Feature Engineering

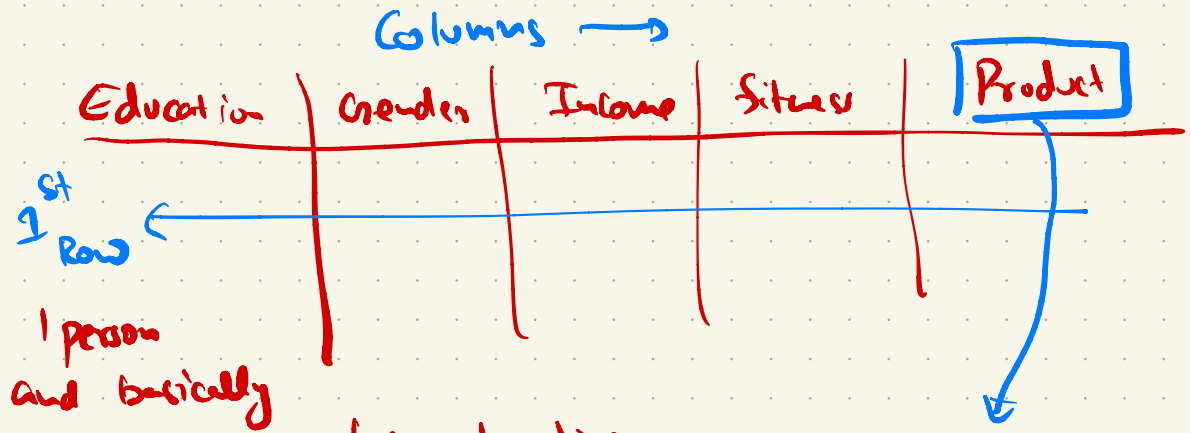
Agenda
1) FE
2) Skewness, Kurtosis
3) Univariate Analysis
   ↳ Converting Cont. data to discrete dat
4) Creating new feature

⇒ Feature Engineering →

Q What is feature ??

Variable, "attributes", "dimension" "Characteristics"

Aerofit Case Study



Rows → Records

Columns → features** in data Science

fields

features are used to predict Target.

Q I want to know whether a person is fit or not?

① understanding fitness

⇒ Small survey → weight & Height.

100 people

Weight	Height	Fit
		0
		1
		0
		1
		0
		1

Domain Expert are generally asked to label the data

Q Is only weight enough?? partially

Q Is only height enough??

⇒ Combination of feature → Can be a good indicator.

Some single feature → Can still be a good indicator

Some single feature → might just be useless

⇒ BMI ⇒ Body Mass Index

⇒ $\frac{W}{H^2}$

New feature are created to simplify/improve the prediction.

W	H	BMI	Fit
x	x	x	
x	x	x	
x	x	x	
x	x	x	

⇒ Domain knowledge → Required to create certain new features.

⇒ Can our Machine Create relevent feature themself. "Machine learning".

⇒ Loan Worthiness Chart →

⇒ Gender, Income, Married, ... → feature

Loan Status → Approved or Not ⇒ Target

Variable Name	Description	Sample Data
Loan_ID	Loan reference number (unique ID)	LP001002; LP001003; ...
Gender	Applicant gender (Male or Female)	Male; Female
Married	Applicant marital status (Married or not married)	Married; Not Married
Dependents	Number of family members	0; 1; 2; 3+
Education	Applicant education/qualification (graduate or not graduate)	Graduate; Under Graduate
Self_Employed	Applicant employment status (yes for self-employed, no for employed/others)	Yes; No
ApplicantIncome	Applicant's monthly salary/income	5849; 4583; ...
CoapplicantIncome	Additional applicant's monthly salary/income	1508; 2358; ...
LoanAmount	Loan amount	128; 66; ...
Loan_Amount_Term	The loan's repayment period (in days)	360; 120; ...
Credit_History	Records of previous credit history (0: bad credit history, 1: good credit history)	0; 1
Property_Area	The location of property (Rural/Semiurban/Urban)	Rural; Semiurban; Urban
Loan_Status	Status of loan (Y: accepted, N: not accepted)	Y; N

⇒ Skewed → log normal.

