Calidad de los Datos: Fundamentos y Aplicaciones

En la era de los datos, tener información no es suficiente: lo que realmente agrega valor es contar con datos confiables, relevantes y bien gestionados. La calidad de los datos no es solo una cuestión técnica; es un factor estratégico que afecta directamente la competitividad, la productividad y la capacidad de innovar de una organización.

Según un estudio de Gartner, las organizaciones pierden un promedio de 12.9 millones de dólares al año debido a problemas de calidad de datos. Estas pérdidas están asociadas a ineficiencias operativas, errores en la toma de decisiones y deterioro de la experiencia del cliente.

por Kibernum Capacitación S.A.

Preguntas de Activación

¿Qué riesgos podría enfrentar una empresa si no cuenta con una arquitectura clara de almacenamiento de datos (como Data Lake, Data Warehouse o Data Mart)?

¿Qué entiendes por "calidad de datos"? ¿La asocias más a un tema técnico, estratégico, o ambos? ¿Por qué?

¿Has trabajado o visto alguna vez una planilla, base de datos o formulario con errores? ¿Qué tipo de problemas observaste y qué consecuencias tuvo?

Objetivos y Principios de la Calidad de Datos

Disponibilidad

Los datos deben estar accesibles cuando se necesiten, permitiendo una toma de decisiones oportuna y eficiente.

Precisión

Representación correcta de la realidad, evitando distorsiones que puedan afectar análisis posteriores.

Relevancia

Adaptación a las necesidades específicas de los usuarios que consumen la información.

Completitud

Presencia de todos los elementos necesarios, sin omisiones críticas que puedan comprometer su utilidad.

Principios Rectores de la Calidad de Datos

Accuracy (Precisión)

Los datos deben reflejar fielmente la realidad. Un error tipográfico en el campo de salario o edad puede cambiar por completo un análisis.

Completeness (Completitud)

Todos los campos críticos deben estar presentes. Por ejemplo, una dirección sin código postal no permite realizar envíos correctos.

Consistency (Consistencia)

El dato debe mantener el mismo valor en todas sus representaciones. Un cliente no puede tener dos nombres distintos en sistemas integrados.

Integrity (Integridad)

El dato debe mantener relaciones correctas dentro de la base. Ejemplo: no puede haber una venta sin cliente asignado.

Más Principios Fundamentales

Reasonableness (Razonabilidad)

Los datos deben tener sentido en su contexto. No es razonable que un cliente tenga 150 años o que un producto pese 0 kg.

Timeliness (Actualización)

El dato debe estar vigente y reflejar los cambios recientes. Una fecha de renovación de contrato desactualizada puede llevar a sanciones.

Uniqueness (Unicidad)

Cada entidad debe estar registrada una única vez. Duplicados afectan reportes y generan redundancia en campañas o costos.

Accessibility (Accesibilidad)

Los datos deben estar disponibles para quienes los necesitan, cuando los necesitan, según permisos definidos por gobernanza.

La Analogía de los Ingredientes

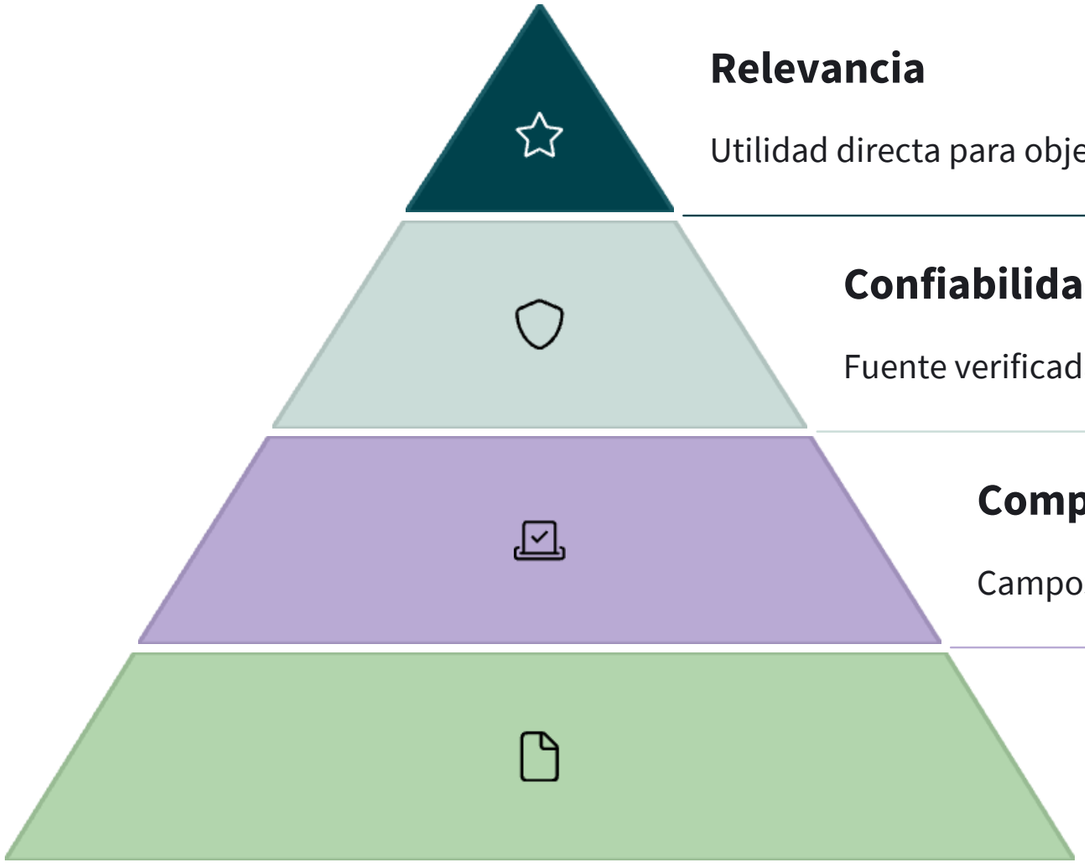
Piensa en los datos como los ingredientes de una receta. Aunque tengas una receta muy bien escrita (los procesos de negocio), si los ingredientes están en mal estado (los datos), el resultado final será un fracaso, sin importar la calidad de la receta.

