

Gerardo Ulises Gonzalez Romero

**Materia:** Introducción a la Ciencia de  
Datos

**Nombre del Profesor:**  
Jaime Alejandro Romero Sierra

**20/10/2025**

[link de Base\\_sucio](#), [link de codigo de limpieza](#), [link de Base limpia](#).

Esta base de datos contiene información sobre el índice de criminalidad en México, esta base tiene como finalidad ser de apoyo para los departamentos policiales de México para su buen uso, en contra de los diferentes crímenes registrados y su respectivo combate de ello.

En esta base encontraremos diversas columnas, las cuales son de gran ayuda para análisis próximos. las columnas con las cuales contamos son las siguientes: Año, Clave\_Ent(Clave de Entidad), Bien jurídico afectado, Tipo de delito, Subtipo de delito, Modalidad, Meses del Año, sexo, Rango de edad.

De igual forma tenemos Renglones, de los 32 estados de México, de esta forma de se obtiene un fácil acceso a la información deseada.

No obstante, la base tiene valores repetidos, sin valor o mal catalogados etc.... aquí mismo se lleva el proceso de limpieza el cual es explicado en cada paso.

### **Significado de cada columna:**

**Año:** Año del crimen,

**Clave\_Ent:** La clave de entidad es una propiedad o un conjunto de propiedades de un tipo de entidad que se utiliza para determinar la identidad de una instancia dentro de un conjunto de entidades en un modelo de datos.

**Bien Jurídico:** El **bien jurídico** se refiere a aquellos bienes, tanto materiales como inmateriales, que son considerados de sumo valor

por la sociedad y que gozan de protección por parte del derecho, ya sea a través de otras ramas del derecho o específicamente del derecho penal.

**Tipo de delito:** Un delito se define como una conducta típica, antijurídica, imputable, culpable y sometida a una sanción penal, que constituye una infracción del derecho penal. Esta conducta puede manifestarse como una acción o una omisión, y debe estar prevista y penada por la ley.

**Subtipo de delito:** El término "subtipo de delito" no se encuentra directamente definido en los contextos proporcionados, pero se puede inferir su significado a partir de la clasificación detallada de los delitos según sus elementos y características.

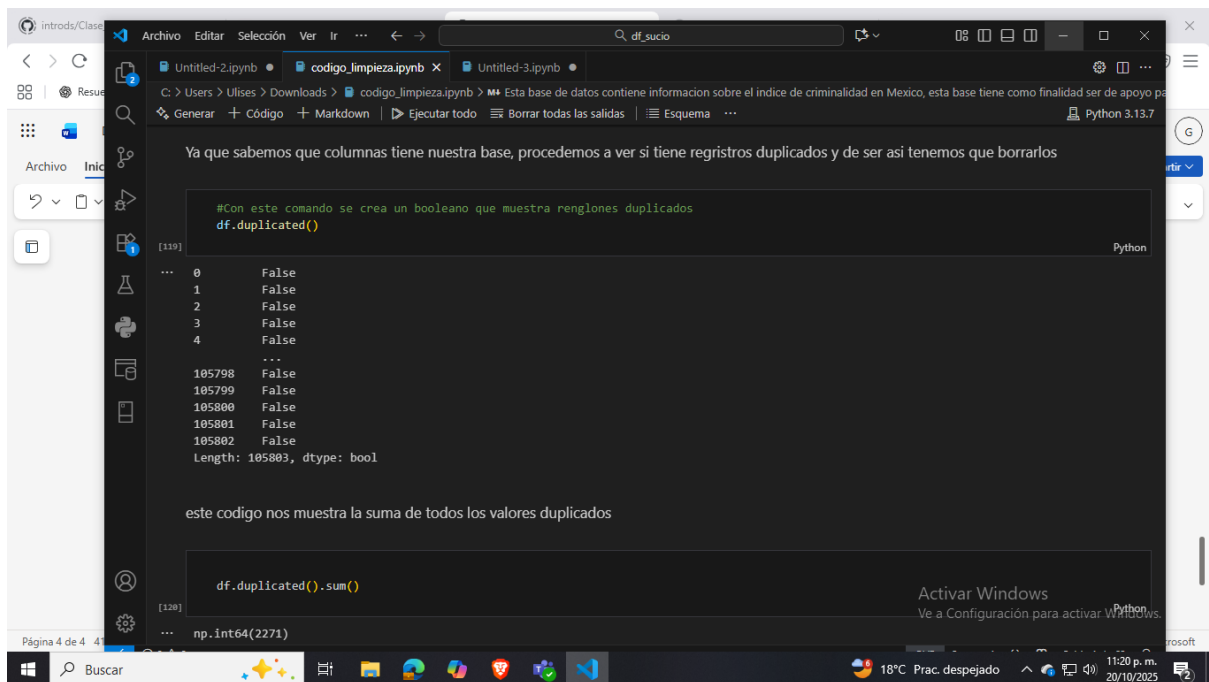
**Modalidad:** se refiere a las circunstancias de lugar, tiempo, modo u ocasión que califican la conducta del sujeto activo y que tienen como función determinar el quantum de la pena, afectando su medida, ya sea aumentándola o disminuyéndola.

**Meses del año:** Enero, febrero, marzo, abril, mayo, junio, Julio, agosto, septiembre, octubre, noviembre, diciembre

**Sexo:** Masculino/Femenino

**Rango de edad:** Menores de edad 0-17, Mayores de edad 18-100

Primeramente, lo primero que hice al limpiar la base de datos fue borrar los duplicados:



The screenshot shows a Jupyter Notebook interface with a dark theme. The notebook has three tabs: 'Untitled-2.ipynb', 'codigo\_limpieza.ipynb' (active), and 'Untitled-3.ipynb'. The active cell contains the following text and code:

Ya que sabemos que columnas tiene nuestra base, procedemos a ver si tiene registros duplicados y de ser así tenemos que borrarlos

```
#Con este comando se crea un booleano que muestra renglones duplicados
df.duplicated()
```

The output of the code is displayed below the cell:

```
...
0      False
1      False
2      False
3      False
4      False
...
105798  False
105799  False
105800  False
105801  False
105802  False
Length: 105803, dtype: bool
```

Below the output, there is another text block:

este código nos muestra la suma de todos los valores duplicados

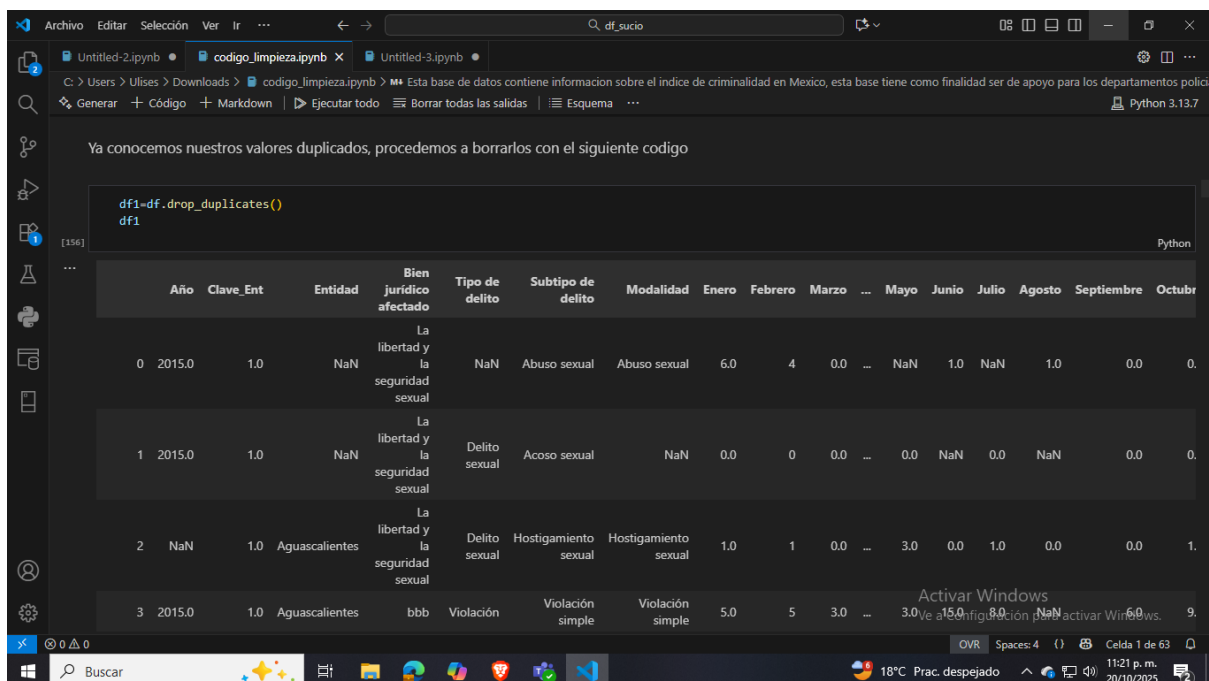
```
df.duplicated().sum()
```

The output of this code is:

```
...
np.int64(2271)
```

The Windows taskbar at the bottom shows the date as 20/10/2025 and the time as 11:20 p.m.

Después de ver que efectivamente contiene duplicados, cedi a la tarea de borrar todos



The screenshot shows the same Jupyter Notebook interface as the previous one. The active cell contains the following text and code:

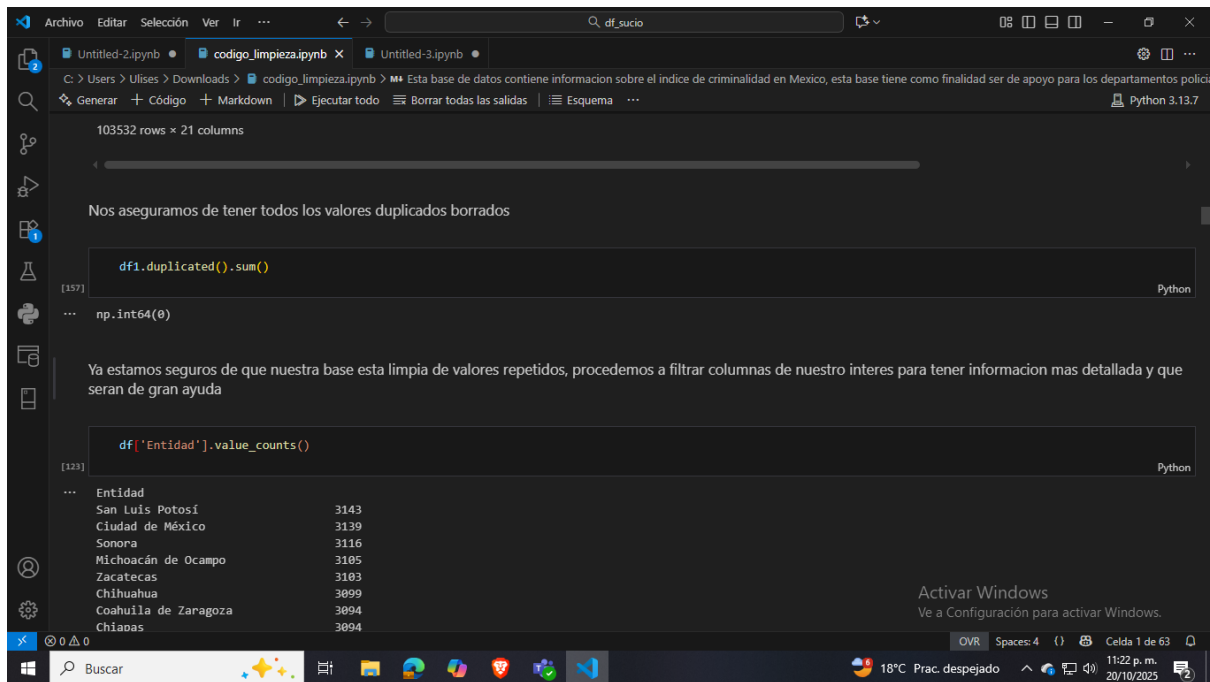
Ya conocemos nuestros valores duplicados, procedemos a borrarlos con el siguiente código

```
df1=df.drop_duplicates()
df1
```

The output of the code is a DataFrame with the following columns: Año, Clave\_Ent, Entidad, Bien jurídico afectado, Tipo de delito, Subtipo de delito, Modalidad, Enero, Febrero, Marzo, Mayo, Junio, Julio, Agosto, Septiembre, Octubre. The first few rows of the DataFrame are:

	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	Mayo	Junio	Julio	Agosto	Septiembre	Octubre
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4	0.0	NaN	1.0	NaN	1.0	0.0	0.
1	2015.0	1.0	NaN	La libertad y la seguridad sexual	Delito sexual	Acoso sexual	NaN	0.0	0	0.0	0.0	NaN	0.0	NaN	0.0	0.
2	NaN	1.0	Aguascalientes	La libertad y la seguridad sexual	Delito sexual	Hostigamiento sexual	Hostigamiento sexual	1.0	1	0.0	3.0	0.0	1.0	0.0	0.0	1.
3	2015.0	1.0	Aguascalientes	bbb	Violación	Violación simple	Violación simple	5.0	5	3.0	3.0	15.0	8.0	NaN	6.0	9.

The Windows taskbar at the bottom shows the date as 20/10/2025 and the time as 11:21 p.m.

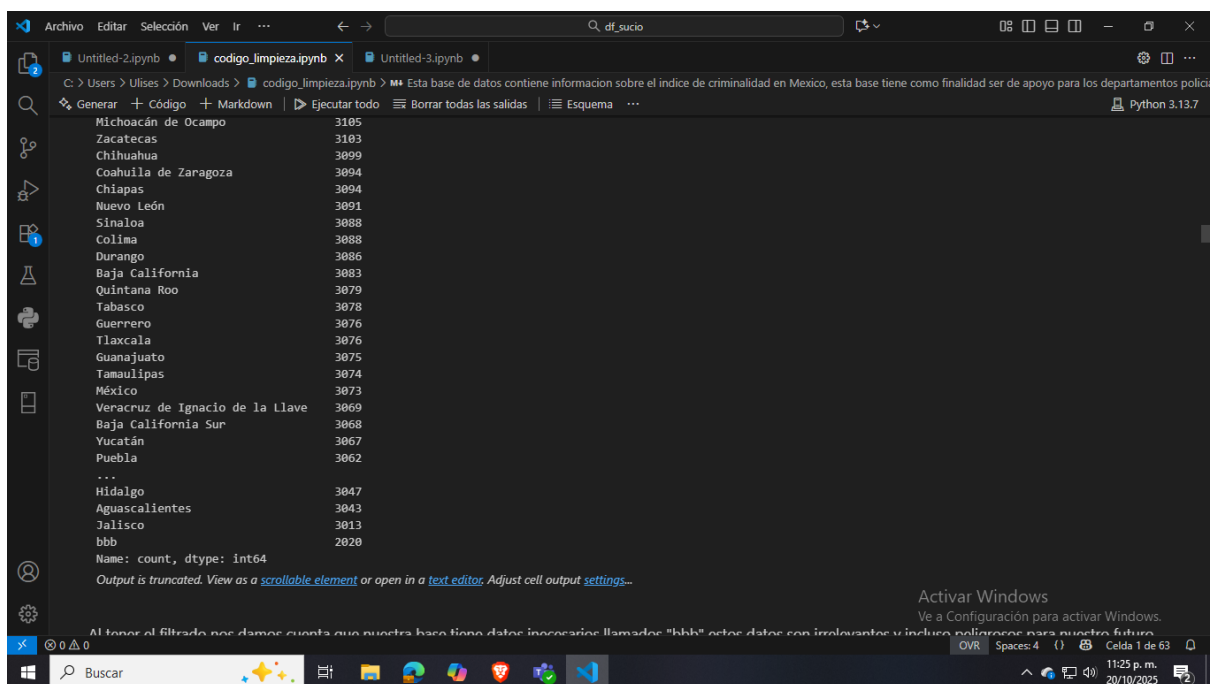


```
df1.duplicated().sum()
np.int64(0)

df['Entidad'].value_counts()
```

Entidad	Count
San Luis Potosí	3143
Ciudad de México	3139
Sonora	3116
Michoacán de Ocampo	3105
Zacatecas	3103
Chihuahua	3099
Coahuila de Zaragoza	3094
Chiapas	3094

Consecutivamente filtre algunas columnas y encuentre valores inválidos



```
df['Entidad'].value_counts()
```

Entidad	Count
Michoacán de Ocampo	3105
Zacatecas	3103
Chihuahua	3099
Coahuila de Zaragoza	3094
Chiapas	3094
Nuevo León	3091
Sinaloa	3088
Colima	3088
Durango	3086
Baja California	3083
Quintana Roo	3079
Tabasco	3078
Guerrero	3076
Tlaxcala	3076
Guanajuato	3075
Tamaulipas	3074
México	3073
Veracruz de Ignacio de la Llave	3069
Baja California Sur	3068
Yucatán	3067
Puebla	3062
...	...
Hidalgo	3047
Aguaascalientes	3043
Jalisco	3013
bbb	2020

Y procedí a borrar cada uno de cada columna donde hubiera uno

```
Archivo Editar Selección Ver Ir ... Q df_sucio
Untitled-2.ipynb • código_limpieza.ipynb X Untitled-3.ipynb •
C:\Users\Ulises\Downloads> código_limpieza.ipynb > Esta base de datos contiene información sobre el índice de criminalidad en Mexico, esta base tiene como finalidad ser de apoyo para los departamentos policia
Generar + Código + Markdown | Ejecutar todo | Borrar todas las salidas | Esquema Python 3.13.7

lista_col=df.columns
for nombre in lista_col:
    print(f"En la columna {nombre} los bbb son: {df[df[nombre] == 'bbb'].shape[0]}")

[162] Python

...
En la columna Año los bbb son: 0
En la columna Clave_Ent los bbb son: 0
En la columna Entidad los bbb son: 2020
En la columna Bien jurídico afectado los bbb son: 2012
En la columna Tipo de delito los bbb son: 0
En la columna Subtipo de delito los bbb son: 2016
En la columna Modalidad los bbb son: 0
En la columna Enero los bbb son: 0
En la columna Febrero los bbb son: 2017
En la columna Marzo los bbb son: 0
En la columna Abril los bbb son: 1996
En la columna Mayo los bbb son: 0
En la columna Junio los bbb son: 0
En la columna Julio los bbb son: 2013
En la columna Agosto los bbb son: 0
En la columna Septiembre los bbb son: 1994
En la columna Octubre los bbb son: 0
En la columna Noviembre los bbb son: 0
En la columna Diciembre los bbb son: 2016
En la columna Sexo/Averiguación previa los bbb son: 2008
En la columna Rango de edad los bbb son: 2018

Activar Windows
Ve a Configuración para activar Windows.

Bien, ya filtramos nuestras columnas y podemos ver que, por lo menos en la mitad de nuestras columnas tiene el dato llamado "bbb"
```

```
Archivo Editar Selección Ver Ir ... Q df_sucio
Untitled-2.ipynb • código_limpieza.ipynb X Untitled-3.ipynb •
C:\Users\Ulises\Downloads> código_limpieza.ipynb > Esta base de datos contiene información sobre el índice de criminalidad en Mexico, esta base tiene como finalidad ser de apoyo para los departamentos policia
Generar + Código + Markdown | Ejecutar todo | Borrar todas las salidas | Esquema Python 3.13.7

df1=df
for i in lista_col:
    df1=df1[df1[i] != 'bbb'] #esta linea de codigo es quien borra el susodicho "bbb"
df1

[163] Python

...

```

	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	...	Mayo	Junio	Julio	Agosto	Septiembre	Octu
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4	0.0	...	NaN	1.0	NaN	1.0	0.0	
2	NaN	1.0	Aguascalientes	La libertad y la seguridad sexual	Delito sexual	Hostigamiento sexual	Hostigamiento sexual	1.0	1	0.0	...	3.0	0.0	1.0	0.0	0.0	
4	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Violación	Violación equiparada	Violación equiparada	0.0	5	3.0	...	3.0	4.0	4.0	4.0	3.0	N
5	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Incesto	Incesto	Incesto	0.0	0	0.0	...	0.0	NaN	0.0	0.0	1.0	
8	2015.0	1.0	Aguascalientes	El patrimonio	Robo sin violencia	Robo a casa habitación	Sin violencia	215.0	176	202.0	...	215.0	228.0	233.0	229.0	202.0	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

```
Activar Windows
Ve a Configuración para activar Windows.

OVR Spaces: 4 11:26 p. m. 20/10/2025
```

```
for i in lista_col:
    print(f"En la columna {i} los bbb son: {df1[df1[i] == 'bbb'].shape[0]}")
```

En la columna Año los bbb son: 0  
En la columna Clave\_Ent los bbb son: 0  
En la columna Entidad los bbb son: 0  
En la columna Bien jurídico afectado los bbb son: 0  
En la columna Tipo de delito los bbb son: 0  
En la columna Subtipo de delito los bbb son: 0  
En la columna Modalidad los bbb son: 0  
En la columna Enero los bbb son: 0  
En la columna Febrero los bbb son: 0  
En la columna Marzo los bbb son: 0  
En la columna Abril los bbb son: 0  
En la columna Mayo los bbb son: 0  
En la columna Junio los bbb son: 0  
En la columna Julio los bbb son: 0  
En la columna Agosto los bbb son: 0  
En la columna Septiembre los bbb son: 0  
En la columna Octubre los bbb son: 0  
En la columna Noviembre los bbb son: 0  
En la columna Diciembre los bbb son: 0  
En la columna Sexo/Averiguación previa los bbb son: 0  
En la columna Rango de edad los bbb son: 0

Finalmente, encontré valores clasificados de manera errónea, lo cual desato el inmediato cambio en todas las columnas donde se presenten

```
df1.info()
```

<class 'pandas.core.frame.DataFrame'>  
Index: 87345 entries, 0 to 105802  
Data columns (total 21 columns):  
# Column Non-Null Count Dtype  
---  
0 Año 82933 non-null float64  
1 Clave\_Ent 82960 non-null float64  
2 Entidad 82891 non-null object  
3 Bien jurídico afectado 82901 non-null object  
4 Tipo de delito 83011 non-null object  
5 Subtipo de delito 82886 non-null object  
6 Modalidad 82922 non-null object  
7 Enero 82961 non-null float64  
8 Febrero 82848 non-null object  
9 Marzo 82963 non-null float64  
10 Abril 82925 non-null object  
11 Mayo 83015 non-null float64  
12 Junio 82944 non-null float64  
13 Julio 73153 non-null object  
14 Agosto 73211 non-null float64  
15 Septiembre 73148 non-null object  
16 Octubre 73203 non-null float64  
17 Noviembre 73235 non-null float64  
18 Diciembre 73165 non-null object  
19 Sexo/Averiguación previa 82924 non-null object  
20 Rango de edad 82851 non-null object

Archivo Editar Selección Ver Ir ... Q df\_sucio

Untitled-2.ipynb • **codigo\_limpieza.ipynb** X Untitled-3.ipynb •

C:\Users\Ulises\Downloads > codigo\_limpieza.ipynb > Esta base de datos contiene informacion sobre el indice de criminalidad en Mexico, esta base tiene como finalidad ser de apoyo para los departamentos policia

Generar + Código + Markdown | Ejecutar todo | Borrar todas las salidas | Esquema Python 3.13.7

Como se puede observar las columnas:Febrero, Abril, Julio, Septiembre y Diciembre son de valor numerico, sin embargo en la base esta establecido como objeto, lo cual me indica de tengo que realizar un cambio de variable

En seguida hacemos uso del comando "unique" para poder visualizar el contenido que tiene la columna "Febrero"

```
df1['Febrero'].unique()
```

```
array(['4', '1', '5', '0', '176', '3', '137', '144', '7', '11', '80',  
      '16', nan, '102', '60', '33', '180', '15', '41', '14', '50', '111',  
      '28', '2', '83', '77', '46', '27', '31', '944', '910', '307',  
      '188', '9', '625', '113', '883', '68', '38', '628', '156', '345',  
      '217', '40', '100', '779', '13', '105', '8', '380', '42', '110',  
      '30', '58', '59', '266', '22', '12', '6', '20', '336', '101', '52',  
      '123', '172', '25', '74', '141', '185', '579', '35', '19', '34',  
      '26', '48', '87', '140', '62', '56', '71', '47', '37', '73', '76',  
      '103', '65', '247', '84', '43', '54', '315', '61', '55', '163',  
      '362', '229', '498', '820', '219', '18', '326', '234', '704', '89',  
      '763', '248', '66', '107', '265', '732', '143', '2021', '859',  
      '175', '646', '191', '112', '587', '678', '211', '32', '10', '63',  
      '23', '82', '232', '256', '44', '45', '377', '250', '638', '99',  
      '318', '257', '21', '70', '57', '403', '192', '613', '253', '158',  
      '69', '72', '301', '324', '166', '79', '521', '132', '431', '29',  
      '122', '78', '577', '106', '97', '86', '118', '1698', '209',  
      '1106', '197', '194', '416', '1078', '568', '2310', '458', '976',  
      '298', '281', '288', '116', '24', '213', '3637', '149', '408',  
      '173', '142', '186', '136', '184', '36', '17', '85', '370', '153',  
      '92', '341', '261', '562', '224', '243', '204', '64', '661',  
      '1102', '200', '207', '40', '75', '100', '168', '100', '460'])
```

Activar Windows  
Ve a Configuración para activar Windows.

OVR Spaces: 4 11:29 p. m. 20/10/2025

Archivo Editar Selección Ver Ir ... Q df\_sucio

Untitled-2.ipynb • **codigo\_limpieza.ipynb** X Untitled-3.ipynb •

C:\Users\Ulises\Downloads > codigo\_limpieza.ipynb > Esta base de datos contiene informacion sobre el indice de criminalidad en Mexico, esta base tiene como finalidad ser de apoyo para los departamentos policia

Generar + Código + Markdown | Ejecutar todo | Borrar todas las salidas | Esquema Python 3.13.7

Se observa que los números están entre comillas, eso indica que son caracteres y no números.

Para cambiar el tipo se ocupa el comando astype( tipo de valor)

int entero

float decimal

En este caso los convertiremos a decimal

```
df1['Febrero']=df1['Febrero'].astype(float)
```

```
df1['Febrero'].unique()
```

```
array([4.000e+00, 1.000e+00, 5.000e+00, 0.000e+00, 1.760e+02, 3.000e+00,  
      1.370e+02, 1.440e+02, 7.000e+00, 1.100e+01, 8.000e+01, 1.600e+01,  
      nan, 1.020e+02, 6.000e+01, 3.300e+01, 1.800e+02, 1.500e+01,  
      4.100e+01, 1.400e+01, 5.000e+01, 1.110e+02, 2.800e+01, 2.000e+00,  
      8.300e+01, 7.700e+01, 4.600e+01, 2.700e+01, 3.100e+01, 9.440e+02,  
      9.100e+02, 3.070e+02, 1.880e+02, 9.000e+00, 6.250e+02, 1.130e+02,  
      8.830e+02, 6.800e+01, 3.800e+01, 6.280e+02, 1.560e+02, 3.450e+02,  
      2.170e+02, 4.000e+01, 1.000e+02, 7.790e+02, 1.300e+01, 1.050e+02,  
      8.000e+00, 3.800e+02, 4.200e+01, 1.190e+02, 3.000e+01, 5.800e+01,  
      5.900e+01, 2.660e+02, 2.200e+01, 1.200e+01, 6.000e+00, 2.000e+01,  
      3.360e+02, 1.010e+02, 5.200e+01, 1.230e+02, 1.720e+02, 2.500e+01])
```

Activar Windows  
Ve a Configuración para activar Windows.

OVR Spaces: 4 11:29 p. m. 20/10/2025



```
df1['Abril']=df1['Abril'].astype(float)

df1['Julio']=df1['Julio'].astype(float)

df1['Septiembre']=df1['Septiembre'].astype(float)

df1['Diciembre']=df1['Diciembre'].astype(float)

Aqui finalmente se observa que los valores acaban de cambiar a tipo numerico decimal

df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 87345 entries, 0 to 105802
Data columns (total 21 columns):
# Column Non-Null Count Dtype
-----
0 Año 82933 non-null float64
1 Clave_Ent 82060 non-null float64
2 Entidad 82891 non-null object
3 Bien jurídico afectado 82901 non-null object
4 Tipo de delito 83011 non-null object
5 Subtipo de delito 82886 non-null object
6 Modalidad 82922 non-null object
7 Enero 82961 non-null float64
8 Febrero 82848 non-null float64
9 Marzo 82963 non-null float64
10 Abril 82925 non-null float64
11 Mayo 83015 non-null float64
12 Junio 82944 non-null float64
13 Julio 73153 non-null float64
14 Agosto 73211 non-null float64
15 Septiembre 73148 non-null float64
16 Octubre 73203 non-null float64
17 Noviembre 73235 non-null float64
18 Diciembre 73165 non-null float64
19 Sexo/Averiguación previa 82924 non-null object
20 Rango de edad 82851 non-null object
dtypes: float64(14), object(7)
memory usage: 14.7+ MB
```

Como pasos finales los valores NaN, no los borre, ya que ese no es el objetivo del proyecto, el cual consiste en recuperar la mayor cantidad de datos posibles, mi solución fue convertir los NaN en 0 de ese modo no representan problemas a futuro

Y como paso final, ya solo reinicié el índice para dar mejor presentación a la base de datos

En conclusión:

La base no represento un desafío, ya que los datos en su mayoría están en una forma en donde yo puedo trabajar sin ningún problema, en cuanto a soluciones, solo aplique lo viste en clase sin mayor

problema, como aprendizaje, me llevo una buena experiencia, para ser la primera base de datos que limpio por mi cuenta propia