# US election results and donor contributions for 2016

## Extract, Transform & Load (ETL) Project

Kathleen Buckingham, Michael Marshall, Richard Pherson, Vibhanshu Shekhar

## Extract

We used two CSV files from Kaggle – donor data for the 2016 election and the 2012 and 2016 US Presidential election. The election data contained election results for both the 2012 and 2016 US Presidential Elections including the proportion of votes cast for Obama and Romney (2012) and Trump and Clinton (2016). The data had 21 columns of which we extracted 5 columns, including: total democratic votes, total republican votes, total votes, state abbreviations and county names. Our focus was on the 2016 election, therefore we extracted information from 2016 and discarded all 2012 data. There were many values that were not useful to our data including percentage of votes per state.

The donor data set provided the names and amounts of contributions for the 2016 US Presidential election campaign. The data had 18 columns, from which we used 4 columns, including: candidate name, contributor amount, contributor receipt date, and contributor state abbreviation. There were many values that were not useful to our analysis including the contributor employer and zip code. Both data sets included the state abbreviation in which we could link the datasets.

We added additional information manually regarding the 7 candidates and their party affiliation.

## Transform

Due to the size of the donor file(166MB), we were unable to push the file to our GitHub repository. While it is possible to accomplish programmatically, due to the minimal level of effort involved, we manually dropped columns from the donor data and uploaded the reduced CSV. To create our tables, we used Pandas and SQL Alchemy to read the CSV’s into Data Frames for further cleanup and organization. It was determined that the data sets did not contain any built-in way to relate the candidates in the donor table to the party-based election results table. As there were only 7 candidates within the data set, a third table was manually created to associate the candidates with their respective political parties.

## Load

We then combined the three datasets into a Postgres database. The data is capable of being related between tables based on State, Political Party, and candidate.

## Lessons Learned

Although the donor data had over a million records, the data appears incomplete. Firstly, there were only 8 candidates of which only one ran on the Democratic ticket. Additionally, the 7 remaining Republican candidates do not reflect the entire field that ran. It may be enough for estimation purposes, but the results of any analysis done with incomplete data such as this may be suspect.

Further, it appears that donors may have had to voluntarily self-identify their State without any reconciliation done on the part of the campaigns. These records indicate a total of 85 different state entries were collected, which is more than the combined list of all US States and Territories.