Mastering Pandas: An in-depth Guide in Data Science Techniques for Researchers

Lecture 1 – 5 April 2023

Overview of Pandas data structures, data manipulation, filtering, and aggregation

NITheCS
National Institute for
Theoretical and Computational Sciences

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

George Leonard defines mastery as a continual journey, and there's no end or "perfection" of the skill—rather, the practice is the point in and of itself.

CSIR
Touching lives through innovation

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

NITheCS
National Institute for
Theoretical and Computational Sciences

Agenda

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- About Me
- Admin
- Mini-School Overview
- Data
- EDA
- Pandas Part 1
- Useful Resources

CSIR
Touching lives through innovation

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

NITheCS
National Institute for
Theoretical and Computational Sciences

About Me

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- Binjamin (pronounced Binyamin) is the Lead Software Engineer at the Centre for High Performance Computing (CHPC), a research centre that forms part of the CSIR
- BEng in Mechatronic Engineering from Stellenbosch University and a MEng degree in Radar Digital Signal Processing from UCT
- Worked in industry focusing on the fields of electronic engineering, data science, and software development
- Currently lead the software engineering architecture, and the design and development of HPC support software
- Lead the annual Coding Summer School which focuses on training researchers in fundamental programming skills to assist with their research.
- Research interest: DSP for IoT, and NLP for African languages – see https://www.masakhane.io/

NITheCS
National Institute for
Theoretical and Computational Sciences

Admin

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- Form - https://forms.office.com/r/QJs4SbGXBB
- All code and lessons – github repo: https://github.com/kode2go/nithecs/tree/main/lecture_01

Mini-School Overview

Timetable:

- Lecture 1 (5 April) - Overview of Pandas data structures, data manipulation, filtering, and aggregation*
- Lecture 2 (12 April) – Aggregation*, data wrangling, merging, joining, and time series analysis
- Lecture 3 (19 April) - Data visualization, statistical modelling, and managing varied datasets types
- Lecture 4 (26 April) - Advanced data analysis techniques, and real-world applications

General Topics*:

- EDA
- Big Data
- Data Pipelines

Overview:

- This mini-school is aimed at those who are familiar with Python and Pandas.
- By the end of the mini-school, attendees should have learnt skills to analyse their datasets more effectively and derive meaningful insights.

Data

https://www.data-to-viz.com/

NITheCS
National Institute for
Theoretical and Computational Sciences

Data

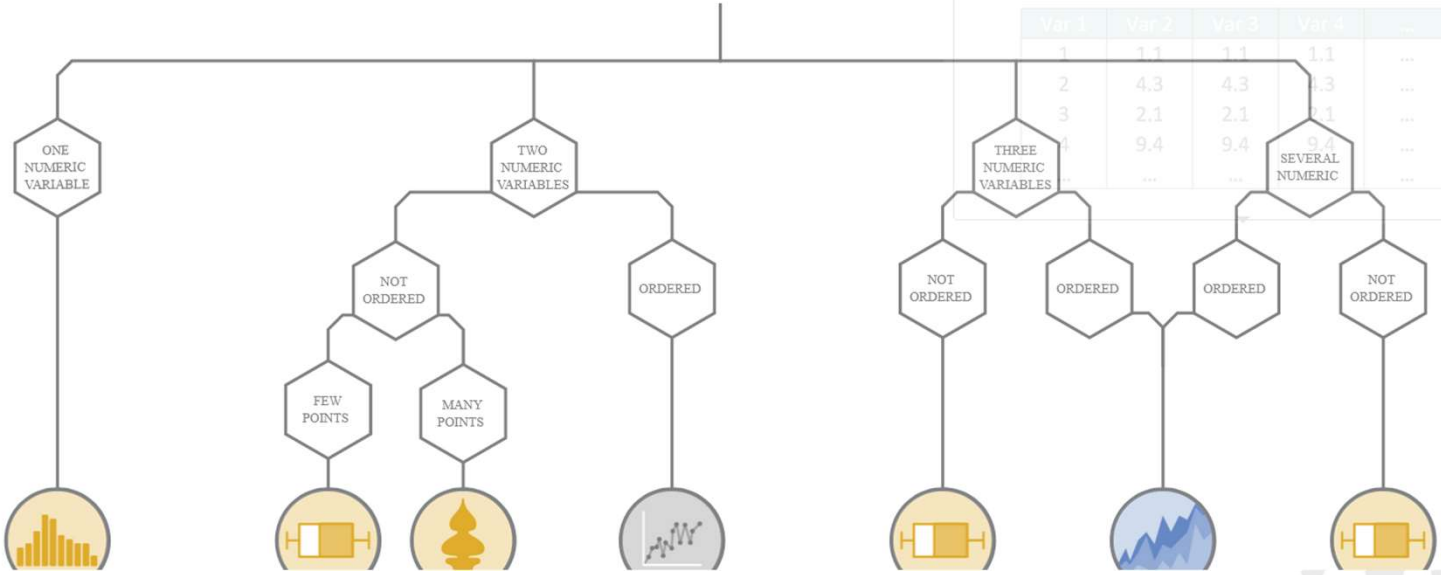
NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

from Data to Viz

EXPLORE STORY ALL CAVEATS POSTER ABOUT CONTACT

What kind of data do you have? Pick the main type using the buttons below. Then let the decision tree guide you toward your graphic possibilities.