Esta analogía nos ayuda a comprender que incluso los mejores procesos y sistemas fallarán si la materia prima informacional no cumple con estándares mínimos de calidad.

Al igual que un chef no puede preparar un plato excelente con ingredientes en mal estado, una organización no puede tomar decisiones acertadas con datos deficientes.

Conceptos Esenciales de la Calidad de Datos

En la gestión moderna de datos, es fundamental comprender que la calidad no es una propiedad universal, sino contextual. Esto significa que un dato puede ser técnicamente correcto y, sin embargo, no aportar valor si no es pertinente para el análisis o la decisión que se requiere tomar.



Relevancia

Utilidad directa para objetivos específicos

Confiabilidad

Fuente verificada y trazabilidad

Completitud

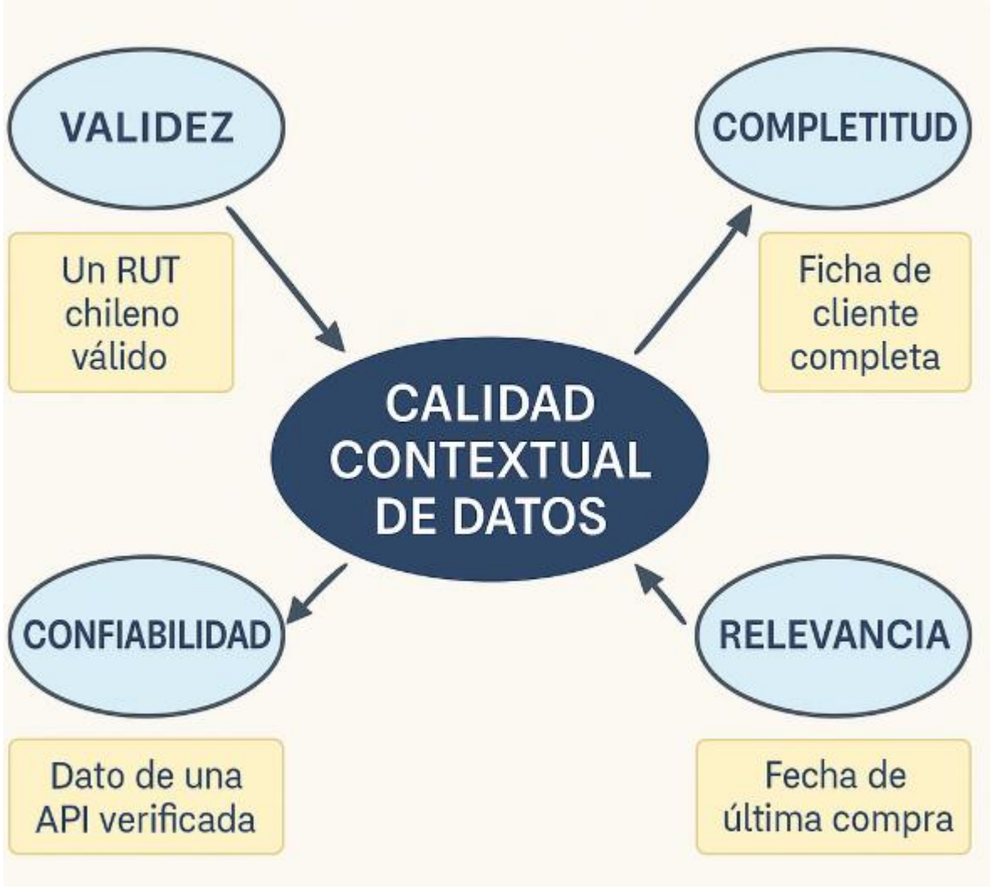
Campos necesarios presentes

Validez

Formato y estructura correctos

Ejemplos Prácticos de Conceptos Clave

Concepto	Definición clara	Ejemplo práctico
Dato válido	Cumple con el formato, tipo de dato y estructura definida para ese campo.	Un RUT chileno que contiene el dígito verificador correcto y estructura esperada.
Dato completo	Todos los campos necesarios están presentes y no vacíos.	Una ficha de cliente con nombre, dirección, comuna y correo electrónico.
Dato confiable	Proviene de una fuente verificada, autorizada o con trazabilidad de origen.	Ingresado desde un sistema de autenticación o proveniente de una API validada.
Dato relevante	Tiene utilidad directa para los objetivos del análisis, operación o decisión.	El campo "fecha de última compra" es útil para segmentar promociones por inactividad.



Dimensiones de la Calidad de Datos

En el contexto de la ingeniería de datos, hablar de calidad no es suficiente sin establecer formas concretas de evaluarla, diagnosticarla y mejorarla. Es aquí donde entran en juego las dimensiones de la calidad de datos.

Precisión

Evalúa si el dato refleja con exactitud la realidad que representa.

Completitud

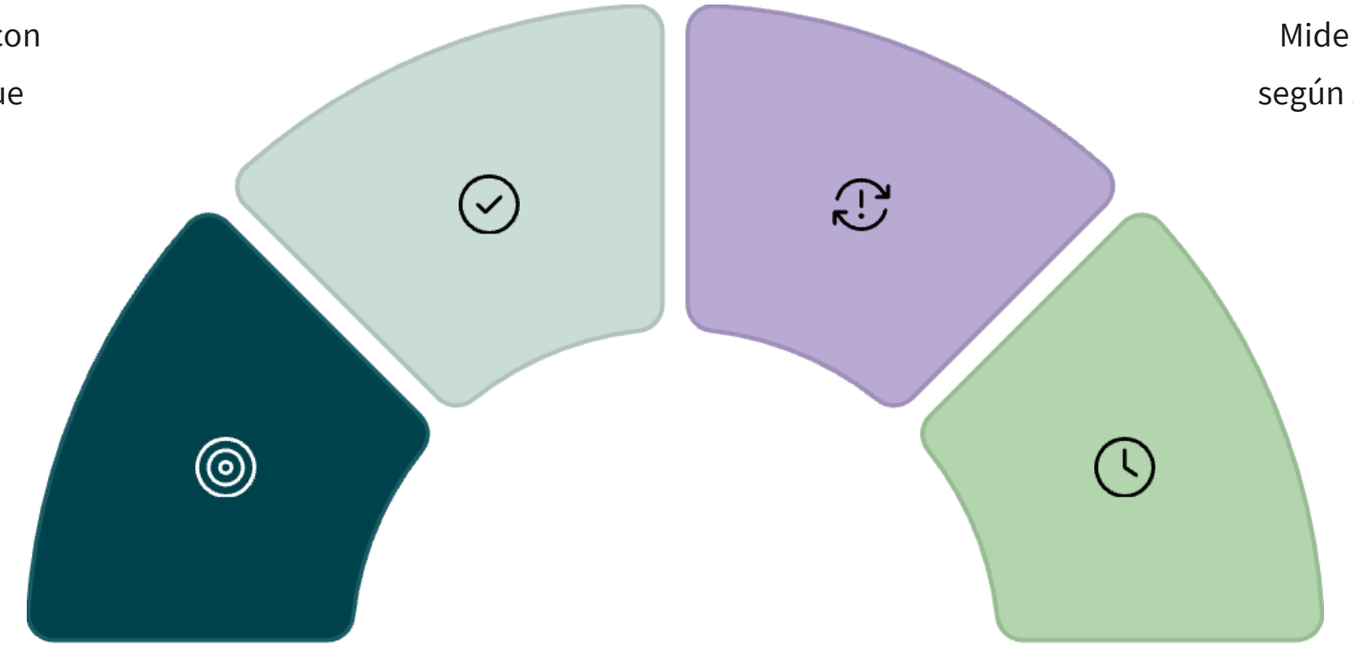
Verifica si todos los datos obligatorios están presentes y no están vacíos.

Consistencia

Examina si los datos mantienen coherencia entre diferentes fuentes o sistemas.

Actualización

Mide si los datos están vigentes según su propósito y fecha de uso.



