

The challenge of missing or duplicate data

The ins and outs of data outliers

Change categorical data to numerical data

Input validation

Review: Clean your data

▶

Video: Wrap-up

2 min

📖

Reading: Glossary terms from week 3

10 min

📝

Quiz: Weekly challenge 3

9 questions

Glossary terms from week 3

Terms and definitions from Course 3, Week 3

- Categorical data:** Data that is divided into a limited number of qualitative groups
- Collective outliers:** A group of abnormal points, following similar patterns and isolated from the rest of the population
- Contextual outliers:** Normal data points under certain conditions but become anomalies under most other conditions
- Data ethics:** Well-founded standards of right and wrong that dictate how data is collected, shared, and used
- Data governance:** A process for ensuring the formal management of a company's data assets
- Deduplication:** The elimination or removal of matching data values in a dataset
- Docstring:** (Refer to **documentation string**)
- Documentation string:** A group of text that explains what a method or function does; also referred to as a “docstring”
- Dummy variables:** Variables with values of 0 or 1 that indicate the presence or absence of something
- Global outliers:** Values that are completely different from the overall data group and have no association with any other outliers
- Heatmap:** A type of data visualization that depicts the magnitude of an instance or set of values based on two colors
- Input validation:** The practice of thoroughly analyzing and double-checking to make sure data is complete, error-free, and high quality
- Joining:** The process of augmenting data by adding values from other datasets; one of the six practices of EDA
- Label encoding:** Data transformation technique where each category is assigned a unique number instead of a qualitative value
- Missing data:** A data value that is not stored for a variable in the observation of interest
- Non-null count:** The total number of data entries for a data column that are not blank
- One-hot encoding:** A data transformation technique that turns one categorical variable into several binary variables
- Outliers:** Observations that are an abnormal distance from other values or an overall pattern in a data population

Terms and definitions from previous weeks

- B**
- Bias:** In data structuring, bias refers to organizing data results in groupings, categories, or variables that are misrepresentative of the whole dataset
- Box plot:** A data visualization that depicts the locality, spread, and skew of groups of values within quartiles
- C**
- Cleaning:** The process of removing errors that may distort your data or make it less useful; one of the six practices of Exploratory Data Analysis (EDA)
- CSV file:** A simple text file that can be easy to import or store in other softwares, platforms, and databases
- D**
- Database (DB) file:** A file type used to store data, often in tables, indexes, or fields
- Data source:** The location where data originates
- Data visualization:** A graph, chart, diagram, or dashboard that is created as a representation of information
- Discovering:** The process data professionals use to familiarize themselves with the data so they can start conceptualizing how to use it; one of the six practices of EDA
- E**
- Exploratory data analysis (EDA):** The process of investigating, organizing, and analyzing datasets and summarizing their main characteristics, often by employing data wrangling and visualization methods; the six main practices of EDA are: discovering, structuring, cleaning, joining, validating, and presenting
- Extracting:** The process of retrieving data out of data sources for further data processing or storage
- F**
- Filtering:** The process of selecting a smaller part of a dataset based on specified values and using it for viewing or analysis
- First-party data:** Data that was gathered from inside your own organization
- G**
- Grouping:** The process of aggregating individual observations of a variable into groups
- H**
- Hypothesis:** A theory or an explanation, based on evidence, that is not yet proven true
- I**
- Info():** Gives the total number of entries, along with the data types—called Dtypes in pandas—of the individual entries
- Int64:** A standard integer data type, representing numbers somewhere between negative nine quintillion and positive nine quintillion
- J**
- Joining:** The process of augmenting data by adding values from other datasets; one of the six practices of EDA
- JSON file:** A data storage file that is saved in a JavaScript format
- M**
- Merging:** A method to combine two (or more) different data frames along a specified starting column(s)
- P**
- PACE:** A workflow data professionals can use to remain focused on the end goal of any given dataset; stands for plan, analyze, construct, and execute
- Presenting:** The process of making a cleaned dataset available to others for analysis or further modeling; one of the six practices of EDA
- S**
- Second-party data:** Data that was gathered outside your organization but directly from the original source
- Slicing:** A method for breaking information down into smaller parts to facilitate efficient examination and analysis from different viewpoints
- Sorting:** The process of arranging data into a meaningful order for analysis
- String:** A sequence of characters and punctuation that contains textual information
- Structuring:** The process of taking raw data and organizing or transforming it to be more easily visualized, explained, or modeled; one of the six practices of EDA
- T**
- Third-party data:** Data gathered outside your organization and aggregated
- V**
- Validating:** The process of verifying that the data is consistent and high quality; one of the six practices of EDA

Mark as completed