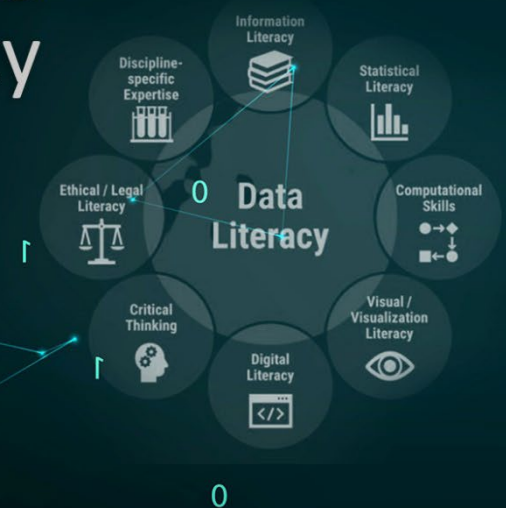




Alfabetización de Datos

Data Literacy



Limpieza de datos con OpenRefine: Guía práctica

Curso: Alfabetización de datos

Cursos de perfeccionamiento profesional – receso académico 2024

Organizado por: **Dirección de investigación - Universidad Tecnológica de Panamá**

Autor: Danny Murillo González

Centro de Investigación, Desarrollo e Innovación en Tecnologías de la Información y las Comunicaciones - CIDITIC

<https://orcid.org/0000-0003-0297-7213>

<https://scholar.google.com/citations?user=YNx08l0AAAAJ&hl=es>

Febrero 2024



OpenRefine



Esta obra está bajo licencia internacional Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0.

<https://ridda2.utp.ac.pa/handle/123456789/18247>

Acerca del curso

El material presentado está relacionado con el tema 4 del curso de Alfabetización de datos con una duración de 40 horas en formato virtual sincrónico organizado por la dirección de Investigación de la universidad Tecnológica de Panamá.

Objetivo: Desarrollar habilidades en la correcta utilización de los datos, abarcando fases como la normalización, manipulación y depuración de estos. Este aprendizaje permitirá realizar un adecuado análisis exploratorio de datos (AED), y posteriormente, transmitir de manera eficaz los resultados a través de visualizaciones comprensibles y dinámicas.

Descripción: La alfabetización de datos (Data Literacy) es la capacidad de explorar, comprender y comunicarse con datos de manera sencilla, significativa y precisa. Este curso contempla las bases de la alfabetización de datos a través de la normalización y evaluación de los tipos de datos, elemento esencial previos a la manipulación y exploración de datos. Se enseñarán habilidades en Data Clean, Tidy Data, Data Wrangling hasta la creación de visualizaciones con gráficos básicos e interactivos online. Ya sea que estés comenzando en el mundo de los datos o desees potenciar tus habilidades actuales, este curso te brindará una base sólida en el dominio de esta temática.

Contenido del Curso:

Tema 1: Introducción a los datos

Tema 2: Documentación de los datos

Tema 3: Manipulación y análisis de datos con Excel

Tema 4: Limpieza de datos con OpenRefine

Tema 5: Análisis Exploratorio de datos (AED)

Tema 6: Fundamentos de Visualización de datos

Tema 7: Gráficos en Datawrapper

Tema 8: Gráficos en Tableau

Contenido

Acerca del curso	2
Open refine.....	4
Carga de datos:.....	4
Crear proyecto.....	5
Operaciones de columnas	7
Facetas.....	10
Facetas de texto	10
Facetas numéricas	11
Facetas de línea de tiempo.....	11
Datos Duplicados (opción 1)	12
Datos Duplicados (opción 2)	17
Rellenar espacios en blanco	21
Unificar datos categóricos	23
Cluster (Agrupar)	29
Extraer datos numéricos de un campo de texto	32
Extraer un texto de una cadena de texto	37
Datos numéricos a categórico, condicional IF	39
Crear valores en una variable con datos de otras variables.....	44
Identificar si un texto esta dentro de una cadena de texto	39
Separar cadenas de texto en columnas.....	¡Error! Marcador no definido.
Guardar líneas de comandos en Open refine.....	48
Exportar documento	50
Bibliografía	51

Open refine

OpenRefine (anteriormente Google Refine) es una herramienta que dispone de un conjunto de características para trabajar con datos tabulares que mejoran la calidad general de un conjunto de datos. Se trata de una aplicación que se ejecuta fuera de su propia computadora como un pequeño servidor web, al que se accede desde un navegador web. Debe pensar en OpenRefine como una aplicación web personal y de acceso privado.

Página de otras opciones de descargas

<https://openrefine.org/download>

Instalación en Windows

Descargue Open refine de

https://openrefine.org/post_download?version=3.7.7&platform=win-with-java

1. Descomprima el archivo descargado.
2. Abra la carpeta **openrefine-win-with-java-3.7.7** y haga doble clic en openrefine.exe
3. Aparecerá una ventana de comando (que no debe cerrar) e inmediatamente después su navegador web mostrará una nueva ventana con la aplicación.]

Ventana de comandos que se habilita antes de mostrar la pantalla en el navegador con OpenRefine. Esta venta no debe cerrarse en ningún momento mientras se utilice la aplicación.

```

C:\Users\danny\Documents\ld x + -
08:20:47.411 [ refine] GET /command/core/get-preference (7ms)
08:20:48.830 [ refine] GET /command/core/get-csrf-token (1419ms)
08:20:48.834 [ refine] POST /command/core/annotate-one-row (4ms)
08:20:48.840 [ refine] GET /command/core/get-history (6ms)
08:20:48.846 [ refine] GET /command/core/get-preference (6ms)
08:20:50.541 [ refine] GET /command/core/get-csrf-token (1695ms)
08:20:50.545 [ refine] POST /command/core/annotate-one-row (4ms)
08:20:50.551 [ refine] GET /command/core/get-history (6ms)
08:20:50.563 [ refine] GET /command/core/get-preference (12ms)
08:20:51.765 [ refine] GET /command/core/get-csrf-token (1202ms)
08:20:51.768 [ refine] POST /command/core/annotate-one-row (3ms)
08:20:51.774 [ refine] GET /command/core/get-history (6ms)
08:20:51.786 [ refine] GET /command/core/get-preference (12ms)
08:20:53.853 [ refine] GET /command/core/get-csrf-token (2067ms)
08:20:53.856 [ refine] POST /command/core/annotate-one-row (3ms)
08:20:53.868 [ refine] GET /command/core/get-history (12ms)
08:20:53.877 [ refine] GET /command/core/get-preference (9ms)
08:20:54.213 [ refine] GET /command/core/get-csrf-token (336ms)
08:20:54.216 [ refine] POST /command/core/annotate-one-row (3ms)
08:20:54.222 [ refine] GET /command/core/get-history (6ms)
08:20:54.235 [ refine] GET /command/core/get-preference (13ms)
08:20:54.879 [ refine] GET /command/core/get-csrf-token (644ms)
08:20:54.882 [ refine] POST /command/core/annotate-one-row (3ms)
08:20:54.888 [ refine] GET /command/core/get-history (6ms)
08:20:54.905 [ refine] GET /command/core/get-preference (17ms)
08:20:55.941 [ refine] GET /command/core/get-csrf-token (1036ms)
08:20:55.943 [ refine] POST /command/core/annotate-one-row (2ms)
08:20:55.948 [ refine] GET /command/core/get-history (5ms)
08:20:55.958 [ refine] GET /command/core/get-preference (10ms)

```

Conjunto de datos a utilizar:

El conjunto de datos tiene información de varias fuentes, entre ellas la práctica del tutorial de Open refine en [Labinoteca](#), con el objetivo de poder realizar más elementos de práctica de limpieza de datos se han integrados diversas variables según la necesidad. Descargar datos del taller en:

<https://ridda2.utp.ac.pa/handle/123456789/18247>

Carga de datos:

Al abrirse el navegador (predeterminado) se muestra la ventana de OpenRefine, en caso de que no se muestre puede abrirla a través del enlace <http://127.0.0.1:3333/>

Los datos que se pueden cargar en OpenRefine se puede hacer desde diversas fuentes en formatos como: TSV, CSV, SV, MS Excel (.xls/.xlsx), JSON, XML, RDF as XML y datos desde Google Docs.

En la opción del **Elegir archivo**, seleccione el archivo **country-data-openRefine.xlsx** y haga click en NEXT, donde se mostrará el conjunto de datos según el formato.

Al dar siguiente se mostrará la **estructura de la tabla, sus variables y algunos datos**, además en la parte inferior de la ventana se muestran con diversas opciones para modificar la lectura de ellos datos antes de cargarlos (ignorar primera fila, seleccionar primera fila, cargar filas en blanco , etc)

Crear proyecto

1. En la esquina superior derecha verá un cuadro de texto en el que puede color el nombre del proyecto; nómbrelo **country-data-clean-01** (el nombre es solo una referencia para el usuario ya que el proyecto se puede volver abrir cuando se necesite).
2. Haga clic en el botón Create Project.

3. Se mostrará la tabla y sus variables, con el número de tablas identificadas 261, solo aparecerán 10 filas que pueden ser modificadas según la cantidad que deseamos mostrar. En esta ventana es posible modificar tanto la estructura de los datos como sus valores.
4. La carga de los datos puede tomar tiempo dependiendo del número de filas.
5. Los datos alineados a la derecha y de color verde representan datos numéricos.
6. Los datos alineados a la izquierda, sin color, son datos tipo texto.

OpenRefine country data openRefine.xlsx [Enlace permanente](#) Abrir... Exportar Ayuda

Facetas / Filtros < **261 filas** Extensiones Wikidata

Deshacer / Rehacer 0 / 0

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas < primera < anterior 1 siguiente > última >

Usar facetas y filtros

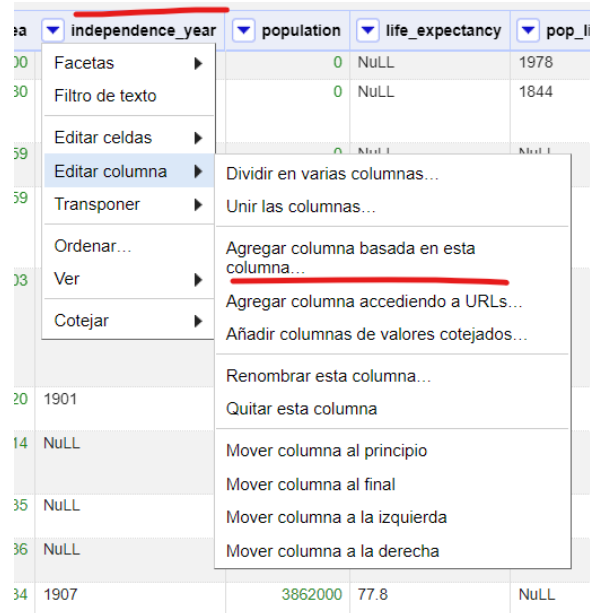
Use las facetas y los filtros para seleccionar subconjuntos de sus datos y trabajar en ellos. Puede encontrar estas opciones en los menús de cada columna.

¿Problemas para comenzar?
[Vea los videos de ayuda](#)

Todo	code	Num	country	continent	region	surface_area	independence_year	population	life_expectancy	pop_life	gnp	gnp_old	local_name	govern
1	AGO	1	Angola	Africa	Central Africa	1246700	1975	12878000	38.3	1966	6648	7964	Angola	Republic
2	BDI	2	Burundi		Eastern Africa	27834	1962	6695000	46.2	1956	903	982	Burundi/burundi	Republic
3	BEN	3	Benin		Western Africa	112622	1960	6097000	50.2	1960	2357	2141	Benin	Republic
4	BFA	4	Burkina Faso		Western Africa	274000	1960	11937000	46.7	1960	2425	2201	Burkina Faso	Republic
5	BWA	5	Botswana		Southern Africa	581730	1966	1622000	39.3	1975	4634	4935	Botswana	Republic
6	CAF	6	Central African Republic		Central Africa	622984	1960	3615000	44	1960	1054	993	Centrafrique/Bé-Africa	Republic
7	CIV	7	Cote d'Ivoire		Western Africa	322463	1960	14786000	45.2	1968	11345	10285	Cote d'Ivoire	Republic
8	CMR	8	Cameroon		Central Africa	475442	1960	15085000	54.8	1964	9174	8596	Cameroon/Cameroon	Republic
9	COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960	51654000	48.8	NULL	6964	2474	Republique Democratique du Congo	Republic
10	COG	10	Congo		Central Africa	342000	1960	2943000	47.4	1990	2108	2287	Congo	Republic

Operaciones de columnas

Podemos modificar las columnas seleccionando la flecha ubicada en cada columna, en este caso nos ubicaremos en la columna **independence_year**, se desplegará un menú con las opciones **EDITAR COLUMNA >** seleccione **AGREGAR columna basada en esta columna**.



Aparece la siguiente ventana, donde le colocaremos por nombre **year_num** y le damos click al botón ACEPTAR.

Añadir columna basada en otra independence_year

Nombre nuevo de la columna:

En error: cambiar a en blanco guardar error copiar valor de la columna original

Expresión: No hay error de sintaxis.

Lenguaje: General Refine Expression Language (GREL)

Previsualización Historial Con estrella Ayuda

row	value	value
1.	NuLL	NuLL
2.	NuLL	NuLL
3.	NuLL	NuLL
4.	NuLL	NuLL
5.	NuLL	NuLL
6.	1901	1901

Al crear la nueva columna con los mismos datos se mostrará la nueva variable a la derecha, los datos alineados a la izquierda indican que OpneRefine los ha identificado como texto, aunque son años.

rea	independence_year	year_num
300	NULL	NULL
780	NULL	NULL
59	NULL	NULL
359	NULL	NULL
903	NULL	NULL
220	1901	1901
14	NULL	NULL
135	NULL	NULL
36	NULL	NULL
334	1907	1907
227	1991	1991
300	1066	1066
273	1921	1921
301	1991	1991
589	1991	1991

Si queremos transformar el tipo de datos, damos click en la flecha de opciones de la columna **year_num**, EDITAR CELDAS > TRANSFORMAR COMUNES > A FECHA.

independence_year	year_num	population	life_expectancy
NULL	0	NULL	19
NULL	0	NULL	18
NULL	600	NULL	19
NULL	2500	NULL	19
NULL	2000	NULL	Nu
1907	1907	3862000	77.8

Los datos de la columna year_num ahora tendrá el año con formato de fecha.

independence_year	year_num
NULL	NULL
NULL	NULL
NULL	NULL
NULL	NULL
NULL	NULL
1901	1901-01-01T00:00:00Z
NULL	NULL
NULL	NULL
NULL	NULL
1907	1907-01-01T00:00:00Z
1991	1991-01-01T00:00:00Z
1066	1066-01-01T00:00:00Z
1921	1921-01-01T00:00:00Z
1991	1991-01-01T00:00:00Z
1991	1991-01-01T00:00:00Z

Facetas

Las facetas son una propiedad de las columnas que permite contabilizar un valor que se repite en la variable la cual puede ser agrupada para contabilizar las veces que se repite ese valor. Las Facetas pueden ser, de texto, numéricas, de líneas de tiempo y gráficas.

Facetas de texto

En la columna **region** daremos click a la flecha hacia abajo, seleccionaremos FACETAS > FACETAS DE TEXTO.

Se mostrará el lado izquierdo la faceta de la variable **region**, con la cantidad de valores que se repiten.

The screenshot shows the OpenRefine interface with a table of 261 rows. The 'region' column is selected, and a facet menu is open. The menu options are: Facetas, Filtro de texto, Editar celdas, Editar columna, Transponer, Ordenar..., Ver, and Cotejar. The 'Facetas' option is expanded, showing: Faceta de texto (highlighted), Faceta numérica, Faceta de línea de tiempo, Faceta gráfica..., Faceta de texto personalizada..., Faceta numérica personalizada..., and Facetas personalizadas.

Todo	code	Num	country	continent	region	surface_area	independence_year	population	life_expectancy	pop_life	gnp	gnp_ok
1	AGO	1	Angola	Africa	Central Africa	1246700	1975	12878000	38.3	1966	6648	7984
2	BDI	2	Burundi		Eastern Africa	27834	1962	6695000	46.2	1956	903	982
3	BEN	3	Benin		Western Africa	112622	1960	6097000	50.2	1960	2357	2141
4	BFA	4	Burkina Faso		Western Africa	274000	1960	11937000	46.7	1960	2425	2201
5	BWA	5	Botswana		Southern Africa	581730	1966	1622000	39.3	1975	4834	4935
6	CAF	6	Central African Republic		Central Africa	622984	1960	3615000	44	1960	1054	993
7	CIV	7	Cote d'Ivoire		Western Africa	322463	1960	14786000	45.2	1968	11345	10285
8	CMR	8	Cameroon		Central Africa	475442	1960	15085000	54.8	1964	9174	8596
9	COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960	51654000	48.8	NULL	6964	2474
10	COG	10	Congo		Central Africa	342000	1960	2943000	47.4	1990	2108	2287

Si le damos click en **CONTEO** en la ventana de faceta **REGION**, podemos ordenar los valores repetidos por cantidad de repeticiones.

En este caso en particular se muestra que el valor **Caribbean** se repite 24 veces.

The screenshot shows the OpenRefine interface with the 'region' facet sorted by count. The 'Caribbean' value is highlighted with a red box, showing it has 24 occurrences. The facet list includes: Caribbean (24), South America (22), Eastern Africa (21), Middle East (18), Southern Europe (18), Western Africa (18), Southern and Central Asia (14), Central Africa (13), Eastern Europe (13), Southeast Asia (11), and Polynesia (10).

Todo	code	Num	country	continent	region	surface_area	independence_year
1	AGO	1	Angola	Africa	Central Africa	1246700	1975
2	BDI	2	Burundi		Eastern Africa	27834	1962
3	BEN	3	Benin		Western Africa	112622	1960
4	BFA	4	Burkina Faso		Western Africa	274000	1960
5	BWA	5	Botswana		Southern Africa	581730	1966
6	CAF	6	Central African Republic		Central Africa	622984	1960
7	CIV	7	Cote d'Ivoire		Western Africa	322463	1960
8	CMR	8	Cameroon		Central Africa	475442	1960
9	COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
10	COG	10	Congo		Central Africa	342000	1960

Facetas numéricas

En la columna **population** daremos click a la flecha hacia abajo, seleccionaremos FACETAS > FACETAS NUMÉRICAS.

Se mostrará el lado izquierdo la faceta de la variable **population** y la distribución de los datos a través de un gráfico de histograma, esta muestra que el rango de los datos esta de 0 a 18000000 y la mayoría de los datos son 0.

The screenshot shows the OpenRefine interface with a numerical facet applied to the 'population' column. The facet is displayed on the left side, showing a histogram with a range from 0 to 1,300,000,000. The main table displays 261 rows of data with columns for code, country, region, surface_area, independence_year, population, life_expectancy, and gnp. The 'population' column values range from 27,834 to 14,786,000.

code	country	region	surface_area	independence_year	population	life_expectancy	pop_life	gnp
AGO	Angola	Central Africa	1246700	1975				3648
BDI	Burundi	Eastern Africa	27834	1962				903
BEN	Benin	Western Africa	112622	1960				2357
BFA	Burkina Faso	Western Africa	274000	1960				2425
BWA	Botswana	Southern Africa	581730	1966				4834
CAF	Central African Republic	Central Africa	622984	1960				1054
CIV	Cote d'Ivoire	Western Africa	322463	1960	14786000	45.2	1968	11345
CMR	Cameroon	Central Africa	475442	1960	15085000	54.8	1964	9174
COD	Congo, The Democratic Republic of the	Central Africa	2344858	1960	51654000	48.8	NULL	6964
COG	Congo	Central Africa	342000	1960	2943000	47.4	1990	2108

Facetas de línea de tiempo

Esta faceta solo puede ser aplicada a los datos de tipo fecha. En la columna **year_num** daremos click a la flecha hacia abajo, seleccionaremos FACETAS > FACETAS DE LINEA DE TIEMPO

Se mostrará el lado izquierdo la faceta de la variable **year_num** y la distribución de los datos a través de un gráfico de histograma, esta muestra que el rango de los datos esta entre los años 1523 y 1993, con mayor cantidad de datos en los años de 1993.

The screenshot shows the OpenRefine interface with a time series facet applied to the 'year_num' column. The facet is displayed on the left side, showing a histogram with a range from 1523 to 1993. The main table displays 187 matching files with columns for code, name, continent, region, surface_area, independence_year, year_num, population, and life_expectancy. The 'year_num' column values range from 1523 to 1993.

code	name	continent	region	surface_area	independence_year	year_num	population	life_expectancy	
AUS	Australia	Australiaand NewZealand	Australiaand NewZealand	7741220	1901			305	
NZL	New Zealand	Australiaand NewZealand	Australiaand NewZealand	270534	1907			NULL	
EST	Estonia	BalticCountries	BalticCountries	45227	1991			362	
GBR	United Kingdom	BalticCountries	BalticCountries	242900	1066			379	
IRL	Ireland	BalticCountries	BalticCountries	70273	1921			NULL	
LTU	Lithuania	BalticCountries	BalticCountries	65301	1991			NULL	
LVA	Latvia	BalticCountries	BalticCountries	64589	1991			375	
ATG	Antigua and Barbuda	Caribbean	Caribbean	442	1981	1981-01-01T00:00:00Z	68000	70.5	1960
BHS	Bahamas	Caribbean	Caribbean	13878	1973	1973-01-01T00:00:00Z	307000	71.1	1960

Datos Duplicados (opción 1)

En el caso de los datos importados a Openrefine hay dos columnas que podemos utilizar para comparar los datos duplicados, **Num y Name** debido a que deben ser datos únicos en la tabla.

Damos click en la flecha al lado de la variable **Num**, FACETAS > FACETAS POR DUPLICADO

261 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

▼	▼	▼	▼	▼	▼	▼	▼	▼
Todo	code	Num	country	continent	region	surface_area	independence_year	popul
1.	AGO	Facetas				1246700	1975	128
2.	BDI	Filtro de texto				27834	1962	66
3.	BEN	Editar celdas				112622	1960	60
4.	BFA	Editar columna				274000	1960	119
5.	BWA	Transponer				581730	1966	16
6.	CAF	Ordenar...				622984	1960	36
7.	CIV	Ver						
8.	CMR	Cotejar						
9.	COD	7 Cote d'Ivoire		Western Africa				47
10.	COG	8 Cameroon		Central Africa				50
		9 Congo, The Democratic Republic of the		Central Africa				16
		10 Congo		Central Africa				29

Facetas personalizadas

- Faceta por palabra
- Faceta por duplicados**
- Faceta log. numérica
- Faceta log. numerica acotada-1
- Faceta por longitud de texto
- Faceta por longitud log. de texto
- Faceta por caracteres Unicode
- Faceta por error
- Faceta por nulo
- Faceta por cuerda vacía
- Faceta por blanco (nulo o cuerda vacía)

En la parte de la izquierda aparecerá una clasificación (FACETA) de la variable **Num** con las facetas identificadas con **TRUE** sí el valor de **NUM** se repite. En este caso se muestran 44 valores que se repiten.

Num		cambiar	
2 elecciones	Ordenar por: A-Z	conteo	
false	217		
true	44		
Facetas por conteo de opciones			

1.	AGO	1	Angola	Africa	Central Africa
2.	BDI	2	Burundi		Eastern Africa
3.	BEN	3	Benin		Western Africa
4.	BFA	4	Burkina Faso		Western Africa
5.	BWA	5	Botswana		Southern Africa
6.	CAF	6	Central		Central

Al darle click en **TRUE** en esta faceta, se mostrará el listado de filas duplicadas por **NUM**, podemos **ordenar** la columna **NUM** por número para poder ver el listado en orden en el cual podemos comparar que en realidad hay valores que se repiten en esta columna (6,7,8,9,10,11...).

