Data Literacy

William John Holden

Contents

Preface

This book started with a casual conversation with Mr. Jim Steddum. I suggested some analytical topics that I think all Warrant Officers should learn. He told me that the Army wants to add "data literacy" topics to Warrant Officer Professional Military Education. I provided an outline with some proposed topics. That outline then formed the basis for this book.

This book contains a lot of code. The reader is *not* expected to be fluent or even familiar with *any* of the programming languages used. These programs are provided as an interactive learning opportunity. Source code in this text is provided along with links to online "playgrounds" where one can edit and run the code using a web browser. This gives the reader an opportunity to work directly with the material, testing ideas and learning new technologies. Another reason for teaching with code is that mathematical notation often hides complexity that one cannot ignore when solving practical problems.

Chapter 1

Introduction

1.1 Data Literacy

The *Wisdom Hierarchy* [1] begins with raw *data*. In context, data form *information*. When aggregated and interpreted through subjective values, information forms *knowledge*. When applied in novel circumstances, knowledge supports *wisdom*. Another model for Data, Information, Knowledge, and Wisdom (DKIW) is that they answer measurements, what, how, and why. The DKIW model is often visualized as the pyramid shown in figure 1.1.

This book is intended to develop *data literacy*, one's ability interpret data into useful information that will support knowledge and wisdom. Studying data requires an understanding of measurements, basic mathematics, statistics, computation, computer programming, databases, and graphs.

The weakest form of knowledge is *anecdata*, based on personal experience and intuition and not scientific rigor. Acecdata can be difficult to refute. Imagine the smoker who insists that *they* have not (yet) experienced any harmful side-effects from smoking.

We use the symbol n to represent the size of a *study*. Studies can be *interventional* (where the researcher actively makes a change to measure the result) or *observational* (where the researcher passively measures results without directly influencing the experiment). Both interventional and observational studies support *empirical knowledge* gathered through experiments.

An $n = 1$ study is effectively anecdata. Small studies risk incorrect conclusions due to *lurking variables* (variables not known or measured by the researcher) or *confounding variables* (variables that interfere with one another). The smoker might observe that lung cancer patients are elderly, claiming that age, not smoking, is the proximate cause.

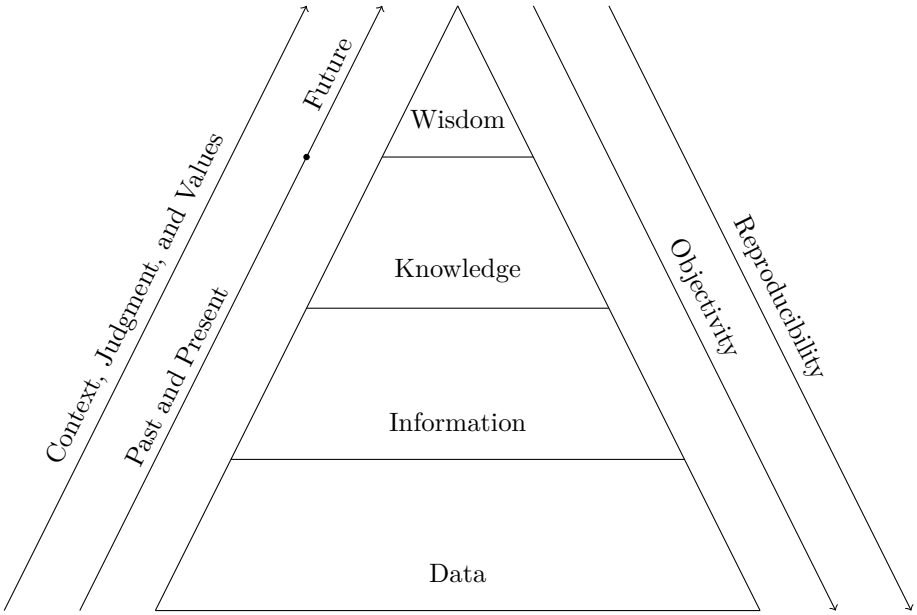


Figure 1.1: The DIKW model shows data, information, knowledge, and wisdom in a hierarchy. Higher levels require subjective valuation. Wisdom is the application of generalized knowledge to form decisions in novel circumstances.

Larger studies seek to *control* for these problems by capturing many observations among many subjects. A large study should also account for *noise* due to sampling. The human population is *approximately* 50% male and 50% female, but in an individual classroom we might have, for example, 11 girls and only 6 boys. This imbalance can be easily explained by random noise. If, on the other hand, a large school has 1110 girls and only 600 boys, then one should be more surprised by this result, assuming our *first principle* that the proportion of men and women should be nearly equal.

Data mining is an effort to distill useful information when one lacks those first principles. For example, could a climate scientist discover the ideal gas law, $PV \propto T$ (pressure and volume are proportionate to temperature) using only weather data? Once the data scientist suspects a relationship among data, a traditional scientist could structure an experiment to *verify* or *falsify* the hypothesis.

Verification and falsification are powerful tools. Both are useful to refute absolute statements. If one says, "all swans are white," then the discovery of only a single black swan falsifies the statement. Likewise, qualified statements with the quantifier "some" or "at least one" are also supported by verification. For example, the statement "some people are seven feet tall" is verifiably true, although not easily as such people are very rare.

A joke goes that an astronomer, a chemists, and a mathematician are on a train to Edinburgh and see a cow. The astronomer says, "all cows are brown." The chemists says, "some cows in Scotland are brown." The mathematician says, "there exists *at least* one cow in Scotland such that *one side* is brown." The level of precision reflects the specificity each field's conclusions. Astronomy is an observational science; interventional studies are obviously impossible at interstellar distances. Chemistry, by contrast, results from centuries of experimentation.

Mathematics and logic, unlike the sciences, primarily use *deductive* reasoning. Deductive reasoning states that an argument must be true if premised upon true assumptions. The sciences use *inductive* reasoning, which develops conclusions from observable evidence. Deductive reasoning produces *facts*; inductive reasoning produces *estimates*.

Mathematical reasoning begins with *axioms* (also known as *postulates* in geometry) that are considered obvious and acceptable without proof. We prove *theorems* from axioms and other theorems (known as *lemmas* when used as intermediates). Theorems are true derivative statements — provided the assumed axioms hold. *Conjectures* are unproven candidate theorems. We have many methods to construct proofs. These mathematical methods include proof by construction (also known as direct proof)[1]

$$p \implies q,$$

proof by contrapositive[2] [3]

$$\neg p \wedge (\neg q \implies \neg p) \implies (p \implies q),$$

proof by contradiction

$$\neg\neg p \implies p,$$

and proof by mathematical induction[4] [5]

[1]The symbol $\implies$ is pronounced "implies" and is called *conditional implication*. $a \implies b$ when b is always true when a is true. One can alternatively read $a \implies b$ as "if a, then b." The *statement* $S = a \implies b$ is false if b is false when a is true. S is still considered true when a is false, regardless of the value of b. An example is the statement, "if it is raining, then I wear a jacket," $r \implies j$. The statement is false if it rains but the speaker does not wear a jacket. The statement remains true if the speaker wears a jacket in the snow or cold without rain.

[2]The symbol $\neg$ is pronounced "not" and indicates negation. $\neg T = F$ and $\neg F = T$.

[3]The symbols $\wedge$ and $\vee$ form the logical "and" and "or", also known as *conjunction* and *disjunction*. The statement $a \vee b$ is an *inclusive or*, meaning the statement is true if a is true, b is true, or both a and b are true. The symbols for the *exclusive or* ("xor"), where exactly one of a and b are true, not both, is $a \oplus b$ and, less commonly, $a \veebar b$.

[4]We will see examples of inductive proofs in sections 5.4 and 6.5.2.

[5]Mathematical induction is, confusingly, not a form of inductive reasoning. Mathematical induction is a form of deductive reasoning.

$$p(i) \wedge (p(k) \implies p(k+1)) \implies p(n).$$

Most analysis lies somewhere between the extremes of data mining and pure mathematics. The sciences use a combination of data and reasoning, especially with statistical methods, to construct, challenge, and refine *models* that *predict* the behavior of the world according to *theories* [2]. Two examples of simple models are the binary classifier and linear model. A *binary classifier* is an example of a model that outputs *categorical* predictions, often producing either true (T) or false (F) outputs. The *accuracy* of the model is the proportion of true positives (TP) and true negatives (TN) of its predictions, which include false positives (FP) and false negatives (FN).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

A *confusion matrix* is a useful representation of the accuracy for a classifier, including classifiers with more than two possible outputs. Rows correspond to the actual categories. Columns correspond to predicted categories. The elements of the matrix are the total number of predictions, grouped by their actual categories.

$$C = \begin{array}{c} \\ a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_n \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & \cdots & p_n \end{array} \\ \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{1n} \\ c_{31} & c_{32} & c_{33} & \cdots & c_{1n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} \end{array}$$

The correct predictions fall on the *diagonal* of the matrix, therefore the accuracy of a predictor is the sum of the diagonal divided by the *grand sum* (the sum of all elements in the matrix).

$$\text{Accuracy} = \frac{\sum_{d=1}^{n} c_{dd}}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}}$$

A *linear model* is an example of a model that outputs *numerical* predictions. The linear model finds some constants $a_1, a_2, \ldots, a_n$ and b for input variables $x_1, x_2, \ldots, x_n$. The model predicts as a *linear combination* of the x-values

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b + \varepsilon$$

where ε is the residual error. *Linear regression* is the procedure for finding the a and b constants that minimize ε. Some algorithms use the *least squares method* to fit linear models. We will return to least squares in section 4.1.

There are many paradigms for implementing our models on a computing machine. *Imperative* programming, visualized in figure 1.2, allows us to represent an *algorithm* (a procedure to compute a result) directly as code. These programs construct knowledge from data in a bottom-up structure.

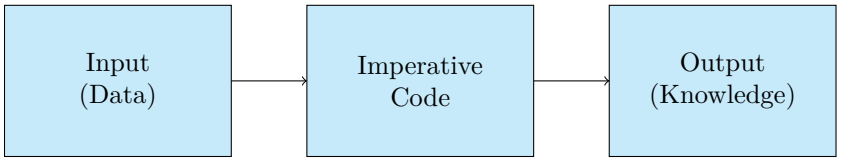


Figure 1.2: Imperative languages are useful to transform input into output. The programmer provides explicit algorithms as instructions to the computing machine.

Declarative computing environments allow the analyst to form a *query*, where a high-level language automatically answers the desired information from the top-down using rules or databases, as illustrated in figure 1.3. These categories are broad generalities. The command `ls *.txt` on a UNIX-like system provides a list of files ending with the `txt` filename extension. The interface presented to the user is declarative, but the algorithm to filter filenames that match the specification is ultimately a set of instructions.

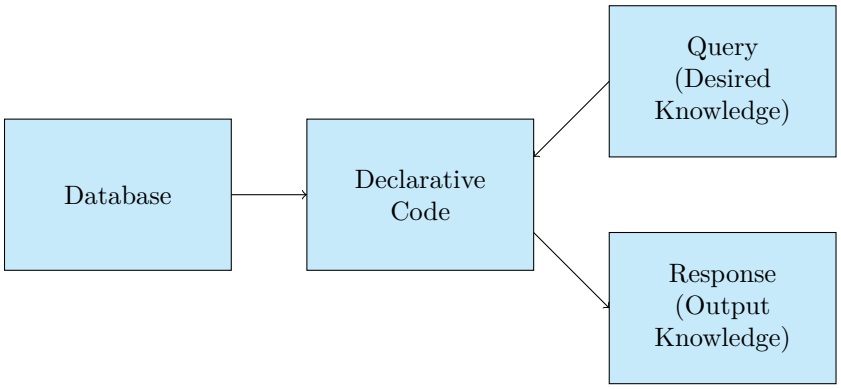


Figure 1.3: Declarative languages, such as Structured Query Language (SQL), regular expressions, and constraint solvers, search for solutions that satisfy the desired query. The user of a declarative system should not need to understand the internal workings necessary to answer queries.

Artificial Intelligence (AI) methods, such as Machine Learning (ML), seek to provide knowledge directly by modeling from data. An ML system is illustrated in figure 1.4. The term "artificial intelligence" is notoriously difficult to define [3, pp. 1–4]. In the past, spellcheck programs were considered AIs [4]. Recently, the public conflates the terms AI with ML, Generative AI, and Large-Language

Models (LLM). The point is that AI is a fundamentally different approach to using a computer from traditional programming. In any AI method, we seek to let the computer program itself by providing data and general rules.

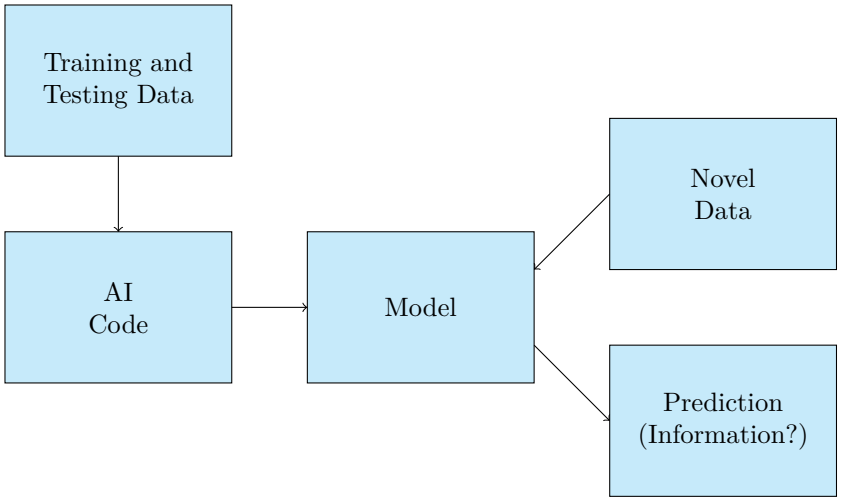


Figure 1.4: AI systems seek to model implicit algorithms by learning from data. A successful AI model is then used to form predictions on novel inputs not seen in the training and testing data. AI models can be difficult or impossible to interpret.

Scientific models of the world consider more than theories and evidence. Science is a social endeavor where consensus ("preponderance of evidence"), peer review before publishing, and reproducibility influence the community's acceptance of new knowledge. By contrast, the models generated from AI methods have no such inputs.

Consider, as a toy example, a bitterly divided community of groups A and B. Local laws forbid businesses from discriminating against the disadvantaged members of group B. The local bank develops a binary classifier to predict whether a loan applicant will default or not. The bank carefully removes any inputs to the classifier's training data that might directly reveal group membership, but in practice the classifier nearly always predicts that the objectively poor and troublesome members of group B will default on their loans. The problem here is that the model lacks the subjective wisdom to form decisions in the context of contradictory values.

Applying our own subjective judgment, anecdotal experiences, contradictory values, vague priorities, and burdensome regulations is a feature, not a bug, of applying wisdom to for decisions from knowledge. Recent excitement in AI methods have led some individuals and organizations to believe that they can automate many roles presently performed by knowledge workers. While some

roles may be deterministic enough to successfully automate, many organizations have forgotten the slogan "the process is the product." A compelling chart, insightful statistic, or accurate prediction is seldom the main value of data-driven analysis. The value of the analysis is in the deep thinking that went into the report. In exploring the data, the analyst must test assumptions, discover relationships, and uncover the hidden structure of their subject.

This book aims to expose the reader to many skills for processing, visualizing, and interpreting data. We will use many different programming languages, some mathematics and statistics, and problems one may encounter. At the end of each chapter, this text provides discussion prompts for group learning and practical exercises for individuals.

1.2 Parameters and statistics

Statistics are the foundation of most data mining, machine learning (ML), and artificial intelligence (AI) methods today. A *statistic* is an estimate of a *parameter*, which is a characteristic of an entire *population*. Statistics are calculated from taking *samples* (subsets) from the population.

For example, suppose we wanted to find the height of the tallest mountain in the world. We might sample $n = 100$ mountains at random from an almanac. Suppose the tallest mountain in our sample is Mount Fuji. Mount Fuji, the tallest mountain in Japan, is 3776 meters tall. We can conclude that the tallest mountain in the world is *at least* 3776 meters tall.

Our estimate is unfortunately quite low. Mount Everest in Nepal, the *highest* mountain in the world, stands 8849 meters above sea level. Mauna Kea in Hawai'i, the *tallest* mountain in the world, stands 4207 meters above sea level and another 6004 meters below. Our estimates of population parameters, *statistics*, generally improve with larger sample sizes, and many statistical methods provide a *margin of error* quantifying sampling error.

One might use statistics to create a *model* to explain a population, based upon sampling data. Models can be useful both for describing the population and also for forming predictions.

1.3 Number Representation

Modern computing machines are *digital* systems which represent numbers as *strings* (lists) of *bits*. A bit has only two possible values: 0 or 1. Today, *bytes* are eight bits long, although there were computers in the past which did not follow this convention.

An unsigned, sixteen-bit integer has $2^{16} = 65\,536$ possible values. The bit string 0011000000111001 represents 12 345.

$$2^0 + 2^3 + 2^4 + 2^5 + 2^{12} + 2^{13} = 1 + 8 + 16 + 32 + 4096 + 8192 = 12\,345$$

Computer engineers have developed several techniques for representing signed (possibly negative) numbers. Some computers use a dedicated sign bit. Others use *one's complement* or *two's complement* representations.

Decimals are still more complex. *Fixed-point* decimals provide some constant amount of digits for the whole and fractional parts of the number. *Floating-point* decimals provide dynamic whole and fractional digits, enabling the computer to represent a large number of decimal digits when possible. Floating-point values are common, but have several well-known pitfalls. In any modern web browser, such as Firefox, press F12 to open the developer console and enter `0.1+0.2`. The result is not `0.3`:

```
>> 0.1+0.2
<- 0.30000000000000004
```

This happens for the same reason that $2 \div 3$ produces a repeating fraction in base 10. The number 2/3 cannot be exactly represented in a string of decimal digits (a sum of powers of ten). Likewise, the number $0.3 = 3/10$ cannot be exactly represented in a string of binary digits.

Computing machines do not process numbers. Rather, they process bit strings which represent numbers.

1.4 Levels of measurement

There are four distinct *levels of measurement* that a value may fit [5]. *Nominal* data is simply names or categories, with no concept of order or distance. A movie might be animated or live-action, a dichotomy without order. Another example might be the film's genre (children, comedy, action, romance, documentary, etc).

Ordinal data has ordering but not distance. Ordinal data might be represented as ordered categories or as numerals, though these numerals do not provide meaningful addition and subtraction. The ratings of a film (G, PG, PG-13, R, and so on) form a ranking, but addition is meaningless (does G + PG-13 = R?) and our concept of distance is weak at best. Another example of ordinal might be the rankings the films receive at an awards ceremony, where one film is the winner and another is the runner-up.

Interval data is numerical data with a concept of distance but not multiplication. The year when a film was produced is an example of interval data. If two films were produced in 2000 and 2010, then it makes sense to say one was made ten years later, but we would not say that the latter film is $\frac{2010}{2000} = 1.005$ times the first.

Ratio data is numerical data with both distance and multiplication. The gross earnings of a film is an example of ratio data. If the 2000 film earned one million dollars and the 2010 film earned two million dollars, then it makes sense to say the second film earned double the first.

Name	Operations	Type
Nominal	$=, \neq$	Categories
Ordinal	$<, >$	Ordered categories
Interval	$+, -$	Numbers with distance
Ratio	$\times, \div$	Numbers with meaningful zero

Interval data might be initially confusing to distinguish from ratio data. One indication is the absence of a meaningful zero. Does zero degrees Celsius or zero degrees Fahrenheit mean the absence of temperature? No. These temperature measurements are simply points along a *scale*. Twenty degrees Celsius is not "twice" ten degrees Celsius; multiplication is not defined on interval data.

Grid coordinates are another example of interval data. One can calculate the distance between two grid coordinates, but we would not say that coordinate 1111 is "half" of coordinate 2222.

Women's pant sizes in the United States, with the confusing size "00," is yet another example of interval data.

Data might be represented in numerical formats when some operations do not make sense. Suppose a political scientist encoded voter's political party as "1", "2", "3", and "4". Is "2" an intermediate value between "1" and "3", or are these actually nominal data where the only arithmetic operations are $=$ and $\neq$? AI methods may form incorrect assumptions about data that domain experts can easily prevent.

1.5 Discretization

Measurements with arbitrarily many decimal digits of precision are *continuous*, whereas measurements with finite steps in between (including categories) are *discrete*. For example, when driving along a road, the house numbers (150 Main Street, 152 Main Street, 154 Main Street…) are discrete; there is no intermediate value between 150 and 151. On the other hand, the grid coordinates associated with each address are continuous; one could (theoretically) specify grid position to the square millimeter, picometer, nanometer, and beyond.

Spoken English has some vocabulary for distinguishing continuous and discrete quantities, although these conventions are not strictly necessary in daily communication.

- How *many* glasses do we have?

- How *much* water do you want?
- I walked *fewer than* 10 km.
- She weighs 5 kg *less than* her sister.

It can be useful to combine continuous measurements into discrete categories. An example might be one's birth date and birth year. No one knows their birth *instant* with subsecond precision. Rather, the year, year and month, or year, month, and day are almost almost always enough information. We even combine years into groups when discussing generations and peer groups. Combining a range of birth years into generational categories is an example of *discretization*.

1.6 Missing values

In practice, *data sets* often have missing values. Different programming languages have substantially different syntax and semantics for representing and handling missing values.

As a small exercise, open Microsoft Excel and enter the values 1, 2, 3, and 5 into cells A1, A2, A3, and A5. Leave cell A4 blank. In cell A6, enter the formula `=PRODUCT(A1:A5)`. The result is $30 = 1 \times 2 \times 3 \times 5$ (the 4 should be missing). Excel did *not* treat the empty cell as an implicit zero and silently ignores the missing value.

Now change cell A4 to `=NA()`. `NA` means "value not available", an explicit indication that a value is not given. The product in cell A6 should update to `#N/A`, which explicitly tells us that there is a problem in the calculation.

Now change cell A4 to `=1/0`. Both cells A4 and A6 should both say `#DIV/0!`, a fault telling us that a division by zero has made further calculation impossible.

Error values propagate from source data through intermediate calculations to final results. If we enter a formula into A7 referencing A6, such as `=SQRT(A6)`, then we will find the same faults in A7 that we see in A6.

Structured Query Language (SQL) databases use the symbol `NULL` to denote missing values. One might build the database *schema* (the structure of the database) to explicitly forbid `NULL` values. For example,

```
CREATE TABLE Run (
  Name TEXT NOT NULL,
  Time INTEGER NOT NULL,
  Distance REAL NOT NULL)
```

defines a table *schema* where each of the three columns must be specified. Many programming languages (including C, Java, and JavaScript) also use the term `null` for variables that do not reference any specific value.

Many programming languages support a `NaN` ("not a number") value in error conditions. One might encounter `NaN` when dividing by zero, subtracting infinities,

and parsing non-numeric words as numbers. Comparisons with `NaN` can be confusing, such as `NaN == NaN` returning *false*.

Some programming languages will automatically *initialize* variables with some zero value. Other languages give some `Undefined` value to uninitialized variables. Still other languages raise an error if no explicit value is assigned to a variable.

1.7 Optional and Sentinel Values

Computer scientist Tony Hoare famously called null references a "billion dollar mistake," explaining that programming languages with *nullable* values contain flaws that might have been prevented in languages that require value initialization [6].

Rust is one of many young languages that provides an *optional* type to express a value which may or may not contain useful information. The form `Some(value)` indicates a usable value. If the programmer wishes to express the absence of a value, they use `None`.

```
use rand::Rng;

fn g() -> Option<f32> {
    let mut rng = rand::rng();
    let x = rng.random_range(0.0..=1.0);
    if x > 0.5 {
        Some(x)
    } else {
        None
    }
}

fn main() {
    match g() {
        Some(x) => println!("g() returned some value, {x}."),
        None => println!("g() returned none.")
    }
}
```

Repeatedly run this program at the Rust Playground[6] and observe that the `g()` function returns `Some(x)` values where $0.5 \leq x \leq 1.0$ and `None`.

The use of the language's type system to express optional values allows Rust to eschew *sentinel values*, which are special values used to control a program. An example of a sentinel value is in the `read()` function of the C programming language. `read()` is used to read data from files, and ordinarily returns the

[6] https://play.rust-lang.org/?&gist=df4c6636ab6ff336dbae5994b7508adc

number of bytes read, but in many cases returns 0 or −1, signaling status to the caller.

1.8 Comma-Separated Values (CSV)

Comma-separated values (CSV) are a well-known situation where sentinel values are especially problematic. A CSV file is a simple method of structuring data in plain-text files. For example, a table such as

x	y	z
Rob	0.74019382956651820	0.3508759018489489
John	0.41331428270607506	0.2936926427452584
David	0.37671743737357277	0.5676190157838865
Frank	0.50270122376380740	0.7939268929144455

can be entered into a CSV file as

```
x,y,z
Rob,0.74019382956651820,0.3508759018489489
John,0.41331428270607506,0.2936926427452584
David,0.37671743737357277,0.5676190157838865
Frank,0.50270122376380740,0.7939268929144455
```

In section 3.8, we will see this same data represented in a more modern format called JSON.

What happens if we add a column where values themselves contain commas? RFC 4180, a specification for CSV, recommends enclosing values containing commas with quotation marks [7]. If the value enclosed with quotation marks also contains a quotation mark, then a doubled quotation mark (`"this field contains ""escaped"" quotation marks"`) *escapes* the inner value for unambiguous parsing.

Unicode and its predecessor, the American Standard Code for Information Interchange (ASCII), provide *control codes* `001C`, `001D`, `001E`, and `001F` for file, group, record, and unit separators, respectively, but these codes are not commonly used. If these codes had been more convenient to type, our world of data might have avoided some of the common pitfalls of CSV and other formats containing sentinel values.

1.9 Strong/weak and static/dynamic typing

Values come in many forms: categorical and numerical, ordered and unordered, discrete and continuous, defined and missing. *Types* can be used to constrain variables to allowable values and applicable operations.

For example, suppose a database indicates how many cars a person owns. It makes no sense to own a fractional or negative car, so we might find an existing type (in this case, whole numbers) or define some new type to model the domain.

Some programming languages offer *dynamic* types that implicitly change the type (*cast*) of values to operate correctly. Go to https://jsfiddle.net or press F12 to open the developer console in most modern browsers. Enter the following into the JavaScript console:

```
>> "5" * 5
<- 25
```

Characters inside quotation marks (`"5"`) are called *strings* and are ordinarily used for text, but JavaScript automatically parses `"5" * 5` as the product of two numerical values and returns `25`.

JavaScript is notoriously inconsistent.

```
>> "5" + 5
<- "55"
```

The resulting string, `"55"`, is the *concatenation* of two strings – perhaps not what one expects.

Many languages and environments seek to automatically parse values. Microsoft Excel and the Python programming language are also dynamic. Other languages, such as Java and Go, are more strict with values and do not automatically change values, especially when the conversion might be "lossy" (where information might be lost, such as approximating the exact value of π as 3.14, or rounding 3.14 to 3, or even changing 3.0 to 3). These languages have both *strong* and *static* typing: the programmer must specify the type of each variable, and lossy type conversions require an explicit cast.

Excel does provide some basic functionality to set number *formats*, but this feature might not stop one from confusing one type of data for another. Excel uses *weak* typing that does prevent one from using unexpected values. Data analysts can benefit greatly by using the appropriate types for the values in their problem.

1.10 Tables, lists, and data frames

Tables of data are structured in *columns* and *rows*, where the rows represent the *individuals* or *observations* in the data set and the columns represent the *features*. For example, a table of employee names might have two columns (the given and surnames) and ten rows, where each row represents one of the ten employees.

In computer science, the terms *list* and *array* both refer to single-column tables, but with different internal memory representation. The distinction is usually unimportant to data analysts.

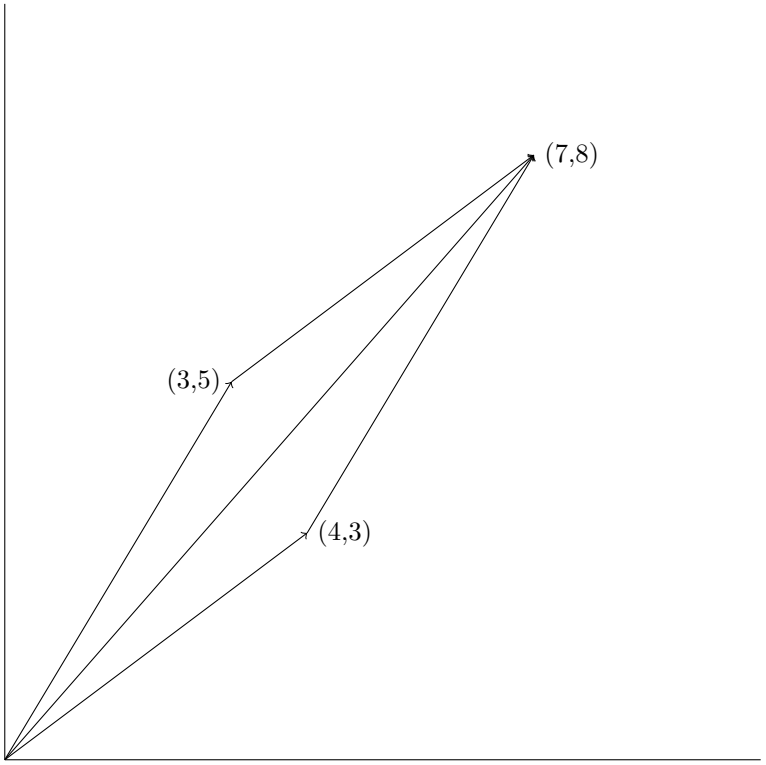


Figure 1.5: Vectors are geometric entities. This plot shows that $(4, 3) + (3, 5) = (3, 5) + (4, 3) = (7, 8)$.

