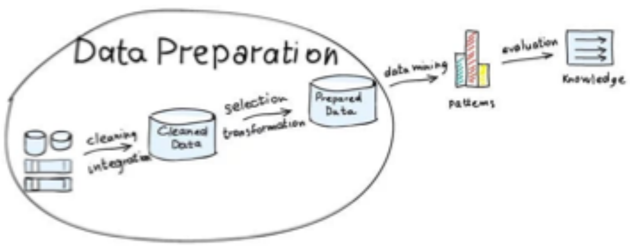


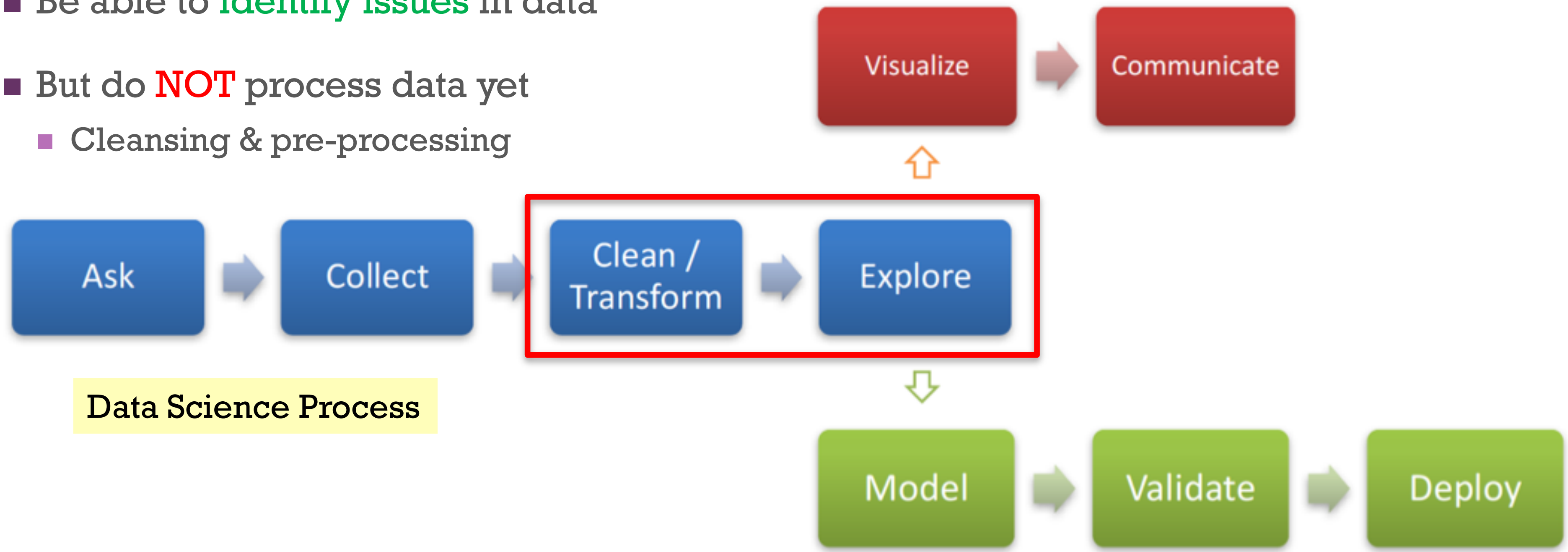
Data Preparation with Python

2110446: Data Science and Data Engineering

Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th

Previous class

- Be able to explore data
- Be able to identify issues in data
- But do NOT process data yet
  - Cleansing & pre-processing



