LECTURE - 31

NORMAL DISTRIBUTION:

A function that represents the distribution of many random variables as a symmetrical bell-shaped curved graph.

→ Also called as Gaussian Distribution, is a Probability distribution that is symmetric about the mean.

→ It shows the data which is near to the mean are more frequent in occurance than the data far from the mean.

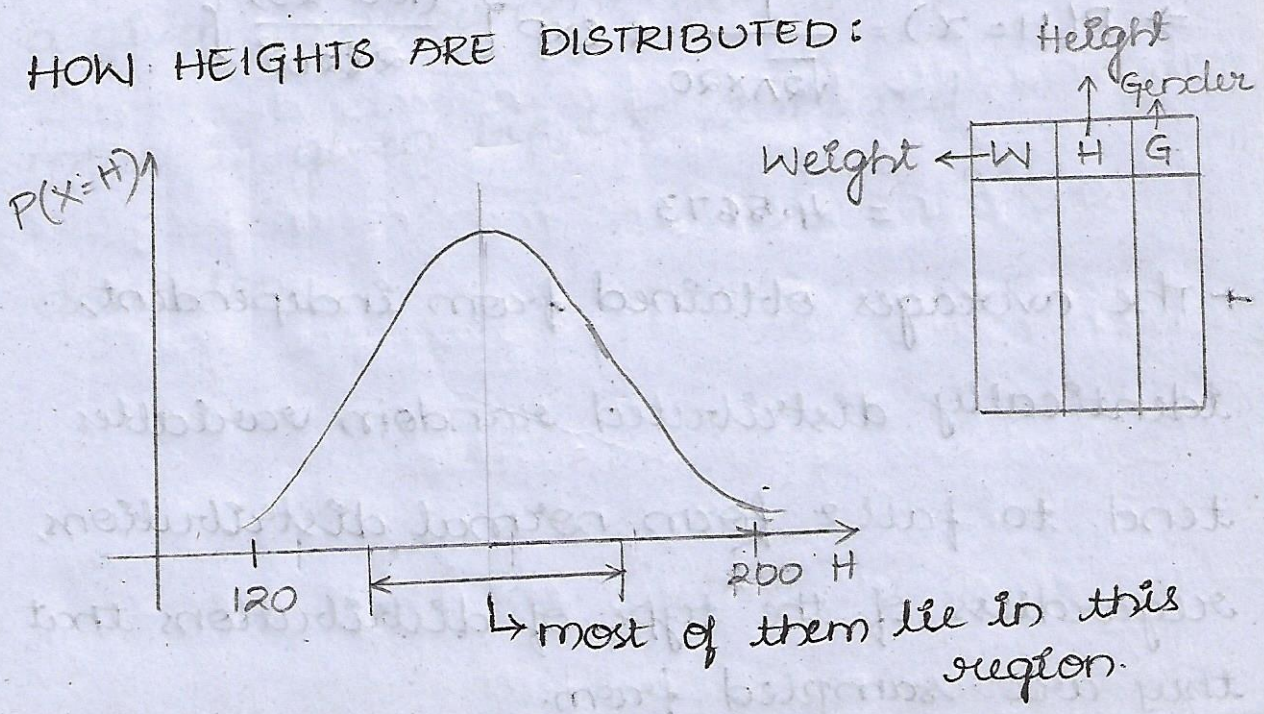
WHY THIS DISTRIBUTION?

→ Many machine learning algorithms works with the variables in the data (COLUMNS) follow the Normal Distribution.

→ In Machine learning, cost function or a neuron potential values are the quantities that are expected to be the sum of many independent processes (such as input features or activation potential of last layer) often have distributions that are nearly normal. One can continue to use parametric statistics knowing gaussian nature of data set.

→ Normal Distribution occurs in nature frequently.

HOW HEIGHTS ARE DISTRIBUTED:



↳ most of them lie in this region.

The Mathematical Representation:

$$H \sim N(\mu, \sigma)$$

Where, H - Height (Random variable)

~ - follows

N - Normal Distribution

μ - mean

σ - Standard deviation.

