36-726: Statistical Practice

Data Wrangling, Data Munging and Data Analysis

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Data Wrangling
  - Transforming “raw” data into data usable for analysis
  - *Documenting data wrangling steps and logic*
- Data Munging
  - Transforming “less raw” data into data usable for analysis
  - *Documenting data munging steps and logic*
- Comments on Data Analysis
  - Different munging for different analyses!
  - *Documenting data analysis steps and logic*

Data Wrangling

Transforming and mapping “raw” data into a format suitable for some sort of analysis

- Highly Iterative
- Highly Task Dependent
- Time Consuming

https://en.wikipedia.org/wiki/Data_wrangling

Typical Data Wrangling Steps

- **Discovering** - Explore the raw data to gain a better understanding of the data: different data is worked and organized in different ways.
- **Structuring** – Raw data is typically unorganized and much of it may not be useful for a particular analysis. Extract the useful parts and organize them (shape/sort/aggregate/…) for easier computation and analysis.
- **Cleaning** – Assuring data quality: Examples include converting all dates to a single format, handling outliers and missing data, removing/merging duplicates, etc.
- **Enriching** – Is additional data available that would enhance the current data set or planned analyses?
- **Validating** – More quality assurance: Similar to **structuring** and **cleaning**, use repetitive sequences of validation rules to assure data consistency as well as quality and security. An example of a validation rule is confirming the accuracy of fields by cross checking data.
- **Publishing** – Make data accessible for analysis (file format, location, etc.).

Be sure to document any and all data wrangling steps and logic!!

https://en.wikipedia.org/wiki/Data_wrangling

Data Munging

- Pretty much synonymous with data wrangling
- When they are used differently…
  - *Data wranging* seems to refer more to processes that start with very raw data
    - (e.g. data scraped from the www)
  - *Data munging* seems to refer to further processing already- or partially-wrangled data with an eye toward new analyses
    - (e.g. linking and further cleaning drivers license and taxpayer lists to count unique persons in the union of the lists, and estimate how many people are missing from both lists)

Typical Data Munging Steps

- The core steps of data munging are the same as data wranging
  - Exploration (Discovering)
  - Transformation (Structuring)
  - Enrichment (Enriching)
  - Validation (Validating, including Cleaning)
- Like data wrangling, data munging is
  - Iterative
  - Highly task-dependent

Be sure to document any and all steps and logic in data munging!!

https://www.talend.com/resources/what-is-data-munging/

Data Exploration/Visualization

- Whether you are looking at completely new/raw data, or you are searching for new relationships in existing data, always begin *(and often return to)* Exploratory Data Analysis (EDA) & Visualization.
- For example:
  - Initial variable definitions (or guesses!)
  - Sample size, record size, dataset format(s), etc.
  - Numerical distribution summaries (center, spread, shape, correlation, etc., etc.)
  - Graphical distribution summaries (boxplots, histograms, scatterplots/scatterplot matrices, trace plots, mosaic plots, fitted curve/surface plots, etc., etc.)

Data Transformation

- Different analyses → different shapes/transformations
- For example:
  - Many analyses benefit from "tidy" data:
    - Each variable forms a column
    - Each observation forms a row
    - Each type of observation forms a table*
  - Even within "tidy" data, some analyses may need
    - *Wide data*: **each row represents one observational unit**, and each column represents a different measurement on that unit
      - (e. g. data from which to calculate a correlation matrix)
    - *Tall data*: **each row represents one measurement**, one column contains the measurement, and the other columns contain the context (e.g. observational unit, occasion, covariate, etc.) for that measurement
      - (e.g. data from which to estimate a regression model)

*this is great from a data consistency pov, but is violated in data sets for multilevel models fitted with lmer() in R, for example…

Data Transformation

- Some other kinds of transformations
  - Converting data from {name:value} pairs to rectangular data frames
  - Combining separate database tables so that data is accessible in one place* (a.k.a. “denormalizing”)
  - Reshaping and aggregrating time series data to dimensions and timespans of interest
  - Filtering (e.g. selecting only certain columns or rows to work with)
  - Splitting and merging columns (e.g. Full name ←→ {First, Middle, Last})
  - Recoding data values (continuous vs discrete, merging categories, etc.)
  - Matching, joining and/or linking data from multiple sources
  - Identifying and removing or merging duplicates
  - Etc.

*this is what is going on with data sets for multilevel models fitted with lmer() in R, for example…

Data Enrichment

- Is additional data available that would enhance the current data set or planned analyses?

- For example:
  - Combining Bureau of Labor Statistics data with Carnegie Mellon job placement data to determine salary ranges for CMU graduates.

- Often a cost (effort) / benefit tradeoff…

Data Validation

- This can involve data cleaning and data structuring/transformation.
- Examples:
  - Correcting, eliminating, or accepting “bad” data values
    - Consistency checks, range checks, logic checks, etc.
    - Deleting records or variables with missing values
    - Imputing missing values, correcting mis-coded values
  - Recoding data values for analysis
    - Treat 12345 rating scale as continuous? Ordered discrete?
  - Matching, linking and de-duplicating/merging dup’s

Comments on Data Analysis

- Different analyses require different munging
- Example:
  - We already saw that correlation may require “wide” data and regression may require “tall” data.
- Example:
  - Suppose we have decided to delete rows with NA’s, and suppose that there are three variables (columns) X, Y and Z with missing values in complementary rows, so that if we delete all the rows with NA’s, there will be no data left.
  - If we plan to do two analyses, one involving the variable X (but not Y or Z) and another involving Z (but not X or Y), we should prepare different data sets for each analysis – one omitting columns Y and Z, and the other omitting columns X and Y – so that when we delete rows with NA’s for each analysis, we will still have data for each analysis.

Documenting Your Work

- *You must carefully document* the steps and logic in your data *wrangling*, *munging* and *analysis*
- Rmarkdown in Rstudio is an excellent tool for this
  - It encourages fully reproducible code
  - It facilitates highly readable code documentation
- I do not think Rmarkdown is the best tool for client-facing communication, but it is excellent for internal, detailed documentation. *I recommend*:
  - Use Word, LaTeX, etc. for client-facing communication
  - Use Rmarkdown internally and for technical appendices

Summary

- Data Wrangling
  - Transforming “raw” data into data usable for analysis
  - *Documenting data wrangling steps and logic*
- Data Munging
  - Transforming “less raw” data into data usable for analysis
  - *Documenting data munging steps and logic*
- Comments on Data Analysis
  - Different munging for different analyses!
  - *Rmarkdown internally; Word, LaTeX, etc. for clients*