Scientific languages, such as Julia and R, often use the term *data frame* (or *dataframe*) as their method for representing tables of data. Data frames often provide rich syntax for row-wise and column-wise operations. By contrast, in an object-oriented language, such as Java and JavaScript, the idiomatic representation of a table is likely an array of objects. We will discuss object-oriented programming in more detail in section 3.7.

1.11 Vectors and matrices

We now quickly mention the terms *vector* and *matrix* here to disambiguate them from other terms already defined.

Arrays, lists, and columns containing numeric data may sometimes be represented with *vectors*. Likewise, tables and data frames might be represented with *matrices*.

A vector is a quantity with both magnitude and direction, generally consisting of two or more elements. A plot demonstrating vector summation is shown in

figure 1.5.

$$\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$$

The above vector $\mathbf{x}$ has three components and length $\sqrt{x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2}$.

A matrix is a collection of vectors used for linear transformations. For example, the three-component *identity matrix*

$$I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

has the property

$$\begin{aligned} I_3\mathbf{x} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= \begin{pmatrix} 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 \\ 0 \cdot x_1 + 1 \cdot x_2 + 0 \cdot x_3 \\ 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 \end{pmatrix} \\ &= \mathbf{x}. \end{aligned}$$

Vectors and matrices form the foundations of *linear algebra*, a rich and powerful branch of mathematics that produces many of the results needed for modern statistics, ML, and AI methods.

Remember that the different in ratio and interval data was that *multiplication* is only defined for ratio data. Similarly, multiplication is well-defined for vectors and matrices, but not on tables of data.

Depending on the problem domain, it may be inappropriate to use matrices and vectors to represent data where such operations are not necessary. Some programming languages use the terms "vector" and "array" interchangeably, or to indicate an array has dynamic vice fixed size. Many programming languages support *tuples* as an alternative representation of a quantity with multiple values.

One must take care to verify an arithmetic operator performs as expected with "vectors" and tuples in different languages. Compare and contrast the semantics of arrays and tuples in Python

```
In [1]: 1 == [1]
Out[1]: False

In [2]: 1 == (1,)
Out[2]: False
```

```
In [3]: [1, 2] + [3, 4]
Out[3]: [1, 2, 3, 4]

In [4]: (1, 2) + (3, 4)
Out[4]: (1, 2, 3, 4)
```

and R

```
> 1 == c(1)
[1] TRUE
> 1 == list(1)
[1] TRUE
> c(1, 2) + c(3, 4)
[1] 4 6
> list(1, 2) + list(3, 4)
Error in list(1, 2) + list(3, 4) :
  non-numeric argument to binary operator
```

Python's behavior is typical of general-purpose programming languages, while R's behavior (see section 3.6.4) is common among scientific languages.

1.12 Complex Numbers

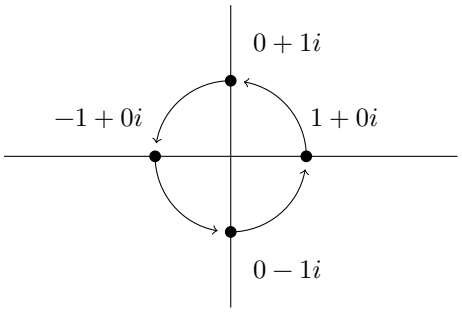


Figure 1.6: Complex numbers are two-dimensional quantities. In this Argand diagram, the x axis is ± 1 and the y axis is $\pm i$. Multiplying a value by i rotates the value among the real and imaginary axes.

Complex numbers, which have "real" and "imaginary" components, are also multidimensional values, but with the property $i = \sqrt{-1}$ and therefore $i^2 = -1$. This means that their multiplication forms a cycle, as shown in figure 1.6.

$$
\begin{aligned}
1 \times i &= i \\
i \times i &= -1 \\
-1 \times i &= -i \\
-i \times i &= 1
\end{aligned}
$$

Many languages provide complex arithmetic for situations that require it. An R example is shown below.

```
> c(1, 1i, -1, -1i) * 1i
[1]  0+1i -1+0i  0-1i  1+0i
```

While complex arithmetic is common in physics and signal progressing, many scientific disciplines have no use cases for complex numbers. If one has a two-dimensional quantity and no application for the multiplication rules shown in figure 1.6, then one should avoid complex types and instead favor arrays.

As an example, suppose one wanted to represent the systolic and diastolic values in blood pressure samples. One *could* use a complex value instead of an array or tuple, but now the values have a concept of multiplication and direction that is not appropriate to the problem domain. Just as we learned to consider which arithmetic operations apply to our data in section 1.4, likewise we must consider the operations applicable to composite values.

1.13 Sets, relations, functions, and algorithms

We now introduce a few terms from *discrete mathematics* that are fundamental to all analysis. A *set* is an unordered collection of *distinct* elements. Sets may be finite or infinite in size. Sets are denoted with curly braces, and the empty set has the special symbol $\emptyset = \{\}$. An example of a set might be

$$W = \{\text{Sunday}, \text{Monday}, \text{Tuesday}, \text{Wednesday}, \text{Thursday}, \text{Friday}, \text{Saturday}\}$$

A *relation* is an association between members of sets. Relations can be used to model any relationship between members any two sets, or even members in the same set. An example might be the relation between integers and elements of W with that many letters, i.e. 6 has a relation on Sunday, Monday, and Friday, 7 has a relation on Tuesday, 8 has a relation on Thursday and Saturday, and 9 has a relation on Wednesday. The term "relation" is seldom used outside of discrete mathematics, but there is a *special case* of a relation that occurs in all mathematical disciplines: *functions*.

A *function* is a relation that uniquely relates members of one set (the *domain*) to another set (the *range*). An example of some functions might be:

$$\begin{aligned}
\text{Translate}\,(\text{Friday}, \text{English}, \text{German}) &= \textit{Freitag} \\
\text{Length}\,(\text{Wednesday}) &= 9 \\
\text{Distance}\,(\text{Thursday}, \text{Tuesday}) &= -2 \\
\text{DaysOfLength}\,(6) &= \{\text{Sunday}, \text{Monday}, \text{Friday}\} \\
\text{Sunday} &= \text{Next}\,(\text{Saturday}) \\
&= \text{Previous}\,(\text{Monday}) \\
&= \text{Previous}\,(\text{Previous}\,(\text{Tuesday})) \\
&= \text{Previous}\,(\text{Next}\,(\text{Sunday}))
\end{aligned}$$

Each of these functions accepts one or more *parameters* as *arguments* and returns the unique corresponding value (if any) from its range. It may appear that the third function, DaysOfLength, has returned three values, but actually this function has returned a single set which contains three values.

Many programming languages use the term "function" as a synonym for *procedure*, *subroutine*, and *method*. Functions are "pure" if they have no side-effects, such as mutating a value outside of the function.

The mathematical definition of the term *algorithm* is the set of instructions necessary to solve a problem. Long division, a procedure for manually dividing numbers, is an example of an algorithm. The term "algorithm" has recently entered the popular lexicon in relation to AI systems. Here, the instructions of the algorithm are part of a model, which is created from data.

1.14 Abstraction and Reification

Take any three-digit decimal (base 10) number, reverse the digits, and their difference will always be divisible by both 9 and 11. For example, $321-123 = 198$; $198 \div 9 = 22$ and $198 \div 11 = 18$.

How and why would this strange property hold? The proof is quite easy using algebra. We change the digits of a three-digit number into variables. Some three-digit number $abc = 100a + 10b + c$.

$$\begin{aligned}
abc - cba &= (100a + 10b + c) - (100c + 10b + a) \\
&= 100a + 10b + c - 100c - 10b - a \\
&= 99a - 99c \\
&= 99(a - c)\square
\end{aligned}$$

By *abstracting* numerals into variables, the claim becomes easy to verify.

Here is another example of abstraction. How does one calculate 30% of 70 without a calculator? First, observe that

$$x\% \text{ of } y = \frac{x \times y}{100}.$$

So, 30% of 70 is $\frac{30 \times 70}{100} = \frac{2100}{100} = 21$.

Abstraction can be a powerful tool for solving problems and developing proofs. In the field of computer networking, countless problems are solved by the pattern, "we have more than one thing, but it is inconvenient to operate more than one of these things, so we built a method to abstractly represent arbitrarily many of these things as super-things."

Abstractions are *leaky* when one must understand the internal details to use the abstraction effectively. For example, surprises such as `0.1 + 0.2 == 0.30000000000000004` in floating-point arithmetic lead to programmers understanding more implementation details than intended; floating-point arithemetic is a leaky abstraction towards representing real numbers (see section 1.3). The object-oriented paradigm (see section 3.7) emphasizes *encapsulation* of both data and code to allow users to use the *interface* exposed by an object without consideration for its implementation.

Reification is the opposite of abstraction: we something specific from something general. For example, suppose we have a language translation function

$$T(x, l_1, l_2)$$

where x is the message to be translated, l_1 is the input language, and l_2 is the output language.

$$T(\text{hello}, \text{English}, \text{French}) = \textit{bonjour}$$

From this general function, we can enclose parameters l_1 and l_2 into a specific function.

$$\begin{aligned} T'(x) &= T(x, \text{English}, \text{French}) \\ T'(\text{goodbye}) &= \textit{au revoir} \end{aligned}$$

Function T' reifies T into a less general form. Such functions might be called *convenience functions* that are provided as a "quality of life" improvement for the user. An example of a convenience function might be `LOG10(number)` in Excel. Excel also provides `LOG(number,[base])` (where `base` defaults to 10 if omitted), but some users may prefer the explicit syntax `LOG10` to improve clarity.

1.15 Discussion prompts

1. If `ls *.txt` is a declarative program, then is `rm *.txt` also a declarative program?
2. Who owns knowledge management?
3. What are good and bad uses for spreadsheets?
4. What is reproducibility? Why would this be important for scientific inquiry?
5. A manager sends an Excel spreadsheet to their employees, telling them to each enter information and send it back. What are some challenges the manager might experience while merging these spreadsheets?
6. The set of natural numbers, $\mathbb{N} = \{0, 1, 2, 3, ...\}$, can be constructed using a *successor function*, $S(n) = n + 1$. If we begin with $0 = 0$, then $1 = S(0)$, $2 = S(1) = S(S(0))$, $3 = S(2) = S(S(1)) = S(S(S(0)))$, and so on. Can we define the reals, $\mathbb{R}$, in such a way? Could we construct a successor function for floating-point approximations of real numbers?
7. CSV makes *appending* data easy: simply write new rows to the end of the file. This structure also makes it easy to *stream* the data record-by-record, but it makes the schema of the data inflexible. One must define the columns in the header of the CSV and continue to use this structure thereafter. Do Excel spreadsheets and SQL databases share this problem? How can organizations store their data when their requirements may change?

1.16 Practical Exercises

1. A sample of $n = 6$ numbers $S = \{7, 17, 37, 47, 67, 97\}$ shares two properties: each value ends in 7, and each value is a prime number. Prove or disprove the statement "if a number ends in 7, then it is prime." Is this type of sampling a common experimental strategy for deductive fields, such as logic and pure mathematics?
2. What method can we use to prove or disprove each of the following statements?
   (a) All numbers that end in 7 are prime.
   (b) No number that ends in 7 are prime.
   (c) Some numbers that end in 7 are prime.
   (d) Some numbers that end in 7 are not prime.
   (e) There are infinitely many numbers that end in 7 and are prime.
   (f) The density of prime primes ending in 7 decreases as $x \to \infty$.
3. There are two equations defining accuracy in section 1.1. Explain why the definition based on a confusion matrix is a generalization of the other

accuracy statistic, which is built from true and false positives and negatives. Which equation is more abstract?

4. Create a small survey using Microsoft Forms (part of Office 365) or Google Forms (part of Google Docs). Compare this experience to the hypothetical manager who gathered information by manually merging spreadsheets.

5. Given a noisy and poorly structured dataset, propose a method of restructuring the data.

6. Discretize the values of a dataset and explain the reasoning.

7. The following Rust program, which one can run at the Rust Playground[7], doubles a value until we exceed the largest representable value. However, the program *appears* to make an arithmetic error at 134 217 730. $67\,108\,864 \times 2 = 134\,217\,728$, not 134 217 730 (no power of two could ever end in zero in decimal). Use https://www.h-schmidt.net/FloatConverter/IEEE754.html to investigate the error.

```rust
fn main() {
    let mut x: f32 = 1.0;
    while x != f32::INFINITY {
        println!("{x}");
        x *= 2.0;
    }
}
```

8. What is the output of the following Java program? Use the Java Playground[8] to experiment.

```java
static void count() {
    float x = 0.0f;
    while (x != x + 1.0f) {
        x += 1.0f;
    }
    System.out.println(x);
}

count();
```

9. An *azimuth* on a magnetic compass conventionally reads 0° when pointed north, 90° for east, 180° for south, and 270° for west. In trigonometry, the angle 0° corresponds to (x, y) position $(1, 0)$ on the unit circle, 90° to $(0, 1)$, 180° to $(-1, 0)$, and 270° to $(0, -1)$. Implement a function A to convert azimuths to angles, another function A^{-1} to convert angles to azimuths, and create test cases to verify that $A^{-1}(A(\theta)) = \theta$.

[7] https://play.rust-lang.org/?gist=82eb9505ef18cf3af0faa2a373c11901
[8] https://dev.java/playground/

Chapter 2

Data Visualization

2.1 Plots

Plots allow us to visualize data. Good plots help us to quickly intuit patterns in the data that might otherwise be difficult to understand.

The term *graph* has different definitions in lower and higher mathematics. We will explain the term "graph" in chapter 6. This text uses the term "plot" as the verb and noun for visualizing data with graphics.

The *bar plot* helps us to compare the count each category in a discrete (or discretized) variable. The *box plot* helps us to see the center and variation of a numerical variable. The *histogram* also helps us to see the center and variation of a numerical variable, often producing the familiar *bell curve* shape, where the height of the curve indicates the count of observations within the range of each "bin." A histogram is essentially a set of bar plots over discretized numerical values.

A *scatter plot* (sometimes called an XY plot) uses x and y axes to show relationships between two variables. One can also color and shape the points to show third and fourth variables. Three-dimensional XYZ plots are sometimes useful, especially in video and interactive presentations.

As a small exercise to experiment with these four plots, go to https://webr.r-wasm.org/latest/ to use the R language in a web browser. R is a programming language for statistics and data visualization.

R includes several built-in data sets. We will use included Motor Trend Cars (`mtcars`) data set.

In the *read-evaluate-print loop* (*REPL*) at the bottom-left of the screen, enter

```
> head(mtcars)
```

to view the column names and first six rows of the Motor Trend Cars (`mtcars`) data set. Enter

```
> mtcars
```

to view the full data set.

Place a question mark before a function or data set name in the REPL to get help in R. Try opening the R help for `mtcars` and `head` with the following commands:

```
> ?mtcars
> ?head
```

2.2 Line Plots

The *line plot* is among the most basic of plots. We seldom see one-dimensional number plots in the sciences, but they are used extensively in elementary education to develop numerical intuition in children. It can also be useful to draw line plots to represent *continuums* of ordered data, including discretized categories.

2.3 Scatter Plots

```
> plot(mtcars$wt, mtcars$mpg)
```

Todo: change base R to ggplot so we can show color.

We can combine two or three orthogonal (perpendicular) line plots to create a *scatter plot*. Scatter plots show data in two or three dimensions. The shape of the plot may reveal patterns and trends in the data.

The term *data* carries weight in this context. Scatter plots *can* be used to represent data in aggregates, but it m

2.4 Bar Plots

Bar plots relate categories to aggregated numerical features of a data set. The categorical group is the *independent* variable. The numerical feature, plotted as the length of the bars, is the *dependent* varible. Independent and dependent variables are sometimes called *free* and *response* variables. In an *interventional* study (where a researcher performs an action to quantify the effect), the independent variable is the item changed directly and the dependent variable is the outcome caused by the change. Bar plots almost always require some amount of "data wrangling," such as the use of SQL aggregate functions (more on this in section 3.5) such as `MIN()`, `MAX()`, `COUNT()`, `SUM()`, and `AVG()`. Figure 2.2 demonstrates a bar plot.

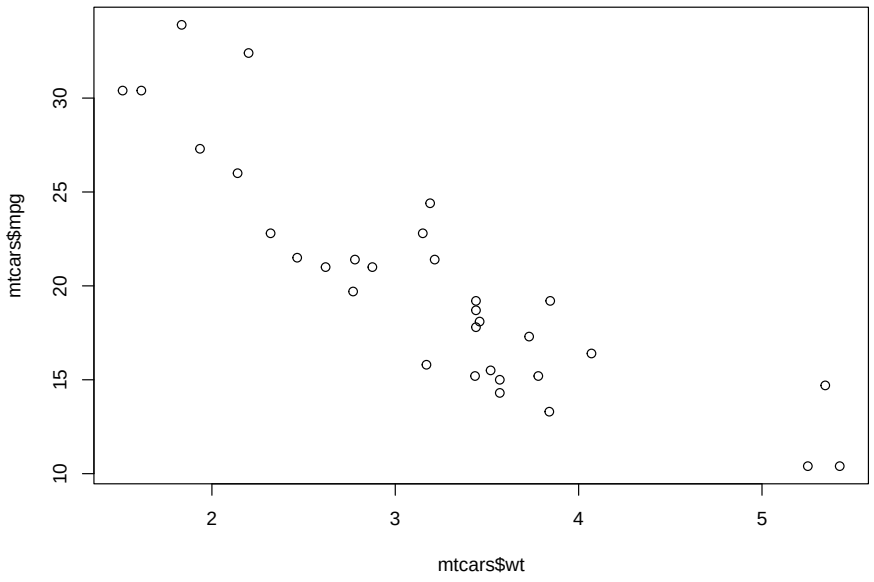


Figure 2.1: todo

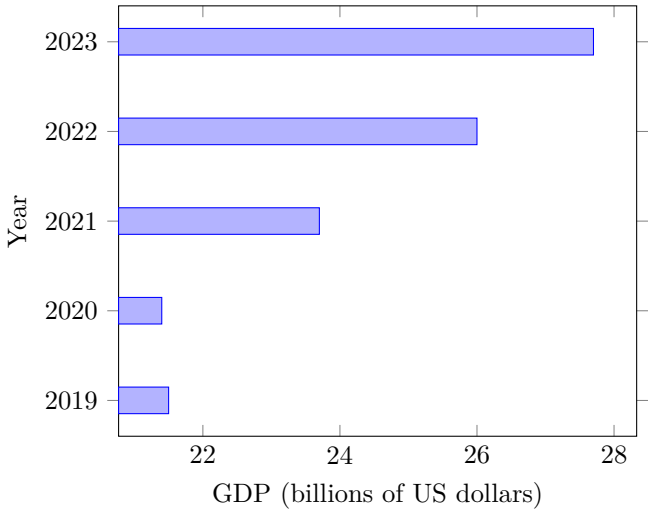


Figure 2.2: A bar plot showing the Gross Domestic Product (GDP) of the United States from 2019–2023, according to <https://tradingeconomics.com/united-states/gdp>. Bar plots summarize information by representing aggregates (such as sums) as functions of categories. Here, years are treated as a categorical variable.

The width of each bar must be uniform. Only the bar height varies. As Area = Width × Height, the exaggerated area of a wide or thin bar will mislead the reader. For example, suppose a bar plot is intended to compare the values $x = h = (3, 10, 11)$, but the bars corresponding to each observation are, respectively, $w = (1, 1, 3)$. The resulting areas are $w \odot h = (3, 10, 22)$ (here, $\odot$ indicates the *element-wise* product of two vectors, also known as a *Hadamard* product). As shown in figure 2.3, the area of the third bar is more than triple that of the second and may confuse the reader.

In the R language, one can create bar plots using the `barplot`. Using https://webr.r-wasm.org/latest/, recreate the figure in 2.2:

```
> gdp = data.frame(Year = 2019:2023, GDP=c(21.5,21.4,23.7,26,27.7))
> gdp
  Year  GDP
1 2019 21.5
2 2020 21.4
3 2021 23.7
4 2022 26.0
5 2023 27.7
> barplot(GDP ~ Year, gdp)
> library(tidyverse)
> ggplot(gdp, aes(x=Year, y=GDP)) + geom_col()
```

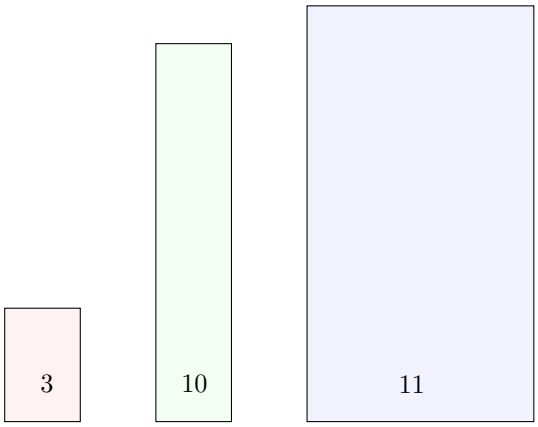


Figure 2.3: The bars of a bar plot should ordinarily have uniform width. This bar plot shows values $x = (3, 10, 11)$, but the width of the third bar makes this observation appear much larger than the others.

2.5 Pareto Charts

A *Pareto chart* is a combination of a sorted box plot and line plot. Pareto charts are frequently used in industrial settings to show cumulative proportions among categories. A typical application for a Pareto chart is to triage the most common causes for a problem to maximize effectiveness of finite resources. Pareto charts are closely associated with the "Pareto Principle," an estimation that 80% of problems are caused by 20% of causes.

A 2014 observational study by Weisenthal et al. surveyed injury rates among CrossFit athletes [8, p. 7]. The researchers report a total of 84 injuries in six movement categories, which are shown in the following table. The movement types are presented in descending order by injury count. The proportion is simply the injury count for the current row divided by the total of injury counts. The cumulative sum, `cumsum` in the R language, is the sum of the current injury proportion and those before.

Movement Type	Injury Count	Proportion	Cumulative Sum
Power Lifting	19	22.619%	22.619%
Gymnastics	17	20.238%	42.857%
Not Associated	16	19.048%	61.905%
Olympic Lifting	14	16.667%	78.572%
Other	13	15.476%	94.048%
Endurance	5	5.952%	100.000%

A Pareto chart for this table is shown in figure 2.4.

R does not provide a native method to produce Pareto charts, but we can do so ourselves with the `ggplot2` library[1]. First, we load the `tidyverse` library[2], which will also include the `dplyr` package[3]. `dplyr` provides the `%>%` infix operator, which anonymously "pipes" the output from one function into the first argument of the next function. We tally each distinct observation, sort them in descending order by count, compute the proportion of the total, and compute the cumulative sum. Finally, we plot the proportions as a box plot and overlay the cumulative sums as a line plot.

This example uses data gathered from CrossFit.com[4]. `dplyr` uses `mutate` as row-wise `map` operation with support for aggregate functions (such as `sum(n)`

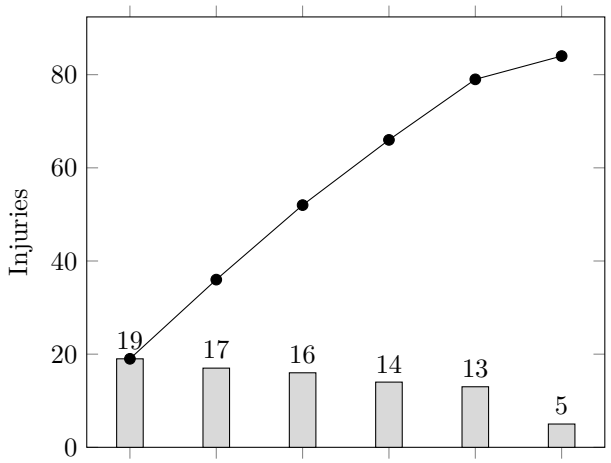


Figure 2.4: A Pareto chart shows the relative and cumulative proportions of discretized quantities, sorted in decreasing incidence. Frequently used in quality control processes, such as Lean Six Sigma, Pareto charts may show that only one or a few causes lead to a significant proportion of problems.

```
  "GW1516", "Oxandrolene", "GW1516", "GW1516", "Clomiphene", "Clomiphene",
  "GW1516", "Turinabol", "GW1516", "GW1516", "GW1516", "RAD140", "GW1516",
  "GW1516", "Stanozolol", "Drostanolone", "GW1516", "Clomiphene",
  "GW1516", "GW1516", "Ostarine", "S-23", "GW1516", "Clomiphene", "GW1516",
  "Meldonium", "GW1516", "GW1516", "5aAdiol", "Stanozolol", "Testosterone",
  "Drostanolone", "GW1516", "GW1516", "Metenolone", "GW1516", "Boldenone",
  "GW1516", "GW1516", "GW1516")
> df = data.frame(Violation = v) %>%
  count(Violation) %>%
  arrange(-n) %>%
  mutate(Proportion = 100.0 * n / sum(n)) %>%
  select(Violation, Proportion) %>%
  mutate(CumSum = cumsum(Proportion))
> df
      Violation Proportion    CumSum
1        GW1516  55.555556  55.55556
2    Clomiphene   8.888889  64.44444
3  Drostanolone   4.444444  68.88889
4     Meldonium   4.444444  73.33333
5    Stanozolol   4.444444  77.77778
6       5aAdiol   2.222222  80.00000
7     Boldenone   2.222222  82.22222
8    Metenolone   2.222222  84.44444
9   Methenolone   2.222222  86.66667
```

```
10     Ostarine   2.222222  88.88889
11  Oxandrolene   2.222222  91.11111
12       RAD140   2.222222  93.33333
13         S-23   2.222222  95.55556
14 Testosterone   2.222222  97.77778
15    Turinabol   2.222222 100.00000
> df %>% ggplot(aes(x = reorder(Violation, -Proportion))) +
    geom_bar(aes(weight = Proportion)) +
    geom_line(aes(y = CumSum, group=1)) +
    xlab("Drug Violation") +
    ylab("Proportion")
```

2.6 Box Plots

A box plot splits data into *quartiles*, where each quartile contains 25% of the observations, and represents the spread of each quartile with a "box and whisker." Box plots are only useful with numerical data. The box is centered at the *median* (sequentially middle) value.

An *interquartile range* (IQR) is the values at the 25th and 75th *percentile* of the values. In the Motor Trend Cars data set there are $n = 32$ values. The first quartile is the first eight values, the second quartile is the second eight values, the third quartile is the third eight values, and the fourth quartile is the remaining eight values. If there are an even number of values, then the boundary between *quantiles* (the specific values marking the boundaries of the quartiles; note the difference in spelling) is taken from the midpoint.

```
> sort(mtcars$mpg)[8:9]
[1] 15.2 15.5
> sort(mtcars$mpg)[24:25]
[1] 22.8 22.8
> quantile(mtcars$mpg)
    0%    25%    50%    75%   100%
10.400 15.425 19.200 22.800 33.900
> 22.800-15.425
[1] 7.375
> IQR(mtcars$mpg)
[1] 7.375
```

When using a boxplot, we traditionally define *outliers*[5] as any value that is 1.5 IQRs below the first quantile or 1.5 IQRs above the third quantile.

$$\text{Outliers}\,(X) = \{x \in X | x < \text{Q1} - 1.5\text{IQR} \vee x > \text{Q3} + 1.5\text{IQR}\}$$

[5]We will see an alternative definition for outliers in section 4.3.

