data

adam okulicz-kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Friday 24th January, 2025 13:33

outline

misc

data types

commenting and syntax

outline

misc

data types

commenting and syntax

outline

misc

data types

commenting and syntax

data basics

- dataset is a matrix
- cols are variables (var), rows are observations (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

dozens of data types/formats/files

- there are
- 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' etc

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 doesnt work, i'd use database/sql connection–clunky
  - and can also scrap with beautifulsoup etc

outline

misc

data types

commenting and syntax

make comments in your code

- for each class we will have ipynb file
- make comments in the electronic code files – you will run electronic files not the printout
- 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 '''
```