US County Data

Adam Curry
Summer Term - DSC680-T302 Applied Data Science (2217-1)

County level data can offer a lot of insight to the health of a community. Some publications have done full analyses to determine which states are "richer/poorer" and categorizes them based on their political affiliation [1][2]. Others have used disparate variables such as income and homicide rates to find the US counties with highest crime rates [3]. Reading these articles made me wonder what other variables could be influencing their analysis. These articles often focus on income, poverty levels, and employment data. But what if private sectors contribute more than public sectors, or if people working from home tend to have higher incomes than those that have to commute. I found a dataset that contains disparate variables including: 2020/2016 presidential voting (at the candidate level), commute methods, (carpool, transit, walk, etc.), income (total and per capita), demographic, COVID cases/deaths (as of 11/2020), employment data (sector, unemployment), and population data. These variables could offer some interesting findings if incorporated together. The political leanings could help determine if a county leans republican or democrat.

Also, since the pandemic, there has been a significant divide in the political climate [4]. Some attribute the divide to the pandemic and have even signaled that COVID may've hurt the election results [9]. I have remained independent throughout my life (voting Democrat, Libertarian, and Republican), and like to think I maintain a diverse group of friends that lean to one side or the other. The contentious election brought some serious divide between my circle of friends, some bickering and insulting each other on social media based on their political leanings.

The purpose of this analysis is to answer some basic hypothesis questions. The point isn't to challenge any sort of policy or political leanings, but simply to find if a dataset with disparate variables

can show influence on the way a county leans politically. Also, since there is voting data attached, I will conduct analysis of the voting records to find COVID's role in the election and to uncover findings using Benford's law. Analysis of election results using Benford's Law has been done several times [4][5][6]. What I will do different, is to attempt to increase the orders of magnitude. The initial dataset from other's researchers tends to be limited all states or to counties within a specific state (i.e. orders of magnitude of 100-1,000). I have not seen an attempt to increase the orders of magnitude and expand to countrywide county voting records, which could increase the orders of magnitude to 10-1,000,000.

Some of the questions this analysis seeks to answer include:

- Hypothesis Question # 1 - Did covid numbers influence the 2020 election?
- Hypothesis Question # 2 - Is it possible to predict if a county is republican or democrat without using non-voting variables (i.e., covid cases/deaths, poverty, profession)?
- Hypothesis Question # 3 - Will classifying counties and combining them with other like counties across the country help increase the levels of magnitude required to meet Benford's Law?
- Hypothesis Question # 4 - Are the observed values the same as the theoretical Benford's law values?

Some questions I anticipate from more technically minded individuals include:

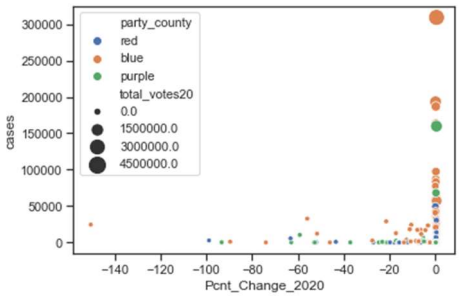
- What steps did you take to validate the county level metrics?
- How can you be certain you met Benford's Law basic principles?
- What were the counties that were outliers in the scatterplot?
- Could removing the outliers help the Covid analysis?
- Why did you leave percent voted in your PCA?
- Would removing LA county from your PCA have any impact on the classification?
- Why didn't you use the "purple" classification in Benford's Law analysis?

- How did you come up with the calculation for Benford's law?
- How did you come up with the chi-squared critical value?
- Did you notice any specific variable that stood out as the biggest influence in the PCA?
- How did you account for missing values in the dataset?

The initial dataset was gathered from Kaggle where the author aggregated data from FiveThirtyEight and DataWorld. I ran some basic cleanup steps to remove null values and duplicate values (see Appendix image A and B). Then I created a few additional variables for the classification model. First, I had to define what constitutes a blue vs red county. To do this, I did a basic conditional statement: if the county voted more republican than democrat in 2020, then they were considered red for 2020; if the county voted more democrat than republican in 2016, then they were considered blue for 2016. Then I added an additional conditional statement that aggregated the two voting terms together to determine if they were "red", "blue", and "purple". Purple would mean they were different both voting terms. The summary statistics showed that I had a lot more republican counties than democrat counties. I was a little worried about these totals, as the blue counties were dwarfed by the number of red counties. Then I remembered that the population of blue counties is most likely larger than red counties, as they tend to be larger cities and coastal regions. Adding the average population revealed that the numbers line up pretty well. It is interesting to see the min and max values range for each category. One "blue" county had 2.3 million votes while another had 490. This seems to be a good indicator that the orders of magnitude increased above 100 - 1,000. In fact, they actually increased to 10 - 1,000,000 (see appendix image C). I also added additional voting metrics: percent of population voted 2020, percent of population voted 2016, percent of republican vote changes from 2016 to 2020, percent of democrat vote changes from 2016 to 2020. These values could show that the county voted more in 2020 than in 2016, which will help in the COVID analysis.

