The challenge of missing or duplicate data

Video: Welcome to week 3
3 min

Video: Methods for handling missing data
8 min

Reading: Data deduplication with Python
20 min

Reading: Follow-along instructions: Work with missing data in a Python notebook
10 min

Lab: Annotated follow-along guide: Work with missing data in a Python notebook
20 min

Video: Work with missing data in a Python notebook
12 min

Video: Remy: A day in the life of a data professional
2 min

Lab: Activity: Address missing data
1h

Lab: Exemplar: Address missing data
20 min

Practice Quiz: Test your knowledge: The challenge of missing or duplicate data
4 questions

The ins and outs of data outliers

Change categorical data to numerical data

Input validation

Review: Clean your data

Follow-along instructions: Work with missing data in a Python notebook

The annotated follow-along notebook linked below is used with the following instructional videos:

1. **Work with missing data in a Python notebook**
2. **Identify and deal with outliers in Python**
3. **Label encoding in Python**
4. **Input validation with Python**

Once you open the notebook, you can skip directly to the code for the relevant video using the links at the top of the notebook.

Accessing and utilizing resources in this section

While watching the videos that follow this reading, you may find it helpful to track the instructor's progress by following along in your own Jupyter notebook. To do so, open the annotated follow-along guide for the videos. The content in this notebook is identical to the content shown in this lesson's instructional videos. In addition to that content, you'll find additional information throughout the notebook. That information is provided to explain the purpose of each concept covered, why the code is written in a certain way, and tips for running the code.

Steps to complete for each video:

1. Read this page of instructions.
2. Open the Annotated follow-along guide: Work with missing data in a Python notebook, which contains a version of the same notebook the instructor will use in the videos.
3. Follow along with the instructor as they go over the code in the notebook.
4. Learn from the instructor and practice running the code in your notebook.

Data dictionary

In the videos shown in this lesson and the following lesson, your notebook will include these datasets:

- **eda_missing_data_dataset1**
- **eda_missing_data_dataset2**
- **eda_outliers_dataset1**
- **eda_outliers_dataset2**
- **eda_outliers_dataset3**
- **eda_label_encoding_dataset**
- **eda_input_validation_joining_dataset1**
- **eda_input_validation_joining_dataset2**

These datasets above represent lightning strike counts in the United States. The data includes data like latitude, longitude, date, city, state, zip code, and lightning strike counts. Each row represents a total lightning strike count on the specified date for a particular location.

Column name	Type	Description
number of strikes	int64	The total count of lightning strikes on a given day
center_point_geom	str	String of characters representing the geographic location based on the lines of latitude and longitude given
longitude	obj	Longitudinal point extracted from center_point_geom
latitude	obj	Latitudinal point extracted from center_point_geom
date	str	The recorded date (format: DD/MM/YYYY)
zip_code	obj	United States postal code of given latitudinal and longitudinal point
city	obj	United States city at given latitudinal and longitudinal point
state	obj	American state where given latitudinal and longitudinal point is located
state_code	obj	Two-letter abbreviation for the given latitudinal and longitudinal point is located

Mark as completed