Example

Thomas Schwarz

The Denver Crime File

- Provided as a csv file
  - Contains data on crimes in Denver, provided by the police department
  - Our question: In which precinct are there the most auto-thefts?

The Denver Crime File

- First order of business:
  - Learn the structure of the data
    - Opening the file will usually get it imported as a spread-sheet, without the details of how data is actually stored.
  - We print out the first five lines

```
count_lines = 0
with open('crime.csv', 'r') as infile:
    for line in infile:
        print(line)
        count_lines += 1
        if count_lines > 5:
            break
```

The Denver Crime File


```
"INCIDENT_ID","OFFENSE_ID","OFFENSE_CODE","OFFENS
E_CODE_EXTENSION","OFFENSE_TYPE_ID","OFFENSE_CATE
GORY_ID","FIRST_OCCURRENCE_DATE","LAST_OCCURRENCE
_DATE","REPORTED_DATE","INCIDENT_ADDRESS","GEO_X"
,"GEO_Y","GEO_LON","GEO_LAT","DISTRICT_ID","PRECI
NCT_ID","NEIGHBORHOOD_ID","IS_CRIME","IS_TRAFFIC"

"2018869789","20188697892399 00","2399","0","theft
-other","larceny","12/27/2018 3:58:00
PM","","12/27/2018 4:51:00 PM","2681 N HANOVER
CT","3178210","1700715","-104.86615590","39.75556
140","5","512","stapleton","1","0"

"2015664356","20156643565441 00","5441","0","traff
ic-accident","traffic-accident","11/13/2015
7:45:00 AM","","11/13/2015 8:38:00 AM","4100
BLOCK W COLFAX
AVE","3129148","1694748","-105.04075970","39.7399
9120","1","122","west-colfax","0","1"
```

The Denver Crime File

- Fields are separated by commata
  - But all contents are in double quotation marks
    - This is probably an artifact of how the file was generated

The Denver Crime File

- We need to skip over the first line
    - Then we find the contents of sixth column and the 15th column in each row

```
count_lines = 0
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        print(data[5], data[14])
        count_lines += 1
        if count_lines > 5:
            break
```

skip over first line

The Denver Crime File

- We need to skip over the first line
    - Then we find the contents of sixth column and the 15th column in each row

```
count_lines = 0
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        print(data[5], data[14])
        count_lines += 1
        if count_lines > 5:
            break
```

split along commata

The Denver Crime File

- We need to skip over the first line
    - Then we find the contents of sixth column and the 15th column in each row

```
count_lines = 0
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        print(data[5], data[14])
        count_lines += 1
        if count_lines > 5:
            break
```

print-out the relevant columns to ensure that we did not miscount

The Denver Crime File

- We need to skip over the first line
    - Then we find the contents of sixth column and the 15th column in each row

```
count_lines = 0
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        print(data[5], data[14])
        count_lines += 1
        if count_lines > 5:
            break
```

stop after five lines

The Denver Crime File

- Result:

```
"larceny" "5"
"traffic-accident" "1"
"larceny" "6"
"larceny" "1"
"all-other-crimes" "5"
"public-disorder" "6"
```

- These are strings that start and end with quotation marks

The Denver Crime File

- We can get rid of the quotation marks using strip

```
for line in infile:
    data = line.split(',')
    offense_cat = data[5].strip('"')
    precinct = data[14].strip('"')
    print(offense_cat, precinct)
```

- And we can make the precinct into an integer

The Denver Crime File

- To count, we create a list of counts
  - We make sure that it is larger than the number of precincts
  - Whenever we find an auto-theft, we increment the corresponding element

```
counters = [0,0,0,0,0,0,0,0,0]
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        offense_cat = data[5].strip('"')
        precinct = int(data[14].strip('"'))
        if offense_cat == 'auto-theft':
            counters[precinct] += 1
print(counters)
```

The Denver Crime File

- But this gives us error!
  - This is where we use exceptions to print out the offending line

```
counters = [0,0,0,0,0,0,0,0,0]
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        offense_cat = data[5].strip('"')
        try:
            precinct = int(data[14].strip('"'))
        except ValueError:
            print(line)
        if offense_cat == 'auto-theft':
            counters[precinct] += 1
print(counters)
```

The Denver Crime File

- We find three lines in error:

```
"2018675045","2018675045299901","2999","1","criminal-mischief-mtr-
veh","public-disorder","9/29/2018 7:00:00 PM","9/29/2018 8:40:00
PM","10/1/2018 10:13:00 PM","2100 BLOCK MARKET ST, ROADSIDE
PARKING","3142947","1700282","-104.99158240","39.75498130","6","612","five-
points","1","0"

"2019553327","2019553327239901","2399","1","theft-
bicycle","larceny","8/27/2019 9:00:00 AM","8/30/2019 9:00:00 AM","8/30/2019
2:00:00 PM","2190 E 11TH AVE , 2ND FLOOR  PARKING
GARAGE","3151510","1692432","-104.96129270","39.73329620","6","622","cheesm
an-park","1","0"

"2019218943","2019218943530900","5309","0","harassment","public-
disorder","4/7/2019 9:10:00 PM","","4/8/2019 1:03:00 PM","1505 N GRANT ST,
COURT
YARD","3145128","1695002","-104.98393210","39.740453","6","621","north-
capitol-hill","1","0"
```

The Denver Crime File

- Since the precinct is causing the problem, we can identify it as data[-5]
  - No exceptions:

```
counters = [0,0,0,0,0,0,0,0,0]
with open('crime.csv', 'r') as infile:
    infile.readline()
    for line in infile:
        data = line.split(',')
        offense_cat = data[5].strip('"')
        try:
            precinct = int(data[-5].strip('"'))
        except ValueError:
            print(line)
        if offense_cat == 'auto-theft':
            counters[precinct] += 1
print(counters)
```

The Denver Crime File

- And this is our answer:
  - [0, 4852, 4120, 6646, 5352, 3199, 2926, 650, 0]
  - Precinct 1 has 4852
  - Precinct 2 has 4120
  - …
- So, Precinct 3 has the most auto-thefts.