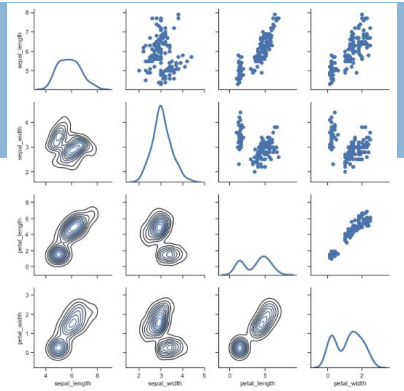
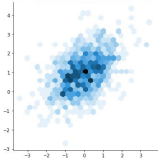
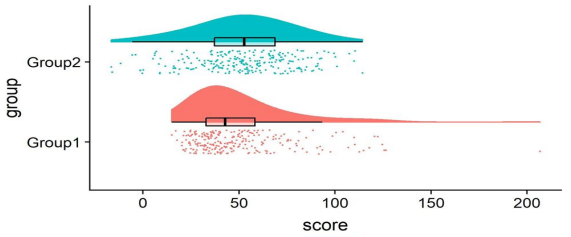
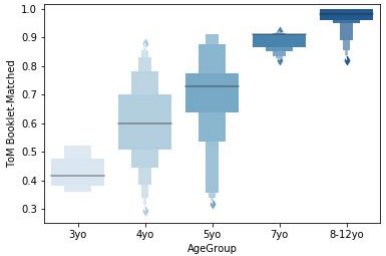
Data Visualization in Python

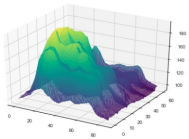
Neuro-Data-Science course 2021







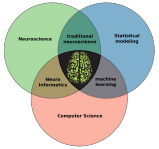




Peer Herholz (he/him) - reviewers/contributors : J. Sanz-Robinson, JB Poline
Postdoctoral researcher - ORIGAMI:Neuro-data-science, MNI, McGill, UdM, BRAMS, UNIQUE
Member - BIDS, ReproNim, Brainhack

@peerherholz

December 09, 2020 - revised July 2021



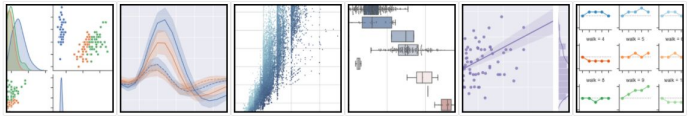
All the graphics...

- Python provides a wide array of options
- Low-level and high-level plotting APIs
- Static images vs. HTML output vs. interactive plots
- Domain-general and domain-specific packages
- Python has many different modules to visualize data, the most prominent/famous/versatile are:
    - MATLAB based plotting system: matplotlib
    - based on matplotlib but nicer looking: Seaborn or Pandas
    - interactive plotting tools: Bokeh and Plotly

matplotlib

pandas

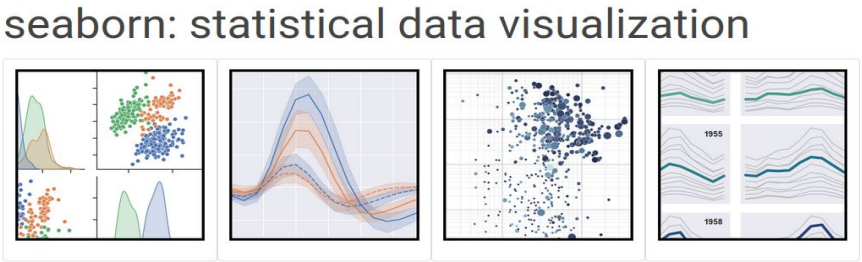
seaborn: statistical data visualization



plotly

bokeh

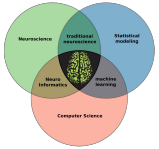
Invest in learning both “high level” and “low level” packages



matplotlib

Why do you need to know the principles / some of matplotlib?

- Because you most often, for article ready quality figures, you will need or want to tweak! If you can’t do it in Seaborn, you will be able to go one level below and do the tweaking in Matplotlib.



All the graphics...

- After this lecture, in the we will use a `jupyter notebook` that lets you explore the concepts

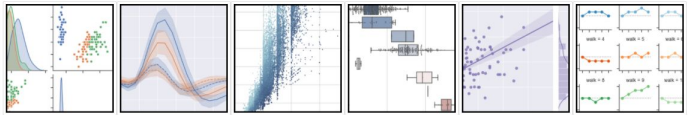
  In the office hours exercises, we'll go over using real data:

  `Python_visualization_for_data.ipynb`

matplotlib

pandas

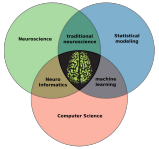
seaborn: statistical data visualization



plotly

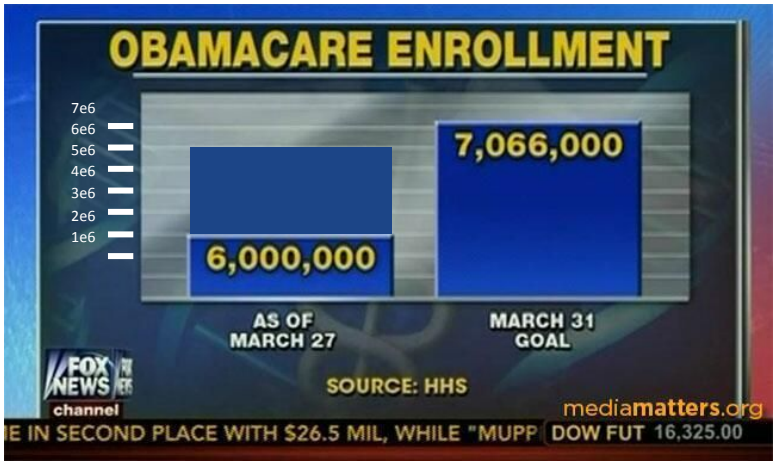
bokeh

https://github.com/neurodatascience/course-materials-2021

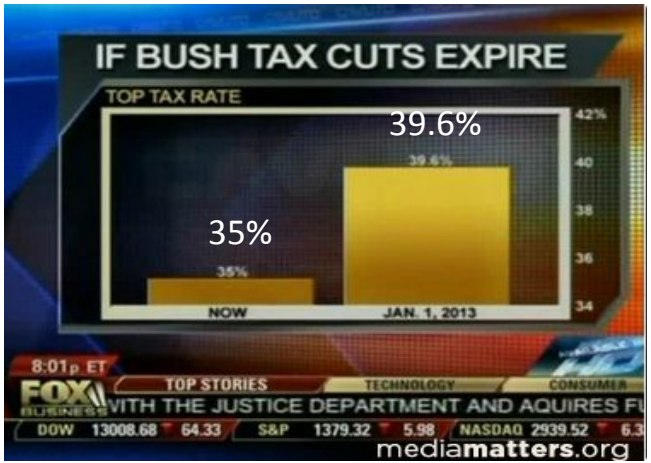


All the graphics...

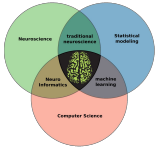
- while this is all fancy and cool, what's a major problem with most graphics?



https://i2.wp.com/flowingdata.com/wp-content/uploads/2014/04/Fox-News-bar-chart-1.jpeg?w=645&ssl=1

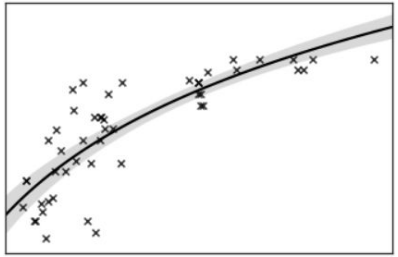


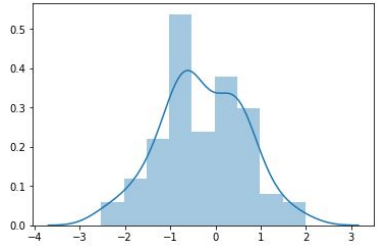
http://forums.applecentral.com/ubbthreads.php/topics/580002/1



Understanding your data: What does it mean?

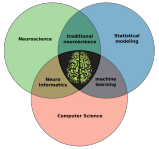
- Do: Learn the structure of your data
  - Data distribution
  - Summary statistics
  - Identify missing data
  - Data structure (dimensions of data)
  - Basic associations





- Don’t: Go on a fishing expedition
  - Search your data for “significant” associations
  - Test every association you can think of until one works
  - hide things and/or emphasize artifacts/wrong things
  - As soon as you have “seen” data : you may be biased in your “hypotheses”

Data analysis by design



Have a hypothesis?

- Data collection
- analysis plan
- preregistration

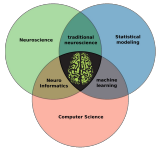
approach

important steps

Don't have a hypothesis?

- Cross-validation
- Left-out test set
- External dataset

The train of stubborn null hypothesis



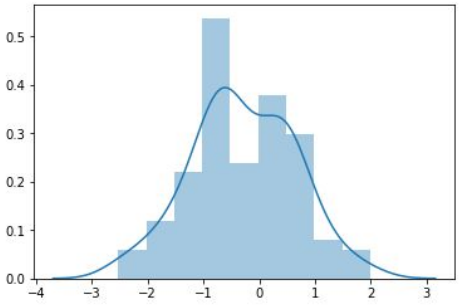
Design study/ collect data → Test null Hypothesis → Planned Alternative Hypotheses → Data exploration

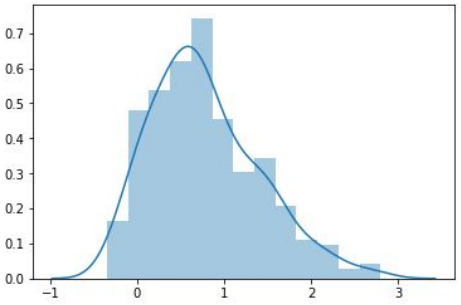
CONSIDERATIONS

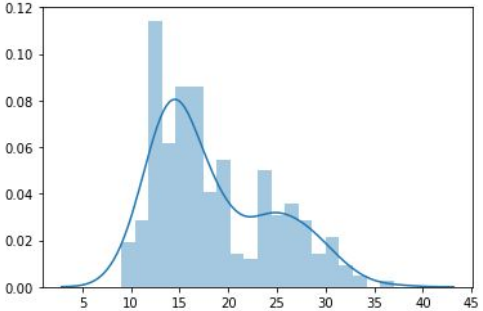
- Sample size
- Publishing null results
- Validation
- Posters and collegiate communication

