DataPreprocess-Oakland Crime Statistics 2011 to 2016

April 7, 2021

1 Oakland Crime Statistics 2011 to 2016

#3

##3.1 Oakland Crime Statistics 2011 to 201620112016csv2016records-for-2015.csv10AgencyCreate TimeLocationArea IdBeatIncident Type IdIncident Type DescriptionEvent NumberClosed TimePriority 5

```
In [44]: import pandas as pd
         import numpy as np
         from collections import Counter
         import matplotlib.pyplot as plt

         path = 'C:/Users/ZL/Desktop/oakland-crime-statistics-2011-to-2016/records-for-2016.csv
         data = pd.read_csv(path, header=0,  engine='python', encoding='utf-8')
         data = data.values
         print('')
         for i in range(data.shape[1]):  #
             counter = Counter(data[:, i])
             print(counter.most_common(5))  # 5
```

```
[('OP', 110827), (nan, 1)]
[('2016-05-06T11:21:13.000', 3), ('2016-01-01T11:56:04.000', 2), ('2016-01-05T15:14:57.000', 2)
[(' INTERNATIONAL BLVD    ', 2156), (' AV&INTERNATIONAL BLVD    ', 1829), (' MACARTHUR BLVD
[('P3', 47425), ('P1', 41419), ('P2', 19610), ('POU', 2173), ('PCW', 194)]
[('04X', 4515), ('08X', 3931), ('26Y', 3511), ('30Y', 3473), ('19X', 3455)]
[(2.0, 86272), (1.0, 24555), (nan, 1)]
[('933R', 10094), ('415', 7883), ('SECCK', 7251), ('10851', 5308), ('911H', 5089)]
[('ALARM-RINGER', 10094), ('SECURITY CHECK', 7251), ('STOLEN VEHICLE', 5308), ('911 HANG-UP', 5
[('LOP160101000003', 1), ('LOP160101000005', 1), ('LOP160101000008', 1), ('LOP160101000007', 1)
[('2016-05-29T00:43:38.000', 3), ('2016-01-02T20:07:50.000', 2), ('2016-01-03T00:56:37.000', 2)
```

Priority122220

```
In [23]: import pandas as pd
         import numpy as np
```

```
        from collections import Counter
        import matplotlib.pyplot as plt

        path = 'C:/Users/ZL/Desktop/oakland-crime-statistics-2011-to-2016/records-for-2016.csv
        data = pd.read_csv(path, header=0, engine='python', encoding='utf-8')
        print(data.describe())  #
```

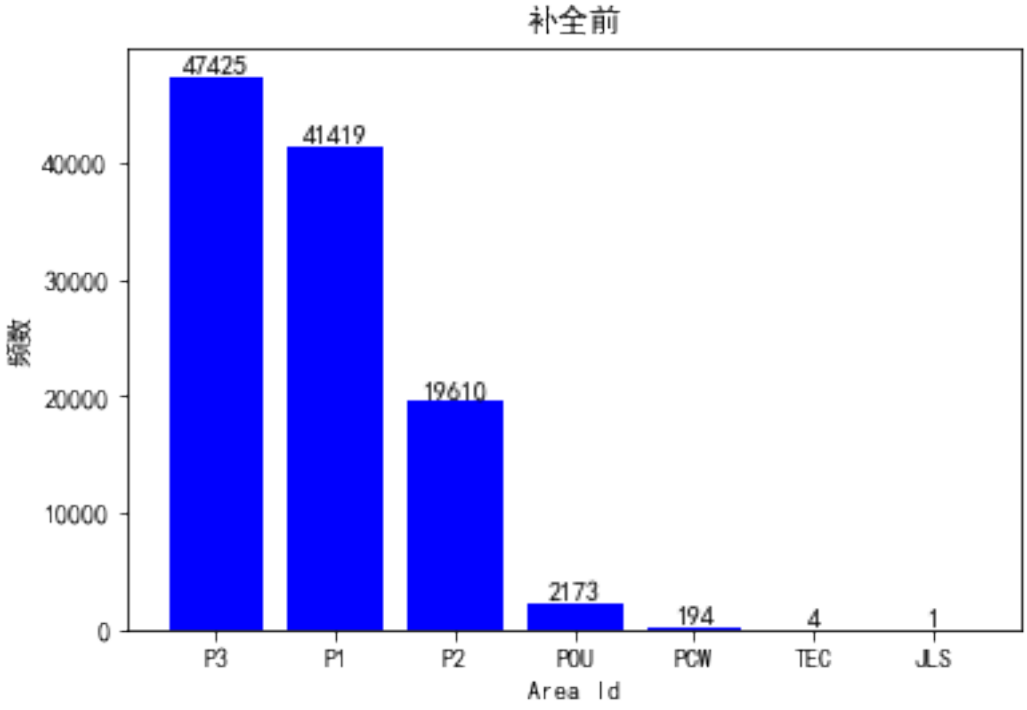
```
            Priority
count  110827.000000
mean        1.778438
std         0.415299
min         1.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
```

