一、项目背景

P2P 借贷平台日益发展，对借贷数据进行分析，对 P2P 平台健康发展，规范信贷制度具有重要意义

二、数据描述

LC.csv 数据集-LC (Loan Characteristics) 表为标的特征表，每支标一条记录。共有 21 个字段，包括一个主键（listingid）、7 个标的特征和 13 个成交当时的借款人信息，全部为成交当时可以获得的信息。

信息的维度比较广，大致可以分为基本信息，认证信息，信用信息，借款信息。

基本信息：年龄、性别；

认证信息：手机认证、户口认证、视频认证、征信认证、淘宝认证；

信用信息：初始评级、历史正常还款期数、历史逾期还款期数；

借款信息：历史成功借款金额、历史成功借款次数、借款金额、借款期限、借款成功日期

三、问题提出

1.用户画像，包含使用平台贷款业务的用户的性别比例，学历水平，是否为旧有用户，年龄分布等信息。

2.资金储备，每日借款金额大概多少？波动有多大？从而公司每日需准备多少资金可以保证不会出现资金短缺？

3.**用户逾期率**，借款人的初始评级、借款类型、性别、年龄等特征对于逾期还款的概率有无显著影响？哪些群体逾期还款率明显较高？

4.**借款利率**，哪些群体更愿意接受较高的借款利率？

四、数据清洗

```
import pandas as pd
import numpy as np
data=pd.read_csv(r'C:\Users\smile\Desktop\p2p\LC.csv')
data.head()
```

	ListingId	借款金额	借款期限	借款利率	借款成功日期	初始评级	借款类型	是否首标	年龄	性别	...	户口认证	视频认证	学历认证	征信认证	淘宝认证	历史成功借款次数	历史成功借款金额	总待还本金	历史正常还款期数	历史逾期还款期数
0	126541	18000	12	18.0	2015-05-04	C	其他	否	35	男	...	未成功认证	成功认证	未成功认证	未成功认证	未成功认证	11	40326.0	8712.73	57	16
1	133291	9453	12	20.0	2015-03-16	D	其他	否	34	男	...	成功认证	未成功认证	未成功认证	未成功认证	未成功认证	4	14500.0	7890.64	13	1
2	142421	27000	24	20.0	2016-04-26	E	普通	否	41	男	...	未成功认证	未成功认证	未成功认证	未成功认证	未成功认证	5	21894.0	11726.32	25	3
3	149711	25000	12	18.0	2015-03-30	C	其他	否	34	男	...	成功认证	成功认证	未成功认证	未成功认证	未成功认证	6	36190.0	9703.41	41	1
4	152141	20000	6	16.0	2015-01-22	C	电商	否	24	男	...	成功认证	成功认证	未成功认证	未成功认证	未成功认证	13	77945.0	0.00	118	14

5 rows × 21 columns

```
data.shape
data.describe()  #查看数值型数据基本特征
data.describe(include='object')
```

(328553, 21)

	ListingId	借款金额	借款期限	借款利率	年龄	历史成功借款次数	历史成功借款金额	总待还本金	历史正常还款期数	历史逾期还款期数
count	3.285530e+05	328553.000000	328553.000000	328553.000000	328553.000000	328553.000000	3.285530e+05	3.285530e+05	328553.000000	328553.000000
mean	1.907948e+07	4423.816906	10.213594	20.601439	29.143042	2.323159	8.785857e+03	3.721665e+03	9.947658	0.423250
std	8.375769e+06	11219.664024	2.780444	1.772408	6.624286	2.922361	3.502736e+04	8.626061e+03	14.839899	1.595681
min	1.265410e+05	100.000000	1.000000	6.500000	17.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000
25%	1.190887e+07	2033.000000	6.000000	20.000000	24.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	0.000000
50%	1.952325e+07	3397.000000	12.000000	20.000000	28.000000	2.000000	5.000000e+03	2.542410e+03	5.000000	0.000000
75%	2.629862e+07	5230.000000	12.000000	22.000000	33.000000	3.000000	1.035500e+04	5.446810e+03	13.000000	0.000000
max	3.281953e+07	500000.000000	24.000000	24.000000	56.000000	649.000000	7.405926e+06	1.172653e+06	2507.000000	60.000000

