



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD CIENCIAS DE LA COMPUTACIÓN

INTRODUCCION A LA CIENCIA DE DATOS

M.C. JAIME ALEJANDRO ROMERO SIERRA

UNIDAD 2: LIMPIEZA DE UNA BASE DE DATOS

ENSUCIADA

ABRAHAM FUENTES LÓPEZ

22/10/2024

Análisis inicial:

Resumen estadístico de la base de datos antes de la limpieza

```
df = pd.read_csv('/content/drive/MyDrive/df_sucio (2).csv')
```

```
df
```

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status	LOCATION	LAT	LO
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	NaN	Adult Arrest	NaN	34.0375	-118.3506
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	NaN	Invest Cont	1000 S FLOWER ST	34.0444	-118.2628
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTIUNIT DWELLING (APARTMENT, DUPLEX, ETC)	NaN	Invest Cont	1400 W 37TH ST	invalid	-118.3002
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	NaN	O	CLOTHING STORE	NaN	Invest Cont	14000 RIVERSIDE DR	34.1576	-118.4387
4	NaN	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200.0	Hollywood	THEFT OF IDENTITY	28.0	M	H	SIDEWALK	NaN	Invest Cont	invalid	34.0944	-118.3277
...
1089038	200212691.0	07/17/2020 12:00:00 AM	NaN	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VAL...	0.0	X	X	HOSPITAL	NaN	Invest Cont	1200 WILSHIRE BL	34.053	-118.2649
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	30.0	M	H	NaN	REVOLVER	Adult Arrest	8TH ST	34.0506	NaN
1089040	241110191.0	08/23/2024 12:00:00 AM	08/23/2024 12:00:00 AM	1530.0	Northeast	VEHICLE - STOLEN	0.0	NaN	NaN	STREET	NaN	Invest Cont	2600 N FIGUEROA ST	34.0884	-118.2298
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	18.0	M	H	STREET	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	invalid	5TH	34.0481	-118.2507
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEAMEANOR (\$399 OR UNDER)	57.0	F	B	VEHICLE PASSENGER/TRUCK	UNKNOWN WEAPON/OTHER WEAPON	Invest Cont	3400 EDGEHILL DR	34.0251	-118.3305

Nuestro Dataframe cuenta con un total de 1,089,043 filas y 15 columnas:

```
# Resumen estadístico de la base de datos antes de limpiarse
```

```
# Cantidad de filas y columnas respectivamente
```

```
df.shape
```

```
(1089043, 15)
```

Total de filas duplicadas:

```
# Total de filas duplicadas encontradas
```

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

```
38269
```

Porcentaje de valores faltantes por columna:

```
# Porcentaje de valores faltantes por columna
df.isnull().mean()*100
```

	0
DR_NO	2.999973
Date_Reported	2.999973
Date_occured	2.999973
Time_occured	2.999973
Area	2.999973
Crime_Code	2.999973
Victim_age	2.999973
Victim_sex	16.520927
Victim_descent	16.535894
Premis	3.059383
Weapon	67.564274
Status	2.999973
LOCATION	2.999973
LAT	2.999973
LON	2.999973

Descripción de los tipos de datos originales:

```
#Descripción de los tipos de datos originales.
df.info()
```

<class 'pandas.core.frame.DataFrame'>			
RangeIndex: 1089043 entries, 0 to 1089042			
Data columns (total 15 columns):			
#	Column	Non-Null Count	Dtype
0	DR_NO	1056372 non-null	float64
1	Date_Reported	1056372 non-null	object
2	Date_occured	1056372 non-null	object
3	Time_occured	1056372 non-null	float64
4	Area	1056372 non-null	object
5	Crime_Code	1056372 non-null	object
6	Victim_age	1056372 non-null	float64
7	Victim_sex	909123 non-null	object
8	Victim_descent	908960 non-null	object
9	Premis	1055725 non-null	object
10	Weapon	353239 non-null	object
11	Status	1056372 non-null	object
12	LOCATION	1056372 non-null	object
13	LAT	1056372 non-null	object
14	LON	1056372 non-null	object
dtypes: float64(3), object(12)			
memory usage: 124.6+ MB			

Proceso de limpieza de datos

Eliminación de las columnas de LOCATION, LAT Y LON

No las ocuparemos ya que se usan para espacios geográficos específicos y aún no aprendemos a usar esas herramientas.

Elimino las columnas de LOCATION, LAT Y LON

de los nombres ya que se usan para espacios geográficos específicos y aún no aprendemos a usar esas herramientas

Ejecutar celda (Ctrl-Enter)

celda ejecutada desde el último cambio

ejecutada por Abraham Fuentes

2:08 (hace 1 hora)

se ha ejecutado en 0.478 s

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	NaN	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	NaN	Invest Cont
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	NaN	Invest Cont
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	NaN	O	CLOTHING STORE	NaN	Invest Cont
4	NaN	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200.0	Hollywood	THEFT OF IDENTITY	28.0	M	H	SIDEWALK	NaN	Invest Cont
...
1089038	200212691.0	07/17/2020 12:00:00 AM	NaN	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	0.0	X	X	HOSPITAL	NaN	Invest Cont
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	30.0	M	H	NaN	REVOLVER	Adult Arrest
1089040	241110191.0	08/23/2024 12:00:00 AM	08/23/2024 12:00:00 AM	1530.0	Northeast	VEHICLE - STOLEN	0.0	NaN	NaN	STREET	NaN	Invest Cont
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	18.0	M	H	STREET	STRONG-ARM (HANDS, FIST, FEET OR BODYLY FORCE)	Invalid
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEAMEANOR (\$399 OR UNDER)	57.0	F	B	VEHICLE, PASSENGER/TRUCK	UNKNOWN WEAPON/OTHER WEAPON	Invest Cont

1089043 rows x 12 columns

Eliminación de las filas con NaN de ciertas columnas

df

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	
4	NaN	08/18/2022 12:00:00 AM	08/17/2020 12:00:00 AM	1200.0	Hollywood	THEFT OF IDENTITY	
...
1089038	200212691.0	07/17/2020 12:00:00 AM	NaN	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	
1089040	241110191.0	08/23/2024 12:00:00 AM	08/23/2024 12:00:00 AM	1530.0	Northeast	VEHICLE - STOLEN	
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEAMEANOR (\$399 OR UNDER)	

1089043 rows x 12 columns

#Elimina las filas con NaN de ciertas columnas

df = df.dropna(subset=['DR_NO'])
df = df.dropna(subset=['Date_Reported'])
df = df.dropna(subset=['Crime_Code'])
df

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	
...
1089038	200212691.0	07/17/2020 12:00:00 AM	NaN	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	
1089040	241110191.0	08/23/2024 12:00:00 AM	08/23/2024 12:00:00 AM	1530.0	Northeast	VEHICLE - STOLEN	
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEAMEANOR (\$399 OR UNDER)	
993923 rows x 12 columns							

Cambio valores nulos a valores que podrían reemplazarlos

df.head()

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	NaN	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	NaN	Invest Cont
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	NaN	Invest Cont
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	NaN	O	CLOTHING STORE	NaN	Invest Cont
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	41.0	M	H	SINGLE FAMILY DWELLING	NaN	NaN

#Cambio valores nulos a valores que podrían reemplazarlos

df['Date_occured'] = df['Date_occured'].fillna("DESCONOCIDO")
df['Victim_descent'] = df['Victim_descent'].fillna("DESCONOCIDO")
df['Weapon'] = df['Weapon'].fillna("SIN_AGRESION")
df.head()

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	SIN_AGRESION	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	SIN_AGRESION	Invest Cont
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	SIN_AGRESION	Invest Cont
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	NaN	O	CLOTHING STORE	SIN_AGRESION	Invest Cont
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	41.0	M	H	SINGLE FAMILY DWELLING	SIN_AGRESION	NaN

Imputacion de valores faltantes

df

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	SIN_AGRESION	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	SIN_AGRESION	Invest Cort
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	SIN_AGRESION	Invest Cort
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	NaN	O	CLOTHING STORE	SIN_AGRESION	Invest Cort
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	41.0	M	H	SINGLE FAMILY DWELLING	SIN_AGRESION	NaN
...
1089038	200212691.0	07/17/2020 12:00:00 AM	DESCONOCIDO	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	0.0	X	X	HOSPITAL	SIN_AGRESION	Invest Cort
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	30.0	M	H	NaN	REVOLVER	Adult Arrest
1089040	241110191.0	08/23/2024 12:00:00 AM	08/23/2024 12:00:00 AM	1530.0	Northeast	VEHICLE - STOLEN	0.0	NaN	DESCONOCIDO	STREET	SIN_AGRESION	Invest Cort
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	18.0	M	H	STREET	STRONG ARM (HANDS, FIST, FEET OR BODY FORCE)	Invalid
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEAMEANOR (\$399 OR UNDER)	57.0	F	B	VEHICLE, PASSENGER/TRUCK	UNKNOWN WEAPON/OTHER WEAPON	Invest Cort

993923 rows x 12 columns

Imputacion de valores faltantes

```
# Victim_age -> media
# Time_occured -> media

# Area -> Moda
# Victim_sex -> Moda
# Premis -> Moda
# Status -> Moda

df['Victim_age'] = df['Victim_age'].fillna(df['Victim_age'].mean())
df['Time_occured'] = df['Time_occured'].fillna(df['Time_occured'].mean())

df['Area'] = df['Area'].fillna(df['Area'].mode()[0])
df['Victim_sex'] = df['Victim_sex'].fillna(df['Victim_sex'].mode()[0])
df['Premis'] = df['Premis'].fillna(df['Premis'].mode()[0])
df['Status'] = df['Status'].fillna(df['Status'].mode()[0])
df
```

df.isnull().sum()

	0
DR_NO	0
Date_Reported	0
Date_occured	0
Time_occured	0
Area	0
Crime_Code	0
Victim_age	0
Victim_sex	0
Victim_descent	0
Premis	0
Weapon	0
Status	0

Eliminación de las filas duplicadas

```
# total de filas duplicadas
df.duplicated().sum()

56387

# Eliminamos las filas duplicadas
df=df.drop_duplicates()
df
```

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wiltshire	VEHICLE - STOLEN	0.0	M	O	STREET	SIN_AGRESION	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	SIN_AGRESION	Invest Court
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	SIN_AGRESION	Invest Court
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING GRAND THEFT (\$950.01 & OVER)	19.0	M	O	CLOTHING STORE	SIN_AGRESION	Invest Court
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	41.0	M	H	SINGLE FAMILY DWELLING	SIN_AGRESION	Invest Court
...
1089034	231115828.0	11/15/2023 12:00:00 AM	11/12/23 0:00	2200.0	Northeast	VEHICLE - STOLEN	0.0	M	DESCONOCIDO	STREET	SIN_AGRESION	Invest Court
1089038	200212691.0	07/17/2020 12:00:00 AM	DESCONOCIDO	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	0.0	X	X	HOSPITAL	SIN_AGRESION	Invest Court
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	30.0	M	H	STREET	REVOLVER	Adult Arrest
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	18.0	M	H	STREET	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)	Invalid
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEMEANOR (\$399 OR UNDER)	57.0	F	B	VEHICLE, PASSENGER/TRUCK	UNKNOWN WEAPON/OTHER WEAPON	Invest Court

```
# total de filas duplicadas
df.duplicated().sum()

0
```

Verificación y eliminación de invalid values

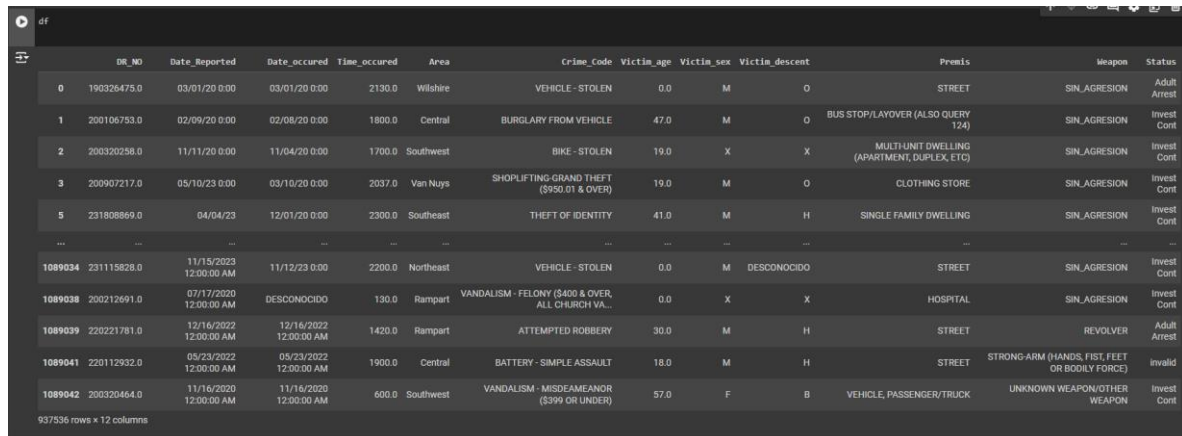
```
# Verificando si tenemos invalid values
for i in lista_col:
    print(f"En la columna {i} los invalid_value son: {df[df[i] == 'bbb'].shape[0]}")

# NO tenemos invalid_value
```

```
En la columna DR_NO los invalid_value son: 0
En la columna Date_Reported los invalid_value son: 0
En la columna Date_occured los invalid_value son: 0
En la columna Time_occured los invalid_value son: 0
En la columna Area los invalid_value son: 0
En la columna Crime_Code los invalid_value son: 0
En la columna Victim_age los invalid_value son: 0
En la columna Victim_sex los invalid_value son: 0
En la columna Victim_descent los invalid_value son: 0
En la columna Premis los invalid_value son: 0
En la columna Weapon los invalid_value son: 0
En la columna Status los invalid_value son: 0
```

Resultados de datos

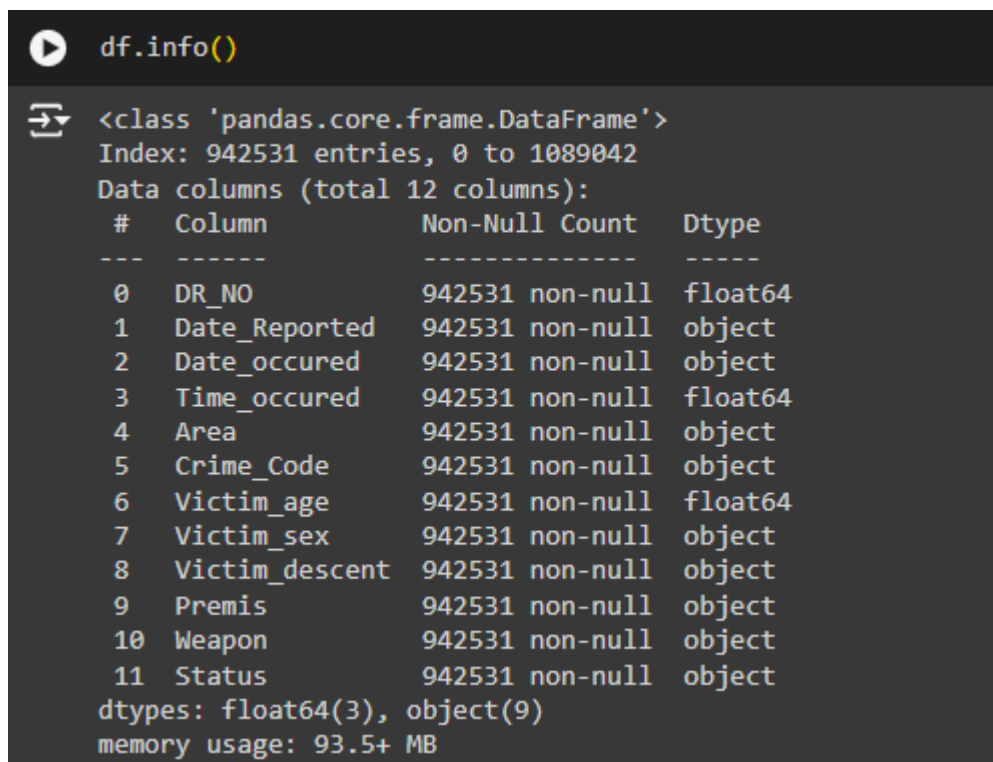
DataFrame final



The screenshot shows a Jupyter Notebook interface with a DataFrame named 'df'. The DataFrame has 12 columns and 937,536 rows. The columns are: DR_NO, Date_Reported, Date_occured, Time_occured, Area, Crime_Code, Victim_age, Victim_sex, Victim_descent, Premis, Weapon, and Status. The rows are indexed from 0 to 1089042. The DataFrame is displayed in a table format with a dark background.

	DR_NO	Date_Reported	Date_occured	Time_occured	Area	Crime_Code	Victim_age	Victim_sex	Victim_descent	Premis	Weapon	Status
0	190326475.0	03/01/20 0:00	03/01/20 0:00	2130.0	Wilshire	VEHICLE - STOLEN	0.0	M	O	STREET	SIN_AGRESION	Adult Arrest
1	200106753.0	02/09/20 0:00	02/08/20 0:00	1800.0	Central	BURGLARY FROM VEHICLE	47.0	M	O	BUS STOP/LAYOVER (ALSO QUERY 124)	SIN_AGRESION	Invest Cont
2	200320258.0	11/11/20 0:00	11/04/20 0:00	1700.0	Southwest	BIKE - STOLEN	19.0	X	X	MULTI-UNIT DWELLING (APARTMENT, DUPLEX, ETC)	SIN_AGRESION	Invest Cont
3	200907217.0	05/10/23 0:00	03/10/20 0:00	2037.0	Van Nuys	SHOPLIFTING-GRAND THEFT (\$950.01 & OVER)	19.0	M	O	CLOTHING STORE	SIN_AGRESION	Invest Cont
5	231808869.0	04/04/23	12/01/20 0:00	2300.0	Southeast	THEFT OF IDENTITY	41.0	M	H	SINGLE FAMILY DWELLING	SIN_AGRESION	Invest Cont
...
1089034	231115828.0	11/15/2023 12:00:00 AM	11/12/23 0:00	2200.0	Northeast	VEHICLE - STOLEN	0.0	M	DESCONOCIDO	STREET	SIN_AGRESION	Invest Cont
1089038	200212691.0	07/17/2020 12:00:00 AM	DESCONOCIDO	130.0	Rampart	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VA...	0.0	X	X	HOSPITAL	SIN_AGRESION	Invest Cont
1089039	220221781.0	12/16/2022 12:00:00 AM	12/16/2022 12:00:00 AM	1420.0	Rampart	ATTEMPTED ROBBERY	30.0	M	H	STREET	REVOLVER	Adult Arrest
1089041	220112932.0	05/23/2022 12:00:00 AM	05/23/2022 12:00:00 AM	1900.0	Central	BATTERY - SIMPLE ASSAULT	18.0	M	H	STREET	STRONG-ARM (HANDS, FIST, FEET OR BODYLY FORCE)	Invalid
1089042	200320464.0	11/16/2020 12:00:00 AM	11/16/2020 12:00:00 AM	600.0	Southwest	VANDALISM - MISDEMEANOR (\$399 OR UNDER)	57.0	F	B	VEHICLE, PASSENGER/TRUCK	UNKNOWN WEAPON/OTHER WEAPON	Invest Cont

Nos quedan 937,536 Filas y 12 Columnas



The screenshot shows the output of the `df.info()` command in a Jupyter Notebook. The output indicates that the DataFrame has 942,531 entries (rows) and 12 columns. The columns are: DR_NO, Date_Reported, Date_occured, Time_occured, Area, Crime_Code, Victim_age, Victim_sex, Victim_descent, Premis, Weapon, and Status. The data types are: float64(3), object(9). The memory usage is 93.5+ MB.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
Index: 942531 entries, 0 to 1089042
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  -
 0   DR_NO           942531 non-null float64
 1   Date_Reported   942531 non-null object  
 2   Date_occured    942531 non-null object  
 3   Time_occured    942531 non-null float64
 4   Area            942531 non-null object  
 5   Crime_Code      942531 non-null object  
 6   Victim_age      942531 non-null float64
 7   Victim_sex      942531 non-null object  
 8   Victim_descent  942531 non-null object  
 9   Premis          942531 non-null object  
10   Weapon          942531 non-null object  
11   Status          942531 non-null object  
dtypes: float64(3), object(9)
memory usage: 93.5+ MB
```

Quedan 3 columnas de tipo float (numérico)

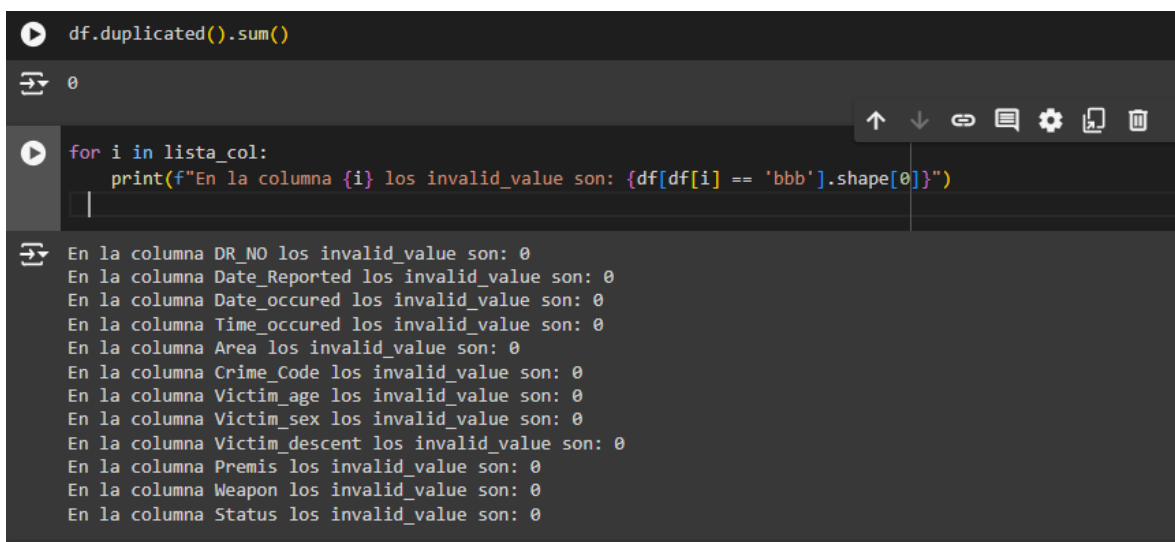
Tabla que muestre el porcentaje de valores faltantes final por columna.



```
df.isnull().mean()*100
```

	0
DR_NO	0.0
Date_Reported	0.0
Date_occured	0.0
Time_occured	0.0
Area	0.0
Crime_Code	0.0
Victim_age	0.0
Victim_sex	0.0
Victim_descent	0.0
Premis	0.0
Weapon	0.0
Status	0.0

Comprobación de que no hay duplicados ni valores inválidos.



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

```
0
```

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

```
En la columna DR_NO los invalid_value son: 0
En la columna Date_Reported los invalid_value son: 0
En la columna Date_occured los invalid_value son: 0
En la columna Time_occured los invalid_value son: 0
En la columna Area los invalid_value son: 0
En la columna Crime_Code los invalid_value son: 0
En la columna Victim_age los invalid_value son: 0
En la columna Victim_sex los invalid_value son: 0
En la columna Victim_descent los invalid_value son: 0
En la columna Premis los invalid_value son: 0
En la columna Weapon los invalid_value son: 0
En la columna Status los invalid_value son: 0
```