In [8]:
```
import pandas as pd

# List of CSV file paths
file_paths = [
    '/Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicago_Crimes_2
    '/Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicago_Crimes_2
    '/Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicago_Crimes_2
    '/Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicago_Crimes_2
]

# Loop through each file and read the first 5 rows and column names
for file_path in file_paths:
    try:
        # Read the CSV file
        df = pd.read_csv(file_path, nrows=5)

        # Display the first 5 rows
        print(f"First 5 rows of {file_path}:")
        print(df)

        # Display the column names
        print(f"Column names in {file_path}:")
        print(df.columns.tolist())
        print("\n" + "-"*50 + "\n")

    except Exception as e:
        print(f"Error reading {file_path}: {e}")
```

```
First 5 rows of /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2001_to_2004.csv:
   Unnamed: 0       ID Case Number                    Date  \
0         879  4786321    HM399414  01/01/2004 12:01:00 AM
1        2544  4676906    HM278933  03/01/2003 12:00:00 AM
2        2919  4789749    HM402220  06/20/2004 11:00:00 AM
3        2927  4789765    HM402058  12/30/2004 08:00:00 PM
4        3302  4677901    HM275615  05/01/2003 01:00:00 AM

                Block  IUCR               Primary Type  \
0    082XX S COLES AVE   840                      THEFT
1      004XX W 42ND PL  2825             OTHER OFFENSE
2  025XX N KIMBALL AVE  1752  OFFENSE INVOLVING CHILDREN
3   045XX W MONTANA ST   840                      THEFT
4   111XX S NORMAL AVE   841                      THEFT

                     Description Location Description  Arrest  ...  Wa
rd  \
0   FINANCIAL ID THEFT: OVER $300           RESIDENCE   False  ...
7.0
1         HARASSMENT BY TELEPHONE           RESIDENCE   False  ...  1
1.0
2   AGG CRIM SEX ABUSE FAM MEMBER           RESIDENCE   False  ...  3
```

```
5.0
3   FINANCIAL ID THEFT: OVER $300               OTHER   False  ...  3
1.0
4  FINANCIAL ID THEFT:$300 &UNDER            RESIDENCE   False  ...  3
4.0

   Community Area  FBI Code  X Coordinate  Y Coordinate  Year  \
0            46.0         6           NaN           NaN  2004
1            61.0        26     1173974.0     1876757.0  2003
2            22.0        20           NaN           NaN  2004
3            20.0         6           NaN           NaN  2004
4            49.0         6     1174948.0     1831051.0  2003

               Updated On   Latitude  Longitude                       L
ocation
0  08/17/2015 03:03:40 PM        NaN        NaN
NaN
1  04/15/2016 08:55:02 AM  41.817229 -87.637328  (41.817229156, -87.637
328162)
2  08/17/2015 03:03:40 PM        NaN        NaN
NaN
3  08/17/2015 03:03:40 PM        NaN        NaN
NaN
4  04/15/2016 08:55:02 AM  41.691785 -87.635116  (41.691784636, -87.635
115968)

[5 rows x 23 columns]
Column names in /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2001_to_2004.csv:
['Unnamed: 0', 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary T
ype', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Bea
t', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', '
Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Locatio
n']

--------------------------------------------------

First 5 rows of /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2008_to_2011.csv:
   Unnamed: 0    ID Case Number                    Date
Block  \
0         388  4785    HP610824  10/07/2008 12:39:00 PM      000XX E 75
TH ST
1         835  4786    HP616595  10/09/2008 03:30:00 AM      048XX W PO
LK ST
2        1334  4787    HP616904  10/09/2008 08:35:00 AM      030XX W MA
NN DR
3        1907  4788    HP618616  10/10/2008 02:33:00 AM  052XX W CHICAG
O AVE
4        2436  4789    HP619020  10/10/2008 12:50:00 PM    026XX S HOMA
N AVE
```

```
   IUCR Primary Type          Description Location Description  Arrest
...  \
0   110     HOMICIDE  FIRST DEGREE MURDER                ALLEY    True
...
1   110     HOMICIDE  FIRST DEGREE MURDER               STREET    True
...
2   110     HOMICIDE  FIRST DEGREE MURDER        PARK PROPERTY   False
...
3   110     HOMICIDE  FIRST DEGREE MURDER           RESTAURANT   False
...
4   110     HOMICIDE  FIRST DEGREE MURDER               GARAGE   False
...

   Ward  Community Area  FBI Code  X Coordinate  Y Coordinate  Year  \
0   6.0            69.0       01A     1178207.0     1855308.0  2008
1  24.0            25.0       01A     1144200.0     1895857.0  2008
2  18.0            66.0       01A     1157314.0     1859778.0  2008
3  37.0            25.0       01A     1141065.0     1904824.0  2008
4  22.0            30.0       01A     1154123.0     1886297.0  2008

               Updated On   Latitude  Longitude                       L
ocation
0  08/17/2015 03:03:40 PM  41.758276 -87.622451  (41.758275857, -87.622
451031)
1  08/17/2015 03:03:40 PM  41.870252 -87.746069   (41.87025207, -87.746
069362)
2  08/17/2015 03:03:40 PM  41.770990 -87.698901  (41.770990476, -87.698
901469)
3  08/17/2015 03:03:40 PM  41.894917 -87.757358  (41.894916924, -87.757
358147)
4  08/17/2015 03:03:40 PM  41.843826 -87.709893  (41.843826272, -87.709
893465)

[5 rows x 23 columns]
Column names in /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2008_to_2011.csv:
['Unnamed: 0', 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary T
ype', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Bea
t', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', '
Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Locatio
n']

________________________________________________

First 5 rows of /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2005_to_2007.csv:
   Unnamed: 0       ID Case Number                    Date  \
0           0  4673626    HM274058  04/02/2006 01:00:00 PM
1           1  4673627    HM202199  02/26/2006 01:40:48 PM
2           2  4673628    HM113861  01/08/2006 11:16:00 PM
3           4  4673629    HM274049  04/05/2006 06:45:00 PM
4           5  4673630    HM187120  02/17/2006 09:03:14 PM
```

```
                 Block  IUCR   Primary Type                    Descripti
on  \
0    055XX N MANGO AVE  2825  OTHER OFFENSE       HARASSMENT BY TELEPHO
NE
1   065XX S RHODES AVE  2017      NARCOTICS            MANU/DELIVER:CRA
CK
2      013XX E 69TH ST  051A        ASSAULT           AGGRAVATED: HANDG
UN
3  061XX W NEWPORT AVE  0460        BATTERY                      SIMP
LE
4      037XX W 60TH ST  1811      NARCOTICS  POSS: CANNABIS 30GMS OR LE
SS

  Location Description  Arrest  ...  Ward  Community Area  FBI Code  \
0            RESIDENCE   False  ...  45.0            11.0        26
1             SIDEWALK    True  ...  20.0            42.0        18
2                OTHER   False  ...   5.0            69.0       04A
3            RESIDENCE   False  ...  38.0            17.0       08B
4                ALLEY    True  ...  13.0            65.0        18

   X Coordinate  Y Coordinate  Year              Updated On   Latitude
\
0     1136872.0     1936499.0  2006  04/15/2016 08:55:02 AM  41.981913
1     1181027.0     1861693.0  2006  04/15/2016 08:55:02 AM  41.775733
2     1186023.0     1859609.0  2006  04/15/2016 08:55:02 AM  41.769897
3     1134772.0     1922299.0  2006  04/15/2016 08:55:02 AM  41.942984
4     1152412.0     1864560.0  2006  04/15/2016 08:55:02 AM  41.784211

   Longitude                       Location
0 -87.771996  (41.981912692, -87.771996382)
1 -87.611920  (41.775732538, -87.611919814)
2 -87.593671  (41.769897392, -87.593670899)
3 -87.780057  (41.942984005, -87.780056951)
4 -87.716745   (41.784210853, -87.71674491)

[5 rows x 23 columns]
Column names in /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2005_to_2007.csv:
['Unnamed: 0', 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary T
ype', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Bea
t', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', '
Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Locatio
n']

--------------------------------------------------

First 5 rows of /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2012_to_2017.csv:
   Unnamed: 0        ID Case Number                    Date  \
0           3  10508693    HZ250496  05/03/2016 11:40:00 PM
1          89  10508695    HZ250409  05/03/2016 09:40:00 PM
```

```
2         197  10508697    HZ250503  05/03/2016 11:31:00 PM
3         673  10508698    HZ250424  05/03/2016 10:10:00 PM
4         911  10508699    HZ250455  05/03/2016 10:00:00 PM

                 Block  IUCR            Primary Type              Descr
iption  \
0   013XX S SAWYER AVE   486                 BATTERY  DOMESTIC BATTERY
SIMPLE
1   061XX S DREXEL AVE   486                 BATTERY  DOMESTIC BATTERY
SIMPLE
2  053XX W CHICAGO AVE   470  PUBLIC PEACE VIOLATION         RECKLESS C
ONDUCT
3    049XX W FULTON ST   460                 BATTERY
SIMPLE
4    003XX N LOTUS AVE   820                   THEFT           $500 AND
UNDER

  Location Description  Arrest  ...  Ward  Community Area  FBI Code  \
0            APARTMENT    True  ...  24.0            29.0       08B
1            RESIDENCE   False  ...  20.0            42.0       08B
2               STREET   False  ...  37.0            25.0        24
3             SIDEWALK   False  ...  28.0            25.0       08B
4            RESIDENCE   False  ...  28.0            25.0        06

   X Coordinate  Y Coordinate  Year              Updated On   Latitude
\
0     1154907.0     1893681.0  2016  05/10/2016 03:56:50 PM  41.864073
1     1183066.0     1864330.0  2016  05/10/2016 03:56:50 PM  41.782922
2     1140789.0     1904819.0  2016  05/10/2016 03:56:50 PM  41.894908
3     1143223.0     1901475.0  2016  05/10/2016 03:56:50 PM  41.885687
4     1139890.0     1901675.0  2016  05/10/2016 03:56:50 PM  41.886297

   Longitude                      Location
0 -87.706819  (41.864073157, -87.706818608)
1 -87.604363   (41.782921527, -87.60436317)
2 -87.758372  (41.894908283, -87.758371958)
3 -87.749516  (41.885686845, -87.749515983)
4 -87.761751  (41.886297242, -87.761750709)

[5 rows x 23 columns]
Column names in /Volumes/PENDRIVE/Projects/Python/Task_2/archive/Chicag
o_Crimes_2012_to_2017.csv:
['Unnamed: 0', 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary T
ype', 'Description', 'Location Description', 'Arrest', 'Domestic', 'Bea
t', 'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate', '
Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Locatio
n']

--------------------------------------------------
```

