Financial Data Science

Lecture 2
Data Preprocessing and Feature Engineering

Liu Peng
liupeng@smu.edu.sg

Why Do We Need to Preprocess Data?

$(x1, x2) \longrightarrow (s(x1), s(x2)) \xrightarrow{f} y$

Check Data Quality First

$\hat{y} = f(x1, x2) = f(s(x1), s(x2))$

Q: how to convert a categorical variable into a numeric one?

Categorical variable

one-hot encoding

cols: k - 1

k: # catg

- Missing value
- Label encoding A/B/C —> 1/2/3
- One-hot encoding A —> [1, 0, 0]
- …

Numeric variable

- Missing value
- Outlier
- Different scale
- …

x2 in [1, 100]

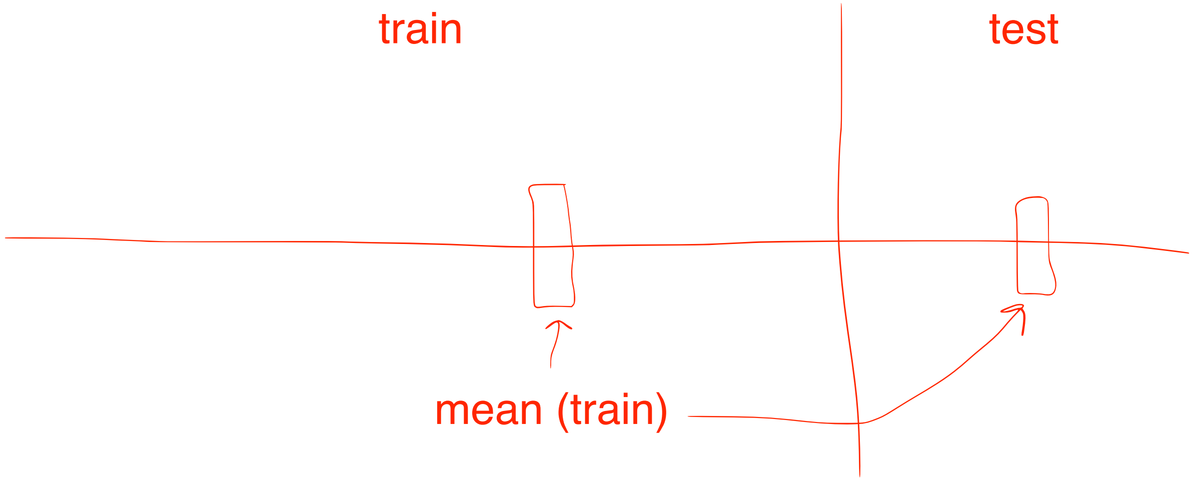
[1, 2, 3, 100]

↓ ↓

f 1.2

Dealing with Missing Data

- Deletion
- Imputation
- Train-test split
- Lookahead bias
- Indicator for missingness



x1	x2	ind
	3	1
2	4	0

In-class quiz

- Q1-3

More on Train-Test Split

- Estimate model performance on unseen data
- Detect overfitting and underfitting
- Ensure fair and unbiased evaluation
- Fix random seed for reproducibility
- Typical split ratio:
- Cross validation

In-class quiz

- Q4-6

Q: how many dimensions are encoded here? 3

Visualization

- Scatter plot
- Bar chart
  - Stacked
  - Side by side
  - Summarizing categorical variable by count or proportion
- Line chart
  - Time series plot
- Histogram/distribution
  - Summarizing numeric variable by histogram
- Boxplot

