Welcome to INFO 2950 (Intro to Data Science)!

Pick up 1 whiteboard, 1 marker, and a few tissues (erasers) on your way in.

Feel free to draw a cat while you wait for class to start.

(Make sure to return these at the end of class!)

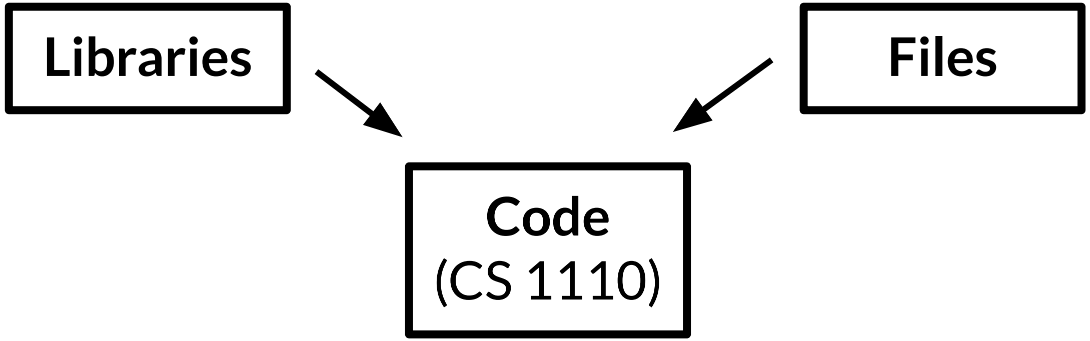
INFO 2950: Intro to Data Science

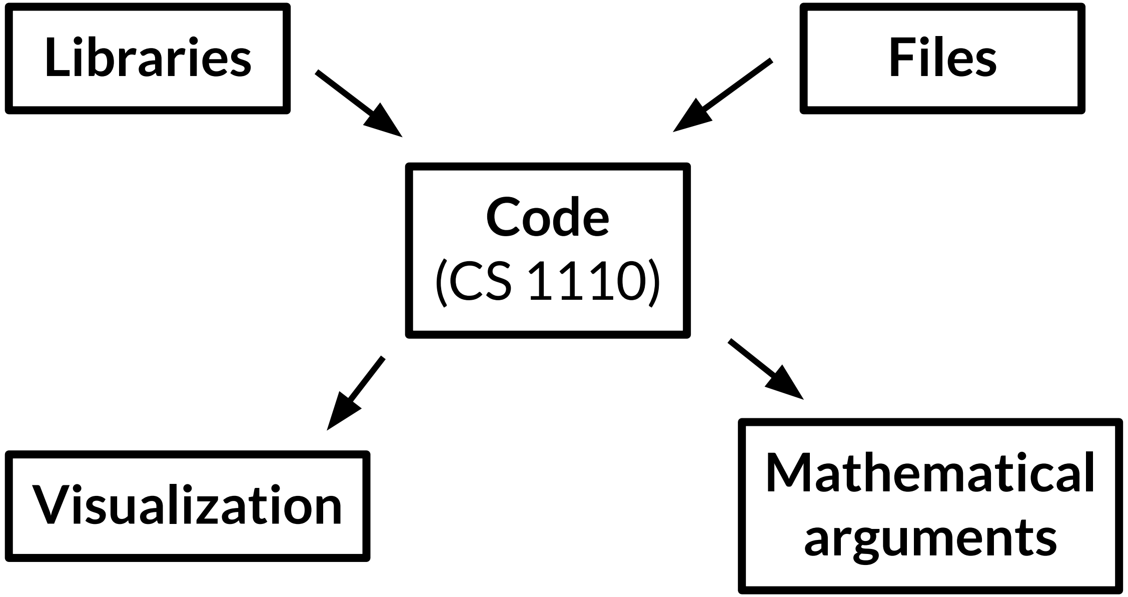
Lecture 2
2023-08-23

Agenda

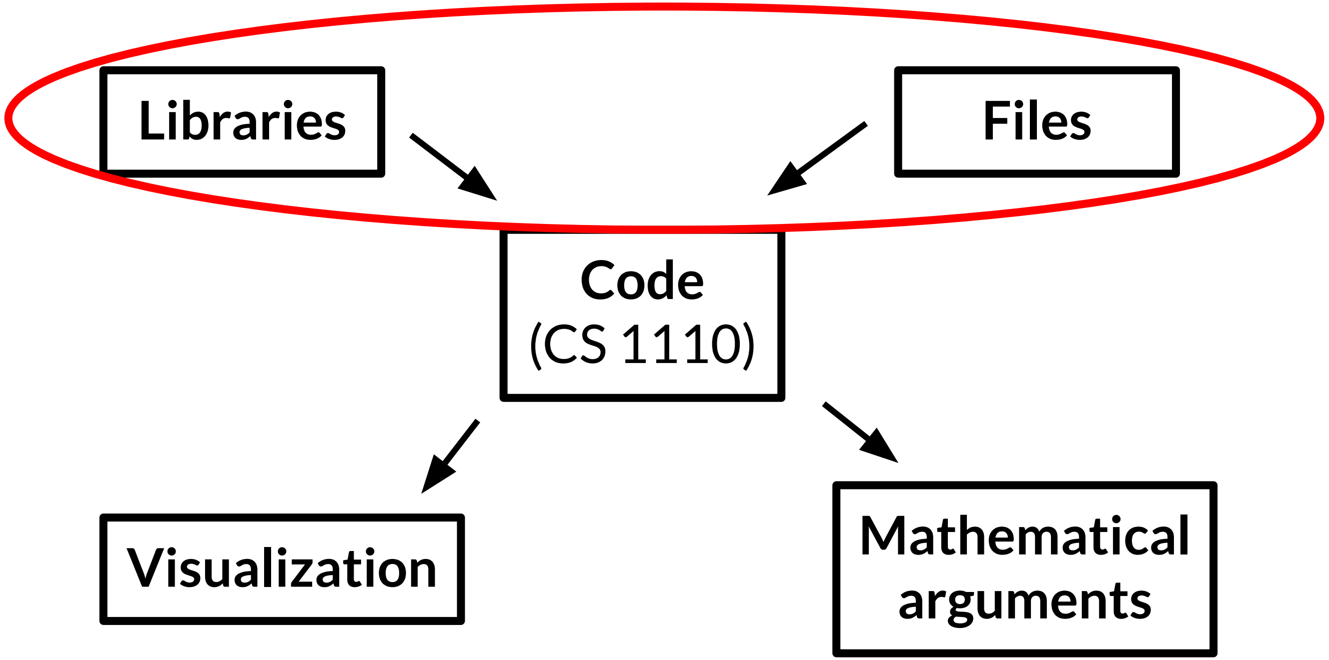
1. Libraries
2. Arrays
3. Importing & file paths
4. Data type conversions
5. SQL
6. Admin

Code
(CS 1110)


Libraries
Files
Code
(CS 1110)

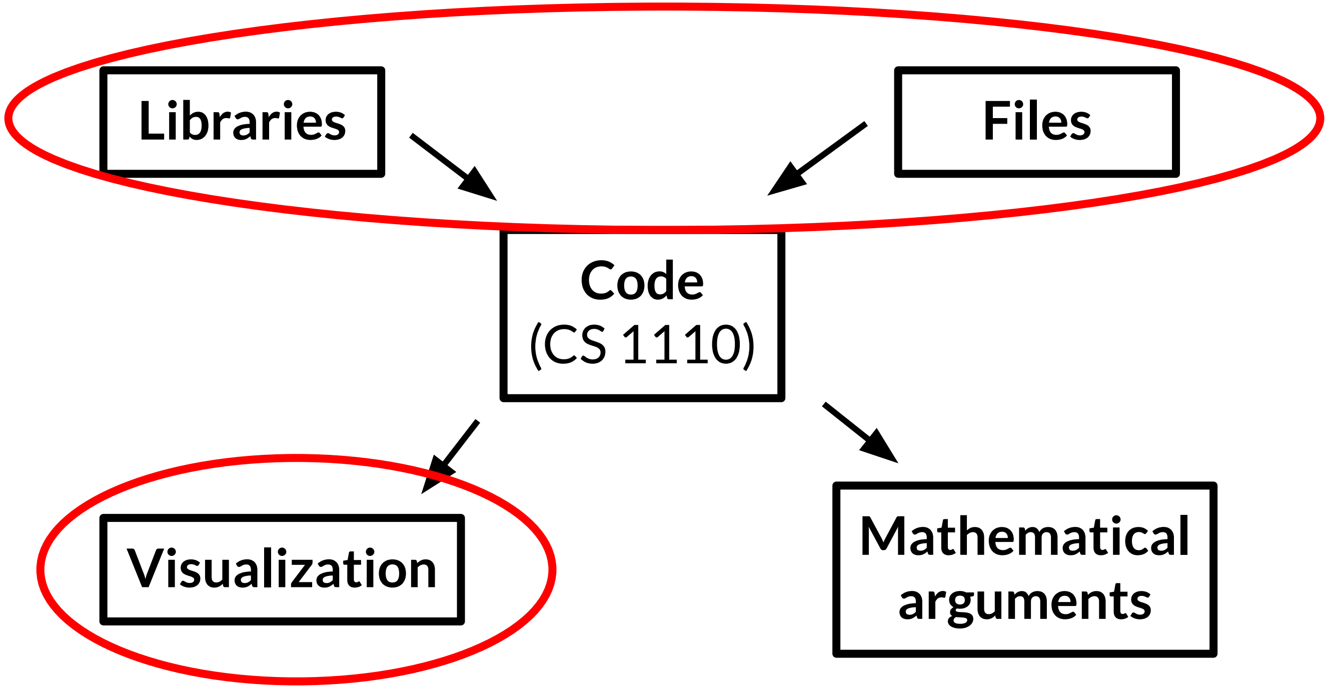

Libraries
Files
Code
(CS 1110)
Visualization
Mathematical arguments

Today


Libraries
Files
Code
(CS 1110)
Visualization
Mathematical arguments

Today

Next Week



Libraries

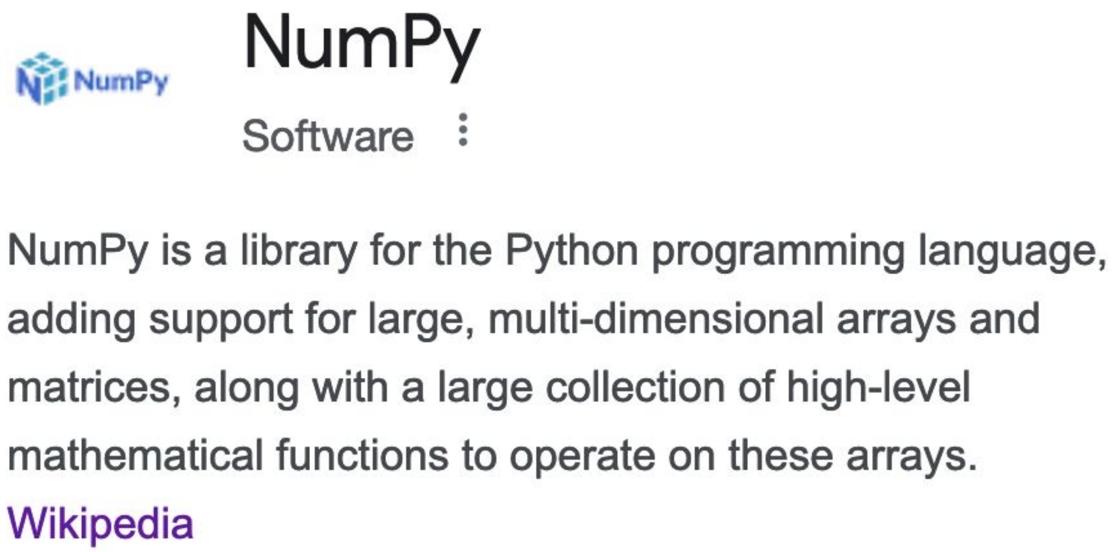
- Libraries give you add-ons to base Python
  → write powerful code in only a few lines!
- Libraries > Packages > Modules

Libraries

- Libraries give you add-ons to base Python
- Libraries > Packages > Modules
- **What code do you write to import a library called *NumPy*?**

Libraries

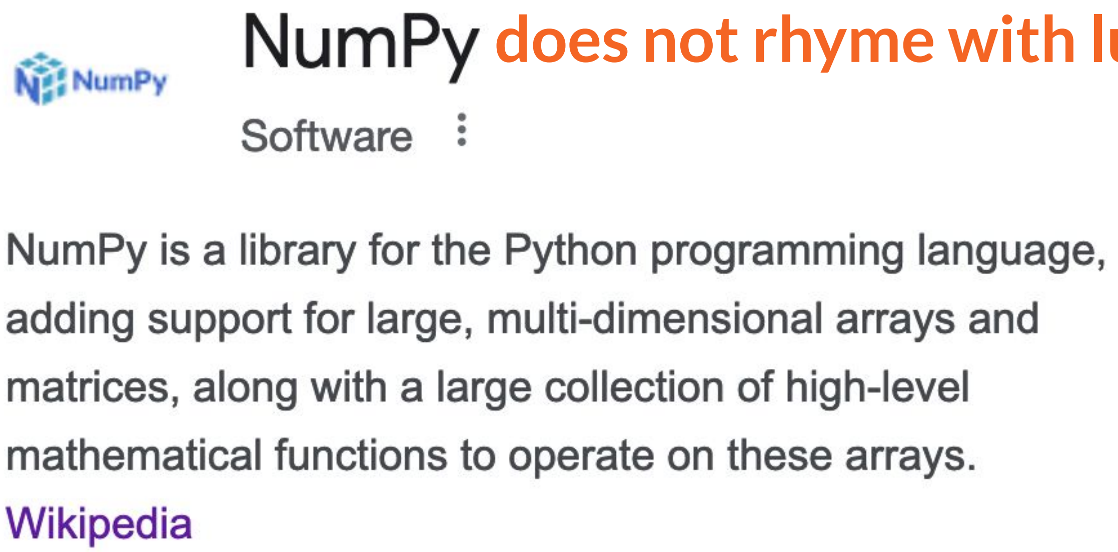
- Libraries give you add-ons to base Python
- Libraries > Packages > Modules
- What code do you write to import a library called **NumPy**?
  - `import numpy`
    OR
  - `import numpy as np`



NumPy

Software

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
Wikipedia



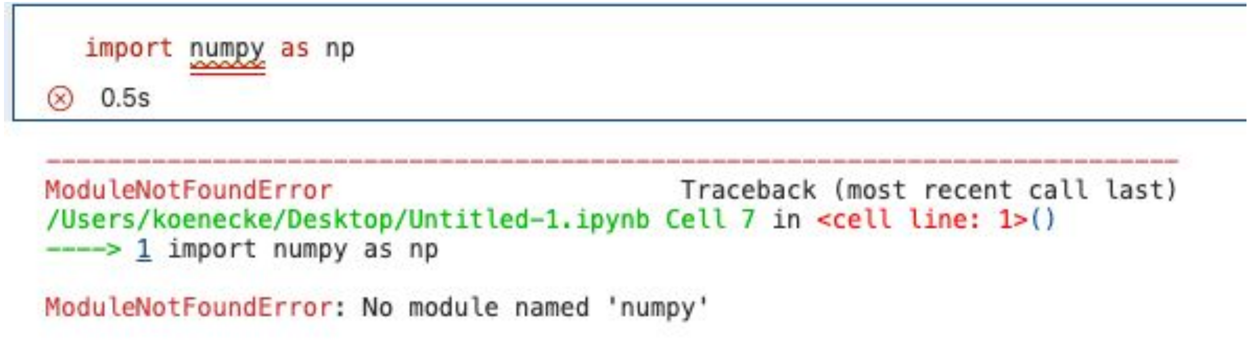
NumPy **does not rhyme with lumpy!!**

Software

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
Wikipedia

Libraries

- Running the correct code gives us an error! Why?

