NumPy

pandas

NumPy

pandas

Lets talk about…

Python Packages

Python Data Analysis Library

"When working with tabular data, such as data stored in spreadsheets or databases, pandas is the right tool for you. Pandas will help you to explore, clean and process your data."

— The Official Pandas Documentation

Python Data Analysis Library

> “When working with tabular data, such as data stored in spreadsheets or databases, pandas is the right tool for you. Pandas will help you to explore, clean and process your data.”
>
> — The Official Pandas Documentation

```
import pandas as pd
```

Pandas: Data Structures

Pandas: Data Structures

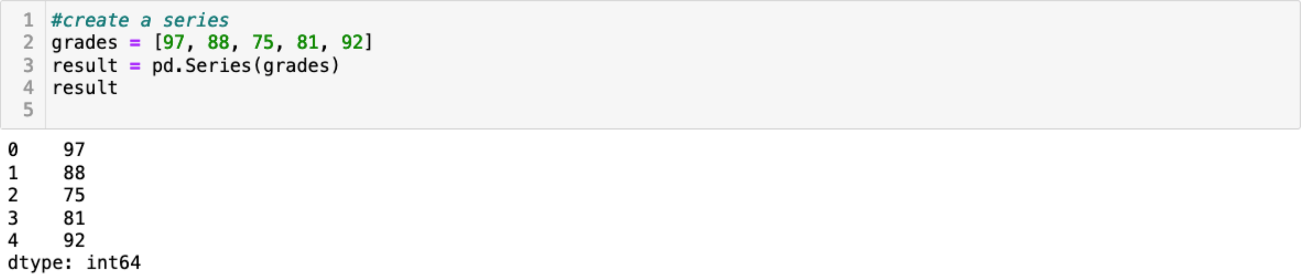
Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.



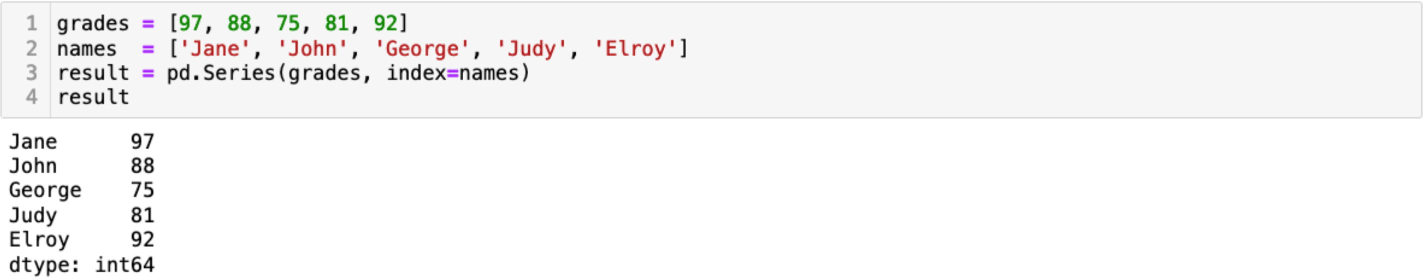
```
#create a series
grades = [97, 88, 75, 81, 92]
result = pd.Series(grades)
result

```

```
0    97
1    88
2    75
3    81
4    92
dtype: int64
```

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.



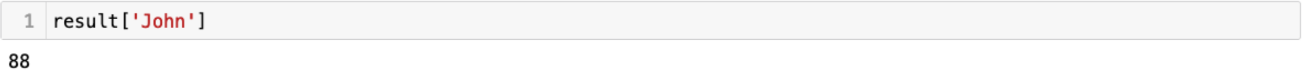
```python
grades = [97, 88, 75, 81, 92]
names  = ['Jane', 'John', 'George', 'Judy', 'Elroy']
result = pd.Series(grades, index=names)
result
```

```
Jane      97
John      88
George    75
Judy      81
Elroy     92
dtype: int64
```

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.



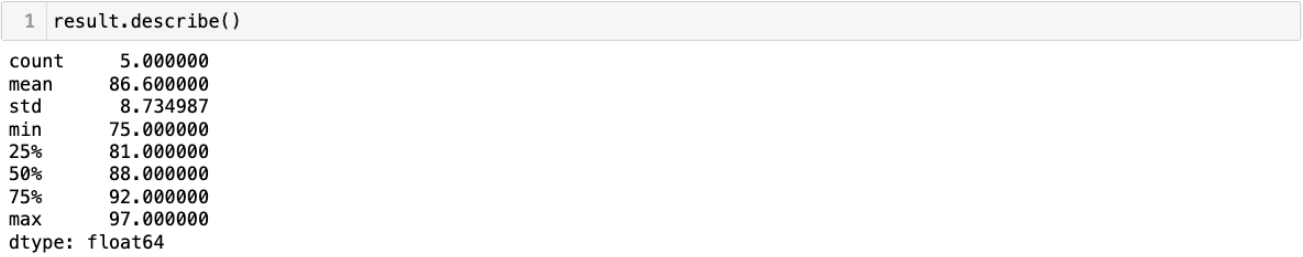
```
result['John']
```

```
88
```

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.



```
result.describe()
```

```
count     5.000000
mean     86.600000
std       8.734987
min      75.000000
25%      81.000000
50%      88.000000
75%      92.000000
max      97.000000
dtype: float64
```

Pandas: Data Structures

››› Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

DataFrame

A dataframe is a data structure that allows us to store tabular data. The columns in a dataframe is a series, and each column can represent a different data type.

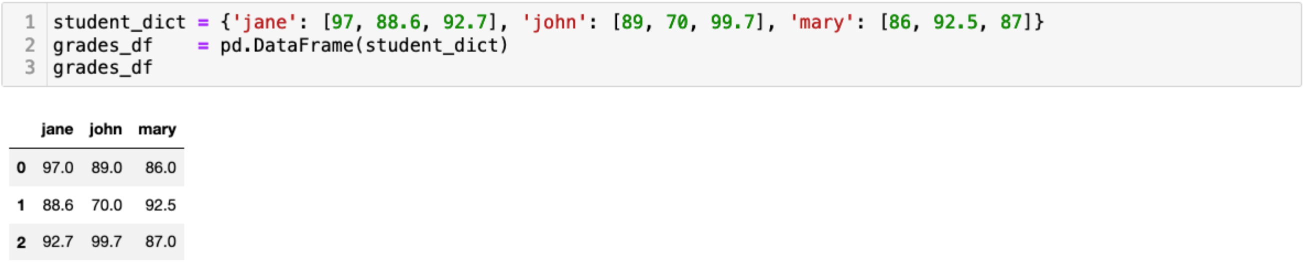
Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

DataFrame

A dataframe is a data structure that allows us to store tabular data. The columns in a dataframe is a series, and each column can represent a different data type.



```python
student_dict = {'jane': [97, 88.6, 92.7], 'john': [89, 70, 99.7], 'mary': [86, 92.5, 87]}
grades_df    = pd.DataFrame(student_dict)
grades_df
```

	jane	john	mary
0	97.0	89.0	86.0
1	88.6	70.0	92.5
2	92.7	99.7	87.0

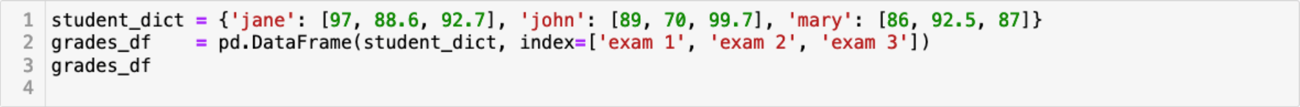
Pandas: Data Structures

Series

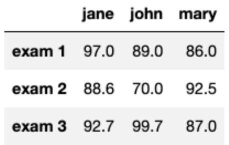
A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

DataFrame

A dataframe is a data structure that allows us to store tabular data. The columns in a dataframe is a series, and each column can represent a different data type.



```python
student_dict = {'jane': [97, 88.6, 92.7], 'john': [89, 70, 99.7], 'mary': [86, 92.5, 87]}
grades_df    = pd.DataFrame(student_dict, index=['exam 1', 'exam 2', 'exam 3'])
grades_df

```



	jane	john	mary
exam 1	97.0	89.0	86.0
exam 2	88.6	70.0	92.5
exam 3	92.7	99.7	87.0

Pandas: Data Structures

Series

A series is one-dimensional (array-like) data structure that contains a sequence of homogenous values. Each value in the sequence has an associated label, called an index.

DataFrame

A dataframe is a data structure that allows us to store tabular data. The columns in a dataframe is a series, and each column can represent a different data type.

Pandas: Data Ingestion

Pandas: Data Ingestion

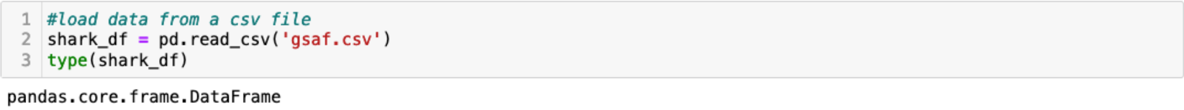
⋙ **Load the data**

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Pandas: Data Ingestion

Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.



```
#load data from a csv file
shark_df = pd.read_csv('gsaf.csv')
type(shark_df)
```

```
pandas.core.frame.DataFrame
```

Pandas: Data Ingestion

⪢ Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

⪢ Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

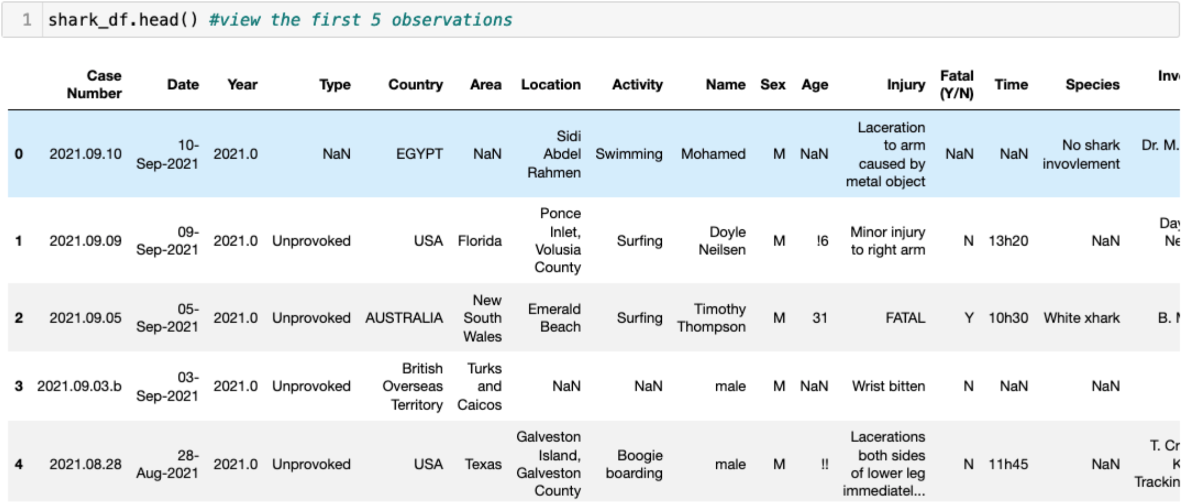
Pandas: Data Ingestion

Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.



```
shark_df.head() #view the first 5 observations
```

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Name	Sex	Age	Injury	Fatal (Y/N)	Time	Species	Inv
0	2021.09.10	10-Sep-2021	2021.0	NaN	EGYPT	NaN	Sidi Abdel Rahmen	Swimming	Mohamed	M	NaN	Laceration to arm caused by metal object	NaN	NaN	No shark invovlement	Dr. M.
1	2021.09.09	09-Sep-2021	2021.0	Unprovoked	USA	Florida	Ponce Inlet, Volusia County	Surfing	Doyle Neilsen	M	!6	Minor injury to right arm	N	13h20	NaN	Day Ne
2	2021.09.05	05-Sep-2021	2021.0	Unprovoked	AUSTRALIA	New South Wales	Emerald Beach	Surfing	Timothy Thompson	M	31	FATAL	Y	10h30	White xhark	B. N
3	2021.09.03.b	03-Sep-2021	2021.0	Unprovoked	British Overseas Territory	Turks and Caicos	NaN	NaN	male	M	NaN	Wrist bitten	N	NaN	NaN	
4	2021.08.28	28-Aug-2021	2021.0	Unprovoked	USA	Texas	Galveston Island, Galveston County	Boogie boarding	male	M	!!	Lacerations both sides of lower leg immediatel...	N	11h45	NaN	T. Cr K Trackin

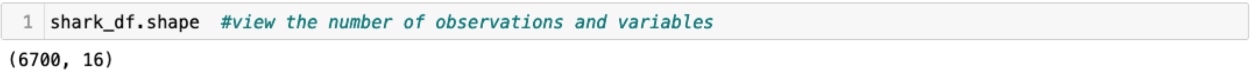
Pandas: Data Ingestion

Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.



```
shark_df.shape  #view the number of observations and variables
```

```
(6700, 16)
```

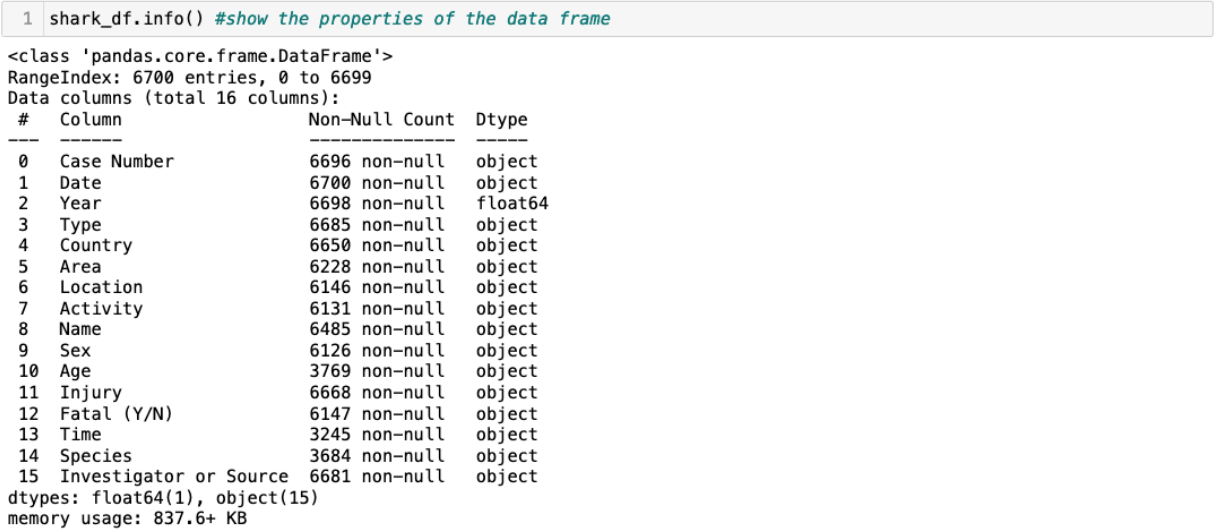
Pandas: Data Ingestion

Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.



```python
shark_df.info() #show the properties of the data frame
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6700 entries, 0 to 6699
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Case Number             6696 non-null   object 
 1   Date                    6700 non-null   object 
 2   Year                    6698 non-null   float64
 3   Type                    6685 non-null   object 
 4   Country                 6650 non-null   object 
 5   Area                    6228 non-null   object 
 6   Location                6146 non-null   object 
 7   Activity                6131 non-null   object 
 8   Name                    6485 non-null   object 
 9   Sex                     6126 non-null   object 
 10  Age                     3769 non-null   object 
 11  Injury                  6668 non-null   object 
 12  Fatal (Y/N)             6147 non-null   object 
 13  Time                    3245 non-null   object 
 14  Species                 3684 non-null   object 
 15  Investigator or Source  6681 non-null   object 
dtypes: float64(1), object(15)
memory usage: 837.6+ KB
```

Pandas: Data Ingestion

Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

Pandas: Data Ingestion

»» Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

»» Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

»» Accessing data

The data frame is made up of observations (rows) and variables (columns). We can select one or more variables and/or observations from the data frame using techniques like slicing and subsetting.

Pandas: Data Ingestion

Load the data

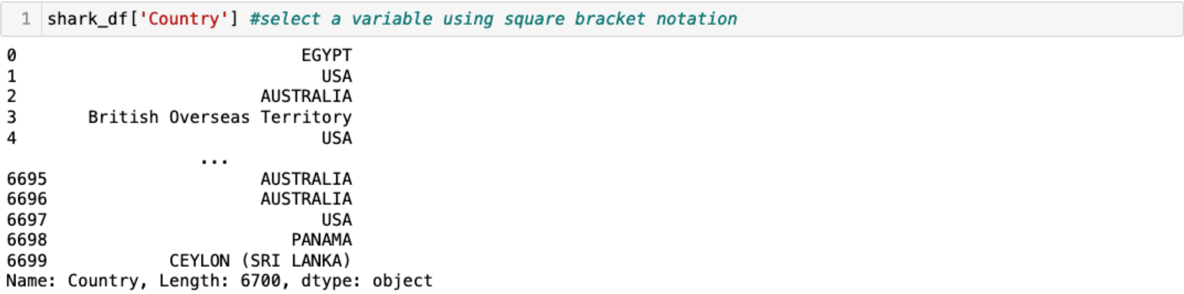
Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

Accessing data

The data frame is made up of observations (rows) and variables (columns). We can select one or more variables and/or observations from the data frame using techniques like slicing and subsetting.



```
shark_df['Country'] #select a variable using square bracket notation
```

```
0                          EGYPT
1                            USA
2                      AUSTRALIA
3      British Overseas Territory
4                            USA
                  ...
6695                   AUSTRALIA
6696                   AUSTRALIA
6697                         USA
6698                      PANAMA
6699            CEYLON (SRI LANKA)
Name: Country, Length: 6700, dtype: object
```

Pandas: Data Ingestion

Load the data

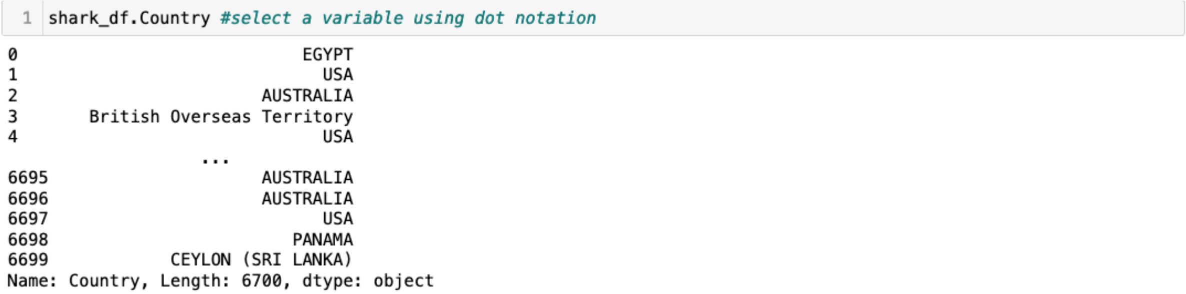
Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

Accessing data

The data frame is made up of observations (rows) and variables (columns). We can select one or more variables and/or observations from the data frame using techniques like slicing and subsetting.



```
shark_df.Country #select a variable using dot notation
```

```
0                            EGYPT
1                              USA
2                        AUSTRALIA
3       British Overseas Territory
4                              USA
                   ...            
6695                     AUSTRALIA
6696                     AUSTRALIA
6697                           USA
6698                        PANAMA
6699            CEYLON (SRI LANKA)
Name: Country, Length: 6700, dtype: object
```

Pandas: Data Ingestion

Load the data

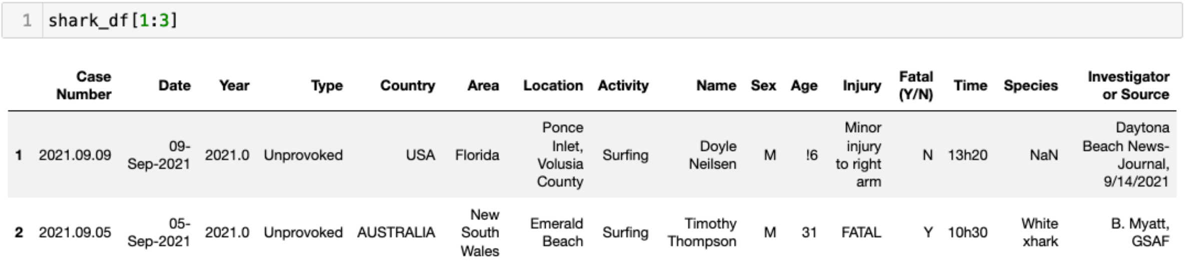
Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

Accessing data

The data frame is made up of observations (rows) and variables (columns). We can select one or more variables and/or observations from the data frame using techniques like slicing and subsetting.



```
shark_df[1:3]
```

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Name	Sex	Age	Injury	Fatal (Y/N)	Time	Species	Investigator or Source
1	2021.09.09	09-Sep-2021	2021.0	Unprovoked	USA	Florida	Ponce Inlet, Volusia County	Surfing	Doyle Neilsen	M	!6	Minor injury to right arm	N	13h20	NaN	Daytona Beach News-Journal, 9/14/2021
2	2021.09.05	05-Sep-2021	2021.0	Unprovoked	AUSTRALIA	New South Wales	Emerald Beach	Surfing	Timothy Thompson	M	31	FATAL	Y	10h30	White xhark	B. Myatt, GSAF

Pandas: Data Ingestion

⋙ Load the data

Load files from our computer by providing the file path to the read_csv() function. If the data is on the web, provide the url.

⋙ Inspect the data

View samples of the data and verify its contents: the number of rows, columns, the data types, etc.

⋙ Accessing data

The data frame is made up of observations (rows) and variables (columns). We can select one or more variables and/or observations from the data frame using techniques like slicing and subsetting.

Data Ingestion: **Web Scraping**

Flat Files on the Web

Text , CSV and Excel files ...pandas also has pd.read_excel()

```
url  = ‘http://’
df   = pd.read csv(url) #load the file from the url
```

Web Scraping

Python provides various libraries that can be used to scrape data from web pages. Web scraping is often used to supplement our data

Application Programming Interface (API)

An API allows us to interact with services on the web. There are both paid and open APIs (free to use); some APIs require authentication and others do not. However, they all have rules that should be followed to ensure that programmers use their resources fairly.

Data Ingestion: **APIs**

Flat Files on the Web

Text , CSV and Excel files …pandas also has pd.read_excel()

```
url  = ‘http://’
df   = pd.read csv(url) #load the file from the url
```

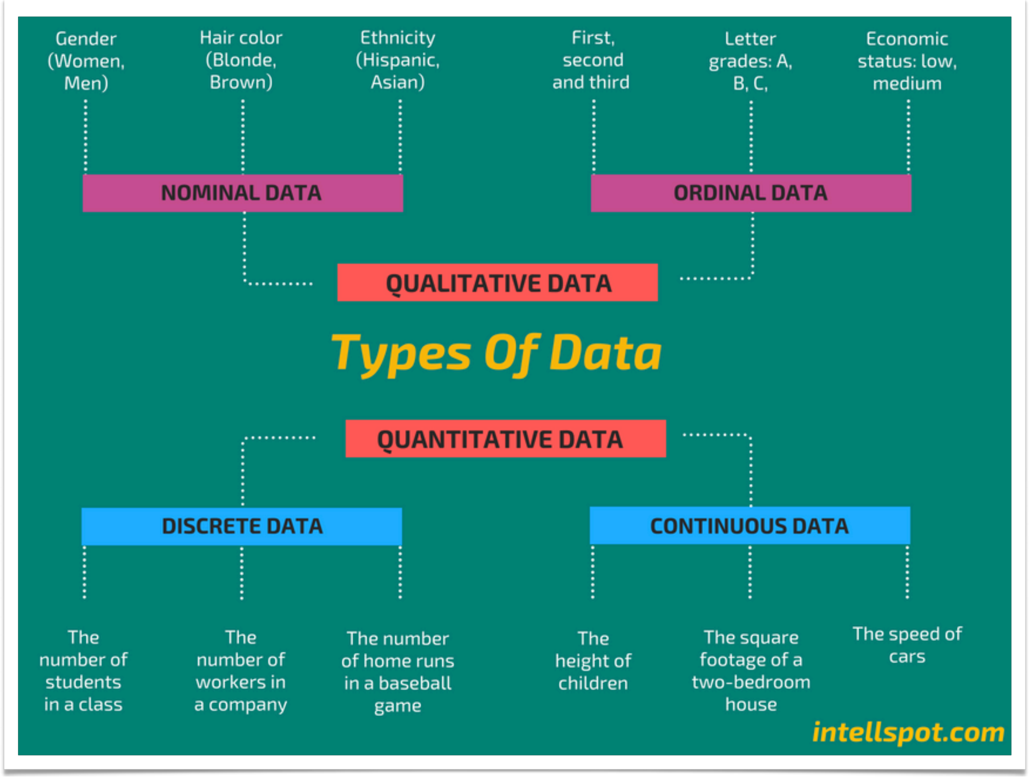
Web Scraping

Python provides various libraries that can be used to scrape data from web pages. Web scraping is often used to supplement our data

Application Programming Interface (API)

An API allows us to interact with services on the web via and endpoint. Endpoints allow users to connect to resources and retrieve data. There are both paid and open APIs (free to use); some APIs require authentication and others do not. However, they all have rules that should be followed to ensure that programmers use their resources fairly.

Types of **Data**


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Source: https://www.intellspot.com/data-types/

Data **Exploration**

Data **Exploration**

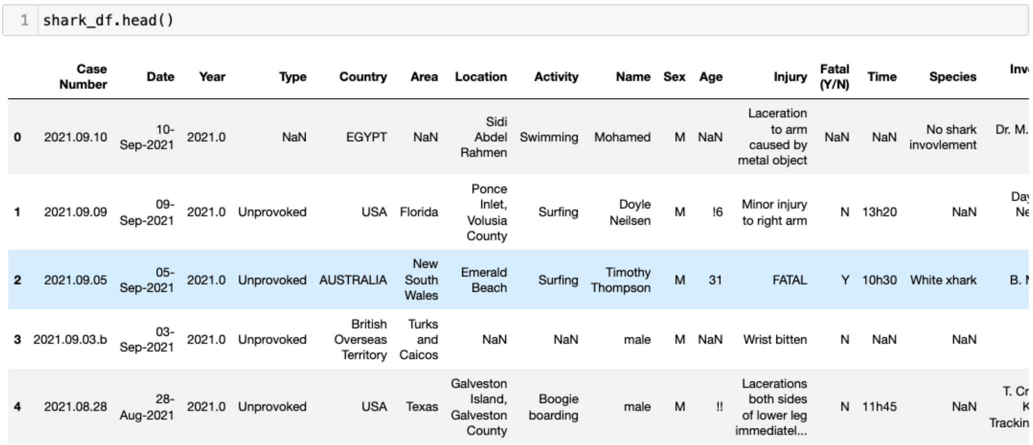
01

Identify any problems with the data. Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.

Data Exploration

01

Identify any problems with the data. Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.



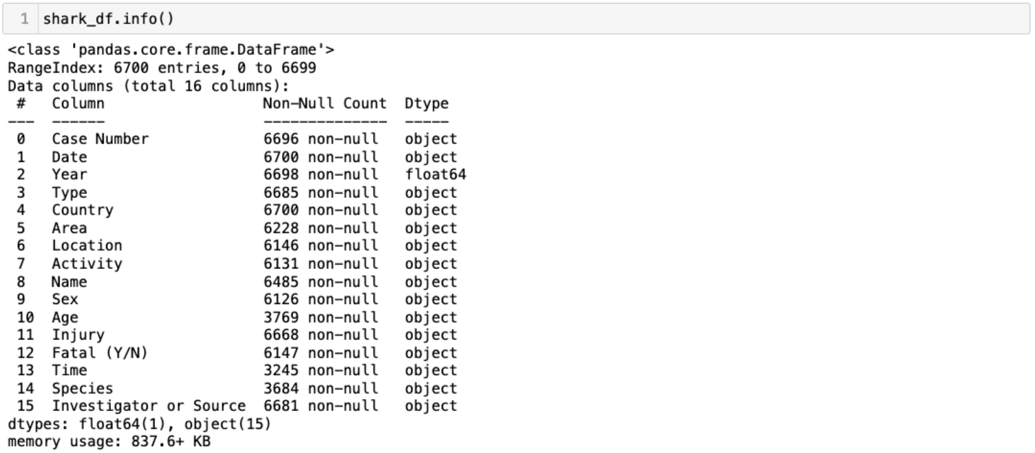
```
shark_df.head()
```

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Name	Sex	Age	Injury	Fatal (Y/N)	Time	Species	Inv
0	2021.09.10	10-Sep-2021	2021.0	NaN	EGYPT	NaN	Sidi Abdel Rahmen	Swimming	Mohamed	M	NaN	Laceration to arm caused by metal object	NaN	NaN	No shark invovlement	Dr. M.
1	2021.09.09	09-Sep-2021	2021.0	Unprovoked	USA	Florida	Ponce Inlet, Volusia County	Surfing	Doyle Neilsen	M	!6	Minor injury to right arm	N	13h20	NaN	Day Ne
2	2021.09.05	05-Sep-2021	2021.0	Unprovoked	AUSTRALIA	New South Wales	Emerald Beach	Surfing	Timothy Thompson	M	31	FATAL	Y	10h30	White xhark	B. N
3	2021.09.03.b	03-Sep-2021	2021.0	Unprovoked	British Overseas Territory	Turks and Caicos	NaN	NaN	male	M	NaN	Wrist bitten	N	NaN	NaN	
4	2021.08.28	28-Aug-2021	2021.0	Unprovoked	USA	Texas	Galveston Island, Galveston County	Boogie boarding	male	M	!!	Lacerations both sides of lower leg immediatel...	N	11h45	NaN	T. Cr K Trackin

Data
Exploration

01

Identify any problems with the data.
Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.



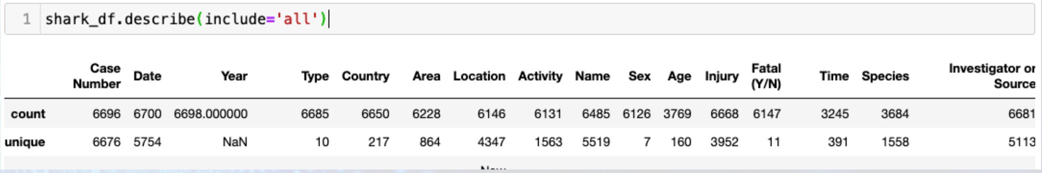
```
shark_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6700 entries, 0 to 6699
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Case Number             6696 non-null   object 
 1   Date                    6700 non-null   object 
 2   Year                    6698 non-null   float64
 3   Type                    6685 non-null   object 
 4   Country                 6700 non-null   object 
 5   Area                    6228 non-null   object 
 6   Location                6146 non-null   object 
 7   Activity                6131 non-null   object 
 8   Name                    6485 non-null   object 
 9   Sex                     6126 non-null   object 
 10  Age                     3769 non-null   object 
 11  Injury                  6668 non-null   object 
 12  Fatal (Y/N)             6147 non-null   object 
 13  Time                    3245 non-null   object 
 14  Species                 3684 non-null   object 
 15  Investigator or Source  6681 non-null   object 
dtypes: float64(1), object(15)
memory usage: 837.6+ KB
```

Data Exploration

01

Identify any problems with the data.
Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.



```
shark_df.describe(include='all')
```

	Case Number	Date	Year	Type	Country	Area	Location	Activity	Name	Sex	Age	Injury	Fatal (Y/N)	Time	Species	Investigator or Source
count	6696	6700	6698.000000	6685	6650	6228	6146	6131	6485	6126	3769	6668	6147	3245	3684	6681
unique	6676	5754	NaN	10	217	864	4347	1563	5519	7	160	3952	11	391	1558	5113

Data **Exploration**

01 **Identify any problems with the data.** Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.

02 **Clean and prepare the data.** Address the issues that were identified, e.g.: resolve the missing data and perform type conversion.

Data **Exploration**

01 **Identify any problems with the data.** Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.

02 **Clean and prepare the data.** Address the issues that were identified, e.g.: resolve the missing data and perform type conversion.

03 **Summarize and visualize the data.** Prepare numerical summaries of the data; view the univariate distribution, the pair-wise correlations, etc.

Data **Exploration**

01 **Identify any problems with the data.** Scrutinize the data and determine if there are: missing values, variables represented with unsuitable data types, duplicates, unusual values.

02 **Clean and prepare the data.** Address the issues that were identified, e.g.: resolve the missing data and perform type conversion.

03 **Summarize and visualize the data.** Prepare numerical summaries of the data; view the univariate distribution, the pair-wise correlations, etc.

04 **Perform outlier detection.** Identify the presence of unusual values within the data and (possibly) remove them using suitable methods such as: IQR or z-score.