Machine Learning Course | Arabic

Data Preprocessing

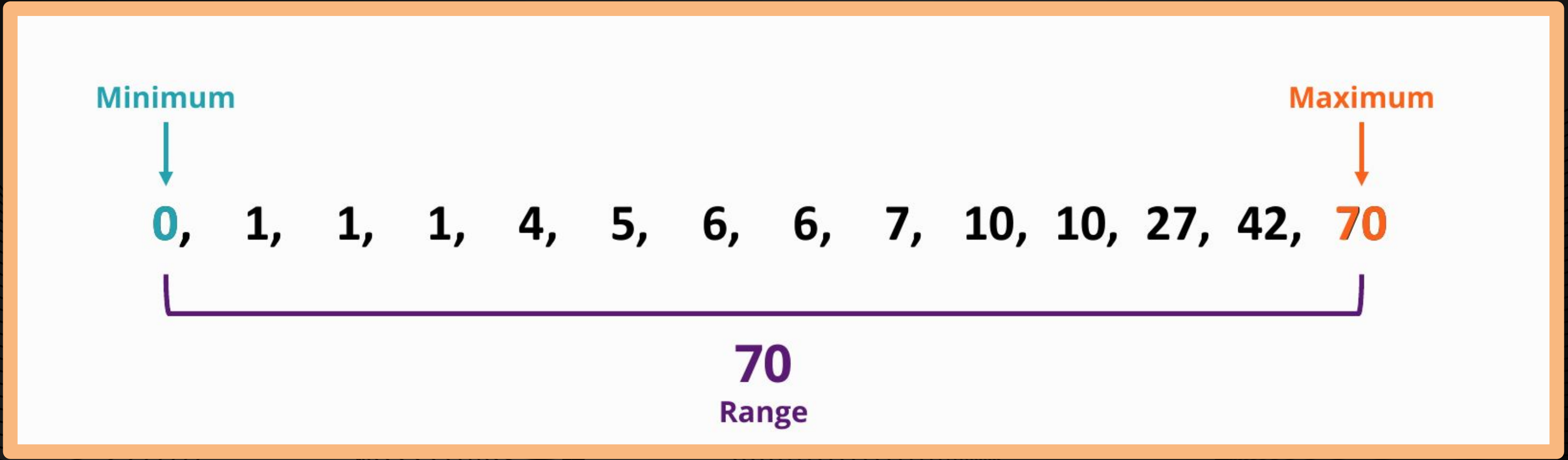
Level - 01

Link to Lecture on Youtube

02

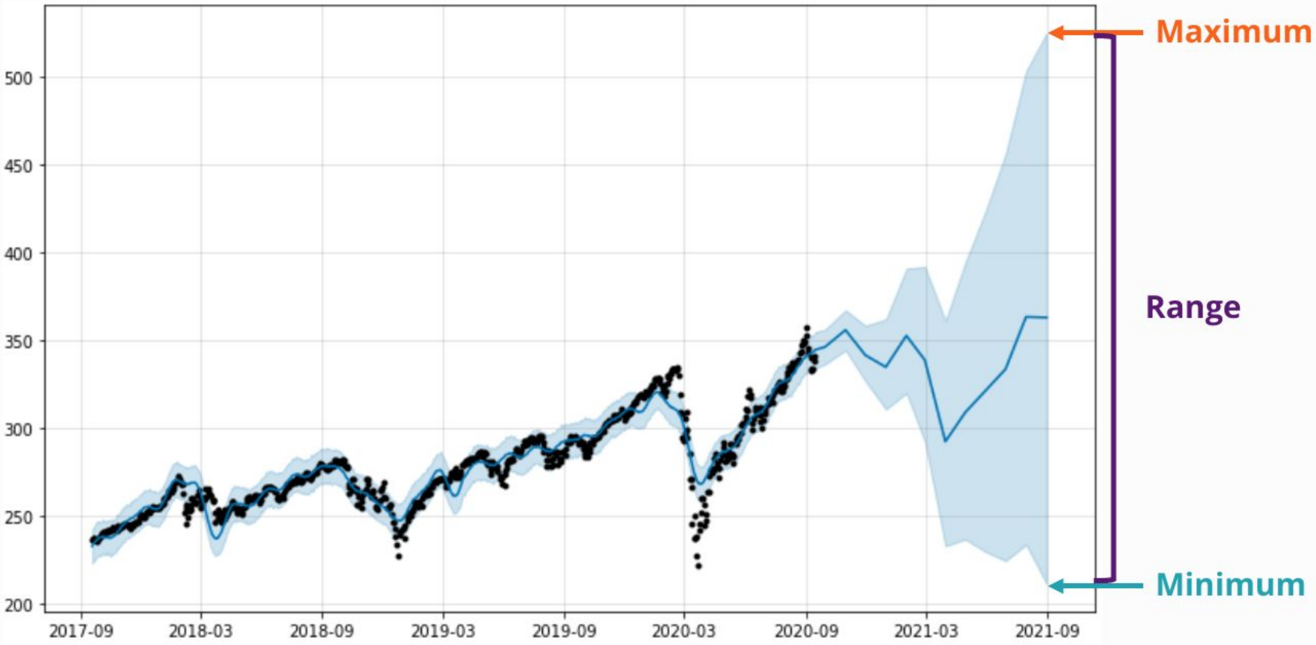
RANGE, Percentiles, IQR, Outliers

Introduction

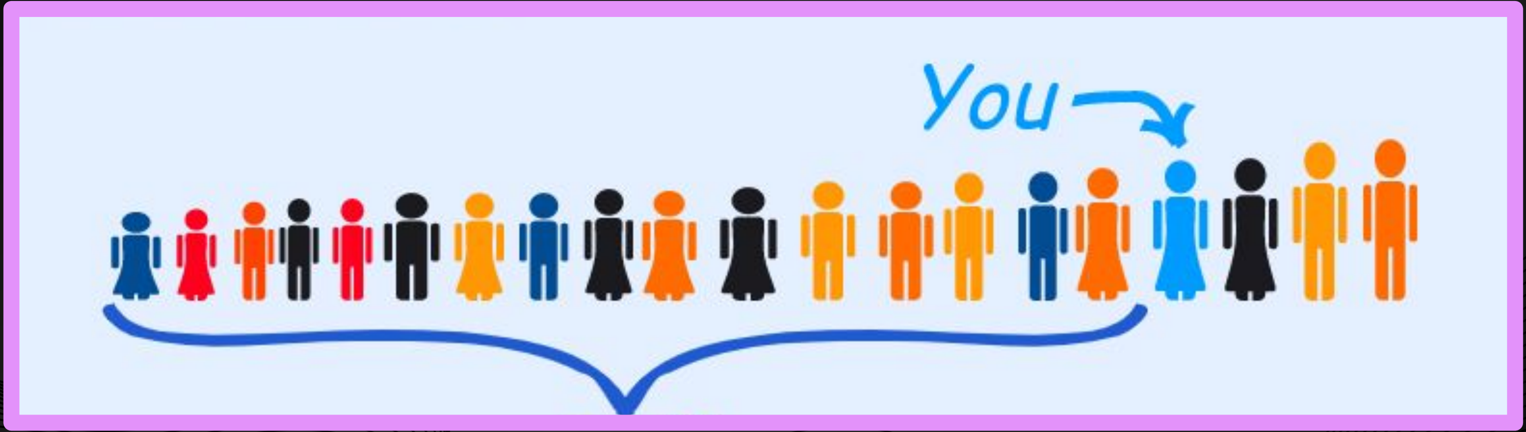


0, 1, 1, 1, 4, 5, 6, 6, 7, 10, 10, 27, 42, 70

Min, Max, Range



Percentiles



This means that **80%** of people are shorter than you.

That means you are at the **80th** percentile.

If your height is **1.85m** then "1.85m" is the 80th percentile height in that group.

Percentiles

Defined as the **percentage** of values that **fall below a particular value** in a set of data scores.

$$R = \frac{P}{100\ (N+1)}$$

R : Rank or sample index

P : Percentile value

Example:

Tracking the weight of children compared to other children of the same age.

