# 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

## 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. 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.

## 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 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."

## 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 housing prices. 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.

## 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.

## Evaluation Methods

Our evaluation will be dependent on a linear correlation of less comfortable living conditions and a change in voters' attitudes. Our null hypothesis before we begin states that in areas will a downward trend of comfort level are more likely to change their voting ideology. We will keep a base confidence level of sixty percent.

## 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 were 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. 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.

## Milestones

It will be extremely important to stay on top of deadlines for this project. As previously mentioned hopefully Trello will 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 the 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. This allows for us to finish with a conclusive final paper instead of a rushed and sloppy final statement.

## Summary of Peer Review session

The most important piece of advice we got from the peer review sessions was “Be specific”. This often applied to the other work that has been done on our topics as we were asked to cite specific studies. However, this applies to all aspects of our project as it will be extremely important to propose a clear and concise hypothesis and plan of action for our project. Working together with other groups is also an important piece of the project. The odds of finding something interesting from the data will be increased if we can learn from the other groups in the class just as we have learned from other studies.