Introduction to Data Science

with Python

Alexis Bogroff

May 26, 2022

Presenter

Alexis Bogroff
Lecturer and Mentor in Data Science
at Paris 1 Panthéon-Sorbonne, ESILV,
Openclassrooms, EM-Lyon

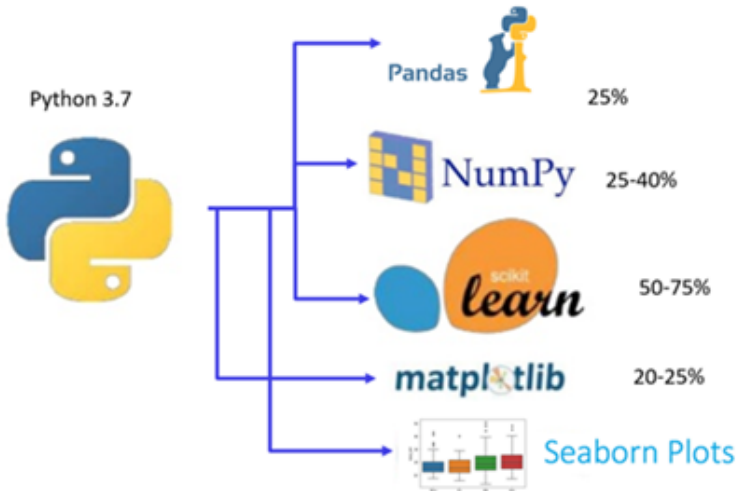
Presenter

Alexis Bogroff
Lecturer and Mentor in Data Science
at Paris 1 Panthéon-Sorbonne, ESILV,
Openclassrooms, EM-Lyon

- 4 years Teaching Assistant and lecturer in VBA, Python for finance, SQL, Data Analysis and Data Science
- 9 months Researcher Assistant at Paris 1 Panthéon-Sorbonne within H2020 European Project
- 1 year Data Scientist at Pléiade Asset Management

Remember from this course

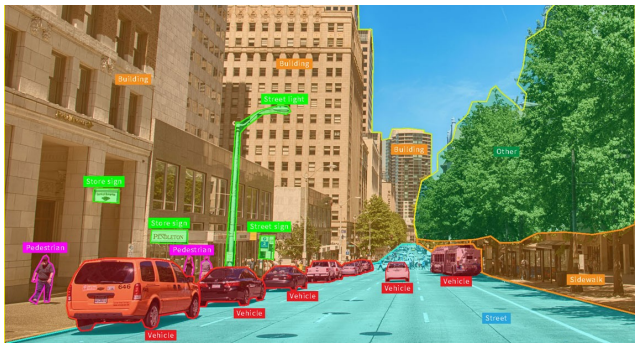
- Overview
- Python
- Pandas
- Data Analysis
- Data Management with Pandas
- Data Visualization
- Predictions

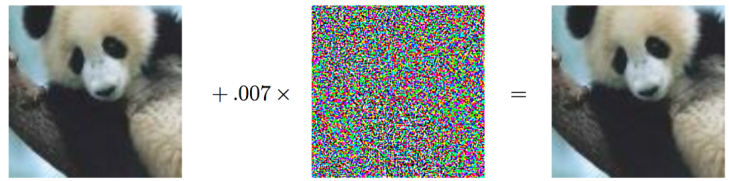


Overview

- Artificial Intelligence, Machine Learning, Deep Learning
- Computer Vision, Recommander systems, NLP
- Programming
- Data Analysis
- Issues, risks, ethics, RGPD
- Ressources

```python
def __init__(self):
    # Booleans
    self.all_cp_registered = False
    self.strats_assessed = False
    # Ckpts map
    self.checkpoints = {
        'count': 0,
        'map': {},
        }

    # Limits
    self.TOLERANCE_ANGLE_STRAIGHT = 15
    self.THRESHOLD_DO_STRAT = .8
    self.STRAT_CIRCLE_TOLERANCE_VAR = 10

    # Structures detected
    self.general_structure = None


def assess_optimal_trajectory(self):
    """
    Assess optimal trajectory
    returns a list of coordinates to follow from initialcp to initialcp
    """
    if self.general_structure == 'circle':  # equi triangle

        # generate 3 intermediate points supposedly on the circle
        # ------------------------------------------------------
        # A is 1, B is 2, C is 3

        # Retrieve pcs positions
        A = self.get_cp_position(1)
        B = self.get_cp_position(2)
        C = self.get_cp_position(3)

        # Compute distance of segments
```





"panda"

57.7% confidence

noise

"gibbon"

99.3% confidence

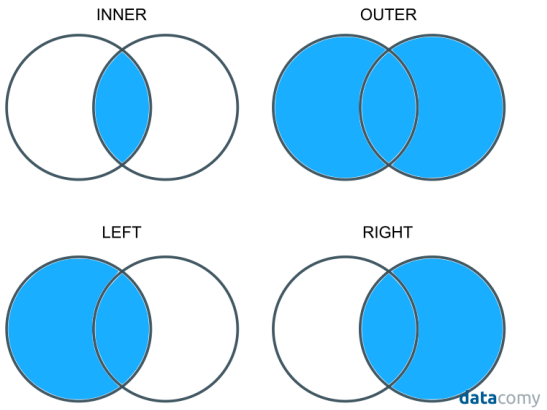
Python

- Programming environment
- Essentials in Python language
  - Data structures
  - Control structures
  - Functions
  - Objects
- Goog practices, coding conventions
- Methods to learn programming


```
# Text and numbers
12          # int (integer)
1.5         # float
'hola'      # str (string)
"hola"
"""hola"""

# Iterables
[42, 58, 209, 42]  # list
(42, 58, 209, 42)  # tuple
{42, 58, 209}      # set
{'name': ['akiko', 'julie'], 'age': [12, 43]}  # dict (dictionary)
```

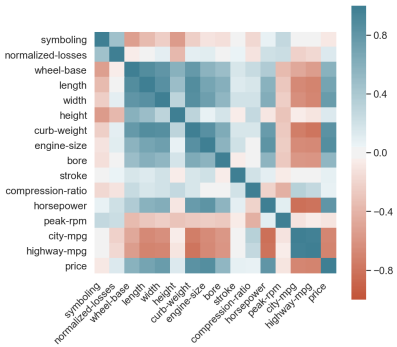
Pandas

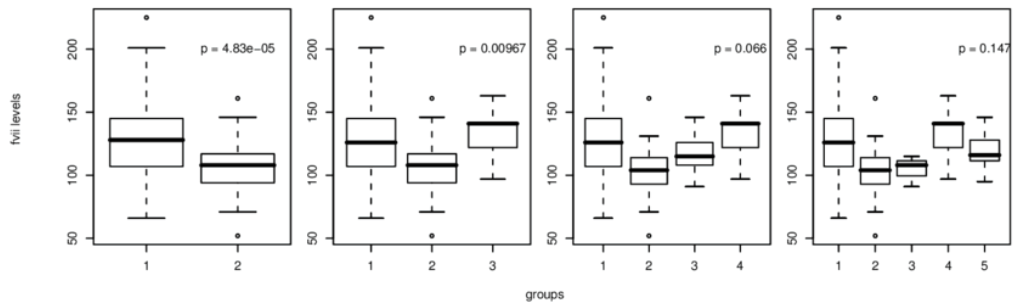
- Core objects
- Masks / Filters
- Basic methods (info, describe) Apply, vectorial operations
- Other useful methods (sort_values, groupby, isna)
- Graphs (.plot, .scatter.plot, .plot.bar, .hist)
- merging DataFrames the right method (outer, indicator=True)
- Pandas profiling



Data Analysis

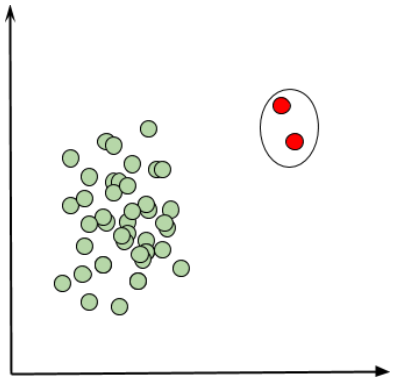
- Measures: centrality, dispersion, IQR
- Pattern analysis
  - Univariate
  - Multivariate
- Data type
  - Qualitative, Quantitative
  - Numbers (Times Series), text, images, music
  - Linear, non-linear
- Correlations
- Statistical laws and tests





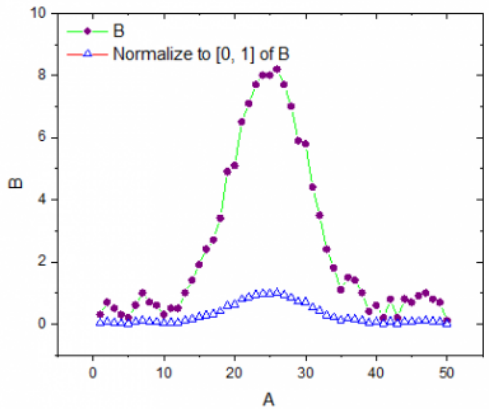
Data Management with Pandas

- Features selection
  - Drop columns, rows (duplicates, constants, useless)
  - Multicolinearity
- NA imputation
  - Missing as the information
  - Reconstruction methods
- Outliers
- Features transformation
  - Logarithm
  - Center and reduce
- Merge, concatenate tables
- Feature engineering
  - One-hot encode
  - Group
  - Filter



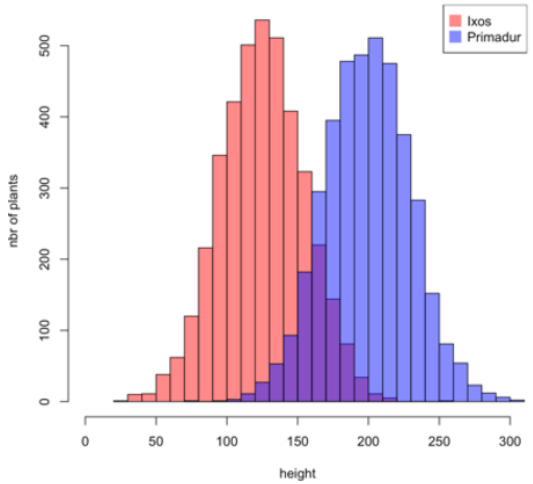
Color
Red
Red
Yellow
Green
Yellow

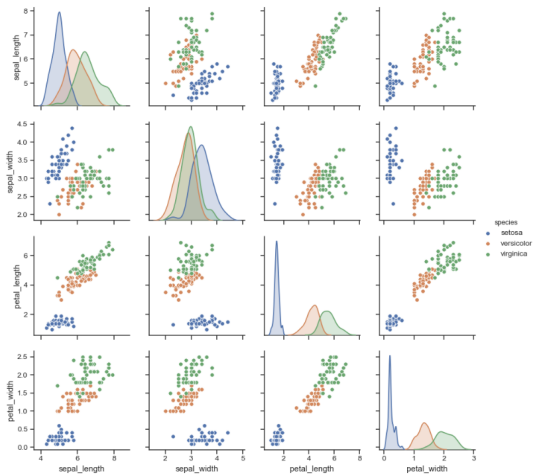
Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1



Data Visualization

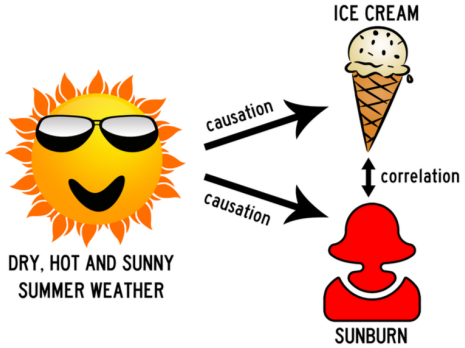
- Why, use cases
- Graph types for univariate analysis
  - Histograms
  - Line plots
  - Lorentz Curve
- Graph types for multivariate analysis
  - Scatter plots
  - Heatmaps
  - Pairplots
- Libraries
  - Matplotlib
  - Seaborn
  - Dash

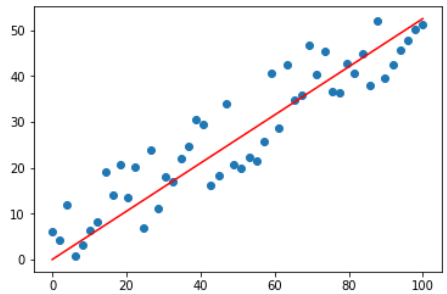


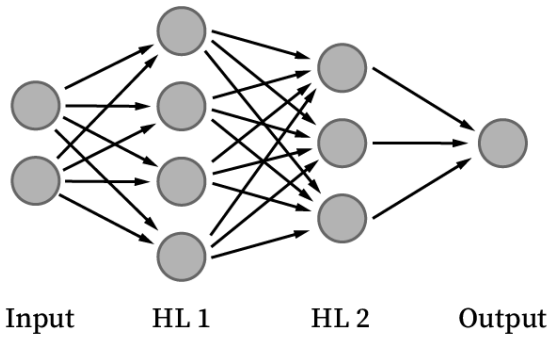


Predictions

- Correlation vs causality
- Use cases
- Regression
- Classification
- Problems types
  - Supervised learning
  - Unsupervised learning
- Models
  - Basic models
  - Training
  - Optimization
- Transfer Learning







How to learn more

- Code:
  - Project (personal or open source[1])
  - Stack Overflow
  - Coding Game
  - Peers
- Data Science:
  - Project
  - MOOC (Andrew Ng. Coursera - *Machine Learning*)
  - Youtube channels
  - Towards Data Science
  - Conferences (retransmitted)
  - Blogs of AI research Labs (GAFAM, OpenAI)
  - Research Papers
  - Books
- Be confident: the harder it is, the stronger your comprehension[2]

[1] Data for Good

[2] "Make It Stick: The Science of Successful Learning" - Peter C. Brown

Job Market

- Your current job might need your skills (automating, analysis, clustering, prediction)
- Data Analyst
- Data Scientist
- Data Engineer
- Machine Learning Engineer
- Project Manager
- Researcher
- Developer
- Sales
- Better communication with developers

Questions

- What are your questions?