

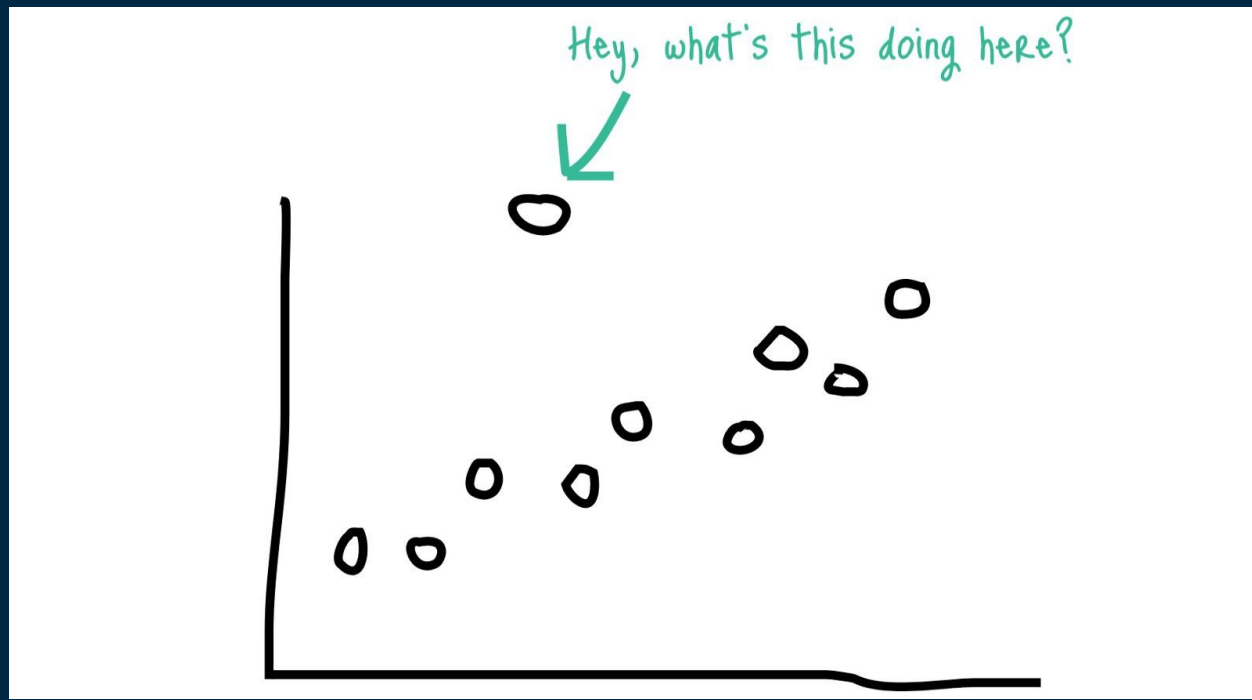
Outlier Detection Analysis

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

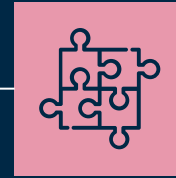
DATA ENTRY ERRORS



02

INSTRUMENT ERRORS

MEASUREMENT ERRORS



03

EXPERIMENTAL ERRORS

DATA EXTRACTION OR