	借款成功日期	初始评级	借款类型	是否首标	性别	手机认证	户口认证	视频认证	学历认证	征信认证	淘宝认证
count	328553	328553	328553	328553	328553	328553	328553	328553	328553	328553	328553
unique	756	6	4	2	2	2	2	2	2	2	2
top	2017-01-25	D	普通	否	男	未成功认证	未成功认证	未成功认证	未成功认证	未成功认证	未成功认证
freq	3558	134860	118103	241090	221946	205546	318448	310052	214429	318947	327401

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 328553 entries, 0 to 328552
Data columns (total 21 columns):
ListingId    328553 non-null int64
借款金额         328553 non-null int64
借款期限         328553 non-null int64
借款利率         328553 non-null float64
借款成功日期       328553 non-null object
初始评级         328553 non-null object
借款类型         328553 non-null object
是否首标         328553 non-null object
年龄           328553 non-null int64
性别           328553 non-null object
手机认证         328553 non-null object
户口认证         328553 non-null object
视频认证         328553 non-null object
学历认证         328553 non-null object
征信认证         328553 non-null object
淘宝认证         328553 non-null object
历史成功借款次数     328553 non-null int64
历史成功借款金额     328553 non-null float64
总待还本金        328553 non-null float64
历史正常还款期数     328553 non-null int64
历史逾期还款期数     328553 non-null int64
dtypes: float64(3), int64(7), object(11)
memory usage: 52.6+ MB
```

可以看到，数据还是比较完整的，无缺失值，异常值，重复值等

#观察一下年龄分布，最小 17 岁，最大 56 岁，平均年龄 29 岁，33 岁以下的占比超过了 75%。说明用户整体还是中青年。

#将年龄分为'15-20', '20-25', '25-30', '30-35', '35-40', '40+'比较合理

#观察一下借款金额分布，最小借款金额为 100 元，最大为 50 万元，平均值为 4424 元，低于 5230 的借款金额占到了 75%。

#说明应该是小额借款比较多。将借款金额分为 0-2000，2000-3000，3000-4000，4000-5000，5000-6000，6000 以上比较合理

所以，数据清洗这步可省略。

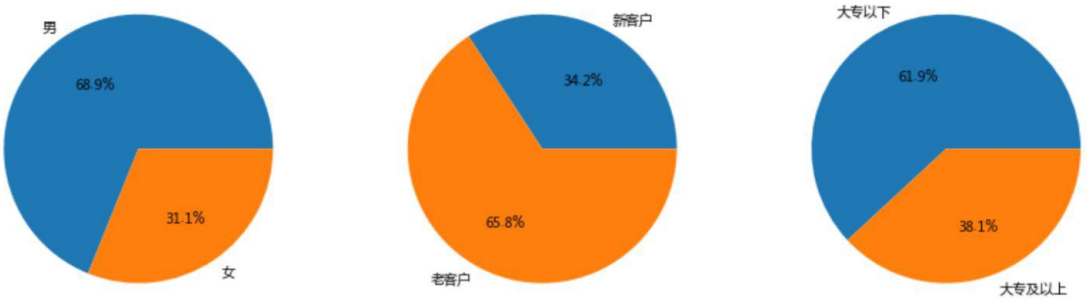
五、数据分析

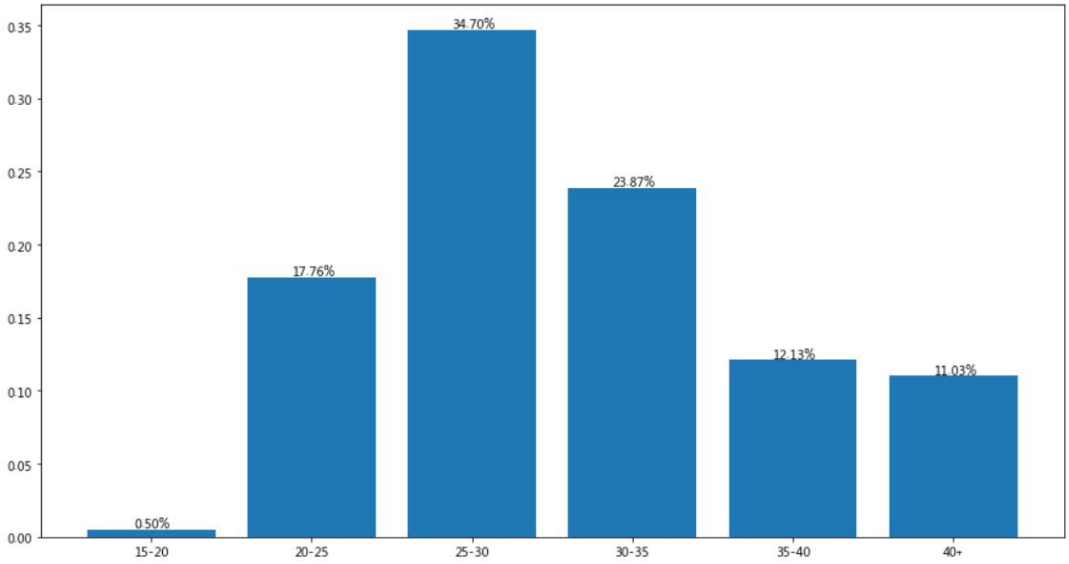
1.**分析用户画像（性别、学历、年龄、是否首标）**

按‘性别’、‘年龄’、‘是否首标’、‘学历认证’字段对‘借款金额’进行加总，用饼图或柱状图将结果可视化

```
import matplotlib.pyplot as plt
#年龄分析
male=data[data['性别']=='男']
female=data[data['性别']=='女']
sex=(male,female)
sex_data=(male['借款金额'].sum(),female['借款金额'].sum())
sex_idx=('男','女')
plt.figure(figsize=(15,6))
plt.subplot(1,3,1)
plt.pie(sex_data,labels=sex_idx,autopct='%.1f%%')
```

```
#新老客户分析
new = LC[LC['是否首标'] == '是']
old = LC[LC['是否首标'] == '否']
newold_data = (new['借款金额'].sum(), old['借款金额'].sum())
newold_idx = ('新客户', '老客户')
plt.subplot(1,3,2)
plt.pie(newold_data, labels=newold_idx, autopct='%.1f%%')

#学历分析
ungraduate = LC[LC['学历认证'] == '未成功认证']
graduate = LC[LC['学历认证'] == '成功认证']
education_data = (ungraduate['借款金额'].sum(), graduate['借款金
额'].sum())
education_idx = ('大专以下', '大专及以上')
plt.subplot(1,3,3)
plt.pie(education_data, labels=education_idx, autopct='%.1f%%')
plt.show()

#年龄分析
ageA = LC.loc[(LC['年龄'] >= 15) & (LC['年龄'] < 20)]
ageB = LC.loc[(LC['年龄'] >= 20) & (LC['年龄'] < 25)]
ageC = LC.loc[(LC['年龄'] >= 25) & (LC['年龄'] < 30)]
ageD = LC.loc[(LC['年龄'] >= 30) & (LC['年龄'] < 35)]
ageE = LC.loc[(LC['年龄'] >= 35) & (LC['年龄'] < 40)]
ageF = LC.loc[LC['年龄'] >= 40]
age = (ageA, ageB, ageC, ageD, ageE, ageF)
age_total = 0
age_percent =[]
for i in age:
    tmp = i['借款金额'].sum()
    age_percent.append(tmp)
    age_total  += tmp
age_percent /= age_total
age_idx = ['15-20', '20-25', '25-30', '30-35', '35-40', '40+']
plt.figure(figsize=(15, 8))
plt.bar(age_idx, age_percent)
for (a, b) in zip(age_idx, age_percent):
    plt.text(a, b+0.001, '%.2f%%' % (b * 100), ha='center',
va='bottom', fontsize=10)
plt.show()
```





结论：

（1）男性客户的贡献的贷款金额占到了 69%，可能的原因是男性更倾向于提前消费且贷款金额较大。

（2）非首标的金额占比达到 66%，说明用户倾向于多次使用，产品粘性较高。

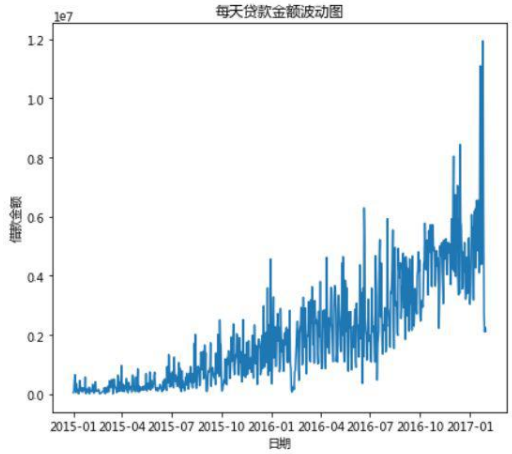
（3）大专以下学历的贷款金额更多，但是由于可能有很多用户并未认证学历，所以数据存在出入。

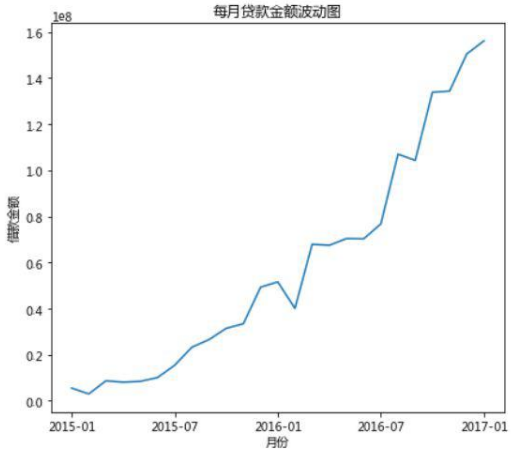
（4）年龄段在 25-30 岁之间的借款金额最多，而 20-35 岁的人群占比超过 75%，是该产品的主力消费人群。

2.**分析资金储备**

每日的借款金额大概多少？波动有多大？公司每日需要准备多少资金可以保证不会出现资金短缺?

```
from datetime import datetime

#分析每日贷款金额的走势
loan = LC[['借款成功日期', '借款金额']].copy()
loan['借款日期'] = pd.to_datetime(loan['借款成功日期'])
loan1 = loan.pivot_table(index='借款日期', aggfunc='sum').copy()
plt.figure(figsize=(15, 6))
plt.subplot(1,2,1)
plt.plot(loan1)
plt.xlabel('日期')
plt.ylabel('借款金额')
plt.title('每天贷款金额波动图')

#分析每月贷款金额的走势
loan['借款成功月份'] = [datetime.strftime(x, '%Y-%m') for x in
loan['借款日期']]
loan2 = loan.pivot_table(index='借款成功月份',
aggfunc='sum').copy()
plt.subplot(1,2,2)
plt.plot(loan2)
plt.xlabel('月份')
plt.xticks(['2015-01','2015-07','2016-01','2016-07','2017-01']
)
plt.ylabel('借款金额')
plt.title('每月贷款金额波动图')
plt.show()

# 对 2017 年 1 月的数据继续进行分析，并求出平均值和标准差
loan3 = loan1.loc['2017-01']
avg = loan3['借款金额'].mean()
std = loan3['借款金额'].std()
print(avg, std)
```





5204663.8 2203394.1435809094

结论：

（1）.每日贷款金额呈现的是一个往上的趋势,但是每天的波动较大。

（2）.每月贷款分析结论：从 2015 年 1 月到 2017 年 1 月，月度贷款金额呈现上升趋势，上升速度随着时间增快。

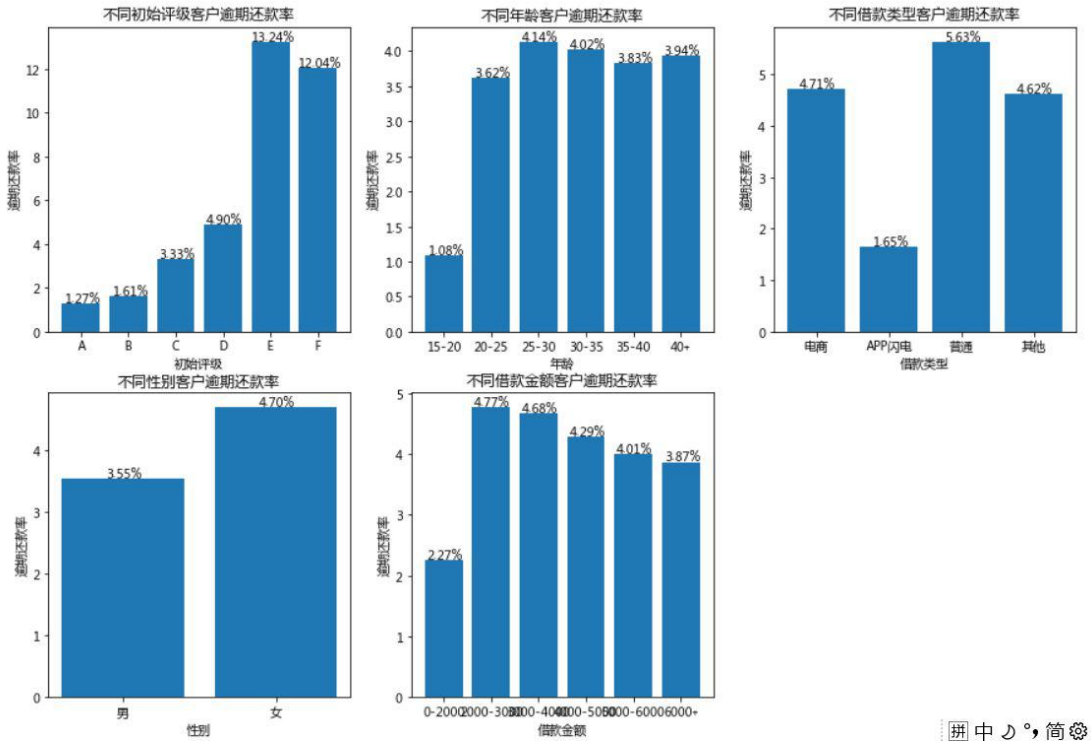
（3）.2017年1月每日的借款金额达到5204664元,标准差为2203394,根据 3σ原则，想使每日借款金额充足的概率达到 99.9%，则每日公式账上需准备 5204664+2203394×3=11814846 元。

3.分析逾期还款率（借款人的初始评级、借款类型、性别、年龄、借款金额等特征）

逾期还款率 = 历史逾期还款期数/（历史逾期还款期数+历史正常还款期数）

```
#初始评级的数据划分
level_idx = ('A','B','C','D','E','F')
lev = []
for i in level_idx:
    temp = LC[LC['初始评级'] == i]
    lev.append(temp)

#借款类型的数据划分
kind_idx = ('电商', 'APP 闪电','普通', '其他')
kind = []
for i in kind_idx:
    temp = LC[LC['借款类型'] == i]
    kind.append(temp)

#不同借款金额的数据划分
amount_idx = ('0-2000', '2000-3000', '3000-4000', '4000-5000',
'5000-6000', '6000+')
amountA = LC.loc[(LC['借款金额'] > 0) & (LC['借款金额'] < 2000)]
amountB = LC.loc[(LC['借款金额'] >= 2000) & (LC['借款金额'] < 3000)]
amountC = LC.loc[(LC['借款金额'] >= 3000) & (LC['借款金额'] < 4000)]
amountD = LC.loc[(LC['借款金额'] >= 4000) & (LC['借款金额'] < 5000)]
amountE = LC.loc[(LC['借款金额'] >= 5000) & (LC['借款金额'] < 6000)]
amountF = LC.loc[(LC['借款金额'] >= 6000)]
amount = (amountA, amountB, amountC, amountD,amountE,amountF)

LC['逾期还款率'] = LC['历史逾期还款期数']/(LC['历史逾期还款期数
']+LC['历史正常还款期数'])*100

#逾期还款率的分析图
def depayplot(i,idx,data,xlabel,title,index):
    depay = []
    for a in data:
        tmp = a[index].mean()
        depay.append(tmp)
    plt.subplot(2,3,i)
    plt.bar(idx, depay)
    for (a, b) in zip(idx, depay):
        plt.text(a, b+0.001, '%.2f%%'% b, ha='center', va='bottom',
fontsize=10)
```

```
    plt.xlabel(xlabel)
    plt.ylabel(index)
    plt.title(title)

plt.figure(figsize=(15, 10))
index = '逾期还款率'
# 根据初始评级对逾期还款率进行分析
depayplot(1,level_idx,lev,'初始评级','不同初始评级客户逾期还款率',index)

# 根据年龄对逾期还款率进行分析
depayplot(2,age_idx,age,'年龄','不同年龄客户逾期还款率',index)

# 根据借款类型对逾期还款率进行分析
depayplot(3,kind_idx,kind,'借款类型','不同借款类型客户逾期还款率',index)

# 根据性别对逾期还款率进行分析
depayplot(4,sex_idx,sex,'性别','不同性别客户逾期还款率',index)

# 根据借款金额对逾期还款率进行分析
depayplot(5,amount_idx,amount,'借款金额','不同借款金额客户逾期还款率',index)

plt.show()
```



结论：

（1）.初始评级对于贷款者的还款能力有比较好的预测作用，EF 两级反转可能是因为样本数量较少，ABCD 四个等级的平均逾期还款率都比较小，而 EF 两级明显增大，故公司对于这两类贷款者要谨慎对待。

（2）.年龄对于逾期率的分布较为平均，25-30 岁的年轻人可以重点关注。

（3）.APP 闪电的逾期还款率明显低于其他三种，故公司可以多考虑与“APP 闪电”借款类型的合作。

（4）.女性的逾期率高于男性，可能是由于生活中男性收入较女性高造成的。

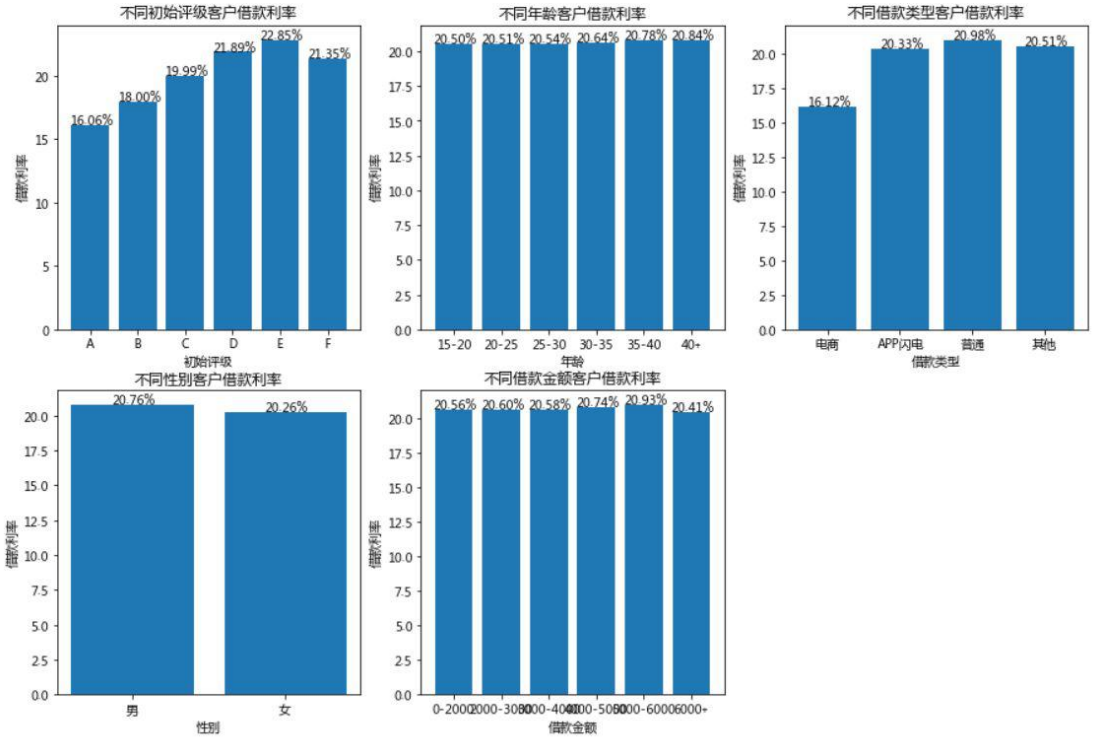
（5）.借款金额在 2000 以下的逾期还款率最低，2000-3000 之间的最高。可以多考虑小额贷款降低逾期风险。

4.分析借款利率（借款人的初始评级、借款类型、性别、年龄、借款金额等特征）

哪些客户群体更愿意接受较高的借款利率？

```
#借款利率的分析图
plt.figure(figsize=(15, 10))
index1 = '借款利率'

# 根据初始评级对借款利率进行分析
```

```
depayplot(1,level_idx,lev,'初始评级','不同初始评级客户借款利率
',index1)

# 根据年龄对借款利率进行分析
depayplot(2,age_idx,age,'年龄','不同年龄客户借款利率',index1)

# 根据借款类型对借款利率进行分析
depayplot(3,kind_idx,kind,'借款类型','不同借款类型客户借款利率
',index1)

# 根据性别对借款利率进行分析
depayplot(4,sex_idx,sex,'性别','不同性别客户借款利率',index1)

# 根据借款金额对借款利率进行分析
depayplot(5,amount_idx,amount,'借款金额','不同借款金额客户借款利
率',index1)

plt.show()
```



结论：

（1）.年龄对于借款利率的分布较为平均，差异性很小。

（2）.初始评级的平均借款利率由小到大排列为 ABCDFDE。

（3）.电商的借款利率明显低于其他三种。

（4）.女性所能接受的借款利率低于男性。

（5）.借款金额对于借款利率的分布较为平均，差异性很小。

对于以上四个问题综合分析 LC 数据集：

(1)、“男性”、“回头客”、“中青年”是拍拍贷用户群体的主要特征。

(2)、每日公司账上需准备 7,283,728 元，方可保证出现当日出借金额不足的可能性小于 0.1%。

(3)、“初始评级”为 D 的群体，借款利率与 E，F 大致相当，但其逾期还款率却只有 E，F 群体的三分之一，相同的收益水平下风险大大降低，应多发展评级为 D 的客户或提高其贷款额度。

(4)、通过“app 闪电”贷款的逾期还款率远低于其他项，约为其他借款类型的三分之一至四分之一，而平均借款利率却和其他项相差不大，证明“app 闪电”是该公司优质的合作方，其所引流来得客户质量很高，“拍拍贷”应与“app 闪电”继续加深合作。

(5)、“电商”中的贷款客户，收益率水平明显较低，逾期率却不低，在该群体中的贷款收益小，风险大。

(6)、从性别上看，男性群体贷款利率较高，逾期风险较小，相较女性一定程度上是更为优质的客户，但并不明显。

本文参考：kesci.com/home/project/

本文来源：https://zhuanlan.zhihu.com/p/111935759