By this definition, the `mtcars` data set contains one outlier in the `mpg` (miles per gallon, a measure of fuel efficiency) column.

```
> q1 = quantile(mtcars$mpg)[2]
> q3 = quantile(mtcars$mpg)[4]
> iqr = IQR(mtcars$mpg)
> subset(mtcars, mpg < q1 - 1.5*iqr | mpg > q3 + 1.5*iqr)
                mpg cyl disp hp drat    wt qsec vs am gear carb
Toyota Corolla 33.9   4 71.1 65 4.22 1.835 19.9  1  1    4    1
```

R provides `boxplot` function to render box plots. Return to the https://webr.r-wasm.org/latest/ site and experiment with this function.

```
> boxplot(mtcars$mpg)
```

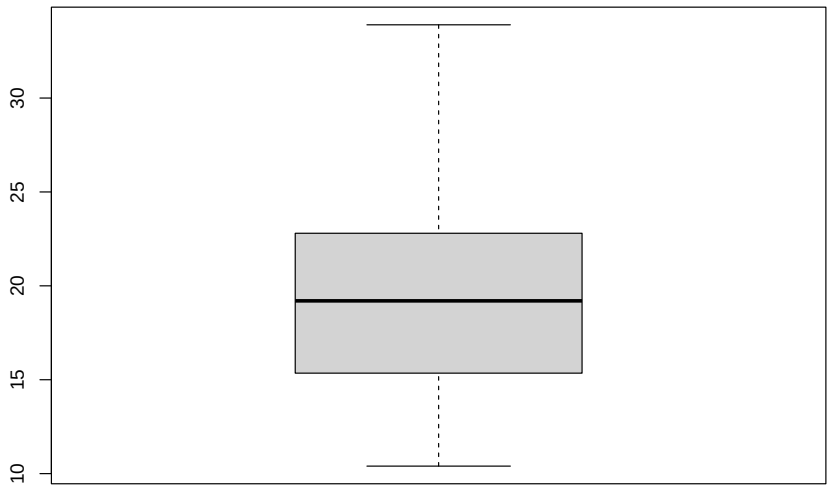


Figure 2.5: R's `boxplot` function creates box-and-whisker plots with four quartiles. Box plots are centered at the median of the data. Outliers may be shown as dots or circles beyond the 0th and 100th percentile markers.

2.7 Histograms

```
> hist(mtcars$mpg)
```

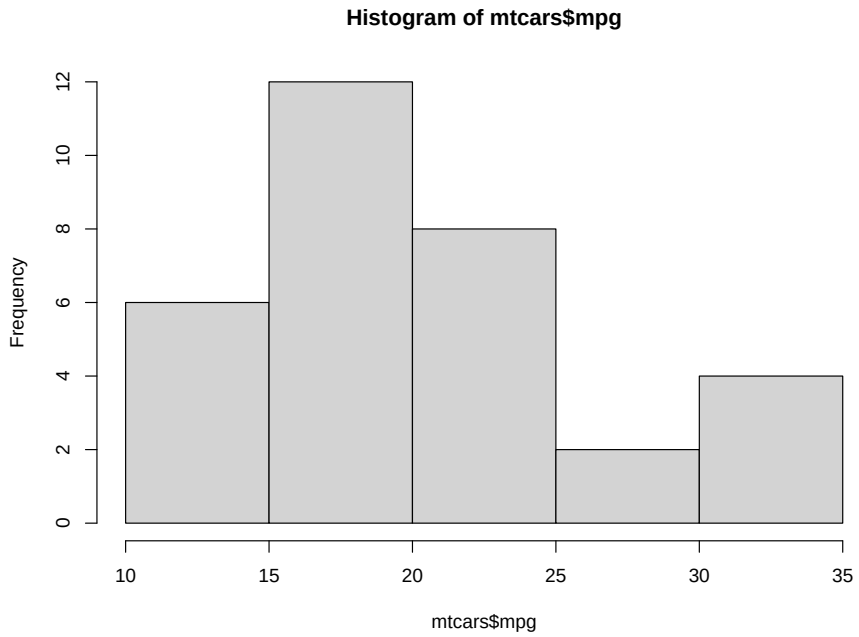


Figure 2.6: todo

2.8 Heat Maps

A *heat map* (sometimes written *heatmap*) is a plot that uses a continuous range of colors to overlay values onto a two-dimensional plot. Heat maps are especially useful for overlaying information onto geographical maps. Figure 2.7 demonstrates a map of Europe where countries are colored by military expenditures as a fraction of their economies.

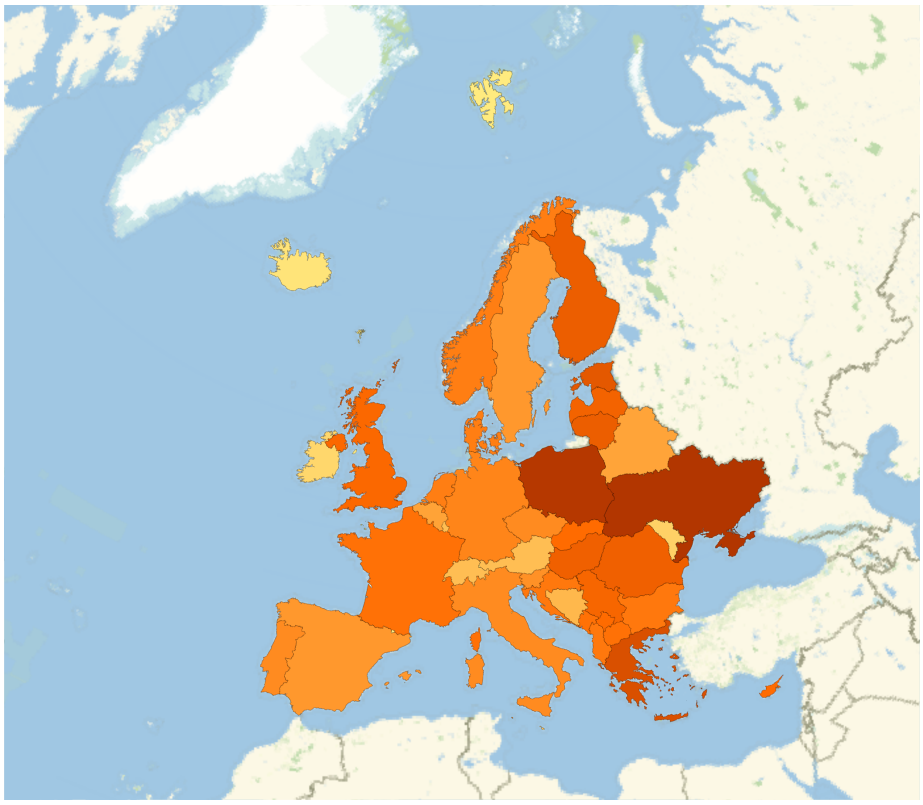


Figure 2.7: A heat map showing military expenditures in European countries. This plot was created in Wolfram Mathematica with input `GeoRegionValuePlot[Flatten[{CountryData[#, "Polygon"] → CountryData[#, "MilitaryExpenditureFraction"]} & /@ CountryData["Europe"]]]`.

2.9 Linear and logarithmic scales

Scientists use the term *order of magnitude* to compare values only by the power of 10. One would say $a = 1.6 \times 10^3$ is three orders of magnitude smaller than $b = 8.3 \times 10^6$, which is to say $b/a \approx 1000$.

The *scale* of an axis, such as in bar plot, is the spacing between values. A *linear*

scale might show marks at 10, 20, 30, 40, and so on. A *logarithmic scale* might show marks at 10, 100, 1000, 10 000, and so on.

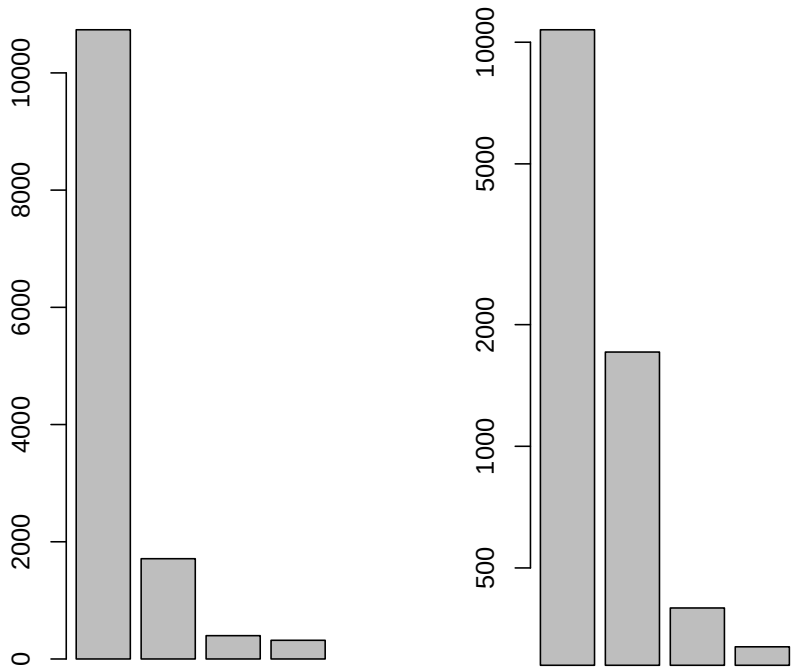


Figure 2.8: These two bar plots show the same data using different scales. The left plot uses a linear scale, where successive marks have a constant *additive* distance. The right plot uses a logarithmic scale, where successive marks have a constant *multiplicative* difference. A logarithmic scale is useful when values differ by orders of magnitude, as the large values obscure differences among the smaller values. Observe that the third and fourth values appear nearly the same on a linear scale, but are clearly different on a logarithmic scale.

Logarithmic scales can be useful for comparing values that differ by more than one order of magnitude. For example, suppose feature of a data set contains categories a, b, c, and d, and the count of each category is

Category	Count
a	10 736
b	1711
c	398
d	319

Return to https://webr.r-wasm.org/latest/ and plot this data with linear and logarithmic scales:

```
> category_counts <- c(10736, 1711, 398, 319)
> category_counts
[1] 10736  1711   398   319
> barplot(category_counts)
> barplot(category_counts, log="y")
```

These plots are shown in figure 2.8.

2.10 Logarithms and exponentiation

A logarithm is the inverse of exponentiation. If

$$a^b = \underbrace{a \times a \times a \times \cdots \times a}_{b \text{ terms of } a} = c,$$

then

$$\log_a c = b.$$

In this case, a is the *base* of the logarithmic, and we read $\log_a c$ as "the log, base a, of c."

Euler's constant, $e \approx 2.718281828459045$, is frequently associated with exponential and logarithmic functions. When the base of a logarithm is equal to e, we call this the *natural logarithm* and express it as

$$\ln x = \log_e x.$$

Plots of e^x and $\ln x$ are shown in figure 2.9.

2.11 Relationships

Figure 2.10 shows exponential and quadratic functions e^x and x^2 on logarithmic scales. The exponential function forms a straight line when plotted this way,

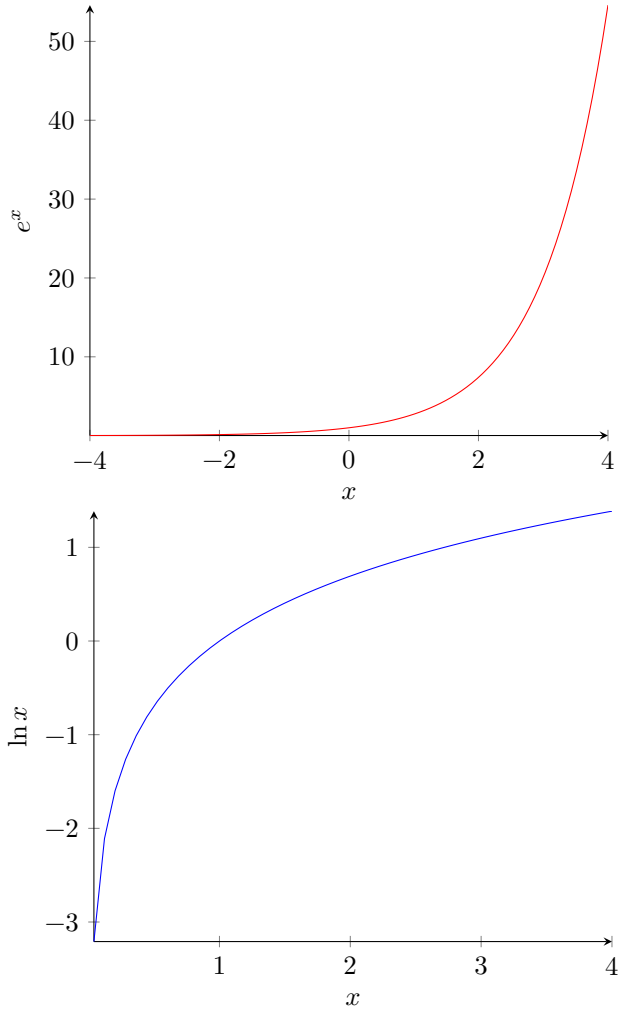


Figure 2.9: The exponential function, e^x, models iterated multiplication and grows quickly. The domain (allowable inputs) for e^x are all real numbers, but the range (possible outputs) are strictly positive reals. It is not possible for e^x to produce a zero or negative output if x is a real number. The logarithmic function, $\ln x$, inverts exponentiation and grows very slowly. The domain and range for $\ln x$ are the reverse of e^x: the domain of $\ln x$ is the positive reals and range is all reals.

but the quadratic does not. This is because $\ln e^x = x$, but $\ln x^2 = 2 \ln x$, which means the plot of x^2 reveals the same slow growth that we saw in figure 2.9. If the exponential function had a base b other than e, then

$$\ln b^x = x \ln b,$$

where $\ln b$ is a constant factor and observable as the slope of the resulting plot.

In his 1954 article *Relation between Weight-Lifting Totals and Body Weight*, Lietzke claims that an athlete's "weight-lifting ability should be proportional to the two-thirds power of the body weight" [9]. Lietzke scales both the x and y axes with a logarithm, where x and y respectively represent bodyweight and weightlifting total. The resulting log-log plot shows straight line with a slope of $0.67 \approx 2/3$ indicates that if

$$\log \text{strength} \propto \frac{2}{3} \log \text{bodyweight}$$

then[6]

$$\text{strength} \propto \text{bodyweight}^{2/3}.$$

Changing the scale on a plot can be a simple but powerful method to develop intuition for the shape of the data. However, one should be cautious of over-generalization. In section 2.13, we will see misleading shape in the plot of a logistic curve, but first we must explain the sigmoid curve in section 2.12.

2.12 Sigmoid Curves

The *sigmoid* function, $\sigma(x)$, can be used to model a system characterized by competing exponential growth and decay[7]. That is, the sigmoid represents a system with limited resources.

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}$$

An example from epidemiology is the spread of a contagion within a population. Initially, very few individuals have the disease, but the rate at which the disease spreads quickly increases as the number of infected members compounds. At the same time, however, the probability that another individual is already infected

[6]The symbol $\propto$ means "is proportional to."

[7]The letter σ has many meanings in mathematics and statistics. In section 4.3, we will introduce variance and standard deviation, which use the symbols σ^2 and σ. Even well-known symbols, such as π and e, have overloaded meanings in this field. One must take care to disambiguate meanings using prose.

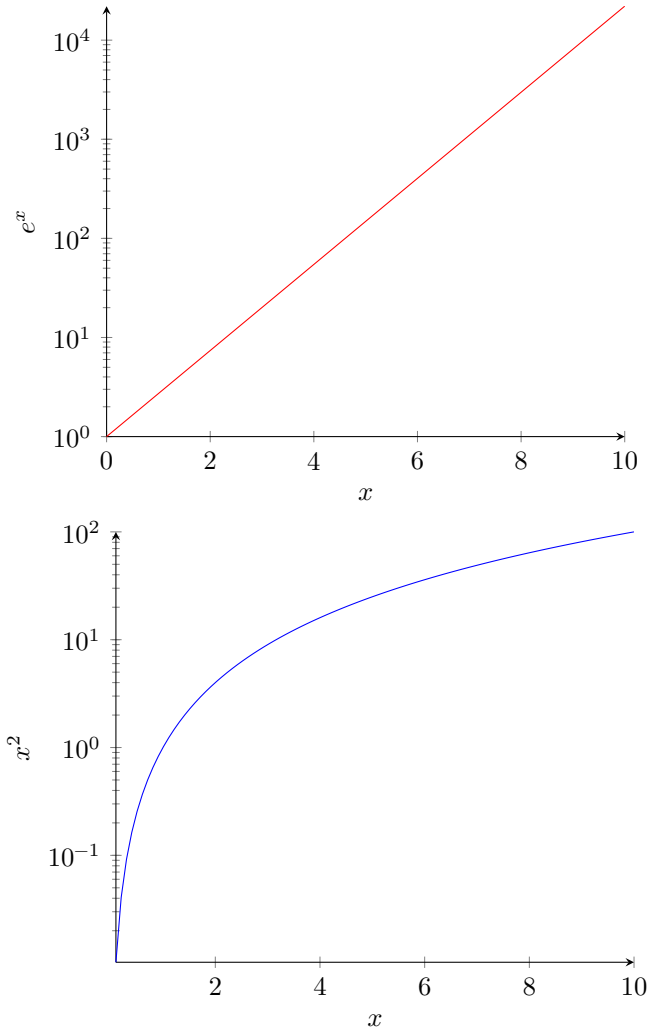


Figure 2.10: The exponential function, e^x, forms a straight line when plotted on a logarithmic scale, as $\ln e^x = x$. By contrast, the quadratic function, x^2, does not form a straight line when plotted on a logarithmic scale. Plotting a fast-growing data series on a log scale is a quick and easy way for the data scientist to "feel" if the curve might fit an exponential behavior or not.

or can resist the contagion also increases, slowing the spread as we reach some *inflection point*, as shown in figure 2.11.

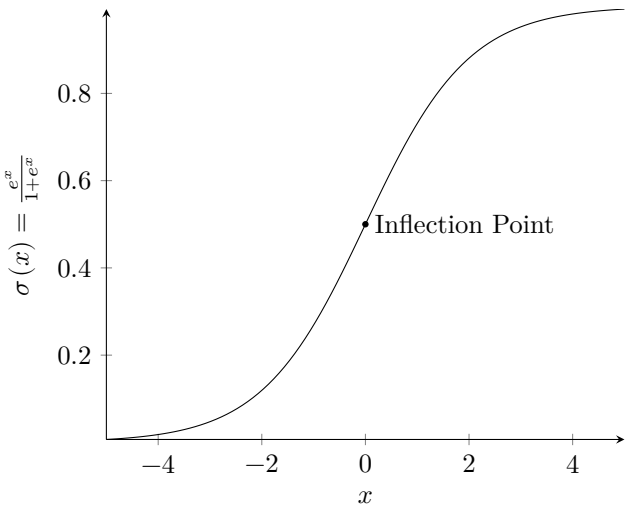


Figure 2.11: todo sigmoid.

2.13 Logistic Curves

The *logistic* function[8] is a parameterized sigmoid function of the form

$$\frac{L}{1 + e^{-k(x-x_0)}}.$$

Figure 2.12 shows a logistic function

$$l(x) = \frac{100}{1 + 2.75e^{-0.4x}}$$

plotted over the domain $0 \leq x \leq 5$. The parameters and domain are chosen carefully to provide a confusing plot. The curve in figure 2.12 forms a mostly straight line. Data sampled from this narrow range might fit a linear model with very little error, but of course this is only because we have zoomed into the center of the curve.

Consider a fad in the world that starts very small, but quickly spreads in popularity as network effects cascade into increased awareness. One might call this a "trend," and the trend line may initially appear linear or even exponential, but as the fad grows so too might shortages, opposition, or regulation slow its growth.

[8] [10]

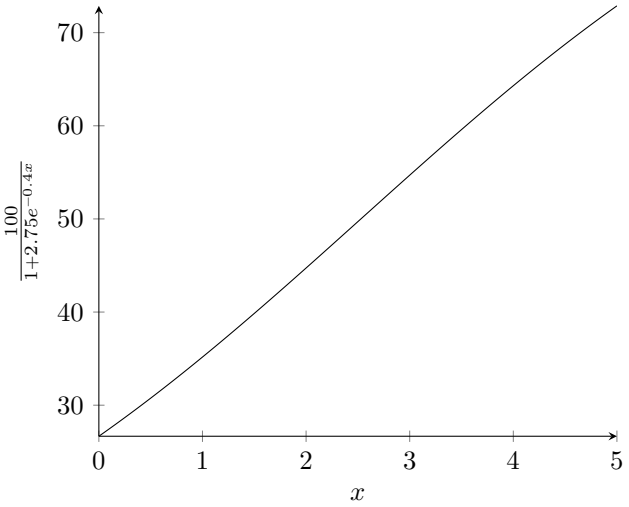


Figure 2.12: todo logistic.

2.14 Discussion prompts

1. Take a close look at the CrossFit.com data in section 2.5 and ask whether this "GW1516" substance is actually the most commonly *abused* substance or whether it is the most commonly *detected* substance. Think of other situations where problems may be over- or under-represented due to the sensitivity of a test.

2. Like a bar plot, a pie chart shows the relative sizes of categorical values. What are some advantages and disadvantages of using pie charts?

3. What are some plot practices, such as inconsistent scales, that would be misleading to the reader?

4. Consider a situation where the sigmoid and logistic curves might reasonably model constrained exponential growth. If one only observes the center of this system, then the slow initial growth and diminishing returns might not be visible in a scatter plot of the data. Discuss graphical and analytical methods one might use to predict the future behavior of the uncertain system.

5. In addition to numerical grades, a teacher wants their students to know their relative standing in comparison to their peers. The teacher wants to minimize how much information students can infer about their classmates, although it is desirable for students to know the central (mean, median, or mode) grades. Which plot technique is better for this task: a bar plot or a box plot?

2.15 Practical exercises

1. Given a dataset, plot the data and explain why this plot technique is appropriate.
2. Be creative and construct intentionally misleading plots, then try to "spot the flaw" in one another's work.
3. Plot our logistic function from section 2.13, $l(x) = \frac{100}{1+2.75e^{-0.4x}}$, on a logarithmic scale and manipulate the domain. Does the logistic function still look linear on a logarithmic scale?
4. Spot the flaw in figure 2.13.

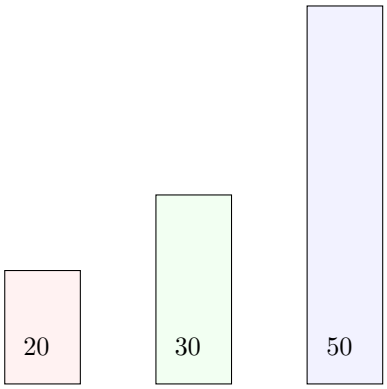


Figure 2.13: These bars represent values 20, 30, and 50.

5. Spot the flaw in figure 2.14.

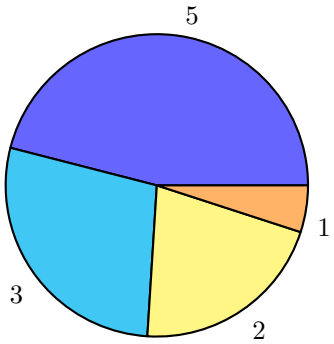


Figure 2.14: These bars represent values 5, 3, 2, and 1.

Chapter 3

Data Operations

3.1 Prose

In this chapter, we will discuss several practical matters for operating on data to extract useful information, but first we should quickly discuss the mechanics of mixing data, code, plots, mathematical notation, and tables, and prose.

Context is king. One must consider the target audience when writing reports from any analysis.

Presentations, such as with Microsoft PowerPoint, are useful as visual aids to speeches. The slides themselves should contain mainly plots, sparse text, and simple tables to summarize information. Slides are a poor medium for presenting raw data, large tables, code, and long passages of prose. Pity statements, organized as bullet points, can be useful for both the speaker and the audience, but full sentences are often not recommended.

Papers are favorable to slides for deep analysis. Papers vary in length. Summaries are short and generally seek to report conclusions without detailed evidence. A senior leader may accept the conclusions presented in an employee's report based on trust in the person, not the persuasiveness of the analysis itself. White papers provide enough evidence to be persuasive on their own merit, although white papers may not provide detailed listings of the data and code used.

The sciences use *notebooks* as a means of presenting prose with in-line code, plots, mathematical notation, and tables. Some examples of notebook interfaces are Jupyter[1] (commonly used with Python, R, Julia, and Scala), Mathematica[2] (the Wolfram Language), and RStudio[3] (R and Python). Notebook interfaces

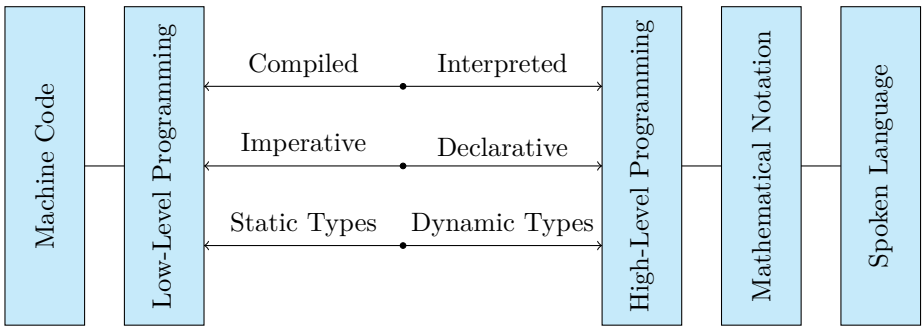
[1] https://jupyter.org
[2] https://www.wolfram.com/mathematica/
[3] https://posit.co/products/open-source/rstudio/

support *literate programming*, a practice of writing code with an emphasis on human understanding above computer interpretation [11].

Written prose and spoken presentation are key to aggregating and processing data into information and then interpreting information into knowledge. Jeff Bezos famously insists upon the use of six-page narratives at Amazon in favor of PowerPoint[4]. The rest of this chapter will focus on technical matters of working with data, but look at how data, code, figures, and mathematical notation are presented throughout. Reflect upon how these may or may not be appropriate when writing and presenting information, depending on format.

3.2 Computability

A joke in computer science says that "you can write C in any language." The joke is literally true. Assuming adequate resources (compute time, memory, storage, and access to necessary inputs and outputs), one could implement a C interpreter in any *Turing-compute* language [12, p. 13] – 17 and execute any C program. Such an endeavor is not theoretical: *virtual machines* and related technologies simulate and emulate entire computing machines, allowing programs to run on systems that they were not designed for.



Programming languages are imprecisely categorized as *low-level* and *high-level*. One should view these terms as a spectrum, not dichotomies. Low-level languages generally require more explicit specification to the machine, allowing for greater control of the computation and often faster. The abstractions available in high-level languages often allow the programmer to code with syntax closer to mathematical notation. Computing in spoken language (or "natural language") has historically failed to satisfy the high expectations popularized in science fiction [13], although recent advances in AI have steadily improved machines' ability to compute results from spoken or written prompts.

Low code platforms, such as Microsoft's Power BI[5], seek to democratize pro-

[4] https://www.youtube.com/watch?v=L227qFemjqI
[5] https://www.microsoft.com/en-us/power-platform/products/power-bi

gramming to non-developers [14]. Low code platforms provide functions to process data into information with prepared analytical and visualization features, generally using minimal programming languages or even visual programming.

Low code platforms can excel at tasks that they were designed for but complicate novel tasks. This tension mirrors the conflict of *narrow* and *general* AI systems. Users trust AI systems, such as Apple's Siri[6], to correctly respond to very specific tasks ("hey Siri, what's the weather tomorrow?") but do not expect these systems to generalize to vague or contextual queries ("hey Siri, recommend a movie for me").

AI systems with natural language input will no doubt continue to advance in the coming years, but we will always need to understand the general methods used to structure, process, transfer, and store our data in novel situations. Where no one has previously described an algorithm to solve new problems, we will likely always turn to code as a reification of our mathematical ideas.

We say that a programming technique is *idiomatic* when the code follows the conventions of the language. Languages often provide features that make non-idiomatic constructs possible, but possibly slow or brittle. Code may be considered not idiomatic because it fails to use a language feature or because it extends the language to provide a feature that was intentionally not implemented. For example, one might implement a matrix in Python as

```
In [1]: [[1,0,0],[0,1,0],[0,0,1]]
Out[2]: [[1, 0, 0], [0, 1, 0], [0, 0, 1]]
```

However, the resulting object is a `list` that does not guarantee contiguous memory layout. Numerical operations with such objects will be significantly slower than the `matrix` object in the Numpy library[7].

```
In [2]: import numpy

In [3]: numpy.matrix([[1,0,0],[0,1,0],[0,0,1]])
Out[3]:
matrix([[1, 0, 0],
        [0, 1, 0],
        [0, 0, 1]])
```

The Python community even has a peculiar term for idiomatic Python code: "pythonic." When learning new programming languages, one should be aware that the new language is much more than new syntax for familiar operations. If an operation feels more tedious or difficult in one language than another, this can be an indication that the new language has different structure. The new language will still compute the computable program, but it may require a changed approach.

[6] https://www.apple.com/siri/
[7] https://numpy.org

3.3 The Relational Algebra

Codd's *relational algebra* is the framework theory describing all modern *database management systems* (DBMS) [15]. The relational algebra can be described with five primitives: *selection* (σ), *projection* (π), the *Cartesian product* ($\times$; also known as the *cross product*), set *union* ($\cup$), and set *difference* ($-$).

Selection takes all or a subset of a table's rows. Projection takes all or a subset of a table's columns. In Structured Query Languages (SQL), selection is specified in the `WHERE` clause and projection is specified in the list of columns immediately after `SELECT`.