The Probability of Normal Distribution is

$$P(X = x) = \frac{1}{\sqrt{2\pi}\sigma} * exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

$\rightarrow H \sim N(160, 20)$

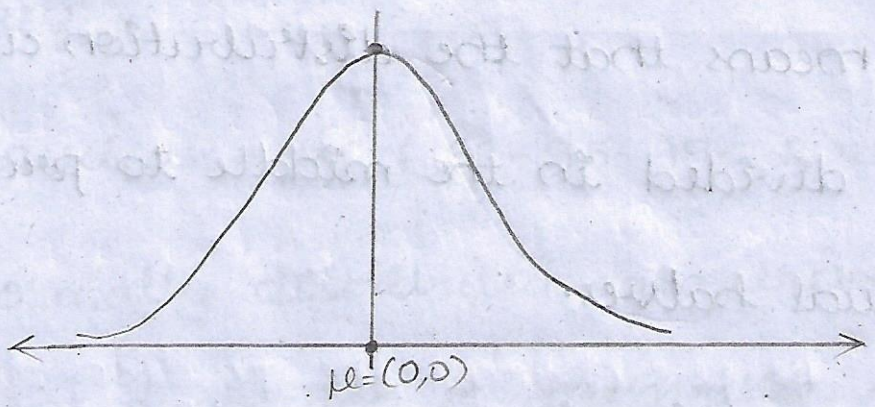
$$\Rightarrow P(H = x) = \frac{1}{\sqrt{2\pi} \times 20} * exp\left\{\frac{-(160-20)^2}{2(20)^2}\right\}$$

$$= 4.5673$$

→ The averages obtained from Independent, Identically distributed random variables tend to fall & form normal distributions, regardless of the type of distributions that they are sampled from.

STANDARD NORMAL DISTRIBUTION:

If the distribution that occurs when a normal random variable has a "mean of zero and a standard deviation of one".



$$X \sim N(\mu = 0, \sigma = 1)$$

→ The standard normal distribution is centered at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation.

PROPERTIES OF NORMAL DISTRIBUTION:

1. It is symmetric :

A normal distribution comes with a perfectly symmetrical shape.

→ This means that the distribution curve can be divided in the middle to produce two equal halves.

→ The symmetric shape occurs when one-half of the observations fall on each side of the curve.

2. The Mean, Median and Mode are equal:

The middle point of a normal distribution is the point with the maximum frequency which means that it possesses the most observations of the variable.

→ The midpoint is also the point where these three measures fall.

→ the measures are usually equal in a perfectly (normal) distribution.

3. It looks like a Bell-shaped curve.

4. $X \sim N(\mu, \sigma)$ → μ & σ can be found easily.

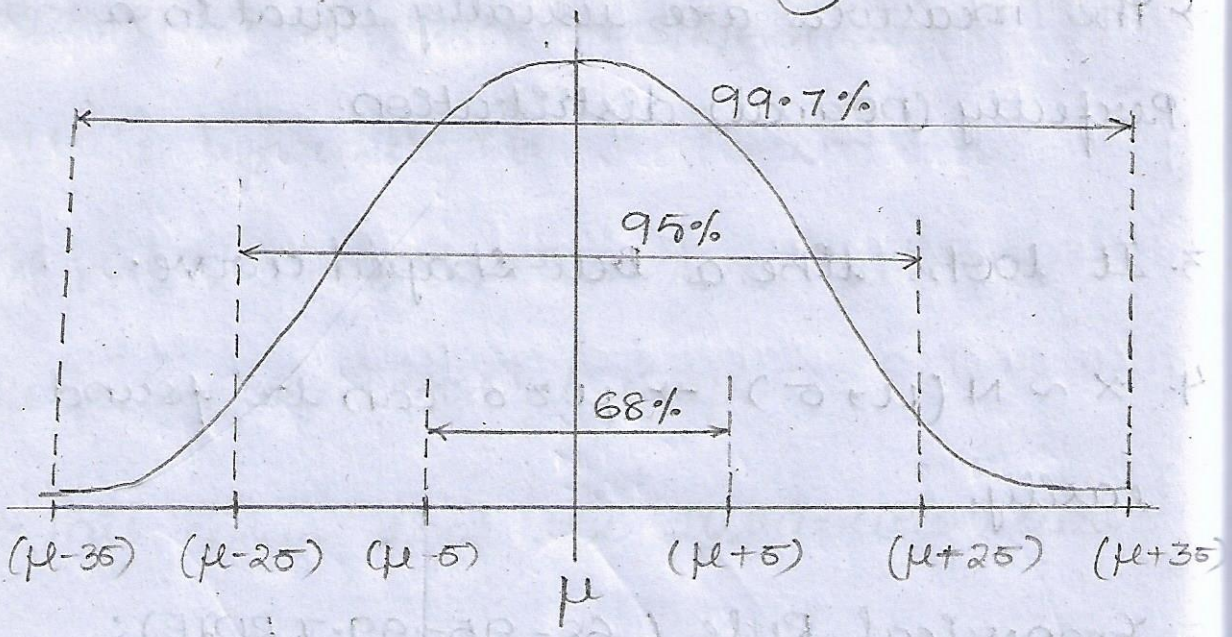
5. Emperical Rule (68-95-99.7 RULE):