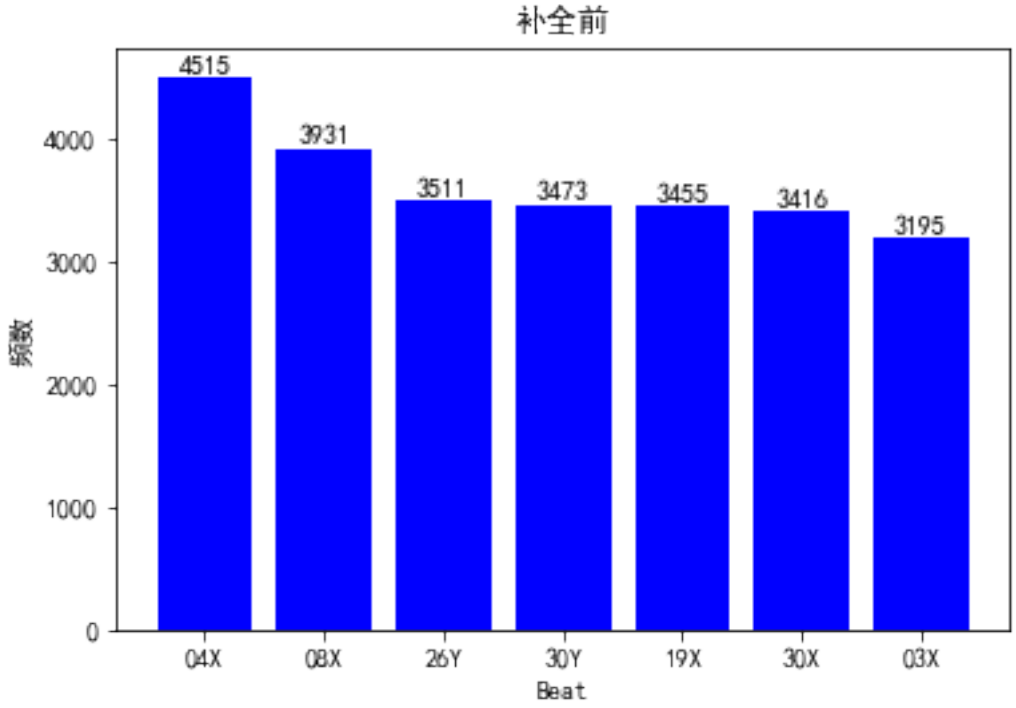
##1.2 Area IdBeat Area IdBeat

```
In [26]: import pandas as pd
         import numpy as np
         from collections import Counter
         import matplotlib.pyplot as plt
         plt.rcParams['font.sans-serif'] = ['SimHei']
         plt.rcParams['axes.unicode_minus'] = False
         path = 'C:/Users/ZL/Desktop/oakland-crime-statistics-2011-to-2016/records-for-2016.csv

         def draw(data,cl,xlabel):
             num = 7
             data = data.values
             counter = Counter(data[:, cl])
             frequency = counter.most_common()  # n
             num_list = []
             name_list = []
             for i in range(num):
                 num_list.append(int(frequency[i][1]))
                 name_list.append(str(frequency[i][0]))
             fig, ax = plt.subplots()
             b = ax.bar(name_list, num_list)
             plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
             for a, b in zip(name_list, num_list):
                 ax.text(a, b + 1, b, ha='center', va='bottom')
             plt.title('')
             plt.xlabel(xlabel)
             plt.ylabel('')
             plt.show()
         data = pd.read_csv(path, header=0, engine='python', encoding='utf-8')
```

```
draw(data, 3, 'Area Id')
draw(data, 4, 'Beat')
```



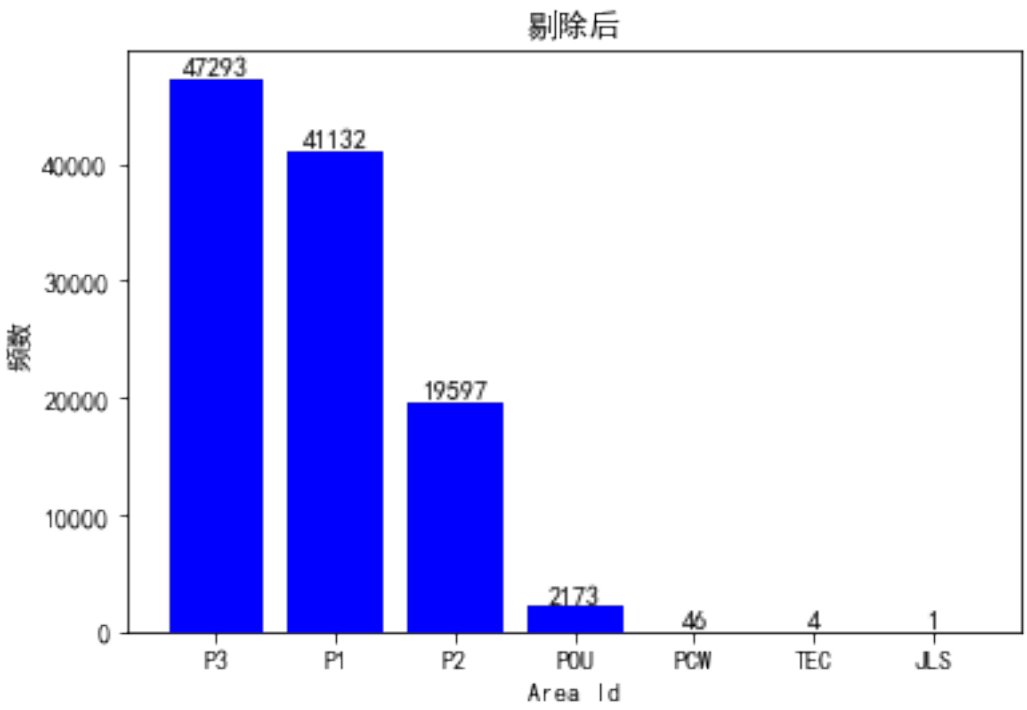


#4 #4.1 39241 Area Id

In [38]:
```
import pandas as pd
import numpy as np
from collections import Counter
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
path = 'C:/Users/ZL/Desktop/oakland-crime-statistics-2011-to-2016/records-for-2016.csv

def draw(data,cl,xlabel):
    num = 7
    data = data.values
    counter = Counter(data[:, cl])
    frequency = counter.most_common()  # n
    num_list = []
    name_list = []
    for i in range(num):
        num_list.append(int(frequency[i][1]))
        name_list.append(str(frequency[i][0]))
    fig, ax = plt.subplots()
    b = ax.bar(name_list, num_list)
    plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
    for a, b in zip(name_list, num_list):
        ax.text(a, b + 1, b, ha='center', va='bottom')
    plt.title('')
    plt.xlabel(xlabel)
    plt.ylabel('')
    plt.show()
data = pd.read_csv(path, header=0,  engine='python', encoding='utf-8')
data_drop = data.dropna()  #
print(''+str(data_drop.shape[0]))
draw(data_drop,3,'Area Id')
```

```
110247
```



##4.2 1010 Area Id150930 Area Id7P3

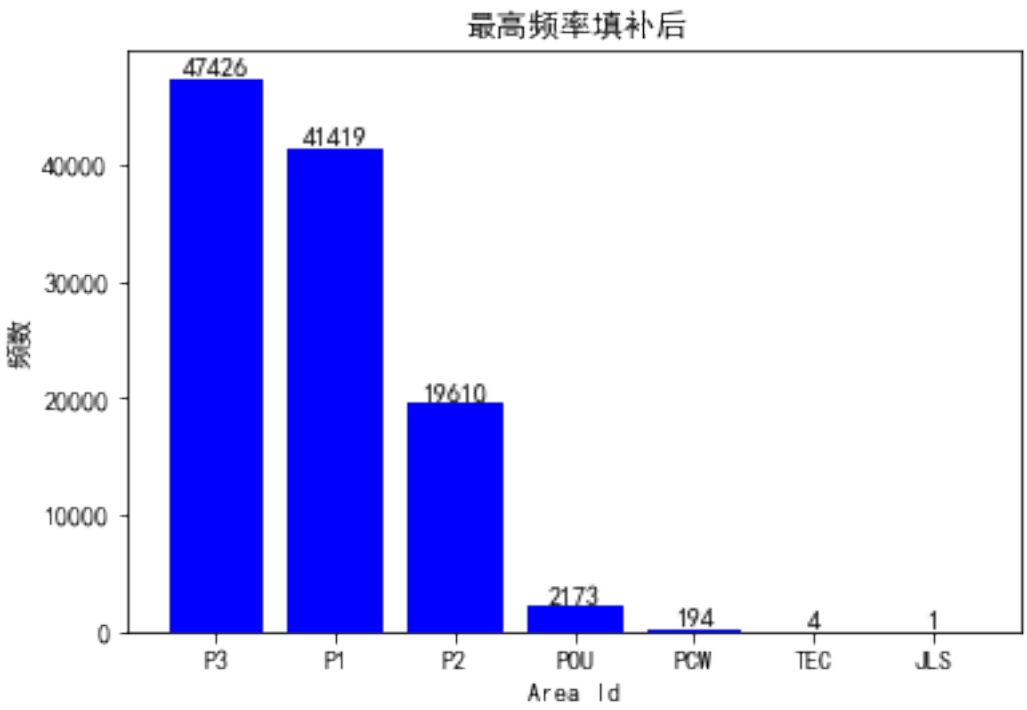
```
In [43]: import pandas as pd
         import numpy as np
         from collections import Counter
         import matplotlib.pyplot as plt
         plt.rcParams['font.sans-serif'] = ['SimHei']
         plt.rcParams['axes.unicode_minus'] = False
         path = 'C:/Users/ZL/Desktop/oakland-crime-statistics-2011-to-2016/records-for-2016.csv

         def draw(data, cl, xlabel):
             num = 7
             data = data.values
             counter = Counter(data[:, cl])
             frequency = counter.most_common()  # n
             num_list = []
             name_list = []
             for i in range(num):
                 num_list.append(int(frequency[i][1]))
                 name_list.append(str(frequency[i][0]))
             fig, ax = plt.subplots()
             b = ax.bar(name_list, num_list)
             plt.bar(range(len(num_list)), num_list, color='blue', tick_label=name_list)
             for a, b in zip(name_list, num_list):
                 ax.text(a, b + 1, b, ha='center', va='bottom')
```

```
    plt.title('')
    plt.xlabel(xlabel)
    plt.ylabel('')
    plt.show()


data = pd.read_csv(path, header=0, engine='python', encoding='utf-8')
data = data.values
max_time = []  #
#
for cl in range(data.shape[1]):
    counter = Counter(data[:, cl])
    counter = counter.most_common()  # listlist
    if counter[0][0] == counter[0][0]:  #
        max_time.append(counter[0][0])
    else:  #
        max_time.append(counter[1][0])
#
data_max = pd.DataFrame(data)
for cl in range(data.shape[1]):
    data_max[cl] = data_max[cl].fillna(max_time[cl])
print(data_max.describe())
draw(data_max,3,'Area Id')
```

```
                5
count  110828.000000
mean        1.778440
std         0.415298
min         1.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
```


最高频率填补后
47426
41419
19610
2173
194
4
1
40000
30000
20000
10000
0
频数
P3
P1
P2
POU
PCW
TEC
JLS
Area Id