A closer look at the PACE model

Let's take a closer look at each stage of the PACE model.

Plan

At the beginning of a project, it is important to establish a solid foundation for success. Here you will define the scope of your project. This is when you will begin by identifying the informational needs of the organization. During the planning stage, you will have the widest viewpoint of a project. By assessing all of the factors and processes involved, you are mapping a path to completion, using your creativity to conceptualize a course of action. Here you will also take special note of tasks that may require an innovative approach within your workflow.

Summary: The planning stage is where you conceptualize the scope of the project and develop the steps that will guide you through the process of completing a project.

Here are a few examples of the types of planning stage tasks:

- Research business data
- Define the project scope
- Develop a workflow
- Assess project and/or stakeholder needs

Analyze

In the analyzing stage, you will interact with the data for the first time. Here you will acquire all of the data you will need for the project. Some datasets could come from primary sources within your organization. Others may need to be collected from secondary sources outside your company. You may even find that you need governmental or open source data. The analyzing stage is also where you will engage in exploratory data analysis or EDA. This involves cleaning, reorganizing and analyzing all of the necessary data for the project.

Summary: The analyzing stage is where you will collect, prepare, and analyze all of the data for your project.

Here are a few examples of the types of analyzing stage tasks:

- Format database
- Scrub data
- Convert data into usable formats

Construct

Just as the name suggests, the construction stage is all about building. In this stage of PACE, you will be building, interpreting, and revising models. Some projects will require machine learning algorithms to uncover correlations within your data. You will use these correlations to uncover information from the data that would otherwise go unused. These relationships can help your organization make informed decisions about the future.

Summary: In the construction stage you will build models that will allow you access to hidden relationships locked within data.

Here are a few examples of the types of construction stage tasks:

- Select modeling approach
- Build models
- Build machine learning algorithms

Execute

In the execution stage, you will put your analysis and construction into action. Here you will deliver your findings to the internal (inside of your organization) and external (outside of your organization) stakeholders. Quite often, this will involve stakeholders from the business-side of the companies you are working with. Presenting your findings is only a part of the execution stage. Stakeholders will provide feedback, ask questions, and make recommendations that you will collect and incorporate.

Summary: In the execution stage you will present the finding of your analysis, receive feedback, and make revisions as necessary.

Here are a few examples of execution stage tasks:

- Share results
- Present findings to other stakeholders
- Address feedback

Choose the workplace scenario you want to complete:

Features by Software	Python	R	Java	C++
Speed	Slower	Depends on configuration and add-ons	Faster	Very fast
Approachability	Easy to learn	Complex	Easy to learn	Complex
Variable	Dynamic	Dynamic	Static	Declarative
Data science focus	Machine learning and automated analysis	Exploratory data analysis and building extensive statistical libraries	Used across projects with open-source assets	Not as widely used but very powerful implementations
Programming Paradigm	Object-oriented	Functional language	Object-oriented	Multi-paradigm (imperative & object-oriented)

Review: Attributes and methods

Python classes are powerful and convenient because they come with built-in features that simplify common data analysis tasks. These features are known as attributes and methods.

- **Attribute**: A value associated with an object or class which is referenced by name using dot notation.
- **Method**: A function that belongs to a class and typically performs an action or operation.

A simpler way of thinking about the distinction between attributes and methods is to remember that attributes are *characteristics* of the object, while methods are *actions* or *operations*.

For example, if the class were Spaceship, then attributes might be:

`name`

`kind`

`speed`

`tractor_beam`

These attributes could be accessed by typing:

```
Spaceship.name
```

```
Spaceship.kind
```

```
Spaceship.speed
```

```
Spaceship.tractor_beam
```

Notice that these characteristics are accessed using only a dot.

On the other hand, methods of the Spaceship class might be:

```
warp()
```

```
tractor()
```

These methods could be used by typing:

```
Spaceship.warp()
```

```
Spaceship.tractor()
```

Notice that methods are followed by parentheses, and it's possible for them to take arguments. For example, `Spaceship.warp(7)` could change the speed of the ship to warp seven.

```python
class Spaceship:
    # Class attribute
    tractor_beam = 'off'

    # Instance attributes
    def __init__(self, name, kind):
        self.name = name
        self.kind = kind
        self.speed = None

    # Instance methods
    def warp(self, warp):
        self.speed = warp
        print(f'Warp {warp}, engage!')

    def tractor(self):
        if self.tractor_beam == 'off':
            self.tractor_beam = 'on'
            print('Tractor beam on.')
        else:
            self.tractor_beam = 'off'
            print('Tractor beam off')
```

```python
# Create an instance of the Spaceship class (i.e. "instantiate")
ship = Spaceship('Mockingbird','rescue frigate')

# Check ship's name
print(ship.name)

# Check what kind of ship it is
print(ship.kind)

# Check tractor beam status
print(ship.tractor_beam)
```

```
Mockingbird
rescue frigate
off
```

Refactoring

The process of restructuring code while maintaining its original functionality

Docstring

A string at the beginning of a function's body that summarizes the function's behavior and explains its arguments and return values

If we use the **or operator** the expression will be True if either of the expressions is true, and False only when both expressions are false.

The **and operator** needs both expressions to be true to return a True result.

The **not operator** inverts the value of the expression that follows it.

Comparators

In Python, you can use comparison operators to compare values. When a comparison is made, Python returns a Boolean result—True or False. Python uses the following comparators:

Operation	Operator
greater than	`>`
greater than or equal to	`>=`
less than	`<`
less than or equal to	`<=`
not equal to	`!=`
equal to	`==`

elif

A reserved keyword that executes subsequent conditions when the previous conditions are not true

```
In [ ]:
def hint_username(username):
    if len(username) < 8:
        print("Invalid username. Must be at least 8 characters long.")
    elif len(username) > 15:
        print('Invalid username. Cannot exceed 15 characters.')
    else:
        print('Valid username.')
```

range()

A Python function that returns a sequence of numbers starting from zero, increments by 1 by default, and stops before the given number

It can be used in while or for loops.

In Python, what type of loop iterates over a sequence of values?

- ○ Else loop
- ○ While loop
- ◉ For loop
- ○ If loop

✓ **Correct**
In Python, a for loop is a piece of code that iterates over a sequence of values.

Use **for loops** when there's a sequence of elements that you want to iterate over.

Use **while loops** when you want to repeat an action until a boolean condition changes.

Booleans are a data type that represents one of two possible states: **True** or **False**.

Correct

Python's `range()` function includes the following parameters: start value, stop value, and step value. The `range()` function returns a sequence of numbers starting from zero; then increments by one, by default; then stops before the given number.

The `range()` function

The `for` loop allows you to create a loop that performs exactly the number of iterations needed for the data structure you're looping over. In other words, whether your iterable sequence contains two, 1,000, or a million elements, you can use the same syntax and don't have to specify the number of iterations you want. However, sometimes you need to perform a task a set number of times, but you don't already have an iterable object to loop over. Or, sometimes you need to generate a known, regular sequence of numbers. This is where the `range()` function is useful.

The `range()` function is a function that takes three arguments: start, stop, step. Its output is an object belonging to the range class. If you only include one argument, it will be interpreted as the stop value. The start and step values by default will be zero and one, respectively. If you include two arguments, they will be interpreted as the start and stop values (again, with step being one by default). Note that the stop value is not included in the range that is returned.

Nested loops

Sometimes you'll need to extract information from nested structures—for example, from a list of lists. One way of doing this is by using nested loops. A nested loop is a loop inside of another loop. You can have an infinite number of nested loops, but it becomes more confusing to read and understand the more nested loops you add.

Here's an example of one loop nested in another:

```
1  students = [['Igor', 'Sokolov'], ['Riko', 'Miyazaki'], ['Tuva', 'Johansen']]
2  for student in students:
3      for name in student:
4          print(name)
5      print()
```

Run

Reset

```
Igor
Sokolov

Riko
Miyazaki

Tuva
Johansen
```

String slice

A portion of a string, also known as a substring, that can contain more than one character

```
In [9]: fruit = 'pineapple'
        fruit[:4]

Out[9]: 'pine'

In [10]: fruit[4:]

Out[10]: 'apple'

In [ ]:
```

To check whether or not a substring is contained in a string, use the keyword `in`.

```
In [11]: 'banana' in fruit

Out[11]: False

In [12]: 'apple' in fruit

Out[12]: True
```

Lists	Strings
Data structure	Data type
Allow duplicate elements	Allow duplicate elements
Allow indexing and slicing	Allow indexing and slicing
Sequences of elements	Sequences of characters

Note: When using parentheses to declare a tuple with just a single element, you must use a trailing comma.

```
1  test1 = (1)
2  test2 = (2,)
3
4  print(type(test1))
5  print(type(test2))
```

Run

Reset

```
<class 'int'>
<class 'tuple'>
```

STRINGS

- Single, double, or triple quotes:

```
empty_str = ''
my_string1 = 'minerals'
my_string2 = "martin"
my_string3 = """
marathon
golfcart
"""

```

Note: Using triple quotes to write a string over multiple lines will insert newlines (`\n`).

```
my_string3 = """
marathon
golfcart
"""

my_string3
```

```
marathon
golfcart
```

Content

- Tuples can contain any data type, and in any combination. So, a single tuple can contain strings, integers, floats, lists, dictionaries, and other tuples.

Note: The following code block is not interactive.

```
my_tuple = (1871, 'all', 'mimsy', ('were', 'the'), ['borogroves'])
```

Mutability

- Tuples are **immutable**. This means that once a tuple is created, it cannot be modified.

`intersection()`

A function that finds the elements that two sets have in common

`union()`

A function that finds all the elements from both sets

`difference()`

A function that finds the elements present in one set, but not the other

`symmetric_difference()`

A function that finds elements from both sets that are mutually not present in the other

iloc[]

A way to indicate in pandas that you want to select by integer-location-based position

loc[]

Used to select pandas rows and columns by name

Operator	Logic
&	and
|	or
~	not

- ✔ Must have more than 20 moons
- ✔ Must NOT have 80 moons
- ✔ Must NOT have a radius less than 50,000 km

```
mask = (planets['moons']>20)& ~(planets['moons']==80) & ~(planets['radius_km']<50000)
```

```
In [9]: mask = (planets['moons'] > 20) & ~(planets['moons'] == 80) & ~(planets['radius_km'] < 50000)
        planets[mask]
```

Out[9]:

	planet	radius_km	moons
5	Saturn	58232	83

groupby()

A pandas DataFrame method that groups rows of the dataframe together based on their values at one or more columns, which allows further analysis of the groups

Built-in aggregation functions

The previous examples demonstrated the `mean()`, `min()`, and `size()` aggregation functions applied to groupby objects. There are many available built-in aggregation functions. Some of the more commonly used include:

- `count()`: The number of non-null values in each group
- `sum()`: The sum of values in each group
- `mean()`: The mean of values in each group
- `median()`: The median of values in each group
- `min()`: The minimum value in each group
- `max()`: The maximum value in each group
- `std()`: The standard deviation of values in each group
- `var()`: The variance of values in each group

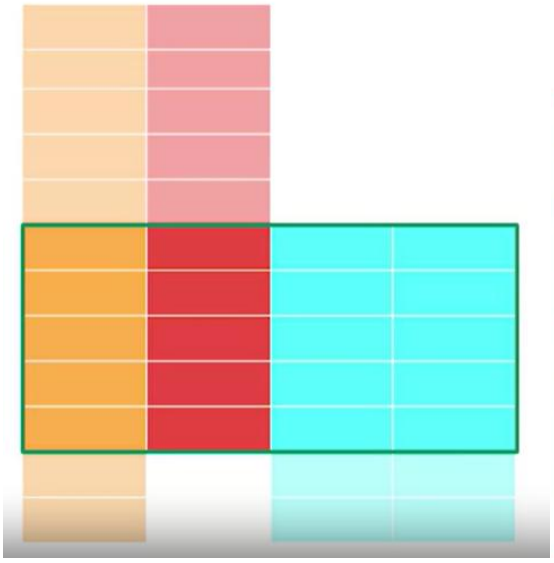
`merge()`

A pandas function that joins two dataframes together; it only combines data by extending along axis one horizontally

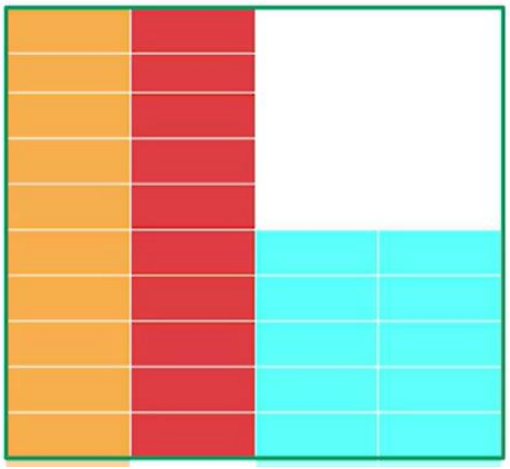
Keys

The shared points of reference between different dataframes—what to match on

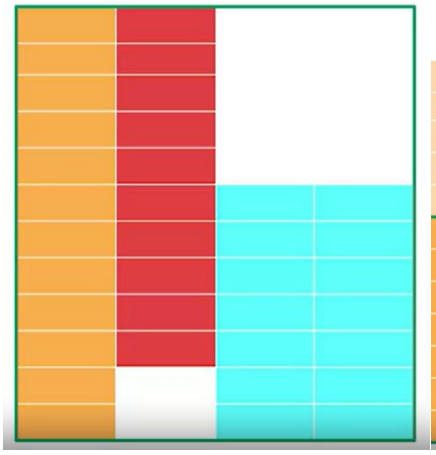
how='inner'



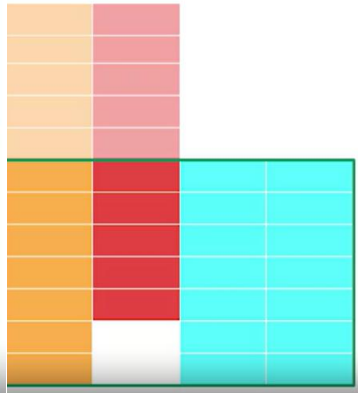
how='left'



how='outer'



how='right'



Discovering

Data professionals familiarize themselves with the data so they can start conceptualizing how to use it

Structuring

The process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled

this EDA practice is called structuring.

Bias (in data structuring)

Organizing data in groupings, categories, or variables that don't accurately represent the whole dataset

Cleaning

The process of removing errors that may distort your data or make it less useful

Joining

The process of augmenting or adjusting data by adding values from other datasets

Validating

The process of verifying that the data is consistent and high quality

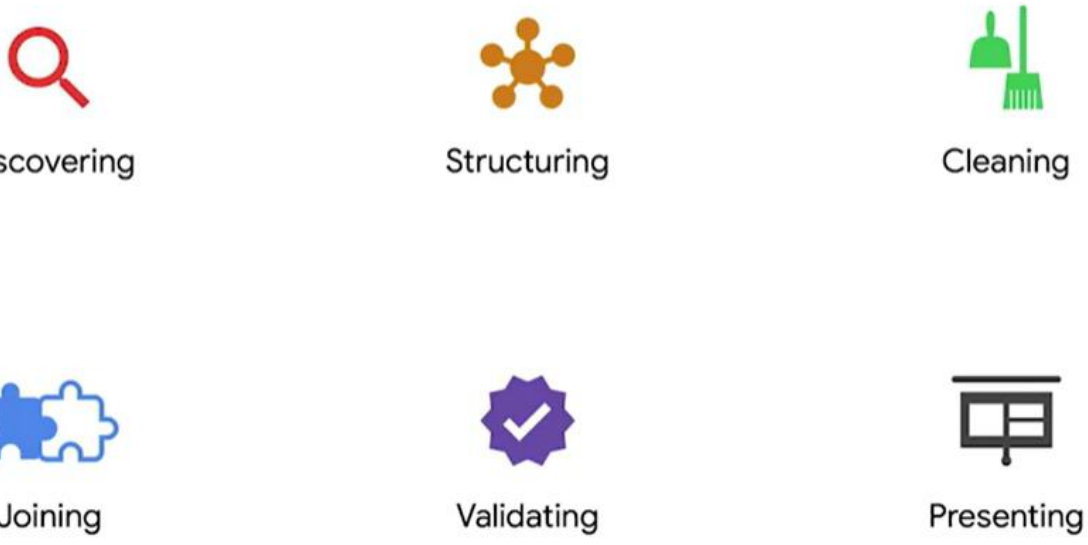
Presenting

Making your cleaned dataset or data visualizations available to others for analysis or further modeling

Data visualization

A graph, chart, diagram, or dashboard that is created as a representation of information

The 6 Practices of EDA



Types of data

- First-party data
- Second-party data
- Third-party data

First-party data

Data that was gathered from inside your own organization

Second-party data

Data that was gathered outside your organization but directly from the original source

Third-party data

Data gathered outside your organization and aggregated

Sorting

The process of arranging data into meaningful order for analysis

Extracting

The process of retrieving data from a dataset or source for further processing

Filtering

The process of selecting a smaller part of your dataset based on specified parameters and using it for viewing or analysis

Slicing

A method for breaking information down into smaller parts to facilitate efficient examination and analysis from different viewpoints

Merging

Method to combine two different data frames along a specified starting column

What to do with missing data

- Request the missing values be filled in by the owner of the data
- Delete the missing column(s), row(s), or value(s)
- Create a NaN category
- Derive new representative value(s)

3 types of outliers

- Global outliers
- Contextual outliers
- Collective outliers

Global outliers

Values that are completely different from the overall data group and have no association with any other outliers

Contextual outliers

Normal data points under certain conditions but become anomalies under most other conditions

Collective outliers

A group of abnormal points that follow similar patterns and are isolated from the rest of the population

Heatmap

A type of data visualization that depicts the magnitude of an instance or set of values based on two colors

Categorical data

Data that is divided into a limited number of qualitative groups

Label encoding

Data transformation technique where each category is assigned a unique number instead of a qualitative value

Input validation

The practice of thoroughly analyzing and double-checking to make sure data is complete, error-free, and high-quality

2. Fill in the blank: If a dataset lacks sufficient information to answer a business question, the process of _____ makes it possible to augment that data by adding values from other datasets.

- (●) joining
- ( ) summing
- ( ) sampling
- ( ) blending

Correct
If a dataset lacks sufficient information to answer a business question, the process of joining makes it possible to augment that data by adding values from other datasets. Joining is most useful if the new data is validated to ensure its format and data entries align and are the same data type as the original dataset.

Basic organizing strategies for a presentation

- Chronological
- Generic-to-specific
- Specific-to-generic

A chronological approach to data visualizations is useful for data that is best understood in a time series.

A generic-to-specific approach helps an audience consider an issue before describing how it affects them.

A specific-to-generic approach is useful to highlight impacts the data can have on a broader scale.

Questions you might get in interviews

- What is your process for cleaning data?
- What tools do you use for creating data visualizations?
- How and why do data visualizations enhance the stories data tells?
- What considerations are top of mind when sharing data stories with non-technical stakeholders?

Statistics

The study of the collection, analysis, and interpretation of data

A data professional might use probability to

- Predict the future rate of return on an investment
- Estimate the annual average sales revenue for a company
- Calculate the margin of error to quantify the uncertainty of an employee satisfaction survey
- Use percentiles to rank median home prices in different cities

A/B testing

A way to compare two versions of something to find out which version performs better

Sample

A subset of the larger population

Inferential statistics

Make inferences about a dataset based on a sample of the data

A/B test steps

- Analyzes a small group of users
- Decide on the sample size

Confidence interval

A range of values that describes the uncertainty surrounding an estimate

Statistical significance

The claim that the results of a test or experiment are not explainable by chance alone

Statistical methods

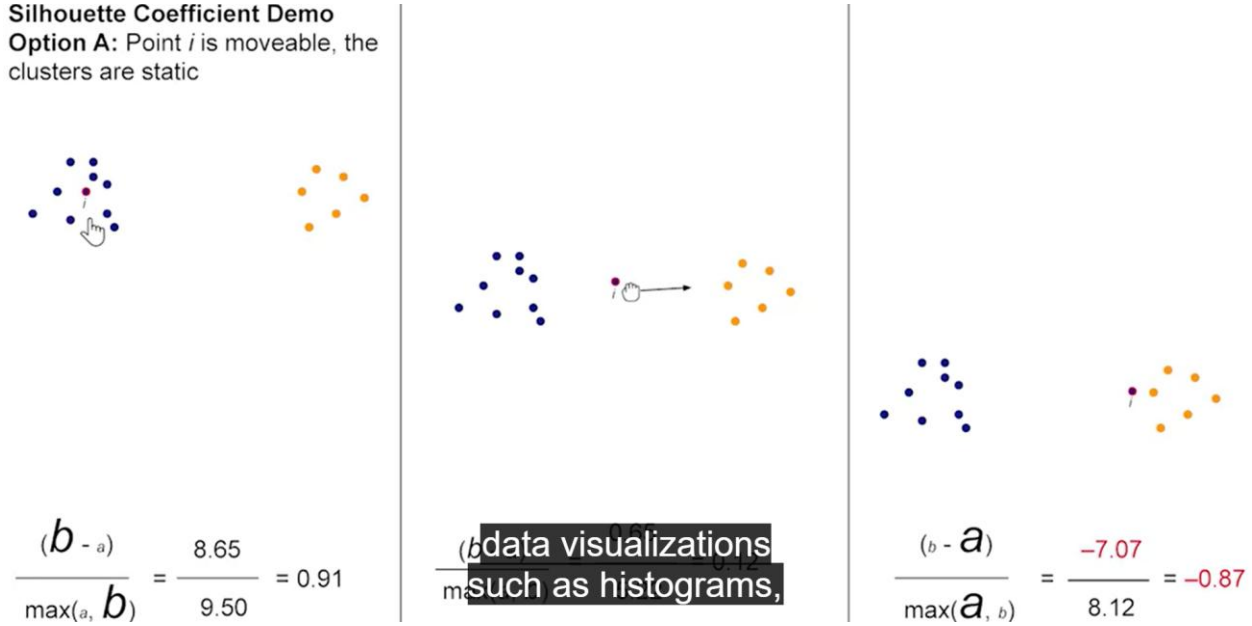
- Descriptive
- Inferential

Descriptive statistics

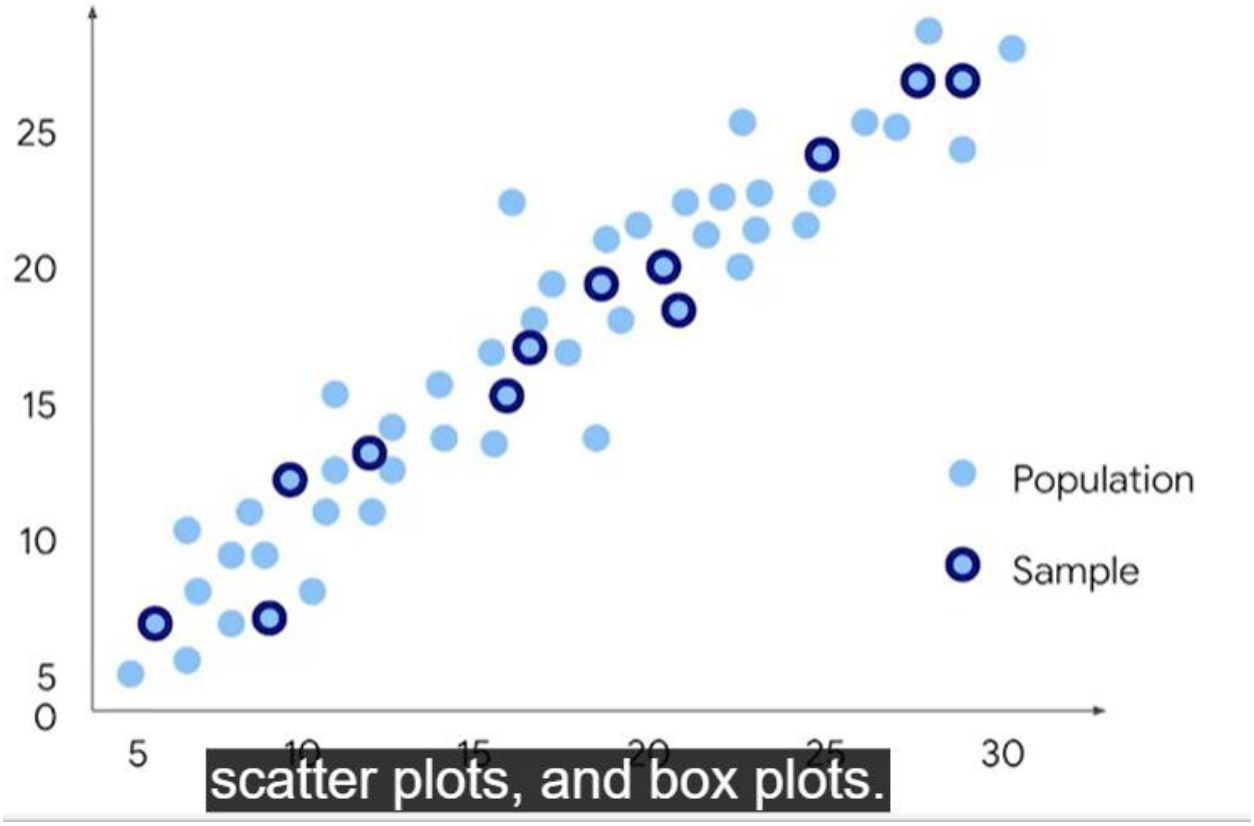
Describe or summarize the main features of a dataset

Forms of descriptive statistics

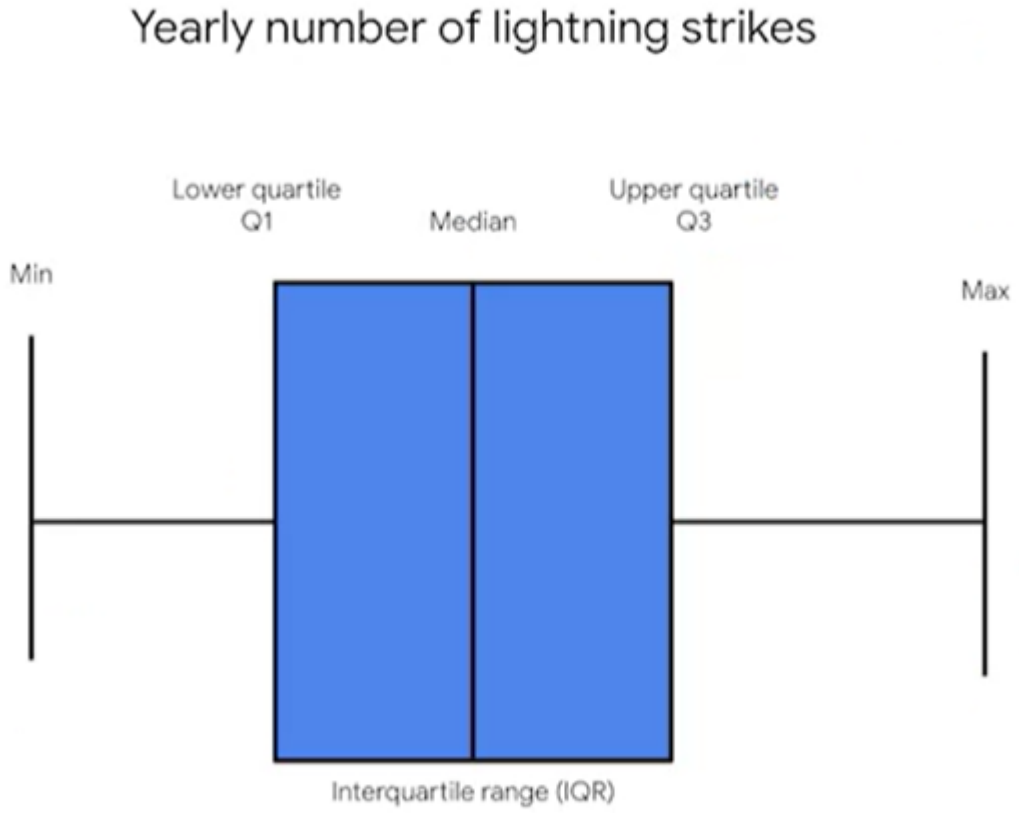
- Visuals, like graphs and tables
- Summary stats



data visualizations such as histograms,



scatter plots, and box plots.



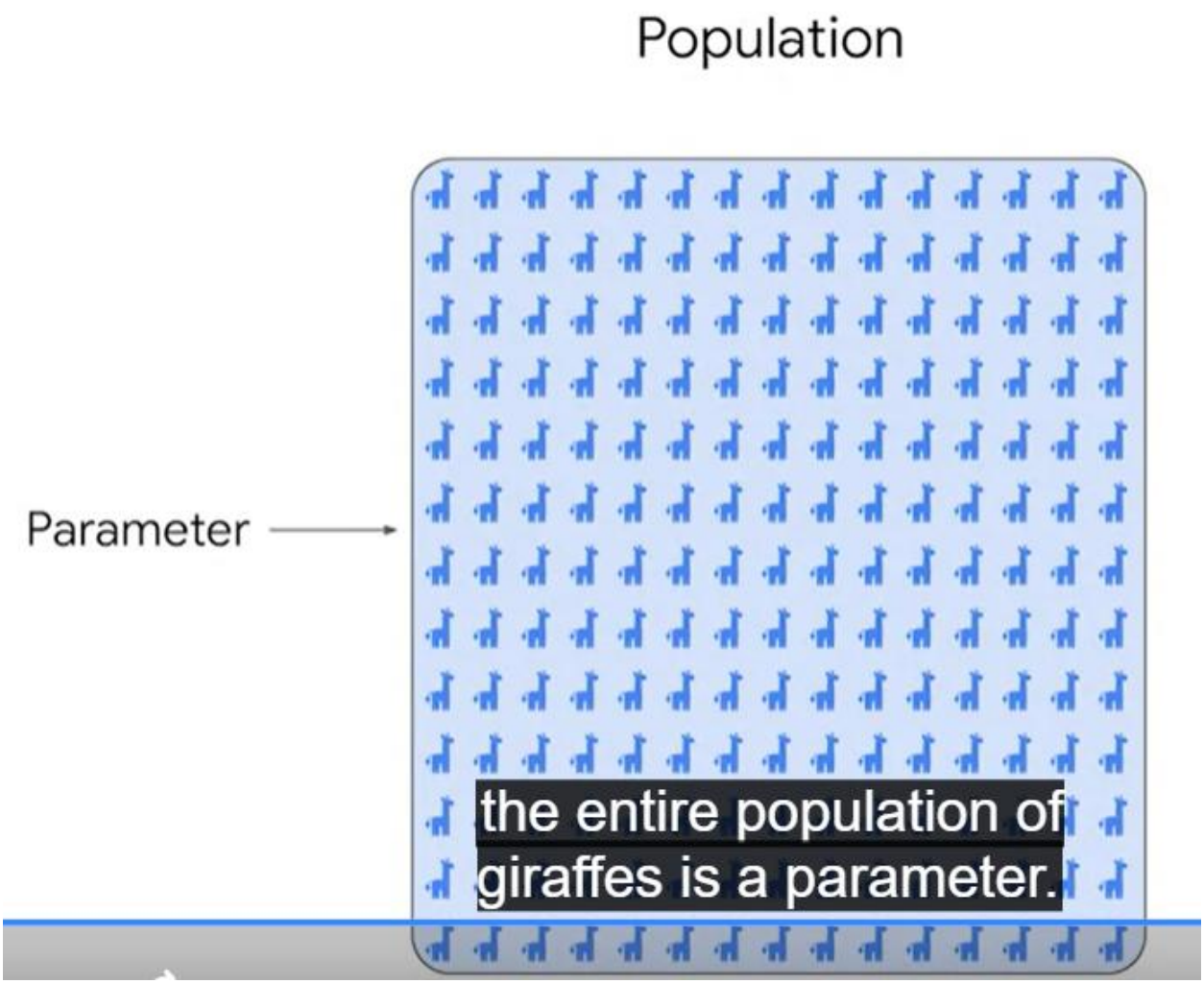
scatter plots, and box plots.

