



## Data Management

### KeepCoding

**Fechas: 12 y 13 de Julio de 2019.**

**Horas: 8h**

**Instructora: Nerea Sevilla**

**Nombre Material: Ejercicios\_DataCleansing\_Trifacta**

## Data Cleansing con Trifacta Wrangler

### Ejercicio 1.-Data Cleansing de Contactos de Empresas

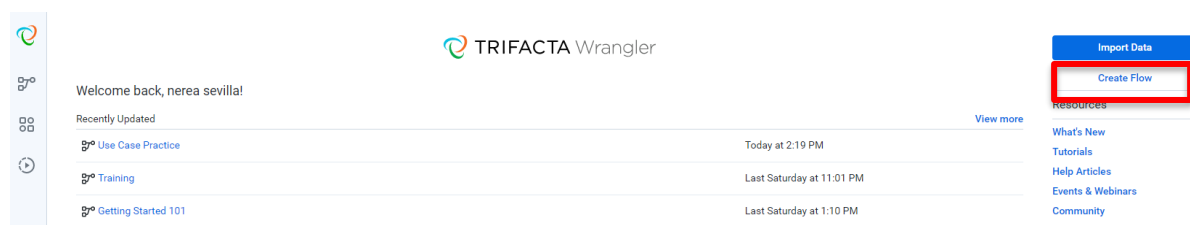
Para empezar la limpieza, utilizaremos un ejemplo sencillo de depuración de datos, vamos a trabajar con fichero Excel entregado como parte del material del módulo de Calidad de datos. El fichero contiene información de empresas con las que queremos realizar acciones de marketing.

#### Paso 1. Crear un nuevo Flujo

En primer crearemos un Flujo: Flows; Create Flow

Un flujo

- Se utiliza para organizar datos relevantes y transformaciones en su flujo de trabajo.
- Permite visualizar las relaciones entre los conjuntos de datos y cómo se vinculan.
- Tiene propagación de cambios automática: esto significa que los cambios realizados en cualquier conjunto de datos se aplican automáticamente y se guardan en todos los conjuntos de datos afectados.



Indicaremos nombre, descripción y pulsaremos CREATE para crear el flujo.



Create Flow

×

Flow Name





Empresas

Flow Description

Relación de Empresas para Acciones de Marketing

Cancel>Create

Ahora tenemos que agregar los datos al flujo de datos.




Flows

Empresas

Relación de Empresas para Acciones de Marketing

...



Add Datasets into this Flow to start wrangling.

Add Datasets

Para ello importaremos el fichero Excel “Empresas.xls” entregado como parte del material del curso, utilizando el botón Add Datasets.

Podemos utilizar un dataset que hayamos importado ya y que aparece en la pantalla de datasets, o importar uno pulsando al botón Import Dataset.



Add Datasets to Flow

Search...

All (12)Imported (12)Reference (0)

<input type="checkbox"/>	lab_2011_transactions.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	lab_2012_transactions.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	lab_2013_transactions.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	lab_2014_transactions.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	lab_2015_transactions.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	zip_to_state_map.csv	Upload	Yesterday at 4:50 PM
<input type="checkbox"/>	lab_customers.csv	Upload	Yesterday at 4:09 PM
<input type="checkbox"/>	WL_Home_Grown_Farmers_Market_Sales 07-16-2...	Upload	Last Saturday at 9:14 ...
<input type="checkbox"/>	WL_Home_Grown_Farmers_Market_Sales 07-15-2...	Upload	Last Saturday at 9:14 ...
<input type="checkbox"/>	US_Farmers_Markets.csv US farmers market exploration dataset	Upload	Last Saturday at 2:20 ...
<input type="checkbox"/>	Customer_Data.csv	Upload	Last Saturday at 1:06 ...

Import Datasets

CancelAdd

El siguiente paso es elegir, en la pantalla que se nos presenta el fichero que queremos cargar.

Import Data and Add to Flow

Upload

Upload from your computer

Drag & drop a file here or

Choose a file

Maximum upload file size: 100MB

0 New Datasets

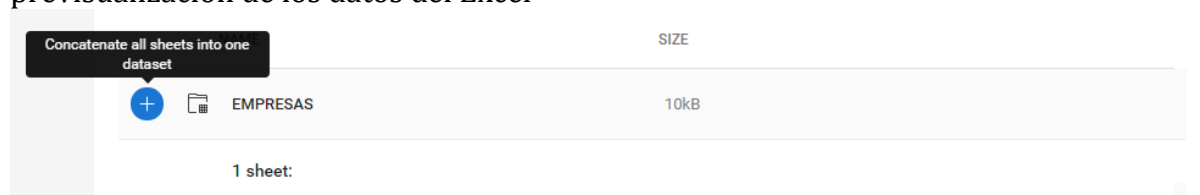
Choose data to import.

Import & Add to FlowCancel

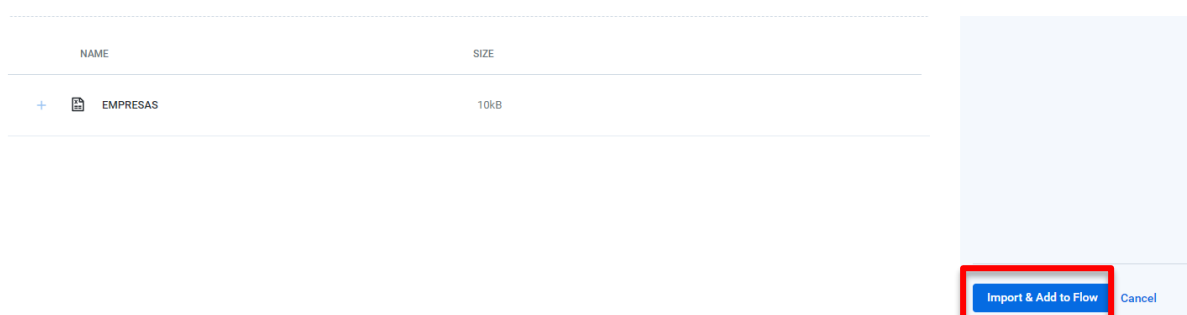


Pulsamos el botón Choose a file, para que aparezca el cuadro de dialogo que nos permitirá añadir el fichero Empresas.xls del dataset que hemos creado.

Confirmamos que queremos importar la hoja Empresas, y se nos muestra una previsualización de los datos del Excel-



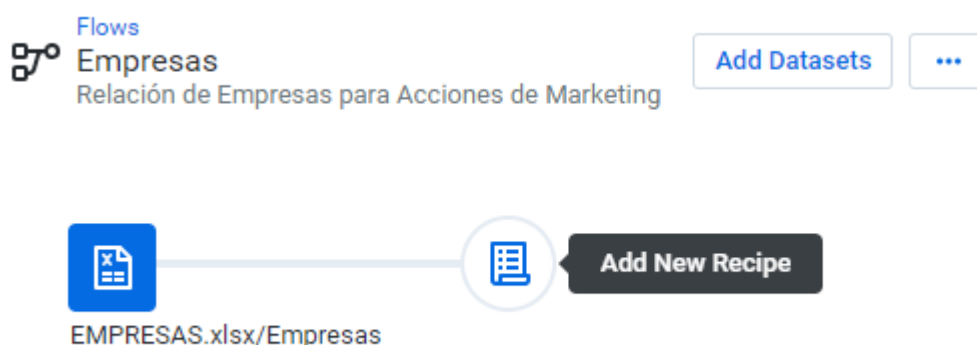
Y pulsamos al botón Importar & Add to Flow, para proceder a cargar los datos y añadir el DataSet al Flujo.



