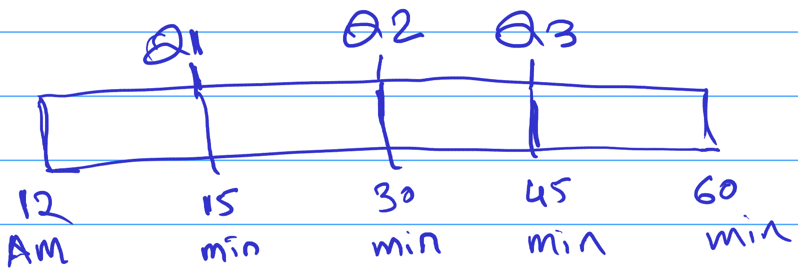
Quantiles & Percentiles

Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to: understand distribution of data, summarize and compare different datasets. They can also be used to identify outliers.

1- Quartile :- Divide the data into four equal parts, Q1 (25th percentile) Q2 (50th percentile) and Q3 (75th percentile)



2- Decile = Divide the data into ten equal part.

3 - Quintile = Divide the data into five equal part

4 - Percentile = Divide the data into 100 equal part.

Note :- 1. Data should be sorted from low to high.
2. you are basically finding the location of an observation. - They are not a actual Value.

Percentile

A percentile is a statistical measure that represents the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile is the value below which 75% of the observations in the dataset fall.

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

PL - Desire percentile location.

P - The percentile rank

N - total No. of observation,

example :-

78	82	84	88	91	93	94	96	98	99
1	2	3	4	5	6	7	8	9	10

$N = 10$ Qu= Find 75th percentile Score

$P = 75$

$$PL = \frac{P}{100}(N+1) = \frac{\cancel{75}^{3}}{\cancel{100}_{4}}(10+1)$$

$$= \frac{3}{4}(11) = \frac{33}{4} = \boxed{8.25} \rightarrow \text{Location}$$

$$= 8^{th}\text{ Location} + 0.25\,(9^{th}\text{ Loc} - 8^{th}\text{ Loc})$$

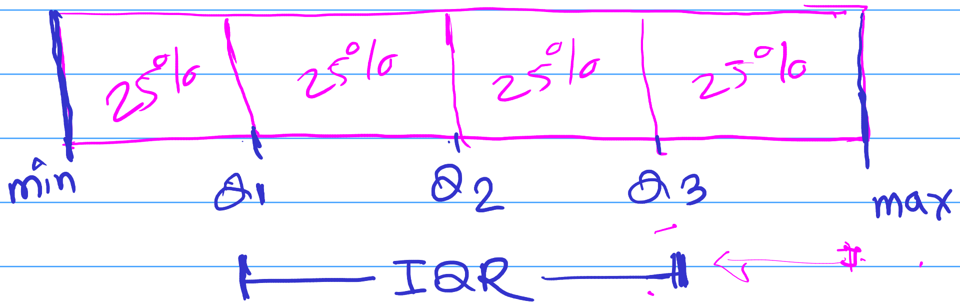
$$= 96 + 0.25(98 - 96)$$

$$= 96 + 0.25(2) = \boxed{96.5} \rightarrow 75^{th}$$

5-Number-Summary ⇒ Box-plot

The five-number summary is a descriptive statistic that provides a summary of a dataset. It consists of five values that divide the dataset into four equal parts, also known as quartiles. The five-number summary includes the following values:

1. **Minimum value**: The smallest value in the dataset.
2. **First quartile (Q1)**: The value that separates the lowest 25% of the data from the rest of the dataset.
3. **Median (Q2)**: The value that separates the lowest 50% from the highest 50% of the data.
4. **Third quartile (Q3)**: The value that separates the lowest 75% of the data from the highest 25% of the data.
5. **Maximum value**: The largest value in the dataset.



[Interquartile Range] IQR

The interquartile range (IQR) is a measure of variability that is based on the five-number summary of a dataset. Specifically, the IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1) of a dataset.

$$IQR = Q_3 - Q_1$$

$$\text{Min.value} = (Q_1 - 1.5\ IQR)$$

$$\text{max Value} = (Q_3 + 1.5 * IQR)$$

$Q_1 = 25\%$ $\quad Q_2 = 50\%$ $\quad Q_3 = 75\%$

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

$\underline{\text{Data}} = [\overset{1}{6}, \overset{2}{213}, \overset{3}{241}, \overset{4}{260}, \overset{5}{280}, \overset{6}{290}, \overset{7}{314}, \overset{8}{321}, \overset{9}{350}, \overset{10}{1500}]$

Outlier Outlier

Location

$Q_1 = \frac{25}{100}(11) = 2.75 = 2^{nd}\text{ location} + 0.75(3^{rd} - 2^{nd})$