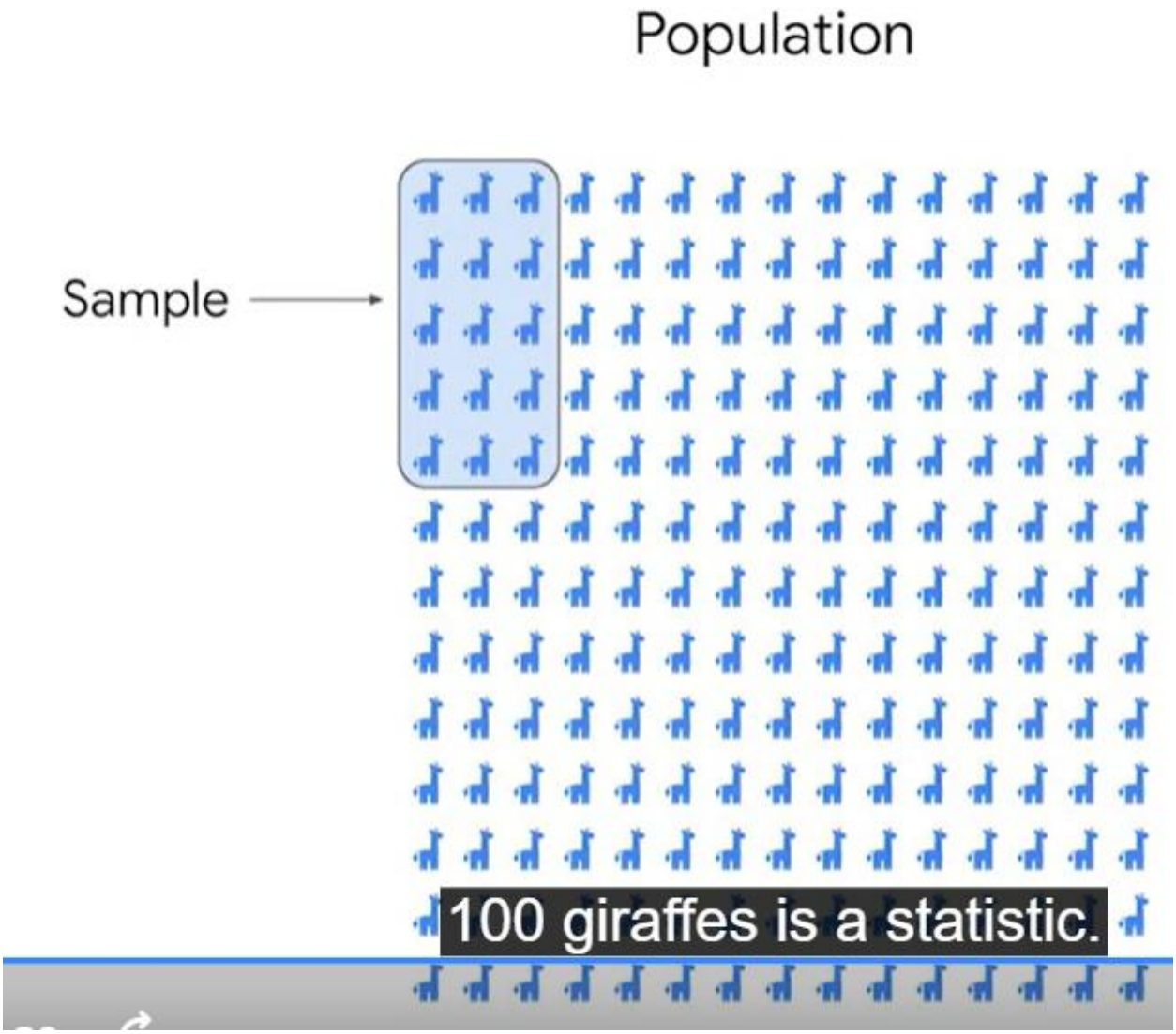
Parameter

A characteristic of a population

Statistic

A characteristic of a sample





Mean

The average value in a dataset

$$\text{Mean} = \frac{\text{(sum of all values)}}{\text{(total \#of all values)}}$$

Median

The middle value in a dataset

5, 7, **8**, 10, 70

If they're an even number
of values in your dataset,

the median is the average
of the two middle values.

5, 7, 8, 10, 70

4, 5, 7, 8, 10, 70

$7+8\div2=15\div2=\boxed{7.5}$

The median is their average, 7.5.

When to use median or mean

- If outliers, use median
- If no outliers, use mean

Home prices

- 9 homes = $100,000
- 1 home = $1,000,000 (outlier)
- Mean home price = $190,000

Home prices

- 9 homes = $100,000
- 1 home = $1,000,000 (outlier)
- Mean home price = $190,000
- Median home price = $100,000

The median gives you a much better idea of the typical value of a home in this

Mode

The most frequently occurring value in a dataset

- No mode: 1, 2, 3, 4, 5
- One mode: 1, 3, 3, 5, 7
- Two modes: 1, 2, 2, 4, 4

Range

The difference between the largest and smallest value in a dataset

Daily temperature (Fahrenheit)

77, 74, 72, 71, 67, 69, 72

Range: 77-67=10

So the range is 10.

Standard deviation

Measures how spread out your values are from the mean of your dataset

Variance

The average of the squared difference of each data point from the mean

Basically, it's the square of the standard deviation.

Sample standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

Sample standard deviation

$$s = \sqrt{\frac{\Sigma (x - \overline{10})^2}{n - 1}}$$

First, find the mean of the dataset.

The Greek letter sigma is a symbol that means sum.

Sample standard deviation

$$s = \sqrt{\frac{(8-10)^2 + (10-10)^2 + (12-10)^2}{3 - 1}}$$

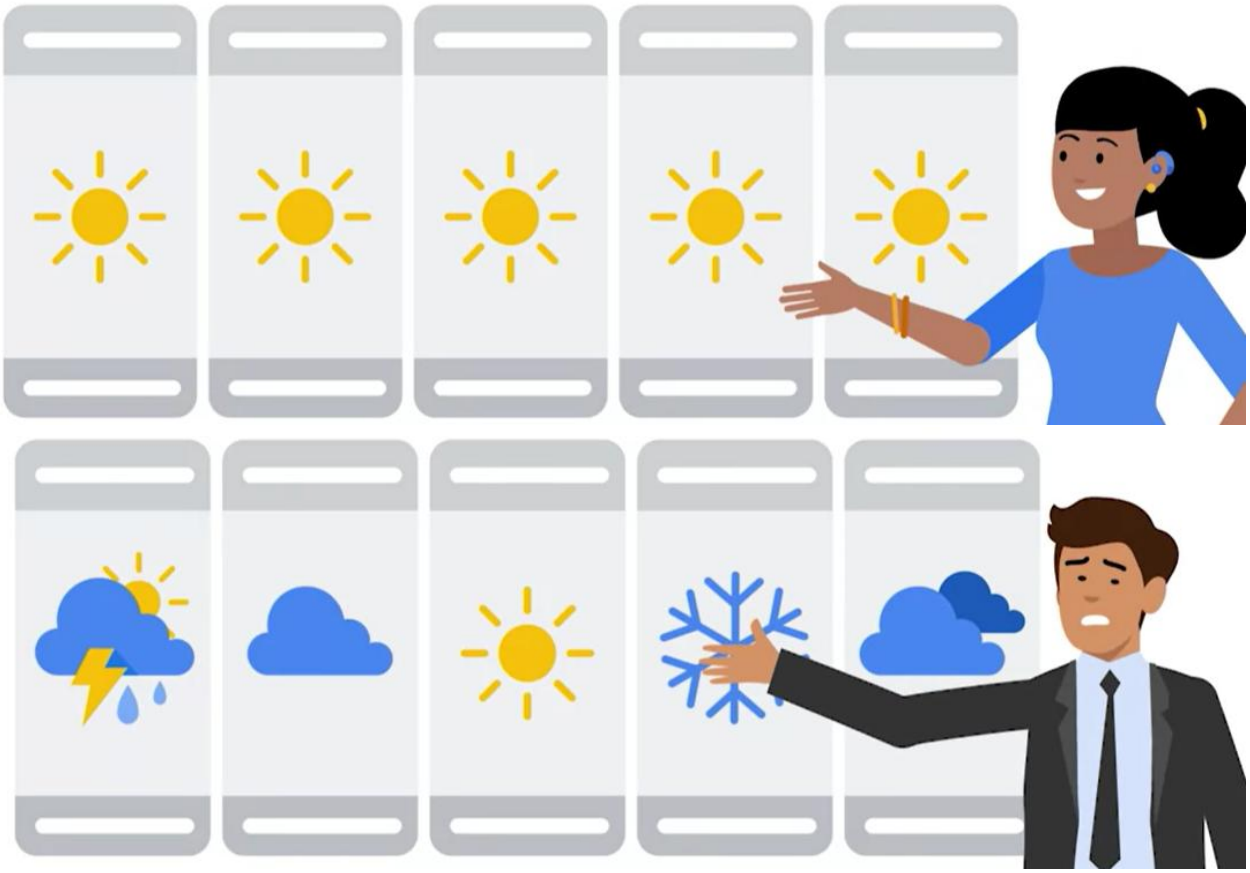
Sample standard deviation

$$s = \sqrt{4}$$

Then our sum of 8 divided by 2 equals 4.

Weather for March

	City A	City B
Mean Temperature	66 degrees	64 degrees
Standard Deviation	3 degrees	16 degrees



Measures of position

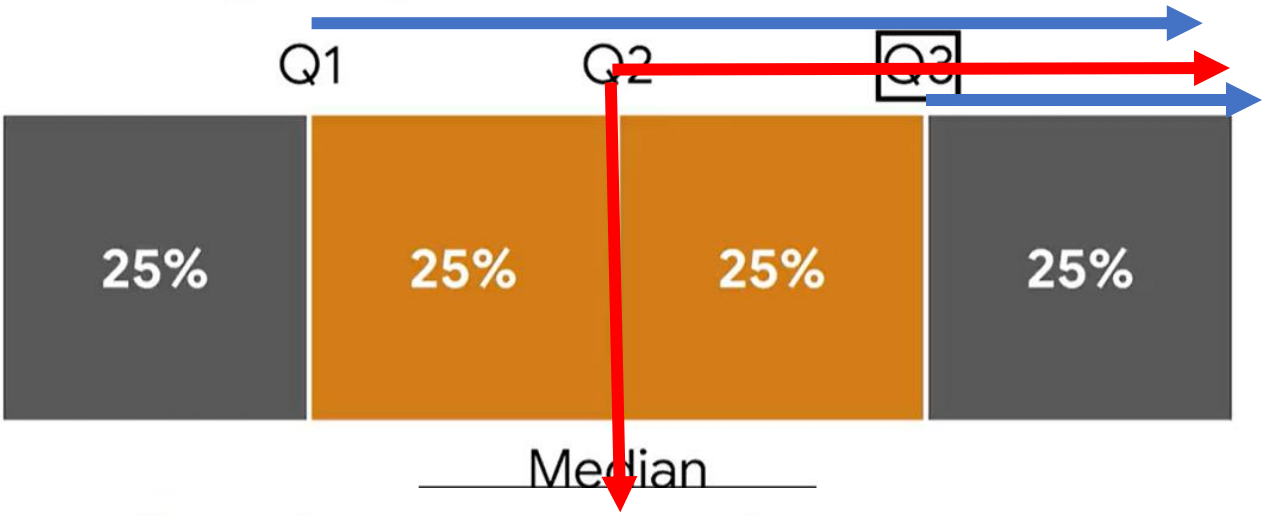
- Percentiles
- Quartiles
- Interquartile range
- Five number summary

Percentile

The value below which a percentage of data falls

Quartile

Divides the values in a dataset into four equal parts



- Q1 = 25th percentile
- Q2 = 50th percentile
- Q3 = 75th percentile

Interquartile range (IQR)

The distance between the first quartile (Q1) and the third quartile (Q3)

Player	#7	#3	#8	#1	#2	#6	#4	#5
Goals scored	11	12	14	18	22	23	27	33

18 and 22, Q2 equals 20.

- Q1 = 25th percentile = 13 goals
- Q3 = 75th percentile = 25 goals

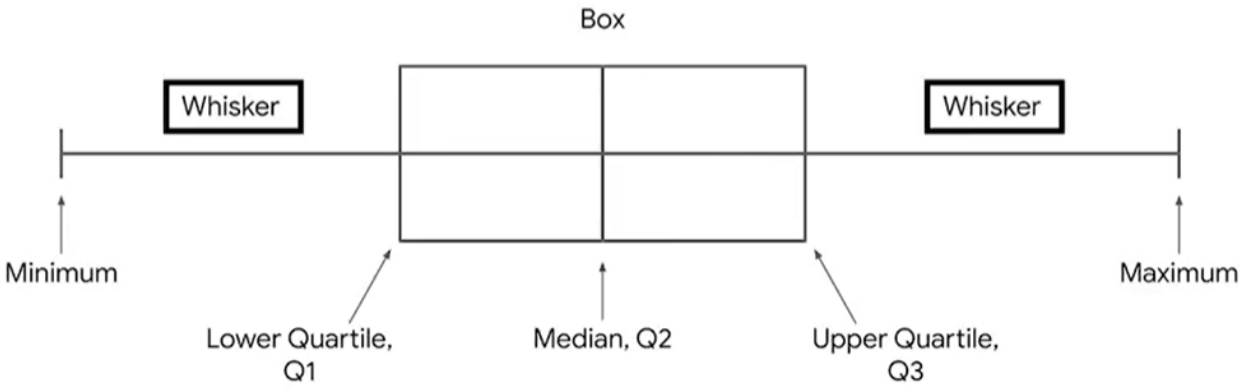
$$IQR = Q3 - Q1$$

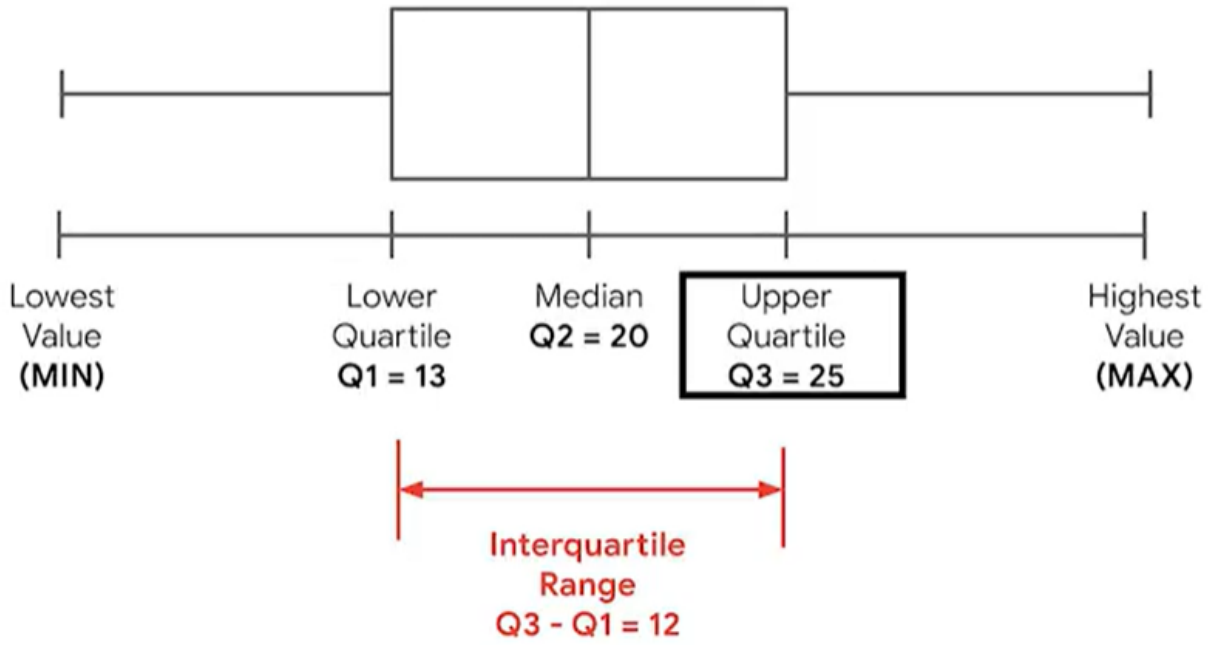
$$IQR = 25 - 13 = 12$$

Five number summary

- The minimum = 11
- The first quartile (Q1) = 13
- The median, or second quartile (Q2) = 20
- The third quartile (Q3) = 25
- The maximum = 38

Box Plot





Literacy rate

The percentage of the population of a given age group that can read and write

`describe()` for a categorical column

- Number of unique values
- Most common value, or mode
- The frequency of the most common value

Functions for stats

- `mean()`
- `median()`
- `std()`
- `min()`
- `max()`

Objective probability

Based on statistics, experiments, and mathematical measurements

Subjective probability

Based on personal feelings, experience, or judgment

Objective probability

- Classical
- Empirical

Classical probability

Based on formal reasoning about events with equally likely outcomes

$$\text{Classical probability} = \frac{\text{Number of desired outcomes}}{\text{Total number of possible outcomes}}$$

Empirical probability

Based on experimental or historical data

$$\text{Empirical probability} = \frac{\text{number of times a specific event occurs}}{\text{total number of events}}$$

Random experiment

A process whose outcome cannot be predicted with certainty

Random experiments

- The experiment can have **more than one possible outcome**
- You can represent each possible outcome **in advance**
- The outcome of the experiment **depends on chance**

$$\text{Classical probability} = \frac{\text{Number of desired outcomes}}{\text{total number of possible outcomes}}$$

Basic rules of probability

- Complement rule
- Addition rule
- Multiplication rule

Types of events

- Mutually exclusive events
- Independent events

Probability notation

- P(A) = probability of event A
- P(B) = probability of event B
- P(A') = probability of **not** event A

Complement of an event

The event not occuring

Complement rule

$$P(A') = 1 - P(A)$$

P(no rain) = 1 - P(rain) = 1 - 0.3 = 0.7

Addition rule
(for mutually exclusive events)

P(A or B) = P(A) + P(B)

$$P(A \text{ or } B) = P(A) + P(B)$$

P(rolling a 2 or rolling a 4) = P(rolling a 2) + P(rolling a 4)

P(rolling a 2 or rolling a 4) = 1/6 + 1/6 = 2/6 = **1/3**

=33%

Independent events

Two events are independent if the occurrence of one event does not change the probability of the other event

Multiplication rule
(for independent events)

P(A and B) = P(A) * P(B)

$$P(A \text{ and } B) = P(A) * P(B)$$

P(1st toss tails and 2nd toss heads) = P(1st toss tails) * P(2nd toss heads)

P (tails and heads) = (0.5) * (0.5) = 0.25

Addition rule applies to mutually exclusive events

Multiplication rule applies to independent events

Conditional probability

The probability of an event occurring given that another event has already occurred

Conditional probability is used in

- Finance
- Insurance
- Science
- Machine learning

Dependent events

Two events are dependent if the occurrence of one event changes the probability of the other event

Conditional probability P(A and B) = P(A) * P(B|A)

- P(A and B) = “probability of event A and event B”
- P(A) = “probability of event A”
- P(B|A) = “probability of event B given event A”

Bayes’ theorem

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

Prior probability

The probability of an event before new data is collected

Posterior probability

The updated probability of an event based on new data

Bayesian statistics (Bayesian inference)

A powerful method for analyzing and interpreting data in modern data analytics

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

- Event A = Rain
- Event B = Cloudy

$$P(\text{Rain}|\text{Cloudy}) = \frac{P(\text{Cloudy}|\text{Rain}) * P(\text{Rain})}{P(\text{Cloudy})}$$

- P(Rain) = 10%
- P(Cloudy) = 40%
- P(Cloudy|Rain) = 50%

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

$$P(\text{Rain}|\text{Cloudy}) = \frac{P(\text{Rain}) * P(\text{Cloudy}|\text{Rain})}{P(\text{Cloudy})}$$

$$P(\text{Rain}|\text{Cloudy}) = \frac{(0.1 * 0.5)}{0.4} = 0.125 = 12.5\%$$

Bayes' theorem (expanded version)

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|\text{not } A) * P(\text{not } A)}$$

False positive

Test result that indicates something is present when it really is not

False negative

Test result that indicates something is not present when it really is

- **Prior probability** = the probability that a person has the medical condition
- **Posterior probability** = the probability that the condition is present GIVEN that the test is positive
- Event A = actually having the medical condition
- Event B = testing positive
- $P(A) = 1\%$
- $P(B|A) = 95\%$
- $P(B|\text{not } A) = 2\%$

$$P(A') = 1 - P(A)$$

$$P(A') = 1 - 0.01 = 0.99 = 99\%$$

- P(not A) = 99%

so the probability of
not A equals 99 percent.

Bayes' theorem (expanded version)

$$P(A|B) = \frac{0.95 * 0.01}{0.95 * 0.01 + 0.02 * 0.99)}$$

$$P(A|B) = 0.324 = 32.4\%$$

Probability distribution

Describes the likelihood of the possible outcomes of a random event

Random variable

Represents the values for the possible outcomes of a random event

Random variables

- Discrete
- Continuous

Discrete random variable

Has a countable number of possible values

Continuous random variable

Takes all the possible values in some range of numbers

Discrete or continuous variables

- COUNT the number of outcomes = discrete
- MEASURE the outcome = continuous

Discrete or continuous distributions

- Discrete distributions represent discrete random variables
- Continuous distributions represent continuous random variables

Sample space

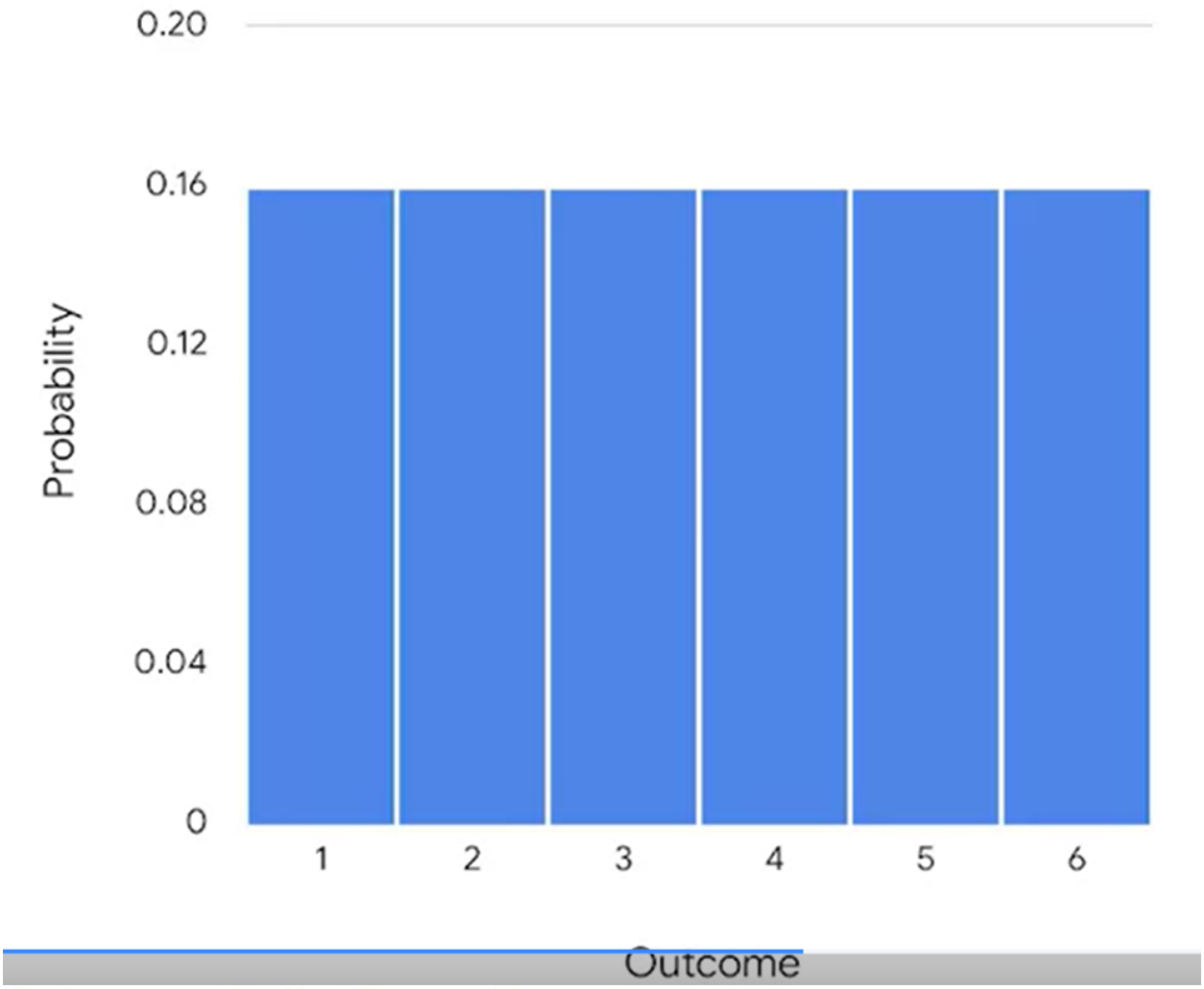
The set of all possible values for a random variable

- Sample space for single coin toss = {Heads, Tails}
- Sample space for single die roll = {1, 2, 3, 4, 5, 6}

Single die roll

- Sample space = {1, 2, 3, 4, 5, 6}
- Probability of each outcome = 16.7%

Outcome of die roll	1	2	3	4	5	6
Probability	1/6	1/6	1/6	1/6	1/6	1/6



Binomial distribution

A discrete distribution that models the probability of events with only two possible outcomes, success or failure

This definition assumes

- Each event is independent
- The probability of success is the same for each event

Mutually exclusive

Two outcomes are mutually exclusive if they cannot occur at the same time

Correct
The binomial distribution models the probability of events with two possible outcomes.

Random experiment

A process whose outcome cannot be predicted with certainty

Random experiments

- The experiment can have more than one possible outcome
- You can represent each possible outcome in advance
- The outcome of the experiment depends on chance

Binomial experiment

- The experiment consists of a number of repeated trials
- Each trial has only two possible outcomes
- The probability of success is the same for each trial
- Each trial is independent

Binomial experiment

- 10 repeated coin tosses
- Two possible outcomes: heads or tails
- The probability of success for each toss is the same: 50%
- The outcome of one coin toss does not affect the outcome of any other coin toss

Binomial experiment

- 100 customer visits
- Two possible outcomes: return or not return
- The probability of success for each customer visit is the same: 10%
- The outcome of one customer visit does not affect the outcome of any other customer visit

Binomial distribution formula

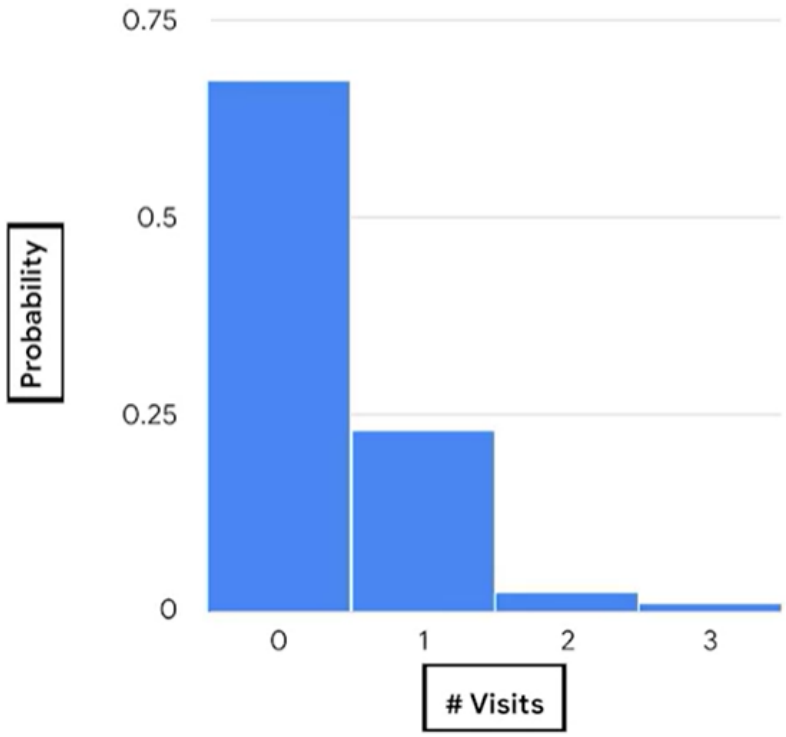
$$P(X=k)= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

- k refers to the number of successes
- n refers to the number of trials
- p refers to the probability of success on a given trial

$$P(X=k)= \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

- P(X=0) = 0.729
- P(X=1) = 0.243
- P(X=2) = 0.027
- P(X=3) = 0.001

Binomial distribution



Discrete probability distributions

- Binomial
- Poisson

Poisson distribution

Models the probability that a certain number of events will occur during a specific time period

Data professionals use the Poisson distribution to model data such as

- Calls per hour for a customer service call center
- Visitors per hour for a website
- Customers per day at a restaurant
- Severe storms per month in a city

Poisson experiment

- The number of events in the experiment can be counted
- The mean number of events that occur during a specific time period is known
- Each event is independent

Poisson distribution formula

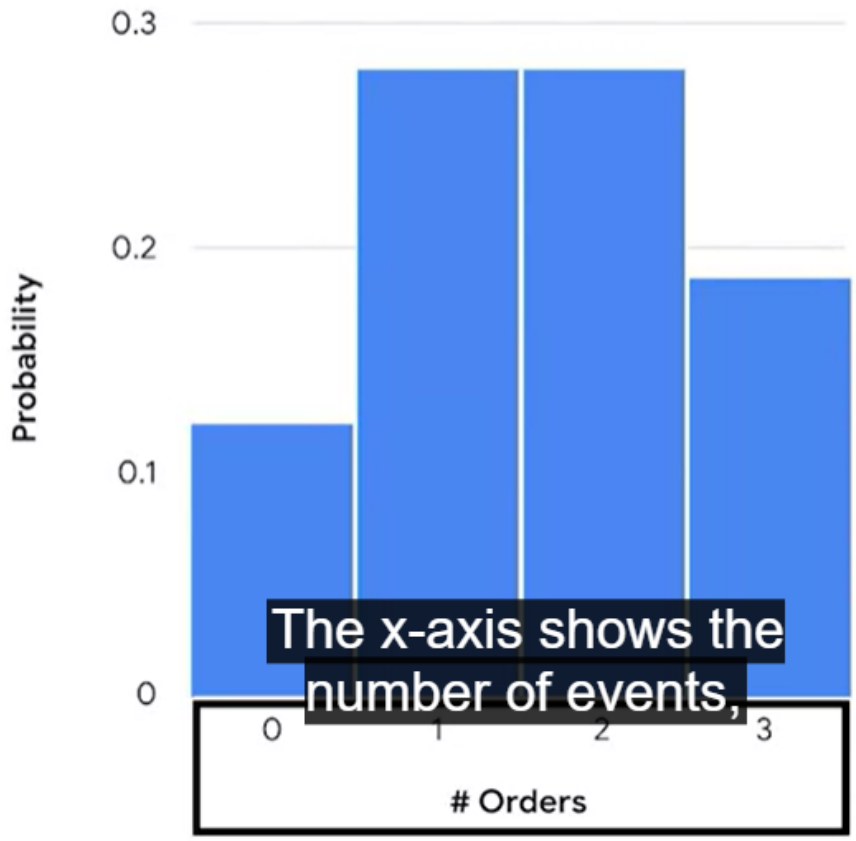
$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- The Greek letter lambda(λ)refers to the mean number of events that occur during a specific time period
- k refers to the number of events
- e is a constant equal to approximately 2.71828
- The exclamation point stands for factorial

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- P(X=0) = 0.1353
- P(X=1) = 0.2707
- P(X=2) = 0.2707
- P(X=3) = 0.1805

Poisson distribution



	Given	Want to Find	Example
Poisson	The average probability of an event happening for a specific time period	The probability of a certain number of events happening in that time period	The probability of getting 12 calls between 2 p.m. and 3 p.m.
Binomial	An exact probability of an event happening	The probability of the event happening a certain number of times in a repeated trial	The probability of getting 8 heads in 10 coin tosses

Normal distribution

A continuous probability distribution that is symmetrical on both sides of the mean and bell-shaped

Normal distribution

- Height
- Weight
- Blood pressure
- IQ scores
- Salaries

blood pressure, IQ scores, salaries, and more.

