Descriptive Statistics

→ Statistics is the science of learning from Data

→ Data is essentially numbers (text/Symbols) which represent some information.

→ It helps to think of data as value of quantitative and qualitative variables.

→ What are the variables types:
- Numerical or Quantitative (continious & Discrete)
- Categorical or Qualitative (always Discrete).
  - Nominal: It is not possible to arrange data in any order.
  - Ordinal: It is possible to arrange data in order.

Example of Nominal data: represent characteristics like Gender (Male / Female / Transgender)
Marital Status — (Married / Unmarried)
Home State — (States of India)

Example of Ordinal Data: Color of Terror alert posible value
Green, Yellow, Orange & Red
↓ Very low Risk
Very high Risk.

ordered Categorical Variable

Player Quality: A class, B class, C class, D class
↓ Very high skill player
Significantly low skill Players.

Descriptive Statistics is method to Quantitatively describing the data

↳ Graphical Representation

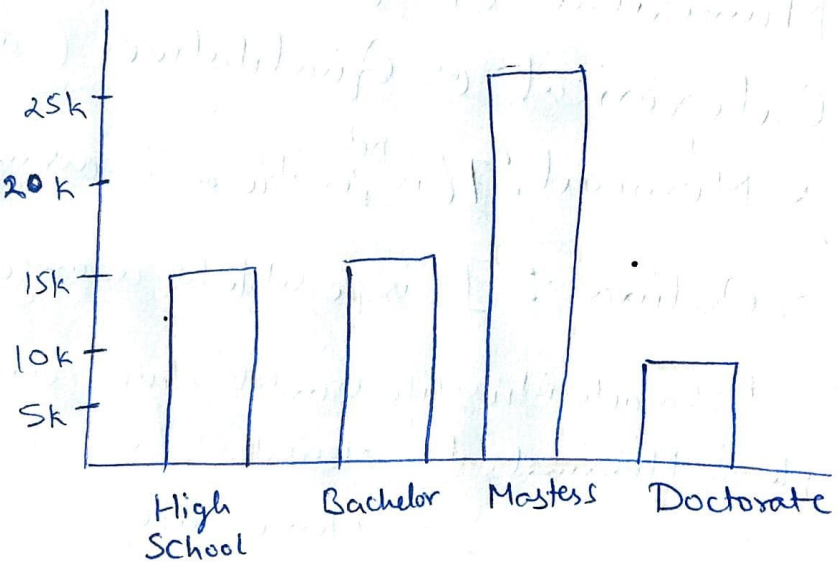
↳ Tabular Representation

↳ Summary Statistics.

Graphical Representation

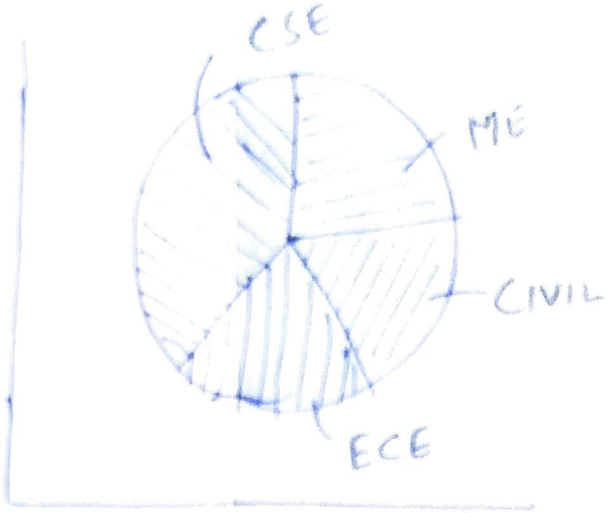
Categorical Variable

↳ BAR CHART



Ordinal Variable (Data)

