Living Comfy – Data Mining Project

CSCI4502 (Spring 2018)

Cooper Timmerman
University of Colorado, Boulder
Boulder, CO 80302
cooper.timmerman@colorado.edu

Elena Ingraham
University of Colorado, Boulder
Boulder, CO 80302
elena.ingraham@colorado.edu

1. Abstract

This project explored the possibility that areas that are wealthier are more politically polarized. Was it possible that the opulence that set these communities apart from their peers subsequently had an effect on their voting habits? Politics are a messy but desirable battleground, with significant rewards upon the presentation of unbiased data and results. Any knowledge gained from data mining in an attempt to answer these questions or those of a similar nature could be valuable to a wide arrange of people and businesses. Our project concluded with us finding no significant correlation between wealth and political tendencies, contrary to what we assumed could be a very plausible outcome.

2. Introduction

Political polarization was measured by the margin the winner won over the runner up in a particular county in presidential elections. The spread of the data included all of the contiguous United States. The wealth of an area was determined by the adjusted gross income stub which was then averaged by zip code and then clustered into groupings of similar tax information by zip code. We embarked this project with the question, are areas where the election winners have higher margins over runner ups wealthier? Is it possible there is a negative correlation and the less wealthy areas are more politically homogeneous? The knowledge gained from this project has a variety of applications, but it would take a lot more hard work and menial cleaning in order to get to a place where we could make any sort of knowledge application.

3. Related Work

In recent years our question has become more popular as republican and democratic parties have been further divided. However, there haven't been any significant findings on this topic, nor have there been any historic analysis (including visualization) on this topic that would lend a reader insight into the trends that have shaped modern day voter styles. The little research that's presentable to the general public is in the form of basic articles, sometimes with simple graphs, rather than an entire map of the country or even by state over a certain period of time. Findings like to comment on how a presidents' votes included a lot of the wealthy population, but not the other way around where we see what percentage of the wealthy population voted one way or the other, as well as poorer classes (based solely on income tax). Additionally, with the wide-spread proliferation of "fake news" and biases, finding an article that doesn't somehow skew the data has grown increasingly more challenging, often nearly impossible. Because of this, we wanted to use real data and statistics collected by none other than the IRS to portray a non-biased representation of income inequality in the US, and how it has shaped elections in the past, as well as the potential it holds to shape future votes.

4. Data Sets

(All datasets downloaded to one or both of team members' computers)

https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-zip-code-data-soi

- Dataset on income tax statistics by zip code. This proved to be a tricky data set to work with because it lacked a few things, and had dozens of categories that we removed after careful consideration and research into what each of them represented. What we came to find is that zip codes aren't enough to plot each location of which. So we needed to first cut down all unneeded categories, condense the file further so that each zip code wasn't represented by 6 separate AGI classes, but rather just one average class, and then combine each zip code with it's appropriate latitude and longitude using the USZipCode Python library functions. What resulted was 11 separate .csv files (for years 2005 through 2015) containing Zip Code, Latitude, Longitude, and Average AGI for each zip code.

https://www.kaggle.com/stevepalley/2016uspresidentialvotebycounty

- Dataset on 2016 election data by county. This is one of the several interesting datasets we plan to overlay on top of our comfortability model to discover unconsidered patterns. We also used the 2004, 2008, 2012 presidential election data. The most important attribute from these data sets were the margin the winner of the election won over the runner up. This was important to determining the political polarity of the county.