$= 213 + 0.75(241 - 213)$

$\boxed{Q_1 = 234}$

location

$Q_2 = \frac{50}{100}(11) = 5.5 = 5^{th} + 0.50(6^{th} - 5^{th})$

$\boxed{Q_2 = 285.5}$

Location

$Q_3 = \frac{75}{100}(11) = 8.25 = 8^{th} + 0.25(9^{th} - 8^{th})$

$= 321 + 0.25(350 - 321)$

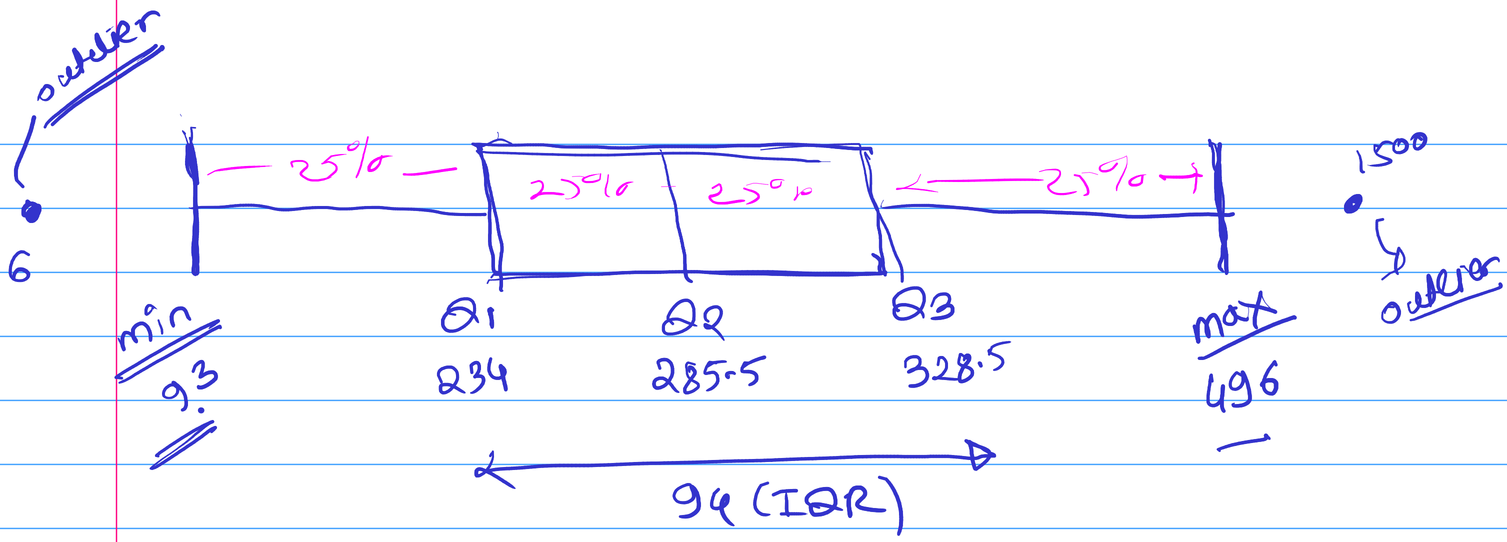
$\boxed{Q_3 = 328.25}$

$IQR = Q_3 - Q_1 = 328.25 - 234$

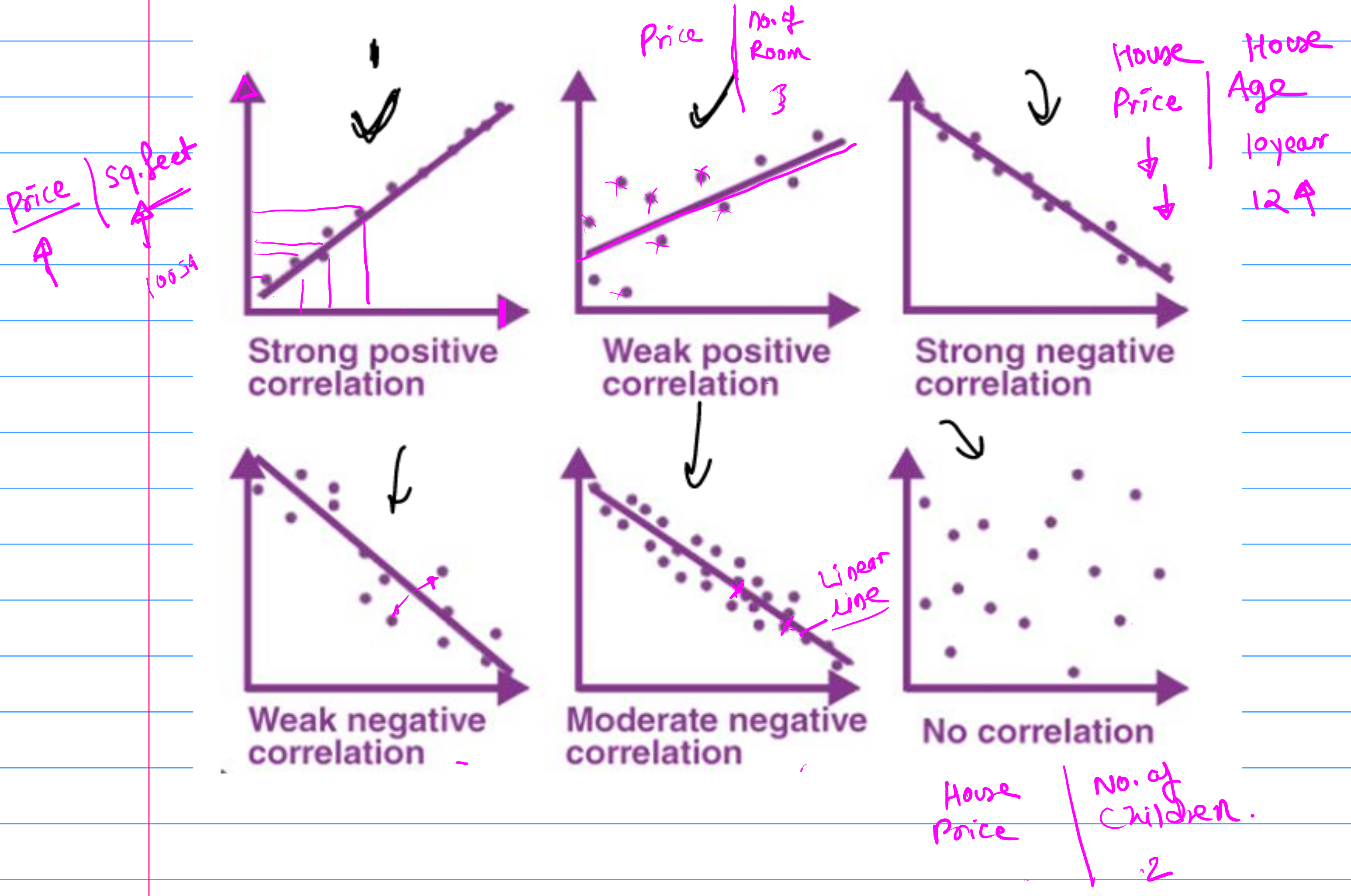
$\boxed{IQR = 94}$

$\min = Q_1 - 1.5(IQR) = 93$

$\max = Q_3 + 1.5(IQR) = 496$

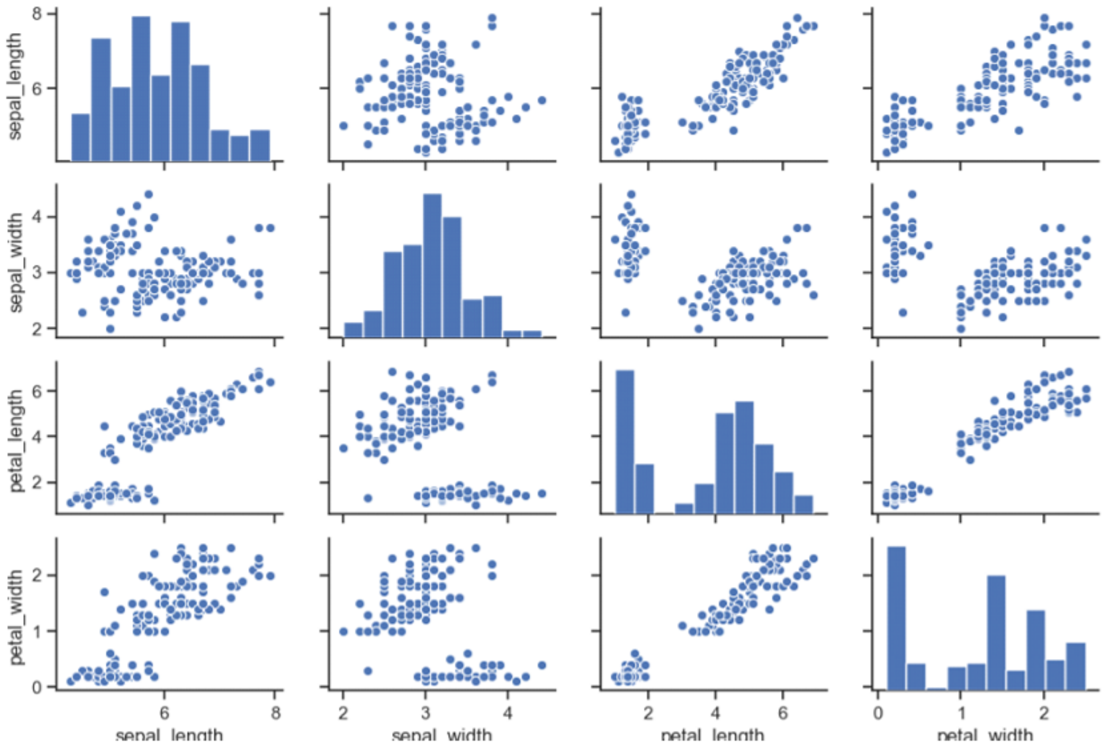


Correlation



- Multivariate Analysis

Pairplot



C1

C1

C1

C2

Type of Random Variable.

Algebric Variable = $x + 5 = 10$ $x = 10 - 5$ $\boxed{x = 5}$

variable

fix value.

Random Variable = x = Rolling a dice.

$$x = \{1, 2, 3, 4, 5, 6\}$$

Discrete

Tosing a coin $x = \{H, T\}$

$x = \{1, 2, 3, 4, 5, 6\}$

Continious

$x = \{0, 10\}$ → 0 to 10

0.1, 0.22, 0.95, 9.75,

➢ Types of Probability Distribution: -

1. Normal or Gaussian Distribution
2. Bernoulli Distribution
3. Uniform Distribution
4. Poisson Distribution
5. Binomial Distribution
6. Log-Normal Distribution

Normal or Gaussian Distribution: -

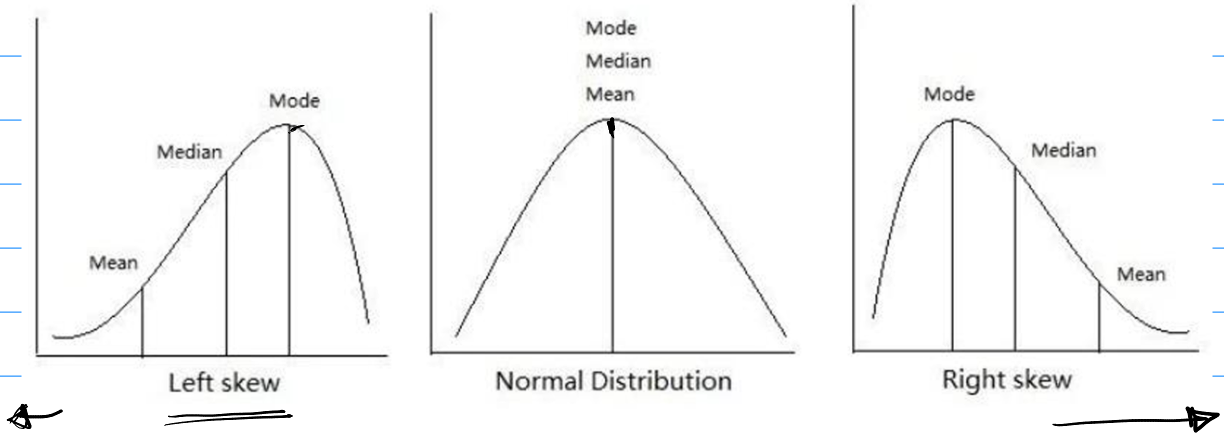
0–10

Probability Distribution function

- it's concerned with Continuous random variables {PDF}
- Normal distributions are symmetrical, but not all symmetrical distributions are normal

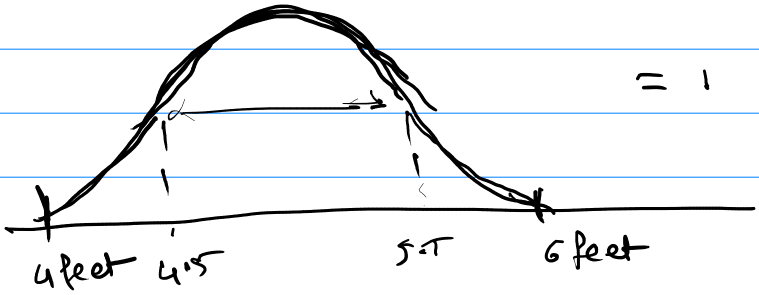
Characteristics of Normal Distribution

- mean = median = mode
- Symmetrical about the center
- Unimodal
- 50% of values less than the mean and 50% greater than the mean



- Height of student in class. / Marks in Test

4 feet ||||||||||| —— 6 feet



= 1

1 mark — 15-60 ~ 20 Mark