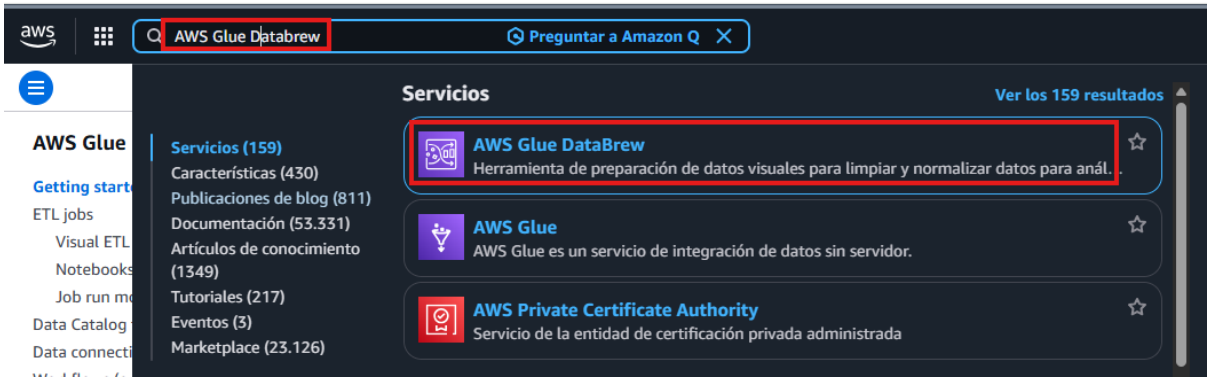
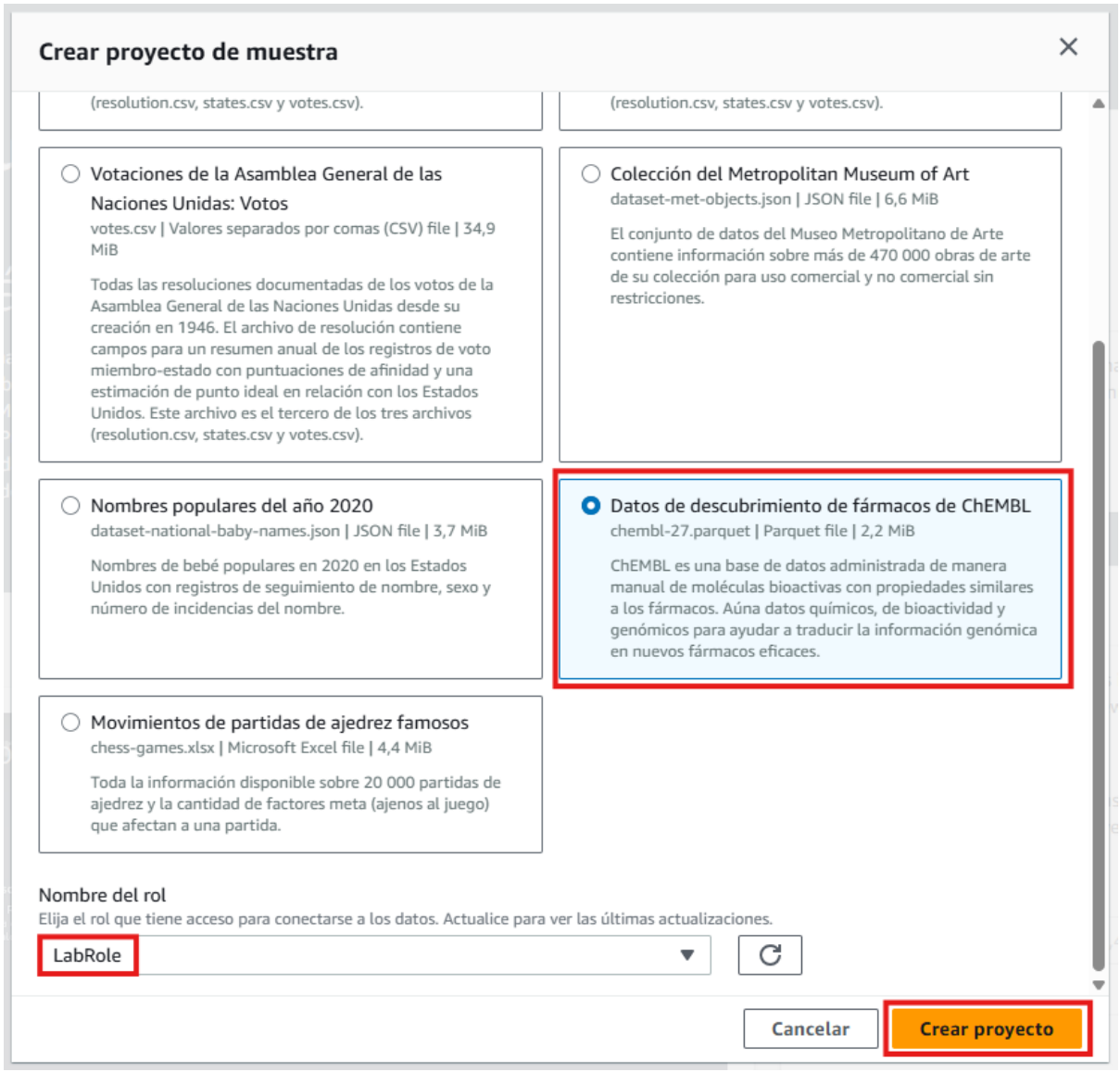
PR_08.3 Dani Gayol Rodríguez

Apartado A

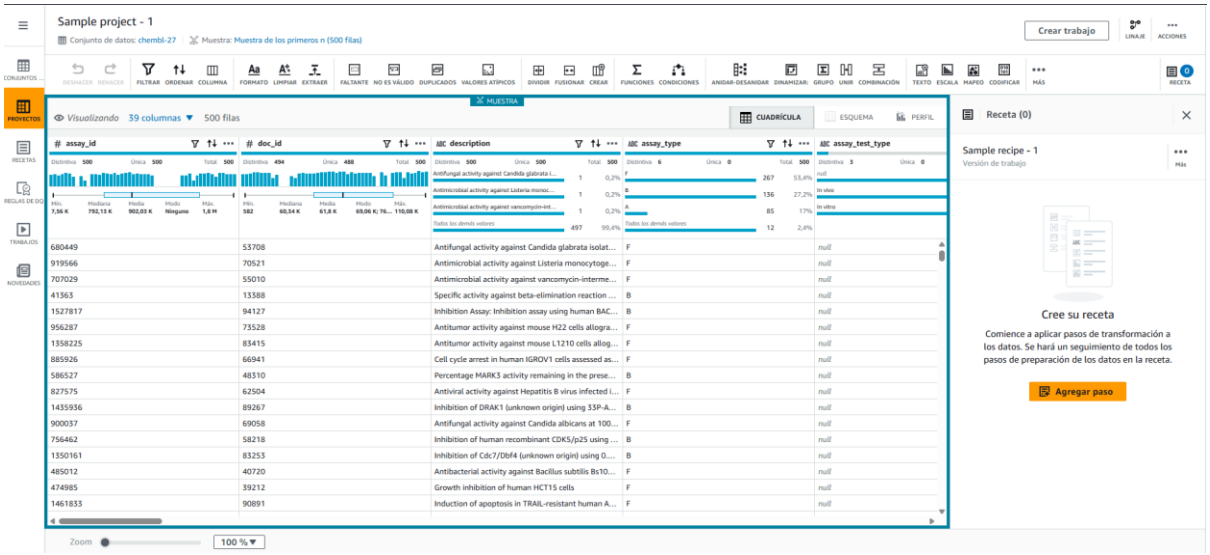
1.) Realiza el ejemplo de Proyecto Databrew que se muestra en el tema UT_08.4_AWS DataBrew entre las páginas 1 y 7.