In [10]:
```
import pandas as pd
```

```
# Load each dataset with the option to skip problematic lines
crime_data_2001_2004 = pd.read_csv('/Volumes/PENDRIVE/Projects/Python/
crime_data_2005_2007 = pd.read_csv('/Volumes/PENDRIVE/Projects/Python/
crime_data_2008_2011 = pd.read_csv('/Volumes/PENDRIVE/Projects/Python/
crime_data_2012_2017 = pd.read_csv('/Volumes/PENDRIVE/Projects/Python/

# Combine all datasets into a single DataFrame
combined_data = pd.concat([crime_data_2001_2004, crime_data_2005_2007,

# Optionally, save the combined data to a new CSV file
combined_data.to_csv('/Volumes/PENDRIVE/Projects/Python/Task_2/archive

# Check the first 5 rows
print(combined_data.head())
```

```
/var/folders/83/sl_nw5nd0pv4njhhbnqkzwvr0000gn/T/ipykernel_2000/2840699
083.py:4: DtypeWarning: Columns (17,20) have mixed types. Specify dtype
option on import or set low_memory=False.
  crime_data_2001_2004 = pd.read_csv('/Volumes/PENDRIVE/Projects/Pytho
n/Task_2/archive/Chicago_Crimes_2001_to_2004.csv', on_bad_lines='skip')
```

```
   Unnamed: 0       ID Case Number                    Date  \
0         879  4786321    HM399414  01/01/2004 12:01:00 AM
1        2544  4676906    HM278933  03/01/2003 12:00:00 AM
2        2919  4789749    HM402220  06/20/2004 11:00:00 AM
3        2927  4789765    HM402058  12/30/2004 08:00:00 PM
4        3302  4677901    HM275615  05/01/2003 01:00:00 AM

                 Block  IUCR               Primary Type  \
0    082XX S COLES AVE  0840                      THEFT
1      004XX W 42ND PL  2825              OTHER OFFENSE
2  025XX N KIMBALL AVE  1752  OFFENSE INVOLVING CHILDREN
3   045XX W MONTANA ST  0840                      THEFT
4   111XX S NORMAL AVE  0841                      THEFT

                     Description Location Description  Arrest  ...  Wa
rd  \
0   FINANCIAL ID THEFT: OVER $300            RESIDENCE   False  ...
7.0
1         HARASSMENT BY TELEPHONE            RESIDENCE   False  ...  1
1.0
2   AGG CRIM SEX ABUSE FAM MEMBER            RESIDENCE   False  ...  3
5.0
3   FINANCIAL ID THEFT: OVER $300                OTHER   False  ...  3
1.0
4  FINANCIAL ID THEFT:$300 &UNDER            RESIDENCE   False  ...  3
4.0

   Community Area  FBI Code  X Coordinate  Y Coordinate    Year  \
0            46.0        06           NaN           NaN  2004.0
1            61.0        26     1173974.0     1876757.0  2003.0
2            22.0        20           NaN           NaN  2004.0
3            20.0        06           NaN           NaN  2004.0
4            49.0        06     1174948.0     1831051.0  2003.0

              Updated On   Latitude  Longitude                      L
ocation
0  08/17/2015 03:03:40 PM        NaN        NaN
NaN
1  04/15/2016 08:55:02 AM  41.817229 -87.637328  (41.817229156, -87.637
328162)
2  08/17/2015 03:03:40 PM        NaN        NaN
NaN
3  08/17/2015 03:03:40 PM        NaN        NaN
NaN
4  04/15/2016 08:55:02 AM  41.691785 -87.635116  (41.691784636, -87.635
115968)

[5 rows x 23 columns]
```

Data Cleaning

Handel Missing Values

In [20]:
```
# Drop rows with missing values
crime_data_2001_2004 = crime_data_2001_2004.dropna()
crime_data_2001_2004
```

Out[20]:

	ID	Case Number	Date	Block	IUCR	Primary Type	Des
1	4676906	HM278933	2003-03-01 00:00:00	004XX W 42ND PL	2825	OTHER OFFENSE	HARAS TELE
4	4677901	HM275615	2003-05-01 01:00:00	111XX S NORMAL AVE	0841	THEFT	FINAN THE &
6	4791194	HM403711	2001-01-01 11:00:00	114XX S ST LAWRENCE AVE	0266	CRIM SEXUAL ASSAULT	PRE
7	4679521	HM216293	2003-03-15 00:00:00	090XX S RACINE AVE	5007	OTHER OFFENSE	WE VIC
9	4680124	HM282389	2003-01-01 00:00:00	009XX S SPAULDING AVE	0840	THEFT	FINAN THEF
...	...	...	...	...	...	...	
1923510	4781176	HM386461	2001-04-01 09:00:00	023XX N LATROBE AVE	0841	THEFT	FINAN THE &
1923511	4671197	HM270817	2003-09-01 00:01:00	045XX N MOBILE AVE	0840	THEFT	FINAN THEF
1923512	4671380	HM269330	2002-08-01 09:00:00	020XX W 82ND PL	0840	THEFT	FINAN THEF
1923513	4782588	HM394550	2001-06-04 00:01:00	087XX S MUSKEGON AVE	0610	BURGLARY	FC
1923514	4673324	HM274913	2002-08-09 15:00:00	067XX S CHAMPLAIN AVE	0840	THEFT	FINAN THEF

1205641 rows × 22 columns

In [21]:
```
# Fill Missing values with Place holders (eg,0,unknown)
```

```
crime_data_2001_2004['X Coordinate'] = crime_data_2001_2004['X Coordin
crime_data_2001_2004['Y Coordinate'] = crime_data_2001_2004['Y Coordin
crime_data_2001_2004['Latitude'] = crime_data_2001_2004['Latitude'].fi
crime_data_2001_2004['Longitude'] = crime_data_2001_2004['Longitude'].
crime_data_2001_2004
```

Out[21]:

	ID	Case Number	Date	Block	IUCR	Primary Type	Des
1	4676906	HM278933	2003-03-01 00:00:00	004XX W 42ND PL	2825	OTHER OFFENSE	HARAS TELE
4	4677901	HM275615	2003-05-01 01:00:00	111XX S NORMAL AVE	0841	THEFT	FINAN THEF &
6	4791194	HM403711	2001-01-01 11:00:00	114XX S ST LAWRENCE AVE	0266	CRIM SEXUAL ASSAULT	PREI
7	4679521	HM216293	2003-03-15 00:00:00	090XX S RACINE AVE	5007	OTHER OFFENSE	WE VIC
9	4680124	HM282389	2003-01-01 00:00:00	009XX S SPAULDING AVE	0840	THEFT	FINAN THEF
...	...	...	...	...	...	...	
1923510	4781176	HM386461	2001-04-01 09:00:00	023XX N LATROBE AVE	0841	THEFT	FINAN THEF &
1923511	4671197	HM270817	2003-09-01 00:01:00	045XX N MOBILE AVE	0840	THEFT	FINAN THEF
1923512	4671380	HM269330	2002-08-01 09:00:00	020XX W 82ND PL	0840	THEFT	FINAN THEF
1923513	4782588	HM394550	2001-06-04 00:01:00	087XX S MUSKEGON AVE	0610	BURGLARY	FC
1923514	4673324	HM274913	2002-08-09 15:00:00	067XX S CHAMPLAIN AVE	0840	THEFT	FINAN THEF

1205641 rows × 22 columns

Convert Data Type

In [22]:
```
crime_data_2001_2004['Date'] = pd.to_datetime(crime_data_2001_2004['Da
crime_data_2001_2004['Year'] = pd.to_numeric(crime_data_2001_2004['Yea
crime_data_2001_2004['Latitude'] = pd.to_numeric(crime_data_2001_2004[
crime_data_2001_2004['Longitude'] = pd.to_numeric(crime_data_2001_2004
crime_data_2001_2004
```

Out[22]:

	ID	Case Number	Date	Block	IUCR	Primary Type	Des
1	4676906	HM278933	2003-03-01 00:00:00	004XX W 42ND PL	2825	OTHER OFFENSE	HARAS TELE
4	4677901	HM275615	2003-05-01 01:00:00	111XX S NORMAL AVE	0841	THEFT	FINAN THEI &
6	4791194	HM403711	2001-01-01 11:00:00	114XX S ST LAWRENCE AVE	0266	CRIM SEXUAL ASSAULT	PREI
7	4679521	HM216293	2003-03-15 00:00:00	090XX S RACINE AVE	5007	OTHER OFFENSE	WI VIC
9	4680124	HM282389	2003-01-01 00:00:00	009XX S SPAULDING AVE	0840	THEFT	FINAN THEF
...	...	...	...	...	...	...	
1923510	4781176	HM386461	2001-04-01 09:00:00	023XX N LATROBE AVE	0841	THEFT	FINAN THEI &
1923511	4671197	HM270817	2003-09-01 00:01:00	045XX N MOBILE AVE	0840	THEFT	FINAN THEF
1923512	4671380	HM269330	2002-08-01 09:00:00	020XX W 82ND PL	0840	THEFT	FINAN THEF
1923513	4782588	HM394550	2001-06-04 00:01:00	087XX S MUSKEGON AVE	0610	BURGLARY	FC
1923514	4673324	HM274913	2002-08-09 15:00:00	067XX S CHAMPLAIN AVE	0840	THEFT	FINAN THEF

1205641 rows × 22 columns

In [26]:
```
print(crime_data_2001_2004.info())  # To check the data types and non-
print(crime_data_2001_2004.head())  # To inspect the cleaned data
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1205641 entries, 1 to 1923514
Data columns (total 22 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   ID                    1205641 non-null  int64
 1   Case Number           1205641 non-null  object
 2   Date                  1205641 non-null  datetime64[ns]
 3   Block                 1205641 non-null  object
 4   IUCR                  1205641 non-null  object
 5   Primary Type          1205641 non-null  object
 6   Description           1205641 non-null  object
 7   Location Description  1205641 non-null  object
 8   Arrest                1205641 non-null  bool
 9   Domestic              1205641 non-null  bool
 10  Beat                  1205641 non-null  int64
 11  District              1205641 non-null  float64
 12  Ward                  1205641 non-null  float64
 13  Community Area        1205641 non-null  float64
 14  FBI Code              1205641 non-null  object
 15  X Coordinate          1205641 non-null  float64
 16  Y Coordinate          1205641 non-null  object
 17  Year                  1205641 non-null  float64
 18  Updated On            1205641 non-null  object
 19  Latitude              1205641 non-null  float64
 20  Longitude             1205641 non-null  float64
 21  Location              1205641 non-null  object
dtypes: bool(2), datetime64[ns](1), float64(7), int64(2), object(10)
memory usage: 195.5+ MB
None
        ID Case Number                Date                     Block  IU
CR  \
1  4676906    HM278933 2003-03-01 00:00:00          004XX W 42ND PL  28
25
4  4677901    HM275615 2003-05-01 01:00:00       111XX S NORMAL AVE  08
41
6  4791194    HM403711 2001-01-01 11:00:00  114XX S ST LAWRENCE AVE  02
66
7  4679521    HM216293 2003-03-15 00:00:00       090XX S RACINE AVE  50
07
9  4680124    HM282389 2003-01-01 00:00:00    009XX S SPAULDING AVE  08
40

          Primary Type                     Description  \
1        OTHER OFFENSE         HARASSMENT BY TELEPHONE
4                THEFT  FINANCIAL ID THEFT:$300 &UNDER
6  CRIM SEXUAL ASSAULT                       PREDATORY
7        OTHER OFFENSE         OTHER WEAPONS VIOLATION
9                THEFT   FINANCIAL ID THEFT: OVER $300

      Location Description  Arrest  Domestic  ...  Ward  Community Area
\
```

```
1               RESIDENCE   False      True  ...  11.0           61.0
4               RESIDENCE   False     False  ...  34.0           49.0
6               RESIDENCE    True      True  ...   9.0           50.0
7  RESIDENCE PORCH/HALLWAY   False     False  ...  21.0           73.0
9               RESIDENCE   False     False  ...  24.0           29.0

   FBI Code  X Coordinate Y Coordinate    Year              Updated On
\
1        26     1173974.0    1876757.0  2003.0  04/15/2016 08:55:02 AM
4        06     1174948.0    1831051.0  2003.0  04/15/2016 08:55:02 AM
6        02     1182247.0    1829375.0  2001.0  08/29/2006 03:46:28 AM
7        26     1169911.0    1844832.0  2003.0  04/15/2016 08:55:02 AM
9        06     1154521.0    1895755.0  2003.0  04/15/2016 08:55:02 AM

    Latitude  Longitude                      Location
1  41.817229 -87.637328  (41.817229156, -87.637328162)
4  41.691785 -87.635116  (41.691784636, -87.635115968)
6  41.687020 -87.608445   (41.687020002, -87.60844523)
7  41.729712 -87.653159  (41.729712374, -87.653158513)
9  41.869772 -87.708180  (41.869772159, -87.708180162)

[5 rows x 22 columns]
```

EDA

Basic Information

In [27]:
```
# Basic information about the dataset
print(crime_data_2001_2004.info())

# Summary statistics for numerical columns
print(crime_data_2001_2004.describe())

# Display first few rows
print(crime_data_2001_2004.head())
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1205641 entries, 1 to 1923514
Data columns (total 22 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   ID                    1205641 non-null  int64
 1   Case Number           1205641 non-null  object
 2   Date                  1205641 non-null  datetime64[ns]
 3   Block                 1205641 non-null  object
 4   IUCR                  1205641 non-null  object
 5   Primary Type          1205641 non-null  object
 6   Description           1205641 non-null  object
 7   Location Description  1205641 non-null  object
 8   Arrest                1205641 non-null  bool
 9   Domestic              1205641 non-null  bool
```

