Week 5

Applying a critical eye to statistical reporting

Dr Andrew Mitchell

a.j.mitchell@ucl.ac.uk

Lecturer in AI and Machine Learning for Sustainable Construction

Misleading Statistics

Misleading Statistics

We'll use a case study of data presented by climate change sceptics to illustrate how even real data and "mathematically correct" statistical analysis can be used to mislead.

Some types of misleading statistics:

- cherry-picking data
- overgeneralization
- faulty causality
- biased sampling
- misleading graphs
- reporting non-statistically-significant results as significant
- reporting statistically-significant but not practically-significant results as meaningful

See *Calling Bullshit - The Art of of Skepticism in a Data-Driven World* [1] for more examples.

“Hiatus in Global Warming"

In the 2010’s a claim began to circulate that data showed that “global mean surface temperature T_S has not risen since 1998, and may have fallen since late 2001” [2].

Along with some controversies about the source of climate data, this became known as ‘ClimateGate’. Similar claims arose again in 2022, showing an apparent pause in climate change from 2015-2022.

These claims *were* based on data - they presented analyses and visualisations of global temperature data which in fact did appear to show a pause or a decrease in global temperature, apparently disproving anthroprogenic climate change.

Let’s take a look back at this data with a critical eye and see whether we find them convincing.

Climate Change measurements

We won't be getting into the science of climate change here, but it's good to understand the basic arguments and sources of evidence.

> To get a complete picture of Earth's temperature, scientists combine measurements from the air above land and the ocean surface collected by ships, buoys and sometimes satellites, too.

> The temperature at each land and ocean station is compared daily to what is 'normal' for that location and time, typically the long-term average over a 30-year period. The differences are called an 'anomalies' and they help scientists evaluate how temperature is changing over time.

> A 'positive' anomaly means the temperature is warmer than the long-term average, a 'negative' anomaly means it's cooler.

Climate Change measurements

[3]

