DAY - 2

*Histogram :

Ages = {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 51, 70, 80,}

1) first we have to sort the numbers
2) create Bins (No. of groups)
3) create Binsize (Size of groups)

Bins can be created by our own.

Eg: [10, 20, 25, 30, 35, 40] min - 10, max - 40

Binsize will be maximum by binsize

i.e, Binsize = $\frac{40}{10}$ - 4

Here, 40 is maximum number in the list and 10 is Binsize.

Suppose, If I want Binsize of 20

I can write it as $\frac{40}{20}$ = 2

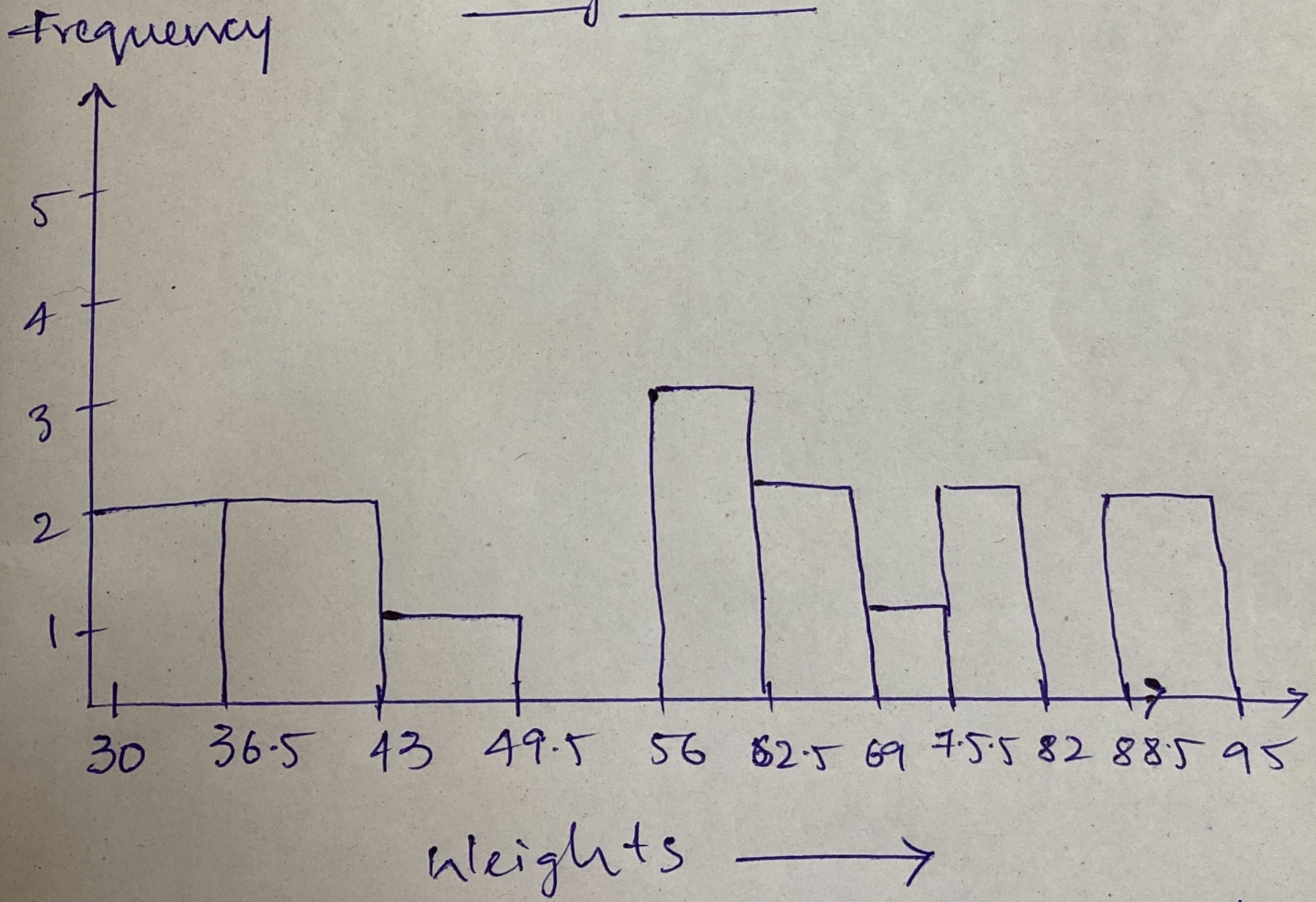
Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95}

