working with a sample ← Statistics → a way to get information out the data.

outcome of stat is always APPROXIMATION

EDA

DESCRIPTIVE STATISTICS [make sense of the data]

[Predictions → close to correct] INFERENTIAL STATISTICS → ML model

Technique for organising and summarizing info.

using
1. Table
2. Graph
3. Charts

Technique for drawing and measuring the reliability of conclusions of the population based on information.

Statistical Data (Stats)

POPULATION ← (maths)

The collection of all individual elements or items under consideration in a stat study.

[UNIVERSAL SET]

You have all the data

eg: INFOSYS → Personality Trait Segmentation → data of all emps.

employee → finite

(Stats) → SAMPLE

A sample is a SUBSET of the population from which info can be obtained.

[SUBSET OF UNIVERSAL SET]

Facebook → % people in the world, loves Chicken Burger ?

→ all the data of entire world.

Get answer → Statistics

Types of data you deal with in Data Science

NUMERICAL
holds mathematical weightage

Continous numerical data *
eg: weight of human

any entity which changes w.r.t time

Discrete numerical data. *
eg: SSC, SAT score

- any entity which remains fixed irrespective of time.
- Always remain in a range (0-100)

Categorical
no mathematical weightage

eg. gender → M, F, T

States in India

- data that classifies the entity.
- The data doesn't hold any mathematical weightage.

Ordinal
holds mathematical weightage

eg:- Feedbacks.

- Ratings. ☆☆☆☆☆
  - mathematical impact is present
  - define category
    - '5 star'
    - '4 star'
    - '3 star'
    - '2 star'
    - '1 star'

Categorical with mathematical weightage

Types of Variables

Qualitative variable
eg: Categorical data.

Quantitative variable
eg: numerical data

Discrete
numbers in a finite range

Continous
numbers in an infinite range

DESCRIPTIVE STATISTICS [Data Exploration Phase / Exploratory Data Analysis (EDA)]

- Descriptive stats is the first step to deal with the SAMPLE
- mindset of a Data Scientist while performing Descriptive Stats include:
  - To check the distribution of data
  - To identify and remove OUTLIERS
  - To identify and deal with INAPPROPRIATE data.
  - To identify and deal with MISSING values in data.
  - To help discover any kind of ASSOCIATION (Association Rule mining) / RELATIONSHIP / PATTERN between two variable in a dataset

Amazon.in shopping website.

- Prashant N → OnePlus 7, headphone, Data Science with Python
  → Hadoop, dumbells, Guitar, guitar picks.
- Ranjan → iPhone XR, earpods.
  → Hadoop, Guitar, guitar tuner

Nikhil ⟶ Hadoop

Guitar

Amazon will recommend Nikhil

Data Exploration Techniques

- Numerical EDA — reading and interpreting numbers.
- Visual EDA — reading graphs and interpreting the same.

Visual EDA

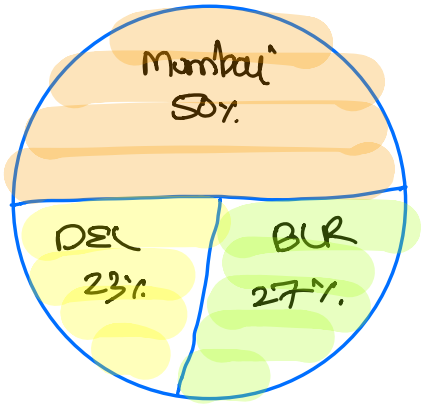
① Dealing with Qualitative data. (Categorical data)

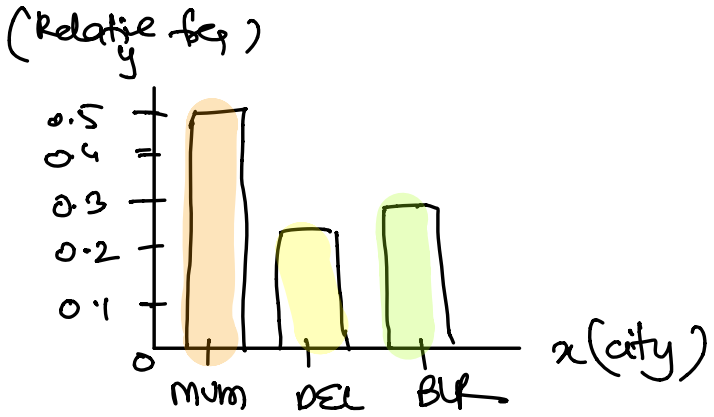
Represent qualitative data using two approach

Graph → Pie, Bar

Frequency Distribution Table

city	Tally	Freq.	R.F
Mum	𝍸 IIII	9	9/18 → 0.5/50%
DEL	IIII	4	4/18 → 0.23/23%
BLR	𝍸	5	5/18 → 0.27/27%
		18	

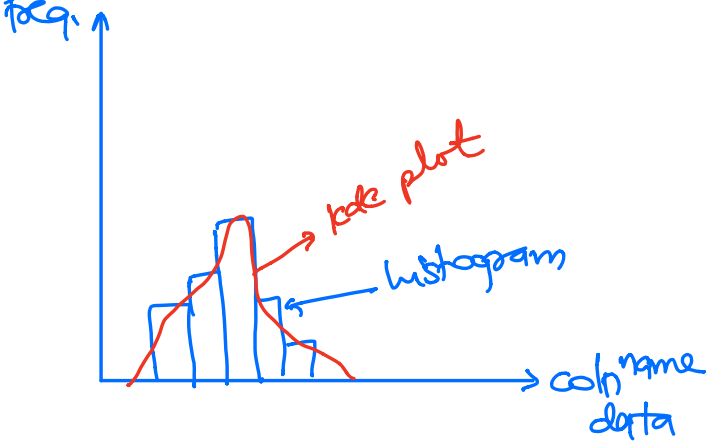




Quantitative Data (numerical data)

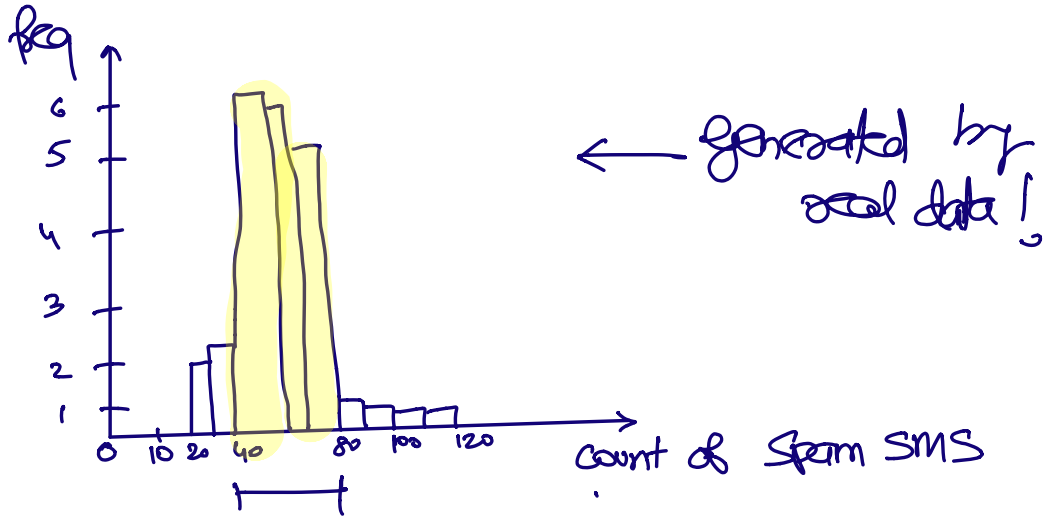
freq dist table

Histograms / KDE plot



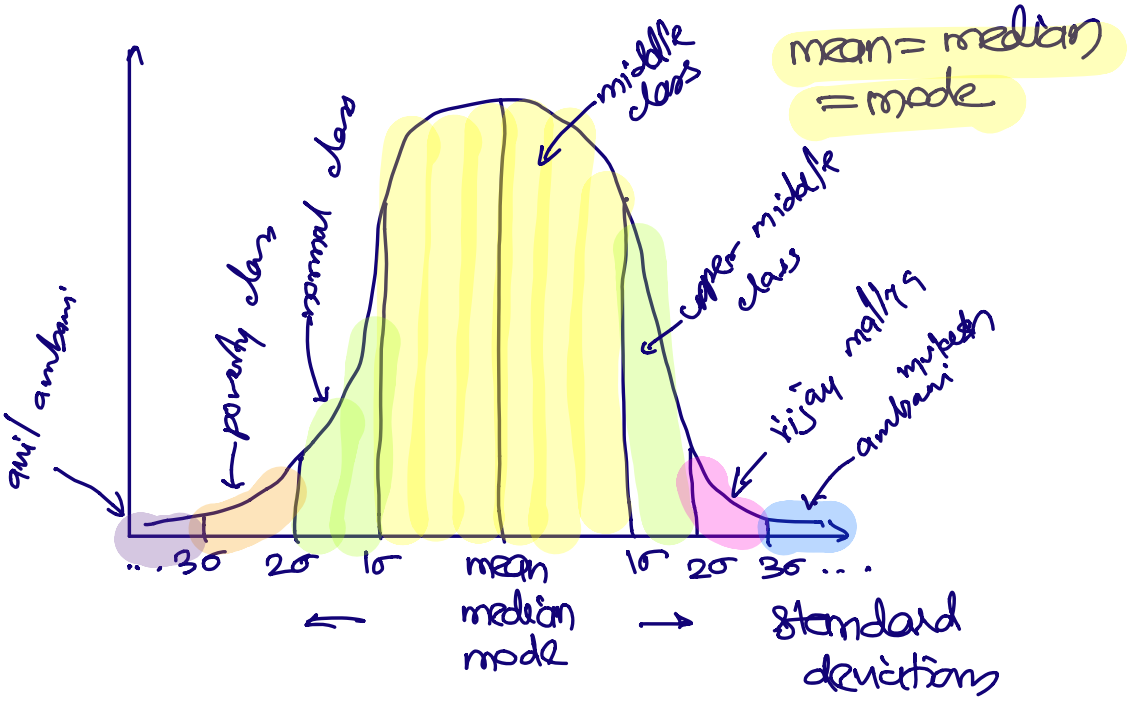
DATA DISTRIBUTION → only applicable for Quantitative data!
(defining the partial nature of the data)

(Numerical → continuous, Discrete)

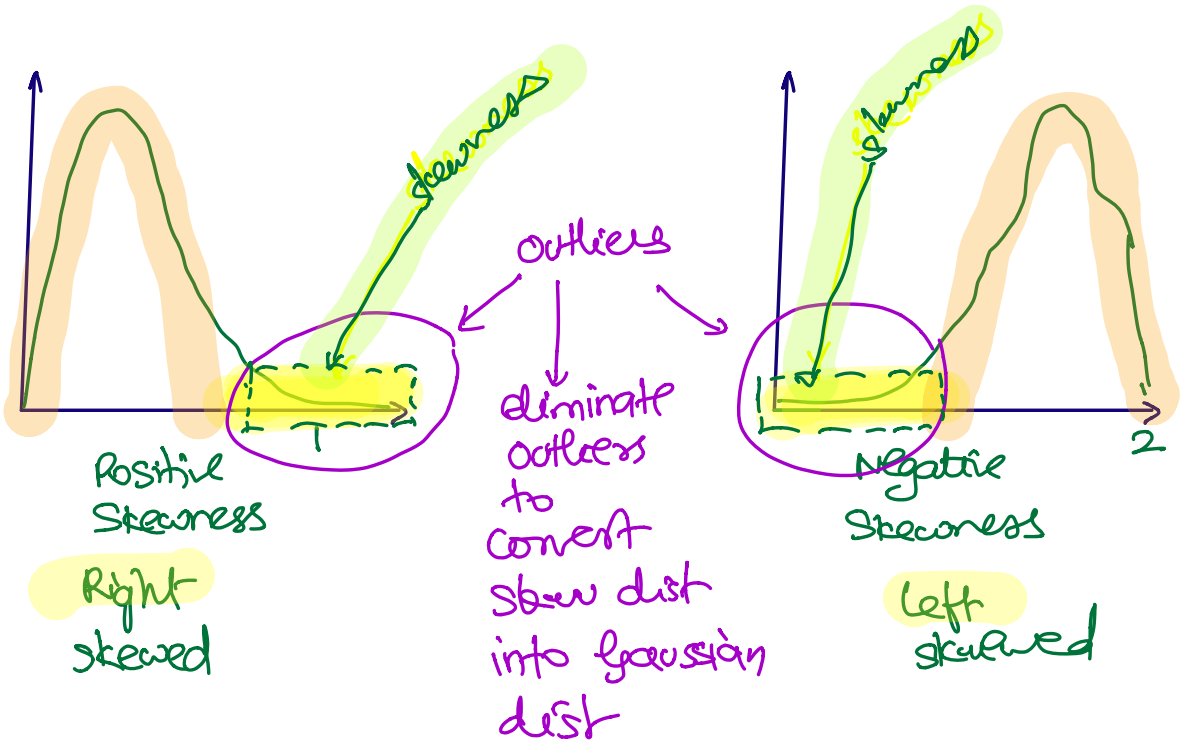


← generated by real data!

ⓐ Gaussian distribution (Normal distribution)



Skeco distribution

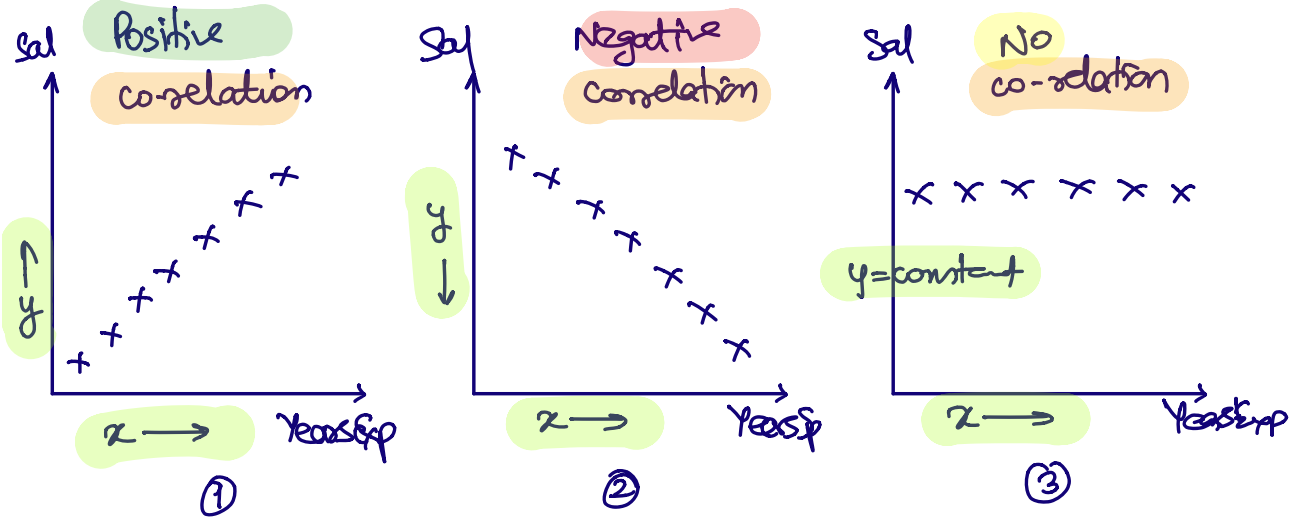


Relationships in data distribution

linear relationship.

present

Regression analysis



OUTLIERS → Outliers are extreme values that affect

Stat perspective ⓐ Gaussian nature of data.

Domain perspective ⓑ Domain nature of the data.

eg: Indian Stock market → NSE (High volatility)

market falling / Crashing ↓

AXIS Bank → Buying

To detect outliers and remove the same using STAT perspective, you can use the concept of Quartiles.

Quartiles

min

Q1 25%

Q2 50% ← mean | median | mode (Purely subject to dist. of data)

Q3 75%

Q4 max

IQR (Remove outlier columnwise)

Inter Quartile Range → $Q_3 - Q_1$

$$lowerRange = Q_1 - (1.5 * IQR)$$

$$upperRange = Q_3 + (1.5 * IQR)$$

Algo :-

① Ensure your column data is sorted in ascending order

② Get value of Q_3 and Q_1 and calc IQR. using the below,

$$IQR = Q_3 - Q_1$$

③ Calc. lower range = $Q_1 - (1.5 * IQR)$

④ Calc. upper range = $Q_3 + (1.5 * IQR)$