Crime, Education, and Teen Pregnancies

Emily McNichols and
Charlotte Hettrich

CS 2316 FALL 2021

What is the relationship between crime, education, and teen pregnancies in Georgia?

Our Hypothesis

As crime increases,

- education benchmarks decrease, measured using
  - school attendance
  - HOPE eligibility
  - high school completers
  - dropout rates
- teen pregnancies increase

Agenda

- Data Collection
- Data Cleaning
- Insights
- Overall Project Findings
- Project Impact

Data Collection

Downloaded Dataset #1: 10 years of education data from Georgia Governor's Office of Student Achievement

Downloaded Dataset #2: 9 years of population data from US Census Bureau, 2020 population data converted from a PDF from Georgia General Assembly

Downloaded Dataset #3: Shapefile from US Census Bureau

Web Collection #1: Scraped data from Georgia Department of Public Health's Online Analytical Statistical Information System, HTML

Web Collection #2: FBI Crime Data API with agency listing and agency offenses summary endpoints, JSON based API

Data Cleaning

- Collecting education datasets to one DataFrame + related inconsistencies
- Column names were tuples in the teen pregnancy data
- Complex nested dictionary to DataFrame
- Formatting datasets to have the counties as the index and the years as the column names
- Reading to SQL + Excel

Insight #1

Counties with the highest crime and teen pregnancy counts

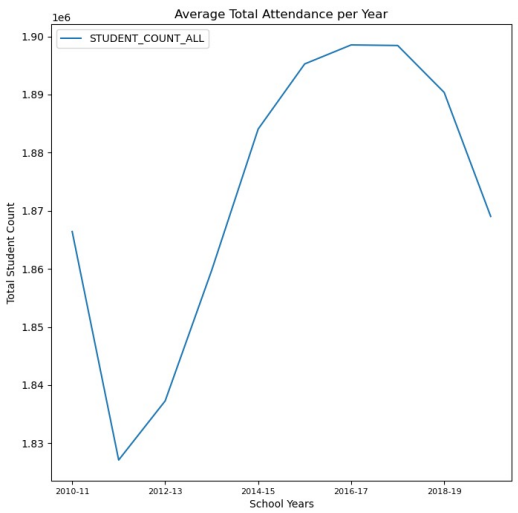
```
for i in range(11):
    t_20_crime = pd.concat([t_20_crime, pd.DataFrame(crime_df[(2010 + i)].nlargest(20).index)], axis = 1)
    t_20_crime.set_index(crime_df[(2010 + i)].nlargest(20).index, inplace = True)
    t_20_crime = pd.concat([t_20_crime, pd.DataFrame(crime_df[(2010 + i)].nlargest(20))], axis = 1)
    t_20_crime.dropna(axis = 0)
    t_20_crime.index = [i for i in range(20)]
```

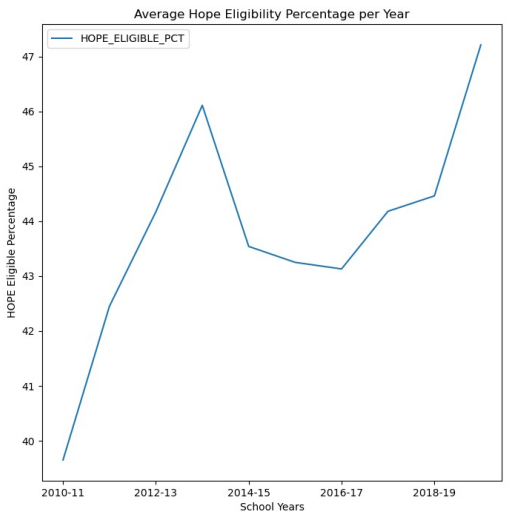
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		County	2010	County	2011	County	2012	County	2013	County	2014	County	2015	County	2016	County	2017	County	2018	County	2019	County	2020
2	1	MUSCOGEE	28959	MUSCOGEE	26805	MUSCOGEE	24548	MUSCOGEE	27031	MUSCOGEE	29005	MUSCOGEE	24891	MUSCOGEE	21629	MUSCOGEE	20321	BIBB	18328	BIBB	9028	BIBB	13369
3	2	GLYNN	6574	GLYNN	6704	FULTON	5879	FULTON	5857	BIBB	18898	BIBB	17568	BIBB	17489	BIBB	17181	MUSCOGEE	16582	MUSCOGEE	8957	DOUGLAS	3992
4	3	FULTON	5712	BIBB	5969	GLYNN	5822	BIBB	5658	FULTON	5811	FULTON	5302	FULTON	5915	FULTON	4788	FULTON	4289	DOUGLAS	4419	TROUP	3820
5	4	BIBB	5502	FULTON	5264	BIBB	5164	GLYNN	5464	GLYNN	5452	DOUGLAS	4496	GLYNN	4111	DOUGLAS	4688	DOUGLAS	4074	FULTON	4257	FULTON	3419
6	5	DOUGLAS	4768	FORSYTH	4402	DOUGLAS	4377	DOUGLAS	4437	WALTON	4920	GLYNN	4359	DOUGLAS	3888	GLYNN	4011	FORSYTH	3893	FORSYTH	3024	COWETA	2749
7	6	FORSYTH	4011	DOUGLAS	4160	TROUP	3826	TROUP	3694	DOUGLAS	4487	TROUP	3983	TROUP	3585	FORSYTH	3768	GLYNN	3658	GLYNN	2842	HALL	2619
8	7	TROUP	3686	TROUP	3739	SPALDING	3390	SPALDING	3322	TROUP	3807	FORSYTH	3160	HALL	3137	HALL	3135	CHATHAM	3254	COWETA	2641	FORSYTH	2484
9	8	HALL	3487	SPALDING	3590	FORSYTH	3256	FORSYTH	3300	HALL	3248	HALL	2992	FORSYTH	3104	HOUSTON	2725	TROUP	3076	HALL	2303	GLYNN	1949
10	9	SPALDING	3444	HALL	3397	HALL	3221	HALL	3244	FORSYTH	3238	TIFT	2597	SPALDING	2736	TIFT	2614	HALL	2877	CHATHAM	2222	SUMTER	1890
11	10	TIFT	2944	SUMTER	3055	WALTON	2880	SUMTER	2955	SPALDING	3072	SPALDING	2534	SUMTER	2581	COWETA	2518	COWETA	2472	SUMTER	1935	BALDWIN	1614
12	11	SUMTER	2895	TIFT	2828	SUMTER	2878	TIFT	2818	SUMTER	2970	WHITFIELD	2502	COWETA	2445	SPALDING	2484	HOUSTON	2241	SPALDING	1477	SPALDING	1483
13	12	GWINNETT	2703	WHITFIELD	2796	WHITFIELD	2681	WHITFIELD	2772	WHITFIELD	2670	COWETA	2491	WHITFIELD	2438	CLAYTON	2434	WHITFIELD	2162	TROUP	1433	CATOOSA	1306
14	13	COWETA	2702	COWETA	2561	TIFT	2664	WALTON	2754	DEKALB	2570	GWINNETT	2391	HOUSTON	2428	WHITFIELD	2182	SUMTER	2105	HOUSTON	1372	WALTON	1236
15	14	WHITFIELD	2520	HOUSTON	2532	COWETA	2546	COWETA	2400	HOUSTON	2520	CLAYTON	2238	TIFT	2188	GWINNETT	2041	SPALDING	1893	WHITFIELD	1344	FAYETTE	1123
16	15	CLAYTON	2284	ROCKDALE	2487	CATOOSA	2273	ROCKDALE	2304	COWETA	2407	WALTON	2218	GWINNETT	2138	CATOOSA	1844	TIFT	1818	LAURENS	1239	LAURENS	991
17	16	CATOOSA	2248	GWINNETT	2343	GWINNETT	2262	CATOOSA	2185	TIFT	2390	HOUSTON	2216	CATOOSA	1955	SUMTER	1814	CLAYTON	1794	TIFT	1220	LEE	914
18	17	ROCKDALE	2186	CLAYTON	2268	ROCKDALE	2144	HOUSTON	2161	CLAYTON	1976	SUMTER	2196	WALTON	1915	TROUP	1712	BALDWIN	1662	LEE	1101	HARALSON	860
19	18	BALDWIN	2089	CATOOSA	2237	HOUSTON	2110	LEE	1953	CATOOSA	1969	CATOOSA	2141	CLAYTON	1854	ROCKDALE	1640	ROCKDALE	1520	FAYETTE	1084	MADISON	783
20	19	HOUSTON	1834	MADISON	1874	CLAYTON	2066	BALDWIN	1876	ROCKDALE	1965	ROCKDALE	1942	BALDWIN	1774	BALDWIN	1503	WARE	1439	HARALSON	1065	BEN HILL	771
21	20	MADISON	1783	BALDWIN	1818	WARE	1632	GWINNETT	1874	GWINNETT	1914	BALDWIN	1921	ROCKDALE	1631	WALTON	1490	WALTON	1437	BALDWIN	1040	BANKS	721

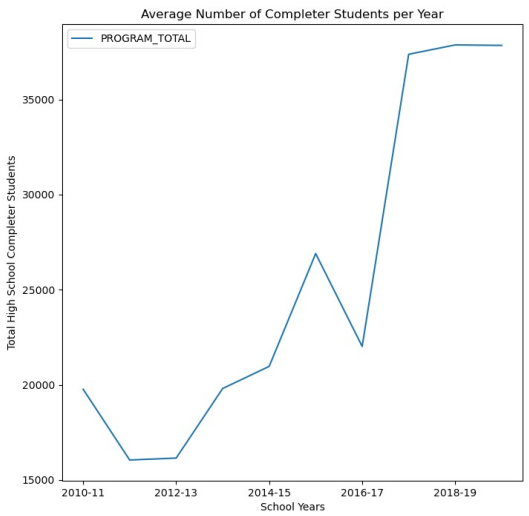
Insight #2

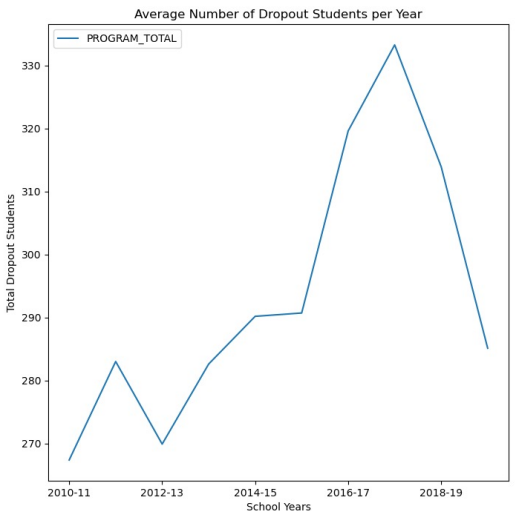
Examining trends in education statistics

```
total_completer_df = pd.read_excel("education_data.xlsx", sheet_name = "total_completer")
total_completer_df = total_completer_df.groupby(["LONG_SCHOOL_YEAR"])["PROGRAM_TOTAL"].mean().to_frame()
total_completer_df = total_completer_df.reset_index()

fig, completer_ax = plt.subplots(1, 1, figsize = (8,8))

total_completer_plt = total_completer_df.plot(kind = "line", x = "LONG_SCHOOL_YEAR", y = "PROGRAM_TOTAL", ax = completer_ax)
total_completer_plt.set_ylabel("Total Completer Students")
total_completer_plt.set_xlabel("School Years")
total_completer_plt.set_title("Average Number of Completer Students per Year")
plt.savefig("completer_plot.png")
plt.show()
```