EXECUTING ERRORS

Most Common Causes of Outliers



04

DATA PROCESSING
ERRORS

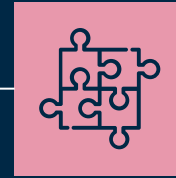
DATA MANIPULATION



05

SAMPLING ERRORS

VARIOUS SOURCES

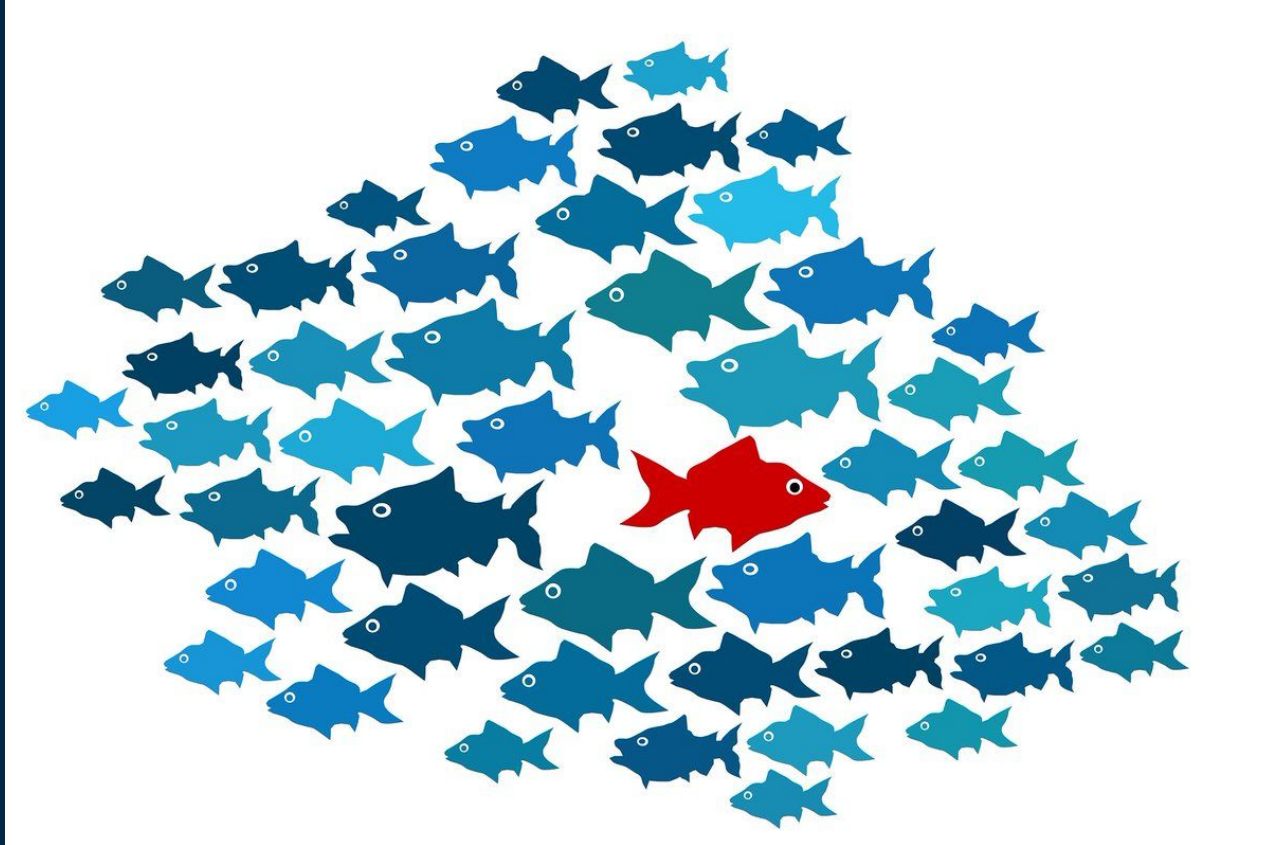


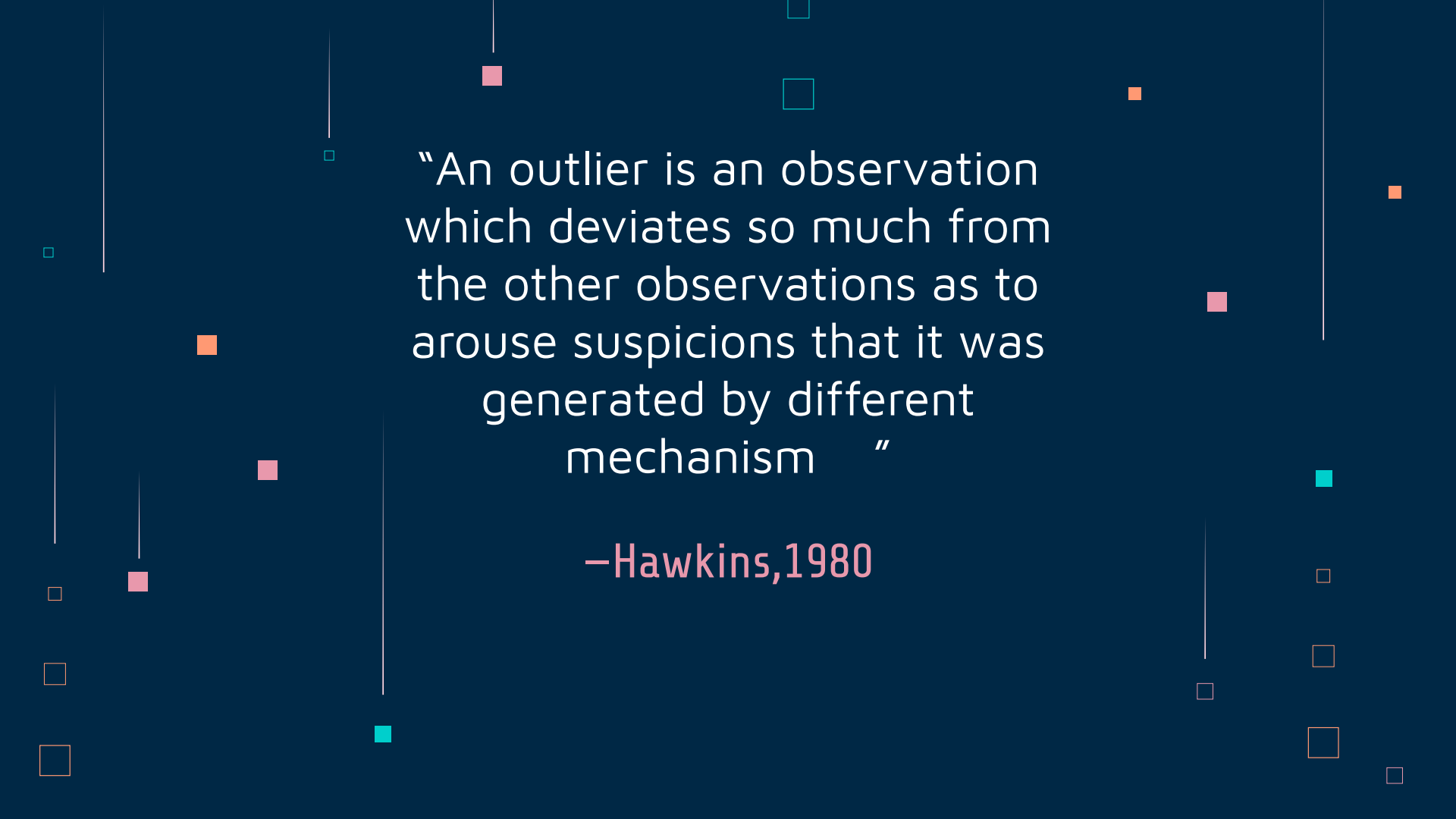
06