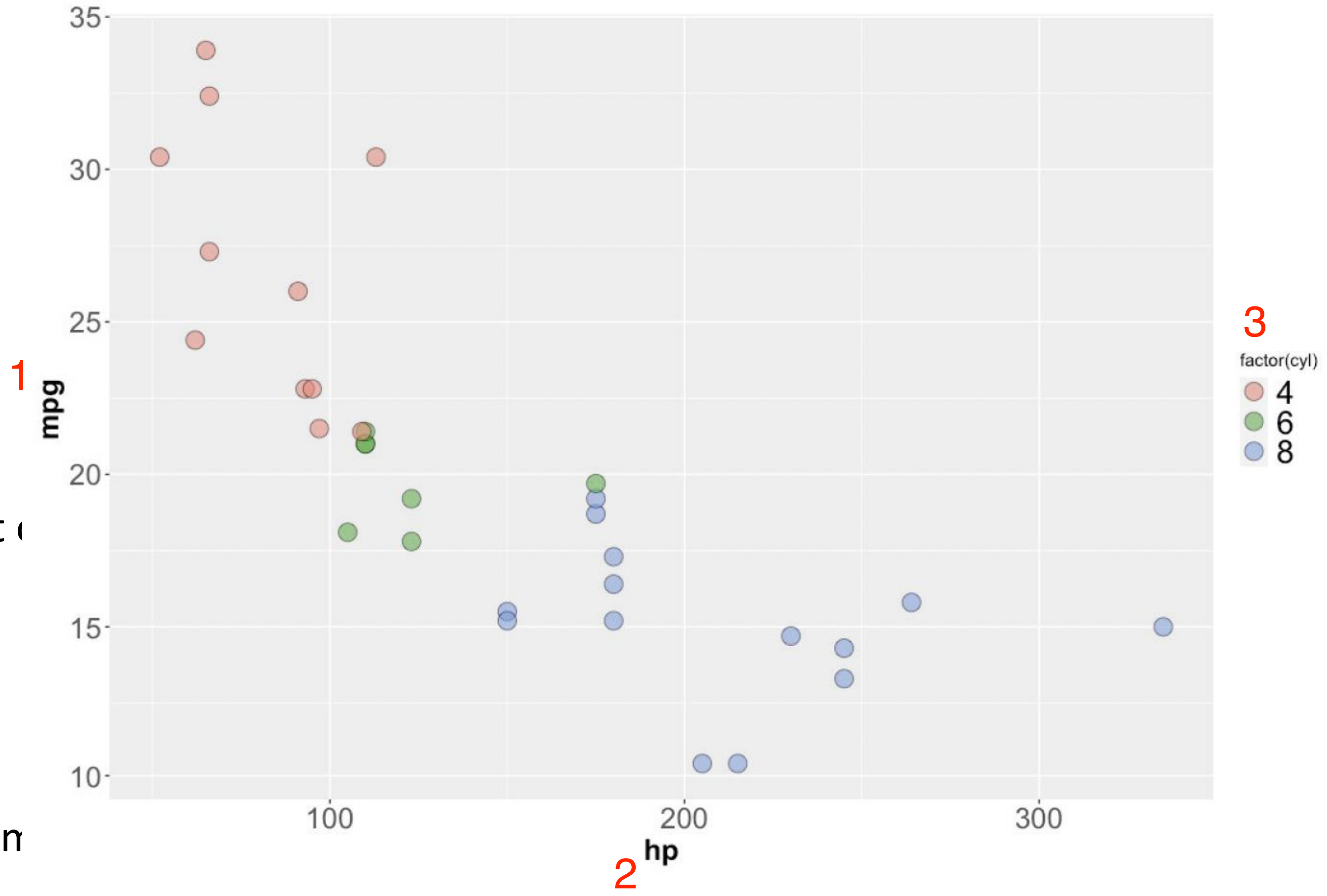


Figure 4.10 – Filling the inner color of the points in the scatter plot

Source: Chapter 4, The Statistics and Machine Learning with R Workshop

Using Distributions

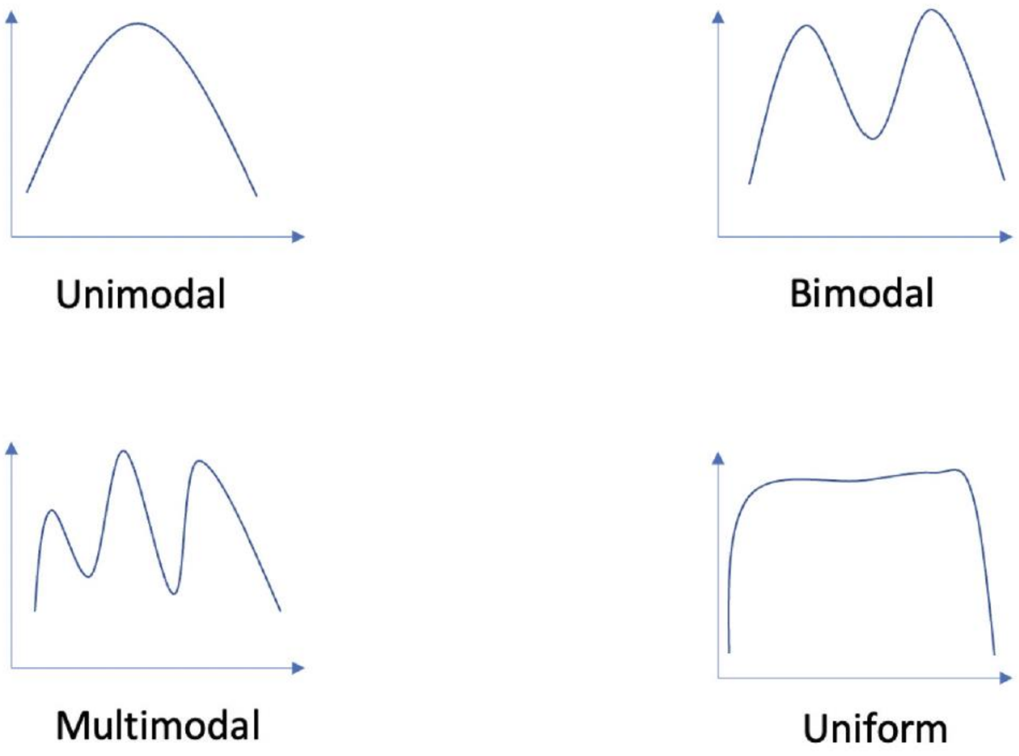


Figure 5.14 – Different types of modalities in a distribution

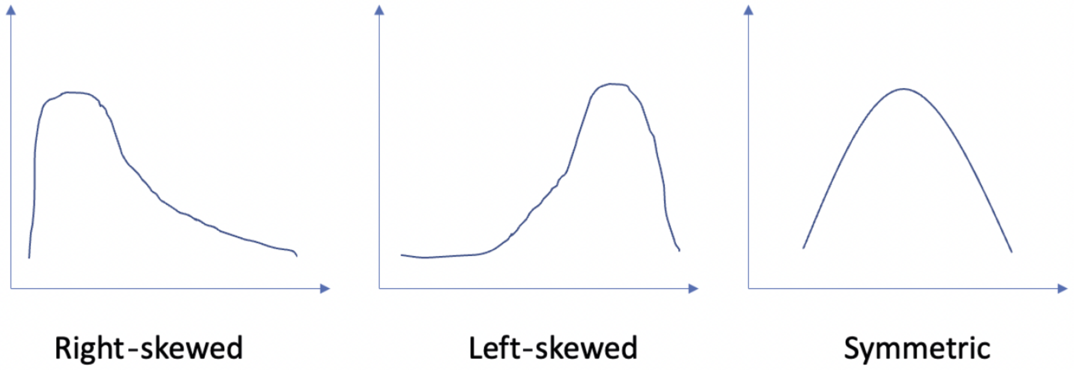


Figure 5.15 – Different types of skewness in a distribution

Source: Chapter 5, The Statistics and Machine Learning with R Workshop

Moving into Two Variables

- Categorical + Categorical: two-way frequency table; confusion matrix
- Categorical + Numeric: side-by-side boxplot
- Numeric + Numeric: scatter plot; correlation plot

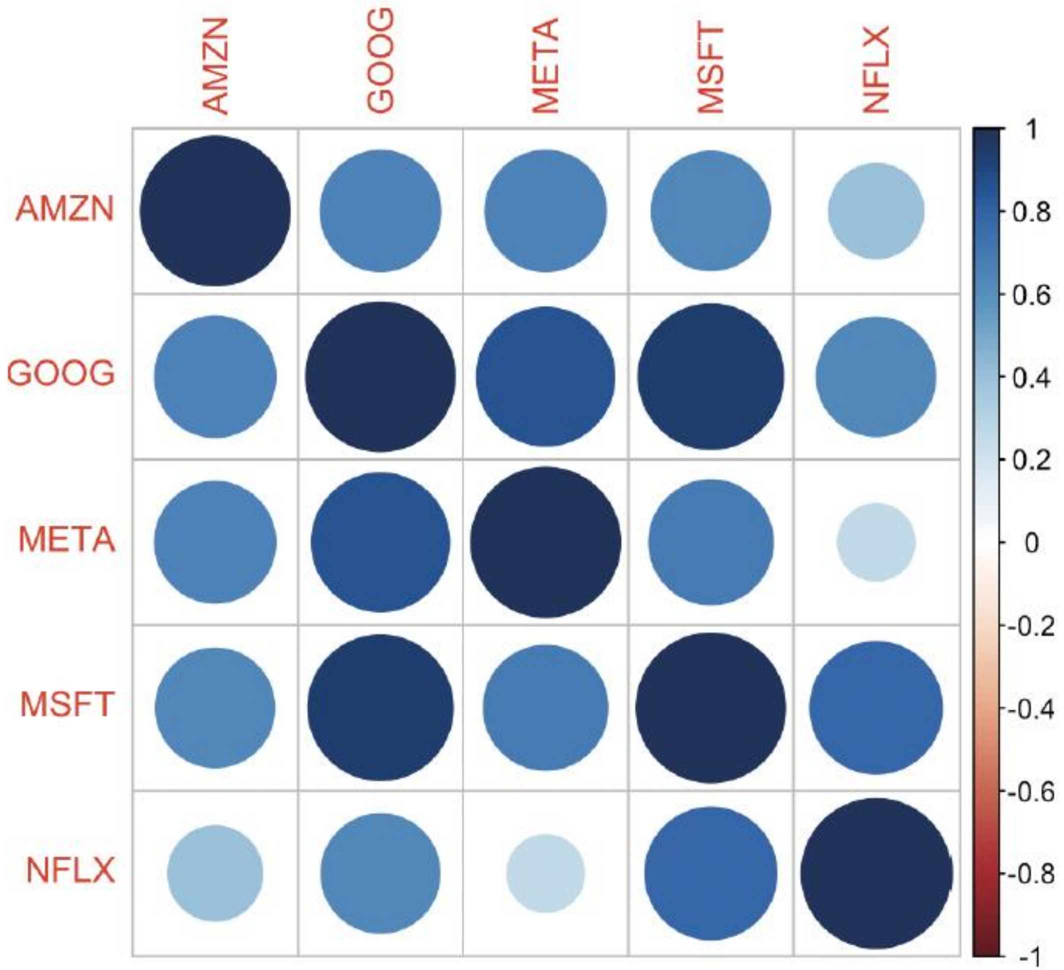


Figure 5.23 – Correlation plot between each pair of stocks

Source: Chapter 5, The Statistics and Machine Learning with R Workshop

In-class quiz

- Q7-9

Feature Engineering

- Process of creating, transforming, or selecting variables to improve model performance
- Transforms raw data into informative inputs for machine learning models
- Transformation
- Binning
- Encoding
- Interaction term
- Aggregation
- Dimension reduction

Feature Transformation

- Build comparable feature ranges
- Standardization
- Min-Max Normalization
- Log transformation

Revisiting Feature Engineering in Tensorflow Playground

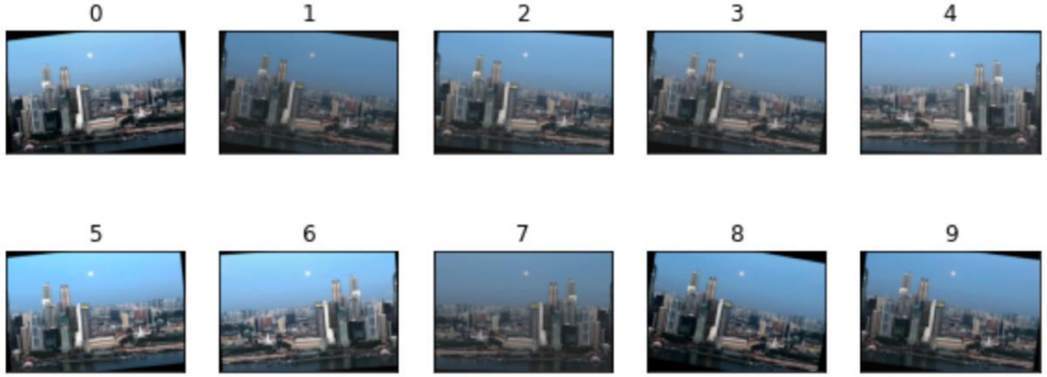
https://playground.tensorflow.org/

In-class quiz

- Q10-12

Feature Engineering in Image Data

- Data augmentation
  - Resizing
  - cropping
  - Rotation
  - Flipping
- Normalization
- Is there use of image data in finance?


0
1
2
3
4
5
6
7
8
9

Feature Engineering in Textual Data

- Lowercasing & Noise Removal
- Stop Words Removal
- Tokenization
- Stemming/Lemmatization
- Text Representation
  - Bag-of-Words (BoW)
  - TF-IDF (Term Frequency–Inverse Document Frequency)
  - N-grams
  - Word Embeddings
- Can we use data augmentation technique for textual data?

Feature Engineering in Financial Data

- Time Series Nature
  - Data often recorded at regular intervals (daily, monthly, quarterly)
  - Trends, seasonality, and cyclic behavior
- Volatility and Noise
  - Markets can be highly volatile; price spikes or drops need careful handling
  - Outlier detection and adjustment are critical
- Price and return
- Technical indicators
- Avoid lookahead bias

Group Discussion

- How can we use image and textual data to improve estimates of asset return and risk in portfolio allocation?
- How to backtest the idea?

Reading materials

- Empirical Asset Pricing via Machine Learning
- Feature Selection and Grouping Effect Analysis for Credit Evaluation

Homework

- First group homework to submit by EOD 27 Apr
- Post learning reflections and questions in the group chat if any
- Review course contents and recording