Quantifying Corruption

Exploring the impact of donations on legislation

Background

Prior to 2010 until the FEC V Citizens United Decision by SCOTUS the FEC tried to limit campaign contributions to US Reps.

SCOTUS determined that individual contributions cannot be limited* .

Neat!

*Subject to some restrictions

Background

This limitless contribution ability means that an individual who has more money can spend more money contributing to a campaign.

The obvious issue being those with more money can pay more to elect people they like.

Goal:

To determine the impact on voting record that these contributions display and to understand if the concerns of equal representation are well founded.

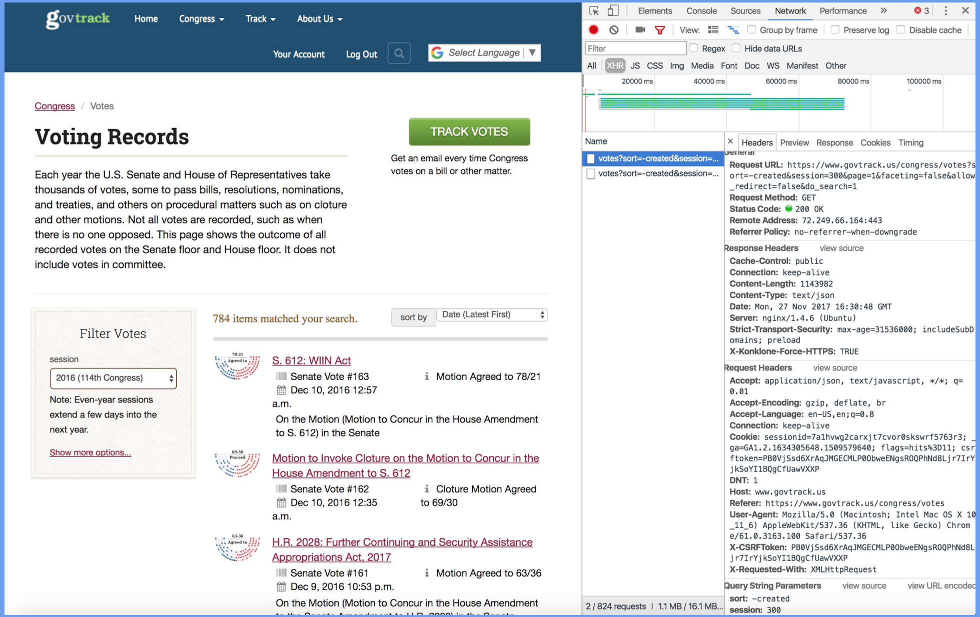
Collecting Data

- A lot of webscraping!
    - Python and BeautifulSoup are amazing

- Data came from govtrack.us & OpenSecrets.org

Collecting Data Cont.

- Issues
  - Multiple links
  - JQuery
  - Download links



Cleaning Donation Data

- Data was nasty, no header
- |2014|,|H2MD08159|,|N00012668|,|Ken Timmerman (R)|,|R|,|MD08|,| |,| |,| |,| |,|RN|,| |
- `:%s/|,|//g` - `sed`
- Then took out name, party, and candidate ID

Cleaning Voting Data

- 705 votes cast
- Data wasn't too nasty
  - 1st line description of the vote then
  - Person ID,state,district,vote,name,party
- Strip out state, district

Donation & Voting Data

- Python script to link the two
    - Voting data has PersonID
    - Donation data has Candidate ID
- Needed for clustering

Building Donation Profiles

- Collect Data
    - Candidate ID
    - Donation Amount
    - Party
- Normalize Data

Building Donation Profiles

- Set Benchmarks
  - Average
  - Standard Deviation
  - Variance
- Group into Categories by Candidate

Building Donation Profiles

- Benchmarks
  - Up to 580.49
  - 580.49 to 1526.17
  - 1526.17 to 4012.47
  - 4012.47 to 10549.24
  - Over 10549.24

Clustering Donation Profiles

- Allows us to look at broader trends
- Effectively filters out anomalous donations