Data Science Process

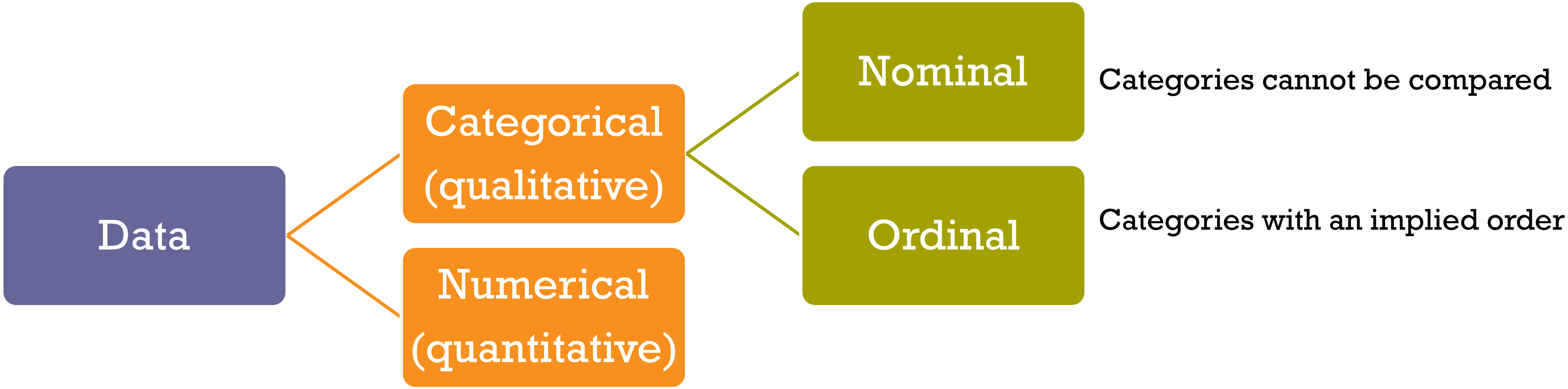
+ Terminology: Data table

inputs target

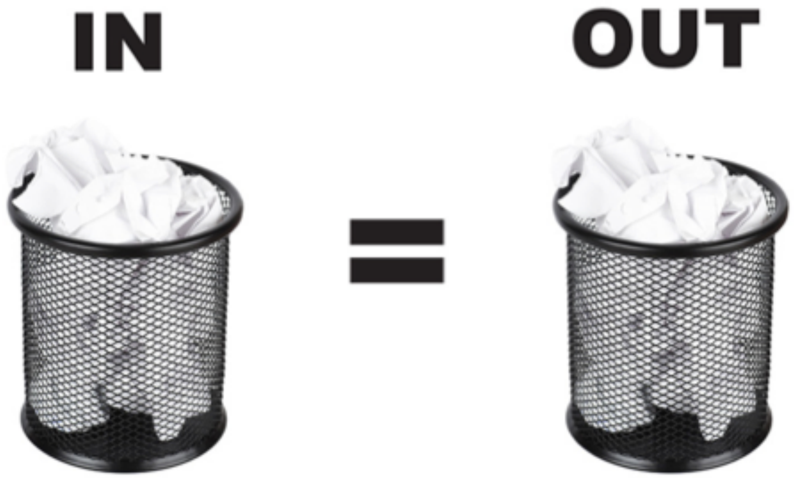
Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

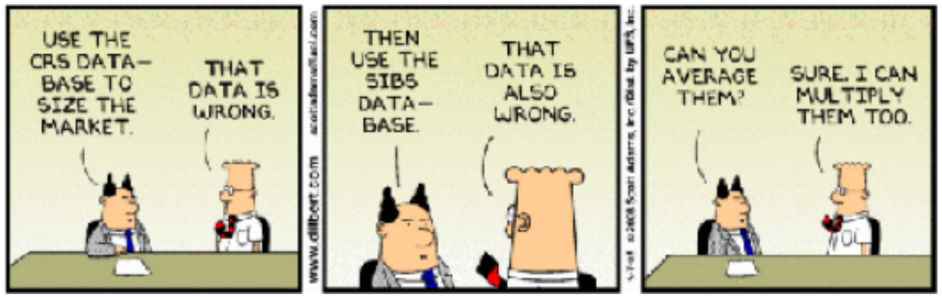
- Row
  - Example, instance, case, observation, subject
- Column
  - Feature, variable, attribute
- Input
  - Predictor, independent, explanatory variable
- Target
  - Output, outcome, response, dependent variable

