Data Visualization

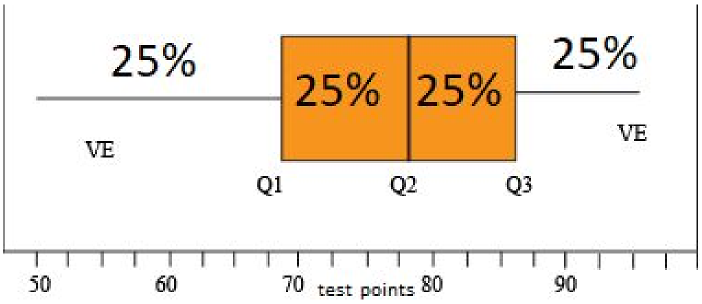
How to visualize data. Graphs, heatmap, pychart,...

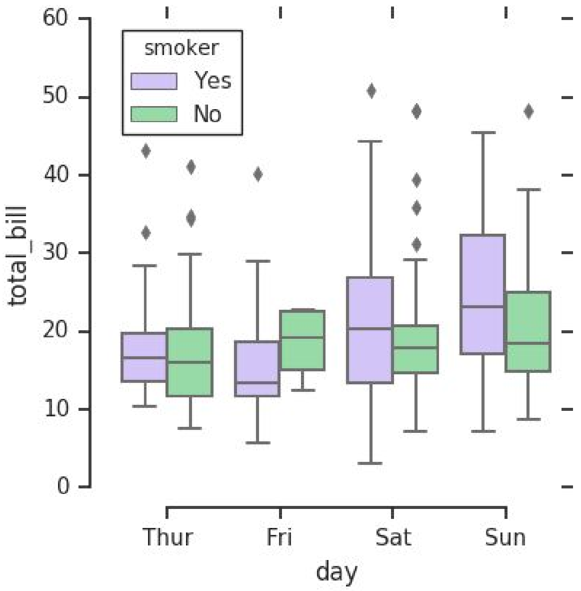
How to call data.

Recommended python package named as **seaborn** [(http://seaborn.pydata.org](http://seaborn.pydata.org)) statistical data visualization, based on **matplotlib**. You can install with pip install seaborn (documentation->Installing and getting started. It requires **numpy,scipy, matplotlib, pandas** package, and recommended **statsmodels**. When you open seaborn website, you can see gallery, tutorials and API.

On gallery section of seaborn web we have a lot of different visualizations:

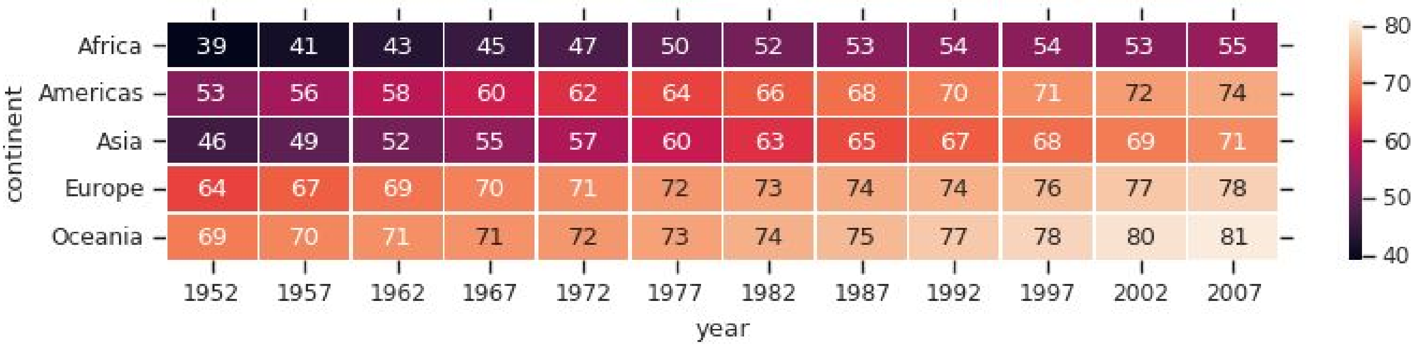
- Boxplots: source is sort the data and get the minimum, maximum, quartiles Q1(25%, lower median half), Q2(50% all data median) and Q3(75%, upper median half) and interquartile range (IQR= Q3 - Q1). Are a useful way to graph data divided into four quartiles, each with the same amount of values. The box diagram does not graph frequency or show individual statistics, but in them we can clearly see where half the data is. It is a good diagram to analyze the asymmetry in the data. Values below Q1–1.5 · IQR or higher than Q3 + 1.5 · IQR are considered outliers



