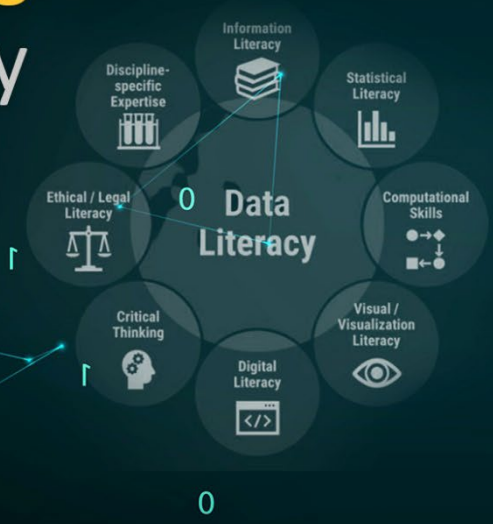
Alfabetización de Datos

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 d e 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

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 deseas 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

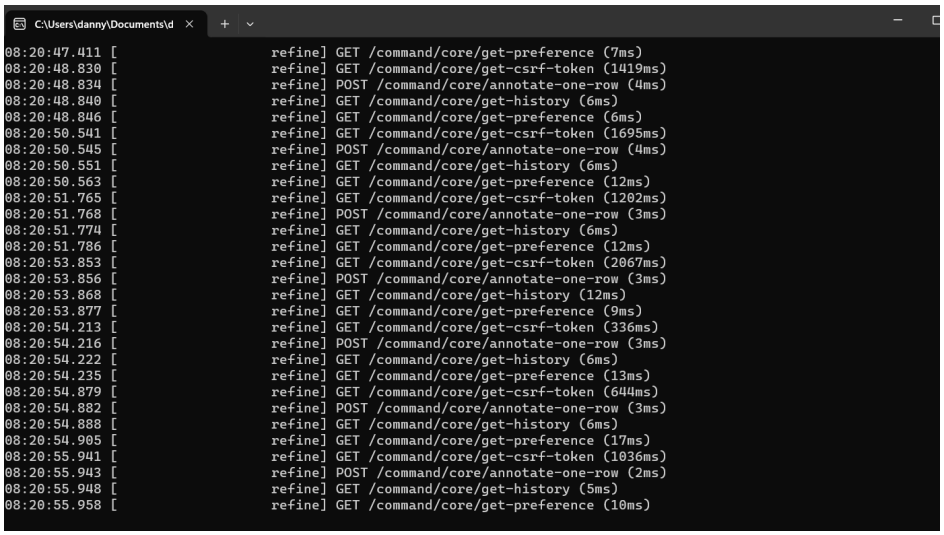
Open refine

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.

Pagina 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.



Conjunto de datos a utilizar:
El conjunto de datos tiene información de varias fuentes, entre ellas la practica 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

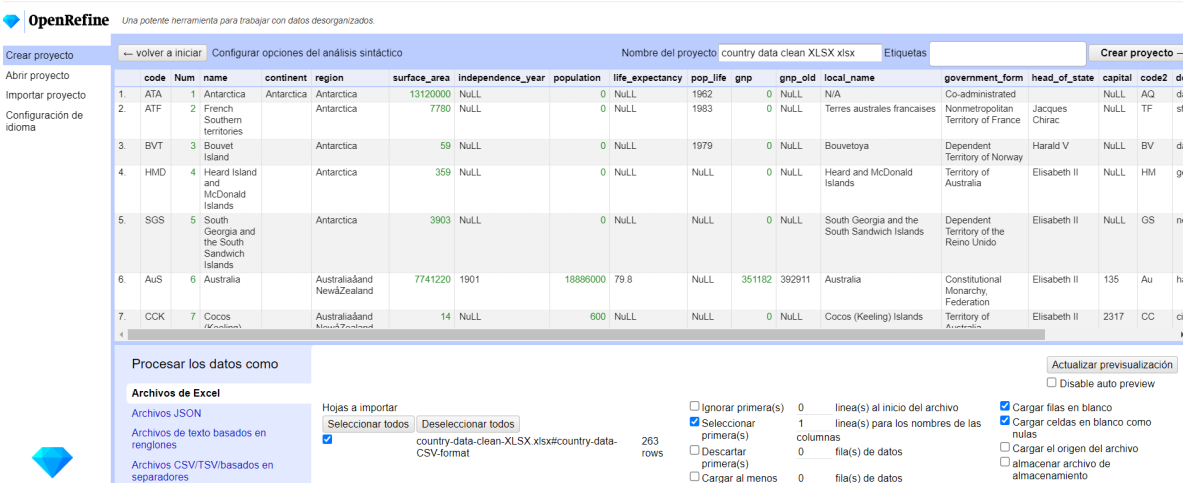
Carga de datos:

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.

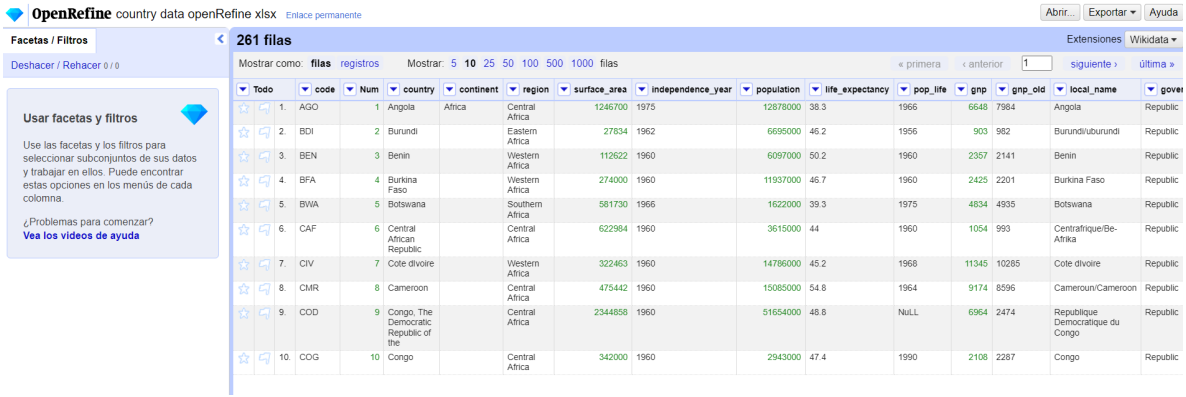
Al dar siguiente se mostrará la **estructura de la tabla, sus variables y algunos datos,** además en la parte inferior de la ventana se muestran con diversas opciones para modificar la lectura d ellos datos antes de cargarlos (ignorar primera fila, seleccionar primera fila, cargar filas en blanco , etc)



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.

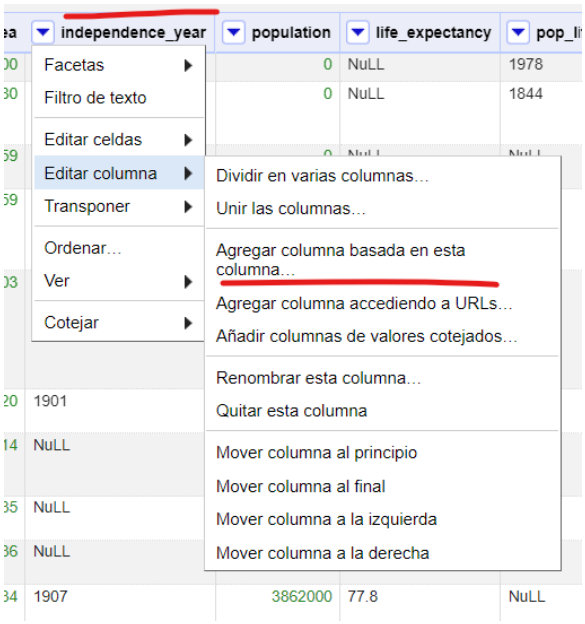
3. Se mostrará la tabla y sus variables, con el número de tablas identificadas 261, solo aparecerán 10 filas que pueden ser modificadas según la cantidad que deseamos mostrar. En esta ventana es posible modificar tanto la estructura de los datos como sus valores.
4. La carga de los datos puede tomar 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.



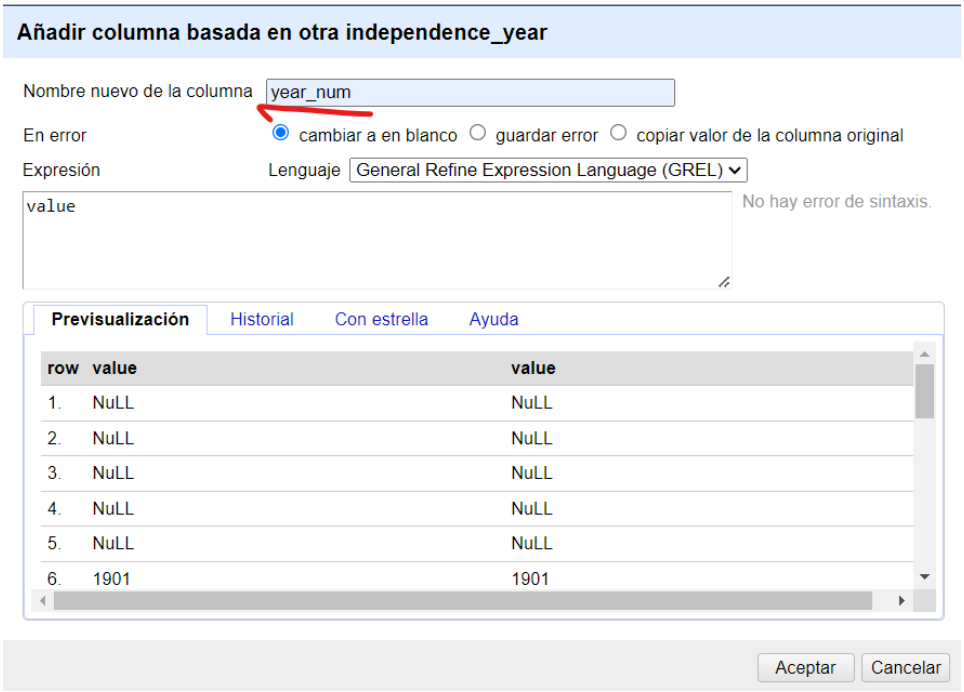
Todo		code	Num	country	continent	region	surface_area	independence_year	population	life_expectancy	pop_life	gnp	gnp_old	local_name	gover
1.	AGO	1	Angola	Africa	Central Africa	1246700	1975	12878000	38.3	1966	6648	7984	Angola	Republic
2.	BDI	2	Burundi		Eastern Africa	27834	1962	6695000	46.2	1956	903	982	Burundi/uburundi	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	4834	4935	Botswana	Republic
6.	CAF	6	Central African Republic		Central Africa	622984	1960	3615000	44	1960	1054	993	Centrafrique/Be-Afrika	Republic
7.	CIV	7	Cote dIvoire		Western Africa	322463	1960	14786000	45.2	1968	11345	10285	Cote dIvoire	Republic
8.	CMR	8	Cameroon		Central Africa	475442	1960	15085000	54.8	1964	9174	8596	Cameroun/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

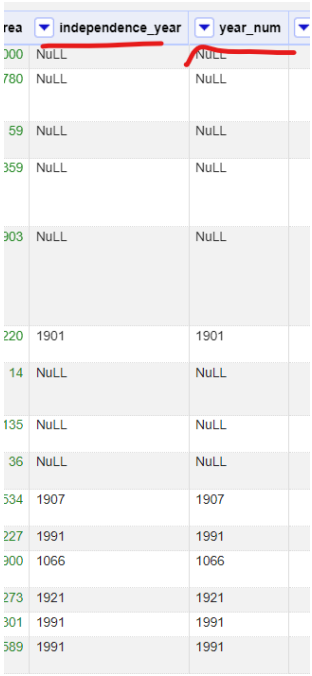
Podemos modificar las columnas seleccionado la flecha ubicada en cada columna, en este caso nos ubicaremos en la columna **independence_year**, se desplegará un menú con las opciones **EDITAR COLUMNA >** seleccione **AGREGAR columna basada en esta columna**.



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

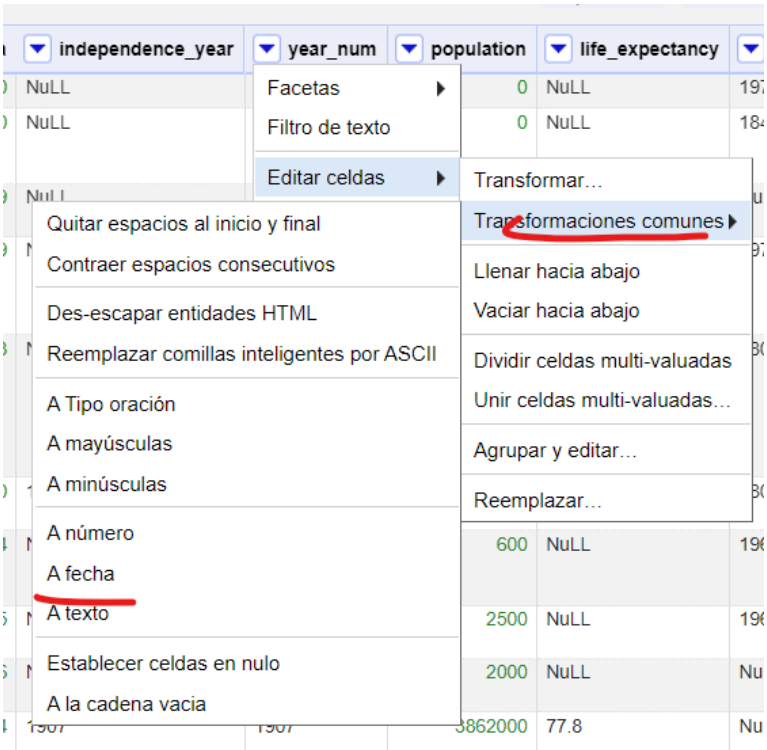


Al crear la nueva columna con los mismos datos se mostrará la nueva variable a la derecha, los datos alineados a la izquierda indican que OpneRefine los ha identificado como texto, aunque son años.

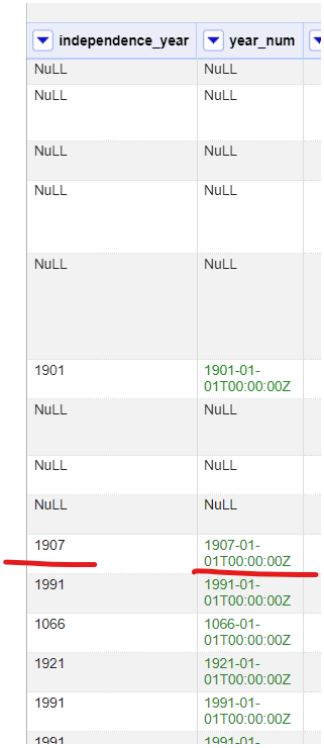


rea	independence_year	year_num
000	NuLL	NuLL
780	NuLL	NuLL
59	NuLL	NuLL
359	NuLL	NuLL
903	NuLL	NuLL
220	1901	1901
14	NuLL	NuLL
135	NuLL	NuLL
36	NuLL	NuLL
534	1907	1907
227	1991	1991
900	1066	1066
273	1921	1921
301	1991	1991
589	1991	1991

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.



Los datos de la columna year_num ahora tendra 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-

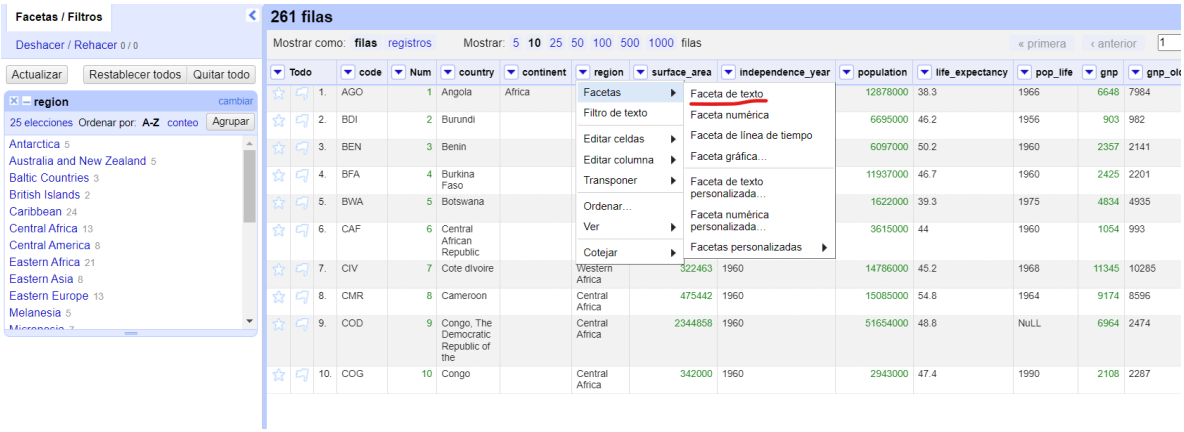
Facetas

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

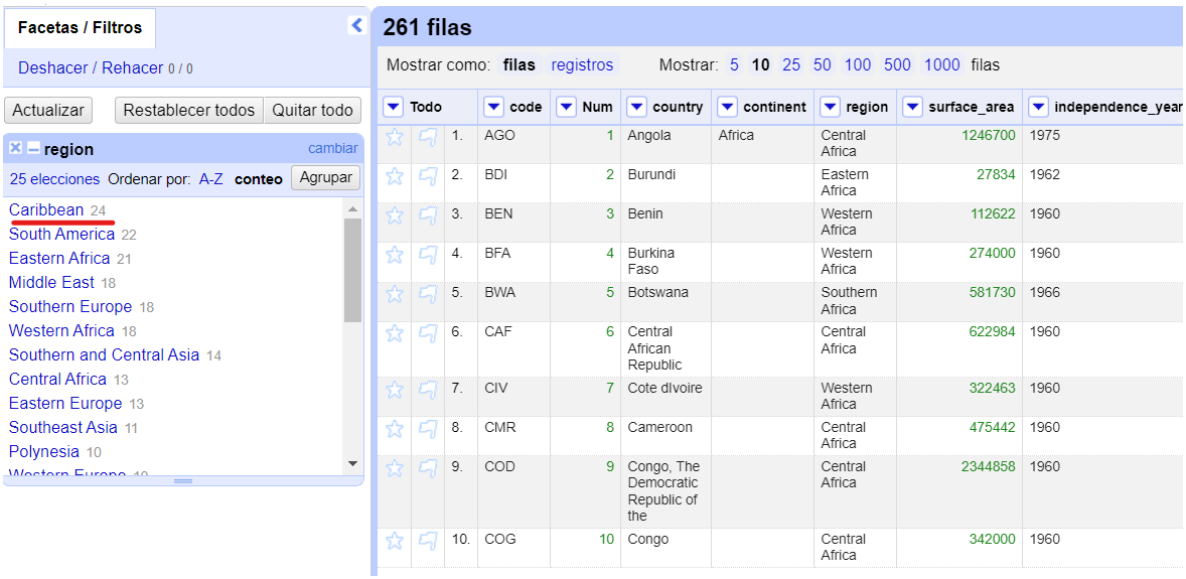
En la columna **region** daremos click a la flecha hacia abajo, seleccionaremos FACETAS > FACETAS DE TEXTO.

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



Se la 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 **Caribben** se repite 24 veces.

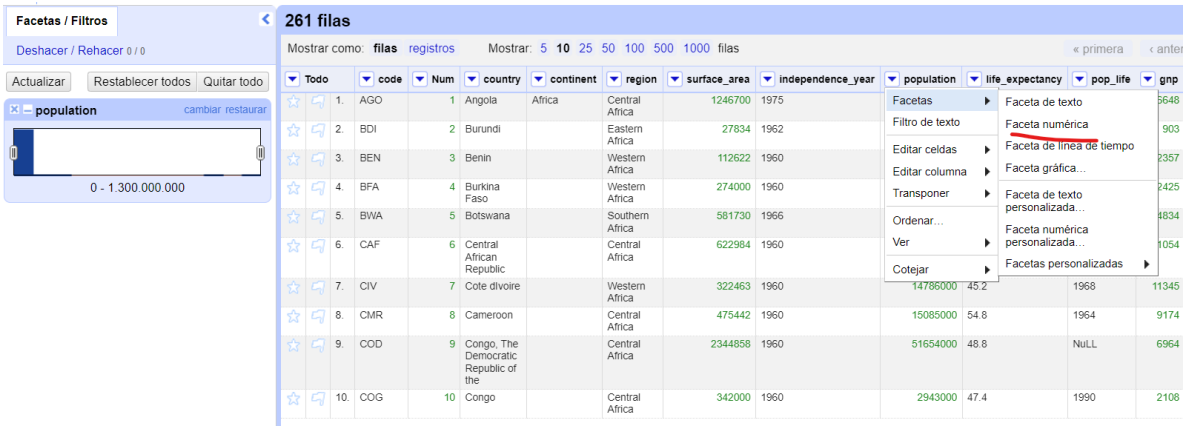


	Todo	code	Num	country	continent	region	surface_area	independence_year
1.		AGO	1	Angola	Africa	Central Africa	1246700	1975
2.		BDI	2	Burundi		Eastern Africa	27834	1962
3.		BEN	3	Benin		Western Africa	112622	1960
4.		BFA	4	Burkina Faso		Western Africa	274000	1960
5.		BWA	5	Botswana		Southern Africa	581730	1966
6.		CAF	6	Central African Republic		Central Africa	622984	1960
7.		CIV	7	Cote dIvoire		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.

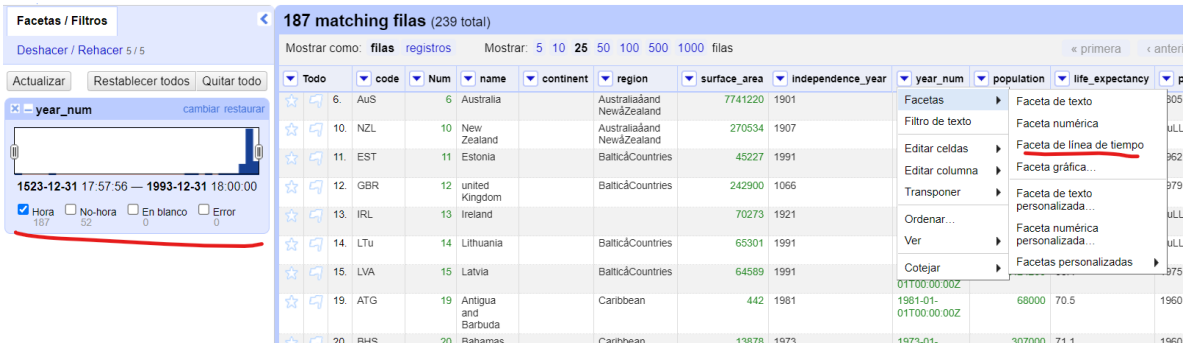
Se mostrará el lado izquierdo la faceta de la variable **population** y la distribución de los datos a través de un gráfico de histograma, esta muestra que el rango de los datos esta de 0 a 18000000 y la mayoría de los datos son 0.



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

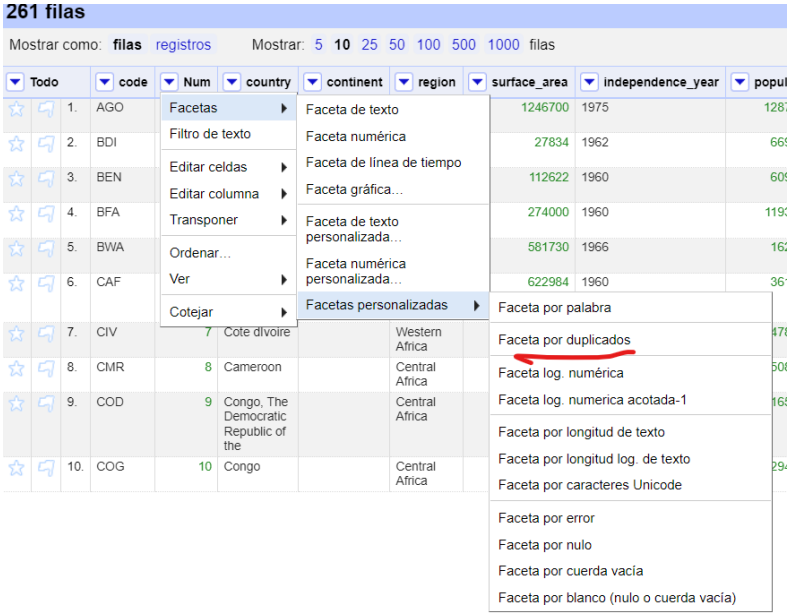
Se mostrará el lado izquierdo la faceta de la variable **year_num** y la distribución de los datos a través de un gráfico de histograma, esta muestra que el rango de los datos esta entre los años 1523 y 1993, con mayor cantidad de datos en los años de 1993.



Datos Duplicados (opción 1)

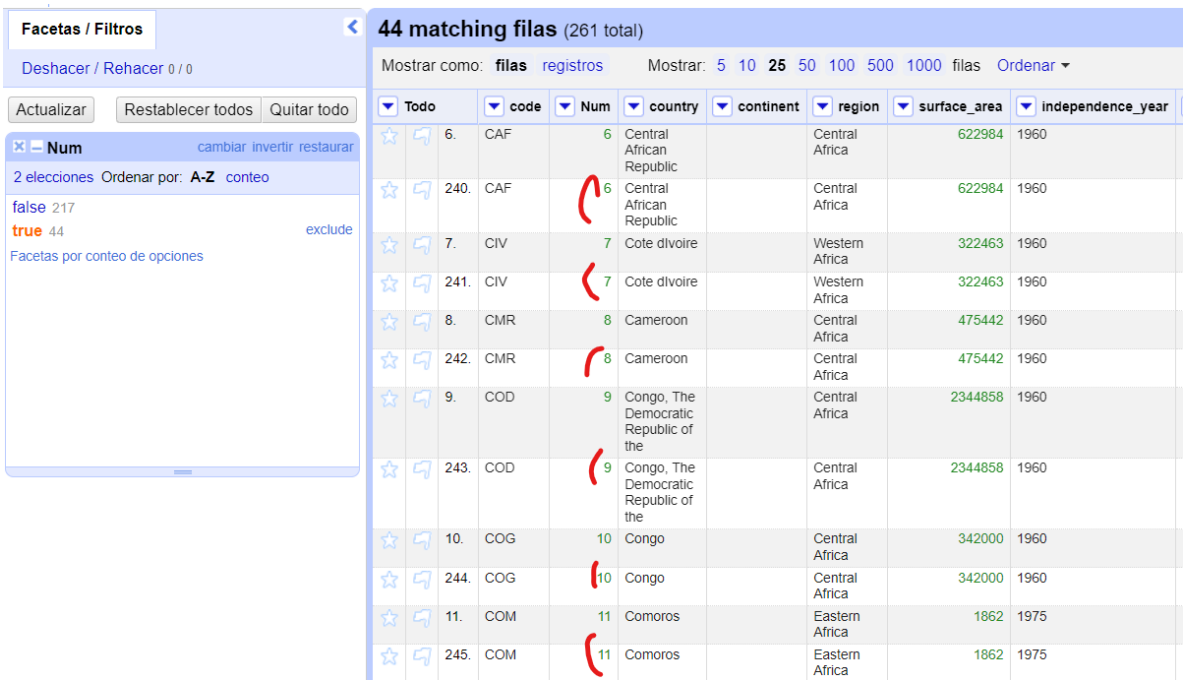
En el caso de los datos importados a Openrefine hay dos columnas que podemos utilizar para comparar los datos duplicados, **Num y Name** debido a que deben ser datos únicos en la tabla.

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

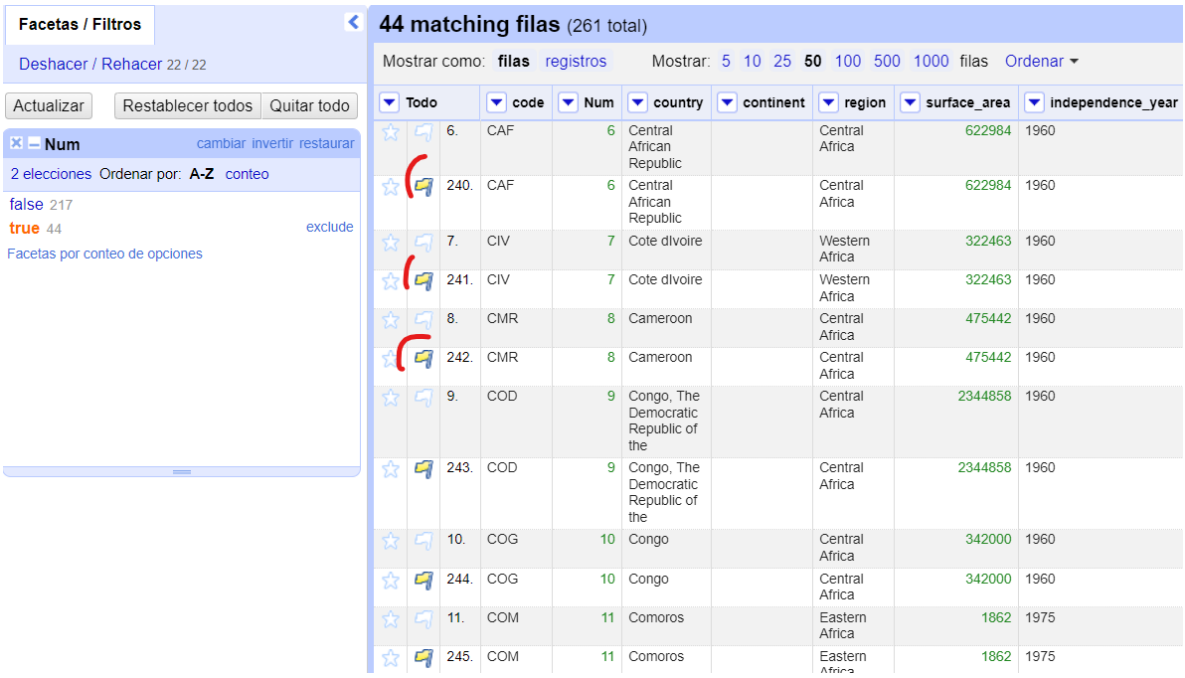


En la parte de la izquierda aparecerá una clasificación (FACETA) de la variable **Num** con las facetas identificadas con **TRUE** sí el valor de **NUM** se repite. En este caso se muestran 44 valores que se repiten.

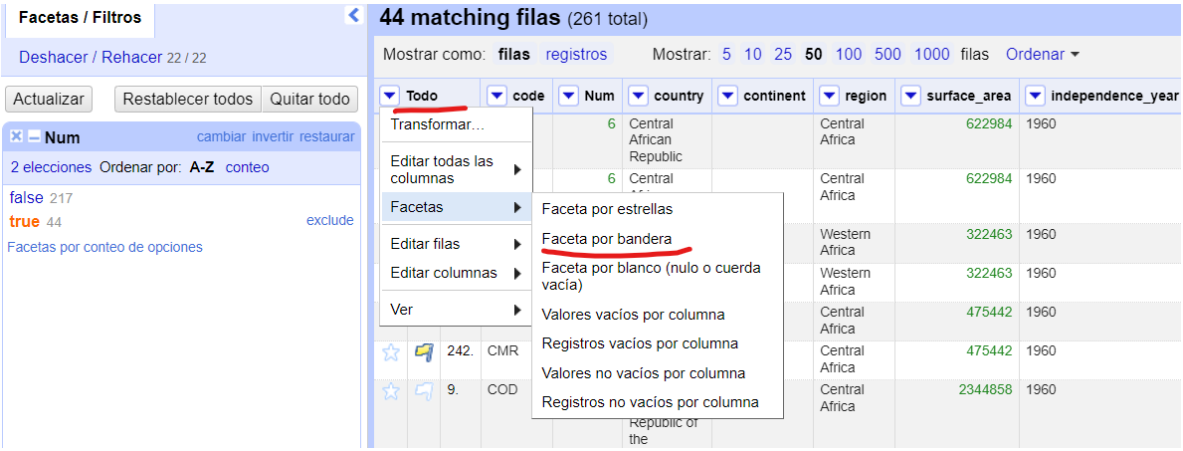
Al darle click en **TRUE** en esta faceta, se mostrará el listado de filas duplicadas por **NUM**, podemos ordenar la columna **NUM** por número para poder ver el listado en orden en el cual podemos comparar que en realidad hay valores que se repiten en esta columna (6,7,8,9,10,11...).



