SUPPLEMENTARY COURSE MATERIAL

INTRODUCTION TO DATA SCIENCE IN PYTHON

Contents

Datasets can be downloaded at: https://github.com/NhatDucHoang/IS_CS_466

Section 1

Introduction to data science

Data science is an interdisciplinary field that combines statistics, mathematics, computer science, programming, and domain knowledge to extract insights and knowledge from structured and unstructured data for the purpose of making informed decisions, predictions, and solving real-world problems.

At its core, data science involves several key activities:

- Collecting and preparing data: Gathering data from various sources, cleaning, and organizing it so it can be analyzed effectively.
- Analyzing data: Applying methods from statistics, machine learning, and AI to discover patterns, trends, relationships, and predictive information in data.
- Interpreting and communicating findings: Translating the results into actionable insights using data visualization, reporting, and storytelling, often tailored for business or scientific purposes.
- Domain expertise: Integrating specialized knowledge about the particular field (e.g., healthcare, finance) to ensure that insights are relevant and actionable.

Data science is applied across a wide range of industries—including healthcare, finance, transportation, marketing, and manufacturing—to support decision-making, optimize operations, personalize services, and discover opportunities and risks.

The data science process typically covers these main steps:

- Capture (data gathering)
- Maintain (data storage and cleaning)
- Process (pattern finding and modeling)
- Analyze (statistical analysis, predictions)
- Communicate (sharing results through visualization and reports)

Data science unifies mathematics, statistics, and computer science, but is distinct from any single one of these fields due to its strong emphasis on extracting practical knowledge from data for decision-making.

The tools and methods used in data science include programming languages (such as Python), algorithms, machine learning, artificial intelligence, and data visualization techniques.

The importance of data science has grown rapidly due to the explosion of available data and the increasing reliance on data-driven decision-making in business and research.
Data science is often described as "the hottest job of the 21st century" because of its critical role in driving innovation and competitive advantage using data analytics.

Business analytics with data science represents a powerful combination at the intersection of business, statistics, and technology, enabling organizations to make smarter, data-driven decisions. Business analytics focuses on leveraging data to understand past performance, diagnose issues, predict trends, and prescribe actions that can drive growth and improve efficiency.

Data science brings advanced computational and analytical techniques—such as machine learning, statistical modeling, and data visualization—to provide deeper insights from increasingly complex and massive datasets. Together, these fields empower companies to not only interpret what has happened and why, but also forecast future outcomes and optimize strategies across areas like marketing, finance, operations, and customer engagement.

Type of data

Data exists in different forms. Some data are quantitative, meaning they are measured and represented with numbers. Quantitative data focuses on measurable quantities and amounts and is commonly analyzed using statistical techniques. Examples include numerical information like height, weight, temperature, heart rate, and sales numbers.

On the other hand, qualitative data refers to non-numeric information that describes characteristics or qualities and is typically examined using approaches such as thematic or content analysis.

Quantitative data can be divided into two main categories: numeric and categorical, and each group contains several subtypes. Numeric data consists of numbers that represent measurable amounts, and these are often accompanied by symbols to show their units. Numeric data is split into continuous and discrete types.

Continuous data includes values that can take on any number within a range, meaning they are drawn from an unlimited set of possibilities. Discrete data, on the other hand, consists of values with a fixed level of precision, resulting in a limited number of possible values.

Categorical data is represented in different forms such as words, symbols, and even numbers. A categorical value is chosen from a finite set of values, and the value does not necessarily

indicate a measurable quantity. Categorical data can be divided into nominal data and ordinal data. For nominal data, the set of possible values does not include any ordering notion, whereas with ordinal data, the set of possible values includes an ordering notion.

Common examples of nominal data include:

- Gender (male, female, non-binary)
- Hair color (brown, blonde, black, red)
- Nationality (American, German, Kenyan, Japanese)
- Marital status (single, married)
- Type of cuisine (Italian, Chinese, Mexican)
- Favorite color (red, blue, green)

Examples of ordinal data include:

- Educational level (high school, bachelor's degree, master's degree, PhD)
- Satisfaction ratings (very satisfied, satisfied, neutral, dissatisfied, very dissatisfied)
- Socioeconomic status (low, middle, high)5
- Likert scale responses (strongly agree, agree, neutral, disagree, strongly disagree)
- Income brackets (under $35,000, $35,000–$54,999, $55,000–$74,999, above $75,000)

Datasets

When working with data, we often organize information into what's called a dataset. A dataset is a collection of related data, typically arranged in a structured form such as a table.

Each row in this table represents an individual data sample or instance—essentially, a single observation or record in the collection. For example, in a dataset of students, each row might correspond to a different student.

The columns of the table are known as features or attributes; these are the measurable properties or characteristics that describe each instance. Features might include age, grade, and major.

Data science with Python

Python is a widely used, versatile programming language known for its readability and simplicity, which makes it especially appealing to both beginners and experienced programmers.

Its popularity in the field of data science stems from its extensive ecosystem of libraries and tools, such as NumPy for numerical operations, pandas for data manipulation, Matplotlib and Seaborn for visualization, and scikit-learn for machine learning.

Python's clear syntax and integrated support for scientific computing enable analysts and researchers to efficiently process, analyze, and visualize large volumes of data.

Microsoft Visual Studio for Python Programming

Microsoft Visual Studio is a comprehensive integrated development environment (IDE) that offers robust support for Python programming, especially on Windows. With Visual Studio, developers can easily create, edit, and run Python applications using features like IntelliSense for code completion, real-time error checking, and advanced debugging tools.

Visual Studio streamlines Python development by supporting multiple interpreters, such as global Python installations, virtual environments, and conda environments, all manageable from within the IDE. Visual Studio provides a user-friendly, feature-rich environment tailored for both beginners and experienced Python developers.

Python's popularity in data science

Python's popularity in data science is largely due to its rich set of specialized libraries, each designed to streamline different parts of the data science workflow. Here are some of the most widely used Python libraries for data science:

- **NumPy**: Essential for numerical computations, NumPy provides fast and efficient operations on large arrays and matrices, along with mathematical functions for tasks such as linear algebra and Fourier transforms.

- **Pandas**: The library for data manipulation and analysis, pandas offers intuitive data structures like DataFrames, making it easy to clean, reshape, and analyze structured data.
- **Matplotlib**: This classic plotting library is highly flexible and can produce a wide range of static, animated, and interactive visualizations. It serves as the backbone for many other visualization libraries.
- **Seaborn**: Built on Matplotlib, Seaborn makes statistical data visualization simpler and more attractive, offering easy tools for plotting complex graphs with minimal code.
- **SciPy**: Focused on scientific and technical computing, SciPy expands on NumPy with modules for optimization, integration, interpolation, statistics, and more.
- **Scikit-learn**: The standard library for machine learning tasks in Python, scikit-learn supplies a broad suite of algorithms for classification, regression, clustering, dimensionality reduction, and model selection.
- **TensorFlow** and **PyTorch**: These libraries are industry leaders for deep learning and neural network development. Both provide GPU acceleration and broad support for building and training complex models.
- **Statsmodels**: This library is widely used for statistical modeling, hypothesis testing, and data exploration, complementing the capabilities of scikit-learn.

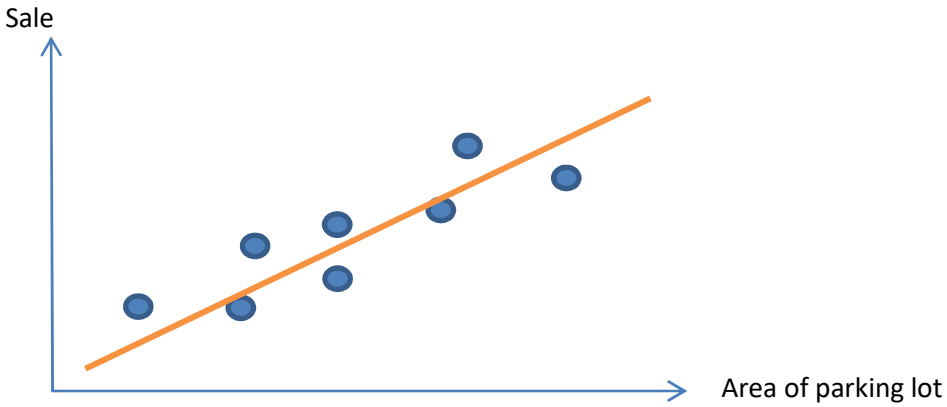
The terminology commonly used in data science

Term	Explanation
Attribute	Another term for predictor; a variable used as input to a model.
Case / Observation	A single unit of analysis (row in data); synonyms: instance, sample, example, record, pattern.
Dependent Variable	Also called response; the variable being predicted in supervised learning.
Estimation	Synonymous with prediction; predicting the value of a continuous variable.
Feature	Synonymous with predictor; an input variable or field.
Holdout Data	Data not used in model fitting, but reserved for assessing model performance; also called validation or test set.
Input Variable	Another word for predictor or feature.

Term	Explanation
Model	A fitted algorithm with its settings, applied to a dataset.
Outcome Variable	Another name for response, dependent, output, or target variable.
Prediction	Forecasting the value of a numerical output variable; also called estimation in regression contexts.
Predictor	Any variable used as input to a model to predict another variable; synonyms: feature, input variable, independent variable, field.
Profile	The set of all measured values for a single observation (e.g., all features for one row).
Record	Observation.
Response	The variable being predicted (usually Y); synonyms: dependent, output, target, or outcome variable.
Sample	In statistics: a collection of observations; in machine learning: a single observation.
Score	The predicted value or class assigned by a model to new data. Scoring means applying a model to new data for prediction.
Supervised Learning	Modeling where the algorithm learns from labeled data to predict an output variable.
Target	Same as response; the variable to be predicted.
Test Data (Test Set)	Data reserved for final model performance assessment; not used during training or model selection
Training Data (Training Set)	Data used to fit (train) the mode.
Unsupervised Learning	Analysis to uncover patterns or structure when no explicit output is being predicted.
Validation Data (Validation Set)	Data used to tune models and select among them, not used for final assessment.
Variable	Any measured quantity, including both input (X) and output (Y) types, in a dataset.

Example 1

A retailer finds that stores with larger parking lots have higher sales. They invest in expanding parking.

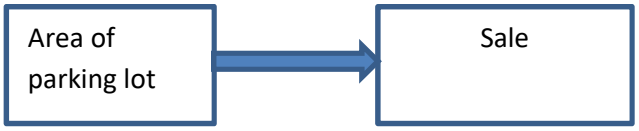


However, after the investment, the retailer finds out that sales do not always increase—because the true cause was that large lots were only built at high-traffic locations, not because bigger parking lots themselves cause more sales.

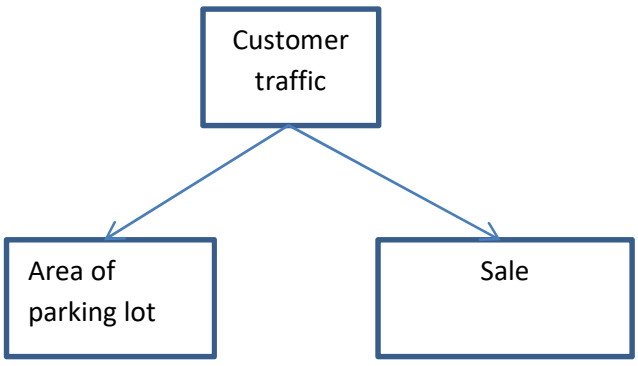
The true underlying reason: Larger lots were constructed because those store locations already experienced higher customer traffic, possibly due to:

- Situated in busy commercial hubs or near highways.
- Serving a larger local population.
- Being anchor stores in large shopping centers.

Here, customer traffic is the real driver of both large lot size and high sales.



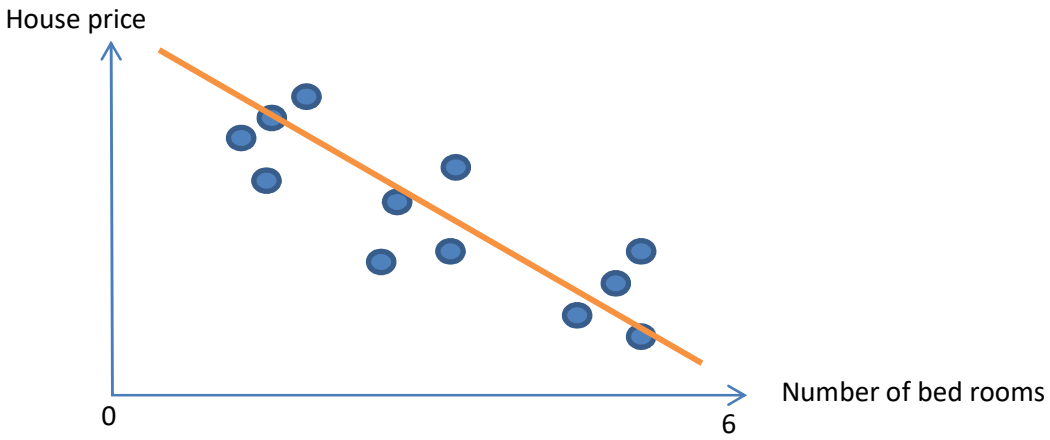
This is what the retailer assumes.



Customer traffic is the real driver.

Example 2

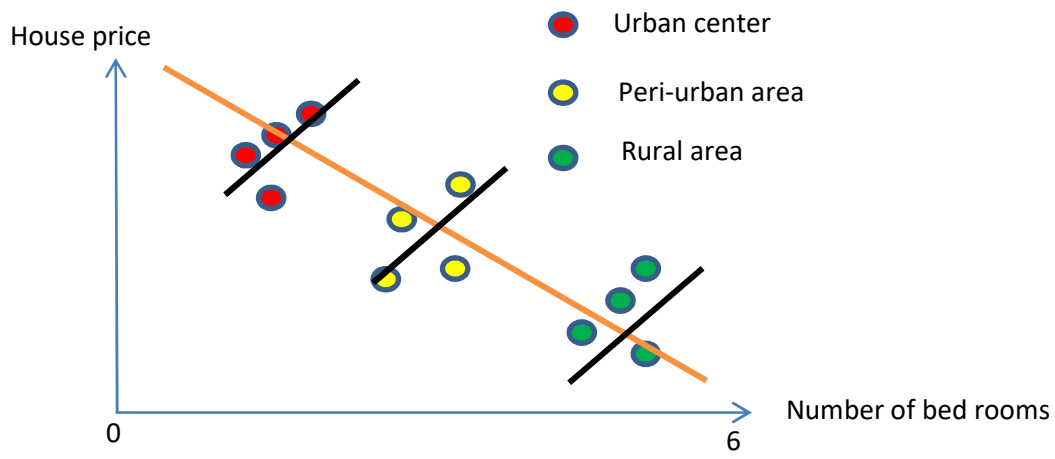
The scatter plot visualizes the relationship between house price (vertical axis) and the number of bedrooms (horizontal axis). The top graph presents all data points together with a single negative trend line. The orange line is a simple linear regression fit showing a negative relationship: as the number of bedrooms increases, the average house price decreases.



In the second scatter plot, the data points by location—urban center (red), peri-urban area (yellow), and rural area (green)—each with its own local trend. Houses are categorized into three groups: urban center, peri-urban area, and rural area. Within each area, house prices generally increase with the number of bedrooms (visualized as upward-tilted black lines for each colored subgroup). The relationship within each subgroup is positive: more bedrooms mean higher prices within the same location type.

While the aggregate data shows a negative trend (more bedrooms => lower price), within each subgroup, the relationship reverses (more bedrooms => higher price).

This paradoxical effect occurs because of a confounding variable: the area (urban, peri-urban, rural). Urban houses have the highest prices per bedroom but tend to have fewer bedrooms. Rural houses have the most bedrooms but the lowest prices per bedroom. When combined, the majority of smaller, more expensive urban houses appear on the left, and larger, cheaper rural houses on the right—creating the misleading negative overall trend.



The data demonstrates the Simpson's Paradox. That is, when individual groups (stratified by a confounder, here "area") show one trend, but the aggregate data shows the opposite.

The key lessons here are:

- Ignoring important subgroups and confounders can lead to incorrect conclusions about causal relationships.
- Always check for hidden variables and analyze relevant subgroups when investigating data relationships, especially in causal inference.

Example 3

A hospital develops a machine learning model to detect a rare but dangerous disease (e.g., Disease X) using two features: blood pressure and cholesterol levels. The dataset consists of many routine patient checkups, but only a small fraction is actual positive cases (patients who truly have Disease X).

Dataset Overview

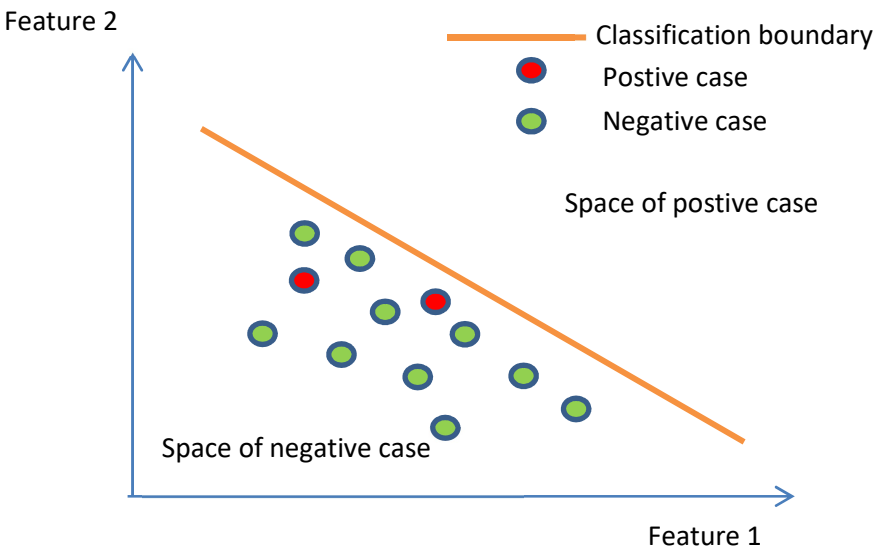
Total patients: 12

Patients with Disease X (positive cases): 2

Healthy patients (negative cases): 10

Suppose the model naively learns to always predict "healthy" regardless of blood pressure or cholesterol.

There are 10 negative cases and 2 positive cases. If a model simply predicts everything as negative, accuracy = 10/12 ≈ 83%. The model's high accuracy belies the fact that it never detects positive cases.



Key lessons:

- High accuracy looks impressive, but it doesn't tell you if the model is actually useful for the real-world goal—identifying positives.
- In many domains (healthcare, security, finance), missing rare positive cases can have significant negative consequences, even if overall accuracy is high.
- Accuracy Paradox warns that in imbalanced data scenarios, high accuracy can be artificially inflated by the majority class.

Example 4: Association Rule Mining (Market Basket Analysis)

Dataset: Transactions of Items Purchased by Customers

Transaction ID	Items Purchased
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Cola
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Cola

Step 1: List Items and Their Supports

- Support is how frequently an item or itemset appears in all transactions.
- Total transactions = 5

Itemset	Count	Support (Count/5)
Bread	4	4/5 = 0.8
Milk	4	4/5 = 0.8
Diaper	4	4/5 = 0.8
Beer	3	3/5 = 0.6
Cola	2	2/5 = 0.4
Bread, Milk	3	3/5 = 0.6
Milk, Diaper	3	3/5 = 0.6
Bread, Diaper	3	3/5 = 0.6
Diaper, Beer	3	3/5 = 0.6

Step 2: Generate Association Rules from Frequent Itemsets

Choose a minimum support threshold, say 0.6, and minimum confidence threshold, say 0.7.

Focus on itemsets with support ≥ 0.6: {Bread}, {Milk}, {Diaper}, {Beer}, {Bread, Milk}, {Milk, Diaper}, {Bread, Diaper}, {Diaper, Beer}

Check confidence for some example rules:

- Rule: Bread → Milk
  Confidence = Support(Bread and Milk) / Support(Bread) = 0.6 / 0.8 = 0.75 (meets minimum confidence)
- Rule: Diaper → Beer
  Confidence = Support(Diaper and Beer) / Support(Diaper) = 0.6 / 0.8 = 0.75 (meets minimum confidence)
- Rule: Milk → Bread
  Confidence = Support(Bread and Milk) / Support(Milk) = 0.6 / 0.8 = 0.75 (meets minimum confidence)

Interpretation:

- If a customer buys Bread, there is a 75% chance they also buy Milk.
- If a customer buys Diaper, there is a 75% chance they also buy Beer.

Example 5

An owner of a store wants to send out emails to customers about the products. He/she wishes to increase the purchase conversion rate.

The purchase conversion rate = the number of purchases / the number of website visitors.

However, the owner does not know whether sending emails can help to increase the purchase conversion.

Data science can help answser the question: Does sending emails increase purchase conversion?

We carry out five steps: (i) Select participants, (ii) Split them into two groups, (iii) The treatment group receives emails and the control group does not receive emails, (iv) Assess the purchase conversion over time, and (v) make decision (confirm or deny the fact the sending email helps increase purchase conversion).

Individual	Treatment outcome	Control outcome
1	1	x
2	1	x
3	1	x
4	1	x
5	0	x
6	x	0
7	x	0
8	x	1
9	x	1
10	x	0

Average outcome of the treatment group = $\frac{1+1+1+1+0}{5} = \frac{4}{5} = 0.8$

Average outcome of the control group = $\frac{0+0+1+1+0}{5} = \frac{2}{5} = 0.4$

Effect = 0.8 – 0.4 = 0.4 → There is a positive effect.

This means sending emails increases the purchase conversion rate by 0.4 (or 40 percentage points) compared to not sending emails.

Let consider the effect of age of the participants. The average age in the treatment group is 30 and the average age in the control group is 36. There is a significant difference in age of the two groups. Thus, age may affect the purchase conversion.

Individual	Age	Treatment outcome	Control outcome
1	20	1	x
2	25	1	x
3	30	1	x
4	35	1	x
5	40	0	x
6	50	x	0
7	35	x	0
8	40	x	1
9	30	x	1
10	25	x	0

To take into account age, we need to estimate counterfactuals via matching. We simply compare the results obtained from the individual of the same age and fill the gap (x) in the data.

Individual	Age	Treatment outcome	Control outcome
1	20	1	1
2	25	1	0
3	30	1	1
4	35	1	0
5	40	0	1
6	50	0	0
7	35	1	0
8	40	0	1
9	30	1	1
10	25	1	0

After estimating the counterfactuals, we estimate the Individual Treatment Effect (ITE):

ITE = Treatment outcome – Control outcome

Then we compute the Average Treatment Effect (ATE) taking into account the factor of age:

$\text{ATE}=\frac{0+1+0+1-1+0+1-1+0+1}{10}=0.2>0$. There is still a positive effect but the ATE here is less than the previously computed value of 0.4.

Individual	Age	Treatment outcome	Control outcome	Individual Treatment Effect (ITE)
1	20	1	1	0
2	25	1	0	1
3	30	1	1	0
4	35	1	0	1
5	40	0	1	-1
6	50	0	0	0
7	35	1	0	1
8	40	0	1	-1
9	30	1	1	0
10	25	1	0	1

Let compute the Conditional Average Treatment Effect (CATE) given age.

$$\text{CATE}(\text{Age} \leq 35) = \frac{0+1+0+1+1+0+1}{7} = 0.57$$

Individual	Age	Treatment outcome	Control outcome	Individual Treatment Effect (ITE)
1	20	1	1	0
2	25	1	0	1
3	30	1	1	0
4	35	1	0	1
7	35	1	0	1
9	30	1	1	0
10	25	1	0	1

$\text{CATE}(\text{Age} > 35) = \frac{-1+0-1}{3} = -0.67$. There is a negative effect on this group.

Individual	Age	Treatment outcome	Control outcome	Individual Treatment Effect (ITE)
5	40	0	1	-1
6	50	0	0	0
8	40	0	1	-1

Recomendation: Send emails to customers aged 35 or below.

Practice

Question 1. Consider the following dataset:

Transaction ID	Items Purchased
1	Apple, Bread, Milk
2	Apple, Diaper, Beer
3	Milk, Diaper, Cola
4	Bread, Milk, Diaper
5	Apple, Milk, Diaper

Compute the confidence for the following rule: Milk → Diaper

Question 2

A pharmaceutical company wants to determine if a newly developed drug effectively cures a disease.

Core Question: Does taking the new drug effectively cure a disease compared to not taking the drug?

Person	Age	Treatment outcome	Control outcome
1	20	1	x
2	25	0	x
3	30	1	x
4	45	0	x
5	50	1	x
6	25	x	1
7	30	x	0
8	45	x	1
9	50	x	0
10	60	x	0

a. Calculate the average cure rate in each group and the increase in cure rate due to the drug.

b. Estimate the counterfactuals.

c. Compute Conditional Average Treatment Effect (CATE) given age: CATE(Age ≥ 50) and CATE(Age < 50). How do you interpret the results?

Question 3.

You are asked to analyze whether a new flu vaccine is effective at preventing flu cases in a population.

1 (No Flu)
0 (Got Flu)

Person	Age	Treatment Outcome (Got Vaccine)	Control Outcome (No Vaccine)
1	18	1	x
2	25	0	x
3	31	1	x
4	37	1	x
5	52	1	x
6	21	x	0
7	35	x	0
8	44	x	1
9	54	x	0
10	62	x	0

a. Calculate the average no-flu (protection) rate in each group and the increase due to vaccination. How do you interpret the results?

b. Estimate the counterfactuals.

c. Compute Conditional Average Treatment Effect (CATE) by age: CATE(Age < 40) and CATE(Age ≥ 40)

Section 2

Statistics

Statistics involves the techniques and principles of gathering, organizing, and examining data to address a research question. Its main purpose is to convert data into insights and a deeper comprehension of our environment. Simply put, statistics is the discipline of gaining knowledge through data.

The process of statistical investigation consists of four steps: (1) defining a statistical question, (2) gathering data, (3) examining the data, and (4) interpreting and sharing the findings.

The three key elements involved in responding to a statistical investigative question:

- Design: Defining the objective or statistical question and devising a plan to collect data that will help answer it
- Description: Organizing and examining the collected data
- Inference: Drawing conclusions and making forecasts from the data to address the statistical question

Descriptive statistics involves techniques used to summarize gathered data, whether it represents a sample or an entire population. These summaries typically include visual representations like graphs, as well as numerical measures such as means and percentages.

Poplation, sample, and statistical inference

- When you analyze a data set during your study, it is because the data represent a group of experimental units that are relevant to your research. In statistical terms, the complete collection of data from all these experimental units is referred to as the population.
- A sample is a portion of data taken from a larger population.
- Statistical inference involves making estimates, predictions, or broader conclusions about the entire population using the data gathered from that sample.

Normal Distribution

The normal distribution, also known as the Gaussian distribution, is a very common continuous probability distribution in statistics and many other fields. Its most notable feature is the bell-shaped probability density curve, which is symmetrical around the **mean value (μ)**.

This distribution is characterized by two main parameters:

- Mean ($\bar{y}$): Defines the central location of the distribution.
- Standard deviation (s): Measures how spread out the data is around the mean.

According to the empirical rule, approximately 68% of values lie within ±1σ, 95% within ±2σ, and 99.7% within ±3σ of the mean.

The variance of a sample of n measurements $y_1, y_2, \ldots, y_n$ is defined as

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}$$

where $\bar{y}$ is the sample mean.

The standard deviation of a set of measurements is the square root of the variance.

The formula for the standard deviation s is:

$$s = \sqrt{s^2}$$

where s^2 represents the variance.

Example 1

The variable cost per apartment (VC) (measured in billion VND) is taken from past sales data as follows:

[1,25; 1,14; 1,19; 1,31; 1,17; 1,20; 1,14; 1,28; 1,19; 1,21]

$$\bar{x} = \frac{1.25 + 1.14 + 1.19 + 1.31 + 1.17 + 1.20 + 1.14 + 1.28 + 1.19 + 1.21}{10}$$

$$= \frac{11.08}{10} = 1.108$$

Use the sample variance formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$s^2 = \frac{0.02876}{10-1} = \frac{0.02876}{9} \approx 0.003196$$

Calculate Standard Deviation s

$$s = \sqrt{0.003196} \approx 0.0565$$

Confidence interval

The confidence interval (CI) of a variable is a range of values determined from sample data that indicates the degree of confidence that the true value of the variable lies within this range with a certain probability (for example, 95%).

Once the mean and standard deviation of the data are known, the confidence interval helps us better understand the uncertainty and risk associated with the data.

Assuming the variable cost per apartment (VC) follows an approximately normal distribution, the 95% confidence interval (CI) for the break-even point (BEP) is calculated using the following formula:

$$CI = \bar{x} \pm M \times \frac{s}{\sqrt{n}}$$

where M = 2.26 is the coeffcient corresponding to a 95% confidence level.

CI = 1.108 ± 2.26×0.0565/ $\sqrt{10}$ = [1.07, 1.15]

Example 1

Outlier detection is an essential step in data cleaning for business analytics, research, and operational analysis. Outliers—values that deviate significantly from the rest of the data—can distort statistical calculations such as the mean and standard deviation, leading to misleading conclusions. Common sources of outliers include data entry errors, measurement mistakes, system error, or rare events.

Consider a scenario where a company is reviewing the completion times for a standard process performed by staff. The data collected for ten recent processes is as follows:

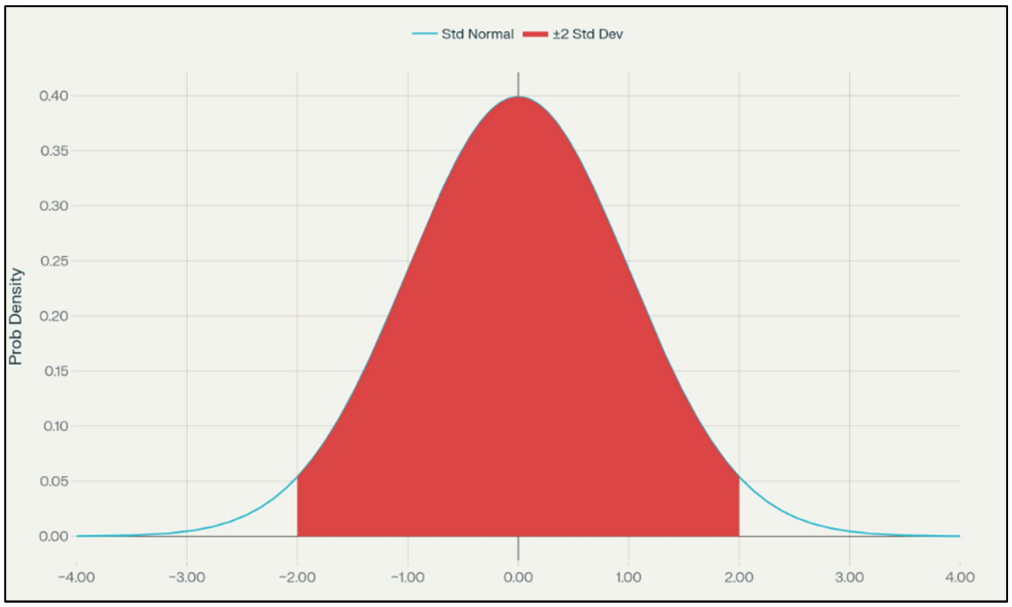
```
data = [10, 12, 11, 9, 12, 8, 10.5, 16, 11, 9.5] (hour)
```

A common approach to identify outliers is to use thresholds based on the standard deviation (std) from the mean:

$$\text{Lower bound} = \bar{x} - k \times s$$
$$\text{Upper bound} = \bar{x} + k \times s$$

where:

- $\bar{x}$ is the sample mean,
- s is the sample standard deviation,
- k is the threshold coefficient (commonly set to 2 or 3).



A threshold of 2×std allows for the detection of moderate outliers. In a normal distribution, roughly 95% of values lie within ±2 standard deviations from the mean. Thus, data points outside this range are statistically uncommon and may be flagged for review.

Mean: 10.90, Std Dev: 2.20

Lower bound: 6.51, Upper bound: 15.29

```
Original = [10, 12, 11, 9, 12, 8, 10.5, 16, 11, 9.5]
```

Cleaned data: [10.0, 12.0, 11.0, 9.0, 12.0, 8.0, 10.5, 11.0, 9.5]

Example 2

The break-even point (BEP) is a fundamental financial concept representing the level of sales at which total revenue exactly equals total costs, resulting in neither profit nor loss. At this critical threshold, a business has generated enough income to fully cover both its fixed and variable expenses, but has not yet begun to realize a profit. Understanding and calculating the break-even point allows companies to set informed sales targets, guide pricing strategies, and evaluate the financial viability of business plans.

BEP is computed as follows:

$$BEP = \frac{FC}{P - VC}$$

where FC is the fixed cost; P is the unit price; VC is the variable of cost of one unit.

A company has undertaken an investment to develop a new urban area. As part of the financial planning for the project, the company must determine how many apartment units need to be sold to reach its break-even point (BEP)—the minimum sales volume required for total revenue to equal total costs, resulting in zero profit or loss.

The fixed cost (FC) for the project is 100 billion VND (one-time costs such as permits, infrastructure, and initial construction).

The unit selling price (P) for each apartment is set at 2.00 billion VND.

The variable cost per apartment (VC)—which includes expenses that vary with the number of units produced, such as materials, labor, and finishing—has been evaluated from past sales as follows (in billion VND):

VC = [1,25; 1,14; 1,19; 1,31; 1,17; 1,20; 1,14; 1,28; 1,19; 1,21]

For each data point:

$$\text{BEP}_i = \frac{FC}{P - VC_i}$$

Where:

- FC = 100
- P = 2.00
- VC_i = Variable cost from the data

BEP = [133.33, 116.28, 123.46, 144.93, 120.48, 125.00, 116.28, 138.89, 123.46, 126.58]

Mean of BEP = 126.87

Std of BEP = 9.45

$CI = \bar{x} \pm M \times \frac{s}{\sqrt{n}}$ where M = 2.26.

CI = 126.87± 2.26×9.45/$\sqrt{10}$ = [121, 134]

Based on historical cost data, the company needs to sell approximately 127 apartments. With 95% confidence, the actual number needed lies between 121 and 133 units, given the observed variability in unit costs.

Example 3

An investor is evaluating two investment options—Fund A and Fund B—by analyzing their historical annual returns over a 7-year period. The investor's goal is to choose the fund that aligns best with their risk tolerance and return requirements.

Fund A is advertised as growth-oriented but with higher risk.

Fund B is marketed as more stable and suitable for conservative investors.

Historical Returns Data

Year	Fund A (%)	Fund B (%)
1	12	8
2	15	9
3	18	7
4	5	8
5	20	10
6	8	7
7	16	9

fund_a_returns = [12, 15, 18, 5, 20, 8, 16]; fund_b_returns = [8, 9, 7, 8, 10, 7, 9]

Fund A: Mean = 13.43, Std. = 5.41

Fund A: Mean = 8.29, Std. = 1.11

Interpretation:

- Fund A offers a higher average annual return but with considerable year-to-year variability (risk).
- Fund B delivers a lower but more stable return.

Example 4

A company is planning to reallocate its investment portfolio to balance risk and return more effectively. The board is considering two portfolio options based on historical return and risk analysis:

- Portfolio X: 60% equities, 30% bonds, 10% real estate.
- Portfolio Y: 40% equities, 50% bonds, 10% real estate.

The company's investment committee seeks data-driven guidance to select the portfolio that fits its moderate risk appetite while ensuring stable contributions to profitability. The company gathers the historical data in the past 5 years as follows:

Year	Equities (%)	Bonds (%)	Real Estate (%)
1	10.5	4.2	6.1
2	13.1	3.8	5.7
3	7.4	4.6	7.2
4	11.0	4.3	6.8
5	8.6	4.1	6.4

Evaluate the company's portfolios by computing the mean and std. of each option.

For every year, calculate the portfolio's annual return by taking a weighted average of the returns of all asset classes, using the portfolio's allocation as weights:

$$\text{Portfolio Return}_{\text{Year n}} = w_1 \times r_{1,n} + w_2 \times r_{2,n} + \ldots + w_k \times r_{k,n}$$

Where:

- w_i = Weight of asset class i in the portfolio
- $r_{i,n}$ = Return of asset class i in year n
- k = Number of asset classes

Example:

If Portfolio X has 60% equities, 30% bonds, 10% real estate:

$$\text{Portfolio X Return}_{\text{Year 1}} = 0.6 \times (\text{Equities}_1) + 0.3 \times (\text{Bonds}_1) + 0.1 \times (\text{Real Estate}_1)$$

Once you have the portfolio return for each year, compute the **mean** (sample average) return:

$$\text{Mean Return} = \frac{1}{N} \sum_{n=1}^{N} \text{Portfolio Return}_n$$

Where N is the number of years.

Next, calculate the **standard deviation** of those annual portfolio returns:

$$\text{Standard Deviation} = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} (\text{Portfolio Return}_n - \text{Mean Return})^2}$$

Year	Equities (%)	Bonds (%)	Real Estate (%)	X	Y
1	10.5	4.2	6.1	8.17	6.91
2	13.1	3.8	5.7	9.57	7.71
3	7.4	4.6	7.2	6.54	5.98
4	11	4.3	6.8	8.57	7.23
5	8.6	4.1	6.4	7.03	6.13

Portfolio X: Mean = 7.98, Std =1.21

Portfolio Y: Mean: 6.79, Std = 0.73

We compare the mean (higher is generally better) and standard deviation (lower means less risk/volatility) of each option.

Decision Implications

The company can use these summary statistics to align its investment strategy:

- If aiming for long-term asset growth and higher surplus, Portfolio X is favorable.
- If seeking smoother results and meeting regular cash flow needs, Portfolio Y is more suitable.

Practice

Question 1. A retail manager is analyzing the daily sales figures (number of units sold) for a specific product over 10 days. Accurate analysis is crucial for inventory planning and avoiding stockouts or overstocks. However, human error in systems can occasionally result in abnormally high or low reported sales.

sales_data = [42, 39, 41, 43, 40, 44, 42, 41, 40, 66]

To ensure meaningful trend analysis and forecasting, the manager decides to remove any data point lying more than 2 standard deviations from the mean.

Question 2. In business and operations research, analyzing average product demand is critical for effective inventory management. Companies often take a sample of demand data and use statistical tools such as mean, standard deviation, and confidence interval to estimate the true average demand.

A store manager collects the weekly demand (units sold) for a particular product over 10 weeks:

Week Number	1	2	3	4	5	6	7	8	9	10
Units Sold	54	61	47	53	57	62	58	55	52	59

The manager wants to estimate, with 95% confidence, the interval in which the true average weekly demand for this product falls.

Question 3. An e-commerce company wants to estimate the average delivery time for its orders. The company collects delivery times (in days) for 10 recent deliveries:

Delivery times = [3.2, 2.8, 3.5, 4.0, 3.7, 3.3, 3.0, 4.2, 3.6, 3.9] (day)

Calculate the mean delivery time, standard deviation, and a 95% confidence interval for the average delivery time based on the sample.

Question 4. An individual is planning for retirement and considering two diversified investment portfolios based on past asset performance. The goal is to achieve steady growth with a moderate risk profile over the next decade.

- Portfolio A: 50% domestic stocks, 30% international stocks, 20% government bonds.
- Portfolio B: 30% domestic stocks, 20% international stocks, 50% government bonds.

The investor wants to compare each portfolio's average annual return and risk, as measured by standard deviation.

Historical data is reported as follows:

Year	Domestic Stocks (%)	International Stocks (%)	Government Bonds (%)
1	11.2	8.4	3.5
2	9.5	7.1	4.0
3	13.3	6.8	3.9
4	7.8	9.2	4.4
5	10.6	10.0	4.2
6	8.7	7.4	3.8

Section 3

Scatterplot for data visualization

When dealing with two quantitative variables, the response variable is typically labeled as *y*, and the explanatory variable as *x*. This labeling is standard because graphs that display their relationship usually place the response variable on the y-axis and the explanatory variable on the x-axis. Such a graph is known as a scatterplot.

Pearson correlation coefficient

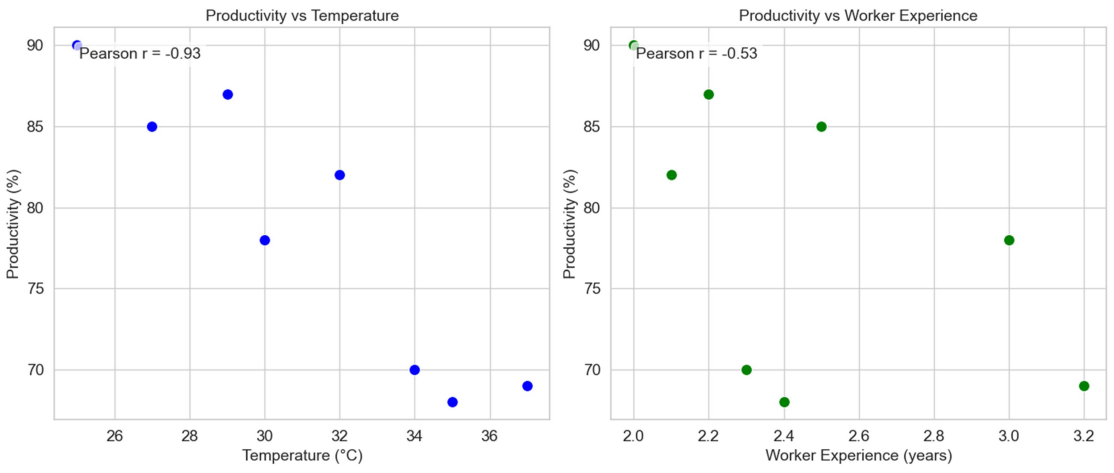
The Pearson correlation coefficient (*r*) is a statistical metric that quantitatively measures the strength and direction of the linear relationship between two variables. Its value ranges from -1 **to +1:** +1 indicates a perfect positive linear correlation, -1 indicates a perfect negative linear correlation, and 0 indicates no linear correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where
n is the number of data points
x and ***y*** are the two variables under consideration.

Example 1

A construction company tracks weekly productivity (measured as the percentage of planned output achieved) over 8 weeks. They want to review and analyze the factors affecting the workers' labor productivity at the construction site.



Week	Workload (%) (Compared to plan)	Average temperature (°C)	Average years of experience of the work team
1	90	25	2
2	85	27	2.5
3	87	29	2.2
4	78	30	3
5	82	32	2.1
6	70	34	2.3
7	68	35	2.4
8	69	37	3.2

Based on the strong negative Pearson correlation coefficient (r) observed between temperature and labor productivity in the data, we can conclude that high environmental temperatures significantly reduce the productivity of construction workers. In contrast, the weak correlation between workers' experience and productivity suggests that, within the observed range, experience alone does not have a strong impact on productivity changes compared to environmental factors such as temperature.

Temperature is an important factor affecting productivity, with productivity decreasing as temperature rises beyond the optimal moderate range (around 24–26°C). Extremely high temperatures cause physiological stress that weakens workers' performance. Based on the data, the company should implement appropriate measures to improve labor productivity.

With a Pearson correlation coefficient of r = 0.91, we can infer that there is a strong positive linear relationship between the manager's experience and productivity in the data set. This means that as the manager's experience increases, productivity tends to increase as well.

	A	B	C	D	E	F	G	H
1	Week	x	y	x-x_m	y-y_m	(x-x_m)*(y-y_m)	(x-x_m)^2	(y-y_m)^2
2	1	**7**	**65**	-2.5	-12.75	31.875	6.25	162.563
3	2	**6**	**69**	-3.5	-8.75	30.625	12.25	76.5625
4	3	**8**	**73**	-1.5	-4.75	7.125	2.25	22.5625
5	4	**9**	**77**	-0.5	-0.75	0.375	0.25	0.5625
6	5	**9**	**80**	-0.5	2.25	-1.125	0.25	5.0625
7	6	**10**	**83**	0.5	5.25	2.625	0.25	27.5625
8	7	**12**	**86**	2.5	8.25	20.625	6.25	68.0625
9	8	**15**	**89**	5.5	11.25	61.875	30.25	126.563
10		x_m	y_m			Sum	Sqrt(sum)	Sqrt(sum)
11		9.5	77.75			154	7.61577	22.1246
12								
13		**r =**	**0.9139667**					

Simple linear regression

Simple linear regression is a statistical method used to model and estimate the relationship between two variables: an independent variable (predictor variable x) and a dependent variable (y). The goal is to find the best-fitting linear equation to the data, which allows predicting the dependent variable based on the independent variable.

$y=B_0+B_1x$ $B_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$ $B_0 = \bar{y} - B_1\bar{x}$ $\bar{x}$ and $\bar{y}$ are the means of x and y.	Where: • y is the predicted value of the dependent variable, • x is the independent or predictor variable, • B_0 is the predicted value of y when x = 0, • B_1 is the slope or regression coefficient (indicating how much y changes for a one-unit change in x)

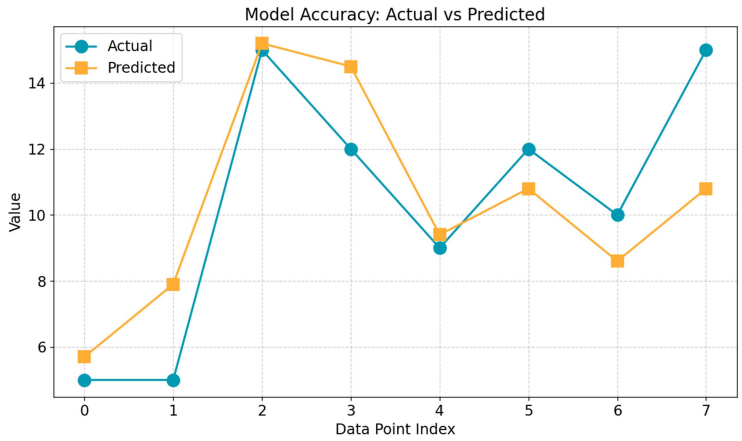
Example 2

We use the regression model to estimate the level of cost overrun based on the management team's experience.

x - the management team's experience (year)	*y* - cost overrun (%)	(x-x_tb)	(y-y_tb)	(x-x_tb)*(y-y_tb)	(x-x_tb)^2
15	5	6.38	-5.38	-34.265625	40.640625
12	5	3.38	-5.38	-18.140625	11.390625
2	15	-6.63	4.63	-30.640625	43.890625
3	12	-5.63	1.63	-9.140625	31.640625
10	9	1.38	-1.38	-1.890625	1.890625
8	12	-0.63	1.63	-1.015625	0.390625
11	10	2.38	-0.38	-0.890625	5.640625
8	15	-0.63	4.63	-2.890625	0.390625
8.63	10.38			-98.875	135.875
Average	Average			Sum	Sum

The results are as follows: B_1 =-0.73 và B_0 =16.65

x	y_p (Predicted) = B0 + B1*x	y (Actual)	e (error) = y - y_p	|e|	Mean error rate (%) = |e|*100/y
15	5.7	5	-0.7	0.7	14.7
12	7.9	5	-2.9	2.9	58.4
2	15.2	15	-0.2	0.2	1.3
3	14.5	12	-2.5	2.5	20.6
10	9.4	9	-0.4	0.4	4.2
8	10.8	12	1.2	1.2	9.8
11	8.6	10	1.4	1.4	13.5
8	10.8	15	4.2	4.2	27.8
				1.7	18.8
				Average	Average



We also use root mean square error (RMSE) to assess the model's performance. RMSE is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_{pi})^2}{N}}$$

where y_i is the observed output data (the i-th forecasted value) and y_{pi} is the i-th output value calculated from the model. N is the number of data points collected (for example, N = 5). The regression model with a smaller RMSE is better.

The coefficient of determination, commonly denoted as r^2, is a key statistic used to evaluate the performance of regression models. r^2 measures the proportion of variance in the dependent variable that can be explained by the independent variable(s) in the model. It provides valuable insight into how well the regression line fits the observed data, a concept known as the model's goodness of fit.

The coefficient of determination (r^2) is also used; it is computed as follows:

$$r^2 = 1 - \frac{SSE}{SSyy}$$

where

$$SSyy = \sum_{i=1}^{N}[(y_i - y_{avg})^2]$$

$$SSE = \sum_{i=1}^{N}[(y_i - y_{pi})^2]$$

The regression model with an r^2 value closer to 1 is more accurate. Conversely, a model with an r^2 value closer to 0 is less reliable.

Metric	Formula
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\left\|e_i\right\|$
Mean Error	$\frac{1}{n}\sum_{i-1}^{n}e_i$
Mean Percentage Error (MPE)	$100\times\frac{1}{n}\sum_{i-1}^{n}\frac{e_i}{y_i}$
Mean Absolute Percentage Error (MAPE)	$100\times\frac{1}{n}\sum_{i=1}^{n}\left\|e_i/y_i\right\|$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i-1}^{n}e_i^2}$

Interval Forecasting with a linear regression model

After establishing the regression model (calculating B_0 and B_1), we can proceed to forecast the confidence interval of the dependent variable (y) based on information from the predictor variable ($x = x_0$) as follows:

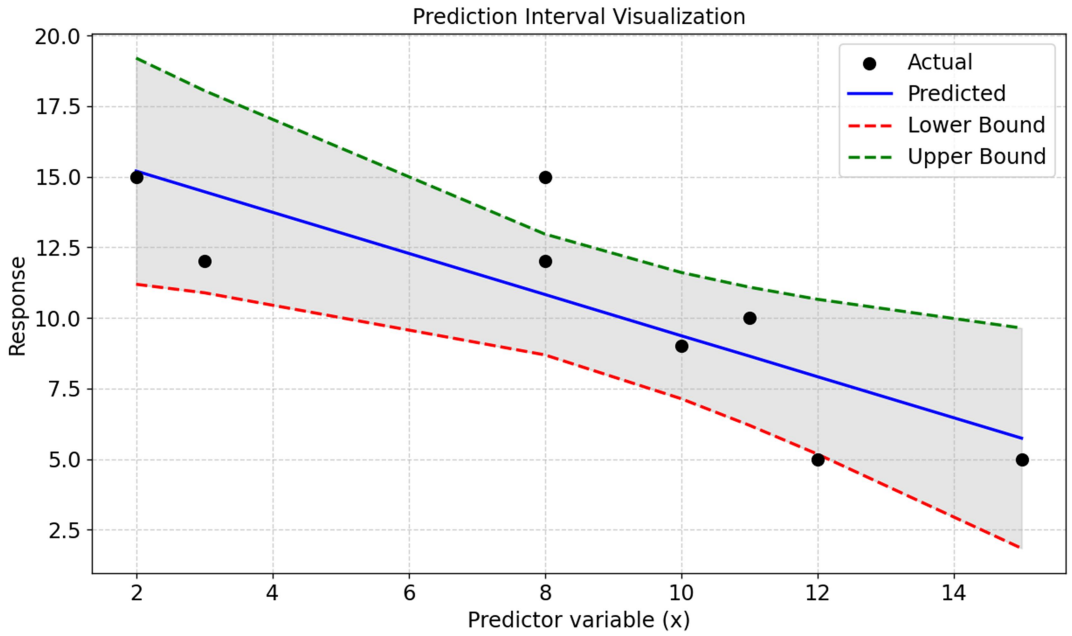
$$B_0+B_1 \text{ x } x_0 \pm \lambda \text{ x } \sqrt{MSR \times (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}})} = y_p \pm \varepsilon$$

where

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$MSR = \frac{\sum_{i=1}^{n} (y_i - y_{p,i})^2}{n-2}$$

$\lambda = t_{\alpha/2,n-2}$ is is looked up in the table.



Example 1

We want to investigate the relationship between years of managerial experience and annual sales revenue among sales managers in a company.

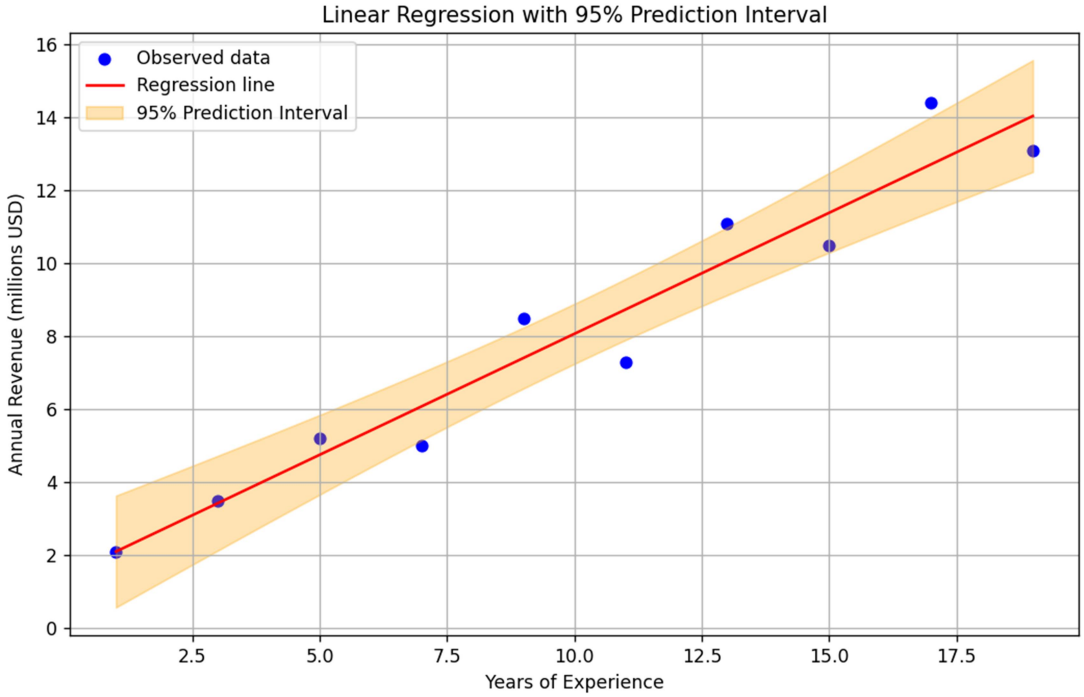
years_of_experience (x) = [1, 3, 5, 7, 9, 11, 13, 15, 17, 19]

annual_revenue (y) = [2.1, 3.5, 5.2, 5.0, 8.5, 7.3, 11.1, 10.5, 14.4, 13.1]

The parameters of the linear regression model $y = B_0 + B_1.x$ are as follows: $B_0 = 1.47$ and $B_1 = 0.74$.

We have MSR = 1.278 and t_crit = 2.306.

x_0 (Years of Experience)	Predicted y (Annual Revenue, million USD)	95% Confidence Interval
1.0	2.100	[0.568, 3.632]
3.0	3.427	[2.128, 4.726]
5.0	4.753	[3.661, 5.846]
7.0	6.080	[5.150, 7.010]
9.0	7.407	[6.570, 8.243]
11.0	8.733	[7.897, 9.570]
13.0	10.060	[9.130, 10.990]
15.0	11.387	[10.294, 12.479]
17.0	12.713	[11.414, 14.012]
19.0	14.040	[12.508, 15.572]



Example 2

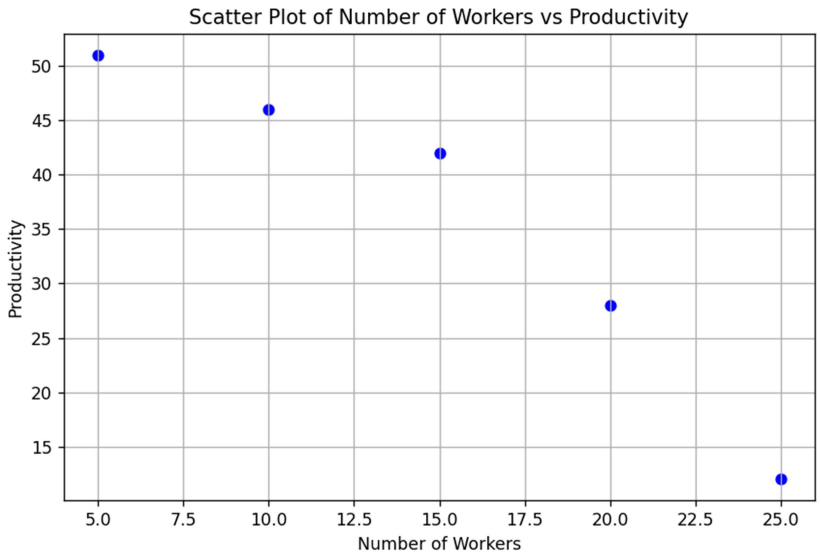
A constructor records the data during the construction process of an acitivity as follows:

Number of workers

workers = [5, 10, 15, 20, 25]

Productivity

productivity = [51, 46, 42, 28, 12]



Using a simple linear regression model, we have $B_0 = 64.6$ and $B_1 = -1.92$. The average error of this model is 3.44%.

We carry out a variable transformation process as follows: $x_t = x^2$; the new model is as follows:

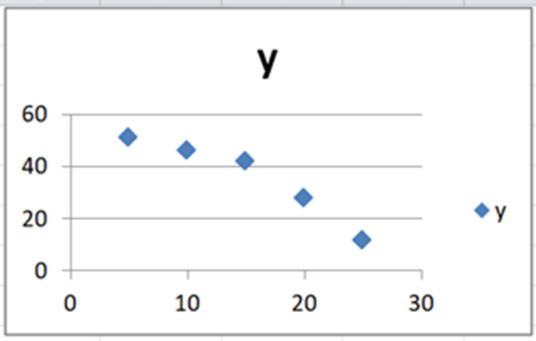
$y = B_0 + B_1.x_t$ with $B_0 = 53.68$ and $B_1 = -0.065$. By doing so, the average error of this model can be reduce to 1.31%.

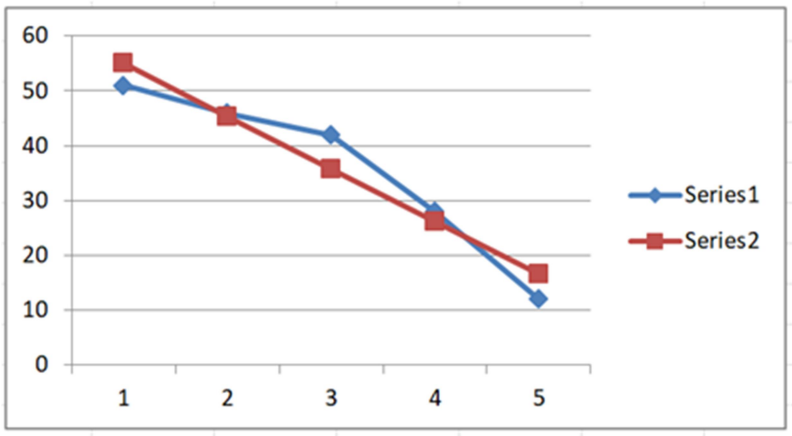
Input variable transformation through nonlinear functions is a powerful technique to enhance the flexibility and performance of linear regression models. Although linear regression assumes a straight-line relationship between the input variables and the target outcome, many real-world relationships are inherently nonlinear. By applying nonlinear transformations—such as polynomial, logarithmic, exponential, or other nonlinear functions—to the original input variables, we create new features that allow the model to capture complex, curved patterns in the data while still maintaining a linear form in terms of the parameters.

This approach broadens the applicability of linear regression by enabling it to approximate nonlinear dependencies without abandoning the interpretability and simplicity that linear models

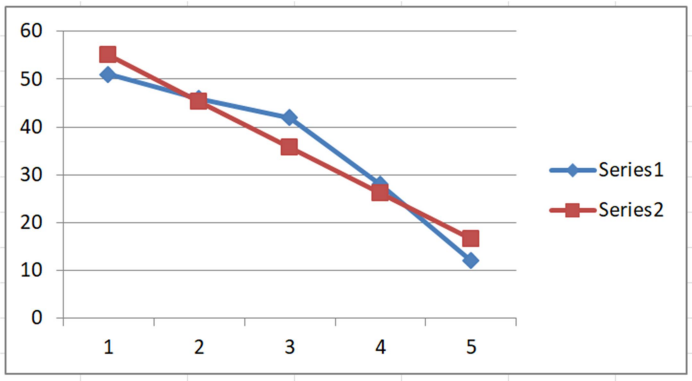
provide. Ultimately, nonlinear transformations of input variables enrich the model's expressive power and can lead to improved accuracy and better understanding of the underlying relationships.

C	D	E	F	G	H	I	J	K
x	**y**	**x*y**	(x-xm)^2					
5	51	255	100					
10	46	460	25					
15	42	630	0					
20	28	560	25					
25	12	300	100					
x_m	**y_m**	SUM						
15	35.8	2205						
SSxy	-480		**x**	**y**	**b0**	**b1**	**yp**	|Delta|
SSxx	250		5	51	64.6	-1.92	55	4
b1	-1.92		10	46	64.6	-1.92	45.4	0.6
b0	64.6		15	42	64.6	-1.92	35.8	6.2
			20	28	64.6	-1.92	26.2	1.8
			25	12	64.6	-1.92	16.6	4.6
							Average	**3.44**





B	C	D	E	F	G	H	I	J	K
x	**y**	**xt = x^2**	**y**	**xt*y**	(xt-xt_m)^2				
5	51	25	51	1275	62500				
10	46	100	46	4600	30625				
15	42	225	42	9450	2500				
20	28	400	28	11200	15625				
25	12	625	12	7500	122500				
		xt_m	**y_m**	**SUM**					
		275	35.8	34025					
SSxy	-15200		**x**	**xt = x^2**	**b0**	**b1**	**yp**	**y**	**|Delta|**
SSxx	233750		5	25	53.68235294	-0.065	52.0567	51	1.05668
b1	-0.065		10	100	53.68235294	-0.065	47.1797	46	1.17968
b0	53.6824		15	225	53.68235294	-0.065	39.0513	42	2.94866
			20	400	53.68235294	-0.065	27.6717	28	0.32834
			25	625	53.68235294	-0.065	13.0406	12	1.04064
								Average	**1.3108**



Commonly used nonlinear transformations for input variables in linear regression include:

- Polynomial transformation: Raising variables to powers (e.g., x^2, x^3) to capture curvilinear relationships such as quadratic or cubic trends.
- Logarithmic transformation: Applying the natural logarithm (logx) to variables, which is especially useful when the relationship between variables is multiplicative or exhibits diminishing returns.
- Exponential transformation: Transforming variables using exponentials (e.g., e^x), suitable when changes in the dependent variable accelerate or decelerate rapidly with increases in the input. Square root transformation:
- Using the square root (sqrt(x)), often helpful when variance increases with the mean or when the effect of the predictor tapers off as it increases.

- Reciprocal transformation: Taking the reciprocal (1/x) to handle certain types of nonlinear decay relationships.
- Power transformation: Raising variables to arbitrary powers (e.g., x^{α}), which generalizes polynomial and root transformations and can help in achieving linearity or stabilizing variance.
- Trigonometric functions: Applying sine, cosine, or other periodic functions to capture cyclic patterns.

Time series forecasting with a linear regression model

Forecasting in business management

Forecasting is a common task in the management of business activities. This process assists us in making decisions related to planning and organizing work.

Objectives of Forecasting

The goal of forecasting is to provide predicted data for a certain variable (for example, the cost of a material, labor costs, etc.) at a specific point in the future. We make forecasts based on an analysis of historical data for that variable.

Types of Forecasting

Forecasting is divided into three applications:

- Short-term forecasting: covers the near future (days, weeks)
- Medium-term forecasting: focuses on the intermediate future (weeks, several weeks, months)
- Long-term forecasting: deals with the distant future (years, many years)

Components of a Time Series

A time series typically consists of the following components:

Trend component (Trend)

Seasonal component (Season)

The decomposition of a time series can be expressed as: $Y_t = T_t + S_t + E_t$

where

- Yt is the value of the variable (of the time series) at time t
- Tt is the trend component
- St is the seasonal component
- Et is the random component (expected to be a small number, close to zero)

The construction of a model is carried out in three steps as follows:

Step 1: Build a univariate regression model to separate the trend component (Trend):

$$Y_{\text{Trend}} = B_0 + B_1 t$$

where Y_{Trend} is the trend component. B_0 and B_1 are the two coefficients of the regression model. t represents the past time points.

Step 2: Remove the trend component from the data series:

$$Y_S = Y - Y_{\text{Trend}}$$

where Y_S is the seasonal component plus the random component. Since the random component is assumed to be approximately zero, we consider:

$$Y_S = \text{Seasonal component}$$

Step 3: Calculate the seasonal coefficients of the series by averaging Y_S at different time points.

To forecast the value of the time series at a future time point, perform the following steps:

To forecast the value of the time series at a future time point, we perform the following steps:

Step 1: Calculate the trend component:

$$Y_{\text{Trend}} = B_0 + B_1 t_{\text{New}}$$

where t_{New} is a future time point.

Step 2: Calculate the seasonal component at time t_{New}, denoted as $S(t_{\text{New}})$.

Step 3: Calculate the value of Y at t_{New} as:

$$Y_{t_{\text{New}}} = Y_{\text{Trend}} + S(t_{\text{New}})$$

Time series forecasting has become an essential tool in both business management and health care by enabling organizations to anticipate future trends and make data-driven decisions.

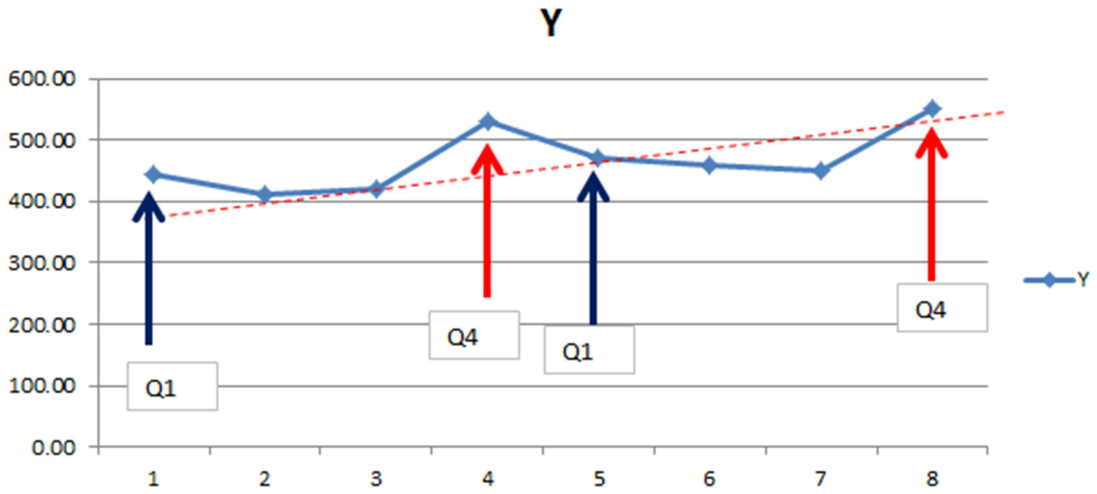
In business management, forecasting is widely used for inventory management, sales prediction, demand planning, cash flow analysis, and financial planning. For example, businesses leverage historical sales data to forecast future product demand, thus optimizing inventory levels and minimizing stockouts or overstock situations.

In health care, time series forecasting aids in predicting patient admissions, managing staff schedules, forecasting hospital resource needs, and monitoring disease outbreaks. Hospitals use

forecast models to anticipate seasonal surges in patient numbers or to prepare for potential epidemics, thereby ensuring better allocation of resources and improved patient care.

Example:

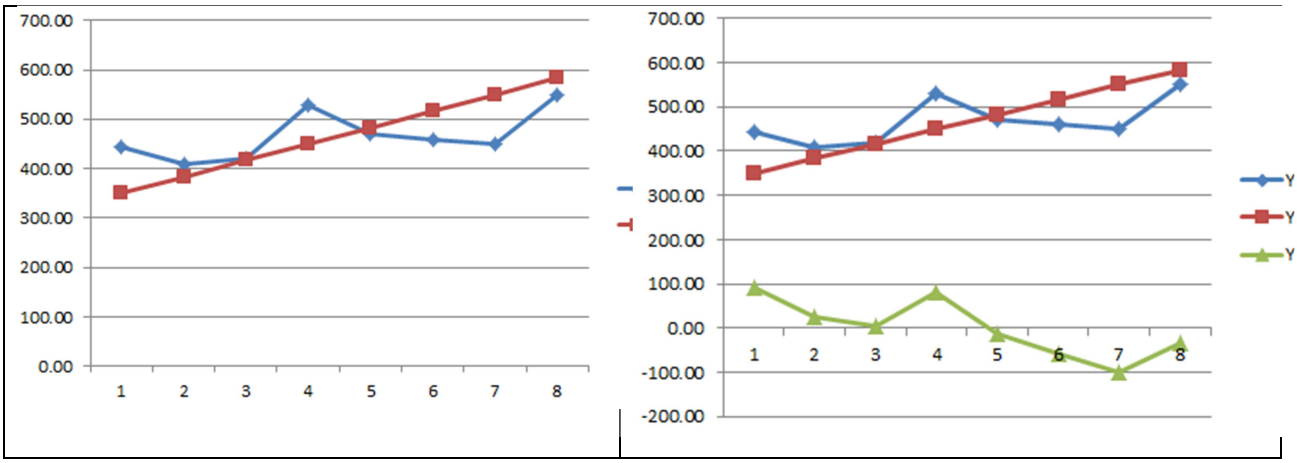
i	Year	Quarter	*t*	*Y*
1	1992	Q1	1.00	443.00
2	1992	Q2	2.00	410.00
3	1992	Q3	3.00	420.00
4	1992	Q4	4.00	530.00
5	1993	Q1	5.00	470.00
6	1993	Q2	6.00	460.00
7	1993	Q3	7.00	450.00
8	1993	Q4	8.00	550.00



Step 1: Construct a simple linear regression model to extract the trend component (assuming there is a linear relationship between *t* and *Y*)

$Y_{Trend} = B_0 + B_1 t$

$B_0 = 316{,}54$ và $B_1 = 33{,}35$.



Step 2: Isolate the seasonal component as follows:

$Y_S = Y - Y_{Trend}$

i	*t*	*Y*	*YTrend*	*Ys*
1	1.00	443.00	349.89	93.11
2	2.00	410.00	383.24	26.76
3	3.00	420.00	416.60	3.40
4	4.00	530.00	449.95	80.05
5	5.00	470.00	483.30	-13.30
6	6.00	460.00	516.65	-56.65
7	7.00	450.00	550.01	-100.01
8	8.00	550.00	583.36	-33.36

Step 3: Compute the seasonal coefficients as follows:

					Ys = Y - YTrend			
19	*i*	*t*	*Y*	*Y Trend*	*Ys*	Quarter		
20	1	1.00	443.00	349.89	93.11	Q1	Quarter	Seasonal Index
21	2	2.00	410.00	383.24	26.76	Q2	Q1	=AVERAGE(E20,E24)
22	3	3.00	420.00	416.60	3.40	Q3	Q2	
23	4	4.00	530.00	449.95	80.05	Q4	Q3	
24	5	5.00	470.00	483.30	-13.30	Q1	Q4	
25	6	6.00	460.00	516.65	-56.65	Q2		
26	7	7.00	450.00	550.01	-100.01	Q3		
27	8	8.00	550.00	583.36	-33.36	Q4		

AVERAGE(number1, [number2], [number3], ...)

Quarter	Seasonal Index
Q1	39.90
Q2	-14.95
Q3	-48.30
Q4	23.35

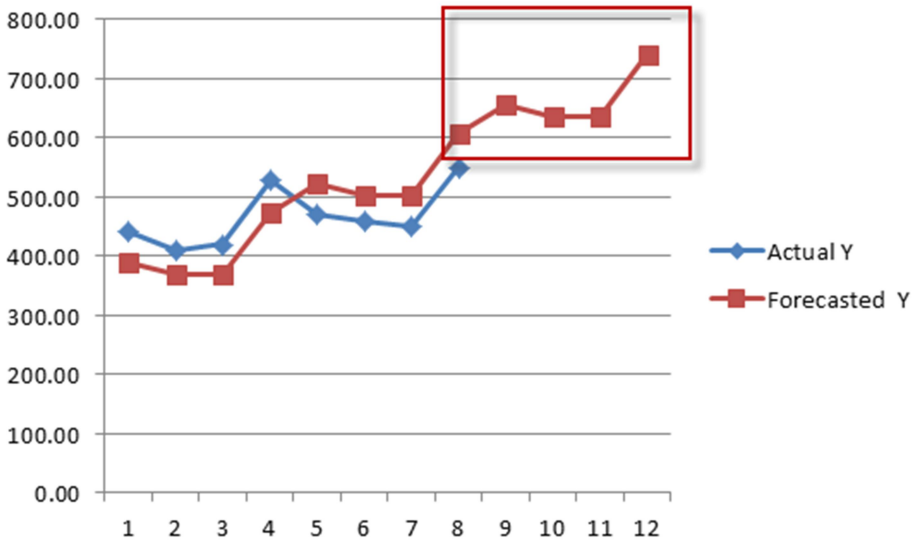
To forecast the value of Y, we carried out the steps as follows:

i	Year	Quarter	*t*	*Y*
1	1992	Q1	1.00	443.00
2	1992	Q2	2.00	410.00
3	1992	Q3	3.00	420.00
4	1992	Q4	4.00	530.00
5	1993	Q1	5.00	470.00
6	1993	Q2	6.00	460.00
7	1993	Q3	7.00	450.00
8	1993	Q4	8.00	550.00
9	1994	Q1	9.00	?
10	1994	Q2	10.00	?
11	1994	Q3	11.00	?
12	1994	Q4	12.00	?

15					b1 =	33.35
16					b0 =	316.54
17						
18	*i*	Year	Quarter	*t*	*Y Trend*	
19	1	1992	Q1	1.00	349.89	
20	2	1992	Q2	2.00	383.24	
21	3	1992	Q3	3.00	416.60	
22	4	1992	Q4	4.00	449.95	
23	5	1993	Q1	5.00	483.30	
24	6	1993	Q2	6.00	516.65	
25	7	1993	Q3	7.00	550.01	
26	8	1993	Q4	8.00	583.36	
27	9	1994	Q1	9.00	=$F$16+$F$15*D27	
28	10	1994	Q2	10.00	650.06	
29	11	1994	Q3	11.00	683.42	
30	12	1994	Q4	12.00	716.77	

i	Year	Quarter	*t*	*Y Trend*	Seasonal Component
1	1992	Q1	1.00	349.89	39.90
2	1992	Q2	2.00	383.24	-14.95
3	1992	Q3	3.00	416.60	-48.30
4	1992	Q4	4.00	449.95	23.35
5	1993	Q1	5.00	483.30	39.90
6	1993	Q2	6.00	516.65	-14.95
7	1993	Q3	7.00	550.01	-48.30
8	1993	Q4	8.00	583.36	23.35
9	1994	Q1	9.00	616.71	39.90
10	1994	Q2	10.00	650.06	-14.95
11	1994	Q3	11.00	683.42	-48.30
12	1994	Q4	12.00	716.77	23.35

Year	Quarter	*t*	*Y Trend*	Seasonal Component	Actual *Y*	Forecasted *Y*
1992	Q1	1.00	349.89	39.90	443.00	389.79
1992	Q2	2.00	383.24	-14.95	410.00	368.29
1992	Q3	3.00	416.60	-48.30	420.00	368.29
1992	Q4	4.00	449.95	23.35	530.00	473.29
1993	Q1	5.00	483.30	39.90	470.00	523.21
1993	Q2	6.00	516.65	-14.95	460.00	501.71
1993	Q3	7.00	550.01	-48.30	450.00	501.71
1993	Q4	8.00	583.36	23.35	550.00	606.71
1994	Q1	9.00	616.71	39.90		=E27+F27
1994	Q2	10.00	650.06	-14.95		635.12
1994	Q3	11.00	683.42	-48.30		635.12
1994	Q4	12.00	716.77	23.35		740.12



Multiple linear regression

Multiple linear regression is a statistical method used to model the relationship between one continuous dependent variable and two or more independent variables. These independent variables can be either numerical or categorical. The main objective is to estimate how each predictor influences the outcome while accounting for the other variables in the model.

The general form of a multiple linear regression equation is:

$$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

where $\hat{Y}$ is the predicted value, β_0 is the intercept, and $\beta_1, \beta_2, ..., \beta_p$ are the coefficients for the independent variables $x_1, x_2, ..., x_p$

Example: Suppose a retail company wants to forecast monthly sales for its stores using several business factors. A multiple linear regression model could use variables such as advertising spend, store floor space, and number of staff to predict sales.

The Regression Equation

Assume the estimated equation from the historical data is:

$$\text{Predicted Monthly Sales} = 8{,}500 + 20 \times \text{Advertising} + 15 \times \text{FloorSpace} + 250 \times \text{Staff}$$

- Predicted Monthly Sales: anticipated sales in USD (in thousands)
- Advertising: monthly advertising spend in USD (in thousands)
- FloorSpace: store size in square meters
- Staff: number of staff employed at the store

Interpreting the Equation

- Intercept ($8,500$): This represents the baseline sales.
- Advertising Coefficient (20): For every additional $1,000 spent on advertising, holding other factors constant, expected monthly sales increase by $20,000.
- FloorSpace Coefficient (15): For each additional 1m² of floor space, with other factors unchanged, sales rise by $15,000.
- Staff Coefficient (250): Hiring one more employee, with other factors held steady, is associated with a $250,000 increase in expected monthly sales.

Business Analytics Interpretation

- The coefficients show the marginal impact of each predictor: more advertising, more store space, and more staff are each independently associated with higher sales.
- Using this model, managers can estimate the expected sales impact of business decisions—like increasing ad budgets or expanding a store.

Practice

Question 1. Given the following data, calculate the Pearson correlation coefficient between Wind Speed and Workload.

Sample	Wind Speed (km/hr)	Workload (%)
1	8	97
2	10	95
3	12	93
4	14	90
5	16	92
6	18	85
7	20	86
8	20	83

Question 2.

Sample	Estimated cost (Bil. VND)	Estimated schedule (Năm)	Number of workers	Average experience of managers (year)	Cost overrun (%)
1	5.0	1.0	100	15	5
2	6.2	1.2	65	12	5
3	12.6	2.7	70	2	15
4	10.8	1.5	120	3	12
5	11.5	1.7	60	10	9
6	13.3	2.8	76	8	12
7	13.6	2.6	75	11	10
8	14.7	3.6	90	8	15

a. Calculate the Pearson correlation coefficients between the factors (total estimated cost, estimated completion time, number of workers, and management team experience) and the cost overrun level.

b. To reduce the occurrence of cost overruns in future projects, what strategies should we implement, and which project factors should we focus on analyzing?

Question 3.

Sample	Wind speed (m/s)	Productivity (m^2/Hr)
1	3.5	40.4
2	3.8	40.3
3	4.3	41.0
4	4.7	39.0
5	6.5	32.0
6	3.8	42.4
7	4.5	44.0
8	5.0	30.0

a. Building a Univariate Regression Model to Forecast Construction Productivity from Wind Speed.

a. With a fitted regression model, you can predict the construction productivity for any given wind speed by substituting the wind speed values into the model equation. For instance, to forecast productivity for wind speeds of 3.6, 3.9, and 4.0

Question 4

Predicting Employee Satisfaction from Training Hours

A company wants to understand the relationship between the number of training hours completed by employees and their satisfaction scores (rated from 1 to 10). By collecting data from various employees, we can apply simple linear regression and calculate prediction intervals.

training_hours = [5, 8, 10, 12, 14, 16, 18, 20, 22, 25]

satisfaction_scores = [4.5, 5.1, 6.2, 6.8, 7.0, 7.5, 8.0, 8.4, 8.7, 9.2]

a. Construction a linear regression model to predict satisfaction_scores from hours.

b. Perform interval estimation using the model with 95% confidence interval (MSR = 0.072 and t_crit = 2.306).

Question 5. A construction company collects a dataset regarding the floor area (1000 m^2) and total cost (Bil. VND) as follows:
The floor area (1000 m^2): 0.5, 0.8, 1.7, 2.2, 2.5
The total cost (Bil. VND): 9, 12, 17, 45, 66
a. Build a simple linear regression to predict the total cost based on the floor area. Compute the average error of the model.

b. Build a simple linear regression using the transformation $x_t = \exp(x)$; assess the prediction accuracy of the newly built model.

Question 6.

A contractor supplying concrete recorded the quantity of fresh concrete ordered (m^3) as follows:

i	Year	Quarter	*t*	*Y (1000 m^3)*
1	2018	Q1	1	30
2	2018	Q2	2	40
3	2018	Q3	3	50
4	2018	Q4	4	70
5	2019	Q1	5	45
6	2019	Q2	6	60
7	2019	Q3	7	80
8	2019	Q4	8	120

a. Build a forecasting model.

b. Predict the concrete orders for 2020.

c. Perform interval prediction of the ordered quantity of fresh concrete for 2020.

Question 7.

The value of construction steel (in 1,000 VND) is provided as follows:

a. Build a forecasting model.

b. Predict the steel price for 2020.

c. Perform interval prediction of the steel price for 2020.

i	Year	Quarter	*t*	*Y (1000 VND)*
1	2018	Q1	1	290
2	2018	Q2	2	300
3	2018	Q3	3	350
4	2018	Q4	4	310
5	2019	Q1	5	375
6	2019	Q2	6	380
7	2019	Q3	7	435
8	2019	Q4	8	410

Question 8.

a) Load the WestRoxbury house price dataset and build a multiple linear regression model to predict property values (TOTAL VALUE) using housing attributes.

b) Using 80% data for the training phase and evaluate the model's performance in the testing phase.

c) Using a scatterplot to visualize the prediction result.

The employed libraries and functions are given below:

Library	Purpose	Required For
pandas	Data manipulation and analysis	Data loading/preprocessing
scikit-learn	Model training, splitting, evaluation (includes `LinearRegression`, `train_test_split`, `mean_squared_error`, `r2_score`)	Model construction, evaluation
matplotlib.pyplot	Data visualization (e.g., scatter plot)	Plotting

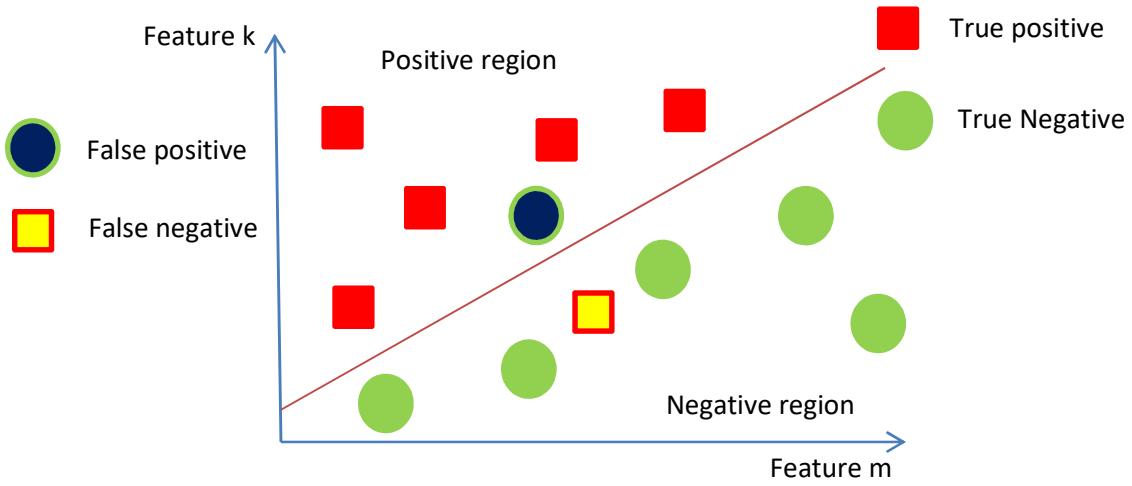
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
```

Section 4

Assessment of a classifier's performance

Assessing a classifier's performance is fundamental in machine learning and involves quantitatively evaluating how well the model distinguishes between classes. Key metrics for this assessment include the false positive rate (the proportion of negative cases incorrectly classified as positive) and the false negative rate (the proportion of positive cases incorrectly classified as negative).

A false positive occurs in binary classification when a test wrongly shows that a condition exists (for example, indicating someone has a disease when they do not). In contrast, a false negative is when a test wrongly suggests the condition is absent even though it is actually present. These represent the two types of mistakes possible in a binary test, as opposed to the two correct outcomes—a true positive and a true negative.



Example: A bank develops a model to decide whether to approve or decline loan applications based on customer information (like income, employment, and credit score). For each new applicant, the model predicts one of two outcomes:

- **Approved** (positive class)
- **Declined** (negative class)

After running the model on a batch of loan applications with known outcomes, we want to see how well the model performed compared to actual decisions.

A confusion matrix is a table used to evaluate the performance of a binary (or multiclass) classification model by comparing its predicted labels to the actual true labels.

Confusion Matrix Structure for Binary Classification

	Predicted: Approved (Positive)	Predicted: Declined (Negative)
Actual: Approved	True Positive (TP)	False Negative (FN)
Actual: Declined	False Positive (FP)	True Negative (TN)

- True Positive (TP): Model correctly predicts "Approved" when the actual status is "Approved".
- False Positive (FP): Model predicts "Approved" but the actual status is "Declined" .
- False Negative (FN): Model predicts "Declined" but the actual status is "Approved".
- True Negative (TN): Model correctly predicts "Declined" when the actual status is "Declined".

Suppose we have test results from 20 loan applications:

- 8 were actually approved, 12 were actually declined.
- The model predicted as follows:

Actual\Predicted	**Approved**	**Declined**	**Total Actual**
Approved	6 (TP)	2 (FN)	8
Declined	3 (FP)	9 (TN)	12
Total Predicted	9	11	20

Interpretation:

- **6** applicants were *correctly approved* (TP)
- **9** applicants were *correctly declined* (TN)
- **3** applicants were *incorrectly approved* (FP)
- **2** applicants were *incorrectly declined* (FN)

The confusion matrix provides a picture of how well the classifier performs.

Example 1

A diagnostic test is considered positive when it indicates that a condition exists, and negative when it shows the condition is not present. To evaluate how reliable a diagnostic test is, one approach is to look at the likelihood of two kinds of mistakes:

- False positive: The test indicates the condition is present when it is actually absent.
- False negative: The test indicates the condition is absent when it is actually present.

Suppose that there was a study involving 5,055 women aged 35 and older evaluated the accuracy of the Triple Blood Test. The results showed that out of 55 cases of Down syndrome, the test would have detected 50.

Down Syndrome Status	POS	NEG	Total
D (Down syndrome present)	50	5	55
ND(unaffected, no Down)	1000	4000	5000
Total	1,050	4,005	5,055

Test Results for Women with Down Syndrome (D):

- 50 women with Down syndrome received a positive test result (true positives).
- 5 women with Down syndrome received a negative test result (false negatives).

Test Results for Women without Down Syndrome (DC):

- 1,000 unaffected women received a positive test result (false positives).
- 4,000 unaffected women received a negative test result (true negatives).

Overall Totals:

- 1,050 women had a positive test result (includes both true positives and false positives).
- 4,005 women had a negative test result (includes both true negatives and false negatives).
- Total women tested: 5,055.

True Positive Rate:

Of 55 women who actually had Down syndrome, 50 tested positive.

True Positive Rate = $50/55 \approx 91\%$

(The test correctly identifies about 91% of Down syndrome cases.)

False Positive Rate:

Of 5,000 women without Down syndrome, 1,000 tested positive by mistake.

False Positive Rate = $1,000/5,000 = 20\%$

(About 20% of unaffected women get a false positive.)

True Negative Rate:

Of 5,000 unaffected women, 4,000 got a correct negative result.

True Negative Rate = 4,000/ 5,000 = 80%

False Negative Rate:

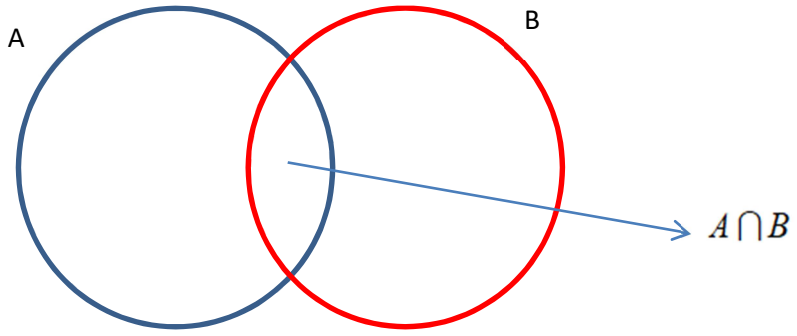
Of 55 women with Down syndrome, 5 got a negative result.

False Negative Rate = 5/55 ≈ 9.1%.

The test is good at detecting Down syndrome. However, it also incorrectly flags many unaffected pregnancies as high risk (a high false positive rate of 20%). While a positive test increases the chance that a fetus has Down syndrome, it does not guarantee it—a large number of positive results are in unaffected pregnancies. This means the test is useful for screening, but any positive result should be followed up with more accurate diagnostic procedures before making decisions.

Conditional probability

Conditional probability refers to the likelihood of an event occurring, given that another related event has already occurred. This concept enables us to update our predictions or understanding based on additional information, which is fundamental in disciplines like statistics, machine learning, and data science.



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Applications of conditional probability are pervasive: it underpins spam filtering (estimating the chance that an email is spam based on its content), disease diagnosis (calculating the probability of a disease given a positive test result), and fraud detection (assessing risk given prior transactions), among many others.

In data science, conditional probability is especially important for model evaluation, Bayesian inference, and feature selection, helping practitioners reason logically about uncertainty and dependencies in data.

An example of conditional probability in business management is the probability that a customer will complete a purchase, given that they have added items to their online shopping cart. Let A be the event "customer completes purchase," and B be the event "customer adds items to cart." The conditional probability P(A|B) measures how likely a purchase is, knowing the customer has initiated the process by adding items.

Example 2

For instance, an e-commerce company determines from its data that 30% of all site visitors add items to their cart (P(B)=0.3), and 12% of all visitors both add items to their cart and complete the purchase (P(A∩B)=0.12). The conditional probability of purchase given items in the cart is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.12}{0.3} = 0.4$$

This means there is a 40% chance a customer will complete a purchase after adding items to the cart. Businesses use this information to improve conversion rates, such as by sending reminder emails or offering discounts to encourage purchase completion.

Example 3

Consider a business that sells two products: Product A and Product B. The marketing team wants to know how likely it is for a customer to purchase Product A if they've already bought Product B. This information can help with targeted cross-selling.

Let's define:

Event A: Customer purchases Product A.

Event B: Customer purchases Product B.

Suppose the data shows:

40% of customers buy Product B (P(B)=0.4).

30% of customers buy both Product A and Product B (P(A∩B)=0.3).

The conditional probability that a customer buys Product A given that they bought Product B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.3}{0.4} = 0.75$$

There is a 75% chance that a customer who purchased Product B will also purchase Product A. This insight allows the marketing team to focus cross-sell efforts on customers who already bought Product B, maximizing the likelihood of additional sales.

Example 4

Down Syndrome Status	POS	NEG	Total
D (Down syndrome present)	50	5	55
ND(unaffected, no Down)	1000	4000	5000
Total	1,050	4,005	5,055

Based on the data, compute the probabity that a Down syndrome given a positive test outcome.

$$P(Down \mid Pos) = \frac{P(Down \cap Pos)}{P(Pos)}$$

P(Pos) = 1050/5055 = 0.21

$P(Down \cap Pos)$ = 50/5055 = 0.0989

$$P(Down \mid Pos) = \frac{0.0989}{0.21} = 0.048$$

In summary, of the women who tested positive, 4.8% actually had fetuses with Down syndrome.

Example 4

Risk Analysis (Credit Scoring)

Let's consider an example of a bank's customer dataset regarding loan defaults and income levels.

Customer	Income	Prior Defaults	Defaulted on Loan?
1	700	Yes	Yes
2	1,200	No	No
3	900	No	Yes
4	1,700	No	No
5	1,500	Yes	No
6	800	Yes	No
7	1,600	No	No
8	600	Yes	Yes
9	1,300	Yes	Yes
10	950	No	No

How to compute the probabilityof default for a customer given his income < 1000 and his record of prior default?

We need to compute P(Default | Income < 1000, Prior Default)

First, let find out who satisfies the criteria: Criteria: Income < $1000 and Prior Defaults = Yes

Customer	Income	Prior Defaults	Defaulted?
1	$700	Yes	Yes
6	$800	Yes	No
8	$600	Yes	Yes

Then, we count the number of defaults among the subgroup.

Calculate Conditional Probability

$$P(\text{Default} \mid \text{Income} < \$1000, \text{Prior Defaults} = \text{Yes}) = \frac{2}{3} \approx 0.67$$

Thus, there is a 67% probability that a customer with income less than $1000 and a history of prior defaults will default on their loan, based on this dataset.

Bayes' Theorem

Bayes' Theorem is a statistical method used to update probability estimates when new information or evidence becomes available, enabling more precise and effective decision-making in uncertain situations.

It is frequently applied in evaluating probabilities related to medical diagnoses—for example, determining the likelihood that a patient will develop cancer after receiving specific test results. This approach is particularly valuable in medicine for understanding the effects of false positives, which occur when a test incorrectly indicates a patient has a condition they actually do not have. Bayes' Theorem provides a way to compute conditional probabilities as follows:

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(B)}$$

$$P(A|B) = \frac{P(A)\cdot P(B|A)}{P(A)\cdot P(B|A)+P(A')\cdot P(B|A')}$$

Example 1

Suppose that 3% of people have a certain kind of diseases. Let's refer to "Event A" as the situation where someone has this disease. Thus, P(A) = 0.03.

A screening test is available for this cancer. This screening test has the following properties:

- The probability that the test correctly identifies someone with the disease is 75%. Thus, P(Pos|A) = 0.75.
- Additionally, the probability that an individual has a positive test result given that he/she does not have the disease is 0.15. Thus, P(Pos| not A) = 0.15.

We are interested in computing the probability P(A|Pos), which quantifies the likelihood that a patient actually has the disease given that the screening test shows a positive result.

$$P(A \mid Pos) = \frac{P(A).P(Pos \mid A)}{P(A).P(Pos \mid A) + P(notA).P(Pos \mid notA)} = \frac{0.03 \times 0.75}{0.03 \times 0.75 + 0.97 \times 0.15} = 0.1339$$

This outcome from Bayes' Theorem shows that, despite a positive result on the screening test, the probability that the patient actually has the disease is still low. In fact, there is a 13.39% chance that the patient has disease when the test result is positive.

Example 2

A medical researcher collected the following data:

Age Group	Gets Flu	Does Not Get Flu	Total
Senior (age ≥ 60)	500	1500	2000
Under 60	2000	6000	8000
Total	2500	7500	10000

Let use the following notation:

- X: Senior person
- Y: Not senior person
- F: Getting the flu

We have the following facts:

- The probability that a person is a senior: P(X) = 2000/10000 = 0.20
- The probability that a senior gets the flu P(F|X) = 500/2000 = 0.25
- The probability that a person is under 60: P(Y) = 8000/10000 = 0.80
- The probability that a person under 60 gets the flu: P(F|Y) = 2000/8000= 0.25
- The probability that a person get the flu: P(F) = 2500/10000 = 0.25

Hence, the probability that a person is a senior given he/she gets the flu is:

$$P(X \mid F) = \frac{P(F \mid X) \times P(X)}{P(F)} = \frac{0.25 \times 0.20}{0.25} = 0.20$$

The probability that a person is under 65 given he/she gets the flu is:

$$P(Y \mid F) = \frac{P(F \mid Y) \times P(Y)}{P(F)} = \frac{0.25 \times 0.80}{0.25} = 0.80$$

The Naive Bayes Classifier with one feature

The naive Bayesian classifier is a family of simple yet effective probabilistic classification algorithms based on Bayes' theorem, designed to assign class labels to data points represented as feature vectors. The defining characteristic of this method is its core assumption that all input features are conditionally independent given the class label—a simplification known as the naive independence assumption, which rarely holds true in real-world data but nonetheless enables efficient computation and surprisingly robust performance in practice. Despite its simplicity, the naive Bayesian classifier is highly scalable, requires minimal parameter tuning, and has found wide application in various domains.

Here, we are interested in the probability that a data point belongs to a certain class C_i, given its set of predictor values $x_1, x_2, ..., x_p$. More generally, if there are m possible classes ($C_1, C_2, ..., C_m$) and each data point has predictors ($x_1, x_2, ..., x_p$), our goal is to calculate $P(C_i \mid x_1, ..., x_p)$. To carry out classification, we calculate the chance of the record falling into each of the possible classes using this approach. The record is then assigned to the class with the highest probability, or a predefined probability threshold may be used to decide if it should be classified as the class of interest.

Example 1: We consider a simple example where we want to classify whether an email is "Spam" or "Not Spam" based on just one discrete input feature: whether the word "Offer" appears in the email.

Email	"Offer" Appears?	Class
1	Yes	Spam
2	Yes	Spam
3	No	Not Spam
4	Yes	Not Spam
5	No	Not Spam

We'd like to compute the posterior probability $P(Spam \mid 'Offer' = Yes)$ as follows:

$$P(Spam \mid 'Offer' = Yes) = \frac{P('Offer' = Yes \mid Spam) \times P(Spam)}{P('Offer' = Yes)}$$

$$P('Offer' = Yes) = P('Offer' = Yes \mid Spam) \times P(Spam) + P('Offer' = Yes \mid NotSpam) \times P(NotSpam)$$

Based on the table, we have the following fact:

$P(Spam) = 2/5$

$P(NotSpam) = 3/5$

$P('Offer' = Yes \mid Spam) = 2/2 = 1$

$P('Offer' = Yes \mid NotSpam) = 1/3$

$$P(Spam \mid 'Offer' = Yes) = \frac{1 \times (2/5)}{P('Offer' = Yes)}$$

$P('Offer' = Yes) = 1 \text{x} (2/5) + (1/3)\ \text{x} (3/5) = 0.6$

$$P(Spam \mid 'Offer' = Yes) = \frac{1 \times (2/5)}{0.6} = 0.67$$

Example 2: Herein, we want to predict whether a customer will buy a premium product based on their membership status (a discrete feature: "Member" or "Non-Member").

Customer	Membership Status	Purchased Premium Product?
1	Member	Yes
2	Non-Member	No
3	Member	Yes
4	Member	No
5	Non-Member	No

We'd like to compute the posterior probability $P(Yes \mid Member)$ and $P(No \mid Member)$ as follows:

$$P(Yes \mid Member) = \frac{P(Member \mid Yes) \times P(Yes)}{P(Member)}$$

$$P(No \mid Member) = \frac{P(Member \mid No) \times P(No)}{P(Member)}$$

Thus, we can express the two quantities of interest as follows:

$P(Yes \mid Member) \propto P(Member \mid Yes) \times P(Yes)$

$P(No \mid Member) \propto P(Member \mid No) \times P(No)$

where the symbol '∝' means "is proportional to."

P(Yes) = 2/5

P(No) = 3/5

P(Member|Yes) = 2/2 = 1

P(Member|No) = 1/3

Thus, P(Yes|Member) ∝ 1 x 2/5 = 2/5 = 0.4

Thus, P(No|Member) ∝ (1/3) x (3/5) = 0.2

Since P(Yes|Member) > P(No|Member), given an input feature = 'Member', the predicted class output is 'Yes'.

The Naive Bayes Classifier with multiple features

We can construct a Naive Bayes classifier for multiple features under the assumption that features are conditionally independent given the class label.

Bayes' theorem for multiple features generally states that:

$$P(Y \mid X_1, X_2) = \frac{P(X_1, X_2 \mid Y) \cdot P(Y)}{P(X_1, X_2)}$$

With the naive assumption, the joint likelihood can be expressed as follows:

$$P(X_1, X_2 \mid Y) = P(X_1 \mid Y) \cdot P(X_2 \mid Y)$$

Example: A manager wants to predict whether a customer will respond to a marketing campaign, based on:

- Membership Status: Member or Non-Member
- Service Used: Frequent or Infrequent

The collected data is presented as follows:

Customer	Membership Status	Service Used	Responded?
1	Member	Frequent	Yes
2	Member	Infrequent	No
3	Non-Member	Infrequent	No
4	Member	Frequent	Yes
5	Non-Member	Frequent	No
6	Member	Infrequent	Yes

We wish to compute:

- P(Yes | Member, Frequent) = ?
- P(No | Member, Frequent) = ?

P(Yes | Member, Frequent) $\propto$ P(Member, Frequent | Yes) x P(Yes)

$\propto$ P(Member| Yes) x P(Frequent | Yes) x P(Yes)

$\propto$ 1 x (2/3) x (3/6) = 0.33

P(No | Member, Frequent) $\propto$ P(Member, Frequent | No) x P(No)

$\propto$ P(Member| No) x P(Frequent | No) x P(No)

$\propto$ (1/3) x (1/3) x (3/6) = 0.06

Thus, P(Yes | Member, Frequent) > P(No | Member, Frequent), the output class of input feature = [Member, Frequent] is Yes.

Practice

Question 1.

Hepatitis C Antibody Screening Test

A diagnostic test is considered positive when it indicates that a condition exists, and negative when it shows the condition is not present.

Suppose a study was conducted involving 3,000 adults at elevated risk (due to medical history or exposures) to assess the accuracy of a screening blood test for Hepatitis C.

Hepatitis C Status	POS	NEG	Total
H (Hepatitis C present)	160	15	175
NH (No Hepatitis C)	250	2,575	2,825
Total	410	2,590	3,000

Compute the True Positive Rate, False Positive Rate, True Negative Rate, and False Negative Rate.

Question 2.

An e-commerce company tracks whether a customer visits the site from a marketing email and whether they make a purchase.

	Purchase: Yes	Purchase: No	Row Total
Email: Yes	3	2	5
Email: No	1	4	5
Col Total	4	6	10

Compute the probability that a customer makes a purchase, given they clicked the email.

Question 3.

A hospital records data on whether incoming patients reported chest pain and whether they were admitted for overnight observation.

Compute the probability that a patient reported chest pain given they were admitted.

	Admitted: Yes	Admitted: No	Row Total
Chest Pain: Yes	3	2	5
Chest Pain: No	2	3	5
Col Total	5	5	10

Question 4.

	Test POS	Test NEG	Total
Has Virus	450	50	500
No Virus	1000	8500	9500
Total	1500	8500	10000

Let use the following notation: V: has virus N: has no virus

Compute P(V), P(N), P(POS), P(POS|V), and P(V|POS)?

Question 5. Consider this dataset:

Customer	Income ($)	Prior Defaults	Defaulted on Loan?
1	1,809	Yes	No
2	1,001	Yes	No
3	1,885	No	No
4	1,709	No	Yes
5	947	Yes	Yes
6	1,649	Yes	No
7	1,616	No	No
8	1,069	No	Yes
9	826	No	Yes
10	818	Yes	No
11	709	Yes	Yes
12	1,204	No	No
13	1,994	No	No
14	1,275	Yes	Yes
15	1,787	No	No
16	1,682	Yes	Yes
17	1,854	Yes	Yes
18	663	No	No
19	1,278	Yes	No
20	1,247	Yes	No

Determine the probability that a customer with income less than $1,000 and a history of prior defaults will default on their loan.

Question 6.

A manager wants to predict if a customer will purchase a new product, using their membership status (the single discrete feature: "Member" or "Non-Member").

Customer	Membership Status	Purchase?
1	Member	Yes
2	Member	No
3	Non-Member	No
4	Member	Yes
5	Non-Member	No
6	Member	Yes

Determine the probabilities of P(Yes|Member) and P(No|Member).

Question 7.

A doctor wants to estimate the probability that a patient has a certain disease, based on the following symptoms:

- Fever: Yes or No
- Cough: Yes or No

Patient	Fever	Cough	Disease Present?
1	Yes	Yes	Yes
2	No	Yes	No
3	Yes	No	Yes
4	No	No	No
5	Yes	Yes	Yes
6	Yes	No	No

Compute P(Yes|Fever = Yes, Cough = Yes) and P(No|Fever = Yes, Cough = Yes).

Question 8.

An operations manager wants to estimate whether a machine needs maintenance based on the following features:

- Shift: Day or Night
- Machine Age: New or Old
- Noise Detected: Yes or No

Record	Shift	Machine Age	Noise Detected	Maintenance Needed?
1	Day	New	No	No
2	Night	Old	Yes	Yes
3	Day	Old	Yes	Yes
4	Night	New	No	No
5	Night	Old	Yes	Yes
6	Day	New	No	No
7	Day	Old	No	Yes
8	Night	New	Yes	No

Compute P(Yes|Night, Old, Yes) and P(No|Night, Old, Yes).