Handling missing data

PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON

Lisa Stuart
Data Scientist

datacamp

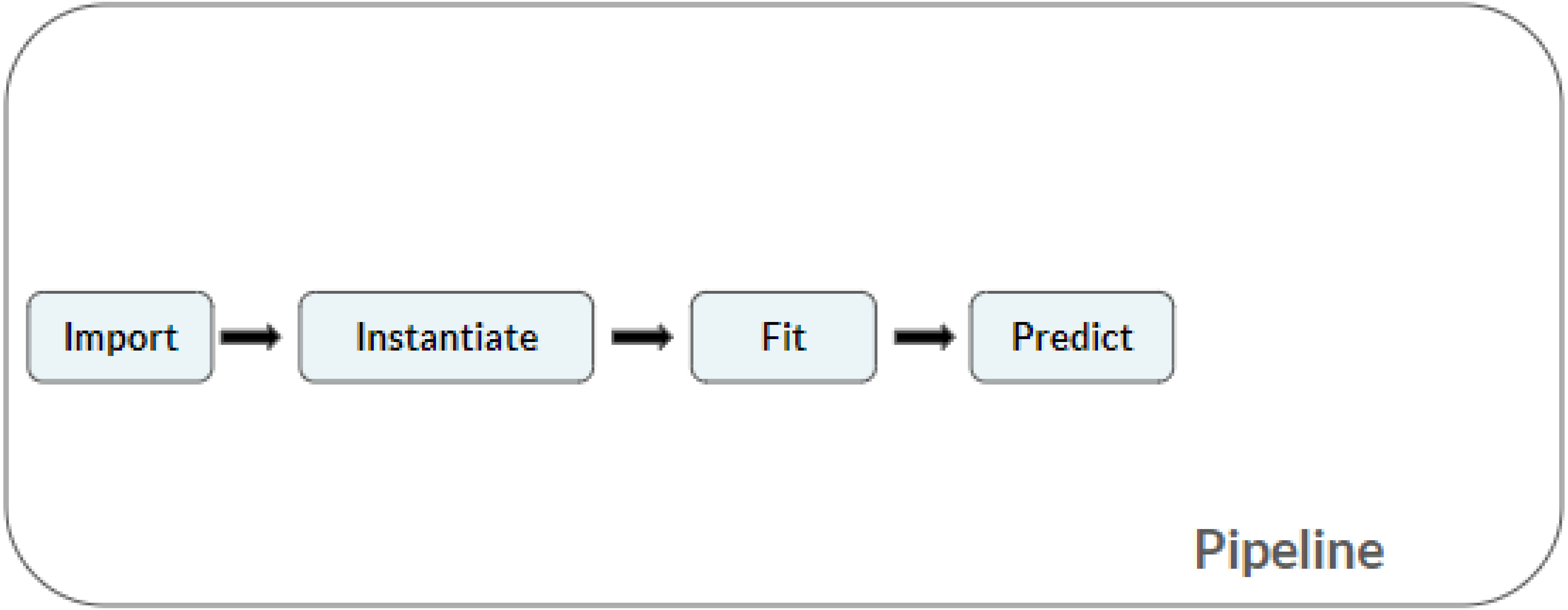
Prerequisites

- Supervised Learning with scikit-learn
- Unsupervised Learning in Python

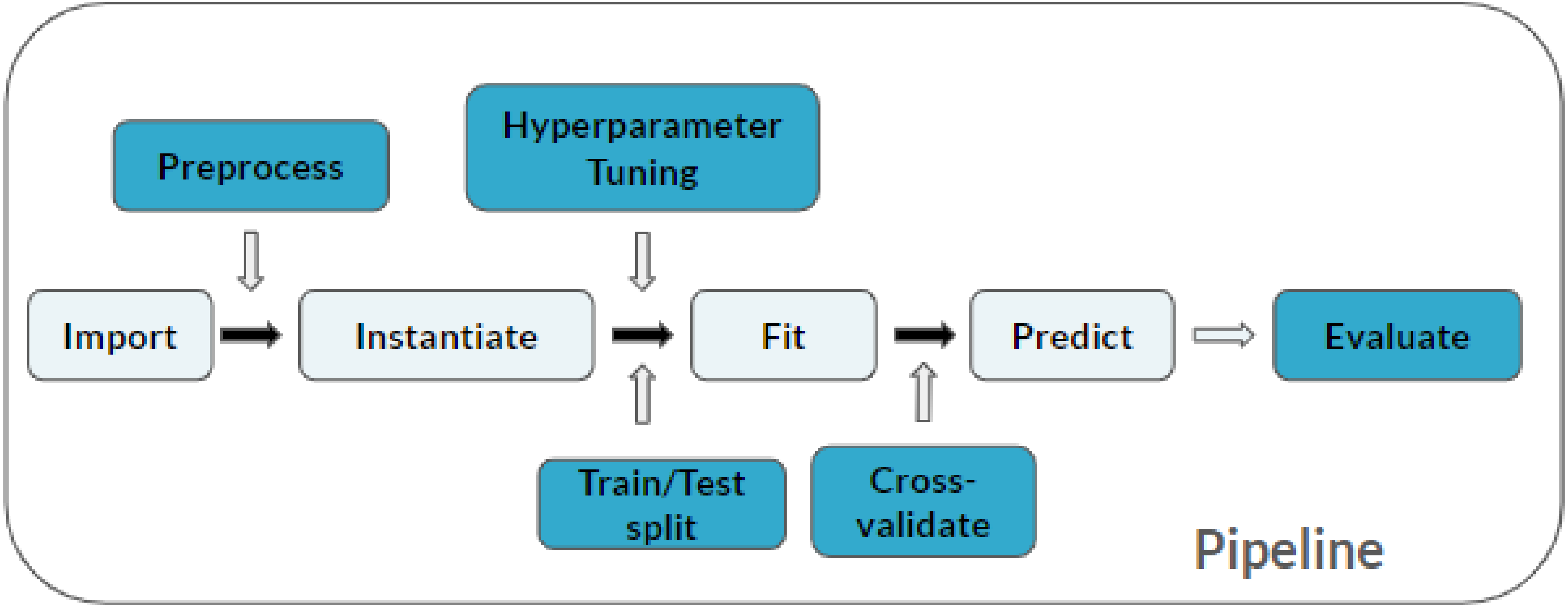
Course outline

- **Chapter 1: Pre-processing and Visualization**
    - Missing data, Outliers, Normalization
- **Chapter 2: Supervised Learning**
    - Feature selection, Regularization, Feature engineering
- **Chapter 3: Unsupervised Learning**
    - Cluster algorithm selection, Feature extraction, Dimension reduction
- **Chapter 4: Model Selection and Evaluation**
    - Model generalization and evaluation, Model selection

Machine learning (ML) pipeline



Our ML pipeline



Missing data

- Impact of different techniques
- Finding missing values
- Strategies to handle

Techniques

1. Omission
   - Removal of rows --> `.dropna(axis=0)`
   - Removal of columns --> `.dropna(axis=1)`
2. Imputation
   - Fill with zero --> `SimpleImputer(strategy='constant', fill_value=0)`
   - Impute mean -> `SimpleImputer(strategy='mean')`
   - Impute median --> `SimpleImputer(strategy='median')`
   - Impute mode --> `SimpleImputer(strategy='most_frequent')`
   - Iterative imputation --> `IterativeImputer()`

Why bother?

- Reduce the probability of introducing bias
- Most ML algorithms require complete data

Effects of imputation

- Depend on:
    - Missing values
    - Original variance
    - Presence of outliers
    - Size and direction of skew
- Omission --> Can remove too much
- Zero --> Bias results downward
- Mean --> Affected more by outliers
- Median --> Better in case of outliers
- Mode and iterative imputation--> Try them out

Function	returns
`df.isna().sum()`	number missing