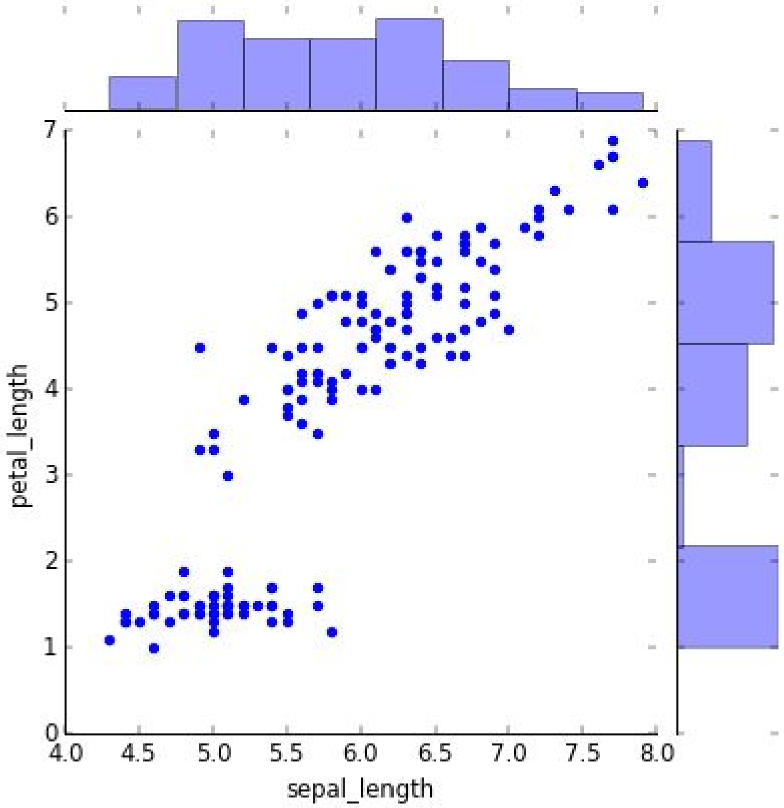


Boxplot shows the three quartile values of the distribution along with extreme values. Whiskers extend to points that lie within 1.5 IQRs of the lower and upper quartile, and then observations that fall outside this range are displayed independently