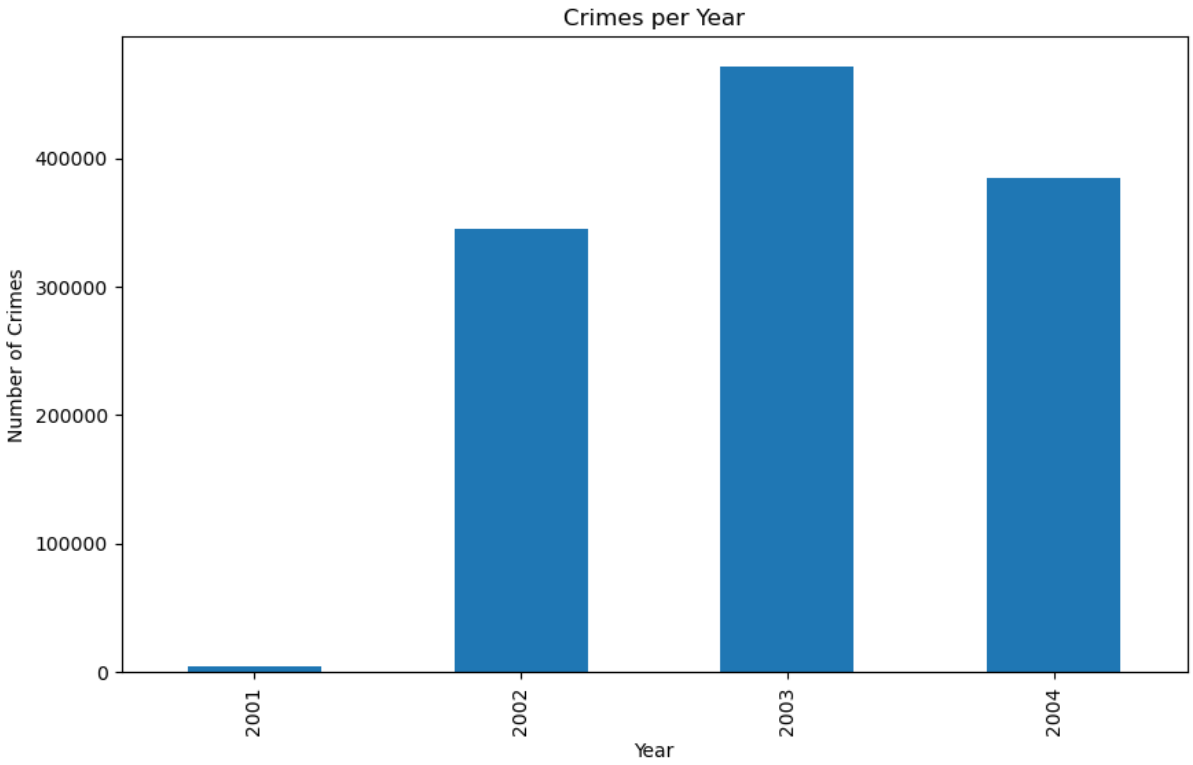
```
 10  Beat                  1205641 non-null  int64
 11  District              1205641 non-null  float64
 12  Ward                  1205641 non-null  float64
 13  Community Area        1205641 non-null  float64
 14  FBI Code              1205641 non-null  object
 15  X Coordinate          1205641 non-null  float64
 16  Y Coordinate          1205641 non-null  object
 17  Year                  1205641 non-null  float64
 18  Updated On            1205641 non-null  object
 19  Latitude              1205641 non-null  float64
 20  Longitude             1205641 non-null  float64
 21  Location              1205641 non-null  object
dtypes: bool(2), datetime64[ns](1), float64(7), int64(2), object(10)
memory usage: 195.5+ MB
None
                 ID                           Date          Beat  \
count  1.205641e+06                        1205641  1.205641e+06
mean   2.883402e+06  2003-08-02 16:03:22.272272640  1.212532e+03
min    6.340000e+02            2001-01-01 00:00:00  1.110000e+02
25%    2.473859e+06            2002-11-20 19:10:00  6.240000e+02
50%    2.857303e+06            2003-07-19 14:15:00  1.113000e+03
75%    3.343199e+06            2004-05-15 23:10:00  1.811000e+03
max    9.231682e+06            2004-12-31 23:59:00  2.535000e+03
std    5.083073e+05                            NaN  7.066955e+02

           District          Ward  Community Area  X Coordinate
Year  \
count  1.205641e+06  1.205641e+06    1.205641e+06  1.205641e+06  1.2056
41e+06
mean   1.134655e+01  2.244538e+01    3.729760e+01  1.164530e+06  2.0030
25e+03
min    1.000000e+00  1.000000e+00    0.000000e+00  1.095509e+06  2.0010
00e+03
25%    6.000000e+00  1.000000e+01    2.300000e+01  1.153200e+06  2.0020
00e+03
50%    1.000000e+01  2.200000e+01    3.200000e+01  1.165887e+06  2.0030
00e+03
75%    1.700000e+01  3.300000e+01    5.600000e+01  1.176244e+06  2.0040
00e+03
max    3.100000e+01  5.000000e+01    7.700000e+01  1.205119e+06  2.0040
00e+03
std    6.955229e+00  1.392323e+01    2.147243e+01  1.599084e+04  7.8756
96e-01

           Latitude     Longitude
count  1.205641e+06  1.205641e+06
mean   4.184317e+01 -8.767175e+01
min    4.164459e+01 -8.792430e+01
25%    4.177124e+01 -8.771278e+01
50%    4.185520e+01 -8.766659e+01
75%    4.190834e+01 -8.762894e+01
max    4.202255e+01 -8.752453e+01
```

```
std    8.603731e-02  5.819350e-02
        ID Case Number                Date                    Block  IU
CR  \
1  4676906    HM278933 2003-03-01 00:00:00          004XX W 42ND PL  28
25
4  4677901    HM275615 2003-05-01 01:00:00       111XX S NORMAL AVE  08
41
6  4791194    HM403711 2001-01-01 11:00:00  114XX S ST LAWRENCE AVE  02
66
7  4679521    HM216293 2003-03-15 00:00:00       090XX S RACINE AVE  50
07
9  4680124    HM282389 2003-01-01 00:00:00    009XX S SPAULDING AVE  08
40

          Primary Type                     Description  \
1        OTHER OFFENSE         HARASSMENT BY TELEPHONE
4                THEFT  FINANCIAL ID THEFT:$300 &UNDER
6  CRIM SEXUAL ASSAULT                       PREDATORY
7        OTHER OFFENSE         OTHER WEAPONS VIOLATION
9                THEFT   FINANCIAL ID THEFT: OVER $300

      Location Description  Arrest  Domestic  ...  Ward  Community Area
\
1                RESIDENCE   False      True  ...  11.0            61.0
4                RESIDENCE   False     False  ...  34.0            49.0
6                RESIDENCE    True      True  ...   9.0            50.0
7  RESIDENCE PORCH/HALLWAY   False     False  ...  21.0            73.0
9                RESIDENCE   False     False  ...  24.0            29.0

   FBI Code  X Coordinate Y Coordinate    Year              Updated On
\
1        26     1173974.0    1876757.0  2003.0  04/15/2016 08:55:02 AM
4        06     1174948.0    1831051.0  2003.0  04/15/2016 08:55:02 AM
6        02     1182247.0    1829375.0  2001.0  08/29/2006 03:46:28 AM
7        26     1169911.0    1844832.0  2003.0  04/15/2016 08:55:02 AM
9        06     1154521.0    1895755.0  2003.0  04/15/2016 08:55:02 AM

    Latitude  Longitude                       Location
1  41.817229 -87.637328  (41.817229156, -87.637328162)
4  41.691785 -87.635116  (41.691784636, -87.635115968)
6  41.687020 -87.608445   (41.687020002, -87.60844523)
7  41.729712 -87.653159  (41.729712374, -87.653158513)
9  41.869772 -87.708180  (41.869772159, -87.708180162)

[5 rows x 22 columns]
```

