Introduction to Pandas

SHARP SIGHT

What you'll learn

- What Pandas is
- What DataFrames are
- How we can use Pandas to manipulate and shape DataFrames
  - i.e., looking ahead at what's to come

What is Pandas?

What is pandas?

- Python package that provides data science tools

- Pandas provides crucial data science data structures
  - DataFrame
  - series

- Also, tools for shaping, processing, and analyzing data
  - aggregation (AKA, grouping)
  - reshaping
  - variable processing

Importing pandas into a program

- Use the import statement to import Pandas into a program
  - This statement is required in order to use pandas

- The alias **`pd`** is standard
  - most programmers use this alias

```
import pandas as pd
```

A quick introduction to DataFrames

Pandas DataFrame

- DataFrames are a special data structure for storing data

- DataFrames have row-and-column structure
  - This structure is one of the key reasons to use DataFrames
  - Most data you work with will have row/column structure

- Pandas provides tools to create DataFrames

- Also, Pandas gives you tools to *manipulate* DataFrames

DataFrames are data organized in a row-and column structure

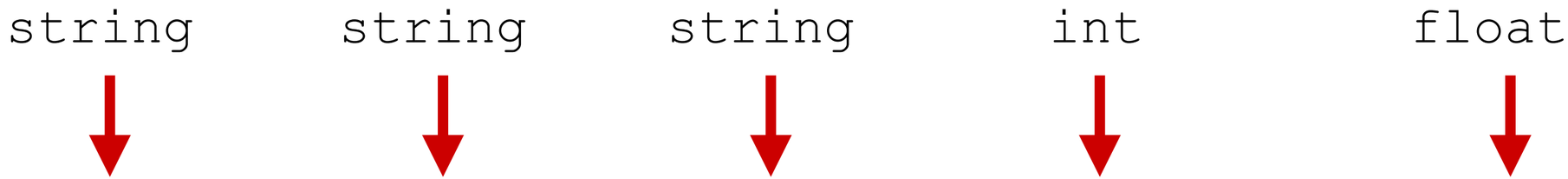
Columns

Rows

country_code	country	continent	population	gdp
USA	United States	North America	323,405,935	18,624,475,000,000
CHN	China	Asia	1,378,665,000	11,190,992,550,230
JPN	Japan	Asia	126,994,511	4,949,273,341,994
DEU	Germany	Europe	82,348,669	3,477,796,274,497
GBR	United Kingdom	Europe	65,595,565	2,650,850,178,102
FRA	France	Europe	66,859,768	2,465,134,297,439

* they are very similar to Excel spreadsheets

In a data frame, each column can be of a different data type



country_code	country	continent	population	gdp
USA	United States	North America	323,405,935	18,624,475,000,000
CHN	China	Asia	1,378,665,000	11,190,992,550,230
JPN	Japan	Asia	126,994,511	4,949,273,341,994
DEU	Germany	Europe	82,348,669	3,477,796,274,497
GBR	United Kingdom	Europe	65,595,565	2,650,850,178,102
FRA	France	Europe	66,859,768	2,465,134,297,439

... So data frames can contain combinations of numerics, booleans, strings, etc

Each row and column of a dataframe has a numeric index

	0	1	2	3	4
	country_code	**country**	**continent**	**population**	**gdp**
0	**USA**	United States	North America	323,405,935	18,624,475,000,000
1	**CHN**	China	Asia	1,378,665,000	11,190,992,550,230
2	**JPN**	Japan	Asia	126,994,511	4,949,273,341,994
3	**DEU**	Germany	Europe	82,348,669	3,477,796,274,497
4	**GBR**	United Kingdom	Europe	65,595,565	2,650,850,178,102
5	**FRA**	France	Europe	66,859,768	2,465,134,297,439

We can use these indexes to take "slices" and subset our data

Rows and columns can also have "labels"

The column labels are the column names

country_code	country	continent	population	gdp
USA	United States	North America	323,405,935	18,624,475,000,000
CHN	China	Asia	1,378,665,000	11,190,992,550,230
JPN	Japan	Asia	126,994,511	4,949,273,341,994
DEU	Germany	Europe	82,348,669	3,477,796,274,497
GBR	United Kingdom	Europe	65,595,565	2,650,850,178,102
FRA	France	Europe	66,859,768	2,465,134,297,439

We can also assign labels to the rows, by assigning a column as the index, etc

We can perform data manipulations on data frames

- Selecting data
  - AKA, slices
- Calculating summary statistics
- Creating new rows and columns
- Visualizing data in the data frame

Recap

Recap of what we learned

- Dataframes are structures that hold data
  - row and column data, like an excel spreadsheet

- Rows and columns have numeric indexes
  - we can use these indexes to "slice" the data and perform data manipulation

- Rows and columns can also have "labels"
  - also used in subsetting and data manipulation

- Dataframes are very flexible and useful
  - you'll learn a lot more about them in future lessons