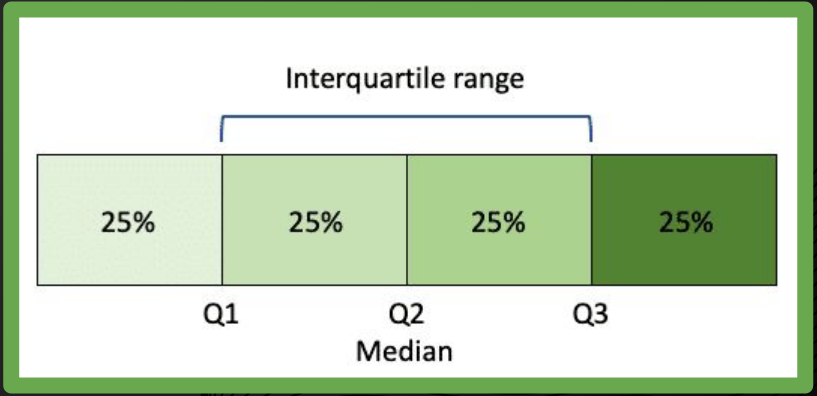
Quantiles

We have three main **Quantiles**:

Q1 : 25th percentile.

Q2 : 50th percentile (**Median**).

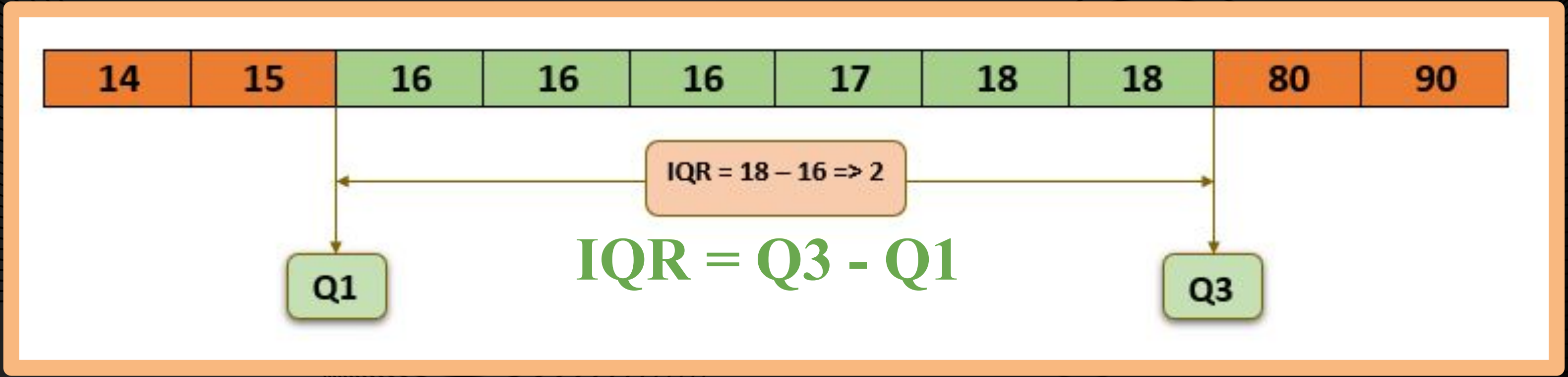
Q3 : 75th percentile.



Interquartile Range IQR

Used to measure the dispersion of values, but it is not affected by outliers. “**Used to handle outliers** ”

IQR contains “**50%**” of the data.



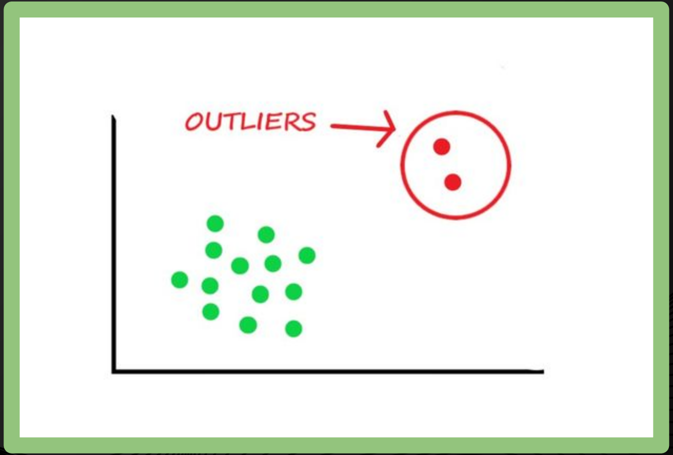
14	15	16	16	16	17	18	18	80	90

$$IQR = Q3 - Q1$$

Outliers

An outlier is a data point within a data set that **lies outside** of the range of most of the other data points.

Should we remove it?



Whenever you find an outlier “stop to think” and analyze it.
Justify your decision to drop / impute / keep it.

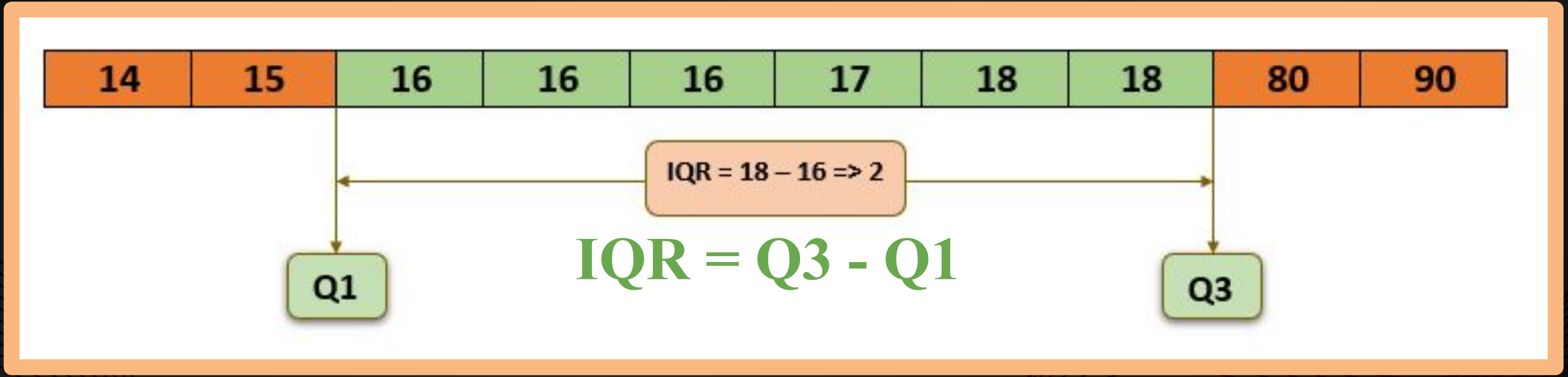
Outliers

You need to **investigate** any outliers carefully before removing them. Outliers often **tell you something** different than central values. (e.g. Age)

Example:

In the distribution of **human height**, **outliers** generally result from specific **genetic conditions**. Some researchers are concerned primarily with these types of conditions, others with the more usual factors that determine heights of 99.7% of adult humans. (Then we keep it here)

Use IQR to Handle Outliers



14	15	16	16	16	17	18	18	80	90

IQR = Q3 - Q1

Lower limit = Q1 - (1.5) IQR

Upper limit = Q3 + (1.5) IQR

Ex.

A survey was given to a random sample of 20 person. They were asked, “**how many textbooks do you own?**” Their responses, were:

[0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, 25]

Length data = 20 sample.

Solution

[0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, 25]

Remember !

$$R = \frac{P}{100\ (N+1)}$$

Median = 10

Q1 => 25% , P = 25

R = 0.25 / (20+1) = 5.25 = 5

Then, Q1 = 8

Q3 => 75% , P = 75

R = 0.75 / (20+1) = 15 .75 = 16

Then, Q3 = 12

Solution

[0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, 25]

Q1 = 8, Q3 =

IQR = Q3 12 **- Q1** = 12 - 8 = 4

$$R = \frac{P}{100}(N+1)$$

Upper limit = Q3 + 1.5 * IQR

Upper limit = 12 + 1.5 * 4 = 18

Lower limit = Q1 - 1.5 * IQR

Lower limit = 8 - 1.5 * 4 = 2

Solution

[0, 0, 2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15, 20, 25]

Data after filtration process:

[2, 5, 8, 8, 8, 9, 9, 10, 10, 10, 11, 12, 12, 12, 14, 15]

Length = 16

Thank You!

Do you have any questions?

Write them in the comments

hozaifazaki99@gmail.com