NATURAL

NOT AN ERROR

Outlier Example

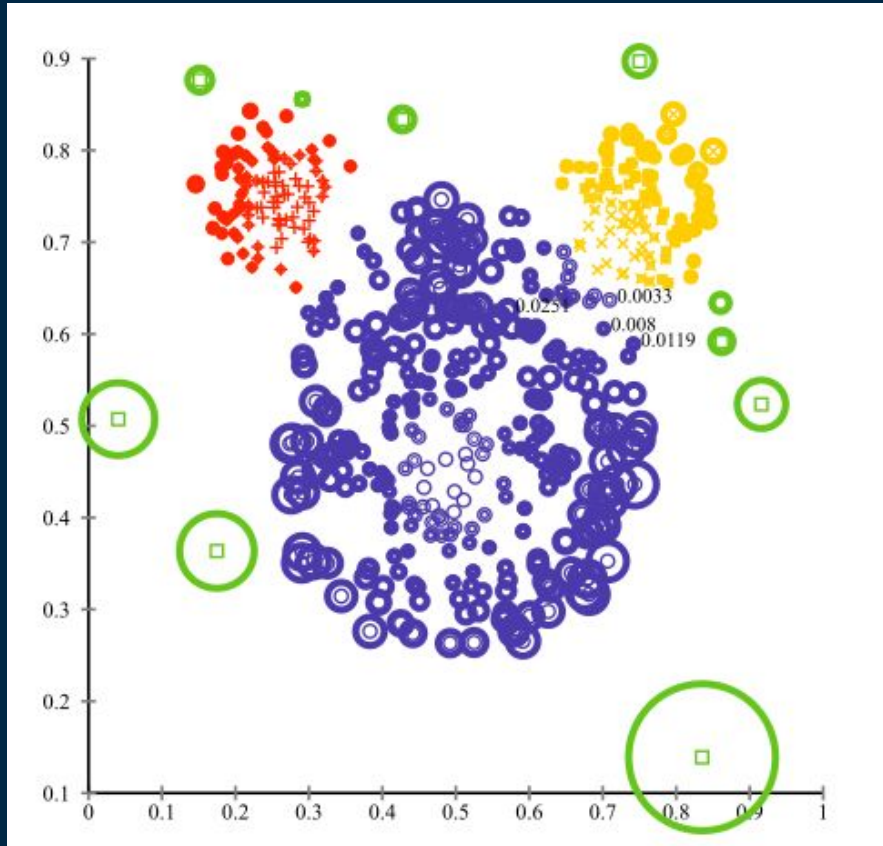


The background is a dark blue gradient. It features several vertical white lines of varying lengths. Scattered throughout are small squares in light blue, orange, and pink. Some squares are