Welcome to the course!

INTRODUCTION TO IMPORTING DATA IN PYTHON

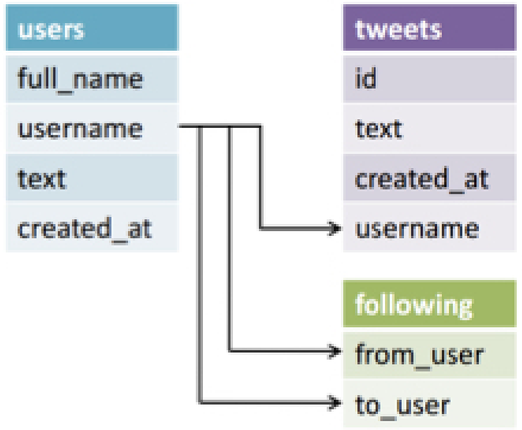
Hugo Bowne-Anderson
Data Scientist at DataCamp

Import data

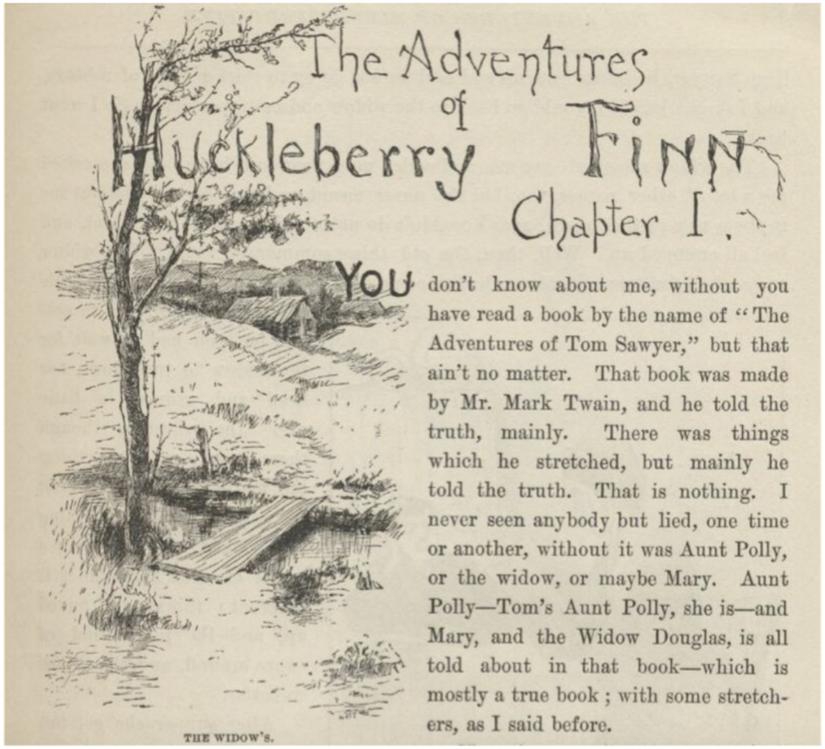
- Flat files, e.g. .txts, .csvs
- Files from other software

Import data

- Flat files, e.g. .txts, .csvs
- Files from other software
- Relational databases



Plain text files



The Adventures of Huckleberry Finn

Chapter I.

YOU don't know about me, without you have read a book by the name of "The Adventures of Tom Sawyer," but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied, one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly—Tom's Aunt Polly, she is—and Mary, and the Widow Douglas, is all told about in that book—which is mostly a true book ; with some stretchers, as I said before.

THE WIDOW'S.

