Tecnológico
de Monterrey

Maestría en Inteligencia Artificial Aplicada

Materia: Ciencia y analítica de datos

Profesor: Jobish Vallikavungal Devassia

Semana 4: Data Analysis with Python (IBM) : Módulo 1

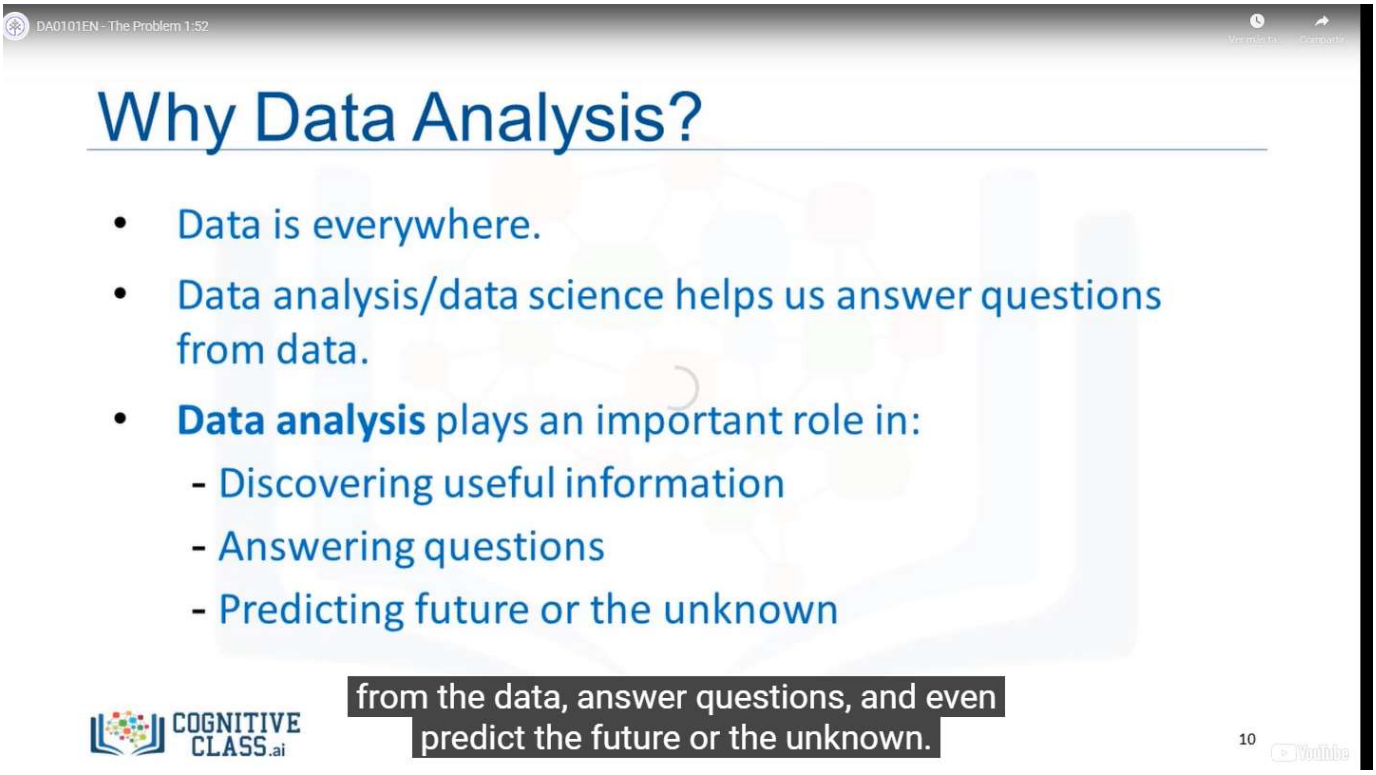
Integrante:

- **Norma de los Ángeles García López - A01228278**

Fecha de entrega: 11 de octubre del 2022

MÓDULO 1

Introduction to Data Analysis with Python



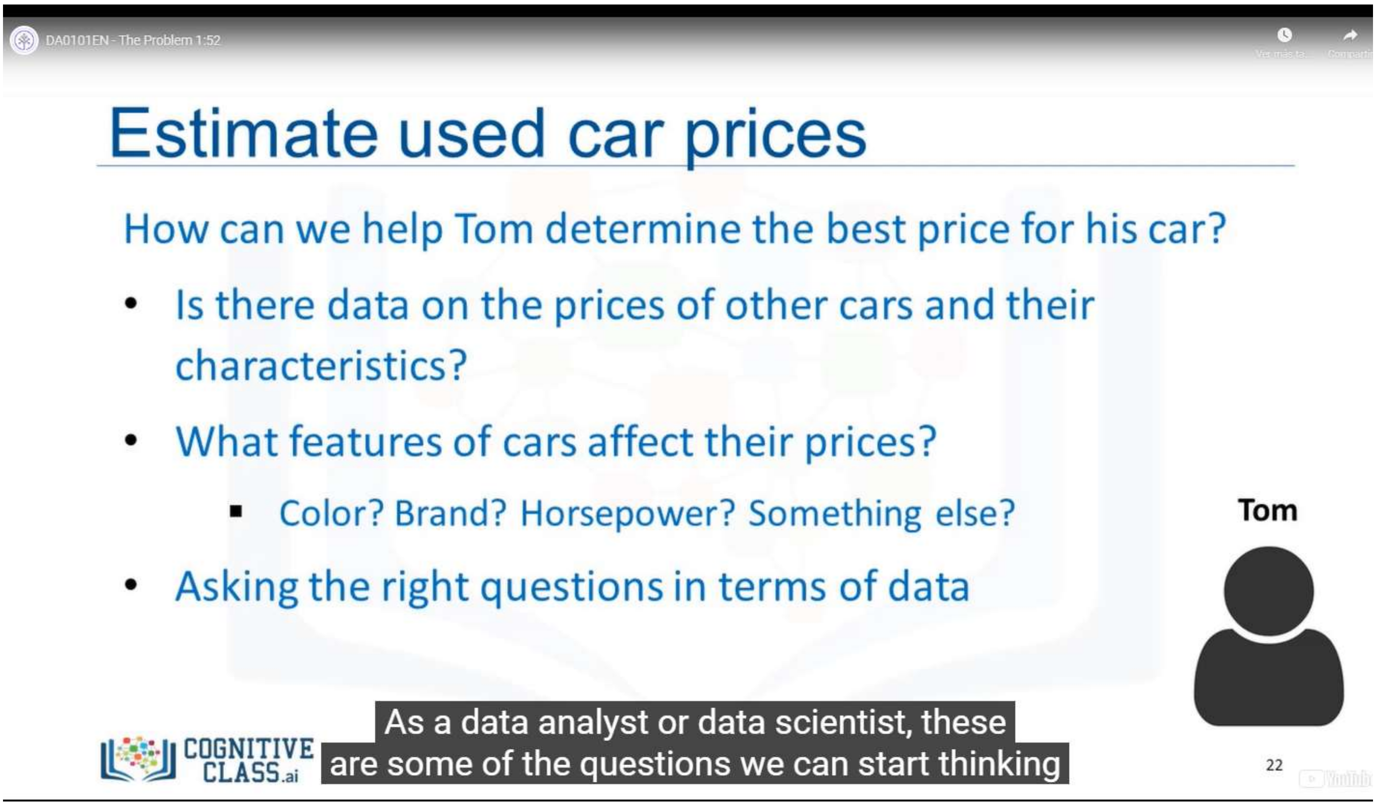
Why Data Analysis?

- Data is everywhere.
- Data analysis/data science helps us answer questions from data.
- **Data analysis** plays an important role in:
  - Discovering useful information
  - Answering questions
  - Predicting future or the unknown

from the data, answer questions, and even predict the future or the unknown.

Problem: Estimate how much money can Tom get from the car?

