Sex, Drugs, and Munchies

Process Book

Ryan Saunders
u0642422@gmail.com
u0642422

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

Document v.1.0

Last edited December 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).

Related Work

We were inspired by many examples demonstrated in class and a few examples outside of class on our own research. Screenshots and in-depth explanations can be found in the appendix.

Questions

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.

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.a in the appendix.

Exploratory Data Analysis

Because we started with more dataset than we could possibly visually deal with, we started by reducing down the dataset to vital categories that we personally believed would have the most interesting results. We created a spreadsheet that showed our initial understanding and organization of this chosen data (Data Review 1). This helped us see that much the dataset could be merged into set categories, as many questionnaire choices followed a similar pattern (something that we didn't realize by viewing the raw data). We then downloaded the raw datasets and converted the data to meaningful values (as explained in the Questions section above). An example of this is shown in Data Review 2.
From there, we further processed the JSON data in javascript so that it could be normalized for D3. We also implemented a small tool that counted the number of values within each dataset for each question. This helped us resolve special cases, and better understand how to visualize the

datasets.

Design Evolution

Alternative designs can be found in the appendix with in-depth explanations and insights into each design (Figure 1-4). Ultimately, we have decided to stick with Figure 1 as our chosen design as was the most flexible option for user exploration.

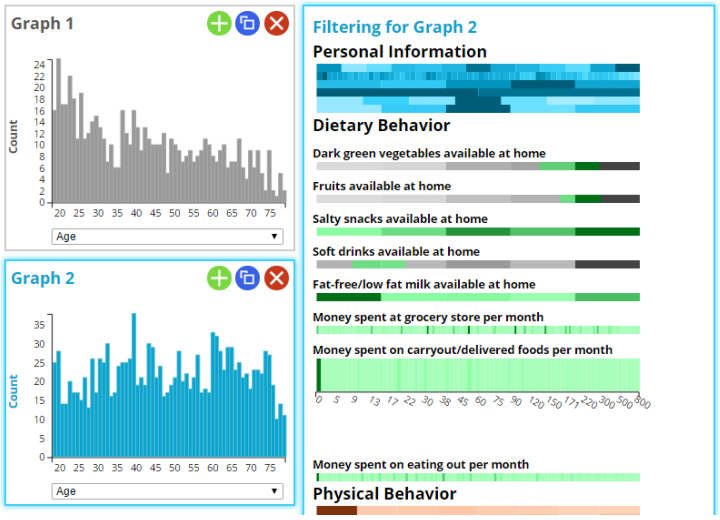
As we began implementing the design in D3 and HTML, we realized the limitations of a horizontal selection of filtering categories (see Figure 1.a). Instead, we decided to use vertical filtering (see Figure 1.b) as it better fit the paradigm of HTML stylesheets, and allowed for a more flexible display of data. We also realized that the graphs presented as area charts would not work for categorical data. Therefore, we switched to a bar chart approach.

We tried to keep our design simple and intuitive to use throughout our development process. Many times we sacrificed greater functionality for elegance and uniformity, but we believe this preserved an inviting visualization.

Implementation

The visualization is meant to allow users to explore correlation among related datasets. There is an output section with user defined graphs for a questionnaire, and an input section that filters the answers of the output graphs based on brushing of datasets. The input and output datasets are the same; however, the input allows interaction among all datasets, whereas the output provides a resulting view of any one dataset. The user is able to compare different filterings and datasets by adding new output graphs, copying previously filtered output graphs, or deleting unneeded graphs. The main idea is to allow for exploration from the user without limiting the results.

Evaluation



It's difficult to state what we've learned from the main visualization, as it allows for a great deal of open exploration; in other words, the results are possibly endless and so the evaluation will always be open ended. However, some interesting results that we discovered were through our own experimentation. One such example is that the histogram of ages for those who have little dark greens and fruits at the home, but plenty of soft drinks shows a clear trend of decline of answers as the age increases (Graph 1 in the figure above). Whereas, the histogram where individuals mostly or always have fruits and vegetables at home, but little soft drinks shows little correlation with age (Graph 2 in the figure above). Whether this shows correlation among healthier foods and older adults or whether it shows unhealthy foods lead to untimely death is beyond the scope of the visualization. But, the flexibility of the system provides interesting results like these.

We believe our visualization fulfill its role quite well. However, one modification that we considered making, but ultimately discarded, was to modify the histogram y-axis in the main visualization output graphs. Currently, the y-axis represents the total count of answers for a particular questionnaire given the input filtering. We considered the idea of instead presenting percentages or ratios, so that it would be easier to compare differences with different filters. However, since we allow for multiple input filtering, this quickly became confusing and decided to stick with absolute answer count.

