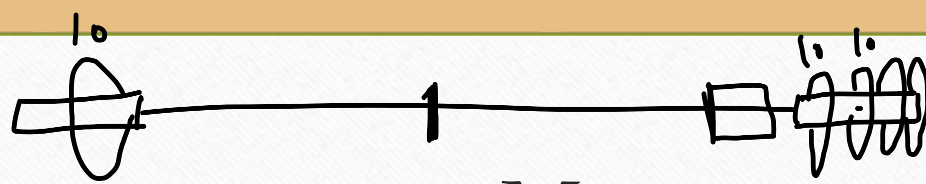


Mean → centre of gravity/mass

- The most common type of Average out there.
- Most influenced by the outliers.
- It changes when symmetricity of the distribution/graph is affected.

$$\bar{x} = \frac{\sum_{i=1}^{N} x}{N}$$

$$\bar{x} = \frac{\sum f}{\sum f}$$

$\sum f$ → frequency

$$\text{Mean} \Rightarrow \frac{\text{sum of obs.}}{\text{No. of obs.}}$$

$[10, 10, 10, 10] \Longrightarrow 10 \Rightarrow \frac{4 \times 10}{4} = 10$

$[10, 10, 10, 100] \Rightarrow 32.5$

⇒ Mean is shifting towards higher value.

⇒ Mean is in love with outliers, so always shifts towards outliers.

$$\begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \\ \end{array}$$
$$[\,1,\ 2,\ 3,\ 4,\ 5,\ 6\,]$$

x

$$\text{Mean} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6}$$

summation

$$\text{Mean} \Rightarrow \frac{\sum_{i=0}^{5} X}{N}$$

x	f
1	2
2	3
3	4

$\rightarrow [1, 1, 2, 2, 2, 3, 3, 3, 3]$

$$\text{Mean} \Rightarrow \frac{1+1+2+2+2+3+3+3+3}{9}$$

↓

Distribution table

$$\Rightarrow \frac{1\times 2 + 2\times 3 + 3\times 4}{9}$$

$$\Rightarrow \frac{\Sigma xf}{\Sigma f} \Rightarrow \text{Mean}$$

df[col].Count_values()

df[col].mean()

↓

pandas

[1 2 3 4 5]

→ 3

[3 2 4 1 5]

Sort ↳ [1 2 3 4 5] ⇒ Median ⇒ 3

Median

↳ Median (50%)

- It is actually the middle value, dividing the data into 2 equal halves.
- Steps:
  - Sort the data in ascending order

Median

n is odd,

Median = $\left(\frac{n+1}{2}\right)^{th}$ observatio

n is even,

$\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}$ o

~~$$\text{Median} = l + \left(\frac{\frac{n}{2} - F}{f}\right) \times h$$~~

- It is ~~not~~ less influenced by outliers.

↳ Special cases

1st 2nd 3th 4th 5th

[11 12 13 14 15]

Median ⇒ 13

---

getting to right position

$$\frac{5+1}{2} \Rightarrow 3^{th} \longrightarrow 13$$

Median

1st 2nd 3rd 3.5 4th 5th 6th

[11 12 13 | 14 15 16] len(list) = 6

$$\frac{3^{rd} + 4^{th}}{2}$$ positions 3.5

$$\left[\frac{\underbrace{\frac{6}{2}}_{3^{rd}} + \underbrace{\left(\frac{6}{2} + 1\right)}_{(3^{rd}) + 1}}{2} \right] \Rightarrow \frac{\overset{3^{rd}}{13} + \overset{4^{th}}{14}}{2} = 13.5$$

[7, 9, 2, 17, 18, 28]

Sort

10 11 12 13 14 15 16

[2, 7, 9 | 17, 18, 28]

Median = $\frac{9+17}{2}$ = 13

(1)

[1, 2, 3, 4, 5]

Median ⇒ 3

(2)

[1 2 3 4 500]

Median ⇒ 3

(3)

[1 2 3 4 500 600]

Median = 3.5

(4)

1	2	3	4	5	6	7	8th
1	2	3	4	500	600	700	800

$$\frac{4^{th} + 5^{th}}{2} = \frac{4 + 500}{2} = 252$$

[B B B A C D E E]

Mode = B

Mode

[B B B A C D E E E]

Modes ⇒ B or E

Bimodal

- It works for both numerical and categorical data.
- Observation with highest frequency is the mode.
- A dataset can be unimodal or multimodal.
- For ungrouped data, we can count or direct observe in the table.
- For grouped data, we use

$$\cancel{\text{Mode} = l_1 + \left(\frac{f_0 - f_{-1}}{2f_0 - f_{-1} - f_1}\right)}$$

df.value_counts()

sns.count_plot()

Imputation of Missing values

→ If outliers aren't present & feature is numerical, mean value is used for imputati-on.

→ If outliers are present, median is used

→ If col is categorical the mode is used!

