Pair B Web Project Plan

Presenting General Social Survey data in an easy-to-use, interactive format

Nathan Mannes and Will Schwarzer

1 The Data

For this project, we will be using data from the General Social Survey (GSS), as made available by the Survey Data and Analysis program at UC Berkeley (`http://sda.berkeley.edu/sdaweb/analysis/;jsessionid=62F8F0097CED444E3D18AD8AE40C740C?dataset=gss14nw`). The GSS is a country-wide survey conducted every few years by the National Opinion Research Center (NORC) at the University of Chicago. The data is freely available through multiple sources, including the NORC itself; we have chosen to use the SDA interface to retrieve our data because we have personal experience with it.

As the set of all survey answers is unreasonably large, we will be using a subset of 954 questions over 59,599 respondents. The questions covered include personal info such as income, family and geographical location; economic concerns such as job security and satisfaction; personal concerns such as religion and happiness; and controversial social issues, such as gun control and abortion, among other topics.

2 The Audience

Because understanding the attitudes and thought processes of the American population is of interest to many, if not most, people, we hope our tool will hold general appeal. However, we have still identified 4 separate target audiences for it, with slightly different requirements for the product.

The first audience is the professional social scientists, professors and students who would use our tool for genuine research. These are the people who could likely extract information from the data themselves; we hope to simply streamline the process and make it more enjoyable and interactive. For these users, the robustness of the data will be paramount, including the accurate presentation of relevant statistical measures. They will also undoubtedly require that the raw data be easily accessible.

The second audience is the professors who would use the tool for class material. They could use it as thought-provoking slide material, as the topic for a class discussion, or even as the subject of a student write-up assignment. For these users, the clarity of the graphing will be most important. They will also require that the graphs be high quality and easily downloaded.

The third audience is the members of the general population who wish to satisfy their own curiosity. They might want to know how many people hold a certain opinion, or how two opinions correlate with each other, or how a certain variable has changed over time. For these people, the simplicity and intuitiveness of the user interface will be most important, as well as the clarity of the results - they will not want the results to be overloaded with meaningless (to them) statistical information.

Finally, the fourth audience is the members of the general population who want to prove a point. Perhaps they want to show a friend that the data is on their side of the argument, or perhaps they want to draw attention on social media to an interesting statistic. For these users, the shareability of the results will be critical. To satisfy them, we will have to not only add options to share to social media, but also ensure that our links are robust - i.e. either that a link to a results page always displays the same results, or that we can generate a specific link for a given query.

3 The Requirements

3.1 Functional requirements

- The system must display a homepage with a description of the site, instructions for using the site and the first step of the query process.
- The system must have a persistent navigation bar with site ID, breadcrumbs, a link to the homepage, and a link to an "about" page.
- The system must display a complete list of all available variables at each appropriate step.
- The system must provide plain-English descriptions of each variable (e.g. CONINC: Family income in constant dollars).
- The system must provide the option for the user to either select a second variable for relationship analysis or not.
- The system must allow the user to control for variables.
- The system must generate relevant statistical data for each query (e.g. r, r^2, n, etc.).
- The system must generate a graph for all queries.
- The system must give the user the option to download the data used in a filtered form (i.e. only including the selected variables and the respondents with non-empty entries for those variables).
- The system must give the user the option to share the results of the query on social media.
- The system must display a results page with all relevant information, including a graph of the data, relevant statistical variables, the option to download the data, and the option to share results to social media.

3.2 Non-functional requirements

- The system shall load each page in < 1 second. (The results page could perhaps load slightly slower, such as within 3 seconds.)
- The system shall succinctly describe on the homepage the purpose of the site and how to use it.
- The system shall have an intuitive homepage layout that quickly and easily answers Krug's 5 questions.
- The system shall have a persistent navigation that passes Krug's Trunk Test.
- The system shall allow the user to quickly and intuitively select variables.
- The system shall have variable descriptions that are appropriate, concise and helpful.
- The system shall have a results page that is well laid-out and aesthetically pleasing.
- The system shall have aesthetically pleasing graphs.
- The system shall have graphs that are appropriate for each type of data.

4 Key Features

Most of the features of our site will be designed to make the querying process as simple and intuitive as possible, while still retaining enough functionality to be useful for more advanced users. These are the features that we think work towards this goal the most.

- A 3-step process for selecting variables and control variables:

(1) The user selects the primary variable they wish to analyze. For simplicity's sake, we will integrate this step into the homepage.

(2) The user selects the second variable, if any, whose relationship to the first they want to analyze.

(3) The user selects any variables whose values they wish to control for (probably 10 variables maximum).

In all of these steps, the ability to quickly and easily make variable selections will be critical to the user experience. For that reason, the following features are also necessary:

- A well-organized variable browsing menu. We do not plan on following the SDA's organization of the variables, as we view it to be somewhat arcane; instead, we will re-organize however seems most intuitive, likely with several hierarchical levels of categorization.
- A fully-functional search bar that works as you would expect. This probably means simply substring-matching the search parameters to the variable names or descriptions; if we find another more sophisticated algorithm that is easily implemented, we will also consider that.

We also plan on making the querying experience as lightweight as possible by automatically choosing the best graph type based on the kinds of variables the user chose. Specifically, these are our currently planned graphs:

- Histograms or bar charts for single-variable queries,
- Response grids for double-categorical-variable queries (i.e. what % answered both this and that),
- Line graphs for categorical-versus-continuous queries,
- Bar charts with averages for continuous-versus-categorical queries, and
- Scatterplots for continuous-versus-continuous queries.

We will also include several more minor features, such as an "about" page with citations and a more detailed site description, the ability to download appropriately filtered data, and the ability to share the results to social media. The features mentioned above are those we have identified as core to the user experience, though, and so are the ones focused on in this plan.