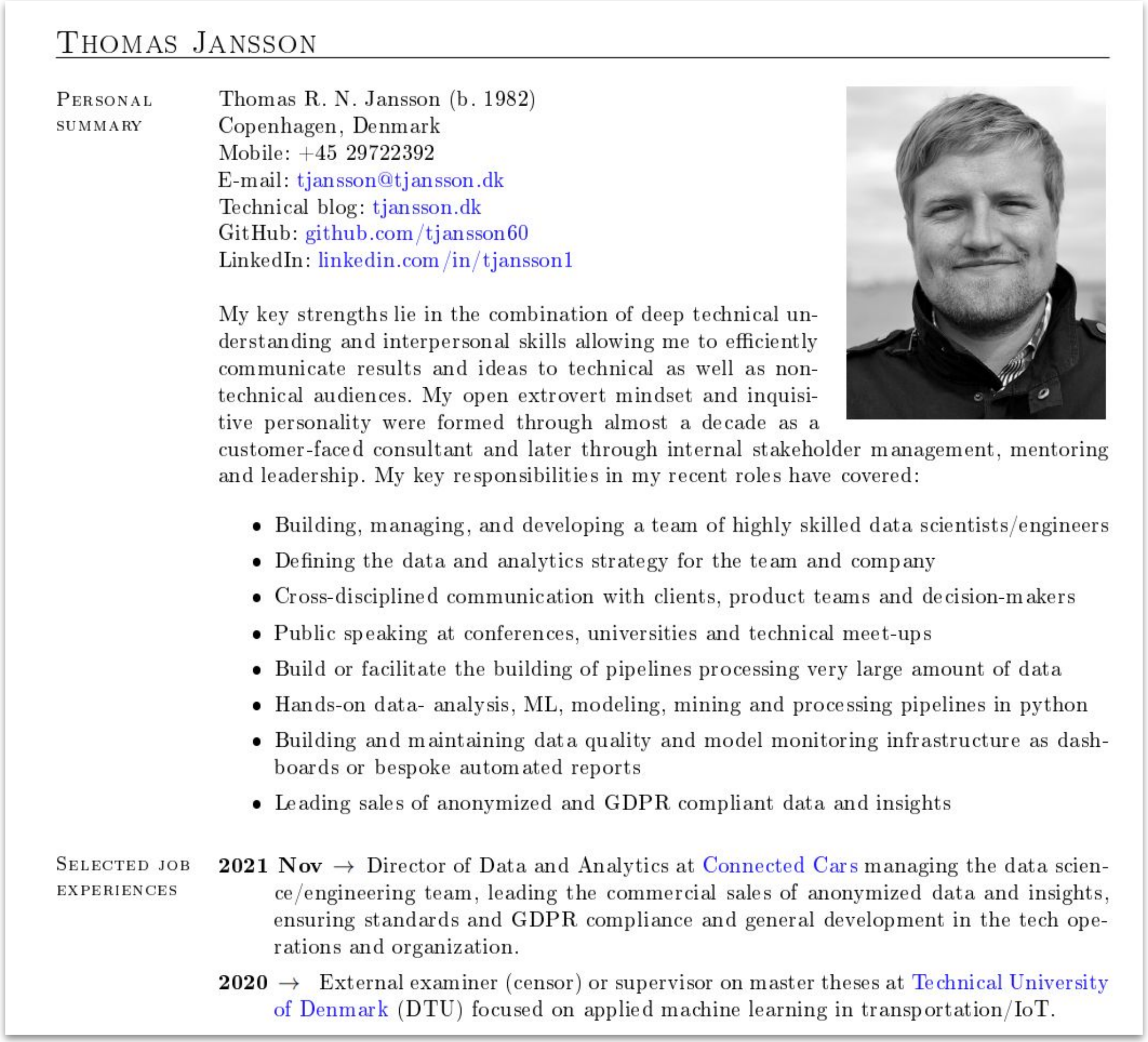
connected cars

Storytelling with data

and how to avoid common pitfalls

About myself



THOMAS JANSSON

PERSONAL SUMMARY

Thomas R. N. Jansson (b. 1982)
Copenhagen, Denmark
Mobile: +45 29722392
E-mail: tjansson@tjansson.dk
Technical blog: tjansson.dk
GitHub: github.com/tjansson60
LinkedIn: linkedin.com/in/tjansson1

My key strengths lie in the combination of deep technical understanding and interpersonal skills allowing me to efficiently communicate results and ideas to technical as well as non-technical audiences. My open extrovert mindset and inquisitive personality were formed through almost a decade as a customer-faced consultant and later through internal stakeholder management, mentoring and leadership. My key responsibilities in my recent roles have covered:

- Building, managing, and developing a team of highly skilled data scientists/engineers
- Defining the data and analytics strategy for the team and company
- Cross-disciplined communication with clients, product teams and decision-makers
- Public speaking at conferences, universities and technical meet-ups
- Build or facilitate the building of pipelines processing very large amount of data
- Hands-on data- analysis, ML, modeling, mining and processing pipelines in python
- Building and maintaining data quality and model monitoring infrastructure as dashboards or bespoke automated reports
- Leading sales of anonymized and GDPR compliant data and insights

SELECTED JOB EXPERIENCES

2021 Nov → Director of Data and Analytics at Connected Cars managing the data science/engineering team, leading the commercial sales of anonymized data and insights, ensuring standards and GDPR compliance and general development in the tech operations and organization.

2020 → External examiner (censor) or supervisor on master theses at Technical University of Denmark (DTU) focused on applied machine learning in transportation/IoT.

Work

- Director of data and analytics at the automotive IoT company Connected Cars
- External examiner (censor) and occasionally co-supervisor on master theses at DTU focused on applied machine learning in transportation/IoT.
- Previously at worked at large companies such as Schlumberger and Mærsk as well as smaller start-up/scale-ups namely Qeye Labs.

Private

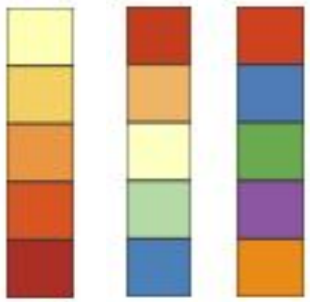
- I am currently quite interested in carbon steel pans, knife sharpening, preparing a perfect espresso puck and youtube videos on meticulously restoring old paintings and powerwashing
- My wife is a musician and she does not work with data
- I have 2 kids aged 7 and 10. They play music, but do not work with data ... yet.

Agenda

1. Colorscales
2. Tabular data
3. How to communicate uncertainty
4. Highlight the change - not the plots
5. Scatterplots
6. Time series data
7. Find anomalies with a glance
8. The right plot to the right people
9. Easily digestible visualizations
10. Disturbing examples
11. Key takeaways and further reading

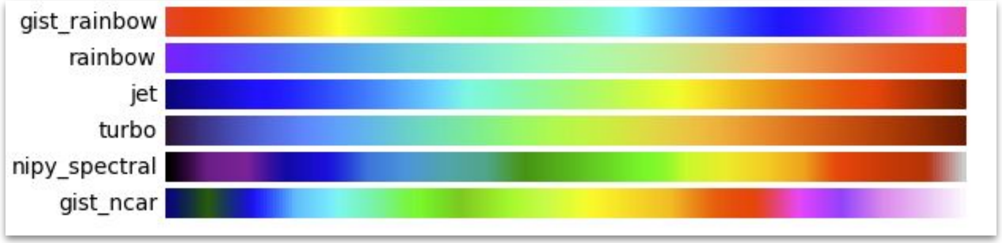
Color Scales

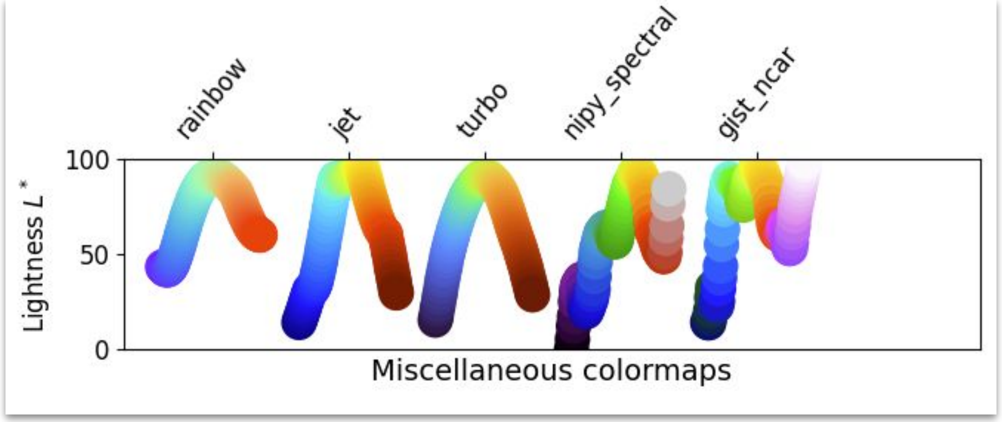
- [Sequential, diverging or qualitative?]

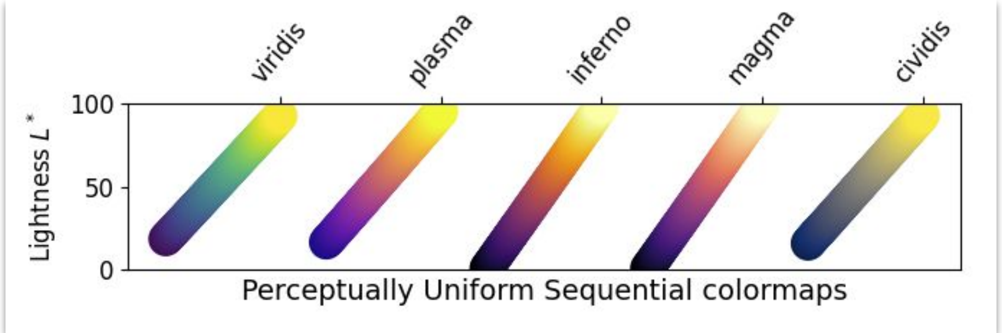


- Color blindness:
  - ~4% of all people: 8% of males, 0.4% females
- B/W friendly?
  - Is the same story told in B/W on paper?
  - ReMarkable/Ebook?
- Projector, online meeting or bad screen friendly?
  - Bright colored lines on white background can often not be seen
  - Cheap screens wash-out colors and can change the conclusions
  - Some online meeting compresses colors harshly
- [Color psychology:]
  - Is green good and red bad? Which culture? Religion?
  - Color scales in shades of company colors can seem beautiful, but might be misleading to the story being told with the data.

https://matplotlib.org/stable/tutorials/colors/colormaps.html





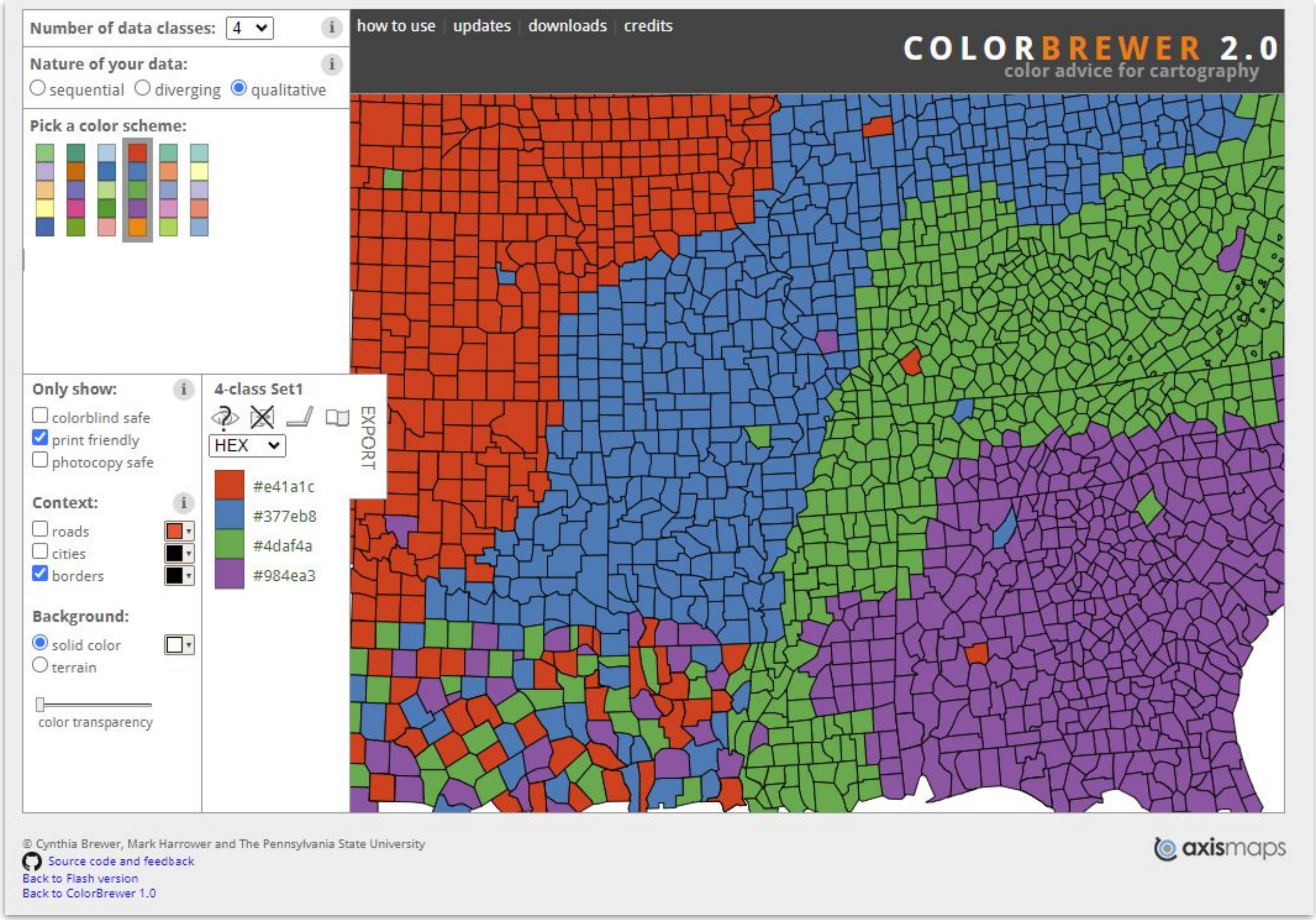


Color Scales - ColorBrewer

colorbrewer2.org

The original ColorBrewer (v1.0) was funded by the NSF Digital Government program during 2001-02, and was designed at the GeoVISTA Center at Penn State (National Science Foundation Grant No. 9983451, 9983459, 9983461).

The design and rebuilding of this new version (v2.0) was donated by Axis Maps LLC, winter 2009 and updated in 2013.



Color Scales - Decision making

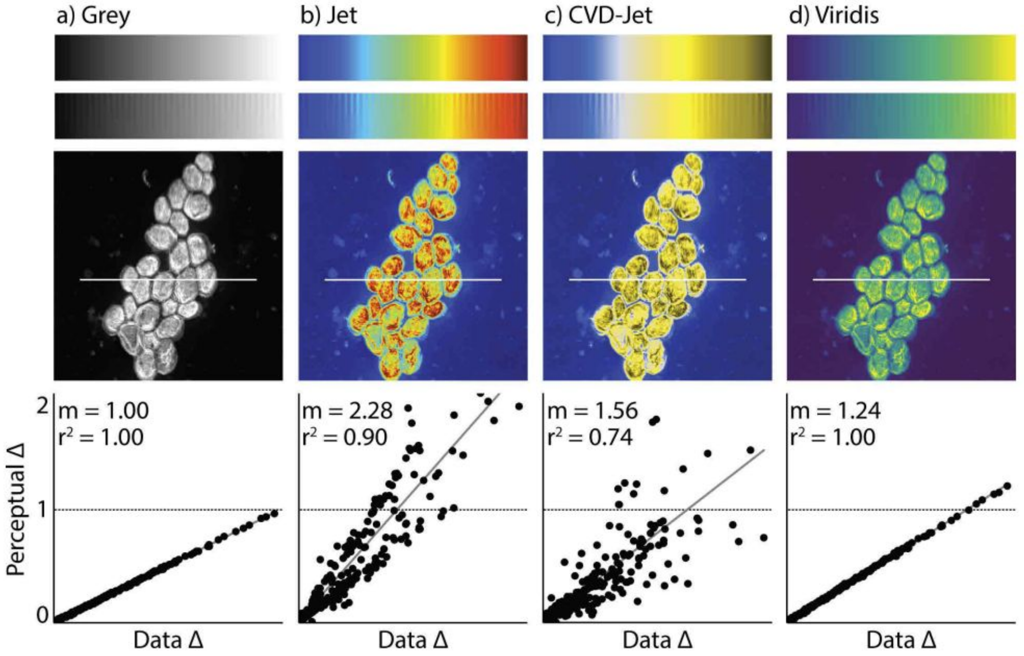


Fig 1. Example of a misleading colormap. Comparison between different colormaps overlaid onto the test image by Kovesi and a nanoscale secondary ion mass spectrometry image. Colormaps are as follows: (a) perceptually uniform grayscale, (b) jet, (c) jet as it appears to someone with red-green colorblindness, and (d) viridis [1], the current gold standard colormap. Below each NanoSIMS image is a corresponding ªcolormap-data perceptual sensitivityº (CDPS) plot, which compares perceptual differences of the colormap to actual, underlying data differences. *m* is the slope of the fitted line and r^2 is the coefficient of determination calculated using a simple linear regression. An example of how the data may be misinterpreted are evident in the bright yellow spots in (b) and (c), which appear to represent significantly higher values than the surrounding regions. However, in fact, the dark red (in b) and dark yellow (in c) actually represent the highest values. For someone who is red-green colorblind, this is made even more difficult to interpret due to the broad, bright band in the center of the colormap with values that are difficult to distinguish.

