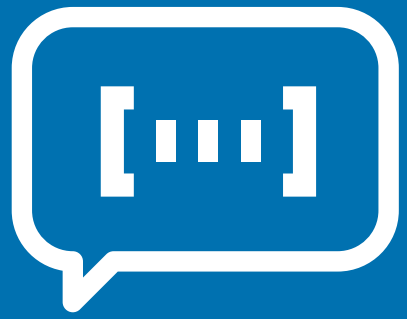
netmind
a bts company

Data Science Basics

Data Quality



DEBATE

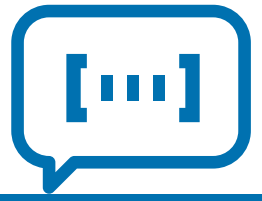
¿Cómo podemos evaluar la calidad de los datos a escala?

OBJETIVO

Obtener visión sobre la calidad de los datos y su perfilado.

INSTRUCCIONES

1. Piensa en los datos que se obtienen de los sensores de las máquinas de una instalación.
2. **Únete** a tu compañero para responder estas **preguntas**:
   - **¿Qué calidad (consideras) presentan esos datos?**
   - **¿Qué aspectos pueden caracterizar esa calidad?**
   - **¿Cómo puede influir esa calidad en los procesos de Machine Learning?**
   - **¿Cómo podemos entender y mejorar esa calidad?**
3. Usa post-its para **capturar** vuestras ideas. Sintetiza 1 post-it para cada pregunta.
4. **Prepárate** para compartir vuestra visión con el resto de la clase.



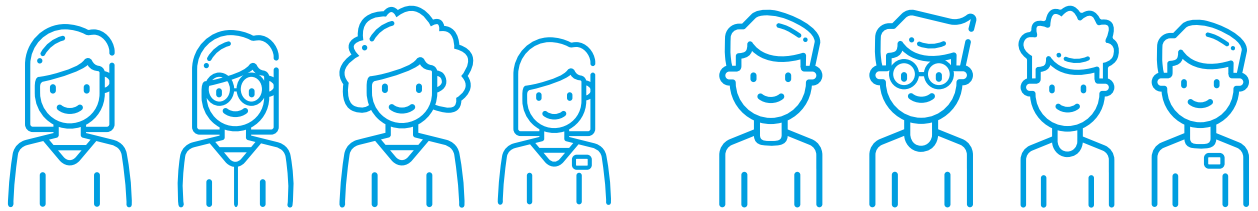
20 min

OBJETIVO

Comparte tu visión

INSTRUCCIONES

1. **Comparte** tu visión con el resto de la clase.
2. Genera unas **conclusiones comunes**.



