Bellwether County Analysis

For more background see here (https://github.com/BuzzFeedNews/2016-11-bellwether-counties).

In [1]:
```
import pandas as pd
from glob import glob
```

Load CQ data

The data files from CQPress contain presidential election results for every county (or comparable geography) in the country.

In [2]:
```
csvs = glob("../data/cq-data/*/*.csv")
```

In [3]:
```
def parse_csv(file_string):
    df = pd.read_csv(
        file_string,
        skiprows=2,
        na_values=["AreaType"]
    ).rename(columns={
        "RepVotesTotalPercent": "rep_pct",
        "DemVotesTotalPercent": "dem_pct"
    }).dropna(subset=["AreaType", "rep_pct"]) # remove counties with no data
    df[["rep_pct", "dem_pct"]] = df[["rep_pct", "dem_pct"]].astype(float)
    return df[[
        "State", "RaceDate", "Area",
        "rep_pct", "dem_pct"
    ]]
```

In [4]:
```
results = pd.concat([ parse_csv(c) for c in csvs ])\
    .sort_values([ "State", "Area", "RaceDate"])\
    .reset_index(drop=True)
```

In [5]:
```
results.head()
```

Out[5]:

	State	RaceDate	Area	rep_pct	dem_pct
0	Alabama	19721107	AUTAUGA	75.17	22.31
1	Alabama	19761102	AUTAUGA	48.32	49.69
2	Alabama	19801104	AUTAUGA	56.87	38.82
3	Alabama	19841106	AUTAUGA	70.07	28.25
4	Alabama	19881108	AUTAUGA	67.13	31.45

In [6]:
```
results["year"] = results["RaceDate"].str.slice(0, 4).astype(int)
```

In [7]:
```
ALL_YEARS = list(range(1972, 2016, 4))
```

Load national results

In [8]:
```
national_results = pd.read_csv(
    "../data/election_results.csv"
).rename(columns={
    "PctRepublican": "pct_rep",
    "PctDemocrat": "pct_dem"
})
```

In [9]:
```
national_results["year"] = national_results["RaceDate"].astype(str).str.slice(
0, 4).astype(int)
```

In [10]:
```
national_results
```

Out[10]:

	RaceDate	pct_rep	pct_dem	year
0	19721107	60.7	37.5	1972
1	19761102	48.0	50.1	1976
2	19801104	50.7	41.0	1980
3	19841106	58.8	40.6	1984
4	19881108	53.4	45.6	1988
5	19921103	37.4	43.0	1992
6	19961105	40.7	49.2	1996
7	20001107	47.9	48.4	2000
8	20041102	50.7	48.3	2004
9	20081104	45.7	52.9	2008
10	20121106	47.2	51.1	2012

Calculate max miss for each county

For each election, calculate the percentage spread between the Republican and Democratic candidates for each county (and nationally). Then, for each county and election, find the biggest “miss” — calculated as the difference between the county spread and the national spread — over the previous four presidential elections.

In [11]:
```
results["rep_dem_spread"] = results["rep_pct"] - results["dem_pct"]
```

In [12]:
```
national_results["rep_dem_spread"] = national_results["pct_rep"] - national_re
sults["pct_dem"]
```

In [13]:
```
results["national_diff"] = pd.merge(
    results,
    national_results,
    on="year",
    how="left",
    suffixes=[".local", ".national"]
).pipe(lambda x: x["rep_dem_spread.local"] - x["rep_dem_spread.national"])
```

In [14]:
```
results.head()
```

Out[14]:

	State	RaceDate	Area	rep_pct	dem_pct	year	rep_dem_spread	national_diff
0	Alabama	19721107	AUTAUGA	75.17	22.31	1972	52.86	29.66
1	Alabama	19761102	AUTAUGA	48.32	49.69	1976	-1.37	0.73
2	Alabama	19801104	AUTAUGA	56.87	38.82	1980	18.05	8.35
3	Alabama	19841106	AUTAUGA	70.07	28.25	1984	41.82	23.62
4	Alabama	19881108	AUTAUGA	67.13	31.45	1988	35.68	27.88

In [15]:
```
def get_max_miss_four(area):
    max_miss = area["national_diff"].abs().rolling(window=4).max()
    max_miss.index = area["year"]
    return max_miss
```

In [16]:
```
max_misses = results.groupby([ "State", "Area" ])\
    .apply(get_max_miss_four)\
    .unstack()\
    .pipe(lambda x: x[x.columns[3:]])
```

In [17]:
```
max_misses.head()
```

Out[17]:

	year	1984	1988	1992	1996	2000	2004	2008	2012
State	**Area**								
	AUTAUGA	29.66	27.88	30.60	37.64	41.47	49.58	55.04	55.04
	BALDWIN	43.09	39.01	39.01	43.97	48.09	51.52	58.65	59.57
Alabama	**BARBOUR**	21.46	12.89	12.89	9.65	7.69	7.69	8.65	8.65
	BIBB	35.54	26.04	20.48	12.66	22.51	42.11	53.04	53.04
	BLOUNT	36.80	23.71	26.49	34.57	43.29	60.14	76.71	77.83

Calculate historical accuracy

For each election, calculate the average Republican-Democrat spread of the five counties with the smallest maximum miss in the four *prior* presidential elections. Compare that number to the overall national Republican-Democrat spread.

In [18]:
```
def calc_historical_accuracy(year):
    closest = max_misses[year - 4].nsmallest(5)
    closest_prev_results = results[
        results["year"] == year
    ].set_index([ "State", "Area" ]).loc[closest.index]
    mean_error = closest_prev_results["national_diff"].mean()
    return mean_error
```

In [19]:
```
for year in ALL_YEARS[4:]:
    acc = calc_historical_accuracy(year)
    print("{0}: {1:.3f}".format(year, acc))
```

```
1988: 0.592
1992: -0.318
1996: 0.272
2000: 5.654
2004: 2.236
2008: 0.066
2012: -0.768
```

The model struggled a lot in 2000 and a little in 2004, but has come within one percentage point of the national results in five of the past seven elections. Based on this analysis, the five counties to watch in the 2016 election are:

In [20]:
```
max_misses[2012].nsmallest(5)
```

Out[20]:
```
State           Area
Minnesota       DAKOTA       1.70
Michigan        MACOMB       1.92
North Carolina  GRANVILLE    1.93
Michigan        CALHOUN      2.30
Iowa            CEDAR        2.44
Name: 2012, dtype: float64
```