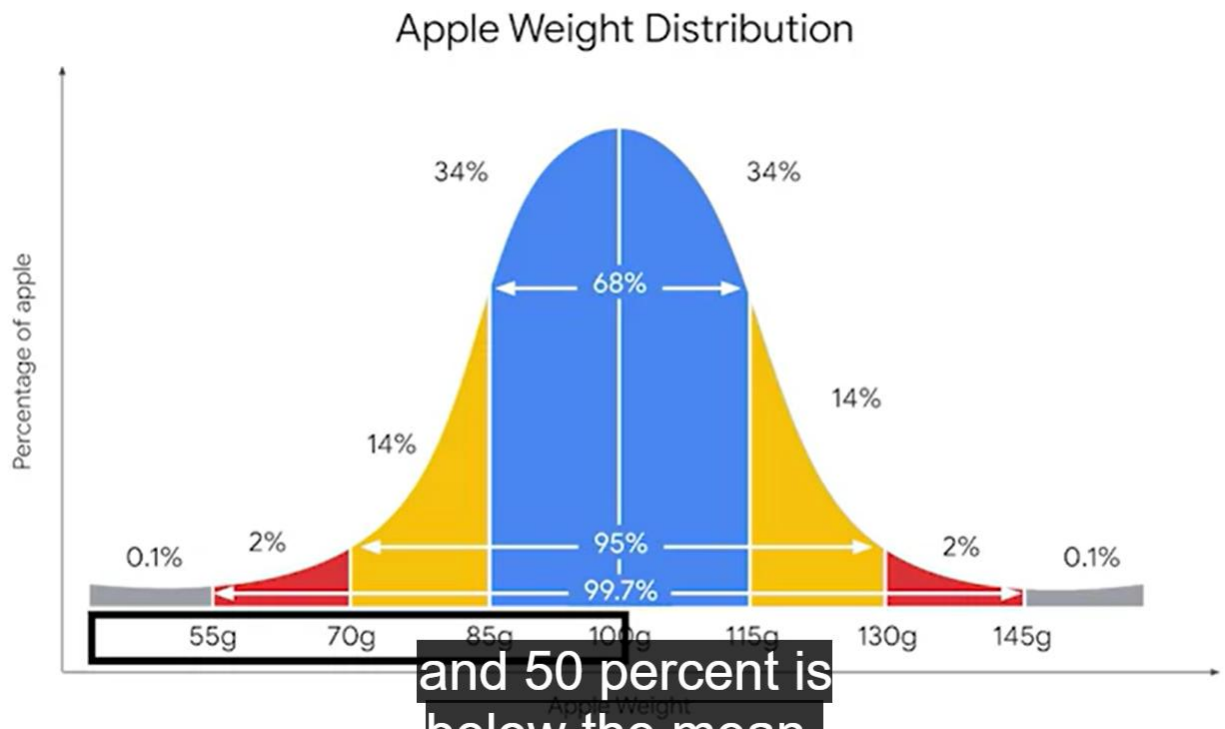
Data professionals use the normal distribution to model datasets in

- Business
- Science
- Government
- Machine learning

science, government, machine learning, and others.

Normal distributions have the following features

- The shape is a bell curve
- The mean is located at the center of the curve
- The curve is symmetrical on both sides of the center
- The total area under the curve equals 1



and 50 percent is below the mean.

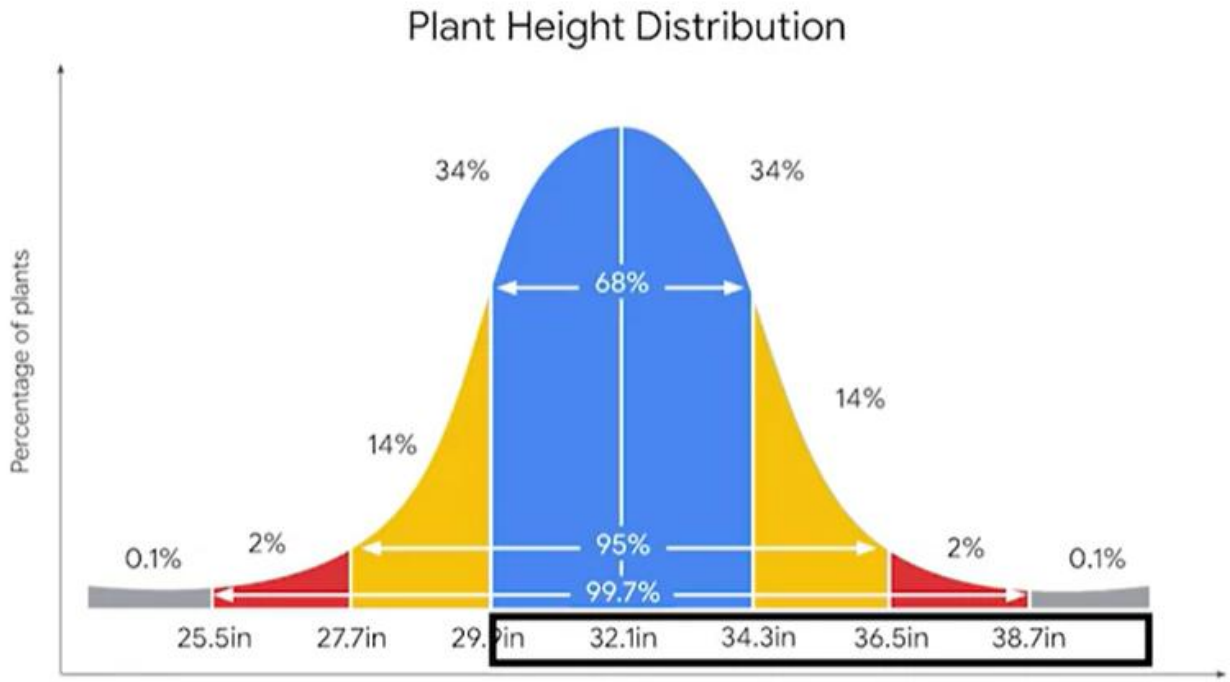
Standard deviation

Calculates the typical distance of a data point from the mean of your dataset

The empirical rule

- 68% of values fall within 1 standard deviation of the mean
- 95% of values fall within 2 standard deviations of the mean
- 99.7% of values fall within 3 standard deviations of the mean

34% + 50% = 84%



greater than 29.9 inches tall.

Used in machine learning for the binary classification of digital images

6 of 12

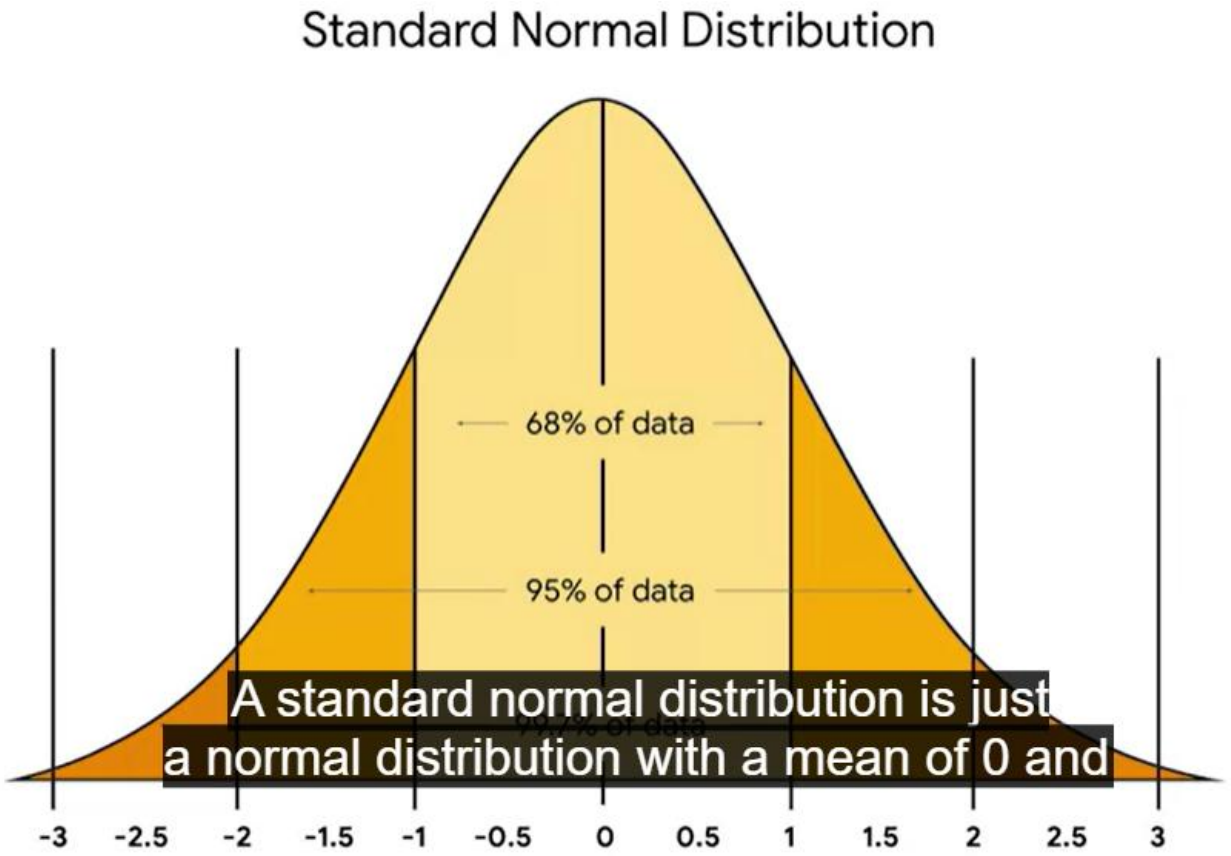
Binomial 2 ›

Z-score

A measure of how many standard deviations below or above the population mean a data point is

Standardization

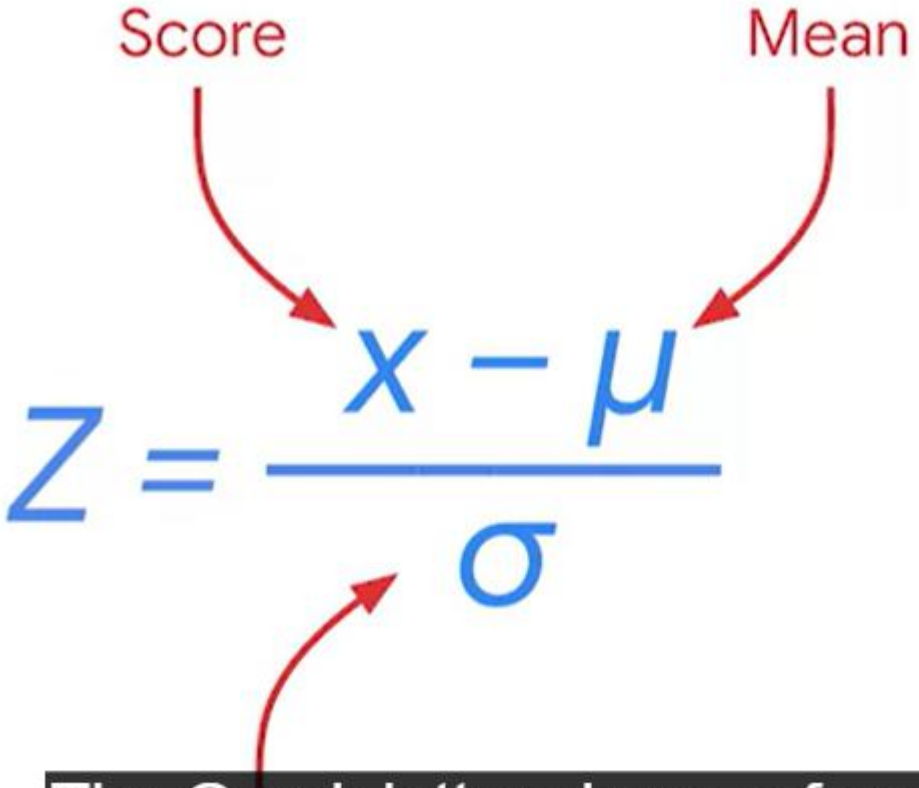
The process of putting different variables on the same scale



A standard normal distribution is just a normal distribution with a mean of 0 and

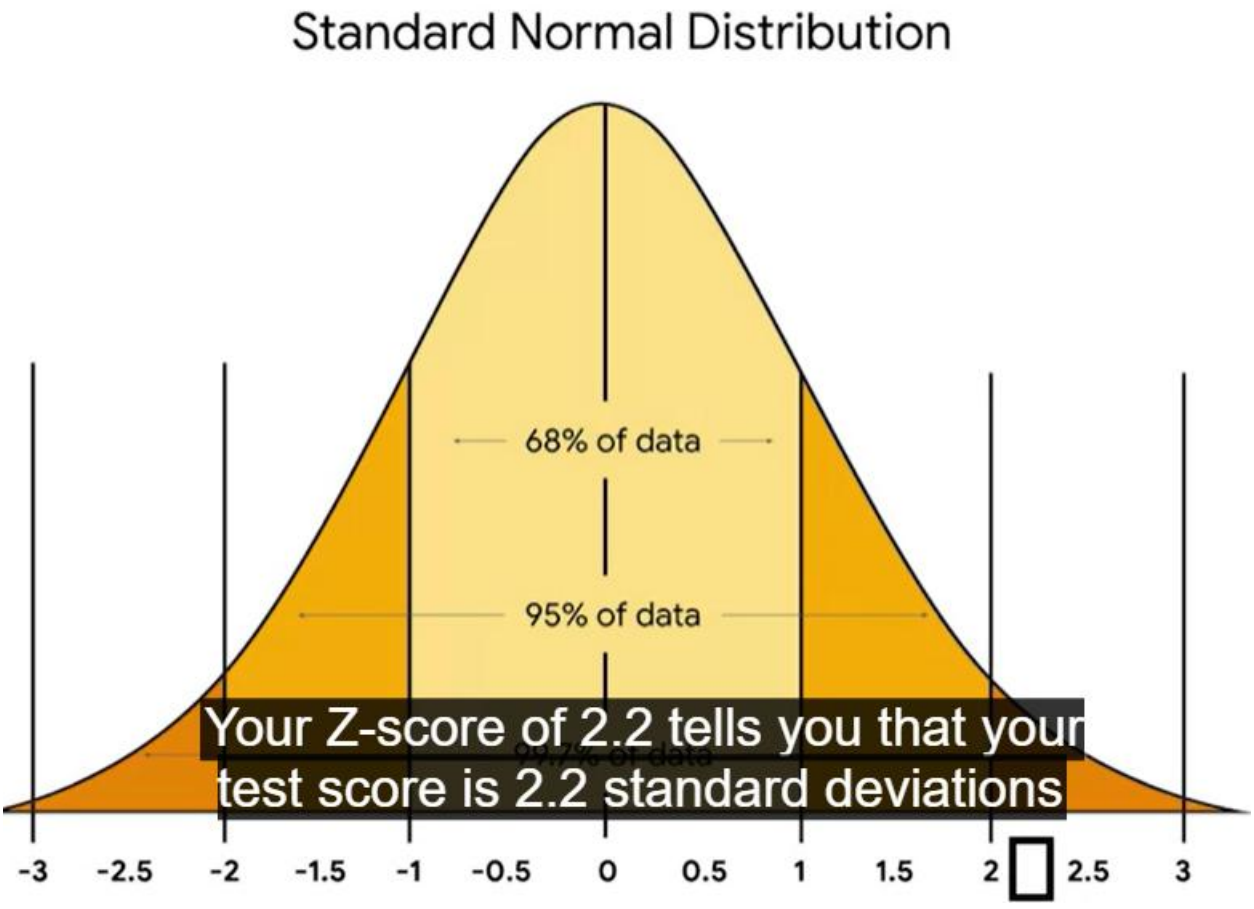
Anomaly detection application

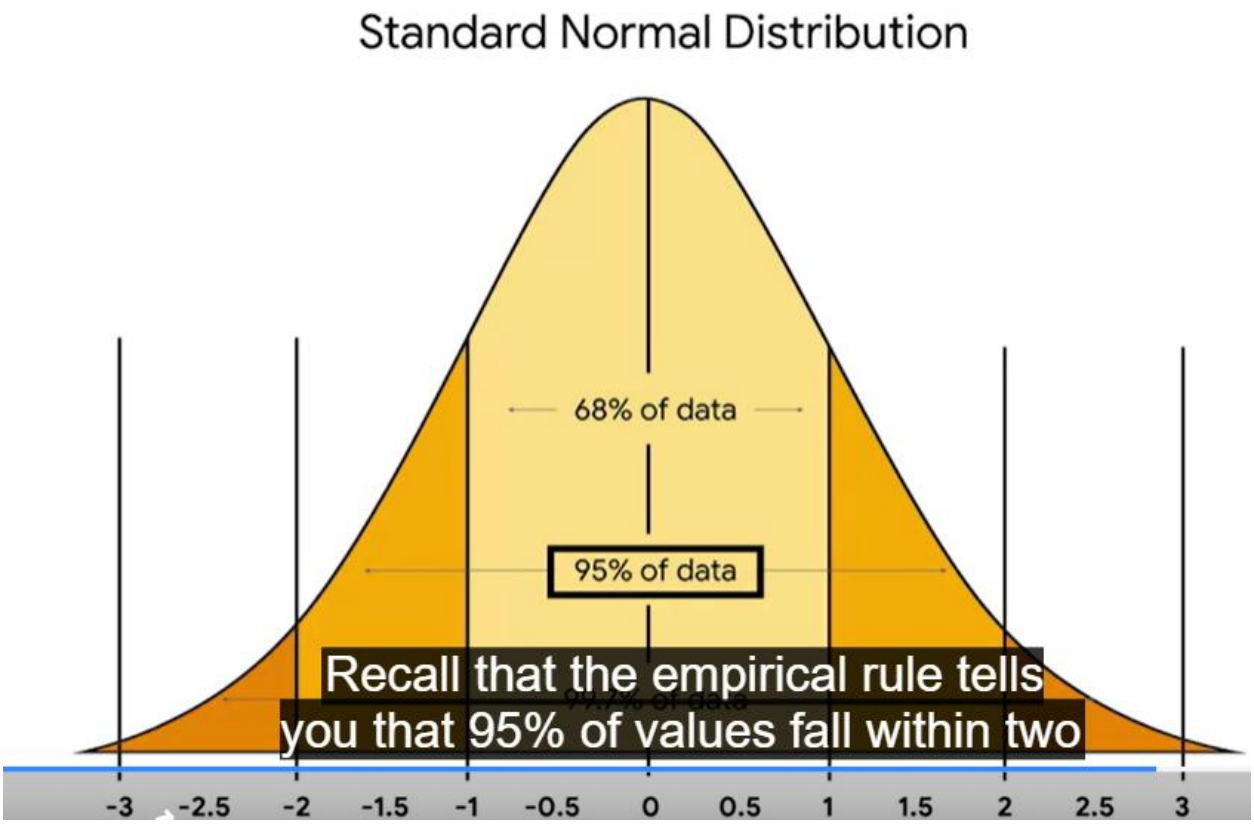
- Fraud in financial transactions
- Flaws in manufacturing products
- Intrusions in computer networks



$$Z = \frac{(133 - 100)}{15}$$

$$Z = \frac{33}{15} \quad Z = 2.2$$





What is the z-score of a data value equal to the mean?

- 1
- 3
- 2
- 0

Correct

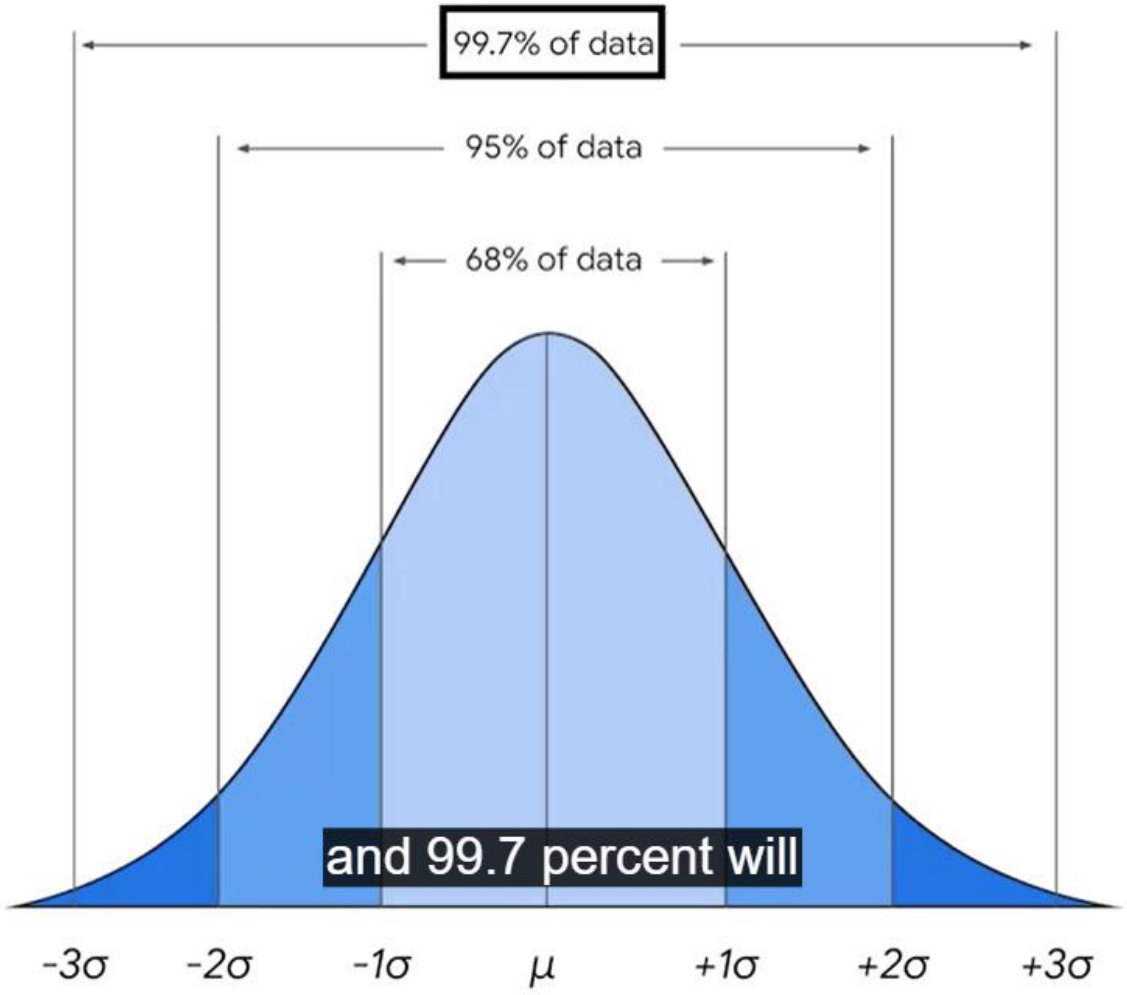
The z-score is 0 if the data value is equal to the mean. A z-score is a measure of how many standard deviations below or above the population mean a data point is.

Python for stats

- SciPy stats
- Statsmodels

The Empirical Rule

- 68% of values fall within 1 standard deviation of the mean
- 95% of values fall within 2 standard deviations of the mean
- 99.7% of values fall within 3 standard deviations of the mean



1 standard deviation below the mean

- Mean - (1 * standard deviation)
- 73 - (1 * 10) = 63

1 standard deviation above the mean

- Mean + 1 * standard deviation
- 73 + (1 * 10) = 83

Z-score

A measure of how many standard deviations below or above the population mean a data point is

Types of statistics

- **Descriptive statistics** summarize the main features of a dataset
- **Inferential statistics** use sample data to draw conclusions about a larger population

The sampling process

First, let's review the main steps of the sampling process:

1. Identify the target population
2. Select the sampling frame
3. Choose the sampling method
4. Determine the sample size
5. Collect the sample data

Simple random sample

Every member of a population is selected randomly and has an equal chance of being chosen

Cluster random sample

Divide a population into clusters, randomly select certain clusters, and include all members from the chosen clusters in the sample

Systematic random sample

Put every member of a population into an ordered sequence. Then, you choose a random starting point in the sequence and select members for your sample at regular intervals.

randomly choose a starting point,

Systematic sample

and pick every fifth name
to be in the sample.

Probability sampling methods

- Simple
- Stratified
- Cluster
- Systematic

Non-probability sampling methods

- Convenience sampling
- Voluntary response sampling
- Snowball sampling
- Purposive sampling

Undercoverage bias

When some members of a population are inadequately represented in the sample

Voluntary response sample

Consists of members of a population who volunteer to participate in a study

Snowball sample

Researchers recruit initial participants to be in a study and then ask them to recruit other people to participate in the study

Purposive sample

Researchers select participants based on the purpose of their study

Statistic vs. parameter

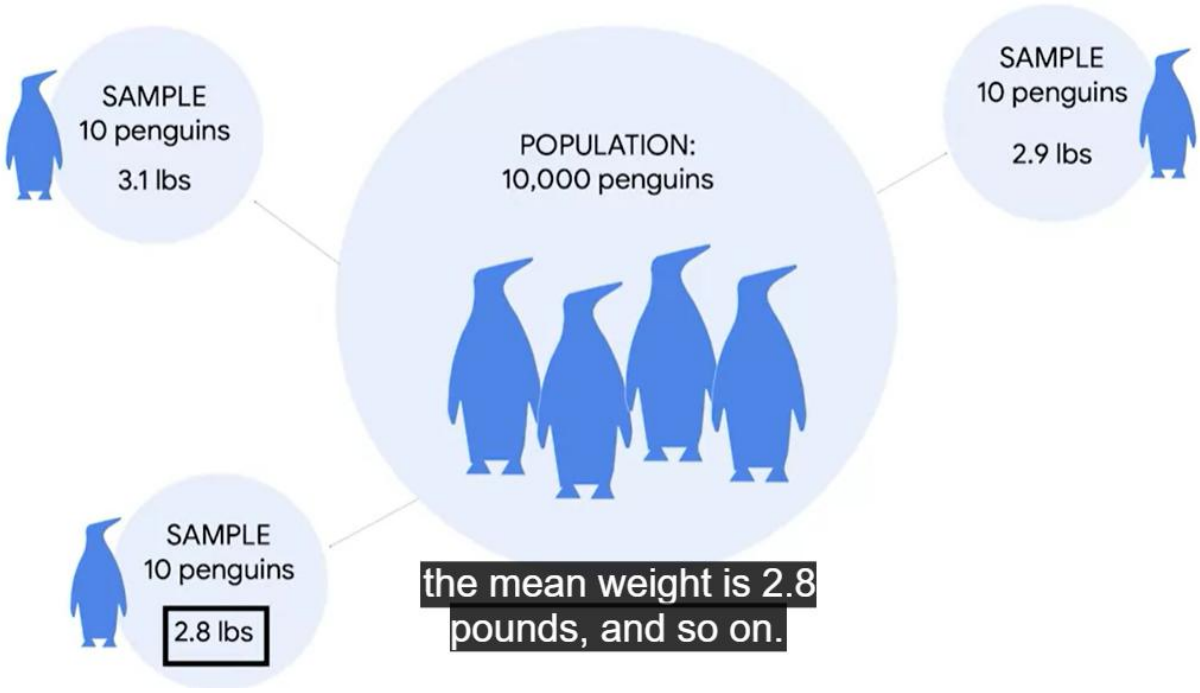
- The mean weight of a random sample of 100 penguins is a **statistic**
- The mean weight of the total population of 10,000 penguins is a **parameter**

Sampling distribution

A probability distribution of a sample statistic

Point estimate

Uses a single value to estimate a population parameter



Sampling variability

How much an estimate varies between samples

Sampling variability

- **Population mean** = 3 lbs
  - **Sample mean** = 3.3 lbs
  - **Sample mean** = 2.8 lbs
  - **Sample mean** = 2.4 lbs

Standard error

- Larger standard error = Sample means are more spread out
- Smaller standard error = Sample means are closer together

Standard error of the mean = $\frac{s}{\sqrt{n}}$

- s: The sample standard deviation
- n: The sample size

Standard error of the mean

$$= \frac{S}{\sqrt{n}}$$

$$= \frac{1}{\sqrt{100}}$$

$$= 0.1 \text{ lbs}$$

As sample size gets larger, standard error gets smaller

Sample size = 10,000 penguins

- **SE** = 0.01

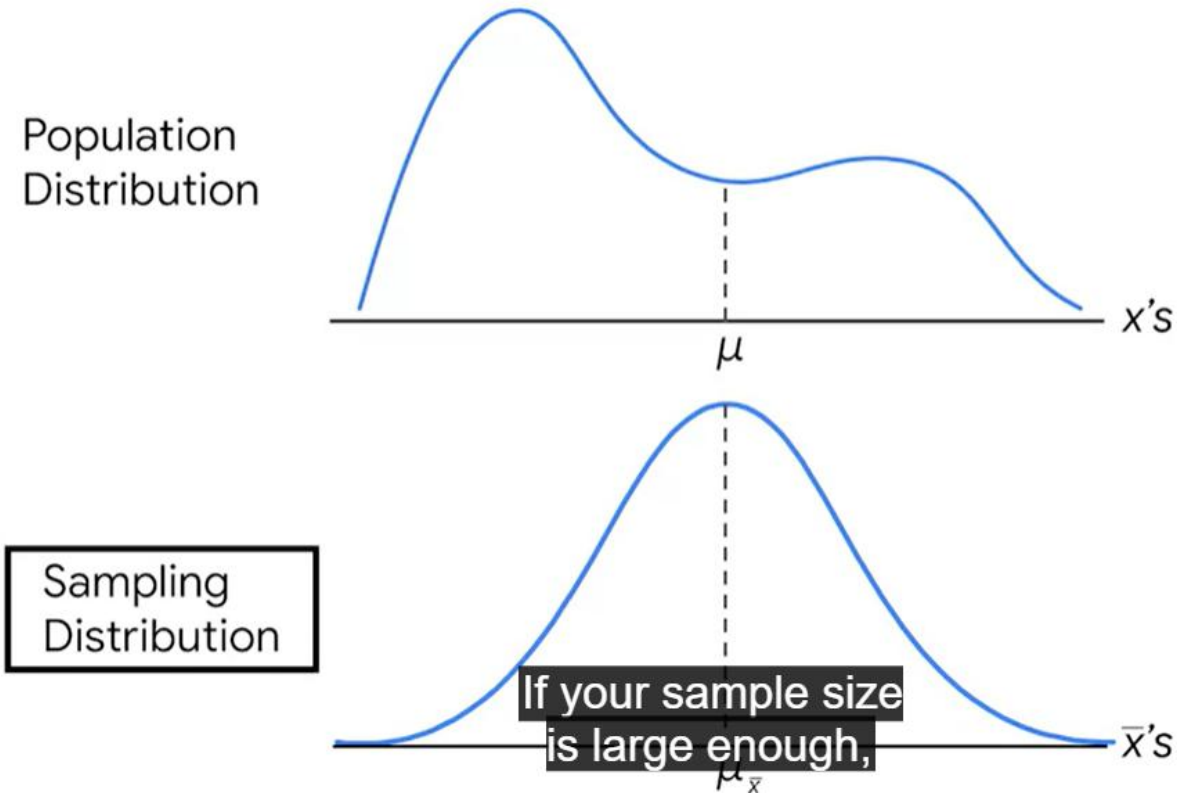
Sample size = 100 penguins

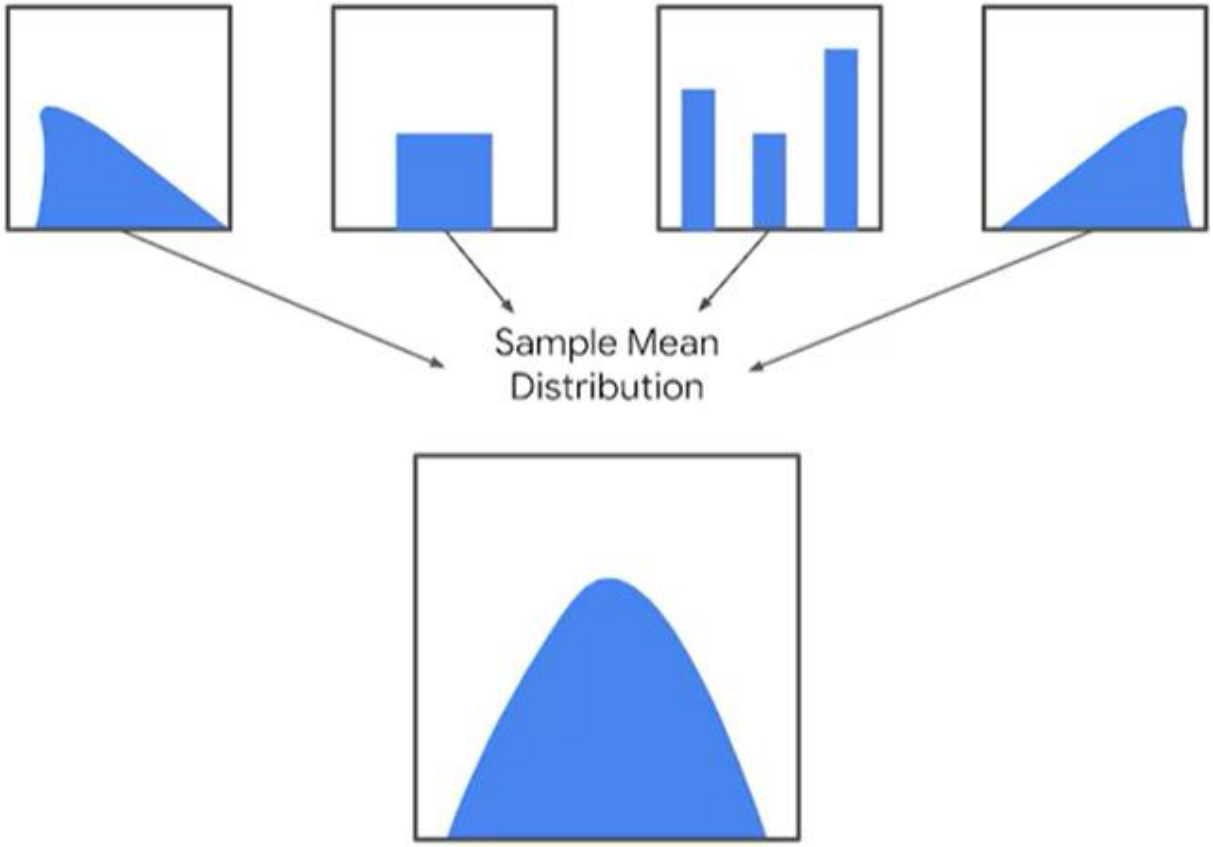
- **SE** = 0.1

will vary with a standard deviation of just 0.01 pounds.

Central Limit Theorem

The sampling distribution of the mean approaches a normal distribution as the sample size increases


Population Distribution
x's
μ
Sampling Distribution
If your sample size is large enough,
x̄'s
μx̄



Normal distribution

Population proportion

The percentage of individuals or elements
in a population that share a certain characteristic

Standard error of the proportion

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

P hat refers to the population proportion, and n refers to the sample size.

For example, suppose you survey 100 teenagers about their sneaker preferences

the population proportion of teens who prefer slip-on sneakers is 10% or 0.1.

In this case, p hat is 0.1 and n is 100.

Standard error of the proportion

$$\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

$$\sqrt{\frac{0.1(1-0.1)}{100}} = 0.03$$

.sample (n, replace, random_state)

- n refers to the desired sample size
- replace indicates whether you are sampling with or without replacement
- random_state refers to the seed of the random number

Sampling with replacement

When a population element can be selected more than one time

Sampling without replacement

When a population element can be selected only one time

Random seed

A starting point for generating random numbers

Interval

sample statistic +/- margin of error

Margin of error

The maximum expected difference between a population parameter and a sample estimate

Penguins interval

30+2=32

30-2=28

Interval: [28 , 32]

Confidence level

Describes the likelihood that a particular sampling method will produce a confidence interval that includes the population parameter

Sales revenue

- "I think we'll do $1,000,000 in sales."
- "Based on a 95 percent confidence level, I estimate that our sales revenue will be between $950,000 and $1,050,000."

Confidence level

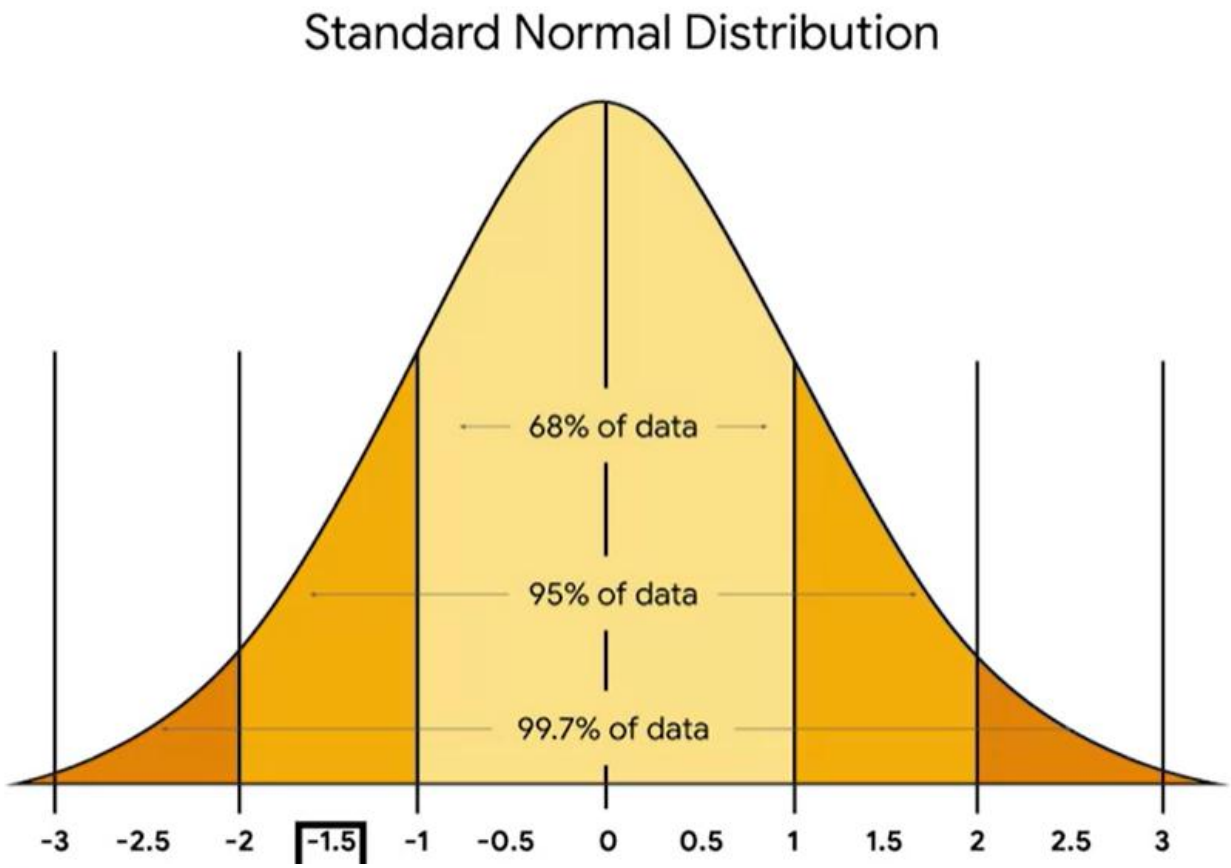
Expresses the uncertainty of the estimation process

Steps for constructing a confidence interval

1. Identify a sample statistic
2. Choose a confidence level
3. Find the margin of error
4. Calculate the interval

Margin of error = z-score * SE

Confidence level	Z-score
90%	1.645
95%	1.96
99%	2.58



Steps for constructing a confidence interval

1. Identify a sample statistic: proportion
2. Choose a confidence level: 95%
3. Find the margin of error: z-score * SE = 1.96 * 0.05 = 0.098
4. Calculate the interval: [45.2% , 64.8%]

Upper limit = Sample proportion + margin of error

0.55 + 0.098 = 0.648

Lower limit = Sample proportion - margin of error

0.55 - 0.098 = 0.452

Confidence interval

95% CI [45.2 , 64.8]

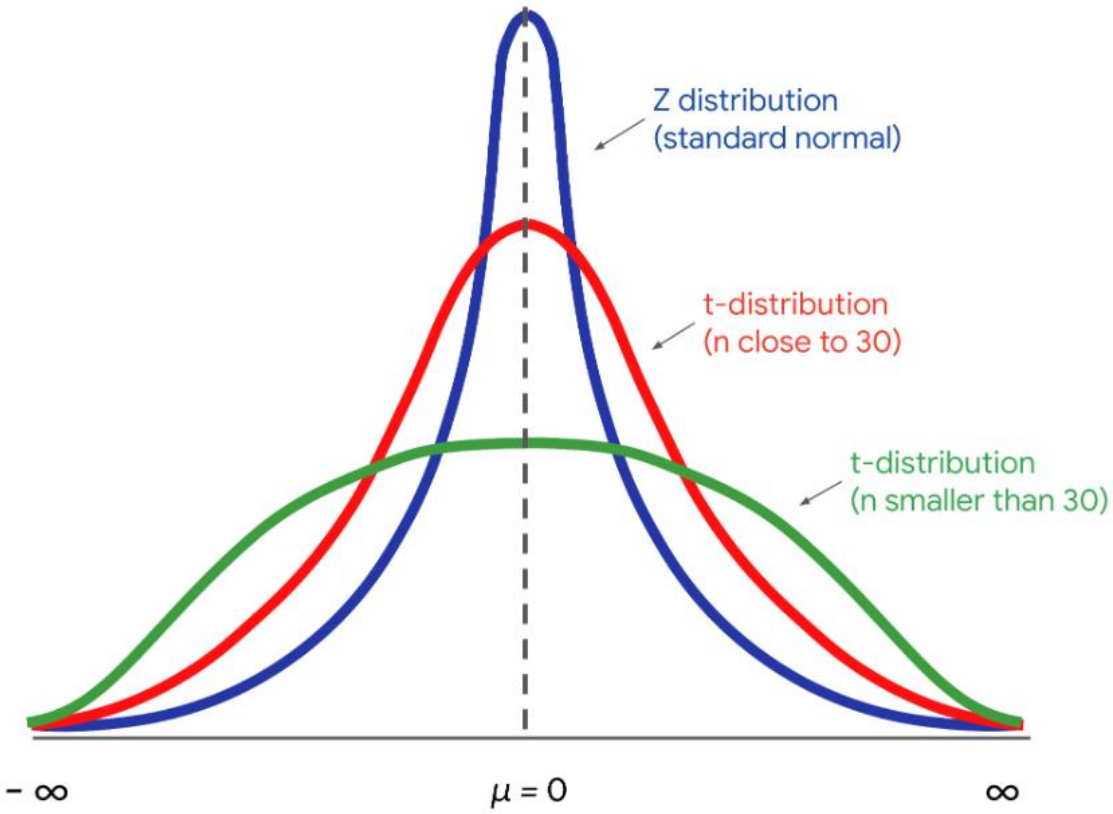
Standard error of the mean:

$$SE = \frac{\sigma^*}{\sqrt{n}}$$

* "σ" when std of population is known, otherwise "s"

Confidence intervals

- Confidence level = 95%
  [20:12 , 20:48] = 36 min
- Confidence level = 99%
  [20:07 , 20:53] = 46 min



Hypothesis testing

A statistical procedure that uses sample data to evaluate an assumption about a population parameter

Statistical significance

The claim that the results of a test or experiment are not explainable by chance alone

Steps for performing a hypothesis test

1. State the null hypothesis and the alternative hypothesis
2. Choose a significance level
3. Find the p-value
4. Reject or fail to reject the null hypothesis

Null hypothesis

A statement that is assumed to be true unless there is convincing evidence to the contrary

Alternative hypothesis

A statement that contradicts the null hypothesis, and is accepted as true only if there is convincing evidence for it

Significance level

The probability of rejecting the null hypothesis when it is true

P-value

The probability of observing results as or more extreme than those observed when the null hypothesis is true

P-value

The probability of observing a difference in your sample means as or more extreme than the difference observed when the null hypothesis is true

A lower p-value means there is stronger evidence for the alternative hypothesis

Decide

Reject or **fail to reject** the null hypothesis

Drawing a conclusion

- If p-value < significance level: **reject** the null hypothesis
- If p-value > significance level: **fail to reject** the null hypothesis

P-value: 1.56%
Significance level: 5%

1.56% < 5%

Conclusion: Reject the null hypothesis

Types of errors in hypothesis testing

- Type I error
- Type II error

Type I error (false positive)

The rejection of a null hypothesis that is actually true

Type II error (false negative)

The failure to reject a null hypothesis which is actually false

To minimize the risk of a type I error, choose a significance level of 1%

	Null hypothesis (H_0)	**Alternative hypothesis (H_a)**
Claims	There is no effect in the population.	There is an effect in the population.
Language	• No effect • No difference • No relationship • No change	• An effect • A difference • A relationship • A change
Symbols	Equality (=, ≤, ≥)	Inequality (≠, <, >)

- Reject the null hypothesis when it's actually true (**Type I error)**
- Reject the null hypothesis when it's actually false (Correct)
- Fail to reject the null hypothesis when it's actually true (Correct)
- Fail to reject the null hypothesis when it's actually false (**Type II error**)

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct Outcome! (True positive)
Fail to reject null hypothesis	Correct Outcome! (True negative)	Type II Error (False negative)

One-sample test

Determines whether or not a population parameter like a mean or proportion is equal to a specific value

Two-sample test

Determines whether or not two population parameters such as two means or two proportions are equal to each other

Test statistic

A value that shows how closely your observed data matches the distribution expected under the null hypothesis

Note: 5% is a conventional choice, and not a magical number. It's based on tradition in statistical research and education. Other common choices are 1% and 10%. You can adjust the significance level to meet the specific requirements of your analysis. A lower significance level means an effect has to be larger to be considered statistically significant.

Pro tip: As a best practice, you should set a significance level before you begin your test. Otherwise, you might end up in a situation where you are manipulating the results to suit your convenience.

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

$$t = \frac{(300 - 305)}{\sqrt{\left(\frac{18.5^2}{40} + \frac{16.7^2}{38}\right)}}$$

$$t = -1.2508$$

Draw a conclusion

- If p-value < significance level: **reject the null hypothesis**
- If p-value > significance level: **fail to reject the null hypothesis**

P-value: 21.48%

Significance level: 5%

21.48%> 5%

Conclusion: Fail to reject the null hypothesis

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Hypotheses

- **Null**: There is no difference in the proportion of satisfied employees in London and Beijing
- **Alternative**: There is a difference in the proportion of satisfied employees in London and Beijing

	London office	Beijing office
Sample size	50	50
Sample proportion	67%	57%

$$Z = \frac{0.67 - 0.57}{\sqrt{0.62\,(1-0.62)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$

$Z = 1.03$ you get a Z-score of 1.03.

Draw a conclusion

- If p-value < significance level: **reject** the null hypothesis
- If p-value > significance level: **fail to reject** the null hypothesis

```
stats.ttest_ind(a,b,equal_var)
```

- **a:** observations from the first sample
- **b:** observations from the second sample
- **equal_var:** indicates whether the population variance of the two samples is assumed to be equal

Questions you might get in interviews

- How would you use statistics to measure the performance of our company?
- How have you used statistical models to solve business problems?
- What are the various factors that go into an experimental design for an A/B test?

"Regression models are super important, powerful tools. You'll be able to answer a wide variety of questions using different types of regression models."

Model assumptions

Statements about the data that must be true to justify the use of particular data science techniques

Continuous Variables	Categorical Variables
Takes on any real value between minimum and maximum value	Have a finite number of possible values
Examples: Product sales Vehicle speed Time spent on webpage	*Examples:* Types of products Educational level

Dependent variable (Y)

The variable a given model estimates, also referred to as a response or outcome variable

$$Y = \text{slope} * X + \text{intercept}$$

Slope

The amount that y increases or decreases per one-unit increase of x

Intercept

The value of y, the dependent variable, when x, the independent variable, equals 0

Positive correlation

A relationship between two variables that tend to increase or decrease together

Negative correlation

An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

Correlation is not causation

Causation

A cause-and-effect relationship where one variable directly causes the other to change in a particular way

Observed values (actual values)

The existing sample of data

Each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable.

Linear regression equation

$$\mu\{Y|X\} = \beta_0 + \beta_1 X$$

Intercept Slope

Betas (β_i) are parameters.

Linear regression estimation

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$= -1 + 5X$$

X	$\hat{y}$
0	-1
1	4
2	9
3	14

For every one-unit increase in X, we get a 5-unit increase in Y

Regression coefficients

The estimated betas in a regression model.
Represented as $\hat{\beta}_i$

Loss function

A function that measures the distance between the observed values and the model's estimated values

Logistic regression

A technique that models a categorical dependent variable based on one or more independent variables

Y	X
Users don't subscribe (Y = 0) **Users subscribe (Y = 1)**	Continuous Minutes the user spends on a webpage

Logistic regression model

$$\mu\{Y|X\} = Prob(Y = 1|X)$$

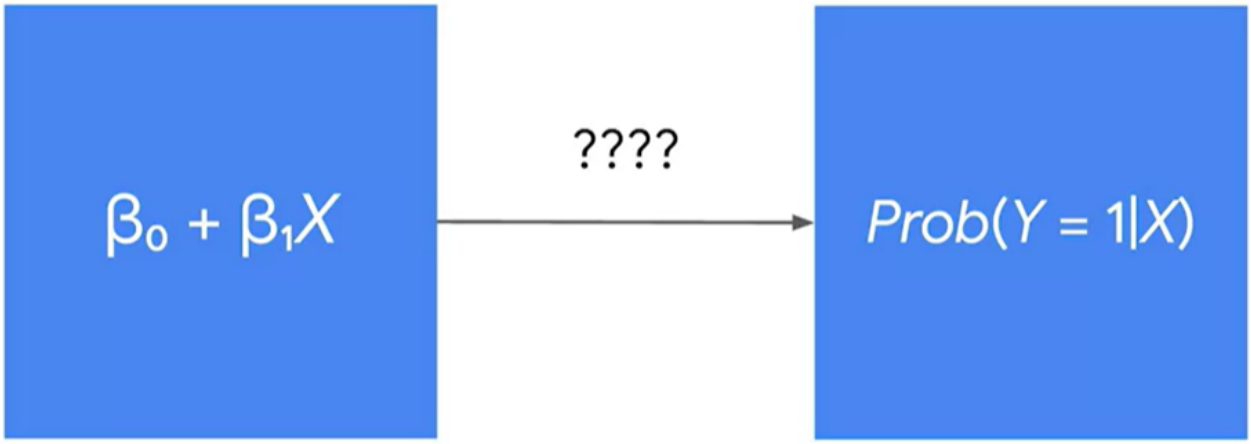
Observations

Number of 0's = 8 Number of 1's = 10

Sum of observations = 10 = Number of 1's

Logistic Regression Model

$$\mu\{Y|X\} = Prob(Y = 1|X) = p$$



$$g(p) = \beta_0 + \beta_1 X$$

Link function

Link function

A nonlinear function that connects or links the dependent variable to the independent variables mathematically

Linear Regression	Logistic Regression
Continuous data (i.e. book sales - 100 books, 200 books, 437 books, etc.)	Categorical data (i.e. newsletter subscription - yes/no)
Estimating the MEAN of y	Estimating the PROBABILITY of an outcome
$\mu(Y\|X) = \beta_0 + \beta_1 X$	$\mu(Y\|X) = Prob(Y = 1\|Y) = p$ $g(p) = \beta_0 + \beta_1 X$

Simple linear regression

A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

Best fit line

The line that fits the data best by minimizing some loss function or error

Predicted values

The estimated Y values for each X calculated by a model

Residual

The difference between observed or actual values and the predicted values of the regression line

Residual = Observed - Predicted

$$\varepsilon_i = y_i - \hat{y}_i$$

Sum of Squared Residuals (SSR)

The sum of the squared differences between each observed value and its associated predicted value

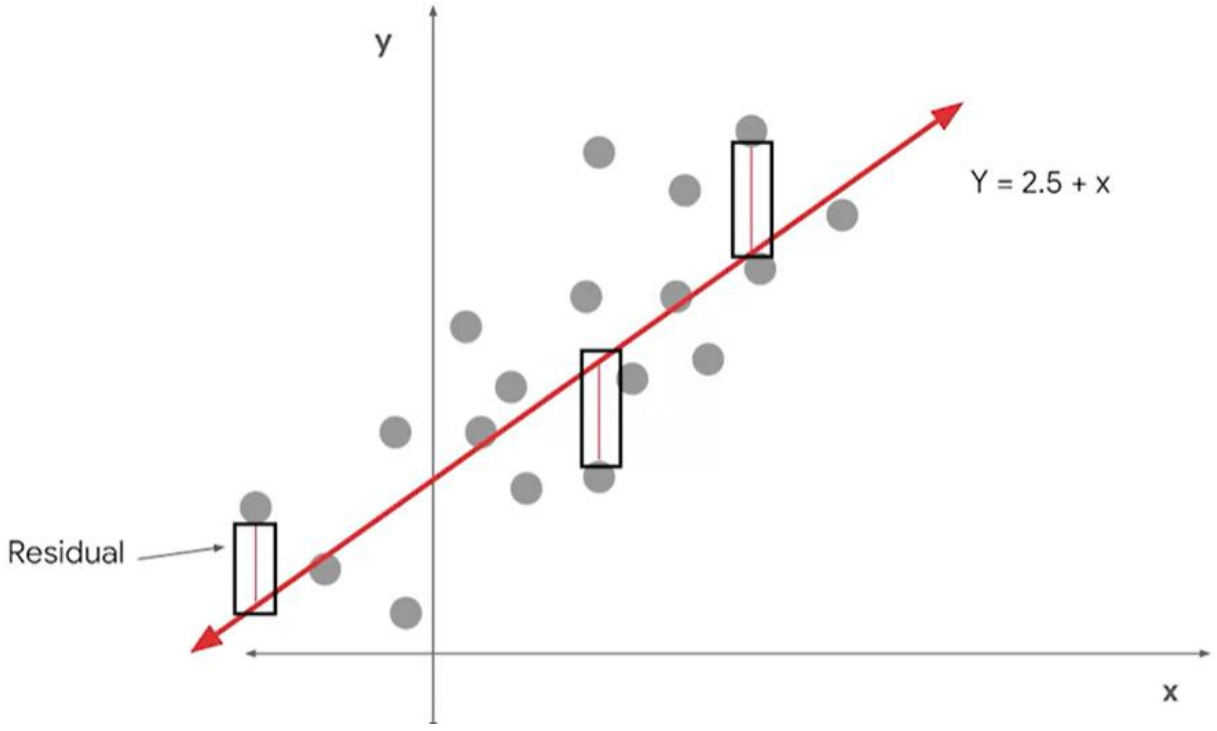
$$\sum_{i=1}^{n} (Observed - Predicted)^2$$

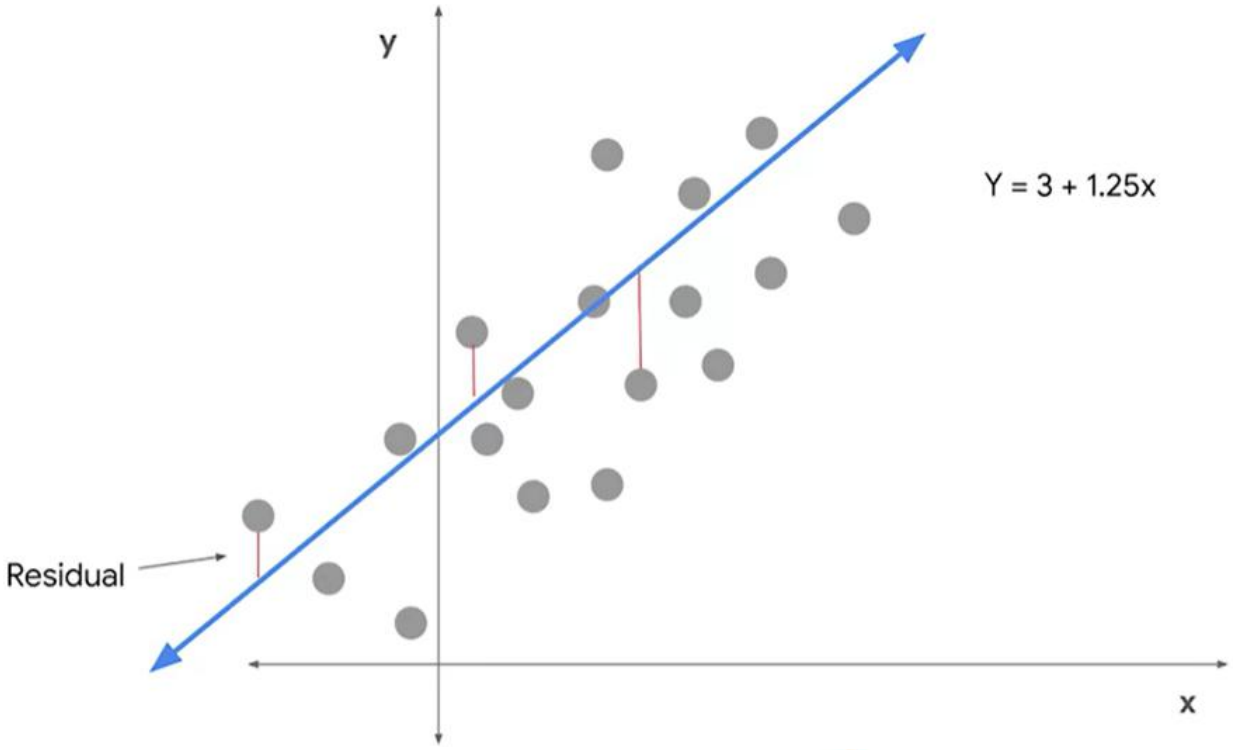
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

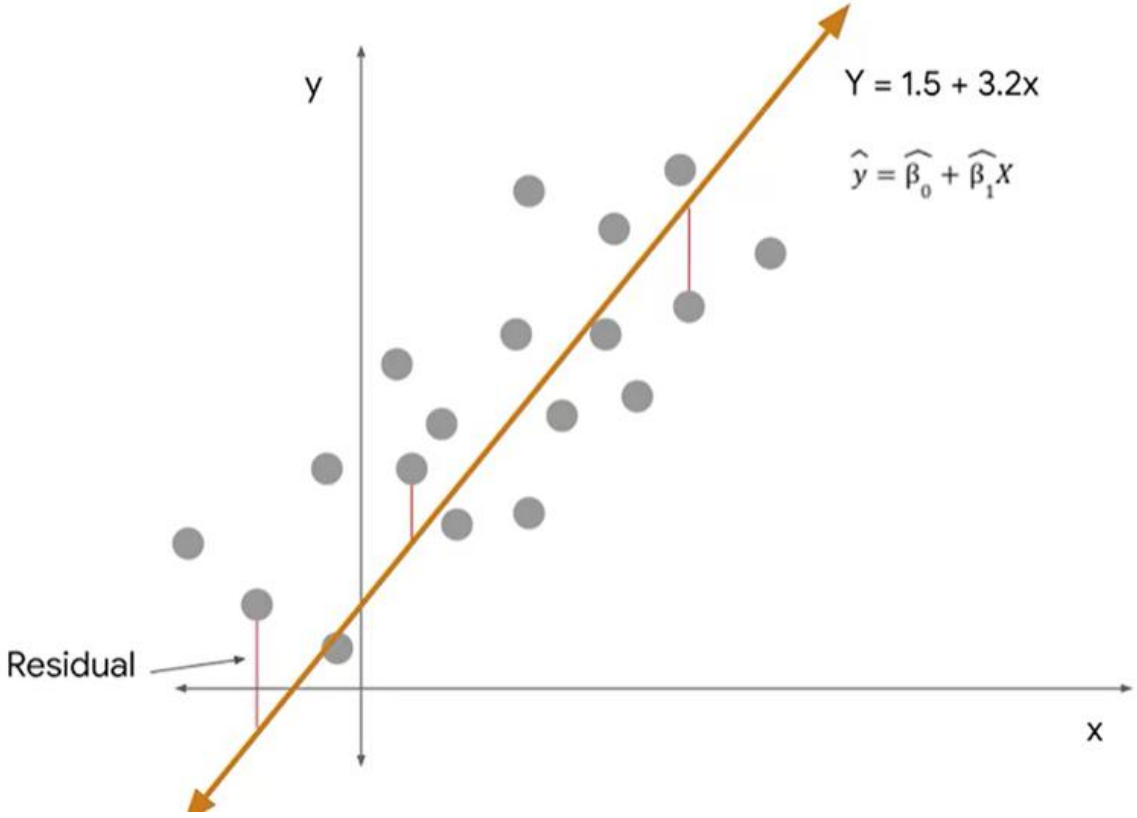
Ordinary Least Squares (OLS)

A method that minimizes the sum of squared residuals to estimate parameters in a linear regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

