Applied Data Science with Python

Data Wrangling

Learning Objectives

By the end of this lesson, you will be able to:

- Learn about a particular style of coding in Pandas called Pandorable
- Comprehend how to load, index, reindex, and merge data using Pandas
- Evaluate various data optimization techniques
- Prepare the data and then format, normalize, and standardize it using data binning

Business Scenario

A retail store wants to analyze the sales data to improve its revenue. It has collected sales data for the past year, including the products sold, sales channels, and customer demographics.

It wants to identify its top-selling products and sales channels, as well as understand its customer base to tailor its marketing strategies. The store also wants to identify any seasonal trends in its sales data to plan its inventory and marketing campaigns.

It plans to use the Python and Pandas libraries to analyze the sales data. It will use indexing, groupby, and plotting to identify its top-selling products and sales channels. It will also use memory optimization techniques to speed up its data analysis. The retail store will use data binning to identify any seasonal trends in its sales data, and use the insights gained to plan its inventory and marketing campaigns accordingly.

Assisted Practices

Let’s understand the topics below using Jupyter Notebooks.

- 11.2_Pandorable or Idiomatic Pandas Code
- 11.3_Loading, Indexing, and Reindexing
- 11.4_Merging
- 11.5_Memory Optimization in Python
- 11.6_Data Pre-Processing: Data Loading and Dropping Null Values
- 11.7_Data Pre-Processing: Filling Null Values
- 11.8_Data Binning: Formatting and Normalization
- 11.9_Data Binning: Standardization
- 11.10_Describing Data

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

Key Takeaways

- Memory optimization techniques, such as converting data types and manipulating columns, can be used to reduce memory usage.
- DataFrames can be aggregated using the groupby() and describe() methods to get summary statistics.
- Indexing and re-indexing of dataframes can be done using set_index(), reset_index(), and specifying indices during data loading.
- Dataframes can be merged using pd.merge() and specifying merge parameters such as "how" and "on".

Key Takeaways

- Pandorable or Idiomatic Pandas code includes indexing, method chaining, memory optimization, groupby and visualization.
- One can deal with missing values while preprocessing data by reading data, dropping null values, and filling null values.
- Some of the data binning techniques are formatting, normalization, and standardization.

Knowledge Check

Knowledge Check

1

How can we access individual rows in a Pandas dataframe?

A. Using the .iloc method

B. Using the .loc method

C. Using the .reindex method

D. Using the .index method

Knowledge Check

1

How can we access individual rows in a Pandas dataframe?

A. Using the .iloc method

B. Using the .loc method

C. Using the .reindex method

D. Using the .index method

The correct answer is **A**

To access individual rows in a Pandas dataframe use the .iloc method by passing the index for it to give the output.

Knowledge Check

2

Which of the following is an example of normalization technique used in data preprocessing?

A. Dropping null values

B. Log scaling

C. Filling null values

D. Setting a threshold value

Knowledge Check

2

Which of the following is an example of normalization technique used in data preprocessing?

A. Dropping null values

B. Log scaling

C. Filling null values

D. Setting a threshold value

The correct answer is **B**

Log scaling is an example of normalization technique used in data preprocessing.

Knowledge Check

3

How can outliers be detected in a dataset?

A. By calculating the mean value of the dataset

B. By calculating the median value of the dataset

C. By calculating the Z score of each data point in the dataset

D. By calculating the standard deviation of the dataset

Knowledge Check

3

How can outliers be detected in a dataset

A. By calculating the mean value of the dataset

B. By calculating the median value of the dataset

C. By calculating the Z score of each data point in the dataset

D. By calculating the standard deviation of the dataset

The correct answer is **C**

Outliers can be detected in a dataset by calculating the Z score of each data point in the dataset.

Thank You