the experts
school of analytics

DATA ANALYSIS AND PROGRAMMING WITH PYTHON

Lecture 3

Data preparation and visualization in Python

26 November 2023

LECTURE PLAN

the experts
school of analytics

1. Working with data in Python
2. Data preparation
    a) Pandas
    b) Pandas vs SQL
3. Data visualization
    a) Matplotlib
    b) Seaborn
    c) Plotly
4. Practice in Jupyter Notebook

DATA PREPARATION

PANDAS

the experts
school of analytics

Pandas is a Python library used for working with data sets.

It has functions for

- analyzing,
- cleaning,
- exploring,
- and manipulating data.

The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

DATA PREPARATION

PANDAS DATAFRAME

the experts
school of analytics

	id	name	date	manner_of_death	armed	age	gender	race	city	state	signs_of_mental_illness	threat_level	flee	body_camera
0	3	Tim Elliot	2015-01-02	shot	gun	53.0	M	Asian	Shelton	WA	True	attack	Not fleeing	False
1	4	Lewis Lee Lembke	2015-01-02	shot	gun	47.0	M	White	Aloha	OR	False	attack	Not fleeing	False
2	5	John Paul Quintero	2015-01-03	shot and Tasered	unarmed	23.0	M	Hispanic	Wichita	KS	False	other	Not fleeing	False
3	8	Matthew Hoffman	2015-01-04	shot	toy weapon	32.0	M	White	San Francisco	CA	True	attack	Not fleeing	False
4	9	Michael Rodriguez	2015-01-04	shot	nail gun	39.0	M	Hispanic	Evans	CO	False	attack	Not fleeing	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4890	5916	Rayshard Brooks	2020-06-12	shot	Taser	27.0	M	Black	Atlanta	GA	False	attack	Foot	True
4891	5925	Caine Van Pelt	2020-06-12	shot	gun	23.0	M	Black	Crown Point	IN	False	attack	Car	False
4892	5918	Hannah Fizer	2020-06-13	shot	unarmed	25.0	F	White	Sedalia	MO	False	other	Not fleeing	False
4893	5921	William Slyter	2020-06-13	shot	gun	22.0	M	White	Kansas City	MO	False	other	Other	False
4894	5924	Nicholas Hirsh	2020-06-15	shot	gun	31.0	M	White	Lawrence	KS	False	attack	Car	False

4895 rows × 15 columns

DATA PREPARATION

PANDAS VS SQL

the experts
school of analytics

SQL	Python
`SELECT * FROM titanic_test_data WHERE pclass = 1`	`titanic_df[titanic_df.pclass == 1]`
`SELECT * FROM titanic_test_data WHERE pclass = 1 OR pclass = 2`	`titanic_df[(titanic_df.pclass == 1) | (titanic_df.pclass == 2)]`
`SELECT * FROM titanic_test_data WHERE pclass IN (1,2)`	`titanic_df[titanic_df.pclass.isin([1,2])]`
`SELECT name FROM titanic_test_data WHERE pclass = 1 AND gender = "male"`	`titanic_df[(titanic_df.pclass == 1) & (titanic_df.gender == "male")]["name"]`
`SELECT name, age FROM titanic_test_data WHERE pclass NOT IN (1,2)`	`titanic_df[~titanic_df.pclass.isin([1,2])] [["name","age"]]`

Pandas and **SQL** are both used for selecting filtering and grouping data.

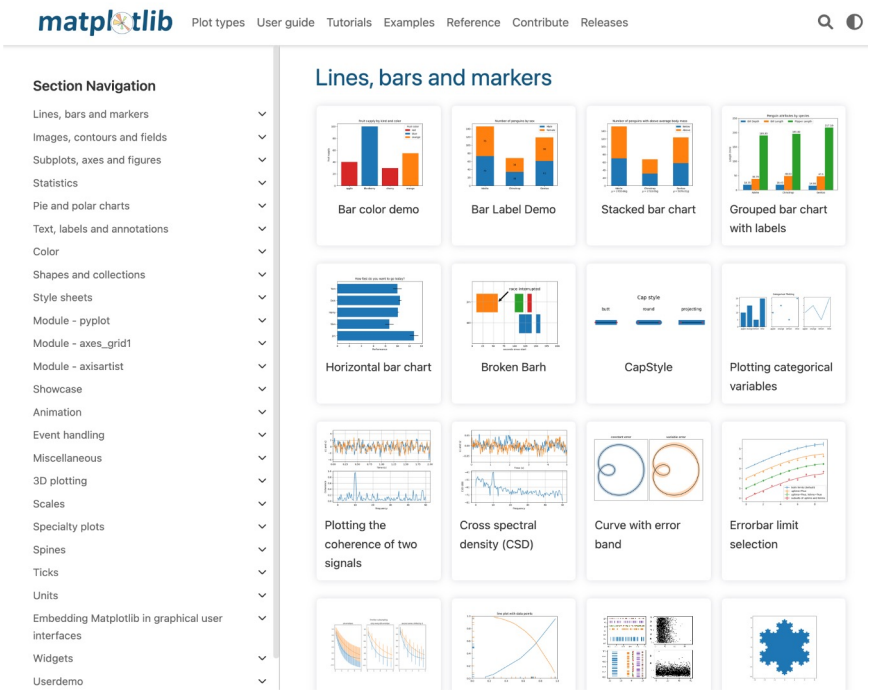
While Pandas has more functionality, SQL performs better on large data sets.

As a part of Python - pandas is very convenient when you need not only extract data, but analyze it.

DATA VISUALIZATION

MATPLOTLIB

the experts
school of analytics

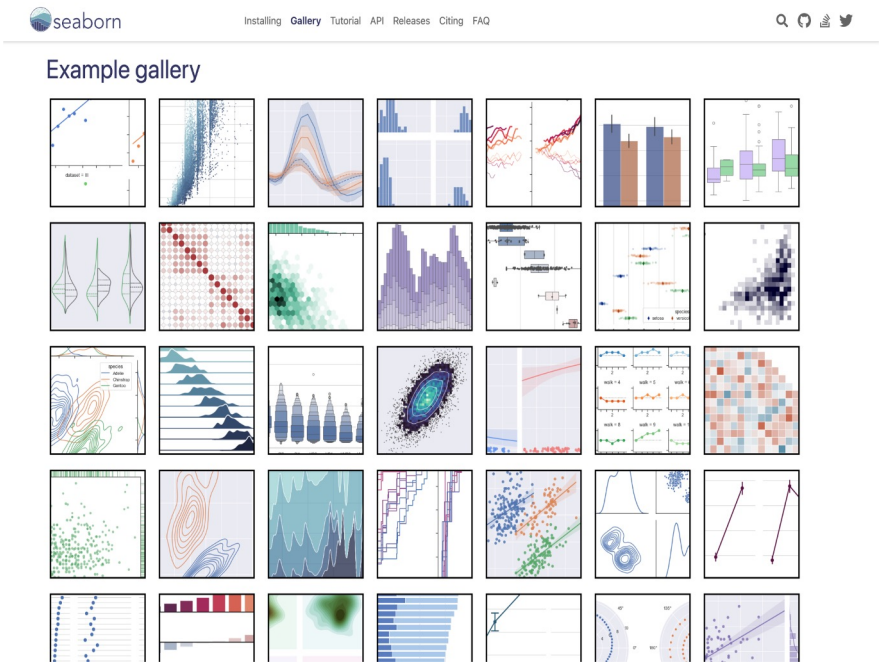


Matplotlib is the most common visualization tool. It perfectly balances simplicity and functionality and has lots of patterns.

DATA VISUALIZATION

SEABORN

the experts
school of analytics

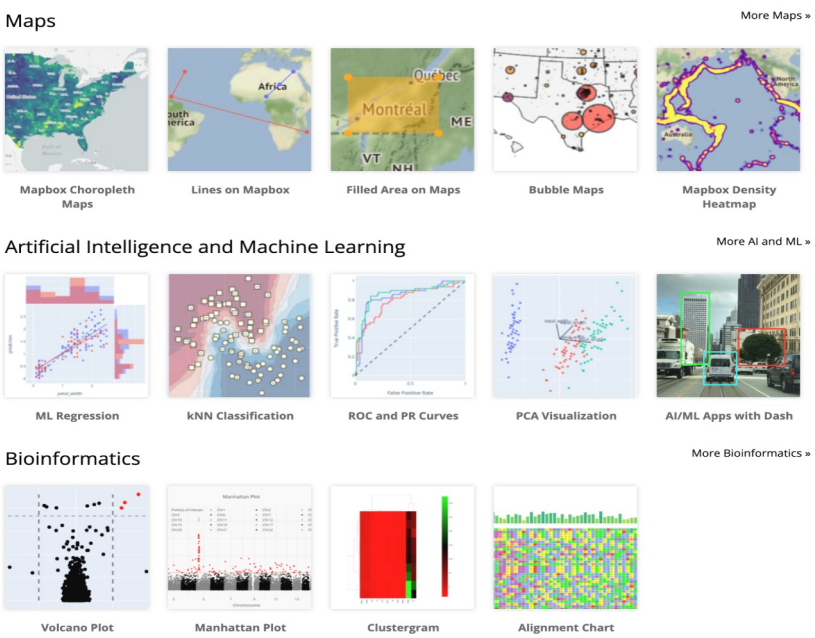


Seaborn is an advanced statistical visualization tool. It has lot's of ready made visualizations in it's gallery that you can reuse for your research.

DATA VISUALIZATION

PLOTLY

the experts
school of analytics



Plotly is the most functional library and it's more dashboard based tool. It is very useful when you want to monitor the same data over time.

PRACTICE IN JUPYTER NOTEBOOK

the experts
school of analytics

Let’s practice in Jupyter Notebook!