In [84]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
from scipy import stats
%matplotlib inline
```

In [23]:

```
data=pd.read_excel('数据集.xlsx')
```

In [24]:

```
data.shape
```

Out[24]:

(16750, 27)

In [25]:

```
data.head(5)
```

Out[25]:

| | 合同金额 | 账龄 | 逾期期数 | 逾期总额 | 实际逾期天数 | 剩余期数 | 产品 | 来源渠道 | 剩余本金 | 逾期本金 | ... | 近三月拨打电话数 | 近一月回款金额 | 近一月回款次数 | 近一月最大回款金额 | 近三月回款金额 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 17000 | M2 | 3 | 4694.29 | 75 | 11 | 1 | b | 15657.3 | 4106.65 | ... | 4 | NaN | NaN | NaN | NaN |
| **1** | 25300 | M2 | 3 | 4403.29 | 75 | 23 | 1 | b | 24359.8 | 2876.21 | ... | 3 | NaN | NaN | NaN | NaN |
| **2** | 18000 | M1 | 2 | 3192.5 | 44 | 10 | 1 | b | 15142.9 | 2912.78 | ... | 4 | NaN | NaN | NaN | 1675.52 |
| **3** | 11500 | M3 | 4 | 4296.64 | 105 | 12 | 1 | b | 11500 | 3686.33 | ... | 4 | NaN | NaN | NaN | NaN |
| **4** | 16900 | C-M1 | 1 | 896.96 | 13 | 21 | 1 | b | 14993.3 | 647.82 | ... | 3 | NaN | NaN | NaN | 1583.52 |

5 rows × 27 columns

In [26]:

```
data.columns
```

Out[26]:

```
Index(['合同金额', '账龄', '逾期期数', '逾期总额', '实际逾期天数', '剩余期数', '产
品', '来源渠道', '剩余本金',
       '逾期本金', '近一月拨打次数', '近一月电话接通次数', '近一月电话拨打天数', '近
一月拨打电话数', '近三月拨打次数',
       '近三月电话接通次数', '近三月电话拨打天数', '近三月拨打电话数', '近一月回款金
额', '近一月回款次数', '近一月最大回款金额',
       '近三月回款金额', '近三月回款次数', '近三月最大回款金额', '下个月回款金额',
'下个月回款次数', '下个月最大回款金额'],
      dtype='object')
```

In [27]:

```
data.describe(include='all').T
```

Out[27]:

| | count | unique | top | freq |
|---|---|---|---|---|
| **合同金额** | 16750 | 628 | 10000 | 2445 |
| **账龄** | 12629 | 38 | C-M1 | 3259 |
| **逾期期数** | 12629 | 27 | 1 | 3524 |
| **逾期总额** | 12629 | 8089 | 0 | 454 |
| **实际逾期天数** | 12629 | 389 | 0 | 393 |
| **剩余期数** | 12629 | 37 | 10 | 1608 |
| **产品** | 12629 | 7 | 1 | 12245 |
| **来源渠道** | 12629 | 2 | b | 12628 |
| **剩余本金** | 12629 | 4631 | 9210.2 | 357 |

In [28]:

```
data.drop(index=10000,inplace=True)
```

In [29]:

```
a=['账龄','产品','来源渠道']
for i in data.columns:
    if i not in a:
        data[i]=data[i].astype(float)
```

In [30]:

```
data
```

Out[30]:

| | 合同金额 | 账龄 | 逾期期数 | 逾期总额 | 实际逾期天数 | 剩余期数 | 产品 | 来源渠道 | 剩余本金 | 逾期本金 | ... | 近三月拨打电话数 | 近一月回款金额 | 近一月回款次数 | 近一月最大回款金额 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 17000.0 | M2 | 3.0 | 4694.29 | 75.0 | 11.0 | 1 | b | 15657.33 | 4106.65 | ... | 4.0 | NaN | NaN | NaN | |
| **1** | 25300.0 | M2 | 3.0 | 4403.29 | 75.0 | 23.0 | 1 | b | 24359.81 | 2876.21 | ... | 3.0 | NaN | NaN | NaN | |
| **2** | 18000.0 | M1 | 2.0 | 3192.50 | 44.0 | 10.0 | 1 | b | 15142.91 | 2912.78 | ... | 4.0 | NaN | NaN | NaN | 16 |
| **3** | 11500.0 | M3 | 4.0 | 4296.64 | 105.0 | 12.0 | 1 | b | 11500.00 | 3686.33 | ... | 4.0 | NaN | NaN | NaN | |
| **4** | 16900.0 | C-M1 | 1.0 | 896.96 | 13.0 | 21.0 | 1 | b | 14993.26 | 647.82 | ... | 3.0 | NaN | NaN | NaN | 15 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |

In [31]:

```
data.duplicated().sum()
```

Out[31]:

0

In [32]:

```
data.info()
```

```
7   来源渠道          12628 non-null  object
8   剩余本金          12628 non-null  float64
9   逾期本金          12628 non-null  float64
10  近一月拨打次数      13231 non-null  float64
11  近一月电话接通次数   13231 non-null  float64
12  近一月电话拨打天数   13231 non-null  float64
13  近一月拨打电话数     13231 non-null  float64
14  近三月拨打次数      15675 non-null  float64
15  近三月电话接通次数   15675 non-null  float64
16  近三月电话拨打天数   15675 non-null  float64
17  近三月拨打电话数     15675 non-null  float64
18  近一月回款金额       6011 non-null   float64
19  近一月回款次数       6011 non-null   float64
20  近一月最大回款金额    6011 non-null   float64
21  近三月回款金额      11181 non-null  float64
22  近三月回款次数      11181 non-null  float64
23  近三月最大回款金额   11181 non-null  float64
24  下个月回款金额       9118 non-null   float64
25  下个月回款次数       9118 non-null   float64
26  下个月最大回款金额    9118 non-null   float64
```

In [33]:

```
data=data.drop(['账龄','来源渠道','产品'],axis=1)
```

In [34]:

```
y_full=data['合同金额'].values
y_full
```

Out[34]:

```
array([17000., 25300., 18000., ..., 15400., 15400., 22000.])
```

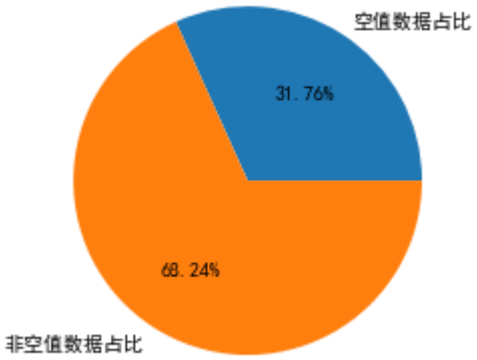
In [35]:

```
data.describe(include='all').T
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 逾期天数 | 12628.0 | 164.839404 | 244.507912 | 0.00 | 19.00 | 51.00 | 159.0000 | 1159.00 |
| 剩余期数 | 12628.0 | 12.726402 | 6.845474 | 1.00 | 8.00 | 11.00 | 19.0000 | 36.00 |
| 剩余本金 | 12628.0 | 12729.448514 | 7649.348904 | 100.00 | 7500.00 | 10702.26 | 17000.0000 | 55900.00 |
| 逾期本金 | 12628.0 | 3999.886188 | 5594.012725 | 0.00 | 1056.92 | 1979.72 | 4218.4100 | 49971.61 |
| 近一一 | | | | | | | | |

In [36]:

```
plt.pie(x = [1002466,2153898],labels = ["空值数据占比","非空值数据占比"],autopct='%.2f%%')
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.show()
```



In [37]:

```
#空值处理
from sklearn.ensemble import RandomForestRegressor
a = data[data[data.columns].isnull().sum(axis=1) > 5]
data.drop(a.index,axis = 0,inplace = True)
```

In [47]:

```
#使用随机森林处理剩下的空值
for index , i in enumerate(data.columns):
    if ((data[i].isnull()).sum()) != 0:
        known = data[data[i].notnull()]
        unknown =data[data[i].isnull()]

        x_train = known.iloc[:,[0,1,2,3,4,5,6,11,12,13,14,18,19,20]]
        y_train = known.iloc[:,index]
        x_test = unknown.iloc[:,[0,1,2,3,4,5,6,11,12,13,14,18,19,20]]

        clf = RandomForestRegressor(n_estimators = 200,random_state=0,max_depth = 15)
        data.loc[data[i].isnull(),i] = clf.fit(x_train,y_train).predict(x_test)
```