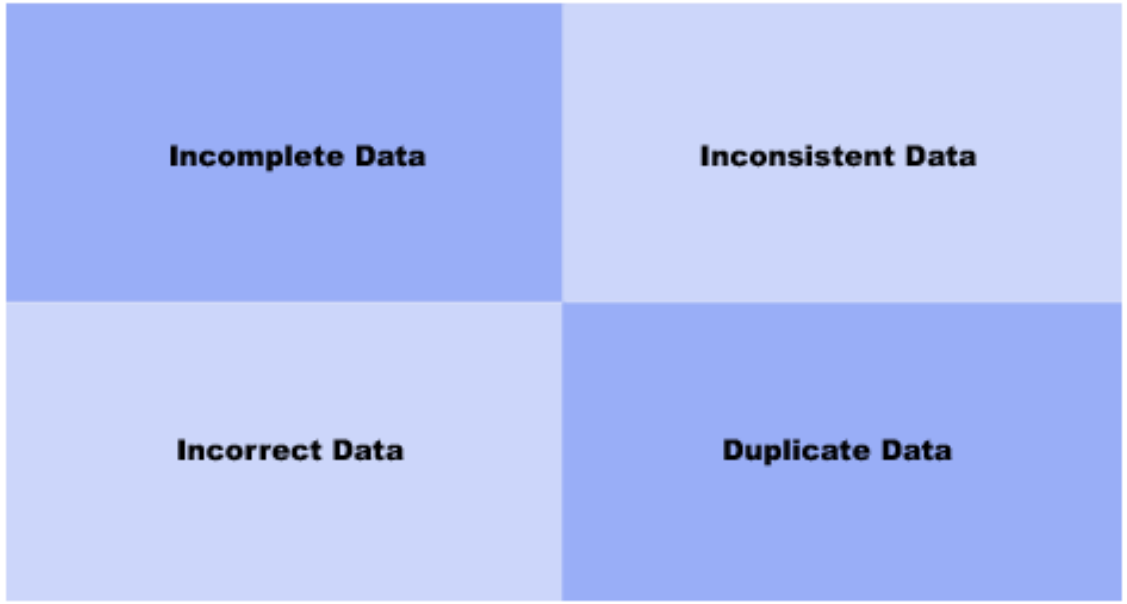
10 min

01

Calidad del dato

Hoy estamos en un mundo de heterogeneidad.
Disponemos de diferentes tecnologías.
Operamos en diferentes plataformas.
Tenemos una gran cantidad de datos que se generan todos los días en todo tipo de organizaciones y empresas.
...
Y tenemos **problemas con los datos**.

Problemas con los datos



Duplicado, inconsistente, ambiguo, incompleto.

Por lo tanto, es necesario **recopilar datos en un solo lugar y limpiarlos**.

¿Por qué es importante la calidad de los datos?

- Los **buenos datos** son **el activo más valioso.**
- Los **malos datos** pueden **dañar seriamente** el negocio y la credibilidad…

Las organizaciones necesitan responder a preguntas como:

- ❑ ¿Qué se nos ha pasado?
- ❑ ¿Cuándo las cosas van mal?
- ❑ Tomar decisiones seguras.

¿Qué es la calidad de los datos?

Source: profisee.com

La calidad de los datos es una percepción o una evaluación de la idoneidad de los datos para cumplir su propósito en un contexto dado.

Se describe por varias dimensiones como:

- **Corrección/ Precisión**: La precisión de los datos es el grado en que los datos capturados describen correctamente la entidad del mundo real.
- **Consistencia**: Se trata de la versión única de la verdad. Coherencia significa que los datos de toda la empresa deben sincronizarse entre sí.
- **Completitud**: Es la medida en que se proporcionan los atributos esperados de los datos.
- **Oportunidad**: Los datos correctos para la persona correcta en el momento correcto son importantes para el negocio.
- **Metadatos**: Datos sobre datos.

Mantenimiento de la calidad de los datos

La calidad de los datos es el resultado del proceso de revisar los datos y depurarlos, estandarizarlos y eliminar la duplicación de registros, además de hacer parte del enriquecimiento de los datos.

1. Mantener datos completos.
2. Limpiar datos estandarizándolos mediante reglas.
3. Usar algoritmos sofisticados para detectar duplicados.
4. Evitar la entrada duplicada de clientes potenciales y contactos.
5. Combinar registros duplicados.
6. Usar roles para la seguridad.

Datos inconsistentes antes de limpiar

Invoice 1

Bill no	CustomerName	SocialSecurityNumber
101	Mr. Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	CustomerName	SocialSecurityNumber
205	Mr. S Aleck	ADWPS10017

Invoice 3

Bill no	CustomerName	SocialSecurityNumber
314	Mr. Stevenson Aleck	ADWPS10017

Invoice 4

Bill no	CustomerName	SocialSecurityNumber
316	Mr. Alec Stevenson	ADWPS10017

Datos consistentes después de la limpieza

Invoice 1

Bill no	CustomerName	SocialSecurityNumber
101	**Mr.** Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	CustomerName	SocialSecurityNumber
20S	**Mr.** Aleck Stevenson	ADWPS10017

Invoice 3

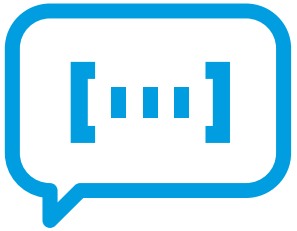
Bill no	CustomerName	SocialSecurityNumber
314	Mr. Aleck Stevenson	ADWPS10017

Invoice 4

Bill no	CustomerName	SocialSecurityNumber
316	Mr. Aleck Stevenson	ADWPS10017

netmind
a bts company

02



Limpieza de datos

Comprender los datos sucios

Hay muchos tipos de errores e incoherencias que pueden contribuir a que los datos estén sucios.

