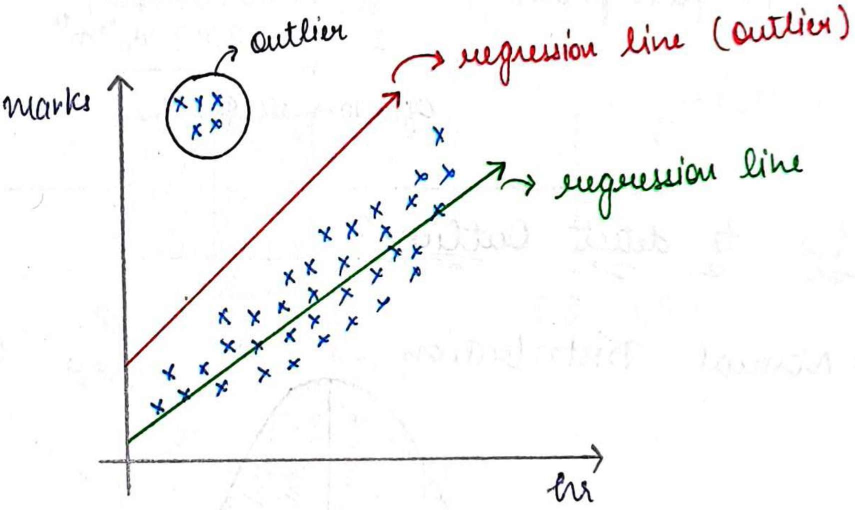
Outliers

What are Outliers?

An outlier is a data point that is significantly different from the rest of the data. Outliers can be much higher or lower than the other data points. They can be caused by measurement or execution errors, bad data collection, or simply show variables not considered when collecting the data.

Hr	marks
-	-
-	-



When is outlier dangerous?

Age

$\boxed{300}$ → Age 300 not possible

Effect of Outliers on ML algorithms

Linear Regression

Logistic Regression

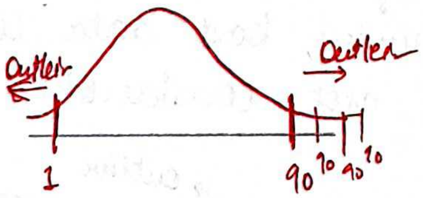
Weights algo.

Ada boost

Deep learning

How to treat Outliers?

- Trimming
  - Data thin
  - fast process
- Capping



after 90 → all 90

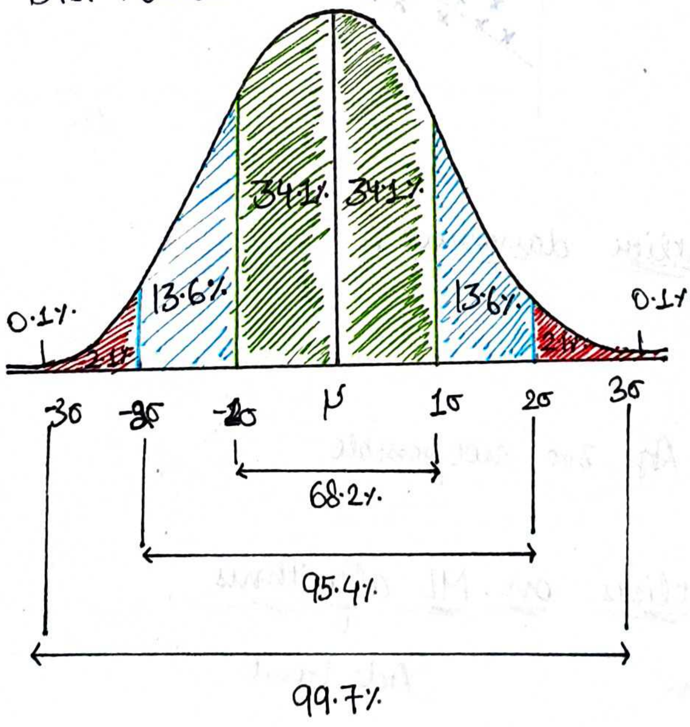
- missing value
- Discritization

90—10

Numerical.

How to detect Outliers?

