# Sex, Drugs, and Munchies

Proposal Document

Ryan Saunders
u0642422@gmail.com
u0642422

Bob Wong
bob.wong@nurs.utah.edu
u0196549

Document v.0.1

Last edited October 2015

Repository: https://github.com/rhyeen/datavisfinalproject

# Background and Motivation

According to the National Institute of Drug Abuse (NIDA) drugs affect the brain by stimulating natural neurotransmitters. Neurotransmitters are the chemical switches which tell brain neurons to turn on or off. Most drugs activate dopamine. Dopamine is the neurotransmitter that regulates movement such as physical activity, emotions, and pleasure. Drugs can have negative long term consequences by habituating a user's neurotransmitter levels. In other words, drug users have to have higher levels of neurotransmitters for the effect of movement, emotions, and pleasure to occur. Another effect of drugs, especially marijuana, is increasing appetite. To investigate the relationship between drugs, pleasures, activity and appetite we are using data garnered by the Center for Disease Control and prevention (CDC).

# Project Objectives

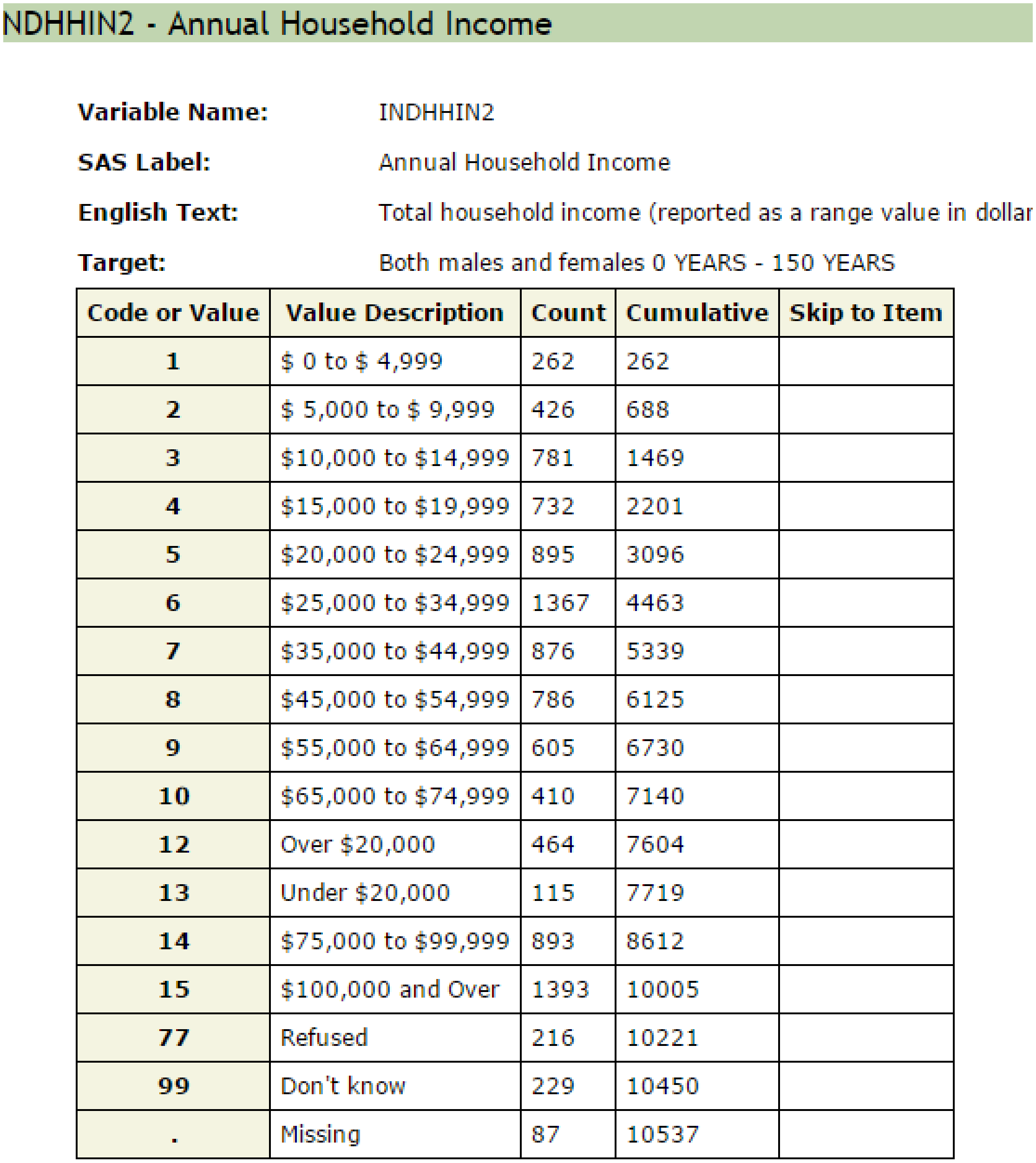
The primary objective for this visualization is to allow the user to investigate the relationships between demographic characteristics and four different behavior constructs (sexual activity, drug usage, food expenditures, and physical activity). The user will be able to select different demographic characteristics such as gender, age, race, income, etc. and see how a national sample of people with those particular characteristics responded to sexual activity, drug usage, physical behaviors and eating habits. Our visualization will also allow the user to pick behaviors (e.g., used marijuana daily for the last month) and see typical user demographics as well as other behaviors like money spent on eating out. The ability to see a wide range of behaviors and to specifically see subsets of demographics will allow the user to investigate the relationships between all the constructs. We will also design the project to incorporate multiform visualizations to allow the user to compare different profiles on the same page at the same time thus capitalizing on the ability of eye vs. memory.

# Data

The data we are using for this project comes from the National Health and Nutritional Examination Survey (NHANES). The NHANES is a multi-year, multi-cohort complex survey conducted by the Center for Disease Control (CDC). The CDC started the survey in the 1960's and recently have been collecting data yearly. NHANES is designed to assess the health and nutritional status of adults and children in the United States. Roughly, 5000 adults and children are assessed each year from across the nation. Data from the survey is used to conduct epidemiological studies and health sciences research, which helps inform sound public health policy, direct and design health programs and services, and expand the health knowledge for the Nation. For our project we will limit the data to adults 18 years and older and less than 80. The data were collected in the 2009 - 2010 time frame and consists of N = 6102. The following link is where the data can be found: http://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Questionnaire&CycleBeginYear=2009.

# Data Processing

The datasets are delivered from the CDC in SAS transport file format. The data are unlabeled values (1's and 2's as opposed to Male's and Female's). The first step of the procedure was to convert the data to be labeled format. This requires conversion of the CDC codebook in PDF format to statistical syntax which will label the data. Once the data is labeled, converting to CSV and JSON was trivial. There have been many data cleanup issues that were not expected like assuming that income would be an ordinal variable when in actuality it is nominal (see table below). Doing histograms of categorical variables is not correct so recoding data will happen prior to visualization. There are certain other variables such as total number of sexual partners/year that needs to be derived from summation of two different variables (female partners/year and male partners/year) since there are many that report as bisexual. Other issues of data cleanup include how to report missing values. The NHANES dataset utilize numerical missing data indicators (usually 77777 or 99999). Having numerical values such as these will drastically impact any kind of data aggregation such as means and also range values when creating histograms. To deal with these values we will set all missing values to NULL.



## NDHHIN2 - Annual Household Income

**Variable Name:** INDHHIN2

**SAS Label:** Annual Household Income

**English Text:** Total household income (reported as a range value in dollar

**Target:** Both males and females 0 YEARS - 150 YEARS

| Code or Value | Value Description | Count | Cumulative | Skip to Item |
|---|---|---|---|---|
| **1** | $ 0 to $ 4,999 | 262 | 262 | |
| **2** | $ 5,000 to $ 9,999 | 426 | 688 | |
| **3** | $10,000 to $14,999 | 781 | 1469 | |
| **4** | $15,000 to $19,999 | 732 | 2201 | |
| **5** | $20,000 to $24,999 | 895 | 3096 | |
| **6** | $25,000 to $34,999 | 1367 | 4463 | |
| **7** | $35,000 to $44,999 | 876 | 5339 | |
| **8** | $45,000 to $54,999 | 786 | 6125 | |
| **9** | $55,000 to $64,999 | 605 | 6730 | |
| **10** | $65,000 to $74,999 | 410 | 7140 | |
| **12** | Over $20,000 | 464 | 7604 | |
| **13** | Under $20,000 | 115 | 7719 | |
| **14** | $75,000 to $99,999 | 893 | 8612 | |
| **15** | $100,000 and Over | 1393 | 10005 | |
| **77** | Refused | 216 | 10221 | |
| **99** | Don't know | 229 | 10450 | |
| **.** | Missing | 87 | 10537 | |

# Visualization Design

We knew that we'd be dealing with complex relationships between multiple datasets. After reviewing the data types, we determined that the best approach to visualizing the data would be to have separate input and output sections. With the input, the user can filter the domain of data in each set. Multiple data sets could be filtered at the same time, to produce an interesting domain. The output would be directly linked to the filter and it would automatically alter the view depending on the filtering. We discussed the benefits of either allowing very refined filtering,or restricting the filtering to a certain set (e.g. only allow users to filter personal information data, such as age, ethnicity, and gender).

We determined that a more flexible tool would be desirable for the user. The resulting design can be found as Figure 1 below. Figure 2 is designated as an optional view, if there is time. Figure 4 and 5 show discarded designs. Figure 6 shows our brainstorm sheet. More detail about each Figure is given as a caption above each Figure.

# Must-Have Features

For our project to be successful, we must meet certain requirements for the input and output views of our visualization tool:

## Input

The user should be able to toggle through the five categories to filter the data. When a user clicks on a category, all other categories should be reduced (distorted) and the selected category should display all datasets for that category. The datasets should be graphed so that the user knows how the brush filtering will function. Finally, the user should be able to filter the data, even across multiple datasets.

## Output

A user should be able to select the x-axis data to be used for charting. Depending on the data, it will either be a bar or area chart with correct labeling. As the input is filtered, the selected chart should update. Users should be able to copy an existing chart, create a new chart, delete an existing chart, change the filtering and data of an existing chart, and rename charts. Multiple charts can display the same dataset, but with different filtering for comparison.

## Optional Features

For our main view (as shown in Figure 1), we have a few optional features that would better engage a user. We would like to create clean transitions between input categories and datasets as the distortion focuses on the selected element. This may require reworking the design, but it could improve the user experience. We would also consider ways for charts to be layered so that multiple filters can be compared on the same chart. This would require the x-axis to the be same, and probably some intelligence about how many charts can be layered as well as how to generate a legend for the layering.

We would also like to provide a simpler view to work with to get a general idea about the differences in demographics and how they affect the other categories. For an optional feature, we would like to develop a view similar to Figure 2.

## Project Schedule

### Oct 25th - 31st:

Prepare data into JSON format, ready for D3.

Begin process book.

### Nov 1st - 7th:

Start developing framework and skeletal views for the input and output.

### Nov 8th - 14th:

Continue to work on process book to be ready for milestone submission.

Link input and output views using a single set of data.

Complete project milestone by the 13th.

### Nov 15th - 21st:

Complete input view, and link all datasets to the output.

Host content on a public website.

### Nov 22nd - 28th:

Complete output view.

Work on additional optional features.

### Nov 29th - Dec 4th:

Cease all work on new features and thoroughly test visualization.

Prepare process book for submission (we would continually work on the book during each week).

Figure 1 - Final design choice split into two components: Input (filtering of a graph) and output (the graphs themselves). The input uses distortion and heat charts to show relevant data that can be brushed. Brushing is linked to the selected graph (in this case, Graph 2). Graphs can be labeled, copied, deleted, and the x-axis can be changed to match any criteria from the five categories (Personal Info, Physical Activity, Sexual Activity, Substance Use, and Dietary Behavior.

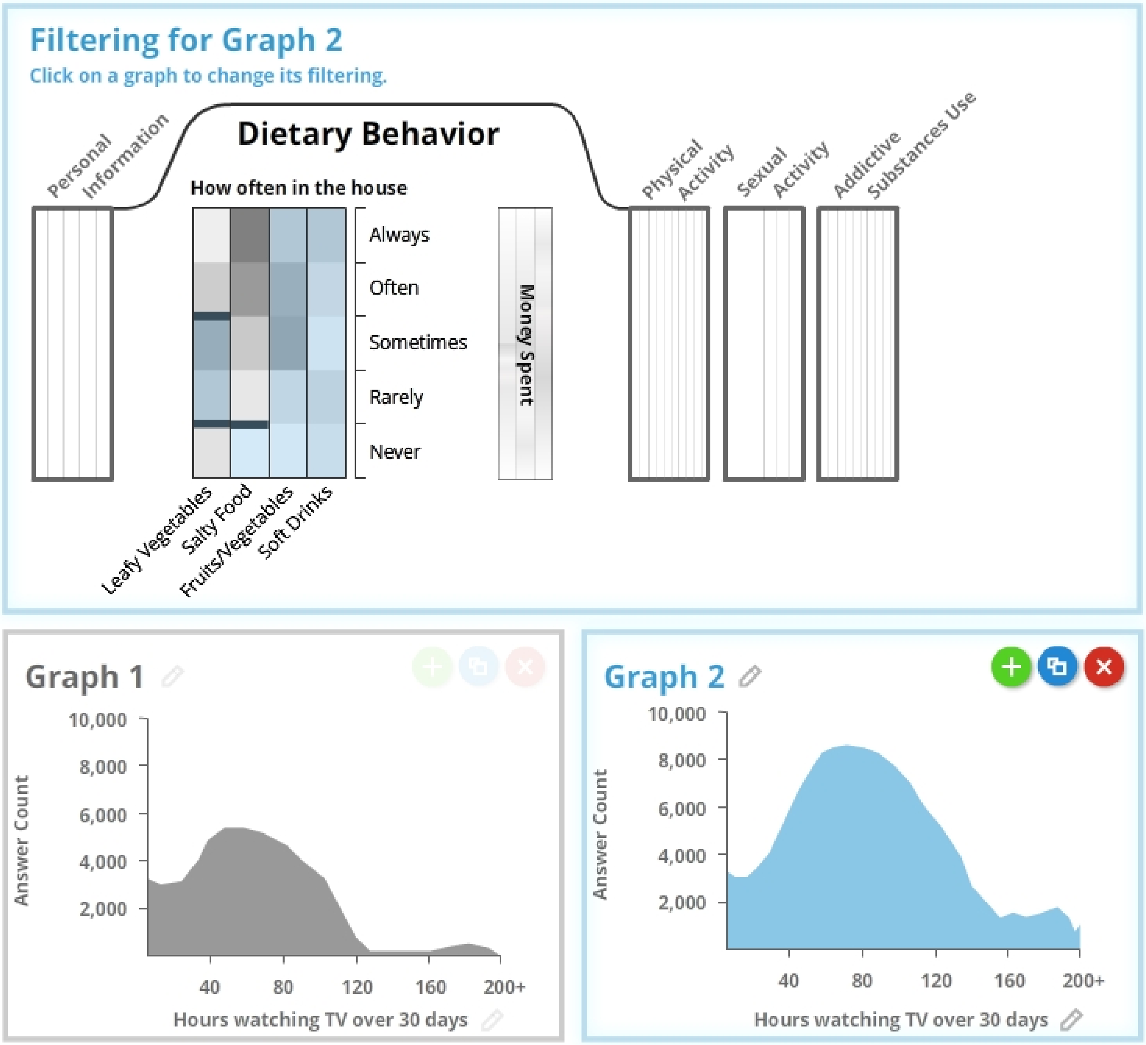
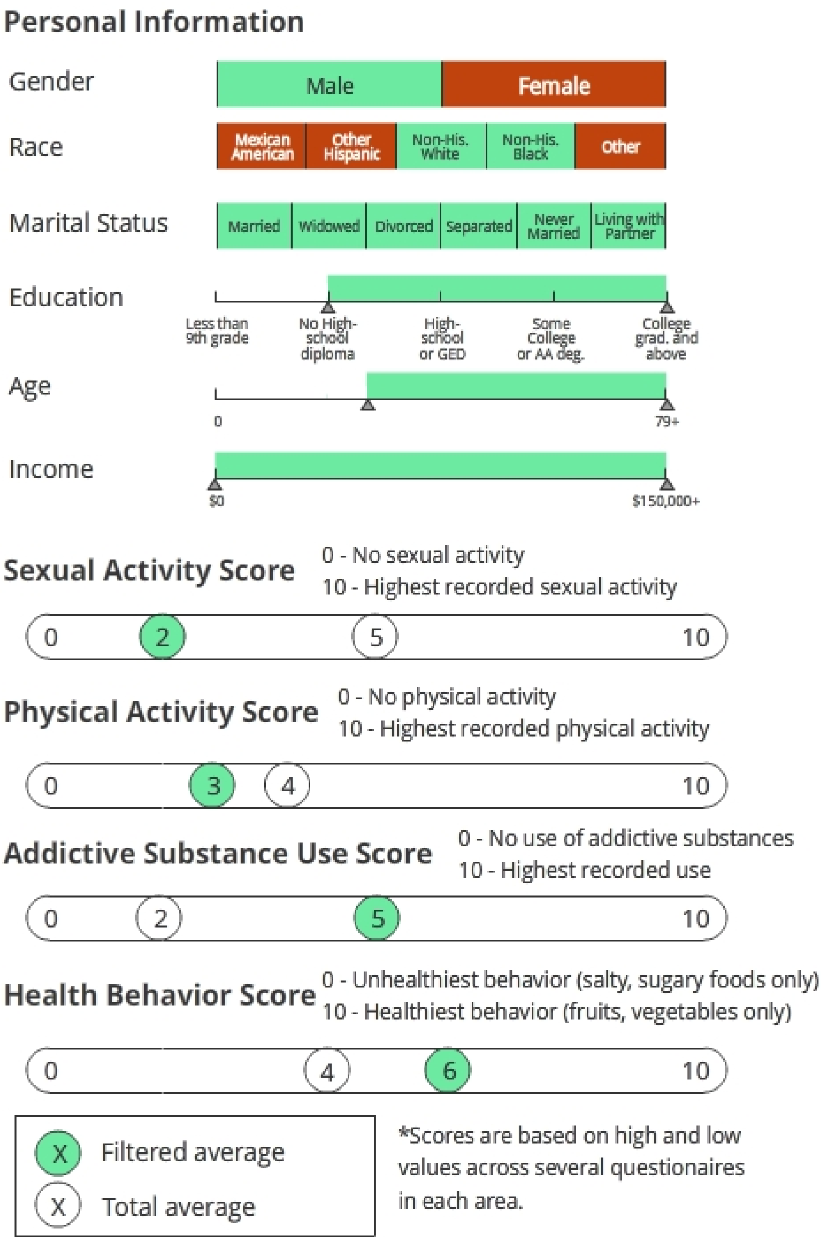


Figure 2 - Optional additional charts that generally show relations between different subsets of the personal information items. Filtering can be done on the personal information, and the colored score is updated to show the change. Scores are computed based on the facts discovered from each of the four remaining categories (excluding Personal Information).



*Scores are based on high and low values across several questionaires in each area.

Figure 3 - Referring to Figure 2, an alternative method to scores would be to display bar charts with the most relevant information from each category given the filter specified. This was one concept design that we considered for this case.

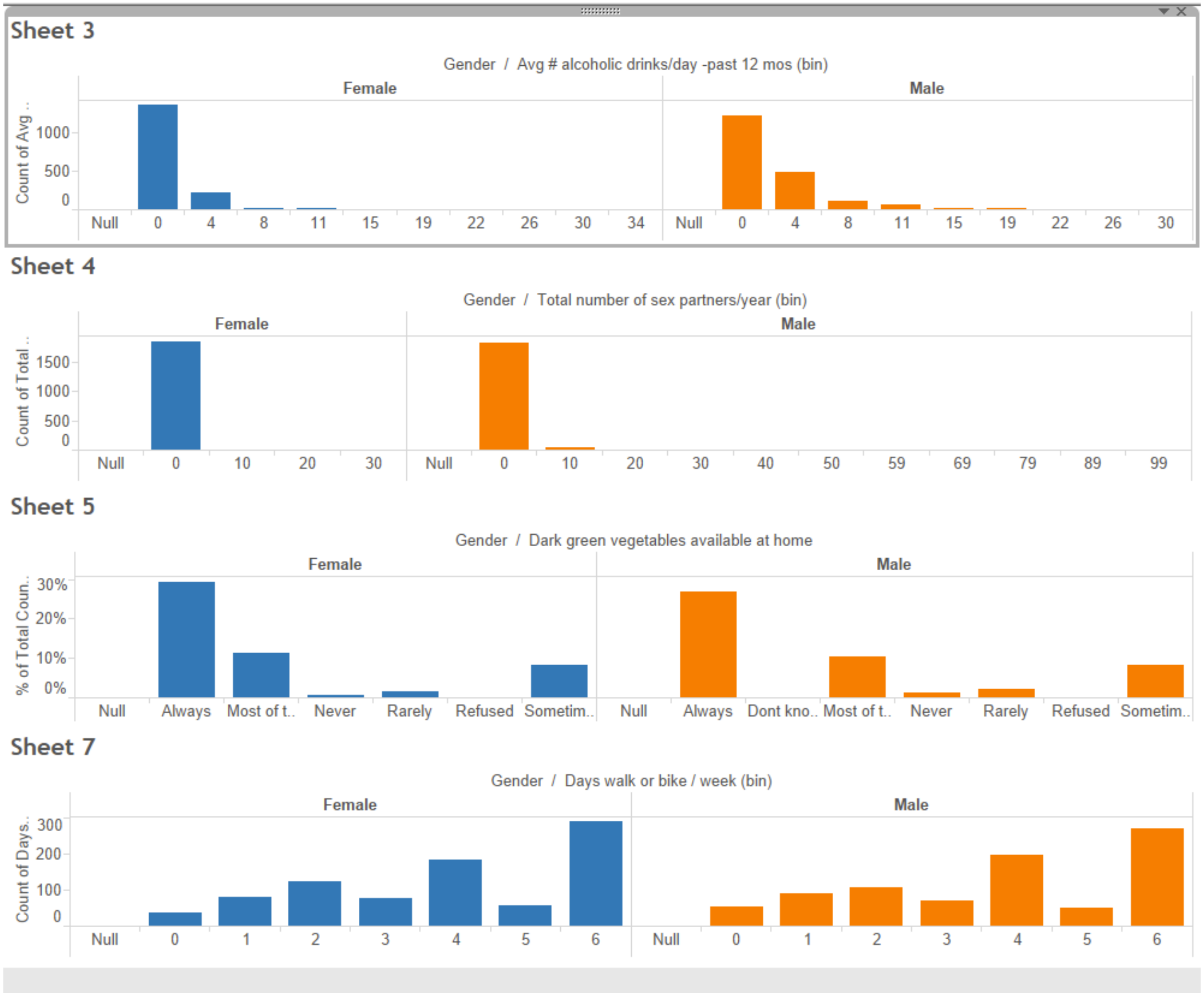
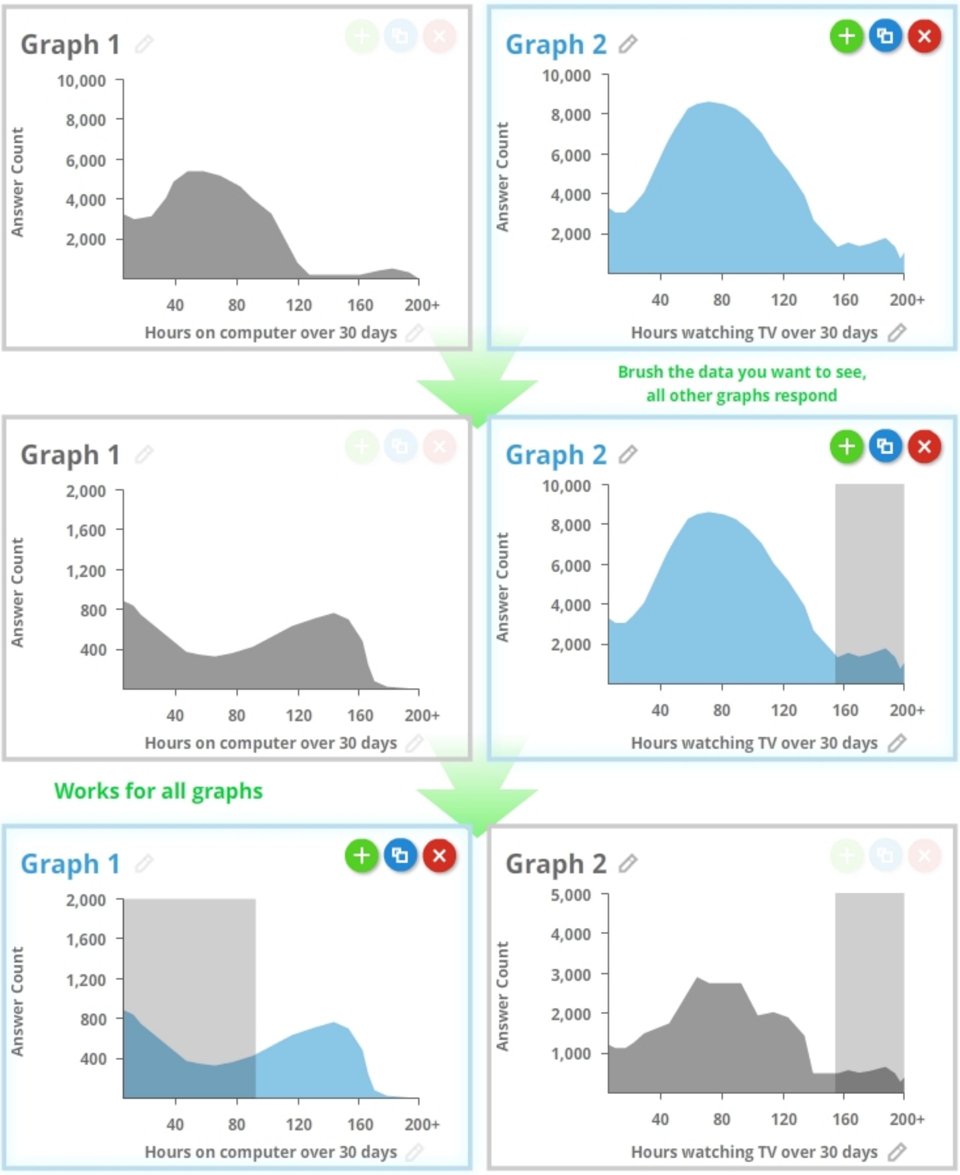


Figure 4 - A discarded design idea. We considered having the user specify the x-axis of different charts based on a given set from the data. From here, the user could brush any graph to filter the domain. This would update all other graphs to match. We discarded the idea as the graphs didn't represent the complexities of data correlation as well as our other design proposals.



Brush the data you want to see, all other graphs respond

Works for all graphs

Figure 5 - A discarded design idea. Taking the idea of Figure 4 and Figure 1 to the extreme, we considered a scatterplot where both the x-axis and y-axis could be changed to match data from the dataset. However, while scrubbing the dataset we realized much of the data is categorical and wouldn't be suitable for a scatterplot. The filtering technique on the left was reused in Figure 2.

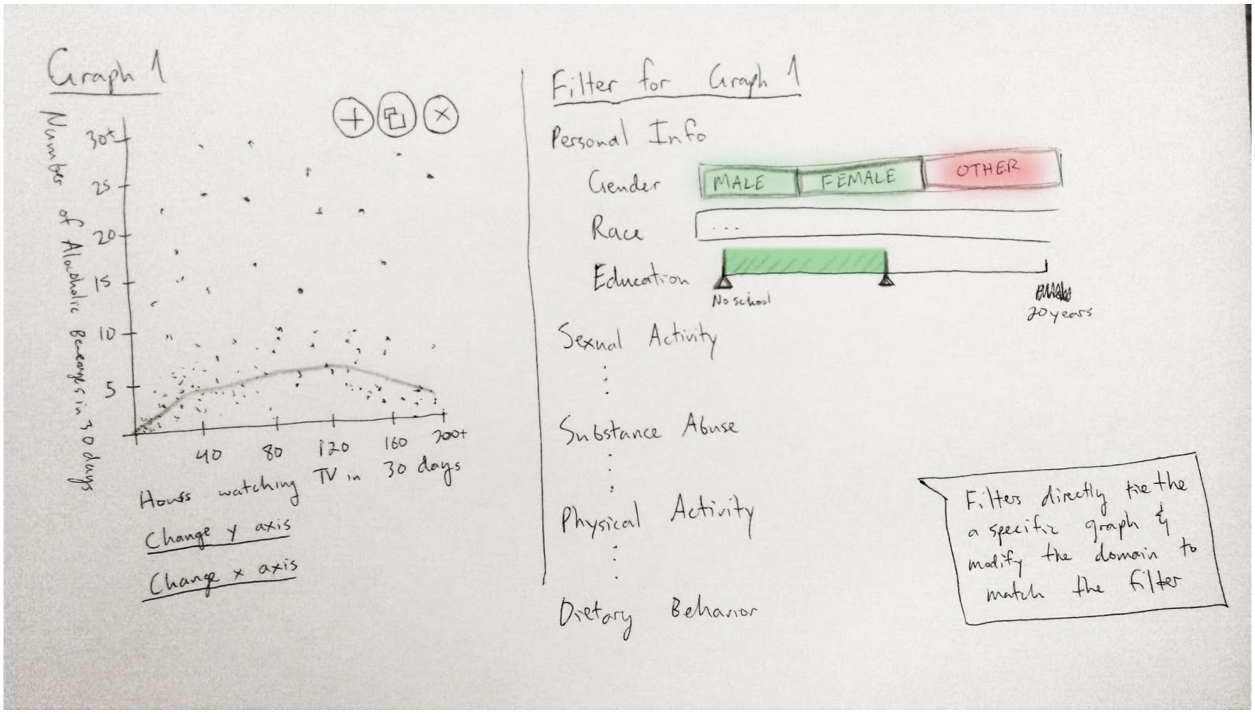
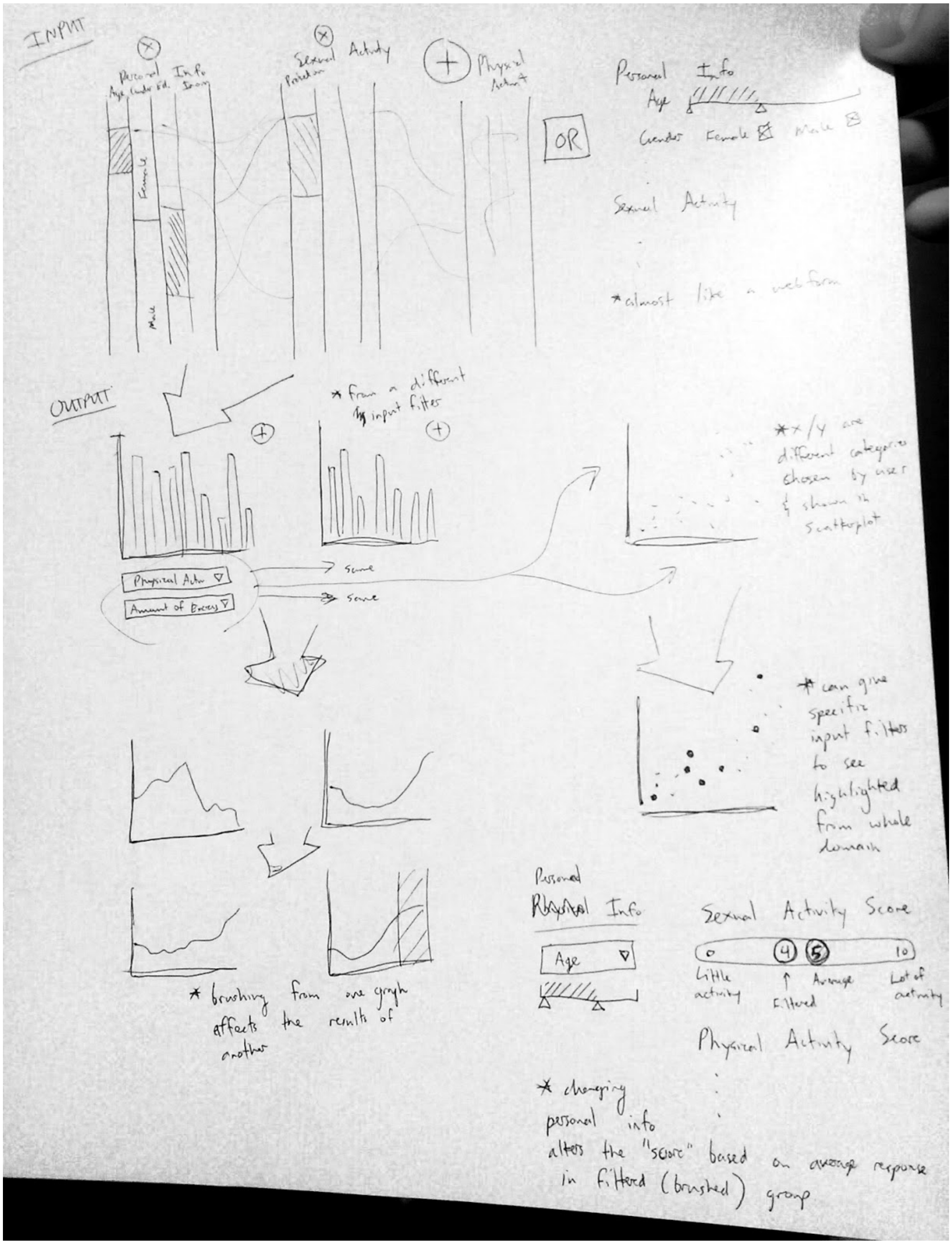


Figure 6 - Our brainstorm sheet.


INPUT
Personal Info
Age Gender Ed. Income
Female
Male
Sexual Activity
Protection
Physical Activity
OR
Personal Info
Age
Gender Female ☒ Male ☒
Sexual Activity
* almost like a web form
OUTPUT
* from a different input filter
Physical Activ ▽
Amount of Beers ▽
Same
Same
* x/y are different categories chosen by user & shown in scatterplot
* can give specific input filters to see highlighted from whole domain
* brushing from one graph affects the results of another
Personal
~~Physical~~ Info
Age ▽
Sexual Activity Score
0
4
5
10
Little activity
Filtered
Average
Lot of activity
Physical Activity Score
* changing personal info alters the "score" based on average response in filtered (brushed) group