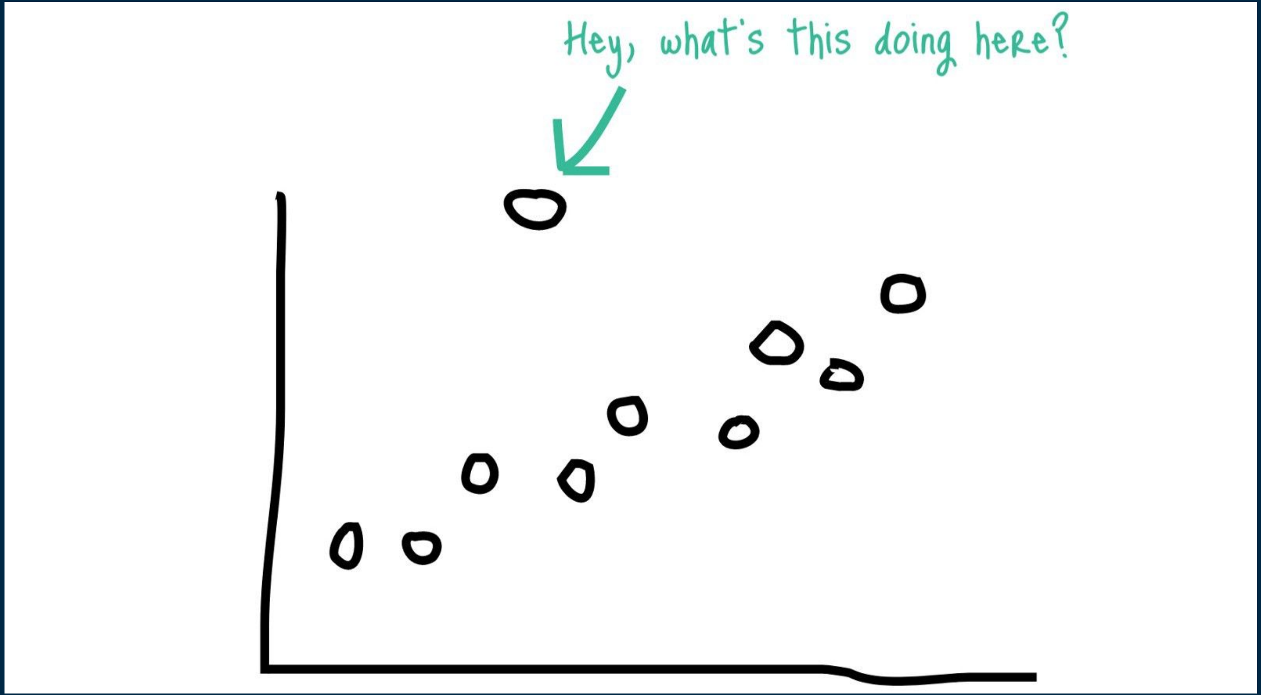
Outlier Detection Analysis

Jiyan Aytek

If your data is bad, your machine learning tools are useless

What is an outlier?



Most Common Causes of Outliers

01

HUMAN ERRORS

DATA ENTRY ERRORS

02

INSTRUMENT ERRORS

MEASUREMENT ERRORS

03

EXPERIMENTAL ERRORS

DATA EXTRACTION OR
EXECUTING ERRORS

Most Common Causes of Outliers

04

DATA PROCESSING ERRORS

DATA MANIPULATION

05

SAMPLING ERRORS

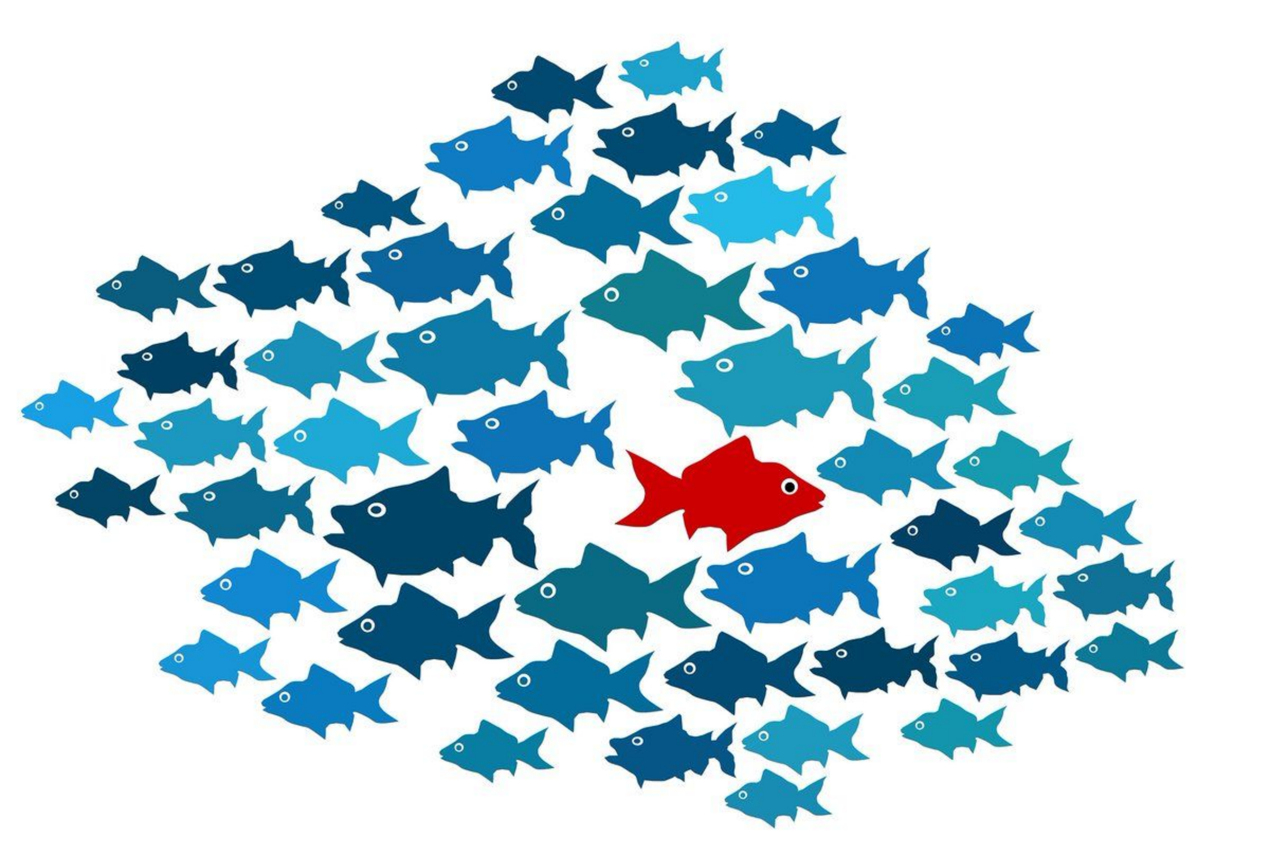
VARIOUS SOURCES

06

NATURAL

NOT AN ERROR

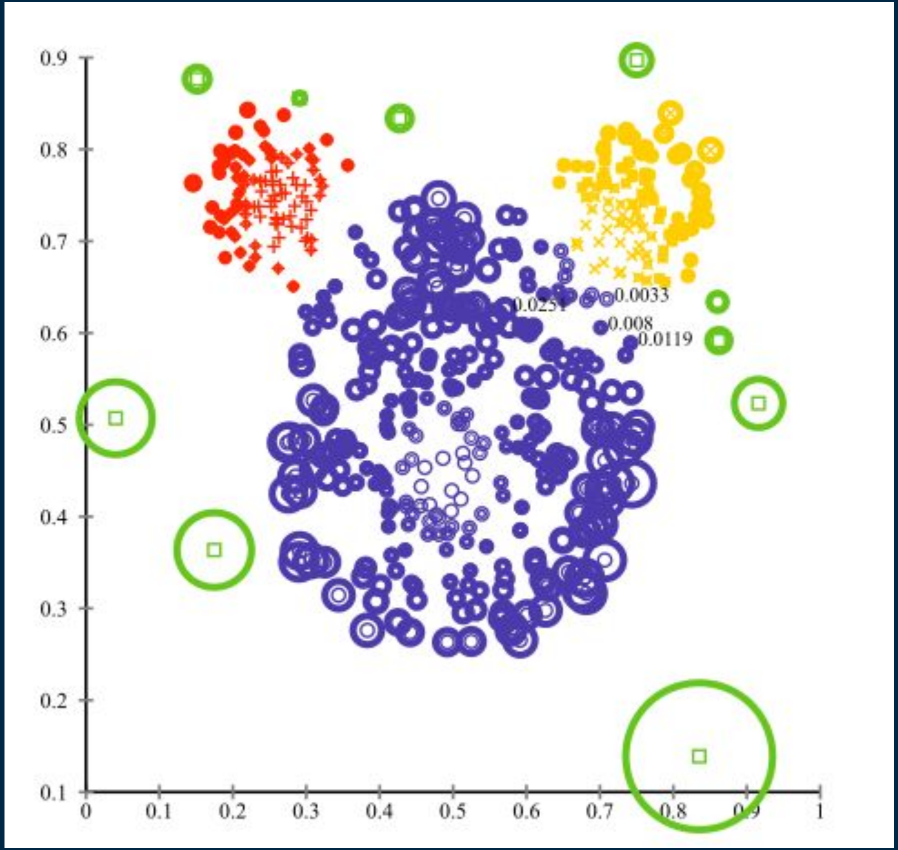
Outlier Example



> “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by different mechanism ”

