INTRODUCTION TO PANDAS DATAFRAMES

365 DataScience

Plan for the next few lectures:

- Introduce the pandas DataFrame object and discuss its purpose
- Create several DataFrames from scratch
- Relate what you have already learnt to what's written in the pandas documentation

365 DataScience

The Pandas DataFrame : Introduction

Series

DataFrame

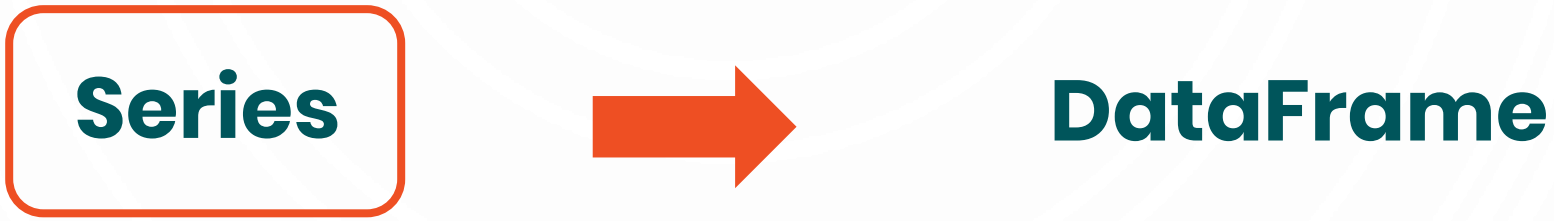
365 DataScience

The Pandas DataFrame : Introduction

Series **DataFrame**

365 DataScience

The Pandas DataFrame : Introduction



365 DataScience

The Pandas DataFrame : Introduction

Series **DataFrame**

365 DataScience

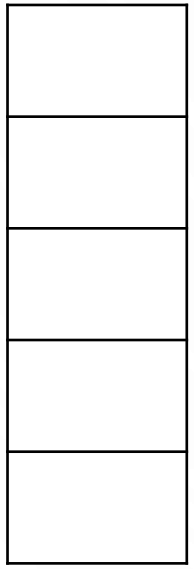
The Pandas DataFrame : Introduction

Series

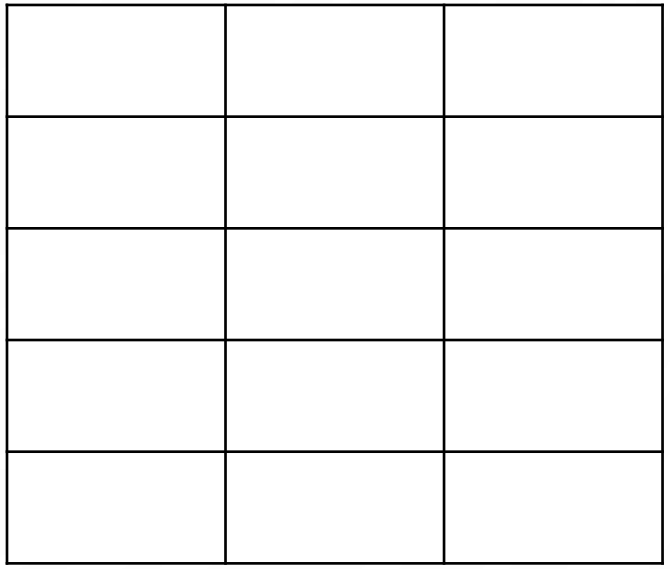
DataFrame

365 DataScience

The Pandas DataFrame : Introduction