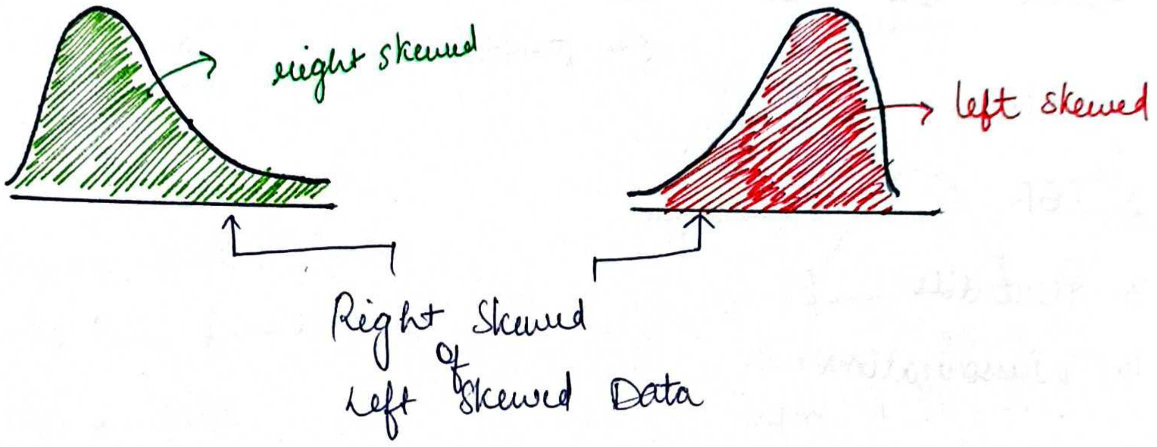
1. Normal Distribution

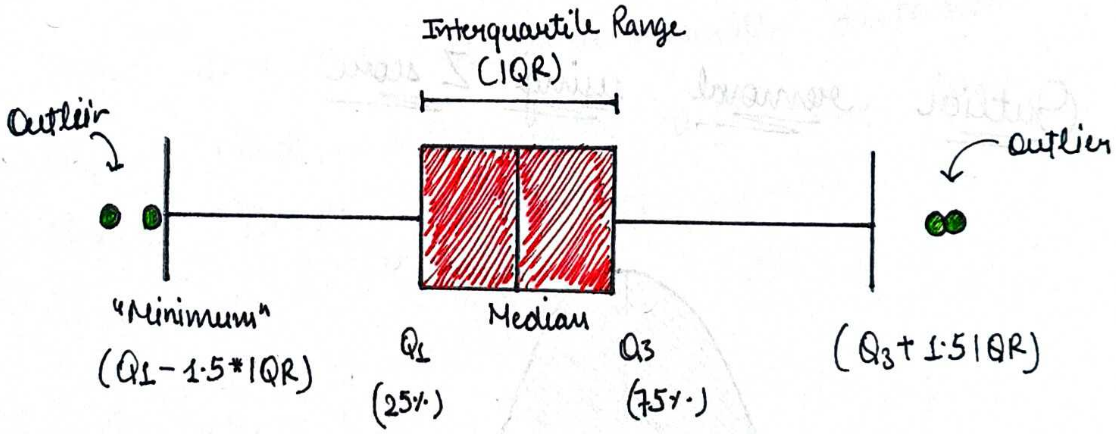


$(u+3\sigma)>$

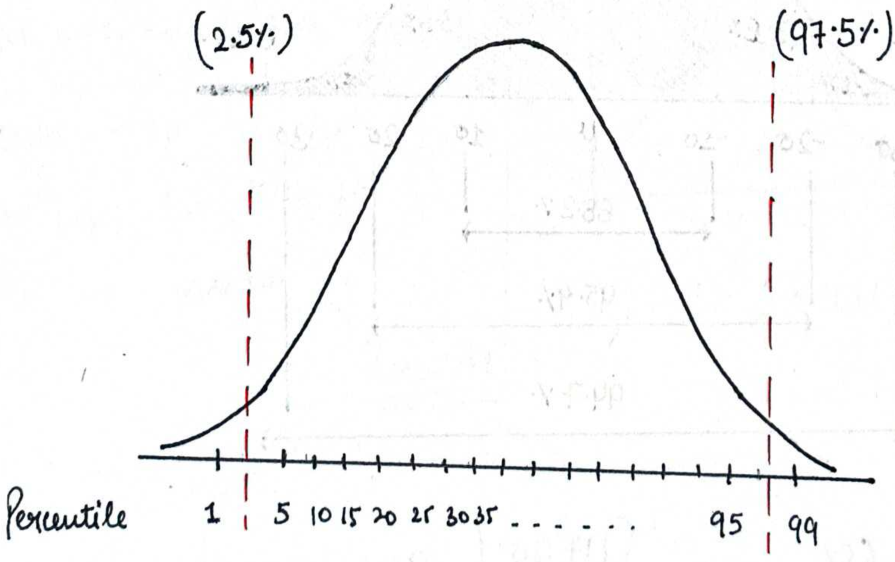
$(u-3\sigma)<$

2. Skewed distribution





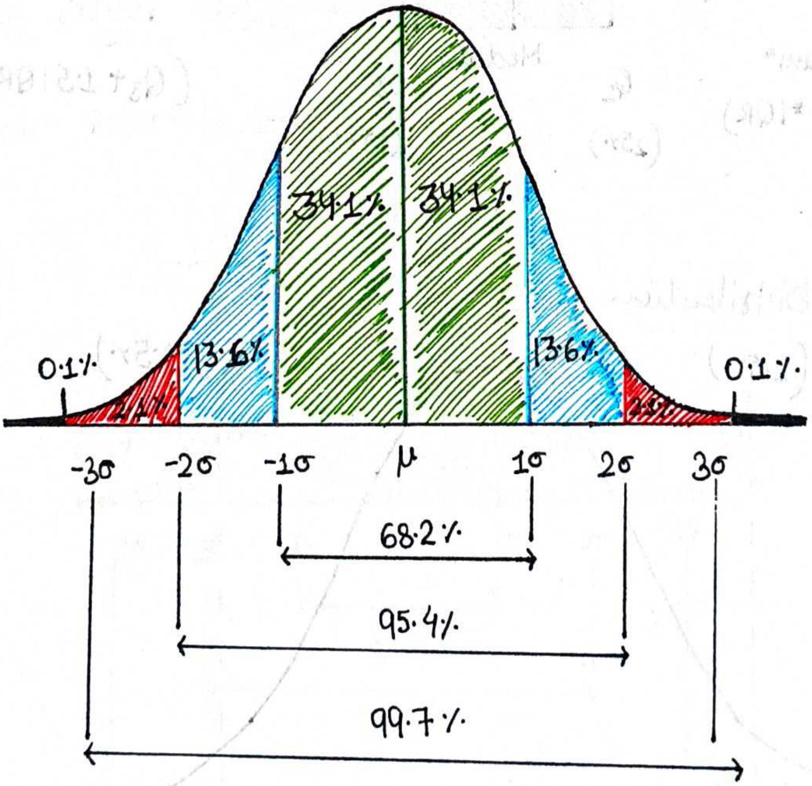
3. Other Distribution



Techniques for outlier Detection and Removal

1. Z-score
2. IQR
3. Percentile
4. Winsorization

Outlier removal using Z score



$\left\{\begin{matrix} \mu + \sigma \\ \mu - \sigma \end{matrix}\right\}$ 68%

$\left\{\begin{matrix} \mu + 2\sigma \\ \mu - 2\sigma \end{matrix}\right\}$ 95%

Z score

$$x' = \frac{x_i - \mu}{\sigma}$$