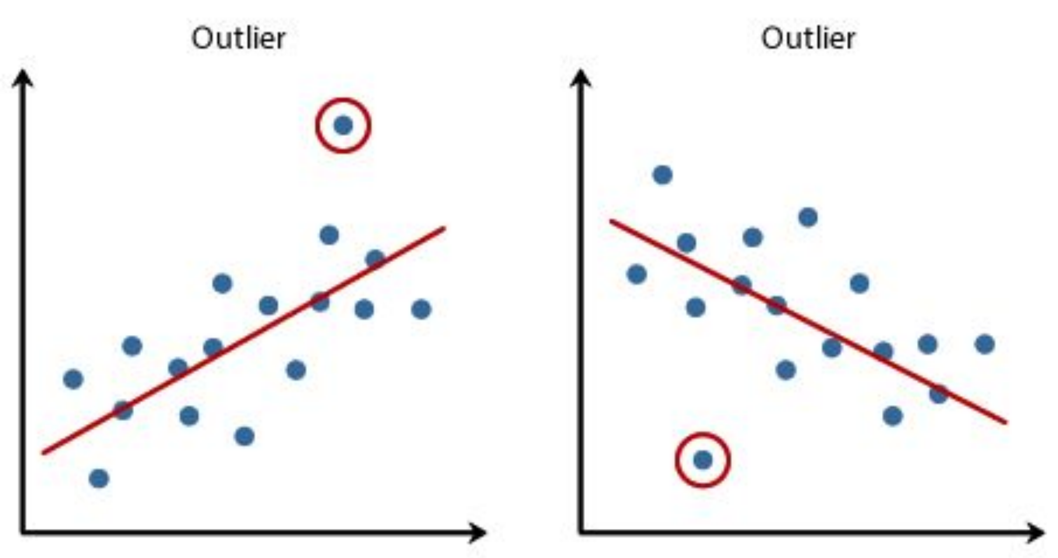
—Hawkins,1980

Clustering

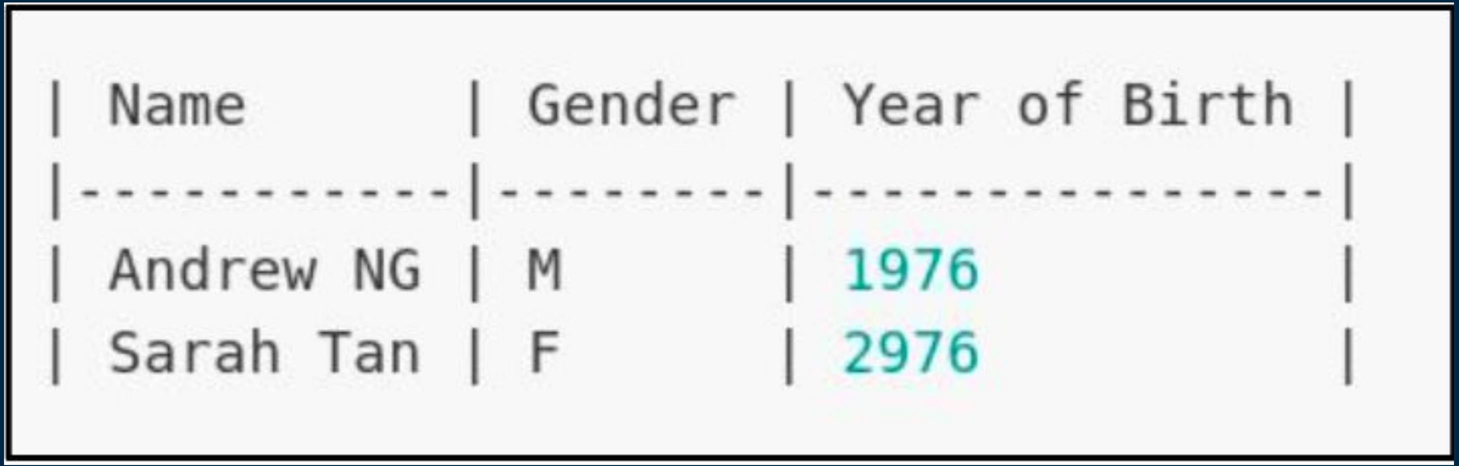


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.0254
0.0033
0.008
0.0119

Regression



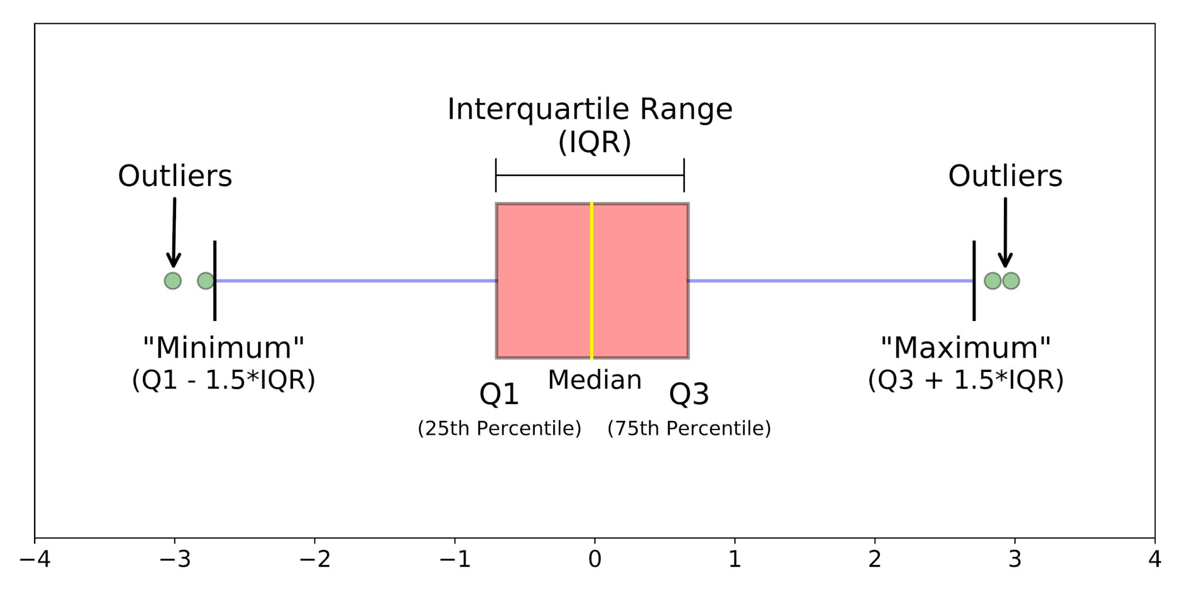
Invalid or Outlier Data

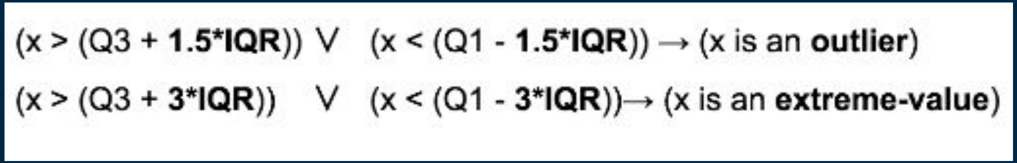


```