Bins = 10

To calculate the Binsize = $\frac{\text{max} - \text{min}}{10} = \frac{65}{10}$

$= \frac{95 - 30}{10} = \frac{65}{10} = 6.5$

Assignment



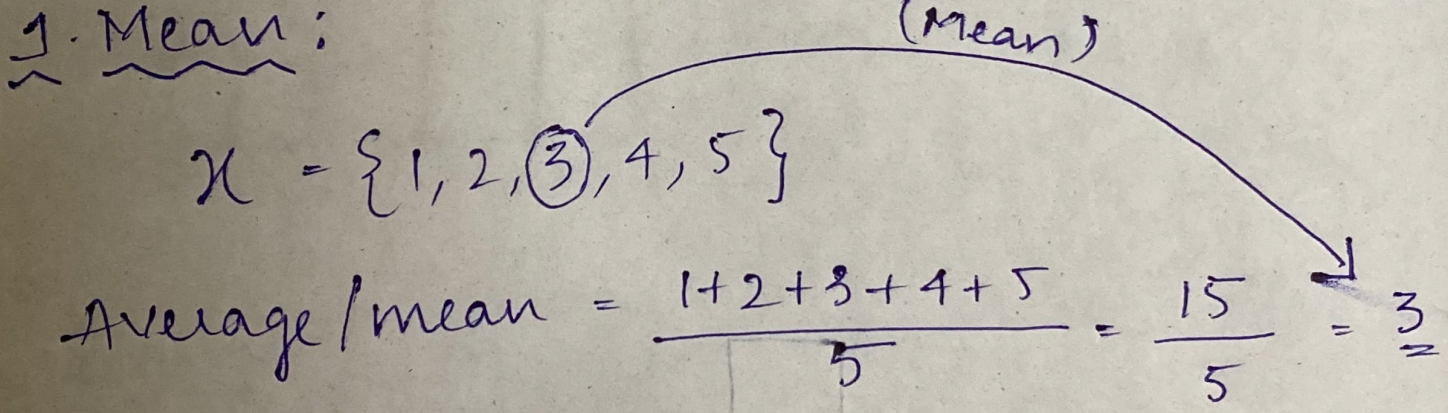
- for smoothening continuous histogram will use pdf (Probability density function)
- for smoothening Discrete continuous histogram will use Pmf (Probability mass function)

* Measure of Central Tendency:

1. Mean
2. Median
3. Mode

A measure of central Tendency is a single value that attempts to describe a set of Data identifying the central position.

1. Mean:



$$\text{Average/mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Mean should be defined based on 2 factor

Population (N) | Sample (n)

Population mean (μ) | Sample mean ($\bar{x}$)

$$= \frac{\sum_{i=1}^{N} x_i}{N}$$

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

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

Age

N = 6

Population Mean (μ) = $\frac{24+23+2+1+28+27}{6}$

$\mu = 14.5$

* Sample Age = {24, 2, 1, 27}

n = 4

Sample Mean ($\bar{x}$) = $\frac{24+2+1+27}{4}$

$\bar{x} = 13.5$

Practical Application (used in feature Engineering)

Age	salary	family size
-	-	-
NAN	-	Nan
-	Nan	-
NAN	NAN	-

In above Table we have lots of null values we can replace the null values with the mean of particular column name so that there will be no loss of Information

Example of Mean:

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

Mean of Age = 29.6
Mean of Salary = 62

Now, we can replace the NAN values with the mean of age & mean of Salary.

Disadvantage: In above Data if an outliers appears the whole data and mean value will be totally Different and cannot find the accurate value. Here Median, comes into the picture:

2) Median: To prevent the outliers from mean.

Steps to find out median:

① sort the numbers

② Find the central tendency

i. if the no. of elements are even we find the average of central elements

ii. if the no. of elements are odd we find the central elements.

Eg: {1, 2, 3, 4, 5, 6, 7, 8, 100, 200}

1st step we have to sort the numbers ✓

2nd step we have to check odd or even ✓

no. of elements = 10 {even}

if even we have to find average of central elements $= \frac{5+6}{2} = 5.5$

Median = 5.5

if suppose the no. of elements are odd

no. of elements = 11 {odd}

{0, 1, 2, 3, 4, 5, 6, 7, 8, 100, 200}

Median = 5 (central element)

