Lecture 2: Cleaning and Transforming Data

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

2/16/2016

Overview

1. Getting Data

2. Cleaning Data

Getting Data: Sources

Quality and Veracity

Of foremost importance is that the source of the data set is reliable, that the data comes with information about how it was collected, and is of high quality.

- File, comma or tab-delimited.
- API
  - See DataSources file for links to APIs with Python wrappers.
  - `pandas.DataFrame.read_json` for JSON (see other pandas options for function to read other formats)
- Webscraping
  - See "Data Science from Scratch" for a good tutorial to write your own web scraper.
  - Or use Scrapy

Not all of these options gives data in a nice table format. Pandas has functionality to read many common formats into a DataFrame, but for unstructured data, you may need to write your own parser.

Getting Data: Formats

XML (eXtensible Markup Language)

Markup construct begins with < and ends with >. Three types of tags:

- start-tags `<section>`
- end-tags `</section>`
- empty-element tags `<line-break />`

Content is the text between the tags. Source: Wikipedia

JSON (JavaScript Object Notation)

Syntax looks like python dictionary. Example (compressed):

```
{ "firstName": "John", "lastName": "Smith",
  "phoneNumbers": [ {"type": "home",
                     "number": "212 555-1234"} ]
}
```

Source: Wikipedia

Tidy Data

Tidy Data

- Each variable in one column, with a human-readable variable name.
- Each observation is in one row.
- One table (dataframe) for each kind of variable.
- Multiple tables linked by a common column.

Code Book

- Describe variables in the data set, including the units.
- Explain summary and transformation choices made in preparing the tidy data set.
- Include a thorough explanation of how the data was collected.

For your analysis to be reproducible, you must record every step of the obtaining and preparing the data.

Missing or Bad Data

- [Re]code missing data.
- Remove duplicates
- Identify bad data
  - Values out of range or of wrong type.
  - Outliers

    Warning: data *cleaning* is not data manipulation, so careful consideration must be given when deciding to remove bad records or outliers.
- Examine the missing data
  - Consider impacts to conclusions and inference.
  - Is there an understandable reason for the data, or does this indicate a problem with data collection?
  - Does the missing data prevent you from answering the question under investigation?

Operations to Consider

- Merge or split tables
- Cast variables into appropriate type
- String manipulation to standardize text variables
- Unify terms
- Create intervals

Exercise

IPython Notebook: Cleaning Data

Using the AgCensus data set from the USDA, learn how to perform common data cleaning operations.

References

- Jeffrey Leek
  Lecture: "The Components of Tidy Data"
- Jacqueline Kazil and Katharine Jarmul
  Data Wrangling with Python, O'Reilly Media (2016)

Recommended Reading

Data Wrangling with Python, Chapters 6-8
Data Science from Scratch, Chapters 9-10

Articles for discussion:
For Big Data Scientists Hurdle to Insights is Janitor Work
The Grammar of Data Science
What Went Wrong in Flint Water Crisis, a cautionary tale about removing outliers.