In [48]:

```
data.isnull().sum()
```

Out[48]:

```
合同金额          0
逾期期数          0
逾期总额          0
实际逾期天数        0
剩余期数          0
剩余本金          0
逾期本金          0
近一月拨打次数       0
近一月电话接通次数     0
近一月电话拨打天数     0
近一月拨打电话数      0
近三月拨打次数       0
近三月电话接通次数     0
近三月电话拨打天数     0
近三月拨打电话数      0
近一月回款金额       0
近一月回款次数       0
近一月最大回款金额     0
近三月回款金额       0
近三月回款次数       0
近三月最大回款金额     0
下个月回款金额       0
下个月回款次数       0
下个月最大回款金额     0
dtype: int64
```

In [49]:

```
data
```

Out[49]:

| | 合同金额 | 逾期期数 | 逾期总额 | 实际逾期天数 | 剩余期数 | 剩余本金 | 逾期本金 | 近一月拨打次数 | 近一月电话接通次数 | 近一月电话拨打天数 | ... | 近三月拨打电话数 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **14** | 10000.00 | 2.0 | 1776.06 | 42.0 | 10.0 | 8412.74 | 1618.21 | 23.0 | 1.0 | 8.0 | ... | 3.0 | |
| **22** | 15500.00 | 2.0 | 2827.79 | 42.0 | 10.0 | 13039.73 | 2508.23 | 19.0 | 1.0 | 10.0 | ... | 3.0 | |
| **68** | 20000.00 | 2.0 | 1432.00 | 51.0 | 22.0 | 18021.16 | 1038.07 | 16.0 | 0.0 | 8.0 | ... | 4.0 | |
| **101** | 10000.00 | 1.0 | 885.81 | 15.0 | 9.0 | 7607.54 | 813.01 | 13.0 | 1.0 | 7.0 | ... | 3.0 | 9 |
| **124** | 7500.00 | 2.0 | 1342.09 | 46.0 | 10.0 | 6309.55 | 1213.66 | 9.0 | 0.0 | 5.0 | ... | 3.0 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **14432** | 3285.00 | 6.0 | 2267.60 | 809.0 | 6.0 | 1693.25 | 1693.25 | 2.0 | 0.0 | 1.0 | ... | 3.0 | |
| **15586** | 16987.54 | 11.0 | 19365.80 | 717.0 | 11.0 | 15645.11 | 15645.11 | 1.0 | 0.0 | 1.0 | ... | 1.0 | |
| **15712** | 29528.68 | 1.0 | 1473.91 | 32.0 | 1.0 | 1370.22 | 1370.22 | 18.0 | 9.0 | 13.0 | ... | 4.0 | 9 |
| **15982** | 12457.53 | 6.0 | 7950.36 | 490.0 | 6.0 | 6407.25 | 6407.25 | 2.0 | 0.0 | 2.0 | ... | 3.0 | 5 |
| **16548** | 23269.34 | 2.0 | 2888.83 | 44.0 | 11.0 | 11328.83 | 1972.22 | 7.0 | 0.0 | 4.0 | ... | 3.0 | 2 |

4761 rows × 24 columns

In [50]:

```
data['下个月回款次数'].describe()
```

Out[50]:

```
count    4761.000000
mean        1.636528
std         1.042780
min         1.000000
25%         1.000000
50%         1.000000
75%         2.000000
max        11.000000
Name: 下个月回款次数, dtype: float64
```

In [51]:

```
data.describe(include='all').T
```

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **二期天数** | 4761.0 | 36.557236 | 56.723446 | 0.00 | 11.000000 | 19.000000 | 46.000000 | 900.00 |
| **剩余期数** | 4761.0 | 12.176644 | 6.923690 | 1.00 | 7.000000 | 10.000000 | 18.000000 | 35.00 |
| **剩余本金** | 4761.0 | 11450.012220 | 6740.074859 | 100.00 | 6647.240000 | 9889.770000 | 15215.060000 | 44370.98 |
| **逾期本金** | 4761.0 | 1461.300015 | 1641.631317 | 0.00 | 784.180000 | 1109.180000 | 1681.910000 | 29001.23 |
| **近一月** | | | | | | | | |

In [52]:

```
#观察逾期与账龄的关系
#逾期分类
data['是否逾期']=(data['逾期总额']>0).astype(int)
```

In [53]:

```
data['是否逾期'].value_counts()
```

Out[53]:

```
1    4325
0     436
Name: 是否逾期, dtype: int64
```

In [77]:

```
data['近三月电话接通次数'].value_counts()
```

```
34.0     9
26.0     8
33.0     7
30.0     5
29.0     4
42.0     4
39.0     3
32.0     2
31.0     2
47.0     1
52.0     1
40.0     1
51.0     1
53.0     1
41.0     1
35.0     1
38.0     1
43.0     1
49.0     1
```

In [56]:

```
data['近三月接通率']=data['近三月电话接通次数']/data['近三月拨打次数']
data['近三月日均拨打次数']=data['近三月拨打次数']/data['近三月电话拨打天数']
data['近三月回款金额率']=data['近三月回款金额']/data['逾期总额']
data['近一月接通率']=data['近一月电话接通次数']/data['近一月拨打次数']
data['近一月日均拨打次数']=data['近一月拨打次数']/data['近一月电话拨打天数']
```

In [69]:

```
data=data[data['近一月接通率']<1]
```

In [70]:

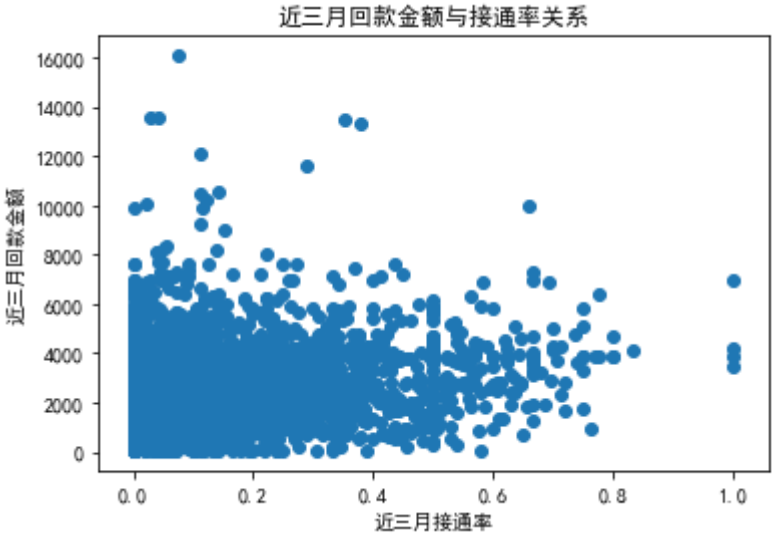
```
data['近三月接通率'].value_counts()
```

Out[70]:

```
0.000000    476
0.333333     47
0.111111     45
0.166667     45
0.500000     44
           ...
0.298611      1
0.347368      1
0.028302      1
0.207317      1
0.078571      1
Name: 近三月接通率, Length: 1146, dtype: int64
```

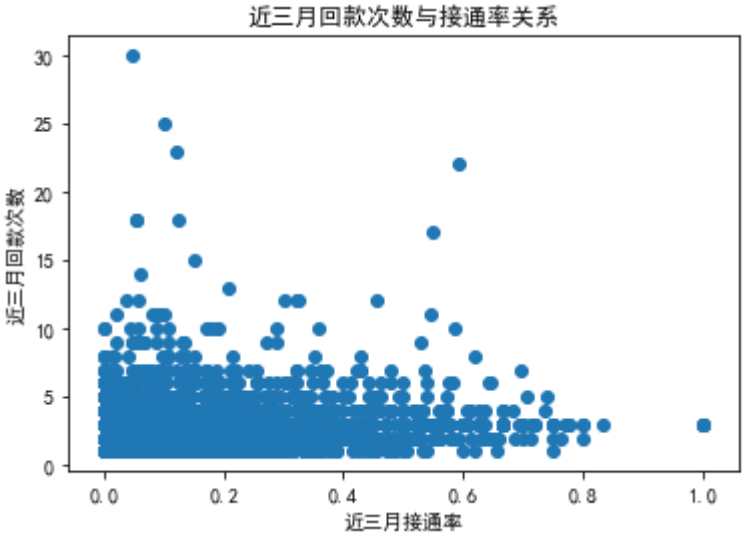
In [71]:

```
x_axis=data.loc[:,'近三月接通率']
y_axis=data.loc[:,'近三月回款金额']
plt.title('近三月回款金额与接通率关系')
plt.xlabel('近三月接通率')
plt.ylabel('近三月回款金额')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```



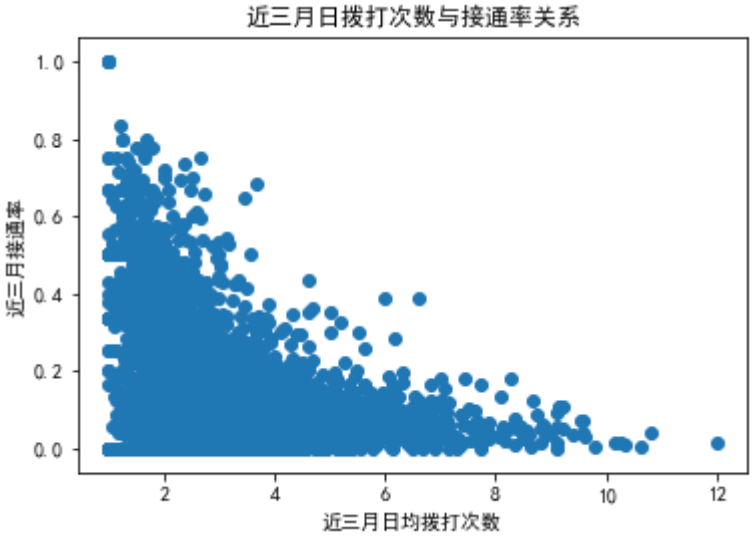

In [72]:

```
x_axis=data.loc[:,'近三月接通率']
y_axis=data.loc[:,'近三月回款次数']
plt.title('近三月回款次数与接通率关系')
plt.xlabel('近三月接通率')
plt.ylabel('近三月回款次数')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```



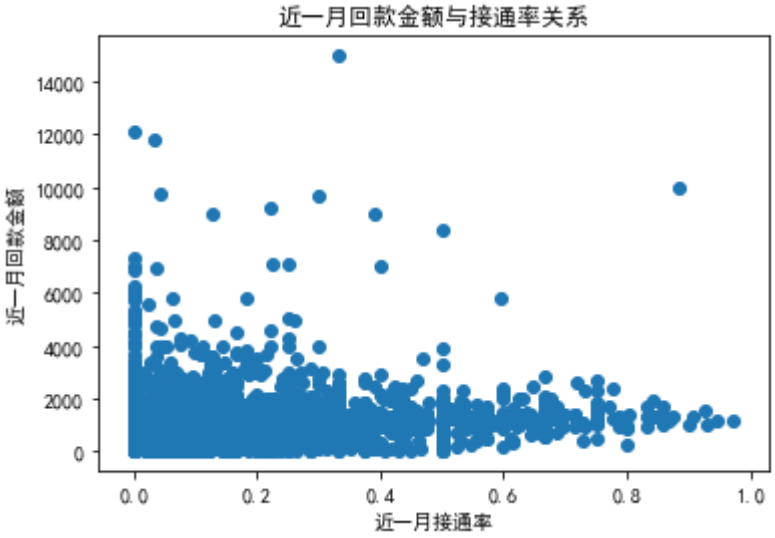
In [73]:

```
x_axis=data.loc[:,'近三月日均拨打次数']
y_axis=data.loc[:,'近三月接通率']
plt.title('近三月日拨打次数与接通率关系')
plt.xlabel('近三月日均拨打次数')
plt.ylabel('近三月接通率')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```



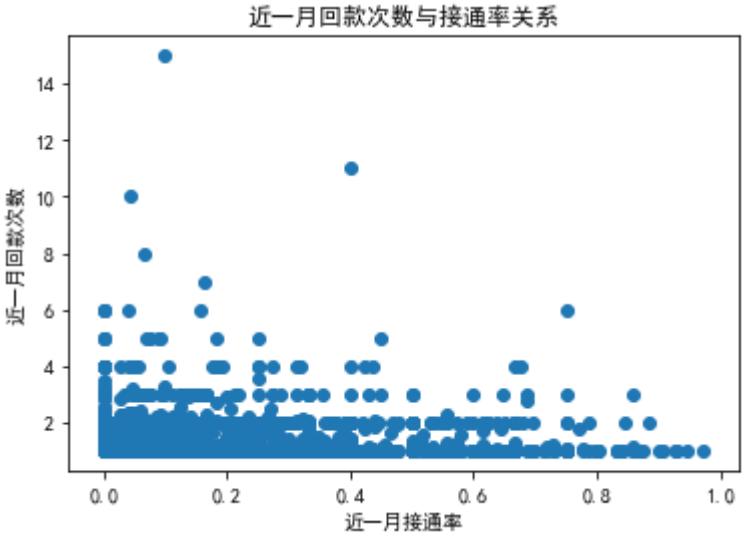

In [74]:

```
x_axis=data.loc[:,'近一月接通率']
y_axis=data.loc[:,'近一月回款金额']
plt.title('近一月回款金额与接通率关系')
plt.xlabel('近一月接通率')
plt.ylabel('近一月回款金额')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```



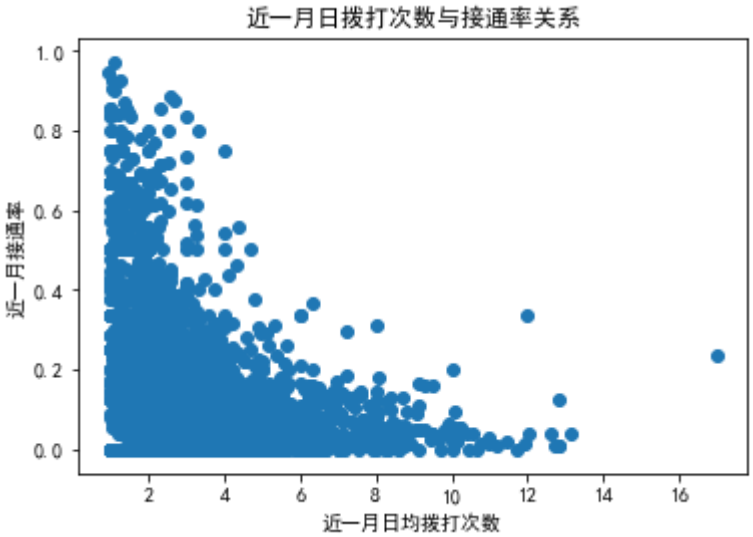
In [75]:

```
x_axis=data.loc[:,'近一月接通率']
y_axis=data.loc[:,'近一月回款次数']
plt.title('近一月回款次数与接通率关系')
plt.xlabel('近一月接通率')
plt.ylabel('近一月回款次数')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```




In [78]:

```
x_axis=data.loc[:,'近一月日均拨打次数']
y_axis=data.loc[:,'近一月接通率']
plt.title('近一月日拨打次数与接通率关系')
plt.xlabel('近一月日均拨打次数')
plt.ylabel('近一月接通率')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```

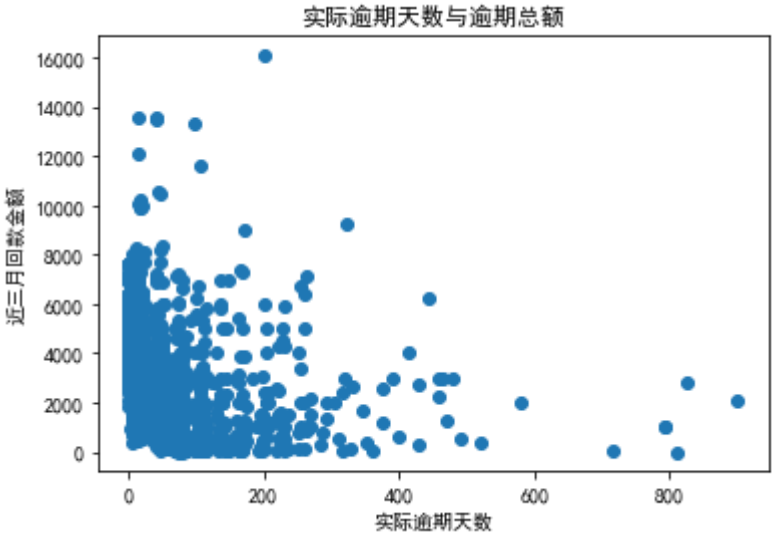



**我们发现无论是近一个月还是近三个月，拨打次数与接通率呈现明显的负相关性，随着**

## 拨打次数的提高，客户接通率变低。

In [82]:

```
x_axis=data.loc[:,'实际逾期天数']
y_axis=data.loc[:,'近三月回款金额']
plt.title('实际逾期天数与逾期总额')
plt.xlabel('实际逾期天数')
plt.ylabel('近三月回款金额')
plt.scatter(x_axis,y_axis,linewidth="1")
plt.show()
```



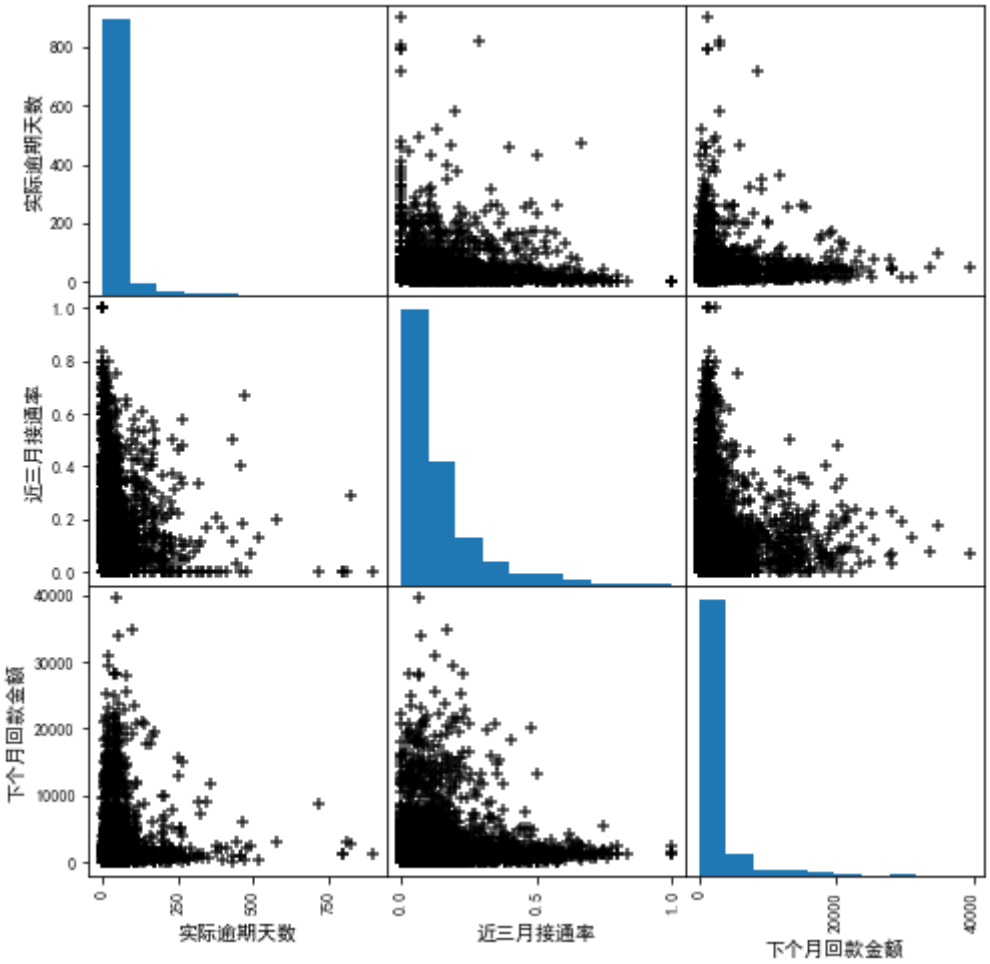

## 由表可知，随着逾期天数的增加，近三个月的回款金额数会减少

In [94]:

```
data_1 = pd.DataFrame(data,columns = ['实际逾期天数','近三月接通率','下个月回款金额'])
pd.plotting.scatter_matrix(data_1,figsize=(8,8),
                   c = 'k',
                  marker = '+',
                  diagonal='hist',
                  alpha = 0.8,
                  range_padding=0.1)
data_1.head()
```

Out[94]:

| | 实际逾期天数 | 近三月接通率 | 下个月回款金额 |
|---|---|---|---|
| **14** | 42.0 | 0.115942 | 2050.185341 |
| **22** | 42.0 | 0.080000 | 2767.433560 |
| **68** | 51.0 | 0.044444 | 3052.465736 |
| **101** | 15.0 | 0.146341 | 1486.894280 |
| **124** | 46.0 | 0.189655 | 1668.972703 |

