**Questions**

* What role do individual contributions play in political candidates winning elections?
  + Do incumbents receive more contributions?
  + Do candidates that receive more contributions win?
  + Contribution dollars per vote
  + What states are candidates from that receive the most contributions?
  + Which congressional districts receive the most contributions?
  + Do incumbents or challengers receive more contributions?

**Goals**

* I wanted to take campaign contribution data and join it to election result data to see if I could answer the above questions.
* Learn / utilize Pyspark
* Work with data frames
  + Create
  + Joins (left and inner)
  + Filtering
  + Aggregation

**Data**

* Election data
  + <https://electionlab.mit.edu/data>
* Campaign contribution data
  + <https://www.fec.gov/data/browse-data/?tab=bulk-data>

**Election results data**

***Data summary***

The election results data consisted of three files; one for Presidential, Senate, and House elections respectively going back to 1976. Each file consisted of all candidates that received votes and the number of votes they received aggregated at the congressional district level. Each dataset was loaded into its own data frame and then the following work was performed on each data frame.

***Columns added***

* Candidate first name and candidate last name
  + These columns were created by splitting the candidate name field
* Is Winner
  + The data did indicate who won the election so I needed to mark the records that received the most votes in each election. I did this by aggregating on the fields outlined below and finding the max votes. I created a data frame of this and then left joined the data frame back to the respective House / Senate / President data frame on all these fields including a join on max votes to votes received by the candidates. A left join was used so I could mark the records that joined as the winners in the original data frames and all other records as non-winners. I then had to account for the fact that Clinton lost the 2016 presidential election despite winning the popular vote and being flagged as the winner in my data frame. To do this I marked Clinton’s 2016 record as a 0 (non-winner) and the current occupant of the White House’s record with a 1.
    - Fields aggregated on
      * Last name
      * Party
      * Election year
      * Office
      * State
      * District

***Additional clean-up***

Since there is not unique ID to join the election data to the campaign contribution data I needed to come up with a way to join the two disparate datasets and settled on using the following fields – last name, party, election year. I couldn’t use the candidate’s full name because of differences in the naming. Example – Hillary Clinton; Clinton, Hillary; Hillary Rodham Clinton. I thought the data would be unique by these fields which would allow me to join the two datasets on a 1:1 basis. However, I found that the data was not unique in a small number of instances – ex John Smith and Bob Smith running for the same seat in the same year for the same party. To remedy this I deduped the data on last name, party, and election year by aggregating each dataset on these fields and removing records that had a count greater than one. I realized that this would affect the analysis, but given the small number of records affected, I deemed it to be acceptable. Once this was performed on each of the three datasets (House, Senate, President) they were ready to be joined to the campaign contribution data.

**Campaign contribution data**

***Overview***

There were three datasets I had to work with for the campaign contribution data. Datasets were provided in two year increments (every even year because there are House elections every two years) and I used files from 2008-2018. For each of these datasets I defined a schema and then loaded the data into said schema. Upon loading the candidate master data I split the CAND\_NAME field into first name and last name fields.

* Individual contributions
  + Line item detail for every contribution. Individual contributions only have to be recorded if they are greater than $200.
* Candidate Master
  + The candidate master file contains one record for each candidate who has either registered with the Federal Election Commission or appeared on a ballot list prepared by a state elections office.
  + A candidate has a unique ID for each office they are running for. For example, Hillary Clinton has a single CAND\_ID for her senate runs and a different CAND\_ID for her presidential runs. The CAND\_ID is the same across years though as long as it is for the same office.
  + Data is not unique by CAND\_ID and CAND\_OFFICE though. The file can contain duplicates of CAND\_ID and CAND\_OFFICE with different CAND\_NAMEs. For example, Hillary Clinton can appear as Hillary Clinton, Hillary Rodham Clinton, and Hillary Rodham Clinton / Timothy Michael Kaine.
* Candidate-committeee linkages
  + Mapping table that linked contributions to the candidate the contribution was for.

***De-dupe candidate data frame***

I needed one record per candidate for each file so I could ultimately join to the elections data. Given the lack of a unique key between the contribution data and the election data I decided I would join on last name, party, and election year. In order to do this I needed to make sure the data was unique and n the few instances where there were duplicates (ex – John Smith and Bob Smith from the same party running for the same seat in the same year) I removed them. I aggregated on the following fields and if the count was greater than one I removed them.

* Fields to dedupe on
  + CAND\_PTY\_AFFILIATION
  + CAND\_OFFICE\_ST
  + CAND\_OFFICE
  + CAND\_OFFICE\_DISTRICT
  + FILE\_YR
  + LAST\_NAME

***Join and consolidate the data***

1. Aggregate contribution data by CMTE\_ID (which is the linking field between contribution data and the candidates).
2. Join candidate data to the campaign linking data to create a new data frame with candidate id and the linking field (CMTE\_ID) so I could then join it to the aggregated contribution data frame (from bullet a).
3. Join the campaign contributions to their respective candidate (data frames from bullets a and b) and aggregate the contributions by candidate id and election year.
4. Join this new aggregated data frame (from bullet c) back to the deduped candidate data frame to get the metadata about the candidate.
5. Flag the records that received the most contributions in each race.
   1. Data was aggregated by the below fields to find the max contributions amount for each race. This had to be done because there was no record to indicate who was in a race against one another. These records are unique for each race.
      1. Office
      2. State
      3. District
      4. Election Year
   2. The above was then left joined back to the data frame it was aggregated off of (data frame from column d) and the records that matched flagged with a 1 in a new “is\_max” column while all other records were given a 0.

**Join candidate contribution data back to the election results**

The final data frames were created by joining each election results data frame (House, Senate, President) to the candidate contribution data frame (data frame from column e above) on the below fields. These data frames that were ultimately created have one record for each candidate / election with corresponding metadata about the candidate, the amount of contributions they received, a flag indicating whether or not they won the election, and a flag indicating whether or not they received the most money of all the candidates in the race.

* Fields joined on
  + Party
  + Last name
  + Election year
  + State
  + District
  + Office

**Displaying results**

These three data frames were then output to CSV files and loaded into tableau for exploration. Once in tableau I could slice and dice the data to answer any question that I had. Here are a couple screenshots of the analysis.



