Process Book

Will Stout, Xuyang Li

October 2020

1 Overview and Motivation

With a growing wealth inequality in America it is becoming increasingly difficult to set yourself up to achieve financial stability. The tricky part is that by the time you graduate college the number of variables you have that change when you might retire is shrinking rapidly. And picking a career in a certain city to stay in long term might just set it in stone. We want to see if there is any significant variance in how the job you choose in a particular city changes how much you might be able to save/invest and therefore when you could retire.

Even though there are many factors included in the universally recognized cost of living index from The Council for Community and Economic Research (namely food, housing, utilities, transportation, health care, and miscellaneous goods and services), there are no categories as easy to compare as rent. Additionally, there are many transparent and trustworthy resources and data sets available that specifically look at average rent/home prices by city. If we wanted to consider the next largest factors for annual spending then the problem either becomes increasingly more about who you are as a person rather than where you live or it is very difficult to find trustworthy data sets.

For instance, how much you spend on health insurance is more about age/medical conditions/whether your workplace provides it or not. For food, it's about what/where you consume on average. Trader Joe's is notorious for offering the same price for an item at every single store they own in America. And for transportation, there are so many different ways to travel that how you decide to travel is more important than in what city you travel. While location may factor into these costs, it doesn't factor in as much as it does for rent, so that's what we're going to focus on.

2 Related Work

The largest inspiration for this project has been `https://www.numbeo.com/cost-of-living/rankings_current.jsp`. A good source for cost of living comparison, but we're looking at creating a more specific and perhaps meaningful comparison tool.

3 Questions

Something we're both curious about is precisely how much variation there is between job salary in different cities, and where is the most cost efficient place to live for a particular job? The question of how much does tax factor into where is the most cost efficient place has taken on more meaning now. Since obtaining all local and state tax information, there is far more variation there than expected. Something new to consider is if we want to create a better way to understand tax outliers like Texas.

4 Data

We plan to get salary data from Payscale.com (job salary survey site), which lets you select a job and find the average salary for that job in a given city. Payscale reports the number of data points they have for each job/city permutation, so we're choosing a select few that have the most survey responses (more than 1,000 responses). Payscale treats all data points the same regardless of how old that data point is and doesn't blend data, so values might seem low since they're not adjusting for inflation, but salaries should be low consistently across different cities. Will already created a web scraper to track the salaries of 14 very popular jobs on Payscale.com, so information will be coming from the web scraper.

In addition to that we're using Zillow.com (apartment/home listing site) data sets that report average rent prices for the 100 most populated cities (by zip code) in America. Their data is blended, modified to reflect normal seasonal changes, and is segmented by month, so we're just going to use their most recent month posted (currently August 2020). And finally we're also going to be using income tax data. Specifically national tax brackets, state income tax rates, local tax rates, Social Security tax, and Medicare tax. We'll collect this by hand, but it shouldn't be that much work.

Some data processing will need to occur, nothing substantial. Payscale tracks the locations of salary reports by city name and is pretty specific about it, i.e. two city centers that are ten miles apart have their data points stored separately. Since we're going to attach salary data to rent data that is organized by zip code, we're going to need to organize which zip codes are within appropriate bounds to consider it a certain city. So the goal is to do that zip code to the city conversion by hand, then in D3 load that data and combine salary and rent data points when city names match.

Sources:
Rental price dataset: `https://www.zillow.com/research/data/`
Wage by occupation dataset: `https://www.payscale.com/research/US/Job`
Tax rates: `https://smartasset.com/taxes/paycheck-calculator`
Zip codes: `https://www.bestplaces.net/`

5 Exploratory Data Analysis

From the beginning we knew we wanted to use a map of the US with only the cities we were including in our data set. A map can be a really great tool for comparing cities that are physically close, but we realized we also wanted to compare all cities against each other in a more visually succinct manner. So we're adding a bubble chart that ranks cities in terms of financial performance. We think users should also be able to compare any two cities side by side with all information we can provide them to give them the most informative experience possible, so that's something else our initial design informed.

6 Design Evolution

We looked at many potential designs. Our main two issues were 1.) how can we create easy to understand visualizations for cities when taking into account their tax and rent. We thought to try mini bar charts and to scale the size of those bar charts depending upon financial efficiency. We decided to use pie charts as easy signifiers of how much salary is going to rent/tax, and to scale them based on how much of your salary you take home. We also wanted users to be able to select if they didn't want to see income after rent/tax, so that will be toggle-able. Then 2.) in wanted to give users an impression of how a city performed against all other cities, we thought to try a bar chart in two variations. The first would show just the best 5 and worst 5 performing cities, and the second showing all cities. We decided to use a bubble chart for visual conciseness. The position of the bubble is moved on a y-scale depending on how cost efficient the city it represents is.

7 Implementation

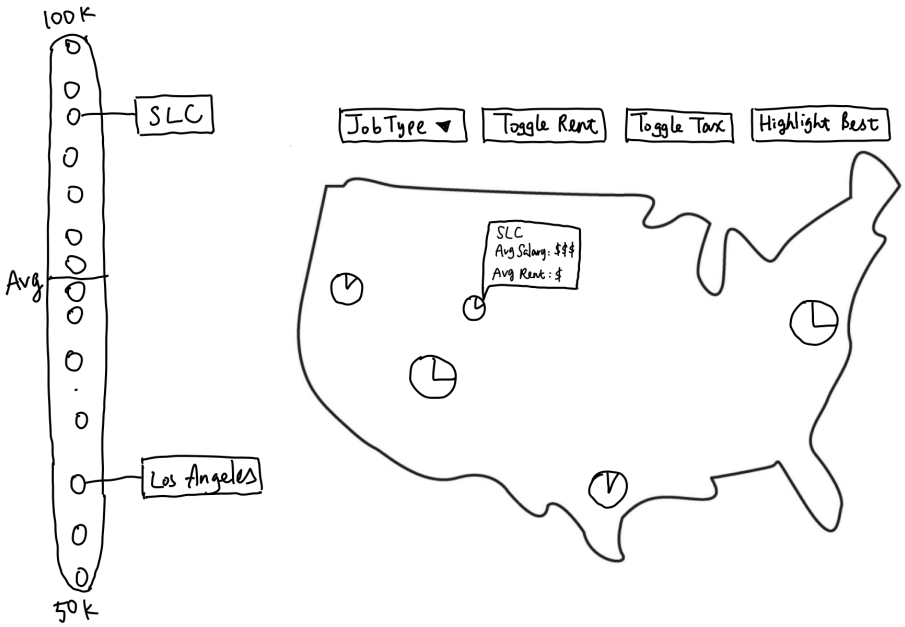

Figure 1: Final Design

We want to give users three different ways of viewing the data. Regionally, nationally, and side-by-side. The map on the right in our design is great for comparing neighboring cities and can portray how cities do within that region. Then the bubble chart on the left is for comparing a city against the whole nation, in an easier to interpret manner. Clicking on a city on the map will highlight that bubble on the bubble chart to make finding the city easier. And finally, we have a stretch goal of side-by-side comparison. You'll be able to click on two cities to see all information we have on them so you can compare two places that perhaps aren't physically close.

8 Evaluation