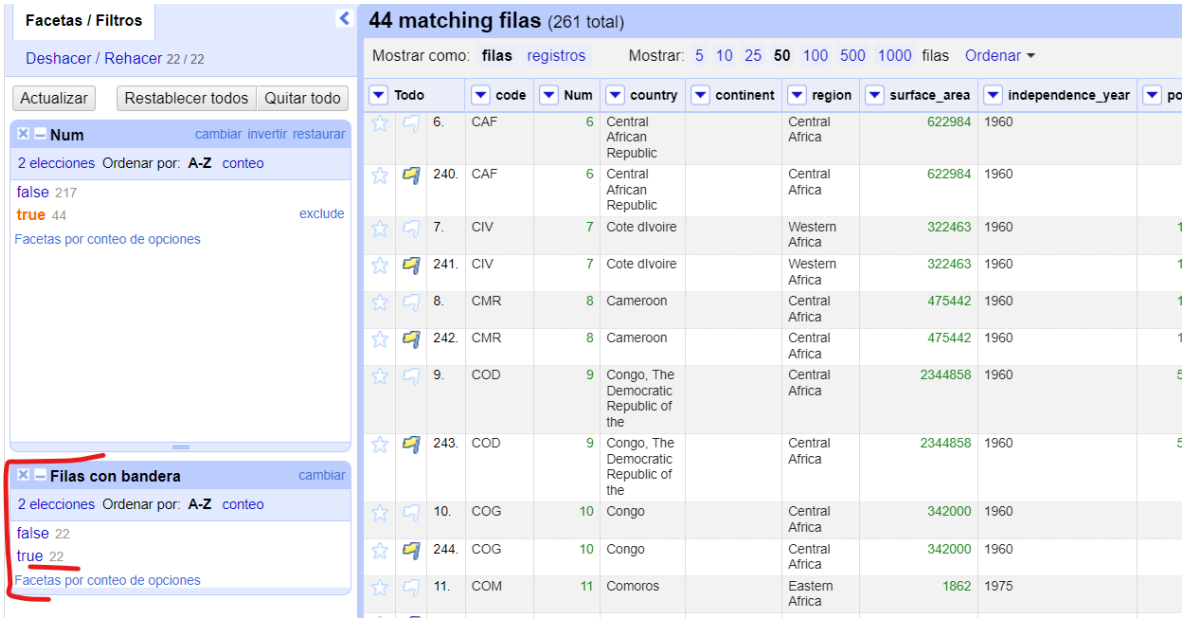
Si queremos eliminar las columnas duplicadas, debemos indicarle a OpenRefine cuales eliminar, para ellos podemos hacer uso de las estrellas o banderas que se muestran en cada fila. En este caso le daremos click a las banderas de todas las filas duplicadas como se muestra en la imagen.



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

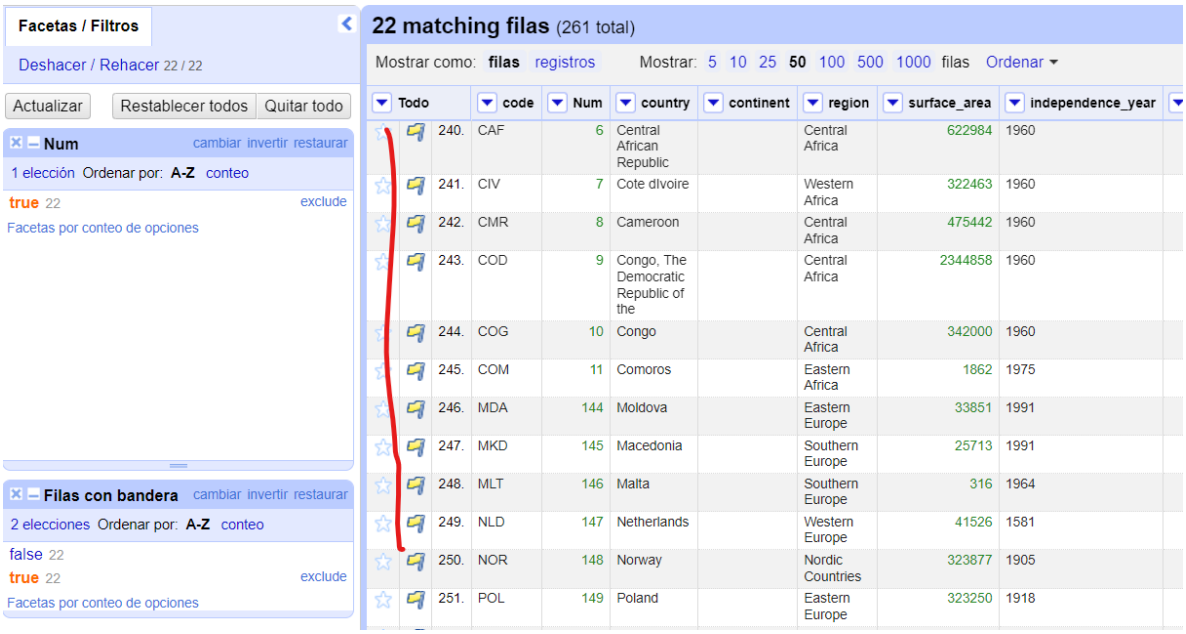


Se mostrará otras facetas de las filas con bandera señaladas con TRUE que en este caso son 22 filas y las que no tienen bandera señaladas con false.

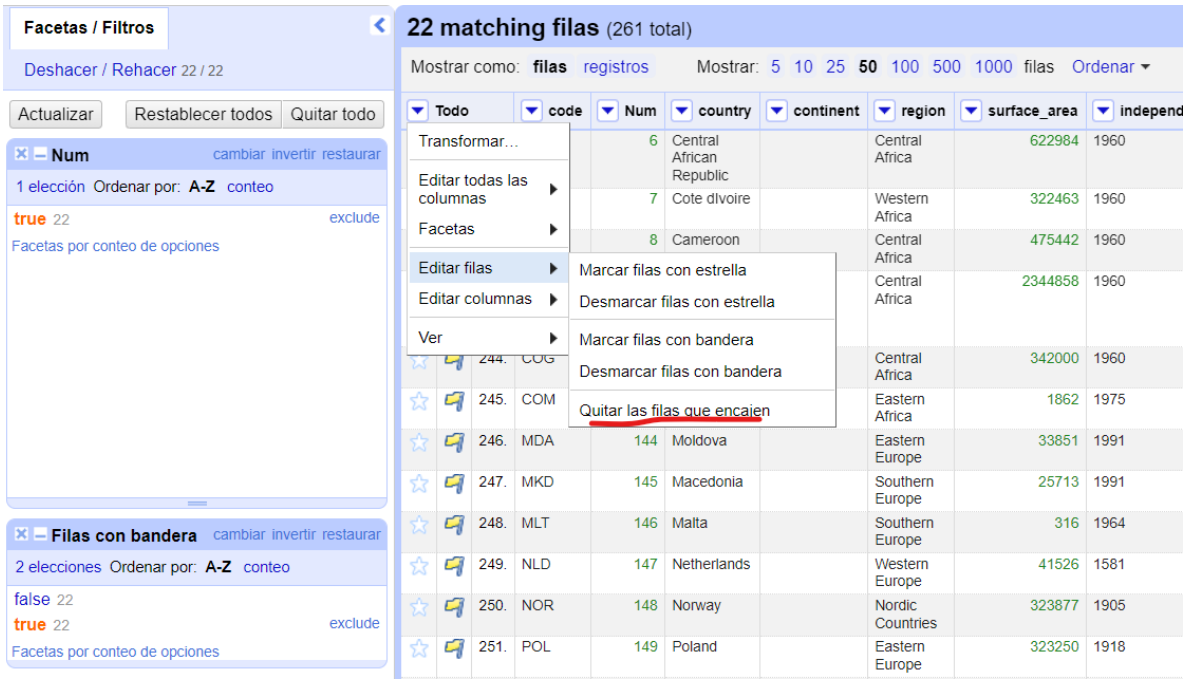


Todo	code	Num	country	continent	region	surface_area	independence_year
6.	CAF	6	Central African Republic		Central Africa	622984	1960
240.	CAF	6	Central African Republic		Central Africa	622984	1960
7.	CIV	7	Cote dIvoire		Western Africa	322463	1960
241.	CIV	7	Cote dIvoire		Western Africa	322463	1960
8.	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

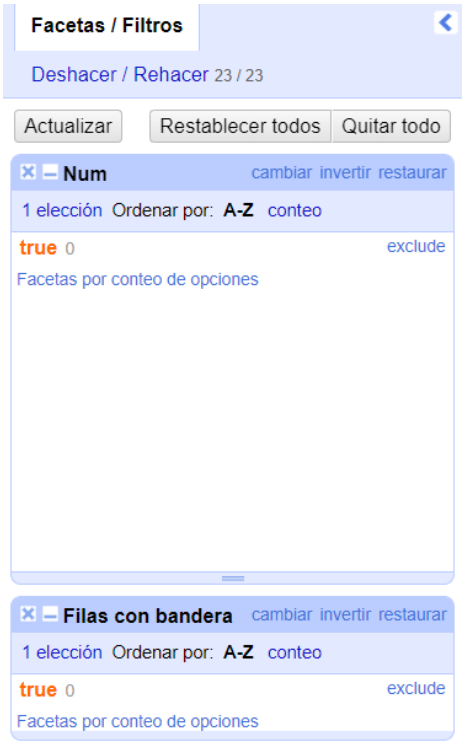
Al hacer click en **TRUE** de la faceta FILAS CON BANDERA se mostrará el listado de las filas que tienen bandera y que fueron seleccionadas como filas duplicadas.



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.



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.

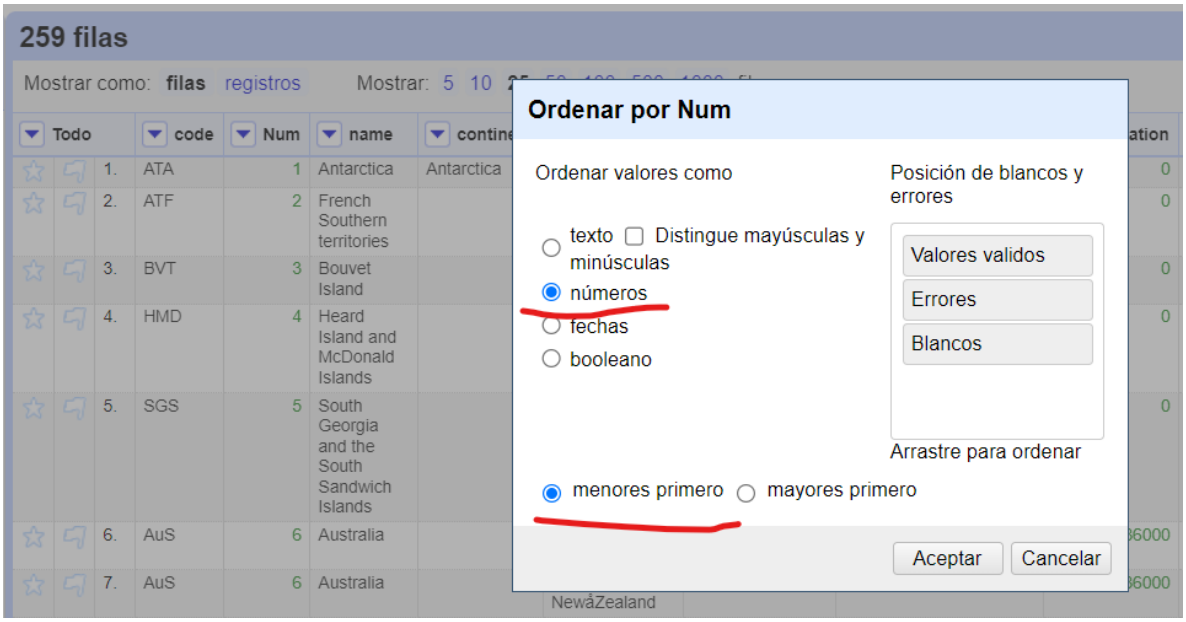


Nota: Aunque el proceso es funcional solo resulta viable si hay pocas filas, pero es un poco complicado con un número considerable de filas a eliminar, ya que habría que seleccionar la opción de bandera o estrella por cada uno, además que al ser un proceso de selección manual se puede cometer errores.

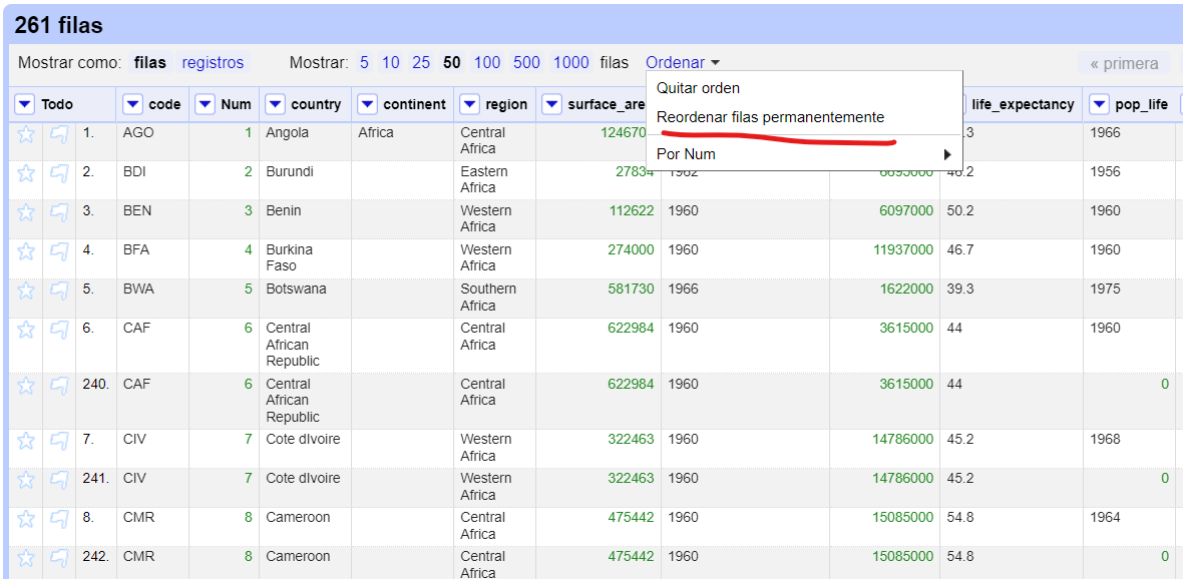
Datos Duplicados (opción 2)

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.



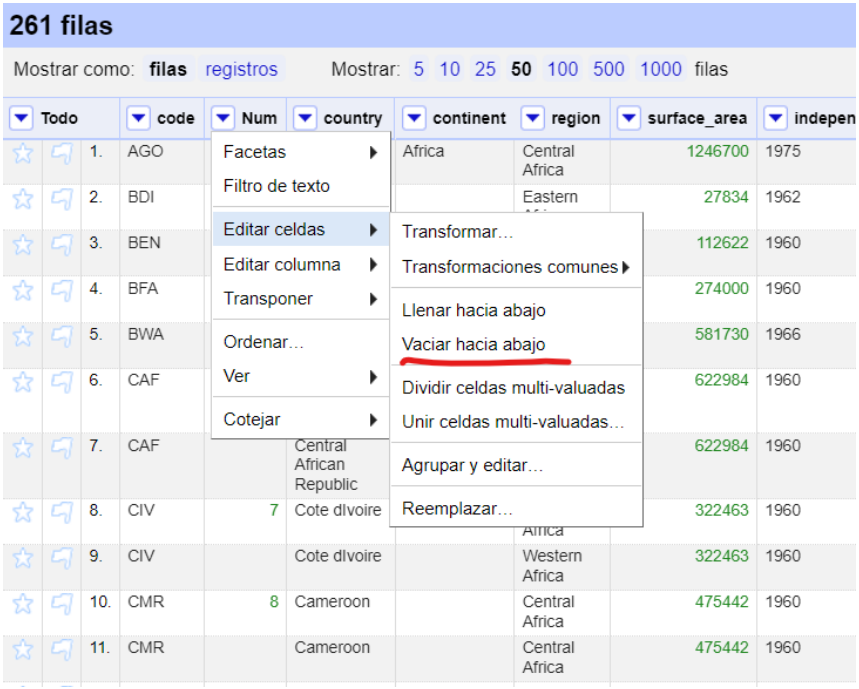
El reordenamiento que se hace es solamente Visual, porque necesitamos decirle a OpenRefine que el ordenamiento debe ser permanente ubicándonos en la parte superior, seleccionar ORDENAR > **Reordenar filas permanentemente.** Las con el mismo número (duplicadas) permanecerán debajo de las filas con el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



	Todo	code	Num	country	continent	region	surface_are			life_expectancy	pop_life
1.		AGO	1	Angola	Africa	Central Africa	124670			.3	1966
2.		BDI	2	Burundi		Eastern Africa	2783				1956
3.		BEN	3	Benin		Western Africa	112622	1960	6097000	50.2	1960
4.		BFA	4	Burkina Faso		Western Africa	274000	1960	11937000	46.7	1960
5.		BWA	5	Botswana		Southern Africa	581730	1966	1622000	39.3	1975
6.		CAF	6	Central African Republic		Central Africa	622984	1960	3615000	44	1960
240.		CAF	6	Central African Republic		Central Africa	622984	1960	3615000	44	0
7.		CIV	7	Cote dIvoire		Western Africa	322463	1960	14786000	45.2	1968
241.		CIV	7	Cote dIvoire		Western Africa	322463	1960	14786000	45.2	0
8.		CMR	8	Cameroon		Central Africa	475442	1960	15085000	54.8	1964
242.		CMR	8	Cameroon		Central Africa	475442	1960	15085000	54.8	0

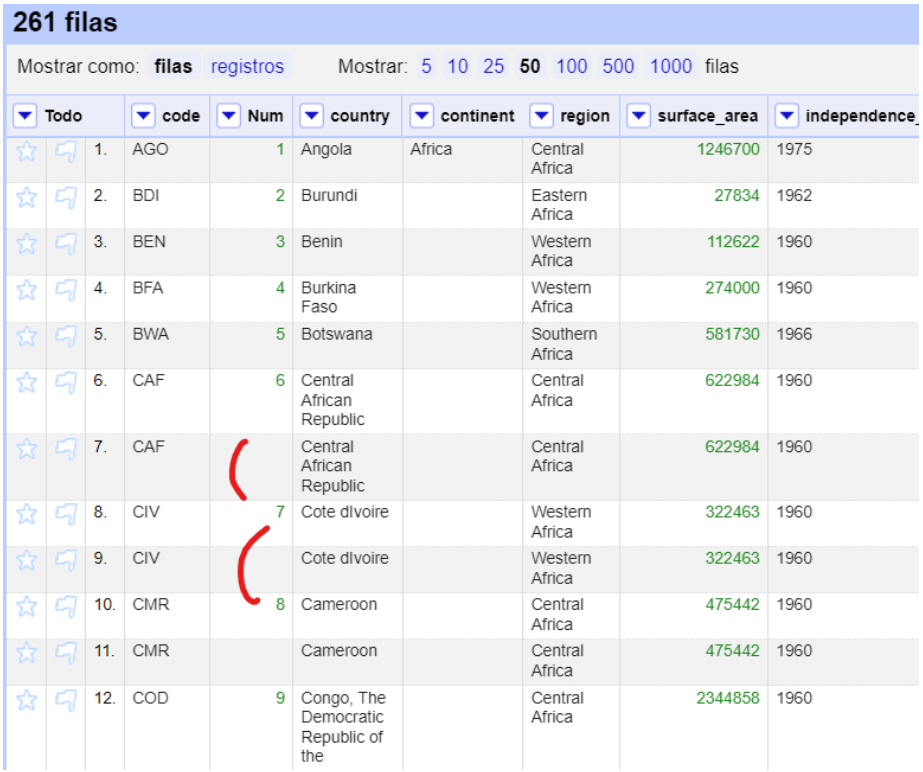
Necesitamos ahora identificar las filas con valores duplicados.