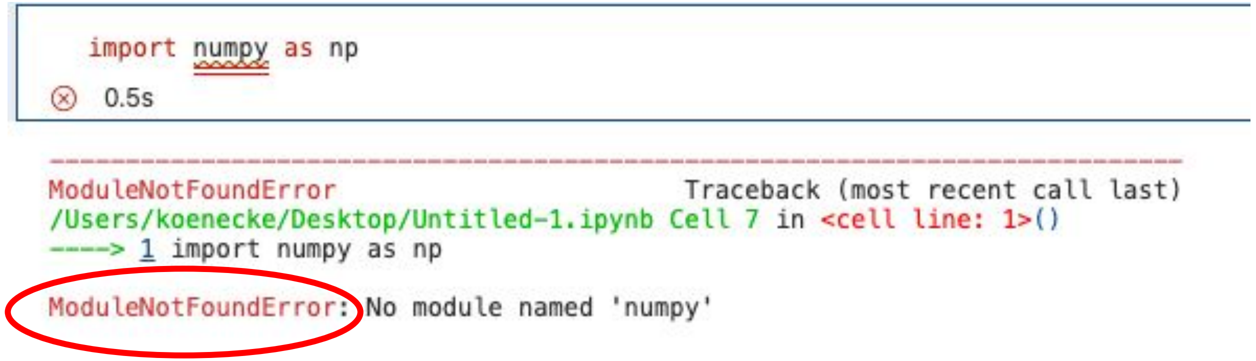


```
import numpy as np
0.5s
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/Users/koenecke/Desktop/Untitled-1.ipynb Cell 7 in <cell line: 1>()
----> 1 import numpy as np

ModuleNotFoundError: No module named 'numpy'
```

Libraries

- Running the correct code gives us an error! Why?

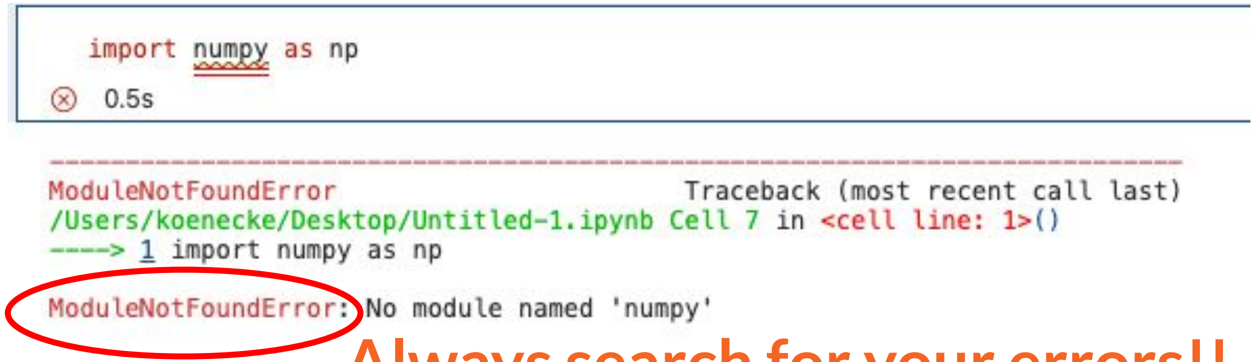


```
import numpy as np
0.5s
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/Users/koenecke/Desktop/Untitled-1.ipynb Cell 7 in <cell line: 1>()
----> 1 import numpy as np

ModuleNotFoundError: No module named 'numpy'
```

Libraries

- Running the correct code gives us an error! Why?



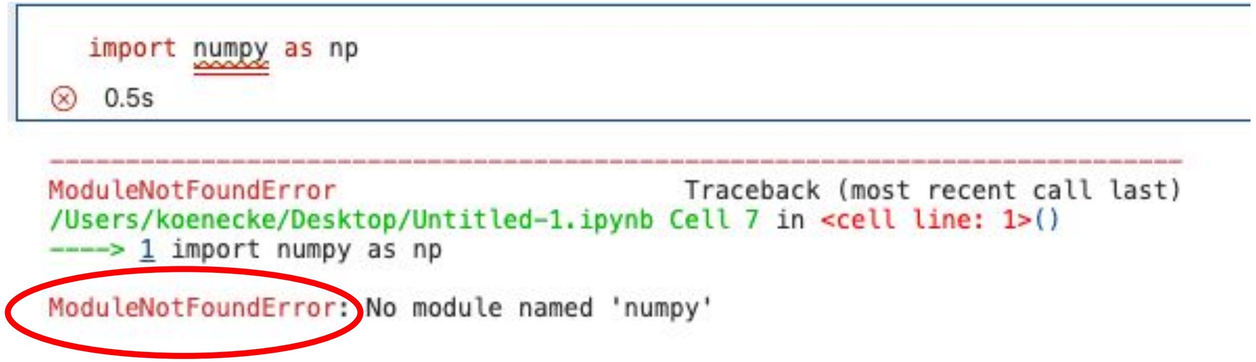
```
import numpy as np
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/Users/koenecke/Desktop/Untitled-1.ipynb Cell 7 in <cell line: 1>()
----> 1 import numpy as np

ModuleNotFoundError: No module named 'numpy'
```

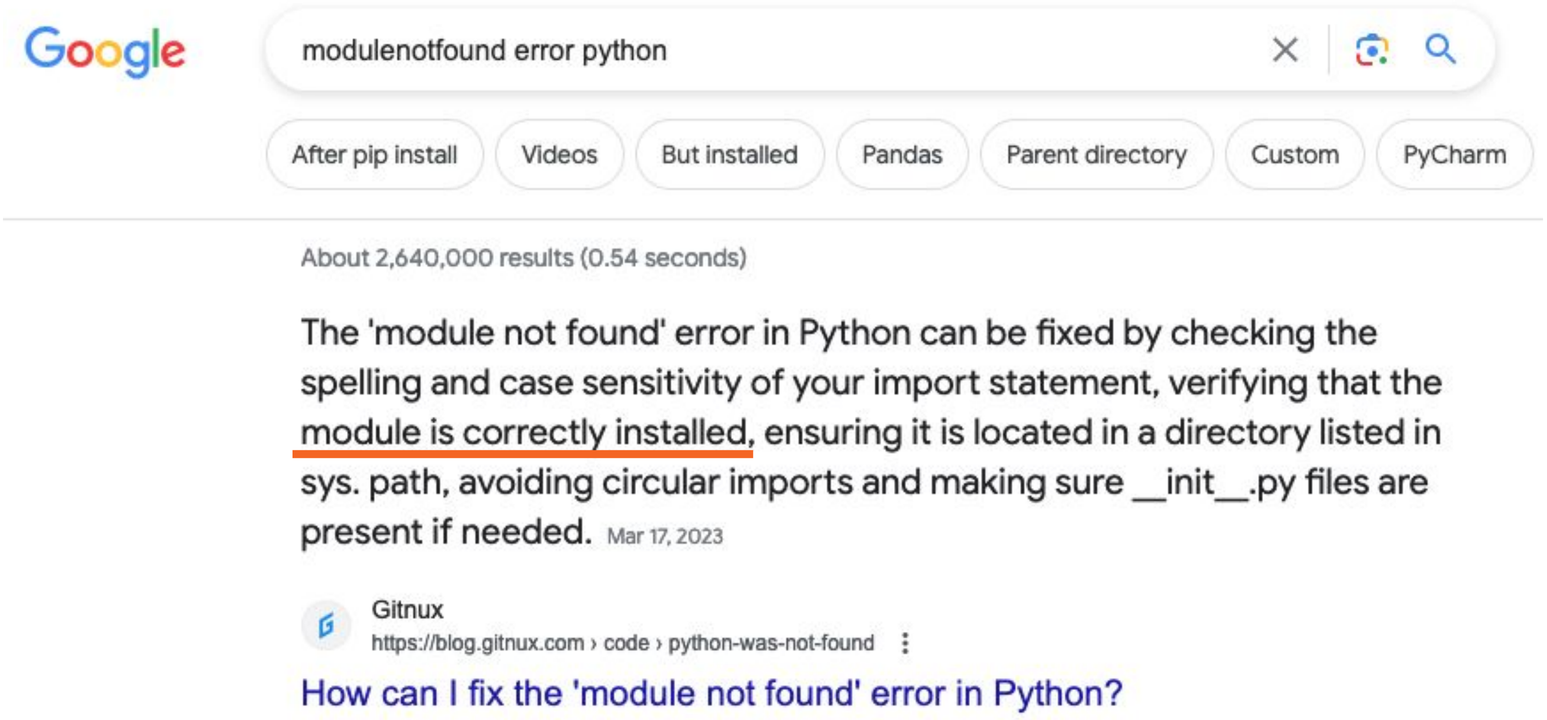
Always search for your errors!!
(Google, StackOverflow, ...)

What search query would you use to figure out this error?

- If we just run import code, it doesn't work! Why?



```
import numpy as np
⊗ 0.5s

---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
/Users/koenecke/Desktop/Untitled-1.ipynb Cell 7 in <cell line: 1>()
----> 1 import numpy as np

ModuleNotFoundError: No module named 'numpy'
```


modulenotfound error python

About 2,640,000 results (0.54 seconds)

The 'module not found' error in Python can be fixed by checking the spelling and case sensitivity of your import statement, verifying that the module is correctly installed, ensuring it is located in a directory listed in sys. path, avoiding circular imports and making sure __init__.py files are present if needed. Mar 17, 2023

Gitnux
https://blog.gitnux.com › code › python-was-not-found

How can I fix the 'module not found' error in Python?

Libraries

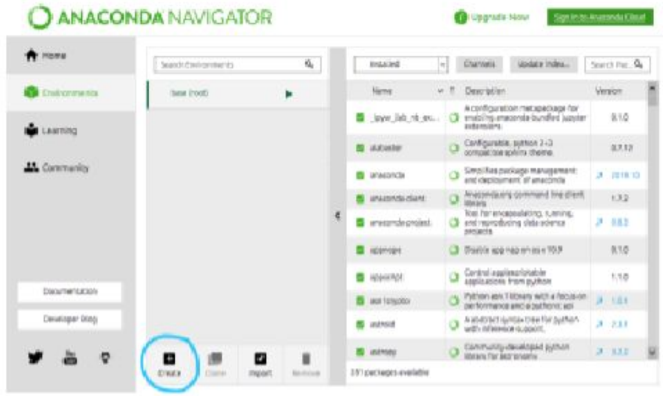
- **Installing != Loading**
- Need to **install** the library before **loading** it in your Python notebook
  - In Anaconda Navigator GUI (click ‘numpy’)
  - In terminal: *conda install numpy*
- Other install methods:
  - conda; pip; within notebook…

Why use Virtual Environments?

INFO 2950 Python Install Instructions.pdf

Download INFO 2950 Python Install Instructions.pdf (243 KB) | Alternative formats

4. **Click the "Create" button** to make a new environment:



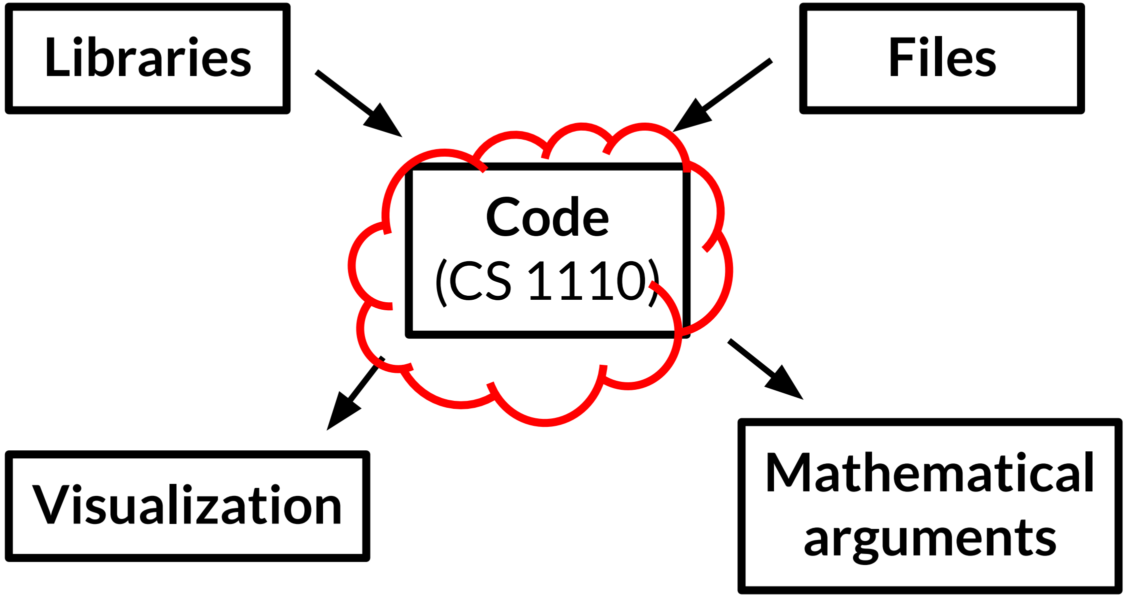
A dialogue box will open. Fill in **info2950** in the name field (the screenshot is older; you may see version 3.10 or 3.11):

Why use Virtual Environments?

- Lots of packages are being actively developed (good!) but depend on different versions of each other (bad!)
- Python isn’t great at dependency management

Why use Virtual Environments?