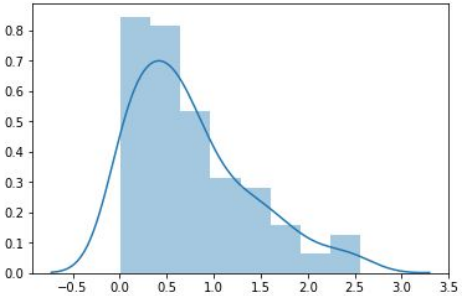
Getting to know your data

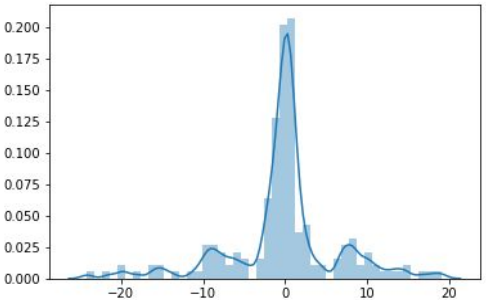
- (real) Data distributions

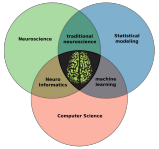






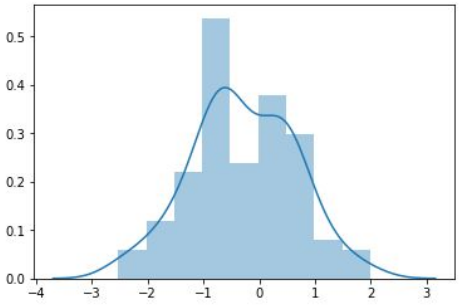


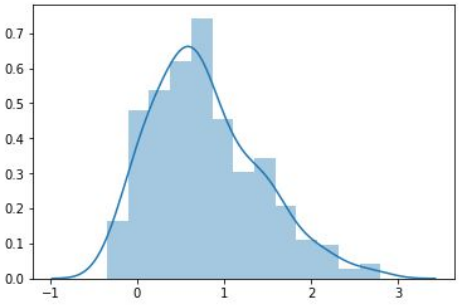


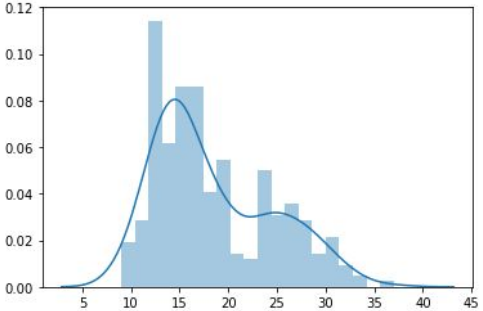


Getting to know your data

- (real) Data distributions







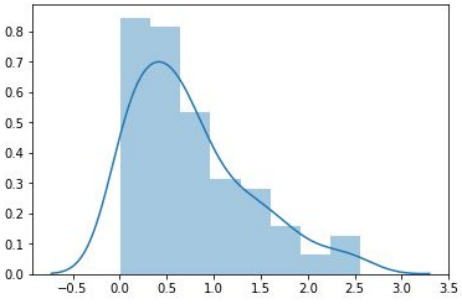
Normal-ish

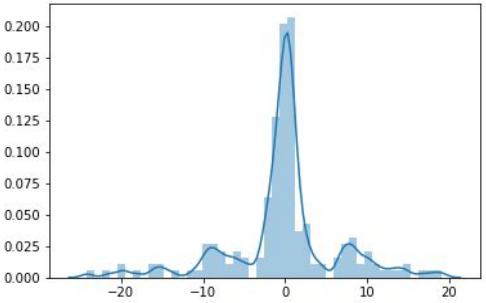
Skewed

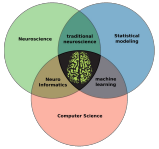
Likely Bimodal

Floor/Ceiling

Laplace-ish



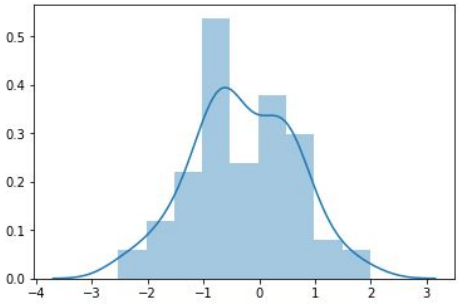


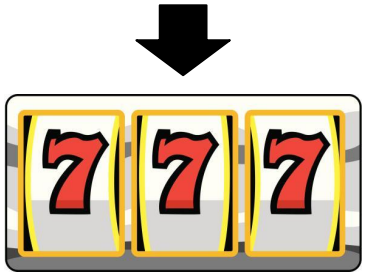


Getting to know your data

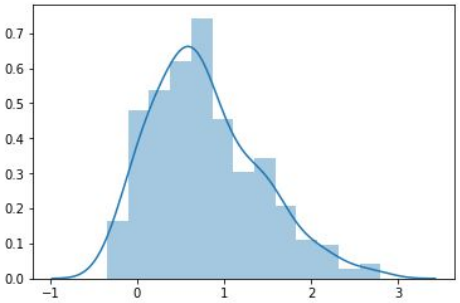
- (real) Data distributions

Normal-ish

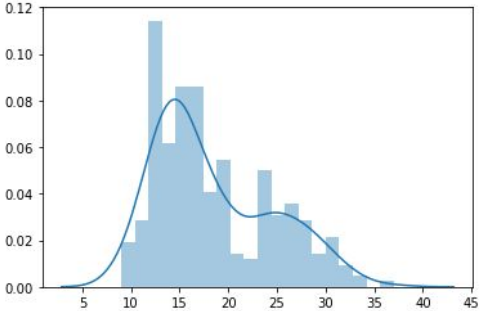




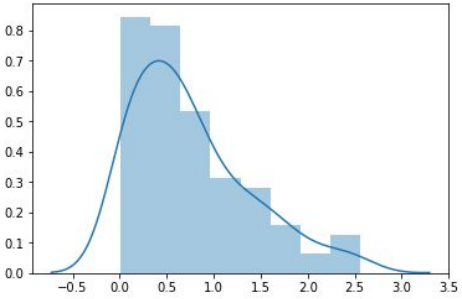
Skewed



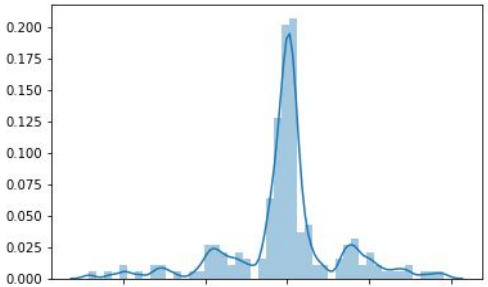
Likely Bimodal

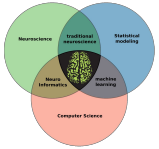


Floor/Ceiling



Laplace-ish



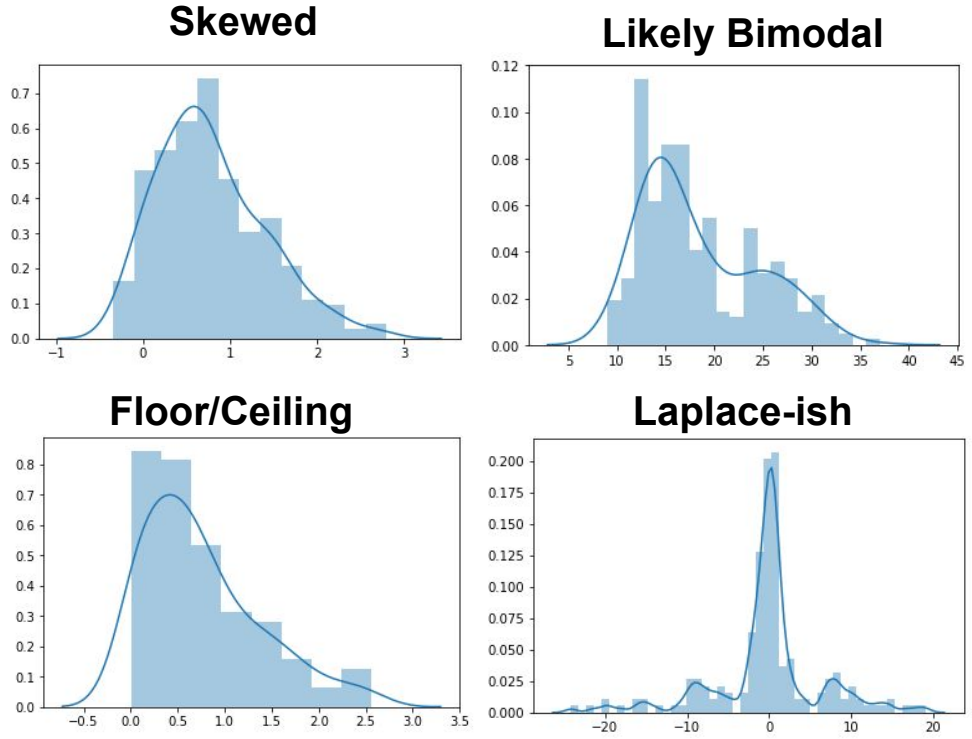


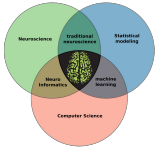
Getting to know your data

- real data distributions and how to tackle them

Strategies