→ Age

27

32

$\boxed{-3 \text{ to } 3}$

Outlier Treatment

1. Trimming

↓

1000 dataset

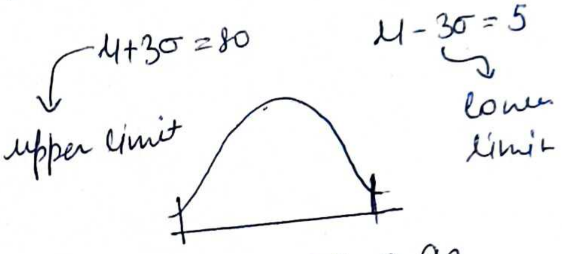
5 row → outlier

↳ Trim from origninal data

995 → data

2. capping →

3 outlier



Outlier → 85, 3, 90

caping → 80, 5, 80

Step1:- Sure Data should be Normally distributed

Step2:- Find upper limit and lower lower

Step3:- Decide → Trimming
→ capping

Percentile

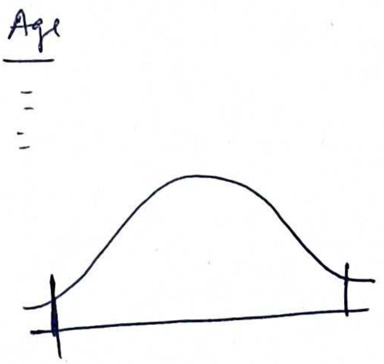
max - 95 → 100%

No age more than 95

min - 10 → 0%

No age less than 10

50% → median



1% → 99%

Percentile

→ remove

→ Cappiy

↳ Winsorization