↳ PIE CHART

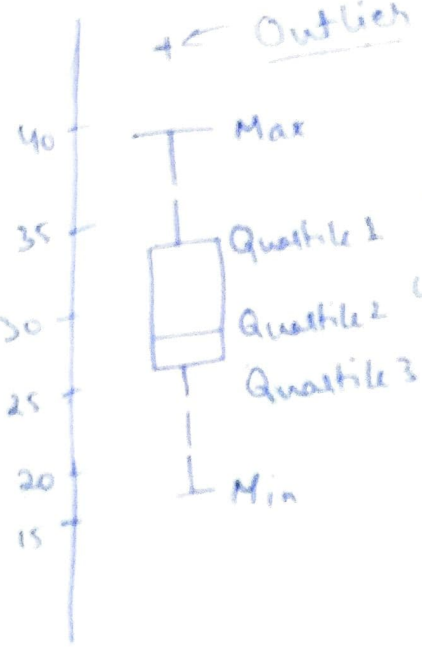


↳ Quantitative Variable

Box & Whisker Plot

~~Box Plot~~

Ex 11, 22, 20, 14, 29, 8, 35
27, 13, 43, 10, 24, 17



Lower Quartile 25%
Upper Quartile 75%

Step 1 - arrange the numbers in ascending order

~~Histogram~~

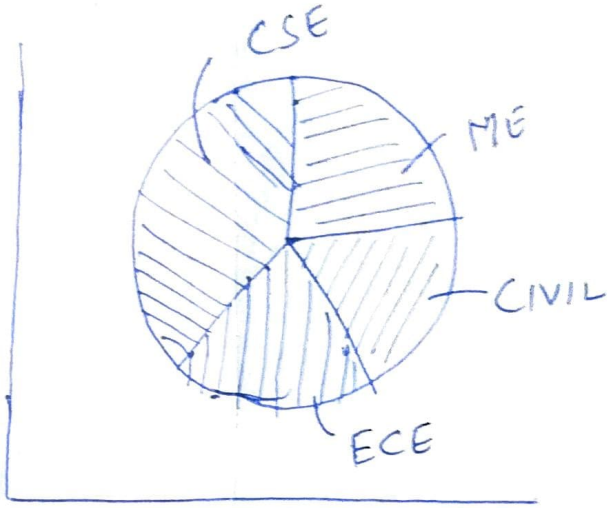
Min
8, 10, 11 | 13, 14, 17 | 20, 22, 24, 27 | 29, 35, 43
12
Median

Find Median Q_1 Q_2 Q_3

Now to check the min & max are not outliers
Now we need to determined the range of numbers in which the outlier could be

$[Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR]$ Any number outside this range is an outlier.

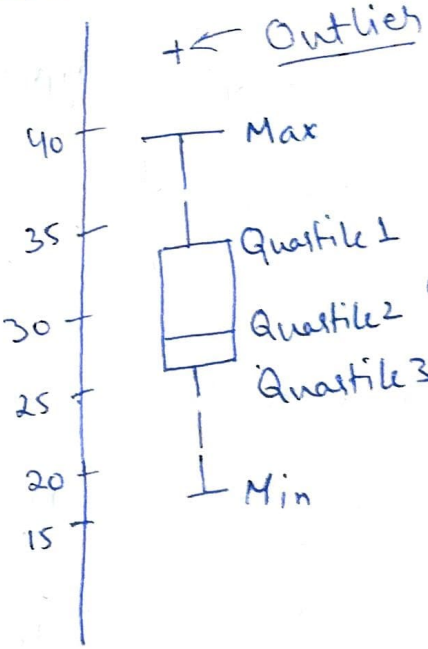
↳ PIE CHART



↳ Quantitative Variable

Box & Whisker Plot

~~Box Plot~~



Lower Quartile 25%

Upper Quartile 75%

Ex 11, 22, 20, 14, 29, 8, 35
27, 13, 49, 10, 24, 17

Step 1 - arrange the numbers in ascending order.

~~Histogram~~

Min ↓ 8, 10, 11 | 13, 14, 17 | (20), 22, 24, 27 | 29, 35, 49 ↓ Max

12 — Median — 28

Find Median Q_1 Q_2

Now to check the min & max are not outliers

Now we need to determined the range of numbers in which the outlier could be

$[Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR]$ Any number outside this range is an outlier.