- Non-parametric tests
- Categorization/encoding
- Normalization (e.g Log-norm)





Getting to know your data

- real data distributions and how to tackle them

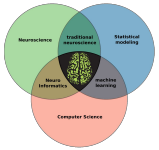
Strategies

- Non-parametric tests
- Categorization/ encoding
- Normalization (e.g Log-norm)

Considerations

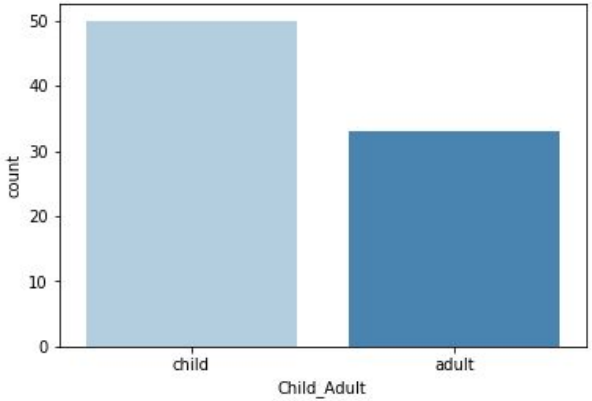
- Degrees of freedom
- Interpretability
- Generalizability

Visualizing your data: Univariate



Categorical

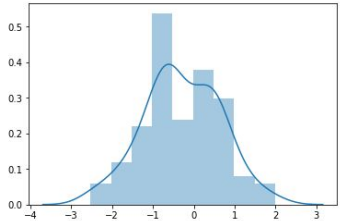
Countplot



Pie Chart (Not recommended)

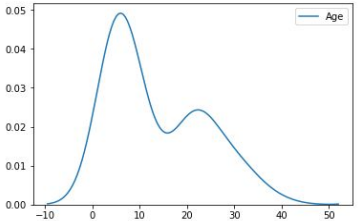
Scalar

Histogram



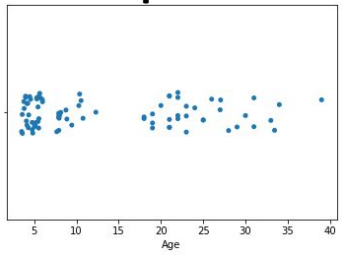
sns.distplot(x)

KDE Plot



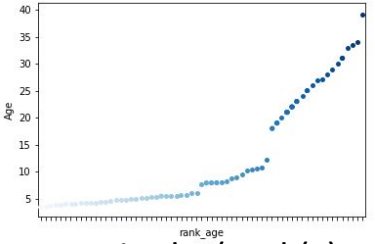
sns.kdeplot(x)

Strip Plot



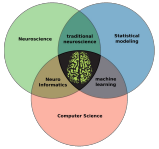
sns.stripplot(x)

Ranked Strip



sns.stripplot(rank(x), x)

Visualizing your data: Bivariate

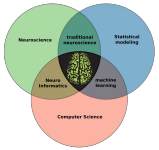


Categorical Categorical

Categorical Scalar

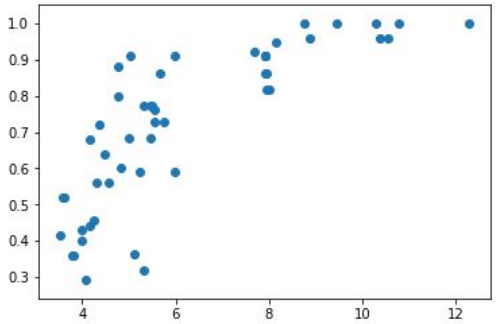
Scalar Scalar

Visualizing your data: Bivariate



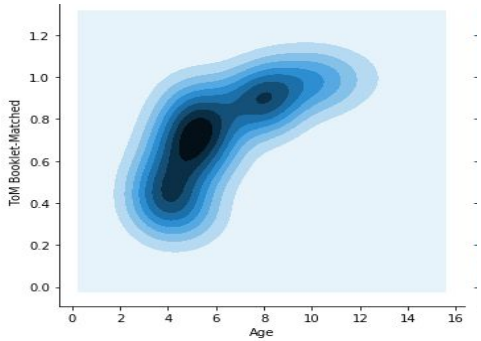
Categorical Categorical	Categorical Scalar	Scalar Scalar