Climate Change measurements

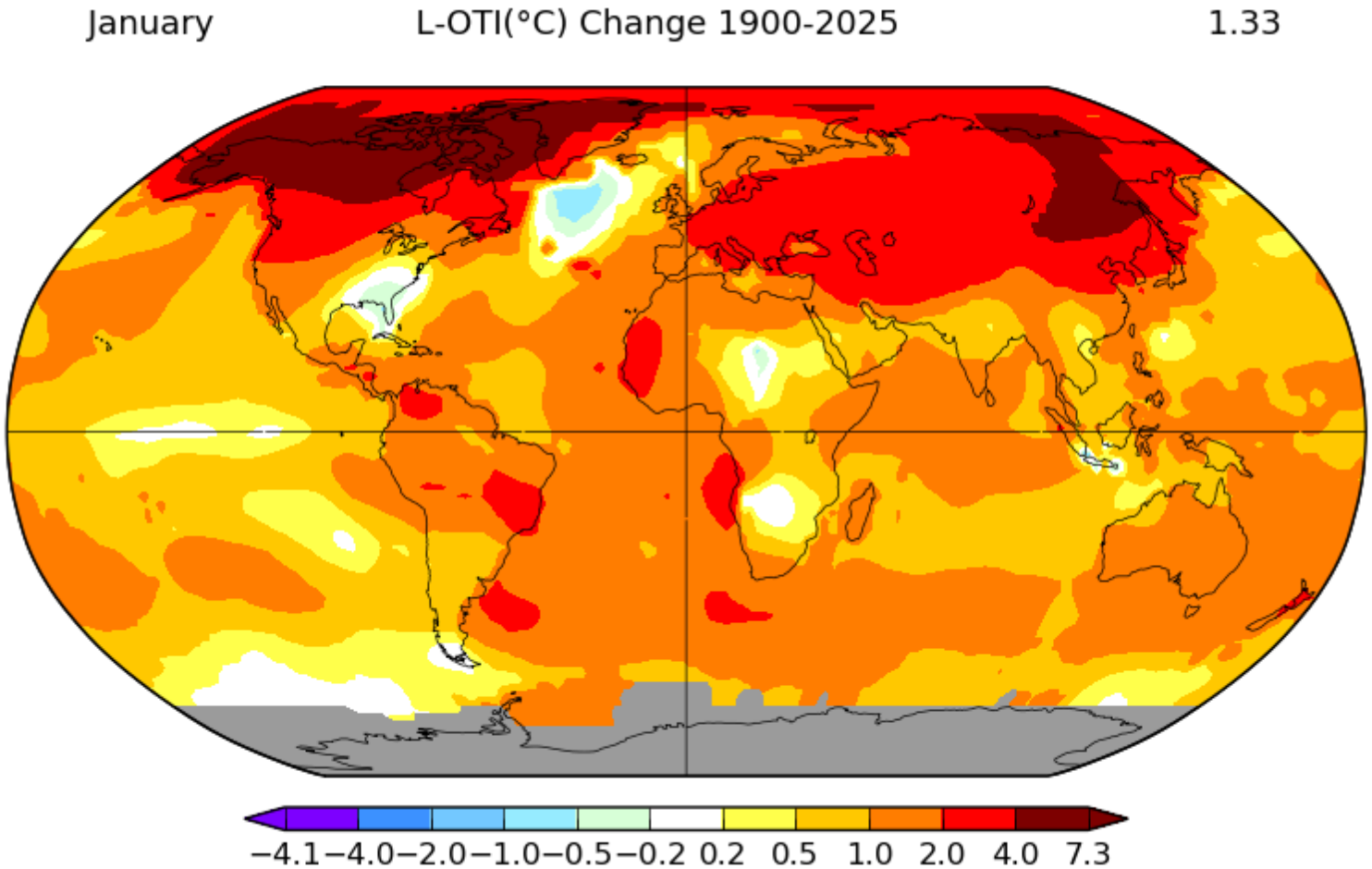


Figure 1: GISS Surface Temperature Analysis. Source: NASA [4]

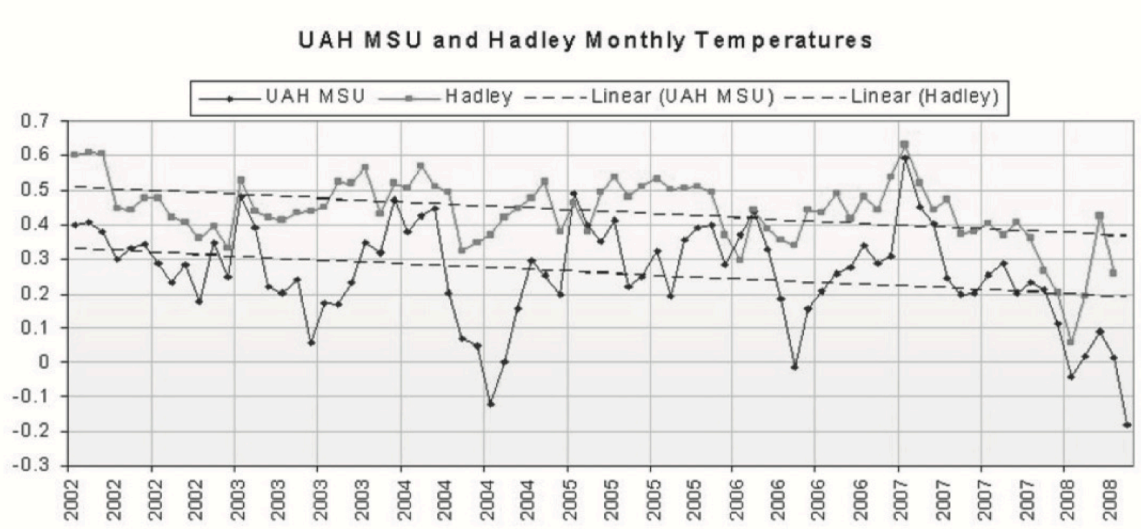
The First Global Warming Pause

- In C. Monckton [2], published in a peer-reviewed journal of The American Physical Society, Christopher Monckton claimed that the mean global temperature data across the four major data sources showed that T_S has not risen between 2001 and 2008.
- He presented a *time series* plot which confirms this.

the conclusion is that, perhaps, **there is no "climate crisis"**, and that currently-fashionable efforts by governments to reduce anthropogenic CO_2 emissions are pointless, may be ill-conceived, and could even be harmful.

The First Global Warming Pause

Figure 1: Mean global surface temperature anomalies (°C), 2001-2008



Since the phase-transition in mean global surface temperature late in 2001, a pronounced downtrend has set in. In the cold winter of 2007/8, record sea-ice extents were observed at both Poles. The January-to-January fall in temperature from 2007-2008 was the greatest since global records began in 1880. Data sources: Hadley Center monthly combined land and sea surface temperature anomalies; University of Alabama at Huntsville Microwave Sounding Unit monthly lower-troposphere anomalies; Linear regressions.

