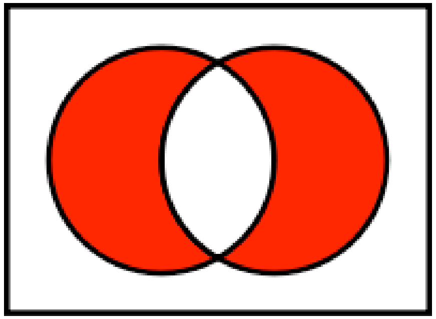
CompSci 6
Introduction to Computer Science



October 18, 2011

Prof. Rodger

Announcements

- Read for next time Chap. 10
- Reading Quiz on Blackboard
  - Due before class next time
- Assignment 4 is due Thursday

- See article “Big Data, Big Responsibility”

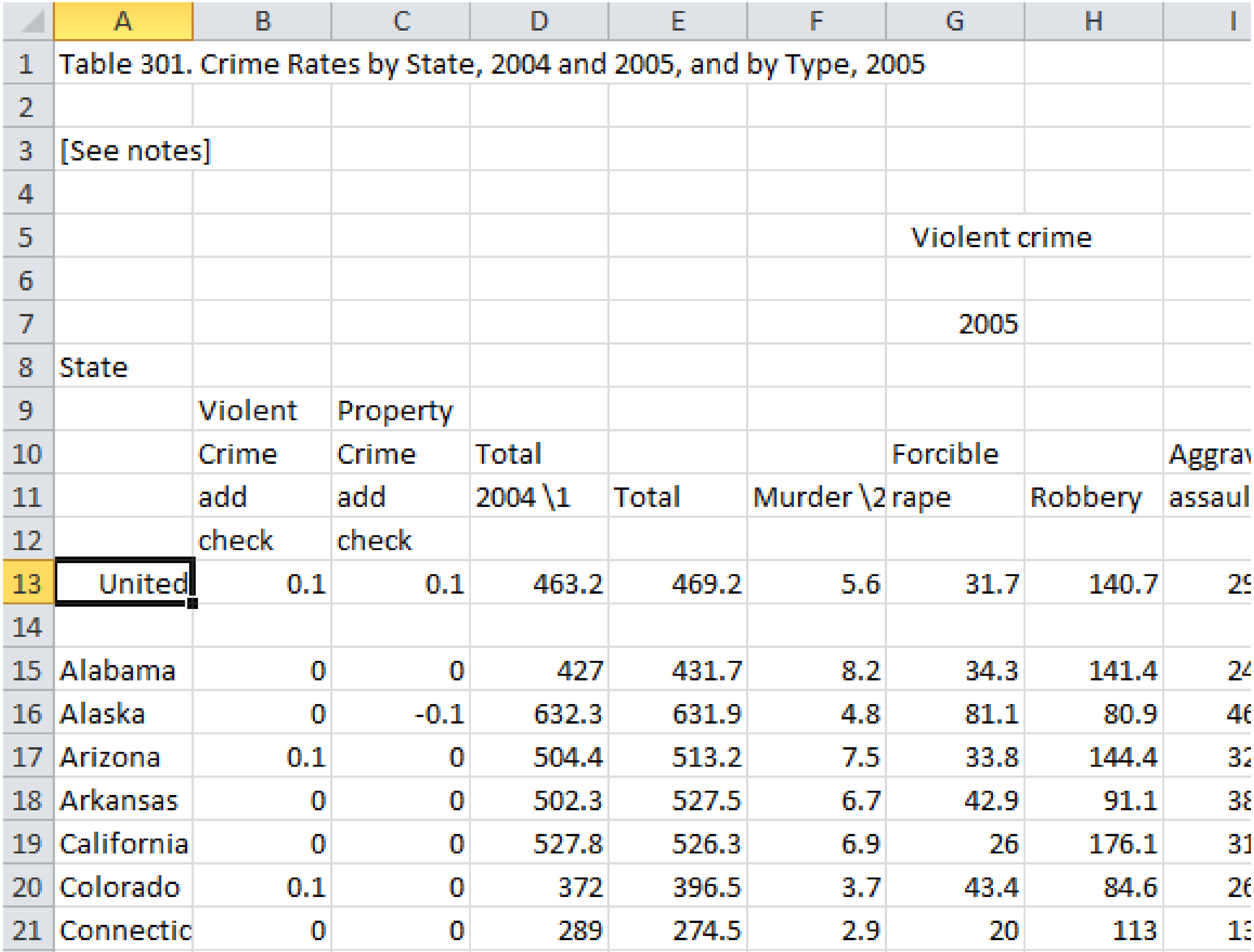
Assignment 4

- Two problems to solve – 2 Python modules
- Create a third Python module