We also wanted to add additional features, such as linking the y-axis of multiple graphs so that they had the same maximum values, or even overlaying two graphs for easy comparison of data.

Appendix

Figure 1.a - 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.

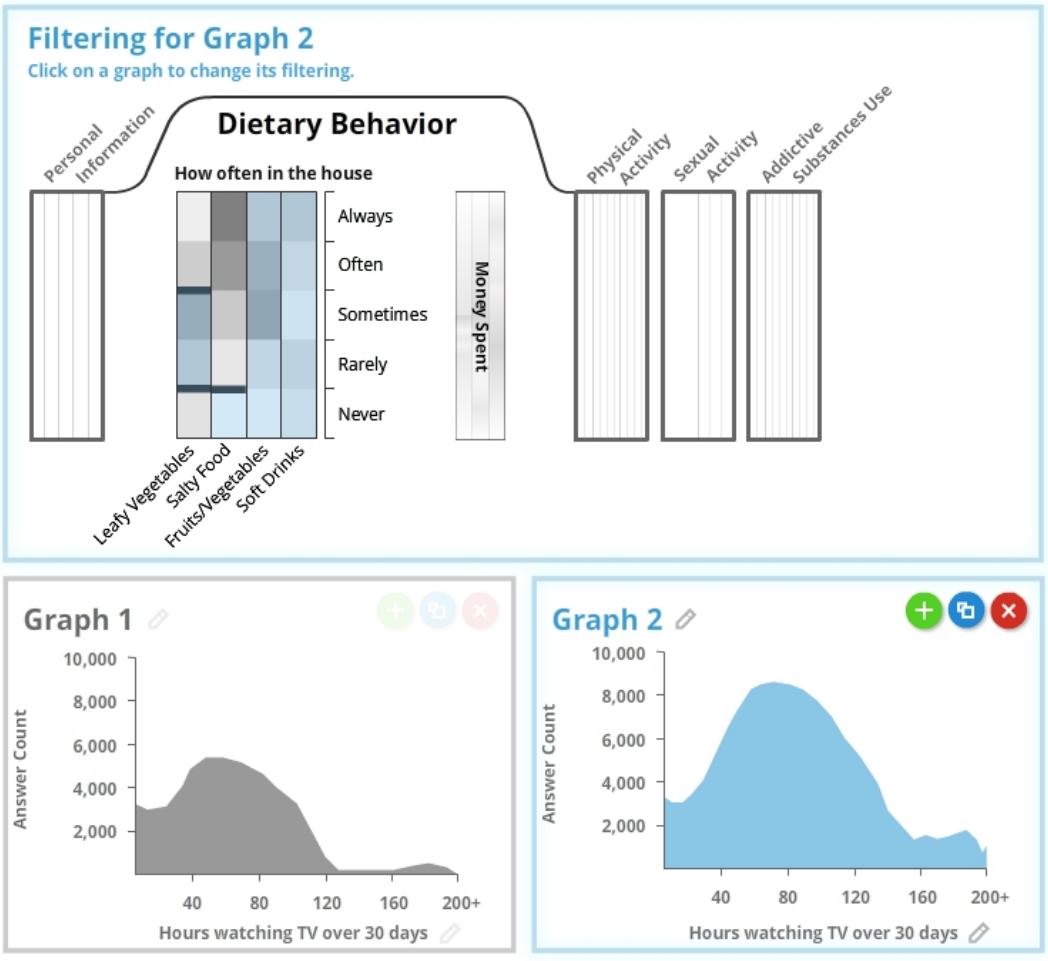


Figure 1.b - Final design implementation derived from Figure 1.a. Notice the switch from horizontal filtering to vertical, and from area charts for graphs to bar charts.