https://doi.org/10.1371/journal.pone.0199239.g001

https://arxiv.org/ftp/arxiv/papers/1712/1712.01662.pdf

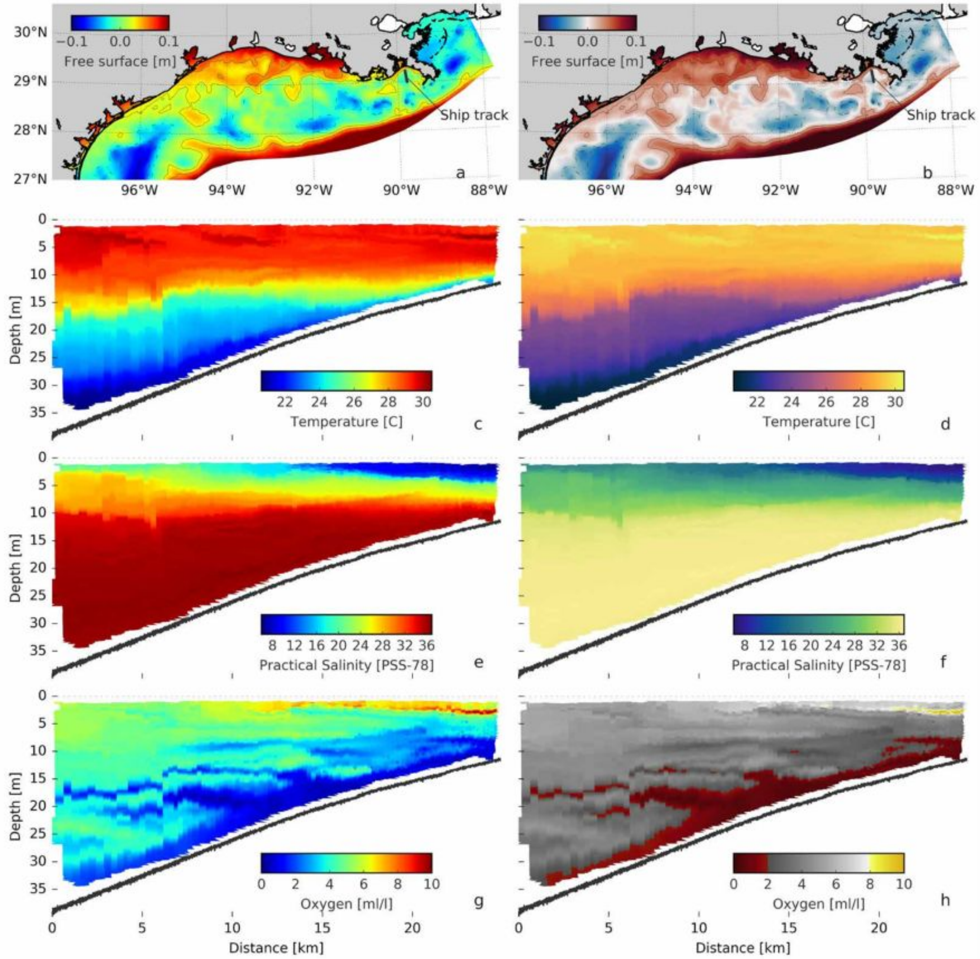


FIGURE 3. Examples of colormap use. The free surface in the northwestern Gulf of Mexico is shown (subplots a and b, where solid line/red are positive values, dash-dot/blue are negative values, and gray solid lines are bathymetric contours) with a ship track indicated. Data collected from an undulating towed vehicle along the ship track are shown are shown in the following rows. Depth vs. distance plots along the track are temperature (c–d), salinity (e–f), and oxygen (g–h). Each property is shown at left in jet and at right in a colormap from cmocean. *Data are recent work of authors DiMarco and Zimmerle for an atlas of oceanographic observations of the mechanisms controlling hypoxia*

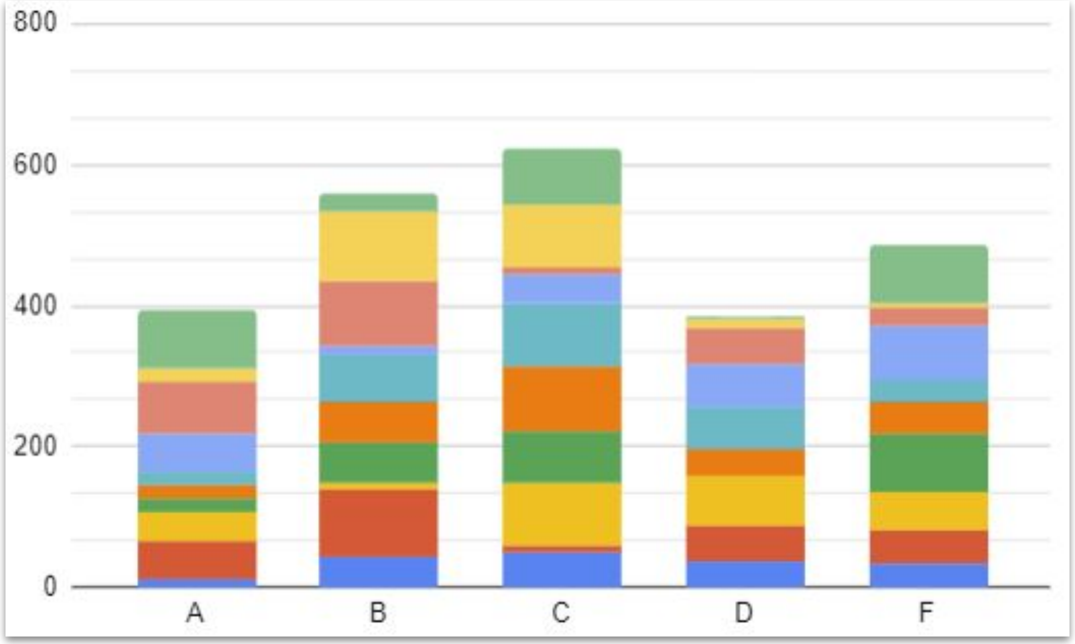
https://www.researchgate.net/publication/307517997_True_Colors_of_Oceanography_Guidelines_for_Effective_and_Accurate_Colormap_Selection

Tabular data

Help the reader/audience get to the conclusion as fast as possible.

What should the eye be attracted to? What is the take-away?

- Right adjusted numbers
- Same number of decimals
- Color-coded values
- Visualize values in any other way than tabular form



10.60993289	44.59277676	50.61605754	38.82107646	33.48668592
56.10516942	93.13326587	10.3314877	49.36636586	47.04618504
39.28700773	9.942695396	86.53846721	68.39615183	55.86812168
21.85301722	58.27251118	73.64551648	0.2535518372	82.22323598
17.95054359	57.85249146	94.46657444	39.79791777	43.15542339
18.94158961	66.7944672	88.92852505	60.31992596	32.8086336
52.47441208	10.80713061	39.2798807	59.5477883	77.27248558
74.7521607	94.40775885	11.05905519	50.82650361	26.33463134
18.04895091	98.76256847	87.86336686	14.63093879	5.582924948
83.89442457	23.27182247	80.70338847	2.582754976	80.75585779

Left aligned and varying number of decimals

10.6099329	44.5927768	50.6160575	38.8210765	33.4866859
56.1051694	93.1332659	10.3314877	49.3663659	47.0461850
39.2870077	9.9426954	86.5384672	68.3961518	55.8681217
21.8530172	58.2725112	73.6455165	0.2535518	82.2232360
17.9505436	57.8524915	94.4665744	39.7979178	43.1554234
18.9415896	66.7944672	88.9285251	60.3199260	32.8086336
52.4744121	10.8071306	39.2798807	59.5477883	77.2724856
74.7521607	94.4077588	11.0590552	50.8265036	26.3346313
18.0489509	98.7625685	87.8633669	14.6309388	5.5829249
83.8944246	23.2718225	80.7033885	2.5827550	80.7558578

Right alignment and fixed number of decimals

10.6099329	44.5927768	50.6160575	38.8210765	33.4866859
56.1051694	93.1332659	10.3314877	49.3663659	47.0461850
39.2870077	9.9426954	86.5384672	68.3961518	55.8681217
21.8530172	58.2725112	73.6455165	0.2535518	82.2232360
17.9505436	57.8524915	94.4665744	39.7979178	43.1554234
18.9415896	66.7944672	88.9285251	60.3199260	32.8086336
52.4744121	10.8071306	39.2798807	59.5477883	77.2724856
74.7521607	94.4077588	11.0590552	50.8265036	26.3346313
18.0489509	98.7625685	87.8633669	14.6309388	5.5829249
83.8944246	23.2718225	80.7033885	2.5827550	80.7558578

Right aligned, fixed number of decimals and color-coded

Tabular data - example from the wild

This happens all the time in the real world. To the left is an example I found last week on arXiv

arXiv:2211.07338v1 [astro-ph.IM] 14 Nov 2022
https://arxiv.org/pdf/2211.07338.pdf

H. Krásná et al.: VLBI Celestial and Terrestrial Reference Frames VIE2022b

Table 10. List of sources with angular separation between ICRF3 and VIE2022b-sx larger than 10 mas.

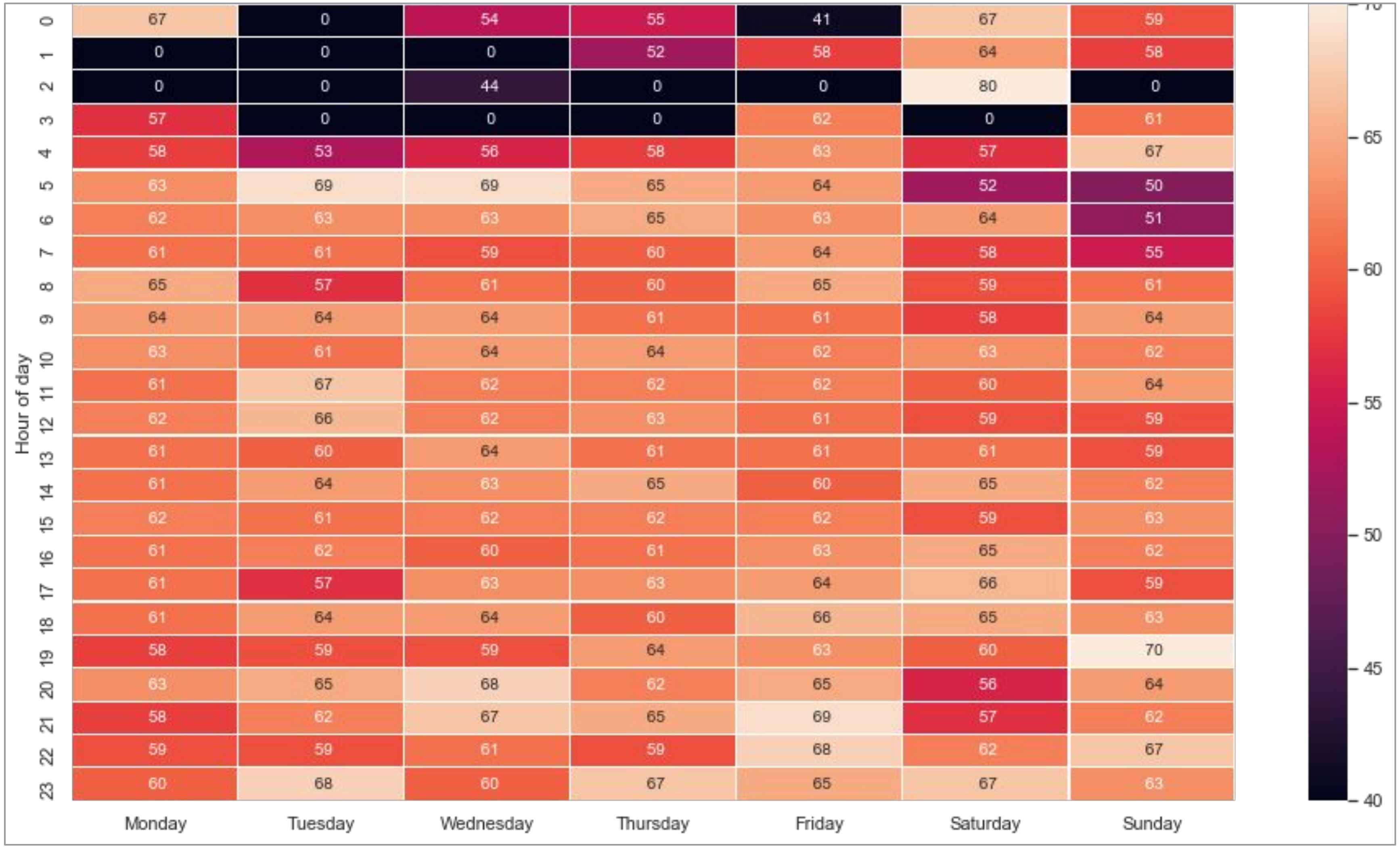
IERS name	IVS name	$\Delta\alpha^*$ [mas]	$\Delta\delta$ [mas]	angular separation [mas]	first obs. [mjd]	last obs. [mjd]	no. of sessions	no. of obs.
0106-391	-	−3.51 ± 4.80	−23.60 ± 18.08	23.86 ± 17.89	58203.3	59460.3	6	64
0134+329	3C48	1.25 ± 0.05	−56.85 ± 0.08	56.87 ± 0.08	48193.8	59378.0	49	1736
0201-440	-	1.34 ± 17.40	−99.25 ± 55.92	99.26 ± 55.92	58143.4	59508.8	4	15
0316+162	CTA21	2.10 ± 0.06	−10.22 ± 0.12	10.44 ± 0.12	50084.5	59378.0	17	1299
0328-060	-	29.73 ± 4.54	−16.02 ± 6.97	33.77 ± 5.19	56874.5	59440.3	8	54
0350+177	-	−6.78 ± 0.84	63.43 ± 1.33	63.79 ± 1.33	57924.7	59405.2	6	116
0512-129	-	−3.68 ± 1.74	9.31 ± 4.41	10.01 ± 4.15	58143.4	59522.9	5	69
0709+008	-	7.36 ± 2.74	7.27 ± 3.12	10.35 ± 2.94	52939.7	58631.3	7	82
0748-378	-	−9.33 ± 10.52	48.22 ± 25.45	49.12 ± 25.07	57011.1	59508.8	8	40
0753-425	-	1.46 ± 0.73	12.36 ± 2.24	12.45 ± 2.22	55370.8	59522.9	7	123
0903-392	-	1.93 ± 4.72	−15.35 ± 14.04	15.47 ± 13.94	57046.0	58981.5	7	32
0932-281	-	6.54 ± 1.79	7.87 ± 4.55	10.23 ± 3.68	50687.3	59508.8	6	99
0951+699	-	12.00 ± 35.25	−4.94 ± 34.56	12.98 ± 35.15	58203.3	58592.8	3	12
1015-314	-	3.58 ± 2.21	−17.51 ± 5.26	17.87 ± 5.17	52305.8	59560.6	8	77
1117-248	-	−12.40 ± 2.21	11.22 ± 3.09	16.72 ± 2.64	50631.3	59463.5	12	71
1306+660	-	−15.08 ± 3.58	−33.07 ± 4.77	36.35 ± 4.59	57011.1	59405.2	8	65
1305-241	-	6.90 ± 8.14	14.74 ± 9.89	16.27 ± 9.60	58158.9	59440.3	5	44
1328+254	-	8.48 ± 0.57	17.13 ± 0.89	19.11 ± 0.83	52408.7	58644.9	6	164
1422+268	-	−2.98 ± 4.87	−12.57 ± 4.64	12.91 ± 4.66	58136.6	58981.5	4	46
1507-246	-	70.00 ± 1.80	−128.92 ± 3.36	146.70 ± 3.08	57924.7	59611.7	8	68
1539-093	-	−29.52 ± 12.78	13.61 ± 10.61	32.50 ± 12.43	50575.3	58981.5	9	36
1612+797	-	7.06 ± 0.64	−7.36 ± 0.74	10.20 ± 0.70	53780.1	58510.3	6	237
1657-298	-	346.60 ± 5.03	−687.18 ± 8.32	769.64 ± 7.76	57973.7	59611.7	7	40
1706-223	-	−3.66 ± 0.64	−14.04 ± 1.73	14.51 ± 1.68	57011.1	58746.6	5	123
1711-251	-	213.09 ± 188.33	−466.99 ± 364.28	513.31 ± 340.50	57596.8	58981.5	7	16
1755+626	-	−21.04 ± 2.94	−41.25 ± 2.63	46.31 ± 2.70	55370.8	59522.9	9	105
1829-106	-	21.41 ± 4.07	−35.84 ± 3.56	41.74 ± 3.70	51731.8	59560.6	10	17
1858-143	-	−2.82 ± 12.25	28.09 ± 16.33	28.23 ± 16.29	58203.3	58981.5	4	23
1934-638	-	−22.59 ± 0.88	2.69 ± 0.72	22.75 ± 0.88	48765.9	59065.7	8	36
2028-204	-	494.59 ± 15.25	−1021.10 ± 32.61	1134.58 ± 30.10	58203.3	59460.3	5	19
2105-212	-	9.91 ± 1.25	−4.23 ± 2.42	10.77 ± 1.49	57011.1	59535.8	8	83
2216-007	-	73.13 ± 2.36	−85.80 ± 3.11	112.73 ± 2.82	56266.8	58644.9	6	80
2219-340	-	13.60 ± 6.72	10.41 ± 18.62	17.13 ± 12.51	57098.3	58981.5	7	36
2318-195	-	10.27 ± 0.73	20.12 ± 1.80	22.59 ± 1.64	58143.4	59460.3	6	101
2346+750	-	−1.74 ± 0.47	10.99 ± 0.63	11.12 ± 0.62	57808.9	59560.6	7	134

Tabular data - ATK example, before

Evaluation of speed changes (ATK - Automatisk Trafikkontrol) on a section of road over a week before and after the installation of the ATK.

The colors darken in the after plot indicating a drop in the average speed after the installation.

Without color coding of the table it would be much harder to get a understanding of the data. The colors helps the reader understand the story told be the data.

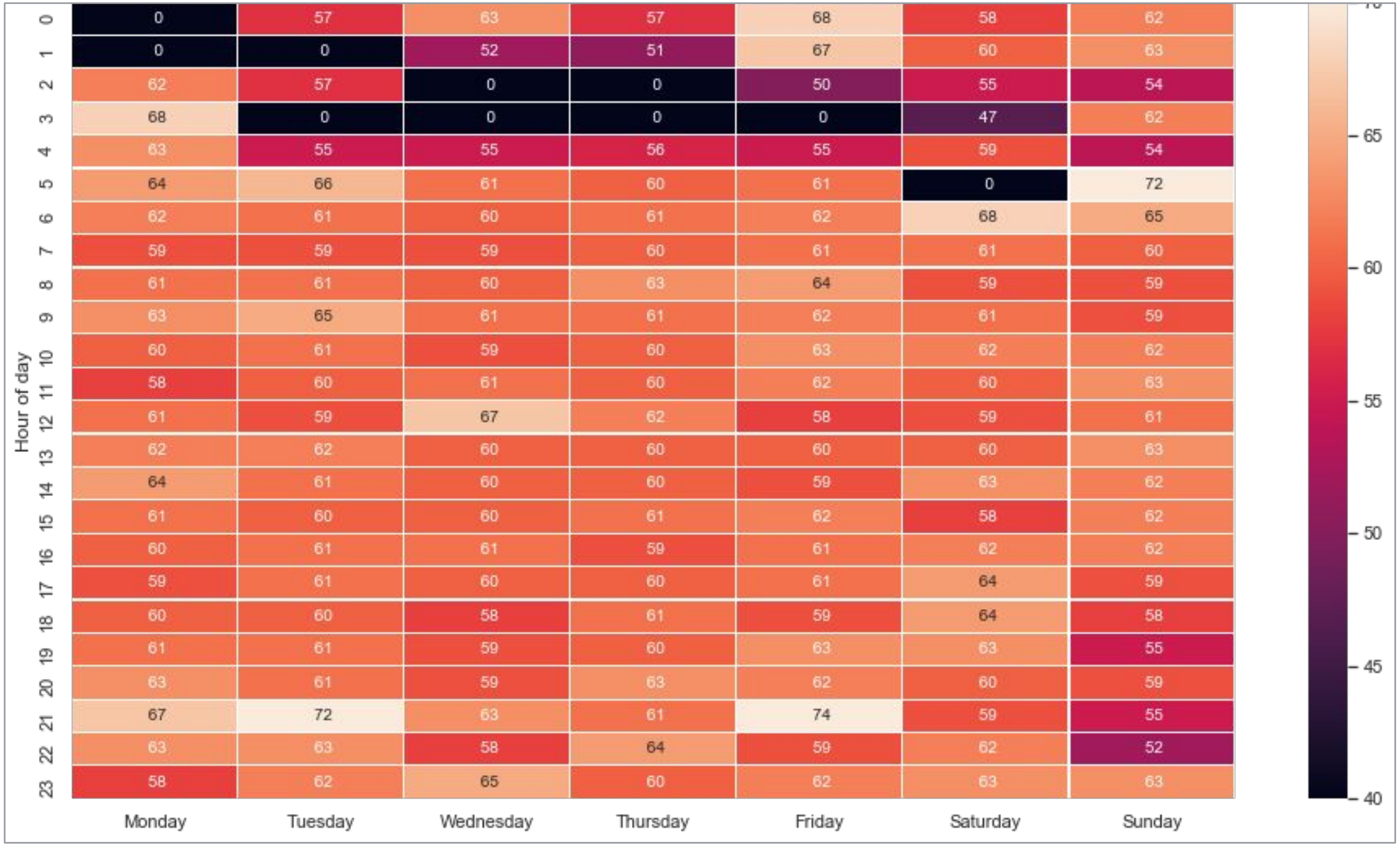


Tabular data - ATK example, after

Evaluation of speed changes (ATK - Automatisk Trafikkontrol) on a section of road over a week before and after the installation of the ATK.

The colors darken in the after plot indicating a drop in the average speed after the installation.

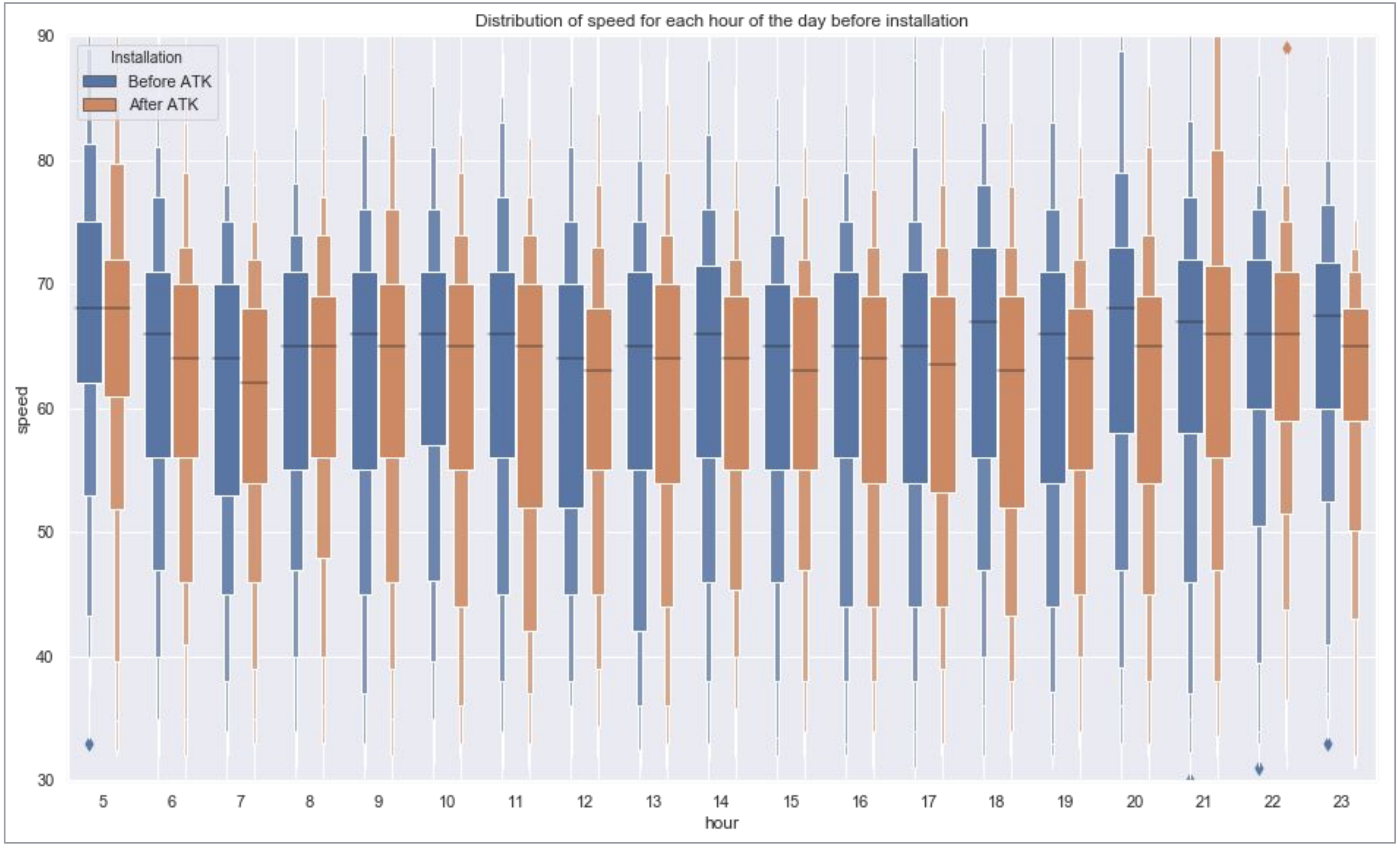
Without color coding of the table it would be much harder to get a understanding of the data. The colors helps the reader understand the story told be the data.



Tabular data - ATK example, comparison

Before the installation (blue) the average speeds are higher and the distributions have more values above 70 km/h.

After (orange) the average speeds are lower and the have less points above the speed limit at 70 km/h.



Using plotting to communicate uncertainty

```python
tips = sns.load_dataset("tips")
tips['Tip [%]'] = (tips['tip']/tips['total_bill'])*100
tips
```

	total_bill	tip	sex	smoker	day	time	size	Tip [%]
0	16.99	1.01	Female	No	Sun	Dinner	2	5.944673
1	10.34	1.66	Male	No	Sun	Dinner	3	16.054159
2	21.01	3.50	Male	No	Sun	Dinner	3	16.658734
3	23.68	3.31	Male	No	Sun	Dinner	2	13.978041
4	24.59	3.61	Female	No	Sun	Dinner	4	14.680765
...	...	...	...	...	...	...	...	...
239	29.03	5.92	Male	No	Sat	Dinner	3	20.392697
240	27.18	2.00	Female	Yes	Sat	Dinner	2	7.358352
241	22.67	2.00	Male	Yes	Sat	Dinner	2	8.822232
242	17.82	1.75	Male	No	Sat	Dinner	2	9.820426
243	18.78	3.00	Female	No	Thur	Dinner	2	15.974441

244 rows × 8 columns

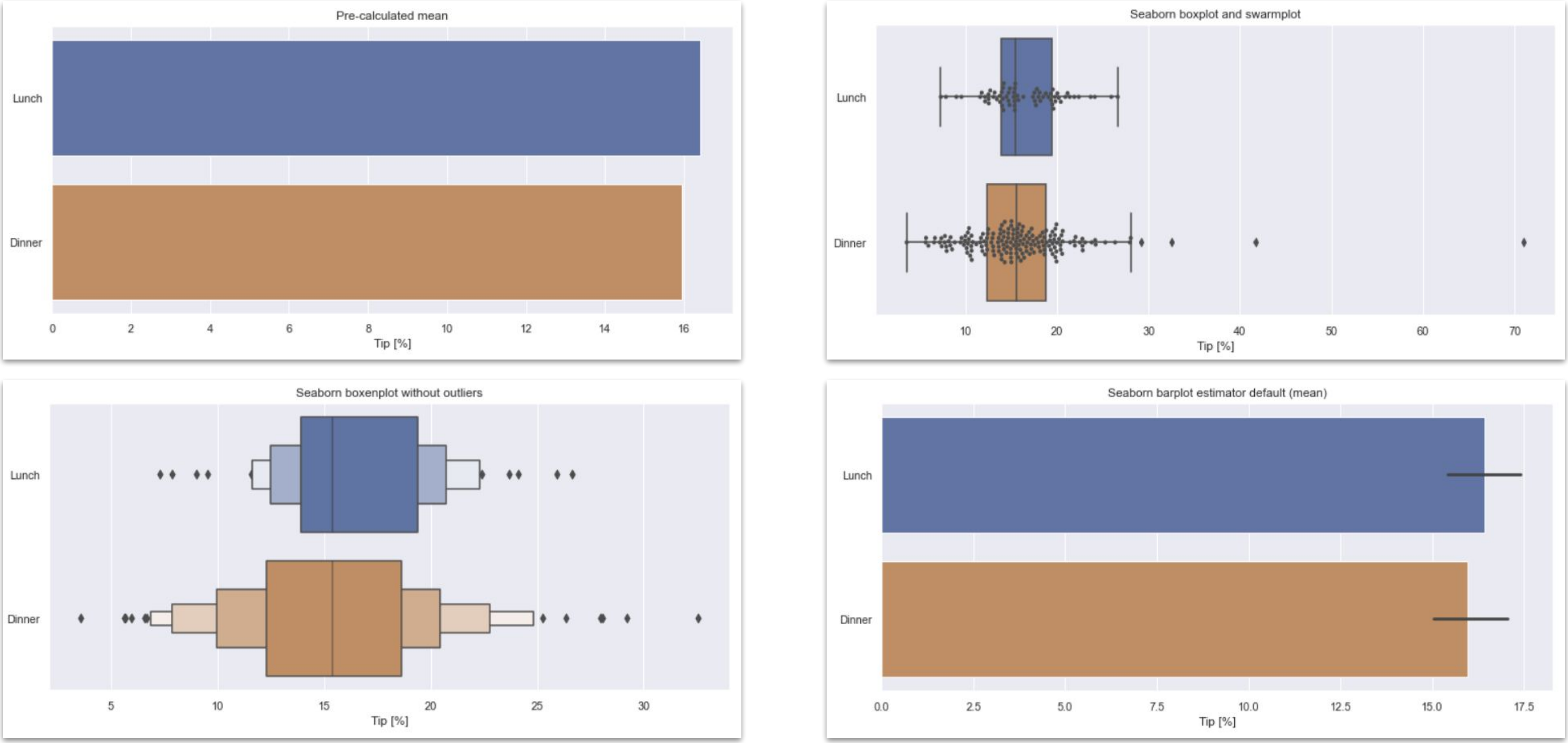
Using the data example “tips” from Seaborn, see:

- github.com/tjansson60/presentation-storytelling-with-data

I will explore the dataset and try to investigate if the tip percentage is larger or smaller at lunch compared to dinner timer.

	count	mean	min	perc_05	quantile_Q1	median	quantile_Q3	perc_95	perc_99	max
time										
Lunch	68	16.412793	7.296137	10.268848	13.914666	15.408357	19.391734	23.220675	26.162351	26.631158
Dinner	176	15.951779	3.563814	7.637882	12.319151	15.540002	18.820878	24.179134	34.846634	71.034483

Using plotting to communicate uncertainty



Using plotting to communicate uncertainty

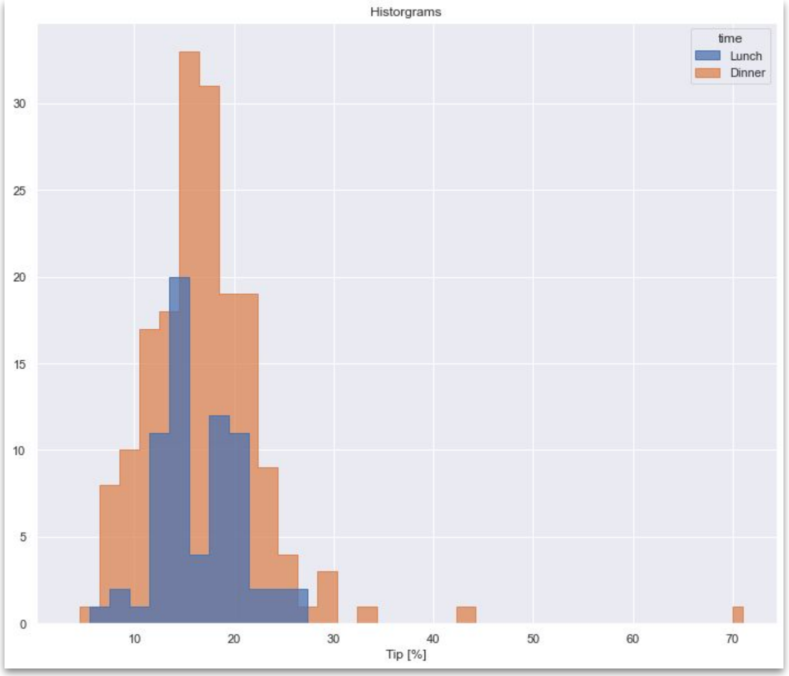
It seems the lunch time tips are bimodal, something that was hard to see in the previous plots.

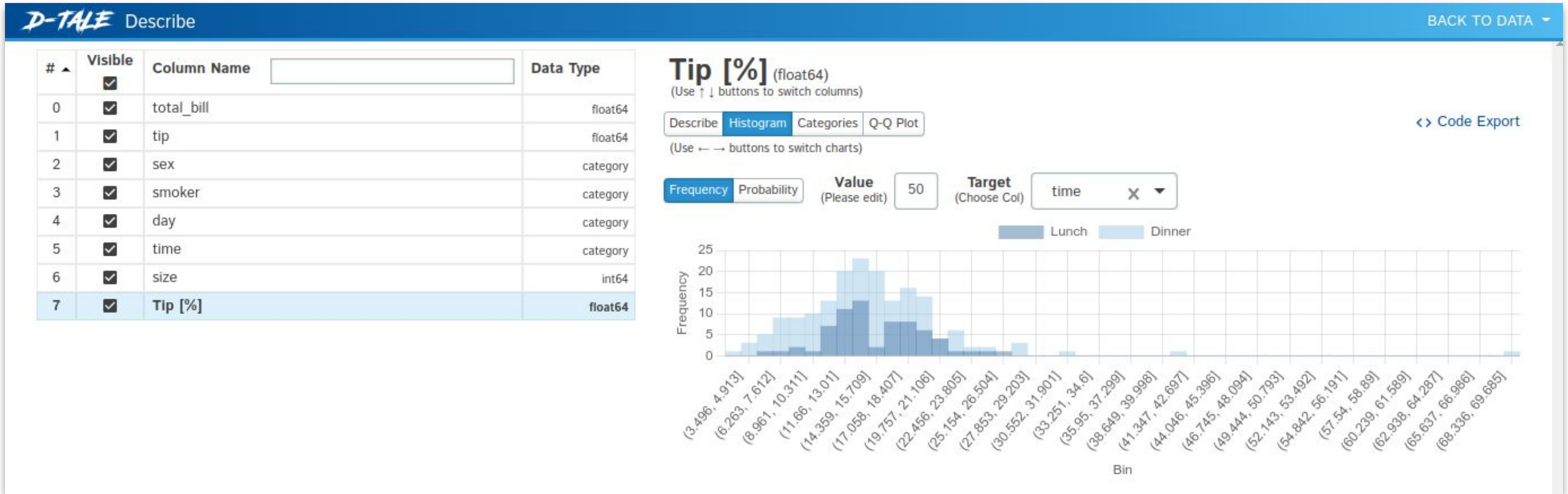
This is why early data exploration is so important!

```python
import seaborn as sns
import dtale

if __name__ == '__main__':
    # Load the data
    tips = sns.load_dataset("tips")
    tips['Tip [%]'] = (tips['tip'] / tips['total_bill']) * 100

    # Start the webserver and show the data
    d = dtale.show(tips, subprocess=False)
    d.open_browser()
```



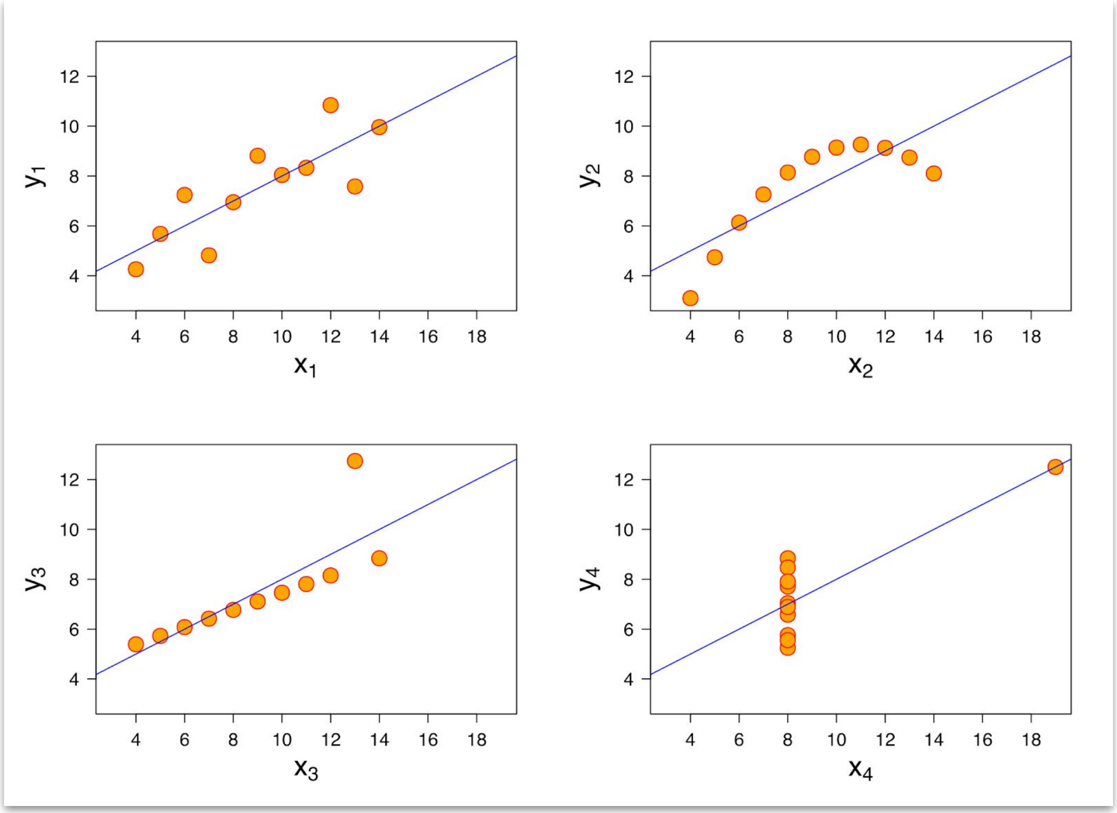


[D-Tale](#) is the combination of a Flask back-end and a React front-end to bring you an easy way to view & analyze Pandas data structures. It integrates seamlessly with ipython notebooks & python/ipython terminals.

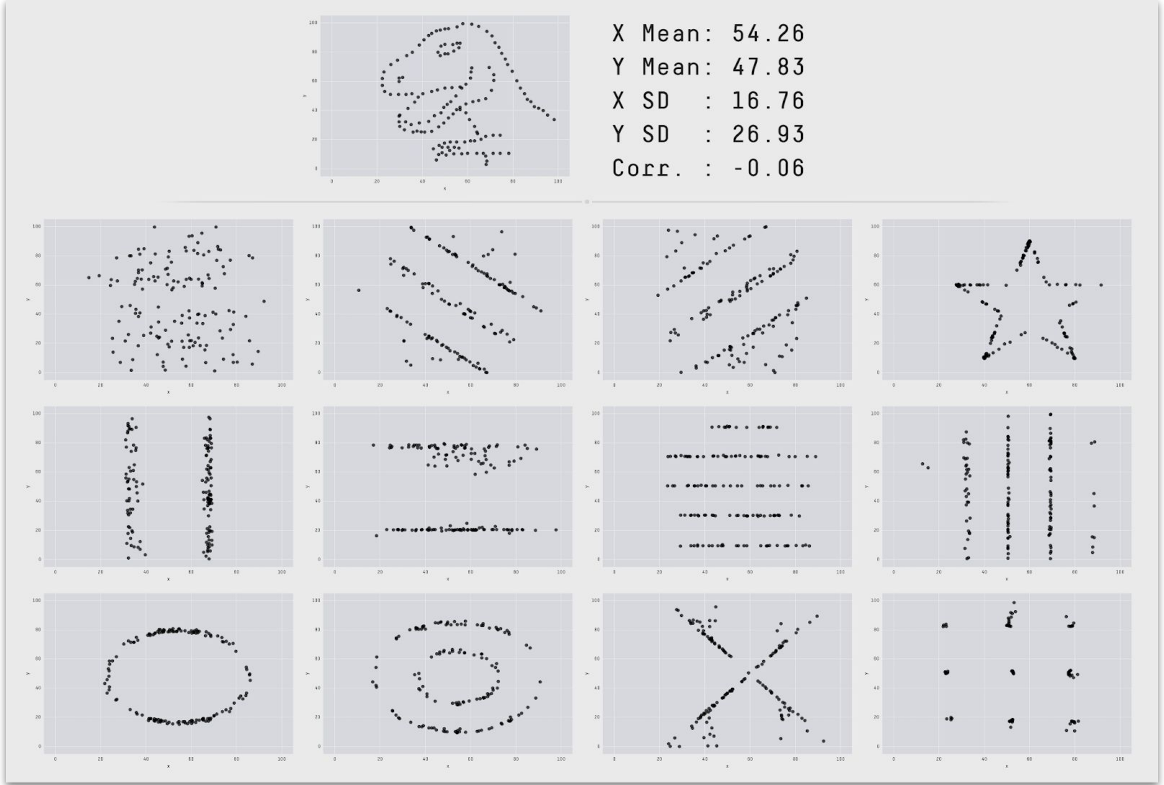
Anscombe's quartet

In the 1970 the statistician Francis Anscombe created 4 dataset that had almost identical descriptive statistics, but very different distributions and visual representations. They had the same:

- Means of x and y
- Variance of x and y
- Correlation between x and y
- Same linear regression and R2

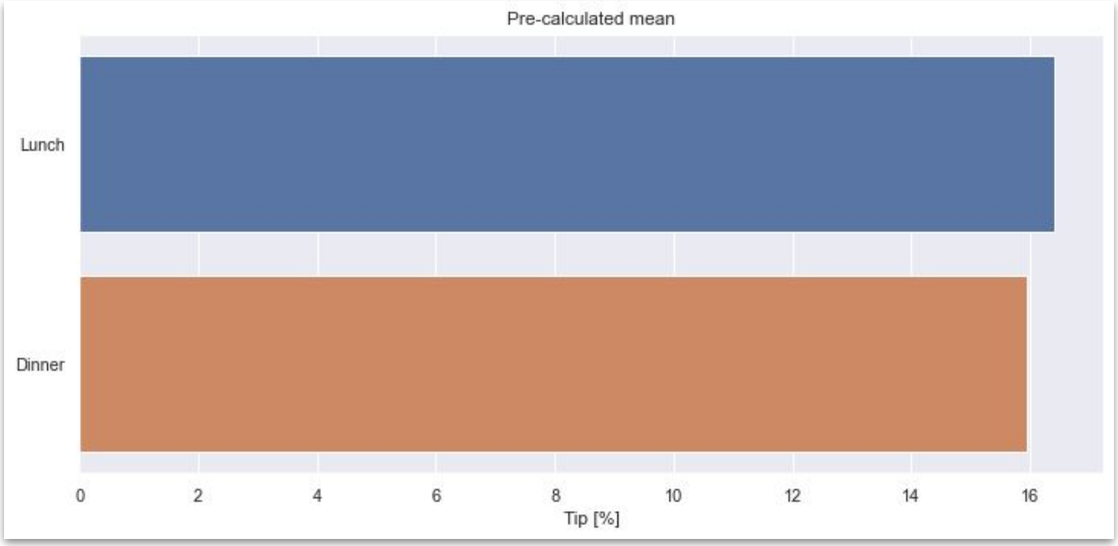


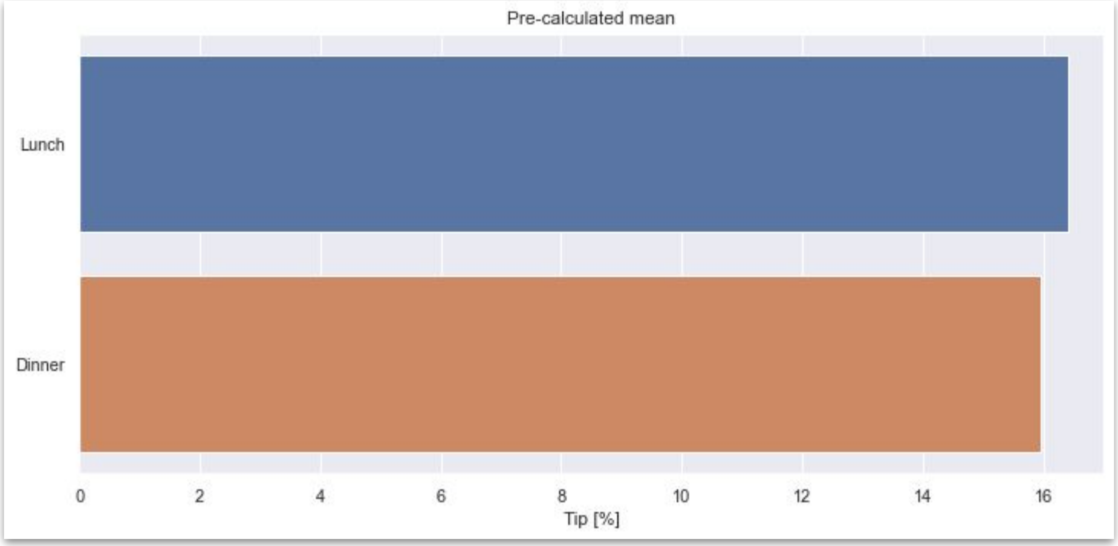
Anscombe's quartet
https://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg#filelinks

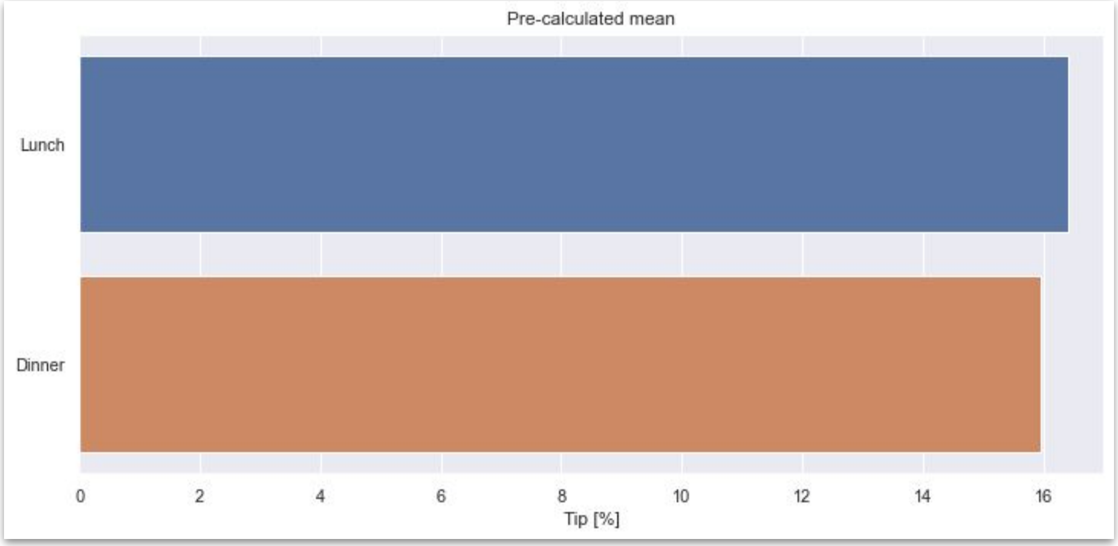


The **Datasaurus Dozen** (2016)
https://www.autodesk.com/research/publications/same-stats-different-graphs

Highlight the change - not the plots





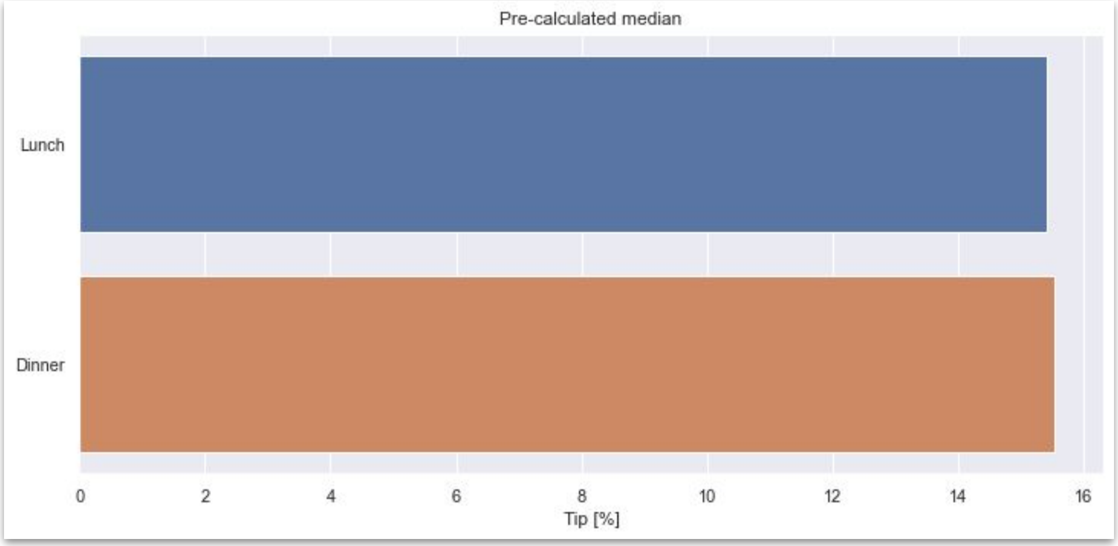


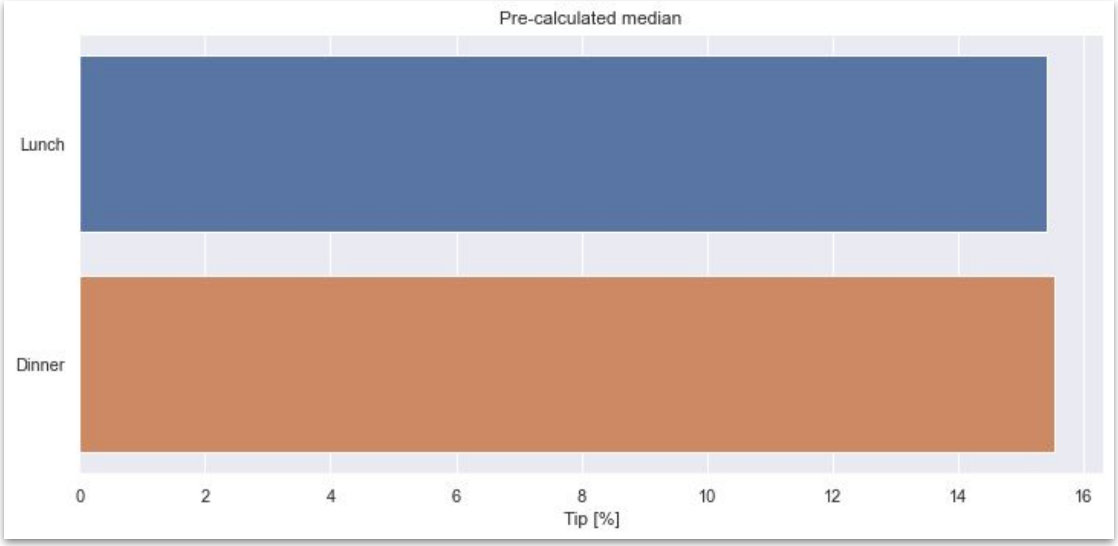
Only the change should be highlighted in the comparison. Avoid:

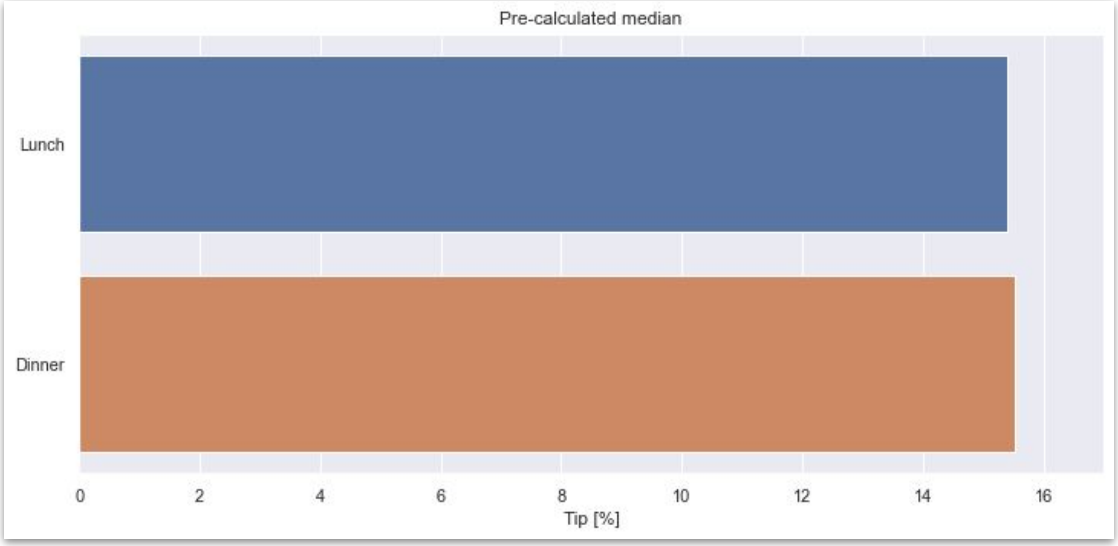
- Slight misalignment of plots
- Scales moving back and forth

Guide the eyes towards the change and not the surroundings.

Highlight the change - not the plots







Only the change should be highlighted in the comparison. Avoid:

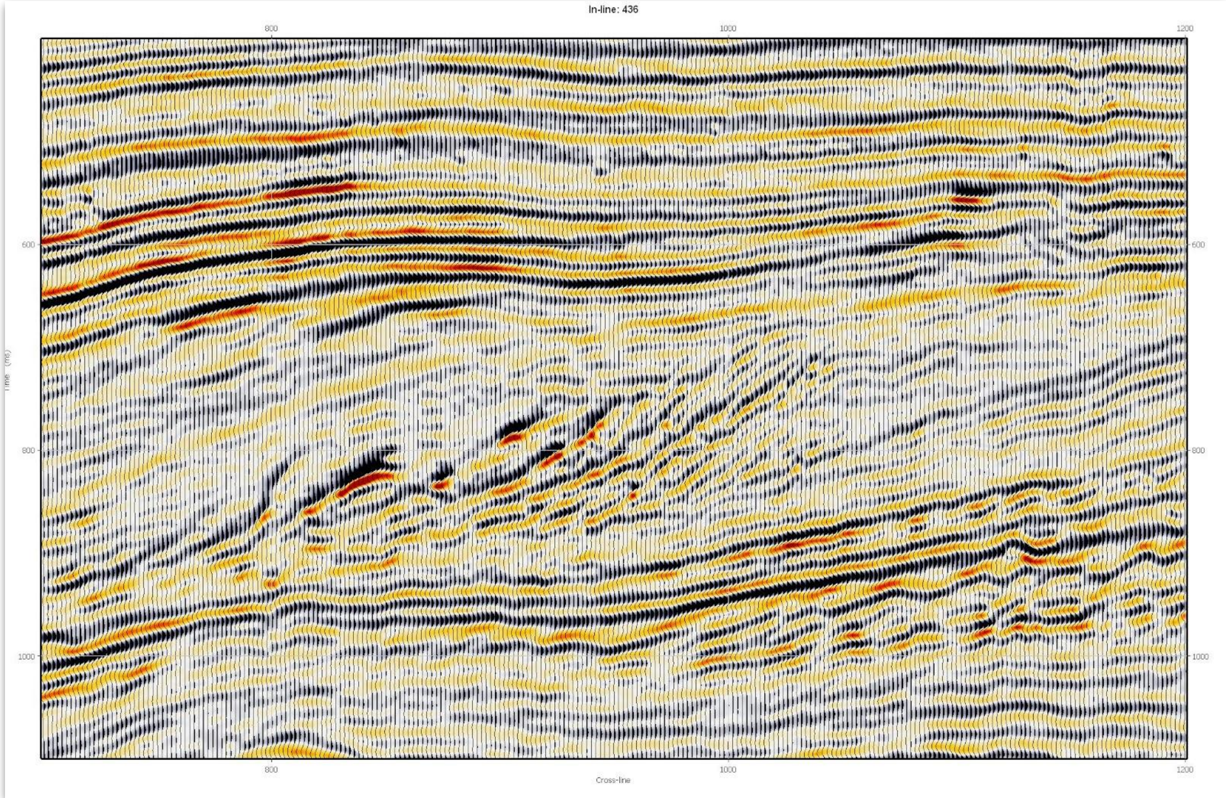
- Slight misalignment of plots
- Scales moving back and forth

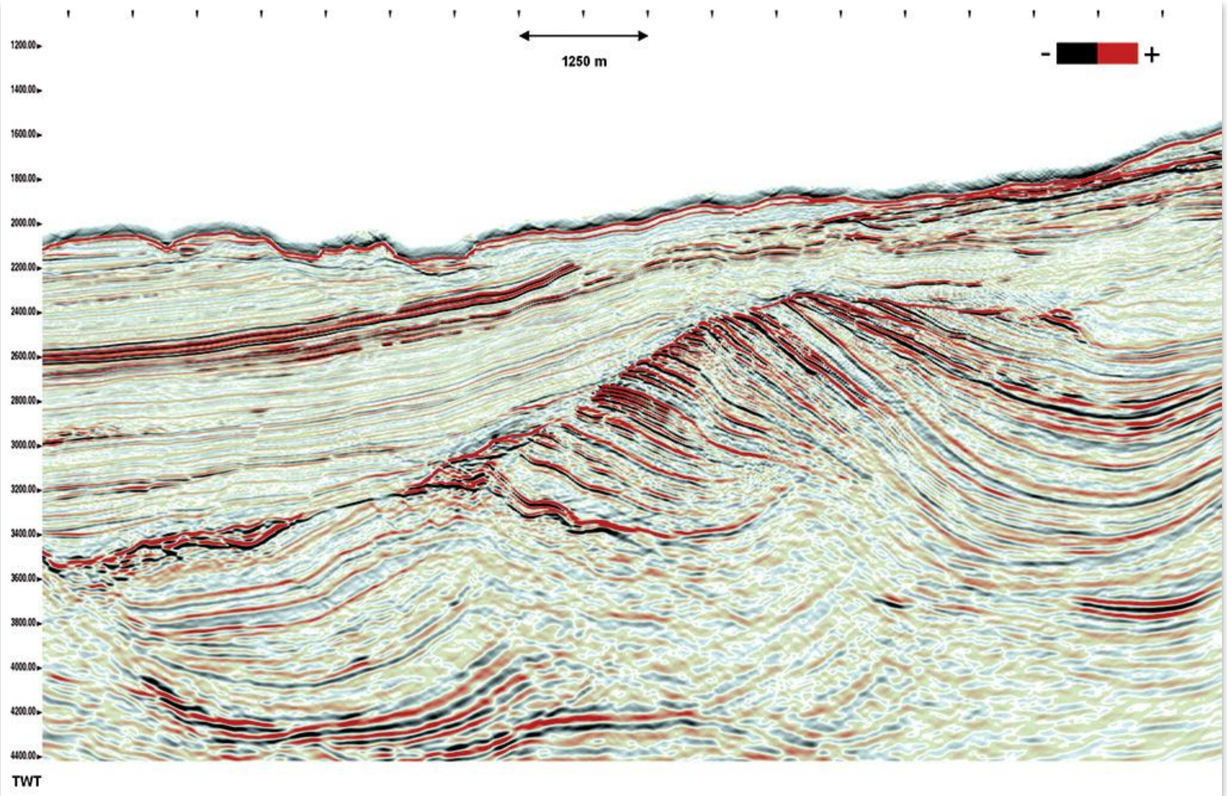
Guide the eyes towards the change and not the surroundings.

Highlight the change - not the plots

Bar plots are simple and can perhaps be over explained, but in complex such as seismic sections this is crucial.

- Left: Seismic section from the free [Project F3 Demo 2020]() seismic dataset plotted using the free [OpendTect]()
- Right: Plots from the [wiki of the Society of Exploration Geophysicists SEG]()

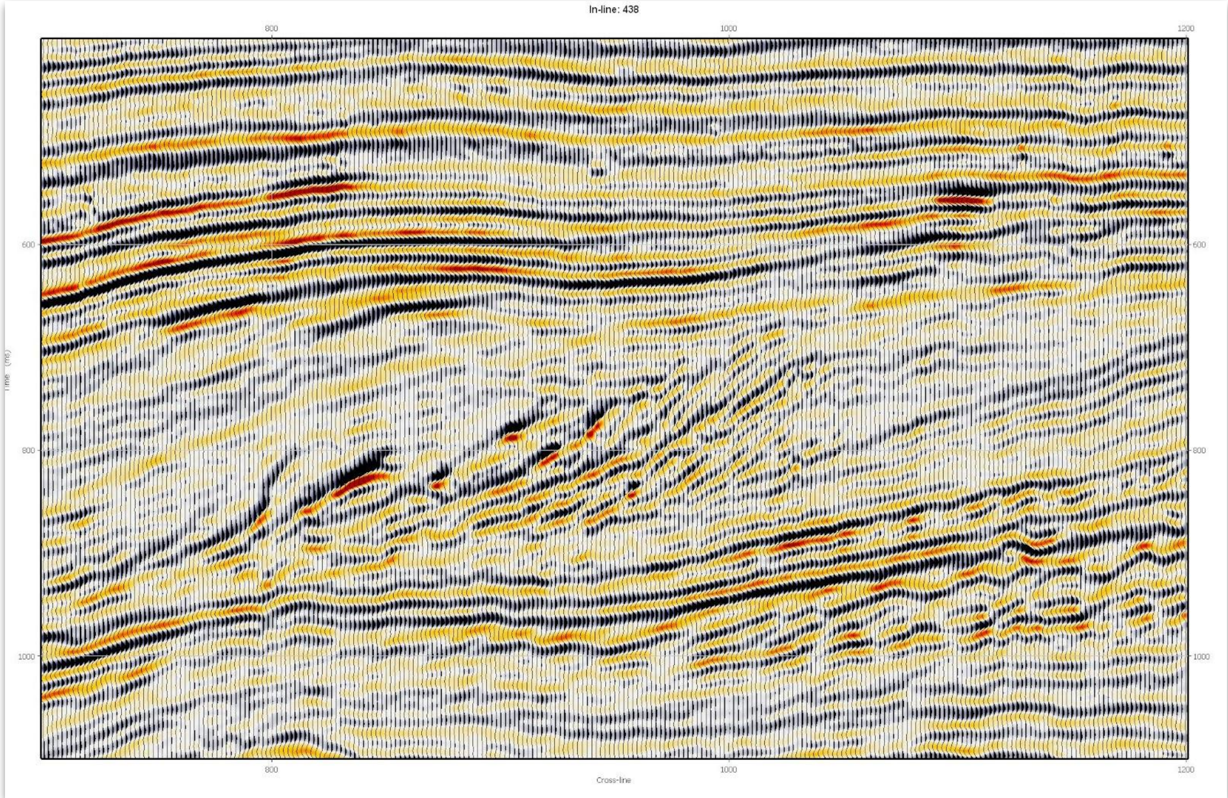


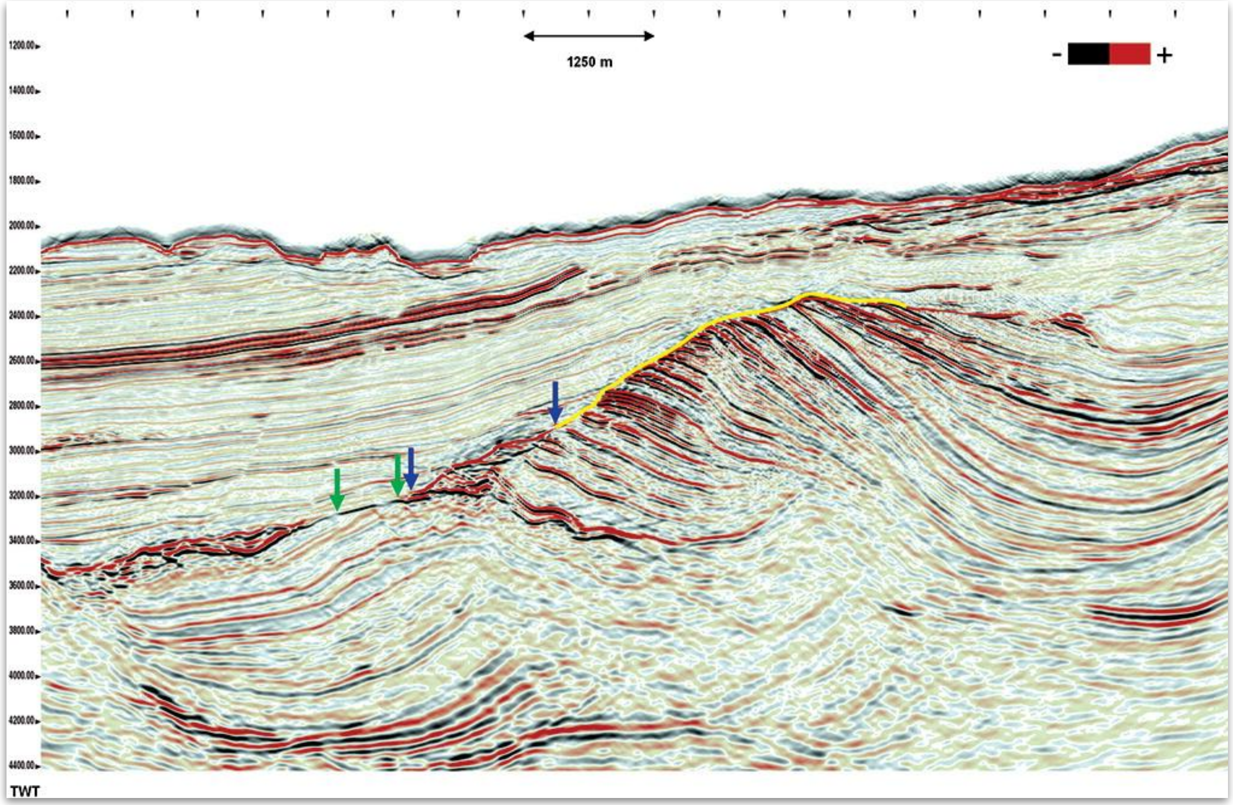


Highlight the change - not the plots

Bar plots are simple and can perhaps be over explained, but in complex such as seismic sections this is crucial.

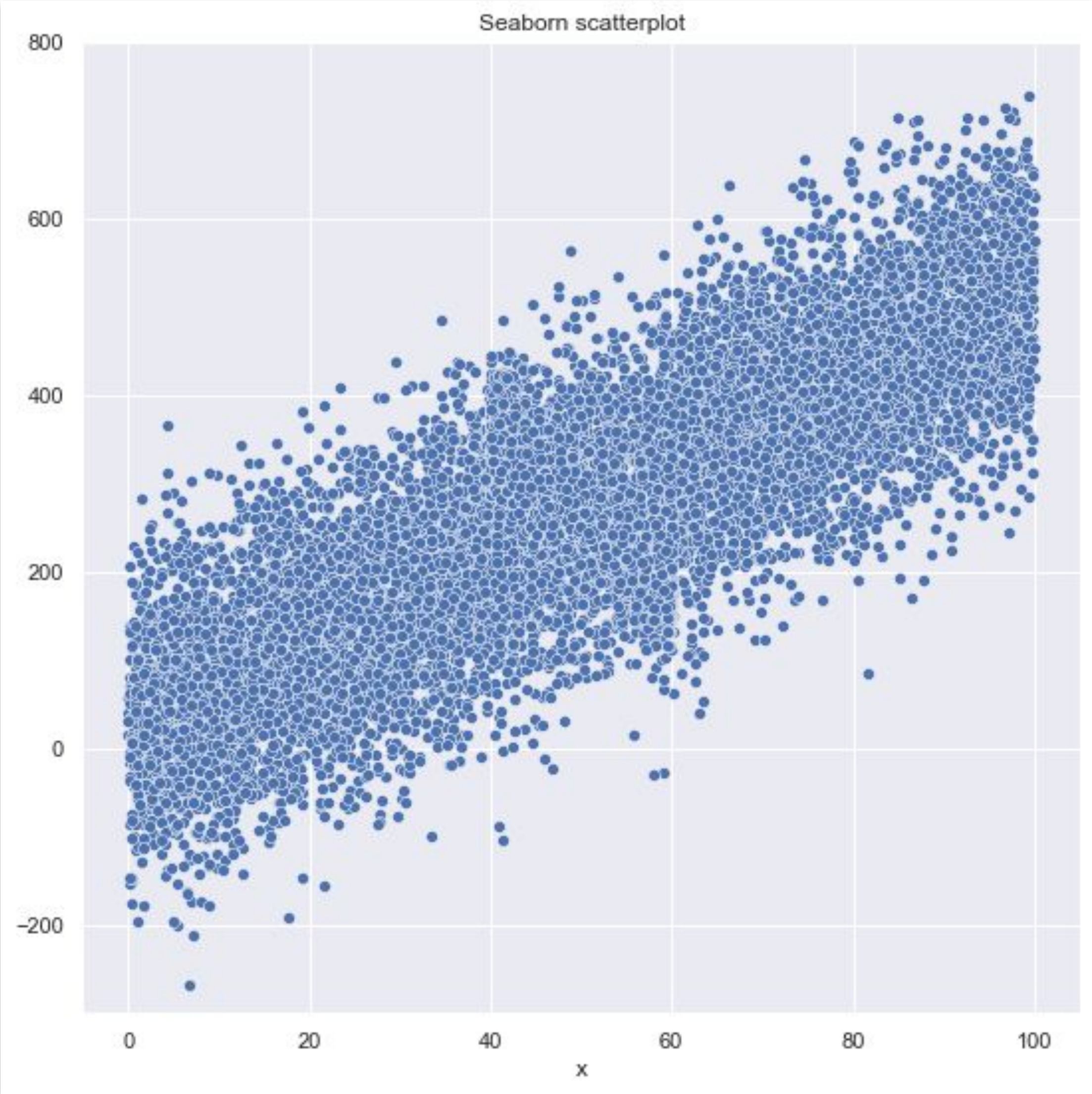
- Left: Seismic section from the free [Project F3 Demo 2020]() seismic dataset plotted using the free [OpendTect]()
- Right: Plots from the [wiki of the Society of Exploration Geophysicists SEG]() are not pixel perfect and it is hard tell what the difference is between the images beside the colored line as everything moves around.





Scatterplots can be deceiving

- 13.000 points
- Seemingly simple conclusion that the data follows a linear relationship with some noise



Scatterplots can be deceiving

- 13.000 points
- Seemingly simple conclusion that the data follows a linear relationship with some noise
- Secondary hidden relationship in data only visible using kde or histogram plots, due to severe overplotting

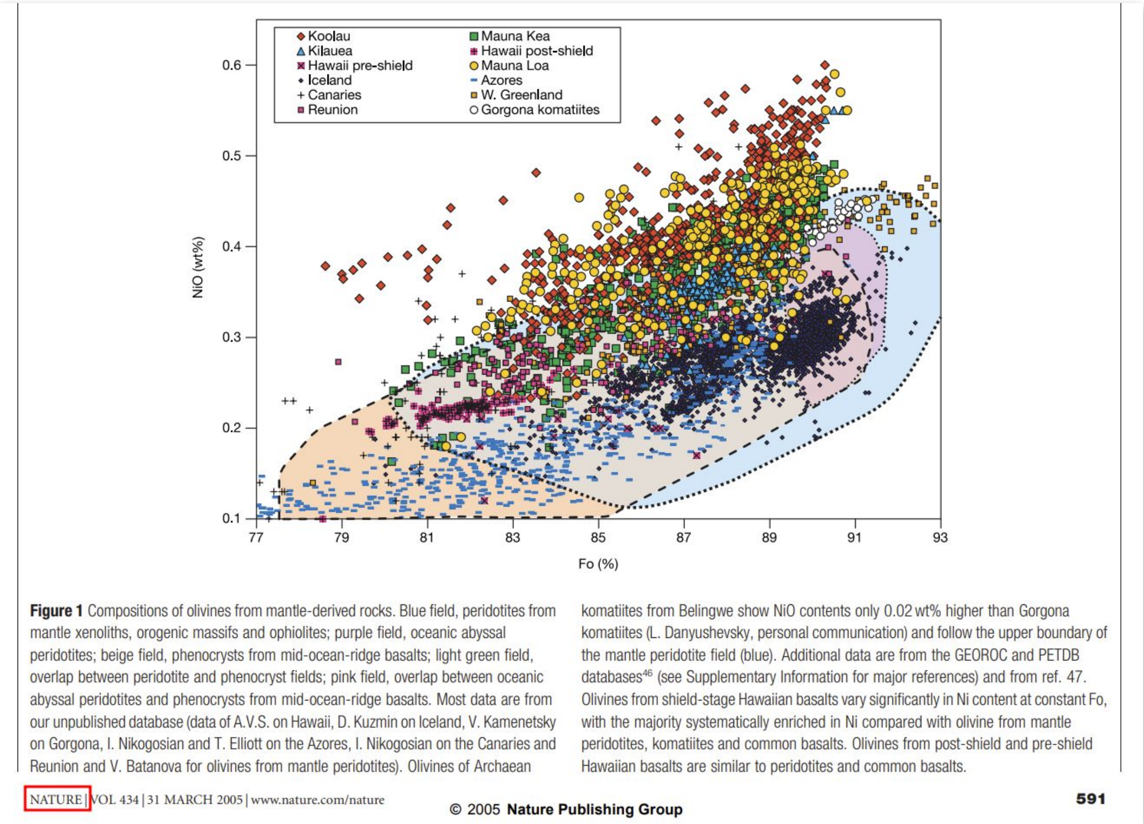
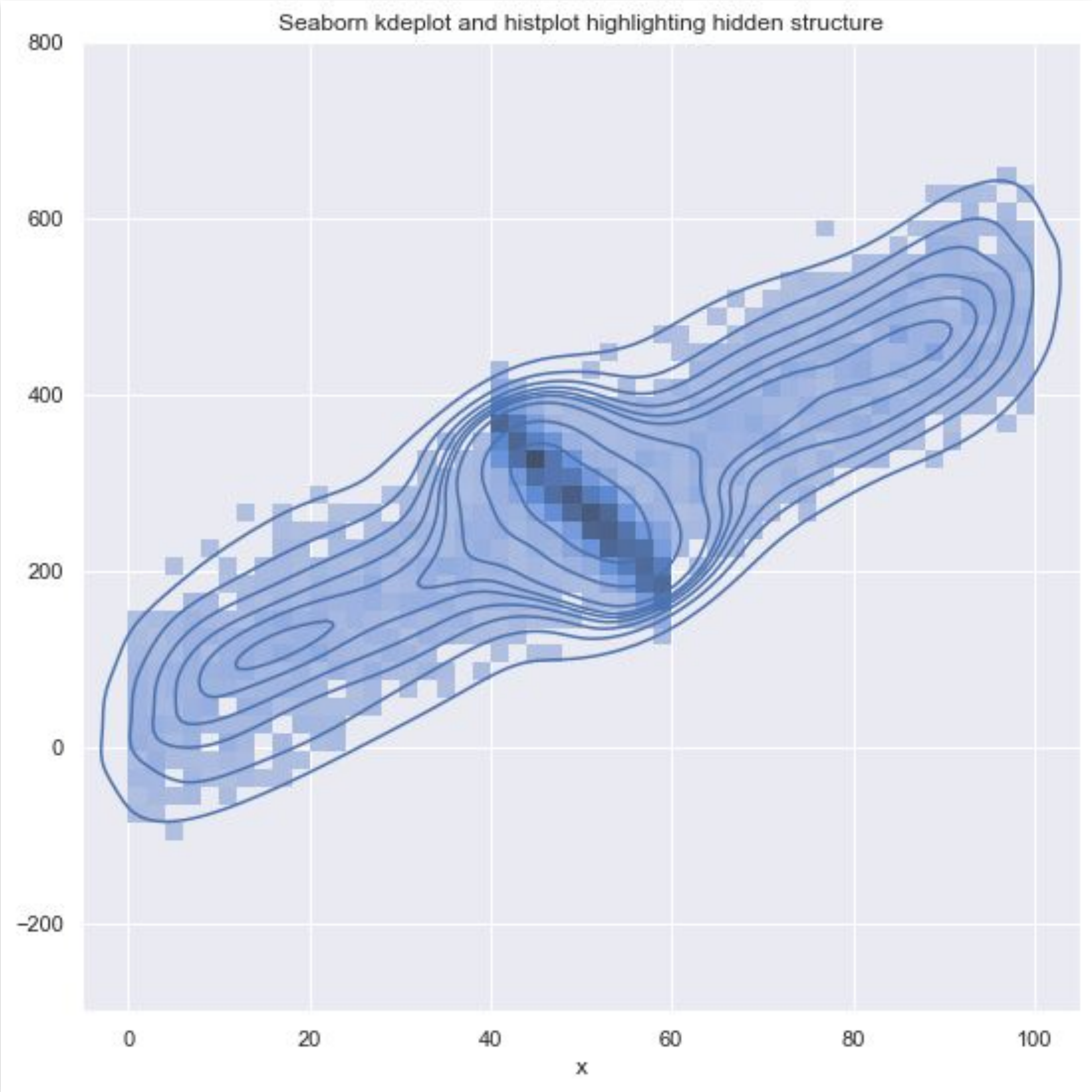


Figure 1 Compositions of olivines from mantle-derived rocks. Blue field, peridotites from mantle xenoliths, orogenic massifs and ophiolites; purple field, oceanic abyssal peridotites; beige field, phenocrysts from mid-ocean-ridge basalts; light green field, overlap between peridotite and phenocryst fields; pink field, overlap between oceanic abyssal peridotites and phenocrysts from mid-ocean-ridge basalts. Most data are from our unpublished database (data of A.V.S. on Hawaii, D. Kuzmin on Iceland, V. Kamenetsky on Gorgona, I. Nikogosian and T. Elliott on the Azores, I. Nikogosian on the Canaries and Reunion and V. Batanova for olivines from mantle peridotites). Olivines of Archaean komatiites from Belingwe show NiO contents only 0.02 wt% higher than Gorgona komatiites (L. Danyushevsky, personal communication) and follow the upper boundary of the mantle peridotite field (blue). Additional data are from the GEOROC and PETDB databases[46] (see Supplementary Information for major references) and from ref. 47. Olivines from shield-stage Hawaiian basalts vary significantly in Ni content at constant Fo, with the majority systematically enriched in Ni compared with olivine from mantle peridotites, komatiites and common basalts. Olivines from post-shield and pre-shield Hawaiian basalts are similar to peridotites and common basalts.

NATURE | VOL 434 | 31 MARCH 2005 | www.nature.com/nature © 2005 Nature Publishing Group 591



Sobolev, A., Hofmann, A., Sobolev, S. et al. An olivine-free mantle source of Hawaiian shield basalts. Nature 434, 590–597 (2005). https://doi.org/10.1038/nature03411

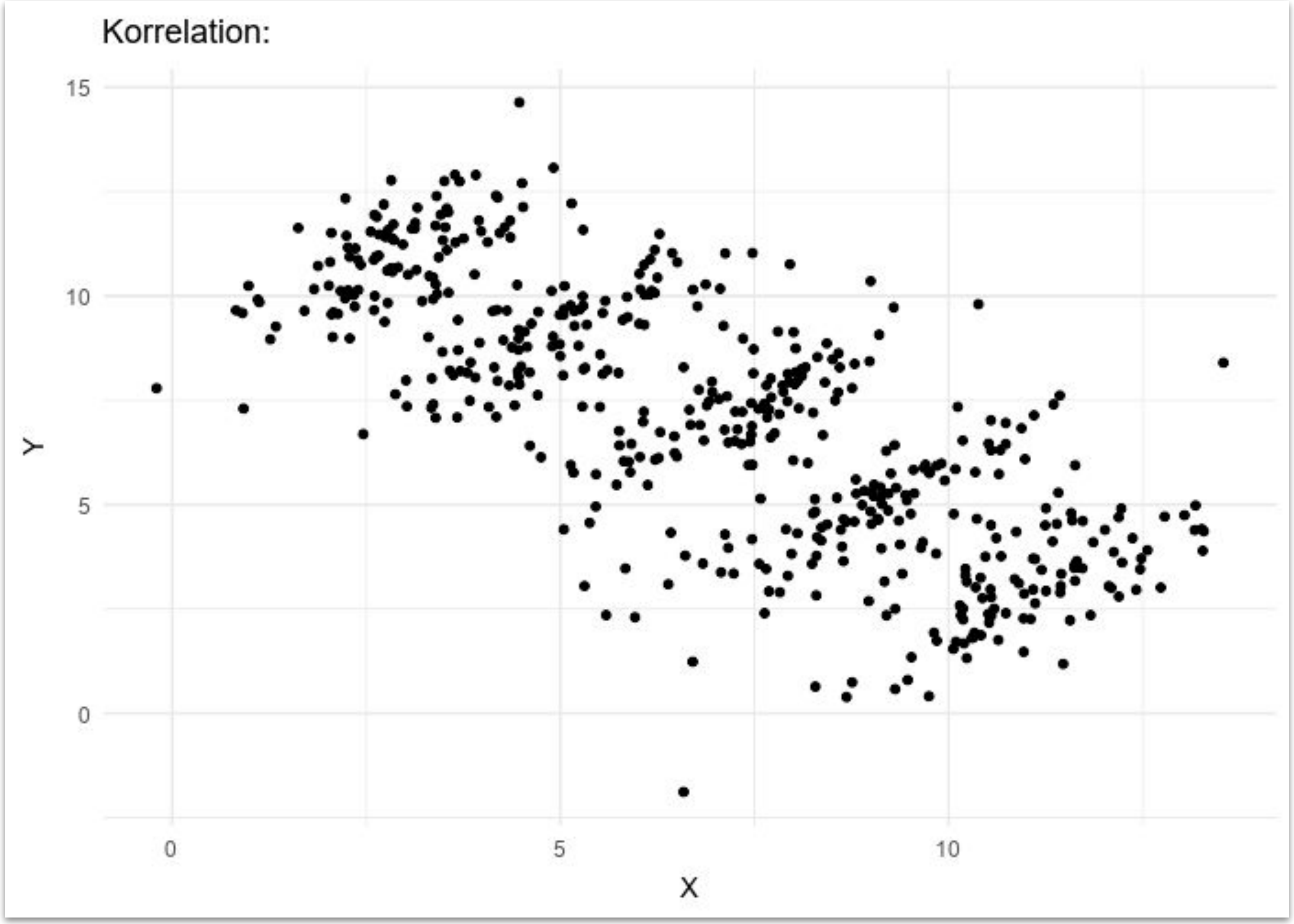
Simpson's paradox

First described by Edward H. Simpson in 1951.

Trends found in subsets of data disappears or reverses when the whole dataset is considered.

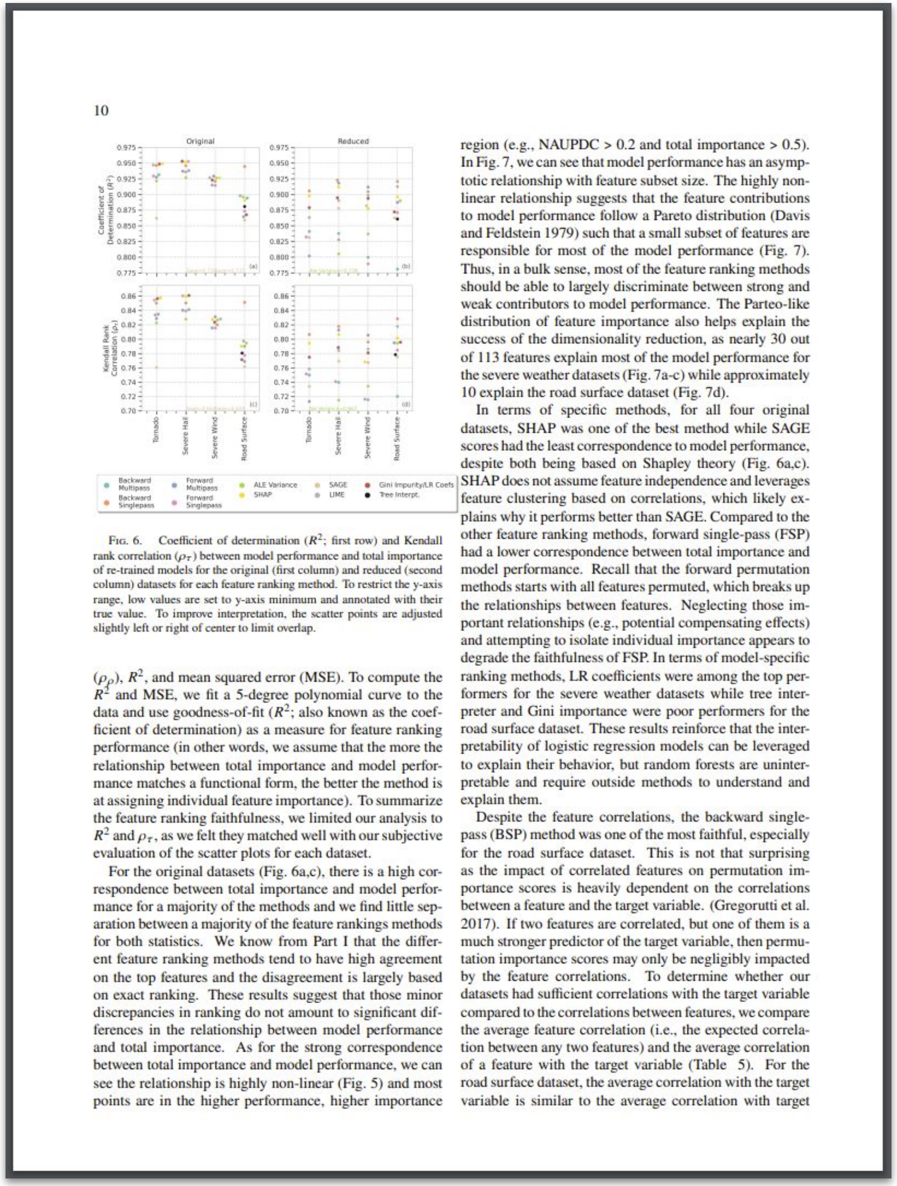
Most common quoted example is from Berkeley in 1973. Statistics from the admission data found that more men were admitted than women when considering all departments. When inspecting the departments individually it was found that women had higher admittance percentages.

Men sought less competitive departments (engineering) whereas women sought more competitive departments (english).



https://commons.wikimedia.org/wiki/File:Simpsons_paradox_-_animation.gif

Size of the figure



10

region (e.g., NAUPDC > 0.2 and total importance > 0.5). In Fig. 7, we can see that model performance has an asymptotic relationship with feature subset size. The highly non-linear relationship suggests that the feature contributions to model performance follow a Pareto distribution (Davis and Feldstein 1979) such that a small subset of features are responsible for most of the model performance (Fig. 7). Thus, in a bulk sense, most of the feature ranking methods should be able to largely discriminate between strong and weak contributors to model performance. The Parteo-like distribution of feature importance also helps explain the success of the dimensionality reduction, as nearly 30 out of 113 features explain most of the model performance for the severe weather datasets (Fig. 7a-c) while approximately 10 explain the road surface dataset (Fig. 7d).

In terms of specific methods, for all four original datasets, SHAP was one of the best method while SAGE scores had the least correspondence to model performance, despite both being based on Shapley theory (Fig. 6a,c). SHAP does not assume feature independence and leverages feature clustering based on correlations, which likely explains why it performs better than SAGE. Compared to the other feature ranking methods, forward single-pass (FSP) had a lower correspondence between total importance and model performance. Recall that the forward permutation methods starts with all features permuted, which breaks up the relationships between features. Neglecting those important relationships (e.g., potential compensating effects) and attempting to isolate individual importance appears to degrade the faithfulness of FSP. In terms of model-specific ranking methods, LR coefficients were among the top performers for the severe weather datasets while tree interpreter and Gini importance were poor performers for the road surface dataset. These results reinforce that the interpretability of logistic regression models can be leveraged to explain their behavior, but random forests are uninterpretable and require outside methods to understand and explain them.

Despite the feature correlations, the backward single-pass (BSP) method was one of the most faithful, especially for the road surface dataset. This is not that surprising as the impact of correlated features on permutation importance scores is heavily dependent on the correlations between a feature and the target variable. (Gregorutti et al. 2017). If two features are correlated, but one of them is a much stronger predictor of the target variable, then permutation importance scores may only be negligibly impacted by the feature correlations. To determine whether our datasets had sufficient correlations with the target variable compared to the correlations between features, we compare the average feature correlation (i.e., the expected correlation between any two features) and the average correlation of a feature with the target variable (Table 5). For the road surface dataset, the average correlation with the target variable is similar to the average correlation with target

Fig. 6. Coefficient of determination (R^2; first row) and Kendall rank correlation (ρ_τ) between model performance and total importance of re-trained models for the original (first column) and reduced (second column) datasets for each feature ranking method. To restrict the y-axis range, low values are set to y-axis minimum and annotated with their true value. To improve interpretation, the scatter points are adjusted slightly left or right of center to limit overlap.

(ρ_ρ), R^2, and mean squared error (MSE). To compute the R^2 and MSE, we fit a 5-degree polynomial curve to the data and use goodness-of-fit (R^2; also known as the coefficient of determination) as a measure for feature ranking performance (in other words, we assume that the more the relationship between total importance and model performance matches a functional form, the better the method is at assigning individual feature importance). To summarize the feature ranking faithfulness, we limited our analysis to R^2 and ρ_τ, as we felt they matched well with our subjective evaluation of the scatter plots for each dataset.

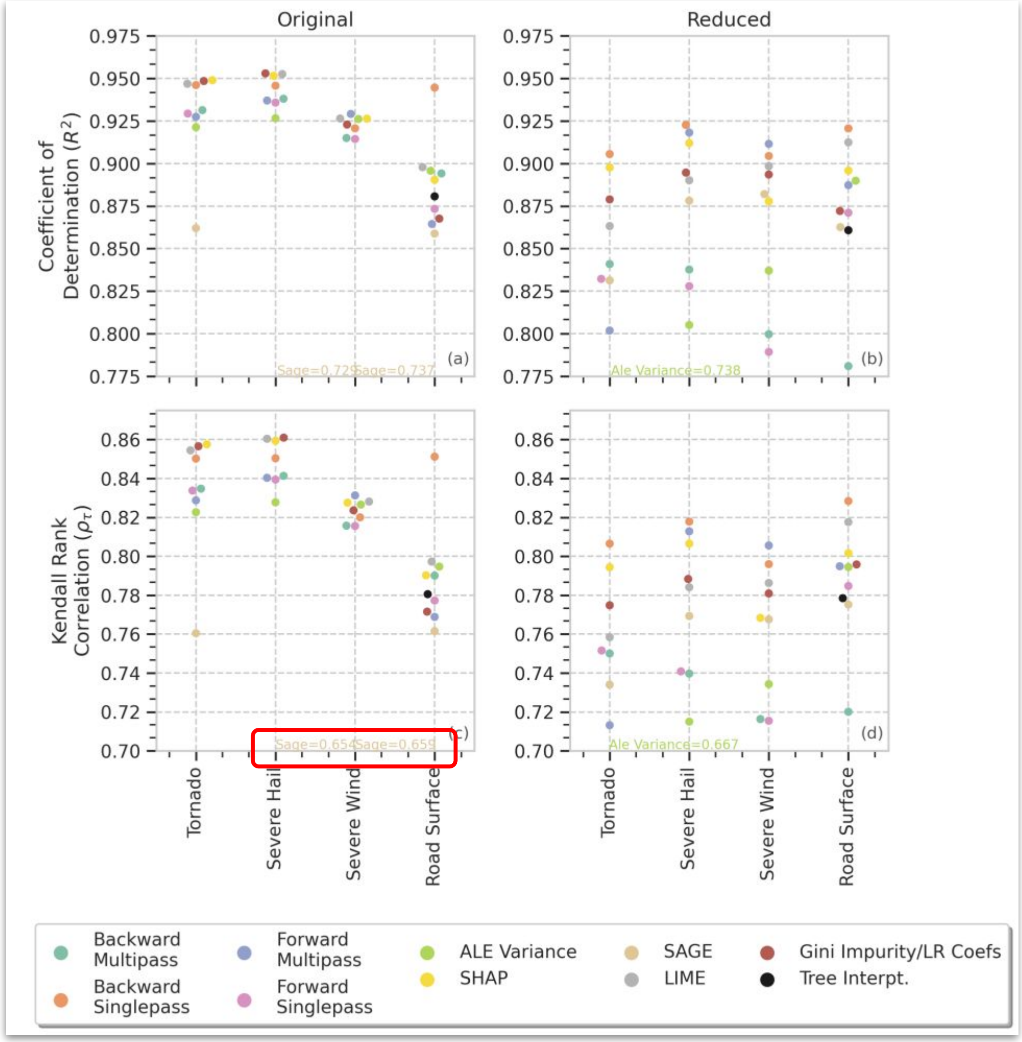
For the original datasets (Fig. 6a,c), there is a high correspondence between total importance and model performance for a majority of the methods and we find little separation between a majority of the feature rankings methods for both statistics. We know from Part I that the different feature ranking methods tend to have high agreement on the top features and the disagreement is largely based on exact ranking. These results suggest that those minor discrepancies in ranking do not amount to significant differences in the relationship between model performance and total importance. As for the strong correspondence between total importance and model performance, we can see the relationship is highly non-linear (Fig. 5) and most points are in the higher performance, higher importance

If the figure is important enough be shown it should be large enough to be read.

Article from 18th of November 2022 on arXiv.

Probably a great article, but practically unreadable figure especially on a e-reader or in print.

arXiv:2211.10378v1 [cs.LG] 18 Nov 2022
https://arxiv.org/pdf/2211.10378.pdf

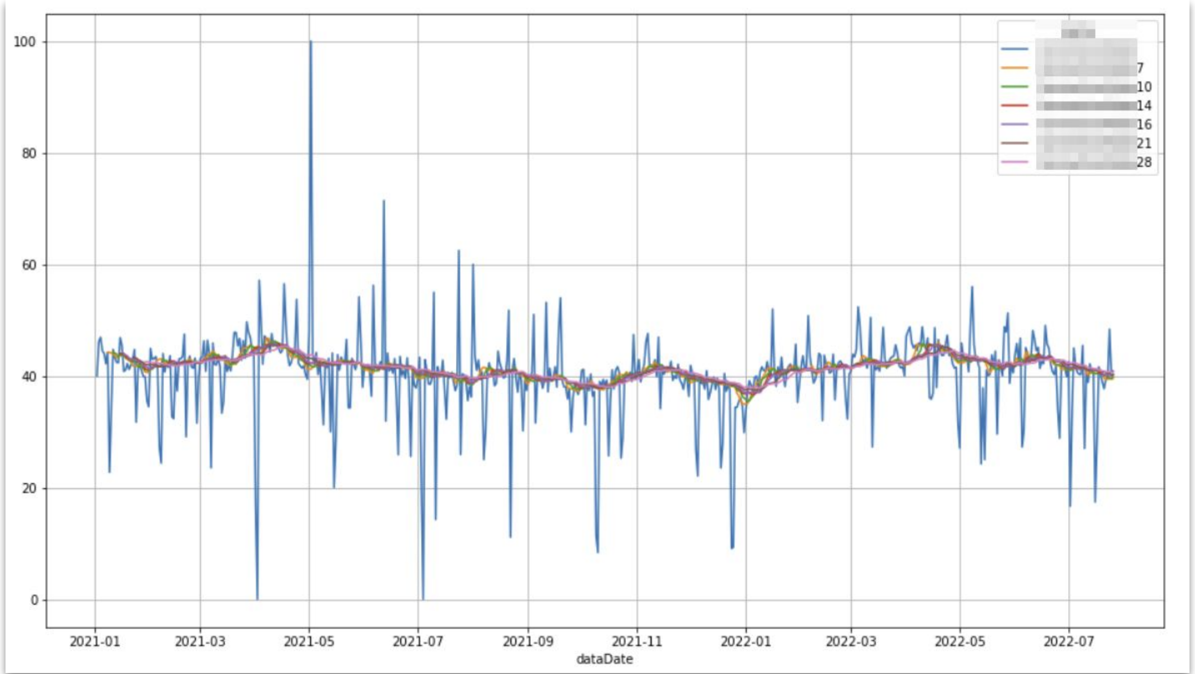


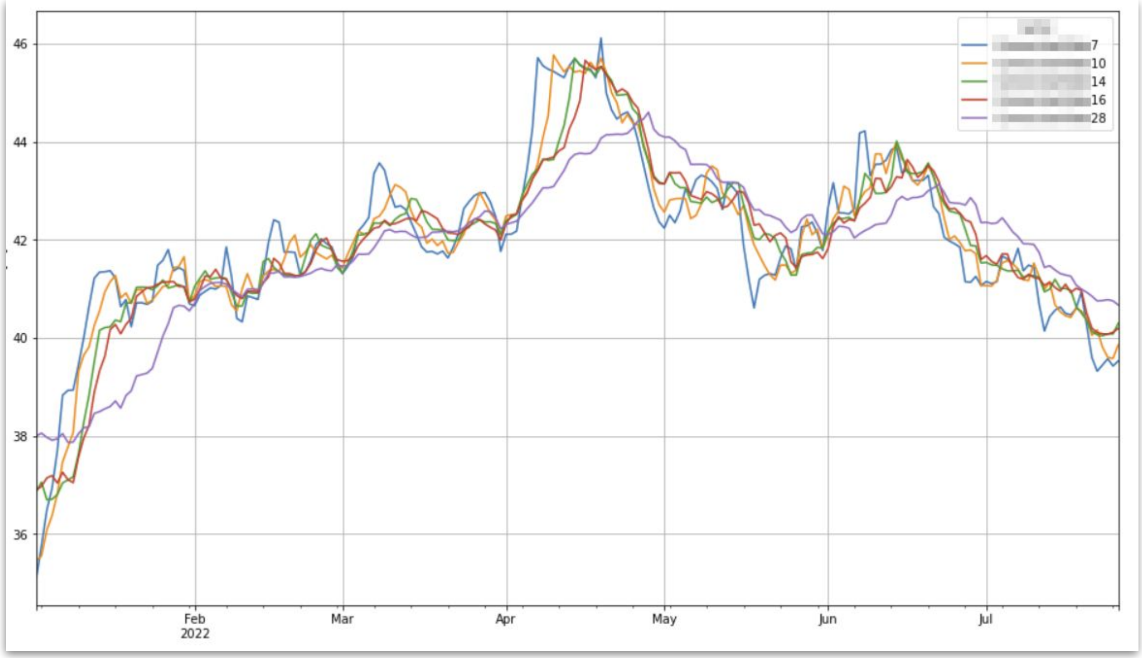
Time series data - resampling

The raw signal is quite nosy and hard to interpret. Daily values, but volatile values in the weekend.

What is the correct resampling to tell the story of the data?

- Removing weekends?
- Running average length?
- Mean or median?
- Centered average?
- Effect of lag?
- Window function?





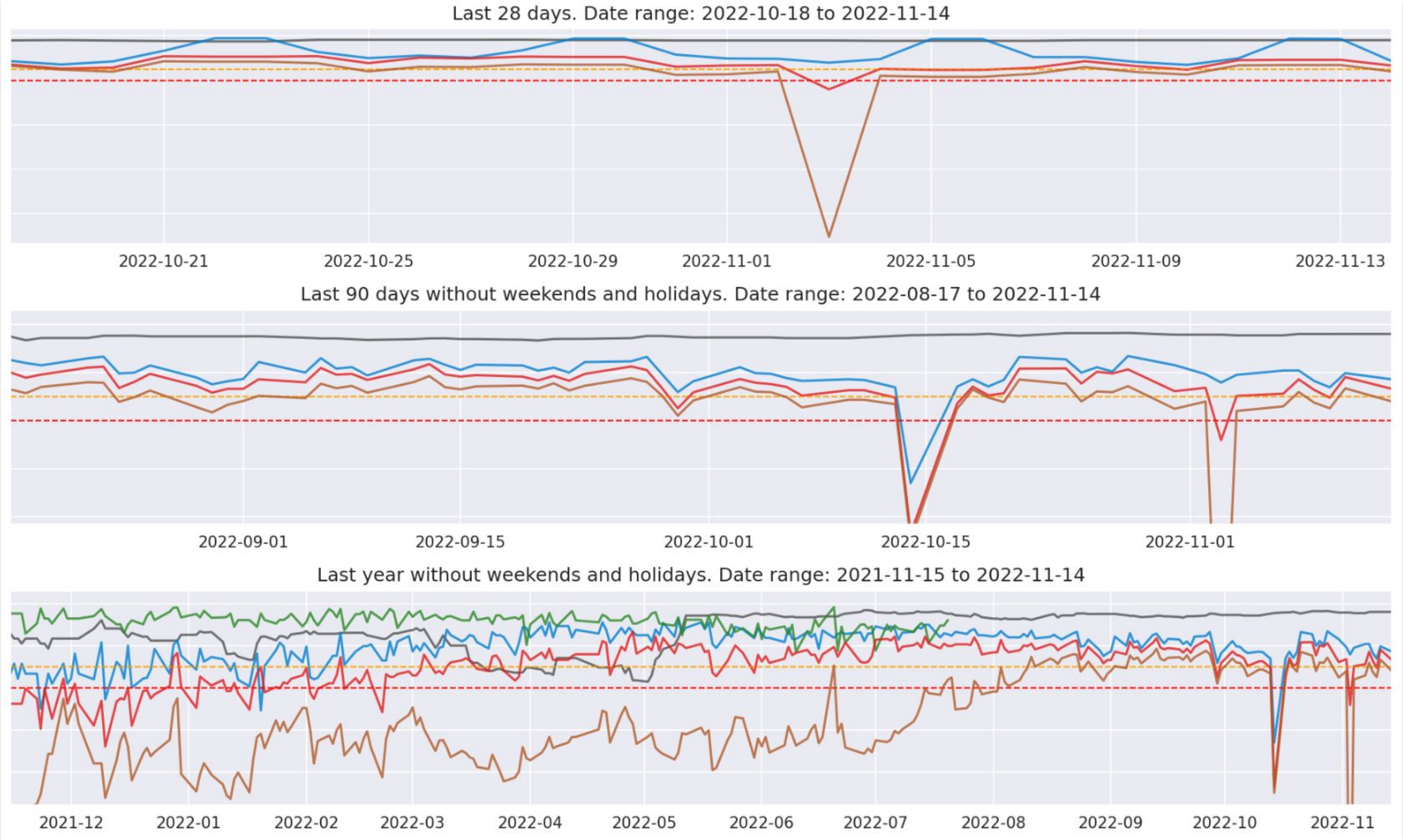
Time series data - multiple timescales

In Connected Cars we monitor more than 1000 data quality checks per car per day. We have a lot of different systems to find anomalies.

In this plot we track some values over time and show the aggregates over three time scales in the same plot:

- The last 28 days
- The last 90 days
- The last year

The idea behind this is that some slow moving developments are not easily detected in short timescales and some fast moving developments are truncated in slow moving timescales, so in order to understand the full development we needed evaluate multiple timescales simultaneously.



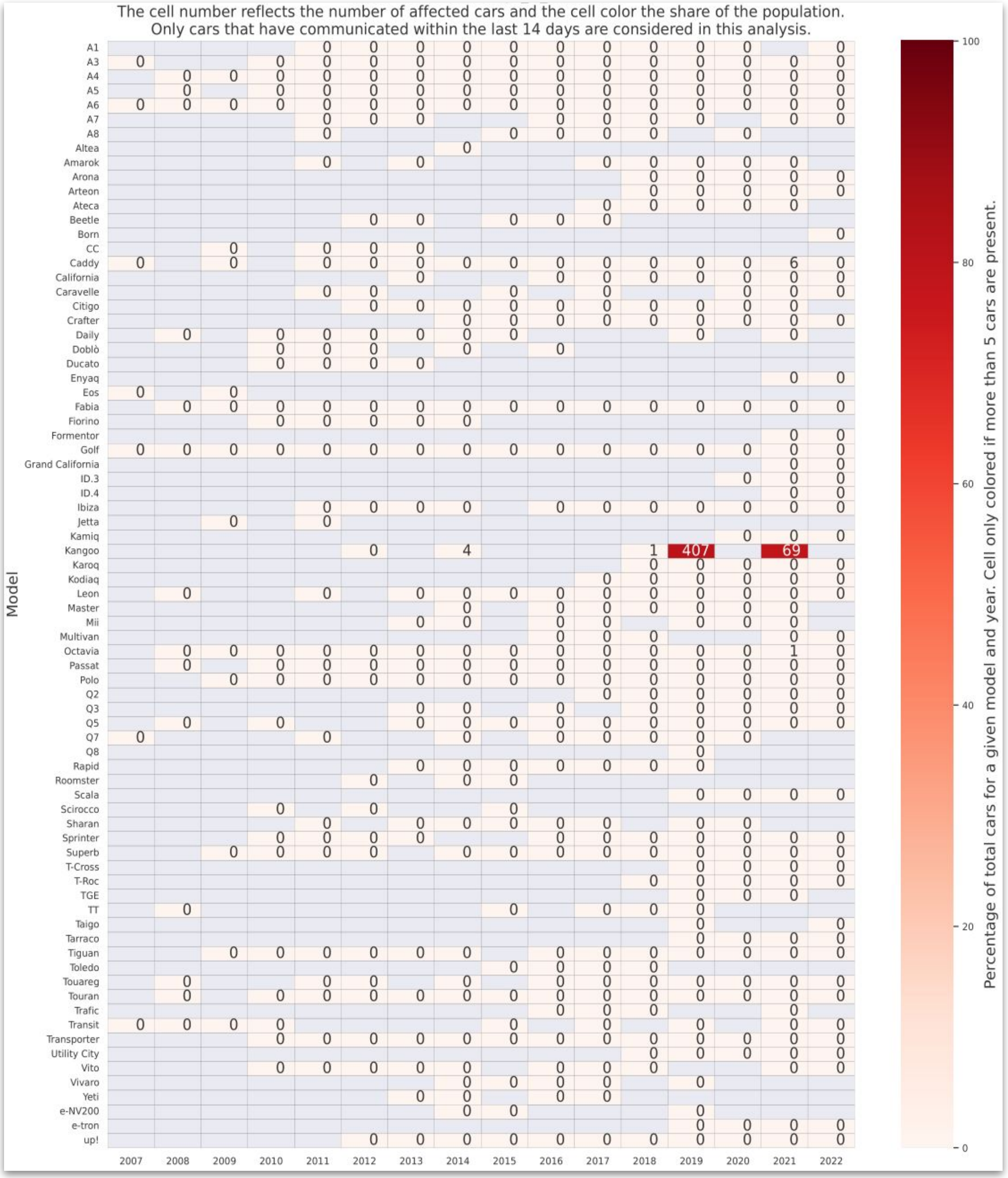
Find anomalies with a glance

Before I started working with data and visualization I was always puzzled by the blinking lights in control rooms. How could that be useful?

People are really good at spotting anomalies, a single red lamp on a wall of yellow lamps will stand out. A glance should be enough to know if something is wrong.

Interactive visualizations are great for viewing data and trends on different time series, using filters and for deep diving into potential issues, but a familiar static dashboard can make it easy to identify anomalies.

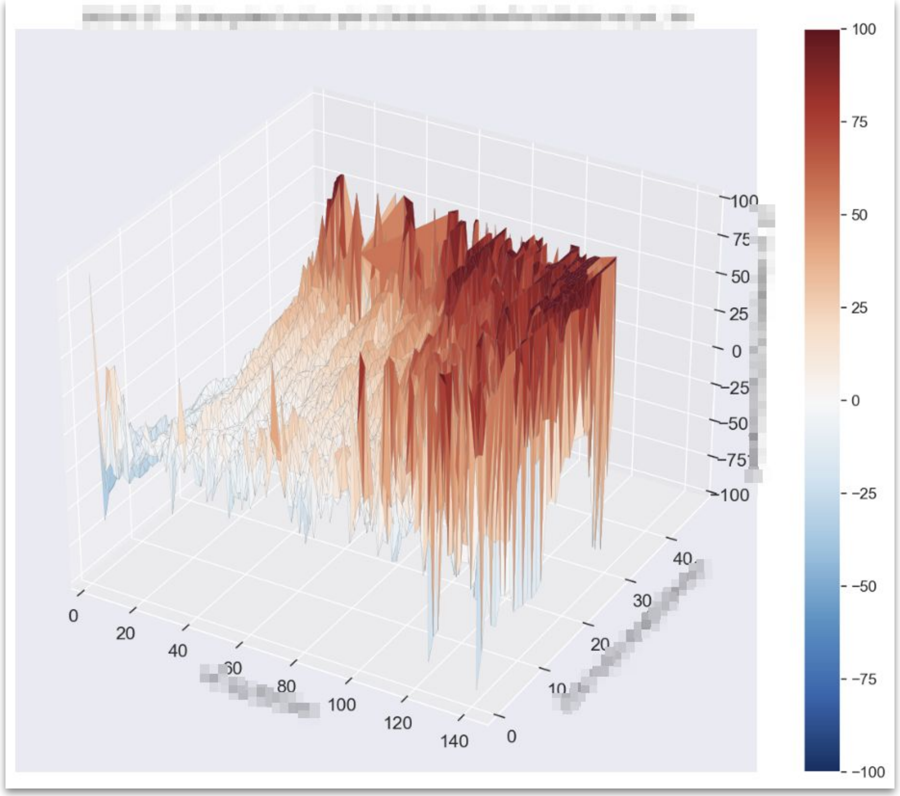
https://www.popularmechanics.com/technology/design/g20681640/control-rooms/

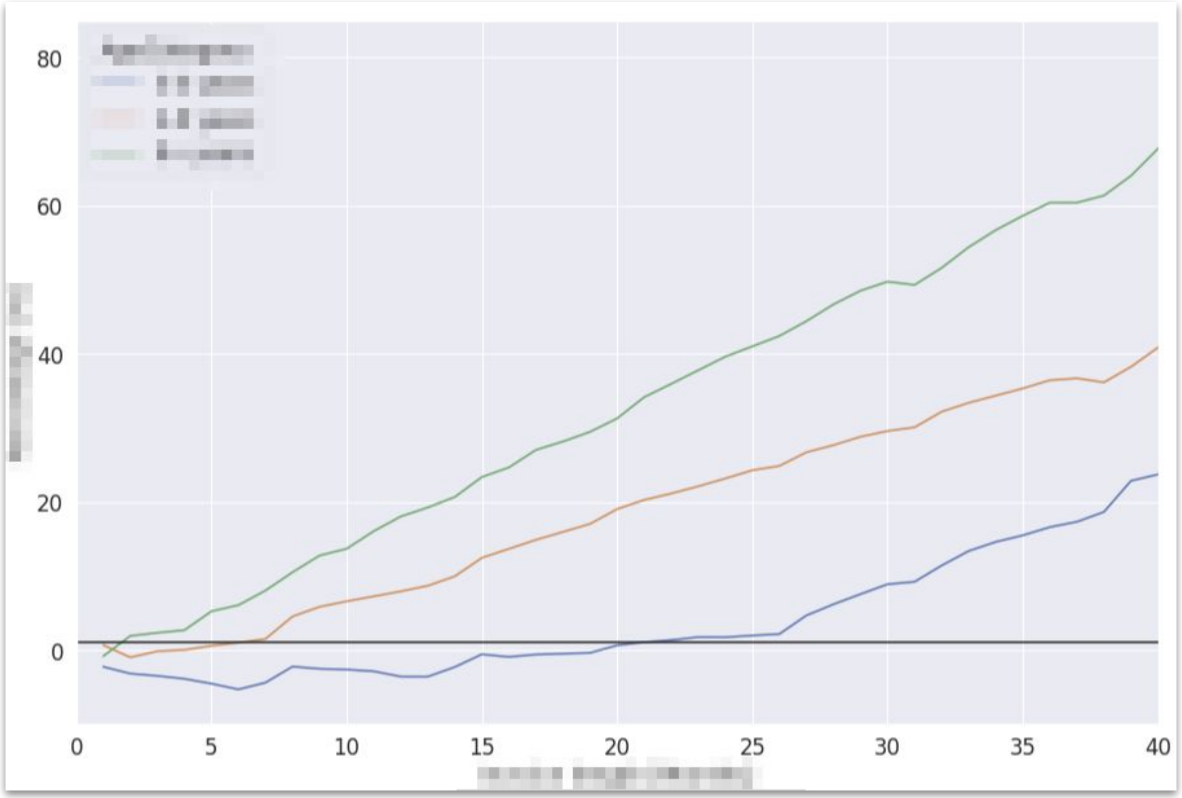


The right plot to the right people

Questions to be answered before deciding on the visualization:

- What is the technical abstraction level or background of the audience?
- What is the story that is trying to be told and to which audience?

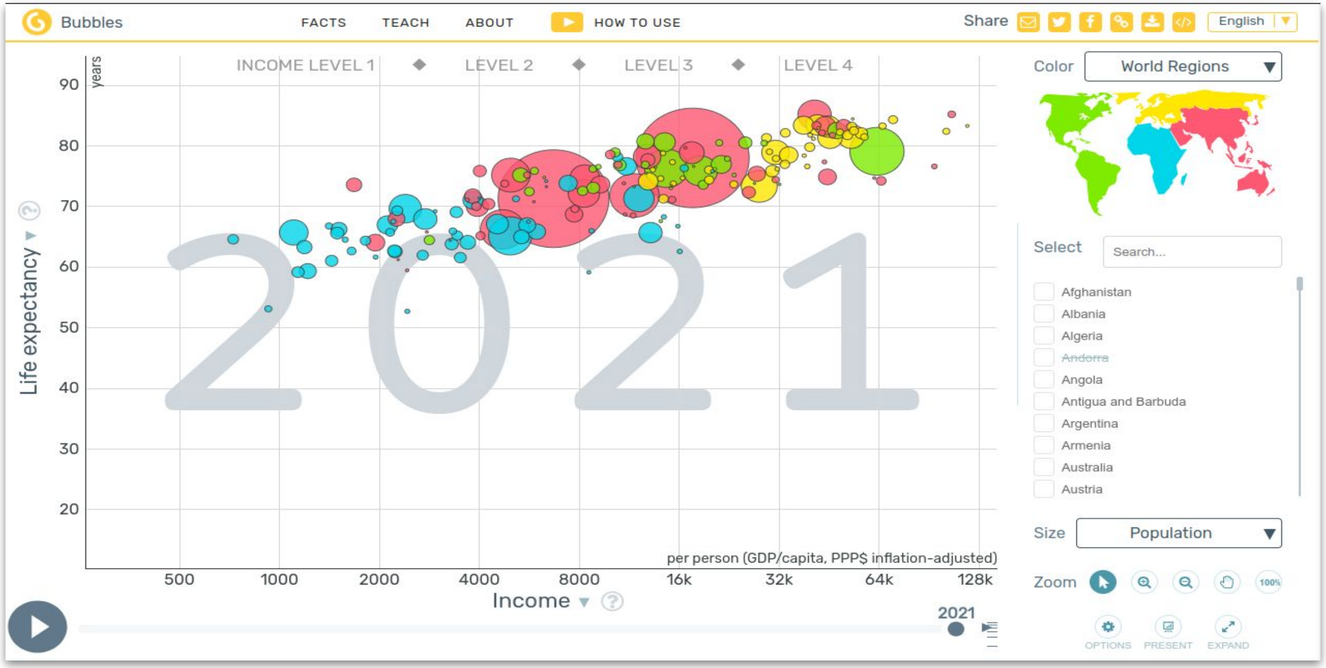




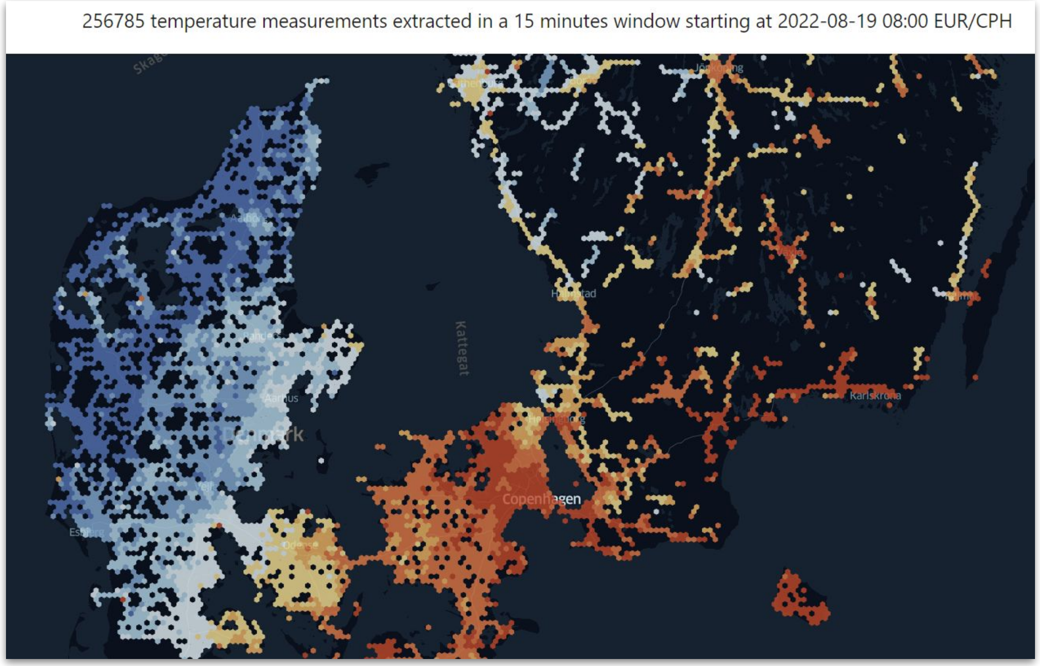
Example of two visualizations of the same data to different audiences

1. A technical plot intended to initiate a conversation about complexity of the data and how the data should be treated.
2. A simpler plot intended to convey the key business takeaways without going into too much detail.

Easily digestible visualizations

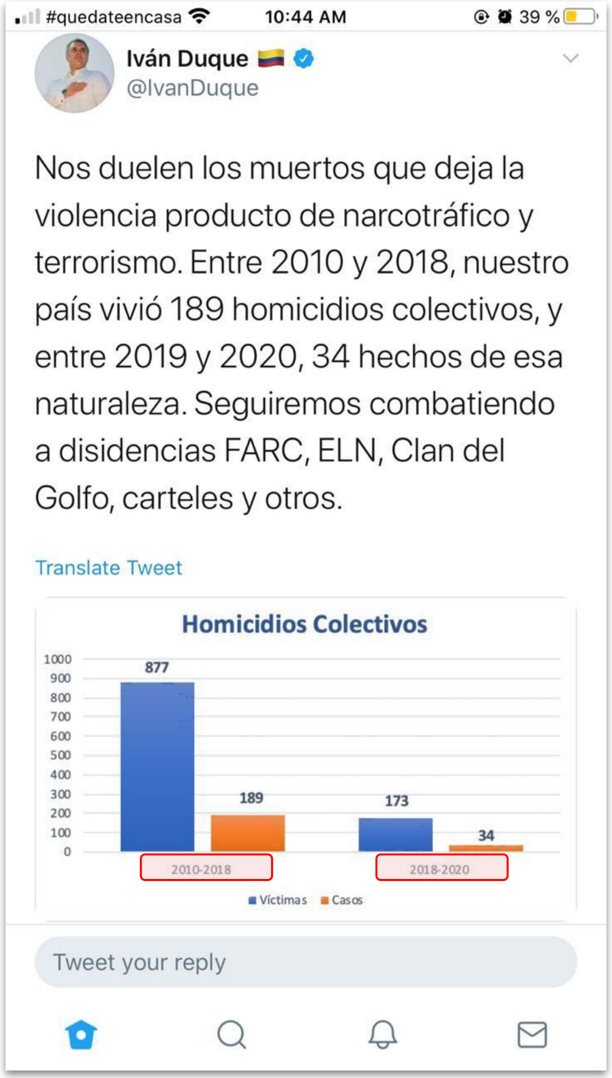


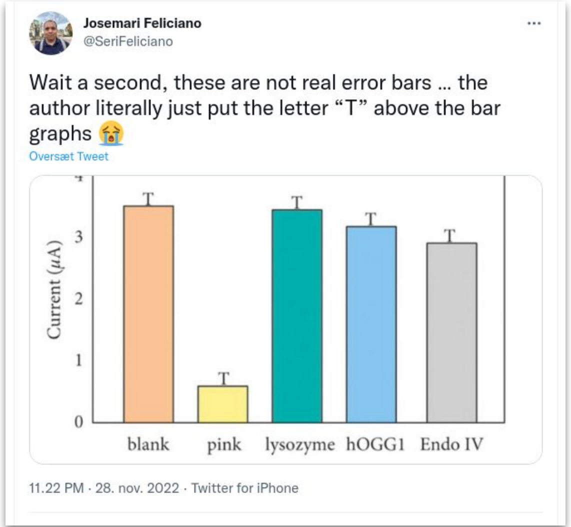
The gapminder visualization. Gapminder was founded in Stockholm on 25 February 2005 by Ola Rosling, Anna Rosling Rönnlund, and Hans Rosling. In 2006, Hans gave his first TED talk, called, “The best statistics you’ve ever seen”. It became one of the most watched TED talks ever.

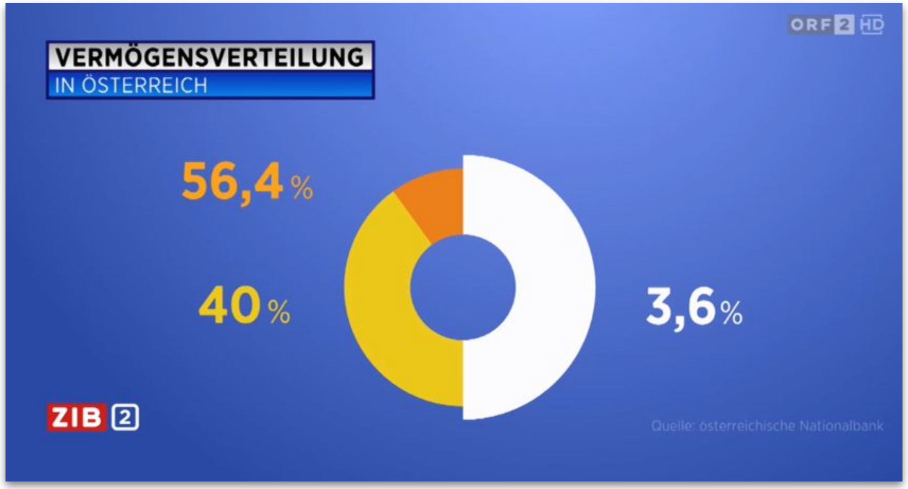


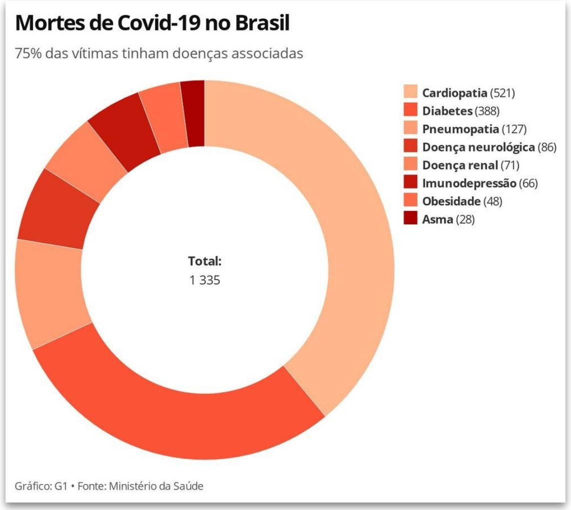
Internal Connected Cars live temperature data map, showing temperature data using kepler.gl

Examples of bad or misleading visualizations









https://viz.wtf/

Key takeaways and further reading

Key takeaways

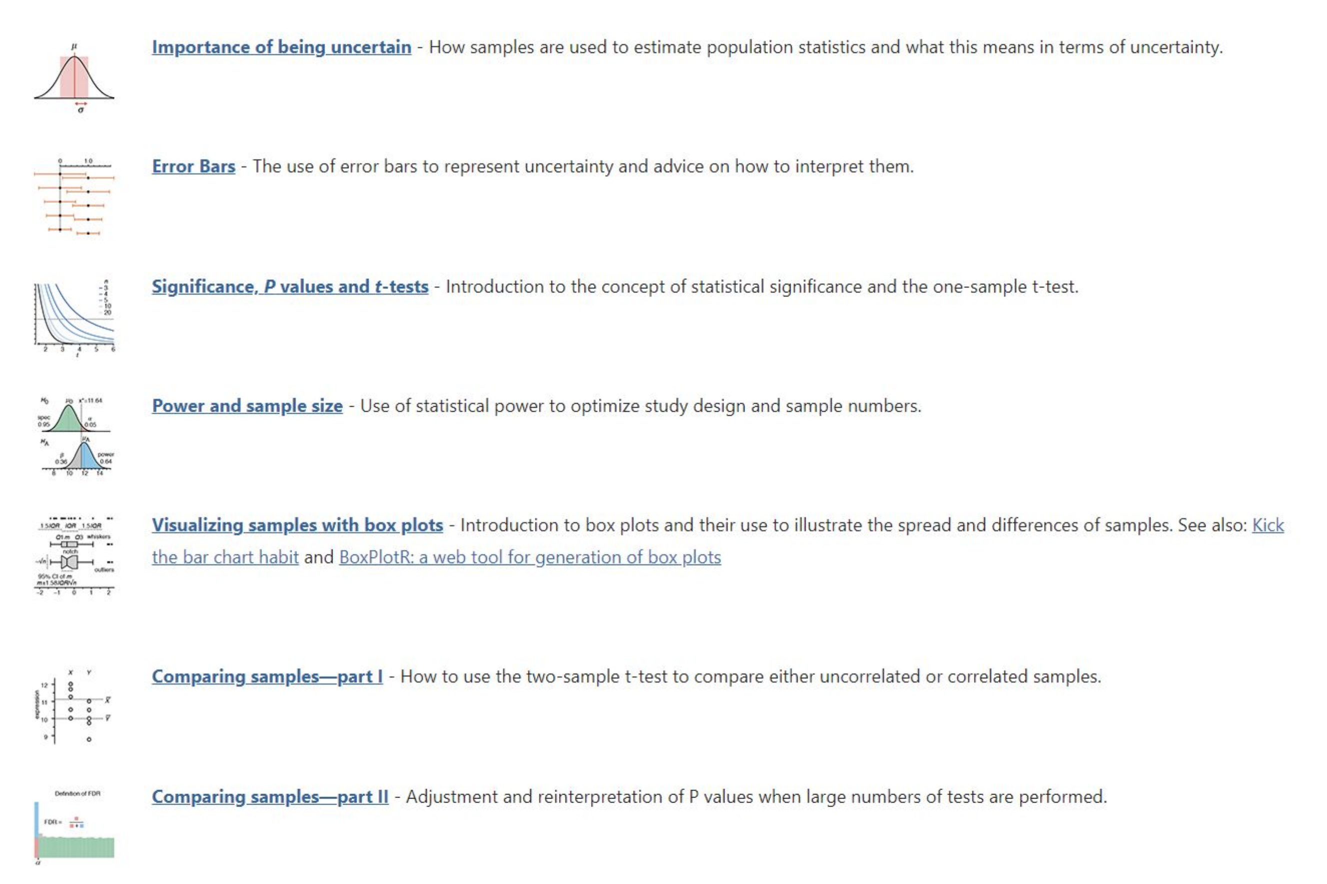
1. Using the right color scale is quite important
2. Help the reader/audience get to the right conclusion as fast as possible.
3. We as data professionals need to help achor an understanding and appreciation for uncertainties in our deliverables.
4. Descriptive statistics should only follow a more exploratory data analysis.

A good visualization should

- Tell a clear story
- Be understandable to the intended audience
- Be large enough to be seen and understood
- Have labels with units and a legend if applicable
- Show a sensible area of data. Outliers should not dominate the ranges of the axis
- Communicate the uncertainty of the results and preemptively address any misinterpretations

DATA VISUALIZATION SOCIETY

https://www.datavisualizationsociety.org



Importance of being uncertain - How samples are used to estimate population statistics and what this means in terms of uncertainty.

Error Bars - The use of error bars to represent uncertainty and advice on how to interpret them.

Significance, *P* values and *t*-tests - Introduction to the concept of statistical significance and the one-sample t-test.

Power and sample size - Use of statistical power to optimize study design and sample numbers.

Visualizing samples with box plots - Introduction to box plots and their use to illustrate the spread and differences of samples. See also: Kick the bar chart habit and BoxPlotR: a web tool for generation of box plots

Comparing samples—part I - How to use the two-sample t-test to compare either uncorrelated or correlated samples.

Comparing samples—part II - Adjustment and reinterpretation of P values when large numbers of tests are performed.

https://www.nature.com/collections/qghhqm/pointsofsignificance

Since September 2013 Nature Methods has been publishing a monthly column on statistics called "Points of Significance."