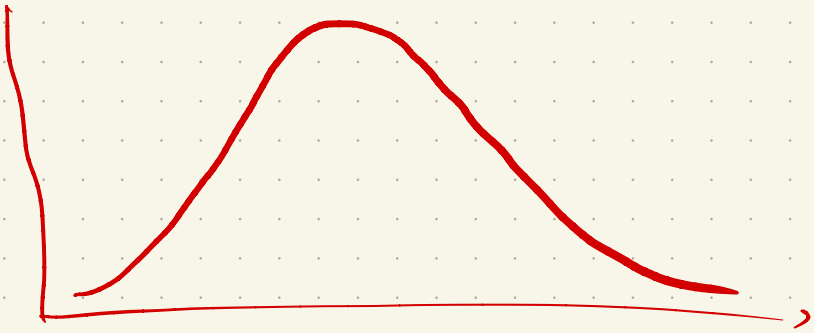
Is the data Skewed ??

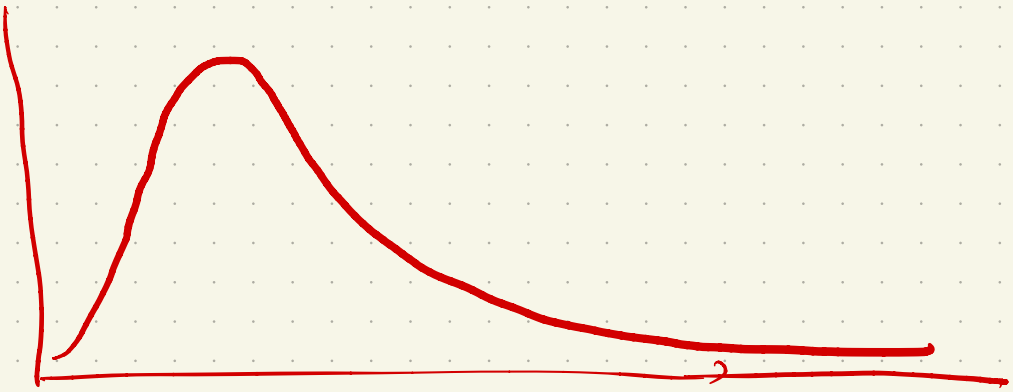
Why this is important ?

① Understand the distribution →

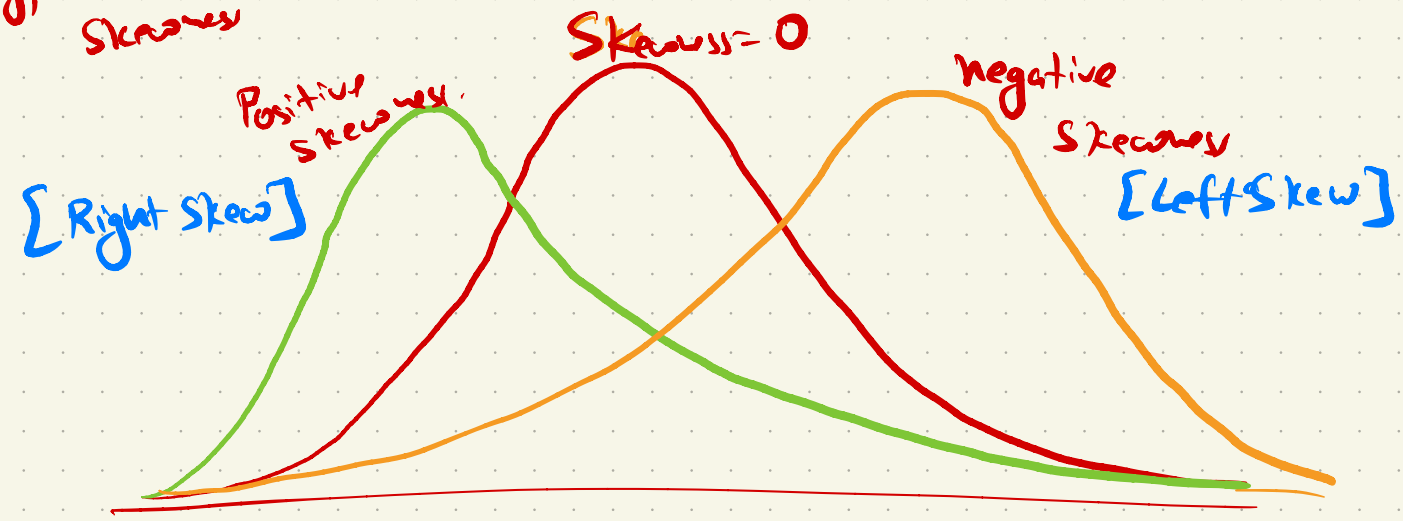
② Skewness impact loan approval (Target)