El primer paso va a ser crear el proyecto de muestra con los datos de CHEMBL

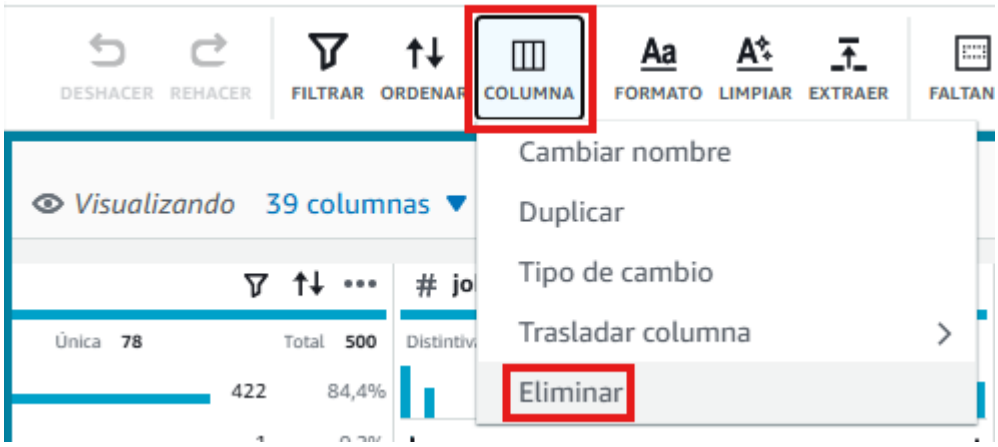


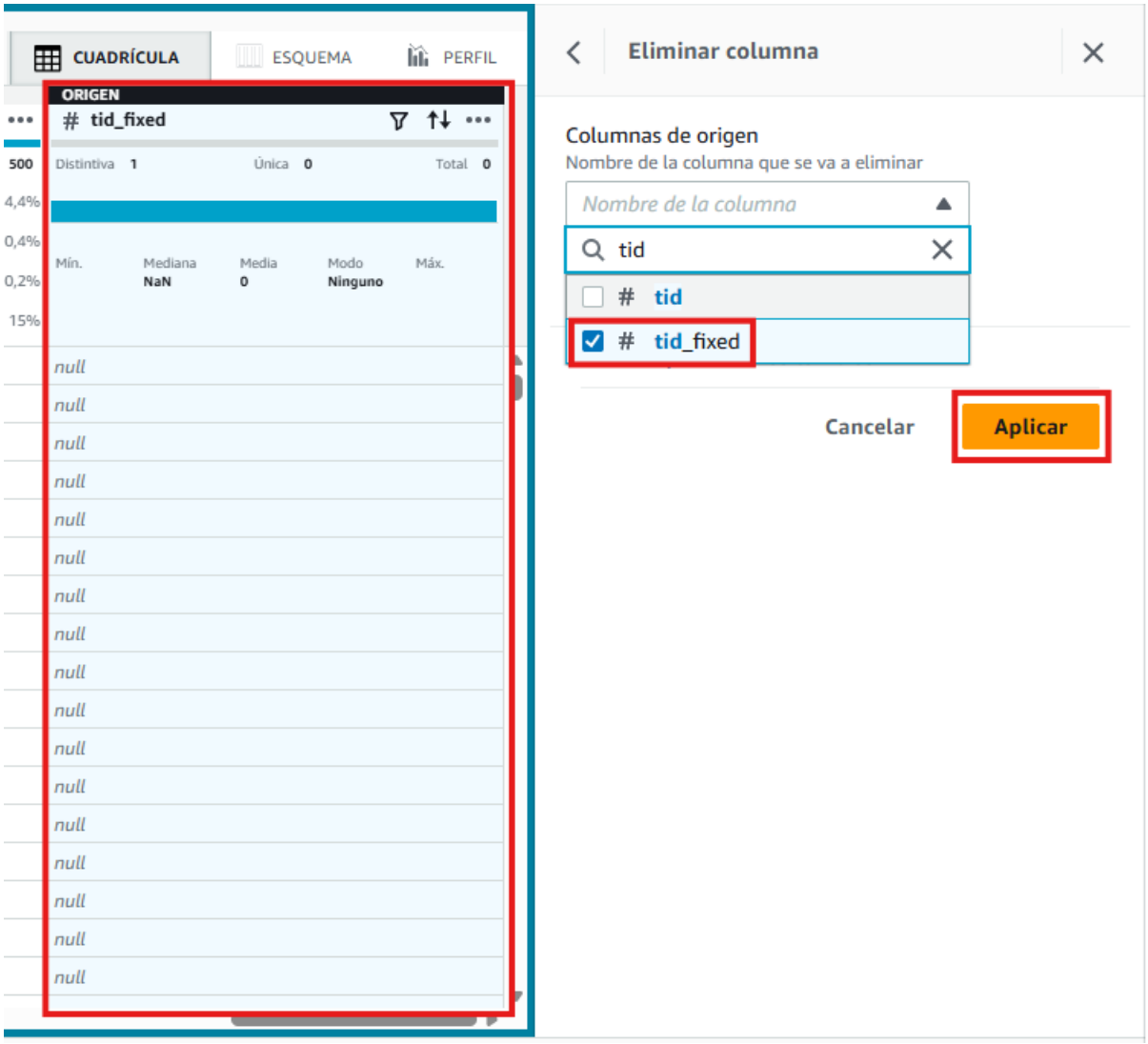


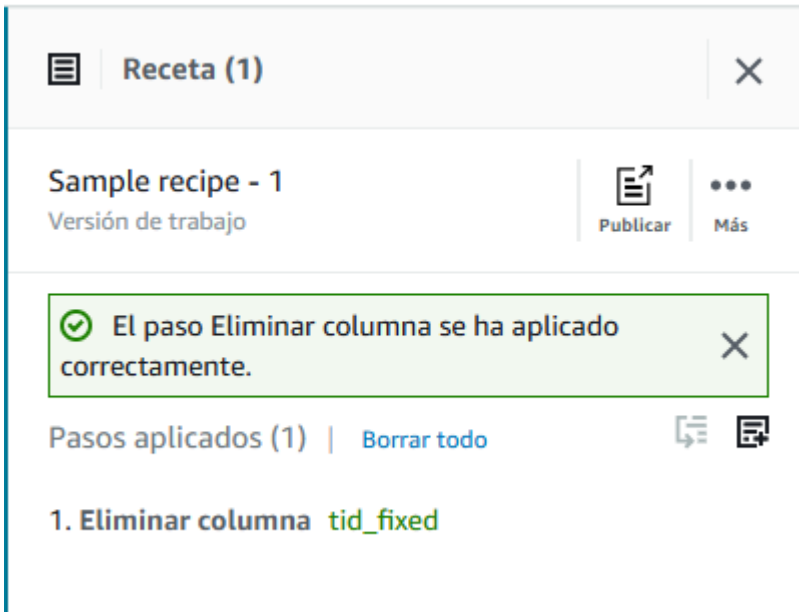
Una vez creado el proyecto, nos manda a interfaz del Databrew:



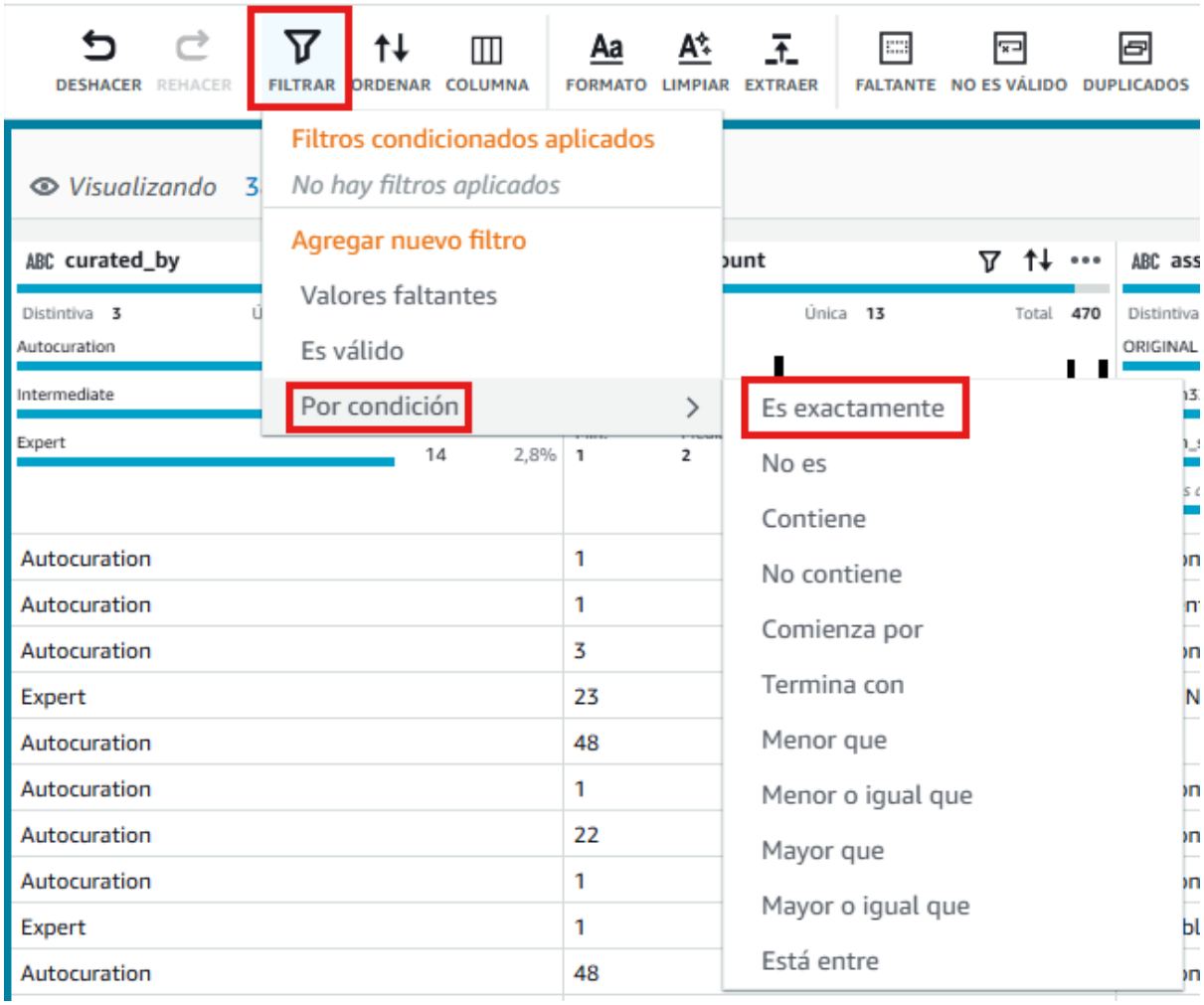
Ahora vamos a seguir con los pasos que nos indica la práctica, el primer paso es eliminar la columna “tid_fixed”

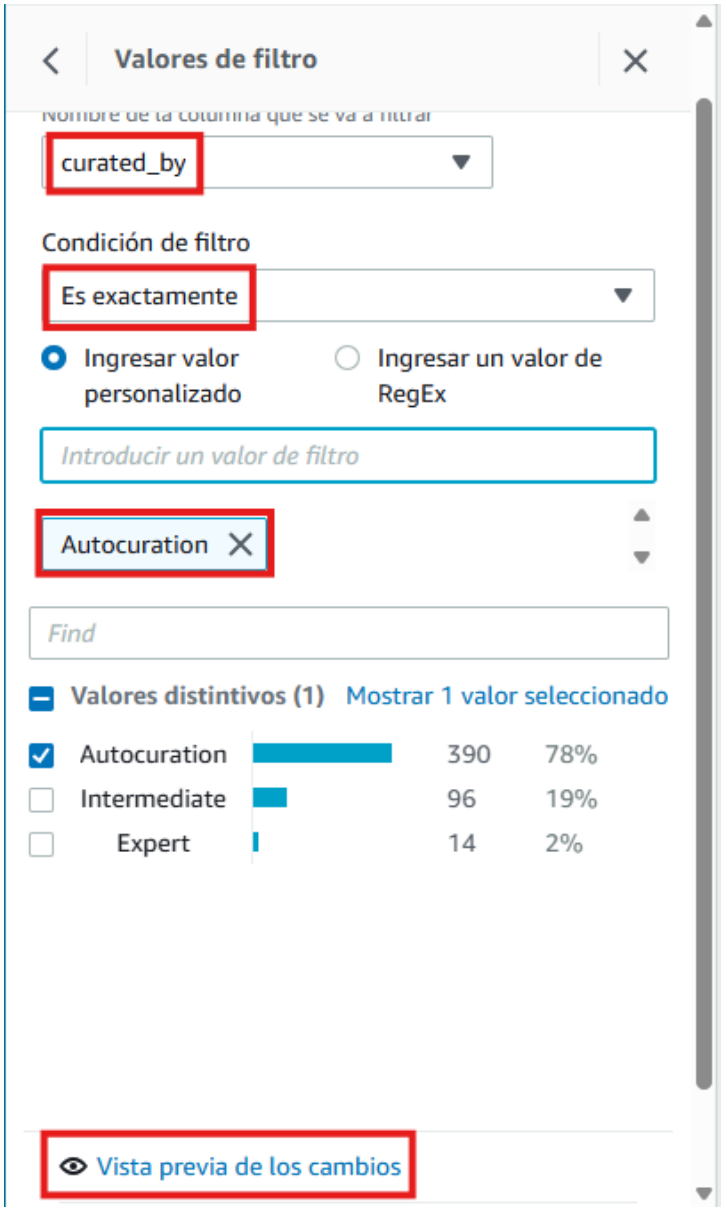


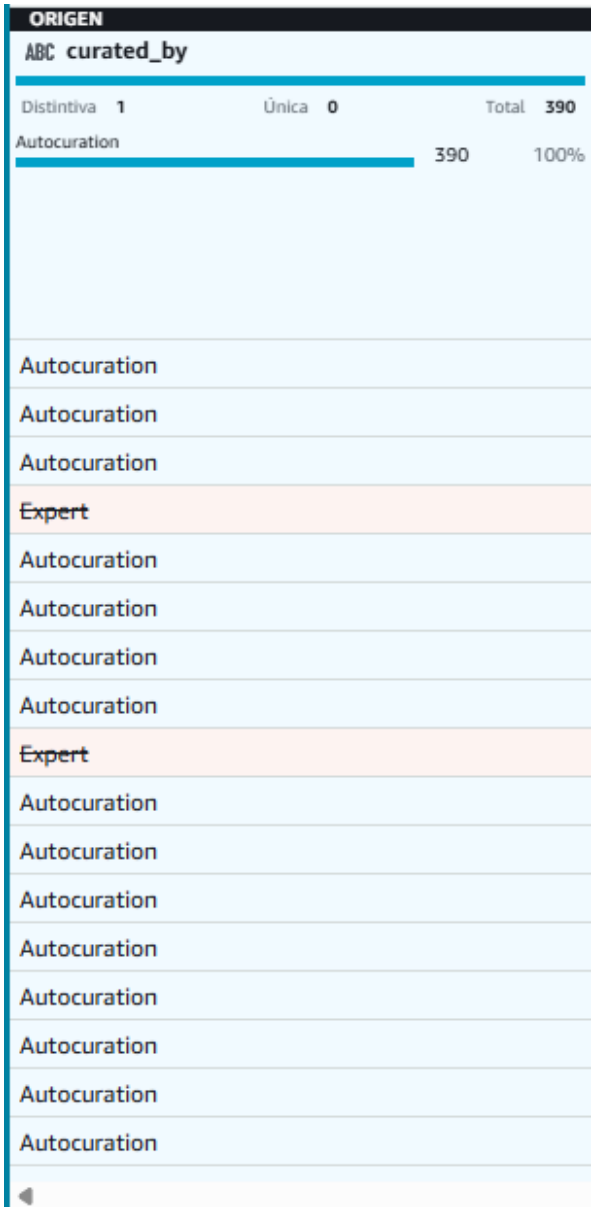




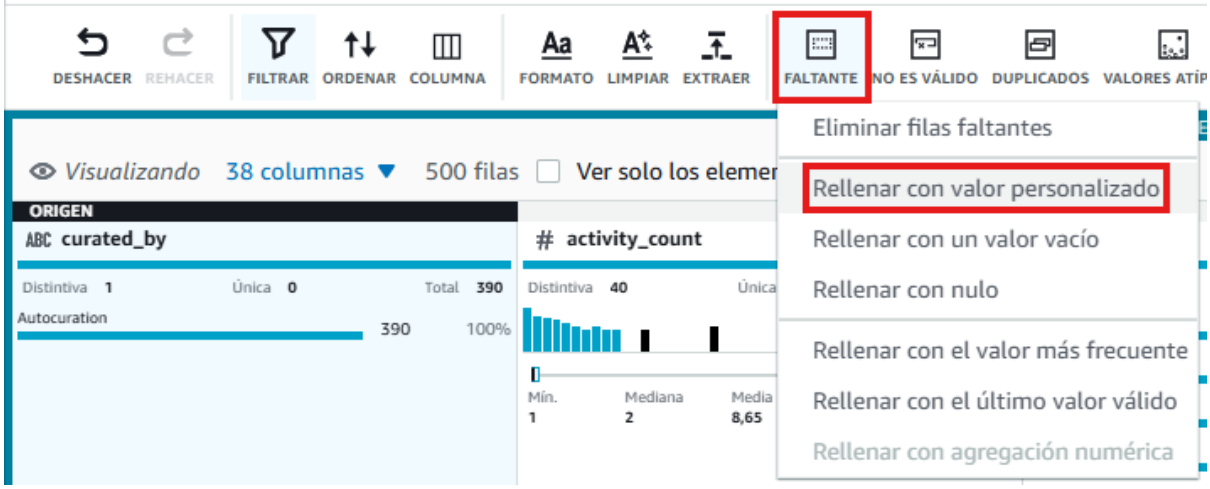
Ahora, vamos a filtrar los datos:

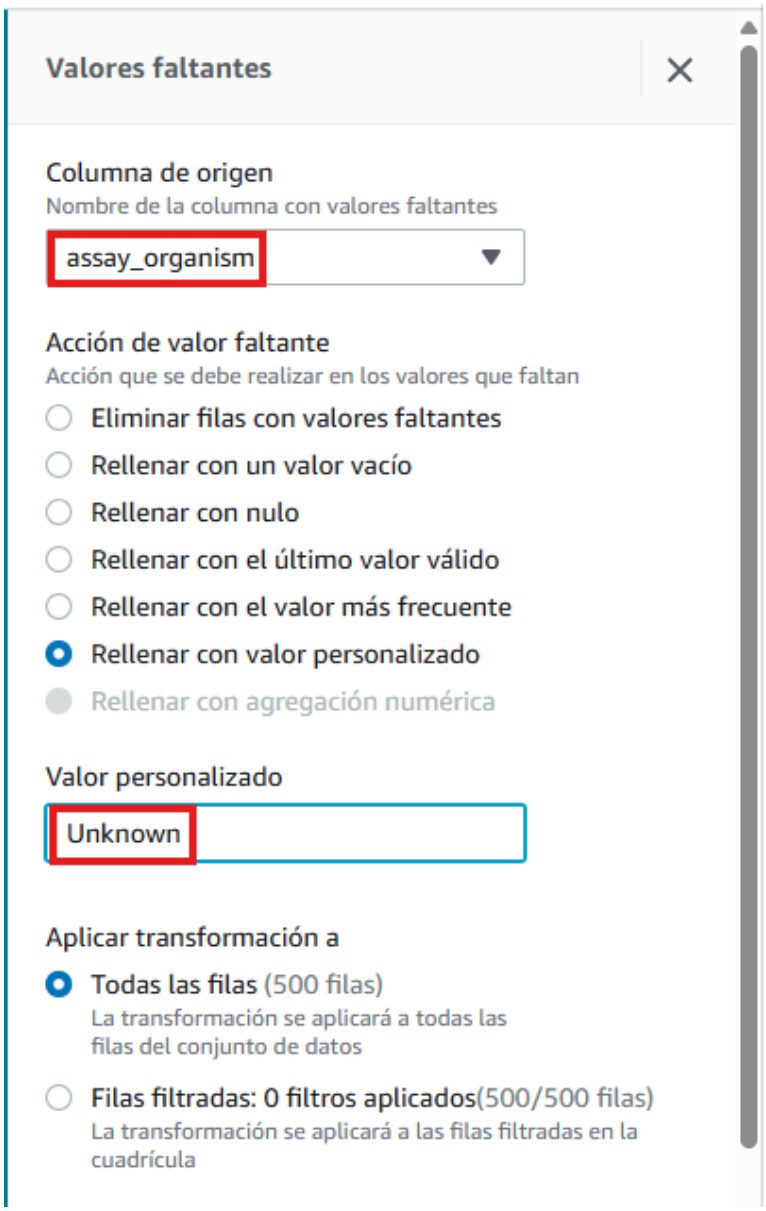


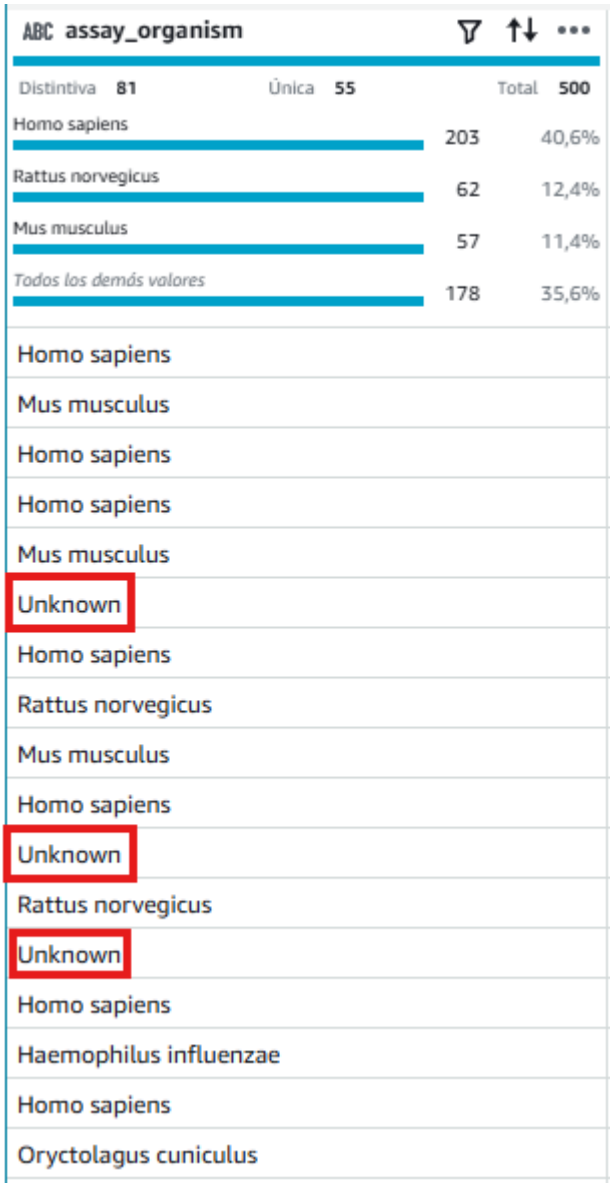

Valores de filtro
Nombre de la columna que se va a filtrar
curated_by
Condición de filtro
Es exactamente
Ingresar valor personalizado
Ingresar un valor de RegEx
Introducir un valor de filtro
Autocuration
Find
Valores distintivos (1)
Mostrar 1 valor seleccionado
Autocuration 390 78%
Intermediate 96 19%
Expert 14 2%
Vista previa de los cambios

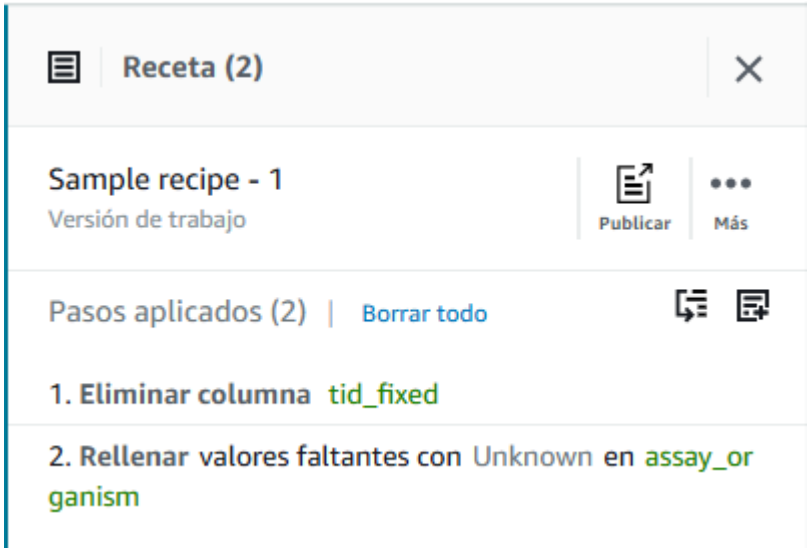


El siguiente paso es rellenar los valores nulos:

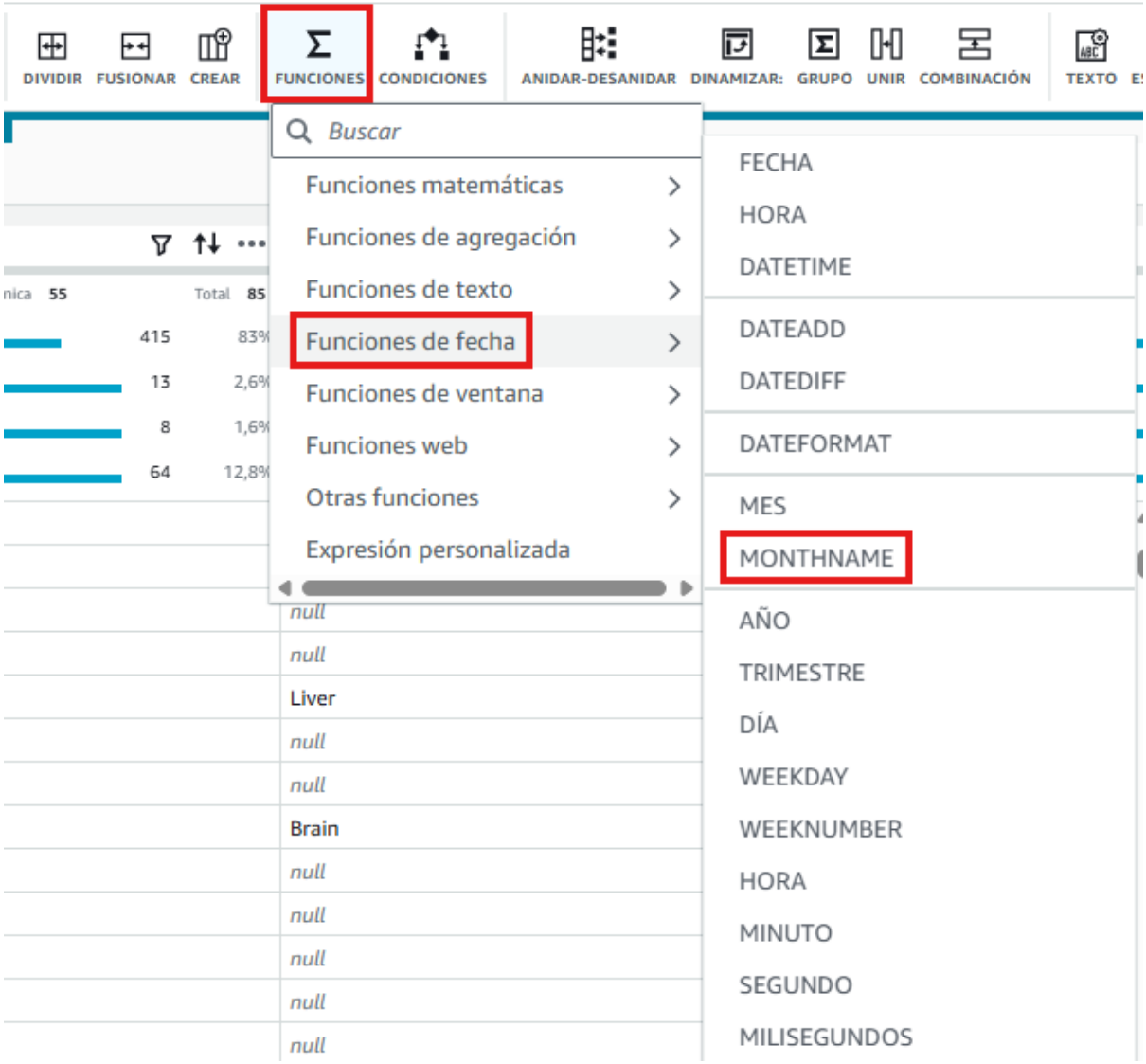


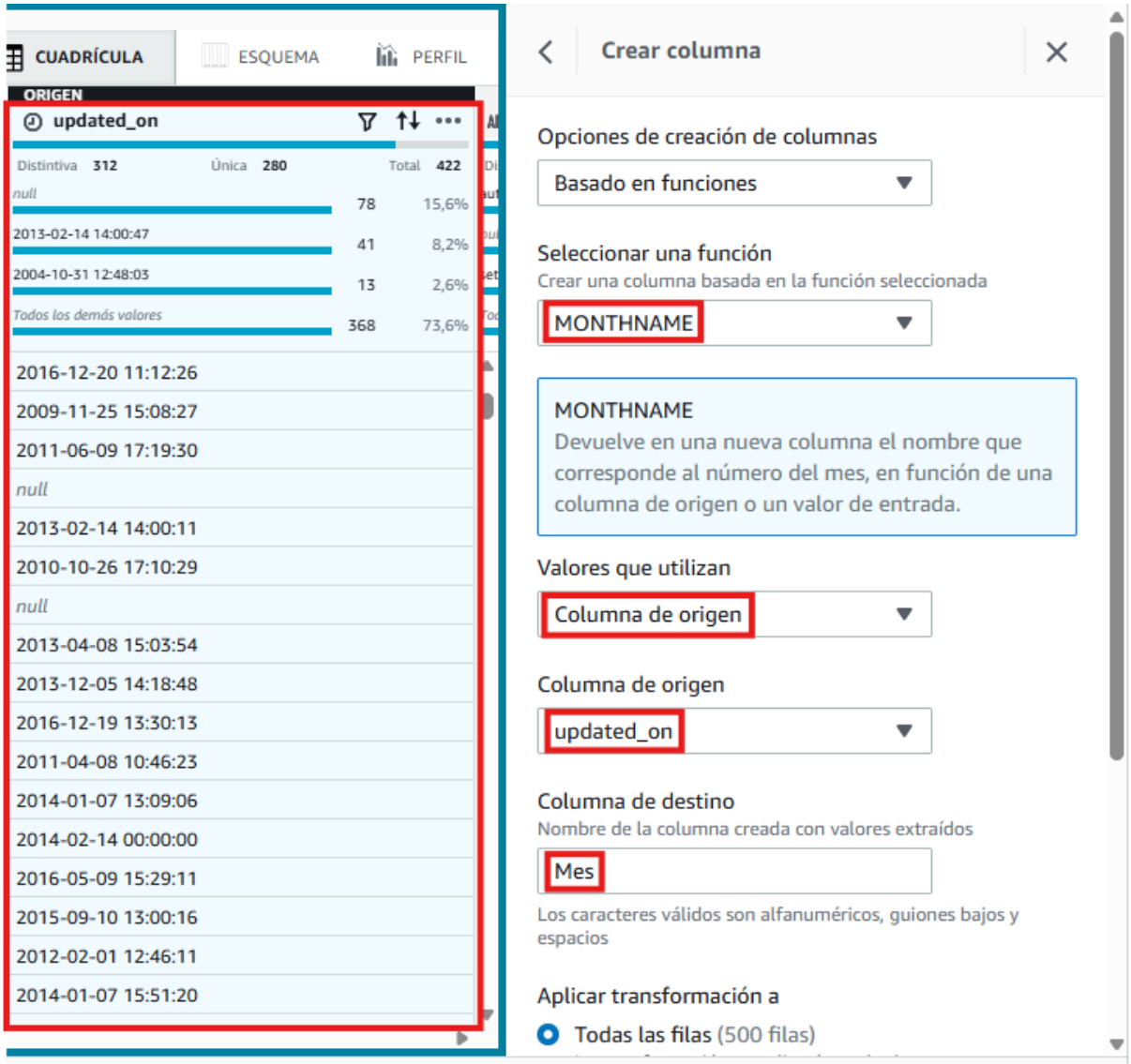

Valores faltantes
Columna de origen
Nombre de la columna con valores faltantes
assay_organism
Acción de valor faltante
Acción que se debe realizar en los valores que faltan
Eliminar filas con valores faltantes
Rellenar con un valor vacío
Rellenar con nulo
Rellenar con el último valor válido
Rellenar con el valor más frecuente
Rellenar con valor personalizado
Rellenar con agregación numérica
Valor personalizado
Unknown
Aplicar transformación a
Todas las filas (500 filas)
La transformación se aplicará a todas las filas del conjunto de datos
Filas filtradas: 0 filtros aplicados(500/500 filas)
La transformación se aplicará a las filas filtradas en la cuadrícula

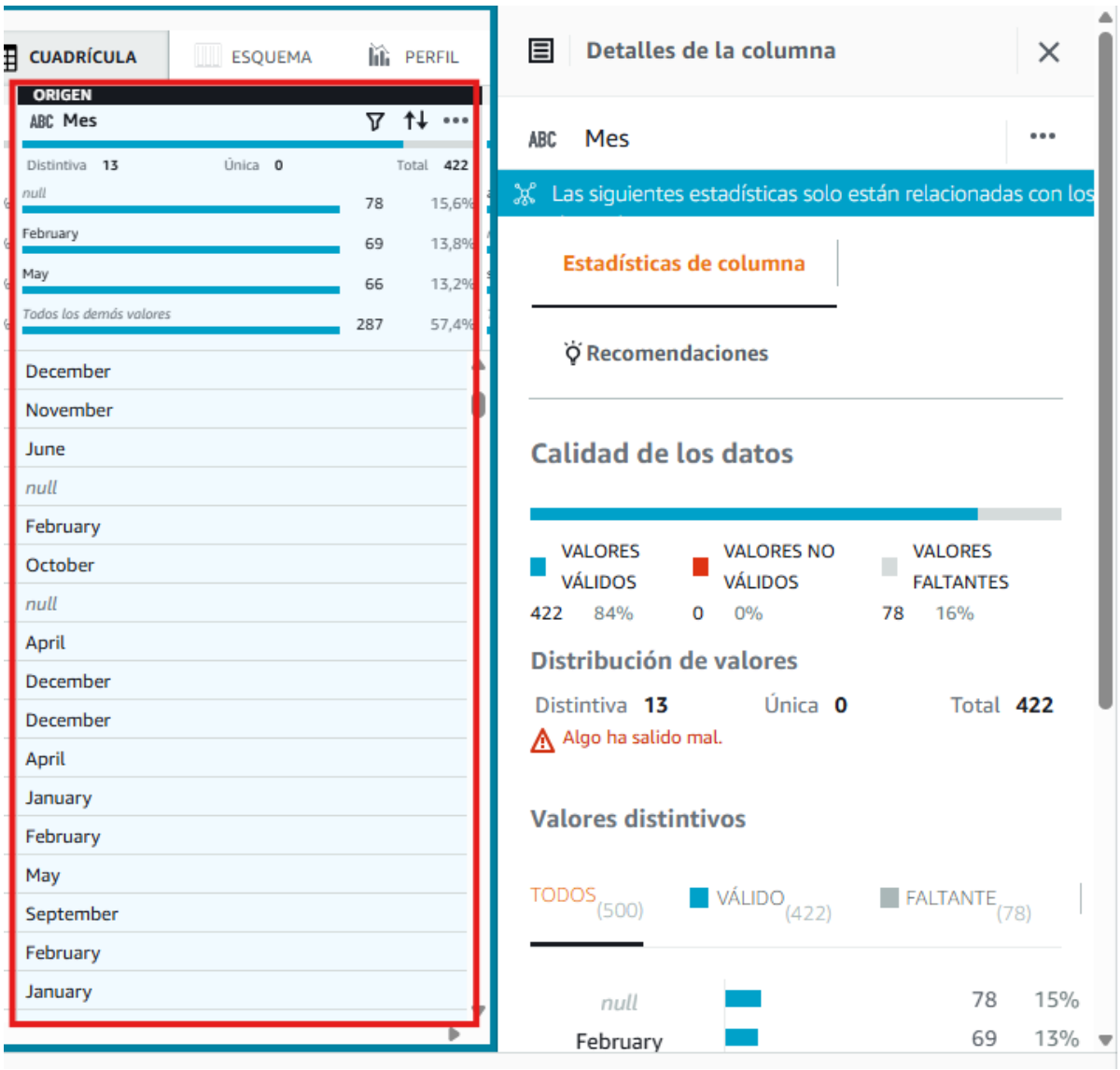




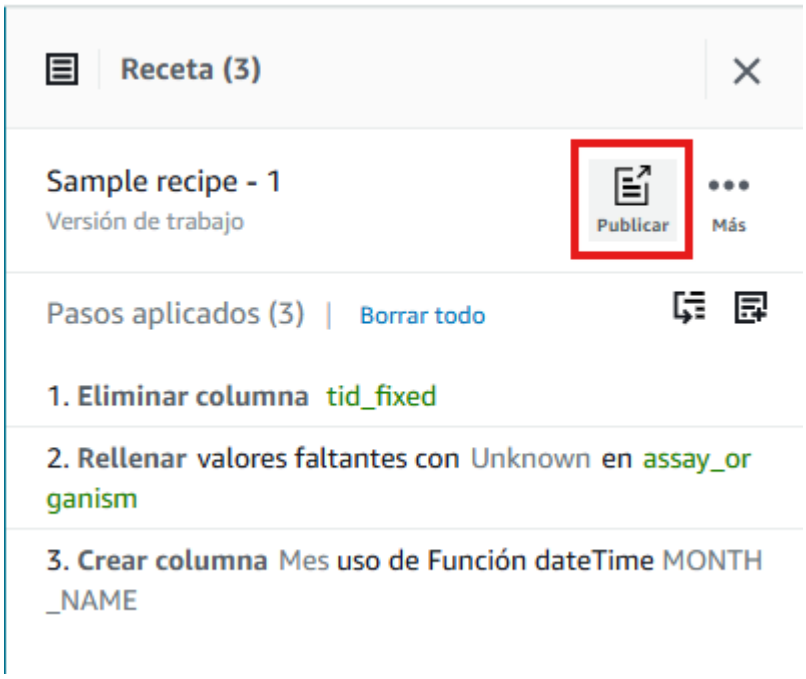
Ahora, vamos a crear una columna nueva:

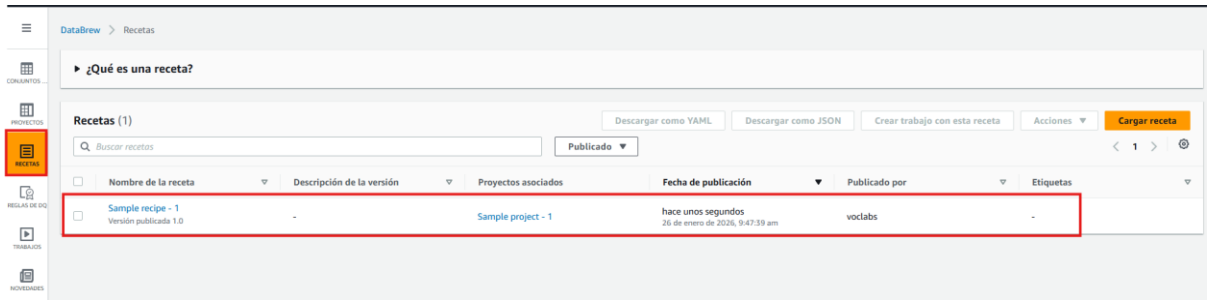

DIVIDIR
FUSIONAR
CREAR
FUNCIONES
CONDICIONES
ANIDAR-DESANIDAR
DINAMIZAR:
GRUPO
UNIR
COMBINACIÓN
TEXTO
Buscar
Funciones matemáticas
Funciones de agregación
Funciones de texto
Funciones de fecha
Funciones de ventana
Funciones web
Otras funciones
Expresión personalizada
FECHA
HORA
DATETIME
DATEADD
DATEDIFF
DATEFORMAT
MES
MONTHNAME
AÑO
TRIMESTRE
DÍA
WEEKDAY
WEEKNUMBER
HORA
MINUTO
SEGUNDO
MILISEGUNDOS
Total 85
415 83%
13 2,6%
8 1,6%
64 12,8%
null
Liver
Brain


CUADRÍCULA
ESQUEMA
PERFIL
ORIGEN
updated_on
Distintiva 312
Única 280
Total 422
null 78 15,6%
2013-02-14 14:00:47 41 8,2%
2004-10-31 12:48:03 13 2,6%
Todos los demás valores 368 73,6%
2016-12-20 11:12:26
2009-11-25 15:08:27
2011-06-09 17:19:30
null
2013-02-14 14:00:11
2010-10-26 17:10:29
null
2013-04-08 15:03:54
2013-12-05 14:18:48
2016-12-19 13:30:13
2011-04-08 10:46:23
2014-01-07 13:09:06
2014-02-14 00:00:00
2016-05-09 15:29:11
2015-09-10 13:00:16
2012-02-01 12:46:11
2014-01-07 15:51:20
Crear columna
Opciones de creación de columnas
Basado en funciones
Seleccionar una función
Crear una columna basada en la función seleccionada
MONTHNAME
MONTHNAME
Devuelve en una nueva columna el nombre que corresponde al número del mes, en función de una columna de origen o un valor de entrada.
Valores que utilizan
Columna de origen
Columna de origen
updated_on
Columna de destino
Nombre de la columna creada con valores extraídos
Mes
Los caracteres válidos son alfanuméricos, guiones bajos y espacios
Aplicar transformación a
Todas las filas (500 filas)

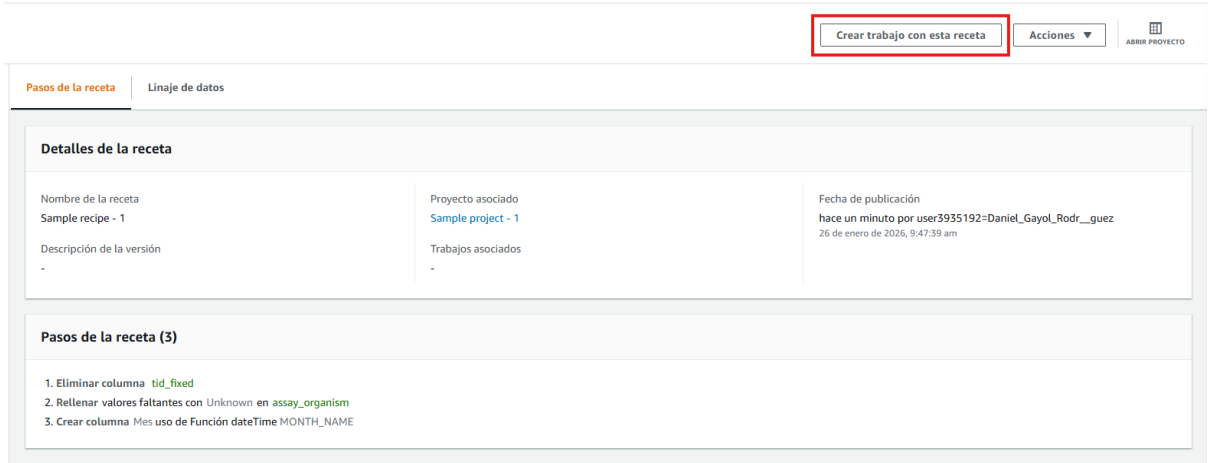


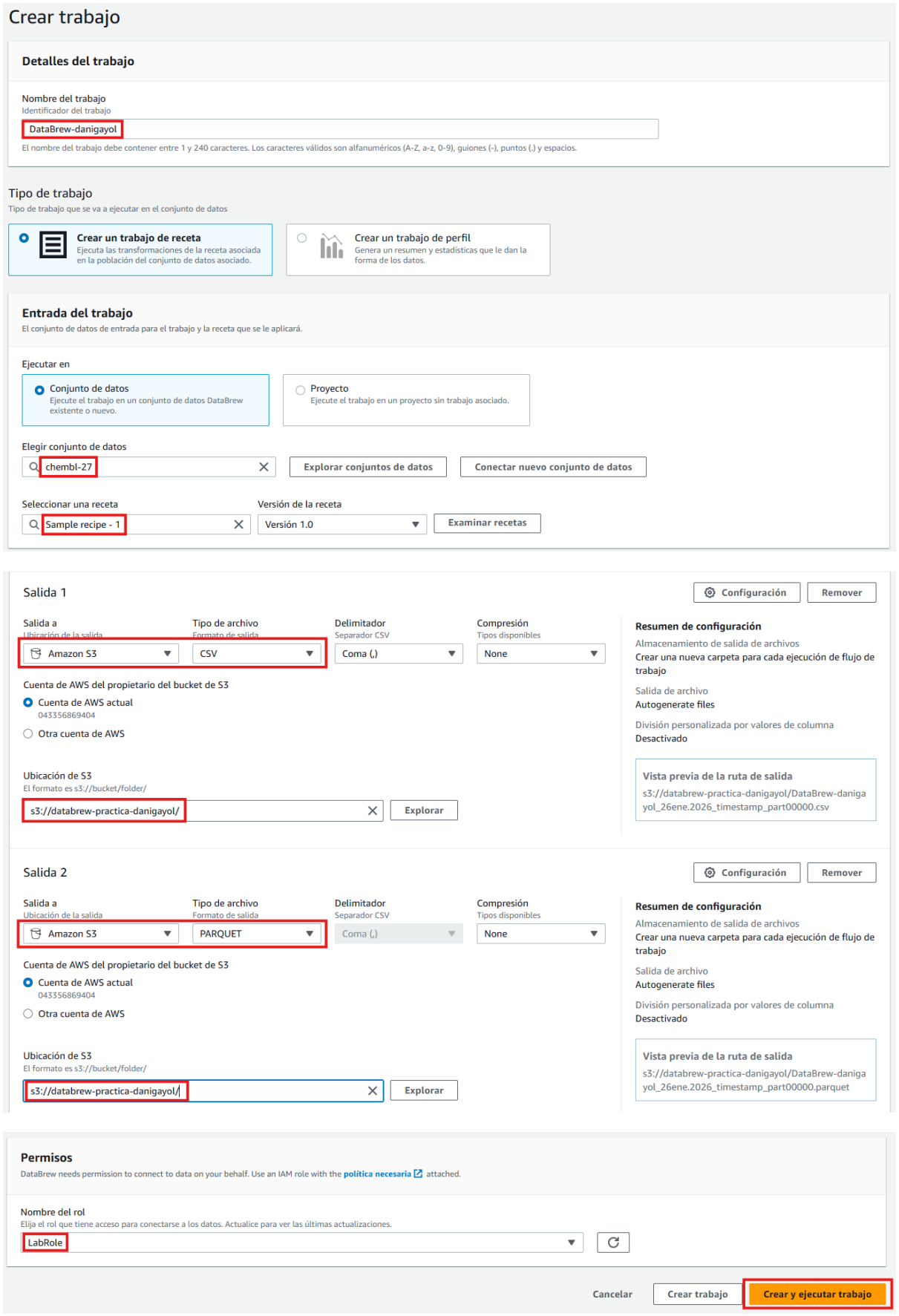
El siguiente paso es publicar la receta, para ello le tenemos que dar al botón de publicar y ya nos aparecerá en el menú de la izquierda de "Recetas"

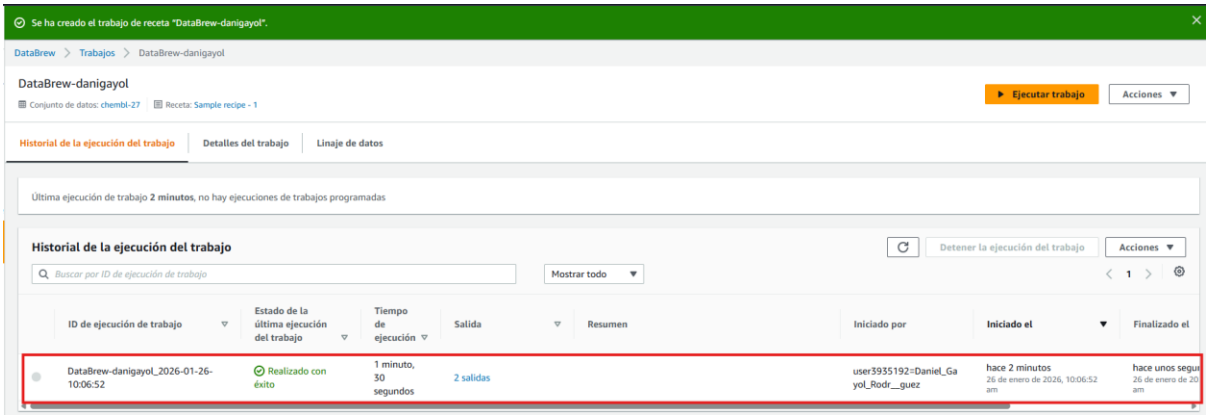




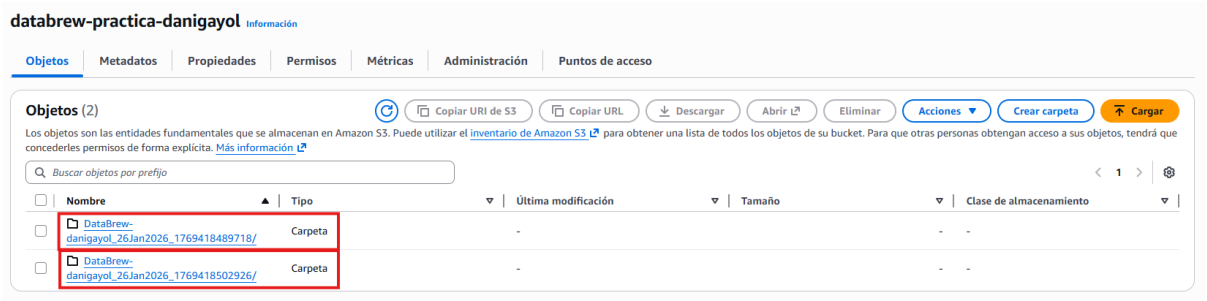
Una vez dentro de la “receta”, le daremos al botón para crear un trabajo:

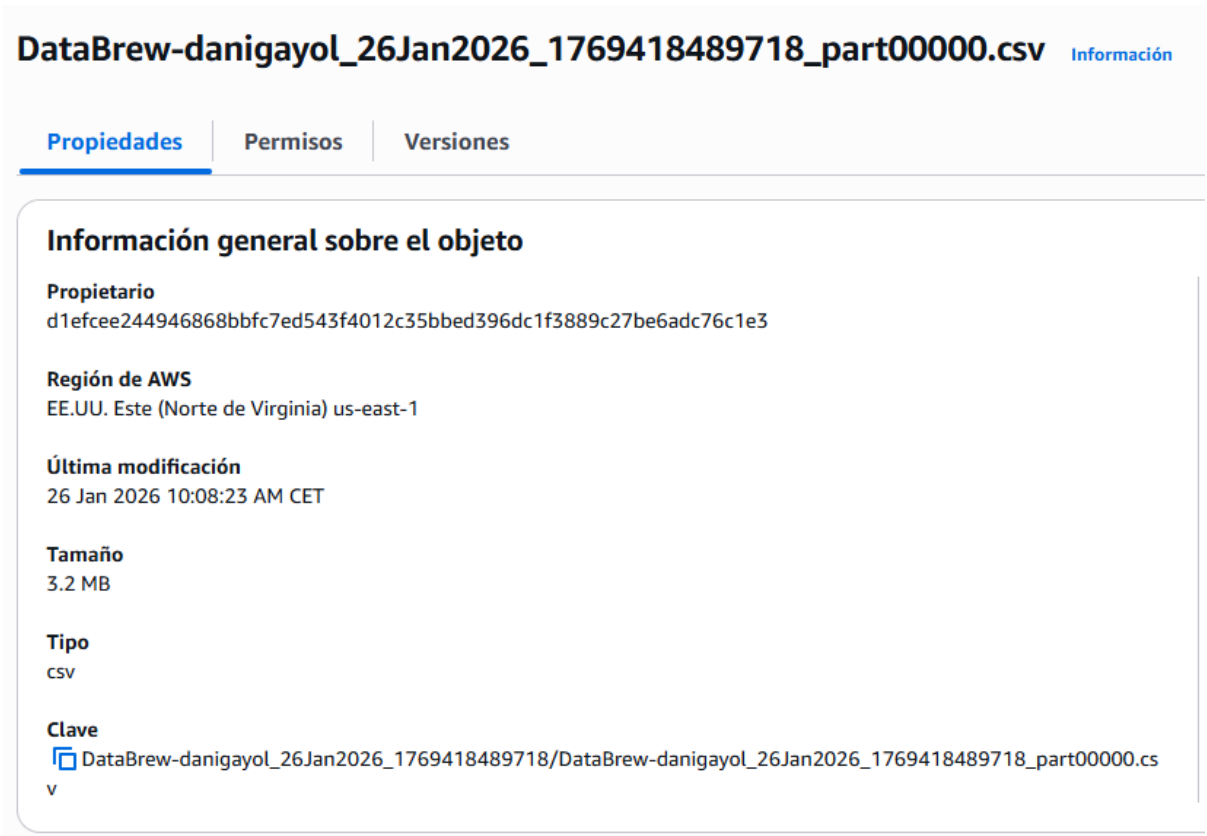



Crear trabajo
Detalles del trabajo
Nombre del trabajo
Identificador del trabajo
DataBrew-danigayol
El nombre del trabajo debe contener entre 1 y 240 caracteres. Los caracteres válidos son alfanuméricos (A-Z, a-z, 0-9), guiones (-), puntos (.) y espacios.
Tipo de trabajo
Tipo de trabajo que se va a ejecutar en el conjunto de datos
Crear un trabajo de receta
Ejecuta las transformaciones de la receta asociada en la población del conjunto de datos asociado.
Crear un trabajo de perfil
Genera un resumen y estadísticas que le dan la forma de los datos.
Entrada del trabajo
El conjunto de datos de entrada para el trabajo y la receta que se le aplicará.
Ejecutar en
Conjunto de datos
Ejecute el trabajo en un conjunto de datos DataBrew existente o nuevo.
Proyecto
Ejecute el trabajo en un proyecto sin trabajo asociado.
Elegir conjunto de datos
chembl-27
Explorar conjuntos de datos
Conectar nuevo conjunto de datos
Seleccionar una receta
Sample recipe - 1
Versión de la receta
Versión 1.0
Examinar recetas
Salida 1
Configuración
Remover
Salida a
Ubicación de la salida
Amazon S3
Tipo de archivo
Formato de salida
CSV
Delimitador
Separador CSV
Coma (,)
Compresión
Tipos disponibles
None
Cuenta de AWS del propietario del bucket de S3
Cuenta de AWS actual
043356869404
Otra cuenta de AWS
Ubicación de S3
El formato es s3://bucket/folder/
s3://databrew-practica-danigayol/
Explorar
Resumen de configuración
Almacenamiento de salida de archivos
Crear una nueva carpeta para cada ejecución de flujo de trabajo
Salida de archivo
Autogenerate files
División personalizada por valores de columna
Desactivado
Vista previa de la ruta de salida
s3://databrew-practica-danigayol/DataBrew-danigayol_26ene.2026_timestamp_part00000.csv
Salida 2
Configuración
Remover
Salida a
Ubicación de la salida
Amazon S3
Tipo de archivo
Formato de salida
PARQUET
Delimitador
Separador CSV
Coma (,)
Compresión
Tipos disponibles
None
Cuenta de AWS del propietario del bucket de S3
Cuenta de AWS actual
043356869404
Otra cuenta de AWS
Ubicación de S3
El formato es s3://bucket/folder/
s3://databrew-practica-danigayol/
Explorar
Resumen de configuración
Almacenamiento de salida de archivos
Crear una nueva carpeta para cada ejecución de flujo de trabajo
Salida de archivo
Autogenerate files
División personalizada por valores de columna
Desactivado
Vista previa de la ruta de salida
s3://databrew-practica-danigayol/DataBrew-danigayol_26ene.2026_timestamp_part00000.parquet
Permisos
DataBrew needs permission to connect to data on your behalf. Use an IAM role with the política necesaria attached.
Nombre del rol
Elija el rol que tiene acceso para conectarse a los datos. Actualice para ver las últimas actualizaciones.
LabRole
Cancelar
Crear trabajo
Crear y ejecutar trabajo



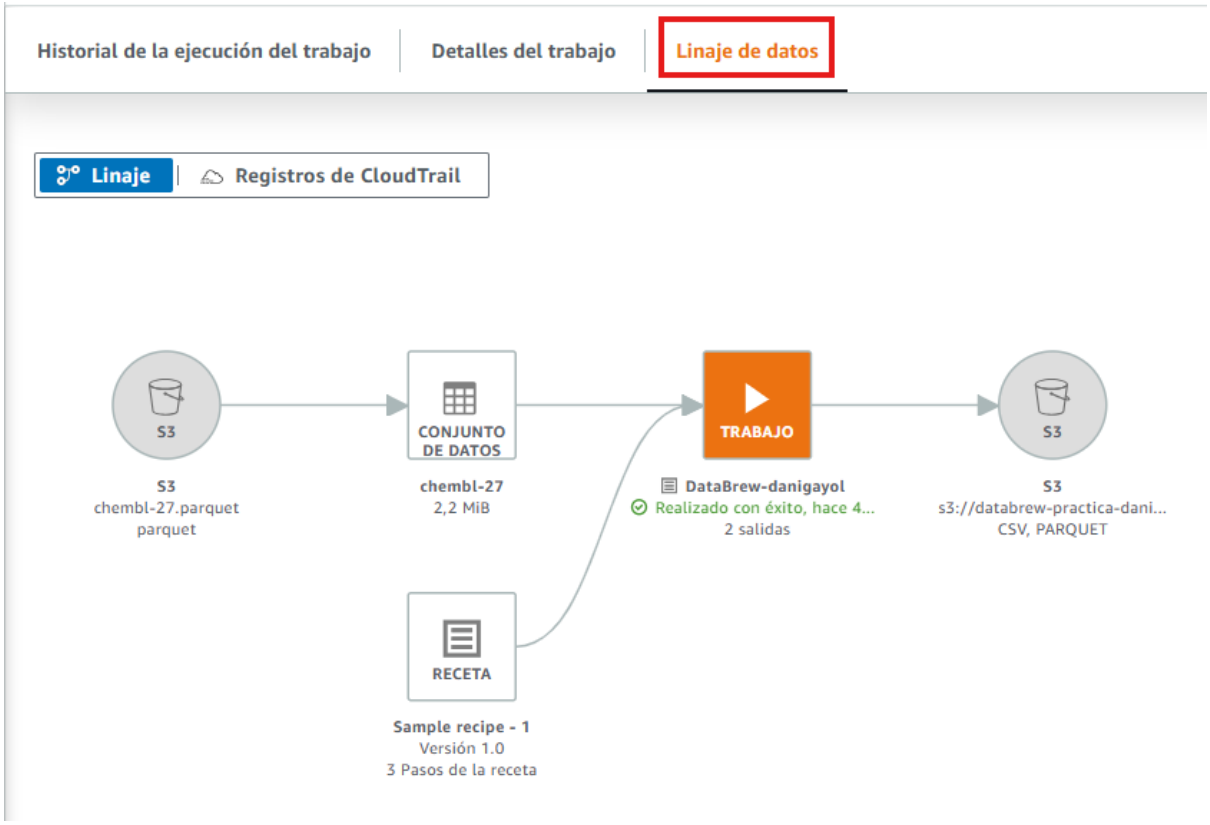
Finalmente, nos vamos a "S3" y ya nos saldrán las carpetas:





2,) Muestra también el linaje resultante de los datos.

Para mostrar el Linaje de Datos, nos vamos al "job" ejecutado anteriormente y entramos en el menú de "Linaje de Datos":


Historial de la ejecución del trabajo
Detalles del trabajo
Linaje de datos
Linaje
Registros de CloudTrail
S3
S3
chembl-27.parquet
parquet
CONJUNTO DE DATOS
chembl-27
2,2 MiB
TRABAJO
DataBrew-danigayol
Realizado con éxito, hace 4...
2 salidas
S3
S3
s3://databrew-practica-dani...
CSV, PARQUET
RECETA
Sample recipe - 1
Versión 1.0
3 Pasos de la receta