706.088 INFORMATIK 1 (VU)

Data Analysis
(part 1)

What is Data Analysis?

A process of **inspecting**, **cleansing**, **transforming** and **modelling data** with the goal of discovering useful information, informing conclusion and supporting decision-making.

- Definition by Wikipedia

Data Analysis Pipeline

Data Extraction
Databases
Files (e.g., CSV)

Data Analysis Pipeline

Data Extraction	Data Cleaning
Databases	Deal with missing values
Files (e.g., CSV)	Outliers and non relevant data

Data Analysis Pipeline

Data Extraction	Data Cleaning	Data Transformation
Databases	Deal with missing values	Combine different data sources
Files (e.g., CSV)	Outliers and non relevant data	Perform calculations

Data Analysis Pipeline

Data Extraction	Data Cleaning	Data Transformation	Analysis
Databases	Deal with missing values	Combine different data sources	Visualization
Files (e.g., CSV)	Outliers and non relevant data	Perform calculations	Exploration / Statistical Analysis

Data Analysis Pipeline

Data Extraction	**Data Cleaning**	**Data Transformation**	**Analysis**	**Action**
Databases	Deal with missing values	Combine different data sources	Visualization	Machine Learning Models
Files (e.g., CSV)	Outliers and non relevant data	Perform calculations	Exploration / Statistical Analysis	Decision making

Useful Third-Party Libraries

Data Extraction	Data Cleaning	Data Transformation	Analysis	Action
pandas	pandas	pandas NumPy	NumPy matplotlib scikit learn	scikit learn

Links: pandas numpy matplotlib scikit learn

Data Extraction (Reading CSV Files)

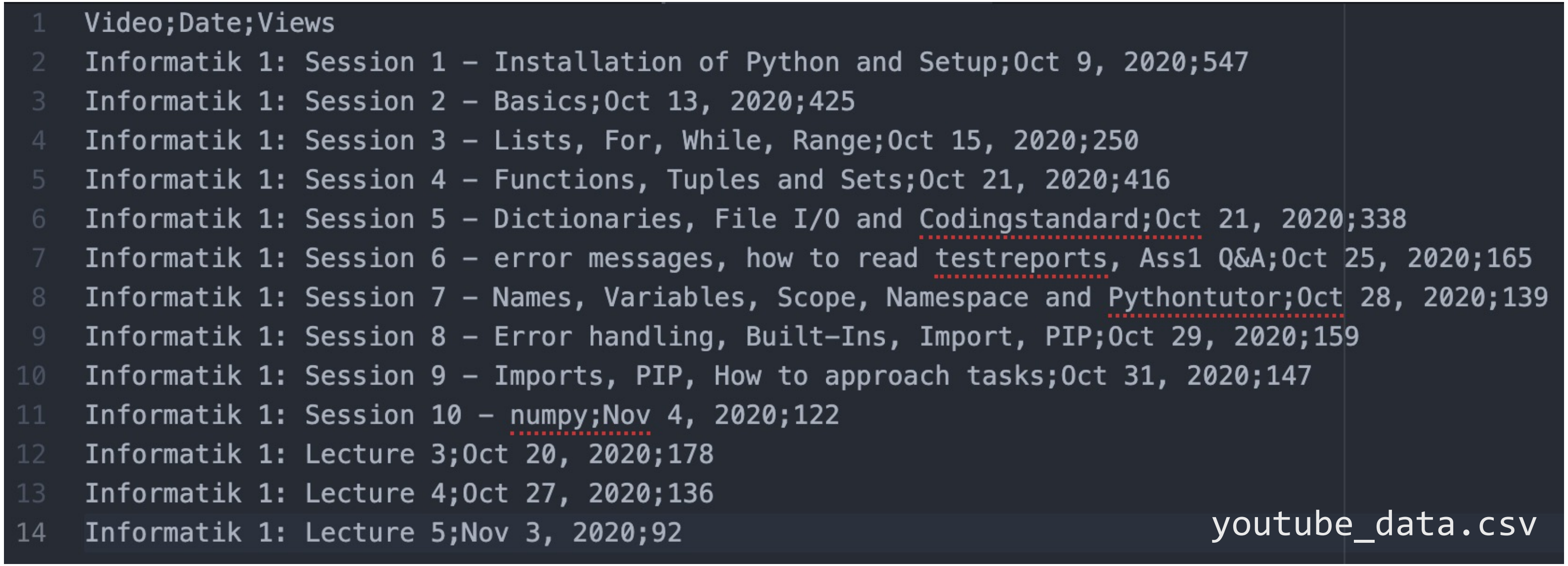
Comma Separated Values (CSV) files are a common way to store data.

Video	**Date**	**Views**
Informatik 1: Session 1	Oct 9, 2020	547
Informatik 1: Session 2	Oct 13, 2020	425
Informatik 1: Session 3	Oct 15, 2020	250
Informatik 1: Session 4	Oct 21, 2020	416
Informatik 1: Session 5	Oct 21, 2020	338

Table: Views of YouTube videos

Data Extraction (Reading CSV Files)

Comma Separated Values (CSV) files are a common way to store data.

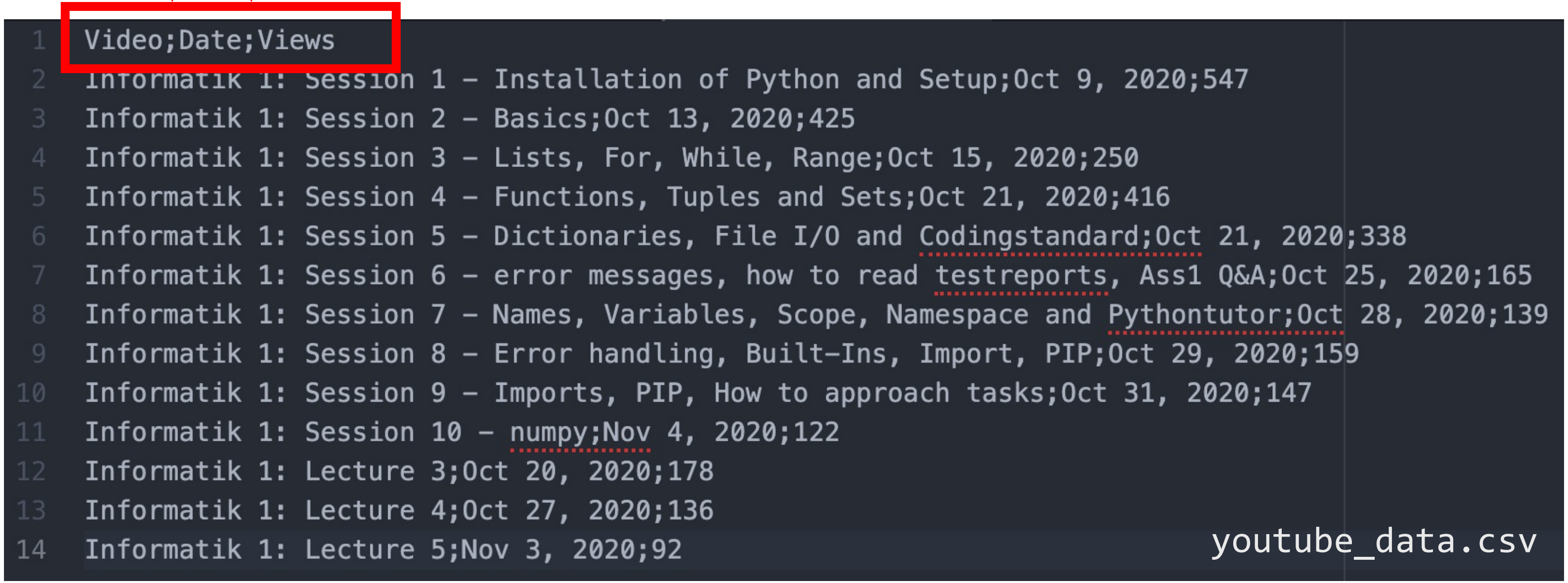


```
Video;Date;Views
Informatik 1: Session 1 – Installation of Python and Setup;Oct 9, 2020;547
Informatik 1: Session 2 – Basics;Oct 13, 2020;425
Informatik 1: Session 3 – Lists, For, While, Range;Oct 15, 2020;250
Informatik 1: Session 4 – Functions, Tuples and Sets;Oct 21, 2020;416
Informatik 1: Session 5 – Dictionaries, File I/O and Codingstandard;Oct 21, 2020;338
Informatik 1: Session 6 – error messages, how to read testreports, Ass1 Q&A;Oct 25, 2020;165
Informatik 1: Session 7 – Names, Variables, Scope, Namespace and Pythontutor;Oct 28, 2020;139
Informatik 1: Session 8 – Error handling, Built-Ins, Import, PIP;Oct 29, 2020;159
Informatik 1: Session 9 – Imports, PIP, How to approach tasks;Oct 31, 2020;147
Informatik 1: Session 10 – numpy;Nov 4, 2020;122
Informatik 1: Lecture 3;Oct 20, 2020;178
Informatik 1: Lecture 4;Oct 27, 2020;136
Informatik 1: Lecture 5;Nov 3, 2020;92
```

youtube_data.csv

Data Extraction (Reading CSV Files)

Delimiter

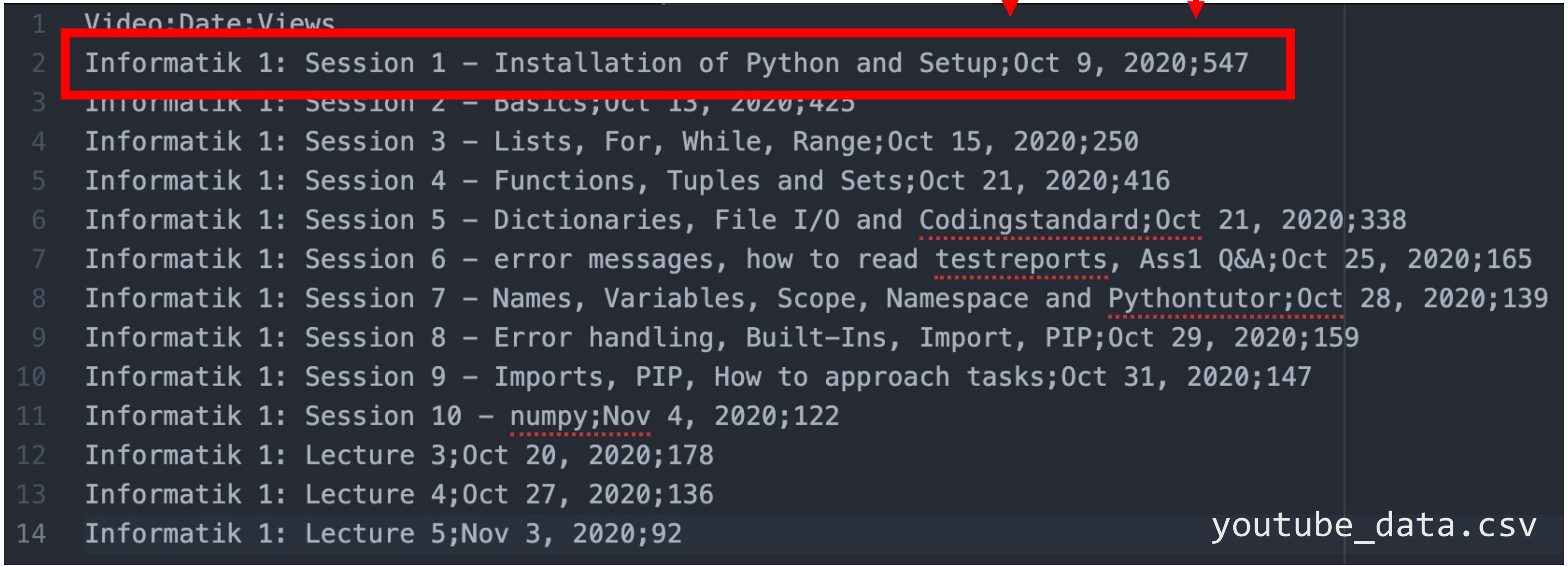


```
Video;Date;Views
Informatik 1: Session 1 – Installation of Python and Setup;Oct 9, 2020;547
Informatik 1: Session 2 – Basics;Oct 13, 2020;425
Informatik 1: Session 3 – Lists, For, While, Range;Oct 15, 2020;250
Informatik 1: Session 4 – Functions, Tuples and Sets;Oct 21, 2020;416
Informatik 1: Session 5 – Dictionaries, File I/O and Codingstandard;Oct 21, 2020;338
Informatik 1: Session 6 – error messages, how to read testreports, Ass1 Q&A;Oct 25, 2020;165
Informatik 1: Session 7 – Names, Variables, Scope, Namespace and Pythontutor;Oct 28, 2020;139
Informatik 1: Session 8 – Error handling, Built-Ins, Import, PIP;Oct 29, 2020;159
Informatik 1: Session 9 – Imports, PIP, How to approach tasks;Oct 31, 2020;147
Informatik 1: Session 10 – numpy;Nov 4, 2020;122
Informatik 1: Lecture 3;Oct 20, 2020;178
Informatik 1: Lecture 4;Oct 27, 2020;136
Informatik 1: Lecture 5;Nov 3, 2020;92
```

youtube_data.csv

Data Extraction (Reading CSV Files)

Delimiter



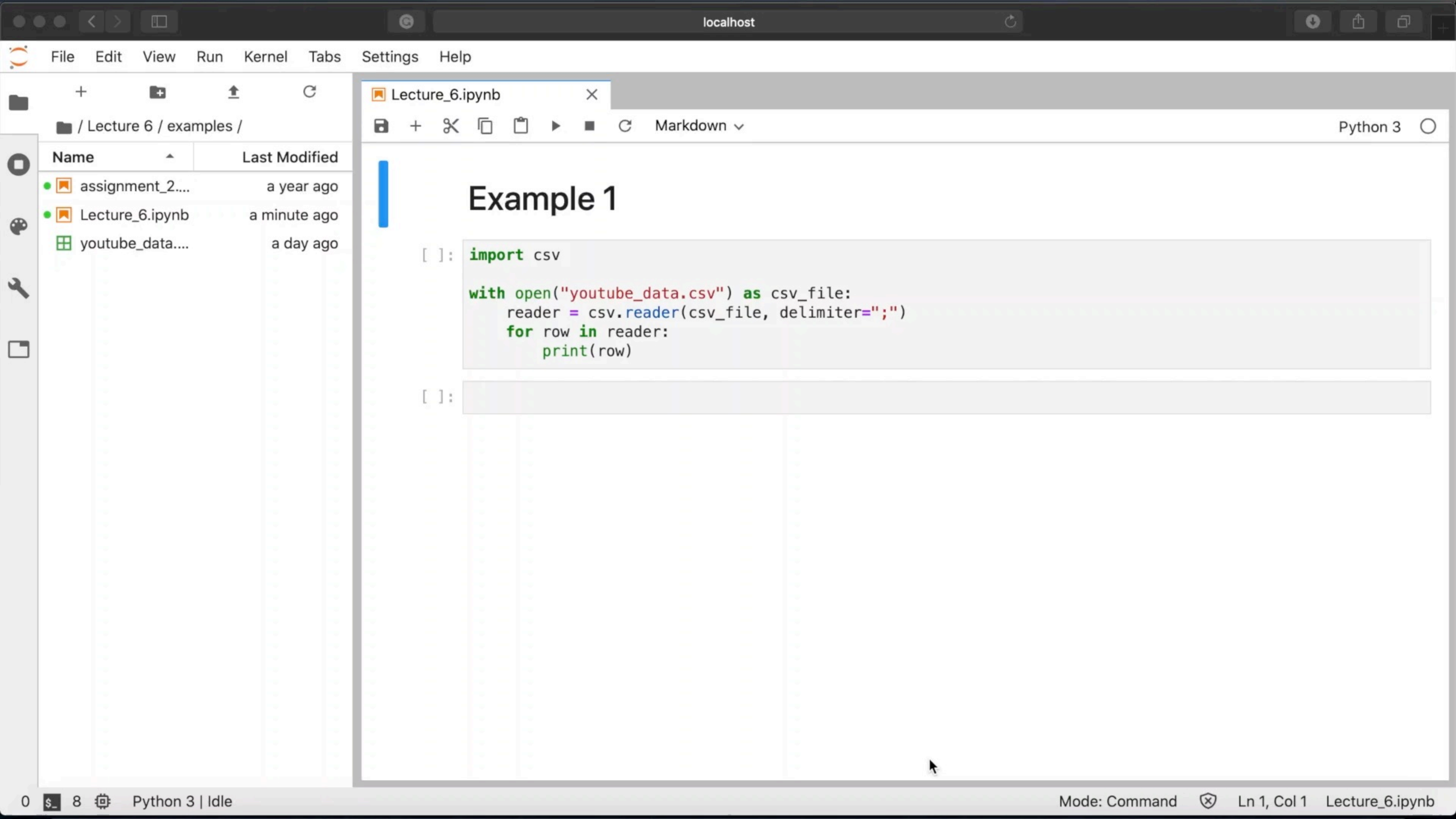
```
Video;Date;Views
Informatik 1: Session 1 – Installation of Python and Setup;Oct 9, 2020;547
Informatik 1: Session 2 – Basics;Oct 13, 2020;425
Informatik 1: Session 3 – Lists, For, While, Range;Oct 15, 2020;250
Informatik 1: Session 4 – Functions, Tuples and Sets;Oct 21, 2020;416
Informatik 1: Session 5 – Dictionaries, File I/O and Codingstandard;Oct 21, 2020;338
Informatik 1: Session 6 – error messages, how to read testreports, Ass1 Q&A;Oct 25, 2020;165
Informatik 1: Session 7 – Names, Variables, Scope, Namespace and Pythontutor;Oct 28, 2020;139
Informatik 1: Session 8 – Error handling, Built-Ins, Import, PIP;Oct 29, 2020;159
Informatik 1: Session 9 – Imports, PIP, How to approach tasks;Oct 31, 2020;147
Informatik 1: Session 10 – numpy;Nov 4, 2020;122
Informatik 1: Lecture 3;Oct 20, 2020;178
Informatik 1: Lecture 4;Oct 27, 2020;136
Informatik 1: Lecture 5;Nov 3, 2020;92
```

youtube_data.csv

CSV Module

```python
import csv

with open("youtube_data.csv") as csv_file:
    reader = csv.reader(csv_file, delimiter=";")
    for row in reader:
        print(row)
```



Example 1

```
[ ]: import csv

     with open("youtube_data.csv") as csv_file:
         reader = csv.reader(csv_file, delimiter=";")
         for row in reader:
             print(row)
```

```
[ ]:
```

Pandas Library

Pandas Library

- Powerful and flexible module for processing data
- Simplifies data cleaning / preparation **a lot!**
- Hides details / complexity of reading / writing data from the programmer
- Integrates NumPy to enable easier calculations

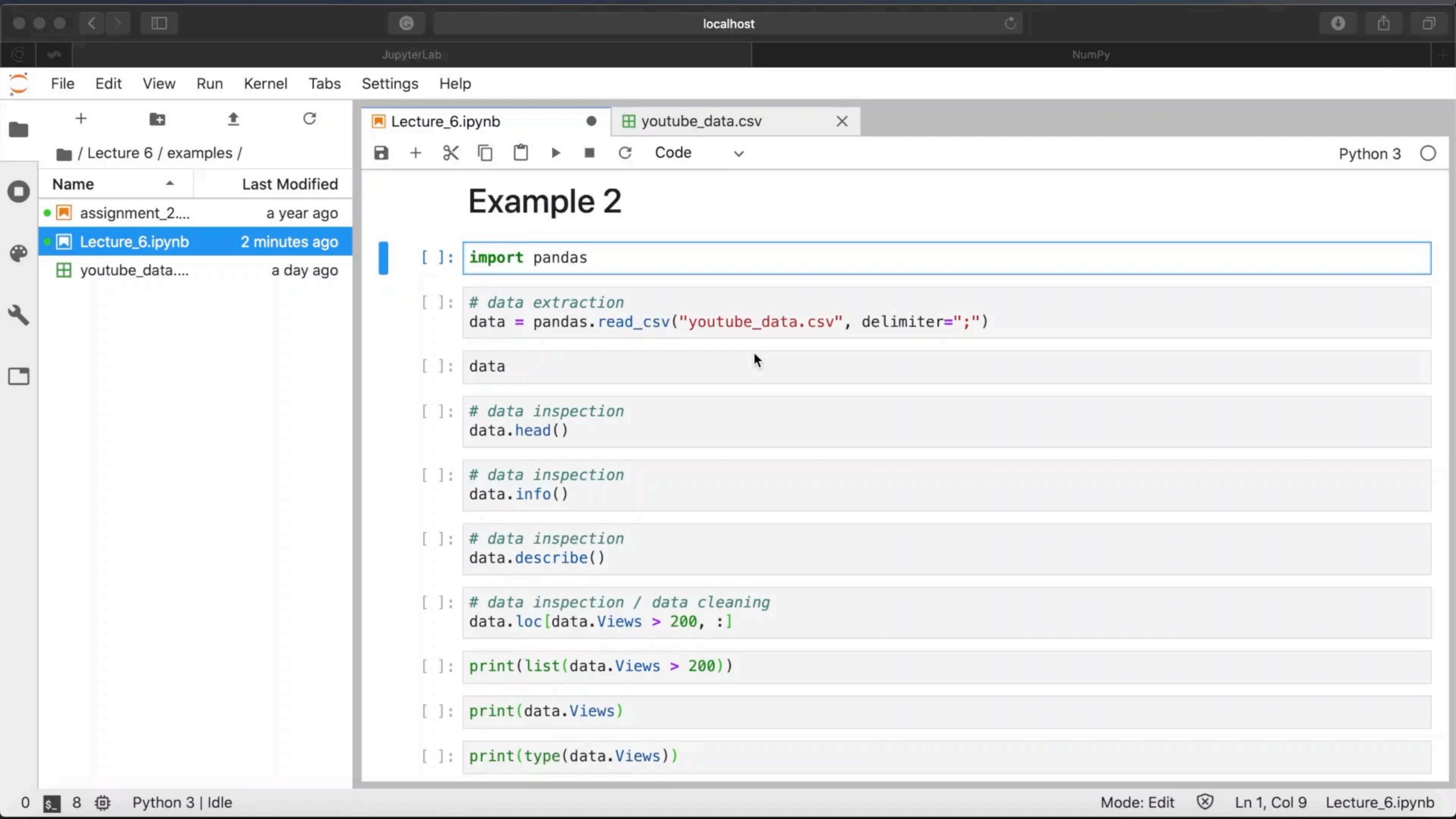
Quick Side Note

Difference between a Module, Package and Library in Python

- **Module** is a file which contains various Python functions and global variables

- **Package** is a collection of modules.

- **Library** is a collection of packages.



Example 2

```
[ ]: import pandas
```

```
[ ]: # data extraction
     data = pandas.read_csv("youtube_data.csv", delimiter=";")
```

```
[ ]: data
```

```
[ ]: # data inspection
     data.head()
```

```
[ ]: # data inspection
     data.info()
```

```
[ ]: # data inspection
     data.describe()
```

```
[ ]: # data inspection / data cleaning
