

Sesión

1

Aprendizajes esperados

Al final de esta sesión verifica que puedas:



Reconocer varias estructuras de datos y archivos, incluyendo hojas de cálculo, archivos CSV y *Dataframes*.

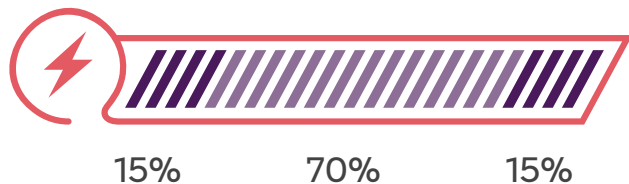


Leer, depurar y escribir código utilizando el paquete *Pandas* de *Python* para leer datos de un archivo CSV a un *Dataframe* de *Pandas*.

Material para la clase

- Anexos 1.1 y 1.2 según el editor de *Python* que utilicen en clase

Duración sugerida



Lo que sabemos, lo que debemos saber



Esta sección corresponde al 15% de avance de la sesión

Inicia la sesión participando en una lluvia de ideas sobre *Python*.



¿Qué recuerdas acerca de su uso?
¿Puedes mencionar algunas funciones y características?

Como ya sabes, *Python* es un lenguaje de programación versátil y muy popular en diferentes áreas. En esta guía aprenderás acerca de su uso en el análisis de datos. *Python* es uno de los lenguajes más utilizados en el mundo del análisis de datos por su simplicidad y su amplio ecosistema de herramientas. A lo largo de esta guía, descubrirás cómo leer archivos y trabajar con datos de manera estructurada.

En grados anteriores, ya has utilizado *Python* para automatizar tareas sencillas, programar juegos interactivos y manipular variables como cadenas de texto y listas de una y dos dimensiones. En esta sesión aprenderás a utilizar un nuevo tipo de variable, pero antes, observa las siguientes imágenes y responde:



¿Qué tienen en común las diferentes representaciones?
¿En qué se diferencian?
¿Qué ventajas y desventajas tiene cada tipo de representación?

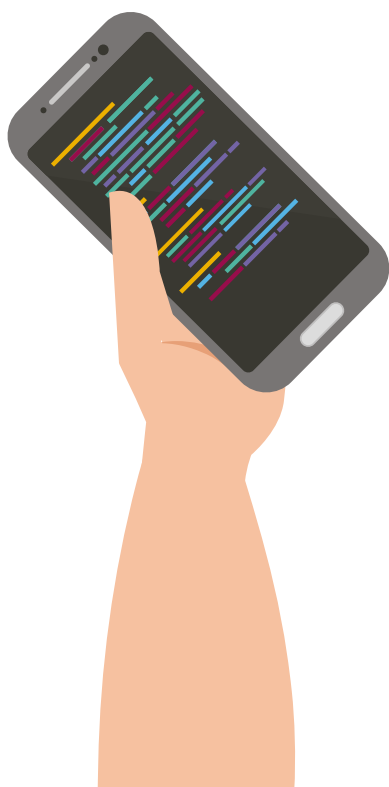


Figura 1. Archivo CSV

```
Rango de edad,Sí sabe,No sabe,No informa,No informa.1
De 0 a 9 años*,40.6,10.31,1.43,
De 10 a 19 años*,97.3,1.37 1.33,
De 20 a 29 años*,96.57 1.71,1.72,
De 30 a 39 años*,95.96,2.69,1.35,
```

Figura 2. Datos en un formato de tabla

A	B	C	D	E
Rango de edad	Sí sabe	No sabe	No informa	No informa.1
De 0 a 9 años*	40.6	10.31	1.43	
De 10 a 19 años	97.3	1.37	1.33	
De 20 a 29 años	96.57	1.71	1.72	
De 30 a 39 años	95.96	2.69	1.35	
De 40 a 49 años	94.5	4.24	1.26	
De 50 a 59 años	92.47	6.28	1.25	
De 60 a 69 años	88.71	10.34	0.96	

Figura 3. Archivo CSV

```
Nombre,Indicador,Año,Valor
Ingresos Altos,Realizó un pago digital,2000,
Ingresos Altos,"Realizó un pago digital, mujer", 2000,
Ingresos Altos,"Realizó un pago digital, hombre",2000,
Ingresos Altos,% mujeres empleadas,2000,88.4271771152308
Ingresos Altos,% hombres empleados,2000,83.8033567952988
Ingresos Altos,% de empleados total,2000,85.8145356958302
Ingresos bajos y medios ,Realizo un pago digital,2000,
```

Figura 4. Datos en un formato de tabla

Nombre	Indicador	Año	Valor
Ingresos Altos	Realizó un pago digital	2000	
Ingresos Altos	Realizó un pago digital, mujer	2000	
Ingresos Altos	Realizó un pago digital, hombre	2000	
Ingresos Altos	% mujeres empleadas	2000	88.42717712
Ingresos Altos	% hombres empleados	2000	83.8033568
Ingresos Altos	% de empleados total	2000	85.8145357
Ingresos bajos y medios	Realizó un pago digital	2000	
Ingresos bajos y medios	Realizó un pago digital, mujer	2000	
Ingresos bajos y medios	Realizó un pago digital, hombre	2000	

Como puedes notar, las Figuras 1 y 2 representan el mismo conjunto de datos. Asimismo, las Figuras 3 y 4 son dos representaciones distintas de otro conjunto de datos.

En este caso, las Figuras 1 y 3 muestran los datos en su formato **separado por comas** (o CSV por las siglas en inglés de “comma separated values”). Este formato es muy usado para guardar grandes archivos de datos porque ocupa poco espacio en memoria.

En cambio, las Figuras 2 y 4 representan los datos en un formato de tabla. Las tablas, por su parte, permiten visualizar los datos de una forma clara y ordenada.

En esta sesión, usarás un paquete de Python muy popular llamado **pandas**, el cual es esencial para leer, procesar y analizar grandes volúmenes de datos. El paquete **pandas** permite leer archivos y guardarlos en un nuevo tipo de estructura de datos llamada **dataframe**.

Un **dataframe** se representa como una tabla que organiza la información en filas y columnas, similar a una hoja de cálculo de Excel. Cada columna puede contener un tipo específico de dato, como números, texto o fechas y cada fila representa un registro. Con un dataframe, puedes realizar operaciones como filtrar, ordenar, sumar, promediar y agrupar datos de forma eficiente. Se presenta en la Figura 5.



Nombre	Indicador
Ingresos Altos	Realizó un pago
Ingresos Altos	Realizó un pago d
Ingresos Altos	Realizó un pago d
Ingresos Altos	% mujeres emple
Ingresos Altos	% hombres empl
Ingresos Altos	% de empleados
Ingresos bajos y medios	Realizó un pago
Ingresos bajos y medios	Realizó un pay
Ingresos bajos y medios	Realizó un p





Figura 5. Dataframe

The diagram shows a table representing a DataFrame. The columns are labeled 'Nombre', 'Edad', 'Grado', and 'Correo'. The rows are labeled '1', '2', '3', and '4'. Arrows point from the labels to the corresponding parts of the table. The label 'Nombres Columnas' points to the column headers. The label 'Nombres Filas' points to the row indices. The label 'Columnas' points to the column headers. The label 'Filas' points to the row indices.

	Nombre	Edad	Grado	Correo
1	María	18	Economía	maria@gmail.com
2	Luis	22	Medicina	luis@yahoo.com
3	Carmen	20	Arquitectura	carmen@gmail.com
4	Antonio	21	Economía	antonio@gmail.com

En la siguiente actividad leerás tu primer archivo csv en *Python*.

Glosario

-  **Archivo CSV:** archivo de texto donde los datos están separados por comas, ideal para almacenar tablas de datos.
-  **Dataframe:** estructura de datos en forma de tabla con filas y columnas, utilizada para organizar y analizar información.
-  **Paquete:** conjunto de herramientas predefinidas en *Python* que facilitan tareas específicas.
-  **Pandas:** paquete de *Python* para manejar, procesar y analizar datos de manera eficiente.

Anexos

Anexo 1.1

Para el desarrollo de esta guía es necesario que descargues Anaconda por lo que debes seguir las siguientes instrucciones:

1. Ingresa en tu navegador al siguiente enlace: <https://www.anaconda.com/download/#windows>
2. Elige tu sistema operativo (las siguientes pasos serán para la instalación en Windows).



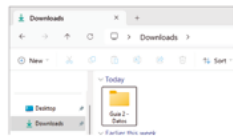
¿Por qué es importante que las personas entiendan cómo funciona el aprendizaje automático?



Anexo 1.2

Para leer archivos en Python, es necesario que tanto el código como el archivo de datos se encuentren en la misma carpeta. A continuación te algunos pasos para seguirlos, pero siempre debes seguir las indicaciones de tu docente.

1. Ingresa a la carpeta "Descargas" (Downloads) de tu computador.
2. Crea una nueva carpeta llamada "Guía 2 - Datos".



Mueve los archivos que descargaste a esta carpeta. Se debe ver así:



Enlace



Archivos de esta sesión:
Leer datos.ipynb / leer datos.py pruebas_saber.csv

Manos a la obra

Conectadas



Esta sección corresponde al 85% de avance de la sesión

En esta actividad, aprenderás algunos comandos básicos para leer e interactuar con datos en Python. Trabajarás en parejas, utilizando un editor de Python como Anaconda, Google Colab o Python-Online, dependiendo de lo que tu docente te indique. Si tienes dudas sobre cómo acceder a alguno de estos editores, consulta el Anexo 1.1, que contiene tutoriales detallados para cada opción.

Para empezar, descarguen los archivos con los que van a trabajar durante la primera parte de la actividad. Siguen el enlace o utilicen el QR y guarden el archivo en el escritorio del computador. Recuerden que los archivos .ipynb se refieren a cuadernos interactivos que pueden utilizar en Anaconda o Google Colab, y los archivos .py se refieren a código de Python que pueden utilizar en editores como Python-Online.

Abran el archivo pruebas_saber.csv haciendo doble clic.



¿Qué observan?
¿Qué esperan ver en Python cuando lo utilicen?

A continuación, tu docente entregará una copia del Anexo 1.2 que corresponda con el editor que están utilizando en el salón. Siguen las indicaciones para cargar el archivo de datos y ejecutar los primeros comandos.

Cuando terminen, es el momento de buscar y explorar datos que les interesen. Para esto, van a utilizar la página de Datos Abiertos del gobierno de Colombia. Recuerden que los Datos Abiertos son información pública que se puede usar, reutilizar y redistribuir libremente.

Sigan los pasos para encontrar y descargar un archivo de datos que te interese a ti y a tu compañero o compañera.

- 1 Ingresen a datos.gov.co y den clic en “Descubre” como se presenta en la *Figura 6*.

Figura 6. Sitio web de datos.gov.co



- 2 Seleccionen el filtro “Conjuntos de datos” como se presenta en la *Figura 7*.

Figura 7. Conjunto de datos



- 3 Seleccionen un tema de su interés en la sección “Clasificación” como en la lista que se presenta en la *Figura 8*.

Figura 8. Selección temática de interés



- 4 Exploren los diferentes conjuntos de datos y hagan clic en uno que capte su atención.
- 5 Den clic en la pestaña “Datos” y observa los resultados como se presenta en la *Figura 9*.

Figura 9. Resultados y presentaciones de datos

ID	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...	ESTU...
0183	DK2018303	CC	COLOMBIA	11	BOGOTÁ	11001	BOGOTÁ D.		11	2834	UNIVERSID. INSTITUCI	ADMINISTR	BOGOTÁ	11001	BOGOTÁ D.	HOTELERIA	BOGOTÁ			
0183	DK2018303	CC	COLOMBIA	76	VALLE	76726	SEYLLA		76	1233	UNIVERSID. UNIVERSID.	CONTADUR	VALLE	76001	CAJÍ	CONTADUR	VALLE			
0183	DK2018301	CC	COLOMBIA	73	TOLIMA	73001	IBAGUÉ	170001934	73001	2829	CORPORAC. INSTITUCI	CONTADUR	BOGOTÁ	11001	BOGOTÁ D.	CONTADUR	BOGOTÁ			
0183	DK2018301	CC	COLOMBIA	68	SANTANDÉ	68001	BARRANCA	160001830	68001	68	2207	INSTITUTO INSTITUCI	INGENIERÍ	SANTANDÉ	68001	BARRANCA	INGENIERÍ	SANTANDÉ		
0183	DK2018301	CC	COLOMBIA	11	BOGOTÁ	11001	BOGOTÁ D.	111001901	11001	11	1338	UNIVERSID. UNIVERSID.	BIBLIOTEC	QUINDÍO	63001	ARMENIA	CIENCIAS E	QUINDÍO		
0183	DK2018301	CC	COLOMBIA	70	SUCRE	70001	ENRIQUELÓ	170001930	70001	70	2829	CORPORAC. INSTITUCI	PSICOLOGÍ	SUCRE	70001	SINCELEJO	PSICOLOGÍ	SUCRE		
0183	DK2018301	CC	COLOMBIA	11	BOGOTÁ	11001	BOGOTÁ D.	311001931	11001	11	1701	PONTIFICIA UNIVERSID.	ADMINISTR	BOGOTÁ	11001	BOGOTÁ D.	ADMINISTR	BOGOTÁ		
0183	DK2018301	CC	COLOMBIA	76	VALLE	76520	PALMIRA		76	2192	UNIVERSID. UNIVERSID.	PSICOLOGÍ	BOGOTÁ	11001	BOGOTÁ D.	PSICOLOGÍ	BOGOTÁ			
0183	DK2018303	CC	COLOMBIA	11	BOGOTÁ	11001	BOGOTÁ D.		11	2723	POLITECNIC. INSTITUCI	ADMINISTR	BOGOTÁ	11001	BOGOTÁ D.	NECOCOCES	BOGOTÁ			
0183	DK2018303	CC	COLOMBIA	76	VALLE	76001	CAJÍ	376001931	76001	76	2749	UNIVERSID. INSTITUCI	ADMINISTR	BOGOTÁ	11001	BOGOTÁ D.	ADMINISTR	VALLE		
0183	DK2018303	CC	COLOMBIA	05	ANTIOQUIA	05000	IBAGUÉ	933000000	05000	05	2782	PUNTADEE INSTITUCI	INGENIERÍ	ANTIOQUIA	05000	SANTA ROSA	INGENIERÍ	ANTIOQUIA		

- 6 Exporten los datos como archivo CSV tal y como se visualiza en la *Figura 10*.

Figura 10. Exportar datos como CSV

Ahora regresen al código, carguen y exploren los datos en *Python*.

Prepárate para compartir tus hallazgos con el resto de la clase al terminar la sesión.

Antes de irnos



Esta sección corresponde al 100% de avance de la sesión

Revisa los aprendizajes de la sesión de forma individual respondiendo las preguntas de forma que mejor reflejen tu progreso:

- 1 ¿Puedes reconocer varias estructuras de datos y archivos, incluyendo hojas de cálculo, archivos CSV y dataframes?
 - Sí
 - Parcialmente
 - Aún no
- 2 ¿Puedes leer, depurar y escribir código utilizando la biblioteca Pandas de *Python* para leer datos de un archivo CSV a un *dataframe* de Pandas?
 - Sí
 - Parcialmente
 - Aún no

Si tus respuestas fueron “Parcialmente” o “Aún no”, vuelve a leer los contenidos. Resalta o subraya los términos que no hayas comprendido. Luego, discute con tus compañeras y compañeros de grupo lo que se hizo en cada momento de la actividad y el rol al que correspondía. Si todavía te quedan dudas, consúltale a tu docente.



CSV



XLS

Para cerrar la sesión, completa el siguiente ticket de salida:

Fecha:

Piensa en el conjunto de datos que descargaste y exploraste en *Python*. ¿Qué preguntas podrías responder con estos datos?

¿Cómo relacionas lo aprendido en esta clase con tus conocimientos previos de *Python*?

