```
import geopandas
```

County Election Data

In [2]:
```
mit_data = pd.read_csv('original_data/countypres_2000-2020.csv',dtype={'county_fips':str})
mit_data = mit_data.rename(columns={'county_fips':'FIPS'})
mit_data = mit_data[~mit_data['FIPS'].isnull()]
mit_data['FIPS'] = mit_data['FIPS'].str.zfill(5)
```

Data Repair: Not all counties have vote totals, so calculate new vote totals based on candidatevotes

2000: North Carolina, Oklahoma; 2004: Oklahoma

In [3]:
```
grp = mit_data.groupby(by=['year','FIPS']).sum().reset_index()
grp = grp.drop(labels=['totalvotes','version'],axis=1)
grp = grp.rename(columns={'candidatevotes':'totalvotes2'})
mit_data = mit_data.merge(grp,on=['year','FIPS'])

mit_data['totalvotes'] = mit_data['totalvotes2']
mit_data = mit_data.drop(labels=['totalvotes2'],axis=1)
print(mit_data.head())
```

```
   year    state state_po county_name   FIPS     office       candidate  \
0  2000  ALABAMA       AL     AUTAUGA  01001  PRESIDENT          AL GORE
1  2000  ALABAMA       AL     AUTAUGA  01001  PRESIDENT  GEORGE W. BUSH
2  2000  ALABAMA       AL     AUTAUGA  01001  PRESIDENT     RALPH NADER
3  2000  ALABAMA       AL     AUTAUGA  01001  PRESIDENT           OTHER
4  2000  ALABAMA       AL     BALDWIN  01003  PRESIDENT          AL GORE

        party  candidatevotes  totalvotes   version   mode
0    DEMOCRAT          4942.0     17208.0  20191203  TOTAL
1  REPUBLICAN         11993.0     17208.0  20191203  TOTAL
2       GREEN           160.0     17208.0  20191203  TOTAL
3       OTHER           113.0     17208.0  20191203  TOTAL
4    DEMOCRAT         13997.0     56480.0  20191203  TOTAL
```

Data Repair: Reclassify Shannon County FIPS as Oglala Lakota County FIPS

In [4]:
```
mit_data.loc[mit_data['FIPS']=='46113','FIPS'] = '46102'
```

Data Repair: Some counties in 2020 list separate tallies for different kinds of ballots

In [5]:
```
mit_data.loc[mit_data.county_name=='MANASSAS PARK CITY']
```

Out[5]:

	year	state	state_po	county_name	FIPS	office	candidate	party	candidatevotes	totalvotes	version	mode
71701	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JOSEPH R BIDEN JR	DEMOCRAT	3137.0	6088.0	20210622	ABSENTEE
71702	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JOSEPH R BIDEN JR	DEMOCRAT	834.0	6088.0	20210622	ELECTION DAY
71703	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JOSEPH R BIDEN JR	DEMOCRAT	21.0	6088.0	20210622	PROVISIONAL
71704	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JO JORGENSEN	LIBERTARIAN	61.0	6088.0	20210622	ABSENTEE
71705	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JO JORGENSEN	LIBERTARIAN	38.0	6088.0	20210622	ELECTION DAY
71706	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	JO JORGENSEN	LIBERTARIAN	1.0	6088.0	20210622	PROVISIONAL
71707	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	OTHER	OTHER	10.0	6088.0	20210622	ABSENTEE
71708	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	OTHER	OTHER	7.0	6088.0	20210622	ELECTION DAY
71709	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	OTHER	OTHER	0.0	6088.0	20210622	PROVISIONAL
71710	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	DONALD J TRUMP	REPUBLICAN	1239.0	6088.0	20210622	ABSENTEE
71711	2020	VIRGINIA	VA	MANASSAS PARK CITY	51685	US PRESIDENT	DONALD J TRUMP	REPUBLICAN	733.0	6088.0	20210622	ELECTION DAY

Out[8]:

	year	state	state_po	county_name	FIPS	office	candidate	party	candidatevotes	totalvotes	version	mode
11889	2020	NaN	NaN	NaN	56043	NaN	OTHER	NaN	71.0	4032.0	NaN	NaN
11890	2020	NaN	NaN	NaN	56045	NaN	DONALD J TRUMP	NaN	3107.0	3560.0	NaN	NaN
11891	2020	NaN	NaN	NaN	56045	NaN	JO JORGENSEN	NaN	46.0	3560.0	NaN	NaN
11892	2020	NaN	NaN	NaN	56045	NaN	JOSEPH R BIDEN JR	NaN	360.0	3560.0	NaN	NaN
11893	2020	NaN	NaN	NaN	56045	NaN	OTHER	NaN	47.0	3560.0	NaN	NaN

Continue with data processing

In [9]:
```
presidential_candidates = {2000:{'gop':'GEORGE W. BUSH','dem':'AL GORE'},
                           2004:{'gop':'GEORGE W. BUSH','dem':'JOHN KERRY'},
                           2008:{'gop':'JOHN MCCAIN','dem':'BARACK OBAMA'},
                           2012:{'gop':'MITT ROMNEY','dem':'BARACK OBAMA'},
                           2016:{'gop':'DONALD TRUMP','dem':'HILLARY CLINTON'},
                           2020:{'gop':'DONALD J TRUMP','dem':'JOSEPH R BIDEN JR'}
                           }
```

In [10]:
```
output_df = pd.DataFrame()
output_df['FIPS'] = mit_data['FIPS'].unique()

years = np.sort(list(presidential_candidates.keys()))

for year in years:
    # Pull this year as a dataframe, pull this year's candidates, and
    # convert year to a string, since it will now be used to name fields
    df=mit_data[mit_data['year']==year]
    candidates = presidential_candidates[year]
    year = str(year)

    # Get candidate info for this year, rename
    gop = df.candidate == candidates['gop']
```

In [8]:
```
mit_data = mit_data[mit_data['year'] != 2020]
mit_data = mit_data.append(out)

mit_data.tail()
```

Out[8]:

	year	state	state_po	county_name	FIPS	office	candidate	party	candidatevotes	totalvotes	version	mode
11889	2020	NaN	NaN	NaN	56043	NaN	OTHER	NaN	71.0	4032.0	NaN	NaN
11890	2020	NaN	NaN	NaN	56045	NaN	DONALD J TRUMP	NaN	3107.0	3560.0	NaN	NaN
11891	2020	NaN	NaN	NaN	56045	NaN	JO JORGENSEN	NaN	46.0	3560.0	NaN	NaN

11892	2020	NaN	NaN	NaN	56045	NaN	JOSEPH R BIDEN JR	NaN	360.0	3560.0	NaN	NaN
11893	2020	NaN	NaN	NaN	56045	NaN	OTHER	NaN	47.0	3560.0	NaN	NaN

Continue with data processing

In [9]:
```
presidential_candidates = {2000:{'gop':'GEORGE W. BUSH','dem':'AL GORE'},
                           2004:{'gop':'GEORGE W. BUSH','dem':'JOHN KERRY'},
                           2008:{'gop':'JOHN MCCAIN','dem':'BARACK OBAMA'},
                           2012:{'gop':'MITT ROMNEY','dem':'BARACK OBAMA'},
                           2016:{'gop':'DONALD TRUMP','dem':'HILLARY CLINTON'},
                           2020:{'gop':'DONALD J TRUMP','dem':'JOSEPH R BIDEN JR'}
                           }
```

In [10]:
```
output_df = pd.DataFrame()
output_df['FIPS'] = mit_data['FIPS'].unique()

years = np.sort(list(presidential_candidates.keys()))

for year in years:
    # Pull this year as a dataframe, pull this year's candidates, and
    # convert year to a string, since it will now be used to name fields
    df=mit_data[mit_data['year']==year]
```

Out[10]:

	FIPS	gop_2000_votes	dem_2000_votes	totalvotes_2000	gop_2000_prc	dem_2000_prc	gop_minus_dem_prc_2000	gop_2004_votes	dem_2004_votes	to
0	01001	11993.0	4942.0	17208.0	69.69	28.72	40.97	15196.0	4758.0	
1	01003	40872.0	13997.0	56480.0	72.37	24.78	47.59	52971.0	15599.0	
2	01005	5096.0	5188.0	10395.0	49.02	49.91	-0.89	5899.0	4832.0	
3	01007	4273.0	2710.0	7101.0	60.17	38.16	22.01	5472.0	2089.0	
4	01009	12667.0	4977.0	17973.0	70.48	27.69	42.79	17386.0	3938.0	

5 rows × 37 columns

In [11]:
```
output_df.to_csv('county_election_data_2000-2020.csv',index=False,float_format='%.2f')
```

