Department of Artificial Intelligence and Data Science

Fundamentals of Data Science (AD45)

Academic Year : 2024-24, Batch 2023

Credits: 3:0:0

Text Books:

Joel Grus, "Data Science from Scratch", 2nd Edition, O'Reilly Publications/Shroff Publishers and Distributors Pvt. Ltd., 2019. ISBN-13: 978- 9352138326

UNIT 1

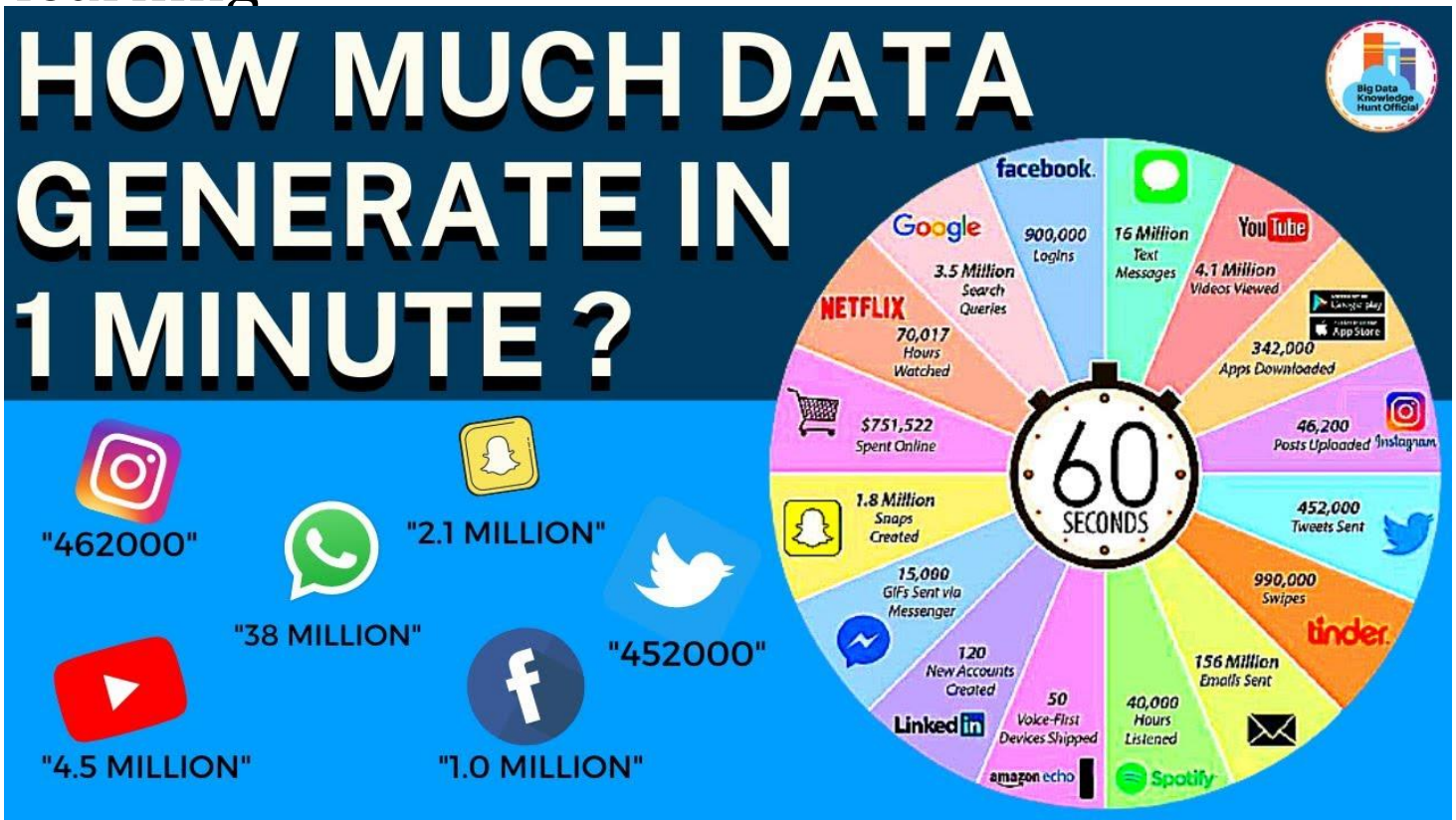
What is Data Science? Types of Data & Data Sources, Visualizing Data, matplotlib, Bar Charts, Line Charts, Scatterplots, Linear Algebra, Vectors, Matrices, Statistics, Describing a Single Set of Data, Correlation, Simpson's Paradox, Some Other Correlational Caveats, Correlation and Causation. Probability, Dependence and Independence, Random Variables, Continuous Distributions, The Normal Distribution.

The Ascendance of Data

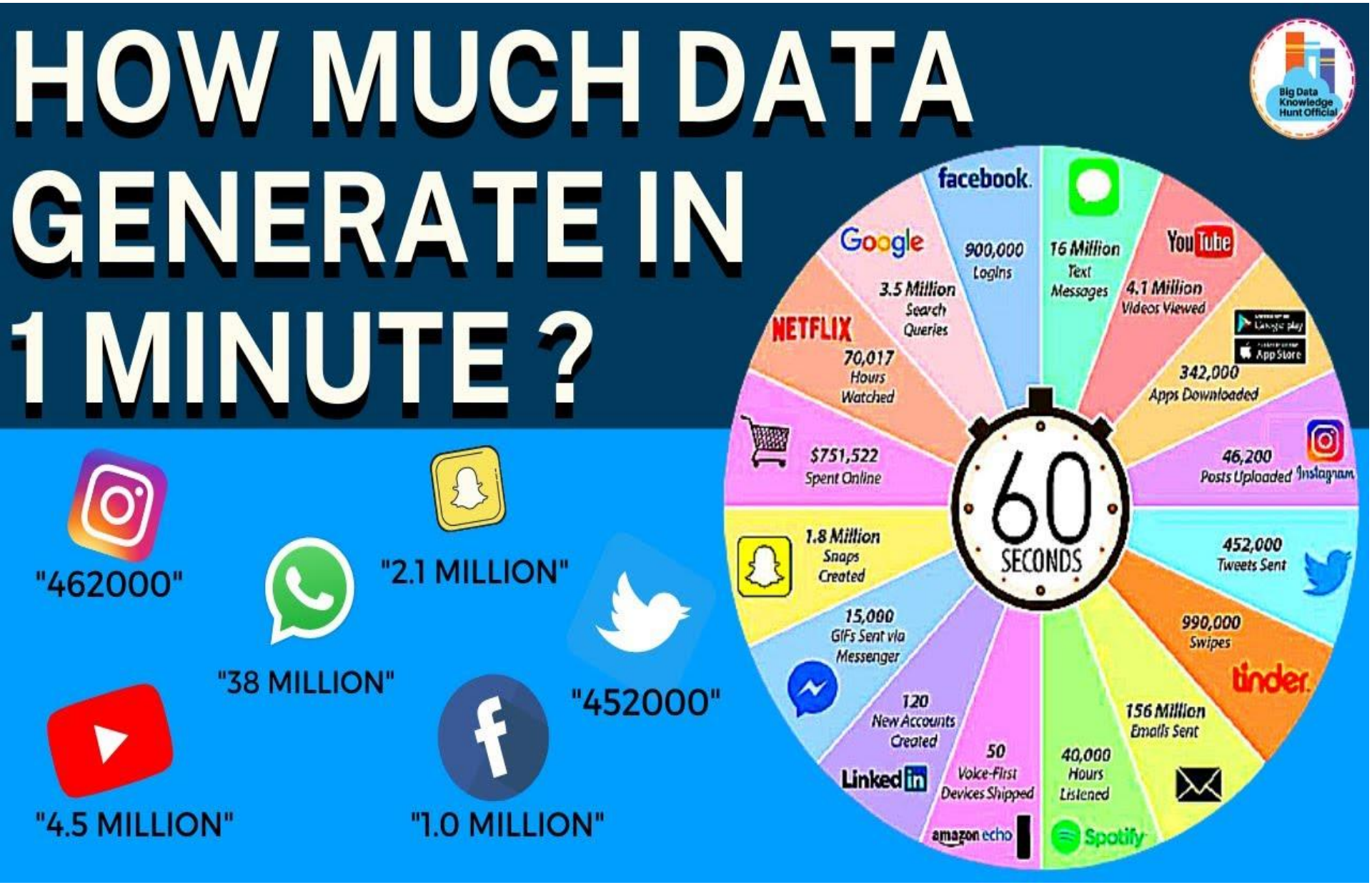
- We live in a world that's drowning in data. Websites track every user's every click.
- Your **smartphone** is building up a record of your location and speed every second of every day. "Quantified selfers" wear pedometers-on-steroids that are always recording their heart rates, movement habits, diet, and sleep patterns.
- **Smart cars** collect driving habits, smart homes collect living habits, and smart marketers collect purchasing habits.
- The **internet** itself represents a huge graph of knowledge that contains (among other things) an enormous cross-referenced encyclopedia; domain-specific databases about movies, music, sports results, pinball machines, memes, and cocktails; and too many government statistics (some of them nearly true!) from too many governments to wrap your head around.

What Is Data Science?

- **What is Data?**
- **Data** is a collection of facts, numbers, measurements, observations. It can be raw or processed and is used to gain insights, make decisions, and drive technologies like AI and machine learning.

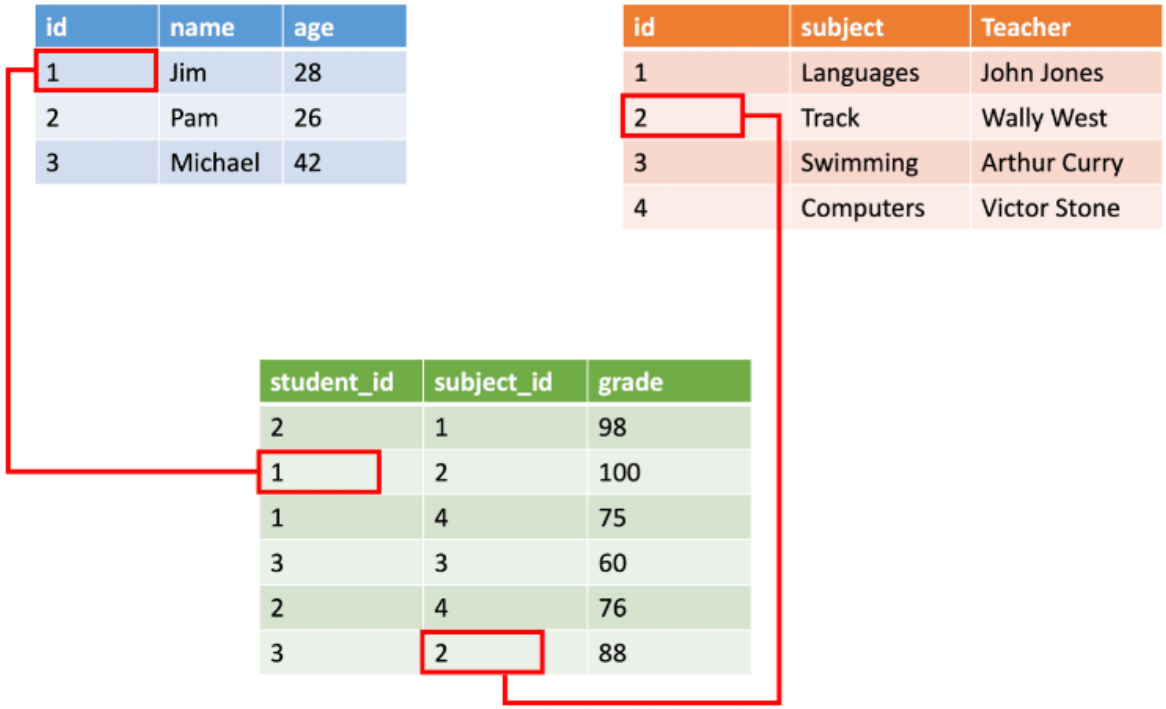


What Is Data Science?



What Is Data Science?: Types of Data

- **Structured Data**
  - Organized and stored in a fixed format (e.g., databases, spreadsheets).
  - Examples:
    - Customer records (Name, Age, Email)
    - Sales data (Price, Quantity, Date)
    - SQL Databases



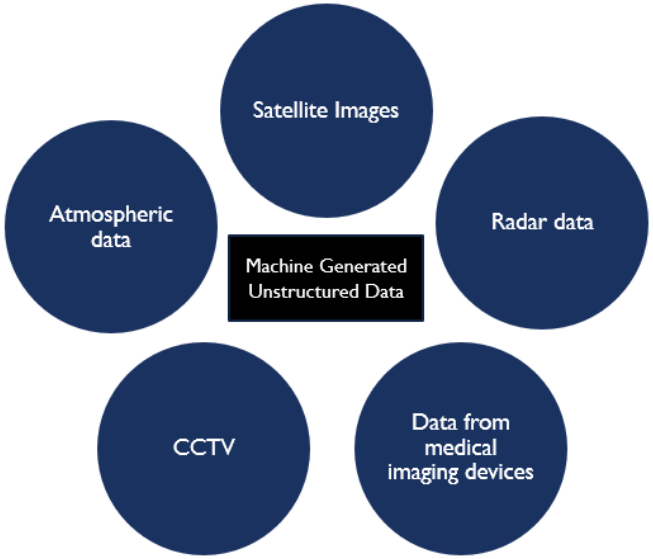
id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

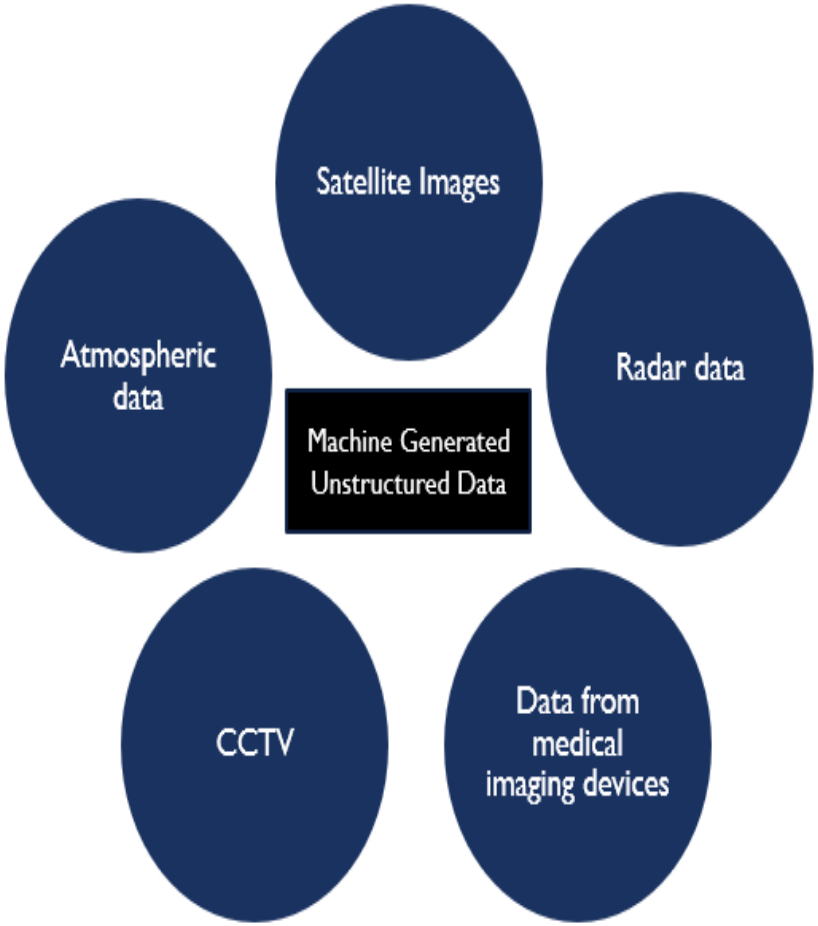
student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

What Is Data Science?: Types of Data

- **Unstructured Data**
- No predefined format, making it harder to process.
- Examples:
  - Text documents, emails, social media posts
  - Images, videos, audio files
  - Sensor data, logs



What Is Data Science?: Types of Data



What Is Data Science?: Types of Data

- **Semi-Structured Data**
- Has some organization but doesn't fit traditional databases.
- Examples:
  - JSON, XML files
  - Emails with metadata (subject, timestamp)

Semi-structured data

```
<University>
 <Student ID="1">
  <Name>John</Name>
  <Age>18</Age>
  <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
  <Name>David</Name>
  <Age>31</Age>
  <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

```
{
  "name": "Jane Smith",
  "email": "jane.smith@example.com",
  "preferences": {
    "newsletter": true,
    "smsNotifications": false
  }
}
```

What Is Data Science?: Sources of Data

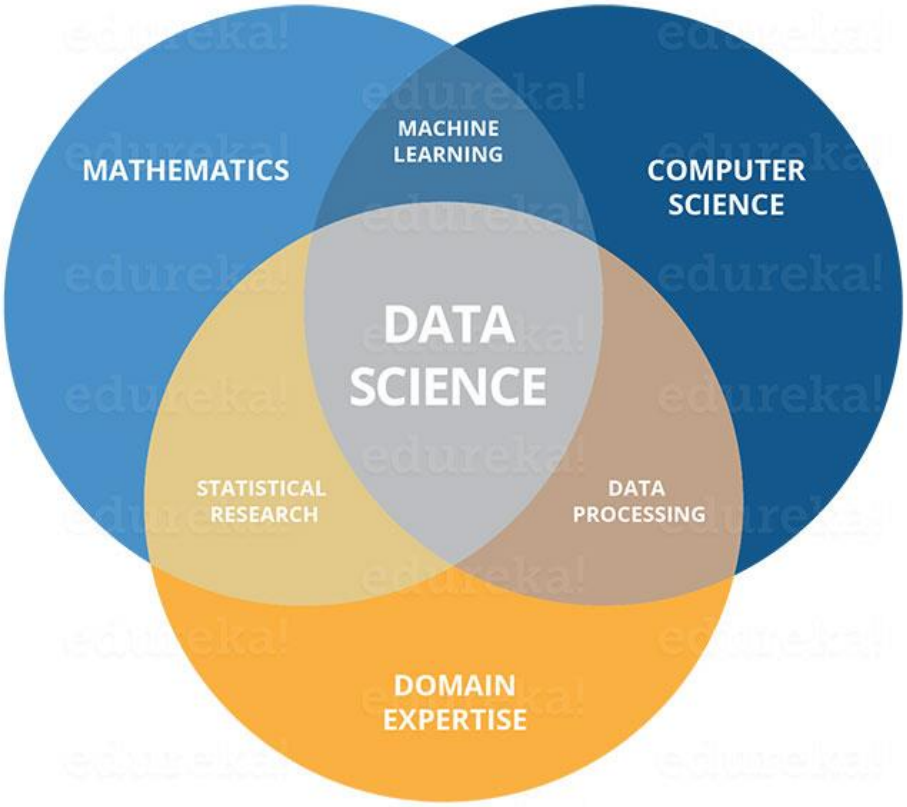
- **Digital Transactions** – Sales records, banking transactions.
- **Social Media** – Posts, likes, shares, and comments.
- **Sensors & IoT Devices** – Temperature sensors, smartwatches.
- **Healthcare Records** – Patient data, medical images.
- **Government & Research** – Census data, scientific experiments.

Why is Data Important?

- **Drives decision-making** – Businesses use data to improve operations.
- **Enables AI & Machine Learning** – Models learn from large datasets.
- **Improves efficiency** – Automates tasks and predicts trends.
- **Personalizes experiences** – Recommender systems (Netflix, Amazon).

What Is Data Science?

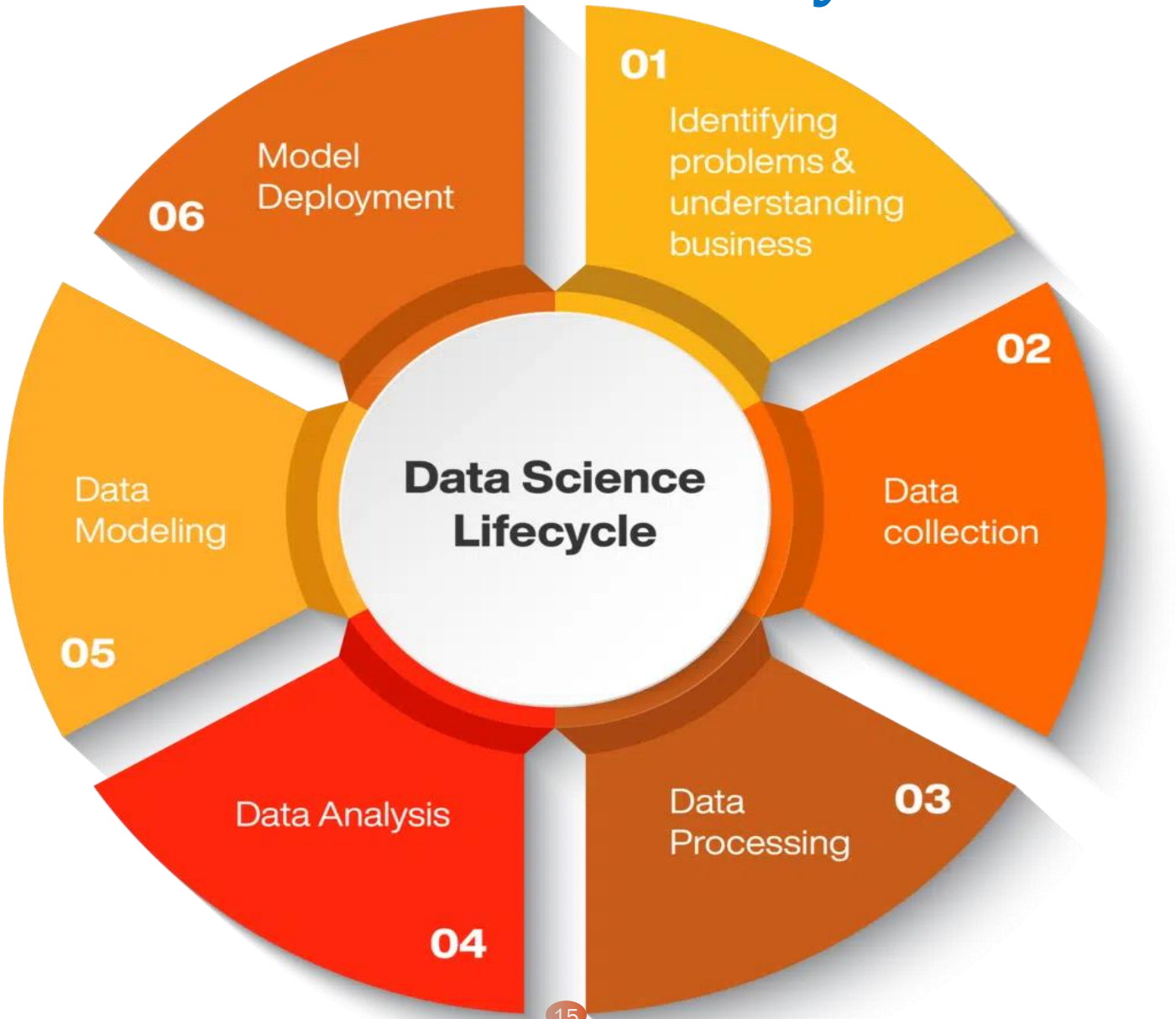
- Data science is the **study** of data to **derive** meaningful insights for businesses. It's a multidisciplinary field that combines **statistics, math, computer engineering, and artificial intelligence** (AI).



What Is Data Science Lifecycle?

- This lifecycle encompasses **Six key stages**, that include—Identify problem, data collection, data pre-processing, data analysis, data modeling, and model deployment and dissemination.
- All of these are crucial in the process of converting raw data into valuable knowledge to support projects and their execution.

What Is Data Science Lifecycle?



What are lambda functions in Python?

- In Python, an anonymous function is a function that is defined without a name.
- While normal functions are defined using the ***def*** keyword in Python, anonymous functions are defined using the ***lambda*** keyword.
- Hence, anonymous functions are also called ***lambda*** functions
- ***How to use lambda Functions in Python?***
- A lambda function in python has the following syntax.

lambda arguments: expression

- Lambda functions can have any number of arguments but only one expression. The expression is evaluated and returned. Lambda functions can be used wherever function objects are required.

Lists

- The most fundamental data structure in Python is the list. A list is simply an ordered collection. It is called as array in other programming languages.

✓ `integer_list = [1, 2, 3]`

✓ `heterogeneous_list = ["string", 0.1, True]`

✓ `list_of_lists = [ integer_list, heterogeneous_list, [] ]`

✓ `list_length = len(integer_list) # equals 3`

✓ `list_sum = sum(integer_list) # equals 6`

Lists

- If you don't want to modify x, you can use list addition:

```
x = [1, 2, 3]
y = x + [4, 5, 6]        # y is [1, 2, 3, 4, 5, 6]; x is unchanged
```

- More frequently we will append to lists one item at a time:

```
x = [1, 2, 3]
x.append(0)        # x is now [1, 2, 3, 0]
y = x[-1]          # equals 0
z = len(x)         # equals 4
```

- It's often convenient to unpack lists when you know how many elements they contain:

```
x, y = [1, 2]     # now x is 1, y is 2
```

although you will get a *ValueError* if you don't have the same number of elements on both sides.

- A common idiom is to use an underscore for a value you're going to throw away:

```
_, y = [1, 2]     # now y == 2, didn't care about the first element
```

Tuples

- Tuples are lists' immutable cousins.
- Pretty much anything you can do to a list that doesn't involve modifying it, you can do to a tuple.
- You specify a tuple by using parentheses (or nothing) instead of square brackets:

```
my_list = [1, 2]
my_tuple = (1, 2)
other_tuple = 3, 4
my_list[1] = 3      # my_list is now [1, 3]
```

```
try:
    my_tuple[1] = 3
except TypeError:
    print("cannot modify a tuple")
```

- Tuples are a convenient way to return multiple values from functions:

```
def sum_and_product(x, y):
    return (x + y), (x * y)

sp = sum_and_product(2, 3)      # sp is (5, 6)
s, p = sum_and_product(5, 10)   # s is 15, p is 50
```

- Tuples (and lists) can also be used for multiple assignment:

```
x, y = 1, 2     # now x is 1, y is 2
x, y = y, x     # Pythonic way to swap variables; now x is 2, y is 1
```

Dictionaries

- Another fundamental data structure is a dictionary, which associates values with keys and allows you to quickly retrieve the value corresponding to a given key:

```
empty_dict = {}                          # Pythonic
empty_dict2 = dict()                     # less Pythonic
grades = {"Joel": 80, "Tim": 95}         # dictionary literal
```

- You can look up the value for a key using square brackets:

```
joels_grade = grades["Joel"]             # equals 80
```

- But you'll get a KeyError if you ask for a key that's not in the dictionary:

```
try:
    kates_grade = grades["Kate"]
except KeyError:
    print("no grade for Kate!")
```

Dictionaries

- You can check for the existence of a key using ***in.*** This membership check is fast even for large dictionaries.

```
joel_has_grade = "Joel" in grades      # True
kate_has_grade = "Kate" in grades      # False
```

- Dictionaries have a get method that returns a default value (instead of raising an exception) when you look up a key that's not in the dictionary:

```
joels_grade = grades.get("Joel", 0)    # equals 80
kates_grade = grades.get("Kate", 0)    # equals 0
no_ones_grade = grades.get("No One")   # default is None
```

- You can assign key/value pairs using the same square brackets:

```
grades["Tim"] = 99                     # replaces the old value
grades["Kate"] = 100                   # adds a third entry
num_students = len(grades)             # equals 3
```

Dictionaries

- you can use dictionaries to represent structured data:

```
tweet = {
    "user" : "joelgrus",
    "text" : "Data Science is Awesome",
    "retweet_count" : 100,
    "hashtags" : ["#data", "#science", "#datascience", "#awesome",
"#yolo"]
}
```

- Besides looking for specific keys, we can look at all of them:

```
tweet_keys   = tweet.keys()     # iterable for the keys
tweet_values = tweet.values()   # iterable for the values
tweet_items  = tweet.items()    # iterable for the (key, value) tuples

"user" in tweet_keys            # True, but not Pythonic

"user" in tweet                  # Pythonic way of checking for keys
"joelgrus" in tweet_values       # True (slow but the only way to check)
```

- Dictionary keys must be “hashable”; in particular, you cannot use lists as keys.
- If you need a multipart key, you should probably use a tuple or figure out a way to turn the key into a string.

defaultdict

- Imagine that you're trying to count the words in a document.
- An obvious approach is to create a dictionary in which the keys are words and the values are counts.
- As you check each word, you can increment its count if it's already in the dictionary and add it to the dictionary if it's not:

```
word_counts = {}
for word in document:
    if word in word_counts:
        word_counts[word] += 1
    else:
        word_counts[word] = 1
```

- You could also use the "forgiveness is better than permission" approach and just handle the exception from trying to look up a missing key:

```
word_counts = {}
for word in document:
    try:
        word_counts[word] += 1
    except KeyError:
        word_counts[word] = 1
```

- A third approach is to use get, which behaves gracefully for missing keys:

```
word_counts = {}
for word in document:
    previous_count = word_counts.get(word, 0)
    word_counts[word] = previous_count + 1
```

defaultdict

- Every one of these is slightly unwieldy, which is why defaultdict is useful.
- A ***defaultdict*** is like a regular dictionary, except that when you try to look up a key it doesn't contain, it first adds a value for it using a zero-argument function you provided when you created it.
- In order to use defaultdicts, you have to import them from collections:

```
from collections import defaultdict

word_counts = defaultdict(int)          # int() produces 0
for word in document:
    word_counts[word] += 1
```

- They can also be useful with list or dict, or even your own functions:

```
dd_list = defaultdict(list)             # list() produces an empty list
dd_list[2].append(1)                    # now dd_list contains {2: [1]}

dd_dict = defaultdict(dict)             # dict() produces an empty dict
dd_dict["Joel"]["City"] = "Seattle"     # {"Joel" : {"City": Seattle"}}

dd_pair = defaultdict(lambda: [0, 0])
dd_pair[2][1] = 1                       # now dd_pair contains {2: [0,
1]}
```

- These will be useful when we're using dictionaries to "collect" results by some key and don't want to have to check every time to see if the key exists yet.

Counters

- A Counter turns a sequence of values into a defaultdict(int)-like object mapping keys to counts:

```
from collections import Counter
c = Counter([0, 1, 2, 0])          # c is (basically) {0: 2, 1: 1, 2: 1}
```

- This gives us a very simple way to solve our word_counts problem:

```
# recall, document is a list of words
word_counts = Counter(document)
```

- A Counter instance has a ***most_common*** method that is frequently useful:

```
# print the 10 most common words and their counts
for word, count in word_counts.most_common(10):
    print(word, count)
```

Sets

- Another useful data structure is set, which represents a collection of distinct elements.
- You can define a set by listing its elements between curly braces:

```
primes_below_10 = {2, 3, 5, 7}
```

- However, that doesn't work for empty sets, as {} already means "empty dict." In that case you'll need to use set() itself:

```
s = set()
s.add(1)       # s is now {1}
s.add(2)       # s is now {1, 2}
s.add(2)       # s is still {1, 2}
x = len(s)     # equals 2
y = 2 in s     # equals True
z = 3 in s     # equals False
```

Sets

- We'll use sets for two main reasons. The first is that in is a very fast operation on sets.
- If we have a large collection of items that we want to use for a membership test, a set is more appropriate than a list:

```
stopwords_list = ["a", "an", "at"] + hundreds_of_other_words + ["yet",
"you"]

"zip" in stopwords_list     # False, but have to check every element

stopwords_set = set(stopwords_list)
"zip" in stopwords_set      # very fast to check
```

- The second reason is to find the distinct items in a collection:

```
item_list = [1, 2, 3, 1, 2, 3]
num_items = len(item_list)                  # 6
item_set = set(item_list)                   # {1, 2, 3}
num_distinct_items = len(item_set)          # 3
distinct_item_list = list(item_set)         # [1, 2, 3]
```

Control Flow

- As in most programming languages, you can perform an action conditionally using if:

```
if 1 > 2:
    message = "if only 1 were greater than two..."
elif 1 > 3:
    message = "elif stands for 'else if'"
else:
    message = "when all else fails use else (if you want to)"
```

- You can also write a ternary if-then-else on one line, which we will do occasionally:

```
parity = "even" if x % 2 == 0 else "odd"
```

- Python has a while loop:

```
x = 0
while x < 10:
    print(f"{x} is less than 10")
    x += 1
```

```
1  x = 3
2  if x % 2 == 0:
3      parity = "even"
4  else:
5      parity = "odd"
6  print(parity)
7
```

✓ 0.4s

odd

```
1  x = 3
2  parity = "even" if x % 2 == 0 else "odd"
3  print(parity)
4
```

✓ 0.4s

odd

Control Flow

- for and in:

```
# range(10) is the numbers 0, 1, ..., 9
for x in range(10):
    print(f"{x} is less than 10")
```

- If you need more complex logic, you can use continue and break:

```
for x in range(10):
    if x == 3:
        continue  # go immediately to the next iteration
    if x == 5:
        break     # quit the loop entirely
    print(x)
```

- This will print 0, 1, 2, and 4.

Truthiness

- Booleans in Python work as in most other languages, except that they're capitalized:

```
one_is_less_than_two = 1 < 2          # equals True
true_equals_false = True == False     # equals False
```

- Python uses the value None to indicate a nonexistent value. It is similar to other languages' null:

```
x = None
assert x == None, "this is the not the Pythonic way to check for None"
assert x is None, "this is the Pythonic way to check for None"
```

- Python lets you use any value where it expects a Boolean. The following are all "falsy":
  - False
  - None
  - [] (an empty list)
  - {} (an empty dict)
  - ""
  - set()
  - 0
  - 0.0

Truthiness

- Pretty much anything else gets treated as True. This allows you to easily use if statements to test for empty lists, empty strings, empty dictionaries, and so on.
- It also sometimes causes tricky bugs if you're not expecting this behavior:

```
s = some_function_that_returns_a_string()
if s:
    first_char = s[0]
else:
    first_char = ""
```

```
1  s = "abc"
2  first_char = s and s[0]
3  print(first_char)
```
✓ 0.4s

```
a
```

- A shorter (but possibly more confusing) way of doing the same is:

```
first_char = s and s[0]
```

- since ***and*** returns its second value when the first is "truthy," and the first value when it's not. Similarly, if x is either a number or possibly None:

```
safe_x = x or 0
```

- is definitely a number, although:

```
safe_x = x if x is not None else 0
```

is possibly more readable.

```
1  x = 5
2  safe_x = x or 0
3  print(safe_x)
```
✓ 0.4s

```
5
```

```
1  x = None
2  safe_x = x or 0
3  print(safe_x)
```
✓ 0.4s

```
0
```

```
1  x = None
2  safe_x = x if x is not None else 0
3  print(safe_x)
4
```
✓ 0.5s

```
0
```

Truthiness

- Python has an all function, which takes an iterable and returns True precisely when every element is truthy, and an any function, which returns True when at least one element is truthy:

```
all([True, 1, {3}])   # True, all are truthy
all([True, 1, {}])    # False, {} is falsy
any([True, 1, {}])    # True, True is truthy
all([])               # True, no falsy elements in the list
any([])               # False, no truthy elements in the list
```

Sorting

- Every Python list has a sort method that sorts it in place.
- You can use the sorted function. which returns a new list:

```
x = [4, 1, 2, 3]
y = sorted(x)      # y is [1, 2, 3, 4], x is unchanged
x.sort()           # now x is [1, 2, 3, 4]
```

- By default, sort (and sorted) sort a list from smallest to largest based on naively comparing the elements to one another.
- If you want elements sorted from largest to smallest, you can specify a ***reverse=True*** parameter. And instead of comparing the elements themselves, you can compare the results of a function that you specify with key:

```
# sort the list by absolute value from largest to smallest
x = sorted([-4, 1, -2, 3], key=abs, reverse=True)  # is [-4, 3, -2, 1]

# sort the words and counts from highest count to lowest
wc = sorted(word_counts.items(),
            key=lambda word_and_count: word_and_count[1],
            reverse=True)
```

List Comprehensions

- List comprehension offers a shorter syntax when you want to create a new list based on the values of an existing list.

The Syntax

newlist = [expression for item in iterable if condition == True]

- The return value is a new list, leaving the old list unchanged.

Condition

- The condition is like a filter that only accepts the items that valuate to True.

Iterable

- The iterable can be any iterable object, like a list, tuple, set etc.

Expression

- The expression is the current item in the iteration, but it is also the outcome, which you can manipulate before it ends up like a list item in the new list.

List Comprehensions

```python
# Based on a list of fruits, you want a new list, containing only the fruits with the letter "a" in the name.

# Without list comprehension you will have to write a for statement with a conditional test inside:

fruits = ["apple", "banana", "cherry", "kiwi", "mango"]
newlist = []

for x in fruits:
    if "a" in x:
        newlist.append(x)

print(newlist)
```

```
['apple', 'banana', 'mango']
```

```python
# With list comprehension you can do all that with only one line of code:

fruits = ["apple", "banana", "cherry", "kiwi", "mango"]

newlist = [x for x in fruits if "a" in x]

print(newlist)
```

```
['apple', 'banana', 'mango']
```

List Comprehensions

- Frequently, you'll want to transform a list into another list by choosing only certain elements, by transforming elements, or both.
- The Pythonic way to do this is with list comprehensions:

```
even_numbers = [x for x in range(5) if x % 2 == 0]  # [0, 2, 4]
squares      = [x * x for x in range(5)]            # [0, 1, 4, 9, 16]

even_squares = [x * x for x in even_numbers]        # [0, 4, 16]
```

- You can similarly turn lists into dictionaries or sets:

```
square_dict = {x: x * x for x in range(5)}  # {0: 0, 1: 1, 2: 4, 3: 9, 4:
16}
square_set  = {x * x for x in [1, -1]}      # {1}
```

- If you don't need the value from the list, it's common to use an underscore as the variable:
-

```
zeros = [0 for _ in even_numbers]       # has the same length as
even_numbers
```

```
1  evenNumbers = [1, 2, 3, 4, 5]
2  zeros = [0 for _ in evenNumbers]
3  print(zeros)
✓ 0.4s
[0, 0, 0, 0, 0]
```

List Comprehensions

- A list comprehension can include multiple ***fors***:

```
pairs = [(x, y)
         for x in range(10)
         for y in range(10)]   # 100 pairs (0,0) (0,1) ... (9,8), (9,9)
```

- and later ***fors*** can use the results of earlier ones:

```
increasing_pairs = [(x, y)                      # only pairs with x < y,
                    for x in range(10)          # range(lo, hi) equals
                    for y in range(x + 1, 10)]  # [lo, lo + 1, ..., hi -
1]
```

Dictionary Comprehension

- Dictionary comprehension is an elegant and concise way to create dictionaries.
- The minimal syntax for dictionary comprehension is:

dictionary = {key: value for vars in iterable}

```
1  # Example 1: Dictionary Comprehension
2  # Consider the following code:
3
4  square_dict = dict()
5  for num in range(1, 11):
6      square_dict[num] = num * num
7  print(square_dict)
8
```

```
{1: 1, 2: 4, 3: 9, 4: 16, 5: 25, 6: 36, 7: 49, 8: 64, 9: 81, 10: 100}
```

```
1  # Now, let's create the dictionary in the above program using dictionary comprehension.
2
3  # dictionary comprehension example
4  square_dict = {num: num * num for num in range(1, 11)}
5  print(square_dict)
6
```

```
{1: 1, 2: 4, 3: 9, 4: 16, 5: 25, 6: 36, 7: 49, 8: 64, 9: 81, 10: 100}
```

Dictionary Comprehension

```
1  #item price in dollars
2  old_price = {'milk': 1.02, 'coffee': 2.5, 'bread': 2.5}
3
4  dollar_to_pound = 0.76
5  new_price = {item: value*dollar_to_pound for (item, value) in old_price.items()}
6  print(new_price)
```

```
{'milk': 0.7752, 'coffee': 1.9, 'bread': 1.9}
```

```
1  original_dict = {'jack': 38, 'michael': 48, 'guido': 57, 'john': 33}
2
3  even_dict = {k: v for (k, v) in original_dict.items() if v % 2 == 0}
4  print(even_dict)
```

```
{'jack': 38, 'michael': 48}
```

Automated Testing and assert

- As data scientists, we'll be writing a lot of code. How can we be confident our code is correct?
- One way is with types (discussed shortly), but another way is with automated tests.
- We will be using assert statements, which will cause your code to raise an AssertionError if your specified condition is not truthy:

```
assert 1 + 1 == 2
assert 1 + 1 == 2, "1 + 1 should equal 2 but didn't"
```

- As you can see in the second case, you can optionally add a message to be printed if the assertion fails.
- It's not particularly interesting to assert that 1 + 1 = 2. What's more interesting is to assert that functions you write are doing what you expect them to:

```
def smallest_item(xs):
    return min(xs)

assert smallest_item([10, 20, 5, 40]) == 5
assert smallest_item([1, 0, -1, 2]) == -1
```

Object Oriented Programming

- Python allows you to define classes that encapsulate data and the functions that operate on them.
- Here we'll construct a class representing a "counting clicker," the sort that is used at the door to track how many people have shown up for the "advanced topics in data science" meetup.
- It maintains a count, can be clicked to increment the count, allows you to read_count, and can be reset back to zero.
- To define a class, you use the class keyword and a PascalCase name.
- By convention, each takes a first parameter, self, that refers to the particular class instance.
- Normally, a class has a constructor, named __init__. It takes whatever parameters you need to construct an instance of your class and does whatever setup you need.
- we construct instances of the clicker using just the class name.

random.seed() and random.randrange()

- The random module actually produces pseudorandom (that is, deterministic) numbers based on an internal state that you can set with ***random.seed*** if you want to get reproducible results.

```
1  random.seed(10) # set the seed to 10
2  print(random.random())
3  random.seed(10) # reset the seed to 10
4  print(random.random())
5
```

✓ 0.5s

```
0.5714025946899135
0.5714025946899135
```

- We'll sometimes use ***random.randrange***, which takes either one or two arguments and returns an element chosen randomly from the corresponding range:

```
random.randrange(10)    # choose randomly from range(10) = [0, 1, ..., 9]
random.randrange(3, 6)  # choose randomly from range(3, 6) = [3, 4, 5]
```

random.shuffle() and random.choice()

- There are a few more methods that we'll sometimes find convenient.
- For example, ***random.shuffle*** randomly reorders the elements of a list:

```
1  up_to_ten = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
2  random.shuffle(up_to_ten)
3  print(up_to_ten)
✓ 0.4s
[10, 9, 5, 3, 6, 4, 2, 1, 8, 7]
```

- If you need to randomly pick one element from a list, you can use ***random.choice***

```
1  my_best_friend = random.choice(["Alice", "Bob", "Charlie"])
2  print(my_best_friend)
✓ 0.3s
Alice
```

- And if you need to randomly choose a sample of elements without replacement (i.e., with no duplicates), you can use ***random.sample***:

```
1  lottery_numbers = range(60)
2  winning_numbers = random.sample(lottery_numbers, 6)
3  print(winning_numbers)
✓ 0.4s
[33, 31, 20, 4, 15, 47]
```

- To choose a sample of elements with replacement (i.e., allowing duplicates), you can just make multiple calls to ***random.choice***:

```
1  four_with_replacement = [random.choice(range(10)) for _ in range(4)]
2  print(four_with_replacement)
✓ 0.5s
[5, 0, 6, 2]
```

zip and Argument Unpacking

- Often we will need to zip two or more iterables together.
- The ***zip*** function transforms multiple iterables into a single iterable of tuples of corresponding function:

```
1  list1 = ['a', 'b', 'c']
2  list2 = [1, 2, 3]
3  # zip is lazy, so you have to do something like the following
4  [pair for pair in zip(list1, list2)]
✓ 0.7s
[('a', 1), ('b', 2), ('c', 3)]
```

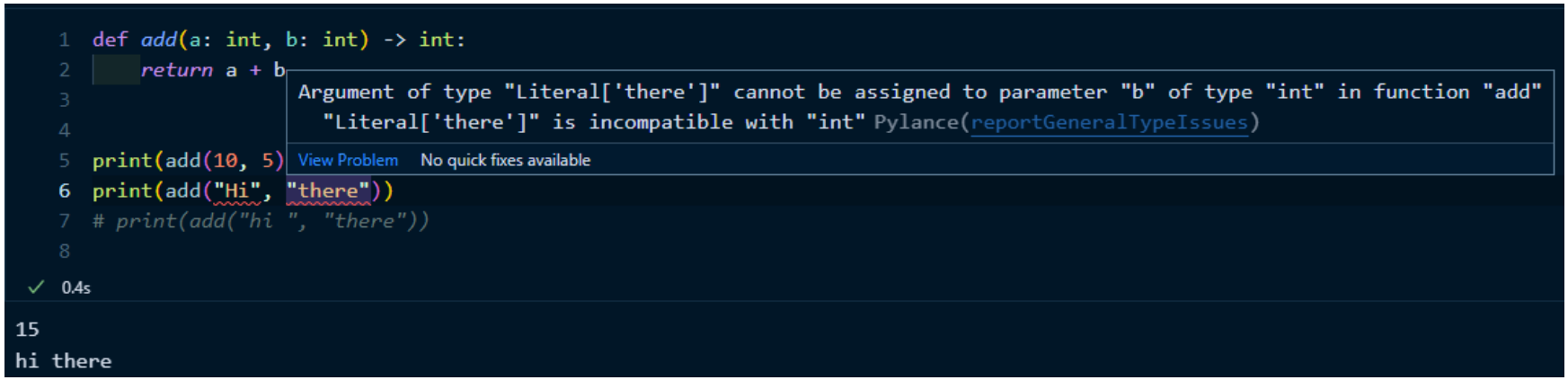
- If the lists are different lengths, zip stops as soon as the first list ends.
- You can also "unzip" a list using a strange trick:

```
pairs = [('a', 1), ('b', 2), ('c', 3)]
letters, numbers = zip(*pairs)
```

- The asterisk (*) performs argument unpacking, which uses the elements of pairs as individual arguments to zip.

Type Annotations

- Python is a dynamically typed language.
- That means that it in general it doesn't care about the types of objects we use, as long as we use them in valid ways.
- Whereas in a statically typed language our functions and objects would have specific types.



- Type annotations allow us to tell Python what we expect to be assigned to names at given points in our application. We can then use these annotations to check the program is in fact doing what we intend.
- However, these type annotations don't actually do anything. You can still use the annotated add function to add strings, and the call to add(10, "five") will still raise the exact same TypeError.

Type Hinting

- In a nutshell: **Type hinting** is literally what the words mean. You hint the type of the object(s) you're using.
- Due to the dynamic nature of Python, inferring or checking the type of an object being used is especially hard.
- This fact makes it hard for developers to understand what exactly is going on in code they haven't written and, most importantly, for type checking tools found in many IDEs that are limited due to the fact that they don't have any indicator of what type the objects are.
- As a result they resort to trying to infer the type with around 50% success rate.

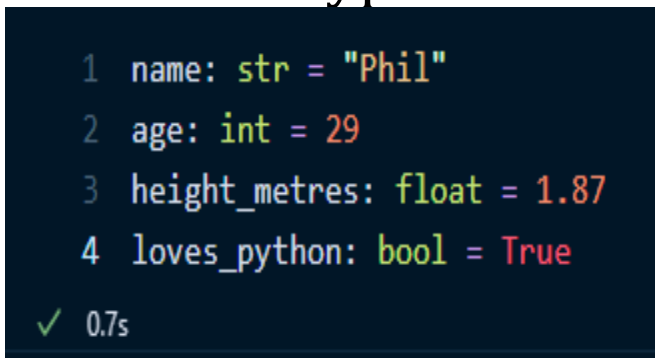
Type Annotations

- There are still (at least) four good reasons to use type annotations in your Python code:

1. Types are an important form of documentation.
2. There are external tools (mypy and pylance) that will read your code, inspect the type annotations, and let you know about type errors before you ever run your code.
3. Having to think about the types in your code forces you to design cleaner functions and interfaces.
4. Using types allows your editor to help you with things like autocomplete and to get angry at type errors.

Type Annotations

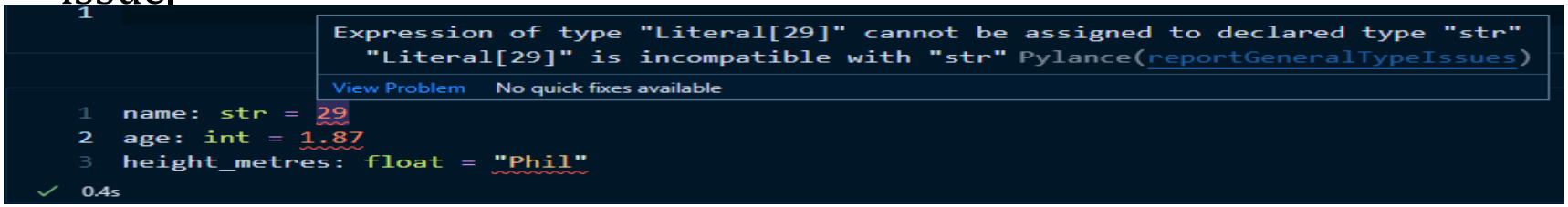
- Let's start by using type hinting to annotate some variables that we expect to take basic types.

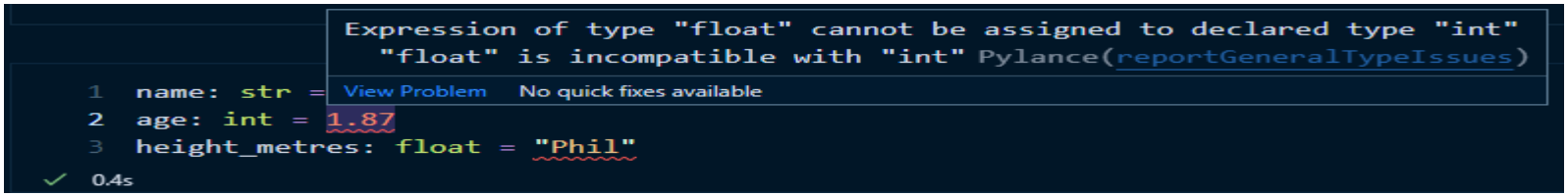


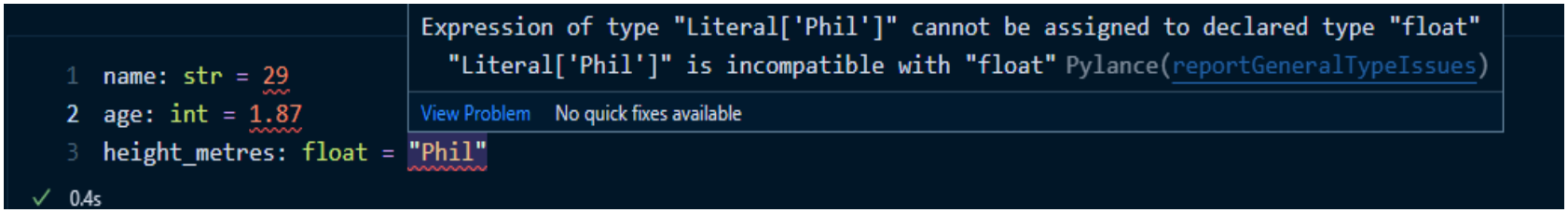
- As you can see from the examples above, we can annotate a variable by adding a colon after the variable name, after which we specify the type.
- Here we've indicated that name is a string, age is an integer, and height_metres should be a float.
- In this case, all of our type annotations align with the values that have been assigned, so we don't have any issues.

Annotating collections

- Now let's look at what happens when things are not what we intended.
- All of our values have gotten jumbled up. Maybe we assigned these values from an iterable that contained the values in the wrong order.
- The errors are actually very helpful. They explain exactly what went wrong, and at the start of each line we get a reference to the file where the error happened, along with a line number.
- Using this information, we can easily track down the source of this issue.

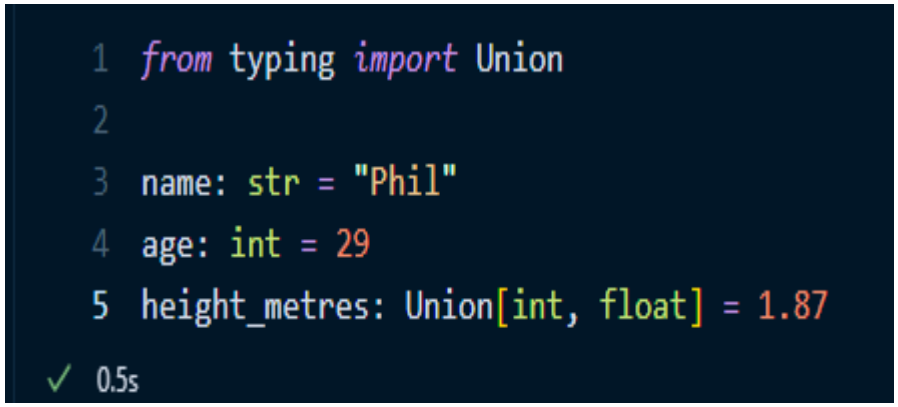






Using *typing* module and *Union*

- Let's say we want to be a little more flexible in how we handle height_metres.
- We only really care that it's a real number, so instead of accepting just floats, I want to also accept integers.
- The way we accomplish this is by using a tool called ***Union*** which we have to import from the ***typing*** module.
- Here we've added ***Union[int, float]*** as a type annotation for ***height_metres***, which means we can accept either integers or floats.
- We can add as many types as we like to this Union by adding more comma separated values between the square brackets.



- The ***typing*** module provides runtime support for type hints.

Using *Any*

- We can also get super flexible and use another tool called ***Any***, which matched any type.
- You should be careful about using ***Any***, because it largely removed the benefits of type hinting.
- It can be useful for indicating to readers that something is entirely generic though.
- We can use ***Any*** like any of the other types:

```
1 from typing import Any
2
3 name: str = "Phil"
4 age: int = 29
5 height_metres: Any = 1.87
✓ 0.4s
```

Annotating collections

- Now that we've looked at annotating basic types, let's talk about how we might annotate that something should be a list, or maybe a tuple containing values of a specific type.
- In order to annotate collections, we have to import special types from the ***typing*** module.
- For lists we need to import ***List*** and for tuples we need to import ***Tuple***.
- As you can see, the names make a lot of sense.

```
1 from typing import List
2
3 names: List = ["Rick", "Morty", "Summer", "Beth", "Jerry"]
✓ 0.5s
```

- Here is an example of a variable using a List annotation:

Annotating collections

- If we wanted to specify which types should in the list, we can add a set of square brackets, much like we did with Union.

```
1  from typing import List
2
3  names: List[str] =  ["Rick", "Morty", "Summer", "Beth", "Jerry"]
✓ 0.6s
```

- If we want, we can allow a variety of types in a list by combining List and Union like this:

```
1  from typing import List, Union
2
3  random_values: List[Union[str, int]] =  ["x", 13, "camel", 0]
✓ 0.5s
```

- When working with tuples, we can specify a type for each item in sequence, since tuples are of fixed length and are immutable.
- For example, we can do something like this:

```
1  from typing import Tuple
2
3  movie: Tuple[str, str, int] = ("Toy Story 3", "Lee Unkrich", 2010)
✓ 0.4s
```

Creating type aliases

- Let's consider a case where we want to store lots of movies. How would we annotate something like that?
- Maybe something like this:

```python
from typing import List, Tuple

movies: List[Tuple[str, str, int]] = [
    ("Finding Nemo", "Andrew Stanton", 2005),
    ("Inside Out", "Pete Docter", 2015),
    ("Toy Story 3", "Lee Unkrich", 2010)
]

```

- This does work, but that type annotation is getting very hard to read.
- That's a problem, because one of the benefits of using type annotations is to help with readability.
- In cases like this where we have complex type annotations, it's often better to define new aliases for certain type combinations. For example, it makes a lot of sense to call each of these tuples a ***Movie***.
- We can do this like so:

```python
from typing import List, Tuple

Movie = Tuple[str, str, int]

movies: List[Movie] = [
    ("Finding Nemo", "Andrew Stanton", 2005),
    ("Inside Out", "Pete Docter", 2015),
    ("Toy Story 3", "Lee Unkrich", 2010)
]
```

Annotating functions

- Now let's get into annotating functions, which is where this kind of tool is most useful.
- The function below is for the subtraction of an integer value, which subtracts two integers and returns the value.
- Here the function above needs to accept two integers 'x' and 'y', but since the rules are not imposed, can take any data type.

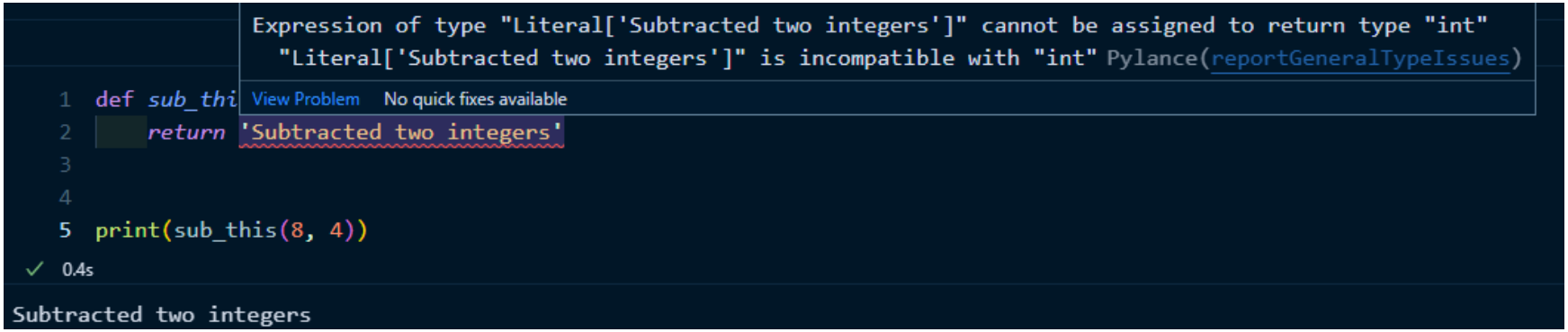
```
1  def sub_this(x, y):
2      return 'Subtraction'
3
4  print(sub_this(8, 'hello'))
5
✓ 0.4s
Subtraction
```

- Also, the return value can be anything where the 'str' value('Subtraction') is returned, but expected was 'int'.
- Let's see the similar implementation of the above program, but by using type hints which can help to implement static type checking and reduce error and bugs in the program quite easily.
- The below code is a simple program that accepts two integers as an input in the parameter and after '**->**' shows the returned data type, which is also an 'int'. However, the function needs to return int, but string 'Subtracted two integers' is returned.
-

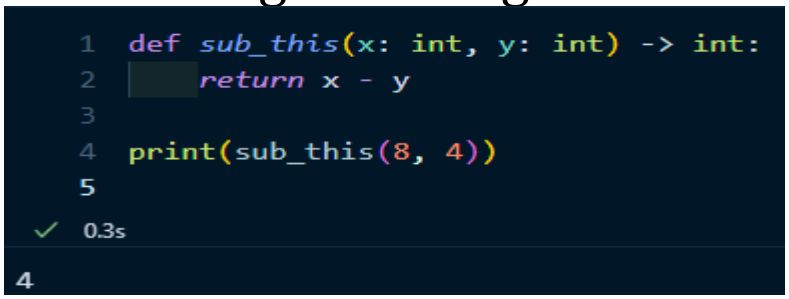
```
1  def sub_this(x: int, y: int) -> int:
2      return 'Subtracted two integers'
3
4
5  print(sub_this(8, 4))
6
✓ 0.5s
Subtracted two integers
```

Annotating functions

- Then the error will be shown, which indicates that the unexpected return value "str" is found and needs to be resolved with "int".



- Let's change the return type to be the subtraction of the two integers so that the integer value gets returned.



- The above results show that the success message gets printed out, and no issues were found.

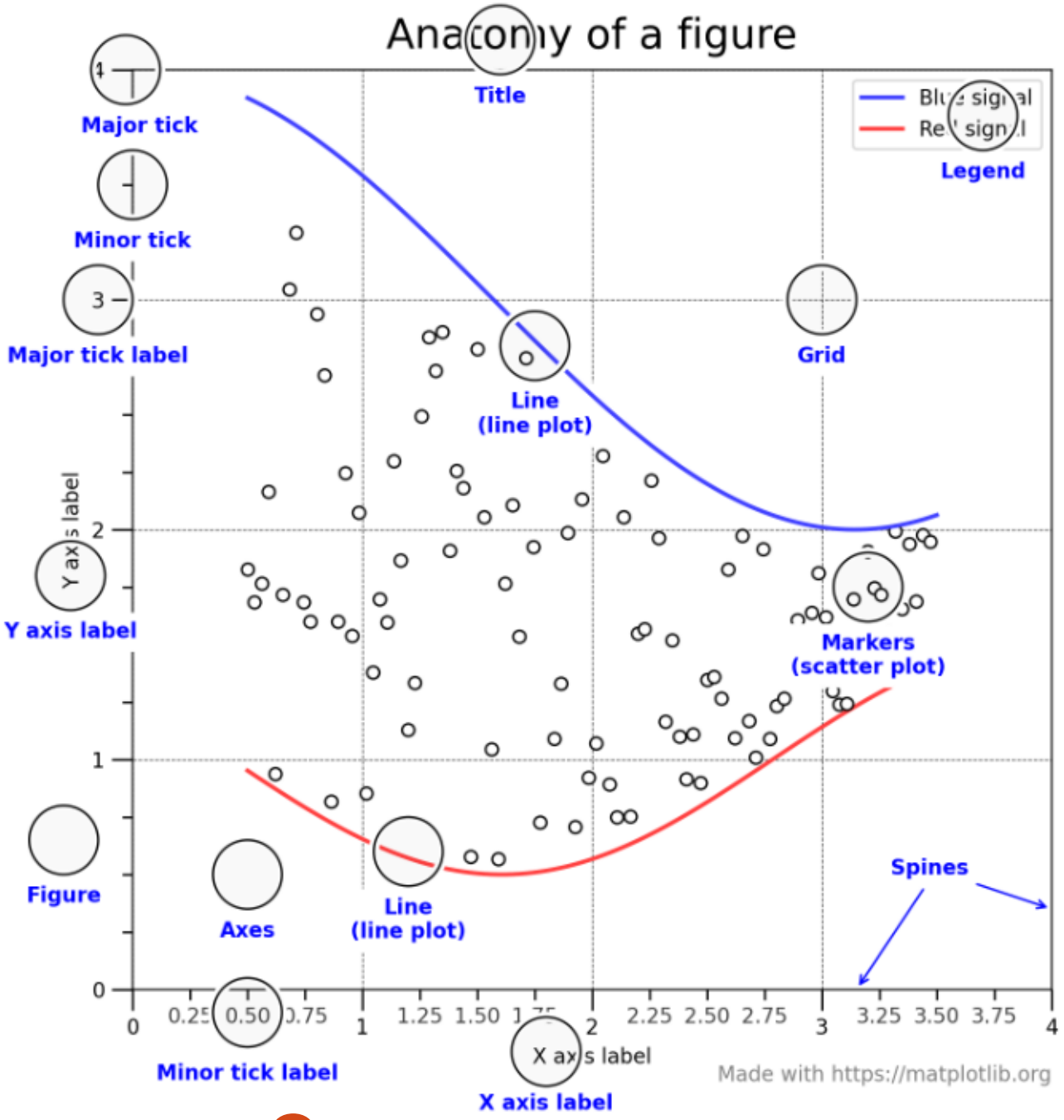
Visualizing Data

- A fundamental part of the data scientist's toolkit is data visualization.
- Data visualization is essential in today's world because it helps people understand complex data quickly and efficiently.
- Here are some key reasons why it is important:
- **Simplifies Complex Data**
- **Enhances Decision-Making**
- **Identifies Trends and Patterns**
- **Saves Time**

matplotlib

- A wide variety of tools exist for visualizing data. We will be using the matplotlib library, which is widely used
- We will be using the ***matplotlib.pyplot*** module.
- In its simplest use, ***pyplot*** maintains an internal state in which you build up a visualization step by step.
- Once you're done, you can save it with ***savefig*** or display it with ***show***.

Matplotlib


Anatomy of a figure
Title
Major tick
Minor tick
Major tick label
Legend
Grid
Line
(line plot)
Y axis label
Markers
(scatter plot)
Figure
Axes
Line
(line plot)
Spines
Minor tick label
X axis label
Made with https://matplotlib.org

Lists

- Probably the most **fundamental data structure in Python is the list,** which is simply an ordered collection.

```
integer_list = [1, 2, 3]
heterogeneous_list = ["string", 0.1, True]
list_of_lists = [integer_list, heterogeneous_list, []]

list_length = len(integer_list)      # equals 3
list_sum    = sum(integer_list)      # equals 6
```

- You can get or set the nth element of a list with square brackets:

```
x = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]

zero = x[0]          # equals 0, lists are 0-indexed
one = x[1]           # equals 1
nine = x[-1]         # equals 9, 'Pythonic' for last element
eight = x[-2]        # equals 8, 'Pythonic' for next-to-last element
x[0] = -1            # now x is [-1, 1, 2, 3, ..., 9]
```

Lists

- If you don't want to modify x, you can use list addition:

```
x = [1, 2, 3]
y = x + [4, 5, 6]        # y is [1, 2, 3, 4, 5, 6]; x is unchanged
```

- More frequently we will append to lists one item at a time:

```
x = [1, 2, 3]
x.append(0)        # x is now [1, 2, 3, 0]
y = x[-1]          # equals 0
z = len(x)         # equals 4
```

- It's often convenient to unpack lists when you know how many elements they contain:

```
x, y = [1, 2]     # now x is 1, y is 2
```

although you will get a *ValueError* if you don't have the same number of elements on both sides.

- A common idiom is to use an underscore for a value you're going to throw away:

```
_, y = [1, 2]     # now y == 2, didn't care about the first element
```

Modules

- You might also do this if your module has an unwieldy name or if you're going to be typing it a lot.
- For example, a standard convention when visualizing data with matplotlib is:

```
import matplotlib.pyplot as plt

plt.plot(...)
```

If you need a few specific values from a module, you can import them explicitly and use them without qualification:

```
from collections import defaultdict, Counter
lookup = defaultdict(int)
my_counter = Counter()
```

Matplotlib

The syntax of the `plot` function is `plot(xl, yl, options)` where:

- `xl` is the list of x-coordinates of points we want to plot
- `yl` is the list of y-coordinates
- other options may specify how we want the plot to look like.

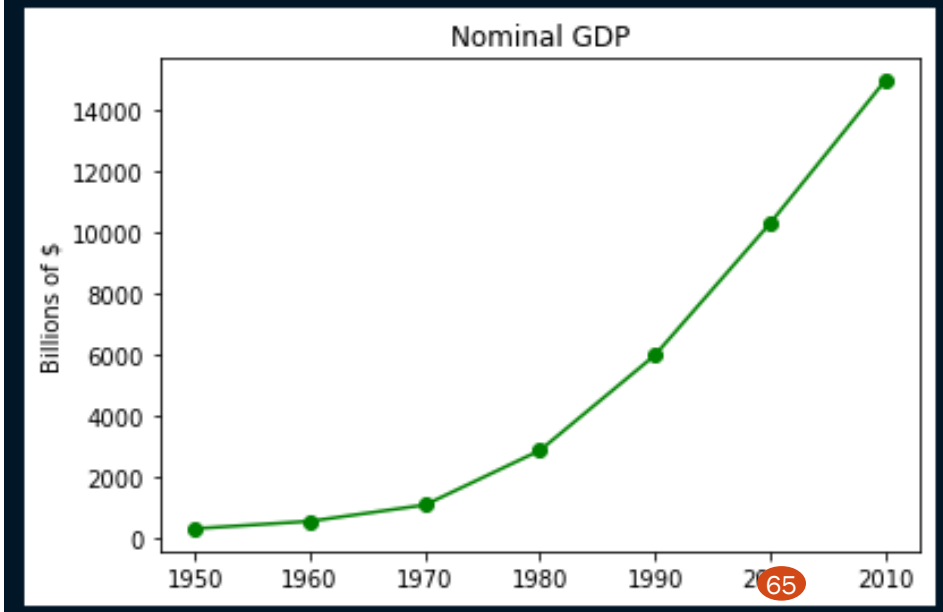
function	usage
`plot`	creates a plot
`title`	sets a title of the plot
`xlabel`	sets label of the x-axis
`ylabel`	sets label of the y-axis
`xlim`	set the range of x values displayed
`ylim`	set the range of y values displayed
`show`	displays the plot

shape	meaning	color	meaning
`'-'`	solid line	`'b'`	blue
`'--'`	dashed line	`'g'`	green
`'.'`	point marker	`'r'`	red
`'o'`	circle marker	`'c'`	cyan
`'s'`	square marker	`'w'`	white
`'+'`	plus marker	`'m'`	magenta
`'x'`	x marker	`'y'`	yellow
`'D'`	diamond marker	`'k'`	black

Visualizing Data

```python
from matplotlib import pyplot as plt
years = [1950, 1960, 1970, 1980, 1990, 2000, 2010]
gdp = [300.2, 543.3, 1075.9, 2862.5, 5979.6, 10289.7, 14958.3]
# create a line chart, years on x-axis, gdp on y-axis
plt.plot(years, gdp, color='green', marker='o', linestyle='solid')
# add a title
plt.title("Nominal GDP")
# add a label to the y-axis
plt.ylabel("Billions of $")
plt.show()

```

✓ 8.7s

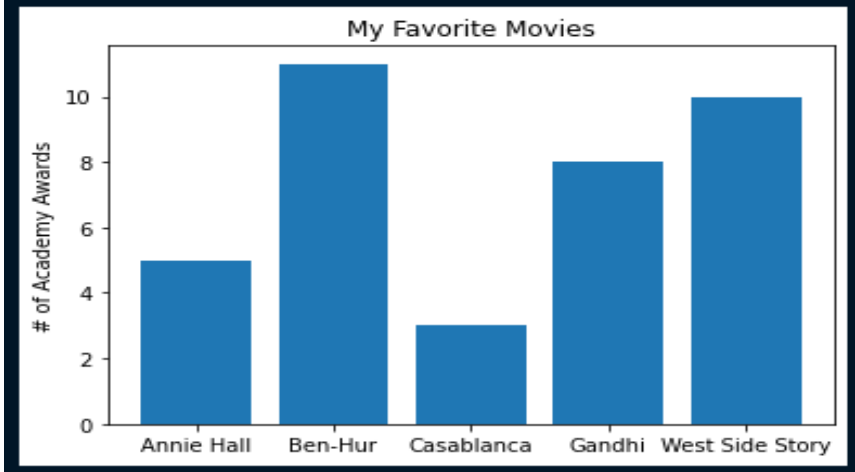


Bar Charts

- A bar chart is a good choice when you want to show how some quantity varies among some discrete set of items.
- ***Syntax: plt.bar(x, height, width, bottom, align)***
- For instance, the below figure shows how many Academy Awards were won by each of a variety of movies.

```
1  from matplotlib import pyplot as plt
2  movies = ["Annie Hall", "Ben-Hur", "Casablanca", "Gandhi", "West Side Story"]
3  num_oscars = [5, 11, 3, 8, 10]
4  # plot bars with left x-coordinates [0, 1, 2, 3, 4], heights [num_oscars]
5  plt.bar(range(len(movies)), num_oscars)
6  plt.title("My Favorite Movies")  # add a title
7  plt.ylabel("# of Academy Awards")  # label the y-axis
8  # label x-axis with movie names at bar centers
9  plt.xticks(range(len(movies)), movies)
10 plt.show()
```

✓ 0.1s

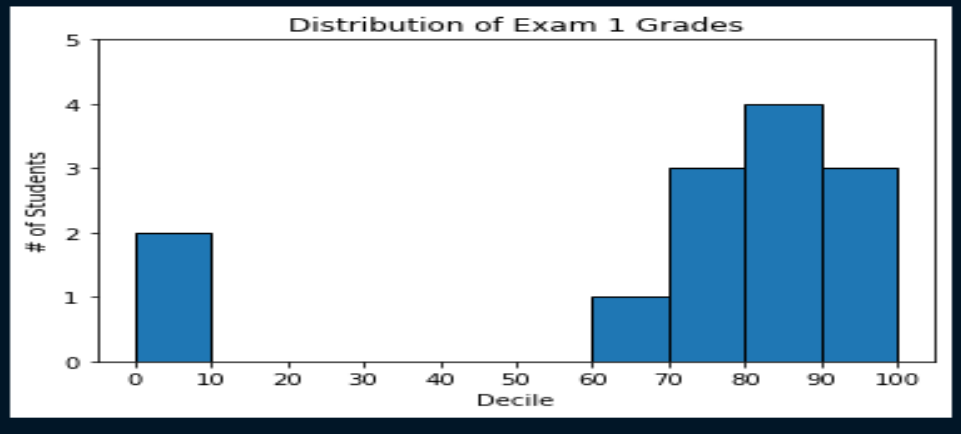


Bar Charts: Histograms

- A bar chart can also be a good choice for plotting histograms of bucketed numeric values, as in the below figure, in order to visually explore how the values are distributed.