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



Todo	code	Num	country	continent	region	surface_area	indepen
1.	AGO			Africa	Central Africa	1246700	1975
2.	BDI				Eastern	27834	1962
3.	BEN					112622	1960
4.	BFA					274000	1960
5.	BWA					581730	1966
6.	CAF					622984	1960
7.	CAF		Central African Republic			622984	1960
8.	CIV	7	Cote dIvoire		Africa	322463	1960
9.	CIV		Cote dIvoire		Western Africa	322463	1960
10.	CMR	8	Cameroon		Central Africa	475442	1960
11.	CMR		Cameroon		Central Africa	475442	1960

Las filas con los valores repetidos ahora no tienen valores.

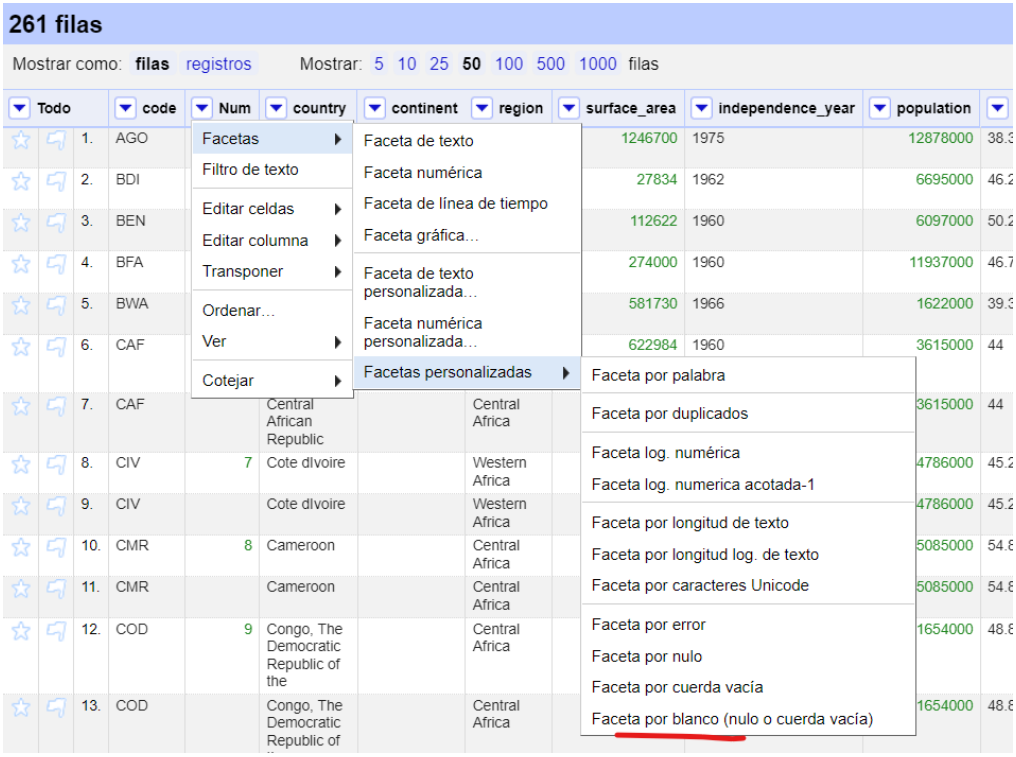


261 filas

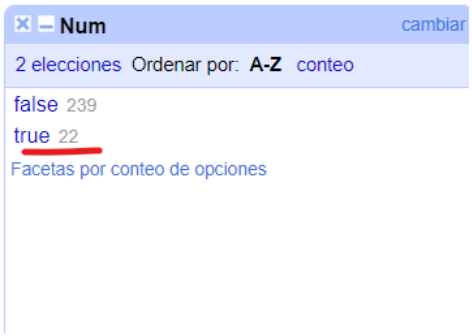
Mostrar como: **filas** registros Mostrar: 5 10 25 **50** 100 500 1000 filas

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		Central African Republic		Central Africa	622984	1960
8.	CIV	7	Cote dIvoire		Western Africa	322463	1960
9.	CIV		Cote dIvoire		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.

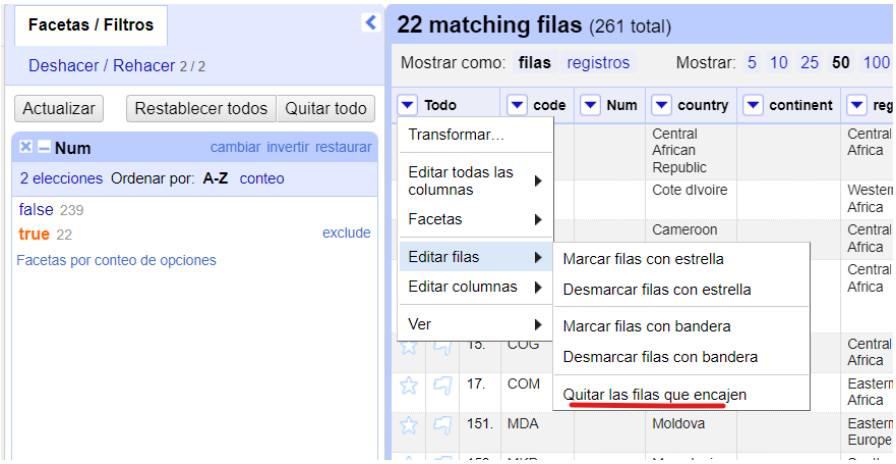


Se mostrará la faceta con los datos vacíos (22) TRUE y los no vacíos (239).

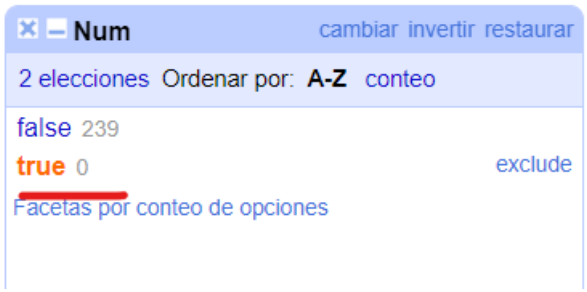


Damos click en **TRUE**, el cual desplegará solo los datos vacíos que en teoría son los duplicados.

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

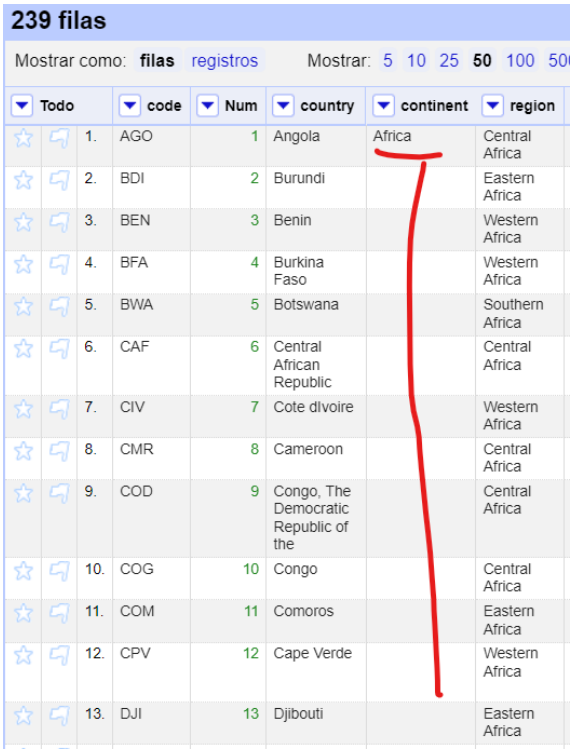


Aparecerán 0 filas en TRUE de la faceta **Num,** indicando que los datos duplicados fueron eliminados. Recuerde darle click al botón **QUITAR TODO** para eliminar las facetas y ver el resultado.



Rellenar espacios en blanco

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



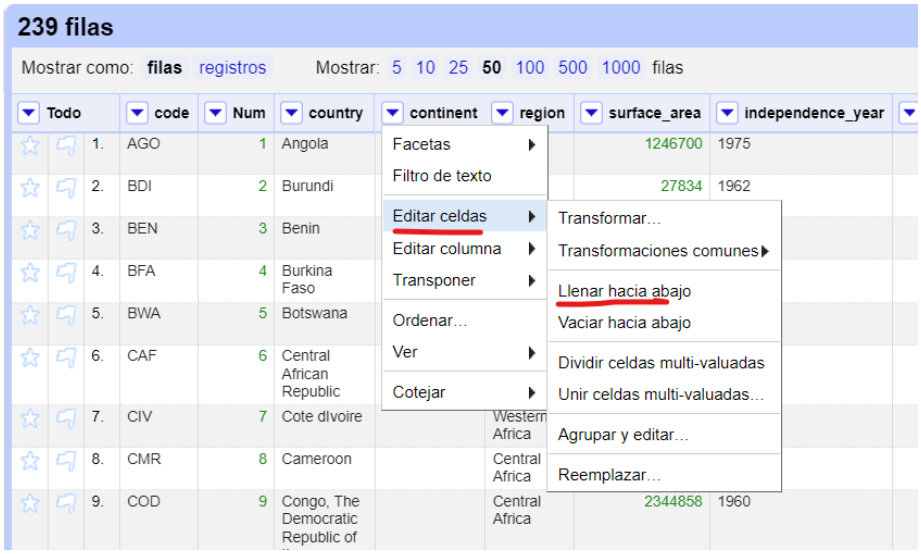
239 filas

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

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 dIvoire		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 aparecen 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 rellenaran con el dato que esta 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.**



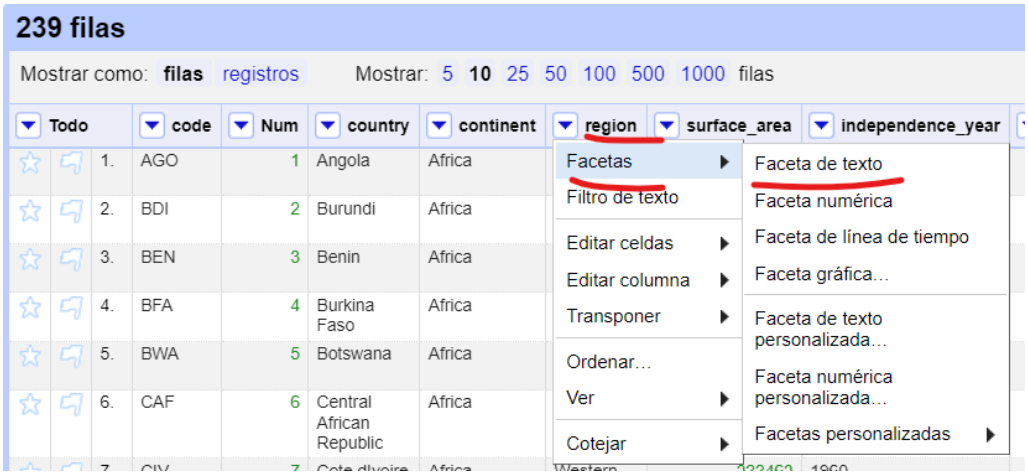
Las filas con espacios en blanco ya fueron llenados con el valor que estaba en la fila posterior a las filas en blanco.

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

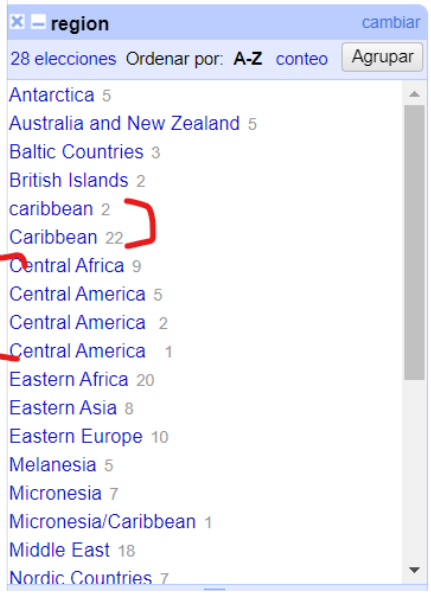
Unificar datos categóricos

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.

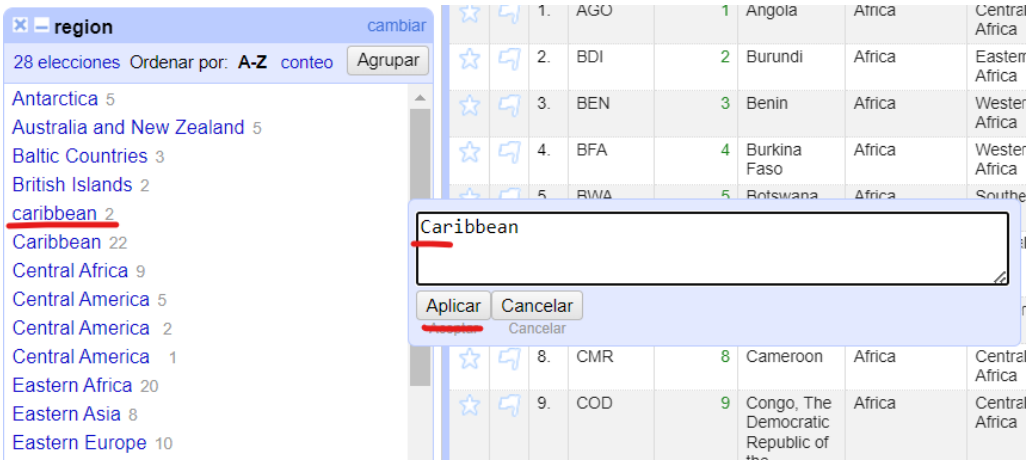
En este ejemplo utilizaremos la variable **region**. Seleccionamos la flecha hacia abajo FACETAS > facetas de texto.



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 **caribeann**, 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.

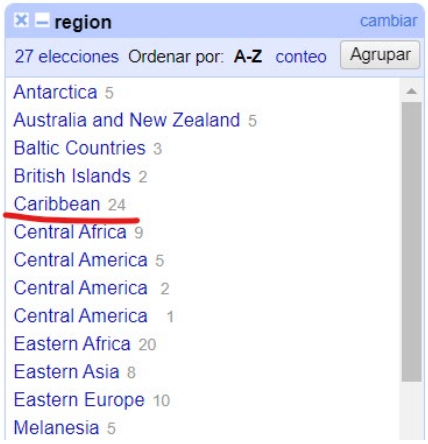


Si nos ubicamos al lado de la categoría **caribbean** vemos que aparecen dos opciones, editar e **include**. Si le damos a include, nos mostrará las filas donde se muestra los datos que tienen **caribbean** en minúscula. En este caso le daremos **EDITAR**.



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.



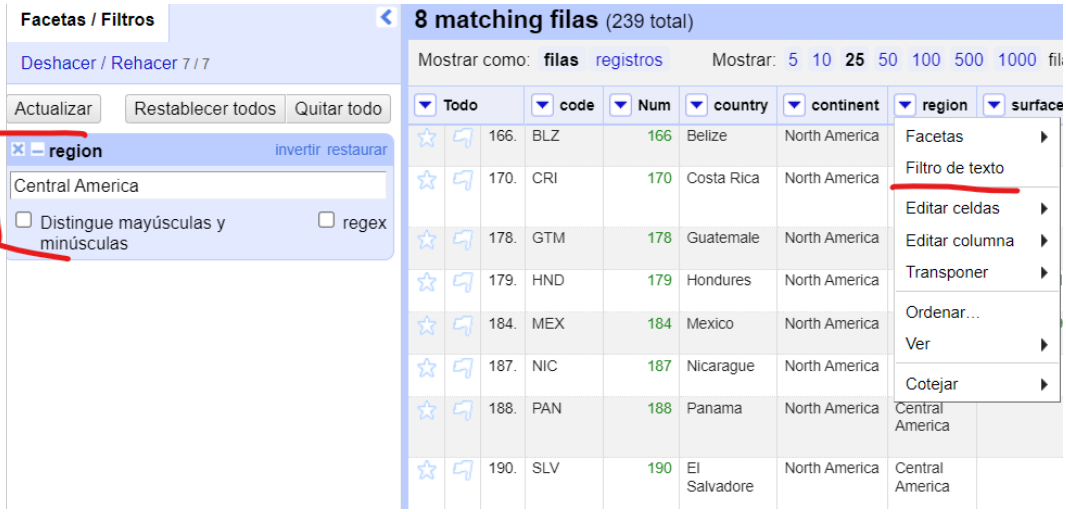
Al realizar los mismos pasos con la categoría **Centra America**, eliminamos los espacios en blanco al final de los tres textos con esa categoría y obtendremos esa categoría con 8 filas.

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.

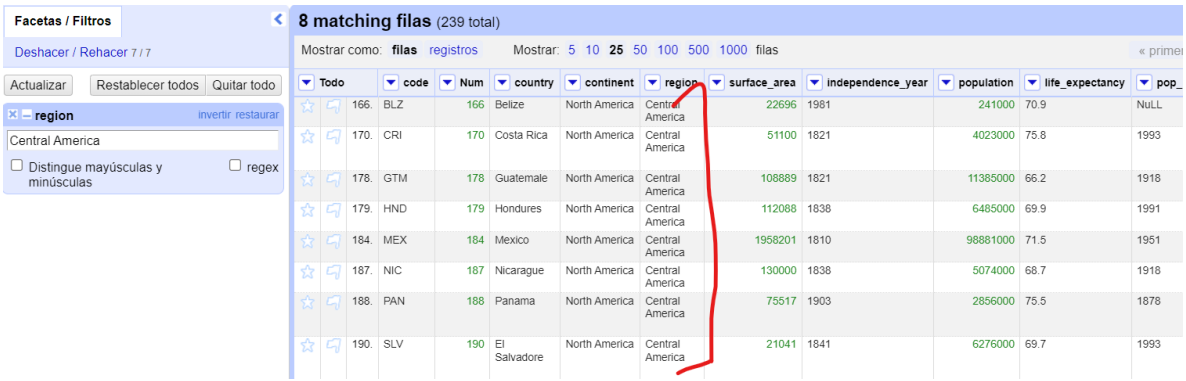
Tomando en cuenta que la categoría **Central America**, y que corregimos en la faceta. Haremos un filtro de estos datos por la variable **región**.

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**.

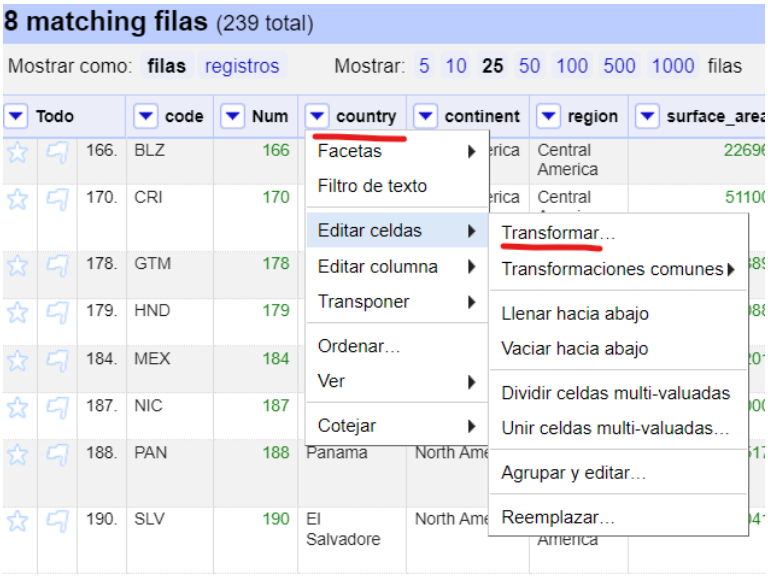


Se mostrará un listado de los países cuya región es **Central América**.

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

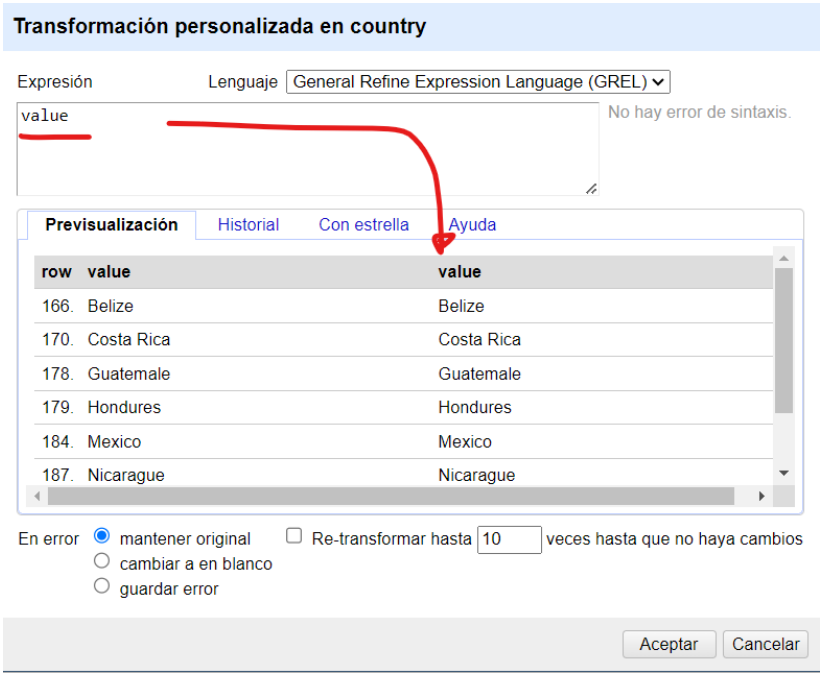


Existen varias formas de reemplazar los datos, para generar un nuevo aprendizaje utilizaremos la función transformar, que se encuentra, en la variable country, click en la flecha, EDITAR CELDAS > TRANSFORMAR.



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.

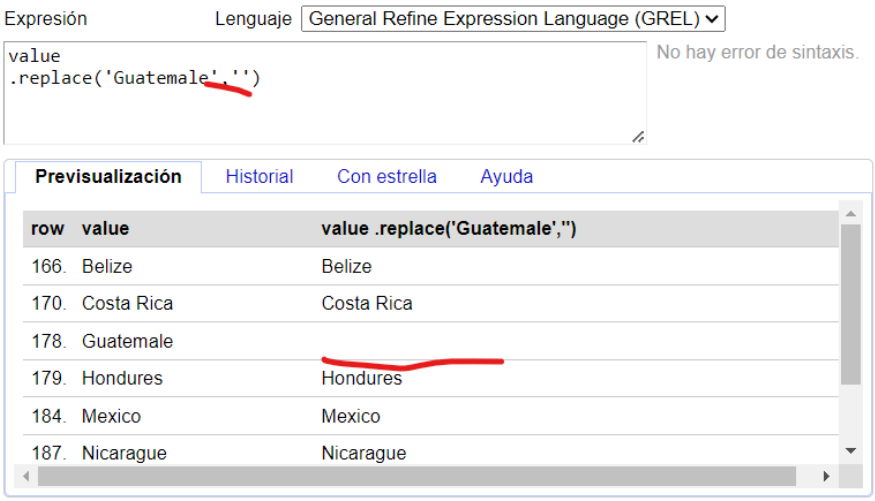


Para este ejemplo utilizaremos la función .repace(“dato actual”, “nuevo dato”).

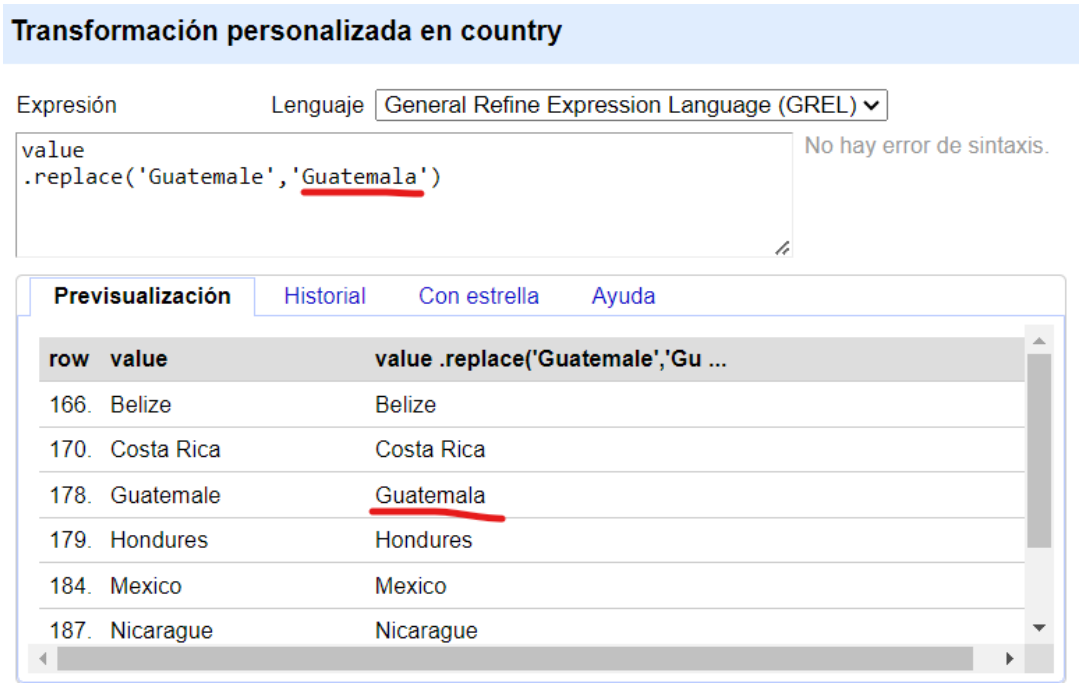
Si escribimos la siguiente expresión:

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

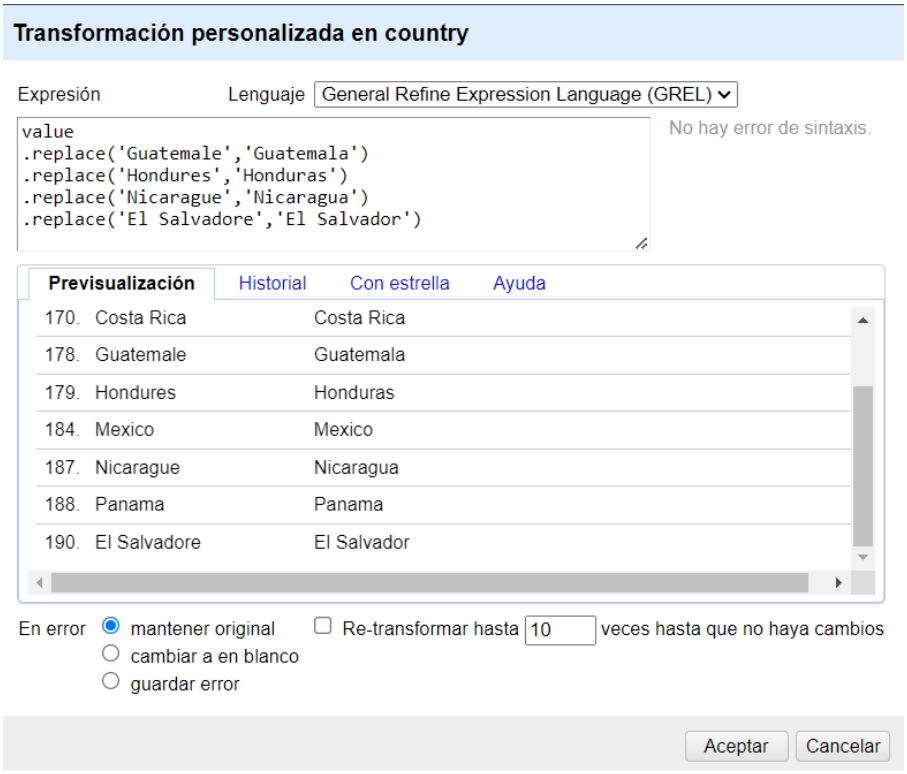
Se modificará el valor de Guatemale por un valor vacío



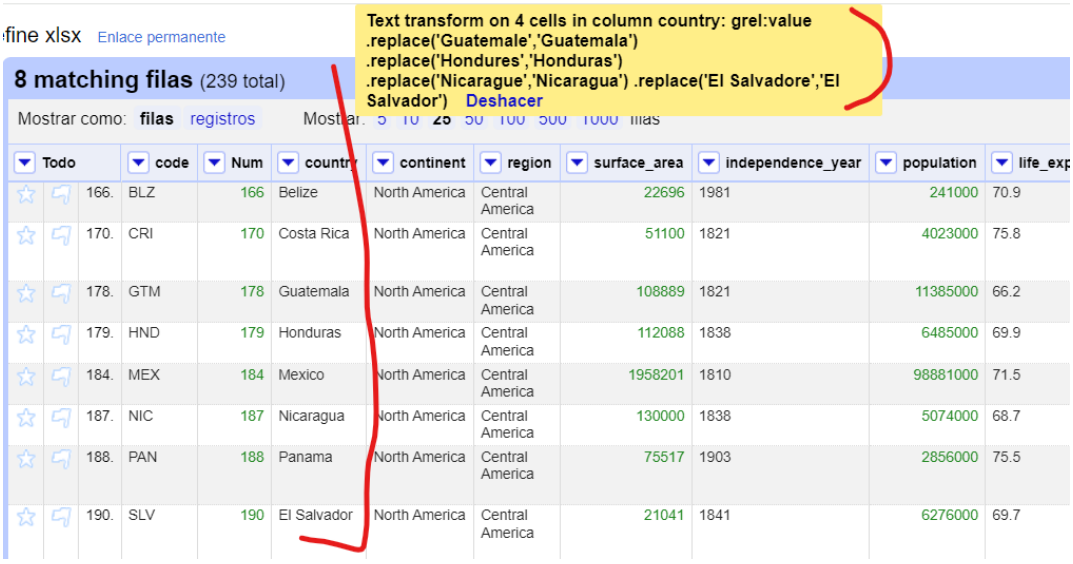
Si completamos la expresión por el valor a reemplazar se mostrará la palabra **Guatemala** de forma correcta.



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



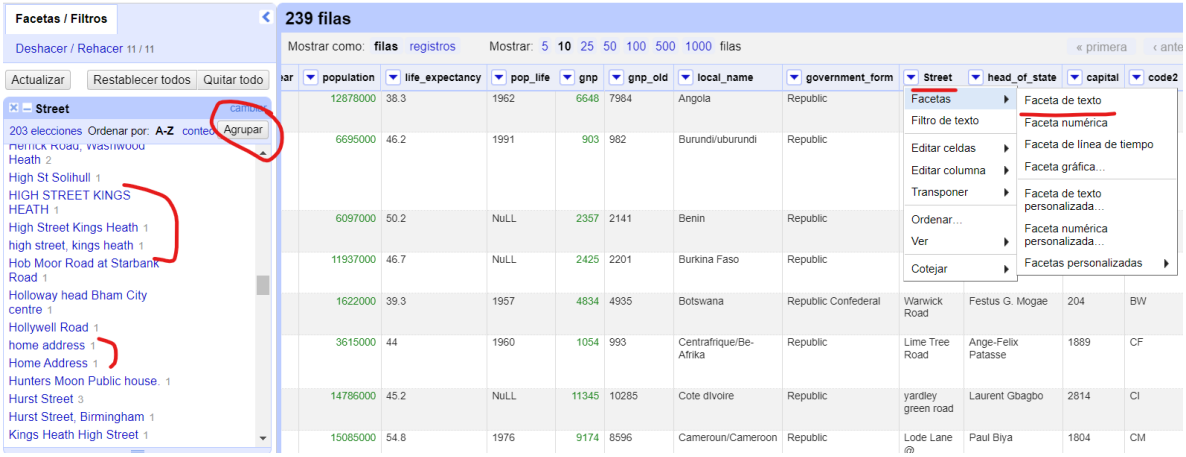
Le damos click a **ACEPTAR** y las filas en la tabla se actualizarán con los nuevos datos.



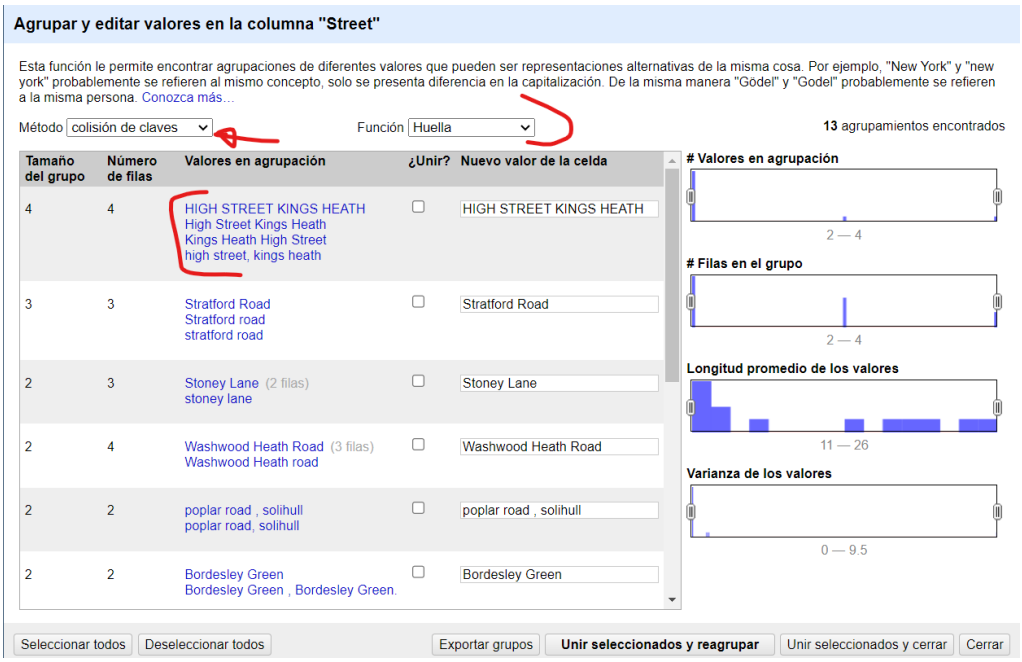
Cluster (Agrupar)

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.

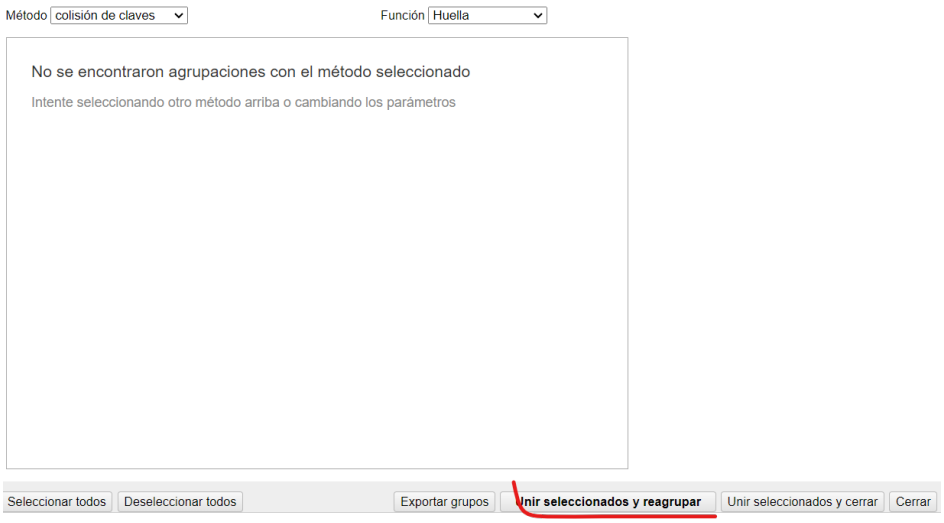


En cada grupo se nos da la opción de modificar los **valores por agrupación** por el **nuevo valor de la celda (el cual puede ser modificado)**, si deseo unir los valores con el mismo dato, debo seleccionar la opción de **UNIR**.

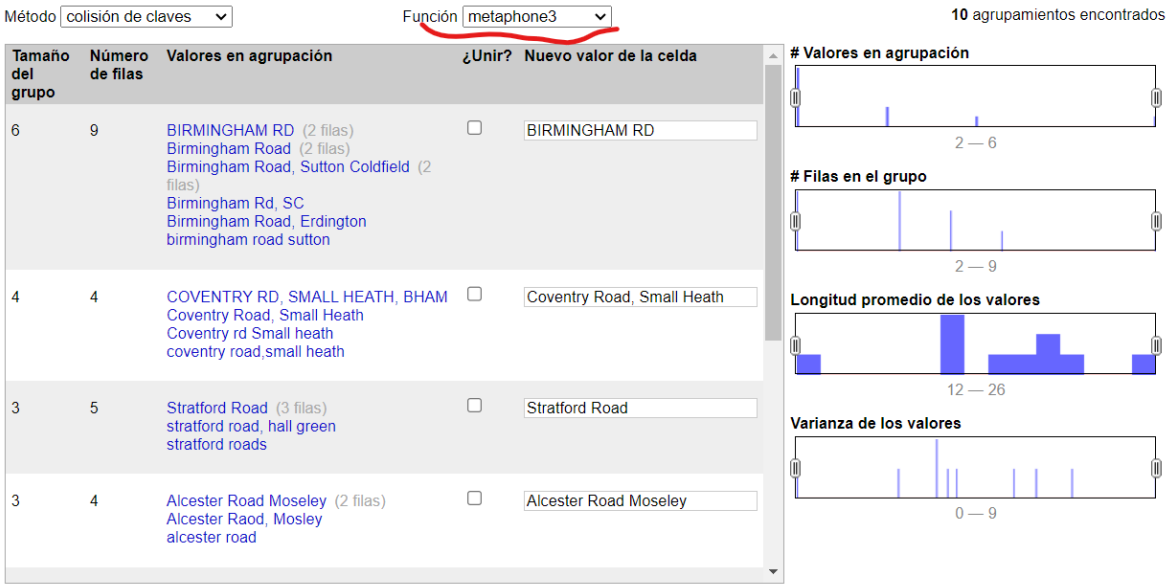
Puedo seleccionar uno por uno de acuerdo con lo que considero compatible según reomendaciones o puedo seleccionar todos.

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	☑	Kings Heath High Street
3	3	Stratford Road Stratford road stratford road	☑	Stratford Road
2	3	Stoney Lane (2 filas) stoney lane	☑	Stoney Lane
2	4	Washwood Heath Road (3 filas) Washwood Heath road	☑	Washwood Heath Road
2	2	poplar road , solihull poplar road, solihull	☑	poplar road , solihull
2	2	Bordesley Green Bordesley Green , Bordesley Green.	☑	Bordesley Green

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



Hay que recordar que el método colisión de claves , **tiene otros algoritmos**, si ahora seleccionamos **methaphone 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**.

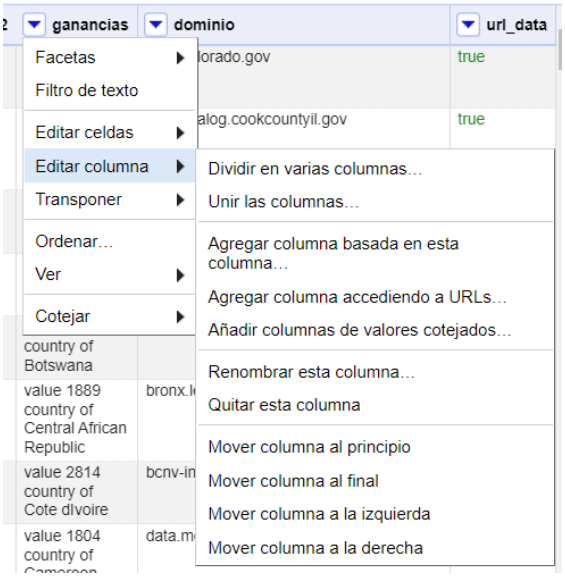


Puedo hacer el mismo proceso con todas las funciones del listado hasta que no aparezcan más datos que se puedan agrupar. Una vez terminemos le damos click al botón **CERRAR.**

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

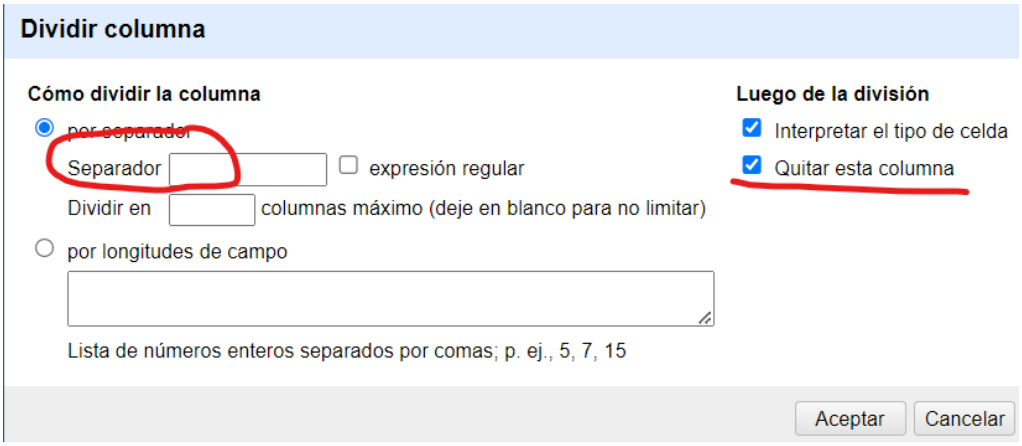
La siguiente practica está orientada en poder obtener el valor numérico que está integrado en esta cadena de texto **value 56 country of Angola** perteneciente a la columna **ganancias**.

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



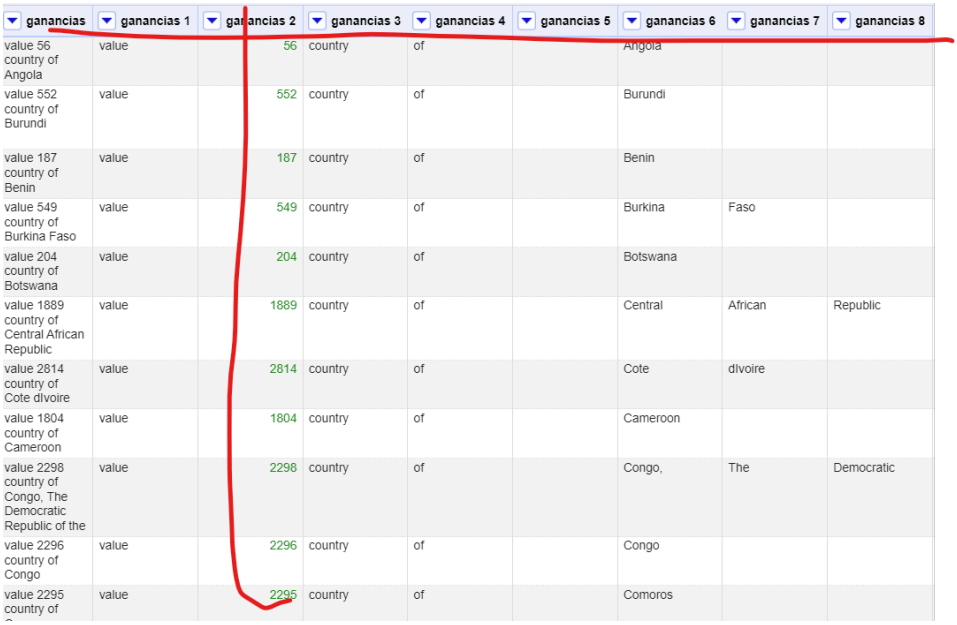
Aparecerá la siguiente ventana donde en el valor del separador debemos colocar solo un espacio en blanco, debido a que la cadena de texto tiene palabras separadas por espacios en blanco.

Desactivar la opción **QUITAR esta columna**.



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

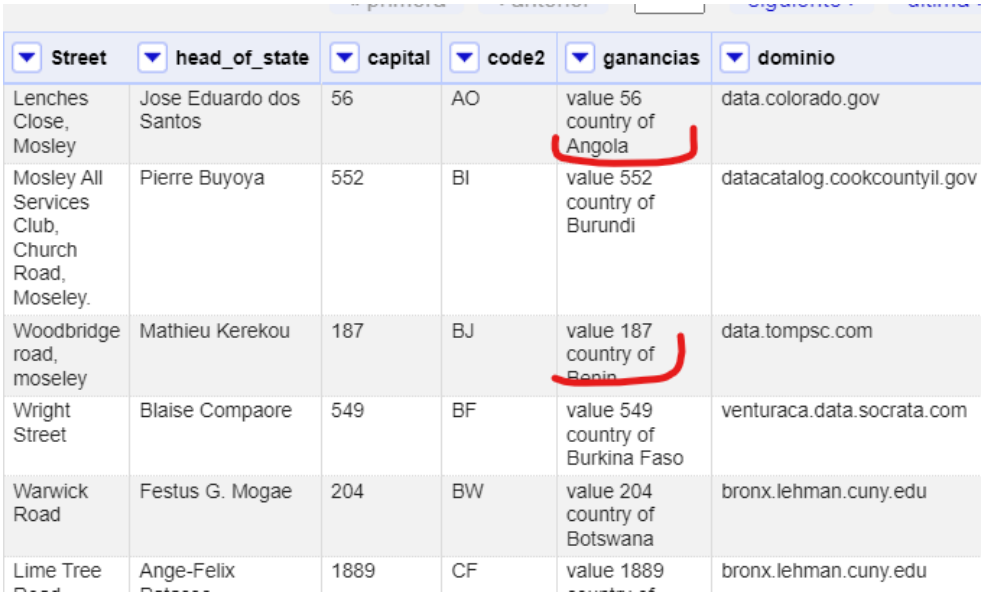
La columna que nos interesa es ganancias 2 que contiene los valores numéricos que contenía la cadena.



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 dIvoire	value	2814	country	of		Cote	dIvoire	
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
value 2296 country of Congo	value	2296	country	of		Congo		
value 2295 country of	value	2295	country	of		Comoros		

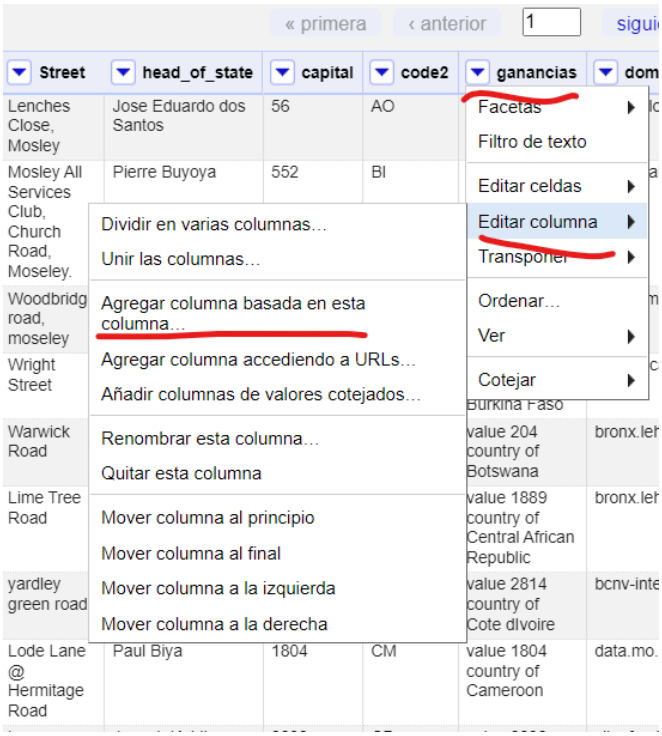
Extraer datos numéricos de un campo de texto (opción 2)

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	Ange-Felix	1889	CF	value 1889	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.



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.

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

Esto le indica a openRefine, selecciónname 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.

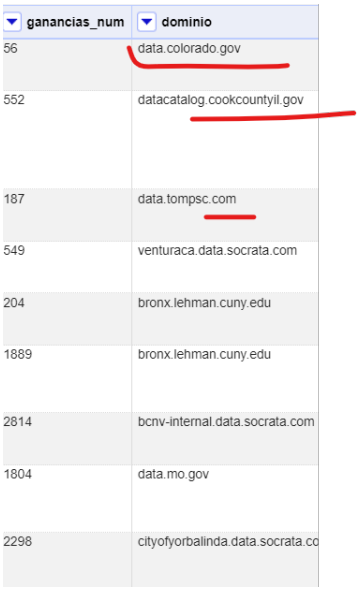
En los valores vemos los números que se han extraído por cada fila.
Le damos al botón **ACEPTAR**.

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

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 dIvoire	2814
CM	value 1804 country of Cameroon	1804
CD	value 2298 country of	2298

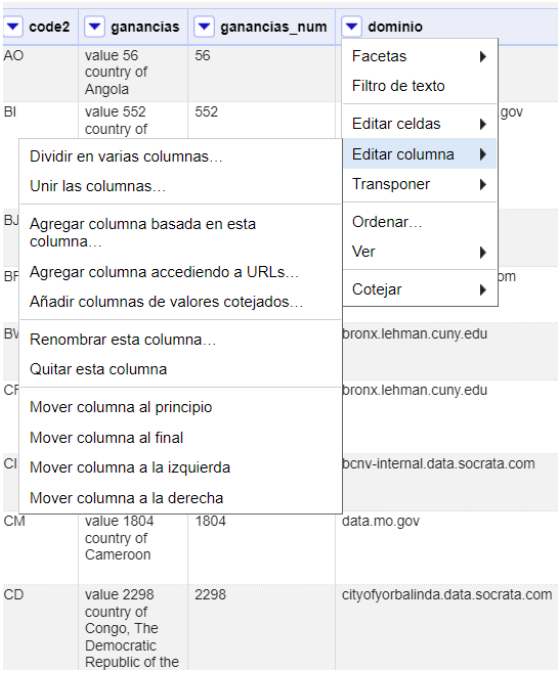
Extraer un texto de una cadena de texto

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.co

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.

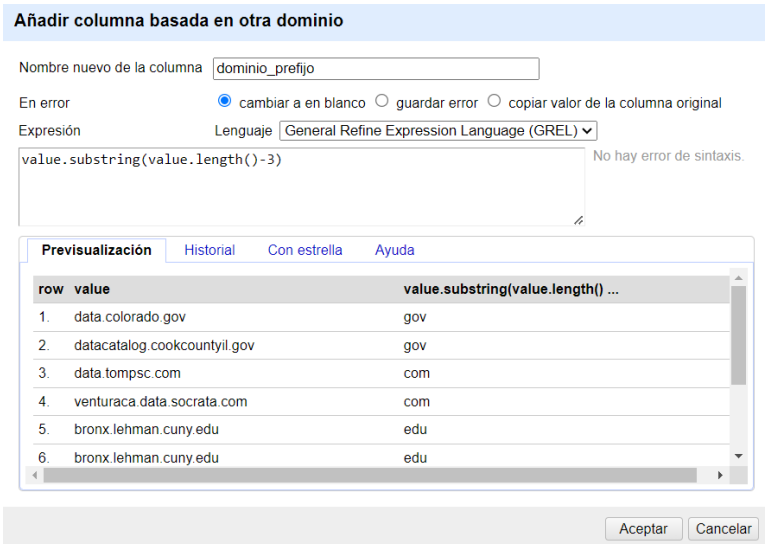


Se abrirá la ventada de nueva columna y colocaremos por nombre **dominio_prefijo**, en Expresión utilizaremos otra expresión regular:

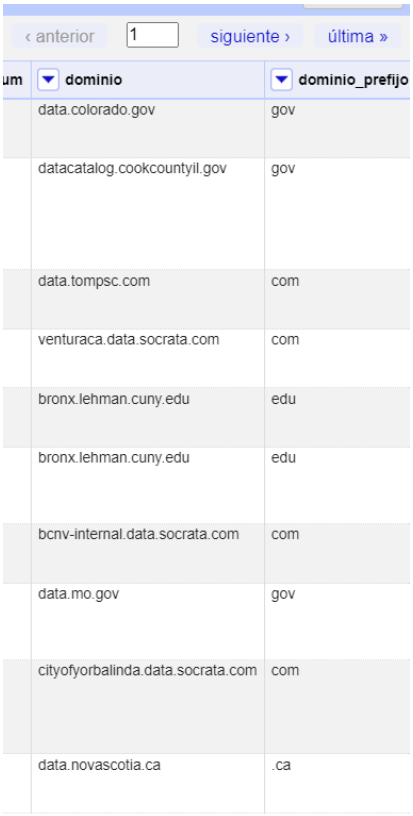
value.substring(value.length()-3)

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.



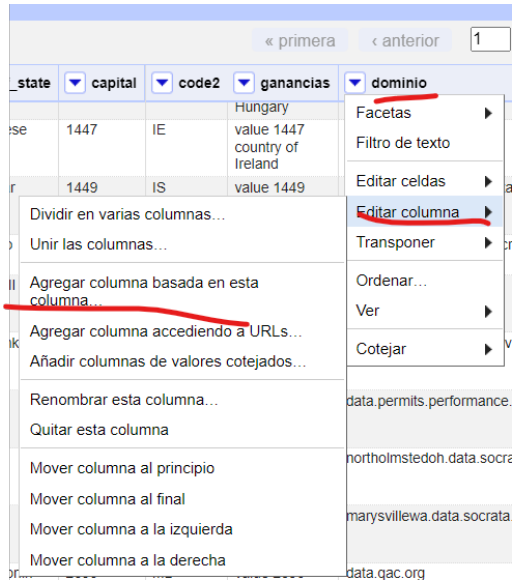
Se creará la columna con los prefijos de las URL



Identificar si un texto está dentro de una cadena de texto

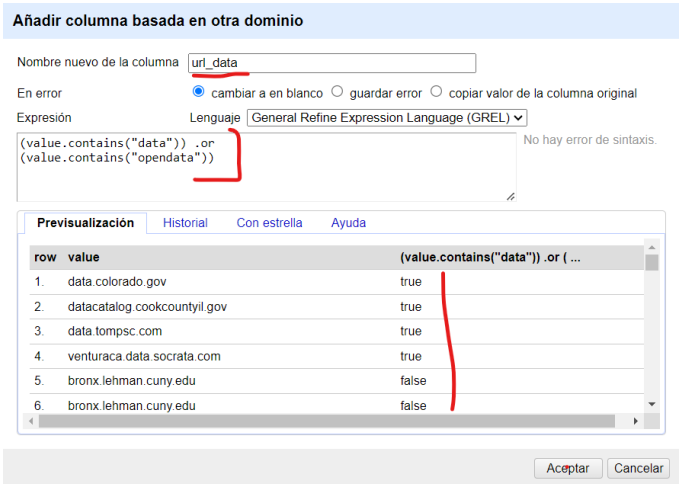
En está practica vamos a crear una nueva variable que identifique si los datos de la variable dominio, proviene de URL con la palabra **data** o **opendata.**

Creamos 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 .contain().
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.



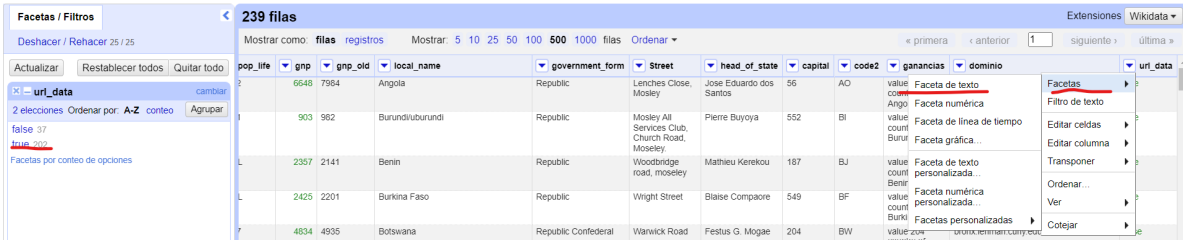
El resultado es una nueva variable **url_data** con los datos **true o false**.

‹ anterior 1 siguiente › última »

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?

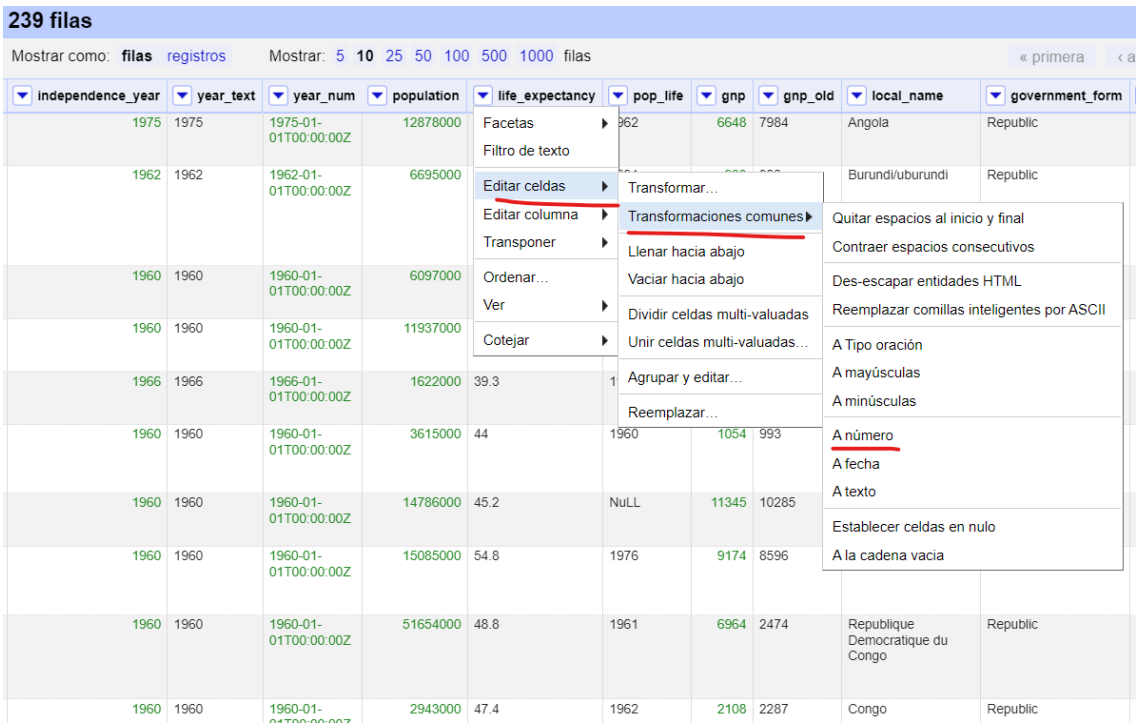
Si generamos una faceta de texto de la variable url_data, podemos obtener como información que de los 239 países de la tabla, **202 almacenan** sus datos en fuentes de datos abiertas con una url estándar como la palabra **data** o **opendata**.



Datos numéricos a categórico, condicional IF

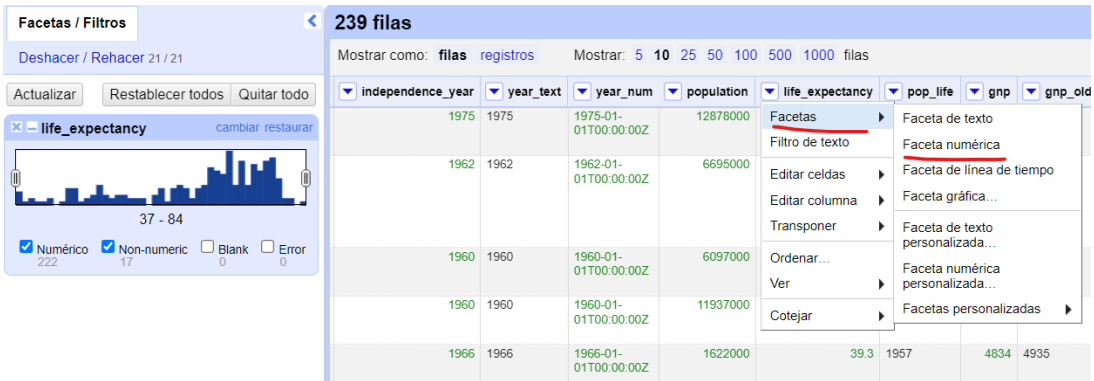
En está practica vamos a crear una nueva variable con valores generados de una condición específica de los datos.

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

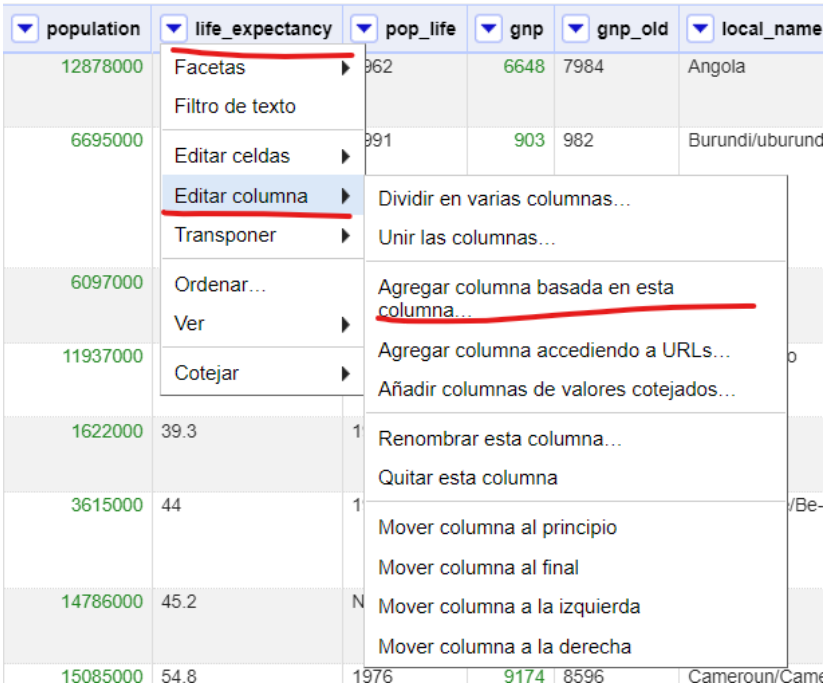


Luego utilizamos las Facetas numéricas en la variable **life_expectancy** , FACETAS > facetas numéricas.

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**

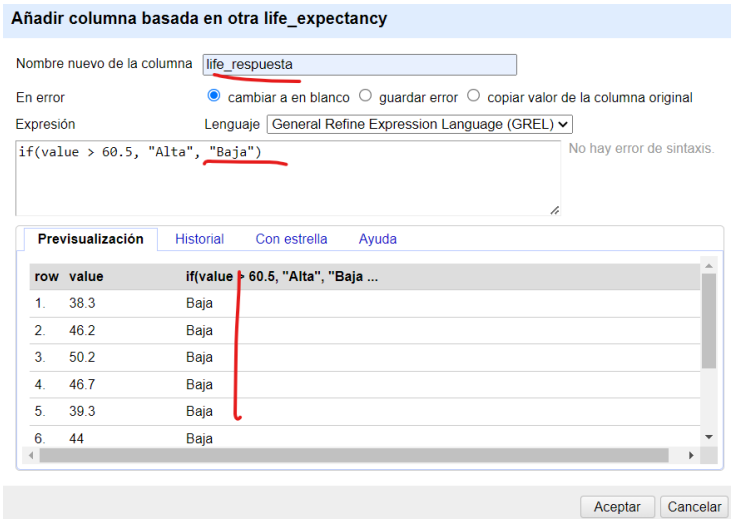


Crearemos una nueva variable que me permite incluir esta condicional.

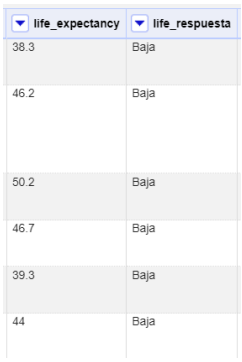


Le colocaremos por nombre a la variable **life_respuesta**.

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**.



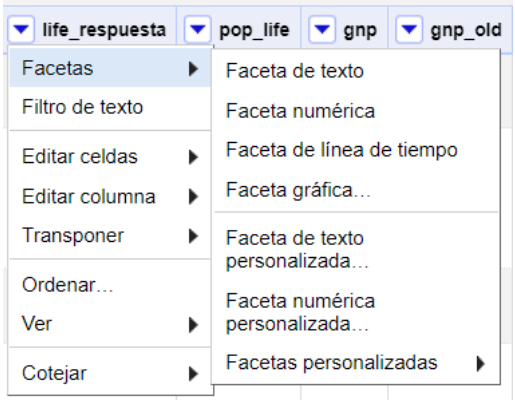
Se muestran los datos de la nueva variable con los valores de **Alta** y **Baja**.



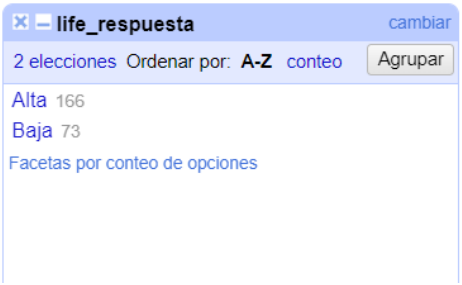
life_expectancy	life_respuesta
38.3	Baja
46.2	Baja
50.2	Baja
46.7	Baja
39.3	Baja
44	Baja

¿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**

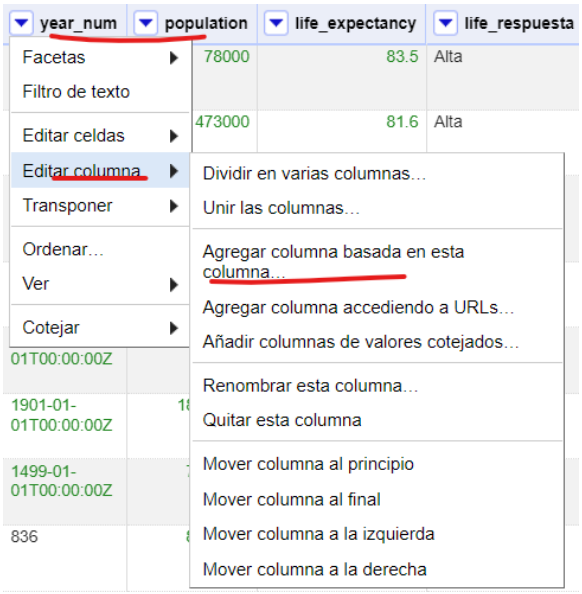


NOTA: Si queremos crear una variable categórica de **life_expectancy** con tres categorías: podemos escribir en EXPRESIÓN al crear la nueva variable el siguiente código.

```
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**.

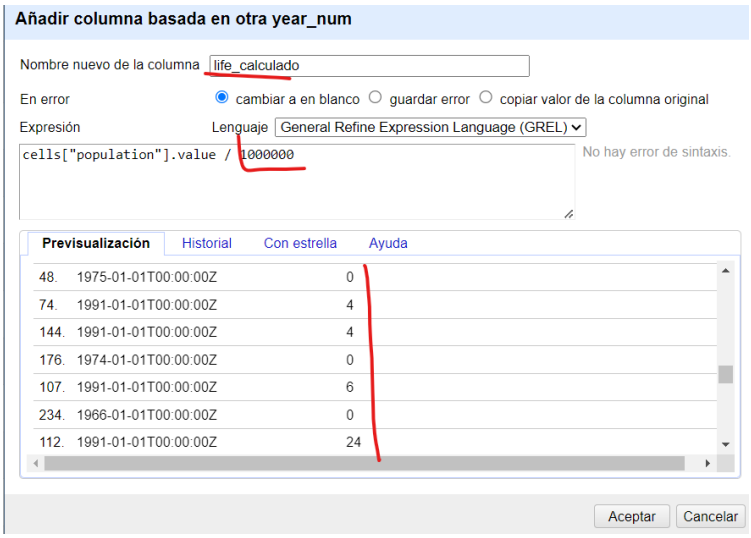


En la ventana colocaremos por nombre a la columna **life_calculado**.

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:

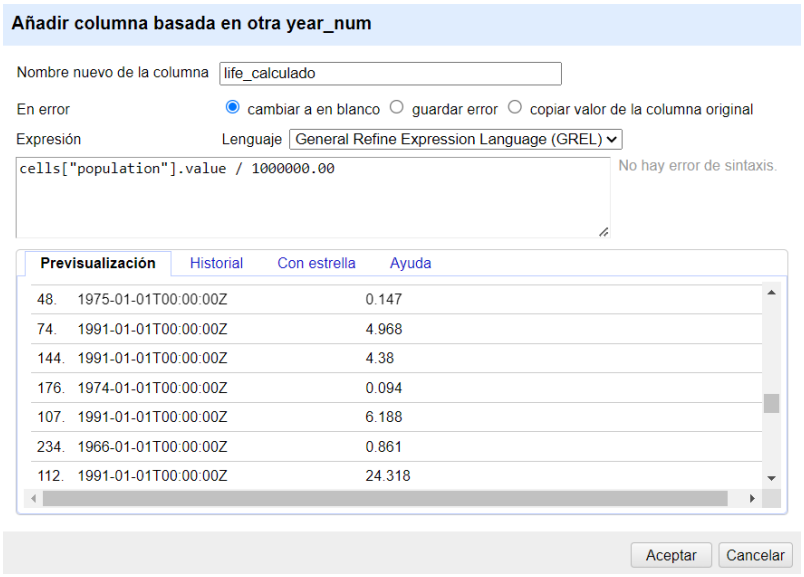
cells["population"].value / 1000000



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

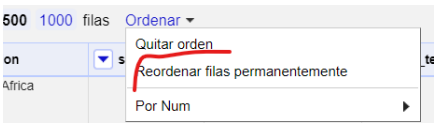
La razón se debe a que openRefine esta dividiendo dos números enteros, por lo que la salida será un número entero. Para que el calculo genere un resultado con número decimal, uno de los dos valores debe ser con número decimal.
El código sería el siguiente:

cells["population"].value / 1000000.00
Con esto tendremos el resultado esperado.



Resultado de la nueva columna **life_calculado**.

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.

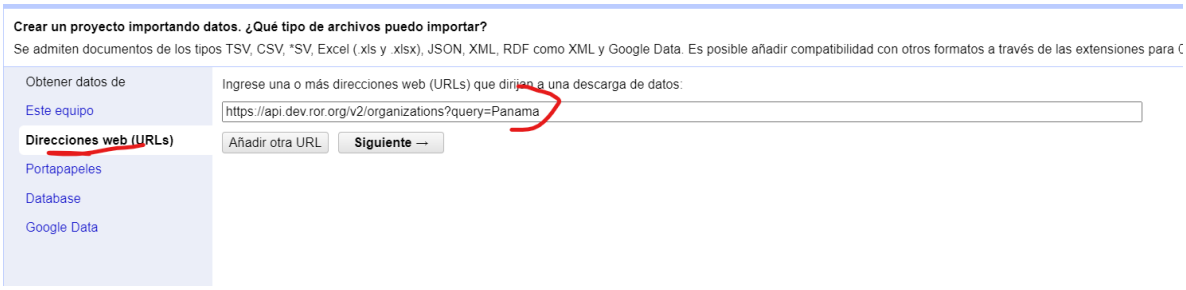


Cargar datos de una URL

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



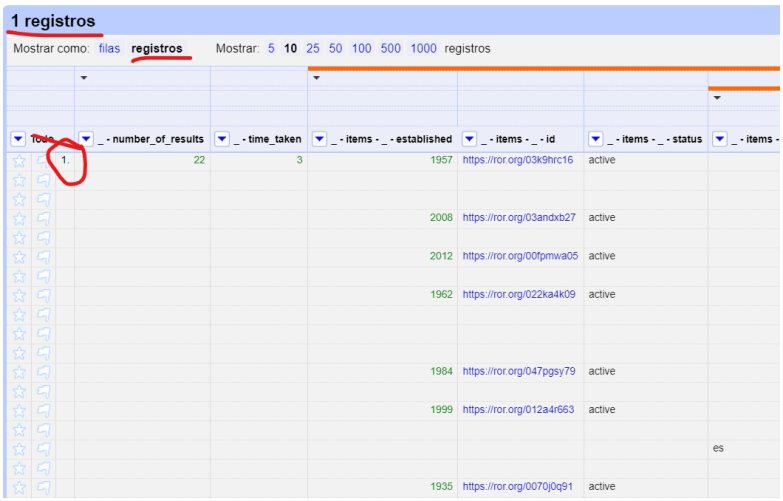
Le damos click al botón siguiente.

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

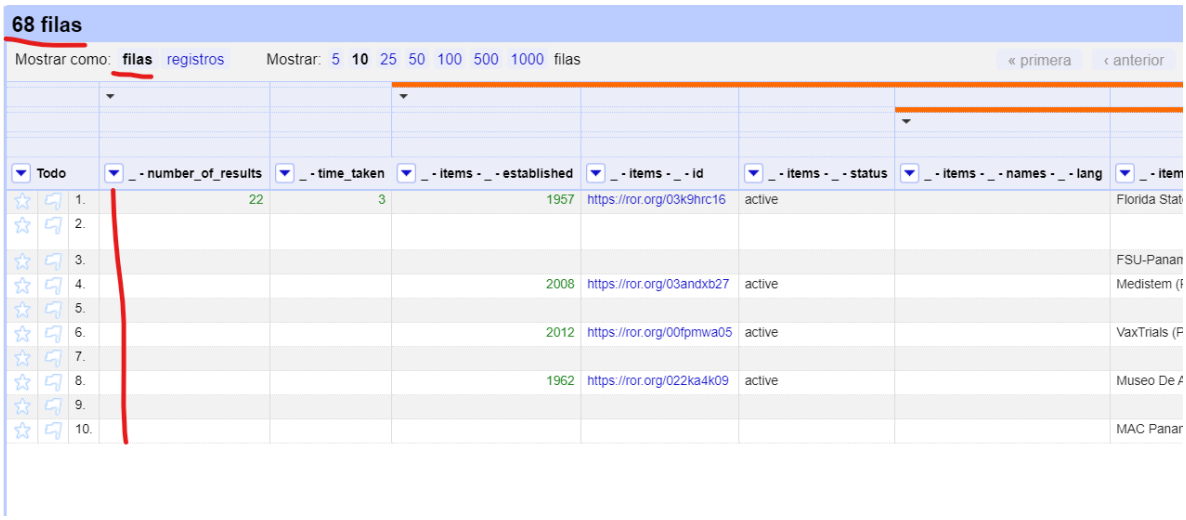
Habilite la opción **CARGAR AL MENOS.**

Le damos al botón CREAR PROYECTO.

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

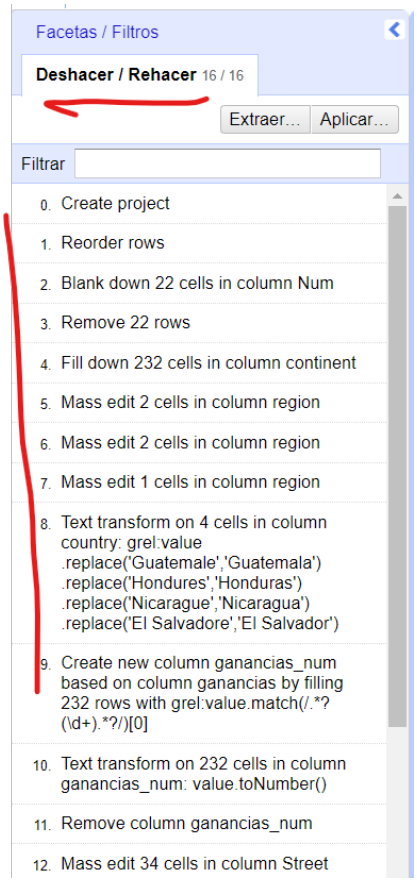


Debemos seleccionar la opción de **filas** para que se muestran todas las filas identificadas, ahora puede eliminar las filas en blanco y las columnas vacías.

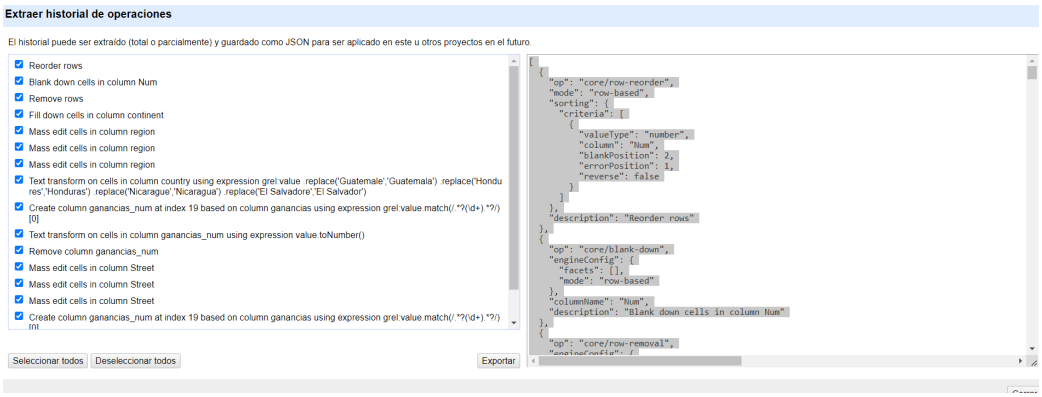


Guardar líneas de comandos en Open refine

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.

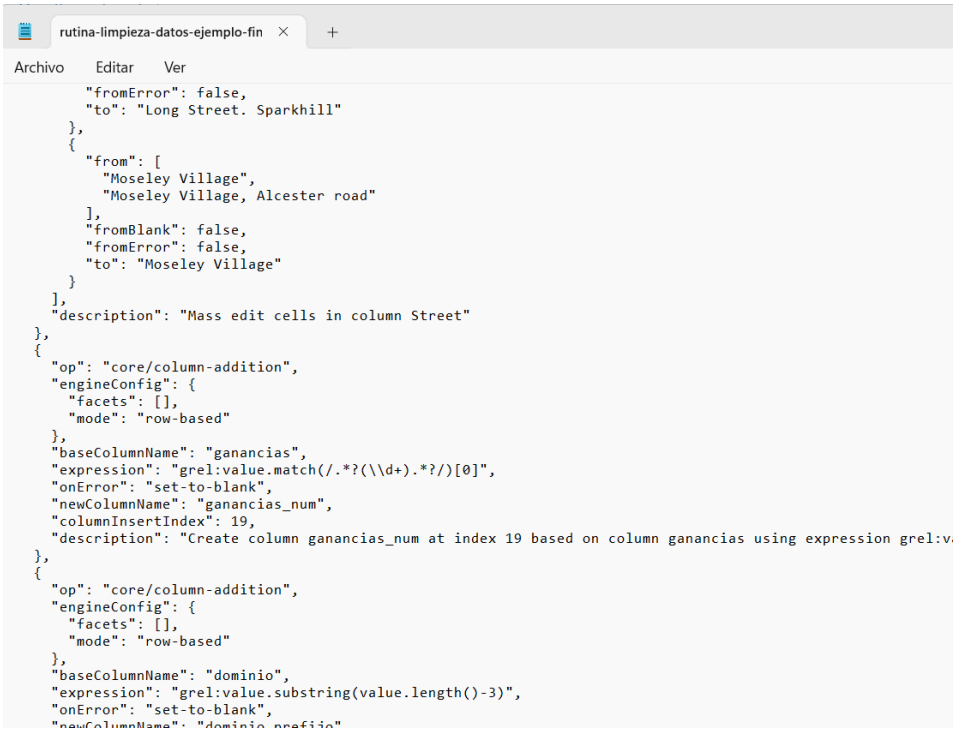


Se abrirá una nueva ventana que mostrará las líneas de comando y el código de los comandos en formato JSON a la derecha.



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

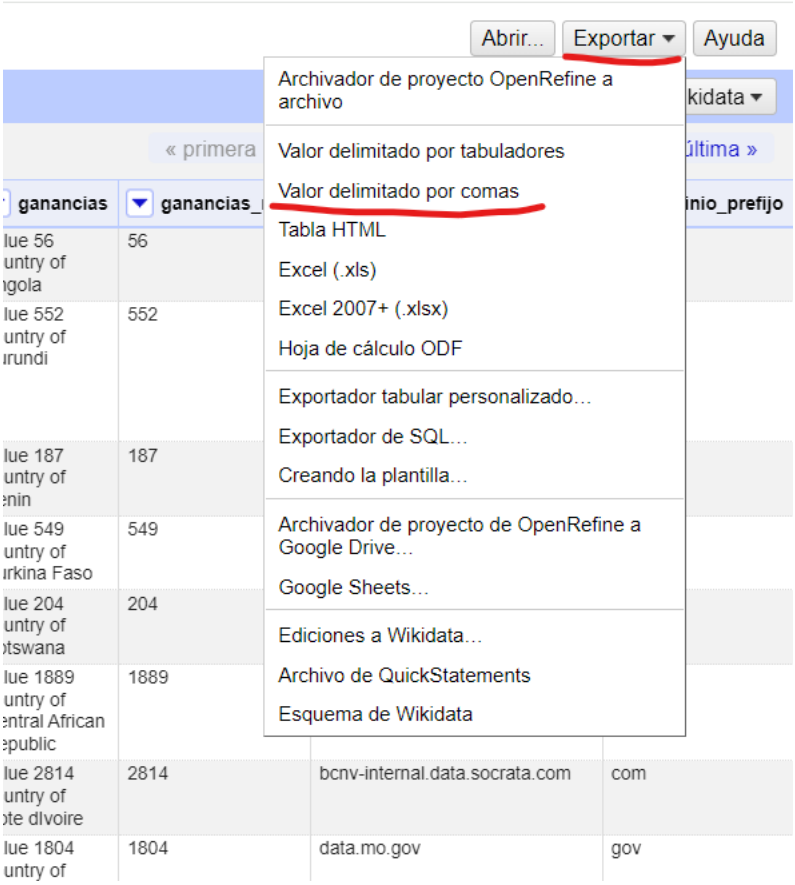
Este código puede ser reutilizado, si el proceso de limpieza con el mismo documento tiene que ser repetido constantemente.



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.



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-aggerate-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