Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA
2020/2021

FACULTY DIEGO PAJARITO

Load data

Using pandas rather than simple python tools

Pycharm
Python
Github
Sublime
QGIS

Libraries
- Pandas
- JSON
- geopandas

Python data analysis library

Data analysis tools for the Python programming language
Open source, BSD-licensed library
High-performance, easy-to-use data structures

Using Conda
```
conda install pandas
```

Using PIP
```
python -m pip install --upgrade pandas
```

Using Conda's graphical interface or Pycharm

Pandas documentation available here: https://pandas.pydata.org/

List:	Dictionary:	Tuple:

['Tokyo',
'Delhi',
'Shanghai',
'Sao Paulo',
'Mexico City',
'Cairo',
'Dhaka',
'Mumbai',
'Beijing',
'Osaka']
	{
	'cities': ['Tokyo','Delhi', 'Shanghai',
	'Sao Paulo', 'Mexico City', 'Cairo',
	'Dhaka', 'Mumbai', 'Beijing', 'Osaka'],
	'population': (37435191, 29399141, 26317104,
	21846507, 21671908, 20484965, 20283552,
	20185064, 20035455, 19222665),
	'source':
	'http://worldpopulationreview.com/world-cities/'
	}
		(37435191,
		29399141,
		26317104,
		21846507,
		21671908,
		20484965,
		20283552,
		20185064,
		20035455,
		19222665)

if/else/elif:	For/While:

`if a == 0:`
`  print('A equals 0')`
`else:`
`  Print('A not 0')`
	`for i in range(10):`
	`  print('i equals: ' + str(i))`

id	name	address	postal_code	lon	lat
1	IAAC main building	Carrer de Pujades 102	08005	2.1932315826416016	41.395747068298895
2	IAAC atelier	Carrer de Pujades 59	08005	2.1919387578964233	41.39522593585012

Get Column / Columns / Row / Rows	Get Subset

`df['column']`
`df[['column1', 'column2']]`
`df[0:1]`
`...`
	`df.head()`
	`df.iloc[ : , 2 ]`
	`df.loc()`
	`...`

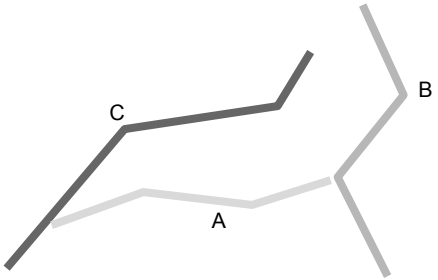
Try to create your own dataframe, add some values and see how you can create it from lists, tuples or dicts.

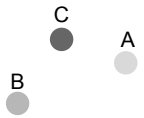
~~Load~~ view data

Using QGIS to see data in the space

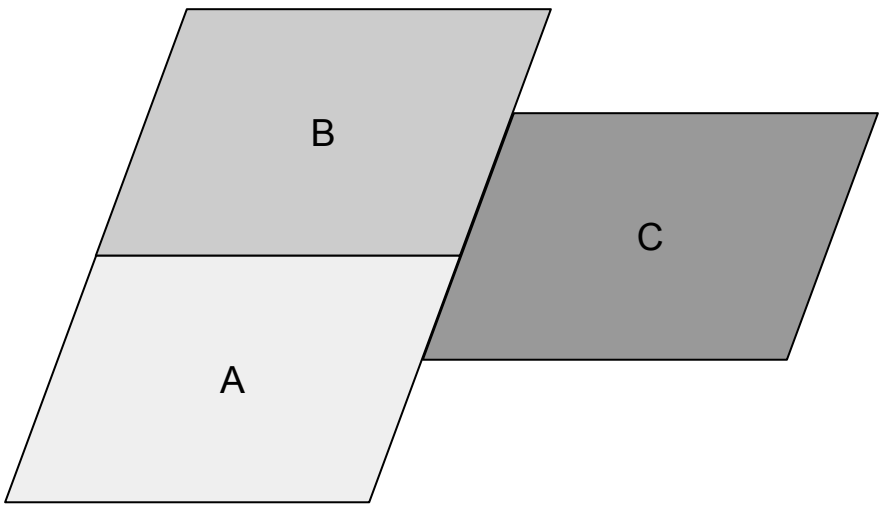
Institute for advanced architecture of Catalonia

GIS data

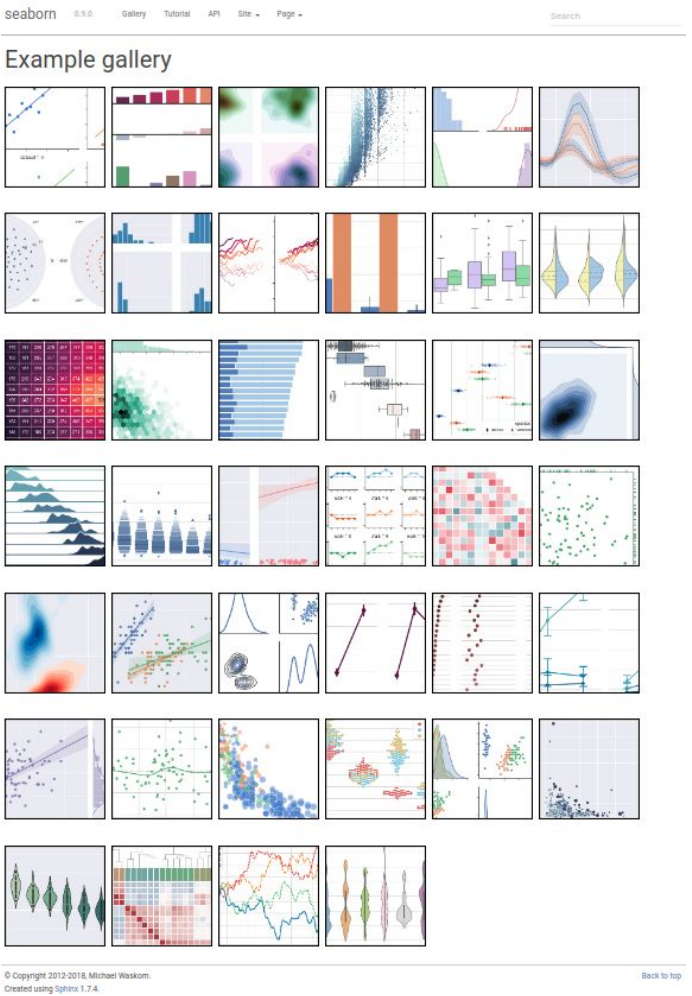




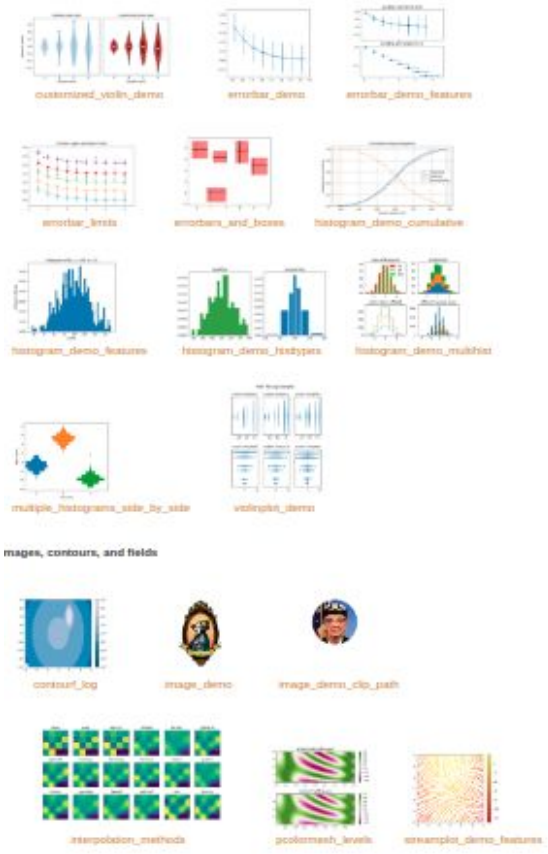
A	B	B	B
C	A	B	B
C	C	A	B
C	C	A	A



A	10
B	20
C	30


seaborn 0.9.0 Gallery Tutorial API Site Page Search
Example gallery
© Copyright 2012-2018, Michael Waskom.
Created using Sphinx 1.7.4.
Back to top