Scatterplot



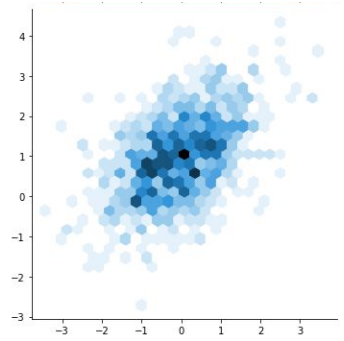
sns.regplot(x,y)
sns.lmplot(x,y)
sns.jointplot(x,y)

Density plot



sns.jointplot(x,y, kind = 'kde')

Hexplot



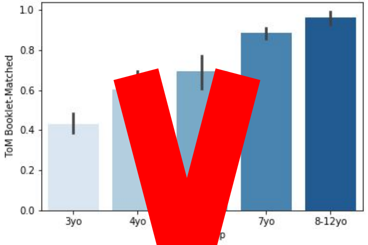
sns.jointplot(x,y, kind = 'hex')

Good for large sample sizes

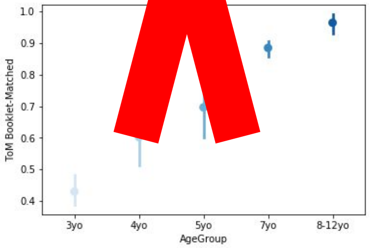
Visualizing your data: Bivariate

Categorical Categorical	Categorical Scalar	Scalar Scalar

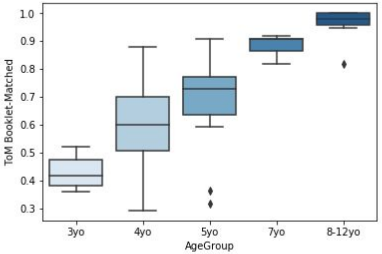
sns.barplot(x,y)



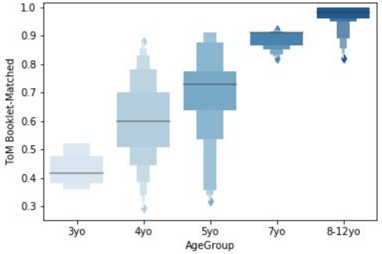
sns.po plot(x,y)



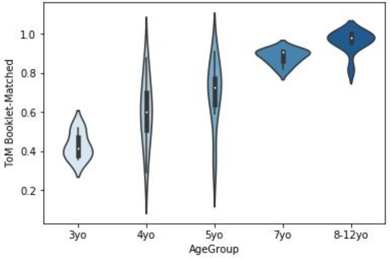
sns.boxplot(x,y)



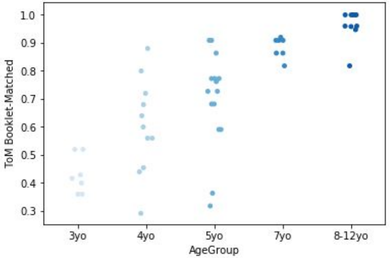
sns.boxenplot(x,y)



sns.violinplot(x,y)



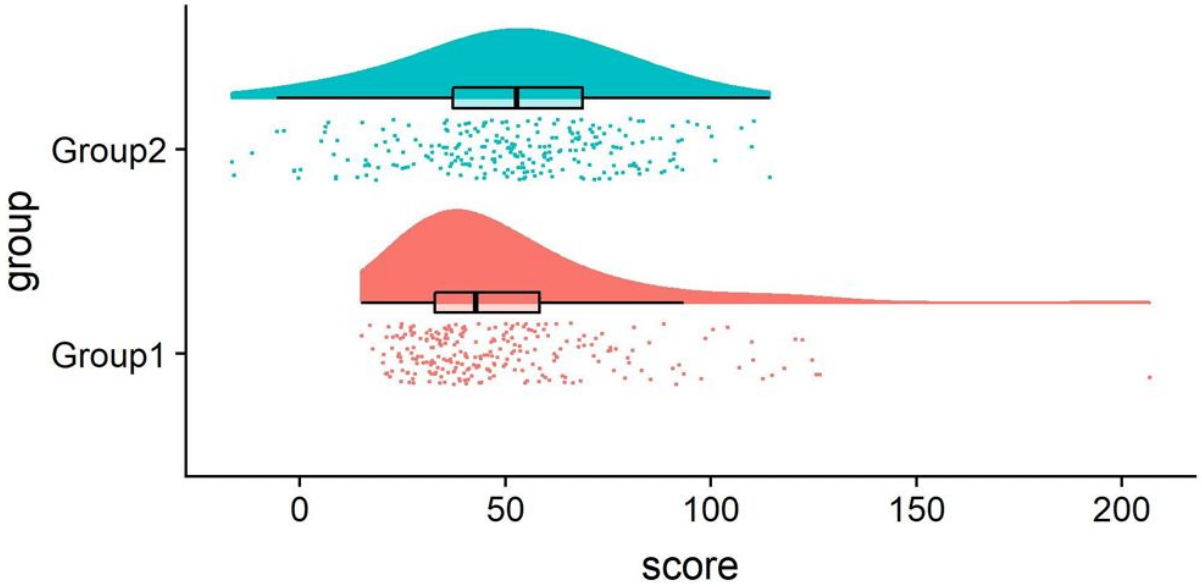
sns.stripplot(x,y)



