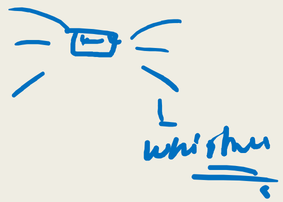
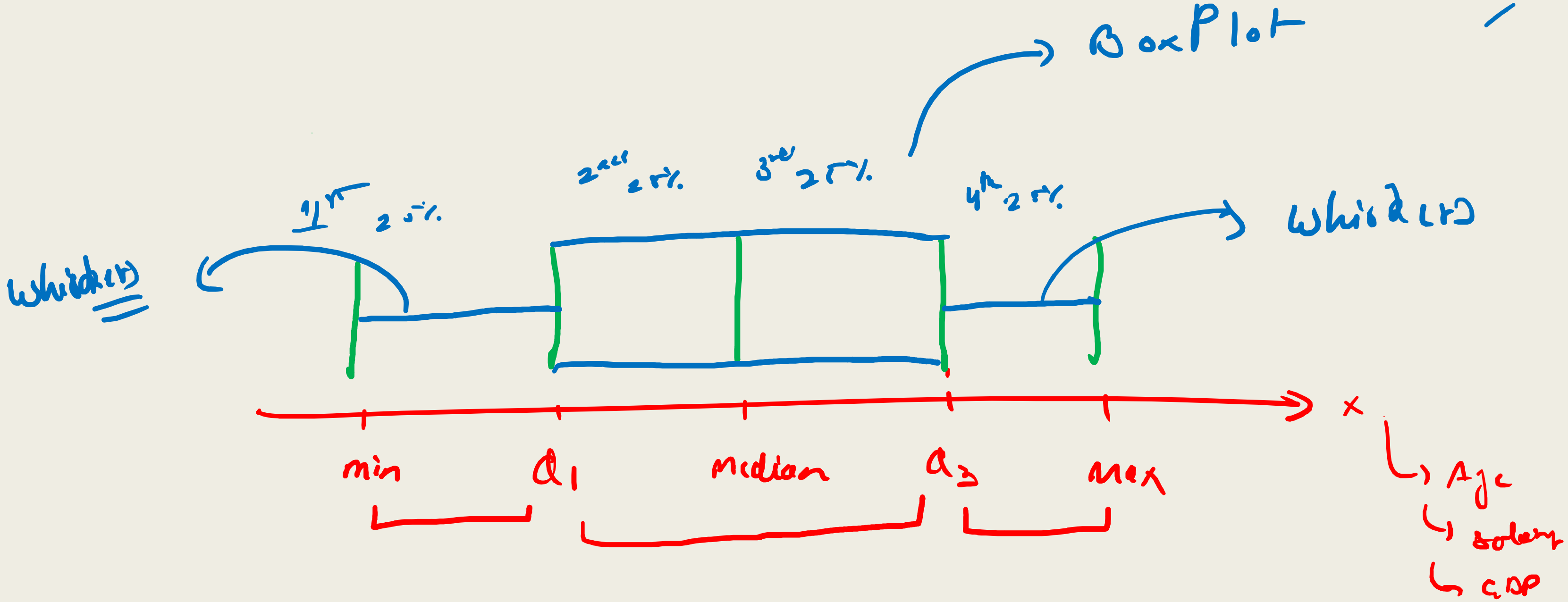
5-pt summary

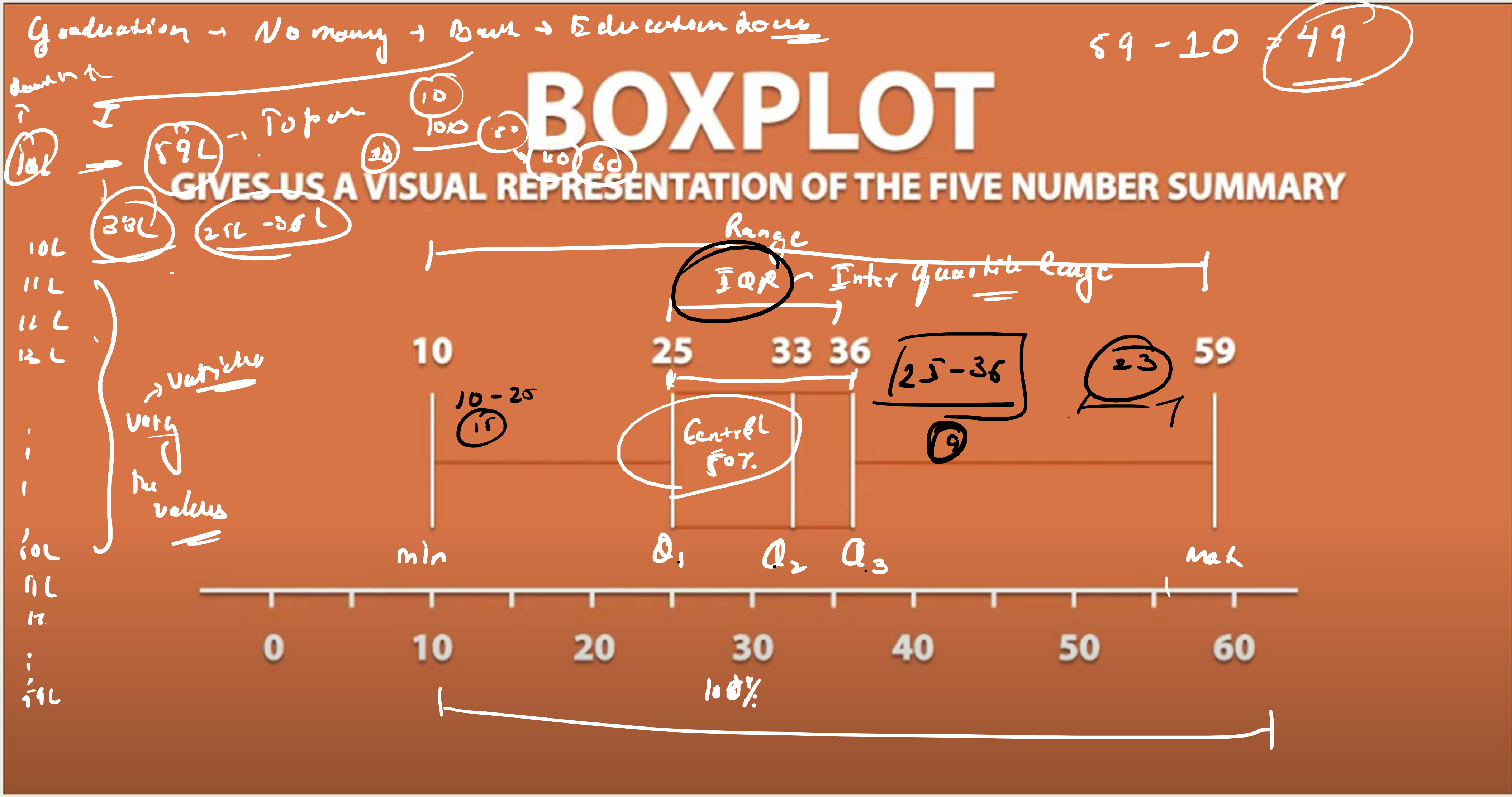
Visual Representation of 5-pt summary

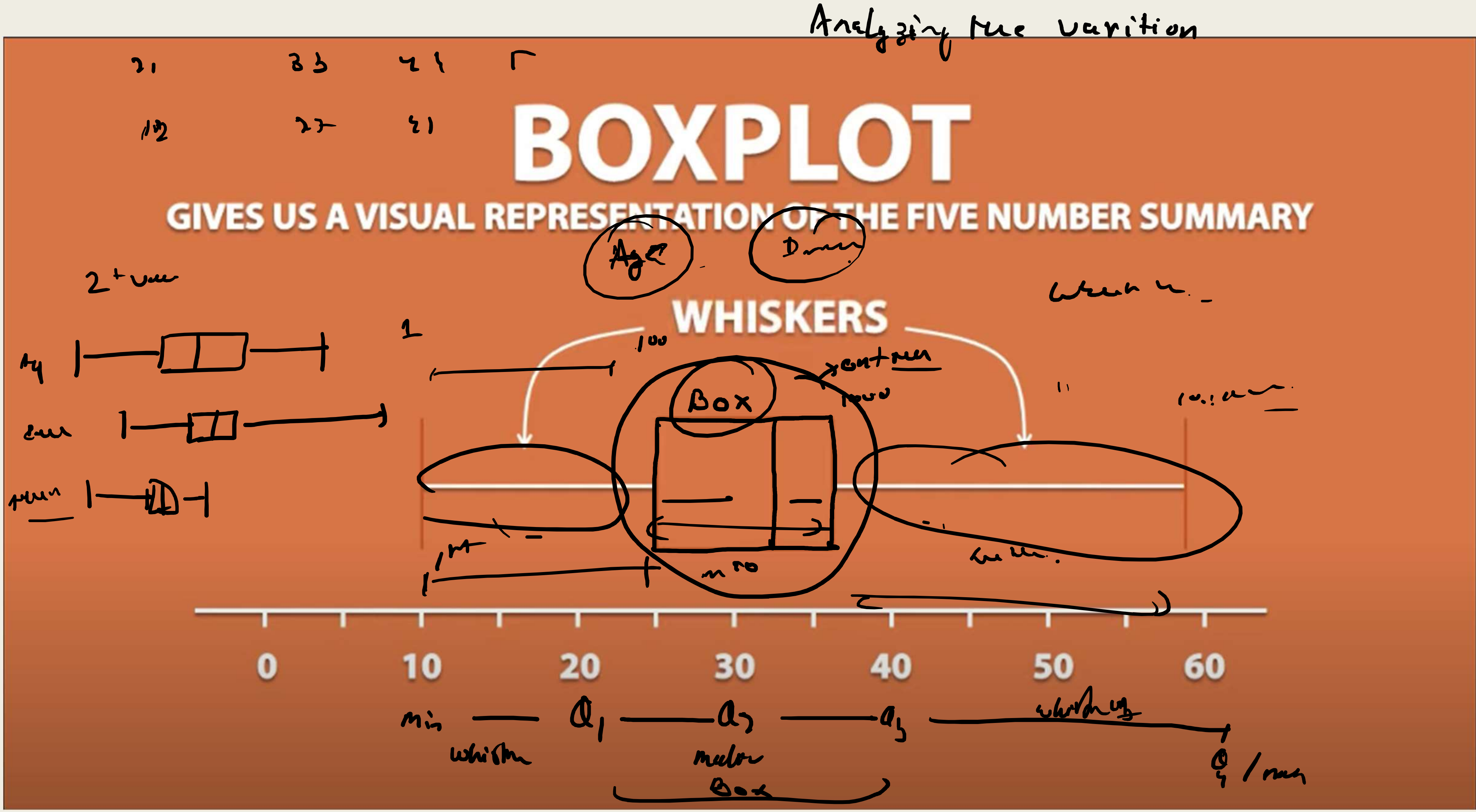
Cat, dog, dwn

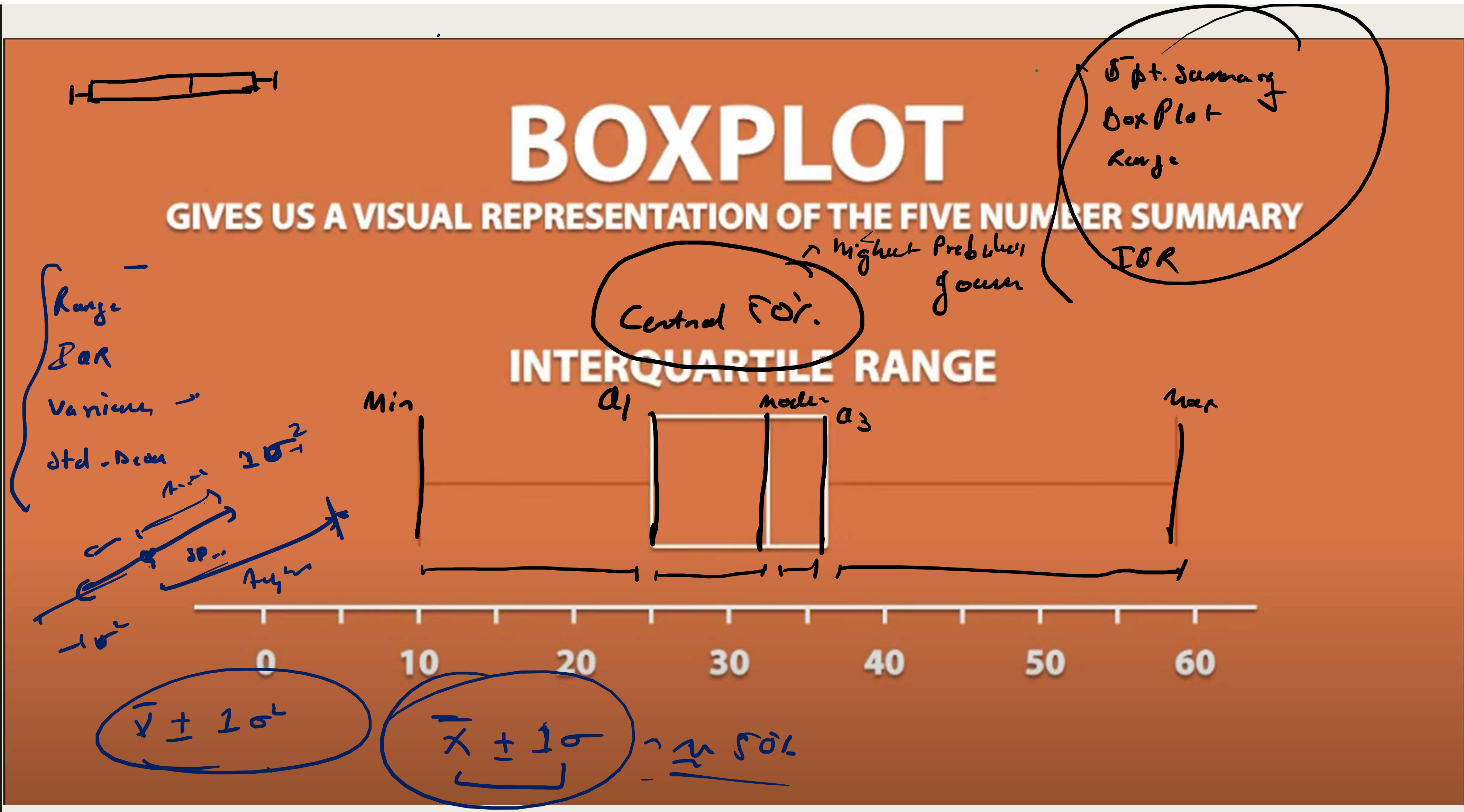




Central Tendency →




Analyzing the variation
BOXPLOT
GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY
Age
WHISKERS
Box
0
10
20
30
40
50
60
min — Q_1 — Q_2 — Q_3 — whiskers — Q_4 / max
whisker
median
Box


BOXPLOT
GIVES US A VISUAL REPRESENTATION OF THE FIVE NUMBER SUMMARY
INTERQUARTILE RANGE
Min
Q1
median
Q3
Max
0
10
20
30
40
50
60
5 pt. summary
Box Plot
Range
IQR
Range
IQR
Variance
std. Dev
$\bar{x} \pm 1\sigma$

→ Dispersion / Spread → Range → End-to-End spread (min - max)

IQR → Central 50% Range (Q_1 - Q_3)

