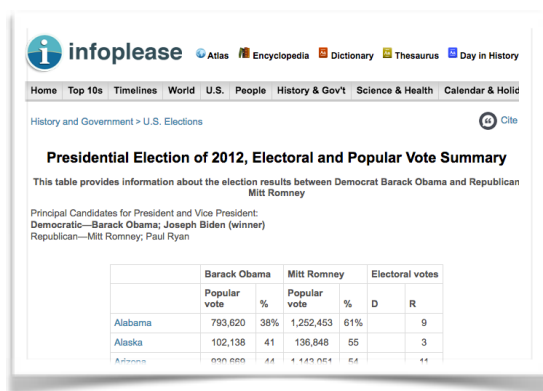


Final Project: 2012 Crime Rate Statistics VS 2012 Presidential Election Party Choice

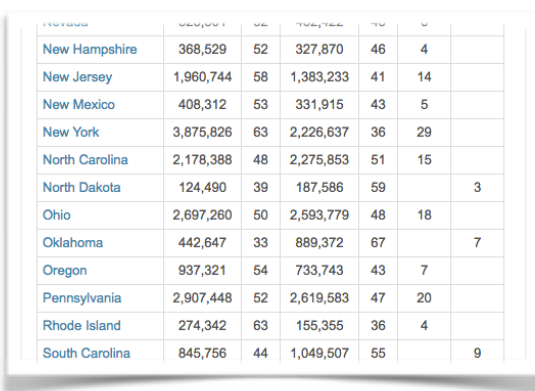
Originally, I had proposed to look at FBI crime rate statistics and tweets containing “can’t sleep” because I wanted to see if states that had a higher ratio of “can’t sleep” tweets correlated with states that had high crime rate statistics. However, I did not end up creating that for my project because of the limitations of the twitter api. Still interested in looking at crime rate statistics, I searched for another idea. Recently, the election had passed and I thought it would be a perfect time to explore some election data. So, I landed upon the idea of looking at 2012 FBI crime rate statistics per state and the 2012 presidential election party choice vote per state. The first question that came to mind was: Does a certain party choice correlate with a higher crime rate statistic? i.e. Do states that vote Democrat have higher crime rate statistics? Upon trying to gather the data for my first question, another question came to mind. Would median household income also play a factor in the crime rate statistics and party choice? Would higher income states have higher crime rates (or lower crime rates) and correlate with a certain party choice? To clarify, I realize correlation does not mean causation. My intention for this project was to see if there were any association between these variables.

Data Sources:



The screenshot shows the infoplease website with a navigation bar at the top. Below the navigation bar, there is a section titled "History and Government > U.S. Elections". The main content area features the heading "Presidential Election of 2012, Electoral and Popular Vote Summary". Below this heading, a brief description states: "This table provides information about the election results between Democrat Barack Obama and Republican Mitt Romney". It also lists the principal candidates for President and Vice President: Democratic—Barack Obama; Joseph Biden (winner) and Republican—Mitt Romney; Paul Ryan. The table below displays the election results for each state, including the popular vote and electoral votes for both candidates.

	Barack Obama		Mitt Romney		Electoral votes	
	Popular vote	%	Popular vote	%	D	R
Alabama	793,620	38%	1,252,453	61%		9
Alaska	102,138	41	136,848	55		3
Arizona	930,680	64	1,415,051	54		11



The screenshot shows a table with 7 columns: State, Democrat, Democrat %, Republican, Republican %, and Electoral votes. The table lists the following states and their corresponding election results:

State	Democrat	Democrat %	Republican	Republican %	Electoral votes
New Hampshire	368,529	52	327,870	46	4
New Jersey	1,960,744	58	1,383,233	41	14
New Mexico	408,312	53	331,915	43	5
New York	3,875,826	63	2,226,637	36	29
North Carolina	2,178,388	48	2,275,853	51	15
North Dakota	124,490	39	187,586	59	3
Ohio	2,697,260	50	2,593,779	48	18
Oklahoma	442,647	33	889,372	67	7
Oregon	937,321	54	733,743	43	7
Pennsylvania	2,907,448	52	2,619,583	47	20
Rhode Island	274,342	63	155,355	36	4
South Carolina	845,756	44	1,049,507	55	9

The datasets came from a variety of places. The first dataset I looked for was for the 2012 election results. I was looking for the states and states’ party choice. Initially, the dataset I was trying to manipulate was the election results from the New York Times exit polls. When I tried use beautiful soup on this dataset I realized I couldn’t complete this this task because the webpage had javascript running behind it and therefore, I need to look for another webpage to scrape. The webpage I finally can be found on this link: “<http://www.infoplease.com/us/government/2012-presidential-election-vote-summary.html>”. I made sure to cross-list the

numbers from the New York Times exit polls to make sure the data was valid. The format it was in was html. There isn't an exact size of records, but the data I took from the webpage was the state, the voting percentage towards the Democrat candidate and the voting percentage toward the Republican candidate. That would be about 51 of each table column.

State	Area	Population	Violent crime	Murder and nonnegotiable manslaughter	Rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft
ALABAMA		2,002,280									
	Area actually reporting	98.9%	16,642	283	981	4,517	11,081	133,153	37,131	88,221	7,801
	Estimated total	100.0%	17,024	285	990	4,582	11,187	134,817	37,524	89,402	7,891
	Metropolitan Statistical Area	1,001,240									
	Area actually reporting	94.1%	2,813	28	177	384	2,244	22,441	5,530	15,763	1,346
	Estimated total	100.0%	2,972	30	186	380	2,378	23,541	5,817	16,518	1,396
	Cities outside metropolitan area	641,490									
	Area actually reporting	99.2%	1,674	27	119	77	1,451	10,438	4,108	5,539	775
	Estimated total	100.0%	1,687	27	120	79	1,462	10,520	4,140	5,603	777
	Nonmetropolitan counties	4,022,022									
	Area actually reporting	99.2%	1,674	27	119	77	1,451	10,438	4,108	5,539	775
	Estimated total	100.0%	1,687	27	120	79	1,462	10,520	4,140	5,603	777
	State Total	4,022,022									
	Area actually reporting	98.9%	16,642	283	981	4,517	11,081	133,153	37,131	88,221	7,801
	Estimated total	100.0%	17,024	285	990	4,582	11,187	134,817	37,524	89,402	7,891
	Rate per 100,000 inhabitants		84.6	14.1	49.0	226.1	553.0	6,678.0	1,856.0	4,456.0	390.0

State	Area	Population	Violent crime	Murder and nonnegotiable manslaughter	Rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft
ALABAMA		2,002,280									
	Area actually reporting	98.9%	16,642	283	981	4,517	11,081	133,153	37,131	88,221	7,801
	Estimated total	100.0%	17,024	285	990	4,582	11,187	134,817	37,524	89,402	7,891
	Metropolitan Statistical Area	1,001,240									
	Area actually reporting	94.1%	2,813	28	177	384	2,244	22,441	5,530	15,763	1,346
	Estimated total	100.0%	2,972	30	186	380	2,378	23,541	5,817	16,518	1,396
	Cities outside metropolitan area	641,490									
	Area actually reporting	99.2%	1,674	27	119	77	1,451	10,438	4,108	5,539	775
	Estimated total	100.0%	1,687	27	120	79	1,462	10,520	4,140	5,603	777
	Nonmetropolitan counties	4,022,022									
	Area actually reporting	99.2%	1,674	27	119	77	1,451	10,438	4,108	5,539	775
	Estimated total	100.0%	1,687	27	120	79	1,462	10,520	4,140	5,603	777
	State Total	4,022,022									
	Area actually reporting	98.9%	16,642	283	981	4,517	11,081	133,153	37,131	88,221	7,801
	Estimated total	100.0%	17,024	285	990	4,582	11,187	134,817	37,524	89,402	7,891
	Rate per 100,000 inhabitants		84.6	14.1	49.0	226.1	553.0	6,678.0	1,856.0	4,456.0	390.0

The second dataset that I used was FBI crime reporting data tables. The FBI has an enormous amount of data. One of the things they keep track of is the crime rates in the US every year. The data table I choose to use can be found here: https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/tables/5tabledatadecpdf/table_5_crime_in_the_united_states_by_state_2012.xls . I was particularly interested in the violent crime rate and property crime rate per state. The crime rate given is a rate per 100,000 inhabitants. There was more interesting data in the tables such as the rate of crimes being not only split by state but by counties i.e. metropolitan areas, cities outside the metropolitan area, and nonmetropolitan counties. To obtain the data you can download the data table as an excel file, then convert it into a csv file to manipulate in python. The records I extracted from this are again about 51 of each table column (states, violent crime rates for that that state, and property crime rates for that that state)

Postal Code	Name	Poverty Est	90% CI Lower	90% CI Upper	Poverty Pct	90% CI Lower	90% CI Upper	Poverty Est	90% CI Lower	90% CI Upper
01	US	48,760,123	48,528,543	48,991,701	15.9	15.8	16.0	16,396,863	16,275,868	16,517,858
1	AL	896,515	880,205	912,825	19	18.7	19.3	306,451	298,963	313,939
2	AK	77,494	74,043	80,945	10.8	10.3	11.3	28,125	26,359	29,891
4	AZ	1,145,911	1,177,400	1,214,462	18.7	18.4	19.0	430,178	418,881	441,475
5	AR	560,928	549,034	572,822	19.6	19.2	20.0	196,464	190,326	202,602
6	CA	6,323,433	6,274,914	6,371,952	17	16.9	17.1	2,164,589	2,135,262	2,193,916
8	CO	688,715	674,161	703,269	13.6	13.3	13.9	220,133	211,406	228,860
9	CT	370,537	359,310	381,764	10.6	10.3	10.9	114,613	108,840	120,386
10	DE	113,316	108,283	118,349	12.7	12.1	13.3	38,462	36,164	40,760
11	DC	112,719	107,296	118,142	18.8	17.9	19.7	51,615	48,610	54,620
12	FL	3,248,276	3,212,213	3,284,329	17.2	17	17.4	1,011,096	989,246	1,032,946
13	GA	1,852,459	1,826,052	1,878,866	19.2	18.9	19.5	672,149	655,922	688,376
15	HI	159,988	153,828	166,148	11.8	11.3	12.3	51,557	48,450	54,664
16	ID	250,203	242,232	258,174	16	15.5	16.5	86,672	82,360	90,984
17	IL	1,847,371	1,823,585	1,871,157	14.7	14.5	14.9	621,972	606,995	636,949

Once I had begun my project I thought about adding more information, such as the census median household income, so that became my third data set. The dataset can be found here : <https://www.census.gov/did/www/saige/data/statecounty/data/2012.html> . The dataset contains state and county estimates for 2012. The particular variable that I was interested in was the estimate of median household income. The data came in an excel file. The way I used it was by converting it into a csv file. The amount of records I took from the data set were about 51 of each table column I needed (state and the state's median household income).

Data Processing Steps:

The first part of my project began by fetching the html for the election results. This was done using a function called `step1_fetch_election`. The url for the 2012 election results was the input. The function read in the url lib request for “<http://www.infoplease.com/us/government/2012-presidential-election-vote-summary.html>” into a variable called `response`. I then read the response html into a variable called `html`. Lastly, I wrote the response html into `step1.html` which generated my output, an output file called `step1.html`. I did the last step so I could parse the html in another function. In order to use this function I called it in main.

```
# 3. Find all table
table = soup.find_all("table")[1]
# print(table)
rows= table.find_all("tr")
# print(rows)

#4. Loop through each row,
#use beautiful soup function and/or regular expressions
state_dict2={}
for row in rows:
    columns=row.find_all("td")
    if len(columns) == 0:
        continue
    state= columns[0].text
    obama_vote_percent = columns[2].text
    romney_vote_percent = columns[4].text
    # print(rows)
    # print(state)
    # print(obama_vote_percent)
    # print(romney_vote_percent)

    state_dict2[state.lower()] = (obama_vote_percent,romney_vote_percent)
return(state_dict2)
```

For the second step, parsing the `step1.html` file, I created a function called `step2_extract_electiondata`. This function read in the `step1.html`, the input of this function, and began to parse it using BeautifulSoup. First, I found the table I needed. Then I found each row in that table and looped through them. While looping through the rows I created a dictionary called `state_dict2` which holds each state, it's Obama vote percentage, and it's Romney vote percentage. Before returning the `state_dict2` I made the state names lowercase to avoid any

issues when using state as a key. Lastly, the output returns `state_dict2`. In order to use this function I call it in main and store it in the variable `election_datadict`, so that I can use it later in main.

```
def read_medianincome_file(filename):
    median_income = dict() # create an empty dictionary
    with open(filename, 'r', newline='') as input_file:
        region_reader = csv.DictReader(input_file, delimiter=',', quotechar='"')
        for row in region_reader:
            median_income[row["Name"].lower()] = row["Median Household Income"]
            #create a dictionary that contains the name of the state as the key
            # make the median household income for that state value
    return median_income
```

Following those two functions, I created another function called `read_medianincome_file`. This reads in 2012 census median household income taking the file `est12US.csv` as input. Before reading the rows of the file, I created a dictionary called

`median_income`. Once I read through the rows of the csv I created the key and value for the `median_income` dictionary. The key for the dictionary was row “Name” which held the state name. I made this row lower case to prevent issues using when using the state as a key. The value for this dictionary was row “Median HouseHold Income” from the `est12US.csv`. To read the csv I used a `csv.DictReader`. This function returned the `median_income` dictionary as an output. I then called this function in my main and stored it in the variable `state_income`.

Once I called the three functions I earlier defined, I opened the input csv data file in my main. The input of this file was the `table_5_crimerate_2012.csv` that held the FBI crime rate statistics for 2012. Next, I opened up a new csv that would later contain all the data I needed to create my visualizations. In my output csv in the fieldnames I created the rows i would need. The rows are as followed: STATE,VIOLENT CRIMES, PROPERTY CRIMES, Democrat, Republican, Party Choice, Median HouseHold Income, Violent Crimes Democrat, Violent

Crimes Republican, Property Crimes Democrat, Property Crimes Republican, and Scaled Median Household Income. I then looped through the rows of the csv and wrote out to them using crime_stats_writer and crime_stats_reader. I then began to clean the data I brought in from the FBI crime rate statistics. The state row contained some digits in the names because of footnotes. In order to get rid of these digits I used a regular expression. I also made the state names lower case so I could use the same key for the dictionaries I created in the functions. I then added the violent crimes and property crimes for each state.

After I cleaned the FBI data that I brought into my output csv, I began to join the dataset together

```
row["Violent Crimes"] = row["Property Crimes"]

if row["STATE"] not in election_datadict: #the fbi had states that weren't necessarily states i.e. Puerto Rico
    continue
row["Democrat"] = (election_datadict[row["STATE"]][0]).rstrip("%") # some numbers had percentages on them
row["Republican"] = (election_datadict[row["STATE"]][1]).rstrip("%") #needed to index what part of the dict
# I needed to make sure I go the right number for each state.

#coding if the state voted democrat and republican
if row["Democrat"] > row["Republican"]: # if the demo number is larger than the rep then
```

using an “if - not in” statement for the dictionaries. This way states that weren’t in the election_datadict would not be used. The reason for this was there were states/

areas in the FBI statistics that aren’t necessarily states and aren’t counted in elections, for example Puerto Rico. Therefore, doing this eliminated areas that posed this problem. Using the states both in the state row and in the election_datadict I created two rows, Democrat and Republican. Row Democrat held the vote percentage for Democrats and row Republican held the percentage for Republicans. I stripped the percentage symbols from both of these, so that I could make later use of the numbers. I then used these two rows to create a row called Party Choice. This row was coded 0 or 1. If the voting percentage was greater for Democrat than Republican then the party choice was given the number 1 which meant the state voted Democrat, if not it was given 0 which meant the state voted Republican. I then use the state_income dictionary created earlier and the row state as a key to it to create the row Median Household Income. I

used a regular expression to remove the commas from the Median Household Income.

```
# used states from my output csv as a key to my dictionary
row["Median Household Income"] = re.sub(r"[^0-9]", "", state_income[row["STATE"]])
# scaling the median household income so it'll look cooler on the map (:
row["Scaled Median Household Income"] = float(row["Median Household Income"]) - 37179 + 1
#keeps track of the rows
```

```
#basically splitting up violent crimes into democrat or republican
#coding the violent crimes rep/dem
if row["Party Choice"] == 1:
    row["Violent Crimes Democrat"] = row["VIOLENT CRIMES"]
else:
    row["Violent Crimes Republican"] = row["VIOLENT CRIMES"]

#to solve visualization issues
# coding the property crimes rep/dem
if row["Party Choice"] == 1:
    row["Property Crimes Democrat"] = row["PROPERTY CRIMES"]
else:
    row["Property Crimes Republican"] = row["PROPERTY CRIMES"]
```

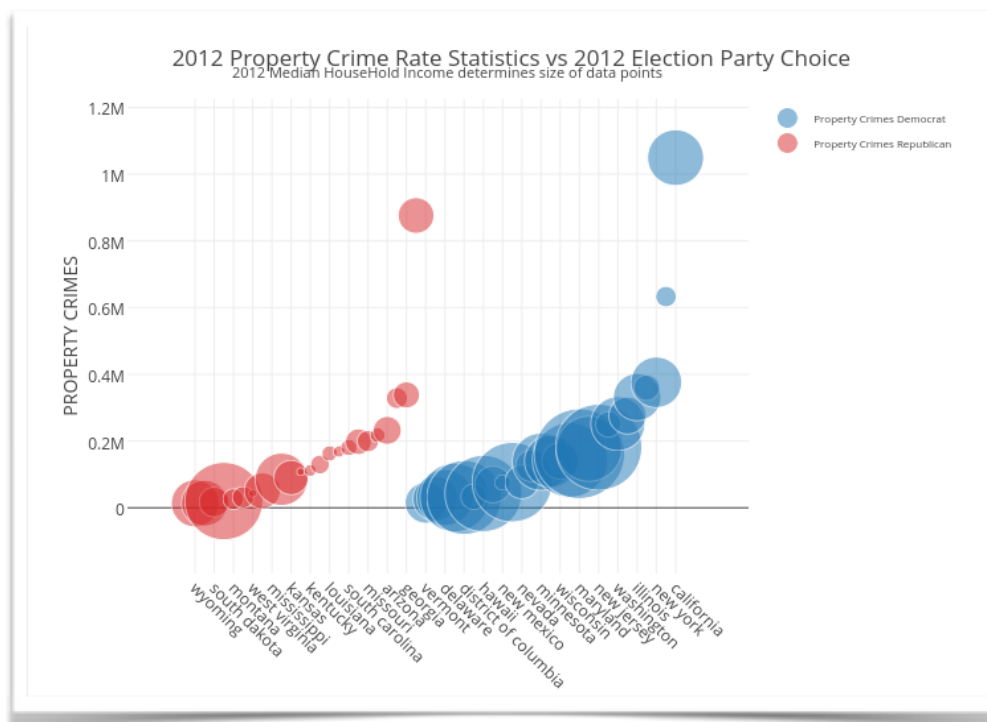
When I first began to create my visualization in [plot.ly](#) I believed that my data was in the correct form need to create the desired scatterplots I wanted. So, in order to solve those issues I used party choice to separate the property crimes and violent crimes into Republican and Democrat crimes. I explained this more in the visualization section. Another, issue that came about was the

visualization of the Median Household Income. In order to show income properly I decided to rescale it by doing: row["Scaled Median Household Income"] = float(row["Median Household Income"]) - 37179 + 1 . I explained more of the changes I made due to visualization in the visualization section.

Visualizations & Findings:

Initially, when I was gathering the data I had begun to generate assumptions on what my data would present. I didn't realize what a difference a visualization would make with the data. The visualization part of this project presented the information in a compelling way that it produced a story out of the data. Since my original intention was to show a relationship between the variables, I felt the best way to represent the data would be to use scatterplots. My choice in platform was [plot.ly](#). Visualizing the data posed some challenges because the way in which I had brought the data to the csv was posing a problem in creating the scatterplots in plot.ly. For example, I coded party choice (Democratic or Republican) 1 or 0. I thought I could plot the states against crime rates and then color the points on the scatter plot blue or red based on the number I assigned them in the code, but because [plot.ly](#) had limitations in doing this I had to re-manipulate my code. I created 4 more columns in my csv. Instead of using 0 or 1 coding, I recoded the violent crimes to be Republican or Democrat and I did the same for property crimes i.e. I had a column called Violent Crimes Democrat and Violent Crimes Republican. Another problem I had was properly showing the median income per household. In order to scale income properly, I decided to create another column, "Scaled Median HouseHold Income", that turned "Median HouseHold Income" into a float and subtracted the minimum household income from that float and then added one. This scaling clarified median household income in the visualization i.e. it proportionally differentiated the size of the data points better. Once I fixed these issues I constructed my two scatterplots.

I produced 2 scatterplots. The first scatterplot is a plot of States vs Property Crime Rate statistics. The data points' size depended on the Scaled Median HouseHold Income per state i.e. the state's circle appeared bigger, relatively, to other states if their Median HouseHold income

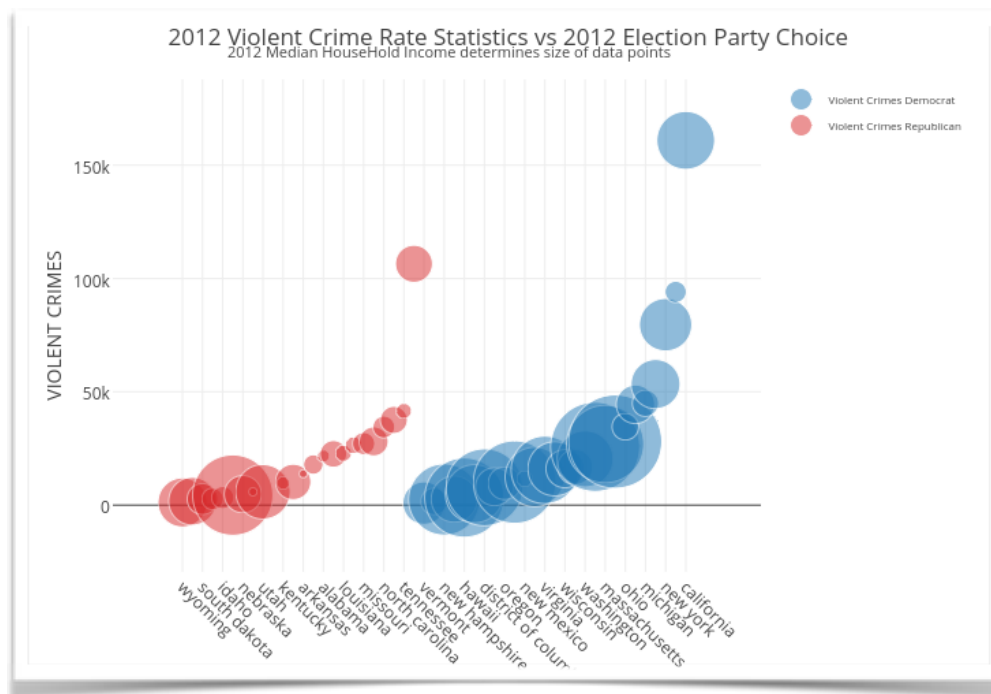


was larger. The color of the data points tells us whether the state voted Democrat or Republican in the 2012 Presidential Election. In the scatter plot you will see two trends based on the color of the party choice of the state. The first notable thing is that in 2012 states that voted Democrat had higher median

household incomes. This was surprising because I assumed that Republican states would have

higher median household incomes since they typically place a heavier weight on economic issues. The second thing I noticed was Democrat states seemed to have slightly higher property crime rates. Under the FBI data declaration, property crime includes the offenses of burglary, larceny-theft, motor vehicle theft, and arson. Perhaps, a reason as to why those states that voted Democrat had higher property crimes can be related to the fact that these states also had higher median household incomes. Therefore, their higher median household incomes may have made them frequent targets of property crimes.

The second scatterplot I produced was plot of States vs Violent Crime Rate statistics. The data points' size, again, depends on the Scaled Median HouseHold Income per state. The color of the data points tell us whether the state voted Democrat or Republican. The difference in median household incomes between Democrat states and Republican states was again noticeable. The notable finding here was states that voted Democrat appear to have much higher violent crime



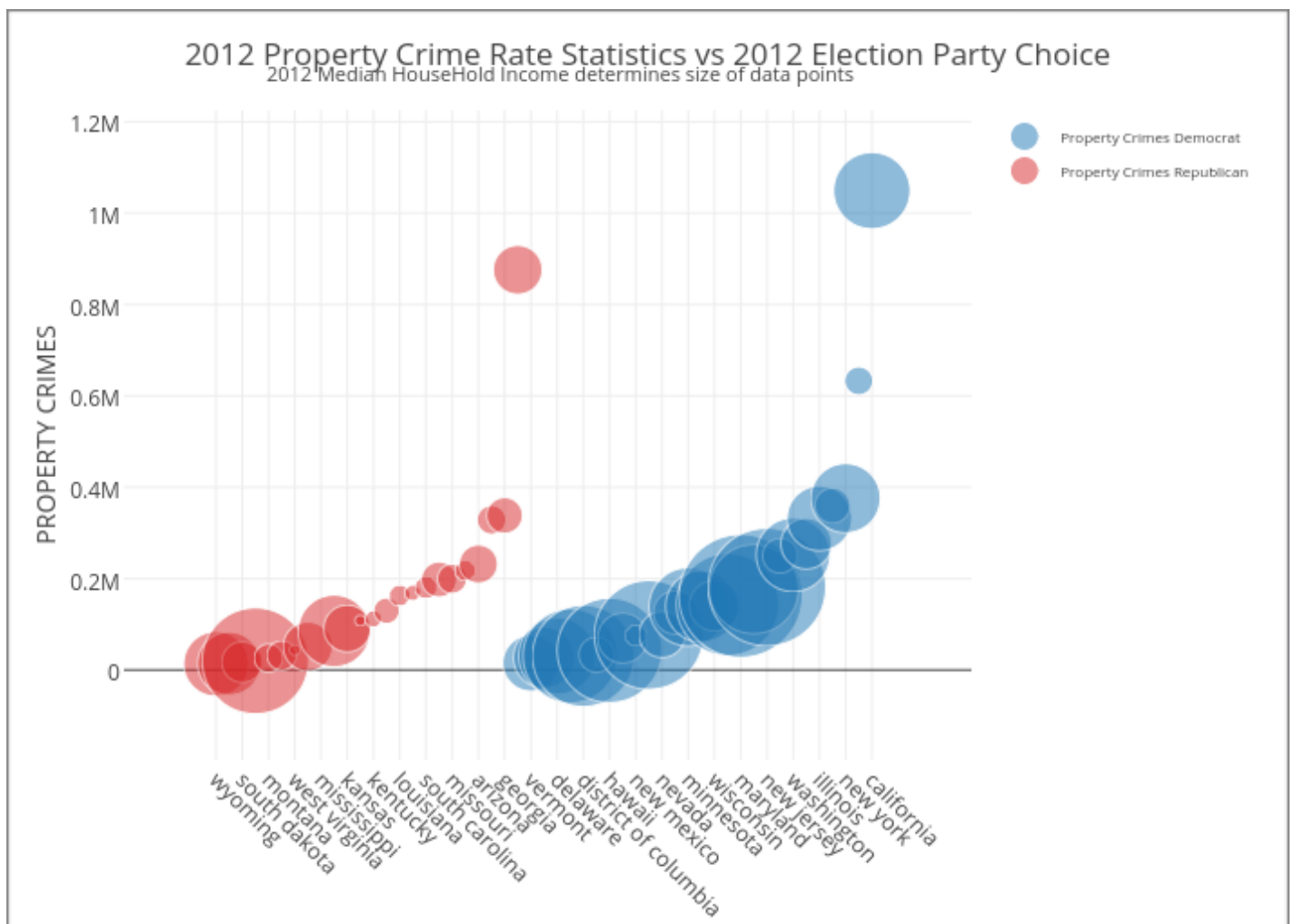
rates than those states that voted Republican. Democrat states also had higher property crime rates. I couldn't think of a potential reason as to why Democrat states had higher violent crime statistics, nonetheless I still think this is an important thing to note.

Looking at both these visualizations then led me to contemplate another story. States where crime rates were lower, also seemed to have lower median household incomes, and voted Republican. Thinking of all of these variables together, perhaps these states vote Republican because economic issues are more present in their daily life. Whereas we see that states with higher crime rates, also seemed to have higher median household incomes, and vote Democrat. Perhaps, these states' higher financial stability allows them to focus on the more social issues and therefore leads them to vote Democrat. I think it would be interesting to do this again with more elections because a larger sample size could show us if these results are significant and stable. After the 2016 election I'm sure there are several data scientist working on models or trying to understand their models and I'm curious to know if at high-level they did something similar to this (of course, it is probably superior to my simple analysis).

Reference List

1. FBI Crime Rate Statistics: https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/tables/5tabledatadecpdf/table_5_crime_in_the_united_states_by_state_2012.xls
2. 2012 Election Results Data: <http://www.infoplease.com/us/government/2012-presidential-election-vote-summary.html>
3. 2012 Census Median Household Income: <https://www.census.gov/did/www/saipe/data/statecounty/data/2012.html>

Appendix:



2012 Violent Crime Rate Statistics vs 2012 Election Party Choice

2012 Median HouseHold Income determines size of data points

