Assignment 3 (W4242)

This assignment is due on Wednesday, October 30th, 6:10pm before class. You can discuss with other people, but make sure write down their names on your homework.

Q1[**Machine Learning**] Recall the UFO dataset in Assignment 1.

(a). To better understand this dataset, perform exploratory data analysis on the infochimps UFO data (`http://www.infochimps.com/tags/ufo`). Specifically, we expect to see the following:

- A boxplot of the duration of UFO sightings of each shape (one boxplot per shape).
- A time series figure with the number of sightings per year (one line per shape).
- A bar chart for sightings by state.
- A custom plot (any plot is fine) designed by you to explore an interesting aspect of the data.

(b). You should now move towards identifying interesting insights from the data:

- Normalize the sightings by state population. What do you observe? Anything interesting?
- Visualize the distributions on a map (some options are basemap, D3, or Google Maps API). Do you notice anything peculiar?
- Now explore the data based on your own intuition. Ask and answer at least two additional questions beyond the basic data analysis we require above.

(c). Given your understanding of the data, your goal is now to build a classifier to predict the shape of a UFO. You have three target classes: circle, triangle, and fireball.

- Select an evaluation metric
- Try two different classification methods and compare the evaluation metric

Q2[Visualization](This is an individual assignment. You may not work in groups.) Your task is to design an infographic for the Kaggle dataset. It can either be static or interactive. While you must use the Kaggle dataset, note that you are free to filter, transform and augment the data as you see fit to highlight the elements that you think are most important in the data set. Feel free to join other datasets to this dataset for your infographic.

Part of this assignment is for you to improve your skills in a tool and document that process. The choice of tools you use is up to you (e.g., R, d3, Illustrator, Processing), as long as it is more complicated than a spreadsheet (i.e., no Excel). Here are some sample tutorials to get you started:

1. D3(Data-Driven Documents):
   - `http://christopheviau.com/d3_tutorial/`
2. R:
   - `http://flowingdata.com/2012/12/17/getting-started-with-charts-in-r/`
   - `http://www.r-bloggers.com/basic-introduction-to-ggplot2/`
   - `http://flowingdata.com/category/tutorials/`
   - `http://www.r-bloggers.com/using-javascript-visualization-libraries-with-r/`
   - `http://www.r-bloggers.com/visualize-large-data-sets-with-the-bigvis-package/`
3. Adobe Illustrator:
   - `http://flowingdata.com/2008/12/16/how-to-make-a-graph-in-adobe-illustrator/`
4. Processing:
   - `http://processing.org/tutorials/`
5. A general one:
   - `http://guides.library.duke.edu/content.php?pid=355157&sid=2976256`

To document your process:

(a) Describe the Infographic you would like to create (4-5 bullet points). This link has perspectives from experts in the field about what makes a good infographic :

`http://marketingland.com/8-experts-talk-about-making-great-infographics-34958`

(b) Describe the tools you will use (2-3 bullet points).

(c) Create an infographic. This infographic should effectively communicate this

data and provide a short write-up (3-4 paragraphs) describing your design.

(d) Describe the process by which you improved (2-3 paragraphs). Includes websites you might have visited, tutorials you might have completed, books you might have read. Be thorough.

As different visualizations can emphasize different aspects of a data set, you should document what aspects of the data you are attempting to most effectively communicate. In short, what story (or stories) are you trying to tell? Just as important, also note which aspects of the data might be obscured or downplayed due to your visualization design.

In your write-up, you should provide a rigorous rationale for your design decisions. Document the visual encodings you used and why they are appropriate for the data. These decisions include the choice of visualization type, size, color, scale, and other visual elements, as well as the use of sorting or other data transformations. How do these decisions facilitate effective communication?

Q3[Reading / Blog response] Read the blog post on mapping data to senses and answer the questions at the end of the post. Include your answer in the comments. Your response must be 2-3 paragraphs long and include a link to one outside source (e.g., paper, article).

`http://columbiadatascience.com/2013/10/13/mapping-data-to-senses/.`