CS 171, Spring 2015
Curtis Stone, David Kaufman, Evan Gastman
Final Project: Heroin Use and Income

Process Book

Original Proposal

Background and Motivation

Although none of us has a specific prior interest or involvement in this topic, all 3 group members found it fascinating and were impressed with the depth of the available data. We read new articles about a recent surge in heroin's popularity in the United States, and wanted to explore the geographic and demographic areas in which this was most prevalent.

Project Objectives

The questions we are trying to answer are:

1. Where is heroin use most prevalent in the United States?
2. What is the relationship between heroin use and income level?
3. What is the relationship between heroin use and other criminal activity?
4. How do the answers to the above questions change as we filter by age, race, and sex?

The motivation is to gain a better understanding of who is using heroin in America today.

Data

Our dataset is from the National Survey on Drug Use and Health, 2012. The survey was conducted by the US Department of Health and Human Services. The data, which is available for free download, can be found by following this link: http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/34933

Data Processing

The data cannot be downloaded in JSON form, so we expect to convert it. The final form of our data will be an array of JSON objects representing individuals, with keys for age, sex, race, state, heroin use, income level, and criminal activity. Due to the massive size of the dataset, we will also need to filter it sustantially. Data processing will be the first step in our project.

Data Display

We plan on using several interactive views to illustrate the interplay among a number of demographic relationships for heroin users, including gender, age, socioeconomic status, and race. The views we are currently considering include: (1) a map of the United States, which shows hotspots of cocaine use via a gradient scale, (2) a bar chart relating-- more specifically-- socioeconomic status and heroin use, (3) a pie chart that depicts the type/severity of criminal activity for the currently selected demographics of heroin users, and (4) a New York Times connected news feed that provides headlines re: national heroin use or specific state heroin use if a state is selected on the map (incorporating NYT Linked Data Application into our app)

So, currently all of our views are connected, which we feel is important in illustrating who is using heroin in the US today and how environments have affected that use.

Must-Have Features

The above views are all ones that we feel are necessary for our project.

Optional Features

We are currently trying to figure out what other views (possibly in the form of scatterplots) could be useful in helping illustrate relationships about heroin use that will be helpful to our audience (in so far as additional views would not be extraneous).

Project Schedule

Proposal due Friday, April 3

April 10, Cleaned up and processed dataset ready for filtering/incorporation into views, as well as research on NYT linked application within our website

April 17, Project Milestone with our core working views

April 24, Inclusion of Optional features

May 5, Project ready for submission
(Throughout work process, update Process Book)
National Survey on Drug Use and Health, 2012
www.icpsr.umich.edu

Data (Cleaning, Wrangling, Filtering)

The dataset was not available for download in JSON form. I downloaded it instead in STATA form. There were initially 3,141 variables, and the file was roughly 350 MB in size. (Almost) every variable corresponds to a survey response. There are 55,160 observations, each of which represents a survey respondent. I read through the survey questions to determine which ones would be most useful for our project.

Still working in STATA, I dropped a large number of variables. Then I looked over the remaining ones and determined which would be most useful. I repeated this process until I was left with 22 variables. Often in the survey there are multiple phrasings or encodings of similar questions, and I opted to keep a few of these to leave our options open later. I also generated some indicator variables that might be useful.

I conducted some exploratory data analysis to get a sense of general trends in the data and determine the best way to represent it.

The variables I decided to use to indicate heroin use are “HERMON” and “HERYR,” signifying respectively whether the respondent has used heroin in the past month and past year. Other variables of interest include mental health and crime indicators, as well as income, sex, and age.

After filtering the main data file, I also created a separate STATA file containing data on heroin users only.

The next step was to convert the data from STATA format to JSON. I initially tried to do this using Python, as our TF suggested, but then found a CSV-to-JSON converter online. STATA conveniently can export data to CSV, so this made conversion rather painless.

