Election Data Analytics Project Write-up
By Daniel He, November 2020

Project Github Repository: https://github.com/danielhe9999/electionanalytics

Introduction: I followed the 2020 election results very closely, and it was an extremely close race by all means. I was curious if there are any interesting correlations between various factors and how people in certain counties voted. For example, are lower socioeconomic counties more likely to vote for a certain candidate? Are differing number of COVID cases truly a political stance? I decided to do this project out of curiosity as well as put my data analytics skills in practice. I wanted to create a predictive model to determine these correlations. If I were on the campaign team for a politician, these insights would be useful in improving my strategy.

Technical Skills Used: Throughout the project, I gained hands-on experience with Python libraries (Pandas, sklearn, BeautifulSoup), Microsoft Excel, and web scraping. Below is my report summarizing my experience and some of the key insights I gained.

Question: How is the 2020 presidential election voter outcome correlated with population size, education level, poverty level, and number of COVID cases in each county?

Data Collection: The goal of the project is to discover any patterns at the county level for the US Election such as correlations with poverty, population, and education to create a predictive model. To gather the data, I came across three websites containing information about the election results, number of COVID cases, and poverty and education statistics. Next, I performed web scraping using the BeautifulSoup library in Python to transform the data into CSV format for easy data manipulation to come.

1. https://www.nytimes.com/interactive/2020/11/03/us/elections/results-alaska.html (Election Results)
2. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/arizona (COVID info)
3. https://www.ers.usda.gov/data-products/county-level-data-sets/ (Poverty and education info)

Data Cleaning/Manipulation: The next step after data collection is to clean the data, which has the goal of making sure there are no sources of error such as duplicates, missing information, or inconsistencies. Since the final CSV contained data from three different sources, I came across multiple inconsistencies:

1. Not all counties in the US are labeled as counties. For example, Alaska had districts and Louisiana has Parishes. Counties in Alaska had mismatching names for all three sources, so they had to be removed from the final dataset.
2. Virginia counties in one of the sources were formatted poorly.
3. Duplicates in county names. There are counties in two different states with the same county name. Therefore, the county was formatted as "County, State Abbreviation" (King, WA)
4. To perform linear regression in my predictive modeling, it would be useful for me to quantify the way a county votes. Therefore, I manipulated the data so there was a numerical spectrum on voting outcomes, with 0 being 100% Biden and 100 being 100% Donald Trump. A number near 50 would indicate a close race in that county.

Sample of first several rows of the final dataset

	County + State	Result	% bachelor's 2014-18	Population Change 2010-2019	% in Poverty	Median Household Income	COVID Cases per 100k
2	McKenzie,ND	84	26.1	2.34311	9.2	76,224	5,358.10
3	Loving,TX	92.5	0	2.0119	3.3	78,637	591.7
4	Williams,ND	84	24.8	1.66404	6.5	79,354	7,339.90
5	Hays,TX	45	37.4	1.45493	13.2	68,180	2,968.40
6	Wasatch,UT	63	40.6	1.44185	5.3	85,380	5,162.70
7	Trousdale,TN	74	16.8	1.43307	18.6	49,280	16,705.10
8	Comal,TX	72	35.3	1.4291	7.1	76,523	1,956.40
9	Kendall,TX	77	40.9	1.4121	7.5	90,937	999.3
10	Hudspeth,TX	68	6.9	1.40726	17.8	42,671	3,254.20
11	Sumter,FL	68	31.1	1.40457	9.3	55,324	2,439.20
12	Dallas,IA	51	49.4	1.40004	5.2	87,267	5,984.80
13	Osceola,FL	43	20.5	1.39251	13.4	50,473	4,442.60
14	Williamson,TX	49	41.1	1.38534	6.4	87,817	1,893.80
15	St. Johns,FL	64	43.7	1.38391	6.6	81,925	2,941.80
16	Forsyth,GA	68	51.7	1.38212	5	105,921	2,300.90
17	Fort Bend,TX	45	46.1	1.37515	7.9	92,310	2,337.40
18	Mountrail,ND	69	20.6	1.36806	10.4	67,162	9,103.80

Analysis Results:

To study different voter groups, I split the dataset into 3 sections: The first includes counties that had at least a 55% majority win for Donald Trump. The second includes counties that had at least a 55% majority win for Joe Biden. The third group are counties that ended in close races, with the winner receiving between 50 and 55% of the vote.

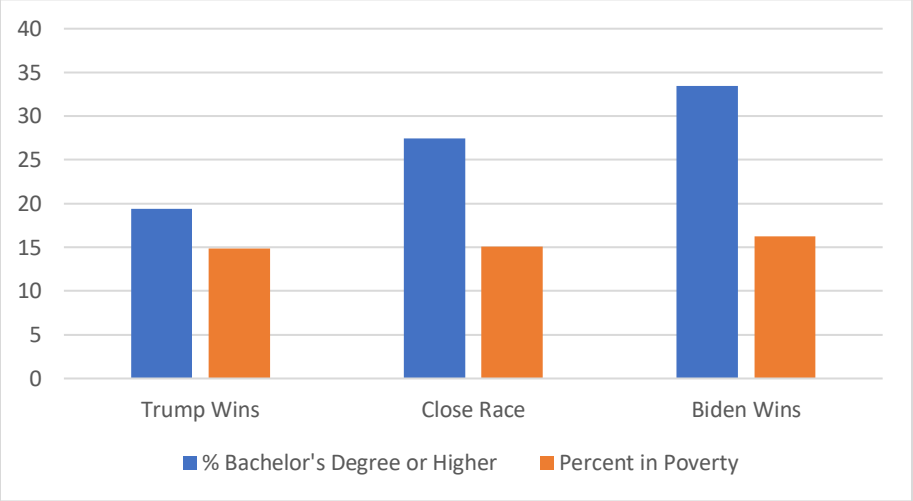
	% Bachelor's or higher	Population Change 2010-2019	Percent in Poverty
Trump wins	19.433	1.00498	14.8616
Close Race	27.4367	1.0339	15.08
Biden wins	33.444	1.0339	16.218

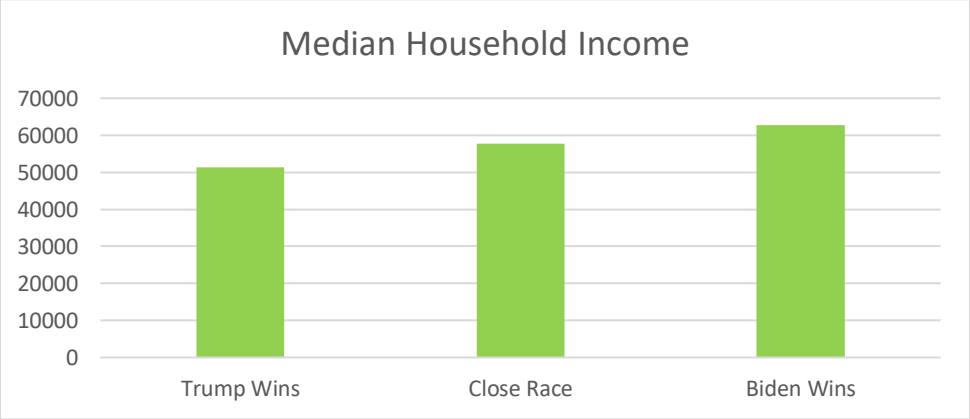
	Median Household Income	COVID cases per 100k
Trump wins	51273.56	4033.686
Close Race	57660.34	3611.25
Biden wins	62711.83	3520.176

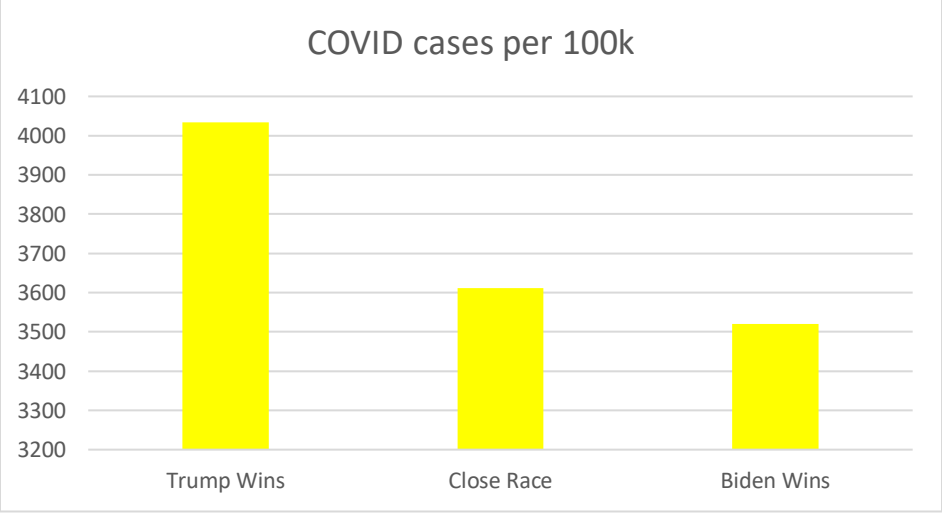
From the results above and visualizations below, it is clear that Biden voters are more educated with an average of 33% of adults holding at least a bachelor's degree compared to Trump voters (19%). Unsurprisingly, Biden voters on average earned more in household income. Counties that are more likely to live in poverty are also more likely to vote for Biden. Finally, counties which voted for Trump have a higher COVID case count per 100k population, which could indicate that mask wearing could be a political issue.

All of these statistics are correlations and do not necessarily indicate causation. However, given that nearly all of the counties' information are in the dataset, the sample size is large and results are quite convincing.

Data Visualizations







Predictive Modeling

Linear Regression to predict how big are the margins of victories for each candidate.

The ***result*** variable is how clear Trump won the county. 0 = Biden landslide, 1 = Trump landslide
x1 = % holding bachelor's degree 2014-2018
x2 = Population change from 2010 - 2019
x3 = Percent of population in poverty
x4 = Estimated median household income
x5 = Number of COVID cases per 100k

Model: predicted ***result*** = 120.3231 – 0.9814***x1*** + 6.7308***x2*** – 1.4382***x3*** – 0.0003***x4*** + 0.0002***x5***

	coef	std err	t	P>|t|	[0.025	0.975]
const	120.3231	3.227	37.282	0.000	113.995	126.651
% bachelor's degree 2014-18	-0.9814	0.034	-28.951	0.000	-1.048	-0.915
Population Change 2010-2019	6.7308	2.862	2.351	0.019	1.118	12.343
Percent in Poverty	-1.4382	0.057	-25.067	0.000	-1.551	-1.326
Median Income	-0.0003	3.14e-05	-10.923	0.000	-0.000	-0.000
Cases per 100k	0.0002	0.000	1.680	0.093	-2.86e-05	0.000

The multiple regression summary indicates that the chances of Trump winning a county is positively correlated with population change and number of COVID cases per 100k and negatively correlated with % residents holding a bachelor's degree or higher, percent in poverty, and median income. All of the covariates except the COVID cases are statistically significant at the 0.05 level.