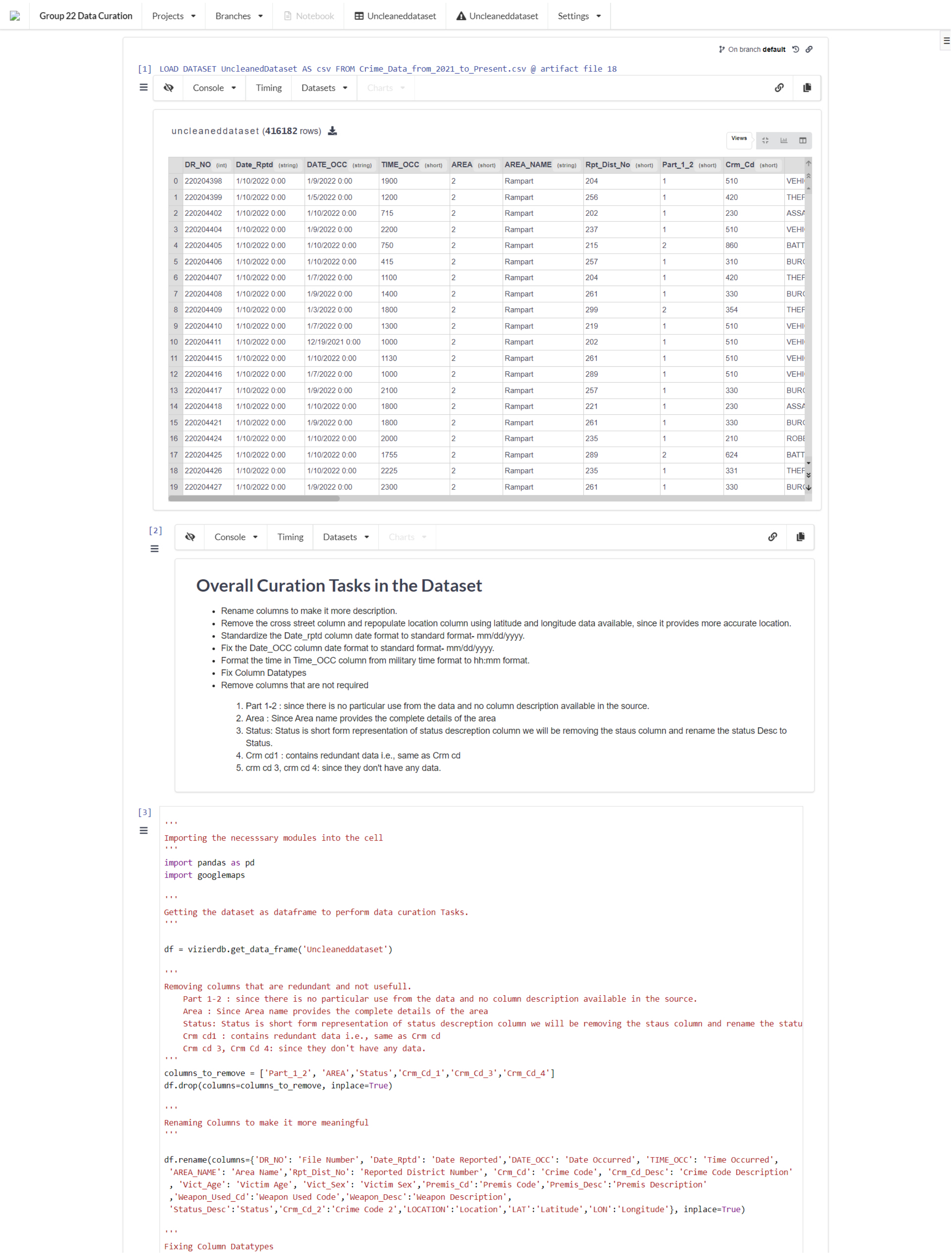


```
[1] LOAD DATASET UncleanedDataset AS csv FROM Crime_Data_from_2021_to_Present.csv @ artifact file 18
```

uncleaneddataset (**416182** rows)

	DR_NO (int)	Date_Rptd (string)	DATE_OCC (string)	TIME_OCC (short)	AREA (short)	AREA_NAME (string)	Rpt_Dist_No (short)	Part_1_2 (short)	Crm_Cd (short)	
0	220204398	1/10/2022 0:00	1/9/2022 0:00	1900	2	Rampart	204	1	510	VEHI
1	220204399	1/10/2022 0:00	1/5/2022 0:00	1200	2	Rampart	256	1	420	THEF
2	220204402	1/10/2022 0:00	1/10/2022 0:00	715	2	Rampart	202	1	230	ASSA
3	220204404	1/10/2022 0:00	1/9/2022 0:00	2200	2	Rampart	237	1	510	VEHI
4	220204405	1/10/2022 0:00	1/10/2022 0:00	750	2	Rampart	215	2	860	BATT
5	220204406	1/10/2022 0:00	1/10/2022 0:00	415	2	Rampart	257	1	310	BUR(
6	220204407	1/10/2022 0:00	1/7/2022 0:00	1100	2	Rampart	204	1	420	THEF
7	220204408	1/10/2022 0:00	1/9/2022 0:00	1400	2	Rampart	261	1	330	BUR(
8	220204409	1/10/2022 0:00	1/3/2022 0:00	1800	2	Rampart	299	2	354	THEF
9	220204410	1/10/2022 0:00	1/7/2022 0:00	1300	2	Rampart	219	1	510	VEHI
10	220204411	1/10/2022 0:00	12/19/2021 0:00	1000	2	Rampart	202	1	510	VEHI
11	220204415	1/10/2022 0:00	1/10/2022 0:00	1130	2	Rampart	261	1	510	VEHI
12	220204416	1/10/2022 0:00	1/7/2022 0:00	1000	2	Rampart	289	1	510	VEHI
13	220204417	1/10/2022 0:00	1/9/2022 0:00	2100	2	Rampart	257	1	330	BUR(
14	220204418	1/10/2022 0:00	1/10/2022 0:00	1800	2	Rampart	221	1	230	ASSA
15	220204421	1/10/2022 0:00	1/9/2022 0:00	1800	2	Rampart	261	1	330	BUR(
16	220204424	1/10/2022 0:00	1/10/2022 0:00	2000	2	Rampart	235	1	210	ROBE
17	220204425	1/10/2022 0:00	1/10/2022 0:00	1755	2	Rampart	289	2	624	BATT
18	220204426	1/10/2022 0:00	1/10/2022 0:00	2225	2	Rampart	235	1	331	THEF
19	220204427	1/10/2022 0:00	1/9/2022 0:00	2300	2	Rampart	261	1	330	BUR(

[2]

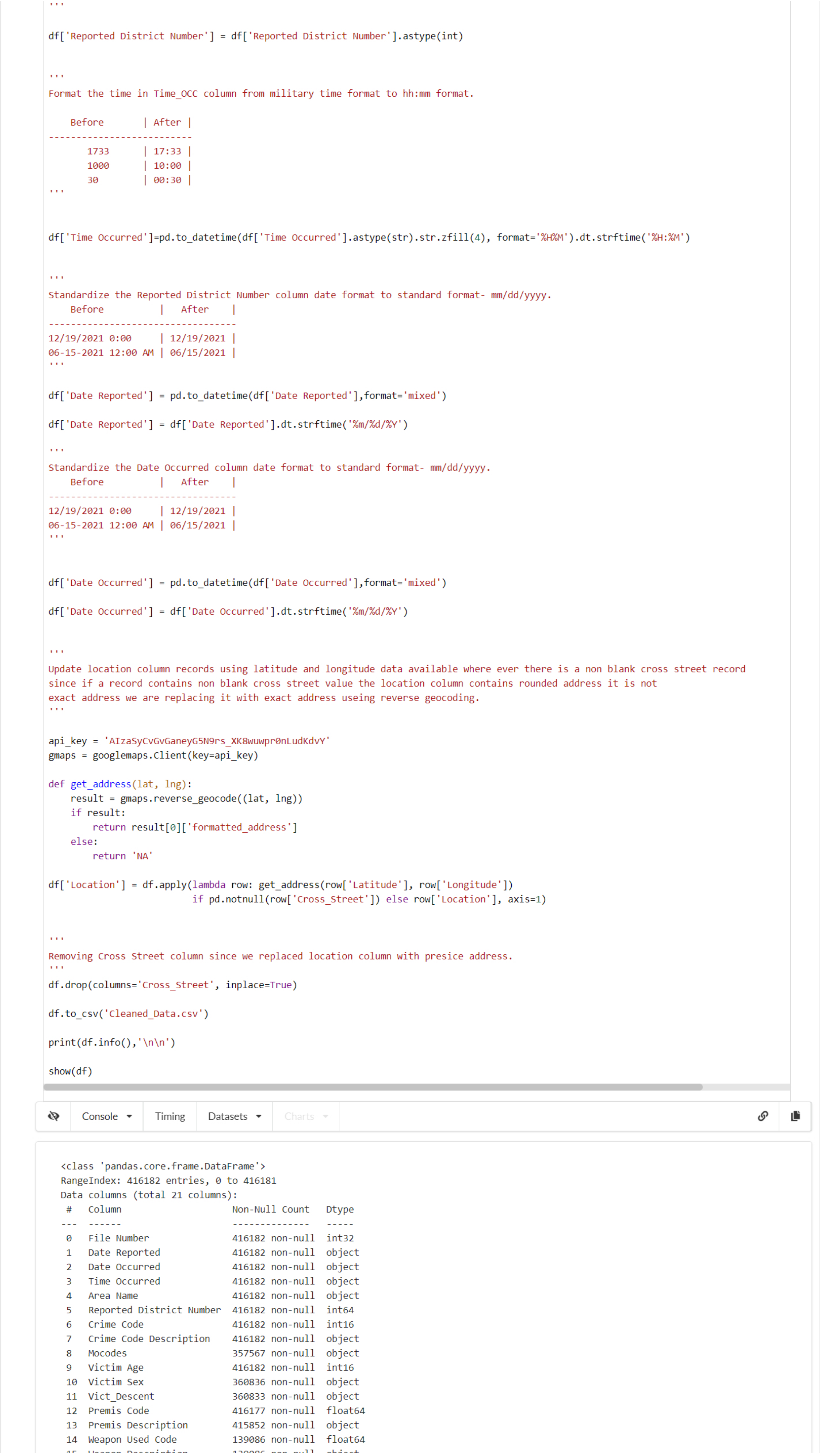
Overall Curation Tasks in the Dataset

- Rename columns to make it more description.
- Remove the cross street column and repopulate location column using latitude and longitude data available, since it provides more accurate location.
- Standardize the Date_rptd column date format to standard format- mm/dd/yyyy.
- Fix the Date_OCC column date format to standard format- mm/dd/yyyy.
- Format the time in Time_OCC column from military time format to hh:mm format.
- Fix Column Datatypes
- Remove columns that are not required

1. Part 1-2 : since there is no particular use from the data and no column description available in the source.
2. Area : Since Area name provides the complete details of the area
3. Status: Status is short form representation of status descreption column we will be removing the staus column and rename the status Desc to Status.
4. Crm cd1 : contains redundant data i.e., same as Crm cd
5. crm cd 3, crm cd 4: since they don't have any data.

[3]

```
'''
Importing the necesssary modules into the cell
'''
import pandas as pd
import googlemaps

'''
Getting the dataset as dataframe to perform data curation Tasks.
'''


df = vizierdb.get_data_frame('Uncleaneddataset')

'''
Removing columns that are redundant and not usefull.
    Part 1-2 : since there is no particular use from the data and no column description available in the source.
    Area : Since Area name provides the complete details of the area
    Status: Status is short form representation of status descreption column we will be removing the staus column and rename the statu
    Crm cd1 : contains redundant data i.e., same as Crm cd
    Crm cd 3, Crm Cd 4: since they don't have any data.
'''
columns_to_remove = ['Part_1_2', 'AREA','Status','Crm_Cd_1','Crm_Cd_3','Crm_Cd_4']
df.drop(columns=columns_to_remove, inplace=True)

'''
Renaming Columns to make it more meaningful
'''


df.rename(columns={'DR_NO': 'File Number', 'Date_Rptd': 'Date Reported','DATE_OCC': 'Date Occurred', 'TIME_OCC': 'Time Occurred',
 'AREA_NAME': 'Area Name','Rpt_Dist_No': 'Reported District Number', 'Crm_Cd': 'Crime Code', 'Crm_Cd_Desc': 'Crime Code Description'
 , 'Vict_Age': 'Victim Age', 'Vict_Sex': 'Victim Sex','Premis_Cd':'Premis Code','Premis_Desc':'Premis Description'
 ,'Weapon_Used_Cd':'Weapon Used Code','Weapon_Desc':'Weapon Description',
 'Status_Desc':'Status','Crm_Cd_2':'Crime Code 2','LOCATION':'Location','LAT':'Latitude','LON':'Longitude'}, inplace=True)

'''
Fixing Column Datatypes
```



```
'''

df['Reported District Number'] = df['Reported District Number'].astype(int)


'''
Format the time in Time_OCC column from military time format to hh:mm format.

    Before       | After |
--------------------------
       1733      | 17:33 |
       1000      | 10:00 |
       30        | 00:30 |
'''


df['Time Occurred']=pd.to_datetime(df['Time Occurred'].astype(str).str.zfill(4), format='%H%M').dt.strftime('%H:%M')


'''
Standardize the Reported District Number column date format to standard format- mm/dd/yyyy.
    Before          |   After    |
----------------------------------
12/19/2021 0:00     | 12/19/2021 |
06-15-2021 12:00 AM | 06/15/2021 |
'''


df['Date Reported'] = pd.to_datetime(df['Date Reported'],format='mixed')

df['Date Reported'] = df['Date Reported'].dt.strftime('%m/%d/%Y')

'''
Standardize the Date Occurred column date format to standard format- mm/dd/yyyy.
    Before          |   After    |
----------------------------------
12/19/2021 0:00     | 12/19/2021 |
06-15-2021 12:00 AM | 06/15/2021 |
'''



df['Date Occurred'] = pd.to_datetime(df['Date Occurred'],format='mixed')

df['Date Occurred'] = df['Date Occurred'].dt.strftime('%m/%d/%Y')


'''
Update location column records using latitude and longitude data available where ever there is a non blank cross street record
since if a record contains non blank cross street value the location column contains rounded address it is not
exact address we are replacing it with exact address useing reverse geocoding.
'''

api_key = 'AIzaSyCvGvGaneyG5N9rs_XK8wuwpr0nLudKdvY'
gmaps = googlemaps.Client(key=api_key)

def get_address(lat, lng):
    result = gmaps.reverse_geocode((lat, lng))
    if result:
        return result[0]['formatted_address']
    else:
        return 'NA'

df['Location'] = df.apply(lambda row: get_address(row['Latitude'], row['Longitude'])
                          if pd.notnull(row['Cross_Street']) else row['Location'], axis=1)


'''
Removing Cross Street column since we replaced location column with presice address.
'''
df.drop(columns='Cross_Street', inplace=True)

df.to_csv('Cleaned_Data.csv')

print(df.info(),'\n\n')

show(df)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 416182 entries, 0 to 416181
Data columns (total 21 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   File Number               416182 non-null  int32
 1   Date Reported             416182 non-null  object
 2   Date Occurred             416182 non-null  object
 3   Time Occurred             416182 non-null  object
 4   Area Name                 416182 non-null  object
 5   Reported District Number  416182 non-null  int64
 6   Crime Code                416182 non-null  int16
 7   Crime Code Description    416182 non-null  object
 8   Mocodes                   357567 non-null  object
 9   Victim Age                416182 non-null  int16
 10  Victim Sex                360836 non-null  object
 11  Vict_Descent              360833 non-null  object
 12  Premis Code               416177 non-null  float64
 13  Premis Description        415852 non-null  object
 14  Weapon Used Code          139086 non-null  float64
 15  Weapon Description        139086 non-null  object
 15  Weapon Description        139086 non-null  object
 16  Status                    416182 non-null  object
 17  Crime Code 2              28089 non-null   float64
 18  Location                  416182 non-null  object
 19  Latitude                  416182 non-null  float32
 20  Longitude                 416182 non-null  float32
dtypes: float32(2), float64(3), int16(2), int32(1), int64(1), object(12)
memory usage: 57.2+ MB
None
```

	File Number	Date Reported	Date Occurred	Time Occurred	Area Name	Reported District Number	Crime Code	Crime Code Description	Mocodes	Victim Age	Victim Sex	Vict_Descent	Premis Code	Premis D
0	220204398	01/10/2022	01/09/2022	19:00	Rampart	204	510	VEHICLE - STOLEN	None	0	None	None	108.0	PARKING L
1	220204399	01/10/2022	01/05/2022	12:00	Rampart	256	420	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	None	0	None	None	101.0	STREE
2	220204402	01/10/2022	01/10/2022	07:15	Rampart	202	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	2004 1817 2021 0913 0334 0445 0400	40	M	H	121.0	YARD (RESIDENTIAL/B
3	220204404	01/10/2022	01/09/2022	22:00	Rampart	237	510	VEHICLE - STOLEN	None	0	None	None	101.0	STREE
4	220204405	01/10/2022	01/10/2022	07:50	Rampart	215	860	BATTERY WITH SEXUAL CONTACT	1822 0216 0400 0522	15	F	H	102.0	SIDEWAL
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
416177	239905949	01/25/2023	01/25/2023	22:00	Newton	1309	210	ROBBERY	0344 0432 1309 0416 0305 0446 0342	29	M	H	108.0	PARKING L
416178	239906039	01/26/2023	01/26/2023	15:10	West Valley	1005	510	VEHICLE - STOLEN	1402	0	M	H	101.0	STREE
416179	239909037	03/03/2023	03/02/2023	20:00	Newton	1383	510	VEHICLE - STOLEN	1402	0	M	W	101.0	STREE
416180	239909747	03/12/2023	03/12/2023	15:00	Rampart	257	626	INTIMATE PARTNER - SIMPLE ASSAULT	2004 2000 0416 0446 0913	30	M	B	710.0	OTHER PRE
416181	239916487	06/04/2023	06/04/2023	19:30	77th Street	1248	510	VEHICLE - STOLEN	1402 1822 0344	0	X	X	101.0	STREE

416182 rows × 21 columns

➕

Connected to vizier @ http://localhost:5001/vizier-db/api/v1/