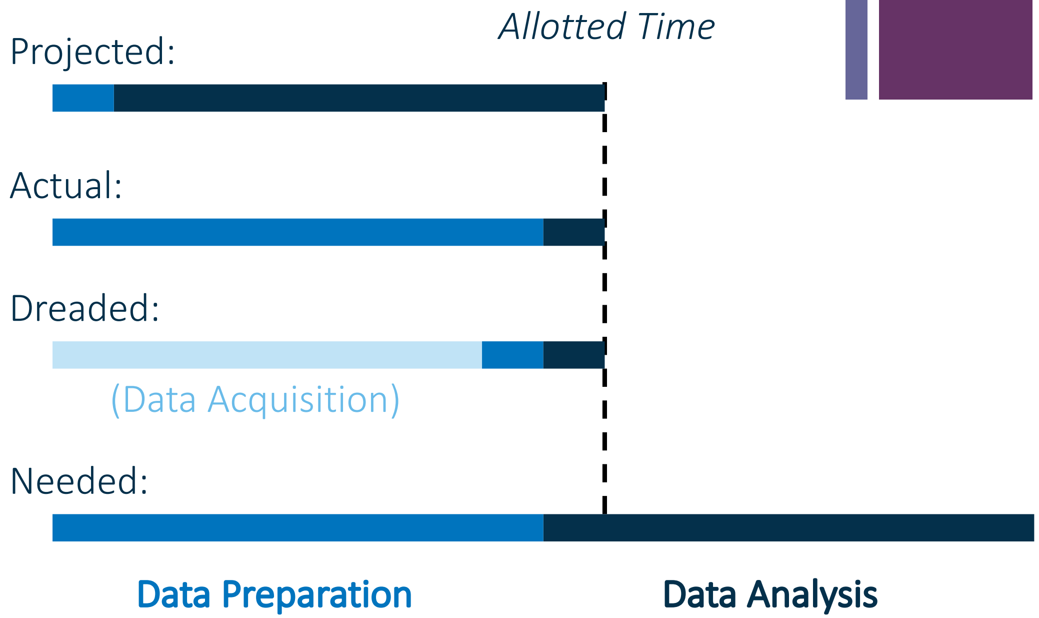
Terminology: Kinds of data



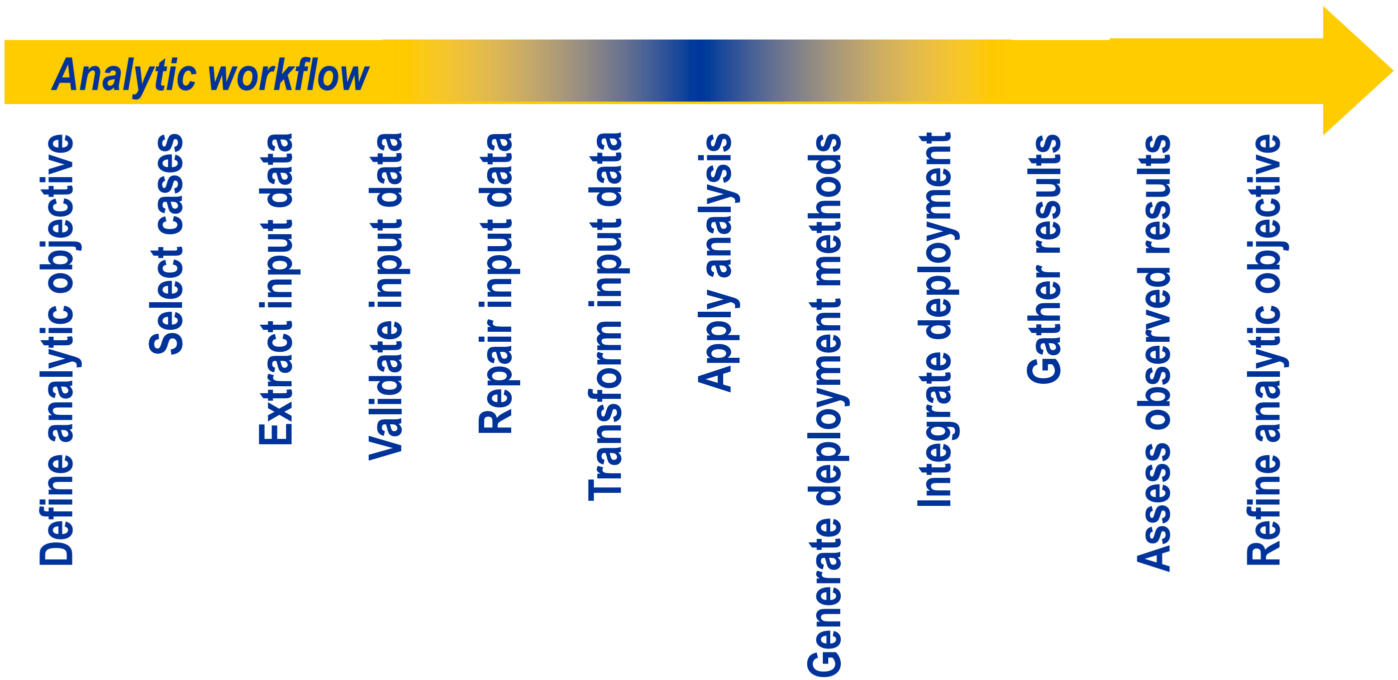
+ Data preparation is very important!



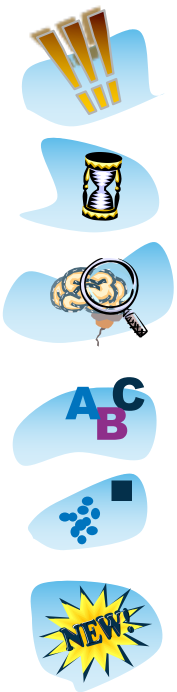




Analytics workflow



+ Data preparation challenges



- Massive data sets
- Temporal infidelity
- Transaction and event data
- Non-numeric data
- Exceptional, extreme, and missing values
- Stationarity

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$


Practice is everything.
Periander
quotefancy

DATAQUEST BLOG

28 DECEMBER 2016 / DATA CLEANING

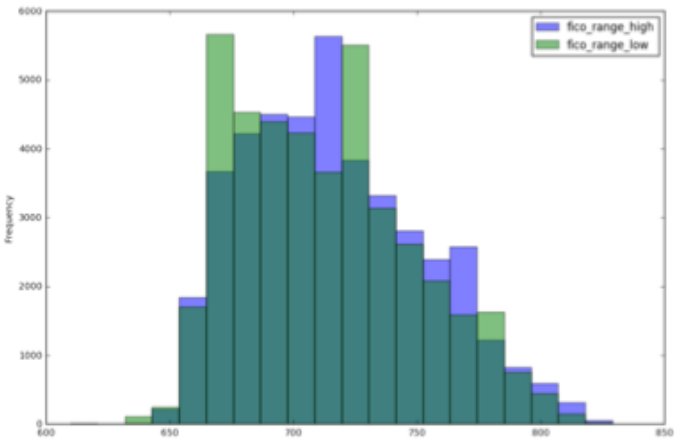
Preparing and Cleaning Data for Machine Learning