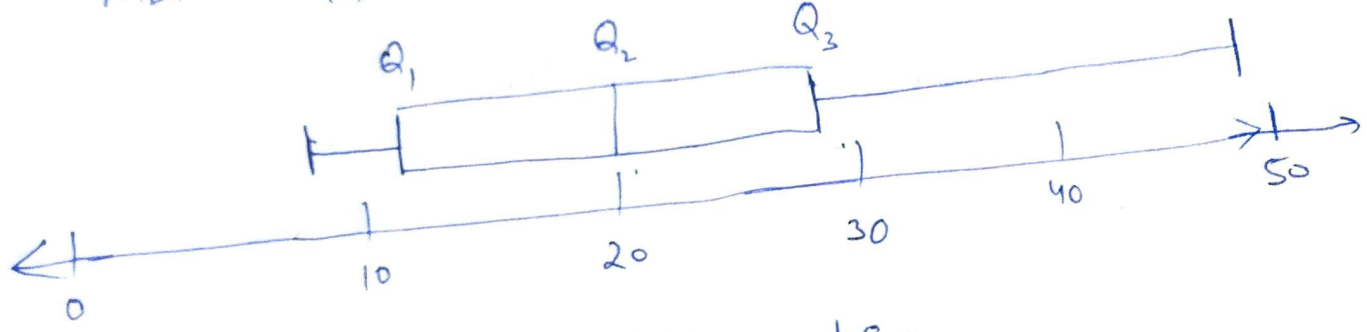
IQR : Inter Quartile Range is the difference between Q_3 & Q_1.

$$IQR = Q_3 - Q_1 = 28 - 12 = 16$$

$$Q_1 - 1.5\,IQR = 12 - 1.5 \times 16 = -12$$

$$Q_3 + 1.5\,IQR = 28 + 1.5 \times 16 = 52$$

$\Rightarrow [-12, 52]$ hence minimum(8) & maximum(43) is not an outlier



If the data have outlier then

Let us take an example

18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22

Step 1 15, 18, 20, 27, 25, 29, 32, 34, 38, 41, 43, 46, 54, 76

15, 18, 20 | 22 25, 29, 32 | 33 | 34, 38, 41 | 43 | 46, 54, 76

Q_1 = 22, Q_2 = 33, Q_3 = 43

$$IQR = Q_3 - Q_1 = 43 - 22 = \boxed{21}$$

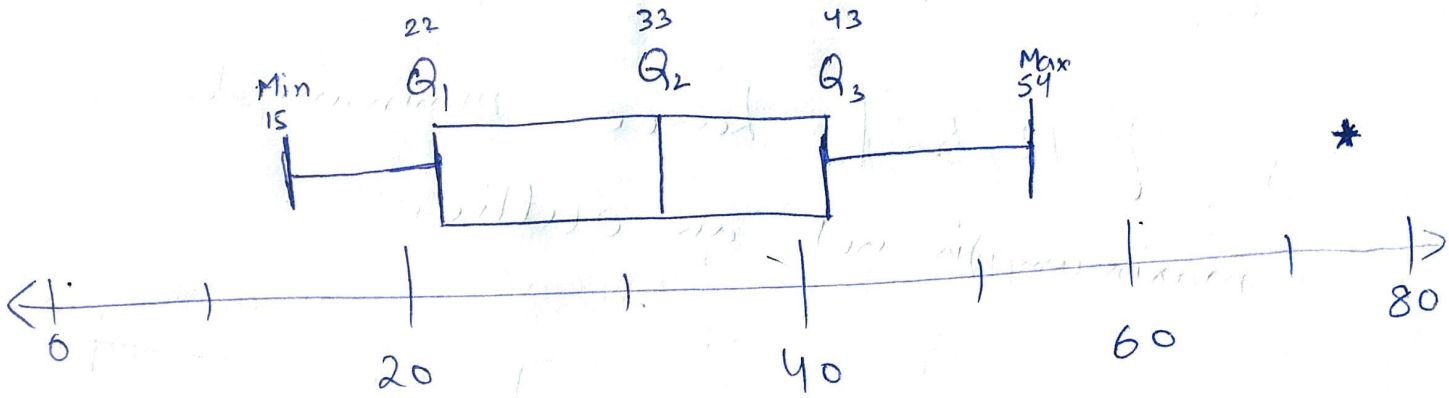
$$[Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR]$$