$Mean_A = 0$

$Mean_B = 0$

A: $-5 \quad 0 \quad +5$

B: $-10 \quad 0 \quad +10$

$Median_A = 0$

$Median_B = 0$

Measures of Spread

- Measures
  - Range
  - Variance
  - Standard Deviation
  - Quartile Range(IQR)

[1 2 3 4 5] Range = 5 − 1 = 4

[1 3 5 10 200]

Range

Range = 200 − 1 = 199

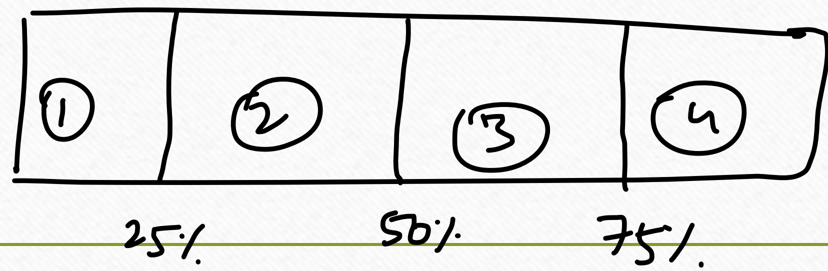
X

- The range is a way of measuring how spread out a set of values are.
- The range only describes the width of the data, not how it's dispersed between the bounds.
- Range is very sensitive to outliers.

$$Range = X_{max} -$$

25% → quantile × 100

percentile



quartile ≠ quantile

encapsulates

Q1

quartile 1

Median

Q2

quartile 2

Q3

quartile 3

$$\text{quantile} \Rightarrow \frac{\text{percentile}}{100}$$

$$\Rightarrow 0.25$$

Quartiles (Q1, Q2, Q3)

max−min

Inter Quartile Range

(between the quartiles)

IQR = Q3−Q1

- Quartiles: The numbers that separate data into 4 equal parts.

- IQR: difference between the 3rd quartile and 1st quartile.
- Less sensitive to outliers

$$\mathrm{IQR} = Q_3 -$$

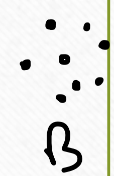
1^{st} half

[1 2 3 4 5 6 7 8 9 10 11]

$Q_1 = 3$ (Median of first half)

II^{nd} half

IQR = 4



Q_2 = Median $\Rightarrow$ 6

IQR = 6

$Q_3 = 9$ (Median of second half)

IQR = 4

$IQR = 9 - 3 = 6$

"Outlier Detection"

$UL = 9 + 1.5 \times 6$
$= 18$
$LL = 3 - 1.5 \times 6 = -6$

[1 2 3 4 5 6 7 8 9 10 11]

$Q1 = 3$

$Q2 = 6$

$Q3 = 9$

$IQR = 6$

$UL \Rightarrow Q3 + \overset{1.5?}{1.5}\,IQR \rightarrow$ beyond it everything is an outlier

$LL \Rightarrow Q1 - 1.5\,IQR \rightarrow$ below it everything is an outlier

Q = [17, 17, 18, 19, 20, 22, 23, 25, 33, 64] outlier

$Q1 = 18$

$Q2 = 21$

$Q3 = 25$

$IQR = 7$

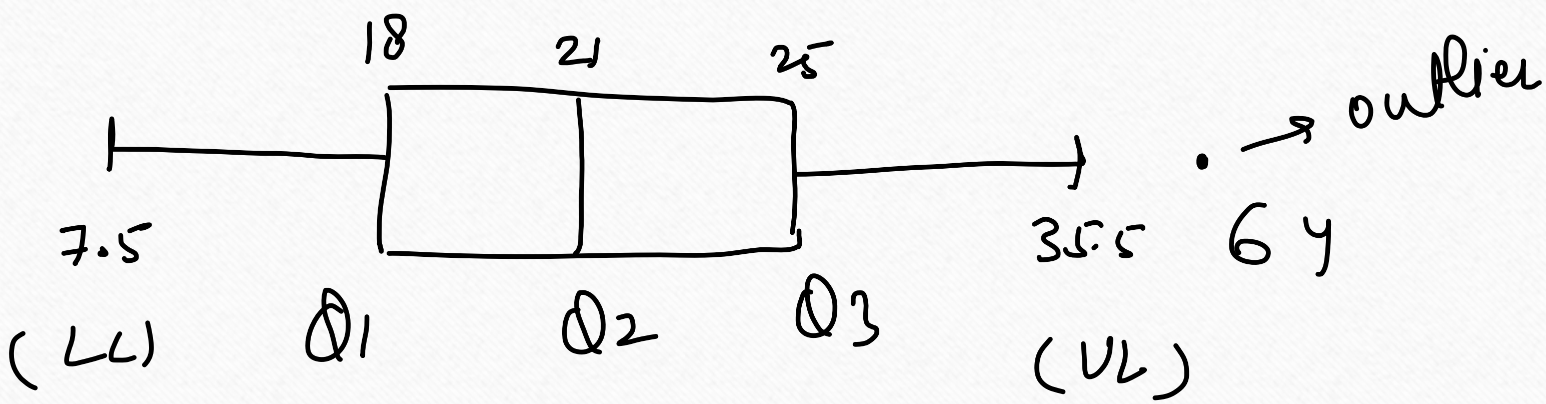
$UL = Q3 + 1.5\, IQR = 25 + 1.5 \times 7 = 35.5$

