Módulo 4: Python para el análisis de datos

En este módulo vamos hablar de Python como herramienta para el análisis de datos. Python es un lenguaje de programación de código abierto, es decir, no necesita licencia para que lo podamos utilizar. Python fue creado a finales de los ochenta por Guido van Rossum y la primera versión salió en 1991. Van Rossum buscó en la creación de Python un lenguaje que fuera intuitivo, entendible y de código abierto, para que, tantos programadores avanzados en otros lenguajes y no programadores, pudieran adaptarse rápidamente a su sintaxis y desarrollar programas de forma rápida y fácil. Como se ha definido en módulos anteriores, la ciencia de datos se encarga de analizar, transformar datos y extraer información de utilidad para la toma de decisiones. Gracias a Python estas tareas se pueden llevar a cabo sin la necesidad de tener conocimientos avanzados de programación, con pocas líneas de código y en entornos de programación amigables que facilitan la programación y visualización de resultados.

Viene con poderosas bibliotecas estadísticas y numéricas como: Pandas, Numpy, Matplotlib, SciPy, scikit-learn, como se describen en la tabla 1.

Nombre	**Funcionalidad principal**
Matplotlib	Visualización de datos.
Pandas	Manipulación de conjuntos de datos.
NumPy	Da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas.
Scikit-learn	Procesamiento de datos y algoritmos de aprendizaje automático
Scipy	Se compone de herramientas y algoritmos matemáticos.

Para comenzar hay que instalar Python, a continuación se muestran los pasos para su instalación en Windows 10, si tiene otro sistema operativo en este enlace están los descargables y los pasos requeridos: https://www.python.org/

Instalación de Python:

1. Ir al enlace https://www.python.org/ y descargar la última versión para Windows 10, como se observa en la ilustración 1.

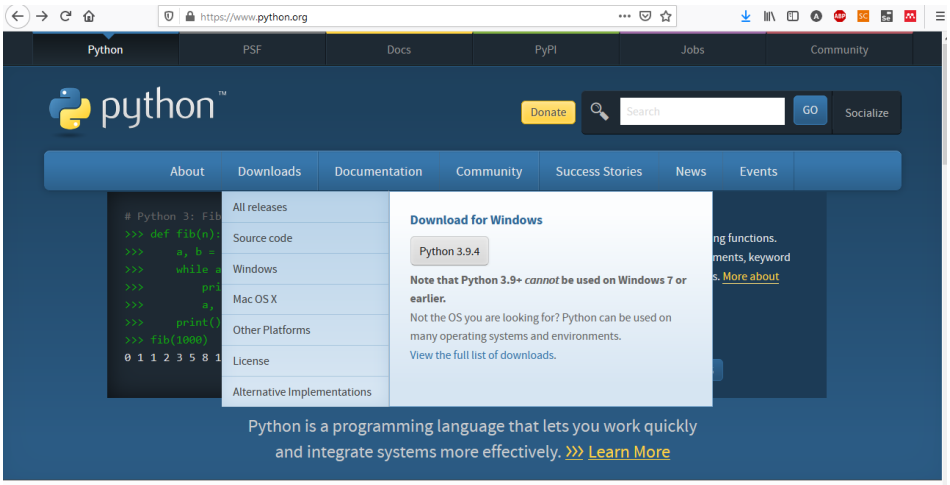


Ilustración 1. Paso 1 de la instalación de Python.

2. Cuando descargue el archivo ejecutable (termina en .exe), se debe dar clic derecho y abrir con propiedades de administrado como se ve en la ilustración 2.

Ilustración 2. Paso 2 de la instalación de Python.

3. Al dar clic en “Ejecutar como administrador” aparece un cuadro como se observa en la ilustración 3, se debe seleccionar la opción “Add Python to PATH” y dar clic en “Install Now”. Luego se debe dar siguiente y el programa comenzará a instalarse y finalizará con una pantalla similar a la ilustración 4.