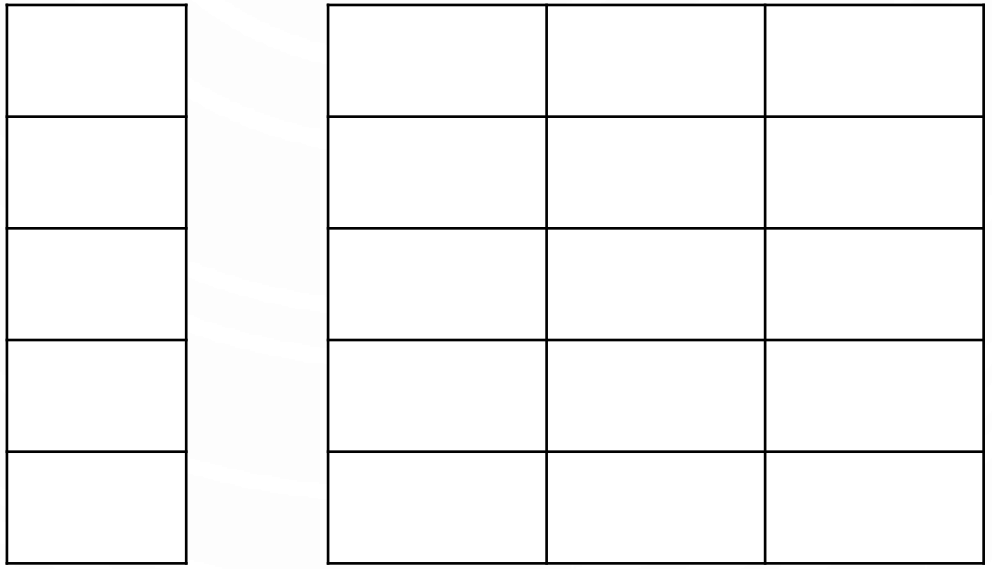
Series



DataFrame

365 DataScience

The Pandas DataFrame : Introduction



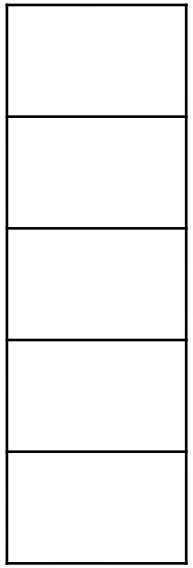
Series DataFrame

365 DataScience

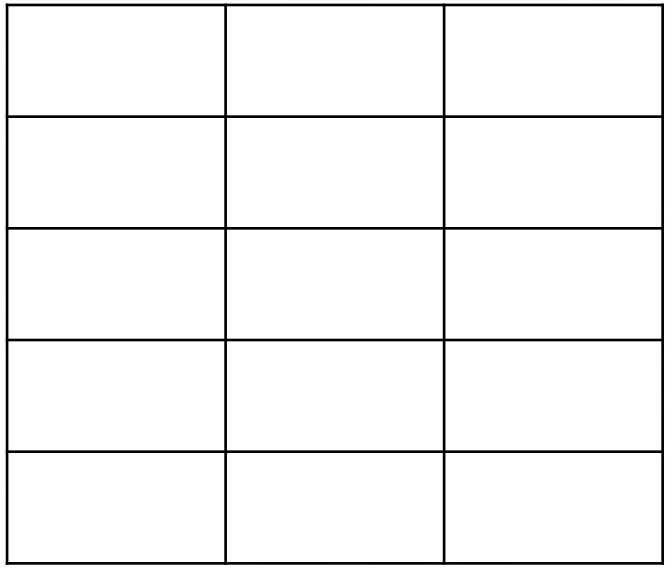
The Pandas DataFrame : Introduction

Single column data

- corresponds to a **single variable**
- information of a **single type**
- we can preserve **data consistency**



Series



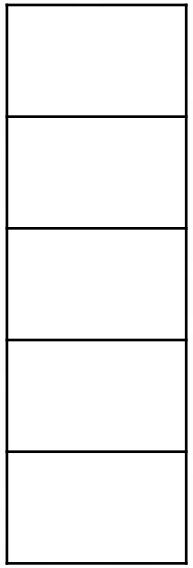
DataFrame

Multi-column data

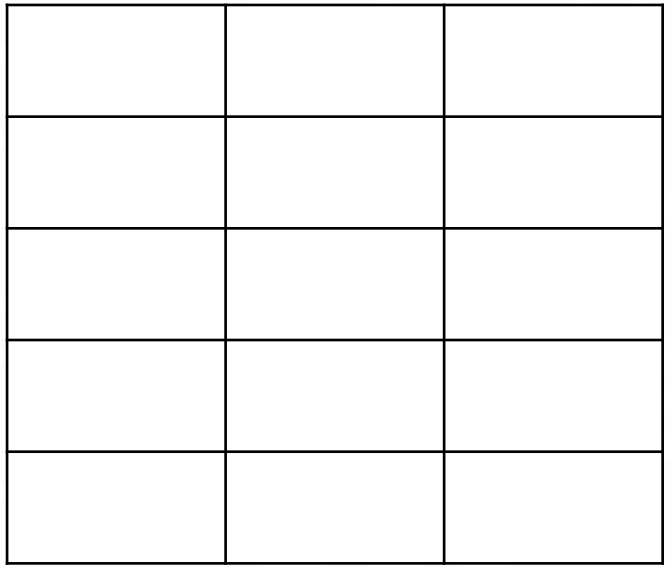
- each column represents a **different** variable
- every column contains data of **its own type**
- the information can potentially **be heterogeneous**
- we can preserve **data consistency**

(we aim for having information of the same type **within** a certain column)