- Lots of packages are being actively developed (good!) but depend on different versions of each other (bad!)
- Python isn’t great at dependency management
- Dependency conflicts (different versions of Python)
- Install packages → mess up your OS?
- Control versioning & where packages go with venv


Libraries
Files
Code
(CS 1110)
Visualization
Mathematical arguments

Lists vs. Arrays

Multiple dimensions through nested lists of arbitrary size	Multiple dimensions, but must be the same length

Lists vs. Arrays

Multiple dimensions through nested lists of arbitrary size	Multiple dimensions, but must be the same length
Indexing multiple dimensions requires [i][j]	Indexing multiple dimensions with [i,j]

<u>Lists</u> vs. <u>Arrays</u>

Multiple dimensions through nested lists of arbitrary size	Multiple dimensions, but must be the same length
Indexing multiple dimensions requires [i][j]	Indexing multiple dimensions with [i,j]
Python checks type of each element = SLOW	Python only checks type once = FAST

Lists vs. Arrays

Multiple dimensions through nested lists of arbitrary size	Multiple dimensions, but must be the same length
Indexing multiple dimensions requires [i][j]	Indexing multiple dimensions with [i,j]
Python checks type of each element = SLOW	Python only checks type once = FAST
Requires for loops for operations on each element	Can do numerical operations "all at once"

Question: List or Array?

[5, “dog”, 7.3]

Question: List or Array?

[5, “dog”, 7.3]

It includes more than one type (integer, string, float), so it is a *list*

Question: List or Array?

```
[ [ 5, 5, 7 ],
  [ 6, 5, 3 ],
  [ 3, 8, 2 ] ]
```

Question: List or Array?

It *could* be a list, but it has only integers so it can also be an array. It has two dimensions.

If it is a list, it has three elements which happen to also be lists.

```
[ [ 5, 5, 7 ],
  [ 6, 5, 3 ],
  [ 3, 8, 2 ] ]
```

0-Indexing in Python/Numpy

```
[[5, 5, 7],
 [6, 5, 3],
 [3, 8, 2]]
```

Numpy: x[0, 2]
Python: x[0][2]

What index is this?

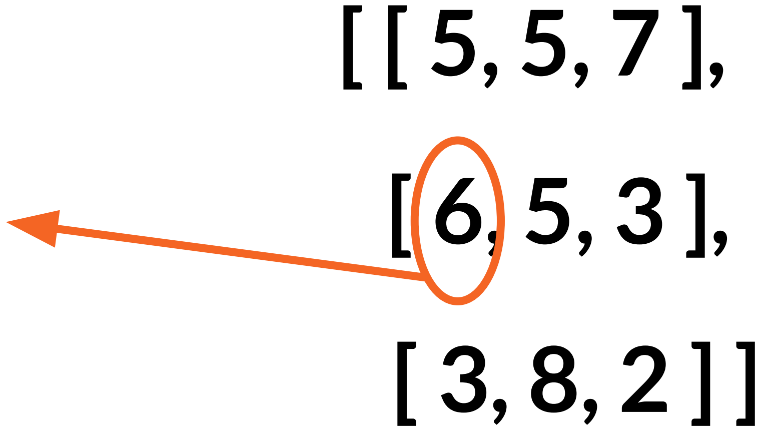
Numpy: x[_, _]
Python: x[_][_]

[[5, 5, 7],

[$\boxed{6}$, 5, 3],

[3, 8, 2]]

What index is this?



```
[[5, 5, 7],
 [6, 5, 3],
 [3, 8, 2]]
```

Numpy: x[1, 0]
Python: x[1][0]

Question: List or Array?

```
[ [ 5, 5, 7 ],
  [ 6, 5 ],
  [ 3, 8, 2 ] ]
```

Question: List or Array?

[[5, 5, 7],
[6, 5],
[3, 8, 2]]

The rows don't have the same number of elements, so it can't be a 2D array. This is a *list*.

Arrays

- Data structure with several items of the same data type

- Seems useful for us…
  - Data frame columns also contain the same type 🧐
  - Linear algebra is the basis for ML

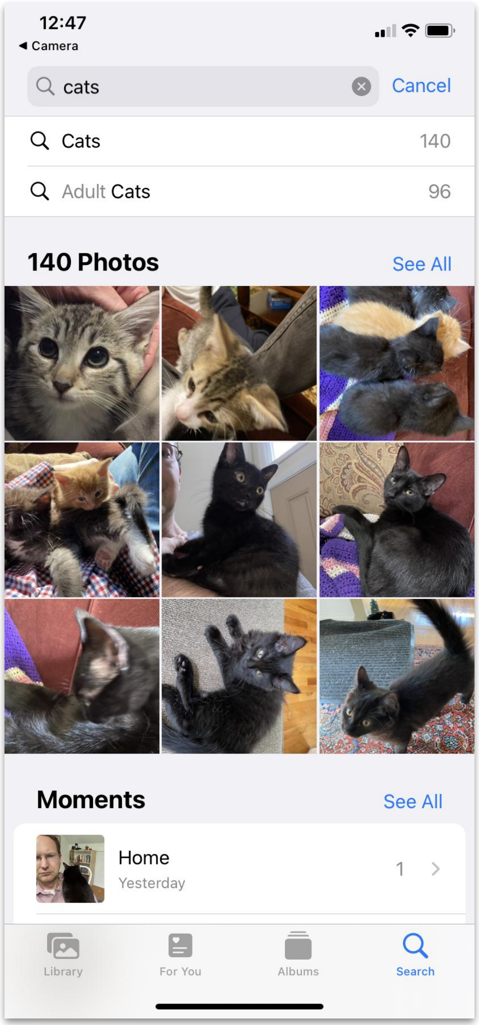
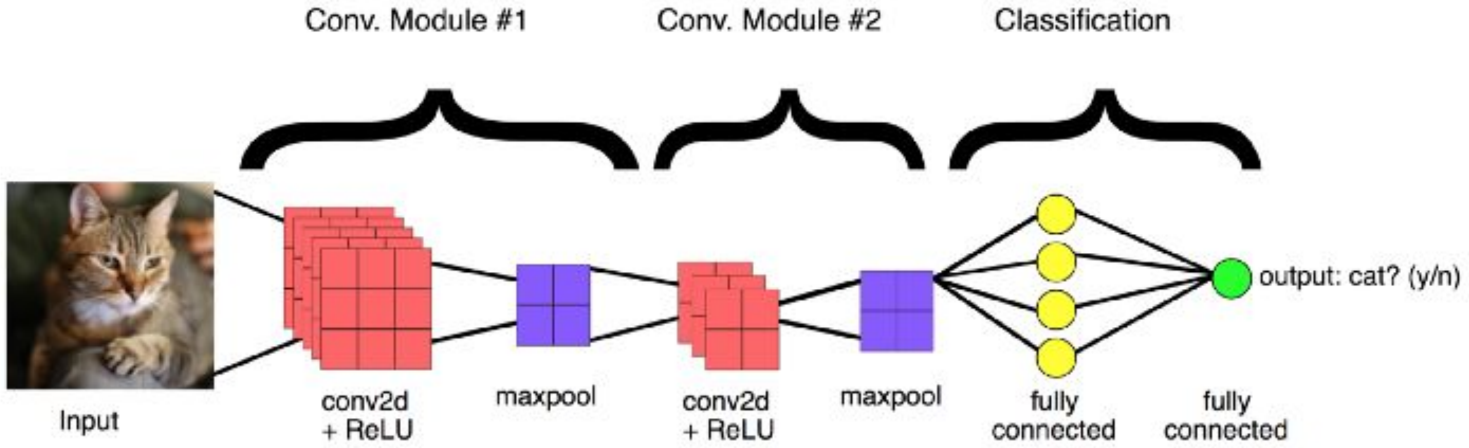


Photo credit: Prof. Mimno

I promise linear algebra is actually very cool and useful



https://towardsdatascience.com/introduction-to-convolutional-neural-network-cnn-de73f69c5b83

Arrays in Python

- We'll use a library!
- Why Numerical Python (NumPy)?
    - Fast & compact storage for arrays
- Do we really need to use NumPy for arrays?
    - Yes. See above + it's easy to use!

Arrays in NumPy

```
>>> import numpy as np

>>> a = np.array([1,2,3])
```

Arrays in NumPy

```
>>> import numpy as np

>>> a = np.array([1,2,3])
```

"nickname" could be anything, **np** is what you will see most often

Arrays in NumPy

```
>>> import numpy as np

>>> a = np.array([1,2,3])
```

just remember to be consistent

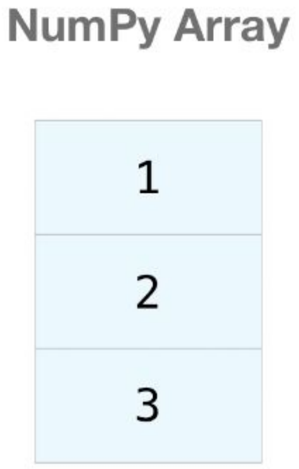
Arrays in NumPy

```
>>> import numpy as np
>>> a = np.array([1,2,3])
```

Draw: what does *a* look like?

Arrays in NumPy

```
>>> import numpy as np

>>> a = np.array([1,2,3])
```



https://numpy.org/devdocs/user/absolute_beginners.html

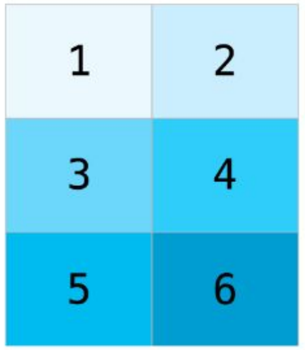
Arrays in NumPy

```
>>> import numpy as np
>>> a = np.array([[1,2],[3,4],[5,6]])
```

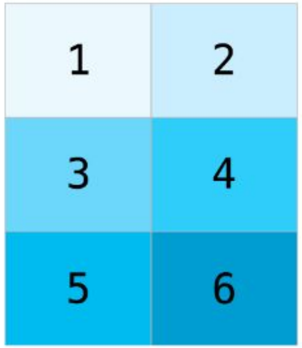
Draw: what does *a* look like?

Arrays in NumPy

```
>>> import numpy as np

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



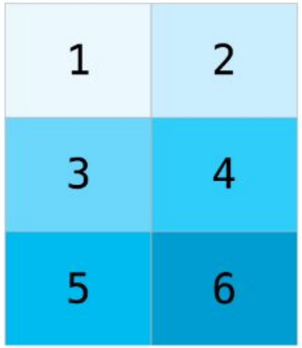
Arrays in NumPy



1	2
3	4
5	6

• # dimensions?	a.ndim
• Size? (# elements)	a.size
• Shape? (? , ?)	a.shape

Arrays in NumPy



# dimensions?	a.ndim	2
Size?	a.size	6
Shape?	a.shape	(3,2)

Arrays in NumPy

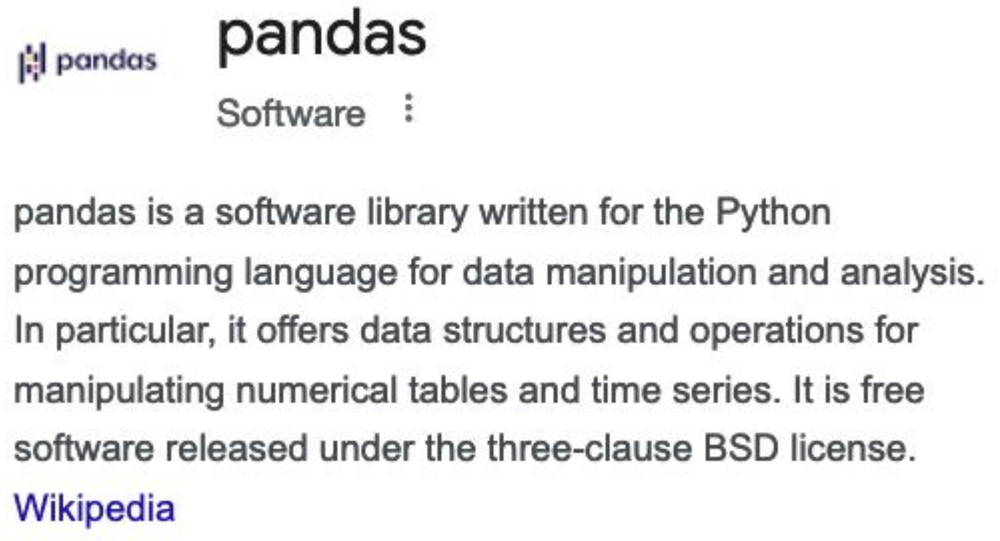
- Higher dimensions: hard for humans, easy for NumPy

```
>>> import numpy as np
>>> array_example = np.array([[[0, 1, 2, 3],
                               [4, 5, 6, 7]],
                              [[0, 1, 2, 3],
                               [4, 5, 6, 7]],
                              [[0 ,1 ,2, 3],
                               [4, 5, 6, 7]]])
```

Arrays in NumPy

- NumPy skill drills in HW1 for practice:
    - Index and slice (just like lists)
    - Arithmetic operations (across arrays, within arrays)
    - Reshape arrays

Another main library needed for data science:



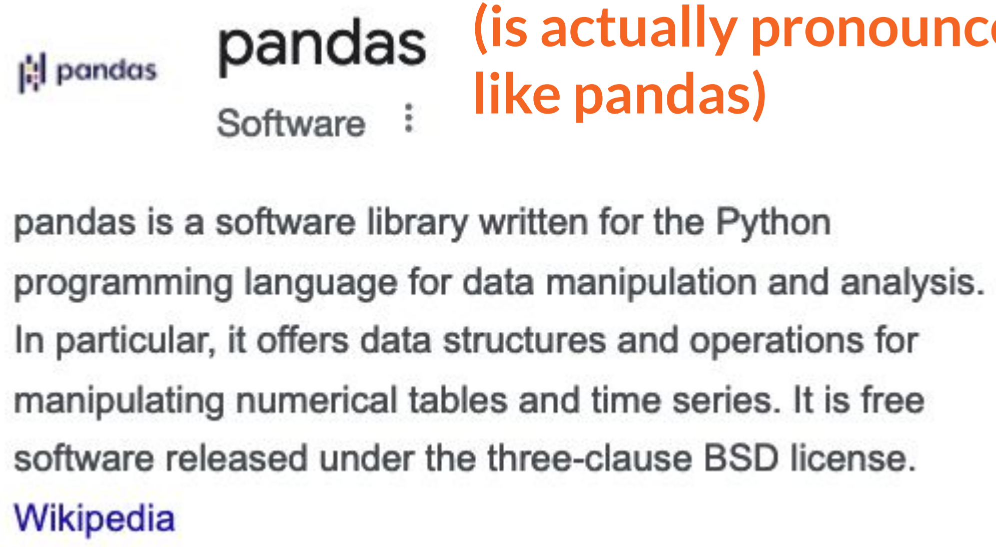
pandas

Software

pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Wikipedia

Another main library needed for data science:



pandas

Software

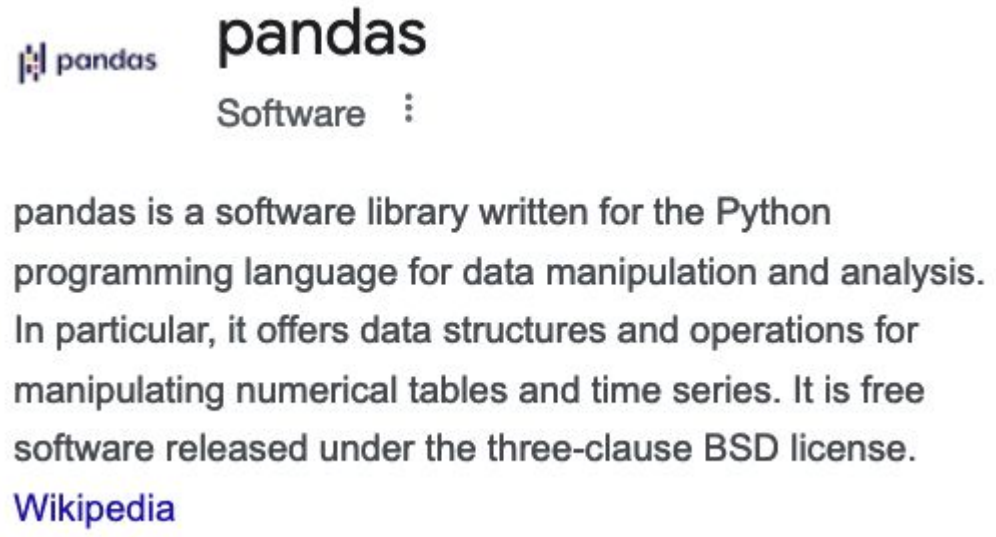
pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Wikipedia

(is actually pronounced like pandas)

https://www.stratascratch.com/blog/how-to-import-pandas-as-pd-in-python/

Another main library needed for dataframes, specifically


pandas

Software

pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

Wikipedia

Libraries for data science

```
>>> import numpy as np

>>> import pandas as pd

>>> a = np.array([1,2,3])
```

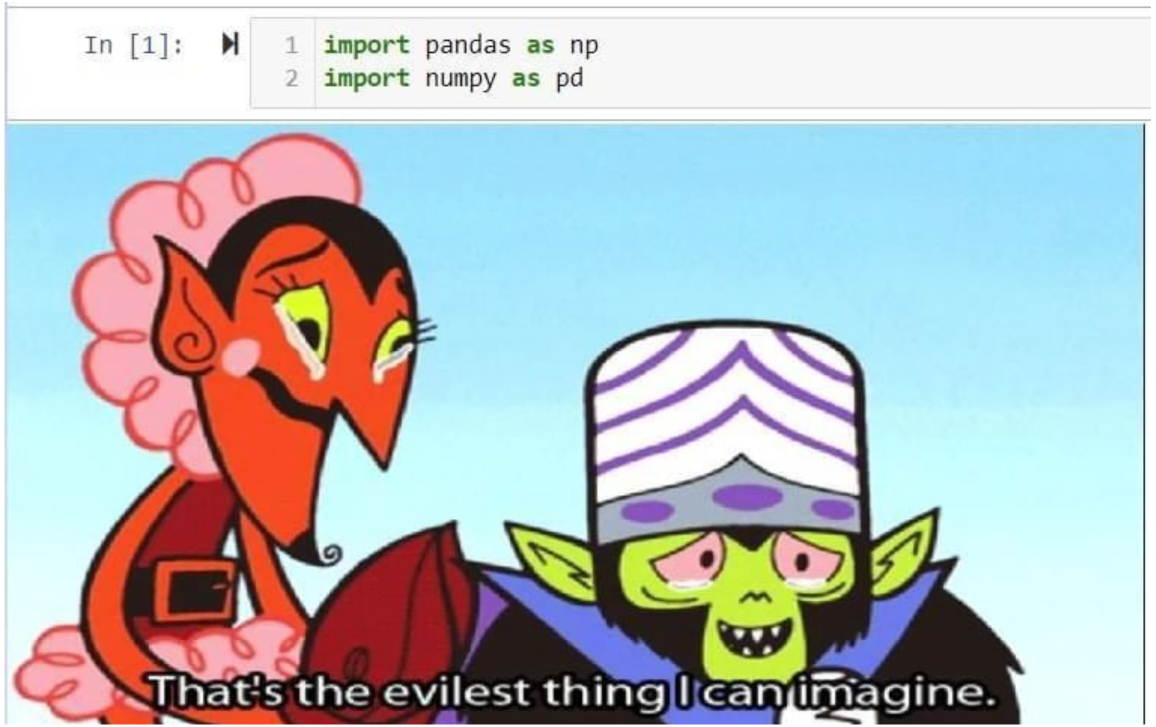
$\left.\right\}$ Best practice: import all your packages at the top of your ipynb

Libraries for data science

```
>>> import numpy as np
>>> import pandas as pd
>>> a = np.array([1,2,3])
```

standard abbreviation for pandas

Explain the meme

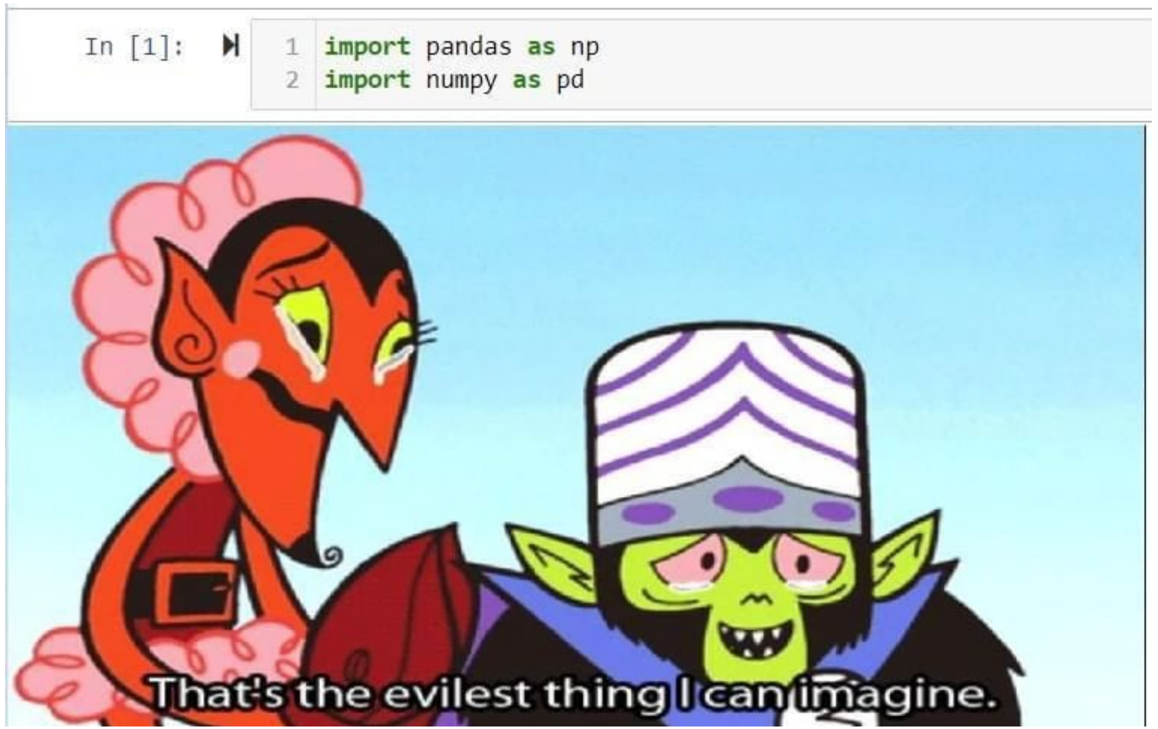


```
In [1]:
import pandas as np
import numpy as pd
```

https://br.ifunny.co/picture/in-import-pandas-as-np-import-numpy-as-pd-that-fNbcBYeW9

Explain the meme

Now typing np.array() will give you a pandas array, and will mess with every programmer's muscle memory expecting np to be numpy



```
In [1]:
import pandas as np
import numpy as pd
```

https://br.ifunny.co/picture/in-import-pandas-as-np-import-numpy-as-pd-that-fNbcBYeW9

1 min break & attendance!

tinyurl.com/3vjrcnws

Looking at data

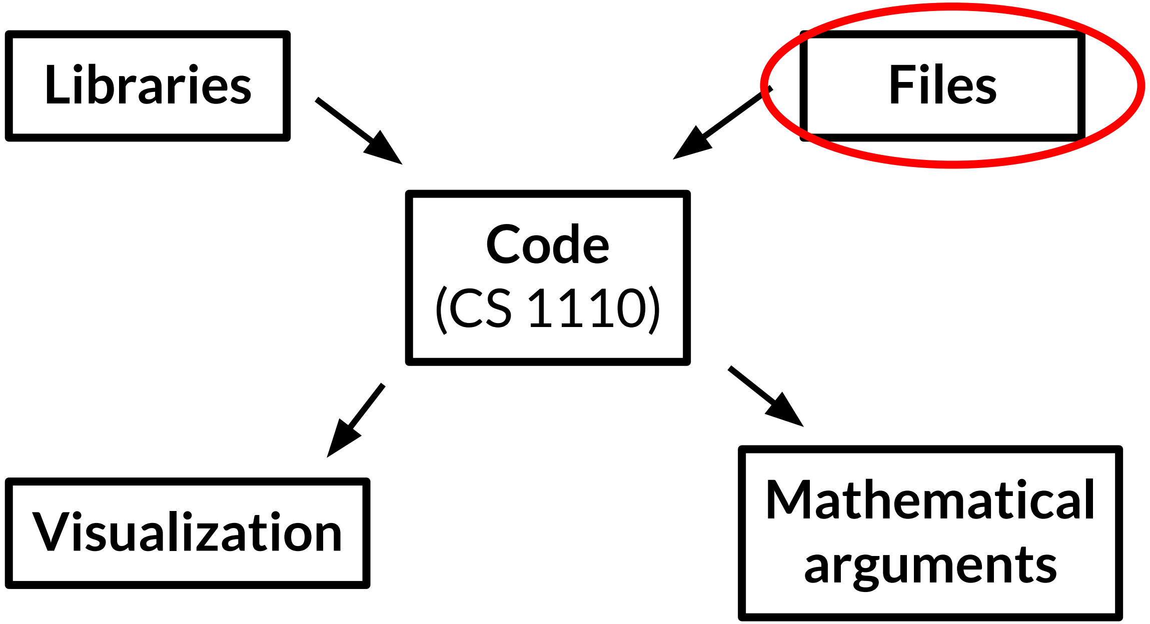
How to find out data attributes from an array:

```
>>> import numpy as np

>>> a = np.array([[1,2],[3,4],[5,6]])

>>> a.shape
```

Looking at data

How to find out data attributes from an array:

```
>>> import numpy as np

>>> [what if a is in a .csv file?]

>>> a.shape
```


Libraries
Files
Code
(CS 1110)
Visualization
Mathematical arguments

Importing data

- Use the Pandas library for dataframes, including csv imports/exports

```
>>> import pandas as pd

>>> a = pd.read_csv(‘data.csv’)

>>> a.shape
```

Importing data

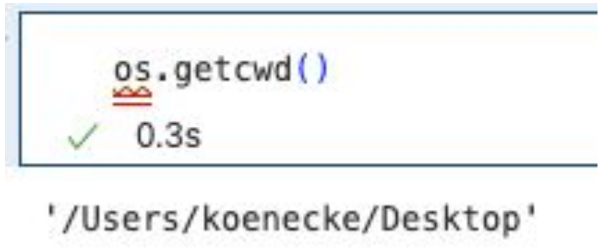
- Use the Pandas library for dataframes, including csv imports/exports

```
>>> import pandas as pd
>>> a = pd.read_csv($\boxed{\text{(‘data.csv’)}}$
>>> a.shape
```

Where is this file?

File systems

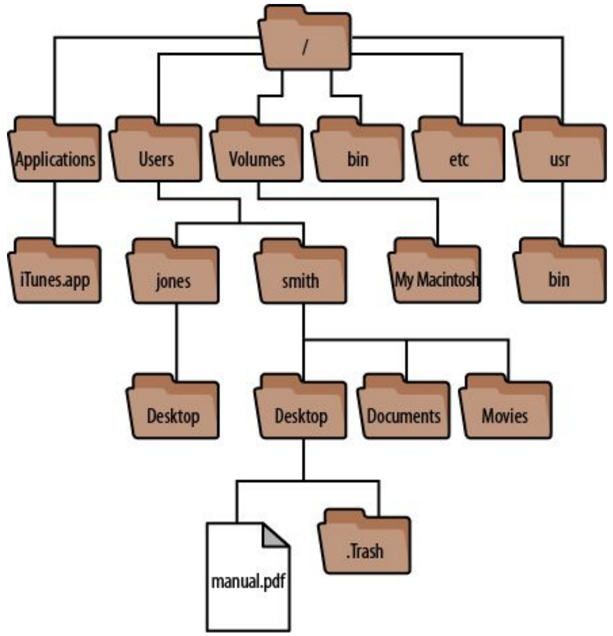
- Your **working directory** (wd) is where your Python script is operating
  - My ipynb is saved on my desktop, so:



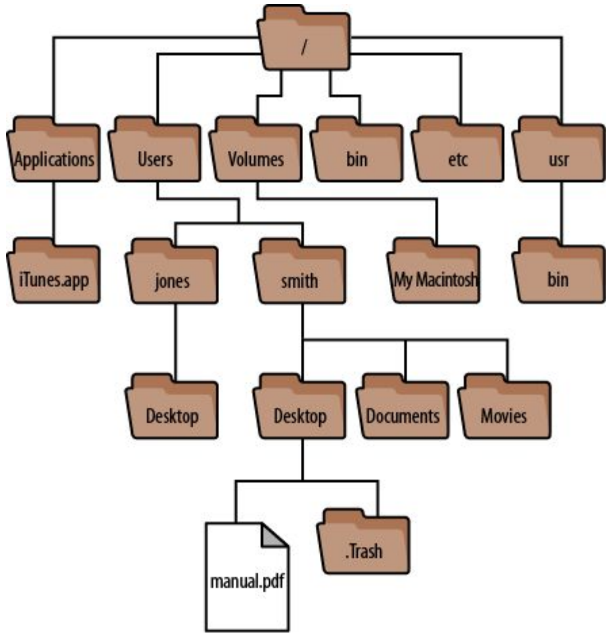
```
os.getcwd()
✓ 0.3s

