Final Report - Draft

Elias White: Where Rose Goes

Data Wrangling:

I was able to get two data sets from the career services office. Each contained data from the last 5 years, the first gave me the all the offers submitted to the career services office, along with the company that gave the student the offer, the major of the student, the location (city and state) of the offer, and, when available, whether the offer was ultimately accepted or rejected. The second gave the location and major data, as well as the salary offered to the student. The data as I received it was pretty messy, with lots of spelling mistakes and inconsistent syntax, so a decent amount of standardization was necessary. This involved standardizing the inputs with some python string manipulation, as well as determining where the misspellings are and retrieving the correct value from a dictionary I created with the correct values.

Because I wanted distance values I build up a little application that aggregates the distinct city – state pairs, queries Google's map API with that information, retrieves and then parses the returned xml string for the latitude, longitude, and county information. All of this constituted a dictionary that I went back through and appended to the dataframe columns with the appropriate latitude, longitude, and county data. I then applied the Haversine big circle distance calculator to determine the distance and bearing for each city – state pair from Rose-Hulman, and these were also appended to the dataframe. Lastly there were several outliers (students accepting jobs in Japan or the Philippines), so for visualizations I discard those values.

Visualizations:

One goal I had was to explore different type of visualizations, specifically maps, and see how effectively they convey information. These are a sampling of the preliminary results:

Libraries used:

- Vincent
- Seaborn
- MatplotLib
- Basemap
- Pysal
- Fiona

Type of Visualizations:

Multivariate kernel density estimation: a nonparametric technique for density estimation i.e., estimation of probability density functions. It can be viewed as a generalization of histogram density estimation with improved statistical properties.

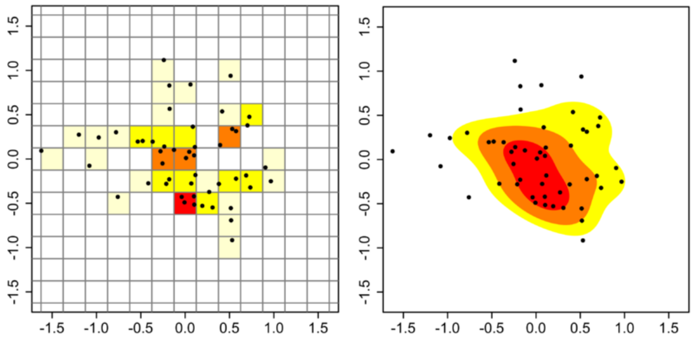
The kernel density estimate is defined to be