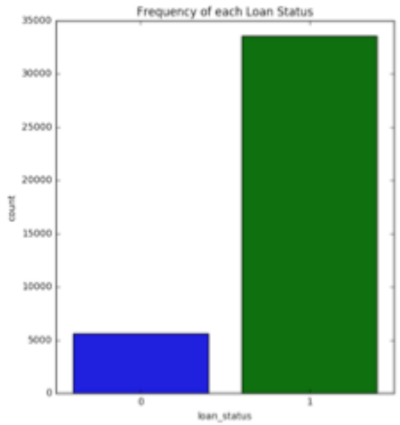
- 1) Examining the Data Set
- 2) Narrowing down columns manually
  - Remove Id's
  - Irrelevant variables
  - Remove zipcode & date
  - Temporal infidelity (data from future)
  - Calculated variables
  - Decide target
    - Select studied cases
    - Distribution of target variables
  - Remove flat values
- 3) Preparing features for ML
  - Preview data
  - Handling missing values
    - Drop unqualified features
  - Investigate categorical features
    - Drop too many unique values (treat as Id)
    - Convert ordinal to numeric
    - Convert categorical to numeric
  - Check all numeric variables

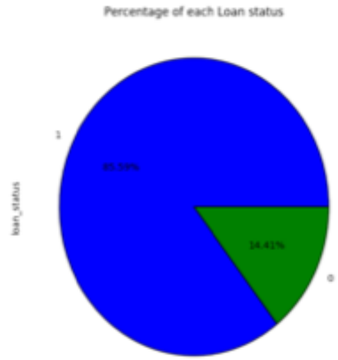
https://www.dataquest.io/blog/machine-learning-preparing-data/

1) Examining the Data Set

- Numerical variables
  - Out of ranges
  - Distribution: histogram
- Categorical variables
  - Miscodes
  - Distribution: frequency table, bar chart
- Target variable
  - Understand proportion of each class: bar chart, pie chart







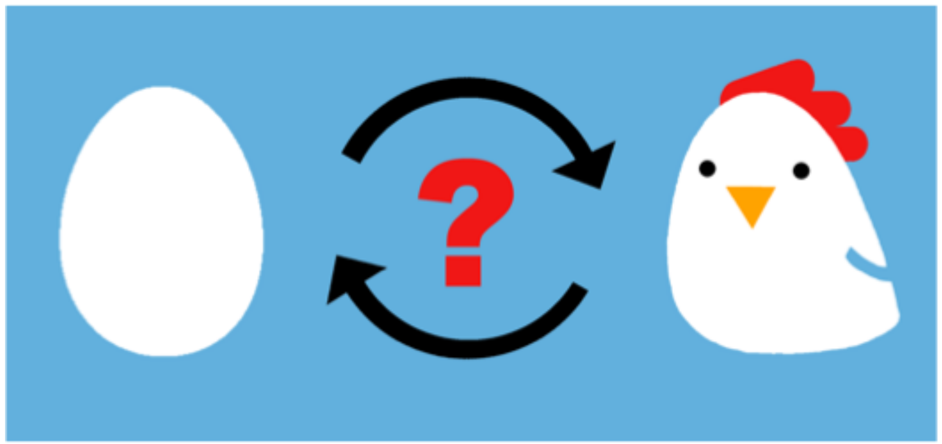
2) Narrowing down columns: Feature understanding is extremely important!

Remove irrelevant features <u>manually</u>

Domain expert

Inputs

Target



2) Narrowing down columns (cont.): Remove unqualified features

- Id's (lack of generalization; overfit)
- Variables with missing values > 50%
- Categorical variables
  - Too many unique values (treat as Id's)
  - Flat values (underfit)
  - Recode, consolidation (grouping)
- Special ways to treat these data
  - Zip code
    - Distance to closet branch
  - Date/time
    - Recency

2) Narrowing down columns (cont.): Temporal Infidelity

- Occurs when the input variables contain information that will be unavailable at the time that the prediction model is deployed.
- Assume that the model will be deployed in July-2017
  - Should we include a variable called "FICO2017", which is calculated at the end of the year?

3) Preparing features for ML (cont.): Impute missing values

? != 0

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

- Numerical variables:
  - Mean
  - Median
- Categorical variables:
  - Mode

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

3) Preparing features for ML (cont.): Categorical to numeric variables

- Ordinal variable
  - Enumerate

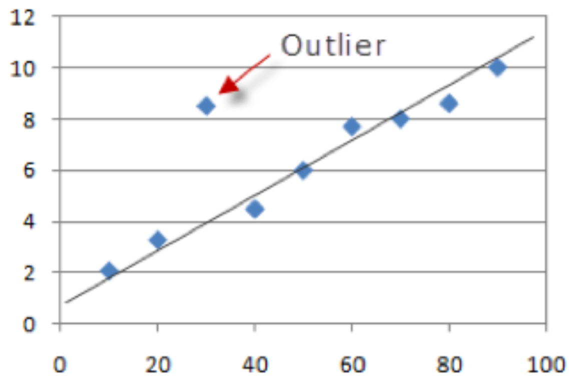
```
"emp_length": {
    "10+ years": 10,
    "9 years": 9,
    "8 years": 8,
    "7 years": 7,
    "6 years": 6,
    "5 years": 5,
    "4 years": 4,
    "3 years": 3,
    "2 years": 2,
    "1 year": 1,
    "< 1 year": 0,
    "n/a": 0
```

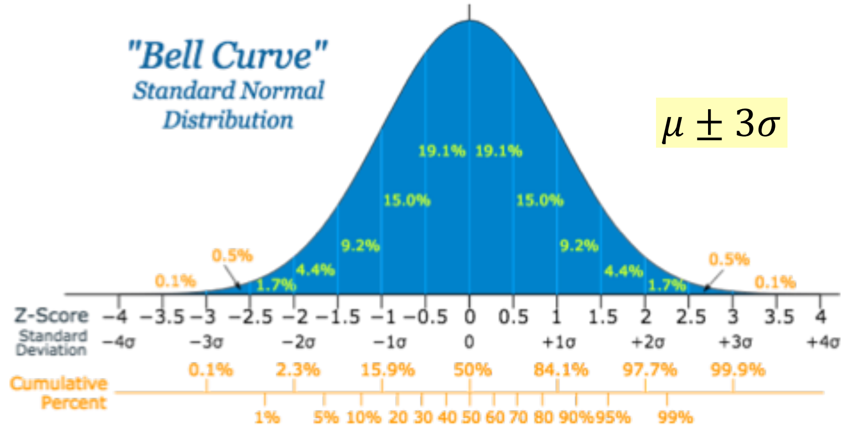
- Nominal variable
  - One-hot vector (dummy codes)
  - Smoothed weight of evidence (SWoE)

Level	D_A	D_B	D_C	D_D	D_E	D_F	D_G	D_H	D_I
A	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0
C	0	0	1	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	0	0	0	0
F	0	0	0	0	0	1	0	0	0
G	0	0	0	0	0	0	1	0	0
H	0	0	0	0	0	0	0	1	0
I	0	0	0	0	0	0	0	0	1

+

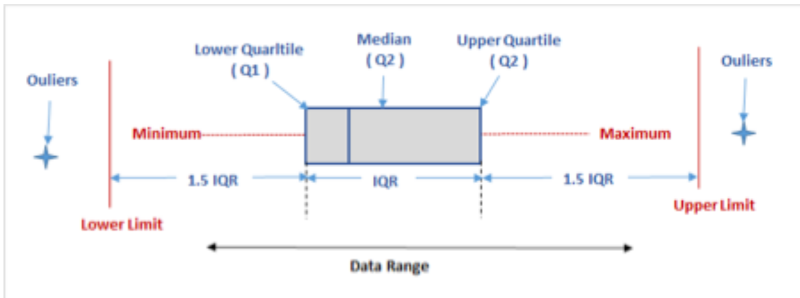
3) Preparing features for ML (cont.): Truncate outliers





