Patterns in Chaos

How Data Visualisation Helps To See the Invisible

yote

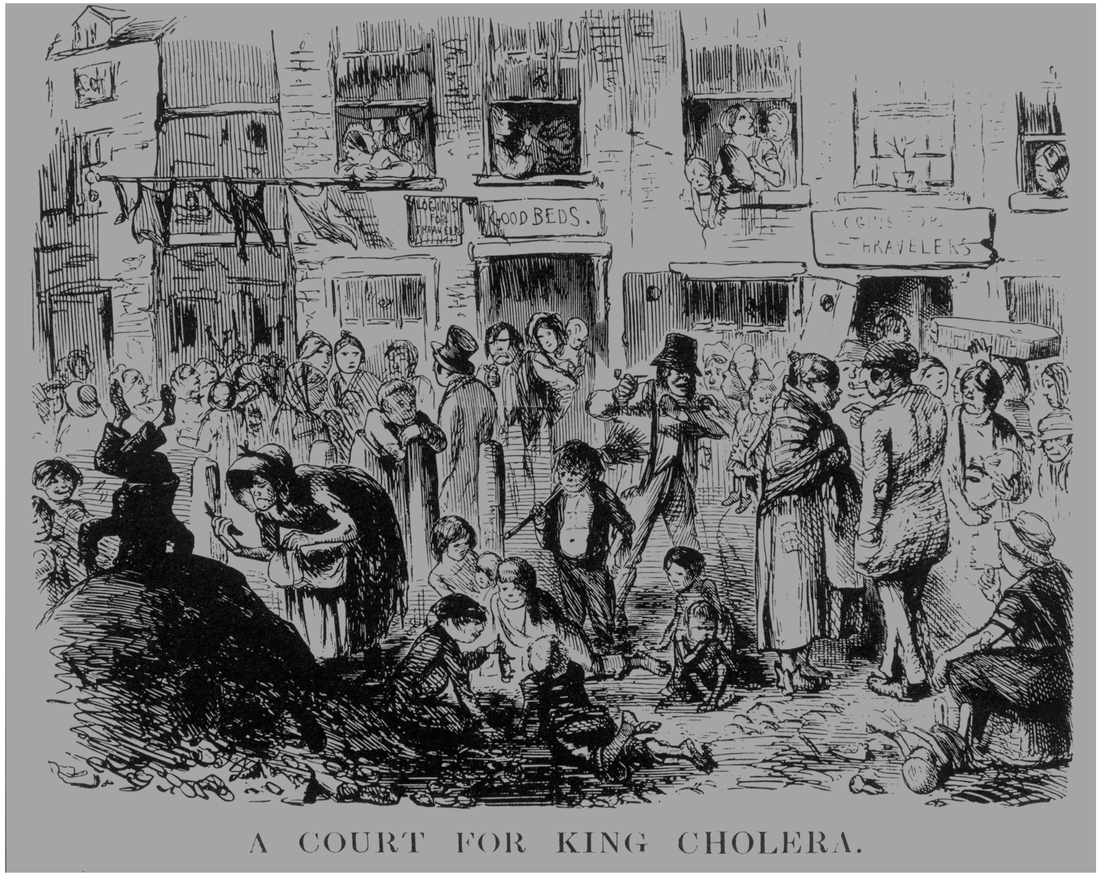
whoami: yote

- Biologist turned biomathematician:
  - PhD in mathematical biology (cancer research)
    - Mostly pencil-and-paper work, but also analysis of larger datasets.
  - Now in industrial research, de-facto doing data science.
    - Effective communication of results to colleagues and nontechnical folks.

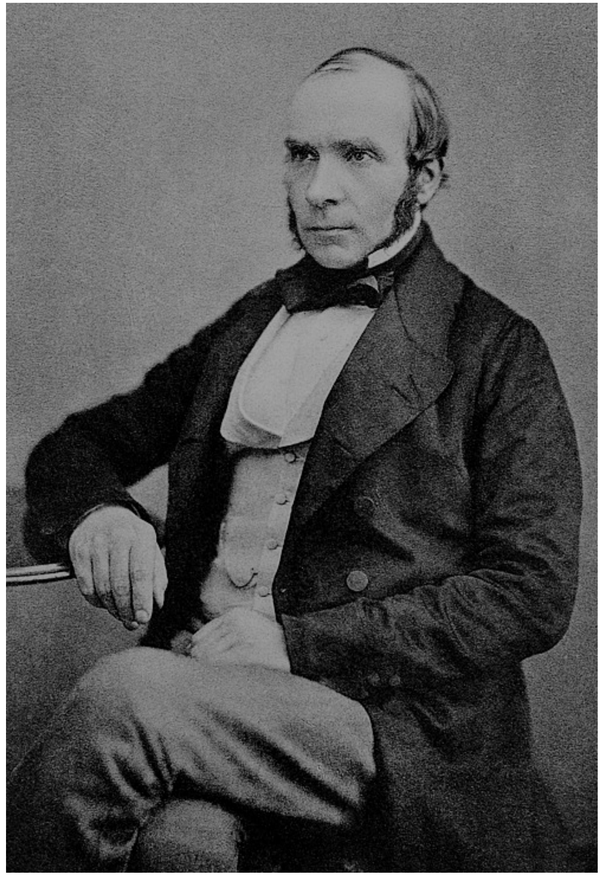
1. Motivating Examples

The 1854 cholera outbreak in Soho, London.

616 people died, but what was to blame?



