# 作业1: 数据探索性分析与数据预处理

322010840 刘聪聪

## 1. 问题描述

自行选择2个数据集进行探索性分析与预处理。

所选数据集：

oakland-crime-statistics-2011-to-2016

## 2.数据集

chicago-building-violations

2011-2016年的奥克兰犯罪统计数据

从2011-2016年，一共6个csv文件

records-for-2011.csv

records-for-2012.csv

records-for-2013.csv

records-for-2014.csv

records-for-2015.csv

records-for-2016.csv

导入库

In [3]:

```
import matplotlib
import numpy as np
import pandas as pd
%matplotlib inline
```

载入数据

In [4]:

```
path_list = []
for i in range(2011,2017):
    path_list.append("data/oakland-crime-statistics-2011-to-2016/records-for-%s.csv"%(str(i)))
data = []
for path in path_list:
    data.append(pd.read_csv(path ))
```

数据的属性

In [5]:

```
print(data[0].dtypes)
```

```
Agency                        object
Create Time                   object
Location                      object
Area Id                      float64
Beat                          object
Priority                     float64
Incident Type Id              object
Incident Type Description     object
Event Number                  object
Closed Time                   object
dtype: object
```

- Agency 机构
- Create Time 建立时间
- Location 位置
- Area Id 区域ID
- Beat
- Priority 优先级
- Incident Type Id 事件类型ID
- Incident Type Description 事件类型描述
- Event Number 事件号
- Closed Time 结束时间

# 3. 数据分析

## 3.1 数据可视化与摘要

## - 可视化分析与摘要 records-for-2011.csv

In [6]:

```
tmp_data = data[0]
```

统计Agency属性每个可能聚会的频数:

In [7]:

```
print(tmp_data.Agency.value_counts(dropna = False))
```

```
OP     180015
NaN         1
Name: Agency, dtype: int64
```

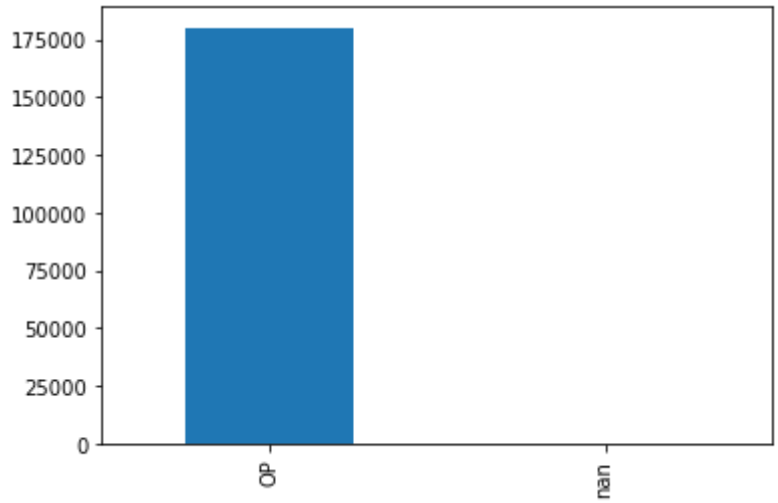
Agency属性数据分布：

In [8]:

```
tmp_data.Agency.value_counts(dropna = False).plot(kind='bar')
```

Out[8]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x195182b0>
```




统计Area Id属性每个可能聚会的频数:

In [9]:

```
attri = "Area Id"
print(tmp_data[attri].value_counts(dropna = False))
```

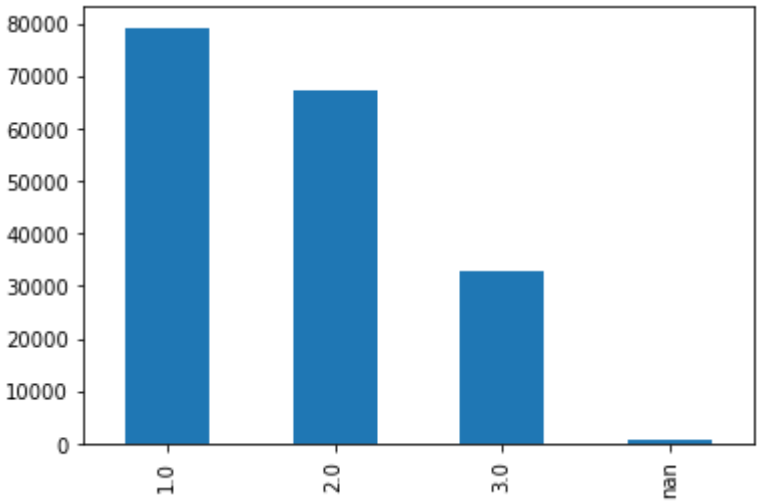
```
1.0    79152
2.0    67261
3.0    32699
NaN      904
Name: Area Id, dtype: int64
```

In [10]:

```
tmp_data[attri].value_counts(dropna = False).plot(kind='bar')
```

Out[10]:

<matplotlib.axes._subplots.AxesSubplot at 0x1956bd60>




统计Beat属性每个可能聚会的频数:

In [11]:

```
attri = "Beat"
print(tmp_data[attri].value_counts(dropna = False))
```

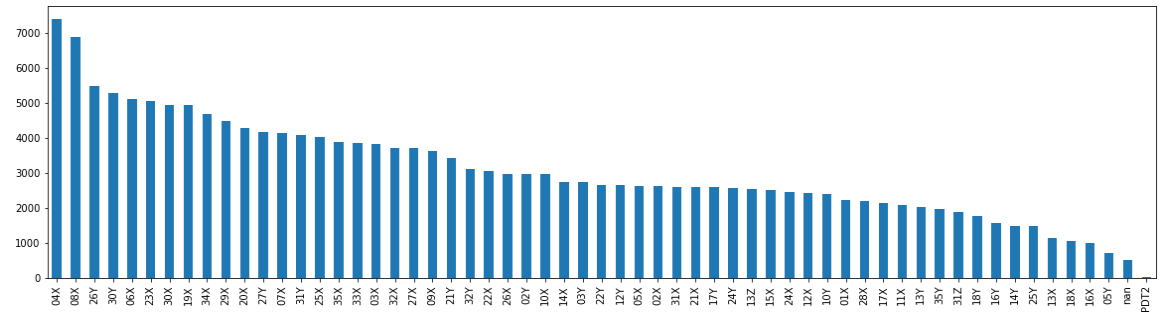
```
04X     7410
08X     6885
26Y     5478
30Y     5295
06X     5119
23X     5051
30X     4956
19X     4955
34X     4673
29X     4483
20X     4287
27Y     4159
07X     4134
31Y     4082
25X     4022
35X     3880
33X     3849
03X     3819
32X     3711
27X     3703
09X     3630
21Y     3435
32Y     3125
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
NaN      520
PDT2      20
Name: Beat, dtype: int64
```

In [12]:

```
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(20,5))
```

Out[12]:

<matplotlib.axes._subplots.AxesSubplot at 0x1c1f6898>



统计Priority属性每个可能聚会的频数:

In [13]:

```
attri = "Priority"
print(tmp_data[attri].value_counts(dropna = False))
```

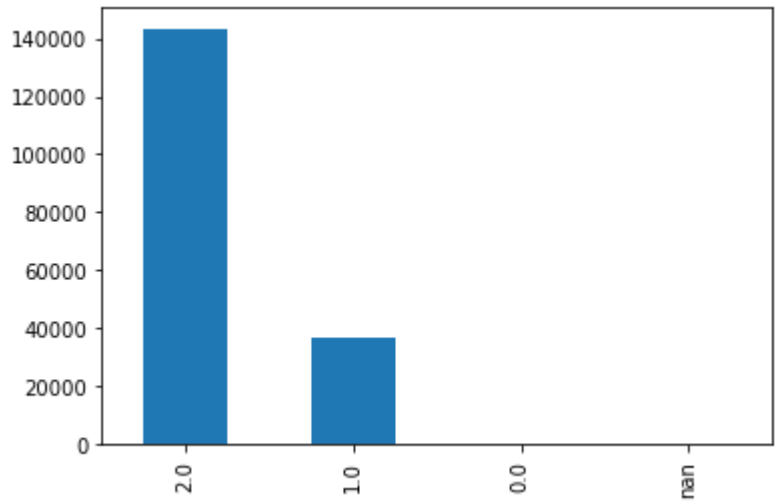
```
2.0    143314
1.0     36699
0.0         2
NaN         1
Name: Priority, dtype: int64
```

In [14]:

```
tmp_data[attri].value_counts(dropna = False).plot(kind='bar')
```

Out[14]:

<matplotlib.axes._subplots.AxesSubplot at 0x1c2ca3e8>




统计Incident Type Id属性每个可能聚会的频数:

In [15]:

```
attri = "Incident Type Id"
print(tmp_data[attri].value_counts(dropna = False))
```

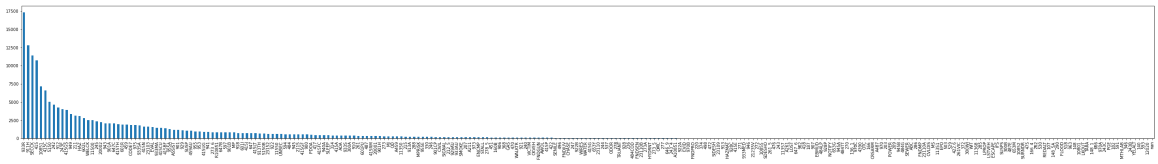
```
933R      17348
911H      12817
SECCK     11393
415       10752
10851      7180
          ...
YELALT        1
140           1
593           1
12020         1
NaN           1
Name: Incident Type Id, Length: 264, dtype: int64
```

In [16]:

```
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(50,6))
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x1ca724d8>
```



统计Priority属性每个可能聚会的频数:

In [17]:

```
attri = "Incident Type Description"
print(tmp_data[attri].value_counts(dropna = False))
```

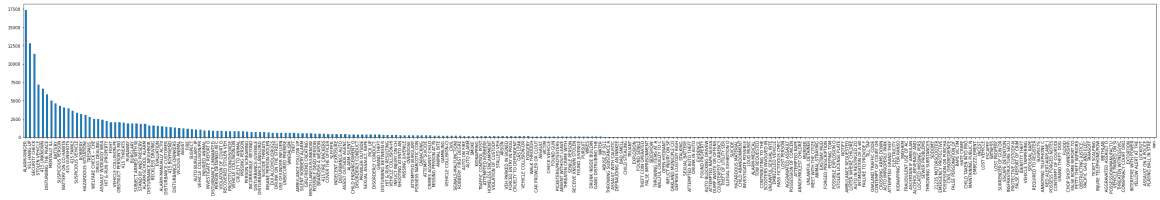
```
ALARM-RINGER              17348
911 HANG-UP               12817
SECURITY CHECK            11393
STOLEN VEHICLE             7180
415 UNKNOWN                6624
                          ...
YELLOW ALERT AT THE           1
LOCKOUT                       1
ASSAULT ON A POLICE           1
PLAYING BALL IN STRE          1
NaN                           1
Name: Incident Type Description, Length: 266, dtype: int64
```

In [18]:

```
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(50,6))
```

Out[18]:

<matplotlib.axes._subplots.AxesSubplot at 0x1cd14028>



## - 可视化分析与摘要 其它年份的csv文件

In [19]:

```
def analysis():
    #matplotlib.pyplot.subplot(2,2,1)
    attri = "Agency"
    print(tmp_data[attri].value_counts(dropna = False))
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar')
    print(" ")

    attri = "Area Id"
    print(tmp_data[attri].value_counts(dropna = False))
    #matplotlib.pyplot.subplot(2,2,2)
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar')
    print(" ")

    attri = "Beat"
    print(tmp_data[attri].value_counts(dropna = False))
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figuresize=(20,5))
    print(" ")

    attri = "Priority"
    print(tmp_data[attri].value_counts(dropna = False))
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar')
    print(" ")

    attri = "Incident Type Id"
    print(tmp_data[attri].value_counts(dropna = False))
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(50,6))
    print(" ")

    attri = "Incident Type Description"
    print(tmp_data[attri].value_counts(dropna = False))
    #tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(50,6))
    print(" ")

#     print(tmp_data.isnull().sum(axis=0))

for i in range(1,5):
    print("### records-for-201%d.csv"%(i+1))
    print("")
    tmp_data = data[i]
    analysis()
```

### records-for-2012.csv

```
OP     187430
NaN         1
Name: Agency, dtype: int64

1.0    101053
2.0     84963
NaN      1415
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
NaN      984
05Y      836
PDT2      28
Name: Beat, dtype: int64

2.0    145504
1.0     41926
NaN         1
Name: Priority, dtype: int64

933R      17216
SECCK     11488
415       11158
911H      10585
10851      8208
          ...
243A          1
VINVER        1
530           1
12020         1
NaN           1
Name: Incident Type Id, Length: 257, 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
VIN VERIFICATION            1
EMBEZZLEMENT BY AN E        1
NaN                         1
Name: Incident Type Description, Length: 259, dtype: int64

### records-for-2013.csv

OP     188051
NaN         1
Name: Agency, dtype: int64

1.0    105216
2.0     80578
NaN      2258
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
NaN     1178
05Y      821
PDT2      18
Name: Beat, dtype: int64

2.0    144859
1.0     43171
0.0        21
NaN         1
Name: Priority, dtype: int64

933R      17859
SECCK     12240
```

```
415       11313
10851      9469
911H       8268
          ...
243B          1
148_5A        1
626_1         1
290           1
NaN           1
Name: Incident Type Id, Length: 254, dtype: int64

ALARM-RINGER            17859
SECURITY CHECK          12240
STOLEN VEHICLE           9469
911 HANG-UP              8268
DISTURBING THE PEACE     6553
                        ...
POSSESS WEAPON AT SC        1
INSFRASTRUCTURE SECU        1
YELLOW ALERT AT THE         1
RAPE VICTIM                 1
IDENTITY THEFT              1
Name: Incident Type Description, Length: 255, dtype: int64

### records-for-2014.csv

OP    187480
Name: Agency, dtype: int64

NaN    177787
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
24X     2843
26X     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
NaN     1217
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
          ...
524           1
484E          1
YELALT        1
A487          1
OTC           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
                        ...
OBTAIN MONEY BY FALS        1
```

```
INTERFERE WITH POWER        1
OAKLAND TRAFFIC CODE        1
ATTEMPTED EXTORTION         1
YELLOW ALERT AT THE         1
Name: Incident Type Description, Length: 258, dtype: int64
```

### records-for-2015.csv

```
OP    192581
Name: Agency, 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
NaN     1325
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
         ...
484E         1
VICE         1
PHONE        1
626_6        1
243B         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
                        ...
TICKET SCALPING             1
DROWNING                    1
POSSESS FORGED NOTES        1
PHONE RPT                   1
RED ALERT-AIRPLANE I        1
Name: Incident Type Description, Length: 262, dtype: int64
```

## 3.2 处理数据缺失

分别使用下列四种策略对缺失值进行处理:

1. 将缺失部分剔除
2. 用最高频率值来填补缺失值
3. 通过属性的相关关系来填补缺失值
4. 通过数据对象之间的相似性来填补缺失值

## - 缺失数据处理 records-for-2011.csv

属性缺失值统计

In [20]:

```
tmp_data = data[0]
print(tmp_data.isnull().sum(axis=0))
```

```
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
```

### Area Id 属性缺失处理

缺失的原因可能是人为失误

- 将缺失部分剔除

In [21]:

```
attri = "Area Id"
d = tmp_data.dropna(subset=[attri])
d
```

Out[21]:

| | Agency | Create Time | Location | Area Id | Beat | Priority | Incident Type Id | Incident Type Description |
|---|---|---|---|---|---|---|---|---|
| **0** | OP | 2011-01-01T00:00:00.000 | ST&SAN PABLO AV | 1.0 | 06X | 1.0 | PDOA | POSSIBLE DEAD PERSON |
| **1** | OP | 2011-01-01T00:01:11.000 | ST&HANNAH ST | 1.0 | 07X | 1.0 | 415GS | 415 GUNSHOTS |
| **2** | OP | 2011-01-01T00:01:25.000 | ST&MARKET ST | 1.0 | 10Y | 2.0 | 415GS | 415 GUNSHOTS |
| **3** | OP | 2011-01-01T00:01:35.000 | PRENTISS ST | 2.0 | 21Y | 2.0 | 415GS | 415 GUNSHOTS |
| **4** | OP | 2011-01-01T00:02:10.000 | AV&FOOTHILL BLVD | 2.0 | 20X | 1.0 | 415GS | 415 GUNSHOTS |
| **...** | ... | ... | ... | ... | ... | ... | ... | .. |
| **180010** | OP | 2011-12-31T23:51:14.000 | WB REDWOOD RD&REDWOOD RD | 1.0 | 22Y | 2.0 | SECCK | SECURITY CHECK |
| **180011** | OP | 2011-12-31T23:52:15.000 | WB 26TH AV | 2.0 | 21X | 2.0 | 415GS | 415 GUNSHOTS |
| **180012** | OP | 2011-12-31T23:53:59.000 | WERNER 40TH STREET WY | 1.0 | 09X | 1.0 | PDOA | POSSIBLE DEAD PERSON |
| **180013** | OP | 2011-12-31T23:55:52.000 | WOOD AV&KEMPTON AV | 1.0 | 08X | 1.0 | 918 | PERSON SCREAMING |
| **180014** | OP | 2011-12-31T23:58:08.000 | WOODSON B ST | 2.0 | 33X | 2.0 | 415GS | 415 GUNSHOTS |

179112 rows × 10 columns

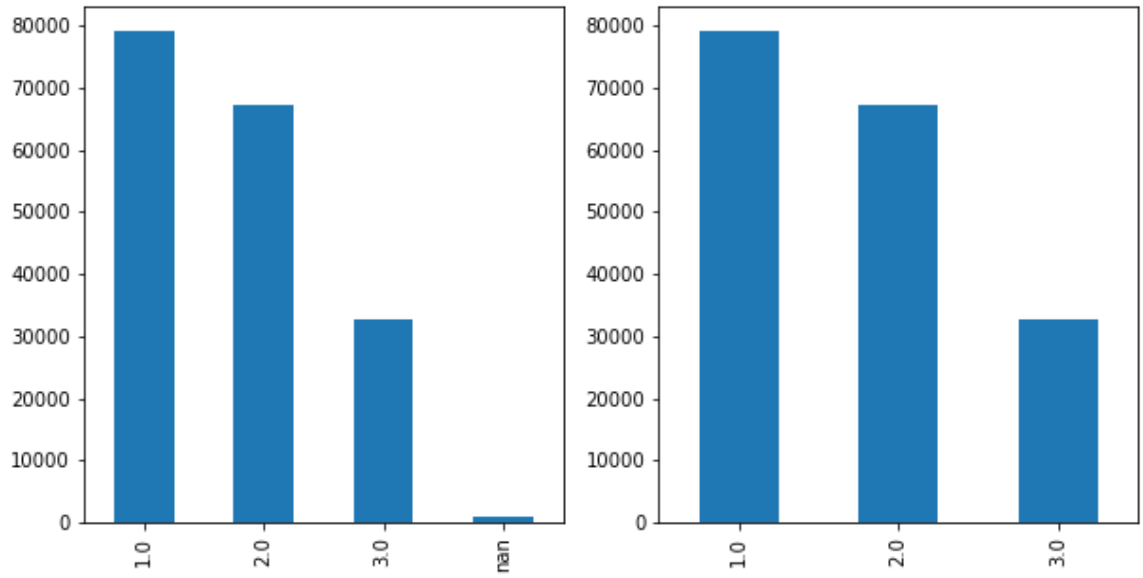
可视化对比

In [22]:

```
matplotlib.pyplot.subplot(1,2,1)
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(10,5))
matplotlib.pyplot.subplot(1,2,2)
d[attri].value_counts(dropna = False).plot(kind='bar')
```

Out[22]:

<matplotlib.axes._subplots.AxesSubplot at 0x1efbd868>




- 用最高频率值来填补缺失值

In [23]:

```
mode = tmp_data[attri].mode()
#int(mode)
f = tmp_data[attri].fillna(int(mode))
print(f)
f.value_counts(dropna = False)
```

```
0         1.0
1         1.0
2         1.0
3         2.0
4         2.0
         ...
180011    2.0
180012    1.0
180013    1.0
180014    2.0
180015    1.0
Name: Area Id, Length: 180016, dtype: float64
```

Out[23]:

```
1.0    80056
2.0    67261
3.0    32699
Name: Area Id, dtype: int64
```

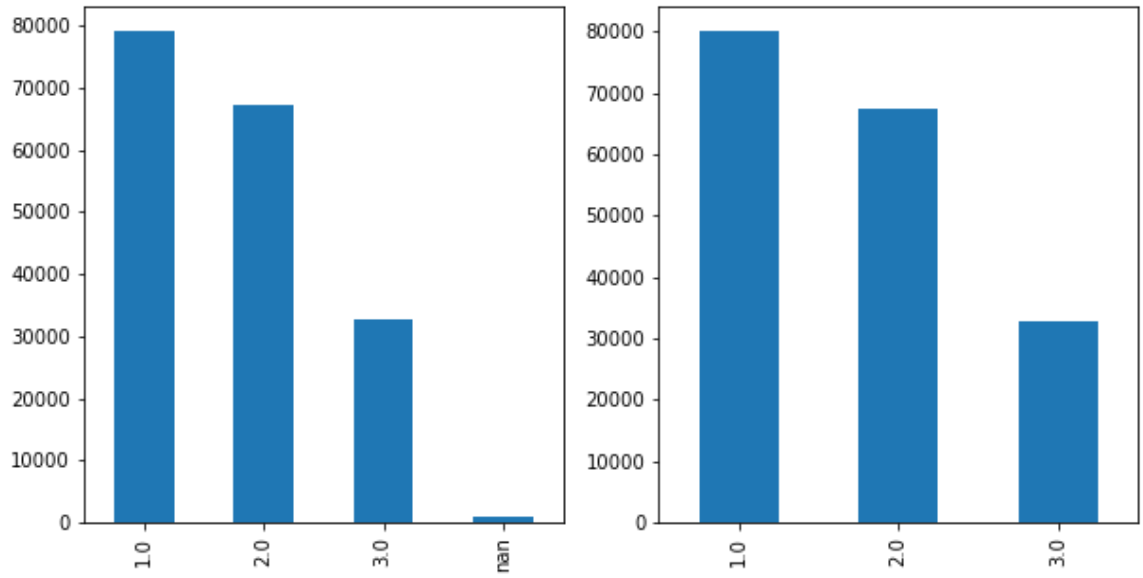
可视化对比

In [24]:

```
matplotlib.pyplot.subplot(1,2,1)
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(10,5))
matplotlib.pyplot.subplot(1,2,2)
f.value_counts(dropna = False).plot(kind='bar',figsize=(10,5))
```

Out[24]:

<matplotlib.axes._subplots.AxesSubplot at 0x1f030d30>




**Beat属性缺失处理**

可能是人为失误导致缺失值

- 将缺失部分剔除

In [25]:

```
attri = "Beat"
d = tmp_data.dropna(subset=[attri])
d
```

Out[25]:

| | Agency | Create Time | Location | Area Id | Beat | Priority | Incident Type Id | Incident Type Description |
|---|---|---|---|---|---|---|---|---|
| **0** | OP | 2011-01-01T00:00:00.000 | ST&SAN PABLO AV | 1.0 | 06X | 1.0 | PDOA | POSSIBLE DEAD PERSON |
| **1** | OP | 2011-01-01T00:01:11.000 | ST&HANNAH ST | 1.0 | 07X | 1.0 | 415GS | 415 GUNSHOTS |
| **2** | OP | 2011-01-01T00:01:25.000 | ST&MARKET ST | 1.0 | 10Y | 2.0 | 415GS | 415 GUNSHOTS |
| **3** | OP | 2011-01-01T00:01:35.000 | PRENTISS ST | 2.0 | 21Y | 2.0 | 415GS | 415 GUNSHOTS |
| **4** | OP | 2011-01-01T00:02:10.000 | AV&FOOTHILL BLVD | 2.0 | 20X | 1.0 | 415GS | 415 GUNSHOTS |
| **...** | ... | ... | ... | ... | ... | ... | ... | .. |
| **180010** | OP | 2011-12-31T23:51:14.000 | WB REDWOOD RD&REDWOOD RD | 1.0 | 22Y | 2.0 | SECCK | SECURITY CHECK |
| **180011** | OP | 2011-12-31T23:52:15.000 | WB 26TH AV | 2.0 | 21X | 2.0 | 415GS | 415 GUNSHOTS |
| **180012** | OP | 2011-12-31T23:53:59.000 | WERNER 40TH STREET WY | 1.0 | 09X | 1.0 | PDOA | POSSIBLE DEAD PERSON |
| **180013** | OP | 2011-12-31T23:55:52.000 | WOOD AV&KEMPTON AV | 1.0 | 08X | 1.0 | 918 | PERSON SCREAMING |
| **180014** | OP | 2011-12-31T23:58:08.000 | WOODSON B ST | 2.0 | 33X | 2.0 | 415GS | 415 GUNSHOTS |

179496 rows × 10 columns

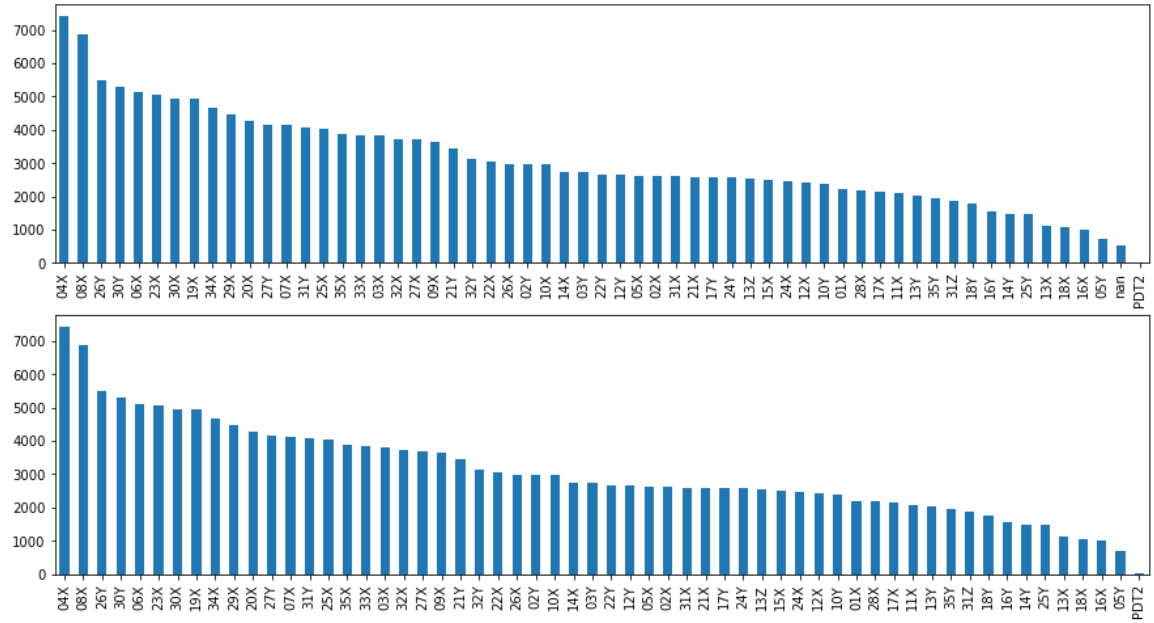
可视化对比

In [26]:

```
matplotlib.pyplot.subplot(2,1,1)
tmp_data[attri].value_counts(dropna = False).plot(kind='bar',figsize=(15,8))
matplotlib.pyplot.subplot(2,1,2)
d[attri].value_counts(dropna = False).plot(kind='bar',figsize=(15,8))
```

Out[26]:

<matplotlib.axes._subplots.AxesSubplot at 0x19556fa0>




## 其它csv文件的缺失值处理

In [27]:

```
def nanHander():
    #剔除包含缺失值的行
    d = tmp_data.dropna(how='any')
    #输出
    print(d)
```

- records-for-2012.csv

In [28]:

```
tmp_data = data[1]
nanHander()
```

```
       Agency              Create Time  Area Id Beat  Priority  \
187256     OP  2012-01-11T13:17:59.000      2.0  28X       2.0
187257     OP  2012-01-01T07:27:25.000      2.0  23X       1.0
187258     OP  2012-01-05T07:21:49.000      1.0  12X       2.0
187260     OP  2012-01-14T18:34:28.000      2.0  33X       2.0
187261     OP  2012-02-15T05:20:20.000      1.0  06X       2.0
...       ...                      ...      ...  ...       ...
187421     OP  2012-12-22T15:08:15.000      1.0  12Y       2.0
187424     OP  2012-11-26T16:55:33.000      2.0  30X       2.0
187426     OP  2012-11-26T22:54:15.000      1.0  25X       1.0
187427     OP  2012-11-29T04:04:54.000      1.0  08X       2.0
187429     OP  2012-11-29T04:41:24.000      2.0  31X       2.0

       Incident Type Id Incident Type Description     Event Number  \
187256              901      VEHICLE COLLISION/PR  LOP120111000450
187257              245      ASSAULT W/DEADLY WEA  LOP120101000403
187258              949        SUSPICIOUS VEHICLE  LOP120105000169
187260             415C               415 UNKNOWN  LOP120114000698
187261              374      DUMP WASTE OR OFFENS  LOP120215000154
...                 ...                      ...              ...
187421             933R              ALARM-RINGER  LOP121222000598
187424             933R              ALARM-RINGER  LOP121126000755
187426              211                   ROBBERY  LOP121126001022
187427            SECCK            SECURITY CHECK  LOP121129000093
187429            SECCK            SECURITY CHECK  LOP121129000099

                  Closed Time  \
187256  2012-01-11T15:47:30.000
187257  2012-01-01T16:02:00.000
187258  2012-01-05T10:26:36.000
187260  2012-01-14T20:25:28.000
187261  2012-02-15T06:15:32.000
...                         ...
187421  2012-12-22T16:32:21.000
187424  2012-11-26T17:12:11.000
187426  2012-11-26T23:26:17.000
187427  2012-11-29T04:44:04.000
187429  2012-11-29T05:39:57.000

                                              Location 1  Zip Codes
187256  {'latitude': '40.874257', 'longitude': '-74.06...     2743.0
187257  {'latitude': '37.767077', 'longitude': '-122.2...     1493.0
187258  {'latitude': '42.262769', 'longitude': '-86.01...     4380.0
187260  {'latitude': '42.120597', 'longitude': '-85.96...     4366.0
187261  {'latitude': '38.626112', 'longitude': '-90.15...    14228.0
...                                                   ...        ...
187421  {'latitude': '38.982036', 'longitude': '-74.94...    14325.0
187424  {'latitude': '42.785336', 'longitude': '-85.97...     1218.0
187426  {'latitude': '37.82844', 'longitude': '-122.27...     1485.0
187427  {'latitude': '39.442458', 'longitude': '-74.50...    14319.0
187429  {'latitude': '26.945458', 'longitude': '-82.35...     3687.0

[123 rows x 11 columns]
```

- records-for-2013.csv

In [29]:

```
tmp_data = data[2]
nanHander()
```

```
       Agency             Create Time                       Location   \
0          OP  2013-01-01T00:00:00.000                         D ST
1          OP  2013-01-01T00:00:05.000                    ARTHUR ST
2          OP  2013-01-01T00:00:50.000                    BRIDGE AV
3          OP  2013-01-01T00:02:16.000              AV&BROOKDALE AV
4          OP  2013-01-01T00:02:47.000             AV&SAN LEANDRO ST
...       ...                      ...                           ...
188046     OP  2013-12-31T23:51:08.000             WHITNEY E 16TH ST
188047     OP  2013-12-31T23:51:59.000  WHITTLE BLVD&HARRINGTON AV
188048     OP  2013-12-31T23:52:03.000   WHITTLE AV&MACARTHUR BLVD
188049     OP  2013-12-31T23:54:20.000       WOOLSLEY, RD&MAKIN RD
188050     OP  2013-12-31T23:58:32.000           WYMAN AV&OLIVE ST

        Area Id Beat  Priority Incident Type Id Incident Type Description  \
0           2.0  33X       1.0             415GS               415 GUNSHOTS
1           2.0  30X       2.0             415GS               415 GUNSHOTS
2           2.0  23X       1.0              243E      BATTERY ON CO-HABITA
3           2.0  29X       2.0             415GS               415 GUNSHOTS
4           2.0  26Y       2.0             415GS               415 GUNSHOTS
...         ...  ...       ...               ...                        ...
188046      2.0  23X       1.0             CODE7      SUBJECT ARMED WITH W
188047      2.0  24X       2.0             415GS               415 GUNSHOTS
188048      2.0  28X       2.0             415GS               415 GUNSHOTS
188049      2.0  31Y       2.0             415GS               415 GUNSHOTS
188050      2.0  34X       2.0             415GS               415 GUNSHOTS

           Event Number              Closed Time
0       LOP130101000002  2013-01-01T00:47:51.000
1       LOP130101000004  2013-01-01T01:30:58.000
2       LOP130101000003  2013-01-01T05:05:50.000
3       LOP130101000005  2013-01-01T01:37:27.000
4       LOP130101000006  2013-01-01T01:33:11.000
...                 ...                      ...
188046  LOP131231000929  2014-01-01T00:24:23.000
188047  LOP131231000930  2014-01-01T00:45:40.000
188048  LOP131231000931  2014-01-01T02:29:02.000
188049  LOP131231000934  2014-01-01T06:03:18.000
188050  LOP131231000935  2014-01-01T03:29:02.000

[185005 rows x 10 columns]
```

- records-for-2014.csv

In [30]:

```
tmp_data = data[3]
d = tmp_data.dropna(subset = ['Agency','Beat','Priority','Event Number','Closed Time','Area Id'
])
print(d)
```

```
     Agency              Create Time  Area Id Beat  Priority Incident Type Id  \
0        OP  2014-01-01T00:00:00.000      1.0  02X         2             415GS
1        OP  2014-01-01T00:00:00.000      2.0  26Y         2             415GS
2        OP  2014-01-01T00:00:00.000      2.0  30Y         2             415GS
3        OP  2014-01-01T00:00:00.000      2.0  30Y         2             415GS
4        OP  2014-01-01T00:01:04.000      2.0  35X         2             CODE7
...     ...                     ...      ...  ...       ...               ...
9821     OP  2014-01-21T04:22:39.000      1.0  04X         2              EVAL
9822     OP  2014-01-21T04:23:14.000      4.0  27X         2              933R
9823     OP  2014-01-21T04:42:14.000      2.0  08X         2               912
9824     OP  2014-01-21T04:56:33.000      1.0  06X         2               314
9825     OP  2014-01-21T05:16:22.000      2.0  34X         2             SECCK

     Incident Type Description     Event Number              Closed Time  \
0                  415 GUNSHOTS  LOP140101000001  2014-01-01T03:22:08.000
1                  415 GUNSHOTS  LOP140101000002  2014-01-01T02:56:31.000
2                  415 GUNSHOTS  LOP140101000004  2014-01-01T00:49:53.000
3                  415 GUNSHOTS  LOP140101000005  2014-01-01T02:51:11.000
4           SUBJECT ARMED WITH W  LOP140101000010  2014-01-01T05:33:22.000
...                          ...              ...                      ...
9821                 EVALUATION  LOP140121000109  2014-01-21T05:08:52.000
9822               ALARM-RINGER  LOP140121000108  2014-01-21T04:58:11.000
9823           SUSPICIOUS PERSON  LOP140121000111  2014-01-21T05:03:36.000
9824           INDECENT EXPOSURE  LOP140121000113  2014-01-21T05:30:02.000
9825              SECURITY CHECK  LOP140121000116  2014-01-21T05:45:15.000

                                          Location 1  Zip Codes
0     {'human_address': '{"address": "LINDEN ST", "c...        NaN
1     {'human_address': '{"address": "AV&INTERNATION...        NaN
2     {'human_address': '{"address": "AV&MACARTHUR B...        NaN
3     {'human_address': '{"address": "MACARTHUR BLVD...        NaN
4     {'human_address': '{"address": "AV&DOWLING ST"...        NaN
...                                                 ...        ...
9821  {'human_address': '{"address": "W GRAND AV", "...        NaN
9822  {'human_address': '{"address": "HIGH ST", "cit...        NaN
9823  {'human_address': '{"address": "BROADWAY", "ci...        NaN
9824  {'human_address': '{"address": "ST&WEST ST", "...        NaN
9825  {'human_address': '{"address": "AV&BANCROFT AV...        NaN

