# Weeks 11 & 12: Storing Data and Final Project

## Weeks 11 & 12: Storing Data and Final Project

You made it! Welcome to the last two weeks of this course. For the last two weeks of the course, we are going to focus on storing and querying data – which is an incredibly important part of data science. It doesn't do a whole lot of good to run analysis that is never stored for later use. It also is highly probably that you will interact with a database or data warehouse for some of your data sources. It is also possible you run into cloud sources and you should understand the basic differences between on-premise and cloud data sources. These last two weeks will focus on doing this with Python and introduce how SQL interacts with Python. The final milestone of your project is also due, which is focused on pulling some data together and storing it in a database. Be advised that the course ends on a week from this Saturday (rather than Sunday), so due dates are adjusted accordingly.

It has been a pleasure to have all of you in this course and I hope you have learned a ton about data wrangling and its importance in the data science project process.

### Contents of the Week

- Overview
- Readings and Tasks
- Supplemental Materials
- Weeks 11 & 12 Discussion/Participation
- Weeks 11 & 12 Exercises
- Project: Milestone 5

### Objectives/Topics

- A. Apply the basics of RDBMS to query databases using Python
- B. Convert data from SQL into a pandas DataFrame
- C. Differences between on-premise and cloud data sources

### 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 8 of *Data Wrangling with Python* (chapter 9 is optional reading this week)

Complete the following:
- Weeks 11 & 12 Discussion/Participation
- Weeks 11 & 12 Exercises
- Project: Milestone 5

# Supplemental Readings

| Readings | Videos |
| --- | --- |

Everything a Data Scientist Should Know About Data Management* (*But Was Afraid to Ask). (Wong, 2019)

How to store Data for your Data Science Process. (Ahmad, 2019)

Cloud Analytics vs On Premise Analytics: Which One Should You Choose? (Sagar, 2019)

# Weeks 11 & 12 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. This question was asked in the first two weeks of the course, but now that you have read and learned more about data wrangling, how would you answer this? What ethical considerations are there for transforming, cleaning and accessing data from various sources? What have you learned in this course that surprised you? What are you still struggling with?
2. Describes the roles and responsibilities of a data wrangler. What are the various job titles someone that wants to focus on data cleaning, prepping, transforming, etc. should apply for?
3. Explain the typical day-to-day activities a data wrangler would perform.
4. What does RDBMS mean and what is it used for?
5. What is the importance of databases for data science?
6. What does it mean when a data source is “in the cloud”? How does this impact the ability to access data?
7. How is an RDMS structured?
8. What is NoSQL? What part does it play in the future of data science?
9. What do each of the following acronyms mean? SQL, DDL, DML, DQL, and DCL? What are the pros/cons of SQL?
10. What are the different types of databases? Why is relational the most popular?
11. What are primary keys and what are they used for? Provide an example.
12. What are joins? What are the different kinds of joins? Provide an example of how you would join multiple tables.
13. How do RDBMS and Pandas work together?

## Weeks 11 & 12 Exercises

Complete the following exercise. 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.

- *Data Wrangling with Python*: Activity 11, page 320

Your exercise is due **Saturday** by Midnight of Week 12. 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 you .py file along with screenshots or PDFs of your ouput (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 5

*Merging the Data and Storing in a Database/Visualizing Data*

Now that you have cleaned and transformed your 3 datasets, you need to load them into a database. You can choose what kind of database (SQLLite or MySQL, Postgre SQL are all free options). You will want to load each dataset into SQL Lite as an individual table and then you must join the datasets together in Python into 1 dataset.

Once all the data is merged together in your database, create 5 visualizations that demonstrate the data you have cleansed. You

### Submission Instructions

You must submit the following:

- Jupyter Notebook File or PDF of your code with Milestone # listed.
- Each transformation and visualization should be clearly labeled.
- Human readable dataset after all transformations should be printed at the end of your notebook.
- 250-500 word summary of what you learned and a summary of the ethical implications.

should have at least 2 visualizations that have data from more than one source (meaning, if you have 3 tables, you must have visualizations that span across 2 of the tables – you are also welcome to use your consolidated dataset that you created in the previous step, if you do that, you have met this requirement).

For the visualization portion of the project, you are welcome to use a python library like Matplotlib, Seaborn, or an R package ggPlot2, Plotly, or Tableau/PowerBI.

PowerBI is a free tool that could be used – Tableau only has a free web author. If your use Tableau/PowerBI you need to submit a PDF with your assignment vs the Tableau/PowerBI file. /p>

Clearly label each visualization. Submit your code for merging and storing in the database, with your code for the visualizations along with a 250-500-word summary of what you learned and had to do to complete the project. In your write-up, make sure to address the ethical implications of cleansing data and your project topic. 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.

**Remember** – your GitHub repository can act as a portfolio for potential employers! I would highly suggest using this to submit your work, so you can fill it with good content that demonstrates the projects you are working on!

Milestone 5 is due **Saturday**, by Midnight of Week 12. Refer to the rubric for more grading detail.

Click the title above to submit your assignment.

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

Term Project Rubric