

data

adam okulicz-kozaryn

`adam.okulicz.kozaryn@gmail.com`

this version: Tuesday 28th January, 2025 12:34

outline

data types

commenting and syntax

outline

data types

commenting and syntax

data basics

- dataset is a matrix
- cols are vars, rows are obs (U/As)
- vars are characteristics of obs
- eg 'edu', 'age', and 'inc' are vars and persons are obs
- each row is a separate person

tens of data types/formats/files

- a basic distinction:
 - software-specific binary files (.dta, .sas7bdat, .sav)
 - generic text files (.txt, .dat, .csv, .tab)
- just google it! eg pandas read csv', pandas export spss'

databases/sql; internet/api

- most data are in databases
 - Oracle, MySQL, NoSQL, MsSQL, etc
 - can use Python to pull directly from databases
- APIs
 - i'd just use API instead of database/sql
 - there is API section towards the end of pandas.ipynb
 - only if for some reason API doesn't work, i'd use database/sql connection—clunky
 - and can also scrap with BeautifulSoup etc

outline

data types

commenting and syntax

make comments in your code

- for each class we have ipynb file
- make comments in ipynb—you will run ipynb not docx etc
- if you do not make comments, you will forget
- use very handy keywords like “TODO”, “BUG”, “LATER”, “FIXME”
- and ctrl-f

commenting

- have preamble (notes, install packages, etc)
- `#comment`

`' ' 'comment`

`block ' ' '`