DATA WITH PANDAS

KARIM

About me

- Abdulkarim Alabdulkarim

- Who?
  - Full stack Engineer
  - Professionally 6 years developer
  - Masters on Information production and systems

Jupyter Notebook

- What?
  - A virtual notebook
  - Markdown and Python
  - Useful for visualizing and quick coding

Jupyter Notebook

- Where?
  - Online
    - Kaggle notebook, Google colab
  - Local
    - JupyterLab, JupyterNotebook, VS Code

What is Pandas

- A Python library for data manipulation and analysis.

- Exploring
- Analyzing
- Cleaning

Pandas Data Structures

- Series.
  - Think row or column

- DataFrame
  - Think table

Show time

- Refer to the “intro” notebook.

When Pandas

- Exploratory Data Analysis (EDA).
- Simplify manual data manipulation tasks.
- Patch Data Manipulation.

Why Pandas

- Open Source
- Is the de facto standard
- Seamlessly integrated with data visualization libraries
  - Seaborn
  - Plotly

Show time!

- Kaggle : Japanese Universities Dataset

- Refer to “EDA” notebook.
- Refer to Data Card

You can also

- Sort
- GroupBy
- Merge
- More Visualization
- Read and write different formats
- Read and Write to SQL
- Webscrape (read html)

You can also

- Speed things with cython
- Use GPU with cudf
- Parallel compute with dask
- Use pyarrow to improve performance

How Pandas

- Three Solutions
  - Loop
  - Apply
  - Vectorize

- Refer to “ThreeApproaches” notebook.
  - Or watch this video instead.

Alternatives

- Polars
  - Designed for speed, with Rust.
  - Lazy Frames.
  - Supports pandas (import and export)

- Spark
  - Designed for actual big data that doesn’t fit the memory.
  - Designed for distributed processing .

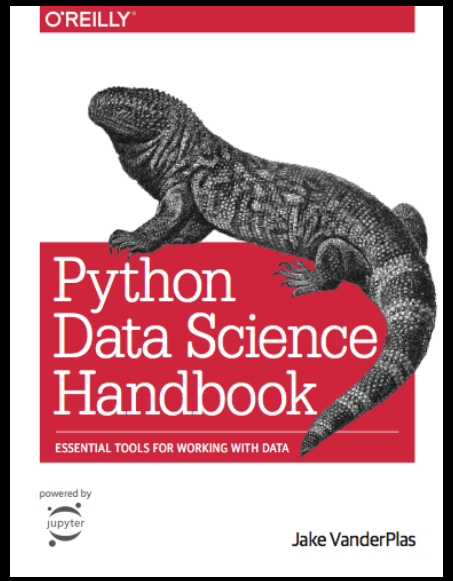
Alternatives

- R
- Juila
- Excel

Resources

- Video:
  - Rob Mulla
- Playlist:
  - Pandas - Corey Schafer
  - 7-day bootcamp: Nick Wan

- Courses:
  - Kaggle Learn
  - UC Berkely – course notes



THANK YOU

ANY OTHER QUESTIONS?