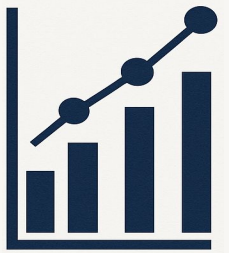
PREPARACIÓN DE DATOS

Integración - Limpieza - Transformación

¿QUÉ VEREMOS HOY?



1. Integración
2. Reconocimiento
3. Limpieza

INTEGRACIÓN DE DATOS

Recopilación de datos de distintas fuentes en un único lugar

Problemas:

- Definición de tipo de datos
- Unión de los mismos datos desde distintas fuentes
- Unificación de formatos

Identificación y descomposición

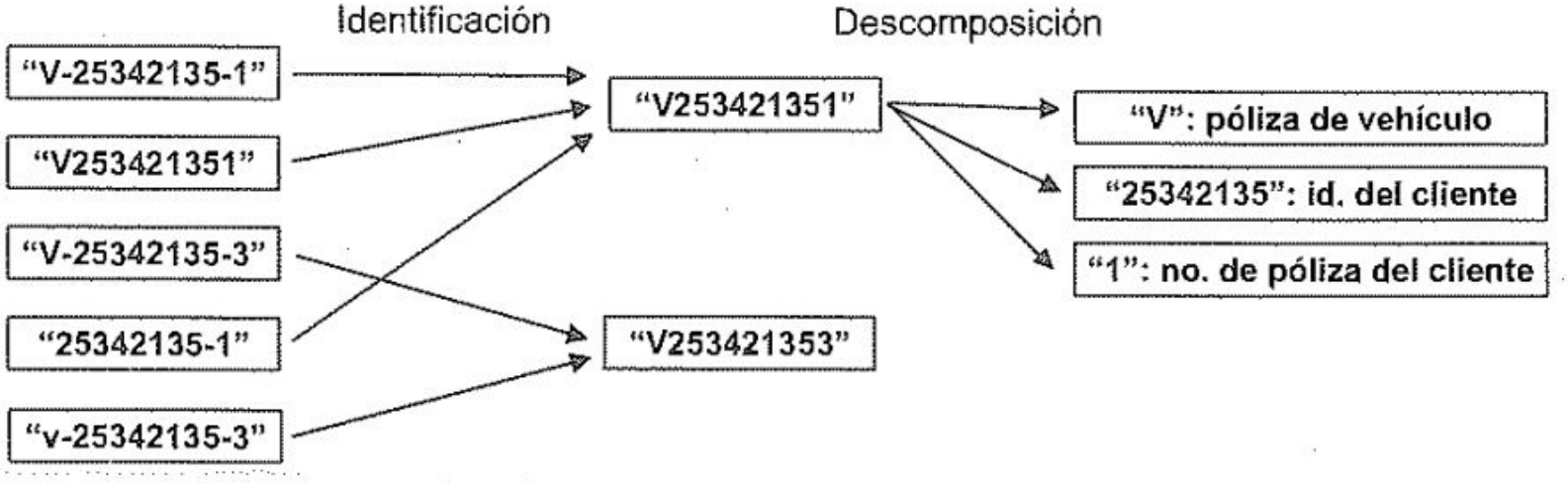


Figura 4.1. Ejemplos de integración: identificación y descomposición.

Combinación de datos desde distintas tablas

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
...		....	...			...		...

Figura 4.2. Ejemplos de integración de atributos de distintas fuentes.