$\mu \pm 3\sigma$

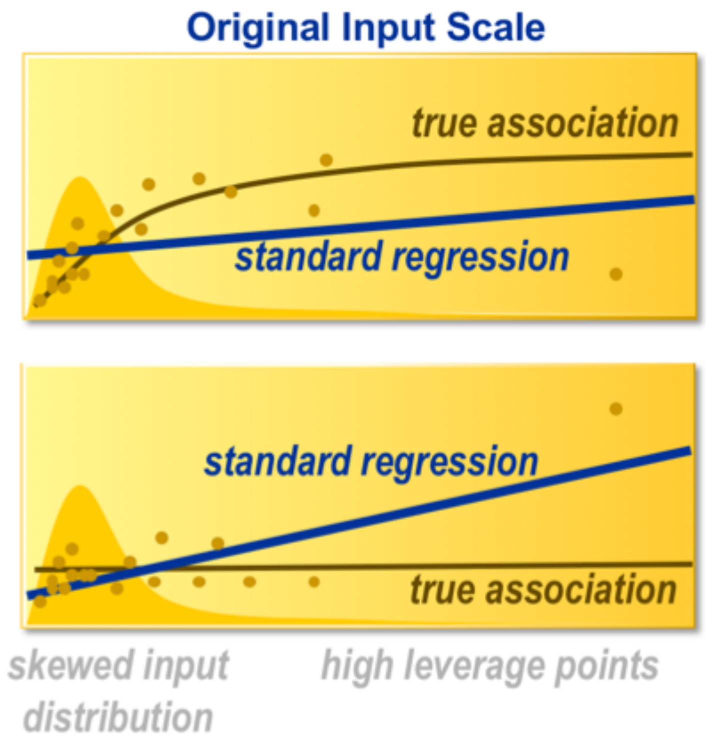
- Outlier, leverage points, extreme values



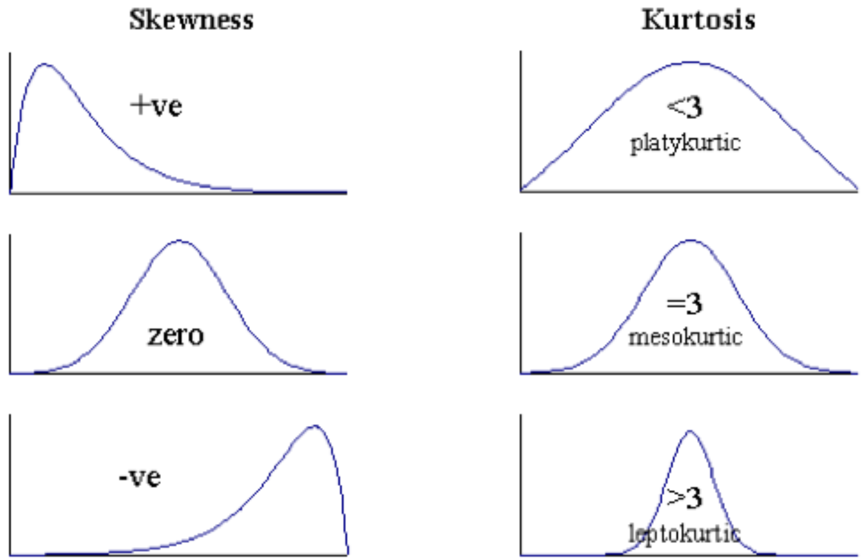
Percentile
1st
2.5th
5th
10th
25th
50th
75th
90th
95th
97.5th
99th

