







# Weeks 9 & 10: Advanced Data Gathering and Visualization

## Weeks 9 & 10: Advanced Data Gathering and Visualization

---

Welcome to Weeks 9 & 10! We are almost to the finish line, with just 4 weeks left of the course. While APIs were introduced in Weeks 7 & 8 – over the course of these two weeks, we will explore them in more depth and practice pulling data from one. APIs are growing in popularity and are a data source in the corporate world you can expect to interact with. We will also look at some basic capabilities of data visualization. While data visualization is a course and a topic all on its own, it can prove useful for determining if you have data cleanliness problems. Outliers and categorical data with issues are easy to spot when creating a simple visual. We also have your fourth project milestone due over these two weeks, where you should be cleansing data you are pulling from an API.

## Contents of the Week

-  Overview
-  Readings and Tasks
-  Supplemental Materials
-  Weeks 9 & 10 Discussion/Participation
-  Weeks 9 & 10 Exercises
-  Project: Milestone 4

## Objectives/Topics



- A** Make use of requests and BeautifulSoup to read various web pages and gather data from them
- B** Perform read operations on XML files and the web using an API
- C** Make use of regex techniques to scrape useful information from a large and messy text corpus
- D** Use Matplotlib to explore and visualize data – which aids in analyzing quality of data

## 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:
  - Chapter 7 of *Data Wrangling with Python*
  - Chapter 9 of *Python for Data Analysis*
-  Complete the following:
  - Weeks 9 & 10 Discussion/Participation
  - Weeks 9 & 10 Exercises
  - Project: Milestone 4

# Supplemental Readings

Readings	Videos 
<a href="#">Application Programming Interface</a> . (Wikipedia, 2019)	
<a href="#">Beautiful Soup</a> . (Python Software Foundation, 2019)	
<a href="#">How To Scrape Web Pages with Beautiful Soup and Python 3</a> . (Tagliaferri, 2019)	
<a href="#">An Introduction to APIs (Application Programming Interfaces) &amp; 5 APIs a Data Scientist must know!</a> (Kaushik, 2016)	

## Weeks 9 & 10 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 are the most common data sources you will run into when doing data science in the real world?
2. Where can you find data and how do you verify its authenticity?
3. What are APIs? How are they used and how important do you think this data source will become?
4. What are the available libraries in Python to get to web data?
5. What are the options for extracting meaning from text on a webpage?
6. What resources are available when interacting with APIs or other public data sources?
7. What is the difference between HTML and XML? What are the pros/cons?
8. How are APIs secured?
9. What is pattern matching? What is greedy vs non-greedy matching?
10. What is the difference between a rest and streaming API?
11. What is the difference between plots and visualizations?
12. What are some popular Python packages for visualizing data?
13. What are some of the most common visualizations for data science? Explain when you would use each one.
14. What is the difference between a histogram and a bar/column chart?
15. What is the best way for visualizing relationships in the data?
16. What are some best practices to follow regarding colors and chart types?

## Weeks 9 & 10 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 9, page 294
2. *Data Wrangling with Python*: Activity 10,

### 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).

page 295

3. Connect to an API of your choice and do a simple data pull - you can use any API - except the API you have selected for your project.

a. In previous versions of this course we have always used Twitter, but with recent organizational changes at Twitter, it has become increasingly difficult to access the free APIs available at Twitter. You are more than welcome to try to use Twitter's API for this portion of the assignment, but please note, there has been some inconsistency experienced when following along with their [documentation](#) posted.

b. Connect to the API and do a "Get" call/operation on the API to return a subset of data from the API

4. Using one of the datasets provided in Weeks 7 & 8, or a dataset of your own, choose 3 of the following visualizations to complete. You must submit via PDF along with your code. You are free to use Matplotlib, Seaborn or another package if you prefer.

- a. Line
- b. Scatter
- c. Bar
- d. Histogram
- e. Density Plot
- f. Pie Chart

Your exercises are due two weeks from Sunday by Midnight of Week 10. Refer to the rubric for more grading detail.

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 4



### *Connecting to an API/Pulling in the Data and Cleaning/Formatting*

Perform at least 5 data transformation and/or cleansing steps to your API 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 you 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 4 is due Sunday, by Midnight of Week 10. 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](#)