Unificación de formatos y medidas

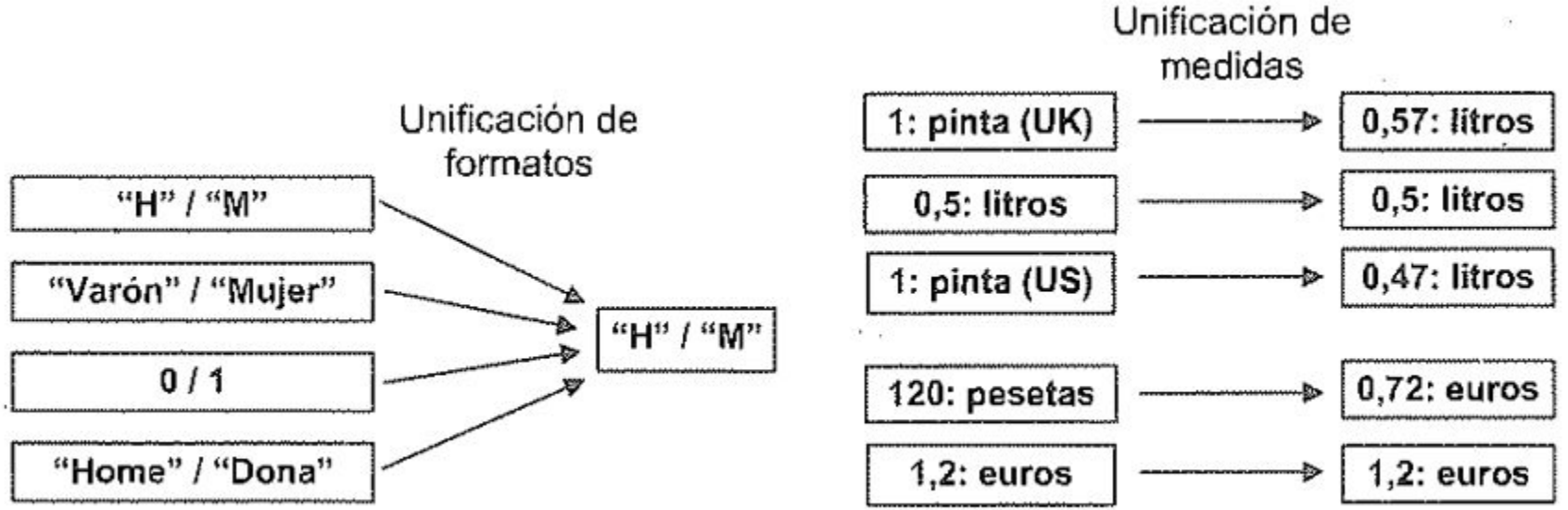


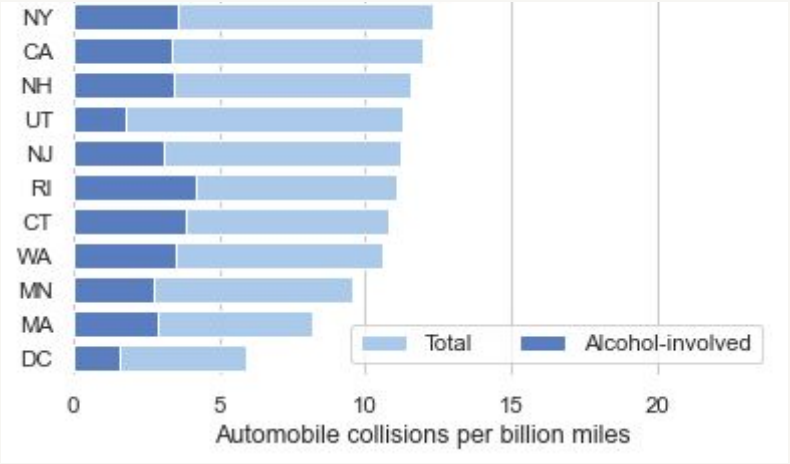
Figura 4.3. Ejemplos de integración: unificación de formatos y medidas.

RECONOCIMIENTO

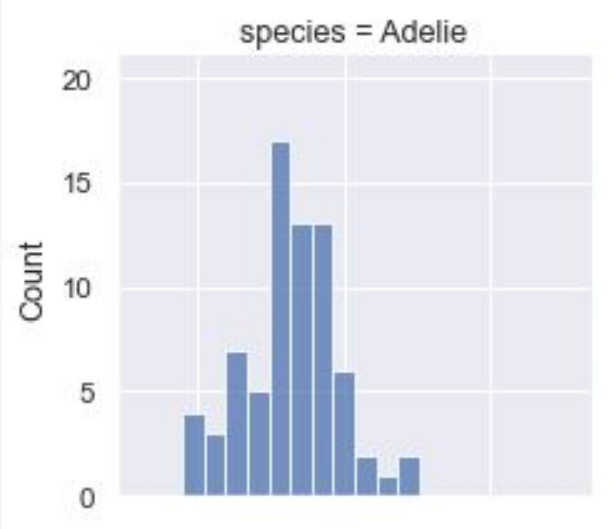
Reconocer el **comportamiento** de los atributos integrados mediante tablas resumen y gráficos

Tabla resumen

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"
...	...	...	...	...	...	...	...	...	...	...



Gráficos de barra



histogramas

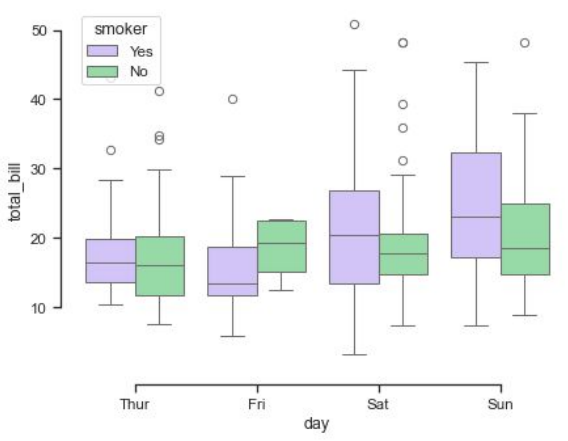
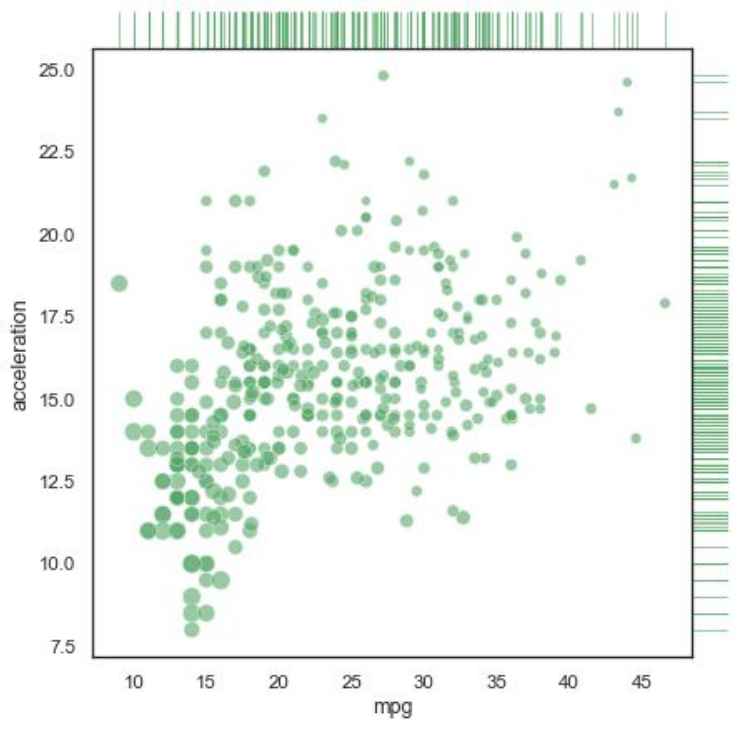
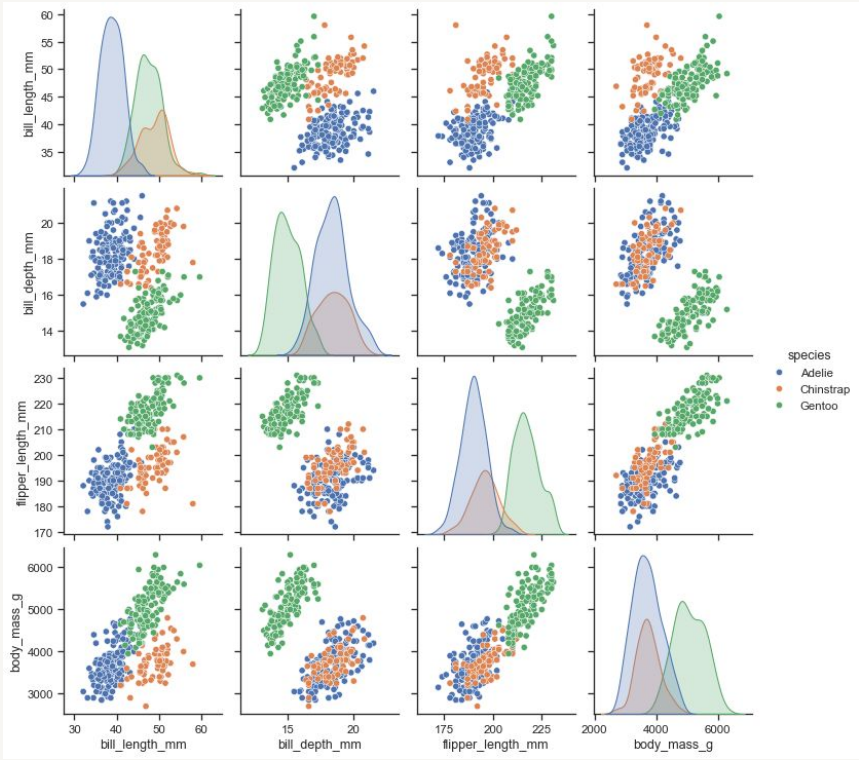