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

3) Preparing features for ML (cont.): Feature transformation

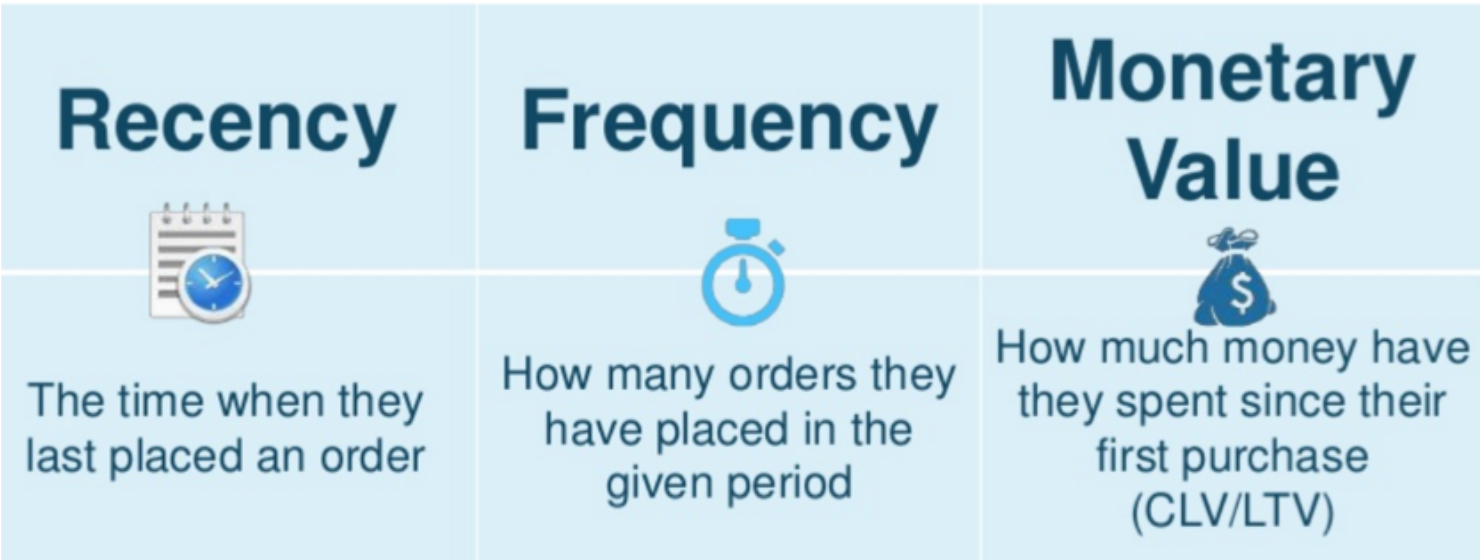


- Skewness
- Example: Salary, Balance in bank account
- **Solutions: Log, Binning**



3) Preparing features for ML (cont.): Feature engineering

- Feature engineering
  - Calculated variables
  - Behavior from transactional data (RFM/RFA)



4) Other preprocessing steps: Train/Test/Validate

Training Data

inputs

target

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

Validation Data

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes

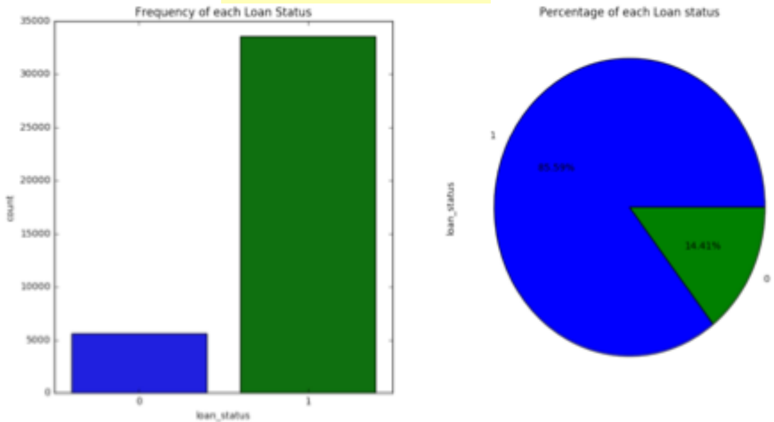
Testing Data

Age	Income	Gender	Province	Purchase
25	25,000	Female	Bangkok	?

4) Other preprocessing steps: Train/Test/Validate (cont.)

Simple random sample

Stratification



+ Other data preparation processes

- Impute missing values
- Outlier detections
- Feature transformation
  - Skewness
- Split train/test
  - Simple random sampling
  - Stratification
- Feature clustering
- Feature selection