Date Analysis

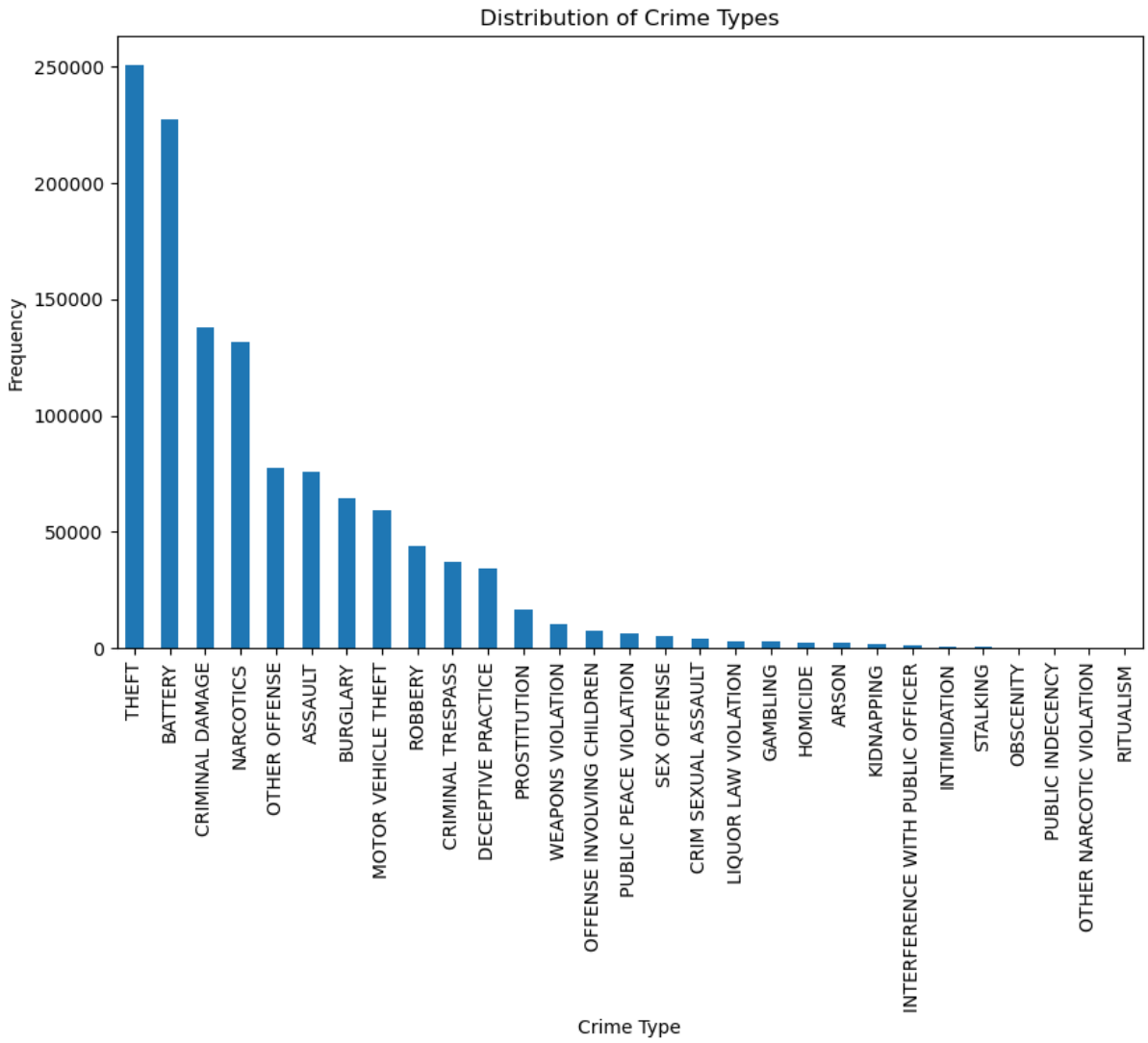
In [31]:
```
# Convert Date to datetime format
crime_data_2001_2004['Date'] = pd.to_datetime(crime_data_2001_2004['Da

# Extract year, month, and day from the Date column
```

```
crime_data_2001_2004['Year'] = crime_data_2001_2004['Date'].dt.year
crime_data_2001_2004['Month'] = crime_data_2001_2004['Date'].dt.month
crime_data_2001_2004['Day'] = crime_data_2001_2004['Date'].dt.day

# Plot crimes per year
plt.figure(figsize=(10, 6))
crime_data_2001_2004.groupby('Year').size().plot(kind='bar')
plt.title('Crimes per Year')
plt.xlabel('Year')
plt.ylabel('Number of Crimes')
plt.show()
```



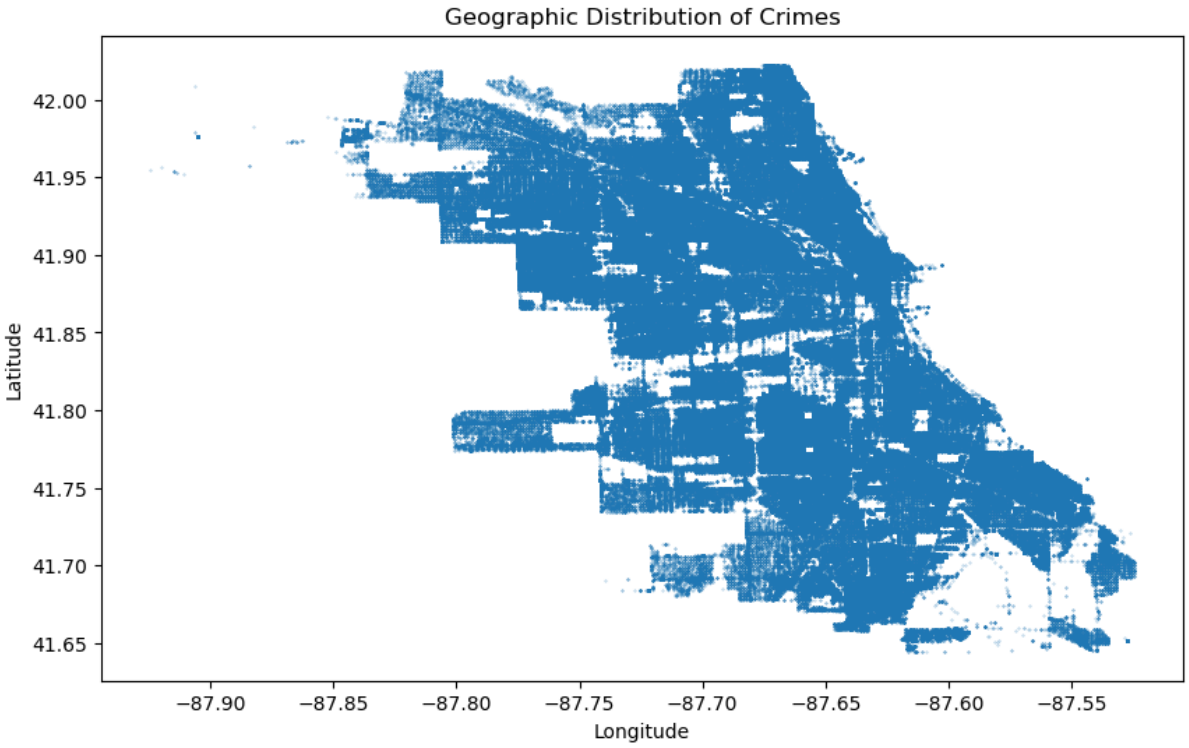
Crime Type Analysis

In [32]:
```
# Distribution of crime types (Primary Type)
plt.figure(figsize=(10, 6))
crime_data_2001_2004['Primary Type'].value_counts().plot(kind='bar')
plt.title('Distribution of Crime Types')
plt.xlabel('Crime Type')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()
```