| Name       | Gender | Year of Birth |
|------------|--------|---------------|
| Andrew NG  | M      | 1976          |
| Sarah Tan  | F      | 2976          |
```

Detection Methods : Uni-variate Methods

Boxplot

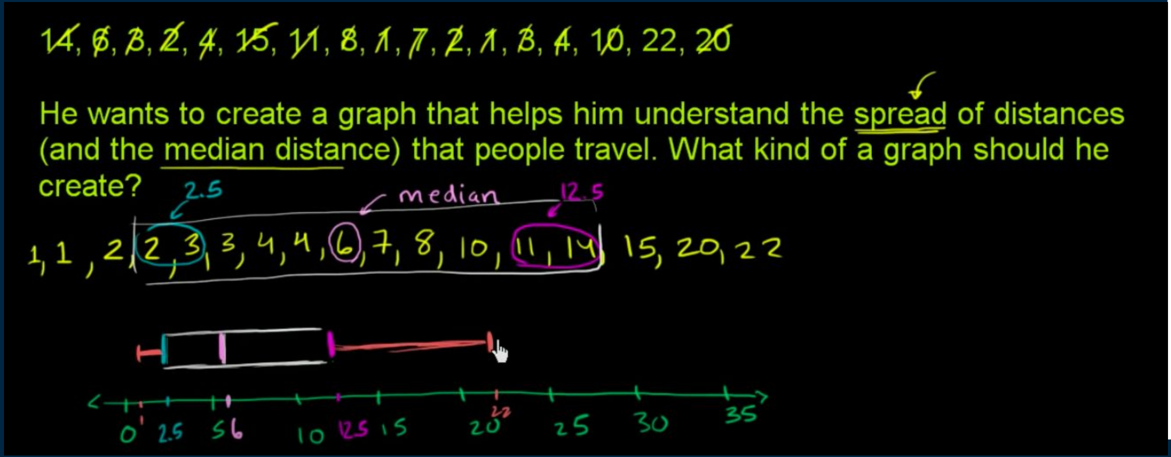




(x > (Q3 + **1.5*IQR**)) V (x < (Q1 - **1.5*IQR**)) → (x is an **outlier**)

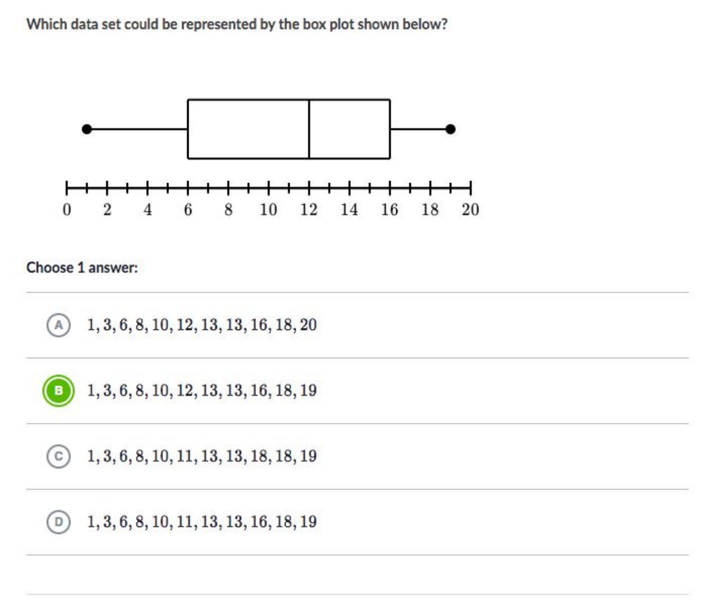
(x > (Q3 + **3*IQR**)) V (x < (Q1 - **3*IQR**))→ (x is an **extreme-value**)

Boxplot



~~14~~, ~~6~~, ~~3~~, ~~2~~, ~~4~~, ~~15~~, ~~11~~, ~~8~~, ~~1~~, ~~7~~, ~~2~~, ~~1~~, ~~3~~, ~~4~~, ~~10~~, 22, ~~20~~

He wants to create a graph that helps him understand the spread of distances (and the median distance) that people travel. What kind of a graph should he create?

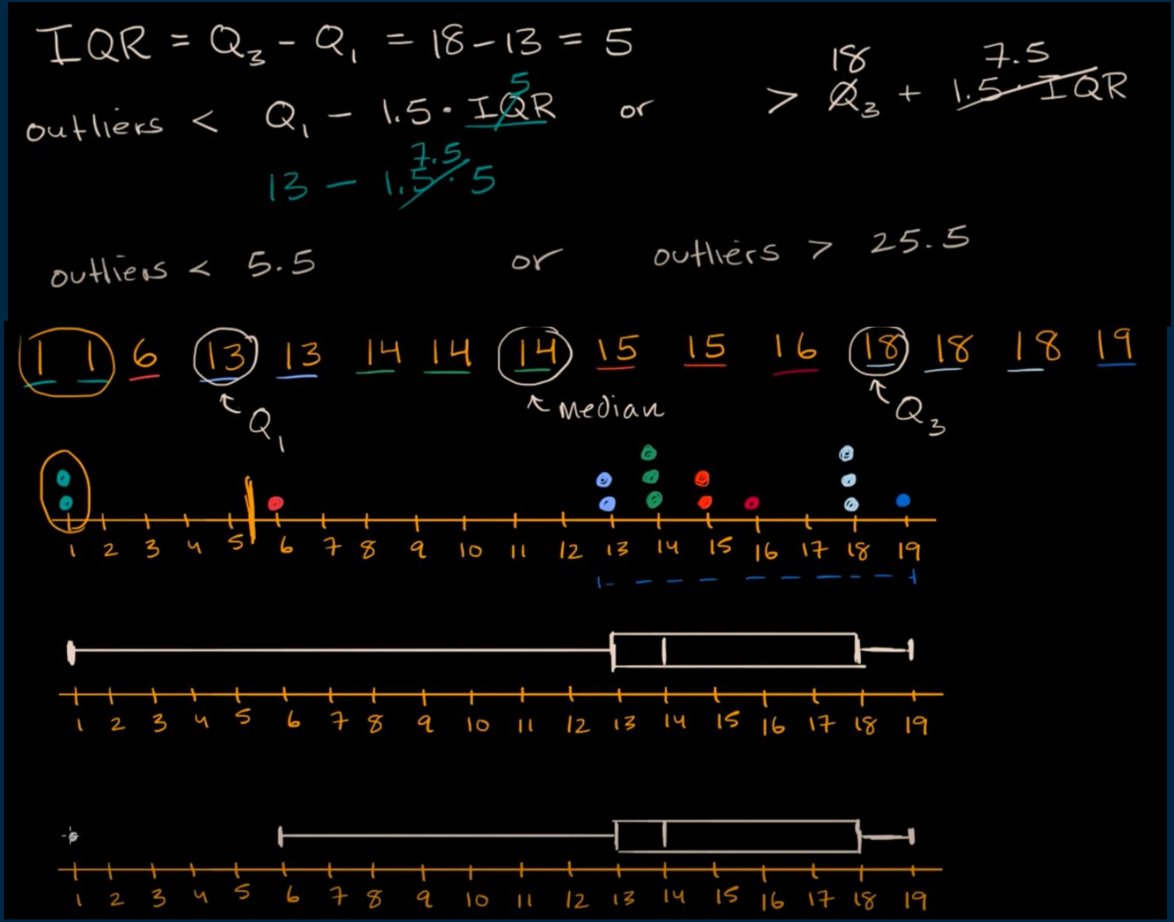


Which data set could be represented by the box plot shown below?