solid, while others are outlines. The text is centered in a white, sans-serif font.

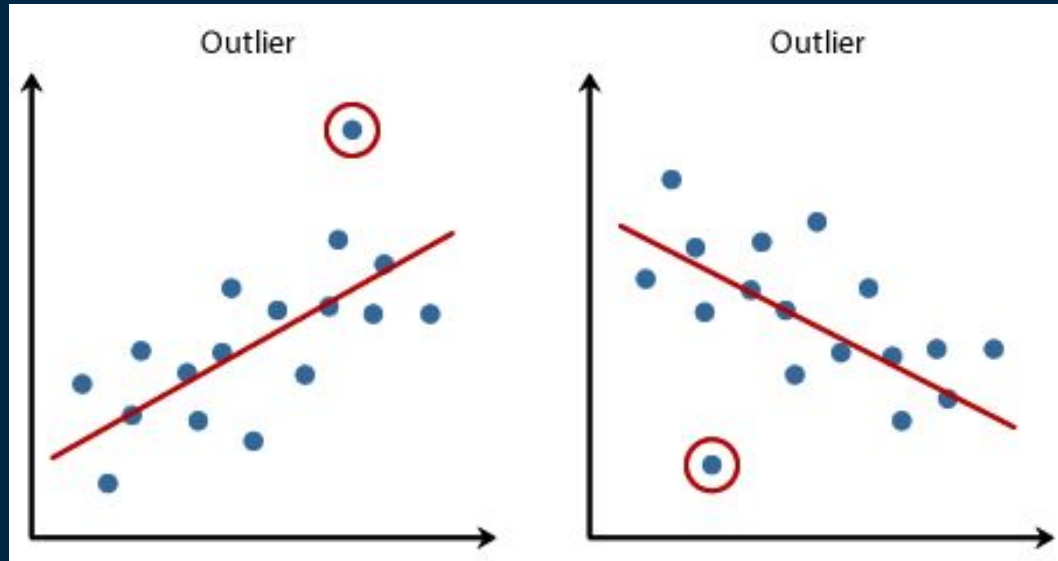
“An outlier is an observation
which deviates so much from
the other observations as to
arouse suspicions that it was
generated by different
mechanism ”