365 DataScience

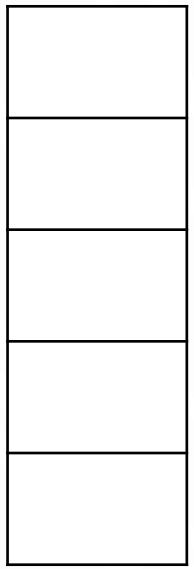


Series

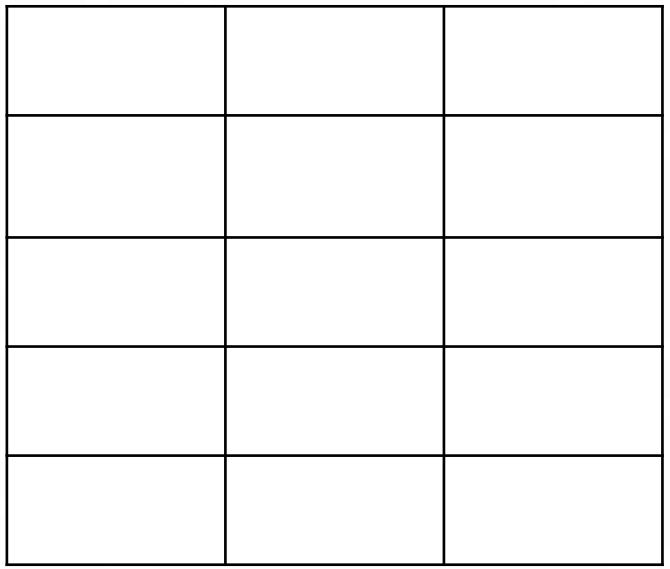


DataFrame

365 DataScience

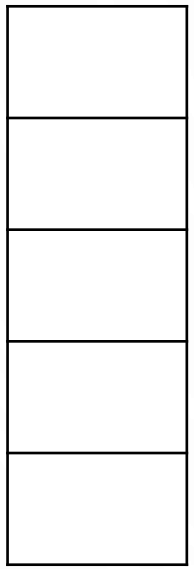


Series

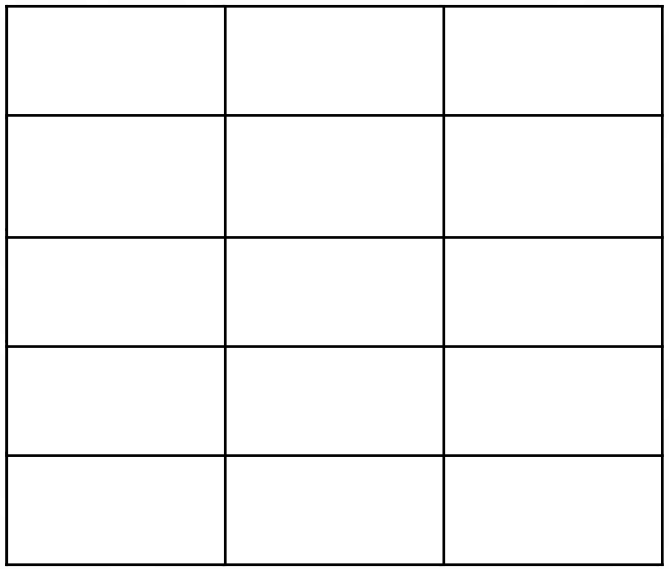


DataFrame

365 DataScience

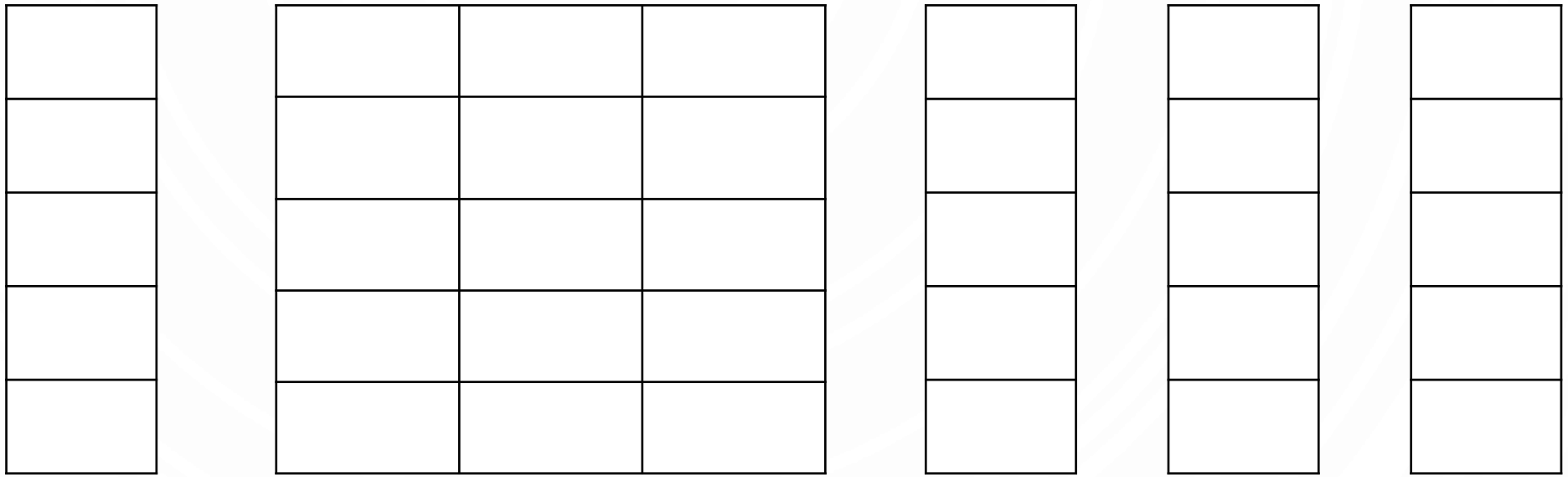


Series



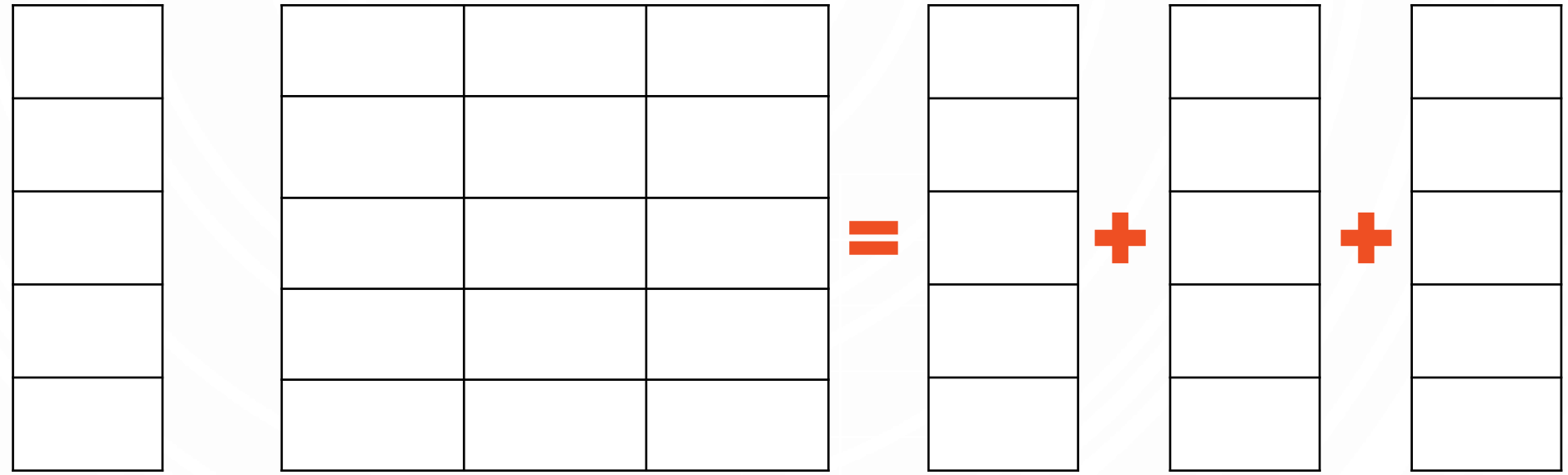
