

A photograph showing several people from behind, sitting at a long table in an office or study room. They are all wearing headphones and looking at their laptops. One laptop screen is clearly visible, displaying what appears to be a video editing interface. The atmosphere is focused and collaborative.

# Repaso final de módulo

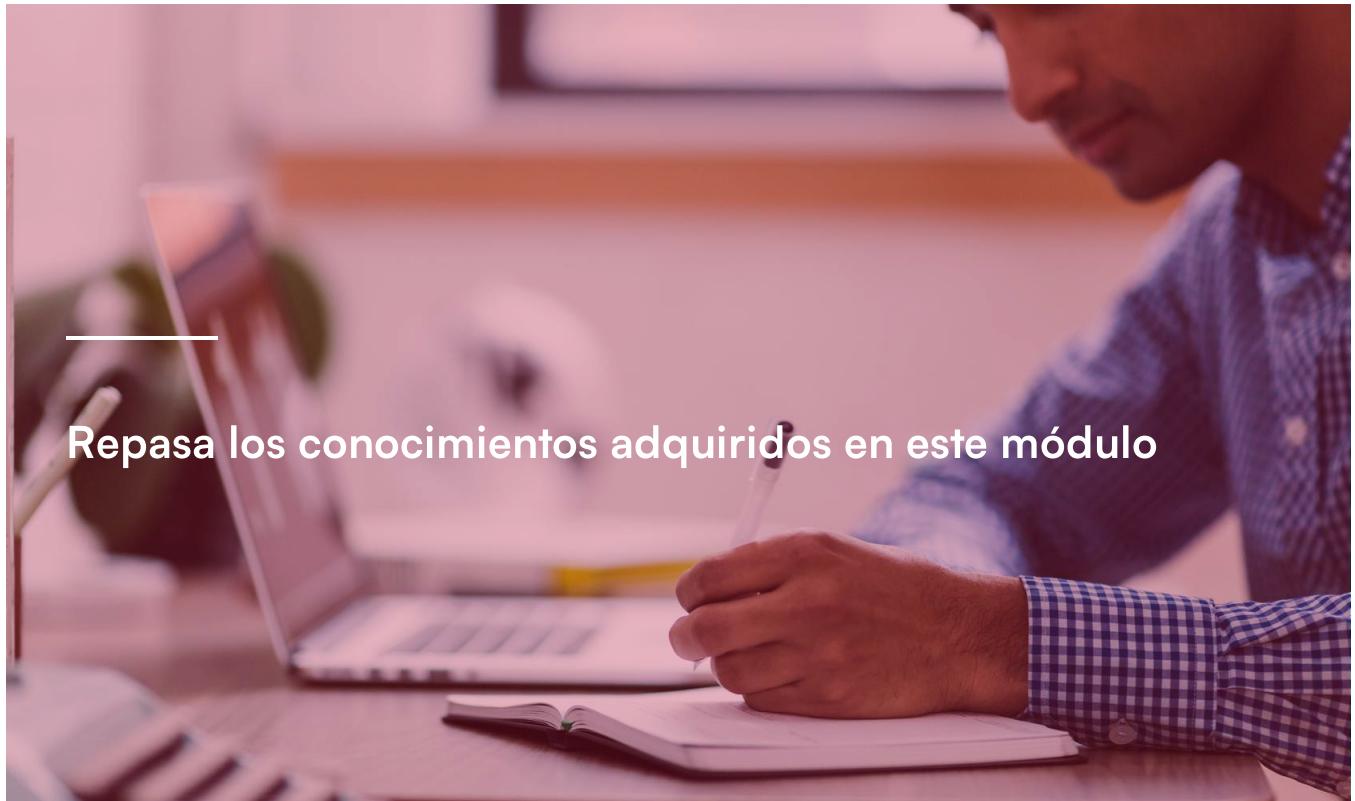


≡ I. Resumen de módulo

≡ II. Caso práctico de repaso con solución

## I. Resumen de módulo

---



**Repasa los conocimientos adquiridos en este módulo**

En este módulo, se han revisado las tecnologías básicas que todo aquel que quiera trabajar en el ámbito tecnológico del *big data* y del análisis de datos necesita conocer para poder entender las técnicas y herramientas más avanzadas en la materia.

Se han cubierto cinco áreas de conocimiento a través de seis unidades de trabajo. La **primera unidad** se ha dedicado al tema de la virtualización de programas y al uso de los comandos de la *shell* de Linux. En muchas ocasiones, el software que se utiliza en este entorno se distribuye ya instalado en una máquina virtual, con el objetivo de evitar que el usuario tenga que dedicar tiempo y esfuerzo a instalarlo. Esto se debe a que la configuración y puesta a punto de este tipo de software suele ser una tarea compleja. Asimismo, en muchos casos, el software que se usa se ejecuta bajo Linux, de manera que para utilizarlo o gestionarlo debe hacerse desde la *shell* de comandos del sistema operativo. Por ello se ha estudiado la *shell* y sus principales comandos y funcionalidades, de manera que el usuario esté familiarizado con dicho entorno.

La **segunda área** de conocimiento que se ha revisado en dos de las unidades ha sido el lenguaje de programación Python y sus librerías para cálculos científicos. Actualmente, Python y el lenguaje R son los lenguajes de programación más utilizados en el ámbito del *big data*. En una de las unidades del módulo se ha descrito el lenguaje de programación Python, estudiando sus principales estructuras sintácticas como lenguaje de propósito general. A continuación, en una unidad independiente, se ha examinado el conjunto de librerías de cálculo científico que dotan a Python de la potencia necesaria para realizar análisis de datos, es decir, las librerías NumPy –para gestión de matrices y cálculos estadísticos–, la librería Matplotlib –para la representación de datos– y la librería Pandas –para facilitar la manipulación de datos–. La combinación de las tres librerías convierte a Python en una poderosa herramienta para llevar a cabo análisis de datos.

La **tercera área** de conocimiento está dedicada a las bases de datos relacionales y al lenguaje SQL. En esta unidad se han revisado los principios conceptuales de las bases de datos relacionales a través del estudio del modelo relacional. A continuación, se ha realizado una introducción al lenguaje de consultas SQL. Este lenguaje permite manipular los datos de una base de datos relacional: creación de tablas, inserción de filas o realización de consultas para recuperar datos. En el ámbito del *big data*, han surgido unas nuevas bases de datos denominadas bases de datos NoSQL, las cuales se caracterizan por seguir principios diferentes a los de las bases de datos relacionales. Sin embargo, para comprenderlas, es necesario entender previamente las bases de datos relacionales. Además, algunas de ellas mantienen ciertas características del mundo relacional e, incluso, determinados lenguajes de consultas son similares a SQL.

La **cuarta área** de conocimiento se refiere a los formatos de almacenamiento de datos. Una tarea básica de cualquier analista de datos es recuperar los datos de diferentes fuentes de información. Esta información, normalmente, se recupera codificada en un formato de datos determinado. En esta unidad se han revisado los principales formatos de datos que se utilizan para codificar la información en internet, CSV, XML y JSON. También se ha descrito cada uno de los formatos: su sintaxis y semántica. Asimismo, se ha realizado una introducción acerca de cómo manipular estos formatos desde Python.

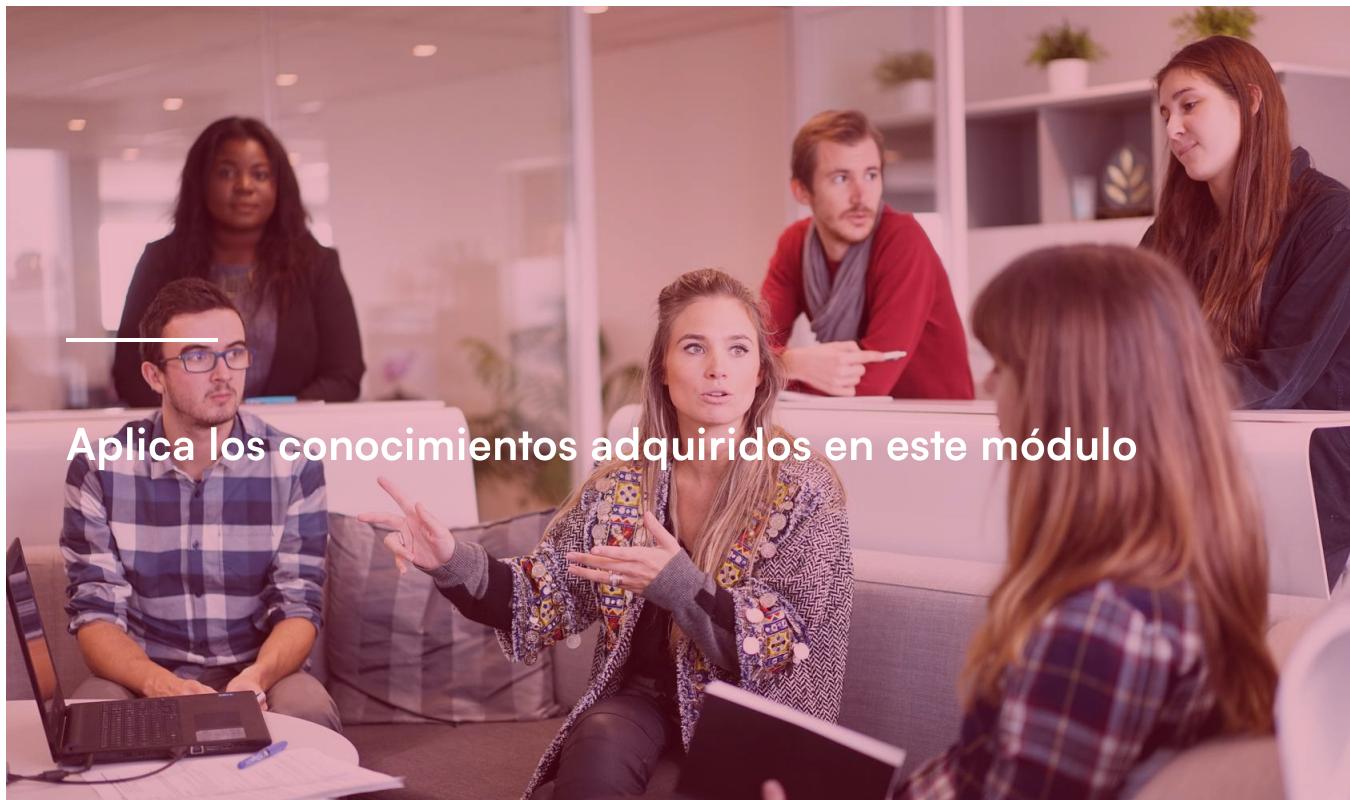
Por último, la **quinta área** de conocimiento que se ha tratado en este módulo tiene como objeto los repositorios digitales de información. Actualmente, la gestión de muchos proyectos informáticos del ámbito del *big data*, y también de otros ámbitos, se desarrolla en repositorios digitales que permiten trabajar de manera colaborativa, como GitHub o Google Drive. Este tipo de entornos ofrece al programador la posibilidad de gestionar las versiones de los programas, la documentación, así como la

organización de un equipo de desarrollo. En muchos casos, este tipo de proyectos requiere la participación de un gran número de personas, por lo que en entornos de trabajo como los citados se usan frecuentemente. Asimismo, este tipo de herramientas se utilizan habitualmente como sistemas de distribución del software o de la documentación que se genera en un proyecto de estas características.

Además, cada unidad del módulo se complementa con un test de preguntas que sirve de repaso a todo lo estudiado y con la propuesta de resolución de un caso práctico.

## II. Caso práctico de repaso con solución

---



**Aplica los conocimientos adquiridos en este módulo**

### ENUNCIADO

Se va a plantear un ejercicio práctico de análisis de datos real a partir de los datos contenidos en un fichero CSV. Para ello, hay que descargar el siguiente archivo “electronic-card-transactions.csv”\*

Este archivo contiene una muestra de datos acerca de transacciones de pago electrónico, agrupadas por comercio y por mes, durante un periodo de 15 años. Cada fila contiene el importe en dólares gastado

(columna 'Data\_value'), el mes y año, y el sector (columna 'Series\_title\_2').

## DATOS

Se pide ejecutar las siguientes tareas sobre dicho documento, en un notebook de Jupyter:

- Crear una función que lea el contenido del fichero y genere una lista de diccionarios, donde cada diccionario contenga los datos de una fila.
- Desarrollar una función *aggregate\_year*, que calcule la suma de los importes, agrupados por año y por sector. Esta función debe retornar un diccionario cuyas claves serán los sectores, y para cada una de ellas su valor será otro diccionario donde cada clave será el año, y su valor la suma de las transacciones durante ese año.
- Construir un *dataframe* a partir de la estructura retornada por *aggregate\_year*, donde los sectores estarán en el eje X y los años en el eje Y.
- Calcular la suma y la media de las transacciones, agrupadas por sector y año.
- Crear una función *plot\_sectors\_by\_date* que reciba como parámetros el *dataframe*, una lista de sectores, y dos años, uno de inicio y otro de fin.

Esta función debe generar una gráfica lineal, donde se muestre la progresión de los datos de los sectores seleccionados, entre las dos fechas.

 ZIP**electronic-card-transactions.csv.zip**

28.9 KB

**VER SOLUCIÓN**

## SOLUCIÓN

Solución disponible en el notebook “solucion\_ejercicio\_repaso.ipynb”.

 ZIP**solucion\_ejercicio\_repaso.ipynb.zip**

25.1 KB