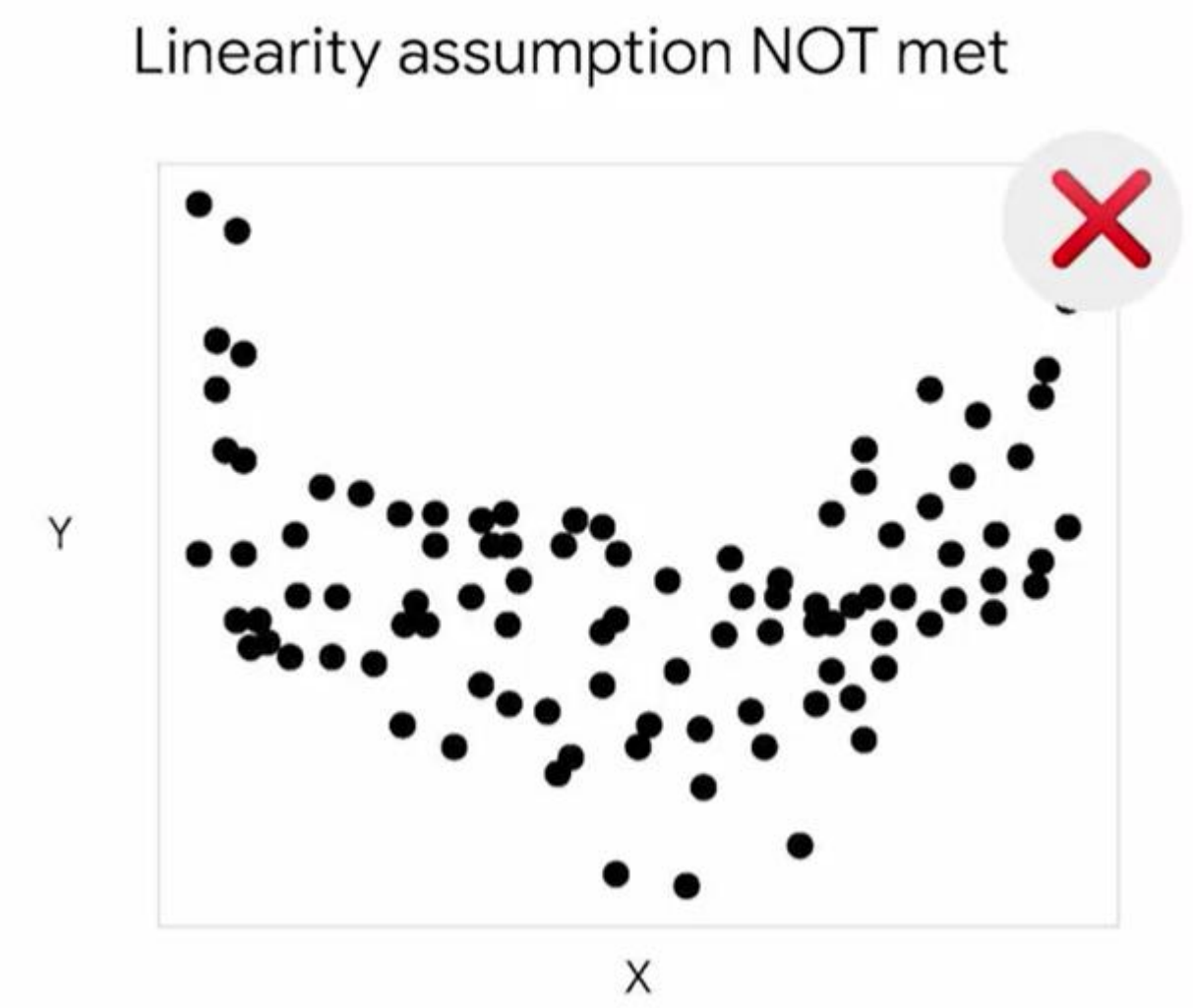


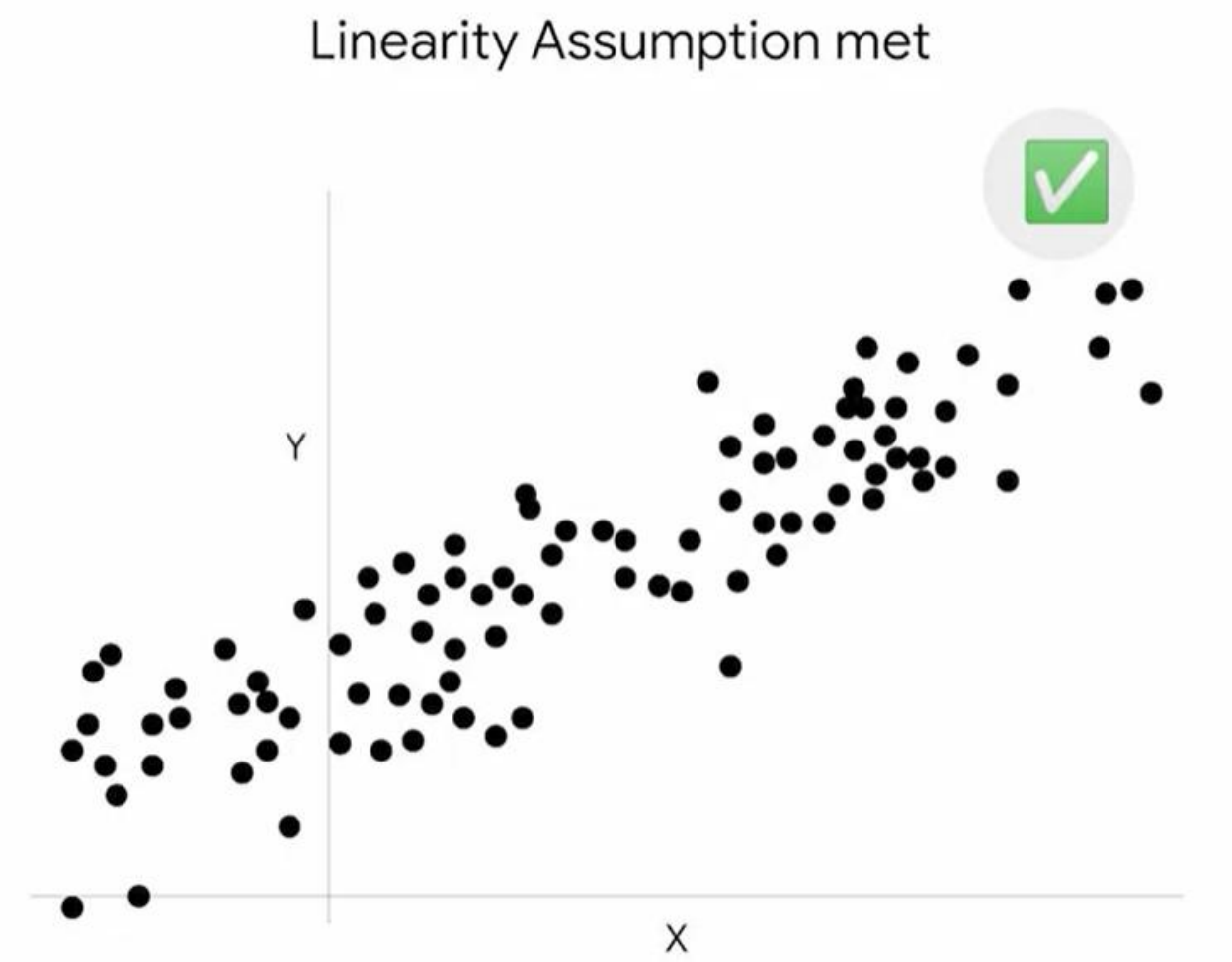

y
Y = 3 + 1.25x
Residual
x


y
Y = 1.5 + 3.2x
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Residual
x

Linear regression assumptions

- Linearity
- Normality
- Independent observations
- Homoscedasticity



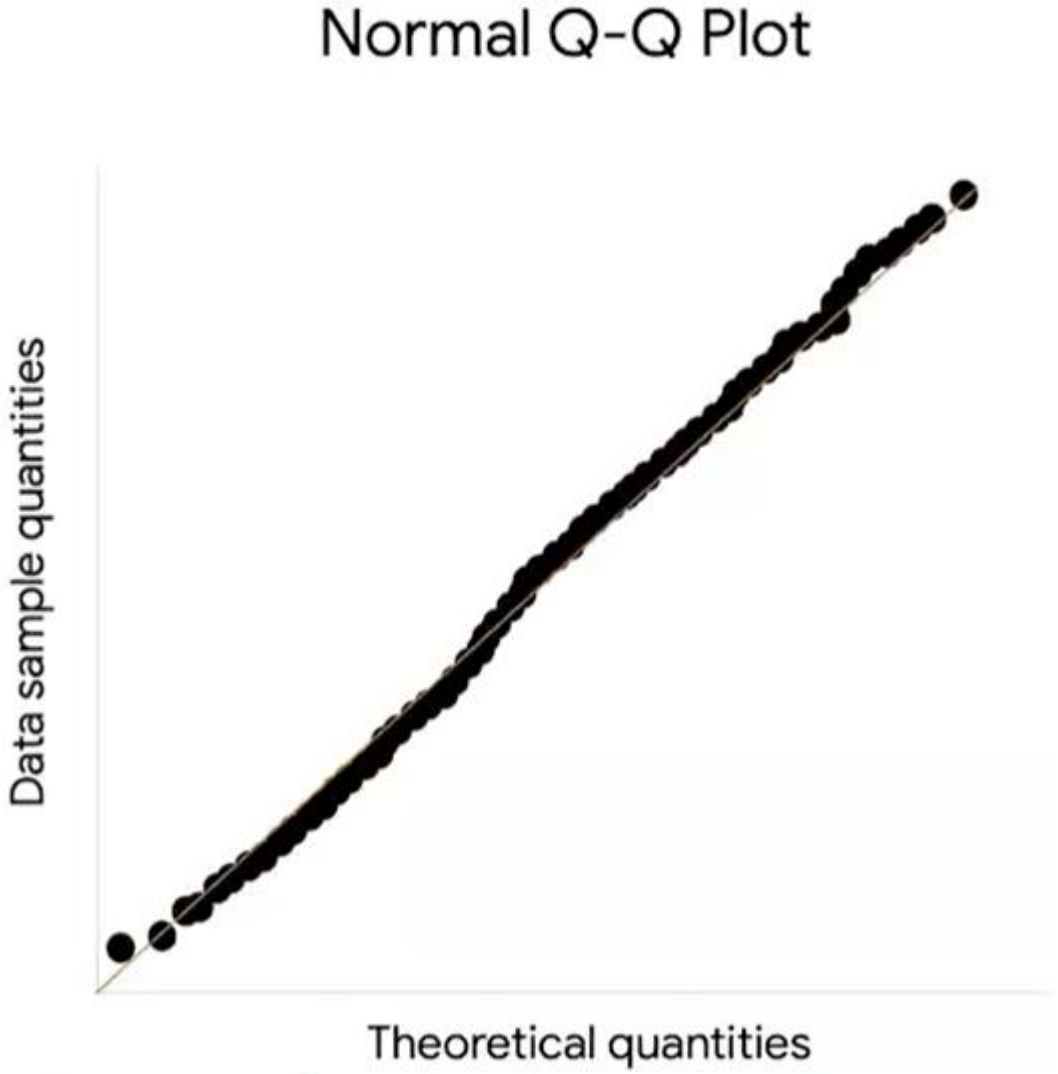


Linearity assumption

Each predictor variable (X_i) is linearly related to the outcome variable (Y)

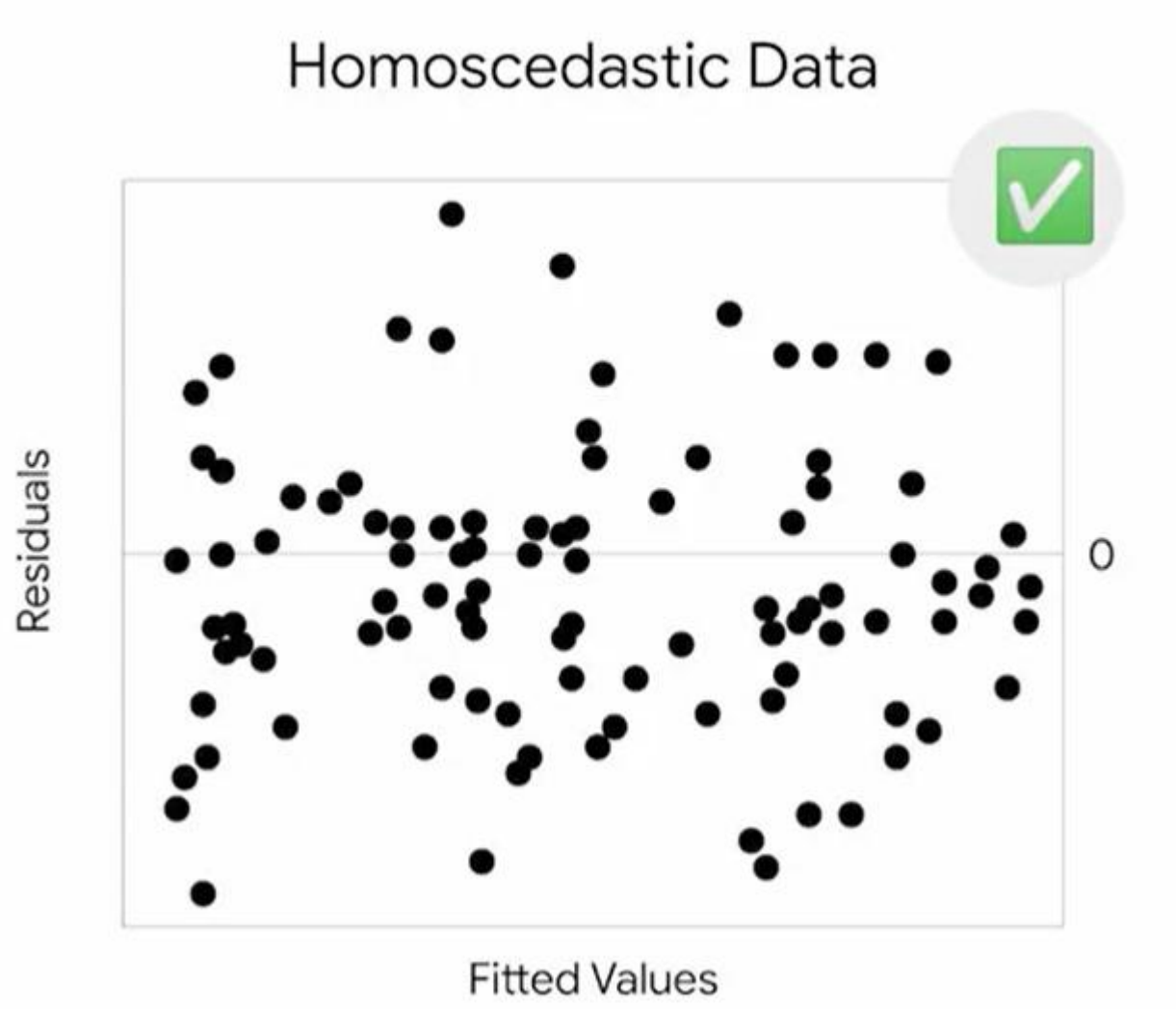
Normality assumption

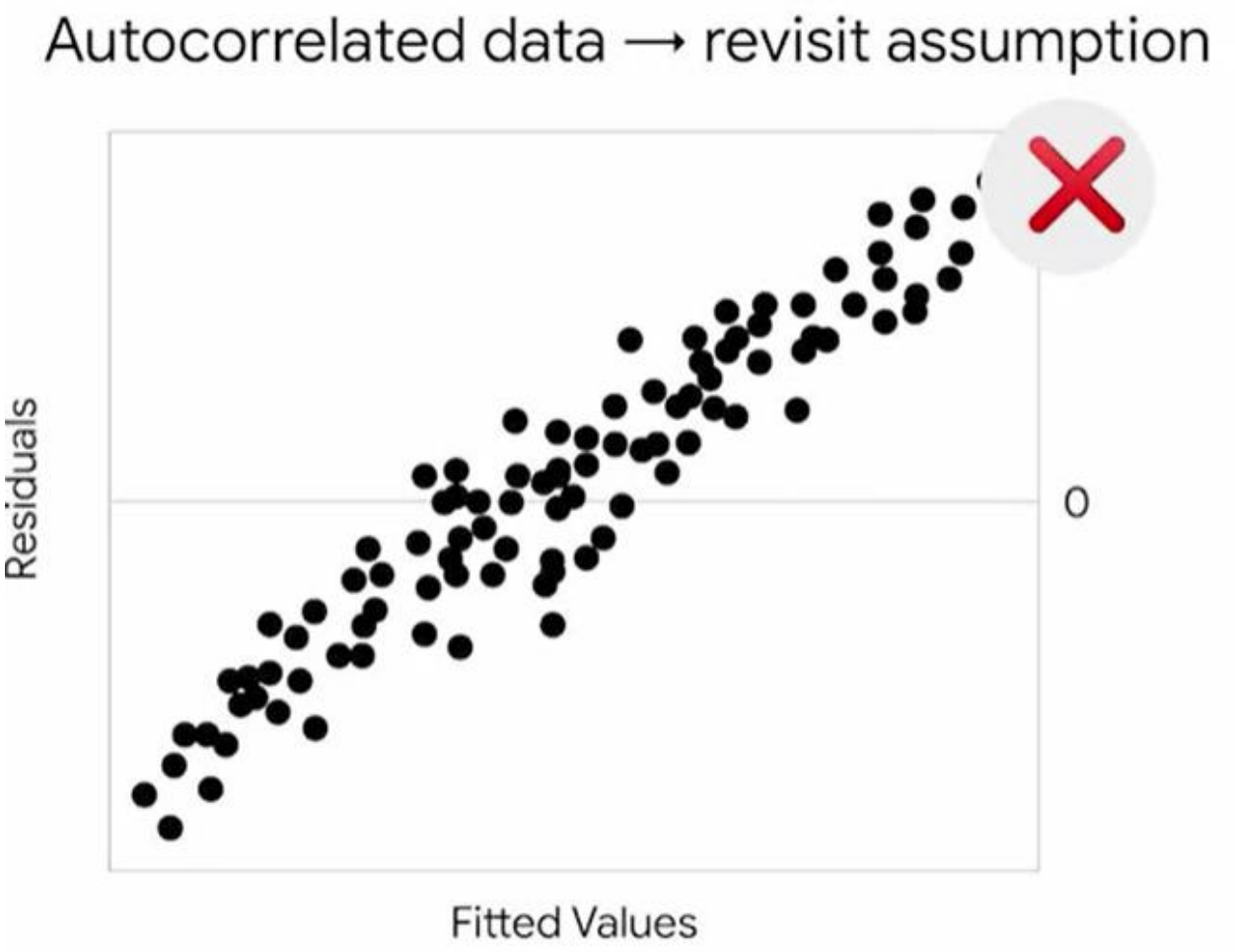
The residuals or errors are normally distributed



Independent observation assumption

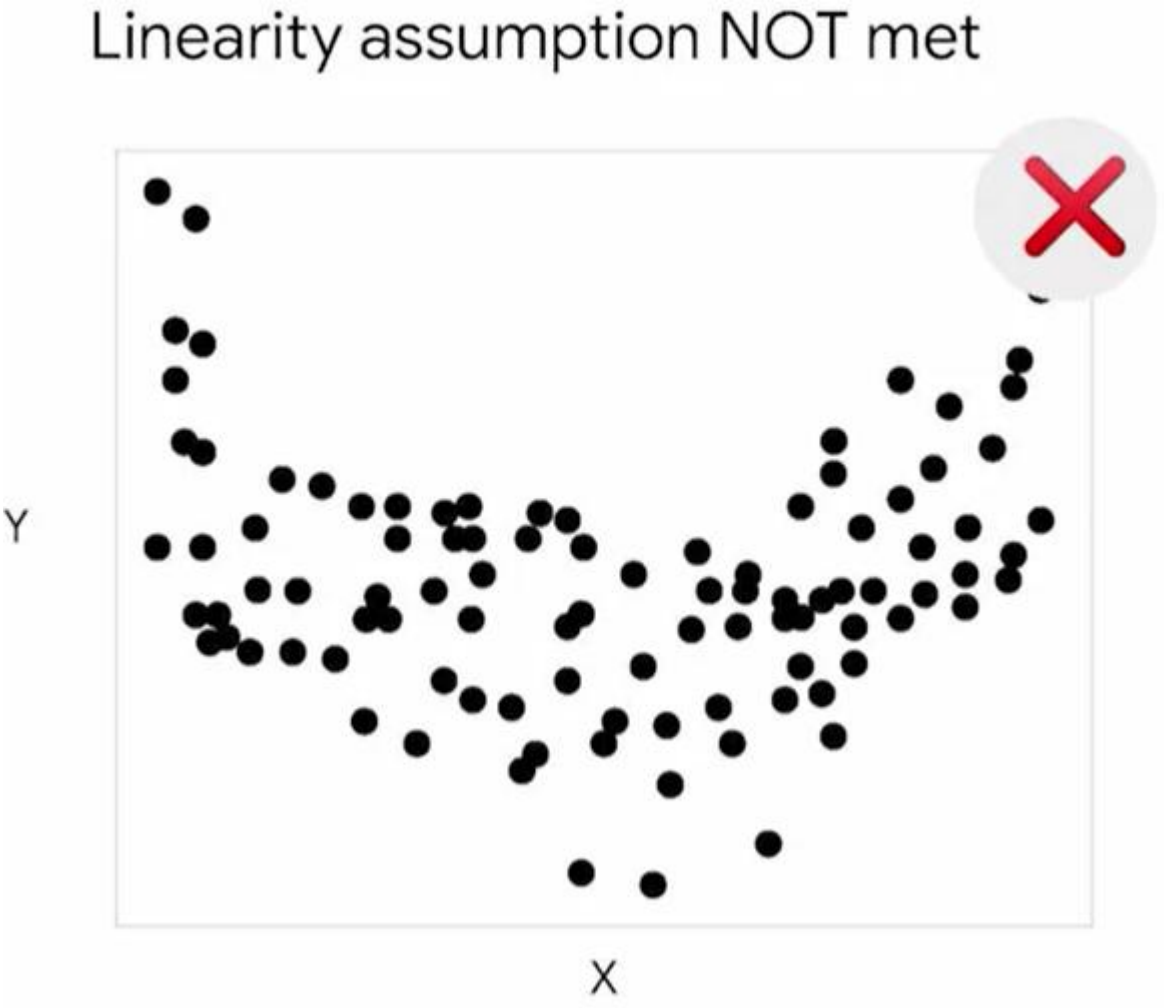
Each observation in the dataset is independent


Homoscedastic Data
Residuals
0
Fitted Values



Homoscedasticity assumption

The variation of the residuals (errors) is constant or similar across the model



Scatterplot matrix

A series of scatterplots that show the relationships between pairs of variables

Y-intercept (β_0)

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1707.2919	205.640	-8.302	0.000	-2112.202	-1302.382
bill_length_mm	141.1904	4.775	29.569	0.000	131.788	150.592

Slope (β_1)

Omnibus:	2.060	Durbin-Watson:	2.067

*Body mass (g) = -1707.30 + 141.19 * bill length (mm)*

Confidence interval

A range of values that describes the uncertainty surrounding an estimate

Confidence band

The area surrounding the line that describes the uncertainty around the predicted outcome at every value of X

Recall that we can represent a simple linear regression line as $y = \beta_0 + \beta_1 X$.

Since regression analysis utilizes **estimation** techniques, there is always a level of uncertainty surrounding the predictions made by regression models. To represent the error, we can actually rewrite the equation to include an error term, represented by the letter ϵ (pronounced "epsilon"): $y = \beta_0 + \beta_1 X + \epsilon$.

Common evaluation metrics

- R^2
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)

R^2 (The coefficient of determination)

Measures the proportion of variation
in the dependent variable, Y,
explained by the independent variable(s), X

Hold-out sample

A random sample of observed data
that is not used to fit the model

R^2 measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X.

- This is calculated by subtracting the sum of squared residuals divided by the total sum of squares from 1.

$$R^2 = 1 - \frac{\text{Sum of squared residuals}}{\text{Total sum of squares}}$$

MSE (mean squared error) is the average of the squared difference between the predicted and actual values.

- Because of how MSE is calculated, MSE is very sensitive to large errors.

MAE: Mean absolute error

MAE (mean absolute error) is the average of the absolute difference between the predicted and actual values.

- If your data has outliers that you want to ignore, you can use MAE, as it is not sensitive to large errors.

body mass (g) = intercept + slope * bill length

Causation

A cause-and-effect relationship where one variable directly causes the other to change in a particular way

Multiple linear regression or multiple regression

A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

$$y = \beta_0 + \beta_1 X$$

y-intercept slope

$$y = \beta_0 + \beta_1 X$$

Number of website clicks

Number of people in the advertisment

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Number of people in the advertisement

Length of the advertisement

Categorical Variable 1	Categorical Variable 2	Categorical Variable 3
Ad Color	**Call to Action**	**Streaming Service**
Black-and-white Color	Call to action No call to action	Service A Service B Service C

One hot encoding

A data transformation technique that turns one categorical variable into several binary variables

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{action} X_{action}$$

Advertisement with call-to-action $\Rightarrow$

$$X_{action} = 1$$

Advertisement with NO call-to-action $\Rightarrow$

$$X_{action} = 0$$

# of categories	# of binary variables
2	1
3	2

$X_{service\ A}$	Service A	Service B	Service C
1	Ad plays on service A	Ad does NOT play on service B	Ad does NOT play on service C
0	Ad does NOT play on service A	Ad plays on EITHER service B OR C	

$X_{service\ A}$	$X_{service\ B}$	Service A	Service B	Service C
1	0	Plays on service A	Does not play on service B	Does not play on service C
0	1	Does not play on service A	Plays on service B	Does not play on service C
0	0	Does not play on service A	Does not play on service B	Plays on service C

Linearity assumption

Each predictor variable (X_i) is linearly related to the outcome variable (Y)

Independent observation assumption

Each observation in the dataset is independent

Normality assumption

The residuals are normally distributed

Homoscedasticity assumption

The variation of the residuals (errors) is constant or similar across the model

No multicollinearity assumption

No two independent variables (X_i and X_j) can be highly correlated with each other

Correct

Feedback: The no multicollinearity assumption states that no two independent variables ($X - i$ and $X - j$) can be highly correlated with each other. This means that $X - i$ and $X - j$ cannot be linearly related to each other.

Concert sale data

- Dependent variable (Y): concert sales
- Independent variables (X):
  - Number of social media followers
  - Number of streams on music platforms
  - Year the artist debuted
  - Cost of the ticket
  - Days until concert

Variance Inflation Factors (VIF)

Quantifies how correlated each independent variable is with all of the other independent variables

Data set
Y : dependent variable
X_1: independent variable 1
X_2: independent variable 2
X_3: independent variable 3

Simple linear regression example

$$\text{sales} = -44 + 2.2 * \text{temperature}$$

$$\text{sales} = \beta_0 + \beta_{\text{temperature}} * X_{\text{temperature}} + \beta_{\text{ad}} * X_{\text{ad}}$$

$$\text{sales} = \beta_0 + \beta_{\text{temperature}} * 75 + \beta_{\text{ad}} * 1$$

$$\text{sales} = \beta_0 + \beta_{\text{temperature}} * 75 + \beta_{\text{ad}} * 0 = \beta_0 + \beta_{\text{temperature}} * 75$$

$$\text{sales} = \beta_0 + \beta_{\text{temperature}} * X_{\text{temperature}} + \beta_{\text{transportation}} * X_{\text{transportation}}$$

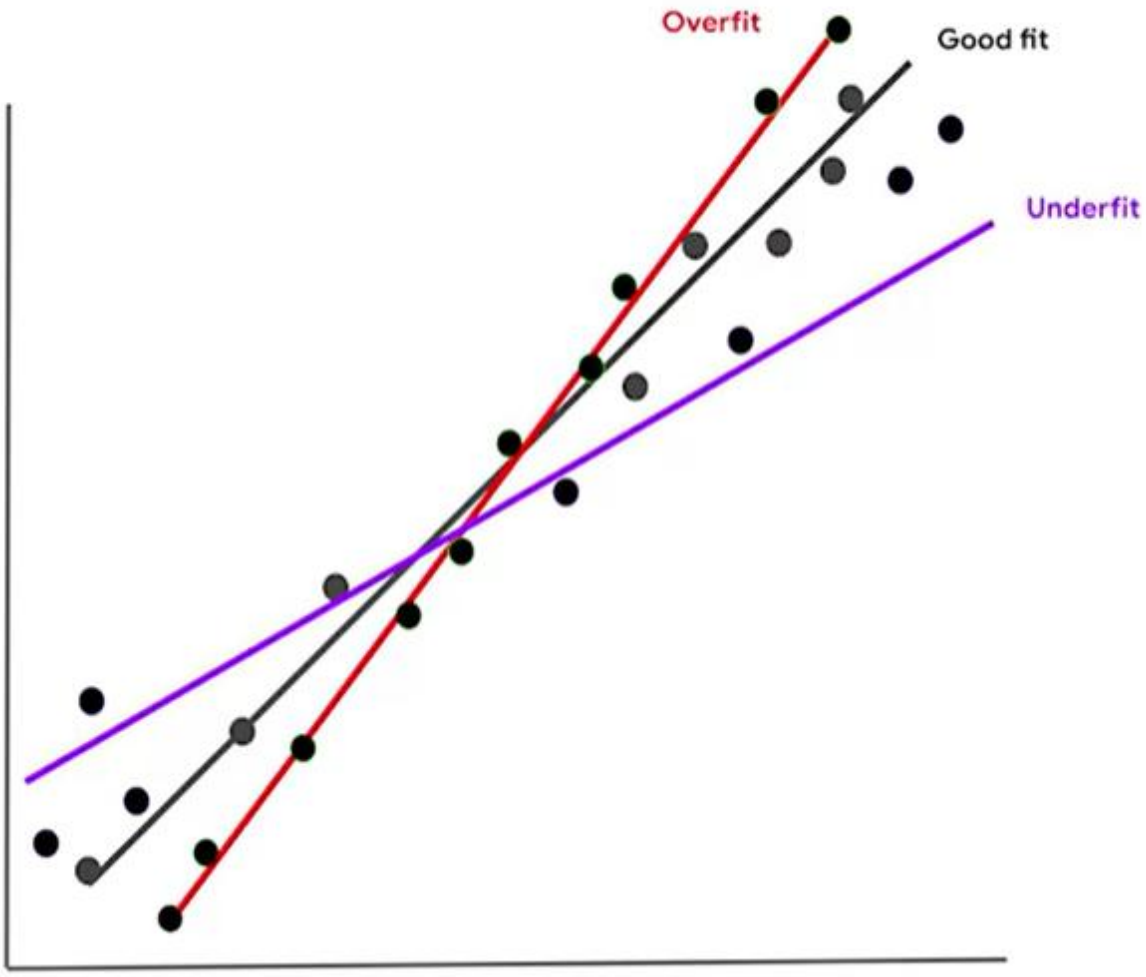
For every 1 degree increase in the temperature, while holding distance to public transportation constant, we expect iced coffee sales to increase by $\beta_{temperature}$

For every 1 kilometer further a store is from public transportation, while holding temperature constant, we expect iced coffee sales to decrease by $\beta_{transportation}$

Interaction term

A term that represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

$$sales = \beta_0 + \beta_{temperature} * X_{temperature} + \beta_{transportation} * X_{transportation} + \beta_{interaction} * (temperature * transportation)$$



Overfitting

When a model fits the observed or training data too specifically, and is unable to generate suitable estimates for the general population

Adjusted R^2

A variation of the R^2 regression evaluation metric that penalizes unnecessary explanatory variables

Adjusted R^2 vs. R^2

- Adjusted R^2 is used to compare models of varying complexity
  - Determine if you should add another variable or not

- R^2 is more easily interpretable
  - Determine how much variation in the dependent variable is explained by the model

Variable selection or feature selection

The process of determining which variables or features to include in a given model

Forward selection

A stepwise variable selection process that begins with the null model, with 0 independent variables, considers all possible variables to add. It incorporates the independent variable that contributes the most explanatory power to the model.

Backward elimination

A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

Extra-sum-of-squares F-test

Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

Bias-variance tradeoff

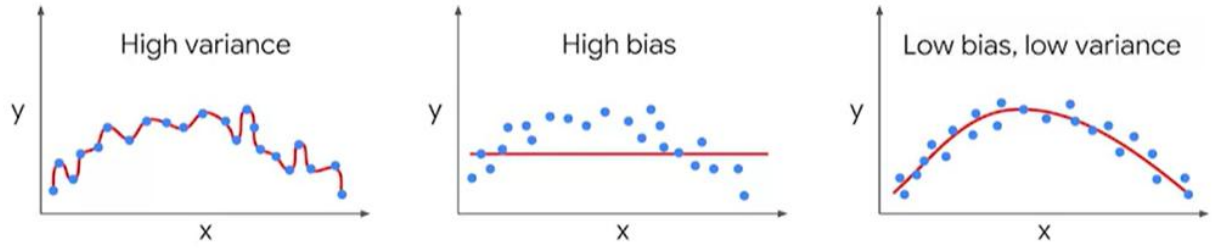
Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

Bias

Simplifies the model predictions by making assumptions about the variable relationships. A highly biased model may oversimplify the relationship, underfitting to the observed data, and generating inaccurate estimates.

