



Universidad Católica  
**San Pablo**

# Preprocesamiento de Datos

Python para Ciencia de Datos

Graciela Meza Lovon, Yessenia Yari Ramos, Alvaro Mamani Aliaga

# Contenido

- ❖ El preprocesamiento de datos es un paso necesario en el análisis de datos.
- ❖ Es el proceso de convertir o mapear datos de una forma sin procesar a otro formato para prepararlos para un análisis posterior.

# Preprocesamiento: Introducción

- ❖ El preprocesamiento de datos es un paso necesario en el análisis de datos.

# Preprocesamiento: Introducción

- ❖ El preprocesamiento de datos es un paso necesario en el análisis de datos.
- ❖ Es el proceso de convertir o mapear datos de una forma sin procesar a otro formato para prepararlos para un análisis posterior.
- ❖

# Preprocesamiento: Introducción

- ❖ El preprocesamiento de datos es un paso necesario en el análisis de datos.
- ❖ Es el proceso de convertir o mapear datos de una forma sin procesar a otro formato para prepararlos para un análisis posterior.
- ❖ El preprocesamiento de datos a menudo se denomina limpieza de datos o data wrangling.

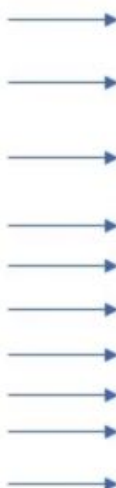
# Preprocesamiento: Introducción

- ❖ La información será cargada en un dataframe.

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40	8.5
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13	3.40	8.3
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13	3.40	7.0

# Preprocesamiento: Introducción

- ❖ La información será cargada en un dataframe.



	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40	8.5
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13	3.40	8.3
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13	3.40	7.0

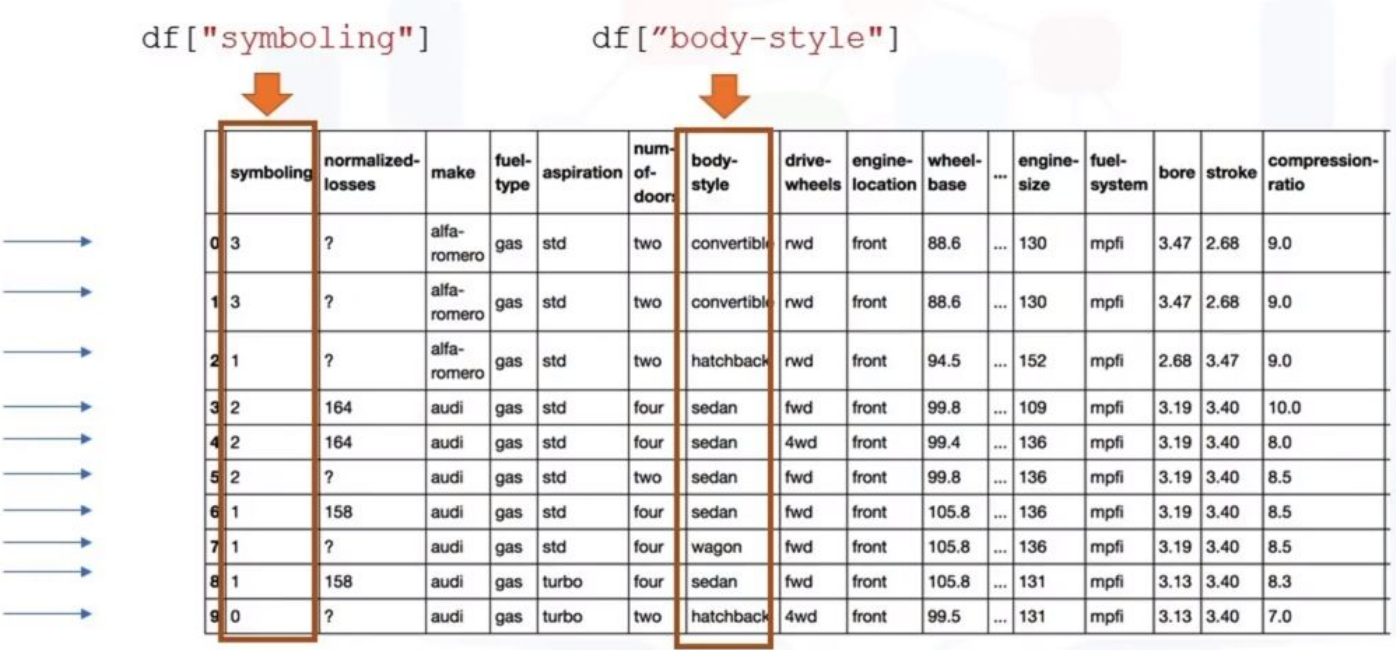


# Preprocesamiento: Introducción

- ❖ La información será cargada en un dataframe.