Some Hypothesis questions



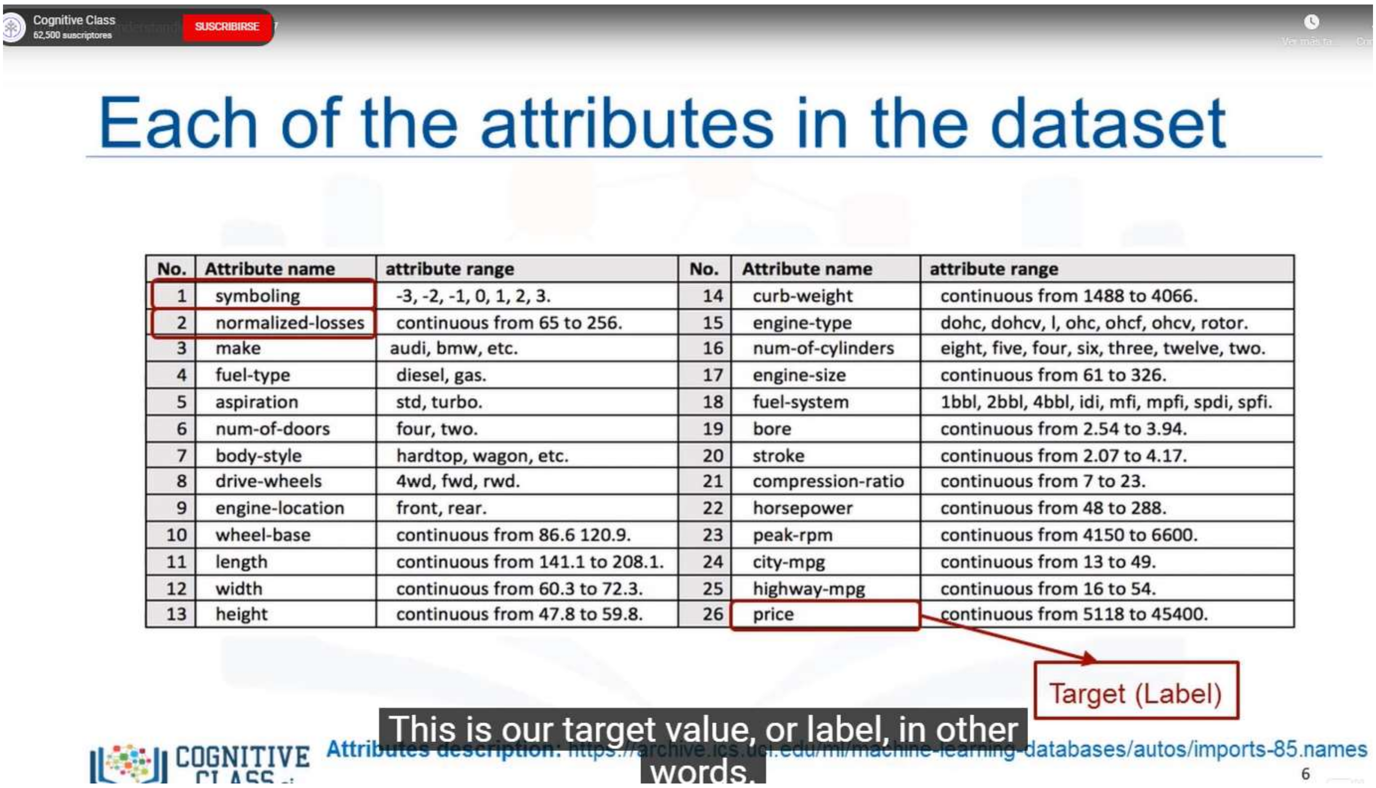
Estimate used car prices

How can we help Tom determine the best price for his car?

- Is there data on the prices of other cars and their characteristics?
- What features of cars affect their prices?
  - Color? Brand? Horsepower? Something else?
- Asking the right questions in terms of data

As a data analyst or data scientist, these are some of the questions we can start thinking

Atributos del dataset



Each of the attributes in the dataset

No.	Attribute name	attribute range	No.	Attribute name	attribute range
1	symboling	-3, -2, -1, 0, 1, 2, 3.	14	curb-weight	continuous from 1488 to 4066.
2	normalized-losses	continuous from 65 to 256.	15	engine-type	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
3	make	audi, bmw, etc.	16	num-of-cylinders	eight, five, four, six, three, twelve, two.
4	fuel-type	diesel, gas.	17	engine-size	continuous from 61 to 326.
5	aspiration	std, turbo.	18	fuel-system	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
6	num-of-doors	four, two.	19	bore	continuous from 2.54 to 3.94.
7	body-style	hardtop, wagon, etc.	20	stroke	continuous from 2.07 to 4.17.
8	drive-wheels	4wd, fwd, rwd.	21	compression-ratio	continuous from 7 to 23.
9	engine-location	front, rear.	22	horsepower	continuous from 48 to 288.
10	wheel-base	continuous from 86.6 120.9.	23	peak-rpm	continuous from 4150 to 6600.
11	length	continuous from 141.1 to 208.1.	24	city-mpg	continuous from 13 to 49.
12	width	continuous from 60.3 to 72.3.	25	highway-mpg	continuous from 16 to 54.
13	height	continuous from 47.8 to 59.8.	26	price	continuous from 5118 to 45400.

Target (Label)

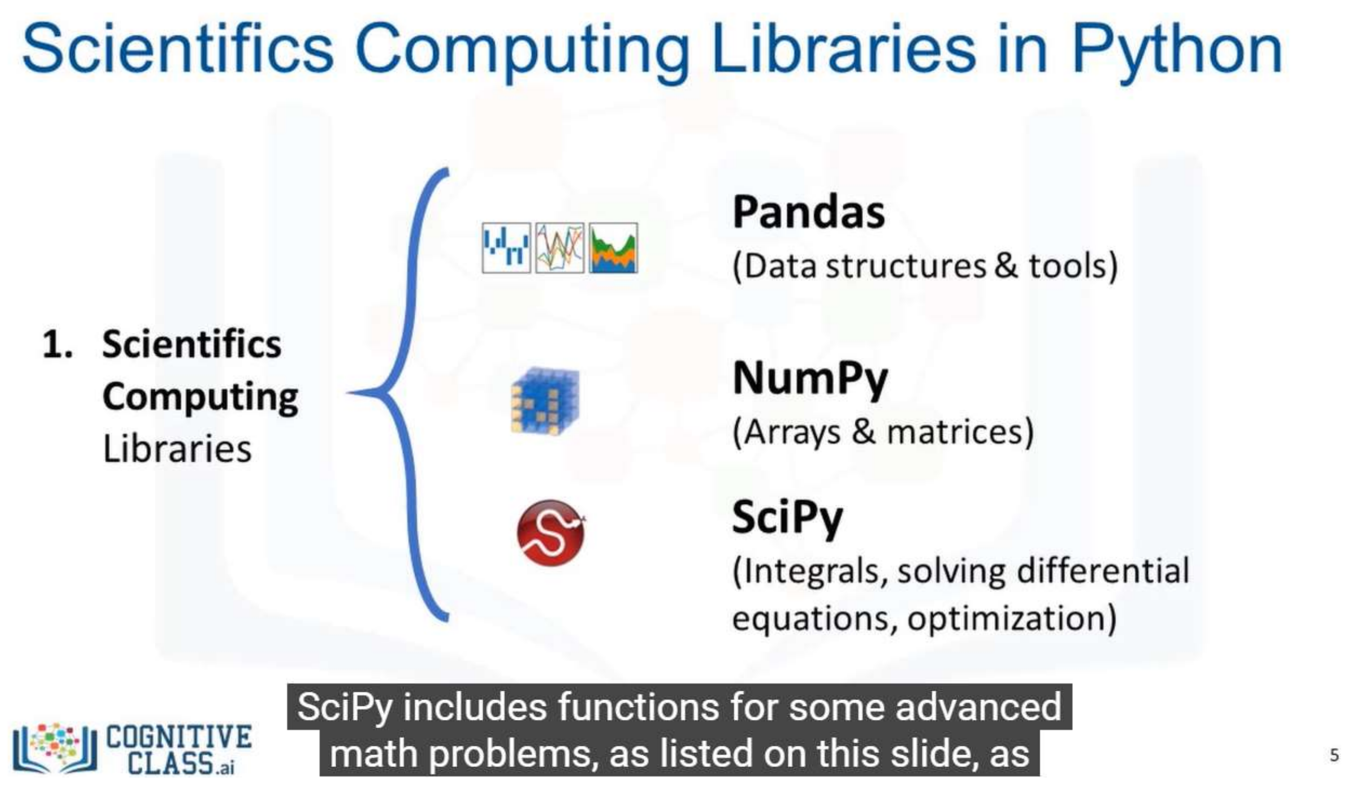
This is our target value, or label, in other words.

Attributes description: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names

The target value is "price" which is the one that we are going to try to estimate with out model, the other variables are the predictors

Python Packages for Data Science

A Python library is a collection of functions and methods that allow you to perform lots of actions without writing any code. The libraries usually contain built-in modules providing different functionalities, which you can use directly.



Scientifics Computing Libraries in Python

1. **Scientifics Computing** Libraries

- **Pandas** (Data structures & tools)
- **NumPy** (Arrays & matrices)
- **SciPy** (Integrals, solving differential equations, optimization)

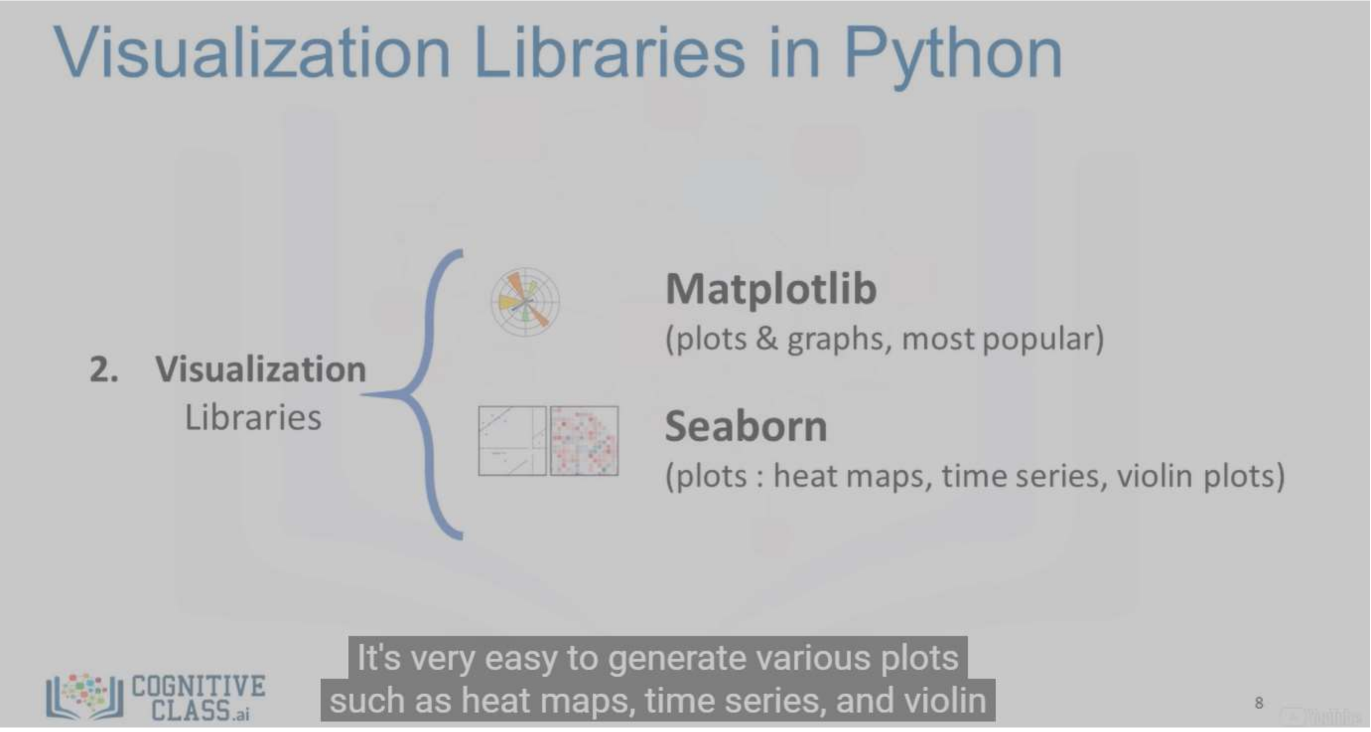
SciPy includes functions for some advanced math problems, as listed on this slide, as

COGNITIVE CLASS.ai

5

Data visualization Methods

Using data visualization methods is the best way to communicate with others, showing them meaningful results of analysis. These libraries enable you to create graphs, charts and maps.



