Final Write-up: Best US Locations for Data Scientists

One Page Spec

Author	**Dave Langer**
Class	**IS608 Spring 2015**
Reviewed/Approved	**Approved by Josh via email**

The Data

The data used for my visualization is collected and published by the US federal government, specifically the Bureau of Labor Statistics (BLS). The datasets that I leveraged were the Occupational Employment Statistics (OES) collated by the BLS in May 2014. The OES covers a large amount of data, the following table summarizes the exact data that I leveraged in my visualization:

Dataset	Parameters
May 2014 OES Data for US States	• The name of each state. • Data was filtered by OES occupational code "15-0000" which designates Computer and Math-related occupations. • The visualization relies on the data for state-wide total employment, median annual salary, and mean annual salary.
May 2014 OES Data for US Cities	• The name of each city (i.e., metropolitan area). • Data was filtered by OES occupational code "15-0000" which designates Computer and Math-related occupations and for the top 25 cities based on total employment for these occupations. • The visualization relies on the data for city-wide total employment, median annual salary, and mean annual salary.

Mapping the OES state data to the visualization using D3 was trivial as the GeoJSON for the US map already contained state names. However, mapping the top 25 cities in the visualization required obtaining the longitude and latitude for each of the 25 cities. To obtain the geo data for the cities I leveraged the GPS Visualizer web site and incorporated the geo data in my R data munging script.

What the Data Shows

The visualization is based on the premise that the OES data for Computer and Math-related occupations is representative of the Data Science profession and presents a choropleth map of the US based on median annual salary for these occupations. Additionally, the visualization also plots the top 25 US cities (i.e., metropolitan areas) for these occupations based on total employment in each city. Lastly, the visualization provide mouse-over tooltips for both states and cities displaying total employment, median annual salary, and mean annual salary for each.

One major limitation of the visualization is that it does not adjust the raw data in terms of cost of living. For example, the Bay Area of California has the highest salaries for the occupations in question, but the cost of living in the Bay Area is far higher than it is in the Seattle area of Washington state. As such, the effective salary for these occupations in Seattle might actually be higher than in the Bay Area.

Why is it Important?

The visualization is arguably important given the media hype regarding the "sexiness" of the Data Science profession, and the resulting explosion of interest in Data Science occupations in the job market. While certain aspects of the visualization are not particularly insightful (e.g., California have the most employment and the Bay Area having the highest salaries), others might be of interest to Data Science job seekers. For example, Washington state ranking higher than California in terms of state-wide median and mean annual salaries. There is potential for this visualization to alter the career plans/trajectories for those interested in pursuing Data Science as a career.