Oakland Crime Statistics 2011 to 2016 数据集分析

该数据集包含从2011年到2016年的数据，在2012年和2014年的csv文件中，比其他csv文件多出'zip code',在具体分析，我们对'zip code'不做考虑，数据中的属性如下

- Agency： 机构
- Create Time： 立案时间
- Location： 案件位置
- Area Id： 区域ID
- Beat： 巡逻区域
- Priority： 案件等级
- Incident Type Id： 事件类型Id
- Incident Type Description： 事件类型描述
- Event Number： 事件编号
- Closed Time： 结案时间

In [35]:

```
import os
import sys
import math
import pandas as pd
import numpy as np
import csv
import json
import pickle
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
import time
%matplotlib inline
```

In [21]:

```
data1 = pd.read_csv('.\data\Oakland\\records-for-2011.csv')
data2 = pd.read_csv('.\data\Oakland\\records-for-2012.csv')
data3 = pd.read_csv('.\data\Oakland\\records-for-2013.csv')
data4 = pd.read_csv('.\data\Oakland\\records-for-2014.csv')
data5 = pd.read_csv('.\data\Oakland\\records-for-2015.csv')
```

```
data6 = pd.read_csv('.\data\Oakland\\records-for-2016.csv')
data1.head()
```

Out[21]:

	Agency	Create Time	Location	Area Id	Beat	Priority	Incident Type Id	Incident Type Description	Event Number	Closed Time
0	OP	2011-01-01T00:00:00.000	ST&SAN PABLO AV	1.0	06X	1.0	PDOA	POSSIBLE DEAD PERSON	LOP110101000001	2011-01-01T00:28:17.000
1	OP	2011-01-01T00:01:11.000	ST&HANNAH ST	1.0	07X	1.0	415GS	415 GUNSHOTS	LOP110101000002	2011-01-01T01:12:56.000
2	OP	2011-01-01T00:01:25.000	ST&MARKET ST	1.0	10Y	2.0	415GS	415 GUNSHOTS	LOP110101000003	2011-01-01T00:07:20.000
3	OP	2011-01-01T00:01:35.000	PRENTISS ST	2.0	21Y	2.0	415GS	415 GUNSHOTS	LOP110101000005	2011-01-01T00:02:28.000
4	OP	2011-01-01T00:02:10.000	AV&FOOTHILL BLVD	2.0	20X	1.0	415GS	415 GUNSHOTS	LOP110101000004	2011-01-01T00:50:04.000

In [23]:

```
data1.shape
```

Out[23]:

```
(180016, 10)
```

In [24]:

```
data2.shape
```

Out[24]:

```
(187431, 11)
```

In [25]:

```
data3.shape
```

Out[25]:

```
(188052, 10)
```

In [27]:

```
data4.shape
```

Out[27]:

```
(187480, 11)
```

In [28]:

```
data5.shape
```

Out[28]:

```
(192581, 10)
```

```
In  [29]:  data6.shape

Out[29]:  (110828, 10)
```

```
In  [11…  data_all=[data1,data2,data3,data4,data5,data6]
```

```
In  [12…  cols1 = list(data1)
          cols2 = list(data2)
          cols3 = list(data3)
          cols4 = list(data4)
          cols5 = list(data5)
          cols6 = list(data6)
          cols_all = [cols1,cols2,cols3,cols4,cols5,cols6]
          print(cols1)
          print(cols2)
          print(cols3)
          print(cols4)
          print(cols5)
          print(cols6)
```

```
['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Number',
'Closed Time']
['Agency', 'Create Time', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Number', 'Closed Tim
e', 'Location 1', 'Zip Codes']
['Agency', 'Create Time', 'Location ', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Numbe
r', 'Closed Time']
['Agency', 'Create Time', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Number', 'Closed Tim
e', 'Location 1', 'Zip Codes']
['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Number',
'Closed Time']
['Agency', 'Create Time', 'Location', 'Area Id', 'Beat', 'Priority', 'Incident Type Id', 'Incident Type Description', 'Event Number',
'Closed Time']
```

数据摘要

对标称数据计算频数

根据每个属性的特点可知，标称属性包括

- Location
- Area Id

- Beat
- Incident Type Id
- Incident Type Description

In [21…

```
for data,cols in zip(data_all,cols_all):
    for col in cols:
        print(data[col].value_counts())
        print('-' * 60)
    print("=" * 60)
```

```
OP    180015
Name: Agency, dtype: int64
------------------------------------------------------------
2011-06-02T00:00:00.000    4
2011-03-27T00:22:41.000    3
2011-09-21T14:05:59.000    3
2011-05-01T18:31:50.000    2
2011-05-12T21:04:34.000    2
                          ..
2011-10-15T10:38:11.000    1
2011-02-02T21:48:32.000    1
2011-02-09T13:58:47.000    1
2011-05-10T09:53:55.000    1
2011-05-02T14:44:02.000    1
Name: Create Time, Length: 179451, dtype: int64
------------------------------------------------------------
 INTERNATIONAL BLVD        3866
 MACARTHUR BLVD            3129
 AV&INTERNATIONAL BLVD     3067
 BROADWAY                  2132
 FOOTHILL BLVD             1791
                           ...
FRUITVALE DAVIS ST            1
43RD STANLEY AV               1
70TH W MACARTHUR BLVD         1
34TH EMBARCADERO WEST         1
28TH CT&COLLEGE AV            1
Name: Location, Length: 32505, dtype: int64
------------------------------------------------------------
1.0    79152
2.0    67261
3.0    32699
Name: Area Id, dtype: int64
------------------------------------------------------------
```

```
04X       7410
08X       6885
26Y       5478
30Y       5295
06X       5119
23X       5051
30X       4956
19X       4955
34X       4673
29X       4483
20X       4287
27Y       4159
07X       4134
31Y       4082
25X       4022
35X       3880
33X       3849
03X       3819
32X       3711
27X       3703
09X       3630
21Y       3435
32Y       3125
22X       3061
26X       2978
02Y       2970
10X       2967
14X       2733
03Y       2726
22Y       2664
12Y       2651
05X       2633
02X       2614
31X       2603
21X       2593
17Y       2582
24Y       2575
13Z       2546
15X       2509
24X       2459
12X       2422
10Y       2383
01X       2210
28X       2191
17X       2133
11X       2087
13Y       2017
```

```
35Y     1956
31Z     1870
18Y     1778
16Y     1561
14Y     1492
25Y     1482
13X     1122
18X     1063
16X      994
05Y      710
PDT2      20
Name: Beat, dtype: int64
------------------------------------------------------------
2.0    143314
1.0     36699
0.0         2
Name: Priority, dtype: int64
------------------------------------------------------------
933R     17348
911H     12817
SECCK    11393
415      10752
10851     7180
         ...
243C         1
970A         1
666          1
243B         1
148_1        1
Name: Incident Type Id, Length: 263, dtype: int64
------------------------------------------------------------
ALARM-RINGER            17348
911 HANG-UP             12817
SECURITY CHECK          11393
STOLEN VEHICLE           7180
415 UNKNOWN              6624
                        ...
CONSPIRACY COURT ORD        1
ASSAULT ON A POLICE         1
EXTORTION                   1
INJURE TELEPHONE/POW        1
OBSTRUCTING JUSTICE-        1
Name: Incident Type Description, Length: 265, dtype: int64
------------------------------------------------------------
LOP111216000789    1
LOP111124000115    1
LOP110305000062    1
```

```
LOP110706000913    1
LOP110922000912    1
                  ..
LOP111126000609    1
LOP110111000539    1
LOP110110000782    1
LOP110710000025    1
LOP110829000767    1
Name: Event Number, Length: 180015, dtype: int64
------------------------------------------------------------
2011-03-04T21:56:33.000    2
2011-12-19T15:28:03.000    2
2011-11-12T00:41:40.000    2
2011-03-24T20:36:06.000    2
2011-11-12T13:48:27.000    2
                          ..
2011-06-11T13:13:19.000    1
2011-11-30T04:04:57.000    1
2011-05-06T17:01:16.000    1
2011-10-22T02:03:28.000    1
2011-09-07T19:55:54.000    1
Name: Closed Time, Length: 179506, dtype: int64
------------------------------------------------------------
============================================================
OP    187430
Name: Agency, dtype: int64
------------------------------------------------------------
2012-06-26T00:00:00.000    8
2012-05-07T00:00:00.000    7
2012-12-02T00:00:00.000    3
2012-04-02T00:00:00.000    3
2012-06-30T00:00:00.000    3
                          ..
2012-03-19T18:09:56.000    1
2012-02-13T14:20:55.000    1
2012-01-04T14:55:02.000    1
2012-07-28T21:20:01.000    1
2012-01-03T16:11:05.000    1
Name: Create Time, Length: 186801, dtype: int64
------------------------------------------------------------
1.0    101053
2.0     84963
Name: Area Id, dtype: int64
------------------------------------------------------------
04X     8088
08X     6691
30Y     5529
```

```
26Y      5374
23X      5301
19X      5158
30X      4988
34X      4965
20X      4682
06X      4676
29X      4606
25X      4396
03X      4380
35X      4291
07X      4235
31Y      3975
09X      3845
32X      3836
21Y      3822
27Y      3701
33X      3697
27X      3685
12Y      3344
32Y      3328
22X      3131
14X      3070
02Y      3043
03Y      3009
26X      2982
10X      2961
13Z      2946
02X      2798
10Y      2727
22Y      2725
24Y      2723
05X      2681
21X      2674
15X      2671
17Y      2635
12X      2491
24X      2483
31X      2482
28X      2321
01X      2193
11X      2165
17X      2127
35Y      1986
13Y      1898
31Z      1849
18Y      1816
```

```
16Y     1680
14Y     1578
25Y     1512
18X     1224
13X     1212
16X     1197
05Y      836
PDT2      28
Name: Beat, dtype: int64
------------------------------------------------------------
2.0    145504
1.0     41926
Name: Priority, dtype: int64
------------------------------------------------------------
933R      17216
SECCK     11488
415       11158
911H      10585
10851      8208
          ...  
285           1
VINVER        1
107           1
243A          1
12020         1
Name: Incident Type Id, Length: 256, dtype: int64
------------------------------------------------------------
ALARM-RINGER            17216
SECURITY CHECK          11488
911 HANG-UP             10585
STOLEN VEHICLE           8208
415 UNKNOWN              6081
                        ...  
ASSAULT ON A POLICE         1
POSSESSION/MANUFACTU        1
ESCAPE DETENTION            1
INCEST                      1
INJURE TELEPHONE/POW        1
Name: Incident Type Description, Length: 258, dtype: int64
------------------------------------------------------------
LOP120324000786    1
LOP120217000984    1
LOP120210001126    1
LOP120729000808    1
LOP120422000303    1
                  ..
LOP120516000031    1
```

```
LOP120427000991    1
LOP120801000774    1
LOP120710000811    1
LOP120709000505    1
Name: Event Number, Length: 187430, dtype: int64
------------------------------------------------------------
2012-05-08T11:29:58.000    2
2012-10-02T20:25:22.000    2
2012-01-03T14:04:54.000    2
2012-11-07T17:27:58.000    2
2012-04-03T01:39:07.000    2
                          ..
2012-12-19T12:52:09.000    1
2012-03-09T19:35:40.000    1
2012-06-11T18:40:12.000    1
2012-05-18T01:17:14.000    1
2012-02-02T22:55:38.000    1
Name: Closed Time, Length: 186874, dtype: int64
------------------------------------------------------------
{'human_address': '{"address": "INTERNATIONAL BLVD", "city": "", "state": "", "zip": ""}'}          3658
{'human_address': '{"address": "MACARTHUR BLVD", "city": "", "state": "", "zip": ""}'}              3335
{'human_address': '{"address": "AV&INTERNATIONAL BLVD", "city": "", "state": "", "zip": ""}'}       3193
{'human_address': '{"address": "BROADWAY", "city": "", "state": "", "zip": ""}'}                    2167
{'human_address': '{"address": "FOOTHILL BLVD", "city": "", "state": "", "zip": ""}'}               1649
                                                                                                    ...
{'human_address': '{"address": "10TH BROOKLYN AV", "city": "", "state": "", "zip": ""}'}               1
{'human_address': '{"address": "OAKLAND GRAND AV&WEST ST", "city": "", "state": "", "zip": ""}'}       1
{'human_address': '{"address": "88TH SHAFTER AV", "city": "", "state": "", "zip": ""}'}                1
{'human_address': '{"address": "98TH APRICOT ST", "city": "", "state": "", "zip": ""}'}                1
{'human_address': '{"address": "ASILOMAR BIRDSALL AV", "city": "", "state": "", "zip": ""}'}           1
Name: Location 1, Length: 35312, dtype: int64
------------------------------------------------------------
4560.0     5
1481.0     3
11164.0    3
4380.0     3
4366.0     2
          ..
14892.0    1
170.0      1
15010.0    1
5463.0     1
2050.0     1
Name: Zip Codes, Length: 150, dtype: int64
------------------------------------------------------------
============================================================
OP    188051
```

```
Name: Agency, dtype: int64
------------------------------------------------------------
2013-01-29T09:16:31.000    18
2013-05-26T00:00:00.000     3
2013-09-20T00:00:00.000     3
2013-07-06T00:00:00.000     3
2013-05-12T00:00:00.000     3
                           ..
2013-03-05T12:00:26.000     1
2013-03-08T14:00:29.000     1
2013-06-20T15:26:12.000     1
2013-11-27T14:50:30.000     1
2013-10-25T13:40:29.000     1
Name: Create Time, Length: 187433, dtype: int64
------------------------------------------------------------
 INTERNATIONAL BLVD           3647
 AV&INTERNATIONAL BLVD        3405
 MACARTHUR BLVD               3002
 BROADWAY                     2036
 FOOTHILL BLVD                1650
                              ...
59TH 55TH AV                     1
BROMLEY ST&PERALTA ST            1
CHAMPION THORNHILL DR            1
18TH AV&SCOVILLE ST              1
HAMPEL AV&KAPHAN AV              1
Name: Location , Length: 36804, dtype: int64
------------------------------------------------------------
1.0    105216
2.0     80578
Name: Area Id, dtype: int64
------------------------------------------------------------
04X     7697
08X     6993
30X     5440
30Y     5439
23X     5279
19X     5211
26Y     5188
34X     5059
06X     4786
20X     4565
29X     4531
25X     4530
03X     4483
07X     4416
31Y     4304
```

```
32X     4194
35X     4053
27Y     4026
21Y     3938
09X     3776
27X     3774
33X     3537
02Y     3522
12Y     3465
32Y     3465
22X     3095
03Y     2899
05X     2896
14X     2881
26X     2787
02X     2713
24X     2710
10X     2702
10Y     2641
22Y     2614
12X     2576
24Y     2571
17Y     2564
15X     2482
13Z     2383
31X     2361
01X     2309
28X     2294
21X     2289
17X     2091
31Z     2047
11X     1964
35Y     1950
13Y     1826
18Y     1817
14Y     1794
16Y     1720
25Y     1537
18X     1387
16X     1255
13X     1209
05Y      821
PDT2      18
Name: Beat, dtype: int64
------------------------------------------------------------
2.0    144859
1.0     43171
```

```
0.0        21
Name: Priority, dtype: int64
------------------------------------------------------------
933R     17859
SECCK    12240
415      11313
10851     9469
911H      8268
         ...  
290          1
209          1
372          1
626_9        1
243B         1
Name: Incident Type Id, Length: 253, dtype: int64
------------------------------------------------------------
ALARM-RINGER            17859
SECURITY CHECK          12240
STOLEN VEHICLE           9469
911 HANG-UP              8268
DISTURBING THE PEACE     6553
                        ...  
KIDNAPPING FOR RANSO        1
IDENTITY THEFT              1
ASSSAULT                    1
POSSESS WEAPON AT SC        1
INCEST                      1
Name: Incident Type Description, Length: 254, dtype: int64
------------------------------------------------------------
LOP131010000795    1
LOP130414000302    1
LOP130808000757    1
LOP131108000773    1
LOP130819000524    1
                  ..
LOP131024000642    1
LOP130507000923    1
LOP130925001174    1
LOP130314000757    1
LOP130922000373    1
Name: Event Number, Length: 188051, dtype: int64
------------------------------------------------------------
2013-02-12T22:52:01.000    4
2013-09-01T17:23:50.000    4
2013-04-26T21:30:39.000    3
2013-12-23T18:18:23.000    3
2013-02-16T15:58:55.000    2
```

```
                          ..
2013-08-17T05:43:27.000    1
2013-06-19T13:04:59.000    1
2013-12-05T21:10:41.000    1
2013-07-06T17:22:44.000    1
2013-12-10T15:53:39.000    1
Name: Closed Time, Length: 187487, dtype: int64
------------------------------------------------------------
============================================================
OP    187480
Name: Agency, dtype: int64
------------------------------------------------------------
2014-10-14T02:45:12.000    14
2014-10-14T02:46:45.000    11
2014-01-01T00:00:00.000     4
2014-09-20T00:00:00.000     4
2014-11-04T14:39:16.000     3
                           ..
2014-05-25T09:46:32.000     1
2014-01-17T11:04:07.000     1
2014-09-02T20:36:48.000     1
2014-06-09T21:03:13.000     1
2014-10-28T18:40:52.000     1
Name: Create Time, Length: 186851, dtype: int64
------------------------------------------------------------
1.0    5031
2.0    3898
5.0     320
4.0     236
3.0     208
Name: Area Id, dtype: int64
------------------------------------------------------------
04X     7868
08X     6723
30X     5539
23X     5485
30Y     5454
26Y     5377
19X     5290
06X     4931
34X     4865
03X     4727
27Y     4653
29X     4645
20X     4639
07X     4617
31Y     4541
```

```
25X     4372
35X     4240
27X     3912
32X     3833
21Y     3784
09X     3625
32Y     3622
02Y     3621
33X     3561
12Y     3214
03Y     3212
14X     2870
26X     2843
24X     2843
02X     2819
22X     2789
24Y     2673
10X     2566
10Y     2537
12X     2516
21X     2502
31X     2486
17Y     2480
05X     2442
13Z     2415
15X     2347
01X     2320
22Y     2297
28X     2186
11X     2092
31Z     2022
35Y     1860
17X     1860
14Y     1772
13Y     1720
18Y     1609
16Y     1495
25Y     1319
13X     1211
18X     1142
16X     1035
05Y      821
PDT2      24
Name: Beat, dtype: int64
------------------------------------------------------------
2    144707
1     42773
```

```
Name: Priority, dtype: int64
------------------------------------------------------------
933R      17799
SECCK     13784
415       11937
911H       9647
10851      8894
          ...  
148_5A        1
484E          1
A487          1
3056          1
524           1
Name: Incident Type Id, Length: 257, dtype: int64
------------------------------------------------------------
ALARM-RINGER            17799
SECURITY CHECK          13784
911 HANG-UP              9647
STOLEN VEHICLE           8894
MENTALLY ILL             7002
                        ...  
FALSE REPORT OF CRIM        1
INSFRASTRUCTURE SECU        1
YELLOW ALERT AT THE         1
VIOLATION OF PAROLE:        1
REQUIRED TO REGISTER        1
Name: Incident Type Description, Length: 257, dtype: int64
------------------------------------------------------------
LOP141114001069    1
LOP140830000862    1
LOP141211000980    1
LOP140828000671    1
LOP140508000828    1
                  ..
LOP140907000584    1
LOP140130000233    1
LOP141222000805    1
LOP140528000786    1
LOP140526000093    1
Name: Event Number, Length: 187480, dtype: int64
------------------------------------------------------------
2014-06-04T16:31:09.000    3
2014-06-20T01:44:34.000    3
2014-04-16T23:24:34.000    2
2014-11-14T11:22:48.000    2
2014-12-06T03:35:12.000    2
                          ..
```

```
2014-11-24T08:37:19.000    1
2014-07-27T18:57:53.000    1
2014-07-30T16:17:42.000    1
2014-09-22T10:40:07.000    1
2014-01-11T00:39:55.000    1
Name: Closed Time, Length: 186913, dtype: int64
------------------------------------------------------------
{'human_address': '{"address": "INTERNATIONAL BLVD", "city": "", "state": "", "zip": ""}'}                   3713
{'human_address': '{"address": "AV&INTERNATIONAL BLVD", "city": "", "state": "", "zip": ""}'}                3290
{'human_address': '{"address": "MACARTHUR BLVD", "city": "", "state": "", "zip": ""}'}                       2812
{'human_address': '{"address": "BROADWAY", "city": "", "state": "", "zip": ""}'}                             1996
{'human_address': '{"address": "FOOTHILL BLVD", "city": "", "state": "", "zip": ""}'}                        1774
                                                                                                             ...
{'human_address': '{"address": "PABLO CORRIDOR", "city": "", "state": "", "zip": ""}'}                          1
{'human_address': '{"address": "83RD HEGENBERGER RD", "city": "", "state": "", "zip": ""}'}                     1
{'human_address': '{"address": "MYRTLE MACARTHUR BLVD&PIEDMONT AV", "city": "", "state": "", "zip": ""}'}       1
{'human_address': '{"address": "37TH 35TH AV", "city": "", "state": "", "zip": ""}'}                            1
{'human_address': '{"address": "73RD OAK ST", "city": "", "state": "", "zip": ""}'}                             1
Name: Location 1, Length: 35131, dtype: int64
------------------------------------------------------------
14519.0    5
27099.0    3
3790.0     3
4560.0     3
28988.0    2
          ..
5456.0     1
29983.0    1
29975.0    1
1870.0     1
24676.0    1
Name: Zip Codes, Length: 160, dtype: int64
------------------------------------------------------------
============================================================
OP    192581
Name: Agency, dtype: int64
------------------------------------------------------------
2015-04-18T13:52:06.000    3
2015-03-28T11:41:05.000    2
2015-02-09T18:22:50.000    2
2015-12-10T11:05:07.000    2
2015-08-20T19:29:17.000    2
                          ..
2015-04-28T10:13:38.000    1
2015-04-03T11:35:09.000    1
2015-10-03T09:53:41.000    1
2015-08-03T20:04:22.000    1
```

```
2015-07-09T08:15:08.000    1
Name: Create Time, Length: 191944, dtype: int64
------------------------------------------------------------
 INTERNATIONAL BLVD           3695
 AV&INTERNATIONAL BLVD        3106
 MACARTHUR BLVD               3105
 BROADWAY                     2407
 FOOTHILL BLVD                1753
                              ...
82ND CAMPBELL ST                 1
36TH SEMINARY AV                 1
100TH N PICARDY DR               1
SUTTER CLAREMONT AV              1
24TH E 10TH ST                   1
Name: Location, Length: 36515, dtype: int64
------------------------------------------------------------
P3     81629
P1     73141
P2     33423
POU     3787
PCW      595
TEC        6
Name: Area Id, dtype: int64
------------------------------------------------------------
04X     8048
08X     6874
30Y     5690
19X     5564
30X     5542
23X     5492
26Y     5449
34X     5172
06X     5056
03X     4983
07X     4910
29X     4599
31Y     4556
25X     4409
35X     4287
20X     4284
27Y     4242
32X     3940
27X     3899
12Y     3868
09X     3831
33X     3790
21Y     3574
```

```
03Y     3512
32Y     3456
14X     3290
02Y     3290
22X     3207
10Y     2937
26X     2802
24X     2733
10X     2705
28X     2579
24Y     2558
13Z     2555
01X     2552
17Y     2551
31X     2535
12X     2516
02X     2515
21X     2511
05X     2464
22Y     2456
15X     2437
35Y     2293
11X     2186
31Z     2127
14Y     1920
17X     1776
13Y     1734
18Y     1604
16Y     1577
25Y     1406
18X     1263
16X     1223
13X     1117
05Y      775
PDT2      35
Name: Beat, dtype: int64
------------------------------------------------------------
2    150162
1     42418
0         1
Name: Priority, dtype: int64
------------------------------------------------------------
933R      18181
SECCK     14809
415       13677
10851      8899
911H       8529
```

```
          ...
PHONE         1
VICE          1
MS            1
626_9         1
REDALT        1
Name: Incident Type Id, Length: 259, dtype: int64
------------------------------------------------------------
ALARM-RINGER            18181
SECURITY CHECK          14809
STOLEN VEHICLE           8899
911 HANG-UP              8529
MENTALLY ILL             8465
                        ...  
ASSSAULT                    1
IDENTITY THEFT              1
TICKET SCALPING             1
FIREARM AT PUBLIC SC        1
FLOOD                       1
Name: Incident Type Description, Length: 261, dtype: int64
------------------------------------------------------------
LOP150730000474    1
LOP150502000259    1
LOP150429000759    1
LOP150429000806    1
LOP150619000474    1
                  ..
LOP150516000043    1
LOP150817001141    1
LOP150430000900    1
LOP151226000122    1
LOP151210000952    1
Name: Event Number, Length: 192581, dtype: int64
------------------------------------------------------------
2015-02-22T16:19:43.000    2
2015-04-12T22:23:59.000    2
2015-06-10T16:05:09.000    2
2015-06-06T19:59:56.000    2
2015-12-26T08:23:49.000    2
                          ..
2015-10-14T19:45:28.000    1
2015-01-24T07:40:33.000    1
2015-01-20T18:22:39.000    1
2015-08-03T00:46:47.000    1
2015-11-09T08:36:32.000    1
Name: Closed Time, Length: 192006, dtype: int64
------------------------------------------------------------
```

```
============================================================
OP    110827
Name: Agency, dtype: int64
------------------------------------------------------------
2016-05-06T11:21:13.000    3
2016-06-15T15:09:14.000    2
2016-01-29T12:42:34.000    2
2016-03-09T13:34:46.000    2
2016-05-22T21:14:30.000    2
                          ..
2016-02-10T17:35:21.000    1
2016-06-21T19:57:33.000    1
2016-03-23T19:04:44.000    1
2016-06-03T11:13:19.000    1
2016-03-14T20:58:45.000    1
Name: Create Time, Length: 110453, dtype: int64
------------------------------------------------------------
 INTERNATIONAL BLVD           2156
 AV&INTERNATIONAL BLVD        1829
 MACARTHUR BLVD               1813
 BROADWAY                     1472
 7TH ST                       1223
                              ...
15TH OUTLOOK AV                  1
73RD 1ST AV                      1
2ND AV&MONTE CRESTA AV           1
76TH AV&HILLSIDE ST              1
TRASK OAK GROVE AV               1
Name: Location, Length: 24046, dtype: int64
------------------------------------------------------------
P3     47425
P1     41419
P2     19610
POU     2173
PCW      194
TEC        4
JLS        1
WAG        1
Name: Area Id, dtype: int64
------------------------------------------------------------
04X     4515
08X     3931
26Y     3511
30Y     3473
19X     3455
30X     3416
03X     3195
```

```
23X    3076
34X    2857
07X    2831
20X    2702
29X    2646
06X    2580
03Y    2562
27Y    2517
25X    2467
31Y    2460
27X    2333
35X    2328
32X    2316
33X    2276
09X    2158
21Y    2100
32Y    2093
12Y    1987
14X    1832
26X    1766
02X    1746
24X    1704
02Y    1659
10Y    1573
10X    1557
22X    1541
17Y    1482
21X    1479
24Y    1454
31X    1439
22Y    1420
13Z    1397
15X    1393
05X    1342
01X    1304
12X    1299
31Z    1268
28X    1261
11X    1208
35Y    1159
18Y    1102
14Y    1027
17X     969
13Y     952
16Y     907
25Y     739
18X     721
```

```
16X      708
13X      630
05Y      408
PDT2      16
Name: Beat, dtype: int64
------------------------------------------------------------
2.0    86272
1.0    24555
Name: Priority, dtype: int64
------------------------------------------------------------
933R     10094
415       7883
SECCK     7251
10851     5308
911H      5089
         ...  
300WI        1
ABC          1
955B         1
OTC          1
407          1
Name: Incident Type Id, Length: 242, dtype: int64
------------------------------------------------------------
ALARM-RINGER            10094
SECURITY CHECK           7251
STOLEN VEHICLE           5308
911 HANG-UP              5089
MENTALLY ILL             4859
                        ...  
EASTBAY MUD                 1
GRAND THEFT: DOG            1
YELLOW ALERT AT THE         1
ALCOHOL,BEVERAGE AND        1
CHILD TAKEN INTO PRO        1
Name: Incident Type Description, Length: 245, dtype: int64
------------------------------------------------------------
LOP160613000974    1
LOP160704000709    1
LOP160424000732    1
LOP160202000716    1
LOP160526000316    1
                  ..
LOP160406001213    1
LOP160608000436    1
LOP160608000016    1
LOP160505000281    1
LOP160706000538    1
```

```
Name: Event Number, Length: 110827, dtype: int64
-------------------------------------------------------------
2016-05-29T00:43:38.000    3
2016-06-25T15:19:22.000    2
2016-07-27T18:14:06.000    2
2016-06-16T15:38:44.000    2
2016-06-10T00:57:44.000    2
                          ..
2016-04-11T03:46:46.000    1
2016-06-17T04:44:38.000    1
2016-05-04T03:24:50.000    1
2016-07-16T12:06:06.000    1
2016-01-12T13:31:20.000    1
Name: Closed Time, Length: 110451, dtype: int64
-------------------------------------------------------------
=============================================================
```

对数值数据计算五数概括以及缺失值

在这个数据集唯一可以认为的数值数据为案件等级，所以计算案件等级的五数概括

In [14…

```
number_data = ['Priority']
for data in data_all:
    print(data[number_data].describe())
```

```
            Priority
count  180015.000000
mean        1.796111
std         0.402916
min         0.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
            Priority
count  187430.000000
mean        1.776311
std         0.416717
min         1.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
            Priority
count  188051.000000
mean        1.770206
```

```
std         0.420967
min         0.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
            Priority
count  187480.000000
mean        1.771853
std         0.419639
min         1.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
            Priority
count  192581.000000
mean        1.779729
std         0.414443
min         0.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
            Priority
count  110827.000000
mean        1.778438
std         0.415299
min         1.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
```

对6个csv的案件等级进行五数概括后发现，最高的案件等级为2.0，最低为0.0，均值大多都在1.7作左右

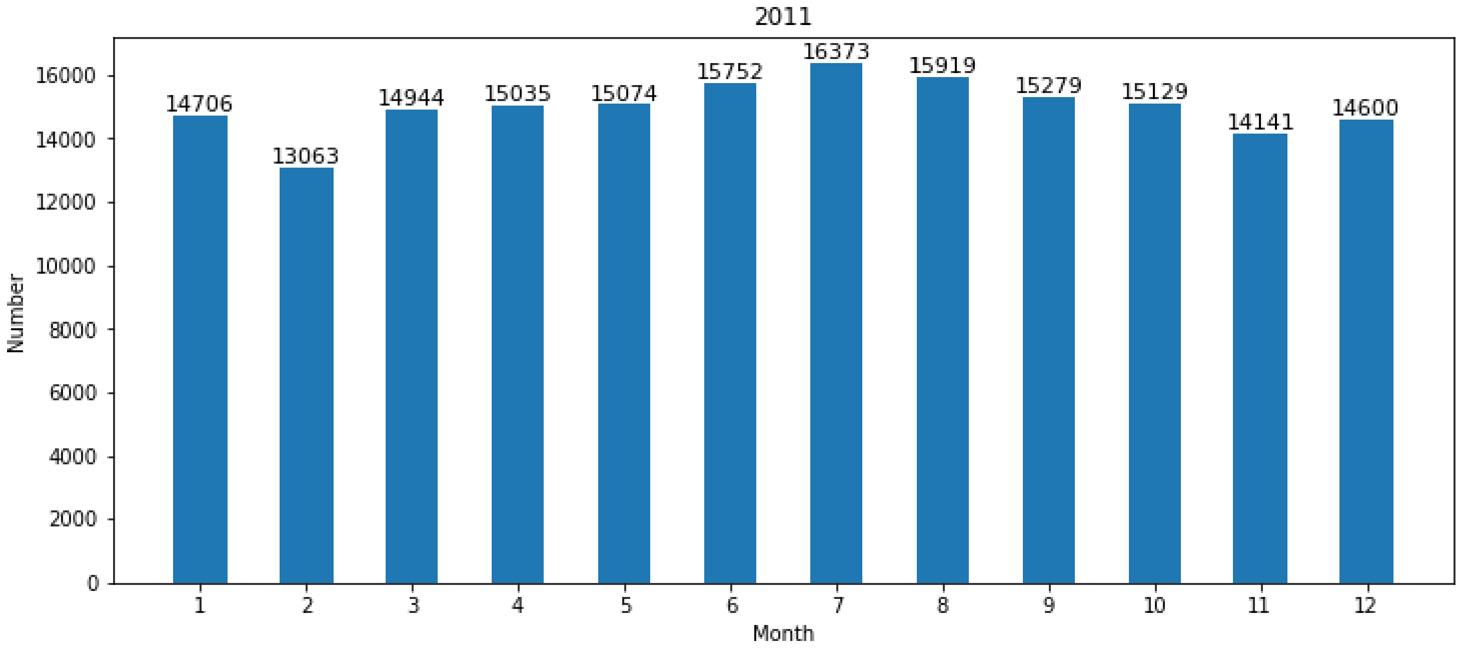
数据可视化

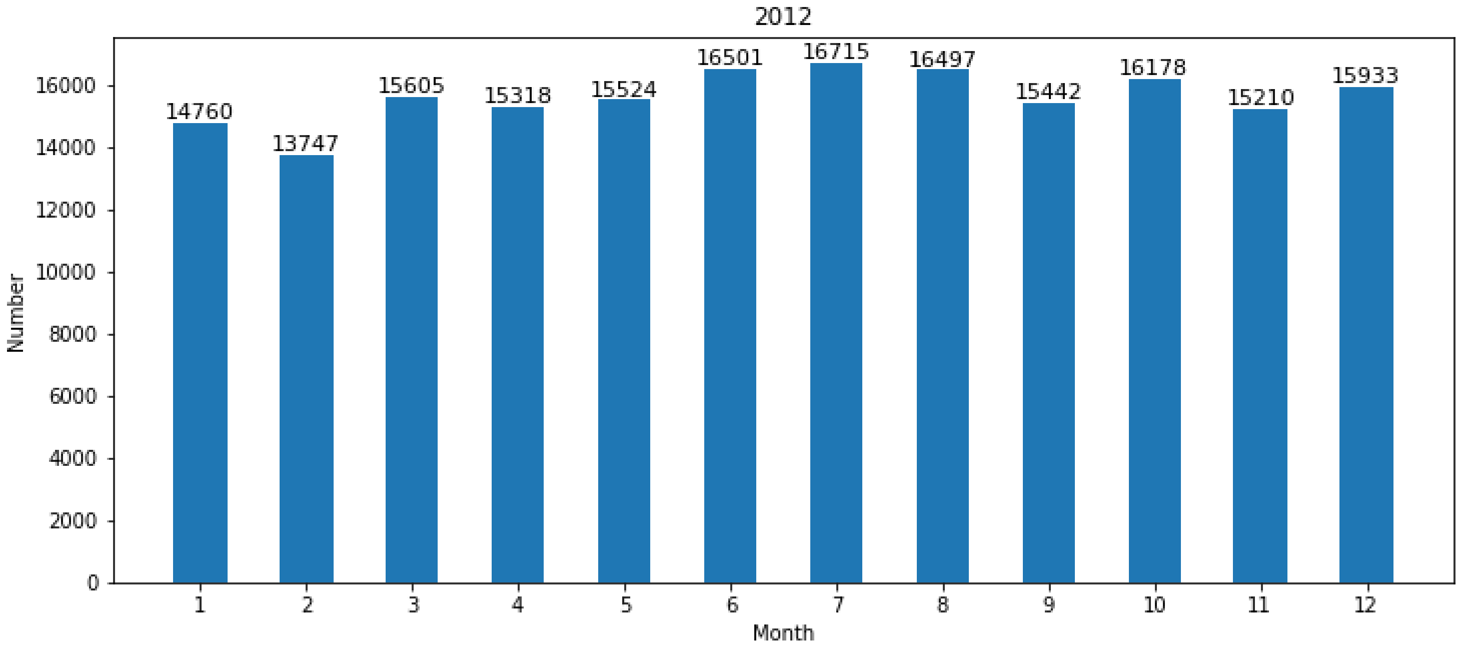
对每年每月立案数量进行可视化分析

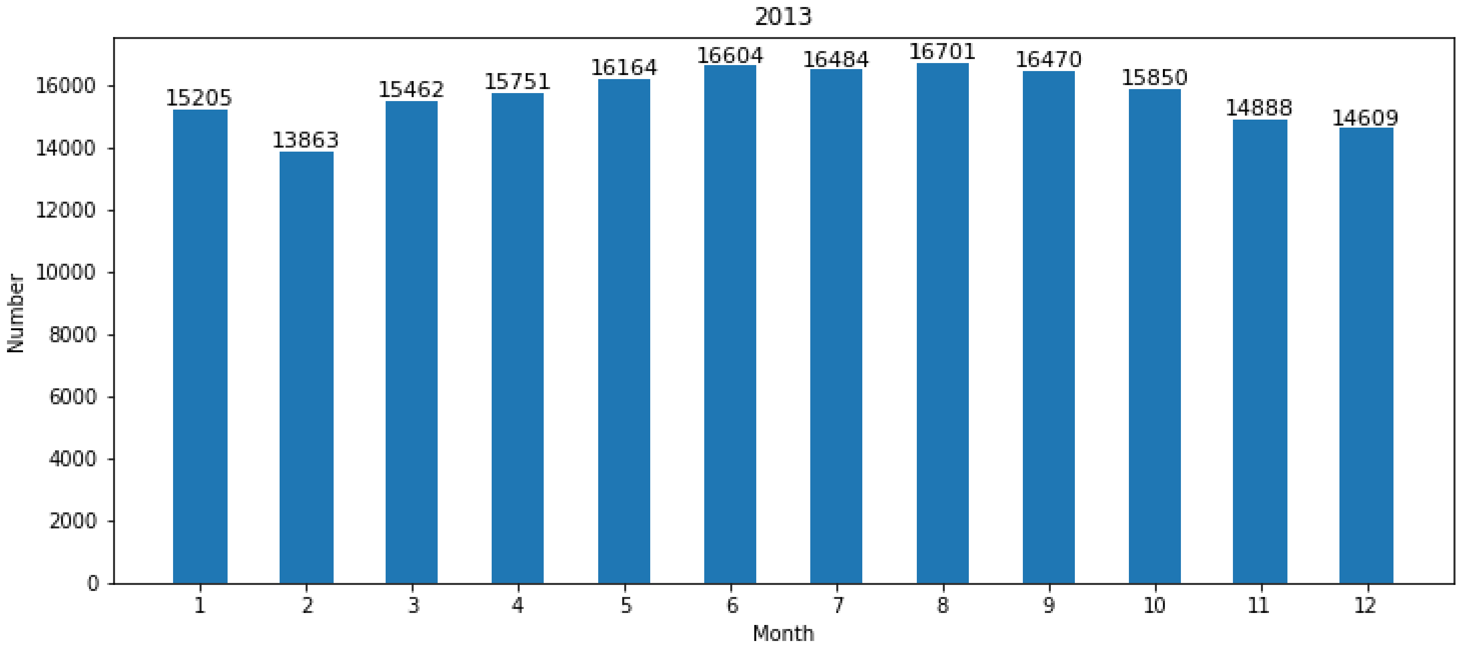
In [93]:

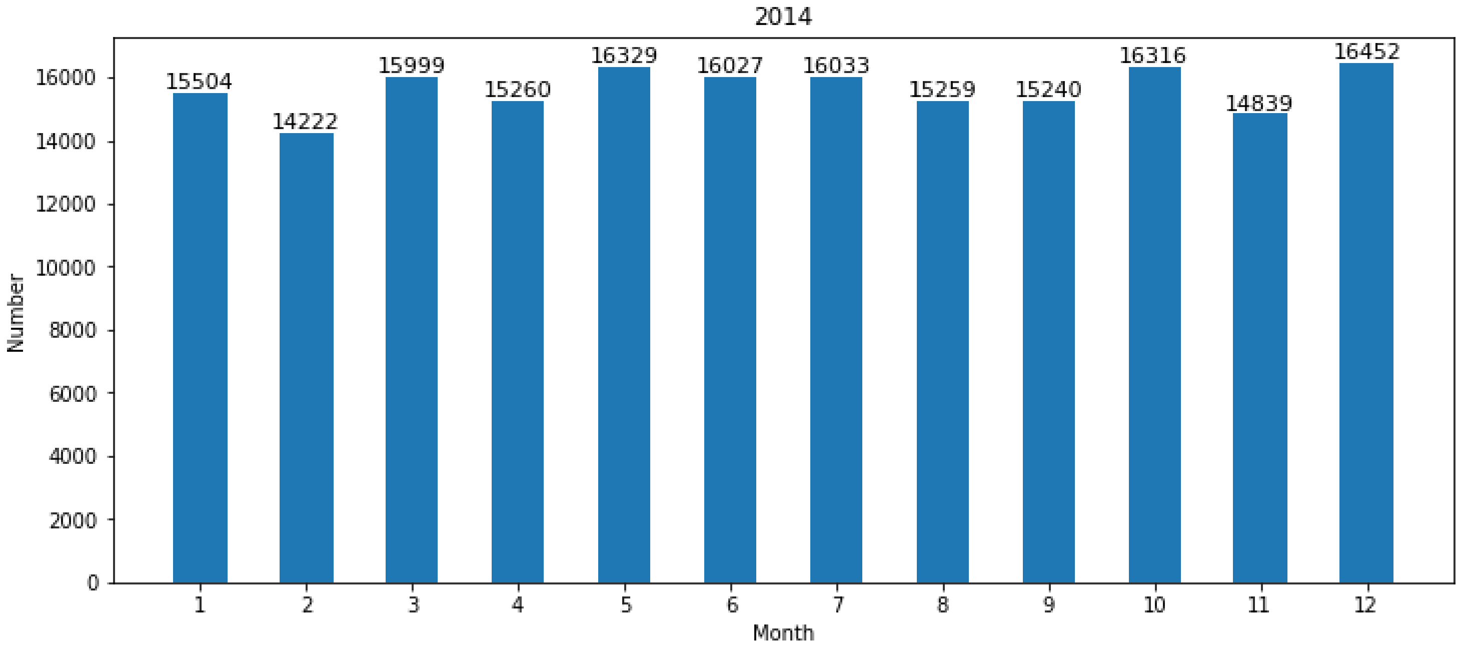
```
t = 0
mon = 0
index = np.arange(12)
k = 1
```

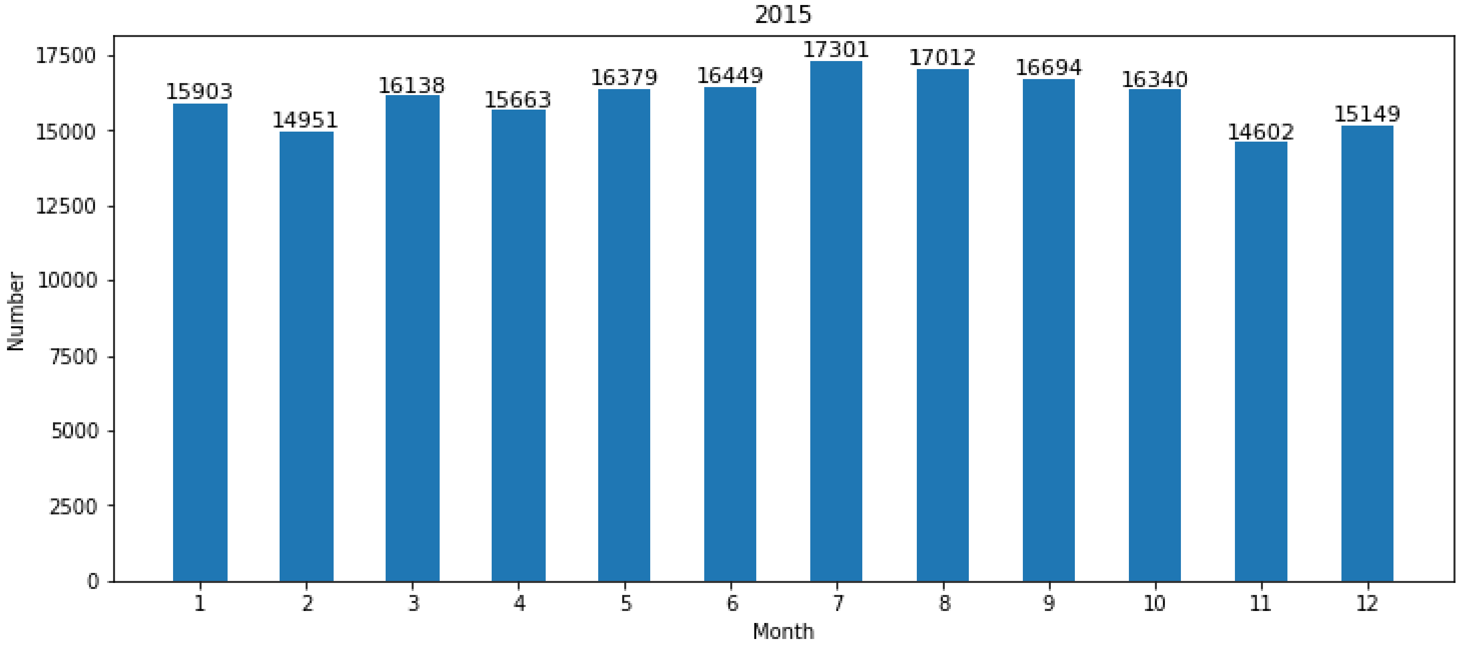
```
year = 2011
for data in data_all:
    lis = data['Create Time']
    lis = lis.dropna()
    lis = lis.values
    mon_count=np.zeros(12)
    for t in lis:
        mon = t[5:7]
        mon_count[int(mon)-1] += 1
    # print(mon_count)
    plt.figure(figsize=(12,5))
    plt.bar(index,mon_count, 0.5, label="mon_count")
    plt.xticks(index,('1','2','3','4','5','6','7','8','9','10','11','12'))
    for a,b in zip(index,mon_count):
        plt.text(a, b+0.05, '%.0f' % b, ha='center', va= 'bottom',fontsize=11)
    plt.xlabel("Month")
    plt.ylabel("Number")
    plt.title(year)
    year=year+1
```

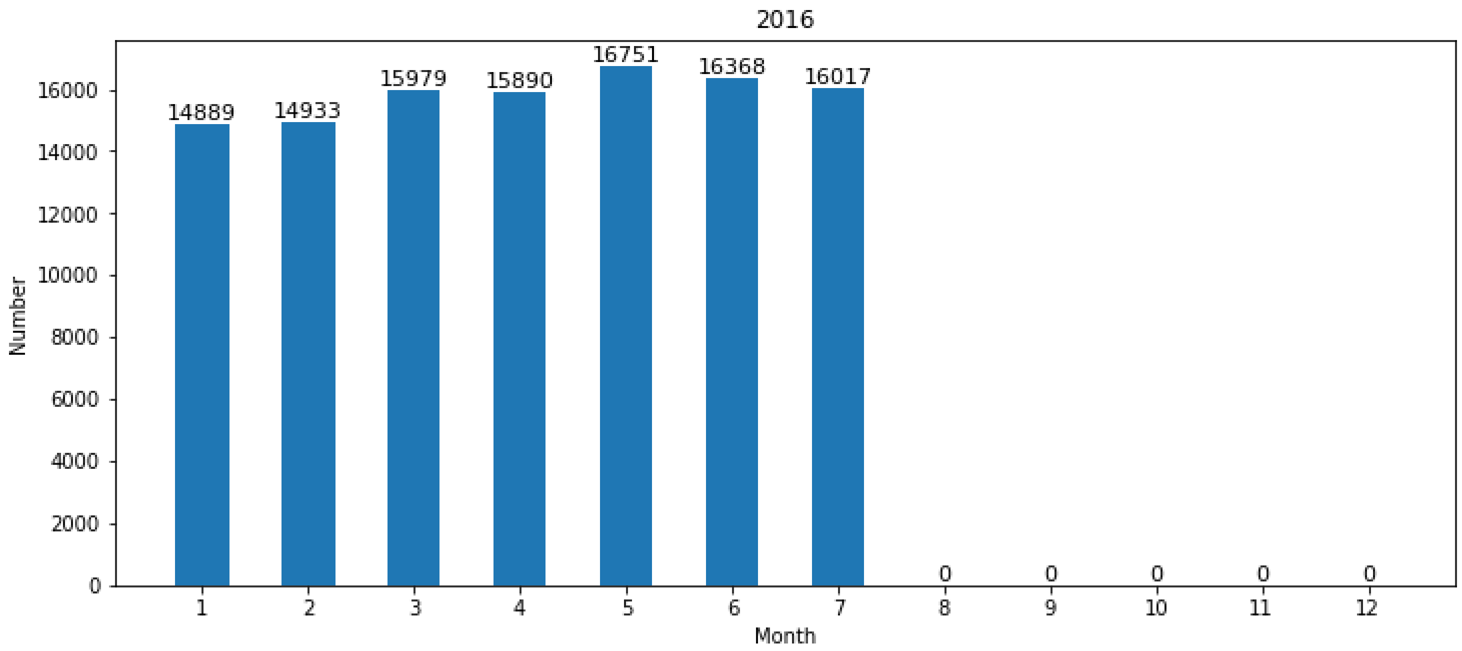



2012
Number
Month
0
2000
4000
6000
8000
10000
12000
14000
16000
1
2
3
4
5
6
7
8
9
10
11
12
14760
13747
15605
15318
15524
16501
16715
16497
15442
16178
15210
15933


2013
Number
Month
0
2000
4000
6000
8000
10000
12000
14000
16000
1
2
3
4
5
6
7
8
9
10
11
12
15205
13863
15462
15751
16164
16604
16484
16701
16470
15850
14888
14609


2014
Number
0
2000
4000
6000
8000
10000
12000
14000
16000
Month
1
2
3
4
5
6
7
8
9
10
11
12
15504
14222
15999
15260
16329
16027
16033
15259
15240
16316
14839
16452


2015
Number
0
2500
5000
7500
10000
12500
15000
17500
Month
1
2
3
4
5
6
7
8
9
10
11
12
15903
14951
16138
15663
16379
16449
17301
17012
16694
16340
14602
15149

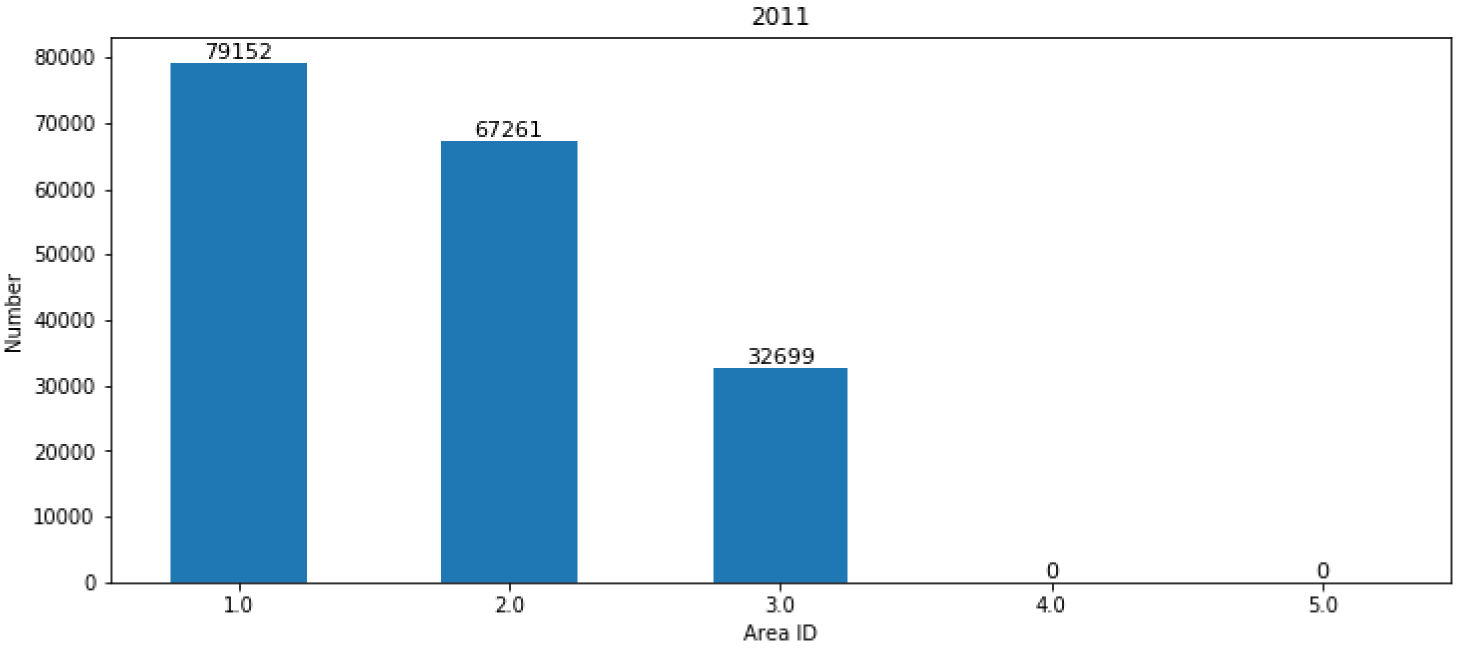


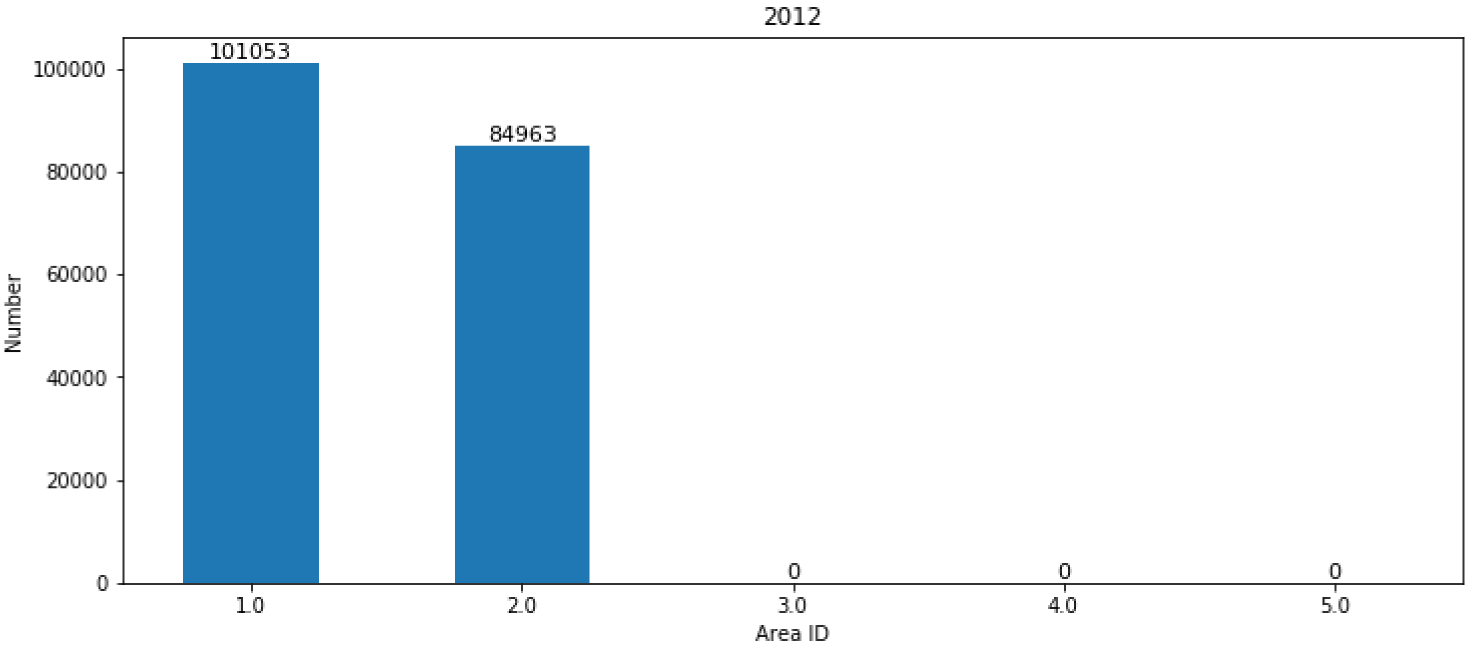
从直方图的结果中可以发现，每年每月立案数量较为平均，且每年每月的案件数量多数在1.5万以上，仅有少数低于1.5万，同时缺失2016年8月之后的数据，这说明在此之后的数据没有进行记录

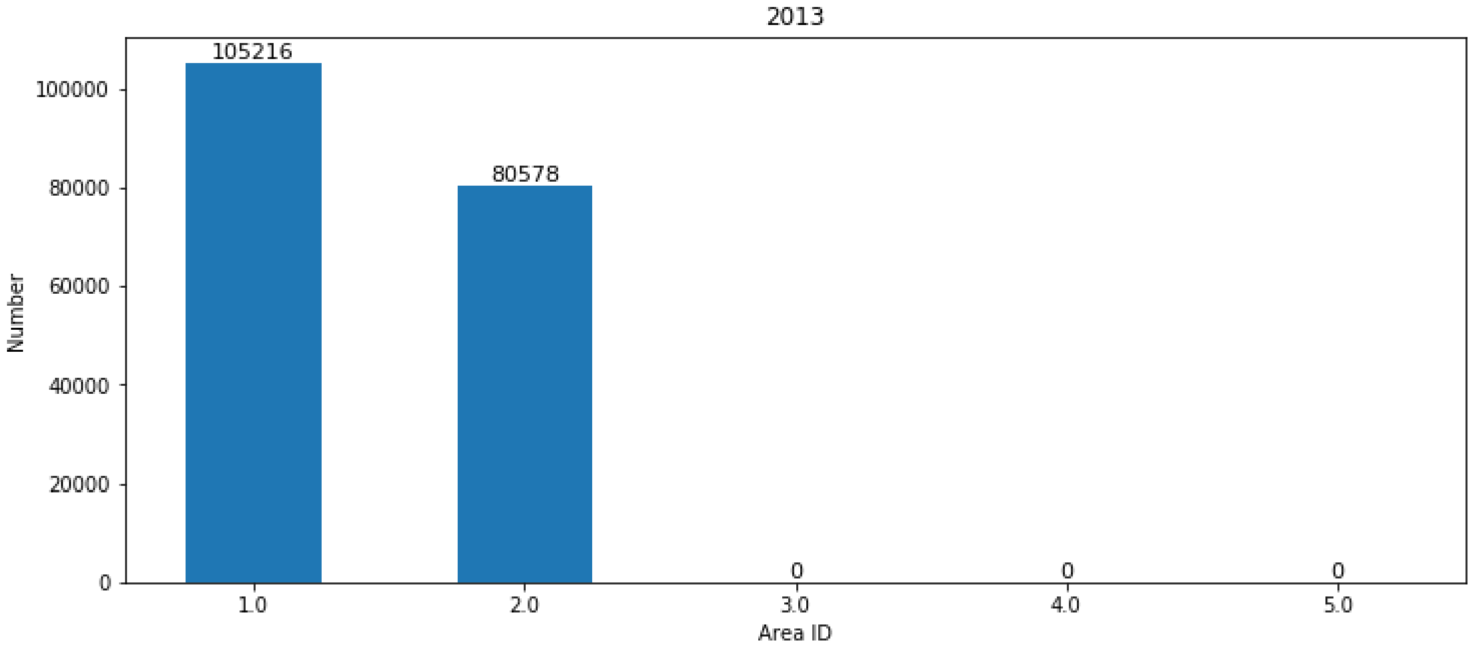
对区域立案数量进行可视化分析

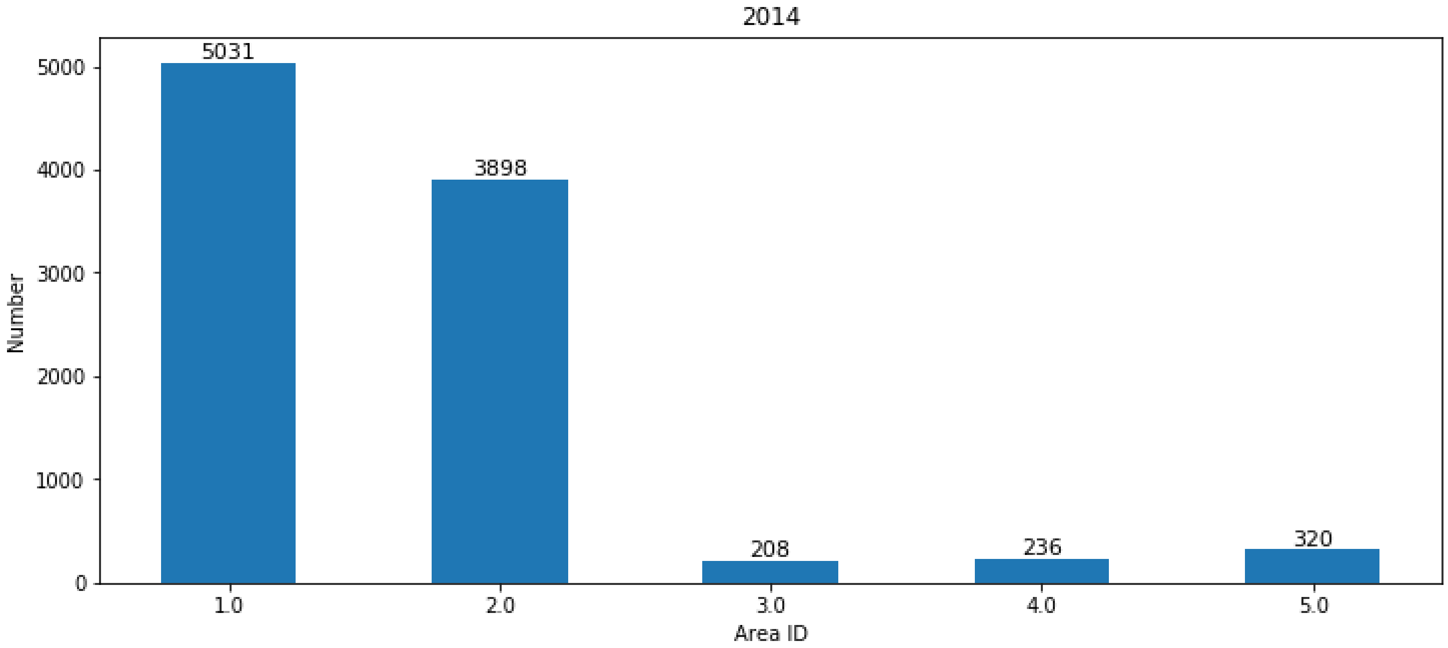
In [13…

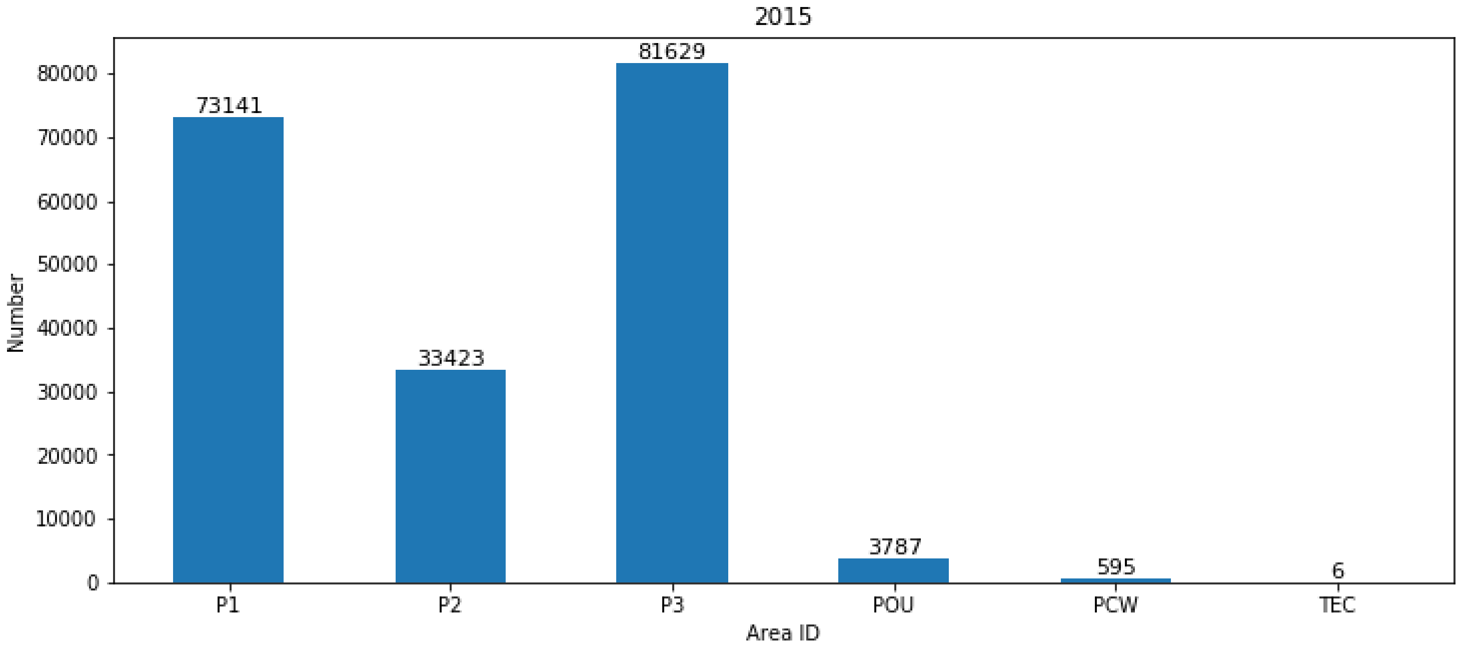
```
'''
P3      81629
P1      73141
P2      33423
POU      3787
PCW       595
TEC         6
'''

index = np.arange(5)
year = 2011
for data in [data1,data2,data3,data4]:
    lis = data['Area Id']
    lis = lis.dropna()
    lis = lis.values
    Area_count=np.zeros(5)
    for t in lis:
        Area_count[int(t)-1] += 1
        # Area_count[int(t)+5] += 1
    # print(mon_count)
    plt.figure(figsize=(12,5))
    plt.bar(index,Area_count, 0.5, label="Area_count")
    plt.xticks(index,('1.0','2.0','3.0','4.0','5.0'))
    for a,b in zip(index,Area_count):
        plt.text(a, b+0.05, '%.0f' % b, ha='center', va= 'bottom',fontsize=11)
    plt.xlabel("Area ID")
    plt.ylabel("Number")
    plt.title(year)
    year=year+1

index = np.arange(6)
for data in [data5,data6]:
    lis = data['Area Id']
    lis = lis.dropna()
    lis = lis.values
    Area_count=np.zeros(6)
    for t in lis:
        if t == 'P1':
            Area_count[0] += 1
        if t == 'P2':
            Area_count[1] += 1
        if t == 'P3':
            Area_count[2] += 1
        if t == 'POU':
            Area_count[3] += 1
        if t == 'PCW':
            Area_count[4] += 1
        if t == 'TEC':
            Area_count[5] += 1
    plt.figure(figsize=(12,5))
    plt.bar(index,Area_count, 0.5, label="Area_count")
    plt.xticks(index,('P1','P2','P3','POU','PCW','TEC'))
    for a,b in zip(index,Area_count):
        plt.text(a, b+0.05, '%.0f' % b, ha='center', va= 'bottom',fontsize=11)
    plt.xlabel("Area ID")
    plt.ylabel("Number")
    plt.title(year)
    year=year+1
```

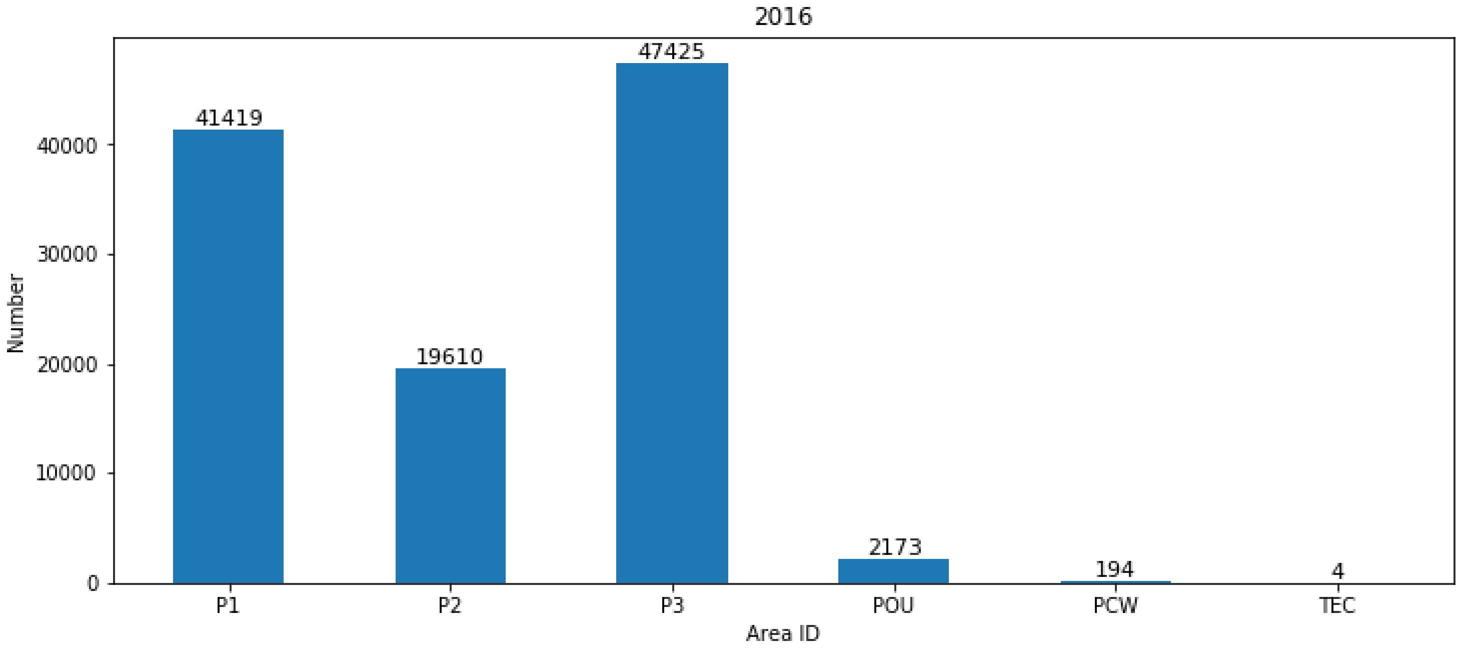

2011
Number
Area ID
79152
67261
32699
0
0
1.0
2.0
3.0
4.0
5.0
0
10000
20000
30000
40000
50000
60000
70000
80000


2012
Number
Area ID
101053
84963
0
0
0
1.0
2.0
3.0
4.0
5.0
0
20000
40000
60000
80000
100000


2013
105216
80578
0
0
0
Number
100000
80000
60000
40000
20000
0
1.0
2.0
3.0
4.0
5.0
Area ID


2014
5031
3898
208
236
320
Number
5000
4000
3000
2000
1000
0
1.0
2.0
3.0
4.0
5.0
Area ID





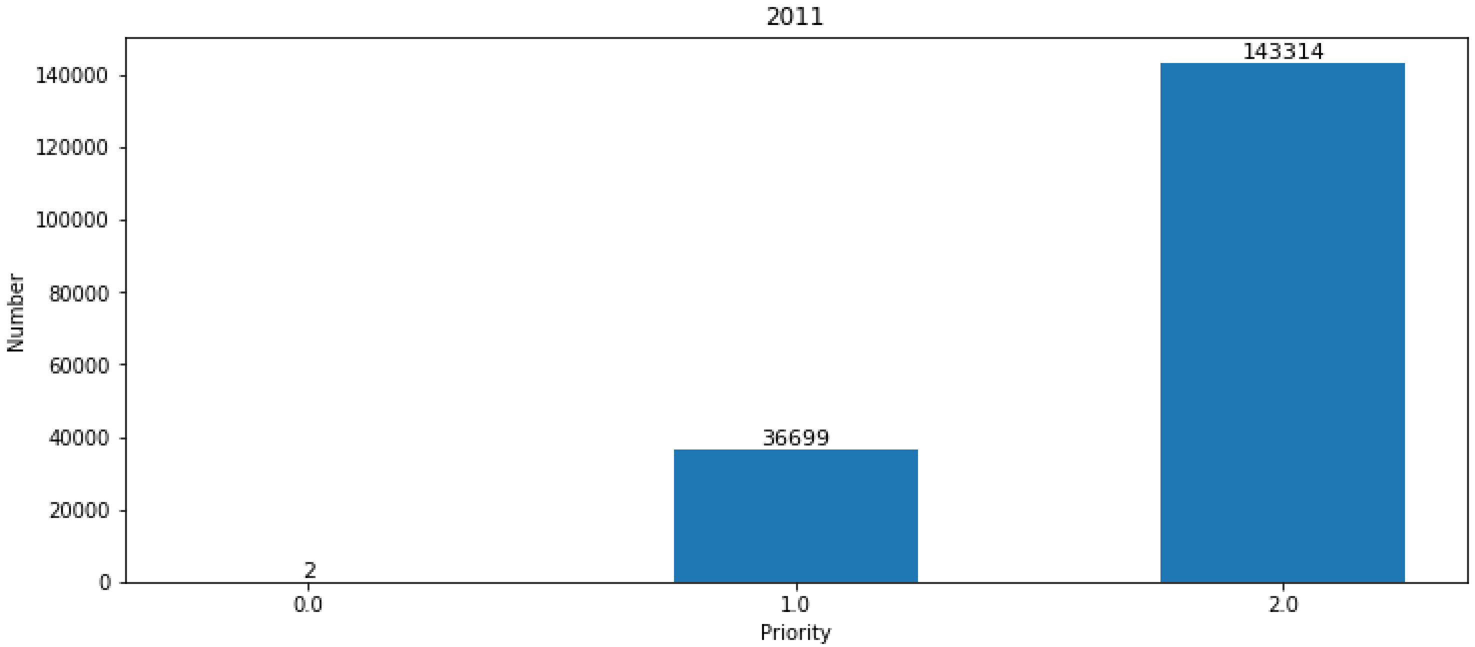
- 在2011年在ID 为1.0的区域，案件数量最多，同时4.0 和 5.0区域没有任何案件

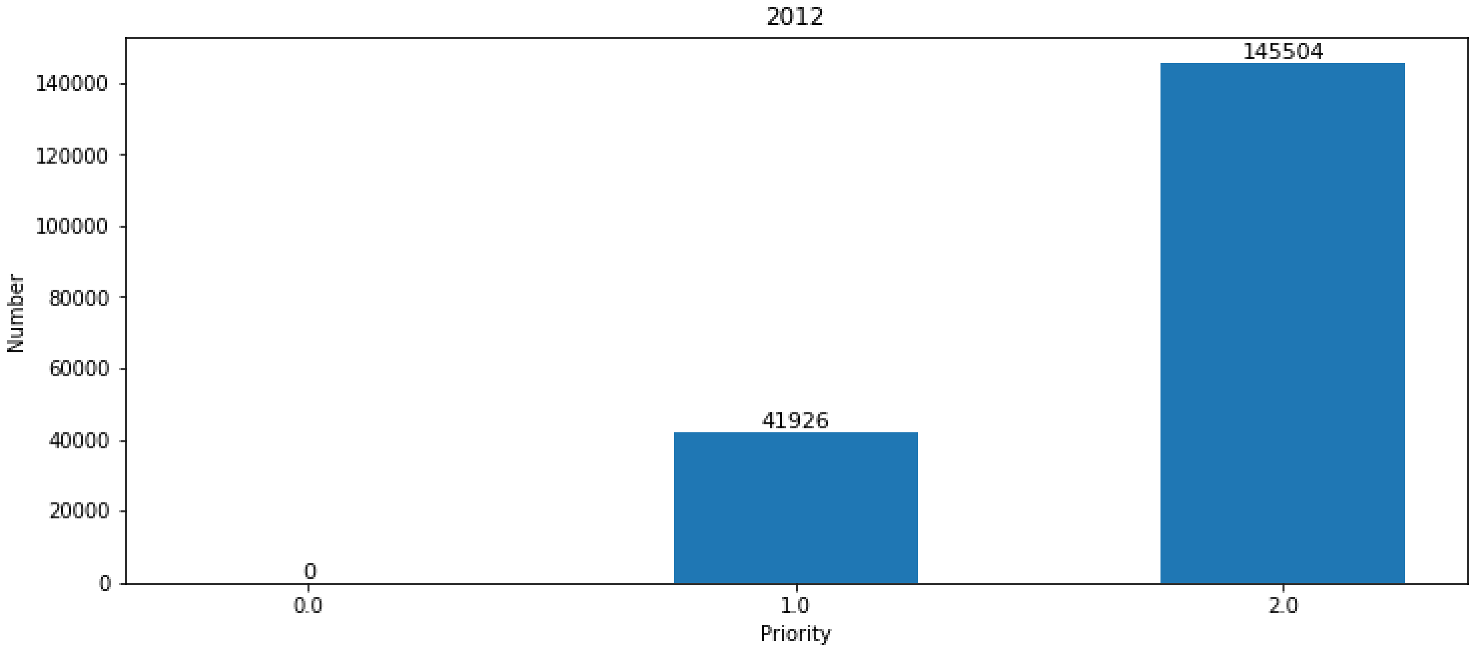
- 当在2012时，相比于2011年1.0区域和2.0区域的案件数量增多，同时3.0区域没有任何案件
- 2013年的案件每个区域的案件数量与2012年的数量相似
- 2014年4.0和5.0也开始有案件
- 2015和2016年的区域ID发现变化
- 在2015年和2016年中P3区域的案件数量最多

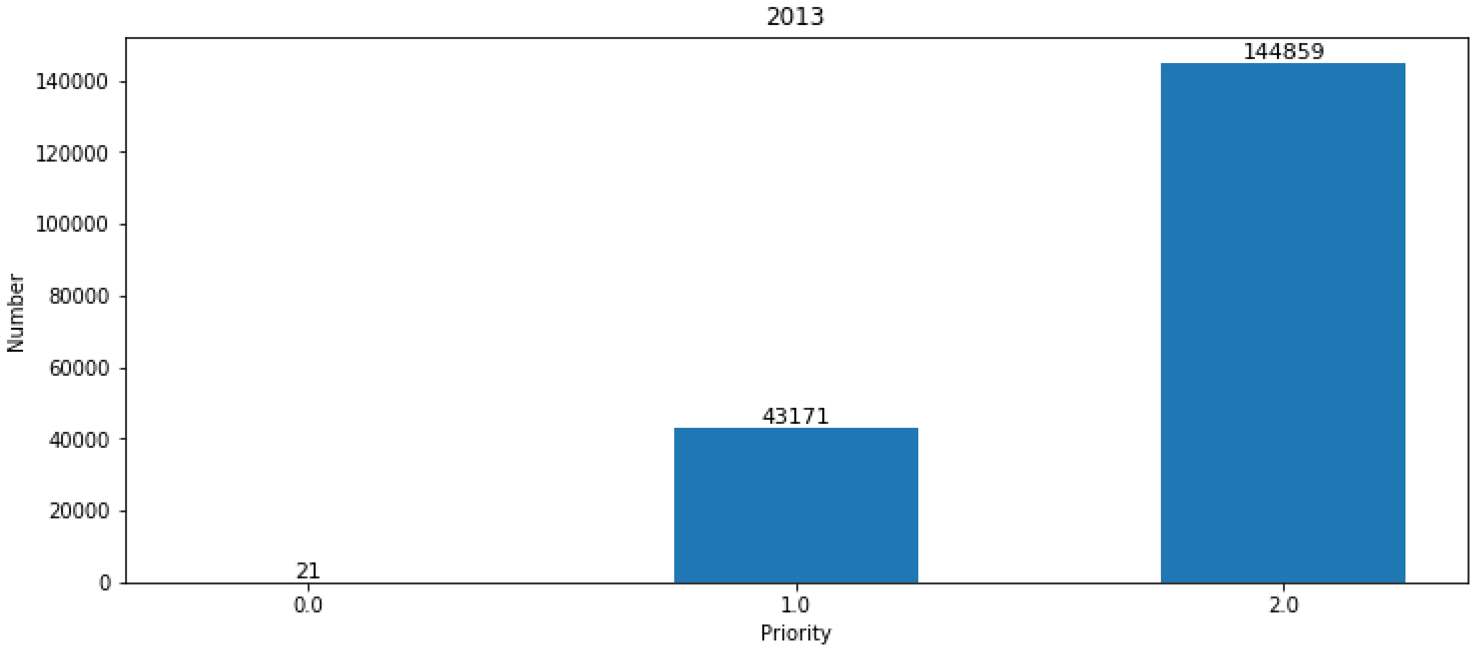
对事件等级进行可视化分析

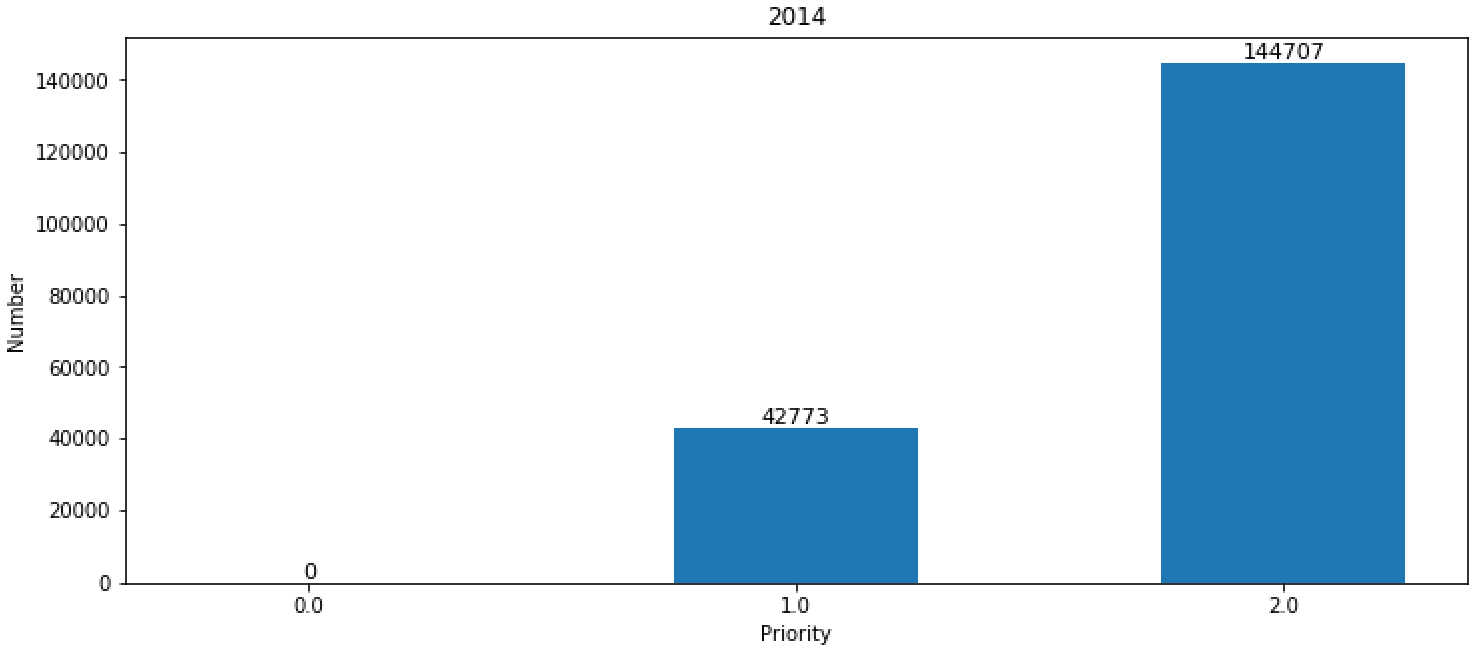
In [22…

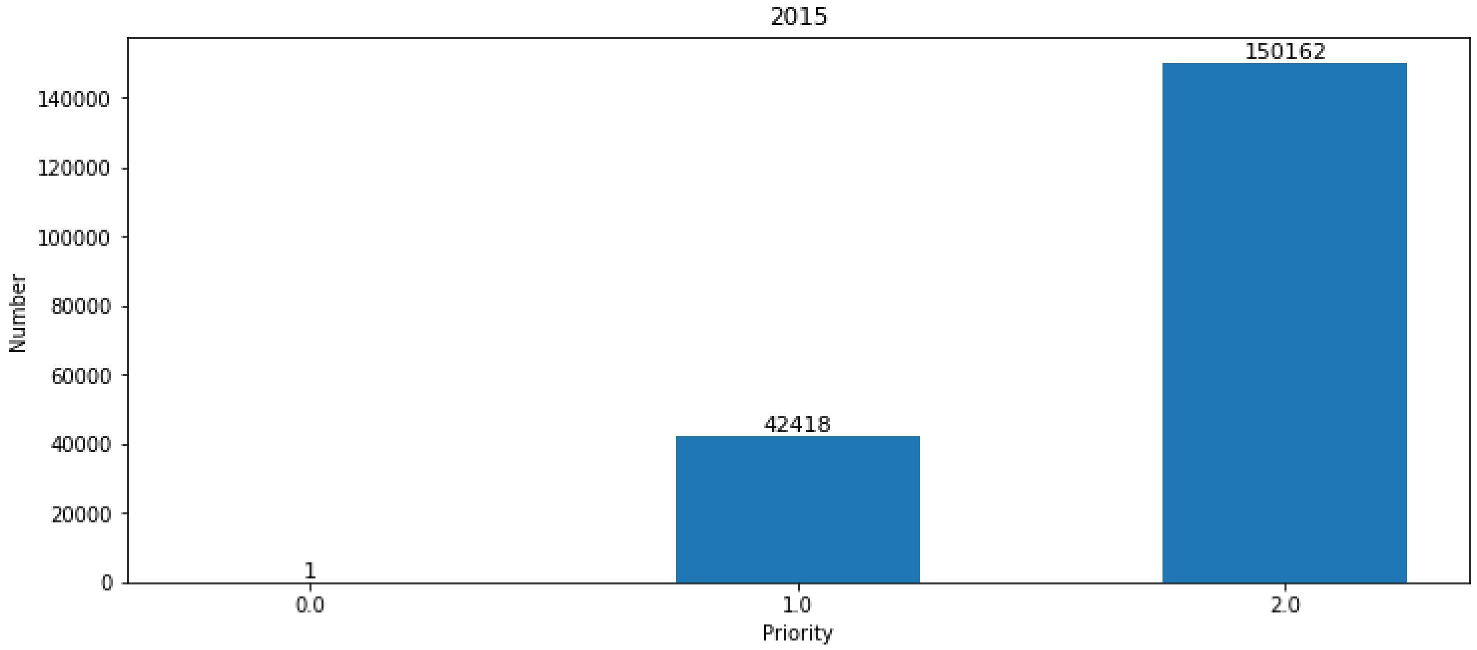
```
year = 2011
index = np.arange(3)
for data in data_all:
    lis = data['Priority']
    lis = lis.dropna()
    lis = lis.values
    count=np.zeros(3)
    for t in lis:
        count[int(t)] += 1
    plt.figure(figsize=(12,5))
    plt.bar(index,count, 0.5, label="count")
    plt.xticks(index,('0.0','1.0','2.0'))
    for a,b in zip(index,count):
        plt.text(a, b+0.05, '%.0f' % b, ha='center', va= 'bottom',fontsize=11)
    plt.xlabel("Priority")
    plt.ylabel("Number")
    plt.title(year)
    year=year+1
```

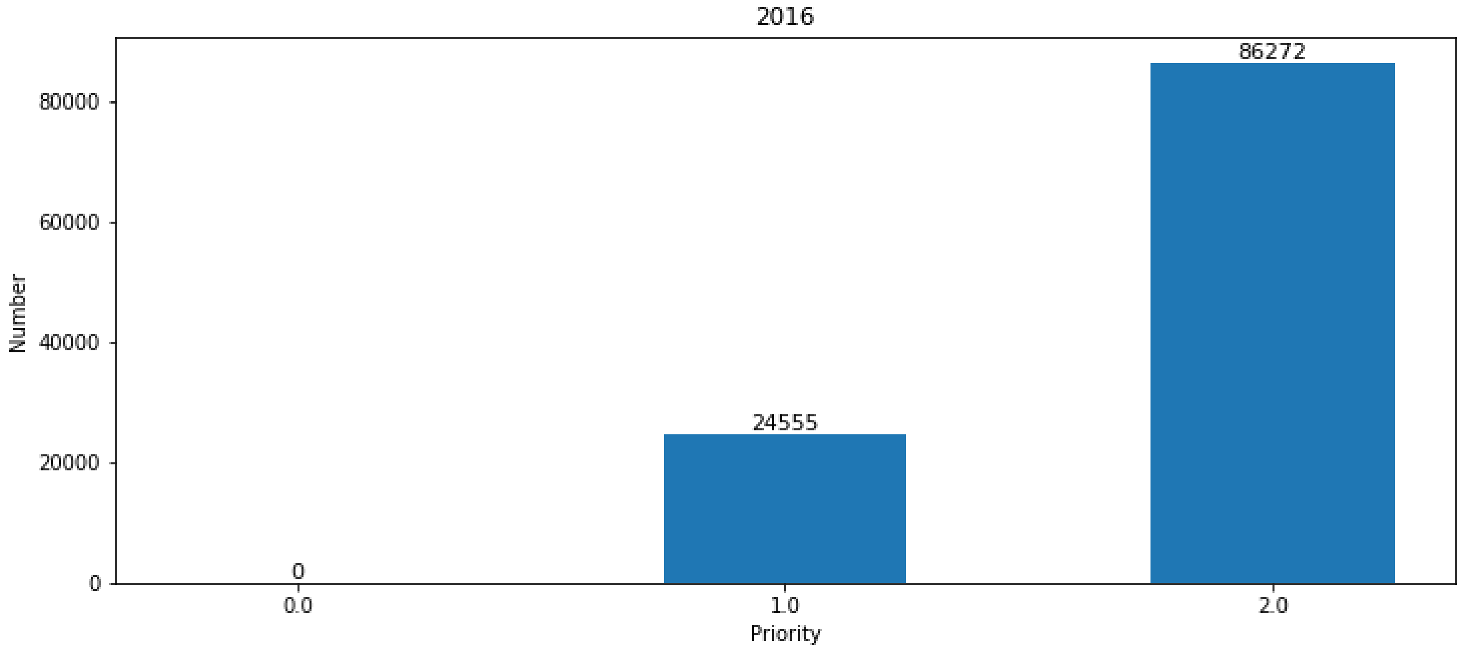

2011
Number
140000
120000
100000
80000
60000
40000
20000
0
2
36699
143314
0.0
1.0
2.0
Priority


2012
Number
140000
120000
100000
80000
60000
40000
20000
0
0
41926
145504
0.0
1.0
2.0
Priority


2013
Number
140000
120000
100000
80000
60000
40000
20000
0
21
43171
144859
0.0
1.0
2.0
Priority


2014
Number
140000
120000
100000
80000
60000
40000
20000
0
0
42773
144707
0.0
1.0
2.0
Priority

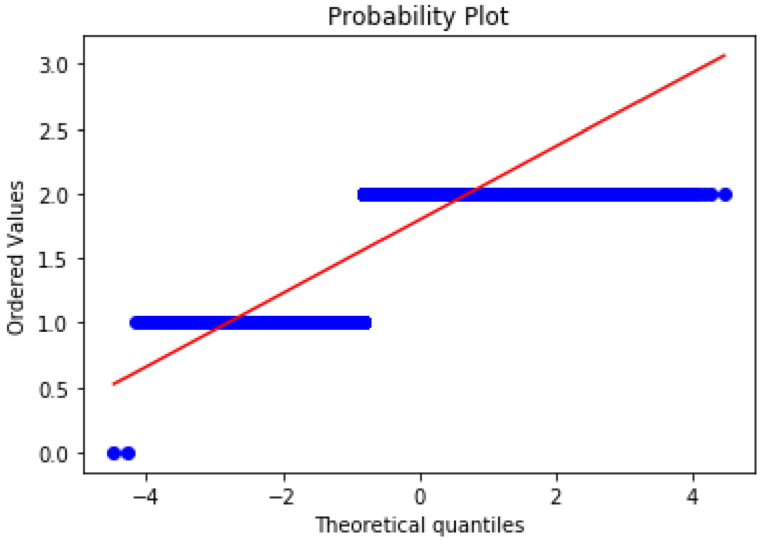


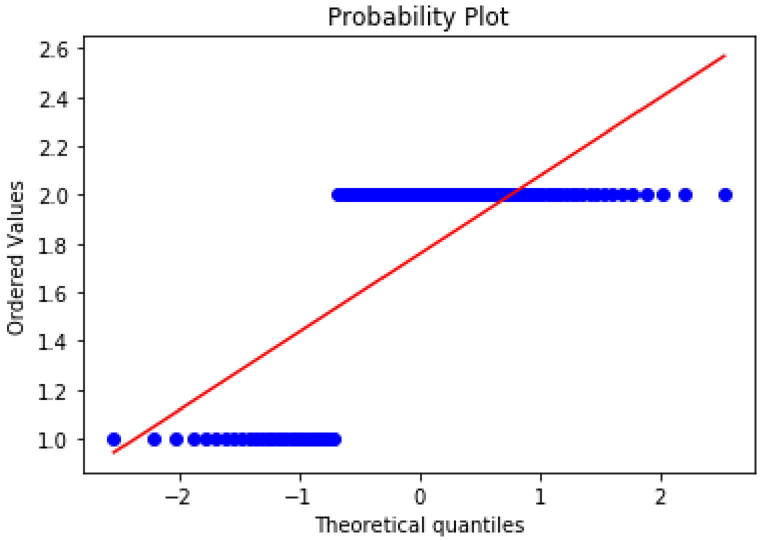


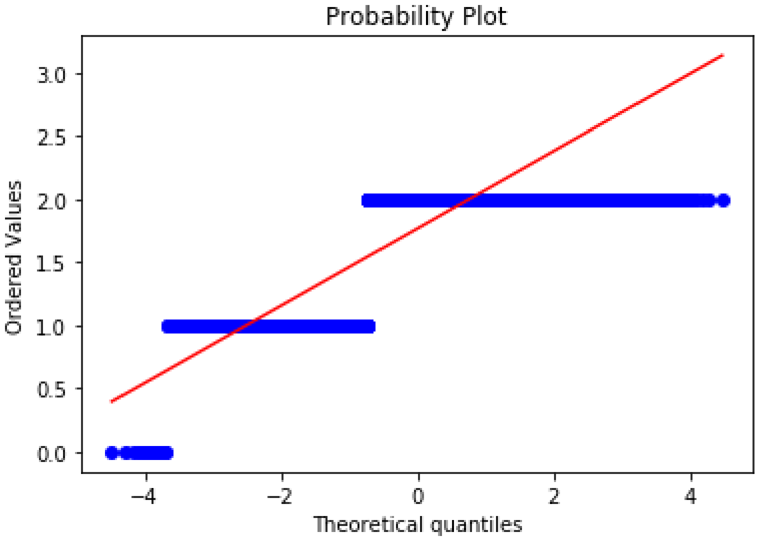
从直方图中可以发现，大多数的案件等级为2.0，等级为1.0的案件十分稀少

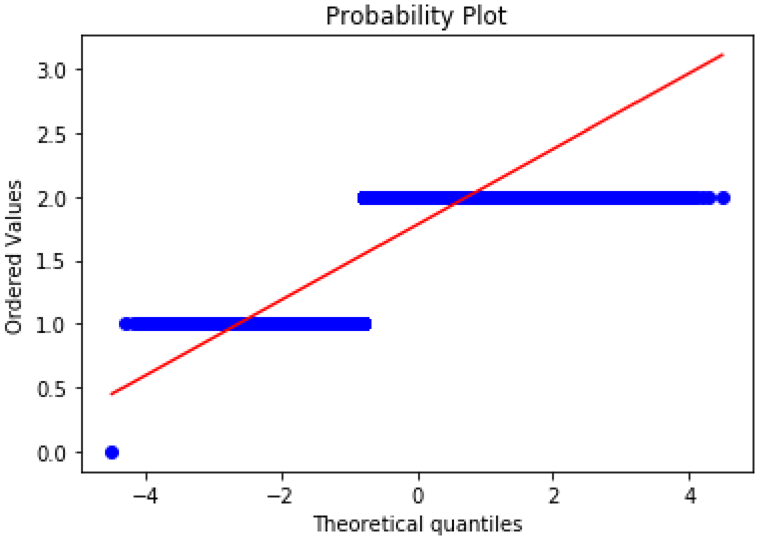
In [15…

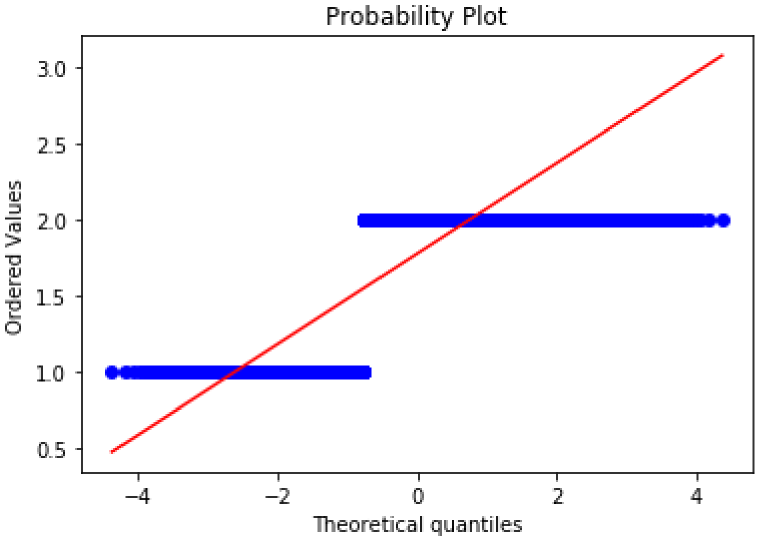
```
for data in data_all:
    data = data.dropna()
    stats.probplot(data['Priority'],dist="norm",plot=plt)
    plt.show()
for data in data_all:
    data.boxplot(column=['Priority'])
    plt.show()
```

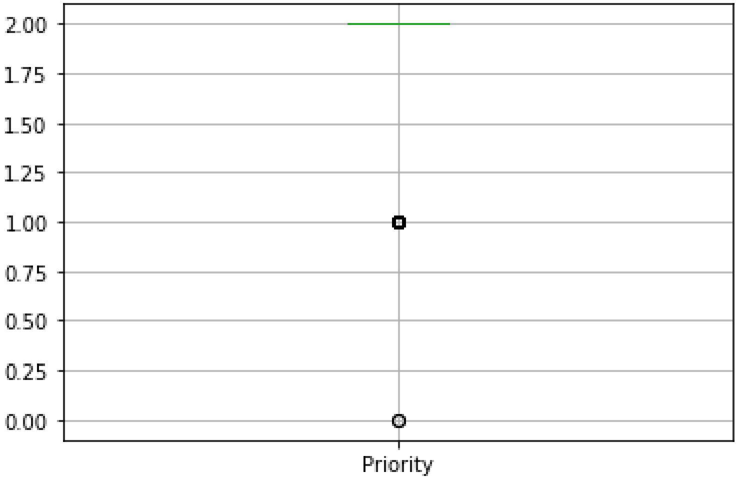


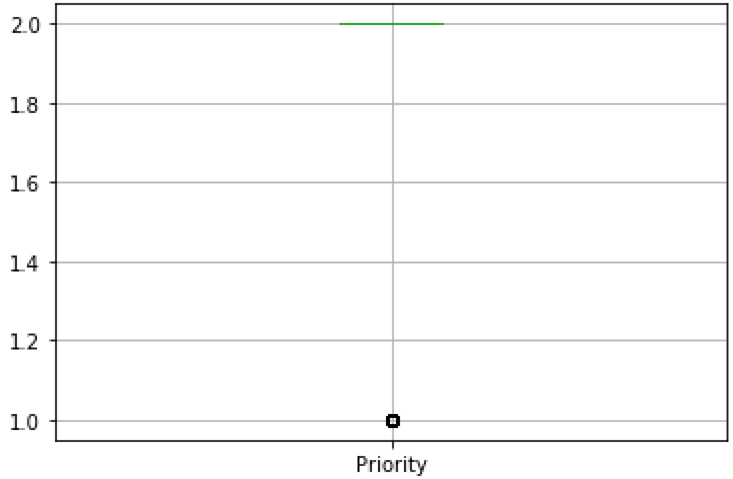


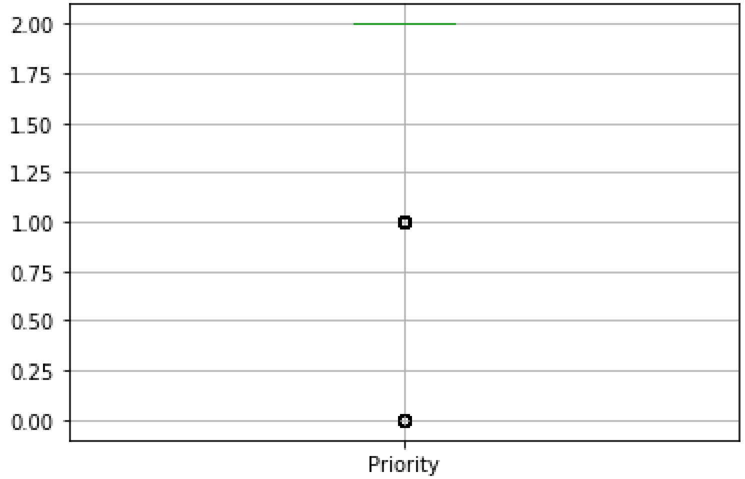

Probability Plot
Ordered Values
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−4
−2
0
2
4
Theoretical quantiles

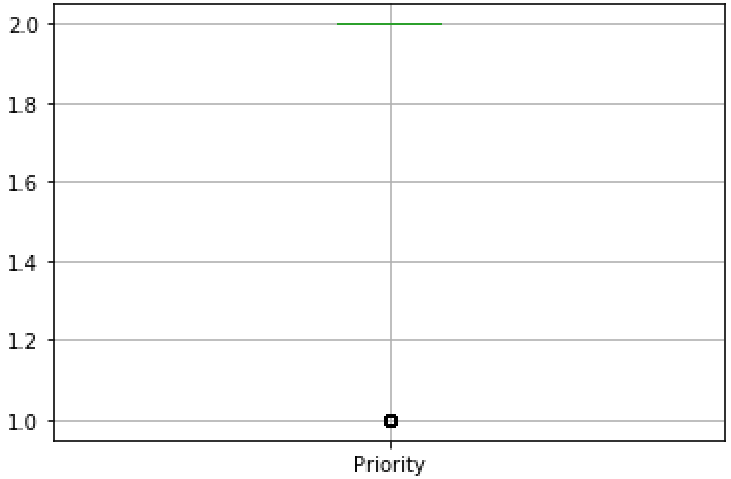

Probability Plot
Ordered Values
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−4
−2
0
2
4
Theoretical quantiles

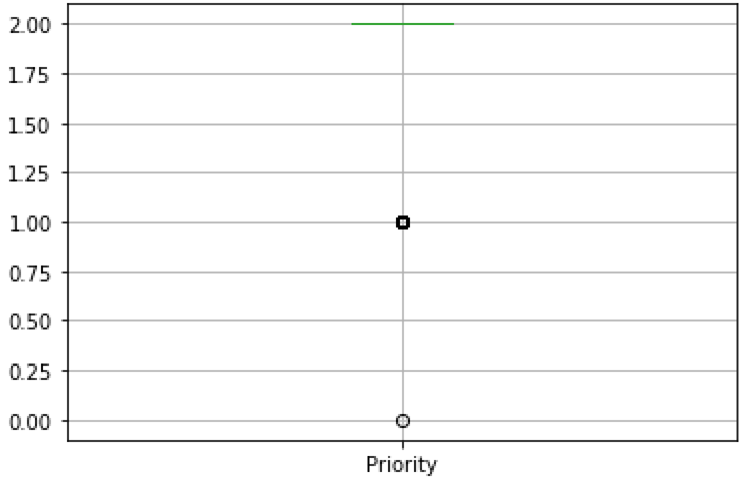

Probability Plot
Ordered Values
3.0
2.5
2.0
1.5
1.0
0.5
−4
−2
0
2
4
Theoretical quantiles

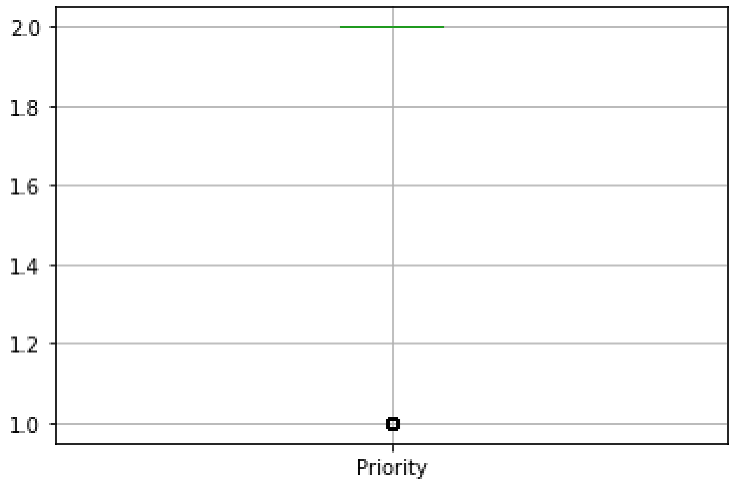

2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Priority


2.0
1.8
1.6
1.4
1.2
1.0
Priority


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Priority


2.0
1.8
1.6
1.4
1.2
1.0
Priority


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Priority



案件等级不服从正态分布，这说明2.0为最低等级的案件，因为高等级的案件很少发现，符合现实生活

数据缺失的处理

In [13…

```
for data,cols in zip(data_all,cols_all):
    print(data.isnull()[cols].sum())
    print("=" * 60)
```

```
Agency                         1
Create Time                    1
Location                       0
Area Id                      904
Beat                         520
Priority                       1
Incident Type Id               1
Incident Type Description      1
Event Number                   1
Closed Time                    7
dtype: int64
============================================================
Agency                            1
Create Time                       1
Area Id                        1415
Beat                            984
Priority                          1
Incident Type Id                  1
Incident Type Description         1
Event Number                      1
```

```
Closed Time                      19
Location 1                       70
Zip Codes                    187256
dtype: int64
============================================================
Agency                          1
Create Time                     1
Location                        0
Area Id                      2258
Beat                         1178
Priority                        1
Incident Type Id                1
Incident Type Description       5
Event Number                    1
Closed Time                     2
dtype: int64
============================================================
Agency                            0
Create Time                       0
Area Id                      177787
Beat                           1217
Priority                          0
Incident Type Id                  0
Incident Type Description       141
Event Number                      0
Closed Time                       0
Location 1                       42
Zip Codes                    187303
dtype: int64
============================================================
Agency                          0
Create Time                     0
Location                        0
Area Id                         0
Beat                         1325
Priority                        0
Incident Type Id                0
Incident Type Description     243
Event Number                    0
Closed Time                     0
dtype: int64
============================================================
Agency                         1
Create Time                    1
Location                       0
Area Id                        1
Beat                         581
```

```
Priority                       1
Incident Type Id               1
Incident Type Description      1
Event Number                   1
Closed Time                    1
dtype: int64
============================================================
```

从缺失数量上发现，主要缺失的Area ID和Beat属性，尤其2014年的csv文件

In [21…

```
for data in data_all:
    data_ = data[['Incident Type Id','Incident Type Description']]
    data_=data_[data_.isnull().T.any()]
    print(data_)
    print("=" * 50)
```

```
       Incident Type Id Incident Type Description
180015              NaN                       NaN
==================================================
       Incident Type Id Incident Type Description
187255              NaN                       NaN
==================================================
       Incident Type Id Incident Type Description
178947              JGP                       NaN
185820              JGP                       NaN
186584              JGP                       NaN
187409              JGP                       NaN
188051              NaN                       NaN
==================================================
       Incident Type Id Incident Type Description
2382                JGP                       NaN
11137               JGP                       NaN
13174               JGP                       NaN
18605               JGP                       NaN
37673               JGP                       NaN
...                 ...                       ...
182424              JGP                       NaN
183100              JGP                       NaN
184135              JGP                       NaN
186580              JGP                       NaN
187323              JGP                       NaN

[141 rows x 2 columns]
==================================================
       Incident Type Id Incident Type Description
1725                JGP                       NaN
1756                JGP                       NaN
```

```
2765                JGP                      NaN
3230                JGP                      NaN
3772                JGP                      NaN
...                 ...                      ...
187356              JGP                      NaN
188118              JGP                      NaN
189202              JGP                      NaN
190735              JGP                      NaN
191221              JGP                      NaN

[243 rows x 2 columns]
==================================================
       Incident Type Id Incident Type Description
110827              NaN                      NaN
==================================================
```

分析 Incident Type Id Incident 和 Type Description，发现类型为JGP的案件其事件描述均为Nan，这可能说明JGP难以描述。

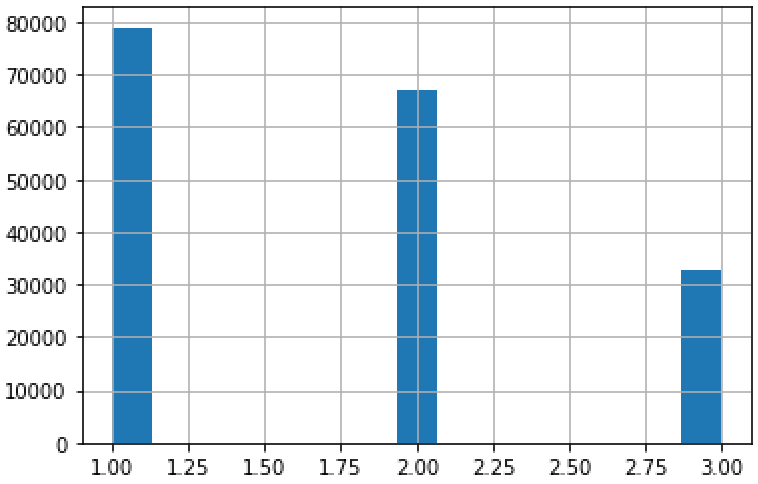
将缺失部分剔除

In [22…

```
del_df = data1.dropna()
del_df['Area Id'].hist(bins = 15)
```

Out[223]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x28503fcfec8>
```

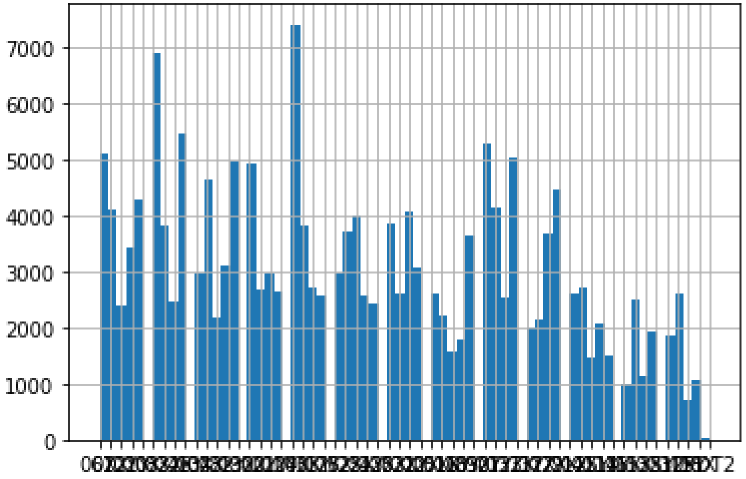


In [22…

```
del_df['Beat'].hist(bins =70)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x285040c3d48>
```

Out[224]:



用最高频率值来填补缺失值

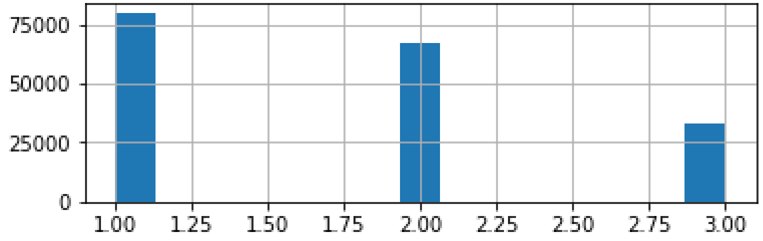
In [17…

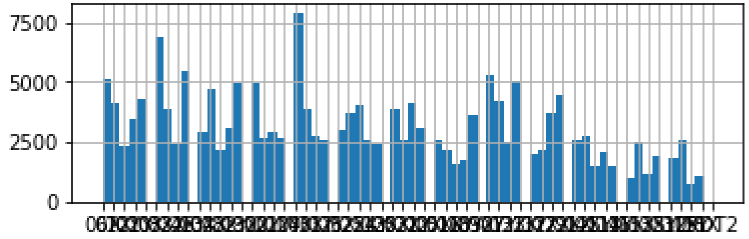
```
for data in data_all:
    fill_max = data.fillna({’Area Id’: data[’Area Id’].mode().item(), ’Beat’: data[’Beat’].mode().item()})
    print(fill_max[’Area Id’].value_counts())
    print("=" * 30)
    print(fill_max[’Beat’].value_counts())
    plt.subplot(2,1,1)
    fill_max[’Area Id’].hist(bins = 15)
    plt.show()
    plt.subplot(2,1,2)
    fill_max[’Beat’].hist(bins =70)
    plt.show()
```

```
1.0     80056
2.0     67261
3.0     32699
Name: Area Id, dtype: int64
==============================
04X      7930
08X      6885
26Y      5478
30Y      5295
06X      5119
23X      5051
30X      4956
```

```
19X      4955
34X      4673
29X      4483
20X      4287
27Y      4159
07X      4134
31Y      4082
25X      4022
35X      3880
33X      3849
03X      3819
32X      3711
27X      3703
09X      3630
21Y      3435
32Y      3125
22X      3061
26X      2978
02Y      2970
10X      2967
14X      2733
03Y      2726
22Y      2664
12Y      2651
05X      2633
02X      2614
31X      2603
21X      2593
17Y      2582
24Y      2575
13Z      2546
15X      2509
24X      2459
12X      2422
10Y      2383
01X      2210
28X      2191
17X      2133
11X      2087
13Y      2017
35Y      1956
31Z      1870
18Y      1778
16Y      1561
14Y      1492
25Y      1482
13X      1122
```

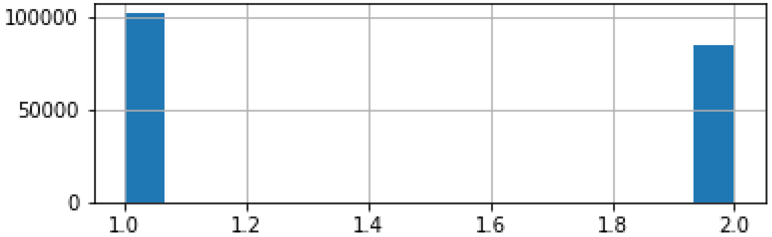
```
18X      1063
16X       994
05Y       710
PDT2       20
Name: Beat, dtype: int64
```

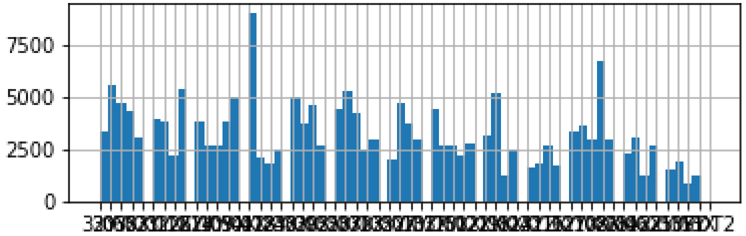




```
1.0     102468
2.0      84963
Name: Area Id, dtype: int64
==============================
04X      9072
08X      6691
30Y      5529
26Y      5374
23X      5301
19X      5158
30X      4988
34X      4965
20X      4682
06X      4676
29X      4606
25X      4396
03X      4380
35X      4291
07X      4235
31Y      3975
09X      3845
32X      3836
21Y      3822
27Y      3701
```

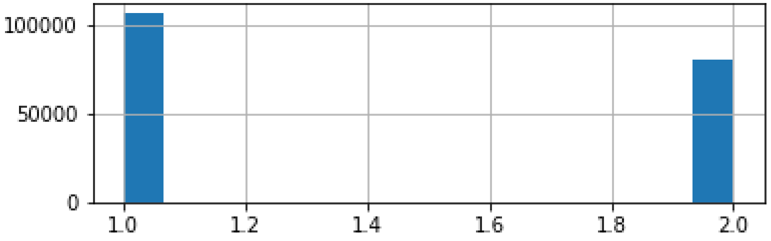
```
33X     3697
27X     3685
12Y     3344
32Y     3328
22X     3131
14X     3070
02Y     3043
03Y     3009
26X     2982
10X     2961
13Z     2946
02X     2798
10Y     2727
22Y     2725
24Y     2723
05X     2681
21X     2674
15X     2671
17Y     2635
12X     2491
24X     2483
31X     2482
28X     2321
01X     2193
11X     2165
17X     2127
35Y     1986
13Y     1898
31Z     1849
18Y     1816
16Y     1680
14Y     1578
25Y     1512
18X     1224
13X     1212
16X     1197
05Y      836
PDT2      28
Name: Beat, dtype: int64
```

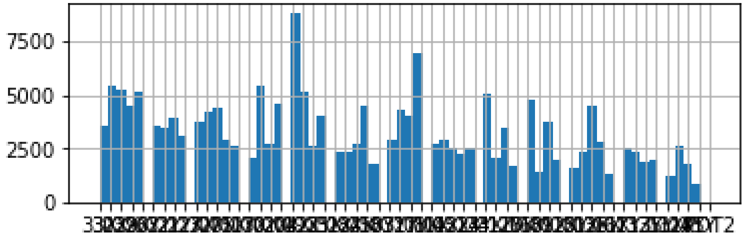




```
1.0    107474
2.0     80578
Name: Area Id, dtype: int64
==============================
04X     8875
08X     6993
30X     5440
30Y     5439
23X     5279
19X     5211
26Y     5188
34X     5059
06X     4786
20X     4565
29X     4531
25X     4530
03X     4483
07X     4416
31Y     4304
32X     4194
35X     4053
27Y     4026
21Y     3938
09X     3776
27X     3774
33X     3537
02Y     3522
12Y     3465
32Y     3465
```

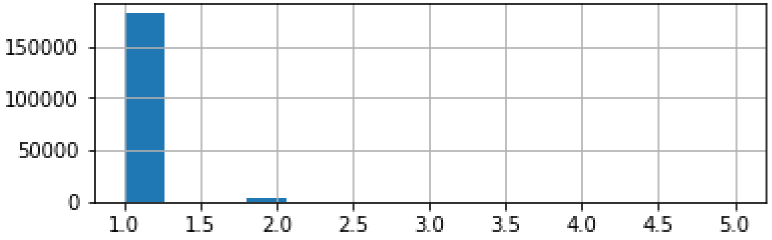
```
22X      3095
03Y      2899
05X      2896
14X      2881
26X      2787
02X      2713
24X      2710
10X      2702
10Y      2641
22Y      2614
12X      2576
24Y      2571
17Y      2564
15X      2482
13Z      2383
31X      2361
01X      2309
28X      2294
21X      2289
17X      2091
31Z      2047
11X      1964
35Y      1950
13Y      1826
18Y      1817
14Y      1794
16Y      1720
25Y      1537
18X      1387
16X      1255
13X      1209
05Y       821
PDT2       18
Name: Beat, dtype: int64
```

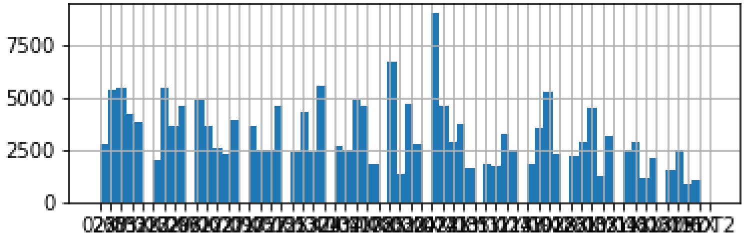




```
1.0    182818
2.0      3898
5.0       320
4.0       236
3.0       208
Name: Area Id, dtype: int64
==============================
04X     9085
08X     6723
30X     5539
23X     5485
30Y     5454
26Y     5377
19X     5290
06X     4931
34X     4865
03X     4727
27Y     4653
29X     4645
20X     4639
07X     4617
31Y     4541
25X     4372
35X     4240
27X     3912
32X     3833
21Y     3784
09X     3625
32Y     3622
02Y     3621
33X     3561
12Y     3214
03Y     3212
14X     2870
26X     2843
24X     2843
02X     2819
22X     2789
```

```
24Y      2673
10X      2566
10Y      2537
12X      2516
21X      2502
31X      2486
17Y      2480
05X      2442
13Z      2415
15X      2347
01X      2320
22Y      2297
28X      2186
11X      2092
31Z      2022
35Y      1860
17X      1860
14Y      1772
13Y      1720
18Y      1609
16Y      1495
25Y      1319
13X      1211
18X      1142
16X      1035
05Y       821
PDT2       24
Name: Beat, dtype: int64
```

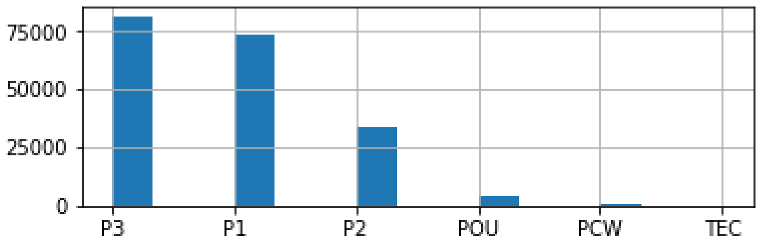


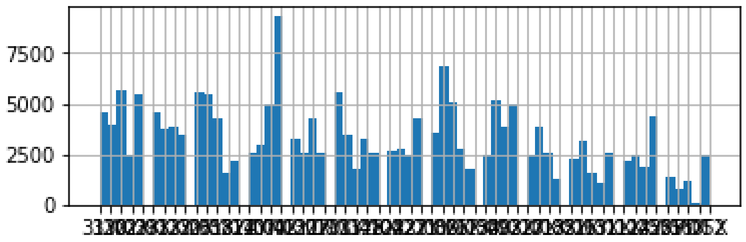


```
P3      81629
```

```
P1     73141
P2     33423
POU     3787
PCW      595
TEC        6
Name: Area Id, dtype: int64
==============================
04X     9373
08X     6874
30Y     5690
19X     5564
30X     5542
23X     5492
26Y     5449
34X     5172
06X     5056
03X     4983
07X     4910
29X     4599
31Y     4556
25X     4409
35X     4287
20X     4284
27Y     4242
32X     3940
27X     3899
12Y     3868
09X     3831
33X     3790
21Y     3574
03Y     3512
32Y     3456
14X     3290
02Y     3290
22X     3207
10Y     2937
26X     2802
24X     2733
10X     2705
28X     2579
24Y     2558
13Z     2555
01X     2552
17Y     2551
31X     2535
12X     2516
02X     2515
```

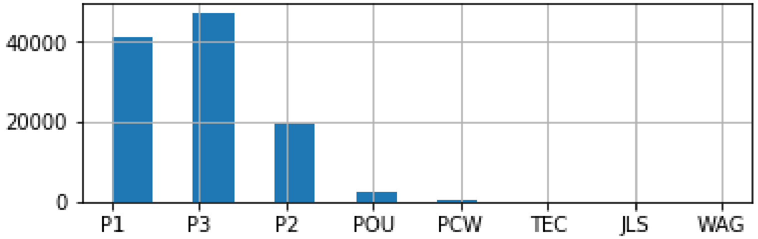
```
21X      2511
05X      2464
22Y      2456
15X      2437
35Y      2293
11X      2186
31Z      2127
14Y      1920
17X      1776
13Y      1734
18Y      1604
16Y      1577
25Y      1406
18X      1263
16X      1223
13X      1117
05Y       775
PDT2       35
Name: Beat, dtype: int64
```

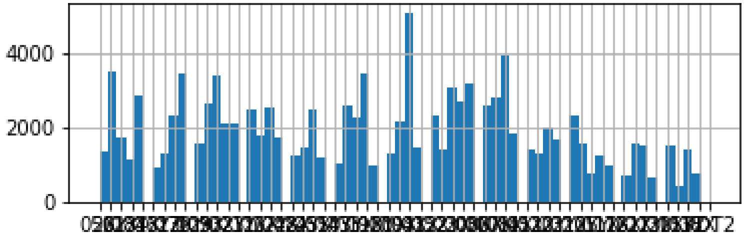




```
P3     47426
P1     41419
P2     19610
POU     2173
PCW      194
TEC        4
JLS        1
WAG        1
Name: Area Id, dtype: int64
==============================
04X        5096
08X        3931
26Y        3511
30Y        3473
19X        3455
30X        3416
03X        3195
23X        3076
34X        2857
07X        2831
20X        2702
29X        2646
06X        2580
03Y        2562
27Y        2517
25X        2467
31Y        2460
27X        2333
35X        2328
32X        2316
33X        2276
09X        2158
21Y        2100
32Y        2093
12Y        1987
14X        1832
26X        1766
02X        1746
24X        1704
02Y        1659
10Y        1573
10X        1557
22X        1541
17Y        1482
21X        1479
24Y        1454
31X        1439
22Y        1420
13Z        1397
15X        1393
05X        1342
01X        1304
12X        1299
31Z        1268
28X        1261
11X        1208
35Y        1159
```

```
18Y      1102
14Y      1027
17X       969
13Y       952
16Y       907
25Y       739
18X       721
16X       708
13X       630
05Y       408
PDT2       16
Name: Beat, dtype: int64
```





用经常发现案件的地方去填充，符合直观感受

通过属性的相关关系来填补缺失值

首先计算相关系数

In [15…

```
for data in data_all:
    x = data.corr()
    print(x)
```

```
           Area Id  Priority
Area Id   1.000000 -0.023366
Priority -0.023366  1.000000
             Area Id  Priority  Zip Codes
Area Id     1.000000 -0.038554   0.023045
Priority   -0.038554  1.000000   0.010370
Zip Codes   0.023045  0.010370   1.000000
           Area Id  Priority
Area Id   1.000000 -0.027769
Priority -0.027769  1.000000
            Area Id  Priority  Zip Codes
Area Id    1.000000 -0.025323        NaN
Priority  -0.025323  1.000000   0.003855
Zip Codes       NaN  0.003855   1.000000
          Priority
Priority       1.0
          Priority
Priority       1.0
```

发现Area Id 和 Priority 的相关关系趋近于0，即基本上不相关，考虑案件位置和案件区域，可能它们之间存在着一些关系，所以我们利用Location 和 Area Id的相关关系来填充缺失值

In [20…

```
loc_area = {}

P = data1.dropna()
loc = P['Location']
area= P['Area Id']
loc = loc.values
area = area.values
for l,a in zip(loc,area):
    loc_area[l] = a

data_1 = data1[['Location','Area Id']]

for i in range(len(data_1)):
    # strr = data_4['Area Id'][i]
    if np.isnan(data_1['Area Id'][i]):
        a = data_1['Location'][i]
        if a in loc_area:
            th = loc_area[a]
            data_1.loc[i, 'Area Id'] = th
print(data_1.isnull()['Area Id'].sum())
```

```
252
```

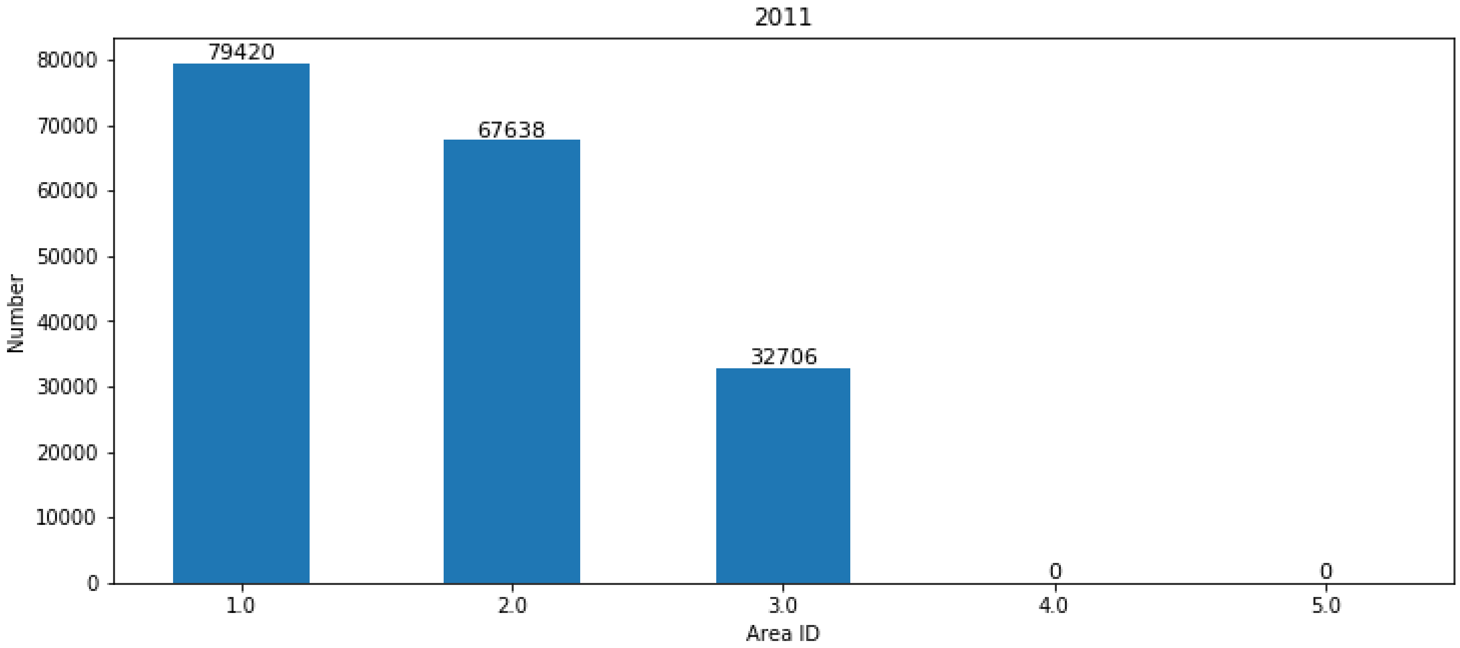
In [22…

```
index = np.arange(5)
lis = data_1['Area Id']
lis = lis.dropna()
lis = lis.values
```

```
Area_count=np.zeros(5)
for t in lis:
    Area_count[int(t)-1] += 1
plt.figure(figsize=(12,5))
plt.bar(index,Area_count, 0.5, label="Area_count")
plt.xticks(index,('1.0','2.0','3.0','4.0','5.0'))
for a,b in zip(index,Area_count):
    plt.text(a, b+0.05, '%.0f' % b, ha='center', va= 'bottom',fontsize=11)
plt.xlabel("Area ID")
plt.ylabel("Number")
plt.title(2011)
```

Out[222]:

```
Text(0.5, 1.0, '2011')
```



相比于之前的直方图，每个区域的案件数量都有所增长，这说明根据案件位置Location，成功的填充了一些缺失数据，但是仍然有252个没有填充上，这可能需要根据其他的csv来进行填充，或者案件位置过于特殊，还没有区域ID。

通过数据对象之间的相似性来填补缺失值

根据巡逻区域的相似性计算填补缺失值

In [22…

```
Beat_area = {}
```

```
P = data1.dropna()
beat = P[’Beat’]
area= P[’Area Id’]
beat = beat.values
area = area.values
for b,a in zip(beat,area):
    Beat_area[l] = a

data_1 = data1[[’Beat’,’Area Id’]]

for i in range(len(data_1)):
    # strr = data_4[’Area Id’][i]
    if np.isnan(data_1[’Area Id’][i]):
        a = data_1[’Beat’][i]
        if a in Beat_area:
            th = Beat_area[a]
            data_1.loc[i, ’Area Id’] = th
print(data_1.isnull()[’Area Id’].sum())
```

```
904
```

发现数据缺失量没有改变，没有填充上，这说明此方法无效