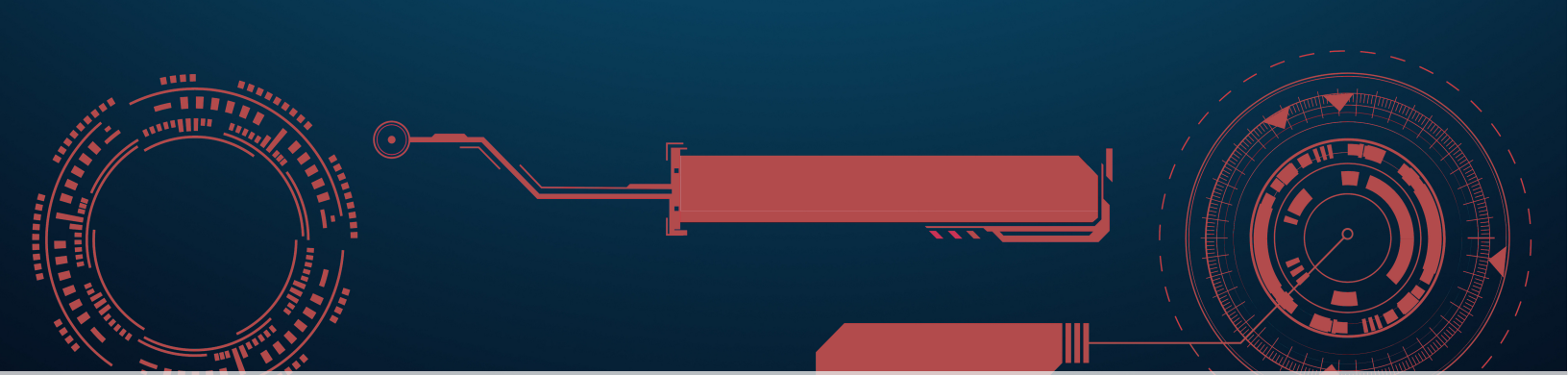


2 Data Visualization and Exploration

Until very recently, most data problems in human endeavours have been linked to **engineering** (that is to say, to the design of objects and machines) and to the **sciences** (namely, the formulation of theories and falsification of hypotheses).

For instance, engineers may equip their machines with sensors and use the data that they collect to assess and evaluate the machines' behaviours under various controlled conditions and, ultimately, to improve their functionality.

Scientists, on the other hand, typically collect data through experimental design to test the validity of their theories. But scientific experiments are expensive;[1] and generate few data points, relatively-speaking.

As data scientist D. Mingle puts it, however, modern data analysis is a different beast:

> Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualization. [27]

In the 21st century, not only is there more data to collect and analyze, but it overwhelmingly comes in a **digital** format (as opposed to the traditional analog paper format) and is mostly derived from **observations** (rather than generated by designed experiment).

Data problems are still solved **empirically**, **theoretically**, and through **computation** and **simulation**, as has been the case historically,[2] but also *via* **data exploration** and **data visualization**.

So what can actually be done with the data, once it has been collected and processed? We think of

- **analysis** as the collection of processes by which we extract actionable insights from the data, and
- **visualization** as the process of presenting data, calculations, and analysis outputs in a visual format.

Visualization of data *prior* to analysis (**data exploration**) can help simplify the analytical process; visualization *following* analysis (**communication**) allows for analysis results to be presented to various stakeholders (see Figure 2.1).

1: The cost of finding the Higgs Boson at CERN's Large Hadron Collider was estimated as 13.25 billion 2012 USD [26].

2: There were exceptions, of course. Richard Feynman, for instance, was known to "solve problems by putting himself in the place of an atom or an electron, essentially asking himself what he would do if he were an atomic or subatomic particle. [28]"

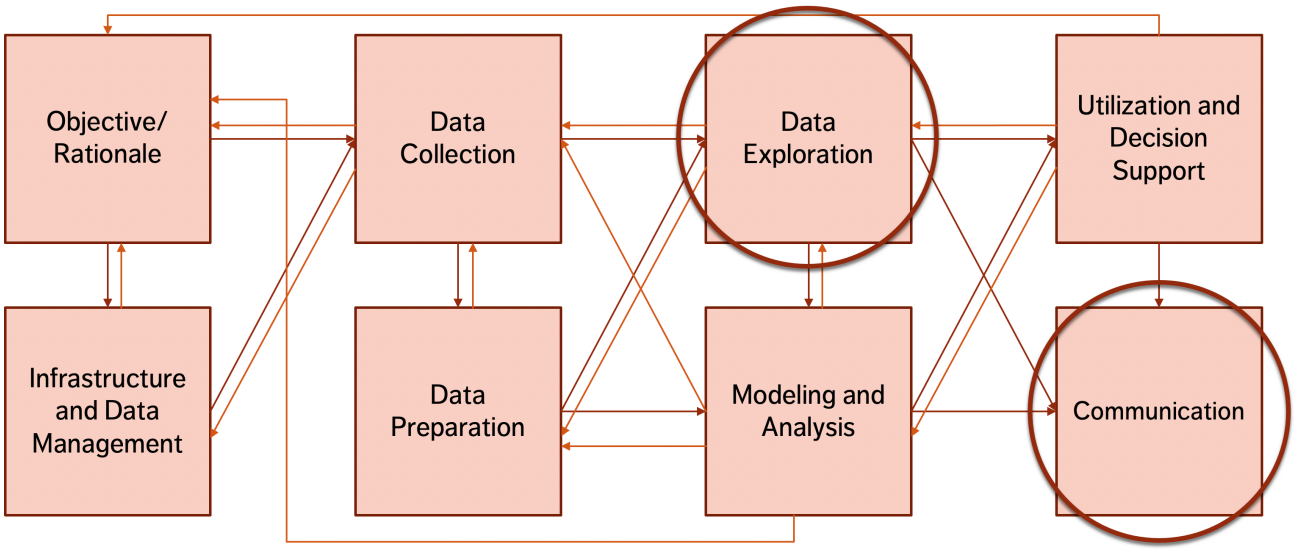


Figure 2.1: The (messy) analytical process is iterative and allows for multiple false starts. Visualization plays a major role in the data exploration and communication stages.

2.1 Exploratory Data Analysis

In this section, we focus on the role of data visualization **prior to analysis**.

Pre-Analysis Uses

Prior to the analysis of the data proper, it is paramount for the data to be explored and for basic questions to be asked (and answered):

- what system does the data represent, in terms of objects, attributes, relationships?
- how does it represent this system? in other words, what is the data model?
- where does the data come from? who collected it and processed it? when did this take place? for what purpose?
- assuming that the data comes in a **flat file** format, what do the rows represent? what about the columns?
- is there enough information (or **metadata**) to answer these questions? where could more information be found?

In the data exploration context, data visualization is typically used to set the stage by helping analysts (as in Figure 2.2):

- detect **invalid entries** and **outliers**;
- shape the **data transformations** (binning, standardization, Box-Cox transformations, dimension reduction, etc.);
- get a **sense for the data** (data analysis as an art form, exploratory analysis), and
- identify **hidden data structures** (clustering, associations, patterns which may inform the next stage of analysis, etc.).

