor

Coping with Outliers

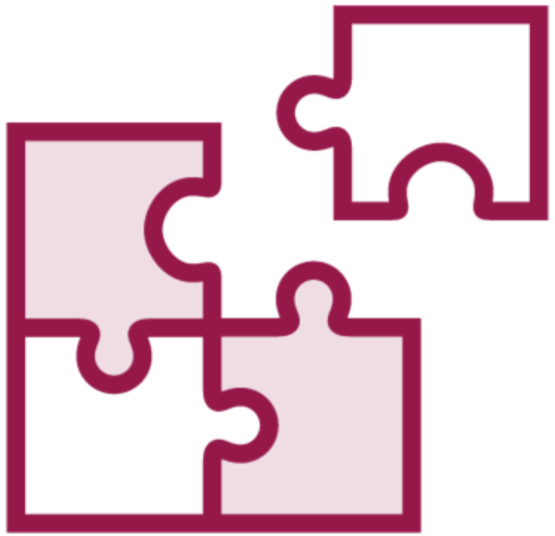


Always start by scrutinizing outliers

If erroneous observation

- Drop if all attributes of that point are erroneous
- Set to mean if only one attribute is erroneous

Coping with Outliers



If genuine, legitimate outlier

- Leave as-is if model not distorted
- Cap/Floor if model is distorted
    - Need to first standardize data
    - Cap positive outliers to +3
    - Floor negative outliers to -3

Demo

Applying different techniques to handle missing values in data

Demo

Detecting and handling outliers in data

Demo

Reading, visualizing, and exploring the dataset

Performing regression analysis for price prediction

Summary

Problems working with data

Dealing with outliers and missing values

Reading, visualizing and exploring a dataset

Building a simple predictive model