Intro to Data Analytics and Visualizations

Lecture 10 – Managing Data
Fall 2014, September17

Outline

1. **Fixing data quality problems**
    - **1.1 Dealing with Missing Values**
    - **1.2 Transformations**
2. Organizing your data for the modeling process

Why Manage?

- Spotted issues in the exploring step (numerical summaries and visualizations)
- Decide what to do about the spotted issues
- Fix
- Re-Organize

  Note: keep track of everything you do.

Missing Data Treatment

- Drop rows if:
  - Not significant proportion of rows;
  - When the same rows have NA across different variables;
- Recode “NA” to meaningful value:
  - When the variable is categorical;
  - When large number of NA’s for one categorical variable;

Missing Values in Numeric Variables

- Missing randomly:
  - Replace (impute) each NA with average of the non-missing values;
  - Replace with better values (better approximation than just the mean).

- Missing systematically:

  - Transform numeric variable to factor; pick meaningful cutoffs for categories; replace NA with “missing” category; (use “cut” R command);

  -Replace with a zero that you keep track of.

Data Transformations

- Why:
  - Data easier to understand;
  - Easier to compare; (eg. Same salary in different states)
  - Easier to model;

- How:
  - Converting continuous to discrete;
  - Normalization and Rescaling;
  - Log Transformations for Skewed and Wide Distributions;

Converting Continuous to Discrete

- When categorical data is better for the model chosen;

- When data patterns are uniform over a few categories (e.g. very young people tend to be on parent's health insurance)

Normalization and Rescaling

- Normalization results in relative values
  - Divide by the median value;
  - Divide by the mean value; (and compare to 1)
- Rescaling results in comparable values
  - Subtract the mean;
  - Divide by the standard deviation.

Log Transformations for Skewed and Wide Distributions

Why?

- Income distribution is skewed to the right typically; values are positive; Other variables follow similar patterns.
- Modeling techniques like bell-shaped distributions.
- Modeling techniques don’t like ranges that are too wide.