A Cartesian product is the multiplication of sets. If $A = \{i, j\}$ and $B = \{x, y, z\}$, then $A \times B = \{(i, x), (i, y), (i, z), (j, x), (j, y), (j, z)\}$. The Cartesian product produces the set of all possible pairwise combinations of elements in each set. These composite values are called *tuples*. Tuples may contain more than two values. If $C = \{c\}$, then

$$A \times B \times C = \{(i, x, c), (i, y, c), (i, z, c), (j, x, c), (j, y, c), (j, z, c)\}.$$

As an exercise, go to the SQLime Playground[8] to use a DBMS named SQLite. Enter the following commands to reproduce the above Cartesian product.

```
CREATE TABLE A (a text);
CREATE TABLE B (b text);
CREATE TABLE C (c text);

INSERT INTO A(a) VALUES ('i'), ('j');
INSERT INTO B(b) VALUES ('x'), ('y'), ('z');
INSERT INTO C(c) VALUES ('c');

SELECT * FROM A CROSS JOIN B CROSS JOIN C;
```

This text views tuples as unordered and "flattened" sets, and therefore Cartesian products are both *commutative* ($R \times S = S \times R$) and *associative* ($R \times (S \times T) = (R \times S) \times T$). Some mathematical texts use a stricter definition for the Cartesian product where the result is a set, which does not "flatten" and therefore provides neither commutativity nor associativity. This text uses the looser definition for compatibility with practical DBMSs, including SQLite. (Mathematics is partly discovered and partly invented.)

Set union, $\cup$, combines two sets. Sets definitionally contain only distinct elements. If $A = \{i, j, k\}$ and $B = \{k, l, m\}$, then

$$A \cup B = \{i, j, k, l, m\}.$$

[8] https://sqlime.org/#deta:mb9f8wq2mq0b

Set difference, $-$, retains the elements of the left set that are not present in the right set.

$$A - B = \{i, j, k\} - \{k, l, m\} = \{i, j\}\,.$$

3.4 Joining Tables

The *join* ($\bowtie$) is a combination of the Cartesian product and selection. For example, suppose we have a tables named `Swim`, `Bike`, and `Run`. Each table has a column that uniquely identifies an athlete. To get a triathletes (the athletes who participate in swimming, cycling, and running), we use an *equijoin* to find the product where the names are equal. Return to the SQLime Playground[9] to demonstrate experiment with the `JOIN` operator.

```
CREATE TABLE IF NOT EXISTS Swim (sn TEXT UNIQUE);
CREATE TABLE IF NOT EXISTS Bike (bn TEXT UNIQUE);
CREATE TABLE IF NOT EXISTS Run (rn TEXT UNIQUE);

INSERT OR IGNORE INTO Swim (sn) VALUES
    ('John'), ('Jane'), ('Luke'), ('Phil');
INSERT OR IGNORE INTO Bike (bn) VALUES
    ('Mary'), ('Alex'), ('Jane'), ('Levi');
INSERT OR IGNORE INTO Run (rn) VALUES
    ('Mike'), ('John'), ('Jane'), ('Sven');

SELECT * FROM Swim, Bike, Run WHERE sn = bn AND sn = rn;
```

There are other syntaxes which achieve the same result using the `ON` and `USING` clauses. As an exercise, try to predict how many rows will return from `SELECT * FROM Swim, Bike, Run` without a `WHERE` clause.

3.5 Grouping and Aggregation

DBMSs provide robust *grouping* functions for operating on related rows. Return to the SQLime Playground[10] and create a small table of hypothetical marathon times.

```
CREATE TABLE IF NOT EXISTS Marathon (rn TEXT UNIQUE,
  time INTEGER,
  gender TEXT CHECK( gender IN ('M', 'F') ));

INSERT OR IGNORE INTO Marathon (rn, time, gender) VALUES
  ('Kyle', 2*60*60 + 14*60 + 22, 'M'),
```

[9] https://sqlime.org/#deta:36fadcq9apak
[10] https://sqlime.org/#deta:32lpfoo57r8g

```
  ('Hank', 2*60*60 + 10*60 + 45, 'M'),
  ('Lily', 2*60*60 + 24*60 + 47, 'F'),
  ('Emma', 2*60*60 + 22*60 + 37, 'F'),
  ('Elle', 2*60*60 + 25*60 + 16, 'F'),
  ('Fred', 2*60*60 + 6*60 + 17, 'M');

SELECT MIN(time) FROM Marathon GROUP BY (gender);
```

`MIN` is one of the *aggregate functions* in SQLite. The `GROUP BY` clause tells the DBMS to split the rows into groups on the `gender` column.

One might be tempted to find the names of our male and female champions with `SELECT rn, MIN(time) FROM Marathon GROUP BY (gender)`. This may work in some DBMSs but there is a subtle bug. It might be obvious that we want the `rn` associated with the `MIN(time)` value, but suppose we change the query to also include `MAX(time)`:

```
SELECT rn, MIN(time), MAX(time) FROM Marathon GROUP BY (gender);
```

Now it is no longer clear which `rn` the query should return. Should the DBMS return the `rn` associated with the `MIN(time)`, the `MAX(time)`, or some other `rn` from the group?

The solution in this particular case is to nest our `MIN(time)` aggregation as a *subquery*.

```
SELECT * FROM Marathon
  WHERE time IN (
    SELECT MIN(time) FROM Marathon GROUP BY (gender));
```

Taking aggregates from aggregates can produce different statistics from those of the original data set. Consider the election of choices A and B by 100 voters as shown in figure 3.1. In elections, the winner may lose the popular vote, as aggregated district votes do not reflect the density within their groups. Aggregation is generally a *lossy* process, where the inputs cannot be reconstructed from the information it produces [16].

The apparent reversal of votes in figure 3.1 is related to *Simpson's Paradox* [17]. TODO: say more about this.

SQL uses the *declarative programming* paradigm, where the language is used to describe the *result* that the user[11] wants while leaving the implementation details to the DBMS. Systems designed for declarative programming often excel in situations that the developer intended but sometimes struggle when the user needs something unusual. For situations where the user needs to specify the detailed process to compute the result, we use the *imperative programming* paradigm. Two specific imperative approaches are *functional* and *object-oriented*

[11] In this context, the "user" is a programmer or data analyst who is "using" the database or programming language

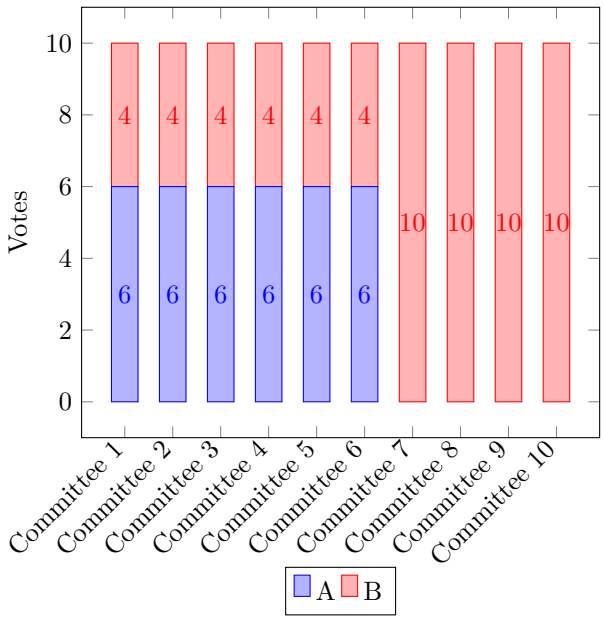


Figure 3.1: This plot shows 100 votes grouped into 10 committees. If each committee is given only one vote, then decision A will receive more committee votes than decision B, having lost the *popular vote* with only 36 votes.

programming. In practice, the distinctions are often blurred by languages and databases that provide functionality from all three.

3.6 Functional Programming

SQL's declarative syntax makes it easy to write select, project, and join (SPJ) queries. SQL grouping and aggregate functions make it possible to perform row-wise and column-wise operations. A *functional* programming language, which emphasizes the use of pure functions (see section 1.13) to express algorithms, provides comparable semantics in the *filter*, *map*, and *reduce* functions.

3.6.1 Filter

Filter works much like the `WHERE` clause: it takes a subset of the rows, based off of a condition. In JavaScript, we might filter an array with:

```
>> ['cat', 'dog', 'fish', 'bird'].filter(v => v.includes('i'))
<- ['fish', 'bird']
```

3.6.2 Map

Map performs the same function over each element of an input set, creating "mappings" to elements of an output set.

```
>> ['fish', 'bird'].map(v => v.toUpperCase())
<- ['FISH', 'BIRD']
```

3.6.3 Reduce

Reduce, also known as *fold*, performs some operation on each element of an input set and returns an *accumulator*, which is passed again to the reduce function with the next input value. To take an array's sum, we use an initial accumulator value of 0.

```
>> 15 + 25 + 35
<- 75
>> [15,25,35].reduce((a, v) => a + v, 0)
<- 75
```

For the array's product, we use 1 for the initial accumulator value.

```
>> 15 * 25 * 35
<- 13125
>> [15,25,35].reduce((a, v) => a * v, 1)
<- 13125
```

Both filter and reduce can be implemented in terms of reduce.

```
>> ['cat', 'dog', 'fish', 'bird'].reduce((a,v) => {
        if (v.includes('i')) {
            a.push(v);
        }
        return a;
    }, [])
<- ['fish', 'bird']
>> ['fish', 'bird'].reduce((a,v) => {
        a.push(v.toUpperCase());
        return a;
    }, [])
<- ['FISH', 'BIRD']
```

In both cases, we use an empty array (`[]`) instead of a numeric identity as our initial accumulator value.

Some languages differentiate `foldl` and `foldr` to differentiate left- and right-associativity. A left-associative function would evaluate $x{\sim}y{\sim}z$ with first $x{\sim}y$ and then $(x{\sim}y){\sim}z$. (In this context, the "~" represents an arbitrary infix operator and has no specific meaning). A right-associative function evaluates $x{\sim}y{\sim}z$ as $x{\sim}(y{\sim}z)$.

3.6.4 Vectorized Functions and Array Programming

A *vectorized function* automatically iterates over array inputs. This approach is related to *array programming*. Automatic vectorization is less common in traditional languages (C, Java, JavaScript) and more common in scientific programming (R, Matlab, Julia). Some examples in the R language, which one can reproduce at https://webr.r-wasm.org/latest/, are:

```
> c(1, 2, 3) + 4
[1] 5 6 7
> c(1, 2, 3) + c(4, 5, 6)
[1] 5 7 9
> sqrt(c(1, 4, 9))
[1] 1 2 3
```

Observe that the pairwise sums in `c(1, 2, 3) + c(4, 5, 6)` are independent. No sum depends on another, and therefore the computing machine can safely perform each operation in *parallel*.

3.6.5 Immutability

Suppose one needs to write a program to sort its input. An obvious solution is to order the inputs directly by *mutating* (changing) the memory in-place. An alternative approach is to copy the input, order the copy, and return the ordered copy.

The Julia language provides both: Julia's `sort!` function mutates its input, while the `sort` function returns a sorted copy, leaving the input unmodified.

The latter approach obviously uses more memory and will likely be slower. Why would one use this approach? **Safety**. If a function "owning" a variable passes an *immutable* (read-only) reference to another function, then the caller can safely reason about the value and state of that variable after the callee returns.

Some languages provide stronger concepts of ownership and immutability than others. The Rust language provides extensive memory safety features [18] by requiring the `mut` keyword to explicitly mark variables and function parameters mutable. Traditionally, languages assumed the opposite and required `const` or `final` keywords to establish invariants (with varying levels of enforcement; Java programmers might be surprised that the `final` keyword does not make an object read-only, but only the *reference to* an object).

Immutability is particularly useful for *thread-safety* in concurrent programming, which we will discuss in section 3.9.

3.7 Object-Oriented Programming

Object-Oriented Programming (OOP) is a technique for modeling both the *data* and associated *code* for a problem together [19] [20]. The data of an object are called *fields* and the code are called *methods*. The specification of the fields and methods of an object is called a *class*. Many programming languages, notably C++, Python, JavaScript, and Python, emphasize OOP as the central design.

Object-orientation comes in many varieties [20]. Many OO languages provide a method inherit data and code from other objects, often in a hierarchy. The following Rust program, which one can run at the Rust Playground[12], demonstrates a `Point` object. The object is defined as a `struct` with two fields, `x` and `y`. The implementation for `Point` adds two methods, a *constructor* (`new`) and a *Manhattan distance* function.

```
struct Point {
    x: f32,
    y: f32,
}

impl Point {
    fn new(x: f32, y: f32) -> Self {
        Self { x, y }
    }

    fn manhattan_distance(&self, other: &Self) -> f32 {
        (self.x - other.x).abs() + (self.y - other.y).abs()
```

[12] https://play.rust-lang.org/?gist=9542264fd12645a4ee1956ab7f890812

```
    }
}

fn main() {
    let a = Point::new(2.0, 4.0);
    let b = Point::new(-2.0, 3.0);
    println!("Distance: {}", a.manhattan_distance(&b));
}
```

Several languages now have special objects that contain only data. In Python, these data-only classes are called *data classes*[13]. In Java, they are called *records*[14]. These data-only classes have several limitations but generally reduce the "boiler-plate" code needed to instantiate, mutate, display, and compare these objects.

3.8 JavaScript Object Notation (JSON)

We introduced CSV in section 1.8 as a method for representing data in a file. *JavaScript Object Notation* (JSON) is an alternative format [21]. JSON's syntax is based on JavaScript. Objects in JSON are key-value pairs. The key of a JSON object must be a double-quoted string. Values can be nested objects, arrays, numbers, strings, Booleans, and `null`. The process of taking a data structure from a program's memory and saving it as JSON is called *serialization*. The inverse, reading an object into memory from a JSON input, is correspondingly *deserialization*. Trailing commas are forbidden.

JSON is much more verbose than CSV, but less verbose than the Extensible Markup Language (XML), which we will not discuss further. JSON is generally "human-readable" and can be authored by hand, although not as easily as CSV. The following code listing shows the table from section 1.8 as JSON.

```
[
    {
        "x": "Rob",
        "y": 0.74019382956651820,
        "z": 0.3508759018489489
    },
    {
        "x": "John",
        "y": 0.41331428270607506,
        "z": 0.2936926427452584
    },
    {
        "x": "David",
        "y": 0.37671743737357277,
```

[13] https://docs.python.org/3/library/dataclasses.html
[14] https://docs.oracle.com/en/java/javase/17/language/records.html

```
        "z": 0.5676190157838865
    },
    {
        "x": "Frank",
        "y": 0.50270122376380740,
        "z": 0.7939268929144455
    }
]
```

JSON documents allow arbitrarily nested and shaped objects, but in many applications it may be undesirable to deserialize records with missing values. Consider if one of the below records were missing a y-value, or if a z value were incorrectly enclosed in double-quotes, thus forming a string instead of a numeral.

Some JSON parsers, such as Rust's Serde[15], allow the programmer to specify the structure of the record before parsing. Libraries may ignore or error when records do not fit the expected shape. One can expect statically-typed languages to require more specification before parsing and dynamically-typed languages to allow greater flexibility at runtime (see section 1.9).

3.9 Parallelism and Concurrency

Parallelism is the ability for a computing machine to perform simultaneous operations. Two tasks are *concurrent* if their order does not matter. Getting dressed in the morning is an example (see figure 3.2). When one dons their pants, shirt, coat, hat, socks, and shoes, one must don socks before shoes, but the order in which one dons shoes and their hat does not. The hat and shoes are concurrent but the socks and shoes are *sequential*. In practice, many programmers confuse the terms *parallel* and *concurrent* as interchangeable.

Concurrent programming can be challenging because one *process* or *thread* (sometimes called *task* or *routine*) might interfere with another, but performance benefits often justify the additional complexity.

Some problems can be partitioned into *subproblems* which can be solved in parallel. Other problems cannot. Some encryption algorithms intentionally *chain* the output from one block into the next. One cannot calculate block n without first calculating block $n-1$, and $n-2$, and so on. The reduce operation applies to this algorithm design.

Other problems can be effortlessly partitioned into subproblems and solved quickly with a *divide-and-conquer* approach. A trivial example might be finding the minimum value in a large dataset. One can partition the dataset, find the minimum value in each partition, and then find the minimum value among those results. This process can be repeated.

[15] https://serde.rs

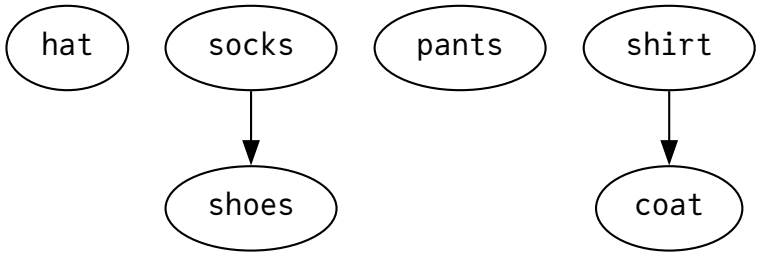


Figure 3.2: Order of operation partially matters when getting dressed. Some clothing items are sequential, but others are concurrent. The system can be modeled as a directed ayclic graph (see section 6.3).

Go to the Go Playground[16] to experiment with a divide-and-conquer `minimum` function in the Go language.

```
package main

import "fmt"

func min(x, y int) int {
    if x <= y {
        return x
    }
    return y
}

func minimum(x []int) int {
    fmt.Println(x)
    n := len(x)
    switch n {
    case 1:
        return x[0]
    case 2:
        return min(x[0], x[1])
    default:
        middle := n / 2
        lower := minimum(x[:middle])
        upper := minimum(x[middle:])
```

[16] https://go.dev/play/p/IOwH08R_z7Z

```
        return min(lower, upper)
    }
}

func main() {
    fmt.Println(minimum([]int{610, 144, 34, 21, 2584, 55, 55}))
}
```

Click the "Run" button several times and observe that the output is completely *deterministic*. Now return to the Go Playground[17] for a slightly modified version of the same program.

```
    default:
        middle := n / 2
        lower := make(chan int)
        upper := make(chan int)
        go func() { lower <- minimum(x[:middle]) }()
        go func() { upper <- minimum(x[middle:]) }()
        return min(<-lower, <-upper)
    }
```

This version constructs two *channels* for communication among concurrent tasks. We use the `go` keyword to create two *Goroutines* (threads in the Go language), which concurrently solve the `minimum` function over subproblems. Finally, we read the results from each channel with `<-lower` and `<-upper` and return. Click the "Run" button several times and observe that the final result is consistent, but the order of operations is not.

The computer industry has recently turned to *Graphical Processing Units* (GPU) as a fast, inexpensive, and energy-efficient method for solving highly parallelizable problems. GPUs were originally designed to draw computer graphics, which extensively use matrix and vector multiplication. These linear transformations can be performed in parallel and GPU makers designed their products to perform many simple calculations in parallel.

3.10 The CAP Theorem

Brewer's *CAP theorem* states that a *distributed system* has at most two qualities of *consistency*, *availability*, and *partition-tolerance* [22]. Consider a system of databases with many replicas. The replicas are consistent if they contain perfect copies of the database, and they are available only they are writable. The distributed system is partition-tolerant if all replicas remain identical, but this is impossible if one allows writes that cannot propagate into the other partition.

The CAP theorem has many practical implications on data integrity and should be considered in design methodology. One must anticipate server and network

[17] https://go.dev/play/p/Vbe7BWrwlku

outages that would create a partition in the distributed in the system and then choose the desired behavior. Can we accept lost database writes when we reconcile after a partition is restored? Or should be accept service outages in order to protect the integrity of the database during an interruption?

A partial solution is to weaken our definition of each quality. Perhaps a system reserves certain rows or columns that are only writable by a specific database, guaranteeing that there will be no conflict if this database continues to write to those changes during a partition. A system might establish some form of confidence intervals in certain data, such as the position of a tracked aircraft with error margins, in recognition that imperfect information might still be useful. Finally, a system might use a quorum model (i.e., 3 of 5 available nodes) to preserve partial availability in the majority partition.

3.11 Discussion Prompts

1. The Excel function `VLOOKUP(lookup_value, table_array, col_index_num, range_lookup)` searches in a table (`table_array`) for a value (`lookup_value`) and returns the value in the numbered column (`col_index_num`)[18]. If `range_lookup` is true, then `VLOOKUP` uses approximate matching, otherwise exact. Assuming one corrects the SQL syntax, what is the *semantic* difference between `VLOOKUP(x, y, z, FALSE)` and the SQL query `SELECT z FROM y WHERE x`? Can we parameterize the SQL statement to produce the same result as `VLOOKUP`?
2. How does the CAP theorem impact intelligence and fires in relation to the command and control (C2) warfighting function (WfF)?
3. Where should unclassified data be stored and processed?
4. What are some methods to prevent conflicts among concurrent writes in a shared database?
5. What can go wrong when altering database schema?

3.12 Practical Exercises

1. Create a custom list in SharePoint that provides multiple views showing grouped and aggregated values.
2. Given a noisy dataset, identify problems in each column that could influence inclusion and exclusion criteria.
3. Define an "embarrassingly parallel" problem and provide both examples and counterexamples.

[18] https://support.microsoft.com/en-us/office/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1

4. In section 3.6.3 we have examples of the filter and map operations implemented in terms of reduce. Later, in our discussion of immutability in section 3.6.5, we learned that a sorting function can either mutate the data in-place or copy the data, leaving the original data unchanged and returning a new data structure with the desired property. Which design are the filter and map operations in section 3.6.3? Rewrite both functions in the other paradigm.

5. The *flatten* operation promotes elements of nested collections to a single container. A flawed example in Julia is

```
julia> reduce(∪, [[:a,:b], [:c,:a], [:d,:a,:b]])
4-element Vector{Symbol}:
 :a
 :b
 :c
 :d
```

   This one-line solution is a *shallow* flatten. It fails on doubly-nested inputs.

```
julia> reduce(∪, [[:a,:b], [:c,:a], [:d,:a,:b], [[:e]]])
5-element Vector{Any}:
 :a
 :b
 :c
 :d
 [:e]
```

   Implement a *deep* flatten that correctly traverses arbitrarily nested inputs. Test that the deep flatten operation correctly handles empty inputs, nested empty inputs (such as `[[[[]]]]`), and duplicates.

Chapter 4

Measures of Central Tendency

4.1 Least squares method

The canonical definition of the *arithmetic mean* for a set of n numbers x is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_4}{n}.$$

Where does this definition come from? Let's open a new workbook in Microsoft Excel and find out.

Leave A1 blank. This will become our *estimate of the mean*. In fields B1 through B9, enter integers 1 through 9. In fields C1 through C9, enter formulas `=B1-$A$1`, `=B2-$A$1`, `=B3-$A$1`, and so on (keeping `$A$1` fixed). In fields D1 through D9, enter formulas `=POWER(C1,2)`, `=POWER(C2,2)`, and so on. In field E1, enter formula `=SUM(D1:D9)`. Finally in field F1, enter the formula `=AVERAGE(B1:B9)`.

Now go to Data, What-If Analysis, Goal Seek. In the Goal Seek dialog, enter `Set cell:` to `E1`, `To value:` to `0`, and `By changing cell:` to `A1`. Click OK. The Goal Seek function runs and should produce a value in A1 near to that in F1. (Goal Seek is not foolproof. Enter `10` into A1 to nudge Excel with a hint if you get a ridiculous answer.)

We have used Goal Seek to minimize the *sum of the squared differences* between our values, x_i, and our estimate of the mean, $\bar{x}$. This *least squares method* dates back to Carl Friedrich Gauss and Adrien-Marie Legendre in the 1800s [23] [24].

Squaring the errors makes the values positive, which prevents underestimates from negating overestimates. One might also consider the *absolute value* (`ABS`

in Excel) as an alternative, but there is a second reason for squaring the errors. Squaring the errors penalizes large errors more than small errors. Accepting small errors but avoiding large errors is the bias that gives the least squares method its strength.

Readers familiar with calculus may recognize that one can find the arithmetic mean, μ, by finding the zero in the *derivative* for the sum of the squared errors (SSE) function. Let X be a sample of size n.

$$X = \{x_1, x_2, \dots, x_n\}.$$

Then the errors of our estimate of the mean, $\bar{x}$, can be found using the error function

$$\operatorname{Err}(\bar{x}) = X - \bar{x} = \{x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}\},$$

and the sum of the squared errors is

$$\operatorname{SSE}(\bar{x}) = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$

To minimize SSE, we take the derivative of SSE in respect to $\bar{x}$ and find its zero.

$$\begin{aligned}
0 &= \frac{\mathrm{dSSE}}{\mathrm{d}\bar{x}} \\
&= \frac{\mathrm{d}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right)}{\mathrm{d}\bar{x}} \\
&= \frac{\mathrm{d}\left((x_1^2 - 2x_1\bar{x} + \bar{x}^2) + (x_2^2 - 2x_2\bar{x} + \bar{x}^2) + \cdots + (x_n^2 - 2x_n\bar{x} + \bar{x}^2)\right)}{\mathrm{d}\bar{x}} \\
&= (-2x_1 + 2\bar{x}) + (-2x_2 + 2\bar{x}) + \cdots + (-2x_n + 2\bar{x}) \\
&= -2x_1 - 2x_2 - \cdots - 2x_n + n\,(2\bar{x}) \\
-2n\bar{x} &= -2x_1 - 2x_2 - \cdots - 2x_n \\
\bar{x} &= \frac{-2x_1 - 2x_2 - \cdots - 2x_n}{-2n} \\
&= \frac{x_1 + x_2 + \cdots + x_n}{n}
\end{aligned}$$

The arithmetic mean is found at $\mu = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$. □

A less-general demonstration of the above proof is given below in the Wolfram Language.

```
In[1]:= X := {x1, x2, x3}

In[2]:= Err[mu_] := X - mu

In[3]:= SSE[mu_] := Total[Err[mu]^2]

In[4]:= SSE'[mu]

Out[4]= -2 (-mu + x1) - 2 (-mu + x2) - 2 (-mu + x3)

In[5]:= Solve[SSE'[mu] == 0, mu]

                 x1 + x2 + x3
Out[5]= {{mu -> ------------}}
                       3
```

4.2 Expected values

The term *average* can refer to *mean*, *median*, and *mode*. Mean only applies to interval and ratio data. Median is simply the middle value among ordinal, interval, and ratio data. Mode is the "commonest" (most frequent) value among nominal, ordinal, interval, and ratio data.

Average	Levels of measurement	Symbols
Mean	Interval, ratio	μ, $\bar{x}$
Median	Ordinal, interval, ratio	(None)
Mode	Nominal, ordinal, interval, ratio	(None)

All three of these *measures of central tendency* enable us to find the *expected value* in a data set, $E(x)$. Population means are assigned the symbol μ. Estimates of the population mean (the sample mean) usually use the name of the sample with a bar, such as $\bar{x}$.

Median and mode can be useful even when analyzing interval and ratio data. Consider a classroom of 10 students who are 6 years old and 1 teacher who is 50 years old. If one selects a random person in the room, what is the expected value for their age? In this case, the modal value (6) is likely a better estimate than the mean value (10).

4.3 The Four Moments

The *four moments* describe the *distribution* of values in a data set. The first moment is the mean. The second moment is *variance*, the expected squared difference of values to the mean. The third moment is *skewness*, the expected

cubed difference of values to the mean. The fourth moment is *kurtosis*, the expected difference of values to the mean raised to the fourth power.

Moment	Name	Definition	Symbol
μ_1	Mean	$E(x)$	μ
μ_2	Variance	$E(x-\mu)^2$	σ^2
μ_3	Skewness	$E(x-\mu)^3$	β_1
μ_4	Kurtosis	$E(x-\mu)^4$	β_2

Variance (σ^2) is calculated from the sum of the squared differences in the random variable (x) and the mean (μ).

$$\begin{aligned}\sigma^2 &= E(x-\mu)^2 \\ &= E(x^2 - 2x\mu + \mu^2) \\ &= E(x^2) - 2E(x)\mu + \mu^2 \\ &= E(x^2) - 2(\mu)\mu + \mu^2 \\ &= E(x^2) - 2\mu^2 + \mu^2 \\ &= E(x^2) - \mu^2.\end{aligned}$$

To calculate the sample variance (s^2) in practice, we use the formula

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

where n is the number of elements in x. Observe that the calculation for this statistic is similar to the least squares method in section 4.1. s^2 is the average squared distance from the mean within the data set.