Visualization Libraries in Python

2. **Visualization** Libraries

- **Matplotlib** (plots & graphs, most popular)
- **Seaborn** (plots : heat maps, time series, violin plots)

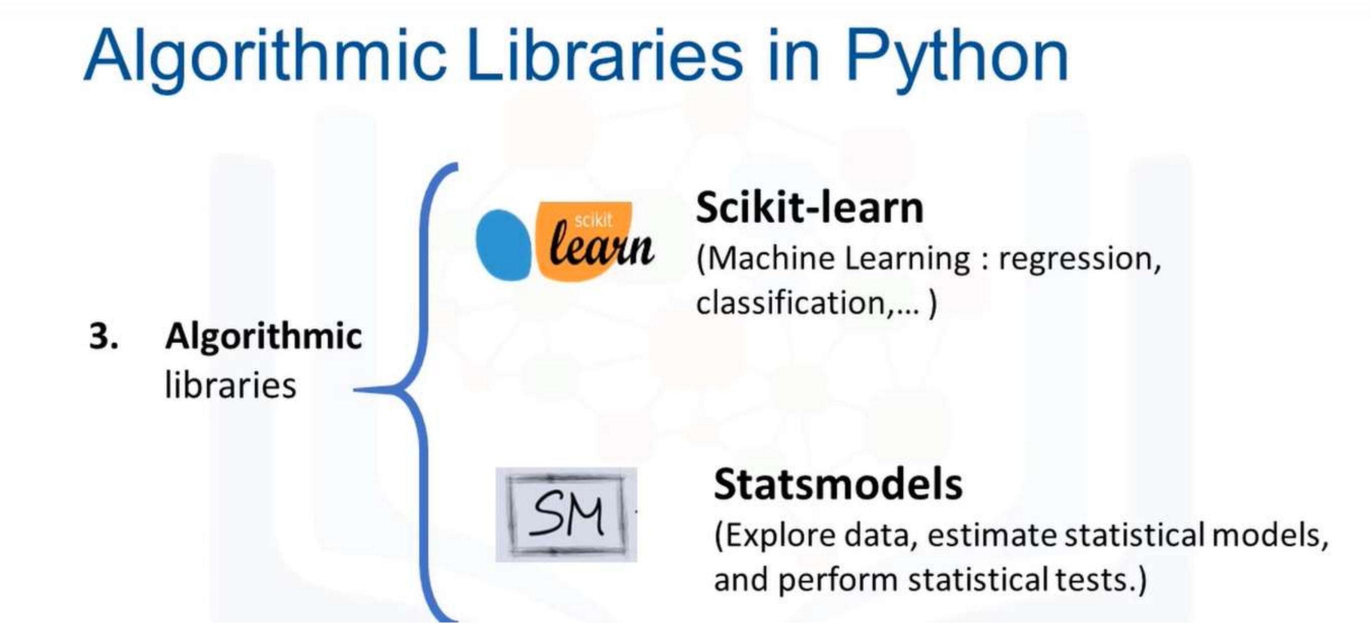
It's very easy to generate various plots such as heat maps, time series, and violin

COGNITIVE CLASS.ai

8

Algorithmic Libraries in Python

With Machine Learning algorithms, we're able to develop a model using our dataset, and obtain predictions.



Algorithmic Libraries in Python

3. **Algorithmic** libraries

Scikit-learn
(Machine Learning : regression, classification,...)

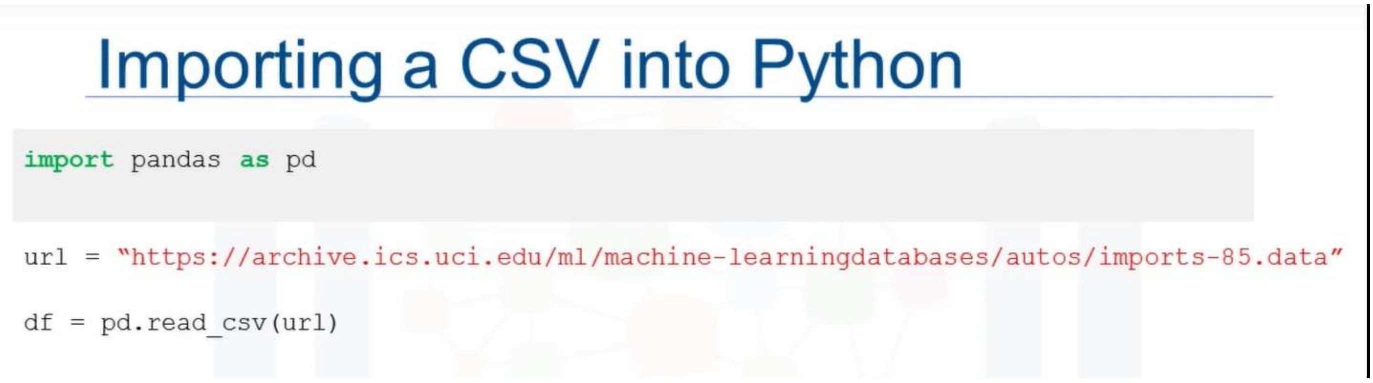
Statsmodels
(Explore data, estimate statistical models, and perform statistical tests.)

Importing and Exporting Data in Python

Data acquisition is a process of loading and reading data into notebook from various sources. To read any data using Python’s pandas package, there are two important factors to consider:

Format is the way data is encoded. We can usually tell different encoding schemes by looking at the ending of the file name. Some common encodings are csv, json, xlsx, hdf and so forth.

The (file) path tells us where the data is stored. Usually it is stored either on the computer we are using, or online on the internet.

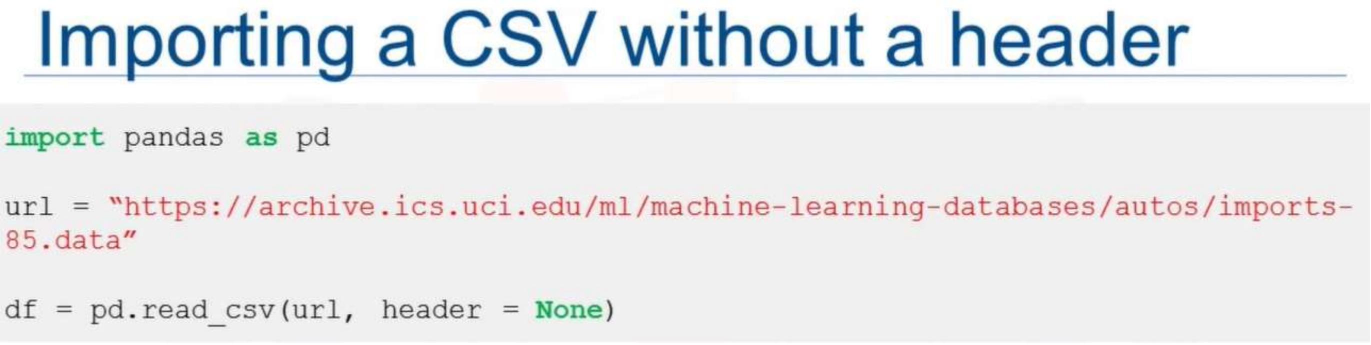


Importing a CSV into Python

```
import pandas as pd

url = “https://archive.ics.uci.edu/ml/machine-learningdatabases/autos/imports-85.data”

df = pd.read_csv(url)
```

