Taller de Python para ciencias de datos

Pablo Belmar Apablaza

Temas

- Bibliotecas mas utilizadas
- Manejo de datos con Python
- Datos como un DataFrame
- Limpieza de datos utilizando python

Bibliotecas

- *Una de las fortalezas que tiene Python es la gran base de datos de bibliotecas externas que tiene para desarrollar aplicaciones en distintas áreas de aplicación.*

Bibliotecas

 *Visualización:*	*Matplotlib* *Seaborn*
 *Análisis de datos:*	*NumPy* *SciPy* *Pandas*
 *Machine learning:*	*Scikit-learn* *LightGBM*
 *Deep learning:*	*Keras* *TensorFlow*
 *Procesamiento de lenguaje natural:*	NLTK Gensim

Bibliotecas

***SciPy**: proporciona rutinas numéricas eficientes fáciles de usar y opera en las mismas estructuras de datos proporcionadas por NumPy. Con SciPy se puede realizar integración numérica, optimización, interpolación, transformadas de Fourier, álgebra lineal, estadística y mucho más. (https://scipy.org/)*

***Pandas**: destaca por lo fácil y flexible que hace la manipulación de datos y el análisis de datos a traves de 2 estructuras de datos principales, las Series para datos en una dimensión y DataFrame para datos en dos dimensiones. (https://pandas.pydata.org/)*

***Scikit-learn**: es una biblioteca para Machine Learning y Análisis de Datos con una gran cantidad de técnicas de aprendizaje automático para realizar aprendizaje supervisado y no supervisado. (https://scikit-learn.org/stable/index.html)*

Bibliotecas

***Keras**: Keras es una biblioteca de alto nivel para trabajar con redes neuronales. El interfaz de Keras es mucho más fácil de usar que el de TensorFlow. (https://keras.io/)*

***TensorFlow**: es una biblioteca desarrollada por Google, para realizar cálculos numéricos mediante diagramas de flujo de datos utilizada para deep learning y otras aplicaciones de cálculo científico. (https://www.tensorflow.org/)*

Pandas - Adquisición de datos

- Existen varios formatos en que se pueden presentar los conjuntos de datos: .csv, .json, .xlsx, entre otros. Los conjuntos de datos pueden estar almacenados en diferentes lugares, en la máquina local o en algún lugar de la red (LAN, Internet, Nube)
- cargar datos dentro de Jupyter Notebook o Python

Pandas - Adquisición de datos

- Se va a trabajar con un conjunto de datos automovilístico, el cual, está con un formato CSV (del inglés, comma separated value)

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

Pandas - Adquisición de datos

La biblioteca **Pandas** es una herramienta util que permite leer diversos conjuntos de datos y convertirlos en un **dataframe.**

Para trabajar con Jupyter notebook, el intérprete de Python se debe instalar la biblioteca de pandas previamente, de manera tal, que cuando se importe la biblioteca no genere un mensaje de error porque no la encuentra.

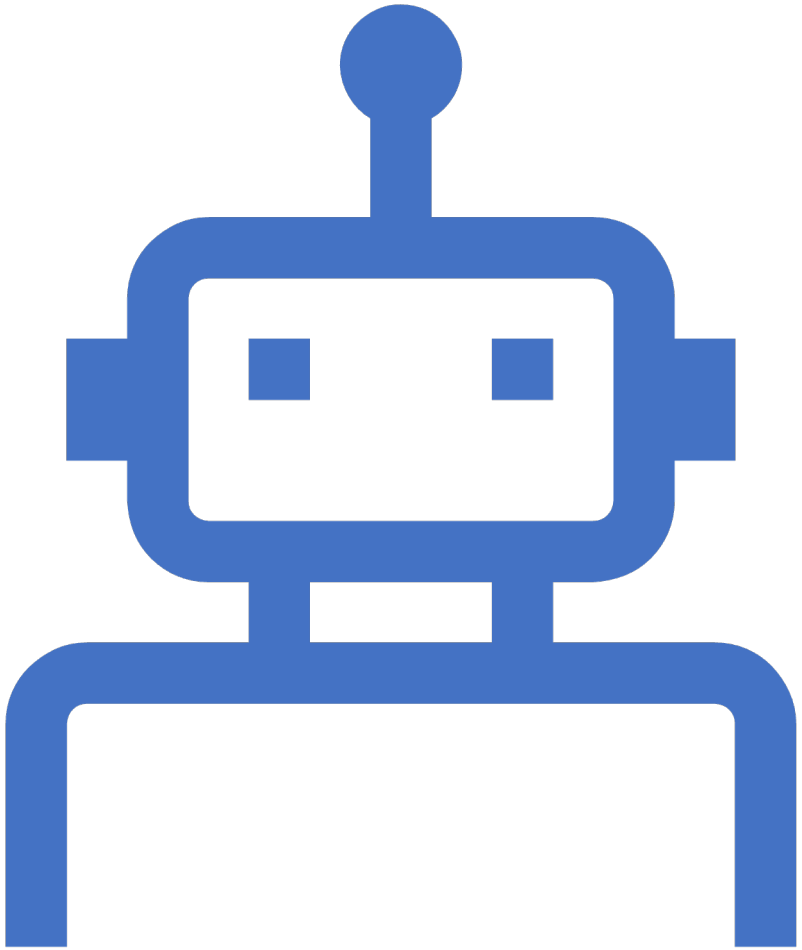
En una terminal escribir lo siguiente: pip install pandas

Vamos!!

Pandas - Limpieza de datos

Consiste en convertir datos desde su formato original a un formato que presenta mejores características para su análisis

¿Cuál es la tasa de consumo de combustible en L/100Km para el automóvil diesel?



Vamos!!