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. Problem Statement/Motivation

There are many people living beyond their means across the United States, but why? We are setting out to discover patterns between regional costs of living and their incomes to develop a model on how "comfortable" people are living regionally. The costs of living will be determined through datasets related to costs of food, household fuels, and more, while the income will be determined by a dataset that includes classification on six different tax brackets per each US zip code. Once this model is created, we plan on overlaying additional interesting datasets to dive deeper into comfortability's correlation with political, health-related, and other categories not necessarily related with comfortability.

2. Literature Survey

Although the definition seems widely accepted, it is important that "living comfortably" is formally defined. We interpreted it as the ratio between how much money one makes versus how much they must pay to live there.

The internet offers many informal lists of the most comfortable cities, as well as several calculators to determine how much the suggested income is to live comfortably in different areas of the country. Most of these lists only use recent data instead of a larger collection, giving only a small piece to the larger picture. Because "living comfortably" is a popular internet article for social media platforms, the lists are not user friendly in the sense that the regions are all on different pages, and there are no visuals to compare regions.

There are numerous datasets made publicly available through data.gov that contain data throughout the past 20+ years on regional average costs of living via food, gasoline, and home utilities. However, there hasn't been much publicly-available development made with the data other than accumulating more data. By combing through all this data and adding quality data visualization, we hope to give this topic some light by giving an alternative to reading through Facebook articles on "living comfortably."

3. Proposed Work

In terms of data collection, we already have the datasets necessary to perform our desired tests. The datasets that have been collected will be described in further detail later in this proposal. The data collected will require a lot of preprocessing which will most likely comprise the majority of our work.

First, we will remove all unnecessary and superfluous attributes from each dataset. For example, we are using a complex income tax dataset with extremely specific tax information that will not lend us any information for our project. Instead these terms are only important for the IRS's records so we will remove them. Once this is complete we will begin to integrate the datasets. The biggest challenge in making the sets compatible will be adjusting the data based on

area. Currently the Income Tax dataset is tracked by zip code, whereas the General Elections datasets are in terms of district. Voting districts are often random, organic shapes and we will need to average some of the data points in order to compare the regions in which they cover.

Once this is finished we will create a model to rank areas of the United States with a score that represents the comfort of living for the average citizen of that area. The measure will be based on a combination of factors including income tax paid and costs of living. The current standard for cost of living measurements is the Cost of Living Index, formerly known as the ACCRA Cost of Living index. Their method involves the collection of data with regards to the cost of goods and services in a specific area at a specific time. This information, in combination with the average rate of consumption of necessary products (such as food and gas), is used to estimate a cost of living for residents in that area.

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. In this data we'll focus on the adjusted gross income and the wages and salaries. This will be the first step to developing our comfortability model.

https://catalog.data.gov/dataset/consumer-price-index-average-price-data

- Dataset on different aspects that influence the cost of living throughout dozens of metropolitan areas of the US. These include food, gas, and home expenses. This is the second half to our comfortability model

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.

https://catalog.data.gov/dataset/national-survey-on-drug-use-and-health-nsduh-2015

- Dataset on the past 9 years of drug/health related surveys throughout the US. This will be another interesting dataset to overlay on top of our comfortability model.

5. Evaluation Methods

Our evaluation will be dependent on a linear correlation between the cost of living and the income tax paid. Our null hypothesis before we begin states that areas with a historic downward trend in comfort level are more likely to change their voting ideology. We will keep a base confidence level of sixty percent.

6. Tools

The tools we use to work through this project will hopefully simplify our workload drastically. In order to cluster data and train datasets we will use the WEKA Workbench. This will be 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 will most likely have to train the WEKA tool to cluster the data into categories. In instances where we will have to use training sets, we will use the 80/20 paradigm as is convention. Jupyter notebook will also be useful for visualization. Not only is it simple and fast to use, it will be helpful when applying the equations that calculate comfort level to the datasets. We will also be using Folium, a mapping visualization tool to be used with Python, to produce visually appealing maps that represent each aspect of our project. This will

allow us to avoid manually plotting on a .svg map file, and instead use Jupyter Notebook with Folium to produce quick, clean results. In order to stay on track and accomplish everything we would like before the deadlines we will be adding 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.

7. Milestones

It will be extremely important to stay on top of deadlines for this project. As previously mentioned before, Trello will help keep us on track. Before the start of spring break we hope to have finished all data cleaning and data pre-processing. This will be March 23, 2018 and although it does not seem like that ambitious of a goal, this work will most likely be the most difficult process of the project. The following week we hope to have the models that will compare our data written. We hope to be finished with this by April 8th. This will be difficult since we will be focused on the midterm during the bulk of the week but both of our weekend schedules are light so we can find time them. By April 17th, 2018 we will have all of our data mined leaving us a good amount of time to thoroughly evaluate the results and look for interesting patterns. By the beginning of May 2018, we plan on having our maps as close to complete as possible. This allows for us to finish with a conclusive final paper instead of a rushed and sloppy final statement.

7.1 Milestones Completed

Since we are using WEKA to help with classification, the first step of our data cleaning process began with converting the files into one usable format. Although this seems like a menial chore, it was critical before we could proceed to anything else. WEKA only accepts files from a limited range of extensions, and our income tax data needed to be converted from an .xls file to a .csv file.

After this was completed we could actually begin to work with the data. The first problem that arose was that of zero values in the data sets. The possible complications of this would arise if we were to use a function that had the possibility of being computed with one of the zero values in the denominator. If this were to happen the function would be undefined and the results would be useless, potentially rendering the classification process as a complete waste of time. One of the easiest ways around this is issue is to add one to every data point. However if the data contains zero values but will not have a function with a possible zero denominator applied to it, adding one to every column is unnecessary.

For the income tax dataset, although it is true that some of the values are zero, first we found the average adjusted gross income (AGI) for each zip code and since every 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 rest of which will be described in the Milestones To-Do portion of the report.

The election data we are using also had many zero values within it as there were some areas that voted purely one way or the other. For this dataset it was necessary to add one to every data point because we are discretizing the data based on the votes cast and we need to be able to use any classifier on just the raw data. This was an easy fix and the data is now usable. However, some issues occurred while trying to classify the data using WEKA. The program ran out of memory and prompted to increase the heap size with an added flag to the command line argument. This event was important because until then we have played with WEKA without a clear intention and we realize now that it will be extremely important to have a strict game plan in order to prevent the system from crashing.

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 nationwide 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. To fix this we pulled a supplementary dataset from the State of Alaska's government website. From there we filled in the missing data and just used the boroughs as a stand in for the county name.

As of right now, all of the data cleaning is completed and the classification process is underway for all of our data. Since the data is so easy to work with now, we should not hit any obstacles in the classification process. However, we are prepared to encounter issues and perform more cleaning if necessary. We also made an important addition to our toolkit to help us with this project. We will now be working with Folium to aid with the visualization portion of the project. Originally, we thought the best method would be to take a blank .svg U.S. map file and fill it in with the data, but Folium has proved to be a much better choice. It allows us to work out of Jupyter Notebooks and the maps are seamlessly beautiful. In order to get familiar with the program, we worked through the public choropleth map example as it is most relevant to our needs. After working with it for a bit we are confident that we can easily use it to visualize our data. More on this process will be covered in the Milestones To-Do section.

7.2 Milestones To-Do

With the success of our previous work on cleaning and preparing the data, we are ready for the next portion of our project. That portion will include further classification for each data set.

Our nationwide income tax dataset includes a lot of extremely useful information that will make it almost trivial to applying binning techniques to each zip code. The attribute that makes this possible is the AGI-stub for each zip code, as previously introduced. This is a number 1-6 that classifies each zip code into six tax brackets with taxable income ranging from less than $25k per year (as the lowest bracket) to over $200k per year (as the highest bracket). With the average we've gained from each zip code, we can easily see how balanced (or unbalanced) the 6 sections of each zip code are. Additionally, we will be classifying each zip code by our own original label to easily compare a state's zip codes by their average AGI. This will give us a good comparison between zip codes across a state. An example of a possible pattern we hope to find would be neighboring zip codes with significant imbalances between their average incomes. Because we have data ranging back twenty years, we plan to implement a historic time lapse of how these averages have transformed through the years.

One of our datasets includes cost of living throughout different areas of the United States. This cost is determined through a myriad of different expenses, including food prices, household fuels (gas and electricity), and other additional costs. However, this dataset proves to be a challenging one to use mainly because of the way it represents geographic spaces. Instead of a consistent use of zip codes or counties, it ranges from broad sweeps like "Northeast" and "Midwest" to "Denver-Aurora-Lakewood, CO" and "San Diego-Carlsbad, CA." Because of this inconsistency, we will have to manually define these areas on our cost-of-living layer of the final map. Because these areas don't include each and every US city, we will likely use the much broader regional areas for the majority of our classification and use more exact locations as we see fit.

For our election dataset, we plan on classifying the results into bins, giving individual districts a "political affiliation score" that would reflect their respective preferences towards a democratic or republican party. As this will be one layer of our multi-layered final map visualization, we plan on using a color-gradient that's easily interpretable

(likely a red to blue gradient). Additionally, we are attempting to provide a historic time lapse on each presidential election year since the year 2000 to see the temporal change in political opinions over the last two decades.

After we finish the classification for each dataset, we will dive deeper into the visualization portion of our project. Our goal is to produce at least three layers that each have different methods of visually classifying US regions. Our cost-of-living and income tax maps will likely be seen as different colored gradients for their separate maps, and another different gradient for the combined model of the two. Additionally, we want to implement the time lapse for each group of data. The election data will be a two-tone, red to blue, gradient to show political preference. For these goals, we will continue to work with Folium to develop beautiful US maps easily interpretable by even the least technologically-savvy person.

Results So Far

As a team we are currently still in the development stage of our map, and some final parts of the classification of each dataset. We predict that once we have implemented the maps, the patterns will be extremely easy to discover and discuss. However, with a lack of the final product, we have only found a few interesting aspects to our datasets just by combing through them and seeing where they need cleaning. Some of our more interesting discoveries include voting districts with a count of zero votes for a specific candidate; this is interesting because it brings up the question of whether voter fraud could have occurred, or if the districts really did unanimously vote one way or another, and why. Once we have our multi-layered map functioning, these questions might be answered, and more patterns will definitely be discovered in the process.