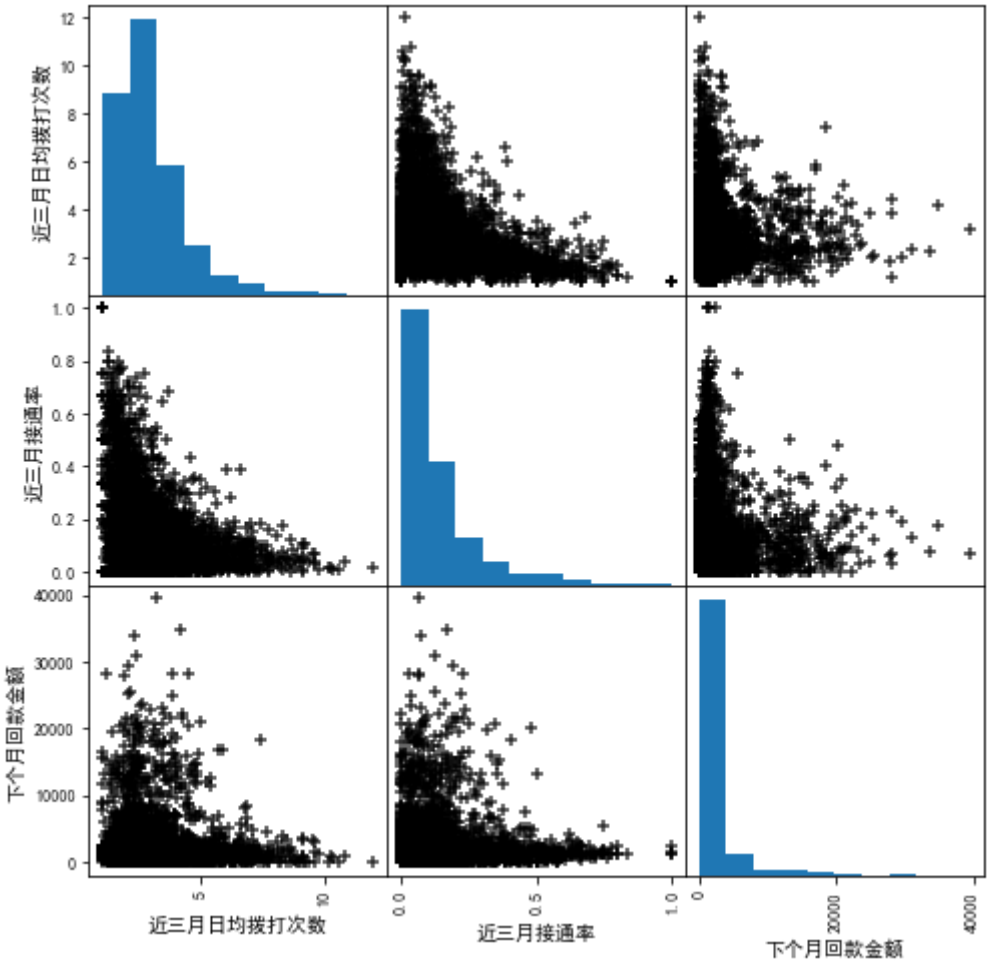


In [93]:

```
data_1 = pd.DataFrame(data,columns = ['近三月日均拨打次数','近三月接通率','下个月回款金额'])
pd.plotting.scatter_matrix(data_1,figsize=(8,8),
                   c = 'k',
                  marker = '+',
                  diagonal='hist',
                  alpha = 0.8,
                  range_padding=0.1)
data_1.head()
```

Out[93]:

| | 近三月日均拨打次数 | 近三月接通率 | 下个月回款金额 |
|---|---|---|---|
| **14** | 2.875000 | 0.115942 | 2050.185341 |
| **22** | 2.000000 | 0.080000 | 2767.433560 |
| **68** | 1.956522 | 0.044444 | 3052.465736 |
| **101** | 3.153846 | 0.146341 | 1486.894280 |
| **124** | 2.416667 | 0.189655 | 1668.972703 |



**通过多方法分析，我们可以知道，一旦客户逾期天数超过半年（200天），客户接通率会很低，而且每天拨打电话次数在五次以内相对来说更不容易使客户厌烦，随着拨打次数的增加，客户下个月回款金额会相对来说减少，所以公司可以把握这个尺寸来提高还款率。**

In [ ]: