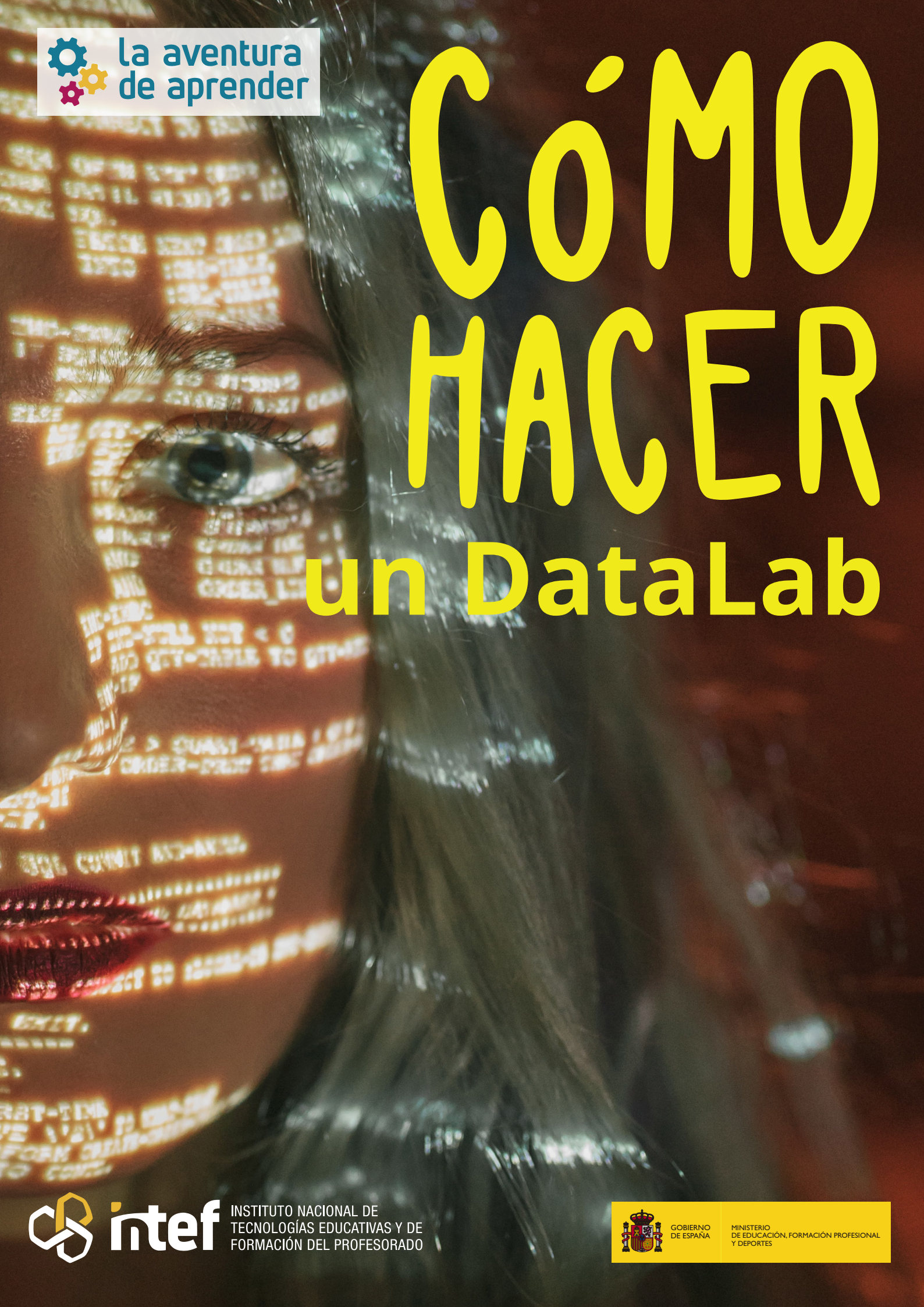
la aventura de aprender

CÓMO HACER un DataLab



intef
INSTITUTO NACIONAL DE TECNOLOGÍAS EDUCATIVAS Y DE FORMACIÓN DEL PROFESORADO

GOBIERNO DE ESPAÑA
MINISTERIO DE EDUCACIÓN, FORMACIÓN PROFESIONAL Y DEPORTES

MINISTERIO DE EDUCACIÓN, FORMACIÓN PROFESIONAL Y DEPORTES
Dirección General de Evaluación y Cooperación Territorial
Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado (INTEF)
Recursos Educativos Digitales

la aventura de aprender

La **Aventura de Aprender** es un espacio de encuentro e intercambio en torno a los aprendizajes para descubrir **qué prácticas, atmósferas, espacios y agentes hacen funcionar las comunidades;** sus porqués y sus cómos o en otras palabras, sus anhelos y protocolos.

Este proyecto parte de unos presupuestos mínimos y fáciles de formular. El primero tiene que ver con la convicción de que **el conocimiento es una empresa colaborativa, colectiva, social y abierta.** El segundo abraza la idea de que **hay mucho conocimiento que no surge intramuros de la academia** o de cualquiera de las instituciones canónicas especializadas en su producción y difusión. Y por último, el tercero milita a favor de que **el conocimiento es una actividad más de hacer que de pensar** y menos argumentativa que experimental.

Estas guías didácticas tienen por objetivo **favorecer la puesta en marcha de proyectos colaborativos que conecten la actividad de las aulas con lo que ocurre fuera del recinto escolar.**

Sin aventura no hay aprendizaje, ya que las tareas de aprender y producir son cada vez más inseparables de las prácticas asociadas al compartir, colaborar y cooperar.

http://laaventuradeaprender.intef.es

Antonio Lafuente
para INTEF
https://intef.es

NIPO (formato html) 164-24-001-7
NIPO (formato PDF) 164-24-002-2
NIPO (formato web) 164-24-010-3
DOI (formato web) 10.4438/LADA_164240103
DOI (formato PDF) 10.4438/LADA030_2024
Por Wenceslao Arroyo Machado para INTEF

Para cualquier asunto relacionado con esta publicación contactar con:
Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado
C/Torrelaguna, 58. 28027 Madrid.
Tfno.: 91-377 83 00. Fax: 91-368 07 09
Correo electrónico: lada@educacion.gob.es

ÍNDICE

Quién hace esta guía

Wenceslao Arroyo Machado es Doctor en Tecnologías de la Información y la Comunicación por la Universidad de Granada (UGR), donde se ha especializado en la evaluación de la ciencia y su atención en medios sociales mediante un enfoque big data.

Esta línea surge como combinación de sus estudios en el Grado de Información y Documentación y el Máster Oficial en Ciencia de Datos e Ingeniería de Computadores, ambos también en la UGR.

Desempeña una participación activa en el ámbito de la investigación, donde ha publicado en revistas internacionales y forma parte del grupo EC3 Research Group. Asimismo, es Chief Operating Officer (COO) de EC3Metrics, Spin-Off de la Universidad de Granada en la que participa dirigiendo la elaboración de informes institucionales e impartiendo docencia de cursos especializados.

También se encarga de la subdirección de #YoSigoPublicando, el proyecto de formación docente de la Universidad de Granada.

INTRODUCCIÓN

De la sociedad de la información a la del conocimiento

En 2022, se estimó que cada minuto se publicaron 66,000 fotografías en Instagram y se subieron 500 horas de vídeos a YouTube. Estas cifras reflejan el volumen masivo de datos que se generan continuamente en nuestra era digital y de los cuales muchas veces no somos conscientes. Cada interacción que realizamos, desde una simple búsqueda hasta un clic en una página web, contribuye a este vasto mar de datos. La velocidad y acumulación de datos es tal que tenemos que introducir periódicamente nuevos términos para simplemente poder cuantificar estas inmensas cantidades. Sin embargo, una pregunta que tenemos que hacernos ante este escenario es: ¿cuántos de estos datos se acaban transformando en información valiosa?

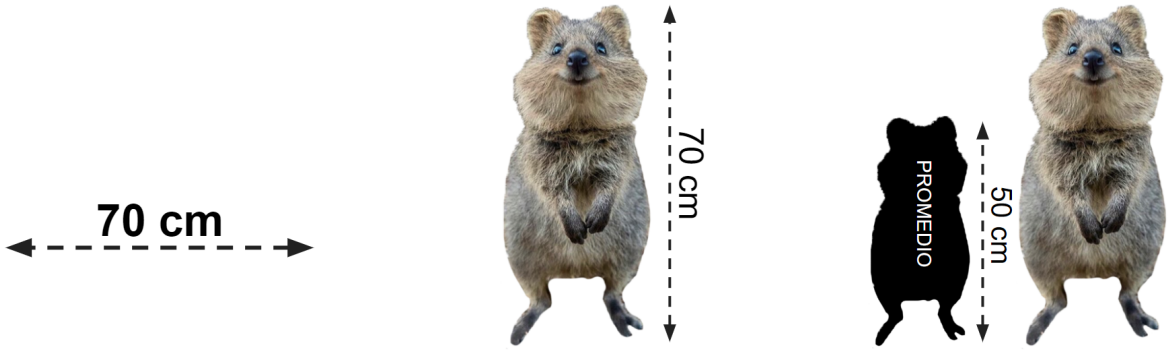
Más allá de lo impresionante de estas cifras y los desafíos asociados en cuanto a su almacenamiento y gestión, este incremento de datos ha tenido un impacto muy notable en la sociedad y la economía. Desde la aparición de la computación en el siglo XX, hemos sido testigos de avances significativos en el ámbito de las Tecnologías de la Información y Comunicación (TIC), con Internet como uno de sus desarrollos más transformadores. Inicialmente, la revolución digital iniciada con la llegada de ordenadores y registros digitales condujo a un aumento en la producción y el acceso a la información, dando origen a lo que conocemos como **sociedad de la información.** Esta fase se caracteriza por una disponibilidad e intercambio de datos e información sin precedentes.

Sin embargo, con el tiempo, la mera disponibilidad de datos e información se ha visto insuficiente, pues sin una correcta interpretación, comprensión y aplicación carecen de utilidad. Es por ello, que se ha producido una transición hacia la denominada **sociedad del conocimiento,** donde el énfasis se encuentra en cómo utilizar y aplicar la información para abordar problemas y tomar decisiones. Por ejemplo, un hospital puede tener miles de registros médicos digitalizados, pero sin las herramientas adecuadas para analizarlos, no sirven para prever y mejorar la atención al paciente. No basta con tener información; necesitamos destilarla en conocimiento para actuar de forma efectiva.

Timo Volz. Licencia: Unsplash

El valor de los datos

Llegados a este punto, y antes de continuar, conviene que ilustremos mejor la distinción entre tres conceptos clave que ya han aparecido y que a menudo se usan de forma indistinta. Se trata del **dato**, la **información** y el **conocimiento.** Consideremos para ello el valor "70 cm". Aislado, este número es simplemente un dato sin contexto. Pero si especificamos que "70 cm" es la altura de un quokka adulto, el dato se convierte en información que nos permite contextualizar un hecho, en este caso la altura de un marsupial concreto. Si además se entiende que un quokka de esa altura es más grande que la mayoría y, por ende, es una rareza o excepción dentro de su especie, hemos adquirido un conocimiento específico sobre el quokka. Esta progresión desde un simple dato hasta una comprensión más profunda ilustra la importancia de interpretar y contextualizar adecuadamente los datos.



DATO	INFORMACIÓN	CONOCIMIENTO
Representación de un atributo que en sí mismo no tiene valor	Contextualización del dato que le otorga significado	Procesamiento de diferentes datos mediante relaciones que dan una comprensión más profunda
Tenemos un dato en bruto que es 70 cm	Sabemos que los 70 cm es la altura de un quokka adulto concreto	Al saber que la altura habitual de un quokka adulto es de 50 cm podemos concluir que nuestro quokka es excepcionalmente alto

Ejemplo para ilustrar la diferencia entre dato, información y conocimiento
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

La complejidad al transformar datos en información útil y luego en conocimiento significa que no es simplemente una cuestión de interpretación. Requiere procesos meticulosos y herramientas adecuadas para extraer significado y valor.

Dicho esto, y considerando la importancia de procesar adecuadamente estos datos, en la era moderna, el término *big data* se ha vuelto cada vez más familiar. Se refiere a enormes conjuntos de datos que superan la capacidad de procesamiento de sistemas tradicionales, demandando nuevas técnicas y herramientas para su análisis. Esta emergencia del *big data* ejemplifica cómo la problemática de gestionar grandes volúmenes de datos ha evolucionado con el tiempo.

Lo que una vez fue considerado un desafío casi insuperable, ahora es parte de nuestra rutina diaria gracias a los avances tecnológicos. Pero, es esencial destacar que la capacidad de extraer información no se limita únicamente a estos enormes conjuntos de datos. **Ya sea en el ámbito del big data o en escalas menores, ahora podemos obtener, procesar y analizar información de manera eficiente.** Lo que hasta hace muy poco era inabarcable, se ha convertido en una cuestión cotidiana. Estamos en una época en la que, independientemente del tamaño de los datos, tenemos el poder y las herramientas para transformarlos en conocimientos valiosos.

Podemos concluir que vivimos en un mundo repleto de datos inexplorados en todos los aspectos. Asimismo, **no importa el campo profesional en el que nos encontremos,** el entendimiento de los datos para su exploración y transformación en conocimiento se ha convertido en una competencia esencial y transversal.

Comprender cómo se extraen, gestionan, procesan y visualizan los datos, no solo suponen competencias clave en esta nueva sociedad, sino que también permiten desarrollar un pensamiento crítico más profundo, una mayor curiosidad y una perspectiva informada sobre el mundo. Es por ello, que con esta guía el objetivo no es crear solo un laboratorio para fomentar y transmitir competencias básicas de procesamiento y tratamiento de datos para generar conocimiento, sino también para desarrollar una mentalidad despierta, colaborativa e innovadora.

A través de esta guía iremos desde los fundamentos más básicos hasta la aplicación real de la ciencia de datos, el campo que engloba a las distintas

Holly Mandarich. Licencia: Unsplash

técnicas involucradas en el procesamiento de los datos. Empezaremos introduciendo este concepto, desgranando tras ello las particularidades y formatos de los datos. Después, nos orientaremos sobre cómo iniciar un proyecto de ciencia de datos, estableciendo desde el entorno de trabajo hasta la formación del equipo necesario. Seguiremos con las etapas cruciales de recopilación y limpieza de datos, esenciales para garantizar la calidad de cualquier análisis. Posteriormente, nos adentraremos en las técnicas de análisis y visualización, y concluiremos con la redacción y distribución de resultados del proyecto.

Términos

- **Nube:** Servidor conectado a internet donde guardamos y accedemos a información o programas sin necesidad de tenerlos o procesarlos en nuestro ordenador.
- **Big data:** Término que hace referencia a una gran cantidad de datos que es complicada de manejar con herramientas normales.
- **Dataset:** Conjunto de datos recopilados en uno o varios archivos.
- **Repositorio:** Archivo digital donde se guardan y organizan archivos o programas para que sean accesibles a otras personas.
- **Curación de datos:** Proceso de organizar, corregir y mejorar datos para que estén listos para ser usados.
- **Minería de datos:** Proceso de buscar patrones o información útil en una gran cantidad de datos.

MATERIALES

A continuación, se incluyen diferentes alternativas para la elección de herramientas. Es recomendable, en la medida de lo posible, optar por un ecosistema concreto al completo (Google Workspace, Microsoft 365, iWork + iCloud o Nextcloud) para favorecer la interoperabilidad entre sus herramientas y lograr una mayor eficiencia y agilidad.

Google Workspace	Microsoft 365	iWork + iCloud	Nextcloud
ALMACENAMIENTO DE DATOS EN LA NUBE			
Google Drive Servicio de almacenamiento en la nube de Google que permite guardar archivos, sincronizarlos entre dispositivos y compartirlos con otros usuarios.	**Microsoft OneDrive** Servicio de almacenamiento en la nube de Microsoft, diseñado para guardar archivos, acceder a ellos desde cualquier dispositivo y compartirlos fácilmente con otros usuarios.	**iCloud** Servicio de almacenamiento en la nube de Apple que permite a los usuarios guardar archivos y acceder a ellos desde cualquier dispositivo Apple, así como desde la web.	**Nextcloud files** Plataforma de colaboración de software libre que proporciona funcionalidades de sincronización y compartición de archivos.
RECOLECCIÓN DE DATOS			
Google Forms Herramienta Google que permite crear encuestas y formularios en línea para recopilar respuestas en tiempo real.	**Microsoft Forms** Herramienta de Microsoft que permite crear encuestas, cuestionarios y formularios y analizar resultados.	**Numbers** Numbers es la hoja de cálculo de Apple y que admite funciones para la creación de formularios.	**Nextcloud Forms** Aplicación de Nextcloud para la creación y gestión de cuestionarios.
PROCESAMIENTO DE DATOS			
Google Sheets Herramienta de hojas de cálculo de Google para la gestión y análisis de datos estructurados en tablas.	**Microsoft Excel** Herramienta de hojas de cálculo de Microsoft para organizar y analizar datos estructurados en tablas.	**Numbers** Numbers es la hoja de cálculo de Apple, diseñada con una perspectiva más visual que otras herramientas.	**Nextcloud Spreadsheet** Herramienta de Nextcloud Office para llevar a cabo análisis de datos con fórmulas y herramientas de hojas de cálculo.

OTRAS HERRAMIENTAS - VISUALIZACIÓN DE DATOS	
Tableau Public Herramienta que permite transformar datos en gráficos y visualizaciones interactivas, permitiendo compartir en línea.	**Datawrapper** Herramienta que permite convertir datos en gráficos en los que presentar información de forma clara y visual.

PASOS

1

En esta sección se describe la construcción de un data lab y el proceso de elaboración de proyectos ilustrando los diferentes procesos clave a considerar en su desarrollo. Pero antes de todo ello presentamos los conceptos ciencia de datos y data lab.

LA CIENCIA DE DATOS

¿QUÉ ES LA CIENCIA DE DATOS?

¿Has visto la película *Moneyball: Rompiendo las reglas?* Narra la historia real del gerente general del equipo de béisbol Oakland Athletics, que, junto con un economista, emplea la estadística para fichar a los mejores jugadores sin sobrepasar un presupuesto modesto. En esencia, esta estrategia se basa en la ciencia de datos: utilizando una base de datos que incluye las habilidades de los jugadores se identifican a los mejores para cada posición sin exceder el presupuesto. Este es uno de los muchos ejemplos de esta disciplina y con el que se entiende fácilmente su potencial.

Podemos definir así la ciencia de datos como un campo interdisciplinario que abarca una variedad de procesos dedicados a la **gestión y el análisis de datos,** independientemente de su volumen. Su virtud se encuentra en la capacidad para rascar en los datos mediante la denominada **minería de datos** para obtener información valiosa que no es tan evidente al examinarlos manualmente, sobre todo cuando nos enfrentamos a un importante volumen de datos. Pero, para revelar esa información oculta y mostrarla de manera clara, es esencial un trabajo previo en el que existen varias etapas y que siguen un orden específico, similar a la producción en cadena de una fábrica.

Sobre todo, se ha de tener en cuenta que, aunque este sea un trabajo de exploración, no se realiza nunca a ciegas. Es por ello, que una cuestión importante es **comprender qué representan los datos** con los que estamos trabajando. ¿Serías capaz de analizar datos de la bolsa de Tokio? Por más avanzado que sea tu dominio en las técnicas de la ciencia de datos, siempre es esencial entender qué información pueden brindar los datos y estar familiarizados con ellos. Esto te permitirá formular las preguntas más pertinentes, identificar problemas en los datos, realizar un análisis adecuado e interpretar correctamente los resultados.

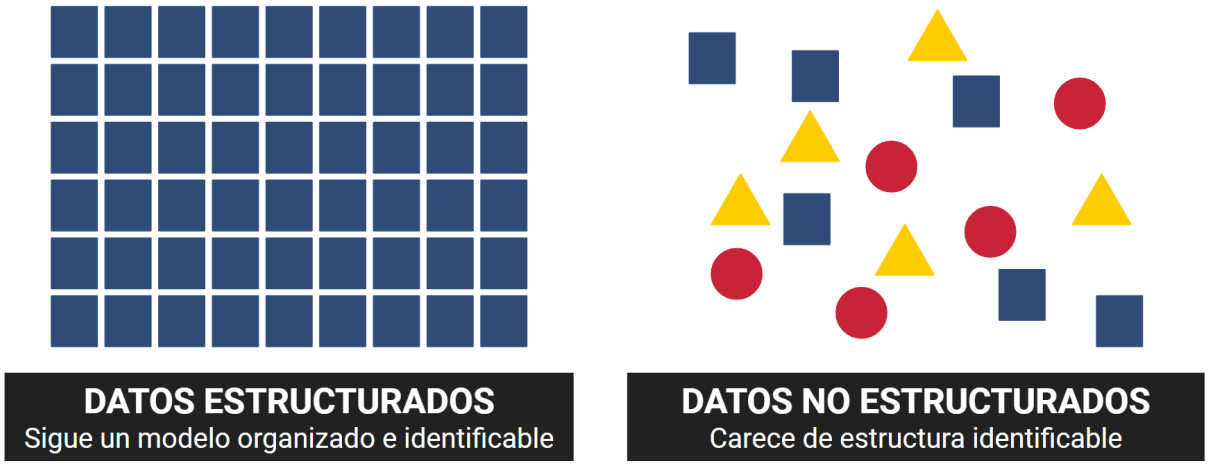
Pixabay. Licencia: CC0

Los datos y sus formatos

Si múltiples son las operaciones que entran dentro de la ciencia de datos, también lo son los propios datos. Una distinción básica en este sentido es la que podemos hacer entre **datos estructurados y no estructurados.** Los primeros son rápidamente reconocibles, pues muestran una organización y estructura fácilmente comprensible e interpretable. Por ejemplo, volviendo a la película de antes, una tabla de datos de jugadores de béisbol en la que cada fila es un jugador y cada columna un atributo de este, como la velocidad o fuerza de lanzamiento. Por su parte, los datos no estructurados combinan diferentes formatos y contenidos sin seguir una estructura tan clara y evidente. Por ejemplo, un video tiene varios componentes, principalmente imágenes y sonido, y estos no tienen una estructura simple como la de tabla, sino que es mucho más compleja.

El caso estándar en la ciencia de datos parte de datos estructurados. De este modo, se suele trabajar con datos que siguen un **modelo de tabla.** Esto es claramente apreciable con las hojas de cálculo de programas como Microsoft Excel, desde las cuales es posible además llevar a cabo muchos de los procesos básicos de la ciencia de datos, inclusive la visualización mediante gráficos. Eso sí, aunque se trabaje con datos estructurados en tablas estos pueden estar almacenados en diferentes formatos de archivo, como el propio archivo de Microsoft Excel (archivos .xls o .xlsx).

No obstante, cabe destacar que este es un formato denominado propietario, lo que quiere decir que sin haber adquirido una licencia de Microsoft Excel no se puede usar esta herramienta y acceder a ellos, algo que puede impedir su uso a otras personas que no dispongan de esta herramienta. Afortunadamente, existen los conocidos



Diferencias entre los modelos de datos estructurados y datos no estructurados
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

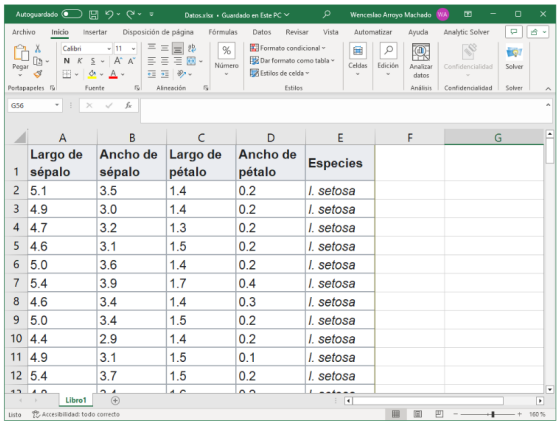
como **formatos abiertos** y que permiten solucionar esta limitación.

El **formato abierto** básico para los datos estructurados es el archivo de texto plano (.txt). En este archivo los datos pueden organizarse en forma de tabla con cada caso o individuo (los jugadores de béisbol) representado en una línea de texto, mientras que cada atributo o variable (velocidad, fuerza de lanzamiento...) se separa con un carácter específico conocido como delimitador. Habitualmente se emplea la coma, punto y coma, o tabulación.

Dependiendo del carácter delimitador utilizado, podemos transformar el archivo de texto en un archivo de valores delimitados por comas (.csv) o por tabulaciones (.tsv). Ambos formatos son estándares en la ciencia de datos, y es común almacenar y compartir datos en estos formatos. La ventaja en ello está en que todas las personas, con independencia de si usan Microsoft Excel, Google Spreadsheet, Numbers o cualquier herramienta de hojas de cálculo, van a poder hacer uso de esos datos sin ninguna limitación.

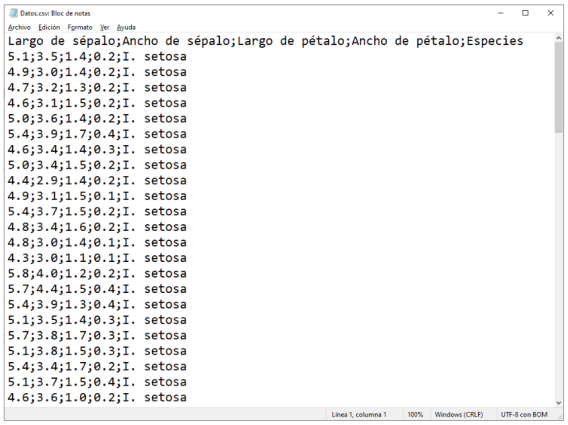
Nota

Durante la guía, al final de cada sección, se incluye un espacio como este siguiendo el desarrollo de un supuesto práctico a modo de ejemplo para ilustrar de todo el proceso de creación de un data lab y, sobre todo, cada una de las fases involucradas en desarrollo de un proyecto.



	A	B	C	D	E
1	Largo de sépalo	Ancho de sépalo	Largo de pétalo	Ancho de pétalo	Especies
2	5.1	3.5	1.4	0.2	I. setosa
3	4.9	3.0	1.4	0.2	I. setosa
4	4.7	3.2	1.3	0.2	I. setosa
5	4.6	3.1	1.5	0.2	I. setosa
6	5.0	3.6	1.4	0.2	I. setosa
7	5.4	3.9	1.7	0.4	I. setosa
8	4.6	3.4	1.4	0.3	I. setosa
9	5.0	3.4	1.5	0.2	I. setosa
10	4.4	2.9	1.4	0.2	I. setosa
11	4.9	3.1	1.5	0.1	I. setosa
12	5.4	3.7	1.5	0.2	I. setosa

FORMATO PROPIETARIO
Archivo *.xlsx* solo legible desde Microsoft Excel



```
Largo de sépalo;Ancho de sépalo;Largo de pétalo;Ancho de pétalo;Especies
5.1;3.5;1.4;0.2;I. setosa
4.9;3.0;1.4;0.2;I. setosa
4.7;3.2;1.3;0.2;I. setosa
4.6;3.1;1.5;0.2;I. setosa
5.0;3.6;1.4;0.2;I. setosa
5.4;3.9;1.7;0.4;I. setosa
4.6;3.4;1.4;0.3;I. setosa
5.0;3.4;1.5;0.2;I. setosa
4.4;2.9;1.4;0.2;I. setosa
4.9;3.1;1.5;0.1;I. setosa
5.4;3.7;1.5;0.2;I. setosa
4.8;3.4;1.6;0.2;I. setosa
4.8;3.0;1.4;0.1;I. setosa
4.3;3.0;1.1;0.1;I. setosa
5.8;4.0;1.2;0.2;I. setosa
5.7;4.4;1.5;0.4;I. setosa
5.4;3.9;1.3;0.4;I. setosa
5.1;3.5;1.4;0.3;I. setosa
5.7;3.8;1.7;0.3;I. setosa
5.1;3.8;1.5;0.3;I. setosa
5.4;3.4;1.7;0.2;I. setosa
5.1;3.7;1.5;0.4;I. setosa
4.6;3.6;1.0;0.2;I. setosa
```

FORMATO LIBRE
Archivo *.csv* sin restricciones en su uso

Diferencias entre datos en archivos de formato propietario y formato libre
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

2 Creación de un Data Lab

¿Qué es un Data Lab?

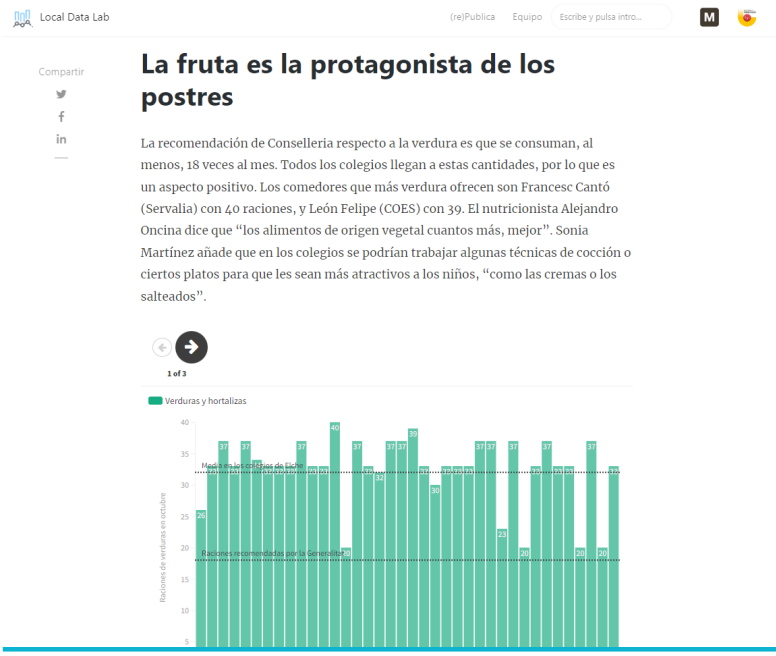
Una vez estamos familiarizados con algunos de los conceptos básicos de la ciencia de datos podemos empezar a construir nuestro data lab o laboratorio de datos. Este es un **espacio común y abierto para el desarrollo de proyectos colaborativos basados en la ciencia de datos** con el objetivo de generar información y fomentar el aprendizaje en torno a los datos. Existen multitud de ejemplos de estos laboratorios de datos, pudiendo encontrar varios vinculados a instituciones y entidades educativas. Uno de ellos es el Data Lab de Periodismo de la Universidad Miguel Hernández, en el que exploran datos relativos a la ciudad de Elche. Es de hecho una práctica común la creación de data labs con el objetivo de aprovechar los datos de la propia institución y generar información valiosa con ellos. Debemos tener claro por ello desde el inicio, si vamos a centrar nuestro data lab en un tema o problema específico o si su enfoque será más abierto.

Búsqueda de espacios

El primer paso para la creación de un data lab se encuentra en la búsqueda de un espacio. En realidad, de dos:

1. Un **espacio físico, seguro y accesible** en donde se desarrollará el laboratorio. Tiene que disponer de un equipamiento mínimo de ordenadores o enchufes para que, en su defecto, las personas que participen puedan llevar sus portátiles. También es indispensable disponer de acceso a internet, pues será donde se encuentre nuestro segundo espacio.

2. Un **espacio en la nube,** para el cual recurriremos principalmente a servicios como los de Google Workspace o Microsoft OneDrive para construir un entorno de trabajo virtual colaborativo.



Ejemplo de proyecto de Data Lab de Periodismo de la Universidad Miguel Hernández.
La fruta es la protagonista de los postres

Elaboración de proyectos

Una vez tengamos nuestros dos espacios, llega el momento de poner en funcionamiento el data lab. Para ello estableceremos un calendario fijo con **sesiones semanales de dos horas** en horario extraescolar, por ejemplo, los miércoles de 17 a 19 horas. Hecho esto, podremos anunciar nuestro data lab y abrir su inscripción, usando para ello el correo electrónico o un formulario web. En este punto tenemos que recordar que **un data lab es un espacio multidisciplinar** en el que son necesarias personas para tareas diferentes al propio análisis de los datos, como la redacción de los resultados, por lo que debemos matizarlo al hacer difusión.

En cuanto contemos con 2 o 3 participantes podremos empezar a poner en marcha el desarrollo del data lab, el cual pasa por la elaboración de proyectos de ciencia de datos. En una primera sesión nos encargaremos de presentar la iniciativa y definir un proyecto en el que trabajar durante los siguientes tres meses, aproximadamente 10 sesiones. La mejor manera para ello es realizar una sesión de ***brainstorming*** de no más de una hora, en las que todas las personas propongan sus ideas abiertamente, valorándose por parte de la persona organizadora su viabilidad. Esto es fundamental ya que un proyecto no debe empezar a rodar si no tenemos claro que podamos tener acceso a los datos y que su análisis es viable. Como recomendación, tendremos preparadas propuestas viables para plantear si no surgen suficientes ideas.

En este punto, lo más habitual es que no tengamos del todo claro cuáles serán los datos concretos con los que trabajaremos o la visualización final que generaremos, pero sí es importante que dejemos claro:

Elaboración de un proyecto de datos
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

1. **Objetivo del proyecto** – ¿Qué vamos a analizar?
2. **Disponibilidad de los datos y complejidad del análisis** – ¿Cómo lo vamos a analizar?

Para ello basta con que respondamos a ambas preguntas de manera muy precisa y concreta. Por ejemplo, para un proyecto podemos (1) querer estudiar cuál es la fruta que más come el alumnado de nuestro centro a la semana durante la primavera y (2) estudiarlo analizando los resultados de una encuesta que haremos al alumnado. Cuanto más podamos precisar y menos demos p**or sentado mejor.** Es posible que durante el desarrollo decidamos poner el foco en un aspecto más concreto o cambiar el método de análisis, pero debemos garantizar un punto de partida firme.

Formación del equipo de trabajo

Nuestro data lab contará con al menos un equipo de trabajo el cual se definirá en la sesión inicial una vez aprobemos el proyecto a realizar. Si disponemos de un grupo de personas amplio **podremos considerar realizar varios proyectos y organizar a las personas en distintos equipos.** En caso de optar por ello, daremos libertad a los participantes para apuntarse libremente al proyecto que más les interese, pero controlaremos que no queden desequilibrados. Aunque el número dependerá en gran medida de la complejidad y tamaño del proyecto, como estimación, cada equipo estará constituido idealmente por entre 3 y 10 personas.

Además de constituir el equipo o equipos, **también dejaremos fijados en esta primera sesión los roles de los participantes,** los cuales son elegidos libremente. Estos roles no implican que sea una persona la que desarrolle en exclusividad una actividad, sino que será la responsable directa de que dicha tarea se realice correctamente y

coordinara para ello al resto del equipo. Un mismo rol puede ser desempeñado por varias personas al igual que una persona puede desempeñar al mismo tiempo varios roles. Los roles indispensables que considerar son los siguientes:

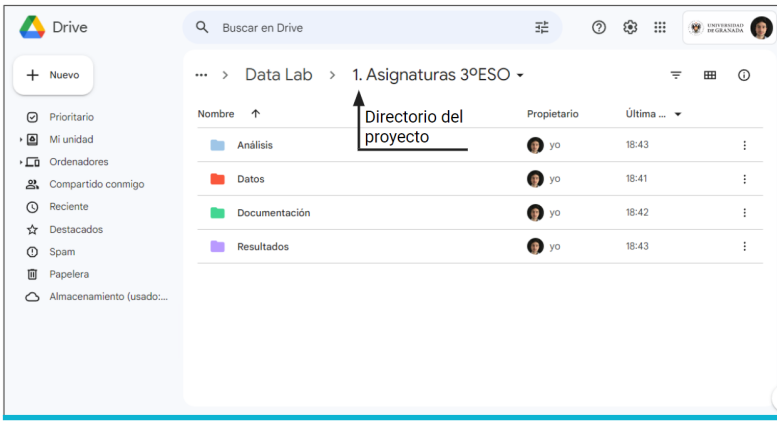
- **Responsable del proyecto (máximo 1 persona)** - se trata del supervisor/a del proyecto, es la persona encargada de revisar que el proyecto transcurra de manera correcta.
- **Rastreador/a de datos (máximo 3 personas)** - responsable de la búsqueda o recopilación de datos.
- **Arquitecto/a de datos (máximo 2 personas)** - responsable de la importación de los datos a la hoja de cálculo y de su limpieza.
- **Analista de datos (máximo 2 personas)** - responsable de examinar y analizar los datos.
- **Especialista en visualizaciones (máximo 1 persona)** - responsable de la generación de los gráficos.
- **Redactor/a de datos (máximo 1 persona)** - responsable de la redacción de una nota comunicando los hallazgos.

Creación del entorno de trabajo

Ya una vez finalizada la sesión, y antes de la siguiente, nos encargaremos de crear el entorno de trabajo colaborativo en la nube para el proyecto o proyectos e iremos agregando a cada una de las personas al mismo. Existen varios directorios básicos que van a estar presentes en todos los proyectos:

- **Análisis -** directorio con la hoja u hojas de cálculo en las que se llevarán a cabo los distintos análisis de los datos.
- **Datos -** directorio en el que se almacenan los datos en bruto, es decir, el fichero o fichero de datos a analizar. Sirve además como copia de seguridad de los datos.
- **Documentación -** directorio en el que se incluyen documentos de texto para la descripción de los datos o cualquier otro material destinado a la toma de notas o documentación.
- **Resultados -** directorio destinado a incluir un archivo de texto con la redacción de los resultados del análisis y el material gráfico generado.

Al trabajar en la nube debemos tener siempre en cuenta los permisos de los distintos directorios y archivos, para lo cual revisaremos y asignaremos permisos específicos de acuerdo con quienes participan en el proyecto. Por norma general, todos los miembros del equipo de trabajo de un proyecto tendrán acceso con permisos de edición a su proyecto, el cual será privado.



Creación de un directorio para un proyecto y subdirectorios básicos en Google Drive
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

Ejemplo

En un instituto de secundaria se ha creado un data lab en el que participa el alumnado bajo la supervisión y orientación del profesorado. En el primer proyecto de este data lab hemos fijado como objetivo abordar la pregunta *"¿cuáles son las asignaturas con mejores calificaciones del último año en 3º de ESO en nuestro instituto?"*, la cual responderemos tras realizar y analizar una encuesta. Tras ello hemos formado un equipo y definido roles. Por último, al terminar la sesión el coordinador del data lab ha creado un directorio en Google Drive con las carpetas "Análisis", "Datos", "Documentación" y "Resultados" e invitado a las personas del equipo.

3

Recopilación de Datos

Con nuestro objetivo y entorno de trabajo ya preparados nos toca localizar los datos con los que vamos a trabajar. Este es el objetivo principal de la segunda sesión, en la cual guiaremos al equipo para que de manera colaborativa localice o recoja los datos con los que trabajaremos en el resto del proyecto. Es necesario que antes de la siguiente sesión el dataset se encuentre disponible en el directorio de la nube con independencia del método escogido:

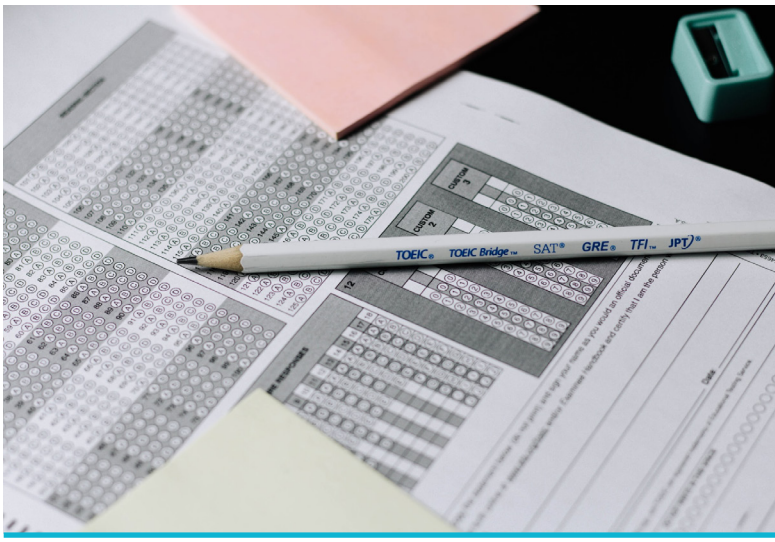
- recogida de datos
- solicitud de datos
- búsqueda de datos públicos

Para saber por cuál optar debemos tener claro si los datos que queremos analizar existen y si estos son accesibles de manera pública o mediante algún proceso de solicitud.

Recogida de datos

¿Tenemos pensado realizar un análisis concreto, pero sabemos de antemano que no existe ningún archivo que contenga los datos o un documento de donde se puedan extraer? En este caso seremos nosotros los que **nos encargaremos de la recogida de los datos,** siguiendo para ello un proceso que variará en función de cada caso. Lo que es fundamental aquí es que, con independencia del método seguido, generemos al final uno o varios archivos con los datos estructurados, ya sea un archivo de texto o una hoja de cálculo.

Si bien existen distintos métodos para la recogida de datos, priorizaremos el desarrollo de cuestionarios. Y es que solo tendremos que preparar un **formulario online** a través de herramientas como Google Forms o Microsoft Forms y solicitar



Nguyen Dang Hoang Nhu. Licencia: Unsplash

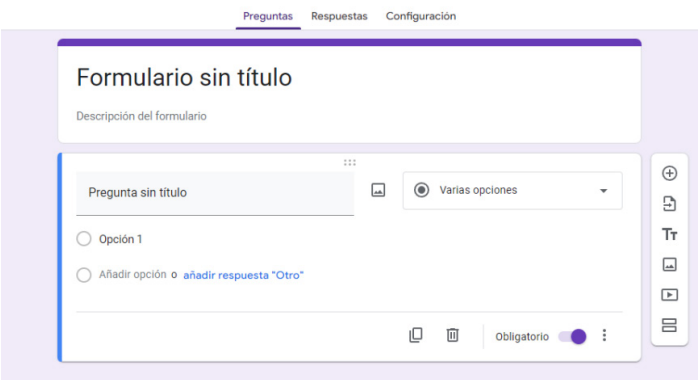
a un grupo de personas que lo rellene de manera online. La ventaja que tiene este procedimiento es que iremos recopilando los datos de las respuestas de manera automática y en un formato estructurado, pudiendo al final generar y exportar automáticamente los datos para analizarlos. Pero para que el análisis se lleve a cabo de manera correcta tenemos que elaborar de manera correcta este cuestionario. La formulación aséptica de las preguntas, su claridad, simplicidad y organización, son las principales consideraciones a tener en cuenta.

Al igual que elaborar un formulario correctamente es crucial, también lo es establecer una **estrategia de difusión.** De nada sirve tener un cuestionario válido si no responde nadie o solo una parte muy específica y que puede sesgar los resultados. Debemos tener claro a que grupo queremos lanzar la encuesta y garantizar que respondan la mayoría. Eso sí, no debemos buscar tener respuesta de todo el mundo, pero sí tener un porcentaje representativo y equilibrado que evite los problemas mencionados de sesgos.

Si este fuese nuestro caso, la sesión se organizaría de la siguiente manera:

1. Presentación de los cuestionarios
2. Diseño de un cuestionario en la herramienta escogida
3. Elaboración de un cuestionario válido y diseño de un plan de difusión

La elaboración del cuestionario es llevada a cabo directamente por el equipo bajo nuestra supervisión y su realización debe completarse antes de la siguiente sesión para poder trabajar con los resultados.



Creación de un cuestionario a través de Google Forms
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

Solicitud de Datos

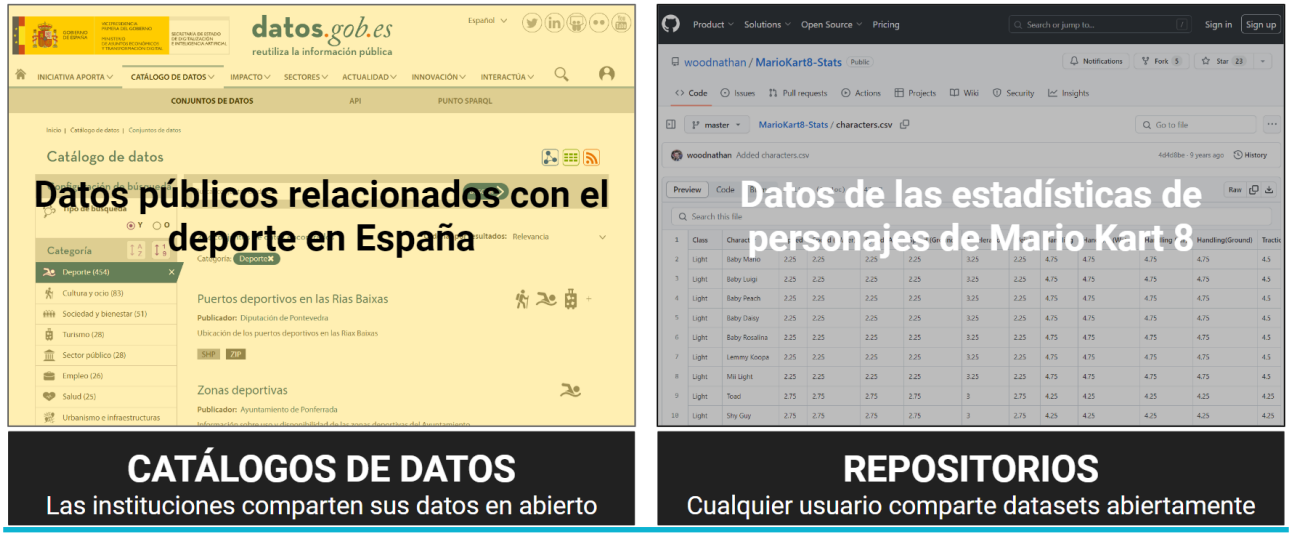
¿Sabemos que los datos que queremos explorar están en un archivo o un documento pero que este no es público? En este caso podemos realizar la solicitud formal de dichos datos. Un ejemplo básico de esta situación sería la solicitud a nuestra institución acerca de las matriculaciones para analizar la evolución del número de matriculados en los últimos cursos. Si bien no se trata de datos públicos, conocemos la existencia de estos y podemos solicitarlos. A diferencia de los formularios online, **este proceso puede requerir de algún esfuerzo extra para adaptar esos datos a un archivo digital de datos estructurado.** Por ejemplo, puede que nos den la información en papel, y nosotros tengamos que introducir esos datos manualmente en una hoja de cálculo. Al igual que con los formularios, garantizaremos que la solicitud y preparación del archivo se realice antes de la siguiente sesión.

Búsqueda de Datos Públicos

¿Los datos que queremos estudiar se encuentran disponibles en abierto en la web? Este es el escenario más rápido y sencillo. Existen una gran cantidad de **fuentes de datos en abierto** a partir de los cuales podemos recuperar datasets que ya se encuentran estructurados y en formatos estándar:

- **Catálogos de datos gubernamentales** - datos sobre información y servicios públicos
  Ejemplo: Datos abiertos del Gobierno de España
- **Repositorios** - plataforma para la subida de datos para ser compartidos en abierto
  Ejemplo: Kaggle
- **Buscadores** - herramientas para la búsqueda de datasets en abierto
  Ejemplo: Google Dataset Search

En este caso organizaremos la sesión para que, una vez seleccionadas y presentadas varias fuentes de datos útiles para el proyecto, el equipo se coordine y localice en ellas un dataset válido.



Los catálogos de datos y repositorios son los principales puntos de acceso a datos en abierto
Wenceslao Arroyo Machado. Licencia: CC BY 4.0

Protección de datos personales

Un aspecto que no debemos descuidar en ninguno de los tres casos descritos es la protección de datos personales, en lo cual incidiremos en la segunda sesión para formar sobre el uso responsable de datos. Desde el momento en que recopilamos información personal hasta su posterior almacenamiento, procesamiento y análisis, cada paso debe estar guiado por principios y prácticas que protejan a las personas. La protección de datos no es simplemente una normativa o una recomendación; es un **derecho fundamental que se encuentra estrechamente vinculado a las libertades individuales.** Ignorarlo o subestimarlo puede tener graves consecuencias, tanto en lo legal como en cuestiones de confianza y seguridad. Es crucial ser conscientes de que cualquier desviación o negligencia en el resguardo o publicación de datos personales puede resultar en sanciones significativas y daños irreparables.

Ejemplo

Para poder responder a la pregunta *"¿cuáles son las asignaturas con mejores calificaciones del último año en 3º de ESO?"* solicitamos los datos al centro, garantizando su uso responsable y de acuerdo a la ley de protección de datos personales, un tema que fue tratado durante la segunda sesión para formar sobre ello.

3. Tratamiento de datos personales

4 Limpieza de Datos

¿Sabías que en proyectos big data el **60% y el 80% del tiempo del proyecto se destina a la limpieza de los datos?** Puede parecer una tarea menor, sobre todo en proyectos en los que se trata con pequeños datasets, pero lo cierto es que gran parte del éxito del proyecto se encuentra en esta fase.

La limpieza de datos consiste en la revisión de los mismos para la **identificación y resolución de problemas.** Imaginemos el siguiente escenario: tras hacer una encuesta a 30 personas generamos un archivo de texto con las respuestas y al importarlo a nuestra hoja de cálculo descubrimos que solo hay 20 respuestas, que en respuestas que deben ser numéricas aparece texto o que respuestas obligatorias aparecen sin rellenar. Identificar todos estos problemas y la raíz de estos para resolverlos se convierte en un ejercicio clave si no queremos tener errores en el posterior análisis e incluso obtener resultados que no son correctos.

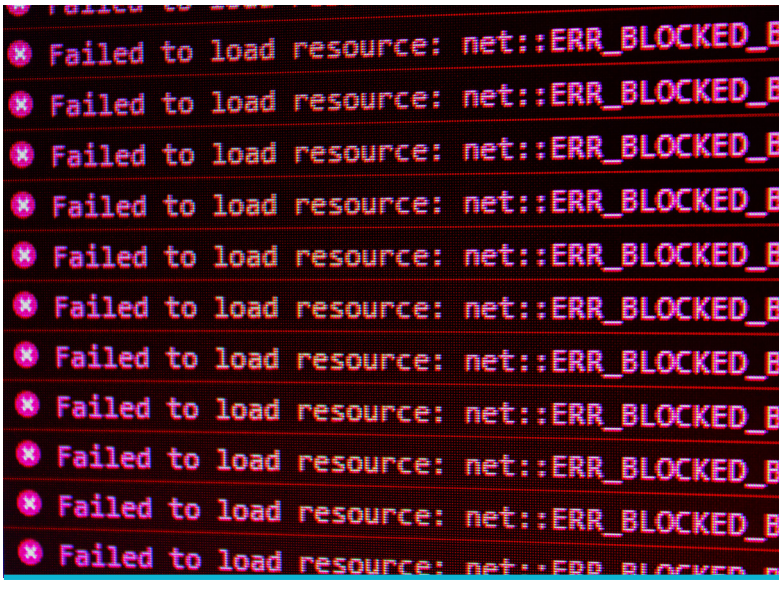
Este es un momento muy valioso para familiarizarse con los datos y su estructura, así como saber reconocer cuáles son los fallos y anticiparse con ellos. Si bien son inabarcables los problemas que podemos encontrar, vamos a centrarnos en los más comunes, los cuales se dividen en dos grandes bloques: **la importación y los problemas de los datos.**

Para ello, dedicaremos las sesiones de la 3 a las 5 a este proceso de la siguiente manera:

1. Presentación a modo de **seminario** de la herramienta escogida y principales procesos de la importación y limpieza de los datos. Podemos dedicar una sesión completa y trabajar en conjunto sobre un dataset de prueba que hayamos descargado de Kaggle.

2. Importación y limpieza del dataset del proyecto por parte del equipo, supervisando en todo momento su correcto desarrollo y la participación activa de todas las personas.

Al final de esta fase tendremos los datos almacenados en una hoja de cálculo en la nube, ya listos para su análisis.



David Pupăză. Licencia: Unsplash

Importación de datos

La fase de importación de datos de un archivo de texto a una hoja de cálculo es uno de los momentos más sensibles, ya que fácilmente puede introducir problemas en los datos que afecten a su análisis si no se realiza con precaución. La solución de estos problemas es muy sencilla, pues solo hay que repetir la importación para revisar y corregir los parámetros de importación.

Los problemas más frecuentes y que debemos considerar siempre son los siguientes:

Error	Descripción	Solución
Codificación	Presencia de caracteres extraños en el texto, lo que puede ser especialmente evidente en caracteres especiales como acentos o eñes. Por ejemplo, en lugar de "película" puedes encontrar "pelÃcula".	Utiliza un formato de codificación correcto cuando importes datos. Por lo general, UTF-8 es una opción segura para muchos conjuntos de datos.
Delimitadores	Los datos no se muestran en columnas o filas correctamente estructuradas. Esto puede manifestarse como datos amontonados o filas que parecen contener demasiados o muy pocos campos.	Cuando importes el archivo, asegúrate de especificar correctamente cuál es el delimitador utilizado. Los archivos .csv, pueden usar comas o puntos y comas.
Formato	Los datos no se muestran o no se comportan como debería. Este es un problema que puede pasar desapercibido hasta que realizas una operación con los datos y obtienes errores. Por ejemplo, los números son interpretados como texto.	Identifica las columnas o campos con problemas de formato y conviértelos al tipo de dato correcto.

Problemas de los datos

Los problemas con los datos son una cuestión más compleja para tratar en cuanto a que estos pueden venir dados por múltiples factores. Por este motivo, **no siempre vamos a poder solucionar los problemas.** Pero que no tengan solución no impide su uso, si no se trata de un dato clave o que pueda hacer que el análisis lleve a resultados erróneos, podremos continuar. Es importante por ello que cuando la importación haya sido realizada con éxito revisemos los datos en profundidad:

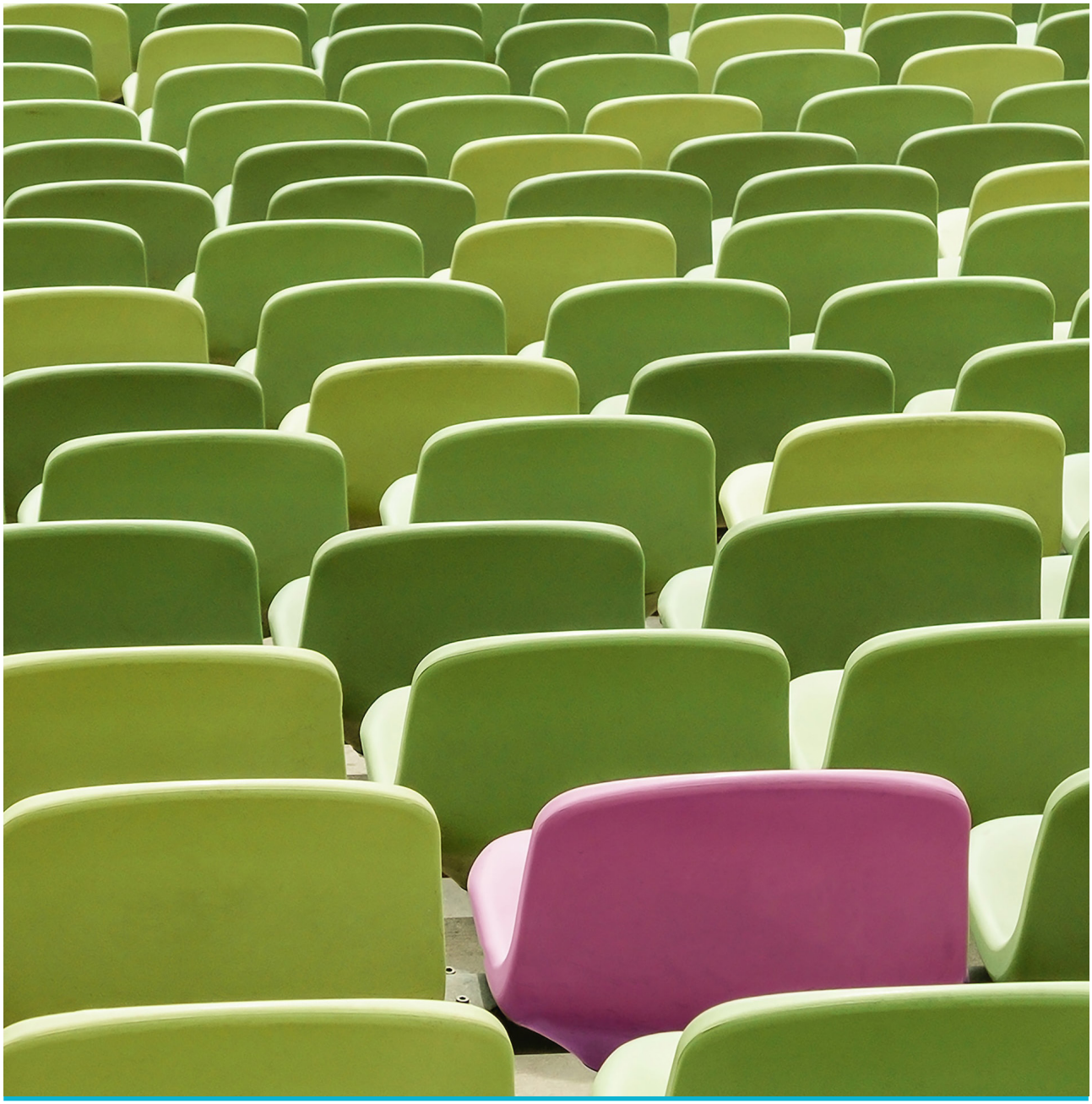
1. Identificando ausencias en los datos, anomalías y valores atípicos
2. Analizando si se tratan de fallos
3. En caso de ser errores, busquemos soluciones.

La presencia de datos anómalos, como **un valor numérico muchísimo más elevado o reducido que el resto, no debe verse siempre como un problema a tratar.** Que en los datos de las calificaciones de un examen todo el alumnado haya obtenido un 5 salvo una persona que tiene un 10, es algo plausible, pero que esa excepción sea un 15 sí es algo que debe ser revisado pues se sale del valor máximo esperado.

Como ejemplo de estos problemas, imagina que estamos trabajando con los datos que hemos generado de una encuesta. Al importar los datos nos encontramos con personas que han rellenado dos veces la encuesta cuando solo pueden una o que varias personas han dado una misma respuesta a una pregunta, pero usando variantes de la misma palabra (por ejemplo SÍ, sí, SI, SII!). En este caso tendremos que eliminar los duplicados y unificar las distintas variantes de la respuesta en una sola.

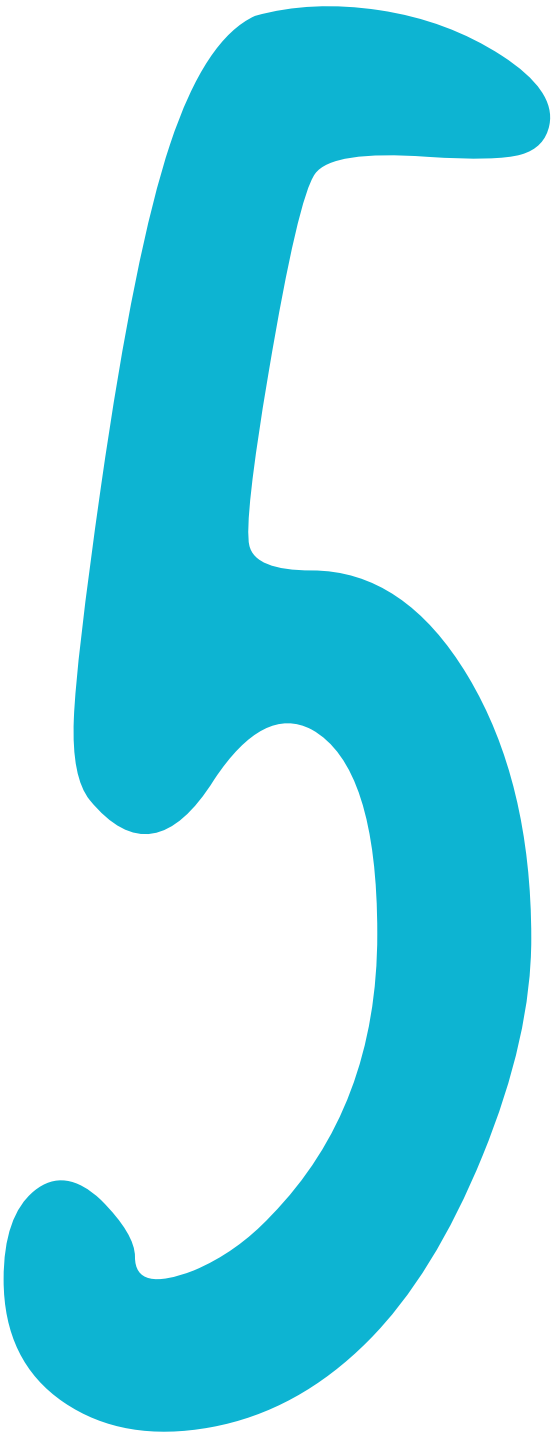
Ejemplo

Tras dedicar una sesión completa a explicar el uso de Google Sheets y los principales problemas de la importación de datos, dedicamos las dos sesiones siguientes a la importación de las notas a la hoja de cálculo y su limpieza. Revisando los datos de las calificaciones detectamos que las notas de una asignatura fueron introducidas dos veces (eliminamos una) y que las calificaciones estaban como valores texto (lo cambiamos a números).



Elena Ramseier. Licencia: Unsplash

Análisis de Datos



Con los datos preparados y revisados en una hoja de cálculo, llega el momento del análisis. Esta actividad la desarrollaremos entre las sesiones 6 y 7 de la siguiente manera:

1. Introducción mediante un **seminario** las principales técnicas de análisis de la herramienta escogida, trabajando en conjunto sobre el dataset de prueba
2. Puesta en común de los hallazgos o cuestiones que hayan podido llamar la atención durante las fases de recogida y limpieza de datos para su exploración
3. Elección de un método para el análisis principal de los datos
4. Exploración y análisis de datos por parte del equipo coordinando su correcta realización y guiando en el proceso
5. Validación de los resultados

Exploración general de los datos

Antes de llevar a cabo el análisis propiamente dicho, con el que pretendemos dar una respuesta clara a nuestro objetivo, es conveniente realizar una **pequeña exploración para comprender y contextualizar mejor nuestros datos** y los correspondientes resultados. Se trata de obtener una panorámica de los datos.

Pongamos un supuesto. Nuestro objetivo es conocer cuántas horas estudian de promedio el alumnado por curso y para ello hemos hecho una encuesta, que no todo el mundo ha respondido. Antes de calcular los distintos promedios, saber cuántas personas han respondido es importante, pues puede darse el caso de cursos que hayan tenido una elevada tasa de respuesta mientras que en otros solo han participado unas pocas personas. Conocer el número de estudiantes y el porcentaje de respuestas por clase nos permitirá saber el tamaño del estudio y si los datos son suficientemente representativos.

Para realizar la exploración debemos revisar los datos y describir su distribución. Esta es otra buena oportunidad para aprender a hacer una lectura e interpretación correcta de los datos. Algunas preguntas que pueden ser útiles para guiar un poco este proceso son:

- ¿Qué representan los datos?
- ¿Cuántos elementos estamos analizando?
- ¿Cuántas variables tienen?
- ¿Faltan datos?
- ¿Los datos están organizados en grupos?
- ¿Existen valores extremos?
- ¿Cuáles son los valores máximos y mínimos?

Vidar Nordli-Mathisen. Licencia: Unsplash

Análisis principal

Cuando nos hayamos familiarizado con los datos y los conozcamos suficientemente daremos paso al análisis principal. En esta ocasión sí partimos con un **objetivo y pregunta clara** a la que queremos dar respuesta. Es por ello por lo que este análisis está mucho más dirigido y debe de haberse fijado de antemano el método a emplear. Existen varios métodos estadísticos básicos y que pueden ser usados en múltiples contextos.

Una cuestión que no debe pasarse por alto en este análisis es la **validación por parte de expertos.** Si estamos analizando la evolución del número de matriculaciones de nuestro centro, una vez tengamos los resultados debemos consultar con la persona responsable para que no solo nos confirme si existen errores en el análisis y sus resultados, sino que nos ofrezca una lectura valiosa de los mismos.

Método	Descripción
Frecuencias	Recuento de veces que aparece un determinado valor en un conjunto de datos.
Porcentajes	Proporción de una cantidad con respecto al total
Promedio	Valor típico o representativo de un conjunto de datos
Correlación	Medida que refleja la relación entre dos variables y cómo de fuerte es esa relación

Ejemplo

Tras realizar un seminario con los métodos de análisis de datos de Google Sheets y poner en común aspectos que han llamado la atención en los datos. El equipo ha llevado a cabo el análisis exploratorio, calculando el total del estudiantado y cuáles son las asignaturas con más aprobados y suspensos, y el análisis principal, calculando el promedio de calificaciones para cada asignatura. Se verificó con el profesorado los valores promedio.

Visualización de Datos

6

Si tuvieses una tabla de datos con la evolución de matriculados durante 50 años, ¿qué harías para determinar si ese valor aumenta o decrece con el tiempo? Existen múltiples maneras para ello, pero lo que sin duda sería más rápido es construir un gráfico de columnas o líneas con dichos valores en el que **de un solo vistazo podamos detectar si hay alguna tendencia y cómo es esta.**

Gracias a la visualización de datos podemos realizar representaciones gráficas de los datos que hacen posible que la información sea rápida y fácilmente comprensible. Mediante la observación podemos detectar patrones o tendencias, como el aumento o descenso de matriculados con el paso del tiempo, valores anómalos, como asignaturas con una mayor cantidad de aprobados o suspensos que el resto.

En las sesiones 8 y 9 llevaremos a cabo las siguientes tareas:

1. Introducción de las visualizaciones de datos y su generación mediante un **seminario** en el que trabajaremos conjuntamente sobre dataset de prueba.

2. Elaboración de visualizaciones por parte del equipo supervisando dicho trabajo.

Jess Bailey. Licencia: Unsplash

Elaboración de visualizaciones

La visualización de datos es un **instrumento indispensable en la ciencia de datos** y cuya elaboración se realiza al mismo tiempo que el análisis, pudiendo ser el principal resultado incluso una visualización. Sin embargo, para que las visualizaciones tengan este poder y resulten reveladoras, es necesaria su correcta elaboración. En el terreno de las visualizaciones entran en juego numerosos elementos que no han de descuidarse para conseguir que aquello que estamos representando sea comprensible con solo una mirada. La correcta elección del tipo de gráfico, el uso de colores distinguibles o evitar el exceso de información son elementos clave para facilitar dicha lectura. En todo momento debemos tratar de hacer que la visualización sea auto explicativa apoyándonos en todos los elementos que tenemos a nuestro alcance al construirla, anteponiendo siempre la sencillez.

Algunos de los gráficos más utilizados son los siguientes, siendo cada uno de ellos idóneo para casos distintos.

Gráfico	Descripción	Caso útil
Columnas 	Representación de valores de datos mediante barras verticales.	Práctico para comparar los valores de una misma variable en diferentes casos o individuos.
Barras 	Representación de valores de datos mediante barras horizontales.	Práctico para comparar los valores de una misma variable en diferentes casos o individuos.
Líneas 	Representación de valores mediante una línea que conecta diferentes puntos.	Útil para representar la evolución de una variable a lo largo del tiempo.
Área 	Representación de valores mediante una línea que conecta diferentes puntos estando relleno el espacio entre esta y el eje horizontal.	Útil para representar la evolución de una variable a lo largo del tiempo, permitiendo remarcar la magnitud o volumen de los datos.
Dispersión 	Representación de puntos en un eje de coordenadas, en el que la posición en X representa el valor de una variable y su posición en Y el de otra variable.	Útil para representar la relación entre dos variables.

Ejemplo

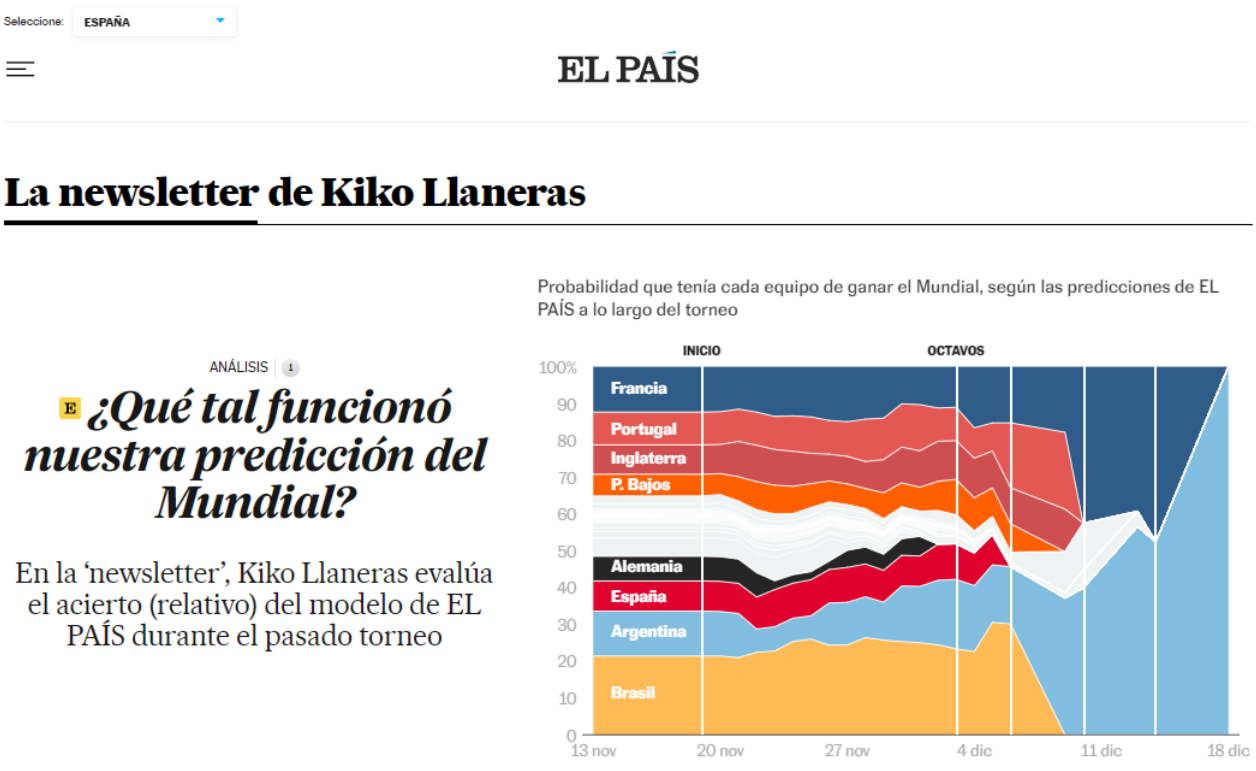
Tras dedicar una sesión al seminario de visualizaciones el equipo elaboró un gráfico de líneas en el que cada punto es una asignatura y la altura de este se corresponde con la nota promedio.

Redacción de los resultados

Una vez finalizado el análisis es crucial **redactar y presentar los resultados de manera comprensible.** Se trata de un texto que debe ajustarse al formato de una nota de prensa y que por lo general va a ser leído por personas que son completas desconocedoras de los datos analizados.

En la sesión 10 nos encargaremos de la:

1. Introducción y consejos para la redacción de los resultados.
2. Elaboración de una nota con los resultados por parte del equipo coordinando su desarrollo.
3. Elección de un método para su difusión.



Ejemplo de un artículo de datos
¿Qué tal funcionó nuestra predicción del Mundial? | La newsletter de Kiko Llaneras | EL PAÍS (elpais.com)

Redacción de la nota

Una parte imprescindible en el texto es la introducción del proyecto. **Cada proyecto supone una historia única** y, por ello, necesita una presentación adecuada. Explicar el contexto y la razón detrás del proyecto no solo ayuda a comprender el escenario de nuestra historia, sino que también prepara a las personas que van a leerla a comprender el significado y la relevancia de los datos que presentaremos.

Con nuestras visualizaciones en mano, **redactaremos descripciones claras y concisas que acompañen a cada gráfico.** Estas descripciones sirven como una guía, ayudando a quienes leen el informe a navegar por la información y entender qué es lo que están viendo. En todo momento, nos esforzaremos por mantener un equilibrio: queremos ser lo suficientemente detallados para que la información sea útil, pero al mismo tiempo, lo suficientemente breves y directos para no abrumar o confundir. Terminamos con una sección de conclusiones que refuerza los puntos clave y de una interpretación final. Debemos considerar siempre un párrafo independiente sobre cómo se ha llevado a cabo el análisis.

Para darle difusión a la nota de prensa podemos hacer uso de los recursos que tengamos a mano:

- Si nuestro data lab o institución tiene una **página web**, puede ser un escaparate para compartir la nota.
- Las **redes sociales** son una buena alternativa y también pueden servir para ello.
- En caso de no disponer de ninguna de estas opciones, podríamos plantear la creación de una web o abrir un perfil en una red social, específicamente para el data lab.

Eso sí, e**n esta sesión solo redactaremos la nota y fijaremos el medio en donde la distribuiremos,** pero será en la sesión final cuando la publiquemos junto a los datos.

Compartir los resultados

Compartir datos es esencial. Al compartir datos, no solo democratizamos el acceso a la información, sino que también favorecemos su reutilización. En nuestro caso no tenemos necesariamente que compartir los datos en bruto que analizamos sino el conjunto final que hemos procesado y empleado para el análisis. Es decir, si estamos analizando la evolución de las personas matriculadas, no compartiremos el listado de personas que fue nuestro punto de partida, el cual además incumpliría la normativa de protección de datos, sino la tabla de frecuencias de personas matriculadas por año que obtuvimos para el análisis. Es además fundamental priorizar formatos abiertos al hacerlo.

Este será el objetivo de la sesión 11, la última del proyecto:

1. Revisión de los datos para garantizar que se encuentran en un formato abierto y libres de información de carácter personal o privada.

2. Publicación de los datos en abierto junto a la nota de resultados.

Hay diversas formas efectivas y seguras de compartir datos. Desde plataformas de almacenamiento de datos en la nube como Google Drive podemos compartir un enlace a los datos empleados. Siempre limitando los permisos para que las personas los descarguen libremente pero no puedan modificar ese fichero concreto que tenemos en la nube. De esta manera podremos incluir dicho enlace en la web de nuestra institución y/o nota de prensa para darle una mayor visibilidad y formalidad, especialmente si se trata de datos institucionales.

Finalmente, es necesario asignar licencias adecuadas a los datos compartidos. Las licencias, como las ofrecidas por **Creative Commons,** proporcionan un marco claro sobre cómo se pueden usar, redistribuir y modificar los datos. Esto protege los derechos del creador original mientras facilita el uso responsable de la información por parte de terceros. Su uso es muy sencillo, pues con seleccionar una licencia concreta e indicarla junto a los datos es suficiente para que quede claro las normas de uso de los mismos.

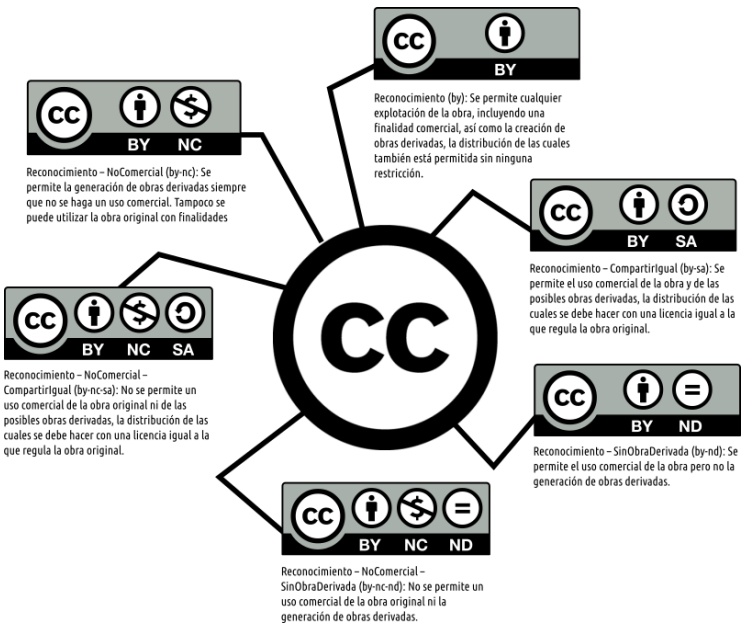


Imagen de Oriana Robles Muñoz https://es.wikipedia.org/wiki/Archivo:Las_seis_licencias_Creative_Commons.png


4 Limpieza de datos
5 Análisis de datos
3 Recopilación de datos
6 Visualización de datos
2 Creación de un Data Lab
7 Redacción de los resultados
8 Compartir los resultados
1 La ciencia de datos

RESUMEN

CONSEJOS

1. **Establece claridad en los objetivos:** Antes de sumergirnos en los datos, debemos asegurarnos de tener claridad sobre lo que estamos tratando de lograr. Un objetivo bien definido guiará nuestras decisiones y nos ayudará a mantener el foco durante todo el proyecto.

2. **Fija un calendario para cada tarea y fase del proyecto:** La gestión del tiempo es esencial para garantizar que el proyecto avance de manera eficiente. Debemos establecer fechas límite realistas y mantenernos al tanto de ellas. Esto no solo ayuda a mantener la estructura, sino que también permite evaluar el progreso y realizar ajustes según sea necesario.

3. **Organiza reuniones periódicas:** Estas reuniones sirven para revisar el avance, abordar problemas y asegurarte de que todos los miembros del equipo estén alineados con los objetivos y las expectativas. La frecuencia dependerá del proyecto, pero es vital que exista un espacio regular para la discusión y el feedback.

4. **Anota todas las decisiones y aspectos que te llamen la atención:** Durante el proceso de limpieza y análisis, encontraremos multitud de hallazgos, tomaremos decisiones y realizaremos observaciones. Anotar estos aspectos nos permitirá tener un registro y facilitará la revisión y el seguimiento a medida que avanza el proyecto.

5. **Ten en cuenta que no siempre encontrarás unos datos a medida:** En el mundo real, es raro encontrar conjuntos de datos que sean perfectos para nuestras necesidades desde el principio. Es esencial ser flexible y estar preparado para adaptar tu enfoque o reconsiderar las hipótesis según los datos disponibles.

6. **Dedica muchos esfuerzos a la limpieza de datos, pero establece límites:** La limpieza y preparación de datos son fundamentales en cualquier proyecto de ciencia de datos. Sin embargo, es importante encontrar un equilibrio. Si dedicamos demasiado tiempo buscando la perfección, podemos retrasar otras fases críticas del proyecto.

7. **Mantén una comunicación entre las distintas fases:** Tenemos que asegurarnos de que haya una comunicación fluida entre las diferentes etapas del proyecto. Por ejemplo, lo que aprendamos en la fase de análisis exploratorio puede ser crucial para las etapas de análisis o redacción. Una comunicación eficiente asegura que todas las fases estén alineadas y se beneficien mutuamente.

8. **Solicita feedback externo:** A veces, una perspectiva externa puede ofrecer comentarios e interpretaciones valiosas o identificar puntos ciegos en nuestro análisis.

RECURSOS

- **Crear tu primer formulario en formularios de google—Centro de aprendizaje de google workspace** https://support.google.com/a/users/answer/9303071?hl=es
- **Guía para el Ciudadano (aepd.es)** https://www.aepd.es/sites/default/files/2020-05/guia-ciudadano.pdf
- **Cómo limpiar sus datos** https://data.europa.eu/elearning/es/module11/#/id/co-01
- **Las diez formas principales de limpiar los datos** https://support.microsoft.com/es-es/office/las-diez-formas-principales-de-limpiar-los-datos-2844b620-677c-47a7-ac3e-c2e157d1db19
- **Importar conjuntos de datos y hojas de cálculo** https://support.google.com/docs/answer/40608?sjid=6116724264195240952-EU
- **Analizar datos - Ayuda de Editores de Documentos de Google** https://support.google.com/docs/answer/9330962?hl=es
- **Lista de funciones de Hojas de cálculo de Google - Ayuda de Editores de Documentos de Google** https://support.google.com/docs/table/25273?hl=es&ref_topic=3105600&sjid=5432195421027101055-EU
- **Google Sheets - Google Cloud Skills Boost** https://www.cloudskillsboost.google/course_templates/196
- **Aprendizaje (tableau.com)** https://www.tableau.com/es-es/learn
- **Cómo elegir el gráfico adecuado para tus datos| Biuwer Analytics** https://biuwer.com/es/blog/como-elegir-el-grafico-adecuado-para-tus-datos/

AWenceslao Arroyo Machado (2024). Cómo hacer un DataLab.

Madrid: Instituto Nacional de Tecnologías Educativas y de Formación del Profesorado (INTEF).

la aventura de aprender