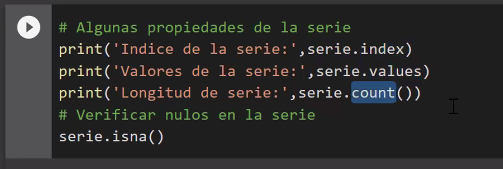
**CLASE 4**



Cuando importamos librerías en Python se le pone un “Apodo”, por ejemplo, a PANDAS se le pone “pd”.

Para crear una serie en PANDAS la “S” tiene que ser mayúscula, es CASE SENSITIVE.

Panas en mas flexible para crear listas. Numpy permite crear un ARRAY uniforme, ya sea numérico, STRING, etc.

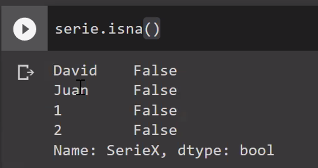


El **count** nos dice cuanto de longitud tiene la SERIE en este caso. Con el ***index*** llamamos a los índices, y con el ***values*** a los valores de la SERIE.

Los “objetos” son los tipos más usuales.

Co el “isna” podemos determinar si tenemos valores NULOS.

Cuando es un atributo no ponemos paréntesis, solo lo hacemos con los métodos.

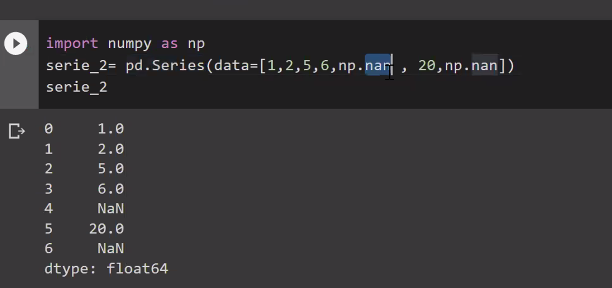


Acá podemos detectar los valores perdidos.

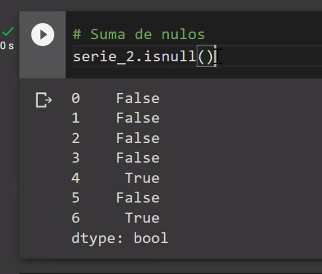


Con el “loc” y el “iloc” ingresamos a la misma celda, pero de distinta manera.

En el “47” podemos filtrar por varios índices, en este caso es el de David (indice) y el “1”(indice).

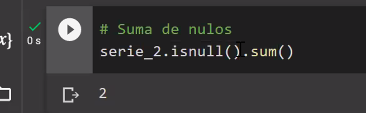


Acá podemos ver una SERIE creada con valores NULOS.

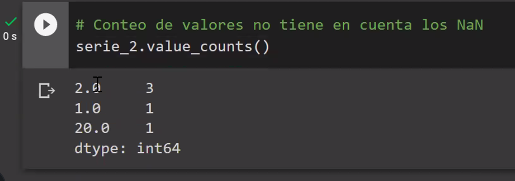


El “isnull” puede ser algo parecido a “isna”. No es practica si queremos saber cuanto valores nulos tenemos en este caso.

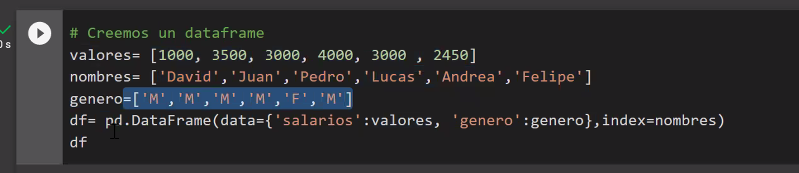
Podemos agregarle una función.



Podemos contar valores.

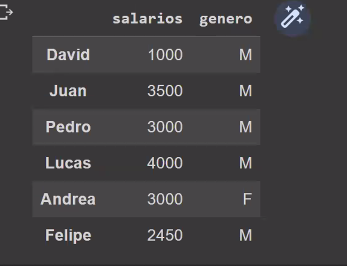


Es la cantidad de valores y sus repeticiones en una SERIE.



Acá creamos el DF.

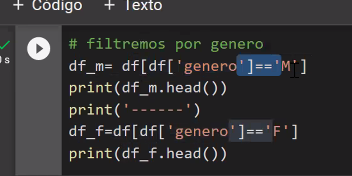
Acá especificamos los valores, el genero en este caso y el índice, según las variables que tenemos como listas.



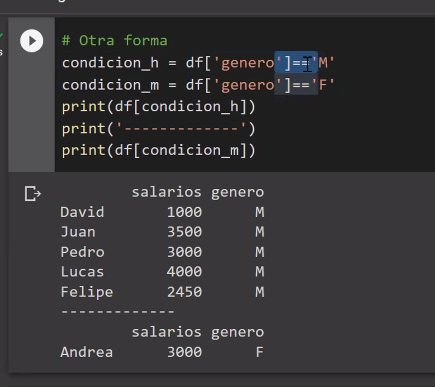
Si no especificamos el índice, pone uno por defecto q es numérico y empieza desde el 0.



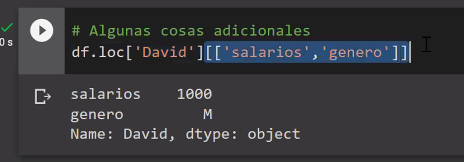
Así tenemos la información de la función con una ayuda.



Acá tenemos un filtro con una comparación condicional.

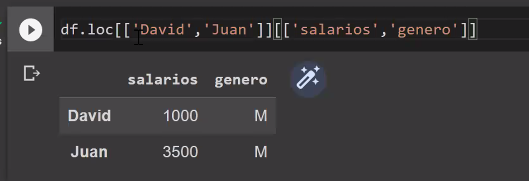


Esta es una forma distinta de filtrar con condicional.



Acá le decimos que dentro del índice “David” nos traiga “salarios” y “generos”.

Cuando queremos entrar al valor de la columna ponemos los dobles corchetes.



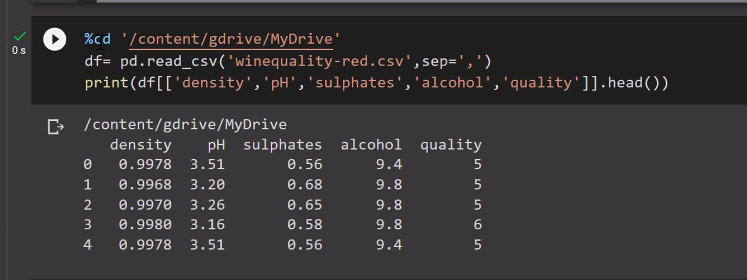
Con el doble corchete les estamos pasando una lista de índices.

LECTURA DE DATOS



No es lo mejor utilizar GOOGLE DRIVE para leer archivos externos. Esto solo conecta.

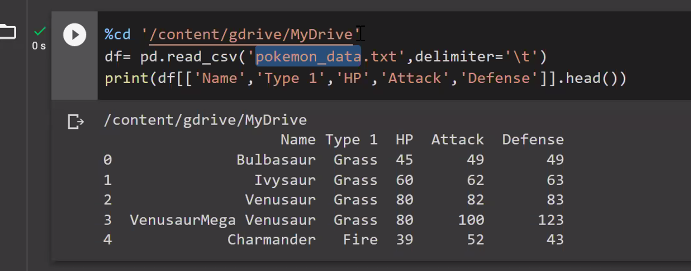
Para leer un archivo de DRIVE tenemos que conectar nuestro DRIVE con COLAB, si que estamos utilizando este editor.



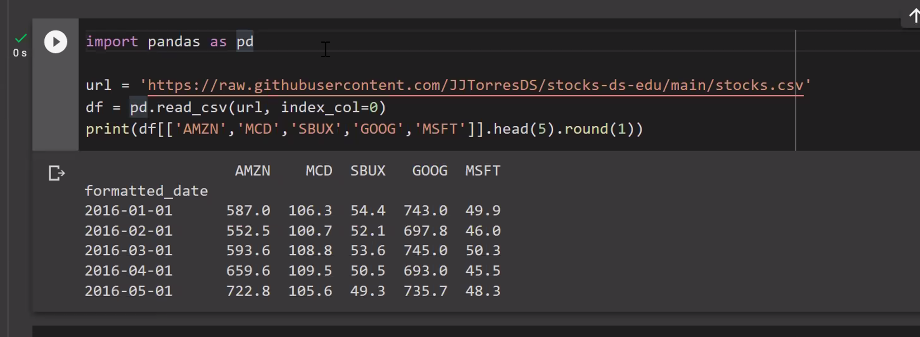
El “%cd” arma la ruta de carpetas donde va a estar el archivo.

En la segunda línea le pasamos el archivo y luego un separador.

Cuando otra persona quiera acceder a esto, si no especificamos que puedan verla, les va a tirar errores.



Podemos crear una subcarpeta dentro “MyDrive”, pero la tenemos que especificar la ruta.



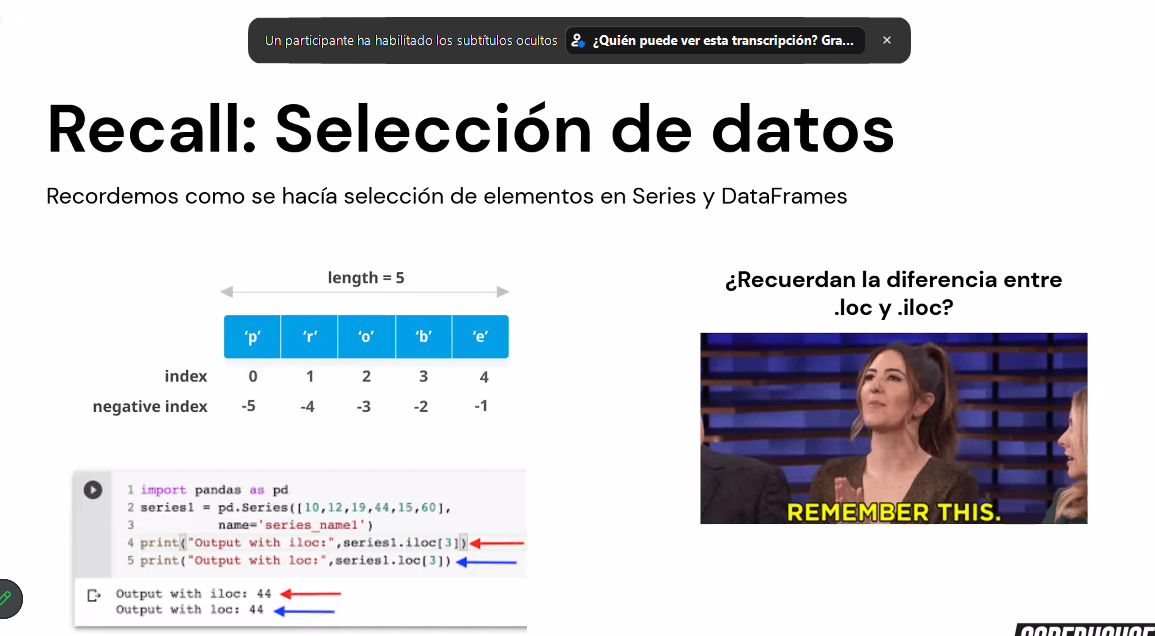
Acá estamos utilizando el repositorio en GitHub, con la URL que nos da para clonar ese archivo indicado.

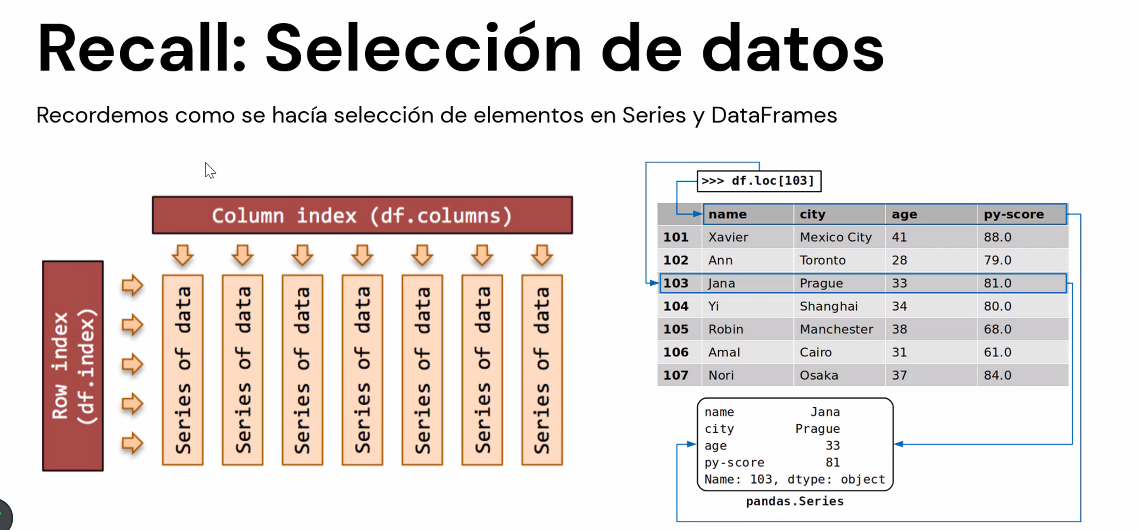
Accedemos al DataSet sin problemas sin necesidad de utilizar autenticación.

**VISUALIZACION**

Hay varias librerías para este fin.

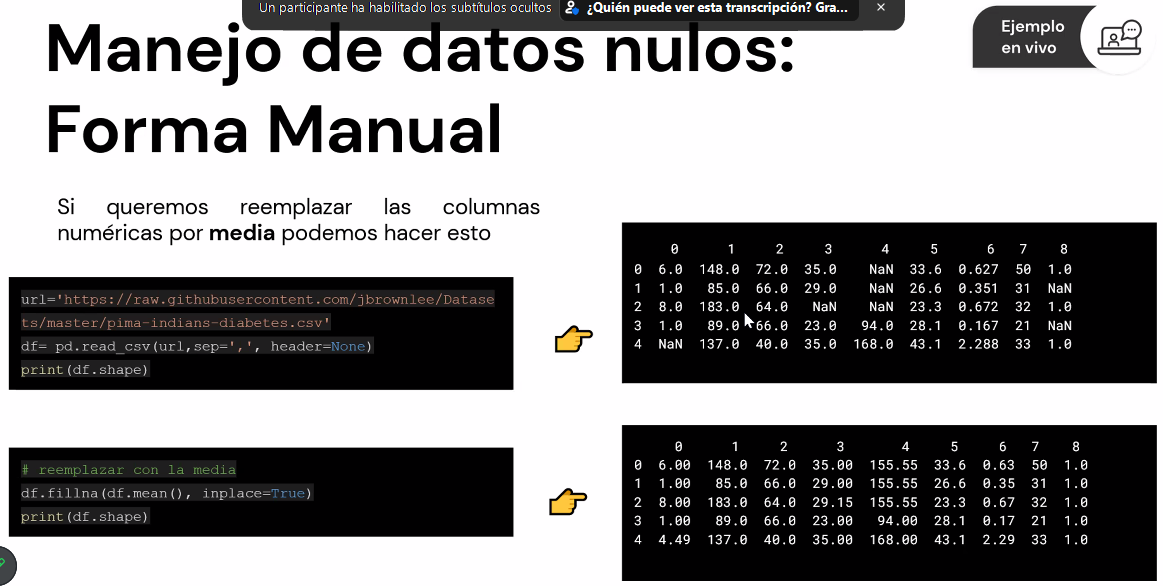
***RECALL DE CLASES ANTERIORES***





Todo esto lo podemos ver en los SLIDES que tenemos de las clases.





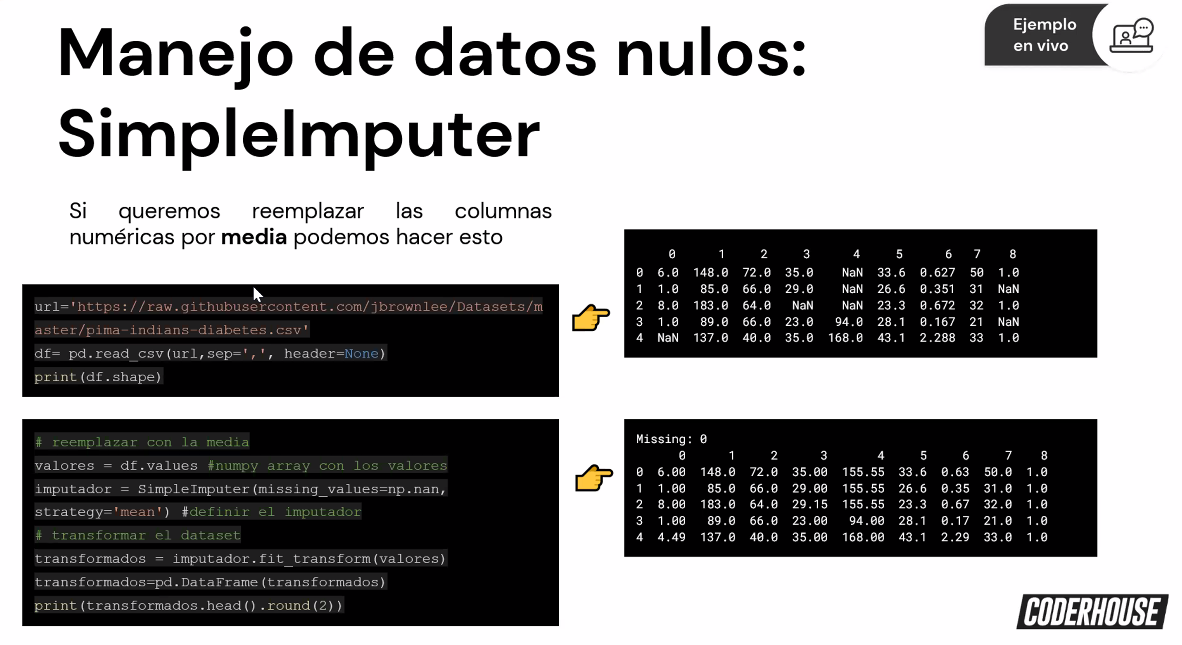
Si empezamos a eliminar INSTACIANCIAS con valores NAN, nos quedaríamos sin datos, en este caso nos quedaríamos sin ningún dato.

