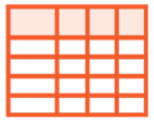
Pandas DataFrame Basics

Mike West
MACHINE LEARNING ENGINEER

Module Overview

- Pandas dataframe overview
- Data types define how data is stored
- Anatomy of the Pandas dataframe
- Data retrieval
- Grouping data in Pandas
- Massaging the Titanic dataset

DataFrame Overview



Pandas dataframes are two-dimensional arrays. All arrays have an index location that's used to find a data point in a dataframe

Pandas and NumPy are two of the core data science libraries used in machine learning

Dataframes are similar in concept to tables in SQL or excel but are much more robust

Real world data is big and NumPy arrays were designed to be fast to accommodate very large datasets

Storing Data Inside Dataframes

Data Types

A data type defines how we store our data in the dataframe

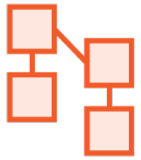
Core Data Types

The int64, the float64, and the object are the main datatypes

The Object

A confusing name for a datatype that stores text

DataFrame Anatomy



Index: The values on the far left-hand side of a dataframe. It can be letters but often you'll see numbers

Columns: The columns are like columns in an excel spreadsheet or a table in a relational database

Data: The stuff inside our dataframe. This is everything inside the table like matrix called the dataframe

The Pandas DataFrame

	UserID	Name	Sex	Age
0	1	Braund, Owen	male	52
1	2	Cumings, John	male	38

Indexing

In Pandas, the process of returning rows and columns is called indexing. This is similar in concept to the SELECT statement in SQL.

The Indexers

iloc

Returns rows and columns by number only

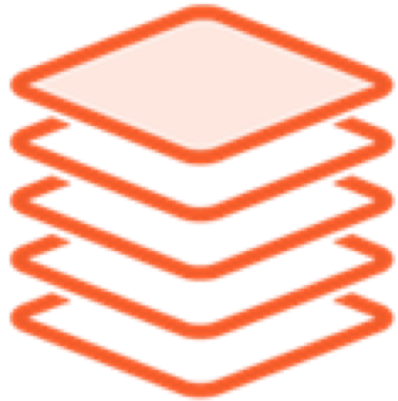
loc

Returns rows and columns by labels or text

Dataset Categorization

Grouping
Slicing rows of your dataframe into distinct groups

SQL Similarity
Dataframe navigation has more functionality than SQL



NumPy
Complexity and speed enhanced due to dataframe architecture

Demo

- Import libraries
- Load the dataframe
- Use loc and iloc on the dataframe
- Sort and group the dataset
- Remove unneeded attribute

Summary

An array is a table like object

Dataframes are built on NumPy Arrays

Data types define how data is stored

Dataframes are indexes, columns, rows

Indexers retrieve data in Pandas

Grouping is fast in Pandas