Más Dimensiones de Calidad

Trazabilidad

Indica si es posible rastrear el origen, la modificación y el uso del dato. Por ejemplo, saber quién modificó un precio y cuándo se hizo.

Accesibilidad

Revisa si los usuarios autorizados pueden acceder al dato necesario cuando lo requieren. Un analista no puede acceder a las métricas clave por restricciones mal configuradas.

Utilidad para métricas

Cada dimensión puede ser convertida en una métrica cuantificable, lo que permite alimentar tableros de control, informes de calidad, y procesos de auditoría de datos.

Introducción Práctica: Explorando la Calidad de datos con Pandas

Ahora vamos a dar un paso hacia la práctica, utilizando ejemplos simples con la librería Pandas.

Para seguir este ejemplo necesitaras:
1. Descargar el archivo CSV
- Haz clic aquí para descargarlo:
- clientes_calidad_demo.csv
- Guárdalo en una carpeta de trabajo, por ejemplo:

C:\Users\TuNombre\Documentos\clase_pandas o una carpeta en tu escritorio.

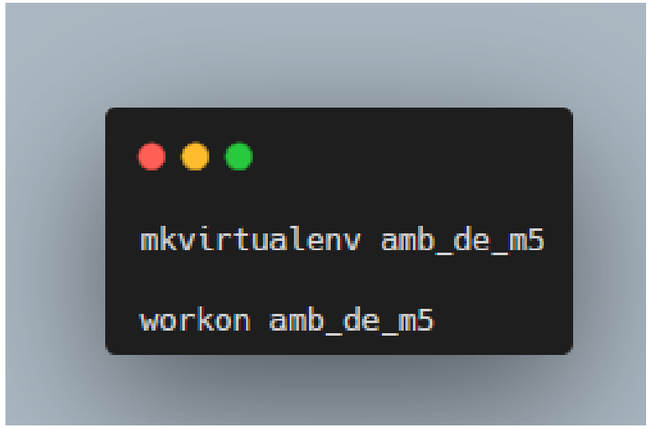
2. Crear un nuevo archivo Python
- Abre VSCode.
- Crea un nuevo archivo: analisis_demo.py.
- Guarda este archivo en la misma carpeta donde descargaste el CSV.

Introducción Práctica: Explorando la Calidad de datos con Pandas

Paso inicial: configurar el entorno

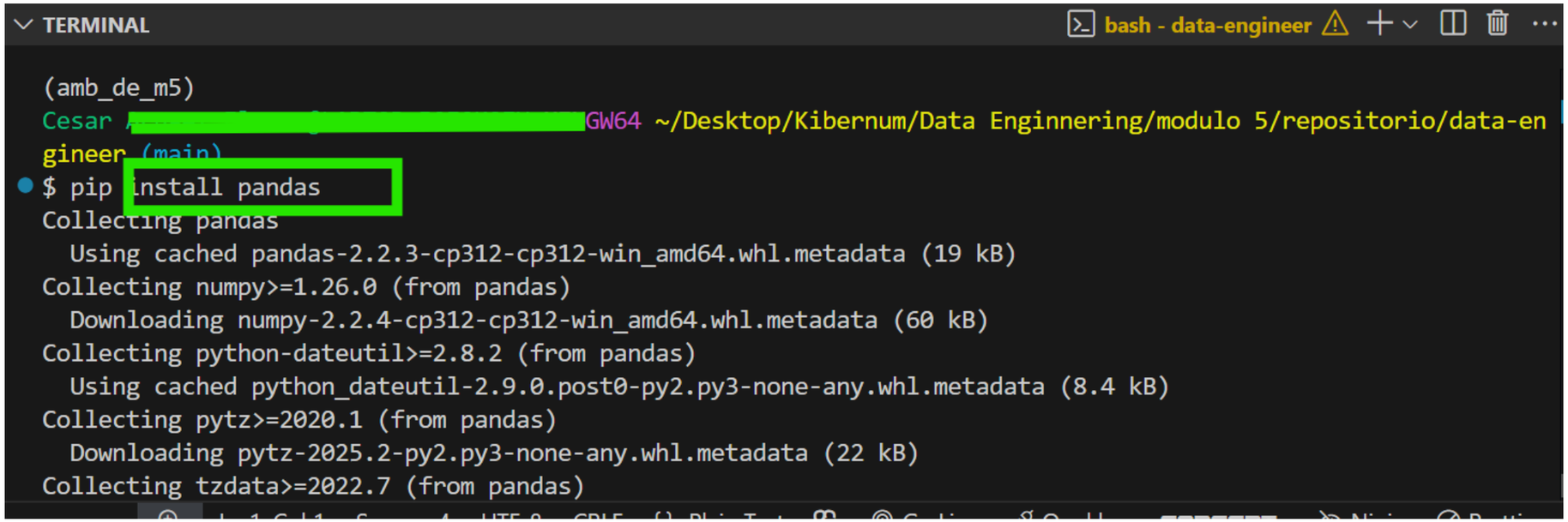
Primero, vamos a crear un entorno virtual. Esto nos permitirá trabajar de forma aislada con las librerías necesarias para el análisis.

Los prints de pantallas son de Windows, Git Bash y virtualenvwrapper-win, puedes ejecutar:



Introducción Práctica: Explorando la Calidad de datos con Pandas

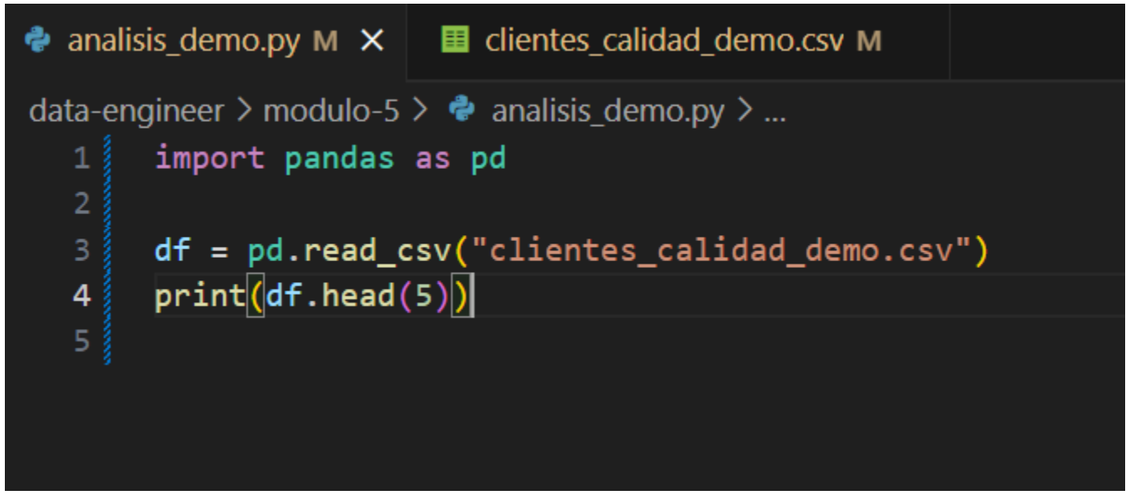
Instalamos Pandas con pip:



Introducción Práctica: Explorando la Calidad de datos con Pandas

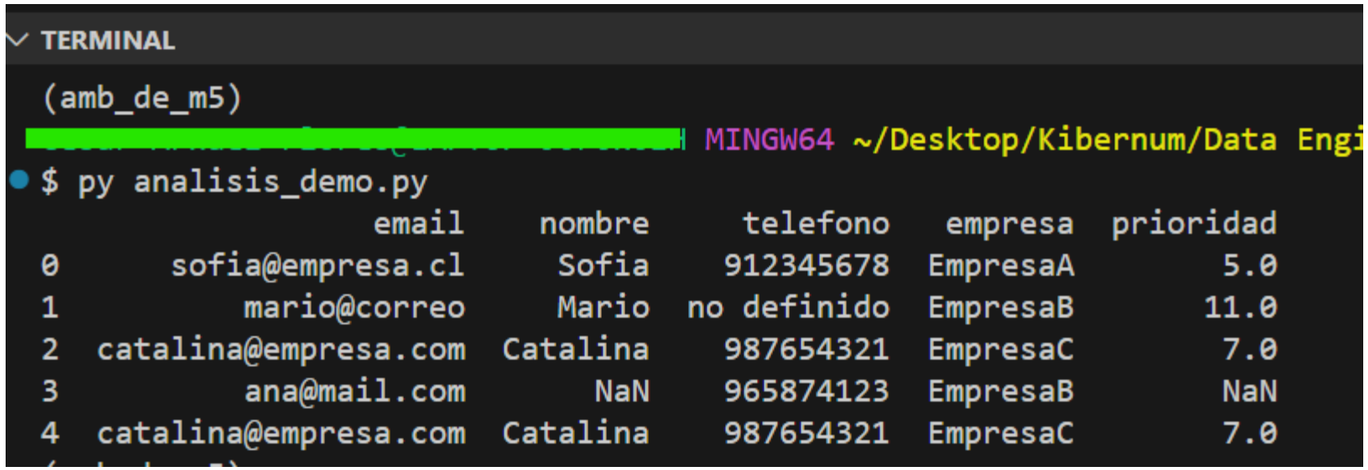
Cargar un archivo CSV y comenzar el análisis

Una vez descargado el archivo clientes_calidad_demo.csv, crea un archivo analisis_demo.py en el editor, en este ejejmplo VSC y escribe lo siguiente:



Introducción Práctica: Explorando la Calidad de datos con Pandas

Esto mostrará las primeras 5 filas del dataset y nos permitirá visualizar qué tipo de datos vamos a analizar.



Ejemplos Básicos de Validación con Pandas

Validar formato de correos correos electrónicos

Utilizando expresiones regulares podemos verificar si los correos tienen un formato válido (usuario@dominio). El resultado se almacena en una nueva columna email_valido.

```
# Priemro creamos una copia patra no modificar el original
# Se recomienda crear una copia del DataFrame original para no modificarlo directamente
df_copia = df.copy()


# Verificar si el campo 'email' es un email válido
# Se considera un email válido si tiene al menos un carácter antes y después del '@' y al m
df_copia['email_valido'] = df_copia['email'].str.contains(r'^\S+@\S+\.\S+$', na=False)
```

	email	nombre	telefono	empresa	prioridad	email_valido
0	sofia@empresa.cl	Sofia	912345678	EmpresaA	5.0	True
1	mario@correo	Mario	no definido	EmpresaB	11.0	False
2	catalina@empresa.com	Catalina	987654321	EmpresaC	7.0	True
3	ana@mail.com	NaN	965874123	EmpresaB	NaN	True
4	catalina@empresa.com	Catalina	987654321	EmpresaC	7.0	True

(amb_de_m5)

Ejemplos Básicos de Validación con Pandas

Medir la completitud por columna

Calculando el porcentaje de valores no nulos por columna, podemos identificar campos con datos faltantes o poco utilizados. Esta operación es comparable a una consulta SQL que utiliza IS NOT NULL.

```
completitud = df.notnull().mean() * 100
print(completitud)
```

```
------
email        100.0
nombre        80.0
telefono     100.0
empresa      100.0
prioridad     80.0
dtype: float64
------
```

Ejemplos Básicos de Validación con Pandas

Detectar registros duplicados duplicados

Contando cuántas filas del dataset están repetidas, podemos identificar problemas que afectan la unicidad y generan errores en reportes o métricas.

```
duplicados = df.duplicated().sum()
print(f"Duplicados detectados: {duplicados}")
```

```
Duplicados detectados: 1
(amb_de_m5)
```

Ejemplos Básicos de Validación con Pandas

```
df_copia['prioridad_valida'] = df_copia['prioridad'].between(1, 10)
print(df_copia)
```

```
df_copia['prioridad_valida'] = df_copia['prioridad'].between(1, 10)
```

Validar rango de valores

Verificando si los valores se encuentran dentro del rango esperado, podemos realizar validaciones comunes para variables categóricas o niveles.

Causas Comunes de Problemas de Calidad

Ingreso manual incorrecto

Errores al digitar datos, muchas veces por falta de validación o capacitación insuficiente. Por ejemplo, ingresar 987654321@ como correo en un formulario.

Duplicación de registros

Cuando un mismo dato se registra más de una vez, sin claves únicas ni controles adecuados. Un cliente aparece dos veces con el mismo correo pero nombres distintos.

Falta de estandarización

Datos que deberían tener un formato uniforme se registran de forma inconsistente. Algunas fechas como 10/01/2024, otras como 2024-01-10.

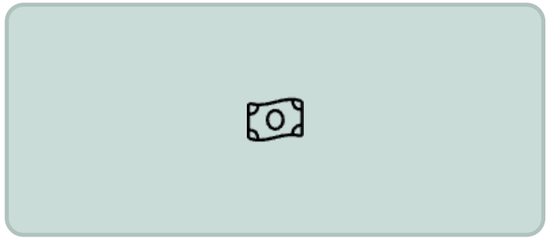
Datos obsoletos

Información que no ha sido actualizada y que ya no representa la situación actual. El número de teléfono de un cliente ya no existe, pero sigue en la base.

Impacto de la Mala Calidad de Datos

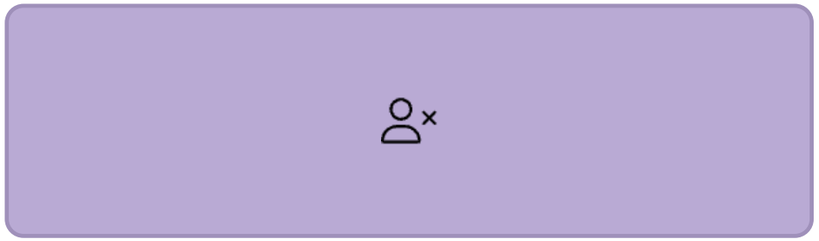
Decisiones erróneas

Basadas en información incorrecta



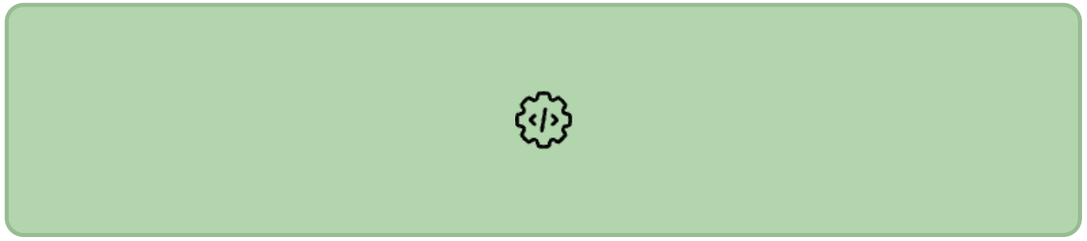
Pérdidas financieras

12.9 millones de dólares anuales



Deterioro de experiencia

Clientes insatisfechos por errores



Ineficiencias operativas

Procesos más lentos y costosos

Características Clave de la Norma ISO 8000

Calidad como activo

Reconoce los datos como activos estratégicos, comparables con recursos financieros o humanos.

Interoperabilidad

Propone formatos y estructuras que permiten que los datos sean compartidos entre distintos sistemas.

Trazabilidad

Requiere poder rastrear el origen, los cambios y el uso de los datos a lo largo de su ciclo de vida.

Validación y consistencia

Exige procesos de verificación estructurados para mantener datos consistentes en toda la organización.

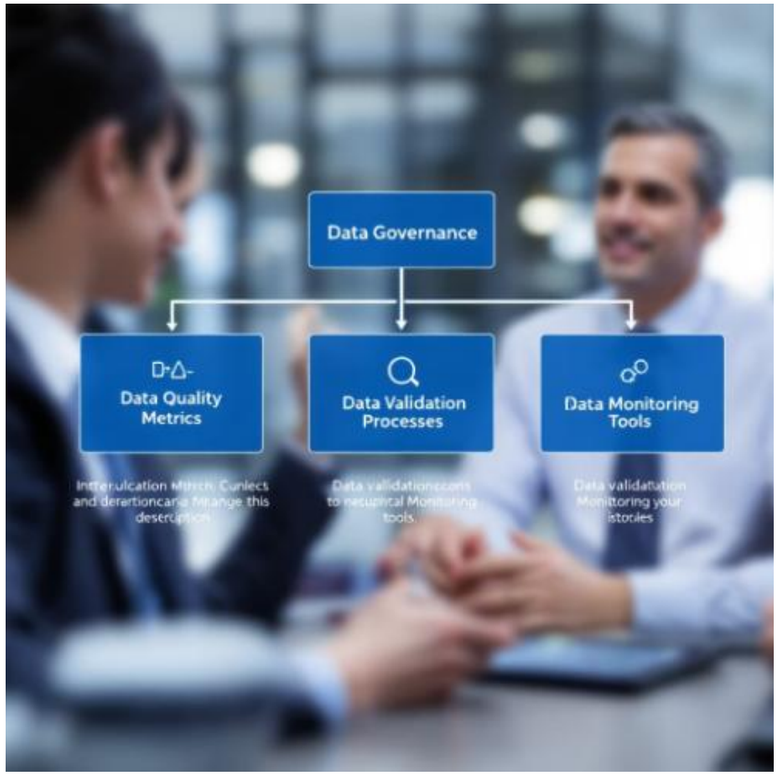
Más Características de ISO 8000

Estandarización de vocabularios

La norma promueve el uso de catálogos, taxonomías y reglas comunes para evitar ambigüedades en la interpretación de los datos. Por ejemplo, si una empresa registra productos con nombres distintos en diferentes áreas (ej. "Notebook HP i5" vs. "HP Intel Core i5"), incumple el principio de vocabulario estandarizado.

Independencia tecnológica

La norma no está atada a herramientas específicas, lo que permite su adopción en diversos entornos tecnológicos. Esto facilita su implementación en organizaciones con infraestructuras heterogéneas y permite que los principios de calidad se mantengan incluso cuando cambian las tecnologías subyacentes.



La ISO 8000 establece un marco estructurado para la gestión de la calidad de los datos, aplicable a diversos contextos como inventarios, información de productos, directorios de clientes o datos de referencia en sistemas empresariales.

Evaluaciones de Calidad de Datos

Una evaluación de calidad de datos (o Data Quality Assessment) no es simplemente una revisión técnica o automatizada: es un proceso integral y estratégico, que busca diagnosticar el estado de los datos con base en criterios objetivos, conocimiento del negocio y buenas prácticas normativas.

Indicadores cuantitativos

Métricas concretas como porcentaje de completitud, duplicación, precisión o validez.

Implementación de mejoras

Acciones correctivas basadas en los hallazgos de la evaluación.

Conocimiento del negocio

Entender qué datos son realmente críticos para los procesos operativos y decisionales.

Cumplimiento normativo

Aplicación de estándares como ISO 8000, políticas internas y buenas prácticas del sector.

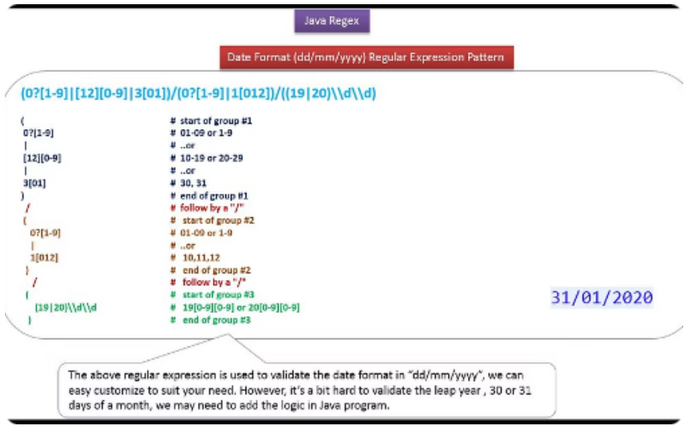
Métricas para Monitoreo de Calidad

Métrica	Fórmula sugerida	Interpretación
Tasa de completitud	*(Campos no nulos / Total de campos) × 100*	¿Qué porcentaje de campos obligatorios están llenos?
Tasa de duplicación	*(Registros duplicados / Total registros) × 100*	¿Qué tan frecuente se repite la misma entidad?
Tasa de precisión	*(Registros correctos / Total registros) × 100*	¿Qué tan bien representan los datos la realidad esperada?
Tasa de valores válidos	*(Valores válidos / Total valores) × 100*	¿Cuántos valores cumplen con su formato o dominio esperado?

Ejemplo de Monitoreo de Calidad

Campo	Completitud (%)	Validez (%)	Duplicados detectados
email	100%	90%	3
nombre	95%	100%	0
prioridad	80%	85%	2

Técnicas y Herramientas para Medir Calidad



Regex / Expresiones regulares

Herramientas potentes para la validación de patrones y formatos en campos como RUT, correo electrónico, números telefónicos y otros datos estructurados. Permiten definir reglas precisas para verificar que los datos cumplan con formatos específicos.

OpenRefine

Aplicación de código abierto que facilita la limpieza interactiva de datos, la estandarización de nombres y la detección de valores atípicos. Especialmente útil para conjuntos de datos de tamaño medio que requieren transformaciones y normalizaciones.

Talend Data Quality

Solución empresarial que ofrece capacidades avanzadas de perfilado automático de datos, detección de duplicados y evaluación de calidad por dimensiones. Incluye funcionalidades de integración con diversas fuentes de datos y generación de informes detallados.

Conclusiones y Reflexiones

1

La calidad como inversión

Invertir en calidad de datos no es un gasto, sino una inversión que reduce costos operativos y mejora la toma de decisiones.

Responsabilidad compartida

La calidad de datos no es solo responsabilidad del departamento de TI, sino de toda la organización.

Mejora continua

La calidad de datos debe monitorearse y mejorarse constantemente, adaptándose a las cambiantes necesidades del negocio.

Formación y cultura

Crear una cultura organizacional que valore la calidad de los datos es tan importante como implementar herramientas técnicas.

Discusión en Clase

- **¿Puede un dato ser correcto, pero irrelevante?**

Por ejemplo, si el campo “profesión” está bien escrito, pero no se utiliza en ningún análisis ni operación, ¿vale la pena mantenerlo? ¿Qué implicancias puede tener?

Enlaces de Interés

- Documentación oficial de Pandas
  https://pandas.pydata.org/docs/

- Guía introductoria en español (oficial de PyData)
  https://pandas.pydata.org/docs/user_guide/index.html

- 10 minutos para Pandas (Quickstart)
  https://pandas.pydata.org/docs/user_guide/10min.html

Preguntas de Reflexión Final

Dimensiones Críticas

¿Qué dimensión de calidad de datos (como completitud, validez, unicidad) consideras más crítica en tu área de interés? ¿Por qué?

Impacto en Decisiones

¿Cómo influye la calidad de los datos en los procesos de toma de decisiones dentro de una organización? Da un ejemplo.

Aplicación Práctica

Ahora que conoces herramientas como Pandas y el marco ISO 8000, ¿qué pasos tomarías como ingeniero/a de datos frente a un archivo que presenta duplicados, errores de formato y datos obsoletos?