Visualizing your data: Bivariate

Categorical Categorical	**Categorical Scalar**	Scalar Scalar

Raincloud Plots



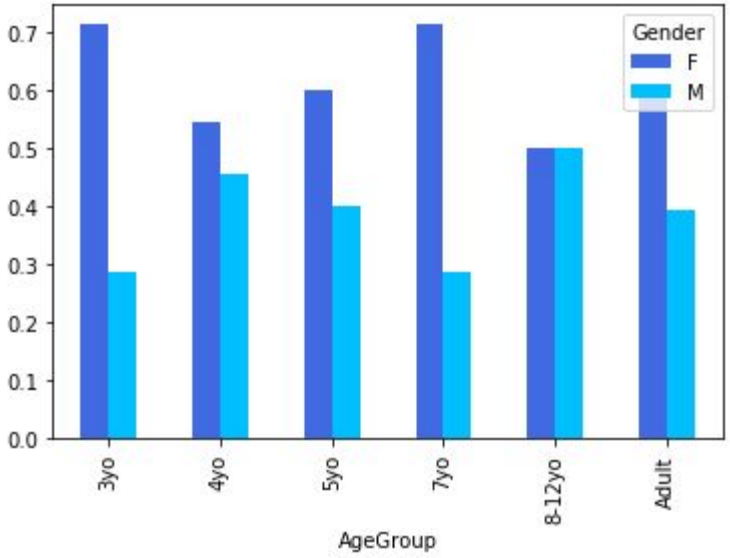
Visualizing your data: Bivariate

Categorical Categorical

Categorical Scalar

Scalar Scalar

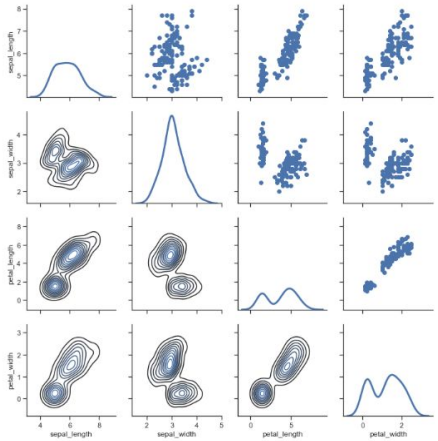
Barplots are okay here!



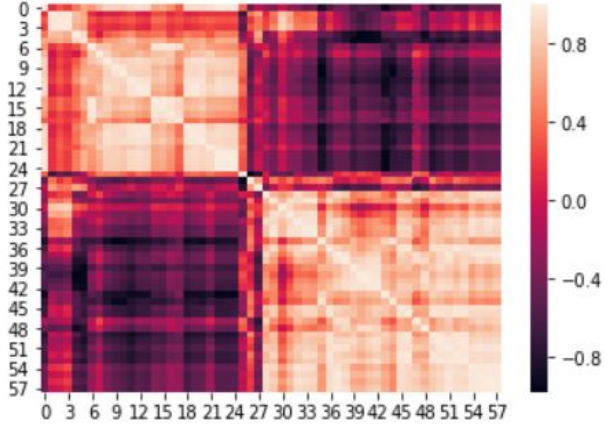
Visualizing your data: Multivariate

- Visualize several pairwise relationships at once with `pairplots`, `heatmaps` or `clustermaps`