The *standard deviation* is the square root of variance,

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

An *outlier* is a value that is very different from other values in the data set. In the set $x = \{15, 75, 79, 10, 7, 54, 4600, 91, 45, 86\}$, one can immediately observe that the value 4600 is substantially different from all of the other values. Outliers can be defined in terms of the mean and standard deviation. Outliers are any values greater than $\bar{x} + 2s$ or less than $\bar{x} - 2s$. We can use the `mean` and `sd` functions with `subset` in the R language at https://webr.r-wasm.org/latest/ to confirm that 4600 is an outlier.

```
> x = c(15, 75, 79, 10, 7, 54, 4600, 91, 45, 86)
> subset(x, x <= mean(x) - sd(x) | x >= mean(x) + sd(x))
[1] 4600
```

We can use skewness ($\mu_3 = \beta_1$) to detect whether the data is imbalanced (skewed) above or below the mean. If skewness is negative then the left tail is longer, if skewness is positive then the right tail is longer, and if skewness is zero then the distribution is equally balanced over the mean. Excel defines its `SKEW` function[1] as

$$\mu_3 = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x - \bar{x}}{s} \right)^3.$$

We can use kurtosis ($\mu_4 = \beta_2$) to detect if a data set contains outliers. The kurtosis of the normal distribution is 3. Karl Pearson defines the *degree of kurtosis* as $\eta = \beta_2 - 3$ [25, p. 181]. Other texts call this *excess kurtosis*. Excel's `KURT` function returns excess kurtosis. If `KURT` returns 0, then the distribution may fit a normal distribution and may contain no outliers. Excel defines its `KURT` function[2] as

$$\mu_4 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

4.4 The Normal Distribution

The *Normal Distribution*, also known as the *Gaussian Distribution*, is a well-known predictor of the probability of continuous outcomes. Parameterized by mean, μ, and standard deviation, σ, the *probability density function* for the normal distribution is

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

$P(x)$ predicts the probability that an observation will have value x. The *standard normal* has parameters $\mu = 0$ and $\sigma = 1$, with skewness $\mu_3 = 0$ and kurtosis $\mu_4 = 3$.

$P(x)$ forms a "bell curve" (see figure 4.1) that one might encounter in a histogram of data, but there are other distributions of data which also form a bell-shaped curve. It is **not** generally safe to immediately assume that data fits a normal distribution when a histogram reflects a bell curve.

Just as *density* in physics is defined as mass per volume, the probability density function is correspondingly a rate. The units for $P(x)$ are probability per value. If we wanted to find mass from the density of a fluid, we would multiply its

[1] https://support.microsoft.com/en-us/office/skew-function-bdf49d86-b1ef-4804-a046-28eaea69c9fa

[2] https://support.microsoft.com/en-us/office/kurt-function-bc3a265c-5da4-4dcb-b7fd-c237789095ab

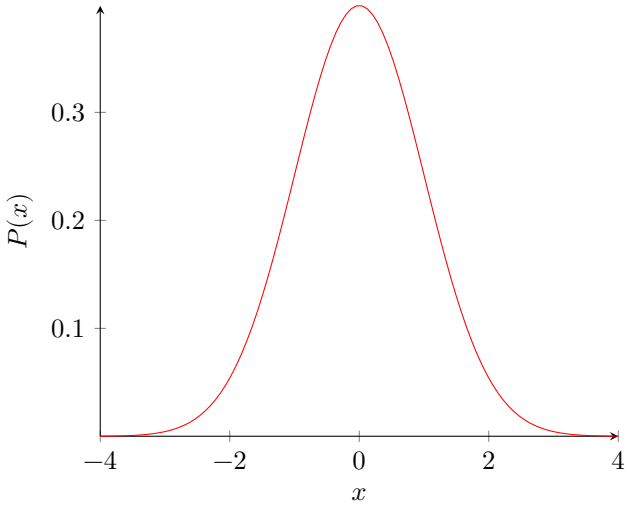


Figure 4.1: This familiar "bell curve" is a plot of the probability density function of the normal distribution. Though the curve appears to touch the horizontal axis in this plot, the values actually approach but never reach zero, even as x continues infinitely far in either direction.

density by the volume of fluid we have. Correspondingly, to take the *probability mass* from $P(x)$, we multiply probabilities by the range of x values.

In a *uniform distribution*, finding the probability mass is as simple as multiplying the probability, $U(x)$, by the range of x values. For example, the probability of getting $0.0 \leq x \leq 0.5$ from a uniform random number generator producing values in the range $[0.0, 1.0]$ is (informally) $U \times (0.5 - 0) = 0.5$. What is the probability of getting *exactly* 0.12345? If U can produce infinitely many digits, then the probability mass is zero. Not approximately zero — exactly zero: we multiplied $U \times (0.12345 - 0.12345) = U \times (0) = 0$.

Probability density is not constant in the normal distribution, therefore to get the probability mass we can use calculus. The integral of the entire space is

$$\int_{-\infty}^{+\infty} P(x)\,dx = 1.$$

The probabilities of an event occurring at $\pm\sigma$, $\pm 2\sigma$, and $\pm 3\sigma$ are approximately 68%, 95%, and 99%.

$$\int_{-1}^{+1} P(x)\,dx \approx 0.68$$
$$\int_{-2}^{+2} P(x)\,dx \approx 0.95$$
$$\int_{-3}^{+3} P(x)\,dx \approx 0.99$$

Just as the probability mass was zero when the range had zero length, we also find that the probability mass for a single value is zero.

$$\int_{t}^{t} P(x)\,dx = 0$$

The R language provides functions `dnorm`, `pnorm`, `qnorm`, and `rnorm`. `dnorm` is the same probability density function shown above as $P(x)$. `pnorm` gives the probability mass (also known as *cumulative probability*) between some range of x values, which are numerical estimations of the definite integrals shown earlier.

```
> integrate(dnorm, -1, 1)
0.6826895 with absolute error < 7.6e-15
> integrate(dnorm, -2, 2)
0.9544997 with absolute error < 1.8e-11
> integrate(dnorm, -3, 3)
0.9973002 with absolute error < 9.3e-07
> integrate(dnorm, -Inf, 0)
0.5 with absolute error < 4.7e-05
> integrate(dnorm, 0.12345, 0.12345)
0 with absolute error < 0
> pnorm(0)
[1] 0.5
> pnorm(1) - pnorm(-1)
[1] 0.6826895
```

The `qnorm` function is the inverse of cumulative probability, allowing us to find an exact point x in the distribution for a probability value $p = P(x)$.

The `rnorm` function generates random numbers which fit a normal distribution. Some computing environments provide only uniformly-distributed random numbers. In Excel, one can use the `qnorm` equivalent (`NORM.INV`) in a clever way to create a normal distribution with

```
=NORM.INV(RAND(),0,1)
```

4.5 Exponential moving averages

An *exponential moving average* (EMA) is a weighted average that creates a bias favoring recent observations. EMAs are used in the financial sector as an implicit means of modeling stock prices with time.

One might also think of EMA as a method to estimate *instantaneous* values in presence of errors. An example application might be a smartwatch using GPS to estimate a runner's pace. As the runner's wrist travels back and forth, the instantaneous velocity of the watch will not match that of the runner. An EMA might be useful to smooth this noise.

EMA is defined *recursively* and parameterized with a weighting multiplier $0 < p < 1$.

$$\text{EMA}\left(x, i\right) = \begin{cases} px_i + (1-p)\text{EMA}\left(x, i-1\right), & \text{if } i > 1. \\ x_1, & \text{otherwise.} \end{cases}$$

EMA can be easily implemented in terms of the `reduce` operation, as shown below in JavaScript.

```
>> x = [7, 8, 9, 10, 9, 8, 7, 9, 11, 13, 15, 17]
>> p = .25
>> x.reduce((ema,v) => p * v + (1-p) * ema, x[0])
<- 12.661770105361938
```

The *base case* is at x_1 in mathematical notation but `x[0]` in JavaScript. This is a matter of convention; the first element of a list is position 1 in math, but 0 in many programming languages.

4.6 Strong and Weak Links

In a *weak-link problem*, the system is harmed by the presence of any defect. The term itself is named for a proverbial chain, which is only as strong as the weakest link. Many safety-, quality-, and process-related problems require one to eliminate weak links.

On the other hand, we sometimes encounter *strong-link problems*, where overall success of a system depends on the very *best* individuals in the population. Olympic athletes are a good example of a strong-link problem: it does not matter that a country has thousands of candidate athletes who did not qualify; what matters is that the national team selects the very best to compete on the world stage.

Venture capital is another example of a strong-link problem. An investor takes risks on many different companies in hopes that one "unicorn" startup will yield a large return, offsetting losses from those unsuccessful startups.

Both weak-link and strong-link problems can be modeled in terms of variance (see section 4.3).

4.7 Inclusion Criteria

The "big" in "big data" refers data mining efforts in very large data sets that one cannot process on small computers with traditional methods. For example, consider an agricultural process which historically measured its yield in volume (such as liters of milk), but later changed this metric to the financial value of the yield (such as a dollars per shipping container, where not all shipping containers deliver the same product). In this hypothetical situation, an analyst might attempt to transform one or both units to a comparable calculated column, but the process may introduce uncertainty.

All studies, large and small, require inclusion criteria. Obvious reasons to exclude samples include:

- Duplicated rows
- Missing values
- Obvious errors (i.e., height and weight entered as "5" and "11")

Outlier analysis can also cast out potential errors, although this technique is inappropriate in strong- and weak-link problems where one's goal is to investigate those outliers.

In some situations, one might decide to exclude certain samples that introduce uncontrolled variance. An example might be a study on 35 automobiles, where 33 cars are new and 2 of the cars are 30 years old. Depending on the study, it might be appropriate to exclude the two old cars from the study. Another example might be a health study of 195 women and only 3 men. The researcher might choose to exclude the men to reduce the dimensionality of the problem. We will discuss the "curse of combinatorics" in the next chapter and see that every additional feature increases the complexity of one's study.

Mark Twain is often quoted to have said, "There are lies, damned lies, and statistics." A researcher can create lies by manipulating inclusion criteria. Many statistical tests output a probability (p), called *p-values*. Scientists engage in *p-hacking* when they seek to manipulate the p-values in a study to coerce a desired result [26]. One must take care to

4.8 Discussion prompts

1. Is four a lot?
2. First battalion has an average ACFT score of 482 while second battalion has an average ACFT score of 491. Which is better?

3. What do we do when statistics show us something that contradicts our values? For example, suppose we discover that Soldiers of a specific demographic have much lower promotion rates than their peers.

4. Is it more important for an organization to think about variance or the 99th percentile?

5. Given a sample set $x = \{5\}$, what is the estimate of the mean ($\bar{x}$), and what is the sample variance (s_x)? That is, what is the expected value ($E(x)$) of a random sample taken from x, and what is the average difference of variables to the expected value? Use software to verify your answer. In the R language, this would be `mean(c(5))` and `sd(c(5))`.

6. A customer-service organization uses *average handling time* (AHT), the expected number of minutes required to complete an interaction, to improve service quality. What are some benefits or risks of this approach? See https://xkcd.com/2899/.

7. Does the Army's standards-based approach hinder our ability to solve strong-link problems requiring unpredictable solutions? How could an ideal institution maintain minimum standards while enabling outliers to flourish?

8. Generate a data frame of random numbers with ten rows and three columns in the R language[3].

   ```
   data.frame(A = rnorm(10), B = rnorm(10), C = rnorm(10))
   ```

   The values of each column should be normally distributed, with a mean of about 0 and a standard deviation of about 1.

   Share the resulting data and split into groups to compete in a p-hacking game. One group should try to find an arbitrary inclusion criteria for B or C that satisfies $\bar{A} \leq -1$. Another group can try to find a different inclusion criteria that makes $\bar{A} \geq 1$. The winner is the group that retains the most rows using the simplest inclusion criteria.

4.9 Practical exercises

1. Calculate the influence that outliers have on different-sized datasets that contain outliers.

2. Calculate the exponential moving average in a small dataset.

3. Given a dataset and experimental result, identify problems caused by analyzing categorical data represented in a numeric form.

4. Given multiple datasets with identical mean and standard deviation, use kurtosis to identify the dataset with more outliers.

[3] https://webr.r-wasm.org/latest/

5. Design or implement an algorithm to incrementally calculate standard deviation, where the estimate of the sample standard deviation is updated with each additional value.
6. Think backwards and try to guess what would be the zeroth moment, μ_0.

Chapter 5

Dimensionality

5.1 Modeling Dimensions

The Myers-Briggs Type Indicator (MBTI) is a well-known model for classifying psychiatric preferences [27]. The model classifies personalities by four *dimensions*. In this context, the term "dimension" does not refer to a *spatial* dimension (left/right, up/down, and forward/backward, ordinarily symbolized in geometry as x, y, and z), but rather as *orthogonal* (independent) *attributes* that characterize a *sample space*, the set of all possible outcomes of a process. Each of the four dimensions of MBTI are *dichotomies* (binary categorical domains). The four dichotomies of MBTI are introversion and extroversion, intuition and sensing, thinking and feeling, and judging and perceiving. In total, there are $2 \times 2 \times 2 \times 2 = 2^4 = 16$ possible *tuples* that can be combined from these inputs.

$$\begin{Bmatrix} \text{I} \\ \text{E} \end{Bmatrix} \times \begin{Bmatrix} \text{N} \\ \text{S} \end{Bmatrix} \times \begin{Bmatrix} \text{T} \\ \text{F} \end{Bmatrix} \times \begin{Bmatrix} \text{J} \\ \text{P} \end{Bmatrix} = \begin{Bmatrix} \text{INTJ} & \text{INTP} & \text{INFJ} & \text{INFP} \\ \text{ISTJ} & \text{ISTP} & \text{ISFJ} & \text{ISFP} \\ \text{ENTP} & \text{ENTP} & \text{ENFJ} & \text{ENFP} \\ \text{ESTJ} & \text{ESTP} & \text{ESFJ} & \text{ESFP} \end{Bmatrix}$$

In this chapter, we explore the nature of the *dimensionality* of data sets. We begin with a brief introduction to *combinatorics*, a field of discrete mathematics for counting and arranging the elements of sets. We see that small input domains combine into large output ranges, complicating data mining efforts and limiting our ability to draw conclusions from even large data corpora. We learn to reason about basic probabilities with the binomial distribution. We introduce Pearson and Chatterjee correlation before showing Principal Component Analysis (PCA), a powerful tool for compressing the dimensions of a data set. Finally, we touch upon the Pareto Frontier, a technique one can use for uncompressible data.

5.2 Combinatorics

Suppose a family has four children, $F = \{a, b, c, d\}$, and a motorcycle. The motorcycle can carry only one passenger, so there are four possible *combinations* of children that you can transport by motorcycle[1]:

$$4 = |\{\{a\}, \{b\}, \{c\}, \{d\}\}| .$$

The family adds a sidecar to the motorcycle and can now transport two children at once. There are now six ways that one can *choose* two elements from a four-element set:

$$6 = |\{a, b\}, \{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}, \{c, d\}| .$$

Recall from section 1.13 that sets are unordered; $\{a, b\}$ is equal to $\{b, a\}$.

Two common notations for the number of possible subsets we can choose are $\binom{n}{r}$ and nCr. The former is favored in higher mathematics, the latter in secondary schools. $\binom{n}{r}$ is read "n choose r" and nCr is read "n combinations taken r at a time."

The family purchases a small car that can transport three passengers:

$$\binom{4}{3} = 4 = |\{a, b, c\}, \{a, b, d\}, \{a, c, d\}, \{b, c, d\}| .$$

The family purchases a larger car that can carry four passengers:

$$\binom{4}{4} = 1 = |\{a, b, c, d\}| .$$

Finally, the family crashes the large car and is left with a bicycle. The bicycle has no capacity to carry passengers, and therefore

$$\binom{4}{0} = 1 = |\{\}| = |\emptyset| .$$

There is only one way to take an empty set from another set.

We now construct a generalized function to count the number of subsets for any combination of r elements taken from a set of size n. Initially, consider the first element in the set. If we choose this element, then we still to select $r - 1$ elements from the remaining $n - 1$ elements. If we do not choose this element,

[1]The vertical bracket notation $|S| = n$ gives the *cardinality* (the size, n) of set S.

then we still must choose r elements from the remaining $n-1$ elements. This gives us *Pascal's formula*, a *recursive* definition for counting combinations.

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r-1}$$

We need at least one *base case* to prevent this function from entering an infinite loop. These identities should be intuitive from the earlier exercise, though the proof for the final case is left as an exercise to the reader.

$$\binom{n}{r} = \begin{cases} 1, & \text{if } n = r. \\ 1, & \text{if } r = 0. \\ n, & \text{if } r = 1. \\ 0, & \text{if } n < r. \end{cases}$$

Implemented in the R language (https://webr.r-wasm.org/latest/),

```
pascal <- function(n,r) {
  if (n < r) {
    return(0)
  } else if (n == r) {
    return(1)
  } else if (r == 0) {
    return(1)
  } else if (r == 1) {
    return(n)
  } else {
    return(pascal(n-1,r) + pascal(n-1,r-1))
  }
}
```

we can reproduce the results of our passengers example. The `sapply` function in R is comparable to the `map` operation (see section 3.6).

```
> sapply(0:4, function(r) pascal(4, r))
[1] 1 4 6 4 1
```

5.3 Permutations

An alternative definition for the combination formula requires *permutations* – ordered subsets. From set $S = \{a, b, c, d\}$ there are twelve two-element permutations, represented here as *tuples*: (a, b), (b, a), (a, c), (c, a), (a, d), (d, a), (b, c), (c, b), (b, d), (d, b), (c, d), and (d, c).

When counting the size of the permutation set of length r chosen from a set of size n, we begin wtih n possible elements for the first tuple element, then

$n-1$ possible elements for the second tuple element, and so on until all r tuple elements are filled.

$$nPr = n \times (n-1) \times (n-2) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}$$

The *permutation formula* is usually defined using the *factorial* function, denoted by the "!" postfix operator.

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1 = \prod_{i=1}^{n} i$$

$0! = 1$ by definition. The intuition for this is the bicycle: there was one way to choose an empty set from a set, and likewise there is one empty tuple of zero ordered elements taken from a set.

The number of $r = n$-length permutations of a set of size n is simply

$$nPn = \frac{n!}{(n-n)!} = \frac{n!}{0!} = \frac{n!}{1} = n!$$

Now we can define the combination formula in terms of the permutation formula. We count the number of permutations but de-duplicate this count, as combinations are unordered. The number of duplicated entries is $rPr = r!$.

$$nCr = \binom{n}{r} = \frac{nPr}{r!} = \frac{\frac{n!}{(n-r)!}}{r!} = \frac{n!}{r!\,(n-r)!}$$

5.4 n choose 2

The case $\binom{n}{2}$ occurs often and deserves special discussion. Using Interactive Python (IPython), we compute the first few terms with *list comprehension*, a form of declarative programming in high-level languages.

```
In [1]: import math

In [2]: [math.comb(n, 2) for n in range(2,12)]
Out[2]: [1, 3, 6, 10, 15, 21, 28, 36, 45, 55]
```

It is not possible to choose two elements from a set of only one, there is only one way to choose two from two, three ways to choose two from three,

$$\{a,b\}, \{a,c\}, \{b,c\} \subset \{a,b,c\},$$

six ways to choose from four,

$$\{a,b\},\{a,c\},\{a,d\},\{b,c\},\{b,d\},\{c,d\} \subset \{a,b,c,d\},$$

and so on. The resulting sequence of integers are called the *triangular numbers*.

$$\begin{aligned} 1 &= 1 \\ 1+2 &= 3 \\ 1+2+3 &= 6 \\ 1+2+3+4 &= 10 \\ 1+2+3+4+5 &= 15 \end{aligned}$$

Intuitively, the difference in $\binom{k+1}{2}$ and $\binom{k}{2}$ is k: if we add a $(k+1)$th element to a set, then we can pair this new element with each of the k existing elements. The generalized form is

$$\binom{n}{2} = 1+2+3+\cdots+(n-1) = \frac{n(n-1)}{2}.$$

We can demonstrate this identity numerically

```
In [3]: [sum(k for k in range(n)) for n in range(2,12)]
Out[3]: [1, 3, 6, 10, 15, 21, 28, 36, 45, 55]

In [4]: [n*(n-1)//2 for n in range(2,12)]
Out[4]: [1, 3, 6, 10, 15, 21, 28, 36, 45, 55]
```

and prove with *mathematical induction*. The basis of induction is the case $n = 2$, where

$$\binom{2}{2} = 1 = \frac{n(n-1)}{2} = \frac{2(2-1)}{2} = \frac{2(1)}{2} = 1.$$

The inductive step is that if $\binom{k}{2} = \frac{k(k-1)}{2}$, then $\binom{k+1}{2} = \frac{(k+1)((k+1)-1)}{2}$. Remembering that $\binom{k+1}{2} - \binom{k}{2} = k$, we find

$$\begin{aligned} \frac{(k+1)((k+1)-1)}{2} - \frac{k(k-1)}{2} &= \frac{(k+1)k}{2} - \frac{k(k-1)}{2} \\ &= \frac{k((k+1)-(k-1))}{2} \\ &= \frac{2k}{2} \\ &= k.\square \end{aligned}$$

An alternative proof is to use algebra from our definition

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

as follows:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{(n)(n-1)(n-2)\cdots(3)(2)(1)}{(2)(n-2)(n-3)\cdots(3)(2)(1)} = \frac{(n)(n-1)}{2}.$$

Yet another proof is to observe the series $1+2+3+\cdots+(n-1)+n$, cleverly reverse the series and add it to itself to form $(n+1)+((n-1)+2)+\cdots+(n+1)$, observe that there are n of these identical terms and the original sum is half that of the second. Though elegant, this proof technique may not as portable to other problems as computational, inductive, and algebraic methods.

5.5 The Curse of Combinatorics

The phrase "the curse of combinatorics" refers to the vast combinatorical spaces that arise naturally from iterated multiplication. Consider a bicycle factory that must manufacture a part in four materials (steel, aluminum, carbon fiber, and titanium), three sizes (small, medium, and large), five styles (road, mountain, touring, utility, and economy), and for five markets (North America, European Union, Latin America, East Asia, and Middle East) which each have different compliance requirements. There are $4 \times 3 \times 5 \times 5 = 300$ distinct variations of this part. Suppose a distributor wants to warehouse 50 of each part, but expects the factory to wait until the part is sold before receiving payment. Should the factory give the the distributor $300 \times 50 = 15\,000$ of this part?

Now suppose an investor wants a rigorous test of the bicycle factory's products. The investor demands that 30 copies of each part be tested in various ways. $300 \times 30 = 9000$ total parts being committed to this study might be unrealistic.

5.6 Satisfiability and Constraint Solvers

The Boolean Satisfiability Problem (SAT) is a class of hard problems that are considered *intractable* because of this "curse of combinatorics" [28]. The SAT problem asks if there is any set of *literals* (reified true or false values) that we can assign to a given set of *variables*, which are combined into the *clauses* of a *formula*. For example, the formula

$$\mathcal{F} = (a \vee b \vee \neg d) \wedge (\neg a \vee c \vee d) \wedge (b \vee \neg c)$$

contains four variables (a, b, c, and d) in three clauses. $\mathcal{F}$ is expressed in the *conjunctive normal form* (CNF), meaning it is the conjunction (logical and) of

clauses. Each clause of a CNF formula is the disjunction (logical or) of Boolean variables. Clauses may contain negated variables. The 2SAT variant of the SAT problem restricts each clause to having exactly two variables. The 3SAT variant requires exactly three variables.

The Wolfram language can solve satisfiability problems[2] [3]:

```
In[1]:= f := (a || b || Not[d]) && (Not[a] || c || d) && (b || Not[c])
In[2]:= SatisfiableQ[f]
Out[2]= True

In[3]:= SatisfiabilityInstances[f, {a, b, c, d}]
Out[3]= {{False, False, False, False}}
```

The literals $a = F$, $b = F$, $c = F$, and $d = F$ satisfy $\mathcal{F}$.

$$\begin{aligned}(F \vee F \vee \neg F) \wedge (\neg F \vee F \vee F) \wedge (F \vee \neg F) = \\ (F \vee F \vee T) \wedge (T \vee F \vee F) \wedge (F \vee T) = \\ (T) \wedge (T) \wedge (T) = T\end{aligned}$$

$\mathcal{F}$ has more than one solution. The following *truth table* enumerates all $2^4 = 16$ possible combinations of true and false literals for a–d.

a	b	c	d	$\mathcal{F}$
T	T	T	T	T
T	T	T	F	T
T	T	F	T	T
T	T	F	F	F
T	F	T	T	F
T	F	T	F	F
T	F	F	T	T
T	F	F	F	F
F	T	T	T	T
F	T	T	F	T
F	T	F	T	T
F	T	F	F	T
F	F	T	T	F
F	F	T	F	F
F	F	F	T	F
F	F	F	F	T

Small instances of SAT are easily solvable by enumerating all 2^n possible sets of literals, but as n grows 2^n quickly becomes too large to search.

[2] https://reference.wolfram.com/language/ref/SatisfiableQ.html.en
[3] https://reference.wolfram.com/language/ref/SatisfiabilityInstances.html

		8	4	2		9	1	
4	3	2		1	5		8	7
9				8		2		4
8					2		7	
	7	4			8			
	2	9			4	5	3	
				7				
	4	3			6		9	
5	8				9	7	2	6

Figure 5.1: This Sudoku puzzle is solvable using Python and Z3.

SAT solvers and constraint solvers are computer programs and languages intended to solve these problems. These solutions are not commonly understood, even among computer scientists [29]. Z3 is one such theorem prover from Microsoft Research[4] [30]. Z3's parenthesized prefix notation resembles that of Lisp languages. Users of constraint solvers may prefer to use more familiar languages, such as Python. Python's list comprehension, which we previously saw in section 5.4, is a useful idiom to declaratively create constraints.

Sudoku is a puzzle with a 9×9 grid of integers in 1–9, as shown in figure 5.1. Each row contains exactly one of 1–9 and each column contains exactly one of 1–9. When partitioned into nine 3×3 squares, each square also contains exactly one of 1–9. The number of valid game configurations is an immense combinatorial space, on the order of $9! \times 8! \times 7! \times \cdots \times 1! \approx 10^{21}$ (the exact number is believed to be higher [31]), yet the following Python program discovers a solution in less than a second using Z3[5].

```
In [1]: from z3 import *
```

[4] https://github.com/Z3Prover/z3

[5] This program is heavily influenced by https://ericpony.github.io/z3py-tutorial/guide-examples.htm

```
In [2]: # Declare Integer variables v[0][0] through v[8][8]. Each variable
   ...: # represents a position of the Sudouku puzzle.
   ...: v = [[Int(f"v{row}{col}") for col in range(1,10)]
   ...:      for row in range(1,10)]
   ...:

In [3]: # Variables have values between 1 and 9, inclusive.
   ...: constraint1 = [And(1 <= v[row][col], v[row][col] <= 9)
   ...:                for row in range(9) for col in range(9)]

In [4]: # The values in each row are distinct.
   ...: constraint2 = [Distinct(v[row]) for row in range(9)]

In [5]: # The values in each column are distinct.
   ...: constraint3 = [Distinct([v[row][col] for row in range(9)])
   ...:                for col in range(9)]

In [6]: # The values in each 3x3 square are distinct.
   ...: constraint4 = [Distinct([v[row + 3 * y][col + 3 * x]
   ...:                          for row in range(3) for col in range(3)])
   ...:                          for y in range(3) for x in range(3)]

In [7]: # Literal assignments for our puzzle input.
   ...: example = [[None, None, 8, 4, 2, None, 9, 1, None],
   ...:            [4, 3, 2, None, 1, 5, None, 8, 7],
   ...:            [9, None, None, None, 8, None, 2, None, 4],
   ...:            [8, None, None, None, None, 2, None, 7, None],
   ...:            [None, 7, 4, None, None, 8, None, None, None],
   ...:            [None, 2, 9, None, None, 4, 5, 3, None],
   ...:            [None, None, None, None, 7, None, None, None, None],
   ...:            [None, 4, 3, None, None, 6, None, 9, None],
   ...:            [5, 8, None, None, None, 9, 7, 2, 6]]

In [8]: constraint5 = [v[row][col] == example[row][col]
   ...:                for row in range(9) for col in range(9)
   ...:                if example[row][col] is not None]

In [9]: # Create and initialize an instance of a Z3 constraint solver
   ...: s = Solver()

In [10]: s.add(constraint1 + constraint2 + constraint3 + constraint4 + constraint5)

In [11]: # Is the problem satisfiable?
    ...: s.check()
Out[11]: sat
```

```
In [12]: m = s.model()

In [13]: [[m.evaluate(v[row][col]) for col in range(9)] for row in range(9)]
Out[13]:
[[6, 5, 8, 4, 2, 7, 9, 1, 3],
 [4, 3, 2, 9, 1, 5, 6, 8, 7],
 [9, 1, 7, 6, 8, 3, 2, 5, 4],
 [8, 6, 5, 1, 3, 2, 4, 7, 9],
 [3, 7, 4, 5, 9, 8, 1, 6, 2],
 [1, 2, 9, 7, 6, 4, 5, 3, 8],
 [2, 9, 6, 8, 7, 1, 3, 4, 5],
 [7, 4, 3, 2, 5, 6, 8, 9, 1],
 [5, 8, 1, 3, 4, 9, 7, 2, 6]]
```

A *reduction* is the process of transforming one problem into another. The Cook-Levin Theorem states that the SAT problem is NP-complete, which means [32]:

1. A Boolean satisfiability problem cannot be solved in polynomial time. There are no known algorithm to solve arbitrary SAT problems of n variables in at most n^k steps, for arbitrary n where k is a constant.
2. While *finding* candidate solutions can requires exponentially many operations, *verifying* candidate solutions requires only polynomial time.
3. Any problem in NP can be reduced to SAT.