[9637 rows x 11 columns]
```

- records-for-2015.csv

In [31]:

```
tmp_data = data[4]
nanHander()
```

```
       Agency              Create Time                      Location Area Id  \
0          OP  2015-01-01T00:01:59.000               S ELMHURST AV         P3
1          OP  2015-01-01T00:02:02.000                    AV&D ST          P3
2          OP  2015-01-01T00:02:06.000                BANCROFT AV          P3
3          OP  2015-01-01T00:03:16.000             MACARTHUR BLVD          P3
4          OP  2015-01-01T00:03:45.000              ST&ADELINE ST          P1
...       ...                      ...                                ...     ...
192576     OP  2015-12-31T23:56:55.000  WHITNEY ST&MOUNTAIN BLVD          P3
192577     OP  2015-12-31T23:57:26.000         WHITTLE CASWELL AV         P3
192578     OP  2015-12-31T23:57:43.000          WILDWOOD E 12TH ST        P3
192579     OP  2015-12-31T23:58:11.000      YERB&HOL AV&HUDSON ST         P1
192580     OP  2015-12-31T23:59:37.000       YERBA ST&SHATTUCK AV         P1

       Beat  Priority Incident Type Id Incident Type Description  \
0       31Y         2              415      DISTURBING THE PEACE
1       32X         2            415GS               415 GUNSHOTS
2       30Y         2             933R               ALARM-RINGER
3       30Y         2            415GS               415 GUNSHOTS
4       02X         2            415GS               415 GUNSHOTS
...     ...       ...              ...                        ...
192576  25X         2            SECCK            SECURITY CHECK
192577  31Y         2            415GS              415 GUNSHOTS
192578  23X         2            SECCK            SECURITY CHECK
192579  12Y         2            SECCK            SECURITY CHECK
192580  11X         2            SECCK            SECURITY CHECK

          Event Number              Closed Time
0       LOP150101000003  2015-01-01T06:23:08.000
1       LOP150101000007  2015-01-01T01:44:40.000
2       LOP150101000004  2015-01-01T02:12:39.000
3       LOP150101000005  2015-01-01T01:53:08.000
4       LOP150101000009  2015-01-01T00:37:09.000
...                 ...                      ...
192576  LOP151231001096  2016-01-01T00:01:15.000
192577  LOP151231001098  2016-01-01T01:15:50.000
192578  LOP151231001100  2016-01-01T00:13:56.000
192579  LOP151231001102  2016-01-01T00:27:40.000
192580  LOP151231001104  2016-01-01T00:29:22.000

[191013 rows x 10 columns]
```

- records-for-2016.csv

In [32]:

```
tmp_data = data[5]
nanHander()
```

```
       Agency              Create Time                Location Area Id Beat  \
0          OP  2016-01-01T00:00:57.000         ST&MARKET ST          P1  05X
1          OP  2016-01-01T00:01:25.000       AV&HAMILTON ST          P3  26Y
2          OP  2016-01-01T00:01:43.000       ST&CHESTNUT ST          P1  02X
3          OP  2016-01-01T00:01:48.000           WALLACE ST          P2  18Y
4          OP  2016-01-01T00:02:05.000              90TH AV          P3  34X
...       ...                      ...                   ...      ...  ...
110822     OP  2016-07-31T23:43:51.000   WENDY'S ST&PINE ST          P1  02Y
110823     OP  2016-07-31T23:45:50.000  WHITMORE ST&WOOD ST          P1  02Y
110824     OP  2016-07-31T23:50:54.000      WHITTLE 69TH AV          P3  26Y
110825     OP  2016-07-31T23:56:29.000    WHITTLE LOOMIS CT          P2  19X
110826     OP  2016-07-31T23:57:31.000       WYMAN LACEY AV          P3  29X

        Priority Incident Type Id Incident Type Description     Event Number  \
0            2.0            415GS              415 GUNSHOTS  LOP160101000003
1            2.0            415GS              415 GUNSHOTS  LOP160101000005
2            2.0            415GS              415 GUNSHOTS  LOP160101000008
3            2.0            415GS              415 GUNSHOTS  LOP160101000007
4            2.0            415GS              415 GUNSHOTS  LOP160101000009
...          ...              ...                       ...              ...
110822       2.0              922       DRUNK ON THE STREET  LOP160731000889
110823       2.0            415GS              415 GUNSHOTS  LOP160731000892
110824       2.0             415N      DISTURBANCE-NEIGHBOR  LOP160731000893
110825       2.0              912         SUSPICIOUS PERSON  LOP160731000895
110826       2.0              415                415 FAMILY  LOP160731000897

                  Closed Time
0       2016-01-01T00:32:30.000
1       2016-01-01T00:48:23.000
2       2016-01-01T00:21:24.000
3       2016-01-01T01:15:03.000
4       2016-01-01T00:54:52.000
...                         ...
110822  2016-08-01T00:36:46.000
110823  2016-07-31T23:58:03.000
110824  2016-08-01T00:08:00.000
110825  2016-08-01T01:33:31.000
110826  2016-08-01T00:16:16.000

[110247 rows x 10 columns]
```