Intro to Data Analytics and Visualizations

Lecture 7 – Exploring Data
Fall 2014, September8

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)

Summary Statistics

- In R: summary(data.frame.name) or summary(variable.name) reports:

*See if:

-there are missing values;

-there are nonsense or very high or very low values;

-variable types are as expected (numeric vs factor);

Spotted Issue: Missing Values

- Ask why are they there?(e.g. incomplete dataset)
- What do they mean? (e.g. information not available; variable value is 0)
- What to do with them? (drop entire rows; code them into a zero; code them into a separate category if variable is factor)

Spotted Issue: Outliers

What are they?

-Invalid value (bad data), too high or too low (i.e. stock price of 1,000,000);

-valid value but out of the typical range;

What to do with them?

-Drop them or not?

Spotted Issue: Data Range

- Data not varied enough to explore relationships between variables
- Data too wide, problem for some modeling methods

Spotted Issue: Units of Variables

- Age in years, weeks, months?

- Salary in hourly or yearly multiple of 1,000?

Solution: good data manuals!

In-class Assignment2: Part1

1) Explore the numerical summaries for all the variables in the "custdata" dataset. Comment on what you observe for each one. Address the common issues we talked about (outliers, units, missing values, data range).

2) Explore the numerical summaries for the "uciCar" dataset we worked with in Lecture 6. (car data) Comment on what you observe.

3) Load the "credit.Rdata", the German credit data. Explore the numerical summary for the variable "Personal.status.and.sex". How do genders compare in terms of marital status? Explore the numerical summary for "Other.debtors/guarantors". What do you glean as information about German loans?