Intro to Data Analytics and Visualizations

Lecture 9 – Exploring Data
Fall 2014, September12

Outline

1. Why explore before modeling?
2. Tools used for exploration
3. Summary statistics
4. **Graphics and Visualizations**

Why explore?

- To spot problems with the data and derive a first feel for what is going on!
- Potential problems with data:
  - Missing variables
  - Missing observations
  - Missing entire subsets
  - Bad values(dirty/inconsistent)
  - Variables need transformation

Tools Used for Exploration

- Summary statistics (descriptive statistics)
- **Visualizations (graphics)**

Visualizations to Explore Data

- Complementary to numerical summaries;
- Can sometimes spot more issues with the data;
- Give an early feel for some relationships between variables;

Tips for Good Visualizations

- Build it then remove anything non-essential;
- Use colors;
- Make it easy to interpret;
- Avoid background patterns and colors;
- Avoid unnecessary or disorganized text;
- Should convey a lot of information about a clear point.
- Pick the right type of graphical display (e.g. pie charts best avoided)

Visualizing Relationships Between Two Variables

Eg: Relationship between age and income; marital status and health insurance; health insurance and income

Line Plot

- Simplest plot, when a variable is a one-to-one function of the other
- The two variables are numeric
- Rarely the case in practice, a typically pure mathematics tool

Scatterplot

- Most common plot to first explore a bivariate relationship between two numeric variables
- Plots dots; each dot has the horizontal coordinate from one variable, and the vertical coordinate from the other variable
- Can have a fitted line/curve to it to approximate the relationship

Bar Charts for Two Categorical Variables

(eg: marital status vs health insurance status (yes/no))

- Stacked bar charts
- Side- by –Side bar charts
- Others: faceted bar charts, filled bar charts (with relative frequencies summing up to 100%)
- Each emphasizes different aspects of the relationship; pick the one you need, or several of them

In-class Assignment2: Part II

1) Install the “hexbin” package; create a hexbin plot for the age and income variables in custdata2 data frame. How does that compare to the scatterplot?
2) Visualize the relationship between the variables “Number of vehicles” and “Income”. What type of chart do you use? What do you see?
3) Visualize the relationship between the variables “income less than 30k” and “recent move”. What type of chart do you select? What do you see?

In-class 2 Submission

- Inclass 2 Part I + II, one R script; By Monday September 15 at 1pm.
- Submit a script with code and comments to Dropbox and your CMDA Git repository.
- Commit and sync your Git.