# Weeks 7 & 8: Data Cleaning and Transforming

## Weeks 7 & 8: Data Cleaning and Transforming

Welcome to Weeks 7 & 8. These two weeks we are going to deep dive into data transformations and cleaning. Some of these techniques will have already been briefly introduced, but the focus for these two weeks is to provide even greater examples and methods for transforming data. As you are working on cleaning data, think about the data you see and generate and how data could be cleaned before it even makes its way into the database. You will submit your third milestone in these two weeks which requires you to do some data clean up on your second data source.

### Contents of the Week

- Overview
- Readings and Tasks
- Supplemental Materials
- Weeks 7 & 8 Discussion/Participation
- Weeks 7 & 8 Exercises
- Project: Milestone 3

### Objectives/Topics

- A Handling Missing or Incomplete Data
- B Removing duplicates, modifying strings, and pivoting data
- C Working with hierarchical data
- D Grouping Data
- E Date/Time Data – converting and handling period frequencies

### 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:

- Chapters 7-8 & 10-11 of *Python for Data Analysis* (chapter 12 is optional reading this week, please note, if you have the 3rd edition of this book, this chapter was removed from the book in this book so disregard)

Complete the following:

- Weeks 7 & 8 Discussion/Participation
- Weeks 7 & 8 Exercises
- Project: Milestone 3

# Supplemental Readings

**Readings**

The Ultimate Guide to Data Cleaning: When the Data is Spewing Garbage. (Elgabry, 2019)

The Art of Cleaning Your Data: Drop that Bad Data Like Obama Drops Mics. (Seif, 2018)

# Weeks 7 & 8 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. How much time does a data scientist typically spend on data wrangling (cleaning and data preparation)? What are some of the reasons for this?
2. How is missing data typically represented in Python or popular libraries like Pandas?
3. How can you fill in missing data? What decisions should you consider before doing this?
4. Discuss some data transformation steps you read about in your reading this week and explain when you would use these techniques.
5. What are regular expressions?
6. What is hierarchical indexing? When would you use this?
7. Why would you need to merge or join data?
8. What is normalized vs denormalized data? What are the pros and cons of each structure?
9. What are pros/cons of SQL? Why isn't it the most common data science language?
10. What is split-apply-combine? Provide an example of how this is used.
11. What does GroupBy mean and what is it used for?
12. What are some examples of data aggregations? What do the various aggregation methods provide?
13. What is a pivot table? What are pivot tables used for and what are the pros/cons?
14. What is time series data? Describe the different types of information that would be grouped into time series data? Why is time series data important?
15. How should time zone data be handled?
16. The time dimension can add a layer of complexity to analyzing data. How do different time periods, like year, quarter, month, etc. play a part in the analysis?
17. What is down sampling? How is it used?

# Weeks 7 & 8 Exercises

The four chapters you read these two weeks focus extensively on cleaning and transforming data.

You can choose from either of these two datasets:

### Submission Instructions

You must submit one consolidated notebook file with the completed exercises. If you are using PyCharm, you

- So Much Data Candy, Seriously. (Ng, 2017)
- The Metropolitan Museum of Art Open Access CSV. (Github, 2019 (this data set has multiple years' worth of data – you can use these files for merging/joining)
- You can also download all of the above data from both sites directly from this link: Weeks 7 & 8 Datasets

For this assignment you need to complete **8** of the following exercises against this data.

> **Note**: You must select at least two methods from each chapter to perform on one of the datasets. You are welcome to do more methods and you do not have to use the same dataset for all 8 methods.

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.

- Chapter 7
  - Filter out missing data
  - Fill in missing data
  - Remove duplicates
  - Transform data using either mapping or a function
  - Replace values
  - Discretization and Binning
  - Manipulate Strings
- Chapter 8
  - Create hierarchical index
  - Combine and Merge Datasets (you will have to either create a new dataset from your existing data or create a relationship between the data I have provided)
  - Reshape
  - Pivot the data
- Chapter 10
  - Grouping with Dicts/Series

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 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

    - Grouping with Functions
    - Grouping with Index Levels
    - Split/Apply/Combine
    - Cross Tabs
- Chapter 11
    - Convert between string and date time
    - Generate date range
    - Frequencies and date offsets
    - Convert timestamps to periods and back
    - Period Frequency conversions

## Project: Milestone 3

*Cleaning/Formatting Website Data*

Perform at least 5 data transformation and/or cleansing steps to your website 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 (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 3 is due Sunday, by Midnight of Week 8. 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