State Election Data

In [23]:
```
mit_data = pd.read_csv('original_data/1976-2020-president.csv',dtype={'state_fips':str})
mit_data = mit_data.rename(columns={'state_fips':'FIPS'})
mit_data = mit_data[~mit_data['FIPS'].isnull()]
mit_data['FIPS'] = mit_data.FIPS.str.zfill(2)
```

In [24]:
```
presidential_candidates = {1976:{'gop':'FORD, GERALD','dem':'CARTER, JIMMY'},
                           1980:{'gop':'REAGAN, RONALD','dem':'CARTER, JIMMY'},
                           1984:{'gop':'REAGAN, RONALD','dem':'MONDALE, WALTER'},
                           1988:{'gop':'BUSH, GEORGE H.W.','dem':'DUKAKIS, MICHAEL'},
                           1992:{'gop':'BUSH, GEORGE H.W.','dem':'CLINTON, BILL'},
                           1996:{'gop':'DOLE, ROBERT','dem':'CLINTON, BILL'},
                           2000:{'gop':'BUSH, GEORGE W.','dem':'GORE, AL'},
                           2004:{'gop':'BUSH, GEORGE W.','dem':'KERRY, JOHN'},
                           2008:{'gop':'MCCAIN, JOHN','dem':'OBAMA, BARACK H.'},
```


State Election Data

In [23]:
```
mit_data = pd.read_csv('original_data/1976-2020-president.csv',dtype={'state_fips':str})
mit_data = mit_data.rename(columns={'state_fips':'FIPS'})
mit_data = mit_data[~mit_data['FIPS'].isnull()]
mit_data['FIPS'] = mit_data.FIPS.str.zfill(2)
```

In [24]:
```
presidential_candidates = {1976:{'gop':'FORD, GERALD','dem':'CARTER, JIMMY'},
                           1980:{'gop':'REAGAN, RONALD','dem':'CARTER, JIMMY'},
                           1984:{'gop':'REAGAN, RONALD','dem':'MONDALE, WALTER'},
                           1988:{'gop':'BUSH, GEORGE H.W.','dem':'DUKAKIS, MICHAEL'},
                           1992:{'gop':'BUSH, GEORGE H.W.','dem':'CLINTON, BILL'},
                           1996:{'gop':'DOLE, ROBERT','dem':'CLINTON, BILL'},
                           2000:{'gop':'BUSH, GEORGE W.','dem':'GORE, AL'},
                           2004:{'gop':'BUSH, GEORGE W.','dem':'KERRY, JOHN'},
                           2008:{'gop':'MCCAIN, JOHN','dem':'OBAMA, BARACK H.'},
                           2012:{'gop':'ROMNEY, MITT','dem':'OBAMA, BARACK H.'},
                           2016:{'gop':'TRUMP, DONALD J.','dem':'CLINTON, HILLARY'},
                           2020:{'gop':'TRUMP, DONALD J.','dem':'BIDEN, JOSEPH R. JR'}
                           }
```


In [26]:
```
# Mitt Romney's name is reversed for Washington for 2012.  This has been fixed in previous versions, but it's
# left here for instructional purposes:

idx = (mit_data['state_po']=='WA') & (mit_data['year']==2012) & (mit_data['party_detailed']=='REPUBLICAN')
mit_data[idx]
```

Out[26]:

	year	state	state_po	FIPS	state_cen	state_ic	office	candidate	party_detailed	writein	candidatevotes	totalvotes	version	notes
3371	2012	WASHINGTON	WA	53	91	73	US PRESIDENT	MITT, ROMNEY	REPUBLICAN	False	1290670	3125516	20210113	NaN

In [27]:
```
# Perform the fix
mit_data.loc[idx,'candidate'] = 'ROMNEY, MITT'
mit_data[idx]
```

Out[27]:

	year	state	state_po	FIPS	state_cen	state_ic	office	candidate	party_detailed	writein	candidatevotes	totalvotes	version	notes
3371	2012	WASHINGTON	WA	53	91	73	US PRESIDENT	ROMNEY, MITT	REPUBLICAN	False	1290670	3125516	20210113	NaN