Variance

Model flexibility and complexity, so the model learns from existing data. A model with high variance can overfit to observed data and generate inaccurate estimates for unseen data.



overfitting | underfitting | good balance

Regularization

A set of regression techniques that shrinks regression coefficient estimates toward zero, adding in bias, to reduce variance

Regularized regression

- Lasso regression
- Ridge regression
- Elastic-net regression

Chi-squared [χ^2] tests will help us determine if two categorical variables are associated with one another, and whether a categorical variable follows an expected distribution.

t tests	X^2 tests
Null & alternative hypotheses	Null & alternative hypotheses
Continuous data	Categorical data

χ^2 Goodness of fit test

Determines whether an observed categorical variable follows an expected distribution

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

χ^2 Goodness of fit test

- Null Hypothesis (H^0)
  - 25 people buy each size of popcorn on any given day
  - The variable follows the expected distribution
- Alternative Hypothesis (H_1)
  - The variable does NOT follow the expected distribution
  - Different numbers of people buy each size of popcorn on any given day

χ^2 Test for independence

Determines whether or not two categorical variables are associated with each other

χ^2 Test for independence

- Variable 1: Weather Precipitation
  No Precipitation
- Variable 2: Popcorn Sales
  100+ people buy popcorn $\leq$ 100 people buy popcorn

Observed Counts

	Rain	No Rain	Totals
100+ popcorn	58	77	135
≤ 100 popcorn	25	115	140
Totals	83	192	275

Analysis of Variance (ANOVA)

A group of statistical techniques that test the difference of means between three or more groups

One-way ANOVA

Compares the means of one continuous dependent variable based on three or more groups of one categorical variable

One-way ANOVA hypotheses

- Null Hypothesis (H_0)

  $H_0 : \mu_{monarch} = \mu_{mourning\ cloak} = \mu_{swallowtail}$

  - The means of each group are equal

- Alternative Hypothesis (H^1)

  $H_1 : \text{NOT } \mu_{monarch} = \mu_{mourning\ cloak} = \mu_{swallowtail}$

  - The means of each group are NOT all equal

Two-way ANOVA

Compares the means of one continuous dependent variable based on three or more groups of two categorical variables

Two-Way ANOVA Hypotheses

	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Species	There is no difference in life spans between the three butterfly species.	There is a difference in life spans between the three butterfly species.
Size	There is no difference in life spans based on butterfly size.	There is a difference in life spans based on butterfly size.
Species & Size Interaction Effect	The effect of species on life span is independent of the butterfly size, and vice versa.	There is an interaction effect between butterfly size and species on life span.

Two-Way ANOVA Hypotheses

	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Color	There is no difference in diamond price based on color.	There is a difference in diamond price based on color.
Cut	There is no difference in diamond price based on cut.	There is a difference in diamond price based on cut.
Color & Cut Interaction Effect	The effect of color on diamond price is independent of the cut, and vice versa.	There is an interaction effect between color and cut on diamond price.

Two-way ANOVA results

- Logarithm of the price is NOT the same for different colors
- Logarithm of the price is NOT the same for different diamond cuts
- There is an interaction effect between the color and cut that impacts the price of the diamond

Post hoc test

Performs a pairwise comparison between all available groups while controlling for the error rate

Covariates are the variables that are not of direct interest to the question to be answered. Analysis of covariance, or ANCOVA, is a statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates.

ANCOVA (Analysis of covariance)

A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to your test

ANCOVA	Linear Regression
Continuous and categorical independent variables	Continuous categorical independent variables
Continuous Y variables	Continuous Y variable
Understand variable relationships	Understand variable relationships
Not focused on covariates	Could be interested in all of the independent variables
Focus on categorical independent variable	Could be interested in predicting Y variable

MANOVA (Multivariate analysis of variance)

- One-way MANOVA
  - One categorical independent variable
- Two-way MANOVA
  - Two categorical independent variables
- Continuous outcome variables

MANCOVA
(Multivariate analysis of covarience)

An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates

One-way ANOVA is a powerful way to determine if there are differences in a continuous outcome variable between groups you're interested in.

Two-way ANOVA allows you to gain similar insights, while incorporating another set of groups as well.

MANCOVA built upon ANOVA and ANCOVA to allow testing of multiple outcome variables of interest.

Logit (log-odds)

The logarithm of the odds of a given probability. So the logit of probability p is equal to the logarithm of p divided by 1 minus p.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Logit in terms of X variables

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Maximum likelihood estimation (MLE)

A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data

Likelihood

The probability of observing the actual data, given some set of beta parameters

Because the observations are independent

$$P(A \text{ AND } B) = P(A) * P(B)$$

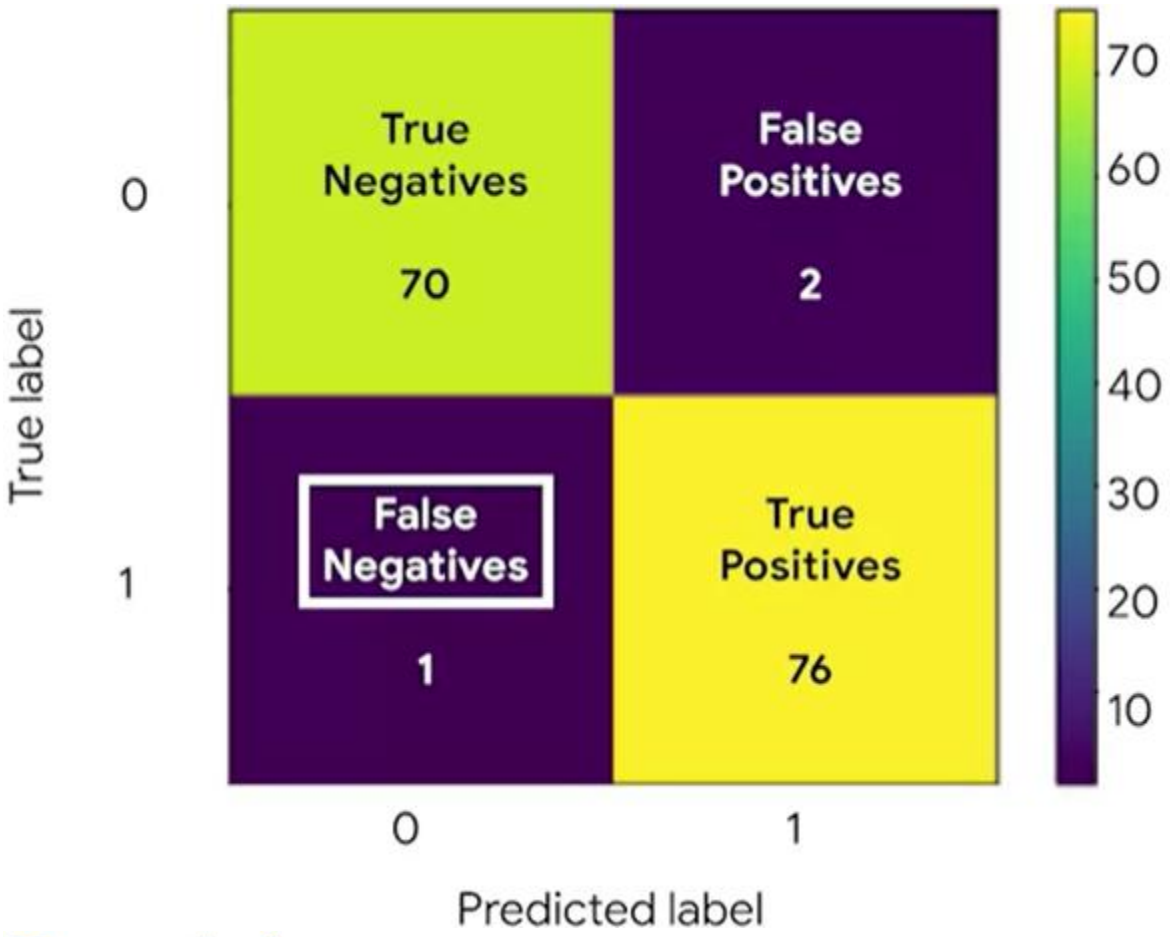
The best logistic regression model estimates the set of beta coefficients that maximizes the likelihood of observing all of the sample data.

Binomial logistic regression assumptions

- Linearity
- Independent observations
- No multicollinearity
- No extreme outliers

Confusion matrix

A graphical representation of how accurate a classifier is at predicting the labels for a categorical variable



Precision

The proportion of positive predictions that were true positives

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Evaluation metrics

- Precision
- Recall
- Accuracy

```
import sklearn.metrics as metrics
metrics.precision_score(y_test,y_pred))
```

Recall

The proportion of positives the model was able to identify correctly

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

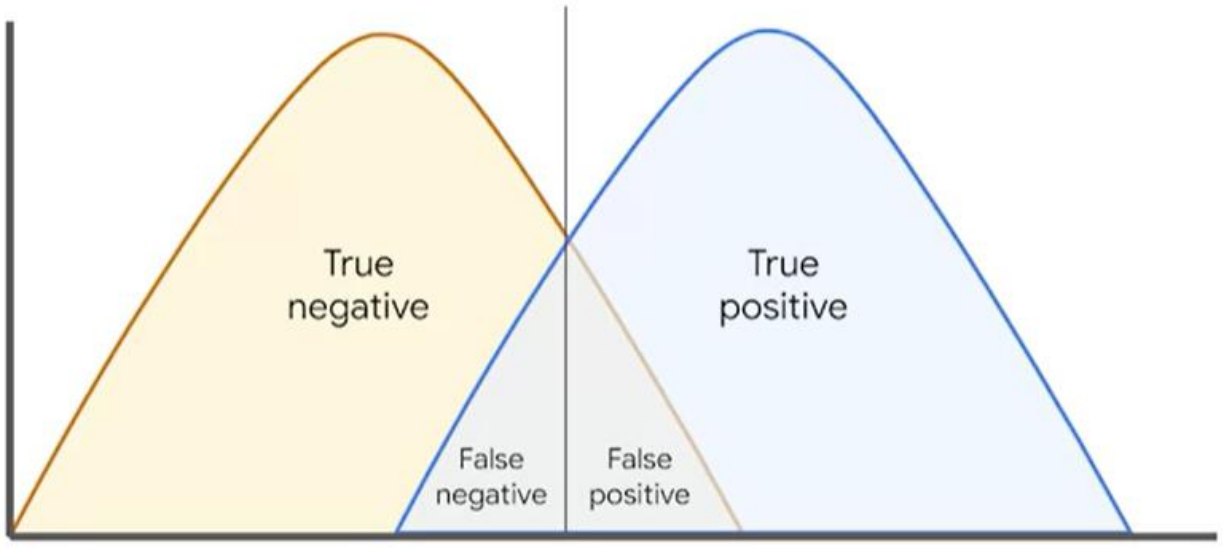
```
metrics.recall_score(y_test,y_pred)
```

Accuracy

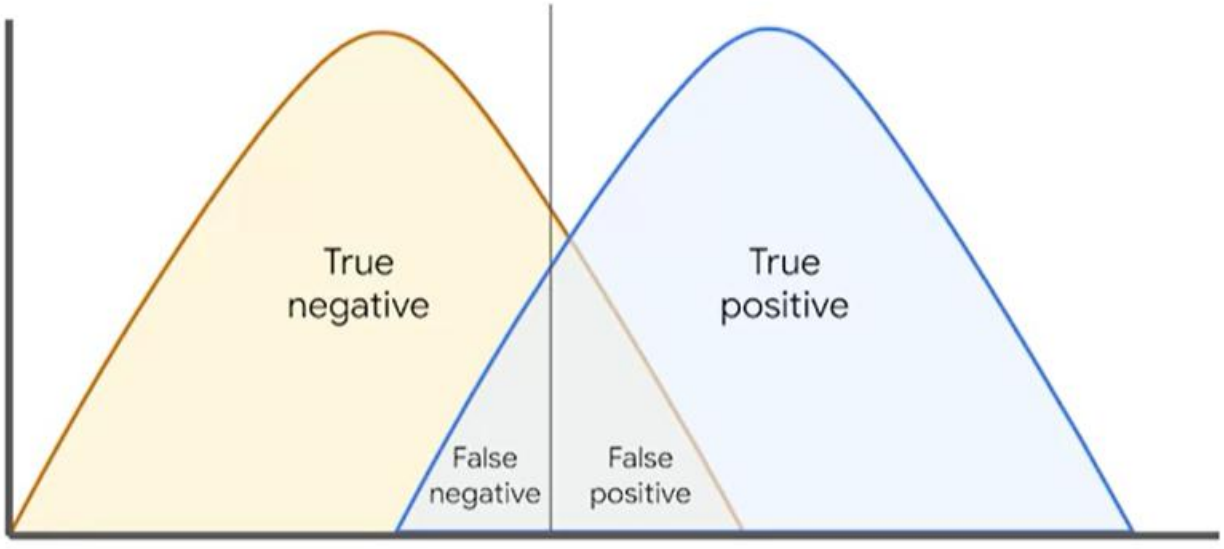
The proportion of data points that were correctly categorized

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

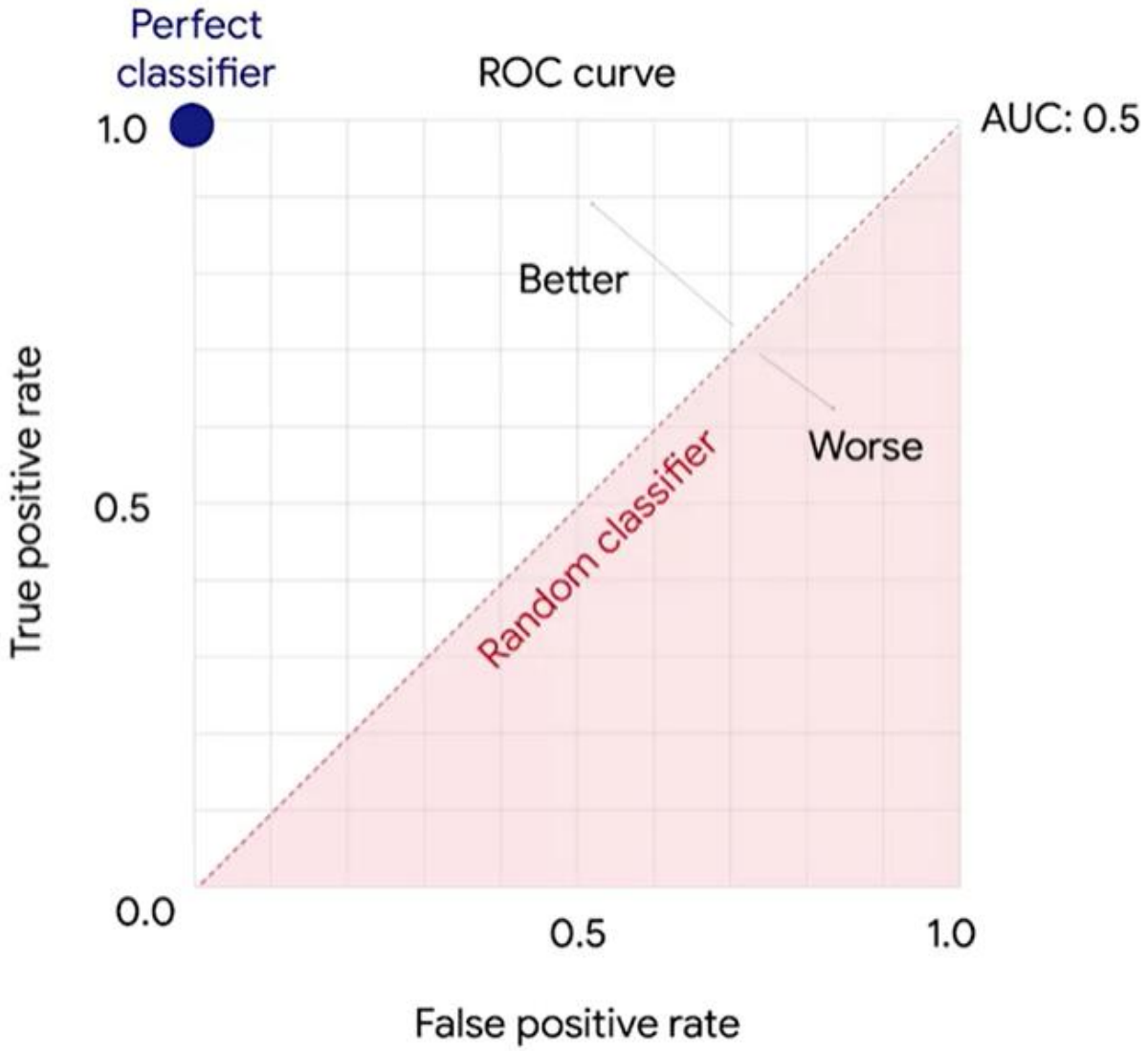
```
metrics.accuracy_score(y_test,y_pred))
```

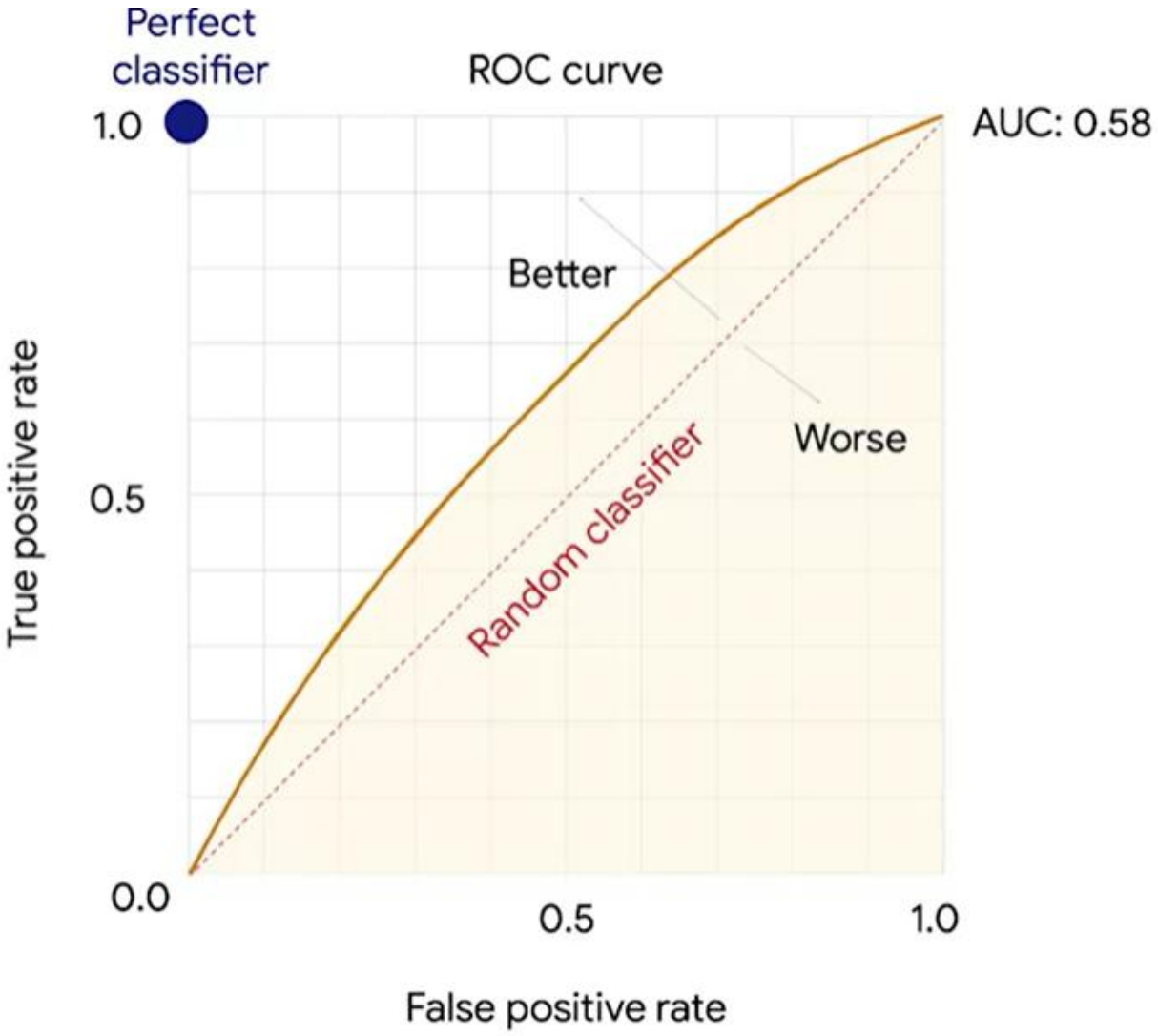


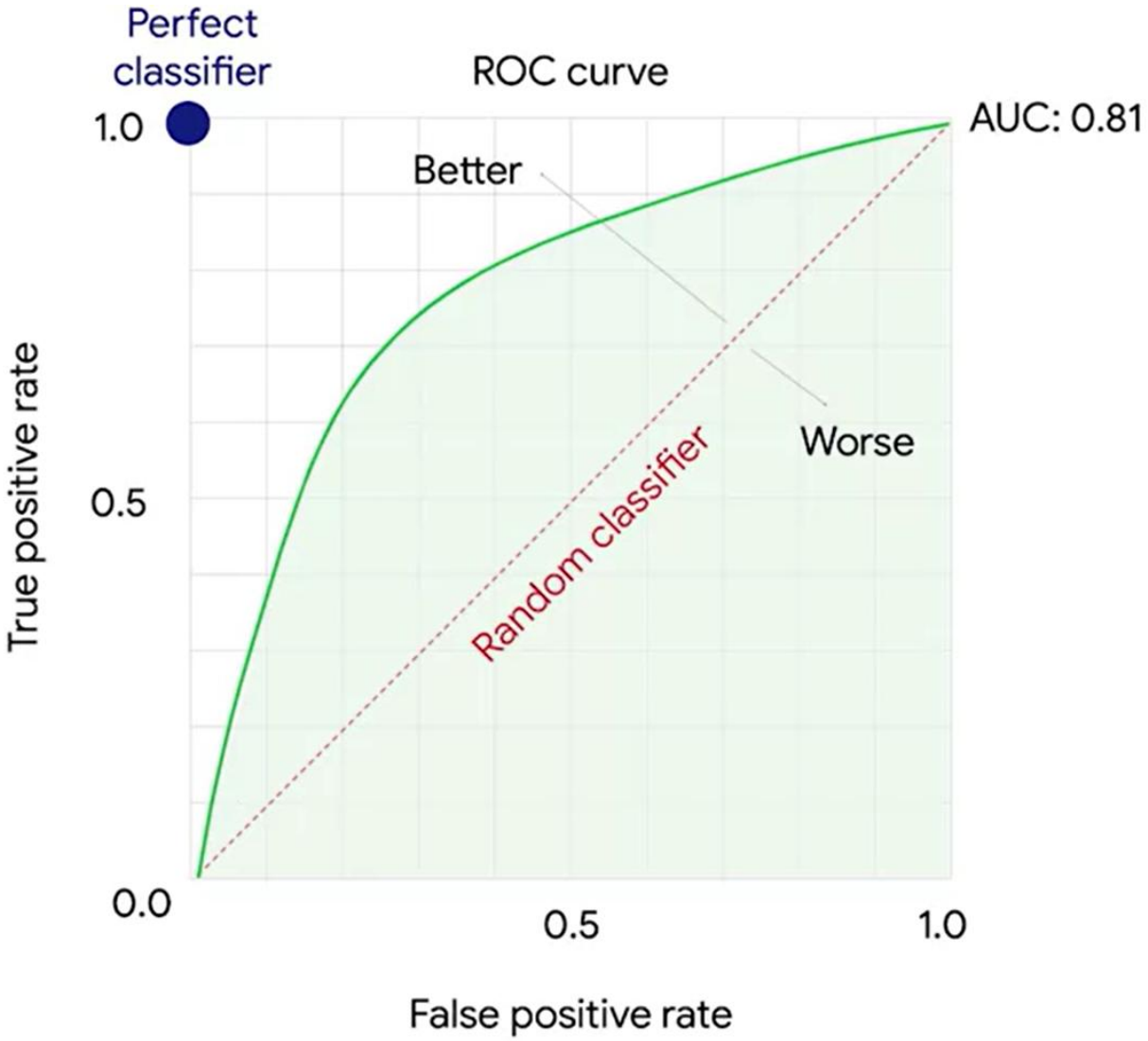
0.5
Threshold

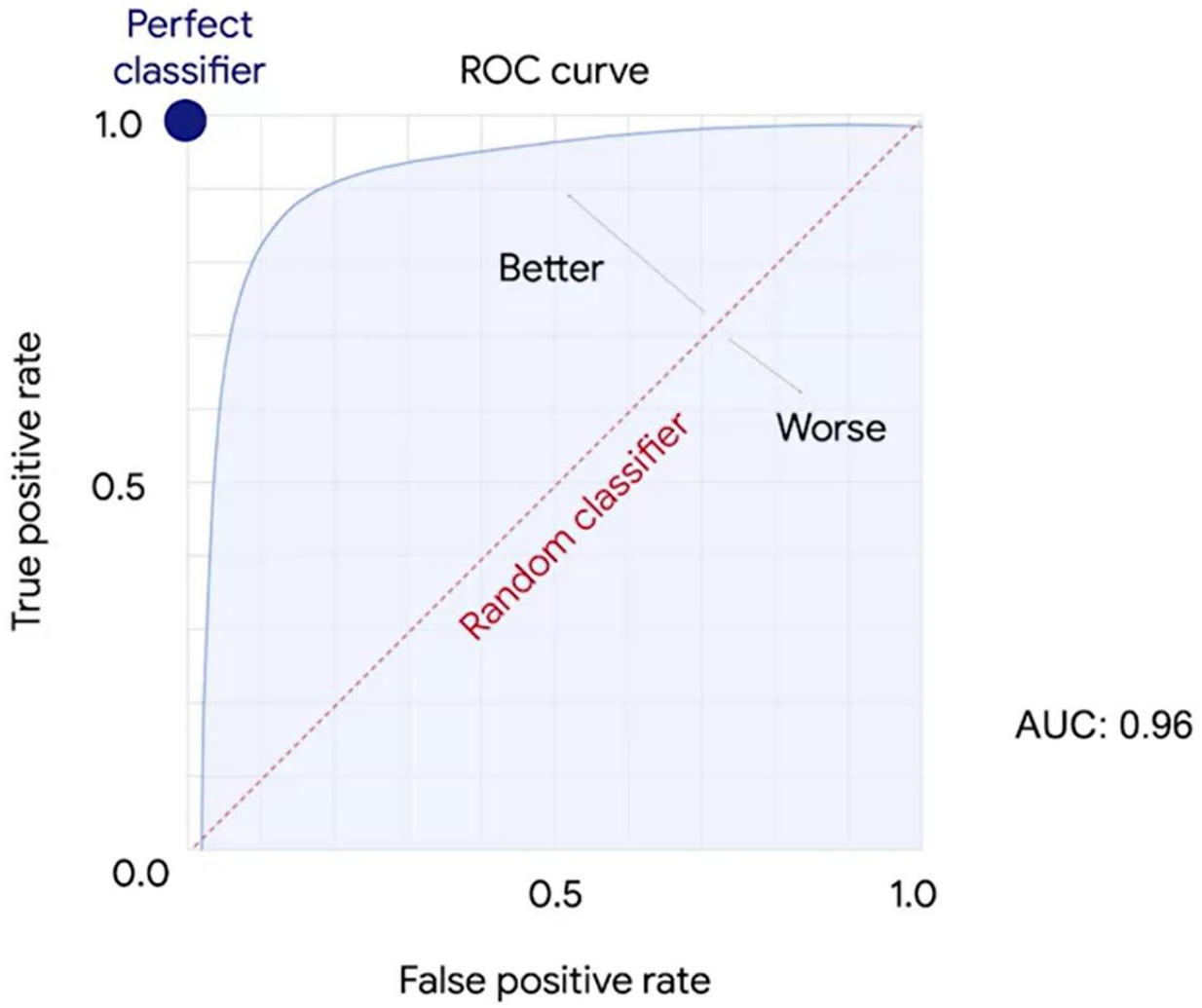


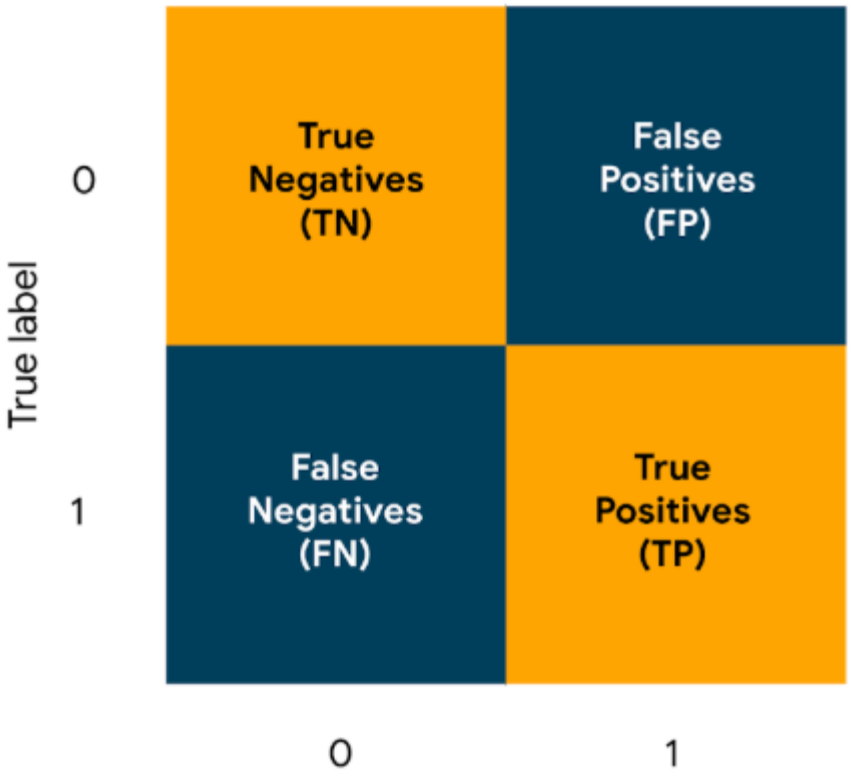
0.4
Threshold


Perfect classifier
ROC curve
AUC: 0.5
1.0
Better
Worse
Random classifier
0.5
True positive rate
0.0
0.5
1.0
False positive rate


Perfect classifier
ROC curve
1.0
AUC: 0.58
Better
Worse
Random classifier
True positive rate
0.5
0.0
0.5
1.0
False positive rate


Perfect classifier
ROC curve
AUC: 0.81
1.0
Better
Worse
Random classifier
0.5
True positive rate
0.0
0.5
1.0
False positive rate


Perfect classifier
ROC curve
1.0
Better
Worse
Random classifier
0.5
True positive rate
0.0
0.5
1.0
False positive rate
AUC: 0.96



The four key parts of a confusion matrix, in the context of binary classification, are the following:

1. **True negatives:**

The count of observations that a classifier correctly predicted as False (0)

2. **True positives:**

The count of observations that a classifier correctly predicted as True (1)

3. **False positives:**

The count of observations that a classifier incorrectly predicted as True (1)

4. **False negatives:**

The count of observations that a classifier incorrectly predicted as False (0)

These counts are useful in computing metrics such as precision, recall, accuracy, and ROC for evaluating logistic regression classifiers.