'/Users/koenecke/Desktop'
```

- What does *'/Users/koenecke/Desktop'* mean?

Hierarchical file systems

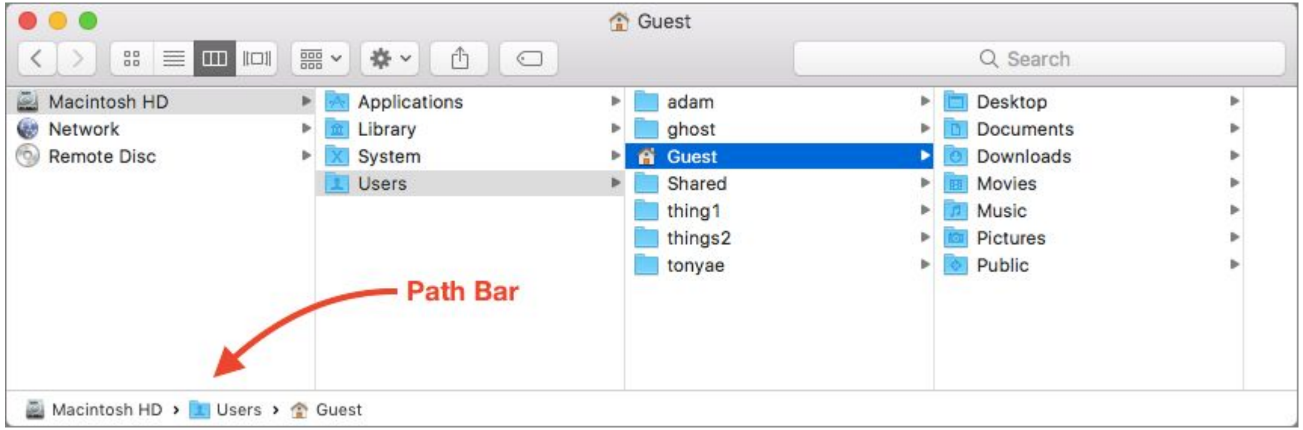


Hierarchical file systems



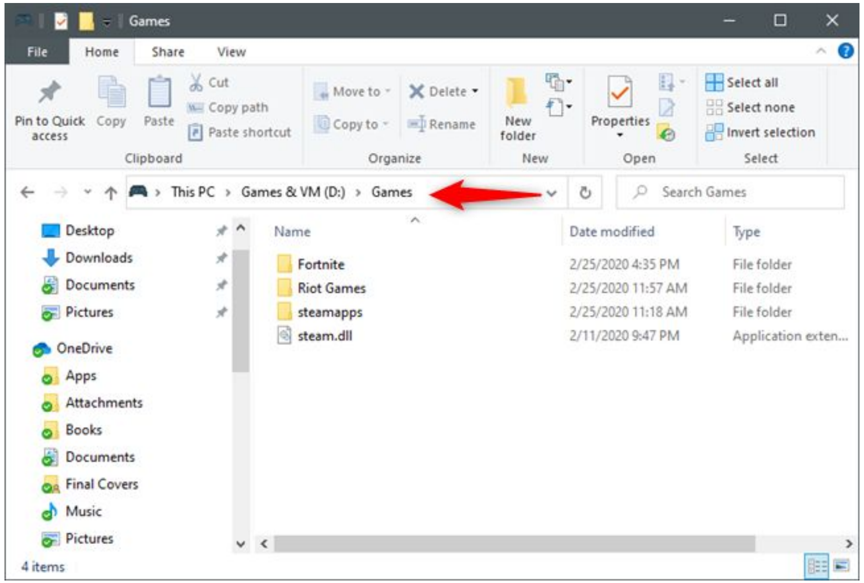
https://whiz.id/en/how-to-make-a-family-tree/ https://www.oreilly.com/library/view/macintosh-terminal-pocket/9781449328962/ch01s03s01.html

Hierarchical file systems



https://mac-fusion.com/finder-secrets-navigating-your-folder-hierarchy-with-the-path-bar/

Hierarchical file systems



Name	Date modified	Type
Fortnite	2/25/2020 4:35 PM	File folder
Riot Games	2/25/2020 11:57 AM	File folder
steamapps	2/25/2020 11:18 AM	File folder
steam.dll	2/11/2020 9:47 PM	Application exten...

https://www.digitalcitizen.life/how-export-directory-tree-folder-windows/

Question:

- My current working directory is *'/Users/ezra/Desktop/2023'*
- I have a file called *'info.csv'* saved in the directory that is the **parent** directory of my working directory
- **What filepath would I use to load *'info.csv'*?**

Question:

- My current working directory is *'/Users/ezra/Desktop/2023'*
- I have a file called *'info.csv'* saved in the directory that is the **parent** directory of my working directory
- **What filepath would I use to load *'info.csv'*?**
  - ***'/Users/ezra/Desktop/info.csv'***

Data frames

```
>>> import pandas as pd
>>> a = pd.read_csv('data.csv')
>>> a.shape
```

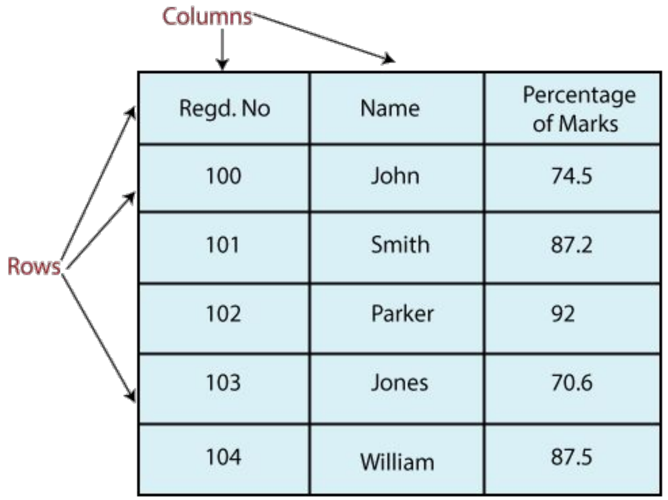
Data frames

```
>>> import pandas as pd
>>> a = pd.read_csv('data.csv')
>>> a.shape
```

Pandas automatically decides each column's data types!

Data frames

- **Recall:** rows & columns, less expressive than spreadsheets

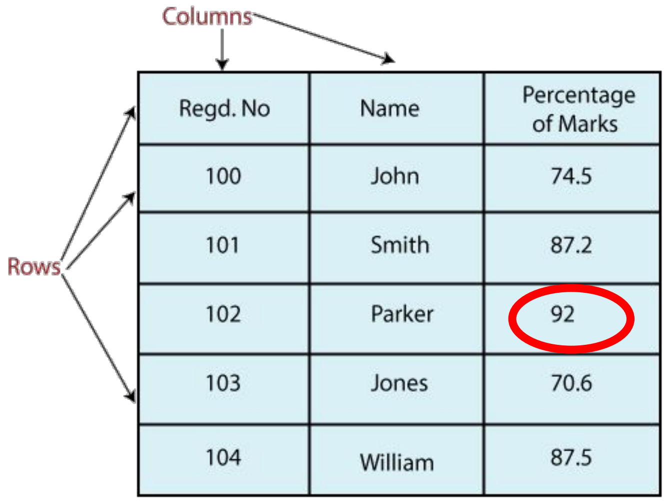


Regd. No	Name	Percentage of Marks
100	John	74.5
101	Smith	87.2
102	Parker	92
103	Jones	70.6
104	William	87.5

https://www.javatpoint.com/python-pandas-dataframe

What's wrong here?

- **Recall:** rows & columns, less expressive than spreadsheets

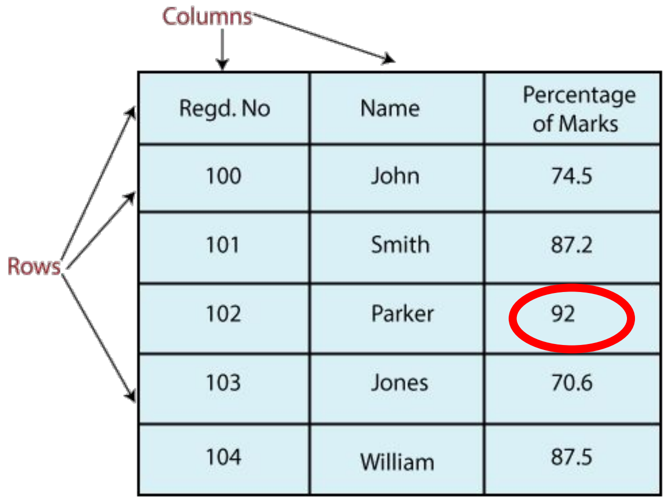

Columns
Rows

Regd. No	Name	Percentage of Marks
100	John	74.5
101	Smith	87.2
102	Parker	92
103	Jones	70.6
104	William	87.5

https://www.javatpoint.com/python-pandas-dataframe

What's wrong here?

- **Recall:** rows & columns, less expressive than spreadsheets

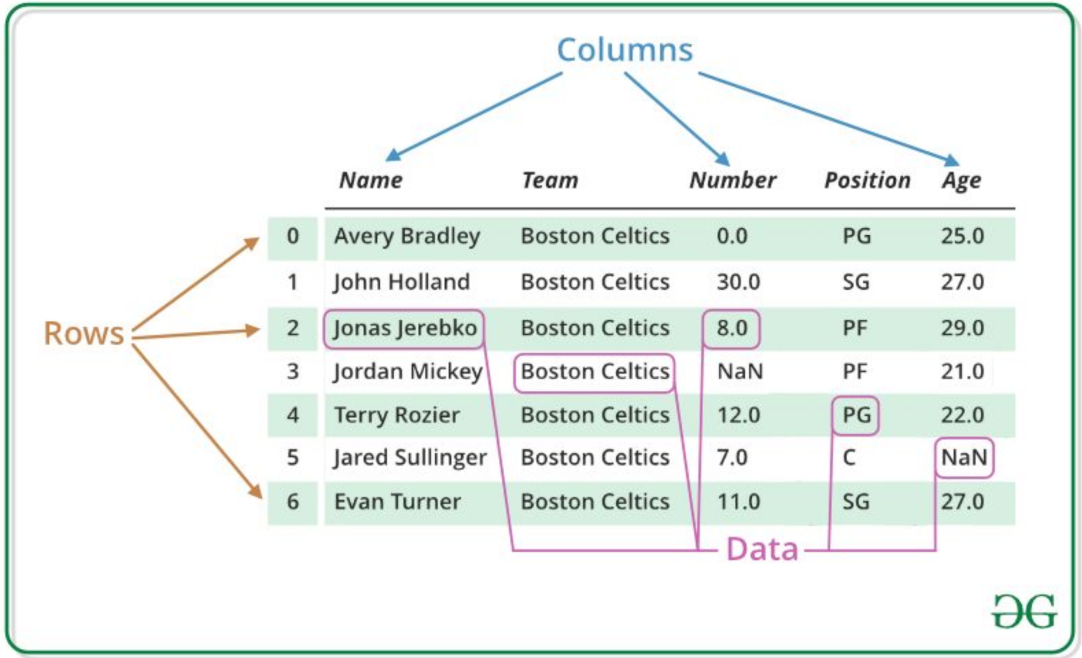

Columns
Rows

Regd. No	Name	Percentage of Marks
100	John	74.5
101	Smith	87.2
102	Parker	92
103	Jones	70.6
104	William	87.5

Int, not a float like the rest of the column!

https://www.javatpoint.com/python-pandas-dataframe

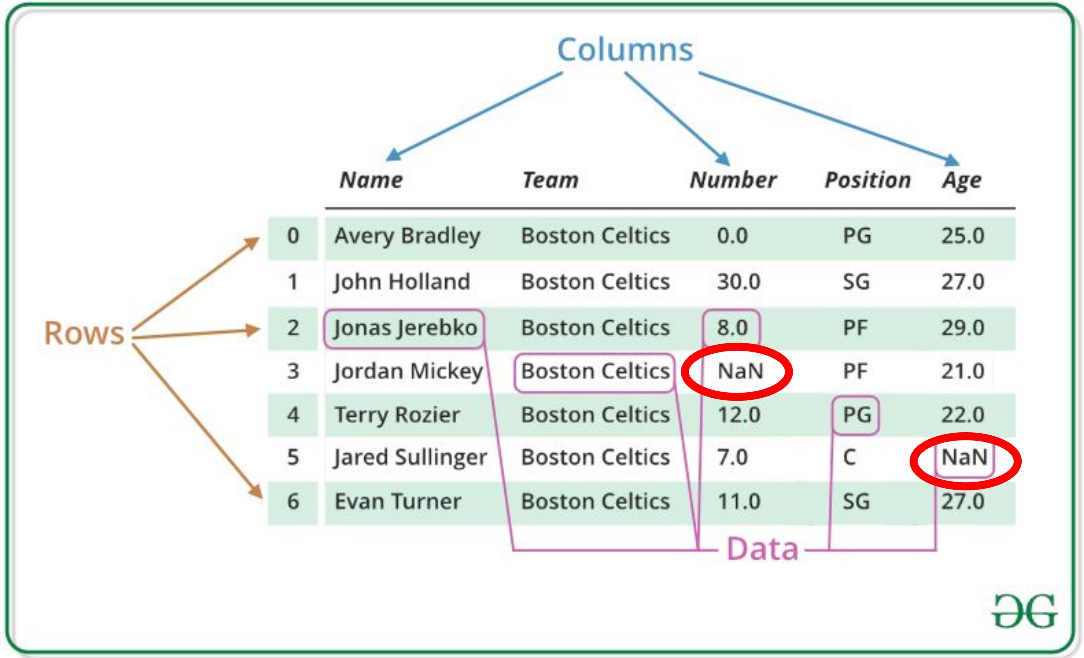
Data Frame (df)



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

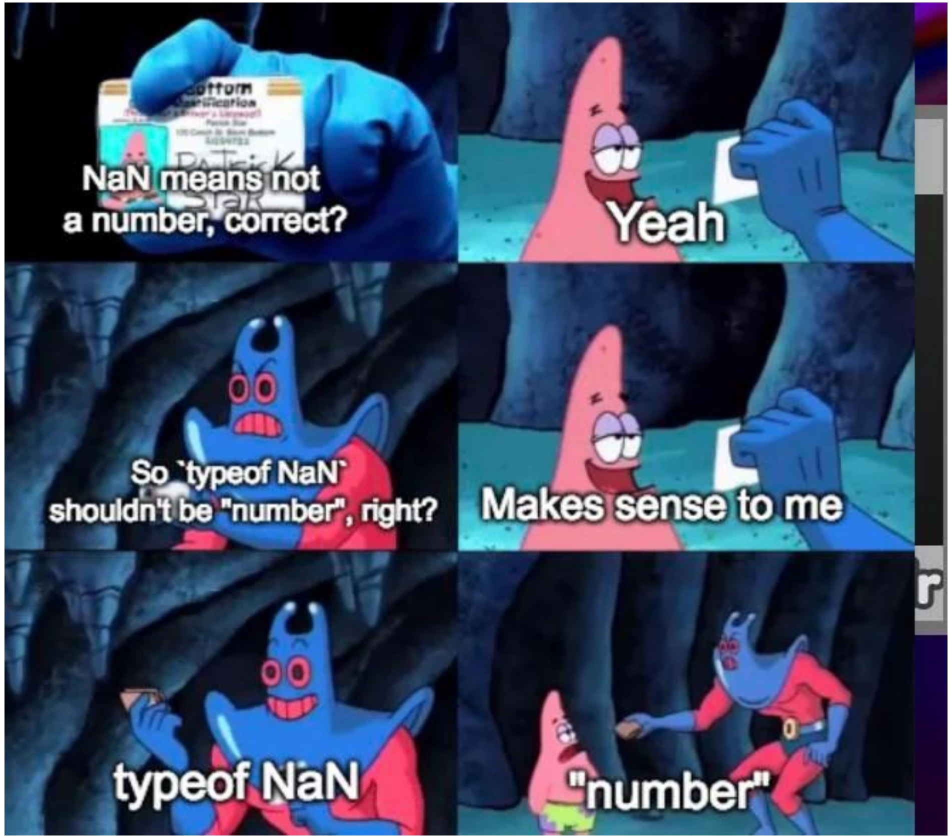
https://www.geeksforgeeks.org/python-pandas-dataframe/

Data Frame (df)



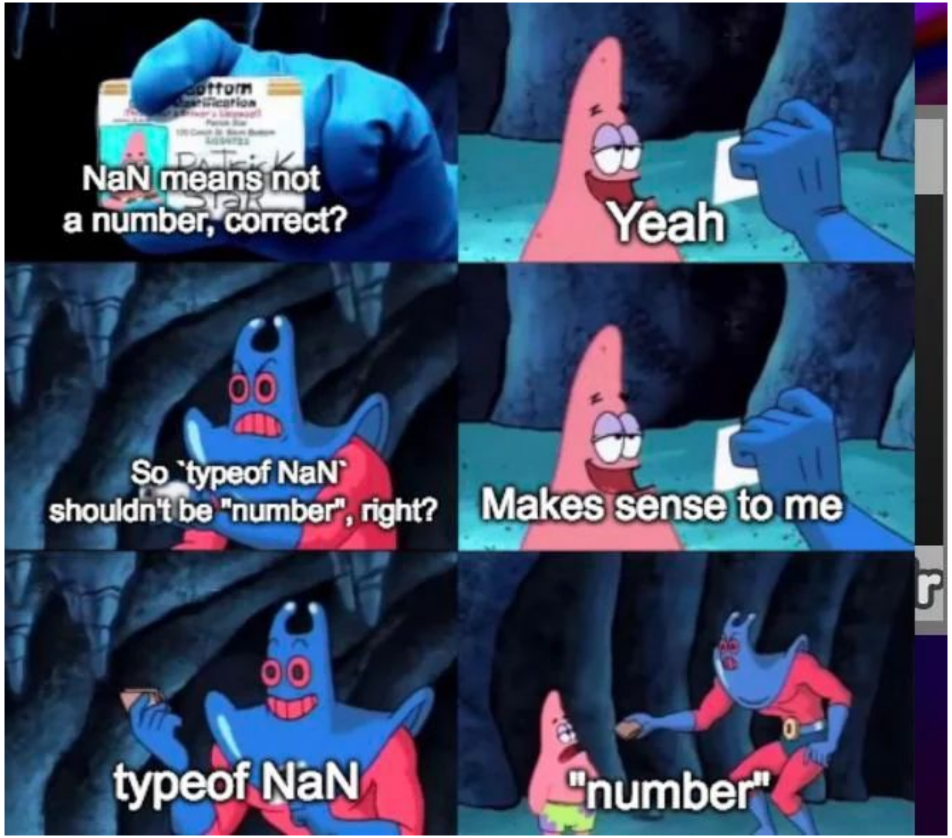
	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

https://www.geeksforgeeks.org/python-pandas-dataframe/


NaN means not a number, correct?
Yeah
So `typeof NaN` shouldn't be "number", right?
Makes sense to me
typeof NaN
"number"

https://www.reddit.com/r/ProgrammerHumor/comments/veqzm1/type_of_nan/

NaN is a float value in Python, so you need to be careful when checking for missing data!!


NaN means not a number, correct?
Yeah
So `typeof NaN` shouldn't be "number", right?
Makes sense to me
typeof NaN
"number"

https://www.reddit.com/r/ProgrammerHumor/comments/veqzm1/type_of_nan/

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`

  - What's the difference?

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`

  - What's the difference?

  - If a is an array of integers…

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`
  - What’s the difference?
  - If a is an array of integers…

```
<class ‘numpy.ndarray’>
```

```
int64
```

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`

  - What's the difference?

  - If **a** is a value (e.g., 5)…

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`

  - What's the difference?
  - If **a** is a value (e.g., 5)…

int

Error: 'int' object has no attribute 'dtype'

Data type conversions

- To check what data type something is:

  `type(a)` versus `a.dtype`

  - What's the difference?

- To convert among str, int, float, bool…
  - Can do this for individual data values, e.g.: `float(a)`
  - Can do this across entire arrays, e.g.: `a.astype(float)`

Show of hands...

```
>>> a = np.array([[1,2],[3,4],[5,6]])
>>> b = 1.0*a
>>> b.dtype
```

- **Is this legal? What happens?**

Yes it’s legal!

```
>>> a = np.array([[1,2],[3,4],[5,6]])
>>> b = 1.0*a
>>> b.dtype
```

- **Output: dtype('float64')**

Question: data types

- What is bool(“False”)?
- What is bool(-100)?

Question: data types

- What is bool(“False”)? **Syntax Error**
- What is bool(-100)? **True**

Question: data types

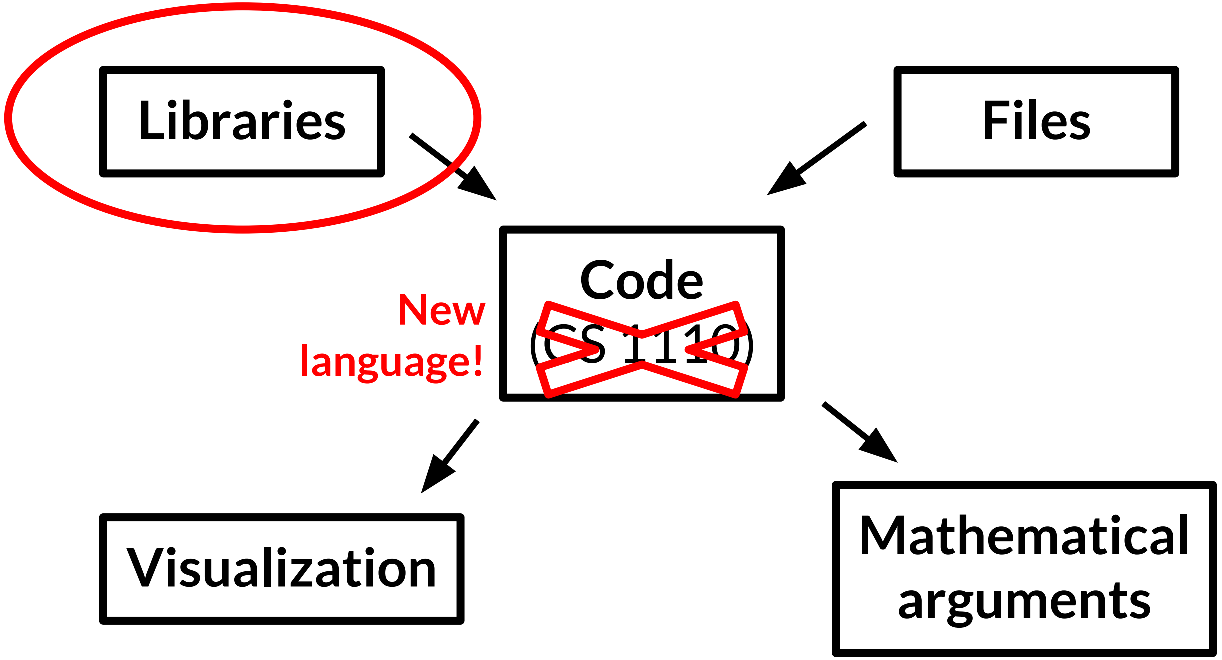
- What is 5 / 2 in **`Python 2`**?
- What about in **`Python 3`**?

Question: data types

- What is 5 / 2 in **`Python 2`**? **2**
- What about in **`Python 3`**? **2.5**

🚨 WARNING 🚨

- Be careful about:
  - Lossy conversion from float → int
  - Converting to Boolean
  - Check: do you have NaNs?
  - Check: do you get errors?


Libraries
Files
Code
~~(CS 1110)~~
New language!
Visualization
Mathematical arguments

Advertisement for new course this semester!

INFO 2310: Interactive Web Application Design and Development

- Introduction to the conceptual, design, and technical aspects of making interactive web applications.
- MERN Stack: MongoDB, Express.js, React, Node.js
- Tues/Thu 10:10am-11:25am, Fri 10:10-11am
- Seats available for IS/ISST majors and IS affiliating students!
- Satisfies category B in Digital Culture and Production concentration and counts as IS major elective.
- Email info2310@cornell.edu to enroll!

1 min break + Interview question

Why should you be suspicious of the number 1,048,576?

Interview question

Why should you be suspicious of the number 1,048,576?

Max # rows in Excel. Files are often bigger, but if opened in Excel to convert to csv, you lose rows!

SQL

https://www.reddit.com/r/ProgrammerHumor/

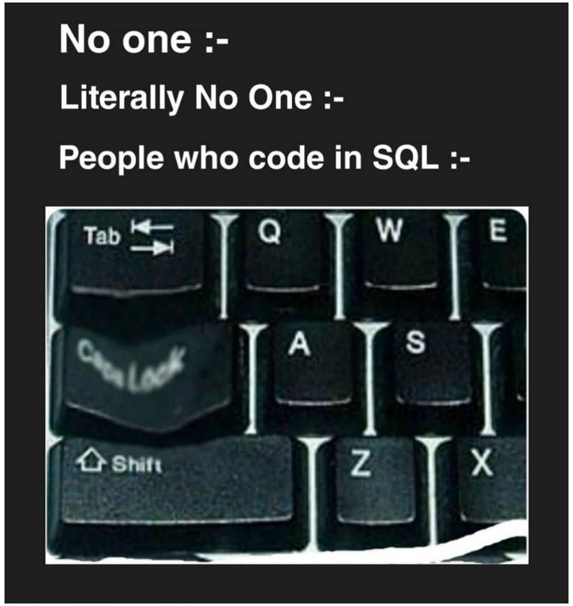
SQL

- **S**tructured **Q**uery **L**anguage
    - A staple question in **data science interviews**
    - Used to interact with databases
- ‘*DuckDB*’ package allows you to write SQL code in Python env
- `>>> import duckdb`

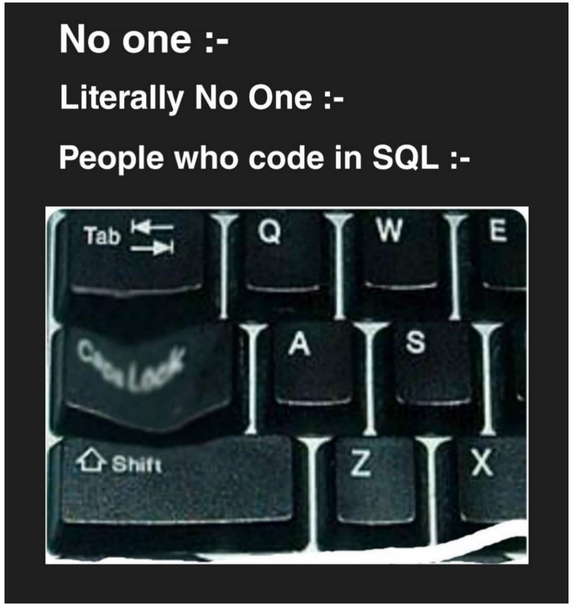
Filtering your data

- Start with a data frame (called a “table” in SQL)
- Making your table smaller = “filtering”
- SQL commands:
    - “SELECT” = Restrict by columns
    - “WHERE” = Restrict by rows

SQL

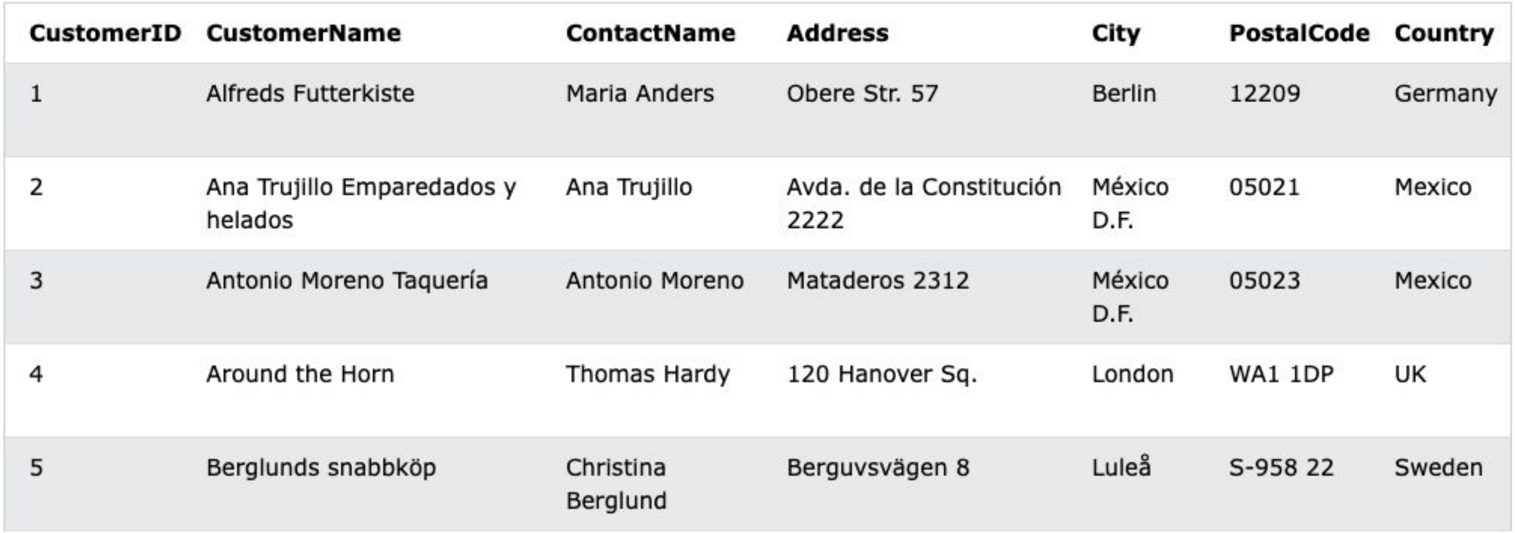

No one :-
Literally No One :-
People who code in SQL :-
Tab
Q
W
E
Caps Lock
A
S
Shift
Z
X

SQL