The P $\neq$ NP problem is an unproven computer science conjecture which states that NP problems, those with the first two characteristics, can be reduced to other NP-complete problems, but cannot be reduced to any P problem. That is, no NP problem can be reduced to a polynomial time solution. Still, modern constraint solvers use many techniques to solve problems quickly, sometimes with a loss of precision. If one can reduce a new and difficult combinatorial problem to satisfiability, then a SAT solver may be able to solve the problem through a declarative interface. Reductions can be difficult. Refer to Dennis Yurichev's *SAT/SMT by Example* for further reading [33].

5.7 Subsets and Venn diagrams

A set intersection ($\cap$) of two sets is the set of all elements present in both sets.

$$S \cap T = \{x | x \in S \wedge x \in T\}.$$

The familiar Venn diagram is commonly used to plot set intersections, but this plot is limited and is frequently misused. Traditionally, the square frame of the plot represents the universal set, U. Each circle of the Venn diagram shows a subset of U along some binary attribute. In figure 5.2, we see a degenerate Venn diagram of a single dimension. Values of U are simply in S or not in S.

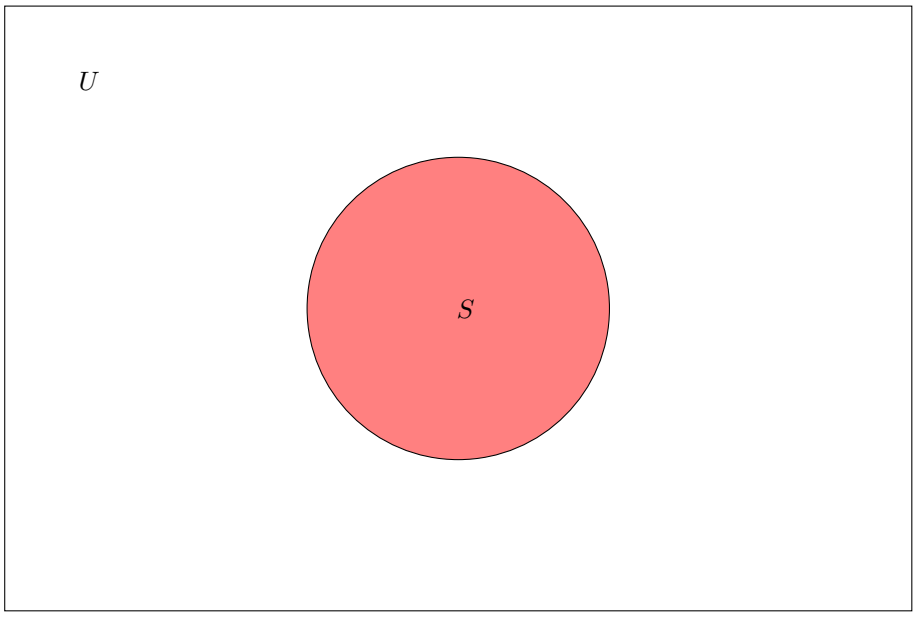


Figure 5.2: A Venn diagram showing a single dimension, $S \subset U$.

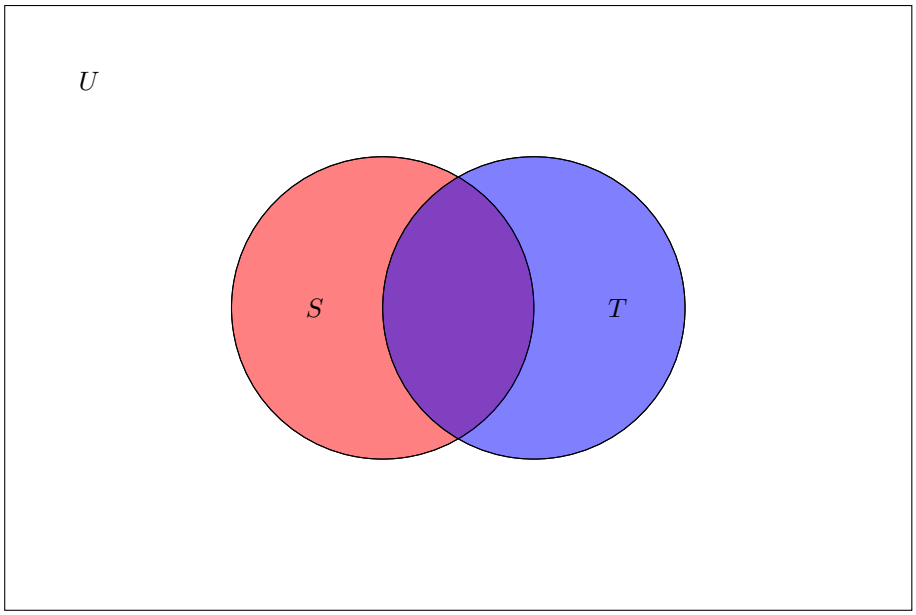


Figure 5.3: A Venn diagram showing two dimensions. The overlap of the circles is the intersection, $S \cap T$.

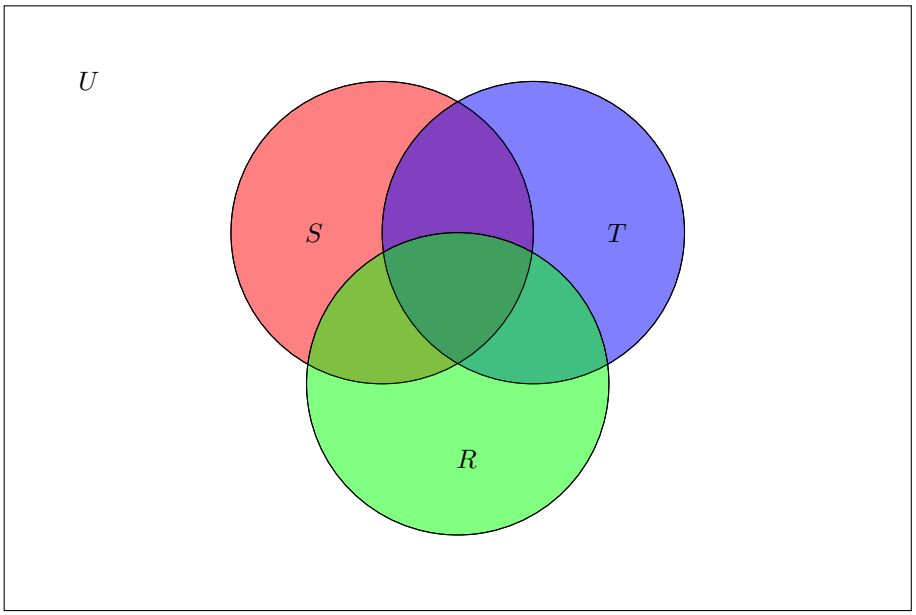


Figure 5.4: A Venn diagram showing two dimensions. The overlap of all three circles is the intersection, $R \cap S \cap T$.

The Venn diagram has its more familiar structure with two and three dimensions, as shown in figures 5.3 and 5.4.

Venn diagrams are not possible in four or more dimensions – at least, not with circles drawn on a two-dimensional plot. The number of subsets of U is two for one dimension (a value is either in S or not in S), four for two dimensions (a value is only in S, only in T, in both, or in neither), eight for three dimensions (R, S, T, $S \cap T$, $S \cap R$, $S \cap T$, $T \cap R$, $R \cap S \cap T$, or none), and so on.

Another challenge one must avoid if using Venn diagrams is that the areas in the plot may not correspond to the relative sizes of the subsets. For example, imagine a Venn diagram showing the sets of bicycle riders and persons with only one foot. The cyclists significantly outnumber the unipeds and their intersection is likely quite small, therefore two circles of equal size may present a misleading graphic.

5.8 Sample spaces

Imagine one wanted to conduct a large study on exercise and health outcomes. Basic demographic variables include age, gender, location, height, weight, and race. Exercise variables in this study include weekly minutes performing cardiovascular training, minutes of resistance training, and minutes of flexibility training. Other

exercise variables in this study include metrics of speed, endurance, strength, flexibility, blood pressure, resting heart rate, body composition, bone density, and sleep duration.

Suppose we discretize (see section 1.5) each continuous variable into discrete categories. For example, we might change the age variable from its numeric values to categories 1–10, 11–20, 21–30, and so on. We separate height into very short, short, average, tall, and very tall. We categorize minutes of weekly training into 0–20, 20–60, 60–120, and 120+. Some variables are divided into very low, low, medium, high, and very high. The process continues until all variables can be represented in discrete (sometimes ordered) categories instead of continuous numeric values.

Variable	Categories
Age	10
Gender	2
Location	5
Height	5
Weight	10
Cardio Minutes	4
Weights Minutes	4
Stretch Minutes	4
Speed	10
Strength	10
Endurance	10
Flexibility	10
Blood Pressure	5
Heart Rate	5
Composition	7
Bone Density	5
Sleep Duration	9

One might expect that, having discretized each variable, it would become easy to draw non-obvious conclusions from a reasonable sample size. Unfortunately, there are $10 \times 2 \times 5 \times 5 \times 10 \times 4 \times 4 \times 4 \times 4 = 320\,000$ possible combinations in the first eight variables alone. Is it unusual for a middle-aged, very tall, very heavy, zero-exercise male living in North America to have average fitness metrics with poor body composition? We would ideally like to sample many such persons, but even in a large study we likely would not have many individuals meeting exactly these characteristics.

Data mining is the search for non-obvious conclusions by analyzing data. Data mining efforts are especially characterized by the lack of *first principles*, meaning the researcher may not have any advance hypothesis about the relationships between variables.

Suppose our fitness research showed that heavy bodyweight predicts poor speed. This is quite obvious and likely not useful. Suppose our fitness research showed that minutes of stretching predicted not only flexibility but also strength and body composition. Such as result is less expected, and might (just as a hypothetical example) lead to a discovery that yoga develops muscle better than its reputation.

Data mining efforts in n-dimensional space are basically always complicated by this "curse of combinatorics." When we repeatedly multiply many variables together, we find that the space of possible combinations becomes so large that even very large samples cover only tiny portions. Our example health study has a total of $10 \times 2 \times 5 \times 5 \times 10 \times 4 \times 4 \times 4 \times 10 \times 10 \times 10 \times 10 \times 5 \times 5 \times 7 \times 5 \times 9 = 25\,200\,000\,000\,000$ possible states in its *sample space*.

5.9 Paradoxes

A *paradox* is a seemingly contradictory statement. Large combinatorial sample spaces sometimes create unexpected situations that may seem paradoxical.

The *birthday paradox* is well-known in computer security. Suppose there are 23 students in a class. What is the probability that any two students share a birthday? One might guess that the probability would be 23/365 until we notice that **any** two students might share a birthday. Student s_1 and s_2 might have the same birthday, s_1 and s_3, s_2 and s_3, and so on.

It is actually easier to calculate the probability that **no** students share a birthday, which we will denote with q. For the first student (s_1), there is a (degenerate) 365/365 probability that s_1 does not have share a birthday with those before because we have not considered any other students. For s_2, there is a 364/365 probability that s_2 has a distinct birthday from s_1. For s_3, there is a 363/365 probability that s_3 has a distinct birthday from both s_1 and s_2. This continues for the remaining students in the class. We multiply these probabilities together to get

$$\begin{aligned} q_{10} &= \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{343}{365} \\ &= \prod_{i=1}^{23} \frac{365 - i + 1}{365} \\ &= 0.492703. \end{aligned}$$

We now take $p = 1 - q$ to find the probability that the event *does* occur and find the likelihood that two of our ten students is

$$p = 1 - q = 1 - 0.492703 = 0.507297.$$

This means that there is more than 50% chance that any two students will share a birthday in a class of 23, a surprising and unintuitive result.

5.10 The Binomial Distribution

We now continue to another example which will demonstrate a limitation of statistical reasoning. Suppose this class of students has a large toy box with 1000 toys. Each time a child removes a toy, the teacher records the toy and the result of a fair coin flip. For example,

Toy	Coin
Shovel	Heads
Racecar	Tails
Robot	Heads
Teacup	Tails

After a very long time, each of the 1000 toys has been taken from the toy box 10 times. The teacher looks over the data and is surprised to find that coin toss has always resulted in tails for each of the ten times that a child has taken the shark toy.

It should be obvious that the shark has nothing to do with the coin flip, yet unlikely events may entice one to assume causal relationships. Consider the sample space of the coin flips. The first flip, c_1, could have been heads or tails. The second flip, c_2, could also have been heads or tails. So far, the sample space contains four possible events, which we will denote HH, HT, TH, and TT. On the third flip, the sample space again doubles in size: HHH, HHT, HTH, HTT, THH, THT, TTH, and TTT. Each additional flip will continue to double the sample space. By the tenth flip, the sample space contains $2^{10} = 1024 \approx 1000$ possible events, of which HHHHHHHHHH is just one.

Upon reflection, it should be hardly surprising that one of one thousand toys would randomly associate with a one-in-one-thousand event. To find the exact chance, we need the *binomial distribution*. The probability of event x occurring in a series of n independent trials of probability p is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

In Excel, we use the `BINOM.DIST` function. In R, `dbinom` in the *probability density function* (PDF) for the binomial distribution. To find the probability that our 1/1024 event occurs *exactly once* in 1000 trials, we find

```
> dbinom(1, 1000, 1/1024)
[1] 0.3679607
> choose(1000,1) * (1/1024)^1 * (1-1/1024)^(1000-1)
[1] 0.3679607
```

As an exercise, reproduce this result in Excel using the formula

```
=BINOM.DIST(1,1000,1/1024,FALSE)
```

We have several options to find the probability that *none* of our 1000 toys associate with ten heads. First, we can use the same `dbinom` and `BINOM.DIST` functions with $x = 0$. Second, we can take the sum of probabilities from the range `x = 1:1000` (the probability of $x = 1$, probability that $x = 2$, and so on) and then subtract this from one.

```
> dbinom(0, 1000, 1/1024)
[1] 0.3764238
> 1-sum(dbinom(1:1000, 1000, 1/1024))
[1] 0.3764238
```

Finally, statistics software often provides a *cumulative distribution function* (CDF) implementation as a shortcut for these summations. In R, this is `pbinom`, but in Excel this is provided in `BINOM.DIST` with the final argument set to `TRUE`.

The toy shark example is intended to demonstrate how *spurious correlations* may occur in large sets of data. The *Texas sharpshooter fallacy* can describe this effect. A sharpshooter fires his pistol at random into a barn wall, then draws circles around clusters of bullet holes and claims to be an expert.

5.11 Causation

One must take care not to confuse correlation with causation. Consider an experiment where seven subjects are each given a fair die and assigned "relationship" numbers such that each pair of subjects (x, y) shares a relationship R_{xy} with every other subject, as shown in figure 5.5 and enumerated in the following table.

Subject	Relationships
1	$R_{12}, R_{13}, R_{14}, R_{15}, R_{16}, R_{17}$
2	$R_{12}, R_{23}, R_{24}, R_{25}, R_{26}, R_{27}$
3	$R_{13}, R_{23}, R_{34}, R_{35}, R_{36}, R_{37}$
4	$R_{14}, R_{24}, R_{34}, R_{45}, R_{46}, R_{47}$
5	$R_{15}, R_{25}, R_{35}, R_{45}, R_{56}, R_{57}$
6	$R_{16}, R_{26}, R_{36}, R_{46}, R_{56}, R_{67}$
7	$R_{17}, R_{27}, R_{37}, R_{47}, R_{57}, R_{67}$

Have each of your seven subjects roll their die. As there are only six sides to the die, we are guaranteed that at least one pair of subjects will receive the same roll. We look at our set of relationships and discover some R_{xy} between those who rolled the same number. We have a correlation, but is there a causal relationship? Of course not.

A *mediating variable* makes establishing causality even more difficult. (todo: say more about mediating variables)

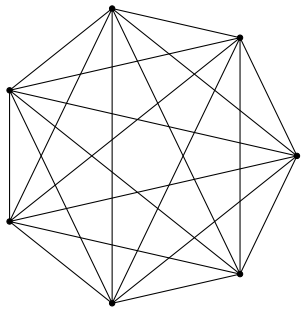


Figure 5.5: A *full mesh* network of 7 elements contains $(7)(7-1)/2 = 21$ connections, as explained in section 5.4.

Confounding factors introduce additional dimensions to a system and can make analysis more complex or, sometimes, impossible. For example, a cohort of hypothetical adult subjects enjoy positive health outcomes in one year having started exercising regularly, improving sleep quality and duration, switching to a healthy diet, stopping smoking, and always wearing blue clothes. Which of these five variables led to improved health? Obviously, the blue clothes did *not* contribute to the health outcomes, but in the presence of the other variables it may be impossible to dismiss a preposterous claim that blue clothes are healthy.

An ideal study should predict, control, and explore the combinatorial space as fully as possible. Consider another hypothetical exercise science study to contrast the benefits of running versus cycling. If the researcher realizes that the runners were significantly younger than the cyclists, then they may not be able to distinguish whether the differences in the cohorts was due to activity or age; the difference in age would be a confounding factor not *controlled* in the study.

While mediating variables and confounding factors can create spurious or misleading correlations, random *noise* in measurements can also obscure or create or exaggerate the relationships between variables. Statistics like correlation are useful to estimate an *effect size* to distinguish signal from noise.

In a data mining effort, the interactions among features in the data set might not be known in advance. *Big data*, large volumes of loosely-related and often semi-structured data, may facilitate the exploration of the n-dimensional space by providing vast numbers of both samples and features. In the following sections, we will learn methods for discovering and compressing relationships among the numerical features.

5.12 Covariance and Correlation

In section 4.3, we defined variance as the average squared difference of a random variable x to its expected value, $\bar{x}$. *Covariance* [34] [35] is a similar statistic for

two variables: covariance is computed from the average product of the differences in x and y to their respective expected values, $\bar{x}$ and $\bar{y}$.

$$\operatorname{cov}(x, y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

If we first *scale* vectors x and y such that their mean is zero and variance is one, then the covariance becomes *correlation*, a simple statistic to interpret.

$$\operatorname{scale}(x) = \frac{x - \bar{x}}{s_x}$$
$$\operatorname{cor}(x, y) = \operatorname{cov}(\operatorname{scale}(x), \operatorname{scale}(y))$$

If the data is not scaled, then the Pearson correlation coefficient (PCC) is

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}.$$

The following Rust program implements both covariance and correlation statistics. One can execute this program at the Rust Playground[6] and reproduce the result in R at https://docs.r-wasm.org/webr/latest/ with `cov(c(5,7,3,6,8), c(65,80,50,70,90))` and `cor(c(5,7,3,6,8), c(65,80,50,70,90))`.

```
fn main() {
    let x = vec![5., 7., 3., 6., 8.];
    let y = vec![65., 80., 50., 70., 90.];
    println!("Covariance: {}", cov(&x, &y).unwrap());
    println!("Correlation: {}", cor(&x, &y).unwrap());
}

fn cov(x: &Vec<f64>, y: &Vec<f64>) -> Result<f64, ()> {
    if x.len() != y.len() {
        return Err(())
    }
    let xm = mean(x);
    let ym = mean(y);
    let n = x.len() as f64;
    let covariance = x.iter().zip(y.iter()).map(|(a,b)| {
        (a - xm) * (b - ym) / (n - 1.0)
    }).sum::<f64>();
    Ok(covariance)
}
```

[6] https://play.rust-lang.org/?gist=1f3b41a17c10c354ee462062772dbd72

```rust
fn cor(x: &Vec<f64>, y: &Vec<f64>) -> Result<f64, ()> {
    cov(&scale(x), &scale(y))
}

fn scale(v: &Vec<f64>) -> Vec<f64> {
    let mu = mean(v);
    let sigma = sd(v);
    v.iter().map(|x| {
        (x - mu) / sigma
    }).collect()
}

fn mean(v: &Vec<f64>) -> f64 {
    v.iter().sum::<f64>() / (v.len() as f64)
}

fn sd(v: &Vec<f64>) -> f64 {
    let mu = mean(v);
    let variance = v.iter().map(|x| {
        (x - mu).powi(2)
    });
    let n = v.len() as f64;
    (variance.sum::<f64>() / (n - 1.0)).sqrt()
}
```

To be more precise, the scaled covariance produces a statistic of **linear** correlation. The correlation of the vector

$$x = (-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5)$$

and its element-wise squares

$$y = x \odot x = (25, 16, 9, 4, 1, 0, 1, 4, 9, 16, 25)$$

is **zero**.

Again using the R language at https://docs.r-wasm.org/webr/latest/,

```
> x = -5:5
> y = x^2
> cor(x,y)
[1] 0
```

Todo: this is a good place to motivate the correlation matrix and cite [36].

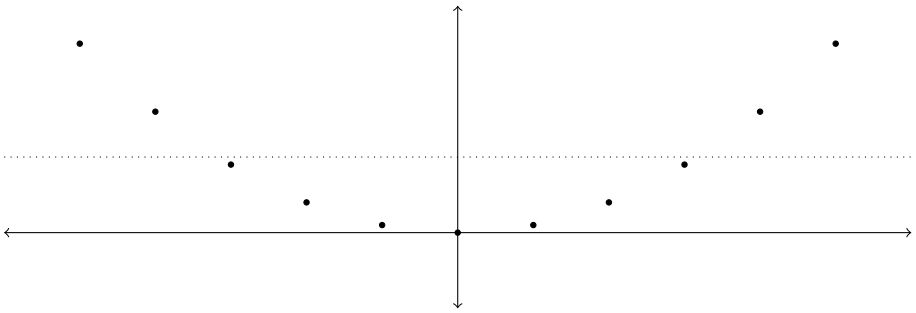


Figure 5.6: The covariance of x and $y = x \odot x$ is zero. The line of best fit for this data is shown on the dotted line, which has a Pearson correlation coefficient of $R^2 = 0$. Having a covariance of zero does not mean that y is completely independent of x, but only that there is no linear dependence.

5.13 Chatterjee's Rank Correlation

Sourav Chatterjee has recently developed and published a new coefficient of correlation [37]. This new statistic, known as ξ and pronounced "xi" or "ksaai", seeks to correlate Y as some arbitrary function of X and produces meaningful metrics on non-linear data.

The algorithm to compute $\xi(X, Y)$ first sorts Y by X, then the *ranks*, r, of the resulting order of Y. If *order* is a list of positions specifying the order of another list, then rank is the order of the order. More formally, r_i is the number of j such that $Y_{(j)} \leq Y_{(i)}$. The statistic is

$$\xi_n(X, Y) = 1 - \frac{3 \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}$$

when there are no ties. If the data set does contain duplicates, then we also use l values, where l_i is the number j such that $Y_{(j)} \geq Y_{(i)}$.

$$\xi_n(X, Y) = 1 - \frac{n \sum_{i=1}^{n-1} |r_{i+1} - r_i|}{2 \sum_{i=1}^{n} l_i (n - l_i)}$$

We earlier saw that for $x = \{x \in \mathbb{R} | -5 \leq x \leq 5\}$ and $y = x \odot x$, the Pearson correlation was $\text{cor}(x, y) = 0$. Using Chatterjee rank correlation, we find $\xi(x, y) = 0.5$.

A Rust implementation of this new ξ statistic is given below and at the Rust Playground[7]. This program is *naively* written for obviousness to the reader, not for speed of execution.

[7] https://play.rust-lang.org/?gist=b9a810274f9567213a5b2a649bd806e8

```rust
fn xicor(x: &[f64], y: &[f64]) -> f64 {
    let n = x.len();

    // Order of x values. This function does not use randomness.
    let mut order: Vec<usize> = (0..n).collect();
    order.sort_by(|&a, &b| x[a].total_cmp(&x[b]));

    // r values are the ranks of the y values. The ith y value is
    // the number of j such that y[j] <= y[i]. The order of r values
    // corresponds to the order of x.
    let r: Vec<_> = order
        .iter()
        .map(|&i| (0..n).filter(|&j| y[j] <= y[i]).count() as f64)
        .collect();

    // l values are just like the r values, only it is y[j] >= y[i].
    let l: Vec<_> = order
        .iter()
        .map(|&i| (0..n).filter(|&j| y[j] >= y[i]).count() as f64)
        .collect();

    // Sum of absolute differences in consecutive r values.
    let rsum = &r.windows(2).map(|ri| (ri[1] - ri[0]).abs()).sum();

    // Sum of l terms for the denominator.
    let lsum = l.iter().map(|&li| li * (n as f64 - li)).sum::<f64>();

    1. - (n as f64 * rsum) / (2. * lsum)
}
```

5.14 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a powerful technique for discovering linear relationships among columns of data and compressing these columns into fewer dimensions [38] [39].

PCA begins with the pairwise correlations among the data set's scaled numerical columns.

```
> head(iris[,-5])
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
```

```
6          5.4         3.9          1.7         0.4
> cor(iris[,-5])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> c = .Last.value
```

We then use a technique from linear algebra called *Singular Value Decomposition* (SVD), which extracts a diagonal matrix D from A where $U'AV = D$, $U'U = I$, and $V'V = I$. We will not discuss the details of this procedure, but will instead leave it to library software.

```
> svd(c)$u
           [,1]        [,2]       [,3]       [,4]
[1,] -0.5210659 -0.37741762  0.7195664  0.2612863
[2,]  0.2693474 -0.92329566 -0.2443818 -0.1235096
[3,] -0.5804131 -0.02449161 -0.1421264 -0.8014492
[4,] -0.5648565 -0.06694199 -0.6342727  0.5235971
> u = .Last.value
```

We multiply our scaled data set by U to compute the principal components, $\text{PC} = XU$.

```
> head(scale(iris[,-5]) %*% u)
         [,1]       [,2]        [,3]         [,4]
[1,] 2.257141 -0.4784238  0.12727962  0.024087508
[2,] 2.074013  0.6718827  0.23382552  0.102662845
[3,] 2.356335  0.3407664 -0.04405390  0.028282305
[4,] 2.291707  0.5953999 -0.09098530 -0.065735340
[5,] 2.381863 -0.6446757 -0.01568565 -0.035802870
[6,] 2.068701 -1.4842053 -0.02687825  0.006586116
> pc = scale(iris[,-5]) %*% u
> head(pc)
         [,1]       [,2]        [,3]         [,4]
[1,] 2.257141 -0.4784238  0.12727962  0.024087508
[2,] 2.074013  0.6718827  0.23382552  0.102662845
[3,] 2.356335  0.3407664 -0.04405390  0.028282305
[4,] 2.291707  0.5953999 -0.09098530 -0.065735340
[5,] 2.381863 -0.6446757 -0.01568565 -0.035802870
[6,] 2.068701 -1.4842053 -0.02687825  0.006586116
```

The resulting matrix can be used for small but accurate linear models. PCA can also reveal unexpected correlations among the data. One can think of the columns of U as new dimensions that might have been hidden among the correlated features of the original data set. The covariance among the principal components is effectively zero.

```
> library(tidyverse)
> cov(pc) %>% zapsmall
         [,1]      [,2]      [,3]      [,4]
[1,] 2.918498 0.0000000 0.0000000 0.0000000
[2,] 0.000000 0.9140305 0.0000000 0.0000000
[3,] 0.000000 0.0000000 0.1467569 0.0000000
[4,] 0.000000 0.0000000 0.0000000 0.0207148
```

The columns are ordered from greatest to least variance. This means that a model might not need all four columns to form accurate predictions, as the later columns account for very little of the variance in the data set.

todo: PCA plot.

5.15 Pareto frontier

A *Pareto frontier* (also known as a *Pareto front*) is a method for visualizing the interaction of two orthogonal (statistically independent) features of a data set.

5/3/1 is a barbell strength training program [40]. This program emphasizes *rep records*, where the lifter is to lift a submaximal mass as many times as possible. This program design adds a second dimension to strength. We say that lifter who progresses from lifting 100 kg for 6 repetitions to 9 repetitions in six months has become stronger, even if the athlete has not directly tested their one-repetition maximum.

Figure 5.7 provides an example of an athlete's rep records over a two-year period in the barbell squat. The frontier, $P(X)$, is visible at the top-right of the scatter plot. If, for example, this lifter were to achieve a 120 kg squat for 8 repetitions, the lift would *dominate* the previous records at $(120, 5)$ and $(116, 8)$, moving the frontier farther from the origin.

A Pareto front only makes sense when the two variables cannot be combined into one. Consider, as an absurd example, a running race where the minutes and seconds of finishing times are recorded in separate columns.

Athlete	Minutes	Seconds
1	18	34
2	19	24
3	20	01

There is no need to compare the three runner's run times in two dimensions: the minutes and seconds are trivially compressible into a single value with no loss of information.