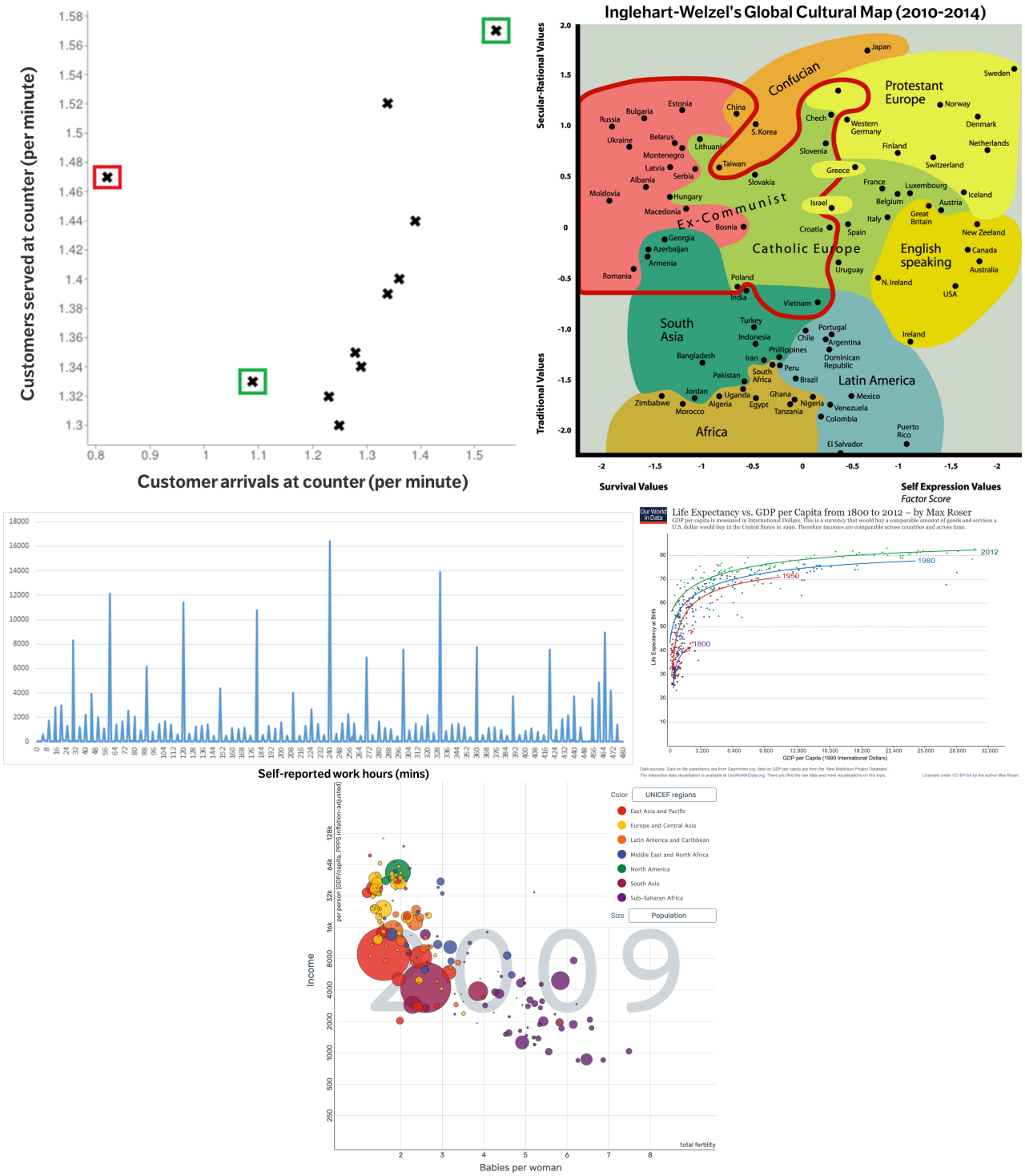


Figure 2.2: Pre-analysis uses of data visualization: outliers in wait time data (personal file, top left); hidden data structure in the Inglehart-Wetzel's Global Cultural Map (2010-2014) (author unknown, top right); heaping in self-reported work hours (personal file, middle left); shaping transformations in life expectancy and income per person data (Our World in Charts, middle right); shaping data transformations in 2009 income relative to fertility rates (Gapminder Foundation).

Data Exploration in Action

Consider a subset of the *algae blooms* dataset consisting of 4 variables (`Cl`, `NO3`, `NH4`, `season`) and 340 observations, found in the *UCI Machine Learning Repository* [29, 30].[3] Its first few observations are as follows:

3: The dataset and its context is also analyzed as a case study in [31] and in [1].

season	Cl	NO3	NH4
winter	60.800	6.238	578.000
spring	57.750	1.288	370.000
autumn	40.020	5.330	346.667
spring	77.364	2.302	98.182
autumn	55.350	10.416	233.700
winter	65.750	9.248	430.000

A **univariate** summary is provided below:

season	statistic	Cl	NO3	NH4
autumn: 80	Min:	0.222	0.000	5.00
spring: 84	Q1:	10.994	1.147	37.86
summer: 86	Median:	32.470	2.356	107.36
winter: 90	Mean:	42.517	3.121	471.73
	Q3:	57.750	4.147	244.9
	Max:	391.500	45.650	24064.00
	NAs:	16	2	2

Is it possible to determine what **system** the data represent from the summary alone? Or where it comes from, why it was collected, and so on? One would need a fair amount of clairvoyance to answer these questions without metadata.[4]

4: Give it a try now; what can you come up with?

As it happens, the algae bloom dataset is a collection of chemical, biological, and physical characteristics related to samples of European rivers taken over a one-year period, with the goal of "protecting rivers and streams by monitoring chemical concentrations and algae communities [30]."

With this context in hand, we understand that `Cl`, `N03`, and `NH4` represent, respectively, the "concentration" of **chlorine**, **nitrate**, and **ammonium** in various European river samples, collected over a one-year period.

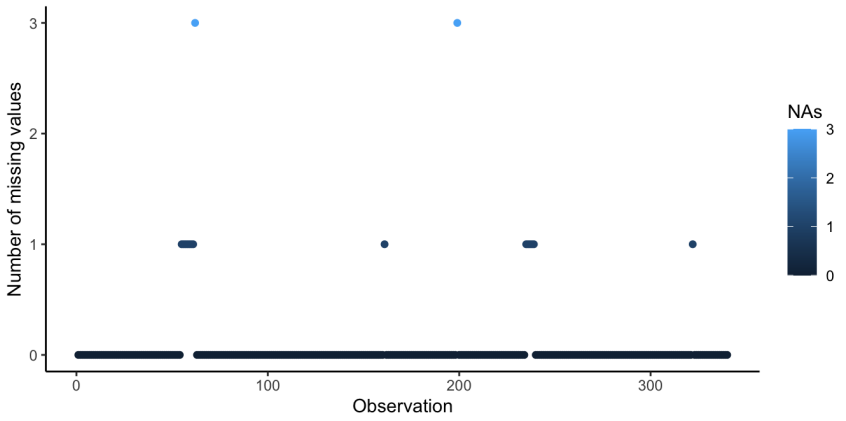
The numerical summary above provides us with a number of item of interests:

1. the distribution of samples during the year seems fairly uniform, with nearly a quarter of the samples collected in each season;
2. there are, respectively, 16, 2, and 2 observations for which the levels of `Cl`, `NO3`, and `NH4` are unavailable (although no information is available to indicate whether any of the observations have multiple missing values);

3. all available measurements are non-negative, as befit concentration levels for various chemical compounds;
4. the measurement ranges for each numerical variable have different magnitudes (≈ 400 for `Cl`, 50 for `NO3`, and 25000 for `NH4`);
5. the jump between the 3rd quartile and the maximum measurement for `Cl` and `NO3` is ofone order of magnitude, but it of two orders of magnitude for `NH4`;
6. and so on.

While a fair amount of insight can be derived from that particular numerical summary, a number of questions remain unanswered. Let us explore this dataset further.

Can we get a more sophisticated understanding than the one provided by the numerical summary? In the figure below, for instance, we see that two of the dataset instances have exactly 3 **missing values** (`Cl`, `NO3`, and `NH4`); the 14 remaining observations with missing values are those for which only Cl is unavailable.