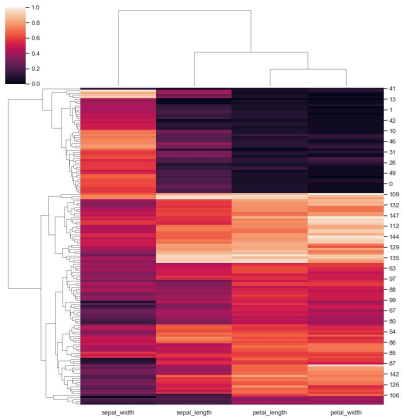
pairplot

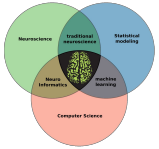


heatmap



clustermap

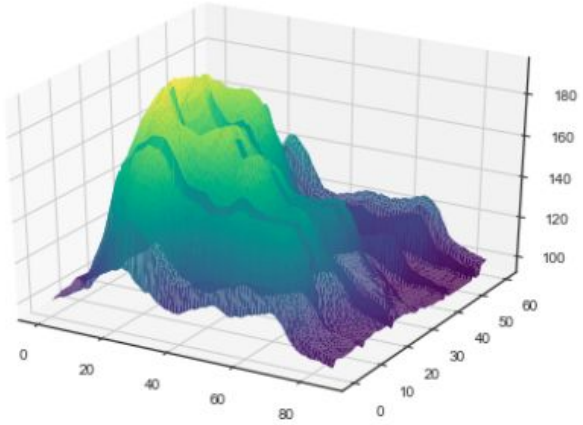




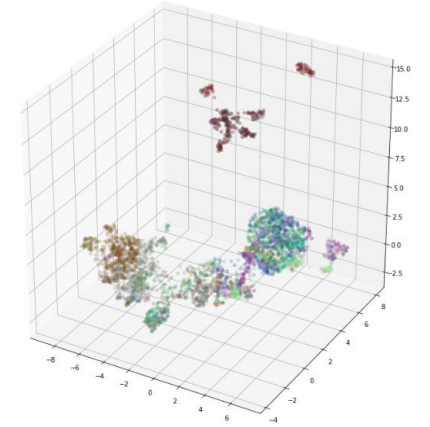
Visualizing your data: Multivariate

- Visualize the interaction between three variables with 3D plots

3D surface plot

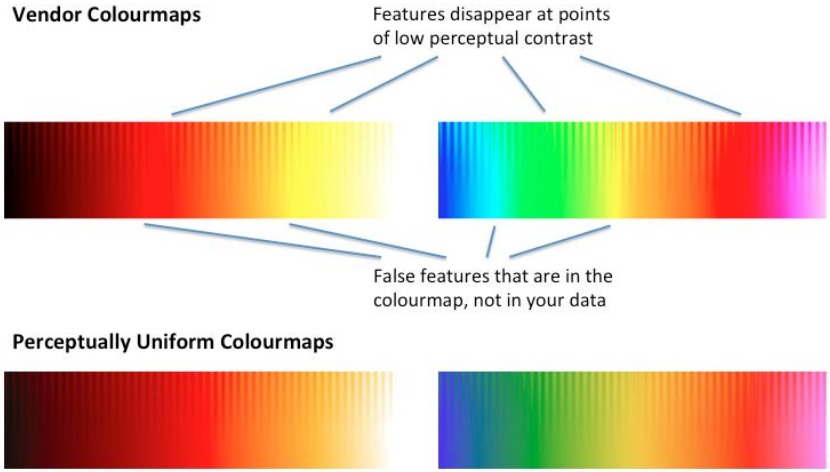


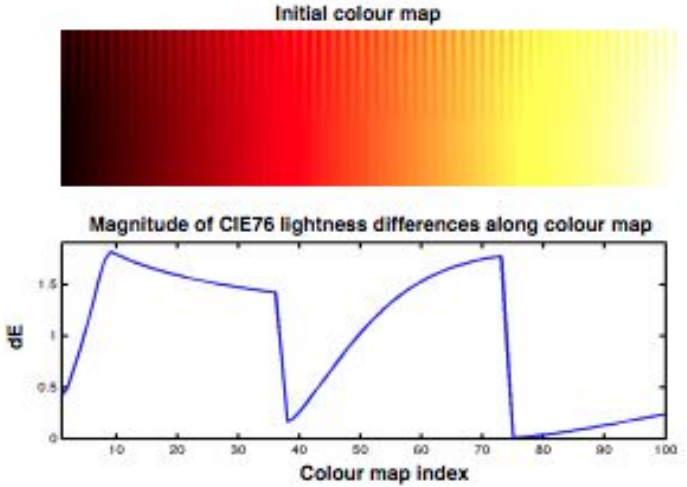
U-MAP + 3D scatter



A word on style

- Don’t use the default style. Be creative! And most important: Be honest!

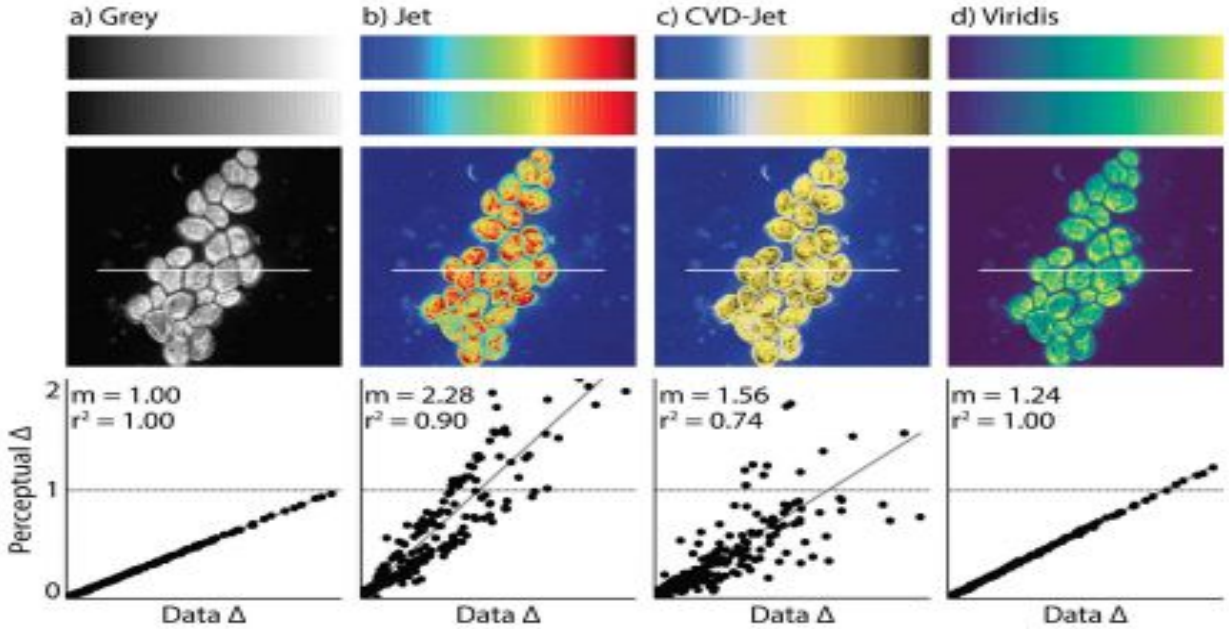




Nunez et al. 2018 *PLoS One*, Kovesi, 2015, Arxiv

A word on style

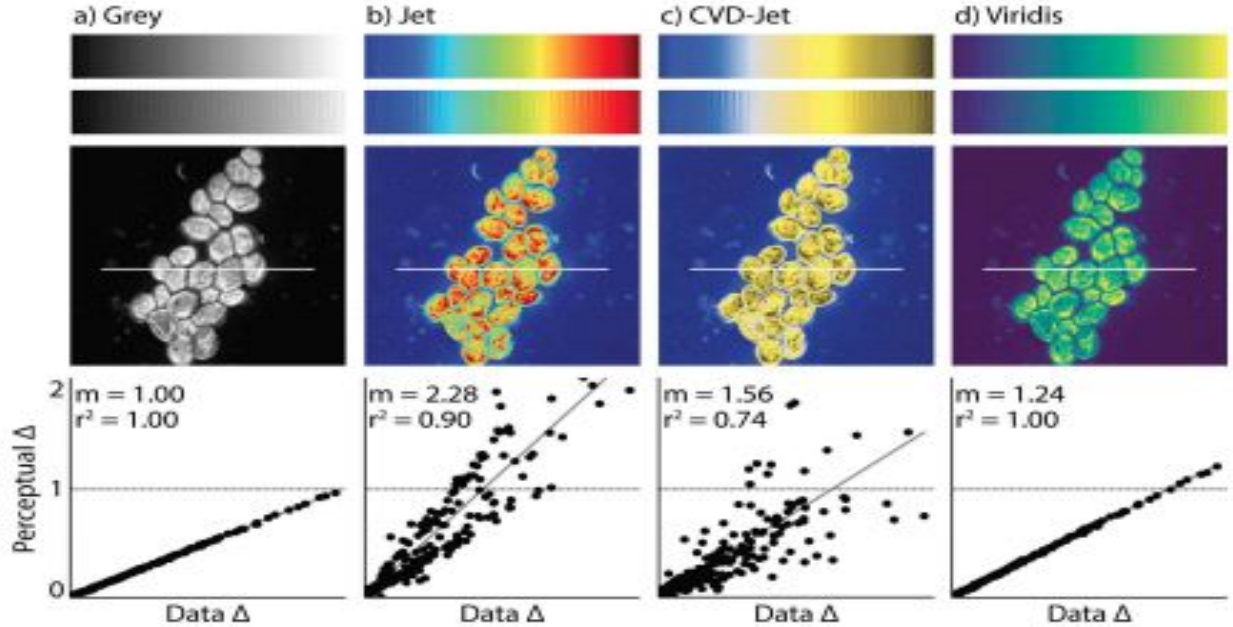
- Don’t use the default style. Be creative! And most important: Be honest!