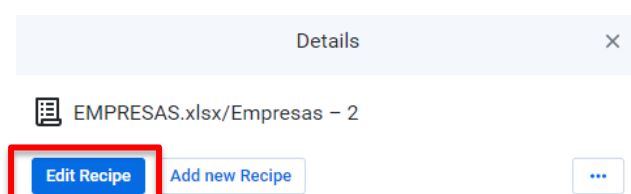
## Paso 2. Construir Receta.

El siguiente paso es indicar a Trifacta la creación del script de DataCleaning o receta, que será el conjunto de pasos que iremos realizando sobre los datos para la limpieza.

Para ello pulsaremos la opción de Add new Recipe.



Una vez creada la receta, pulsaremos al botón Edit Recipe para ir aplicando las diferentes acciones que nos permitan preparar y limpiar el dataset.





Al editar la receta podemos empezar a preparar nuestros datos aplicando diferentes acciones o pasos.

### Paso 3. Transformaciones.

En la pantalla de transformaciones de Trifacta se nos muestra información visual sobre los datos y genera sugerencias para aplicar a los datos.

En la pantalla de transformaciones hay dos formas de ver los datos: Grid y Columns



Grid

La vista de cuadrícula le permite ver los datos en una vista tabular, la mayor de las transformaciones se realizan desde esta vista.

La vista columnas permite ver la distribución de las columnas, se utiliza sobre todo para realizar varias operaciones sobre columnas de características similares, como por ejemplo cambiar los tipos de columnas.

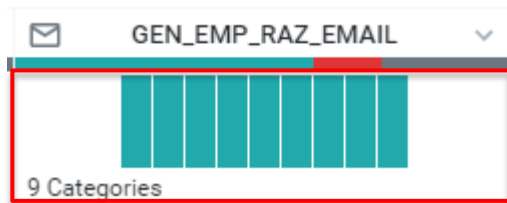


De entrada identifica correctamente el nombre de las columnas y el tipo de dato, si es integer o texto, formato CP (ZIP) formato email (@), etc...

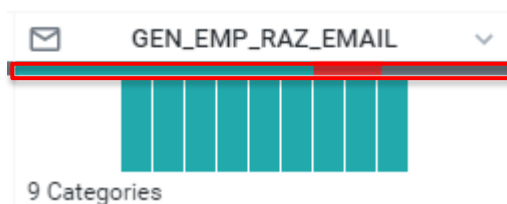
El primer paso para manejar cualquier conjunto de datos nuevo es perfilar los datos y descubrir su contenido.

Trifacta nos proporciona 2 herramientas que facilita el descubrimiento de los tipos de columna.

**1.-Histograma**, que visualiza el profiling de información de cada columna.



**2.- La barra de calidad de datos.**



Es la barra que aparece debajo de las etiquetas de nombres de los campos, con diferentes colores.

1. Si es verde, representa la cantidad de datos correctos. Los datos coinciden con el formato de la celda.
2. Si es negra, indica que hay algún problema o error de calidad de datos con valores erróneos o defectuosos. cantidad de datos incorrectos. Los datos no coinciden con el formato de la celda.
3. Si es roja, indica que existen valores inválidos.

La primera transformación la haremos sobre la columna GEN\_EMP\_RAZ\_NUM\_VIA que no tiene datos y la barra de calidad de datos nos marca todos los registros como inválidos, pulsaremos encima de la columna y elegiremos borrar la columnas de las sugerencias aportadas por Trifacta, pulsando Add.





Source	to be dropped	Preview
SSN	GEN_EMP_RAZ_TLF	SSN GEN_EMP_RAZ_TLF1 # GEN_EMP_RAZ_TLF2
12 Categories		14 Categories 915.03M - 945.89M
875264755		875264755 null
875241800		875241800 null
8745 36 18 02		8745 36 18 02 null
875291500		875291500 null
875214150		875214150 null
875144400		875144400 null
875465629		875465629 null
876719300		876719300 null
874037352/915031700		874037352 915031700
875254000		875254000 null
875290636		875290636 null
875225850		875225850 null
875893015/945893027		875893015 945893027
875371200		875371200 null
875890334		875890334 null

**Replace**  
`/` with `\*` in GEN\_EMP\_RAZ\_TLF  
`{delim}` with `\*` in GEN\_EMP\_RAZ\_TLF  
**Split on values matching** [See all](#)  
`/`  
`/` starting after `{digit}{9}` ending before `{digit}{9}`  
`/` starting after `874037352` ending before `915031700`  
**Extract values matching** [See all](#)  
`/`  
`/` starting after `{digit}{9}` ending before `{digit}{9}`  
`/` starting after `874037352` ending before `915031700`

Como resultado de la operación, obtenemos una nueva columna GEN\_EMP\_RAZ\_TLF2, con el segundo teléfono de contacto para la empresa.

Otra operación que realizaremos sobre el teléfono principal es quitar los espacios en blanco para que todos los teléfonos tengan el mismo formato, para ello en el campo hacemos clic en el espacio en blanco del registro invalido y elegimos de las sugerencias la que nos ofrece de Replace, para eliminar el espacio en blanco.

Source	to be dropped	Preview
SSN	GEN_EMP_RAZ_TLF1	SSN GEN_EMP_RAZ_TLF1 # GEN_EMP_RAZ_TLF2
14 Categories		14 Categories 915.03M - 945.89M
875264755		875264755
875241800		875241800
8745-36-18-02		8745361802
875291500		875291500
875214150		875214150
875144400		875144400
875465629		875465629
876719300		876719300
874037352		874037352
875254000		875254000
875290636		875290636
875225850		875225850
875893015		875893015
875371200		875371200
875890334		875890334

**Split on values matching** [See all](#)  
` ` 3 times  
`{delim}`  
` `   
**Extract values matching** [See all](#)  
`{delim}`  
` `   
`{delim}` starting after `{digit}{4}` ending before `{digit}{2}`  
**Count values matching** [See all](#)  
`{delim}`  
` `   
`{delim}` starting after `{digit}{4}` ending before `{digit}{2}`  
**Replace**  
`{delim}` with `\*` in GEN\_EMP\_RAZ\_TLF1  
[Edit](#) [Add](#)

Ya sólo nos queda un registro invalido, lo que haremos es aceptar la sugerencia de Trifacta para marcar como nulo el teléfono que es incorrecto.

Source	to be dropped	Preview
SSN	GEN_EMP_RAZ_TLF1	SSN GEN_EMP_RAZ_TLF1 # GEN_EMP_RAZ_TLF2
13 Categories		13 Categories 915.03M - 945.89M
875264755		875264755 null
875241800		875241800 null
8745361802		null
875291500		875291500 null
875214150		875214150 null
875144400		875144400 null
875465629		875465629 null
876719300		876719300 null
874037352		874037352 915031700
875254000		875254000 null
875290636		875290636 null
875225850		875225850 null
875893015		875893015 945893027
875371200		875371200 null

**Delete rows**  
with mismatched values in GEN\_EMP\_RAZ\_TLF1  
**Keep rows**  
with mismatched values in GEN\_EMP\_RAZ\_TLF1  
**Create a new column**  
flag mismatched values in GEN\_EMP\_RAZ\_TLF1  
**Set**  
mismatched values to NULL()  
[Edit](#) [Add](#)





Ahora prepararemos el campo Nombre de Contacto GEN\_EMP\_RAZ\_CON\_NOMBRE.

El primer paso a realizar es quitar los espacios en blanco del principio del campo, añadiendo la sugerencia que nos ofrece Trifacta de quitar los espacios del principio del campo {start}{delim}, mediante la función Replace.

Source	to be dropped	Preview
GEN_EMP_RAZ_CON_NOMBRE		GEN_EMP_RAZ_CON_NOMBRE
12 Categories		12 Categories
OSCAR MUSITU		OSCAR MUSITU
LAURA IBARROLA		LAURA IBARROLA
ALICIA SANZ PINEDO (RRHH)		ALICIA SANZ PINEDO (RRHH)
MIGUEL HERNAEZ DUQUE		MIGUEL HERNAEZ DUQUE
IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO)		IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO)
LOURDES ALZOLA MILIKUA		LOURDES ALZOLA MILIKUA
RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO)		RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO)
JAVIER TREVIÑO IZQUIERDO		JAVIER TREVIÑO IZQUIERDO
BLANCA IBISATE		BLANCA IBISATE
JAVIER ANUNCIBAY MARTINEZ		JAVIER ANUNCIBAY MARTINEZ
VICTOR GALLAGA MOZOS		VICTOR GALLAGA MOZOS
IGNACIO VALLE SOLANO		IGNACIO VALLE SOLANO

La siguiente transformación que haremos es separar los nombres, apellidos y cargo, para ello elegimos los espacios en blanco, y elegimos el paso de división sugerido por Trifacta.

Source	to be dropped	Preview
GEN_EMP_RAZ_CON_NOMBRE		GEN_EMP_RAZ_CON_NOMBRE1 GEN_EMP_RAZ_CON_NOMBRE2 GEN_EMP_RAZ_CON_NOMBRE3 GEN_EMP_RAZ_CON_NOMBRE4
12 Categories		11 Categories 12 Categories 9 Categories
OSCAR MUSITU		OSCAR MUSITU null null null
LAURA IBARROLA		LAURA IBARROLA null null null
ALICIA SANZ PINEDO (RRHH)		ALICIA SANZ PINEDO (RRHH) null null null
MIGUEL HERNAEZ DUQUE		MIGUEL HERNAEZ DUQUE null null null
IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO)		IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO) null null null
LOURDES ALZOLA MILIKUA		LOURDES ALZOLA MILIKUA null null null
RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO)		RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO) null null null
JAVIER TREVIÑO IZQUIERDO		JAVIER TREVIÑO IZQUIERDO null null null
BLANCA IBISATE		BLANCA IBISATE null null null
JAVIER ANUNCIBAY MARTINEZ		JAVIER ANUNCIBAY MARTINEZ null null null
VICTOR GALLAGA MOZOS		VICTOR GALLAGA MOZOS null null null
IGNACIO VALLE SOLANO		IGNACIO VALLE SOLANO null null null

Obtendremos 4 nuevas columnas, con la división del campo origen.

Cambiamos las etiquetas de los nombres de las 4 columnas. Haciendo click en la columna y eligiendo del menú contextual Rename.

GEN_EMP_RAZ_CON_NOMBRE1	GEN_EMP_RAZ_CON_NOMBRE2
OSCAR MUSITU	OSCAR MUSITU
LAURA IBARROLA	LAURA IBARROLA
ALICIA SANZ PINEDO (RRHH)	ALICIA SANZ PINEDO (RRHH)
MIGUEL HERNAEZ DUQUE	MIGUEL HERNAEZ DUQUE
IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO)	IDOIA BUJANDA RODRIGUEZ (ADMINISTRATIVO)
LOURDES ALZOLA MILIKUA	LOURDES ALZOLA MILIKUA
RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO)	RAQUEL MADARIAGA LOMAS (ADMINISTRATIVO)
JAVIER TREVIÑO IZQUIERDO	JAVIER TREVIÑO IZQUIERDO
BLANCA IBISATE	BLANCA IBISATE
JAVIER ANUNCIBAY MARTINEZ	JAVIER ANUNCIBAY MARTINEZ
VICTOR GALLAGA MOZOS	VICTOR GALLAGA MOZOS
IGNACIO VALLE SOLANO	IGNACIO VALLE SOLANO

Cambiamos la etiqueta de la primera columna de las 4 nuevas por GEN\_EMP\_RAN\_CON\_NOMBRE.



Preview				Option	
RBC	GEN_EMP_RAZ_CON_NOMBRE	RBC	GEN_EMP_RAZ_CON_NOMBRE2	RBC	GEN_EMP_RAZ_CON_NOMBRE3
				Manual rename	
				Specify the new name for each column	
				Columns (1)	
				GEN_EMP_RAZ_CON_NOMBRE1	
				GEN_EMP_RAZ_CON_NOMBRE	
				Cancel Add	

Repetimos la misma operación para las otras 3 nuevas columnas. Le aplicamos las etiquetas GEN\_EMP\_RAZ\_CON\_APELLIDO1, para el primer apellido del nombre de contacto de la empresa.

Preview				Option	
RBC	GEN_EMP_RAZ_CON_APELLIDO1	RBC	GEN_EMP_RAZ_CON_NOMBRE3	RBC	GEN_EMP_RAZ_CON_NOMBRE2
				Manual rename	
				Specify the new name for each column	
				Columns (1)	
				GEN_EMP_RAZ_CON_NOMBRE2	
				GEN_EMP_RAZ_CON_APELLIDO1	
				Cancel Add	

GEN\_EMP\_RAZ\_CON\_APELLIDO2, para el segundo apellido del nombre de contacto de la empresa.

Preview				Option	
RBC	GEN_EMP_RAZ_CON_APELLIDO2	RBC	GEN_EMP_RAZ_CON_NOMBRE4	RBC	GEN_EMP_RAZ_CON_NOMBRE3
				Manual rename	
				Specify the new name for each column	
				Columns (1)	
				GEN_EMP_RAZ_CON_NOMBRE3	
				GEN_EMP_RAZ_CON_APELLIDO2	
				Cancel Add	

Y por último, el cargo de la persona de contacto.