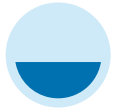
Valores perdidos

Son muy frecuentes.

Puede que a tu conjunto de datos le falten varios años de datos, que sólo tenga alguna información sobre un cliente, o que no contenga toda la gama de productos de tu empresa.

Los valores perdidos pueden tener múltiples efectos en tu análisis.

La falta de grandes porciones de datos cruciales puede causar **sesgos en tus resultados**.



Valores atípicos

Valores que se salen de la norma y no son representativos de los datos.

Pueden deberse a un error tipográfico o a circunstancias excepcionales.

Importante diferenciar los verdaderos valores atípicos de las situaciones extremas informativas.

Pueden sesgar tus resultados, sugiriendo en última instancia una respuesta errónea.

Duplicados

Pueden representar en exceso una entrada en tu análisis, lo que te llevaría a una **conclusión errónea**.

Ten cuidado con los duplicados que aparecen más de una vez y con los que contienen información contradictoria o actualizada.

Datos erróneos

Puedes tener mal escrito el nombre de un cliente, un número de producto incorrecto, información obsoleta o datos mal etiquetados.

A veces puede ser difícil determinar si tus datos son erróneos,

¡Importante verificar tu fuente!

Tu análisis es tan bueno como tus datos.

Inconsistencias

Las incoherencias se presentan de muchas formas.

Puede haber entradas de datos incoherentes, lo que puede indicar una errata o un error.

- Edad de cliente atrasada, un ingrediente que cambia de número de identificación o un producto con dos precios simultáneos, merece la pena que eches un vistazo más de cerca para asegurarte de que todo es correcto.

Otro tipo problemático de incoherencia es la incoherencia en el formato de los datos.

- Los distintos valores pueden comunicarse en unidades diferentes (kilómetros vs. millas vs. pulgadas), en estilos diferentes (Mes Día, Año vs. pulgadas). Día-Mes-Año), en distintos tipos de datos (flotantes frente a enteros), o incluso en distintos tipos de archivo (.jpg frente a .png).

Estas incoherencias harán difícil, si no imposible, que tu código interprete los valores correctamente.

Esto puede dar lugar a un análisis incorrecto o a que tu código no se ejecute en absoluto.

Exploración y preprocesamiento de datos

Es crucial que comprendas tu conjunto de datos antes de utilizarlo en un análisis complejo. Para desarrollar este tipo de comprensión de tus datos, debes realizar una exploración de los mismos.

1. Ante todo, **examina la fuente** de tu conjunto de datos y determina si tiene algún sesgo o agenda que pueda afectar a la calidad o fiabilidad de tus datos.
2. Conoce el **contexto de tus datos** y cualquier otro factor que pueda haber afectado a tus datos y que no esté contabilizado internamente.
3. Determina **cuántas variables diferentes tienes**. En un conjunto de datos con formato de tabla, las variables suelen ser las columnas, mientras que cada entrada de datos es una fila.
4. Determina **cuántas categorías diferentes** tienes dentro de cada variable. Por ejemplo, si una de tus variables es Tipo de fruta, deberías tener una buena idea de cuántos tipos diferentes de fruta están representados en esta variable.
5. Observa las **estadísticas resumidas** de cada columna, incluidas la media, la mediana, la varianza y la desviación típica.
6. Si es posible, **traza cada variable y tantos pares de variables como puedas** y observa realmente los gráficos. Busca cualquier desviación, valor atípico, tendencia o correlación que merezca la pena investigar más a fondo.
7. Si tu conjunto de datos forma parte de una base de datos relacional, **fíjate bien en las relaciones** y asegúrate de que entiendes cómo se relacionan entre sí las distintas tablas.
8. Si procede, utiliza la **función de perfiles de la librería pandas** para generar un informe de perfil. Esto te proporcionará información valiosa sobre tu conjunto de datos.

Importancia de la exploración y el preprocesamiento

Todo esto puede parecer superfluo para tu análisis previsto, pero hay algunas razones importantes por las que siempre debes seguir estos pasos.

1. Gran **comprensión de los límites** de tu conjunto de datos, lo cual es esencial si quieres confiar en los resultados de tus análisis finales.

