

ITESM

# MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

SEGUNDO TRIMESTRE

APUNTES PERSONALES Data Science  
with Python

MODULO 1

NOMBRE: VILLALPANDO GUERRERO  
JIRAM CESAR

## MODULO 1

### El problema

Data Análisis es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones. El análisis de datos tiene múltiples facetas y enfoques, que abarca diversas técnicas en una variedad de nombres, en diferentes negocios, la ciencia, y los dominios de las ciencias sociales. Los datos se coleccionan y analizan para indagar en cuestiones, probar conjeturas o refutar teorías.

Ejemplo: Cómo poder usar los datos para llegar a la conclusión de qué producto comprar o qué decisión tomar con base a los datos.

## Entendiendo la data

En todas los data sets hay atributos que tienen y que serán distintos o distinguirán los valores. Por ejemplo:

## 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 to 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.

Cada atributo tiene su rango. Hay que tener en cuenta que la mayoría de dataframes vienen en CSV que significa valores separados por comas

## Paquetes de data science para Python

Algunas paqueterías relevantes para Python y de las más usadas son:

## 1. Scientifics Computing Libraries



### Pandas

(Data structures & tools)



### NumPy

(Arrays & matrices)



### SciPy

(Integrals, solving differential equations, optimization)

Para data visualización se recomiendan las siguientes librerías, estas permitirán compartir los datos de una manera más visible y que podrán ser usadas para tomar decisiones en cualquier tipo de proyecto u organización

## 2. Visualization Libraries



### Matplotlib

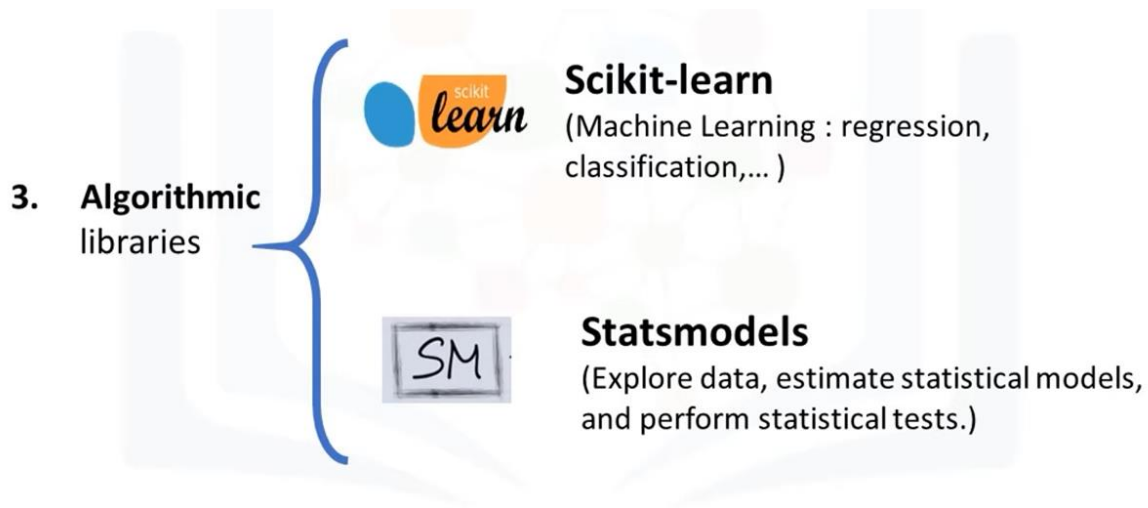
(plots & graphs, most popular)



### Seaborn

(plots : heat maps, time series, violin plots)

Otras librerías para algoritmos y más técnicas son las siguientes, que permitan realizar un estudio más preciso de los que se requiere



### Importando y exportando los datos en Python

El código Python en un módulo obtiene acceso al código en otro módulo por el proceso de importarlo. La instrucción `import` es la forma más común de invocar la maquinaria de importación, pero no es la única manera. Funciones como `importlib`.

Hay factores a considerar como: formato (csv, json, etc) y también el path en donde se encuentra nuestro dataset

En pandas se puede leer mediante el path y después usar: `pd.read_csv(url)`

Para imprimir

Para poder imprimir podemos usar `df.head(n)`, para los primeros datos

Podemos también agregar headers en pandas

Mediante `df.columns=headers`

Exportar

Para exportar un dataframe a csv, ejemplo: `df.to_csv(ruta)`

Para otros formatos:

▶ <b>Data Format</b>	<b>Read</b>	<b>Save</b>
▶ csv	<code>pd.read_csv()</code>	<code>df.to_csv()</code>
▶ json	<code>pd.read_json()</code>	<code>df.to_json()</code>
▶ Excel	<code>pd.read_excel()</code>	<code>df.to_excel()</code>
▶ sql	<code>pd.read_sql()</code>	<code>df.to_sql()</code>

## Empezar a analizar en Python

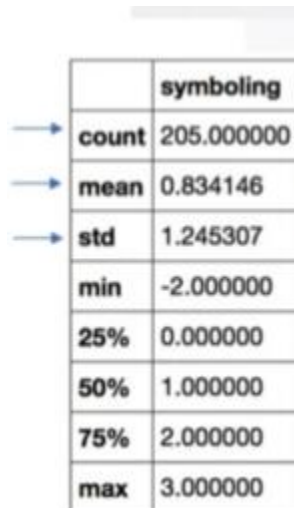
Ejemplos de insights que se deben de tener en cuenta

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 <a href="#">datetime</a> module in Python's standard library)	time data.

Un DataFrame es una estructura de datos con dos dimensiones en la cual se puede guardar datos de distintos tipos (como caracteres, enteros, valores de punto flotante, factores y más) en columnas. Es similar a una hoja de cálculo o una tabla de SQL o el data.

Para describir estadísticamente podemos usar: `df.describe()`

Y nos regresará lo siguiente:



	symboling
count	205.000000
mean	0.834146
std	1.245307
min	-2.000000
25%	0.000000
50%	1.000000
75%	2.000000
max	3.000000

Dataframe.describe(include="all")

Nos indicará más atributos como:

	count	20
→	unique	Ni
→	top	Ni
→	freq	Ni
	mean	0.0
	std	1.0
	min	-2
	25%	0.0
	50%	1.0
	75%	2.0
	max	3.0

Dataframe.info()

Mostrará las primeras y ultimas 30 filas del dataframe





## Evaluación

### Question 1

1/1 point (graded)

What does CSV stand for?

☒ Comma-separated values

☐ Car sold values

☐ Car state values

☐ None of the above



Submit

You have used 2 of 2 attempts

### Question 2

0/1 point (graded)

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

☒ Row

☐ Column

☐ Each element in the dataset



Submit

You have used 2 of 2 attempts



---

## Question 3

1/1 point (graded)

What is the name of what we want to predict?

☒ Target

☐ Feature

☐ Dataframe



Save

Submit

You have used 1 of 2 attempts

---

## Question 4

1/1 point (graded)

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

☒ `df.head()`

☐ `df.tail()`



Submit

You have used 1 of 1 attempt



## Question 5

1/1 point (graded)

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

☒ `df.dtypes`

☐ `df.head()`

☐ `df.tail()`



Save

Submit

You have used 1 of 2 attempts

## Question 6

1/1 point (graded)

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

☒ `df.describe()`

☐ `df.head()`

☐ `df.tail()`



Save

Submit

You have used 1 of 2 attempts

## Question 7

0/1 point (graded)

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



Submit

You have used 1 of 1 attempt