

DS2000
1/17 (Tue)

Admin

- HW1 due Fri 9pm
- Quiz 1 4:30 today - 9:50 am Fri

Agenda

1. Data Sources
2. File Processing
3. Python

Recap ~ result?

7	//	2	3
7	/	2	3.5
7	%	2	1
6	%	2	0
8	%	2	0

`print(2/3)` 0.66666666

`print(round(2/3, 2))`

0.67
//, % usually ints
ok with floats

`7.0 // 2.5` 2

1. Data Sources

1. gather data
2. computation
3. communication

from user (input)

from a file

↳ ~.txt

DS Integrity

- How was the data collected?
- Do we have confidence in accuracy?
- What can I share?

Today's data:

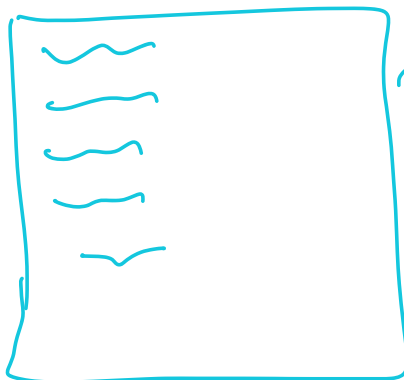
- Salaries of Boston employees
- Source: data.boston.gov

Salary, overtime, title, name

↳ don't share

2. File Processing

Our file



boston-salaries.txt

↳ one number per line
Salary from 2021

Q: What do we want to know?

- differential in salaries (highest vs. 2nd highest) ...
- avg
- median
- range
- min
- max
- upper/lower quartile
- std dev
- change over years
- mode

In Python: no input!

- name of file (constant)
- open the file

with open ("~", "r") as infile:

↳ python vocab

↳ read (string)

↳ name of file (string)

- read a line (one at a time)

x = infile.read line()

↳ x is variable (string)

↳ read one line

x = int(infile.readline())

↳ x is an int

- save .txt file in same folder as .py file