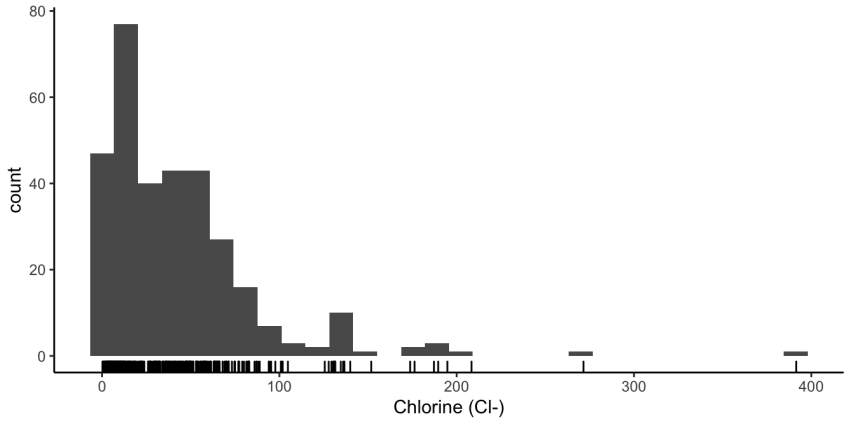


The same chart also shows that there are two contiguous blocks of observations with missing values (around observations 66 and 225, roughly speaking). This suggests that there could be **data collection issues** – a group of measurement slips might have been misplaced, or a student intern might have misunderstood how to test the water samples, say – but as no information is available about the process, we can at best provide **possible** (if not potential) explanations for the existence of this pattern, which may only prove to be an artifact of the sorting process, in the final analysis.

We can also expand our understanding of the various measurements by plotting **univariate distributions** instead of relying on their respective 6-point numerical summaries. For instance, the non-missing values of `Cl` (of which there are 324) range from 0.222 to 391.500, with a mean level of 42.517 and a median of 32.470. What does any of this mean, in practice?

From the value of the **median**, we know that half the measurements fall between 0.222 and 32.470, and half fall between 32.470 and 391.500. The observations of the second half fall in a longer interval than those of the first

half, so we would expect the measurement to be **denser in the high-level regime** than in the low-level one. This is borne out by the histogram of `Cl` measurements.



The numerical summary hints at the presence of **outliers** in these measurements (since the median is substantially smaller than the mean), and the visual display provides a clear picture – the measurements above 250 are quite likely to be outliers (either due to **measurement errors** or because they are indicative of particularly **unrepresentative sampling sites**).

The small clusters of observations around the 150 and the 200 marks in the image above are also suspicious, but it could simply be the reality of the measurements in the field. Perhaps these measurements were taken downstream of some chemical factory, say?

Without observation-specific context, it is nearly impossible to gauge how likely the above explanation holds, but at the very least, the chart highlights potential problem areas that any eventual analysis will have to address.

Additional charts (for data containing all the observations and for data in which certain outliers ahve been removed) are presented in Figure 2.3; is enough information present to get a good understanding of the data **prior** to analysis? Do any questions/problems naturally arise from this exploration?

At the exploration stage, the **chart types** and the **chart aesthetics** are secondary – the main objective is to help analysts **get a sense** for the data.[5]

5: We used `ggplot2` to produce high-quality R charts for the algae blooms data (see Chapters 11 and 12 for details).

It is only when data insights (obtained through analysis and/or visualization) need to be **communicated** to an audience that **storytelling** and **design** must be taken into consideration (see Chapters 4, 5, 7, and 8).

Presently, we introduce some of the more commonly-used (and simplest) visualization procedures; we will also discuss other methods in Chapter 9 (*Visualization Toolkit*).

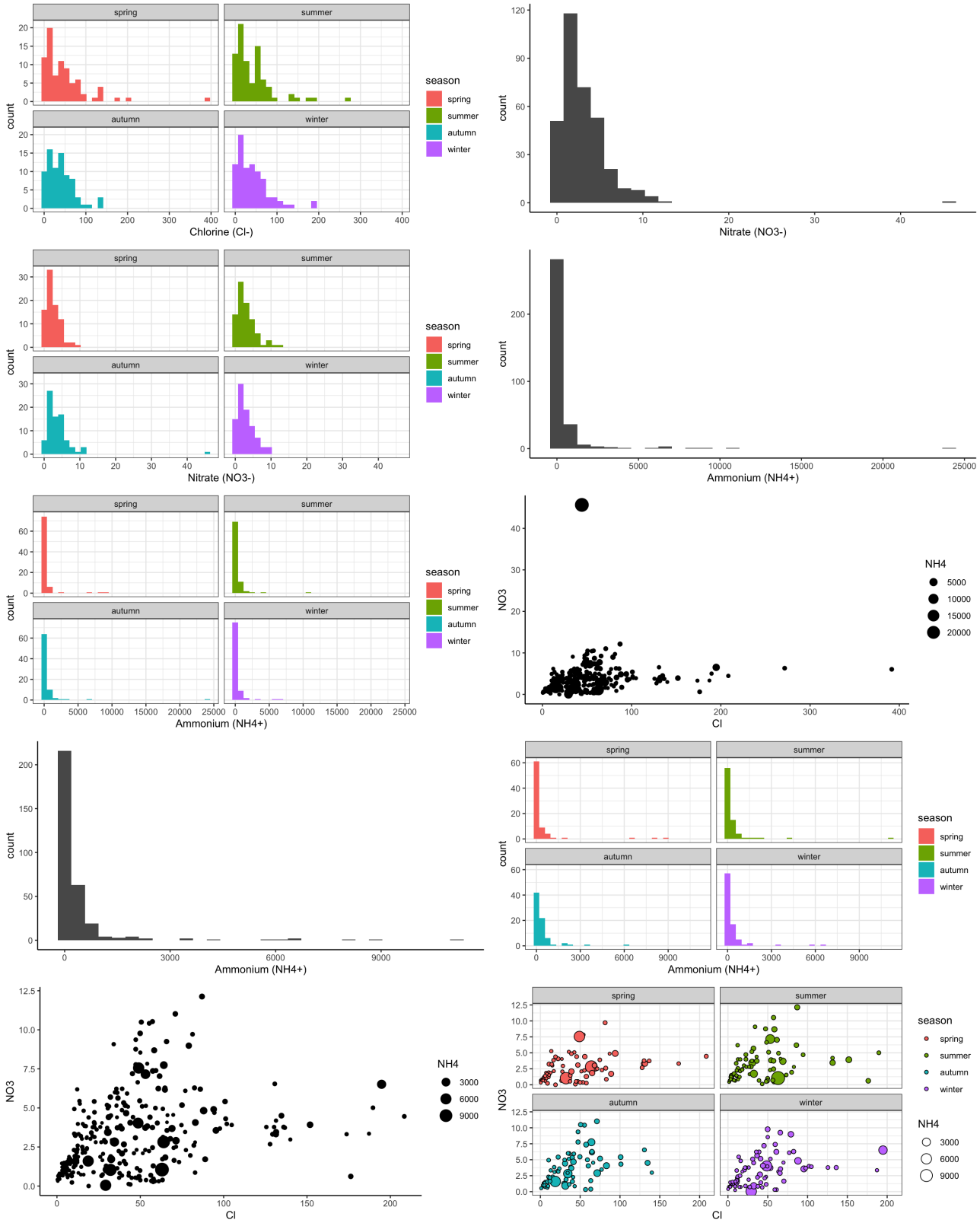


Figure 2.3: Exploratory data analysis of the algae blooms dataset: all observations (first three rows); with outliers removed (2 bottom rows).

2.2 Workhorse Data Visualizations

Some types of simple charts are used extensively for data exploration (as well as for data communication and storytelling).[6] We will be providing more in-depth explanations in Chapter 9 (*Visualization Toolkit*).

6: Popularity is no promise of quality, however, which is why we will not be discussing the ubiquitous *pie chart*.

Rug Plots

Rug plots, also known as comb charts, indicate on the number line the presence of a numerical value in the data; the gaps, therefore, display the absence of *those* values in the data. They are **univariate** data visualizations.