```
from matplotlib import pyplot as plt
from collections import Counter
grades = [83, 95, 91, 87, 70, 0, 85, 82, 100, 67, 73, 77, 0]
# Bucket grades by decile, but put 100 in with the 90s
histogram = Counter(min(grade // 10 * 10, 90) for grade in grades)
plt.bar([x + 5 for x in histogram.keys()],  # Shift bars right by 5
        histogram.values(),  # Give each bar its correct height
        10,  # Give each bar a width of 10
        edgecolor=(0, 0, 0))  # Black edges for each bar
plt.axis([-5, 105, 0, 5])  # x-axis from -5 to 105,
# y-axis from 0 to 5
plt.xticks([10 * i for i in range(11)])  # x-axis labels at 0, 10, ...,100
plt.xlabel("Decile")
plt.ylabel("# of Students")
plt.title("Distribution of Exam 1 Grades")
plt.show()
```

0.1s



Bar Charts

- The third argument to ***plt.bar*** specifies the bar width. Here we chose a width of 10, to fill the entire decile.
- We also shifted the bars right by 5, so that, for example, the "10" bar (which corresponds to the decile 10–20) would have its center at 15 and hence occupy the correct range.
- We also added a black edge to each bar to make them visually distinct.
- The call to ***plt.axis*** indicates that we want the x-axis to range from –5 to 105 (just to leave a little space on the left and right), and that the y-axis should range from 0 to 5.
- And the call to ***plt.xticks*** puts x-axis labels at 0, 10, 20, …, 100.

Control Flow

- As in most programming languages, you can perform an action conditionally using if:

```
if 1 > 2:
    message = "if only 1 were greater than two..."
elif 1 > 3:
    message = "elif stands for 'else if'"
else:
    message = "when all else fails use else (if you want to)"
```

- You can also write a ternary if-then-else on one line, which we will do occasionally:

```
parity = "even" if x % 2 == 0 else "odd"
```

- Python has a while loop:

```
x = 0
while x < 10:
    print(f"{x} is less than 10")
    x += 1
```

```
1  x = 3
2  if x % 2 == 0:
3      parity = "even"
4  else:
5      parity = "odd"
6  print(parity)
7
```

✓ 0.4s

odd

```
1  x = 3
2  parity = "even" if x % 2 == 0 else "odd"
3  print(parity)
4
```

✓ 0.4s

odd

Control Flow

- for and in:

```
# range(10) is the numbers 0, 1, ..., 9
for x in range(10):
    print(f"{x} is less than 10")
```

- If you need more complex logic, you can use continue and break:

```
for x in range(10):
    if x == 3:
        continue  # go immediately to the next iteration
    if x == 5:
        break     # quit the loop entirely
    print(x)
```

- This will print 0, 1, 2, and 4.

zip and Argument Unpacking

- Often we will need to zip two or more iterables together.
- The ***zip*** function transforms multiple iterables into a single iterable of tuples of corresponding function:

```
1  list1 = ['a', 'b', 'c']
2  list2 = [1, 2, 3]
3  # zip is lazy, so you have to do something like the following
4  [pair for pair in zip(list1, list2)]
✓ 0.7s
[('a', 1), ('b', 2), ('c', 3)]
```

- If the lists are different lengths, zip stops as soon as the first list ends.
- You can also "unzip" a list using a strange trick:

```
pairs = [('a', 1), ('b', 2), ('c', 3)]
letters, numbers = zip(*pairs)
```

- The asterisk (*) performs argument unpacking, which uses the elements of pairs as individual arguments to zip.

List Comprehensions

- List comprehension offers a shorter syntax when you want to create a new list based on the values of an existing list.

The Syntax

newlist = [expression for item in iterable if condition == True]

- The return value is a new list, leaving the old list unchanged.

Condition

- The condition is like a filter that only accepts the items that valuate to True.

Iterable

- The iterable can be any iterable object, like a list, tuple, set etc.

Expression

- The expression is the current item in the iteration, but it is also the outcome, which you can manipulate before it ends up like a list item in the new list.

List Comprehensions

- Frequently, you'll want to transform a list into another list by choosing only certain elements, by transforming elements, or both.

```
even_numbers = [x for x in range(5) if x % 2 == 0]  # [0, 2, 4]
squares      = [x * x for x in range(5)]            # [0, 1, 4, 9, 16]
```

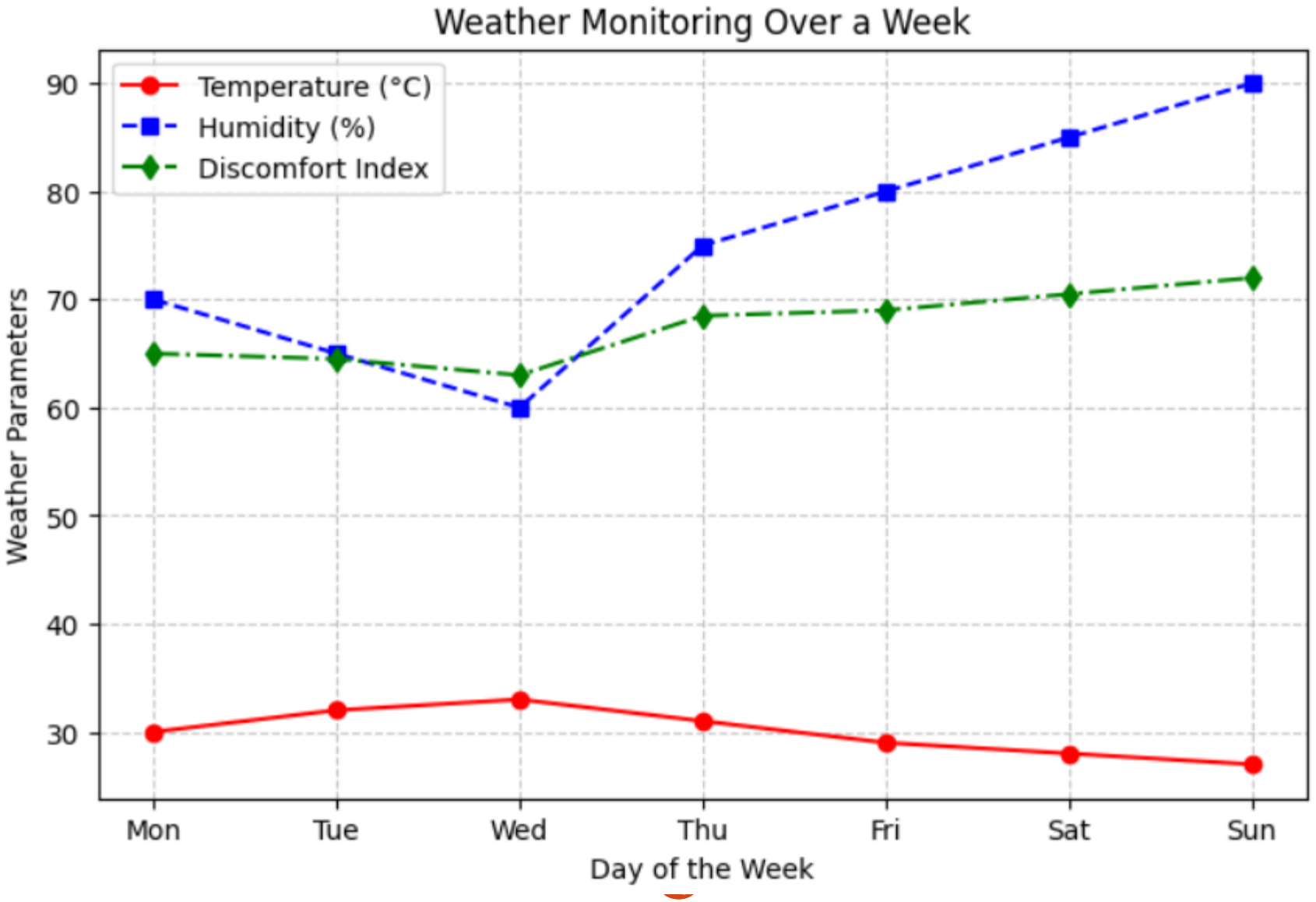
- list comprehensions:

```
even_squares = [x * x for x in even_numbers]        # [0, 4, 16]

square_dict = {x: x * x for x in range(5)}  # {0: 0, 1: 1, 2: 4, 3: 9, 4:
16}
square_set  = {x * x for x in [1, -1]}      # {1}
```

Multiple Line Charts: using plot()

```
import matplotlib.pyplot as plt
# Sample data: Days of the week
days = ["Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"]
temperature = [30, 32, 33, 31, 29, 28, 27] # Temperature (°C)
humidity = [70, 65, 60, 75, 80, 85, 90] # Humidity (%)
# Discomfort Index (A rough measure of heat discomfort:Temp + 0.5 * Humidity)
discomfort_index = [t + 0.5 * h for t, h in zip(temperature, humidity)]
plt.figure(figsize=(8, 5)) # Create the line plot
plt.plot(days, temperature, 'r-o', label="Temperature (°C)")  # Red solid line with circles
plt.plot(days, humidity, 'b--s', label="Humidity (%)")  # Blue dashed line with squares
plt.plot(days, discomfort_index, 'g-.d', label="Discomfort Index")  # Green dot-dashed line with diamonds
plt.xlabel("Day of the Week")
plt.ylabel("Weather Parameters")
plt.title("Weather Monitoring Over a Week")
plt.legend() # Adding legend
plt.grid(True, linestyle="--",) # Grid for better readability
plt.show()
```

Multiple Line Charts

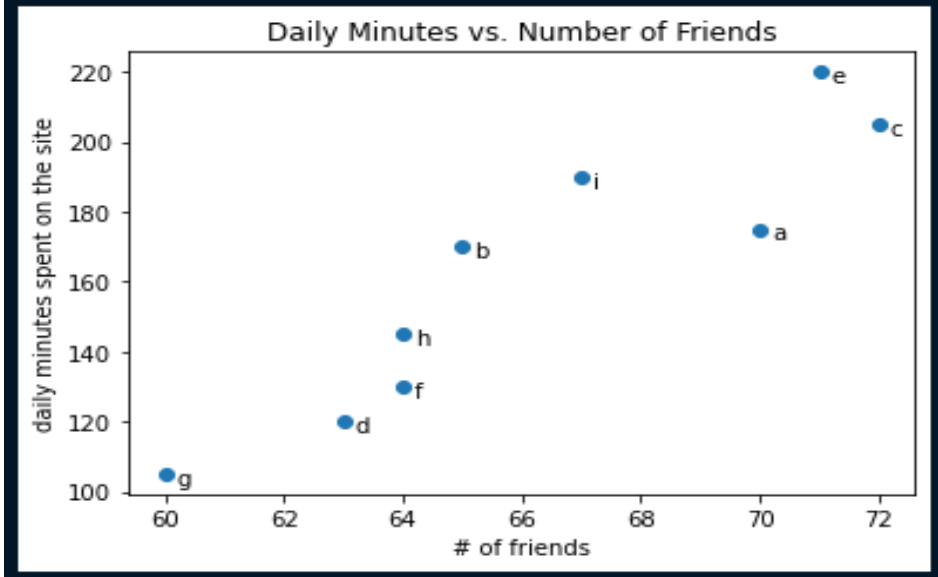


Scatterplots

- A scatterplot is the right choice for visualizing the relationship between two paired sets of data.
- For example, the below figure illustrates the relationship between the number of friends your users have and the number of minutes they spend on the site every day:

```python
from matplotlib import pyplot as plt
friends = [70, 65, 72, 63, 71, 64, 60, 64, 67]
minutes = [175, 170, 205, 120, 220, 130, 105, 145, 190]
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i']
plt.scatter(friends, minutes)
# label each point
for label, friend_count, minute_count in zip(labels, friends, minutes):
    plt.annotate(label,
                 xy=(friend_count, minute_count),  # Put the label with its point
                 xytext=(5, -5),  # but slightly offset
                 textcoords='offset points')
plt.title("Daily Minutes vs. Number of Friends")
plt.xlabel("# of friends")
plt.ylabel("daily minutes spent on the site")
plt.show()

```

✓ 0.1s



Linear Algebra

- Linear algebra is the branch of mathematics that deals with vector spaces.

Vectors

- Physical quantity with magnitude and no direction is known as a **scalar** quantity and a physical quantity with magnitude and directions is known as a **vector** quantity.
- Vectors are objects that can be added together to form new vectors and that can be multiplied by scalars (i.e., numbers), also to form new vectors
- Vectors, are often a useful way to represent numeric data.
- **For example**, if you have the heights, weights, and ages of a large number of people, you can treat your data as **three-dimensional vectors** [height, weight, age].
- height_weight_age = [70, # *inches,*
  170, # *pounds,*
  40] # *years*

Vectors

For example, in a 2D space, a vector might look like:

$$v = [3, 4]$$

In a 3D space:

$$w = [1, -2, 5]$$

Vectors

Size (sq.ft)	Number of Rooms	Price ($)
850	2	180,000
1,200	3	250,000
1,500	3	300,000
1,800	4	350,000
2,000	4	400,000
2,500	5	500,000
1,100	2	210,000
1,400	3	275,000
1,700	4	320,000
2,200	4	450,000

Vectors

Uses of Vectors in Data Science

Vectors are fundamental in **data science and machine learning**. Here's why:

1. Representing Data Points

- Each **data point** in a dataset can be represented as a vector.
- Example: A house dataset with features:

$$\text{House} = [\text{Size (sq ft)}, \text{Number of rooms}, \text{Price}]$$

$$[1500, 3, 200000]$$

Each row is a **vector** in a high-dimensional space.

2. Feature Engineering

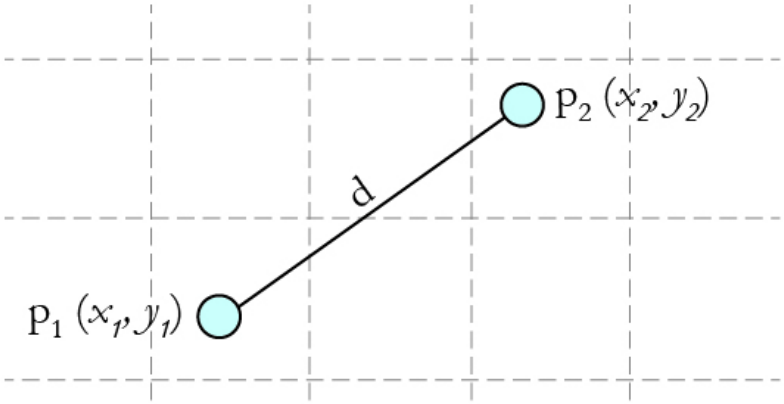
- In machine learning, vectors are used to represent **features** (input variables).
- Example: In an **image dataset**, a 28x28 grayscale image can be represented as a **784-dimensional vector** (flattened pixels).

Vectors

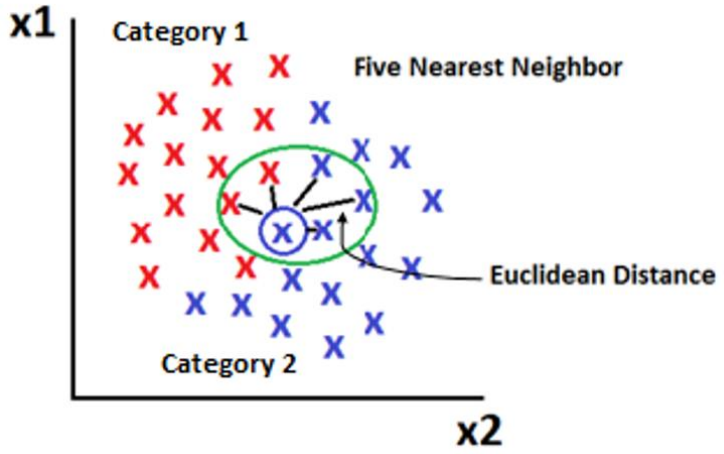
3. Distance Calculation (Similarity & Clustering)

- **Euclidean distance** between two vectors is used to measure similarity.
- Example: In **K-Nearest Neighbors (KNN)**, vectors represent points in space, and the algorithm finds the closest ones.

$$\text{Distance}(v, w) = \sqrt{(v_1 - w_1)^2 + (v_2 - w_2)^2 + ... + (v_n - w_n)^2}$$



Euclidean distance (d) $= \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

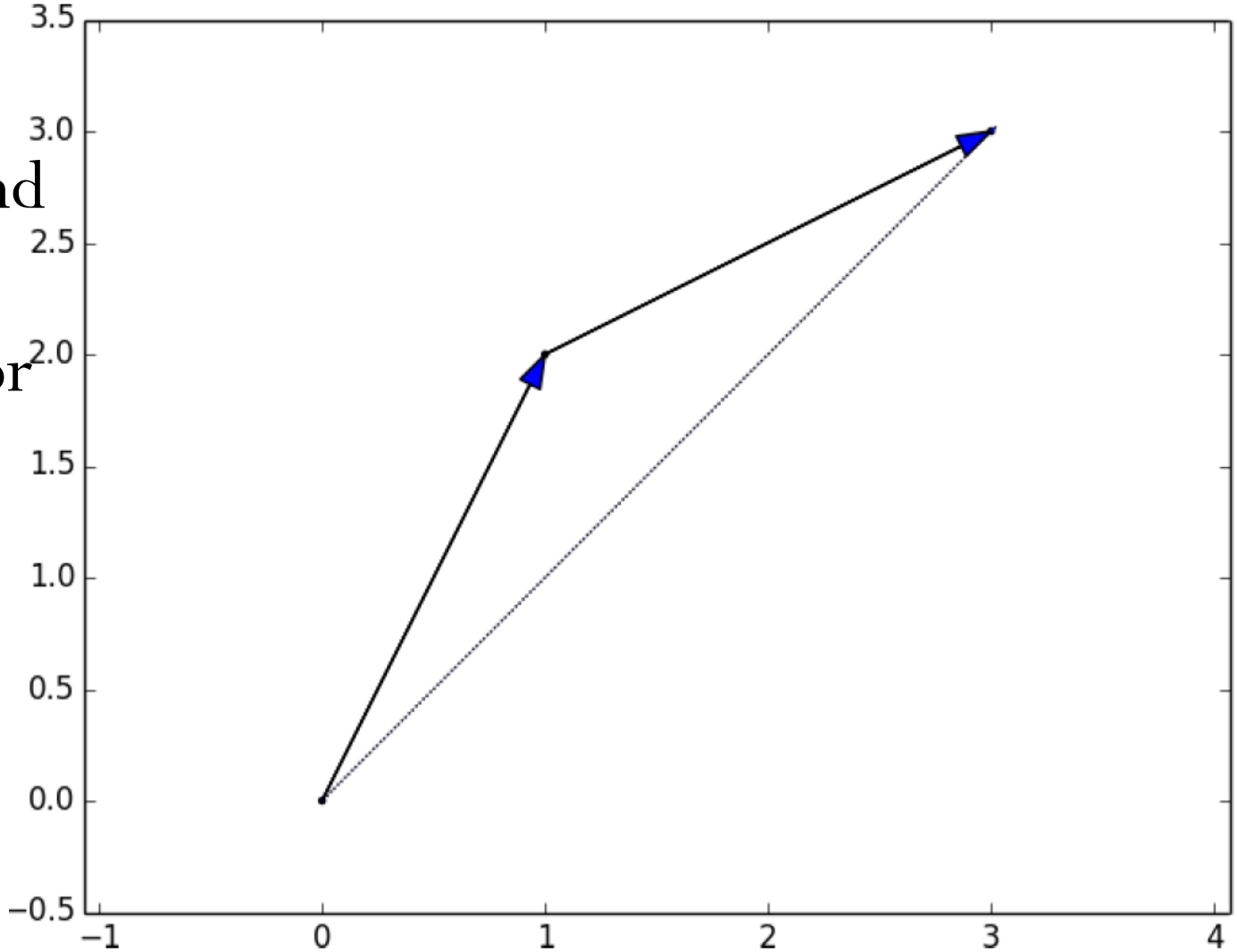


Vectors

- **Perform arithmetic on vectors.**
- Because Python lists aren't vectors (and hence provide no facilities for vector arithmetic), we'll need to build these arithmetic tools ourselves.
- Add two vectors. Vectors add component wise.
- This means that if two vectors **v and w** are the same length, their sum is just the vector whose first element is v[0] + w[0], whose second element is v[1] + w[1], and so on.
- (If they're not the same length, then we're not allowed to add them.)

Adding two Vectors

For example, adding the vectors [1, 2] and [2, 1] results in [1 + 2, 2 + 1] or [3, 3], as shown in Figure 4-1.



Adding two vectors

Add two Vectors

- If the Vectors are of not the same length

```
def vector_add(v, w):
  """adds corresponding elements"""
  return [v_i + w_i   for v_i, w_i in zip(v, w)]
print(vector_add([5, 7, 9], [4, 5, 6]))
```

Subtract two Vectors

- To subtract two vectors we just subtract the corresponding elements:

```
def vector_subtract(v, w):
  """subtracts corresponding elements"""
  return [v_i - w_i   for v_i, w_i in zip(v, w)]
print(vector_subtract([5, 7, 9], [4, 5, 6]))
```

Component wise sum a list of Vectors

- We'll also sometimes want to component wise sum a list of vectors—that is, create a new vector whose first element is the sum of all the first elements, whose second element is the sum of all the second elements, and so on:

```
def vector_sum(vectors):
    """sums all corresponding elements"""
    result = vectors[0] # start with the first vector
    for vector in vectors[1:]: #loop over the others
        result = vector_add(result, vector) #add to result
        return result
print(vector_sum([[1, 2], [3, 4], [5, 6], [7, 8]]))
```

Multiply a vector by a scalar

- We'll also need to be able to multiply a vector by a scalar, which we do simply by multiplying each element of the vector by that number:

```
def scalar_multiply(c, v):
    """c is a number, v is a vector"""
    return [c * v_i for v_i in v]
print(scalar_multiply(2, [1, 2, 3]))
```

Compute Component wise Means

- This allows us to compute the component wise means of a list of (same sized) vectors:

```
def vector_mean(vectors):
  """compute the vector whose ith element is the mean of the
  ith elements of the input vectors"""
  n = len(vectors)
  return scalar_multiply(1/n, vector_sum(vectors))
print(vector_mean([[1, 2], [3, 4], [5, 6]]))
```

