



Tema 6. Calidad de datos y limpieza

Índice



- 1 **Introducción**
- 2 **Medidas de calidad en los datos**
- 3 **Orígenes de problemas en datos**
- 4 **Proceso de limpieza de datos**
- 5 **Beneficios de la calidad de datos**

Introducción - Clases de datos

1 Categóricos - Cualitativos

Son todos aquellos que contestan a la pregunta '¿Cuál?' – '¿Cuáles?'.

En este grupo se encuentran los datos que indican color, sentimiento, experiencias, datos de entrevistas, etc.

2 Numéricos - Cuantitativos

Datos que se refieren a un número.

Ingresos anuales, edad o peso de una persona. Calificación en un examen.

Introducción - Clases de datos - Categóricos

1 Nominal

Descripción de una categoría. Podemos utilizar un **adjetivo** para definir la categoría. Los valores nominales NO tienen un orden.

Un ejemplo podría ser la nacionalidad de una persona. 'español', 'inglés', 'alemán', etc.

2 Ordinal

Valores que tienen un orden en sí. Por ejemplo: Riesgo 'alto', 'medio', 'bajo'.

3 Binario

Un caso especial y muy utilizado ya que representa valores 'sí' – 'no' o 'verdadero' – 'falso'.

Introducción - Clases de datos - Numérico

1 Discretos

Datos numéricos enteros que solo se pueden expresar con una cifra. Por ejemplo, el número de votos que ha tenido un país en Eurovisión. Puede tener 100 ó 101 puntos, pero no 100,23 puntos.

2 Continuos

Datos numéricos que pueden recibir cualquier valor. Por ejemplo, la altura o el peso de una persona.

















- a Intervalo** Variables numéricas cuyos valores representan magnitudes y la distancia entre números en su escala es igual
- b Razón** Como el anterior, pero cuentan con un cero absoluto. Por ejemplo, nº de desviaciones de la media.

Clasificación pilotos F1

		Clasificación		Nombre	Nacionalidad				
1º	PUNTOS 161			NED Max Verstappen RED BULL (268)		ÚLTIMO GP 1º	VICTORIAS 5	POLES 7	MEJOR PUESTO 1º (x 5)
2º	PUNTOS 113			MON Charles Leclerc FERRARI (212)		ÚLTIMO GP 3º	VICTORIAS 0	POLES 0	MEJOR PUESTO 2º
3º	PUNTOS 107			MEX Sergio Pérez RED BULL (268)		ÚLTIMO GP 8º	VICTORIAS 0	POLES 0	MEJOR PUESTO 2º (x 3)
4º	PUNTOS 101			GBR Lando Norris MCLAREN (154)		ÚLTIMO GP 2º	VICTORIAS 1	POLES 0	MEJOR PUESTO 1º
5º	PUNTOS 93			ESP Carlos Sainz FERRARI (212)		ÚLTIMO GP 5º	VICTORIAS 1	POLES 0	MEJOR PUESTO 1º
6º	PUNTOS 53			AUS Oscar Piastri MCLAREN (154)		ÚLTIMO GP 4º	VICTORIAS 0	POLES 0	MEJOR PUESTO 4º (x 3)
7º	PUNTOS 44			GBR George Russell MERCEDES AMG F1 (79)		ÚLTIMO GP 7º	VICTORIAS 0	POLES 0	MEJOR PUESTO 5º
8º	PUNTOS			GBR		ÚLTIMO GP	VICTORIAS	POLES	MEJOR PUESTO