Choose 1 answer:

- (A) 1, 3, 6, 8, 10, 12, 13, 13, 16, 18, 20
- [✓ (B) 1, 3, 6, 8, 10, 12, 13, 13, 16, 18, 19]
- (C) 1, 3, 6, 8, 10, 11, 13, 13, 18, 18, 19
- (D) 1, 3, 6, 8, 10, 11, 13, 13, 16, 18, 19

Identifying outliers in a dataset (with Boxplot)



$IQR = Q_3 - Q_1 = 18 - 13 = 5$

outliers $< Q_1 - 1.5 \cdot \overset{5}{\cancel{IQR}}$ or $> \overset{18}{\cancel{Q_3}} + \overset{7.5}{\cancel{1.5 \cdot IQR}}$

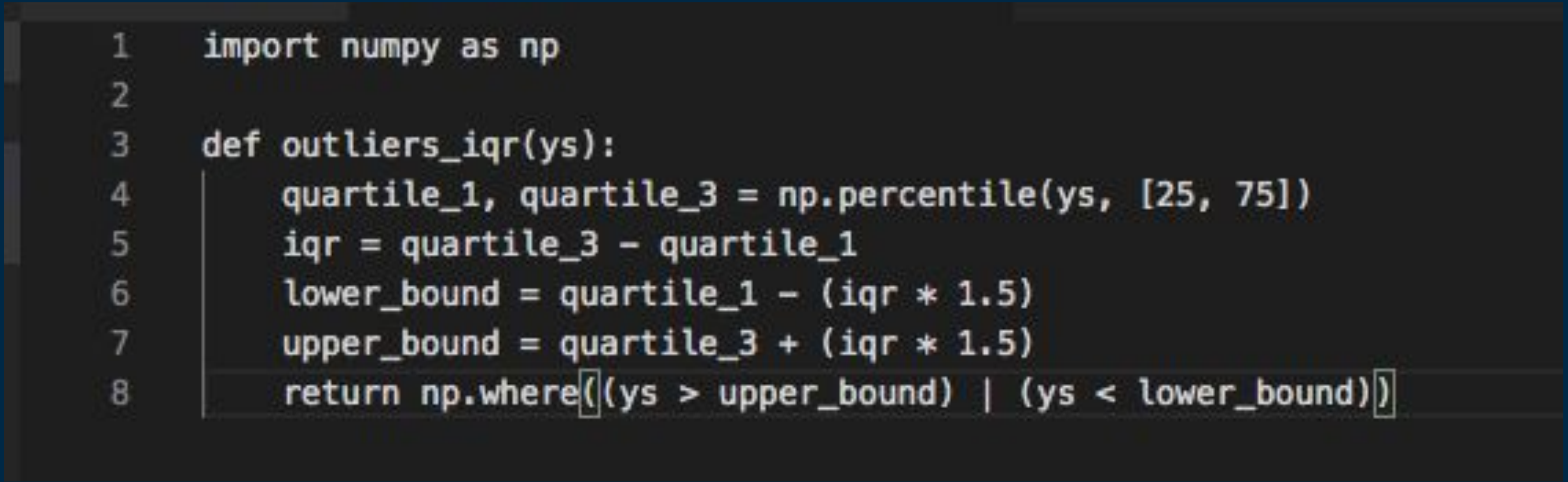
$13 - \overset{7.5}{\cancel{1.5 \cdot 5}}$

outliers < 5.5 or outliers > 25.5

1 1 6 13 13 14 14 14 15 15 16 18 18 18 19

↖ Q_1 ↖ Median ↖ Q_3

Boxplot



```python
import numpy as np

def outliers_iqr(ys):
    quartile_1, quartile_3 = np.percentile(ys, [25, 75])
    iqr = quartile_3 - quartile_1
    lower_bound = quartile_1 - (iqr * 1.5)
    upper_bound = quartile_3 + (iqr * 1.5)
    return np.where((ys > upper_bound) | (ys < lower_bound))
```