The numerical order in which the values appear is (possibly) different from the order in which the values appear in the data. Furthermore, if two observations have the same value, they are plotted on top of one other; it is thus impossible to determine exactly **how many** observations are represented by the markings.

Figure 2.4: The rug plot of an artificial dataset.

For continuous variables,[7] the corresponding distribution may then be inferred from the density of the marks, assuming that enough observations are displayed: where should new observations be expected to land, according to the chart in Figure 2.4?

7: Where observed values are unlikely to be identical.

Bar Charts and Histograms

Bar charts and **histograms** are also univariate visualizations.

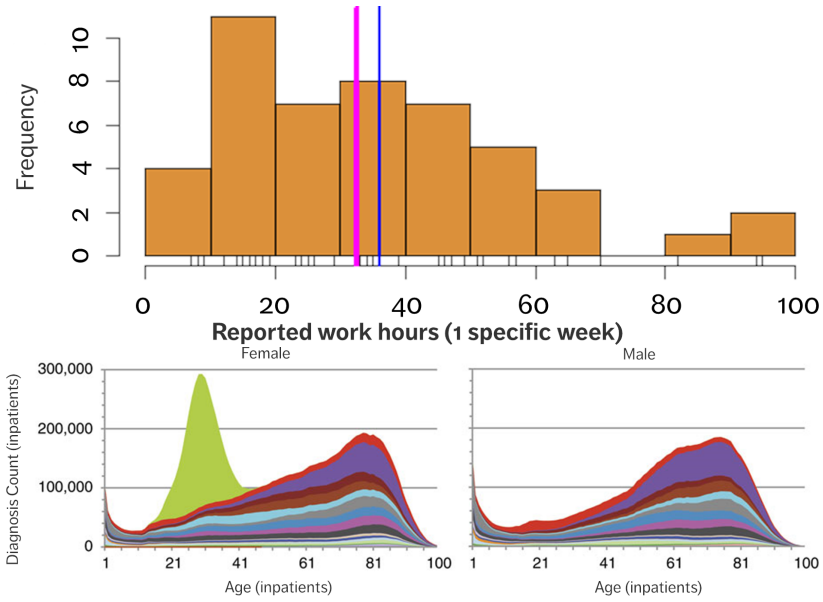


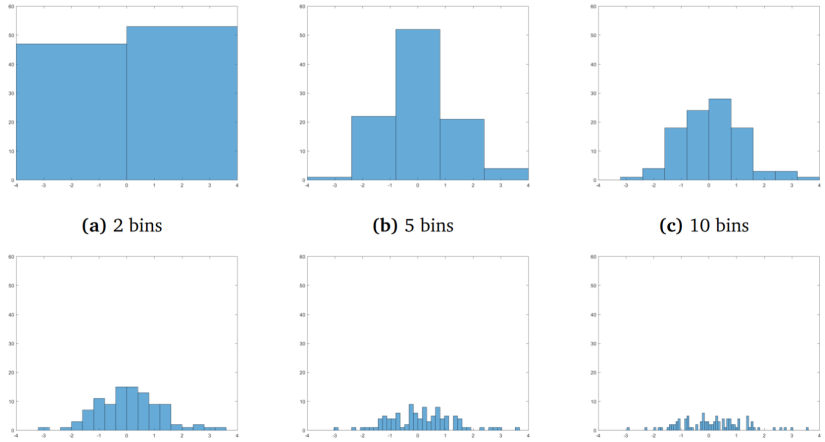
Figure 2.5: Histogram of reported work hours (personal file, top); stacked histograms of diagnosis count by age by medical conditions among Danish inpatients from 1996-2010 [32].

They are easy to read, in the sense that they look like something right out of a high-school class: we tally the number of observations that fall in a class and produce a bar whose height is **proportional** to that tally.[8]

They can be displayed vertically or horizontally (there is some evidence to suggest that horizontal charts are more effective, see Section 4.2, *Gestalt Principles*), and adorned with **added information** (rug, median, mean, on histograms for instance, etc.), or stacked on top of one another.[9]

Histograms make it easy to spot suspected **outliers** (such as the three employees who reported working more than 80 hours a week in Figure 2.5), **centrality**, **spread**, and **skewness**, they are useful for side-by-side comparisons at a glance, and they provide a detailed visual representation of a dataset variable in its entirety.

On the flip side, there is no agreed-upon rule to determine the number of bins (the classes) that should be used in a histogram (see 2.6).



(a) 2 bins **(b)** 5 bins **(c)** 10 bins

(d) 20 bins **(e)** 50 bins **(f)** 100 bins

Figure 2.6: Effect of binning on frequency count charts (height determines frequency counts, not area); the data consists of 100 normally distributed points, with $\mu = 0$ and $\sigma = 1$.

Line Charts

Line charts also have a classic, easy-to-read feel – they look like Joe and Jane Q. Public's **expectation** of a data visualization.

But the right look does not necessarily convey a **valid point**:[10] in Figure 2.7, for instance, the middle right chart shows a high but spurious correlation between two time series.[11] This harmless example highlights the danger of conflating correlation and causation.

Scatterplots

While we can always find ways to display multivariate variables using rug plots, histograms (see the algae blooms charts in the previous section), and line charts, they were ostensibly designed to be univariate charts. Not so **scatterplots**, who require at least 2 variables, as at the bottom of Figure 2.7.

8: If the classes are determined by the levels of a (non-ordinal) categorical variable, the corresponding chart is a **bar chart**; if they are determined by the levels of an ordered numerical variable (whether discrete or continuous), the corresponding chart is a **histogram**.

9: Somewhat depressingly, they are some of the simplest graphical representations, but a majority still does not know how to read them.

10: To be fair, that is a problem that is not only limited to line charts.

11: The yearly number of pool drownings per year in the U.S. and the yearly number of film in which actor Nicolas Cage appeared.

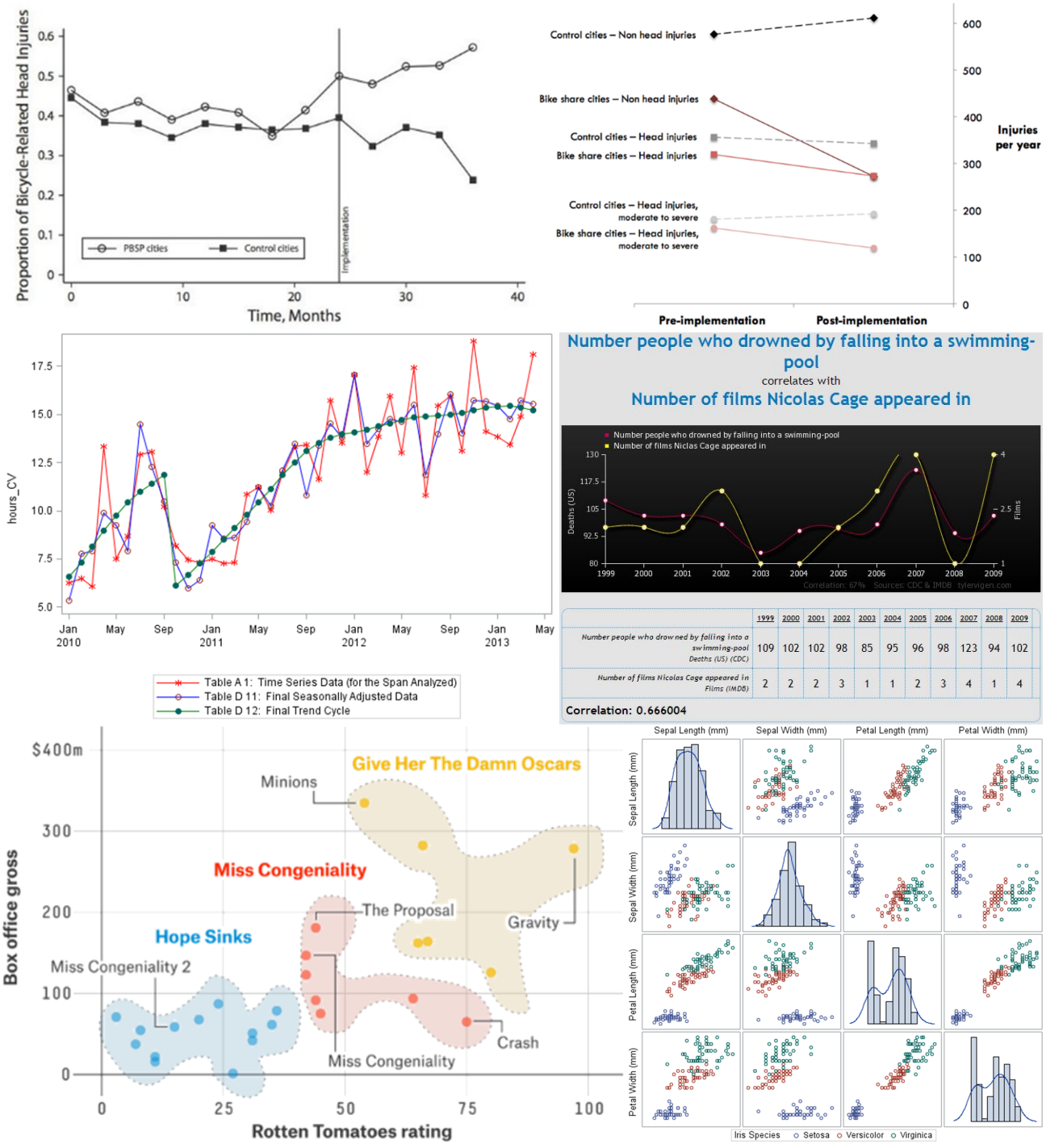


Figure 2.7: Line charts: proportion of all bicycle-related injuries that were classified as head injuries among cities with public bike share programs and control cities, centered on intervention date (vertical line), in North American cities (top left, [33]); data from study show declines in all injuries, including head injuries after bike share system implemented – because head injuries decline less than other injuries, they are now a larger proportion of all injuries (top right, Teschke); time series decomposition of a transit time indicator (middle left, personal file); number of people who drowned by falling into swimming pools in the United States from 1999 to 2009 against the number of films the actor Nicolas Cage appeared in on a yearly basis over the same time period (middle right, author unknown). Scatterplot clusters in Sandra Bullock movies (bottom left, FiveThirtyEight.com), and scatterplot matrix (with histograms on the diagonal) of Anderson's 1936 Iris dataset recording the measurements of 4 characteristics for a sample of 50 individual flowers from each of 3 iris species (bottom right, SAS).

On the positive side of the ledger, scatterplots make it easy to spot (suspected) outliers or clusters (along the two variables that are displayed), and they provide a **detailed** visual representation of two variables at a time; we can also display pairwise scatterplots of all dataset variables into a **scatterplot matrix**, in which the diagonal and upper triangular elements can be used to showcase other dataset features (histograms, correlations, etc.).

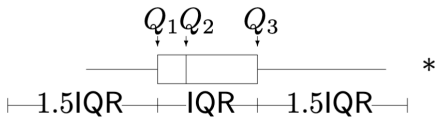
On the flip side, focusing on only two variables might **obscure** important relationships in the data; moving to a scatterplot matrix to sidestep *this* issue increases the complexity of the chart, which can quickly become unreadable with an increase in the number of variables, but even then, relationships involving 3 or more variables may remain invisible.

Boxplots

The **boxplot** is a quick and easy way to present a graphical summary of a univariate distribution.

We draw a box along the observation axis, with endpoints at the lower and upper **quartiles**, and with a "belt" at the **median**.[12] Then, we plot a line extending from Q_1 to the smallest value less than $1.5(Q_3 - Q_1)$ to the left of Q_1 and from Q_3 to the largest value less than $1.5(Q_3 - Q_1)$ to the right of Q_3. Any suspected outlier is plotted separately, as shown below:

12: See [1] for a definition of these concepts.



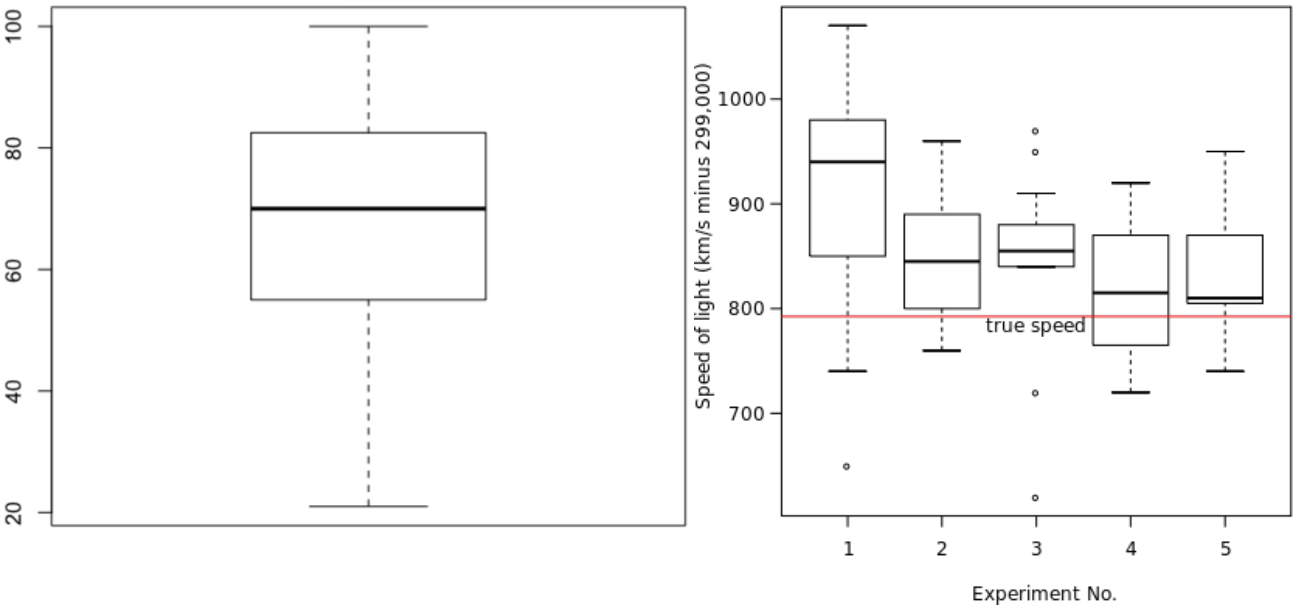


Figure 2.8: Boxplots of the speed of light computed experimentally in the famous Michelson-Morley experiments (Wikipedia). The true speed of light has been superimposed on the data; outlying values appear as dots (right); boxplot of the final grades in a probability and statistics course (left).

2.3 Representing Multivariate Observations

It is not unusual to see modern datasets with 5, 10, 100, or even 1000+ variables. **High-dimensionality** brings a host of problems (such as the **curse of dimensionality**, which is remedied by feature selection and dimension reduction [1, ch.23]); from a data visualization perspective, the challenge is that at most two attributes can be represented by position in the plane. How can we then represent other crucial elements on a flat computer screen or a piece of paper?

Potential solutions include using a **3-dimensional physical display** (such as one produced by 3D printing), or, more reasonably, one of the following **visual elements**:

- marker size
- marker colour
- colour intensity and value
- marker texture
- line orientation
- marker shape
- motion/movie

These elements do not always mix well – there can be "too much of a good thing", so to speak. In practice, human brains can reasonably be hoped to integrate 4 or 5 design elements in a chart (including 2 reserved for position), together with a motion component; the use of additional design elements tends to complicate matters and confuse the reader more than anything. Efficient design is as much art as it is science: while we do want to highlight **multivariate** relationships, we do need to keep things "parsable."[13]

13: Less is more, as long as "less" is enough.

In Figure 2.9, we provide a two-dimensional scatterplot display of a subset of the NASA CM1 dataset.

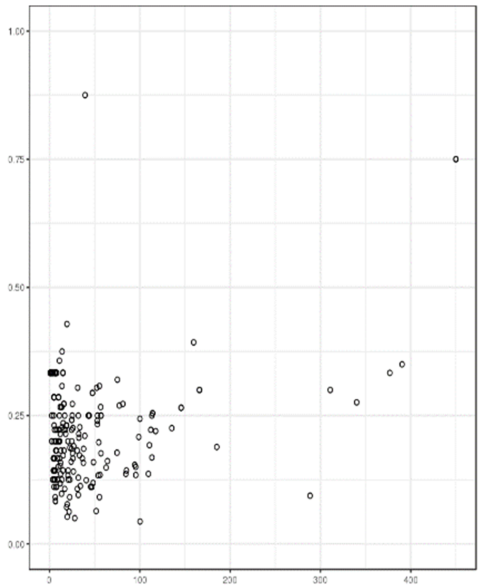


Figure 2.9: Scatterplot of a subset of the NASA CM1 dataset (minimalist).

We went out of our way to find a display that is as **uninformative** as it is dull, and we think that you will agree that we succeeded in our attempt.[14] The addition of data-aligned design elements does not only make the chart more appealing – it also allows for **insight discovery** (see Figure 2.10).

14: Who could blame anyone for skipping over this chart entirely on their way to more entertaining fare? Is there **anything** of even remote interest that can be said about the chart and the underlying data?

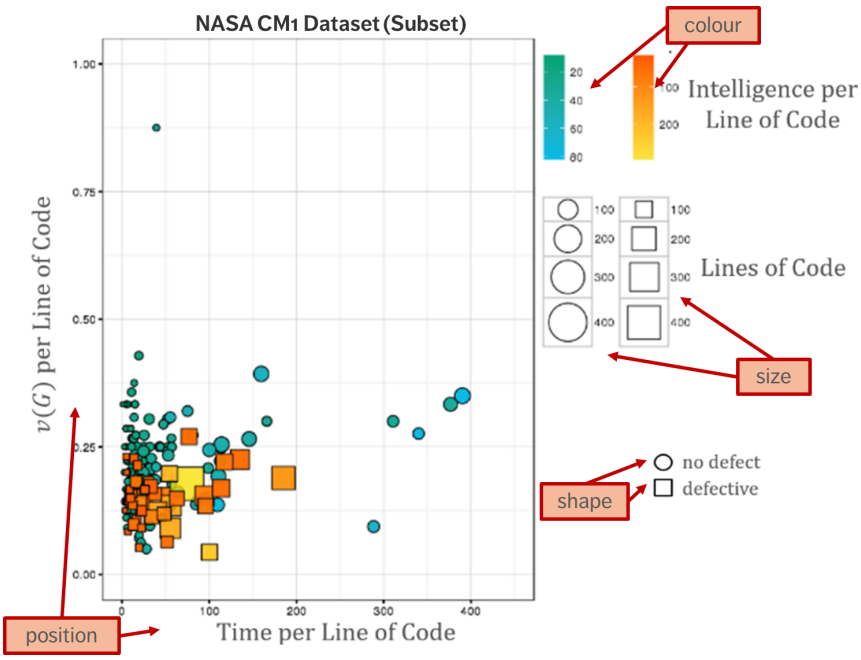


Figure 2.10: Bubble chart of a subset of the NASA CM1 dataset (with 5 variables).

For instance, we can see that defective components tend to contain more lines of code than non-defective components, *on average*, and that components for which more time was spent per line of code tend to be non-defective.[15]

15: True, this might not come as much of a surprise, but even such simple insights as these remain unreachable from Figure 2.9 alone.

Example: Bubble Chart Health and Wealth of Nations (see Figure 1.4).

- **Data:**
  - 2012 life expectancy in years
  - 2012 inflation adjusted GDP/capita in USD
  - 2012 population for 193 UN members and 5 other countries
- **Some Questions and Comparisons:**
  - Can we predict the life expectancy of a nation given its GDP/capita?[16]
  - Are there outlier countries?[17]
  - Are countries with a smaller population healthier?[18]
  - Is continental membership an indicator of health and wealth levels?[19]
  - How do countries compare against world values for life expectancy and GDP per capita?[20]
- **Multivariate Elements:** position for health and wealth, bubble size for population, colour for continental membership, and labels to identify the nations.

16: The trend is linear: Expectancy $\approx 6.8\times$ ln GDP/capita + 10.6

17: Botswana, South Africa, and Vietnam, among others, at a glance.

18: Bubble size seems uncorrelated with the axes' variates.

19: There seems to be a clear divide between Western Nations (and Japan), most of Asia, and Africa.

20: The vast majority of countries fall in three of the quadrants. There are very few wealthy countries with low life expectancy. China sits near the world values, which is expected for life expectancy, but more surprising when it comes to GDP/capita – compare with India.

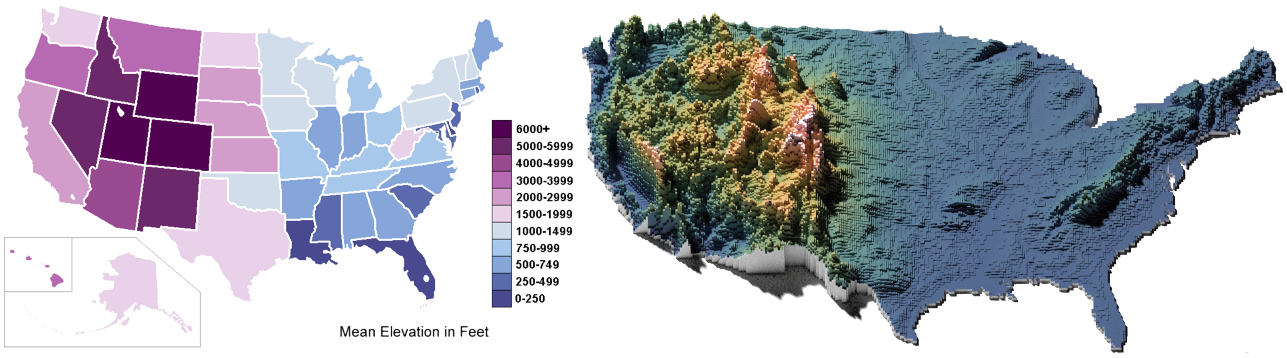


Figure 2.11: Mean elevation by U.S. state, in feet (left, source unknown); high resolution elevation map (right, by twitter user `@cstats`).

- **Comments:**
  - Are life expectancy and GDP/capita appropriate proxies for health and wealth?
  - A fifth element could also be added to a screen display: the passage of time. In this case, how do we deal with countries coming into existence (and ceasing to exist as political entities)?

Example: Choropleth Map Mean Elevation by U.S. State (see Figure 2.11).

- **Data:** 50 observations, ranging from sea level (0-250) to (6000+)
- **Some Questions and Comparisons:**
  - Can the mean elevation of the U.S. states tell us something about the global topography of the U.S.? [21]
  - Are there any states that do not "belong" in their local neighbourhood, elevation-wise? [22]
- **Multivariate Elements:** geographical location (position) and elevation (purple-blue colour gradient as the marker for mean elevation in the chart on the left; height and colour gradient in the chart on the right).
- **Comments:**
  - Is the 'mean' the right measurement to use for this map?[23]
  - Would there be ways to include other variables in this chart?[24] .
  - What is going on with the scale in the legend?[25]

21: Western states have higher mean elevation, probably due to the presence of the Rockies; Eastern coastal states are more likely to suffer from rising water levels, for instance.

22: West Virginia and Oklahoma seem to have the "wrong" shade – is that an artefact of the colour gradient and scale?

23: It depends on the author's purpose.

24: Population density with texture, for instance

25: No idea,,,

Example: Network Diagram Lexical Distances (see Figure 2.12).

- **Data:**
  - speakers and language groups for 43 European languages
  - lexical distances between languages
- **Some Questions and Comparisons:**
  - Are there languages that are lexically closer to languages in other lexical groups than to languages in their own groups?[26]
  - Which language has the most links to other languages?[27]

26: French is lexically closer to English than it is to Romanian, say.

27: English has 10 links.

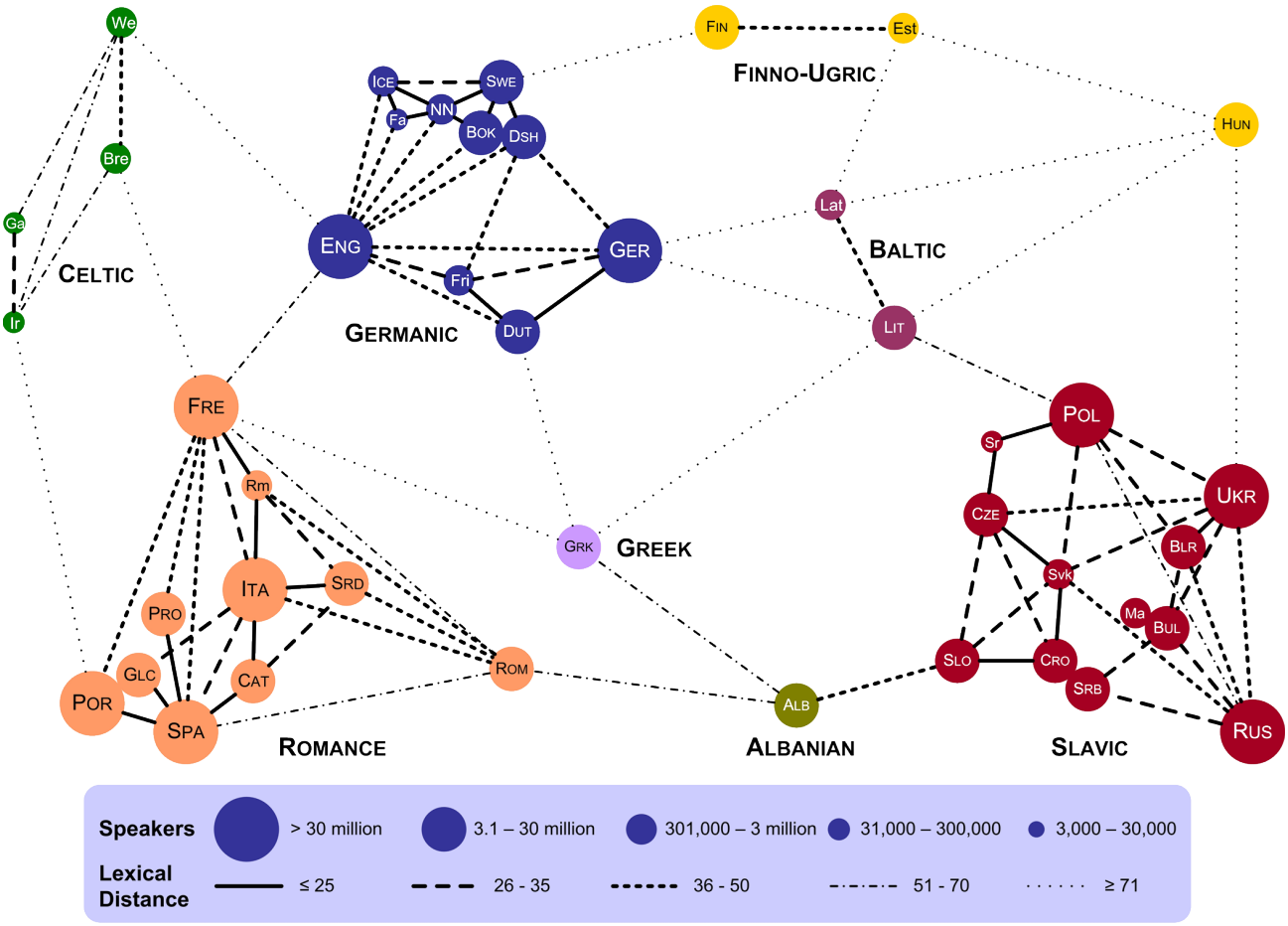


Figure 2.12: Lexical distances of European languages [34].

- Are there languages that are lexically close to multiple languages in other groups?[28] .
- Is there a correlation between the number of speakers and the number of languages in a language group? [29]
- Does the bubble size refer only to European speakers? [30]

- **Multivariate Elements:**
  - colour and cluster for language group
  - line style for lexical distance
  - bubble size for number of speakers

- **Comments:**
  - How is lexical distance computed? Is it actually a "distance" in the mathematical sense?
  - Some language pairs are not joined by links – does this mean that their lexical distance is too large to be rendered?
  - Are the actual geometrical distances meaningful? For instance, Estonian is closer to French in the chart than it is to Portuguese... is it also lexically closer?[31]

28: Greek is lexically close to 5 groups

29: Language groups with more speakers tend to have more languages.

30: Portuguese seems to have as many speakers as French? Worldwide, that may be the case, but in Europe that is definitely not so.

31: Full disclosure, we are not sure if "lexically" is a word... but you know what we mean.

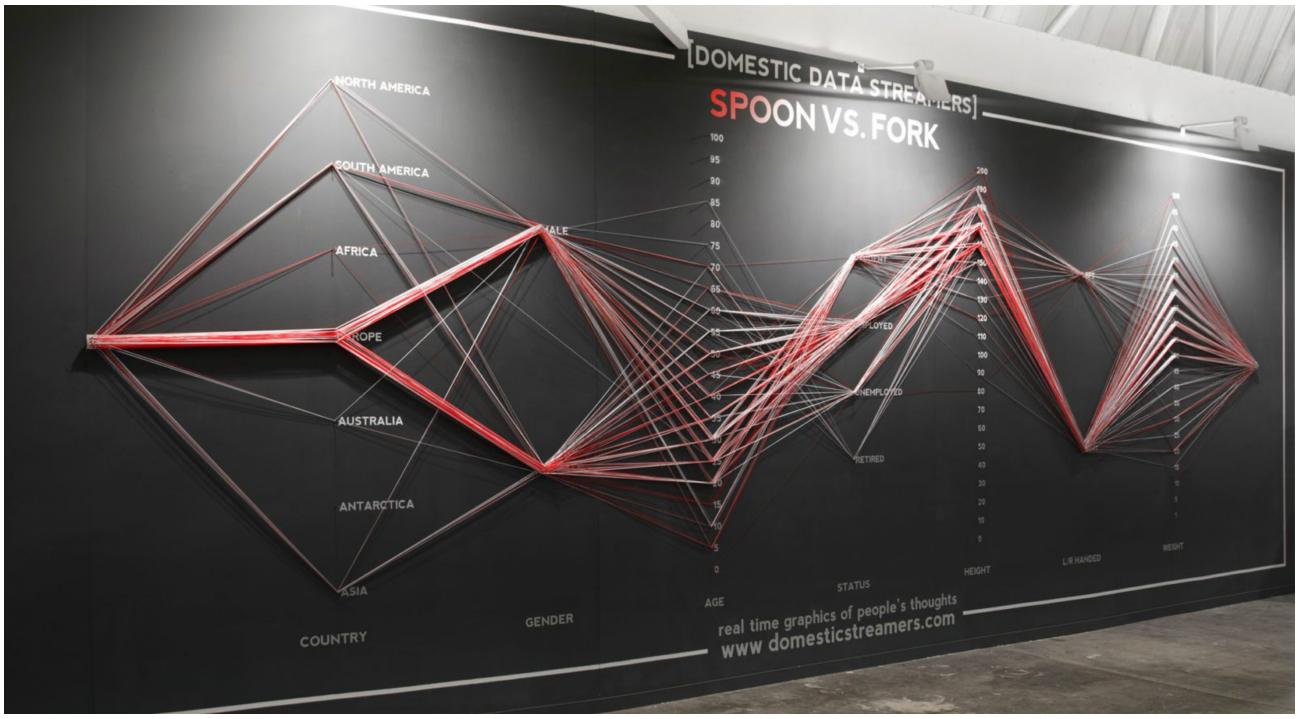


Figure 2.13: Physical visualization of people's demographics [Domestic Data Streamers].

