

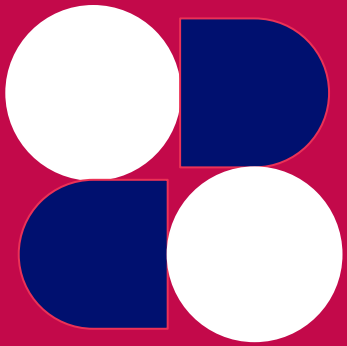


Programa PUCP DATA LITERACY

Formación tecnológica

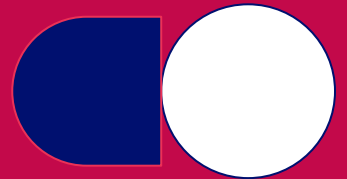
Dirección Académica de Planeamiento y Evaluación
Dirección de Gestión del Talento Humano





SESIÓN 3:

Principios de limpieza de datos





Al finalizar la sesión, lograrás:

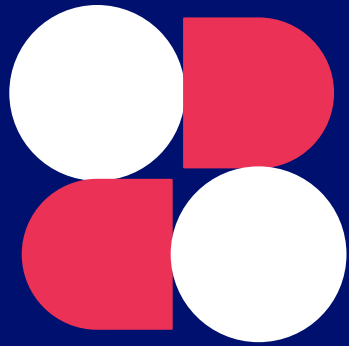
- Justificar la importancia de la limpieza de datos en la obtención de resultados precisos.
- Conocer un framework para la limpieza de datos



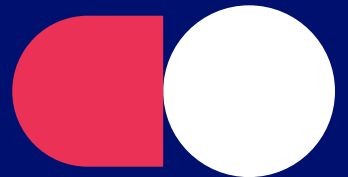
Contenido:

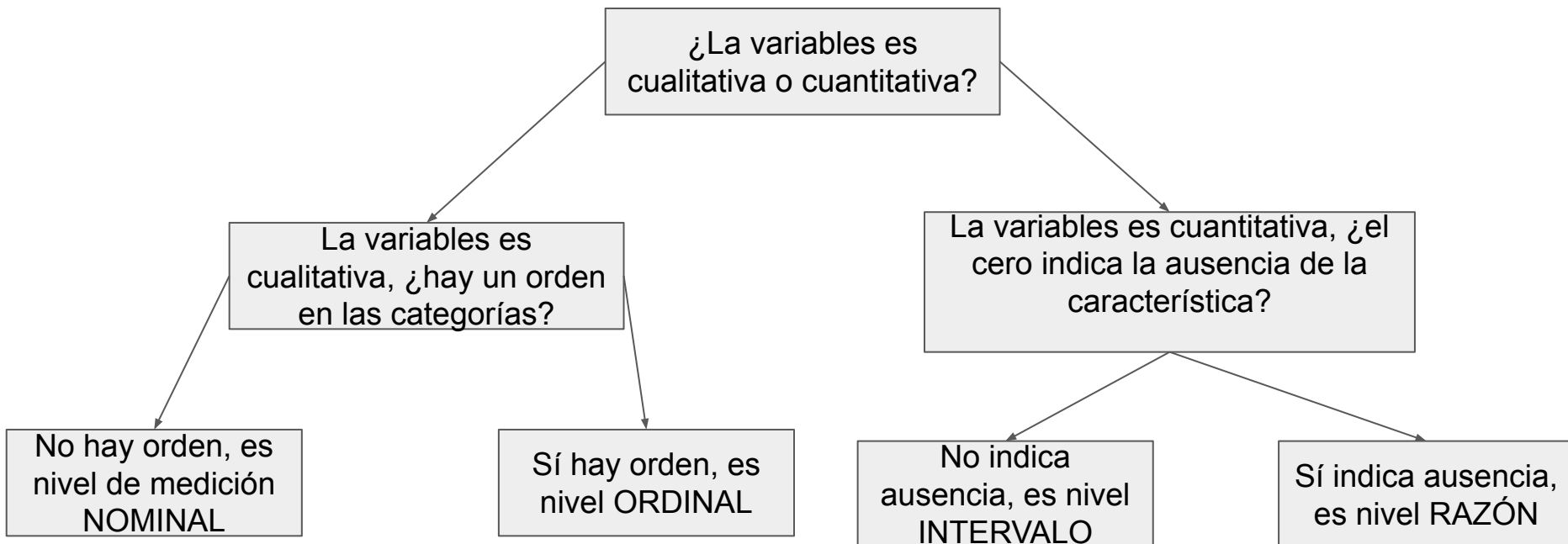


- 1 Repaso de los niveles de medición
- 2 Proceso de análisis de datos
- 3 Definición limpieza de datos
- 4 Framework de limpieza de datos



Repaso niveles de medición







Variables nominales

- Nombre del estudiante
- Título de un artículo
- Nombre de un programa académico
- Nombre de los software que tiene la universidad
- Idiomas que ofrece el centro de idiomas
- Nombre de las revistas
- Ciudad de residencia
- Estado civil

Variables ordinales

- Nivel de satisfacción de los estudiantes con los profesores
- Nivel alcanzado de estudio de los profesores
- Semestre en que se encuentra el estudiante
- Grado de Disponibilidad de Sistemas Críticos
- Evaluación de la Seguridad de la Red
- Nivel de Experiencia del Personal de TI
- Nivel de Urgencia de Incidentes de TI



Ejemplos



Variables Intervalo

- Temperatura de los salones
- Puntuación en una Prueba de Aptitud Académica
- Fecha de inicio del semestre de cada semestre

Variables razón

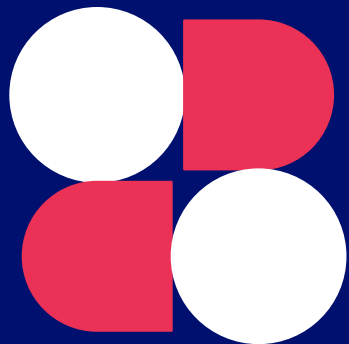
- Número de créditos completados
- Edad de los estudiantes
- Número de años trabajados en la universidad
- Número de publicaciones por facultad
- Presupuesto de inversión en infraestructura por año
- Tasa de retención de estudiantes
- Número de libros en cada sección de la biblioteca
- Distancia recorrida por los estudiantes para llegar a la universidad
- Tiempo Medio de Graduación
- Horas impartidas por profesor



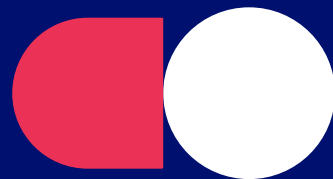
Resumen



Tipo de variable	Niveles de medición	Propiedades					
		Distinguir observaciones	Hay un orden	Igualdad de intervalos	Igualdad en las proporciones	Cero	Operaciones matemáticas
cualitativas	nominal	Sí	No	No	No		
	ordinal	Sí	Sí	No	No		
cuantitativas	intervalo	Sí	Sí	Sí	No	0 es arbitrario	sumar, restar
	razón	Sí	Sí	Sí	Sí	0 significa la ausencia	sumar, restar ,multiplicar, dividir



Proceso de analizar datos





En el día a día



Consumiendo
datos



Tomar decisiones
Basadas en datos



Presentar



Evaluar



Analizar



Reconocer el problema u
oportunidad



Formular hipótesis



Definir y
recolectar datos



Limpiar y procesar





Reconocer el problema o cuál es la oportunidad

- Identificar cuál es el problema y por qué es importante
- Identificar a las partes interesadas en el problema (desde líderes hasta analistas)
- Definir qué tipos de decisiones se pueden tomar a partir del resultado del análisis





Formular hipótesis



Definir muy bien el problema y cuáles son las principales variables que vamos a incluir en el análisis





Formular hipótesis



Guiarse con los resultados de análisis previos

- Qué datos utilizaron
- Qué tipo de análisis
- Qué resultados obtuvieron

Otros recursos:

- Internet
- Libros
- Conversar con analista
- Gestión de conocimiento de la universidad





Encuadre del problema



- ¿Has definido claramente un problema u oportunidad que sea importante para tu área/ organización?
- ¿Has considerado múltiples formas/ alternativas de resolver el problema?
- ¿Has identificado a los interesados para el problema?
- ¿Estás seguro de que la manera en que planeas resolver el problema resonará con los interesados y que utilizarán los resultados para tomar una decisión?





Encuadre del problema

- ¿Tienes claro qué decisión se debe tomar y quién la tomará en base a los resultados de tu análisis una vez que se resuelva el problema?
- ¿Has comenzado con una definición amplia del problema, pero luego lo has reducido a un problema muy específico con una formulación clara sobre la pregunta que se abordará, los datos que se aplicarán, el tipo de análisis y los posibles resultados?





Encuadre del problema

- ¿Tienes a alguien que pueda ayudarte a resolver ese tipo particular de historia analítica?
- ¿Has buscado si hay hallazgos o experiencias previas relacionadas con este problema, ya sea dentro o fuera de tu organización?
- ¿Has revisado tu definición de problema en base a lo que has aprendido de tu revisión de hallazgos previos?





Preparación de datos



- Encontrar los datos que necesitamos
- **Limpiar los datos**
- Reestructurar los datos
- Combinar fuentes de datos





Preparación de datos



30%-80% del tiempo puede destinarse a reorganizar, reestructurar, limpiar los valores nulos, duplicados y errores tipográficos



Analizar los datos



Descriptiva

¿Qué está pasando
o qué pasó?

Inferencial

¿Qué pasó con el resto?

Diagnóstico

¿Por qué está
pasando o pasó?

Predictiva

¿Qué es lo más probable que
pase?

Prescriptiva

¿Qué debemos hacer?



Presentar datos y tomar decisiones basadas en datos



- Comprender la audiencia
- Identificar los objetivos
- Seleccionar las visualizaciones adecuadas
- Contextualizar los datos
- Destacar puntos clave
- Fomentar la participación y el debate
- Tomar decisiones basadas en datos



Repaso



slideswith.com/PREPASOS3

Leer



Analiza



Medir



Describir





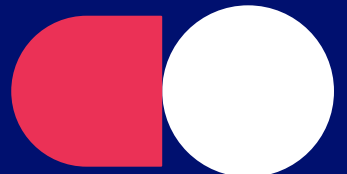
Respuestas a repaso



- ¿Existe una fórmula mágica única para limpiar los datos? R/ No
- Este no es un problema de limpieza de datos R/ El tamaño de la muestra
- Puedo realizar un análisis de datos sin realizar limpieza y tengo la seguridad de que los resultados no van a presentar errores R/ Falso
- Conocer el número de estudiantes matriculados el año pasado hace parte de R /
Análítica descriptiva
- Responder ¿Por qué es más probable que algunos estudiantes no caigan en bajo rendimiento? hace parte de la analítica R/ Diagnóstico
- ¿Qué familias tienen mayor probabilidad de no pagar a tiempo la matrícula? Esta pregunta la puedo resolver con la analítica R/ Predictiva



Limpieza de datos





¿Qué es limpiar los datos?



Definición:

Proceso de corregir o eliminar datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos dentro de un conjunto de datos.

Cuándo es común que ocurra:

- En una sola fuente de datos, especialmente en variables de texto y digitadas manualmente
- Al combinar varias fuentes de datos



¿Por qué es importante limpiar los datos?



- Si los datos son incorrectos, los resultados (desde las visualizaciones hasta los resultados de los algoritmos) no van a ser confiables, aunque parezcan correctos.
- El riesgo es alto si se piensa tomar decisiones con datos sin limpieza



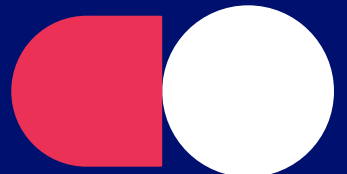
¿Existe una fórmula para limpiar los datos?



No existe una forma única que indique cuáles son los pasos exactos del proceso de limpieza de datos y cómo hacerlos porque los procesos variarán de un conjunto de datos a otro.

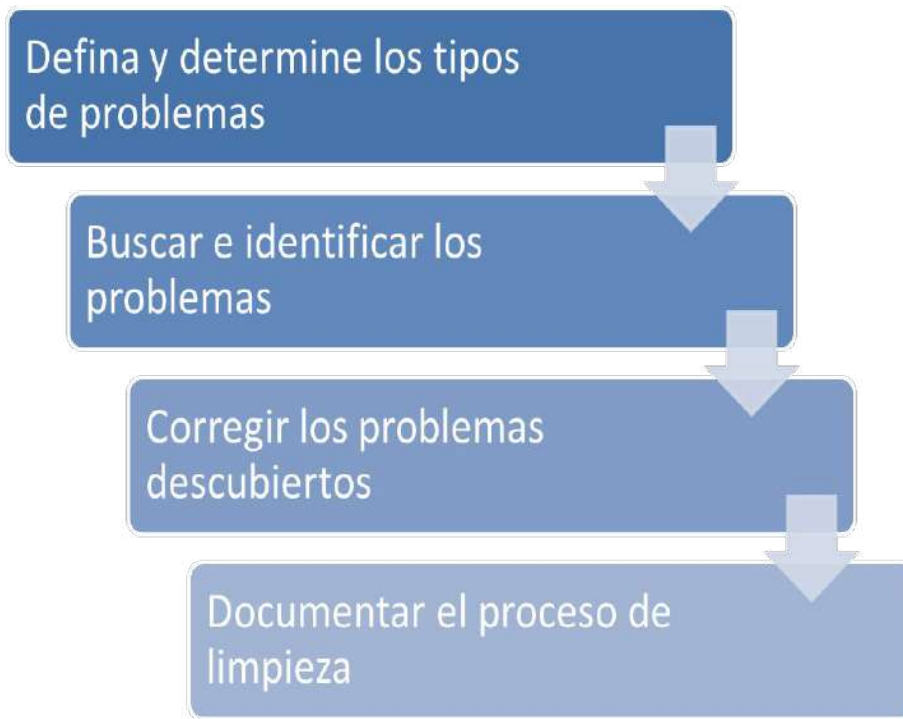


Framework limpieza de datos





Framework de limpieza de datos





Principales problemas



Observaciones duplicadas

Valores inconsistentes

Valores perdidos

Valores fuera del
diccionario de variables

Errores de digitación

Formato de la variable no
coincide con su tipo



Observaciones duplicadas



Observaciones duplicadas





Observaciones duplicadas



Eliminar observaciones duplicadas



Observaciones duplicadas



ID estudiante	Fecha	Actividad
123	15-may-2024	Clase de piano
123	15-may-2024	Psicología
123	15-may-2024	Fútbol

¿Está o no duplicada?

Depende de lo que queramos responder



Valores perdidos



Valores perdidos



- Identifique dónde se encuentran los valores perdidos y si es necesario recodifíquelos correctamente.
- Intente entender la naturaleza de los datos perdidos.
- Determine cuál es el mejor método para tratarlos

ID	Edad	Sexo	Ciudad
123	19	F	Lima
456		F	NA
789	24	999	Lima



Valores perdidos



Intente entender la naturaleza de los datos perdidos

Atrición por un proceso natural o social

Pérdida aleatoria de información

Patrón de salto en la encuesta

Falta intencional/sistemática como parte del
proceso de recopilación de datos



Determine cuál es el mejor método para tratarlos.

Reconocer su existencia y no efectuar tratamiento alguno. Pero tenerlo en cuenta en la aplicación de los modelos y el análisis.

Eliminar observaciones con valores perdidos

Imputación de datos: sustituir valores perdidos en un campo por otros empleando:

- Media
- Mediana
- Moda

Modelos estadísticos



Errores de digitación

Convenciones de denominación

Lima
Ciudad de Lima

Representaciones diferentes

Sí
Si
si
sí

13-03-2024
03-13-2024
3 mar 2024

Espacios vacíos

“Mujer”
“ Mujer ”
“Mujer“
“ Mujer”

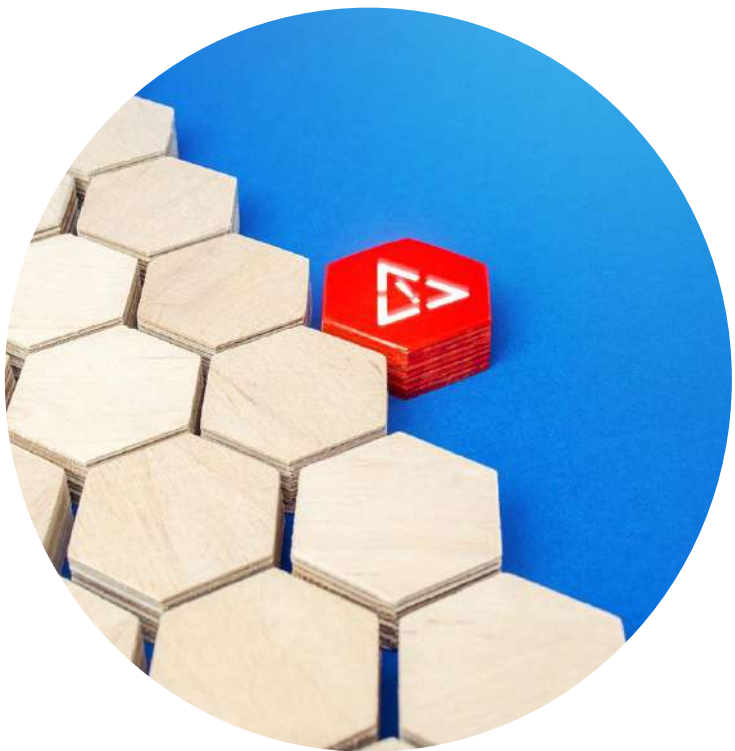


Errores de digitación



Soluciones

- Estandarice las opciones de respuesta desde la fuente de origen
- Intente recolectar los datos usando medios digitales
- Estandarice la digitación:
 - Uso de tildes
 - Uso de mayúsculas
 - Digitación de fechas
- Con los datos:
 - Utilice la función reemplazar para homogenizar respuestas y eliminar espacios vacíos



Valores inconsistentes



Valores inconsistentes



- Estos valores se hallan analizando la consistencia de las respuestas entre variables que se encuentran relacionadas-
- Analizando los rangos lógicos
- La detección de estos valores inconsistentes dependerá del conocimiento y la lógica de la persona

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
60	11/11/1956	60
47	7/04/1940	77



Valores inconsistentes



Edad	Número de hijos
10	9
15	0
25	1

Nombre	Estatura
Andrés	-1.5
Daniela	1.65
María	1.77



Valores inconsistentes



Posibles soluciones

- Convertir los valores inconsistentes en valores perdidos.
- Sin embargo, si tiene acceso a la fuente de datos original (encuesta), puede revisar las respuestas y corregir si es necesario.

Edad	Fecha de nacimiento	Edad calculada
27	1990/01/17	27
60	11/11/1956	60
47	7/04/1940	77



Valores por fuera del diccionario de datos



Un diccionario de datos es documento de referencia que nos sirve para identificar las variables que componen la base de datos. Nos da información de nombres, definiciones y atributos.

Por ejemplo:

- Nombre de la variable : Sexo
- Descripción: Hace referencia al sexo de la persona
- Valores que puede tomar
 - 1: Mujer
 - 0: Hombre



Valores por fuera del diccionario de datos



Nombre	Estado civil
Paola	1
Esteban	4
Cesar	5
Carlos	8
Mayra	2

Código Estado Civil	
1	Soltero
2	Casado
3	Separado o divorciado
4	Unión libre
5	viudo



Formato no coincide con el tipo de dato



Formato no coincide



Edad
18
41
23
50
17

- La edad es una variable cuantitativa y en debería tener un formato numérico, si tiene un formato de texto puede ocasionar problemas para realizar cálculos y estimar modelos.
- Todo depende de la herramienta empleada.
- Solución: verifique el formato y haga las correcciones pertinentes



Framework de limpieza de datos

Defina y determine los tipos de problemas



Buscar e identificar los problemas



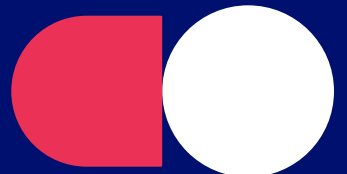
Corregir los problemas descubiertos



Documentar el proceso de limpieza



Actividad grupal





Actividad



1

Tamaño de grupos: unipersonal

2

Descripción: seleccione un excel con el que trabaje y reflexione:

¿Se ha presenta alguno de los problemas de limpieza descritos y qué hice para resolverlo?

¿Alguno de los problemas de limpieza descritos no los he revisado en mis datos y creo que podría hacerlo?

3

Tiempo de realización: 5 minutos

4

Socialización: 3 personas como mínimo



Gracias