Facetas / Filtros

Deshacer / Rehacer 0 / 0

Actualizar Restablecer todos Quitar todo

Num cambiar invertir restaurar

2 elecciones Ordenar por: A-Z conteo

false 217

true 44 exclude

Facetas por conteo de opciones

44 matching filas (261 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas Ordenar ▾

Todo	code	Num	country	continent	region	surface_area	independence_year
☆	6. CAF	6	Central African Republic		Central Africa	622984	1960
☆	240. CAF	6	Central African Republic		Central Africa	622984	1960
☆	7. CIV	7	Cote d'Ivoire		Western Africa	322463	1960
☆	241. CIV	7	Cote d'Ivoire		Western Africa	322463	1960
☆	8. CMR	8	Cameroon		Central Africa	475442	1960
☆	242. CMR	8	Cameroon		Central Africa	475442	1960
☆	9. COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
☆	243. COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
☆	10. COG	10	Congo		Central Africa	342000	1960
☆	244. COG	10	Congo		Central Africa	342000	1960
☆	11. COM	11	Comoros		Eastern Africa	1862	1975
☆	245. COM	11	Comoros		Eastern Africa	1862	1975

Si queremos eliminar las columnas duplicadas, debemos indicarle a OpenRefine cuales eliminar, para ellos podemos hacer uso de las estrellas o banderas que se muestran en cada fila. En este caso le daremos click a las banderas de todas las filas duplicadas como se muestra en la imagen.

Facetas / Filtros

Deshacer / Rehacer 22 / 22

Actualizar Restablecer todos Quitar todo

Num cambiar invertir restaurar

2 elecciones Ordenar por: A-Z conteo

false 217

true 44 exclude

Facetas por conteo de opciones

44 matching filas (261 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas Ordenar ▾

Todo	code	Num	country	continent	region	surface_area	independence_year
☆	6. CAF	6	Central African Republic		Central Africa	622984	1960
☆	240. CAF	6	Central African Republic		Central Africa	622984	1960
☆	7. CIV	7	Cote d'Ivoire		Western Africa	322463	1960
☆	241. CIV	7	Cote d'Ivoire		Western Africa	322463	1960
☆	8. CMR	8	Cameroon		Central Africa	475442	1960
☆	242. CMR	8	Cameroon		Central Africa	475442	1960
☆	9. COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
☆	243. COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
☆	10. COG	10	Congo		Central Africa	342000	1960
☆	244. COG	10	Congo		Central Africa	342000	1960
☆	11. COM	11	Comoros		Eastern Africa	1862	1975
☆	245. COM	11	Comoros		Eastern Africa	1862	1975

Hacemos click en la columna **TODO**, y en la flecha seleccionamos FACETAS > FACETAS POR BANDERA

Se mostrará otras facetas de las filas con bandera señaladas con TRUE que en este caso son 22 filas y las que no tienen bandera señaladas con false.

Al hacer click en **TRUE** de la faceta FILAS CON BANDERA se mostrará el listado de las filas que tienen bandera y que fueron seleccionadas como filas duplicadas.

The screenshot shows the OpenRefine interface with 22 matching files. The 'Facetas / Filtros' panel on the left shows two facets: 'Num' and 'Filas con bandera'. The main table has columns: 'Todo', 'code', 'Num', 'country', 'continent', 'region', 'surface_area', and 'independence_year'. A red arrow points to the 'Todo' column header.

Todo	code	Num	country	continent	region	surface_area	independence_year
240.	CAF	6	Central African Republic		Central Africa	622984	1960
241.	CIV	7	Cote d'Ivoire		Western Africa	322463	1960
242.	CMR	8	Cameroon		Central Africa	475442	1960
243.	COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960
244.	COG	10	Congo		Central Africa	342000	1960
245.	COM	11	Comoros		Eastern Africa	1862	1975
246.	MDA	144	Moldova		Eastern Europe	33851	1991
247.	MKD	145	Macedonia		Southern Europe	25713	1991
248.	MLT	146	Malta		Southern Europe	316	1964
249.	NLD	147	Netherlands		Western Europe	41526	1581
250.	NOR	148	Norway		Nordic Countries	323877	1905
251.	POL	149	Poland		Eastern Europe	323250	1918

En este momento podemos eliminar las filas. Hacemos click en la columna **TODO**, y en la flecha seleccionamos **EDITAR FILAS > QUITAR LAS FILAS QUE ENCAJEN**.

The screenshot shows the OpenRefine interface with the 'Todo' column context menu open. The option 'Quitar las filas que encajen' is highlighted in red. The 'Facetas / Filtros' panel on the left shows the same two facets as before.

Todo	code	Num	country	continent	region	surface_area	independ
Transformar...		6	Central African Republic		Central Africa	622984	1960
Editar todas las columnas		7	Cote d'Ivoire		Western Africa	322463	1960
Facetas		8	Cameroon		Central Africa	475442	1960
Editar filas					Central Africa	2344858	1960
Editar columnas					Central Africa		
Ver					Central Africa	342000	1960
244.	COG				Central Africa	342000	1960
245.	COM				Eastern Africa	1862	1975
246.	MDA	144	Moldova		Eastern Europe	33851	1991
247.	MKD	145	Macedonia		Southern Europe	25713	1991
248.	MLT	146	Malta		Southern Europe	316	1964
249.	NLD	147	Netherlands		Western Europe	41526	1581
250.	NOR	148	Norway		Nordic Countries	323877	1905
251.	POL	149	Poland		Eastern Europe	323250	1918

En ambas facetas **Num** y **Filas con bandera** aparecerán las opciones de true con 0 valores. Hacer click en el botón **QUITAR TODO** para eliminar las facetas y mostrar el listado de filas (239) sin los registros duplicados.

The screenshot shows the 'Facetas / Filtros' (Facets / Filters) panel in OpenRefine. At the top, there is a 'Deshacer / Rehacer 23 / 23' (Undo / Redo 23 / 23) indicator. Below it are three buttons: 'Actualizar' (Update), 'Restablecer todos' (Reset all), and 'Quitar todo' (Remove all). The panel contains two facets:

- Num**: Shows '1 elección' (1 selection), 'Ordenar por: A-Z' (Sort by: A-Z), and 'conteo' (count). It has a 'true 0' status and an 'exclude' button. Below the facet is the text 'Facetas por conteo de opciones' (Facets by option count).
- Filas con bandera**: Shows '1 elección' (1 selection), 'Ordenar por: A-Z' (Sort by: A-Z), and 'conteo' (count). It has a 'true 0' status and an 'exclude' button. Below the facet is the text 'Facetas por conteo de opciones' (Facets by option count).

Nota: Aunque el proceso es funcional solo resulta viable si hay pocas filas, pero es un poco complicado con un número considerable de filas a eliminar, ya que habría que seleccionar la opción de bandera o estrella por cada uno, además que al ser un proceso de selección manual se puede cometer errores.

Datos Duplicados (opción 2)

En esta segunda opción de eliminar datos duplicados, seleccionamos nuevamente la variable **Num**.

Damos click en la flecha al lado de la variable **Num**, ORDENAR. Se abrirá la ventana Ordenar por **Num**, seleccionamos, números y menores primero, según imagen y ACEPTAR.

259 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000

Ordenar por **Num**

Ordenar valores como

- texto Distingue mayúsculas y minúsculas
- números**
- fechas
- booleano

Posición de blancos y errores

Valores validos

Errores

Blancos

Arrastre para ordenar

menores primero mayores primero

Aceptar Cancelar

El reordenamiento que se hace es solamente Visual, porque necesitamos decirle a OpenRefine que el ordenamiento debe ser permanente ubicándonos en la parte superior, seleccionar **ORDENAR > Reordenar filas permanentemente**. Las con el mismo número (duplicadas) permanecerán debajo de las filas con el número similar.

261 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas Ordenar ▾ « primera

Quitar orden

Reordenar filas permanentemente

Por Num ▶

	code	Num	country	continent	region	surface_area	life_expectancy	pop_life	
1.	AGO	1	Angola	Africa	Central Africa	124670	3	1966	
2.	BDI	2	Burundi		Eastern Africa	27834	42	1956	
3.	BEN	3	Benin		Western Africa	112622	6097000	50.2	1960
4.	BFA	4	Burkina Faso		Western Africa	274000	11937000	46.7	1960
5.	BWA	5	Botswana		Southern Africa	581730	1622000	39.3	1975
6.	CAF	6	Central African Republic		Central Africa	622984	3615000	44	1960
240.	CAF	6	Central African Republic		Central Africa	622984	3615000	44	0
7.	CIV	7	Cote d'Ivoire		Western Africa	322463	14786000	45.2	1968
241.	CIV	7	Cote d'Ivoire		Western Africa	322463	14786000	45.2	0
8.	CMR	8	Cameroon		Central Africa	475442	15085000	54.8	1964
242.	CMR	8	Cameroon		Central Africa	475442	15085000	54.8	0

Necesitamos ahora identificar las filas con valores duplicados.

Seleccionamos en la columna **NUM**, EDITAR CELDAS > VACIAR HACIA ABAJO.

261 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 **50** 100 500 1000 filas

▼ Todo	▼ code	▼ Num	▼ country	▼ continent	▼ region	▼ surface_area	▼ indepen
☆	1. AGO	Facetas		Africa	Central Africa	1246700	1975
☆	2. BDI	Filtro de texto			Eastern	27834	1962
☆	3. BEN	Editar celdas				112622	1960
☆	4. BFA	Editar columna				274000	1960
☆	5. BWA	Transponer				581730	1966
☆	6. CAF	Ordenar...				622984	1960
☆	7. CAF	Ver	Central African Republic			622984	1960
☆	8. CIV	Cotejar	Cote d'Ivoire			322463	1960
☆	9. CIV		Cote d'Ivoire	Africa	Western Africa	322463	1960
☆	10. CMR		Cameroon		Central Africa	475442	1960
☆	11. CMR		Cameroon		Central Africa	475442	1960

Las filas con los valores repetidos ahora no tienen valores.

261 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 **50** 100 500 1000 filas

▼ Todo	▼ code	▼ Num	▼ country	▼ continent	▼ region	▼ surface_area	▼ independence
☆	1. AGO	1	Angola	Africa	Central Africa	1246700	1975
☆	2. BDI	2	Burundi		Eastern Africa	27834	1962
☆	3. BEN	3	Benin		Western Africa	112622	1960
☆	4. BFA	4	Burkina Faso		Western Africa	274000	1960
☆	5. BWA	5	Botswana		Southern Africa	581730	1966
☆	6. CAF	6	Central African Republic		Central Africa	622984	1960
☆	7. CAF	7	Central African Republic		Central Africa	622984	1960
☆	8. CIV	7	Cote d'Ivoire		Western Africa	322463	1960
☆	9. CIV	8	Cote d'Ivoire		Western Africa	322463	1960
☆	10. CMR	8	Cameroon		Central Africa	475442	1960
☆	11. CMR		Cameroon		Central Africa	475442	1960
☆	12. COD	9	Congo, The Democratic Republic of the		Central Africa	2344858	1960

Nuevamente en la variable **Num**, vamos a **FACETAS > Facetas personalizadas > facetas por blanco**.

261 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 **50** 100 500 1000 filas

▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
▼	code	Num	country	continent	region	surface_area	independence_year	population	▼
☆	1.	AGO	Facetas	Faceta de texto		1246700	1975	12878000	38.8
☆	2.	BDI	Filtro de texto	Faceta numérica		27834	1962	6695000	46.2
☆	3.	BEN	Editar celdas	Faceta de línea de tiempo		112622	1960	6097000	50.2
☆	4.	BFA	Editar columna	Faceta gráfica...		274000	1960	11937000	46.7
☆	5.	BWA	Transponer	Faceta de texto personalizada...		581730	1966	1622000	39.3
☆	6.	CAF	Ordenar...	Faceta numérica personalizada...		622984	1960	3615000	44
☆	6.	CAF	Ver	Facetas personalizadas					
☆	7.	CAF	Cotejar	Faceta por palabra				3615000	44
☆	8.	CIV	7	Faceta por duplicados	Central African Republic				
☆	8.	CIV		Faceta log. numérica	Central Africa			4786000	45.2
☆	9.	CIV		Faceta log. numerica acotada-1	Western Africa			4786000	45.2
☆	10.	CMR	8	Faceta por longitud de texto	Western Africa			5085000	54.8
☆	11.	CMR		Faceta por longitud log. de texto	Central Africa			5085000	54.8
☆	11.	CMR		Faceta por caracteres Unicode	Central Africa			5085000	54.8
☆	12.	COD	9	Faceta por error	Congo, The Democratic Republic of the			1654000	48.8
☆	12.	COD		Faceta por nulo	Central Africa				
☆	13.	COD		Faceta por cuerda vacía	Congo, The Democratic Republic of			1654000	48.8
☆	13.	COD		Faceta por blanco (nulo o cuerda vacía)	Central Africa				

Se mostrará la faceta con los datos vacíos (22) **TRUE** y los no vacíos (239).

Num cambiar

2 elecciones Ordenar por: **A-Z** conteo

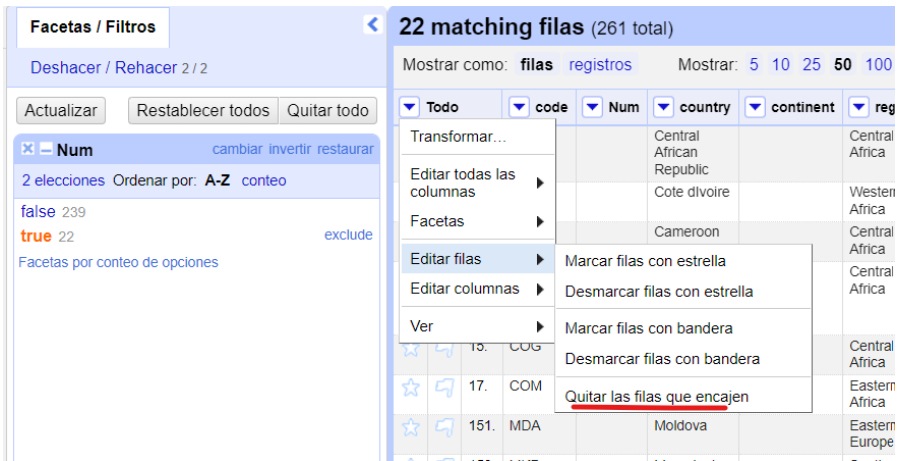
false 239

true 22

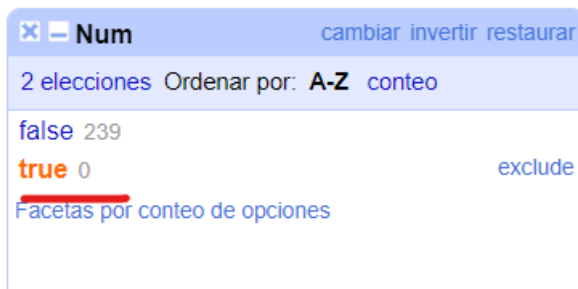
Facetas por conteo de opciones

Damos click en **TRUE**, el cual desplegará solo los datos vacíos que en teoría son los duplicados.

En la primera columna **Todos**, seleccionamos, **Editar filas > Quitar las filas que encajen**.



Aparecerán 0 filas en TRUE de la faceta **Num**, indicando que los datos duplicados fueron eliminados. Recuerde darle click al botón **QUITAR TODO** para eliminar las facetas y ver el resultado.



Rellenar espacios en blanco

Si observamos los datos de la columna **continent**, vemos que aparece el nombre del continente solo en una fila, esto se debe a que quien creó los datos solo colocó el nombre del continente al inicio de cada grupo de países con el mismo continente, por lo que existen datos en blanco.

239 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500

▼ Todo	▼ code	▼ Num	▼ country	▼ continent	▼ region	
☆	1.	AGO	1	Angola	Africa	Central Africa
☆	2.	BDI	2	Burundi		Eastern Africa
☆	3.	BEN	3	Benin		Western Africa
☆	4.	BFA	4	Burkina Faso		Western Africa
☆	5.	BWA	5	Botswana		Southern Africa
☆	6.	CAF	6	Central African Republic		Central Africa
☆	7.	CIV	7	Cote d'Ivoire		Western Africa
☆	8.	CMR	8	Cameroon		Central Africa
☆	9.	COD	9	Congo, The Democratic Republic of the		Central Africa
☆	10.	COG	10	Congo		Central Africa
☆	11.	COM	11	Comoros		Eastern Africa
☆	12.	CPV	12	Cape Verde		Western Africa
☆	13.	DJI	13	Djibouti		Eastern Africa

Para rellenar estos datos con los valores de los **continentes**, es importante resaltar que los datos deben aparecer en orden, como en este caso donde están ordenados por la variable **Num**, ya que de esa forma podemos asegurarnos de que los espacios en blancos se rellenen con el dato que está por encima de él, como la columna **continent** que inicia con el valor de **Africa** y debajo, los datos en blanco.

Nos ubicamos en la variable **continent**, desplegamos el menú en la flecha hacia abajo, EDITAR CELDAS > **llenar hacia abajo**.

239 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

▼	▼	▼	▼	▼	▼	▼	▼
Todo	code	Num	country	continent	region	surface_area	independence_year
☆	1.	AGO	1	Angola		1246700	1975
☆	2.	BDI	2	Burundi		27834	1962
☆	3.	BEN	3	Benin			
☆	4.	BFA	4	Burkina Faso			
☆	5.	BWA	5	Botswana			
☆	6.	CAF	6	Central African Republic			
☆	7.	CIV	7	Cote d'Ivoire	Western Africa		
☆	8.	CMR	8	Cameroon	Central Africa		
☆	9.	COD	9	Congo, The Democratic Republic of	Central Africa	2344858	1960

Facetas

Filtro de texto

Editar celdas

Editar columna

Transponer

Ordenar...

Ver

Cotejar

Transformar...

Transformaciones comunes

Llenar hacia abajo

Vaciar hacia abajo

Dividir celdas multi-valuadas

Unir celdas multi-valuadas...

Agrupar y editar...

Reemplazar...

Las filas con espacios en blanco ya fueron llenados con el valor que estaba en la fila posterior a las filas en blanco.

239 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

▼	▼	▼	▼	▼	▼	▼	▼
Todo	code	Num	country	continent	region	surface_area	independence_year
☆	1.	AGO	1	Angola	Africa	1246700	1975
☆	2.	BDI	2	Burundi	Africa	27834	1962
☆	3.	BEN	3	Benin	Africa	112622	1960
☆	4.	BFA	4	Burkina Faso	Africa	274000	1960
☆	5.	BWA	5	Botswana	Africa	581730	1966
☆	6.	CAF	6	Central African Republic	Africa	622984	1960
☆	7.	CIV	7	Cote d'Ivoire	Africa	322463	1960
☆	8.	CMR	8	Cameroon	Africa	475442	1960
☆	9.	COD	9	Congo, The Democratic Republic of the	Africa	2344858	1960
☆	10.	COG	10	Congo	Africa	342000	1960
☆	11.	COM	11	Comoros	Africa	1862	1975
☆	12.	CPV	12	Cape Verde	Africa	4033	1975
☆	13.	DJI	13	Djibouti	Africa	23200	1977
☆	14.	DZA	14	Algeria	Africa	2381741	1962
☆	15.	EGY	15	Egypt	Africa	1001449	1922
☆	16.	ERI	16	Eritrea	Africa	117600	1993

Unificar datos categóricos

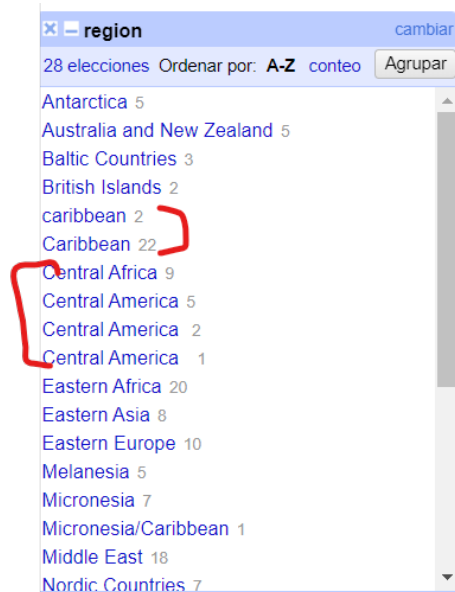
Uno de los problemas comunes en los datos, principalmente en datos de tipos categóricos es que los nombres de las categorías no estén unificados, un espacio en blanco antes o después del texto, mayúscula o minúscula en la misma categoría genera diferencias en la categorización de los datos.

En este ejemplo utilizaremos la variable **region**. Seleccionamos la flecha hacia abajo FACETAS > facetas de texto.



239 filas						
Mostrar como: filas registros						
Mostrar: 5 10 25 50 100 500 1000 filas						
▼	▼	▼	▼	▼	▼	▼
Todo	code	Num	country	continent	region	surface_area
★	1.	AGO	1	Angola	Africa	
★	2.	BDI	2	Burundi	Africa	
★	3.	BEN	3	Benin	Africa	
★	4.	BFA	4	Burkina Faso	Africa	
★	5.	BWA	5	Botswana	Africa	
★	6.	CAF	6	Central African Republic	Africa	
★	7.	CMR	7	Cameroon	Africa	

Se mostrarán los datos categóricos en la ventana de facetas **región** con el número de filas que repiten esa valor o categoría. En esta pestaña podemos identificar que hay dos categorías **caribbean**, una que inicia con la letra mayúscula (22 FILAS) y otra que no (2 filas), también en la categoría **Central America** hay tres categorías, y aunque el texto es igual, podemos ver que el número al final del nombre no tiene el mismo espaciado, por lo que es probable que existan espacios en blanco en el texto.



region	count
Antarctica	5
Australia and New Zealand	5
Baltic Countries	3
British Islands	2
caribbean	2
Caribbean	22
Central Africa	9
Central America	5
Central America	2
Central America	1
Eastern Africa	20
Eastern Asia	8
Eastern Europe	10
Melanesia	5
Micronesia	7
Micronesia/Caribbean	1
Middle East	18
Nordic Countries	7

Si nos ubicamos al lado de la categoría **caribbean** vemos que aparecen dos opciones, editar e **include**. Si le damos a include, nos mostrará las filas donde se muestra los datos que tienen **caribbean** en minúscula. En este caso le daremos **EDITAR**.

The screenshot shows the OpenRefine interface. On the left, the 'region' facet is expanded, and 'caribbean' is selected and underlined in red. The main table displays the following data:

1.	AGO	1	Angola	Africa	Central Africa
2.	BDI	2	Burundi	Africa	Eastern Africa
3.	BEN	3	Benin	Africa	Western Africa
4.	BFA	4	Burkina Faso	Africa	Western Africa
5.	BWA	5	Botswana	Africa	South Africa
8.	CMR	8	Cameroon	Africa	Central Africa
9.	COD	9	Congo, The Democratic Republic of the	Africa	Central Africa

A dialog box is open over the table, showing the text 'Caribbean' in a text input field, with 'Aplicar' and 'Cancelar' buttons below it.

Se desplegará una ventanita que nos da la opción de modificar el texto, cambiamos la letra “c” a mayúscula y le damos **APLICAR**.

Aparecerá una sola categoría **Caribbean** con 24 filas con es categoría.

