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.

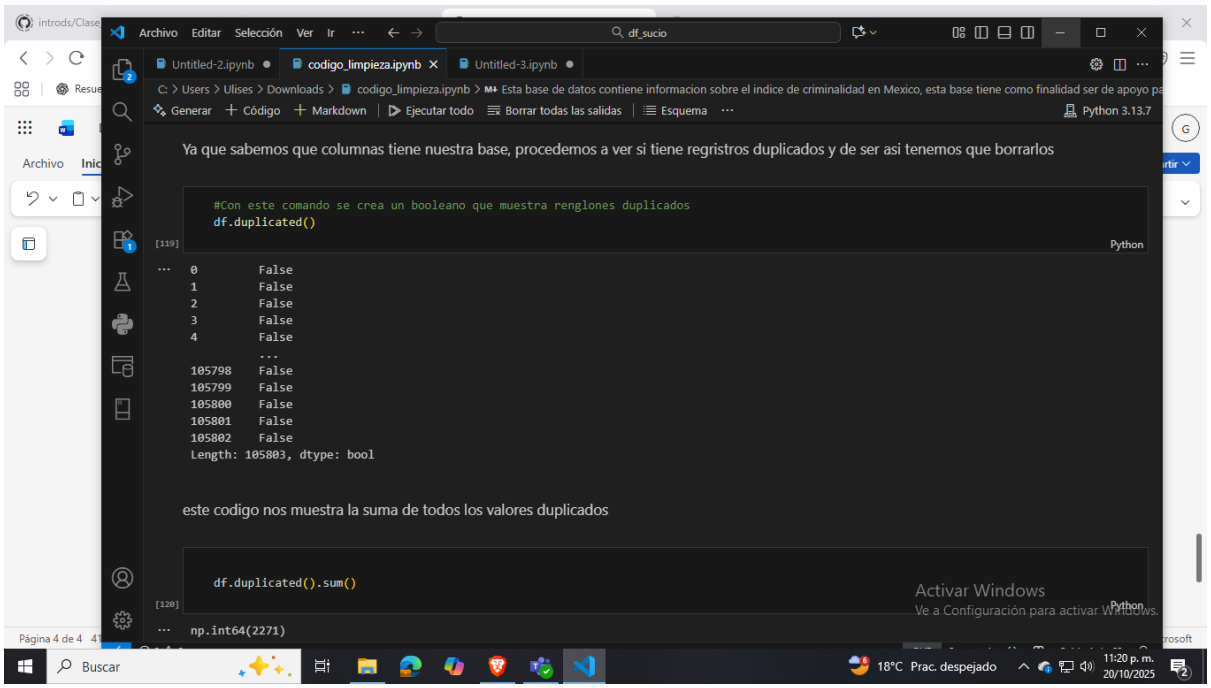
Modalidad: se refiere a las circunstancias de lugar, tiempo, modo u ocasión que califican la conducta del sujeto activo y que tienen como función determinar el quantum de la pena, afectando su medida, ya sea aumentándola o disminuyéndola.

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

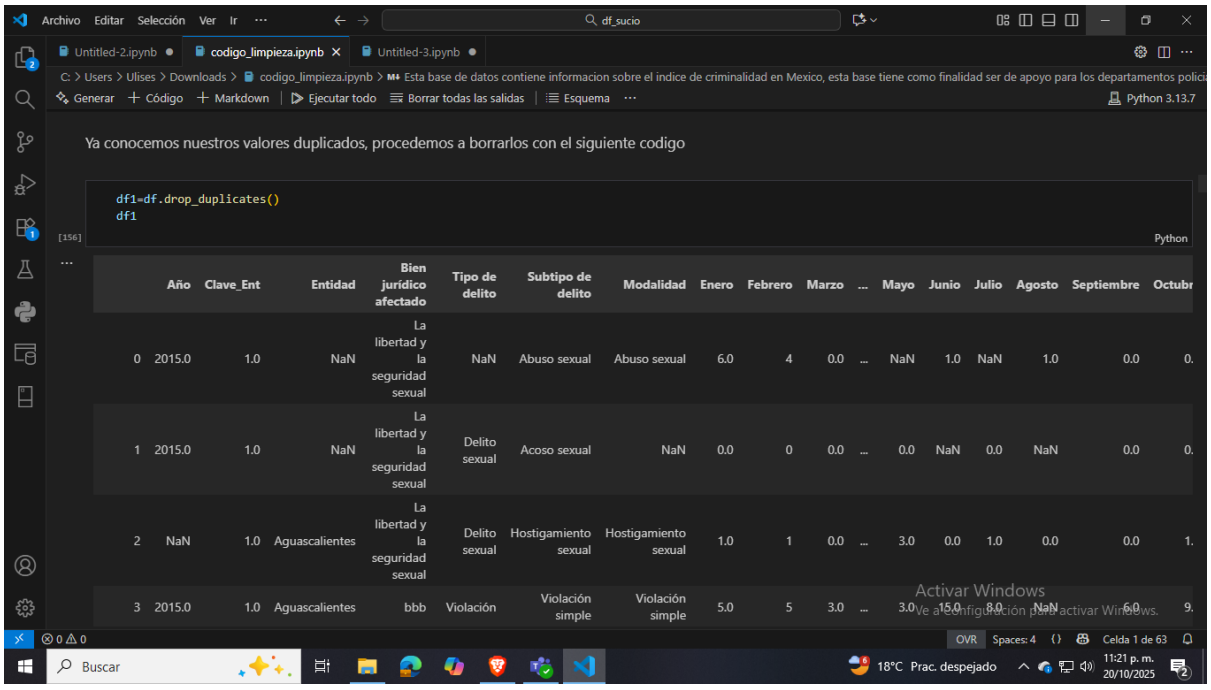
Sexo: Masculino/Femenino

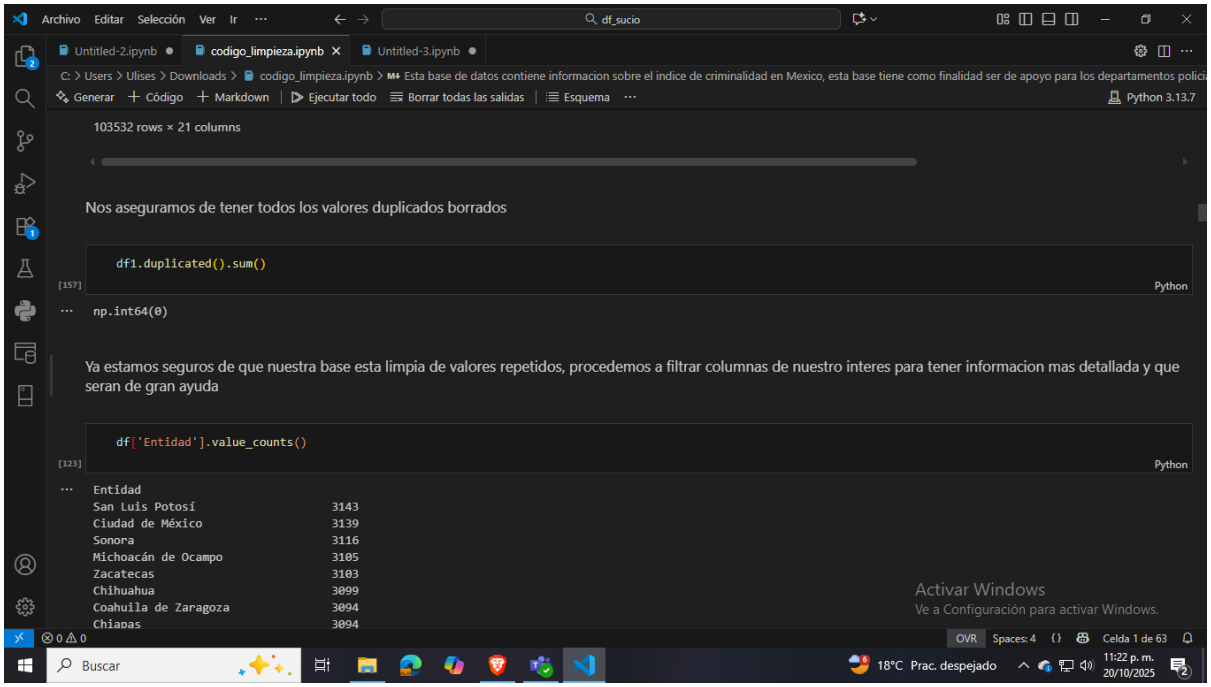
Rango de edad: Menores de edad 0-17, Mayores de edad 18-100

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

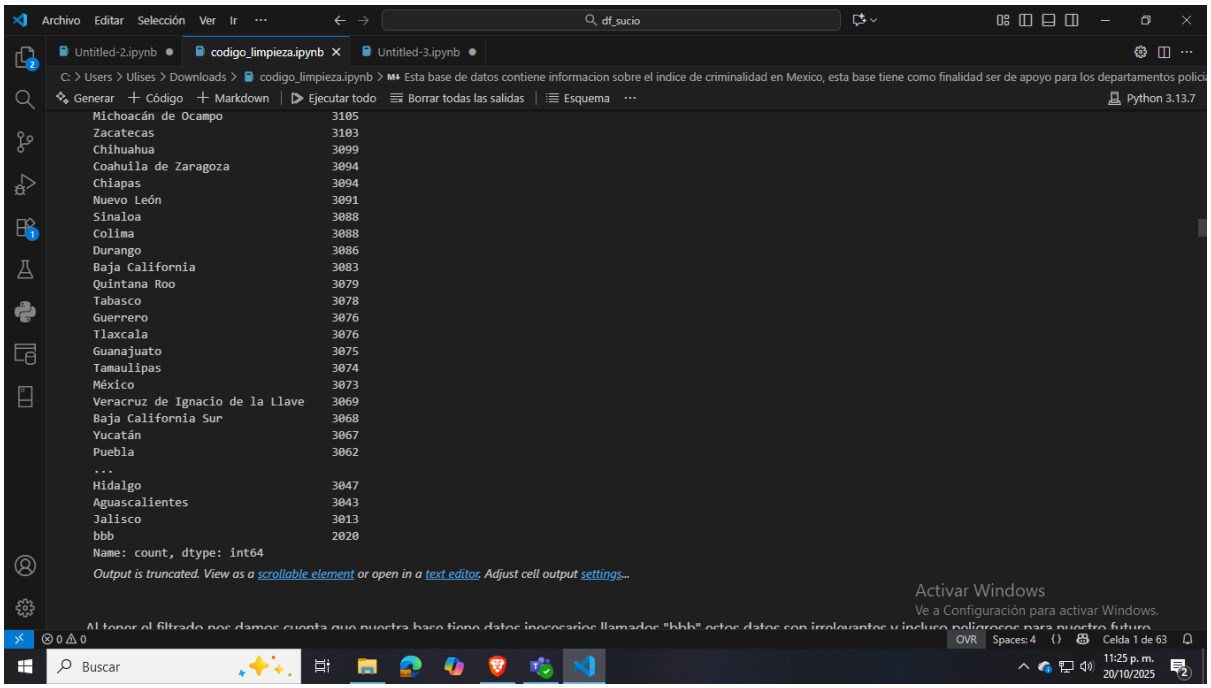


Después de ver que efectivamente contiene duplicados, cedi a la tarea de borrar todos

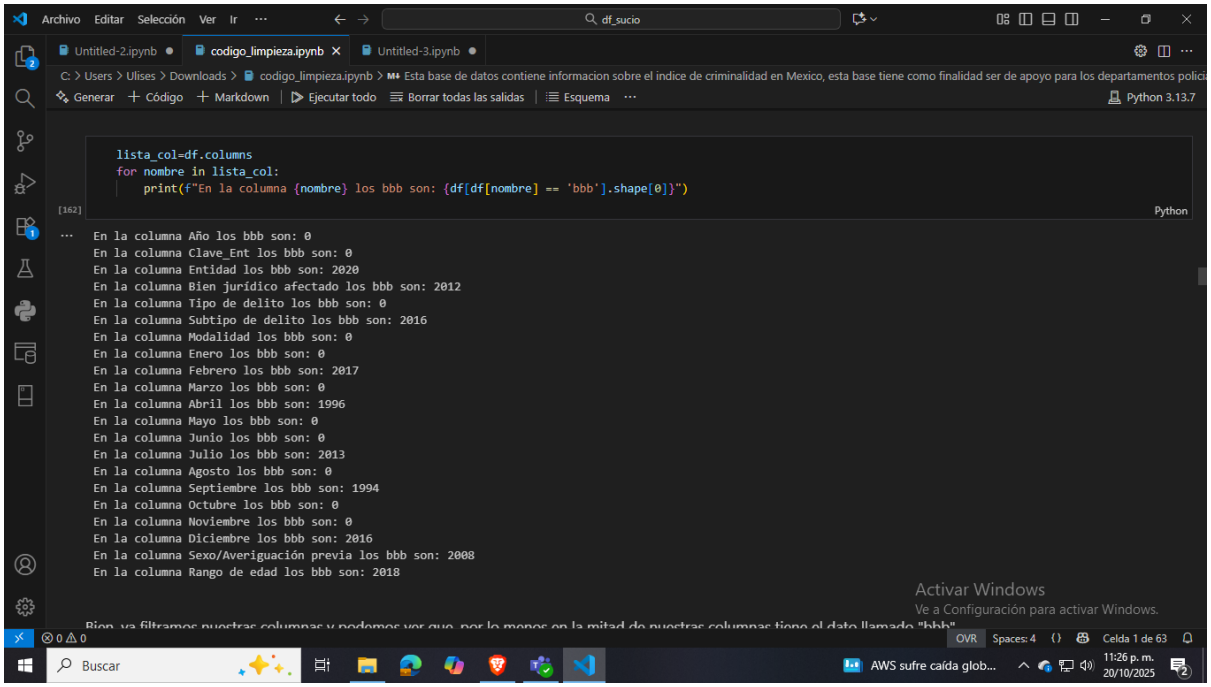


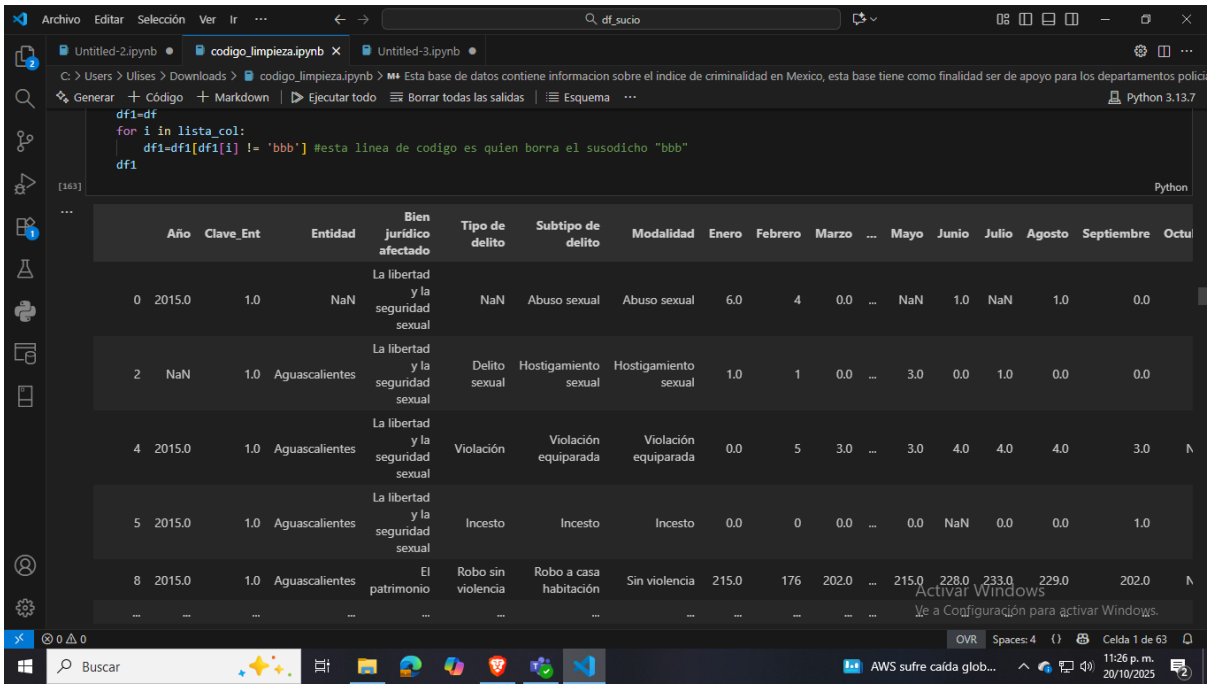


Consecutivamente filtre algunas columnas y encontre valores inválidos

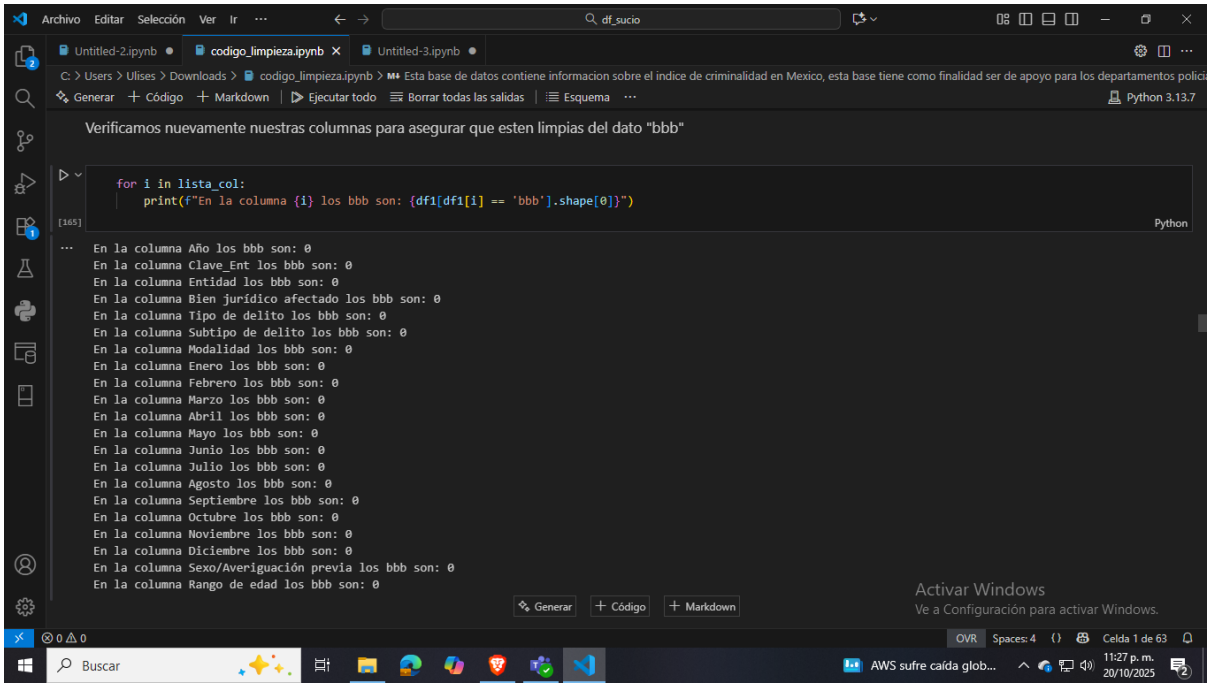


Y procedí a borrar cada uno de cada columna donde hubiera uno

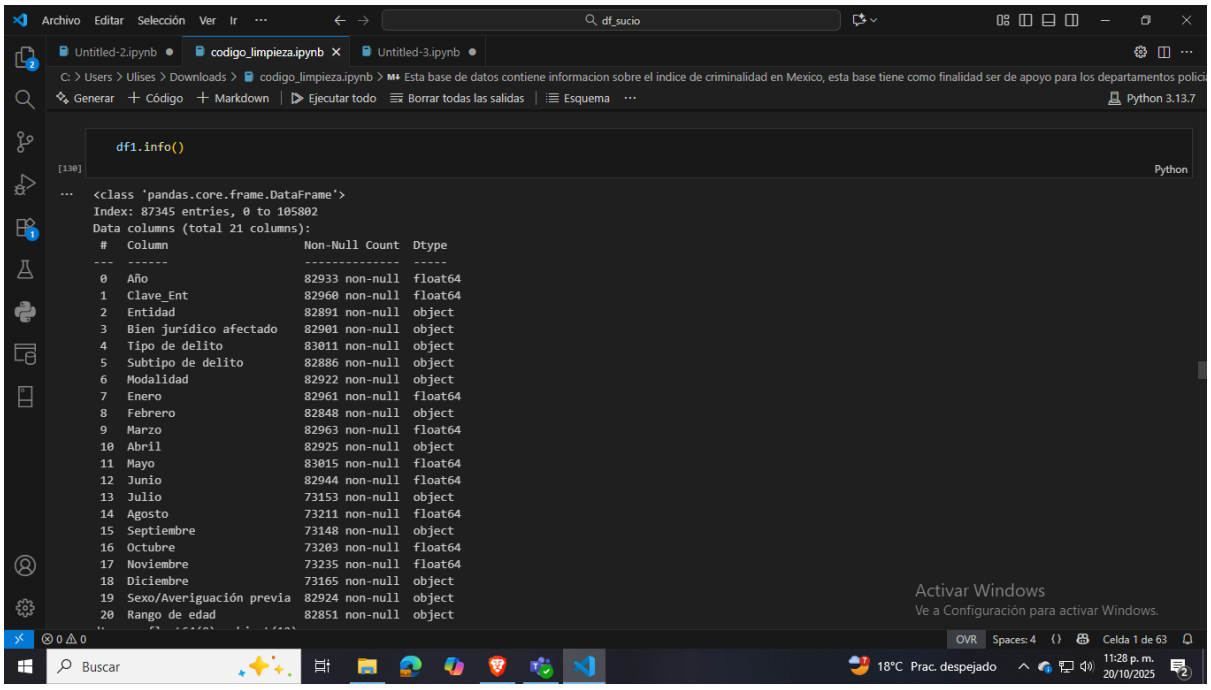


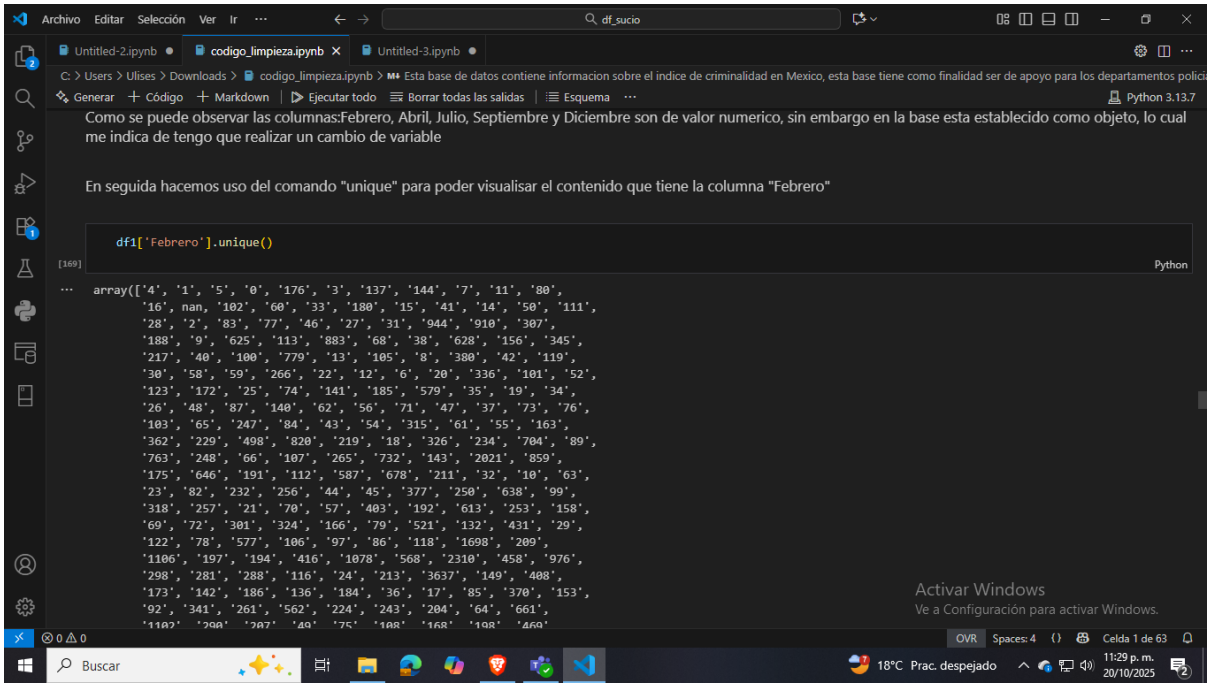


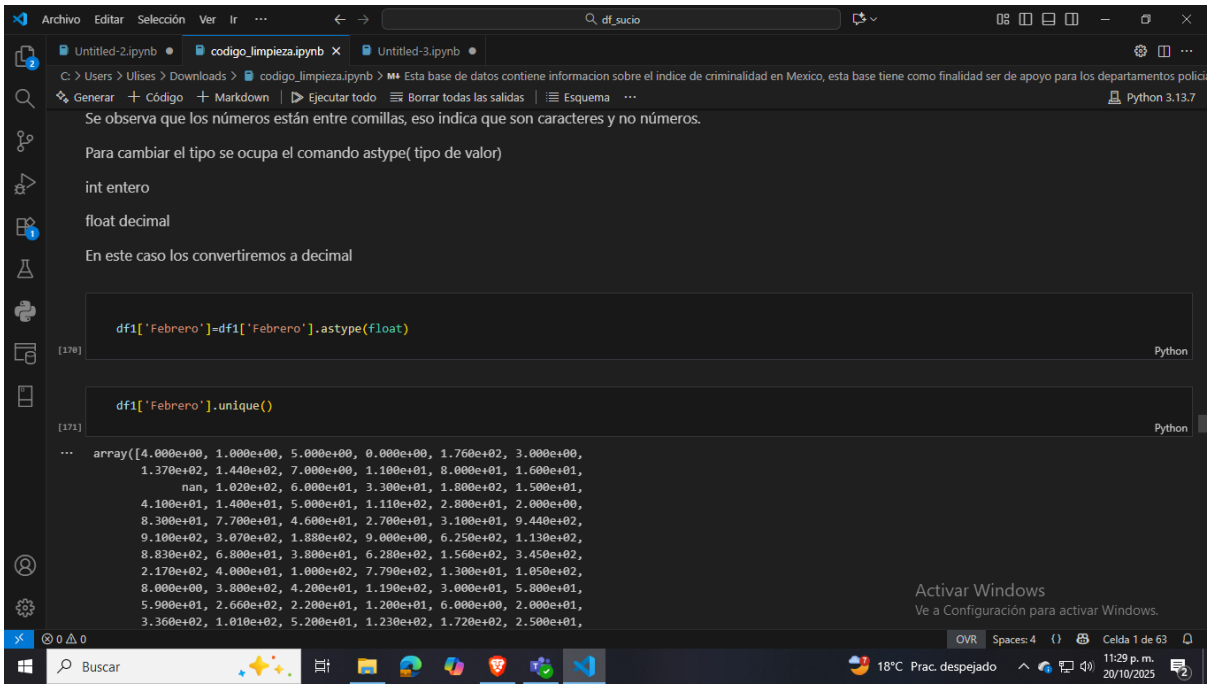
	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	...	Mayo	Junio	Julio	Agosto	Septiembre	Octu
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4	0.0	...	NaN	1.0	NaN	1.0	0.0	
2	NaN	1.0	Aguascalientes	La libertad 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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

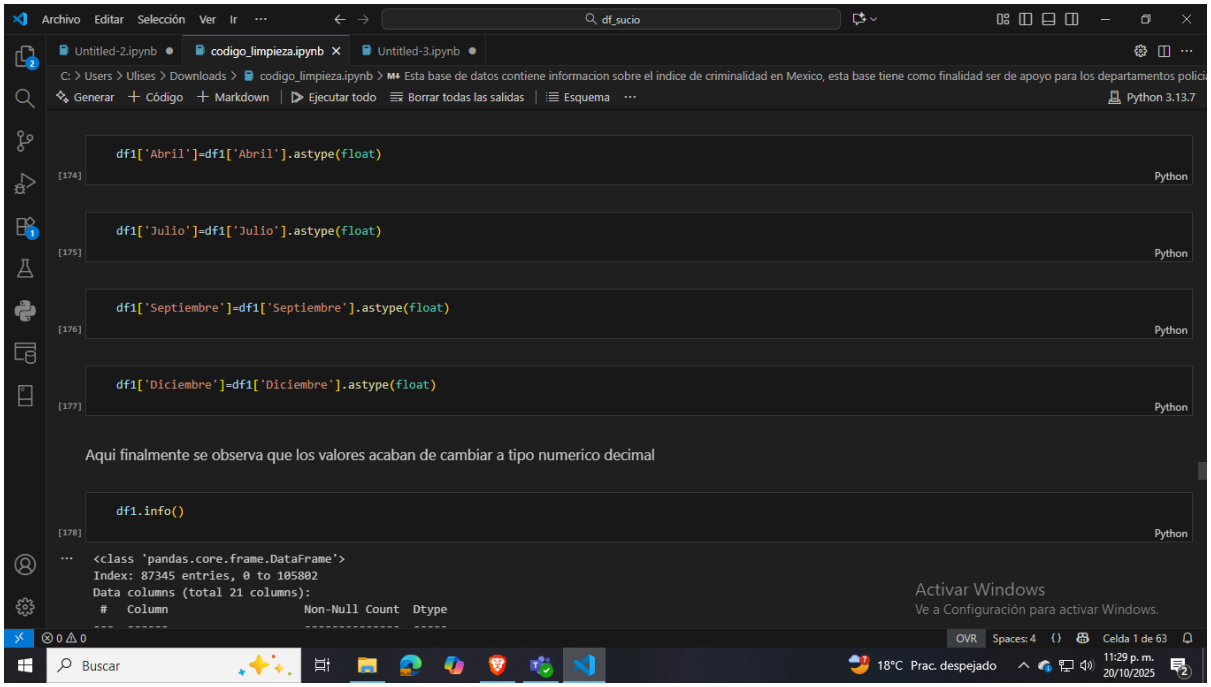


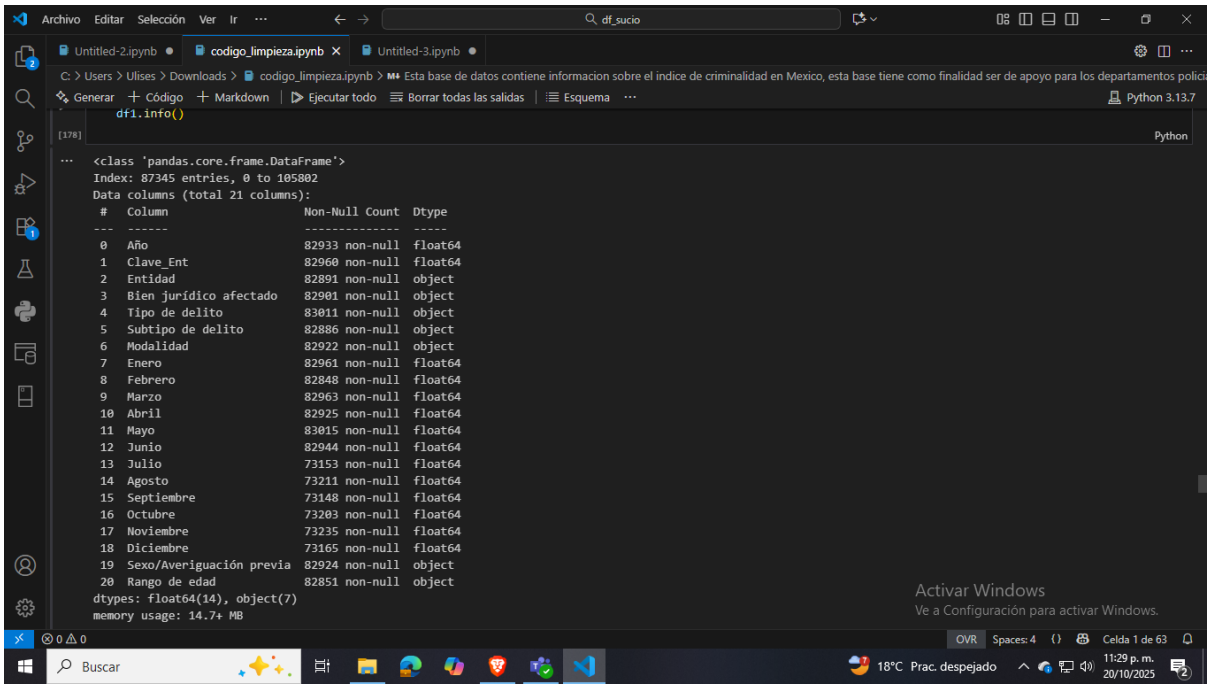
Finalmente, encontré valores clasificados de manera errónea, lo cual desato el inmediato cambio en todas las columnas donde se presenten



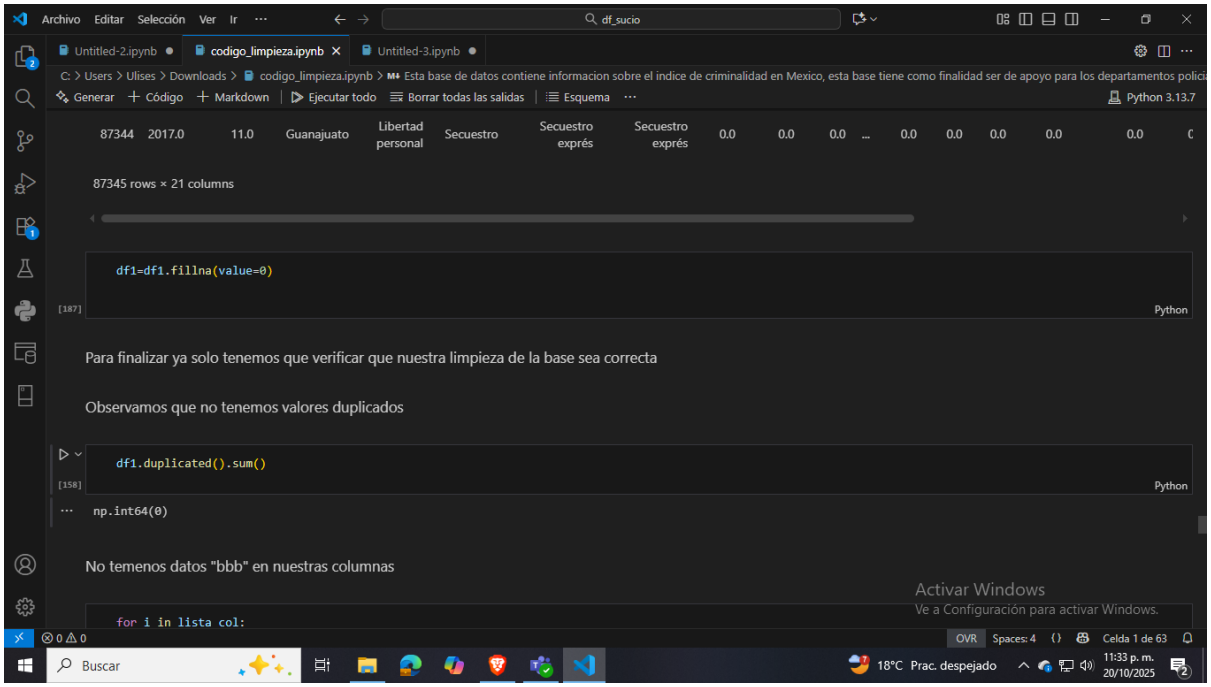

Archivo Editar Selección Ver Ir
df_sucio
Untitled-2.ipynb
codigo_limpieza.ipynb
Untitled-3.ipynb
Generar + Código + Markdown Ejecutar todo Borrar todas las salidas Esquema
Python 3.13.7
Como se puede observar las columnas:Febrero, Abril, Julio, Septiembre y Diciembre son de valor numerico, sin embargo 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 visualisar el contenido que tiene la columna "Febrero"
df1['Febrero'].unique()
[169]
Python
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', '119',
'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',
Activar Windows
Ve a Configuración para activar Windows.
OVR Spaces: 4 Celda 1 de 63
Buscar
18°C Prac. despejado
11:29 p. m.
20/10/2025


Archivo Editar Selección Ver Ir
df_sucio
Untitled-2.ipynb
codigo_limpieza.ipynb
Untitled-3.ipynb
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)
[170]
Python
df1['Febrero'].unique()
[171]
Python
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 Celda 1 de 63
Buscar
18°C Prac. despejado
11:29 p. m.
20/10/2025

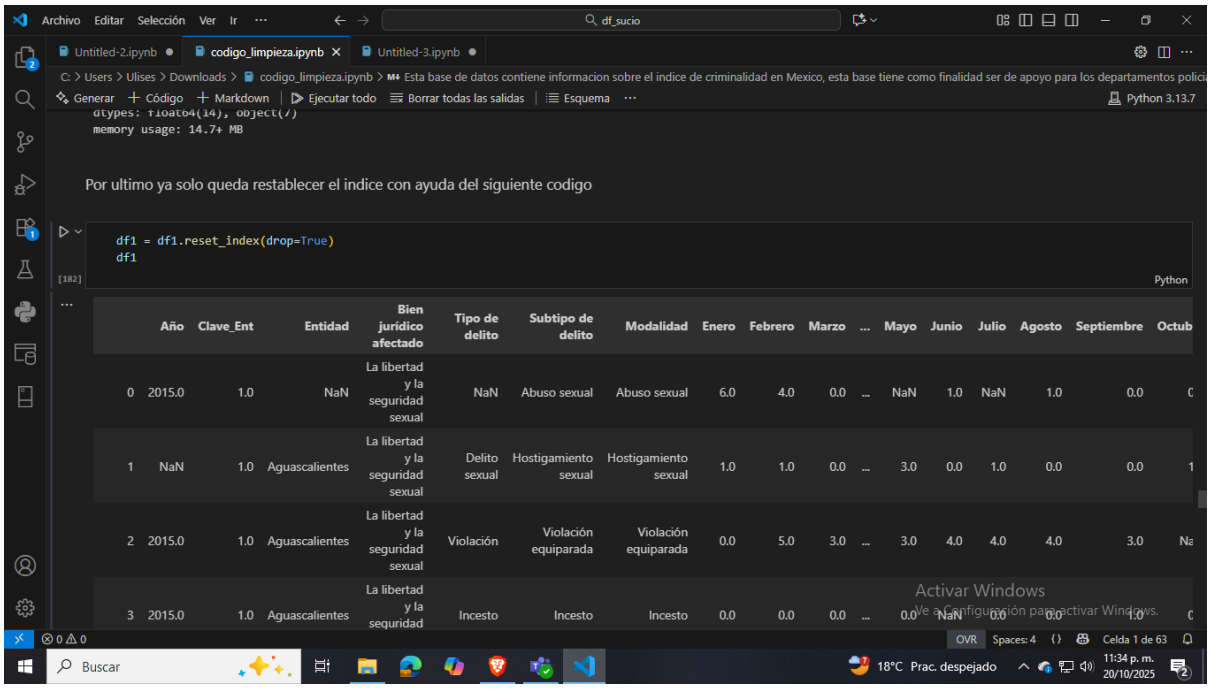




Como pasos finales los valores NaN, no los borre, ya que ese no es el objetivo del proyecto, el cual consiste en recuperar la mayor cantidad de datos posibles, mi solución fue convertir los NaN en 0 de ese modo no representan problemas a futuro



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



	Año	Clave_Ent	Entidad	Bien jurídico afectado	Tipo de delito	Subtipo de delito	Modalidad	Enero	Febrero	Marzo	...	Mayo	Junio	Julio	Agosto	Septiembre	Octub
0	2015.0	1.0	NaN	La libertad y la seguridad sexual	NaN	Abuso sexual	Abuso sexual	6.0	4.0	0.0	...	NaN	1.0	NaN	1.0	0.0	C
1	NaN	1.0	Aguascalientes	La libertad y la seguridad sexual	Delito sexual	Hostigamiento sexual	Hostigamiento sexual	1.0	1.0	0.0	...	3.0	0.0	1.0	0.0	0.0	1
2	2015.0	1.0	Aguascalientes	La libertad y la seguridad sexual	Violación	Violación equiparada	Violación equiparada	0.0	5.0	3.0	...	3.0	4.0	4.0	4.0	3.0	Na
3	2015.0	1.0	Aguascalientes	La libertad y la seguridad	Incesto	Incesto	Incesto	0.0	0.0	0.0	...	0.0	NaN	0.0	0.0	0.0	C

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