In normally distributed data, there is a constant proportion of ~~distribution~~ of distance lying under the curve between the mean and specific number of standard deviations from the mean.

→ 68% of all the cases fall within $\pm$ 1 standard deviation from the mean.

→ 95% of all the cases fall within $\pm$ 2 standard deviations from the mean.

→ 99.7% of all the cases fall within $\pm$ 3 standard deviations from the mean.

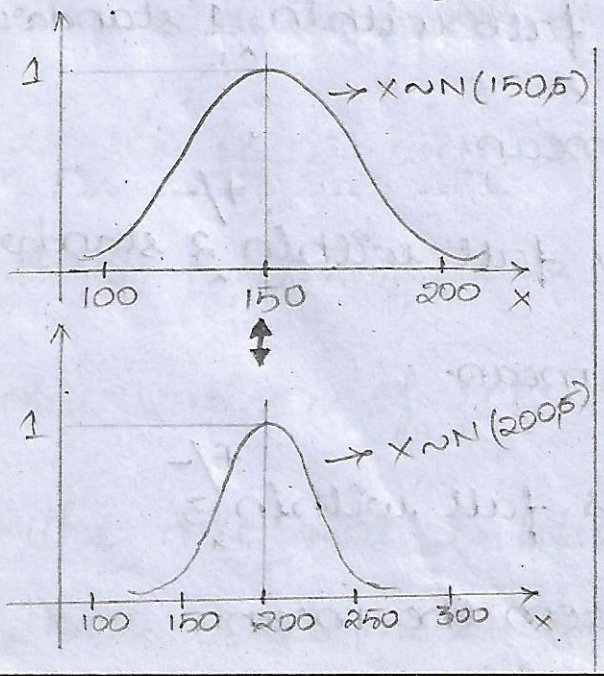


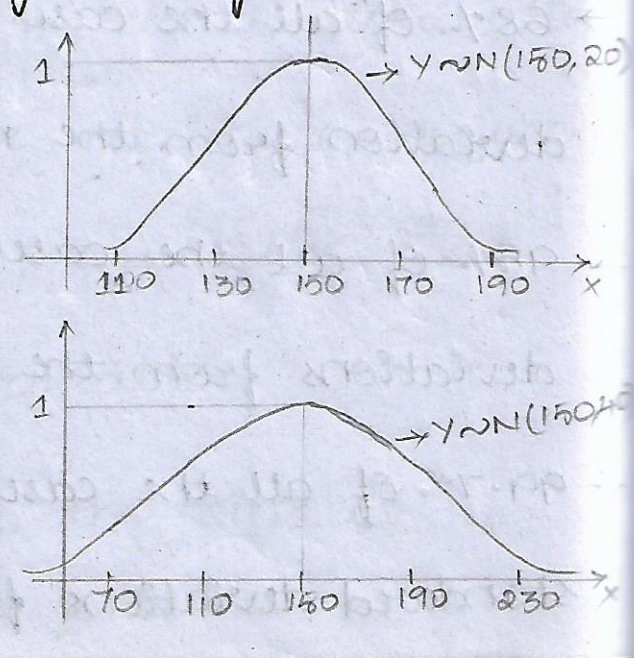
CASE - 1:

CHANGE IN μ : changes the position of the graph either in '+/-' direction where in the rest are constant.

CASE - 2:

CHANGE IN σ : Spread gets Increased as the values get changed.