Now that the dataset was defined and cleaned, I was able to begin my analysis. The first hypothesis question was related to voter turnout and COVID data. To start, I explored the correlation metrics between the voting and covid data. Total votes were highly correlated to covid case data, but were also negatively correlated to the percentage change year over year (same with covid deaths) (see appendix image D). Since the "percent change year over year" was negatively correlated with total cases and deaths, I wanted to see how this looked on a scatterplot along with population.

As you can see, most of the change is focused around zero, with some of the negative influenced counties showing smaller populations, and higher case ranges. This data contained cases up to 11/2021, so these counties could've been shown less interest in voting as a result of the pandemic.



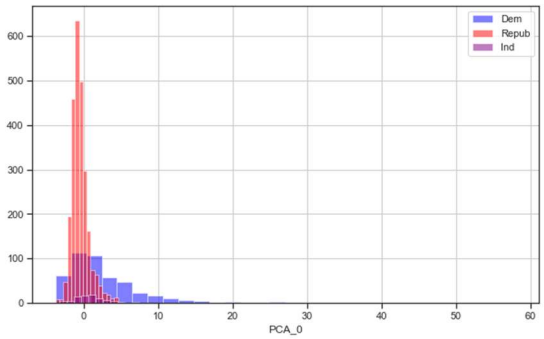
Next, I ran these metrics through a multilinear regression model, to see which variables had the most influence on the Pcnt_Change_2020, broken out by county party type.

Hypothesis Question # 1 - Did covid numbers influence the 2020 election?

It's hard to make a determination based on two variables (cases and deaths), but based on these findings, COVID cases and deaths don't appear to have much impact on the overall election results. Broken into three categories shows a range of R^2 0.165 - 0.269. The p-value shows COVID cases may've impacted the change in votes in red counties (p-value < .000), but in every other county type, COVID cases and deaths were all greater than .05.

The next step was to classify the data using non-voting variables. This was done using a combination of techniques. First, I conducted a principal component analysis to reduce the number of features to two (down from > 30). Then I ran the data through sklearn's DecisionTreeClassifier and RandomForestClassifier.

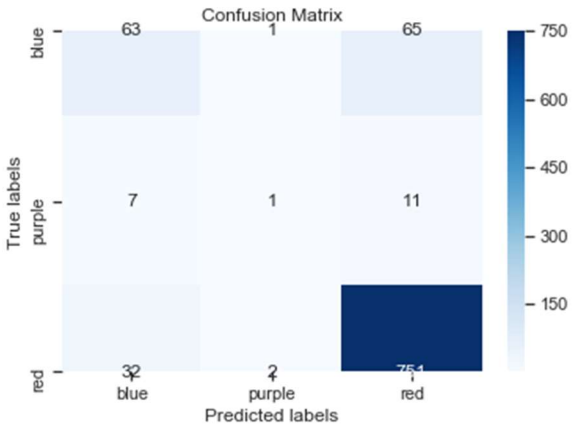
The principal component analysis revealed that component one had the most influence on the categories. This histogram shows the data are heavily right skewed. The histogram in the appendix (image E) shows the data are symmetric. Therefore, I chose to limit to only two components, instead of adding a third, as this would just add noise to the data.



After trimming my dimensions down to two, I ran the data through a decision tree and random forest to predict if the county was democrat or republican (based on my defined logic of percent vote 2016 vs 2020). I found that the republican party frequency was easy to predict, but the democrat and independent counties weren't. This is most likely a result of the number of "red" counties vs. the number of "blue" counties.

Hypothesis Question # 2 - Is it possible to predict if a county is republican or democrat without using non-voting variables (i.e., covid cases/deaths, poverty, profession)?