The screenshot shows the OpenRefine interface. The 'region' facet is expanded, and 'Caribbean' is selected and underlined in red. The main table is not visible in this view.

Al realizar los mismos pasos con la categoría **Centra America**, eliminamos los espacios en blanco al final de los tres textos con esa categoría y obtendremos esa categoría con 8 filas.

The screenshot shows the OpenRefine interface. The 'region' facet is expanded, and 'Central America' is selected and underlined in red. The main table is not visible in this view.

Reemplazar datos

Algunas veces es necesario reemplazar datos que están en los datos, ya sea porque están erróneos o porque se desean modificar, el detalle es que estos datos pueden ser únicos en la tabla.

Tomando en cuenta que la categoría **Central America**, y que corregimos en la faceta. Haremos un filtro de estos datos por la variable **región**.

Nos ubicamos en la variable **región**, damos click en la flecha y seleccionamos **filtro de texto**. Se mostrará una ventana de filtro y escribiremos la palabra **Central America**.

Facetas / Filtros

Deshacer / Rehacer 7 / 7

Actualizar Restablecer todos Quitar todo

region invertir restaurar

Central America

Distingue mayúsculas y minúsculas regex

8 matching filas (239 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

	code	Num	country	continent	region	surface
166.	BLZ	166	Belize	North America	Central America	
170.	CRI	170	Costa Rica	North America	Central America	
178.	GTM	178	Guatemala	North America	Central America	
179.	HND	179	Honduras	North America	Central America	
184.	MEX	184	Mexico	North America	Central America	
187.	NIC	187	Nicaragua	North America	Central America	
188.	PAN	188	Panama	North America	Central America	
190.	SLV	190	El Salvador	North America	Central America	

Se mostrará un listado de los países cuya región es **Central América**.

En los datos mostrados vemos que existen errores en los nombres de algunos países así que es necesario reemplazarlos.

Facetas / Filtros

Deshacer / Rehacer 7 / 7

Actualizar Restablecer todos Quitar todo

region invertir restaurar

Central America

Distingue mayúsculas y minúsculas regex

8 matching filas (239 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

	code	Num	country	continent	region	surface_area	independence_year	population	life_expectancy	pop_...
166.	BLZ	166	Belize	North America	Central America	22696	1981	241000	70.9	NULL
170.	CRI	170	Costa Rica	North America	Central America	51100	1821	4023000	75.8	1993
178.	GTM	178	Guatemala	North America	Central America	108889	1821	11385000	66.2	1918
179.	HND	179	Honduras	North America	Central America	112088	1838	6485000	69.9	1991
184.	MEX	184	Mexico	North America	Central America	1958201	1810	98881000	71.5	1951
187.	NIC	187	Nicaragua	North America	Central America	130000	1838	5074000	68.7	1918
188.	PAN	188	Panama	North America	Central America	75517	1903	2856000	75.5	1878
190.	SLV	190	El Salvador	North America	Central America	21041	1841	6276000	69.7	1993

Existen varias formas de reemplazar los datos, para generar un nuevo aprendizaje utilizaremos la función transformar, que se encuentra, en la variable country, click en la flecha, EDITAR CELDAS > TRANSFORMAR.

8 matching filas (239 total)

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

Todo	code	Num	country	continent	region	surface_area	
☆	166.	BLZ	166	Facetas	merica	Central America	22696
☆	170.	CRI	170	Filtro de texto	merica	Central	51100
☆	178.	GTM	178	Editar celdas			88
☆	179.	HND	179	Editar columna			88
☆	184.	MEX	184	Transponer			0
☆	187.	NIC	187	Ordenar...			00
☆	188.	PAN	188	Ver	Panamá	North America	11
☆	190.	SLV	190	Cotejar	El Salvador	North America	4

Se mostrará una ventana que muestran varios elementos:

- Expresión = indica la formula a construir para reemplazar los datos.
- Lenguaje = se utiliza un lenguaje de expresiones regulares propio de OpenRefine llamado GREL.
- Previsualización: a medida que vayamos construyendo la expresión la columna **value**, se modifica con el nuevo valor.

Transformación personalizada en country

Expresión Lenguaje **General Refine Expression Language (GREL)**

value No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value
166.	Belize	Belize
170.	Costa Rica	Costa Rica
178.	Guatemala	Guatemala
179.	Honduras	Honduras
184.	Mexico	Mexico
187.	Nicaragua	Nicaragua

En error mantener original Re-transformar hasta veces hasta que no haya cambios
 cambiar a en blanco
 guardar error

Aceptar Cancelar

Para este ejemplo utilizaremos la función `.repace("dato actual", "nuevo dato")`.

Si escribimos la siguiente expresión:

value

.replace('Guatemale','')

Se modificará el valor de Guatemale por un valor vacío

Expresión Lenguaje

```
value
.replace('Guatemale','')
```

No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value .replace('Guatemale','')
166.	Belize	Belize
170.	Costa Rica	Costa Rica
178.	Guatemale	
179.	Hondures	Hondures
184.	Mexico	Mexico
187.	Nicarague	Nicarague

Si completamos la expresión por el valor a reemplazar se mostrará la palabra **Guatemala** de forma correcta.

Transformación personalizada en country

Expresión Lenguaje

```
value
.replace('Guatemale','Guatemala')
```

No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value .replace('Guatemale','Gu ...
166.	Belize	Belize
170.	Costa Rica	Costa Rica
178.	Guatemale	Guatemala
179.	Hondures	Hondures
184.	Mexico	Mexico
187.	Nicarague	Nicarague

Podemos añadir todos los valores a reemplazar, añadiendo una nueva fila. Es necesario dar **enter** para añadir nuevas filas para reemplazar los datos.

Transformación personalizada en country

Expresión Lenguaje General Refine Expression Language (GREL)

```
value
.replace('Guatemale','Guatemala')
.replace('Hondures','Honduras')
.replace('Nicarague','Nicaragua')
.replace('El Salvadore','El Salvador')
```

No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

170.	Costa Rica	Costa Rica
178.	Guatemale	Guatemala
179.	Hondures	Honduras
184.	Mexico	Mexico
187.	Nicarague	Nicaragua
188.	Panama	Panama
190.	El Salvadore	El Salvador

En error mantener original Re-transformar hasta veces hasta que no haya cambios
 cambiar a en blanco
 guardar error

Aceptar Cancelar

Le damos click a **ACEPTAR** y las filas en la tabla se actualizarán con los nuevos datos.

fine.xlsx [Enlace permanente](#)

Text transform on 4 cells in column country: `grel:value`
`.replace('Guatemale','Guatemala')`
`.replace('Hondures','Honduras')`
`.replace('Nicarague','Nicaragua')` `.replace('El Salvadore','El Salvador')` **Deshacer**

Mostrar como: **filas** registros Mostrar: 5 10 20 50 100 500 1000 filas

Todo	code	Num	country	continent	region	surface_area	independence_year	population	life_exp	
☆	166.	BLZ	166	Belize	North America	Central America	22696	1981	241000	70.9
☆	170.	CRI	170	Costa Rica	North America	Central America	51100	1821	4023000	75.8
☆	178.	GTM	178	Guatemala	North America	Central America	108889	1821	11385000	66.2
☆	179.	HND	179	Honduras	North America	Central America	112088	1838	6485000	69.9
☆	184.	MEX	184	Mexico	North America	Central America	1958201	1810	98881000	71.5
☆	187.	NIC	187	Nicaragua	North America	Central America	130000	1838	5074000	68.7
☆	188.	PAN	188	Panama	North America	Central America	75517	1903	2856000	75.5
☆	190.	SLV	190	El Salvador	North America	Central America	21041	1841	6276000	69.7

Cluster (Agrupar)

Si realizamos una **faceta de texto** de la variable Street tal cual como la practica de facetas, podemos identificar que existen datos que son parecidos, pero difieren en una letra, mayúsculas, minúsculas o en algunos casos tiene otras palabras.

En lugar de modificar los datos en las facetas utilizaremos el botón AGRUPAR en la faceta para que Open Refine pueda a través de un algoritmo identificar los posibles datos similares.

Se abrirá una nueva ventana que muestra las agrupaciones automáticas detectadas según el método **colisión de claves** y la función **Huella**.

Cada grupo muestra los valores en agrupación similares.

Agrupar y editar valores en la columna "Street"

Esta función le permite encontrar agrupaciones de diferentes valores que pueden ser representaciones alternativas de la misma cosa. Por ejemplo, "New York" y "new york" probablemente se refieren al mismo concepto, solo se presenta diferencia en la capitalización. De la misma manera "Godel" y "Godel" probablemente se refieren a la misma persona. [Conozca más...](#)

Método: **colisión de claves** Función: **Huella** 13 agrupamientos encontrados

Tamaño del grupo	Número de filas	Valores en agrupación	¿Unir?	Nuevo valor de la celda
4	4	HIGH STREET KINGS HEATH High Street Kings Heath Kings Heath High Street high street, kings heath	<input type="checkbox"/>	HIGH STREET KINGS HEATH
3	3	Stratford Road Stratford road stratford road	<input type="checkbox"/>	Stratford Road
2	3	Stoney Lane (2 filas) stoney lane	<input type="checkbox"/>	Stoney Lane
2	4	Washwood Heath Road (3 filas) Washwood Heath road	<input type="checkbox"/>	Washwood Heath Road
2	2	poplar road , solihull poplar road, solihull	<input type="checkbox"/>	poplar road , solihull
2	2	Bordesley Green Bordesley Green , Bordesley Green.	<input type="checkbox"/>	Bordesley Green

En cada grupo se nos da la opción de modificar los **valores por agrupación** por el **nuevo valor de la celda (el cual puede ser modificado)**, si deseo unir los valores con el mismo dato, debo seleccionar la opción de **UNIR**.

Puedo seleccionar uno por uno de acuerdo con lo que considero compatible según recomendaciones o puedo seleccionar todos.

Tamaño del grupo	Número de filas	Valores en agrupación	¿Unir?	Nuevo valor de la celda
4	4	HIGH STREET KINGS HEATH High Street Kings Heath Kings Heath High Street high street, kings heath	<input checked="" type="checkbox"/>	Kings Heath High Street
3	3	Stratford Road Stratford road stratford road	<input checked="" type="checkbox"/>	Stratford Road
2	3	Stoney Lane (2 filas) stoney lane	<input checked="" type="checkbox"/>	Stoney Lane
2	4	Washwood Heath Road (3 filas) Washwood Heath road	<input checked="" type="checkbox"/>	Washwood Heath Road
2	2	poplar road , solihull poplar road, solihull	<input checked="" type="checkbox"/>	poplar road , solihull
2	2	Bordesley Green Bordesley Green , Bordesley Green.	<input checked="" type="checkbox"/>	Bordesley Green

Una vez terminado de unir damos click al botón **UNIR SELECCIONADOS Y REGRESAR**, aparecerá el cuadro vacío sin recomendación.

Método Función

No se encontraron agrupaciones con el método seleccionado
Intente seleccionando otro método arriba o cambiando los parámetros

Hay que recordar que el método colisión de claves , **tiene otros algoritmos**, si ahora seleccionamos **metaphone 3**, veremos que este detecto otros posibles cluster para agrupar, el cual debo verificar, **unir** y dar click en el botón **unir seleccionados y reagrupar**.

