DataCamp
We're hiring!

PAID COURSE

Cleaning Data in Python

Replay Course

4 hours | 17 Videos | 58 Exercises | 8,020 Participants | 4700 XP

STATEMENT OF ACCOMPLISHMENT

Download

This course is part of these tracks:

Data Analyst with Python

Data Scientist with Python

Importing & Cleaning Data with Python

Python Developer

Daniel Chen
Data Science Consultant at Lander Analytics

Daniel is a Software Carpentry instructor and a doctoral student in Genetics, Bioinformatics, and Computational Biology at Virginia Tech, where he works in the Social and Decision Analytics Laboratory under the Biocomplexity Institute. He received his MPH at the Mailman School of Public Health in Epidemiology and is interested in integrating hospital data in order to perform predictive health analytics and build clinical support tools for clinicians. An advocate of open science, he aspires to bridge data science with epidemiology and health care.

See More

COLLABORATOR(S)

Hugo Bowne-Anderson

Yashas Roy

PREREQUISITES

- Intro to Python for Data Science
- Intermediate Python for Data Science

DATASETS

- Air quality
- DOB job application filings
- Ebola
- Gapminder
- Tuberculosis
- Tips
- NYC Uber data

Course Description

A vital component of data science involves acquiring raw data and getting it into a form ready for analysis. In fact, it is commonly said that data scientists spend 80% of their time cleaning and manipulating data, and only 20% of their time actually analyzing it. This course will equip you with all the skills you need to clean your data in Python, from learning how to diagnose your data for problems to dealing with missing values and outliers. At the end of the course, you'll apply all of the techniques you've learned to a case study in which you'll clean a real-world Gapminder dataset!

1 Exploring your data FREE 100%

So you've just got a brand new dataset and are itching to start exploring it. But where do you begin, and how can you be sure your dataset is clean? This chapter will introduce you to the world of data cleaning in Python! You'll learn how to explore your data with an eye for diagnosing issues such as outliers, missing values, and duplicate rows.

▷	Diagnose data for cleaning	50 xp
</>	Loading and viewing your data	50 xp
</>	Further diagnosis	50 xp
▷	Exploratory data analysis	50 xp
☰	Calculating summary statistics	50 xp
</>	Frequency counts for categorical data	100 xp
▷	Visual exploratory data analysis	50 xp
</>	Visualizing single variables with histograms	100 xp
</>	Visualizing multiple variables with boxplots	100 xp
</>	Visualizing multiple variables with scatter plots	100 xp

HIDE CHAPTER DETAILS Completed

2 Tidying data for analysis 100%

Here, you'll learn about the principles of tidy data and more importantly, why you should care about them and how they make subsequent data analysis more efficient. You'll gain first hand experience with reshaping and tidying your data using techniques such as pivoting and melting.

▷	Tidy data	50 xp
☰	Recognizing tidy data	50 xp
</>	Reshaping your data using melt	100 xp
</>	Customizing melted data	100 xp

▷	Pivoting data	50 xp
</>	Pivot data	100 xp
</>	Resetting the index of a DataFrame	100 xp
</>	Pivoting duplicate values	100 xp
▷	Beyond melt and pivot	50 xp
</>	Splitting a column with .str	100 xp
</>	Splitting a column with .split() and .get()	100 xp

HIDE CHAPTER DETAILS Completed

3 Combining data for analysis 100%

The ability to transform and combine your data is a crucial skill in data science, because your data may not always come in one monolithic file or table for you to load. A large dataset may be broken into separate datasets to facilitate easier storage and sharing. Or if you are dealing with time series data, for example, you may have a new dataset for each day. No matter the reason, it is important to be able to combine datasets so you can either clean a single dataset, or clean each dataset separately and then combine them later so you can run your analysis on a single dataset. In this chapter, you'll learn all about combining data.

▷	Concatenating data	50 xp
</>	Combining rows of data	100 xp
</>	Combining columns of data	100 xp
▷	Finding and concatenating data	50 xp
</>	Finding files that match a pattern	100 xp
</>	Iterating and concatenating all matches	100 xp
▷	Merge data	50 xp
</>	1-to-1 data merge	100 xp
</>	Many-to-1 data merge	100 xp

</>	Many-to-many data merge	100 xp

HIDE CHAPTER DETAILS

Completed

4 Cleaning data for analysis

100%

Here, you'll dive into some of the grittier aspects of data cleaning. You'll learn about string manipulation and pattern matching to deal with unstructured data, and then explore techniques to deal with missing or duplicate data. You'll also learn the valuable skill of programmatically checking your data for consistency, which will give you confidence that your code is running correctly and that the results of your analysis are reliable!

▷	Data types	50 xp
</>	Converting data types	100 xp
</>	Working with numeric data	100 xp
▷	Using regular expressions to clean strings	50 xp
</>	String parsing with regular expressions	100 xp
</>	Extracting numerical values from strings	100 xp
</>	Pattern matching	100 xp
▷	Using functions to clean data	50 xp
</>	Custom functions to clean data	100 xp
</>	Lambda functions	100 xp
▷	Duplicate and missing data	50 xp
</>	Dropping duplicate data	100 xp
</>	Filling missing data	100 xp
▷	Testing with asserts	50 xp
</>	Testing your data with asserts	100 xp

HIDE CHAPTER DETAILS Completed

5 **Case study** 100%

In this final chapter, you'll apply all of the data cleaning techniques you've learned in this course towards tidying a real-world, messy dataset obtained from the Gapminder Foundation. Once you're done, not only will you have a clean and tidy dataset, you'll also be ready to start working on your own data science projects using the power of Python!

Putting it all together	**50 xp**
Exploratory analysis	**50 xp**
Visualizing your data	**100 xp**
Thinking about the question at hand	**100 xp**
Assembling your data	**100 xp**
Initial impressions of the data	**50 xp**
Reshaping your data	**100 xp**
Checking the data types	**100 xp**
Looking at country spellings	**100 xp**
More data cleaning and processing	**100 xp**
Wrapping up	**100 xp**
Final thoughts	**50 xp**

HIDE CHAPTER DETAILS Completed

LEARN

Courses

RESOURCES

Community

- Skill Tracks
- Career Tracks
- Pricing

- RDocumentation
- Teach

GROUPS

- For Business
- For Academics

ABOUT

- Company
- Jobs
- Press
- Privacy Policy
- Terms of Use

DataCamp
We're hiring!

DataCamp offers interactive R and Python courses on topics in data science, statistics, and machine learning. Learn from a team of expert teachers in the comfort of your browser with video lessons and fun coding challenges.

LEARN MORE