diagrama de caja

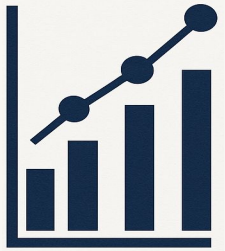
diagramas de dispersión



matriz de gráficas



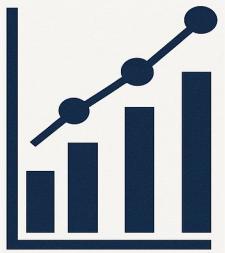
LIMPIEZA DE DATOS



En este paso debemos detectar y solucionar los problemas que presentan los datos después de integrarlos

- Selección de atributos
- Valores duplicados
- Valores faltantes
- valores erróneos
- Outliers

LIMPIEZA DE DATOS



SELECCIÓN DE ATRIBUTOS

id	nombre	apellido	dni	fecha_nac	edad	sexo	cp	localidad	cant vehículos
1	Juan	Gomez	47512698	05/08/03	22	M	2000	Rosario	0
2	María	López	48510680	25/02/04	21	F	2132	Funes	0
3	Lisa	Pérez	47695717	11/09/03	22	F	2000	Rosario	0
4	José	Díaz	48614410	13/04/04	21	M	2152	Granadero Baigorria	0

Los ID no aportan valor

La fecha de nacimiento y la edad aportan la misma información

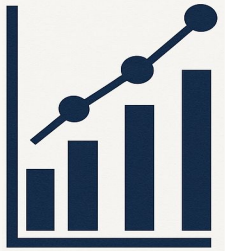
Si la columna tiene un único valor, entonces no aporta valor

id	nombre	apellido	dni	fecha_nac	edad	sexo	cp	localidad	cant vehículos
1	Juan	Gomez	47512698	05/08/03	22	M	2000	Rosario	0
2	María	López	48510680	25/02/04	21	F	2132	Funes	0
3	Lisa	Pérez	47695717	11/09/03	22	F	2000	Rosario	0
4	José	Díaz	48614410	13/04/04	21	M	2152	Granadero Baigorria	0

Normalmente nombre, apellido y dni no tienen importancia

Si los códigos refieren a objetos con más información (claves foráneas), entonces integrar estos datos asociados

LIMPIEZA DE DATOS

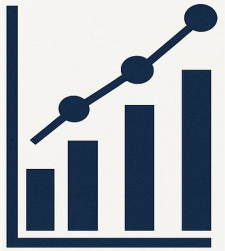


VALORES DUPLICADOS

Acciones:

- Validar que son registros duplicados. En ese caso deben ser eliminados.
- Si son registros independientes que contienen los mismos valores, entonces se deben mantener.

LIMPIEZA DE DATOS

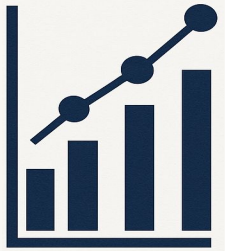


VALORES FALTANTES

Acciones:

- Ignorar (no realizar ninguna acción)
- Eliminar la columna (filtrar o reemplazar)
- Filtrar la fila (filtrar o reemplazar)
- Reemplazar valor:
  - con la media/mediana (valores numéricos)
  - con la moda (valores categóricos)
  - con valor aleatorio
- Segmentar
- Esperar

LIMPIEZA DE DATOS



VALORES ERRÓNEOS

Casos:

Patente
SA561HK

Sexo
H

Edad
236

año
2998

¿Qué hacer?

Se trata como un valor faltante y se resuelve como vimos anteriormente

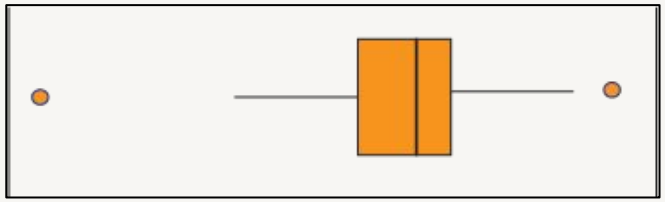
LIMPIEZA DE DATOS

OUTLIERS

Valores estadísticamente atípicos que son correctos pero complican el análisis de los datos

$$Limite\ Sup = Q_3 + 1.5 \cdot IQR$$

$$Limite\ Inf = Q_1 - 1.5 \cdot IQR$$



$$Outliers = \mu \pm 3 \cdot \sigma$$