Preview

RBC	GEN_EMP_RAZ_CON_CARGO	RBC	GEN_EMP_RAZ_COD_PROV
2 Categories		3 Categories	
null		Araba	
null		Araba	
(RRHH)		Araba	
null		ARABA	
(ADMINISTRATIVO)			
null		Araba	
(ADMINISTRATIVO)		Araba	
null		BIZKAIA	
null		Araba	
null		Araba	
null		Araba	
null		Araba	
null		Araba	
null		Araba	

Option

Manual rename

Specify the new name for each column

Columns (1)

GEN\_EMP\_RAZ\_CON\_NOMBRE4

GEN\_EMP\_RAZ\_CON\_CARGO

Cancel

Add

Recordar que en la receta se pueden eliminar pasos realizados para volver a estados anteriores de los datos.

Run Job

New Step

Recipe

1 Delete GEN\_EMP\_RAZ\_NUM\_VIA

2 Split GEN\_EMP\_RAZ\_TLF on delimiters matching '/' into 2 columns

3 Replace matches of '{delim}' from GEN\_EMP\_RAZ\_TLF1 with ''

4 Set GEN\_EMP\_RAZ\_TLF1 to IFMISMATCHED(\$col, ['SSN'], 'N/A')

5 Replace matches of '{start}{delim}' from GEN\_EMP\_RAZ\_CON\_NOMBRE with ''

6 Split GEN\_EMP\_RAZ\_CON\_NOMBRE on delimiters matching '' into 4 columns

7 Rename GEN\_EMP\_RAZ\_CON\_NOMBRE1 to 'GEN\_EMP\_RAZ\_CON\_NOMBRE'

8 Rename GEN\_EMP\_RAZ\_CON\_NOMBRE2 to 'GEN\_EMP\_RAZ\_CON\_APELLIDO1'

9 Rename GEN\_EMP\_RAZ\_CON\_NOMBRE3 to 'GEN\_EMP\_RAZ\_CON\_APELLIDO2'

10 Rename GEN\_EMP\_RAZ\_CON\_NOMBRE4 to 'GEN\_EMP\_RAZ\_CON\_CARGO'

Par el cargo de la persona de contacto vamos eliminar los paréntesis y dejar sólo el texto. Para ello seleccionamos el paréntesis del comienzo y añadimos la sugerencia de Trifacta de sustituirlo por vacío.



Source	to be dropped	Preview	
RBC	GEN_EMP_RAZ_CON_CARGO	RBC	GEN_EMP_RAZ_CON_CARGO
2 Categories		2 Categories	
null		null	
null		null	
null		null	
RRHH)		RRHH)	
ADMINISTRATIVO)		ADMINISTRATIVO)	
ADMINISTRATIVO)		ADMINISTRATIVO)	
null		null	
null		null	
null		null	
null		null	
null		null	
null		null	
null		null	

**Replace**  
`\\` with `\*` in GEN\_EMP\_RAZ\_CON\_CARGO  
[Edit](#) [Add](#)

**Extract values matching** [See all](#)  
`\\`  
`\\` starting after `{start}` ending before `{upper}{4}`  
`\\` starting after `{start}` ending before `R`

**Split on values matching** [See all](#)  
`\\`  
`\\` starting after `{start}` ending before `{upper}{4}`  
`\\` starting after `{start}` ending before `R`

Repetimos la misma operación con el paréntesis del final del GEN\_EMP\_RAZ\_CON\_CARGO.

Source	to be dropped	Preview	
RBC	GEN_EMP_RAZ_CON_CARGO	RBC	GEN_EMP_RAZ_CON_CARGO
2 Categories		2 Categories	
null		null	
null		null	
null		null	
RRHH)		RRHH	
ADMINISTRATIVO)		ADMINISTRATIVO	
ADMINISTRATIVO)		ADMINISTRATIVO	
null		null	
null		null	
null		null	
null		null	
null		null	
null		null	
null		null	

**Replace**  
`\\` with `\*` in GEN\_EMP\_RAZ\_CON\_CARGO  
[Edit](#) [Add](#)

**Split on values matching** [See all](#)  
`\\`  
`\\` starting after `O` ending before `{end}`  
`\\` starting after `{upper}{14}` ending before `{end}`

**Extract values matching** [See all](#)  
`\\`  
`\\` starting after `O` ending before `{end}`  
`\\` starting after `{upper}{14}` ending before `{end}`

**Y ya tenemos separados los campos correspondientes al Contacto de la Empresa.**

Pasamos a la preparación del campo Email, para tratar los registros inválidos, para mostrar sólo dichos registros. Primero seleccionamos los registros inválidos, haciendo clic en la barra de calidad, sobre el color rojo y posteriormente, seleccionamos la opción Rows que aparece en el margen inferior derecho de la pantalla.



Preview					
SSN	GEN_EMP_RAZ_TLF1	#	GEN_EMP_RAZ_TLF2	GEN_EMP_RAZ_EMAIL	RBC
14 Categories	915.03M - 945.89M		9 Categories	11 Categories	12 Categories
875893015		945893027	wesiz@inauxa.es/ - inauxa@sea.es	VICTOR	GALLAGA
875890334		null	-		null

13 Columns 2 Rows 5 Data Types

Show only affected ☒ Rows

Se puede observar que uno de los registros tiene dos emails. Vamos a hacer Split por "/" para separarlos. Add en la sugerencia de TRIFACTA. **"Split on value matching"**.

Source	to be dropped	Preview			
GEN_EMP_RAZ_EMAIL		GEN_EMP_RAZ_EMAIL1	GEN_EMP_RAZ_EMAIL2	RBC	GEN_EMP_RAZ_EMAIL
9 Categories		10 Categories	1 Category		11 Categories
weso@aaf.es		weso@aaf.es	null		OSCAR
wesanciero@omegaelevator.com		wesanciero@omegaelevator.com	null		LAURA
wesola@sagola.com		wesola@sagola.com	null		ALICIA
wesin@motorgorbea.net.bmw.es		wesin@motorgorbea.net.bmw.es	null		MIGUEL
wesdaria@tubacex.com		wesdaria@tubacex.com	null		IDAIA
weslex@parklex.com		weslex@parklex.com	null		LOURDES
wesmeo@ikastola.eus		wesmeo@ikastola.eus	null		RAQUEL
wes.vitoria@uribesalgo.com		wes.vitoria@uribesalgo.com	null		JAVIER
wesiz@inauxa.es/ - inauxa@sea.es		wesiz@inauxa.es	inauxa@sea.es		BLANCA
weslle@tauxme.es		weslle@tauxme.es	null		JAVIER
			null		IGNACIO

Replace

'(delim-ws)' with '\*' in GEN\_EMP\_RAZ\_EMAIL

'/' with '\*' in GEN\_EMP\_RAZ\_EMAIL

Split on values matching

'(delim-ws)'

'/'

'(delim-ws)' starting after '(email)' ending before '(url)'

