Janet Montgomery
MVIS 5301 – Statistical Applications for Visualization
Final Project
28 February 2015

An Exploration of STEM Involvement by Women

Visualization Link: http://mica.thisisja.net/STEM_Assessment.html

"Don't be afraid of hard work. Nothing worthwhile comes easily. Don't let others discourage you or tell you that you can't do it. In my day I was told women didn't go into chemistry. I saw no reason why we couldn't." – Gertrude B. Elion, Nobel Prize Winner in Physiology or Medicine

Topic Selection

Gertrude B. Elion's words echo a past generation's perspective towards women but she in turn advocates equality between men and women. During the initial MICA boot camp, I analyzed the allocation of Science, Technology, Engineering, and Mathematics (STEM) funding within the Federal Government. For this project, I utilized a piece of that STEM funding analysis as the starting point and built off of it to learn more about STEM initiatives and how different demographics were represented within the education and employment data.

Target Audience

The target audience for this project analysis is the general public. I wanted to provide visualizations that someone without a deep understanding of STEM could use to understand the overall sex breakout and within sectors.

Design and Analysis Process

I started the data collection from the education perspective, using National Center for Education Statistics (NCES) data, but I learned that there are differing opinions regarding the classification of STEM post-secondary degrees. Knowing that I wanted to examine employment data, I started with the Bureau of Labor and Statistics (BLS) website, where there is an official recommendation to the Office of Management and Budget (OMB) of STEM Standard Occupational Classification (SOC) codes. I gathered several additional mappings including a mapping of SOC to Classification of Instructional Programs (CIP) Codes, which classify post-secondary degrees, as well as a 2000 to 2010 mapping of SOC and CIP codes.

From the NCES website, I pulled Integrated Postsecondary Education Data System (IPEDS) student enrollment data, but ultimately decided to use conferred degree data. The decision to use conferred degree data was due in part to the CIP Code data being classified at the smallest degree of specificity (6 digit CIP Codes), which provided a link to the SOC codes. I learned the BLS employment data was done in partnership with the Census (via the Current Population Survey). After examining the data available and its caveats between BLS and Census, I decided to utilize the Census' EEO Tabulation. For each data files, I completed standard data cleaning focusing on not available, empty values, special characters, and ensured at least one column was labeled identically for merging reasons. For any data prior to 2010, I translated the SOC and CIP codes to their equivalent 2010 codes to allow for data comparison and categorized them as STEM or Not STEM. For the STEM codes, I also captured their 2 and 4 digit equivalents.

As expected, challenges arose during data cleaning and analysis. One challenge was the Census EEO Tabulation does not capture teachers based on their level and subject matter, because of survey design. This leaves a known hole in the employment analysis. I considered including the teacher data but since the STEM was a small (categorically) component of the overall numbers, I thought it would improperly skew. I looked for an analysis that showed what percentage of the overall teachers numbers were STEM, but was unable to find the breakout. I opted to use Census over BLS data because it represented a larger sample size and allowed for a more granular mapping. Additionally with nature of survey data, I had to be conscious of imputation bias and sample selection bias.

Another challenge I faced was mapping the employment and education data together. In some of the mapping artifacts, codes were rolled up (e.g., instead of 11-1130, 11-1131, 11-1132, it was noted as 11-113x) which resulted in additional data manipulation/translation. Because of the data's many-to-many relationship, the mapping of values at the lowest specificity (6 digit code) proved more difficult than initially planned. I decided to examine the 2 and 4 digit code level relationships, which reduced the duplication of codes. Another challenge was the volume of CIP codes – 2 digit (48 codes), 4 digit (421 codes), 6 digit STEM (883 codes), 6 digit non-STEM (740 codes), which played into design considerations.

To work through this analysis, I used both R and Excel. I manipulated and merged the data using both tools but completed most of the analyses and initial correlations, descriptive statistics, number summaries, and plotting in R (e.g., histograms and scatterplots). With an understanding of the data, I began the translation into visual form. I sketched the basic concepts

of the visualization which turned into the outline of the user journey – from overview to sector specificity.

Creating the Visuals

Early in the analysis and design process, I decided that my goal was to make an interactive exploratory visualization using HighCharts. I projected that it would likely result in a significant number of charts in order to show the sector differences. Because of this and the desire for an easily accessible visualization, I decided to utilize common chart types. This minimizes the time the user spends on learning how to read the chart and instead focuses that energy on exploring the differences between sectors and demographics. This resulted in a visualization built with the following charts: bar, column, stacked bar and column, line, area, and scatterplots.

Conclusion

Looking at the analysis as a whole, I believe it accomplishes the goal – providing an analysis the general public could use to understand where women have a strong presence and where growth is needed. Although it initially presents a lot of information, the user is able to zero-in on sectors of interest, with a summary for comparison. Understanding where women still represent a significant minority allows the public to be better informed in community, school, political, and/or funding discussions. In the future, if I continued this analysis, I would like to examine the changes in education and employment over time as well as breaking out the data by other demographics such as geographic location and race.

Sources

- Census Data - http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
- Census STEM Report - http://www.census.gov/prod/2013pubs/acs-24.pdf
- BLS Employment Data– http://www.bls.gov/data/#employment
- NCES IPEDS– http://nces.ed.gov/ipeds/
- NCES CIP Site - http://nces.ed.gov/ipeds/cipcode/resources.aspx?y=55
- STEM Funding - http://catalog.data.gov/dataset/2010-federal-stem-education-inventory-data-set
- SOC Definitions and Mapping - http://www.bls.gov/soc/

Screenshot of Visualization