Coefficient of Variation (CV)

Lecture

CV is the ratio of Standard deviation to its mean.

ex- Salary | experience | skillset.

$$CV = \left(\frac{\text{Standard deviation}}{\text{mean}}\right) \times 100$$

Covariance & Correlation

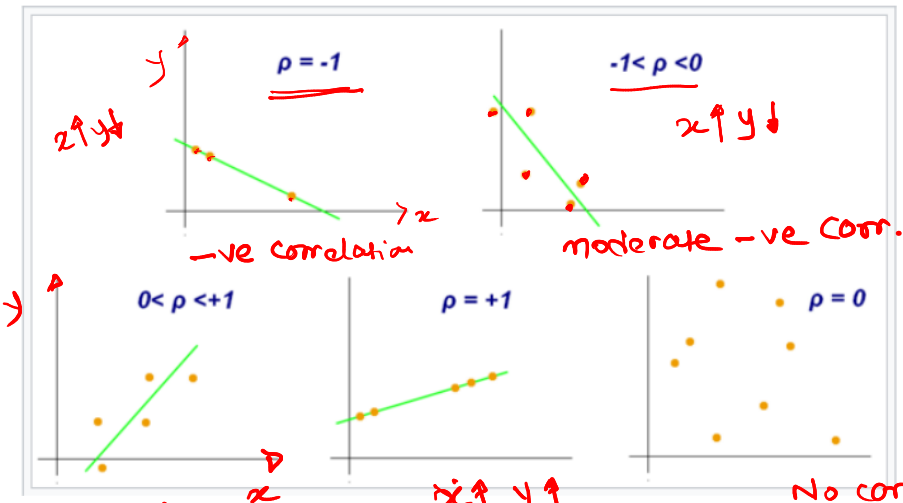
Range
$-\infty$ to $+\infty$

Range
-1 to 1

Range = -1 to 0 to 1

A. Pearson correlation coefficient - The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Linear line
$(y = Ax + b)$



-ve correlation

moderate -ve corr.

No correlation

x↑ y↑ moderate +ve corr

x↑ y↑ +ve corr

House price

Price	sq. feet Area
↑	↑

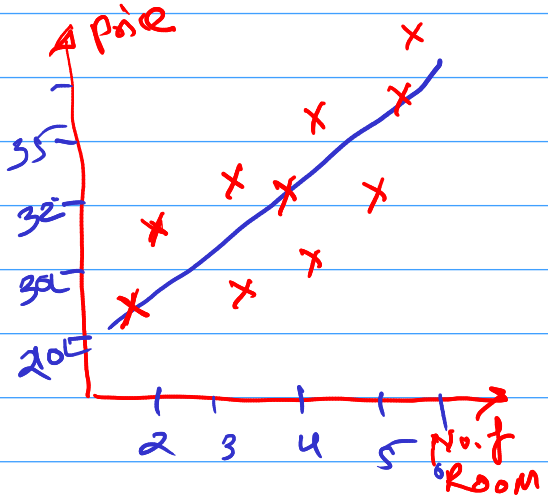
Strongly +ve corr.

P = 1

Price	No. of Room	Sq. Feet
30L	3	1000
32L	4	1000

Moderately +ve

P = 0 to 1



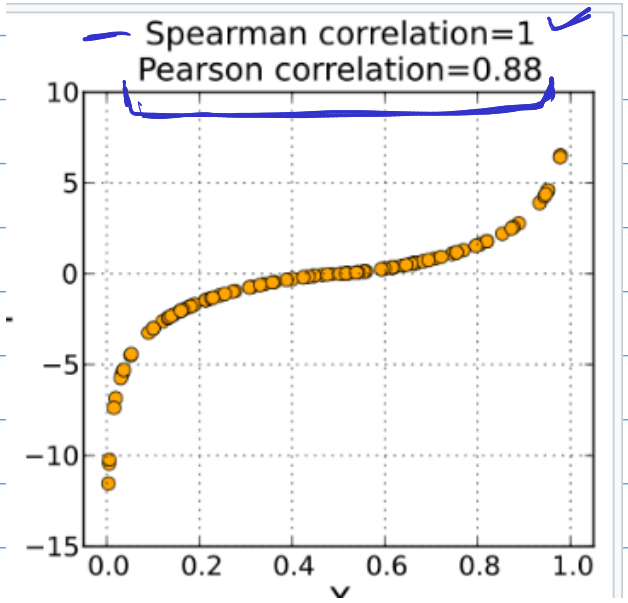
Price	Age of House	Sq feet	No. of Room
10L	10 year	1000	4
5L	20 year	1000	4