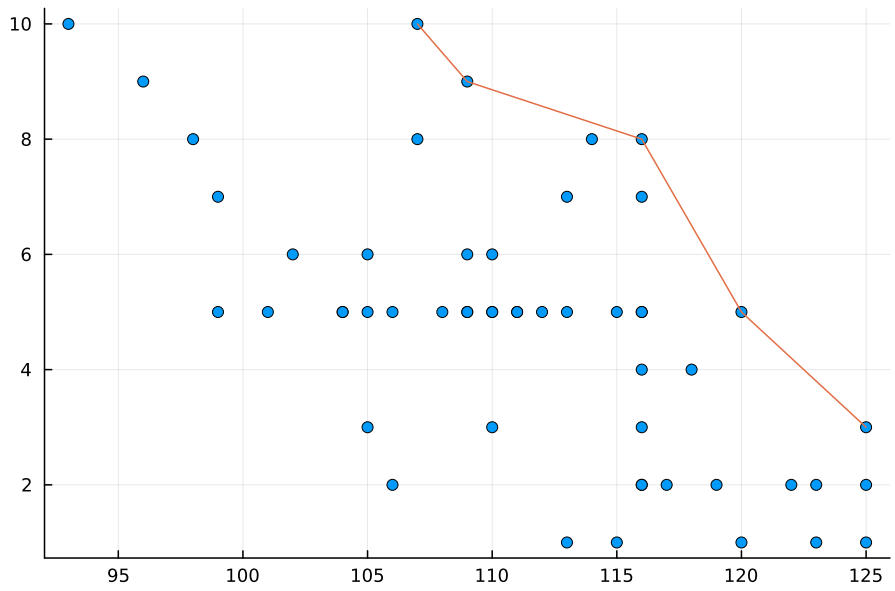


Figure 5.7: The points along the red line form the Pareto front for this data set.

In the case of the rep records shown in figure 5.7, there is a general negative correlation between mass and repetitions. This relationship can be estimated with Brzycki's formula (among others) [41], which states

$$\text{Predicted 1-RM} = \frac{\text{Weight Lifted}}{1.0278 - 0.0289x},$$

where x is the number of repetitions performed. Strong correlations in the columns of a data set present an opportunity to compress the data, thus reducing dimensionality, and search for non-obvious insights where one lacks first principles.

5.16 Discussion Prompts

1. https://www.tylervigen.com/spurious-correlations curates an entertaining collection of spurious correlations. However, not all spurious correlations might be so obvious. What are some principals we should apply to either trust or be skeptical of statistical evidence?
2. Conduct a classroom competition of "Catch the cheaters!" at https://primerlearning.org. Discuss the winning and losing strategies, then watch https://www.youtube.com/watch?v=XTcP4oo4JI4.

3. Read the interactive article https://www.mayerowitz.io/blog/mario-meets-pareto [42]. Discuss the compromises inherent in a multi-objective optimization problem.
4. Controversial topics may involve several dimensions. Advocates for one position may claim on one basis in dimension x, where the opposition's counterclaim is in dimension y. Discuss a contemporary impasse with orthogonal or irreconcilable aspects.
5. The Monty Hall problem is a notoriously unintuitive probability question. In the problem, a game show host hides a prize behind one of three doors. The guest is asked to guess which door has the prize. The host then opens one of the two unselected doors, which never contains the prize, and asks the guest if they would like to change their guess. Should the guest keep their original guess, or should the change to the unopened door? Some strategies to decide might be:
   (a) Play the game several times and tally results.
   (b) Implement the game in software to generate a large number of results quickly.
   (c) Attempt to deduce the problem using mathematical reasoning.
   (d) Change the assumptions of the game, such as adding more doors or more prizes.

5.17 Practical Exercises

1. Use nested `sapply` statements to improve `sapply(0:4, function(r) pascal(4, r))`. Iterate `pascal(n, r)` over $0 \leq n \leq 10$ and $0 \leq r \leq n$, generating the first 11 lines of Pascal's Triangle. Compare the result to `sapply(0:10, function(n) choose(n, 0:n))`. Why does the built-in `choose` function accept ranges (`0:n`) when our own `pascal` function does not?

2. About one in twenty white males have some form of color blindness. About 70.2% of the U.S. military report themselves as white, and about 82.8% as male. Let $P(C|W \cap M) = 0.05$, $P(W) = 0.702$, and $P(M) = 0.828$. If a Command gives a briefing to twelve random generals each year, what is the probability that one or more of those generals is color blind? (Naively assume, for the sake of simple calculation in this exercise, that women and non-whites are never color blind.) Assume further that W and M are independent and that

$$P(W \cap M) = P(W)P(M) = 0.581256,$$

therefore

$$P(C|W \cap M) = \frac{P(C \cap W \cap M)}{P(W \cap M)}$$

and consequently

$$P(C \cap W \cap M) = P(C|W \cap M)P(W \cap M) = (0.05)(0.581256) = 0.0290628.$$

Use this value for p in your `dbinom` calculation. Based upon this result, is it wise to depend on color-coded graphics in a presentation?

3. Come up with a creative way to draw a four-dimensional Venn diagram.
4. Use Excel to reproduce the zero correlation between x and $y_1 = x \odot x$ from section 5.12. Now update the y column to $y_2 = x \odot x + x = (20, 12, 6, 2, 0, 0, 2, 6, 12, 20, 30)$. What is $\text{cor}(x, y_2)$?
5. Use Excel's line of best fit feature to construct a linear models between both x and y_1 and also x and y_2. Observe that the y-intercept in both models is 10. Try to figure out where this constant comes from.
6. The empty set, $\emptyset$, is a set containing no elements. Its cardinality is zero.

$$|\emptyset| = 0$$

If a set contains an empty set,

$$Z = \{\emptyset\}$$

then is $|Z|$ equal to 0 or 1?

Chapter 6

Graph Theory

6.1 Vertices, edges, and paths

A *graph* (G) is a collection of *vertices* (V; also known as *nodes* or *points*) and the *edges* (E; also known as *relations* or *lines*; see section 1.13) connecting them.

$$G = \{V, E\}$$

Edges are conventionally named for the vertices they connect. For example, let $V = \{u, v\}$ and $E = \{(u, v)\}$, then $G = \{V, E\}$ is a graph with two vertices, u and v, and one edge, (u, v).

Graphs can be used to model any form of *network* where the elements of sets bear relations. Graphs can be directed, where the source and destination vertices in an edge are significant ($(u, v) \neq (v, u)$), or undirected ($(u, v) = (v, u)$). The number of edges associated with a vertex is its *degree*. In directed graphs, we distinguish *in-degree* (the number of edges leading into the vertex) and *out-degree* (the number of edges originating from the vertex).

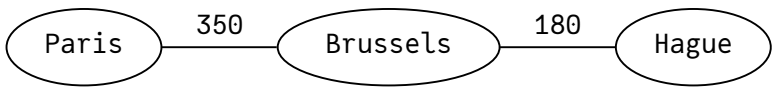


Figure 6.1: Graphs are conventionally visualized as circles (vertices) connected by lines (edges).

A *path* is a series of edges that *transitively* connect two vertices that are not directly connected. We say that $u \rightsquigarrow v$ (u leads to v) if the graph contains some path from u to v.

The edges of a graph may have a *distance function* (δ; also known as *weight* and *cost*), which relates each edge with some real number. Such graphs are *weighted graphs*. For example, suppose a high-speed railroad has train stations in Paris, Brussels, and the Hague. The distance from Paris to Brussels is about 350 km and the distance from Brussels to Hague is another 180 km. The total path length from Paris to Hague via Brussels is therefore 350 km + 180 km = 530 km.

$$\begin{aligned}\delta(\text{Paris}, \text{Hague}) &= \delta(\text{Paris}, \text{Brussels}) + \delta(\text{Brussels}, \text{Hague}) \\ &= 350\,\text{km} + 180\,\text{km} \\ &= 530\,\text{km}.\end{aligned}$$

Let us quickly reproduce this result using a *graph database* named Neo4j. Go to https://console.neo4j.org and click the "Clear DB" button. Enter the below *Cypher query* into the input field and press the triangle-shaped execute button.

```
CREATE
    (paris:CITY {name:"Paris"}),
    (brussels:CITY {name:"Brussels"}),
    (hague:CITY {name:"Hague"}),
    (paris)-[:ROUTE {dist:350}]->(brussels),
    (brussels)-[:ROUTE {dist:180}]->(hague);
```

This command created three vertices and two edges. Verify and visualize this graph with the below query. Neo4j's Cypher language uses `MATCH` as its selection (σ) operator.

```
MATCH (x:CITY)
RETURN x;
```

Finally, query the database for a path of any length (`*`) connecting Paris to Hague, returning the path (`p`) and its cumulative distance. Refer to section 3.6 for a description of the reduce operation.

```
MATCH (:CITY {name:'Paris'})-[p:ROUTE*]->(:CITY {name:'Hague'})
RETURN p, REDUCE(length=0, e IN p | length + e.dist) AS distance;
```

6.2 Connectivity and distance

A *complete* graph is *fully connected*, having paths between all pairs of vertices. One's model for distance may be challenged in incomplete graphs, which contain partitions.

What is the driving distance from Paris, France to Sydney, Australia? There is no route (no path) connecting Europe, by ground, to Australia, and therefore

there is no real number to quantify the distance.

Depending on the application, one might represent an unreachable node as having infinite distance, as we will in section 6.5.3 with Dijkstra's algorithm. One might also use a special, non-numeric values, as described in our discussion of missing values in section 1.6.

6.3 Special cases of graphs

A *directed acyclic graph* (DAG) is a special case of a directed graph. A *cycle* (also known as a *loop*) occurs in a directed graph when there is any path from some vertex to itself. For example, let G be a directed graph where

$$G = \{\{a, b, c, d\}, \{(a, b), (b, c), (c, d), (c, a)\}\}.$$

The edges (a, b), (b, c), and (c, a) form a cycle. Vertices a, b, and c together form a *connected component*. A directed graph with one or more connected components is not a DAG, but the graph of components (where vertices of the component subgraph are merged into a "super vertex") is itself a DAG.

A *tree* is another special case of an acyclic graph. Trees are often drawn in a vertical hierarchy where each *child node* has one *parent*, and the only parent node with no parent is called the *root node*. One's ancestral family tree is an instance of a tree. Without a time machine, it is impossible to one to form a hereditary loop with an ancestor.

6.4 Representation

Graphs are modeled using either an *adjacency list* or an *adjacency matrix*. An adjacency list might look like

$$\begin{aligned} \text{adj}\,(\text{Paris}) &= \{\text{Brussels}\} \\ \text{adj}\,(\text{Brussels}) &= \{\text{Paris}, \text{Hague}\} \\ \text{adj}\,(\text{Hague}) &= \{\text{Brussels}\}. \end{aligned}$$

A separate data structure would be necessary to represent edge weights. *Dictionary* types, such as `dict` in Python and `map` in Go, can be convenient implementations for both adjacency lists and edge weight functions.

An adjacency matrix represents edges with weights in a single data structure, such as

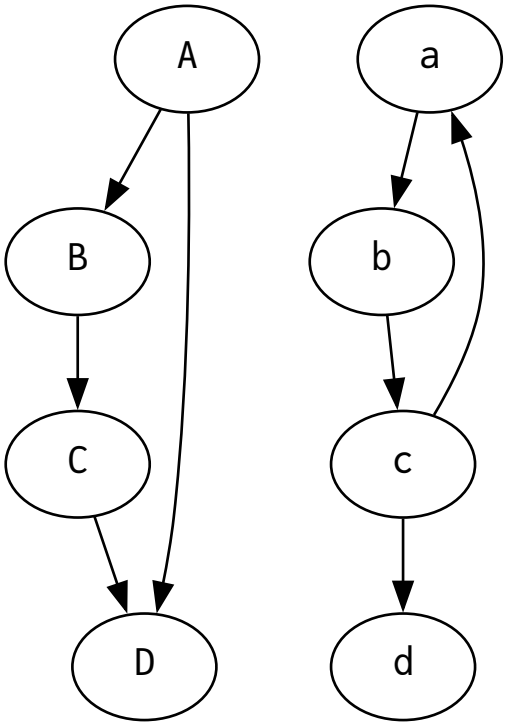


Figure 6.2: A directed acyclic graph (DAG) may contain no cycles. The left graph is a DAG. The right graph is not because the edges connecting a, b, and c form a loop.

$$E = \begin{array}{r} \\ \text{Paris} \\ \text{Brussels} \\ \text{Hague} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Paris} & \text{Brussels} & \text{Hague} \end{array} \\ \begin{bmatrix} 0 & 350 & \infty \\ 350 & 0 & 180 \\ \infty & 180 & 0 \end{bmatrix} \end{array}.$$

In this example, the distance of a vertex to itself is defined as zero,

$$\delta(u, u) = 0$$

and the distance between vertices is considered infinite if those vertices are not directly connected by an edge.

$$(u, v) \notin E \implies \delta(u, v) = \infty$$

The $\in$ operator and its negation, $\notin$, tests whether an object is an "element of" a set; $\in$ is read "in" and $\notin$ is read "not in." The symbol $\implies$ is for *conditional implication* and is read "implies." If the *statement* on the left of $\implies$ is true, then the statement on the right must also be true.

6.5 Search algorithms

6.5.1 Depth-first search

Imagine a video game where the player searches the kingdom for treasure. The player has no knowledge of where the treasure might be, so from a starting point they fully explore the forest, mountains, and sea. Both the starting point and the sea connect to opposite sides of the city. Upon entering the city from the sea-side, the player explores the city and discovers the treasure. This is an example of a *depth-first search* (DFS).

Go to the Go Playground[1] to run the following DFS implementation, written in Go. This implementation uses a *recursive* definition of the DFS function (the DFS function invokes itself as it explores the graph). The function uses an external data structure (`quest`) to identify which vertices have already been discovered. Upon successful search, the function prints its position in the graph as the recursive calls "unwind."

```
package main

import "fmt"

var g = map[string][]string{
    "start":     []string{"forest", "mountains", "sea", "city"},
```

[1] https://go.dev/play/p/AuH2qOgSG-c

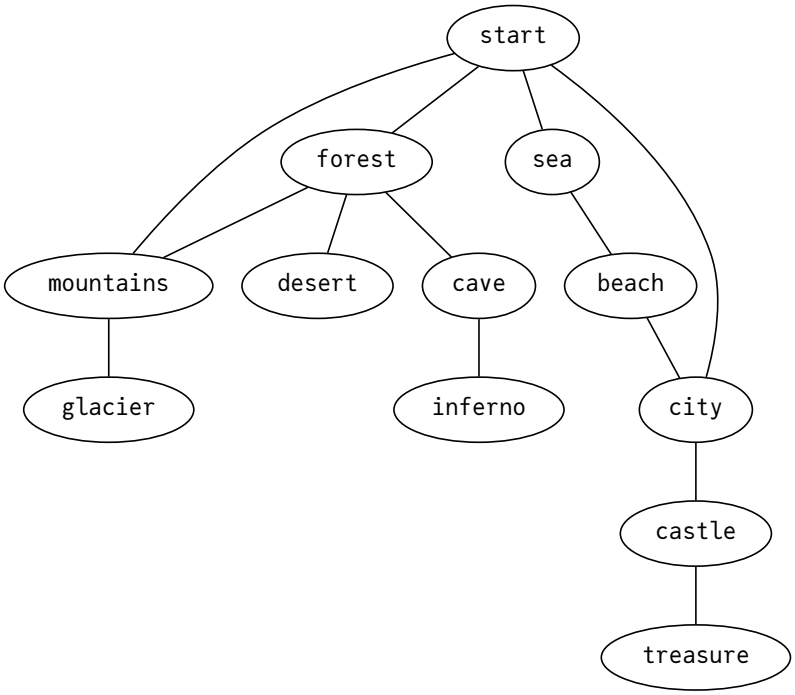


Figure 6.3: There are many paths from the starting point to the treasure in this kingdom.

```
    "forest":    []string{"start", "mountains", "desert", "cave"},
    "mountains": []string{"start", "forest", "glacier"},
    "desert":    []string{"forest"},
    "cave":      []string{"forest", "inferno"},
    "inferno":   []string{"cave"},
    "glacier":   []string{"mountains"},
    "sea":       []string{"start", "beach"},
    "beach":     []string{"sea", "city"},
    "city":      []string{"beach", "start", "castle"},
    "castle":    []string{"city", "treasure"},
    "treasure":  []string{"castle"}}

var quest map[string]string = make(map[string]string)

func dfs(src, dst string) bool {
    quest[src] = "discovered"

    if src == dst {
        fmt.Printf("Discovered %s:", dst)
        return true
    }

    for _, neighbor := range g[src] {
        if quest[neighbor] != "discovered" {
            if dfs(neighbor, dst) == true {
                fmt.Printf(" %s", src)
                return true
            }
        }
    }

    return false
}

func main() {
    dfs("start", "treasure")
    fmt.Println()
}
```

The program should output `Discovered treasure: castle city beach sea start`. DFS successfully discovers the treasure, but we have no guarantee that this algorithm will find the *optimal* (shortest) path.

6.5.2 Breadth-first search

Imagine the protagonist of our hypothetical adventure game was not a lone wanderer, but rather a field marshal commanding a large army. This army explores one region at a time, holding each area as adjacent units proceed into their respective area. The army incrementally expands the radius of the search *frontier* in a search technique called *breadth-first search* (BFS). Once one unit discovers the treasure, we are certain that no shorter path was possible thanks to an *invariant* in our search algorithm.

Maintaining an invariant is essential for *mathematical induction*, where we establish that some *predicate* P is true for the *base case* $P(0)$ and that $P(k)$ implies $P(k+1)$ and therefore $P(n)$ is true for all $n > 0$. The proof for the correctness of a BFS follows:

1. Along the frontier of radius $r = 0$, the BFS algorithm on graph G has not discovered a path from u to v. $\delta(u, v)$ is therefore at *closest* $r = 1$.
2. At $r = 1$, BFS has not found v and therefore $2 \leq \delta(u, v)$.
3. At $r = 2$, BFS has not found v and therefore $3 \leq \delta(u, v)$.
4. ⋮
5. At $r = k$, BFS has not found v and therefore $k + 1 \leq \delta(u, v)$.
6. ⋮
7. At $r = n$, BFS has located v and therefore $n = \delta(u, v)$. □

Go to the Go Playground[2] and run the following BFS implementation, written in Go. This implementation uses an *iterative* BFS function. The BFS function does not invoke itself. Instead, the procedure adds unexplored vertices to a queue and records the "parent" of each vertex. We "unwind" the resulting tree from child to parent nodes to construct the shortest path.

```
func bfs(src, dst string) map[string]string {
    parent := map[string]string{src: src}
    queue := []string{src}
    for len(queue) > 0 {
        position := queue[0]
        queue = queue[1:]
        if position == dst {
            break
        }
        for _, neighbor := range g[position] {
            if _, ok := parent[neighbor]; !ok {
                parent[neighbor] = position
                queue = append(queue, neighbor)
            }
        }
    }
    return parent
```

[2] https://go.dev/play/p/yMIcmcsK_V9

```
}

func main() {
    tree := bfs("start", "treasure")
    fmt.Printf("Discovered treasure:")
    position := "treasure"
    for position != tree[position] {
        position = tree[position]
        fmt.Printf(" %s", position)
    }
    fmt.Println()
}
```

This program should output `Discovered treasure: castle city start`. BFS finds the shortest path between two vertices by *hop count*, but it does not consider edge weights. In the following section, we will find that we can often explore graphs much faster by ordering our breadth-first traversal by cumulative path cost.

6.5.3 Dijkstra's algorithm

Dijkstra's algorithm uses a *priority queue* to visit nodes from shortest to longest path [43]. For this reason, Dijkstra's algorithm is also known as the *shortest-path first* (SPF). Like BFS, Dikjstra's algorithm is a *greedy algorithm* that discovered a globally optimal solution by repeatedly making locally optimal decisions. Let us turn to the Python language to demonstrate Dijkstra's algorithm on our same treasure-hunting graph, this time with edge weights. Run this program at https://www.python.org/shell/.[3]

```
from heapq import *
from collections import defaultdict

g = dict(
    [
        ("start", ["forest", "mountains", "sea", "city"]),
        ("forest", ["start", "mountains", "desert", "cave"]),
        ("mountains", ["start", "forest", "glacier"]),
        ("desert", ["forest"]),
        ("cave", ["forest", "inferno"]),
        ("inferno", ["cave"]),
        ("glacier", ["mountains"]),
        ("sea", ["start", "beach"]),
        ("beach", ["sea", "city"]),
```

[3]Python *requires* tab characters where other languages might accept spaces and tabs interchangably. Copying this program from a PDF will likely not work. A plain text verion of this program is available at https://github.com/wjholden/Data-Literacy/blob/main/dijkstra.py.

```
        ("city", ["beach", "start", "castle"]),
        ("castle", ["city", "treasure"]),
        ("treasure", ["castle"]),
    ]
)

w = dict(
    [
        (("start", "forest"), 70),
        (("start", "mountains"), 60),
        (("start", "sea"), 54),
        (("start", "city"), 81),
        (("forest", "start"), 42),
        (("forest", "mountains"), 51),
        (("forest", "desert"), 56),
        (("forest", "cave"), 63),
        (("mountains", "start"), 71),
        (("mountains", "forest"), 38),
        (("mountains", "glacier"), 72),
        (("desert", "forest"), 93),
        (("cave", "forest"), 19),
        (("cave", "inferno"), 17),
        (("inferno", "cave"), 71),
        (("glacier", "mountains"), 25),
        (("sea", "start"), 49),
        (("sea", "beach"), 88),
        (("beach", "sea"), 79),
        (("beach", "city"), 29),
        (("city", "beach"), 30),
        (("city", "start"), 33),
        (("city", "castle"), 36),
        (("castle", "city"), 39),
        (("castle", "treasure"), 76),
        (("treasure", "castle"), 76),
    ]
)

def dijkstra(src, dst):
    explored = set()
    distance = defaultdict(lambda: float('inf'))
    previous = {'start': None}
    distance[src] = 0
    queue = []
    heappush(queue, (0, src))
    while queue:
        _, current = heappop(queue)
```

```
        if current == dst:
            path = []
            parent = current
            while parent in previous:
                path.append(parent)
                parent = previous[parent]
            return path, distance[dst]
        if current in explored:
            continue
        explored.add(current)
        for neighbor in g[current]:
            d = distance[current] + w[(current, neighbor)]
            if neighbor not in explored and d < distance[neighbor]:
                distance[neighbor] = d
                previous[neighbor] = current
                heappush(queue, (d, neighbor))
    print("No path found")

path, distance = dijkstra("start", "treasure")
print("Path =", ', '.join(path))
print("Distance =", distance)
```

This program should output `Path found: 193`, and `Path: treasure castle city start`. The shortest path from `start` to `treasure` has a total path cost of 193.

Neo4j produces the same result. We reconstruct our graph in the Cypher language at https://console.neo4j.org:

```
CREATE
    (start:LOCATION {name:'start'}),
    (inferno:LOCATION {name:'inferno'}),
    (beach:LOCATION {name:'beach'}),
    (castle:LOCATION {name:'castle'}),
    (forest:LOCATION {name:'forest'}),
    (mountains:LOCATION {name:'mountains'}),
    (desert:LOCATION {name:'desert'}),
    (cave:LOCATION {name:'cave'}),
    (glacier:LOCATION {name:'glacier'}),
    (sea:LOCATION {name:'sea'}),
    (city:LOCATION {name:'city'}),
    (treasure:LOCATION {name:'treasure'}),
    (start)-[:CONN {distance:70}]->(forest),
    (start)-[:CONN {distance:60}]->(mountains),
    (start)-[:CONN {distance:54}]->(sea),
    (start)-[:CONN {distance:81}]->(city),
    (inferno)-[:CONN {distance:71}]->(cave),
    (beach)-[:CONN {distance:79}]->(sea),
```

```
    (beach)-[:CONN {distance:29}]->(city),
    (castle)-[:CONN {distance:39}]->(city),
    (castle)-[:CONN {distance:76}]->(treasure),
    (city)-[:CONN {distance:30}]->(beach),
    (city)-[:CONN {distance:33}]->(start),
    (city)-[:CONN {distance:36}]->(castle),
    (treasure)-[:CONN {distance:76}]->(castle),
    (forest)-[:CONN {distance:42}]->(start),
    (forest)-[:CONN {distance:51}]->(mountains),
    (forest)-[:CONN {distance:56}]->(desert),
    (forest)-[:CONN {distance:63}]->(cave),
    (mountains)-[:CONN {distance:71}]->(start),
    (mountains)-[:CONN {distance:38}]->(forest),
    (mountains)-[:CONN {distance:72}]->(glacier),
    (desert)-[:CONN {distance:93}]->(forest),
    (cave)-[:CONN {distance:19}]->(forest),
    (cave)-[:CONN {distance:17}]->(inferno),
    (glacier)-[:CONN {distance:25}]->(mountains),
    (sea)-[:CONN {distance:49}]->(start),
    (sea)-[:CONN {distance:88}]->(beach);
```

and then we can use the built-in `shortestPath` function to obtain the same result.

```
MATCH
    (src:LOCATION {name:'start'}),
    (dst:LOCATION {name:'treasure'}),
    path = shortestPath((src)-[:CONN*]->(dst))
RETURN
    path,
    REDUCE(d=0, e in relationships(path) | d + e.distance)
    AS distance;
```

Neo4j should also report a distance of 193.

As an exercise, change `->(dst)` to `-(dst)` and re-run the query with this change. The total distance is now 145. The difference in `()-[]->()` and `()-[]-()` is that one is a directed edge, the other is undirected. With `-()` instead of `->()`, Neo4j treats all edges in the graph as undirected. Neo4j allows *parallel edges* that connect the same two vertices.

6.5.4 Informed search with A*

DFS, BFS, and Dijkstra's algorithms are all *uninformed* search algorithms. The A* algorithm is an *informed* search algorithm: it uses some *heuristic* to explore its frontier ordered by minimum estimated distance to the destination [44].

A* can solve hard problems that are not immediately recognizable as searches. The canonical example is the 8-piece puzzle problem [3, pp. 111–117]. An 8-piece

puzzle arranges 8 movable square tiles into a 3×3 grid. One position is empty. The challenge is to slide the pieces until all pieces are in order.

$$\begin{bmatrix} 5 & 4 & 1 \\ & 2 & 8 \\ 3 & 6 & 7 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & \end{bmatrix}$$

The choice of heuristic function influences the path A* chooses as it searches for a solution. One heuristic function for the 8-piece puzzle problem returns the count of mismatched pieces. If two pieces are out of place then the heuristic distance is 2, if three pieces are out of place then the heuristic distance is 3, and so on. An alternative, more sophisticated heuristic function, uses the *Manhattan distance* (also known as *Taxicab distance*) of each puzzle piece to its destination. Manhattan distance is the sum of unsigned differences of each dimension between two coordinates in n-dimensional space. The intuition is that a taxicab cannot fly in a straight line, but rather has to corner the rectangular blocks of Manhattan.

$$d_T(p, q) = \sum_{i=i}^{n} |p_i - q_i|$$

For example, if $p = (x_1, y_1)$ and $q = (x_2, y_2)$, then $d_T(p, q) = |x_1 - x_2| + |y_1 - y_2|$. A heuristic function for an A* solution to the 8-piece puzzle problem uses Manhattan distance to quantify the closeness of the current game state to the desired solution. This approach works because progress *towards* the solution ultimately results in a global solution. The design of the heuristic function is key to the informed search technique.

Some problems contain *local extrema* (local but not global minimal or maximal values) that might stop the search at a suboptimal solution. If the problem has an extremely large space (there are too many candidate solutions to search exhaustively), then it may be acceptable to accept a "good-enough" local best candidate solution as an approximation for the global optimum. The *local search* technique uses an *ensemble* of *search agents* which independently search *neighborhoods* of the problem space. A local search could be built upon A* searches from different origins. Ideally, the A*-based local search should explore the problem with reasonable depth, mobility, and coverage [45].

Interestingly, the 8-piece and comparable 15-piece puzzle problems are not guaranteed to be solvable. Exactly half of all possible $9! = 362\,880$ and $15! = 1\,307\,674\,368\,000$ permutations are reachable from any random 8- and 15-piece puzzle game state [46]. n-piece puzzles ($n \geq 3$) contain two partitions and are therefore classified as *bipartite* graphs. The puzzle is unsolvable if started in the partition that does not contain the solution. We must understand that search algorithms might not be able to solve a problem if the graph of the problem space contains a partition.

6.5.5 A* and the Stable Marriage Problem

We will demonstrate the A* informed search algorithm on the *Stable Marriage Problem* [47]. The Stable Marriage Problem seeks to pair the members of two equal-sized sets to one another based upon their mutual preferences. This problem and its solution are applied to many practical situations, including the Army Talent Alignment Process (ATAP); see https://www.youtube.com/watch?v=9mEBe7fzrmI for an official and detailed explanation. Explained with marriages, the problem has all of the men rank all of the women from most to least preferred. Correspondingly, the women also rank all of the men from most to least preferred.

$$M = \begin{array}{r} \\ \text{Man 1} \\ \text{Man 2} \\ \text{Man 3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Woman 1} & \text{Woman 2} & \text{Woman 3} \end{array} \\ \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix} \end{array}$$

$$W = \begin{array}{r} \\ \text{Woman 1} \\ \text{Woman 2} \\ \text{Woman 3} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Man 1} & \text{Man 2} & \text{Man 3} \end{array} \\ \begin{bmatrix} 3 & 1 & 2 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{bmatrix} \end{array}$$

Each man is paired to one woman. The set of matching is considered "stable" when there is no "*rogue couple*": where a man x who prefers some other woman y to his current wife *and* where woman y prefers man x to her current husband. (The solution does not need to give everyone their first preference. It is acceptable for some person to prefer someone other than their current spouse; instability occurs only when that person requites.) In the above example, the pairing $[1, 2, 3]$ (man 1 paired to woman 1, man 2 with woman 2, 3 with 3) is stable. The pairing $[3, 2, 1]$ (man 1 paired to woman 3, man 2 with woman 2, 3 with 1) is not stable because man 1 perfers woman 2 to his current spouse (woman 3) and woman 2 prefers man 1 to her current spouse (man 2).