No podemos dejar valores perdidos, tenemos que manejarlos de alguna forma, ya sea eliminar una FEATURES (columna), o solo sumar los valores reales ignorando los NAN.

Si reemplazamos por 0, puede no ser una solución, puede perjudicar el promedio, por ejemplo.

La MEDIANA es el valor que separa el DS en los mismos valores (existentes) hacia arriba y hacia abajo. Es el valor del medio.

La MODA es el valor que más se repite.

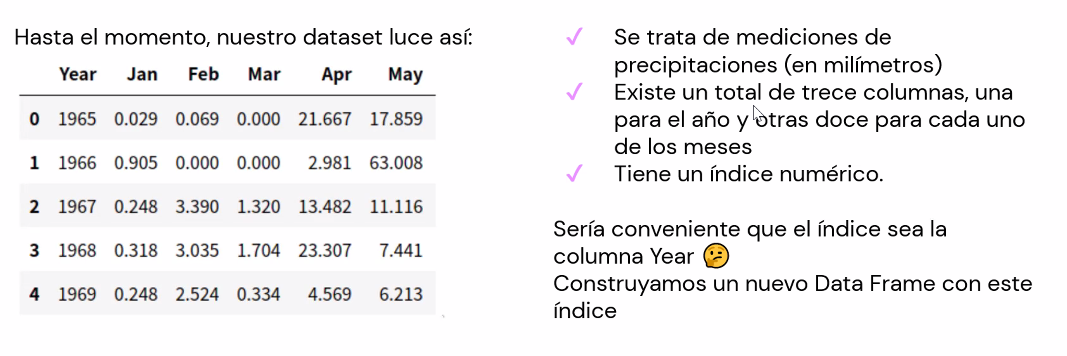


Si el valor perdido esta especificado de otra forma, no con el NAN, utilizamos el “SimpleImputer”.

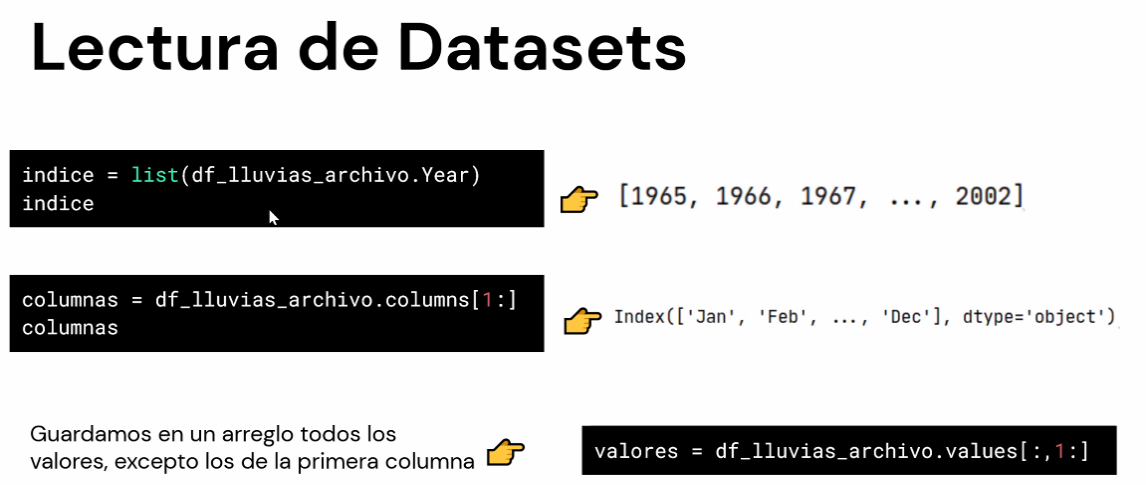
Estos son formas de completar los valores perdidos.

FUNCION DE AGREGACION

Tomar una serie de datos y nos arroja un valor.



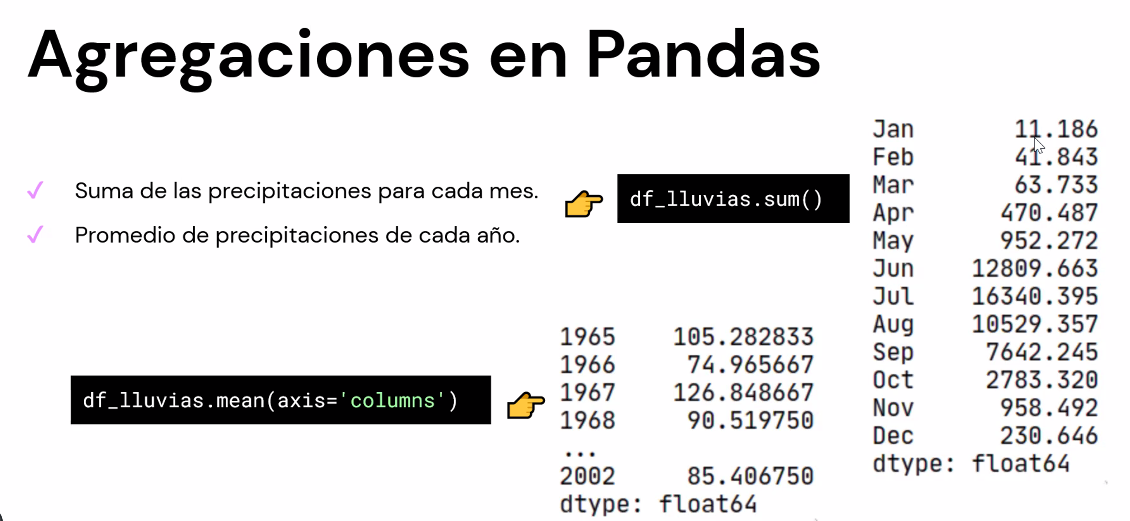
En este caso sería mejor q YEAR sea índice.



Ahí le decimos q las columnas son todas menos la YEAR, q en este caso se convierte en INDEX.



Ahora podemos hacer las AGREGACIONES.

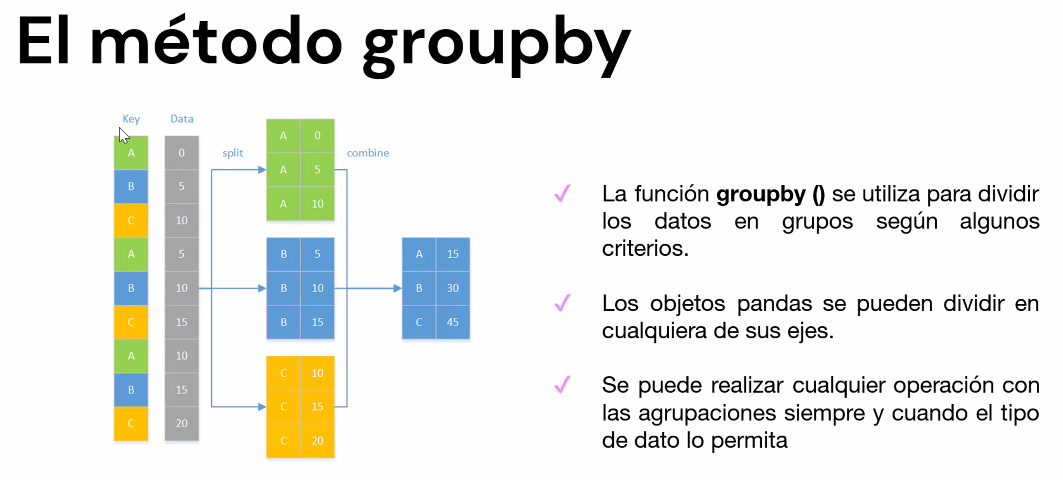


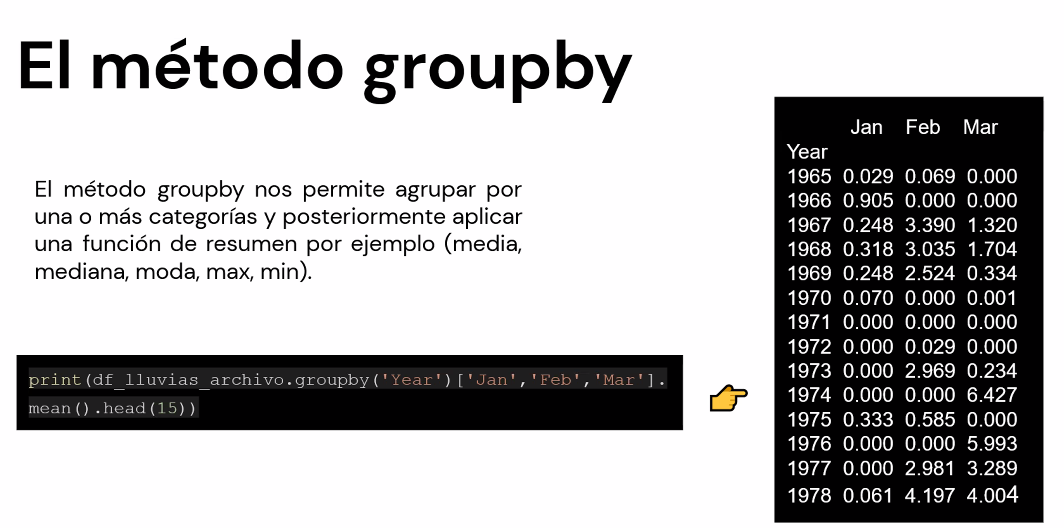


Con esta función nos resume el DataSet. No incluyamos el “round”. El DESCRIBE, tiene ***percentil***. Son el 25%, 50% y el 75%.