6. Skewness and Kurtosis:

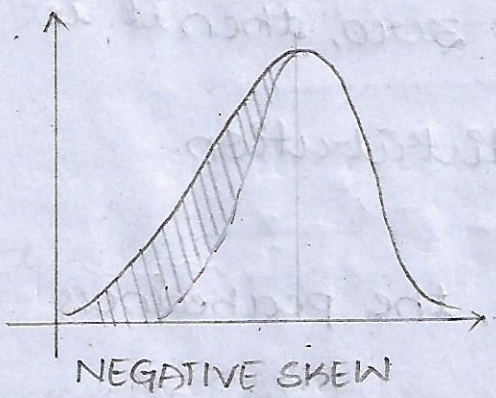
Skewness and kurtosis are coefficients that measures how different a distribution is ~~for~~ from a normal distribution.

SKEWNESS:

Skewness is the measures of the symmetry of a normal distribution.

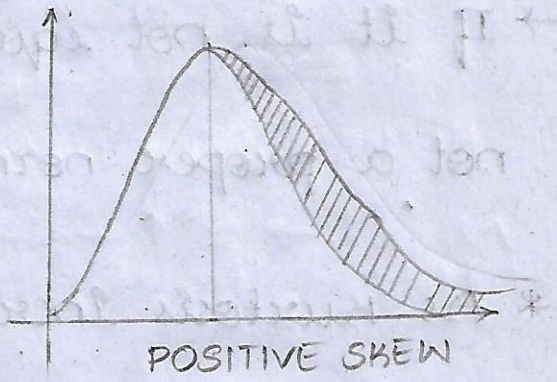
KURTOSIS:

Kurtosis measures the thickness of the tail ends relative to the tails of a normal distribution.



↓

Negative skew is due to the outliers

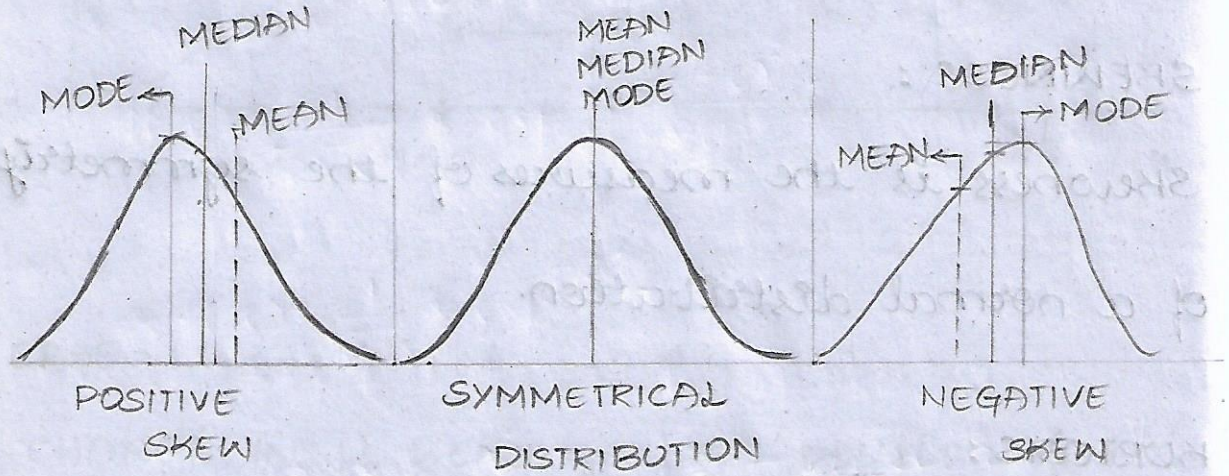


↓

Positive skew is due to the outliers.

* OUTLIER'S ATTRACT MEAN.

A General relationship of mean and median under differently skewed unimodal distribution:



NOTE:

→ The kurtosis value is exactly equal to zero in a normal distribution.

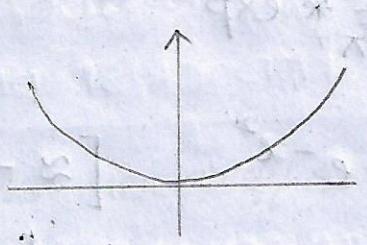
→ If it is not equal to zero, then it is not a proper normal distribution.

