Universidad de Sonora
División de Ciencias Exactas y Naturales
Departamento de Física
Física computacional
Python
Actividad 3
Iveth R. Navarro

29 de enero de 2021
Puerto Peñasco, Sonora, México

Resumen

La presente es una práctica cuyo objetivo es el dominio de la biblioteca *Pandas* para la manipulación y análisis de datos.

1. Introducción

El análisis de datos es la ciencia que se encarga de examinar un conjunto de datos con el propósito de sacar conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas. De ahí que sea tan importante para un físico el aprender del mismo. El análisis de datos puede ser optimizado haciendo uso de diferentes recursos, por ejemplo, Python. A su vez, en Python podemos gozar de la herramienta *Pandas*.

Pandas es una herramienta de manipulación de datos de alto nivel desarrollada por Wes McKinney. Es construido con el paquete Numpy y su estructura de datos clave es llamada el DataFrame. El DataFrame te permite almacenar y manipular datos tabulados en filas de observaciones y columnas de variables.

En esta actividad hacemos uso de dichas herramientas para el análisis de una lista de datos.

2. Desarrollo

Como Pandas es la librería objetivo de aprendizaje, se describirán brevemente algunos detalles sobre la misma.

- **Carácterísticas principales**
  - Define nuevas estructuras de datos basadas en los arrays de la librería NumPy pero con nuevas funcionalidades.
  - Permite leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL.
  - Permite acceder a los datos mediante índices o nombres para filas y columnas.
  - Ofrece métodos para reordenar, dividir y combinar conjuntos de datos.
  - Permite trabajar con series temporales.
  - Realiza todas estas operaciones de manera muy eficiente.
- **Tipos de datos**
  - Series: Estructura de una dimensión.
  - DataFrame: Estructura de dos dimensiones (tablas).
  - Panel: Estructura de tres dimensiones (cubos).

Estas estructuras se construyen a partir de arrays de la librería NumPy, añadiendo nuevas funcionalidades.

A continuación se describe la puesta en práctica de este conocimiento.

2.1. Actividad

La base de datos a analizar es la que se relaciona con la actividad 1: Información climatológica de, en mi caso, la estación 26072, Puerto Peñasco, Sonora. Lo primero que se realizó es un data frame habiendo importado la librerpia Pandas al programa. Las primeras compilaciones fueron solicitudes de los 10 primeros datos de la la base de datos y los 10 últimos al data frame creado. Analizando, lo más destacable es que en ambas solictudes se puede notar un número considerable de registros nulos en la evaporación. Al abrir la base de datos manualmente, pude constatar que efectivamente se tiene un registro nulo de evaporación hasta el año de 1986 (los registros comienzan desde 1952). Por otra parte, que en el 2015 haya tantos casos nulos seguidos, parece ser más un caso aislado.

Luego, se solicitó un análisis de forma y al compilar, se obtuvo que la base de datos tiene 17602 renglone sy 5 columnas. Al abrir la base, este dato se puede constatar.

Lo siguiente fue solicitar información al data frame. La observación aquí es que la compilación arrojó que la base de datos no contenía ningun dato nulo. Sin embargo, como ya comentamos anteriormente, sí que los hay y en gran cantidad en la columna de evaporación.

Lo siguiente fue crear un nuevo data frame para mantener el primero intacto. Luego, uno más para eliminar los datos nulos. Al compilar este último, pude verificar su funcionalidad, ya que, en comparación con los dos anteriores, no contuvo ningún dato nulo, dándo solución al problema anterior.

Luego, pasamos a cambiar el formato de los datos. Cuando los datos no sean números, podemos leer NaN en pantalla.

Seguido de esto, hacemos un conteo de datos faltantes.

Lo siguiente fue un análisis estadístico que arrojó los siguientes resultados.

	Precip	Evap	Tmax	Tmin
count	17568.000000	9008.000000	16608.000000	16648.000000
mean	0.229559	6.810391	28.885013	15.586893
std	2.699255	2.859273	6.992284	7.390738
min	0.000000	0.000000	8.000000	-8.000000
25%	0.000000	4.600000	23.000000	10.000000
50%	0.000000	6.900000	29.200000	14.400000
75%	0.000000	8.800000	34.800000	21.500000
max	162.000000	18.000000	44.500000	32.200000

Lo siguiente fue indicarle a Python que reconozca el formato de las fechas. Para ser más detallistas, también se agregaron columnas para distingiuir meses y años. También se corroboró que los meses y años estuvieran fueran números enteros.

3. Conclusión

En lo personal, me pareció muy interesante la simplicidad del código al solicitar un análisis estadístico. Me dejó muy sorprendida y me hizo pensar en las posibilidades. A su vez, me sorprendió también que el darle formato a la fecha resultara más complejo que un análisis estadístico y creo que fue la parte que más se me dificultó.

En general, me pareció una práctica muy entretenida y que disfruté llevar a cabo y la carga me pareció adecuada para la semana. No hubo alguna cosa que me aburriera y no creo que haya algo que mejoraría. Aunque Python sigue siendo nuevo para mí, le asigno un nivel bajo de complejidad.

4. Bibliografía

- learnpython.org. (s. f.). Pandas Basics - Learn Python - Free Interactive Python Tutorial. learnpython. Recuperado 28 de enero de 2021, de `https://www.learnpython.org/es/Pandas%20Basics`
- QuestionPro. (s. f.). Análisis de Datos. Recuperado 28 de enero de 2021, de `https://www.questionpro.com/es/analisis-de-datos.html`
- Alberca, A. S. (2020, 4 octubre). La librería Pandas. Aprende con Alf. `https://aprendeconalf.es/docencia/python/manual/pandas/`