(https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Punch-A_Court_for_King_Cholera.png)



(https://en.wikipedia.org/wiki/John_Snow#/media/File:John_Snow.jpg)

The 1854 cholera outbreak in Soho, London.

616 people died, but what was to blame?

(https://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg)

(https://www.merrittcartographic.co.uk/cholera.html)

The 1854 cholera outbreak in Soho, London.

616 people died, but what was to blame?

(https://learn.arcgis.com/de/projects/map-a-historic-cholera-outbreak/GUID-3CA27B42-F093-4FE5-B4DA-72D15877BE8A-web.png)

(https://upload.wikimedia.org/wikipedia/commons/thumb/c/cd/Pump_Handle_-_John_Snow_.jpg/1920px-Pump_Handle_-_John_Snow_.jpg)

Anscombe’s Quartet teaches us to look.

(Francis Anscombe, 1973)

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

(https://en.wikipedia.org/wiki/Anscombe's_quartet)

What did we just observe?

„Der Mensch ist ein Augentier.“

- 90% of all information transmitted to the human brain are visual. *(Potter et al., 2014, Atten Percept Psychophys.)*
- 50% of our brain is related to vision, and visual information accounts for 66% of its electrical activity at 2-3 billion firings per seconds. *(Fixot et al., 1957, Am J Ophthalmol.)*
- The human brain processes images 60’000 times faster than text *(Vogel et al., 1986).*
- But some forms of visualisations seem to be more effective than others…
- … so how can we do a good job at data visualisation?

2. Two Simple Rules of Thumb

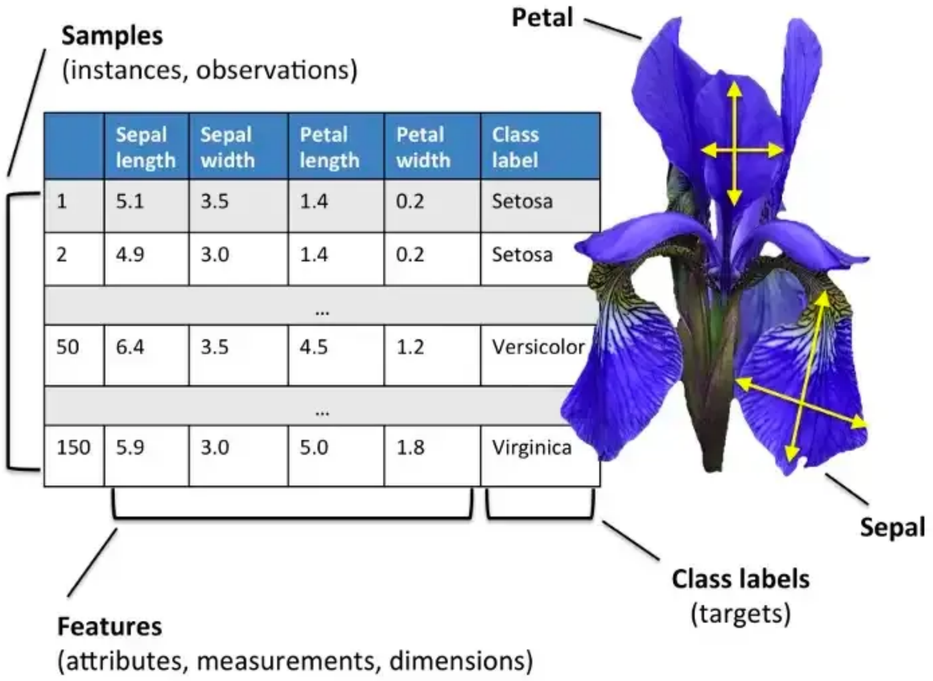
Two simple rules already go a long way.
(Disclaimer: Personal opinion.)

- Simplicity
  - Start with the simplest analyses possible (single variables, subsets, …).
    - Gradually build up complexity of the analysis (and your understanding).
  - Simple kinds of plots first: Always try to plot the raw data (or parts of it) first.
- Mindfulness
  - Every plot is supposed to test / make one point, which you should be keenly aware of before.
    - Every choice you make for this plot follows from this goal.
    - Eliminate everything not required for this („Chart Junk“).
  - Be aware of what a specific kind of plot can tell you…
    - … and what not! (Possibilities for follow-up analyses!)

Let’s apply these rules to the Iris dataset.
(Ronald Fisher, 1936)

Iris Versicolor **Iris Setosa** **Iris Virginica**

(https://raw.githubusercontent.com/RubixML/Iris/master/docs/images/iris-species.png)



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

(https://eminebozkus.medium.com/exploring-the-iris-flower-dataset-4e000bcc266c)

One of **many possible objectives:** Telling apart species using measurements.

We start simple.

We begin with only one variable and directly show the raw data.

- Some variables seem to differ more strongly between species.
- But there’s no way to always tell all three species apart.
- What can these plots **not** tell us?

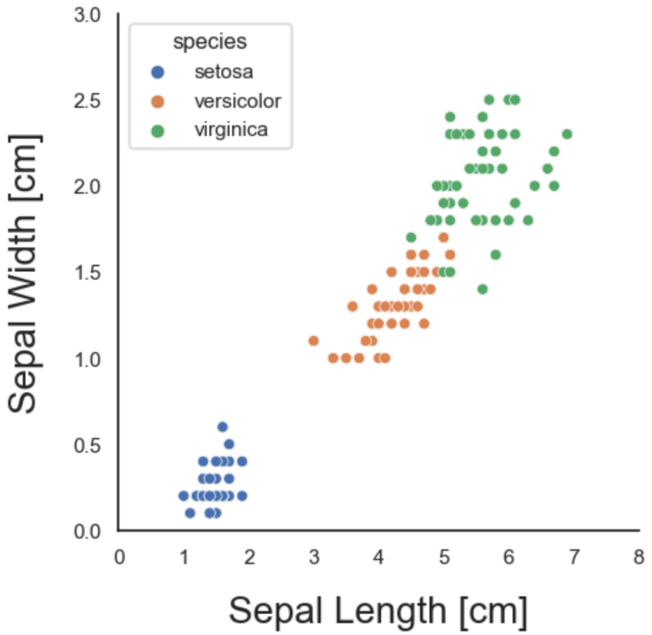
In 2-d, some more patterns emerge.