```
[1.3333333333333333, 2.0]
```

Dot Product

- The dot product of two vectors is the sum of their component wise products:

1. Dot Product Formula

The **dot product** of two vectors v and w in an **n-dimensional space** is given by:

$$v \cdot w = v_1 \cdot w_1 + v_2 \cdot w_2 + \ldots + v_n \cdot w_n$$

In summation notation:

$$v \cdot w = \sum_{i=1}^{n} v_i \cdot w_i$$

where:

- v and w are vectors.
- v_i and w_i are the **corresponding elements** of each vector.

Dot Product

- The dot product of two vectors is the sum of their component wise products:

2. Example Calculation

Let's calculate the dot product of the two vectors:

$$v = [1, 2, 3], \quad w = [4, 5, 6]$$

Using the formula:

$$(1 \times 4) + (2 \times 5) + (3 \times 6)$$

$$= 4 + 10 + 18 = 32$$

So, the dot product $v \cdot w = 32$.

Dot Product

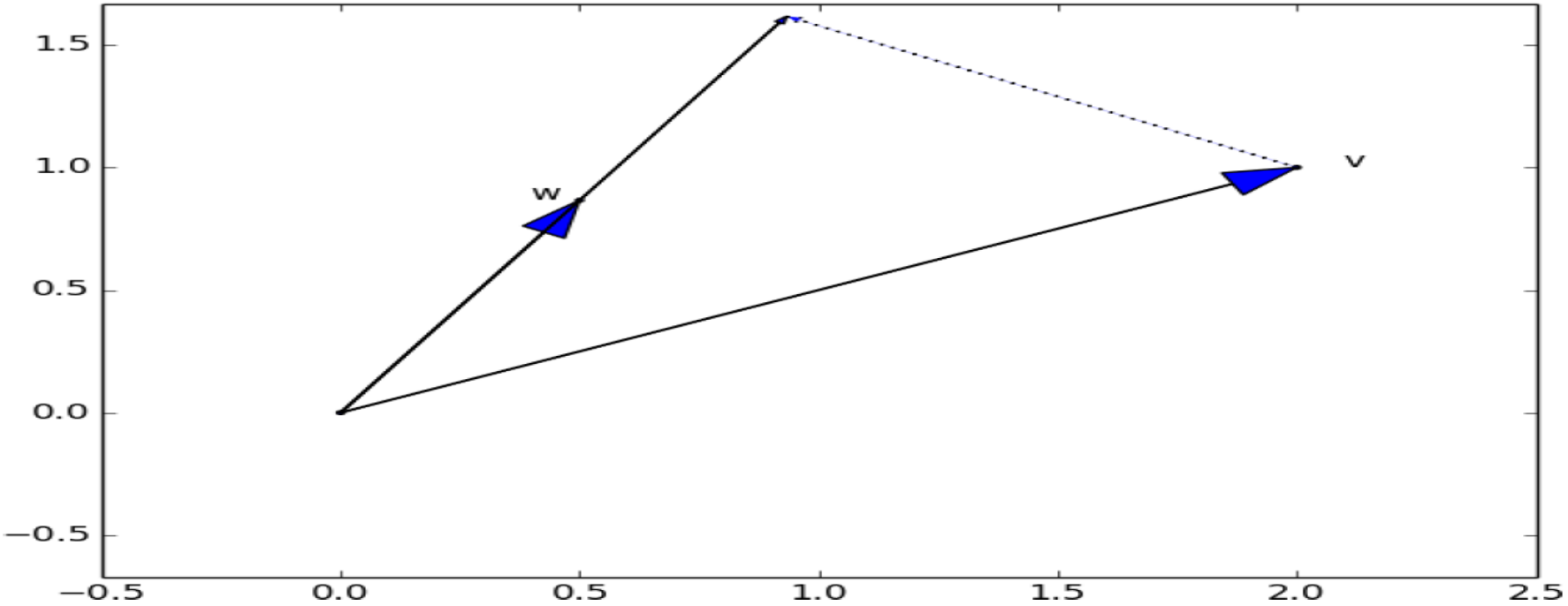
- The dot product of two vectors is the sum of their component wise products:

```
def dot(v, w):
  """v_1 * w_1 + ... + v_n * w_n"""
  return sum(v_i * w_i for v_i, w_i in zip(v, w))
print(dot([1, 2, 3], [4, 5, 6]))
```

```
32
```

Dot Product

- If w has magnitude 1, the dot product measures how far the vector v extends in the w direction.
- For example, if ***w = [1, 0]***, then ***dot(v, w)*** is just the first component of v.
- Another way of saying this is that it's the length of the vector you'd get if you projected ***v*** onto ***w***.



The dot product as vector projection

Sum of Squares

- Using this, it's easy to compute a vector's sum of squares:

```
def sum_of_squares(v):
  """v_1 * v_1 + ... + v_n * v_n"""
  return dot(v, v)
print(sum_of_squares([1, 2, 3]))
```

Compute Magnitude(length)

- Using this, it's easy to compute a vector's sum of squares:

Mathematical Formula of Magnitude (Euclidean Norm)

For a vector **v** with **n** components:

$$v = [v_1, v_2, ..., v_n]$$

The magnitude (or Euclidean norm) is given by:

$$||v|| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}$$

Example Calculation

Let's take a vector:

$$v = [3, 4]$$

Using the formula:

$$||v|| = \sqrt{3^2 + 4^2} = \sqrt{9 + 16} = \sqrt{25} = 5$$

Compute magnitude (or length)

```
def dot(v, w):
  """v_1 * w_1 + ... + v_n * w_n"""
  return sum(v_i * w_i for v_i, w_i in zip(v,w))

def sum_of_squares(v):
  """v_1 * v_1 + ... + v_n * v_n"""
  return dot(v, v)
```

```
import math
def magnitude(v):
  return math.sqrt(sum_of_squares(v))
# math.sqrt is square root function
print(magnitude([1,2,3]))
```

Compute Distance Between Two Vectors

- To compute the distance between two vectors, defined as:

These functions calculate the **Euclidean distance** between two vectors v and w. This is the straight-line distance between two points in space.

Mathematical Formula for Euclidean Distance

For two vectors:

$$v = [v_1, v_2, ..., v_n]$$

$$w = [w_1, w_2, ..., w_n]$$

The squared distance is:

$$\text{squared\_distance}(v, w) = (v_1 - w_1)^2 + (v_2 - w_2)^2 + ... + (v_n - w_n)^2$$

The **Euclidean distance** is obtained by taking the square root:

$$\text{distance}(v, w) = \sqrt{(v_1 - w_1)^2 + (v_2 - w_2)^2 + ... + (v_n - w_n)^2}$$

Compute Distance Between Two Vectors

- To compute the distance between two vectors, defined as:

Example Calculation

Let's compute the distance between **v** = **[3,4]** and **w** = **[1,2]**.

Step 1: Compute (v - w)

$$(3-1, 4-2) = (2, 2)$$

Step 2: Compute Squared Distance

$$(2)^2 + (2)^2 = 4 + 4 = 8$$

Step 3: Compute Euclidean Distance

$$\sqrt{8} \approx 2.83$$

Compute Distance Between Two Vectors

- To compute the distance between two vectors, defined as:

$$\sqrt{(v_1 - w_1)^2 + \ldots + (v_n - w_n)^2}$$

```
def vector_subtract(v, w):
  """subtracts corresponding elements"""
  return [v_i - w_i   for v_i, w_i in zip(v, w)]


def squared_distance(v, w):
  """(v_1 - w_1) ** 2 + ... + (v_n - w_n) ** 2"""


def distance(v, w):
  return math.sqrt(squared_distance(v, w))
print(distance([3,4],[1,2]))
```

Compute Distance Between Two Vectors

- Other way

```
def distance(v, w):
  return magnitude(vector_subtract(v, w))
print(distance([3,4],[1,2]))
```

Matrices

- A matrix is a two-dimensional collection of numbers.
- We will represent matrices as lists of lists, with each inner list having the same size and representing a row of the matrix.
- If ***A*** is a matrix, then ***A[i][j]*** is the element in the ***ith*** row and the ***jth*** column. Per mathematical convention, we will frequently use capital letters to represent matrices. For example:

Vector using List and assert

```
from typing import List
Vector=List[float]
def add(v: Vector, w: Vector) -> Vector:
  """Adds corresponding elements"""
   assert len(v) == len(w), "vectors must be the same length"
  return [v_i+w_i for v_i, w_i in zip(v, w)]
print(add ([5, 7, 9], [4, 5, 6]))
```

Assert keyword

Definition and Usage

The `assert` keyword is used when debugging code.

The `assert` keyword lets you test if a condition in your code returns True, if not, the program will raise an AssertionError.

You can write a message to be written if the code returns False, check the example below.

More Examples

Example

Write a message if the condition is False:

```
x = "welcome"

#if condition returns False, AssertionError is raised:
assert x != "hello", "x should be 'hello'"
```

Matrices

- For example:

```
# Another type alias
Matrix = List[List[float]]

A = [[1, 2, 3],  # A has 2 rows and 3 columns
     [4, 5, 6]]

B = [[1, 2],     # B has 3 rows and 2 columns
     [3, 4],
     [5, 6]]
```

Get Row

- If a matrix has n rows and k columns, we will refer to it as an n × k matrix.
- We can (and sometimes will) think of each row of an n × k matrix as a vector of length k, and each column as a vector of length n:

```
# Get a row from matrix
from typing import List

Vector= List[float]
Matrix = List[List[float]]
def get_row(A: Matrix , i:int)->Vector:
  """ Return ith row"""
  return A[i]
print("Display the first row of matrix")
print(get_row([[1,2,3],[4,5,6]],0))
print("Display the second row of matrix")
print(get_row([[1,2,3],[4,5,6]],1))
```

```
Display the first row of matrix
[1, 2, 3]
Display the second row of matrix
[4, 5, 6]
```

Get Column

```
#Get Column from a Matrix
from typing import List

Vector=List[float]
Matrix = List[List[float]]

def get_col(A:Matrix,j:int) -> Vector:
  return [A_i[j] for A_i in A]

print("Display First Column of Matrix")
print(get_col([[1,2,3],[4,5,6]],0))

print("Display First Column of Matrix")
print(get_col([[1,2,3],[4,5,6]],1))

print("Display First Column of Matrix")
print(get_col([[1,2,3],[4,5,6]],2))
```

```
Display First Column of Matrix
[1, 4]
Display First Column of Matrix
[2, 5]
Display First Column of Matrix
[3, 6]
```

Matrix rows and columns

- Given this list-of-lists representation, the matrix ***A*** has ***len(A)*** rows and ***len(A[0])*** columns, which we consider its shape:

```
# Get shape of a matrix

from typing import Tuple

def shape(A:Matrix) -> Tuple:
  n_row=len(A)
  n_col=len(A[0])
  return n_row,n_col

print(shape([[1,2,3],[4,5,6],[8,9,10]]))
```

```
(3,3)
```

Lambda Functions

```
1  # Here is an example of lambda function that doubles the input value.
2
3  # Program to show the use of lambda functions
4  double = lambda x: x * 2
5
6  print(double(5))
7
8  # In the above program, lambda x: x * 2 is the lambda function.
9  # Here x is the argument and x * 2 is the expression that gets evaluated and returned.
```

✓ 3.9s

```
10
```

Create Matrix

- We'll also want to be able to create a matrix given its shape and a function for generating its elements.
- We can do this using a nested list comprehension

```
from typing import List
from typing import Callable

Vector=List[float]
Matrix = List[List[float]]

def make_matrix(num_rows, num_cols, entry_fn):
"""returns a num_rows num_cols,(i,j)th entry is entry_fn(i, j)"""
  return [[entry_fn(i, j) # given i, create a list
  for j in range(num_cols)] # [entry_fn(i, 0), ... ]
  for i in range(num_rows)] # create one list for each i
print(make_matrix(2,3, lambda i,j:i*j))
```

```
[[0, 0, 0], [0, 1, 2]]
```

Identity Matrix

- you could make a 5 × 5 identity matrix (with 1s on the diagonal and 0s elsewhere) like so:

```
def is_diagonal(i, j):
  """1's on the 'diagonal', 0's everywhere else"""
  return 1 if i == j else 0
identity_matrix = make_matrix(5, 5, is_diagonal)
print(identity_matrix)
```

Statistics

- Statistics refers to the mathematics and techniques with which we understand data.
- Required for large data set .
- Imagine staring at a list of 1 million numbers.
- For that reason we use statistics to distill and communicate relevant features of our data.

Describing a Single Set of Data

- One obvious description of any dataset is simply the data itself:

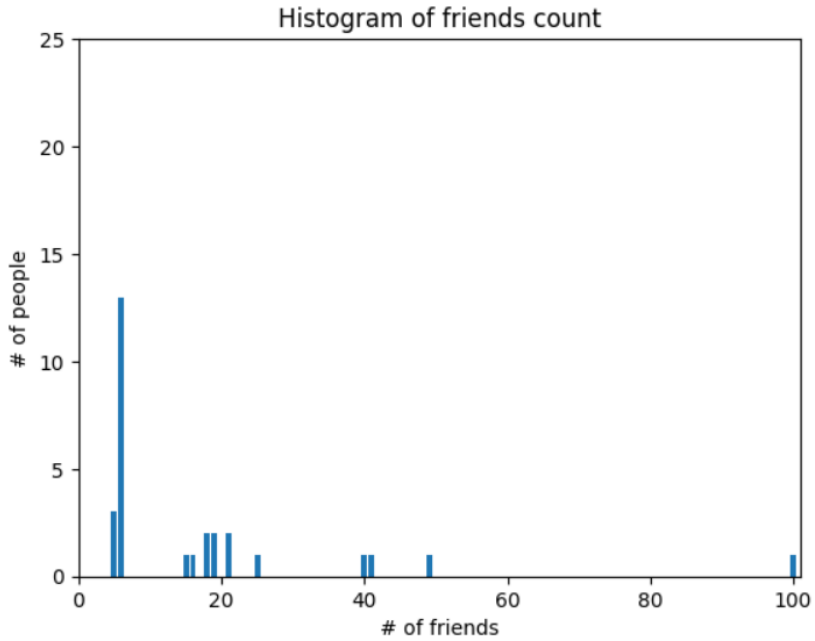
```
num_friends = [100, 49, 41, 40, 25,
               # ... and lots more
              ]
```

- For a small enough dataset, this might even be the best description.
- But for a larger dataset, this is unwieldy and probably opaque. (Imagine staring at a list of 1 million numbers.)
- For that reason, we use statistics to distill and communicate relevant features of our data.

Histogram using Counter and plt.bar()

```
import matplotlib.pyplot as plt
from collections import Counter
num_friends=[100,49,41,40,25,21,21,19,19,18,18,16,15,6,6,6,6,6,6
             ,6,6,6,6,6,6,6,5,5,5]

friends_count=Counter(num_friends)
print(friends_count)
xs=range(101)
ys=[friends_count[x] for x in xs]
plt.bar(xs,ys)
plt.axis([0,101,0,25])
plt.title("Histogram count")
plt.xlabel("# of friends")
plt.ylabel("# of people")
plt.show()
```

```
Counter({6: 13, 5: 3, 21: 2, 19: 2, 18: 2, 100: 1, 49: 1, 41: 1, 40: 1, 25: 1, 16: 1, 15: 1})
```



Largest and Smallest Values

- Unfortunately, this chart is still too difficult to slip into conversations. So generate some statistics. The simplest statistic is the number of data points:
- **Interested in the largest and smallest values**:

```
num_friends=[100,49,41,40,25,21,21,19,19,18,18,16,15,6,6,6,6
,6,6,6,6,6,6,6,6,6,5,5,5]
num_points = len(num_friends)
print(num_points)
largest_value = max(num_friends)
print(largest_value)

smallest_value = min(num_friends)
print(smallest_value)

sorted_values = sorted(num_friends)
print(sorted_values)
```

```
29
100
5
[5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 15, 16, 18, 18, 19, 19, 21, 21, 25, 40, 41, 49, 100]
```

Central Tendencies → Mean

- Central tendency is a statistical concept in data science that describes the center or typical value of a dataset. **It helps summarize a large set of data points by identifying a single representative value.**
- Usually, we'll want some notion of where our data is centered.
- Most commonly we'll use the mean (or average), which is just the sum of the data divided by its count:

```
num_friends=[100,49,41,40,25,21,21,19,19,18,18,16,15,6,6,6,6,6,6
             ,6,6,6,6,6,6,6,5,5,5]
def mean(x):
   return sum(x) / len(x)
mean(num_friends)
```

17.06896551724138

- If you have two data points, the mean is simply the point **halfway** between them.
- As you add more points, the **mean shifts around**, but it always depends on the value of every point.

Median

- We'll also sometimes be interested in the median, which is the **middlemost value** (if the number of data points is odd) or the aver**age of the two middle-most values** (if the number of data points is even).
- For instance, if we have five data points in a **sorted vector x**, the median is x[5 // 2] or x[2].
- If we have six data points, we want the average of x[2] (the third point) and x[3] (the fourth point).
- Notice that—unlike the mean—the **median doesn't fully depend on every value in your data.**
- For example, if you make the largest point larger (or the smallest point smaller), the middle points remain unchanged, which means so does the median.
- The mean is skewed due to an extreme outlier.
- The median provides a fairer estimate

Median

```
num_friends=[100,49,41,40,25,21,21,19,19,18,18,16,15,6,
6,6,6,6,6,6,6,6,6,6,6,5,5,5]
def median(v):
  """finds the 'middle-most' value of v"""
  n = len(v)
  sorted_v = sorted(v)
  midpoint = n // 2
  if n % 2 == 1:
    # if odd, return the middle value
    return sorted_v[midpoint]
  else:
# if even, return the average of the middle values
    lo = midpoint - 1
    hi = midpoint
    return (sorted_v[lo] + sorted_v[hi]) / 2
median(num_friends) # 6.0
```

Percentiles, Quartiles and Quantiles

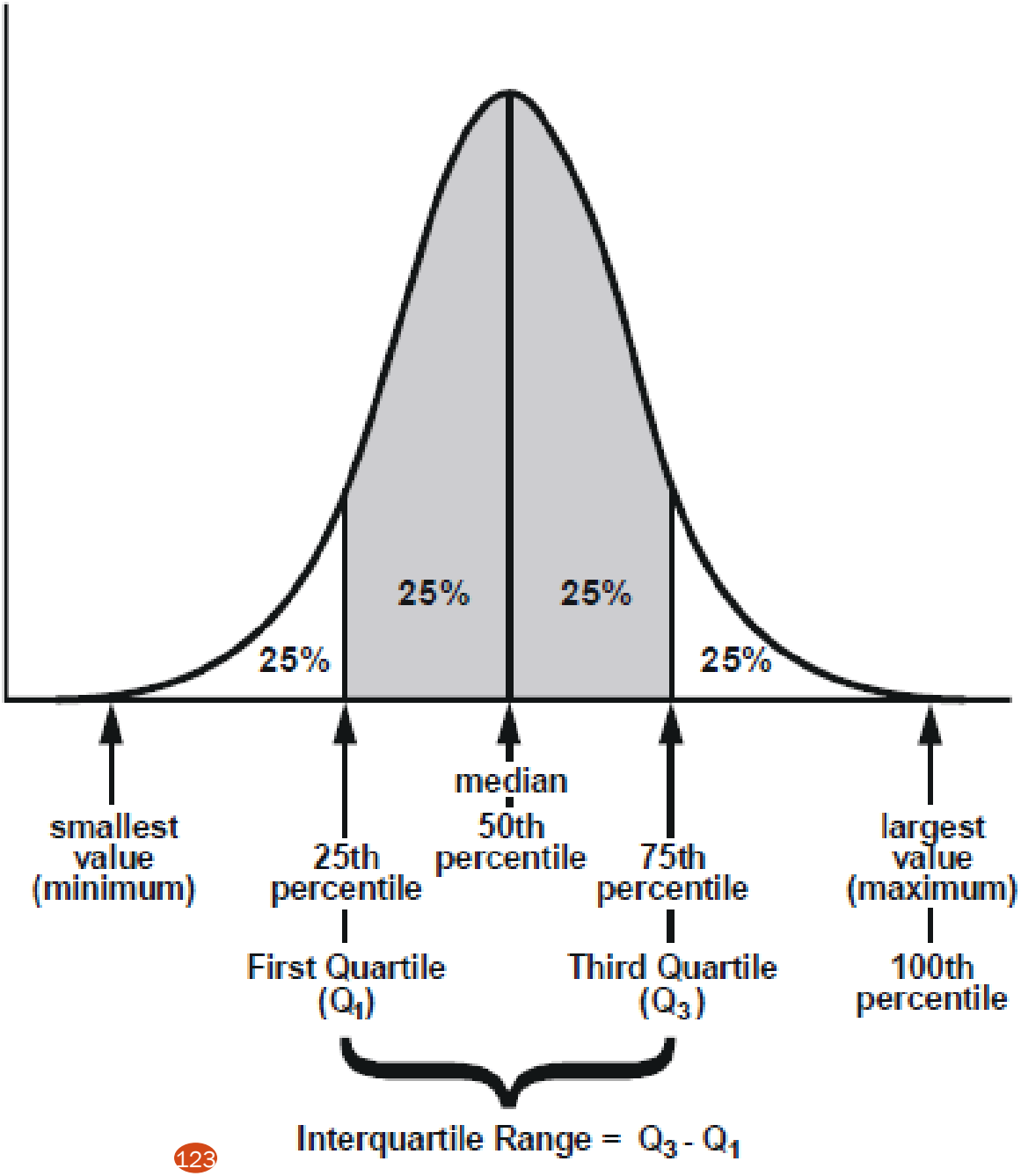
- In **statistics**, **percentiles, quartiles, and quantiles** are used to divide data into different parts to **analyze distributions effectively**. These terms are closely related but differ in the number of divisions they create. Here's a simple definition of each:
- ***Percentiles:*** Range from 0 to 100.
  - **Use Case**: Used in test scores, income distribution, and healthcare analysis.
- ***Quartiles:*** Range from 0 to 4.
- Breakdown:
  - Q1 (25th percentile): The value below which 25% of data falls.
  - Q2 (50th percentile or Median): The middle value of the dataset.
  - Q3 (75th percentile): The value below which 75% of data falls.
  - Q4 (100th percentile): The maximum value in the dataset.
  - Use Case: Used in box plots, income distribution, and performance analysis.

Percentiles, Quartiles and Quantiles

- ***Quantiles:*** Range from any value to any other value.
- Percentiles and quartiles are types of quantiles.
- Some types of quantiles even have specific names, including:
- 4-quantiles are called ***quartiles***.
- 5-quantiles are called ***quintiles***.
- 8-quantiles are called ***octiles***.
- 10-quantiles are called ***deciles***.
- 100-quantiles are called ***percentiles***.

Median

This diagram is called a Box Plot Distribution or Quartile Distribution. It represents the interquartile range (IQR) and the percentile breakdown of a dataset in a normal distribution curve.



Quantile

- A generalization of the median is the quantile, which represents the value under which a certain percentile of the data lies (the median represents the value under which 50% of the data lies):

```
num_friends          =          [70,60,66,49,41,40,25,21,21,19,19,18,18,
16,15,6,6,6,6,6,6,6,6,6,6,6,5,5,5,5]
print(sorted(num_friends))
print("length of list:")
print(len(num_friends))
def quantile(x, p):
  """returns the pth-percentile value in x"""
  p_index = int(p * len(x))
  print("p_index=", p_index)
  return sorted(x)[p_index]
print("10% of Quantile is",quantile(num_friends, 0.10))
print("25% (Q1) Quantile is",quantile(num_friends, 0.25))
print("75% (Q3) Quantile is", quantile(num_friends, 0.75))
print("90% of Quantile is",quantile(num_friends, 0.90))
```

Mode

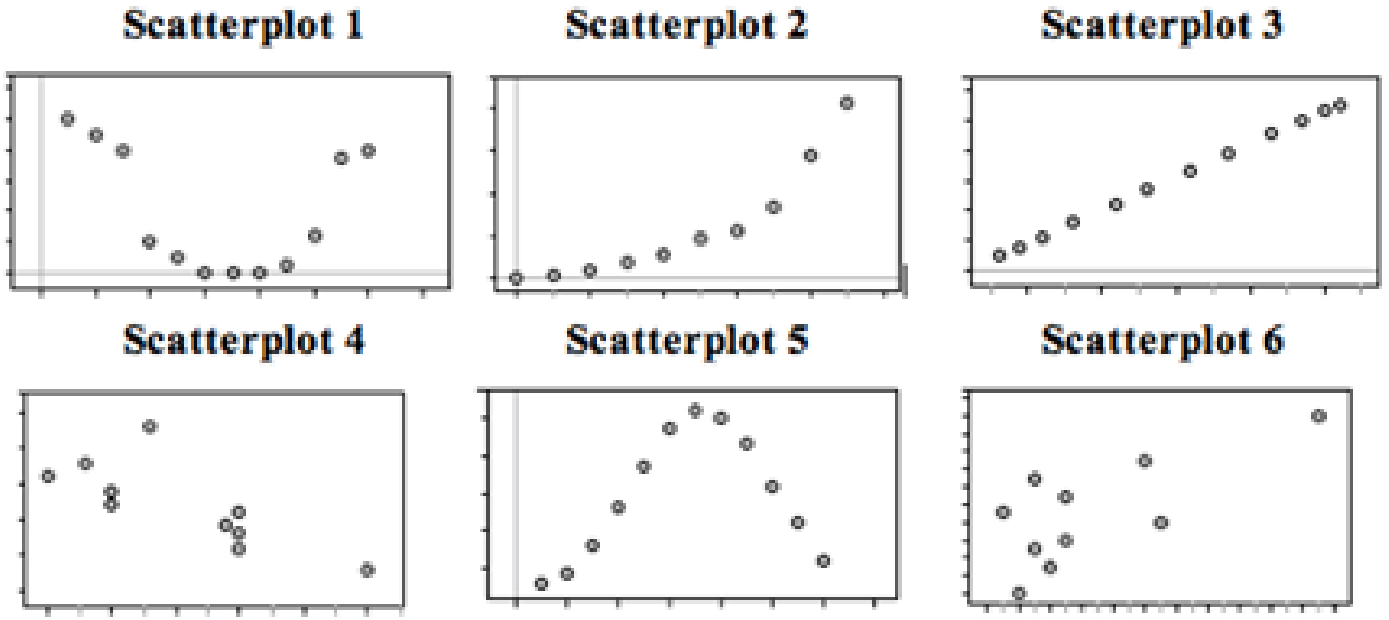
- Mode is defined as a value that occurs most frequently in a dataset.
- For Example, In {6, 9, 3, 6, 6, 5, 2, 3}, the Mode is 6 as it occurs most often.
- The Mode of data set A = {100, 80, 80, 95, 95, 100, 90, 90,100 ,95 } is 80, 90, 95, and 100 because all the four values are repeated twice in the given set.
- The Mode of data set B = {11, 12, 12, 14, 15, 16, 16, 16, 16, 18, 20, 20, 24, 26}
- Here, we get 16 four times, 12 and 20 twice each, and other terms only once. Hence, the Mode for a given set of data is 16.

Mode

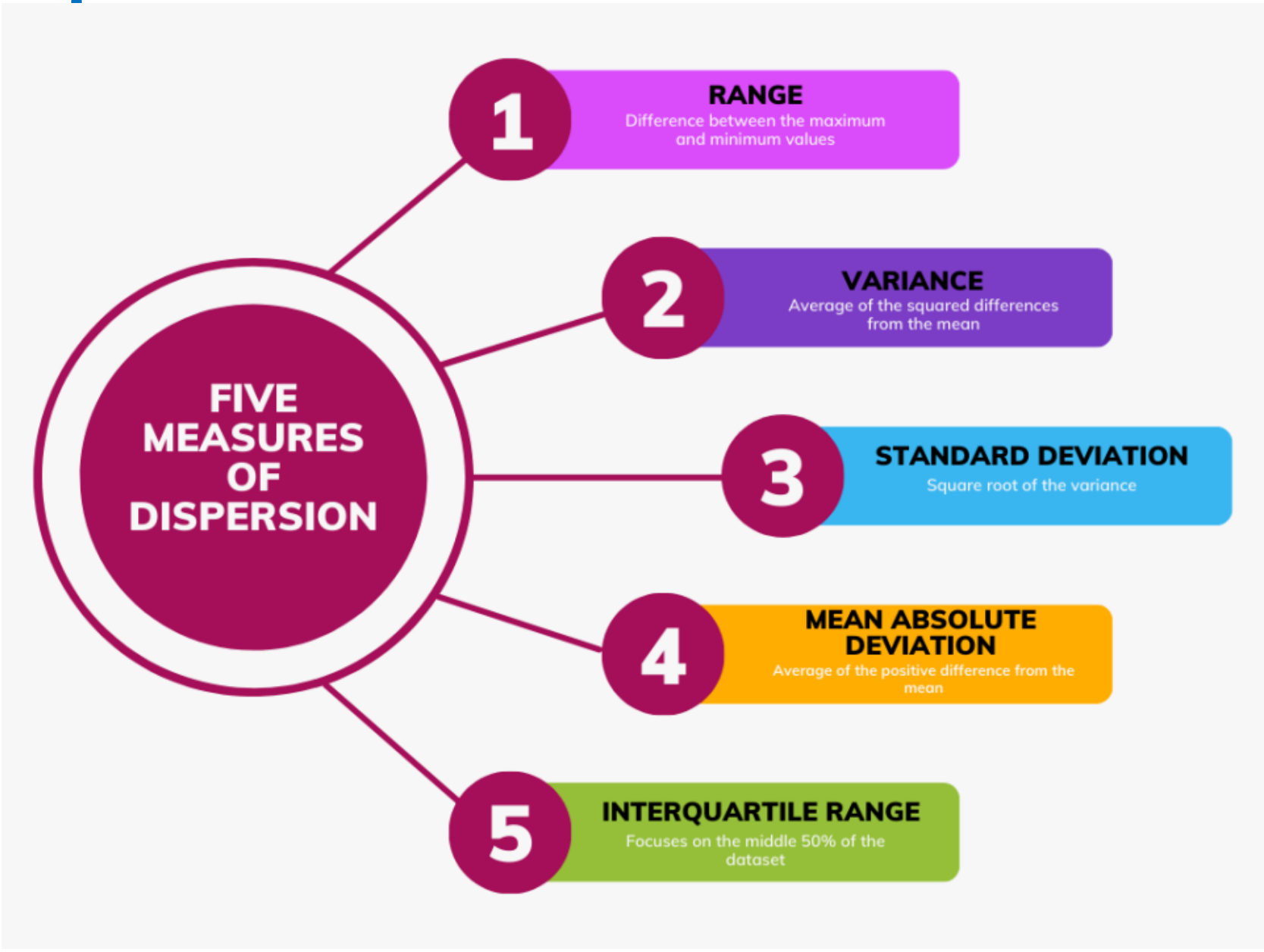
```
from collections import Counter
num_friends =
[100,49,41,40,25,21,21,19,19,18,18,16,15,6,6,6
,5,5,5]

def mode(x):
  """returns a list, might be more than one
mode"""
  counts = Counter(x)
  print(counts)
  print(type(counts))
  max_count = max(counts.values())
  return [x_i for x_i, count in counts.items()
if count == max_count]
mode(num_friends) # 1 and 6
```

Dispersion

- Dispersion refers to measures of how spread out our data is.
- It measures how much the data points differ from the central value (mean, median, or mode).
- Understanding dispersion helps in assessing the reliability, consistency, and distribution of data.



Dispersion


FIVE MEASURES OF DISPERSION
1
RANGE
Difference between the maximum and minimum values
2
VARIANCE
Average of the squared differences from the mean
3
STANDARD DEVIATION
Square root of the variance
4
MEAN ABSOLUTE DEVIATION
Average of the positive difference from the mean
5
INTERQUARTILE RANGE
Focuses on the middle 50% of the dataset

Dispersion

Key Measures of Dispersion

1. **Range**
   - Difference between the maximum and minimum values.
   - Formula:

$$\text{Range} = \max(X) - \min(X)$$

   - **Example:** If dataset = {2, 5, 8, 10}, then **Range = 10 - 2 = 8.**
2. **Variance (σ^2)**
   - Measures the average squared deviation from the mean.
   - Formula (for population variance):

$$\sigma^2 = \frac{\sum(X_i - \mu)^2}{N}$$

   - **Example:** If dataset = {2, 4, 6}, Mean (μ) = 4,

$$\sigma^2 = \frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{3} = \frac{4+0+4}{3} = 2.67$$

Dispersion

3. **Standard Deviation (σ or s)**
   - Square root of variance.
   - Formula:

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

   - It provides dispersion in the same unit as the data.
4. **Interquartile Range (IQR)**
   - Measures the spread of the middle 50% of the data.
   - Formula:

$$IQR = Q3 - Q1$$

   - **Example:** If Q1 = 25 and Q3 = 75, then **IQR = 75 - 25 = 50.**
5. **Mean Absolute Deviation (MAD)**
   - Average absolute difference from the mean.
   - Formula:

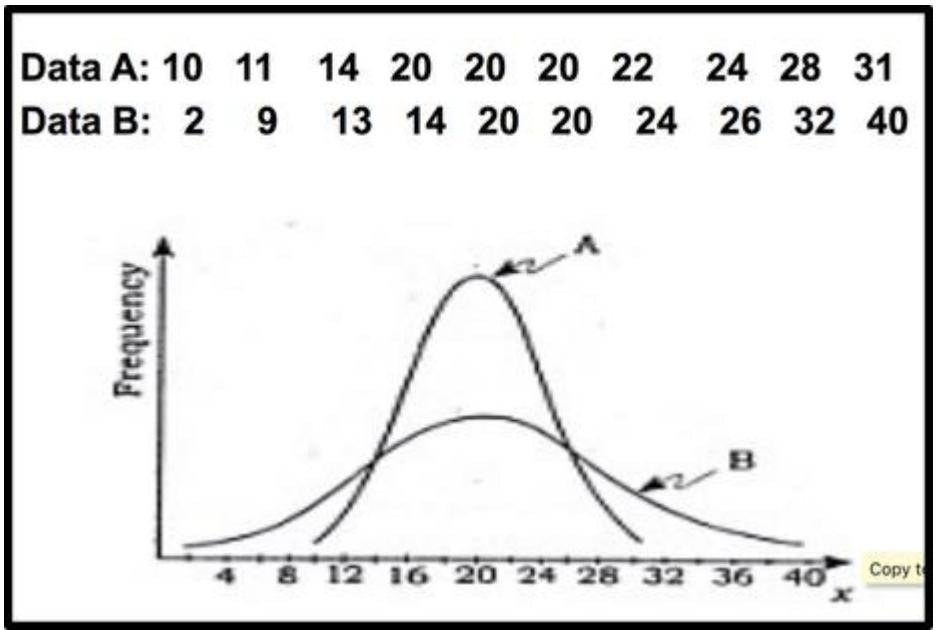
$$MAD = \frac{\sum |X_i - \mu|}{N}$$

Dispersion Scenario: Exam Scores Analysis

- Imagine a teacher has the **math exam scores** of two different classes (Class A and Class B). She wants to analyze which class has more consistent performance.
- **Dataset:**
- **Class A Scores**: [85, 87, 90, 88, 86, 89, 91, 87, 88, 90]
- **Class B Scores**: [75, 95, 60, 80, 100, 55, 85, 90, 50, 105]

Dispersion Scenario: Exam Scores Analysis

- Imagine a teacher has the **math exam scores** of two different classes (Class A and Class B). She wants to analyze which class has more consistent performance.
- **Dataset:**
- **Class A Scores**: [85, 87, 90, 88, 86, 89, 91, 87, 88, 90]
- **Class B Scores**: [75, 95, 60, 80, 100, 55, 85, 90, 50, 105]



Class A Scores: [85, 87, 90, 88, 86, 89, 91, 87, 88, 90]
Class B Scores: [75, 95, 60, 80, 100, 55, 85, 90, 50, 105]

Step 1: Calculate Measures of Dispersion

We will calculate **range**, **variance**, and **standard deviation** for both classes

Class A:

- **Range** = Max - Min = 91 - 85 = **6**
- **Variance** = 3.16
- **Standard Deviation** = $\sqrt{3.16} \approx$ **1.78**

Class B:

- **Range** = Max - Min = 105 - 50 = **55**
- **Variance** = 376.67
- **Standard Deviation** = $\sqrt{376.67} \approx$ **19.4**

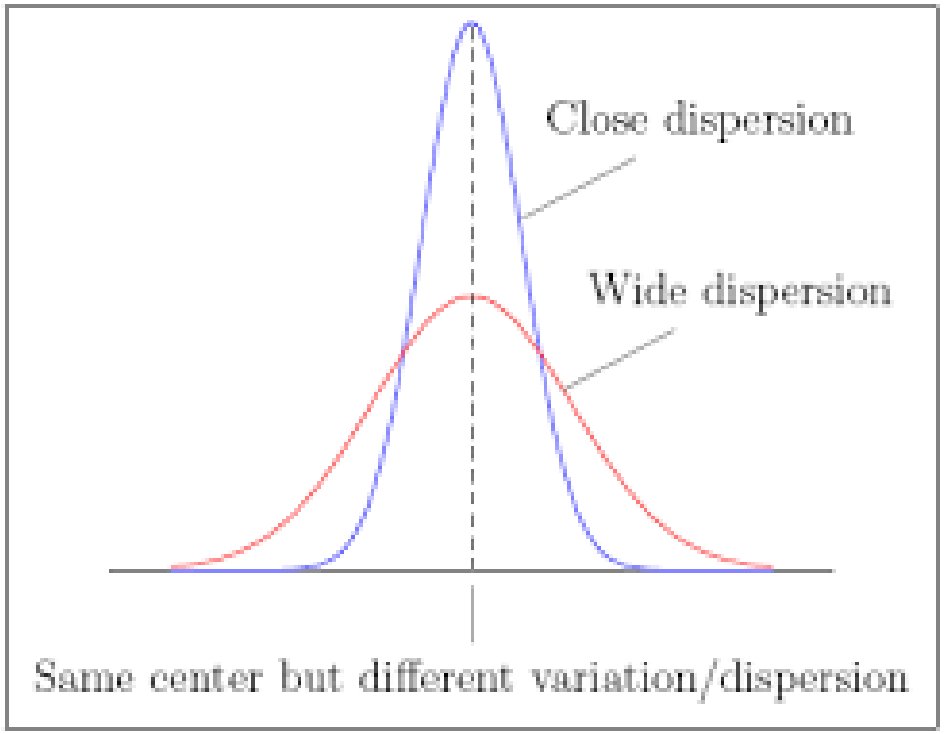
Dispersion Scenario: Exam Scores Analysis

Step 2: Interpret the Results

- **Class A has a lower standard deviation (1.78)** → The scores are **consistent** and close to the mean.
- **Class B has a high standard deviation (19.4)** → The scores are **widely spread** with some students scoring very high and others very low.

Class A has **low dispersion**, meaning students perform **consistently** around the average score.

Class B has **high dispersion**, meaning there is **more variability**, with some students excelling and others struggling.



Dispersion

- Dispersion refers to measures of how spread out our data is.
- Typically they're statistics for which values near zero signify not spread out at all and large values (whatever that means) signify very spread out.
- For instance, a very simple measure is the **range**, which is just the difference between the largest and smallest elements:

```
# "range" already means something in Python, so we'll
use a different name
num_friends =
[100,49,41,40,25,21,21,19,19,18,18,16,15,6,6,6,5,5,5]
def data_range(x):
  return max(x) - min(x)
data_range(num_friends) # 99
```

Variance

- The term variance refers to a statistical measurement of the spread between numbers in a data set.
- More specifically, variance measures how far each number in the set is from the mean and thus from every other number in the set.
- Variance is often depicted by this symbol: $\boldsymbol{\sigma^2}$.

Understanding Variance

- In statistics, variance measures variability from the average or mean.
- It is calculated by taking the differences between each number in the data set and the mean, then squaring the differences to make them positive, and finally dividing the sum of the squares by the number of values in the data set.
- Variance is calculated by using the following formula:

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

where:

$x_i = i^{th}$ data point

$\bar{x} =$ Mean of all data points

$n =$ Number of data points

- The use of n − 1 instead of n in the formula for the sample variance is known as Bessel's correction, which corrects the bias in the estimation of the population variance

Variance

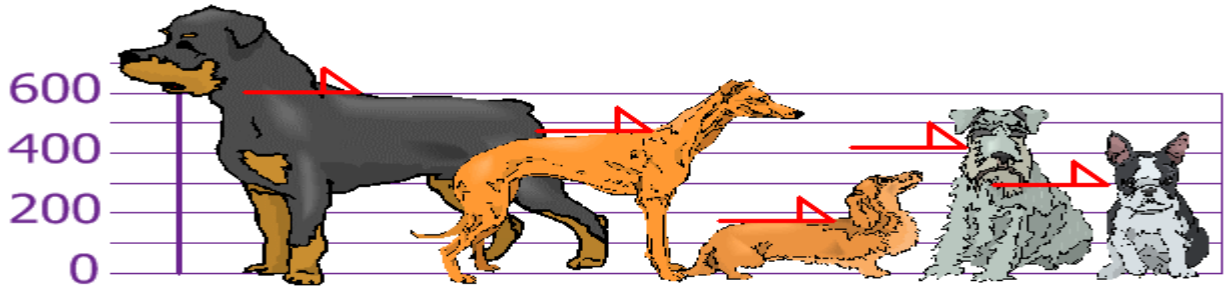
```
# Given dataset
num_friends = [100, 49, 41, 40, 25, 21, 21, 19, 19, 18, 18, 16,
15, 6, 6, 6, 5, 5, 5]

# Step 1: Calculate the mean
mean_value = sum(num_friends) / len(num_friends)

# Step 2: Compute squared differences from the mean
squared_diffs = [(x - mean_value) ** 2 for x in num_friends]

# Step 3: Calculate variance
population_variance = sum(squared_diffs) / len(num_friends)  #
Population variance
sample_variance = sum(squared_diffs) / (len(num_friends) - 1)  #
Sample variance (ddof=1)

# Print results
print(f"Population Variance: {population_variance:.2f}")
print(f"Sample Variance: {sample_variance:.2f}")
```

Standard Deviation

- The Standard Deviation is a measure of how spread out numbers are.
- Its symbol is σ (the greek letter sigma)
- The formula is easy: it is the square root of the Variance. So now you ask, "What is the Variance?

Standard Deviation

- Example
- Measure the heights of dogs (in millimeters):



- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.
- Find out the Mean, the Variance, and the Standard Deviation.
- Your first step is to find the Mean:

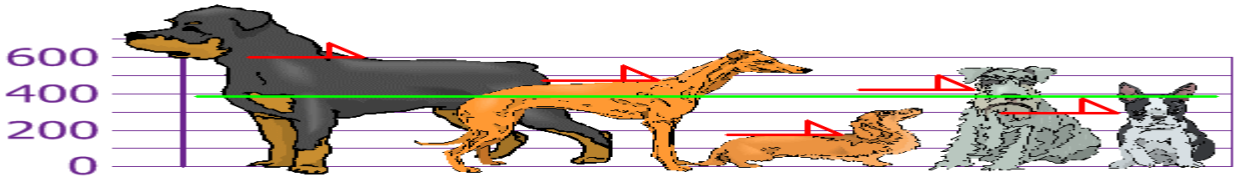
$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

- $$= \frac{1970}{5}$$

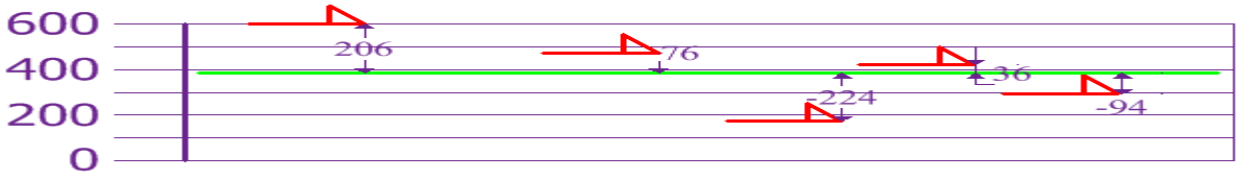
$$= 394$$

Standard Deviation

- so the mean (average) height is 394 mm. Let's plot this on the chart:



- Now we calculate each dog's difference from the Mean:



- To calculate the Variance, take each difference, square it, and then average the result:

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

- So the Variance is 21,704

Standard Deviation

- And the Standard Deviation is just the square root of Variance, so:

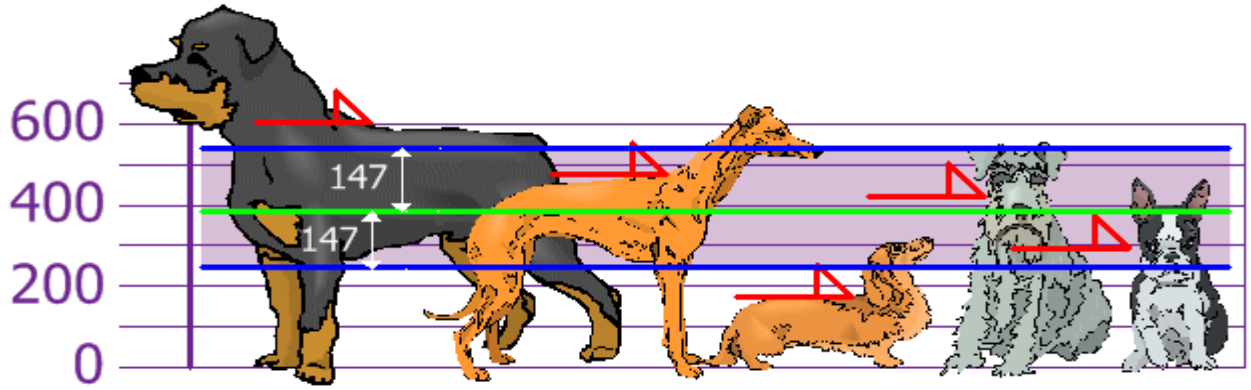
Standard Deviation

$\sigma = \sqrt{21704}$

$= 147.32...$

$= \mathbf{147}$ *(to the nearest mm)*

- Show which heights are within one Standard Deviation (147mm) of the Mean:



- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
- Rottweilers are tall dogs. And Dachshunds are a bit short, right?

Standard Deviation

```
num_friends = [100, 49, 41, 40, 25, 21, 21, 19, 19, 18, 18, 16, 15, 6, 6, 6, 5, 5, 5]
# Step 1: Calculate the mean
mean_value = sum(num_friends) / len(num_friends)
# Step 2: Compute squared differences from the mean
squared_diffs = [(x - mean_value) ** 2 for x in num_friends]
# Step 3: Calculate variance
population_variance = sum(squared_diffs) / len(num_friends)  # Population variance
sample_variance = sum(squared_diffs) / (len(num_friends) - 1)
print(f"Population Variance: {population_variance:.2f}")
print(f"Sample Variance: {sample_variance:.2f}")
std_var=math.sqrt(sample_variance)
print("Standard Deviation",std_var)
```

Difference between Percentile Values

- A more robust alternative computes the difference between the 75th percentile value and the 25th percentile value:

```
num_friends = [100, 49, 41, 40, 25, 21, 21, 19,
19, 18, 18, 16, 15, 6, 6, 6, 5, 5, 5]
def interquartile_range(x):
   return quantile(x, 0.75) - quantile(x, 0.25)
print(interquartile_range(num_friends))
```

- which is quite plainly unaffected by a small number of outliers.

Correlation and Covariance

- Both **correlation** and **covariance** measure the relationship between two variables.

Covariance

- Variance measures how a **single variable** deviates from its mean, covariance measures how **two variables** vary from their means.
- Covariance formula is a statistical formula which is used to assess the **relationship between two variables**.
- It helps to know whether the two variables vary together or change together.
- The covariance is denoted as Cov(X,Y) and the formulas for covariance are given below.

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values.

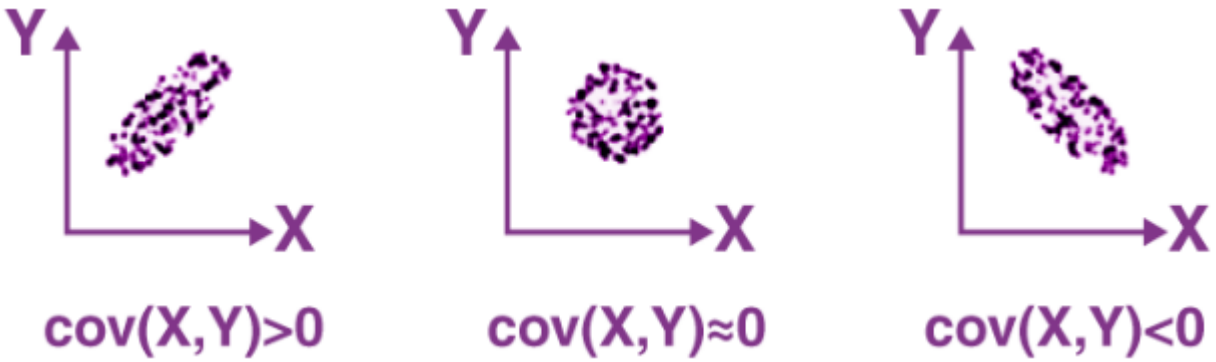
$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

Standard Deviation

```
num_friends = [100, 49, 41, 40, 25, 21, 21, 19, 19, 18, 18, 16, 15, 6, 6, 6, 5, 5, 5]
# Step 1: Calculate the mean
mean_value = sum(num_friends) / len(num_friends)
# Step 2: Compute squared differences from the mean
squared_diffs = [(x - mean_value) ** 2 for x in num_friends]
# Step 3: Calculate variance
population_variance = sum(squared_diffs) / len(num_friends)  # Population variance
sample_variance = sum(squared_diffs) / (len(num_friends) - 1)
print(f"Population Variance: {population_variance:.2f}")
print(f"Sample Variance: {sample_variance:.2f}")
std_var=math.sqrt(sample_variance)
print("Standard Deviation",std_var)
```

Covariance of X and Y

- Below figure shows the covariance of X and Y.



- If ***cov(X, Y)*** is greater than zero, then we can say that the covariance for any two variables is **positive** and both the variables move in the **same direction**.
- If ***cov(X, Y)*** is less than zero, then we can say that the covariance for any two variables is **negative** and both the variables move in the **opposite direction**.
- If ***cov(X, Y)*** is **zero**, then we can say that there is **no relation** between two variables.

Calculate the covariance and interpret the results for the following Temperature and Ice Cream sales application scenario

Given data:

Day	Temperature (X)	Ice Cream Sales (Y)
1	70	300
2	75	350
3	80	400
4	85	450
5	90	500
6	95	550
7	100	600

Step 1: Calculate the means

$$\bar{X} = \frac{2+3+4+5+6+7}{6} = \frac{27}{6} = 4.5$$

$$\bar{Y} = \frac{60+65+70+75+80+85}{6} = \frac{435}{6} = 72.5$$

Calculate the covariance and interpret the results for the following Temperature and Ice Cream sales application scenario

Given data:

Day	Temperature (X)	Ice Cream Sales (Y)
1	70	300
2	75	350
3	80	400
4	85	450
5	90	500
6	95	550
7	100	600

Step 2: Calculate the covariance

$$\text{Cov}(X,Y) = \frac{1}{6}\left[(2-4.5)(60-72.5) + (3-4.5)(65-72.5) + (4-4.5)(70-72.5) + (5-4.5)(75-72.5) + (6-4.5)(80-72.5) + (7-4.5)(85-72.5)\right]$$

Breaking it down:

$$\text{Cov}(X,Y) = \frac{1}{6}\left[(-2.5)(-12.5) + (-1.5)(-7.5) + (-0.5)(-2.5) + (0.5)(2.5) + (1.5)(7.5) + (2.5)(12.5)\right]$$

$$\text{Cov}(X,Y) = \frac{1}{6}\left[31.25 + 11.25 + 1.25 + 1.25 + 11.25 + 31.25\right]$$

$$\text{Cov}(X,Y) = \frac{87.5}{6} = 14.58$$

Result:

$$\text{Covariance} = 14.58$$

Calculate the covariance and interpret the results for the following Temperature and Ice Cream sales application scenario

Interpretation:

- **Positive Covariance:** The covariance is **positive** (1071.43), indicating a **direct positive relationship** between **temperature** and **ice cream sales**.

Calculate the covariance and interpret the results for the following Temperature and Heating Bills Application scenario

Scenario:

- In colder months, heating bills tend to be higher because people use more energy to warm their homes.
- In warmer months, heating bills are lower because people do not need to use the heater as much.

Given data (temperature in °C and heating bill in $):

Month	Temperature (°C)	Heating Bill ($)
1	-5	200
2	0	180
3	5	150
4	10	120
5	15	100

Step 1: Calculate the means

$$\bar{X} = \frac{-5 + 0 + 5 + 10 + 15}{5} = \frac{25}{5} = 5$$

$$\bar{Y} = \frac{200 + 180 + 150 + 120 + 100}{5} = \frac{750}{5} = 150$$

Calculate the covariance and interpret the results for the following Temperature and Heating Bills Application scenario

Month	Temperature (°C)	Heating Bill ($)
1	-5	200
2	0	180
3	5	150
4	10	120
5	15	100

Step 2: Calculate the covariance

$$\text{Cov}(X,Y) = \frac{1}{5}\left[(-5-5)(200-150) + (0-5)(180-150) + (5-5)(150-150) + (10-5)(120-150) + (15-5)(100-150)\right]$$

Breaking it down:

$$\text{Cov}(X,Y) = \frac{1}{5}\left[(-10)(50) + (-5)(30) + (0)(0) + (5)(-30) + (10)(-50)\right]$$

$$\text{Cov}(X,Y) = \frac{1}{5}\left[-500 - 150 + 0 - 150 - 500\right]$$

$$\text{Cov}(X,Y) = \frac{-1300}{5} = -260$$

Interpretation:

- **Negative Covariance**: The covariance is **negative** (-260), indicating that there is an **inverse relationship** between **temperature** and **heating bill**. As the temperature rises (warmer months), the heating bill decreases, which makes sense because people need to use less heating.

Covariance

- Recall that dot sums up the products of corresponding pairs of elements.
- When corresponding elements of x and y are either both above their means or both below their means, a positive number enters the sum.
- When one is above its mean and the other below, a negative number enters the sum.
- Accordingly, a "large" positive covariance means that x tends to be large when y is large and small when y is small.
- A "large" negative covariance means the opposite—that x tends to be small when y is large and vice versa.
- A covariance close to zero means that no such relationship exists.
- Nonetheless, this number can be hard to interpret, for a couple of reasons:
- Its units are the product of the inputs' units (e.g., friend-minutes-per-day), which can be hard to make sense of. (What's a "friend-minute-per-day"?)
- If each user had twice as many friends (but the same number of minutes), the covariance would be twice as large. But in a sense, the variables would be just as interrelated. Said differently, it's hard to say what counts a
- For this reason, it's more common to look at the correlation, which divides out the standard deviations of both variables.

```
# Sample data: Advertising budget and Sales revenue
advertising_budget = [10, 20, 30, 40, 50]  # Budget
sales_revenue_Pos = [200, 400, 600, 800, 1000]  # Revenue
sales_revenue_Neg = [2000,1500, 1200, 800, 100]  # Revenue
def calculate_covariance(x, y)->float:
    # Step 1: Calculate means of x and y
    mean_x = sum(x) / len(x)
    mean_y = sum(y) / len(y)

    # Step 2: Calculate covariance using the formula
    covariance = sum((x[i] - mean_x) * (y[i] - mean_y) for i in
range(len(x))) / len(x)

    return covariance
c=calculate_covariance(advertising_budget, sales_revenue_Pos)
#c=calculate_covariance(advertising_budget, sales_revenue_Neg)
if c== 0:
    print("Looks good!")
elif c < 0:
    print("Negative Covariance")
else:
    print("Positive Covariance")
# Print the result
print(f"Covariance between advertising budget and sales revenue:
{c:.2f}")
```

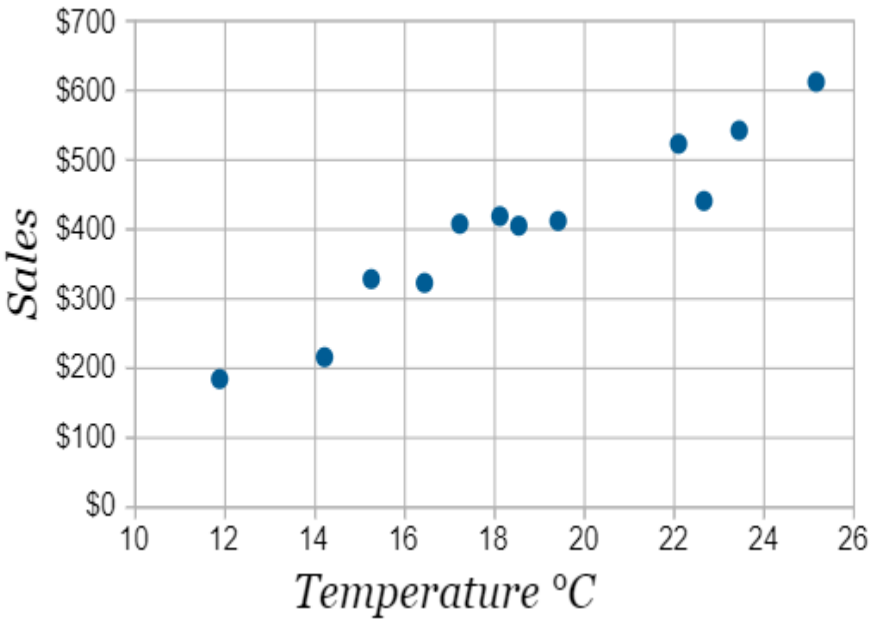
Correlation

- When two sets of data are strongly linked together we say they have a High Correlation.
- Correlation is Positive when the values increase together, and
- Correlation is Negative when one value decreases as the other increases
- A correlation is assumed to be linear (following a line).

Example: Ice Cream Sales

- The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. Here are their figures for the last 12 days:

Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



- We can easily see that warmer weather and higher sales go together. The relationship is good but not perfect.

Correlation

- Covariance is a measure to show the extent to which given two random variables change with respect to each other.
- Correlation is a measure used to describe how strongly the given two random variables are related to each other.

Correlation:

The **correlation coefficient** r is calculated by dividing the covariance by the product of the standard deviations of x and y:

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

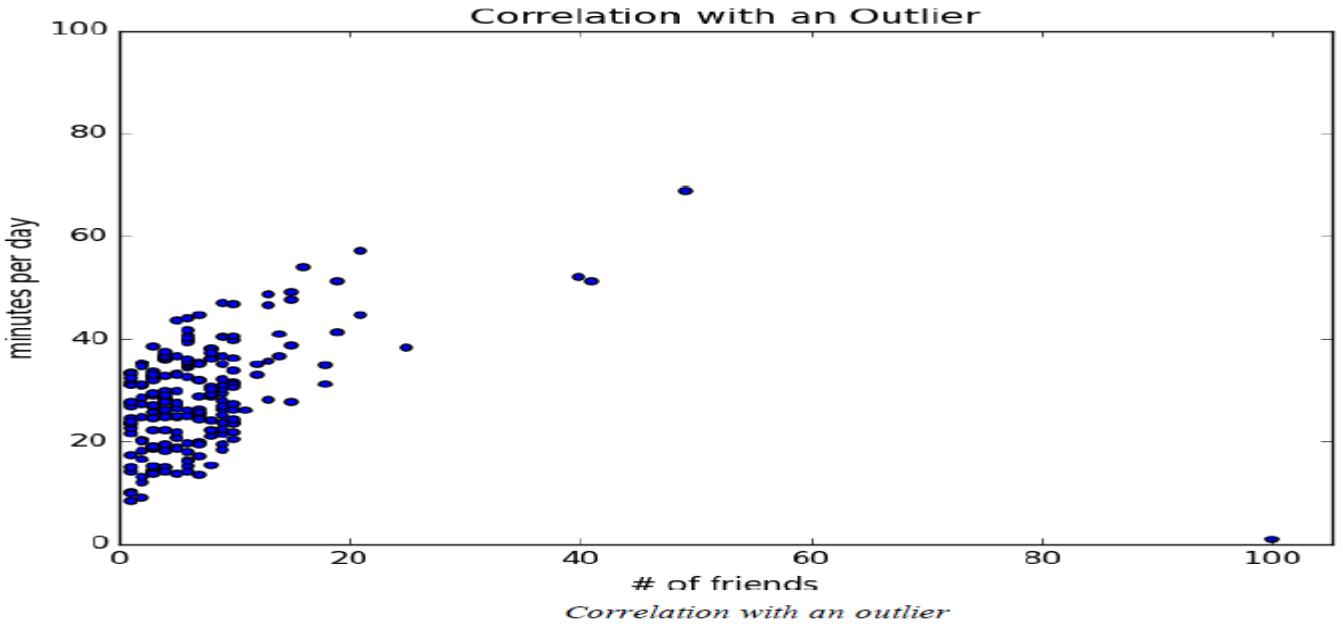
- r ranges between -1 and +1.
  - $r = 1$: Perfect positive correlation (they move together).
  - $r = -1$: Perfect negative correlation (they move oppositely).
  - $r = 0$: No linear relationship.

Correlation

```
import statistics
def calculate_correlation(x, y):
    if len(x) != len(y):
        raise ValueError("x and y must have the same length.")
    mean_x = statistics.mean(x)
    mean_y = statistics.mean(y)
    numerator = sum((x_i - mean_x) * (y_i - mean_y) for x_i, y_i in zip(x, y))
    denominator = ( (sum((x_i - mean_x) ** 2 for x_i in x) * sum((y_i - mean_y) ** 2 for y_i in y))  ** 0.5    )
    return numerator / denominator if denominator != 0 else 0
num_friends = [85,42,41,40,25,21,21,19,19,18,120,28,15,6,6,6,6,6,6,6,6,6]
daily_minutes =
[1,68.7,1.25,52.08,38.36,44.45,57.13,51.4,41.42,31.30,35,59,39.43,14.18,35.24,40.13,41.32,35.45,36.07,43.44,23.5
5,24.6]
correlation = calculate_correlation(num_friends, daily_minutes)
print(f"Pearson correlation coefficient: {correlation:.2f}")
if correlation <=0:
  print("No correlation!")
elif 0.1 < correlation < 0.25:
  print("Weak Correlation")
elif 0.25 < correlation < 0.5:
  print("Moderate Correlation")
elif 0.5 < correlation < 0.75:
  print("Strong Correlation")
elif 0.75 < correlation < 1:
  print("Very Strong Correlation")
else:
  print("Invalid")
```

Correlation

- The correlation is unitless and always lies between –1 (perfect anticorrelation) and 1 (perfect correlation).
- A number like 0.25 represents a relatively weak positive correlation.



Correlation with an outlier

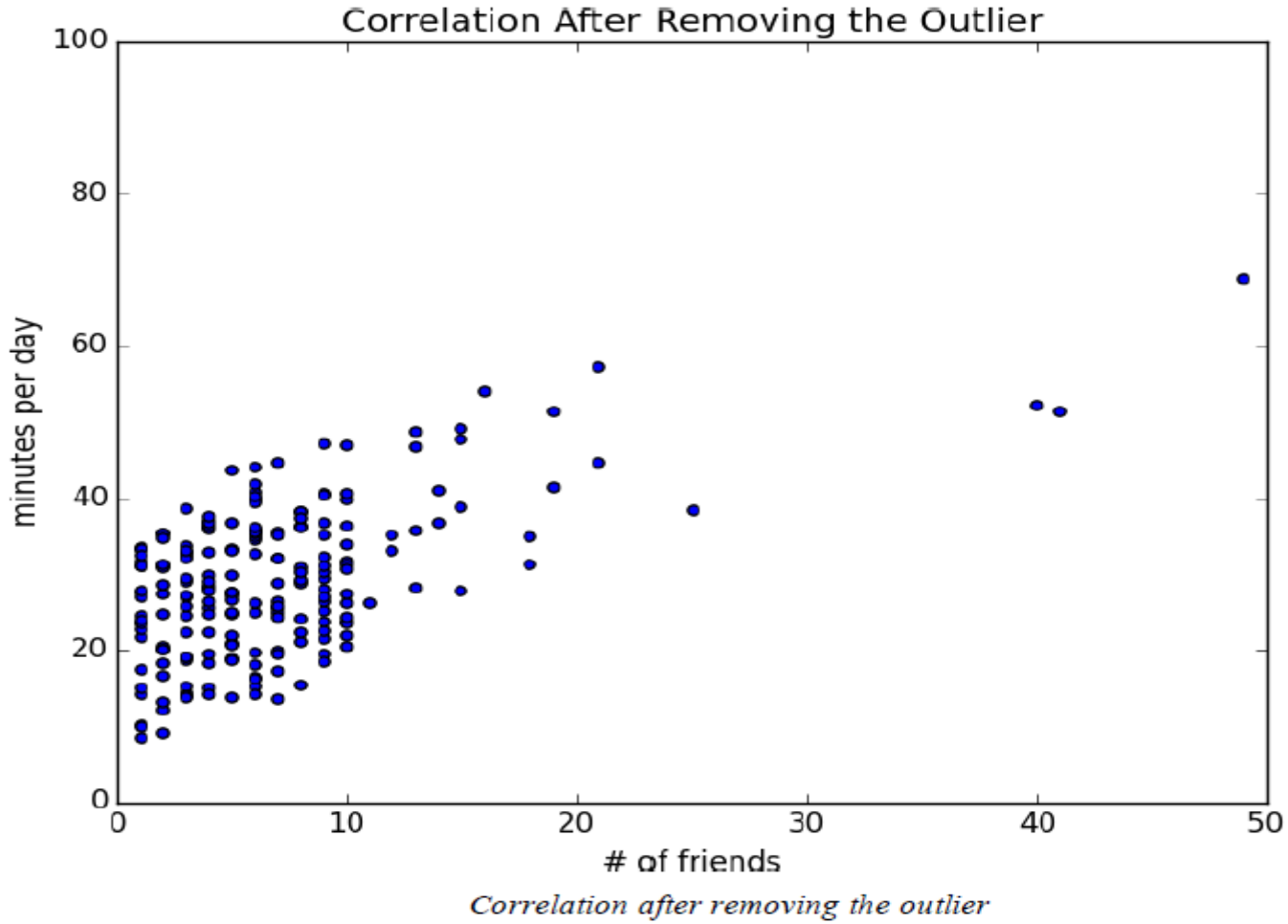
- Let's examine our data. The person with 100 friends (who spends only 1 minute per day on the site) is a huge outlier, and correlation can be very sensitive to outliers.
- What happens if we ignore him?

Correlation

```
import statistics
num_friends = [60,42,41,40,25,21,21,19,19,18,120,28,15,6,6,6,6,6,6,6,6,6]
outlier = num_friends.index(60) # index of outlier
print(outlier)
def calculate_correlation(x, y):
    """Calculates the Pearson correlation coefficient between two lists x and y."""
    if len(x) != len(y):
        raise ValueError("x and y must have the same length.")
    mean_x = statistics.mean(x)
    mean_y = statistics.mean(y)
    numerator = sum((x_i - mean_x) * (y_i - mean_y) for x_i, y_i in zip(x, y))
    denominator = ((sum((x_i - mean_x) ** 2 for x_i in x) * sum((y_i - mean_y) ** 2 for y_i in
y)) ** 0.5    )
    return numerator / denominator if denominator != 0 else 0
daily_minutes =
[1,68.7,1.25,52.08,38.36,44.45,57.13,51.4,41.42,31.30,35,59,39.43,14.18,35.24,40.13,41.32,35.4
5,36.07,43.44,23.55,24.6]
daily_minutes_good = [x for i, x in enumerate(daily_minutes) if i != outlier]
num_friends_good = [x for i, x in enumerate(daily_minutes) if i != outlier]
print(daily_minutes_good)
print(calculate_correlation(num_friends_good, daily_minutes_good))
```

Correlation without Outlier

- Without the outlier, there is a much stronger correlation.



Correlation after removing the outlier

Simpson's Paradox

- One common surprise when analyzing data is Simpson's paradox, in which correlations can be misleading when confounding variables (is an external variable that affects both the independent variable (the one you manipulate) and the dependent variable (the one you measure), creating a misleading or biased relationship between them.)are ignored.
- For example, imagine that you can identify all of your members as either East Coast data scientists or West Coast data scientists. You decide to examine which coast's data scientists are friendlier:

Coast	# of members	Avg. # of friends
West Coast	101	8.2
East Coast	103	6.5

- It certainly looks like the West Coast data scientists are friendlier than the East Coast data scientists.

Simpson's Paradox

- If you look only at people with PhDs, the East Coast data scientists have more friends on average. And if you look only at people without PhDs, the East Coast data scientists also have more friends on average!

Coast	Degree	# of members	Avg. # of friends
West Coast	PhD	35	3.1
East Coast	PhD	70	3.2
West Coast	No PhD	66	10.9
East Coast	No PhD	33	13.4

Simpson’s Paradox

- **Simpson's Paradox** occurs when a trend appears in different groups of data but disappears or reverses when the data is combined. This paradox demonstrates how misleading conclusions can arise if we fail to consider underlying subgroups in a dataset.
- In data science, Simpson’s Paradox is a critical reminder of the importance of analyzing data at multiple levels and understanding potential confounding factors that might affect the results.

Correlation and Causation

- **Causation** implies that one event directly affects another. It answers the question:
- **Does a change in one variable cause a change in another variable?**
- For causation to be established, you need to rule out confounding variables and establish a direct cause-and-effect relationship.
- So, ***"correlation is not causation,"***
- Nonetheless, this is an important point—if x and y are strongly correlated, that might mean that x causes y, that y causes x, that each causes the other, that some third factor causes both, or nothing at all.
- Consider the relationship between ***num_friends*** and ***daily_minutes***. It's possible that having more friends on the site causes DataSciencester users to spend more time on the site.
- **First might** be the case if each friend posts a certain amount of content each day, which means that the more friends you have, the more time it takes to stay current with their updates.
- **Second**, its possible that the more time users spend arguing in the DataSciencester forums, the more they encounter and befriend likeminded people.
- That is, spending more time on the site causes users to have more friends.
- A **third possibility** is that the users who are most passionate about data science spend more time on the site (because they find it more interesting) and more actively collect data science friends (because they don't want to associate with anyone else).

Probability

- Probability is a way of quantifying the uncertainty associated with events chosen from some universe of events.
- The universe consists of all possible outcomes.
- And any subset of these outcomes is an event; for example, “the die rolls a 1” or “the die rolls an even number.”
- Notationally, we write P(E) to mean “the probability of the event E.”
- We’ll use probability theory to build and evaluate models.

Dependence and Independence

- Two events E and F are dependent if knowing something about whether E happens gives us information about whether F happens (and vice versa).
- Otherwise, they are independent.
- For instance, if we flip a fair coin twice, knowing whether the first flip is heads gives us no information about whether the second flip is heads or tail. These events are independent.
- On the other hand, knowing whether the first flip is heads certainly gives us information about whether both flips are heads or not. These two events are dependent.
- Mathematically, we say that two events E and F are independent if the probability that they both happen is the product of the probabilities that each one happens:

$$P(E, F) = P(E)P(F)$$

Conditional Probability

Conditional probability measures the probability of an event occurring, given that another event has already occurred. It is a fundamental concept in probability theory and statistics.

The conditional probability of event A, given that event B has occurred, is denoted by $P(A \mid B)$ and is defined by the formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{if} \quad P(B) > 0$$

Where:

- $P(A \mid B)$: Conditional probability of A given B
- $P(A \cap B)$: Joint probability that both A and B occur
- $P(B)$: Probability that event B occurs

Conditional Probability

Example:

Suppose you draw a card from a standard deck of 52 playing cards.

Let:

- Event A: The card drawn is a King
- Event B: The card drawn is a face card (King, Queen, or Jack)

In a standard deck:

- $P(A) = \frac{4}{52}$ (since there are 4 Kings)
- $P(B) = \frac{12}{52}$ (since there are 12 face cards: 4 Kings, 4 Queens, and 4 Jacks)
- $P(A \cap B) = \frac{4}{52}$ (since all 4 Kings are also face cards)

Now, calculate $P(A \mid B)$:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{4}{52}}{\frac{12}{52}} = \frac{4}{12} = \frac{1}{3}$$

So, given that the card drawn is a face card, the probability that it is a King is $\frac{1}{3}$.

Bayes Probability

The conditional probability of E given F, denoted $P(E|F)$, means the probability that event E happens, knowing that event F has already happened.

From the definition of conditional probability:

$$P(E|F) = \frac{P(E,F)}{P(F)}$$

Here:

- $P(E,F)$: The probability that both events E and F occur.
- $P(F)$: The probability that event F occurs.

Since $P(E,F)$ (joint probability of E and F) can also be written as:

$$P(E,F) = P(F|E) \cdot P(E)$$

This gives:

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{P(F)}$$

This is the first key step toward Bayes' Theorem.

Bayes Probability

Now, we need to break down $P(F)$ (the probability that event F happens).

The event F can happen in two ways:

1. F happens *and* E happens (denoted $P(F, E)$).
2. F happens *and* E does *not* happen (denoted $P(F, \neg E)$).

So, we can write:

$$P(F) = P(F, E) + P(F, \neg E)$$

Using conditional probability, we can express these as:

$$P(F) = P(F|E) \cdot P(E) + P(F|\neg E) \cdot P(\neg E)$$

Here:

- $P(F|E)$: Probability that F happens given E.
- $P(F|\neg E)$: Probability that F happens given that E *does not* happen.
- $P(\neg E)$: Probability that E does not happen ($P(\neg E) = 1 - P(E)$).

Bayes Probability

Now, substitute $P(F)$ back into the conditional probability formula:

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{P(F|E) \cdot P(E) + P(F|\neg E) \cdot P(\neg E)}$$

This is **Bayes' Theorem**, and it allows us to "reverse" conditional probabilities (i.e., switch E and F).

Bayes Probability

Simple Example: Medical Test for a Rare Disease

Imagine there is a rare disease that affects 1 out of every 10,000 people. You take a test that says you have the disease, but how accurate is that test really? Let's break it down using Bayes' Theorem.

- Event D: You have the disease.
- Event T: The test says you are positive for the disease.

We want to calculate: **What is the chance you actually have the disease if the test says you are positive?** This is $P(D|T)$, the probability of having the disease, given a positive test result.

Given Information:

1. $P(D)$: Probability that a random person has the disease = 1/10,000 = 0.0001.
2. $P(\neg D)$: Probability that a random person does *not* have the disease = 0.9999.
3. $P(T|D)$: Probability that the test is positive if you have the disease = 0.99.
4. $P(T|\neg D)$: Probability that the test is positive even if you don't have the disease (false positive) = 0.01.

Bayes Probability

Applying Bayes' Theorem:

Now, plug these values into the formula:

$$P(D|T) = \frac{P(T|D) \cdot P(D)}{P(T|D) \cdot P(D) + P(T|\neg D) \cdot P(\neg D)}$$

Substitute the numbers:

$$P(D|T) = \frac{(0.99) \cdot (0.0001)}{(0.99) \cdot (0.0001) + (0.01) \cdot (0.9999)}$$

$$P(D|T) = \frac{0.000099}{0.000099 + 0.009999}$$

$$P(D|T) = \frac{0.000099}{0.010098}$$

$$P(D|T) \approx 0.0098 \text{ or } 0.98\%$$

Conclusion:

Even though the test is 99% accurate, because the disease is so rare, the chance that you actually have the disease after testing positive is **only about 1%**. This is because false positives happen more frequently than true positives when dealing with rare conditions.

Random Variables

- A random variable is a variable whose possible values have an associated **probability distribution**. A very simple random variable equals 1 if a coin flip turns up heads and 0 if the flip turns up tails.
- A more complicated one might measure the number of heads you observe when **flipping a coin 10 times** or a value picked from range(10) where each number is equally likely.
- The associated distribution gives the probabilities that the variable realizes each of its possible values. The coin flip variable equals 0 with probability 0.5 and 1 with probability 0.5.
- The range(10) variable has a distribution that assigns probability 0.1 to each of the numbers from 0 to 9.

Random Variables

- This program simulates tossing two coins 10 times, each time recording the result of the tosses using random library. The goal is to keep track number of the following conditions: Both Coins Show Heads, First Coin Shows Heads, At Least One Coin Shows Heads.

```
import random
def coin_toss():
    return random.choice(["heads", "tails"])
both_heads = 0
at_least_one_head = 0
first_is_heads = 0
random.seed(0)
for _ in range(10):  # Simulate 10 tosses of two coins
    first_coin = coin_toss()
    second_coin = coin_toss()
    print(first_coin,second_coin)

    if first_coin == "heads":
        first_is_heads += 1
    if first_coin == "heads" and second_coin == "heads":
        both_heads += 1
    if first_coin == "heads" or second_coin == "heads":
        at_least_one_head += 1
print(f"Both are heads: {both_heads}")
print(f"First is heads: {first_is_heads}")
print(f"At least one is heads: {at_least_one_head}")
```

Continuous Distributions

- A coin flip corresponds to a **discrete distribution**—one that associates positive probability with discrete outcomes(whole numbers).
- **Continuous Distribution** is distributions across a continuous outcomes. (real numbers.)
- For example, the **uniform distribution** puts equal weight on all the numbers between 0 and 1.

```
def uniform_pdf(x):
  return 1 if x >= 0 and x < 1
else 0
print(uniform_pdf(-0.3))
```

Continuous Distributions

- There are infinitely many numbers between 0 and 1,.
- Represent a continuous distribution with a **probability density function (PDF)** such that the probability of seeing a value in a certain interval equals the integral of the density function over the interval.

1. Cumulative Distribution Function (CDF), gives the probability that a ***random variable is less than or equal to a certain value***.

```
def uniform_cdf(x):
  "returns the probability that a uniform random variable is <= x"
  if x < 0: return 0 # uniform random is never less than 0
  elif x < 1: return x # e.g. P(X <= 0.4) = 0.4
  else: return 1 # uniform random is always less than 1
print(uniform_cdf(0.5))
```

Continuous Distributions

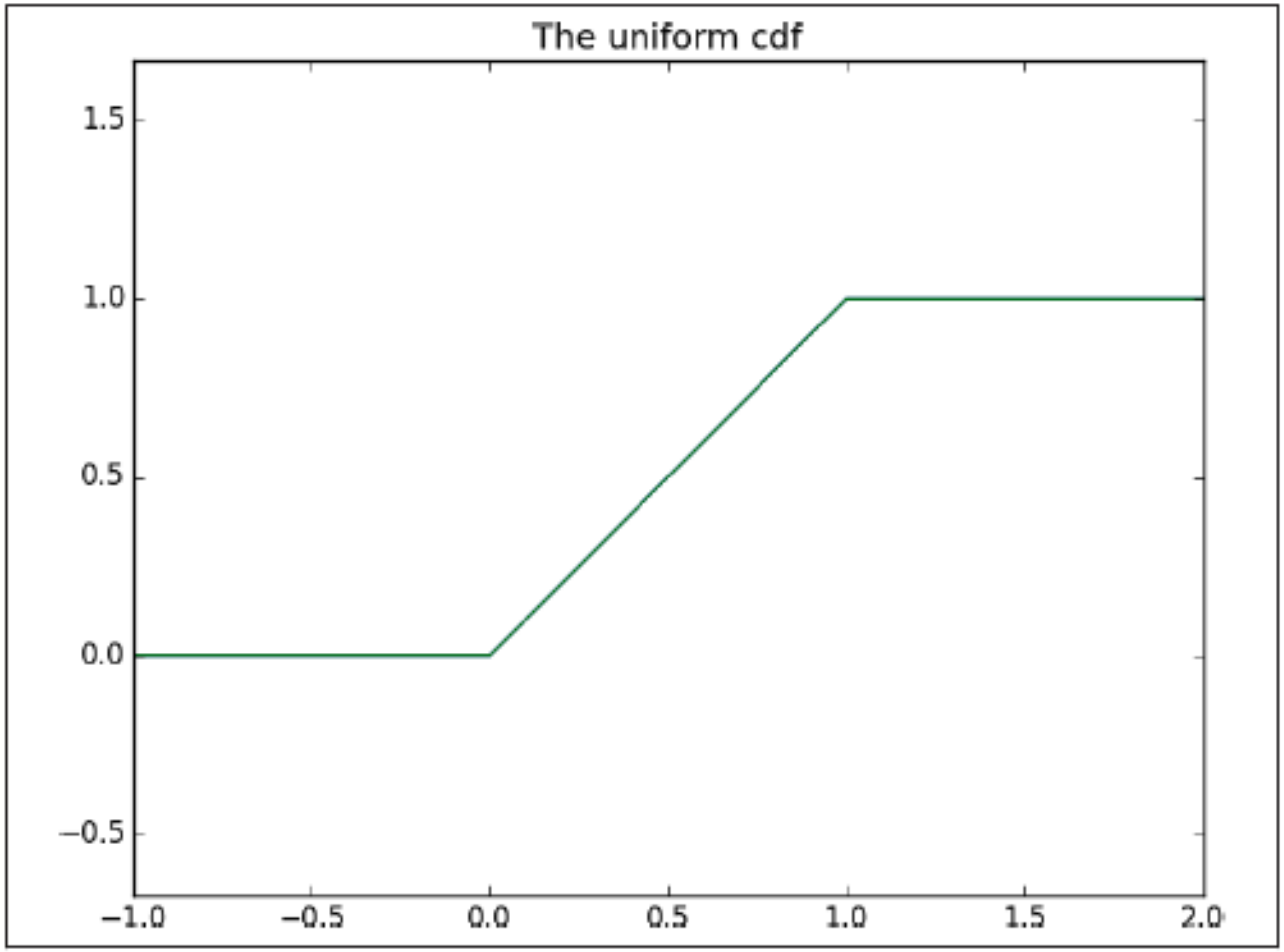


Figure 6-1. The uniform cdf

PDF for Normal Distribution

- The normal distribution (also called the Gaussian distribution) is a continuous probability distribution.
- The **PDF** describes the probability distribution of a continuous random variable at any specific value x in the interval.
- The normal distribution is the classic **bell curve–shaped** distribution and is completely determined by two parameters: its mean ***μ (mu)*** and its standard deviation ***σ (sigma).***
- The **mean** indicates where the bell is centered, and the **standard deviation** how "wide" it is.
- It has the PDF:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The Normal Distribution

- It has the PDF:

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

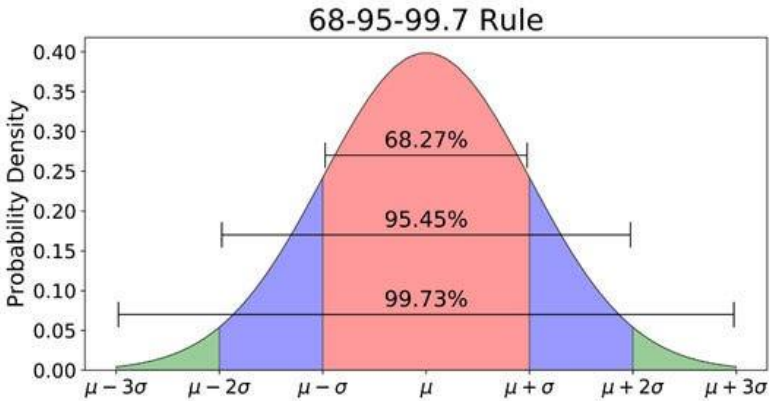
- Where:
- Probability density function value at
- Mean (average) of the distribution
- Standard deviation (a measure of spread or dispersion)

- The term $\frac{1}{\sigma\sqrt{2\pi}}$ is a **normalizing constant** that ensures the total area under the curve equals 1.
- The term $\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ controls the shape of the curve:
  - If x is close to μ, the exponent is small, so the PDF value is high.
  - As x moves away from μ, the exponent becomes larger and the PDF value approaches zero.

The Normal Distribution

Key Characteristics of Normal Distribution:

- **Symmetry**: The distribution is symmetric around the mean .
- **Bell-Shaped Curve**: The highest point is at , and the curve tapers off as you move away from the mean.
  - A small σ makes the curve narrow and tall.
  - A large σ makes the curve wide and flat.
- **68-95-99.7 Rule (Empirical Rule):**
  - 68% of the data lies within 1 standard deviation ().
  - 95% of the data lies within 2 standard deviations ().
  - 99.7% of the data lies within 3 standard deviations ().Variance (square of the standard deviation)
- **PDF** describes the shape of a distribution and shows where values are most likely to occur.



The Normal Distribution

Need for Normal Distribution in Data Science:

1.Modeling Real-World Phenomena:

Many real-world processes follow normal distribution (e.g., height, IQ scores, measurement errors).

2.Anomaly Detection:

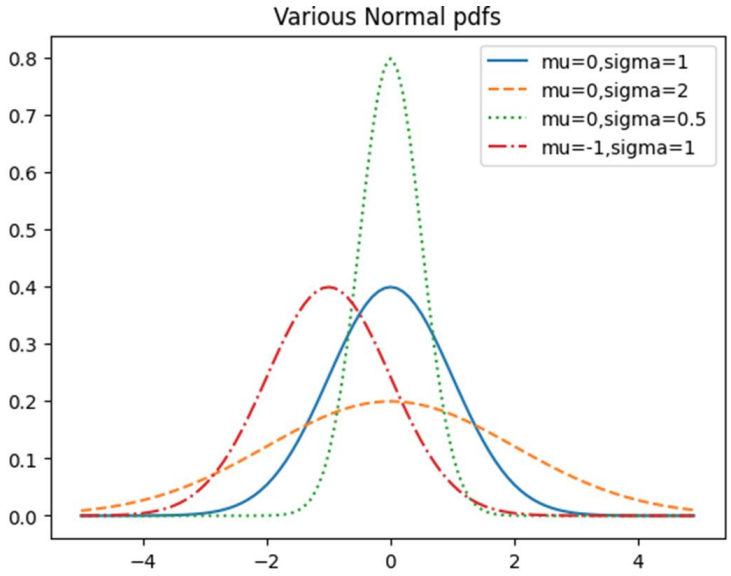
Deviations from the expected normal distribution can indicate anomalies (e.g., fraud detection).

3.Error and Residual Analysis:

In regression models, normal distribution of residuals is a key assumption for accurate model performance.

The Normal Distribution

```
import math
def normal_pdf(x, mu=0, sigma=1):
  sqrt_two_pi = math.sqrt(2 * math.pi)
  return (math.exp(-(x-mu) ** 2 / 2 / sigma ** 2) / (sqrt_two_pi * sigma))
xs = [x / 10.0 for x in range(-50, 50)]
print(xs)
plt.plot(xs,[normal_pdf(x,sigma=1) for x in xs],'-',label='mu=0,sigma=1')
plt.plot(xs,[normal_pdf(x,sigma=2) for x in xs],'--',label='mu=0,sigma=2')
plt.plot(xs,[normal_pdf(x,sigma=0.5) for x in xs],':',label='mu=0,sigma=0.5')
plt.plot(xs,[normal_pdf(x,mu=-1) for x in xs],'-.',label='mu=-1,sigma=1')
plt.legend()
plt.title("Various Normal pdfs")
plt.show()
```



The CDF for Normal Distribution

For a **normal distribution**, the CDF expression.

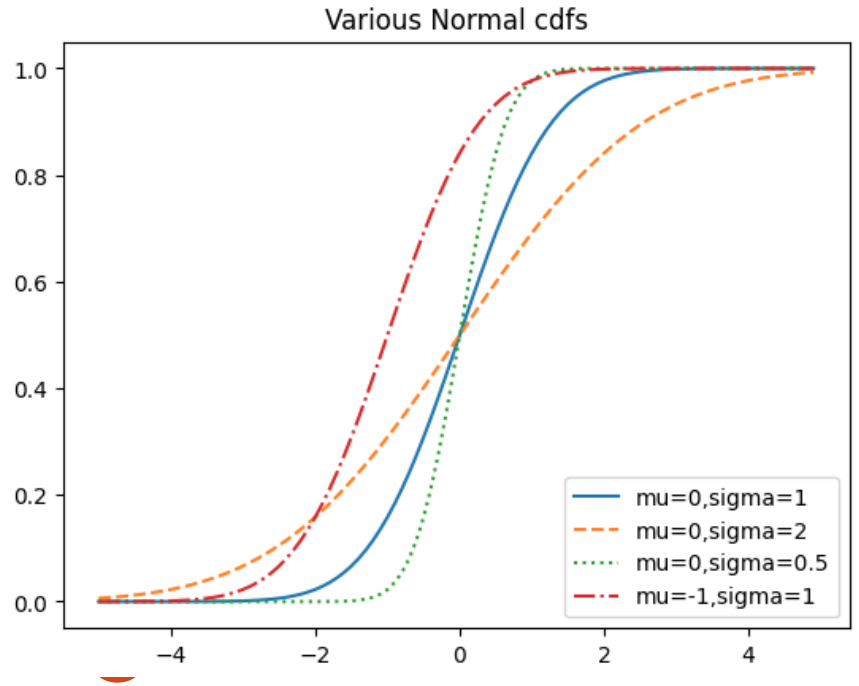
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

- The CDF for the normal distribution cannot be written in an "elementary" manner, but we can write it using Python's ***math.erf*** error function:

CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

The Normal Distribution

```
from matplotlib import pyplot as plt
import math
def normal_cdf(x, mu=0,sigma=1):
  return (1 + math.erf((x - mu) / math.sqrt(2) / sigma)) / 2
xs = [x / 10.0 for x in range(-50, 50)]
plt.plot(xs,[normal_cdf(x,sigma=1) for x in xs],'-',label='mu=0,sigma=1')
plt.plot(xs,[normal_cdf(x,sigma=2) for x in xs],'--',label='mu=0,sigma=2')
plt.plot(xs,[normal_cdf(x,sigma=0.5) for x in xs],':',label='mu=0,sigma=0.5')
plt.plot(xs,[normal_cdf(x,mu=-1) for x in xs],'-.',label='mu=-1,sigma=1')
plt.legend(loc=4) # bottom right
plt.title("Various Normal cdfs")
plt.show()
```



The Normal Distribution

•**PDF (Probability Density Function)**:
•Shows the relative likelihood of different delivery times.
•It does not give cumulative probabilities but instead shows how dense the distribution is at each point.
•**CDF (Cumulative Distribution Function)**:
•Used to calculate the cumulative probability that the delivery time is less than or equal to a certain value.
•Useful for finding probabilities over a range of values (like delivery between 30 and 40 minutes).

The Normal Distribution

- Sometimes we'll need to invert ***normal_cdf*** to find the value corresponding to a specified probability.
- There's no simple way to compute its inverse, but ***normal_cdf*** is continuous and strictly increasing, so we can use a binary search

```
def inverse_normal_cdf(p: float, mu: float = 0, sigma: float = 1, tolerance: float = 0.00001) -> float:
    """Find approximate inverse using binary search"""
    # if not standard, compute standard and rescale
    if mu != 0 or sigma != 1:
        return mu + sigma * inverse_normal_cdf(p, tolerance=tolerance)
    low_z = -10.0 # normal_cdf(-10) is (very close to) 0
    hi_z = 10.0 # normal_cdf(10) is (very close to) 1
    while hi_z - low_z > tolerance:
        mid_z = (low_z + hi_z) / 2 # Consider the midpoint
        mid_p = normal_cdf(mid_z) # and the CDF's value there
        if mid_p < p:
            low_z = mid_z # Midpoint too low, search above it
        else:
            hi_z = mid_z # Midpoint too high, search below it
    return mid_z
```

- The function repeatedly bisects intervals until it narrows in on a Z that's close enough to the desired probability.

The Central Limit Theorem

- *Central Limit Theorem: A* random variable is defined as the average of a large number of independent and identically distributed random variables is itself approximately normally distributed.

In particular, if $x_1, ..., x_n$ are random variables with mean μ and standard deviation σ, and if n is large, then:

is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Equivalently (but often more usefully),

$$\frac{(x_1 + ... + x_n) - \mu n}{\sigma\sqrt{n}}$$

is approximately normally distributed with mean 0 and standard deviation 1.

The Central Limit Theorem

1. `bernoulli_trial(p)` – Performing a Single Bernoulli Trial

A **Bernoulli trial** is a random experiment with only two possible outcomes: **success (1)** or **failure (0)**.

- It's named after the Swiss mathematician Jacob Bernoulli.
- The trial has a fixed probability p of success.

  For example:

  - Flipping a coin: $p = 0.5$ (50% chance of getting heads).
  - Checking if a lightbulb is defective, where $p = 0.05$ is the probability it's defective.

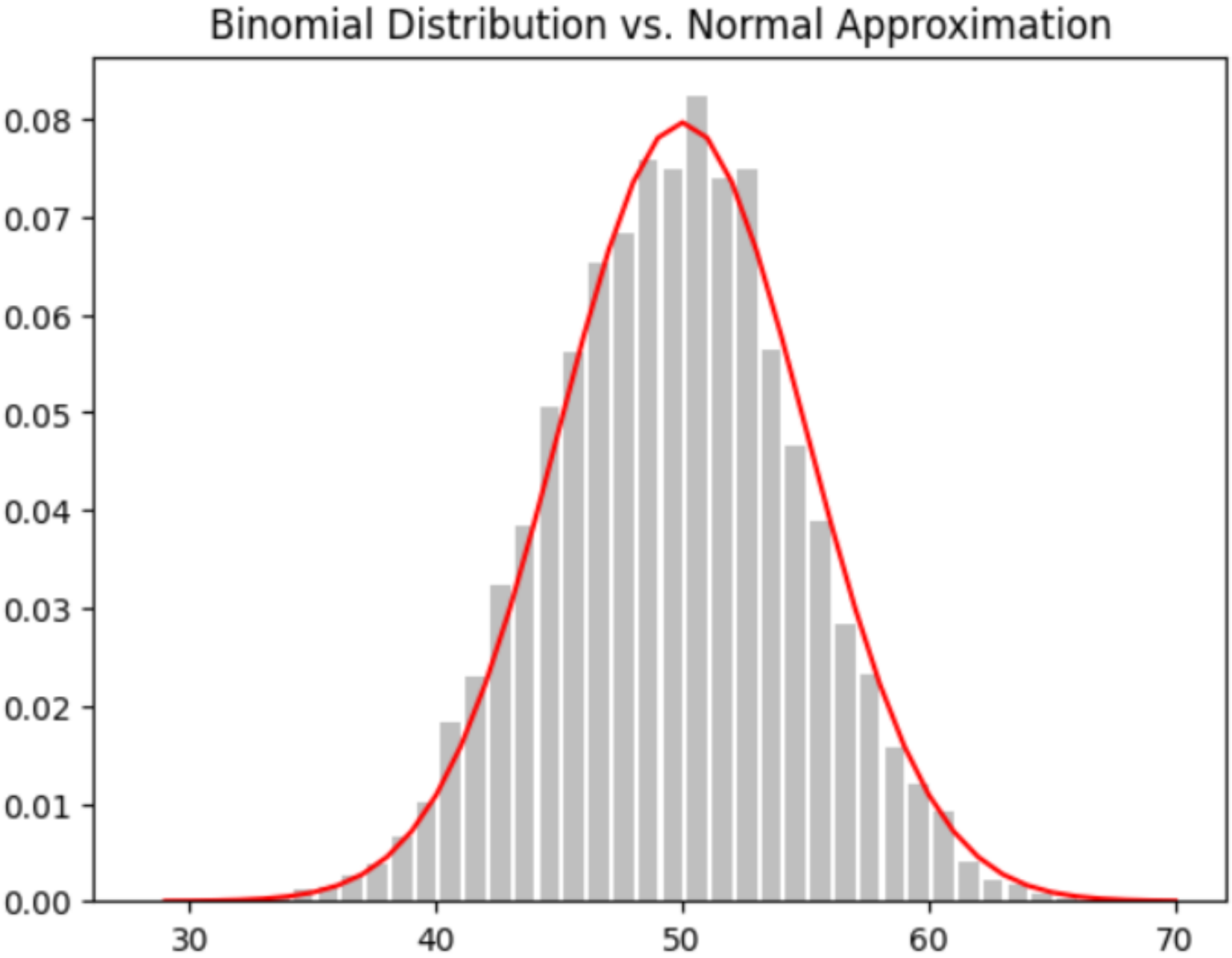
2. `binomial(n, p)` – Counting the Number of Successes in n Trials

The **binomial distribution** models the number of successes in n independent Bernoulli trials, each with success probability p.

This function simulates n Bernoulli trials and returns the total number of successes.

```
import random
import math
from collections import Counter
import matplotlib.pyplot as plt
# Function to perform a single Bernoulli trial with success probability p
def bernoulli_trial(p):
    return 1 if random.random() < p else 0
# Function to perform n Bernoulli trials and return the number of successes
def binomial(n, p):
    return sum(bernoulli_trial(p) for _ in range(n))
# Cumulative distribution function for a normal distribution
def normal_cdf(x, mu=0, sigma=1):
    return (1 + math.erf((x - mu) / math.sqrt(2) / sigma)) / 2
```

```
# Function to generate and plot the binomial and normal approximation
def make_hist(p, n, num_points):
    data = [binomial(n, p) for _ in range(num_points)]
    # Create a histogram of the binomial data
    histogram = Counter(data)
    plt.bar( [x - 0.4 for x in histogram.keys()], [v / num_points for v in
histogram.values()], 0.8,
        color='0.75‘  )
        # Calculate mean and standard deviation for the normal approximation
    mu = p * n
    sigma = math.sqrt(n * p * (1 - p))
    # Plot the normal approximation as a line chart
    xs = range(min(data), max(data) + 1)
    ys = [normal_cdf(i + 0.5, mu, sigma) - normal_cdf(i - 0.5, mu, sigma) for i in xs]
    plt.plot(xs, ys, color='red')
    plt.title("Binomial Distribution vs. Normal Approximation")
    plt.show()
# Example usage
make_hist(p=0.5, n=100, num_points=10000)
```


Binomial Distribution vs. Normal Approximation
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
40
50
60
70