Intro to Data Wrangling

- Gathering Data
  - Best Practice: Download files via script
    - Scalable & Reproducible
- Assessing Data
  - Data Quality
    - Missing Data
    - Invalid Data
    - Inaccurate Data
    - Inconsistent Data

    ] Factors that affect quality
  - Tidiness
    - Structural form of dataset
  - Types of Assement
    - Visual: Look at data in table
    - Programmatic: Look at pandas .info() & equivalent functions
- Cleaning Data
  - Programmatic Data Cleaning
    1. Define
    2. Code
    3. Test
- Finish w/ Reassess & Iterate
- Wrangling vs. EDA vs. ETL
  - Exploratory Data Analysis (EDA): Simple visualization & analysis to maximize potential of later analysis
  - Extract Transform - Load Process (ETL): Copying data from one or more sources into single dataset

Gathering Data

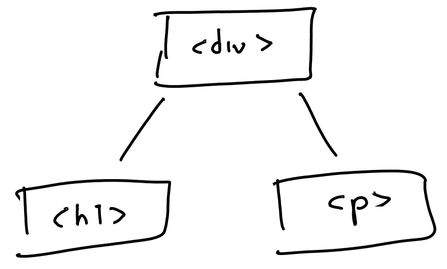
- Flat Files
  - Entries occupy one line w/ specific delimiters
- Web Scraping
  - HTML Elements
    - Heading Elements: used for section heading

      `<h1> This is a Heading. </h1>`
    - Paragraph Elements: used for standard blocks of text

      `<p> This is a paragraph. </p>`
    - Span Elements: used to group text w/in block of text

      `<p> This block has a <span> element </span> </p>`
    - Image Elements: used to embed images

      `<img src="image.jpg" alt="Text Description" />`
  - Trees




```
<div>
    <h1> Heading </h1>
    <p> Body of text </p>
</div>
```

- HTML w/ Python

```
import requests
response = requests.get(url)
with open('title.html', mode='wb') as file:
    file.write(response.content)
```

] Save to file

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(response.content, 'lxml')
soup.find('TITLE')
```

] Parse file

- DOWNLOADING FILES FROM THE INTERNET
  - HTTP IS A REQUEST / RESPONSE PROTOCOL

- TEXT FILE STRUCTURE
  - ENCODING: SCHEME FOR CONVERTING CHARACTER SET BITS TO NUMBERS AND LETTERS
  - CHARACTER SETS: COLLECTION OF CHARACTERS AVAILABLE TO USE W/IN A SYSTEM
- APPLICATION PROGRAMMING INTERFACE (API)
- JSON FILE STRUCTURE
  - GREAT FOR ACCESSING COMPLICATED DATA ARCHITECTURE
  - INTERPRETED AS DICTIONARIES IN PYTHON
- STORING DATA
  - RELATIONAL DATABASE STRUCTURE
    - FETCHED W/ STRUCTURED QUERY LANGUAGE (SQL)
    - DATABASE TABLES
      - DIVIDES INTO ROW & COLUMNS
      - DESCRIPTIVE NAMES FOR COLUMNS, ALL SAME DATATYPE
  - SQL COMMON FUNCTIONS

```
SELECT *
FROM demo.orders
```
] ~ READ ALL FROM DEMO ORDERS INTO EXPLORABLE TABLE

- OTHER FILE FORMATS
  - FLAT FILES
  - HTML FILES
  - JSON FILES
  - TXT FILES
  - EXCEL
  - PICKLE
  - HDFS
  - SAS *
  - STAT *
- DATA GATHERING CAN BE ITERATIVE

ACCESSING DATA

- TWO FACTORS TO CONSIDER
  - DATA QUALITY ISSUES: MISSING, DUPLICATE, OR INCORRECT DATA
  - LACK OF TIDINESS: STRUCTURAL ISSUES THAT MAY SLOW OR INHIBIT CLEANING, ANALYSIS, OR VISUALIZING
- CASE STUDY: DIABETES DRUG PHASE II TRIAL
  - DIRTY VS MESSY DATA
    - DIRTY DATA: ISSUES W/ CONTENT (i.e., LOW QUALITY)
    - MESSY DATA: ISSUES W/ STRUCTURE (i.e., UNTIDY)
  - TYPES VS. STEPS
    - TYPES
      - VISUAL
      - PROGRAMMATIC
    - STEPS
      - DETECTING AN ISSUE
      - DOCUMENTING AN ISSUE
- VISUAL ASSESSMENT
  - DIRECTED: MANUALLY LOOK THROUGH EACH ROW
  - NON-DIRECTED: LEISURE STROLL THROUGH TABLE
- ASSESSING VS EXPLORING

- DATA QUALITY DIMENSIONS

1. COMPLETENESS
2. VALIDITY
3. ACCURACY
4. CONSISTENCY

- PROGRAMMATIC ASSESSMENT
  - VALUE COUNTS
  - SAMPLE

- TIDINESS: VISUAL ASSESSMENT
  - EACH VARIABLE FORMS A COLUMN
  - EACH OBSERVATION FORMS A ROW
  - EACH TYPE OF OBSERVATION FORMS A TABLE

CLEANING DATA

- THREE STEPS
  1. DEFINE
  2. CODE
  3. TEST

- DATA CLEANING PROCESS
  1. MAKE COPY
  2. ADDRESS MISSING DATA
  3. CLEAN FOR TIDINESS
  4. CLEAN FOR QUALITY

PROJECT OVERVIEW

- GOAL: WRANGLE WeRateDogs TWITTER DATA
- DATA: ENHANCED TWITTER ARCHIVE
  - TWEET TEXT
  - RATING
  - DOG NAME
  - RETWEET & FAV. COUNTS
  - IMG PREDICTED BREED
- STEP 1: GATHERING DATA