$$[22 - 1.5 \times 21,\ 43 + 1.5 \times 21]$$

$$[-9.5,\ 74.5]$$

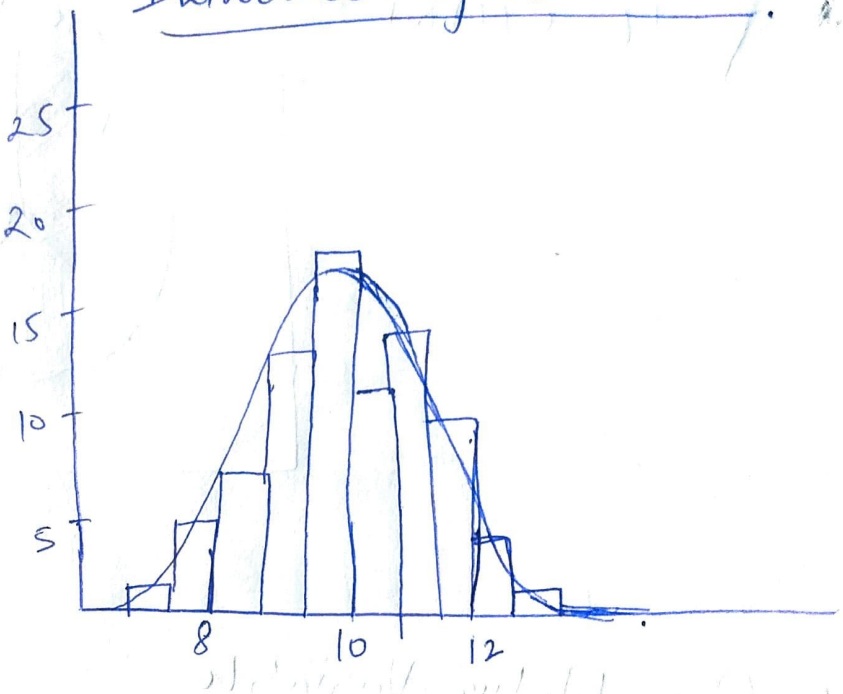
Min 15 is in range

Max 76 is outside the range



Histogram

Introduced by Karl Pearson.



How to Choose Number of bins.

Scott's rule

$$\hat{\Delta}_b = 3.49\,\hat{\sigma}\,n^{-1/3}$$

$\hat{\sigma}$: estimate of Standard deviation

n : Number of data points

Freedman-Diaconis Rule.

$$\hat{\Delta}_b = 2\left(q_{75} - q_{25}\right) n^{-1/3}$$

where $Q_{75} - Q_{25}$ = IQR (Inter Quartile Range.)

"Optimal Data-Based Binning of Histogram" by Kelvin H. Knuth}, arXiv: physics/0605197.

(1) Number of bins k can be assigned directly by choosing bin width h as

$$k = \left\lceil \frac{max - min}{h} \right\rceil .$$

(2) Square - Root Rule

$$k = \lceil \sqrt{n} \rceil$$

(3) Sturges formula

$$k = \lceil \log_2 n \rceil + 1$$

(4) Rice Rule

$$k = \lceil 2\sqrt[3]{n} \rceil$$

(5) Doane's Formula: Doane's modifies Sturges rule to improve the performance for non Gaussian data

$$k = 1 + \log_2(n) + \log_2\left(1 + \frac{|g_1|}{\sigma_{g_1}}\right)$$

g_1 : 3rd Moment (Skewness) of data

$$\sigma_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$$

Shimazaki & Shinomoto's Rule.

$$\arg\min_h \frac{2\bar{m} - v}{h^2}$$

where $\bar{m}$ & v are mean & biased variance of a histogram with bin-width h

$$\bar{m} = \frac{1}{k}\sum_{i=1}^{k} m_i \quad \& \quad v = \frac{1}{k}\sum_{i=1}^{k}(m_i - \bar{m})^2$$