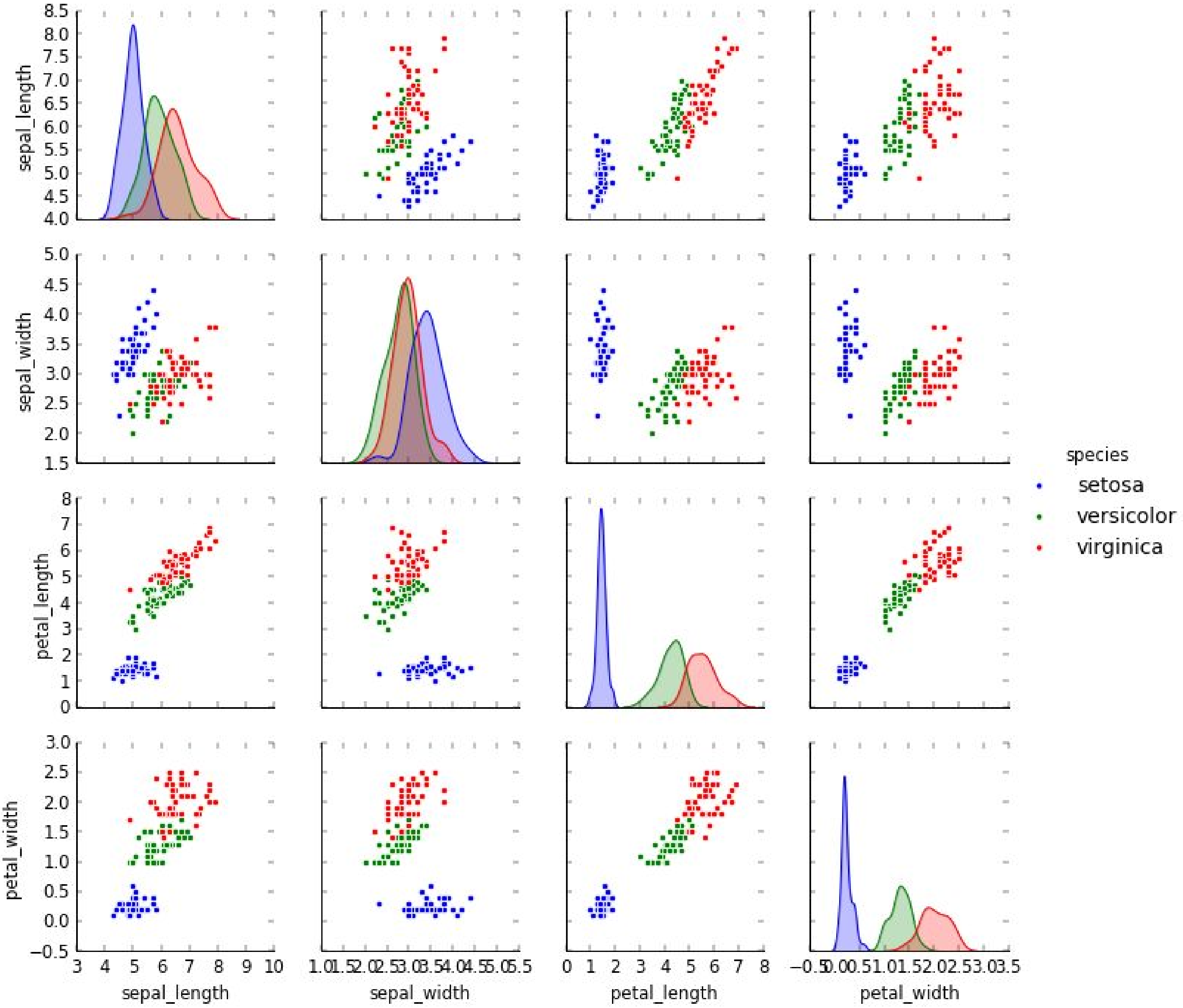
❏ Heatmaps (annotated): color represents values, eg. pivot table. Color represents values of the data. There is an example for Life expectancy with a pivot table



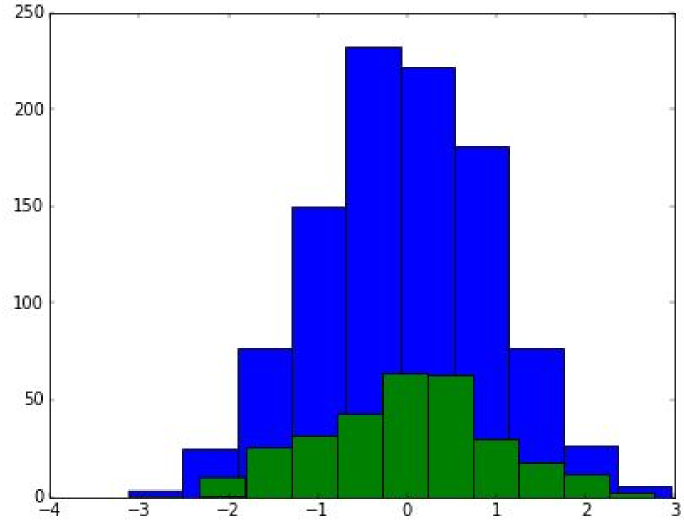
❏ Jointplot: Visualize dataset structure, e.g: hexagonal bin plot with marginal distributions. focuses on single relationship.