—Hawkins, 1980

Clustering



Regression



Invalid or Outlier Data

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

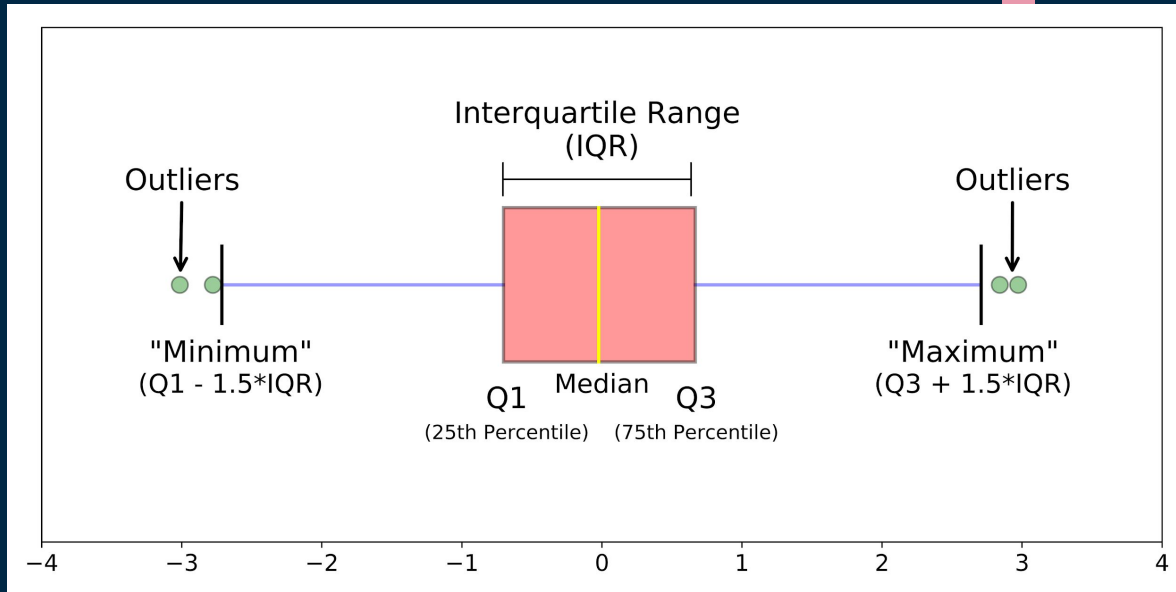
A cluster of small squares in the top right corner, including cyan, pink, and orange colors, some with solid fills and others as outlines.

Detection Methods :

Uni-variate Methods

A small cluster of squares in the bottom left corner, including orange and cyan colors, some with solid fills and others as outlines.

Boxplot



$(x > (Q3 + 1.5 \cdot IQR)) \vee (x < (Q1 - 1.5 \cdot IQR)) \rightarrow (x \text{ is an outlier})$

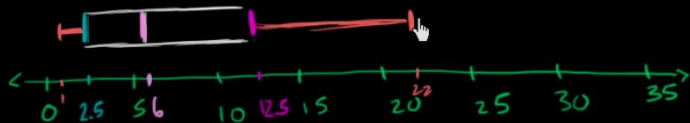
$(x > (Q3 + 3 \cdot IQR)) \vee (x < (Q1 - 3 \cdot IQR)) \rightarrow (x \text{ is an extreme-value})$

Boxplot

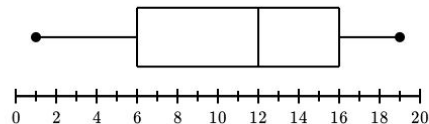
14, 6, 3, 2, 4, 15, 11, 8, 1, 7, 2, 1, 3, 4, 10, 22, 20

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?

