

«Talento Tech»

Data Analytics

con Python

Clase 04





Clase N° 4 | Calidad de Datos

Temario:

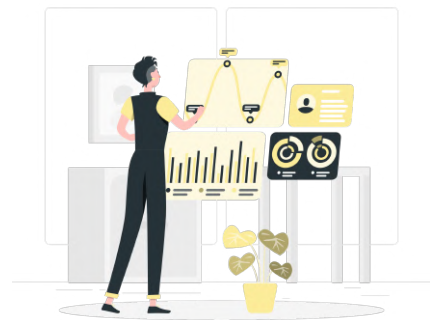
1. Concepto de calidad de los datos y su importancia.
2. Identificación y tratamiento de valores nulos y duplicados.

Objetivos de la clase:

- Conocer los conceptos principales de la calidad de datos y su importancia en la analítica de datos.
- Aprender a identificar valores nulos y duplicados. Considerar las diferentes estrategias para tratarlos.

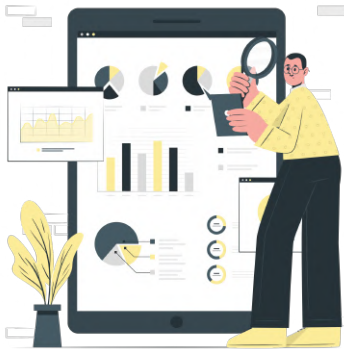
Calidad de los Datos

La **calidad de los datos** es un concepto fundamental en el campo de la analítica de datos, ya que se refiere a la **medida en que los datos son aptos para su propósito específico**. La calidad de los datos no solo implica que los datos sean precisos, sino que también deben ser completos, consistentes, relevantes y válidos. Esta multifacética dimensión se desglosa en varios atributos, que exploraremos a continuación:



1. **Precisión:** Los datos deben reflejar con exactitud la realidad. Por ejemplo, en un conjunto de datos de ventas, si el precio de un producto se registró incorrectamente, esto afectará la precisión de cualquier análisis posterior sobre ingresos.
2. **Compleitud:** Se refiere a cuán completos son los datos. En un cuestionario, si ciertos campos permanecen sin respuesta (nulos), esto puede llevar a una falta de información necesaria para un análisis robusto. Por ejemplo, si un cliente no proporciona su edad, y este dato es vital para segmentar el mercado, la completitud se ve afectada.
3. **Consistencia:** Los datos deben ser consistentes entre diferentes fuentes y registros. Por ejemplo, si en una base de datos se registra que un cliente vive en "Calle A" y en otra base de datos se registra como "Avenida A", esta inconsistencia puede causar problemas al intentar combinar o comparar datos de diferentes fuentes.
4. **Validez:** Esto implica que los datos se ajustan a ciertos criterios o restricciones predefinidos. Por ejemplo, un campo que requiere fechas debe contener entradas que sean efectivamente fechas y no texto. Si un campo de fecha tiene un valor como "2024-31-12", este no es válido en el formato estándar de fecha.

5. **Relevancia:** Los datos deben ser relevantes para el análisis o la tarea en cuestión. La recopilación de datos que no sirven para responder a las preguntas o toparse con los problemas específicos que se están tratando es una pérdida de recursos.



La importancia de la **calidad de los datos** no puede subestimarse. Las decisiones empresariales, desde la estrategia de marketing hasta las decisiones de inversión, se basan en gran medida en los análisis de datos. Si los datos son defectuosos, las conclusiones serán erróneas, lo que podría resultar en decisiones poco acertadas. Un ejemplo claro se encuentra en el lanzamiento de nuevos productos: basar la estrategia de lanzamiento en datos de ventas inflados o incorrectos puede llevar a sobreproducción, pérdida de recursos y desequilibrio en el inventario.

Para asegurar la calidad de los datos en un análisis con Python, se recomienda utilizar la **biblioteca Pandas** para explorar y evaluar nuestros datos.

Por ejemplo, se pueden utilizar métodos como `info()` para obtener un resumen del DataFrame y identificar el tipo de datos, así como `isnull().sum()` para contar el número de valores nulos. Esto proporciona una buena base para entender si nuestros datos están listos para el análisis.

Identificación y Tratamiento de Valores Nulos y Duplicados

El manejo de **valores nulos y duplicados** son procesos fundamentales para mantener la calidad de los datos antes de realizar cualquier análisis. Estos problemas son comunes y pueden surgir en cualquier conjunto de datos, particularmente al recolectar información de diversas fuentes.

Valores Nulos

Los valores nulos, que **representan la ausencia de información**, pueden presentarse por diversas razones, como errores durante la recolección de datos, problemas de formulario o simplemente porque algunos encuestados eligen no responder. La presencia de valores nulos puede complicar análisis estadísticos y modelos predictivos, ya que muchos algoritmos no pueden funcionar correctamente con datos faltantes.

Existen varias estrategias para tratar los valores nulos:

1. **Eliminación:** En algunos casos, puede ser apropiado eliminar filas o columnas enteras que contienen valores nulos si estos son pocos y no afectan significativamente al análisis.
2. **Imputación:** Consiste en rellenar los valores nulos con un valor adecuado. Metodologías comunes incluyen la imputación con la media, mediana o moda de la columna. En el caso de datos más complejos, se puede utilizar métodos de imputación más avanzados como algoritmos de machine learning.
3. **Predicción:** En algunos casos, se pueden construir modelos predictivos que estimen los valores nulos en función de otros datos disponibles.



Duplicados

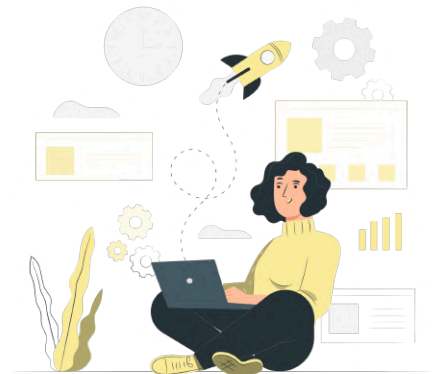
Los duplicados son **registros que aparecen más de una vez en un conjunto de datos**. Esto puede ocurrir debido a errores en la entrada de datos, sistemas de registro múltiples o combinaciones de datasets. La existencia de duplicados puede distorsionar la interpretación de los datos, como inflar recuentos o crear sesgos en el análisis.

Para manejar los duplicados, se sigue el siguiente proceso:

1. **Identificación:** Identificar qué filas son duplicadas.
2. **Eliminación:** Eliminar las filas duplicadas manteniendo solo la primera ocurrencia.

Reflexión final

La identificación y tratamiento de valores nulos y duplicados son pasos esenciales en el proceso de limpieza de datos que aseguran que nuestro análisis sea realizado con conjuntos de datos de alta calidad. Este enfoque no solo mejora la **fiabilidad** de nuestras conclusiones, sino que también potencia la **efectividad** de las decisiones empresariales basadas en esos datos. Al prestar atención a estos aspectos dentro de la analítica de datos, podemos crear análisis más robustos y precisos que verdaderamente reflejan la realidad y apoyan la toma de decisiones estratégicas.



Próximos Pasos

- Los desafíos de la raw data.
- Técnicas para limpiar datos: eliminación de duplicados, caracteres no deseados, corrección en los tipos de datos.

Ejercicios prácticos



Actividad 1: Identificación de Valores Nulos y Duplicados

Contexto



Para esta actividad, tu mentor será Matías, quien compartirá sus conocimientos sobre la limpieza de datos y te guiará en el proceso. Tu tarea consiste en cargar datos en un DataFrame usando Pandas y aplicar los métodos específicos para identificar filas con datos faltantes y duplicados, lo que te ayudará a entender cómo se trabaja con los datos en un entorno profesional.

Objetivos

Practicar la carga de datos en un DataFrame utilizando Pandas.
Identificar y contar valores nulos en diferentes columnas del DataFrame.
Detectar y contar registros duplicados.
Generar un informe que resuma los hallazgos de la limpieza de datos.

Ejercicio práctico

1. Descargar los datos en formato CSV y cargarlos en un DataFrame en Python usando Pandas. Utilizar el método `isnull()` para identificar las filas con datos faltantes y contar el número de valores nulos por columna.
2. Usar el método `duplicated()` para identificar las filas duplicadas y contar cuántas filas duplicadas hay en total.

3. Crear un informe que incluya:
 - a. La cantidad total de registros en el DataFrame.
 - b. La cantidad total de valores nulos por columna.
 - c. La cantidad de filas duplicadas.
4. Crear un dataframe de los registros duplicados para que se vea el contenido.

Sets de datos

- [satis_clientes](#)

¿Por qué importa esto en SynthData?

La identificación de valores nulos y duplicados es vital en el trabajo de análisis de datos. En SynthData, utilizamos datos limpios y precisos para tomar decisiones inteligentes. Al aprender a reconocer estos problemas en los datos, brindarás un gran apoyo al equipo en la generación de informes y análisis más confiables.

Actividad 2: Exploración y limpieza preliminar con Python.

Contexto



En esta actividad, trabajarás guiado por Luis, nuestro Analista de BI, con una planilla de Google Sheets que registra la temperatura corporal de un grupo de personas durante 10 días. Debes realizar un examen preliminar para identificar datos problemáticos, como duplicados y valores nulos. Esta práctica te ayudará a dar los pasos previos a la limpieza de datos, un componente vital en cualquier proyecto de análisis.

Objetivos

Identificar los datos duplicados y nulos en el conjunto.
Analizar el mejor tratamiento para los datos anómalos y nulos.

Ejercicio práctico

Crear un dataframe a partir de la planilla de cálculo y efectuar un examen preliminar.
Identificar los datos:

- a. duplicados
- b. nulos

Analizá cuál sería el mejor tratamiento para los datos anómalos (fuera de rango) y nulos en los siguientes contextos:


		Ignorar	Eliminar	Reemplazar	Analizar
Estudio estadístico	anómalos				
	nulos				
Seguimiento de un paciente	anómalos				
	nulos				
Detección de casos de contagio	anómalos				
	nulos				

Sets de datos

[Actividad 2](#)

¿Por qué importa esto en SynthData?

La exploración y limpieza de datos son pasos esenciales para garantizar que cualquier análisis posterior sea robusto y significativo. En SynthData, nos enfrentamos a datos en diferentes formatos, y es crucial asegurarnos de que estén listos para ser analizados. Esta actividad te enseñará a detectar problemas comunes en conjuntos de datos, como duplicados y valores nulos, y cómo abordarlos adecuadamente.

 Estos ejercicios son una simulación de cómo se podría resolver el problema en este contexto específico. Las soluciones encontradas no aplican de ninguna manera a todos los casos.

Recordá que las soluciones dependen de los sets de datos, el contexto y los requerimientos específicos de los stakeholders y las organizaciones.



Buenos Aires
aprende
Agencia de Políticas para el Futuro

BA Buenos
Aires
Ciudad