In [1]:
```
# Imports
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# reading and printing all the columns
df = pd.read_csv('F_PoliceIncident.csv')
pd.set_option('display.max_columns', None)
#pd.set_option('display.max_rows', None)
df.info()
```

```
C:\Users\jules\anaconda3\lib\site-packages\IPython\core\interactiveshell.py:344
4: DtypeWarning: Columns (1) have mixed types.Specify dtype option on import or
set low_memory=False.
  exec(code_obj, self.user_global_ns, self.user_ns)

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 848467 entries, 0 to 848466
Data columns (total 20 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   Year of Incident         848467 non-null  int64
 1   Watch                    848467 non-null  object
 2   Call (911) Problem       809734 non-null  object
 3   Type of Incident         848467 non-null  object
 4   Type of Location         847324 non-null  object
 5   Reporting Area           847682 non-null  float64
 6   Beat                     848333 non-null  float64
 7   Division                 848333 non-null  object
 8   Sector                   848282 non-null  float64
 9   Council District         846785 non-null  object
 10  Year of Occurrence       848467 non-null  int64
 11  Month of Occurence       848467 non-null  object
 12  Day of the Week          848467 non-null  object
 13  Time of Occurrence       848467 non-null  object
 14  Person Involvement Type  817550 non-null  object
 15  Victim Type              809575 non-null  object
 16  Hate Crime Description   847270 non-null  object
 17  Drug Related Incident    810528 non-null  object
 18  Penal Code               848467 non-null  object
 19  Zip Code                 845036 non-null  float64
dtypes: float64(4), int64(2), object(14)
memory usage: 129.5+ MB
```

In [2]:
```
sumOfNa = df.isna().sum()
sumOfNa
```

Out[2]:
```
Year of Incident              0
Watch                         0
Call (911) Problem        38733
Type of Incident              0
Type of Location           1143
Reporting Area              785
Beat                        134
Division                    134
Sector                      185
Council District           1682
Year of Occurrence            0
Month of Occurence            0
Day of the Week               0
Time of Occurrence            0
Person Involvement Type    30917
Victim Type                38892
Hate Crime Description      1197
Drug Related Incident      37939
Penal Code                    0
Zip Code                   3431
dtype: int64
```

test 2 missing values

In [18]:
```
df = pd.read_csv('F_PoliceIncident_DealMissing.csv')
```

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 848467 entries, 0 to 848466
Data columns (total 19 columns):
 #   Column                   Non-Null Count   Dtype  
---  ------                   --------------   -----  
 0   Year of Incident         848467 non-null  int64  
 1   Watch                    848467 non-null  int64  
 2   Type of Incident         848467 non-null  object 
 3   Type of Location         847324 non-null  object 
 4   Reporting Area           847682 non-null  float64
 5   Beat                     848333 non-null  float64
 6   Division                 848333 non-null  object 
 7   Sector                   848333 non-null  float64
 8   Council District         846785 non-null  object 
 9   Year of Occurrence       848467 non-null  int64  
 10  Month of Occurence       848467 non-null  object 
 11  Day of the Week          848467 non-null  object 
 12  Time of Occurrence       848467 non-null  object 
 13  Person Involvement Type  817550 non-null  object 
 14  Victim Type              809575 non-null  object 
 15  Hate Crime Description   847270 non-null  object 
 16  Drug Related Incident    810528 non-null  object 
 17  Penal Code               848467 non-null  object 
 18  Zip Code                 848467 non-null  int64  
dtypes: float64(3), int64(4), object(12)
memory usage: 123.0+ MB
```

In [19]:
```
sumOfNa = df.isna().sum()
sumOfNa
```

Out[19]:
```
Year of Incident              0
Watch                         0
Type of Incident              0
Type of Location           1143
Reporting Area              785
Beat                        134
Division                    134
Sector                      134
Council District           1682
Year of Occurrence            0
Month of Occurence            0
Day of the Week               0
Time of Occurrence            0
Person Involvement Type    30917
Victim Type                38892
Hate Crime Description      1197
Drug Related Incident      37939
Penal Code                    0
Zip Code                      0
dtype: int64
```

In [20]:
```
index = df[df['Beat'].isna()].index
```

In [21]:
```
index
```

Out[21]:
```
Int64Index([ 30205,  36586,  41326,  41658,  42927,  43471,  51137,  51596,
             51951,  54578,
            ...
            799838, 807625, 807731, 814790, 815872, 816320, 818774, 839325,
            839413, 846996],
           dtype='int64', length=134)
```

In [22]:
```
df.drop(index, inplace=True)
```

In [23]:
```
df.shape
```

Out[23]:
```
(848333, 19)
```

In [24]:
```
sumOfNa = df.isna().sum()
sumOfNa
```

Out[24]:
```
Year of Incident              0
Watch                         0
Type of Incident              0
Type of Location           1143
Reporting Area              651
Beat                          0
Division                      0
Sector                        0
Council District           1548
Year of Occurrence            0
Month of Occurence            0
Day of the Week               0
Time of Occurrence            0
Person Involvement Type    30917
Victim Type                38891
Hate Crime Description      1197
Drug Related Incident      37939
Penal Code                    0
Zip Code                      0
dtype: int64
```

In [25]:
```
df.mode()
```

Out[25]:

	Year of Incident	Watch	Type of Incident	Type of Location	Reporting Area	Beat	Division	Sector	Council District	Year o Occurrence
0	2020	1	BMV	Highway, Street, Alley ETC	1217.0	521.0	NORTHEAST	530.0	D2	202(

In [45]:
```
df.to_csv('Final_PoliceIncident.csv', encoding='utf-8',sep=',', index=False)
```

In [ ]:

Final Perge

In [29]:
```
df = pd.read_csv('F_PoliceIncident_DealMissing.csv')
```

In [30]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 848333 entries, 0 to 848332
Data columns (total 19 columns):
 #   Column                   Non-Null Count   Dtype 
---  ------                   --------------   ----- 
 0   Year of Incident         848333 non-null  int64 
 1   Watch                    848333 non-null  int64 
 2   Type of Incident         848333 non-null  object
 3   Type of Location         848333 non-null  object
 4   Reporting Area           848333 non-null  int64 
 5   Beat                     848333 non-null  int64 
 6   Division                 848333 non-null  object
 7   Sector                   848333 non-null  int64 
 8   Council District         846785 non-null  object
 9   Year of Occurrence       848333 non-null  int64 
 10  Month of Occurence       848333 non-null  object
 11  Day of the Week          848333 non-null  object
 12  Time of Occurrence       848333 non-null  object
 13  Person Involvement Type  817416 non-null  object
 14  Victim Type              809442 non-null  object
 15  Hate Crime Description   847136 non-null  object
 16  Drug Related Incident    810394 non-null  object
 17  Penal Code               848333 non-null  object
 18  Zip Code                 848333 non-null  int64 
dtypes: int64(7), object(12)
memory usage: 123.0+ MB
```

In [31]:
```
sumOfNa = df.isna().sum()
sumOfNa
```

Out[31]:
```
Year of Incident              0
Watch                         0
Type of Incident              0
Type of Location              0
Reporting Area                0
Beat                          0
Division                      0
Sector                        0
Council District           1548
Year of Occurrence            0
Month of Occurence            0
Day of the Week               0
Time of Occurrence            0
Person Involvement Type    30917
Victim Type                38891
Hate Crime Description      1197
Drug Related Incident      37939
Penal Code                    0
Zip Code                      0
dtype: int64
```

In [33]:
```
index = df[df['Council District'].isna()].index
```

In [34]:
```
df.drop(index, inplace=True)
```

In [44]:
```
sumOfNa = df.isna().sum()
sumOfNa
```

Out[44]:
```
Year of Incident              0
Watch                         0
Type of Incident              0
Type of Location              0
Reporting Area                0
Beat                          0
Division                      0
Sector                        0
Council District              0
Year of Occurrence            0
Month of Occurence            0
Day of the Week               0
Time of Occurrence            0
Person Involvement Type       3
Victim Type                   0
Hate Crime Description        0
Drug Related Incident      36818
Penal Code                    0
Zip Code                      0
dtype: int64
```

In [41]:
```
index = df[df['Hate Crime Description'].isna()].index
```

In [42]:
```
index
```

Out[42]:
```
Int64Index([   879,    969,    993,   1365,   1568,   2519,  26380,  27724,
             27810,  28898,
            ...
            841152, 841265, 842696, 843283, 843996, 844041, 844182, 844444,
            845362, 847786],
           dtype='int64', length=1071)
```

In [43]:
```
df.drop(index, inplace=True)
```