An ROC curve plots two key concepts

1. **True Positive Rate:** equivalent to **Recall.** The formula for True Positive Rate is as follows:

$$True\ Positive\ Rate = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

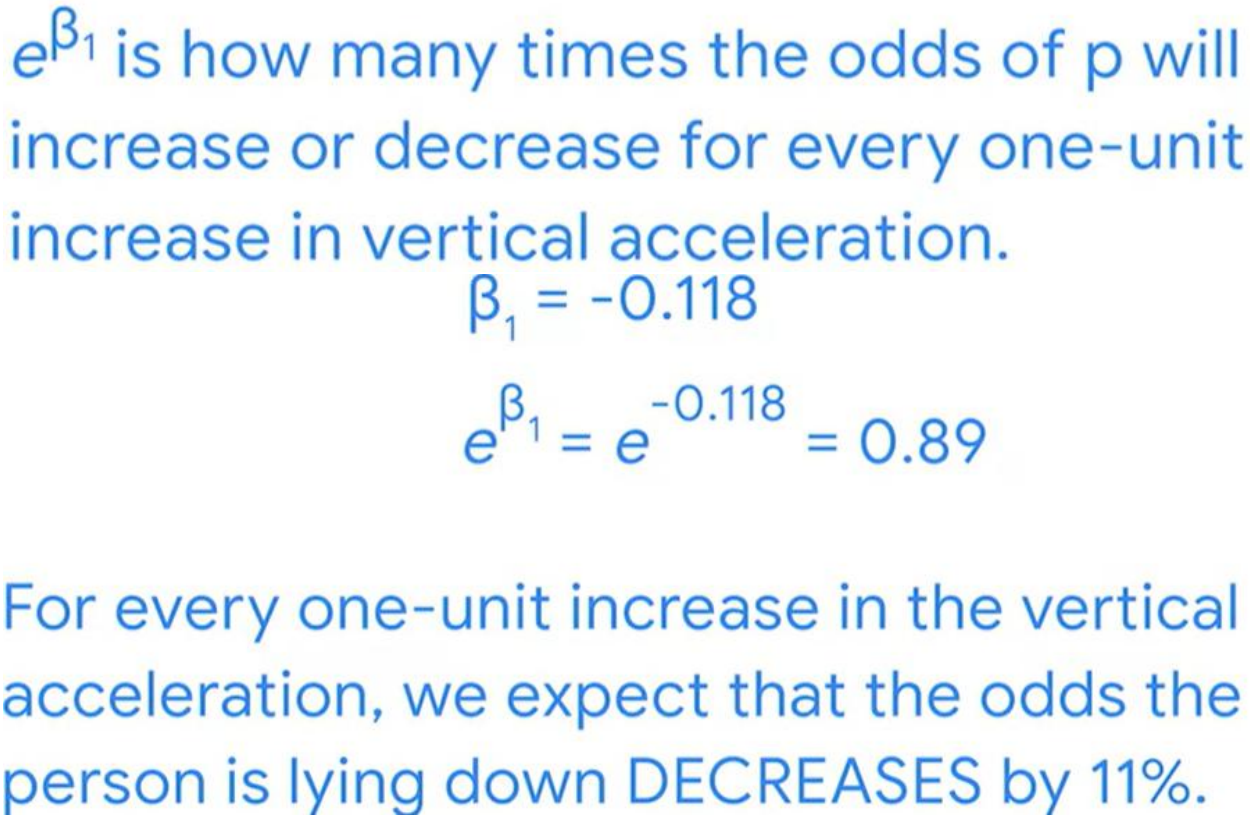
2. **False Positive Rate:** The ratio between the False Positives and the total count of observations that should be predicted as False. The formula for False Positive Rate is as follows:

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives\ +\ True\ Negatives}$$

For each point on the curve, the x and y coordinates represent the False Positive Rate and the True Positive Rate respectively at the corresponding threshold.

AUC

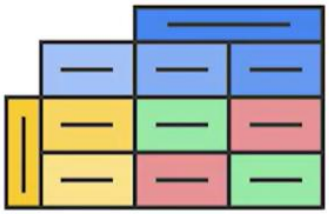
AUC stands for area under the ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0.0 to 1.0. A model whose predictions are 100% wrong has an AUC of 0.0, and a model whose predictions are 100% correct has an AUC of 1.0. An AUC smaller than 0.5 indicates that the model performs worse than a random classifier (i.e. a classifier that randomly assigns each example to True or False), and an AUC larger than 0.5 indicates that the model performs better than a random classifier.



$$\beta_1 = 0.25$$

$$e^{\beta_1} = e^{0.25} = 1.28$$

For every one-unit increase in X, we expect that the odds Y being 1 to INCREASE by 28%.
For every one-unit increase in X, **holding other variables constant**, we expect the odds of Y being 1 to increase by 28%.



Confusion Matrix

0.85

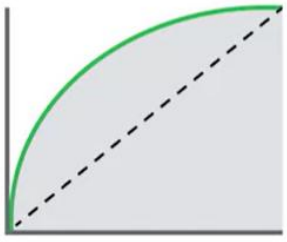
Accuracy

0.818

Precision

0.9

Recall



ROC/AUC

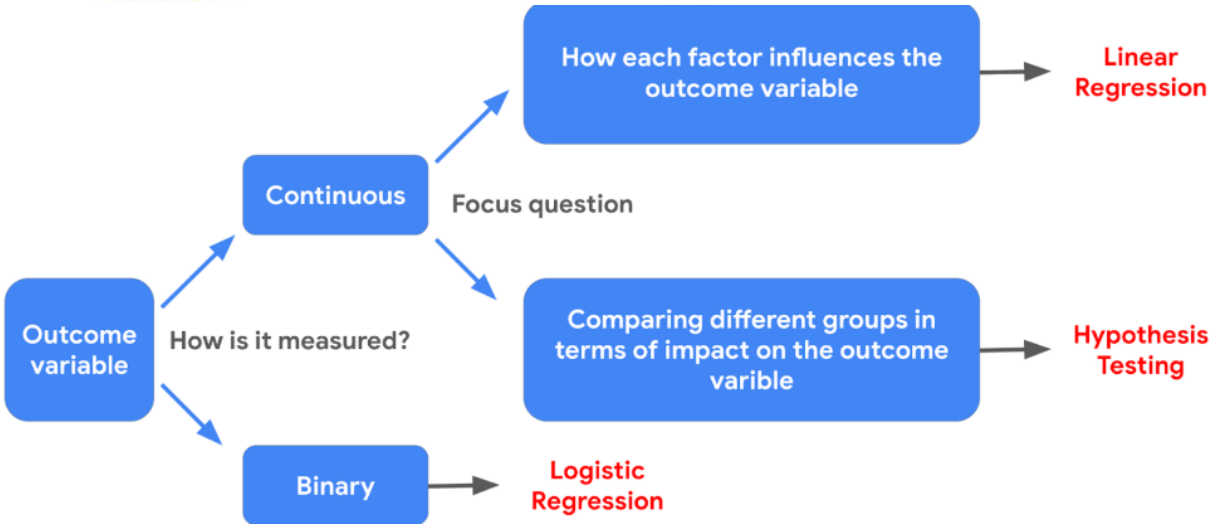
Other metrics

- AIC
- BIC

"Logistic regression coefficients report in percentages how much a factor increases or decreases the **likelihood of an outcome.**"

Evaluating logistic regression

- P-value
- Confusion matrices
- Precision
- Recall
- Accuracy
- ROC/AUC
- AIC
- BIC

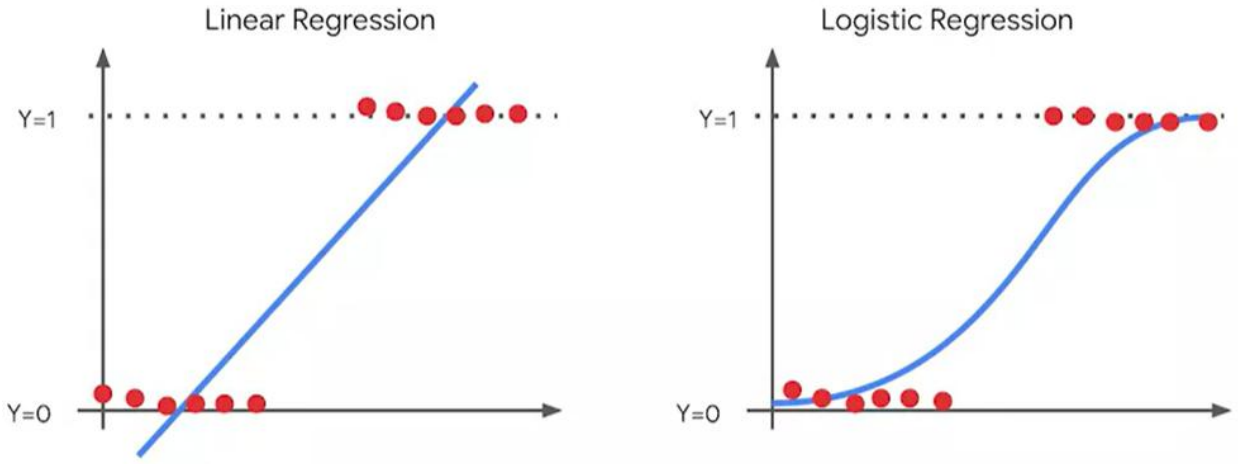


Questions you might get in interviews

- What kind of assumptions do we have for linear regression?
- What should we do if there are outliers?
- How do you determine whether outliers are influential points?
- How do you check multicollinearity and what should be done if there is multicollinearity?

Machine learning

The use and development of algorithms and statistical models to teach computer systems to analyze and discover patterns in data



Supervised machine learning

Uses labeled datasets to train algorithms to classify or predict outcomes

THIS OCCURS MORE OFTEN, IT IS MORE USED. PREDICTION

X DATA ↓

Labeled data

Height (cm)	Bird
45	penguin
101	penguin
179	ostrich
271	ostrich
115	penguin
76	penguin
244	ostrich
63	penguin
228	ostrich

← Y data

↑ labels

Supervised machine learning

A category of machine learning that uses labeled datasets to train algorithms to classify or predict outcomes

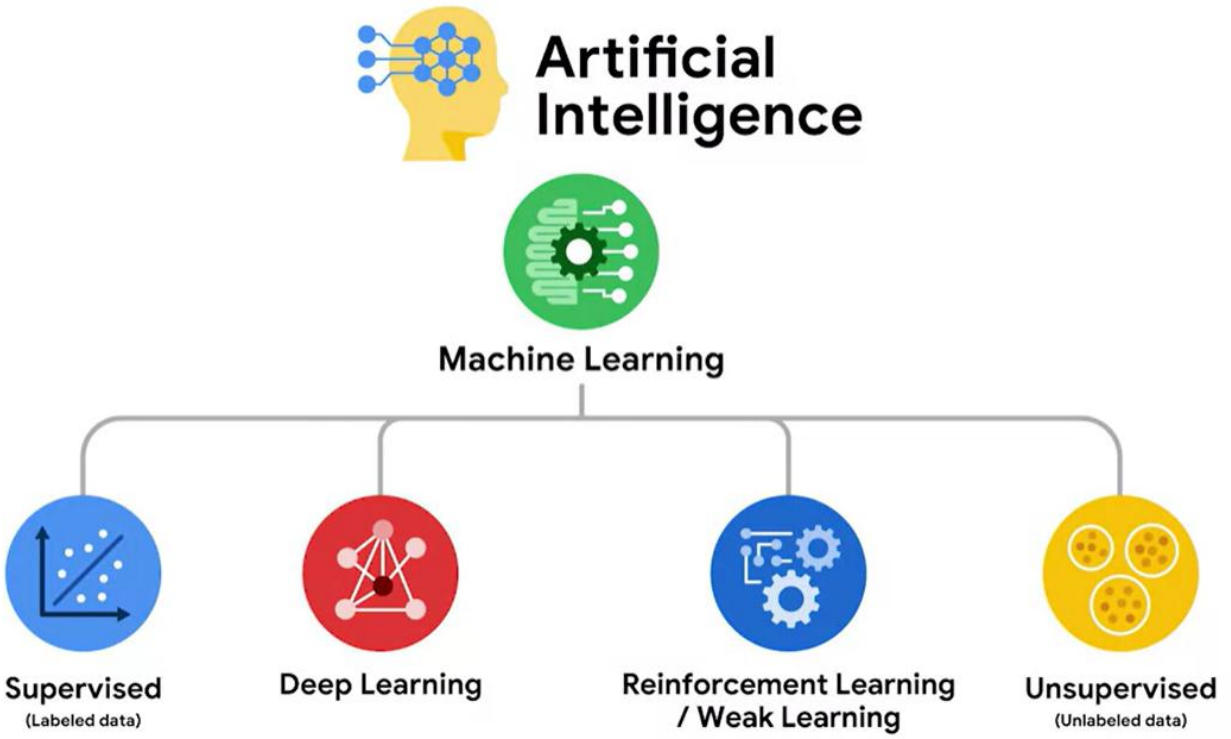
Unsupervised machine learning

Uses algorithms to analyze and cluster unlabeled datasets

Unlabeled data

Height (cm)	Bird
45	
101	
179	
271	
115	
76	
244	
63	
228	

their similarity based
on patterns detected by



Continuous features

Features that can take on an infinite and uncountable set of values

Categorical variables

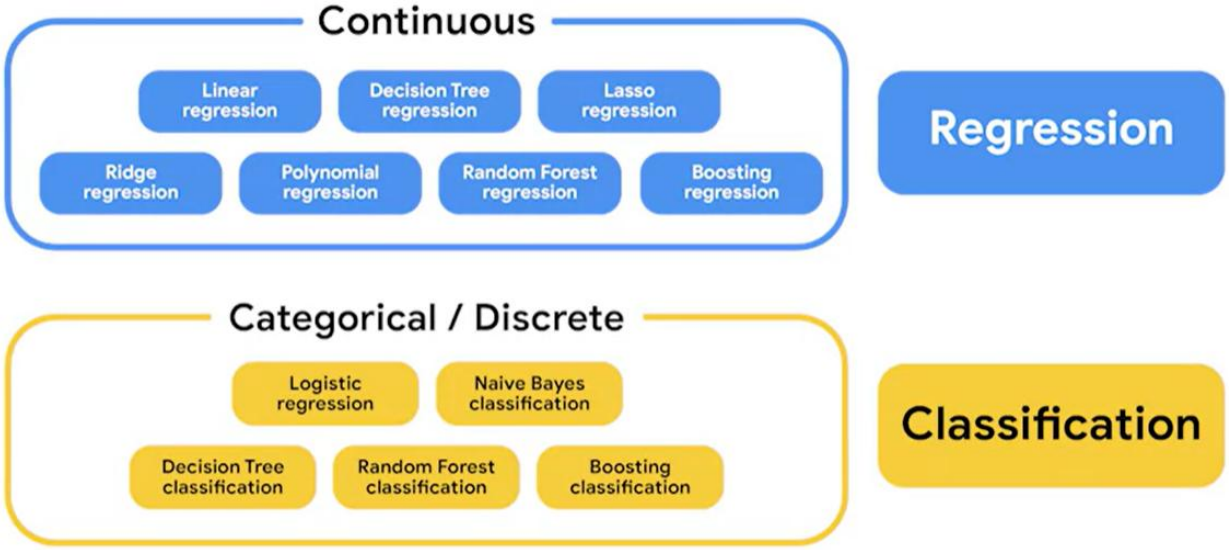
Variables that contain a finite number of groups or categories

Discrete features

Features with a countable number of values between any two values

Continuous variables

Variables that can take on an infinite and uncountable set of values



Recommendation systems

Unsupervised learning techniques that use unlabeled data to offer relevant suggestions to users

Goal: To quantify how similar one thing is to another

Content-based filtering

Comparisons are made based on attributes of content

A.

Song	Beat	Key	BPM	Piano?	Acoustic guitar?
A	rock	F maj	74	yes	no
B	reggaeton	D min	100	no	yes
C	rock	B ♭ maj	72	yes	no

B.

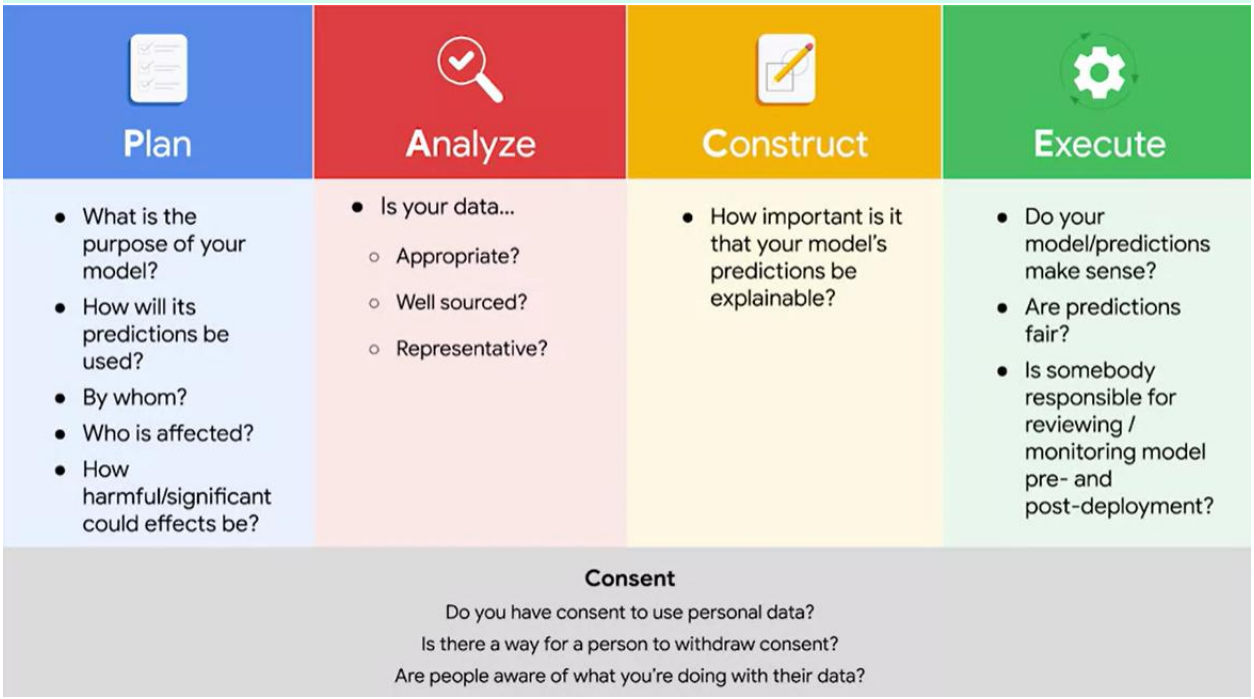
Song	Beat	Key	BPM	Piano?	Acoustic guitar?
A	rock	F maj	74	yes	no
B	reggaeton	D min	100	no	yes
C	rock	B ♭ maj	72	yes	no

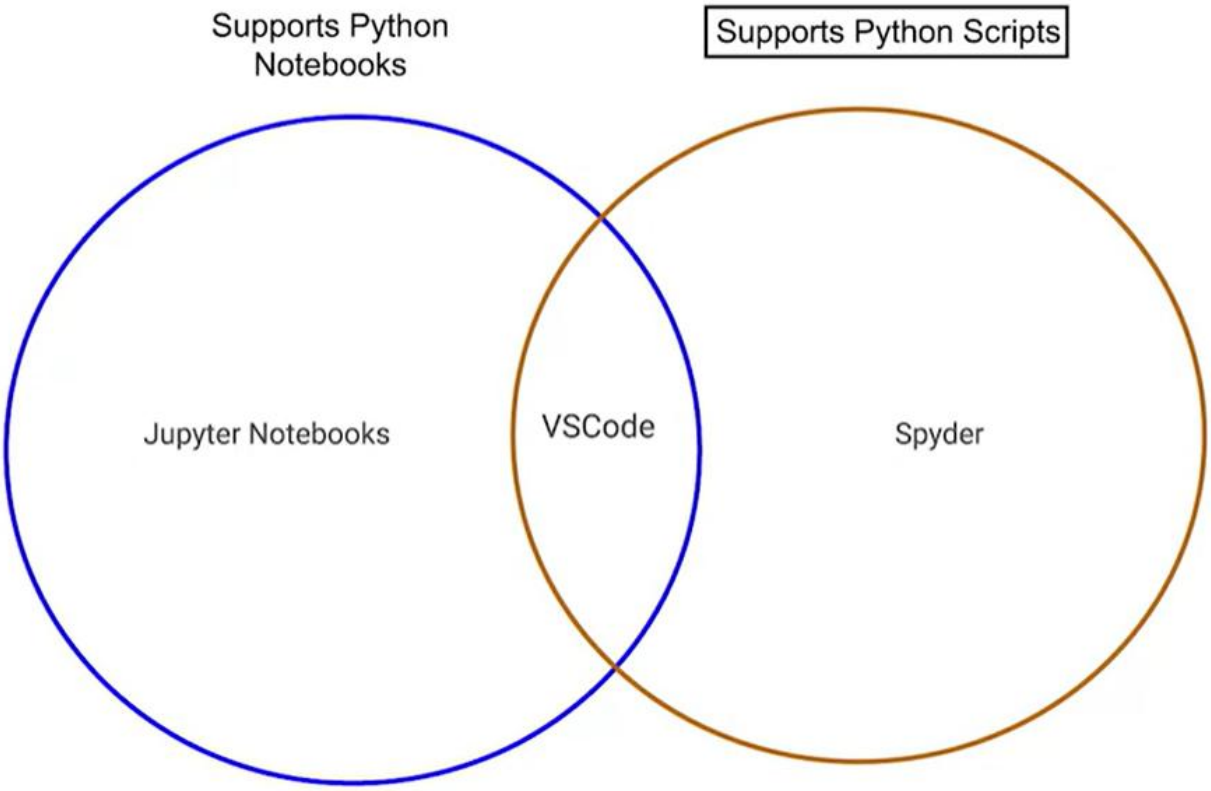
Collaborative filtering drawbacks

- Requires LOTS of data to work
- Requires lots of data from each user
- Data is sparse

Content-based systems are ineffective at making recommendations across content types. This is because different content types rarely share the same features, so user preferences about one content type often cannot be applied to another.

When several users actively like or dislike content by rating it or giving it a review, this enables collaborative filtering. A recommendation system using collaborative filtering makes comparisons based on who else likes a piece of content. Then, it will suggest that same content to others with similar preferences.




Supports Python Notebooks
Supports Python Scripts
Jupyter Notebooks
VSCode
Spyder

Feature engineering

The process of using practical, statistical, and data science knowledge to select, transform, or extract characteristics, properties, and attributes from raw data

Three general categories of feature engineering

- Selection
- Transformation
- Extraction

Feature selection

Select the features in the data that contribute the most to predicting your response variable

Feature transformation

Modifying existing features in a way that improves accuracy when training the model

Feature extraction

Taking multiple features to create a new one that would improve the accuracy of the algorithm

Customer churn

Business term that describes how many and at what rate customers stop using a product or service, or stop doing business with a company altogether

Feature extraction

The process of taking two or more features and using them to create a brand new feature

Naive Bayes

A supervised classification technique that is based on Bayes' Theorem with an assumption of independence among predictors

Posterior probability

The probability of an event occurring after taking into consideration new information

Likelihood of a predictor x given a class c

Class prior probability

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Posterior probability

Predictor prior probability

IMPORTANT METHODS IN PYTHONS

SET

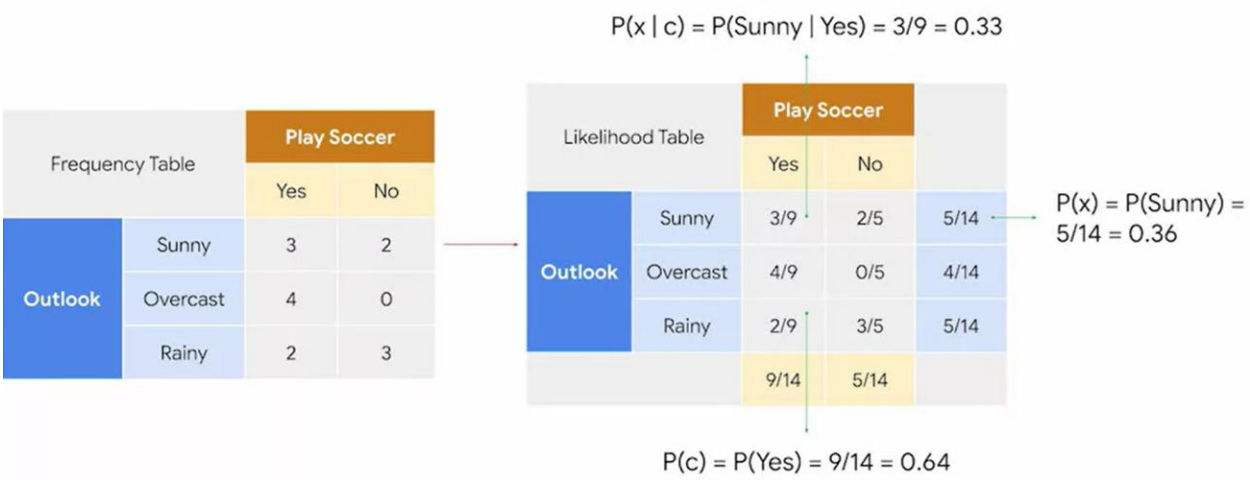
- add()
- clear()
- pop()
- union()
- issuperset()
- issubset()
- intersection()
- difference()
- isdisjoint()
- setdiscard()
- copy()

LIST

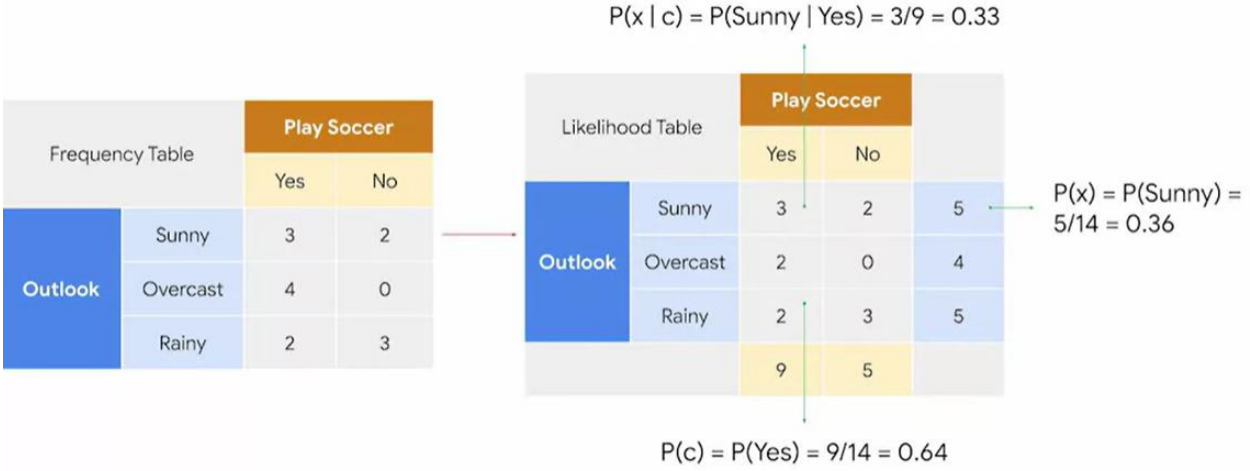
- append()
- copy()
- count()
- insert()
- reverse()
- remove()
- sort()
- pop()
- extend()
- index()
- clear()

DICTIONARY

- copy()
- clear()
- fromkeys()
- items()
- get()
- keys()
- pop()
- values()
- update()
- setdefault()
- popitem()



Posterior Probability: P(c | x) = P(Yes | Sunny) = 0.33 x 0.64 ÷ 0.36 = 0.60



Posterior Probability: P(c | x) = P(**No** | Sunny) = 0.40 x 0.36 ÷ 0.36 = 0.40

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Precision

Proportion of positive predictions
that were correct to all positive predictions

$$Precision = \frac{TP}{TP + FP}$$

Recall

Proportion of actual positives that were
identified correctly to all actual positives

$$Recall = \frac{TP}{TP + FN}$$

F1 score

The harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Accuracy

The number of correct predictions divided by the total number of predictions

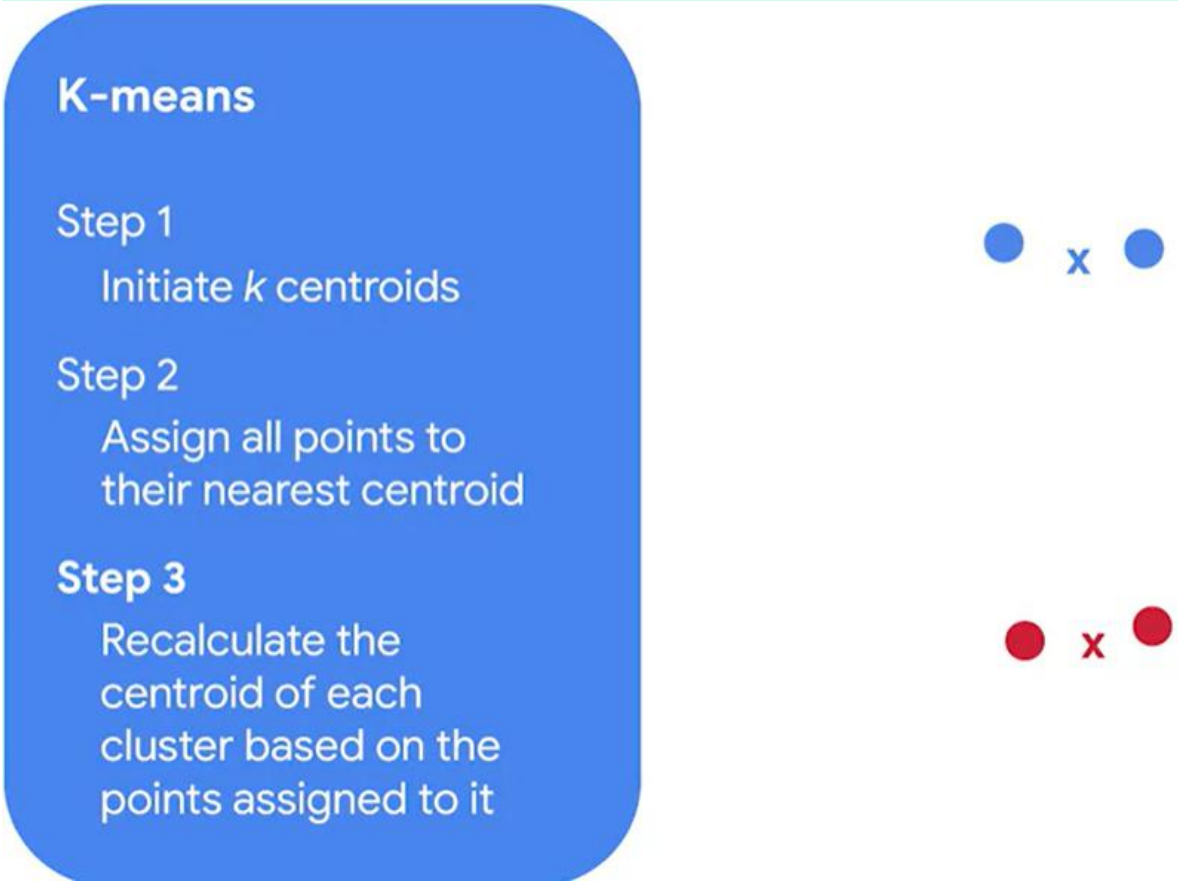
$$Accuracy = \frac{\#\ Correct\ Predictions}{\#\ Total\ Predictions}$$

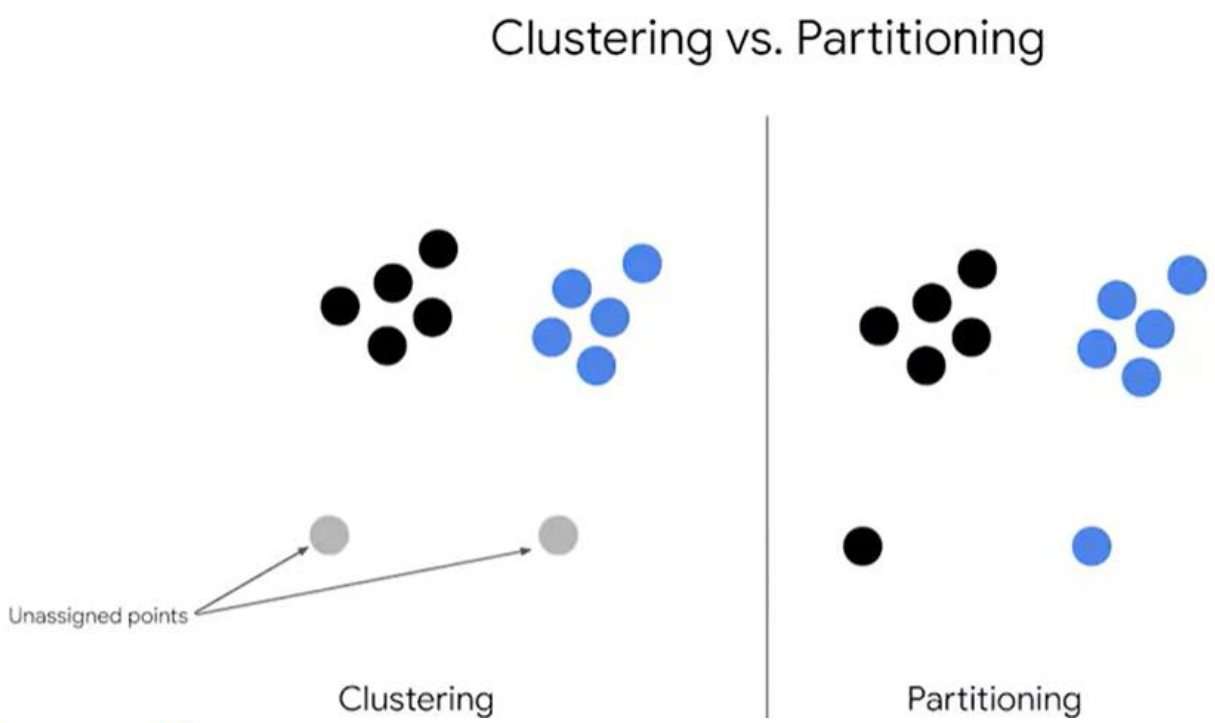
Centroid

The center of a cluster determined by the mathematical mean of all the points in that cluster

Correct

When using k-means, the value for k is a decision that the modeler makes. Sometimes the data professional will have an idea about the number of clusters necessary for a project. Other times, it will be necessary to try different values to determine which one provides the best results.





