pandas

Guilan university
Data Science

created by

Athena Hassani & Alireza Sheikhiy

What is Pandas?

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

How to install Pandas

The most straightforward ways to acquire Pandas are through pip or conda and anaconda navigator.

Install using pip:

```
pip install pandas
```

Install using conda:

```
conda install pandas
```

what we are going to talk about...

- Series
- DataFrames
- DataFrames and Series useful functions
- Selecting
- Insertion
- Updating
- Removing
- Sorting
- Missing Data
- Multi Indexing (Advanced Indexing)
- Reshaping and Pivot Tables
- Combining Series and Dataframes
- Duplicate Labels
- Data Mapping
- Grouping
- I/O Tools
- The str accessor and string methods
- Options and Settings

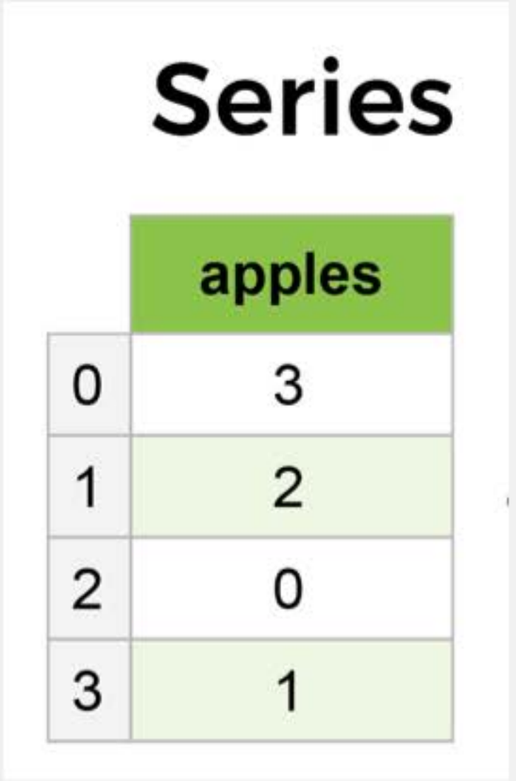
Adidas & Nike products

The main dataset we have is a list of Adidas and Nike sport sneakers.
This dataset contains features below:

- Product Name
- Product ID
- Listing Price
- Sale Price
- Discount
- Brand
- Rating
- Reviews.

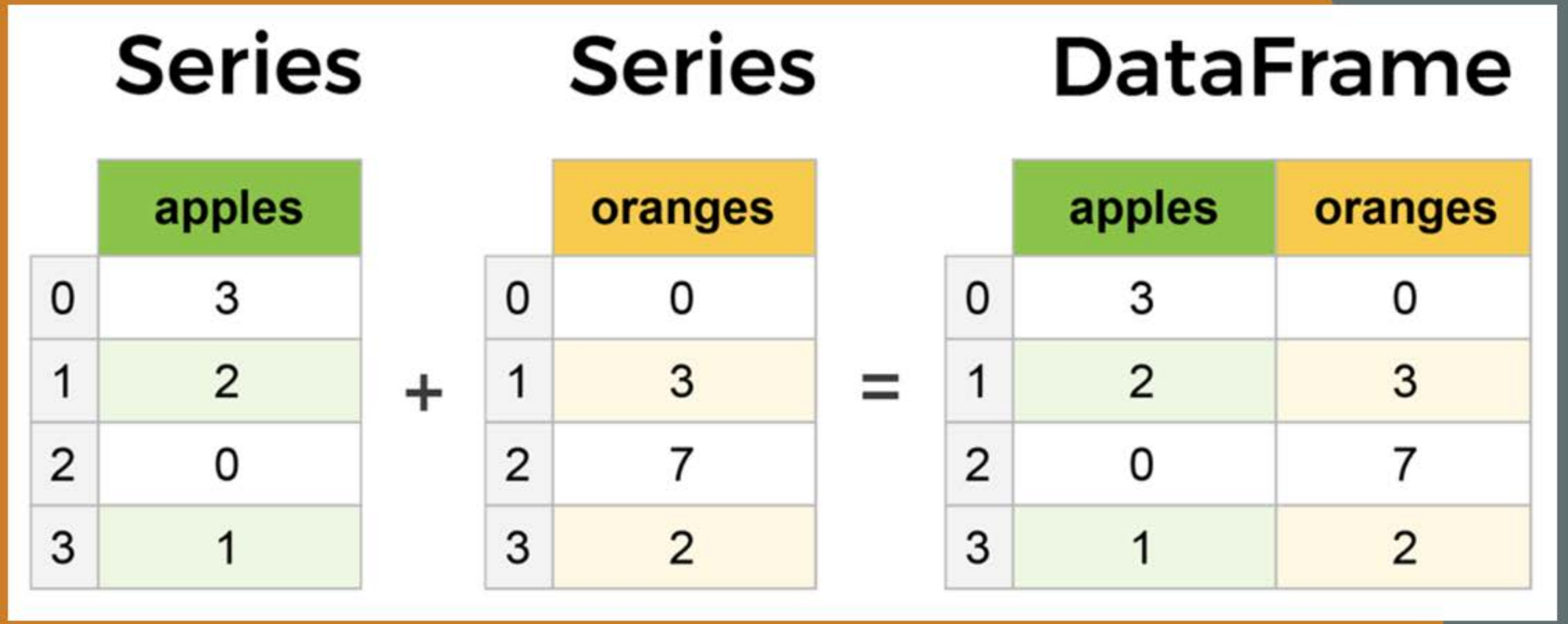
Series

A Pandas Series is like a column in a table. It is a one-dimensional array holding data of any type.



DataFrame

DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object.


Series
Series
DataFrame
apples
oranges
apples
oranges
0 3
1 2
2 0
3 1
+
0 0
1 3
2 7
3 2
=
0 3 0
1 2 3
2 0 7
3 1 2

Selecting

A **select** command is a statement that shows information in datasheet view. We can show or access specific split of series and dataframe objects like cells, rows and columns.

Insertion

Insertion is used to add records or columns to a dataframe.

Updating

Updating is used to update or change a value or values of specific cells, rows or columns.

Removing

Removing is used to remove specific cells, rows or columns or any split of a dataframe or series.

Sorting

Sorting is the process of arranging data into meaningful order so that you can analyze it more effectively.

Missing Data

In Pandas missing data is represented by *NaN*.

While *NaN* is the default missing value marker in pandas.
For reasons of computational speed and convenience, we need to be able to easily detect this value with data of different types.

Multi Indexing

Multi-index allows you to label your records by more than one column in your index. It is a multi-level or hierarchical object for pandas object.

Reshaping & Pivot Tables

Reshape is useful to massage a DataFrame into a format where one or more columns are identifier variables, while all other columns, considered measured variables, are “pivoted” to the row axis, leaving just two non-identifier columns, “variable” and “value”.

Combining Series & Dataframes

The Series and DataFrame objects in pandas are powerful tools for exploring and analyzing data.
Part of their power comes from a multifaceted approach to combining separate datasets.
With pandas, you can merge, join, and concatenate your dataframes, allowing you to unify and better understand your data as you analyze it.

Duplicate Labels

Index objects are not required to be unique; you can have duplicate row or column labels

Data Mapping

The pandas *map()* function from Series is used to substitute each value in a Series with another value, that may be derived from a function, a dict or a Series . Since DataFrame columns are series, you can use *map()* to update the column and assign it back to the DataFrame.

Grouping

The *groupby()* is a powerful method in pandas that allows you to group data based on a single column or more.
You can apply many operations to a groupby object, including aggregation methods like *sum()*, *mean()*, and *count()*.

I/O Tools

pandas IO Tools is the API that allows you to save the contents of Series and DataFrame objects to the clipboard, objects, or files of various types. It also enables loading data from the clipboard, objects, or files.

The str accessor & string methods

Accessors in Pandas provide functions specific to a particular data type.

The *str* accessor is the one for string operations.

Options and Settings

The options and settings in Pandas are used to manipulate the data present in the DataFrames. To deal with unreadable data and to customize the various aspects like behavior, data, analysis, etc.

References

The pandas documents in Pydata website

https://pandas.pydata.org/docs

pandas

Guilan university
Data Science

created by

Athena Hassani & Alireza Sheikhiy

What is Pandas?

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

How to install Pandas

The most straightforward ways to acquire Pandas are through pip or conda and anaconda navigator.

Install using pip:

```
pip install pandas
```

Install using conda:

```
conda install pandas
```

what we are going to talk about...

- Series
- DataFrames
- DataFrames and Series useful functions
- Selecting
- Insertion
- Updating
- Removing
- Sorting
- Missing Data
- Multi Indexing (Advanced Indexing)
- Reshaping and Pivot Tables
- Combining Series and Dataframes
- Duplicate Labels
- Data Mapping
- Grouping
- I/O Tools
- The str accessor and string methods
- Options and Settings

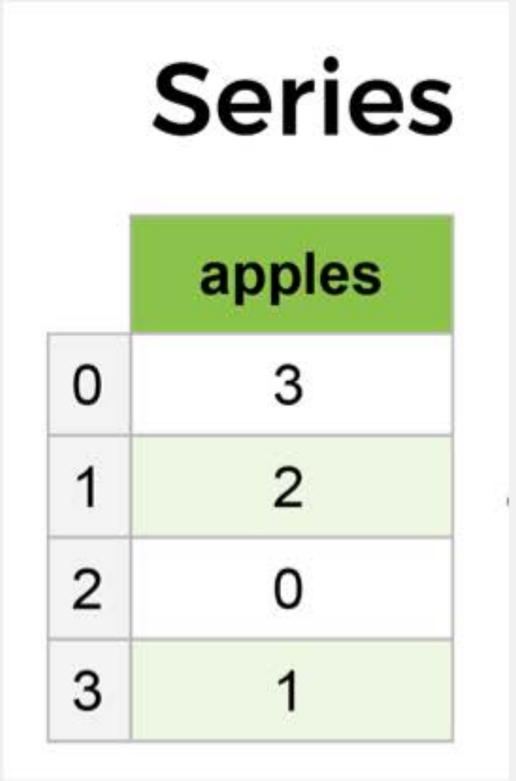
Adidas & Nike products

The main dataset we have is a list of Adidas and Nike sport sneakers.
This dataset contains features below:

- Product Name
- Product ID
- Listing Price
- Sale Price
- Discount
- Brand
- Rating
- Reviews.

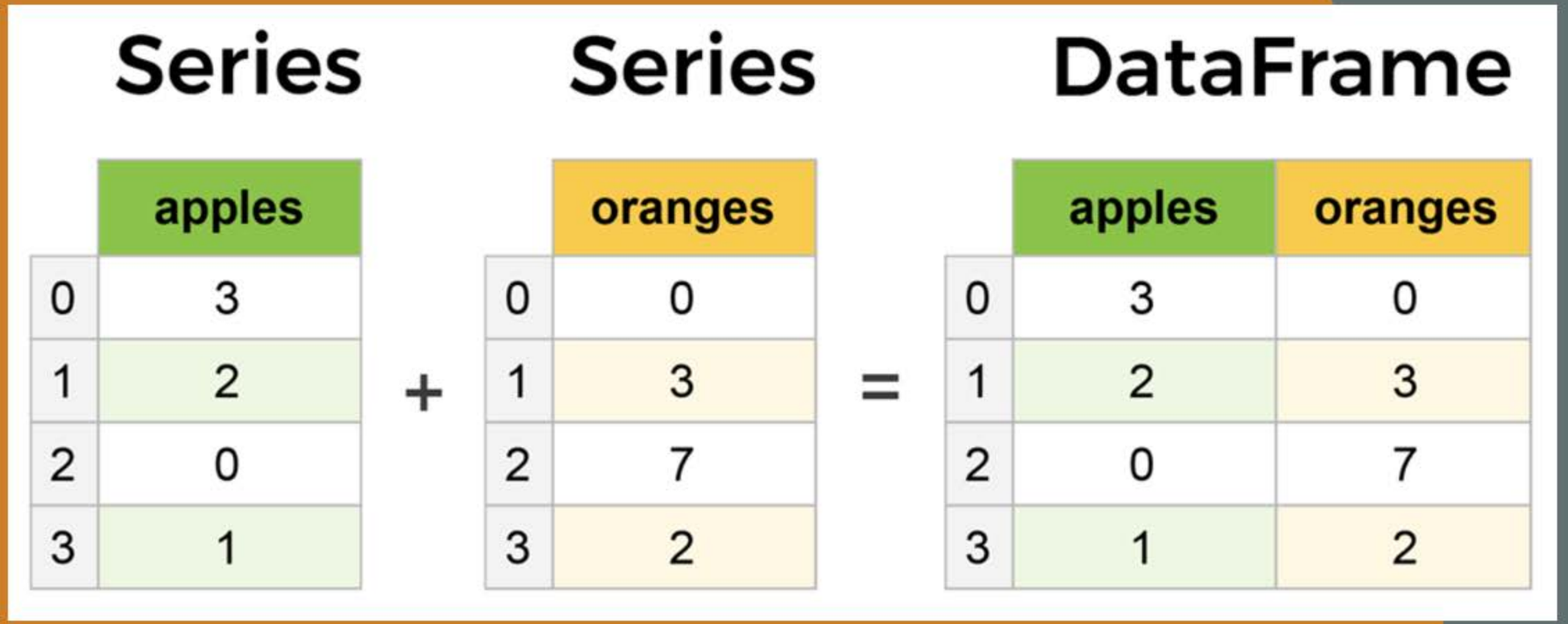
Series

A Pandas Series is like a column in a table. It is a one-dimensional array holding data of any type.



DataFrame

DataFrame is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object.


Series
Series
DataFrame
apples
oranges
apples
oranges
0 3
1 2
2 0
3 1
+
0 0
1 3
2 7
3 2
=
0 3 0
1 2 3
2 0 7
3 1 2

Selecting

A **select** command is a statement that shows information in datasheet view. We can show or access specific split of series and dataframe objects like cells, rows and columns.

Insertion

Insertion is used to add records or columns to a dataframe.

Updating

Updating is used to update or change a value or values of specific cells, rows or columns.

Removing

Removing is used to remove specific cells, rows or columns or any split of a dataframe or series.

Sorting

Sorting is the process of arranging data into meaningful order so that you can analyze it more effectively.

Missing Data

In Pandas missing data is represented by *NaN*.

While *NaN* is the default missing value marker in pandas.
For reasons of computational speed and convenience, we need to be able to easily detect this value with data of different types.

Multi Indexing

Multi-index allows you to label your records by more than one column in your index. It is a multi-level or hierarchical object for pandas object.

Reshaping & Pivot Tables

Reshape is useful to massage a DataFrame into a format where one or more columns are identifier variables, while all other columns, considered measured variables, are “pivoted” to the row axis, leaving just two non-identifier columns, “variable” and “value”.

Combining Series & Dataframes

The Series and DataFrame objects in pandas are powerful tools for exploring and analyzing data.
Part of their power comes from a multifaceted approach to combining separate datasets.
With pandas, you can merge, join, and concatenate your dataframes, allowing you to unify and better understand your data as you analyze it.

Duplicate Labels

Index objects are not required to be unique; you can have duplicate row or column labels

Data Mapping

The pandas *map()* function from Series is used to substitute each value in a Series with another value, that may be derived from a function, a dict or a Series . Since DataFrame columns are series, you can use *map()* to update the column and assign it back to the DataFrame.

Grouping

The *groupby()* is a powerful method in pandas that allows you to group data based on a single column or more.
You can apply many operations to a groupby object, including aggregation methods like *sum()*, *mean()*, and *count()*.

I/O Tools

pandas IO Tools is the API that allows you to save the contents of Series and DataFrame objects to the clipboard, objects, or files of various types. It also enables loading data from the clipboard, objects, or files.

The str accessor & string methods

Accessors in Pandas provide functions specific to a particular data type.

The *str* accessor is the one for string operations.

Options and Settings

The options and settings in Pandas are used to manipulate the data present in the DataFrames. To deal with unreadable data and to customize the various aspects like behavior, data, analysis, etc.

References

The pandas documents in Pydata website

https://pandas.pydata.org/docs