bar_demo
fill_demo
fill_demo_features
line_demo_dash_control
line_styles_reference
linestyles
marker_fillstyle_reference
marker_reference
marker_reference
scatter_with_legend
Shapes and collections
artist_reference
path_patch_demo
scatter_demo
Statistical plots
boxplot_color_demo
boxplot_demo
boxplot_demo
boxplot_vs_violin_demo
bxp_demo
bxp_demo


customized_violin_demo
errorbar_demo
errorbar_demo_features
errorbar_limits
errorbars_and_boxes
histogram_demo_cumulative
histogram_demo_features
histogram_demo_histtypes
histogram_demo_multihist
multiple_histograms_side_by_side
violinplot_demo
Images, contours, and fields
contourf_log
image_demo
image_demo_clip_path
interpolation_methods
pcolormesh_levels
streamplot_demo_features

Statistics

Descriptive

Variables can take multiple values and change across the time. We use variables to describe objects, so they need a clear definition. Some elements to consider when defining variables are:

Name
Description
Purpose
Domain

And, therefore, data structure

The domain refers to the kind of values a variable can take. Domains define the most convenient data structure to use and are as diverse as variables.
Some examples for domains and data structures are:

Boolean	True / False
Numeric	Int: …, -3, -2, -1, 0, 1, 2, 3, ... Float: -∞, …, 0.0, …, ∞
Char	‘a’ ‘word’ ‘something else !”·$%&(/’ ‘../meaningful/text/file.ext’
Objects	Lists [ ] Tuple () Dict {‘key’: ‘value} List of lists [[ ], [ ]] Dataframe

Intrinsic features of the data set. These statistics set the path for the type of analysis to perform and the way to combine with other datasets

class, size, rows, columns, keys, range, ...

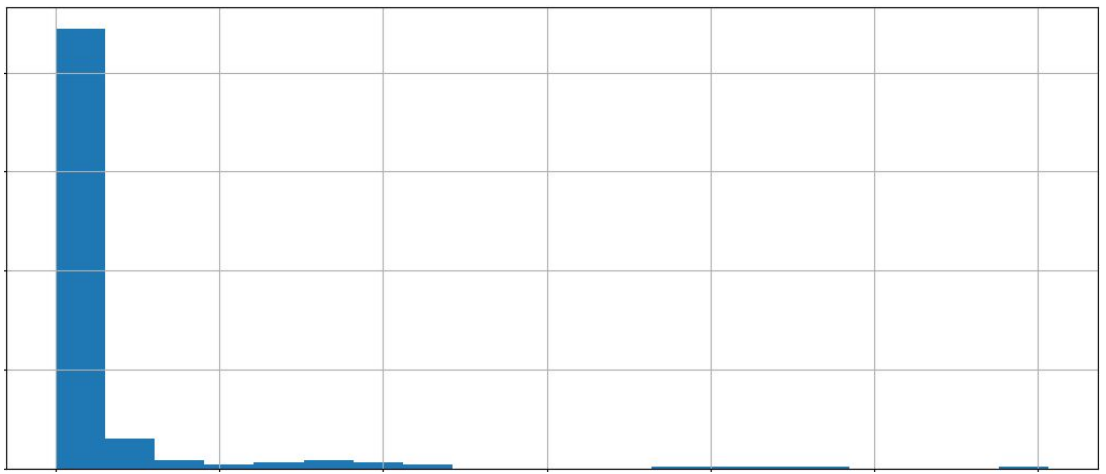
Explore the documentation and find these (and other) features for the country population dataset.

