① Primary - Pivot, Group by
② Desc + Inf Analytics ↓ (Stats Basic)
③ Bridge to ML

ML ready
① Missing Values
② Outlier → ① Identify ② Treat
③ Transformation (transform to Normal) (makes it analytics)

Data Types
measure
Numeric - int, float ...
String - Categorical
→ Dimension
Date ← Dimension (Y, M, Q, W, Day)
Numerical
D1, D2, D3
- time

Case Study

LOAN Prediction

1000
Life Insurance
Personal Loan

→ 100
Home Insurance
Housing Loans

Q1 → target

① Binary Classification
- Log Reg
- DT class
- RT class

(1000) All → Q1 (100) → (Y)
Home loan
1 Y 0.9
2 N 0.3
3 Y 0.8
4 N 0.1
5
6
.
.
1000 N 0.2

0 - 1 (→ Y/N)
not more

≥ 0.5 - Y
< 0.5 - N
Cut
→ Binary Classification

Demographics.
Age, Value,
Loans, status
Eligibility.
Banking
Dependents - Co Applic
Credit Score

614 → 122 (Y) → House loan
Existing customer

Risky - Lender

A | B
① with tolls - Highway
km/hr = 350 Km → 3 Hrs
② without tolls - village
= 250K. → 4 hrs.
km/hr
Cost
Cost - time
time - Cost

↓ Reduce time of Petrol cost.
↑ increase time as cost.

Categorical Inference

Count of Loan_ID		Loan_Status				
Education	Gender	N	Y	Grand Total		
Graduate	Female	31	61	92	GraduateFemale	66.3%
Graduate	Male	105	271	376	GraduateMale	72.1%
Graduate	(blank)	4	8	12	Graduate(blank)	66.7%
Not Graduate	Female	6	14	20	Not GraduateFemale	70.0%
Not Graduate	Male	45	68	113	Not GraduateMale	60.2%
Not Graduate	(blank)	1	0	1	Not Graduate(blank)	0.0%
Grand Total		192	422	614	Grand Total	68.7%

322/422

→ 11:38 - Be Back
11:38

Missing Values

1) Identify - Count Blank
2) a) Exclude - loose info
   b) Impute
      - Categorical → Mode, Constant (missing)
      - Numerical → mean – Normal, median – skewed, other (advanced)

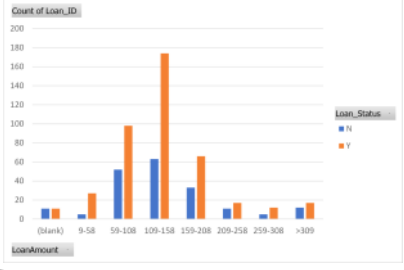
mean, median, mode

(Normal) (Not Normal)

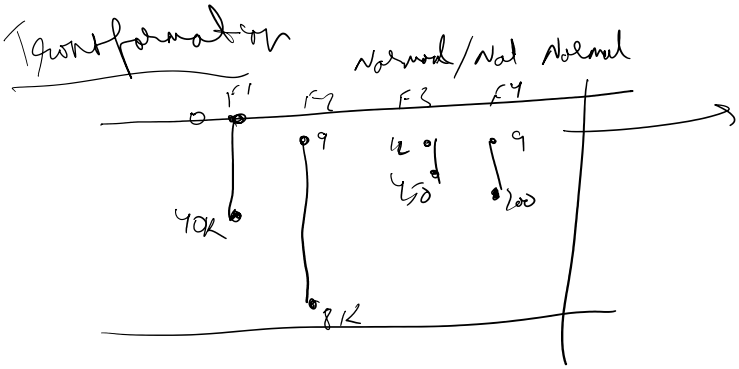
Outliers - Numerical

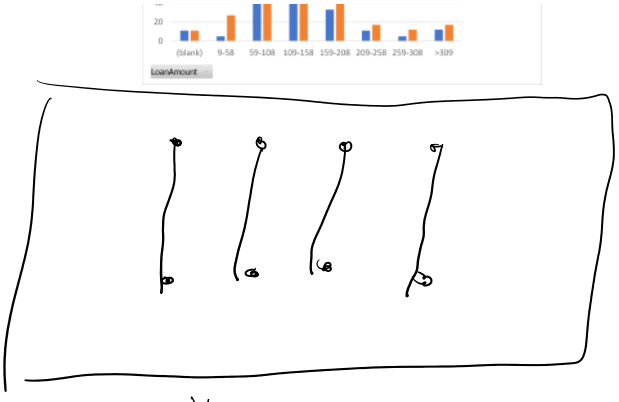
>450 → Replace 450

<5 → 5



Transformation





log →

2) min-max

3) standard