Es un método muy utilizado en el DataScience.

Pasemos a otro método.

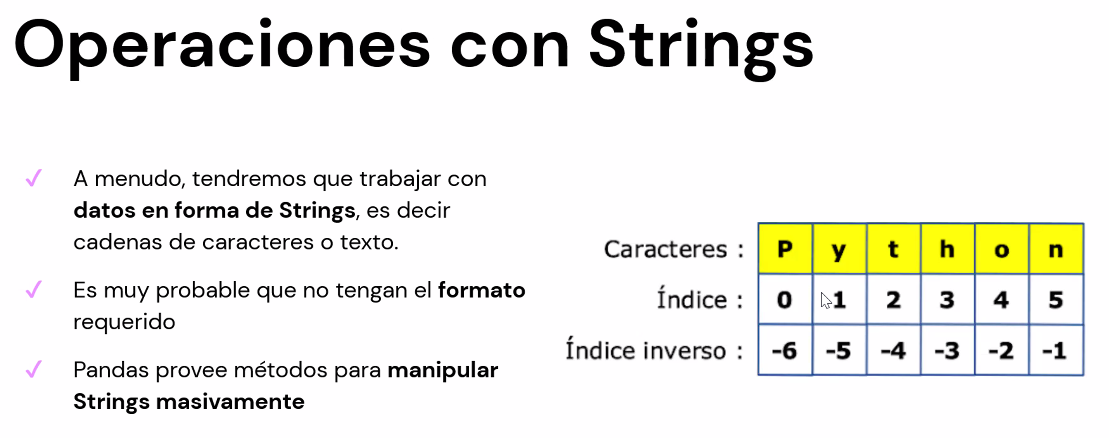


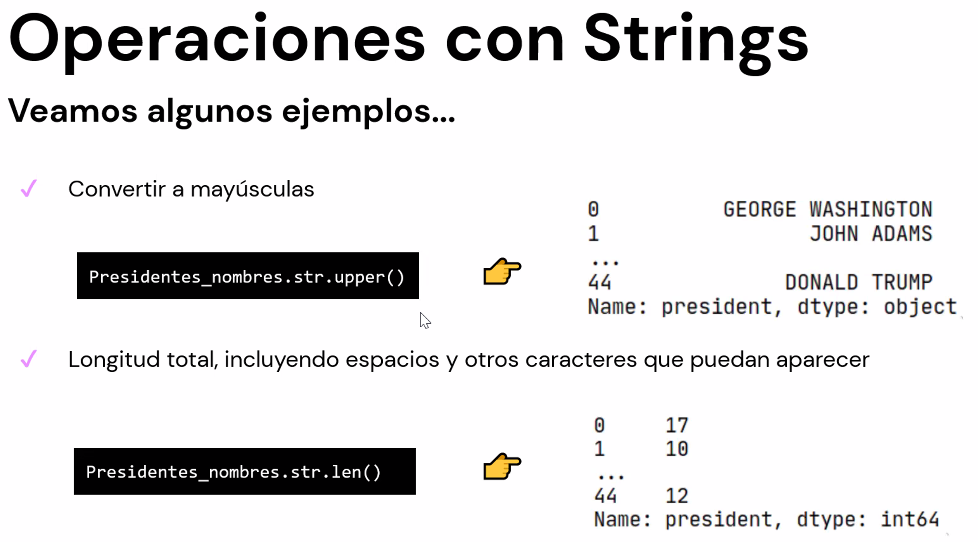


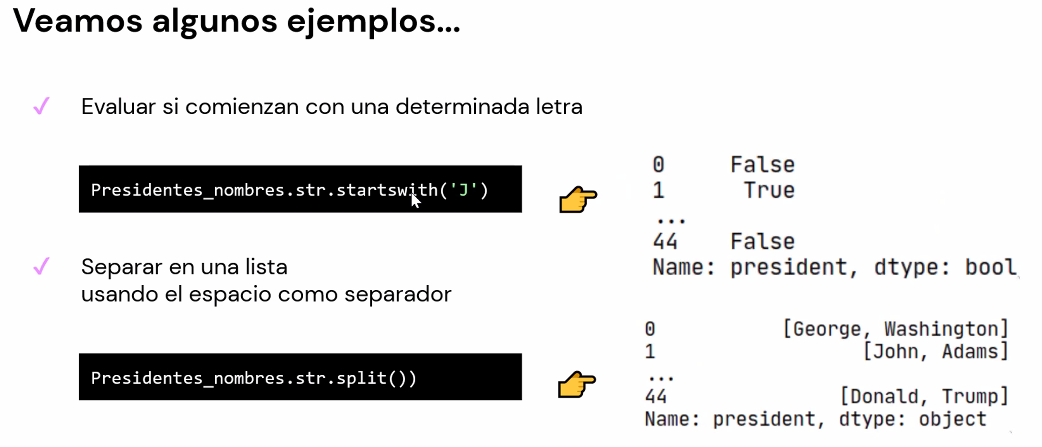
Acá le decimos q agrupe por año, pero solo algunas columnas que son los meses. Nos calcula media y nos muestra los primeros 15.

Es como la “pivot table” de EXCEL.

Si no especificamos las columnas, nos toman todas.

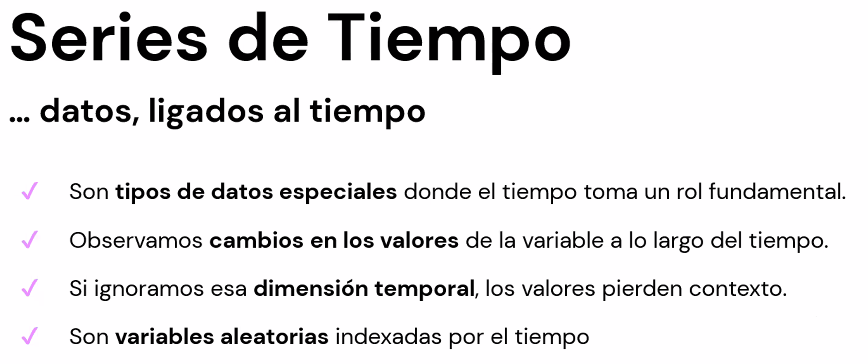






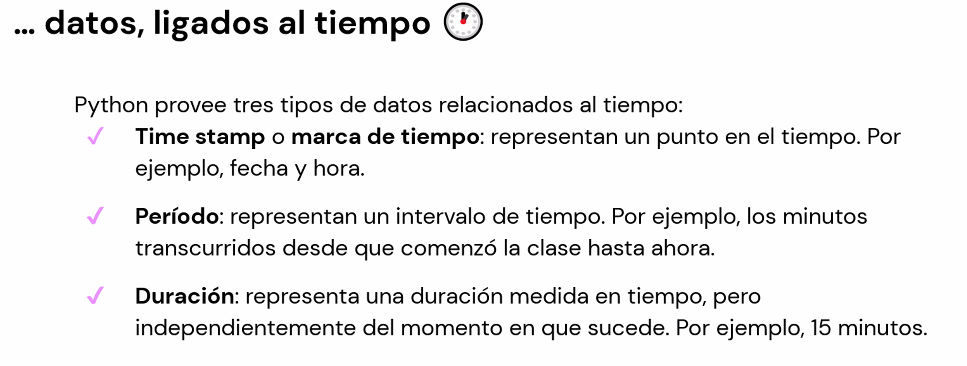
Podemos aplicar filtros y separadores, si este último lo dejamos vacío, separa por espacios vacíos.

**SERIES DE TIEMPO**

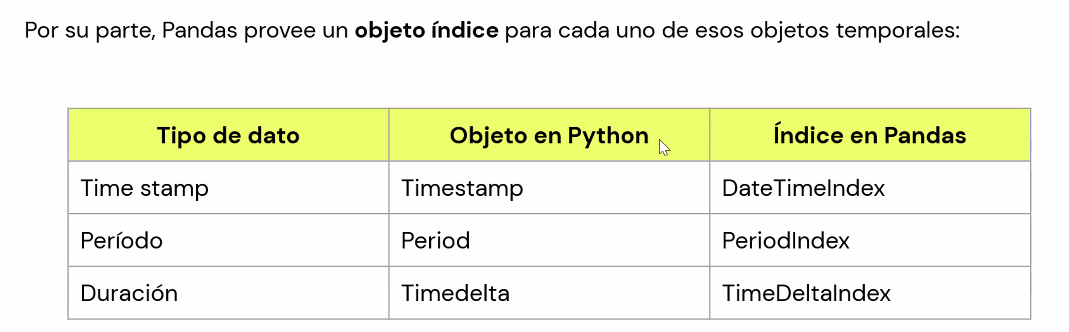


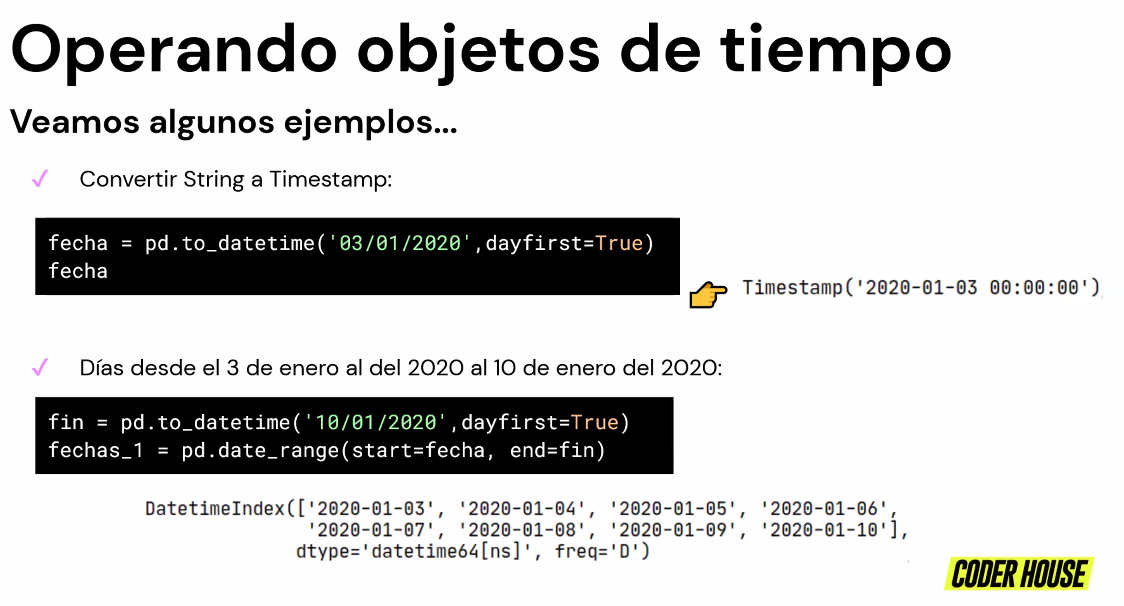
Hay una variable TIEMPO q es fundamental para la correcta interpretación del análisis.

La DIMENSION TEMPORAL es relevante para el análisis.



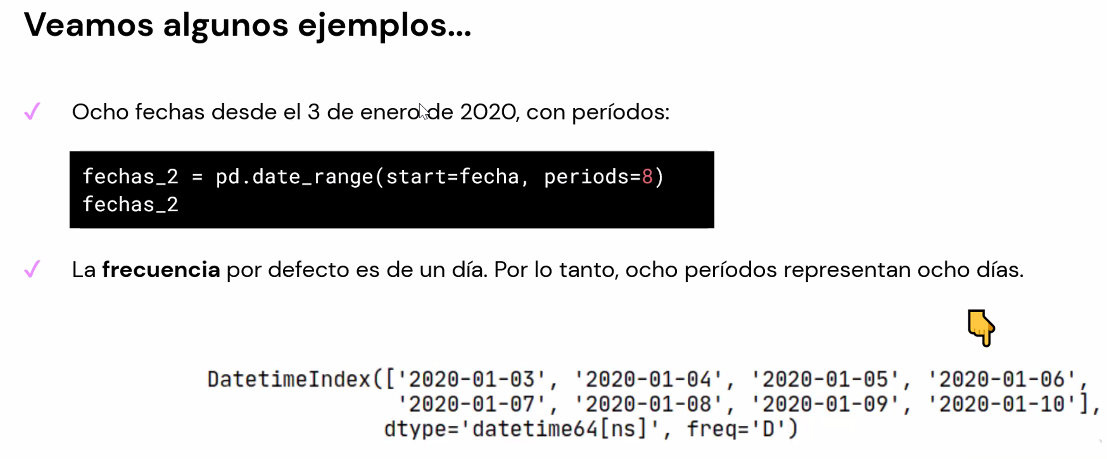
La DURACION es cuánto tiempo duro un proceso.



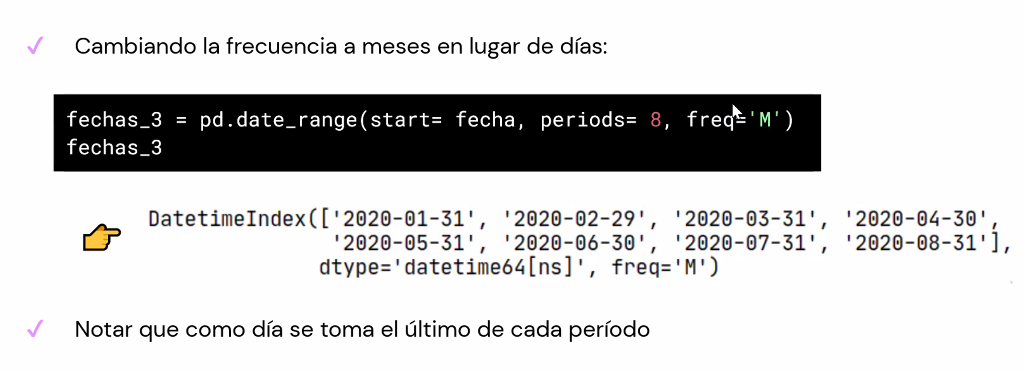


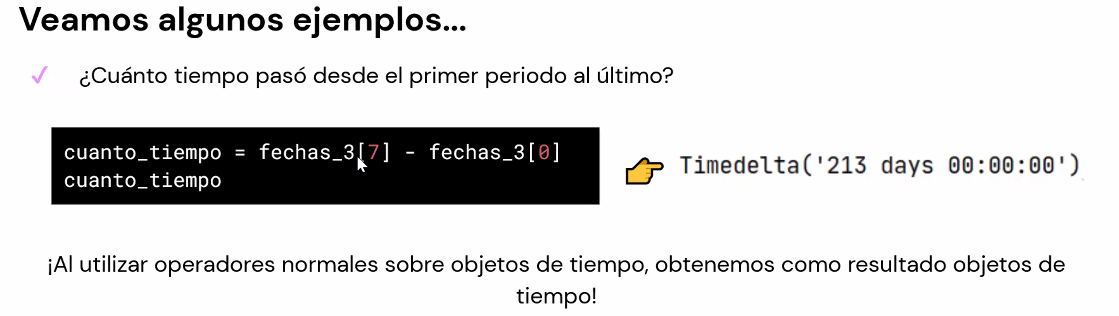
El parámetro “dayfirst”, anteponemos el día antes que el año, o sea, en el inicio del TIMESTAMP.

En el segundo ejemplo, podemos ver los valores q hay en el medio del rango establecido.

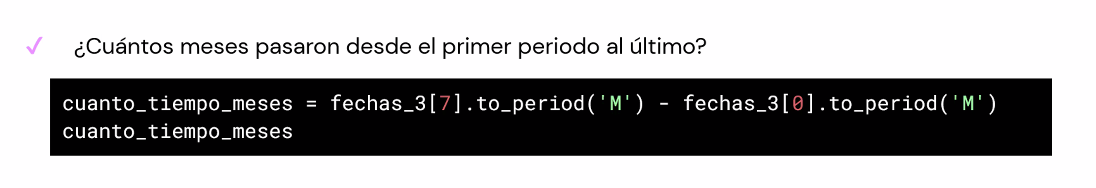


Esta es otra forma de establecer un rango.





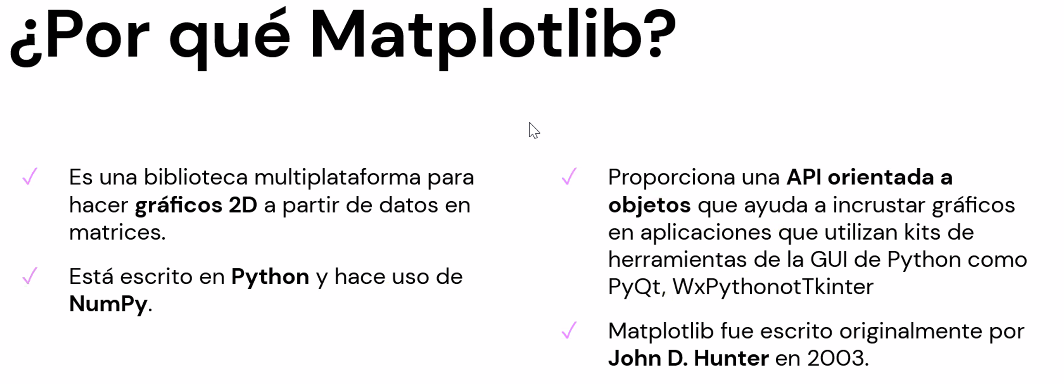
Por defecto cuenta los días en este caso.



Podemos especificar meses.