Once everything was in JSON, I separated and organized the data into 3 files: all data, users-only data, and metadata. “All data” is simply the filtered STATA file converted to JSON. “Users-only” only includes responses from the ~200 respondents who reported using heroin in the past year. Having this smaller dataset rather than filtering every time

will increase the efficiency and speed of our project. This makes sense, as our main views only provide information on heroin users. The “metadata” file contains the data collapsed by income brackets. I used STATA to tabulate this information and copied it into JSON. Having this file will allow us to avoid looping through the 55,160-observation dataset each time. I think this decision makes sense, as the graph has a very small finite number of possible permutations.

At this point, I wrote JavaScript functions to wrangle the JSON data into the various forms we would need for our visualizations. This primarily involved aggregating individual-level data into objects containing counts for each category.

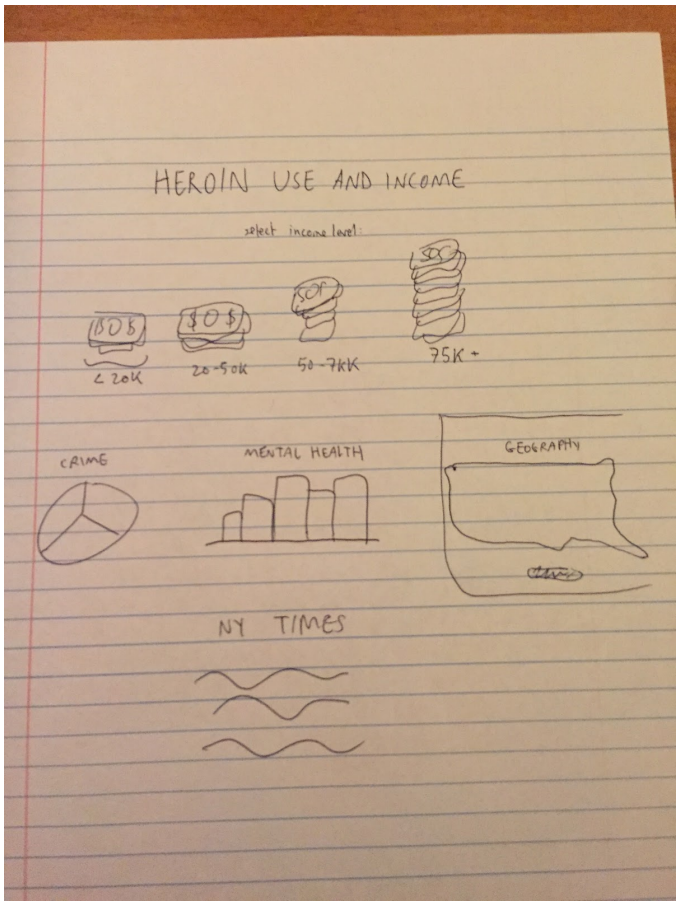
Design Evolution

The first major evolution in the design was shifting away from focusing on geography to focusing on income. We had to abandon the state-by-state map aspect of our project because we discovered that our dataset did not include the respondent’s state, for confidentiality reasons. At first this felt like a setback, but we now we feel that the relationship between heroin use and income is more relevant, compelling, and original than the geographical approach. The NSDUH also *does* provide some aggregate information broken down by state, but it is not heroin-specific. If we have time at the very end we may include a (small) map, but this would be a standalone feature and we do not consider it essential.

To represent income, and give this relationship a central place in our visualization, we will include “money stacks” signifying income brackets, allowing users to filter the data based on this.