The short answer is "yes". Using PCA, I was able to reduce the features from > 30 down to two components. Joining the components back to the counties revealed that red counties were easier to predict. This is most likely a result of the number of counties in the population, which means I may not have enough blue counties for this classification. Perhaps several years' worth of data could help. Regardless, the model's average accuracy was good: the decision tree predicted 83% of the records correct and 87% correct in the random forest. The image on the right is a confusion matrix of predicted values.



Finally, I created a function that would allow the data to be sliced to view Benford's distribution on my choice of variables. It would then plot the observed values compared to the theoretical first digit values and calculate and return the chi-squared to determine if the distribution were significant. To begin, I used variables: 2020 votes, 2016 votes, covid data, and Total Pop broken out by county type. The first three variables are potentially "human entered" numbers and the TotalPop is a naturally occurring number(non-human). I chose TotalPop to demonstrate how the distribution should look, if there are any findings that violate Benford's law in the voting/covid analysis. Some assumptions that must be met concerning Benford's Law - the numbers need to be random and not assigned, with no imposed minimums or maximums; the numbers should cover several orders of magnitude, and the dataset should be large; recommendations in the literature call for 100 to 1,000 samples as a minimum, though Benford's law has been shown to hold true for datasets containing as few as 50 numbers [5].

Hypothesis Question # 3 - Will classifying counties and combing them with other like counties across the country help increase the levels of magnitude required to meet Benford's law?

As a novice in this area, it appears the answer is "yes". Using county level data across the country, instead of county level data at the state level increases the orders of magnitude from 10 - 1000 to 10 to 1,000,000. With my current understanding, I THINK this data set qualifies based on the orders of magnitude and the numbers should be random.

Hypothesis Question # 4 - Are the observed values the same as the theoretical Benford's law values?

Overall, the answer is yes. It appears most of the observations are less than the critical chi-squared value (15.51), meaning the difference in observed versus theoretical values is not statistically significant. However, there were a few items that stood out. Several of the red county variables were above the 15.51 threshold. I found this odd, especially with values such as "TotalPop", as this shouldn't be a "human" influenced number. This makes leads me to three possible outcomes: a - the data is questionable (remember I took this from a Kaggle dataset), b - I don't fully grasp the concept of Benford's law, c - the population is not significant and/or my definition of red and blue states isn't robust enough. If its option a, then I am super frustrated at this point in my analysis. If it's option b, so be it... at least I have some sort of understanding of Benford's law and can apply it on future projects (potentially work-related projects around fraud). If its option c, then I can move on to the rest of the analysis. However, I want to explore an untouched dataset using only the votes as they stand for every county. So, my final step in this project was to analyze the data without the classification I applied earlier.

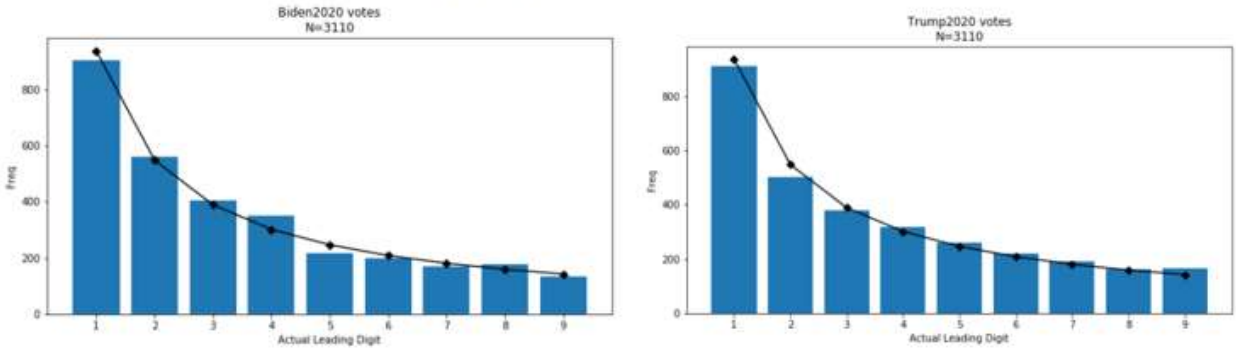
The findings revealed that the chi-squared test statistic is only greater than 15.51 on one of the eight calculations, which is Biden2020. Meaning, the difference in the distribution is statistically significant and the theoretical values are different from the observed values (i.e. “smoking gun” crowd would be happy with this). However, this doesn't mean there was election fraud, as I've pointed out, there are several expert statisticians who have indicated the use of Benford's law in election data isn't