`df["symboling"]`

`df["body-style"]`




	symboling	normalized-losses	make	fuel-type	aspiration	num-of-door	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0
5	2	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40	8.5
6	1	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
7	1	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
8	1	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13	3.40	8.3
9	0	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13	3.40	7.0

# Preprocesamiento: Introducción

## ❖ Cambiar los atributos

```
df['symboling'] = df['symboling'] + 1
```



	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio
0	4	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
1	4	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0
2	2	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0
3	3	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.40	10.0
4	3	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.40	8.0
5	3	?	audi	gas	std	two	sedan	fwd	front	99.8	...	136	mpfi	3.19	3.40	8.5
6	2	158	audi	gas	std	four	sedan	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
7	2	?	audi	gas	std	four	wagon	fwd	front	105.8	...	136	mpfi	3.19	3.40	8.5
8	2	158	audi	gas	turbo	four	sedan	fwd	front	105.8	...	131	mpfi	3.13	3.40	8.3
9	1	?	audi	gas	turbo	two	hatchback	4wd	front	99.5	...	131	mpfi	3.13	3.40	7.0

## Preprocesamiento: Manejo de Información Faltante

❖ ¿Qué es un dato faltante?:

- Cuando no se almacena ningún valor de datos para la característica de una observación en particular, decimos que esta característica tiene un valor faltante.
- Normalmente, el valor que falta en el conjunto de datos aparece como ?, "N/A", 0, una celda en blanco.

[illegible]

# Preprocesamiento: Manejo de Información Faltante

- ❖ Eliminar datos mediante `dataframes.dropna()`

highway-mpg	price
...	...
20	23875
22	NaN
29	16430
...	...

# Preprocesamiento: Manejo de Información Faltante

- ❖ Eliminar datos mediante `dataframes.dropna()`

highway-mpg	price
...	...
20	23875
22	NaN
29	16430
...	...



highway-mpg	price
...	...
20	23875
29	16430
...	...

`axis=0` drops the entire row

`axis=1` drops the entire column

# Preprocesamiento: Manejo de Información Faltante

- ❖ Eliminar datos mediante `dataframes.dropna()`

highway-mpg	price
...	...
20	23875
22	NaN
29	16430
...	...



highway-mpg	price
...	...
20	23875
29	16430
...	...

`axis=0` drops the entire row  
`axis=1` drops the entire column

```
df.dropna(subset=['normalized-losses'], axis = 0, inplace = True)
```

# Preprocesamiento: Manejo de Información Faltante

- ❖ Reemplazar datos mediante `dataframe.replace(valor_faltante, nuevo valor)`

normalized-losses	make
...	...
164	audi
164	audi
NaN	audi
158	audi
...	...

# Preprocesamiento: Manejo de Información Faltante

- ❖ Reemplazar datos mediante `dataframe.replace(valor faltante, nuevo valor)`

normalized-losses	make
...	...
164	audi
164	audi
NaN	audi
158	audi
...	...



normalized-losses	make
...	...
164	audi
164	audi
162	audi
158	audi
...	...



# Preprocesamiento: Manejo de Información Faltante

- ❖ Reemplazar datos mediante `dataframe.replace(valor faltante, nuevo valor)`

normalized-losses	make
...	...
164	audi
164	audi
NaN	audi
158	audi
...	...



normalized-losses	make
...	...
164	audi
164	audi
162	audi
158	audi
...	...

```
media = df['normalized-losses'].mean()
```

```
df['normalized-losses'].replace(np.nan, media, inplace = True)
```

# Preprocesamiento: Formato de los Datos

- ❖ Los datos generalmente se recopilan de diferentes lugares por diferentes personas que pueden almacenarse en diferentes formatos.

# Preprocesamiento: Formato de los Datos

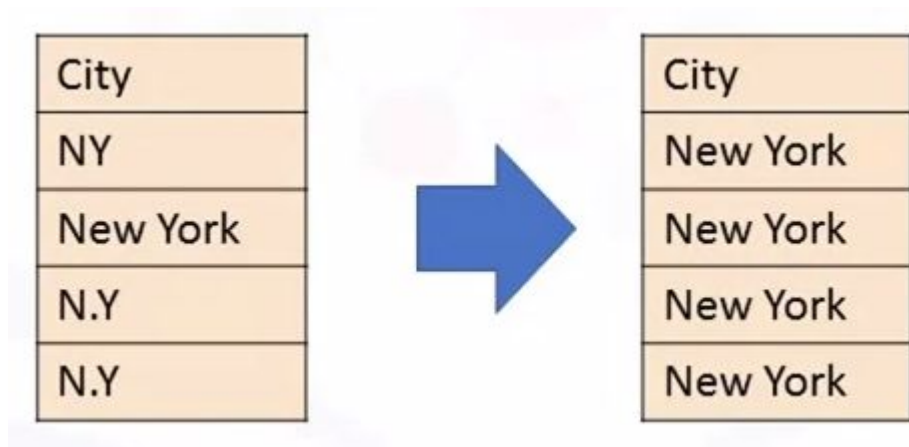
- ❖ Los datos generalmente se recopilan de diferentes lugares por diferentes personas que pueden almacenarse en diferentes formatos.
- ❖ El formateo de datos significa llevar los datos a un estándar común de expresión que permite a los usuarios hacer comparaciones significativas.

# Preprocesamiento: Formato de los Datos

- ❖ Los datos generalmente se recopilan de diferentes lugares por diferentes personas que pueden almacenarse en diferentes formatos.
- ❖ El formateo de datos significa llevar los datos a un estándar común de expresión que permite a los usuarios hacer comparaciones significativas.
- ❖ Como parte de la limpieza del conjunto de datos, el formateo de datos asegura que los datos sean consistentes y fácilmente comprensibles.

# Preprocesamiento: Formato de los Datos

- ❖ Sin formato: confuso, difícil de agrupar y comparar.
- ❖ Con formato: mayor claridad, fácil de agrupar y comparar



```
df['fuel-type'].replace("gas", "Gas", inplace = True)
df.head()
```

# Preprocesamiento: Formato de los Datos

- ❖ Convertir los datos a una unidad más usada, etc.
  - Ejemplo, convertir mpg a L/km en el conjuntos de datos de autos

city- mpg	highway- mpg	price
21	27	13495.0
21	27	16500.0
19	26	16500.0
24	30	13950.0

# Preprocesamiento: Formato de los Datos

- ❖ Convertir los datos a una unidad más usada, etc.
  - Ejemplo, convertir mpg a L/km en el conjuntos de datos de autos

city- mpg	highway- mpg	price		city- L/km	highway- mpg	price
21	27	13495.0		8.928040	27	13495.0
21	27	16500.0	➔	8.928040	27	16500.0
19	26	16500.0		8.077750	26	16500.0
24	30	13950.0		10.203474	30	13950.0

# Preprocesamiento: Formato de los Datos

- ❖ Convertir los datos a una unidad más usada, etc.
  - Ejemplo, convertir mpg a L/100km en el conjuntos de datos de autos

city- mpg	highway- mpg	price		city- L/km	highway- mpg	price
21	27	13495.0	➔	8.928040	27	13495.0
21	27	16500.0		8.928040	27	16500.0
19	26	16500.0		8.077750	26	16500.0
24	30	13950.0		10.203474	30	13950.0

```
df['city-mpg'] =df['city-mpg']/2.35214  
df.rename(columns={'city-mpg': 'city-L/km'}, inplace=True)
```



# Preprocesamiento: Formato de los Datos

- ❖ Tipos de datos incorrectos.

- Ejemplo, el tipo de dato asignado al precio es object, cuando debería ser un entero

```
df['price'].tail()
```

```
200    16845
```

```
201    19045
```

```
202    21485
```

```
203    22470
```

```
204    22625
```

```
Name: price, dtype: object
```

# Preprocesamiento: Formato de los Datos

- ❖ Existen diferentes tipos de datos en Pandas
  - Objects : “Hola”, “mio”, ...
  - Int64: 123, 45
  - Float64: 12.5, 45.999
  - Otros tipos

# Preprocesamiento: Formato de los Datos

- ❖ Para identificar el tipo de dato:
  - Se usa `dataframe.dtypes()`
- ❖ Para convertir un dato a otro tipo
  - Se usa `dataframe.astype()`

```
df['price'] = df['price'].replace(" ", np.nan, inplace=True)
```

```
df['price'] = df['price'].astype(float)
```

```
df['price'].tail()
```

```
200    16845.0
201    19045.0
202    21485.0
203    22470.0
204    22625.0
Name: price, dtype: float64
```

*¡Gracias!*