Standard Deviation

Using Standard Deviation

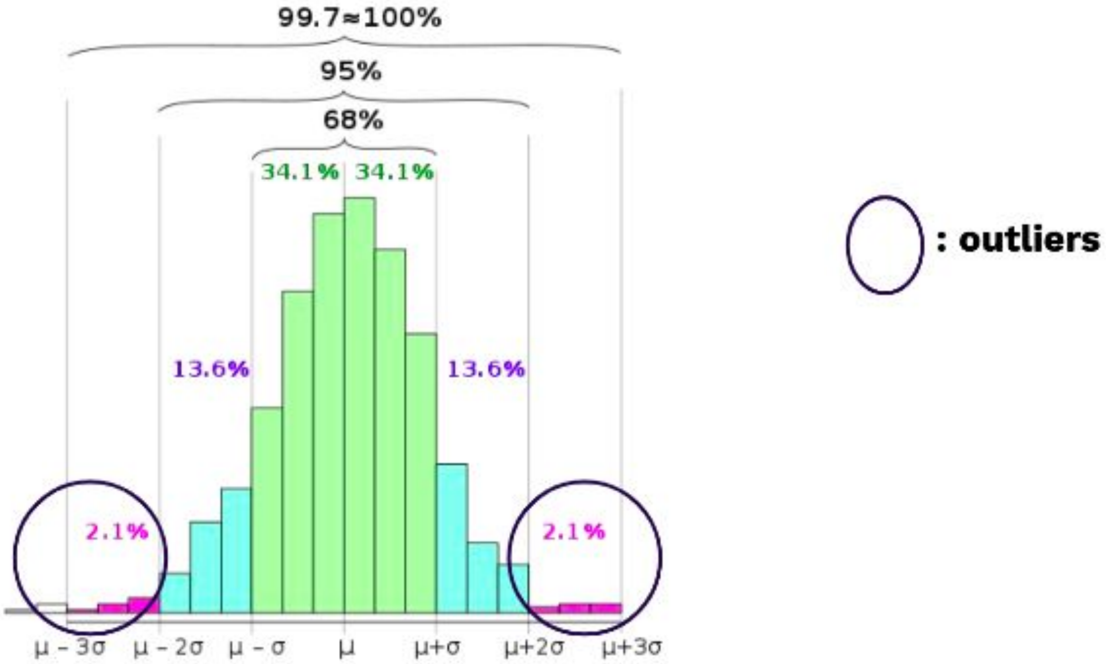
For value X=x:

$(x > (\mu + \mathbf{2\sigma})) \lor (x < (\mu - \mathbf{2\sigma})) \rightarrow$

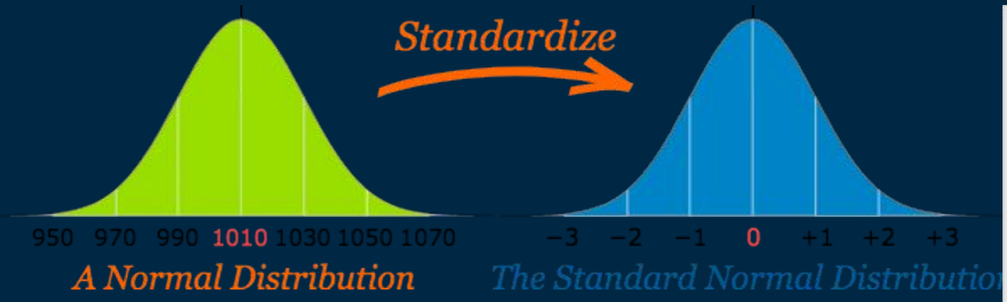
(x is an **outlier**)

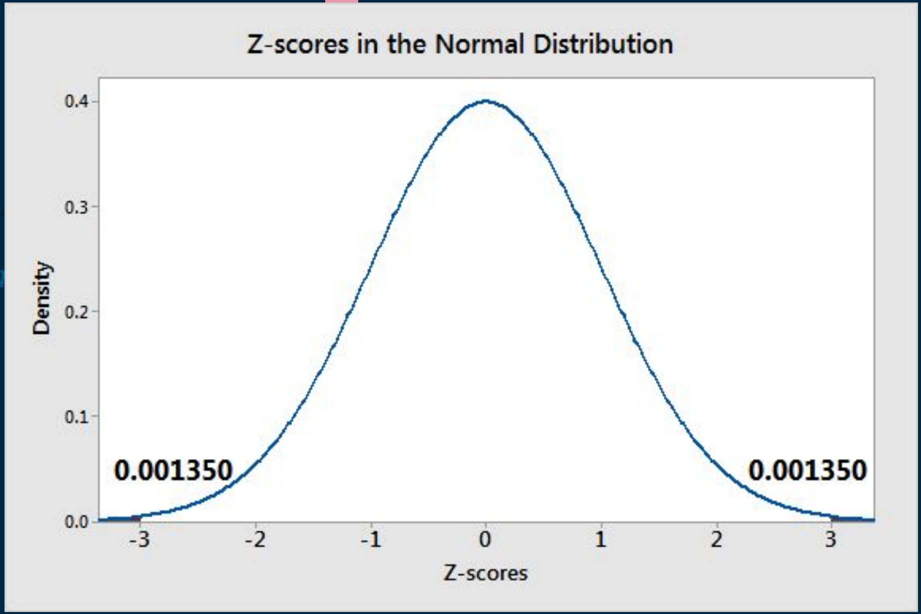
$(x > (\mu + \mathbf{3\sigma})) \lor (x < (\mu - \mathbf{3\sigma})) \rightarrow$