advised. Also, I only spot checked the dataset. There may be data discrepancies missing or incorrect values. However, it is interesting that at the county level, where it appears the principals of Benford's law are met, there is a significance in observed voting totals versus theoretical voting totals for the 2020 election.

```
count of Biden2020 - 3110
[936.20328651 547.64381566 388.55947085 301.39014046 246.25367521
 208.20451575 180.3549551  159.08434481 142.30579564]

Biden2020 - full - votes

expected counts: [936.20328651 547.64381566 388.55947085 301.39014046 246.25367521
 208.20451575 180.3549551  159.08434481 142.30579564]
actual counts: [903 559 405 350 216 199 176 170 132]
Chi-squared Test Statistic = 15.673
Critical value at a P-value of 0.05 is 15.51.
count of Trump2020 - 3110
[936.20328651 547.64381566 388.55947085 301.39014046 246.25367521
 208.20451575 180.3549551  159.08434481 142.30579564]
```



The "next steps" I could take for this analysis would require a much stronger understanding of first digit analysis. Entire research papers have been dedicated to analyzing second digit and even last digit analysis [6][7]. Perhaps I will take some time to understand what each digit could represent and take the Benford's Law analysis a step further. For the classification analyses, I would need additional years' worth of data to get a robust model. The limited dataset I have doesn't seem to offer enough observations to properly classify non-red counties.

Appendix/References:

[1] Rich Exner, cleveland.com. (2015, September 18). *Blue states, red states; politics and household income*. cleveland. https://www.cleveland.com/datacentral/2015/09/blue_states_red_states_rich_st.html.

[2] Hendrickson, M. (2021, June 29). *Are the 10 Poorest U.S. States Really Republican?* Forbes. https://www.forbes.com/sites/markhendrickson/2012/06/07/are-the-10-poorest-u-s-states-really-republican/?sh=496d1ceb6e06.

[3] Bryant, A. (2017, June 5). *10 U.S. counties with the highest murder rate*. Police1. https://www.police1.com/ambush/articles/10-us-counties-with-the-highest-murder-rate-kerWgaEUmxJkn74J/.

[4] Dimock, M., & Wike, R. (2020, November 13). *America is exceptional in the nature of its political divide*. Pew Research Center. https://www.pewresearch.org/fact-tank/2020/11/13/america-is-exceptional-in-the-nature-of-its-political-divide/.

[5] Chatziapostolou, E. (2020, August 23). *Fraud detection using Benford's Law (Python Code)*. Medium. https://towardsdatascience.com/frawd-detection-using-benfords-law-python-code-9db8db474cf8.

[6] standupmaths. (2020, November 10). *Why do Biden's votes not follow Benford's Law?* YouTube. https://www.youtube.com/watch?v=etx0k1nLn78.

[7] Golbeck, J. (2020, November 14). *Benford's Law Does Not Prove Fraud in the 2020 US Presidential Election*. Medium. https://jengolbeck.medium.com/benfords-law-does-not-prove-fraud-in-the-2020-us-presidential-election-cc81715bfbda.

[8] *Benford's Law*. from Wolfram MathWorld. (n.d.). https://mathworld.wolfram.com/BenfordsLaw.html.

[9] https://fivethirtyeight.com/features/how-much-did-covid-19-affect-the-2020-election/

Image A:

Out[241]:

	county	state	percentage16_Donald_Trump	percentage16_Hillary_Clinton	total_votes16	votes16_Donald_Trump	votes16_Hillary_Clinton	percentage2
0	Unassigned	AL	None	None	None	None	None	
1	Unassigned	AK	None	None	None	None	None	
2	Unassigned	AZ	None	None	None	None	None	
3	Unassigned	AR	None	None	None	None	None	
4	Unassigned	CA	None	None	None	None	None	
5	Unassigned	CO	None	None	None	None	None	
6	Unassigned	CT	None	None	None	None	None	
7	Unassigned	DE	None	None	None	None	None	
8	Unassigned	FL	None	None	None	None	None	
9	Unassigned	GA	None	None	None	None	None	
10	Unassigned	HI	None	None	None	None	None	
11	Unassigned	ID	None	None	None	None	None	

Image B:

Out[243]:

	county	state	percentage16_Donald_Trump	percentage16_Hillary_Clinton	total_votes16	votes16_Donald_Trump	votes16_Hillary_Clinton	percen
0	Androscoggin	ME	0.509	0.415	55340.0	28189.0	22975.0	
1	Aroostook	ME	0.555	0.383	34963.0	19419.0	13377.0	
2	Cumberland	ME	0.337	0.601	171249.0	57697.0	102935.0	
3	Franklin	ME	0.482	0.427	16382.0	7900.0	7001.0	
4	Hancock	ME	0.428	0.504	31983.0	13682.0	16107.0	
5	Kennebec	ME	0.481	0.444	65999.0	31753.0	29296.0	
6	Knox	ME	0.397	0.540	23021.0	9148.0	12440.0	
7	Lincoln	ME	0.454	0.478	21432.0	9727.0	10241.0	
8	Oxford	ME	0.521	0.391	31094.0	16214.0	12172.0	
9	Penobscot	ME	0.519	0.409	80228.0	41601.0	32832.0	
10	Piscataquis	ME	0.591	0.339	9144.0	5403.0	3098.0	
11	Sagadahoc	ME	0.431	0.494	21596.0	9304.0	10679.0	

Image C:

party_county	cnt_	sum_blue20	sum_blue16	sum_red20	sum_red16	sum_total16	sum_total20	max_total_votes16	min_total_votes16	max_total_votes20	min_total_votes20	AvgPopPerCounty	AvgRedChangePcnt	AvgBlueChangePcnt
blue	455	455	455	-	-	64,410,457	72,622,203	2,314,275	490.0	4,139,895	282	376015.6	12.84%	4.71%
purple	89	58	31	31.000	58.000	8,670,360	9,762,802	1,201,934	186.0	2,063,663	194	241026.705	0.27%	5.82%
red	2566	0	0	2,566.000	2,566.000	54,206,267	62,503,529	625,720	64.0	601,261	66	49793.7922	15.59%	5.27%

Image D:

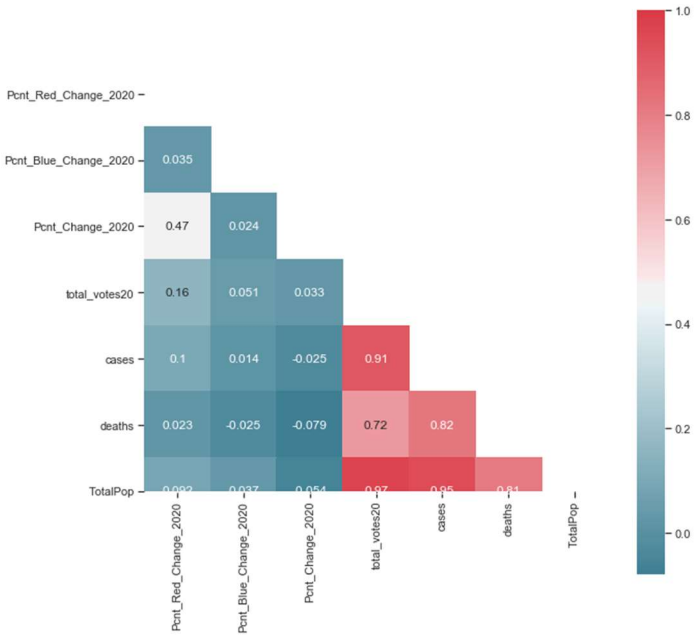
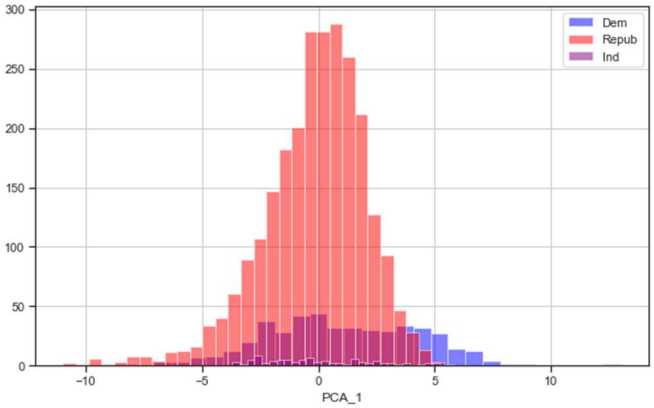

Pcnt_Red_Change_2020
Pcnt_Blue_Change_2020
Pcnt_Change_2020
total_votes20
cases
deaths
TotalPop
0.035
0.47
0.024
0.16
0.051
0.033
0.1
0.014
-0.025
0.91
0.023
-0.025
-0.079
0.72
0.82
1.0
0.8
0.6
0.4
0.2
0.0

Image E:


Dem
Repub
Ind
300
250
200
150
100
50
0
−10
−5
0
5
10
PCA_1