Inertia

Sum of the squared distances between each observation and its nearest centroid

Correct
For an effective clustering model, the clusters should be clearly identifiable. Within each intracluster, the points are close to each other; within each intercluster, there is lots of empty space.

Correct
For an effective clustering model, the clusters should be clearly identifiable. Within each intracluster, the points are close to each other; within each intercluster, there is lots of empty space.

Correct
For an effective clustering model, the clusters should be clearly identifiable. Within each intracluster, the points are close to each other; within each intercluster, there is lots of empty space.

$$Inertia = \sum_{i=1}^{n} (x_i - c_k)^2$$

Silhouette score

The mean of the silhouette coefficients of all the observations in the model

$$S = \frac{(b-a)}{\max(a, b)}$$

The labels attribute will enable them to get the cluster assignments. It returns a list of values that is the same length as the training data. Each value corresponds to the number of the cluster to which that point is assigned.

Tree-based learning

A type of supervised machine learning that performs classification and regression tasks

Decision tree

Flow-chart-like supervised classification model, and a representation of various solutions that are available to solve a given problem, based on the possible outcomes of related choices

Correct

In a decision tree, nodes are where decisions are made, and they are connected by edges. At each node, a single feature of the data is considered and decided on. Edges direct from one node to the next during this process. Eventually, all relevant features will have been resolved, resulting in the classification prediction.

Root node

The first node of the tree, where the first decision is made

Decision node

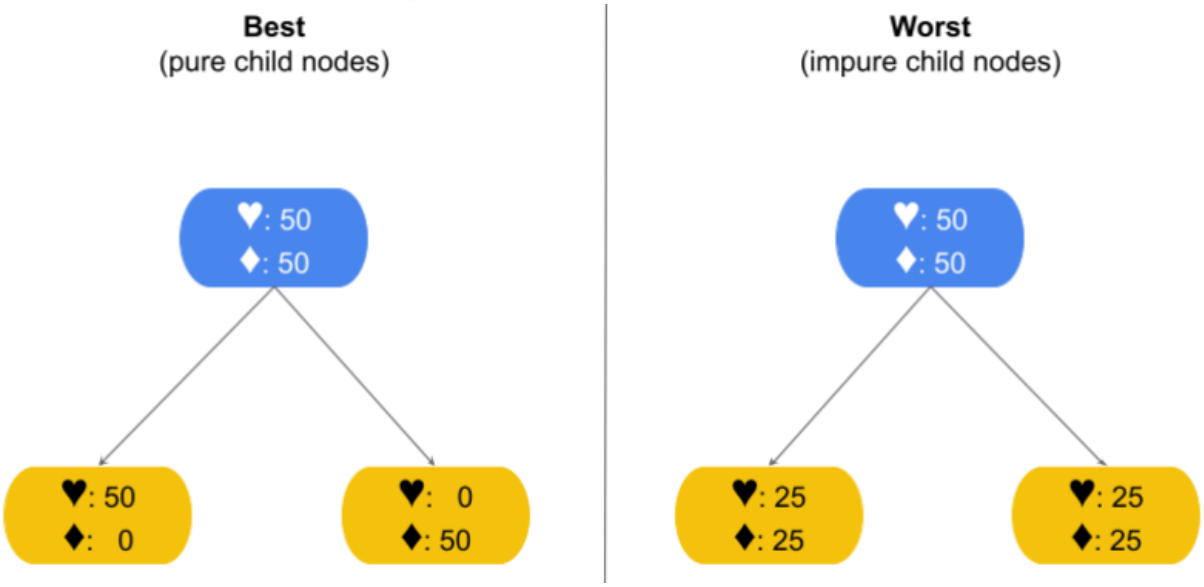
Nodes of the tree where decisions are made

Leaf node

The nodes where a final prediction is made

Child node

A node that is pointed to from another node



$$Gini\ impurity = 1 - \sum_{i=1}^{N} P(i)^2$$

In the case of the fruits example, this becomes:

$$Gini\ impurity = 1 - P(apple)^2 - P(grape)^2$$

$$= 1 - (\frac{number\ of\ apples\ in\ node}{total\ number\ of\ samples\ in\ node})^2 - (\frac{number\ of\ grapes\ in\ node}{total\ number\ of\ samples\ in\ node})^2$$

For the "**red=yes**" child node:

Gini impurity $= 1 - (1/3)^2 - (2/3)^2$

$= 1 - 0.111 - 0.444$

$= 0.445$

And for the "**red=no**" child node:

Gini impurity $= 1 - (3/4)^2 - (1/4)^2$

$= 1 - 0.5625 - 0.0625$

$= 0.375$

Total Gi:

$$= (\frac{number\ of\ samples\ in\ LEFT\ child}{number\ of\ samples\ in\ BOTH\ child\ nodes}) * Gi_{left\ child} + (\frac{number\ of\ samples\ in\ RIGHT\ child}{number\ of\ samples\ in\ BOTH\ child\ nodes}) * Gi_{right\ child}$$

$= (3/7 * 0.445) + (4/7 * 0.375)$

$Gi_{total} = 0.405$

Repeat this process for every split option

This same process is repeated for every split option. The fruit example has nine options (A–I):

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

Max depth

Defines how "long" a decision tree can get

Min samples leaf

Defines the minimum number of samples for a leaf node

GridSearch

A tool to confirm that a model achieves its intended purpose by systematically checking every combination of hyperparameters to identify which set produces the best results, based on the selected metric

Model validation

The set of processes and activities intended to verify that models are performing as expected

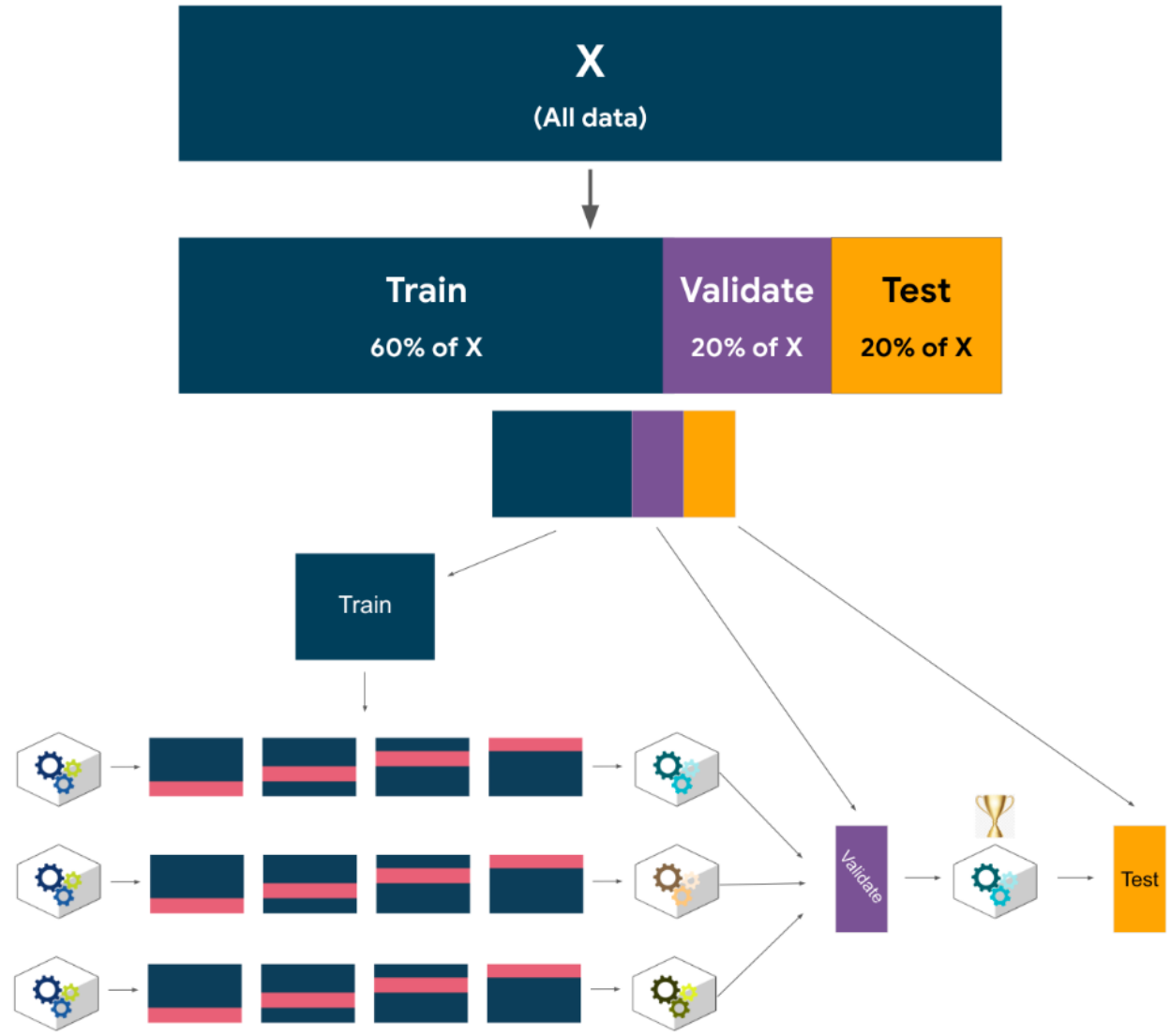
Cross-validation

A process that uses different portions of the data to test and train a model on different iterations

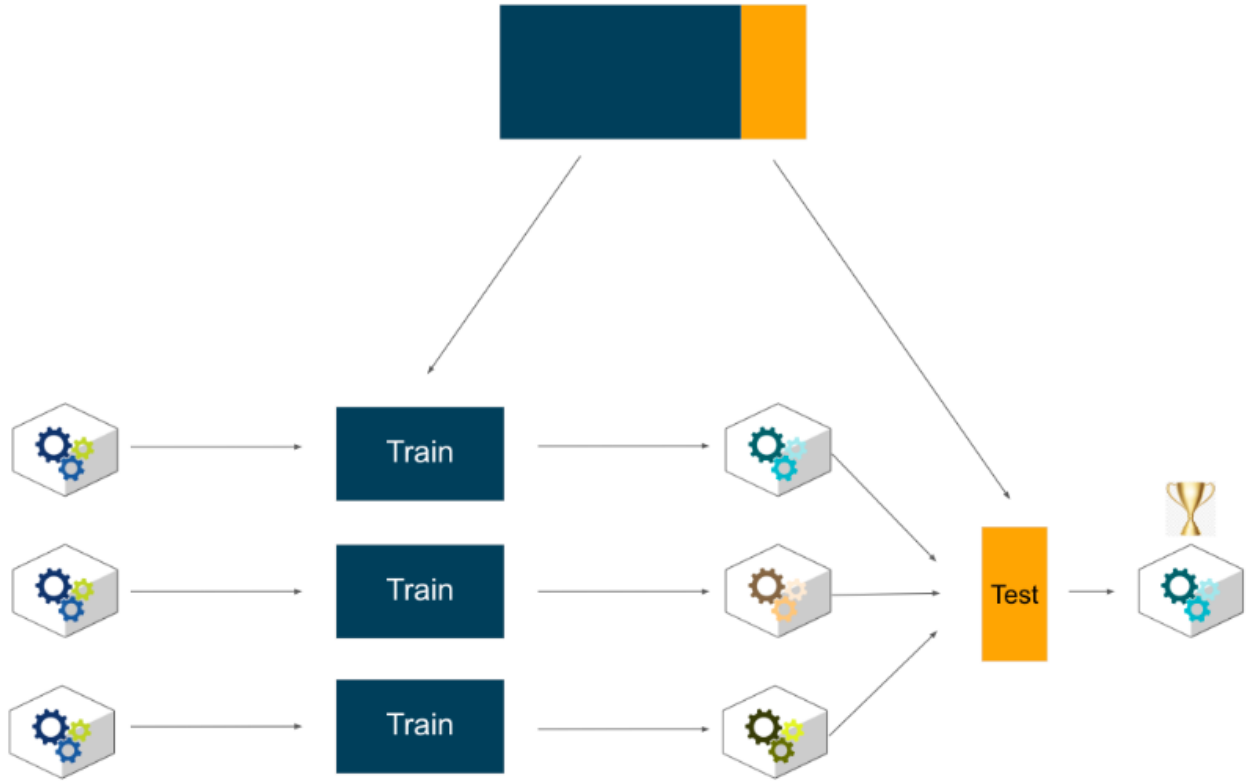
Validation sets

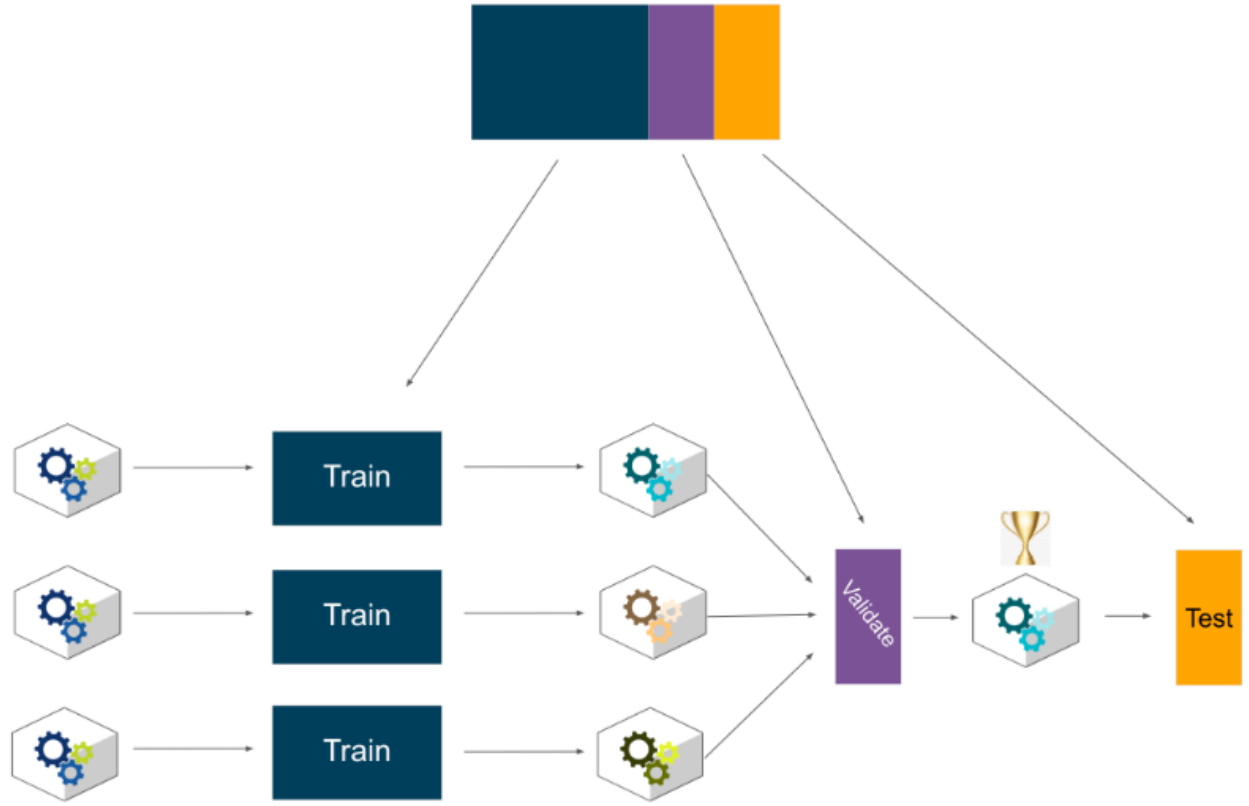
The simplest way to maintain the objectivity of the test data is to create another partition in the data—a validation set—and save the test data for after you select the final model. The validation set is then used, instead of the test set, to compare different models.

Here is one common way of splitting data, but note that these proportions are not required. You can split to whichever ratios make the most sense for your use case.



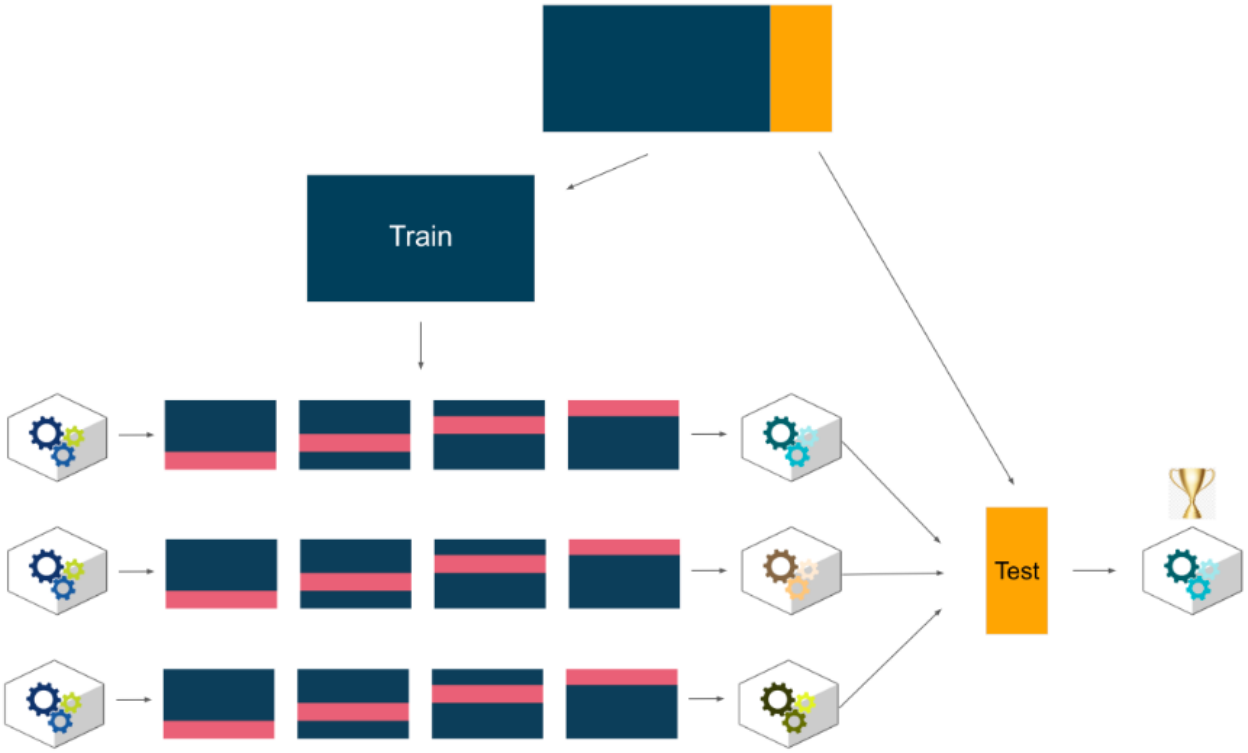
A.





B. C.

D.



Hyperparameter tuning

Changing parameters that directly affect how the model trains, before the learning process begins

Min samples leaf

Defines the minimum number of samples for a leaf node

Max depth

Defines how “long” a decision tree can get

GridSearch

A tool to confirm that a model achieves its intended purpose by systematically checking every combination of hyperparameters to identify which set produces the best results, based on the selected metric

Ensemble learning (or “ensembling”)

Aggregating their outputs to make a prediction

Base learner

Each individual model that comprises an ensemble

Weak learner

A model that performs slightly better than randomly guessing

Bagging

Bootstrap + aggragating

Random forest

Ensemble of decision trees trained on bootstrapped data with randomly selected features

min_samples_leaf

Defines the minimum number of samples for a leaf node

min_samples_leaf

A split can only occur if it guarantees a minimum number of observations in the resulting nodes

min_samples_split

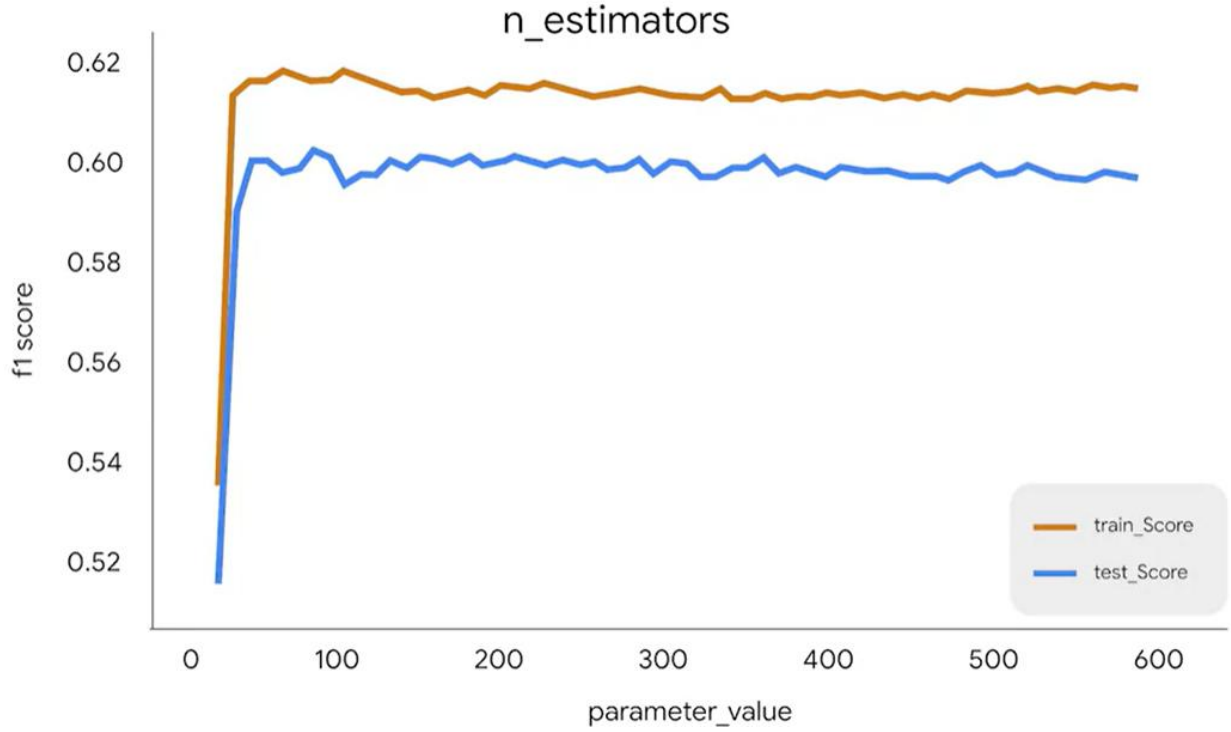
Controls threshold below which nodes become leaves

max_features

Specifies the number of features that each tree randomly selects during training

n_estimators

Specifies the number of trees your model will build in its ensemble



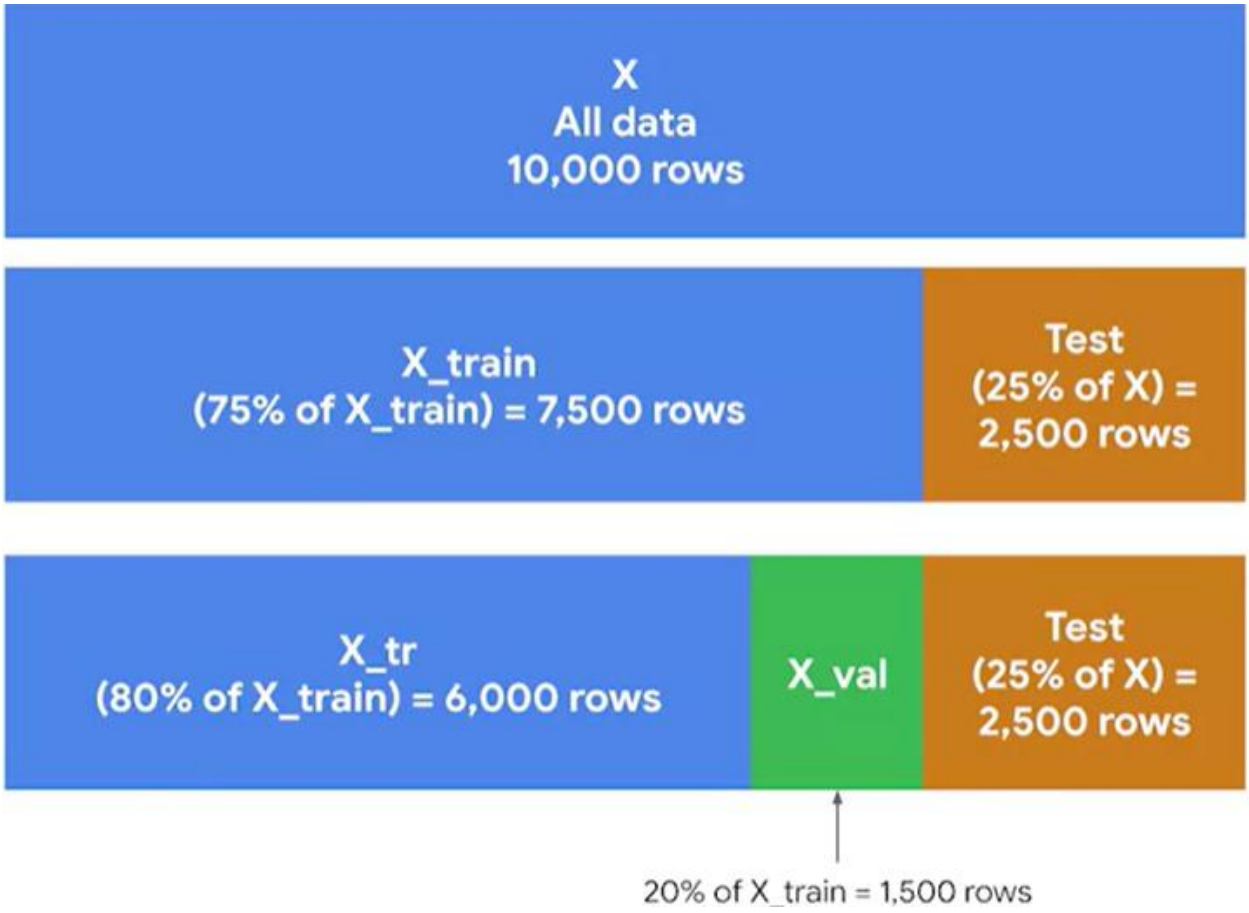
```
%%time
```

A magic command that gives you the runtime of the cell it's entered in

Magic commands ("magics")

Commands that are built into Python to simplify common tasks

Magic commands always begin with either '%' or '%%'



En resumen:

- `axis=0` : Elimina filas con valores faltantes.
- `axis=1` : Elimina columnas con valores faltantes.

In a random forest, bootstrapped data is used to train the ensemble of decision-tree base learners. Bootstrapping refers to sampling with replacement. So, a random forest model will grow each of its trees by taking a random subset of the available features in the training data, then splitting each node at the best feature available to that tree.

Boosting

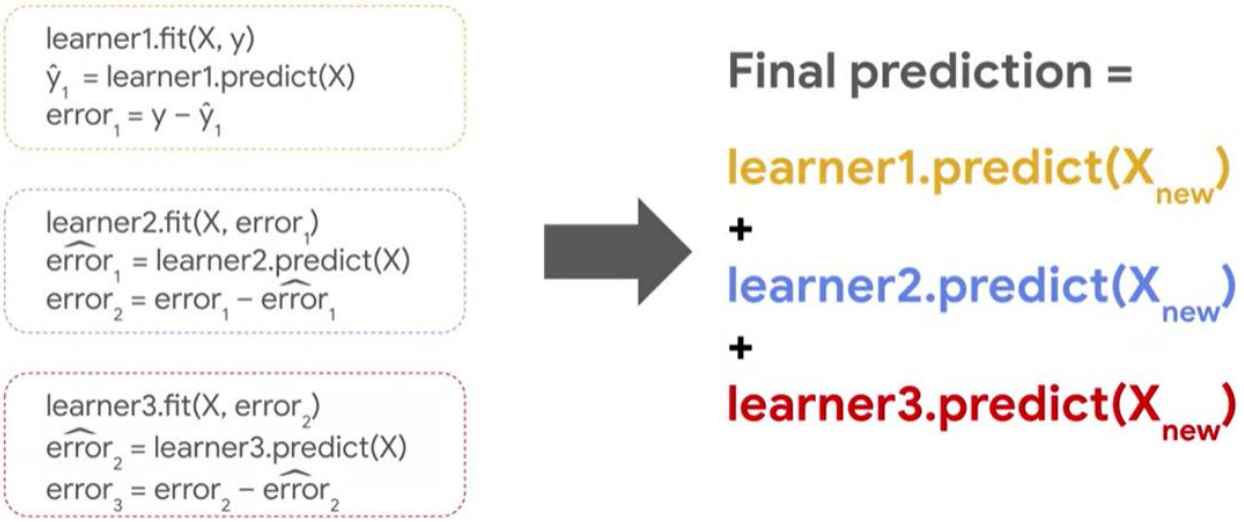
Technique that builds an ensemble of weak learners sequentially, with each consecutive learner trying to correct the errors of the one that preceded it

Adaptive boosting (AdaBoost)

A boosting methodology where each consecutive base learner assigns greater weight to the observations incorrectly predicted by the preceding learner

Gradient boosting

A boosting methodology where each base learner in the sequence is built to predict the residual errors of the model that preceded it



Gradient boosting machines (GBMs)

Model ensembles that use gradient boosting

Black-box model

Any model whose predictions cannot be precisely explained

Extrapolation

A model's ability to predict new values that fall outside of the range of values in the training data

Disadvantages of gradient boosting machines (GBMs)

- Tuning many hyperparameters can be time-consuming
- Difficult to interpret
- Have difficulty with extrapolation
- Prone to overfitting if too many hyperparameters are tuned

XGBoost

Extreme gradient boosting, an optimized GBM package

`max_depth`

Controls how deep each base learner tree will grow

- Typical values: 2-10

n_estimators

Maximum number of base learners your ensemble will grow

- Typical values: 50-500

learning_rate (shrinkage)

How much weight is given to each consecutive tree's prediction in the final ensemble

- Typical values: 0.01-0.3

min_child_weight

A tree will not split a node if it results in any child node with less weight than this value

- (0–1): interpreted as a percentage (e.g., 0.1 = 10% of training data)
- 1+: interpreted as a sample weight (≅number of observations)