Glossary terms from module 2

Terms and definitions from Course 3, Module 2

Box plot: A data visualization that depicts the locality, spread, and skew of groups of values within quartiles

CSV file: A simple text file that can be easy to import or store in other softwares, platforms, and databases

Database (DB) file: A file type used to store data, often in tables, indexes, or fields

Data source: The location where data originates

Extracting: The process of retrieving data out of data sources for further data processing or storage

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

Grouping: The process of aggregating individual observations of a variable into groups

Hypothesis: A theory or an explanation, based on evidence, that has not yet been refuted

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

JSON file: A data storage file that is saved in a JavaScript format

Merging: A method to combine two (or more) different data frames along a specified starting column(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

Third-party data: Data gathered outside your organization and aggregated

Terms and definitions from previous modules

B

Bias: In data structuring, bias refers to organizing data results into groupings, categories, or variables that are misrepresentative of the whole dataset

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)

D

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

J

Joining: The process of augmenting data by adding values from other datasets; one of the six practices of EDA

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

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

V

Validating: The process of verifying that the data is consistent and high quality; one of the six practices of EDA