DataFrame

365 DataScience



Series DataFrame

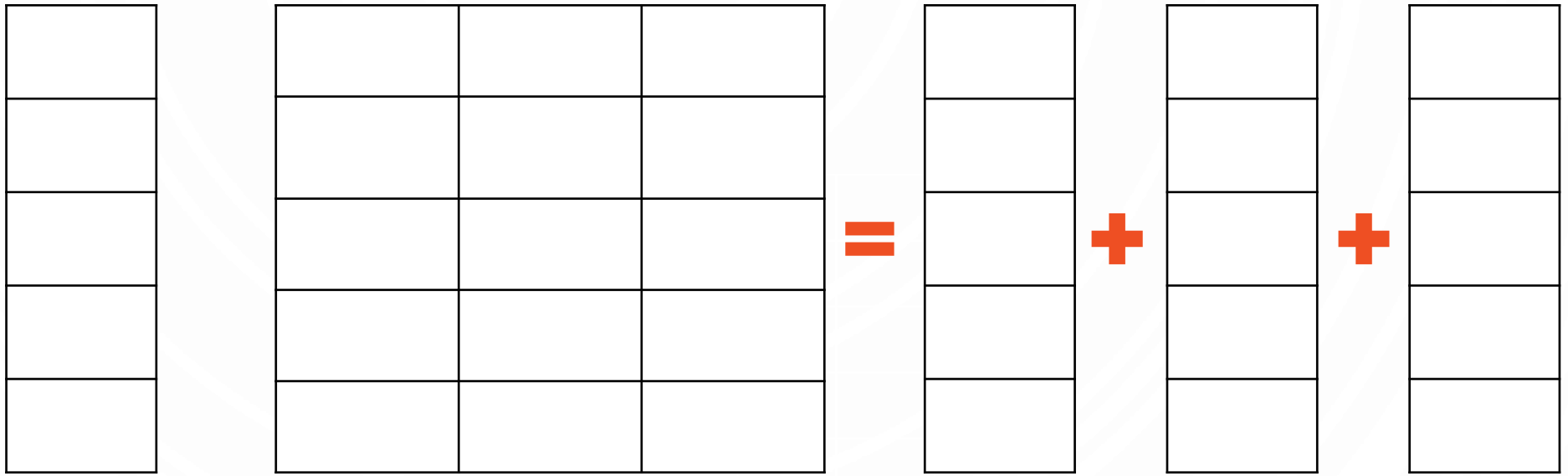
365 DataScience



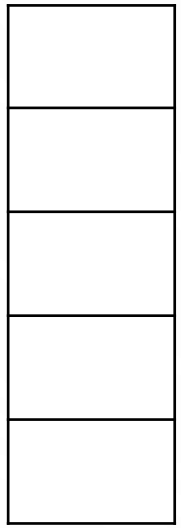
Series DataFrame

365 DataScience

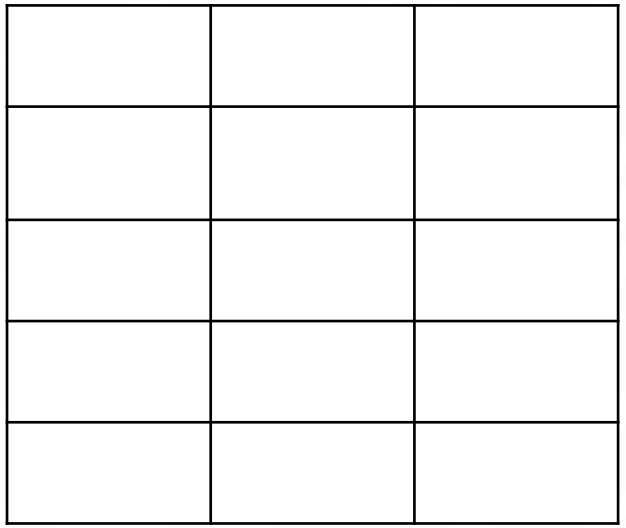
Any characteristic of a Series, is also applicable to the separate columns of a DataFrame.



