**Clase 06 - Introducción a la manipulación de datos con Pandas (Parte II)**

INTRODUCCIÓN A LA LIBRERÍA EN PANDAS

Fuentes de datos disponibles

* En esta pagina encontrarás muchas más APIs para descargar información: <https://github.com/public-apis/public-apis#finance>

SERIES Y DATA FRAME

Manejo de datos nulos

* Los **datos nulos** son uno de los problemas más comunes a los que se enfrenta un Data Scientist
* Son un **problema** porque muchos algoritmos no están diseñados para trabajar con ellos y pueden **disminuir el performance** en general de las tareas de aprendizaje automático
* Son más **comunes** de lo que las personas piensan.
* Pueden surgir por malos esquemas de muestreo, falla de sensores, renuencia a responder en encuestas, malos diseños de captura de datos

Para poder manejar estos datos, en general, se usan las siguientes estrategias:

* Introducir un **valor constante** para los nulos o una categoría llamada Desconocido en variables categóricas
* Reemplazar por un valor seleccionado al azar de los otros registros:
  + Usar la **media, mediana o moda** para rellenar el valor
  + Valor estimado usando un **modelo**

Podemos aplicar estas técnicas de forma manual o usando

[**SimpleImputer** de **ScikitLearn**](https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html)

Agregaciones en Pandas

* Pandas permite realizar agregaciones sobre Data Frames enteros o porciones del mismo.

El método Describe

Este método nos ayuda a…

* Obtener un breve resumen del Data Frame con describe()
* Redondear los valores de un Data Frame con el método round()
* Si **transponemos** el Data Frame antes de aplicar describe, obtenemos el resumen según el año

El método Groupby

* La función **groupby ()** se utiliza para dividir los datos en grupos según algunos criterios.
* Los objetos pandas se pueden dividir en cualquiera de sus ejes.
* Se puede realizar cualquier operación con las agrupaciones siempre y cuando el tipo de dato lo permita