Variable bin widths

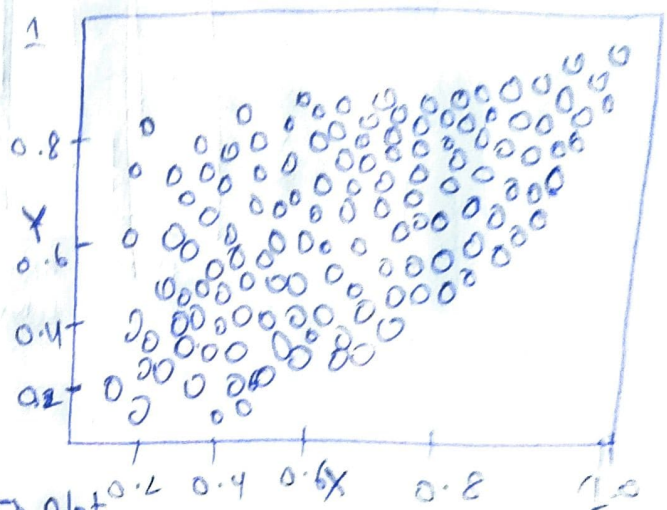
where number of samples in each bin is expected to be approximately equal.

Graphical Representation: Multiple Variable

* Scatter Plots: Two Quantitative Variables (correlation)

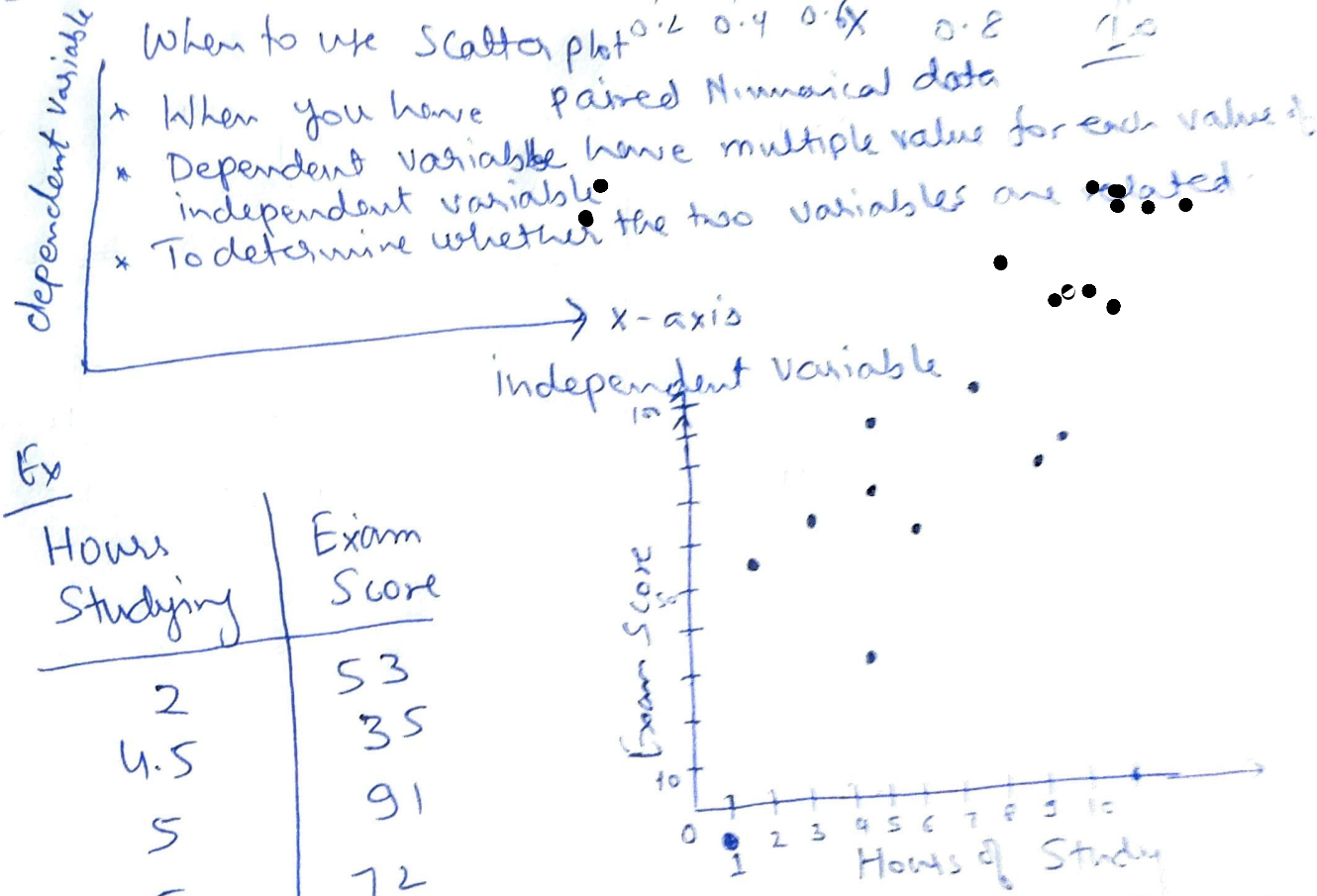
It is used to Show a relationship between two variables.