5. Main Techniques Applied

The majority of the time and effort spent on this project occurred during the data cleaning phase. This part of the project was the least rewarding and the most time consuming. Unfortunately, the brave men and women at the IRS did not plan on our team using their data set for this project. A lot of the zip codes had discrepancies that were only uncovered after specialized cleaning. For example, our first attempt to rid all datasets of superfluous columns in a mass sweep failed after realizing that columns had slightly different naming formalities (like "agi_stub" vs "agi_class"). Once these issues were accounted for, we ran into further issues where some zip codes weren't valid at all, returning null values when ran through the USZipCode search function for latitude and longitude. The Statistics of Income Division for these datasets actually states in their documentation that the "SOI did not attempt to correct any zip codes on the returns." Because of this lack of consideration for the proper formatting of zip codes, many zip codes didn't contain all 6 AGI classes, some would overlap with the next zip codes AGI classes, and many zip codes didn't exist at all. Unfortunately, some valid zip codes didn't include all 6 classes, and were consequently removed to avoid inaccurate representations of their averages. Although it is true that some of the values for number of returns per AGI class were zero, first we found the average adjusted gross income for each zip code and since every valid zip code had at least one tax return the average returned was a non-zero number. The adjusted gross income information was delineated in brackets which originally caused some zero values. For example, there were zip codes that had no residents making over $200,000 a year. Since taking the average removed all zero values we did not have to add one to every data point and have begun the classification process.

The election data had its own issues. One of the most interesting things we learned during the data cleaning process was that Alaska is an oddity of sorts when it comes to how the state is divided. Instead of using counties, as is the standard, Alaska uses boroughs. This provided us with an interesting hole in our election dataset. Since there are no actual counties, the county name read as just "Alaska" for the 2016 election data and was completely blank for the 2012 election. We were then obligated to use a supplementary dataset to fill in the missing values. After this we realized the data would be incompatible with income tax data and we decided to exclude it altogether. From there we decided to focus on just the contiguous United States in order to move forward with the data. Alaska, as well as Hawaii, would just pollute the results collected and it seemed like a consolation to include them at the risk of corrupting our other results.

There were also issues within the election dataset which involved special keys being used in county names. It was important to keep the county names so that we had a greater degree of freedom in merging it with other datasets if need be. We made the mistake of removing the state name from the tax data and had to merge it to a dataset of states, zip codes and geographical coordinates which was a time consuming and easily avoidable process. Instead of the total removal of all the county names we had to do a simple search query of all the county names that interrupted the csv file parser with special characters. The most common interferer was the period as in the name St. Claire. Unfortunately, we missed the apostrophe in the name O'Brien and spent a unnecessary amount of time trying to figure out the issue. This was a silly mistake but a great learning opportunity in the discipline of data cleaning.

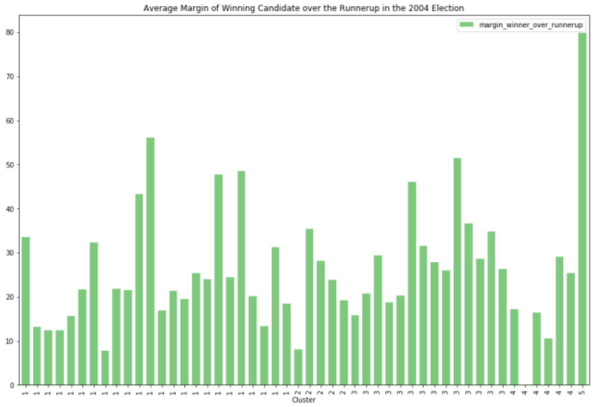
The main method we applied to our income tax dataset was the K-Means clustering algorithm. We were very familiar with the process of this algorithm since we covered it extensively in class as well as on various homework assignments that helped us practice. Before beginning the process, we read through the descriptions of other clustering algorithms. Since we were using