365 DataScience



Series

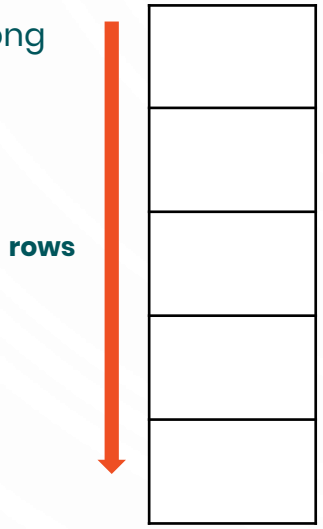


DataFrame

365 DataScience

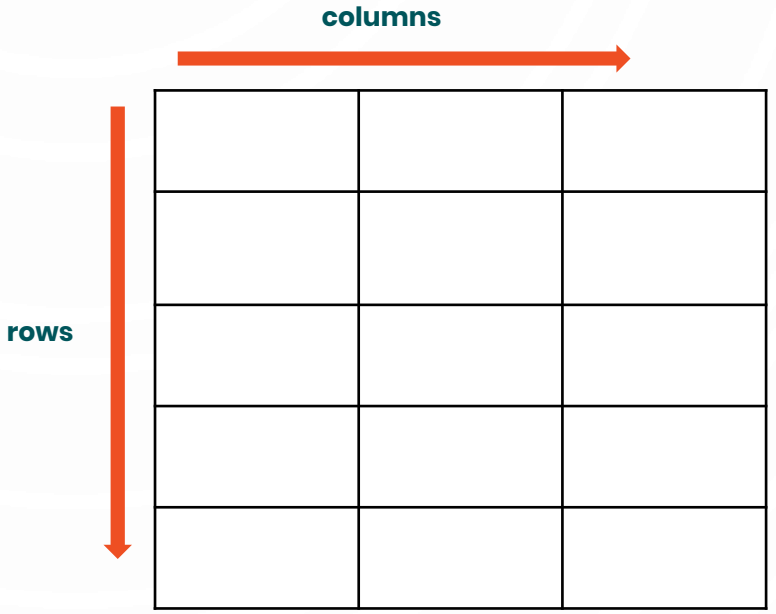
One-dimensional data structure

- contains values along **a single** axis (rows)



Series

Two-dimensional data structure

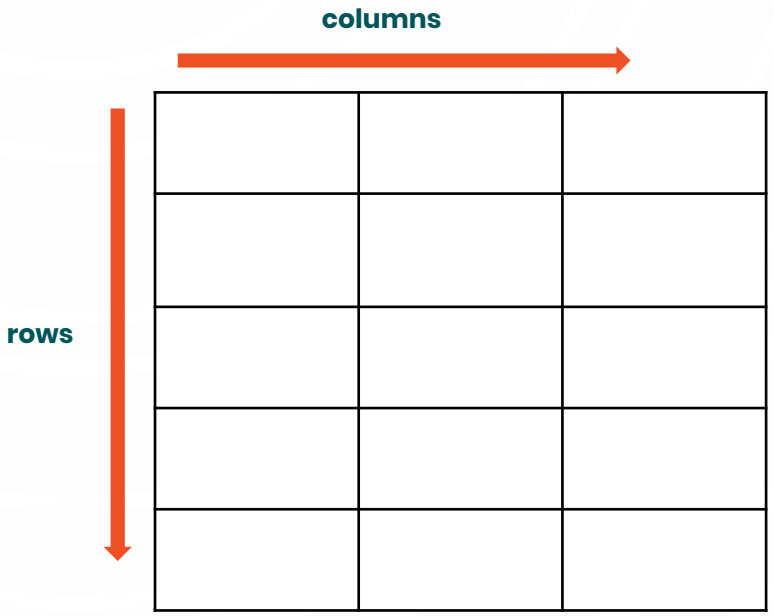


DataFrame

- contains values along **two** axes (rows and columns)
- represents a **tabular** structure
- it is the closest Python analogue to a standard two-dimensional dataset, or a spreadsheet
- can have both row and column **labels**

365 DataScience

Two-dimensional
data structure

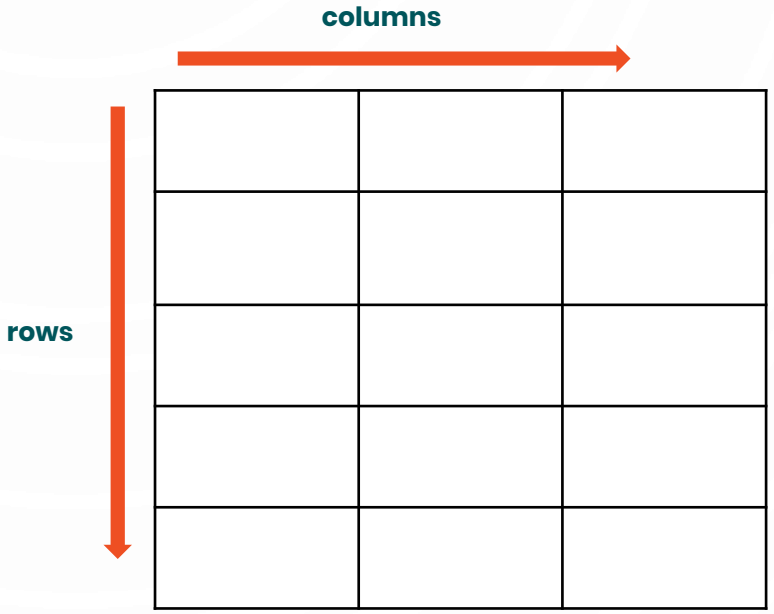


- contains values along **two** axes (rows and columns)
- represents a **tabular** structure
- it is the closest Python analogue to a standard two-dimensional dataset, or a spreadsheet
- can have both row and column **labels**

DataFrame

365 DataScience

2-D Matrix = **Two-dimensional** NumPy array = **Two-dimensional** data structure



DataFrame

- contains values along **two** axes (rows and columns)
- represents a **tabular** structure
- it is the closest Python analogue to a standard two-dimensional dataset, or a spreadsheet
- can have both row and column **labels**

365 DataScience

2-D Matrix = **Two-dimensional** NumPy array = **Two-dimensional** data structure

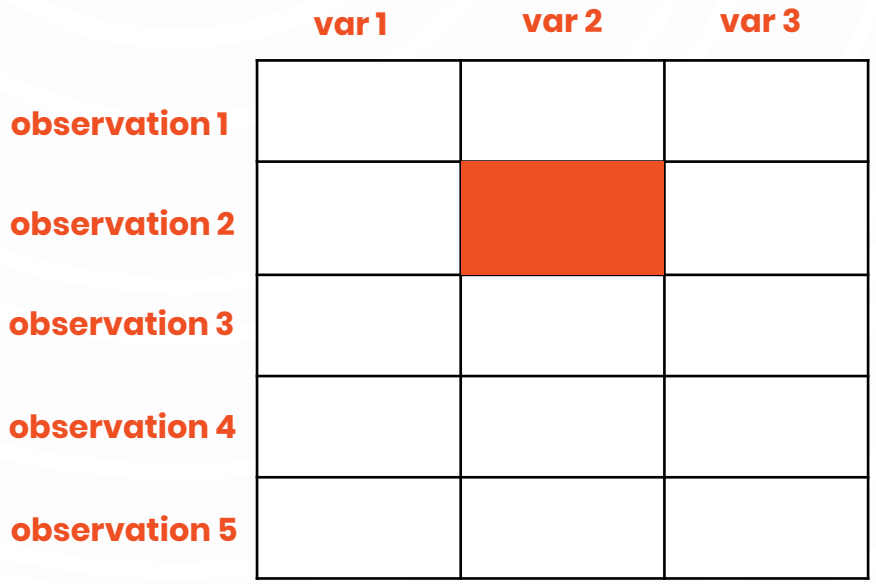
	var 1	var 2	var 3
observation 1			
observation 2			
observation 3			
observation 4			
observation 5			

DataFrame

- contains values along **two** axes (rows and columns)
- represents a **tabular** structure
- it is the closest Python analogue to a standard two-dimensional dataset, or a spreadsheet
- can have both row and column **labels**

365 DataScience

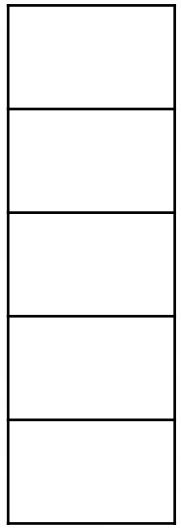
2-D Matrix = **Two-dimensional** NumPy array = **Two-dimensional** data structure



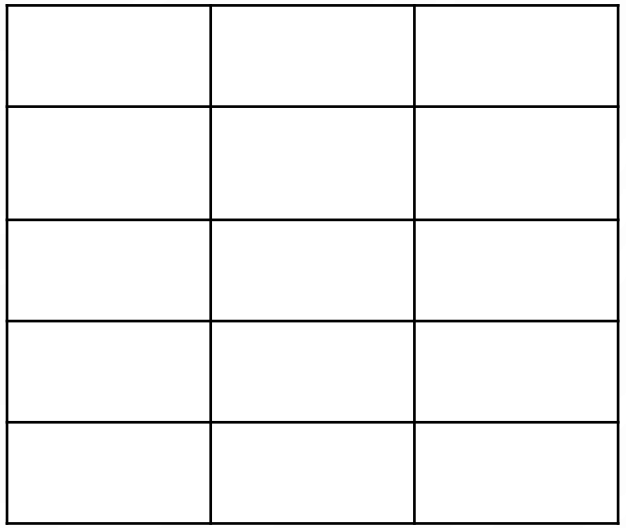
DataFrame

- contains values along **two** axes (rows and columns)
- represents a **tabular** structure
- it is the closest Python analogue to a standard two-dimensional dataset, or a spreadsheet
- can have both row and column **labels**