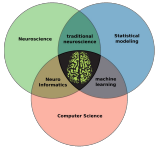
Nunez et al. 2018 *PLoS One*, Kovesi, 2015, Arxiv

A word on style

- Consider color choices (perceptually uniform, colorblind-sensitive, is color necessary?)



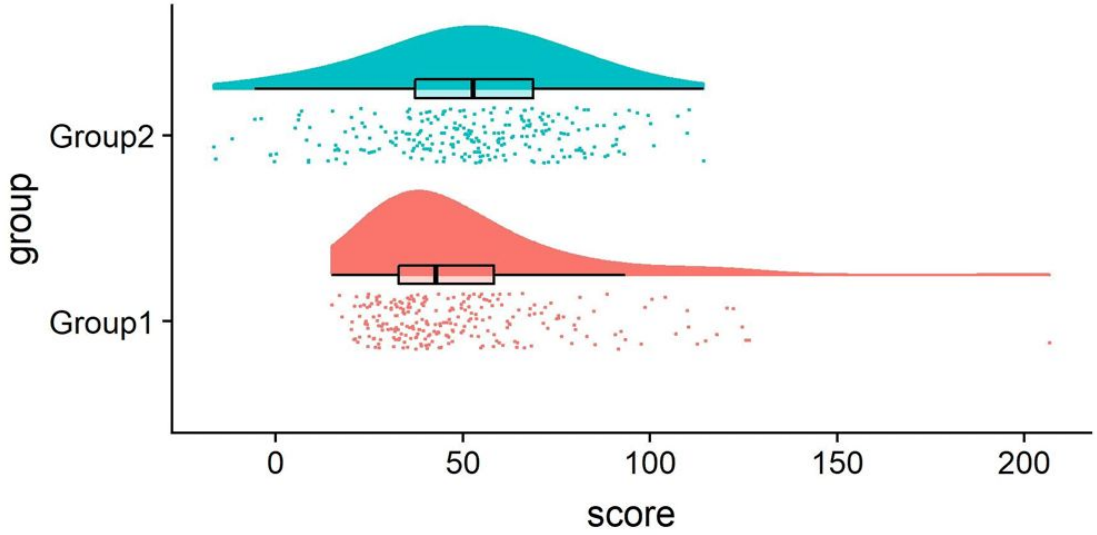
Nunez et al. 2018 *PLoS One*, Kovesi, 2015, Arxiv

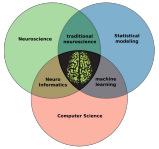


A word on style

- Save time by using python for all aspects of plot creation (Everything is customizable!)

- Show me the data!!

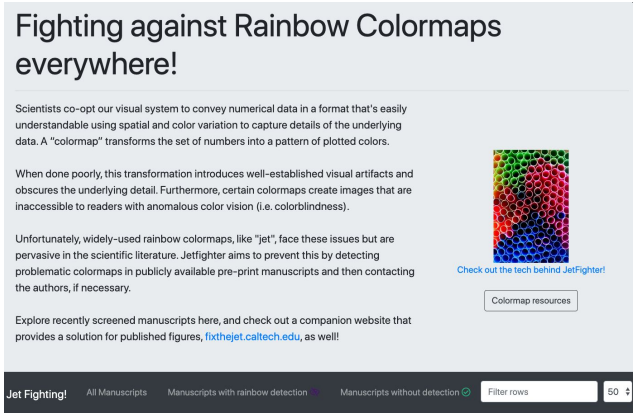


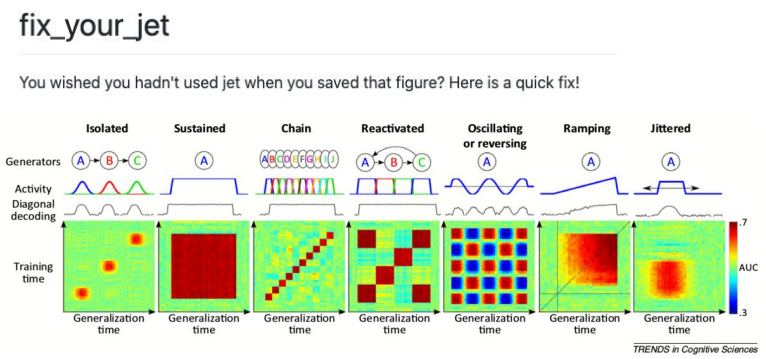


A word on style

- Check out the very helpful and cool new homepage https://python-graph-gallery.com/ to see how you can create different kinds of graphs
- familiarize yourself with different color maps: https://matplotlib.org/3.1.0/tutorials/colors/colormaps.html

- don't use jet!





Conclusions

- Visualization is a useful tool to figure out what is going on in data.
- Your figures should be reproducible (code + data).
- Display your data without forcing a perception of it.
- Be careful that the trends you see in exploratory analyses don’t bias your hypotheses and tests.
- It is good to know both the high- and low-level libraries!