

Weeks 1 & 2: What is Data Wrangling?

Syllabus Acknowledgment



After you have read the Syllabus, please click the Syllabus acknowledgment button above. This will take you to a one-question quiz asking if you have completed the task. Once you have selected your response, click save and submit. Upon completion, you will have access to all Week 1 material.

Weeks 1 & 2: What is Data Wrangling?









Welcome to Weeks 1 & 2 of DSC 540 – which is all about data! Where we find data, what we do once we have found data and making it clean and useful. I highly encourage everyone to review the GitHub repositories for each of the books we are using in this course. Data Wrangling with Python is going to be more focused on the types of data we will interact with and Python for Data Analysis is going to spend more time on transforming, cleaning, and preparing data for use. Both books were recently released and are written using Python 3. This course runs in two-week blocks, so try not to get overwhelmed by the readings or the exercises, pace yourself and come up with a plan to accomplish all the deliverables on time. These first two weeks are going to be focused on the basics of Python, what can be done in Python, why data wrangling is so important, as well as prepare you for what we will cover in this course.

We also recently updated the course to use the 3rd edition the the *Python for Data Analysis* textbook. As you can imagine, keeping up with an open source programming language is challenging, and almost as soon as a book is published, there are immediately changes that are made that can't be published in print. Please refer to the links below often when you run into issues in the printed text. While the course still has references to the 2nd edition right now, if you are able to purchase the 3rd edition, I highly recommend it as it will provide better, more efficient code solutions and keep you more current.

Make sure you check out the GitHub & other sites for our textbooks:

- [Data Wrangling_\(Author's GitHub\)](#)
- [Data Wrangling_\(Pakt GitHub\)](#)
- [Python for Data Analysis](#)
- [Python for Data Analysis, 3E \(Open Edition\)](#)

Contents of the Week

-  Overview
-  Helpful Data Science Resources
-  Readings and Tasks
-  Supplemental Materials
-  Term Project Information
-  Introduction Post
-  Weeks 1 & 2 Discussion/Participation
-  Weeks 1 & 2 Exercises

Objectives/Topics




- A** Define the importance of data wrangling in data science
- B** Manipulate data structures available in Python
- C** Compare the different implementations of the inbuilt Python data structures
- D** Compare Python's advanced data structures
- E** Utilize data structures to solve real-world problems
- F** Make use of OS file-handling operations
- G** What is an IDE?
- H** Why should Python be used for data wrangling?

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 1-2 of *Data Wrangling with Python*
 - Chapters 1-3 of *Python for Data Analysis*
-  Read the Term Project Instructions in the Term Project link in the course menu
-  Complete the following:
 - Introduction Post
 - Weeks 1 & 2 Discussion/Participation
 - Weeks 1 & 2 Exercises
 - Project: Milestone 1

Supplemental Readings

Readings 	Videos 
Data Wrangling Versus ETL: What's the Difference? (Zheng, 2017)	
What Exactly is Data Wrangling? (Kumar, 2017)	
The Value is in the Data (Wrangling) . (Haight, 2017)	
Interpreter (Computing) . (Wikipedia, 2019)	
First Steps with Python . (Kearney, n. d.)	

Term Project Information

You will have a term project during this course that has 5 milestones. More details about the

project are provided in the Term Project link in the course menu on the left-hand side of the course page.

Weeks 1 & 2 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 Jupyter notebook? How is it used? Why would you choose this IDE over another IDE?
2. What is data wrangling? Why is it important?
3. What options exist for data wrangling? Why use Python? What are the pros/cons of using Python vs another tool?
4. What is the difference between a list and an array? When would you use either?
5. What are tuples? What are they used for?
6. What is an iterator? How does it relate to a loop?
7. What is a stack and what is it used for?
8. What is a queue? What is it used for?
9. What are basic file operations? Gives examples of some in Python and how they are used.
10. What is structured vs unstructured data? How do you analyze data that isn't structured?
11. What is data profiling? What are the differences between Syntactic and Semantic Profiling?
12. What are the key libraries in Python? What do libraries provide?
13. What are the differences between Python 2 and Python 3? Why do both exist?
14. What is an IDE? What do you use an IDE for?
15. What are some terms/jargon that are used regarding data science/python most individuals aren't familiar with?
16. What does it mean to say Python is an interpreted language? What does it mean to say Python is object oriented?
17. What are functions in Python? What are anonymous (Lambda) functions?
18. What ethical considerations are there for transforming, cleaning and accessing data from various sources?
19. What are the key differences between data wrangling and ETL? Those with a reporting background are probably very familiar with the acronym, ETL (Extract Transform Load), but how does it differ from what we are calling data wrangling and why does that difference matter to data science?

Weeks 1 & 2 Exercises



1. Install the latest versions of either [Docker](#) or [Anaconda](#). Your book Data Wrangling with Python uses Docker, however, you are welcome to use whichever distributor you feel comfortable with.

2. Create a Jupyter notebook where you create a list, iterate over the list and sort your results, generate random numbers, add to the list, and then print your results.

3. Create a line chart with Matplotlib and the following data file.

a. Data file: [world-population.xlsm](#)

b. (Hint: *Python for Data Analysis 2nd Edition*: Page 19-50, *Python for Data Analysis 3rd Edition*: Page 281 & *Data Wrangling with Python*: Preface)

4. Complete the following activities:

a. *Data Wrangling with Python*: Activity 1
page 17

b. *Data Wrangling with Python*: Activity 2
page 31

c. *Data Wrangling with Python*: Activity 3
page 49

d. *Data Wrangling with Python*: Activity 4
page 59

Your exercises are due two weeks from Sunday by Midnight of Week 2. 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 files 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](#)