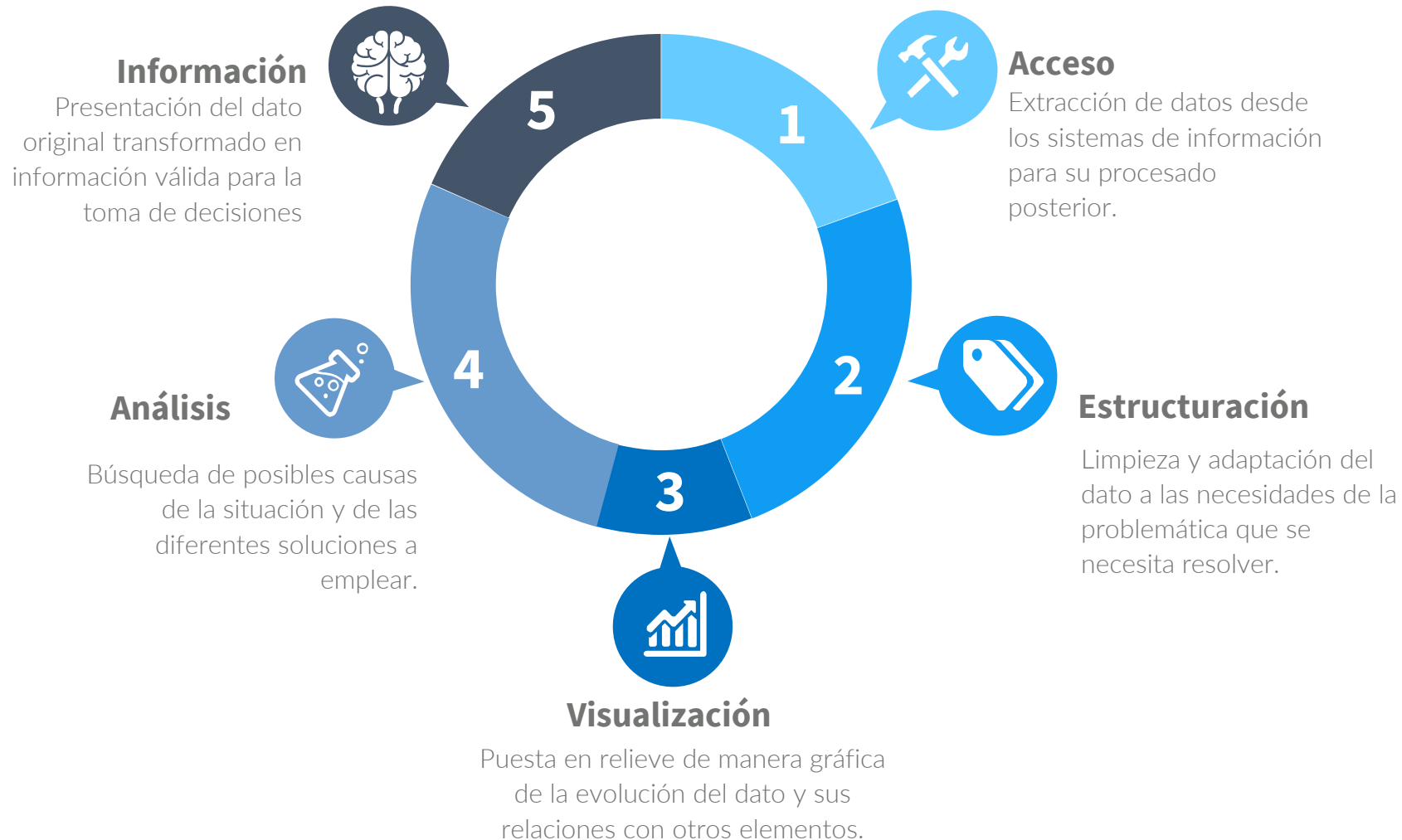
Clasificación pilotos F1

Clasificación	Nombre	Nacionalidad
---------------	--------	--------------

1º	PUNTOS 161		 NED Max Verstappen RED BULL (268)	ÚLTIMO GP 1º	VICTORIAS 5	POLES 7	MEJOR PUESTO 1º (x 5)
2º	PUNTOS 113		 MON Charles Leclerc FERRARI (212)	ÚLTIMO GP 3º	VICTORIAS 0	POLES 0	MEJOR PUESTO 2º
3º	PUNTOS 107		 MEX Sergio Pérez RED BULL (268)	ÚLTIMO GP 8º	VICTORIAS 0	POLES 0	MEJOR PUESTO 2º (x 3)
4º	PUNTOS 101		 GBR Lando Norris MCLAREN (154)	ÚLTIMO GP 2º	VICTORIAS 1	POLES 0	MEJOR PUESTO 1º
5º	PUNTOS 93		 ESP Carlos Sainz FERRARI (212)	ÚLTIMO GP 5º	VICTORIAS 1	POLES 0	MEJOR PUESTO 1º
6º	PUNTOS 53		 AUS Oscar Piastri MCLAREN (154)	ÚLTIMO GP 4º	VICTORIAS 0	POLES 0	MEJOR PUESTO 4º (x 3)
7º	PUNTOS 44		 GBR George Russell MERCEDES AMG F1 (79)	ÚLTIMO GP 7º	VICTORIAS 0	POLES 0	MEJOR PUESTO 5º
8º	PUNTOS		 GBR	ÚLTIMO GP	VICTORIAS	POLES	MEJOR PUESTO

Introducción – Proceso toma decisiones

Convertimos datos en información para la toma de decisiones



MEDIDAS DE CALIDAD DE DATOS



Medidas de calidad de datos

- + Precisión
- + Completitud
- + Consistencia
- + Actualidad
- + Unicidad

PRECISIÓN

Exactitud de los datos en relación con la realidad, indicando en qué medida se alinean con la información del mundo real que intentan representar.

Un alto nivel de precisión garantiza que los datos sean confiables y representativos de la situación real.

En una base de datos de registros médicos, la precisión se refiere a la corrección de los diagnósticos y tratamientos registrados.



La cantidad de datos disponibles en relación con los datos esperados. Los datos incompletos pueden surgir debido a errores en la recolección de datos, información sin relevancia temporal o limitaciones del sistema.

En una base de datos de clientes, la integridad se refiere a la presencia de valores para todos los campos obligatorios, como nombre, dirección y número de teléfono, en todos los registros de clientes

CONSISTENCIA

La uniformidad y coherencia de los datos a través de diferentes fuentes.

La inconsistencia puede producirse por variaciones, convenciones de nombre o definiciones de datos en conflicto.

En una base de datos de inventario, la consistencia se refiere a la uniformidad de las unidades de medida utilizadas.



ACTUALIDAD

La relevancia temporal de los datos en relación con el análisis y su marco temporal.

En una base de datos de transacciones financieras, la actualidad se refiere a la fecha en que se registraron las transacciones.



UNICIDAD

La ausencia de duplicados en los datos.

En una base de datos de empleados, la unicidad se refiere a la presencia de un solo registro para cada empleado.



Orígenes de problemas en datos



+ Errores en la introducción de datos

Despistes en la entrada manual de datos conducen a errores como faltas de ortografía, de tipografía o valores incorrectos. El error humano, la falta de formación o mecanismos de validación inadecuados contribuyen a este tipo de errores



+ Datos incompletos o falta de datos

Los datos pueden estar incompletos por varias razones. La falta de un proceso adecuado de recolección, omisiones en la entrada de datos o limitaciones del sistema son algunos de ellos.

Orígenes de problemas en datos



+ Transformación y manipulación

Transformaciones de datos, como las agregaciones, cálculo o conversión de datos pueden introducir errores si no se implementan de manera correcta. Funciones incorrectas, asunciones erróneas o errores en el proceso de manipulación de datos son causas frecuentes de este tipo de errores.



+ Integración

Al combinar datos de múltiples fuentes, las inconsistencias pueden surgir como consecuencia de diferentes formatos o estructuras de datos.

Orígenes de problemas en datos



+ Almacenamiento y transferencia

Sistemas de almacenamiento poco confiables llevan a pérdidas o corrupción de datos o accesos no autorizados. Eventos como fallos en los sistemas o el hardware pueden resultar en pérdida de datos.



+ Gobernanza y documentación

Inadecuadas prácticas de gobernanza, junto con la falta de estándares o definiciones de datos, la falta de documentación puede producir graves problemas de datos.

Orígenes de problemas en datos



+ Cambios y actualizaciones

Conforme el tiempo pasa, los datos evolucionan, hay cambios en las reglas de negocio y los sistemas se actualizan. Todo ello impacta sobre la calidad de datos.



+ Datos externos

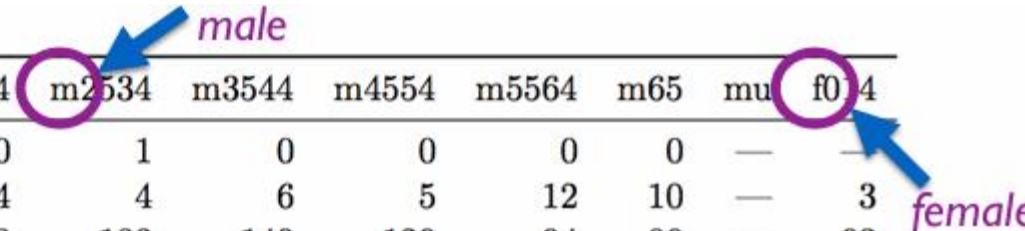
Al combinar datos de múltiples fuentes, las inconsistencias pueden surgir como consecuencia de diferentes formatos o estructuras de datos.

Otros orígenes de problemas en datos

- + Encabezados de columnas son valores, no nombres de variables.
- + Múltiples variables almacenadas en una sola columna.
- + Variables almacenadas tanto en filas como en columnas.
- + Una observación aparece almacenada en más de una tabla.

Otros orígenes de problemas en datos

- Encabezados de columnas son valores, no nombres de variables.



country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f0
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, f1524, f2534 and so on. These are not shown to conserve space. Note the mixture of 0s and missing values (—). This is due to the data collection process and the distinction is important for this dataset.

Otros orígenes de problemas en datos

+ Múltiples variables almacenadas en una sola columna.

042023051019001000001001SRAQUEL	PINEDA	GARCIA	F00000000	N
042023051019001000001001TNOEMI	AGUIRRE	QUINTANA	F00000000	S
042023051019001000001002SANTON	BILBAO	RUIZ	M00000000	N
042023051019001000001002TIÑAKI	ARRIETA	PEREZ	M00000000	S
042023051019001000001003SEDUARDO	REDONDO	SAEZ DE BURUAGA	M00000000	N
042023051019001000001003TIGONE	MARTINEZ DE LUNA	UNANUE	F00000000	S
042023051019001000001004SMIKEL GOTZON	ALUTIZ	PEREZ	M00000000	N
042023051019001000001004TMARIA JOSE	ELICES	PEREZ	F00000000	S
042023051019001000001005TRICARDO	GONZALEZ DE HEREDIA	LOPEZ DE VICUÑA	M00000000	S
042023051019001000001006TJUDITH	RUIZ DE AZUA	GONZALEZ DE MENDIBIL	F00000000	N
042023051019001000001007TGORKA	PRADAS	RUBIO	M00000000	N
042023051019001000001008TADELI	FERNANDEZ	FERNANDEZ	F00000000	N
042023051019001000001009TMARTA	RUIZ DE ARCAUTE	CORREO	F00000000	N
042023051019001000001010TOSCAR	CADAVID	LOPEZ DE LUZURIAGA	M00000000	N
042023051019001000001011TMERTXE	RUIZ DE ARGANDOÑA	VILLAR	F00000000	N
042023051019001000002001SDANIEL SEBASTIAO	M'DONBAXE	MALONGUI	M00000000	N
042023051019001000002001TALBERTO	LASARTE	BOVEDA "TITO"	M00000000	S
042023051019001000002002SLUIS	ALDAY	ALZOLA	M00000000	N
042023051019001000002002TELISABET	ZUBIZARRETA	ARRUABARRENA	F00000000	S
042023051019001000002003SAINHOA	ALTUNA	GARCIA DE SALAZAR	F00000000	N
042023051019001000002003TALEJANDRO	CRESPO	ANTOLIN	M00000000	N
042023051019001000002004TITSASO	MUSITU	DOMINGUEZ	F00000000	N
042023051019001000002005TAITOR	SAN MARTIN	ALANGUA	M00000000	N
042023051019001000002006TIZASKUN	ARRATIBEL	MURGUIONDO	F00000000	N
042023051019001000002007TRAUL	MUSITU	IRIGOYEN	M00000000	N
042023051019001000002008TALMIKE	URIBARRI	ARAMENDI	F00000000	N
042023051019001000002009TJESUS MARIA	DELGADO	GARCIA	M00000000	N
042023051019001000002010TAYALA	ALONSO	PEREZ DE VILLARREAL	F00000000	N
042023051019001000002011TRAMON	URTURI	PEREZ DE ARRILUCEA	M00000000	N

Otros orígenes de problemas en datos

- + Variables almacenadas tanto en filas como en columnas.

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

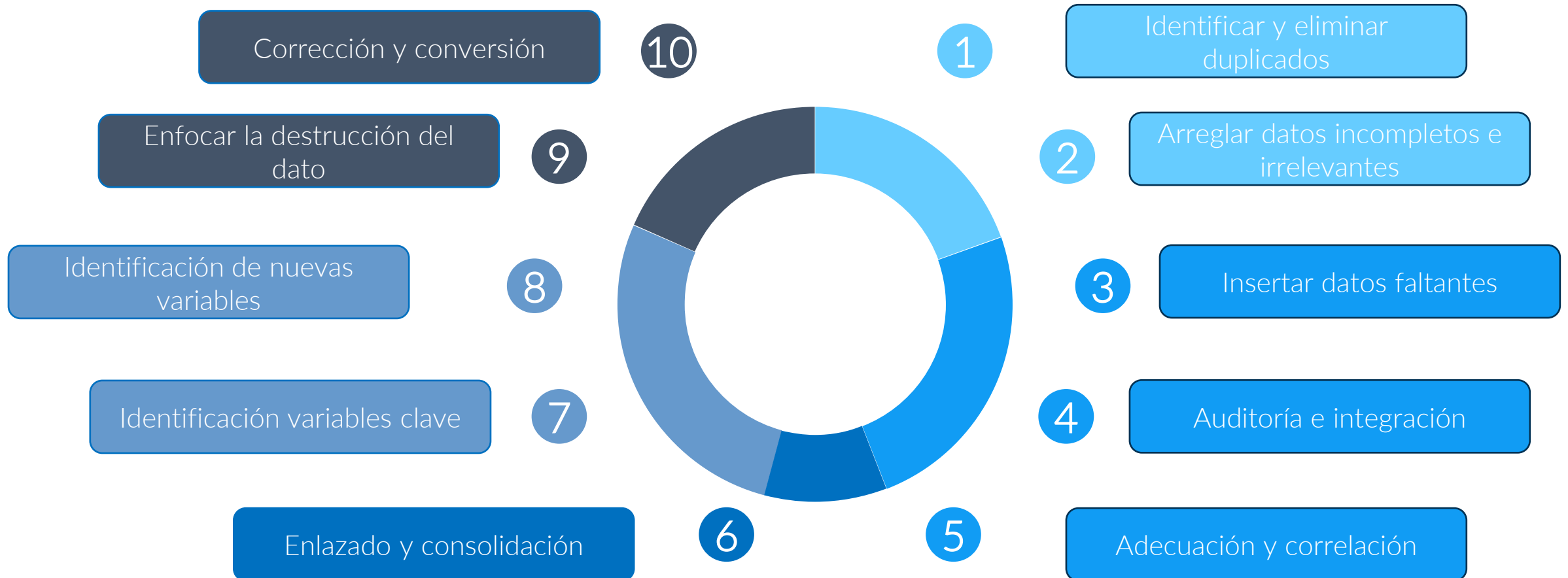
Otros orígenes de problemas en datos

- + Una observación aparece almacenada en más de una tabla.

- Song				- Rank each week		
id	artist	track	time	id	date	rank
1	2 Pac	Baby Don't Cry	4:22	1	2000-02-26	87
2	2Ge+her	The Hardest Part Of ...	3:15	1	2000-03-04	82
3	3 Doors Down	Kryptonite	3:53	1	2000-03-11	72
4	3 Doors Down	Loser	4:24	1	2000-03-18	77
5	504 Boyz	Wobble Wobble	3:35	1	2000-03-25	87
6	98~0	Give Me Just One Nig...	3:24	1	2000-04-01	94
7	A*Teens	Dancing Queen	3:44	1	2000-04-08	99
8	Aaliyah	I Don't Wanna	4:15	2	2000-09-02	91
9	Aaliyah	Try Again	4:03	2	2000-09-09	87
10	Adams, Yolanda	Open My Heart	5:30	2	2000-09-16	92
11	Adkins, Trace	More	3:05	3	2000-04-08	81
12	Aguilera, Christina	Come On Over Baby	3:38	3	2000-04-15	70
13	Aguilera, Christina	I Turn To You	4:00	3	2000-04-22	68
14	Aguilera, Christina	What A Girl Wants	3:18	3	2000-04-29	67
15	Alice Deejay	Better Off Alone	6:50	3	2000-05-06	66

Table 13: Normalised billboard dataset split up into song dataset (left) and rank dataset (right). First 15 rows of each dataset shown; **genre** omitted from song dataset, **week** omitted from rank dataset.

Proceso de limpieza de datos



Técnicas de limpieza de datos

1 Eliminación de duplicados

Identificación y eliminación de filas duplicadas en función de los atributos seleccionados.

El proceso de limpieza de datos implica fusionar o eliminar dichos duplicados para garantizar informes de ventas precisos

2 Imputación de datos

Cuando faltan datos de alguna de las variables, hay que decidir si se elimina el registro completo o se pueden crear datos basados en otros registros.

El proceso de limpieza de datos implica fusionar o eliminar dichos duplicados para garantizar informes de ventas precisos

Técnicas de limpieza de datos

3 Corrección de datos incorrectos

Utilizar reglas de validación de datos, controles de coherencia y revisión manual si es necesario.

El objetivo principal detrás de la validación de datos es comprobar que están destinados al uso previsto.

4 Manejo de datos atípicos

Identificando valores atípicos mediante métodos estadísticos como la puntuación Z o el IQR, luego decida si limitarlos, transformarlos o eliminarlos.

Técnicas de limpieza de datos

5 Validación de coherencia de datos

Utilizar reglas de validación de dato para comprobar las relaciones y la coherencia entre atributos.

En una BBDD de inventario, validar que el valor total del stock coincida con la suma de los valores de los artículos individuales.

6 Transformaciones de datos

Codificar datos categóricos o crear términos de interacción basados en necesidades analíticas.

En un sistema de recomendación, se aplica codificación one-hot a las categorías de productos para convertirlas a un formato adecuado para algoritmos de aprendizaje automático.



Samsung: Data Entry Error Cost \$105 Billion

In 2018, a Samsung Securities employee in South Korea made a “fat-finger” error mistaking won (South Korea’s currency) for shares, paying out 1,000 Samsung Securities shares to workers instead of 1,000 won per share in dividends.

That single human error cost the technology company \$300 million in the end! (For a short period of time, the company had issued a mind-blowing \$105 billion worth of shares, but that was fixed within 37 minutes, according to IEEE.) Ultimately, Samsung Securities paid dividends worth 1,000 times the value of each share to 2,018 of its employees.

Similar fat-finger problems can afflict any organization without protocols in place to protect itself. In the case of Samsung Securities, if an assurance process sent the data to another employee or automatically checked the range, they could’ve avoided the error. Had proper processes been in place before the employee paid out the shares, a prompt would have shown the error. The situation could have been averted and the loss avoided with some fairly simple data processes.



WSJ PRO VENTURE CAPITAL

Home News Data Sectors Newsletters

TECH

Uber Shortchanged New York City Drivers by Millions of Dollars

The ride-hailing company says it is refunding the money after miscalculating its commissions

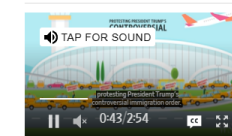
By Greg Bensinger

Updated May 23, 2017 2:36 pm ET | WSJ PRO

Share

Resize

33



From sexual harassment allegations to social media opposition, Uber's steered off the road. WSJ's Lee Hawkins explains the company's missteps. Photo/Video: Drew Evans/The Wall Street Journal (Originally published March 30, 2017)

Uber Technologies Inc. on Tuesday said it mistakenly underpaid New York City drivers for the past 2½ years, an error that will likely cost it tens of millions of dollars. It is the second time in three months the ride-hailing company has acknowledged it deprived workers of their proper earnings.

Under the terms of its November 2014 nationwide driver

agreement, Uber was meant to take its commission, generally 25%, from U.S. drivers based on fares after any taxes and fees were deducted. Uber said that, instead, in New York City it calculated a higher cut using the full fare before accounting for sales tax and a local injury-compensation fund fee.

Uber told The Wall Street Journal it would refund the money plus interest

TOP NEWS & ALERTS

Pete Sonsini, Early Investor in Databricks, Gets Closer to Launching New VC Firm



Emerging Form of Targeted Cancer Therapy Draws Venture Capital



Renegade Partners Nets \$200 Million in VC Deals



	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Errores fonéticos/ortográficos

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Errores fonéticos/ortográficos

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Errores fonéticos/ortográficos

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Errores fonéticos/ortográficos

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN \$	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Errores fonéticos/ortográficos

Valor equivocado

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Valor nulo

Errores fonéticos/ortográficos

Valor equivocado

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Valor nulo

Valor inexistente

Errores fonéticos/ortográficos

Valor equivocado

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Valor nulo

Valor inexistente

Errores fonéticos/ortográficos

Valor equivocado

Estructura
inconsistente

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Mezclar letras y números

Estructura inconsistente

Caracteres no imprimibles

Valor nulo

Valor inexistente

Errores fonéticos/ortográficos

Valor equivocado

Estructura
inconsistente

	CRM	REDES SOCIALES	HISTORIA CLÍNICA	CENSO	BBDD PROPIA
Identificación	CC 123-045-9	CC 123-045-9abc	1234-0897		CC1230459
ID	12453598	45354867	19098736	19900047-X	654897
Nombre	Juan Juse	JUAN S	Juan Sanchez	Juan Sanchez Toro	Sanchez Toro J.
Correo	jjst@gmail.comn		jjst81@gmail.com	Jjst81@gmal.com	jjsti81.com
Dirección	Pza. Mestre Ripoll 8	Plaza del Mestre Ripoll nº 8	P. M. Ripoll	Plaza del Mestre Ripoll 8, 50	8, Pz. Mestre Ripoll
Teléfono	687-081-900	687081900	P. M. Ripoll	+ 34687-081-900	xxx-xxx-xxx
Fecha Nacimiento	09/12/2000	9-Diciembre-00	2000/12/09		2000/09/12

Valor erróneo

Mezclar letras y números

Estructura inconsistente

Beneficios de la calidad de datos



+ Mejora en la eficiencia de los datos

La calidad de datos garantiza que los conjuntos de datos sean precisos y completos antes del análisis.

Esto genera datos sin errores que se necesitan para futuras investigaciones o entrenamiento del modelo de aprendizaje automático, en última instancia, ahorrando tiempo y recursos.

No solo ahorra tiempo y recursos, sino que incluso puede ayudar a evitar errores causados por datos incorrectos. Al identificar las imprecisiones desde el principio, las empresas pueden evitar que los errores empeoren y realizar cambios cruciales antes de que sea demasiado tarde.

Beneficios de la calidad de datos



+ Revela nuevos conocimientos

Ayuda a las empresas a descubrir patrones y relaciones ocultos en sus datos que pueden haber pasado desapercibidos anteriormente. Esto puede brindarles una comprensión más completa de sus operaciones y los factores que impulsan su éxito.

Con este conocimiento, las empresas pueden tomar mejores decisiones para generar crecimiento y rentabilidad.

Beneficios de la calidad de datos



+ Precisión mejorada

Los datos limpios conducen a conocimientos precisos. La limpieza de datos establece una base sólida para un análisis y una toma de decisiones precisos al eliminar errores y estandarizar formatos.

Los datos correctamente limpios agilizan el proceso de análisis, minimizando el tiempo dedicado a la detección y corrección de errores durante las etapas analíticas.