# Qué es la Limpieza de datos o data cleansing

La **limpieza de datos** , también conocida como **data cleansing** o **data cleaning** , es el proceso de identificar y corregir errores, inconsistencias y datos incompletos o duplicados en un conjunto de datos. Es una parte fundamental del análisis de datos y la preparación de datos para su posterior uso en aplicaciones y análisis.

La calidad de los datos es esencial para garantizar la precisión y confiabilidad de cualquier análisis o modelo que se realice. Los datos sucios o no confiables pueden conducir a resultados incorrectos y conclusiones sesgadas, lo que puede afectar negativamente la toma de decisiones y los resultados comerciales.

## En esta página

# Importancia de la limpieza de datos o data cleansing

La limpieza de datos, también conocida como data cleansing, es un paso crítico e indispensable en el proceso de análisis de datos y toma de decisiones basadas en datos. La importancia de la limpieza de datos radica en varios aspectos fundamentales:

- **Precisión en el análisis** : Los datos sucios o no confiables pueden llevar a resultados incorrectos y sesgados. Al limpiar los datos y corregir errores, se asegura que los resultados del análisis sean precisos y confiables.
- **Toma de decisiones informada** : La calidad de los datos influye directamente en la calidad de las decisiones que se toman. Una toma de decisiones basada en datos de alta calidad es más informada y puede llevar a resultados más exitosos.
- **Confianza en los resultados** : Los datos limpios y precisos inspiran confianza en los resultados del análisis. Tener datos confiables respalda las conclusiones y recomendaciones que se obtienen a partir del análisis.
- **Eficiencia en el análisis** : La limpieza de datos evita retrabajos y pérdidas de tiempo al garantizar que los datos estén listos para su análisis. Los analistas pueden centrarse en interpretar y utilizar los datos, en lugar de lidiar con problemas de calidad de datos.

- **Reducción de riesgos** : Los datos erróneos pueden llevar a decisiones desfavorables o incluso costosas. Al limpiar los datos, se reduce el riesgo de tomar decisiones basadas en información incorrecta.
- **Mejora de la productividad** : La limpieza de datos puede automatizarse en muchos casos, lo que permite un procesamiento más rápido y una mayor productividad en el análisis.
- Cumplimiento normativo y legal: En ciertas industrias, como la salud y las finanzas, es crucial mantener la precisión y confidencialidad de los datos para cumplir con las regulaciones y normas legales.
- **Mayor competitividad** : Las organizaciones que priorizan la limpieza de datos pueden tomar decisiones más acertadas y estar mejor preparadas para enfrentar los desafíos comerciales y superar a la competencia.
- **Optimización de recursos** : La limpieza de datos ayuda a evitar el uso de recursos en datos incorrectos o duplicados, lo que puede resultar en ahorros significativos.
- **Facilitación del aprendizaje automático** : Los modelos de machine learning se benefician de datos limpios y de alta calidad. La limpieza de datos prepara los datos para su uso en algoritmos de aprendizaje automático y mejora la precisión de los modelos.

En resumen, la limpieza de datos es un proceso esencial para garantizar que los datos utilizados en el análisis sean confiables, precisos y útiles. Es un paso crítico para obtener información valiosa, tomar decisiones informadas y obtener ventaja competitiva en el mundo actual impulsado por los datos.

# Proceso de la limpieza de datos o data cleansing

El proceso de limpieza de datos, también conocido como data cleansing, es una serie de pasos sistemáticos que se llevan a cabo para identificar, corregir y eliminar errores, inconsistencias y datos incompletos o duplicados en un conjunto de datos. Este proceso es esencial para asegurar que los datos utilizados en el análisis sean precisos, confiables y estén listos para su uso en aplicaciones y toma de decisiones. A continuación, se describen los pasos típicos del proceso de limpieza de datos:

- **Recopilación de datos** : El primer paso es recopilar los datos de diferentes fuentes, como bases de datos, archivos, sistemas en línea, etc.
- **Exploración inicial** : Se realiza una exploración inicial del conjunto de datos para identificar problemas comunes, como valores faltantes, errores tipográficos, valores atípicos, etc.

- **Detección de valores faltantes** : Se identifican y registran los valores faltantes en el conjunto de datos. Los valores faltantes pueden ser eliminados, reemplazados por valores estimados o tratados de manera adecuada según el contexto.
- **Tratamiento de valores atípicos** : Los valores atípicos o valores que se desvían significativamente del resto de los datos se analizan y se decide si deben ser corregidos, eliminados o conservados según el contexto del análisis.
- **Corrección de errores** : Se buscan y corrigen errores tipográficos y otros errores evidentes que puedan afectar la precisión de los datos.
- **Eliminación de duplicados** : Se identifican y eliminan registros duplicados en el conjunto de datos para evitar distorsiones en los resultados del análisis.
- **Normalización y estandarización** : Los datos se ajustan al mismo formato y unidad para facilitar su comparación y análisis.
- **Validación de datos** : Se verifica que los datos cumplan con las reglas y restricciones establecidas, y que sean coherentes y precisos.
- **Verificación cruzada** : Los datos se comparan con otras fuentes o bases de datos para garantizar su consistencia y precisión.
- **Documentación** : Todas las acciones realizadas durante el proceso de limpieza de datos se documentan para garantizar la

transparencia y reproducibilidad del trabajo realizado.

- **Evaluación de la calidad** : Se evalúa la calidad general de los datos limpios para asegurar que estén listos para su uso en análisis y toma de decisiones.

Es importante destacar que el proceso de limpieza de datos puede ser un proceso iterativo, donde se realizan múltiples rondas de revisión y corrección para asegurar la calidad de los datos. Además, el proceso de limpieza de datos puede variar según el contexto y los requisitos específicos del análisis. Un enfoque cuidadoso y sistemático en la limpieza de datos es esencial para garantizar que los datos utilizados en el análisis sean confiables y puedan proporcionar información valiosa para la toma de decisiones.

# Técnicas de limpieza de datos o data cleansing

Existen diversas técnicas de limpieza de datos, también conocidas como data cleansing, que se utilizan para identificar y corregir errores, valores atípicos y datos incompletos en un conjunto de datos. A continuación, se presentan algunas de las técnicas más comunes:

1. **Eliminación de valores faltantes** : Los valores faltantes en un conjunto de datos pueden ser problemáticos para el análisis.

Una técnica común es eliminar las filas que contienen valores faltantes. Sin embargo, esta técnica debe usarse con precaución, ya que puede llevar a la pérdida de información relevante.

2. **Imputación de valores faltantes** : En lugar de eliminar las filas con valores faltantes, se pueden reemplazar esos valores por estimaciones basadas en otros datos o técnicas estadísticas, como el promedio, la mediana o la interpolación.

3. **Detección y manejo de valores atípicos** : Los valores atípicos son observaciones inusuales o extremas que pueden afectar negativamente el análisis. Es importante identificar y manejar estos valores adecuadamente, ya sea eliminándolos, transformándolos o sustituyéndolos por valores más adecuados.

4. **Corrección de errores tipográficos** : Los errores tipográficos y de escritura pueden provocar datos inconsistentes. Se pueden aplicar técnicas de corrección ortográfica y normalización de texto para solucionar estos problemas.

5. **Eliminación de duplicados** : Los datos duplicados pueden generar resultados sesgados y conducir a conclusiones incorrectas. La identificación y eliminación de duplicados es fundamental para asegurar la calidad de los datos.

6. **Normalización de datos** : La normalización es el proceso de transformar los datos a un formato estándar y uniforme. Esto puede incluir la conversión de letras a mayúsculas o

minúsculas, la estandarización de unidades de medida o la representación de fechas en un formato consistente.

7. **Validación cruzada** : La validación cruzada implica comparar los datos en un conjunto con los datos en otro conjunto o fuente para asegurar la consistencia y precisión de los datos.
8. **Consistencia de formatos y tipos de datos** : Asegurar que los datos en el conjunto tengan formatos y tipos de datos coherentes es importante para evitar errores en el análisis.
9. **Tratamiento de datos desactualizados** : En algunas ocasiones, los datos pueden estar desactualizados. Se debe considerar la actualización de los datos para reflejar la información más reciente.
10. **Estándares de limpieza y reglas de validación** : Establecer estándares y reglas de validación para los datos es esencial para garantizar la calidad de los datos y mantener la consistencia en futuras actualizaciones.

Cabe destacar que las técnicas de limpieza de datos pueden variar según el tipo de datos y el contexto del análisis. Es importante aplicar estas técnicas de manera cuidadosa y estar atento a los efectos que puedan tener en los resultados del análisis. La limpieza de datos es un proceso iterativo que requiere tiempo y esfuerzo, pero es fundamental para garantizar que los datos sean confiables y precisos para la toma de decisiones y análisis adecuado.

# Herramientas de limpieza de datos

Existen diversas herramientas de limpieza de datos o data cleansing disponibles en el mercado que facilitan y agilizan el proceso de limpiar y preparar datos para su análisis. Estas herramientas ofrecen una variedad de funciones y características para identificar y corregir errores, eliminar duplicados, manejar valores faltantes y mejorar la calidad general de los datos. Algunas de las herramientas más populares de limpieza de datos son las siguientes:

- **OpenRefine** : OpenRefine es una herramienta de código abierto que permite limpiar y transformar grandes conjuntos de datos de forma interactiva. Ofrece capacidades avanzadas de limpieza, como la detección y corrección de errores tipográficos, la eliminación de espacios en blanco y la normalización de valores.
- **Trifacta Wrangler** : Trifacta Wrangler es una plataforma de preparación de datos que proporciona una interfaz visual para limpiar y dar formato a datos complejos y desordenados. Utiliza técnicas de aprendizaje automático para sugerir transformaciones y patrones comunes en los datos.
- **DataWrangler** : DataWrangler es otra herramienta de preparación de datos de código abierto desarrollada por el

equipo de investigación de Stanford. Permite la manipulación y limpieza de datos mediante una interfaz interactiva y fácil de usar.

- **Talend Data Preparation** : Talend Data Preparation es una herramienta integral que ofrece capacidades de limpieza y preparación de datos en tiempo real. Permite la integración de datos de múltiples fuentes y la limpieza a gran escala.
- **Open Data Kit (ODK)** : ODK es una plataforma de recolección de datos móviles que también proporciona funcionalidades básicas de limpieza y validación de datos recopilados a través de formularios móviles.
- **Microsoft Excel** : Aunque no es una herramienta dedicada exclusivamente a la limpieza de datos, Excel ofrece varias funciones y herramientas útiles para la limpieza básica de datos, como la eliminación de duplicados, la validación de datos y el uso de fórmulas para transformar y limpiar los datos.
- **Python y pandas** : Python es un lenguaje de programación muy utilizado en el análisis de datos, y la biblioteca pandas proporciona funciones y métodos para limpiar y transformar datos. Es una opción poderosa y flexible para realizar limpieza de datos personalizada.

# Problemas de no realizar limpieza de datos o data

# cleansing

La falta de usar data cleansing o limpieza de datos puede llevar a una serie de problemas que afectan la calidad y confiabilidad de los análisis y decisiones basadas en datos. Algunos de los problemas más comunes que surgen cuando no se realiza una limpieza adecuada de los datos son los siguientes:

- **Inexactitud en los resultados del análisis** : Datos sucios o inexactos pueden llevar a resultados incorrectos en el análisis, lo que puede conducir a decisiones erróneas y no informadas.
- **Duplicados y redundancias** : La presencia de datos duplicados o redundantes puede distorsionar los resultados del análisis y llevar a conclusiones erróneas.
- **Datos incompletos** : Valores faltantes o datos incompletos pueden afectar la integridad de los resultados y la precisión de los modelos de análisis.
- **Errores tipográficos y formatos inconsistentes** : Datos con errores tipográficos o en formatos inconsistentes pueden causar problemas en el procesamiento y análisis de los datos.
- **Sesgo en los resultados** : Datos no representativos o sesgados pueden influir en los resultados del análisis y llevar a conclusiones sesgadas o inexactas.
- **Pérdida de tiempo y recursos** : La falta de limpieza de datos puede llevar a un uso ineficiente de tiempo y recursos en la

corrección de errores y en la repetición del análisis.

- **Decisiones incorrectas** : Las decisiones basadas en datos incorrectos o no confiables pueden tener un impacto negativo en la eficiencia y eficacia de una organización.
- **Incumplimiento normativo** : En ciertas industrias, como la salud y las finanzas, el uso de datos no limpios puede llevar a incumplimientos de normas y regulaciones.
- **Pérdida de oportunidades** : Datos no limpios pueden ocultar información valiosa y oportunidades comerciales que de otro modo podrían haber sido identificadas.
- **Falta de confianza en los datos** : La falta de limpieza y calidad en los datos puede generar desconfianza entre los usuarios y analistas, lo que puede limitar el uso efectivo de los datos en la toma de decisiones.

En resumen, la limpieza de datos es un paso crucial en el proceso de análisis de datos. La falta de limpieza puede llevar a problemas significativos que afectan la precisión, confiabilidad y eficacia del análisis y las decisiones basadas en datos. Es esencial dedicar tiempo y recursos para garantizar que los datos utilizados en el análisis sean precisos, completos y confiables.

**Quizá te pueda interesar:**

- ¿Qué es una ETL?
- ¿Qué es el business intelligence?

- ¿Qué es el business analytics?