UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 5: Data Collection and Cleaning

Course Plan

Module Titles
Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Module 4 – Pandas
Current Focus: Module 5 – Data Collection and Cleaning
Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop (No Content)
Module 8 – Time Series
Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Using Python libraries for gathering and preparing the data:
  - Discuss types of data
  - Reading and saving data
  - Cleaning up problem data using Pandas
  - Handling missing data
  - Getting data from the web

Topics for this Module

- **5.1** Types of Data
- **5.2** Gathering Data
- **5.3** Preparing Data
- **5.4** Web Scraping
- **5.5** Resources and Homework

Module 5 – Section 1

Types of Data

Types of Data

- Quantitative
  - Measured quantities
  - Results of experiments
  - Scalars or vectors

- Qualitative
  - Categories/labels
  - Text

- Semi-quantitative
  - Orderings

Patterns in Data

- Clusters and correlations
- Points form a line, curve, surface, shape
- Quantities have a distribution
- Complex e.g. represent English sentences

Types of Data Analysis Questions (Ref. Jeff Leek)

- **Descriptive**: What are the main features of the dataset?
- **Exploratory**: What previously unknown relationships exist in the data?
- **Inferential**: What hypotheses do we have about the world and how might the data allow us to test them?

Types of Data Analysis Questions (cont'd)

- **Predictive**: What future events can we predict?
- **Causal**: What happens to one variable when you change another (usually requires a randomized study)
- **Mechanistic**: What is the underlying cause-and-effect mechanism?

Module 5 – Section 2

Gathering Data

Sources of Data - Internal

- Transactional
- Systems health
- Financial
- Concepts or classifications
- Documents or other text
- Email
- Devices

Sources of Data - External

- Financial markets
- Events and news feeds
- Social media: Twitter, LinkedIn, Facebook
- Location-based: cellphones, tracking devices
- Social and economic databases
- Open Data
- 3rd-party data vendors

Data Collection & Sampling

- Data Collection Studies
  - Observational
    - Prospective
    - Retrospective
  - Experimental
- We will study techniques for reducing bias and sampling properly in the statistics course
- Also designing studies, but most corporate data is retrospective

Cognitive Bias

- Concept of Cognitive Bias:

- List of Cognitive Biases:

- Clustering Illusion:

Module 5 – Section 3

Preparing Data

Tidy Data Makes It Easier (Hadley Wickham)

- One variable per column
- One observation per row
- Tables hold elements of only one kind
- Column names are easy to use and informative
- Obvious mistakes in the data have been removed
- Variable values are internally consistent
- Appropriate transformed variables have been added
- Reference: “Tidy Data”, Hadley Wickham, Journal of Statistical Software:

Metadata

- Data about data
- Examples:
  - Database table and column names
  - Tags in HTML and XML
  - Field labels on web pages
  - Timestamps
  - Data ownership and access information

Python for Data Munging

- Cleansing, Cleaning, Wrangling, Transforming: Getting your data in shape for analysis
- Python is a good choice because of its generality
- Pandas is particularly useful for this

Loading/Saving DataFrames

- `import pandas as pd`
- CSV: `pd.read_csv(), pd.to_csv()`
- Excel: `pd.read_excel(), pd.to_excel()`
- Relational tables: `pd.read_sql(), pd.to_sql()`

XL/Wings and ExcelPython

- Tools for live interaction with Excel:
  - xlwings.org
  - github.com/ericremoreynolds/excelpython (as of 2016, ExcelPython has been integrated into xlwings)

Working with Relational DBMS’s

- The Python community is gravitating toward SQLAlchemy:
- SQLAlchemy has both a simple connector to a variety of RDBMS’s and a full Object-Relational Mapper

Reading/Writing NoSQL (Some Examples)

- MongoDB: `import pymongo`
- CouchBase: `import couchbase`
- HDFS: `import hdfs`
- HBase: `import happybase`

Working with Missing Data

- Default N/A is NaN (Not a number)
- `dropna()`
- `fillna()`
- `isnull()`
- `notnull()`
- `na_values=[‘NULL’]` # option in read_csv

Combining Data

- `merge()` # general join
- `join()` # less typing if joining on indexes
- `concat()` # like R cbind/rbind
- `a.combine_first(b)` # splice together overlapping data: a’s values prioritized, use values from b to fill holes