Let’s focus on the sepals first.

- Sepals don’t seem to carry all too much information.
- Some ways of encoding categorical data are easier to visually grasp.
- Some ways (like size) bias attention towards certain categories.
- Size seems to suggests a continuous rather than a categorical trait…

In 2-d, some more patterns emerge.

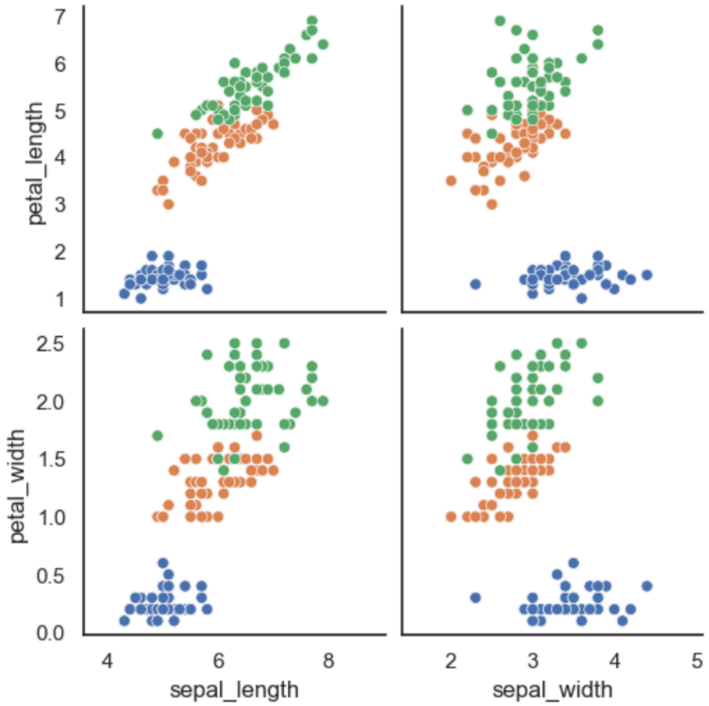
Now let's examine the petals.



- Petals make all three species (nearly perfectly) distinguishable.
- Interestingly, there's a clean correlation, consistent across species — why?
- What do we still need to look at?

In 2-d, some more patterns emerge.

Four plots remain.



- Petal and sepal traits seems to be rather independent — why?
- Four variables required us to look at six plots already…

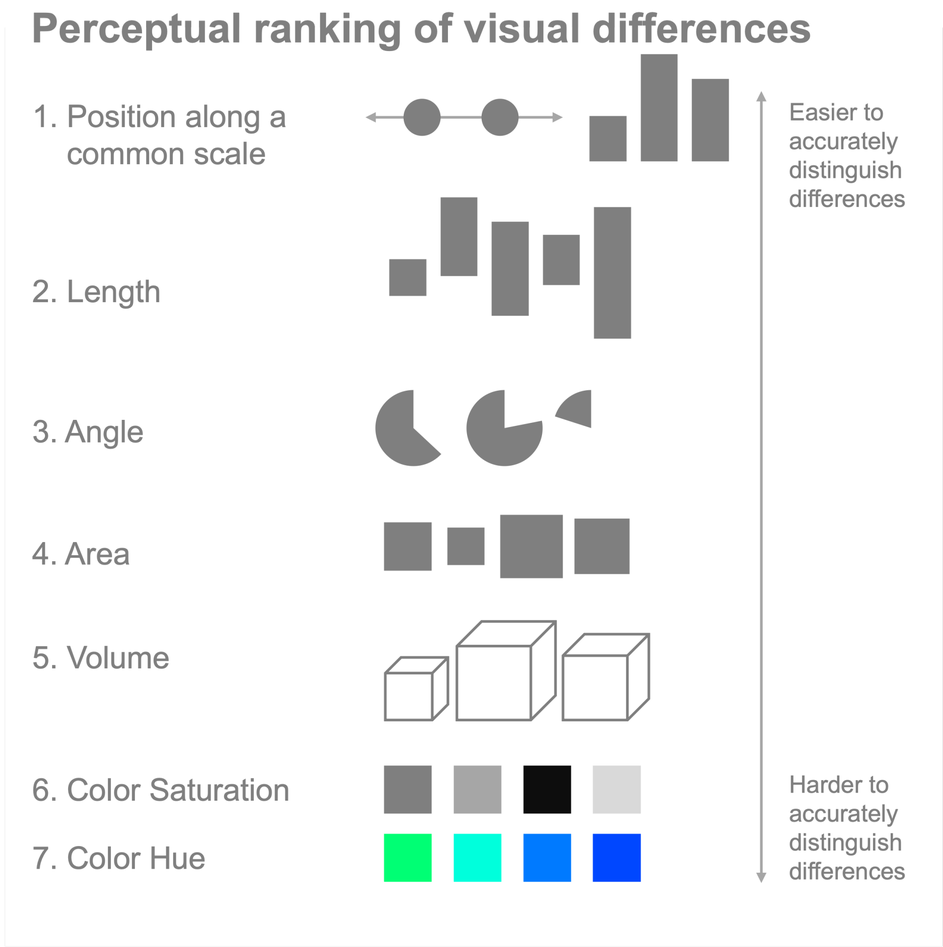
What did we just observe?

- We have distinguished categorical vs continuous variables.
  - They seem to work best with different kinds of visual cues.
    - There is psychological research on what works best!
- More variables means more plots to inspect.
  - There are ways around that.

3. Results From Psychology

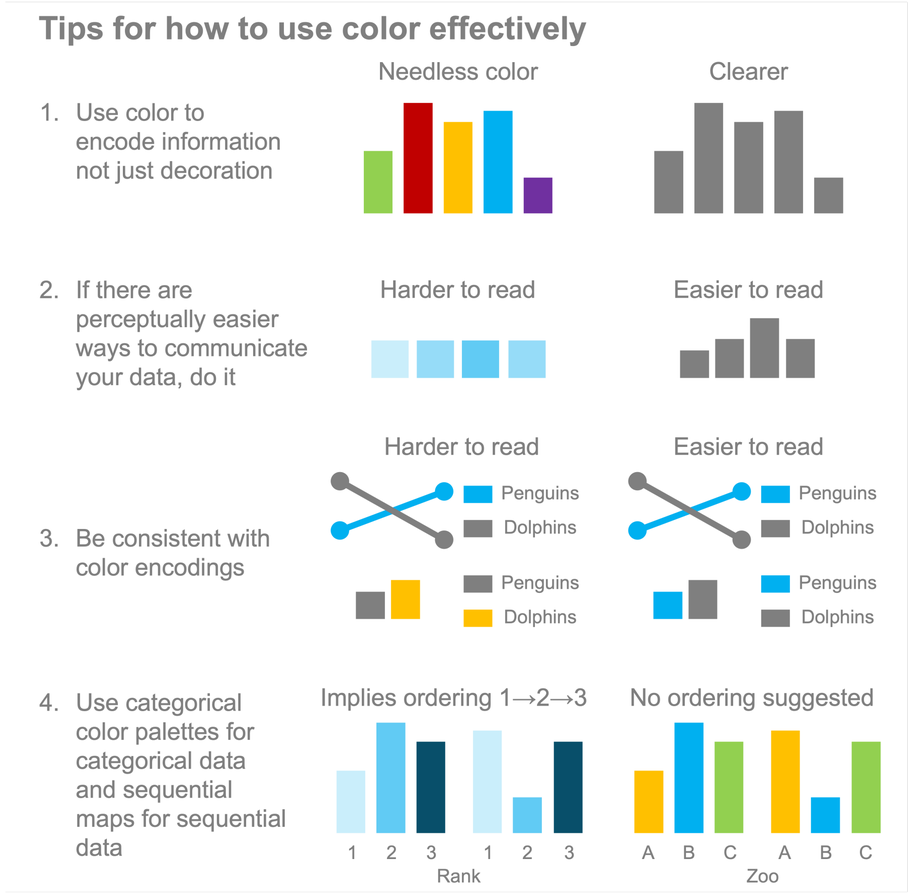
Some visualisations are better than others.

Cleveland & McGill (1985), for the case of continuous variables.



Colour is immensely powerful.

Use it sparingly (for categories) and try to not mislead or confuse.



(https://www.practicaldatascience.org/notebooks/class_5/week_1/1.1.2_effective_plotting_practices.html)

4. The Curse Of Dimensionality

Some first ideas for dealing with it.

- Preselecting features based on univariate inspection…
  - …and hoping for the best.
- Dropping very strongly correlated (sets of) predictors, as they only carry identical information.
  - We'll see an example of this soon.
- Trying to let a Machine Learning model figure it out by itself.
- But clever ways of visualisation can also work as well :).

The classic Diamonds Dataset

(modified from https://www.kyle-w-brown.com/diamonds-prediction/diamonds-data.html, originally sourced from the Loose Diamonds Search Engine)

Format

A data frame with 53940 rows and 12 variables:

- `carat`: Weight of the diamond (0.2–5.01).
- `cut`: Quality of the cut (`Fair`, `Good`, `Very Good`, `Premium`, `Ideal`).
- `color`: Diamond color, from `D` (best) to `J` (worst).
- `clarity`: A measurement of how clear the diamond is (`I1` (worst), `SI2`, `SI1`, `VS2`, `VS1`, `VVS2`, `VVS1`, `IF` (best)).
- `depth`: Width of top of diamond relative to widest point (43–95).
- `table`: Total depth percentage = `z` / mean(`x`, `y`) = 2 * `z` / (`x` + `y`) (43–79).
- `price`: Price in US dollars ($326–$18,823).
- `x`: Length in mm (0–10.74).
- `y`: Width in mm (0–58.9).
- `z`: Depth in mm (0–31.8).

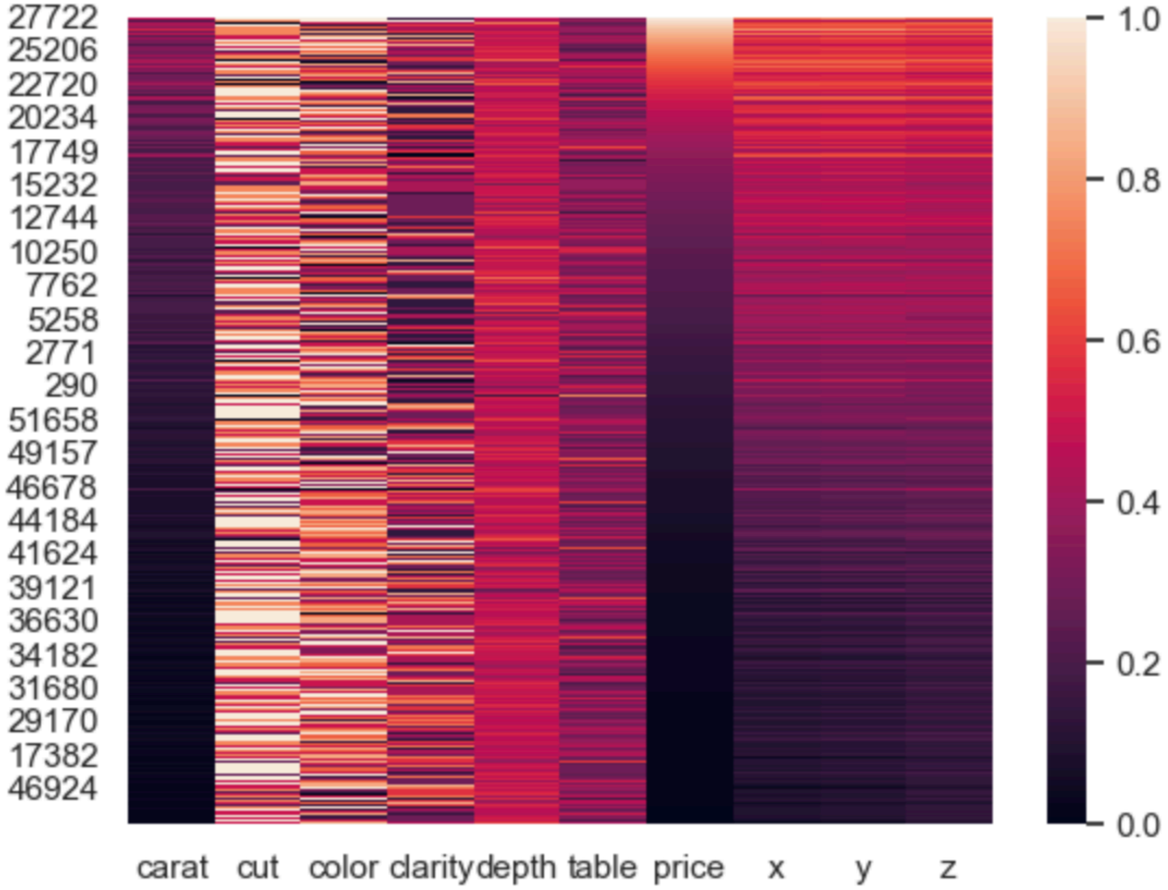
```
diamonds
```

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	5	6	2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	4	6	3	59.8	61.0	326	3.89	3.84	2.31
2	0.23	2	6	5	56.9	65.0	327	4.05	4.07	2.31
3	0.29	4	2	4	62.4	58.0	334	4.20	4.23	2.63
4	0.31	2	1	2	63.3	58.0	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...	...
53897	0.72	5	7	3	60.8	57.0	2757	5.75	5.76	3.50
53898	0.72	2	7	3	63.1	55.0	2757	5.69	5.75	3.61
53899	0.70	3	7	3	62.8	60.0	2757	5.66	5.68	3.56
53900	0.86	4	3	2	61.0	58.0	2757	6.15	6.12	3.74
53901	0.75	5	7	2	62.2	55.0	2757	5.83	5.87	3.64

53902 rows × 10 columns

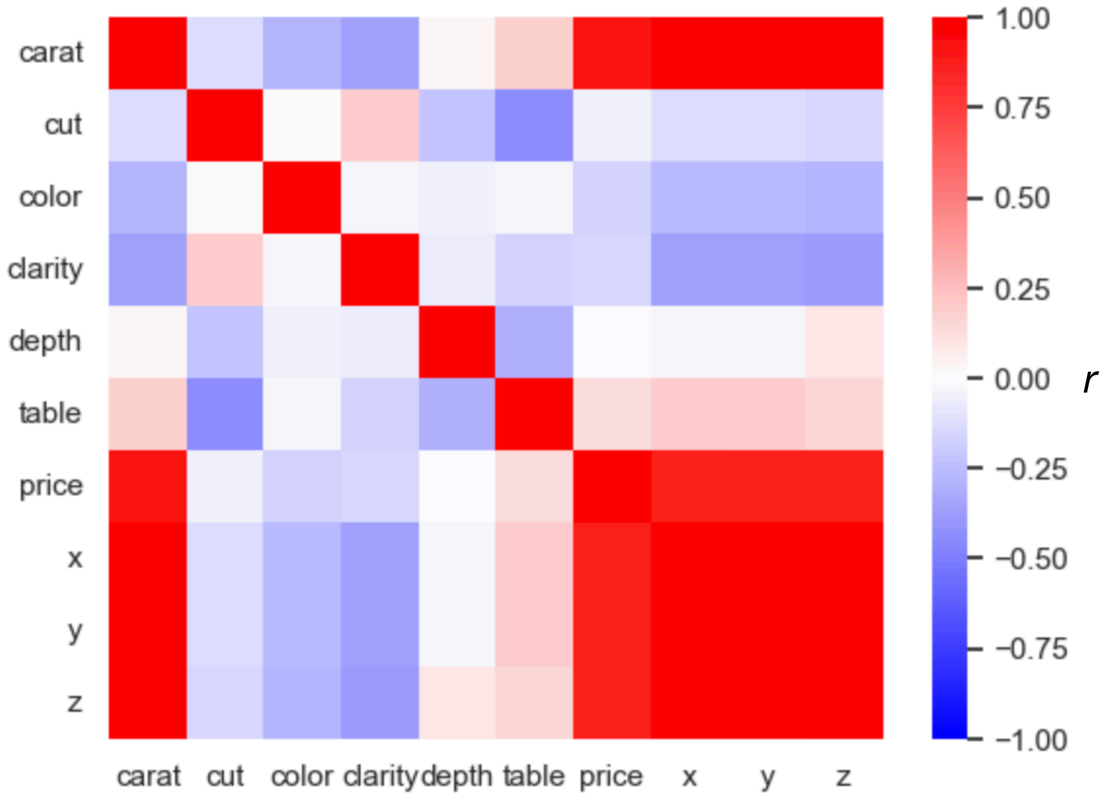
A heatmap is a nice „map“ of the dataset.

(We transformed all variables to 0…1 and sorted entries by price.)



- Price, carat, x, y, z seem to be strongly correlated with each other.
- Cut, colour, clarity, depth seem not strongly correlated.
  - But it is well-known that these factors influence pricing.
  - What are we missing?
- Assessing correlation patterns does not scale (30’000 human genes…).

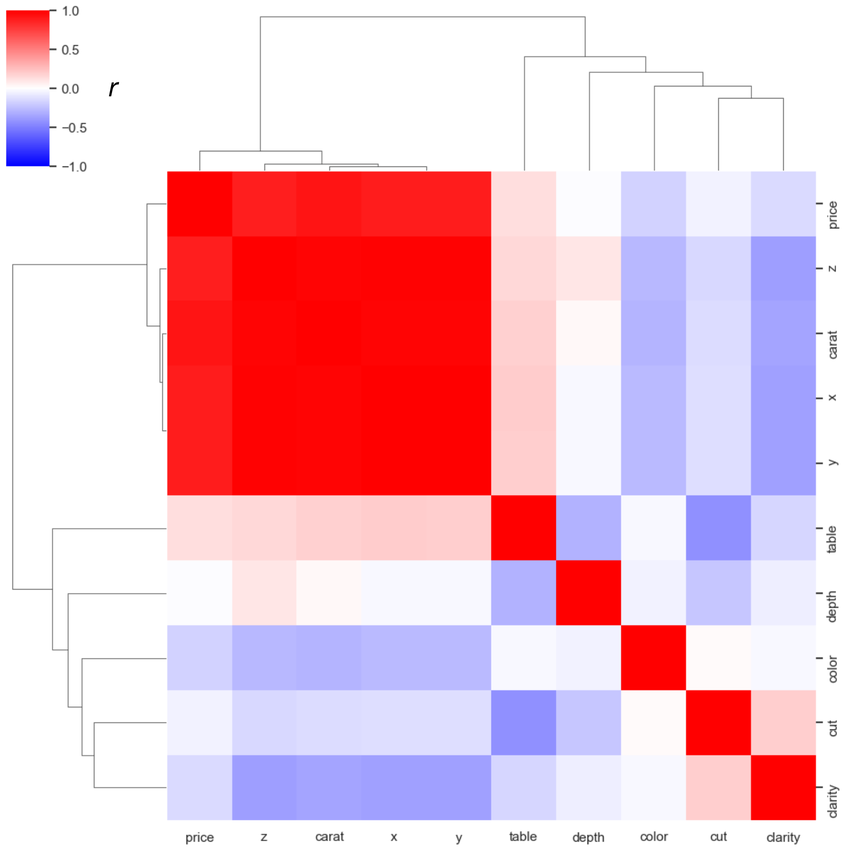
A heatmap of correlation coefficients can help.



- We use a different colormap now — why?
- Correlation coefficients confirm the pattern we just observed by inspection.

A clustermap tries to group correlated features.

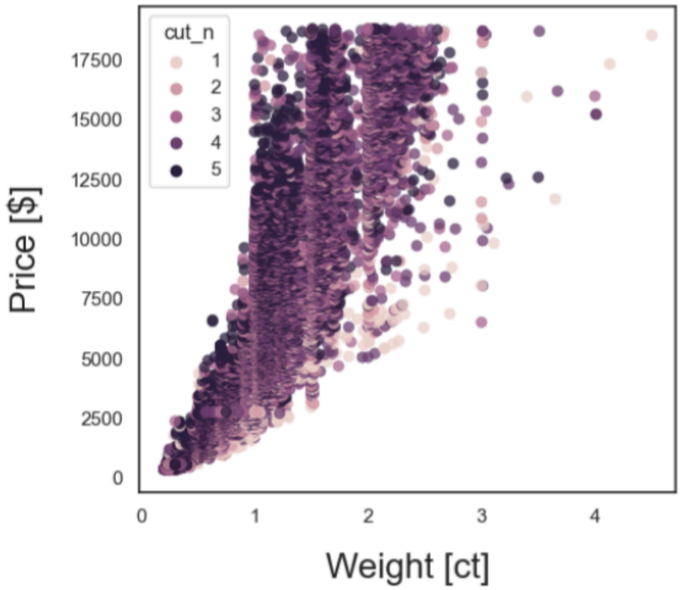
(Cave: There is some mathematical machinery behind this, so know what you are doing.)

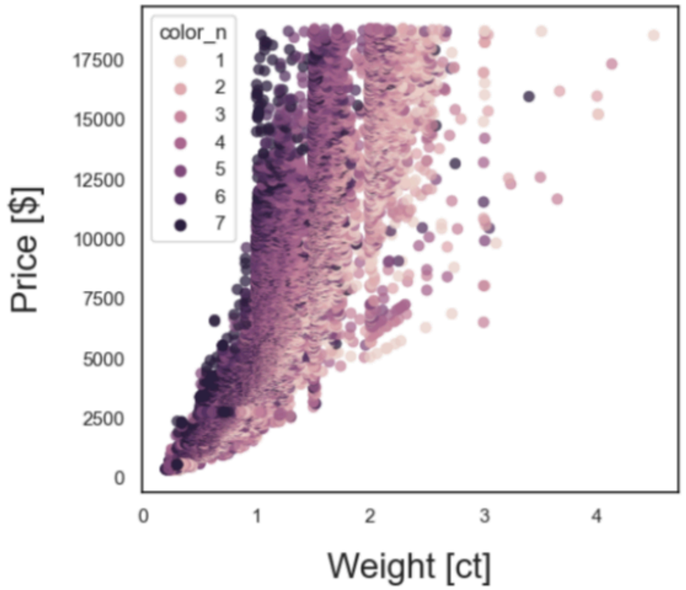


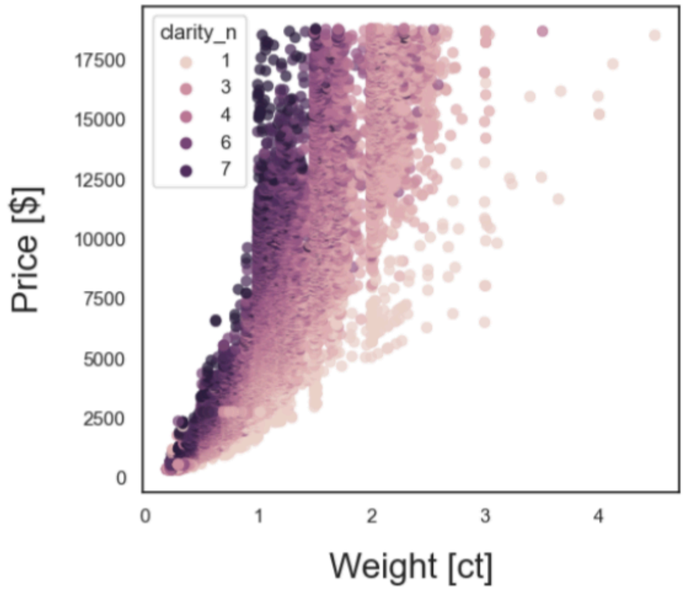
- Clustering of correlation coefficients again confirms the pattern we just observed by inspection.
- However, price still seems uncorrelated with cut, colour, clarity.
  - Why?

Cut, colour and clarity are secondary factors.

Also note the sharp increase in price at exactly 1ct ;).



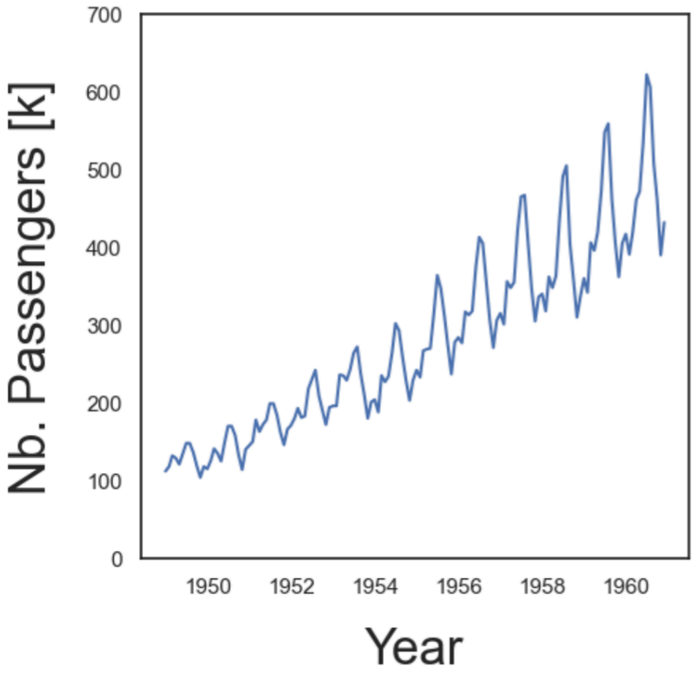


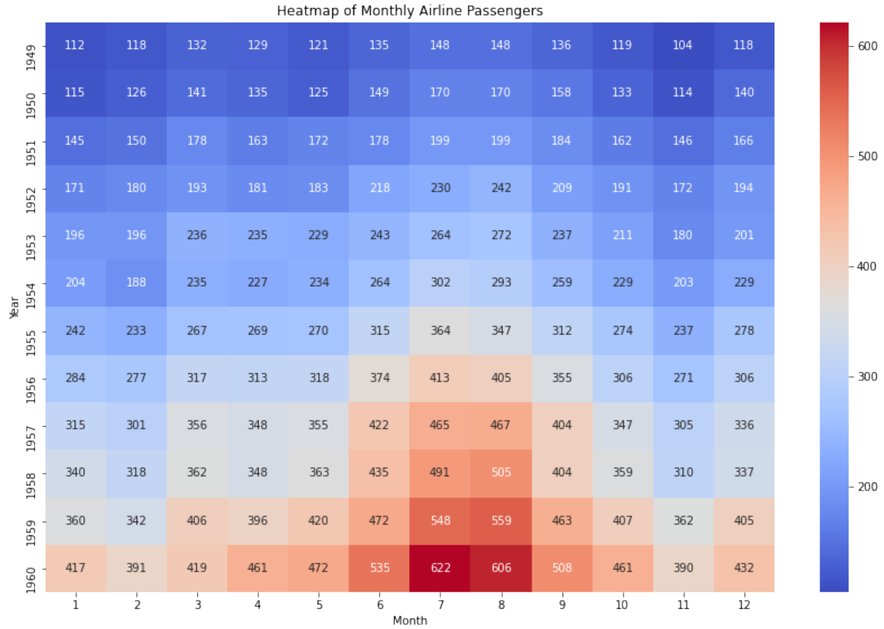


5. Heatmaps for Time Series Data

Airline passenger data shows seasonality.

Box, Jenkins & Reinsel (1976)

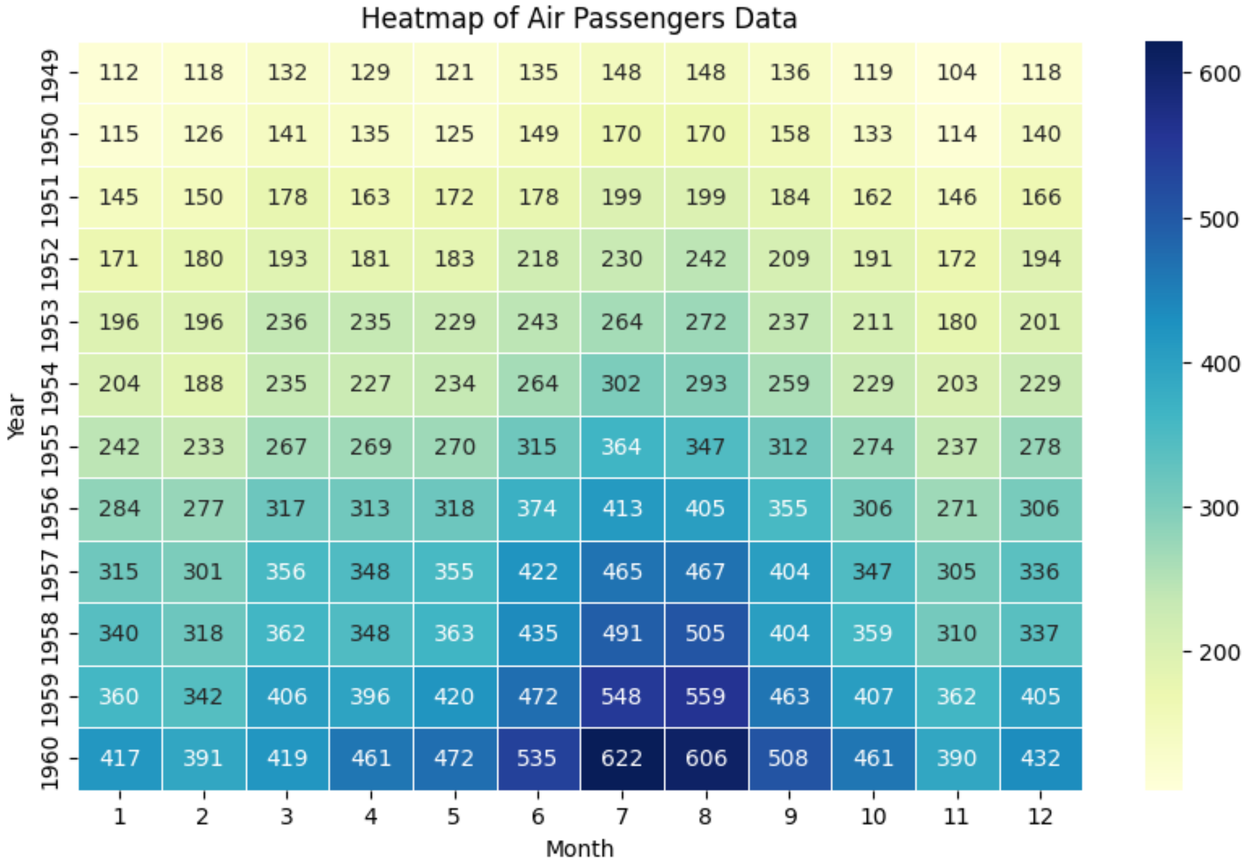




(https://medium.com/@kylejones_47003/time-series-visualization-for-business-analysis-with-python-5df695543d4a)

- Numbers increase between years.
- July and August seem to dominate within a year.
- Is there any design choice you disagree with?

A continuous scale is more appropriate.



(https://python.plainenglish.io/understanding-time-series-data-35ae64d95fe8)

- What can we **not** see?
  - Trends within weeks, days, …
  - Does the ratio between months stay consistent over years?

6. Conclusions

Conclusions

- „Der Mensch ist ein Augentier.“
- Start with the **simplest** plots / analyses imaginable, and work as closely to the raw data as possible.
  - Abstractions / more complex plots do have their place, but should only be used when really necessary.
- Data can be continuous or categorical. Choose type of visualisation accordingly.
  - Keep in mind that some types of visualisation can be suggestive…
  - …and some are just Bad (TM).
- Be **mindful** of what a particular plot
  - can do (-> eliminate everything unnecessary),
  - and what not (-> follow-up questions).

Thanks for your attention!

- Questions always very welcome!
- Feel free to hit me up:
  - On Telegram: @GermanCoyote
  - On Matrix: @yote:catgirl.cloud
  - And of course in person…