



# Data Wrangling

Data Science: Jordan Boyd-Graber  
University of Maryland

DECEMBER 29, 2017

## Administrivia

- Course Staff: Apoorva / Pedro
- Moodle self-enrollment now working
- Questions?

## Big Picture

- Data are messy (this isn't so messy!)
- The first step to doing anything cool is using data
- Need to use common sense and brute force often
- You'll see more in first homework (up tomorrow, due Sept. 8)

## First Steps: Get Data

- From FEC
- Odd formatting

## First Steps: Get Data

- From FEC
- Odd formatting
- Today: pure Python (easier with Pandas), will help expose level of Python you'll need

**Look at file ...**

## Look at file ...

- Periods instead of commas (vice versa)
- Odd New York parties
- Semi-colon delimiters
- Includes totals

## Read in Data



## Read in Data

```
from csv import DictReader
votes = list(DictReader(open("2012pres.csv", 'r'),
                        delimiter=";"))
```

## Read in Data

## Read in Data

```
from csv import DictReader
votes = list(DictReader(open("2012pres.csv", 'r'),
                        delimiter=";"))
```

**How many votes were cast?**

**How many votes were cast?**

Total votes 129085410

## How many votes were cast?

Total votes 129085410

```
total_votes = sum(int(x["TOTAL VOTES #"].replace(".", ""))  
                  for x in votes if x["TOTAL VOTES #"])
```

**What state had the largest numerical margin between first and second place?**

**What state had the largest numerical margin between first and second place?**

Largest numerical margin 3014327 in California



**What state had the largest numerical margin between first and second place?**

Largest numerical margin 3014327 in California

```
margins = {}  
for ss in set(x["STATE"] for x in votes):  
    margins[ss] = winner(votes, ss)[1] - second(votes, ss)[1]  
num_margin = argmax(margins)  
print("Largest numerical margin %i in %s" %  
      (max(margins.values()), num_margin))
```

**What state had the largest percentage margin between first and second place?**

**What state had the largest percentage margin between first and second place?**

Largest percentage margin 48.04 in Utah

**What state had the largest percentage margin between first and second place?**

Largest percentage margin 48.04 in Utah

```
margins = {}
for ss in set(x["STATE"] for x in votes
               if x["STATE"] != "District of Columbia"):
    margins[ss] = winner(votes, ss)[2] - \
        second(votes, ss)[2]
num_margin = argmax(margins)
print("Largest percentage margin %f in %s" %
      (max(margins.values()), num_margin))
```

**What state had the largest numerical third party vote (and for whom)?**

**What state had the largest numerical third party vote (and for whom)?**

Johnson had largest third party vote in California with 143221

## What state had the largest numerical third party vote (and for whom)?

Johnson had largest third party vote in California with 143221

```
all_third_vote = {}
top_third_vote = {}
for ss in set(x["STATE"] for x in votes):
    try:
        all_third_vote[ss] = \
            dict((x["LAST NAME"],
                  parseint(x["GENERAL RESULTS"])))
        for x in votes
        if x["STATE"] == ss
            and x["LAST NAME"] not in kMAJOR
            and x["LAST NAME"])
    except ValueError:
        all_third_vote[ss] = {}
    if all_third_vote[ss]:
        top_third_vote[ss] = max(all_third_vote[ss].val
```

**What state had the largest percentage vote (and for whom)?**



**What state had the largest percentage vote (and for whom)?**

Johnson had largest third party percent in New Mexico with 3.55

## What state had the largest percentage vote (and for whom)?

Johnson had largest third party percent in New Mexico with 3.55

```
all_third_vote = {}
top_third_vote = {}
for ss in set(x["STATE"] for x in votes):
    try:
        all_third_vote[ss] = \
            dict((x["LAST NAME"],
                  parseint(x["GENERAL RESULTS"])))
        for x in votes
        if x["STATE"] == ss
            and x["LAST NAME"] not in kMAJOR
            and x["LAST NAME"])
    except ValueError:
        all_third_vote[ss] = {}
    if all_third_vote[ss]:
        top_third_vote[ss] = max(all_third_vote[ss].val
```

## Summary

- Data are messy
- Easier with formatted data (e.g., csv)
- Need basic data structures
- Check whether answers are reasonable

## Next Time ...

- Lecture: make sure to do reading
- Probability foundations (if you found today boring ...)
- Math needed for the course (quiz likely)