Source: Project Gutenberg

Table data

titanic.csv

	Name	Sex	Cabin	Survived
	Braund, Mr. Owen Harris	male	NaN	0
	Cumings, Mrs. John Bradley	female	C85	1
	Heikkinen, Miss. Laina	female	NaN	1
	Futrelle, Mrs. Jacques Heath	female	C123	1
	Allen, Mr. William Henry	male	NaN	0

[1] Source: Kaggle

Table data

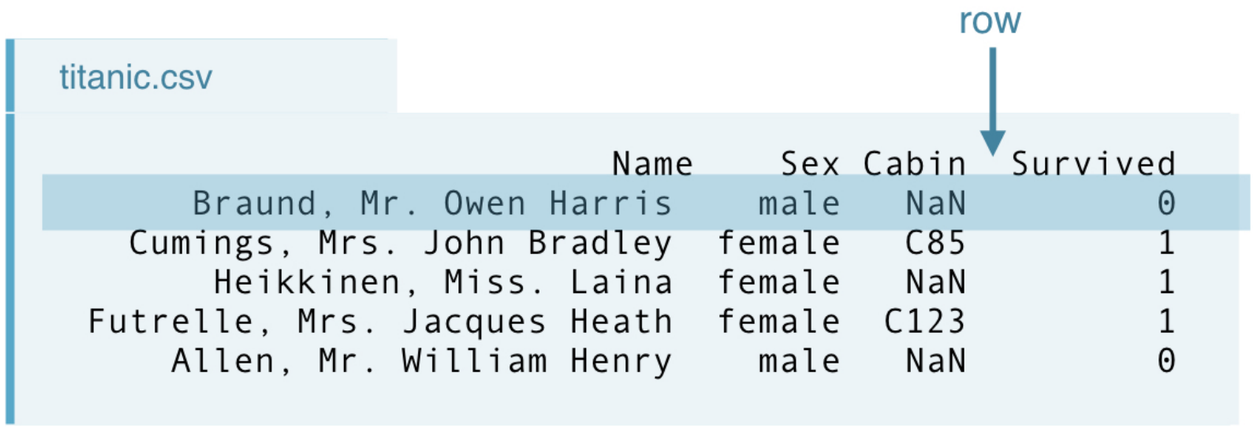
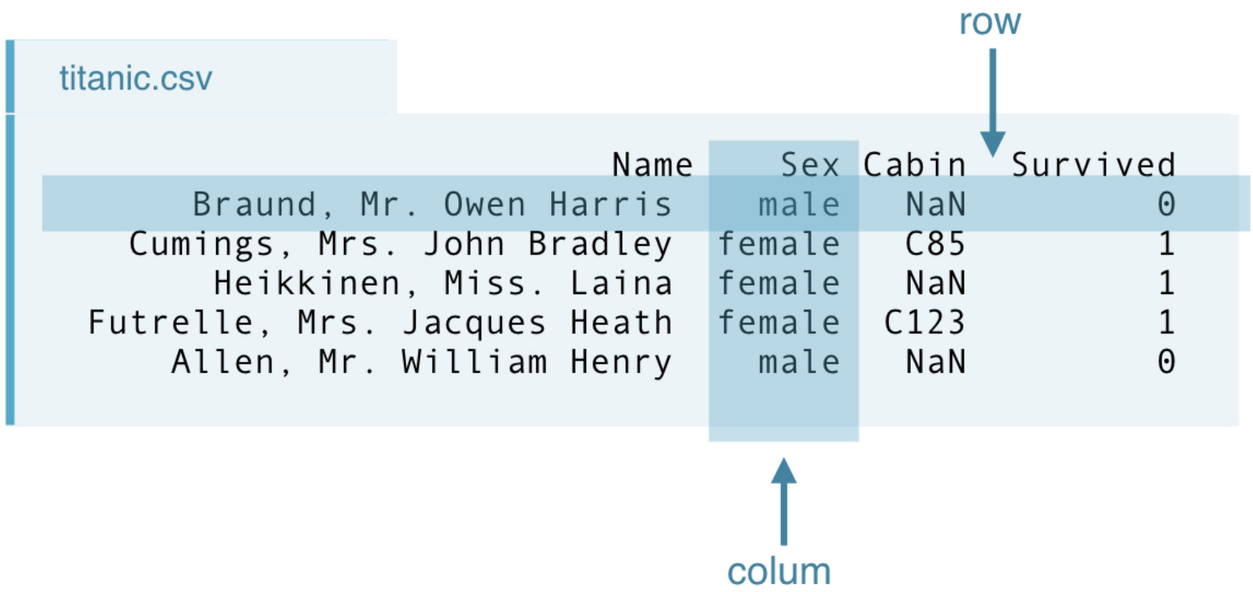

row
titanic.csv
Name Sex Cabin Survived
Braund, Mr. Owen Harris male NaN 0
Cumings, Mrs. John Bradley female C85 1
Heikkinen, Miss. Laina female NaN 1
Futrelle, Mrs. Jacques Heath female C123 1
Allen, Mr. William Henry male NaN 0

Table data



- Flat file

Reading a text file

```
filename = 'huck_finn.txt'

file = open(filename, mode='r')  # 'r' is to read

text = file.read()

file.close()
```

Printing a text file

```
print(text)
```

```
YOU don't know about me without you have read a book by
the name of The Adventures of Tom Sawyer; but that
ain't no matter. That book was made by Mr. Mark Twain,
and he told the truth, mainly. There was things which
he stretched, but mainly he told the truth.  That is
nothing. never seen anybody but lied one time or
another, without it was Aunt Polly, or the widow, or
maybe Mary. Aunt Polly--Tom's Aunt Polly, she is--and
Mary, and the Widow Douglas is all told about in that
book, which is mostly a true book, with some
stretchers, as I said before.
```

Writing to a file

```
filename = 'huck_finn.txt'
file = open(filename, mode='w')  # 'w' is to write
file.close()
```

Context manager with

```
with open('huck_finn.txt', 'r') as file:
    print(file.read())
```

```
YOU don't know about me without you have read a book by
the name of The Adventures of Tom Sawyer; but that
ain't no matter. That book was made by Mr. Mark Twain,
and he told the truth, mainly. There was things which
he stretched, but mainly he told the truth.  That is
nothing. never seen anybody but lied one time or
another, without it was Aunt Polly, or the widow, or
maybe Mary. Aunt Polly--Tom's Aunt Polly, she is--and
Mary, and the Widow Douglas is all told about in that
book, which is mostly a true book, with some
stretchers, as I said before.
```

In the exercises, you'll:

- Print files to the console
- Print specific lines
- Discuss flat files

Let's practice!

INTRODUCTION TO IMPORTING DATA IN PYTHON

The importance of flat files in data science

INTRODUCTION TO IMPORTING DATA IN PYTHON

Hugo Bowne-Anderson
Data Scientist at DataCamp

Flat files

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2.
3101282,7.925,,S
```

Flat files

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2.
3101282,7.925,,S
```

Name	Gender	Cabin	Survived
Braund, Mr. Owen Harris	male	NaN	0
Cumings, Mrs. John Bradley	female	C85	1
Heikkinen, Miss. Laina	female	NaN	1
Futrelle, Mrs. Jacques Heath	female	C123	1
Allen, Mr. William Henry	male	NaN	0

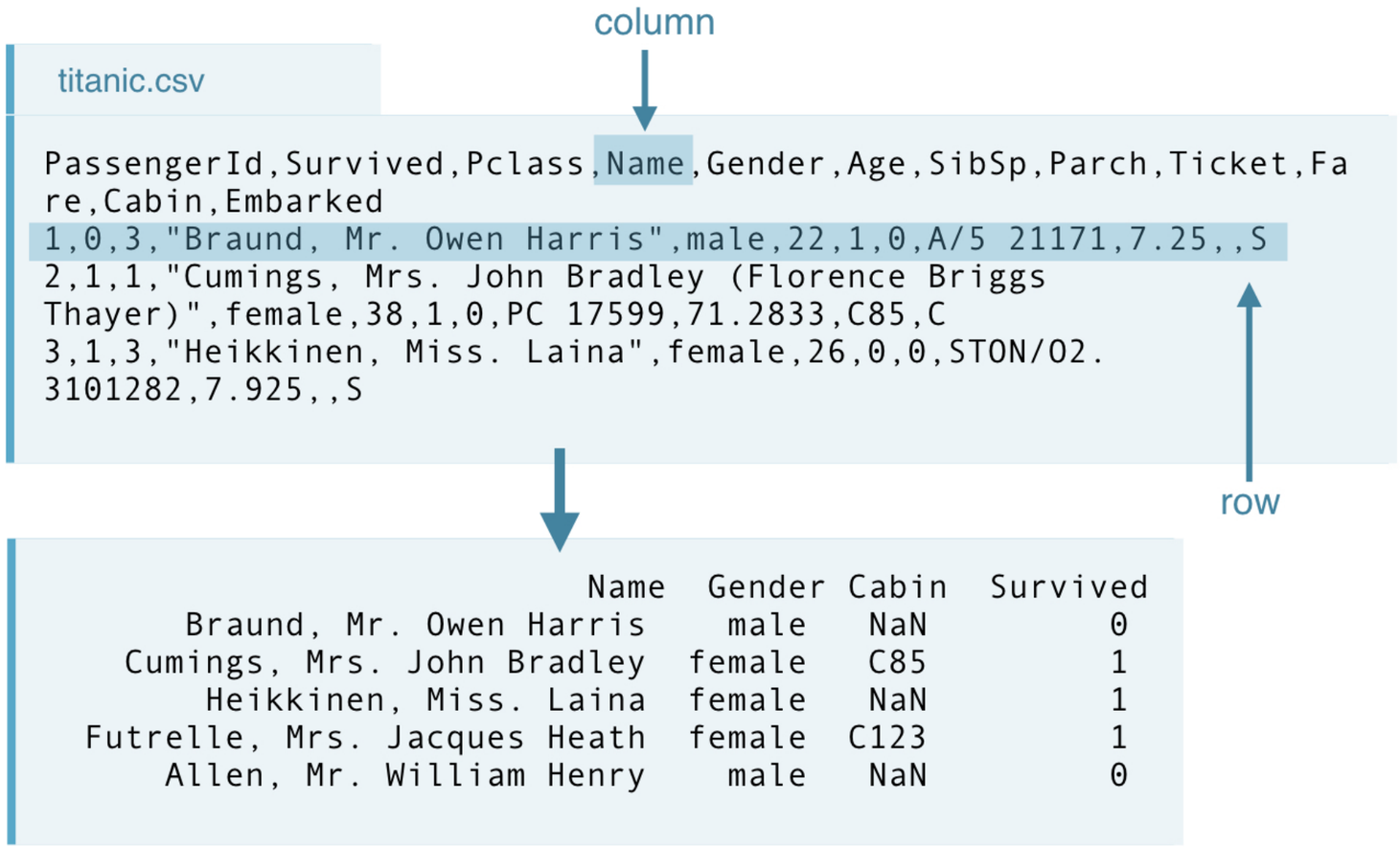
Flat files

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2.
3101282,7.925,,S
```

Name	Gender	Cabin	Survived
Braund, Mr. Owen Harris	male	NaN	0
Cumings, Mrs. John Bradley	female	C85	1
Heikkinen, Miss. Laina	female	NaN	1
Futrelle, Mrs. Jacques Heath	female	C123	1
Allen, Mr. William Henry	male	NaN	0

Flat files


column
titanic.csv
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2.
3101282,7.925,,S
row
Name Gender Cabin Survived
Braund, Mr. Owen Harris male NaN 0
Cumings, Mrs. John Bradley female C85 1
Heikkinen, Miss. Laina female NaN 1
Futrelle, Mrs. Jacques Heath female C123 1
Allen, Mr. William Henry male NaN 0

Flat files

- Text files containing records
- That is, table data
- Record: row of fields or attributes

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
```

Flat files

- Text files containing records
- That is, table data
- Record: row of fields or attributes
- Column: feature or attribute

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
```

Flat files

- Text files containing records
- That is, table data
- Record: row of fields or attributes
- Column: feature or attribute

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fa
re,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs
Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
```

Header

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC
17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,
35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,
27,0,2,347742,11.1333,,S
```

Header

titanic.csv