  - Identify common functions and put in the third module – maximize this!
  - Example function – reading from a file
  - Use import to read this module into the two programs
- Questions?

Back to Example: Which state has the highest murder rate?

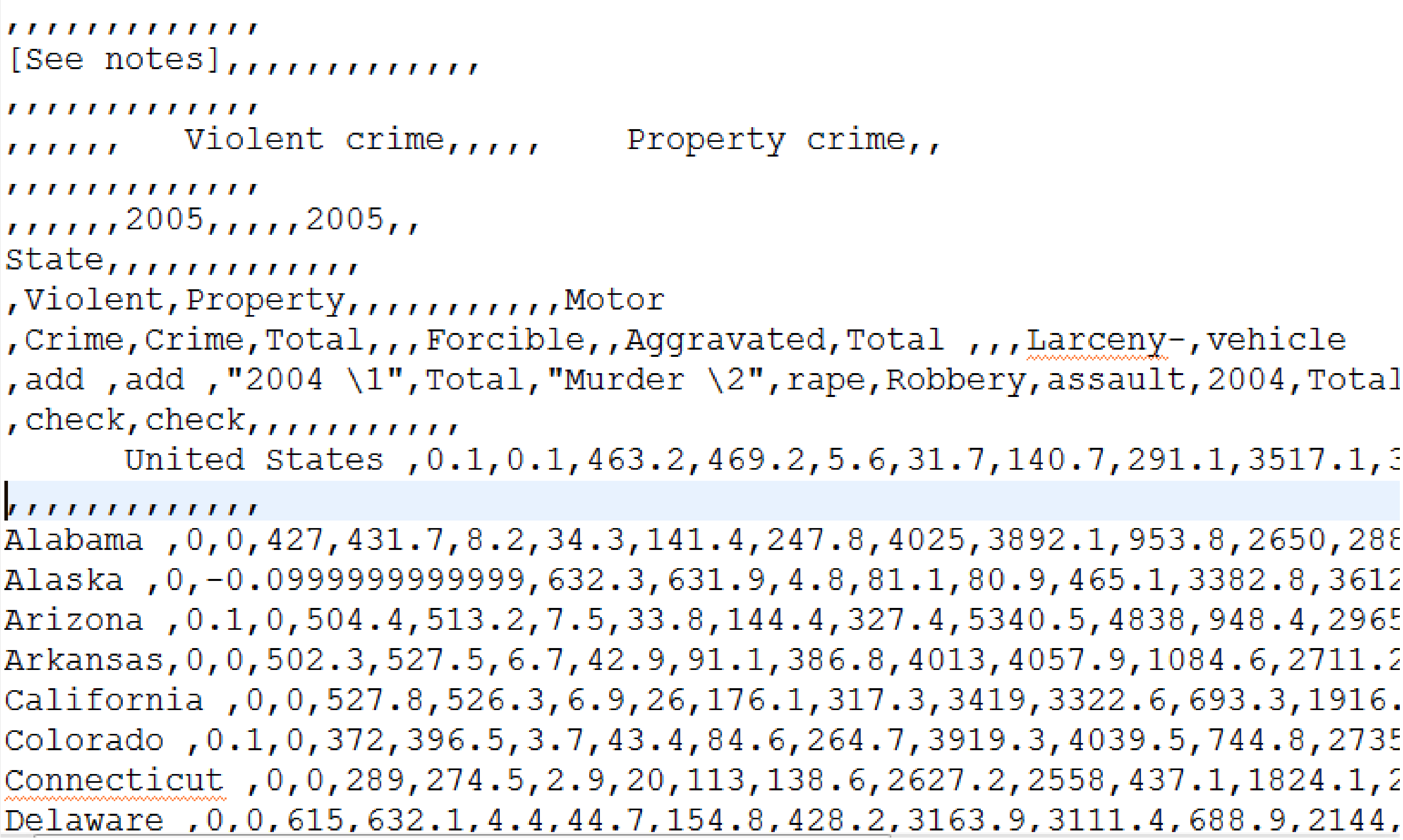
- Get datafile
  - Infochimps – Crime rate by states 2004 and 2005
- Csv file – separator is comma
- Data may be messy – look at it
- How do we figure out which state had the highest/lowest murder rate?