In [28]:
```
output_df = mit_data.loc[:,['state','state_po','FIPS']]
output_df = output_df.drop_duplicates()

years = np.sort(list(presidential_candidates.keys()))

for year in years:
    # Pull this year as a dataframe, pull this year's candidates, and
    # convert year to a string, since it will now be used to name fields
    df=mit_data[mit_data['year']==year]
    candidates = presidential_candidates[year]
    year = str(year)

    # Get candidate info for this year, rename
    gop = df.candidate == candidates['gop']
    gop = df.loc[gop,['state_po','candidatevotes']]
    gop = gop.groupby('state_po').sum()
    gop = gop.rename(columns={'candidatevotes':'gop' + '_' + year + '_votes'})
    dem = df.candidate == candidates['dem']
    dem = df.loc[dem,['state_po','candidatevotes','totalvotes']]
    dem = dem.groupby('state_po').sum()
    dem = dem.rename(columns={'candidatevotes':'dem' + '_' + year + '_votes'})
    dem = dem.rename(columns={'totalvotes':'totalvotes' + '_' + year})
```

```
    # Write this information to the output dataframe and calculate some fields
    output_df = output_df.merge(gop,on='state_po',how='left')
    output_df = output_df.merge(dem,on='state_po',how='left')
    output_df['gop_' + year + '_prc'] = np.round(100 * output_df['gop_' + year + '_votes'] / output_df['totalvotes_' + year],
    output_df['dem_' + year + '_prc'] = np.round(100 * output_df['dem_' + year + '_votes'] / output_df['totalvotes_' + year],
    output_df['gop_minus_dem_prc_' + year] = output_df['gop_' + year + '_prc'] - output_df['dem_' + year + '_prc']

output_df.head()
```

Out[28]:

	state	state_po	FIPS	gop_1976_votes	dem_1976_votes	totalvotes_1976	gop_1976_prc	dem_1976_prc	gop_minus_dem_prc_1976	gop_1980_vot
0	ALABAMA	AL	01	504070	659170	1182850	42.61	55.73	-13.12	6541
1	ALASKA	AK	02	71555	44058	123574	57.90	35.65	22.25	861
2	ARIZONA	AZ	04	418642	295602	742719	56.37	39.80	16.57	5296
3	ARKANSAS	AR	05	267903	498604	767535	34.90	64.96	-30.06	4031
4	CALIFORNIA	CA	06	3882244	3742284	7803770	49.75	47.95	1.80	4522

5 rows × 75 columns

In [30]:
```
output_df.to_csv('state_election_data_1976-2020.csv',index=False,float_format='%.2f')
```

Creating a GeoPackage

In [16]:
```
fn = "zip://original_data/cb_2020_us_county_20m.zip"
county_df = geopandas.read_file(fn,dtype={'GEOID':str})
county_df = county_df.drop(columns=['STATEFP','COUNTYFP','COUNTYNS','AFFGEOID','ALAND','AWATER','LSAD','NAMELSAD'])
county_df.head()
```

Out[16]:

	GEOID	NAME	STUSPS	STATE_NAME	geometry
0	01061	Geneva	AL	Alabama	POLYGON ((-86.19348 31.19221, -86.12541 31.182...
1	08125	Yuma	CO	Colorado	POLYGON ((-102.80377 40.00255, -102.79358 40.3...
2	17177	Stephenson	IL	Illinois	POLYGON ((-89.92647 42.50579, -89.83759 42.504...
3	28153	Wayne	MS	Mississippi	POLYGON ((-88.94335 31.82456, -88.91046 31.826...
4	34041	Warren	NJ	New Jersey	POLYGON ((-75.19261 40.71587, -75.17748 40.764...

In [17]:
```
election_df = pd.read_csv('county_election_data_2000-2020.csv',dtype={'FIPS':str})
election_df.head()
```

Out[17]:

	FIPS	gop_2000_votes	dem_2000_votes	totalvotes_2000	gop_2000_prc	dem_2000_prc	gop_minus_dem_prc_2000	gop_2004_votes	dem_2004_votes	t
0	01001	11993.0	4942.0	17208.0	69.69	28.72	40.97	15196.0	4758.0	
1	01003	40872.0	13997.0	56480.0	72.37	24.78	47.59	52971.0	15599.0	
2	01005	5096.0	5188.0	10395.0	49.02	49.91	-0.89	5899.0	4832.0	
3	01007	4273.0	2710.0	7101.0	60.17	38.16	22.01	5472.0	2089.0	
4	01009	12667.0	4977.0	17973.0	70.48	27.69	42.79	17386.0	3938.0	

5 rows × 37 columns

In [18]:
```
county_df = county_df.merge(election_df,how='left',left_on='GEOID',right_on='FIPS')
```

In [19]:
```
county_df.to_file("election.gpkg", layer='county', driver="GPKG")
```