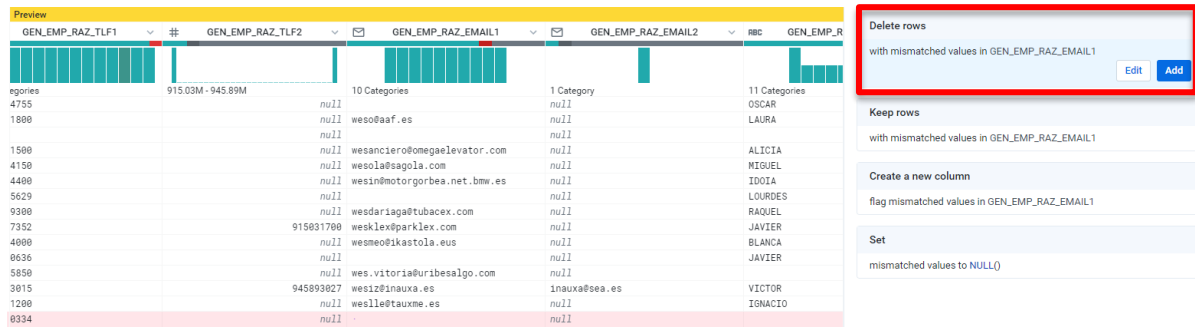
Extract values matching

'(delim-ws)'

'/'

'(delim-ws)' starting after '(email)' ending before '(url)'

Ahora sólo nos queda un registro inválido para el email. Se trata de un registro sin dato, y del que tampoco tenemos el teléfono, lo eliminamos, añadiendo la sugerencia de TRIFACTA.



Haciendo click en la columna GEN\_EMP\_RAZ\_COD\_PROV, utilizaremos la sugerencia que nos permite realizar la operación que deseamos realizar.

Source

to be dropped

RBC

GEN\_EMP\_RAZ\_COD\_PROV

3 Categories

Araba

Araba

Araba

Araba

Araba

BIZKAIA

Araba

Araba

Araba

RBC\_EMP\_RAZ\_COD\_PROV

BIZKAIA

Patterns

Upper}{Lower}{4}

Upper}{5}

Upper}{7}

Show pattern details...

Suggestions

Rename

Rename GEN\_EMP\_RAZ\_COD\_PROV to 'GEN\_EMP\_RAZ\_COD\_PROV'

Set

missing values to \*

Delete columns

GEN\_EMP\_RAZ\_COD\_PROV

Create a new column

ANY(GEN\_EMP\_RAZ\_COD\_PROV)

Values to columns

for each unique value in GEN\_EMP\_RAZ\_COD\_PROV

Text Format

Lowercase GEN\_EMP\_RAZ\_COD\_PROV

Uppercase GEN\_EMP\_RAZ\_COD\_PROV

Edit

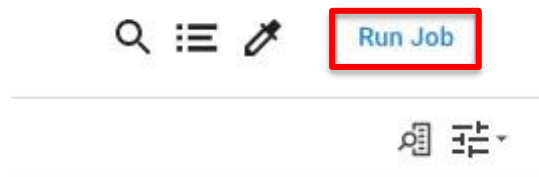
Add

### Ejecución de Trabajo (Run Job).

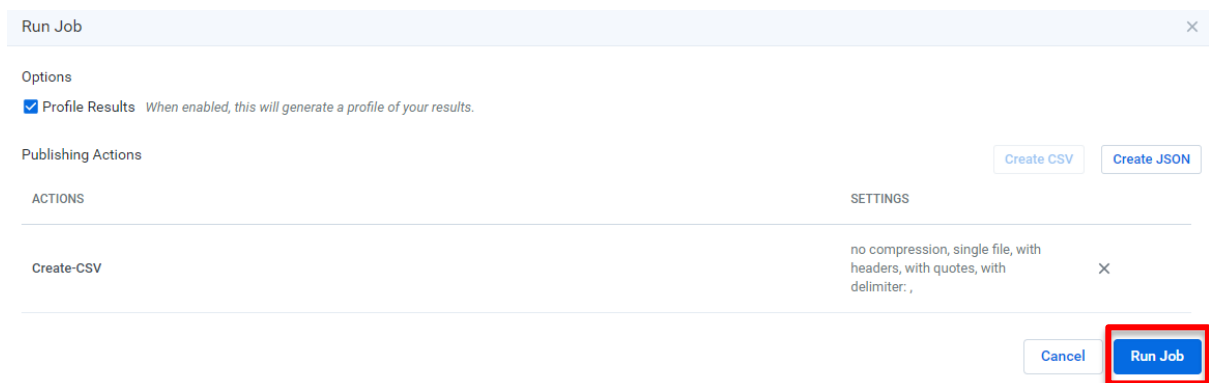
The diagram illustrates a data processing workflow. It begins with a box labeled "SOURCE DATA" containing a grid of small rectangles. An arrow labeled "IMPORT" points from this box to a central circular processing area. Inside this circle are three icons: a document, a circular arrow, and a refresh symbol. An arrow labeled "OUTPUT" points from the central circle to a box labeled "NEW RESULTS DATA", which also contains a grid of small rectangles.



Una vez que tenemos realizada la receta, se aplican las transformaciones a los datos de entrada para generar la salida.  
Para ejecutar un trabajo, elija el botón Ejecutar trabajo en la parte superior derecha.



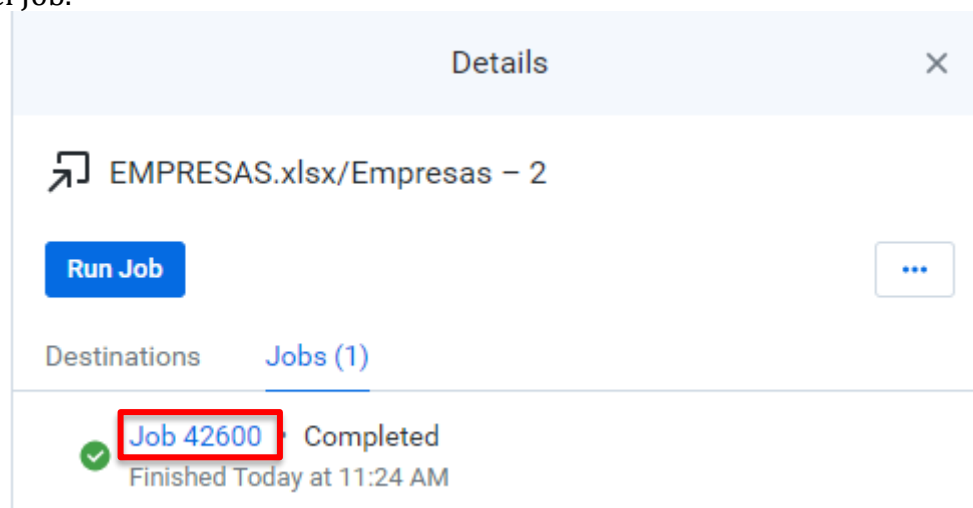
En la versión libre de Trifacta, solo se permite generar salida CSV o crear JSON.



Al pulsar en el botón Run Job, comenzará la ejecución del trabajo que aplica las transformaciones y nos generará la salida y unos resultados del profiling.



Una vez finalizada la ejecución del trabajo podemos visualizar los resultados, haciendo clic en el Job.





Los resultados se muestran en una pantalla de Resumen, en la pestaña de Output, podemos descargar el fichero csv generado por la ejecución del trabajo, que podremos integrar por ejemplo a nuestro CRM.

Empresas > EMPRESAS.xlsx/Empresas - 2

Job 42600

Finished Today at 11:24 AM

Download results

...

Overview

Output Destinations

Profile

Dependencies

Name	Status	Duration
EMPRESAS.xlsx/Empresas - 2.csv	Completed	<div>View details</div> <div>...</div> <div>Download result</div> <div>Create imported dataset</div>





## Ejercicio 2.-Data Cleaning de Datos Abiertos

Utilizar el fichero de datos abiertos del Gobierno de Canarias (<https://opendata.gobiernodecanarias.org/dataset/observacion-de-cetaceos-autorizados>). El dataset contiene observaciones de cetáceos por parte de empresas autorizadas en las costas de las islas. Son datos recogidos por diferentes empresas y por esta razón se registran datos con errores de calidad de datos. Descargaremos el fichero con formato XLS, en el csv presenta problemas con las letras acentuadas.

Crearemos el Flujo (Flow) e importaremos el dataset con el fichero xls para cargar los datos. Posteriormente crearemos una receta nueva y la editaremos para comenzar a realizar transformaciones para preparar el fichero de salida.

Al editar la receta, entramos en la pantalla de transformaciones, visualizándose el profiling que Trifacta realiza de las columnas del dataset y nos presenta la información, reconociendo las columnas del fichero y los tipos de campo.

El primer paso de transformación es eliminar todos los campos que no son de contacto como Situación Administrativa, Código de Identificación y Signatura.

El siguiente paso es obtener todos los campos de contacto que están mezclados en la columna Teléfonos/Internet. Teléfonos, correo electrónico, fax y web, utilizando una división por el carácter punto y coma (;)



The source table has a column 'Teléfonos/Internet' with values like 'Teléfonos: 636 271 841,928 770 059;'. The 'Replace' panel shows a rule to replace 'with ' in Teléfonos/Internet'.

Así obtenemos 4 campos diferentes. Vamos a transformar los números de Teléfono, realizando también una división, pero antes quitamos el texto “Teléfonos:”

The source table is split into four columns: 'Teléfonos/Internet1', 'Teléfonos/Internet1', 'Teléfonos/Internet2', and 'Teléfonos/Internet2'. The 'Extract values matching' panel shows a rule to extract phone numbers from the first column.

Reemplazamos el carácter “ - ” por “,” para poder obtener un patrón de separación el campo teléfono.

The source table shows phone numbers with hyphens replaced by commas. The 'Replace with' panel shows a rule to replace hyphens with commas.

Y ahora ya podemos obtener tres campos los números de teléfonos dividiendo el campo origen utilizando el carácter “,”

The source table is split into three columns: 'Teléfonos/Internet1', 'Teléfonos/Internet5', and 'Teléfonos/Internet6'. The 'Split on values matching' panel shows a rule to split the phone numbers on commas.

Ahora quitamos espacios en blanco en todas las columnas de teléfono.

The source table shows phone numbers with spaces removed. The 'Split on values matching' panel shows a rule to split the phone numbers on commas.

Cambiamos el nombre de las etiquetas a los campos.



Preview					Option	
ZIP	Teléfono_1	ZIP	Teléfonos/Internet6	#	Teléfonos/Internet7	RBC
70 Categories	30 Categories			609.53M - 900.71M		40 Categories
922480130	null					
629190382	null					
616214878	null					
666897788	null					Correos: catanarhaba@hotmail.com
922212627	null					Correos: antonia@jamelias.es
922716864	null					Faxes: 922.724.852
922794169	667668931					Correos: administracion@temerifed
922798844	666542888					Correos: info@barcostenerife.com
6865673611	670642864					null

Examinamos los registros que nos han quedado inválidos para los números de teléfonos. En el caso de la columna Teléfono\_1, reemplazamos el registro con el número invalido por el valor NULL

Source					to be dropped		Preview		Delete rows
ZIP	Teléfono_1	ZIP	Teléfono_1	ZIP	Teléfono_2	#	Teléfono_3	RBC	
70 Categories	70 Categories	30 Categories		609.53M - 900.71M		40 Categories			with mismatched values in Teléfono_1
666244080	666244080								
928851448	928851448								
660958282	660958282								
928540133	928540133								
617435237	617435237								
928150010	928150010								
928270748	928270748								
928243685	928243685								
928151266	928151266								
928355981	928355981								
928243685	928243685								
636271841	636271841								

En el caso de los registros inválidos del teléfono\_2, sustituimos el carácter ‘.’ por NULL.

Source					to be dropped		Preview		Split on values matching
ZIP	Teléfono_2	ZIP	Teléfono_2	#	Teléfono_3	RBC			
30 Categories	32 Categories			609.53M - 900.71M		40 Categories			See all
null	null								“.” 2 times
null	null								“{delim}”
null	null								“.”
null	null								
655991903	655991903								
629201978	629201978								
609161299	609161299								
687518544	687518544								
609507186	609507186								
609161299	609161299								
928770059	928770059								
649925918	649925918								
609359020	609359020								
928770059	928770059								
null	null								
928772222	928772222								
null	null								
679162031	679162031								
679162031	679162031								
616472265	616472265								

Ya hemos realizado todos los pasos en la receta para obtener los teléfonos de contacto de las empresas de avistamiento de cetáceos.

Ahora vamos a realizar los pasos de transformación para obtener el email, el fax y la página web de contacto.

Extraemos el email de la columna Teléfonos\_internet2.

Source		Preview	
RBC	Teléfonos/Internet2	☐ Teléfonos/Internet1	RBC Teléfono
40 Categories		34 Categories	
-		null	null
-		null	null
-		null	null
Correos:	nauticapuertojandia@yahoo.es	nauticapuertojandia@yahoo.es	-
-		null	-
Correos:	spirit.reservas@gmail.com, jzaera@gmail.com	spirit.reservas@gmail.com	- Faxes: 928 15
Correos:	sonia@multifinanzas.es	sonia@multifinanzas.es	- Faxes: 928 27
Correos:	canariasyachts@telefonica.net	canariasyachts@telefonica.net	- Faxes: 928 24
Correos:	administracion@grupotomassosa.com	administracion@grupotomassosa.com	- Páginas web: I
Correos:	bahiacat@bahiacat.com	bahiacat@bahiacat.com	- Páginas web: I
Correos:	canariasyachts@telefonica.net	canariasyachts@telefonica.net	- Faxes: 928 24
Correos:	zenonmanero@gmail.com	zenonmanero@gmail.com	-
Correos:	lorisalmone@hotmail.com	lorisalmone@hotmail.com	- Páginas web: I
Correos:	francarvajal66@gmail.com	francarvajal66@gmail.com	-
-		null	null
-		null	null
null			-
Correos:	info@taxibootcanarias.com	info@taxibootcanarias.com	-

Y renombramos columna.

Preview

email

RBC

Teléfono

34 Categories

25 Categories

34 Categories	25 Categories
null	null
null	null
null	null
nauticapuertojandia@yahoo.es	.
null	null
null	null
spirit.reservas@gmail.com	· Faxes: · 928 · 15
sonia@multifinanzas.es	· Faxes: · 928 · 27
canariasyachts@telefonica.net	· Faxes: · 928 · 24
administracion@grupotomassosa.com	· Páginas web: r

Option

Manual rename

Specify the new name for each column

Columns (1)




Teléfonos/Internet1

email

Cancel

Add

Eliminamos la columna Teléfonos/email porque ya no la necesitamos.

Drop Column			
RBC	Teléfonos/Internet2	email	RBC
 <p>40 Categories</p> <ul style="list-style-type: none"> <li>-</li> <li>-</li> <li>-</li> <li>null</li> <li>Correos: nauticapuertojandia@yahoo.es</li> <li>Correos: spirit.reservas@gmail.com,jzaera@gmail.com</li> <li>Correos: sonia@multifinanzas.es</li> <li>Correos: canariasyachts@telefonica.net</li> <li>Correos: administracion@grupotomassosa.com</li> <li>Correos: bahiacat@bahiacat.com</li> <li>Correos: canariasyachts@telefonica.net</li> <li>Correos: zenonmanero@gmail.com</li> <li>Correos: lorisalmone@hotmail.com</li> <li>Correos: francarvajal66@gmail.com</li> <li>null</li> <li>Correos: info@taxiboatcanarias.com</li> </ul>	 <p>34 Categories</p> <ul style="list-style-type: none"> <li>null</li> <li>null</li> <li>null</li> <li>nauticapuertojandia@yahoo.es</li> <li>null</li> <li>null</li> <li>spirit.reservas@gmail.com</li> <li>sonia@multifinanzas.es</li> <li>canariasyachts@telefonica.net</li> <li>administracion@grupotomassosa.com</li> <li>bahiacat@bahiacat.com</li> <li>canariasyachts@telefonica.net</li> <li>zenonmanero@gmail.com</li> <li>lorisalmone@hotmail.com</li> <li>francarvajal66@gmail.com</li> <li>null</li> <li>null</li> <li>info@taxiboatcanarias.com</li> <li>null</li> </ul>	 <p>25 Categories</p> <ul style="list-style-type: none"> <li>null</li> <li>null</li> <li>null</li> <li>-</li> <li>null</li> <li>null</li> <li>-</li> <li>Faxes: 928 15</li> <li>Faxes: 928 27</li> <li>Faxes: 928 24</li> <li>Páginas web: 4</li> <li>Páginas web: 1</li> <li>Faxes: 928 24</li> <li>-</li> <li>Páginas web: 1</li> <li>-</li> <li>null</li> <li>null</li> <li>-</li> <li>null</li> </ul>	<p><b>ABC Teléfonos/Internet2</b></p> <p><code>{upper}{lower}{4}: {digit}{3} {digit}{3} {digit}{3}</code></p> <p><code>{upper}{lower}{6}: {email},{email}</code></p> <p><a href="#">Show pattern details...</a></p> <p><b>Suggestions</b></p> <p>Split on values matching</p> <p>..</p> <p>..</p> <p><b>Rename</b></p> <p>Rename Teléfonos/Internet2 to <b>TeléfonosVInternet2</b></p> <p><b>Delete columns</b></p> <p>Teléfonos/Internet2</p> <p><a href="#">Edit</a> <a href="#">Add</a></p>

Transformamos el campo Teléfono/Internet3, eliminando el Texto “Faxes:”



Source	to be dropped	Preview			
RBC	Teléfonos/Internet3	RBC	Teléfonos/Internet3	RBC	Teléfonos/Internet4

Extract values matching

See all

Faxes: `

(alpha)+`

Faxes: ` starting after `(start)` ending before `928`

Replace

Faxes: ` with ` in Teléfonos/Internet3

Edit Add

first occurrence of `(start) (alpha)+` with ` in Teléfonos/Internet3

Y ahora quitamos espacios en blanco entre los números.

Source	to be dropped	Preview			
#	Teléfonos/Internet3	#	Teléfonos/Internet3	RBC	Teléfonos/Internet4

Replace

(delim) ` with ` in Teléfonos/Internet3

Edit Add

` with ` in Teléfonos/Internet3

Split on values matching

See all

` 2 times

(delim) `

`

Todos los números de fax, que contienen registros inválidos los reemplazamos por valores nulos.

Source	to be dropped	Preview			
#	Teléfonos/Internet3	#	Teléfonos/Internet3	RBC	Teléfonos/Internet4

Delete rows

with mismatched values in Teléfonos/Internet3

Keep rows

with mismatched values in Teléfonos/Internet3

Create a new column

flag mismatched values in Teléfonos/Internet3

Set

mismatched values to NULL()

Edit Add

mismatched values to 0

Cambiamos el nombre de la etiqueta como último paso de la transformación de la columna.

Preview					
#	Fax	RBC	Teléfonos/Internet4	RBC	Puerto Base

Option

required

Manual rename

Specify the new name for each column

Columns (1)

Add

Teléfonos/Internet3

Fax

Cancel Add

La transformación para el último campo de contacto, es obtener la columna con la información de la página web.



Source: ABC | Preview: ABC | Puerto Base: ABC

9 Categories | 1 Category | 15 Categories | 77 Categories

Extract values matching: `'(url)'`

Replace: first occurrence of `'(start){url}'` with `'` in Teléfonos/Internet4

y cambiamos el nombre de la etiqueta de la columna por web.

Preview: web | Puerto Base: ABC

Option: Manual rename

Specify the new name for each column

Columns (1): Teléfonos/Internet1

web

Eliminamos la columna Teléfonos/Internet4.

Por último, realizamos las transformaciones necesarias sobre el titular de la empresa.

Vamos a obtener CIF/NIF de la columna Titular

Source: ABC | Preview: ABC | Titular1: # | Capacidad máxima: 0 - 205

77 Categories | 77 Categories

Extract values matching: `'(alpha-numeric)+(alpha-numeric)+'`

Replace: `'(alpha-numeric)+(alpha-numeric)+'` with `'` in Titular

Count values matching: `'(alpha-numeric)+(alpha-numeric)+'`

Después de obtener el CIF/NIF, cambiamos el nombre de la etiqueta de la columna.

Preview: CIF/NIF | # | Capacidad máxima: 0 - 205

Option: Manual rename

Specify the new name for each column

Columns (1): Titular1

CIF/NIF



N.I.F. 33478426V	33478426V	FRANCISCO JAVIER TORRES GOMARIZ
B35522788	B35522788	DOLPHIN AND WHALES, S.L.
B35838937	B35838937	MULTIACUATIC, S.L.
B35125335	B35125335	CANARIAS YACHTS, S.L.
B35422104	B35422104	EXCURSIONES SUPERCAT S.L.
B35512508	B35512508	BAHIA CAT, S.L.
B35125335	B35125335	CANARIAS YACHTS, S.L.
N.I.F. 01621959E	01621959E	Fernando Zenón Manero Rodriguez
B76091792	B76091792	LINEAS SALMÓN, S.L.
B76118850	B76118850	AVENTURA SAGITARIUS STAR SL
N.I.F. 01621959E	01621959E	Fernando Zenón Manero Rodriguez
N.I.F. 43284887Z	43284887Z	Manuel Quesada Calero

'N.I.F.' with '' in Titular1

Edit Add

Count values matching See all

'N.I.F.'

'N.I.F.' after '{start}' before '33478426'

'N.I.F.' after '{start}' before '{digit}{8}'

Una vez obtenido el CIF/NIF, solo nos queda obtener el Nombre de la Razón Social del Titular, para ello utilizaremos una división de campos (Split) utilizando como carácter separador el carácter '-'.

Source	to be dropped	Preview
RBC	Titular	RBC Titular1 RBC
77 Categories		77 Categories 77 Categories
C.I.F. B35954130 FUERTEVENTURA SPORT FISHIG, S.L.		C.I.F. B35954130 FUERTEVENTURA SPORT
C.I.F. B35767375 Perdomo Santana, S.L.		C.I.F. B35767375 Perdomo Santana, S.L.
C.I.F. B76140128 INSTITUTO CANARIO DE ESTUDIOS DE LA NATURALEZA Y APLICACIONES INDUSTRIALES, S.L.		C.I.F. B76140128 INSTITUTO CANARIO D

Split on values matching See all

'(delim)' starting after '{digit}(8)' ending before '{upper}(13)'

'(delim)' starting after '{digit}(8)' ending before '{upper}+'

Como último paso eliminamos el campo Titular1, porque no la necesitamos.

Drop Column	RBC Titular1	RBC Titular2
77 Categories		77 Categories
C.I.F. B35954130 FUERTEVENTURA SPORT FISHIG, S.L.		FUERTEVENTURA SPORT FISHIG, S.L.
C.I.F. B35767375 Perdomo Santana, S.L.		Perdomo Santana, S.L.
C.I.F. B76140128 INSTITUTO CANARIO DE ESTUDIOS DE LA NATURALEZA Y APLICACIONES INDUST		INSTITUTO CANARIO DE ESTUDIOS DE LA NATURALEZA Y APLICACIONES INDUST
C.I.F. B35917657 ESCUELA NAUTICA MORRO JABLE, S.L.		ESCUELA NAUTICA MORRO JABLE, S.L.
C.I.F. B48403323 LAMBOK, S.L.		LAMBOK, S.L.
N.I.F. 33478426V FRANCISCO JAVIER TORRES GOMARIZ		FRANCISCO JAVIER TORRES GOMARIZ
C.I.F. B35522788 DOLPHIN AND WHALES, S.L.		DOLPHIN AND WHALES, S.L.
C.I.F. B35838937 MULTIACUATIC, S.L.		MULTIACUATIC, S.L.
C.I.F. B35125335 CANARIAS YACHTS, S.L.		CANARIAS YACHTS, S.L.
C.I.F. B35422104 EXCURSIONES SUPERCAT S.L.		EXCURSIONES SUPERCAT S.L.
C.I.F. B35512508 BAHIA CAT, S.L.		BAHIA CAT, S.L.
C.I.F. B35125335 CANARIAS YACHTS, S.L.		CANARIAS YACHTS, S.L.
N.I.F. 01621959E Fernando Zenón Manero Rodriguez		Fernando Zenón Manero Rodriguez
C.I.F. B76091792 LINEAS SALMÓN, S.L.		LINEAS SALMÓN, S.L.
C.I.F. B76118850 AVENTURA SAGITARIUS STAR SL		AVENTURA SAGITARIUS STAR SL
N.I.F. 01621959E Fernando Zenón Manero Rodriguez		Fernando Zenón Manero Rodriguez
N.I.F. 43284887Z Manuel Quesada Calero		Manuel Quesada Calero
C.I.F. B76166230 NAUTICS TRAVEL CANARIAS, S.L.		NAUTICS TRAVEL CANARIAS, S.L.
C.I.F. B04761763 ALANDALUS PARTNERS, S.L.		ALANDALUS PARTNERS, S.L.
C.I.F. B76248467 OCEAN CAT SLU		OCEAN CAT SLU
N.I.F. 43616967C MIGUEL BELLO MESA JOSE		MIGUEL BELLO MESA JOSE

RBC Titular1

{upper}. {upper}. {upper}. {hex} 83

{upper}. {upper}. {upper} {digit}{8} {upper} 14

{upper}. {upper}. {upper} {hex} 7

{upper}. {upper}. {upper} {upper} {digit}{7} {upper} 6

Show pattern details...

Suggestions

Split on values matching

' ' 3 times

' '

' '

Rename

Rename Titular1 to 'Titular1'

Delete columns

Titular1

Edit Add

Y renombramos etiqueta de columna Titular2 por Titular.

Preview	RBC CIF/NIF	#	Capacidad máxima
77 Categories		0 - 205	
FUERTEVENTURA SPORT FISHIG, S.L.	B35954130		
Perdomo Santana, S.L.	B35767375		
INSTITUTO CANARIO DE ESTUDIOS DE LA NATURALEZA Y APLICACIONES INDUSTRIALES, S.L.	B76140128		
ESCUELA NAUTICA MORRO JABLE, S.L.	B35917657		
LAMBOK, S.L.	B48403323		
FRANCISCO JAVIER TORRES GOMARIZ	33478426V		
DOLPHIN AND WHALES, S.L.	B35522788		
MULTIACUATIC, S.L.	B35838937		
CANARIAS YACHTS, S.L.	B35125335		

Option required

Manual rename

Specify the new name for each column

Columns (1)

Titular2

Titular1

Cancel Add

Por último, ejecutamos el trabajo para aplicar todos los pasos de la transformación al dataset original que hemos importado.



Run Job

Options

☒ Profile Results When enabled, this will generate a profile of your results.

Publishing Actions

Create CSV

Create JSON

ACTIONS

SETTINGS

Create-CSV

no compression, single file, with headers, with quotes, with delimiter: ,

X

Cancel

Run Job

Al finalizar la ejecución del trabajo, podremos descargar el csv con el resultado de aplicar todas las transformaciones que hemos realizado.

Cetaceos > observacion-de-cetaceos-autorizados.xls/Relación por Isla y Municipio - 2

Job 42637

Finished Today at 2:20 PM

Download results

...

Overview

Output Destinations

Profile

Dependencies

Name	Status	Duration
observacion-de-cetaceos-autorizados.xls/Relación por Isla y Municipio - 2.csv	Completed	<div>View details</div> <div>...</div>

Download result

Create imported dataset