* As kurtosis increases, the peakedness in a distribution increases and viceversa

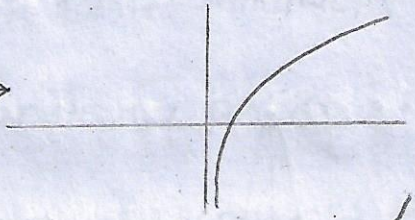
Why does the Normal Distribution has a bell-shaped curve?

→ As most of the continuous data values in a normal distribution tend to cluster around the mean and fwether a value is from the mean, the less likely it is occured.

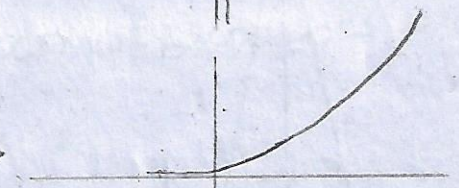
$f(x) = x^2 \longrightarrow$



$f(x) = \log x \longrightarrow$

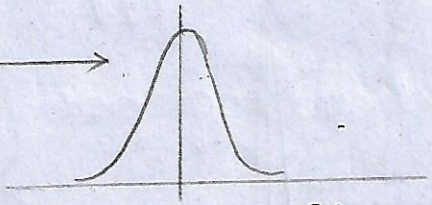


$f(x) = e^{+x} \longrightarrow$



$f(x) = e^{-x} \longrightarrow$

$f(x) = 1/x^2 \text{ (OR) } x^{-2} \longrightarrow$



$$\Rightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma} * \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

Probability distribution of a random variable is stated above.

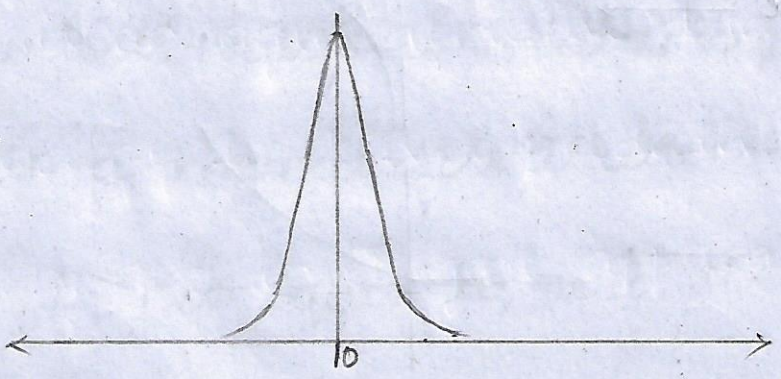
The Standard ~~deviation~~ Normal Distribution:

$X \sim N(0,1)$ then

$$f(x) = \frac{1}{\sqrt{2\pi} \times 1} * \exp\left\{\frac{-(x-0)^2}{2(1)^2}\right\}$$

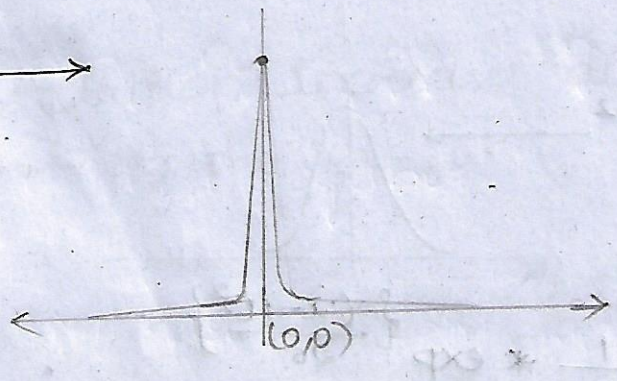
$$\Rightarrow f(x) = \frac{1}{\sqrt{2\pi}} * \exp^{-x^2/2}$$

$$\Rightarrow f(x) \approx e^{-x^2} \quad [\approx - \text{equivalent} \not\equiv$$



As $x \uparrow \rightarrow x^2 \uparrow \rightarrow -x^2 \downarrow \rightarrow e^{-x^2} = ?$

$e^{-x^2} \longrightarrow$



DATA ANALYTICS FRAMEWORK:

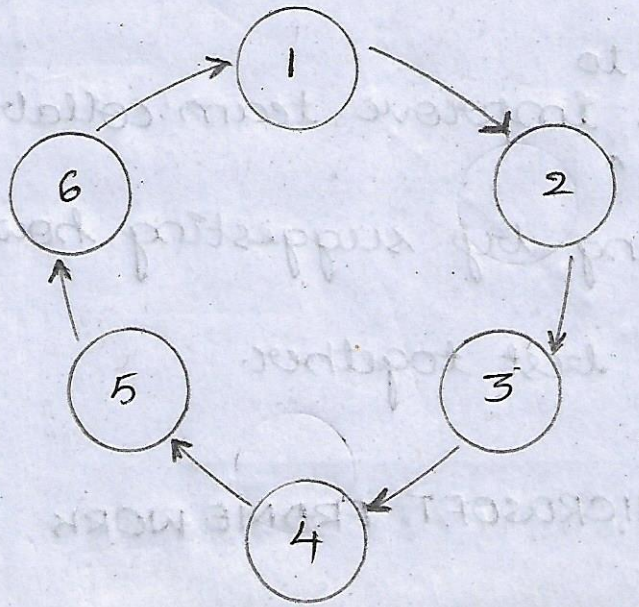
It empowers business entrepreneurs to analyze huge data.

→ Without using these advanced tools and frameworks, it becomes quite difficult for business to find conventional analytics and intelligence solutions for their business.

1. CRISP-DM : CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING.

→ It is open standard process model that describes common approaches used by data mining experts.

→ It is most widely-used analytics model.



1. BUSINESS UNDERSTANDING
2. DATA UNDERSTANDING
3. DATA PREPARATION
4. MODELLING → Machine learning
5. EVALUATION
6. DEPLOYMENT → Flask, Devops by AWS CLOUD

> NOTE:
> NO GUARATEE FOR CRISP-DM FRAMEWORK.

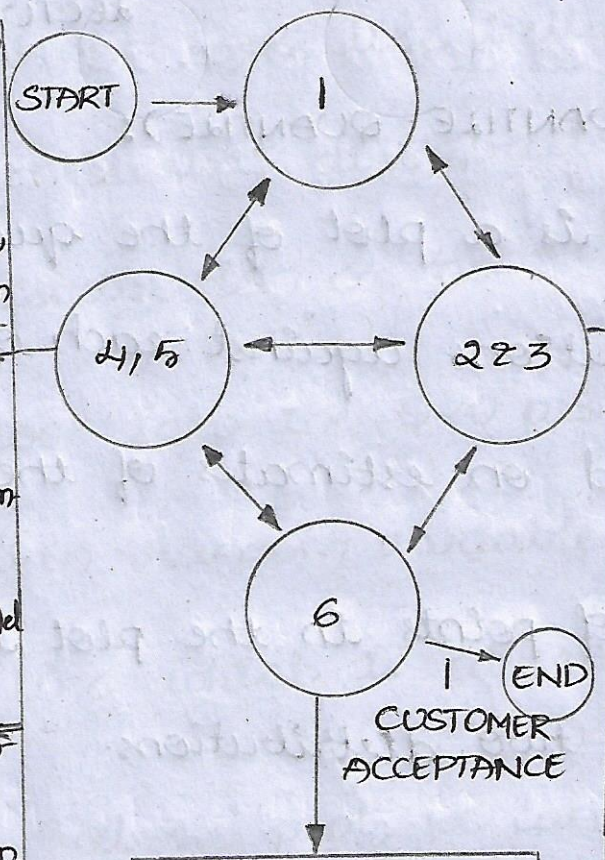
2. TDSP : TEAM DATA SCIENCE PROCESS

→ TDSP is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently.

→ TDSP helps to improve team collaboration and learning by suggesting how team roles work best together.

→ It is a MICROSOFT FRAME WORK

DATA SCIENCE LIFE CYCLE



FEATURE ENGINEERING:
Transform, Binning, Text, Image, Temporal, Feature selection

MODEL TRAINING:
Algorithms, Ensemble parameter tuning retraining model management

MODEL EVALUATION:
cross validation, model reporting A/B Testing.

→ DATA SOURCE:
On-Premises Vs Cloud Database Vs Files

→ PIPELINE:
Streaming Vs Batch Low Vs High frequency

→ ENVIRONMENT:
On-premises Vs cloud Database Vs Data Lake Vs... small Vs Medium Vs Big Data

→ WRANGLING, EXPLORATION & CLEANING:
structured Vs Unstructured Data validation & clean up visualization