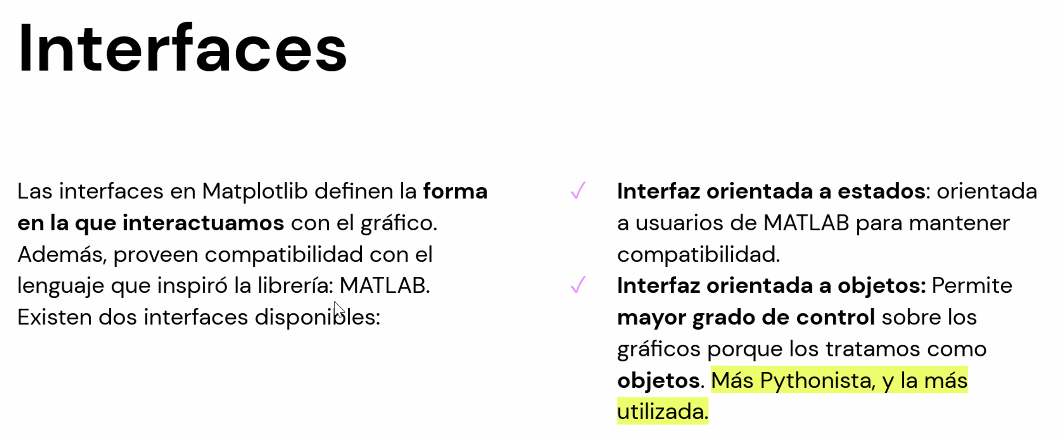
**VISUALIZACIONES**

MATPLOTLIB



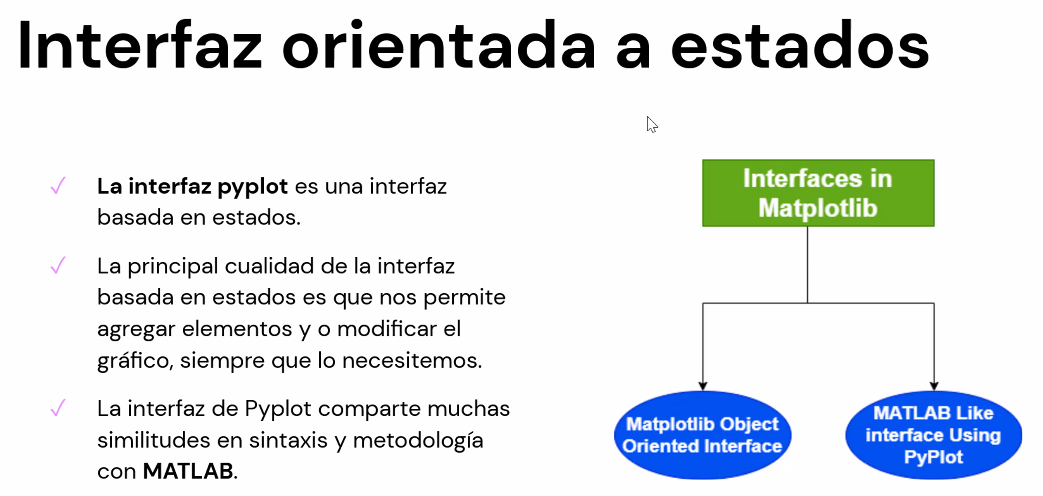
Son buenas para visualizaciones básicas.



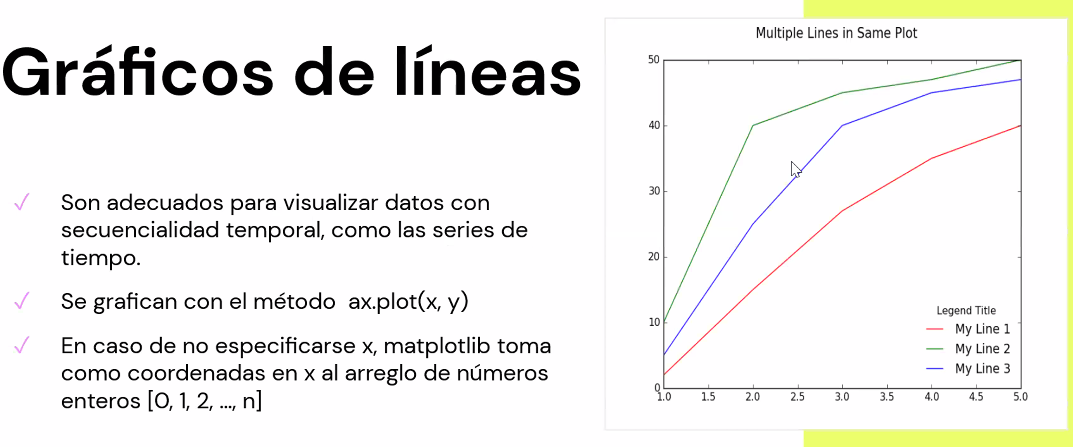


Este curso utilizara la ORIENTADA A OBJETOS.

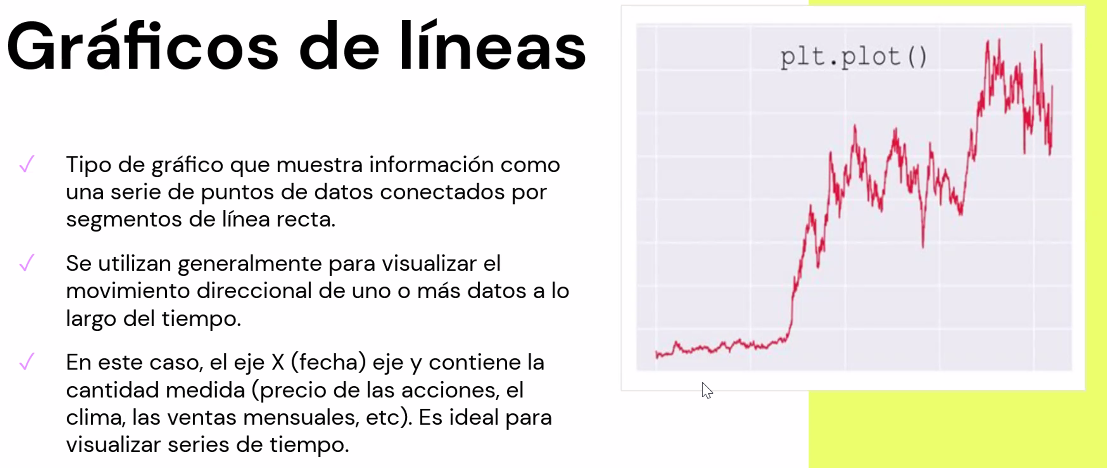
**PYPLOT**



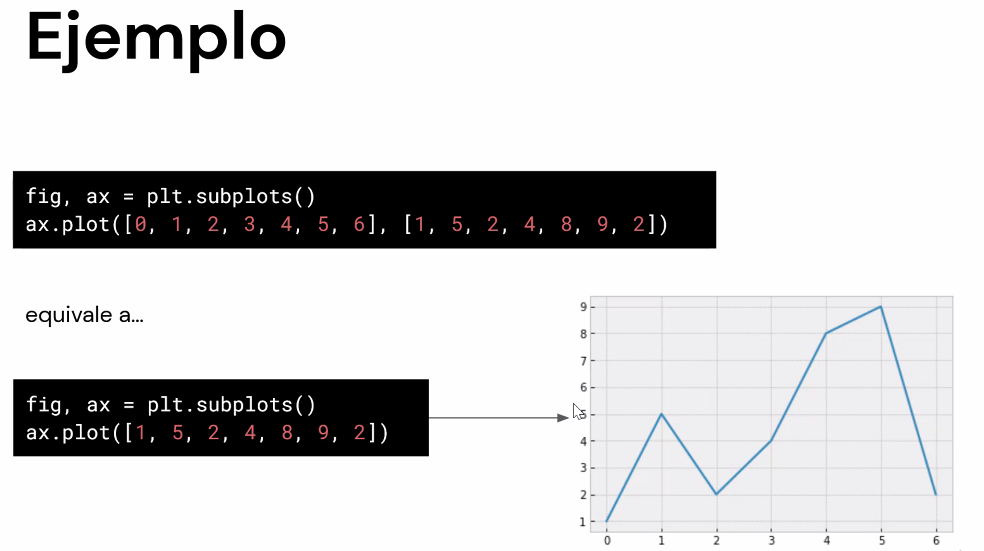
**TIPOS DE GRAFICOS**



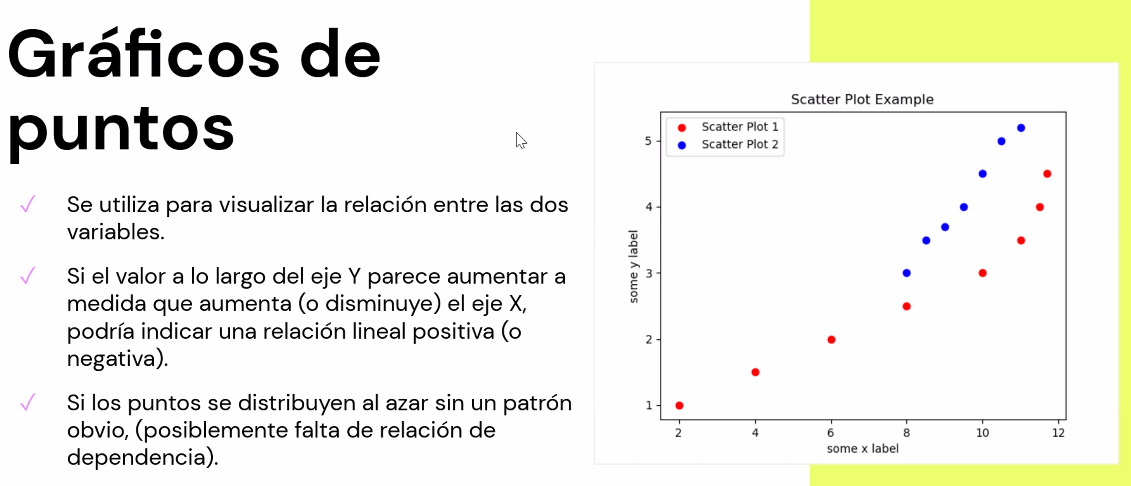
Si no especificamos el grafico, por defecto nos muestra el grafico de líneas.