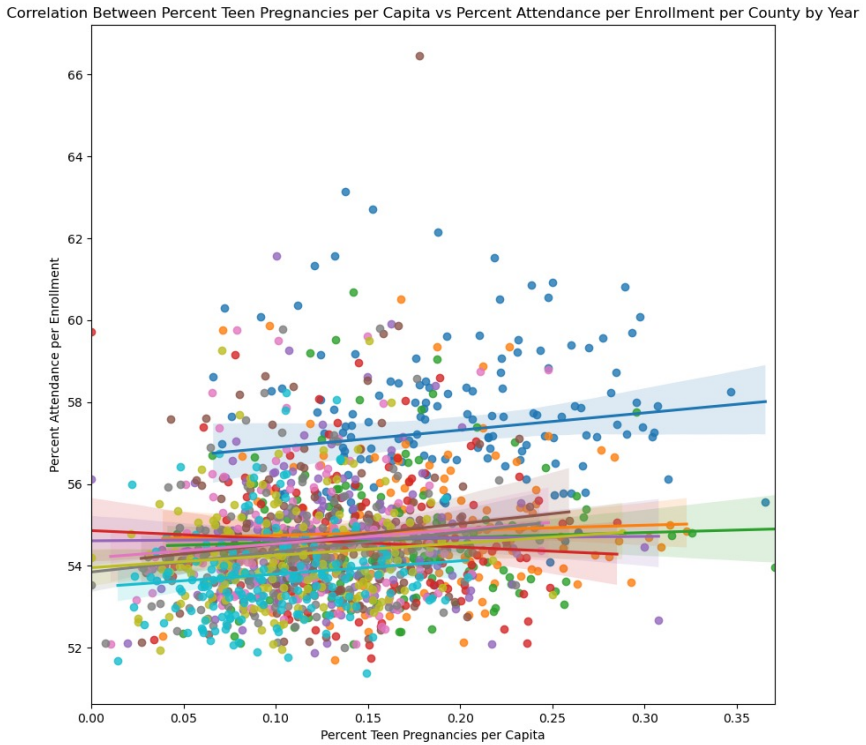


Insight #3

Correlating teen pregnancy with school attendance

```
fig, prego_corr_ax = plt.subplots(1, 1, figsize = (8,8))

for i in range(2011, 2021):
    prego_corr_ax = sns.regplot(x=percent_prego_df[i], y=county_attendance_perc_df[i])
```

```
correlation_prego_attend_df_by_county = percent_prego_corr_df.corrwith(county_attendance_perc_df, axis = 1).to_frame()
correlation_prego_attend_df_by_county.rename(columns = {0: "Correlation Between Teen Pregnancy and Attendance by County"},
                                             inplace = True)
correlation_prego_attend_df_by_year = percent_prego_corr_df.corrwith(county_attendance_perc_df, axis = 0).to_frame()
correlation_prego_attend_df_by_year.rename(columns = {0: "Correlation Between Teen Pregnancy and Attendance by Year"},
                                           inplace = True)
```



(

	Correlation Between Teen Pregnancy and Attendance by County
APPLING	0.748228
ATKINSON	0.779880
BACON	0.405524
BAKER	0.304595
BALDWIN	-0.076747
...	...
WHITFIELD	0.901143
WILCOX	0.366195
WILKES	0.403846
WILKINSON	0.189754
WORTH	0.502704

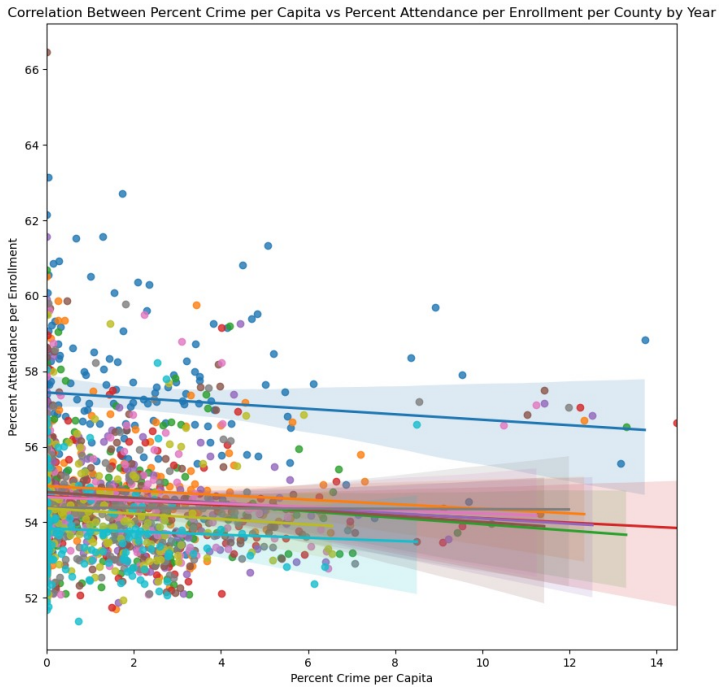
	Correlation Between Teen Pregnancy and Attendance by Year
2011	0.138014
2012	0.047467
2013	0.047148
2014	-0.073458
2015	0.012173
2016	0.129375
2017	0.103452
2018	0.155493
2019	0.108214
2020	0.104833

Insight #4

Correlating crime with school attendance

```
fig, crime_corr_ax = plt.subplots(1, 1, figsize = (8,8))

for i in range(2011, 2021):
    crime_corr_ax = sns.regplot(x=percent_crime_corr_df[i], y=county_attendance_perc_df[i], ax = crime_corr_ax)
```

```
correlation_crime_attend_df_by_county = percent_crime_corr_df.corrwith(county_attendance_perc_df, axis = 1).to_frame()
correlation_crime_attend_df_by_county.rename(columns = {0: "Correlation Between Crime and Attendance by County"},
                                             inplace = True)
correlation_crime_attend_df_by_year = percent_crime_corr_df.corrwith(county_attendance_perc_df, axis = 0).to_frame()
correlation_crime_attend_df_by_year.rename(columns = {0: "Correlation Between Crime and Attendance by Year"},
                                          inplace = True)
```



(

County	Correlation Between Crime and Attendance by County
APPLING	0.220429
ATKINSON	0.234698
BACON	0.109570
BAKER	0.238548
BALDWIN	-0.079012
...	...
WHITFIELD	0.780415
WILCOX	NaN
WILKES	-0.208226
WILKINSON	-0.034733
WORTH	-0.377859

	Correlation Between Crime and Attendance by Year
2011	-0.091478
2012	-0.088338
2013	-0.126381
2014	-0.094989
2015	-0.101193
2016	-0.093978
2017	-0.057193
2018	-0.005153
2019	-0.078675
2020	-0.048735

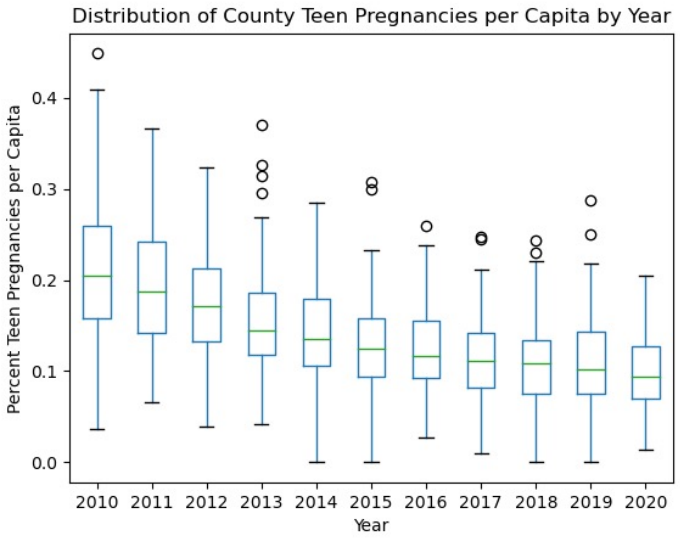
Insight #5

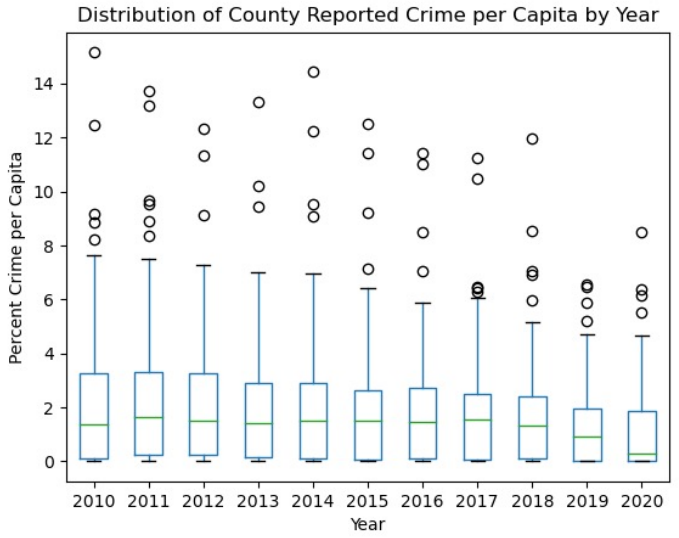
Observing trends with distributions

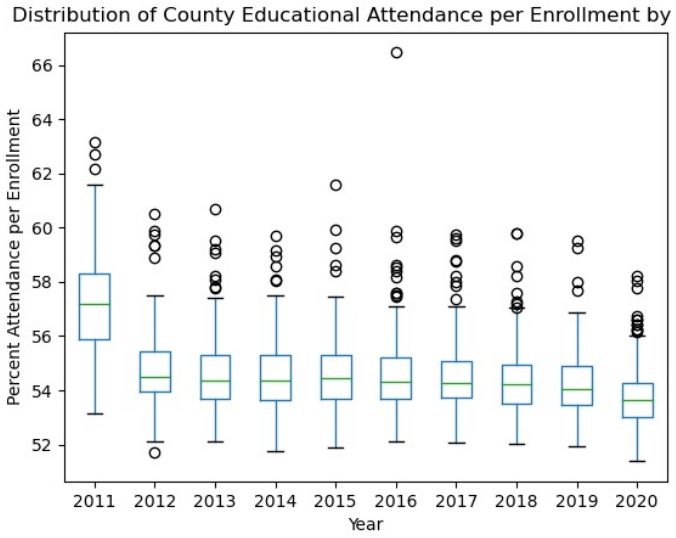
```
# Get per capita percentages for both teen pregnancies and crime
percent_prego_df = ((prego_df / pop_df) * 100).round(4)
percent_crime_df = ((crime_df / pop_df) * 100).round(4)
```

Crime per capita by county

County	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
APPLING	3.1690	3.5051	3.0460	3.3097	2.2845	2.9691	3.8800	4.0677	4.9768	2.7086	1.4097
ATKINSON	0.0000	1.6268	0.3150	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5389	2.3896
BACON	2.8295	3.4814	3.6771	4.1529	4.9025	5.5739	5.6585	4.6319	4.0155	2.1677	3.3573
BAKER	0.0583	0.5436	1.0676	0.0000	1.4294	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
BALDWIN	4.5707	4.0243	3.3520	4.0625	3.4677	4.2174	3.9197	3.3427	3.7021	2.3168	3.6850
...	...	...	...	...	...	...	...	...	...	...	...
WHITFIELD	2.4527	2.7167	2.6042	2.6984	2.5901	2.4135	2.3351	2.0934	2.0739	1.2846	0.1167
WILCOX	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
WILKES	0.4043	0.8212	0.4957	0.4234	0.7639	0.8475	1.0501	1.0041	0.8313	1.0944	6.1370
WILKINSON	0.0000	0.0000	0.0000	0.0000	0.0000	0.0220	0.0222	0.0000	0.0000	0.0000	0.0000
WORTH	1.7841	1.7246	1.5692	2.4986	2.8468	2.5619	2.6470	2.4454	2.4136	2.4794	2.1507





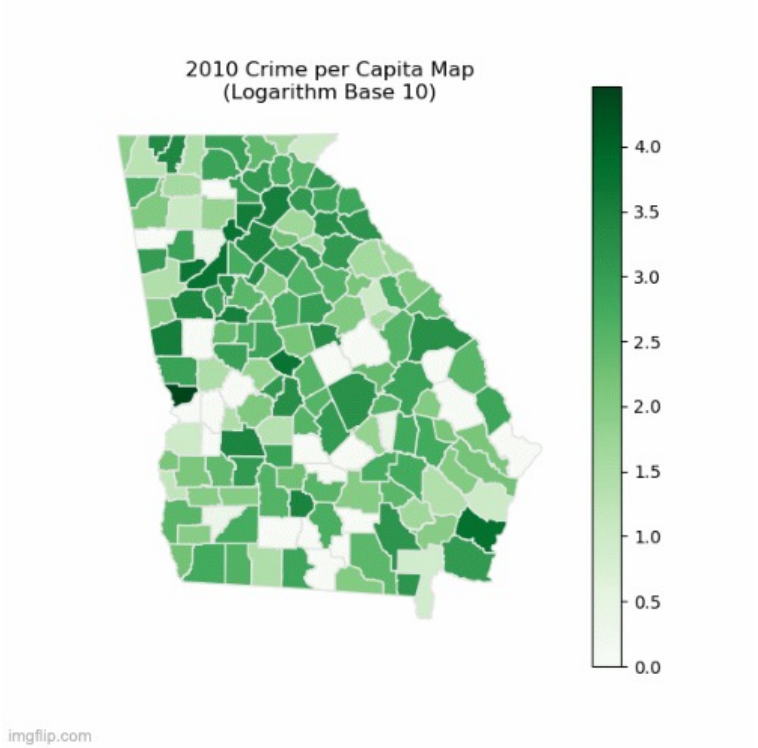


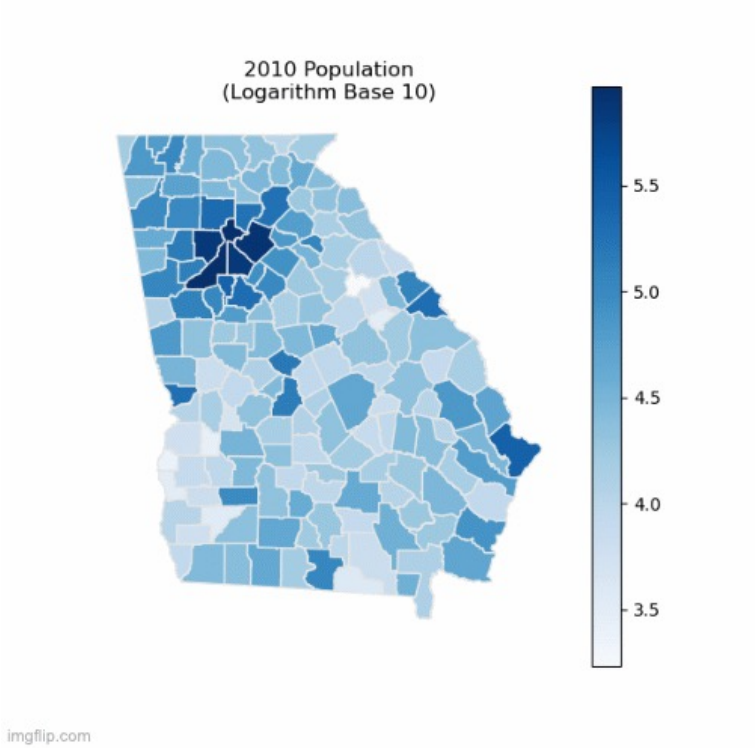
Insight #6

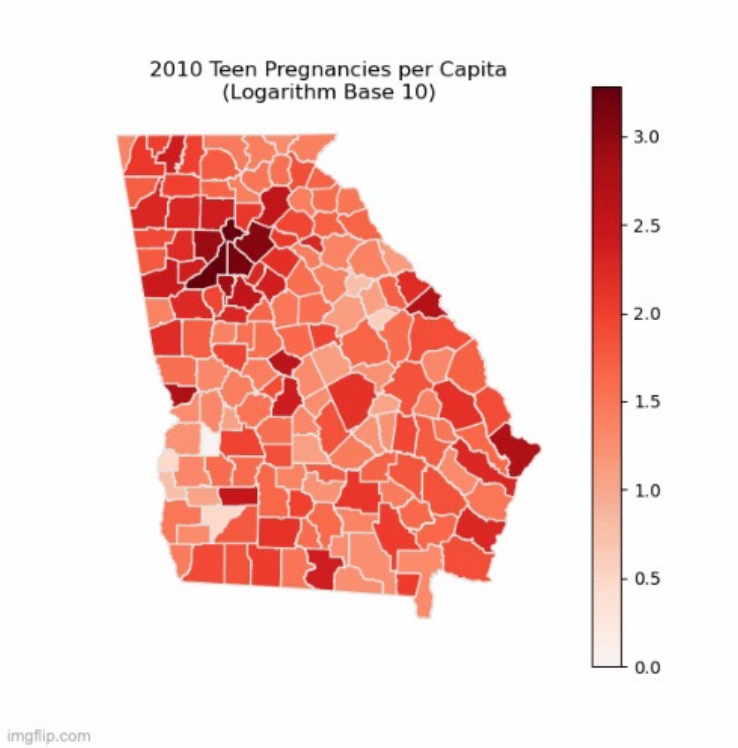
Mapping out crime and teen pregnancies per capita

```
crime_map_df = gdf.merge(crime_df_log, left_on = ['NAME'], right_index = True)

for i in range(2010, 2021):
    fig, ax = plt.subplots(1, figsize = (6,6))
    crime_map_df.plot(column = i, cmap = 'Greens', linewidth = 1, ax = ax, edgecolor = '0.9', legend = True)
    ax.axis('off')
    plt.title(str(i) + ' Crime per Capita Map\n(Logarithm Base 10)')
    fig.savefig(str(i) + "crime_map.png")
    fig.show()
```

```
crime_df = crime_df.astype(float)
crime_df_log = crime_df.where(crime_df == 0, lambda x: np.log10(x), inplace = False)
```



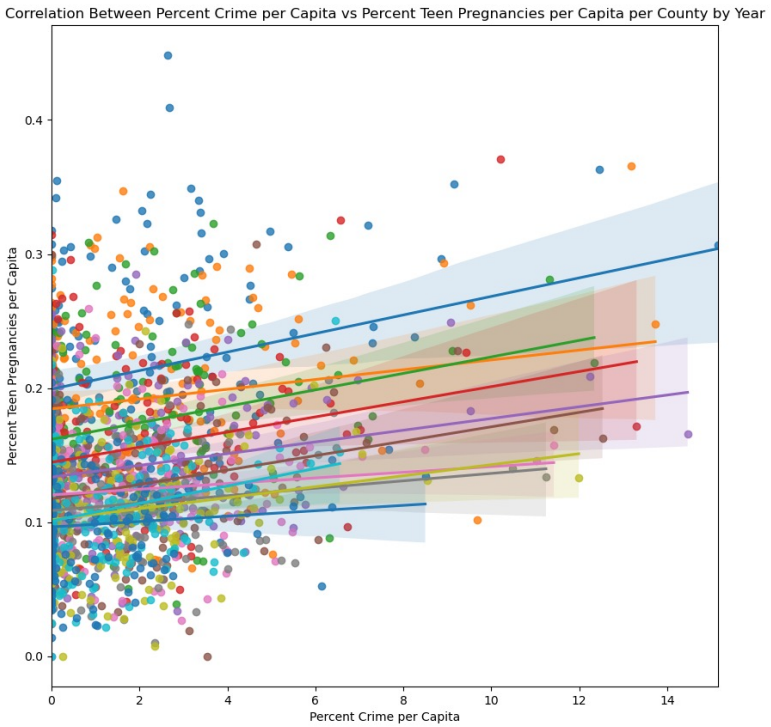




Insight #6

Supplementing side by side comparison with a direct correlation plot

```
fig, prego_crime_ax = plt.subplots(1, 1, figsize = (8,8))

for i in range(2010, 2021):
    prego_crime_ax = sns.regplot(x=percent_crime_df[i], y=percent_prego_df[i], ax = prego_crime_ax)
```



Overall Project Findings

Trends:

- Teen pregnancy, dropout rate is decreasing
- HOPE eligibility, graduation, attendance is increasing
- Crime has been relatively stable

Results:

- Population/population density primary factor in crime and teen pregnancies
- There is no strong correlation between crime, education, and teen pregnancies
- Crime, teen pregnancy, and education likely Pareto distributions

Project Impact

Beyond Material:

- Extra modules + deeper knowledge of Pandas
- Greater acclimation to our operating system

Greatest Challenge: Data collection + cleaning

Have a wonderful holiday!