`df['feature'].mean()`	feature mean
`.shape`	row, column dimensions
`df.columns`	column names
`.fillna(0)`	fills missing with 0
`select_dtypes(include = [np.number] )`	numeric columns
`select_dtypes(include = ['object'] )`	string columns
`.fit_transform(numeric_cols)`	fits and transforms

Effects of missing values

What are the effects of missing values in a Machine Learning (ML) setting? Select the answer that is **true**:

- Missing values aren't a problem since most of `sklearn` 's algorithms can handle them.
- Removing observations or features with missing values is generally a good idea.
- Missing data tends to introduce bias that leads to misleading results so they cannot be ignored.
- Filling missing values with zero will bias results upward.

Effect of missing values: answer

What are the effects of missing values in a Machine Learning (ML) setting? The correct answer is:

- **Missing data tends to introduce bias that leads to misleading results so they cannot be ignored.** (Filling missing values by testing which impacts the variance of a given dataset the least is the best approach.)

Effects of missing values: incorrect answers

What are the effects of missing values in a Machine Learning (ML) setting?

- Missing values aren't a problem... (Most of `sklearn` 's algorithms cannot handle missing values and will throw an error.)
- Removing observations or features with missing values... (Unless your dataset is large and the proportion of missing values small, removing rows or columns with missing data usually results in shrinking your dataset too much to be useful in subsequent ML.)
- Filling missing values with zero will bias results upward.(It's the opposite, filling with zero will bias results downward.)

Let's practice!

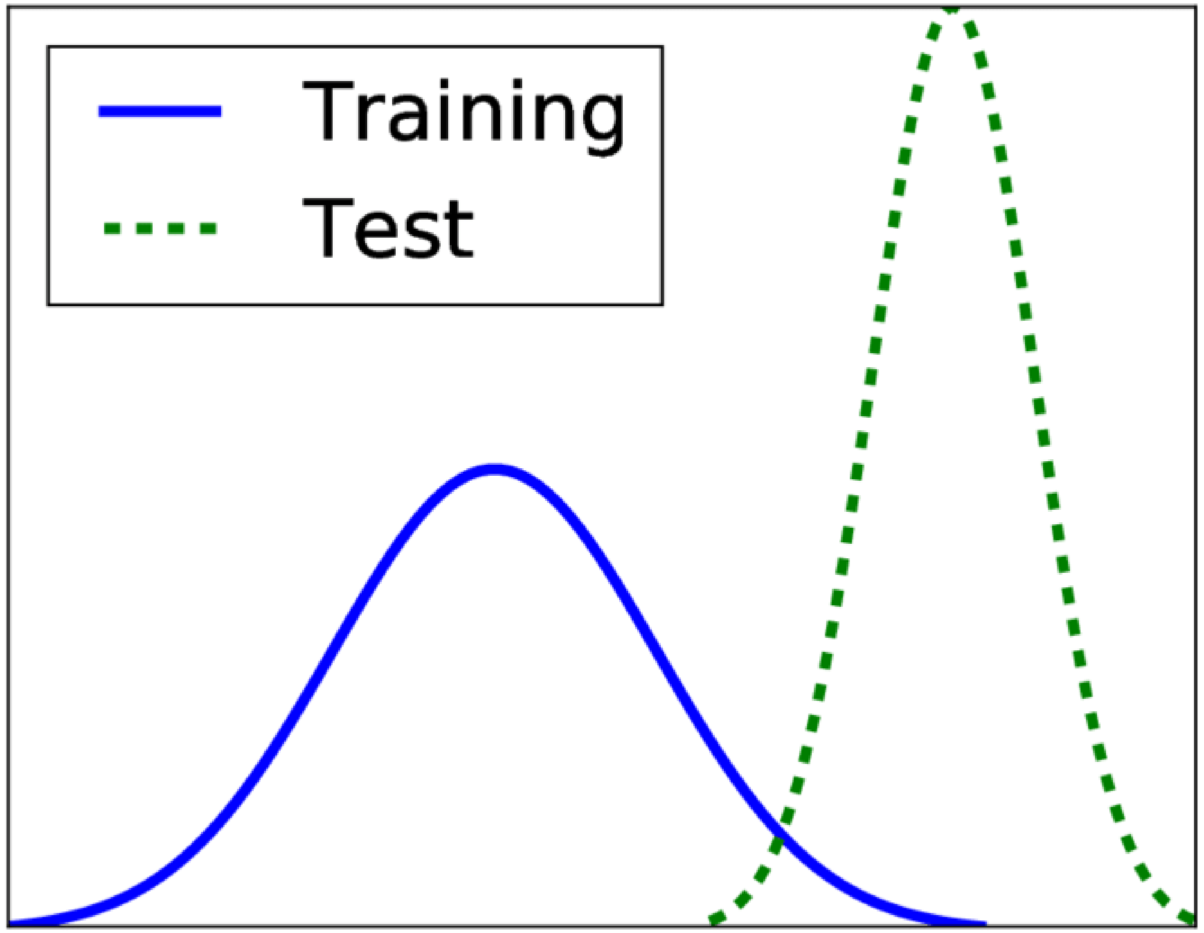
PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON

Data distributions and transformations

PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON

Lisa Stuart
Data Scientist

Different distributions



[1] https://www.researchgate.net/figure/Bias-Training-and-test-data-sets-are-drawn-from-different-distributions_fig22_330485084

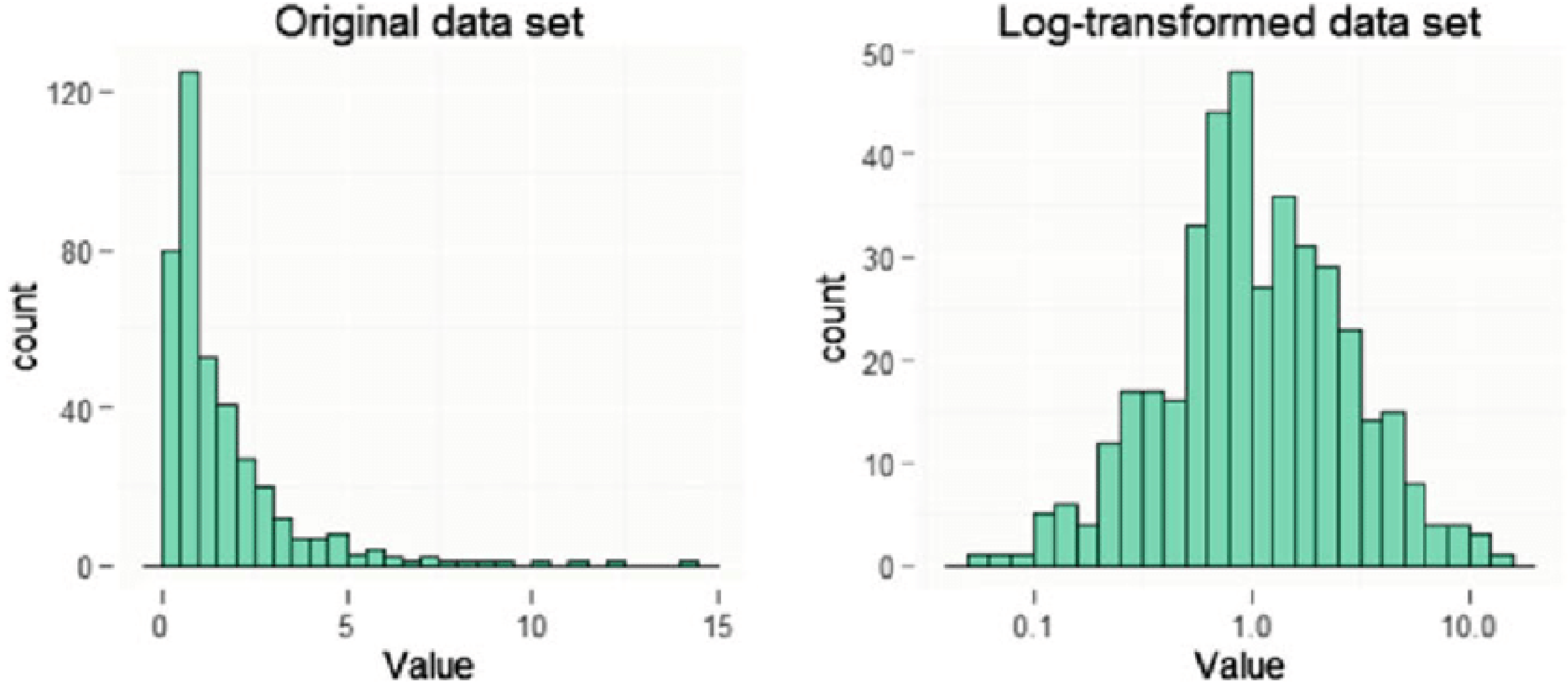
Train/test split

```
train, test = train_test_split(X, y, test_size=0.3)


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

`sns.pairplot()` --> plot matrix of distributions and scatterplots

Data transformation



[1] https://www.researchgate.net/figure/Example-of-the-effect-of-a-log-transformation-on-the-distribution-of-the-dataset_fig20_308007227

Box-Cox Transformations

```
scipy.stats.boxcox(data, lmbda= )
```

`lmbda` (p)	x^p	**transform**
-2	$x^{-2} = 1/2$	reciprocal square
-1	$x^{-1} = 1/x$	reciprocal
-0.5	$x^{-1/2} = 1/\sqrt{x}$	reciprocal square root
0.0	$\log(x)$	log
0.5	$x^{1/2} = \sqrt{x}$	square root
1	$x^1 = x$	no transform
2	$x^2 = x$	square

Let's practice!

PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON

Data outliers and scaling

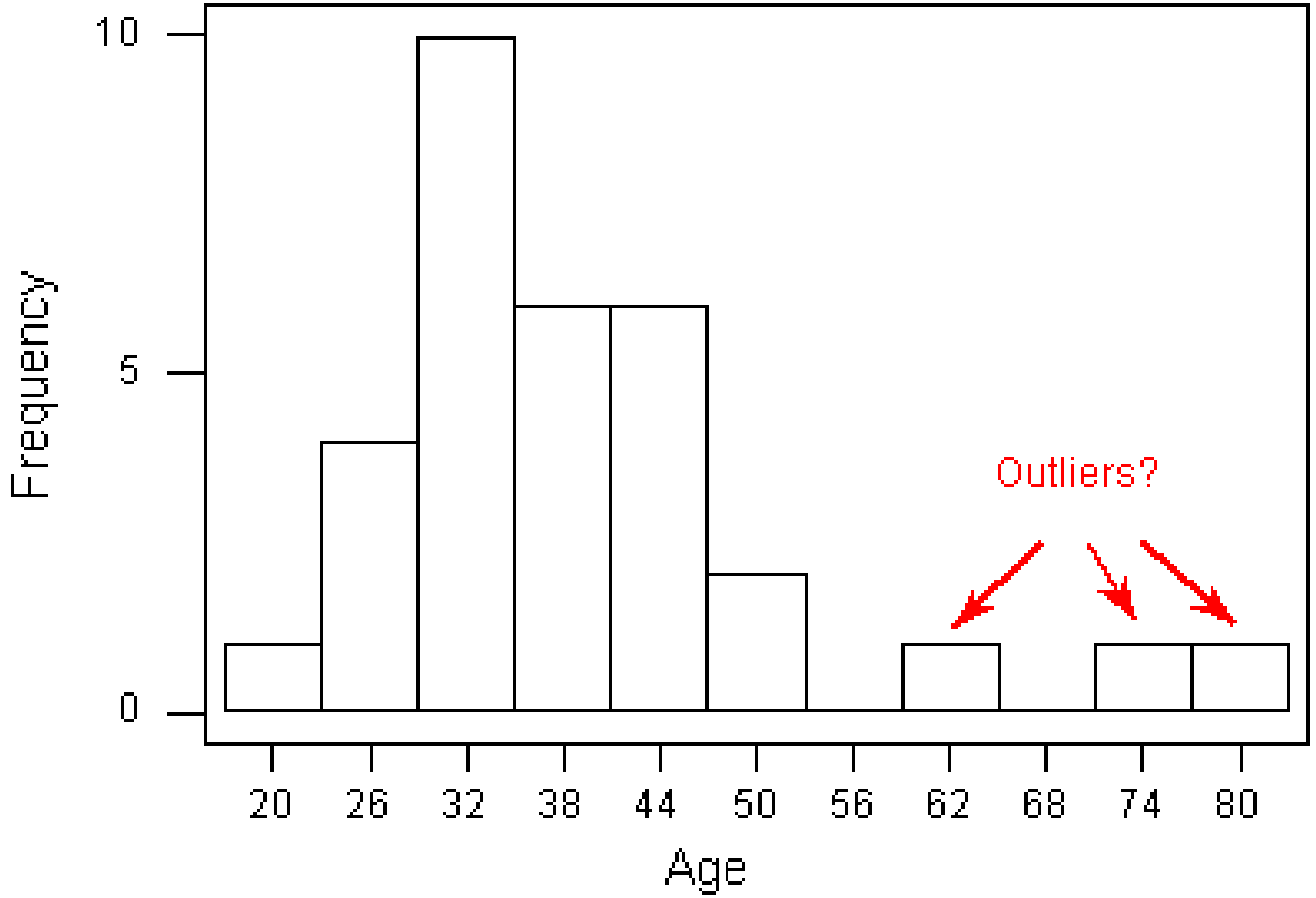
PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON

Lisa Stuart
Data Scientist

datacamp

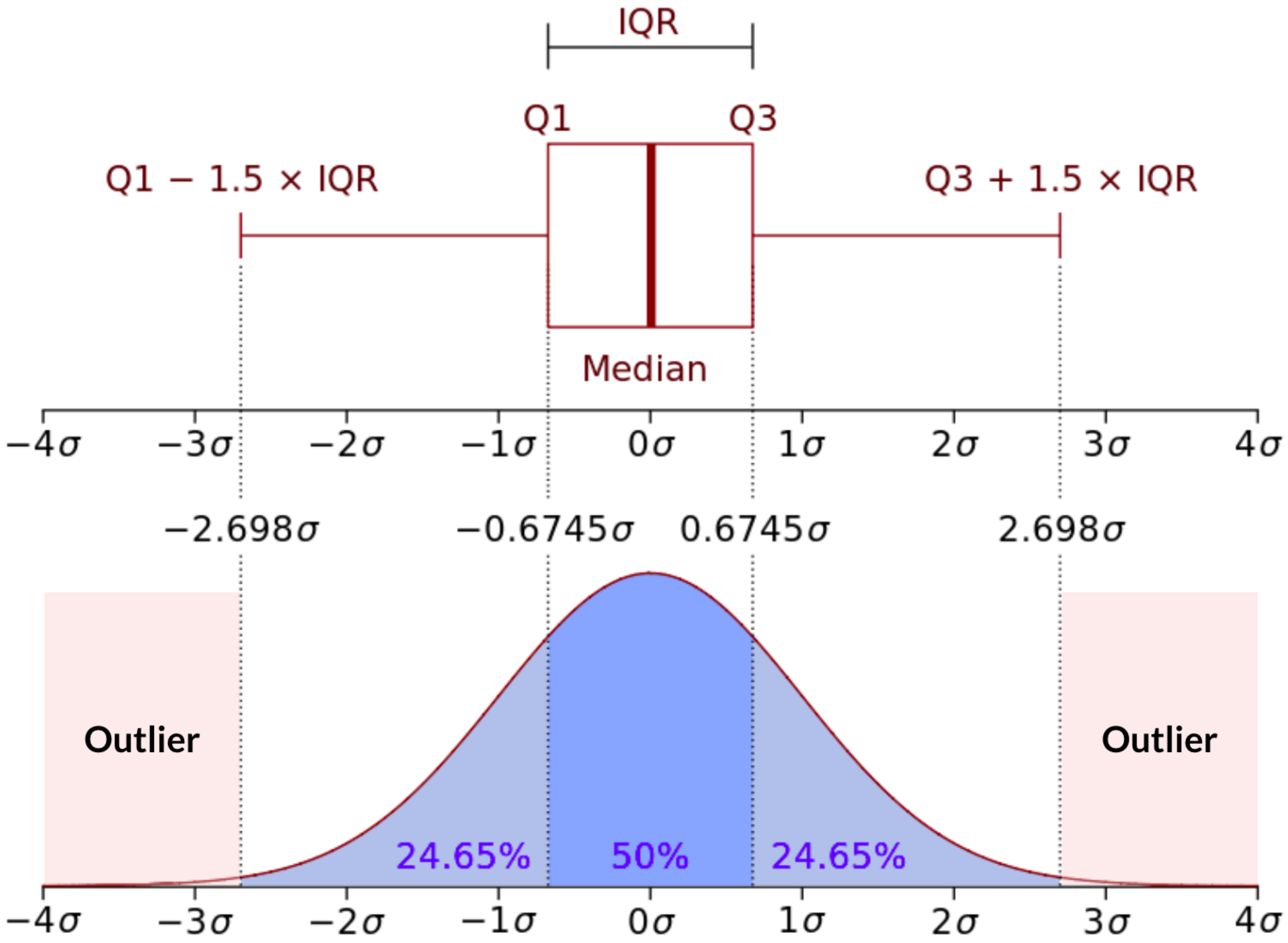
Outliers

- One or more observations that are distant from the rest of the observations in a given feature.



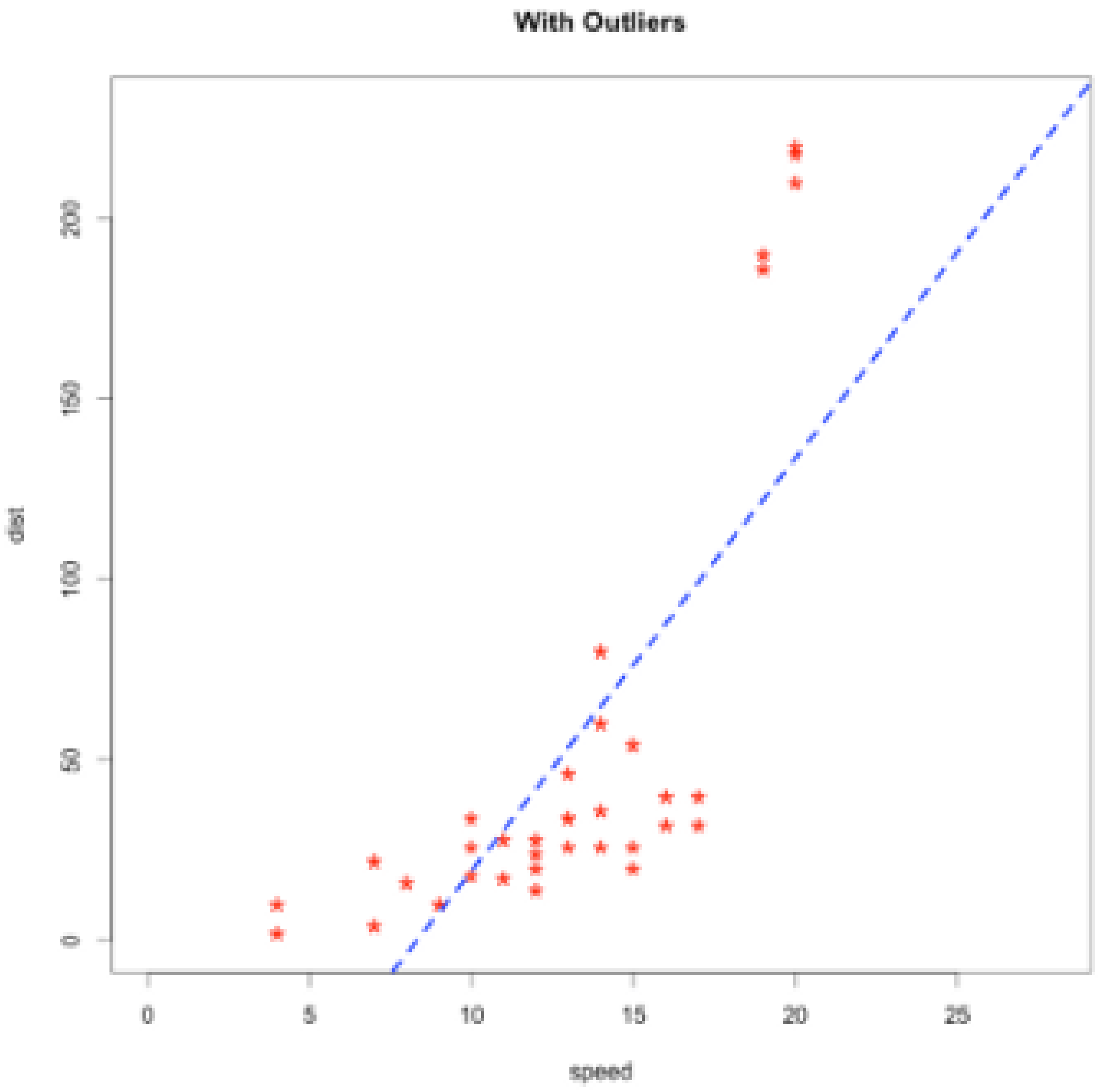
[1] https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/understanding-outliers/

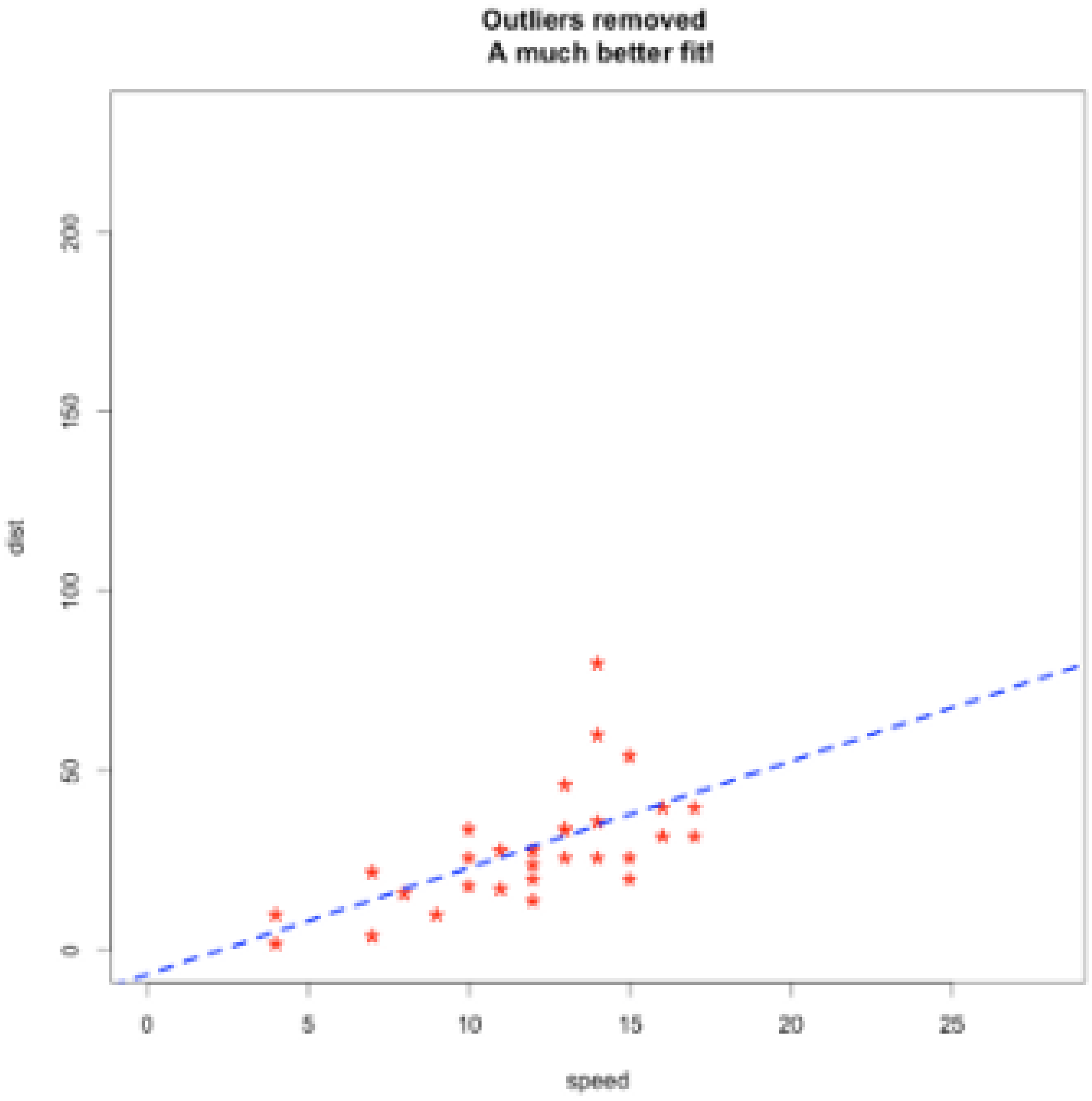
Inter-quartile range (IQR)



[1] By Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285

Line of best fit



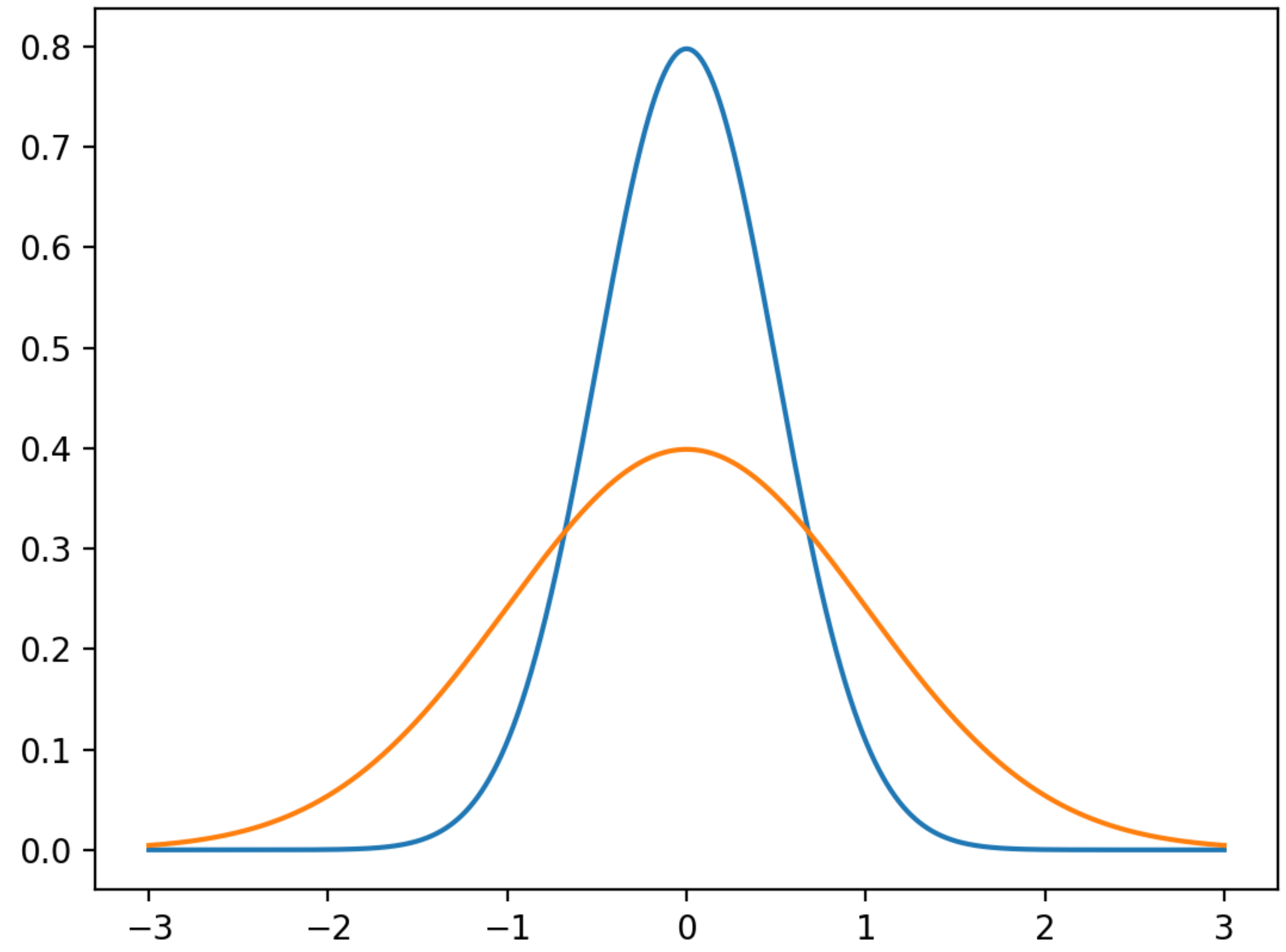


[1] https://www.r-bloggers.com/outlier-detection-and-treatment-with-r/

Outlier functions

Function	returns
`sns.boxplot(x= , y='Loan Status')`	boxplot conditioned on target variable
`sns.distplot()`	histogram and kernel density estimate (kde)
`np.abs()`	returns absolute value
`stats.zscore()`	calculated z-score
`mstats.winsorize(limits=[0.05, 0.05])`	floor and ceiling applied to outliers
`np.where(condition, true, false)`	replaced values

High vs low variance



[1] https://machinelearningmastery.com/a-gentle-introduction-to-calculating-normal-summary-statistics/

Standardization vs normalization

- Standardization:
  - Z-score standardization
  - Scales to mean 0 and sd 1

$$z = \frac{x_i - \mu}{\sigma}$$

- Normalization:
  - Min/max normalizing
  - Scales to between (0, 1)

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

[1] https://medium.com/@rrfd/standardize-or-normalize-examples-in-python-e3f174b65dfc

Scaling functions

- `scikit-learn.preprocessing.StandardScaler()` --> (mean=0, sd=1)
- `sklearn.preprocessing.MinMaxScaler()` --> (0,1)

Outliers and scaling

How should outliers be identified and properly dealt with? What result does min/max or z-score standardization have on data? Select the statement that is **true**:

- An outlier is a point that is just outside the range of similar points in a feature.
- In a given context, outliers considered anomalous are helpful in building a predictive ML model.
- Mix/max scaling gives data a mean of 0, an SD of 1, and increases variance.
- Z-score standardization scales data to be in the interval (0,1) and improves model fit.

Outliers and scaling: answer

How should outliers be identified and properly dealt with? What result does min/max or z-score standardization have on data? The correct answer is:

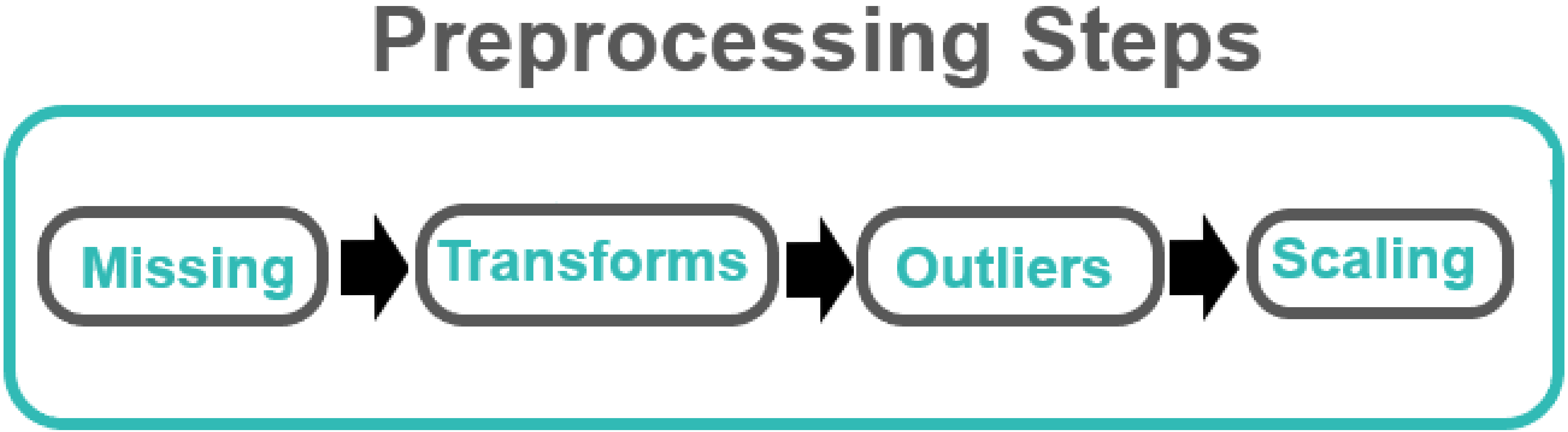
- **In a given context, outliers considered anomalous are helpful in building a predictive ML model.** (Data anomalies are common in fraud detection, cybersecurity events, and other scenarios where the goal is to find them.)

Outliers and scaling: incorrect answers

How should outliers be identified and properly dealt with? What result does min/max or z-score standardization have on data?

- An outlier is just outside the range of similar points in a feature. (A point is not suspected of being an outlier until more than 1.5 times beyond the IQR.)
- Mix/max scaling gives data a mean of 0, an SD of 1, and increases variance. (Min/max scaling scales data to be in the interval (0,1) and it depends on the original data whether or not variance is increased or decreased.)
- Z-score standardization scales data to be in the interval (0,1) and improves model fit. (Z-score standardization scales the data to have mean 0 and sd of 1, which can improve model fit.)

One last thing...



Let's practice!

PRACTICING MACHINE LEARNING INTERVIEW QUESTIONS IN PYTHON