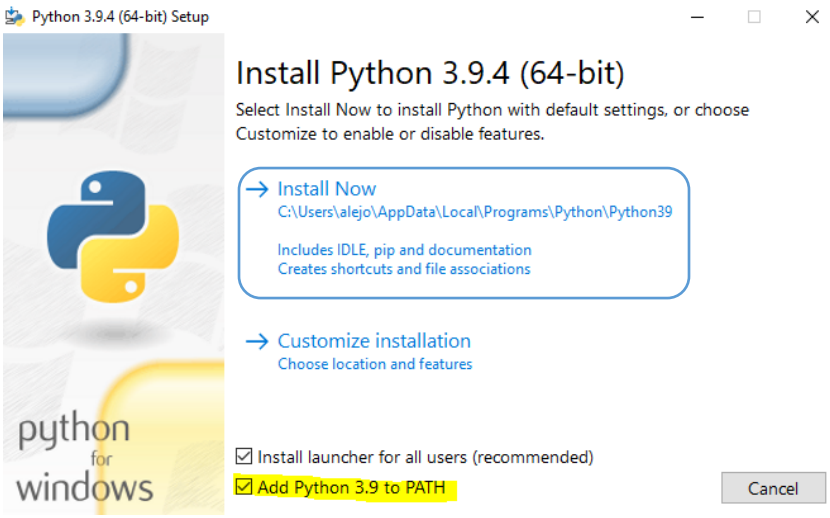


Ilustración 3. Paso 3 de la instalación de Python.

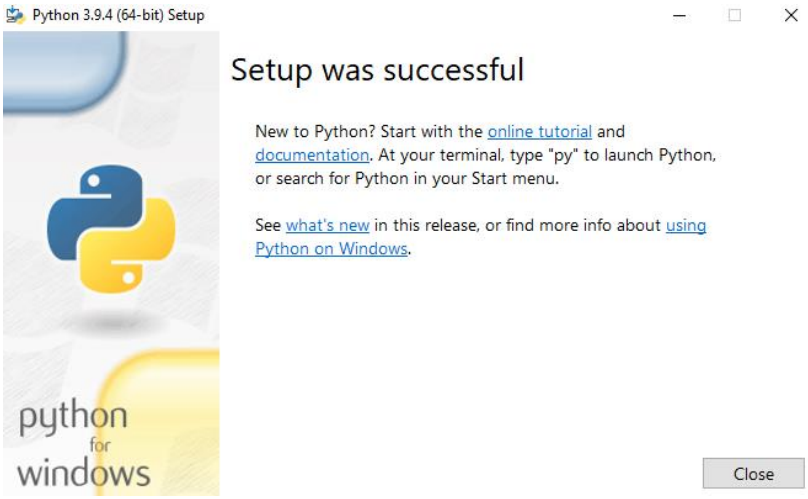


Ilustración 4. Mensaje final de instalación de Python.

4. Ahora para verificar Python quedó instalado en el equipo se debe abrir una terminal (cmd) y escribir el comando “python --version” y deben obtener una respuesta del terminal con la versión de Python instalada en su computador, similar a como se observa en la ilustración 5.

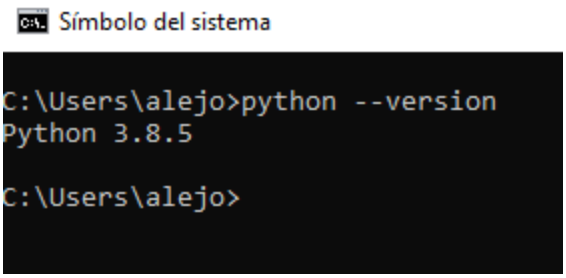


Ilustración 5. Versión de Python en el cmd.

También es necesario instalar una distribución gratuita de Python llamada anaconda que cuenta con todas las librerías que se van a emplear en este módulo, se describen los pasos a continuación:

1. Ingresar a https://www.anaconda.com/products/individual#Downloads y descargar la versión para Windows 10, como se observa en la ilustración 6 (si tiene otro sistema operativo deben descargar el archivo correcto).

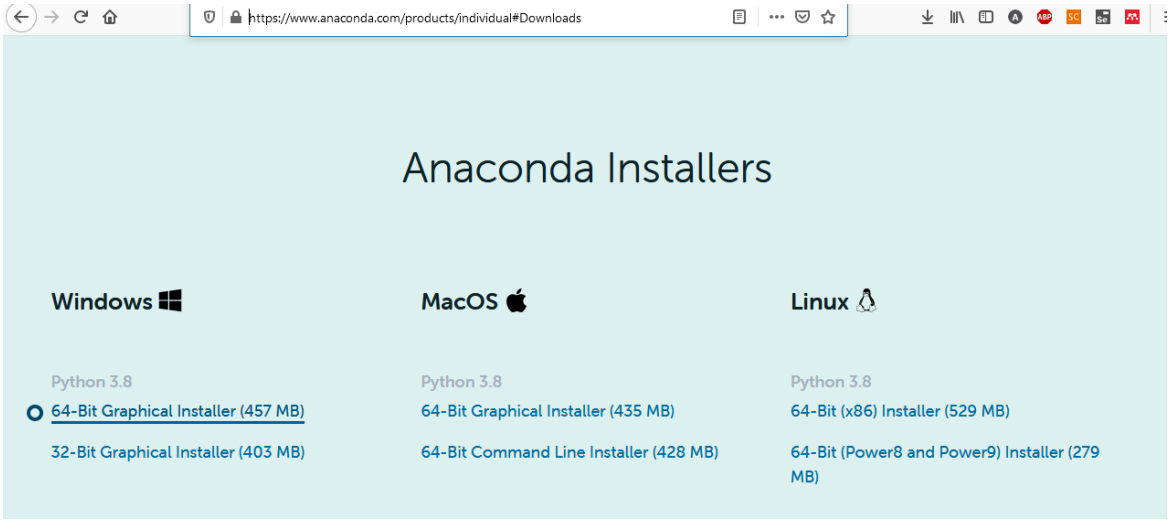


Ilustración 6. Descarga de Ananconda.

2. Se debe ejecutar el instalador como administrador, como se observa en la ilustración 7.

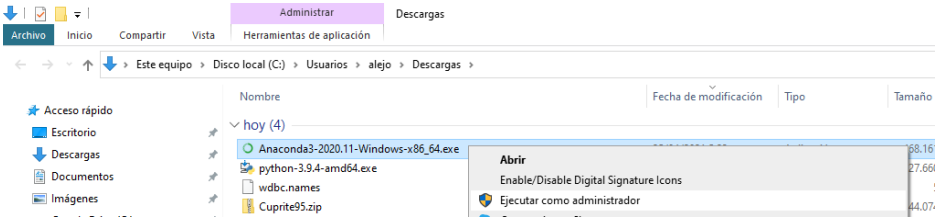
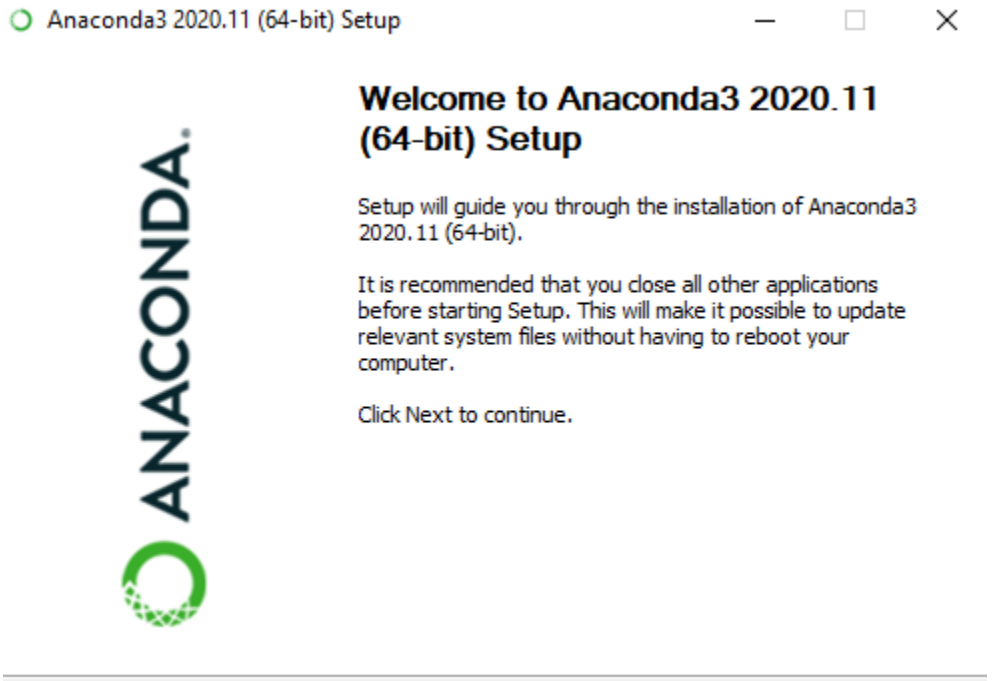
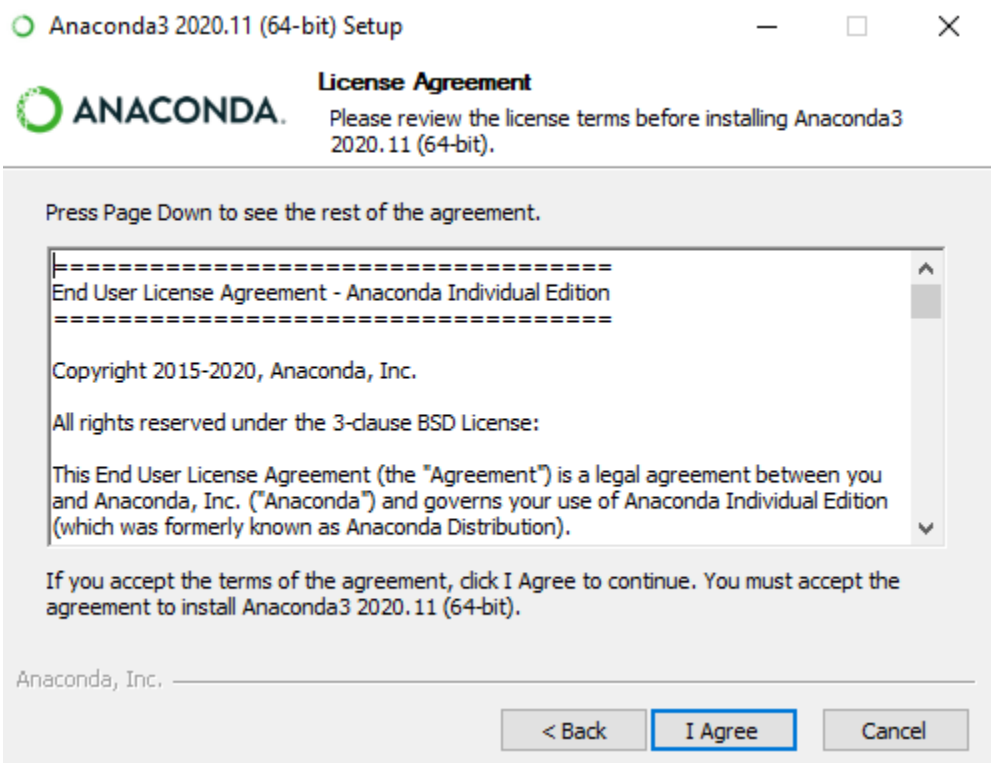
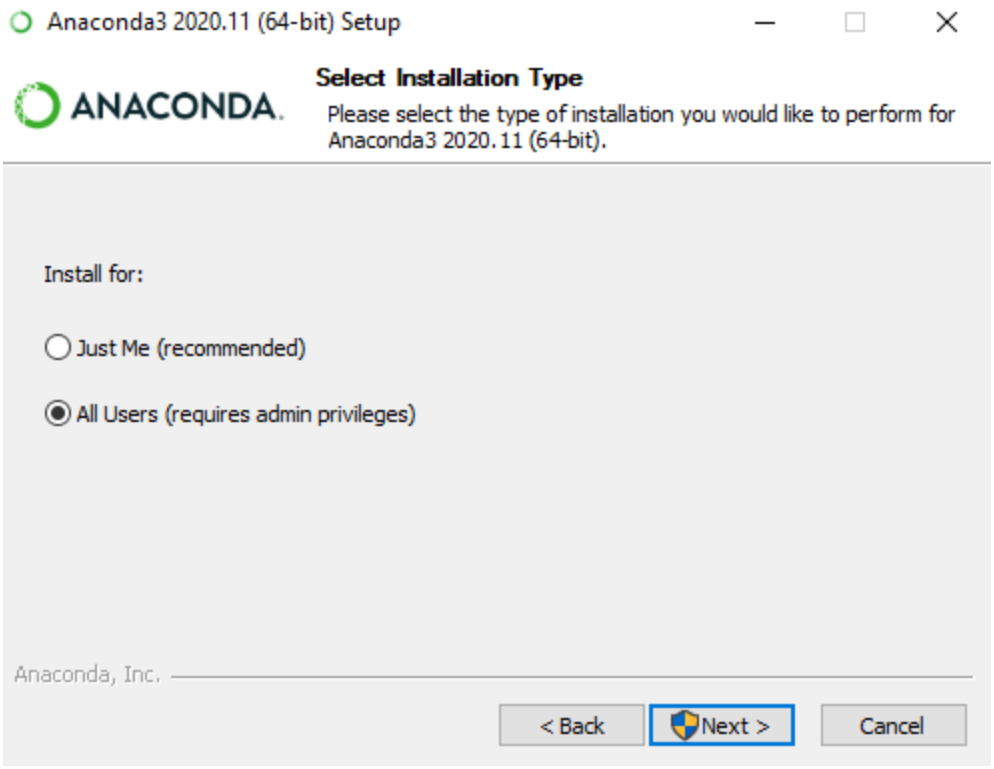


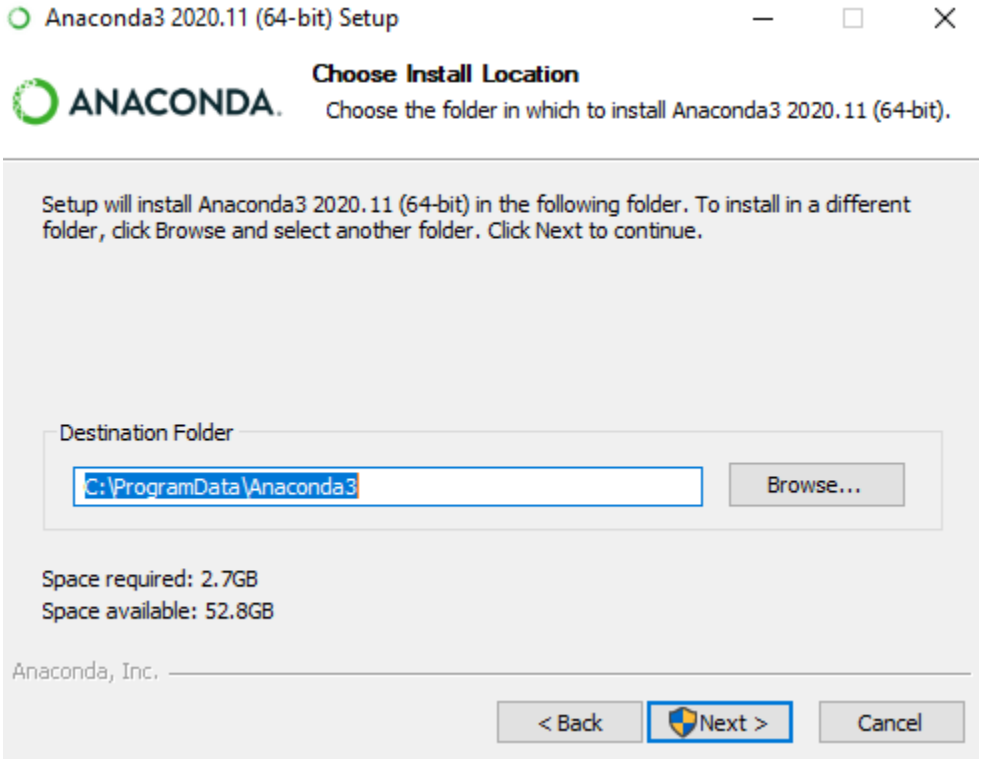
Ilustración 7. Ejecutar Anaconda como administrador.

3. Luego va aparecer el asistente de instalación y se deben seleccionar las opciones que aparecen en las ilustraciones siguientes:









4. Ahora se debe ejecutar la aplicación Jupyter que viene instalada en Anaconda, esta aplicación permite crear y ejecutar código en Python y guardarlo en unos archivos portables que finalizan con la extensión .ipynb. Para abrir Jupyter debemos ir a inicio y buscar “Anaconda Prompt”, como se observa en la ilustración 8.

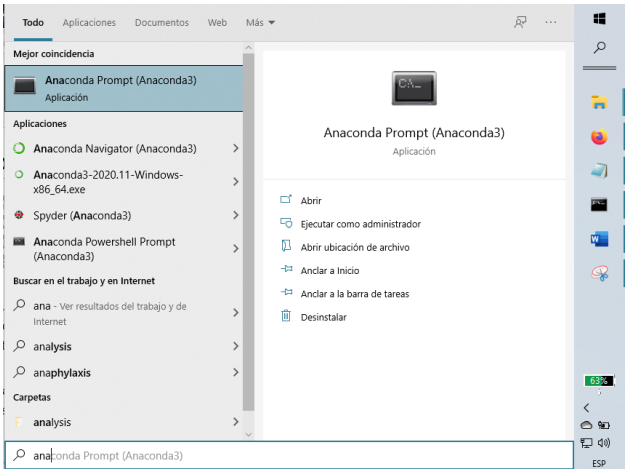


Ilustración 8. Anaconda Prompt.

5. Luego de abrir Ananconda Prompt, aparecerá una terminal, se debe escribir el siguiente comando “Jupyter Notebook” y dar enter, como se observa en la ilustración 9.

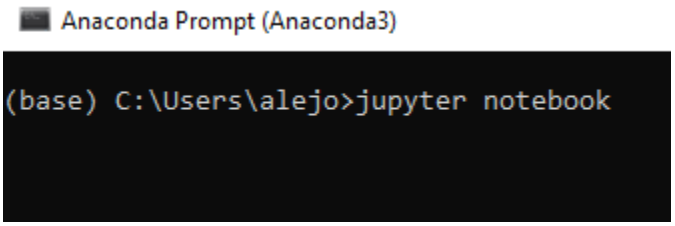


Ilustración 9. Jupyter Notebook.

6. Luego se abrirá una ventana similar a la que se muestra en la ilustración 10, acá se puede buscar en las carpetas de nuestro computador algún archivo con extensión .ipynb como se muestra en los vídeos de este módulo.

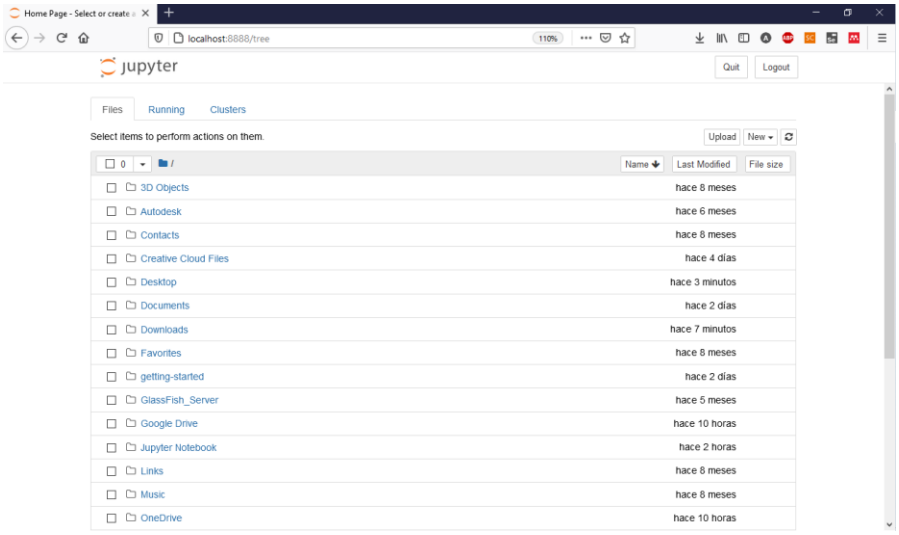


Ilustración 10. Interfaz Jupyter Notebook.

Ahora vamos hablar de hablar de la librería pandas, que permite cargar datos que están almacenados en un archivo .csv. Los archivos CSV (del inglés comma-separated values) son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma). El uso de esta librería y otras librerías, los comandos básicos para estadística descriptiva y las instrucciones para realizar las gráficas básicas con Python se encuentran explicados con detalle en los vídeos de este módulo.

Referencias:

Página oficial de Python: https://www.python.org/

Página oficial de Anaconda: https://www.anaconda.com/