Point to Point

→ Mean, Median → Spread

↳ Standard Deviation, Variation → measure of spread from the Center

1-4 = 3 = 1 N

1, 2, 2.7, 4, 5, 6, 7

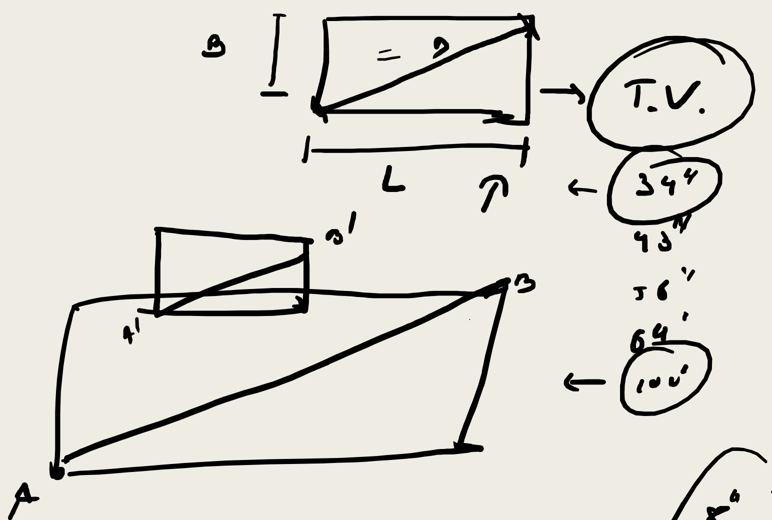
1-4 = -3

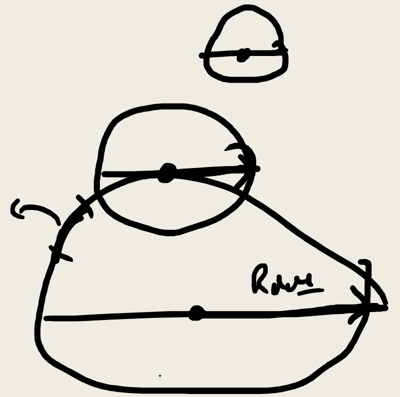
4 ± 3

7-4 = 3

4 ± 3

1-4
2-4
3-4





$2\pi R^2$

Radius

Diameter = 2 * Radius

Pizza: 8", 10", 12", 16"

