Lecture 5: Data Visualization
Modeling Social Data, Spring 2017
Columbia University

Eshan Agarwal

February 17, 2017

1 Guest Lecture

1.1 What is data visualization?

- "The use of computer-generated interactive visual representations to amplify cognition."
- The goal of visualizations is to develop and support hypothesis, and to inspire and convince others

1.2 Exploratory data analysis

There are multiple methods of showing data

- Plain data – very difficult to perceive patterns and make sense of data
- Summary-statistics, e.g. mean, median, etc. – can get an intuitive understanding
- Plot Data – can identify patterns and trends

To create a visualization, we must record data, analyze data, and communicate the findings to the community.

1.3 What makes "good" visualizations?

Identify which dimensions of data should be matched to which areas

1.3.1 Design Principles

- Expressiveness
  - express all the facts/info in the set of data, nothing less and nothing more
  - choose correct type of chart – ex. a scatter plot cannot express a one to many relation
- Effectiveness
  - More effective if information can be conveyed more quickly
  - Easy to understand

Some things to remember about visualizations:

- Tell the truth and nothing but the truth – i.e. Don't lie!
- Use encodings that people decode better
- Not all visual encodings are equal
  - Some may contain biases for example

1.4 Visualization Effectiveness

Steven's Power law

$$S = I^p$$

where S is the perceived sensation, I is the physical intensity and p is the exponential relation.

1.4.1 Perception Biases

- Area - we underestimate large areas over small areas
- Perception of shock increases quicker than the actual level of shock
- We perceive length and position well, much better than color saturation or pie charts (Cleveland and McGill Experiment)
- We can actually rank human perception biases

1.4.2 Data Types

- **Normal** – non intrinsic ordering – eye color, gender, etc.
- **Ordinal** – contains a natural ordering – socioeconomic class, month, etc.
- **Quantitative** – is described numerically

1.4.3 Other Decisions

Color

- How should I color my plot? – not all colors are equal
- Choose colors that maintain distinguishability
- Small changes in color should correlate with proportional changes in value

Tools

- There is a tradeoff between speed and expressiveness
- Declarative Encoding Languages
  - program by describing *what* not *how*
  - separate specification from execution
  - examples are: HTML/CSS, SQL, D3
  - Advantages
    - faster iteration
    - performance
    - reuse-ability
    - portability
  - Disadvantages
    - debugging is difficult
- The Grammar of Graphics
  - Set of principles for graphical APIs
  - "Don't give a pie, give primitives to make a pie and more"
  - Provide small tools that provide more flexibility and customization

2 ggplot

Visit the Jupyter Notebook for exact source code, some observations are listed here.

- Purpose of a plot is to communicate a 10 word point
- Use geoms to represent data points, use aes() function to add aesthetics, variables, axes, etc.
- Pipe commands using '+' not '%>%' – the data frame will default to first argument
- When using aes(), order of arguments doesn't matter – they are added just as descriptors
- **geom_histogram()** – implicit stat counting happening, then maps the count for you
  - Be sure to specify the number of bins – If you don't, some random number will be assumed and a warning will be thrown
  - Identifies categorical variabels and maps them to bins if needed
- **geom_smooth()** – fits a model to the data
  - specify method="lm" to force linear model
- **geom_density()** – fill in plot
- Be careful what to include in aes() – if it is a constant, best to keep it out of the aesthetic mappings
- Be careful when using xlim – may either zoom into the plot, or eliminate points from computation
- R will default categorical axes to alphabetical ordering – often it is better to change this to something that conveys pattern/point better