(x is an **extreme-value**)



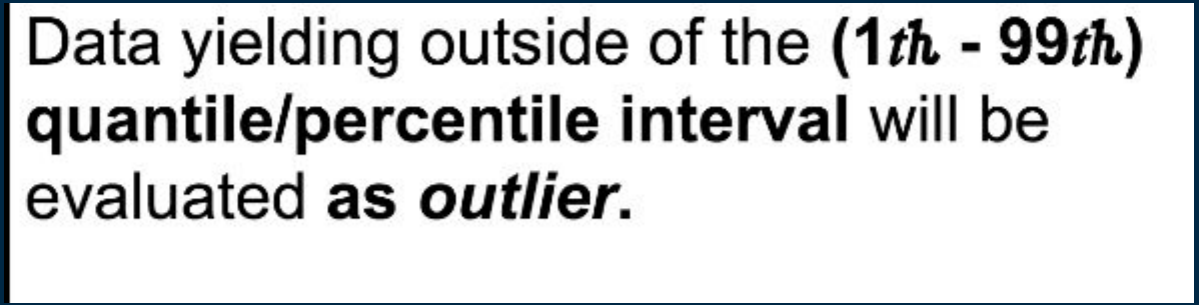
Z Score





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

Hard Edges Method



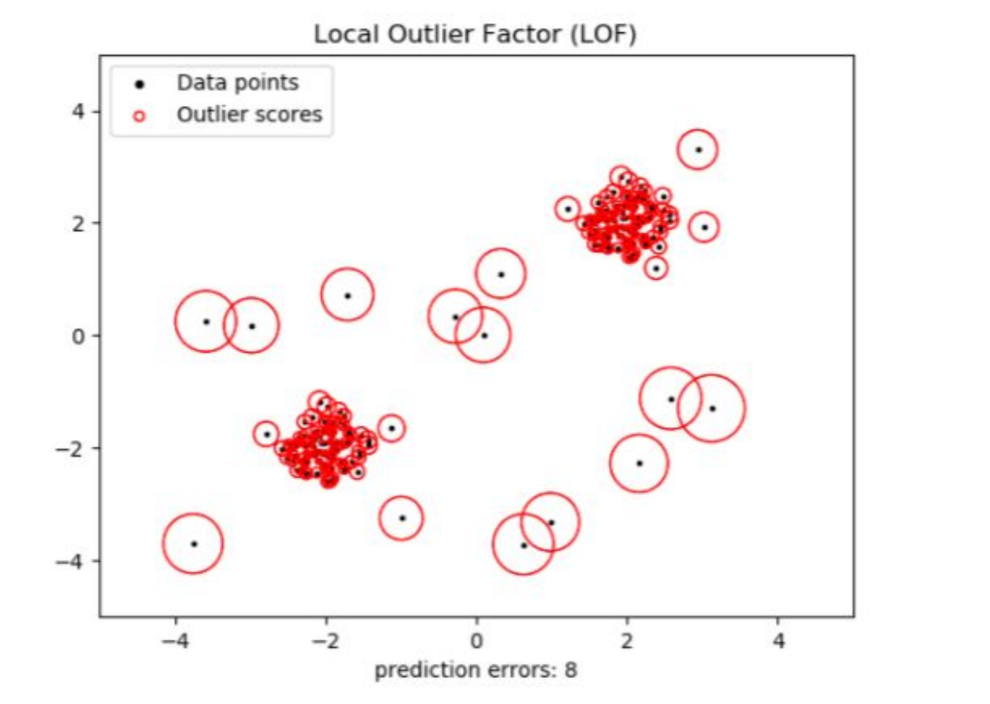
Data yielding outside of the **(1*th* - 99*th*) quantile/percentile interval** will be evaluated **as *outlier*.**

Why use Hard Edges Method?

- No calculate std, mean,median
- Basic and quick
- Appropriate for big dataset (for example : 300.000 rows)

Detection Methods : Multi-variate Methods

LOCAL OUTLIER FACTOR



https://towardsdatascience.com/local-outlier-factor-for-anomaly-detection-cc0c770d2ebe

THANK YOU