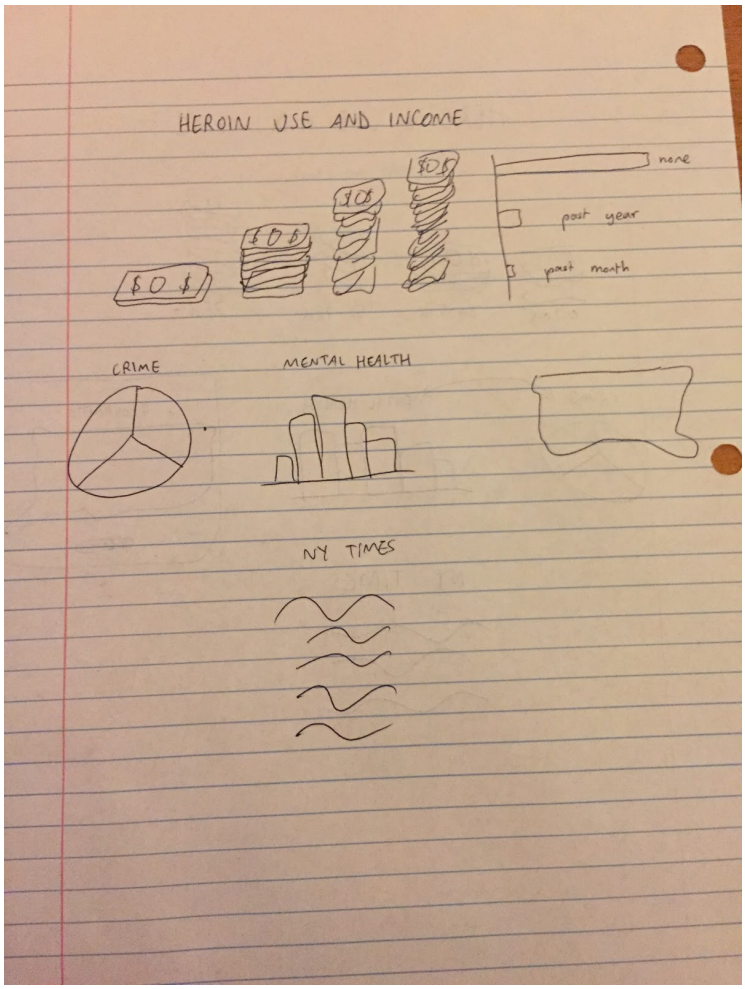
Looking at the dataset, it was clear that mental health and substance treatment would be a fascinating area to explore and represent visually. So we added another bar chart for this. To represent criminal activity among heroin users, we thought a pie chart would work well. (Image 2.)

Image 2.



After a little more thought, we realized we were missing a key piece of information that would provide important context for the existing visualizations. This was the number of heroin users vs. total population in each income bracket. This information is all contained in the “metadata” file. The simplest way to represent this is another bar chart, with options to hide/display the total number, which will dwarf the number of recent heroin users. (Image 3).

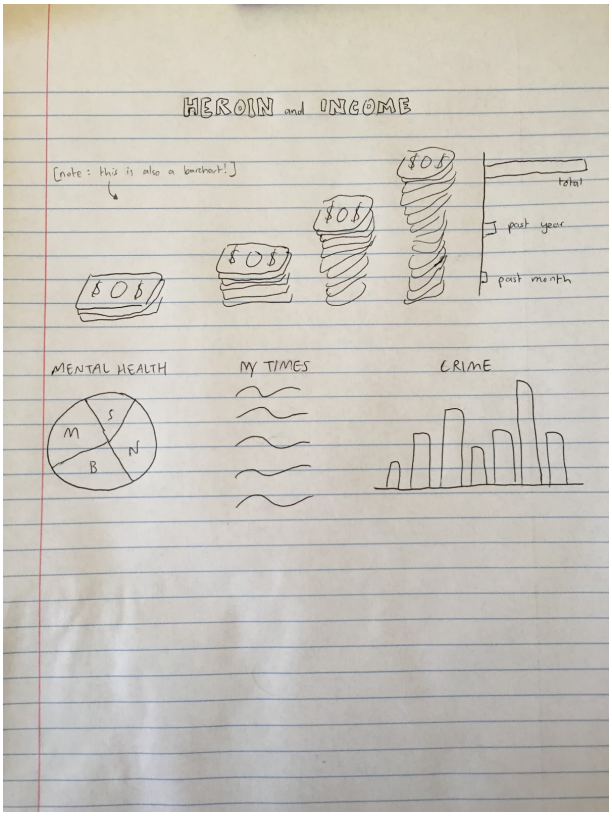
Image 3.