③ Outlier Understanding.



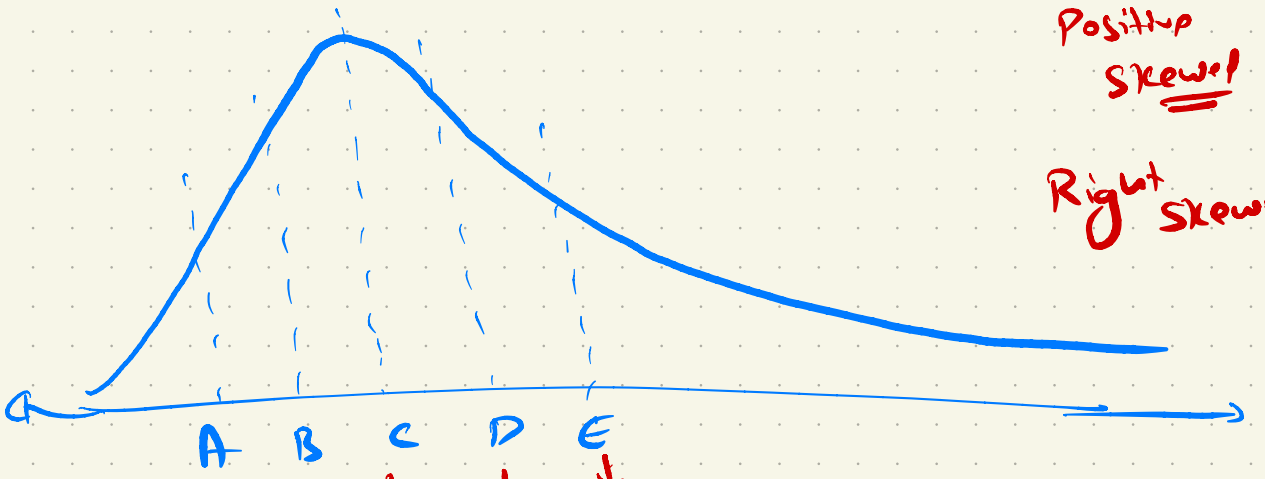


Types of Skewness



There are multiple maths formula for skewness →

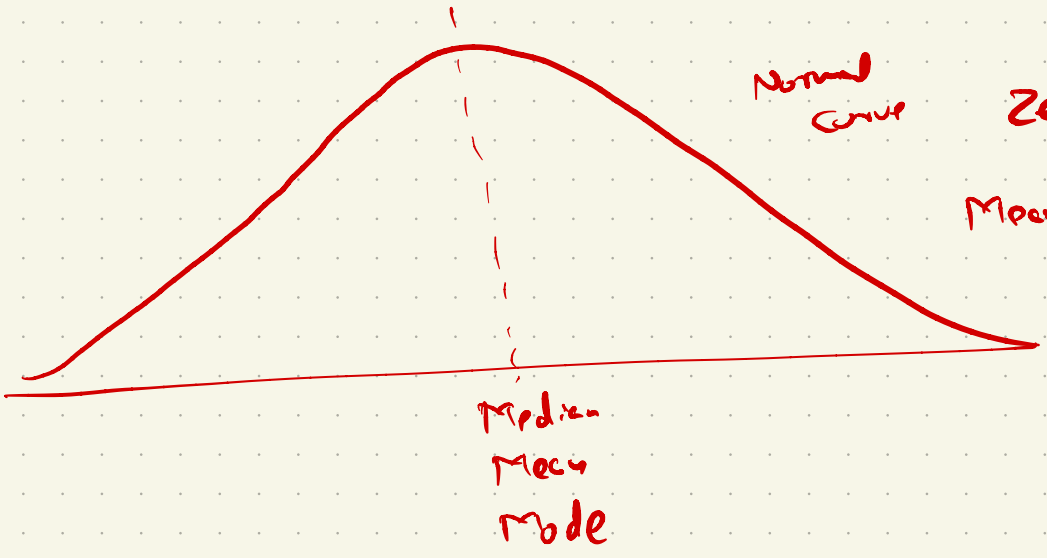
$$\text{Skew} = \frac{3(\text{mean} - \text{mod})}{\sigma}$$



Positive Skewed

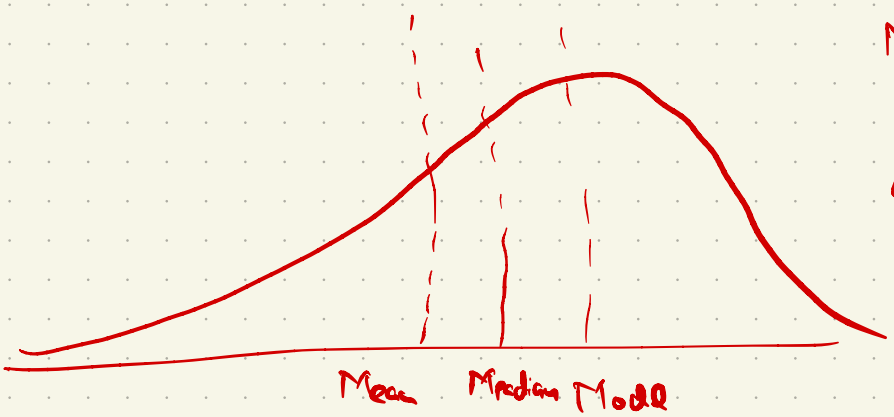
Right Skewed

Right Skew → Mean is on right of Median.



Normal Curve

Zero Skewed



Negative Skewed

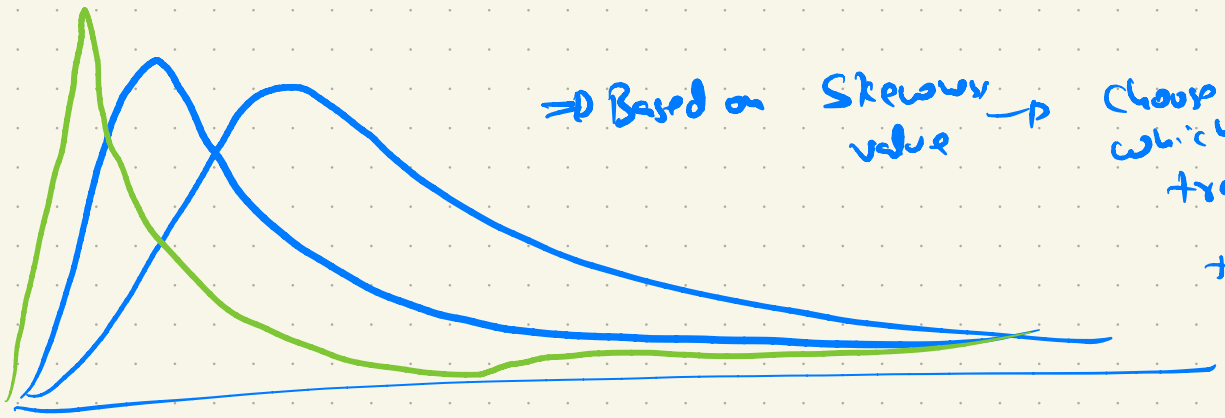
Left Skewed

Skewness → tells us the shape of the distribution

μ → mean

σ → Spread of dist

Skewness → Shape of distribution.