Example: Parallel Coordinates Data Strings (see Figure 2.13).

- **Data:**
  - demographic information for a conference's participants
  - preference for using a spoon (red string) or a fork (white string) for breakfast
- **Some Questions and Comparisons:**
  - Where is the conference most likely to be taking place?[32]
  - Were there more men or women at the conference?[33]
  - Can we see a correlation between age and height, or height and weight?[34]
- **Multivariate Elements:**
  - seven demographic measurements (country (?), gender, age, status, height, left/right handed, weight)
  - colour to represent the preference (spoon vs. fork)
- **Comments:**
  - There does not seem to a link between the colour of the string and the measurements – it seems as though the spoon vs. fork question is a red herring (or at least a way to get attendees to participate in the hands-on data collection exercise)
  - Are the selected scales reasonable? Why include heights below 60cms?

32: Most strings go through Europe (conference taking place in Barcelona, in fact).

33: Seems about 50-50, but note the one string bypassing the gender axis completely – evidence of sub-optimal questionnaire design?

34: There is likely to be one, but the inclusion of "status" and "left/right handedness" between the pairs of interest makes it impossible to tell.

2.4 Communicating Analysis Results

Beyond data exploration, the crucial element of data presentations is that they need to help **convey the insight** (or the message); they should be clear, engaging, and (more importantly) readable. Our ability to think of questions (and to answer them) is in some sense limited by **what we can visualize**.[35]

35: There is always a risk that if certain types of visualization techniques dominate in evidence presentations, the kinds of questions that are particularly well-suited to providing data for these techniques will come to dominate the landscape, which will then affect data collection techniques, data availability, future interest, and so forth.

Selecting a Chart Type

The choice of visualization methods is strongly dependent on the analysis objective, that is, on the **questions that need to be answered**. Presentation methods should not be selected randomly (or simply from a list of easily-produced templates).

In Figure 2.14, F. Ruys suggests various types of visual displays that can be used, depending on the objective(s):

- who is involved?
- where is the situation taking place?
- when is it happening?
- what is it about?
- how/why does it work?
- how much?

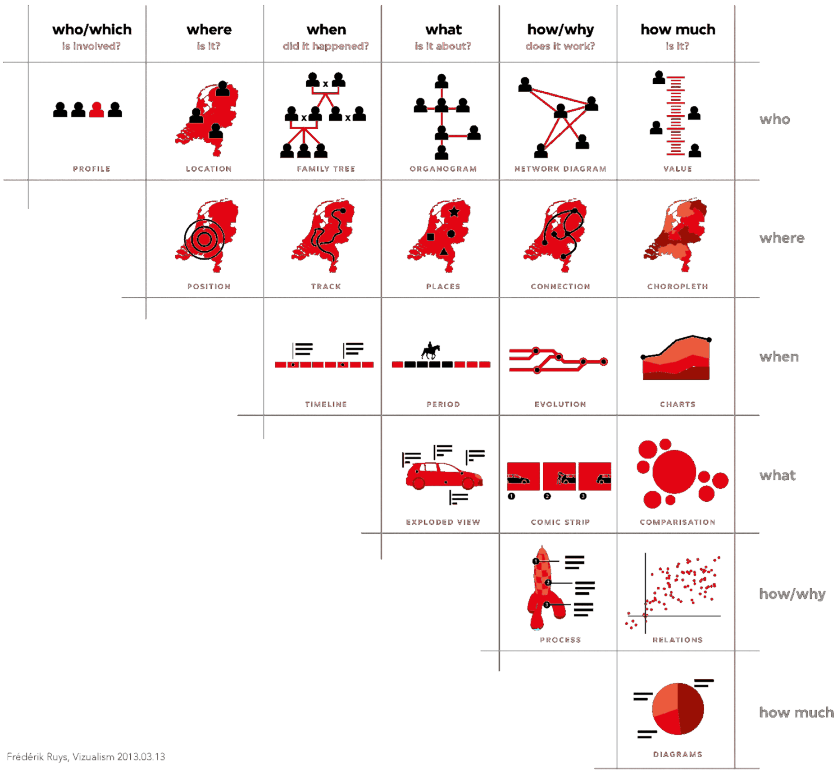


Figure 2.14: Data visualization suggestions, by question type (F. Ruys, Vizualism.nl).

A communication dashboard should at the very least be able to produce the following types of display:

- **univariate diagrams** (word clouds, box plots, histograms, bar charts, etc.)
- **multivariate charts** – comparison and relation (scatterplots, bubble charts, parallel coordinate charts, decision trees, cluster plots, trend plots, etc.)
- **choropleth maps** (heat maps, classification maps, etc.)
- **network diagrams** and connection maps (association rule networks, phrase nets, etc.)
- and a few others.

We will discuss this further in Chapter 9 (*Visualization Toolkit*).

Basic Rules of Data Communication

FlowingData suggests some **basic rules** to help make sure that the audience gets the most of **data communication** attempts (we will revisit this topic in Chapter 8, *Effective Storytelling Visuals*).

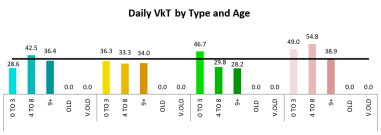
1. **Check the data:** are there outliers, spikes, anomalies, unexpected things?;
2. **Explain encoding:** don't assume the reader knows what everything on the chart means;[36]
3. **Label axes:** knowing the scale is important;
4. **Include units:** eliminate the need for guesswork;
5. **Keep your geometry in check:** circles and 2D shape are sized by **area**, bars by **length**;[37]
6. **Include your sources:** protect yourself, and let those who want to dig deeper do so;
7. **Consider your audience:** a poster can be wordy, a presentation should be minimalist.

The basic rules are simply **guidelines**: they can be bent if doing so helps get the point across. This requires knowing the audience and a host of other things that are not directly related to data and charts (see Chapter 7, *Stories and Storytelling*).

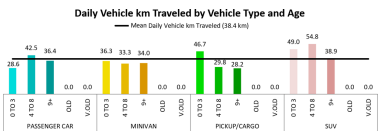
Integrating data and words usually helps to **convey the message** and adding design elements can enhance our understanding of the data., but J. Bertin suggests, in *Semiology of Graphics* [35], that not all **retinal variables** are equally effective when it comes to convey or represent information.

Data displays are not just about picking a random visualization method; we may need to **experiment** to find the optimal choice for the given context.[38] The communication outcomes will vary depending on the structure of the data and the (combinations of) questions; engaging the services of a **Tsarina of Common Sense** may help streamline the process (more on this later).

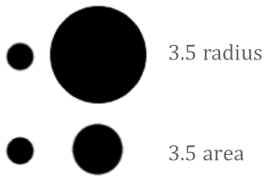
36: There is very little chance that non-specialist audiences will know what the acronyms and colours represent in the image below:



Much better to be specific:



37: Sizing by radius (top) and area (bottom):



38: Although there is value in creating random charts **as well** to ensure that we do not get stuck in a rut, such as it is.