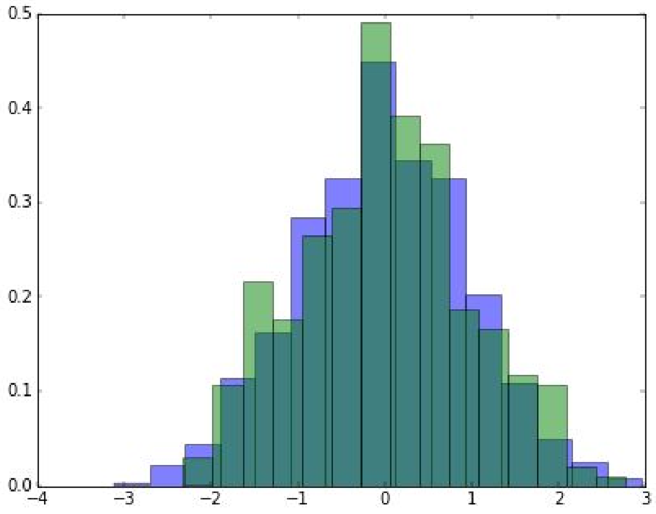


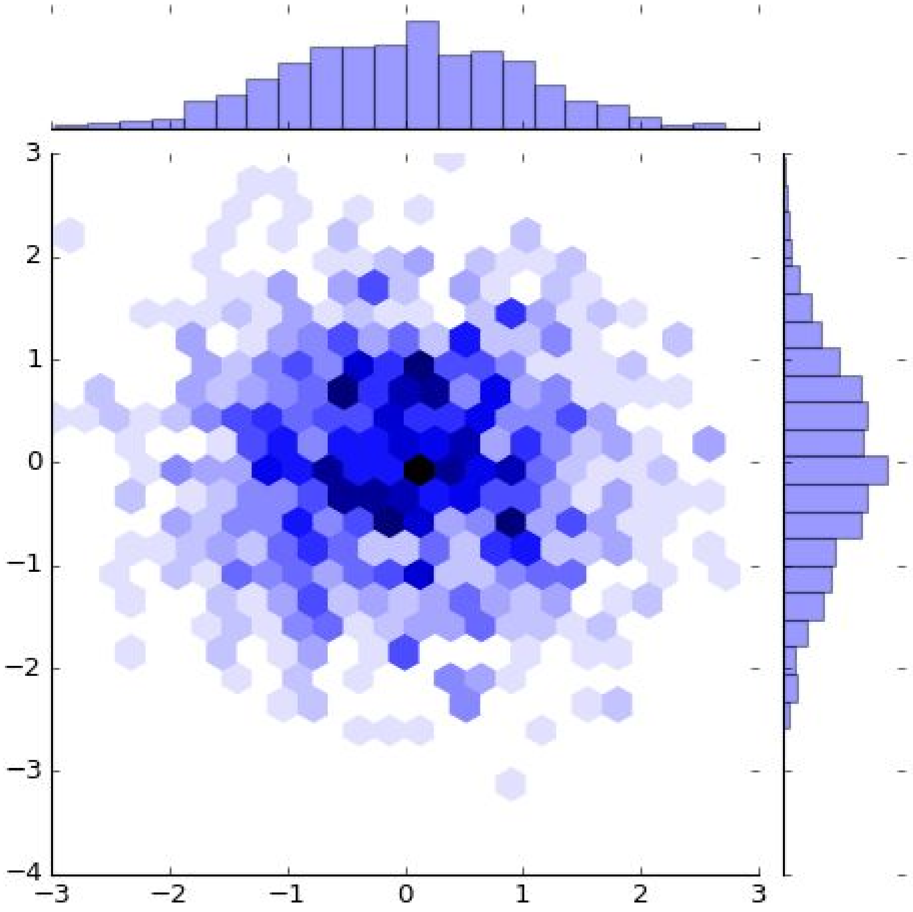
- Pairplot: takes a broader view, showing all pairwise relationships and the marginal distributions, optionally conditioned on a categorical variable.



- Histograms: accurate representation of the distribution of numerical data, is an estimate of the probability distribution of a continuous variable. First step is to "bin" or "bucket" the range of values, that is, divide the entire range of values into a series of non overlapping intervals, that must be adjadcent, but not required of equal size; then count how many values fall into each interval. Histogram can be normalized to display relative frequencies; it shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1. When bins are not equal witdh, vertical axis is not a frequency, is a frequency density.



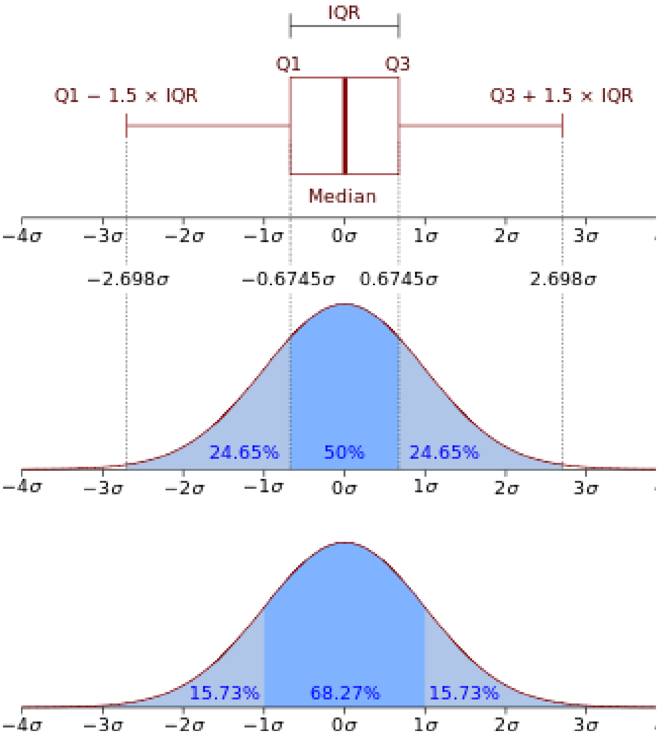




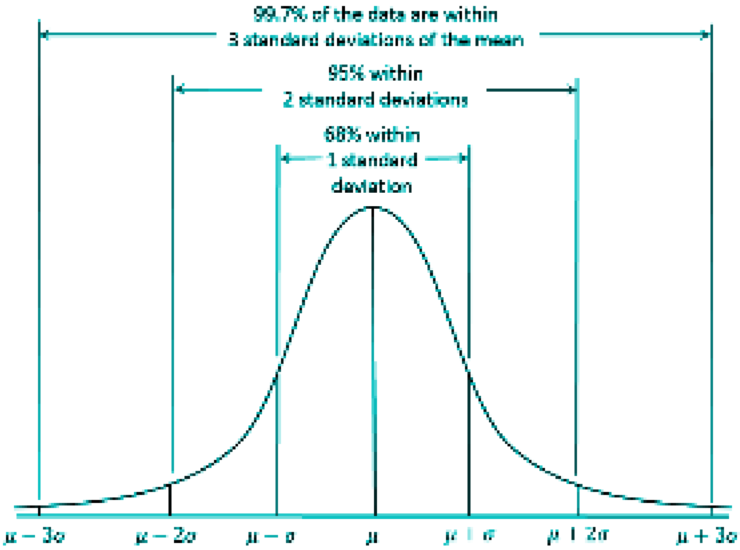
Good references for more info were Wikipedia.

- Kernel plots: Kernel Density Estimation (KDE) is a non-parametric way to estimate the probability density function(PDF) of a random variable. Is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. In some fields such as signal processing and econometrics it is also termed the Parzen–Rosenblatt window method.

Foundations for non discrete distribution



Standard Distribution with associated boxplot



Normal distribution: standard deviation and coverage

Foundations for discrete distribution (kernels)

- .
-