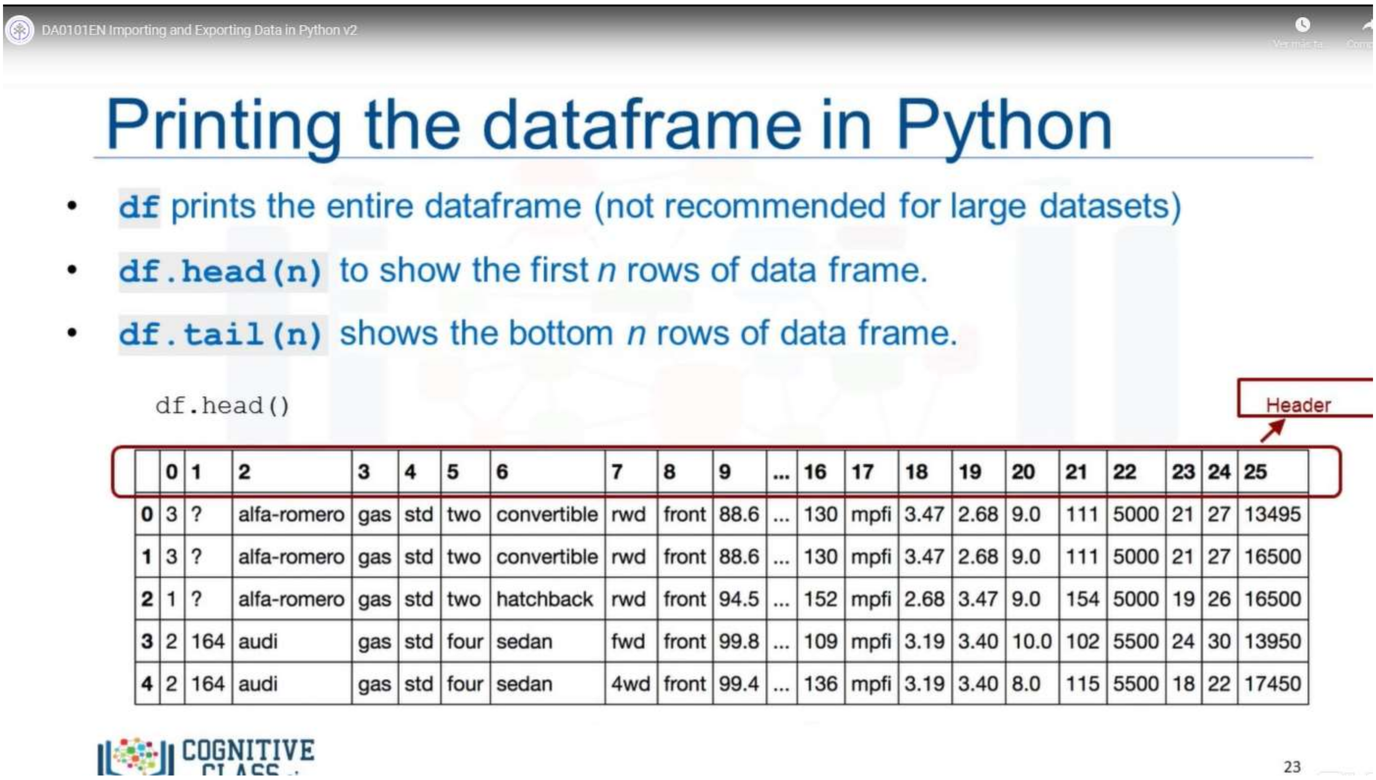


Importing a CSV without a header

```
import pandas as pd

url = "https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data"

df = pd.read_csv(url, header = None)
```

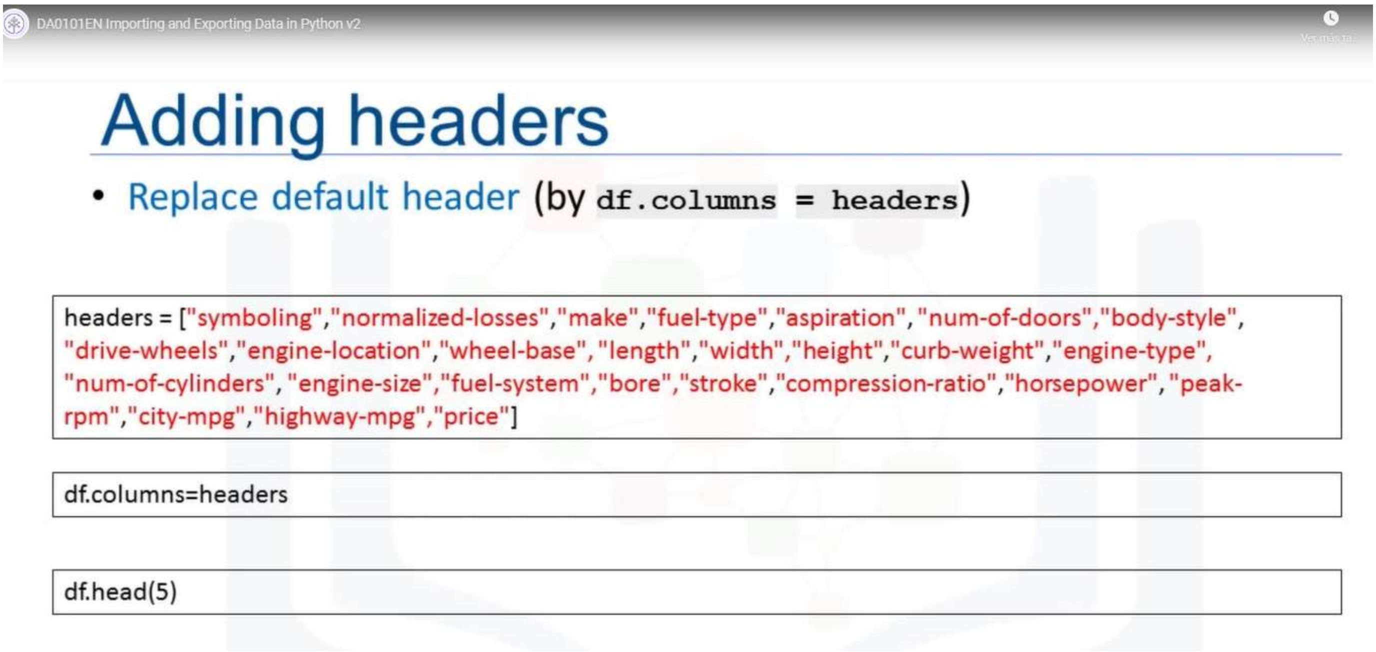


Printing the dataframe in Python

- `df` prints the entire dataframe (not recommended for large datasets)
- `df.head(n)` to show the first *n* rows of data frame.
- `df.tail(n)` shows the bottom *n* rows of data frame.

```
df.head()
```

	0	1	2	3	4	5	6	7	8	9	...	16	17	18	19	20	21	22	23	24	25
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	13495
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	111	5000	21	27	16500
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	154	5000	19	26	16500
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0	102	5500	24	30	13950
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0	115	5500	18	22	17450



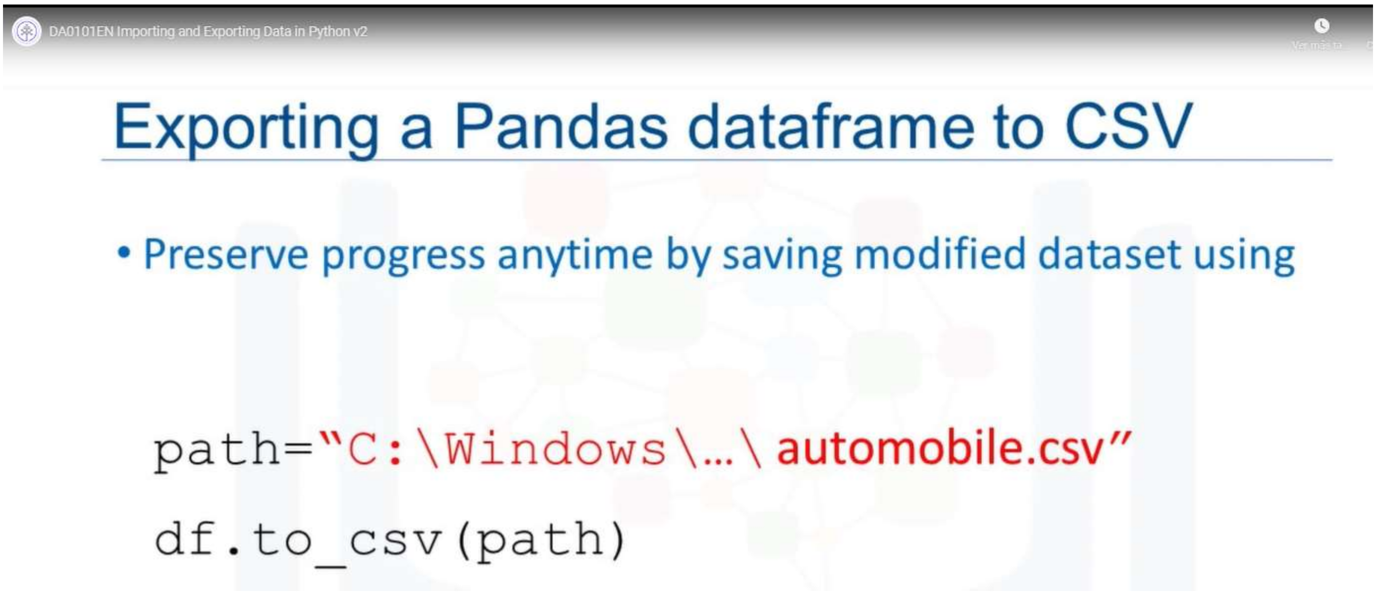
Adding headers

- Replace default header (by `df.columns = headers`)

```
headers = ["symboling","normalized-losses","make","fuel-type","aspiration", "num-of-doors","body-style",
"drive-wheels","engine-location","wheel-base", "length","width","height","curb-weight","engine-type",
"num-of-cylinders", "engine-size","fuel-system","bore","stroke","compression-ratio","horsepower", "peak-
rpm","city-mpg","highway-mpg","price"]
```

```
df.columns=headers
```

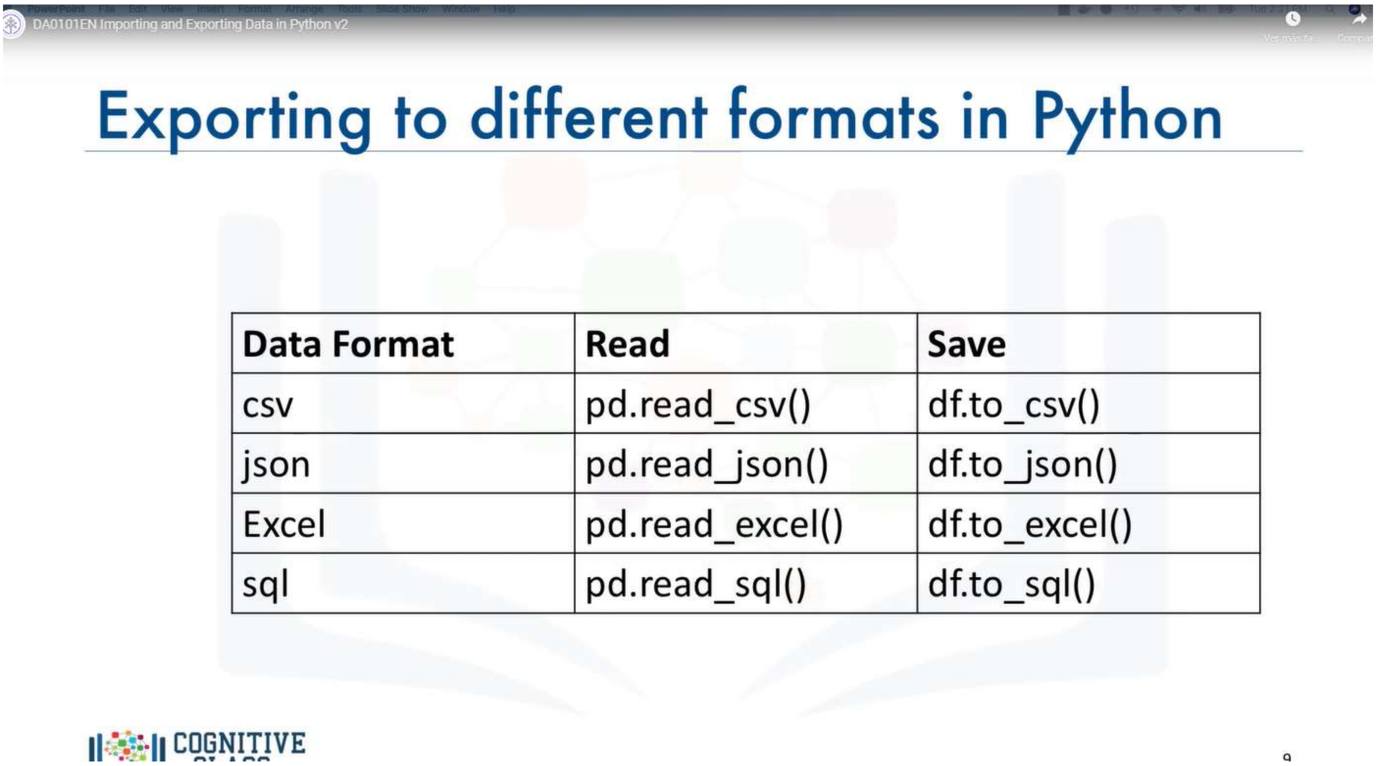
```
df.head(5)
```



Exporting a Pandas dataframe to CSV

- Preserve progress anytime by saving modified dataset using

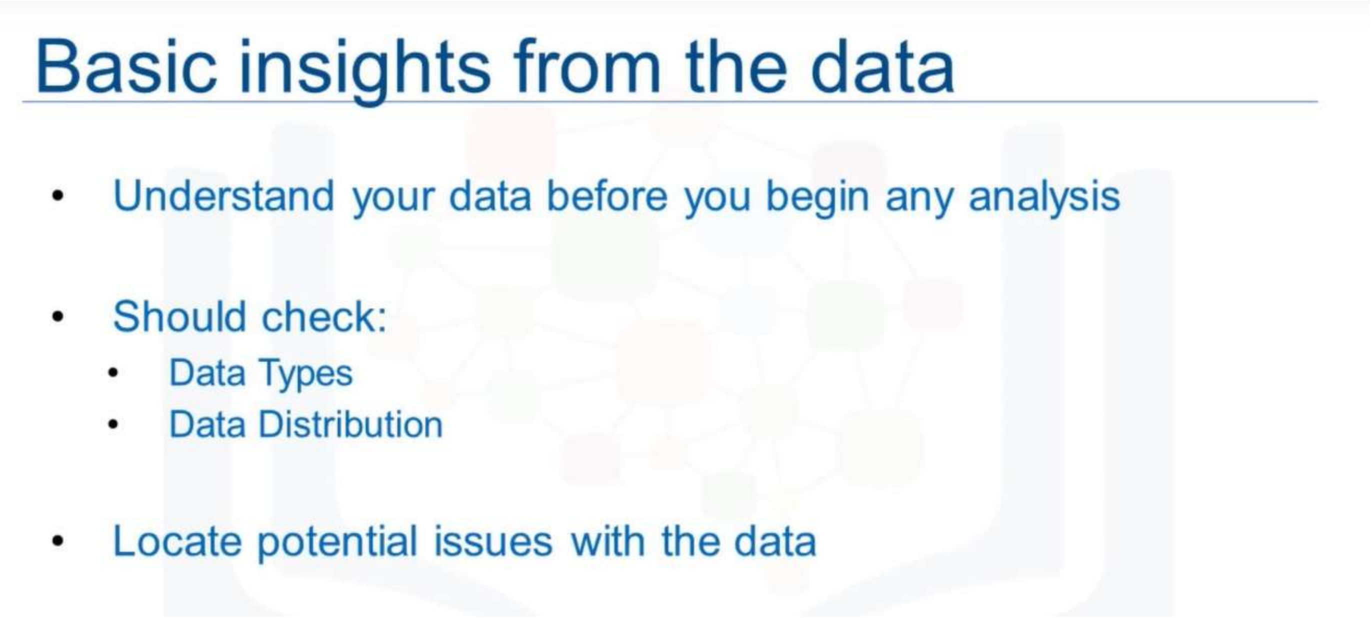
```
path="C:\Windows\...\ automobile.csv"
df.to_csv(path)
```



Exporting to different formats in Python

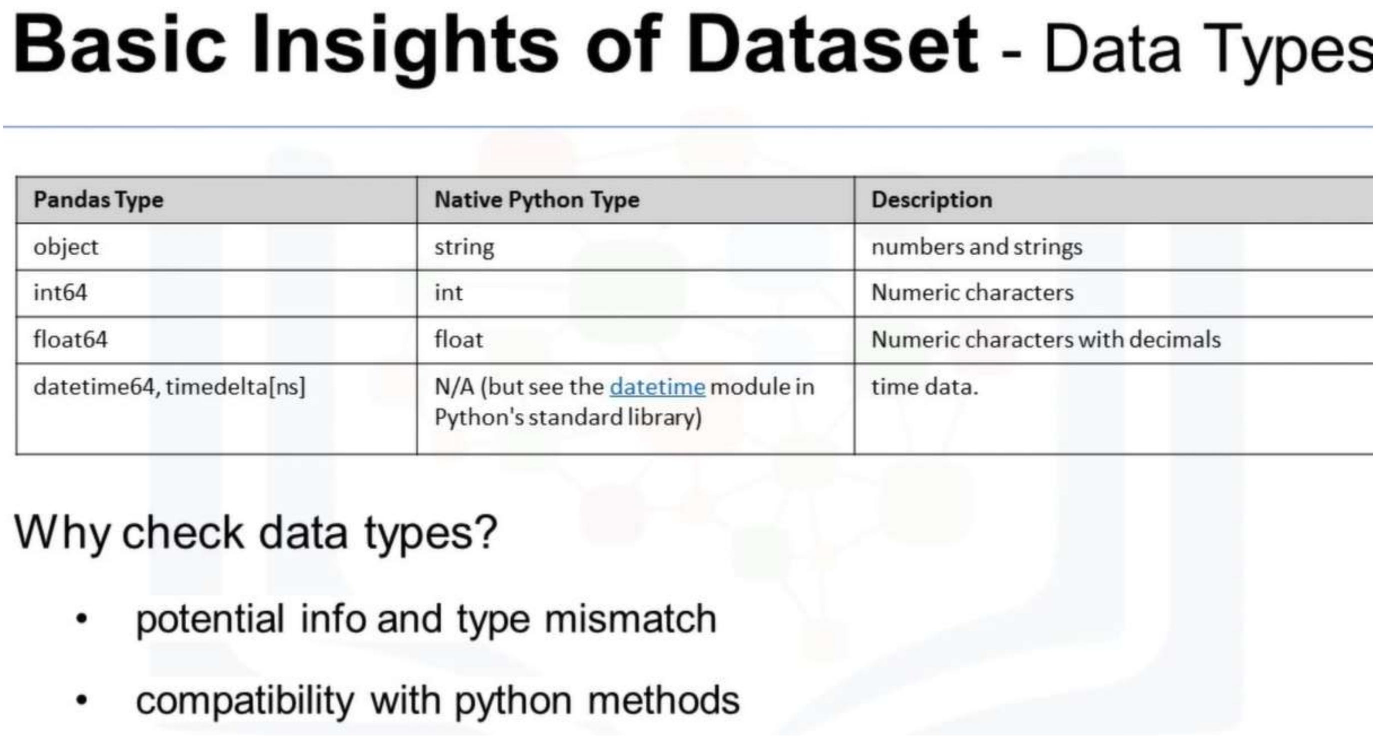
Data Format	Read	Save
csv	pd.read_csv()	df.to_csv()
json	pd.read_json()	df.to_json()
Excel	pd.read_excel()	df.to_excel()
sql	pd.read_sql()	df.to_sql()

Getting Started Analyzing Data in Python



Basic insights from the data

- Understand your data before you begin any analysis
- Should check:
  - Data Types
  - Data Distribution
- Locate potential issues with the data



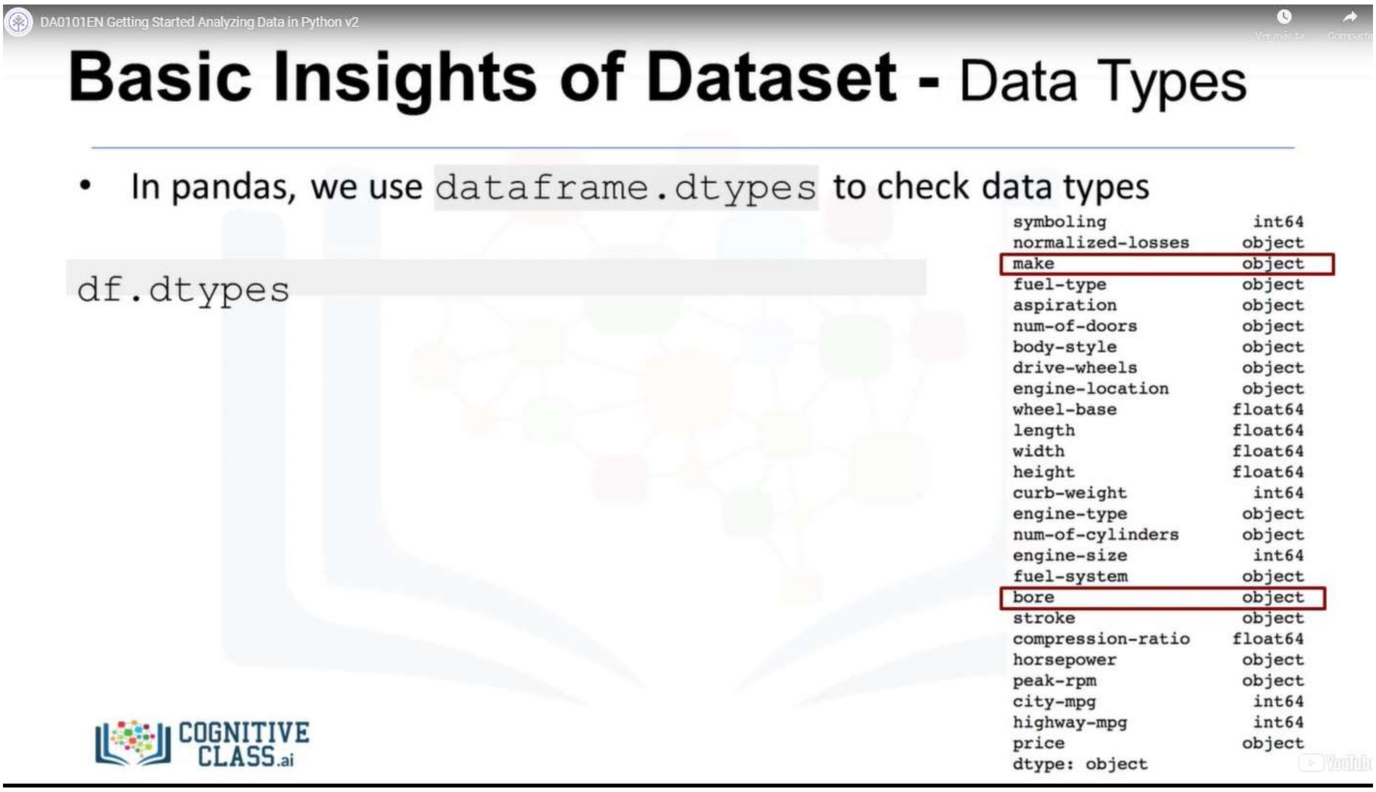
Basic Insights of Dataset - Data Types

Pandas Type	Native Python Type	Description
object	string	numbers and strings
int64	int	Numeric characters
float64	float	Numeric characters with decimals
datetime64, timedelta[ns]	N/A (but see the datetime module in Python's standard library)	time data.

Why check data types?

- potential info and type mismatch
- compatibility with python methods

When the "dtype" method is applied to the data set, the datatype of each column is returned in a Series. A good data scientist's intuition tells us that most of the data types make sense. The make of cars, for example, are names, so this information should be of type object.



DA0101EN Getting Started Analyzing Data in Python v2

Basic Insights of Dataset - Data Types

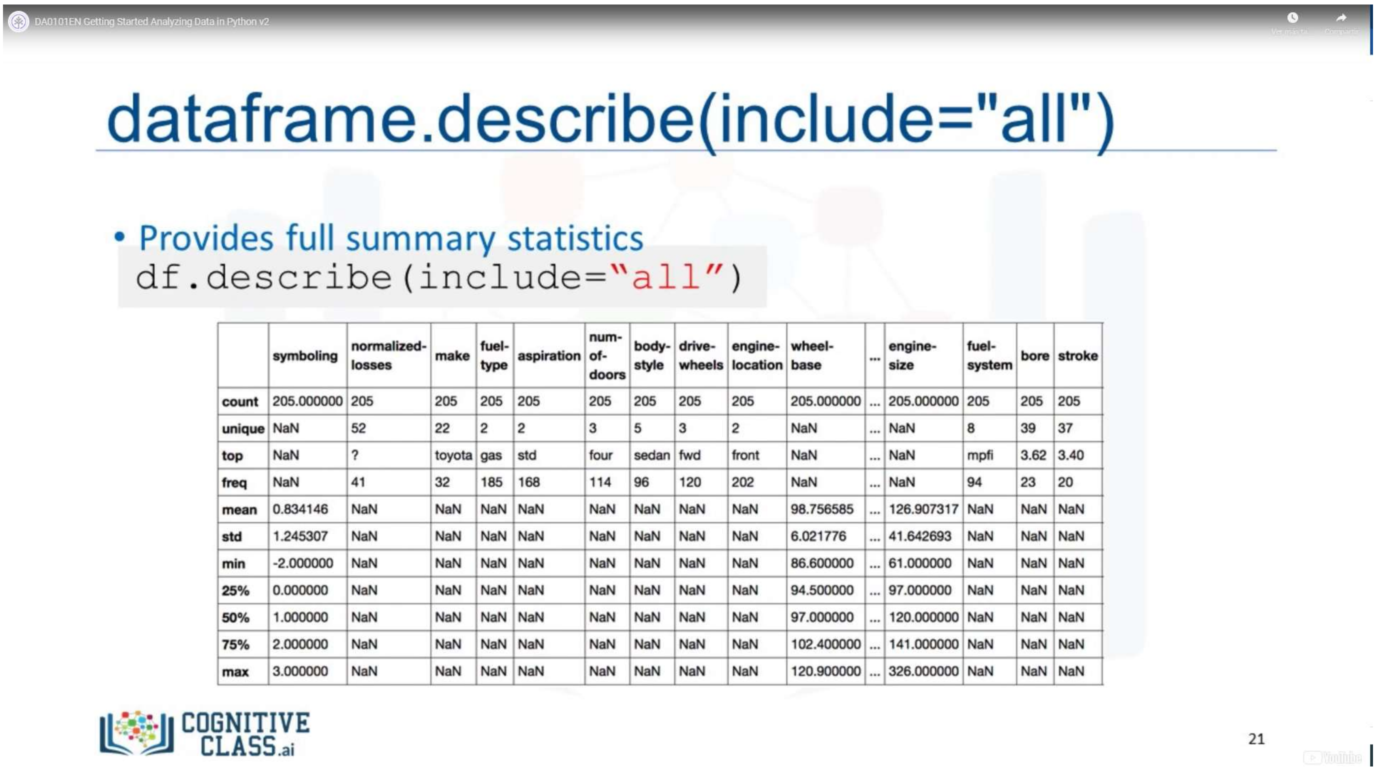
- In pandas, we use `dataframe.dtypes` to check data types

```
df.dtypes
```

```
symboling              int64
normalized-losses     object
make                  object
fuel-type             object
aspiration            object
num-of-doors          object
body-style            object
drive-wheels          object
engine-location       object
wheel-base           float64
length               float64
width                float64
height               float64
curb-weight            int64
engine-type           object
num-of-cylinders      object
engine-size            int64
fuel-system           object
bore                  object
stroke                object
compression-ratio    float64
horsepower            object
peak-rpm              object
city-mpg               int64
highway-mpg            int64
price                 object
dtype: object
```

COGNITIVE CLASS.ai

Now we would like to check the statistical summary of each column to learn about the distribution of data in each column. The statistical metrics can tell the data scientist if there are mathematical issues that may exist, such as extreme outliers and large deviations. The data scientist may have to address these issues later. To get the quick statistics, we use the describe method. It returns the number of terms in the column as "count", average column value as "mean", column standard deviation as "std", the maximum and minimum values, as well as the boundary of each of the quartiles.



DA0101EN Getting Started Analyzing Data in Python v2

dataframe.describe(include="all")

- Provides full summary statistics

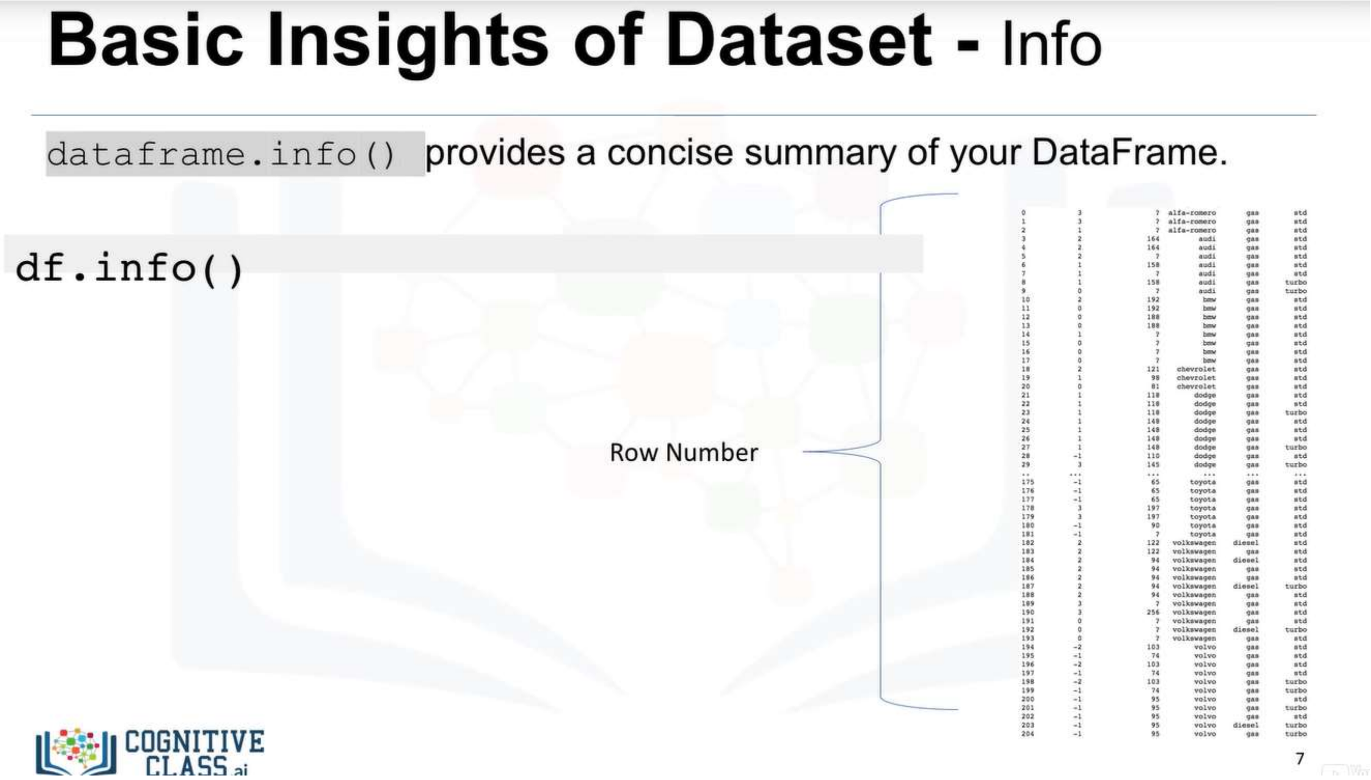
```
df.describe(include="all")
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke
count	205.000000	205	205	205	205	205	205	205	205	205.000000	...	205.000000	205	205	205
unique	NaN	52	22	2	2	3	5	3	2	NaN	...	NaN	8	39	37
top	NaN	?	toyota	gas	std	four	sedan	fwd	front	NaN	...	NaN	mpfi	3.62	3.40
freq	NaN	41	32	185	168	114	96	120	202	NaN	...	NaN	94	23	20
mean	0.834146	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	98.756585	...	126.907317	NaN	NaN	NaN
std	1.245307	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6.021776	...	41.642693	NaN	NaN	NaN
min	-2.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	86.600000	...	61.000000	NaN	NaN	NaN
25%	0.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	94.500000	...	97.000000	NaN	NaN	NaN
50%	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	97.000000	...	120.000000	NaN	NaN	NaN
75%	2.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	102.400000	...	141.000000	NaN	NaN	NaN
max	3.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	120.900000	...	326.000000	NaN	NaN	NaN

COGNITIVE CLASS.ai

21

"Unique" is the number of distinct objects in the column, "top" is the most frequently occurring object, and "freq" is the number of times the top object appears in the column. Some values in the table are shown here as "NaN", which stands for "not a number". This is because that particular statistical metric cannot be calculated for that specific column data type.



Basic Insights of Dataset - Info

`dataframe.info()` provides a concise summary of your DataFrame.

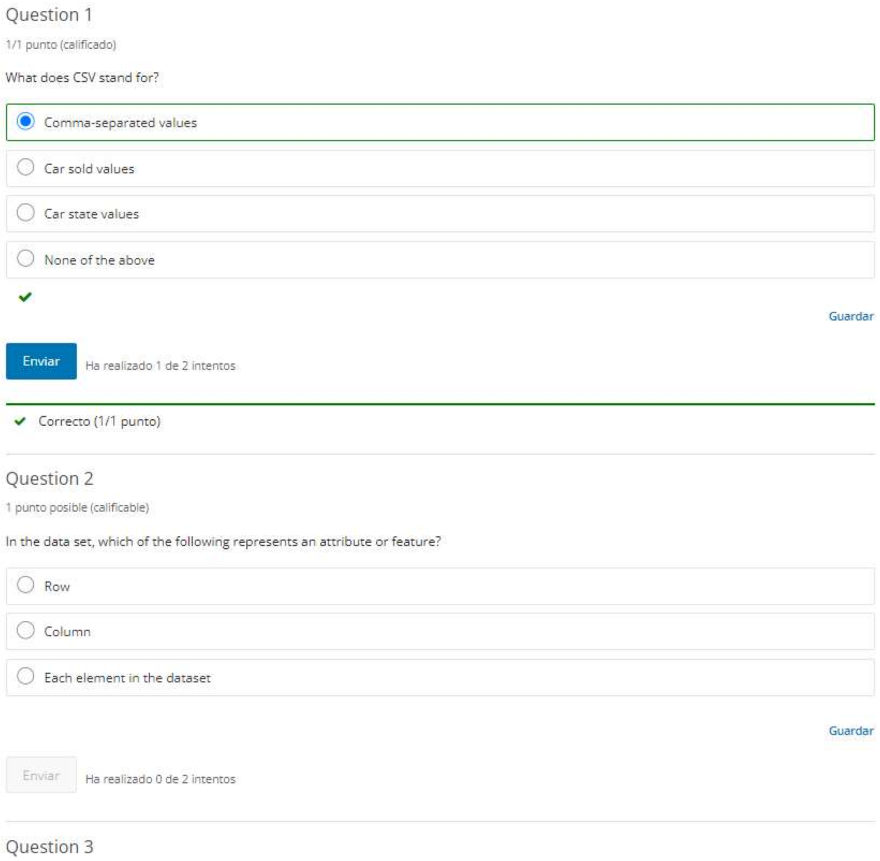
```
df.info()
```

```
0       3       ?  alfa-romero     gas    std
1       3       ?  alfa-romero     gas    std
2       1       ?  alfa-romero     gas    std
3       2     164         audi     gas    std
4       2     164         audi     gas    std
5       2       ?         audi     gas    std
6       1     158         audi     gas    std
7       1       ?         audi     gas    std
8       1     158         audi     gas  turbo
9       0       ?         audi     gas  turbo
10      2     192          bmw     gas    std
11      0     192          bmw     gas    std
12      0     188          bmw     gas    std
13      0     188          bmw     gas    std
14      1       ?          bmw     gas    std
15      0       ?          bmw     gas    std
16      0       ?          bmw     gas    std
17      0       ?          bmw     gas    std
18      2     121    chevrolet     gas    std
19      1      98    chevrolet     gas    std
20      0      81    chevrolet     gas    std
21      1     118        dodge     gas    std
22      1     118        dodge     gas    std
23      1     118        dodge     gas  turbo
24      1     148        dodge     gas    std
25      1     148        dodge     gas    std
26      1     148        dodge     gas    std
27      1     148        dodge     gas  turbo
28     -1     110        dodge     gas    std
29      3     145        dodge     gas  turbo
..    ...     ...          ...     ...    ...
175    -1      65       toyota     gas    std
176    -1      65       toyota     gas    std
177    -1      65       toyota     gas    std
178     3     197       toyota     gas    std
179     3     197       toyota     gas    std
180    -1      90       toyota     gas    std
181    -1       ?       toyota     gas    std
182     2     122   volkswagen  diesel    std
183     2     122   volkswagen     gas    std
184     2      94   volkswagen  diesel    std
185     2      94   volkswagen     gas    std
186     2      94   volkswagen     gas    std
187     2      94   volkswagen  diesel  turbo
188     2      94   volkswagen     gas    std
189     3       ?   volkswagen     gas    std
190     3     256   volkswagen     gas    std
191     0       ?   volkswagen     gas    std
192     0       ?   volkswagen  diesel  turbo
193     0       ?   volkswagen     gas    std
194    -2     103        volvo     gas    std
195    -1      74        volvo     gas    std
196    -2     103        volvo     gas    std
197    -1      74        volvo     gas    std
198    -2     103        volvo     gas  turbo
199    -1      74        volvo     gas  turbo
200    -1      95        volvo     gas    std
201    -1      95        volvo     gas  turbo
202    -1      95        volvo     gas    std
203    -1      95        volvo  diesel  turbo
204    -1      95        volvo     gas  turbo
```

7

Another method you can use to check your dataset is the dataframe.info function. This function shows the top 30 rows and bottom 30 rows of the dataframe.

Graded Review Questions



Question 1

1/1 punto (calificado)

What does CSV stand for?

● Comma-separated values

○ Car sold values

○ Car state values

○ None of the above

✓

Ha realizado 1 de 2 intentos

✓ Correcto (1/1 punto)

Question 2

1 punto posible (calificable)

In the data set, which of the following represents an attribute or feature?

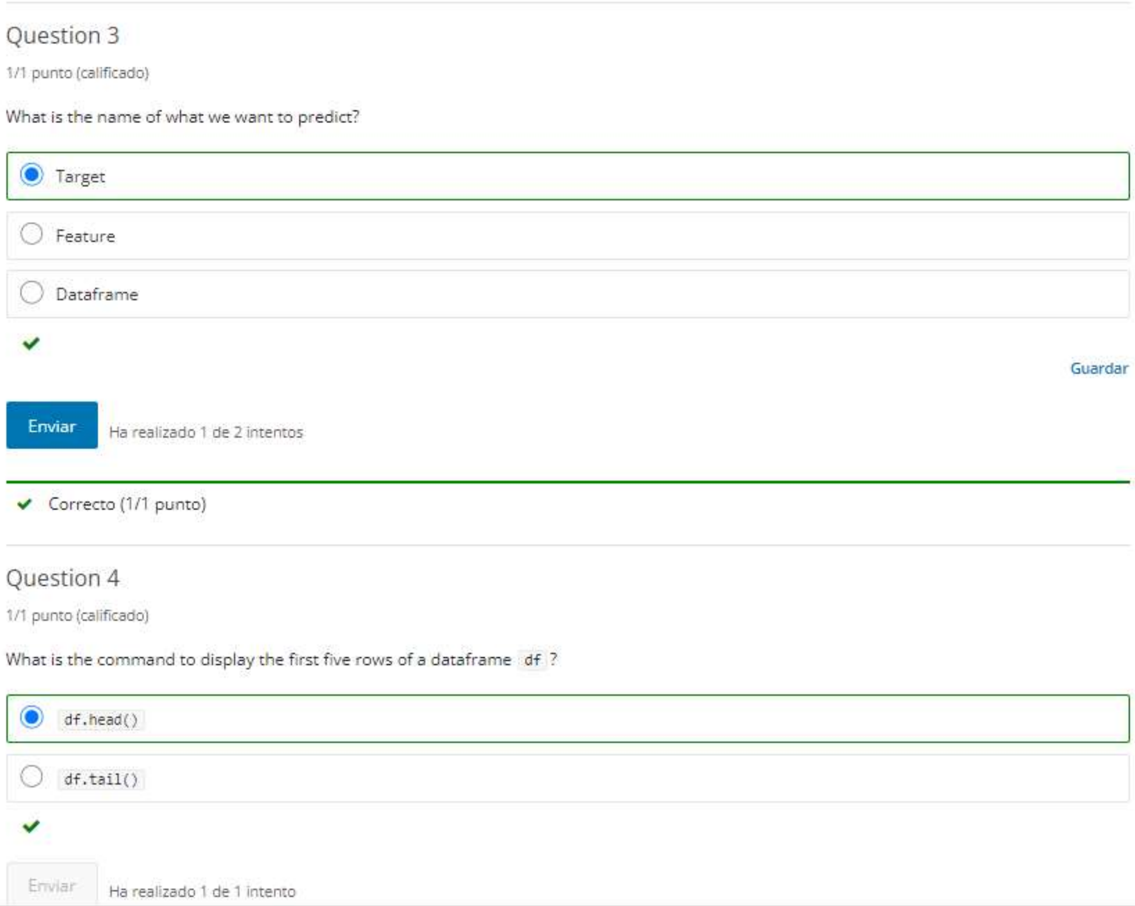
○ Row

○ Column

○ Each element in the dataset

Ha realizado 0 de 2 intentos

Question 3



Question 3

1/1 punto (calificado)

What is the name of what we want to predict?

● Target

○ Feature

○ Dataframe

✓

Ha realizado 1 de 2 intentos

✓ Correcto (1/1 punto)

Question 4

1/1 punto (calificado)

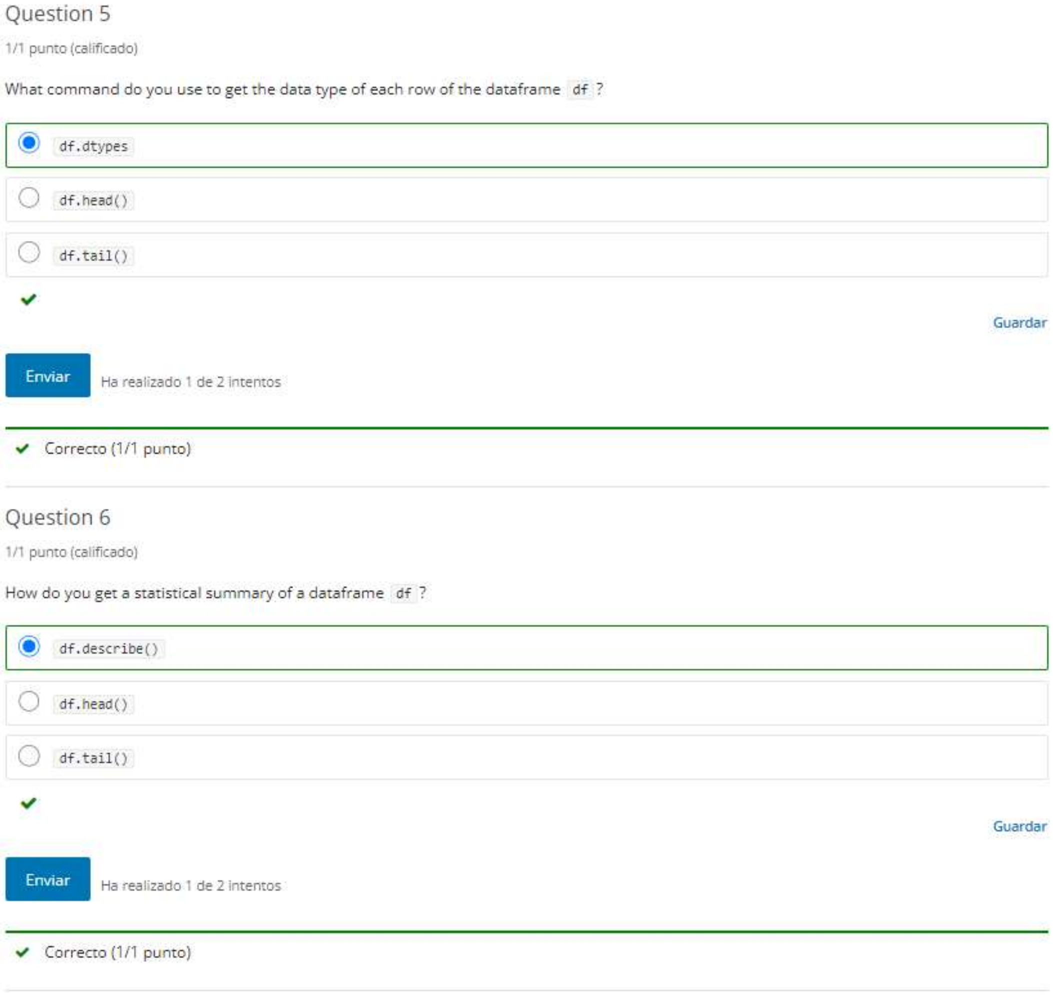
What is the command to display the first five rows of a dataframe `df` ?

● `df.head()`

○ `df.tail()`

✓

Ha realizado 1 de 1 intento



Question 5

1/1 punto (calificado)

What command do you use to get the data type of each row of the dataframe `df` ?

● `df.dtypes`

○ `df.head()`

○ `df.tail()`

✓

Ha realizado 1 de 2 intentos

✓ Correcto (1/1 punto)

Question 6

1/1 punto (calificado)

How do you get a statistical summary of a dataframe `df` ?

● `df.describe()`

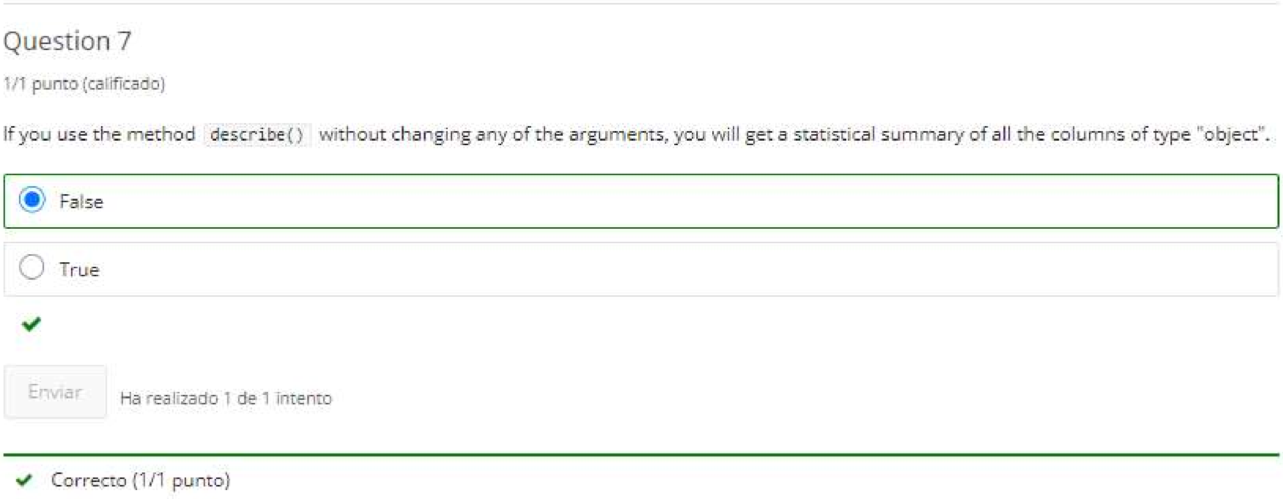
○ `df.head()`

○ `df.tail()`

✓

Ha realizado 1 de 2 intentos

✓ Correcto (1/1 punto)



Question 7

1/1 punto (calificado)

If you use the method `describe()` without changing any of the arguments, you will get a statistical summary of all the columns of type "object".

● False

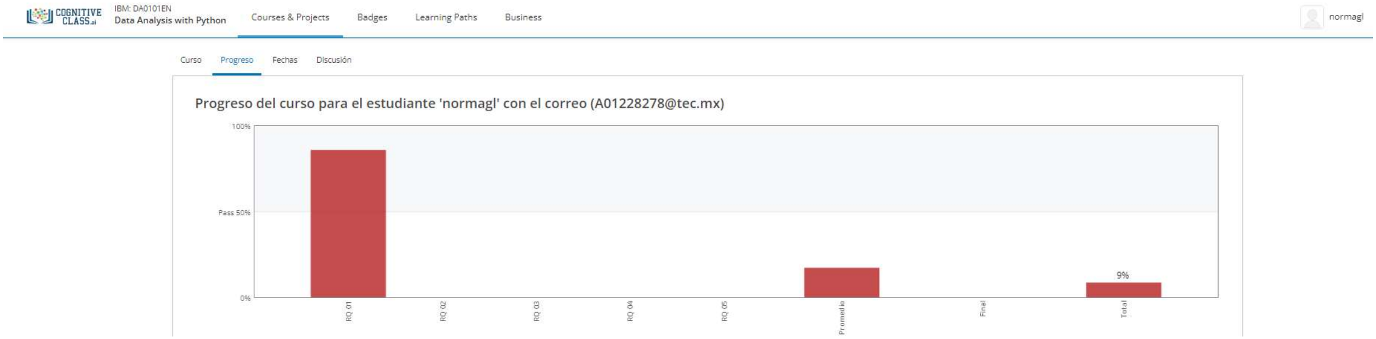
○ True

✓

Ha realizado 1 de 1 intento

✓ Correcto (1/1 punto)

PROGRESO



Progreso del curso para el estudiante 'normagl' con el correo (A01228278@tec.mx)

Productos de pago de Colab - Cancelar contratos