```
PassengerId,Survived,Pclass,Name,Gender,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC
17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,
35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,
27,0,2,347742,11.1333,,S
```

File extension

- .csv - Comma separated values
- .txt - Text file
- commas, tabs - Delimiters

Tab-delimited file

MNIST.txt

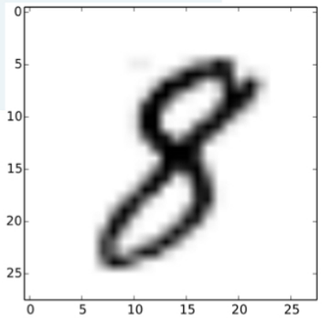
```
pixel149  pixel150  pixel151  pixel152  pixel153
0         0         0         0         0
86        250       254       254       254
0         0         0         9         254
0         0         0         0         0
103       253       253       253       253
0         0         5         165       254
0         0         0         0         0
0         0         0         0         0
0         0         0         0         41
253       253       253       253       253
```

Tab-delimited file

MNIST.txt

```
pixel149  pixel150  pixel151  pixel152  pixel153
0         0         0         0         0
86        250       254       254       254
0         0         0         9         254
0         0         0         0         0
103       253       253       253       253
0         0         5         165       254
0         0         0         0         0
0         0         0         0         0
0         0         0         0         41
253       253       253       253       253
```

MNIST image



How do you import flat files?

- Two main packages: NumPy, pandas

  - Here, you'll learn to import:
    - Flat files with numerical data (MNIST)
    - Flat files with numerical data and strings (titanic.csv)

Let's practice!

INTRODUCTION TO IMPORTING DATA IN PYTHON

Importing flat files using NumPy

INTRODUCTION TO IMPORTING DATA IN PYTHON

Hugo Bowne-Anderson
Data Scientist at DataCamp

Why NumPy?

- NumPy arrays: standard for storing numerical data

Why NumPy?

- NumPy arrays: standard for storing numerical data
- Essential for other packages: e.g. scikit-learn

- loadtxt()
- genfromtxt()

Importing flat files using NumPy

```
import numpy as np
filename = 'MNIST.txt'
data = np.loadtxt(filename, delimiter=',')
data
```

```
[[   0.    0.    0.    0.    0.]
 [  86.  250.  254.  254.  254.]
 [   0.    0.    0.    9.  254.]
 ...,
 [   0.    0.    0.    0.    0.]
 [   0.    0.    0.    0.    0.]
 [   0.    0.    0.    0.    0.]]
```

Customizing your NumPy import

```
import numpy as np
filename = 'MNIST_header.txt'
data = np.loadtxt(filename, delimiter=',', skiprows=1)
print(data)
```

```
[[   0.    0.    0.    0.    0.]
 [  86.  250.  254.  254.  254.]
 [   0.    0.    0.    9.  254.]
 ...,
 [   0.    0.    0.    0.    0.]
 [   0.    0.    0.    0.    0.]
 [   0.    0.    0.    0.    0.]]
```

Customizing your NumPy import

```
import numpy as np
filename = 'MNIST_header.txt'
data = np.loadtxt(filename, delimiter=',', skiprows=1, usecols=[0, 2
print(data)
```

```
[[   0.    0.]
 [  86.  254.]
 [   0.    0.]
 ...,
 [   0.    0.]
 [   0.    0.]
 [   0.    0.]]
```

Customizing your NumPy import

```
data = np.loadtxt(filename, delimiter=',', dtype=str)
```

Mixed datatypes

titanic.csv

Name	Gender	Cabin	Fare
Braund, Mr. Owen Harris	male	NaN	7.3
Cumings, Mrs. John Bradley	female	C85	71.3
Heikkinen, Miss. Laina	female	NaN	8.0
Futrelle, Mrs. Jacques Heath	female	C123	53.1
Allen, Mr. William Henry	male	NaN	8.05

[1] Source: Kaggle

Mixed datatypes

titanic.csv

```
                          Name  Gender Cabin  Fare
      Braund, Mr. Owen Harris    male   NaN   7.3
  Cumings, Mrs. John Bradley  female   C85  71.3
      Heikkinen, Miss. Laina  female   NaN   8.0
Futrelle, Mrs. Jacques Heath  female  C123  53.1
    Allen, Mr. William Henry    male   NaN  8.05
```

Let's practice!

INTRODUCTION TO IMPORTING DATA IN PYTHON

Importing flat files using pandas

INTRODUCTION TO IMPORTING DATA IN PYTHON

Hugo Bowne-Anderson
Data Scientist at DataCamp

What a data scientist needs

- Two-dimensional labeled data structure(s)
- Columns of potentially different types
- Manipulate, slice, reshape, groupby, join, merge
- Perform statistics
- Work with time series data

Pandas and the DataFrame

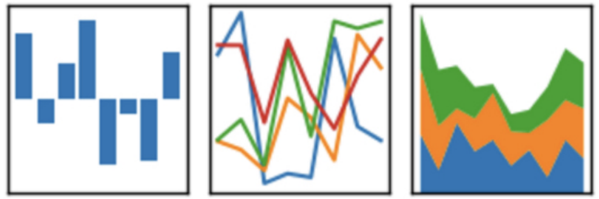
Wes McKinney

Pandas and the DataFrame

Wes McKinney

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Pandas and the DataFrame

What problem does *pandas* solve?

Python has long been great for data munging and preparation, but less so for data analysis and modeling. *pandas* helps fill this gap, enabling you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

- DataFrame = pythonic analog of R's data frame

Pandas and the DataFrame

Manipulating pandas DataFrames

- Exploratory data analysis
- Data wrangling
- Data preprocessing
- Building models
- Visualization
- Standard and best practice to use pandas

Importing using pandas

```
import pandas as pd
filename = 'winequality-red.csv'
data = pd.read_csv(filename)
data.head()
```

```
   volatile acidity  citric acid  residual sugar
0              0.70         0.00             1.9
1              0.88         0.00             2.6
2              0.76         0.04             2.3
3              0.28         0.56             1.9
4              0.70         0.00             1.9
```

```
data_array = data.values
```

You'll experience:

- Importing flat files in a straightforward manner
- Importing flat files with issues such as comments and missing values

Let's practice!

INTRODUCTION TO IMPORTING DATA IN PYTHON

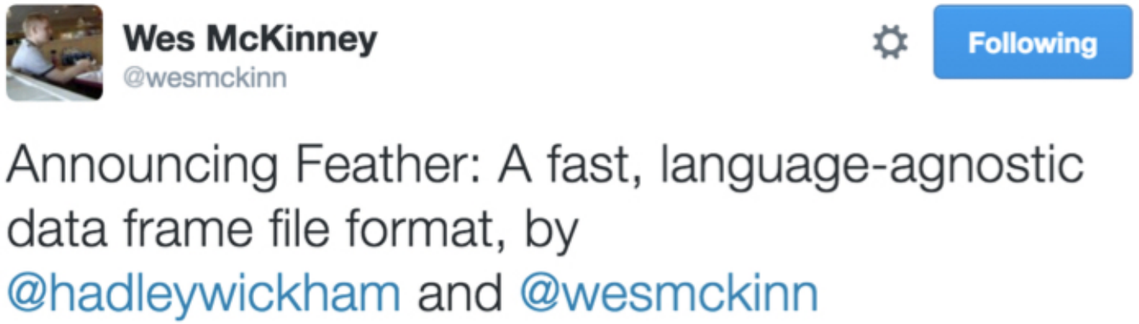
Final thoughts on data import

INTRODUCTION TO IMPORTING DATA IN PYTHON

Hugo Bowne-Anderson
Data Scientist at DataCamp

Next chapters:

- Import other file types:
  - Excel, SAS, Stata
- Feather



- Interact with relational databases

Next course:

- Scrape data from the web
- Interact with APIs

Let's practice!

INTRODUCTION TO IMPORTING DATA IN PYTHON