365 DataScience



Series

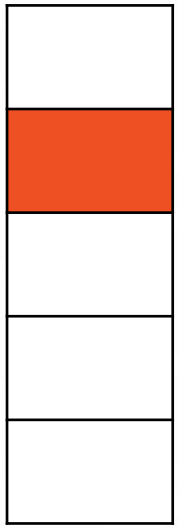


DataFrame

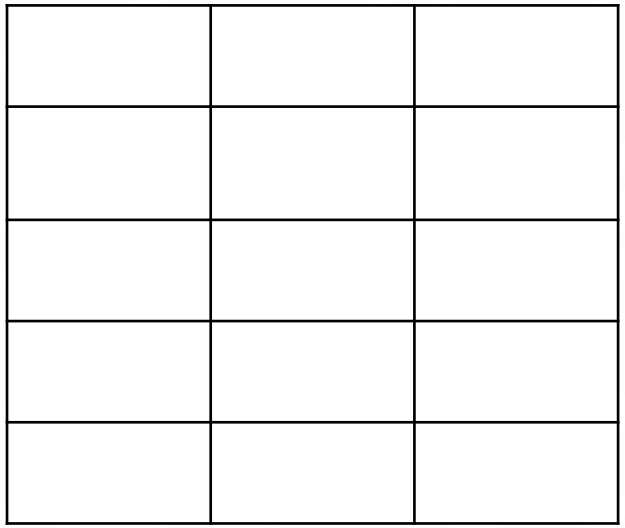
365 DataScience

Single point
of reference

row index



Series

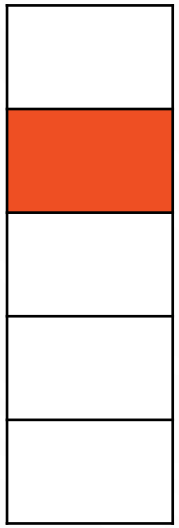


DataFrame

365 DataScience

Single point
of reference

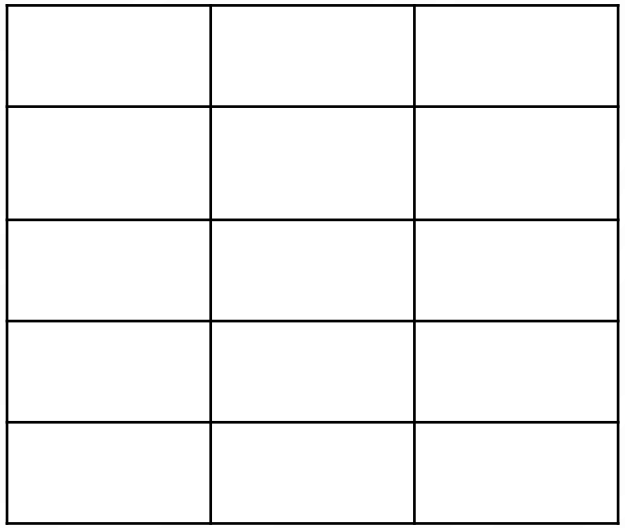
row index



Series

Two points
of reference

row index

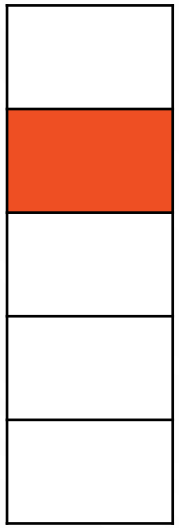


DataFrame

365 DataScience

Single point
of reference

row index

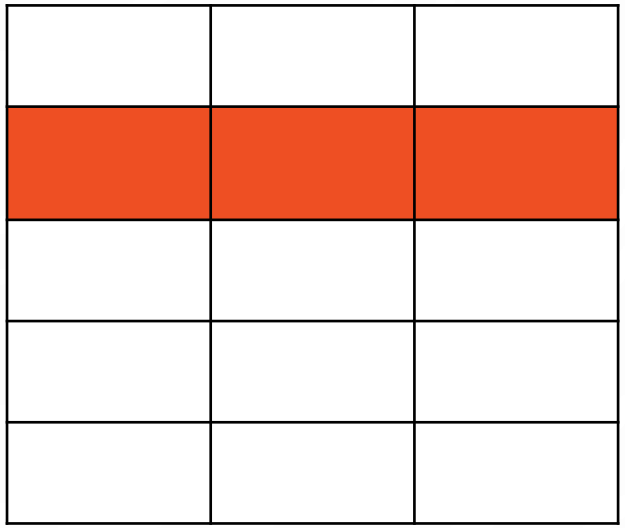


Series

Two points
of reference

column index

row index

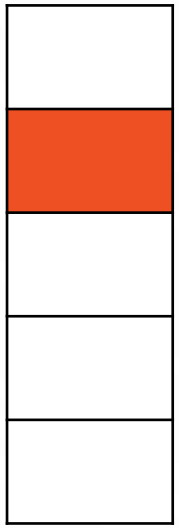


DataFrame

365 DataScience

Single point
of reference

row index

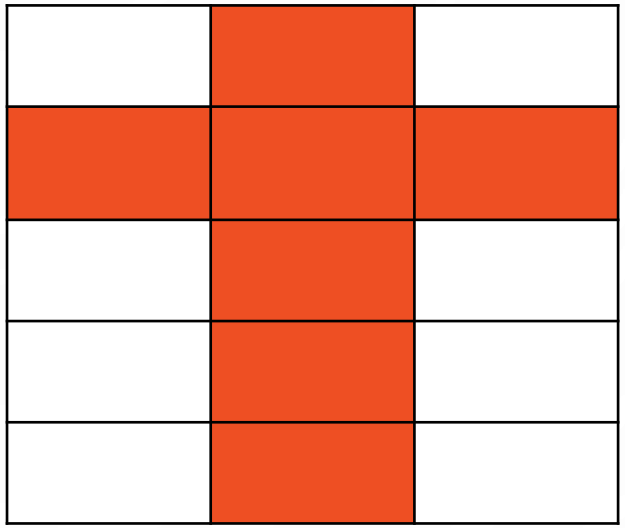


Series

Two points
of reference

column index

row index

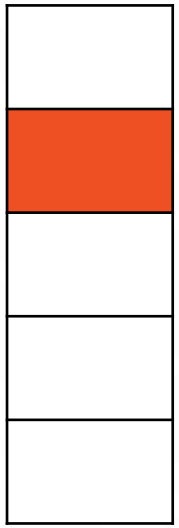


DataFrame

365 DataScience

Single point
of reference

row index

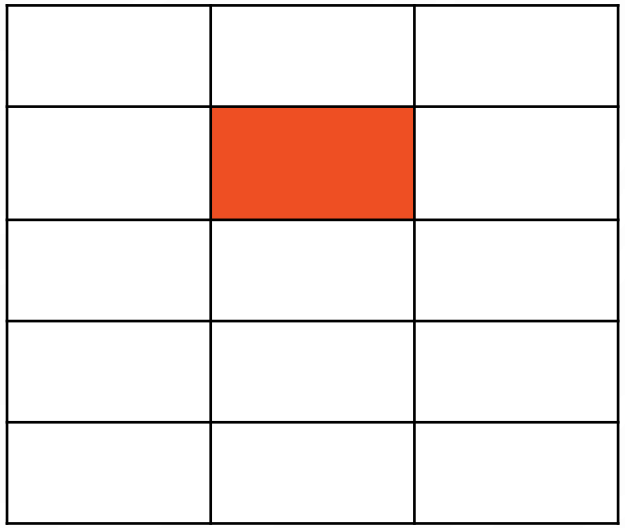


Series

Two points
of reference

column index

row index

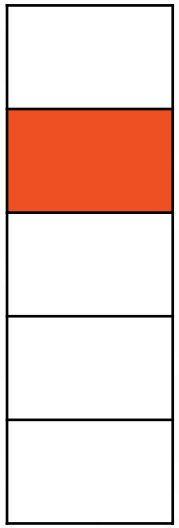


DataFrame

365 DataScience

Single point
of reference

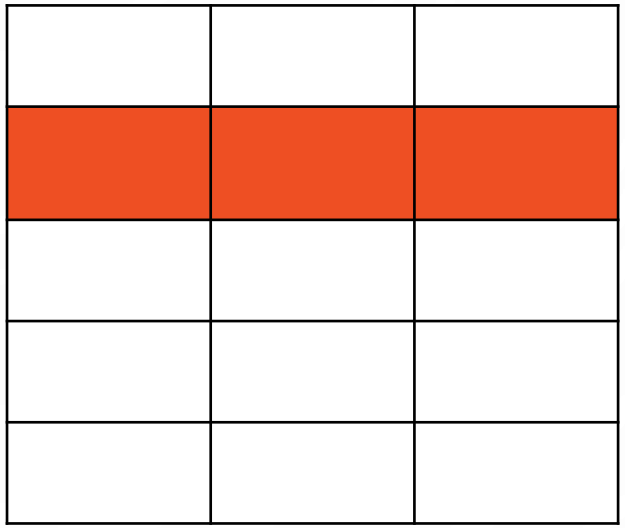
row index



Series

Single point
of reference

row index

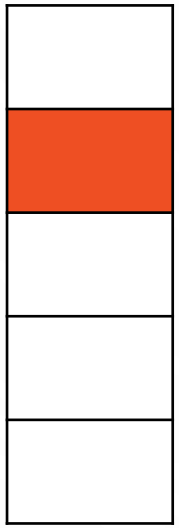


DataFrame

365 DataScience

Single point
of reference

row index

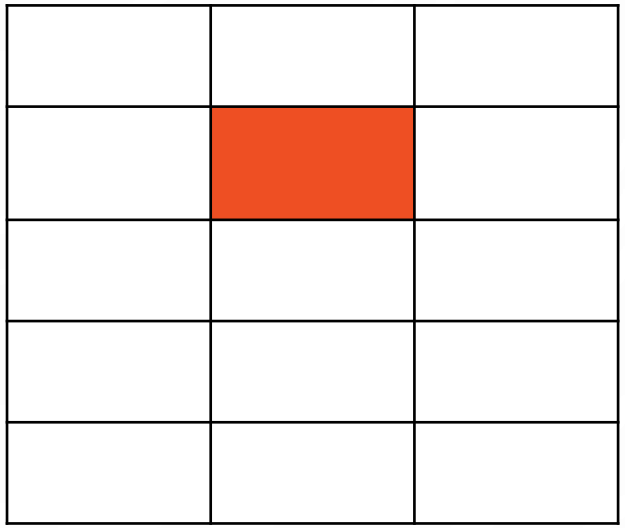


Series

Two points
of reference

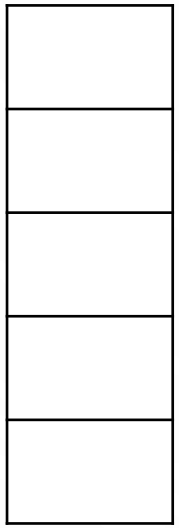
column index

row index

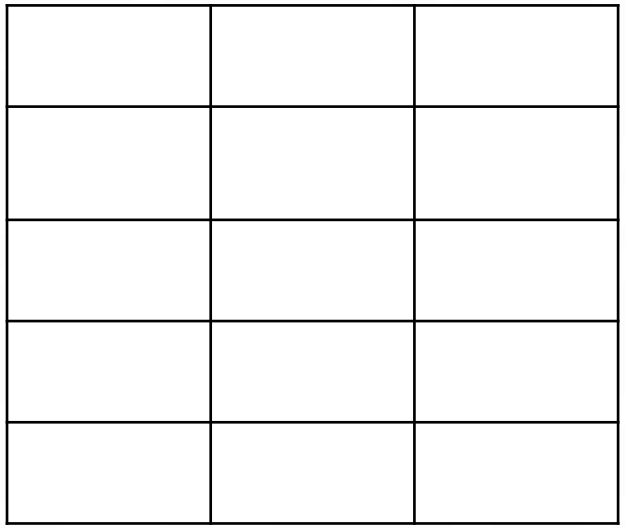


DataFrame

365 DataScience



Series



DataFrame

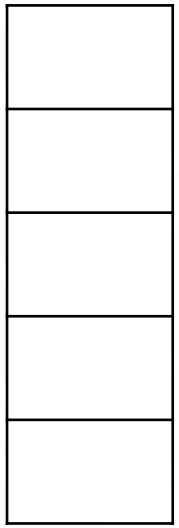
365 DataScience

Series and DataFrames as Programming Objects

A Powerful Version of the Python List

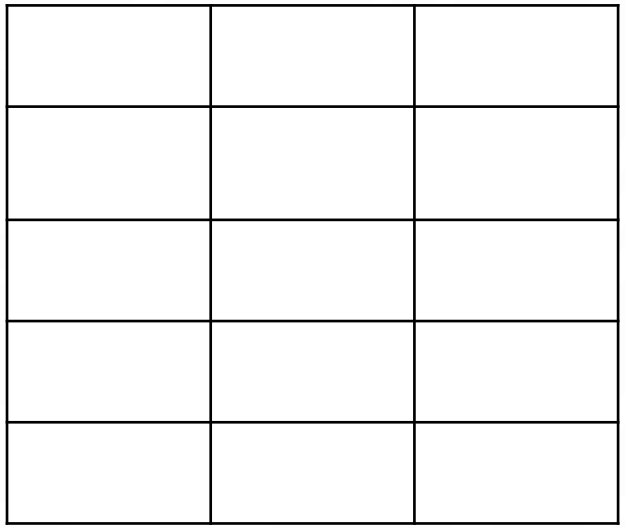
+ Python Dictionary Features

- Its indexing relates to **the keys of a dictionary**
- It also allows us to extract the desired parts of the given dataset more quickly and efficiently



Series

An Enhanced Python Dictionary



DataFrame

- We can provide a whole object that contains the values of an entire column **to the dictionary keys**
- Inherits the characteristics of the Dictionary class
- Plus a lot more features and functionalities

365 DataScience