2. Puede indicarte **tendencias y análisis importantes** que no habías considerado antes. Éstos tienen el potencial de añadirse a tus análisis previstos o de presentar un factor de complicación que debes tener en cuenta.

3. Es tu **primera pista** sobre dónde puedes tener datos sucios. Puede ser la primera vez que veas ese valor atípico, que te des cuenta de que hay el doble de categorías de las que debería haber, o que descubras que el método de recogida de datos era distinto el año pasado que el anterior. Todos estos son datos fundamentales que deberían animarte a sentir curiosidad por tu conjunto de datos.

Buenas prácticas y consejos

- Almacena los datos brutos por separado.
- Documenta tu código de limpieza.
- ¡Cuidado con las consecuencias imprevistas!
- Mantén un registro de limpieza de datos.
- Escribe funciones reutilizables.

03

Perfilado de datos

Contexto

Source: astera.com

En el proceso de diseño de un almacén de datos nos enfrentamos a situaciones como:

- **Varias inconsistencias** de datos en la fuente, como registros faltantes o valores NULL.
- La columna que se eligió como clave principal **no es única** en toda la tabla.
- el diseño del **esquema no es coherente** con los requisitos del usuario final.
- Cualquier otro **concern** con los datos, que debe haberse solucionado desde el principio.

Solucionar los problemas de calidad de datos significaría realizar cambios en los paquetes de flujo de datos ETL, limpiar las inconsistencias identificadas, etc.

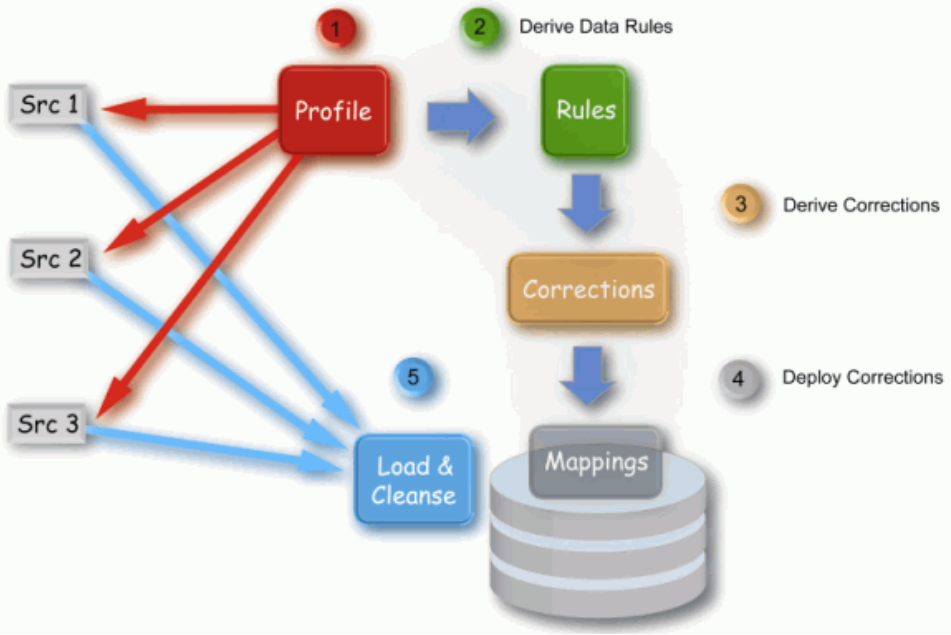
Esto a su vez dará lugar a una gran cantidad de trabajo por hacer.

El re-trabajo significará costos adicionales para el empresa, tanto en términos de tiempo, como de esfuerzo.

...

Entonces, **¿qué hacer en estos casos?**

Solución



Source: oracle.com

- En lugar de tener que solucionar el problema, sería mejor detectarlo **desde el principio** antes de que se convierta en un problema.
- Después de todo, "**MEJOR PREVENIR QUE CURAR**".
- ¡Necesitamos la **creación de perfiles de datos!**

¿Qué es el perfilado de datos?

Es el proceso de **examinar y analizar estadísticamente** el contenido de una fuente de datos y, por lo tanto, recopilar información sobre los datos.

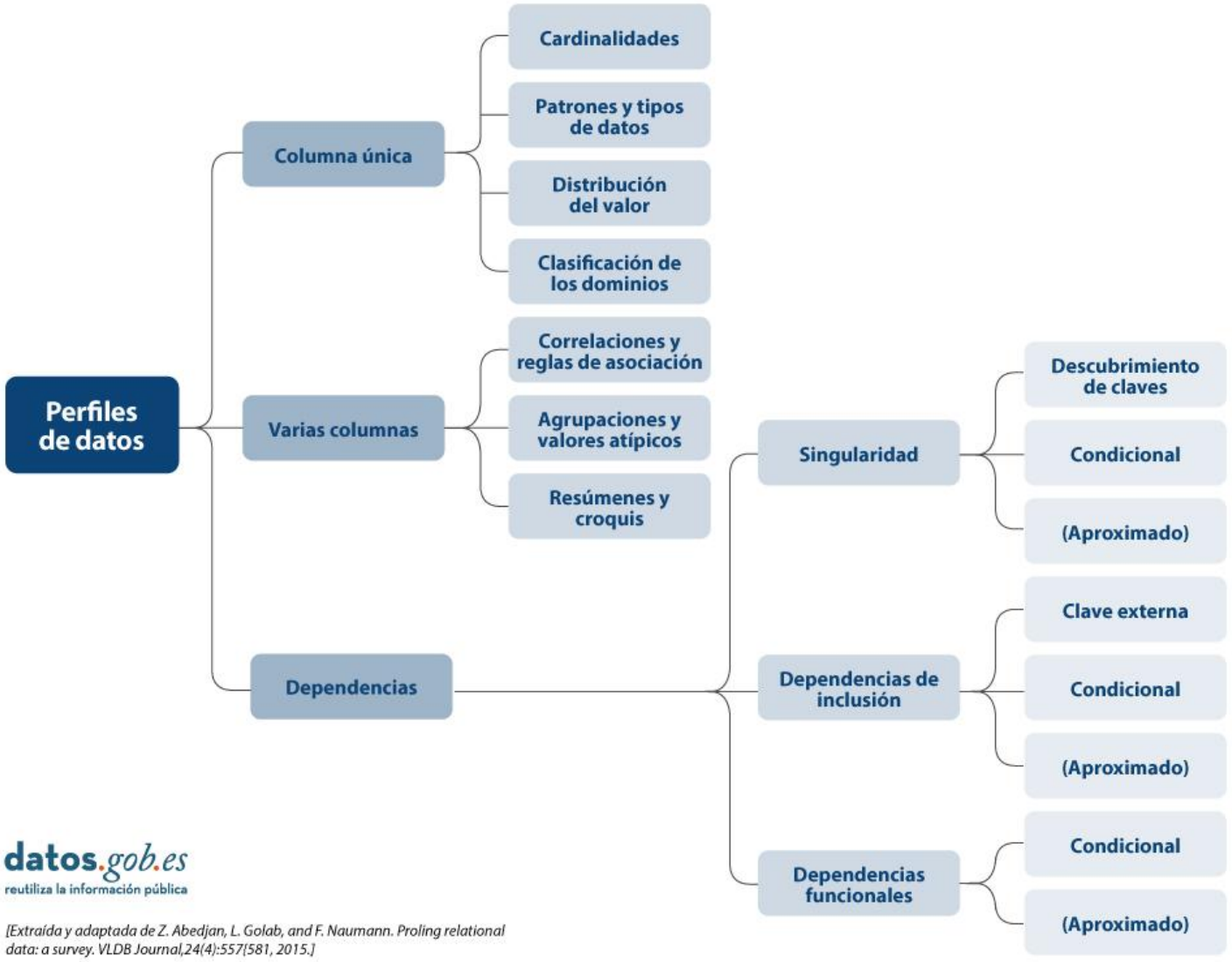
Consiste en **técnicas** utilizadas para analizar los datos que tenemos, en cuanto a precisión y exhaustividad. Es una técnica indispensable durante el análisis exploratorio de datos (**EDA**).

Su objetivo final es proporcionar un **entendimiento claro y detallado** de la **estructura**, **contenido** y **calidad** de los datos, lo que es esencial antes de su uso en cualquier aplicación.

