
DNB PUBLIC
Open Source Workshop
Data analyse met (Python) Pandas
24 september 14:00 – 17:00
DeNederlandscheBank
EUROSYSTEEM

Data analysis met Pandas

14:00	Theorie/achtergrond informatie
14:15	Gebruik Pandas
15:15	Voorbereiden case study
15:30	Case study uitvoeren
16:45	Resultaten bespreken

Wat is Pandas?

"Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language."

Een Python package, dat gebouwd is boven op het NumPy package

- Beperkt het schrijven van code Python voor het ontsluiten, manipuleren en verwerken van gestructureerde data
- Makkelijke visualisatie via Matplotlib
- Slechts 2 data structuren: Series en DataFrame

Pandas DataFrame

Tabel met 1 of twee dimensies:

- Series: 1-dimensionaal
- DataFrames: 2-dimensionaal (axis=0 zijn de rijen, axis=1 zijn de kolommen)
  - DataFrame is opgebouwd als meerdere series met 1 index

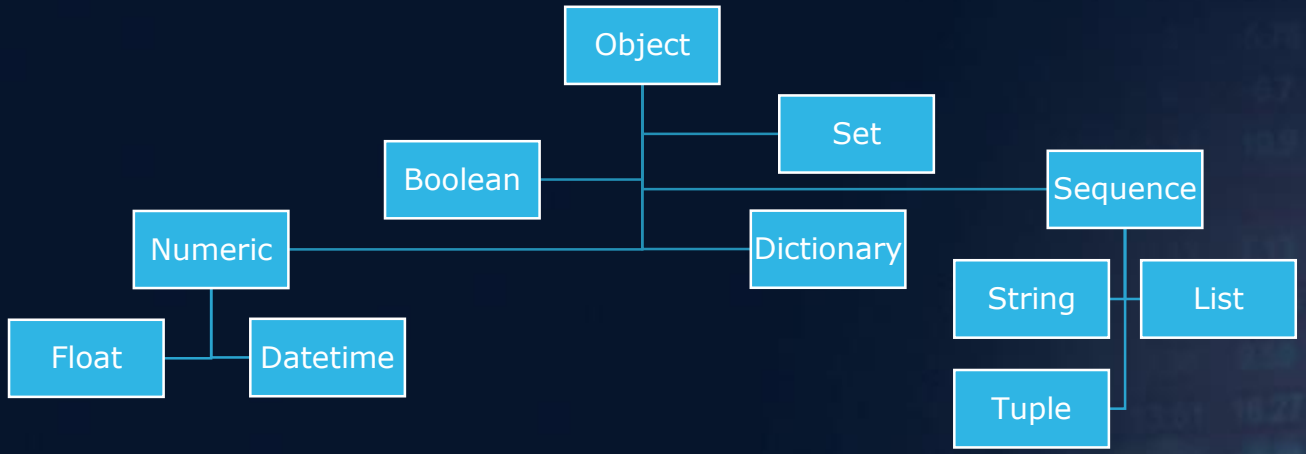
Gebruik van makkelijk te interpreteren namen voor de kolommen/rijen

- Kolommen ≃ Variabelen
- Rijen ≃ Observaties

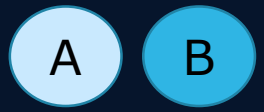
Overige functionaliteit:

- Lege data velden worden NaN (np.nan).
- Toevoegen/verwijderen van kolommen in een DataFrame moet makkelijk gaan
- Data samenvoegen gebaseerd op de indexes
- Zijn niet gemaakt om van vorm te veranderen

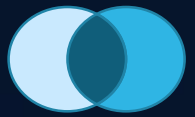
Datatypes in Python



Data combineren



Tabellen A en B willen we samenvoegen op basis van gelijke waarden uit specifieke kolommen.



Outer
Alle datapunten meenemen

Inner
Alleen de overlappende data

Left/right
Alle datapunten uit de linker/rechter helft meenemen en de overlappende data

Als de matches niet uniek zijn zullen alle combinaties terug komen in het samengevoegde dataframe (many-to-many), dit kan leiden tot hele grote datasets

Data analysis met Pandas: Stappen

Data analysis met Pandas: Case study

Covid maatregelen in VS:
- Paycheck Protection Program: Leningen aan kleinbedrijf om mensen in dienst te houden

Zijn deze leningen op de juiste plek terecht gekomen?
- Op de locaties waar de afname in mobiliteit het grootst was

Extra:
- In de sectoren waar het aantal uren dat er gewerkt is het meeste achteruit ging

Data analysis met Pandas: Case study

Databestanden:
- PPP_Loan_Data.csv De individuele leningen uit het PPP (voor grote leningen is er een loan_range)
- Global_Mobility_Report.csv Dagelijkse Google mobility indexes vanaf februari
- Uszips.xlsx Lijst met postcodes en sensuscodes voor de VS

Code:
- Workshop_code.ipynb De code die we behandeld hebben tijdens de workshop, handig te gebruiken als 'cheatsheet'
- CaseStudyBegin.ipynb Code om de data in te inlezen voor Colab

Colab: *https://github.com/mnijhuis-dnb/open_source_workshop*
RANS: *G:\Algemeen_Kopieergebied\OSW_Pandas*
RAN: Data op Sharepoint

Extra:
- US_Hours_Worked.json Geaggregeerde aantal gewerkte uren per week per maand vanaf 2019 per sector in duizendtallen seizoensgecorrigeerd
- NAICS.json Sectornamen die bij de sectorcodes (NAICS) horen