



Caso 134 Limpieza Profunda

01 ETL y Modelo de Datos. PQ Limpieza de datos de 1º grado, técnicas de depuración sobre filas y columnas: reducción, formateo y rellenar

Jose Ignacio González Gómez

Departamento de Economía, Contabilidad y Finanzas - Universidad de La Laguna

www.jggomez.eu

R24/25

Ejercicio Basado: *EL ADN de Power Query – Miguel Caballero y Fabian Torres*

Contenido

Resumen y objetivos	1
1 Presentación.....	2
2 Se Pide	4
3 Técnicas de depuración, principales ajustes propuestos	5
3.1 Configuración de la consulta	5
3.2 Detalles del proceso de depuración	5
3.2.1 Filas filtradas de la columna nº de pedidos.....	6
3.2.2 Dividir columna SKU por delimitador	6
3.2.3 Valor reemplazado Columna Categoría de Descuento y columna Ingresos	7

Resumen y objetivos

Power Query. Creando parámetros de conexión

Proceso Transformación: Entre algunos ajustes a realizar básicos como: Quitar filas superiores, encabezados promovidos, quitar columnas, quitar errores, columnas con nombre cambiado, tipo de datos cambiado, etc.. hay otros a prestar atención como: Filtrar filas que contienen errores o incongruencias. Dividir columna por nº caracteres y por delimitador. Reemplazar valores, poner En Mayúsculas Cada Palabra, etc..

1 Presentación

Disponemos de un fichero de datos (134_LimpiezaProfunda.xlsx) relacionados con la gestión de pedidos, como podemos observar es necesario llevar un detallado proceso de limpieza y transformación para poder disponer de un fichero de trabajo correctamente estructurado.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1															
2															
3															
4															
5															
6	Número de Pedido - Tip Compra	SKU	Categoría de Descuento	Ingresos	Unidades	Costo del Producto	Costo de Envío	Costo Empaque	Fecha de Envío	Fecha de Llegada	País	Ciudad	Código País		
19	Total Resumen:				133.4171201237										
20															
21															
22															
23															
24	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
27	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx
28	98094261Normal	CC01-2	Non	105.06	2	8.55	4.55	2	36509	36529	PARAGUAY	Asunción	PA		
29	98094260Normal	L01-5	Non	61.5	2	8.25	3.25	2	36510	36529	PERU	Lima	PE		
30	98094258Normal	CB01-2	Non	222.029999999999	1	7.99	4.99	3	36511	36528	PARAGUAY	Asunción	PA		
31	98094259Devolución	CB01-2	Non	666.089999999999	3	5.61	5.61	2	36511	36529	PARAGUAY	Asunción	PA		
32	98094257Normal	CC02-1	Non	80.16	3	9.99	3.99	3	36512	36528	URUGUAY	Montevideo	UR		
33	98094256Normal	L01-5	Non	30.75	1	5.14	6.14	3	36512	36528	URUGUAY	Montevideo	UR		

Ilustración 1

Es decir, nos encontramos con diversas casuísticas relacionadas con el proceso de depuración de datos como son: *filas y columnas vacías, filas de error, encabezados no correspondiente, totales y subtotales, columnas que no nos interesa, tipos de datos cruzados, categorías de dos campos combinadas, manejo de minúsculas y mayúsculas no apropiado*, entre otros, tal y como vemos en las ilustraciones correspondientes:

	A	B	C	D	E	F
1						
2						
3						
4						
5						
6	Número de Pedido - Tip Compra	SKU	Categoría de Descuento	Ingresos	Unidades	Costo del Producto
19	Total Resumen:			133.4171201237		
20						
21						
22						
23						
24	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
27	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx
28	98094261Normal	CC01-2	Non	105.06	2	
29	98094260Normal	L01-5	Non	61.5	2	
30	98094258Normal	CB01-2	Non	222.029999999999	1	

Ilustración 2

	G	H	I	J	K	L	M	N	O
1									
2									
3									
4									
5									
6	Costo de Envio	Costo Empaque	Fecha de Envio	Fecha de Llegada	País	Ciudad	Código País		
19									
20									
21	####	####	####	####	####	####	####		
22	####	####	####	####	####	####	####		
23	####	####	####	####	####	####	####		
24	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!		
27	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxxx	xxx. xxx	
28	4.55	2	36509	36529	PARaguay	Asunción	PA		
29	3.25	2	36510	36529	PERÚ	Lima	PE		
30	4.99	3	36511	36528	PARaguay	Asunción	PA		

Ilustración 3

A estos datos se deberá aplicar lo que conocemos como Limpieza de datos de primer grado, técnicas de depuración que incluye las siguientes tareas y acciones del tipo:

1. Reducción

- Exclusión de filas y filtrar por null, total, quitar filas superiores, quitar errores, etc..
- Separar y/o Quitar columnas...
- Promover encabezado

2. Formateo de Datos

- Transformación a nivel de columnas
- Reemplazar valores
- Dividir columnas
- Asignación de nombres

3. Rellenar

4. Transposición

En este caso, algunas otras consideraciones a prestar especial atención son:

- Filtrar filas en la columna pedidos, especialmente la que contienen errores o incongruencias.
- La columna Numero de pedido Tipo de compra debemos tenerla separada, por un lado Numero de Pedido y por otro el tipo de compra. Dividir columna por nº caracteres, columna Nº de pedidos
- Arreglar la columna SKU es el código de producto que solo debe contener todos los caracteres antes del guion, de forma que aquellos que están después del guion no valen, deben desaparecer, es decir CC01-2 debe quedar transformada en CC01. Dividir columna por delimitador Columna código SK
- Reemplazar valores
 - En la columna Descuentos se ha detectado un error que hay que corregir, Balck Fraday por Black Friday
 - Especial atención requiere la columna ingresos, encontraremos valores “extraños” (p.e 2,2203E+16) o que no corresponden porque son números extremadamente grandes y la empresa “vende” productos de precios que no

son tan elevados, esto ocurre porque la configuración del ordenador toma el punto (.) como separador de miles y en realidad necesitamos que sea separador de decimales. Propuesta cambiar el punto (.) por la coma (,) y remplazarlo a número decimal.

- La columna País debemos estandarizarla, los elementos que la componen están escritos con mayúsculas y minúsculas sin ningún orden por esta razón, se debe estandarizar.

2 Se Pide

Llevar a cabo el proceso de ETL de forma que el resultado final obtenido sea similar al presentado a continuación

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
	Número de Pedido	Tip Compra	SKU	Categoría de Descuento	Ingresos	Unidades	Costo del Producto	Costo de Envío	Costo Empaque	Fecha de Envío	Fecha de Llegada	País	Ciudad	Código País
1	98094261	Normal	CC01	Non	105,06	2	8,55	4,55	2	15/12/1999	04/01/2000	Paraguay	Asunción	PA
2	98094260	Normal	L01	Non	61,5	2	8,25	3,25	2	16/12/1999	04/01/2000	Perú	Lima	PE
3	98094258	Normal	CB01	Non	222,03	1	7,99	4,99	3	17/12/1999	03/01/2000	Paraguay	Asunción	PA
4	98094259	Devolución	CB01	Non	666,09	3	5,61	5,61	2	17/12/1999	04/01/2000	Paraguay	Asunción	PA
5	98094257	Normal	CC02	Non	80,16	3	9,99	3,99	3	18/12/1999	03/01/2000	Uruguay	Montevideo	UR
6	98094256	Normal	L01	Non	30,75	1	5,14	6,14	3	18/12/1999	03/01/2000	Uruguay	Montevideo	UR
7	98094255	Normal	CB01	Non	666,09	3	7,14	4,14	2	19/12/1999	08/01/2000	Argentina	Rosario	AR
8	98094253	Normal	L01	Non	61,5	2	8,58	3,58	2	20/12/1999	08/01/2000	Ecuador	Quito	EC
9	98094254	Normal	L02	Non	120,6	3	9,55	5,55	2	20/12/1999	04/01/2000	Venezuela	Caracas	VE

Ilustración 4

	A	B	C	D	E	F	G	H
	Número de Pedido	Tip Compra	SKU	Categoría de Descuento	Ingresos	Unidades	Costo del Producto	Costo de Envío
1	98094261	Normal	CC01	Non	105,06	2	8,55	4,55
2	98094260	Normal	L01	Non	61,5	2	8,25	3,25
3	98094258	Normal	CB01	Non	222,03	1	7,99	4,99
4	98094259	Devolución	CB01	Non	666,09	3	5,61	5,61
5	98094257	Normal	CC02	Non	80,16	3	9,99	3,99

	I	J	K	L	M	N
	Costo Empaque	Fecha de Envío	Fecha de Llegada	País	Ciudad	Código País
1	2	15/12/1999	04/01/2000	Paraguay	Asunción	PA
2	2	16/12/1999	04/01/2000	Perú	Lima	PE
3	3	17/12/1999	03/01/2000	Paraguay	Asunción	PA
4	2	17/12/1999	04/01/2000	Paraguay	Asunción	PA
5	3	18/12/1999	03/01/2000	Uruguay	Montevideo	UR

Ilustración 5

3 Técnicas de depuración, principales ajustes propuestos

3.1 Configuración de la consulta

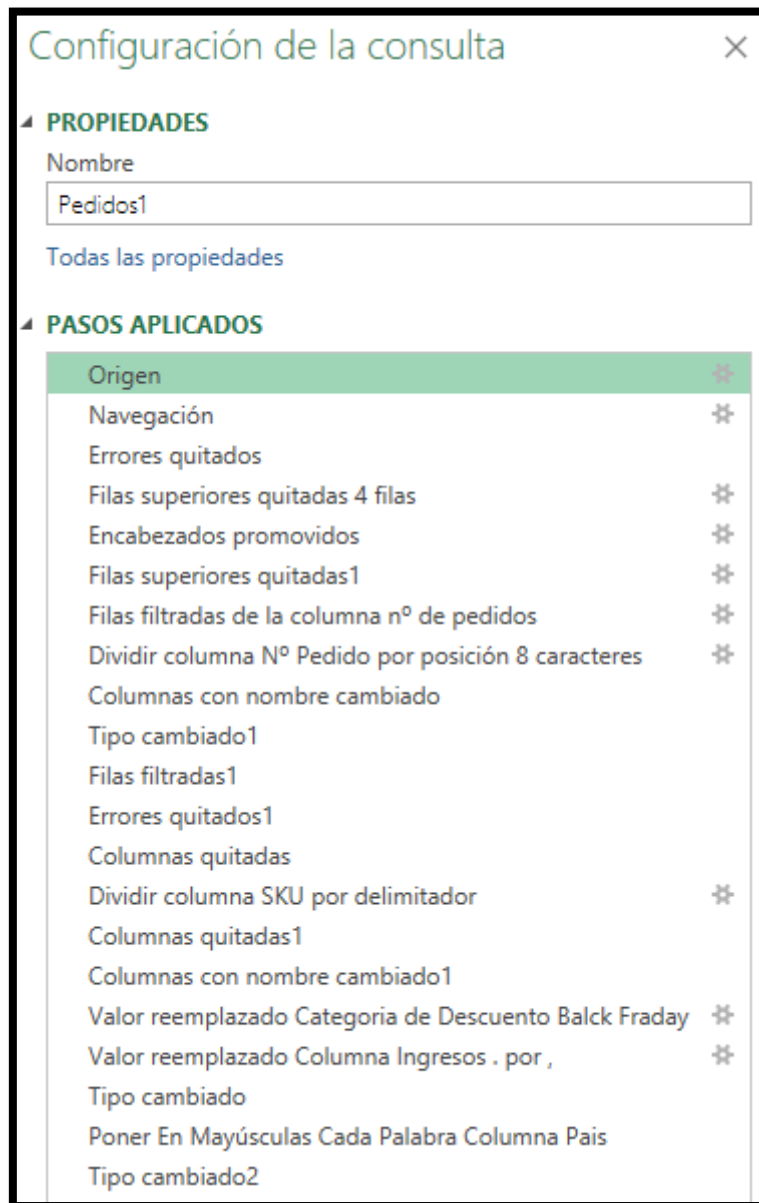


Ilustración 6

3.2 Detalles del proceso de depuración

Algunos detalles del proceso de depuración son:

3.2.1 Filas filtradas de la columna nº de pedidos

Filtrar filas

Aplique una o más condiciones de filtro a las filas de la tabla.

☐ Básico ☒ **Uso avanzado**

Conservar filas en las que

Y/o	Columna	Operador	Valor
	Número de Pedido...	no es igual a	null
☐ Y	Número de Pedido...	no es igual a	****
☐ Y	Número de Pedido...	no es igual a	Subtotal 1999, Primera Año
☐ Y	Número de Pedido...	no es igual a	Total Resumen:
☐ Y	Número de Pedido...	no es igual a	xxxx

Ilustración 7

3.2.2 Dividir columna SKU por delimitador

Dividir columna por delimitador

Especifique el delimitador utilizado para dividir la columna de texto.

Seleccione o escriba el delimitador

--Personalizado--

-

Dividir en

☐ Delimitador situado más a la izquierda

☐ Delimitador situado más a la derecha

☒ Cada aparición del delimitador

► Opciones avanzadas

Carácter de comillas

"

☐ Dividir con caracteres especiales

Insertar carácter especial

Ilustración 8

3.2.3 Valor reemplazado Columna Categoría de Descuento y columna Ingresos

Reemplazar los valores

Reemplace un valor con otro de las columnas seleccionadas.

Valor que buscar

Reemplazar con

▸ Opciones avanzadas

Ilustración 9

Reemplazar los valores

Reemplace un valor con otro de las columnas seleccionadas.

Valor que buscar

Reemplazar con

▸ Opciones avanzadas

Ilustración 10