Perfilado de datos

1. La elaboración de perfiles de datos nos ayuda a realizar una **evaluación exhaustiva** de la calidad de los datos.
2. Ayuda al **descubrimiento de anomalías** en los datos.
3. Nos ayuda a **comprender** el contenido, la estructura, las relaciones, etc. sobre los datos en la fuente de datos que estamos analizando.
4. Nos ayuda a saber si los datos existentes **se pueden aplicar** a otras áreas o propósitos.
5. Nos ayuda a comprender los diversos **problemas/desafíos** que podemos enfrentar en un proyecto de base de datos mucho antes de que comience el trabajo real. Esto nos permite tomar **decisiones tempranas** y actuar en consecuencia.
6. También se utiliza para **evaluar y validar** metadatos.

Tipos de perfilado de datos



[Extraída y adaptada de Z. Abedjan, L. Golab, and F. Naumann. Proling relational data: a survey. VLDB Journal,24(4):557{581, 2015.]

Ref: https://datos.gob.es/

Ejemplos comunes de análisis a realizar

- **Calidad de los datos:** analizar la calidad de los datos en la fuente de datos.
- **Valores NULL:** buscar la cantidad de valores NULL/NAN en un atributo.
- **Claves candidatas:** el análisis de la medida en que ciertas columnas son distintas brindará información útil al desarrollador para la selección de claves candidatas.
- **Selección de clave principal:** para verificar si la columna de clave candidata no viola los requisitos básicos de no tener valores NULL o valores duplicados.
- **Valores de cadena vacíos:** una columna de cadena puede contener NULL o incluso valores de cadena vacíos que pueden crear problemas más adelante.
- **Longitud de cadena:** un análisis de la longitud más grande y más corta posible, así como la longitud de cadena promedio de una columna de tipo cadena, puede ayudarnos a decidir qué tipo de datos sería el más adecuado para dicha columna.
- **Identificación de cardinalidad:** las relaciones de cardinalidad permiten observar si una cantidad grande de registros tiene el mismo valor..
- **Formato de datos:** A veces, el formato en el que se escriben ciertos datos en algunas columnas puede o no ser fácil de usar.

Herramientas

Ofrecen informes detallados y visualizaciones avanzadas en Python:

- **Pandas Profiling**
  - https://pypi.org/project/pandas-profiling/
- **YData Profiling**
  - https://pypi.org/project/ydata-profiling/

ACTIVIDAD

Evaluación de calidad de datos

OBJETIVO

Evaluar la calidad de datos de las ventas de productos

INSTRUCCIONES

Se quiere hacer una evaluación de calidad de datos sobre las ventas (sales) y pagos (payments). Para ello se requiere hacer un análisis de los siguientes puntos:

- Calidad de los datos
- Selección de clave principal
- Identificación de cardinalidad
- Obtener media, varianza y desviacion Estandar, covarianza, correlacion
- Mejorar la calidad.

"Estadística Descriptiva con Python y Pandas": https://coderhook.github.io/Descriptive%20Statistics