If there is no outliers → use MEAN
If there is outliers → use MEDIAN

3) Mode : Frequently occuring elements

Eg: {1, 2, 2, 3, 3, 3, 4, 5}

- mode = 3

{1, 2, 2, 2, 2, 3, 3, 3, 3, 5}

- mode = 2 & 3

Mode is used for categorical variables
If NAN values are more than the original Data we have to delete the particular row

Types of flowers
Lily
-
Rose
Rose
Sunflower
-
Rose
-
Sunflower

Rose is the mode
∴ so rose can be placed in NAN values

* Measure of Dispersion :

1.) variance (σ^2) (sigma^2)

2.) standard deviation (σ) (sigma)

1.) variance :

Population variance (σ^2)

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

Sample variance (s^2)

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

Eg: $\{1, 2, 3, 4, 5\}$

$$\mu = \frac{15}{5} = 3$$

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

$$= \frac{4+1+0+1+4}{5}$$

$$= \frac{10}{5}$$

$$= 2$$

$\{1, 2, 3, 4, 5, 6, 80\}$

$$\mu = \frac{101}{7} = 14.4$$

$$\sigma^2 = \frac{[(1-14.4)^2 + (2-14.4)^2 + (3-14.4)^2 + (4-14.4)^2 + (5-14.4)^2 + (6-14.4)^2 + (80-14.4)^2]}{7}$$

$$= 719.10$$

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

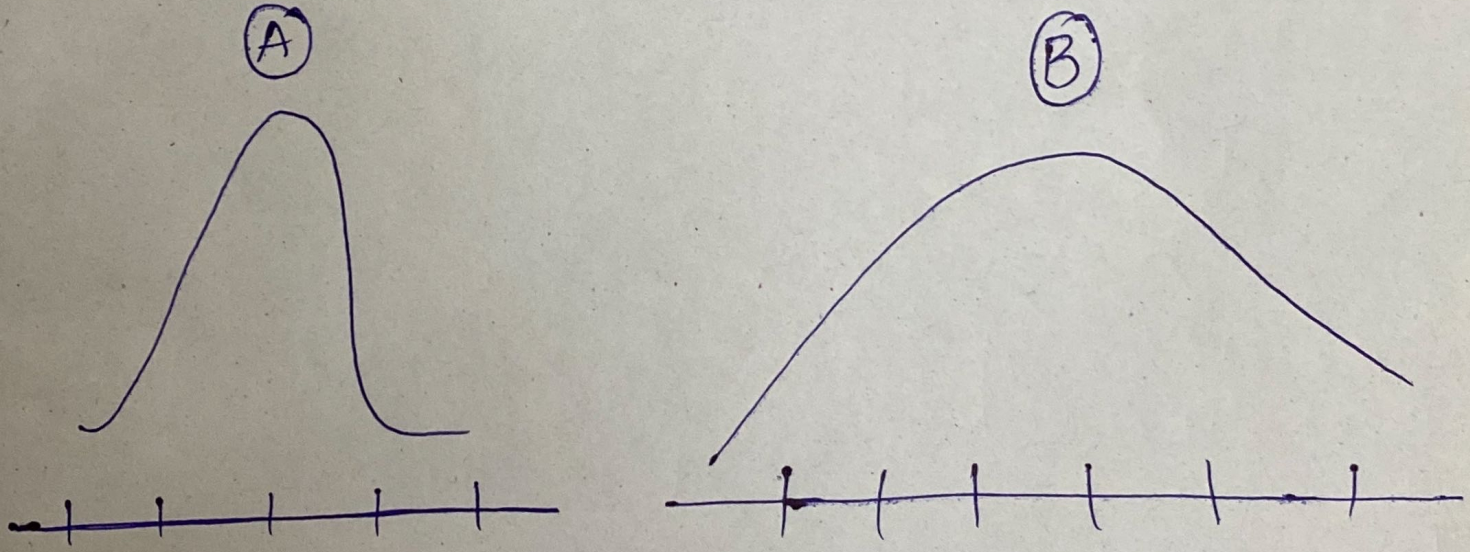
$\{1, 2, 3, 4, 5\}$

$$\mu = \frac{15}{5} = 3$$

$$\sigma^2 = 2$$

$$\sigma = \sqrt{2} = 1.41$$

It Basically tells us How many standard Deviation is away from the mean



Variance will be higher in B
Standard deviation will also be in B

* Percentiles and Quartiles:

Percentage = {1, 2, 3, 4, 5, 6, 7, 8}

% of even no's = $\frac{\text{no. of even numbers}}{\text{Total no of numbers}}$

$= \frac{4}{8} = 0.5$

$= 50\%$

Percentiles: It is the value below which a certain percentage of observation lie.