Figure 2: Mean global surface temperature anomalies, 2001-2008. Source: C. Monckton [2]

The New Global Warming Pause

Nearly a decade later, talk of a pause has re-emerged with claims in the media such as:

> contrary to the dogma which holds that a rise in carbon dioxide inescapably heats up the atmosphere **global temperature has embarassingly flatlined for more than seven years even as CO_2 levels have risen**. [5]

Again, the claim comes from a blog post written by Christopher Monckton titled *The New Pause Lengthens to 7 years 10 months* [6]. Let's look in depth at the data used to make this claim.

The New Global Warming Pause

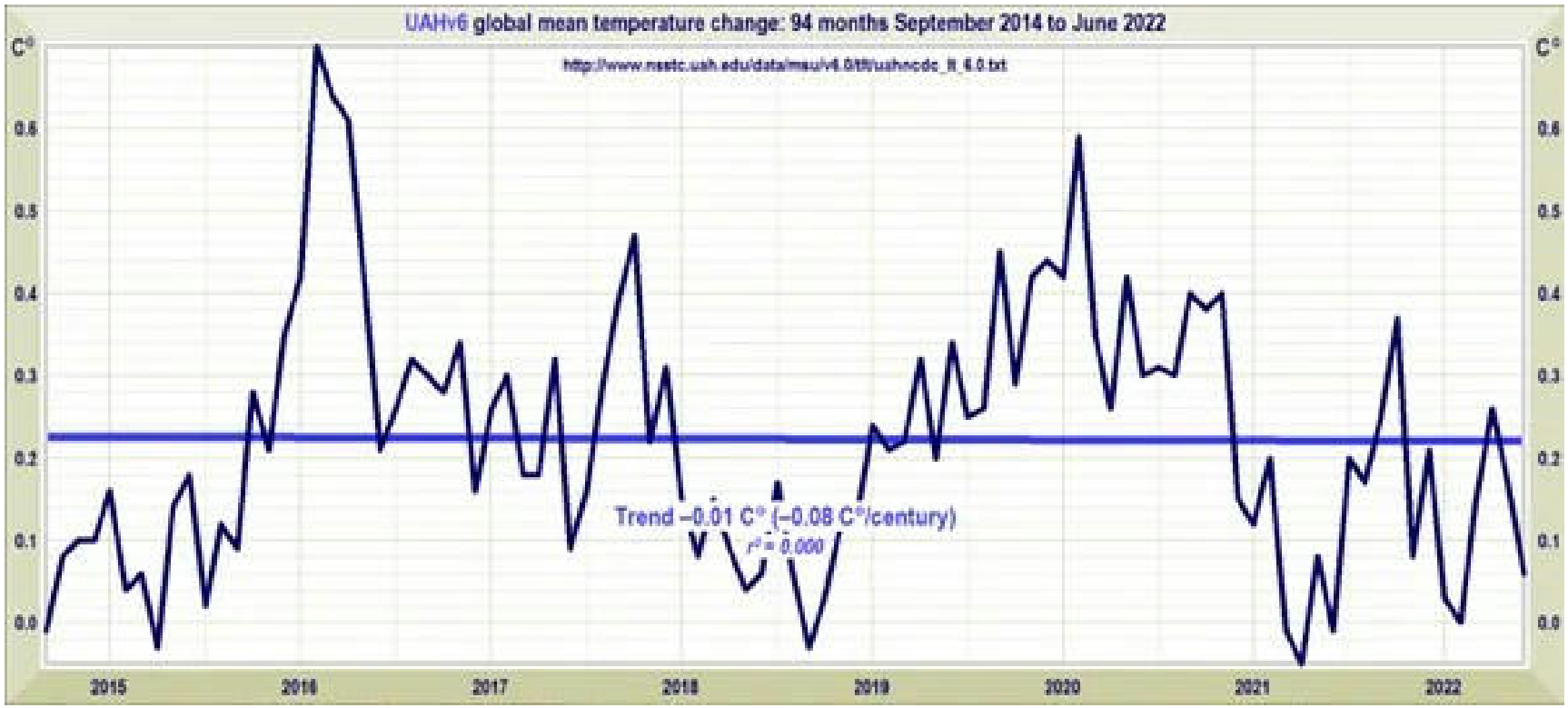


Figure 3: "This Pause [...] is, as always, not cherry-picked. It is derived from the UAH monthly global mean lower-troposphere temperature anomalies as the period from the earliest month starting with which the least-squares linear-regression trend to the most recent month for which data are available does not exceed zero." [6]

The New Global Warming Pause

So, what is wrong with this presentation? Why might it be misleading?

The New Global Warming Pause

Figure 4: Annual global surface temperature data from ERA5, along with Carbon Brief's estimate of annual 2022 temperatures based on the first six months of the year and the linear trend over the 2015 to 2022 period. Warming since pre-industrial is calculated using the Berkeley Earth dataset for the period prior to 1979. [7]

Cherrypicking Data

Looking at these eight years in isolation ignores the larger context.

A slightly different eight-year period - 2011 to 2018 rather than 2015 to 2022 - would offer the opposite conclusion, namely that **global warming had massively accelerated to a rate of 5.6C per century**.

Same as the prior plot, but showing annual global surface temperature data from 2000 and the trend over the 8-year period from 2011 through to 2018. [7]

Cherrypicking Data

In reality, both of these are acts of “cherry-picking” - overemphasising short-term **variability.**

Also note that Monckton picks his time periods carefully - the first ‘pause’ is from 2001 to 2008. Next, he shows the data from 2015 to 2022 - **so what happened from 2008 to 2015**? That is left out.

Finding spurious patterns within natural variance

So the questions we should ask, from a statistics perspective are:

- How large is the expected variability over any given period?
- Does the apparent downward trend in the period 2015-2022 fit within this variability, meaning we might just be looking at what is effectively noise?
- Or is the trend large enough to be seen without this random variability?

Figure 5: Same as the prior plots, but highlighting the years from 2015 onward compared to the 1979-2022 trend. [7]

> The fluctuations in recent years are well within the range of expected variability, and do not indicate any departure from the long-term warming trend in surface temperatures the world has experienced over the past 50 years.

Finding spurious patterns within natural variance

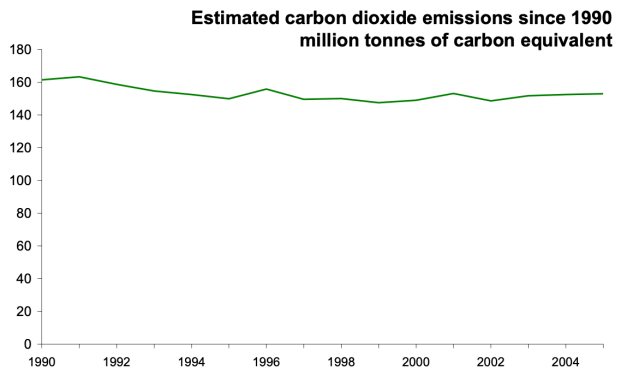
> The acceleration started from below the trendline and brought temperatures well above it, while the pause started above the trendline and brought temperatures back down to around what would be expected for 2021 and 2022.
>
> [7]

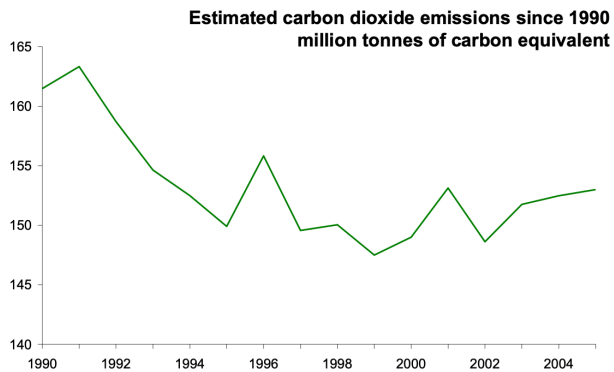
Zooming out further makes the *trend* very clear.

The 'pause' periods fit well within the natural variability. By intentionally focusing in on the periods which decrease effectively by chance over a short period of time, we can make the data appear to show a trend which is not there.

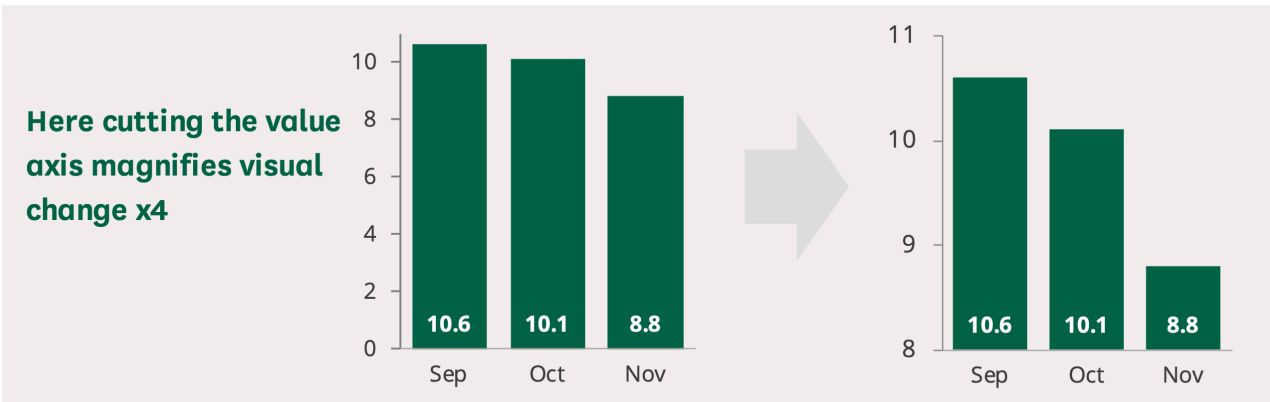
Same as the prior plots, but including Berkeley Earth data from 1850 through 2021. [7]

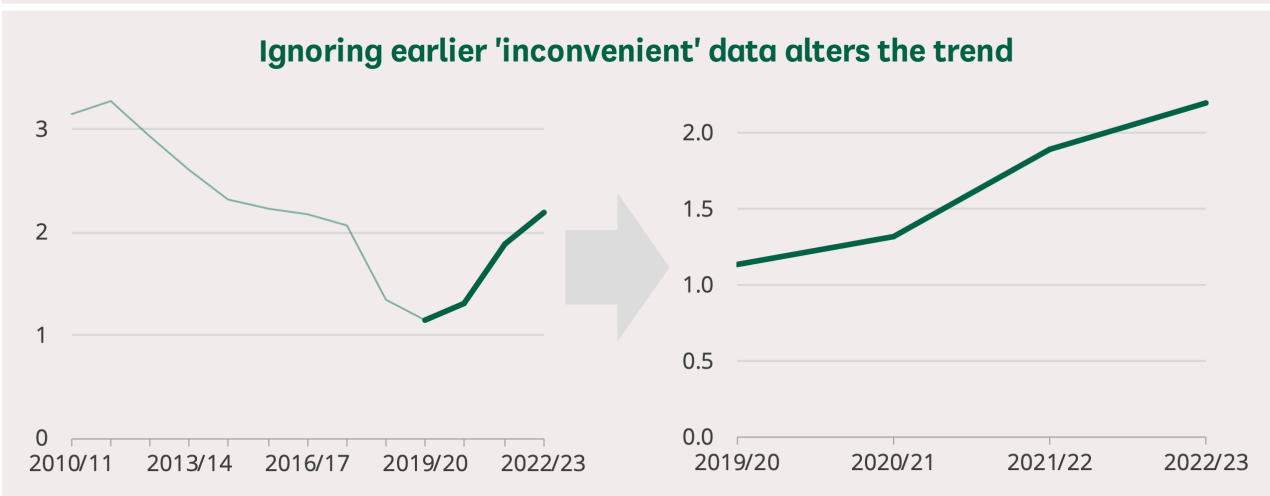
Other types of misleading or inappropriate visualisations





Other types of misleading or inappropriate visualisations





Other types of misleading or inappropriate visualisations

Figures from [8]

Concluding Thoughts

When presented with a statistical analysis or visualisation, what questions should we ask?

How can we make sure we're thinking about the data critically?

Further Reading

[8]. *How to spot spin and inappropriate use of statistics* (Research Briefing No. 4446). UK House of Commons Library.

[9]. *Statistical literacy guide: How to read charts* (Research Briefing No. SN04445). UK House of Commons Library.

References

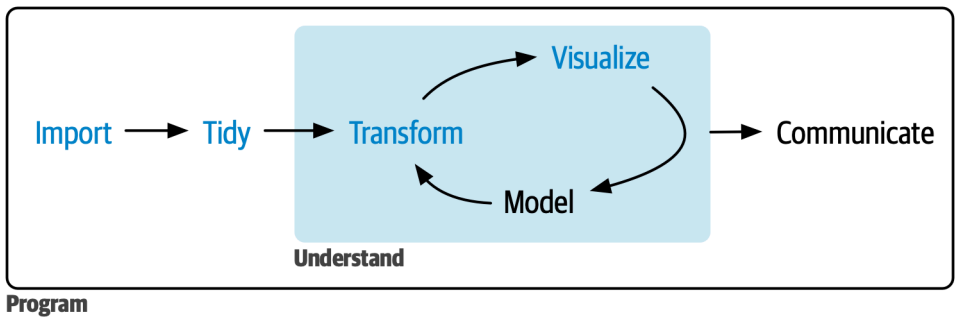
The Data Analysis Workflow

The process of statistical modeling

There is a set of steps that we generally go through when we want to use our statistical model to test a scientific hypothesis:

1. Specify your question of interest
2. Identify or collect the appropriate data
3. Prepare the data for analysis
4. Determine the appropriate model
5. Fit the model to the data
6. Criticize the model to make sure it fits properly
7. Test hypothesis and quantify effect size
8. Communicate your analysis

Data Analysis Workflow



Import

Throughout, we have been using the `tidyverse` library of packages for data analysis.

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

```
library(tidyverse)
```

Import


dplyr
tibble
tidyr
ggplot2
readr
stringr
forcats
lubridate
purrr

Import

- There are tools for reading data from almost any source:
  → `read_csv()`, `read_excel()`, `read_rds()`, ...
- When we load a dataset with a `tidyverse()` function, it will return a `tibble`

```
data <- read_csv("data/Apple_Emissions/greenhouse_gas_emissions.csv")
```

Tidy

The same data can be represented in multiple ways. Here's the same data organized three different ways:

Each dataset shows the same values of four variables: country, year, population, and number of documented cases of TB (tuberculosis), but each dataset organizes the values in a different way.

```
table1
```

```
# A tibble: 6 × 4
  country      year  cases population
  <chr>       <dbl>  <dbl>      <dbl>
1 Afghanistan  1999    745   19987071
2 Afghanistan  2000   2666   20595360
3 Brazil       1999  37737  172006362
```

Tidy

```
4 Brazil       2000  80488  174504898
5 China        1999 212258 1272915272
6 China        2000 213766 1280428583
```

```
table3
```

```
# A tibble: 6 × 3
  country      year rate
  <chr>       <dbl> <chr>
1 Afghanistan  1999 745/19987071
2 Afghanistan  2000 2666/20595360
3 Brazil       1999 37737/172006362
4 Brazil       2000 80488/174504898
5 China        1999 212258/1272915272
6 China        2000 213766/1280428583
```

Tidy

```
table2
```

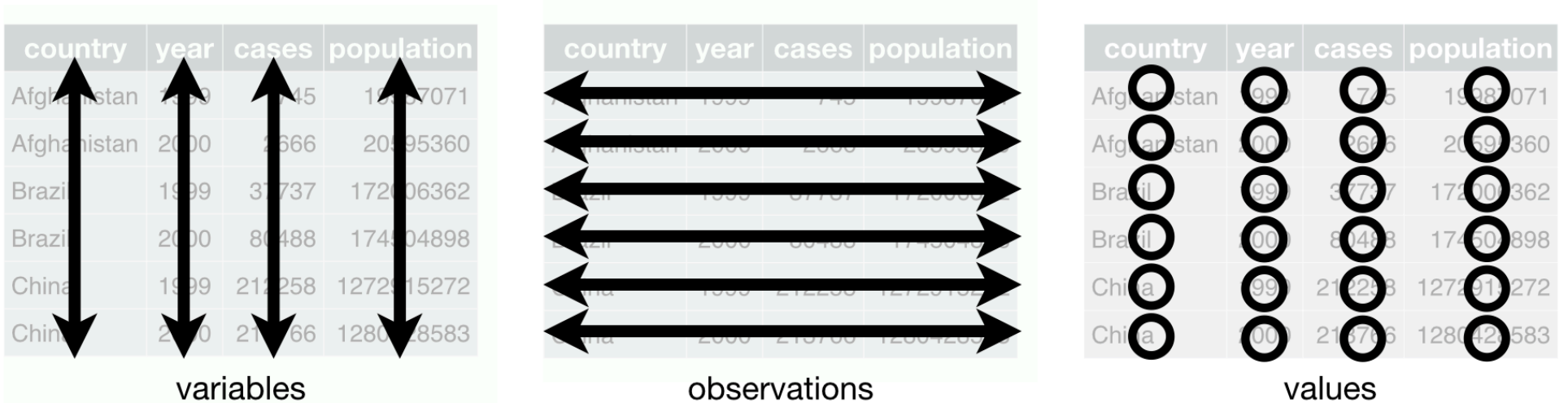
```
# A tibble: 12 × 4
   country      year type           count
   <chr>       <dbl> <chr>          <dbl>
 1 Afghanistan  1999 cases            745
 2 Afghanistan  1999 population   19987071
 3 Afghanistan  2000 cases           2666
 4 Afghanistan  2000 population   20595360
 5 Brazil       1999 cases           37737
 6 Brazil       1999 population  172006362
 7 Brazil       2000 cases           80488
 8 Brazil       2000 population  174504898
 9 China        1999 cases          212258
10 China        1999 population 1272915272
```

Tidy

```
11 China        2000 cases          213766
12 China        2000 population 1280428583
```

- There are three rules that make a dataset tidy:
  1. Each variable is a column; each column is a variable.
  2. Each observation is a row; each row is an observation.
  3. Each value is a cell; each cell is a single value.

Tidy



Why ensure your data is tidy?

Tidy

1. There's a general advantage to picking one consistent way of storing data. If you have a consistent data structure, it's easier to learn the tools that work with it because they have an underlying uniformity.
2. There's a specific advantage to placing variables in columns because it allows R's vectorized nature to shine. That makes transforming tidy data feel particularly natural.

So, our first task after importing the data is to make sure it's tidy. In addition to the rules above, this can also include things like:

- ensure the data types are correct
- clean up the column names
- make sure we know what the variables represent

For the .csv data we loaded, our column names can be a bit difficult to work with since they have spaces in them. We can use a function from the `janitor` package to clean these:

Tidy

```
data <- data |>
  janitor::clean_names()
data
```

```
# A tibble: 127 × 6
   fiscal_year category            type            scope   description
emissions
         <dbl> <chr>               <chr>           <chr>   <chr>
<dbl>
 1        2022 Corporate emissions Gross emissions Scope 1 Natural ga…
39700
 2        2022 Corporate emissions Gross emissions Scope 1 Fleet vehi…
12600
 3        2022 Corporate emissions Gross emissions Scope 1 Other (R&D…
2900
```

Tidy

```
 4        2022 Corporate emissions Gross emissions Scope … Electricity
0
 5        2022 Corporate emissions Gross emissions Scope … Steam, hea…
3000
 6        2022 Corporate emissions Gross emissions Scope 3 Business t…
113500
 7        2022 Corporate emissions Gross emissions Scope 3 Employee c…
134200
 8        2022 Corporate emissions Gross emissions Scope 3 Upstream f…
10600
 9        2022 Corporate emissions Gross emissions Scope 3 Work from …
7500
10        2022 Corporate emissions Gross emissions Scope 3 Transmissi…
0
# ℹ 117 more rows
```

Transform

We've dealt with data transformations quite a bit already. This includes operations like calculating the mean for different groups, or for multiple groups:

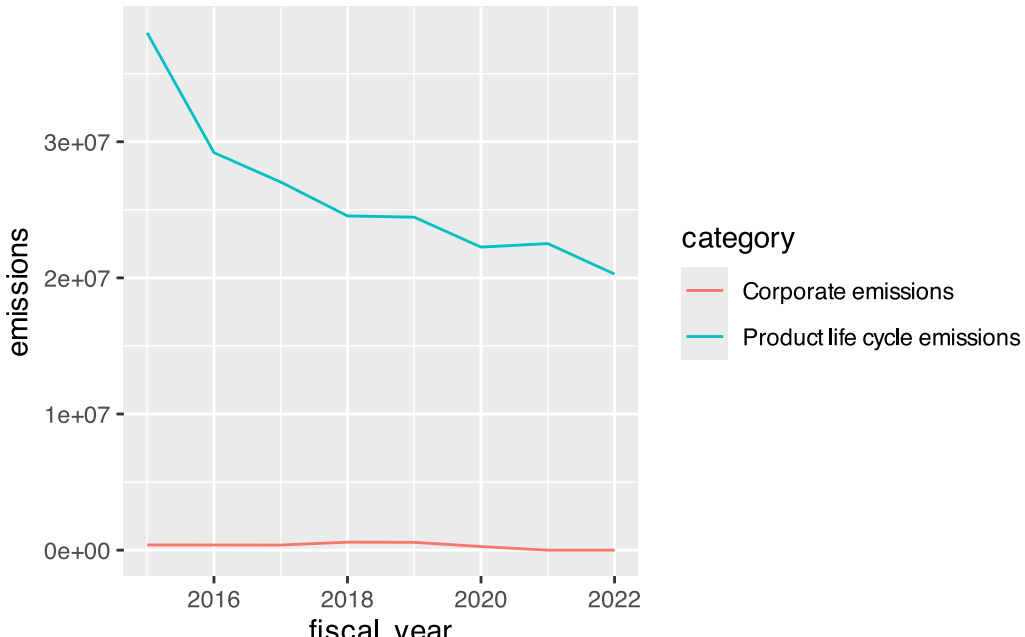
```
data |>
  group_by(category) |>
  summarise(
    mean_emissions = mean(emissions, na.rm = TRUE),
  )
```

```
# A tibble: 2 × 2
  category                      mean_emissions
  <chr>                                  <dbl>
1 Corporate emissions                   35594.
2 Product life cycle emissions        5630000
```

Visualize

```
data |>
  group_by(category, fiscal_year) |>
  summarise(emissions = sum(emissions, na.rm = TRUE)) |>
  ggplot(aes(x = fiscal_year, y = emissions, color = category)) +
  geom_line()
```

Visualize


emissions
3e+07
2e+07
1e+07
0e+00
2016
2018
2020
2022
fiscal_year
category
Corporate emissions
Product life cycle emissions

Communicate

This is where we will dive into using Quarto. Start by downloading the Apple Emissions dataset from Moodle and open RStudio.

We'll go through how to create and write a full analysis in a `.qmd` file using this dataset.

Refer to our lecture notes specifically on using Quarto

Assignment - Deceptive Visualisation

References

[1] C. T. Bergstrom and J. D. West, *Calling Bullshit: The Art of Scepticism in a Data-Driven World*, First published by Allen Lane. in An Allen Lane Book. London, UK USA Canada Ireland Australia: Penguin Books, 2021.

[2] C. Monckton, "Climate Sensitivity Reconsidered," *Physics and Society*, vol. 37, no. 3, pp. 6–19, 2008, [Online]. Available: https://higherlogicdownload.s3.amazonaws.com/APS/a05ec1cf-2e34-4fb3-816e-ea1497930d75/UploadedImages/Newsletter_PDF/july08.pdf

[3] R. Pidcock, "Explainer: How Do Scientists Measure Global Temperature?." Accessed: Feb. 12, 2025. [Online]. Available: https://www.carbonbrief.org/explainer-how-do-scientists-measure-global-temperature/

[4] NASA, "GISS Surface Temperature Analysis (v4): Global Maps." Accessed: Feb. 12, 2025. [Online]. Available: https://data.giss.nasa.gov/gistemp/maps/

References

[5] M. Phillips, "Sri Lanka Shows the Danger of Green Dogma," *The Times*, Jul. 2022, Accessed: Feb. 12, 2025. [Online]. Available: https://www.thetimes.com/article/sri-lanka-shows-the-danger-of-green-dogma-sf69m752q

[6] C. Monckton, "The New Pause Lengthens to 7 Years 10 Months." Accessed: Feb. 12, 2025. [Online]. Available: https://wattsupwiththat.com/2022/07/02/the-new-pause-lengthens-to-7-years-10-months/

[7] Z. Hausfather, "Factcheck: No, Global Warming Has Not `Paused' over the Past Eight Years." Accessed: Feb. 12, 2025. [Online]. Available: https://www.carbonbrief.org/factcheck-no-global-warming-has-not-paused-over-the-past-eight-years/

[8] P. Bolton, "How to Spot Spin and Inappropriate Use of Statistics," Jun. 2023. Accessed: Feb. 12, 2025. [Online]. Available: https://researchbriefings.files.parliament.uk/documents/SN04446/SN04446.pdf

References

[9] P. Bolton, “Statistical Literacy Guide: How to Read Charts,” Sep. 2007. Accessed: Feb. 12, 2025. [Online]. Available: https://researchbriefings.files.parliament.uk/documents/SN04445/SN04445.pdf