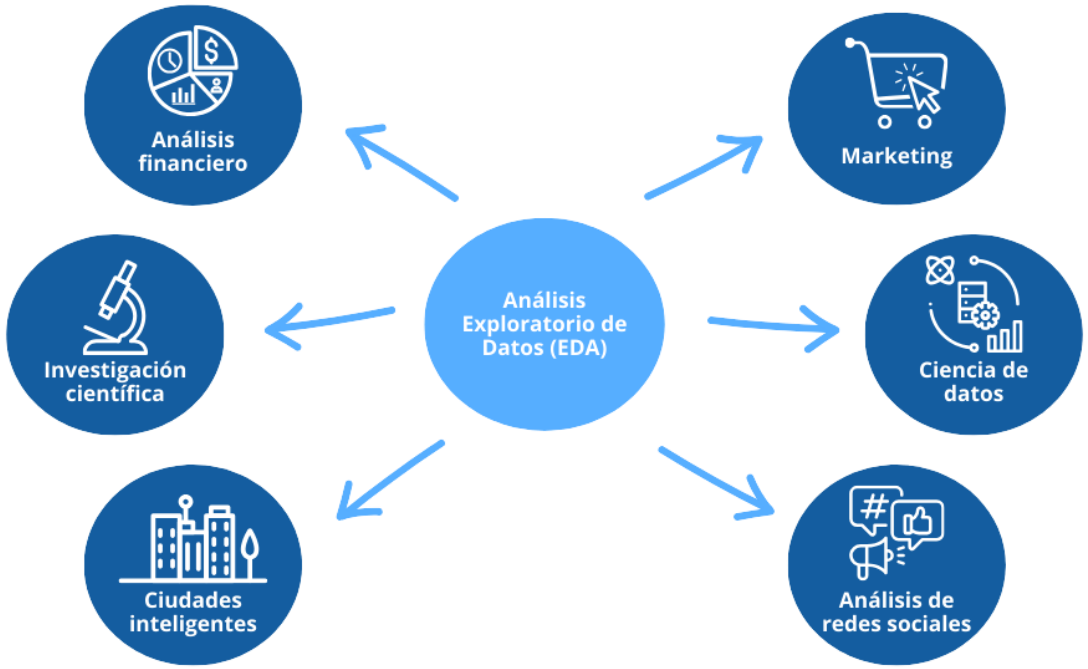
Columnas sales:

- orderNumber
- orderLineNumber
- orderDate
- shippedDate
- requiredDate
- customerNumber
- employeeNumber
- productCode
- status
- comments
- quantityOrdered
- priceEach
- sales_amount
- origin

Columnas payments:

- customerNumber
- checkNumber
- paymentDate
- amount

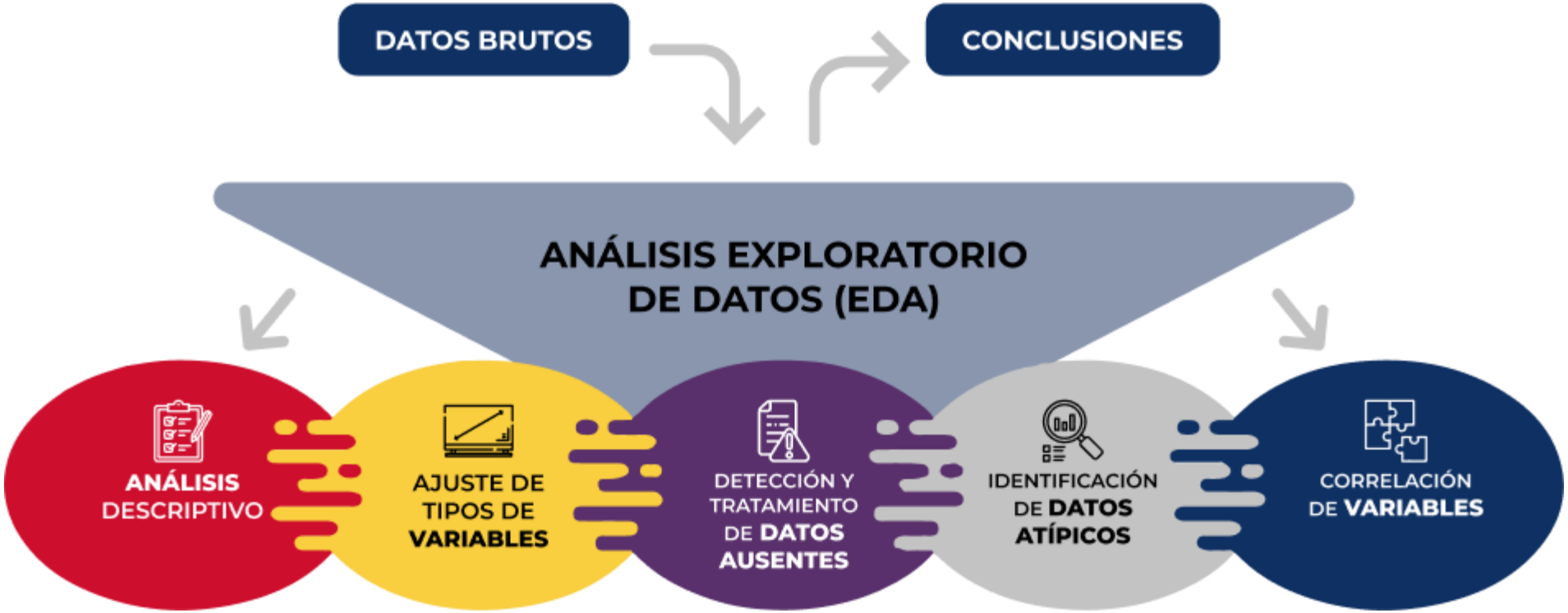
Análisis Exploratorio de Datos - EDA



El AED consiste en aplicar un conjunto de técnicas estadísticas dirigidas a **explorar, describir y resumir la naturaleza de los datos**, de tal forma que podamos garantizar su **objetividad e interoperabilidad**.

Identificar posibles **errores**, revelar la presencia de **valores atípicos**, comprobar la **relación entre variables** (correlaciones) y su posible redundancia, así como realizar un **análisis descriptivo** de los datos mediante representaciones gráficas y resúmenes de los aspectos más significativos.

Análisis Exploratorio de Datos - EDA



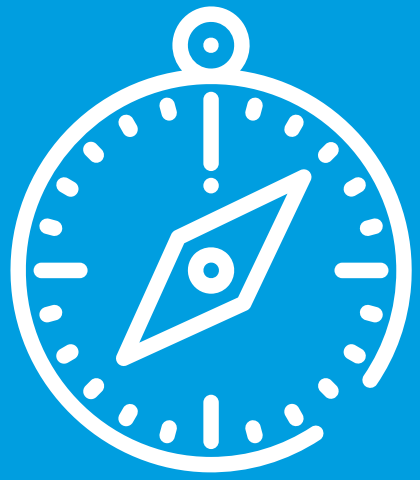
Ref: https://datos.gob.es/

Objetivos

- Entender la estructura de los datos
- Identificar patrones y relaciones
- Detectar valores atípicos
- Explorar distribuciones y resúmenes estadísticos
- Guiar la toma de decisiones en la limpieza de datos
- Facilitar la selección de modelos y enfoques analíticos
- Generar hipótesis
- Comunicar resultados de manera efectiva

Proceso

- **Recopilación de datos:** Antes de realizar el EDA, es necesario tener acceso a los datos que se van a analizar. Esto podría involucrar la recolección de datos, la importación de conjuntos de datos existentes o la conexión a bases de datos.
- **Exploración inicial:** Al comienzo del EDA, se realiza una exploración inicial para obtener una comprensión básica del conjunto de datos. Esto puede incluir la revisión de la estructura de los datos, el tipo de variables presentes y la identificación de posibles problemas, como valores faltantes o inconsistentes.
- **Visualización de datos:** Se utilizan diversas herramientas gráficas, como histogramas, diagramas de dispersión, diagramas de caja y gráficos de barras, para visualizar la distribución de variables y explorar relaciones entre ellas. Estas visualizaciones proporcionan una perspectiva intuitiva de los datos.
- **Estadísticas descriptivas:** Se calculan medidas estadísticas descriptivas, como la media, la mediana, la desviación estándar y cuartiles, para resumir las características numéricas de las variables. Estas estadísticas proporcionan una descripción cuantitativa de la tendencia central y la dispersión de los datos.
- **Identificación de valores atípicos:** Se buscan y analizan valores atípicos o extremos que podrían indicar errores en la recopilación de datos o revelar patrones interesantes. Técnicas como los diagramas de caja y los gráficos de dispersión pueden ser útiles en este contexto.
- **Análisis de relaciones:** Se exploran las relaciones entre variables, utilizando herramientas como matrices de dispersión o mapas de calor de correlación. Esto ayuda a comprender la interacción entre diferentes aspectos de los datos.
- **Generación de hipótesis:** A medida que se exploran los datos, pueden surgir hipótesis sobre patrones o tendencias interesantes que podrían ser investigadas más a fondo en etapas posteriores del análisis.
- **Iteración y refinamiento:** El proceso de EDA es iterativo. A medida que se descubren más aspectos de los datos, se pueden realizar ajustes en la exploración y se pueden formular nuevas preguntas para guiar el análisis continuo.



Next steps

¡Muchas gracias!

Nos vemos en: