Argentina Programa

Introducción Pandas

Que es un Data Set?

Un data set, también conocido como conjunto de datos, es una colección de datos organizados y estructurados que se utilizan en el análisis, la investigación o la creación de modelos y algoritmos en diversas disciplinas, como la ciencia de datos, la inteligencia artificial, el aprendizaje automático y la estadística.

Un data set puede contener diferentes tipos de información, como números, texto, imágenes, audio u otros formatos de datos. Estos datos se recopilan y almacenan en un formato coherente y estructurado para que puedan ser procesados y analizados de manera eficiente.

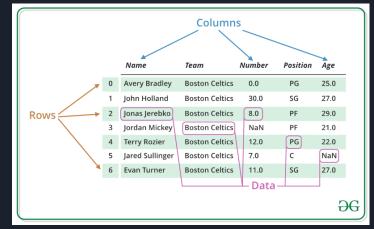
Pandas

Pandas es una biblioteca de Python que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento. Es ampliamente utilizada en el campo de la ciencia de datos y análisis para manipular y analizar datos de manera eficiente.



Tipos de Datos De Pandas

- Series: Es una estructura de datos unidimensional similar a un array o una lista con etiquetas de datos, llamadas índices. Puedes pensar en ella como una columna en una hoja de cálculo.
- DataFrame: Es una estructura de datos bidimensional que se asemeja a una tabla de bases de datos o una hoja de cálculo con filas y columnas. Un DataFrame puede contener múltiples series y cada una de ellas puede tener un tipo de dato diferente.



Funciones para Exploración de Datos

Head() y tail(): Muestra las primeras (por defecto 5) o las últimas filas del DataFrame

shape: Devuelve una tupla con el número de filas y columnas del DataFrame.

info(): Proporciona información sobre el DataFrame, incluyendo el tipo de datos de cada columna y si hay valores nulos.

describe(): Proporciona estadísticas resumidas de las columnas numéricas, como la media, la mediana, el mínimo, el máximo y los percentiles.

value_counts(): Cuenta las ocurrencias únicas en una columna.

Funciones Para Exploración de Datos

isnull() y sum(): Verifica si hay valores nulos y cuenta cuántos hay por columna.

corr(): Calcula la matriz de correlación entre las columnas numéricas del DataFrame.