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



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)

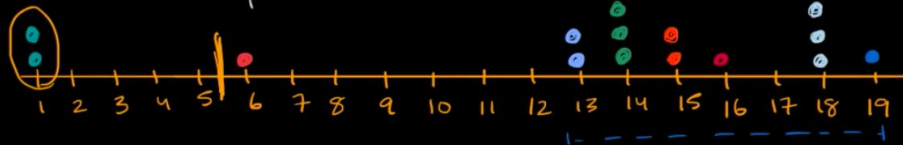
$$IQR = Q_3 - Q_1 = 18 - 13 = 5$$

$$\text{outliers} < Q_1 - 1.5 \cdot \cancel{IQR}^5 \quad \text{or} \quad > Q_3 + \cancel{1.5 \cdot IQR}^{7.5}$$
$$13 - \cancel{1.5 \cdot 5}^{7.5} \quad \text{or} \quad 18 + 7.5$$

$$\text{outliers} < 5.5 \quad \text{or} \quad \text{outliers} > 25.5$$

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

← Q_1 ← Median ← Q_3



Boxplot

```
1 import numpy as np
2
3 def outliers_iqr(ys):
4     quartile_1, quartile_3 = np.percentile(ys, [25, 75])
5     iqr = quartile_3 - quartile_1
6     lower_bound = quartile_1 - (iqr * 1.5)
7     upper_bound = quartile_3 + (iqr * 1.5)
8     return np.where((ys > upper_bound) | (ys < lower_bound))
```

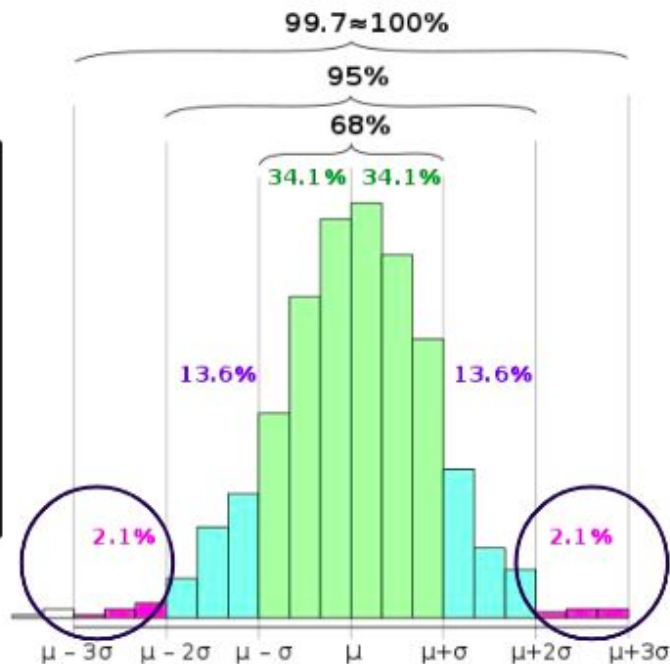
Standard Deviation

Using Standard Deviation

For value $X=x$:

$(x > (\mu + 2\sigma)) \vee (x < (\mu - 2\sigma)) \rightarrow$
(x is an **outlier**)

$(x > (\mu + 3\sigma)) \vee (x < (\mu - 3\sigma)) \rightarrow$
(x is an **extreme-value**)

