# Weeks 3 & 4: Understanding Packages

Weeks 3 & 4: Understanding Packages

Welcome to Weeks 3 & 4 – by this time in the program you should be familiar with the concept of packages in Python. They are bundles of standard functions and operations that can be used to accelerate a method or calculation so you don't have to write tons and tons of code to do the same thing. Some of the most popular Python packages are Pandas, NumPy, Matplotlib, etc. and this course is going to give you a nice overview of how to use these packages when you want to analyze and transform your data. After the first two weeks of the course you should feel more comfortable using a tool like Jupyter Notebook, which will be incredibly important as we move forward in this course. If you have questions about how to use Jupyter, please email your instructor, post in Teams, or discuss with your instructor and peers via Teams.

These two weeks will also start to introduce data transformation and data cleansing techniques – which will be covered fully throughout the rest of the course in various weeks. Your first milestone for your project is required for these two weeks – and might be the hardest!

### Contents of the Week

- Overview
- Readings and Tasks
- Supplemental Materials
- Weeks 3 & 4 Discussion/Participation
- Weeks 3 & 4

### Objectives/Topics

- A Create and manipulate one-dimensional and multi-dimensional arrays
- B Create and manipulate pandas DataFrames and series objects
- C Plot and visualize numerical data using the Matplotlib

### Weekly Resources

1. Python Standard Library
2. GitHub
3. Anaconda
4. Jupyter Notebook

Exercises

Project: Milestone 1

library

D Apply matplotlib, NumPy, and pandas to calculate descriptive statistics from a DataFrame/matrix

E Perform subsetting, filtering, and grouping on pandas DataFrames

F Apply Boolean filtering and indexing from a DataFrame to choose specific elements

G Perform JOIN operations in pandas that are analogous to the SQL command

H Identify missing or corrupted data and choose to drop or apply imputation techniques on missing or corrupted data

I Pandas Structures

# Readings and Tasks

Here are your tasks for this week:

Read the following:

- Preface and Chapters 3 & 4 of *Data Wrangling with Python*
- Chapter 5 of *Python for Data Analysis*

Complete the following:

- Weeks 3 & 4 Discussion/Participation
- Weeks 3 & 4 Exercises
- Project: Milestone 1

# Supplemental Materials

**Readings**

**Videos**

Data Wrangling with Pandas. (Raza, 2018)

Wrangling Data with Pandas. (Guo, 2017)

Pandas Cheat Sheet: Data Science and Data Wrangling with Python. (Willems, 2017)

Intro to Data Science Part 1: Numpy and Pandas. (Souterre, 2018)

# Weeks 3 & 4 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 a series? When would you use a series?
2. What is a dataframe? How is it used?
3. What are NumPy arrays? What is a vectorized operation?
4. What are vectors? Why are they important?
5. What are different types of arrays?
6. What is the DOT method and Bracket method? What are the differences between these methods?
7. What are descriptive statistics?
8. Define probability, discrete and continuous distributions. Describe the differences between these distributions and when each can be used.
9. Why is visualizing data important? What are the advantages of using data visualizations?
10. What is subsetting?
11. What is the difference between aggregating, transforming and filtering data?
12. How should outliers/missing values be handled in data preparation?
13. What are the differences between concatenating, joining, and merging data?

## Weeks 3 & 4 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 5, page 116

2. *Data Wrangling with Python:* Activity 6, page 171

3. Create a series and practice basic arithmetic steps

    a. Series 1 = 7.3, -2.5, 3.4, 1.5

        i. Index = ‘a’, ‘c’, ‘d’, ‘e’

    b. Series 2 = -2.1, 3.6, -1.5, 4, 3.1

        i. Index = ‘a’, ‘c’, ‘e’, ‘f’, ‘g’

    c. Add Series 1 and Series 2 together and print the results

    d. Subtract Series 1 from Series 2 and print the results

Your exercises are due two weeks from Sunday by Midnight of Week 4. 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 a 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 1

***Identify Datasets***

The first milestone of this project will be to select the data you want to work with. You will need to select 3 different data sources that have different file types of information – and the data will need to have a relationship between them. If one doesn’t exist, you will

### Submission Instructions

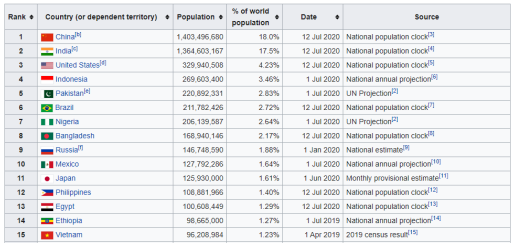
Click the title above to submit your assignment.

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

Term Project Rubric

have to create one. It is likely you will need to manipulate the data to create a relationship. Finding the data, you want to work with for this project, will likely be the hardest part of the project. You must have one of each of the following types of datasets – and you need a minimum of 1000 rows across all datasets. You need a total of 30 columns across the 3 datasets you select.

- CSV/Excel/PDF or another flat file source.
- Website you want to pull data from--you will want to identify a website that has data stored in a table, similar to the screenshot below.



| Rank | Country (or dependent territory) | Population | % of world population | Date | Source |
|---|---|---|---|---|---|
| 1 | China[3] | 1,403,496,680 | 18.0% | 12 Jul 2020 | National population clock[3] |
| 2 | India[4] | 1,364,603,167 | 17.5% | 12 Jul 2020 | National population clock[4] |
| 3 | United States[5] | 329,940,508 | 4.23% | 12 Jul 2020 | National population clock[5] |
| 4 | Indonesia | 269,603,400 | 3.46% | 1 Jul 2020 | National annual projection[6] |
| 5 | Pakistan[6] | 220,892,331 | 2.83% | 1 Jul 2020 | UN Projection[2] |
| 6 | Brazil | 211,782,426 | 2.72% | 12 Jul 2020 | National population clock[7] |
| 7 | Nigeria | 206,139,587 | 2.64% | 1 Jul 2020 | UN Projection[2] |
| 8 | Bangladesh | 168,940,146 | 2.17% | 12 Jul 2020 | National population clock[8] |
| 9 | Russia[7] | 146,748,590 | 1.88% | 1 Jan 2020 | National estimate[9] |
| 10 | Mexico | 127,792,286 | 1.64% | 1 Jul 2020 | National annual projection[10] |
| 11 | Japan | 125,930,000 | 1.61% | 1 Jun 2020 | Monthly provisional estimate[11] |
| 12 | Philippines | 108,881,966 | 1.40% | 12 Jul 2020 | National population clock[12] |
| 13 | Egypt | 100,608,449 | 1.29% | 12 Jul 2020 | National population clock[13] |
| 14 | Ethiopia | 98,665,000 | 1.27% | 1 Jul 2019 | National annual projection[14] |
| 15 | Vietnam | 96,208,984 | 1.23% | 1 Apr 2019 | 2019 census result[15] |

- API you will pull data from.

Some places you can find datasets are listed below:

- Tableau Community
- Kaggle Datasets
- Data.Gov
- Science.Gov
- Data.Gov.UK
- NORC
- European Social Survey
- API List
- PrommableWeb
- Public APIs
- OpenWeatherMap

Wikipedia is a good source to find data that is in a table - and the structure of the HTML is usually very similar.

There are no restrictions on what dataset you use, other than you cannot use the specific datasets used in the book(s).

For this first milestone, you need to submit the following:

- 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)
- The relationships between them, or the relationship you will make between them
- What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words
- Project Subject Area: Describe your project in 1-2 sentences
- Data Sources:
    - Flat File:
        - Description
        - Link or Flat File uploaded
    - API:
        - Description
        - Link
    - Website:
        - Description
        - Link
- Relationships
    - Describe how the data from each source is connected (see example below).
    - If there isn't an obvious relationship, explain how you will make one
- 250 Words describing how you plan to tackle the project, what the data means, the ethical implications of your project scenario/topic, and what challenges you might face.

Submit via a PDF to the assignment link.

Example:

Example:

> In case you are confused what is meant by a relationship between the data sources here is an example (this is a very simple example and I would expect your datasets to have more variables)
>
> CSV File: Contains a list of stores by store ID and other metadata about the stores
>
> Website: Contains a list of store locations, by location ID and store ID and the various departments each store has by department ID.
>
> API: Contains the transactions at each store – contains a transaction ID and store ID.
>
> All 3 of these data sources are related by Store ID. The CSV file has a 1 to many relationship with the Website by StoreID and has a one to many relationship with the API data by StoreID as well.

Milestone 1 is due Sunday, by Midnight of Week 4. Refer to the rubric for more grading detail.