Si no es con serie tiempo, no deberíamos utilizar el grafico de líneas.

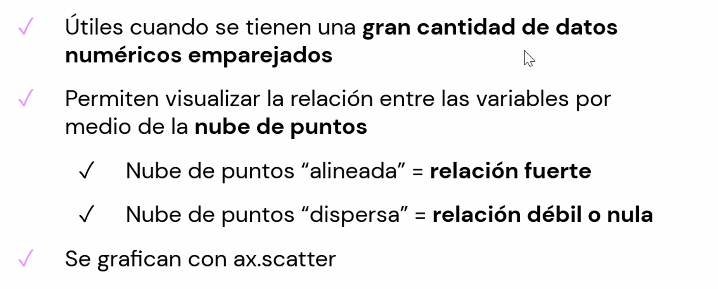


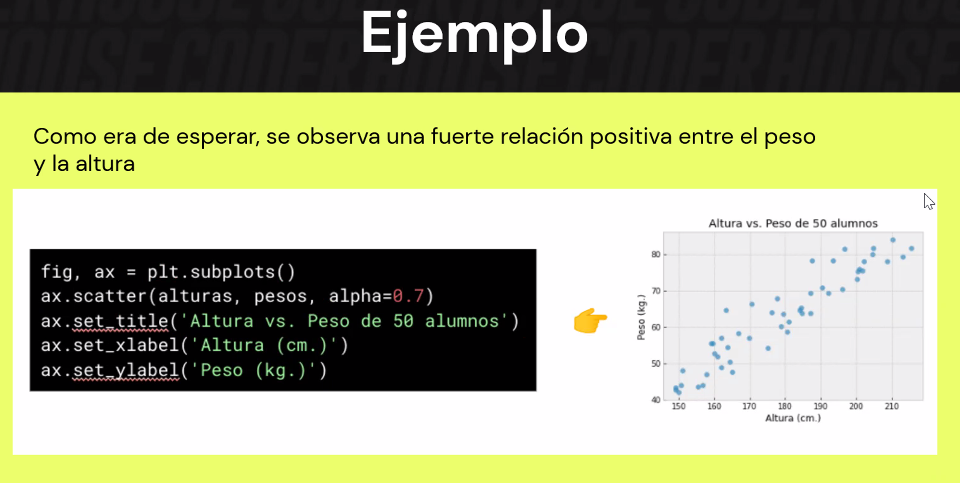
**GRAFICOS DE PUNTOS**

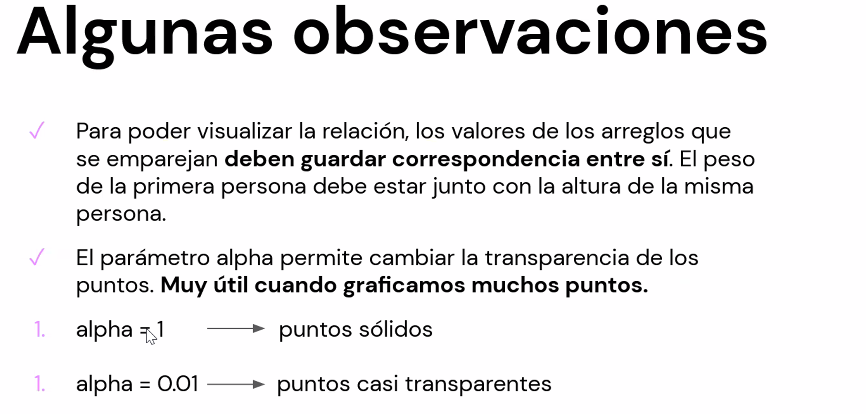


Vemos la relación entre dos variables. Su nombre en inglés, y como debería llamarlo, “SCATTER PLOT”.

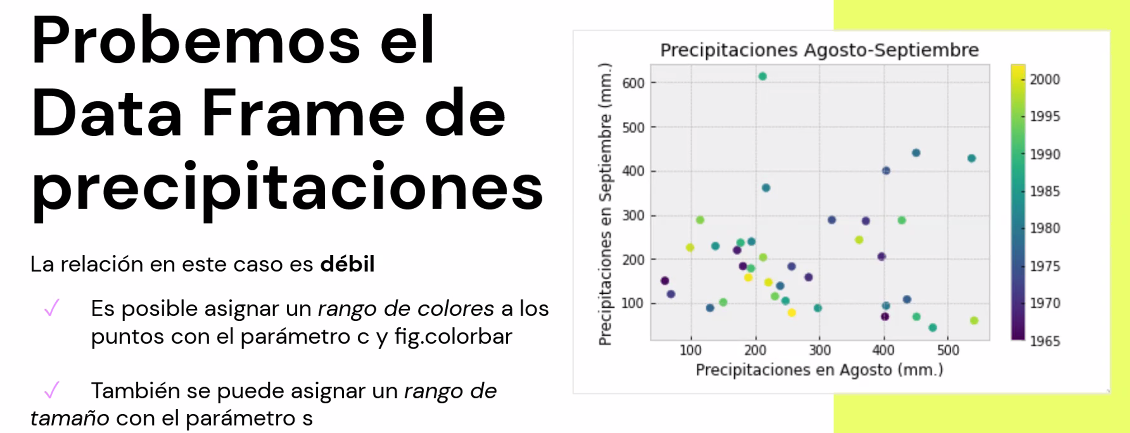
En este ejemplo, podemos ver q estas dos variables están relacionadas.





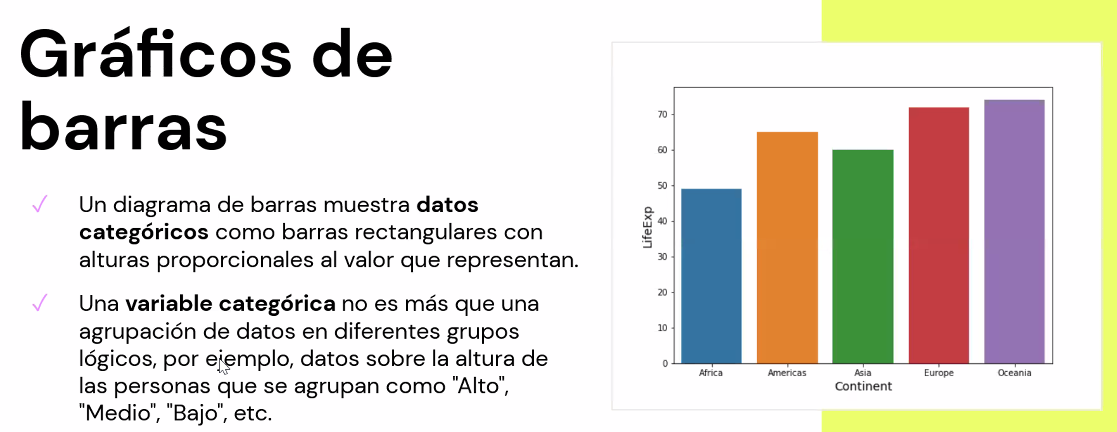


Es útil cuando hay puntos que se pisan con otros, con muchos datos interpuestos uno con otros.

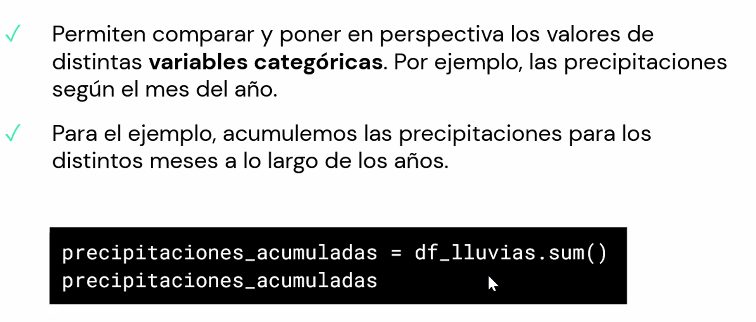


Acá no relación entre las variables.

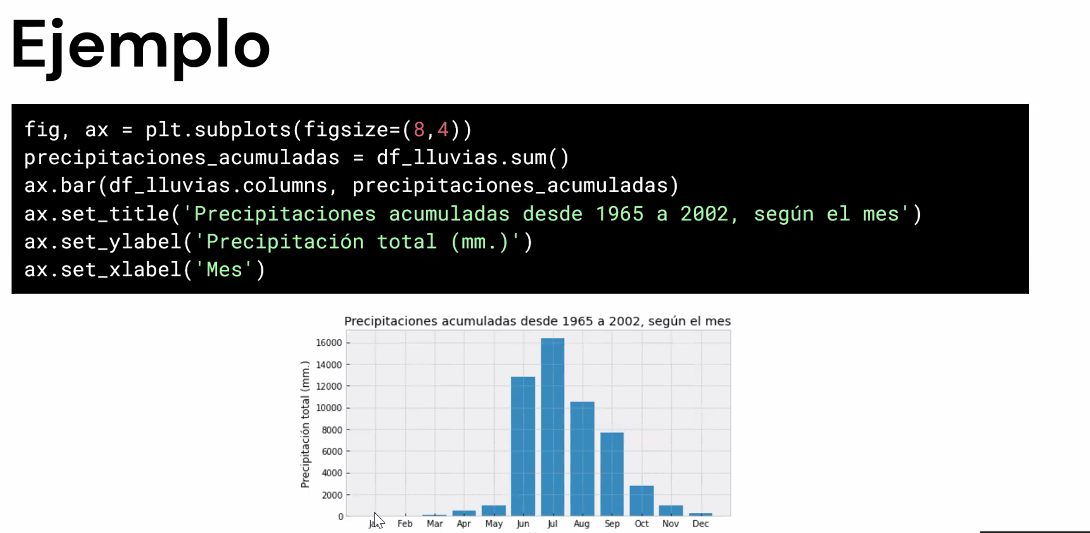
Estas son dos FEATURES. Aportan el mismo valor, podemos eliminar alguno y tener mayor poder predictivo.