Método: colisión de claves Función: metaphone3 10 agrupamientos encontrados

Tamaño del grupo	Número de filas	Valores en agrupación	¿Unir?	Nuevo valor de la celda
6	9	BIRMINGHAM RD (2 filas) Birmingham Road (2 filas) Birmingham Road, Sutton Coldfield (2 filas) Birmingham Rd, SC Birmingham Road, Erdington birmingham road sutton	<input type="checkbox"/>	BIRMINGHAM RD
4	4	COVENTRY RD, SMALL HEATH, BHAM Coventry Road, Small Heath Coventry rd Small heath coventry road,small heath	<input type="checkbox"/>	Coventry Road, Small Heath
3	5	Stratford Road (3 filas) stratford road, hall green stratford roads	<input type="checkbox"/>	Stratford Road
3	4	Alcester Road Moseley (2 filas) Alcester Raod, Mosley alcester road	<input type="checkbox"/>	Alcester Road Moseley

Valores en agrupación

Filas en el grupo

Longitud promedio de los valores

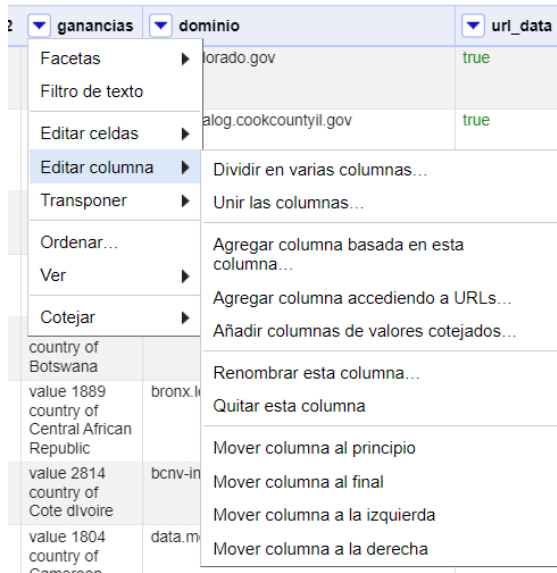
Varianza de los valores

Puedo hacer el mismo proceso con todas las funciones del listado hasta que no aparezcan más datos que se puedan agrupar. Una vez terminemos le damos click al botón **CERRAR**.

Extraer datos numéricos de un campo de texto (opción 1)

La siguiente practica está orientada en poder obtener el valor numérico que está integrado en esta cadena de texto **value 56 country of Angola** perteneciente a la columna **ganancias**.

Nos ubicamos en la columna **ganancias** , EDITAR COLUMNAS, DIVIDIR EN VARIAS COLUMNAS



Aparecerá la siguiente ventana donde en el valor del separador debemos colocar solo un espacio en blanco, debido a que la cadena de texto tiene palabras separadas por espacios en blanco.

Desactivar la opción **QUITAR esta columna**.

Dividir columna

Cómo dividir la columna

por separador

Separador expresión regular

Dividir en columnas máximo (deje en blanco para no limitar)

por longitudes de campo

Lista de números enteros separados por comas; p. ej., 5, 7, 15

Luego de la división

Interpretar el tipo de celda

Quitar esta columna

Se mostrarán todos los datos separados por columnas, cuyos nombres son ganancias 1, ganancias 2... ganancias 9.

La columna que nos interesa es ganancias 2 que contiene los valores numéricos que contenía la cadena.

ganancias	ganancias 1	ganancias 2	ganancias 3	ganancias 4	ganancias 5	ganancias 6	ganancias 7	ganancias 8
value 56 country of Angola	value	56	country	of		Angola		
value 552 country of Burundi	value	552	country	of		Burundi		
value 187 country of Benin	value	187	country	of		Benin		
value 549 country of Burkina Faso	value	549	country	of		Burkina Faso		
value 204 country of Botswana	value	204	country	of		Botswana		
value 1889 country of Central African Republic	value	1889	country	of		Central African Republic		
value 2814 country of Cote d'Ivoire	value	2814	country	of		Cote d'Ivoire		
value 1804 country of Cameroon	value	1804	country	of		Cameroon		
value 2298 country of Congo, The Democratic Republic of the	value	2298	country	of		Congo, The Democratic Republic of the		
value 2296 country of Congo	value	2296	country	of		Congo		
value 2295 country of Comoros	value	2295	country	of		Comoros		

Extraer datos numéricos de un campo de texto (opción 2)

Uno de los ejemplos más comunes en la limpieza de datos es que existen variables donde hay información de texto y número, donde solo queremos extraer el campo numérico.

Street	head_of_state	capital	code2	ganancias	dominio
Lenches Close, Mosley	Jose Eduardo dos Santos	56	AO	value 56 country of Angola	data.colorado.gov
Mosley All Services Club, Church Road, Moseley.	Pierre Buyoya	552	BI	value 552 country of Burundi	datacatalog.cookcountyil.gov
Woodbridge road, moseley	Mathieu Kerekou	187	BJ	value 187 country of Benin	data.tompsc.com
Wright Street	Blaise Compaore	549	BF	value 549 country of Burkina Faso	venturaca.data.socrata.com
Warwick Road	Festus G. Mogae	204	BW	value 204 country of Botswana	bronx.lehman.cuny.edu
Lime Tree Road	Ange-Felix Patassé	1889	CF	value 1889 country of Central African Republic	bronx.lehman.cuny.edu

Para ellos nos vamos ubicar en la columna ganancias, le damos click a la flecha, seleccionamos **EDITAR COLUMNA** > Agregar columna basada en esta columna.

Street	head_of_state	capital	code2	ganancias	dominio
Lenches Close, Mosley	Jose Eduardo dos Santos	56	AO	Facetas	
Mosley All Services Club, Church Road, Moseley.	Pierre Buyoya	552	BI	Filtro de texto	
Woodbridge road, moseley	Mathieu Kerekou	187	BJ	Editar celdas	
Wright Street	Blaise Compaore	549	BF	Editar columna	
Warwick Road	Festus G. Mogae	204	BW	Transponer	
Lime Tree Road	Ange-Felix Patassé	1889	CF	Ordenar...	
yardley green road	Patrice Talon	2814	CM	Ver	
Lode Lane @ Hermitage Road	Paul Biya	1804	CM	Cotejar	

Se abrirá una nueva ventana que nos pedirá el nombre de la nueva columna y un espacio llamado **EXPRESIÓN** que nos permite colocar la expresión del valor a mostrar, en este caso solo aparecerá la palabra **value**, que indica el mismo valor de la columna.

Añadir columna basada en otra ganancias

Nombre nuevo de la columna

En error cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value
1.	value 56 country of Angola	value 56 country of Angola
2.	value 552 country of Burundi	value 552 country of Burundi
3.	value 187 country of Benin	value 187 country of Benin
4.	value 549 country of Burkina Faso	value 549 country of Burkina Faso
5.	value 204 country of Botswana	value 204 country of Botswana
6.	value 1889 country of Central African Republic	value 1889 country of Central African Republic

Si queremos que OpenRefine seleccione los números integrados en cada valor, utilizaremos expresiones regulares, una de ellas ya definida como: **value.match(/.*?(\d+).*/)[0]**

Esto le indica a openRefine, seleccióname los valores de cada columna que coinciden (value.match) con, cadena de texto al inicio (.*?), números sin importar el número de dígitos y que finalice con una cadena de texto (.*/).

[0] esto indica que del resultado, que es un vector, extrae solo el dato en la posición 0.

En los valores vemos los números que se han extraído por cada fila.

Le damos al botón **ACEPTAR**.

Añadir columna basada en otra ganancias

Nombre nuevo de la columna

En error cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value.match(/.*?(\d+).*/)[0]
1.	value 56 country of Angola	56
2.	value 552 country of Burundi	552
3.	value 187 country of Benin	187
4.	value 549 country of Burkina Faso	549
5.	value 204 country of Botswana	204
6.	value 1889 country of Central African Republic	1889

Aceptar Cancelar

Al darle aceptar vemos que se ha creado la columna **ganancias_num**, con los valores solo numéricos.

I	code2	ganancias	ganancias_num
	AO	value 56 country of Angola	56
	BI	value 552 country of Burundi	552
	BJ	value 187 country of Benin	187
	BF	value 549 country of Burkina Faso	549
	BW	value 204 country of Botswana	204
	CF	value 1889 country of Central African Republic	1889
	CI	value 2814 country of Cote d'Ivoire	2814
	CM	value 1804 country of Cameroon	1804
	CD	value 2298 country of	2298

Extraer un texto de una cadena de texto

Tenemos una variable llamada dominio que contiene diversas URL por país, queremos identificar el prefijo del dominio que corresponde a los tres últimos caracteres.

ganancias_num	dominio
56	data.colorado.gov
552	datacatalog.cookcountyil.gov
187	data.tompsc.com
549	venturaca.data.socrata.com
204	bronx.lehman.cuny.edu
1889	bronx.lehman.cuny.edu
2814	bcnv-internal.data.socrata.com
1804	data.mo.gov
2298	cityofyorbalinda.data.socrata.com

Vamos a crear una nueva columna basada en los datos de la columna dominio, damos click a la flecha en la variable, EDITAR COLUMNA > Agregar columna basada en nueva columna.

code2	ganancias	ganancias_num	dominio
AO	value 56 country of Angola	56	
BI	value 552 country of	552	gov
Bj			
BF			pm
BV			bronx.lehman.cuny.edu
CF			bronx.lehman.cuny.edu
CI			bcnv-internal.data.socrata.com
CM	value 1804 country of Cameroon	1804	data.mo.gov
CD	value 2298 country of Congo, The Democratic Republic of the	2298	cityofyorbalinda.data.socrata.com

Se abrirá la ventana de nueva columna y colocaremos por nombre **dominio_prefijo**, en Expresión utilizaremos otra expresión regular:

value.substring(value.length()-3)

Esta expresión indica, selecciona de la cadena el valor (**value.substring**), contabilizando el ancho del texto, menos tres caracteres del final (**value.length()-3**).

Con esto se seleccionan los prefijos, gov, com, edu, etc.

Añadir columna basada en otra dominio

Nombre nuevo de la columna

En error cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje No hay error de sintaxis.

Previsualización Historial Con estrella Ayuda

row	value	value.substring(value.length() ...
1.	data.colorado.gov	gov
2.	datacatalog.cookcountyil.gov	gov
3.	data.tompsc.com	com
4.	venturaca.data.socrata.com	com
5.	bronx.lehman.cuny.edu	edu
6.	bronx.lehman.cuny.edu	edu

Aceptar Cancelar

Se creará la columna con los prefijos de las URL

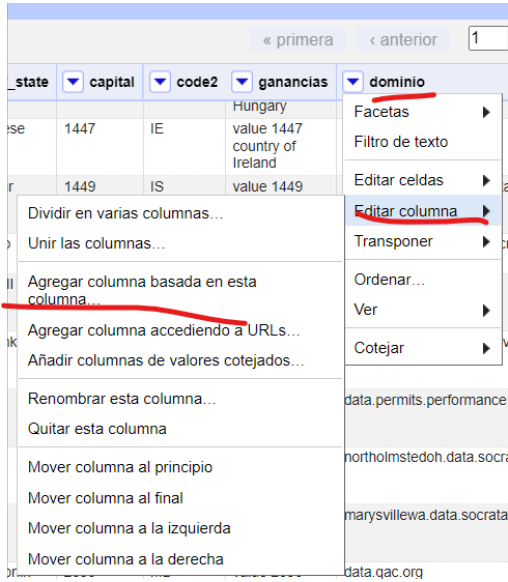
< anterior 1 siguiente > última »

um	dominio	dominio_prefijo
	data.colorado.gov	gov
	datacatalog.cookcountyil.gov	gov
	data.tompsc.com	com
	venturaca.data.socrata.com	com
	bronx.lehman.cuny.edu	edu
	bronx.lehman.cuny.edu	edu
	bcnv-internal.data.socrata.com	com
	data.mo.gov	gov
	cityofyorkbalinda.data.socrata.com	com
	data.novascotia.ca	.ca

Identificar si un texto está dentro de una cadena de texto

En esta práctica vamos a crear una nueva variable que identifique si los datos de la variable dominio, proviene de URL con la palabra **data** o **opendata**.

