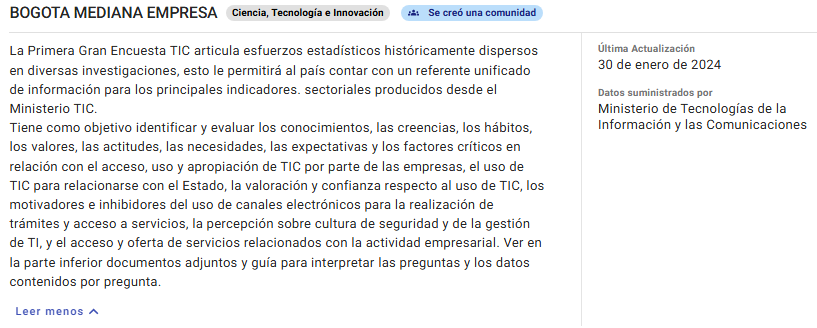
# DATOS ESCOGIDOS

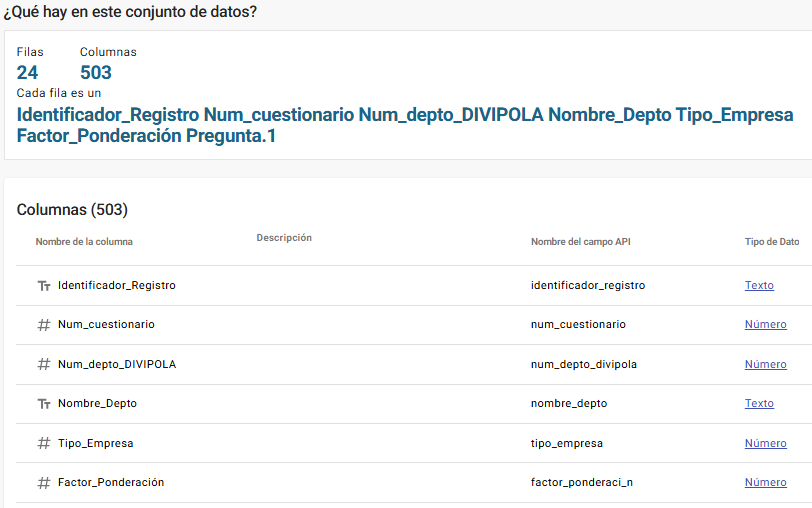
<https://www.datos.gov.co/browse?q=bogot%C3%A1%20mediana%20empresa%20&sortBy=relevance>

## 1.1. DESCRIPCIÓN









## 1.2. VIDEO DE VISUALIZACIÓN DE LOS DATOS EN LA PLATAFORMA WEB.

<https://drive.google.com/file/d/1-esNQ0yy-WG1xmxofasd005LtAZ15nJV/view?usp=drive_link>

## 1.3. ELEMENTOS ADJUNTOS:

1.3.1. DICCIONARIO\_EMPRESAS\_ENCUESTA\_TIC\_2017

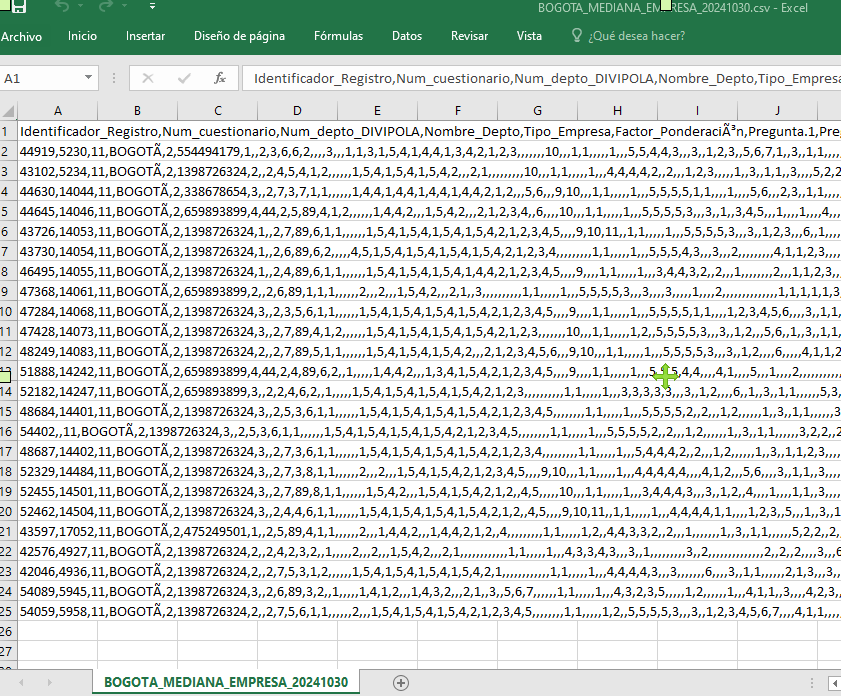
<https://drive.google.com/file/d/1AZwgEO1YYYeyRuGCjdFxahJf9n2tRXFF/view?usp=drive_link>

1.3.2. Instrumento\_empresas\_VF.PDF

<https://drive.google.com/file/d/1AZwgEO1YYYeyRuGCjdFxahJf9n2tRXFF/view?usp=drive_link>

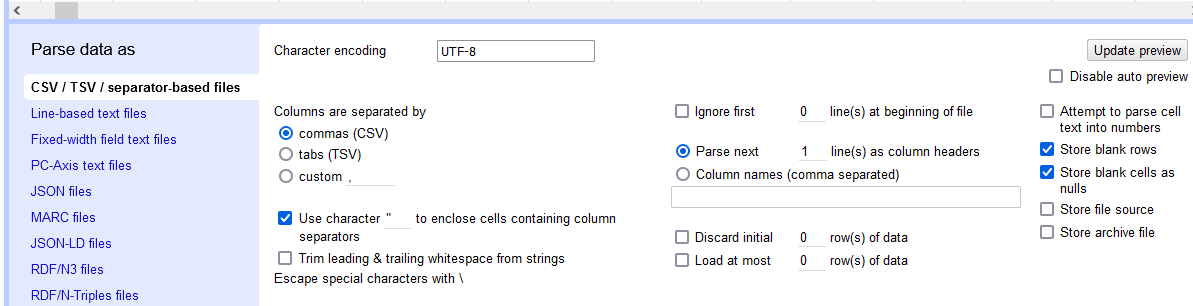
# 2. TRABAJANDO LOS DATOS

## 2.1. ASPECTO INICIAL DEL ARCHIVO.

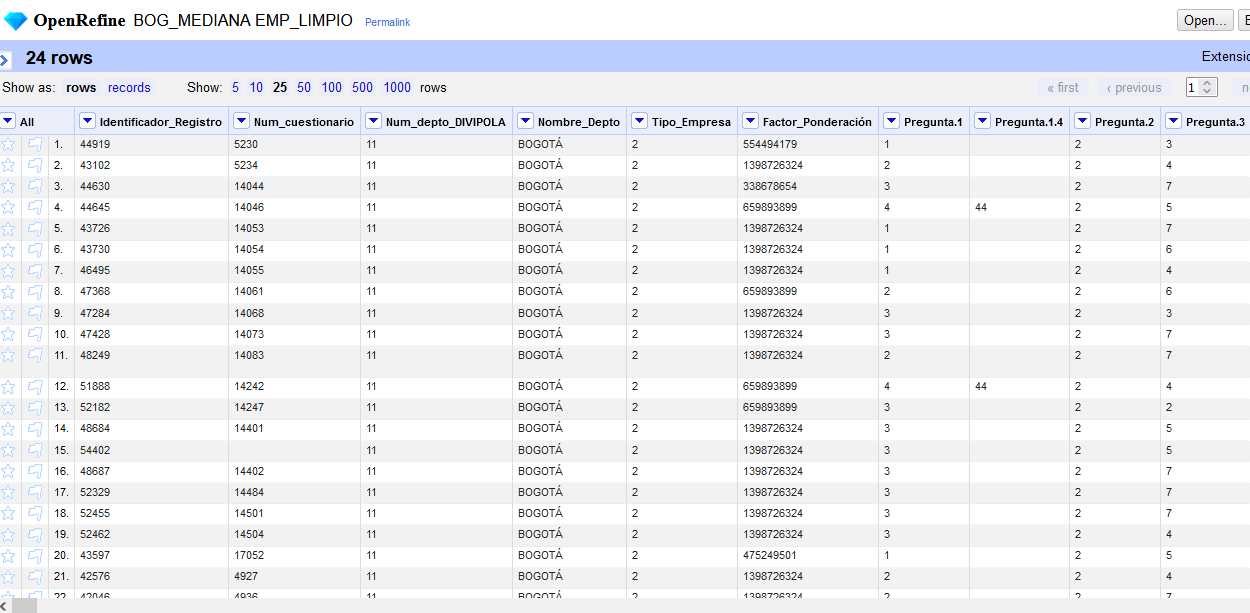


## 2.2. LIMPIEZA DE DATOS USANDO OpenRefine.

### 2.2.1. Configuración inicial del proyecto en OpenRefine software.



### 2.2.2. Pantallazo del proyecto recién creado.



### 2.2.3. Registro de modificaciones hechas o no al conjunto de datos.

#### 2.2.3.1. Columnas eliminadas:

A continuación, se registran las columnas eliminadas y el motivo:

* **“Identificador\_Registro”:** Números de cinco dígitos, no repetido en ninguna de las filas y que quizá identifique a quien realizó la encuesta. En todo caso, al no hacer parte del conjunto de datos que conforman las respuestas a la encuesta, no parecen aportar información relevante.
* **“num\_cuestionario”:** Números de cinco dígitos, no repetido en ninguna de las filas y que seguro identifican el número de cuestionario de cada encuesta. En todo caso, al no hacer parte del conjunto de datos que conforman las respuestas a la encuesta, no parecen aportar información relevante.
* **“Num\_depto\_DIVIPOLA”:** El número 11 aparece en todas las filas. Este número es el asignado a la ciudad de Bogotá DC, para referirse a ella. Además de la redundancia; la columna se elimina porque desde el comienzo se sabía que todo el conjunto de datos se refiere a esta ciudad.
* Nombre\_Depto
* **“Tipo\_Empresa”**: El número 2 aparece en todas las filas. Este número es el asignado para indicar las medianas empresas. Además de la redundancia; la columna se elimina porque desde el comienzo se sabía que todo el conjunto de datos se refiere a medianas empresas.

#### 2.2.3.2. Columnas NO eliminadas:

A continuación, se registran las columnas NO eliminadas y el motivo:

**\* “Factor\_Ponderación”:** Números de 9 o 10 dígitos que si se repiten en alguna(s) otras filas. No sabemos a qué ponderación hagan referencia, pero quizá sea relevante. Además, que con que se esté repitiendo en varios registros; ya se pueden sacar porcentajes y proporciones de esta columna (dato).

\* Todas las columnas destinadas a espacios para las respuestas de la encuesta se conservaron; aunque estuvieran en blanco; ya que, de algún modo en este conjunto de datos, la ausencia de registros en los campos también brinda información (Opciones de respuesta a las que nadie opto).

# 3. TRABAJANDO CON EL ARCHIVO ADJUNTO A LOS DATOS (DICCIONARIO\_EMPRESAS\_ENCUESTA\_TIC\_23017.xlsx).

## 3.1. Consulta Web:

Además de mi archivo con datos; tengo un archivo en excel que es el diccionario para entender mis columnas y sus datos. Indícame cómo proceder con este diccionario y si acaso puedo usar esa información de excel para hacer algún tipo de automatización en Python para lectura e interpretación.

## Copilot

Vamos a sacarle el máximo provecho a tu diccionario de datos en Excel. Aquí tienes los pasos a seguir y cómo usar esa información para automatizar la lectura e interpretación en Python.

### 1. Revisar y Entender el Diccionario de Datos

* Abre tu archivo de diccionario en Excel.
* Asegúrate de que las columnas estén claramente etiquetadas, indicando el nombre de la columna en tu dataset principal, el tipo de datos, y quizás una breve descripción de cada campo.

### 2. Importar el Diccionario de Datos a Python

Puedes usar la librería pandas para leer el archivo Excel en Python:

python

Copiar

import pandas as pd

# Lee el archivo Excel que contiene tu diccionario de datos

diccionario = pd.read\_excel('ruta\_a\_tu\_diccionario.xlsx')

print(diccionario)

### 3. Automatizar la Interpretación de Datos

Basándote en el diccionario, puedes automatizar la interpretación de datos para asegurar consistencia y precisión. Aquí tienes un ejemplo:

python

Copiar

# Leer tu archivo principal de datos

datos = pd.read\_csv('ruta\_a\_tu\_archivo\_de\_datos.csv')

# Recorre cada columna en el diccionario y aplica las transformaciones necesarias

for index, row in diccionario.iterrows():

columna = row['Nombre de la Columna']

tipo\_dato = row['Tipo de Dato']

if tipo\_dato == 'numérico':

datos[columna] = pd.to\_numeric(datos[columna], errors='coerce')

elif tipo\_dato == 'fecha':

datos[columna] = pd.to\_datetime(datos[columna], errors='coerce')

# Puedes añadir más condiciones según sea necesario

print(datos.info())

### 4. Validación y Limpieza de Datos

Asegúrate de validar y limpiar los datos usando la información del diccionario:

python

Copiar

# Identificar y manejar datos faltantes

datos = datos.dropna() # O utiliza otras técnicas de imputación

# Verificar la coherencia de los datos según el diccionario

# Por ejemplo, asegurarte de que las fechas estén dentro de un rango válido

# Resumir los datos limpiados

print(datos.describe())

Este enfoque te permite utilizar tu diccionario de datos para automatizar y estandarizar la lectura y limpieza de datos, garantizando que toda la información sea interpretada de acuerdo a las definiciones y tipos especificados en tu diccionario.

¿Qué opinas? ¿Hay algún otro detalle específico en el que te gustaría profundizar? 😊

## 3.1. Creación del proyecto y pegar código:

