Project 2: ETL Challenge

This project was completed by Indrajith, Felix, and Emily.

This project will deliver a database with data about US elections. It has been aggregated at the “state-level”. By doing so, a link was made among all datasets.

1. What insights can we discover by mapping this data from the past 4 elections, specially voter turnout in the past election cycles?
2. Is there any correlation between candidates and state voting patterns?
3. Is there any correlation between candidates and Twitter profile?

**Extract**: First the following 4 data sources are loaded. Csv files from Kaggle and dataverse can be found in the Resources folder tabs.

1. 2016 US Election: <https://www.kaggle.com/benhamner/2016-us-election#county_facts.csv>
2. 1976-2016 US Election: <https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/42MVDX/MFU99O&version=5.0>
3. 2012 & 2016 US Election: <https://www.kaggle.com/joelwilson/2012-2016-presidential-elections>
4. 2020 US Election:<https://www.kaggle.com/radustoicescu/2020-united-states-presidential-election>

**Transform:**

1. Isolate data for the 2008 through 2016 Presidential elections
2. Organize data by State, Year and Election Results

1976-2016 US Election (MIT\_CSV):

* Organize the data to fit into the following:
  + Election\_Year
  + State
  + Candidate\_Name
  + Party\_Affiliation
  + Votes
* Isolate 2008 US Election data and convert it into csv file
* Small data cleanup within the file, such as rename the column names
* Drop any rows that have missing data

2012 & 2016 votes:

* Create a table and manually upload data into the following:
  + Election\_Year
  + State
  + Candidate\_Name
  + Party\_Affiliation
  + Votes
* Total votes for 2012 and 2016 election year
* Select only 3 columns (state\_abbr, 2012 votes, 2016 votes) and reformat it into another table

**Merge the two datasets together**

**Load:**

We created the tables in postgreSQL.