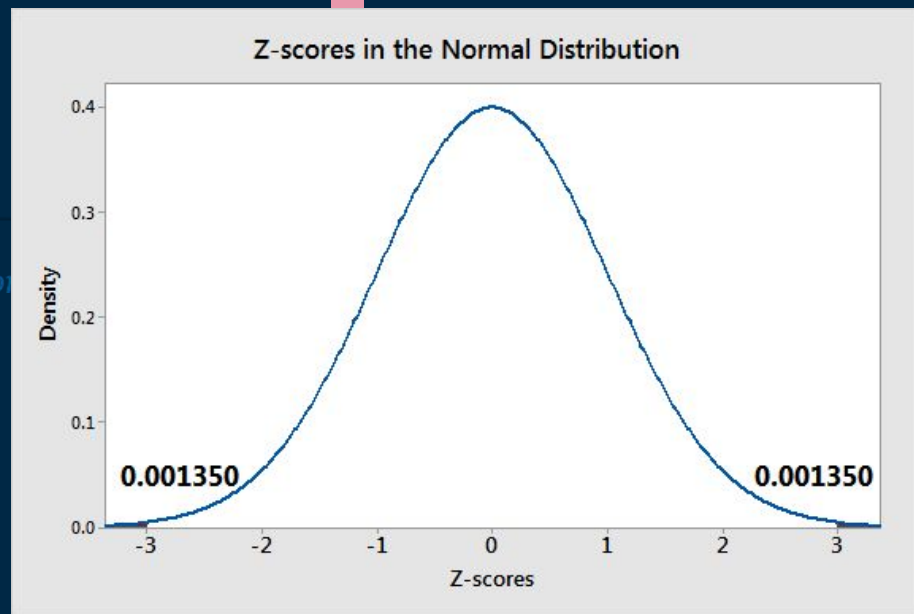


○ : outliers

Z Score



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



Hard Edges Method

Data yielding outside of the (**1th** - **99th**)
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)

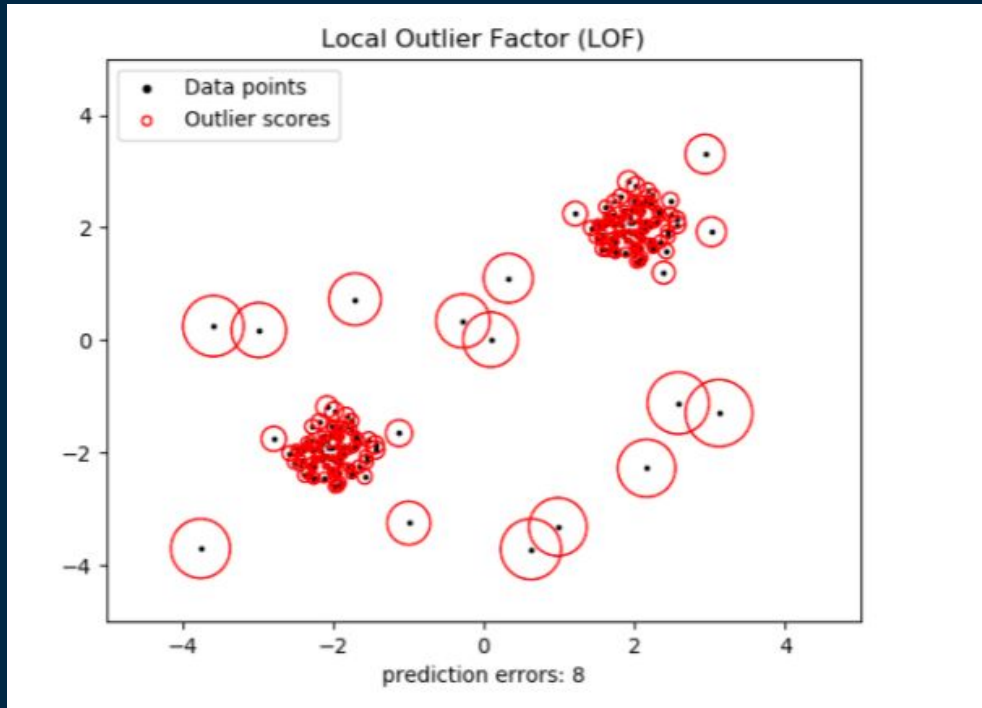
A cluster of small squares in the top right corner, including cyan, pink, and orange colors, some with solid fills and others as outlines.

Detection Methods :

Multi-variate Methods

A small cluster of squares in the bottom left corner, including an orange outline and a cyan solid square.

LOCAL OUTLIER FACTOR



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

The background is a dark blue field decorated with a pattern of small squares and thin vertical lines. The squares are in three colors: light blue, orange, and pink. Some squares are solid, while others are hollow. The vertical lines are thin and white, extending from the top or bottom of the frame. The text 'THANK YOU' is centered in the middle of the image.

THANK YOU