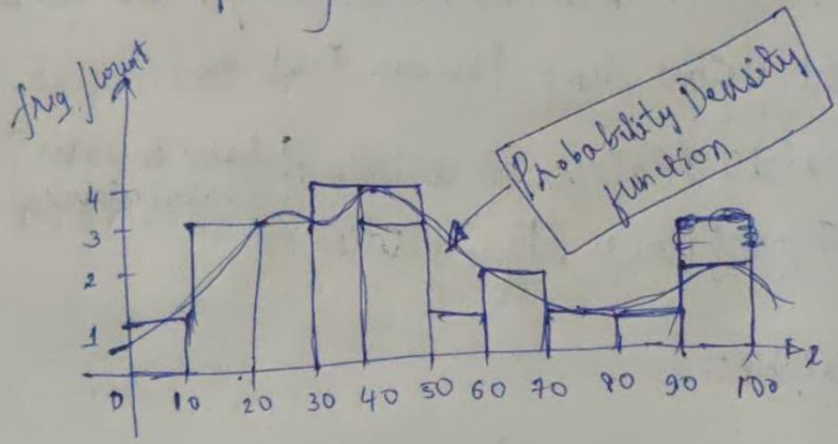
Pan card - Qualitative statistics

Histograms:-

1) Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

1) Sorting the no
2) Bins - No of groups
3) Bin size - Size of bins

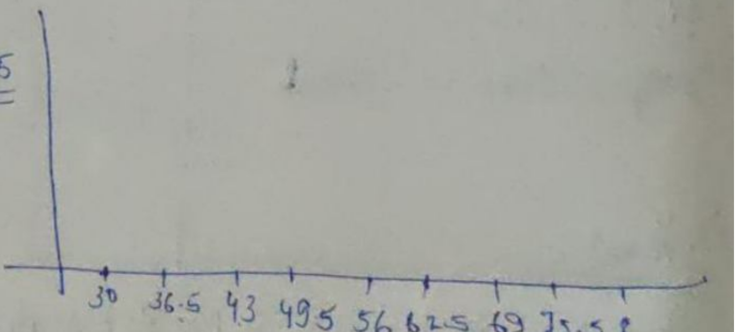
No of bins = $\frac{Max}{10} = \frac{100}{10} = 10$



2) Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

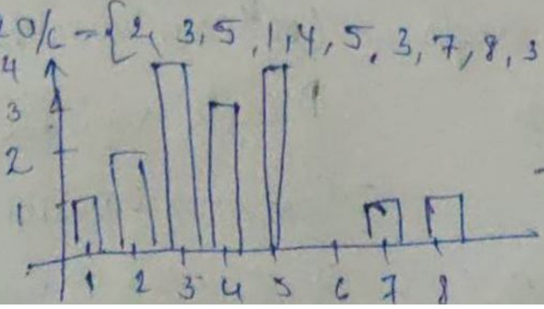
Bins = 10

No of bins = $\frac{95-30}{10} = \frac{65}{10} = 6.5$



Discrete stats

No of bank a/c - {2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5}



→ Probality Mass functions.

Measure of Central Tendency → A measure of central tendency is a single value that attempts to describe a se of data identifying the central position.

1) Mean
2) Median
3) Mode

Mean:-

$\boxed{\text{Population Mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N}}$ $\boxed{\text{sample mean } (\bar{x}) = \sum_{i=1}^{n} x_i/n}$

$\boxed{N \geq n}$ but, $\boxed{\mu \geq \bar{x}, \ \bar{x} \geq \mu}$

Population Ages = {24, 23, 2, 1, 28, 27}

$N = 6$

$\mu = \frac{24+23+2+1+28+27}{6} = 17.5$;

$\boxed{\mu = 17.5}$

Sample age = {24, 2, 1, 27}

$n = 4$

$\bar{x} = \frac{24+2+1+27}{4} = \frac{\overset{13.5}{54}}{4}$

$\boxed{\bar{x} = 13.5}$

Practical Application [Feature Engineering]

Age	Salary	Family Size.
–	–	–
NAN	–	–
–	NAN	–
–	–	–
–	–	–
–	–	NAN

Instead dropping NAN, replace NAN value with mean value.

If NAN dropped, there may be chances of losses of data

Age	Salary
24	45
28	50
29	NAN
NAN	60
31	75
36	80
NAN	NAN.

$$Age(x) = \frac{24+28+29+31+36}{7}$$

$$=$$

$$Salary(x) = \frac{45+50+60+75+80}{7}$$

$$=$$

Replace Age(NAN) with
Salary (NAN) with

Median:-

Steps to find Median

1) Sort the no

2) Find the central no {If no of elements are even, then avg of centre)
(If no of elements are odd, then central elements)

{1, 2, 3, 4, 5, 6, 7, 8, 100, 120}

Median $= 5+6/_2 = 5.5$

With outliear - Median.

Without outliner Mean.

Mode :- Most frequent occuring elements.

1) {1, 2, 3, 3, 4, 5, 6} → Mode = 3

2) {1, 2, 3, 2, 2, 3, 3, 4, 5, 6} → Mode = {2, 3}

Mode used with Categorical variable.

Measure of dispersion:-

1) Variance (σ^2) → Spread of Data
2) Standard deviation (σ) →

Variance

Population variance (σ^2)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Sample variance (s^2)

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

Standard Deviation :- ($\sqrt{\sigma^2}$)

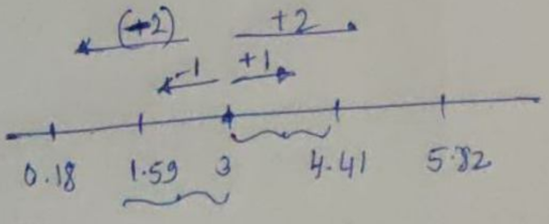
$= \{1, 2, 3, 4, 5\}$

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

$$= \frac{4+1+0+1+4}{5} = 10/5 = 2$$

$$\boxed{\sigma^2 = 2}$$

So, Std. deviation $= \sqrt{\sigma^2} = \sqrt{2} = 1.41$ //


(-2)
+2
-1
+1
0.18
1.59
3
4.41
5.82