







Weeks 5 & 6: Data Formats/Data Structures/Data Sources

Weeks 5 & 6: Data Formats/Data Structures/Data Sources

Welcome to Weeks 5 & 6. These two weeks we continue to explore different data structures, and we also start to focus on different data sources. There are some data sources that make data wrangling easy – and then there are some, like a PDF that can be extremely challenging to pull data from. There are also times the data you want isn't accessible at all – and these weeks will provide some techniques for how you might go about getting to the data. Our data cleansing techniques will continue with emphasis on data formats, and dealing with incomplete or inaccurate data. You will submit your second milestone in these two weeks which requires you to do some data clean up on your first data source. Remember – if you have questions, reach out in Teams!

Contents of the Week

-  Overview
-  Readings and Tasks
-  Supplemental Materials
-  Weeks 5 & 6 Discussion/Participation
-  Weeks 5 & 6 Exercises
-  Project: Milestone 2

Objectives/Topics

- A** Read CSV, Excel and JSON files into pandas DataFrames
- B** Read PDF documents and HTML tables into pandas DataFrames
- C** Perform basic web scraping and using powerful yet easy to use libraries such as BeautifulSoup
- D** Extract structured and textual information from portals
- E** Clean and handle real-life messy data
- F** Prepare data for data analysis by formatting data in the format required by downstream systems
- G** Identify and remove outliers from data
- H** Quick look at APIs
- I** Interacting with Databases

Weekly Resources

- 1 [Python Standard Library](#)
- 2 [GitHub](#)
- 3 [Anaconda](#)
- 4 [Jupyter Notebook](#)

Readings and Tasks

Here are your tasks for this week:



Read the following:

- Preface and Chapters 5 & 6 of *Data Wrangling with Python*
- Chapter 6 of *Python for Data Analysis*



Complete the following:

- Weeks 5 & 6 Discussion/Participation
- Weeks 5 & 6 Exercises
- Project: Milestone 2

Supplemental Readings

Readings	Videos
Intro to Data Science Part 2: Data Wrangling . (Souterre, 2018)	
Data Wrangling with Pandas . (Singh, 2019)	
A Comprehensive Introduction to Data Wrangling . (Tomar, 2016)	

Weeks 5 & 6 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is the difference between text based and non-text-based data sources? What are the benefits of each and when would one be used over another?
2. How does someone choose the data source they will store data in? What are advantages of one source type over another? What is a delimited format?
3. What is Tabula? What is it used for?
4. What is web scraping? Why is this important in data science? What are some tools to do web scraping with?
5. What is HTML? What is it used for?
6. What are generator expressions? What are some examples and what are they used for?
7. How do we handle messy data? Are there some rules we should always follow?
8. How do we handle incomplete data or outliers?
9. Why do we format data? Are there any data formatting rules we should always apply regardless of the dataset?
10. What is a Z score? How is it used?
11. What is fuzzy matching? What are some challenges you might face when using fuzzy matching?
12. Define the following terms and explain how they apply to data wrangling:
 - a. Indexing
 - b. Type interface and data conversion
 - c. Datetime parsing
 - d. Iterating
 - e. Unclean data issues
 - f. Serialization

Weeks 5 & 6 Exercises



Complete the following exercises. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

1. *Data Wrangling with Python*: Activity 7, page 207

2. *Data Wrangling with Python*: Activity 8, page 233

3. Insert data into a SQL Lite database – create a table with the following data below that you will create yourself (Hint on how to create the SQL: *Python for Data Analysis 2nd edition* page 191, *Python for Data Analysis 3rd Edition*: Page 199):

a. Name, Address, City, State, Zip, Phone Number

b. Add at least 10 rows of data and submit your code with a query generating your results.

Your exercises are due two weeks from Sunday by Midnight of Week 6. Refer to the rubric for more grading detail.

Submission Instructions

You must submit one consolidated notebook file with the completed exercises. If you are using pycharm, you must submit your .py file along with screenshots or PDFs of your output (code results after the code has been executed). If you submit via GitHub, you must submit either a PDF or notebook file. Do not submit any zip files.

Click the title above to submit your assignment.

View the rubric for this Assignment by clicking on the link below:

[Exercise Rubric](#)

Project: Milestone 2



Cleaning/Formatting Flat File Source

Perform at least 5 data transformation and/or cleansing steps to your flat file data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone.

- Replace Headers
- Format data into a more readable format
- Identify outliers and bad data
- Find duplicates
- Fix casing or inconsistent values
- Conduct Fuzzy Matching

Make sure you clearly label each transformation (Step #1, Step #2, etc.) in your code and describe what it is doing in 1-2 sentences. You can submit a Jupyter Notebook or a PDF of your code. If you submit a .py file you need to also include a PDF or attachment of your results.

Milestone 2 is due Sunday, by Midnight of Week 6. Refer to the rubric for more grading detail.

Submission Instructions

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation should be labeled with description or what it is doing.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed.

Click the title above to submit your assignment.

View the rubric for this Assignment by clicking on the link below:

[Term Project Rubric](#)