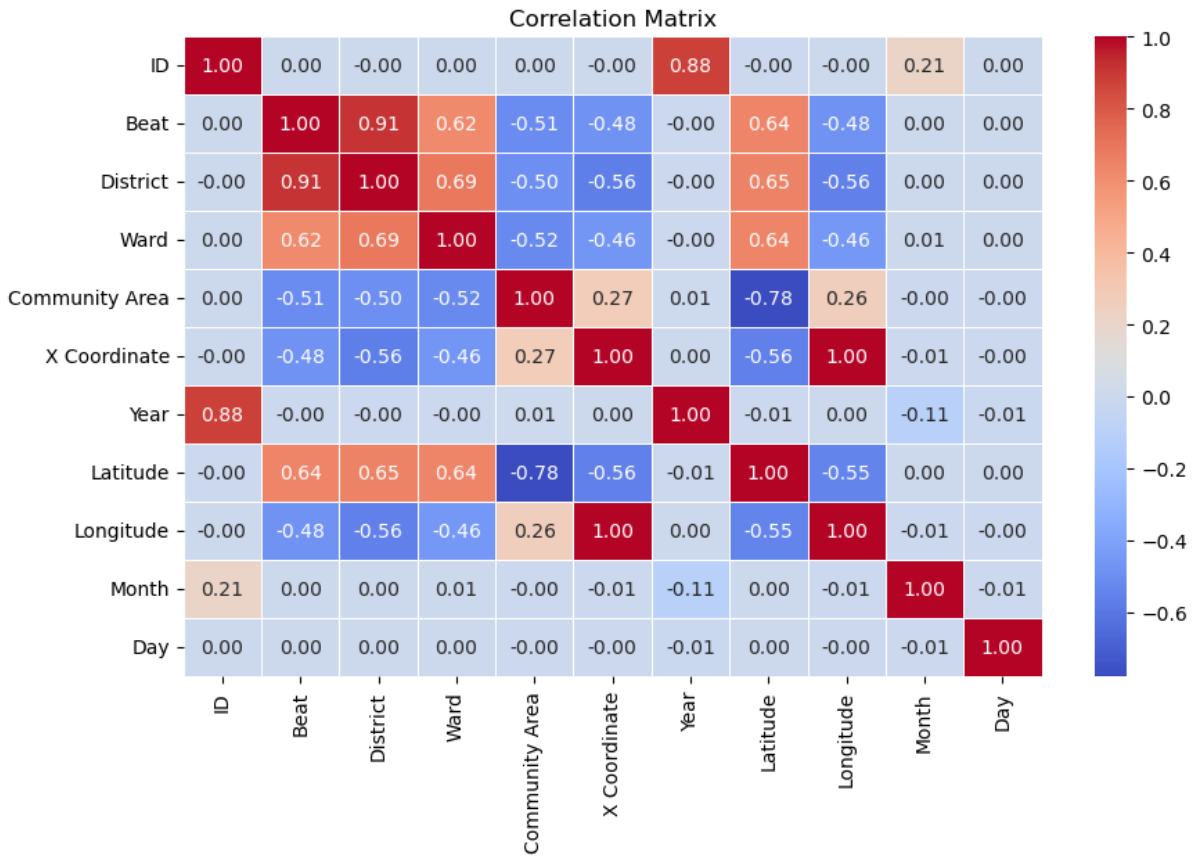
Geospatial Analysis

In [33]:
```
# Scatter plot of crimes based on latitude and longitude
plt.figure(figsize=(10, 6))
plt.scatter(crime_data_2001_2004['Longitude'], crime_data_2001_2004['L
plt.title('Geographic Distribution of Crimes')
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.show()
```



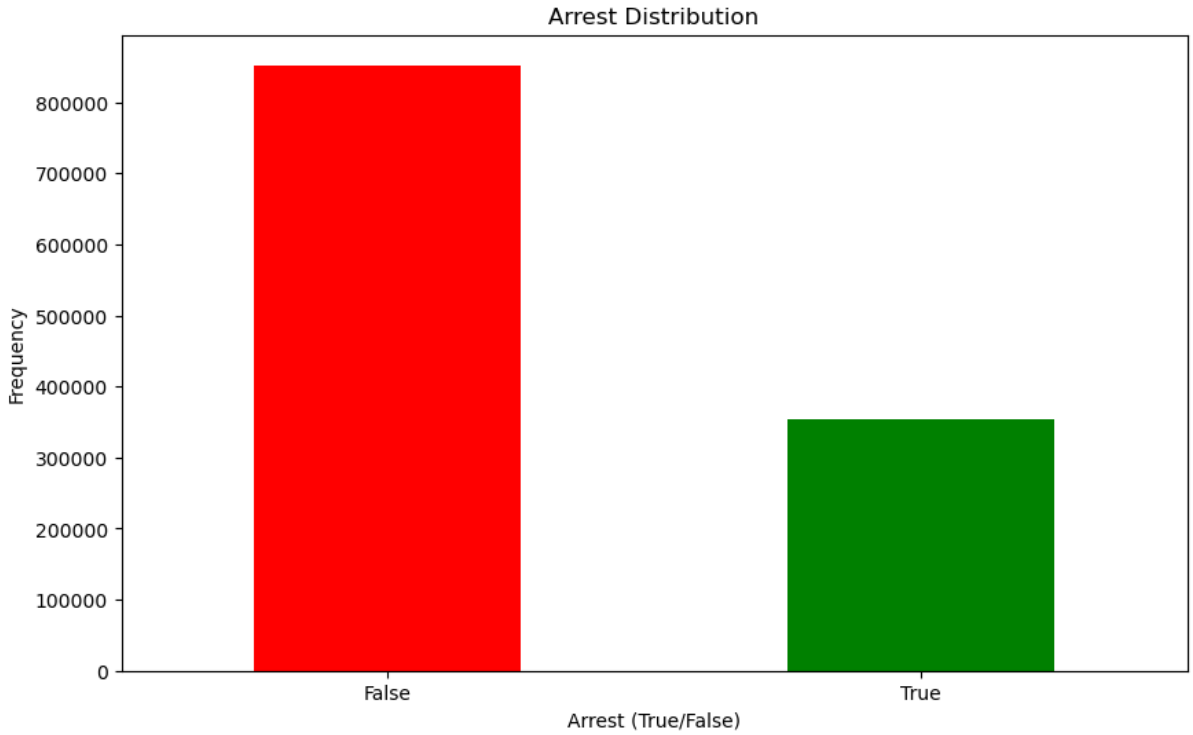
Correlation Analysis

In [35]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Filter only numeric columns
numeric_data = crime_data_2001_2004.select_dtypes(include=['number'])

# Correlation matrix
corr_matrix = numeric_data.corr()

# Plotting the heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linew
plt.title('Correlation Matrix')
plt.show()
```



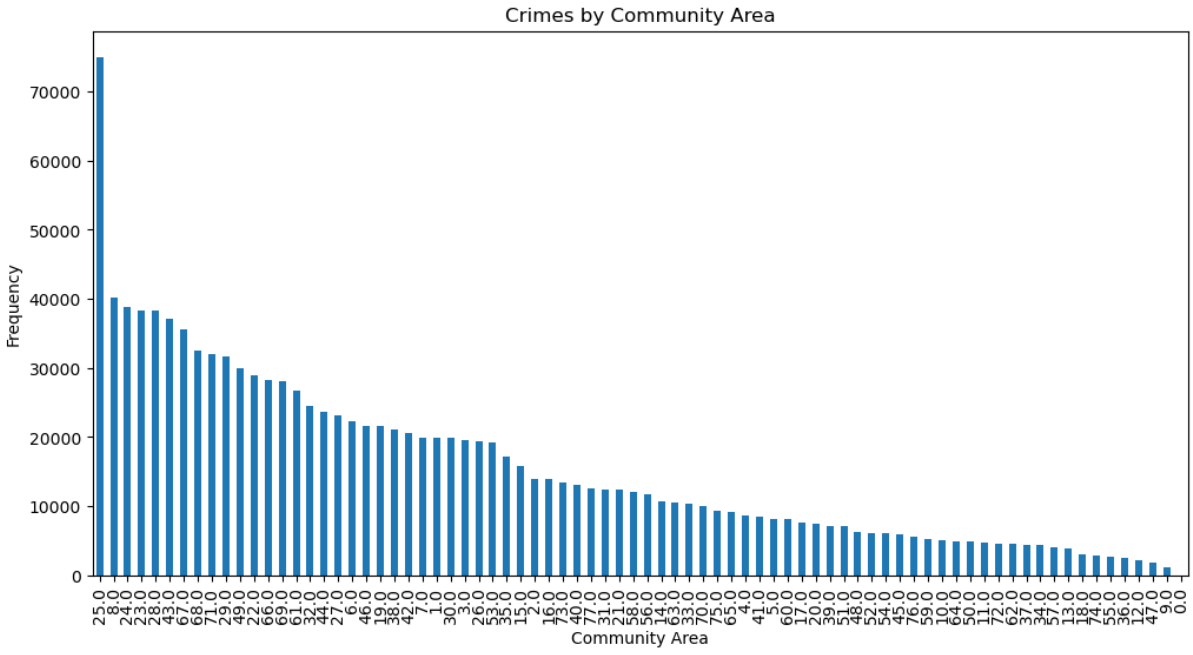
Distribution of Arrest Data

In [36]:
```
# Plot distribution of arrests
plt.figure(figsize=(10, 6))
crime_data_2001_2004['Arrest'].value_counts().plot(kind='bar', color=[
plt.title('Arrest Distribution')
plt.xlabel('Arrest (True/False)')
plt.ylabel('Frequency')
plt.xticks(rotation=0)
plt.show()
```



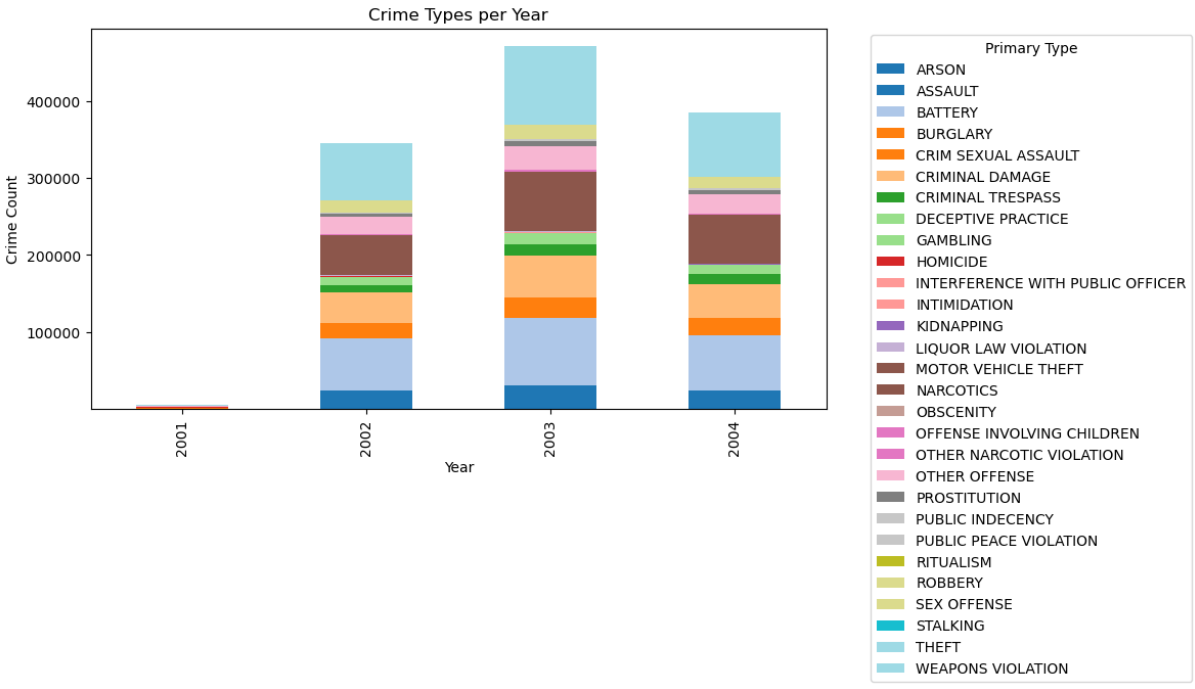
Community Area Analysis

In [37]:
```
# Crimes by Community Area
plt.figure(figsize=(12, 6))
crime_data_2001_2004['Community Area'].value_counts().plot(kind='bar')
plt.title('Crimes by Community Area')
plt.xlabel('Community Area')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()
```



Bivariate Analysis (Year vs. Crime Type)

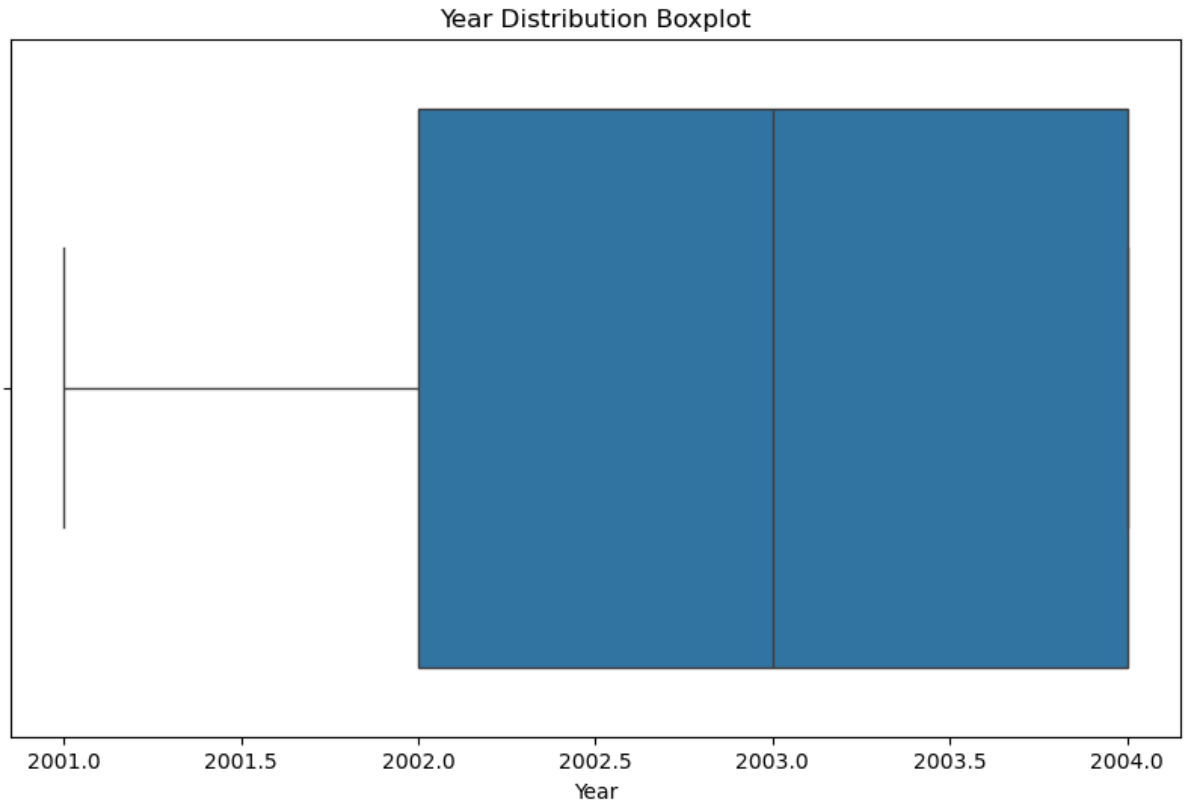
In [40]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

# Assuming crime_data_2001_2004 is your DataFrame
# Create a pivot table for stacked bar chart
pivot_data = crime_data_2001_2004.groupby(['Year', 'Primary Type']).si

# Plot the stacked bar chart
pivot_data.plot(kind='bar', stacked=True, figsize=(12, 6), colormap='t

plt.title('Crime Types per Year')
plt.xlabel('Year')
plt.ylabel('Crime Count')
plt.xticks(rotation=90)

# Adjust legend to be outside of the plot
plt.legend(title='Primary Type', bbox_to_anchor=(1.05, 1), loc='upper

plt.tight_layout()  # Adjust layout to avoid clipping of legend
plt.show()
```

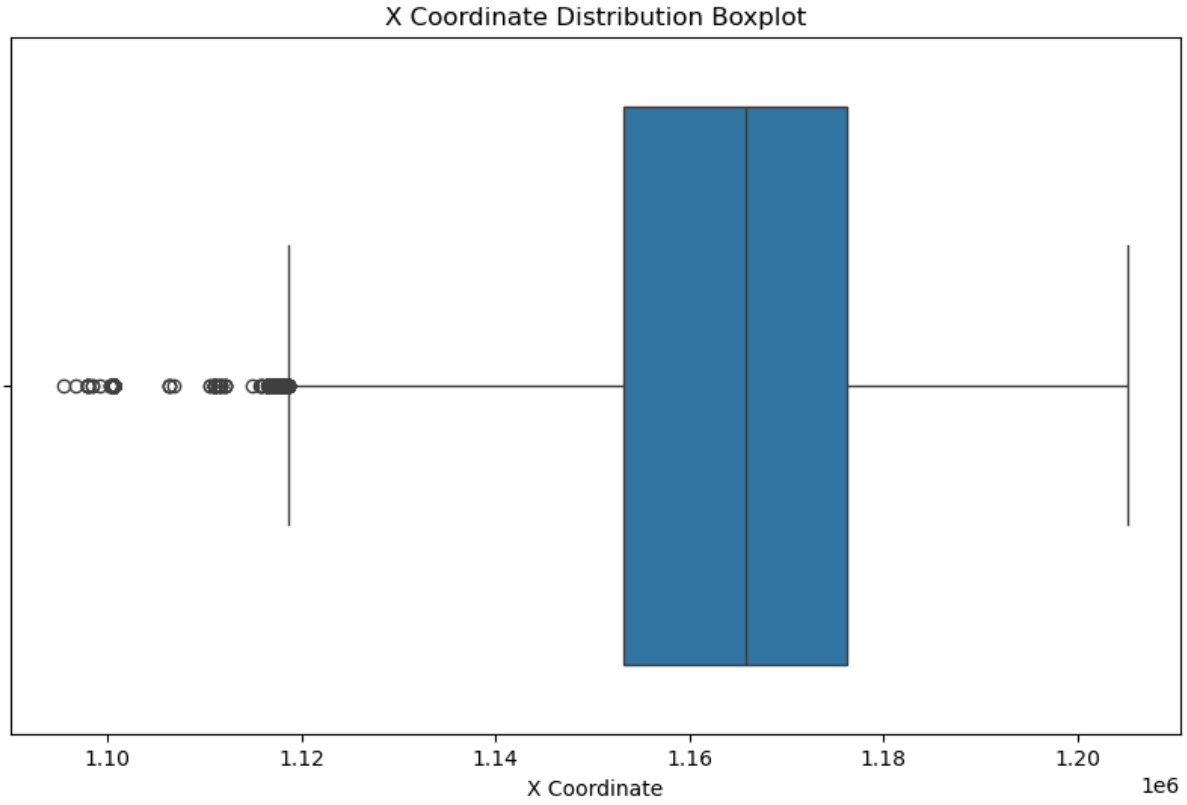


Outlier Detection

In [41]:
```
# Boxplot for year
plt.figure(figsize=(10, 6))
sns.boxplot(x=crime_data_2001_2004['Year'])
plt.title('Year Distribution Boxplot')
plt.show()
```

```
# Boxplot for coordinates
plt.figure(figsize=(10, 6))
sns.boxplot(x=crime_data_2001_2004['X Coordinate'])
plt.title('X Coordinate Distribution Boxplot')
plt.show()
```





Machine Learning

Data Preprocessing

In [44]:
```
# Preprocess data (check for missing values, handle them, etc.)
combined_data.dropna(inplace=True)  # Drop rows with missing values (i

# Convert the 'Date' column to datetime format
combined_data['Date'] = pd.to_datetime(combined_data['Date'], errors='

# Extract useful features from 'Date' (e.g., year, month, day of week)
combined_data['Year'] = combined_data['Date'].dt.year
combined_data['Month'] = combined_data['Date'].dt.month
combined_data['DayOfWeek'] = combined_data['Date'].dt.dayofweek

# Encode categorical features (e.g., 'Primary Type' and 'Location Desc
combined_data = pd.get_dummies(combined_data, columns=['Primary Type',
```

```
/var/folders/83/sl_nw5nd0pv4njhhbnqkzwvr0000gn/T/ipykernel_2000/1900518
97.py:5: UserWarning: Could not infer format, so each element will be p
arsed individually, falling back to `dateutil`. To ensure parsing is co
nsistent and as-expected, please specify a format.
  combined_data['Date'] = pd.to_datetime(combined_data['Date'], errors
='coerce')
```

Feature Engineering

In [45]:
```
# Define features and target variable
X = combined_data[['Year', 'Month', 'DayOfWeek'] + [col for col in com
y = combined_data['Arrest']  # Target variable (whether an arrest was
```

Model Training

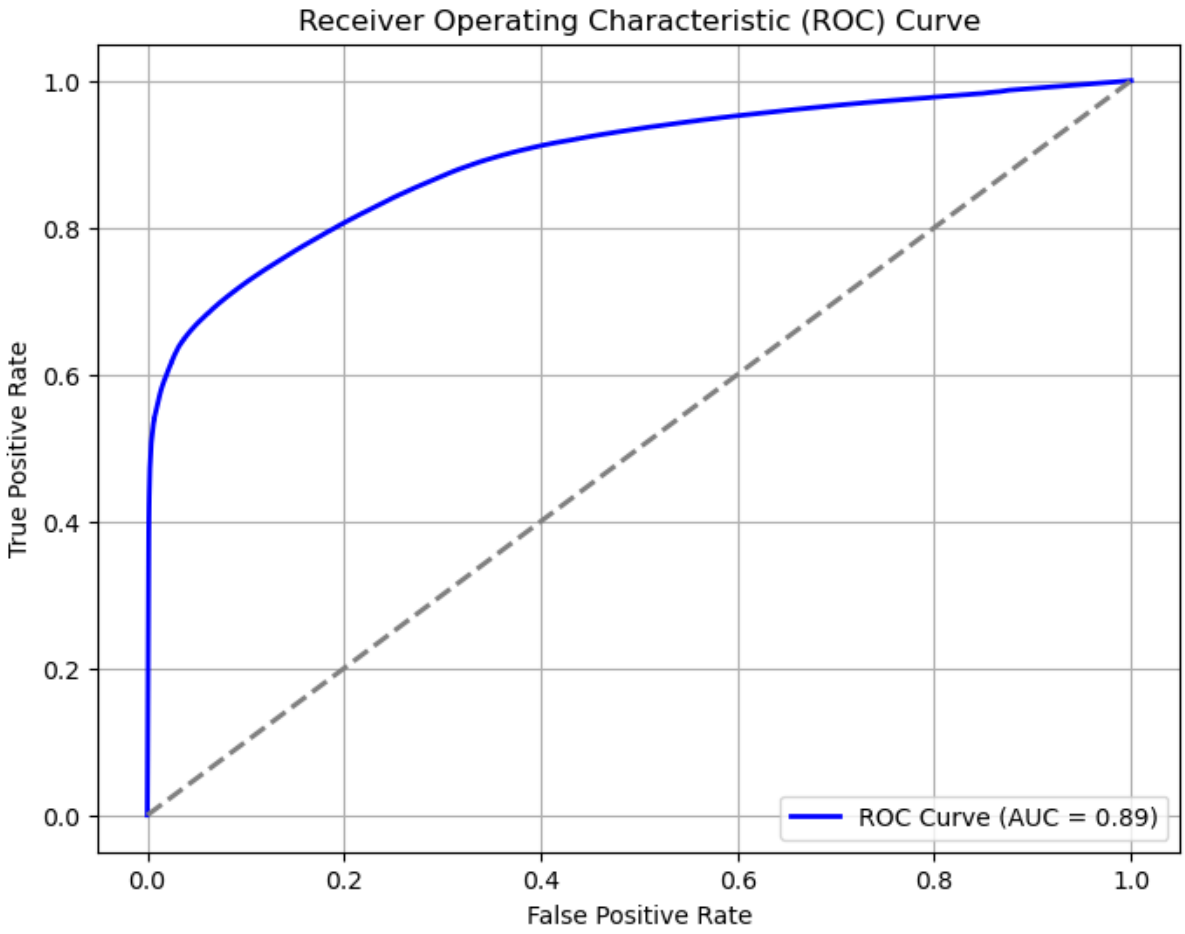
In [46]:
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.

# Train a Random Forest model
model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)

# Predict on the test set
y_pred = model.predict(X_test)

# Evaluate the model's performance
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy: {accuracy}')
```

```
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8743396804714764
              precision    recall  f1-score   support

       False       0.87      0.96      0.92   1025939
        True       0.88      0.65      0.74    403104

    accuracy                           0.87   1429043
   macro avg       0.88      0.80      0.83   1429043
weighted avg       0.87      0.87      0.87   1429043
```

In [49]:

```
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, roc_auc_score

# Get the probabilities for the positive class (True)
y_probs = model.predict_proba(X_test)[:, 1]

# Compute the ROC curve and AUC score
fpr, tpr, thresholds = roc_curve(y_test, y_probs)
roc_auc = roc_auc_score(y_test, y_probs)

# Plot the ROC curve
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, label=f'ROC Curve (AUC = {roc_auc:.2f})', color='bl
plt.plot([0, 1], [0, 1], color='gray', linestyle='--', lw=2)  # Diagon
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc='lower right')
plt.grid()
plt.show()
```


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.89)