     data.loc[data.Views > 200, :]
```

```
[ ]: print(list(data.Views > 200))
```

```
[ ]: print(data.Views)
```

```
[ ]: print(type(data.Views))
```

NumPy Package

NumPy

Fundamental package for scientific computing
“MATLAB in Python”

Contains:

- Powerful N-dimensional array (list) object
- Sophisticated functions
- Useful linear algebra (matrix and vector products, etc.)
- Random number generation

NumPy Arrays

- NumPy’s main object is the multidimensional array
- In NumPy dimensions are called *axes*

```
import numpy as np

vector = np.array([1, 2, 3])         # one axis / dimension
matrix = np.array([[1, 2, 3],        # two axes / dimensions
                   [4, 5, 6]])

print(matrix.shape)            # prints: (2, 3)
print(matrix.ndim)             # prints: 2
```

NumPy Operators / Functions

Standard operators (+, -, *, …) are elementwise

```
participants_weight = np.array([50, 61, 75, 70])
participants_height = np.array([1.60, 1.50, 1.73, 1.80])

bmi = participants_weight / participants_height ** 2
```

NumPy Operators / Functions

Universal functions

```
random_numbers = np.array([0, 1, 2])
np.exp(random_numbers)        # exponential, elementwise
np.sqrt(random_numbers)       # square root, elementwise
np.sin(random_numbers)        # trigonometric sine, elementwise
```

See docs.scipy.org for full list of available functions

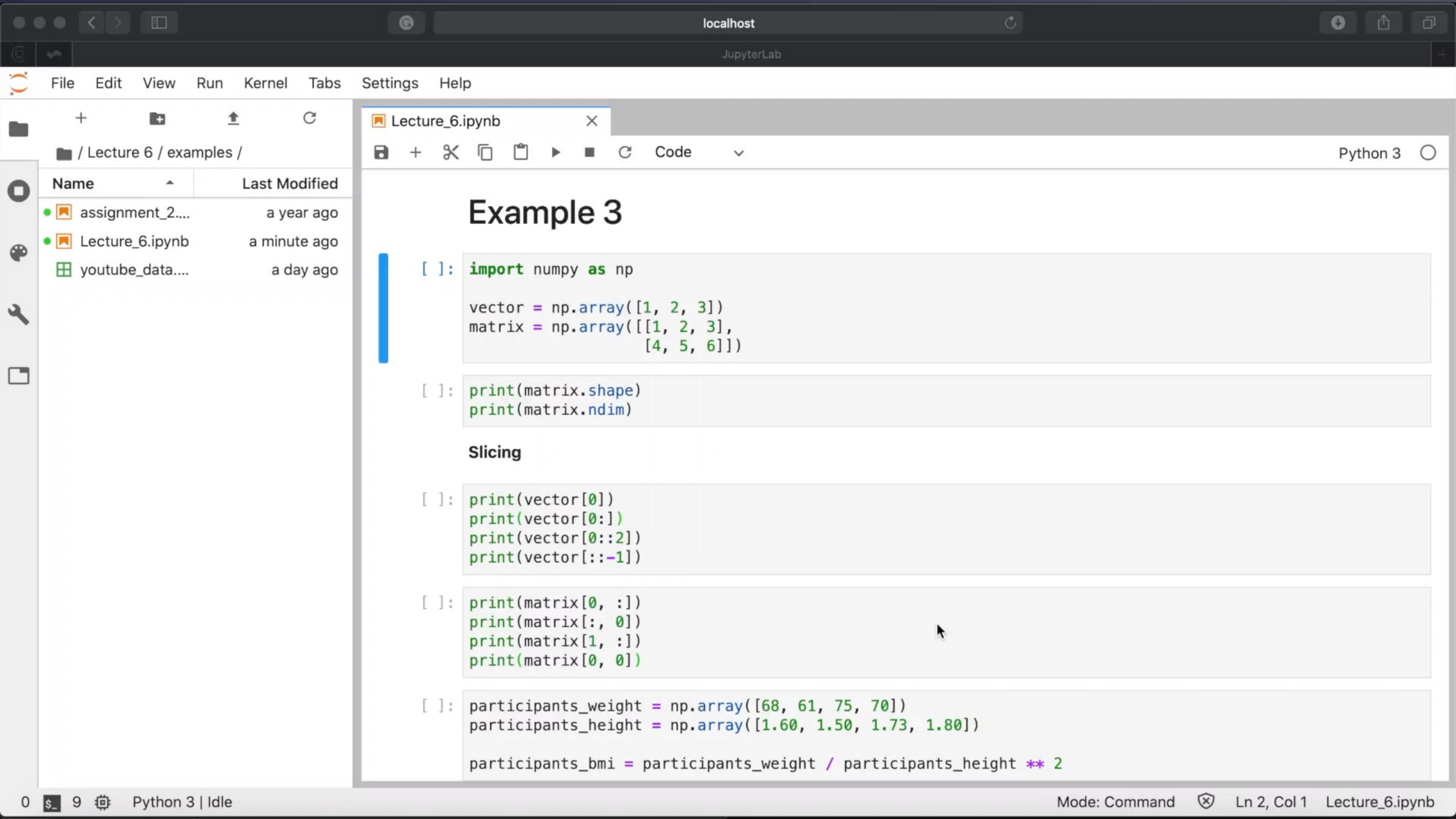
NumPy Array Methods

NumPy array is a **class** and implements some handy methods

```
participants_weight = np.array([50, 61, 75, 70])

print(participants_weight.mean())   # prints: 64.0
print(participants_weight.max())    # prints: 75
print(participants_weight.min())    # prints: 50
```

See docs.scipy.org for full list of available methods



Example 3

```
[ ]: import numpy as np

     vector = np.array([1, 2, 3])
     matrix = np.array([[1, 2, 3],
                        [4, 5, 6]])
```

```
[ ]: print(matrix.shape)
     print(matrix.ndim)
```

Slicing

```
[ ]: print(vector[0])
     print(vector[0:])
     print(vector[0::2])
     print(vector[::-1])
```

```
[ ]: print(matrix[0, :])
     print(matrix[:, 0])
     print(matrix[1, :])
     print(matrix[0, 0])
```

```
[ ]: participants_weight = np.array([68, 61, 75, 70])
     participants_height = np.array([1.60, 1.50, 1.73, 1.80])

     participants_bmi = participants_weight / participants_height ** 2
```

Thanks!