Limpieza y Transformacion

Latex

Oscar Painen Briones

2024

Universidad de Aysén

Contenidos

Integración

1 Integración

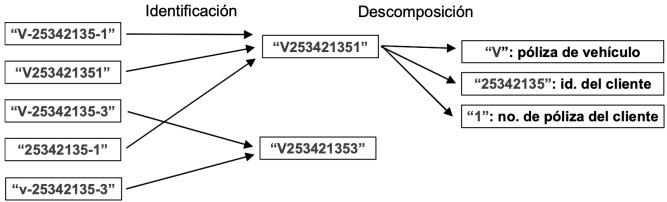
La **integración de datos** tiene como objetivo unificar información de diversas fuentes, asegurando que los datos del mismo objeto se combinen correctamente, y que los datos de objetos distintos se mantengan separados.

Errores comunes

1 Integración

- Unificación incorrecta de objetos distintos.
- Separación de datos del mismo objeto, lo que genera fragmentación.

Se ilustran ejemplos de estos desafíos, que muestra los problemas al identificar correctamente los objetos a partir de identificadores externos.



Errores comunes

1 Integración

Cuando se integran dos fuentes diferentes de datos de distintos objetos suele suceder que puedan aparecer datos faltantes o datos inconsistentes.

DNI	EDAD	COD.POSTAL	ESTADO	AÑOS_CARNÉ
...	....	...	...	...
25342135	35	46019	Casado	13
98525925	23	28004	Soltero	1
...	....	...	...	...

DNI	FECHA_NAC	CIUDAD	CASADO	CARNÉ
...	....	...	...	...
77775252	1/1/1950	Benitatxell	SÍ	A2
25342135	18/11/1971	Valencia	NO	B1
...	....	...	...	...

Fuente 1

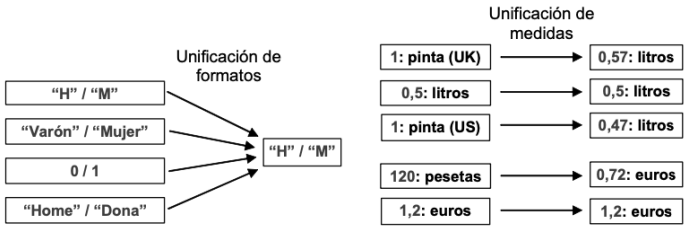
Fuente 2

DNI	EDAD	FECHA_NAC	CIUDAD	COD_POSTAL	ESTADO	CASADO	AÑOS_CARNÉ	CARNÉ
...		....	...			...		...
25342135	35	18/11/1971	Valencia	46019	Casado	NO	13	B1
98525925	23	-	-	28004	Soltero	-	1	-
77775252	-	1/1/1950	Benitatxell	-	-	SÍ	-	A2
...		....	...			...		...

Errores comunes

1 Integración

Ejemplos de unificación de formatos y medidas en el proceso de integración, como la transformación de códigos, géneros y unidades monetarias entre diferentes bases de datos.



Universidad de Aysén

Contenidos

2 Reconocimiento

Reconocimiento

2 Reconocimiento

Una vez integrados los datos, se realiza el **reconocimiento** de características a través de informes estadísticos que resumen atributos numéricos y nominales.
Estos informes son útiles para detectar problemas de calidad de datos como valores nulos o distribuciones anómalas.

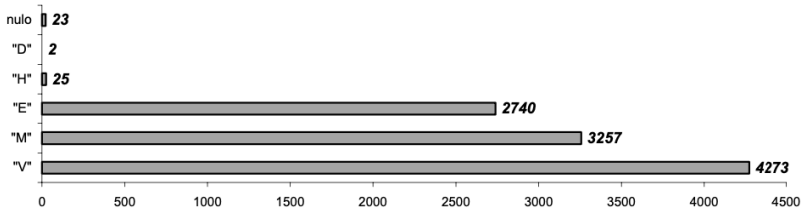
Resumen de atributos de una base de datos

La Tabla resume las características de una base de datos: número total de datos, valores nulos, distintos, media, moda y desviación estándar.

Atributo	Tabla	Tipo	# total	# nulos	# dists	Media	Desv.e.	Moda	Min	Max
Código postal	Cliente	Nominal	10320	150	1672	-	-	"46003"	"01001"	"50312"
Sexo	Cliente	Nominal	10320	23	6	-	-	"V"	"E"	"M"
Estado civil	Cliente	Nominal	10320	317	8	-	-	Casado	"Casado"	"Viudo"
Edad	Cliente	Numérico	10320	4	66	42,3	12,5	37	18	87
Total póliza p/a	Póliza	Numérico	17523	1325	142	737,24€	327€	680€	375€	6200€
Asegurados	Póliza	Numérico	17523	0	7	1,31	0,25	1	0	10
Matrícula	Vehículo	Nominal	16324	0	16324	-	-	-	"A-0003-BF"	"Z-9835-AF"
Modelo	Vehículo	Nominal	16324	1321	2429	-	-	"O. Astra"	"Audi A3"	"VW Polo"
...	...	...	...	...	...	...	...	...	...	...

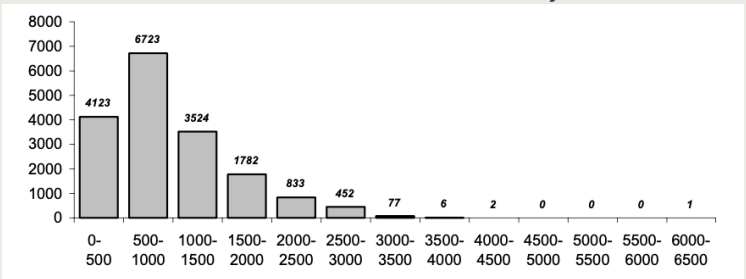
Histograma de distribución de frecuencias

El **histograma** representa la distribución de frecuencias del atributo ”sexo”, mostrando posibles valores anómalos.



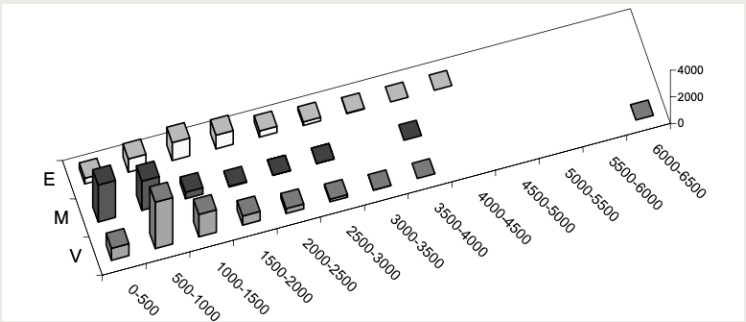
Otras representaciones gráficas de atributos

Distribución de valores anómalos en otro conjunto de datos.



Reconocimiento

2 Reconocimiento

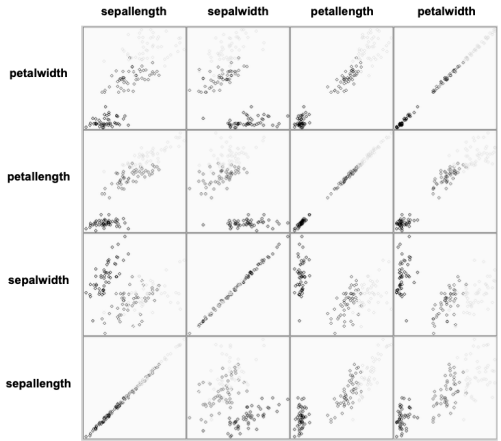


Histograma de otro atributo relevante.

Reconocimiento

2 Reconocimiento

Matriz de gráfica de dispersión mostrando la correlación entre diferentes características de flores.



Universidad de Aysén

Contenidos

Valores faltantes

3 Valores faltantes o Erroneos

Los **valores faltantes** son aquellos que no están presentes en los datos y pueden generar sesgos en el análisis si no se gestionan adecuadamente.
Es importante identificar si estos valores faltantes son relevantes o no existen en la realidad, y decidir si deben ser reemplazados o eliminados.

Universidad de Aysén

Valores erróneos

3 Valores faltantes o Erroneos

Los **valores erróneos** o anómalos pueden distorsionar los resultados del análisis. La detección de estos valores depende del formato y el origen de los datos.
Un ejemplo típico es el formato incorrecto de una matrícula de vehículo o la asignación incorrecta de fechas en registros históricos.

Universidad de Aysén

Valores faltantes o erróneos

3 Valores faltantes o Erroneos

Acciones sobre estos datos

- **Ignorar**, algunos algoritmos no tienen problemas con datos faltantes.
- **Eliminar** toda la columna, depende de la proporción de nulos. **Solución extrema.**
- **Filtrar** la fila provoca sesgo en los datos.
- **Reemplazar** el valor de forma manual o por algún valor que preserve la media/varianza. Se puede predecir.
- Modificar la política de datos.

Universidad de Aysén

Limpieza y Transformacion

Thank you for listening!