$$\hat{f}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i)$$

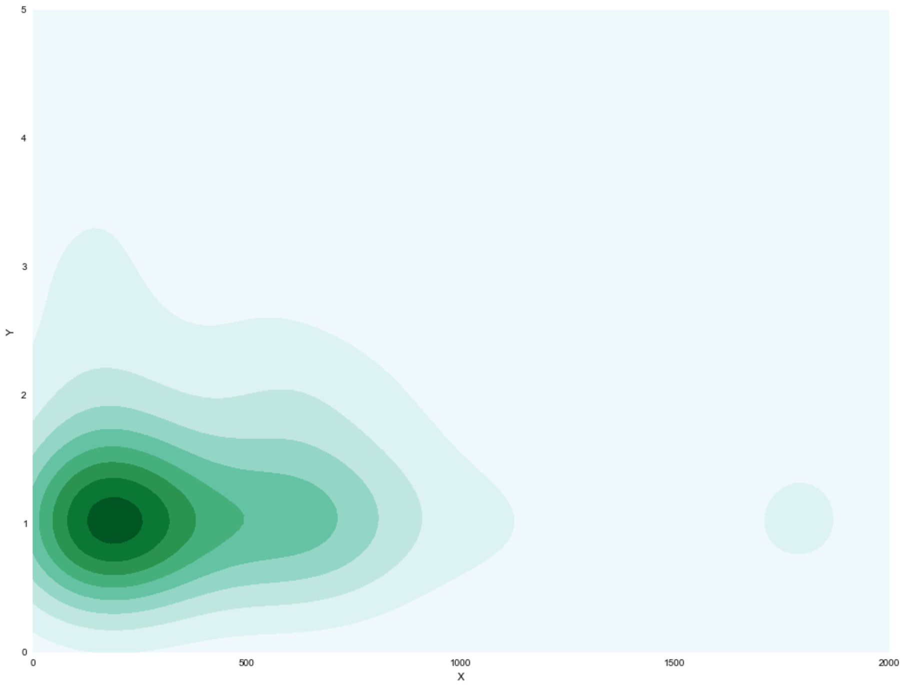
Where

- $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$, $i = 1, 2, \ldots, n$ are d-vectors;
- $\mathbf{H}$ is the bandwidth (or smoothing) $d \times d$ matrix which is symmetric and positive definite;
- K is the kernel function which is a symmetric multivariate density;
- $K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$.

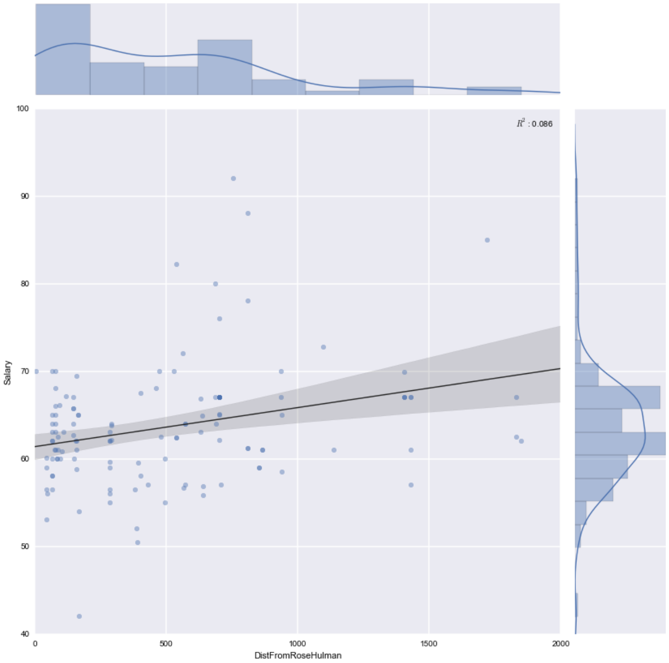
Below is a visual desciption of what this means. The left is the 2-D histogram, the right is the multivariate kde plot:



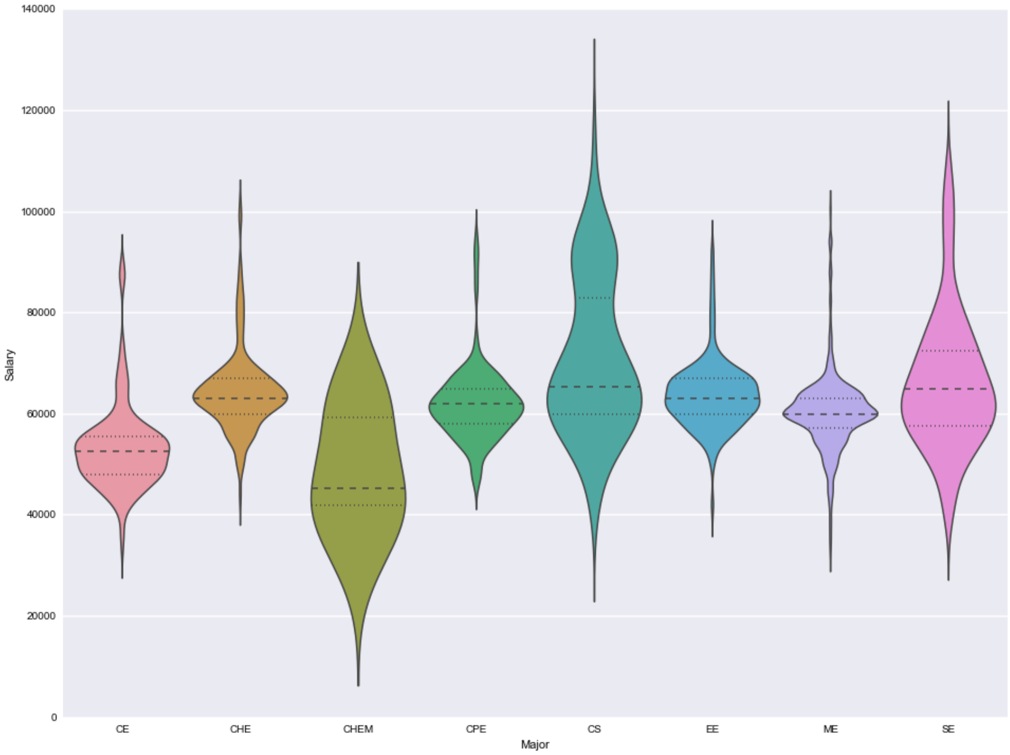
Here is a mkde plot I generated with the y axis showing the number of people accepting a job at a given location (capped at 5) and x axis being distance from Rose-Hulman:



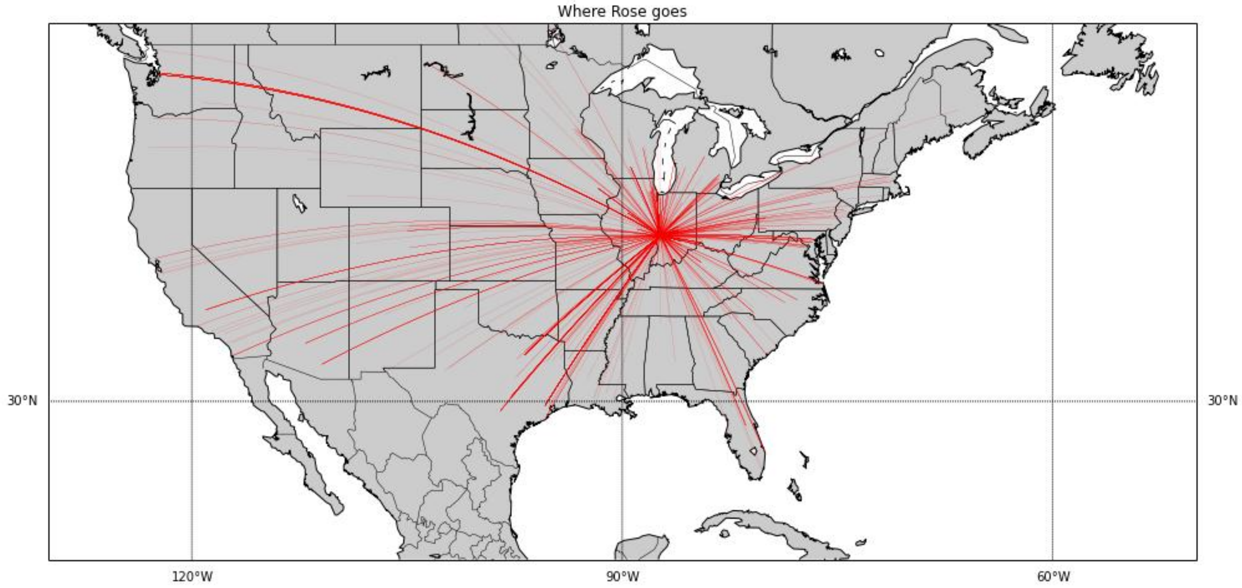
I’ve also done some regression plots. This one shows, for EE’s, the distance from Rose on the x axis and salary on the y axis. It shows that salaries increase as diance from Rose increases:



I also have done some violin plots, which is a combination boxplot and kernel density plot, providing some visual probability indications using width. This one shows salary versus major:



I have a number of others, but as this is the draft final report and I initially was talking about mapping the last image I'll include simply shows a map of the US with all the great circle arc drawn to locations when Rose-Hulman students have been offered a job in the past 5 years:


Where Rose goes
30°N
30°N
120°W
90°W
60°W