→ Based on Skewness value → Choose which transformation to apply

Skewness → ① Data preparation → transformation

② Decision making can be improved.

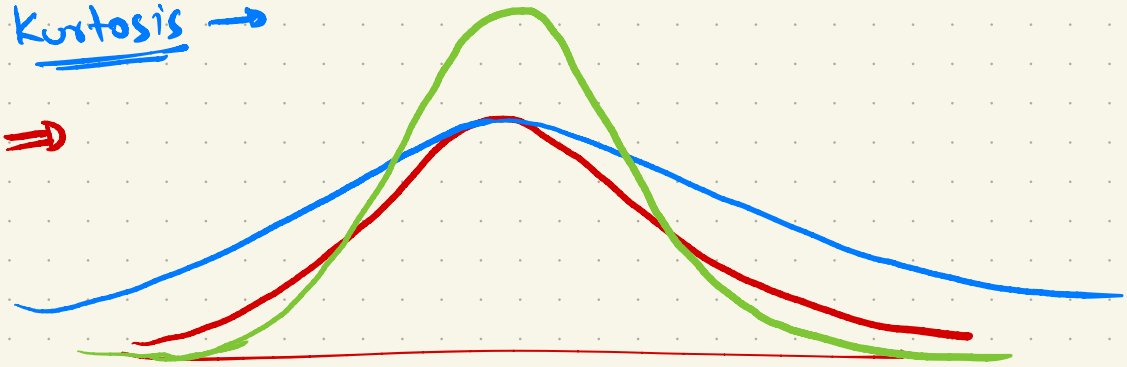
③ Understand data & patterns.

Q → mean → Central tendency

σ → Spread

Skew → Shape of data.

Kurtosis →



⇒ σ → Spread of data is.

⇒ Kurtosis → Tells us how peaked the data is.

↳ helps understanding data
- Outlier detection

⇒ Pointy → More Concentrated → Lepto kurtic
Normal → → Meso kurtic
Flat → → Platy kurtic

⇒ Skewness + kurtosis → Interview Topics.

1
Still used if it is very extreme.

⇒ Univariate Analysis →

⇒ Analyse impact of 1 variable against target.

Total income → Loan

A Salary 20L
Loan 50L →
Loan Term 1yr

B Salary 10L Loan 50L →
Term 20 yrs

⇒ Applicant should be able to pay EMI

⇒ EMI → Total loan Amount / Loan Year × 12

⇒ Domain knowledge → income impact Loan Status

→ we checked income → didn't see result ✗

→ we convert it to total income → ✗

→ Is income enough to pay the loan → EMI, able to pay

→ able to pay to Loan Status → ✓

You are working on a dataset for predicting house prices. You have a feature 'YearBuilt' representing the year a house was built.
It does not contain any missing values. To improve the model's performance, what feature engineering technique can you apply to 'YearBuilt'?

⇒ • Feature Engineering

• Skewness + kurtosis → to understand data

• Univariate analysis
  - Charts
  - Test to validate
  - Binning
  - Create new feature

• FE is an art →

• The more you practice → the better your model becomes

Collab Link : https://colab.research.google.com/drive/1kLzjBZWRJS_drg98Piv8RZ4Hvd7JBkBv?usp=sharing

This blog has pretty simple but good starter problems and their solutions
https://kindsonthegenius.com/blog/hypothesis-testing-solved-examplesquestions-and-solutions/
You can also refer to Ch- 9 -11 of this book
https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf?_gl=1*1nfheaw*_ga*MTc0ODY0ODkwNy4xNzAyMzE2ODk4*_ga_T746F8B0QC*MTcwMjMxNjg5OC4xLjEuMTcwMjMxNzM4Ni41NS4wLjA.