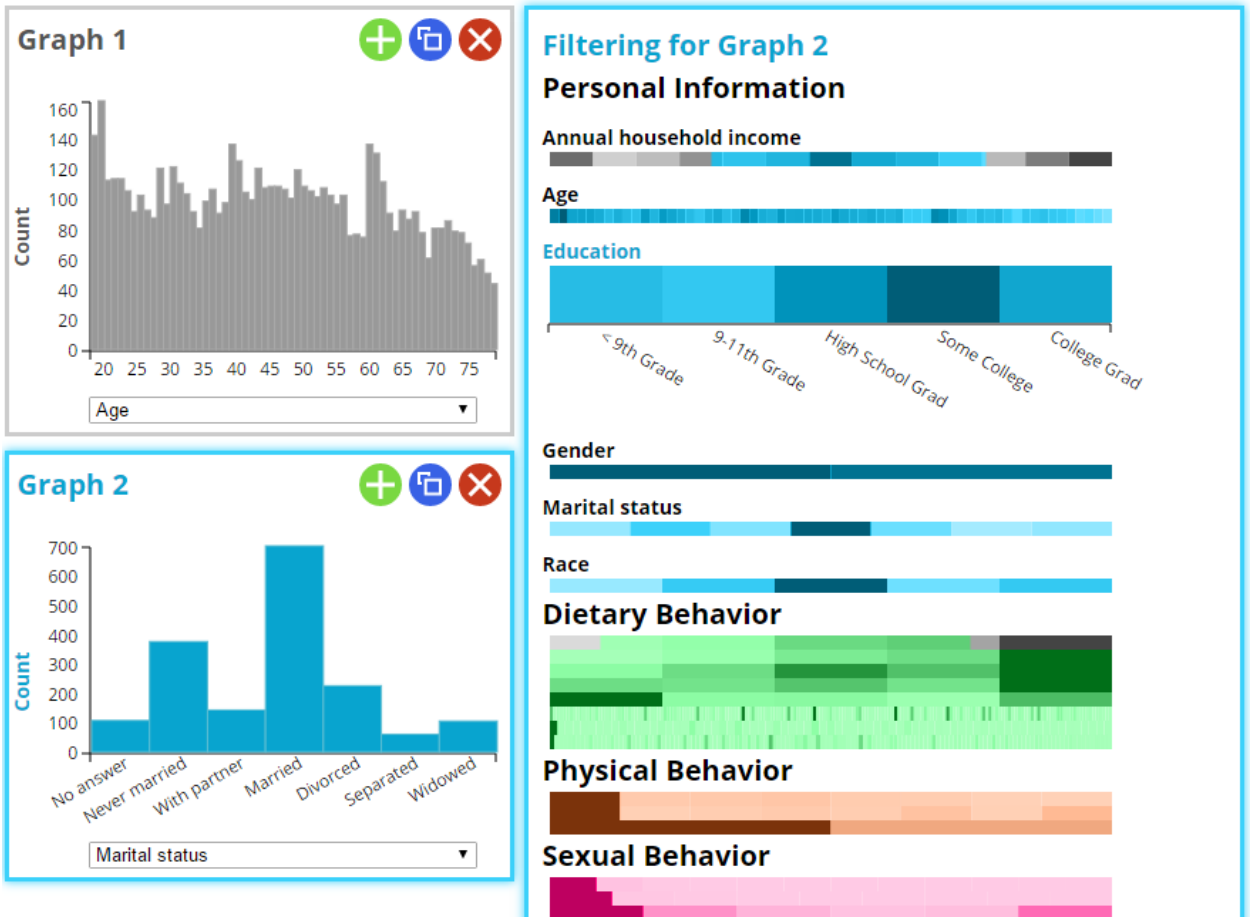


Figure 2 - Optional additional charts that generally show relations between different subsets of the personal information items. Filtering can be done on the personal information, and the colored score is updated to show the change. Scores are computed based on the facts discovered from each of the four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

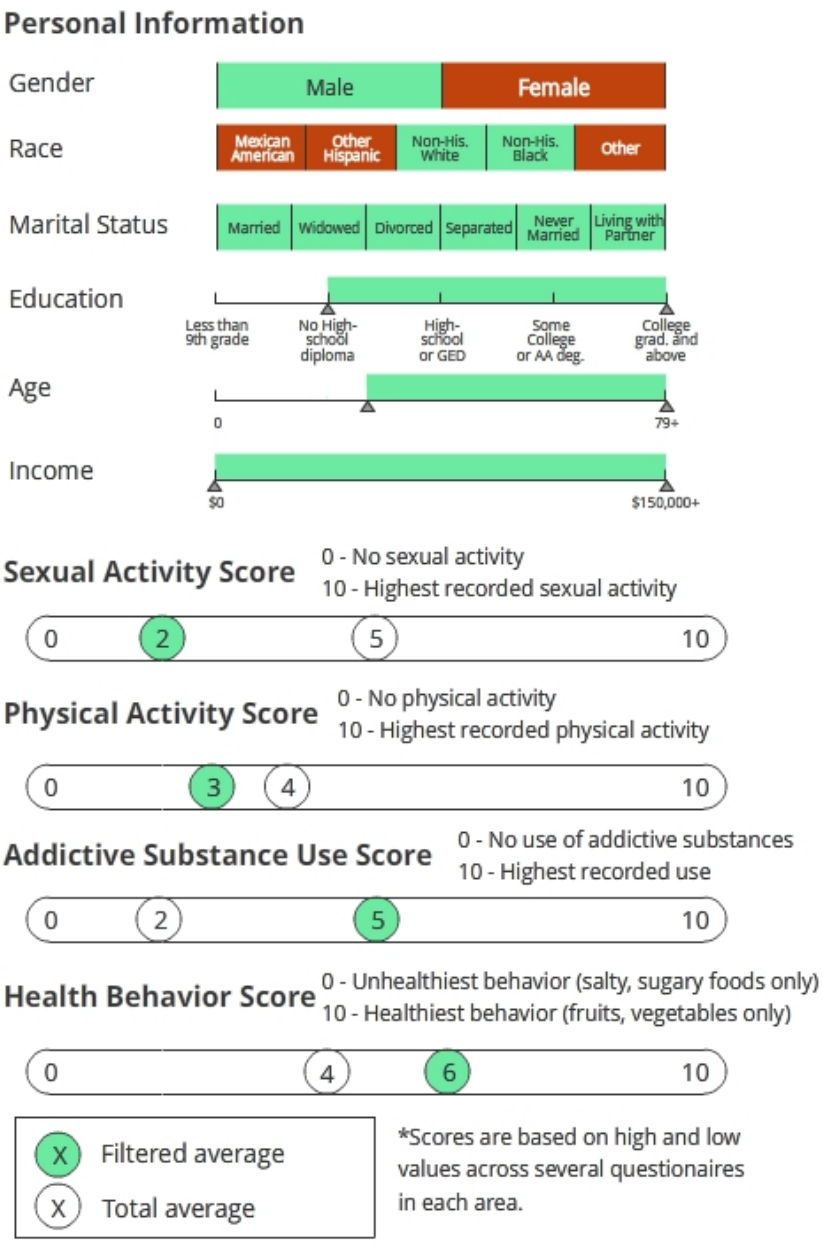
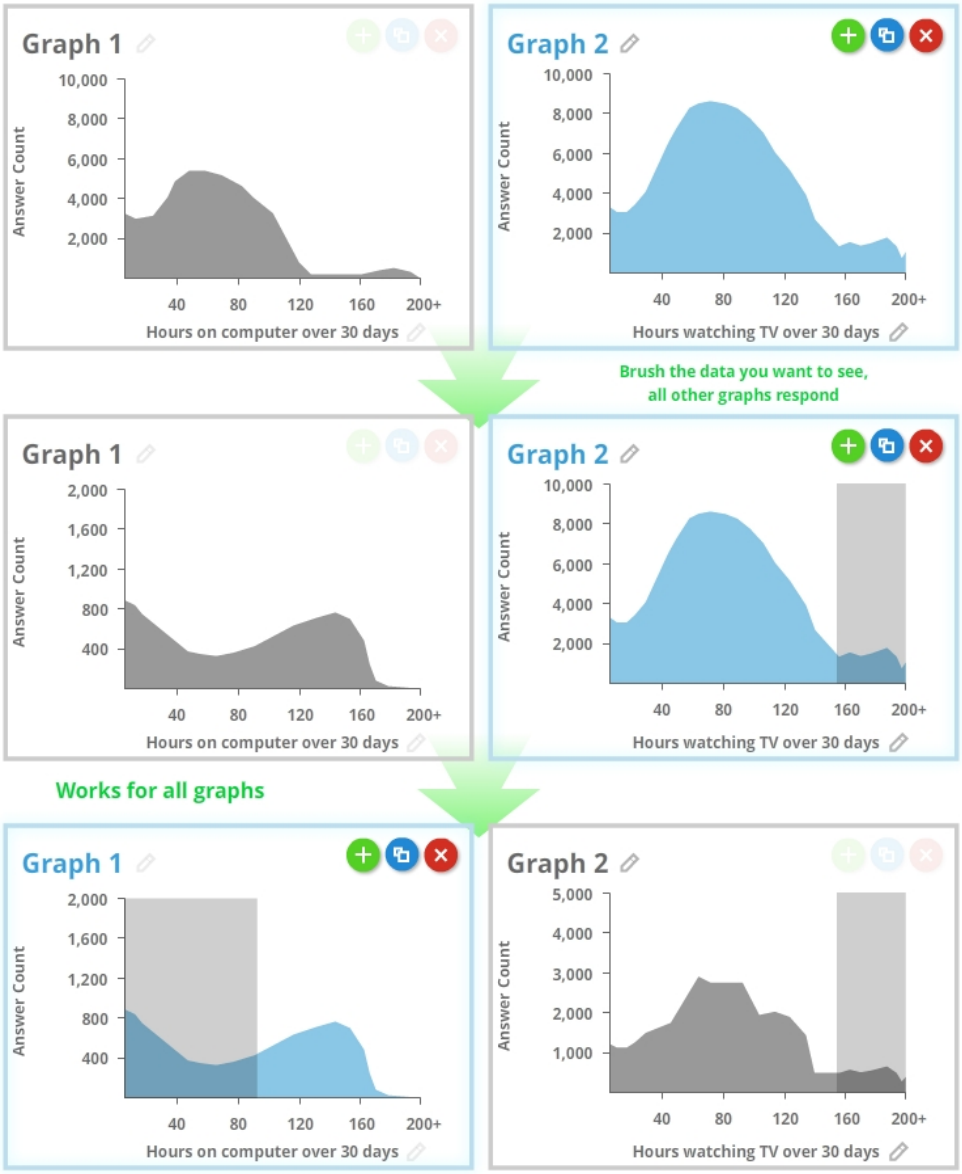


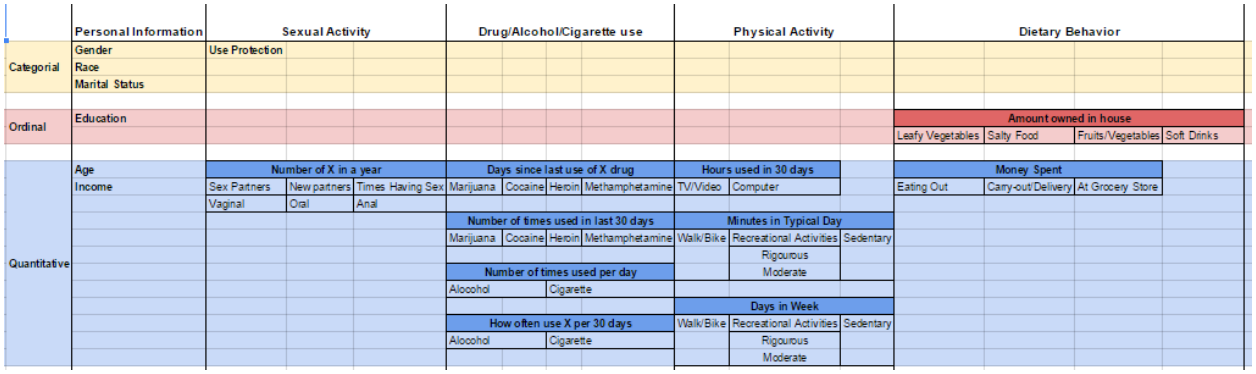
Figure 3.a - Referring to Figure 2, an alternative method to scores would be to display bar charts with the most relevant information from each category given the filter specified. This was one concept design that we considered for this case.

Figure 3.b - Ultimately, we took the ideas of Figure 3.a and decided to make a secondary visualization to showcase one particular story: the difference in dietary behavior of males and females. The d3 code used to develop this visualization comes from Dr. Naushad Pasha Puliyambalath: http://bl.ocks.org/NPashaP/96447623ef4d342ee09b. However, we decided that our visualization too closely resembled Dr. Puliyambalath's work, and ultimately removed this secondary visualization from the final project.

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.



Data Review 1 -

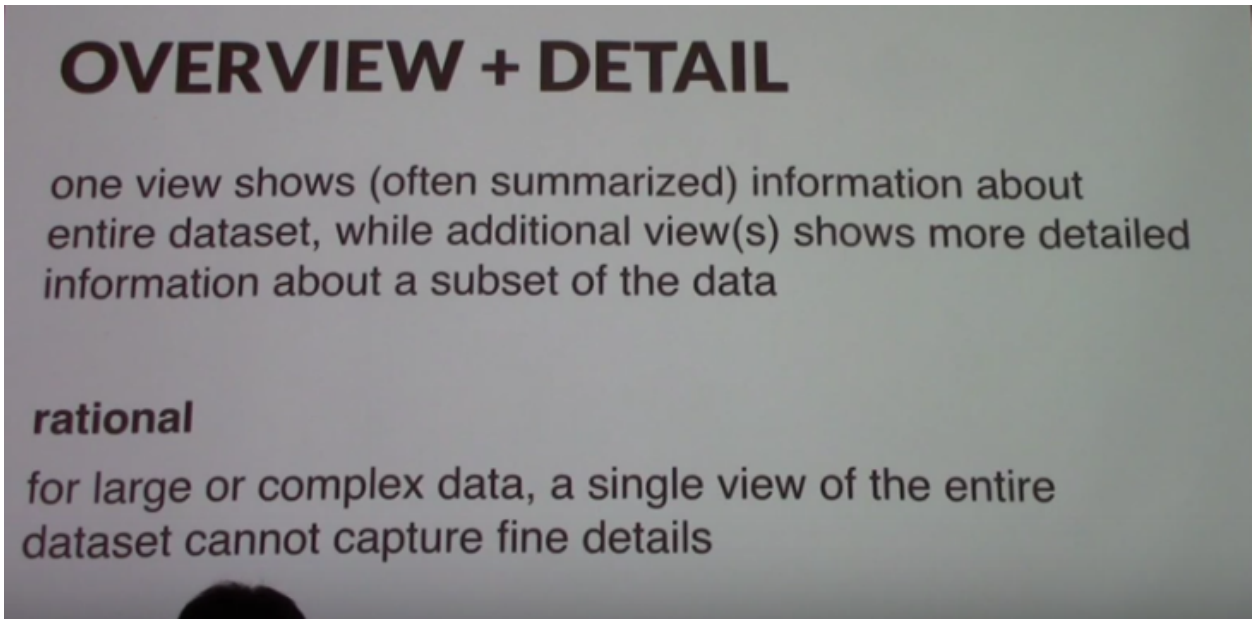


	Personal Information	Sexual Activity			Drug/Alcohol/Cigarette use				Physical Activity			Dietary Behavior			
Categorial	Gender	Use Protection													
	Race														
	Marital Status														
Ordinal	Education											Amount owned in house			
												Leafy Vegetables	Salty Food	Fruits/Vegetables	Soft Drinks
Quantitative	Age	Number of X in a year			Days since last use of X drug				Hours used in 30 days			Money Spent			
	Income	Sex Partners	New partners	Times Having Sex	Marijuana	Cocaine	Heroin	Methamphetamine	TV/Video	Computer		Eating Out	Carry-out/Delivery	At Grocery Store	
		Vaginal	Oral	Anal											
					Number of times used in last 30 days				Minutes in Typical Day						
					Marijuana	Cocaine	Heroin	Methamphetamine	Walk/Bike	Recreational Activities	Sedentary				
										Rigourous					
					Number of times used per day					Moderate					
					Alocohol	Cigarette									
									Days in Week						
					How often use X per 30 days				Walk/Bike	Recreational Activities	Sedentary				
					Alocohol	Cigarette				Rigourous					
										Moderate					

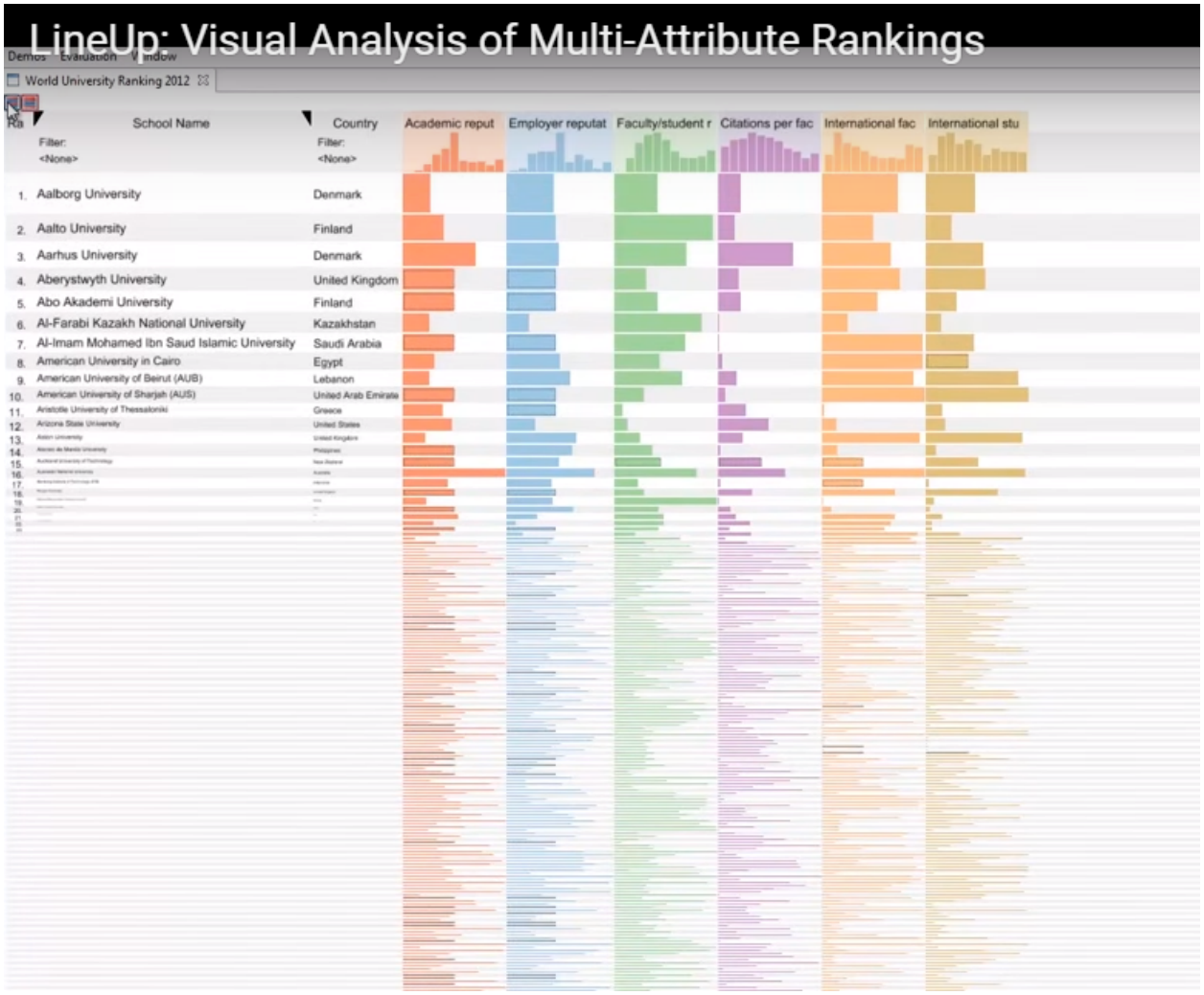
Data Review 2 -

Respon	Data re	Gender	Age in y	Race/H	Educati	Marital	Annual	Ev
51663	6	Female	18	Other Hisp			$20,000 to	N
51924	6	Male	18	Mexican A			$100,000 a	N
52011	6	Male	18	Non-Hispa			$20,000 to	N
52016	6	Female	18	Mexican A			$25,000 to	N
52046	6	Male	18	Non-Hispa			$100,000 a	Ye
52049	6	Male	18	Non-Hispa			$25,000 to	N
52075	6	Male	18	Other Rac			$35,000 to	Ye
52077	6	Female	18	Non-Hispa			$15,000 to	
52123	6	Female	18	Other Hisp			$25,000 to	N
52253	6	Female	18	Non-Hispa			$75,000 to	Ye
52298	6	Male	18	Non-Hispa			Dont know	
52319	6	Female	18	Other Rac			$35,000 to	N
52347	6	Female	18	Mexican A			$100,000 a	N
52612	6	Male	18	Mexican A			$35,000 to	Ye
52623	6	Female	18	Non-Hispa			$20,000 to	Ye
52639	6	Male	18	Non-Hispa			$100,000 a	Ye
52742	6	Male	18	Non-Hispa			$25,000 to	N
52816	6	Male	18	Non-Hispa			$55,000 to	N
52930	6	Female	18	Non-Hispa			$55,000 to	N

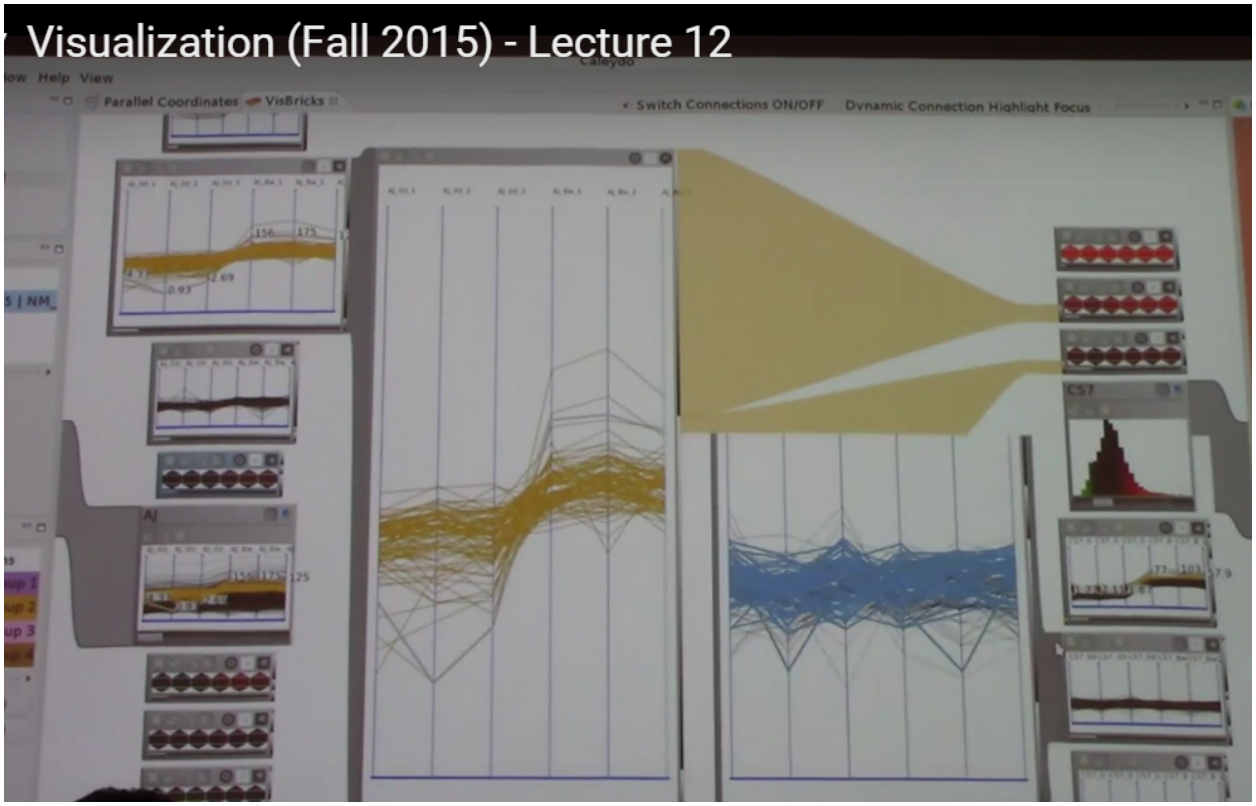
Building from concepts from class:



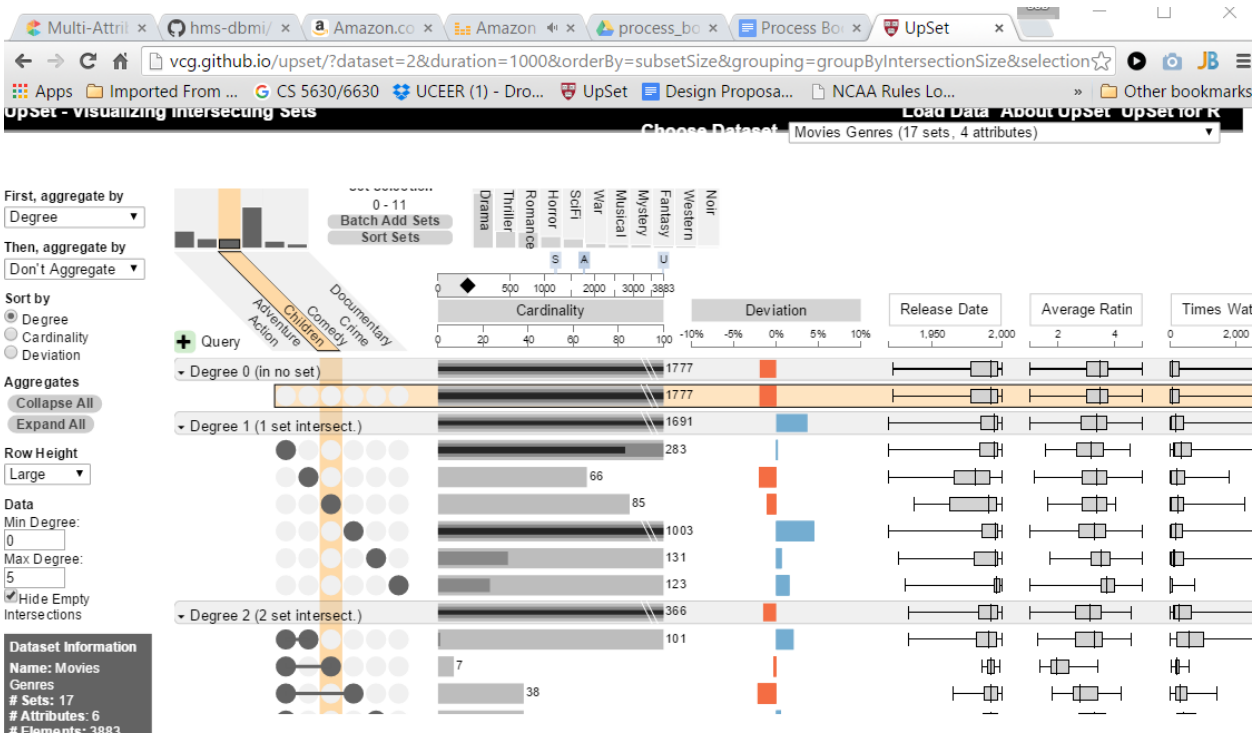
Related work 1 - We thought it would be good to have the ability to look at the data overall, and if on-focus the graphs would be "zoomed" and more detail would appear. Contrarily, if the object (graph) is not-on-focus you could still see the graph in a "diffuse" appearance without detail and clarity. The diffuse appearance would allow more graphs to be on the screen at the same time as the on-focus graph and give the on-focus graph context. The "diffuse" appearance is illustrated below.



Related work 2 - We really liked having multiple views on one screen with ability to filter and subset. Something similar to the graph that was demonstrated in class below.



Related work 3 - Since our dataset has a lot of categorical data, and less continuous variables, we thought we could do something similar using push buttons to filter and subset data as demonstrated in class in the figure below.



Related work 4 - An additional feature that we may include is a dashboard that allows quick comparisons between genders on drug, alcohol, food behaviors etc. The dashboard will be designed similarly to the figure below. The layout will always have gender as the pie chart. And the bar graphs will always have frequency counts. The user will be able to pick which question they want to display.