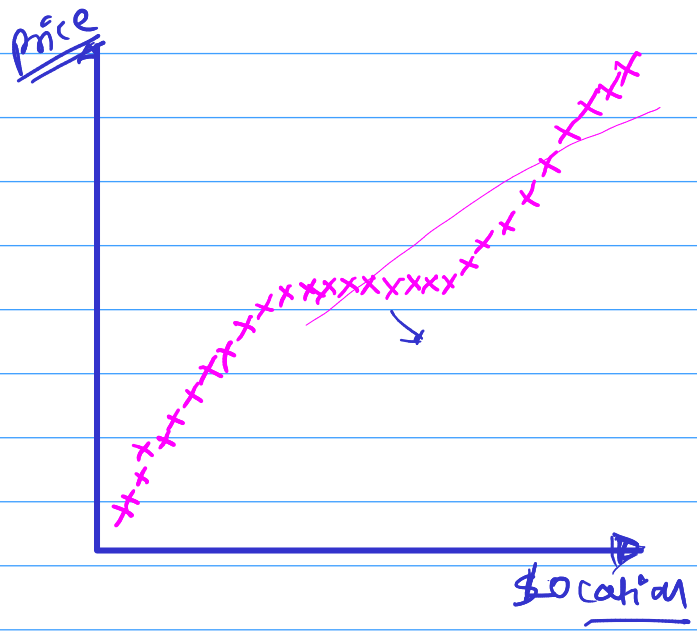
House Price	No. of people

Non Correlation

B. Spearman's rank correlation coefficient:- A correlation can easily be drawn as a scatter graph, but the most precise way to compare several **pairs of data** is to use a statistical test - this establishes whether the correlation is really significant or if it could have been the result of chance alone.

Spearman's Rank correlation coefficient is a technique which can be used to summarise the strength and direction (negative or positive) of a relationship between two variables. The result will always be between 1 and minus 1.





Graphical Representation.

1. Univariate Analysis — single feature
2. Bivariate Analysis — two features
3. Multivariate Analysis — Many.

Categorical Univariate Analysis

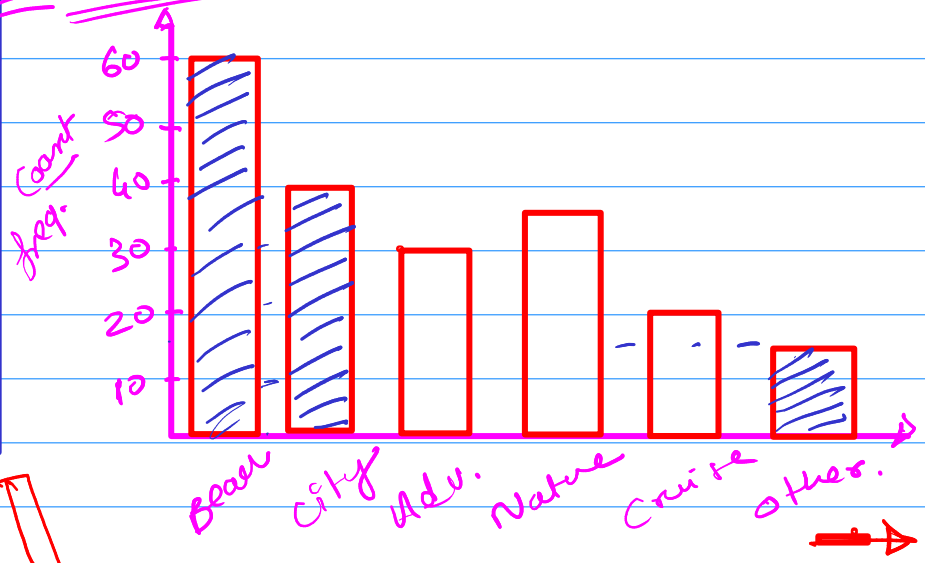
↳ Frequency Distribution Table.

Ex - 200 students - Vacation

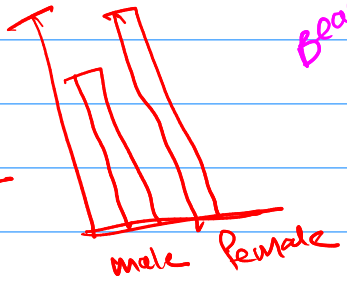
Beach, City, Adventure, Nature, Cruise and others.

Type of Vacc.	Frequency (Value_counts)
Beach	60
City	40
Adv.	30
Nature	35
Cruise	20
Other	15

= Bar Chart



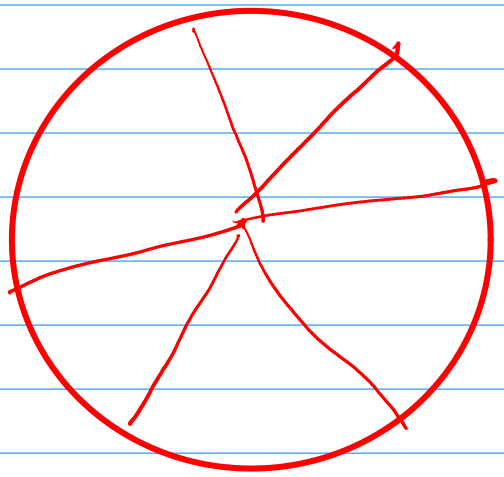
male	30
female	45



Relative freq. = $\frac{\text{freq}}{\text{Total No.}}$

100%

Type of Vacc.	Frequency (Value_counts)	Relative freq.	
Beach	60 = 60/200	0.3	30%
City	40	0.2	20%
Adv.	30	0.15	15%
Nature	35	0.175	17.5%
Cruise	20	0.1	10%
Other	15	0.075	7.5%
		= 1	100%

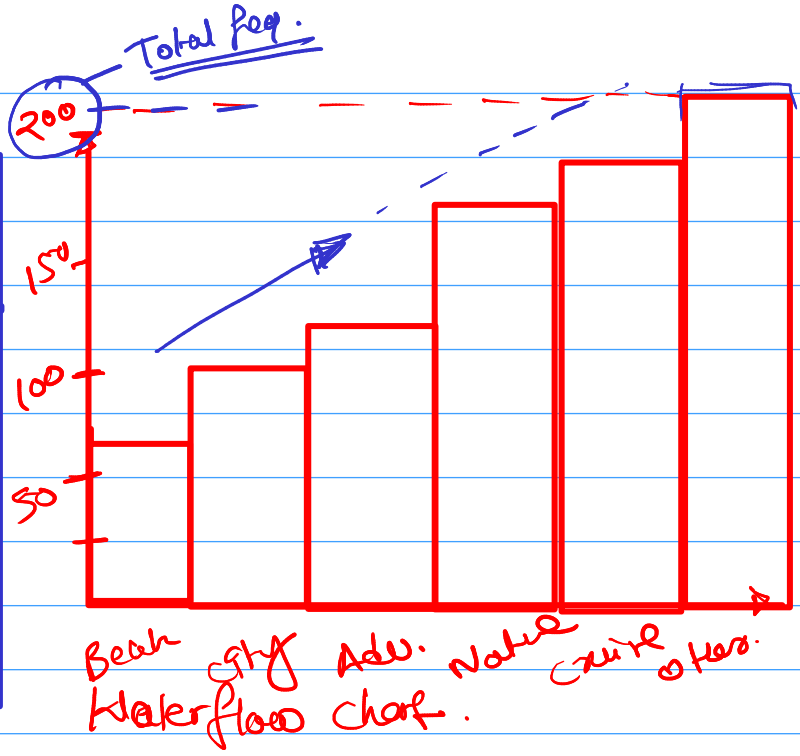


Pie-chart

Cumulative freq

first freq + 2nd freq.

Type of Vac.	Frequency (Value_counts)	Cumulative freq.
Beach	60	60
City	40	100
Adv.	30	130
Nature	35	165
Cruise	20	185
Other	15	200



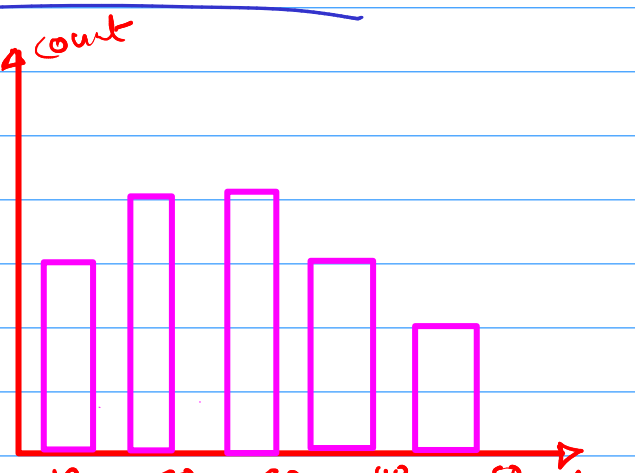
Waterflow chart.

Numerical → Feature Engg → Categorical

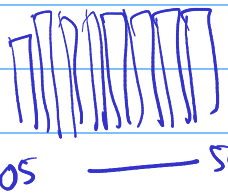
Age	Age	total
20	20	
21	05	
32	06	
38	07	
05	08	100
10	09	
20	10	
35	11	
40	12	

Bins

Age-group
0-10
11-20
21-30
31-40
41-50

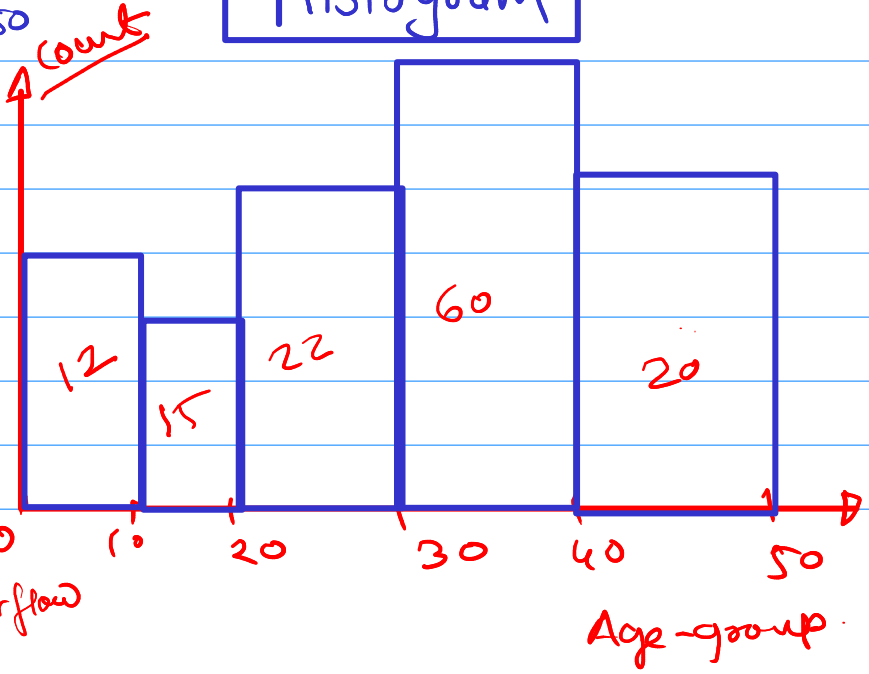


Bar chart.



Histogram

No of Stud	Bins - Age group	freq
12	0-10	12
15	10-20	27
22	20-30	49
60	30-60	109
20	60-50	129



* waterflow

School 1st - Boys → Age.

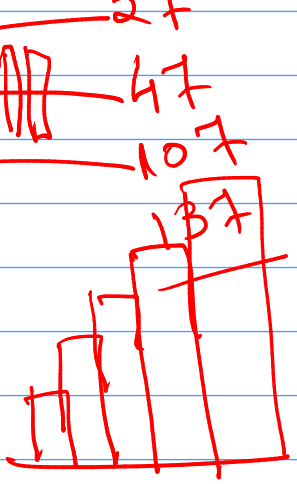
Age group	Student (freq)
0 - 10	12
10 - 20	15
20 - 30	20
30 - 40	60
40 - 50	30

Age (Numerical)

1 - 12
2 - 15
3 -
4
5
6
.
.
.
50

Relative freq

12
27
47
107
137



Categorical — Categorical

↳ Contingency Table / CrossTab

Titanic

Survived | Pclass

Pclass

Survived	1	2	3
0	42	31	63
1	71	118	13

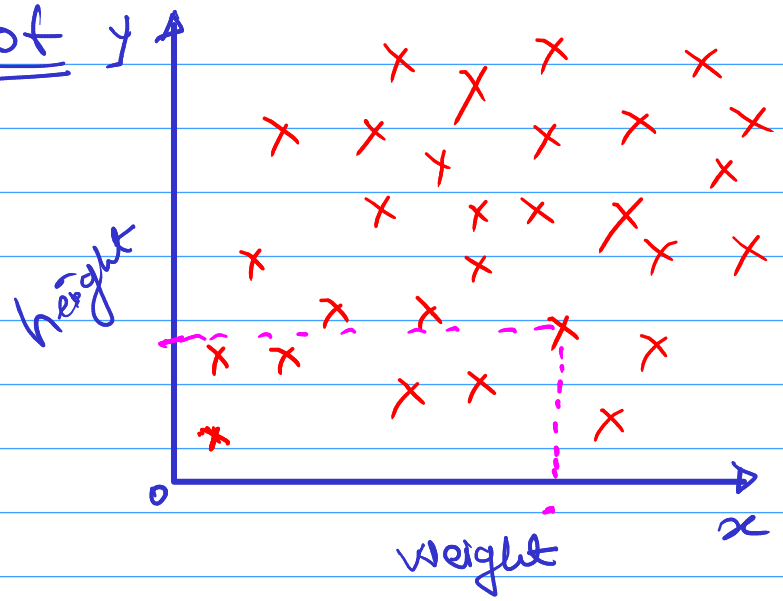
Survived	sex
0	Male
1	Female

Sex

Survived	Male	Female
0	23	69
1	74	189

Numerical & Numerical

Scatter plot



Categorical & Numerical / Numerical & Categorical

Age gr	0-10	11-20	21-30
Male	32	41	110
Female	15	18	120