Data as a spreadsheet



	A	B	C	D	E	F	G	H	I
1	Table 301. Crime Rates by State, 2004 and 2005, and by Type, 2005								
2									
3	[See notes]								
4									
5							Violent crime		
6									
7							2005		
8	State								
9		Violent	Property						
10		Crime	Crime	Total			Forcible		Aggra
11		add	add	2004 \1	Total	Murder \2	rape	Robbery	assaul
12		check	check						
13	United	0.1	0.1	463.2	469.2	5.6	31.7	140.7	2
14									
15	Alabama	0	0	427	431.7	8.2	34.3	141.4	24
16	Alaska	0	-0.1	632.3	631.9	4.8	81.1	80.9	46
17	Arizona	0.1	0	504.4	513.2	7.5	33.8	144.4	32
18	Arkansas	0	0	502.3	527.5	6.7	42.9	91.1	38
19	California	0	0	527.8	526.3	6.9	26	176.1	31
20	Colorado	0.1	0	372	396.5	3.7	43.4	84.6	26
21	Connectic	0	0	289	274.5	2.9	20	113	13

Data as a .csv file



```
,,,,,,,,,,,,,
[See notes],,,,,,,,,,,,,

,,,,,,,,,,,,,
,,,,,,   Violent crime,,,,,    Property crime,,

,,,,,,,,,,,,,
,,,,,,2005,,,,,2005,,
State,,,,,,,,,,,,,
,Violent,Property,,,,,,,,,,,Motor
,Crime,Crime,Total,,,Forcible,,Aggravated,Total ,,,Larceny-,vehicle
,add ,add ,"2004 \1",Total,"Murder \2",rape,Robbery,assault,2004,Total
,check,check,,,,,,,,,,,
      United States ,0.1,0.1,463.2,469.2,5.6,31.7,140.7,291.1,3517.1,3
,,,,,,,,,,,,,
Alabama ,0,0,427,431.7,8.2,34.3,141.4,247.8,4025,3892.1,953.8,2650,288
Alaska ,0,-0.0999999999999,632.3,631.9,4.8,81.1,80.9,465.1,3382.8,3612
Arizona ,0.1,0,504.4,513.2,7.5,33.8,144.4,327.4,5340.5,4838,948.4,2965
Arkansas,0,0,502.3,527.5,6.7,42.9,91.1,386.8,4013,4057.9,1084.6,2711.2
California ,0,0,527.8,526.3,6.9,26,176.1,317.3,3419,3322.6,693.3,1916.
Colorado ,0.1,0,372,396.5,3.7,43.4,84.6,264.7,3919.3,4039.5,744.8,2735
Connecticut ,0,0,289,274.5,2.9,20,113,138.6,2627.2,2558,437.1,1824.1,2
Delaware ,0,0,615,632.1,4.4,44.7,154.8,428.2,3163.9,3111.4,688.9,2144,
```

Problem: Which state has highest murder rate? Lowest rate?

1. Read in the .csv file -> processFile(file)
2. Find the row and column where “Murder” starts -> getColumnRow
3. Get the one column of items for the “Murder” column
4. Convert/clean the column into float numbers
5. Find the max/min values
6. Find the states with max/min

Step 1 – Process data file

- processFile(file) returns a lists of lists
- Each inner list is one row of the data
- Example for Row 15

  Alabama ,0,0,427,431.7, …

  to

  [“Alabama”, “0”, “0”, “427”, “431.7”, …]

  – What appears if there isn’t anything in an entry? That is ,,

2. Find (row, column) where “Murder” appears

- getColumnRow(data, word)
  - data is the list of lists, word is word to find
  - returns the row and column (as a tuple) where the word first appears.
- First let’s write a helper function to focus on just one row, given one row (a list) return the position the word appears in the list or -1 if it does not appear
  - columnNumber(data, word)
- Then write getColumnRow

3. Get the items in the column below the word

- Now we know what row and column “Murder” appears in
- getColumn is given data (the list of lists) and startpos the (row, column) where “murder” appears
- getColumn returns a list of strings of all values below this word

4. Convert/clean the column into float numbers

- At this point, everything in the column is either a number as a string “4.7” or an empty string ‘’
- convertToNums returns a list of floats of those strings that are valid numbers

- Classwork!

5. Find the max min values

- This is easy if we have a list of numbers

6. Find the corresponding states

- getStatesWithRate returns a list of states that correspond to that number

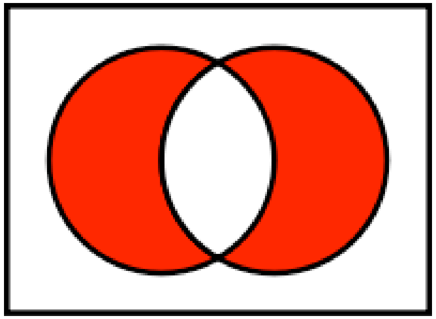
- Classwork!

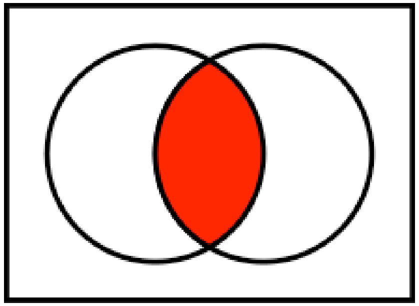
Python Sets

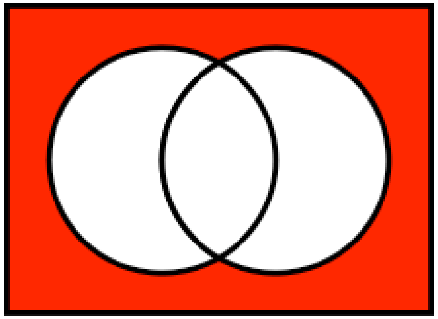
- Set – unordered collection of distinct items
  - Unordered – can look at them one at a time, but cannot count on any order
  - Distinct - one copy of each
- Operations on sets:
  - Modify: add, clear, remove
  - Create a new set: difference(-), intersection(&), union (|), symmetric_difference(^)
  - Boolean: issubset <=, issuperset >=
- Can convert list to set, set to list

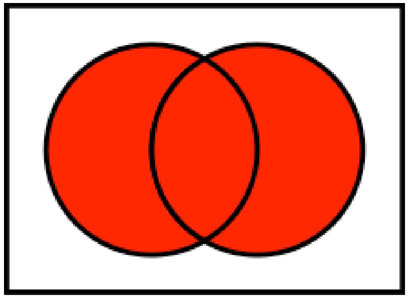
Set Operations from pictures

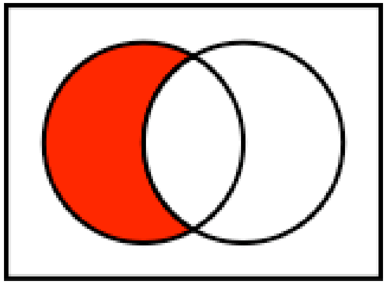
http://en.wikipedia.org/wiki/File:Venn0111.svg











Set Examples

```
poloClub = set(['Mary', 'Laura', 'Dell'])
rugbyClub = set(['Fred', 'Sue', 'Mary'])
print [w for w in poloClub.intersection(rugbyClub)]
print [w for w in poloClub.union(rugbyClub)]
```

More Set Examples

```
lista = ['apple', 'pear', 'fig', 'orange', 'strawberry']
listb = ['pear', 'lemon', 'grapefruit', 'orange']
listc = [x for x in lista if x in listb]
listd = list(set(lista)|set(listb))
print listc
print listd
```

- See setExample.py
- Classwork!