Crearemos una nueva variable basada en la columna dominio utilizando EDITAR COLUMNA > AGREGAR COLUMNA BASADA EN ESTA COLUMNA

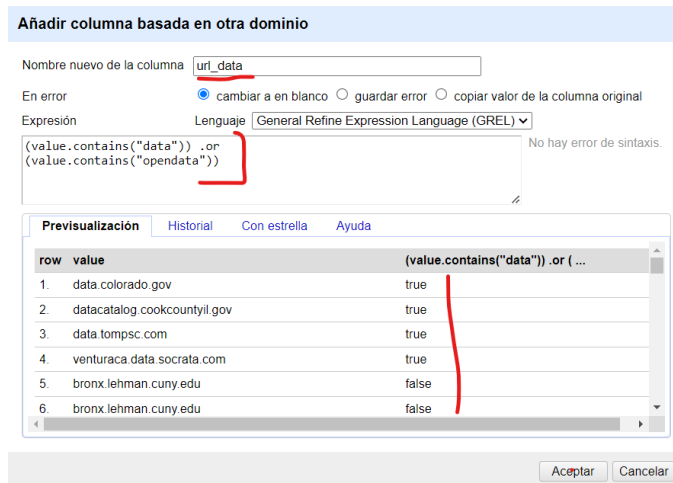


Le daremos el nombre de **url_data** a la nueva columna, para verificar que el valor de la URL en la variable dominio contienen las palabras buscadas utilizaremos la función `.contains()`.

En el espacio de Expresión colocaremos:

`value.contains("data") .or (value.contains("opendata"))`

Que colocará el valor de true si el valor del dominio contiene el texto buscado como se muestra en la imagen inferior.



El resultado es una nueva variable **url_data** con los datos **true** o **false**.

dominio	url_data
vusd.data.socrata.com	true
tetoncountyid.data.socrata.com	true
delaware-dshsst.data.socrata.com	true
data.kcmo.org	true
highways.hidot.hawaii.gov	false
data.permits.performance.gov	true
northolmstedoh.data.socrata.com	true
marysvillewa.data.socrata.com	true
data.qac.org	true
unioncityhi-reporting.data.socrata.com	true
www.data.va.gov	true
www.pivcide.pr	false
ntis.data.commerce.gov	true

¿Para qué sería funcional generar esta información?

Si generamos una faceta de texto de la variable **url_data**, podemos obtener como información que de los 239 países de la tabla, **202 almacenan** sus datos en fuentes de datos abiertas con una url estándar como la palabra **data** o **opendata**.

The screenshot shows the OpenRefine interface with a table of 239 rows. The 'url_data' column is faceted, showing 202 'true' and 37 'false' values. The table columns are: pop_life, gnp, gnp_old, local_name, government_form, Street, head_of_state, capital, code2, ganancias, and dominio. The facet for 'url_data' is expanded, showing the distribution of values.

Datos numéricos a categórico, condicional IF

En esta práctica vamos a crear una nueva variable con valores generados de una condición específica de los datos.

Transformamos los datos de la variable **life_expectancy** en formato de número.

239 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas « primera <

independence_year	year_text	year_num	population	life_expectancy	pop_life	gnp	gnp_old	local_name	government_form
1975	1975	1975-01-01T00:00:00Z	12878000	Facetas	362	6648	7984	Angola	Republic
1962	1962	1962-01-01T00:00:00Z	6695000	Filtro de texto				Burundi/uburundi	Republic
1960	1960	1960-01-01T00:00:00Z	6097000	Editar celdas					
1960	1960	1960-01-01T00:00:00Z	11937000	Editar columna					
1966	1966	1966-01-01T00:00:00Z	1622000	Transponer					
1960	1960	1960-01-01T00:00:00Z	3615000	Ordenar...	1960	1054	993		
1960	1960	1960-01-01T00:00:00Z	14786000	Ver	NULL	11345	10285		
1960	1960	1960-01-01T00:00:00Z	15085000	Cotejar	1976	9174	8596		
1960	1960	1960-01-01T00:00:00Z	51654000	Reemplazar...	1961	6964	2474	Republique Democratique du Congo	Republic
1960	1960	1960-01-01T00:00:00Z	2943000		1962	2108	2287	Congo	Republic

Luego utilizamos las Facetas numéricas en la variable **life_expectancy**, FACETAS > facetas numéricas.

El rango de número de esta variable según el histograma es de 37 a 84. Solo a manera de esta práctica sumaremos estos datos $37+84 = 121$ y lo dividiremos entre 2, $121*2 = 60.5$.

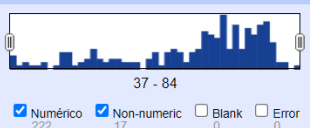
Este número nos servirá para indicarle a través de condicionales que, si el valor es mayor que 60.5, colocar como valor **Alta**, y si es menor, **Baja**

Facetas / Filtros

Deshacer / Rehacer 21 / 21

Actualizar Restablecer todos Quitar todo

life_expectancy cambiar restaurar



37 - 84

Numérico Non-numeric Blank Error

222 17 0 0

239 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas

independence_year	year_text	year_num	population	life_expectancy	pop_life	gnp	gnp_old
1975	1975	1975-01-01T00:00:00Z	12878000	Facetas			
1962	1962	1962-01-01T00:00:00Z	6695000	Filtro de texto			
1960	1960	1960-01-01T00:00:00Z	6097000	Editar celdas			
1960	1960	1960-01-01T00:00:00Z	11937000	Editar columna			
1966	1966	1966-01-01T00:00:00Z	1622000	Transponer			
				Ordenar...			
				Ver			
				Cotejar			
					39.3	1957	4834 4935

Crearemos una nueva variable que me permite incluir esta condicional.

population	life_expectancy	pop_life	gnp	gnp_old	local_name
12878000	39.3	1962	6648	7984	Angola
6695000	44	1991	903	982	Burundi/uburund
6097000	45.2	N			
11937000	45.2	N			
1622000	45.2	N			
3615000	44	1			/Be-
14786000	45.2	N			
15085000	54.8	1976	9174	8596	Cameroun/Came

Le colocaremos por nombre a la variable **life_respuesta**.

En **Expresión** colocaremos el siguiente código:

if(value > 60.5, "Alta", "Baja")

esto significa, si el valor de la columna actual es mayor a 60.5, colocar **Alta**, sino, **Baja**.

Damos click al botón **aceptar**.

Añadir columna basada en otra life_expectancy

Nombre nuevo de la columna:

En error: cambiar a en blanco guardar error copiar valor de la columna original

Expresión: No hay error de sintaxis.

Lenguaje:

Previsualización Historial Con estrella Ayuda

row	value	if(value > 60.5, "Alta", "Baja ...
1.	38.3	Baja
2.	46.2	Baja
3.	50.2	Baja
4.	46.7	Baja
5.	39.3	Baja
6.	44	Baja

Aceptar Cancelar

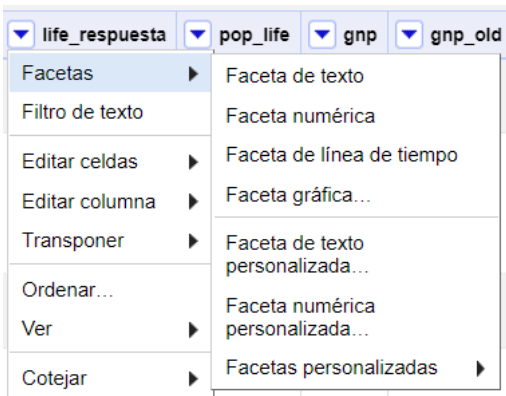
Se muestran los datos de la nueva variable con los valores de **Alta** y **Baja**.

life_expectancy	life_respuesta
38.3	Baja
46.2	Baja
50.2	Baja
46.7	Baja
39.3	Baja
44	Baja

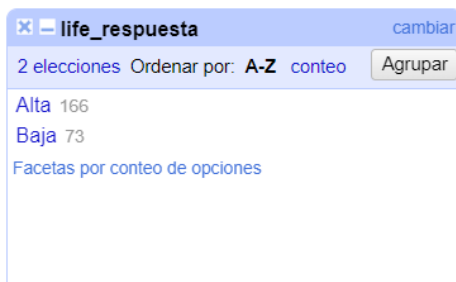
¿Para qué se transforman datos numéricos a una nueva variable con datos tipo texto?

La respuesta es , para poder categorizarlos o hacer facetas de texto para identificar los datos que están por encima del valor referenciado 60.5:

Facetamos los datos de la variable **life_respuesta**, **FACETAS > facetas de texto**.



Identificamos que hay **166** países con el valor **Alto**, que en conclusión tienen un **life_expectancy** por encima de **60.5**



NOTA: Si queremos crear una variable categórica de **life_expectancy** con tres categorías: podemos escribir en EXPRESIÓN al crear la nueva variable el siguiente código.

```
if(cells["life_expectancy"].value > 40.3 , "Alta",
if(cells["life_expectancy"].value < 40.3 , "Media", "Baja"))
```

Crear valores en una variable con datos de otras variables

Creamos una nueva columna utilizando como ejemplo la columna **year_num**.

year_num	population	life_expectancy	life_respuesta
Facetas	78000	83.5	Alta
Filtro de texto			
Editar celdas	473000	81.6	Alta
Editar columna	<ul style="list-style-type: none"> Dividir en varias columnas... Unir las columnas... Agregar columna basada en esta columna... Agregar columna accediendo a URLs... Añadir columnas de valores cotejados... Renombrar esta columna... Quitar esta columna Mover columna al principio Mover columna al final Mover columna a la izquierda Mover columna a la derecha 		
1901-01-01T00:00:00Z			
1499-01-01T00:00:00Z			
836			

En la ventana colocaremos por nombre a la columna **life_calculado**.

Resalto que los datos de la columna corresponden a la de **year_num**, sin embargo, podemos hacer referencia a una columna en particular utilizando la función **cells [nombre de la celda].value**.

En este ejemplo utilizaremos la columna **population** y dividiremos este valor entre 1000000. El código sería el siguiente:

`cells["population"].value / 1000000`

Añadir columna basada en otra year_num

Nombre nuevo de la columna:

En error: cambiar a en blanco guardar error copiar valor de la columna original

Lenguaje:

Expresión: No hay error de sintaxis.

Previsualización | Historial | Con estrella | Ayuda

48.	1975-01-01T00:00:00Z	0
74.	1991-01-01T00:00:00Z	4
144.	1991-01-01T00:00:00Z	4
176.	1974-01-01T00:00:00Z	0
107.	1991-01-01T00:00:00Z	6
234.	1966-01-01T00:00:00Z	0
112.	1991-01-01T00:00:00Z	24

Aceptar Cancelar

A ver el resultado del cálculo, vemos que el resultado de la división que debiera tener números decimales, pero no ocurre de esa forma.

La razón se debe a que openRefine esta dividiendo dos números enteros, por lo que la salida será un número entero. Para que el calculo genere un resultado con número decimal, uno de los dos valores debe ser con número decimal.

El código sería el siguiente:

cells["population"].value / 1000000.00

Con esto tendremos el resultado esperado.

Añadir columna basada en otra year_num

Nombre nuevo de la columna

En error cambiar a en blanco guardar error copiar valor de la columna original

Expresión Lenguaje

No hay error de sintaxis.

Previsualización	Historial	Con estrella	Ayuda
48.	1975-01-01T00:00:00Z	0.147	
74.	1991-01-01T00:00:00Z	4.968	
144.	1991-01-01T00:00:00Z	4.38	
176.	1974-01-01T00:00:00Z	0.094	
107.	1991-01-01T00:00:00Z	6.188	
234.	1966-01-01T00:00:00Z	0.861	
112.	1991-01-01T00:00:00Z	24.318	

Resultado de la nueva columna **life_calculado**.

life_calculado
12.878
6.695
6.097
11.937
1.622
3.615
14.786
15.085

Nota: puede que el calculo no se haga correctamente si hasta aquí has seguido todo el ejemplo, esto puede solucionarse eliminando el Orden de los valores que se había creado permanentemente, seleccionado QUITAR ORDEN.