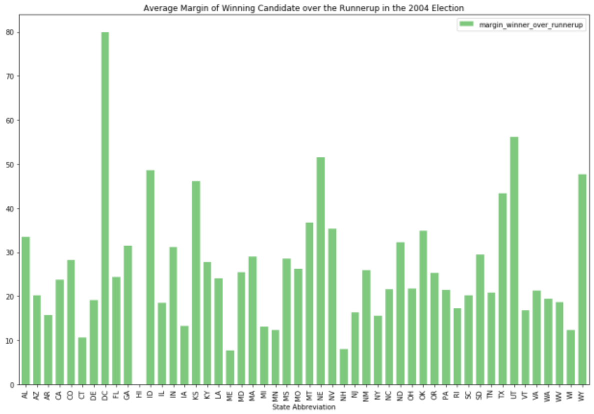
WEKA we did not need to write our own code to cluster the code which gave us more motivation to explore other algorithms. However, the K-Means algorithm is one of the most popular clustering methods for a reason. We briefly considered using the EM algorithm but found the complexity to be unnecessary and the benefits to be slim considering the fact that we knew how many clusters we desired before we began the process. From there WEKA did the rest -- well, we wish it could have been that easy. The first problem we encountered with WEKA had to do with memory shortages. The program was being run through the terminal and when it crashed, it suggested running the program with a flag that would increase the memory heap. This was not as helpful as we would have liked and had to use the garbage cleaner functionality in the WEKA GUI chooser to free up some space. Since we were using very large datasets it was important to be able to free up as much space as possible before running it. The first computer we decided to run the program on failed miserably. We ran one K-Means model for fourteen hours, and then another for fourteen hours and finally one model for twenty six hours. All of which yielded no results. This really reinforced the importance of starting the project early. After this debacle, we decided to switch to a computer with more RAM. And without the least bit of struggle, WEKA ran without hiccup. This was an excellent lesson in the beauty and heartbreak experienced in data mining. The tools available make the process extremely streamlined, only if you are familiar with the limitations of the programs you are using as well as having the correct hardware on your end.

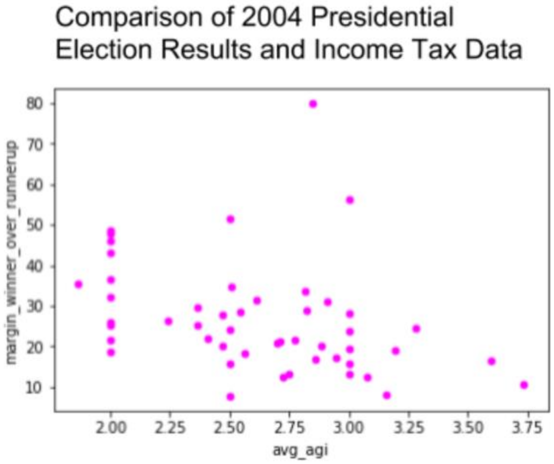
6. Key Results

The most important results of this project had little to do with the correlation, or lack thereof between our datasets, but really came down to the experience and proficiency we gained by using the data mining tools this class exposed us to. The first time we played with WEKA the program seemed complex and inaccessible but now we feel completely confident exploring the tool further, possibly in our careers.

Results from our data analysis were somewhat unexpected. We went into the project naively assuming there would be a correlation between the income data and the election data. However, we found no strong correlation between the margin a winner won by in a particular county and the adjusted gross income cluster from the income tax data. In retrospect, income might not be the most important quality to analyze the shaping of voter patterns over the years, but we were only able to make that type of decision after doing the data mining. If we were to do this project again and had more than a semester to conduct our experiments, we're confident that some patterns unconsidered could be uncovered. It just takes the initial guess of what could influence the behaviors of voters, and the dive into various datasets to find an answer. We were disappointed that we did not discover some sort of beer and diapers anomaly but we now realize examples like that are important because they are rare and require immense amounts of time and data. However the case of Washington D.C. is interesting because it does follow the hypothesis aforementioned in the introduction. It is an area of high wealth and high political hogenemy which is very interesting and could have a variety of applications that will be mentioned later. It is also an outlier in the data set which makes it even more intriguing.

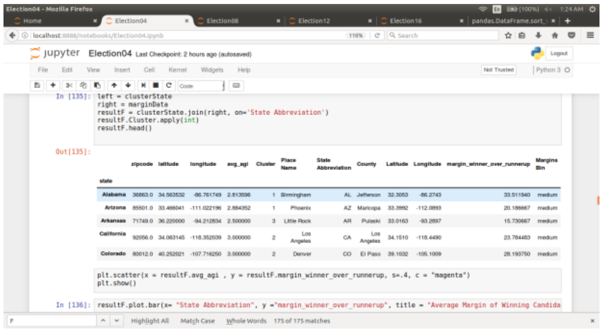






7. Tools

The tools we used to work through this project simplified our workload drastically. In order to cluster data and train datasets, we used the WEKA Workbench. It was especially helpful in organizing the datasets in terms of area of the United States. Since some of the data is by zip code and others are organized by district we had to train the WEKA tool to cluster the data into categories. In instances where we had to use training sets, we used the 80/20 paradigm as is convention we learned from this class. Jupyter Notebook was also useful for visualization. Not only is it simple and fast to use, it was helpful when applying the equations that calculate comfort level to the datasets.



```
In [135]: left = clusterState
          right = marginData
          resultF = clusterState.join(right, on='State Abbreviation')
          resultF.Cluster.apply(int)
          resultF.head()
```