1, 2, 2.7, 4, 5, 6, 7

$$\frac{(1-4)+(2-4)+(2.7-4)+(4-4)+(5-4)+(6-4)+(7-4)}{7}$$

1-1

1-4

-3

+3

Avg. distance from the mean

↳ Std Deviation

$$\frac{\Sigma \text{ dist}}{N}$$

$$\frac{\Sigma(x_i - \bar{x})}{N}$$

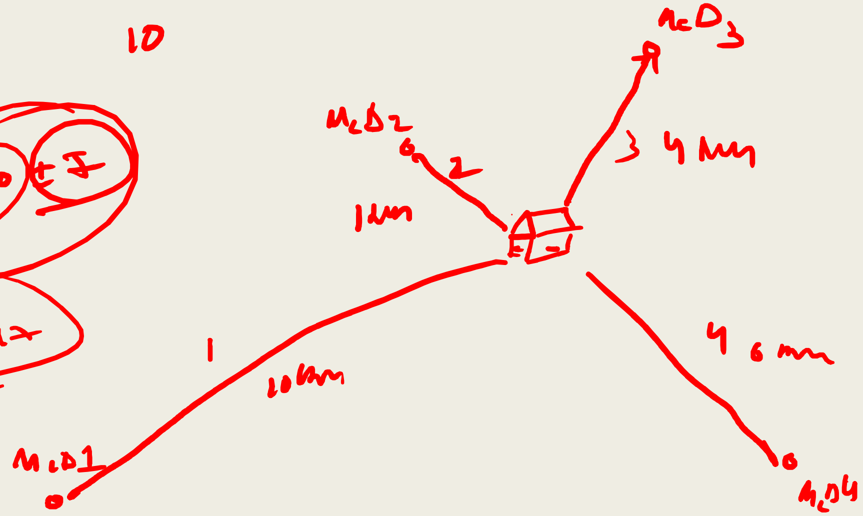
Burger

McD

10 + 7

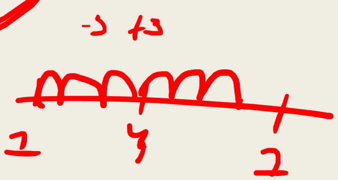
3 - 17

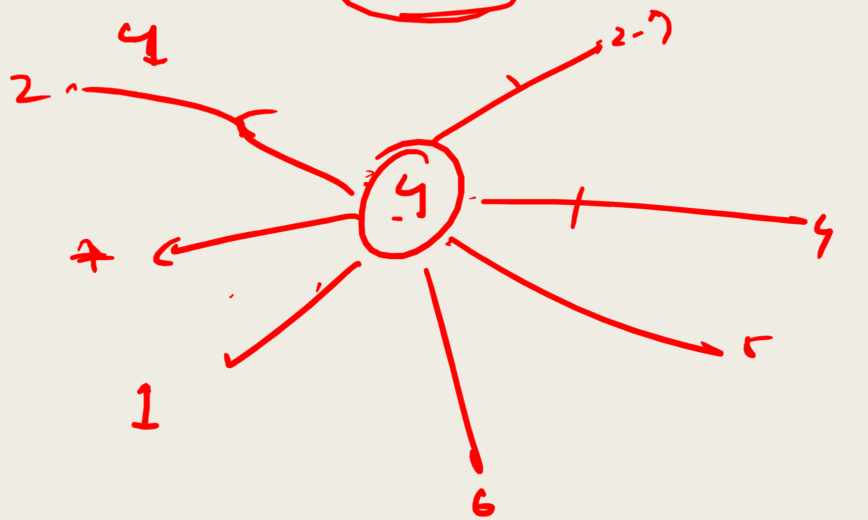
10



$$\frac{1, 4, 6, 10}{4}$$

5 km





$$\frac{\sum (x_i - \bar{x})}{N}$$ → 3 use

-ve sign

$N = 3$

→ How to remove -ve sign → sq. root (square) $(-4)^2 = \sqrt{16} = 4$

Modulus

Square

	Q_1	Q_2	Q_3
x_1	1	3	5
x_2	4	6	10
x_3	7	9	15
$\bar{X} =$	4	6	10
$x_1 - \bar{x} \Rightarrow$	-3	-3	-5
$x_2 - \bar{x} \Rightarrow$	0	0	0
$x_3 - \bar{x} \Rightarrow$	+3	+3	+5
	0	0	0

Always zero - $\sum (x_i - \bar{x})$

near zero

• Modulus $\rightarrow$ $\dfrac{\Sigma |x_i - \bar{x}|}{N}$

$\sqrt{10} \rightarrow \pm 4$

$|1-4| + |2-4| \ldots$

↳ Mean Absolute distance Avg. → X → wrong information

$d \times a$

• square → $\dfrac{\Sigma (x_i - \bar{x})^2}{N}$ ⇒ Variance (σ^2)

$\Sigma \sqrt{(x_i - \bar{x})^2}$ $((x_i - \bar{x})^2)^{1/2}$

Avg. squared distance from the mean

$\Sigma (x_i - \bar{x}) = 0$

• Sq. root [square] ⇒ $\sqrt{\dfrac{\Sigma (x_i - \bar{x})^2}{N}}$ → Std Deviation (σ) sigma

↓

Avg distance from the mean

$|1-4| = 3$

$|7-4| = 3$ → $\frac{6}{2}$ - 3 →

2. Measure of Spread / Dispersion

1. Standard deviation

Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

$$\text{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x - \mu)^2}$$

In Python :

Population STD = pstdev()

$$\text{S.D.} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

Sample STD = stdev()

Measure of Spread / Dispersion

2. Variance

Variance is a square of average distance between each quantity and mean. That is, it is square of standard deviation.

Population Variance (σ^2)

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

Sample Variance (S^2)

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

In Python : Population Var = pvariance()

Sample Variance = variance()

3. Range

Range is one of the simplest techniques of descriptive statistics. It is the difference between lowest and highest value.

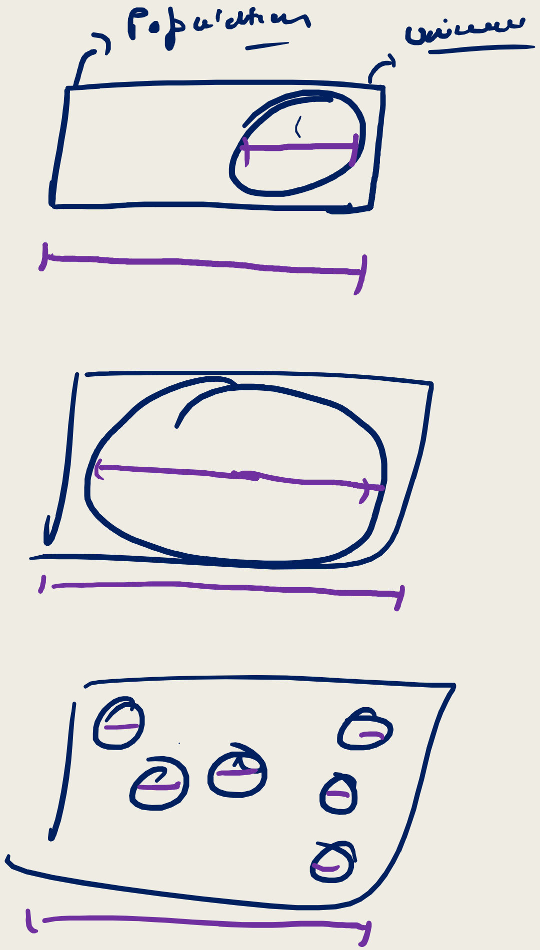
Range = Maximum - Minimum

4. IQR (Interquartile Range)

In statistics and probability, quartiles are values that divide your data into quarters provided data is sorted in an **ascending** order.

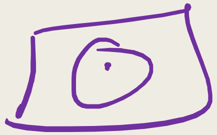
IQR = Q3 – Q1

Sample → Part of Population



Population size > Sample size

Sample size < Population size



Sample std
<
Population std.

→ pop X

$\bar{x}$

Sample

→ $\bar{x} = 100 = \mu$

→ $\sigma = 50$ → < pop std

→ for sure

close → Pop std

Inflate → $\boxed{\text{degree of freedom (d.f)}}$

→ Degrees of freedom

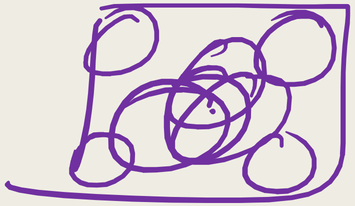
$$\sum(x_i - \bar{x}) = 0$$

$$\sum(x_i - \mu) = 0$$

becomes true

10
15
20

30
60
100



$\bar{X} = 15$

$\mu = 16$

Number of fixed values = d.f.

$\boxed{d.f. = N-1}$

10 → 9

100 → 99

503 → 502

$\bar{x}$		μ	
10 - 15 =	-5	10 - 16 =	-6
15 - 15 =	0	15 - 16 =	-1
20 - 15 =	+5	23 - 16 =	+7
	0		0 ≠ 0

$\bar{x}$ fixed

Taking the freedom of changing only 1 value

such test

Sample spread < Population spread ✓

d.f. = N−1

d.f. < N

Sample spread ≈ Population spread

↓

Inflate

If we decrease the denominator keeping the Numerator fixed

then the overall value will Increase

$$\uparrow \quad \rightarrow \sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}} \Rightarrow \sqrt{\frac{\sum (x_i - \bar{x})^2}{d.f. \downarrow}}$$

$$\hookrightarrow \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

→ $\frac{1}{1} = 1.00$

→ $\frac{1}{2} = 0.50$

→ $\frac{1}{3} = 0.34$

→ $\frac{1}{4} = 0.25$

→ $\frac{1}{5} = 0.20$

Measure of Spread / Dispersion

Steps to find out the IQR

1. Order the data from least to greatest
2. Find the median
3. The left side of median is lower half and right side of the data is upper half.
4. Calculate the median of both the lower and upper half of the data (Called Q1 and Q3 respectively)
5. The IQR is the difference between the upper and lower medians

(Note: When we write down Minimum, Maximum, Q1, Q2 (Median) and Q3, this is called 5-point summary or 5 number summary)

$IQR = Q_3 - Q_1$