The Stable Marriage Problem is known to be solvable with a simple and predictable algorithm by Gale and Shapley [47]. First, each man proposes to his most-preferred woman. Women who receive multiple proposals maintain only their most-preferred proposal and reject the others. Second, each rejected man proposes to his next-preferred woman; women receiving new proposals continue to maintain only the one most-preferred. The process continues until no man is rejected in a round. Finally, the women accept their most-preferred proposal and the algorithm terminates.

A *reduction* is a method of solving one problem by restating it in terms of another. Reductions can be a powerful but difficult technique for solving hard problems. By reducing the stable marriage problem to a graph search problem, we can solve the problem with an A* algorithm. Our A* solution will be slower and less efficient than the Gale-Shapley algorithm, but it will demonstrate a method for solving difficult problems using general and reusable techniques. The 8-piece

puzzle problem could likely also be solved with a very fast and direct algorithm, but it may be very difficult for us to discover this solution. Likewise, there may be an optimal algorithm to solve a Rubik's cube, but given an A* solver and a fast computer we may be able to find an economically-acceptable solution.

Our informed search algorithm is implemented in Julia. This is comparable our implementation of Dijkstra's algorithm (see section 6.5.3), but instead of searching for a named destination this function instead uses its heuristic function. If the heuristic function returns the value zero, then the search is considered successful and the program terminates. This A* program also prints its search in the DOT graph description language, which can be rendered as a graphic using the GraphViz program.

```
using DataStructures

function informed_search(source, edges::Function, heuristic::Function)
    println("digraph {")

    pq = PriorityQueue()
    visited = Set()
    enqueue!(pq, source⇒heuristic(source))

    while !isempty(pq)
        u = dequeue!(pq)
        push!(visited, u)
        println("\"$(u)\" [color=\"blue\"];")

        if heuristic(u) == 0
            println("}")
            return
        end

        for v ∈ edges(u)
            if v ∉ visited && !haskey(pq, v)
                enqueue!(pq, v⇒heuristic(v))
                println("\"$(u)\" -> \"$(v)\";")
            end
        end
    end

    error("Failed to find a solution.")
end
```

The heuristic function for the stable marriage function seeks to quantify and differentiate instability by returning the sum of the squared distance (see section 4.1) of a rogue couple's candidate and current preferences.

```
function stability(men::Matrix, women::Matrix, matching)
```

```
    n = length(matching)
    wife = matching
    husband = Dict(values(matching) .⇒ keys(matching))
    metric = 0

    for man ∈ 1:n
        for woman ∈ 1:n
            # Candidate and current preferences for the man
            x1, x2 = men[man, woman], men[man, wife[man]]
            # Candidate and current preferences for the woman
            y1, y2 = women[woman, man], women[woman, husband[woman]]
            # The matching is unstable if, and only if, both the man
            # and the woman prefer each other to their current matches.
            if x1 < x2 && y1 < y2
                metric += (x1 - x2)^2
                metric += (y1 - y2)^2
            end
        end
    end

    return metric
end
```

The size of our problem comes from the number of possible matchings. If there are n men and n women, then the first man can be matched to n women, the second man can be matched to $n-1$ women, and so on until the last man can only be matched to the only remaining woman. Thus,

$$\text{Problem size} = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 3 \cdot 2 \cdot 1 = n!$$

A problem of factorial size is a large combinatorial problem. We will not attempt to visit all possible nodes in the problem space. Instead, our edge function will generate three permutations of the current position by switching two matchings.

```
using StatsBase

function e(u)
    v = Set()
    for _ ∈ 1:3
        x = copy(u)
        # Sample, without replacement, two random elements to swap.
        y = sample(collect(eachindex(u)), 2, replace=false)
        x[y[1]], x[y[2]] = x[y[2]], x[y[1]]
        push!(v, x)
    end
    return v
```

```
end
```

Our ranking matrices for men, M, and women, W are taken from example 2 from Gale and Shapley [47].

$$M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \\ 2 & 1 & 3 & 4 \\ 4 & 2 & 3 & 1 \end{bmatrix}$$

$$W = \begin{bmatrix} 3 & 4 & 2 & 1 \\ 3 & 1 & 4 & 2 \\ 2 & 3 & 4 & 1 \\ 3 & 2 & 1 & 4 \end{bmatrix}$$

The Julia language has a compact notation to create matrix literals.

```
M = [1 2 3 4; 1 4 3 2; 2 1 3 4; 4 2 3 1]
W = [3 4 2 1; 3 1 4 2; 2 3 4 1; 3 2 1 4]
```

We construct a *closure* to encapsulate M and W with our stability metric function into the heuristic function.

```
h(u) = stability(M, W, u)
```

This syntax declares a unary function `h` that will enable our A* to navigate matrices M and W without direct knowledge of either.

Finally, we invoke the A* informed search algorithm to navigate the Stable Marriage Problem as a graph, starting from pairings $[4, 3, 2, 1]$ (man 1 matched to woman 4, man 2 to woman 3, 3 to 2, and 4 to 1).

```
julia> informed_search([4,3,2,1], e, h)
digraph {
"[4, 3, 2, 1]" [color="blue"];
"[4, 3, 2, 1]" -> "[4, 2, 3, 1]";
"[4, 3, 2, 1]" -> "[4, 1, 2, 3]";
"[4, 3, 2, 1]" -> "[4, 3, 1, 2]";
"[4, 3, 1, 2]" [color="blue"];
"[4, 3, 1, 2]" -> "[4, 2, 1, 3]";
"[4, 3, 1, 2]" -> "[4, 1, 3, 2]";
"[4, 2, 1, 3]" [color="blue"];
"[4, 2, 1, 3]" -> "[2, 4, 1, 3]";
"[4, 2, 1, 3]" -> "[1, 2, 4, 3]";
"[2, 4, 1, 3]" [color="blue"];
"[2, 4, 1, 3]" -> "[2, 1, 4, 3]";
"[2, 4, 1, 3]" -> "[3, 4, 1, 2]";
"[3, 4, 1, 2]" [color="blue"];
}
```

(Note: this *stochastic* algorithm uses randomness in the `sample` operation. Outputs are not deterministic. In rare cases, this procedure may not discover the one and only solution, $[3, 4, 1, 2]$. See https://github.com/wjholden/Data-Literacy/blob/main/StableMarriageSearch.jl for an expanded version of this program which uses a seeded random number generator for reproducibility.)

We can input this `digraph` data into https://dreampuf.github.io/GraphvizOnline/ to visualize the search tree, as shown in figure 6.4.

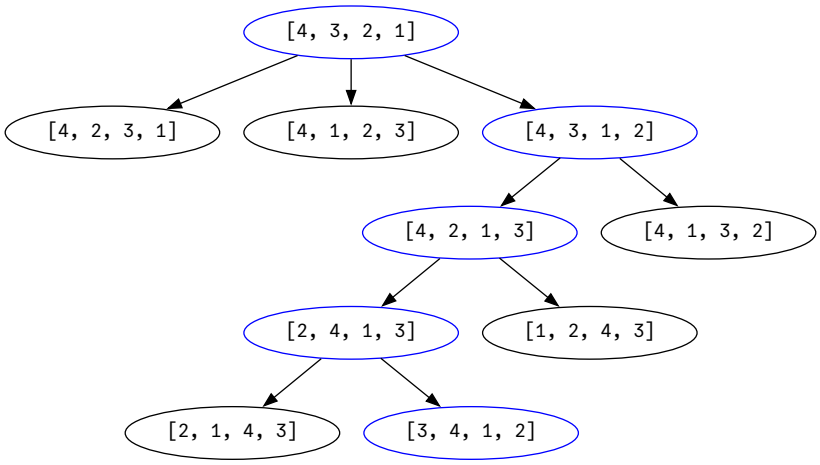


Figure 6.4: A search tree of the Stable Marriage Problem, reduced to an informed search that is solvable with A*.

Again, we applied informed search to the Stable Marriage Problem as an exercise in artificial intelligence methods. Though slow, an informed search can navigate difficult problems with very little information: a simple heuristic function tells A* whether it has gotten closer or farther from the solution. This technique can be useful for solving challenging problems where an optimal solution is not known. Moreover, we can also apply informed search to *intractable* problems where computational complexity forces us to accept approximate solutions as a compromise.

6.6 Centrality

6.6.1 Degree

The average person (ten of twelve) in the social graph shown in figure 6.5 has *fewer* friends the average friend count of their friends. These statistics are summarized in the following table.

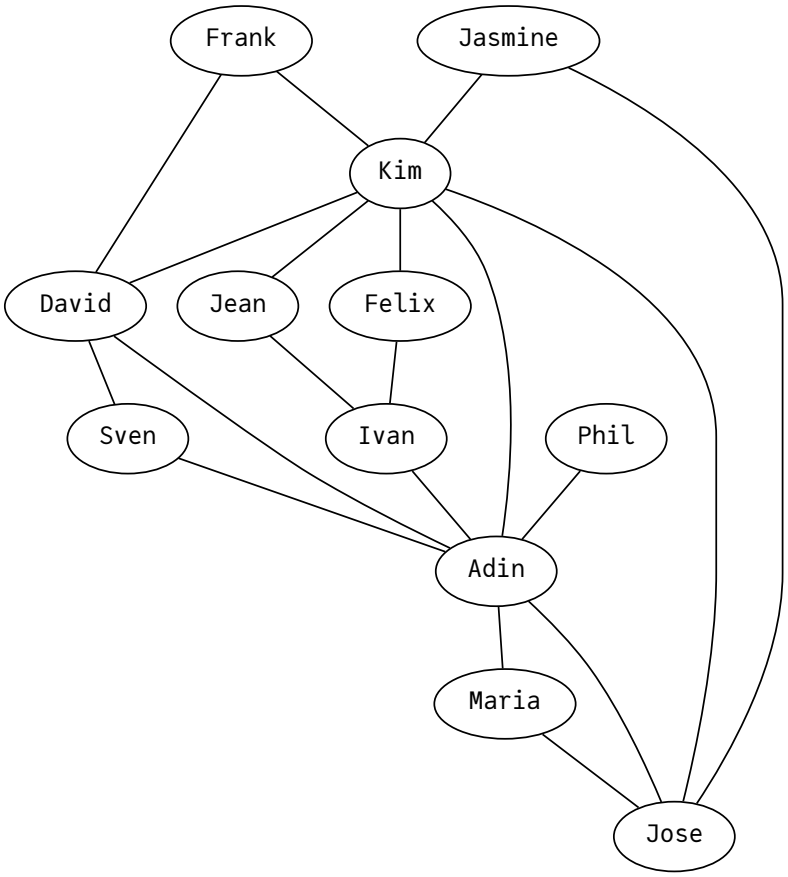


Figure 6.5: Arithmetic in graphs can produce unintuitive results. In this graph, 10 of 12 vertices has a lower degree than the average among its adjacent nodes.

Name	Friends	Average friend's friend count
Frank	2	5
Jasmine	2	5.5
Kim	7	3.14
David	3	3.67
Sven	2	5
Felix	2	5
Jean	2	5
Ivan	3	3.67
Adin	7	3.14
Maria	2	5.5
Jose	4	4.5
Phil	1	7

How can this be possible? We must be careful with definitions. The average number of friends in this graph is 3.08, but we are not looking at the entire graph. Instead, we are only looking at a subgraph. Look closely at Phil. Phil has only one friend, Adin, and Adin is highly-connected. In fact, Adin has more friends than all of his friends, except Kim.

Highly-connected nodes, with outlier degree, can be particularly important in many graph applications. Transportation networks are an example: congestion at any major airport "hub" can quickly spread to adjacent airports and beyond.

Degree centrality is a simple and intuitive graph statistic [48]. Simply count the in-degree, out-degree, or both, and use this metric to discover important nodes in the graph.

6.6.2 Closeness

Closeness centrality, C_C, is also fairly simple to compute from the sum of unweighted distances from each node to each other node [49]. Closeness centrality has been used to study criminal and terrorist networks [50].

$$C_C(x) = \frac{1}{\sum_{y \in v} \delta(x, y)}$$

The following Rust program uses an unweighted invocation of Dijkstra's algorithm in the `petgraph` crate[4]. Run this program at the Rust Playground[5].

```
use petgraph::algo::dijkstra;
use petgraph::prelude::*;
```

[4]The term *crate* is peculiar to Rust. In this context, the term is interchangeable with "library," which is a more common programming term.

[5]https://play.rust-lang.org/?gist=e67ef3bc05daab21dd73a7869093d9cb

```
use petgraph::visit::NodeRef;
use petgraph::Graph;

fn main() {
    let mut graph: Graph<(), (), Undirected> = Graph::new_undirected();
    let a = graph.add_node(());
    let b = graph.add_node(());
    let c = graph.add_node(());
    let d = graph.add_node(());
    let e = graph.add_node(());
    graph.extend_with_edges(&[(a, b), (b, c), (c, d), (d, e)]);
    let graph = graph;
    closeness(&graph);
}

fn closeness<N, E>(graph: &Graph<N, E, Undirected>) {
    for u in graph.node_indices() {
        let delta = dijkstra(&graph, u.id(), None, |_| 1);
        let n = graph.node_count() as f64;
        let distances = delta.values().cloned().sum::<i32>() as f64;
        let closeness = (n - 1.0) / distances;
        println!(
            "Closeness score for vertex {} is {}.",
            u.index(),
            closeness
        );
    }
}
```

The output of this program should be

```
Closeness score for vertex 0 is 0.4.
Closeness score for vertex 1 is 0.5714285714285714.
Closeness score for vertex 2 is 0.6666666666666666.
Closeness score for vertex 3 is 0.5714285714285714.
Closeness score for vertex 4 is 0.4.
```

As an exercise, draw the graph described in this program (either by hand or using software) and assess if the centrality statistics seem intuitive.

6.6.3 Betweenness

Betweenness centrality, C_B, seeks to improve upon closeness by considering edge weights [51]. The betweenness for a vertex x is calculated the proportion of shortest paths $\sigma(s,t)$ that include x as an intermediary node along the path.

$$C_B(x) = \sum_{s \neq x, x \neq t, s \neq t} \frac{\sigma(s, t|x)}{\sigma(s, t)}$$

The computational complexity of so many pathfindings becomes significant in large graphs and dense graphs [52] [53].

6.6.4 PageRank

Google's well-known *PageRank* algorithm uses a creative edge weighting function based on a page's importance or quality [54]. The metric is parameterized with a dampening factor, d, and requires multiple iterations.

$$\mathrm{PR}(u) = \sum_{v \in B_u} \frac{\mathrm{PR}(v)}{L(v)}$$

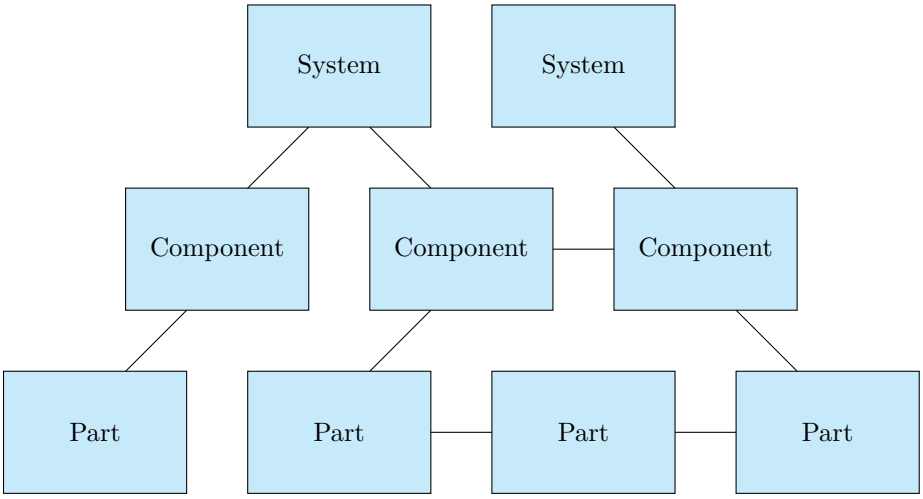
This peculiar concept of a page's "reputation" is central to the efficacy of the Google search engine.

There are many more measures of centrality in graphs beyond degree centrality, closeness, betweenness, and PageRank. Depending on the application, any or none of these metrics may be useful. The analyst studying a graph might consider these metrics "man-made"; they were created by scientists and mathematicians seeking to model a problem. These metrics can be customized to the user's needs to best approach their unique problem domain.

6.7 Discussion prompts

1. A graph can be represented with an adjacency list or a matrix. What are the advantages and disadvantages of each approach?
2. What algorithm can be used to solve the "seven ways to Kevin Bacon" problem?
3. Is a Gantt chart a graph? How can one find the critical path of a project if represented as a graph?
4. If the distance from Paris to Sydney is infinitely far, then can we use some *greater infinity* to represent the distance from London to Sydney?
5. Think of a practical problem that can be modeled as a graph, but where the four discussed measures of center (degree centrality, closeness, betweenness, and PageRank) are not effective. As a discussion point, consider whether values immediately associated with vertices and edges dominate their importance, or if some extrinsic network effect has a greater effect.

6. A manufacturer sells systems that are made of components. Those components are assembled from atomic parts. Many parts are interchangeable with other parts, and many components are interchangeable with other components. How can the manufacturer discover unused or duplicative parts and components?



6.8 Practical exercises

1. Convert currency exchange rates from multiplication to addition using a logarithm, then prove that infinite arbitration is impossible given a set of exchange rates and Bellman-Ford implementation.
2. Define a topological sorting and relate it to a workplace problem.
3. Define the Traveling Salesman Problem (TSP) and explain the computational difficulty of this problem.
4. Determine the minimum paving needed to fully connect a tent complex using a list of coordinates and a Prim or Kruskal implementation.
5. Simulate an infection model in a dense social graph where edge weights represent probability of infection.

References

[1] J. Rowley, "The wisdom hierarchy: Representations of the DIKW hierarchy," *Journal of Information Science*, vol. 33, no. 2, pp. 163–180, 2007, doi: 10.1177/0165551506070706. Available: https://doi.org/10.1177/0165551506070706

[2] G. Glassman, "Science is successful prediction," presented at the Broken science initiative epistemology camp, 2024. Available: https://www.youtube.com/watch?v=bpQyjZoF5EA

[3] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach (4th edition)*. Pearson, 2020. Available: http://aima.cs.berkeley.edu/

[4] K. W. Church and W. A. Gale, "Probability scoring for spelling correction," *Statistics and Computing*, vol. 1, pp. 93–103, 1991.

[5] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946, doi: 10.1126/science.103.2684.677

[6] T. Hoare, "Null references: The billion dollar mistake," presented at the QCon, 2009. Available: https://www.infoq.com/presentations/Null-References-The-Billion-Dollar-Mistake-Tony-Hoare/

[7] Y. Shafranovich, "Common Format and MIME Type for Comma-Separated Values (CSV) Files." in Request for comments. RFC 4180; RFC Editor, Oct. 2005. doi: 10.17487/RFC4180. Available: https://www.rfc-editor.org/info/rfc4180

[8] B. M. Weisenthal, C. A. Beck, M. D. Maloney, K. E. DeHaven, and B. D. Giordano, "Injury rate and patterns among CrossFit athletes," *Orthop. J. Sports Med.*, vol. 2, no. 4, p. 2325967114531177, Apr. 2014, Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4555591/

[9] M. H. Lietzke, "Relation between weight-lifting totals and body weight," *Science*, vol. 124, no. 3220, pp. 486–487, 1956, doi: 10.1126/science.124.3220.486. Available: https://www.science.org/doi/abs/10.1126/science.124.3220.486

[10] B. Shulman, "Math-alive! Using original sources to teach mathematics in social context," *PRIMUS*, vol. 8, no. 1, pp. 1–14, 1998, doi: 10.1080/10511979808965879. Available: https://doi.org/10.1080/10511979808965879

[11] D. E. Knuth, "Literate programming," *The Computer Journal*, vol. 27, no. 2, pp. 97–111, Jan. 1984, doi: 10.1093/comjnl/27.2.97. Available: https://doi.org/10.1093/comjnl/27.2.97

[12] G. Michaelson, "Programming paradigms, turing completeness and computational thinking," *The Art, Science, and Engineering of Programming*, vol. 4, Feb. 2020, doi: 10.22152/programming-journal.org/2020/4/4. Available: https://programming-journal.org/2020/4/4/

[13] E. W. Dijkstra, "On the foolishness of 'natural language programming'," 1978. Available: http://www.cs.utexas.edu/users/EWD/ewd06xx/EWD667.PDF

[14] A. C. Bock and U. Frank, "Low-code platform," *Business & Information Systems Engineering*, vol. 63, pp. 733–740, 2021, doi: 10.1007/s12599-021-00726-8. Available: https://link.springer.com/article/10.1007/s12599-021-00726-8

[15] E. F. Codd, "A relational model of data for large shared data banks," *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970, doi: 10.1145/362384.362685

[16] S. Cai, B. Gallina, D. Nyström, and C. Seceleanu, "Data aggregation processes: A survey, a taxonomy, and design guidelines," *Computing*, vol. 101, pp. 1397–1429, 2019, doi: 10.1007/s00607-018-0679-5. Available: https://link.springer.com/content/pdf/10.1007/s00607-018-0679-5.pdf

[17] E. H. Simpson, "The interpretation of interaction in contingency tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 13, no. 2, pp. 238–241, Dec. 2018, doi: 10.1111/j.2517-6161.1951.tb00088.x. Available: https://doi.org/10.1111/j.2517-6161.1951.tb00088.x

[18] White House Office of the National Cyber Director, "Back to the building blocks: A path toward secure and measurable software," 2024. Available: https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/02/Final-ONCD-Technical-Report.pdf

[19] T. Rentsch, "Object oriented programming," *SIGPLAN Not.*, vol. 17, no. 9, pp. 51–57, Sep. 1982, doi: 10.1145/947955.947961. Available: https://doi.org/10.1145/947955.947961

[20] S. Klabnik and C. Nichols, *The rust programming language*. USA: No Starch Press, 2018.

[21] T. Bray, "The JavaScript Object Notation (JSON) Data Interchange Format." in Request for comments. RFC 8259; RFC Editor, Dec. 2017. doi: 10.17487/RFC8259. Available: https://www.rfc-editor.org/info/rfc8259

[22] E. Brewer, "CAP twelve years later: How the 'rules' have changed," *Computer*, vol. 45, no. 2, pp. 23–29, 2012, doi: 10.1109/MC.2012.37

[23] A. M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris, France: Firmin Didot, 1805, pp. 72–75.

[24] A. M. Legendre, "On least squares," in *A source book in mathematics*, New York, NY, USA: McGraw-Hill, 1929, pp. 576–579.

[25] K. Pearson, "'Das fehlergesetz und seine verallgemeiner-ungen durch Fechner und Pearson*.' A rejoinder," *Biometrika*, vol. 4, no. 1–2, pp. 169–212, Jun. 1905, doi: 10.1093/biomet/4.1-2.169. Available: https://doi.org/10.1093/biomet/4.1-2.169

[26] J. M. Wicherts, C. L. S. Veldkamp, H. E. M. Augusteijn, M. Bakker, R. C. M. van Aert, and M. A. L. M. van Assen, "Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking," *Frontiers in Psychology*, vol. 7, 2016, doi: 10.3389/fpsyg.2016.01832. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01832

[27] I. B. Myers, *The Myers-Briggs type indicator: Manual (1962)*. Palo Alto: Consulting Psychologists Press, 1962.

[28] M. R. Garey and D. S. Johnson, *Computers and intractability; a guide to the theory of NP-completeness*. USA: W. H. Freeman & Co., 1990.

[29] M. Hořeňovský, "Modern SAT solvers: Fast, neat and underused (part 1 of N)." Aug. 03, 2018. Available: https://codingnest.com/modern-sat-solvers-fast-neat-underused-part-1-of-n/

[30] L. de Moura and N. Bjørner, "Z3: An efficient SMT solver," in *Tools and algorithms for the construction and analysis of systems*, C. R. Ramakrishnan and J. Rehof, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 337–340. doi: 10.1007/978-3-540-78800-3_24

[31] B. Felgenhauer and F. Jarvis, "Enumerating possible sudoku grids," 2005, Available: http://www.afjarvis.org.uk/sudoku/sudoku.pdf

[32] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the third annual ACM symposium on theory of computing*, in STOC '71. New York, NY, USA: Association for Computing Machinery, 1971, pp. 151–158. doi: 10.1145/800157.805047. Available: https://doi.org/10.1145/800157.805047

[33] D. Yurichev, "SAT/SMT by Example." May 12, 2024. Available: https://smt.st

[34] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, Jun. 1938, doi: 10.1093/biomet/30.1-2.81. Available: https://doi.org/10.1093/biomet/30.1-2.81

[35] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, Nov. 1945, doi: 10.1093/biomet/33.3.239. Available: https://doi.org/10.1093/biomet/33.3.239

[36] M. Friendly, "Corrgrams: Exploratory displays for correlation matrices," *The American Statistician*, vol. 56, no. 4, pp. 316–324, 2002, doi: 10.1198/000313002533. Available: https://www.datavis.ca/papers/corrgram.pdf

[37] S. Chatterjee, "A new coefficient of correlation," *Journal of the American Statistical Association*, vol. 116, no. 536, pp. 2009–2022, 2021, doi: 10.1080/01621459.2020.1758115. Available: https://arxiv.org/abs/1909.10140

[38] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901, doi: 10.1080/14786440109462720. Available: https://doi.org/10.1080/14786440109462720

[39] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, pp. 417–441, 1933, doi: dx.doi.org/10.1037/h0071325. Available: https://www.cis.rit.edu/~rlepci/Erho/Derek/Useful_References/Principal%20Components%20Analysis/Hotelling_PCA_part1.pdf

[40] J. Wendler, *5/3/1: The simplest and most effective training system to increase raw strength*. Lulu.com, 2011.

[41] M. Brzycki, "Strength testing—predicting a one-rep max from reps-to-fatigue," *Journal of Physical Education, Recreation & Dance*, vol. 64, no. 1, pp. 88–90, 1993, doi: 10.1080/07303084.1993.10606684

[42] A. Mayerowitz, "Mario meets Pareto." https://www.mayerowitz.io/blog/mario-meets-pareto, 2024.

[43] E. W. Dijkstra, "A note on two problems in connexion with graphs," in *Edsger Wybe Dijkstra: His Life,Work, and Legacy*, 1st ed.New York, NY, USA: Association for Computing Machinery, 2022, pp. 287–290. doi: 10.1145/3544585.3544600

[44] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968, doi: 10.1109/TSSC.1968.300136

[45] D. Schuurmans and F. Southey, "Local search characteristics of incomplete SAT procedures," *Artificial Intelligence*, vol. 132, no. 2, pp. 121–150, 2001, doi: 10.1016/S0004-3702(01)00151-5

[46] Wm. W. Johnson and W. E. Story, "Notes on the '15' puzzle," *American Journal of Mathematics*, vol. 2, no. 4, pp. 397–404, 1879, doi: 10.2307/2369492. Available: http://www.jstor.org/stable/2369492

[47] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962, doi: 10.1080/00029890.1962.11989827

[48] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978, doi: https://doi.org/10.1016/0378-8733(78)90021-7. Available: https://www.sciencedirect.com/science/article/pii/0378873378900217

[49] A. Bavelas, "Communication patterns in task-oriented groups," *The Journal of the Acoustical Society of America*, vol. 22, no. 6, pp. 725–730, Nov. 1950, doi: 10.1121/1.1906679. Available: https://doi.org/10.1121/1.1906679

[50] V. E. Krebs, "Mapping networks of terrorist cells," in *Connections*, 2002, pp. 43–52. Available: http://www.orgnet.com/MappingTerroristNetworks.pdf

[51] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977, Available: http://www.jstor.org/stable/3033543. [Accessed: Sep. 26, 2024]

[52] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001, doi: 10.1080/0022250X.2001.9990249. Available: http://snap.stanford.edu/class/cs224w-readings/brandes01centrality.pdf

[53] U. Brandes and C. Pich, "Centrality estimation in large networks," *International Journal of Bifurcation and Chaos*, vol. 17, no. 7, pp. 2303–2318, 2007, Available: https://www.uni-konstanz.de/mmsp/pubsys/publishedFiles/BrPi07.pdf

[54] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, 1998, doi: https://doi.org/10.1016/S0169-7552(98)00110-X. Available: https://www.sciencedirect.com/science/article/pii/S016975529800110X