Out[135]:

state	zipcode	latitude	longitude	avg_agi	Cluster	Place Name	State Abbreviation	County	Latitude	Longitude	margin_winner_over_runnerup	Margins Bin
Alabama	36863.0	34.563532	-86.761749	2.813596	1	Birmingham	AL	Jefferson	32.3053	-86.2743	33.511940	medium
Arizona	85501.0	33.466041	-111.022196	2.884352	1	Phoenix	AZ	Maricopa	33.3992	-112.0893	20.186667	medium
Arkansas	71749.0	36.220000	-94.212834	2.500000	3	Little Rock	AR	Pulaski	33.0163	-93.2897	15.730667	medium
California	92056.0	34.063145	-118.352039	3.000000	2	Los Angeles	CA	Los Angeles	34.1510	-118.4490	23.784483	medium
Colorado	80012.0	40.252021	-107.716250	3.000000	2	Denver	CO	El Paso	39.1032	-105.1009	28.193750	medium

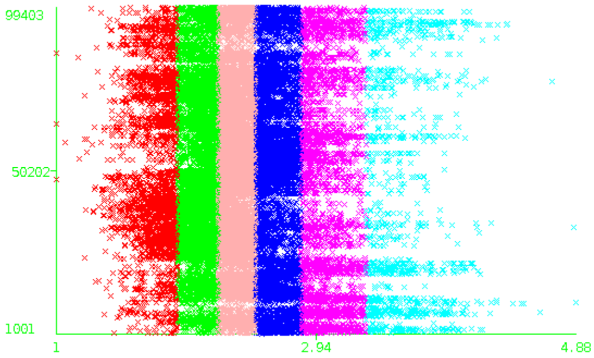
```
plt.scatter(x = resultF.avg_agi , y = resultF.margin_winner_over_runnerup, s=.4, c = "magenta")
plt.show()
```

```
In [136]: resultF.plot.bar(x= "State Abbreviation", y ="margin_winner_over_runnerup", title = "Average Margin of Winning Candida
```

While we thought we would end up using a Python mapping library known as Folium, we switched to a different library after Folium didn't offer a "choropleth" type of map that could easily read in zip code data. We instead chose to work with MapBoxGL, a separate Python library which is much more suited for integrating large data sets and visualizing them. After learning the basics of this new system of mapping, the most efficient way to show all data points across the country was to have zip code tied to coordinates (comprised of latitude and longitude based at the center of the particular zip code) and a single point was given for each zip code that held data on the AGI level, and its given cluster value. In order to stay on track and accomplish everything we wanted to before the deadlines we added Trello to our toolkit. Trello is a program that makes it easy to add and edit components of the project that we need to accomplish.

8. Visualization

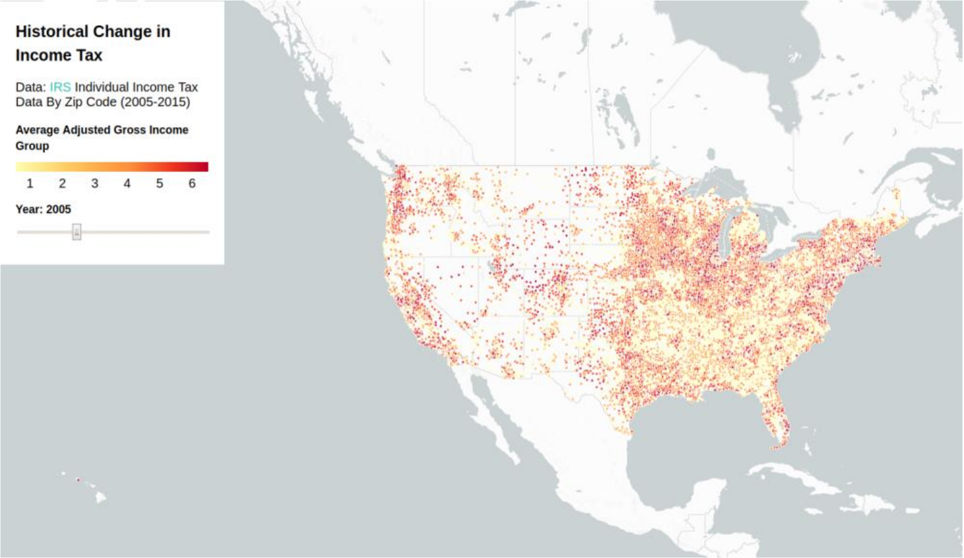
For our project we were able to produce a map of the US's nearly 40,000 zip codes and their trends associated with income tax over the last 10 years. Each plot is a latitude/longitude point at the center of each zip code in the United States, with a color denoting the group of average AGI a zip code was in based on the clustering algorithm we ran using WEKA. The clustering outcome is shown below, and we found it super interesting to see the distribution of wealth among the thousands of zip codes in America. The X-axis depicts the average AGI for each zip code, which fell in a range from 1.0 at the lowest to 4.87 at the highest (a small area in lower Manhattan was essentially tied with a small island community off the coast of Miami, FL known as Fisher Island as the two wealthiest zip codes in the United States). The Y-axis is the zip codes from 01001 to 99403.



Additionally, we've included some screenshots of our United States map of each zip code, colored according to the cluster that WEKA assigned to them. This process was extremely lengthy, because the goal was to produce a webpage with a slider based on the year that could show each set of results, but the Javascript behind being able to show 10 different maps without refreshing the page turned out to be much more challenging than expected, and wasn't fully perfected. As you can see, each year above the slider states "2005", but the following images are 2007, 2008, and 2009, respectively. This is the most interesting snapshot to analyze because it shows the change of incomes in the same time as both the election of President Barack Obama as well as the United States Housing Bubble Crisis of 2008, where a large majority of Americans suffered financially. Within a few years, these sufferings were partially stabilized, as we can see with the increase of dark red spots after the recession in the 2009 Map (#3). For further analysis of these maps we've included the code and instructions for viewing them in our Github repo.

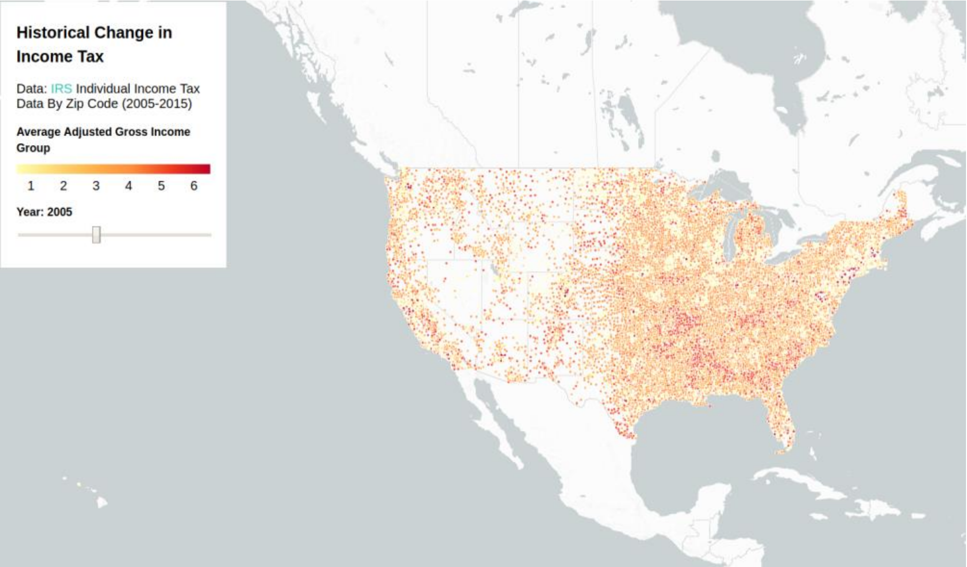
(Images together on following page)

#1 − 2007


Historical Change in Income Tax

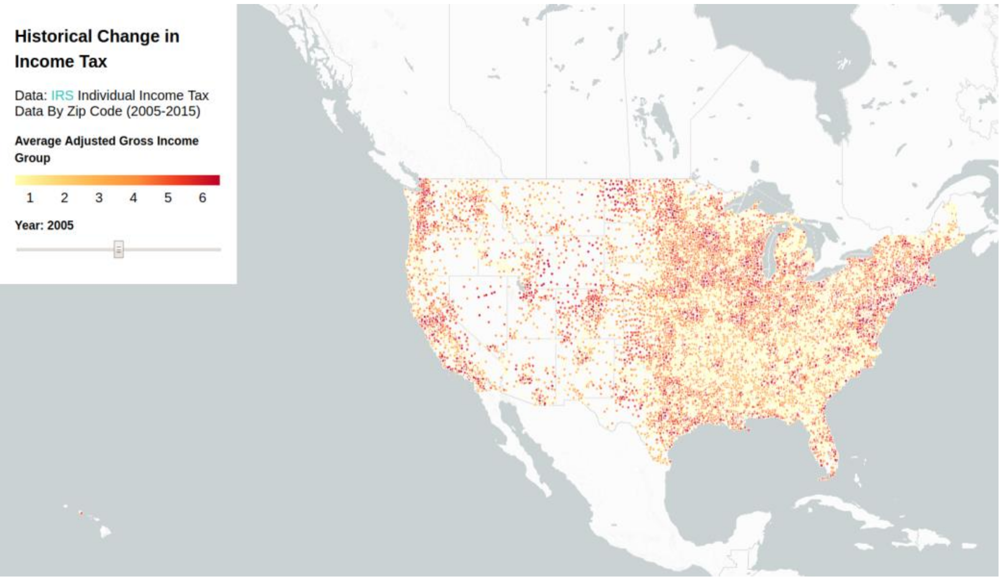
Data: IRS Individual Income Tax Data By Zip Code (2005-2015)

#2 − 2008


Historical Change in Income Tax

Data: IRS Individual Income Tax Data By Zip Code (2005-2015)

#3 − 2009


Historical Change in Income Tax

Data: IRS Individual Income Tax Data By Zip Code (2005-2015)

9. Application

The discovery that wealthy areas in the United States and significant election margins have no correlation is nonetheless a valuable discovery. Presidential candidates have been known often to tour (or ignore) areas based on their wealth, but our findings indicate that this selective rallying is redundant because there is little to no correlation that would predict a regions vote based on their income.

Additionally, those looking for a region to relocate (based solely on the level of income they could expect) could easily interpret this map and review the most recent years that indicate areas of high income. As stated before, this search could obviously be improved with datasets on health, costs of living, and real estate prices if more time was allocated for the study. But to get a rough sense of what regions offer high income in comparison to others, our map visualization provides a valuable resource that is easily interpretable by even the least tech-savvy citizen of the US.

Outliers could also serve as a means to discover more knowledge from these data sets. When embarking on this project we maintained the mindset that averages and modes were the best way to extract patterns and information. However, near the end of this class we learned the importance of outliers and the possible knowledge that can be gained from analyzing them. Collective outliers could be especially important in finding areas of interest for campaign managers and politicians alike. Reinvestigating the outliers in this data set would bring even more knowledge. For example, we noticed just by looking and the data in a bar chart that Washington D.C. was the most politically homogeneous. This makes sense because of the importance of politics in Washington D.C. but was still an interesting fact that could be investigated further. Washington D.C. is an outlier in this case but it could be valuable to apply this information to political action. It is possible that senators are out of touch because they live far away from their home states in a politically homogeneous area and campaign managers could make sure they make stops in towns with higher political polarity.

We also extracted a rating level for each state on a scale of very liberal to very conservative based on voting margins. This information could be applied to voting prediction models by analyzing which state hover in the middle and are likely to flip. This could be important for politicians as well as business owners. For example if a conservative state is becoming increasingly liberal, the owner of a Legal Marijuana distributor could look to invest in the newly liberal state before the market is oversaturated.