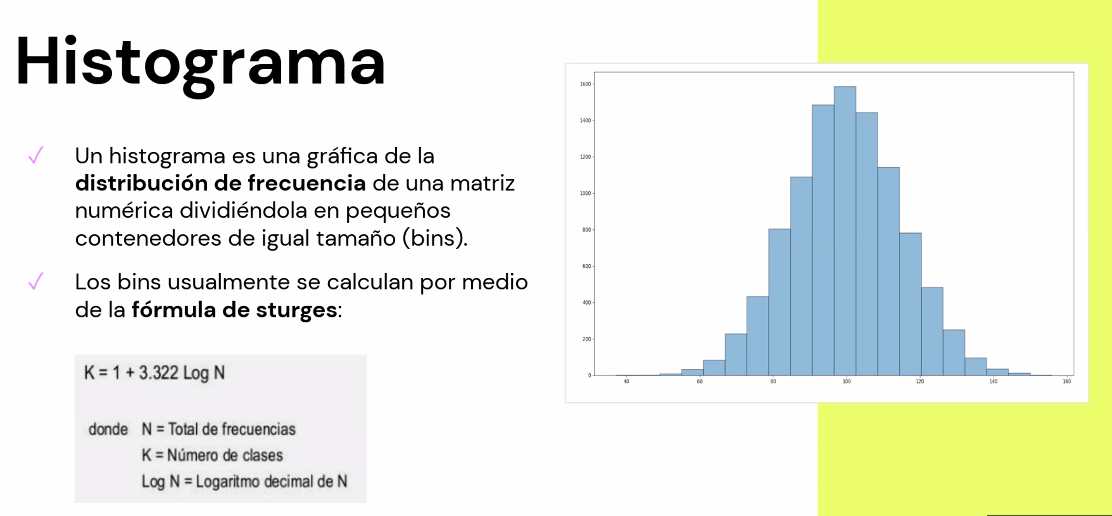
Lo usamos cuando tenemos variables de tipo categórica.



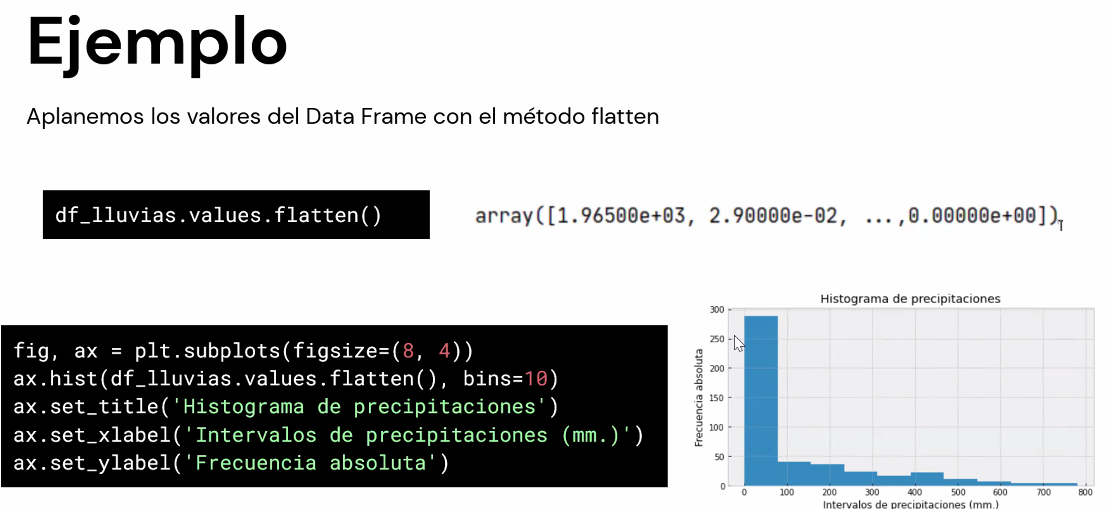
Cualquiera de las funciones q vimos pueden ser PLOTEADAS por el grafico de barras.



**HISTOGRAMA**



Es parecido al de BARRAS, pero en realidad son distintos. Nos sirve para buscar un SESGO. El SESGO es cuando la orientación del grafico va para la izquierda o a la derecha.



Acá podemos ver que el SESGO esta hacia la izquierda.

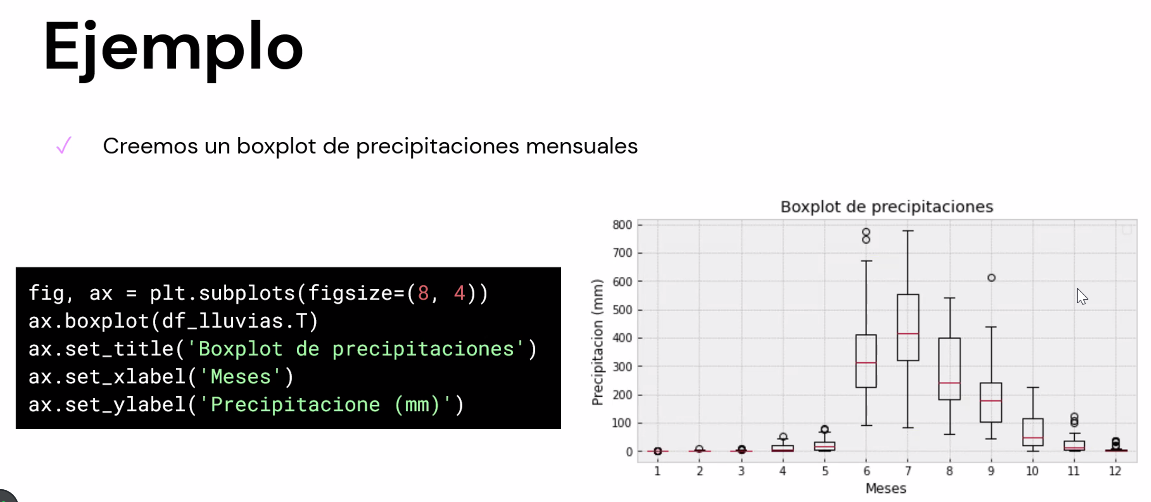
**BOXPLOT**

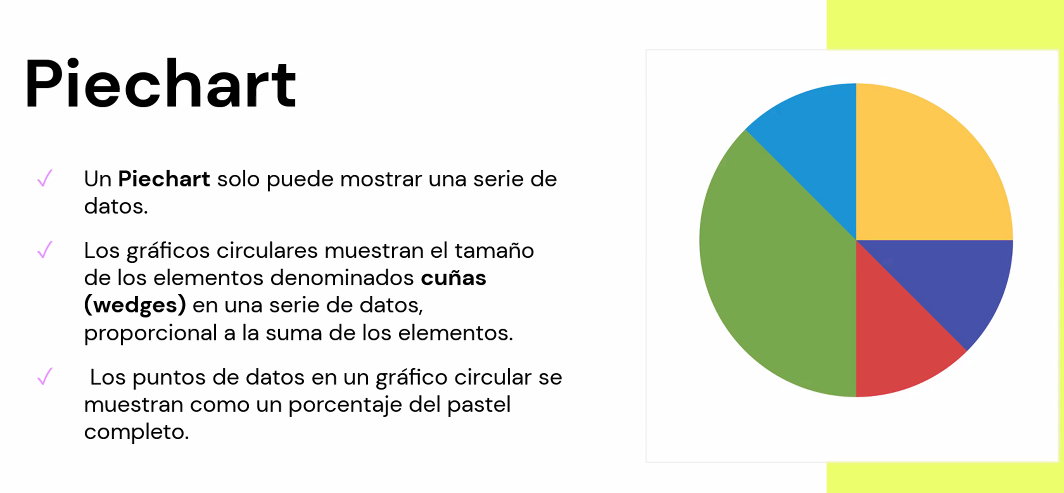


Nos muestra la mediana de nuestro FEATURE como primer paso, después completa el percentil 25, hacia arriba y hacia abajo dentro de la CAJA (zona pintada de verde), que es donde están los datos más cercanos a la MEDIANA. El resto está hacia afuera.

En este caso el ejemplo es asimétrico.

El BIGOTE va a llegar hasta el máximo y el mínimo, la recomendación es que quede en el 5 y en el 95 percentil, todo lo que queda se pueda eliminar (Outliers). Como ejemplo, si fueran edades, dejaríamos afuera a los más jóvenes y más viejos. Dejamos valores afuera para que la predicción del modelo sea mas precisa.





Esto muestra como se distribuyen los datos dentro de una variable categórica.