Numeric | Categoric | Num & Cat | Maps | Network | Time series



CSIR
Touching lives through innovation

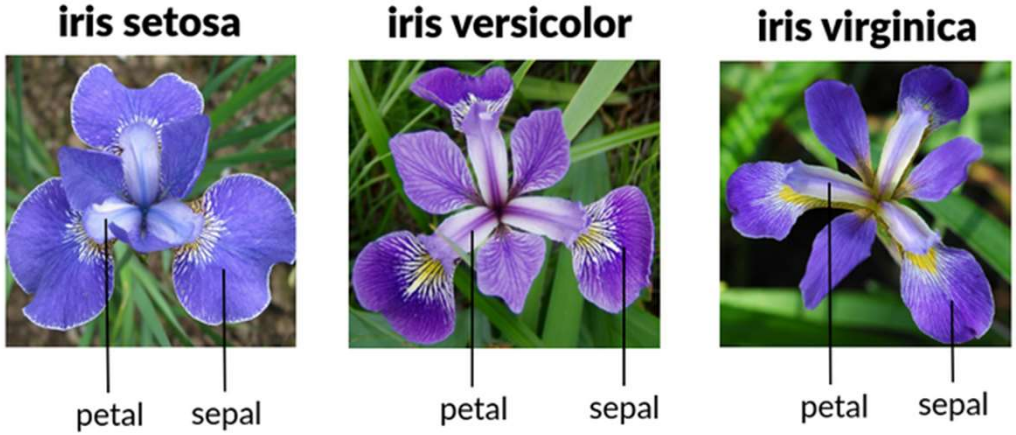
science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

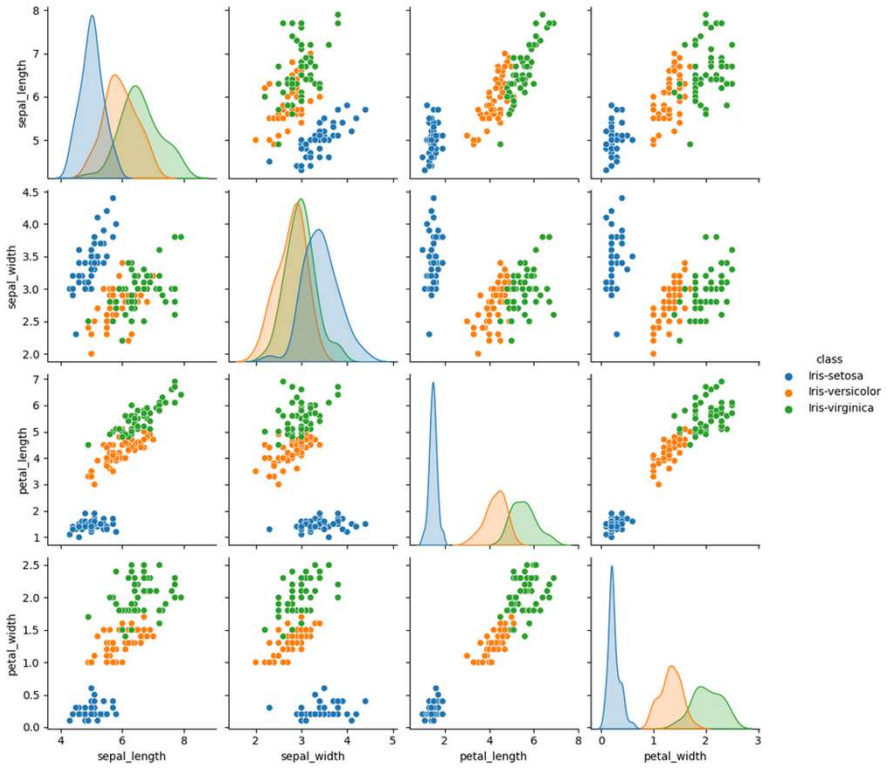
NITheCS
National Institute for
Theoretical and Computational Sciences

EDA

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- Exploratory data analysis on the Iris dataset
- 1936 by the statistician and biologist Ronald Fisher





CSIR
Touching lives through innovation

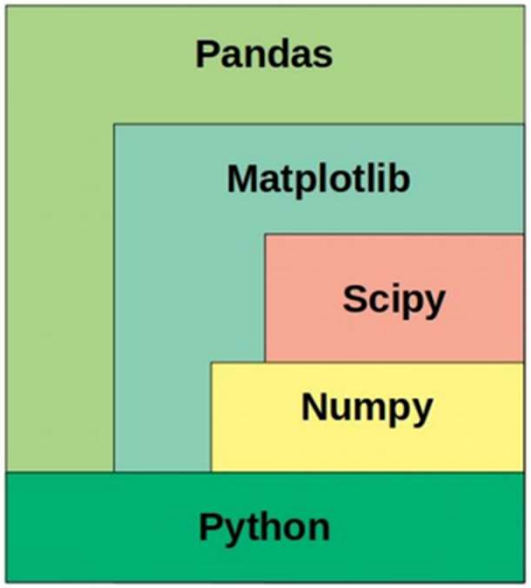
science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

NITheCS
National Institute for
Theoretical and Computational Sciences

Pandas Part 1

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- Pandas – "Panel Data" – econometrics term for data sets
- Numpy -> Pandas
- Pandas Data Structures and Functionality
- Pandas Profiling



CSIR
Touching lives through innovation
science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

NITheCS
National Institute for
Theoretical and Computational Sciences

Useful Resources Fundamentals

NICIS
NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM
CHPC

- DataCamp (https://www.datacamp.com/)
- Kaggle (https://www.kaggle.com/)
- Towards Data Science (https://towardsdatascience.com/)
- Analytics Vidhya (https://www.analyticsvidhya.com/)
- Real Python (https://realpython.com/)
- Dataquest (https://www.dataquest.io/)
- Python Data Science Handbook (https://jakevdp.github.io/PythonDataScienceHandbook/)