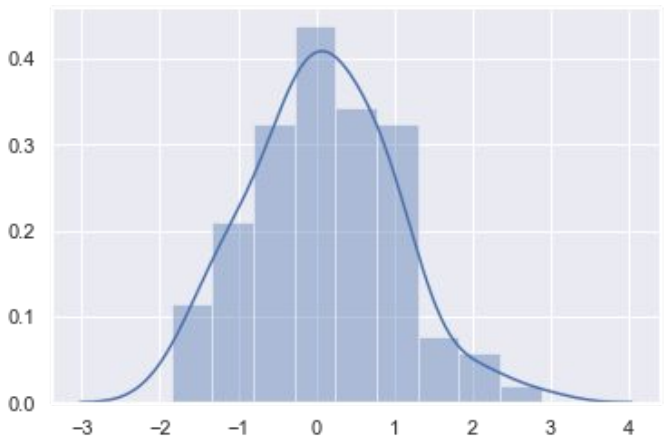
Serve to describe a given dataset in terms of tendencies and data dispersion.
These statistics gives the overall picture of a dataset

What can we say about country name and population?

Central tendency	Mean Mode Median Frequency (integer, classes)
Dispersion	Range Variance Standard Deviation

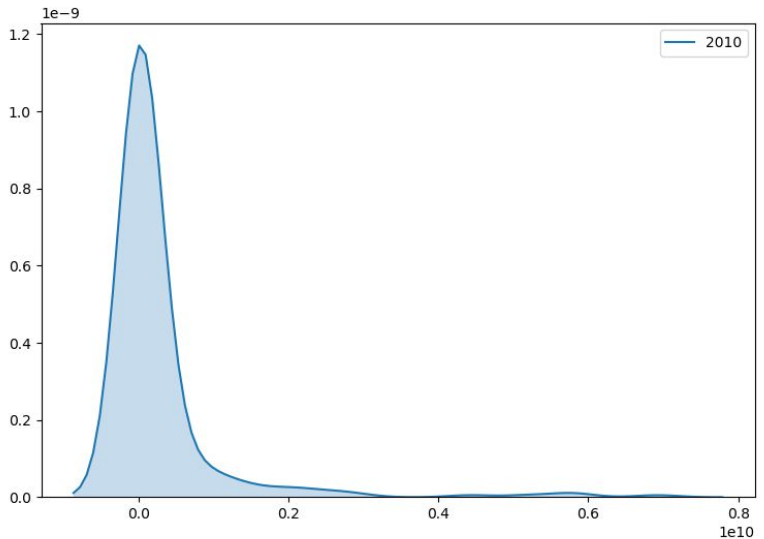
We can use either matplotlib or seaborn to create the histograms. **Histograms** are representations of the distribution of numeric data. After establishing ranges, the numeric values **falling into these** ranges and counted and represented in a plot. The equivalent diagram for non-numeric data is based on **frequency tables**.



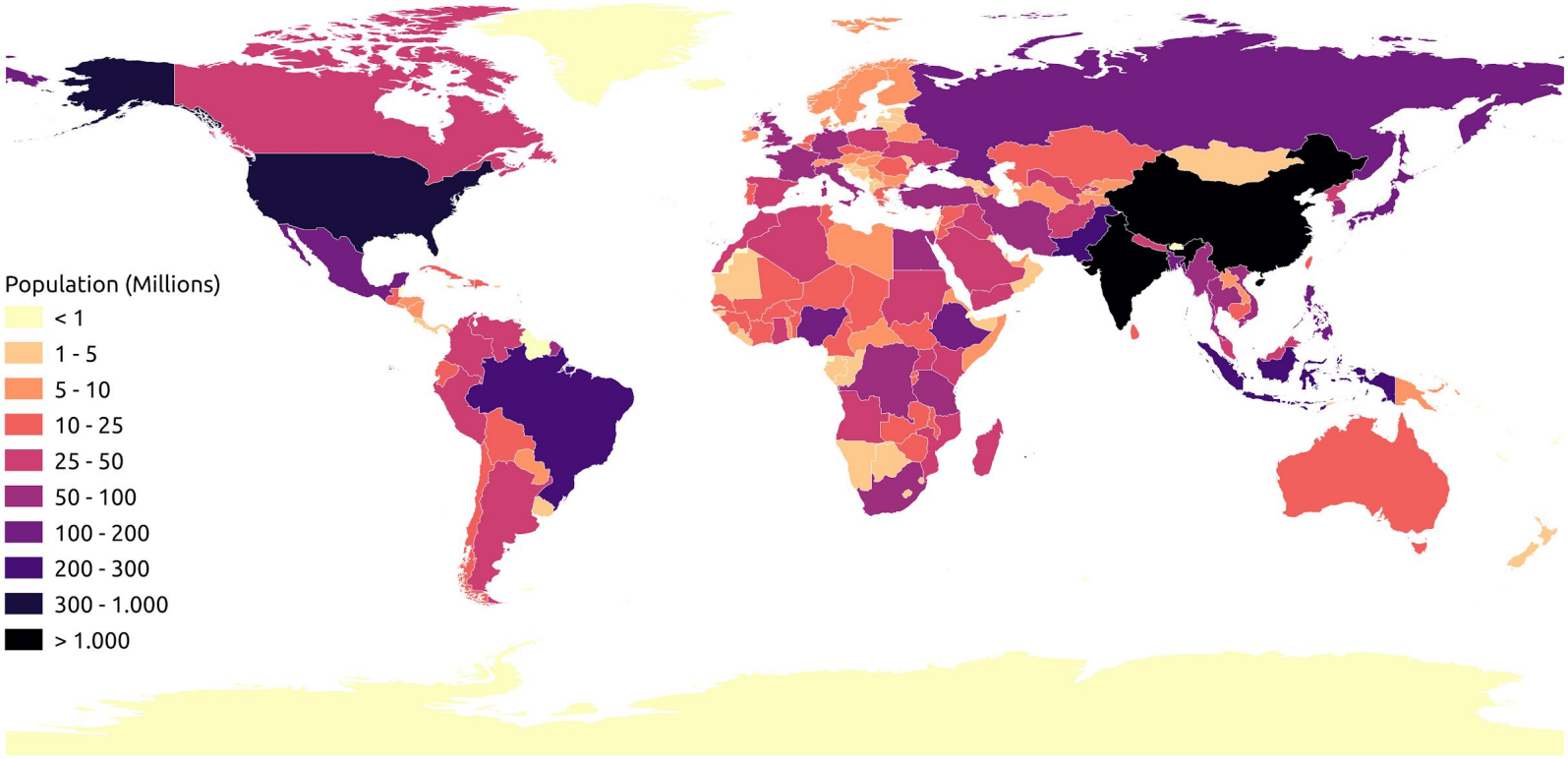


Normal distribution is a common reference for numeric distribution. It assumes a random process to generate data that is usually not the case. It is built from the mean and standard deviation values that can be overlayed to the histogram.

We can use either matplotlib or seaborn to create the histograms, probability distribution and density charts.



QGIS offers mapping capabilities and map integration beyond the use of geometries and shapes. This tasks also implies manual set up for data management (potentially optimised through python scripts).



To go beyond in Pandas

https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html#min

https://matplotlib.org/3.1.1/users/index.html

https://towardsdatascience.com/matplotlib-tutorial-learn-basics-of-pythons-powerful-plotting-library-b5d1b8f67596

https://seaborn.pydata.org/tutorial.html

https://elitedatascience.com/python-seaborn-tutorial

And QGIS

https://docs.qgis.org/3.4/en/docs/training_manual/index.html

Just to feed your curiosity:

Bivariate and pairwise plots as well as Spatial Clustering and geometry generator

Institute for advanced architecture of Catalonia | BARCELONA

MASTER IN CITY & TECHNOLOGY
DIGITAL TOOLS AND BIG DATA
2019/2020

FACULTY DIEGO PAJARITO