In [1]:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

In [2]:

```
df = pd.read_excel('数据集.xlsx',encoding='utf-8')
```

In [3]:

```
df
```

Out[3]:

	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	M2	3	4694.29	75	11	1	b	15657.3	4106.65	...	4	NaN	NaN	NaN
1	25300	M2	3	4403.29	75	23	1	b	24359.8	2876.21	...	3	NaN	NaN	NaN
2	18000	M1	2	3192.5	44	10	1	b	15142.9	2912.78	...	4	NaN	NaN	NaN
3	11500	M3	4	4296.64	105	12	1	b	11500	3686.33	...	4	NaN	NaN	NaN
4	16900	C-M1	1	896.96	13	21	1	b	14993.3	647.82	...	3	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16745	11000	M9	7	7060.03	295	7	1	b	6260.05	6260.05	...	12	NaN	NaN	NaN
16746	11000	M4	2	2041.43	143	2	1	b	1691.55	1691.55	...	6	NaN	NaN	NaN
16747	15400	M9	7	10691.2	295	7	1	b	9197.14	9197.14	...	5	NaN	NaN	NaN
16748	15400	M10	8	12213.5	325	8	1	b	10461.5	10461.5	...	2	NaN	NaN	NaN
16749	22000	M10	11	15351	325	20	1	b	18674.5	9828.29	...	3	NaN	NaN	NaN

16750 rows × 27 columns

In [4]:

```
df.shape
```

Out[4]:

```
(16750, 27)
```

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16750 entries, 0 to 16749
Data columns (total 27 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   合同金额        16750 non-null  object
 1   账龄          12629 non-null  object
 2   逾期期数        12629 non-null  object
 3   逾期总额        12629 non-null  object
 4   实际逾期天数      12629 non-null  object
 5   剩余期数        12629 non-null  object
 6   产品          12629 non-null  object
 7   来源渠道        12629 non-null  object
 8   剩余本金        12629 non-null  object
 9   逾期本金        12629 non-null  object
 10  近一月拨打次数     13232 non-null  object
 11  近一月电话接通次数  13232 non-null  object
 12  近一月电话拨打天数  13232 non-null  object
 13  近一月拨打电话数    13232 non-null  object
 14  近三月拨打次数     15676 non-null  object
 15  近三月电话接通次数  15676 non-null  object
 16  近三月电话拨打天数  15676 non-null  object
 17  近三月拨打电话数    15676 non-null  object
 18  近一月回款金额     6012 non-null   object
 19  近一月回款次数     6012 non-null   object
 20  近一月最大回款金额  6012 non-null   object
 21  近三月回款金额     11182 non-null  object
 22  近三月回款次数     11182 non-null  object
 23  近三月最大回款金额  11182 non-null  object
 24  下个月回款金额     9119 non-null   object
 25  下个月回款次数     9119 non-null   object
 26  下个月最大回款金额  9119 non-null   object
dtypes: object(27)
memory usage: 3.5+ MB
```

In [6]:

```
df.columns
```

Out[6]:

```
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 [7]:

```
df.duplicated().sum()
```

Out[7]:

0

In [8]:

```
df.loc[10000]
```

Out[8]:

```
合同金额              合同金额
账龄                  账龄
逾期期数              逾期期数
逾期总额              逾期总额
实际逾期天数          实际逾期天数
剩余期数              剩余期数
产品                  产品
来源渠道              来源渠道
剩余本金              剩余本金
逾期本金              逾期本金
近一月拨打次数        近一月拨打次数
近一月电话接通次数    近一月电话接通次数
近一月电话拨打天数    近一月电话拨打天数
近一月拨打电话数      近一月拨打电话数
近三月拨打次数        近三月拨打次数
近三月电话接通次数    近三月电话接通次数
近三月电话拨打天数    近三月电话拨打天数
近三月拨打电话数      近三月拨打电话数
近一月回款金额        近一月回款金额
近一月回款次数        近一月回款次数
近一月最大回款金额    近一月最大回款金额
近三月回款金额        近三月回款金额
近三月回款次数        近三月回款次数
近三月最大回款金额    近三月最大回款金额
下个月回款金额        下个月回款金额
下个月回款次数        下个月回款次数
下个月最大回款金额    下个月最大回款金额
Name: 10000, dtype: object
```

In [9]:

```
df.drop([10000],inplace=True)
```

In [10]:

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

In [11]:

```
df.describe(include='all')
```

Out[11]:

	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
count	16749.000000	12628	12628.000000	12628.000000	12628.000000	12628.000000	12628.0
unique	NaN	37	NaN	NaN	NaN	NaN	6.0
top	NaN	C-M1	NaN	NaN	NaN	NaN	1.0
freq	NaN	3259	NaN	NaN	NaN	NaN	12245.0
mean	18624.729573	NaN	3.599541	5157.835481	164.839404	12.726402	NaN
std	8532.845317	NaN	3.489975	7443.908405	244.507912	6.845474	NaN
min	1090.000000	NaN	0.000000	0.000000	0.000000	1.000000	NaN
25%	10500.000000	NaN	1.000000	1346.400000	19.000000	8.000000	NaN
50%	19000.000000	NaN	2.000000	2559.660000	51.000000	11.000000	NaN
75%	23269.340000	NaN	5.000000	5364.587500	159.000000	19.000000	NaN
max	63600.230000	NaN	32.000000	75091.310000	1159.000000	36.000000	NaN

11 rows × 27 columns

In [12]:

```
df.fillna(0,inplace=True)
df.head()
```

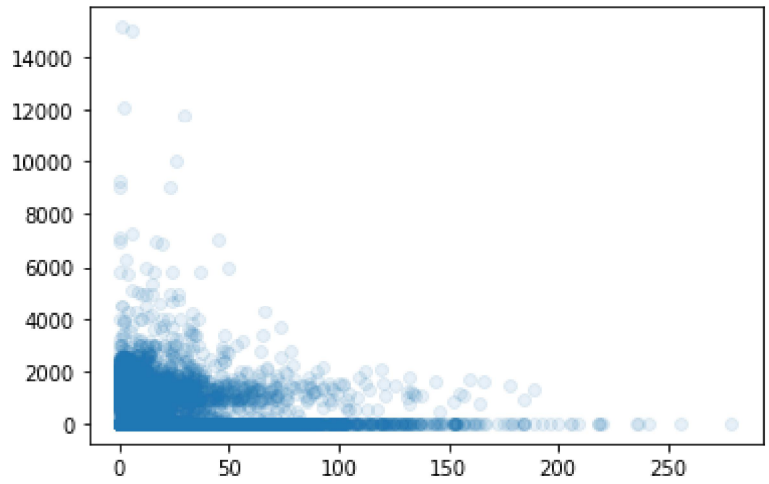
Out[12]:

	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	近回
0	17000.0	M2	3.0	4694.29	75.0	11.0	1	b	15657.33	4106.65	...	4.0	0.0	0.0	0.0	
1	25300.0	M2	3.0	4403.29	75.0	23.0	1	b	24359.81	2876.21	...	3.0	0.0	0.0	0.0	
2	18000.0	M1	2.0	3192.50	44.0	10.0	1	b	15142.91	2912.78	...	4.0	0.0	0.0	0.0	167
3	11500.0	M3	4.0	4296.64	105.0	12.0	1	b	11500.00	3686.33	...	4.0	0.0	0.0	0.0	
4	16900.0	C-M1	1.0	896.96	13.0	21.0	1	b	14993.26	647.82	...	3.0	0.0	0.0	0.0	158

5 rows × 27 columns

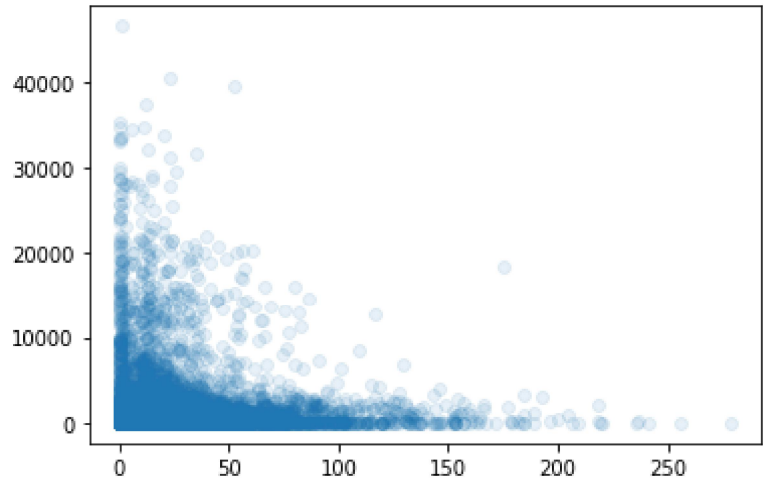
In [13]:

```
x = df['近一月拨打次数']
y = df['近一月回款金额']
plt.scatter(x,y,alpha=0.1)
plt.show()
```



In [14]:

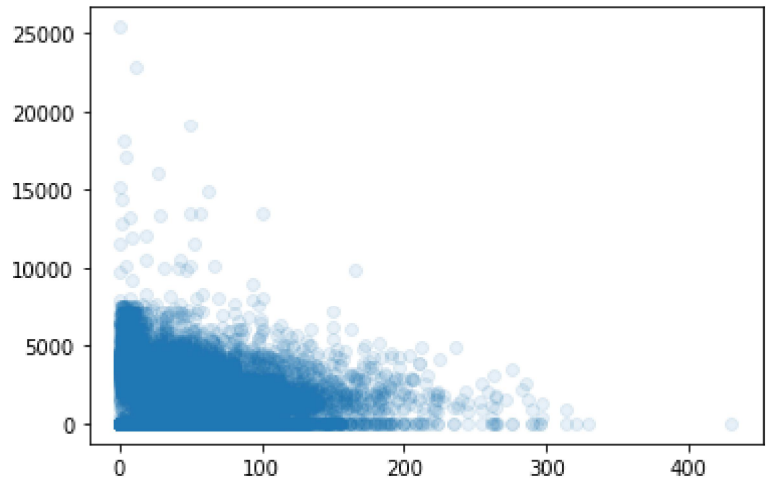
```
x = df['近一月拨打次数']
y = df['下个月回款金额']
plt.scatter(x,y,alpha=0.1)
plt.show()
```



从近一个月拨打电话数、近一个月回款金额和下个月回款金额来看，拨打次数在0-100时，用户还款较多，从时间上来看，下个月不还款的人数降低

In [15]:

```
x = df['近三月拨打次数']
y = df['近三月回款金额']
plt.scatter(x,y,alpha=0.1)
plt.show()
```



从三个月拨打次数和回款金额的关系图来看，拨打次数超过300次，回款金额与回款数量基本不会增加，应该时用户出现抵触的心理，所以三个月的电话拨打次数在200次以内就可以

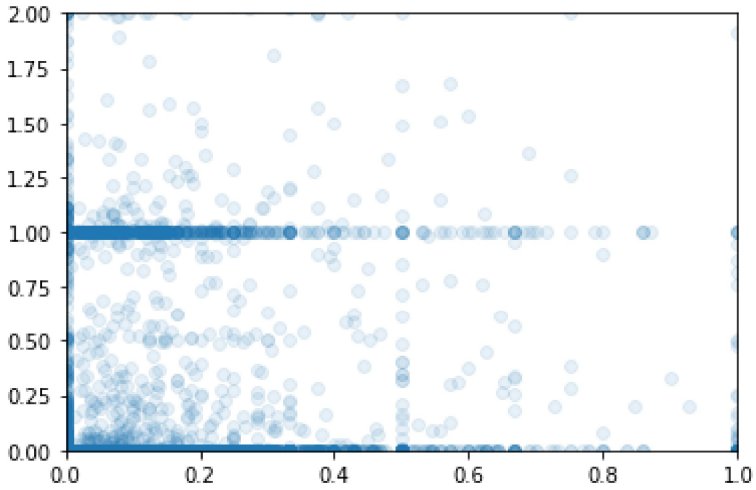
In [30]:

```
df['近一个月回款率'] = df['近一月回款金额']/df['逾期总额']
df['近三个月回款率'] = df['近三月回款金额']/df['逾期总额']
df['下个月回款率'] = df['下个月回款金额']/df['逾期总额']

df['近一个月接通率'] = df['近一月电话接通次数']/df['近一月拨打次数']
df['近三个月接通率'] = df['近三月电话接通次数']/df['近三月拨打次数']
```

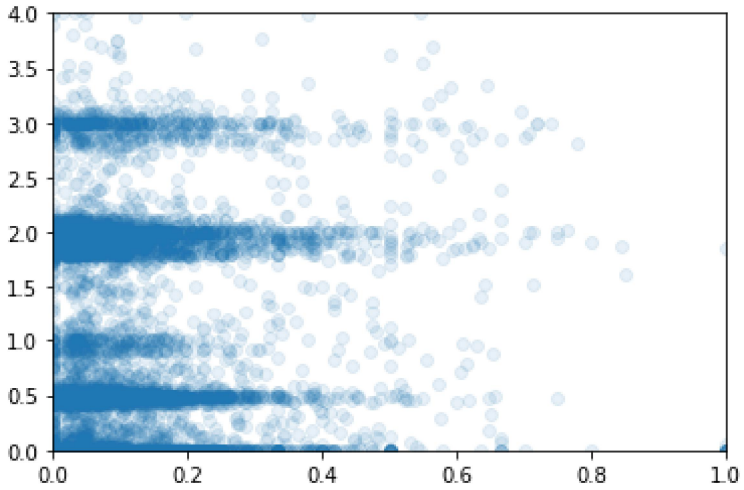
In [46]:

```
x = df['近一个月接通率']
y = df['近一个月回款率']
plt.xlim(0,1)
plt.ylim(0,2)
plt.scatter(x,y,alpha=0.1)
plt.show()
```



In [41]:

```
x = df['近三个月接通率']
y = df['近三个月回款率']
plt.xlim(0,1)
plt.ylim(0,4)
plt.scatter(x,y,alpha=0.1)
plt.show()
```



In [ ]:

```

```