The next development was realizing we should switch the visualizations for mental health and crime, for two reasons. First, crime has more possible values and pie charts tend to work well with only a small number of options. Second, thinking of the crime categories as being exclusive or summing to 1 makes little sense, and thus using them as pieces of a whole pie would not work well. The mental health treatment and substance treatment variables, on the other hand, break up nicely into 4 categories: just one, just the other, both, neither.

Later, looking at the “money stacks,” we realized another potential area for interactivity. We could turn the stacks themselves into a bar chart, which would change dynamically as the user selected different ways of bracketing income. This feels much better than having a large static image in the center of our visualization. (Image 4.)

Image 4.



Early Implementation Stages

For the first checkpoint, we visualized the “crime” data as a bar chart with vertical bars, implemented the totalVis as horizontal bars, and the mental health vis as a pie chart. At this point, none of these views are interactive. As mentioned before, we realize that it will be a challenge of totalVis to show the extremely large bar (total people surveyed) as well as the other small bars. As a stopgap, we decided to only encode the smaller bars (# of heroin users in the past month and # of users in the past year), and we encoded them as percentages of the total number surveyed. These percentages are extremely small, (~0.5%). We are not confident that this is the right way to go, but thought it was a

better option to turn in for our milestone than just visualizing the raw numbers, especially since we have not had time to implement a solution to the big-bar problem. As of right now, our plan is to implement a toggle option where the user can decide whether the big bar (total people surveyed) is displayed or not. If not, the scale will update so the two small bars become visible.

Milestone

April 17th: We submitted milestone 1! Our milestone contained prototypes of the “crime” and “total” views, and a preview of the “mental health” pie chart. The filters and interactivity have not yet been added.

Meeting

On April 25th, we met with our TF, Daniel Haehn. Daniel was happy with our milestone and we discussed next steps moving forward. We came away with a clear list of tasks and things to implement:

- Change view layout. This means ordering our various views on the page to create the simplest and most elegant layout possible.

- Add dollar bills. This means adding a new view in the place of the income bracket selection buttons.

- Usage view with total bar with interruption. We wanted to change “totalvis” to encode population totals rather than percentages. However, this would result in one bar always being much, much larger than the other two, since the total population dwarfs the number of heroin users in all cases. Initially, we had thought of letting the user toggle whether the large bar would be shown, so they could see two views, one essentially of one giant bar, and another of the two smaller bars. However, in consulting with Daniel, we decided that this wasn’t an effective way to visualize this data. As a way around this issue, Daniel suggested adding an “interruption” to the largest bar. He showed us a paper that uses this technique. We agreed that this was a better solution.

- Add readable labels. To make our graphs easier to understand, Daniel suggested adding labels and/or tooltips.

- Add supplementary filters. This is what makes our project interactive and engaging. Beyond the central feature, namely filtering by income, we also want to allow the user to filter the data based on sex and age. This will give the user a much fuller picture of who is using heroin.

- Add *New York Times* headlines. The aim here is to put the data into the context of current events. This feature is not interactive with the others, and does not make use of our dataset. It is therefore lower priority than the others, but we hope to include it to add a sense of relevance and urgency to the data being displayed. In order to implement this feature, we will use the New York Times API to pull recent headlines based on relevant keywords.

- Eye candy. Finally, we will add all sorts of eye candy to make our visualizations pop. This will include a landing page, stylish titles, and nice looking buttons and filtering methods.

We also discussed the possibility of comparing heroin to other drugs. This is an interesting possibility, but we felt that it was less essential to our project than the above features. Daniel agreed.

Overall we found the meeting very helpful and it left us optimistic about the possibilities for our project.

Division of Work

We divided the tasks set out in the meeting among the team. Curt will handle constructing the new “dollar bills” visualization, as well as adding the filtering functionality to the project. Evan will work on eye candy and the New York Times feed. David will add interactivity to the pie chart and work on adding an “interruption” to the totals view.

Money Vis

Since radio buttons are boring, I thought it would make sense to display the income bracket selectors as stacks of dollar bills. To give our central feature interactivity, I decided it made sense to implement this as a bar chart itself. I did not want this to simply be a static chart, however.

Since the main relationship in our project is between heroin and income, it made sense to give the user more options on how to filter income. Our dataset provided two different systems of bracketing income, so I decided to allow the user to select which of these to display in the Money Vis. The view would therefore either consist of 4 stacks of bills or 7.

To implement this, I needed to return to the data. This is because the income brackets were originally part of our “metadata” file, and when I wrangled this for the first time I decided only to include one system of bracketing.

As before, I used STATA to filter the dataset, then converted the updated file from .dta to CSV and finally to JSON. I saved the information broken down by income brackets in a second metadata file. This second file is parallel to the first, but with 7 brackets instead of 4. We use the queue.js library to facilitate the loading of several data files.

The next step was building another bar chart. This chart would encode the average income in each bracket. It would be special in two regards: it would be clickable, allowing the user to filter the data to the selected bracket, and it would have images of dollar bills instead of traditional solid color bars.

The main code for the bar chart was straightforward. Making it clickable involved firing an event that was listened for in index.html.

Adding the image was more challenging. After some Googling, I determined the best option for images on bar charts in d3 was the svg “pattern” element. As it turns out, the svg pattern element is very confusing and bad. After many failures with pattern, I decided to return to Google. Eventually I discovered the svg “image” element, which is infinitely better than pattern. I familiarized myself with the use of this element and then found a nice cartoon image of money online (Image 5). I set it to have a flexible aspect ratio and then encoded it with the same height and width as the bars in the chart.

Finally I added buttons to allow the user to clear income selection and switch between 7 money stacks and 4.

Image 5.

Filtering

The next stage was adding filtering--the primary source of interactivity in our visualization. We used an object to keep track of which filters were currently applied at a given time. Filtering by income occurs when one of the money stacks is clicked. I also added buttons to allow the user to filter based on sex and age.

“Interruption” of TotalVis chart

We took a look at the paper that Daniel suggested as a way to solve the problem of one bar (representing the total people surveyed) being much, much larger than our others (representing the number of heroin users in the last year and the last month). The former can be as large as ~55,000, while the latter value is never larger than 227. However, this solution did not seem optimal because it seemed like the scope of the paper was much broader than the problem we needed to solve, and implementing such a new concept (without much documentation) seemed too time-consuming for the size of our problem. In addition, we found several other good solutions. The one we chose involves creating two different scales on the same (x-)axis. The first scale takes up the left half of graph and is scaled from zero to the second-highest value, while the second scale takes up the right half of the graph and is scaled from the second-highest value to the highest value. As a result, we are able to plot all three bars on the same graph in a way that allows the user to actually see the relationship between the smaller bars, while also understanding that the larger bar is significantly bigger. For clarity, we made the ticks on the scale on the right-hand side purple to distinguish from the scale on the left-hand side. We also made the larger bar purple to signify that it is being scaled according to the purple scale. We did the same with the scale on the left, making its ticks orange.

Interactivity of pie chart

There are many examples of interactive pie charts which provided a good starting point for adding interactivity to our pie chart. However, all of the examples we found construct the pie chart as a stand-alone visualization. That is, none are instantiated as a new object as part of a multiple coordinated view. We wanted to pursue the latter approach and coordinate the views using event handlers in order to maintain the integrity of the structure of our visualization. Doing this, however, was a little more challenging to implement. Specifically, to transition the pie chart, we used .attrTween to call a custom Tween function in order to transition the chart appropriately. However, when using .attrTween, the "this" context for the function that is called is the current DOM element. This was a little problematic in our implementation, because the Tween function didn't have any previously declared variables in scope. To overcome this, we simply (although the solution didn't present itself for a while) copied the necessary variables (specifically D3's arc generator) into the Tween function. We then stored the displayed angles in this._current and used d3.interpolate to transition between the stored angles and the new angles of the pie chart.

An unexpected side-effect of this approach is that when the pie chart has no data (e.g. we are filtering for specific incomes, age ranges and gender with no heroin users), the chart had bugs updating. Not only would the chart fail to draw when there was no data, on further updates, the chart still didn't draw correctly. This was because we were trying to interpolate from one arc to the next, and when the previous arc wasn't drawn, it couldn't interpolate to the next arc. To fix this, we added a condition where if the pie chart is passed no data, it does not update. Rather, we hide the pie chart by making its opacity 0. This is the behavior we want, because if there is no data, it doesn't make sense to draw a pie chart to visualize it. It also matches the behavior of the crimeVis, where the bars disappear if there is no data. On update, if there is data, we unhide the pie chart (by making its opacity 1) and updating. This avoids the problem of interpolating between a broken graph and new one.

Evaluation

This project certainly furthered our understanding about the power of visualizations. A number of trends that are now easily discernible in our views were not available to us from our dataset. Consider the trend we found about heroin use by income group. The 75k+ income group had the greatest number of users per income group. One or two of us voiced the sentiment that they imagined heroin use would be more divided among the lower income groups; however, this was not the case. Last week, we came across a

New York Times article discussing heroin use in the United States and how people from higher income groups were using heroin more than people from lower income groups but wealthier people are more sheltered from society and as a result are less likely to be seen as users. Another trend we found when exploring our data in the visualization was use fairly split between men and women (men did use more than women and their criminal records were more numerous and serious), but we imagined that this would not be the case. We realized that a number of times we did not have a real basis for of assumptions, but our feelings of surprise when looking at the data were telling and evoked interesting questions (e.g. is a person in a relationship more or less likely to use heroin if their partner is using?). Ultimately we are quite proud of the final project we have produced because of the numerous trends one can glean from playing around with the filters.

When we were first discussing views for the project, we discussed incorporating some sort of New York Times/RSS feed to add context to our project for our users and encourage them to continue to explore the topic of interest. We think this was a better idea than we at first gave it credit for. It was super simple to implement, but it gives the whole visualization more credibility, and connects to all those people immediately affected when the see articles about heroin in New York and Vermont, for example.

There are a number of ways we could further develop our visualization that we have discussed amongst one another as we have worked on our project:

1) Incorporate a time slider into our visualization. We were not able to add this time filter because this data was not available to us. However, part of the reason we decided to do this project was out of concern for the growth of heroin use in the country. While the numbers in 2012 marked the highest numbers in the country's history up until that point, showing the upward climb of heroin use would have been a compelling addition to our project.
2) Add other drugs to compare and contrast with heroin use. On the one hand, our focus on this project concerned the demographics of people using heroin in the country and we would not want to take away from our focus by adding a number of other drugs; however, if we were to come up with a way to incorporate the

demographics for other drug users, we might be able to discover some trends particular to heroin users.

3) Although we did not have the data available to us, as you can see from our first proposal, we had hoped to add a map of the United States to show hotspots of heroin use. In conjunction with the time slider, such a map could add a lot of interesting information and room for exploration for our viewers. Then we could link the articles from the rss feed to the map by state.

These avenues for future development indicate something else we have learned over the course of our project, namely the dataset you choose is extremely important. It must give you data points not only about your general topic, but particular to the questions you hope to answer.

Thank you for your time & enjoy our project :)