$LL = Q1 - 1.5\, IQR = 18 - 1.5 \times 7$

$= 7.5$

Box plot → Documentation of Seaborn/ Matplotlib

What is UL & LL ?



Checking for Outliers

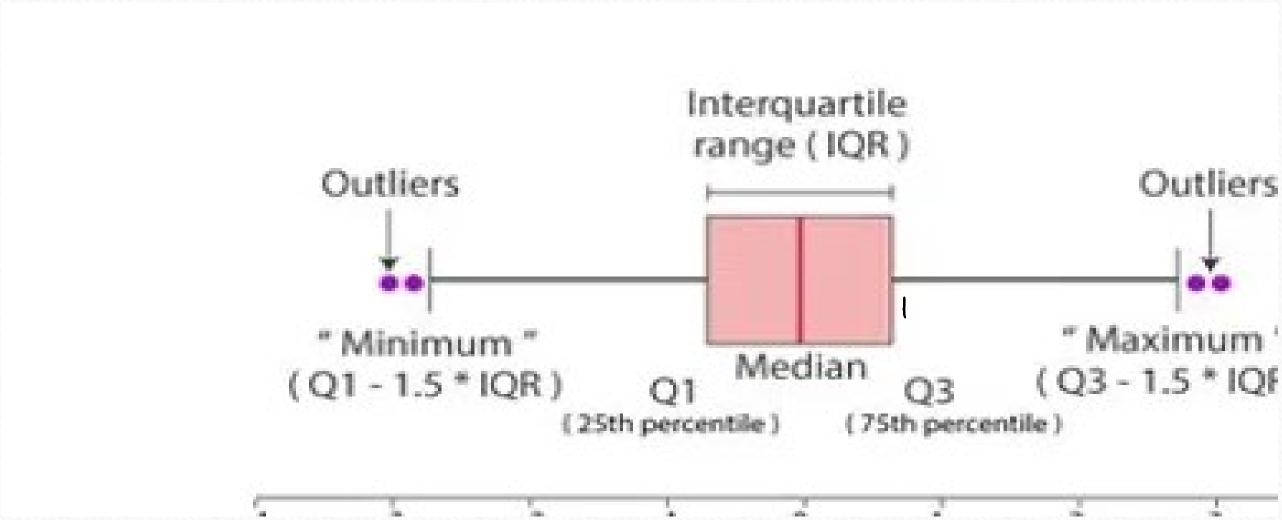
- With Range, Outlier detection isn't possible but with IQR it is possible!!

$$L = Q1 - (1.5 * IQR)$$

$$H = Q3 + (1.5 * IQR)$$

- Where L is the lower outlier
- H is the higher outlier
- Q1 and Q3 are the average values of those quart

Box Plot & 5 Number Summary



VK : [80, 0, 1, 40, 25]

MSD: [40, 41, 40, 39, 40]

[10, 10, 10, 10]

[10, 0, 5, -10, 0]

Variance

40

40

- The variance is a way of measuring spread, and it's the average of the distance of values from the mean squared.

population mean

sample mean

(1) $$\text{VAR} = \frac{1}{n}\sum_{i=0}^{n}(x - \mu)^2$$

length

(2) $$\text{VAR} = \frac{1}{n-1}\sum_{i=0}^{n}(x - \bar{x})^2$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$v = \frac{\sum_{i=1}^{n} x_i^2}{n} - \mu^2$$

- The problem with the variance is that it can be quite difficult to think about spread in terms of distances squared.

(2) → sample variance → estimating pop. variance from sample

$[1 \quad 2 \quad 3 \quad 4 \quad 5] \rightarrow$ Mean $\Rightarrow 3$

variance $\Rightarrow \frac{1}{n} \Sigma (x-\mu)^2$

avg

$$= \frac{1}{n}\left[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2\right]$$

$$= \frac{1}{5}\left[4+1+0+1+4\right] \Rightarrow \frac{10}{5} = 2$$

$[-5 \quad 0 \quad 5] \Rightarrow$ python $\rightarrow$ variance

$$\frac{(-5-0)^2 + (0-0)^2 + (5-0)^2}{3} = \frac{50}{3}$$

$$\downarrow$$

$$\frac{50}{3-1} = \frac{50}{2} = 25$$

python $\Rightarrow$ 25