Ver 1.0

GARP®
RAI™ | Risk and AI

RAI

Video Slides

Video Content Slides

- Module 2: Tools & Techniques
  - Ch1: Introduction to Tools & Techniques
  - Ch2: Unsupervised Learning
  - Ch3: Supervised Learning - Part 1
  - Ch4: Supervised Learning – Part 2
  - Ch5: Semi-Supervised Learning
  - Ch6: Reinforcement Learning
  - Ch7: Model Estimation
  - Ch8: Model Performance Evaluation
  - Ch9: Natural Language Processing
  - Ch10: Generative Artificial Intelligence
- Module 5: Data & AI Model Governance
  - Model Governance
  - Model Validation

Click on a chapter or topic to go to associated slides

GARP®
RAI™ | Risk and AI

GARP®
RAI™ | Risk and AI

Tools and Techniques

Introduction to Tools and Techniques

Dr. Michael Dowling, Professor of Finance, Dublin City University Business School

Machine Learning

- Machine learning (ML)
  - Is an aspect of artificial intelligence (AI) and a set of tools for data analysis and building models.
  - Covers a range of techniques in which a model is trained to recognize patterns in data.
  - Includes prediction and classification models.
  - Has gained significant popularity in recent years and has grown alongside advances in computing power and huge increases in the amount of data available to analysts.

- ML offers advantages over traditional econometrics methods in areas such as:
  - Handling big data
  - Handling non-linearity
  - Reducing dimensionality
  - Handling missing data

Machine Learning vs Classical Econometrics

- Classical econometrics usually involves a hypothesis that the data-generating process (DGP) can be approximated based on theory whereas machine learning (ML) treats the DGP as unknown.
- ML focus is not on inference, but on the ability to produce out-of-sample predictions.
- ML terminology tends to deviate from that of classical statistics.

Types of ML Methodologies

- ML methodologies can be categorized as:
  - Unsupervised
  - Supervised
  - Semi-supervised
  - Reinforcement learning.

- Unsupervised learning
  - Concerned with recognizing patterns in data.
  - Each observation has a vector of features but no corresponding output value to predict.
  - Generally, involves clustering the data or finding a way to reduce the dimensionality of the data.

Types of ML Methodologies

- Supervised learning
  - Concerned with prediction and classification.
    - When the value of a numerical variable (e.g., the price of a house) is to be predicted, this is called a prediction problem.
    - When an observation is to be classified (e.g., a loan is to be classified as “will repay” or “will default”), this is called a classification problem.
  - Each observation has a vector of features and an associated output or label.
  - The algorithm learns from the “labeled” data with the aim of producing accurate predictions of the target value for new, unseen, and unlabeled instances.

Types of ML Methodologies

- Semi-supervised learning
  - Similar to supervised learning, the objective of semi-supervised learning is to make predictions.
  - Only part of the available data is labeled (i.e., some observations do not contain an associated output or label.)
  - The algorithm uses the labeled data to determine patterns within the explanatory variables and to predict "pseudo-labels" for the unlabeled observations.

Types of ML Methodologies

- Reinforcement learning
  - Concerned with making a series of decisions to achieve a goal.
  - No explicit labels. Instead, feedback is provided in the form of a “reward” during the learning process, which encourages a desired behavior but without giving explicit instructions to the learner.
  - Unlike both unsupervised and supervised learning, the “output” from reinforcement learning applications is a recommended action given the circumstances rather than a prediction, classification, or cluster.

Different Data Types

Classification Basis	Data Type	Sub Type	Example
Primary data Types	Numerical data	Continuous data	Income, age, temperature
		Discrete data	Number of daily emails received Number of pages in books
	Categorical data	Nominal data: Data with no natural ordering implied.	Marital status, gender
		Ordinal data: Data with implied natural ordering	Educational level, credit ratings, Likert scale responses (strongly agree, agree, neutral, disagree, strongly disagree)
Organization of data	Structured: Data that can be organized in rows and columns		Census data, mortgage loan origination, and performance data
	Unstructured: Unorganized data. Most alternative data belong to this category		Textual data: Prospectus, books, news releases, transcripts of interviews and earnings calls, etc. Audiovisual: Podcast recordings, voice messages, webinars recordings, maps, photographs, paintings
	Semi-structured: Data has some structure, but it is not a fixed format.		Email, CSV files, HTML files etc.
Longitudinal vs. Cross-sectional data	Longitudinal data: Contains data spanning several time periods		Stock prices and other financial data series, Economic data series
	Cross-sectional data: Data is for a single point in time or time period.		2024 annual income of different individuals Product sales in May 2026 for different companies

Scales of Data Measurement

- Scales of measurement refer to the different levels of measurement.
  - Nominal: data without any inherent order (e.g., eye color and blood type).
  - Ordinal: data with a meaningful order but not having clear intervals between values (e.g., customer satisfaction ratings of dissatisfied, neutral, satisfied).
  - Interval: data has a meaningful order with a consistent interval between values, but no true zero-point, implying that a value of zero does not represent absence of the quantity being measured (e.g., Celsius temperature measurements, calendar dates, longitude, and latitude.)
  - Ratio: data has a natural ordering with a clear interval between values and a true zero point (e.g., height in centimeters and distance traveled).

Data Cleaning

- Data cleaning is important because of the errors potentially associated with the collection process.

- The main reasons for data cleaning are:
  - Inconsistent recording
  - Unwanted observations
  - Duplicate observations
  - Outliers

Data Cleaning

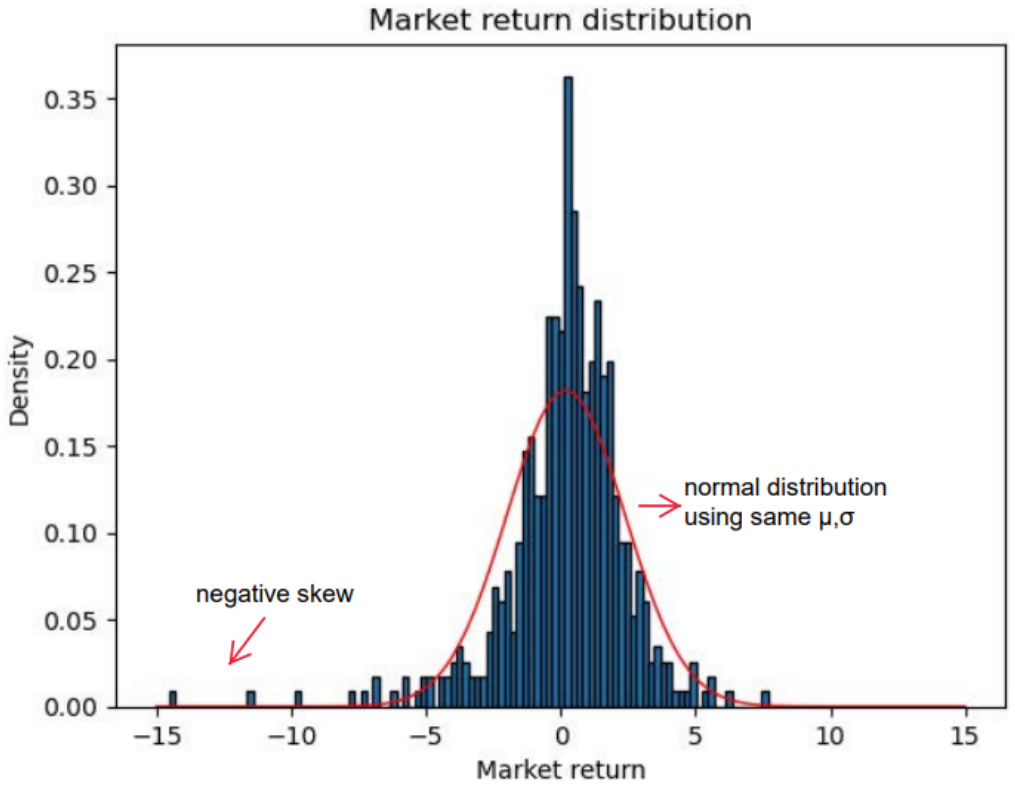
- Missing data is the most common problem encountered during the data-preparation stage.
  - Structurally missing or just unavailable
  - Informative vs. noninformative missingness
- Methods to handle missing data include:
  - Imputation (e.g., a missing feature could be replaced with the mean or median of the features)
  - Missing observations can be estimated in some way from observations on other features (e.g., some predictive models, such as tree-based techniques can account for missing data.)

Data Visualization

- The next step in exploratory data analysis is data visualization.
- Data visualization translates information into a visual context, such as a graph.
- Data visualization can be of great help during the data preparation phase to grasp patterns and identify potential problems, such as outliers.
- During the exploratory phase, data visualization techniques can be used to achieve a reasonable understanding of the shape of the distribution of one or more variables.

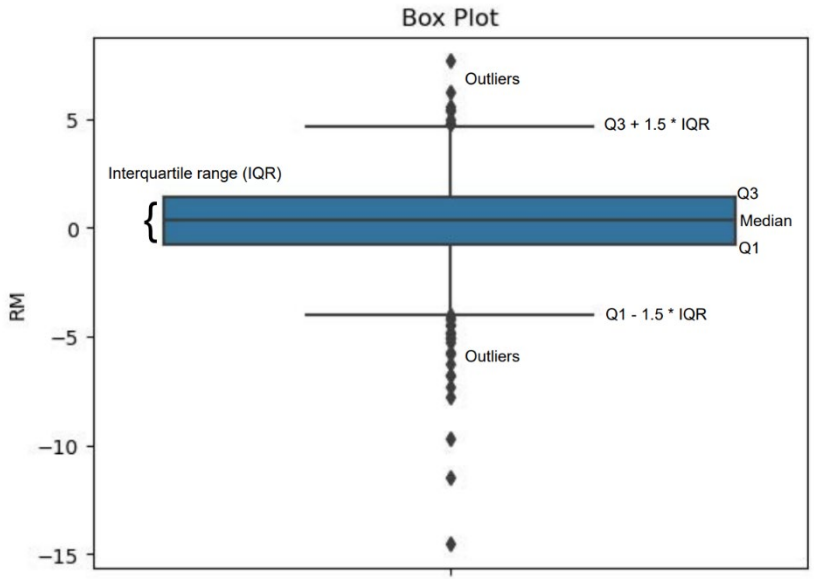
Data Visualization

- A histogram is a count of how many observations fall within specified divisions (bins) of the x-axis.



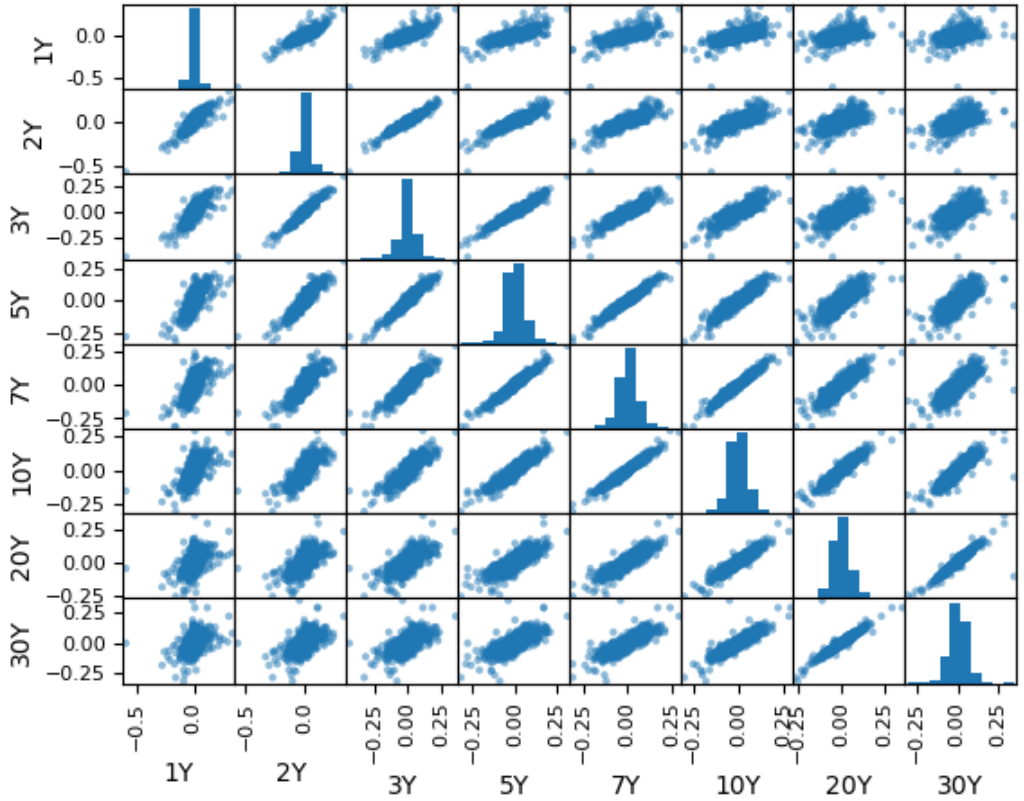
Data Visualization

- A box-and-whiskers plot, sometimes simply called boxplot, can be employed to detect outliers. It is a graphical summary of a distribution.



Data Visualization

- Scatter plots show the relationship among different attributes



Encoding

- Although quantitative data can be directly used as input to a model, qualitative information needs to be transformed in a way that is suitable for statistical analysis.
- The process of transforming non-numeric information into numbers is sometimes termed encoding.

Data Type Converted	Example	Mapped to
Categorical - Nominal	Marital Status	Single (0/1), Married (0/1), Divorced (0/1) etc.
Categorical – Ordinal	Credit Ratings	AAA (6), AA(5), A(4), BBB (3), BB (2), B (1), Unrated (0)
Numerical	Market capitalization	Micro-Cap (<$250MM) (0) Small ($250MM -$2B) (1) Medium ($2B - $10B) (2) Large ($10B-$200B) (3) Mega Cap ($200B and above) (4)

Data Scaling

- Many ML approaches require all the variables to be measured on the same scale; otherwise, the technique will not be able to determine the parameters appropriately and the results will be dominated by the feature with the largest magnitude.

- There are broadly two methods to achieve rescaling:
  - Standardization
    - Scaled the variables to have a mean of zero and a variance of one.

$$\tilde{x}_{ij} = \frac{x_{ij} - \hat{\mu}_i}{\hat{\sigma}_i}$$

  - Normalization
    - Scaled the variables into a boundary of zero and one.

$$\tilde{x}_{ij} = \frac{x_{ij} - x_{i,min}}{x_{i,max} - x_{i,min}}$$

Data Scaling

- Whether standardization or normalization is preferable will depend on the characteristics of the data.

- Standardization is typically preferred when the data contain outliers, because normalization would squash the data points into a tight range that is uncharacteristic of the original data.