Transforming Data

- Reshaping/pivoting
- Removing duplicates
- Mappings
- Discretization/binning
- Detecting/filtering outliers
- Random sampling
- Indicator/dummy variables

Working with Strings

- Strings and string functions
- Unicode
- Regular Expressions (RE’s)
- Vectorized string functions

Module 5 – Section 4

Web Scraping

Types of Websites

- Static pages
- Dynamic pages
- APIs

Web Page Contents

- HTML
- XML
- CSS
- JavaScript
- Also
  - Images
  - Semantic web markup
  - Microformats

HTML

```
<a href="http://globalmusicdepot.com/store/ca/by-brand.html" class="level-top">
<span>Our Brands</span>
</a>
<ul class="level0">
<li class="level1 nav-1-1 first">
<a href="http://globalmusicdepot.com/store/ca/by-brand/albion-amps.html">
<span>Albion Amps</span>
</a>
</li>
<li class="level1 nav-1-2">
<a href="http://globalmusicdepot.com/store/ca/by-brand/analysis-plus.html">
<span>Analysis Plus</span>
</a>
</li>
<li class="level1 nav-1-3 parent">
<a href="http://globalmusicdepot.com/store/ca/by-brand/angel-lopez.html">
<span>Angel Lopez</span>
</a>
```

JavaScript

```
var xdebug = (function() {

    // Get the content in a cookie
    function getCookie(name) {

        // Search for the start of the cookie
        var prefix = name + "=",
            cookieStartIndex = document.cookie.indexOf(prefix),
            cookieEndIndex;

        // If the cookie is not found return null
        if (cookieStartIndex == -1) {
            return null;
        }

        // Look for the end of the cookie
        cookieEndIndex = document.cookie.indexOf(";", cookieStartIndex + prefix.length);
        if (cookieEndIndex == -1) {
            cookieEndIndex = document.cookie.length;
        }

        // Extract the cookie content
        return unescape(document.cookie.substring(cookieStartIndex + prefix.length, cookieEndIndex));
    }
```

JSON

- Almost all computer languages have a built-in JSON-like data structure
  - JavaScript: Object
  - Python: Dict
  - Java: HashMap

- See Python `json` package

XML

- Extensible Markup Language

```
<name>Joe Cool</name>
<age>34</age>
<status>cool</status>
<girlfriends></girlfriends>
```

- See Python `xml` package

Scraping the Easy Way

- Google Chrome Web Scraper extension:
- import.io:
- Spooky Stuff:

Python Libraries for Web Scraping

- urllib2
- BeautifulSoup
- scrapy
- requests
- Scrapemark
- RoboBrowser
- lxml

Hands-On

- PyData Book 2nd Edition ch06, ch07

- Advanced techniques for self-study: ch08, ch09

Module 5 – Section 5

Resources and Homework

Resources

- Useful blog entry on web scraping with Python:

- The scrapy web scraping library:

- The urllib URL-handling library:

- The requests web scraping library

- The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)

Resources (cont'd)

- Import.io web scraper:
- Beautiful soup HTML/XML parser:
- Good blog on web scraping:
- Python Regular Expressions Cheat Sheet:

Assigment 2: Who Survived the Titanic?

- For this assignment, we will analyze the open dataset with real data on the passengers aboard the Titanic
- Download the data from Kaggle website: file **"train.csv"**
- The definition of all variables can be found on the same page, in the Data Dictionary section
- Read the data from the file into pandas dataframe
- Analyze, clean and transform the data to answer the following question:
  - What categories of passengers were most likely to survive the Titanic disaster?

Assignment 2 (cont'd)

- You might include the following attributes in your analysis:
  - Passenger age
  - Passenger gender
  - Cabin class the passenger travelled in (variable 'ticket class')
- What other attributes did you use for the analysis? Explain how you used them. Provide a complete list of all attributes used.
- Did you engineer any attributes? If yes, explain the rationale and how the new attributes were used in the analysis?
- If you have excluded any attributes from the analysis, provide an explanation why you believe they can be excluded
- How did you treat missing values? Provide a detailed explanation in the comments.
- Submit Jupyter Notebook with your solution via BlackBoard prior to the next class

Next Class

- Descriptive Statistics
- Data Visualization
- Visualizing data with matplotlib

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies