UNIVERSIDAD DE SONORA
Licenciatura en Física para la clase de Física Computacional
10 de Febrero de 2019
Grupo 3

Introducción a la biblioteca Pandas de Python

Castillo Espinoza Aarón Sebastián

I.. Introducción

Pandas es una librería de python destinada al análisis de datos, que proporciona unas estructuras de datos flexibles y que permiten trabajar con ellos de forma muy eficiente. Pandas ofrece las siguientes estructuras de datos:

- Series: Son arrays unidimensionales con indexación (arrays con índice o etiquetados), similar a los diccionarios. Pueden generarse a partir de diccionarios o de listas.
- DataFrame: Son estructuras de datos similares a las tablas de bases de datos relacionales como SQL.
- Panel, Panel4D y PanelND: Estas estructuras de datos permiten trabajar con más de dos dimensiones. Dado que es algo complejo y poco utilizado trabajar con arrays de más de dos dimensiones no trataremos los paneles en estos tutoriales de introducción a Pandas.

II.. Actividades desarrolladas

El propósito de esta actividad era que conociéramos mas a fondo la biblioteca de Pandas, para esto descargue una base de datos del Servicio Meteorológico Nacional sobre la ciudad de Bahía de Kino. En ella se encontraba una recopilación de las lluvias, temperatura máxima y temperatura mínima de cada día desde el año de 1974 hasta el 2011. Una vez teniendo estos datos proseguí a realizar el análisis de los datos con ayuda de las siguientes funciones con las que cuenta Pandas.

Para empezar guarde los datos en un DataFrame al cual denomine como ”df”:

- df.dtypes(): Este nos dice el tipo de datos con el que estamos trabajando, pues a pesar de que la tabla contiene datos numéricos, Pandas no los reconoce como tal en la mayoría de los casos, por eso es recomendable utilizar esta función al inicio de cada programa de análisis de datos y así evitarnos errores posteriores.
- pd.to_datetime: Una vez que sabemos con que tipo de variables es necesario convertirlas a el tipo de variables que necesitamos. ”pd.to_datetime.es una función de Pandas que se encarga de convertir los datos de nuestra columna ”FECHA.a una variable del tipo fecha para poder trabajar con ella.
- dt.month y dt.year: Una vez convertida la fecha a una variable de datetime utilizamos dt.month y dt.year para obtener el número de mes y año de cada renglón
- pd.to_numeric Ahora el resto de los datos los convertimos a datos numéricos con la función pd.to_numeric. Con este es necesario agregar un seguro pues algunos datos de las columnas pueden aparecer como ’Nulo’ y estos no los puede leer Pandas. El seguro que se aplica es ’coerce’ que convierte todos aquellos datos inválidos para su lectura en datos a datos del tipo ’NA’ que le dice a Pandas que ignore ese dato para no causar error.
- df.head() y df.tail(): Df.head() nos muestra los primeros n filas del DataFrame, mientras que df.tail() muestra las ultimas n filas del DataFrame. Por defecto muestra los primeros 5 pero bien pueden cambiarse por los que uno desee.
- df.mean(): Esta función del DataFrame calcula el promedio de los valores de la columna de datos especificada. Si no se especifica mostrará los de todas las columnas.
- df.std: Con esta función calculamos la desviación estándar de la columna especificada. Si no se especifica, mostrara el valor calculado de cada columna.

- df.median: Parecido a la promedio, esta función nos dará como resultado el valor de la mediana de la columna especificada o bien, de todas columnas si no se especifica alguna.
- df.max() y df.min(): Estas funciones del DataFrame nos entregan el valor máximo y mínimo de la columna especificada o en su defecto, de todas las columnas.
- df.describe(): Esta función lo que hace es una combinación de todas las anteriormente mencionadas, lo que hace se conoce como un análisis exploratorio de datos
- df.unique(): Con esta función cuenta la cantidad de datos únicos para la columna especificada, o de todas las columnas si no se especifica ninguna. En este caso fue utilizada para saber el numero total de años con los que cuenta el DataFrame.
- df.sum(): Se utiliza para realizar una suma de todos los datos que se encuentran en una columna especifica. Esta función me fue útil al momento de calcular las precipitaciones mensuales y acumuladas así como en el calculo de las temperaturas máximas y mínimas promedio por mes.

También fue necesario el uso de loops de la forma ”for-in-range”que me permitieron la optimización del código pues no fue necesaria la escritura de cada operación sino adecuar el rango correcto para que el loop se realizara las veces necesarias. En mi programa utilice 3 loops diferentes, uno con rango de 0 a 12 para calcular la precipitación mensual. Otro tuvo rango de 1974 a 2011 que son los años en los cuales se hizo la recolección de datos; este lo utilice para calcular las temperaturas mínimas y máximas promedio de cada mes. El ultimo loop que utilice tuvo el mismo rango que el anterior solo que este fue utilizado para calcular la precipitación promedio anual y también para saber la cantidad de datos que se tenían para cada año, pues no todos los años contaban con la misma cantidad de datos y debía calcular la precipitación promedio anual en relación con la cantidad de datos que se tenían por año.

III.. Reflexión

En esta actividad pude desarrollar de manera mas completa el conocimiento adquirido en la practica no. 2 pues esta vez la empecé desde 0, a diferencia de la anterior en la que el profesor nos proporciono un código en el cual basarnos. Esta practica se baso en la búsqueda de información y ejemplos en los cuales apoyarnos en la escritura de nuestro programa.

Practicamos nuestra habilidad al momento de hacer el análisis de datos, poder discernir entre la información que nos sera útil de toda la recabada y como trabajar con esta para dar respuesta a las preguntas planteadas. descubrí que el mes mas lluvioso en Bahía de Kino es el mes de agosto, seguido de septiembre y julio. Agosto también es el mes en, promedio, más caluroso medido desde 1974, así como enero es el mas frió.

A pesar de la falta de algunos datos de unos años o bien años completos en los que no se tiene ninguna medición logre hacer un análisis de temperaturas bajas y altas promedio por mes, así como de las lluvias o precipitaciones que se presentaron en ese periodo.

La falta de datos es una gran lastima y dificulta mucho el trabajo de todo aquel que desee hacer un análisis del área en cuestión pues no existe manera de recuperar todos esos datos perdidos. Aunque existen métodos para intentar estimarlos, estos nunca serán tan precisos como una buena medición.