```
SELECT column1, column2, ...
FROM table_name
WHERE condition;
```

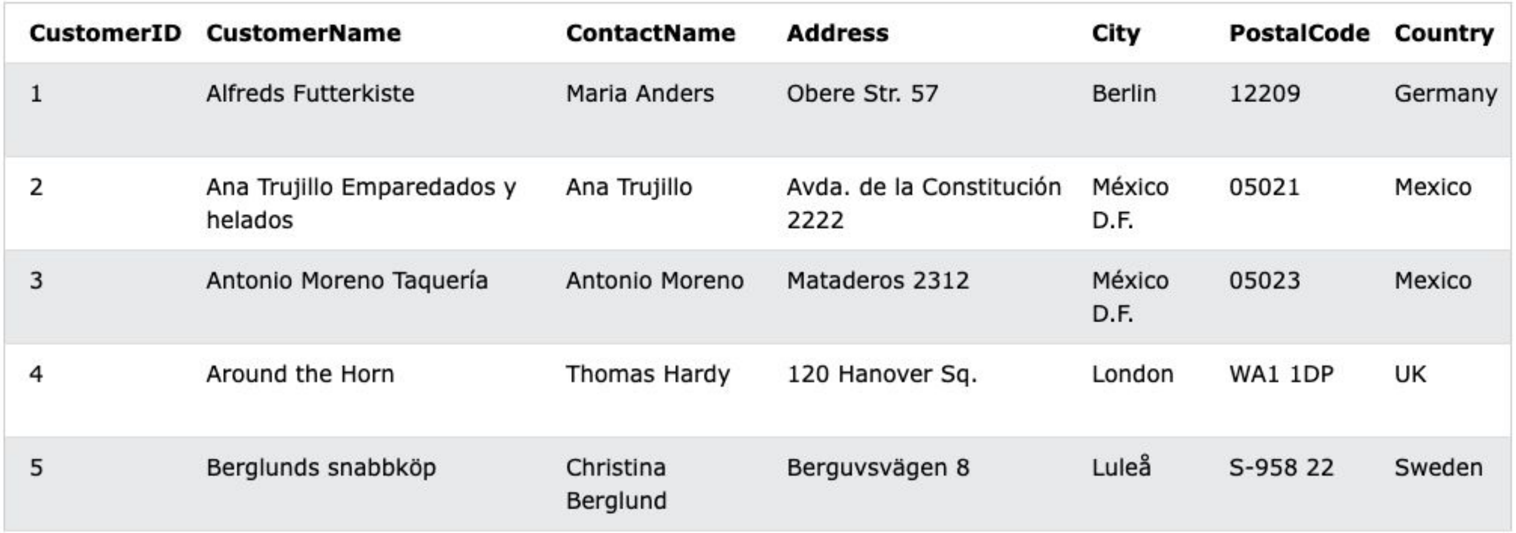
Table named `Customers`



CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

Table named `Customers`

FROM Customers;

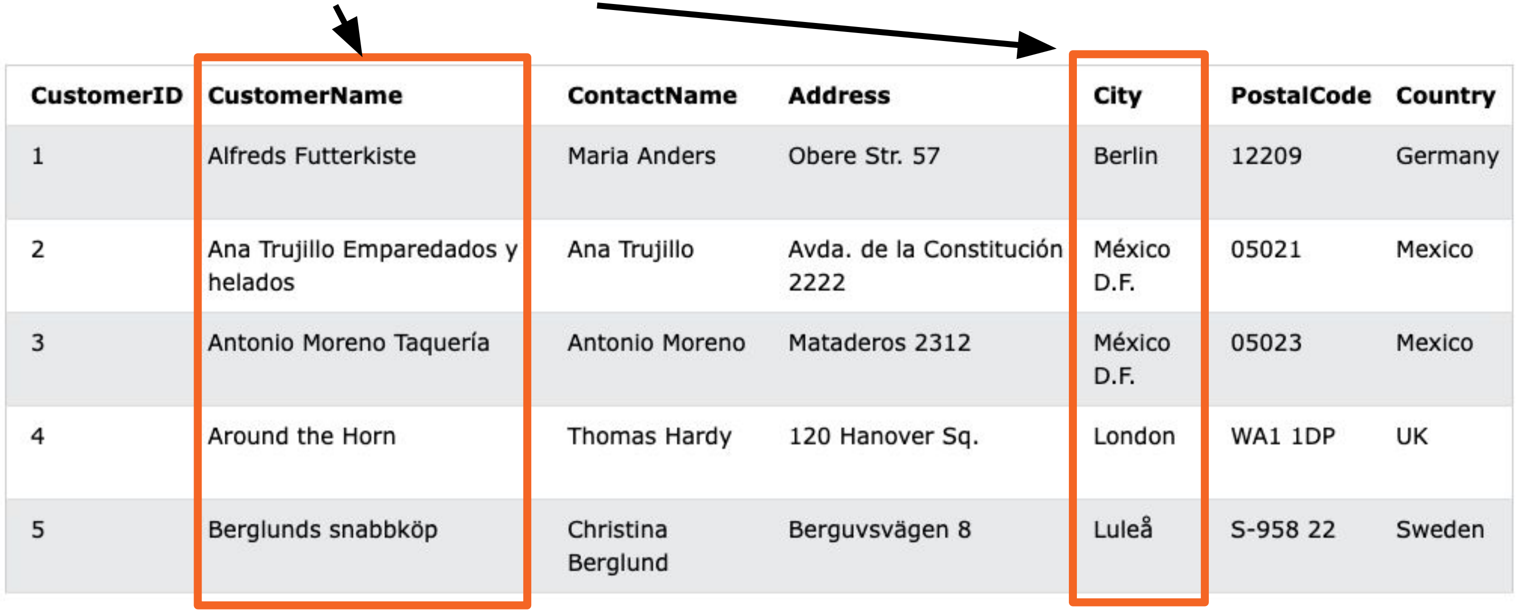


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

Table named `Customers`

SELECT CustomerName, City **FROM** Customers;

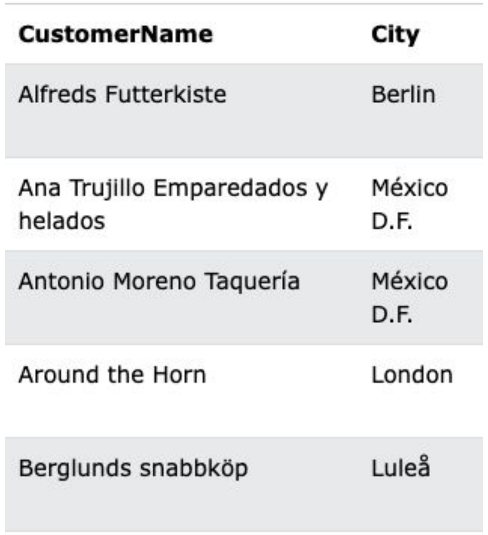


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

Table named `Customers`

SELECT CustomerName, City **FROM** Customers;

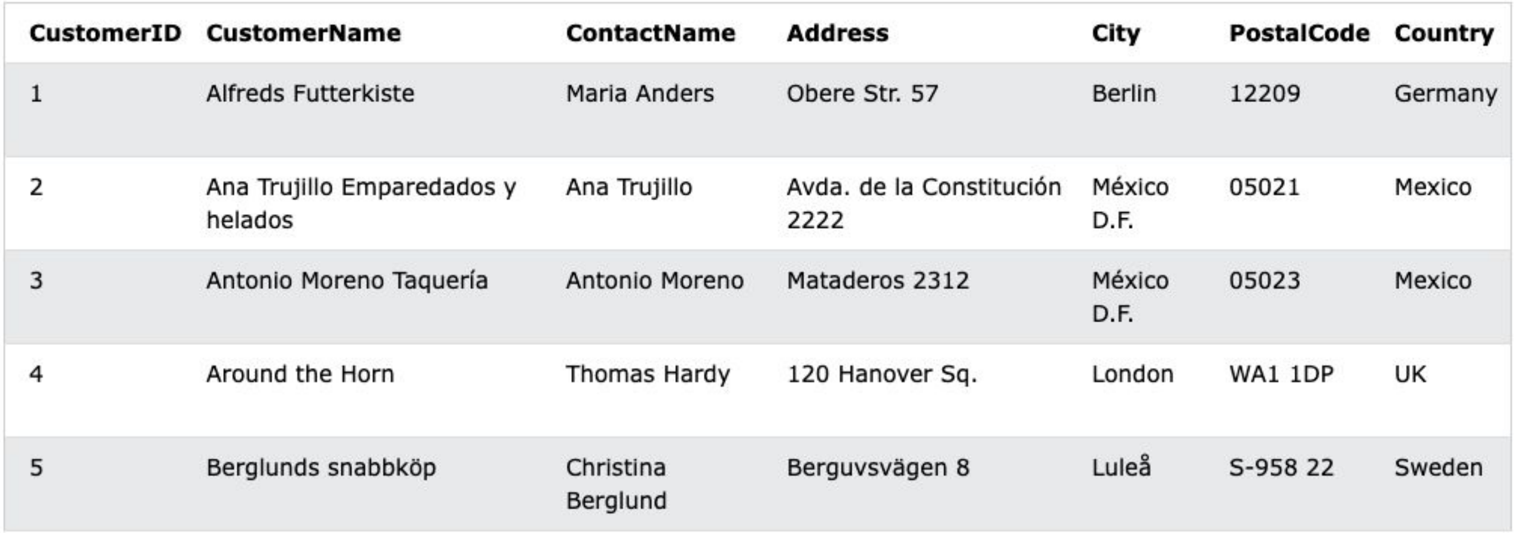
This new table is the output of our SQL query!



CustomerName	City
Alfreds Futterkiste	Berlin
Ana Trujillo Emparedados y helados	México D.F.
Antonio Moreno Taquería	México D.F.
Around the Horn	London
Berglunds snabbköp	Luleå

Table named `Customers`

SELECT * FROM Customers;



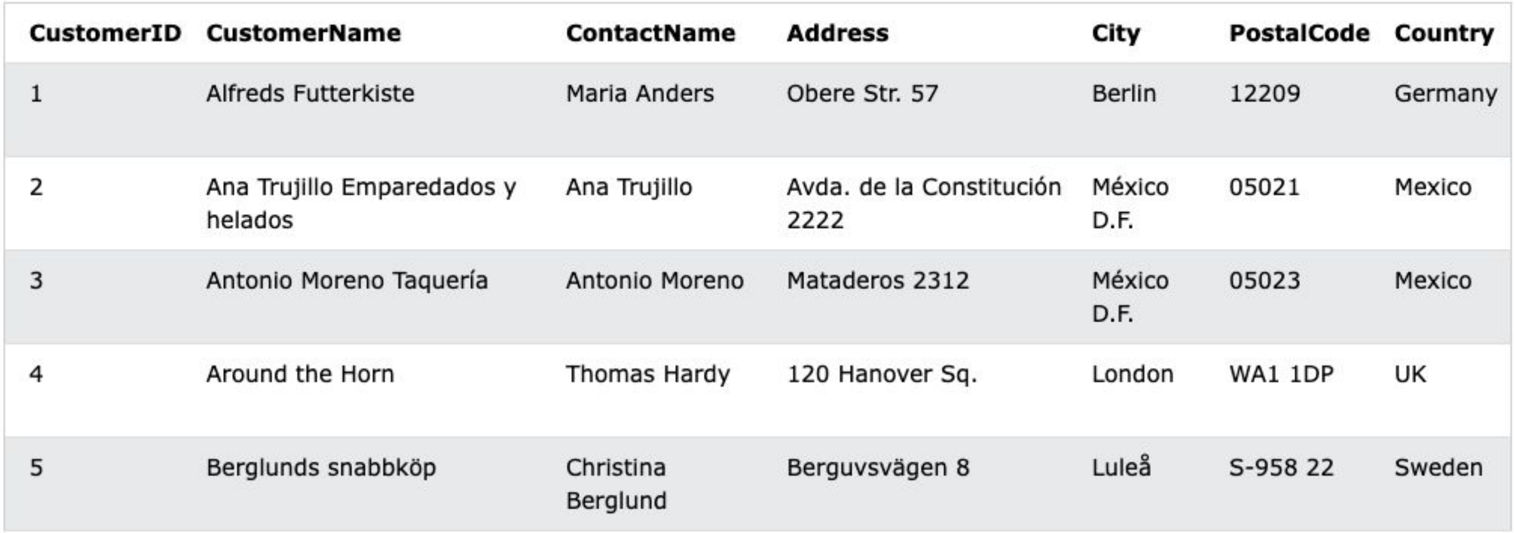
CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

Table named `Customers`