A Symbol, usually a dot is used to Show a data pair.



When to use Scatter plot

* When you have paired Numerical data
* Dependent variable have multiple value for each value of independent variable
* To determine whether the two variables are related



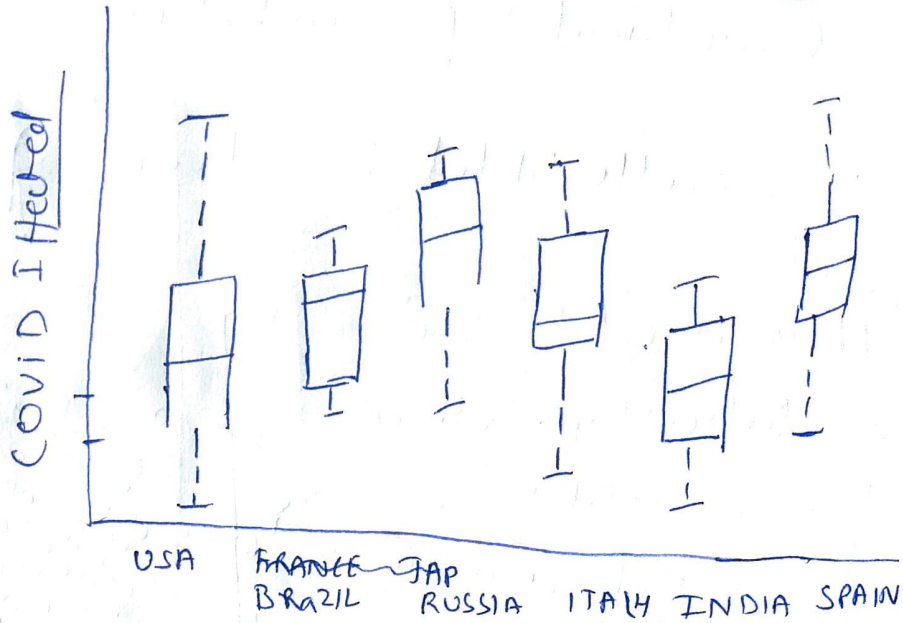
Ex

Hours Studying	Exam Score
2	53
4.5	35
5	91
5	72
6	60
3	62
10	85
9.5	78
8	99

Find the Relationship between the variables. if exist!

Matrix Scatter Plot: To represent Correlation between more than two variables

Box Plots


COVID Healed
USA
FRANCE
JAP
BRAZIL
RUSSIA
ITALY
INDIA
SPAIN

Contingency Table : 2 Categorical Variables with frequency of occurance as the theme

Results		WORK EX	
		Y	N
MBA	Y	22	7
	N	32	17

How many of Employees have MBA before joining the organization

How many of Employees have work experience befor joining the organization.

Ex Smoker/Non Smoker in Org:

joint dist.

Gendy	Smoker	Non-Smoker	Total
Male	7 [illegible]	[illegible]	[illegible] 57.14%
Female	3 [illegible]	[illegible]	[illegible]
Total	10 [illegible]	[illegible]	20 [illegible]

Relative frequency

P(smoking | Male), P(smoking) dist.

Marginal