# Guía práctica para el desarrollo de bases de datos socioeconómicas con Python y Excel


## Autores:

**Joshua G. Aranda Martínez**
**Melisa Roa Rubio**
*Facultad de Economía. UNAM.*
**Karol Granados Martínez**
*Laboratorio Nacional de Resiliencia Costera. UNAM.*
**Citlalin Martínez Córdova**
**Sophie Ávila Foucat**
*Instituto de Investigaciones Económicas. UNAM.*



2023

INSTITUTO DE INGENIERÍA UNAM

Laboratorio Nacional de Resiliencia Costera

## Responsables del programa:

**Sophie Ávila Foucat.** *Instituto de Investigaciones Económicas. UNAM*
**Alec Torres Freyermuth.** *Instituto de Ingeniería. UNAM*
**Fernando Nuno Dias Marques Simões.** *Facultad de Ciencias. UMDI-Sisal. UNAM*

## Autores:

**Joshua G. Aranda Martínez**
**Melisa Roa Rubio**
*Facultad de Economía. UNAM.*
**Karol P. Granados Martínez**
*Laboratorio Nacional de Resiliencia Costera. UNAM.*
**Citlalin Martínez Córdova**
**Sophie Ávila Foucat**
*Instituto de Investigaciones Económicas. UNAM.*

## Diseño:

Karol Granados Martínez

## Imágenes:

Joshua G. Aranda Martínez
Melisa Roa Rubio

## Cita:

Aranda-Martínez, J., Roa-Rubio, M., Granados-Martínez K.P., Martínez Córdova, C., Ávila-Foucat, S.V. (2023) Guía práctica para el desarrollo de bases de datos socioeconómicas con Python y Excel. [PDF]

## Licencia:

Guía práctica para el desarrollo de bases de datos socioeconómicos con Python y Excel, Joshua G. Aranda Martínez, Melisa Roa Rubio, Karol Granados Martínez, Citlalin Martínez Córdova y Sophie Ávila Foucat, D.R.,©, 2023,UNAM -CC -BY-NC.

Este documento se realizó con el apoyo del Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) 1V3000123: Resiliencia de socioecosistemas costeros asociados al turismo ante COVID 19 y cambio climático

# Índice general

Guía de uso ..... 3
Introducción ..... 4
1. Las fuentes de datos ..... 5
1.1 Definición del problema ..... 5
1.2 La escala del problema ..... 5
1.3 Definición de variables ..... 6
1.4 Repositorios de datos ..... 6
2. Breve introducción a Python 3 ..... 11
2.1 ¿Qué es Python 3? ..... 11
2.2 ¿Cómo podemos emplear Python 3? ..... 11
2.3 Los scripts de Python ..... 12
3. Extracción y preparación de los datos ..... 13
3.1 Preparación de los datos ..... 13
3.1.1 Los datos con acentos ..... 13
3.1.2 Diseño de la base de datos ..... 15
3.2 Durante la extracción de datos ..... 16
3.2.1 Selección y filtrado de datos ..... 16
3.2.2 Manejo de NaN ..... 20
4. Gestión y manejo de datos ..... 23
4.1 Información disponible vs información solicitada ..... 23
4.2 Creación del metadato ..... 24
4.2.1 Pasos para la creación de un metadato. ..... 24
5. Ejemplo de obtención de datos con Excel ..... 26
Referencias ..... 34

# Guía de uso

La presente guía tiene como objetivo ilustrar el proceso que se llevó a cabo durante la recopilación de información socioeconómica para el proyecto "*Resiliencia de socioecosistemas costeros asociados al turismo ante COVID 19 y cambio climático*" , con el fin de que pueda resultar útil para la realización de actividades análogas en futuros proyectos de investigación.

Producto de la experiencia adquirida a lo largo del proyecto, buscamos presentar algunos de los métodos que nos parecieron más adecuados para recopilar, ordenar y limpiar la información solicitada de manera eficiente, segura y oportuna. El lenguaje de programación Python 3, es una de las herramientas que se pueden utilizar durante el proceso de preparación y limpieza de datos de manera más rápida y efectiva, al reducir el potencial de error humano y permitirnos usar conjuntos de datos mucho más grandes de los que podríamos manipular de forma más rudimentaria.

Asimismo, trataremos de señalar algunas de las particularidades que a menudo presentan las fuentes de información en las que nos basamos para construir las bases de datos; con algunas advertencias sobre su uso que pueden resultar útiles para quien aún no se haya familiarizado con ellas. En última instancia, consideramos que es de utilidad tener una introducción a la búsqueda de datos para la investigación en áreas interdisciplinarias y de las ciencias sociales, ya que contar con datos de gran escala resulta esencial para llevar a cabo cualquier tipo de análisis.

# Introducción

Para lograr una comprensión completa de la capacidad de resiliencia de los ecosistemas costeros, es necesario abordar enfoques transdisciplinarios[1] capaces de entender las dinámicas sociales, económicas y ambientales de estos desafíos. Para ello, es crucial contar con una amplia gama de datos e información, que permitan la creación de repositorios de información y que funcionan como un depósito centralizado de datos.

El propósito fundamental de este repositorio en el marco del proyecto es facilitar el acceso a la información y promover la colaboración entre diferentes entidades y personas interesadas. Sin embargo, en esta guía enfocaremos la atención a la obtención de información socioeconómica relevante proveniente de diferentes fuentes para siete sitios costeros de estudio (Agiabampo, Arrecife Alacranes, Celestún, Copalita-Huatulco, Dos Bocas, Laguna de Términos y Sisal), con el objetivo de dar este primer paso en la creación de bases de datos. El documento se dividirá en cuatro capítulos principales y un anexo, cuya secuencia sigue la del proceso seguido para la recopilación de dichos datos:

1. **Fuentes de datos**
   En este capítulo haremos una reseña de las principales fuentes de datos y estructura de sus variables.

2. **Introducción a Python**
   En este capítulo se dará una breve descripción de lo que es y para que se puede utilizar el lenguaje de Python para el desarrollo de bases de datos.

3. **Preparación y obtención de datos**
   En este capítulo exploraremos la parte más técnica del proceso, que incluirá la extracción y el ordenamiento de los datos según accesibilidad de datos

4. **Gestión y manejo de datos**
   En este último capítulo se muestran los posibles pasos a seguir una vez que se han construido las bases de datos.

5. **Ejemplo d obtención de datos en Excel**
   Se presentarán ejemplos de los formatos empleados y algunos métodos para manipular los datos de manera segura.

[1] La transdisciplina es un proceso para la producción de conocimiento que involucra diferentes perspectivas desde diferentes disciplinas y de manera conjunta donde se construyen teorías y metodologías a partir de esa colaboración, con el objetivo de atender problemas complejos (Klein, 2008).

# 1. Las fuentes de datos

## 1.1 Definición del problema

Cualquier investigación que se realiza debe partir de una o más preguntas cuya potencial respuesta será planteada a manera de hipótesis, misma que será contrastada con las evidencias que sean recolectadas a lo largo de la investigación.

Es importante recordar que partimos de la premisa de que es necesario definir preguntas de investigación e hipótesis antes de iniciar el proceso de recopilación de datos. Lo anterior constituye la etapa previa e indispensable para el inicio de la recolección de datos.

## 1.2 La escala del problema

Cuando se habla de la realización de proyectos de investigación, en especial aquellos relacionados con factores de nivel macro[2] —para el proyecto de investigación en el que nos basamos para la realización del presente texto se pretende hacer un análisis a nivel socioecosistema—, la mayoría de las fuentes de datos son de instituciones públicas. Ya sea que se trate de una fuente como el Instituto Nacional de Estadística y Geografía (INEGI), cuya principal ocupación es la recopilación de dicha información, o que se busque información más especializada que puede ser recolectada y publicada por otros organismos de carácter público.

Otras fuentes de información relevante son los datos recolectados por el propio grupo de investigación, aunque por supuesto esto dependerá de la escala del proyecto y de los recursos con los que se cuente. Asimismo, el proceso de diseño de un proyecto de análisis en el que se vaya a realizar la recolección primaria de datos escapa al propósito del presente manual. *En áreas de investigación como las ciencias sociales o colaboraciones interdisciplinarias, la mayoría de las veces se hace uso de datos previamente recolectados por agencias gubernamentales a una escala mayor a la que podríamos, de manera realista, obtener por nuestros propios medios.*[3]

---

[2] En contraposición con los análisis de escala micro, que pueden referirse a un individuo, un hogar o una empresa, en la que nuestra información probablemente tenga que ser recolectada como parte del estudio pues es casi imposible que existan fuentes previas sobre dichos temas.

[3] Por supuesto siempre habrá excepciones a la regla, como los estudios de pequeña escala que han realizado algunos economistas conductuales, pero, en lo general, dado que este tipo de proyectos de investigación suelen tener un enfoque de mayor escala, se torna imposible la recolección interna de los datos. Siendo así, más que el enfoque disciplinario que se emplee, el uso de fuentes tendrá más que ver con la escala del problema a analizar.

## 1.3 Definición de variables

Teniendo claridad sobre nuestro problema y la escala en la que se requiere, es posible trazar las líneas generales de un conjunto de variables que permita realizar nuestro estudio y resolver las preguntas planteadas. Por supuesto, no hay garantía de que podamos hallar los datos correspondientes a cada variable que resulte relevante a la investigación. Este es un problema con el que nos enfrentaremos a menudo y al que será necesario adaptarse.

Cuando una variable no se encuentra como tal, se puede estimar con base en otras variables que se encuentran disponibles, o que pueda ser reemplazada por otra que cumpla un papel similar. Es importante asegurarse de haber realizado una búsqueda exhaustiva de variables en las fuentes que puedan resultar relevantes. Por ejemplo, una variable que no reporte el INEGI quizá sea reportada por el Consejo Nacional de Población (CONAPO) o viceversa.

## 1.4 Repositorios de datos

Las principales fuentes de datos y organismos que reportan información socioeconómica y demográfica utilizadas en la realización del este proyecto son las siguientes:

- Instituto Nacional de Estadística y Geografía (INEGI)

El **INEGI** es un órgano constitucional autónomo cuya tarea implica la recopilación de información sobre cuestiones socioeconómicas, así como información geográfica, que también puede resultar valiosa. Debido a ello, es a menudo la principal fuente de la que podemos obtener información si nuestra investigación tiene un carácter económico, demográfico o interdisciplinario en el que se requiera información que tenga que ver con censos o estimaciones de ese tipo de variables.

Es muy probable que cualquier investigador (y quienes le apoyen) conozca el INEGI y esté familiarizado con el tipo de datos que es posible hallar en su sitio. El principal obstáculo que llega a presentarse cuando se quiere realizar una consulta en el INEGI es que su sitio web está en constante remodelación. Lo que implica que a menudo se tendrá que seguir un camino ligeramente distinto para llegar al mismo conjunto de datos según la fecha en que se consulte la información. Aunque la forma de acceder suele ser más o menos intuitiva en todas las versiones del sitio, a veces la forma más sencilla es usar la función de “Chat uno a uno” que provee el propio INEGI. En donde los miembros del equipo pueden orientar al usuario al conjunto de datos correcto si es que éste está teniendo problema para encontrarlos dentro del sitio.

A continuación, se muestra un sintetizado del tipo de información que contiene los Censos de Población y Vivienda (INEGI, 2020), ya que proporcionan datos de:

- Población: Proporciona información sobre la cantidad de habitantes en cada entidad federativa, municipio y localidad del país, desglosada por sexo, edad, estado civil, educación, ocupación y otros indicadores relevantes.

- Características socioeconómicas: Ofrece datos sobre la situación laboral, ingresos, educación, salud y otros aspectos socioeconómicos de la población mexicana.

- Viviendas: Incluye información acerca del tipo de vivienda, servicios básicos, materiales de construcción, tenencia de la vivienda y otras características relacionadas con el hábitat de la población.

- Migración: Permite analizar los flujos migratorios y conocer el lugar de residencia anterior y posterior de las personas, así como las razones de migración.

- Población indígena: Proporciona datos específicos sobre la población indígena en México, como su distribución geográfica, idiomas hablados y otras características culturales.

La Encuesta Nacional de Ocupación y Empleo (ENOE) de INEGI, 2023, recopila una variedad de información relacionada con la ocupación y el empleo en México. Ofrece información acerca de:

- Empleo: Proporciona información sobre la población económicamente activa, la ocupación, el sector económico, el tipo de empleo (formal o informal), las horas trabajadas y la duración del empleo.

- Desempleo: Ofrece datos sobre la población desocupada, la tasa de desempleo, las razones de desempleo y la duración del desempleo.

- Características laborales: Incluye información sobre el salario y los ingresos laborales, la calidad del empleo, la seguridad social, los beneficios laborales y la sindicalización.

- Informalidad: Permite analizar la proporción de empleo informal en la economía, así como sus características y condiciones laborales.

- Educación y formación: Proporciona datos sobre el nivel educativo alcanzado, la capacitación laboral y las certificaciones profesionales.

- ### Consejo Nacional de Población (CONAPO)

El **CONAPO** es un órgano gubernamental cuya tarea tiene que ver con la alineación de las políticas de gobierno con la planeación demográfica, por lo que reporta algunas variables relacionadas con estas tareas.

Si bien CONAPO no se enfoca en el reporte de datos como sí lo hace el INEGI, puede ocurrir que en una investigación que involucre factores demográficos se tenga que recurrir a datos provistos por el CONAPO. Como mucha de la información provista por organismos gubernamentales, pueden consultarse varios de sus datos desde el **portal de datos abiertos del gobierno de México**.

Entre la información que proporciona CONAPO se encuentra:

- El Índice de Marginación por municipio (1990-2015) (CONAPO, 2020a) permite evaluar la situación socioeconómica de los municipios en diferentes momentos a lo largo de un periodo de 25 años con información cada cinco años.

- El Índice de Marginación por localidad (1990-2010) (CONAPO, 2020a) se enfoca en proporcionar información más detallada a nivel de localidad sobre la situación socioeconómica en un periodo de 20 años con información cada cinco años.

Ambos índices se basan en la metodología desarrollada por la CONAPO, que considera indicadores como la educación, el acceso a servicios básicos, el empleo, la vivienda, la salud, entre otros, para calcular el grado de marginación.

- ### Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL)

EL **CONEVAL** es un organismo público descentralizado de nivel federal cuya principal tarea es la evaluación de la política social y la medición de la pobreza en el país. En consecuencia, la información que genera puede resultar de enorme importancia para cualquier investigación que trate temas sociales, de pobreza o vulnerabilidad. La información incluida en la base de datos del CONEVAL comprende diversos aspectos, como:

- Indicadores de pobreza: Incluye mediciones de la pobreza en términos de ingreso, carencias sociales y otros aspectos relacionados con las condiciones de vida de la población.
- Variables socioeconómicas: Datos sobre características demográficas, nivel educativo, acceso a servicios básicos, empleo, salud, vivienda, entre otros factores que influyen en la situación de pobreza.

- Desagregación geográfica: Los datos están desglosados por estados y municipios.

- Servicio de Información Agroalimentaria y Pesquera (SIAP)

El **SIAP** es un órgano administrativo desconcentrado que forma parte de la Secretaría de Agricultura y Desarrollo Rural (SADER). Su objetivo es generar y publicar información estadística y geográfica en materia agroalimentaria, además de participar en la coordinación en dicha materia entre diversos niveles de gobierno.

El SIAP contiene un repositorio importante y altamente detallado sobre actividades agroalimentarias, sean de carácter agrícola, ganadero, pesquero o derivado de algún producto primario. En lo general, si se requieren datos sobre el sector primario, el SIAP podría ser un repositorio extremadamente útil para consultar. Entre la información que provee se encuentran las siguientes variables:

**Variables de actividad agrícola**

- Superficie sembrada: Se refiere a la superficie, medida en hectáreas, que fue sembrada en el año de la medición.

- Superficie cosechada: Se refiere a la superficie medida en hectáreas que fue cosechada en el año de la medición, y permite estimar si hubo pérdidas en términos de la superficie.

- Valor de la producción: Se presenta en miles de pesos y representa el valor total de la producción agrícola del municipio dado.

**Variables de actividad pecuaria**

- Producción: Se presenta en toneladas y está desgranada a nivel municipal y separada por tipo de producto.

- Precio: Se presenta en pesos por kilogramo, y también se separa por tipo de producto y por municipio.

- Valor de la producción: Se refiere al valor total de la producción del tipo de producto pecuario dado, se presenta en miles de pesos.

- Secretaría de Turismo (SECTUR)

La **SECTUR** es una de las principales secretarías del nivel federal, cuya tarea es la de promover y regular el turismo. Si bien la recolección de información no es su principal tarea, sí publica algunos datos a los que no se puede acceder en otros repositorios. Si bien la mayoría de sus datos se encuentran en el ya mencionado portal de datos abiertos del gobierno federal, a veces puede resultar útil acceder al sitio web de la secretaría y explorar las ligas que allí aparecen.

- Banco Mundial (BM)

El **BM** es una institución financiera internacional que ofrece financiamiento a naciones en vías de desarrollo para promover el crecimiento y el desarrollo en los mismos. Como parte de sus tareas, reporta un conjunto bastante amplio de datos económicos y puede representar una alternativa útil a la recopilación de datos en organismos gubernamentales, si bien la mayoría de sus estadísticas se basan en lo que reporta cada país de manera individual.

- Comisión Económica para América Latina y el Caribe (CEPAL)

La **CEPAL** es una institución dependiente de la Organización de Naciones Unidas (ONU) cuyo objetivo es promover el desarrollo económico y la cooperación entre los países de la región y entre éstos y el resto del mundo. Como parte de dicha misión reporta un conjunto amplio de datos, de nuevo primando aquellos que tienen que ver con temas socioeconómicos y de desarrollo.

**Explora cada una de las fuentes de datos que se presentaron anteriormente. Tan solo debes dar clic en los nombres de las fuentes de datos en color MORADO**

# 2. Breve introducción a Python 3

En la introducción mencionamos que, a lo largo de esta guía, emplearíamos el lenguaje de programación Python 3 para llevar a cabo algunos procesos que nos ayudarían en la preparación y limpieza de los datos que hayamos encontrado durante nuestra búsqueda. Ha llegado el momento de explicar con algo más de detalle qué es Python 3 y por qué hemos decidido incluir ejemplos de su uso en esta guía.

## 2.1 ¿Qué es Python 3?

Python es un lenguaje de programación de código abierto desarrollado por Guido Van Rossum y publicado en su primera versión en 1994.[4] Es un lenguaje extremadamente versátil y veloz, que además resulta amigable con los nuevos usuarios debido a su sintaxis intuitiva. Debido a todo ello, su popularidad ha crecido considerablemente desde su lanzamiento, para el análisis y procesamiento de datos aplicados en diferentes áreas de investigación de la ciencia.

## 2.2 ¿Cómo podemos emplear Python 3?

Para usar Python 3 es necesario obtener el código desde **el sitio oficial de Python**[5] (lo más recomendable), o desde otro lugar que también lo ofrezca. Habiendo realizado la descarga de la última versión estable, y que dependerá de nuestro sistema operativo (se ofrecen versiones para Windows, Mac OS, Linux/UNIX y otros), debemos de proceder a su instalación. Esta se llevará a cabo como la de cualquier otro archivo ejecutable.

Al finalizar, podremos acceder a Python desde la consola de nuestro sistema operativo, aunque usualmente resulta más sencilla la utilización de este lenguaje mediante el uso de un tipo de software conocido como Entorno de Desarrollo Integrado —abreviado por sus siglas en inglés como IDE—, mismo que ofrece ciertas funciones que facilitan el uso de Python y lo hacen aún más amigable a quienes no estén familiarizados con la programación.

Para los usuarios de Windows, Visual Studio Code, desarrollado por Microsoft puede resultar una muy buena opción al ser de uso gratuito, tener facilidades para detectar y apoyarnos con la sintaxis de varios lenguajes de programación entre los

[4] Amazon, "¿Qué es Python?", en https://aws.amazon.com/es/what-is/python/, consultado el 4 de octubre de 2023.
[5] https://www.python.org/downloads/

que se incluye Python. Asimismo, al ser desarrollado por el propio Microsoft, no nos requiere acudir con desarrolladores desconocidos que puedan poner en riesgo la integridad de nuestros equipos.

## 2.3 Los scripts de Python

Para emplear Python, lo primero que debemos hacer es crear, u obtener, un *script*. Esencialmente, un script contiene una serie de instrucciones escritas en el lenguaje de programación que estamos usando, que en este caso es Python, que el ordenador puede interpretar y ejecutar en resultados deseados.

```
SYS_SALIDA equ 1

section .data
    msg db "Hello, World",0x0a
    len equ $ - msg ;longitud de msg

section .text
global _start ;para el linker
_start: ;marca la entrada
    mov eax, 4 ;llamada al sistema (sys_write)
```

*Figura 1. Ejemplo de script que se escribe por primera vez, " Hello, World".*

Si bien los usuarios más experimentados en Python pueden escribir de cero sus propios scripts, es probable que para quienes se inician en su uso sea más sencillo adaptar scripts ya hechos a sus necesidades. Este proceso es una parte fundamental en el mundo de la programación y el análisis de datos;  y es uno de los motivos por los cuales se incluyen varios ejemplos en esta guía. ES importante mencionar que esta guía no pretende enseñar Python de manera intensiva, pero di dará un enfoque de uso de la herramienta aplicado a datos socioeconómicos. Así que recomendamos ampliamente a quien esté interesado que busque aprender más sobre este lenguaje.

# 3. Extracción y preparación de los datos

En este apartado damos pautas generales de los problemas que pueden aparecer a la hora de crear un repositorio de datos y posibles formas de resolverlos. Es importante recalcar este último punto, ya que en el mundo de procesamiento de información existen numerosas formas de aproximarnos a un problema. Hay una serie de puntos antes de comenzar, que permite hacer más sencilla la extracción y posterior utilización de la información que queramos. **Es importante mencionar que los ejemplos y códigos mostrados a continuación están basados en alguna de las variables para los municipios, localidades o estados de los sitios de estudio mencionados previamente.**

## 3.1 Preparación de los datos

### 3.1.1 Los datos con acentos

Los acentos y caracteres especiales en la información contenida en una base de datos pueden generar problemas significativos en términos de búsqueda, comparación y procesamiento de datos. En primer lugar, la presencia de acentos puede dificultar la búsqueda eficiente de información, ya que los motores de búsqueda y algoritmos de procesamiento de texto pueden no reconocer correctamente las variantes acentuadas y no acentuadas de una misma palabra. Esto puede llevar a resultados de búsqueda incompletos o inexactos.

Esto es especialmente relevante en términos de uniformidad y coherencia de la información almacenada. Algunas bases de datos y sistemas de búsqueda son sensibles a las diferencias con y sin acentos, lo que podría llevar a resultados no deseados si no se establece una convención clara. Al mantener todo en un estándar, se asegura una estructura uniforme y homogénea que facilita la manipulación, búsqueda y análisis de los datos.

En resumen, la eliminación de acentos tiene como finalidad mejorar la precisión en la búsqueda de la información deseada. Esto previene situaciones donde podríamos no encontrar, por ejemplo, el nombre de una localidad debido a la ausencia, o a la presencia, de acentos. Realizar búsquedas que no arrojen resultados correctos debido a errores de escritura sería equivalente a obtener falsos positivos, es decir, la información existe, pero no se ha localizado de manera precisa y adecuada.

En general, se recomienda no emplear acentos cuando se hace un manejo computarizado de datos. Los ordenadores no siempre están equipados para reconocer ciertos caracteres, lo que puede generar problemas más adelante. Asimismo, puede ser buena idea homogeneizar también el uso de mayúsculas y minúsculas, pues a menudo las computadoras y programas que empleemos harán distinción entre ambos conjuntos de caracteres.

## ¿Cómo se soluciona?

Existen muchos métodos para resolver un problema de este tipo, que podrían incluir incluso hacer una revisión "manual", o de caso por caso para asegurarse que se respeten las convenciones establecidas. Para la mayoría de las bases de datos habrá cambios significativos que llevarían mucho tiempo poderse modificar. En consecuencia, sugerimos el uso de código Python para homogeneizar de manera automática la información.

### Código en Python

1. Se importan las bibliotecas necesarias: *pandas* para manejar los datos y *xlswriter* para crear un nuevo archivo Excel.
2. Definir una función llamada 'sin_tilde' que toma una cadena de texto y reemplaza las vocales con tildes por sus equivalentes sin tildes, tanto en minúsculas como en mayúsculas.
3. Leer el archivo .xlsx , definiendo antes el directorio de trabajo.
4. Selección de la ubicación y el nombre del archivo de salida, es decir, el que no contiene acentos
5. Iniciar un contador para llevar las iteraciones en cada fila.
6. Crear un bucle para mantener el nombre de las columnas en la nueva hoja resultado, aunque sin tildes.
7. Crear un segundo bucle para operar a través de las filas en el DataFrame original (*df.iterrows*), después en cada fila y columna aplicar la función *sin_tilde* a cada valor de celda.
8. Escribir los valores procesados en la hoja de salida y aumenta el contador para avanzar a la siguiente fila.
9. Finalmente se cierra el archivo Excel de salida después de completar la escritura de los datos procesados.

**Encontrarás el ejemplo del código en la siguiente página:**

```python
# Librerias
import pandas as pd
import xlsxwriter

# Función para eliminar tildes
def sin_tilde(MiCadena):
    replacements = (
    ("á", "a"),
    ("é", "e"),
    ("í", "i"),
    ("ó", "o"),
    ("ú", "u"),
    )
    for a, b in replacements:
        MiCadena = MiCadena.replace(a, b).replace(a.upper(), b.upper())
    return MiCadena

# Directorio donde se encuentra la informacion
%cd "F:\direccion de donde se guardo la base a utilizr"

# Lee el archivo Excel
df = pd.read_excel("nombre de la base.xls", sheet_name="nombre de la hoja")

# Directorio donde se gurdara la salida
workbook = xlsxwriter.Workbook("F:\direccion de donde se quiere guardar los datos de
salida\nombre de la base de salida.xlsx")
worksheet = workbook.add_worksheet("nombre de la hoja de salida")

# Contador para realizar la función para cada fila y columna
count = 0

# Mantener el nombre de las columnas de datos de salida
for i, column_name in enumerate(df.columns):
    worksheet.write(count, i, sin_tilde(column_name))
count += 1

# Aplicación de la función en los datos
for index, row in df.iterrows():
    for i in range(0, df.shape[1]):
        worksheet.write(count, i, sin_tilde(str(row[i])))
    count += 1

workbook.close()
```

*Figura 2. Código para homogenizar datos y eliminar caracteres especiales.*

### 3.1.2 Diseño de la base de datos

Antes de "vaciar" los datos que hemos preparado en un formato adecuado para nuestro trabajo, es necesario considerar la forma en que vamos a emplear la información. Así como preguntarnos quién tendrá acceso y qué objetivos buscamos conseguir al construir esta base de datos (de nuevo tenemos que recordar las preguntas de investigación).

Idealmente, las bases que construyamos contendrán toda la información relevante que haya sido posible encontrar a lo largo del proceso de búsqueda. Sin embargo, definir cuál es la información verdaderamente relevante puede ser más difícil de lo que parece. Por ejemplo, a lo largo del proceso puede ser que una variable desestimada —y por tanto excluida— previamente resulte necesaria, y volver a

realizar la búsqueda y extracción de dicha variable, aunque posible, retrasaría el proceso de investigación.

A su vez, es importante balancear esta posibilidad con la necesidad de que nuestras bases de datos sean legibles y accesibles a los miembros del equipo de investigación que las requieran. Y considerar que la inclusión de una cantidad enorme de datos y variables irrelevantes puede tener un efecto adverso sobre la legibilidad y utilidad.

Otro tanto puede decirse de los metadatos, que podemos considerar como información *sobre* la información, aunque en este caso lo mejor es incluir tanta información como sea posible. Prestando especial atención en la claridad de las fuentes de la que fueron extraídos los datos, el número de versiones y disponibilidad, así como el momento en que se realizó la descarga o extracción de datos.

Finalmente, con el fin de mantener la accesibilidad de nuestras bases de datos, se recomienda hacer un molde que le dé una forma *convencional*[6] a todas las bases de datos que vayamos a construir para un proyecto, de tal suerte que resulte sencillo navegar por cada una conociendo su estructura general. Esto es de enorme importancia cuando son varios los colaboradores de un proyecto que están participando en el proceso de construcción de las bases de datos, ya que la consistencia a lo largo del trabajo es esencial.

## 3.2 Durante la extracción de datos

### 3.2.1 Selección y filtrado de datos

Con la base de datos libre de las acentuaciones, se carga y se le aplican filtros para obtener la información necesaria para la investigación en curso.

#### ¿Cómo se soluciona?

La manera más práctica de seleccionar la información objetivo de la investigación es partir primero desde los filtros más generales hacia los filtros más específicos. Por ejemplo, en este proyecto se desea información a nivel municipal que muestre las condiciones socioeconómicas de la región, por lo que de esta manera el primer filtro o selección será la de los municipios para después seleccionar las series o datos socioeconómicos que sean propios de la investigación.

[6] Por supuesto que aquí nos referimos a que la forma dependerá de un formato establecido dentro del grupo de trabajo. Debido a que hay muchas formas lógicas de ordenar un conjunto de datos y reportar la información relevante, es recomendable que se tome una decisión conjunta

**Código en Python**

1. Importar las bibliotecas *pandas* para el manejo de datos, *numpy* para operaciones numéricas y *xlsxwriter* para la generación de archivos .xlsx.
2. Importar la base de datos que se utilizará, definiendo previamente el directorio de trabajo.
3. Se realiza un análisis exploratorio para identificar el contenido de las columnas y la naturaleza de los datos.
4. Siguiendo un enfoque de agrupación desde lo más amplio hasta lo más específico, se seleccionan las regiones relevantes para el proyecto.
5. Por precaución, se verifica la extracción adecuada de las regiones objetivo comparándola con el conjunto de datos extraídos.
6. Hay dos posibles escenarios si no concuerdan los datos objetivo y real:
    a. Primero, la región realmente no está en la base de datos.
    b. Segundo, la región está presente, pero la forma en que se busca puede ser incorrecta debido a diferencias en mayúsculas, minúsculas, espacios, puntos o guiones.[7] De ser el caso:
        i. Una forma de confirmar la escritura correcta de la región es mediante análisis exploratorios que revelen patrones en los datos. Esto se logra con experiencia y consulta de diversas fuentes de datos.
        ii. Es recomendable mantener una lista de nombres alternativos para cada región a medida que se trabaja con diferentes bases de datos. Esto reduce los falsos positivos al buscar información.
7. Agrupar las regiones, como municipios y entidades federativas, en categorías más generales para facilitar la investigación.
8. Agregar el año, o período temporal al que pertenece la información para contextualizarla temporalmente.
9. Renombrar las columnas con etiquetas más cortas y descriptivas, facilitando la búsqueda y comprensión de los datos.
10. Reorganizar las columnas en la base de datos, siguiendo un orden de izquierda a derecha que refleje la agrupación de lo general a lo particular.
11. La base de datos resultante se guarda en formato .xlsx, asegurándose de establecer previamente el directorio de trabajo.

[7] Si bien se realizó un proceso de limpieza de los datos con vistas a eliminar problemas relacionados con la acentuación, siempre existe la posibilidad de que haya habido algún error en nuestro código, o que por el motivo que sea la información no haya sido modificada con éxito. Asimismo, es posible que otro tipo de errores no previstos anteriormente ocurran. Dependiendo cuál sea el caso, puede ser útil hacer una exploración manual de los datos para identificar el problema antes de buscar una solución automatizada.

```python
# Librerias
import pandas as pd

# Importar base de datos

%cd "F:\Beca Conacyt\Bases de datos_crudas\UE"
df = pd.read_excel("data_2009.xlsx", sheet_name= "concentrado")

# Conocer las columnas o series que contiene el archivo
df.columns

# Extracción de la data por municipio

df1 = df[df['Municipio'].isin(["Ahome","Calkini","Candelaria loxicha",
            "Carmen","Celestun","Champoton","Escarcega","Huatabampo","Hunucma","Palizada",
            "Paraiso","Pluma hidalgo","Progreso","San agustin loxicha",
            "San carlos yautepec","San francisco ozolotepec",
            "San juan mixtepec",
            "San juan mixtepec -dto. 08 -", "San juan mixtepec -dto. 26 -",
            "San juan ozolotepec","San marcial ozolotepec","San mateo piñas","San mateo rio hondo",
            "San miguel del puerto","San miguel suchixtepec","San pedro el alto","San pedro huamelula",
            "San pedro mixtepec",
            "San pedro mixtepec -dto. 22 -", "San pedro mixtepec -dto. 26 -",
            "San pedro pochutla","San sebastian rio hondo","Santa maria huatulco",
            "Santa maria ozolotepec","Santa maria tonameca","Santiago xanica","Santo domingo ozolotepec",
            "Oaxaca", "Sinaloa", "Sonora", "Tabasco", "Yucatan", "Campeche"])]

# Confirmar que todos los municipios se hayan seleccionado adecuadamente

df1['Municipio'].unique().size #Este solo funciona como un apartado general, pero puede conducir a malas
interpretaciones

objetivo = ["Ahome","Calkini","Candelaria loxicha",
            "Carmen","Celestun","Champoton","Escarcega","Huatabampo","Hunucma","Palizada",
            "Paraiso","Pluma hidalgo","Progreso","San agustin loxicha",
            "San carlos yautepec","San francisco ozolotepec",
            "San juan mixtepec",
            "San juan mixtepec -dto. 08 -", "San juan mixtepec -dto. 26 -",
            "San juan ozolotepec","San marcial ozolotepec","San mateo piñas","San mateo rio hondo",
            "San miguel del puerto","San miguel suchixtepec","San pedro el alto","San pedro huamelula",
            "San pedro mixtepec",
            "San pedro mixtepec -dto. 22 -", "San pedro mixtepec -dto. 26 -",
            "San pedro pochutla","San sebastian rio hondo","Santa maria huatulco",
            "Santa maria ozolotepec","Santa maria tonameca","Santiago xanica","Santo domingo ozolotepec",
            "Oaxaca", "Sinaloa", "Sonora", "Tabasco", "Yucatan", "Campeche"]


realidad = d['Municipio'].unique()

realidad = d['Municipio'].unique()

faltantes = set(objetivo) - set(realidad)
print(faltantes)

# Confirmar si los datos faltantes efectivamente estos datos no se tienen o hubo un error en el código

valores_a_buscar = faltantes

resultados = df1[df1['Municipio'].isin(valores_a_buscar)]


if not resultados.empty:
    print("Los valores existen en la columna:")
    print(resultados)
else:
    print("Los valores no existen en la columna.")
```

*Figura 3.Primera parte de código para la selección y filtrado de variables (continua en la siguiente página).*

```python
# Confirmar que los municipios efectivamente pertenezcan a la entidad federativa adecuada

df1.columns

municipios_a_filtrar = realidad
entidades_a_filtrar = ["Oaxaca", "Sinaloa", "Sonora", "Tabasco", "Yucatan", "Campeche"]

for municipio, entidad in zip(municipios_a_filtrar, entidades_a_filtrar):
    df2 = df1.drop(df1[(df1['Municipio'] == municipio) & (df1['Entidad'] != entidad)].index)
print(df2)


# Añadir el nombre al observatorio que pertenecen

df2.columns

conditionlist = [
    (df2['Entidad'] == 'Campeche'),
    ((df2['Entidad'] == 'Sonora') | (df2['Entidad'] == 'Sinaloa')),
    (df2['Municipio'] == 'Progreso'),
    (df2['Municipio'] == 'Calkini'),
    (df2['Municipio'] == 'Celestun'),
    (df2['Entidad'] == 'Oaxaca'),
    (df2['Entidad'] == 'Tabasco'),
    (df2['Municipio'] == 'Hunucma')
                 ]

choicelist = ['Laguna de Terminos', 'Agiabampo',
              'Arrecife Alacranes', 'Celestun',
              'Celestun',
              'Copalita', 'Dos Bocas', 'Sisal'
              ]

df2['id_observatorio'] = np.select(conditionlist, choicelist)

# Añadir el año al que pertenece la data

df2['AÑO'] = '2009'

# Cambiar el nombre de las variables
df2.columns

df3 = df2.rename(columns={'Entidad' : 'id_estado',
                          'Municipio' : 'id_municipio',
                          'AÑO' : 'año',
'A111A Producción bruta total (miles de pesos)' :   'Produccion Bruta Total',
'Actividad Económica' : 'Actividad Economica',
'UE Unidades económicas' :  'Numero de Unidades Economicas',
                          })

df3.columns

# Ordenar las variables

df3=df3[['id_observatorio', 'id_estado', 'id_municipio',
        'año', 'Actividad Economica',
       'Numero de Unidades Economicas', 'Produccion Bruta Total']]


# Convertir el observatorio en INDEX

df4=df3.set_index('id_observatorio')


# Guardar en formato .xlsx

%cd "F:\Directorio"

df4.to_excel('nombre_base_de_datos.xlsx', sheet_name = 'nombre_de_la_hoja')
```

*Figura 4. Segunda parte de código para la selección y filtrado de variables.*

### 3.2.2 Manejo de NaN

Controlar una base de datos que contiene valores NaN (Not a Number) es esencial para mantener la integridad y utilidad de los datos. Los valores NaN son una representación utilizada en informática y análisis numérico para denotar valores que no son numéricos o que no pueden ser expresados como números válidos en operaciones matemáticas. Suelen aparecer en bases de datos cuando los datos están incompletos o no están disponibles. Para gestionar adecuadamente estos valores, se pueden seguir tres enfoques clave.

En primer lugar, es importante identificar los valores NaN en la base de datos. Esto se puede lograr mediante consultas específicas que busquen y muestren registros con valores faltantes. A continuación, se pueden tomar decisiones informadas sobre cómo tratar esos valores. Segundo, se debe considerar la opción de reemplazar los valores NaN con un valor predeterminado o estimado. Esto ayuda a mantener la integridad del conjunto de datos y permite realizar análisis coherentes. Sin embargo, es fundamental elegir un valor de reemplazo adecuado y documentar el proceso para evitar distorsiones en los resultados.

Por último, se puede optar por omitir los registros con valores NaN en análisis específicos. Algunos algoritmos y modelos de análisis pueden manejar la ausencia de datos, mientras que otros pueden requerir una imputación previa. En cualquier caso, se debe evaluar cuidadosamente si omitir registros afectará la validez de los resultados y si es necesario aplicar técnicas de imputación para mantener la representatividad de los datos. En resumen, controlar una base de datos con valores NaN implica identificar, reemplazar o manejar adecuadamente estos valores para garantizar la coherencia y utilidad de los análisis y decisiones basados en los datos.

#### ¿Cómo se soluciona?

En el marco de este proyecto, se ha optado por la alternativa de señalar únicamente los valores NaN. Esto se logra a través de una observación atenta de los valores NaN, ya sea mediante una simple observación inicial o a través de un análisis más exhaustivo y profundo. No obstante, es fundamental ejercer precaución, dado que en algunas bases de datos los valores NaN no son siempre declarados de manera evidente, lo que implica la necesidad de llevar a cabo una búsqueda activa para identificarlos de manera adecuada.

**Código en Python**

1. Se busca examinar cada una de las filas de cada columna con el fin de determinar qué valores existen, y ubicar los NaN que no han sido identificados.
    a. En este proceso si los valores NaN están establecidos no se declaran valores únicos, de manera que se omiten todos los puntos siguientes hasta el último.
2. Se utiliza el método *apply* para aplicar una función en este caso a lo largo de todas las columnas, y se le aplica el método *unique*, que devuelve los valores únicos que viven en cada columna
3. Una vez encontrando imperfecciones en la uniformidad de datos, se puede visualizar los NaN no identificados, tales como puntos (.), guiones (-) , ceros (0) , etc.
4. Se define el valor no identificado como NaN para ser definido como tal.
5. Por último, se contabilizan los valores nulos o NaN , y así se observa su correcta especificación.

```
# Analizar que tipo de valores viven en cada columna
valores_unicos_por_columna = df.apply(lambda col: col.unique())
print("Valores únicos por columna:")
print(valores_unicos_por_columna)


# Convertir los caracteres que representan NaN a efectivamente NaN
df1 = df.replace({"-" : np.NaN}).copy()


# Mostrar todos los valores nulos o NaN
df1.isnull().sum
```

*Figura 5. Código para identificar y reemplazar valores únicos a NaN.*

**TIP**

La inclusión de valores como "-" o "0" suele estar vinculada a valores tipo NaN en una base de datos; sin embargo, la determinación del símbolo que haga alusión a datos faltantes o ausentes depende del criterio del investigador. Por lo tanto, resulta esencial examinar detenidamente la documentación relacionada con la base de datos para comprender con precisión su significado.

- Un ejemplo común es el valor "-", el cual por lo general se asocia con datos faltantes o situaciones en las que ciertos datos no son aplicables. El primer caso representa una ausencia de información, mientras que la segunda categoría indica que los datos carecen de aplicabilidad o relevancia en ese contexto específico.

- Otro ejemplo se refiere al uso del valor "0", el cual con frecuencia tiene dos interpretaciones distintas: en primer lugar, puede indicar un valor real igual a cero; en segundo lugar, puede señalar la total inexistencia de una característica particular.

# 4. Gestión y manejo de datos

Una vez que se tienen construidas las bases de datos, es necesario mantener la claridad y la transparencia de los datos, lo que representaría una buena gestión y manejo de estos.

## 4.1 Información disponible vs información solicitada

En el ámbito de la gestión de datos, es común encontrarse con la disparidad entre la información disponible y la información solicitada en las bases de datos. Esta disparidad puede surgir por varias razones, una de las más comunes es encontrar bases de datos incompletas.

A menudo, estos detalles se observan al comparar diferentes estados, municipios o localidades. Para afrontar esta problemática, resulta esencial tener claras las delimitaciones de cada región, ya que algunos sitios pueden contar con datos exhaustivos y otros, por el contrario, podrían carecer de información.

Esta situación puede generar desafíos al tratar de obtener una imagen precisa y completa de la situación de la región. Al realizar análisis comparativos o tomar decisiones basadas en datos, es importante tener en cuenta esta disparidad. Este conocimiento permitirá no solo advertir sobre las restricciones que podrían surgir al trabajar con dichos datos, sino también comprender las posibles implicaciones contenidas. En consecuencia, se plantea el problema de la visibilidad de datos en el siguiente apartado.

### ¿Cómo se soluciona?

Se debe crear una función que me indique todos los valores nulos, de esta manera debieron especificarse correctamente todos los valores NaN.

**Código en Python**

1. Importar las librerías que se utilizarán (ojo: si se declararon anteriormente no es necesario la inclusión)
2. Leer el archivo .xlsx , no olvidando definir el directorio de trabajo.
3. Se define una función llamada '*calcular_porcentaje_nan (dataframe)*' que calcula el porcentaje de información completos en un base de datos.
4. Por último, se llama a la función creada para la obtención del porcentaje de datos completos.

```python
# Librerias
import pandas as pd

# Directorio
%cd "F:\Beca Conacyt\Bases datos_limpias\ingresoso"


# Base de datos a utilizar
datos = pd.read_excel("2005_2023_ingresos_tipoempleo_trabajadorindependientes_entidad.xlsx",
sheet_name= "Tabulado")

# Defines la función
def calcular_porcentaje_nan(dataframe):
    total_valores = dataframe.size
    valores_nan = dataframe.isnull().sum().sum()
    porcentaje_nan = (valores_nan / total_valores) * 100
    porcentaje_completo = 100 - porcentaje_nan
    return porcentaje_completo

# Aplicas la función
porcentaje_completo = calcular_porcentaje_nan(datos)
print(f"El porcentaje de completitud de los datos es: {porcentaje_completo}%")
```

*Figura 6. Código de identificación de vacíos en una base de datos.*

## 4.2 Creación del metadato

Un metadato es la información que describe y proporciona contexto sobre otros datos. En esencia, son datos sobre los datos. Los metadatos brindan detalles y características que permiten comprender y administrar los datos de manera más efectiva. Pueden incluir información sobre la estructura, el significado, la fuente, la calidad, el formato, la temporalidad y otros atributos de los datos.

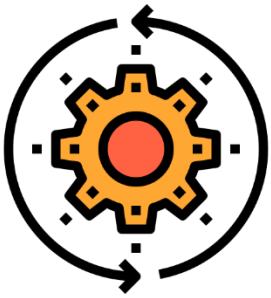
La importancia de los metadatos radica en su capacidad para mejorar la confiabilidad y la transparencia de los datos. Ayudan a evitar malentendidos y errores de interpretación al proporcionar detalles sobre los datos extraídos. Además, los metadatos son esenciales para facilitar la colaboración entre equipos multidisciplinarios, ya que establecen un lenguaje común para la comunicación.

### 4.2.1 Pasos para la creación de un metadato.

- Identificar los elementos clave: se debe determinar qué aspectos son esenciales para comprender y utilizar los datos de manera efectiva. Esto puede incluir detalles de cada variable o campo de datos, explicando su significado, unidad de medida, además de información sobre la fuente de información, la fecha de creación o su periodicidad de actualización.

- Definir estándares: hay que establecer cómo se presentarán los metadatos y qué estándares de nomenclatura o formatos de fecha se seguirán para mantener la coherencia y la claridad, y sobre todo la uniformidad entre diferentes bases de datos.

- Indicar limitaciones: Debe indicarse cualquier tipo de anomalía en los datos o datos faltantes o inclusive condiciones en el uso de la data para determinadas investigaciones. Esto con el fin de advertir a cualquiera que haga uso de ella las limitaciones de información.



**Ejemplo**

Un metadato cumple la función de proporcionar información valiosa al usuario que examina la base de datos. Para crearlo, es posible emplear una tabla en formato Excel en la que las columnas representan las distintas características de cada variable incluida en el conjunto de datos.

En este caso, hemos empleado encabezados como "Variable" para identificar la variable, "Descripción" para describir el tipo de información contenida, "Unidad de medida" para indicar el formato del dato, "Frecuencia" para señalar la regularidad con la que se presenta el dato, "Cobertura a nivel observatorio" para determinar si la variable es accesible para todos los observatorios, "Cobertura geográfica" para especificar el nivel máximo de desagregación geográfica, "Cobertura temporal" para establecer el período de registros, "Fuente" para identificar la fuente de los datos, "URL de acceso rápido" para enlazar directamente a la fuente primaria o secundaria, y "Fecha de actualización" para indicar la última vez que la base de datos se modificó.

A continuación, un ejemplo de llenado del cuadro metadato:

| *Variable* | *Descripcion* | *Unidad de medida* | *Frecuencia* |
|---|---|---|---|
| ECOD01_IDH | Índice de Desarrollo Humano | Índice | Quinquenal |
| ECOD02_INS | Índicador de Salud | Índice | Quinquenal |
| ECOD03_INE | Índicador de Educación | Índice | Quinquenal |
| ECOD04_INI | Índicador de Ingreso | Índice | Quinquenal |

| *Cobertura nivel observatorio* | | | | | | | *Cobertura geográfica* | *Cobertura temporal* |
|---|---|---|---|---|---|---|---|---|
| Agiabampo | Arrecife Alacranes | Celestun | Copalita | Dos Bocas | Laguna de | Sisal | Municipal | 2020-2022 |
| Agiabampo | Arrecife Alacranes | Celestun | Copalita | Dos Bocas | Laguna de Términos | Sisal | Municipal | 2020-2022 |
| Agiabampo | Arrecife Alacranes | Celestun | Copalita | Dos Bocas | Laguna de Términos | Sisal | Municipal | 2020-2022 |
| Agiabampo | Arrecife Alacranes | Celestun | Copalita | Dos Bocas | Laguna de Términos | Sisal | Municipal | 2020-2022 |

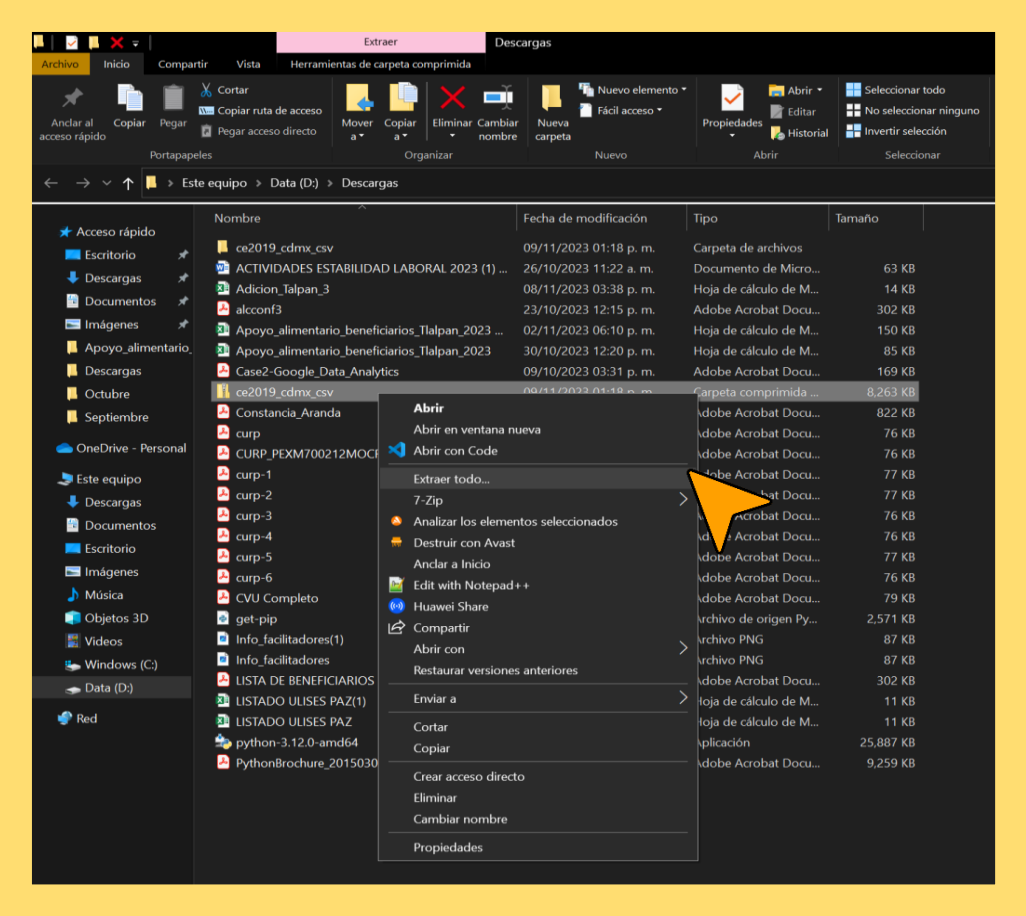
| *Fuente* | *URL de acceso rápido* | *Fecha de actualización* |
|---|---|---|
| PNUD. (2020). *Índice de Desarrollo Humano* | https://www.idhmunicipalmexico.org/ | 8/23/2023 |

# 5. Ejemplo de obtención de datos con Excel

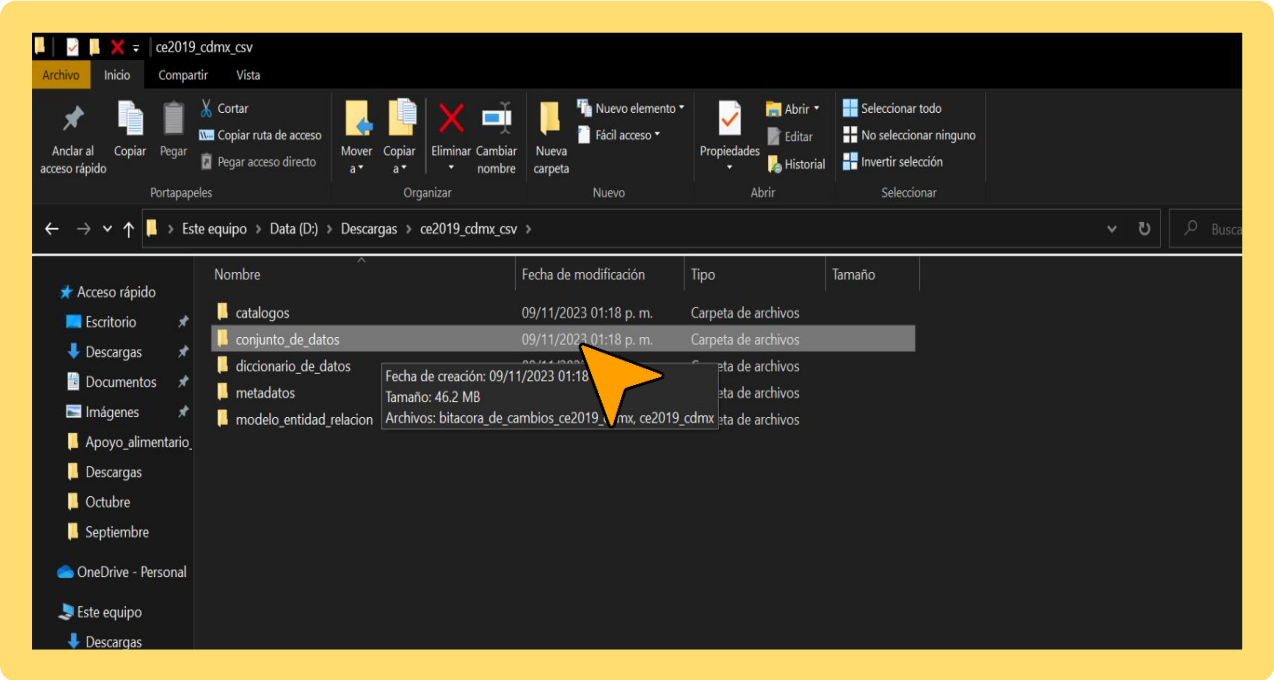
En este capítulo final incluiremos algunos puntos técnicos que podrían ayudar a solidificar lo presentado anteriormente en esta guía. Como una alterativa al uso de Python se presenta el siguiente ejemplo de cómo desarrollar una base de datos disponible en internet usando hojas de Excel.

En este ejemplo, emplearemos únicamente herramientas de hojas de cálculo, y obtendremos los datos de cinco alcaldías (Benito Juárez, Coyoacán, Tlalpan, Tláhuac, y Milpa Alta) de la Ciudad de México. Recopilaremos los datos de Producto Bruto Total, Inversión Total y Total de Ingresos disponible en el Censo Económico 2019:
https://www.inegi.org.mx/contenidos/programas/ce/2019/Datosabiertos/ce2019_cdmx_csv.zip

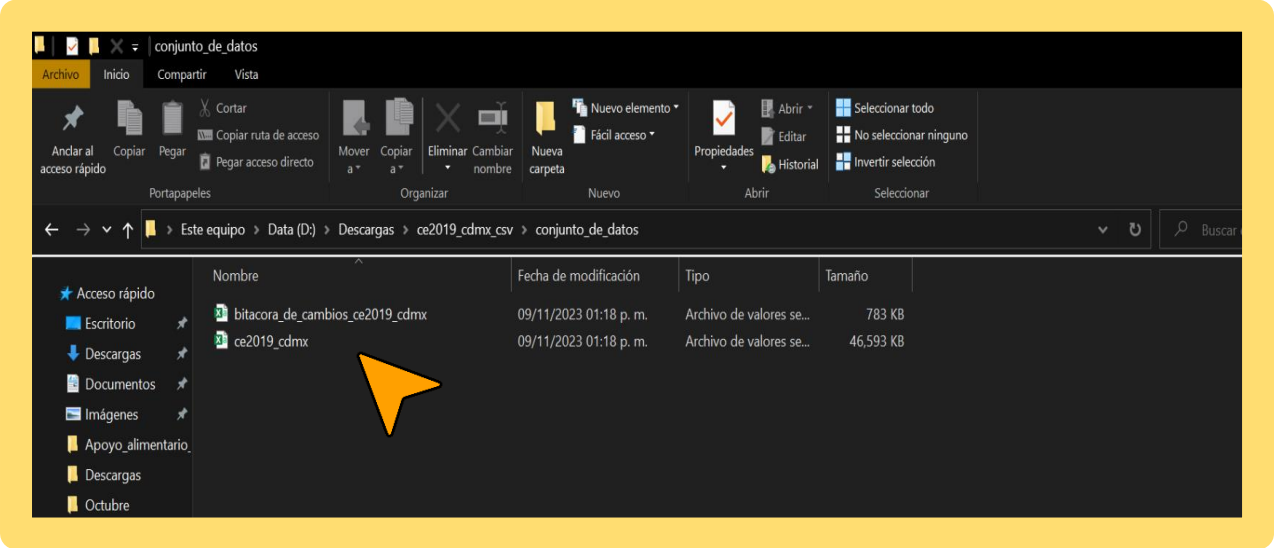
El primer paso es obtener el conjunto de compreso en un archivo de tipo ZIP. Esto implica que antes de poder hacer uso de los datos libremente, tendremos que dar clic derecho a nuestro archivo y seleccionar la opción de “Extraer todo”.



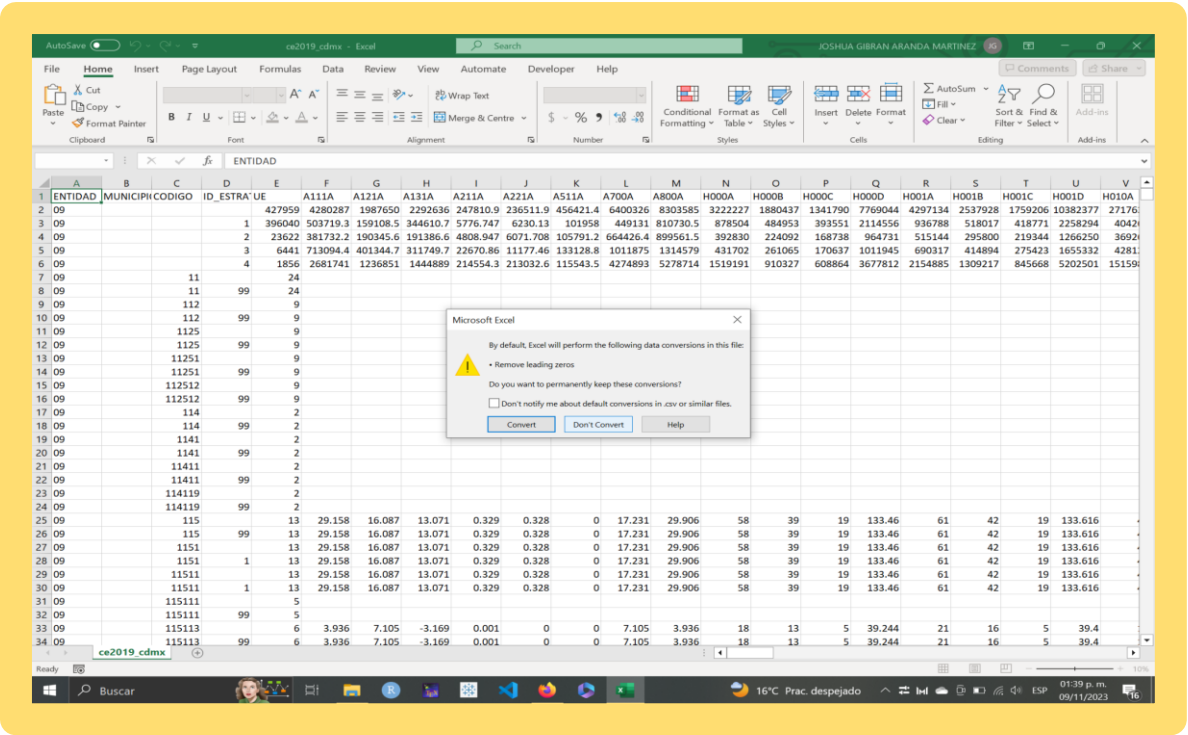
Habiendo extraído los ficheros del archivo ZIP, es posible continuar con el proceso. El conjunto de datos se encontrará anexa en la carpeta llamada “conjunto_de_datos”.



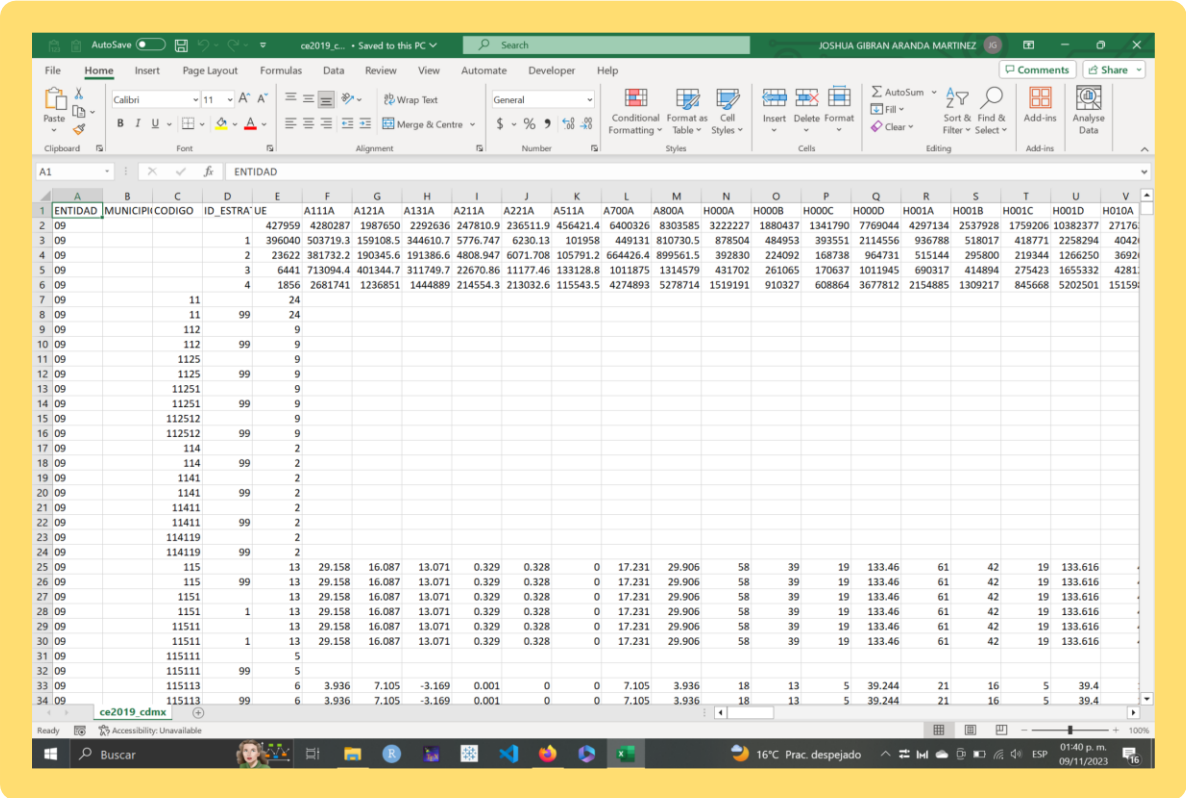

Dentro de la carpeta se encuentra la bitácora de cambios correspondiente a los datos. Para el ejercicio actual, se utilizará el conjunto de datos, pero siempre es recomendable que se guarde una bitácora de cambios que permita aclarar posibles discrepancias que puedan hallarse a lo largo del uso del conjunto de datos.




Se debe seleccionar y abrir el archivo “ce2019_cdm x”:



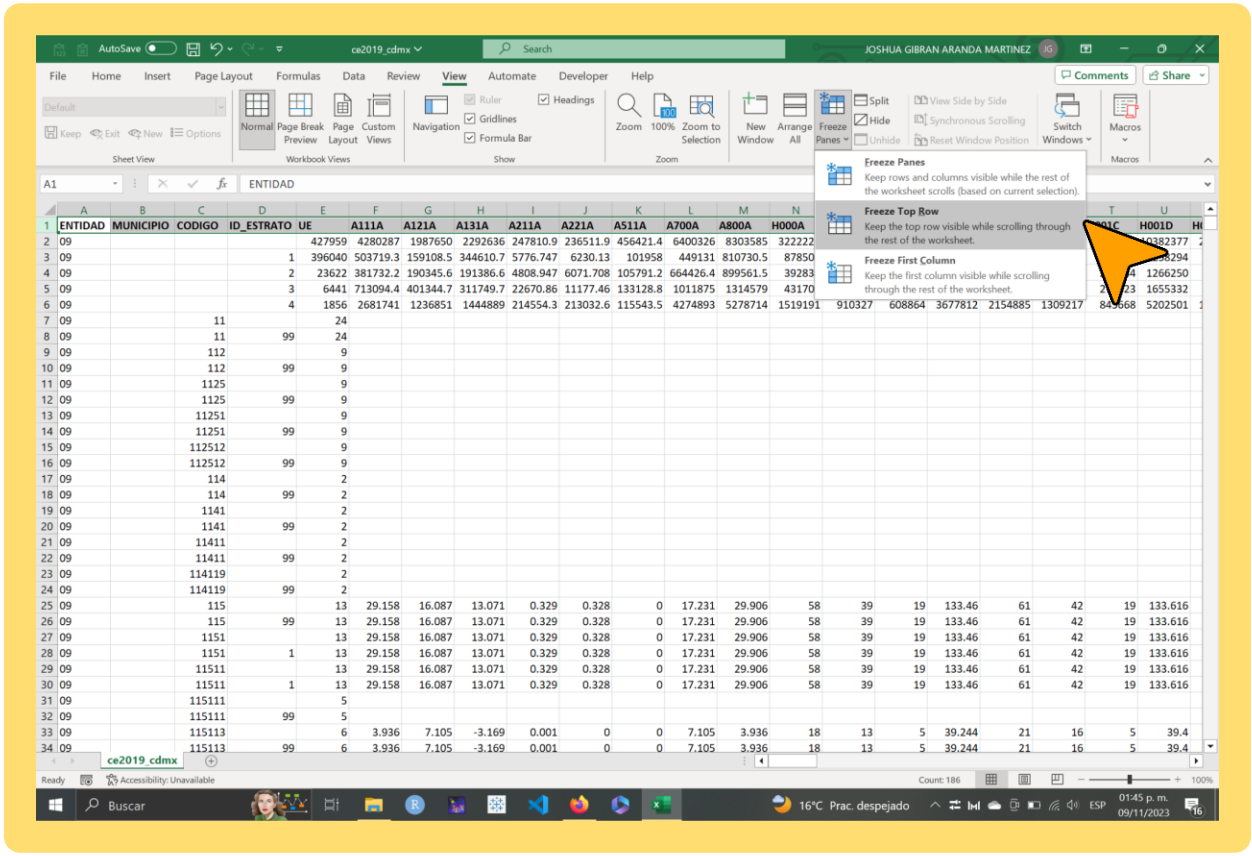

En Excel se suelen borrar los ceros a la izquierda. En casos como este, se recomienda no realizar esa conversión, pues éstos a menudo se emplean para denotar ciertas claves, ya sea geográficas o sectoriales.



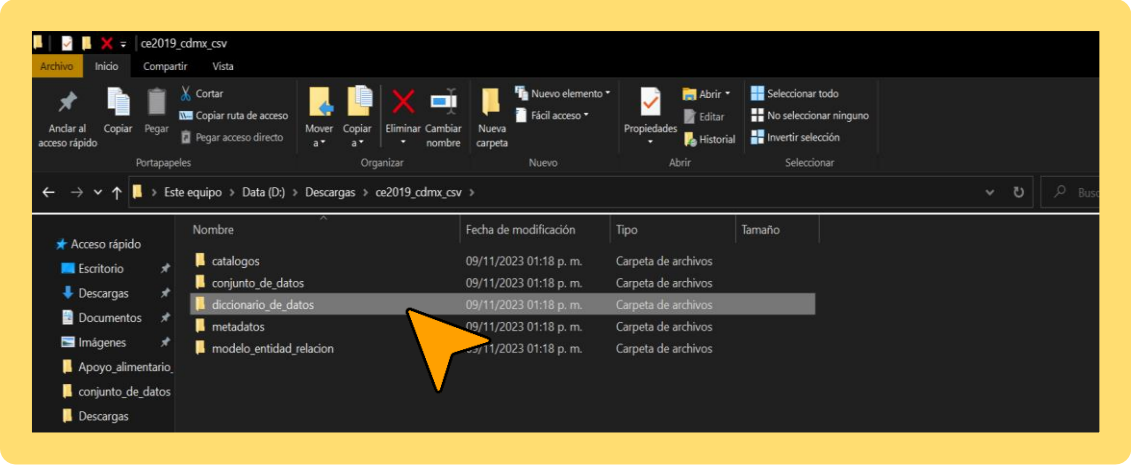

Como puede verse, el conjunto de datos es grande y las etiquetas no nos dicen mucho más allá que la entidad seleccionada (09) que corresponde a la Ciudad de México.

Algunos de los nombres de las columnas no alcanzan a desplegarse completamente. Con el fin de visualizar más claramente la información, son recomendables dos funciones comunes en los programas de hojas de cálculo: una, es ajustar el ancho de columna de tal suerte que se pueda ver con claridad cuál es el título de cada columna, y la otra es la de envolver el texto.

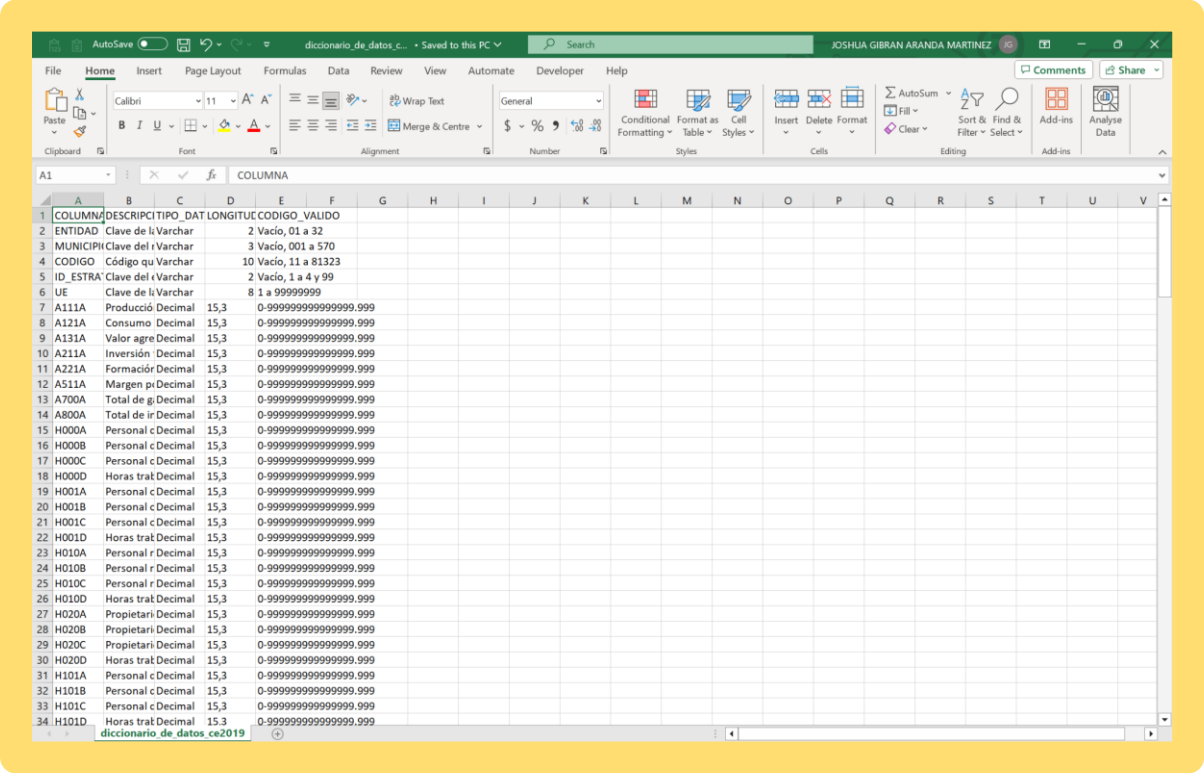
Asimismo, se recomienda remarcar en negritas la primera fila del encabezado, para denotar su importancia como títulos de columna y hacer más fácil la navegación. Un último paso es congelar la primera fila, de tal suerte que al navegar a lo largo de nuestro documento no perdamos de vista el nombre de la columna. Esta opción está disponible en la pestaña de Ver de Excel, seleccionando “congelar primera fila”:




Para este punto la navegación se facilita, pero se debe identificar qué significan los títulos de las columnas que tenemos. Para ello, se debe regresar a la carpeta original, donde se encuentra el archivo llamado “diccionario_de_datos”.

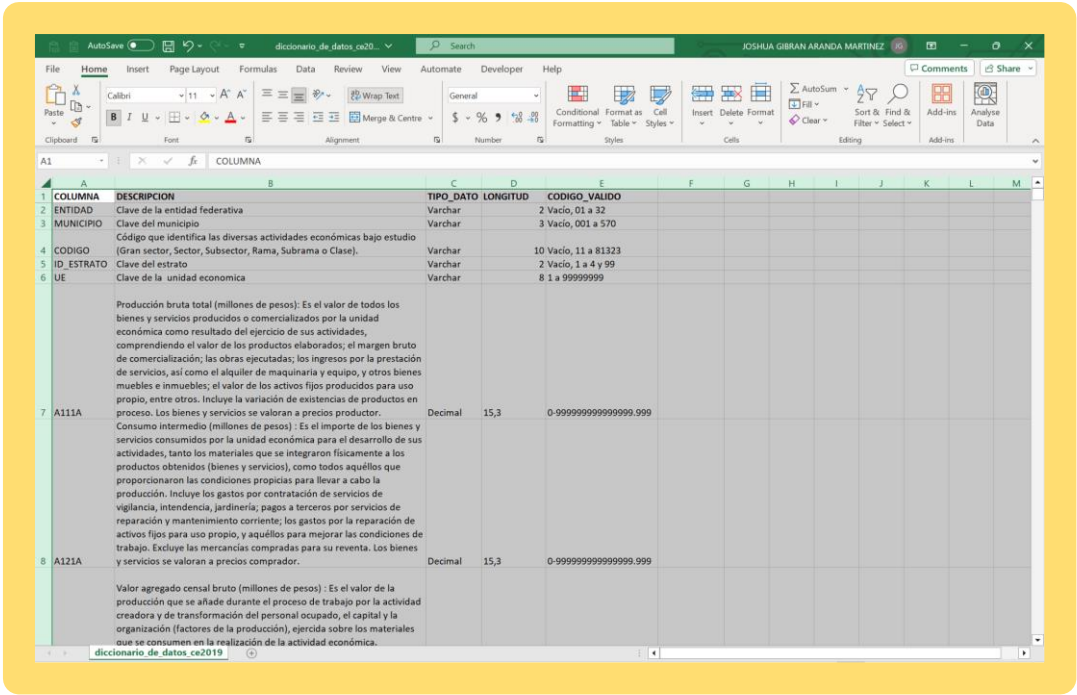



Un diccionario de datos permite identificar un conjunto nuevo de datos, en especial en casos donde las variables contenidas emplean números clave.




Como se realizó previamente, se recomienda facilitar la navegación añadiendo negritas, congelando la primera fila, ajustando el ancho de columna y haciendo uso de la función “Envolver”, que es especialmente útil cuando se tienen celdas con textos demasiado largos para que el ajuste del ancho de columna nos sea útil.

Por supuesto que hacer la columna bastante más ancha permite leer con más facilidad el texto envuelto.



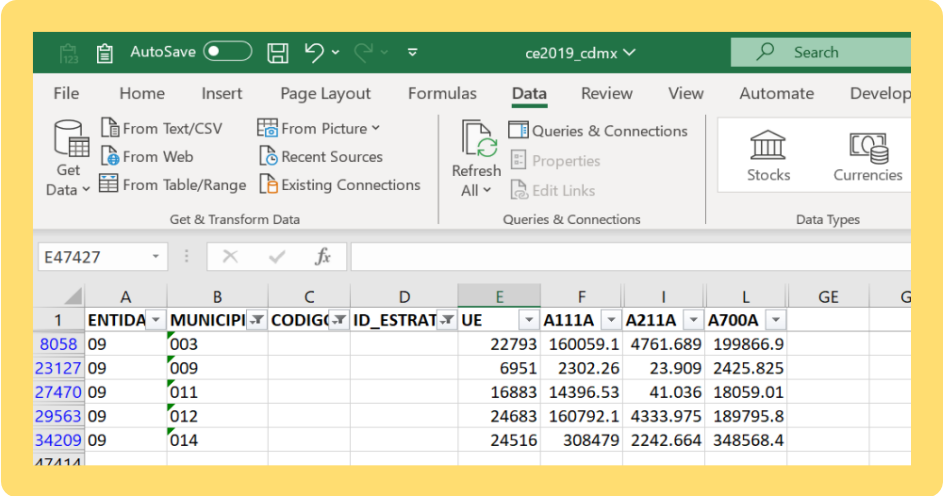
| COLUMNA | DESCRIPCION | TIPO_DATO | LONGITUD | CODIGO_VALIDO |
|---|---|---|---|---|
| ENTIDAD | Clave de la entidad federativa | Varchar | 2 | Vacío, 01 a 32 |
| MUNICIPIO | Clave del municipio | Varchar | 3 | Vacío, 001 a 570 |
| CODIGO | Código que identifica las diversas actividades económicas bajo estudio (Gran sector, Sector, Subsector, Rama, Subrama o Clase). | Varchar | 10 | Vacío, 11 a 81323 |
| ID_ESTRATO | Clave del estrato | Varchar | 2 | Vacío, 1 a 4 y 99 |
| UE | Clave de la unidad economica | Varchar | 8 | 1 a 99999999 |
| A111A | Producción bruta total (millones de pesos): Es el valor de todos los bienes y servicios producidos o comercializados por la unidad económica como resultado del ejercicio de sus actividades, comprendiendo el valor de los productos elaborados; el margen bruto de comercialización; las obras ejecutadas; los ingresos por la prestación de servicios, así como el alquiler de maquinaria y equipo, y otros bienes muebles e inmuebles; el valor de los activos fijos producidos para uso propio, entre otros. Incluye la variación de existencias de productos en proceso. Los bienes y servicios se valoran a precios productor. | Decimal | 15,3 | 0-999999999999999.999 |
| A121A | Consumo intermedio (millones de pesos) : Es el importe de los bienes y servicios consumidos por la unidad económica para el desarrollo de sus actividades, tanto los materiales que se integraron físicamente a los productos obtenidos (bienes y servicios), como todos aquéllos que proporcionaron las condiciones propicias para llevar a cabo la producción. Incluye los gastos por contratación de servicios de vigilancia, intendencia, jardinería; pagos a terceros por servicios de reparación y mantenimiento corriente; los gastos por la reparación de activos fijos para uso propio, y aquéllos para mejorar las condiciones de trabajo. Excluye las mercancías compradas para su reventa. Los bienes y servicios se valoran a precios comprador. | Decimal | 15,3 | 0-999999999999999.999 |
| | Valor agregado censal bruto (millones de pesos) : Es el valor de la producción que se añade durante el proceso de trabajo por la actividad creadora y de transformación del personal ocupado, el capital y la organización (factores de la producción), ejercida sobre los materiales que se consumen en la realización de la actividad económica. | | | |

Habiendo ajustado las columnas para que éstas resulten más legibles, se puede buscar la clave de las variables que se quieran recopilar. En este caso, las claves corresponden con A111A, A211A, y A700A. Teniendo identificadas las claves, se puede realizar la búsqueda en el conjunto de datos del censo económico. Sin embargo, se recomienda ocultar las columnas que no son útiles o prioritarias. Esto permite conservar la información en caso de que sea necesario consultarla más tarde.

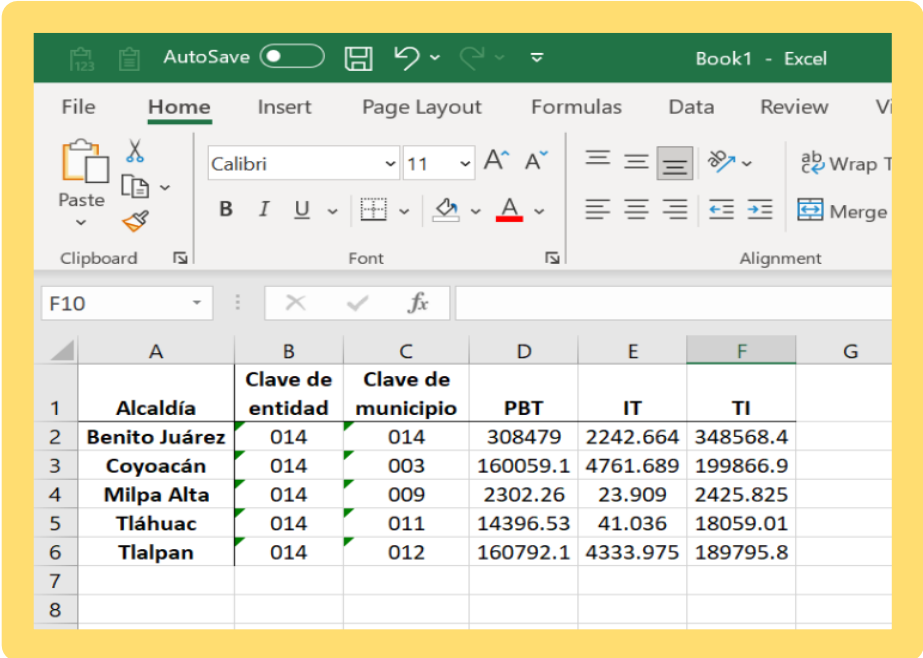
Ahora bien, se deben identificar los códigos de municipio que correspondan con las alcaldías objetivo. En este caso, sólo se incluye el código, así que se deben buscar cuáles son los correctos. Para las bases de datos de INEGI, la nomenclatura es uniforme, y para el caso de la Ciudad de México podemos hallar los códigos de municipio en la liga:
https://cuentame.inegi.org.mx/monografias/informacion/df/territorio/div_municipal.aspx?tema=me&e=09

| | |
|---|---|
| Benito Juárez | 014 |
| Coyoacán | 003 |
| Milpa Alta | 009 |
| Tláhuac | 011 |
| Tlalpan | 012 |

En la base de datos se deben emplear filtros para que sólo se muestren los datos relevantes. Consultando el diccionario de datos, se aprecia que la información viene dividida también en sectores y estratos. Para este ejemplo sólo se necesita el dato general, por lo que con el filtro se seleccionaran los lugares vacíos



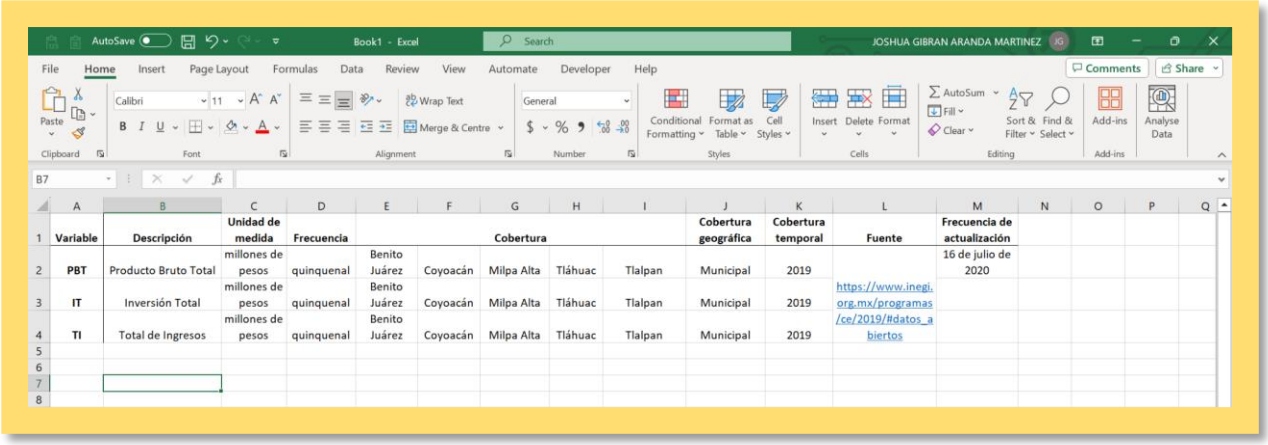
| | ENTIDA | MUNICIPI | CODIGO | ID_ESTRAT | UE | A111A | A211A | A700A |
|---|---|---|---|---|---|---|---|---|
| 8058 | 09 | 003 | | | 22793 | 160059.1 | 4761.689 | 199866.9 |
| 23127 | 09 | 009 | | | 6951 | 2302.26 | 23.909 | 2425.825 |
| 27470 | 09 | 011 | | | 16883 | 14396.53 | 41.036 | 18059.01 |
| 29563 | 09 | 012 | | | 24683 | 160792.1 | 4333.975 | 189795.8 |
| 34209 | 09 | 014 | | | 24516 | 308479 | 2242.664 | 348568.4 |



| | Alcaldía | Clave de entidad | Clave de municipio | PBT | IT | TI |
|---|---|---|---|---|---|---|
| 2 | Benito Juárez | 014 | 014 | 308479 | 2242.664 | 348568.4 |
| 3 | Coyoacán | 014 | 003 | 160059.1 | 4761.689 | 199866.9 |
| 4 | Milpa Alta | 014 | 009 | 2302.26 | 23.909 | 2425.825 |
| 5 | Tláhuac | 014 | 011 | 14396.53 | 41.036 | 18059.01 |
| 6 | Tlalpan | 014 | 012 | 160792.1 | 4333.975 | 189795.8 |

En este ejemplo se etiqueto con el nombre de cada alcaldía, incluyendo la clave del municipio y la de la entidad, con el fin de que se puedan emplear para rastrear los datos en la base original si ello fuese necesario.

En el nombre de la hoja se recomienda incluir el año y la fuente (CE, para Censo Económico). Esto puede ser muy útil si se trabajará con varios años y fuentes.



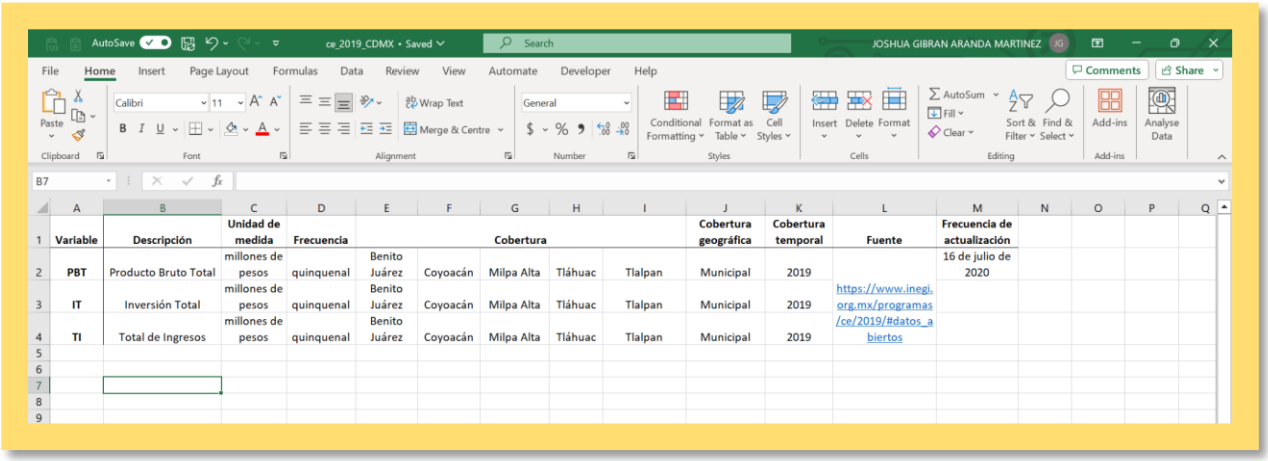
| Variable | Descripción | Unidad de medida | Frecuencia | Cobertura | | | | | Cobertura geográfica | Cobertura temporal | Fuente | Frecuencia de actualización |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PBT | Producto Bruto Total | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | https://www.inegi.org.mx/programas/ce/2019/#datos_abiertos | 16 de julio de 2020 |
| IT | Inversión Total | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | | |
| TI | Total de Ingresos | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | | |

Es importante incluir los metadatos relevantes a la información que se incluye en la base de datos, para que pueda consultarse con claridad. El metadato de la base de datos ya sería utilizable, para finalmente nombrar el archivo.

Para ello, cada institución o grupo de trabajo puede diseñar su nomenclatura estandarizada. Si ésta existe, será necesario respetarla, y si no la hay, es recomendable que se diseñe una:

El nombre del archivo de metadato se puede nombrar como: ce_2019_CDMX.xlsx, incluyendo así el nombre de la fuente, el año y el área geográfica.

Finalmente se recomienda no usar ningún acento, espacio o carácter especial, pues éstos podrían no ser legibles para todos los ordenadores y sistemas.



| Variable | Descripción | Unidad de medida | Frecuencia | Cobertura | | | | | Cobertura geográfica | Cobertura temporal | Fuente | Frecuencia de actualización |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PBT | Producto Bruto Total | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | https://www.inegi.org.mx/programas/ce/2019/#datos_abiertos | 16 de julio de 2020 |
| IT | Inversión Total | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | | |
| TI | Total de Ingresos | millones de pesos | quinquenal | Benito Juárez | Coyoacán | Milpa Alta | Tláhuac | Tlalpan | Municipal | 2019 | | |

# Referencias


- Amazon, "¿Qué es Python?", en https://aws.amazon.com/es/what-is/python/, consultado el 4 de octubre de 2023.

- Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO). (s.f.). Laguna de Términos-Pantanos de Centla. Consultado en http://www.conabio.gob.mx/conocimiento/regionalizacion/doctos/rhp_090.html

- Consejo Nacional de Población (CONAPO). (2020a). Índice de Marginación por municipio 1990-2015. Consultado en https://www.gob.mx/conapo/documentos/indices-de-marginacion-1990-2010

- Consejo Nacional de Población (CONAPO). (2020b). Índice de marginación por municipio. Consultado en https://www.gob.mx/conapo/documentos/indices-de-marginacion-2020-284372

- Consejo Nacional de Población (CONAPO). (2020c). Índice de Marginación por Localidad, 2020. Consultado en https://datos.gob.mx/busca/dataset/indice-de-marginacion-carencias-poblacionales-por-localidad-municipio-y-entidad/resource/a6d095ec-ba03-4549-ad4a-367a5941dc07

- Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL). (2020). Pobreza a nivel municipio 2010-2020. Consultado en https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx

- Instituto Nacional de Estadística y Geografía (INEGI). (2016). Estudio de información integrada de la Cuenca Laguna de Términos y otras. Consultado en: https://www.inegi.org.mx/contenido/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825087456_1.pdf

- Instituto Nacional de Estadística y Geografía (INEGI). (2019). Censos Económicos 2019. Recuperado de https://www.inegi.org.mx/programas/ce/2019/

- Instituto Nacional de Estadística y Geografía (INEGI). (2020). Censo de Población y Vivienda 2020. Consultado en: https://www.inegi.org.mx/programas/ccpv/2020/

- Instituto Nacional de Estadística y Geografía (INEGI). (2023). Encuesta Nacional de Ocupación y Empleo (ENOE) - 15 años y más. Consultado en: https://www.inegi.org.mx/programas/enoe/15ymas/#Tabulados

- CONABIO, "90. Laguna de términos - Pantanos de Centla", Consultado en: http://www.conabio.gob.mx/conocimiento/regionalizacion/doctos/rhp_090.html,

- INEGI, "Censos Económicos 2019", tabulados en línea. Consultado en: https://www.inegi.org.mx/programas/ce/2019/#Datos_abiertos

- INEGI, "Indicadores Laborales para los municipios de México", tabulados en línea, Consultado en: https://www.inegi.org.mx/programas/ilmm/#Datos_abiertos

- Klein, J. (2008). Evaluation of Interdisciplinary and Transdisciplinary Research. A literature Review. American Journal of Preventive Medicine, 116-123.

- SIAP, “Anuario Estadístico de la Producción Agrícola”, tabulados en línea. Consultado en: https://nube.siap.gob.mx/cierreagricola/

- SIAP, “Anuario Estadístico de la Producción Ganadera”, tabulados en línea. Consultado en: https://nube.siap.gob.mx/cierre_pecuario/