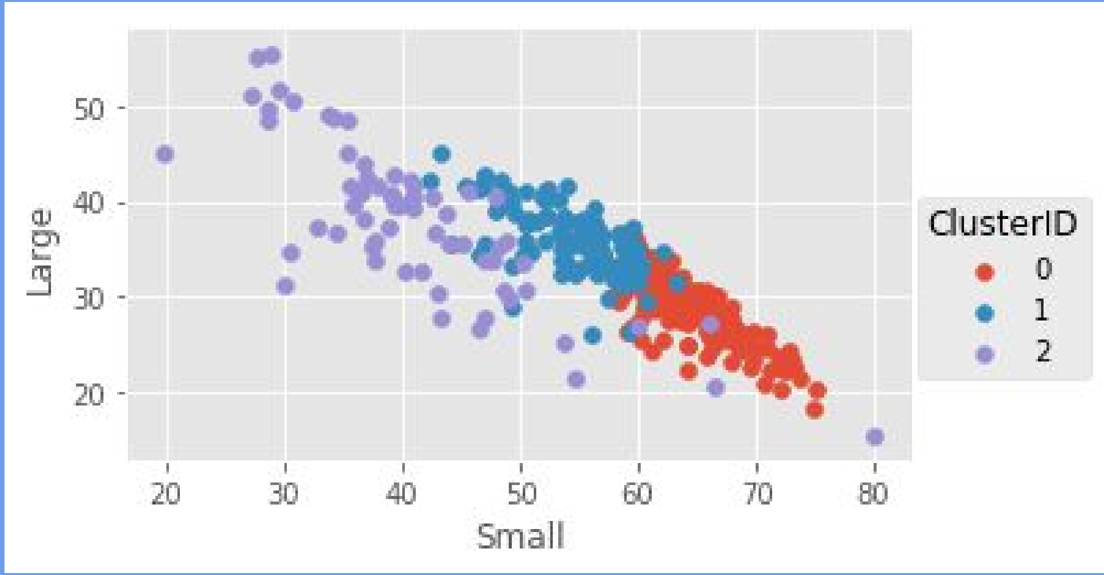
Clustering Donation Profiles

- Picking a value of k

Visualizing Clusters

Large: Top 2 donation Categories

Small: Bottom 2 Donation Categories

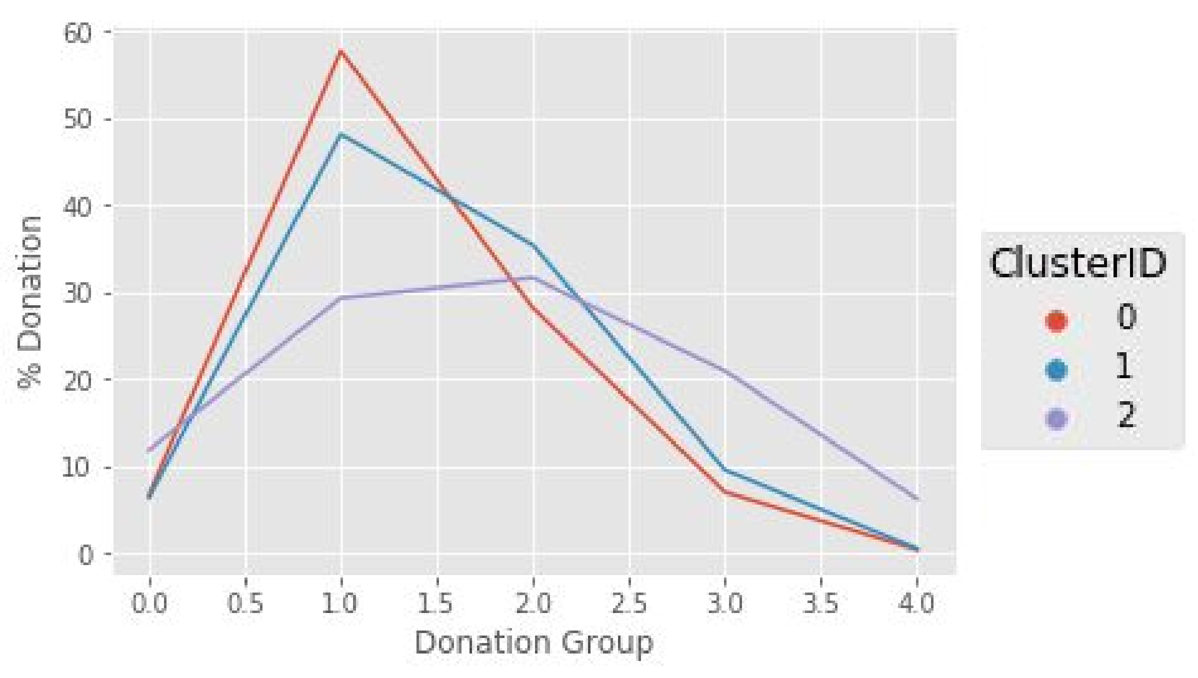


Cluster 0: mostly small contributions

Cluster 1: Some of small and large contributions

Cluster 2: Mostly Large Contributions

Visualizing Clusters



Looking in terms of percentages.

Donation Categories

0: 0 to 580.49

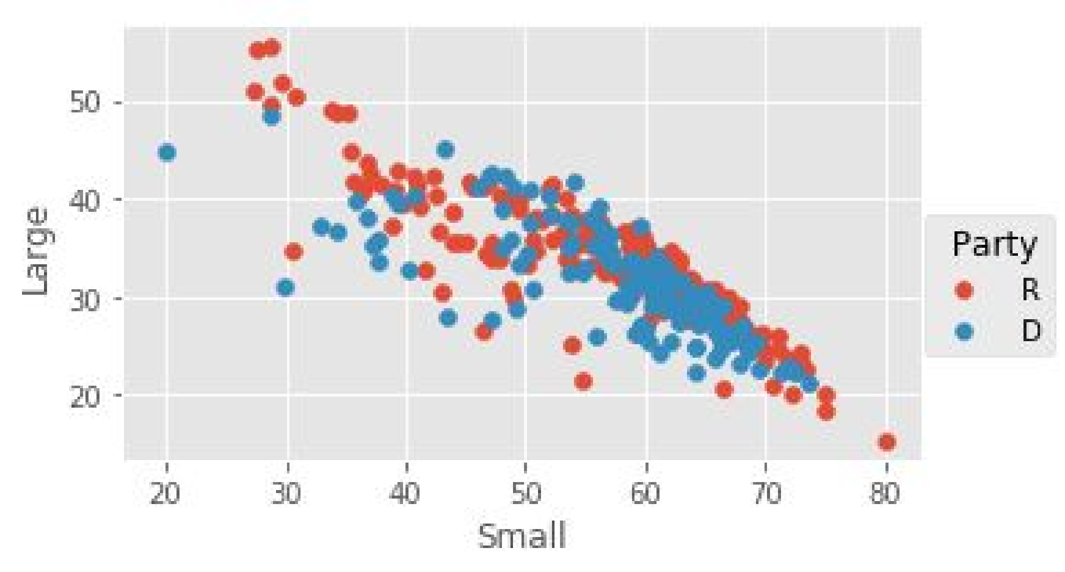
1: 580.49 to 1526.17

2: 1526.17 to 4012.47

3: 4012.47 to 10549.24

4: over 10549.24

Visualizing Clusters



Although the Republican candidates have a broader range and tend to receive larger single contributions neither party is limited to a single cluster and both recieve a range of donations

Testing For Significance

- To get a sense for how each cluster voted, we averaged all votes for each cluster
- An average value of 1 means all members voted yes

Testing For Significance

- Three aggregate voting records
- Each cluster votes on the same bills, so we have data on paired “events”
- Paired T-Test

Testing For Significance

- T-Test for every pair of clusters
- Highest p-value between clusters: ~0.000051

```
In [6]:  t_test_pv = np.empty((3, 3))
         for i, cvl1 in enumerate(cluster_voting):
             for j, cvl2 in enumerate(cluster_voting):
                 t_test_pv[i][j] = stats.ttest_rel(cvl1, cvl2)[1]

         t_test_pv

Out[6]: array([[            nan,   5.10779906e-05,   4.11646508e-14],
               [  5.10779906e-05,             nan,   3.73281923e-12],
               [  4.11646508e-14,   3.73281923e-12,             nan]])
```

Testing For Significance

- This suggests there's less than a ~0.005% chance differences are due to chance
- **But,** the t-test assumes that the observations follow a normal distribution

Testing For Significance

- We can try a different tact using logistic regression
- Train a logistic regression model on the data, using cluster ID as a categorical variable
- The magnitude of the 2nd and 3rd coefficients is our confidence (those clusters are different)

Testing For Significance

- The percentage of bills that showed significance

```
[[ 0.          0.93019943  0.9031339 ]
 [ 0.93019943  0.          0.96153846]
 [ 0.9031339   0.96153846  0.        ]]
```

Conclusions

- Explicitly stating your question (and workflow) is important
- Real data is messy
  - Collecting it is not always straightforward
- Model assumptions matter

What We Didn’t Do

- We didn’t prove a causal relationship between donations and voting behavior
- We didn’t consider the identity of donors
- We didn’t examine the content of bills

Are Politicians Corrupt?

- Ya