```sql
SELECT * FROM Customers;
```

* is a wildcard token that selects all columns



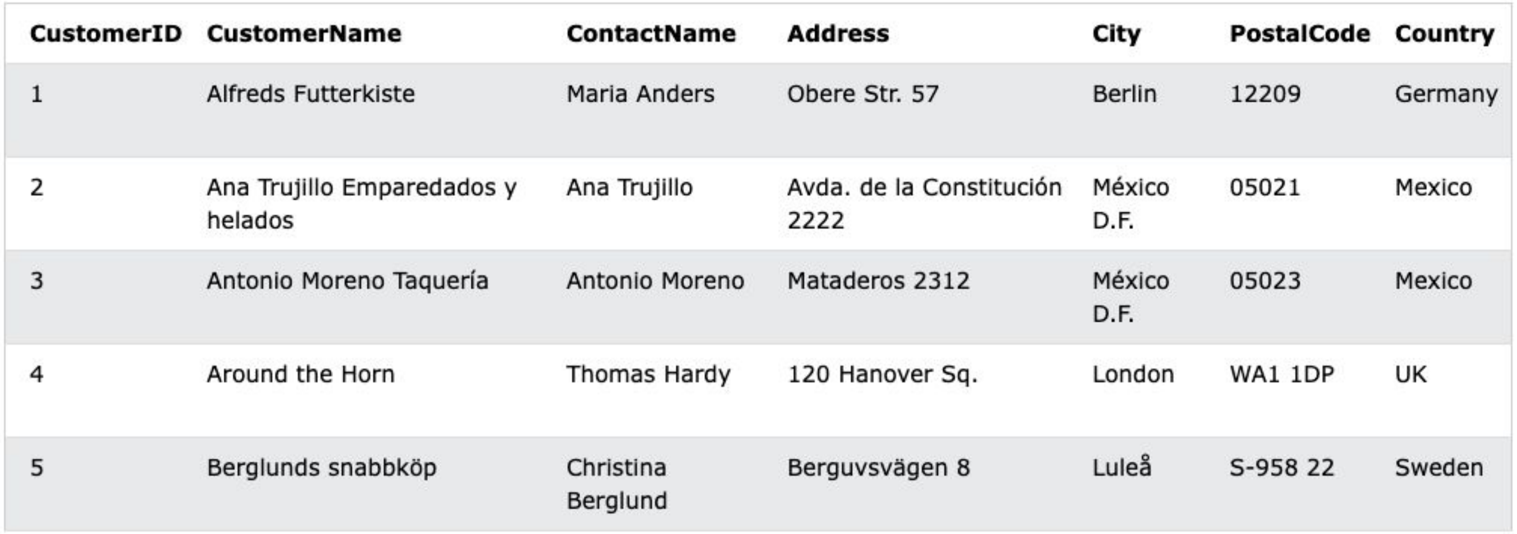
CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

https://medium.com/google-cloud/sql-is-structured-query-language-but-the-structure-can-be-different-in-different-databases-3f38b90c319

Table named `Customers`

```
SELECT * FROM Customers WHERE Country='Mexico';
```

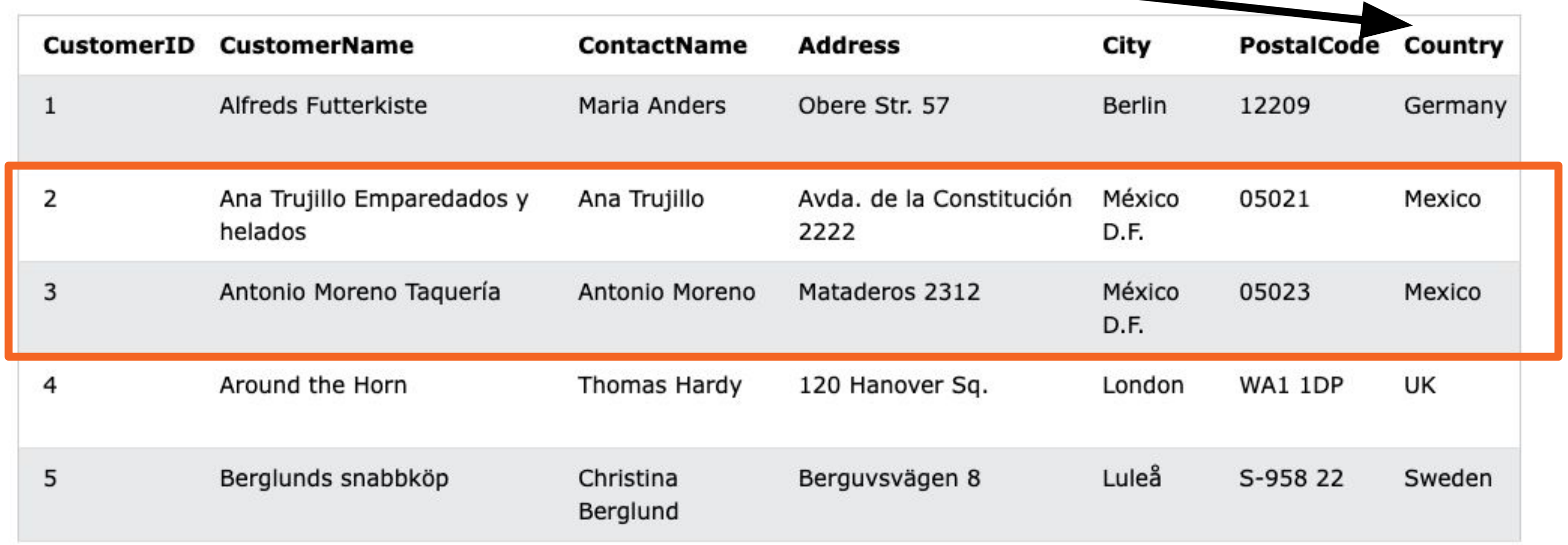


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

Table named `Customers`

```
SELECT * FROM Customers WHERE Country='Mexico';
```

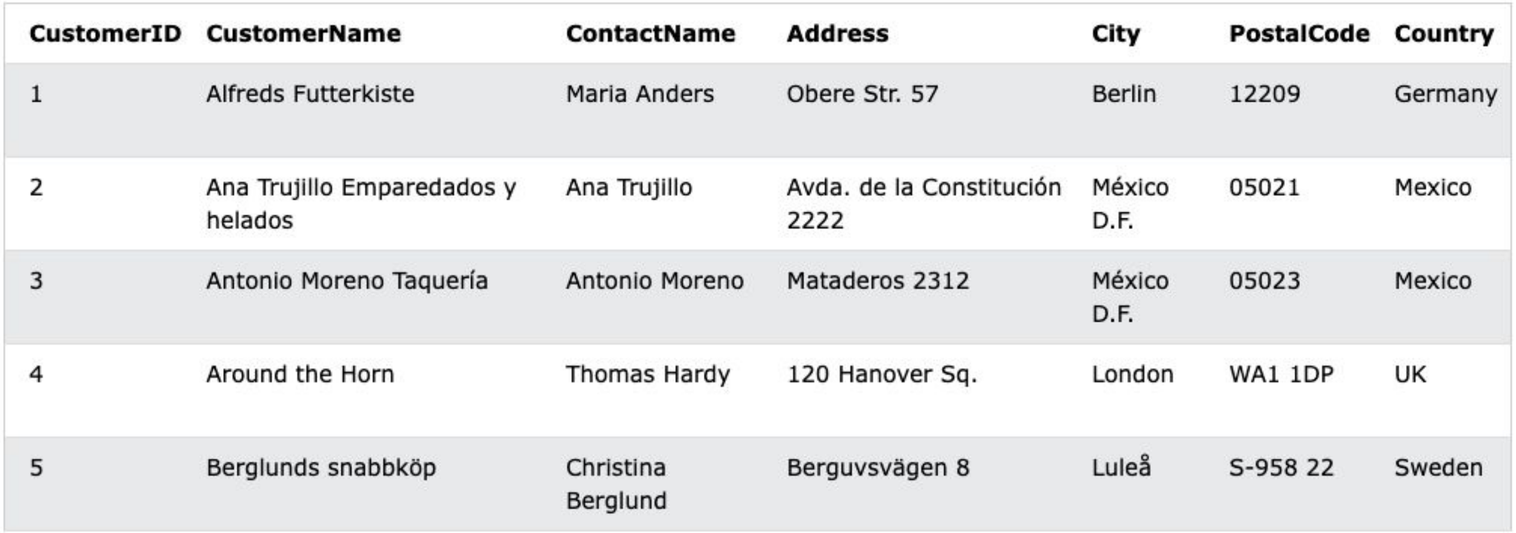


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

Table named `Customers`

Be careful! Single = in SQL is double == in Python

```
SELECT * FROM Customers WHERE Country='Mexico';
```

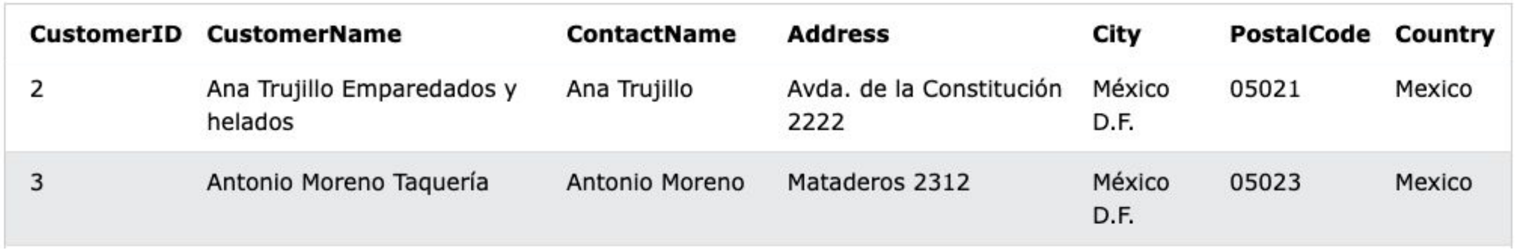


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

Table named `Customers`

```
SELECT * FROM Customers WHERE Country='Mexico';
```

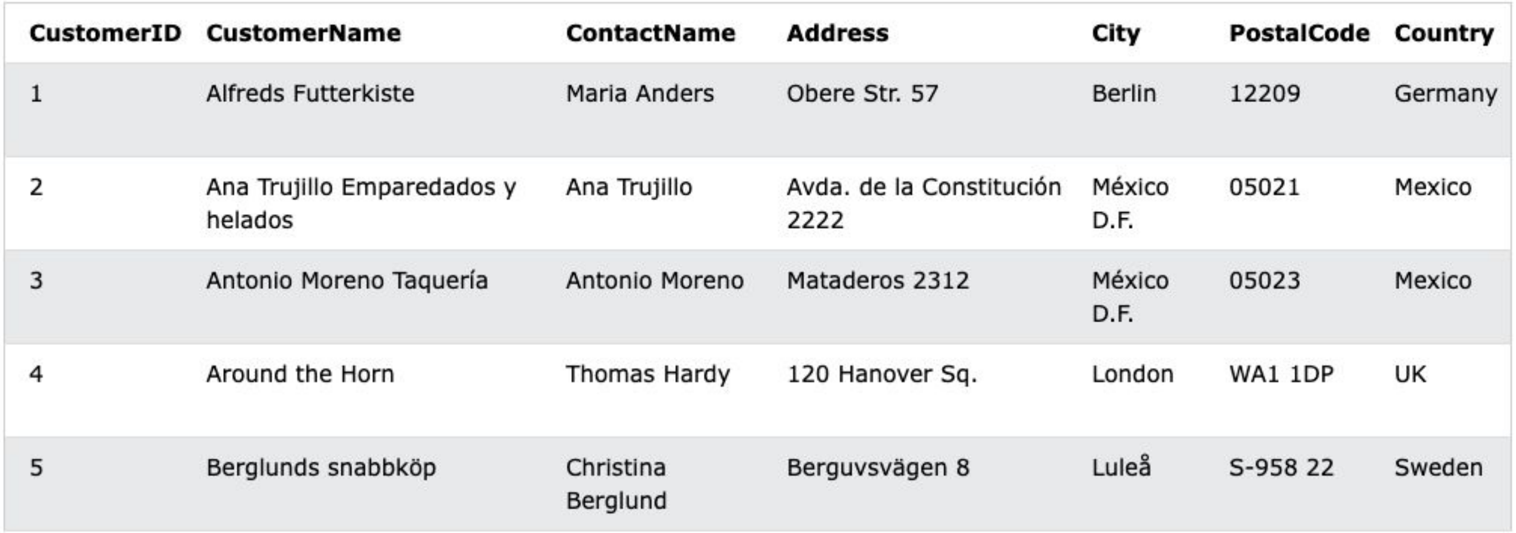


CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico

This new table is the output of our SQL query!

How many rows and cols in result?

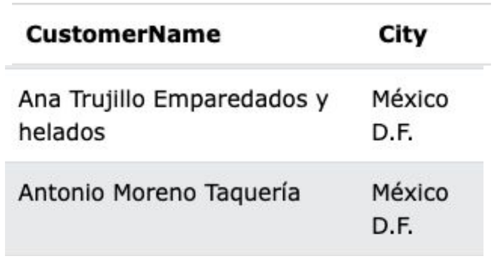
```
SELECT CustomerName, City FROM Customers WHERE Country='Mexico';
```



CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

https://www.w3schools.com/sql/sql_examples.asp

```
SELECT CustomerName, City FROM Customers WHERE Country='Mexico';
```



CustomerName	City
Ana Trujillo Emparedados y helados	México D.F.
Antonio Moreno Taquería	México D.F.

2 rows, 2 columns

Filtering your data

- Are there non-SQL ways to filter your data frame?
- Yes, but they’re messier to write & harder to read
- SQL commands are transferable logic to other statistical programming languages
  - R
  - SAS

Manipulating your data

- Table “Fruits” shows quarterly profit for each product
- **How do we get each fruit’s half-year profit?**

Product	**Q1**	**Q2**
Apple	$100	$20
Banana	$50	$2
Cantaloupe	$600	$500

Manipulating your data

- To define a new column using existing columns, use arithmetic operators for values and **AS** for new column name

- **SELECT** Product, Q1 + Q2 **AS** H1 **FROM** Fruits;

Product	**Q1**	**Q2**
Apple	$100	$20
Banana	$50	$2
Cantaloupe	$600	$500

Product	**H1**
Apple	$120
Banana	$52
Cantaloupe	$1100

Manipulating your data

- **Draw the output:**
- **SELECT** 2*Q1 **AS** DoubledQ1 **FROM** Fruits;

Product	**Q1**	**Q2**
Apple	$100	$20
Banana	$50	$2
Cantaloupe	$600	$500

Table named `Fruits`

- **SELECT** 2*Q1 **AS** DoubledQ1 **FROM** Fruits;

Product	**Q1**	**Q2**
Apple	$100	$20
Banana	$50	$2
Cantaloupe	$600	$500

➡

DoubledQ1
$200
$100
$1200

Manipulating your data

- Today: how to sum **across** columns → more columns (“manipulating”)

- Next time: how to sum **within/across** columns → a single stat! (“summarizing” / “aggregating”)

Interview Question: data types

- What does print(1.81e308) display?
- Why?

Interview Question: data types

- What does print(1.81e308) display?
  inf
- Why? **System overflow at 2^{1024}**

Question: what do the below have in common?

- arithmetic with *nan*
- sqrt(x) for negative x
- 1e314 - 1e314

Question: what do the below have in common?

- arithmetic with *nan*
- sqrt(x) for negative x
- 1e314 - 1e314

They all yield **nan**!

HW1

- Install Python 3 Anaconda by Friday 08/25/2022
    - Recommend VSCode for IDE

- Be able to answer:
    - What version of Python did you install?
    - What is a conda environment?
    - Why might you use different conda environments?
    - What IDE are you using?

- Can use slip days (10 total for the semester)

1. Cap your marker
2. Return marker & whiteboards to each of their bins
3. Throw your tissues in the trash