Eg: Dataset {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9, 9, (10), 11, 11, 12}

What is the percentile rank of 10?

Percentile rank of x = $\frac{\text{no. of values below x}}{n}$

$= \frac{16}{20} = 80\%$

We can say 10 is greater than 80% of entire Distribution.

percentile rank of 8 ?

$= \frac{9}{20} = 0.45 = 45\%$

we can say 45%. of entire distribution is lower than 8

Percentile rank of 6 ?

$= \frac{7}{20} = 0.35 = 35\%$

we can say 6 is greater than 35%. of entire distribution

* What is the value that exists at 25%.

$$\text{value} = \frac{\text{percentile}}{100} \times n$$

$$= \frac{25}{100} \times 20$$

$$= 5 \ (5^{th} \text{ Index})$$

* value that exists at 95%.

$$\text{value} = \frac{95}{100} \times 20$$

$$= 19 \ (19^{th} \text{ Index})$$

= (12) $ from Dataset 3

* 5 Number Summary:

1. Minimum
2. first Quartile (25%.) (Q_1)
3. Median
4. Third Quartile (75%.) (Q_3)
5. Maximum

→ By all these we can remove outliers using Box plot.

Eg: {1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

[Lower fence ⟷ higher fence]

Lower fence = $Q_1 - 1.5\,(IQR)$

$IQR = Q_3 - Q_1$ → it is called as Inter Quartile range

Higher fence = $Q_3 + 1.5\,(IQR)$

$Q_1 = \frac{25}{100} \times (n+1)$ (or) (u) $n = 21$

$= \frac{25}{100} \times 21 = 5.25$ → Index

Average of 5 & 6th Index = (3) 25%.

$Q_3 = \frac{75}{100} \times n+1 \ (21)$

(75%)

$= 15.75 \longrightarrow$ Index

Average of 15 & 16 th Index $= \frac{7+8}{2} = 7.5$

$Q_3 = \underline{\underline{7.5}}$

Lower fence $= 3 - 1.5(7.5-3)$

$= 3 - 1.5(4.5)$

$= -3.65$

Higher fence $= 7.5 + 1.5(7.5-3)$

$= 7.5 + 1.5(4.5)$

$= 14.25$

So, now we have values from $[-3.65 \leftrightarrow 14.25]$

{1,2,2,2,3,3,3,4,5,5,5,6,6,6,6,7,8,8,9,~~27~~}

we have to remove outliers before -3.65 and after 14.25

Now, we got the values and we can plot the box plot for visualization

Minimum = 1

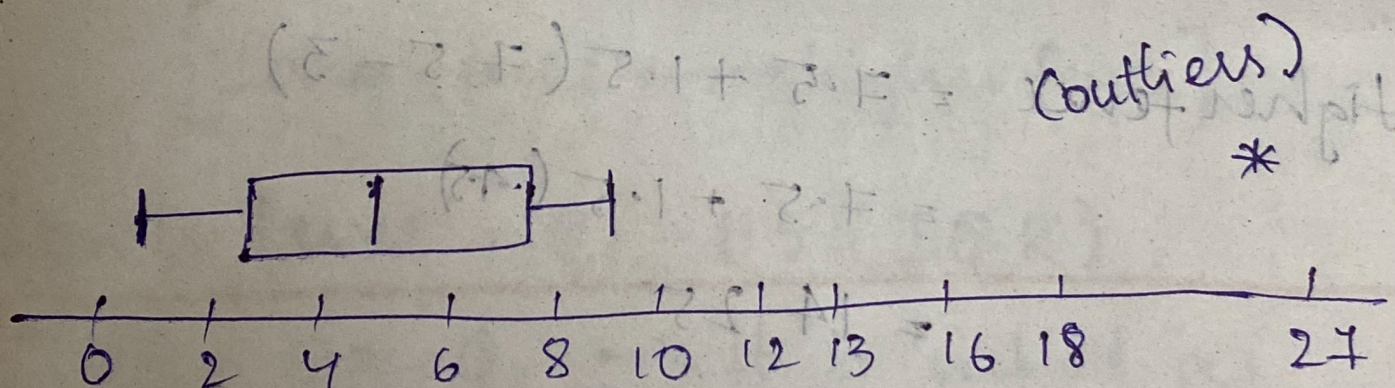
$Q_1 = 3$

Median = 5

$Q_3 = 7.5$

Maximum = 9

Box plot



To treat outliers we use lower and higher fence.