Edder Gerardo Hernández Jiménez

# CAPÍTULO I. PROBLEMATIZACIÓN DEL OBJETO DE ESTUDIO

## 1.1 Planteamiento del problema

## El auge del Data Science ha generado una alta demanda de profesionales en este campo, impulsada por el crecimiento exponencial de datos desde la pandemia. Sin embargo, en el estado de Chiapas, la oferta formativa en esta área es limitada y poco especializada, en contraste con entidades como Ciudad de México, Jalisco o Nuevo León. Esta carencia impide aprovechar herramientas modernas para el análisis de datos públicos como la ENOE, restringiendo el desarrollo de soluciones innovadoras ante problemáticas estructurales como la informalidad laboral. Ante esta brecha, surge la necesidad de aplicar técnicas de Data Science para analizar el mercado laboral mexicano y extraer conocimiento útil para la toma de decisiones.

## 1.2 Objetivos

### 1.2.1 Objetivo general

Identificar patrones y segmentos de población que enfrentan dificultades en el mercado laboral mexicano, a partir de la aplicación de metodologías del *Data Science* a los datos de la Encuesta Nacional de Ocupación y Empleo (ENOE) de México.

### 1.2.2 Objetivos específicos

1. Diseñar un flujo de trabajo (pipeline) en R que integre técnicas del Data Science adaptadas a las limitaciones de los microdatos de la ENOE.
2. Describir los principales métodos de análisis de datos utilizados en *Data Science*, que se pueden aplicar en el estudio del mercado laboral en México.
3. Aplicar técnicas de análisis exploratorio a los datos de la ENOE, análisis de regresión y clustering, para identificar patrones y grupos de población con mayores dificultades en el mercado laboral en México.

### 1.2.3 Preguntas de investigación

1. ¿Qué es el *Data Science*?
2. ¿Cuáles son los principales métodos de análisis de datos utilizados en *Data Science*, que se pueden aplicar en el estudio del mercado laboral en México?
3. ¿Cuáles son los grupos de población que enfrentan mayores dificultades y limitaciones en el mercado laboral en México a partir del análisis realizado a los datos de la ENOE?

## 1.3 Justificación

La ENOE proporciona datos esenciales sobre empleo, desempleo, ingresos y condiciones laborales en México, a partir de una muestra representativa a nivel nacional. Esta información permite identificar grupos vulnerables dentro del mercado laboral, como aquellos sin acceso a empleos formales. Dado que el mercado laboral influye directamente en la economía, la productividad y la calidad de vida, contar con datos actualizados es clave para reducir desigualdades y tomar decisiones informadas. En este contexto, el Data Science se presenta como una herramienta eficaz para analizar grandes volúmenes de datos, detectar patrones complejos y desarrollar modelos predictivos que ayuden a comprender y mejorar las condiciones del mercado laboral mexicano.

## 1.4 Limitaciones del estudio

El presente trabajo de investigación se delimita temporalmente en el periodo de julio a diciembre de 2023. El enfoque principal de la investigación se fundamenta en una revisión documental exhaustiva, respaldada por un análisis descriptivo cuantitativo, es decir, aplicar técnicas del *Data Science* a los datos recopilados por la Encuesta Nacional de Ocupación y Empleo (ENOE) de México, la cual es aplicada de manera trimestral. Por lo tanto, se analizarán los datos recopilados durante el III y IV trimestre del año 2023, es decir, un semestre, con el propósito de identificar patrones y segmentos en el mercado laboral de la población.

# CAPÍTULO 4. DISEÑO METODOLOGICO

## 4.1 Enfoque de investigación

### 4.1.1 Enfoque adoptado

## La investigación adopta un enfoque híbrido que combina una revisión documental exhaustiva (de naturaleza cualitativa) con un análisis descriptivo cuantitativo. Esta elección metodológica responde a los objetivos del proyecto y a la naturaleza estadística de los datos proporcionados por la ENOE. La revisión documental sirve como base conceptual, mientras que el análisis cuantitativo permite identificar patrones y relaciones significativas dentro del mercado laboral.

## 4.2 Población y Muestra

### 4.2.1 Población objetivo por ENOE

Con relación al INEGI, a través de la ENOE (2022), se detalla que su población objetivo está conformada por las personas residentes en viviendas particulares dentro del territorio nacional mexicano. Por otro lado, atendiendo a las características económicas, se recopilan datos de la población de 12 años y más, aunque los indicadores que se difunden mensual y trimestralmente se generan para la población de 15 años y más.

### 4.2.2 Muestra ENOE

#### 4.2.2.1 Características

#### La ENOE es una encuesta por muestreo representativa de la población mexicana, basada en una selección aleatoria estratificada de hogares y personas. Cada hogar seleccionado proporciona información sobre todos sus integrantes, asegurando una cobertura geográfica equilibrada mediante el uso de ponderadores.

#### 4.2.2.2 Limitantes de la muestra

# No obstante, la muestra presenta limitaciones inherentes a su diseño. Existen posibles errores de muestreo y no respuesta, debido a la naturaleza probabilística y la posibilidad de que algunos individuos no participen. Además, al tratarse de una encuesta auto reportada, los datos pueden estar sujetos a sesgos personales o imprecisiones en la memoria del encuestado. Estas limitaciones deben ser consideradas al analizar e interpretar los resultados derivados de la ENOE.

# CAPÍTULO 5. ANÁLISIS DEL MERCADO LABORAL MEXICANO

## 5.1 Análisis de la ENOE

Es importante comprender que la Encuesta Nacional de Ocupación y Empleo (ENOE) es una fuente valiosa de datos que se utiliza para analizar el mercado laboral mexicano. La ENOE contiene datos relacionados con la población económicamente activa en México. En particular, se enfoca en recopilar información sobre el empleo y el desempleo en el país. Los datos incluyen información demográfica como edad, género, nivel educativo, así como datos sobre la situación laboral de las personas, incluyendo su ocupación, ingresos, sector económico, entre otros (ENOE, 2022).

Asimismo, el INEGI a través de la ENOE busca cuantificar la población de 15 años y más, en relación con su clasificación de dos grupos mutuamente excluyentes: Población económicamente activa (PEA) y la población no económicamente activa (PNEA), al mismo tiempo estos dos grupos se subdividen en otros 2 grupos de forma individual, como se plasma en la figura 5.1.

Figura 5.1 Distribución de la población en la ENOE



Nota: Elaboración propia basada en el INEGI (ENOE).

La diferencia clave entre los cuatro grupos radica en su participación en la fuerza laboral y su disposición para trabajar, es posible identificar a los 4 grupos como:

* Ocupados: Personas que tienen un empleo remunerado y están trabajando activamente.
* Desocupados: Personas que están sin trabajo, pero están buscando activamente empleo y están disponibles para trabajar.
* Disponibles: Aquellos que no están trabajando en el momento de la encuesta, pero están dispuestos a trabajar si surge la oportunidad, aunque no están buscando activamente empleo.
* No disponibles: Personas que no están trabajando y tampoco están buscando empleo activamente en el momento de la encuesta, debido a razones como estudios, jubilación, responsabilidades familiares u otras circunstancias personales.

### 5.1.1 Estructura de la base de datos

Los cuestionarios dados por la ENOE se aplican de manera trimestral a lo largo de los años y su información se subdivide en 5 tablas como se describe en la tabla 5.1, las cuales son accesibles a través de su sitio web https://www.inegi.org.mx/programas/enoe/15ymas/#microdatos.

Tabla 5.1 Tablas de la ENOE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tabla de datos | Descripción | Abreviatura | Trimestre | Año |
| VIVT123.CSV | Tabla de vivienda | VIVT | 1 | 2023 |
| HOGT123.CSV | Tabla de hogares | HOGT | 1 | 2023 |
| SDEMT123.CSV | Tabla de sociodemográfico | SDEMT | 1 | 2023 |
| COE1T123.CSV | Tabla de cuestionario de ocupación y empleo I | COE1T | 1 | 2023 |
| COE2T123.CSV | Tabla de cuestionario de ocupación y empleo II | COE2T | 1 | 2023 |

Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

Teniendo en cuenta a la ENOE (2022), sus tablas juegan un papel importante a la hora de subdividir la información con relación al mercado laboral del pueblo mexicano.

1. Tabla de Vivienda: Esta tabla contiene información sobre las viviendas seleccionadas en la muestra, incluyendo su identificación y número de hogares que la conforman. La tabla de vivienda se relaciona con la tabla de hogar a través del número de hogares que conforman cada vivienda.
2. Tabla de Hogar: Esta tabla contiene información sobre los hogares seleccionados en la muestra, incluyendo el número de personas que conforman cada hogar. La tabla de hogar se relaciona con la tabla sociodemográfico a través de la información sobre la población que conforma cada hogar.
3. Tabla Sociodemográfico: Esta tabla contiene información sobre la población, incluyendo su edad, sexo, estado civil, nivel educativo, entre otros. La tabla sociodemográfica se relaciona con las tablas de cuestionario de ocupación y empleo parte I y parte II a través de la información sobre la población económicamente activa y no económicamente activa.
4. Tabla de Cuestionario de Ocupación y Empleo (COE) parte I y II: Estas tablas contienen información sobre la situación económica de la población, incluyendo su situación laboral, sector económico, ingresos, entre otros.

Asimismo, dentro del conjunto de las presentes tablas, es importante considerar que las tablas COE1 y COE2 en conjunto, presentan una variación entre ampliado y básico, esta variación significa que una será más corta y la otra más extensa. Los periodos trimestrales específicos de aplicación para estas variaciones se pueden observar en la tabla 5.2, es muy importante tener a consideración esta información debido a las limitaciones del estudio que se estableció desde un principio.

Tabla 5.2 Cuestionario ampliado y básico por trimestres.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AÑO | T1 | T2 | T3 | T4 |
| 2005 | Ampliado | Ampliado | Ampliado | Ampliado |
| 2006 | Ampliado | Ampliado | Básico | Básico |
| 2007 y 2008 | Básico | Ampliado | Básico | Básico |
| 2009 a la fecha | Ampliado | Básico | Básico | Básico |

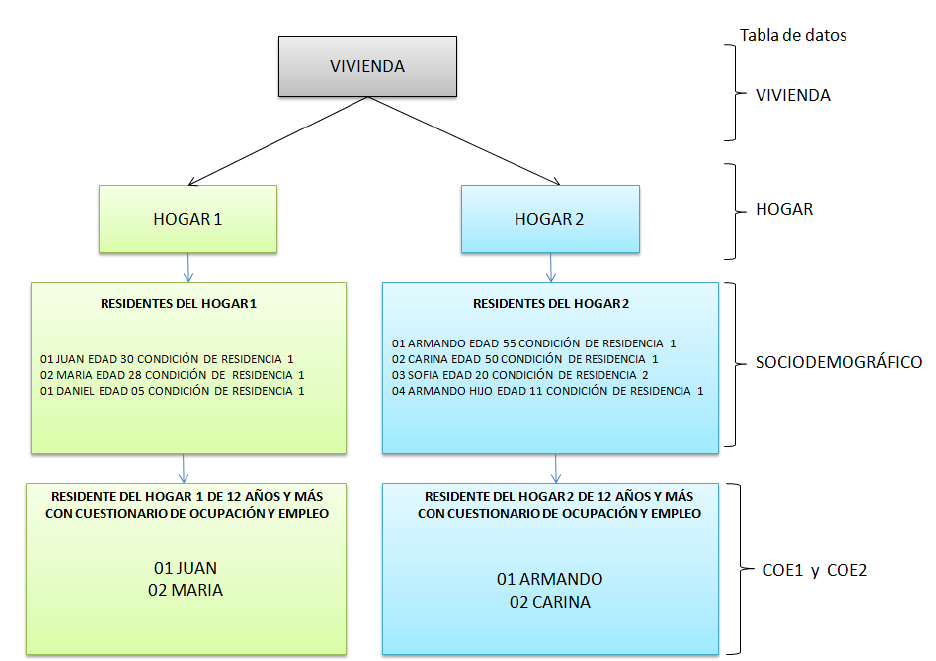
Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

No menos importante, es necesario explicar la secuencia de aplicación de cada uno los cuestionarios y su respectico almacenamiento. Como plantea la ENOE (2023):

Por cada vivienda existe uno o más hogares y por cada hogar con entrevista lograda se tiene uno o más residentes (características sociodemográficas), es decir, para cada hogar se tiene el registro de sus residentes en la tabla de sociodemográfico (SDEM) y a su vez, para cada residente con edad de 12 años cumplidos o más, se tiene un registro del Cuestionario de Ocupación y Empleo, el cual se almacena en las tablas COE1 y COE2. (Registros, párr. 2)

Conforme a lo anterior, la ENOE indica que el almacenamiento de los datos obtenidos a través de sus cuestionarios se organiza en las cinco tablas siguiendo el orden de la figura 5.2:

Figura 5.2 Secuencia de registros



Nota: ENOE (2023). Extraído de https://acortar.link/XYR5Tf

Cabe resaltar que, en algunas ocasiones no se logra realizar la entrevista, lo que hace que la secuencia de datos termine en la tabla de hogar. Además, hay casos en los que se obtiene la entrevista, pero si los integrantes del hogar están ausentes (condición de residencia 2) o son menores de 12 años, la secuencia se detiene en la tabla sociodemográfica. En caso contrario, la secuencia continúa hasta las tablas COE1 y COE2. Este detalle es importante, ya que permite determinar qué registros son "válidos" y cuáles no, en relación con el contexto de la población objetivo de la ENOE.

### 5.1.2 Estructura de los conjuntos de datos

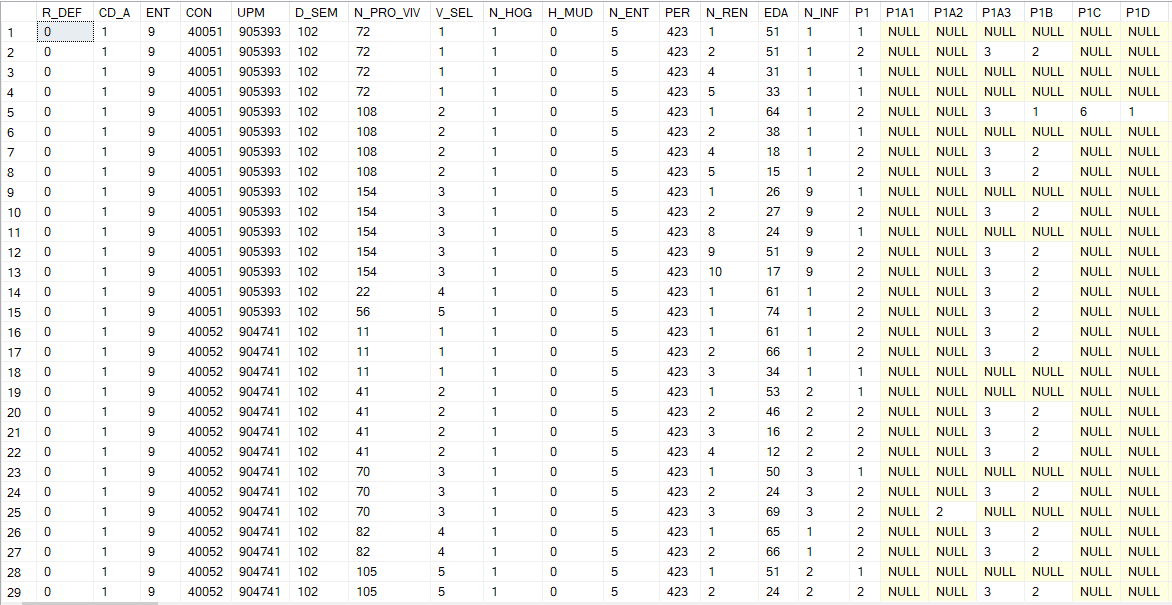
De acuerdo con la ENOE (2022), sus datos se recopilan mediante encuestas realizadas en viviendas particulares en todo el territorio nacional mexicano. Las entrevistas se llevan a cabo cara a cara con los residentes de estas viviendas. La población objetivo para esta encuesta son las personas residentes en viviendas particulares, y la información se obtiene a través de cuestionarios estructurados.

La información recopilada por los cuestionarios es almacenada en 5 tablas donde cada tabla contiene varias columnas que representan atributos específicos, como una pregunta y a su vez estas contienen múltiples filas que hacen referencia a los múltiples registros capturados como se logra observar en la figura 5.3.

Figura 5.3 Estructura del conjunto de datos\_1

Variables / Atributos / Columnas

De la tabla COE1



Registros / filas

De la tabla COE1

Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

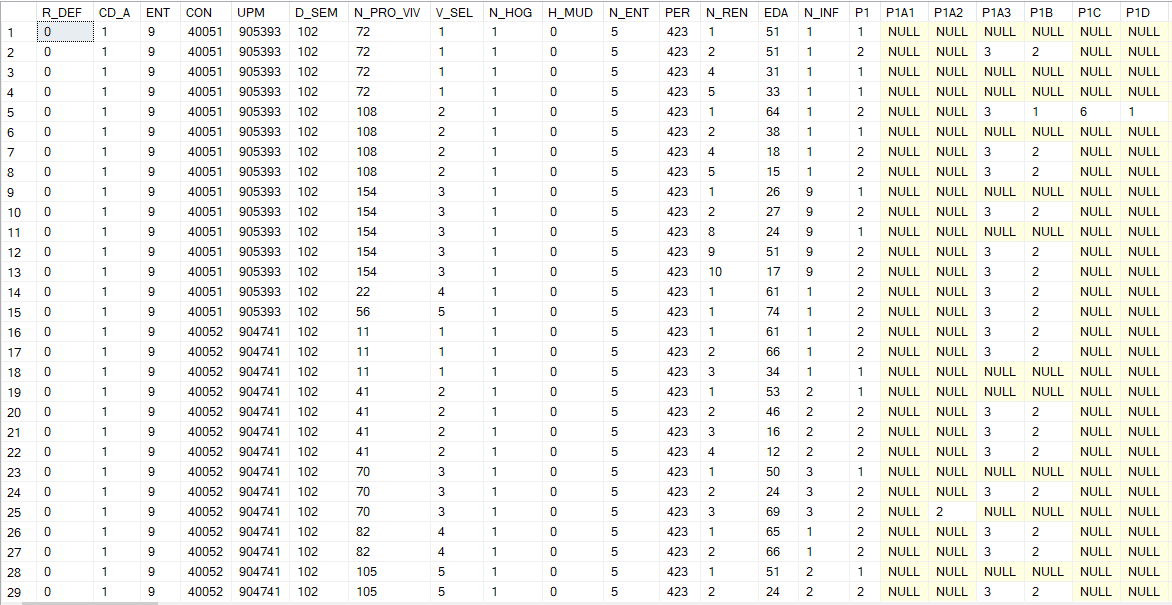
Para simplificar la estructura de las tablas y evitar que se vuelvan excesivamente complejas, el INEGI utilizó un sistema de clave-valor para representar columnas y filas. Por ejemplo, en la figura 5.4 se muestra la tabla COE1 con la variable "entidad", representada por la clave "ENT". En este caso, los registros almacenan el valor 9, que no es solo un número; en este contexto, representa una entidad específica: la "Ciudad de México".

Figura 5.4 Estructura del conjunto de datos\_2

Columna

Clave: ENT

Valor: Entidad



Registros

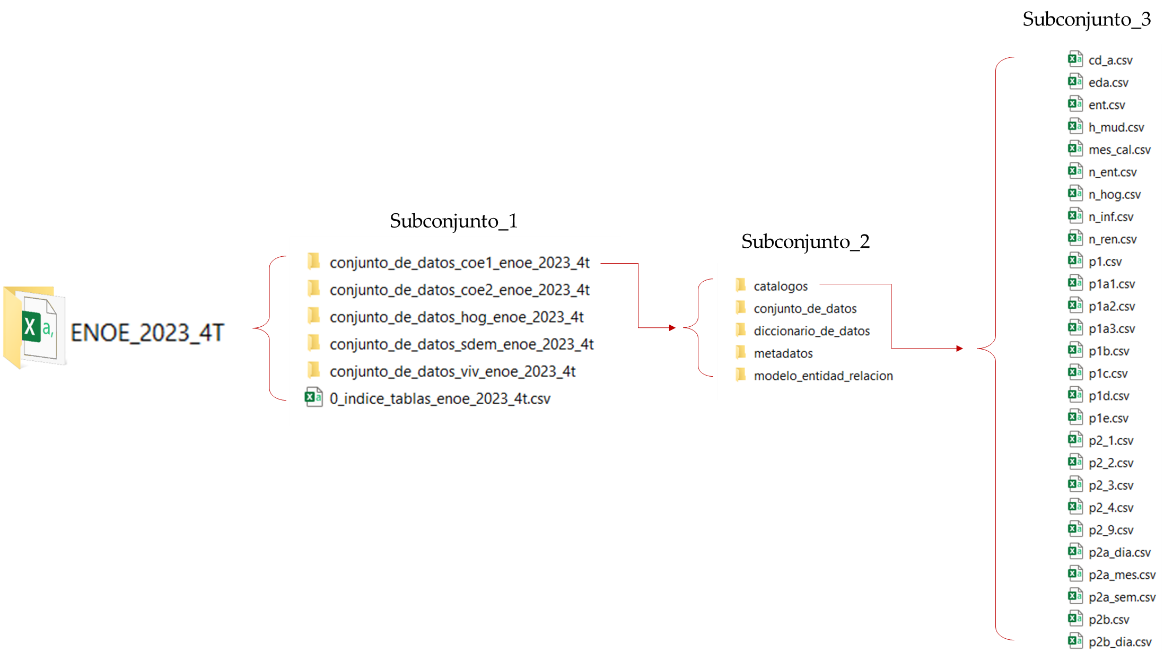
Clave: 9

Valor: Ciudad de México

Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

Estos pares de clave-valor para las columnas y filas están organizados en conjuntos de datos conocidos como catálogos de datos y diccionarios de datos, respectivamente. Por ello, es fundamental conocer el contenido del conjunto de datos de la ENOE para cada trimestre. En particular, el contenido de los archivos correspondientes al cuarto trimestre del año 2023 se refleja en la figura 5.5.

Figura 5.5 Estructura del conjunto de datos\_3



Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

En el ejemplo anterior, se reflejó el contenido de los archivos correspondientes a la tabla COE1 y sus catálogos, es decir, todos los archivos que contienen los pares clave-valor para los registros de dicha tabla. Es necesario explicar el contenido de cada subconjunto:

* Subconjunto 1: Contiene cinco carpetas, cada una destinada a una de las tablas de la base de datos, y un archivo Excel que describe simplemente el nombre de cada tabla.
* Subconjunto 2: Incluye el contenido de cada tabla de la base de datos. Es importante destacar que este subconjunto es idéntico para todas las tablas. En este ejemplo específico, se aborda el subconjunto 2 de la tabla COE1, el cual contiene cinco carpetas, cada una con un propósito específico, como se muestra en la tabla 5.3.
* Subconjunto 3: Contiene de 1 a N archivos, dependiendo de la carpeta seleccionada del subconjunto 2. La cantidad de archivos varía según el propósito específico de cada carpeta seleccionada previamente.

Tabla 5.3 Contenido\_Subconjunto 2.

|  |  |
| --- | --- |
| Subconjunto 2 | Descripción |
| catalogos | Contiene los pares clave-valor de los registros (filas) |
| conjunto\_de\_datos | Contiene todos los registros en un formato  de tipo *Dataframe* |
| diccionrio\_de\_datos | Contiene los pares clave-valor de las columnas (cabeceras)) |
| metadatos | Contiene información extra (*Feedback*) |
| modelo\_entidad\_relacion | Contiene el modelo ER de la base de datos |

Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

### 5.1.3 Indicadores estratégicos

En el contexto de la ENOE (2022), los indicadores estratégicos son métricas que permiten filtrar de manera clara y precisa distintos aspectos de la población mexicana. Para definir y analizar estos indicadores, la ENOE utiliza criterios específicos que permiten determinar cómo se agrupan las distintas categorías de la población mexicana. Los criterios de selección y las condiciones para cada variable se detallan en el documento “Reconstrucción de variables de la ENOE”.

Un ejemplo de estos indicadores es el que permite identificar a la población en sus agrupaciones principales: la población total, la población menor de 15 años y la población de 15 años y más. Para utilizar los indicadores estratégicos, es necesario aplicar los criterios especificados por el INEGI, como se observa en la tabla 5.4.

Tabla 5.4 Criterios de agrupaciones principales

|  |  |
| --- | --- |
| Población | Criterio |
| Población total | R\_DEF=00 y (C\_RES=1 o 3) |
| Población menor de 15 años | R\_DEF=00 y (C\_RES=1 o 3) y  ((EDA>=00 y EDA<=15) o EDA=99) |
| Población de 15 años y más | R\_DEF=00 y (C\_RES=1 o 3) y  (EDA>=15 y EDA<=98) |

Nota: Elaboración propia basada en la documentación técnica de la ENOE (2023).

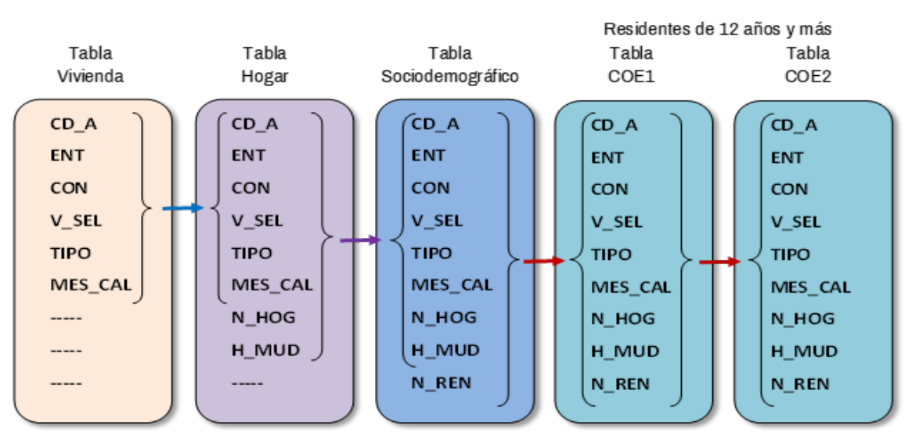
Los criterios para los indicadores estratégicos son reglas o especificaciones que ayudan a identificar quiénes se incluyen en un grupo y quiénes no. La sintaxis reflejada en la tabla anterior puede parecer compleja, ya que estas especificaciones deben aplicarse en el entorno de trabajo, o más precisamente, en el software particular que se esté utilizando. La sintaxis utiliza operadores lógicos, de asignación y de comparación para determinar qué registros se filtran en relación con una o más columnas, como se abordó en el capítulo 2 sobre programación.

Por ejemplo, los criterios para la población total se pueden interpretar como: "Todos los registros con entrevista completada (R\_DEF = 00) y, al mismo tiempo, todos los registros en los que la condición de residencia sea igual a residente habitual o nuevo residente (C\_RES = 1 o 3)". A medida que se crean más subconjuntos de grupos de población, los criterios aumentan significativamente e incluso se deben combinar criterios de un subconjunto con otro. Toda esta información se detalla a profundidad en el documento "Reconstrucción de variables de la ENOE".

## 5.2 Creación de la Base de Datos

Los documentos "ENOE\_Conociendo la base de datos" y "ENOE\_Estructura de la base de datos", proporcionados por el INEGI, detallan cada una de las 5 tablas y sus respectivas propiedades. La comprensión de estos documentos es fundamental, ya que permiten la identificación precisa de la estructura de las tablas, sus columnas, la nomenclatura utilizada, los tipos de datos manejados, los rangos admitidos y los posibles valores almacenados. Estos detalles son esenciales para comprender la cantidad y naturaleza de las variables, así como para identificar las claves primarias y foráneas, y entender la relación entre las tablas, como se observa en la figura 5.6.

Figura 5.6 Estructura BD de la ENOE

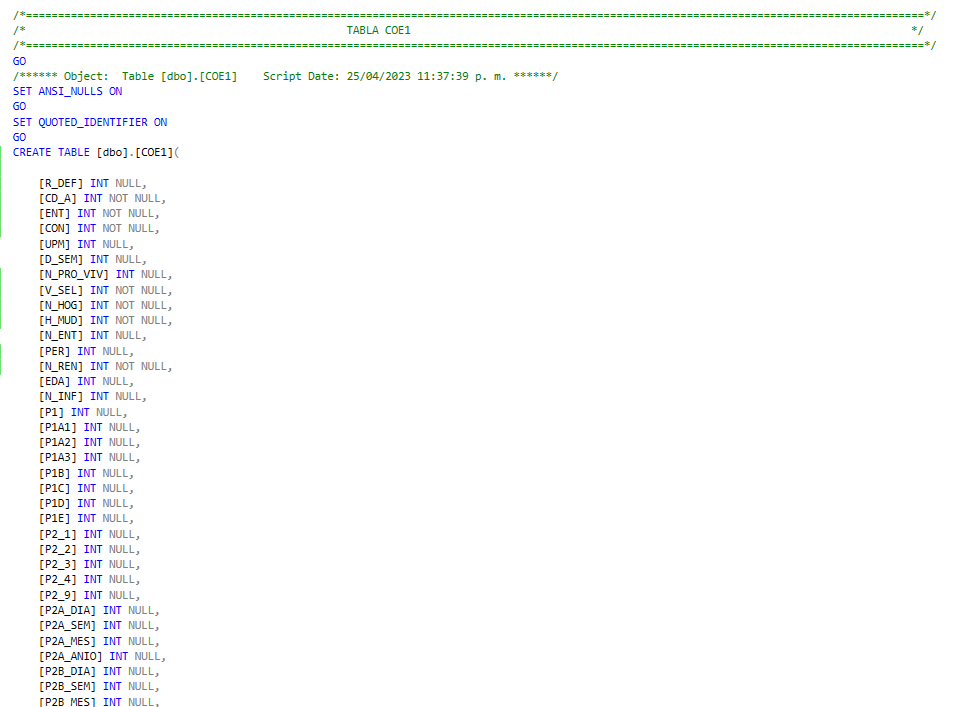
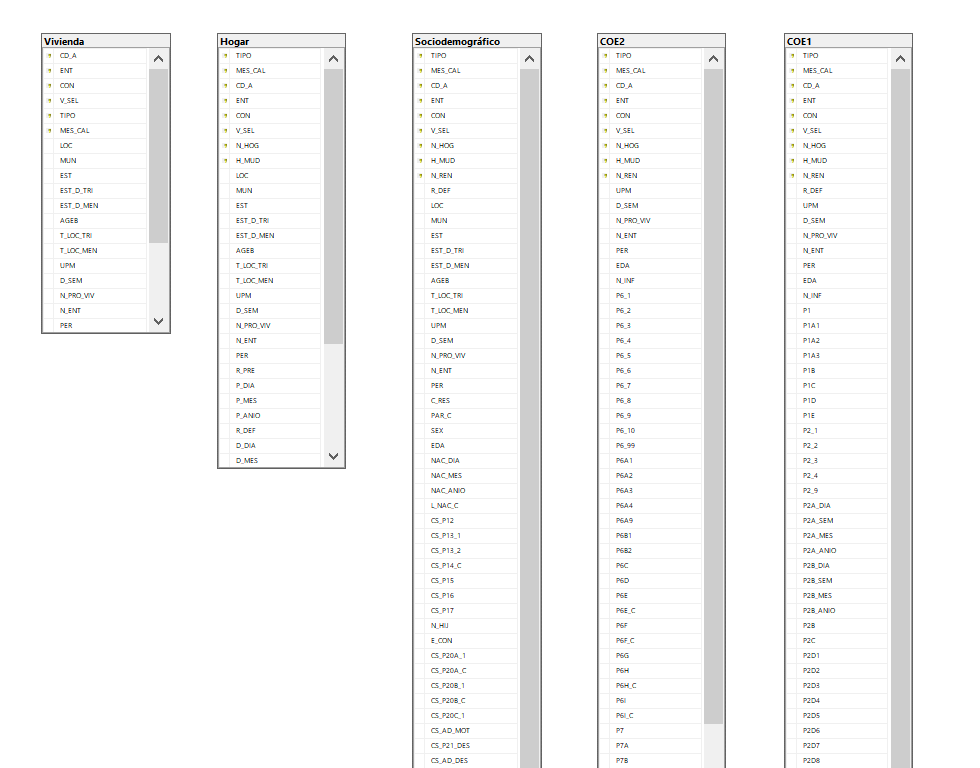


Nota: Figura 5.2 ENOE (2022). Consideraciones importantes (p. 3)

### 5.2.1 Creación de tablas

La creación de la base de datos se efectuó utilizando Microsoft SQL Server, un sistema de gestión de bases de datos relacionales. En conjunto con SQL Server Management Studio (SSMS), reconocido como una herramienta de gestión de bases de datos para SQL, se procedió a la creación de las tablas, sus respectivas columnas y sus claves primarias, como se muestra en la figura 5.7. Este procedimiento se repitió para cada una de las 5 tablas dentro de la ENOE, teniendo en cuenta las propiedades únicas de cada tabla y sus respectivas claves primarias distintas. Se utilizo la documentación pertinente y el análisis previo realizado en este apartado.

Figura 5.7 Creación de tablas

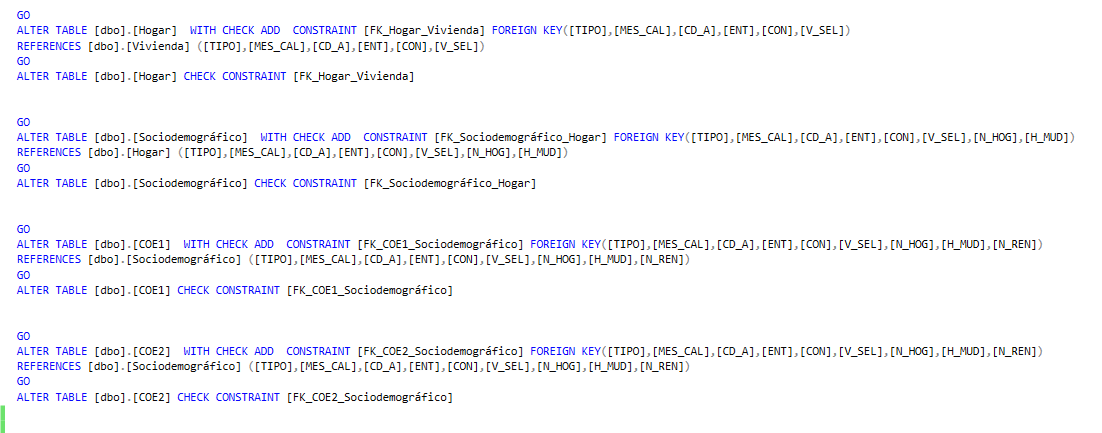
 

Nota: Elaboración propia con base a la documentación técnica de la ENOE.

### 5.2.2 Creación de relaciones

Después, se avanzó hacia la creación de las llaves foráneas, una acción que se puede observar claramente en la figura 5.8 del proceso. Estas llaves son esenciales porque permiten establecer relaciones entre las distintas tablas de la base de datos. Esto resulta fundamental para mantener la integridad de los datos y asegurar que la información esté conectada de manera coherente y significativa.

Figura 5.8 Creación de llaves foráneas



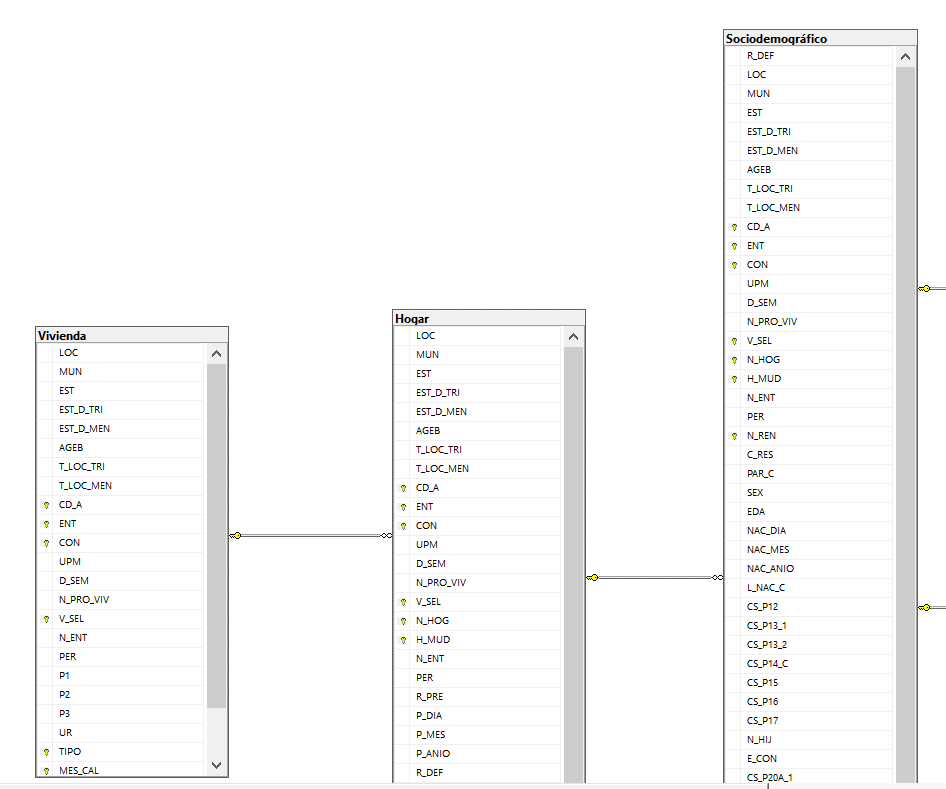
Nota: Elaboración propia con base a la documentación técnica de la ENOE.

Es importante no pasar por alto el tipo de relaciones que una base de datos relacional puede contener, tanto 1:1, 1:N y N:N, cada una cuenta con sus respectivas características que se abordaron en el apartado 2.2.3.2.2. Es posible reconocer que las primeras 3 tablas de la ENOE son una relación 1 a muchos (1:N), debido a la siguiente lógica:

* Tabla vivienda-hogar:
* Una vivienda puede contener uno o más hogares.
* Uno más hogares solo pueden pertenecer a una vivienda.
* Tabla hogar-sociodemográfico:
* Un hogar puede albergar uno o más residentes.
* Uno más residentes solo pueden pertenecer a un hogar.

Las relaciones representadas en la figura 5.9 se derivaron del análisis anterior. Además, se fundamentó con la descripción detallada proporcionada por la ENOE. Sin embargo, la documentación de la ENOE no profundiza en una perspectiva exhaustiva sobre el manejo de bases de datos relacionales, específicamente en lo que respecta al uso de llaves foráneas.

Figura 5.9 Relaciones 1:N



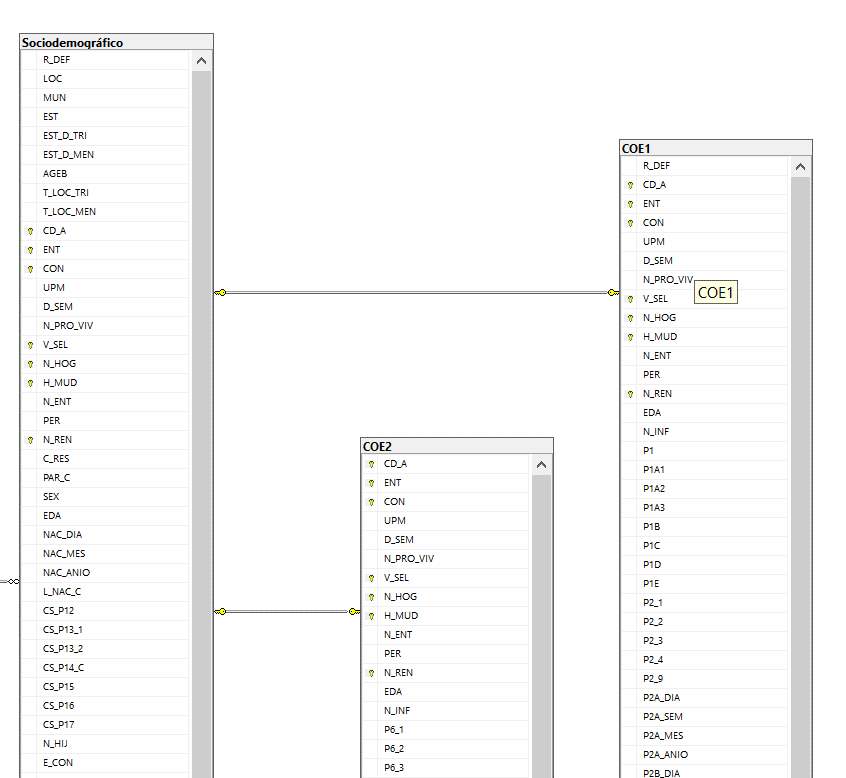
Nota: Elaboración propia con base a la documentación técnica de la ENOE.

Asimismo, se cuenta con relaciones 1 a 1 como se observa en la figura 5.10, en donde la lógica se centra en que:

* Un residente >= 15 años puede responder solo una vez el cuestionario COE1
* Un residente >= 15 años puede responder solo una vez el cuestionario COE2

Es importante tener en cuenta que las tablas COE1 y COE2 forman una sola tabla en conjunto. No obstante, el INEGI optó por manejarlas por separado debido a la gran cantidad de atributos que poseen respectivamente. Para los fines de esta investigación, se ha respetado esta estructura de trabajo establecida por el INEGI.

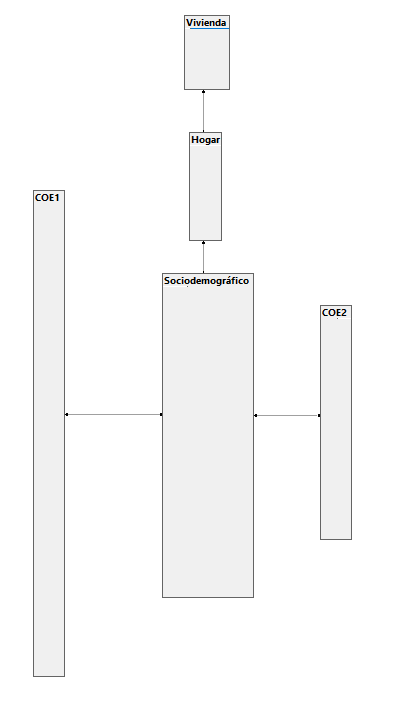
Figura 5.10 Relaciones 1:1



Nota: Elaboración propia con base a la documentación técnica de la ENOE.

Como última instancia, se procedió a la generación del diagrama E-R, el cual se muestra detalladamente en la figura 5.11. Este diagrama es una herramienta esencial que ofrece una visión entendible del comportamiento de la base de datos. A través de él, es posible identificar de manera clara y precisa qué tablas establecen comunicación directa entre sí y cuáles no.

Figura 5.11 Modelo E-R



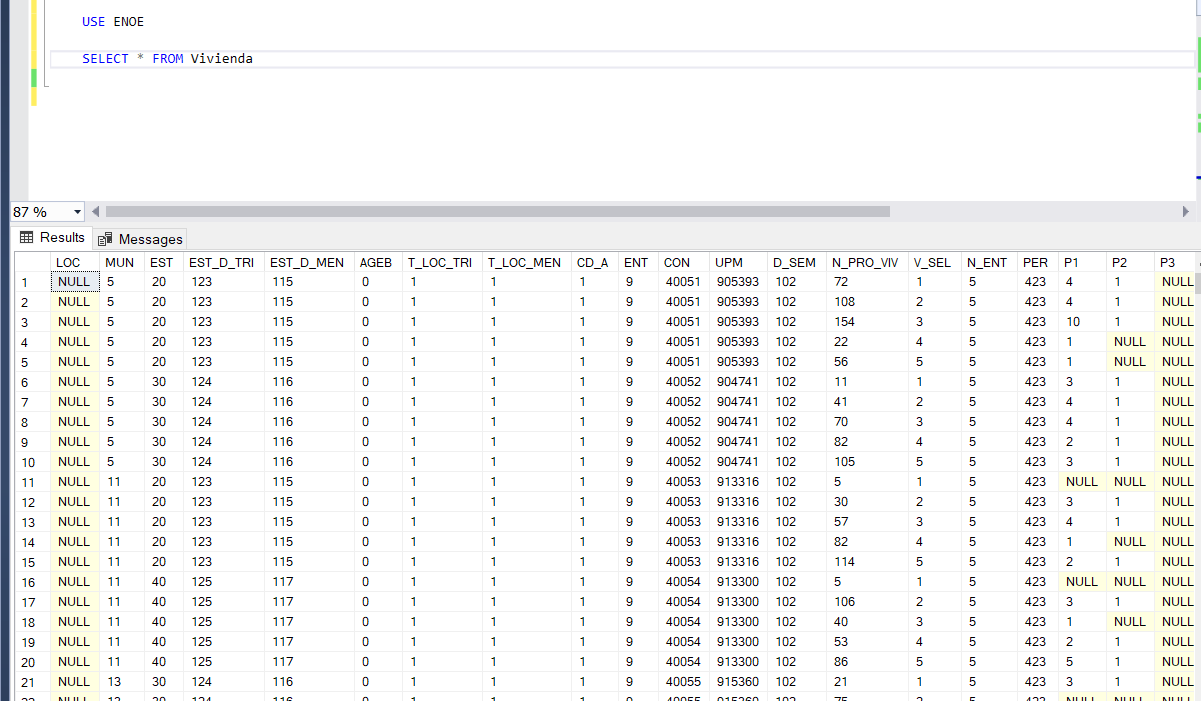
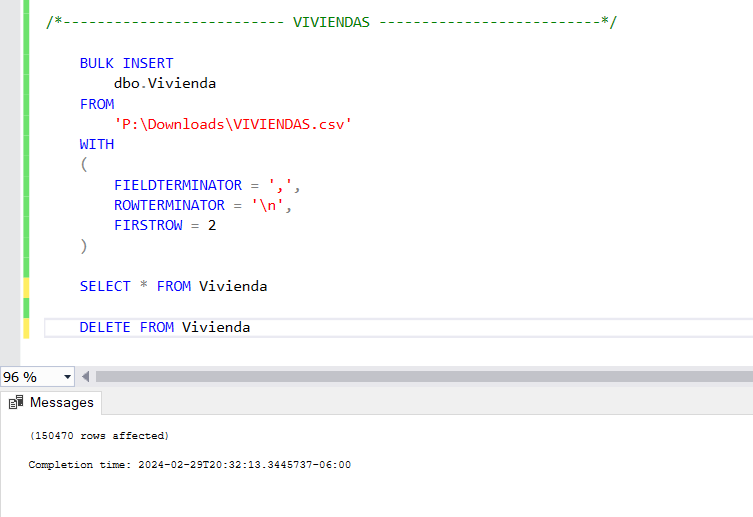
Nota: Elaboración propia con base a la documentación técnica de la ENOE.

### 5.2.3 Inserción de registros

La correcta creación de las tablas constituye un paso crucial en el proceso de registrar información en una base de datos. Es esencial considerar aspectos como el tipo de atributos manejados por cada tabla, tales como números enteros, decimales, fechas, cadenas de texto, entre otros. Dado que en el área del *Data Science* se suelen manejar grandes volúmenes de registros, es fundamental buscar métodos eficientes para su inserción. La ENOE ofrece sus registros en varios formatos; en este caso particular, se optó por trabajar con el formato .CSV debido a su flexibilidad y facilidad de uso, lo que permite un manejo más ágil y directo de los datos.

En SQL Server, la inserción masiva de datos mediante un archivo externo se llevó a cabo según lo ilustrado en la figura 5.12. Asimismo, se realizó una consulta para visualizar los registros, con el fin de confirmar que los 150,470 registros de viviendas correspondientes al cuarto trimestre del año 2023 se insertaron correctamente. Esta inserción se realizó con todas las tablas y los trimestres correspondientes que se analizan en el presente trabajo de investigación.

Figura 5.12 Insertar registros

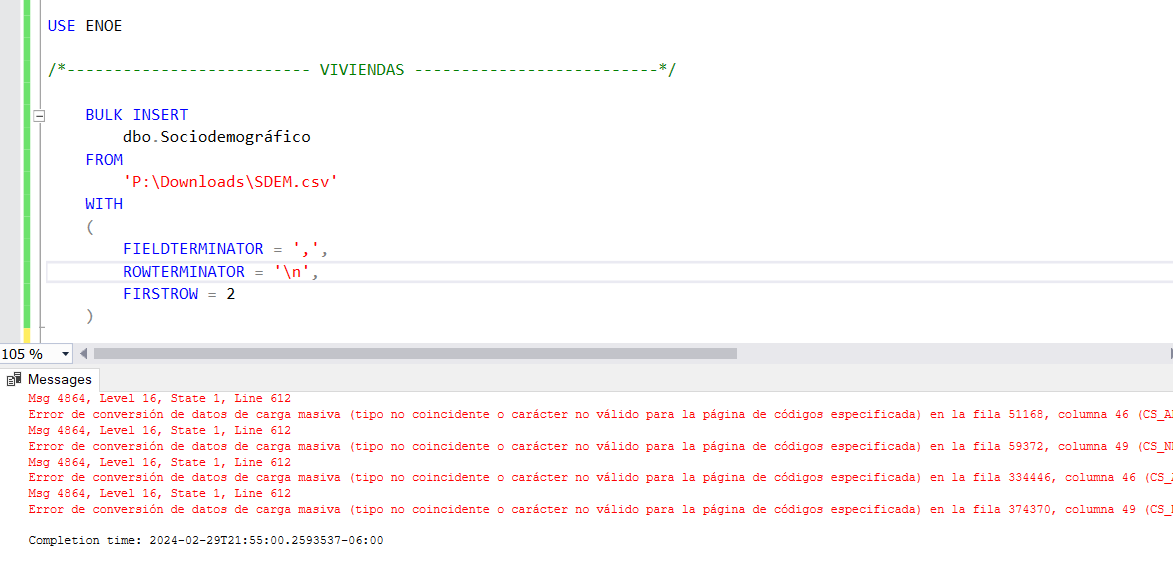


Nota: 150,470 inserciones registradas en la tabla Vivienda.

### 5.2.4 Tratamiento de los datos

Es frecuente encontrar errores como el que se muestra en la figura 5.13 al trabajar con la inserción de registros en bases de datos. En estas situaciones, la labor de un científico de datos adquiere relevancia, ya que es fundamental haber analizado previamente las variables que conforman la base de datos. Este análisis permite identificar si es necesario realizar ajustes como reemplazar algunos datos, modificar formatos o abordar posibles valores faltantes, entre otros aspectos. En el caso específico mencionado, se encontró un error de formato, donde los valores de una columna no coincidían con el formato esperado.

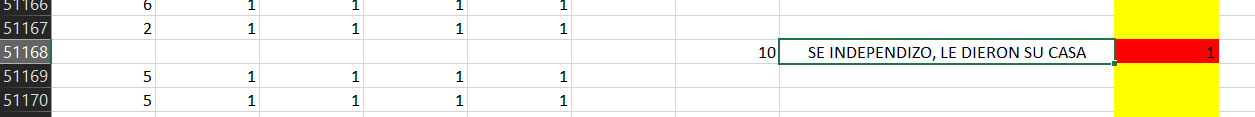
Figura 5.13 Errores de inserción



Nota: Elaboración propia.

Retomando el punto anterior, un científico de datos busca soluciones ante situaciones como la presente y analiza el contexto a fondo. Al explorar el error dentro de la figura 5.14, se observó que la falla dentro de la tabla sociodemográfico estaba en el registro 51,168, en la columna 46 "cs\_p21\_des". Sin embargo, al examinar la celda señalada como la responsable del error (celda roja), se constató que no había ningún problema aparente. Esto se debió a que el formato de la columna era numérico y el valor almacenado también lo era. En consecuencia, se procedió a analizar las columnas adyacentes en busca de pistas adicionales.

Figura 5.14 Análisis del error de inserción



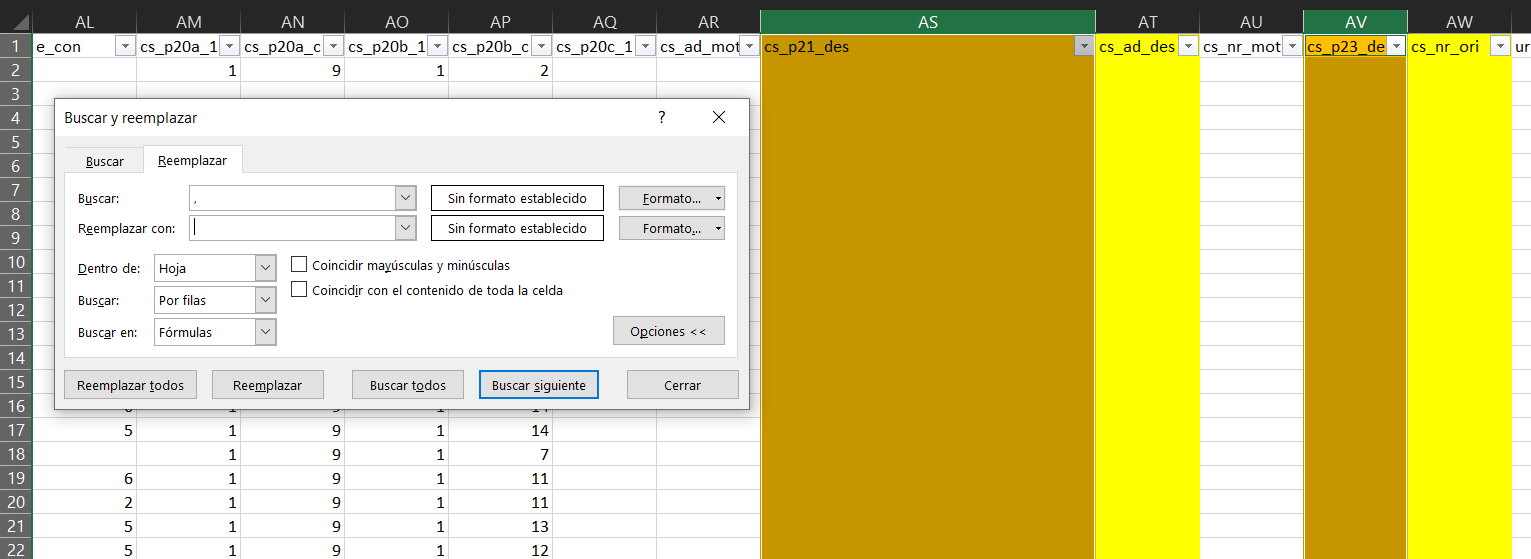
Nota: Elaboración propia.

Se descubrió que la columna vecina almacena información con formato de cadenas de texto, en donde sus valores contienen oraciones de distintos tipos, sin embargo, algunos valores manejan “,”, por ejemplo “Se independizo, le dieron su casa”, el problema radica en que la inserción por archivos .CSV toma en cuenta a las “,” como el delimitador entre columnas, por ello es muy importante manejar estos casos.

Este tipo de situaciones también ocurren en registros que almacenan caracteres especiales o muy largos, así que es fundamental hacer un proceso previo de limpieza y tratamiento de datos, este proceso está directamente relacionado con EDA.

El tratamiento inicial para estos datos fue sencillo como se observa en la figura 5.15 porque solo se reemplazó ciertos caracteres, es decir, se remplazó todas las “,” por un carácter vacío. Es importante resaltar que el tratamiento de datos siempre será variado y será de acuerdo al tipo de situación que se presenta.

Figura 5.15 Tratamiento de datos



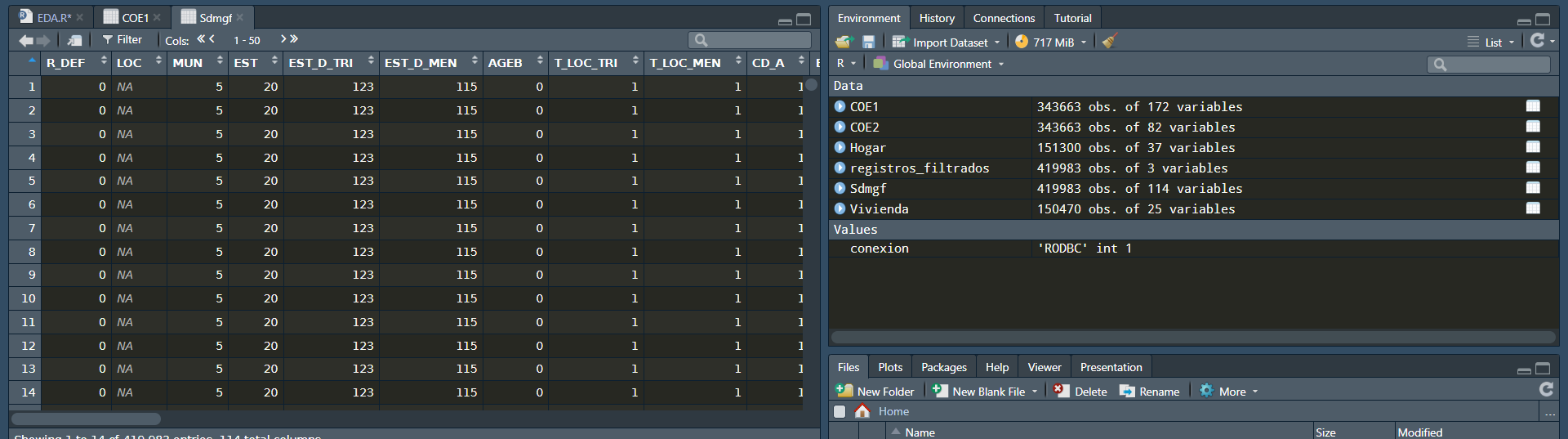
Nota: Elaboración propia.

## 5.3 Importación de Base de datos a R

En el presente código en R, se utilizó la paquetería "RODBC" para interactuar con la base de datos desde el entorno de R. Primero, se instaló la paquetería mediante install.packages("RODBC") y luego se activó con library(RODBC). Posteriormente, se estableció una conexión a la base de datos especificada, utilizando las credenciales proporcionadas. Se ejecutaron consultas SQL para extraer los datos de las diferentes tablas de la base de datos, como "Vivienda", "Hogar", "Sociodemográfico", "COE1" y "COE2", y se almacenaron los resultados en objetos de *dataframes* en R. Finalmente, se visualizó la información de cada tabla utilizando la función "*View*()", lo que facilitó la inspección y exploración de los datos de manera interactiva dentro de R. La paquetería "RODBC" proporciono una conexión con la base de datos relacional desde R, permitiendo ejecutar consultas SQL y extraer datos de manera eficiente para su análisis y visualización dentro del entorno de R.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21 | #Instalación de paqueteria  install.packages("RODBC")  #Ativar paqueteria  library(**RODBC**)  #Creación de la conexión a nuestra base de datos  conexion <- odbcConnect("conexionR",uid = "sa", pwd = "Password")  #Almacenamiento de las tablas a través de una consulta SQL  **Vivienda** <- sqlQuery(conexion, "SELECT \* FROM Vivienda")  **Hogar** <- sqlQuery(conexion, "SELECT \* FROM Hogar")  **Sdmgf** <- sqlQuery(conexion, "SELECT \* FROM Sociodemográfico")  **COE1** <- sqlQuery(conexion, "SELECT \* FROM COE1")  **COE2** <- sqlQuery(conexion, "SELECT \* FROM COE2")  #Visuaalización la información  **View**(**Vivienda**)  **View**(**Hogar**)  **View**(**Sdmgf**)  **View**(**COE1**)  **View**(**COE2**) |

Asimismo, es posible apreciar los objetos *dataframes* que se crearon y que ahora están almacenados dentro de R, como se puede apreciar en la figura 5.16.

**Figura 5.16 Tablas de SQL convertidas a *Dataframes*

Nota: 5 Tablas de SQL convertidas en *dataframes*.