500 1000 filas Ordenar ▾

on

Africa

Quitar orden

Reordenar filas permanentemente

Por Num ▶

Cargar datos de una URL

En este ejemplo cargaremos los datos de una URL, específicamente de una URL del API de la página ror.org.

Nos ubicamos en la opción DIRECCIONES WEB URL y copiamos el siguiente enlace:

<https://api.dev.ror.org/v2/organizations?query=Panama>

Crear un proyecto importando datos. ¿Qué tipo de archivos puedo importar?
Se admiten documentos de los tipos TSV, CSV, *SV, Excel (.xls y .xlsx), JSON, XML, RDF como XML y Google Data. Es posible añadir compatibilidad con otros formatos a través de las extensiones para C

Obtener datos de
Este equipo
Direcciones web (URLs)
Portapapeles
Database
Google Data

Ingrese una o más direcciones web (URLs) que dirijan a una descarga de datos:
https://api.dev.ror.org/v2/organizations?query=Panama

Añadir otra URL Siguiente →

Le damos click al botón siguiente.

Se mostrará la siguiente ventana con la estructura en formato JSON.

Habilite la opción **CARGAR AL MENOS**.

— volver a iniciar Configurar opciones del análisis sintáctico Nombre del proyecto organizations Etiquetas Crear proyecto —

```
{
  number_of_results: 22,
  time_taken: 128,
  items: [
    {
      admin: {
        created: {
          date: 2018-11-14,
          schema_version: 1.0
        },
        last_modified: {
          date: 2021-04-06,
          schema_version: 2.0
        }
      }
    }
  ]
}
```

Procesar los datos como

Archivos JSON

Archivos de texto basados en renglones

Archivos CSV/TSV/basados en separadores

Archivos de texto con campos de anchura fija

PC-Axis text files

Cargar al menos

Mantener las celdas vacías

Quitar espacios al inicio y final de las celdas

Procesar texto de celdas en números

Cargar el origen del archivo

almacenar archivo de almacenamiento

Especifique primero una ruta. Actualizar previsualización

Disable auto preview

0 fila(s) de datos

Le damos al botón CREAR PROYECTO.

Al cargar el proyecto, OpenRefine Nos indicará que hay un solo **registro**.

1 registros

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 registros

	_ - number_of_results	_ - time_taken	_ - items - _ - established	_ - items - _ - id	_ - items - _ - status	_ - items -
1.	22	3	1957	https://ror.org/03k9hrc16	active	
			2008	https://ror.org/03andxb27	active	
			2012	https://ror.org/00fpmwa05	active	
			1962	https://ror.org/022ka4k09	active	
			1984	https://ror.org/047pgsy79	active	
			1999	https://ror.org/012a4r663	active	es
			1935	https://ror.org/0070j0q91	active	

Debemos seleccionar la opción de **filas** para que se muestran todas las filas identificadas, ahora puede eliminar las filas en blanco y las columnas vacías.

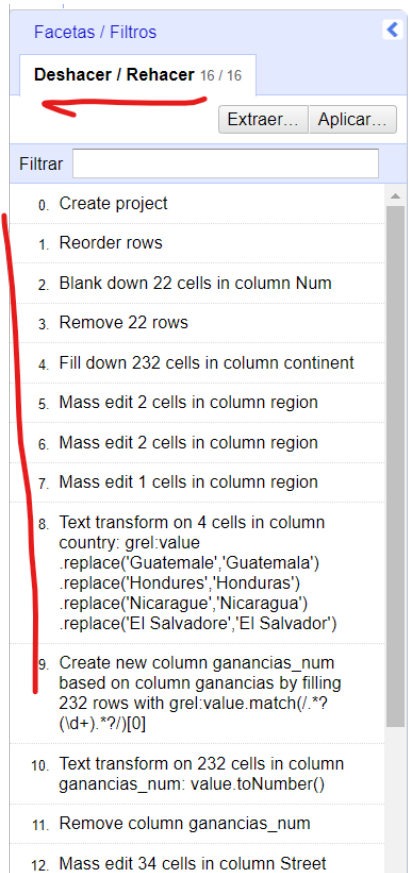
68 filas

Mostrar como: **filas** registros Mostrar: 5 10 25 50 100 500 1000 filas « primera < anterior

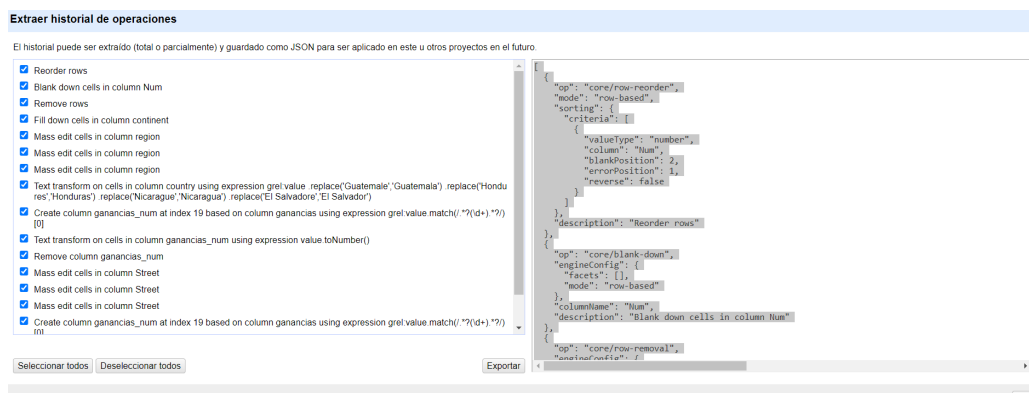
	_ - number_of_results	_ - time_taken	_ - items - _ - established	_ - items - _ - id	_ - items - _ - status	_ - items - _ - names - _ - lang	_ - item
1.	22	3	1957	https://ror.org/03k9hrc16	active		Florida Stat
2.							
3.							FSU-Panan
4.			2008	https://ror.org/03andxb27	active		Medistem (f
5.							
6.			2012	https://ror.org/00fpmwa05	active		VaxTrials (F
7.							
8.			1962	https://ror.org/022ka4k09	active		Museo De A
9.							
10.							MAC Panar

Guardar líneas de comandos en Open refine

En OpenRefine hay una opción de Deshacer y rehacer, esta contiene los comandos ejecutados para la limpieza de datos. Si queremos guardar estos datos para volverlos a ejecutar le damos click a la opción EXTRAER.



Se abrirá una nueva ventana que mostrará las líneas de comando y el código de los comandos en formato JSON a la derecha.



Seleccionamos este código, lo copiamos y lo guardamos en un archivo TXT o bloc de notas.

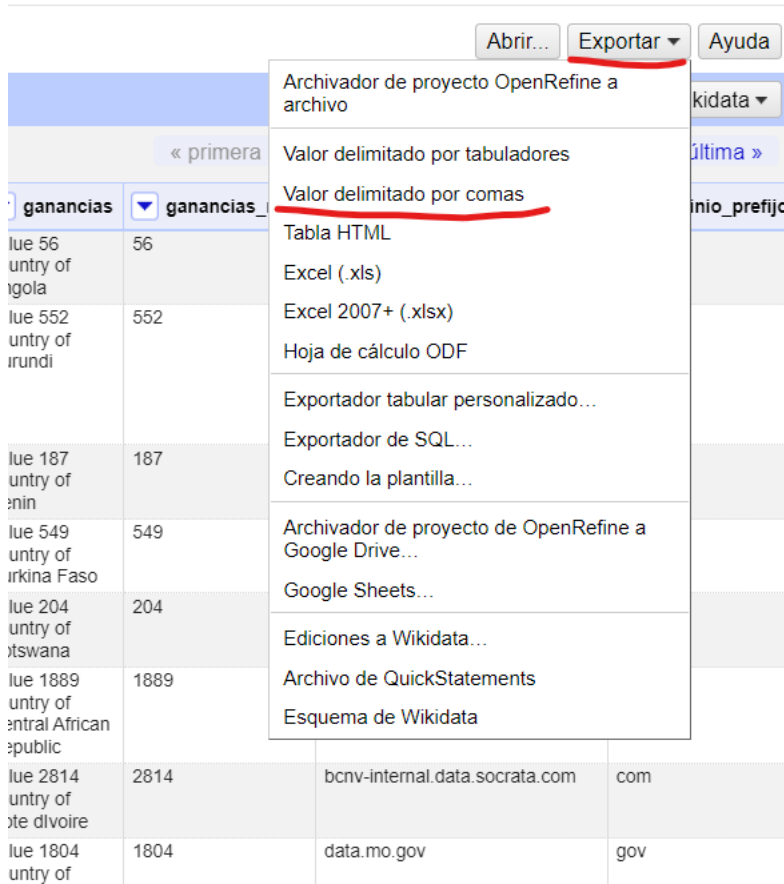
Este código puede ser reutilizado, si el proceso de limpieza con el mismo documento tiene que ser repetido constantemente.

```
rutina-limpieza-datos-ejemplo-fin × +
Archivo Editar Ver
{
  "fromError": false,
  "to": "Long Street. Sparkhill"
},
{
  "from": [
    "Moseley Village",
    "Moseley Village, Alcester road"
  ],
  "fromBlank": false,
  "fromError": false,
  "to": "Moseley Village"
}
],
"description": "Mass edit cells in column Street"
},
{
  "op": "core/column-addition",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "baseColumnName": "ganancias",
  "expression": "grel:value.match(/.*?(\\d+).*/)[0]",
  "onError": "set-to-blank",
  "newColumnName": "ganancias_num",
  "columnInsertIndex": 19,
  "description": "Create column ganancias_num at index 19 based on column ganancias using expression grel:v
},
{
  "op": "core/column-addition",
  "engineConfig": {
    "facets": [],
    "mode": "row-based"
  },
  "baseColumnName": "dominio",
  "expression": "grel:value.substring(value.length()-3)",
  "onError": "set-to-blank",
  "newColumnName": "dominio_abbrev"
}
```

Exportar documento

Para exportar el documento con los datos limpios damos click en el botón EXPORTAR ubicado en la parte superior derecha del OpenRefine. Seleccionamos valor delimitado por comas.

Le colocamos por nombre **country-data-clean-2024-02-05** y le damos click al botón guardar.



The screenshot shows the OpenRefine interface with the 'Exportar' dropdown menu open. The menu options are:

- Archivador de proyecto OpenRefine a archivo
- « primera
- Valor delimitado por tabuladores
- Valor delimitado por comas** (highlighted with a red underline)
- Tabla HTML
- Excel (.xls)
- Excel 2007+ (.xlsx)
- Hoja de cálculo ODF
- Exportador tabular personalizado...
- Exportador de SQL...
- Creando la plantilla...
- Archivador de proyecto de OpenRefine a Google Drive...
- Google Sheets...
- Ediciones a Wikidata...
- Archivo de QuickStatements
- Esquema de Wikidata

The background shows a table with columns 'ganancias' and 'ganancias_'. The data rows are partially visible, showing values like 56, 552, 187, 549, 204, 1889, 2814, and 1804.

Bibliografía

<https://openrefine.org/docs/manual/installing>

<https://programminghistorian.org/es/lecciones/limpieza-de-datos-con-OpenRefine>

<https://docs.gbif.org/openrefine-guide/3.0/es/>

<https://www.altergeosistemas.com/blog/2014/01/12/normalizacion-datos-openrefine/>

<https://labinoteca.com/2017/01/03/data-cleaning-con-open-refine/>

<https://www.juntadeandalucia.es/datosabiertos/portal/tutoriales/usar-openrefine.html>

<https://openup.org.za/blog/openrefine-aggregate-tutorial>

https://odl.ischool.uw.edu/openrefine_tutorial/

<http://www.padjo.org/tutorials/open-refine/clustering/>

<https://josepvalles.wordpress.com/2013/12/29/chuleta-guia-tutorial-open-refine-google-refine/>

<https://ror.readme.io/docs/openrefine-reconciler>