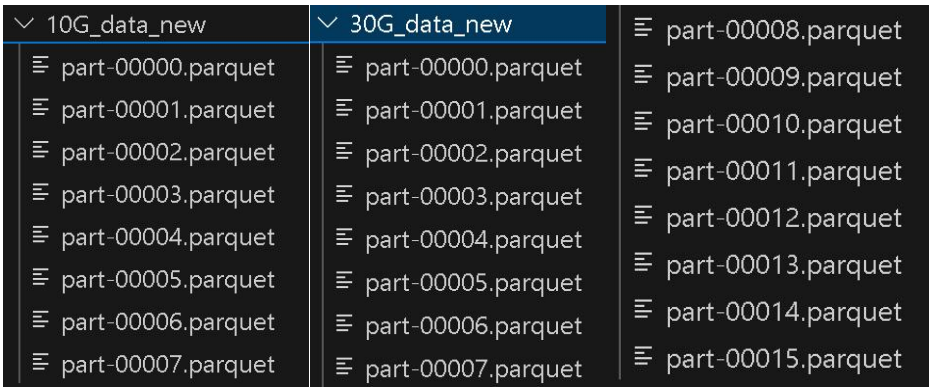
# 互评作业 1：数据探索性分析与预处理

李明泽 3220241546

## 一、 数据预处理

### 1. 数据读取




10G 数据集（左） 30G 数据集（右）

构建$concat.py$函数，负责对 10G 数据集和 30G 数据集中的数据进行拼接，其中由于 30G 数据集的数据量过大，所以采用先使用$drop.duplicates$对重复值进行删除，再进行拼接的方法最终得到 10G 数据集的 45,000,000 条数据（未删除重复值），30G 数据集的 135,000,000 条数据（未删除重复值）。（其中 10G 数据集的每个子数据集包含 5625000 条数据，30G 数据集的每个子数据集包含 8437500 条数据）

### 2. 缺失值统计

10G 数据集中一共有 45,000,000 条数据（未删除重复值），经过缺失值统计分析，各个数据项没有发现缺失情况。

```
==========================================
10G数据集预处理开始

[缺失值统计]
                          缺失数量  缺失比例(%)
id                           0      0.0
last_login                   0      0.0
age                          0      0.0
income                       0      0.0
gender                       0      0.0
country                      0      0.0
address                      0      0.0
purchase_history             0      0.0
is_active                    0      0.0
registration_date            0      0.0
login_history                0      0.0
last_login_year              0      0.0
last_login_month             0      0.0
last_login_day               0      0.0
registration_date_year       0      0.0
registration_date_month      0      0.0
registration_date_day        0      0.0
```

30G 数据集中一共有 135,000,000 条数据（未删除重复值），经过缺失值统计发现，各个数据项没有发现缺失情况。

```
==========================================
30G数据集预处理开始

[缺失值统计]
                        缺失数量  缺失比例(%)
id                         0      0.0
last_login                 0      0.0
age                        0      0.0
income                     0      0.0
gender                     0      0.0
country                    0      0.0
address                    0      0.0
purchase_history           0      0.0
is_active                  0      0.0
registration_date          0      0.0
login_history              0      0.0
last_login_year            0      0.0
last_login_month           0      0.0
last_login_day             0      0.0
registration_date_year     0      0.0
registration_date_month    0      0.0
registration_date_day      0      0.0
```

### 3. 重复值统计

使用$df.duplicated(subset = 'user_name', keep = 'first').sum()$函数进行重复值统计，再使用$df.drop_duplicates(subset = 'user_name', keep = 'first', inplace = True)$函数删除重复值，得到剩下的删除重复值数据。

10G 数据集中各个子集包含的重复数据如图所示：

```
part-00000.parquet                  part-00004.parquet
[重复值统计] 37710                  [重复值统计] 37300
[删除重复值后数据量统计] 5587290    [删除重复值后数据量统计] 5587700
part-00001.parquet                  part-00005.parquet
[重复值统计] 37452                  [重复值统计] 37359
[删除重复值后数据量统计] 5587548    [删除重复值后数据量统计] 5587641
part-00002.parquet                  part-00006.parquet
[重复值统计] 37318                  [重复值统计] 37490
[删除重复值后数据量统计] 5587682    [删除重复值后数据量统计] 5587510
part-00003.parquet                  part-00007.parquet
[重复值统计] 37721                  [重复值统计] 37329
[删除重复值后数据量统计] 5587279    [删除重复值后数据量统计] 5587671
```

| 统计项 | 数值 |
| --- | --- |
| 总重复数据量 | 299,679 条 |
| 删除重复后总数据量 | 44,700,321 条 |

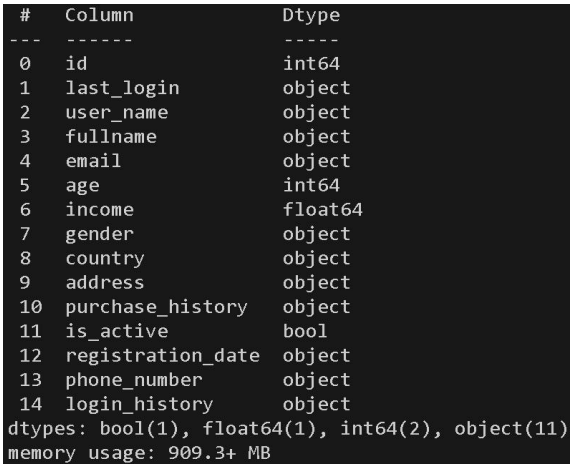
30G 数据集中各个子集包含的重复数据如图所示:

```
part-00000.parquet                  part-00008.parquet
[重复值统计] 83430                  [重复值统计] 83531
[删除重复值后数据量统计] 8354070    [删除重复值后数据量统计] 8353969
part-00001.parquet                  part-00009.parquet
[重复值统计] 83607                  [重复值统计] 83646
[删除重复值后数据量统计] 8353893    [删除重复值后数据量统计] 8353854
part-00002.parquet                  part-000010.parquet
[重复值统计] 82973                  [重复值统计] 83255
[删除重复值后数据量统计] 8354527    [删除重复值后数据量统计] 8354245
part-00003.parquet                  part-000011.parquet
[重复值统计] 83345                  [重复值统计] 83870
[删除重复值后数据量统计] 8354155    [删除重复值后数据量统计] 8353630
part-00004.parquet                  part-000012.parquet
[重复值统计] 83093                  [重复值统计] 83201
[删除重复值后数据量统计] 8354407    [删除重复值后数据量统计] 8354299
part-00005.parquet                  part-000013.parquet
[重复值统计] 83649                  [重复值统计] 83666
[删除重复值后数据量统计] 8353851    [删除重复值后数据量统计] 8353834
part-00006.parquet                  part-000014.parquet
[重复值统计] 84173                  [重复值统计] 83191
[删除重复值后数据量统计] 8353327    [删除重复值后数据量统计] 8354309
part-00007.parquet                  part-000015.parquet
[重复值统计] 83628                  [重复值统计] 82898
[删除重复值后数据量统计] 8353872    [删除重复值后数据量统计] 8354602
```

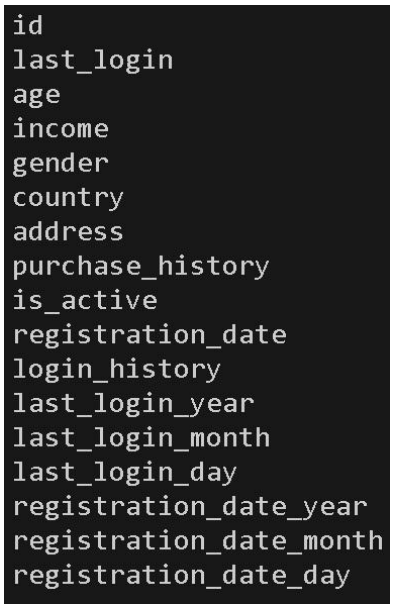
| 统计项 | 数值 |
| --- | --- |
| 总重复数据量 | 1,334,156 条 |
| 删除重复后总数据量 | 133,663,844 条 |

## 4. 异常值处理

数据集中包含如下数据项:



```
 #   Column            Dtype
---  ------            -----
 0   id                int64
 1   last_login        object
 2   user_name         object
 3   fullname          object
 4   email             object
 5   age               int64
 6   income            float64
 7   gender            object
 8   country           object
 9   address           object
 10  purchase_history  object
 11  is_active         bool
 12  registration_date  object
 13  phone_number      object
 14  login_history     object
dtypes: bool(1), float64(1), int64(2), object(11)
memory usage: 909.3+ MB
```



```
id
last_login
age
income
gender
country
address
purchase_history
is_active
registration_date
login_history
last_login_year
last_login_month
last_login_day
registration_date_year
registration_date_month
registration_date_day
```

原数据集数据项（左）预处理后数据项（右）

其中$email$、$fullname$、$user\_name$、$phone_number$将在步骤 4 中进行匿名化处理，本次作业对$age$、$income$、$gender$、$country$、$chinese\_address$、$purchase\_history$、$is_active$进行异常值分析。

$age$：保留 18-100 范围的数据，并且按照'18-25', '26-35', '36-45', '46-55', '55+'的范围进行分组，为后续可视化分析做准备。

$income$：过滤负值，保留收入为正的数据，并且按照'低', '中低', '中高', '高'对收入高低进行分组，为后续可视化分析做准备。

$gender$：统一格式，将 male、female、m、f 等转化为‘男’、‘女’对应的存储格式，并且对数据中包含的“其他”、“未定义”统一重定义为“其他”。

$chinese\_address$：针对数据内容，匹配首个省级行政区，如北京|上海|天津|重庆|河北等。

$purchase\_history$：解析购买历史，提取购买金额、购买物品种类；其中购买金额存在超过两位浮点数的情况，采用截取法只保留前两位小数，其余部分舍弃。

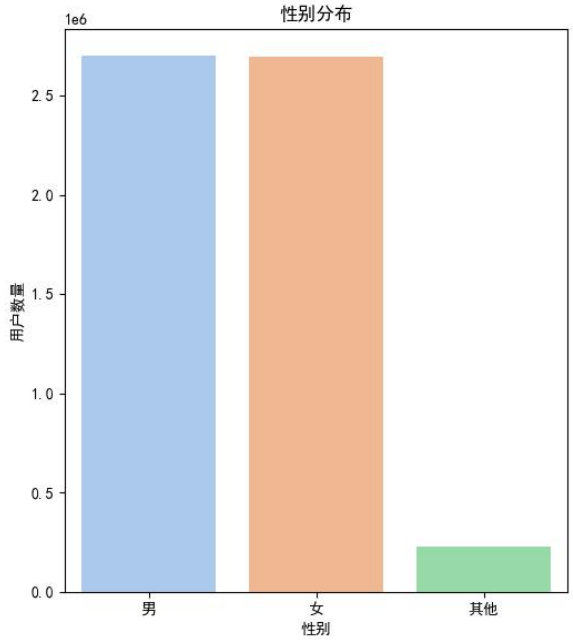
$is\_active$：将用户活跃标准转换为布尔类型，方便后续可视化分析。

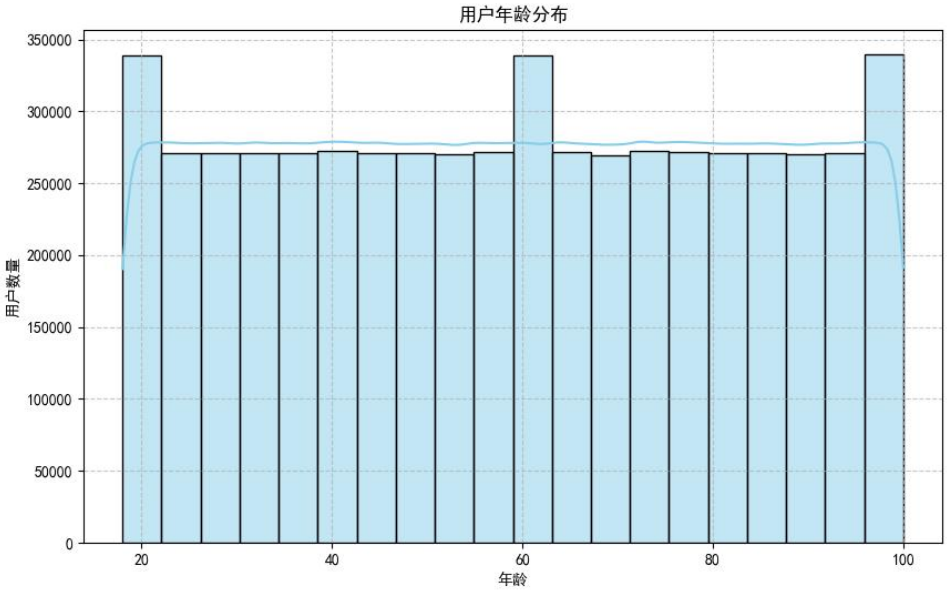
5. 数据匿名化处理

对'$user\_name$', '$chinese\_name$', '$email$', '$phone\_number$'数据项进行匿名化处理，防止用户隐私信息泄露。

## 二、 数据探索性分析和可视化分析(全部可视化分析结果见末尾)

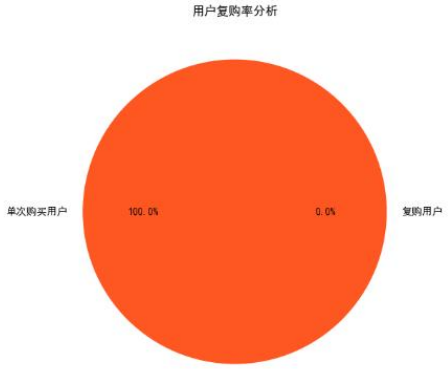
## 1. 基本柱状图



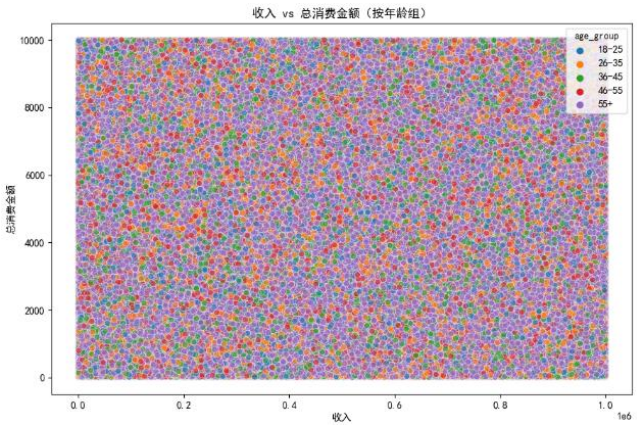




## 2. 基本饼状图


性别比例
其他
4.00270223618%
男
48.00053238869%
47.99676537514%
女

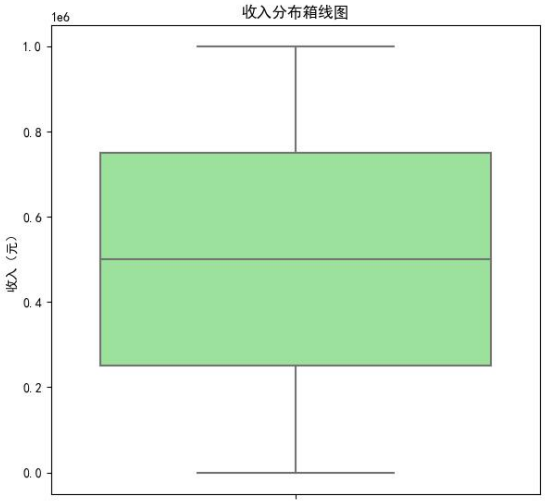

用户复购率分析
单次购买用户
100.0%
0.0%
复购用户

## 3. 热力图


收入 vs 总消费金额（按年龄组）
age_group
18-25
26-35
36-45
46-55
55+
总消费金额
收入
1e6

## 4. 词云


## 5. 箱线图




## 三、 建立用户画像

经过上述数据集预处理过程，构建包括：“年龄中位数、年龄分布、性别分布、高频国家、高频城市、收入平均水平、收入分布、消费平均水平、消费频率、高频消费产品、活跃度平均水平、信誉度平均水平、信誉等级分布”的用户画像：

年龄中位数：调用$median$函数，计算数据集中$age$数据项的中位数；

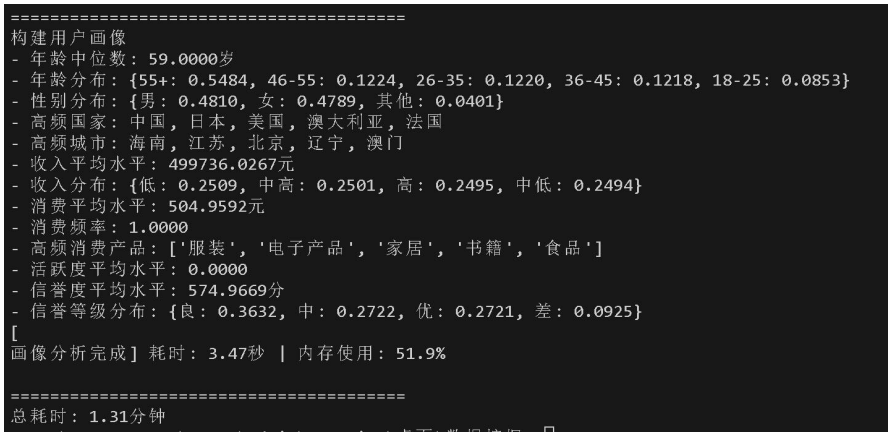
年龄分布：$df['age\_group'].value\_counts(normalize = True).to\_dict()$，按照数据预处理中对$age$数据项的分组处理结果，计算年龄分组情况；

性别分布：$df['gender'].value\_counts(normalize=True).to\_dict()$，按照数据预处理中对$gender$数据项的分组处理结果，计算性别分组情况；

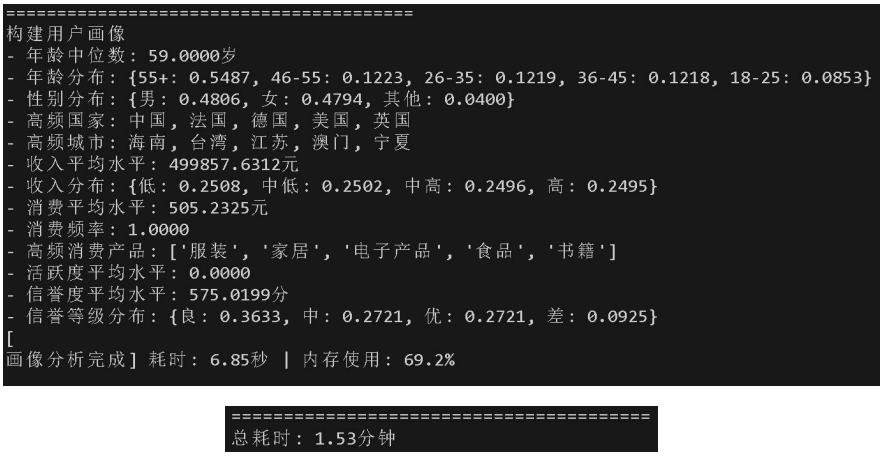
高频国家、高频城市：$df['country'].value\_counts().head(5).index.tolist()$，统计数据集中出现次数最多的前五个国家和城市；

收入平均水平：调用$mean$函数，计算收入的平均值；

消费频率：$len(valid\_amounts) / len(df)$，计算数据集中用户的消费频率。



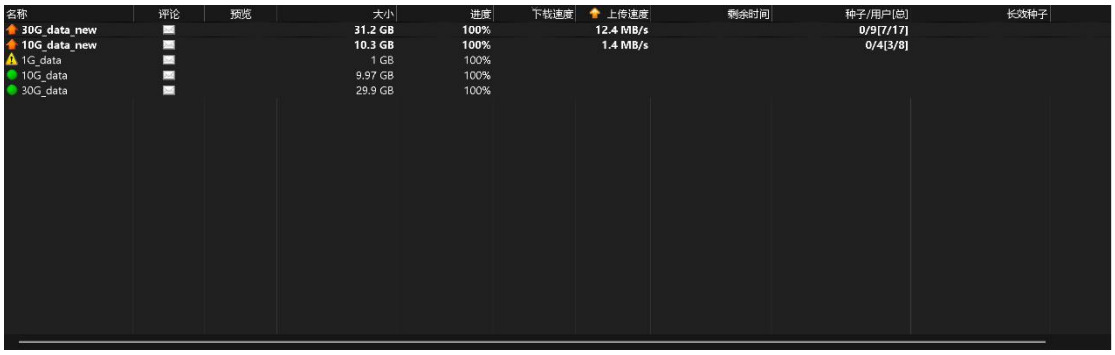

10G 数据集用户画像




30G 数据集用户画像

# 实验结果总览

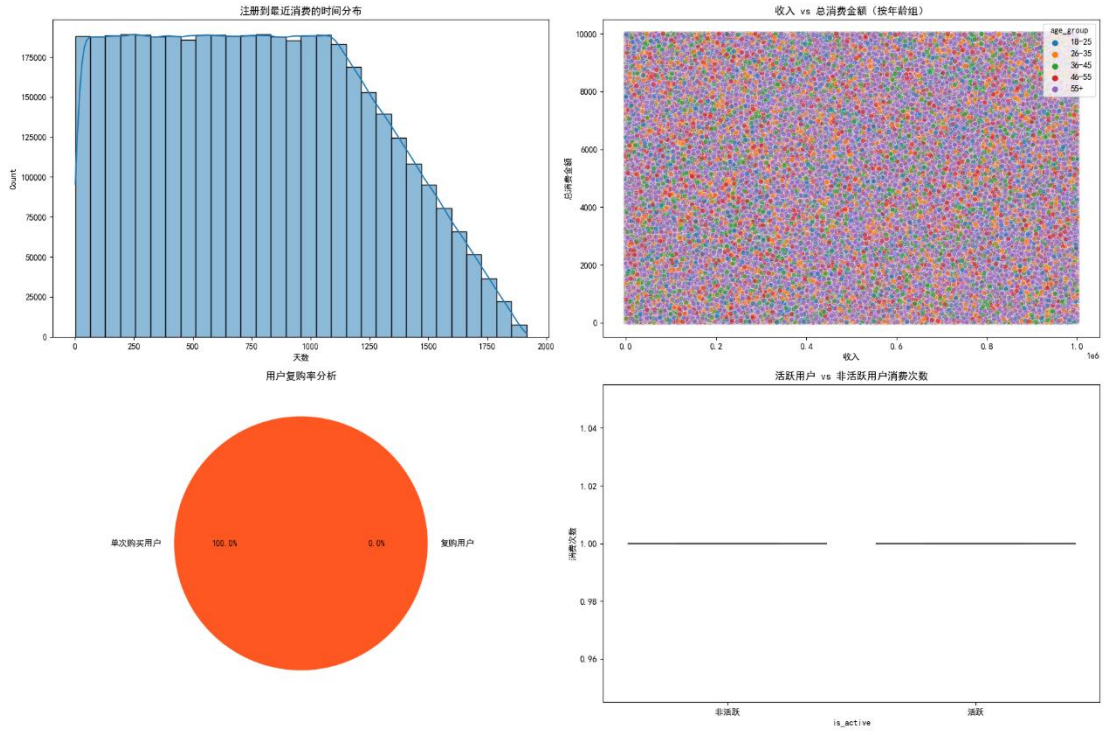
## 下载数据集（上传）

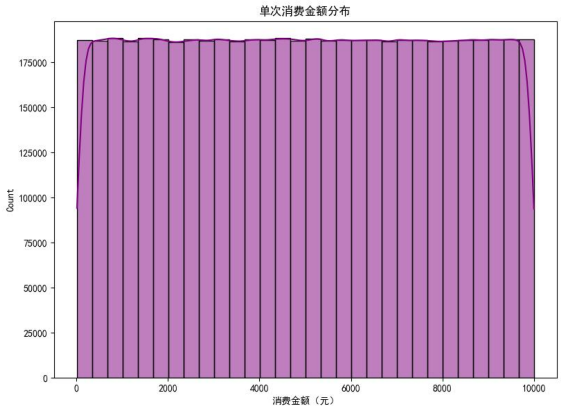


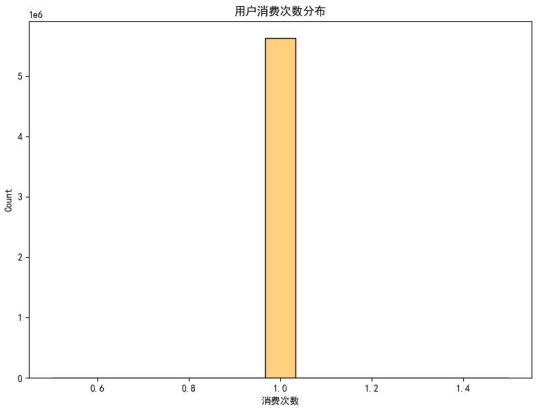

## 对 10G 数据集进行处理

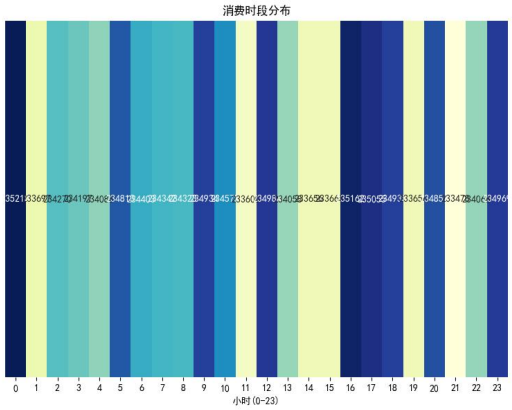
```
==========================================
10G数据集预处理开始

[缺失值统计]
                          缺失数量  缺失比例(%)
id                           0       0.0
last_login                   0       0.0
age                          0       0.0
income                       0       0.0
gender                       0       0.0
country                      0       0.0
address                      0       0.0
purchase_history             0       0.0
is_active                    0       0.0
registration_date            0       0.0
login_history                0       0.0
last_login_year              0       0.0
last_login_month             0       0.0
last_login_day               0       0.0
registration_date_year       0       0.0
registration_date_month      0       0.0
registration_date_day        0       0.0
```

```
[重复值统计] 0
[删除重复值后数据量统计] 5625000
[预处理完成] 耗时: 103.89秒 | 内存使用: 98.3%

可视化分析 10G数据集
- 平均登录频率: 6.0次
- 用户复购率: 0.00%
- 中位生命周期: 552.0天
- 用户留存分布: >1年: 0.5706, 3-12月: 0.3341, 1-3月: 0.0731, 1周-1月: 0.0222, <1周: 0.0000

========================================
构建用户画像
- 年龄中位数: 59.0000岁
- 年龄分布: {55+: 0.5487, 36-45: 0.1222, 26-35: 0.1220, 46-55: 0.1217, 18-25: 0.0853}
- 性别分布: {男: 0.4800, 女: 0.4800, 其他: 0.0400}
- 高频国家: 日本, 巴西, 美国, 法国, 印度
- 高频城市: 新疆, 香港, 广西, 台湾, 山东
- 收入平均水平: 500068.7920元
- 收入分布: {高: 0.2500, 中高: 0.2500, 中低: 0.2500, 低: 0.2500}
- 消费平均水平: 5003.8467元
- 消费频率: 1.0000
- 高频消费产品: ['智能手表', '上衣', '裤子', '益智玩具', '米面']
- 活跃度平均水平: 0.4998
[
画像分析完成] 耗时: 16.13秒 | 内存使用: 89.3%

========================================
总耗时: 7.11分钟
```

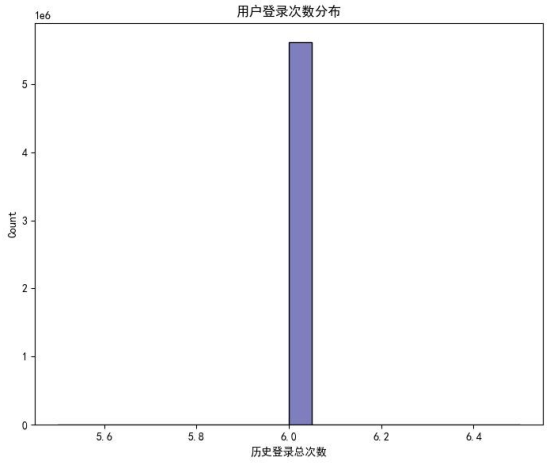


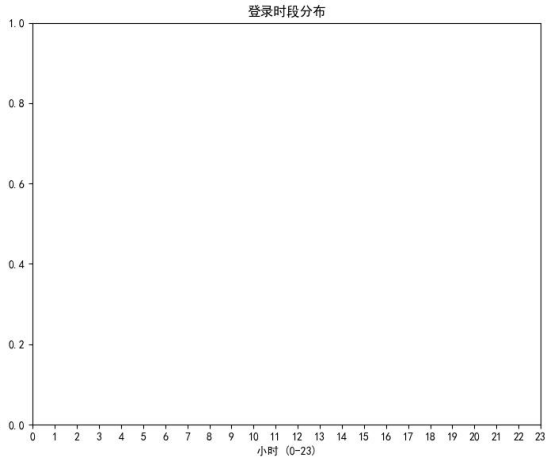

单次消费金额分布
Count
消费金额（元）

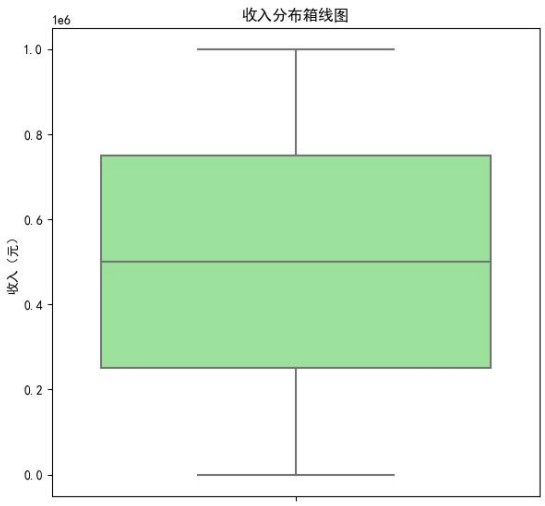

用户消费次数分布
1e6
Count
消费次数

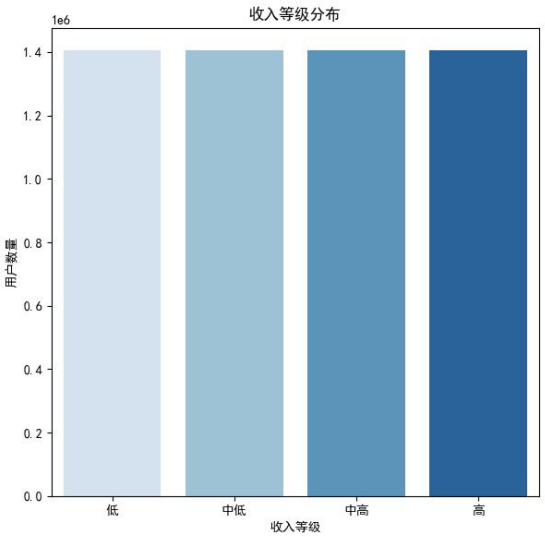

消费时段分布
小时(0-23)

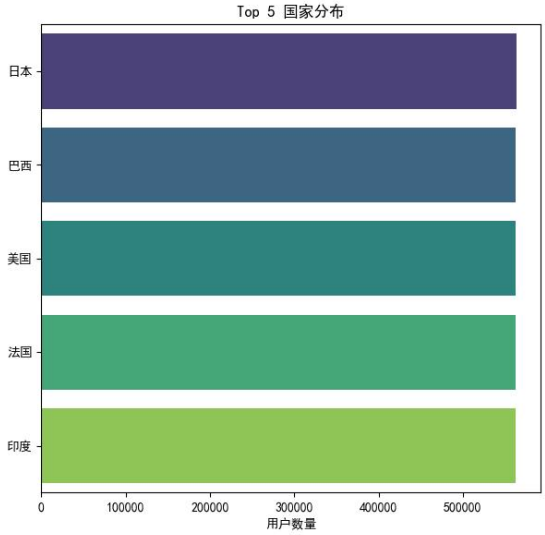

消费类别词云
智能手表
益智玩具
上衣裤子
米面

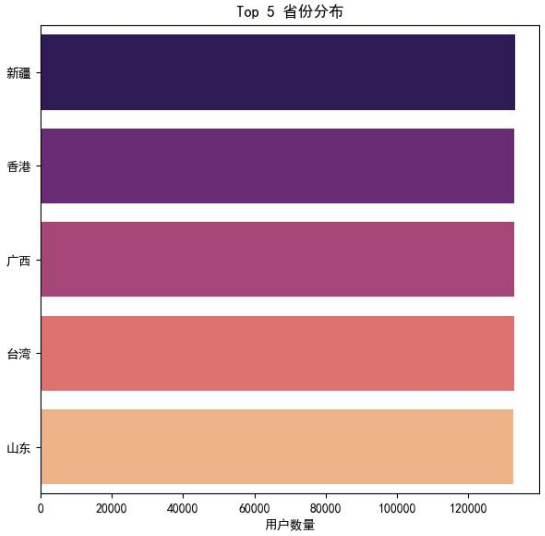

用户登录次数分布
1e6
Count
历史登录总次数

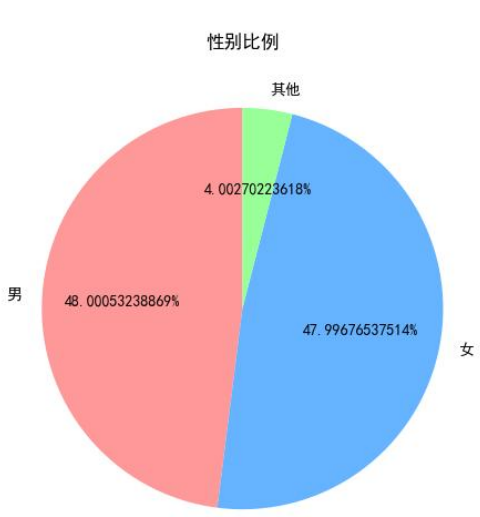

登录时段分布
小时（0-23）

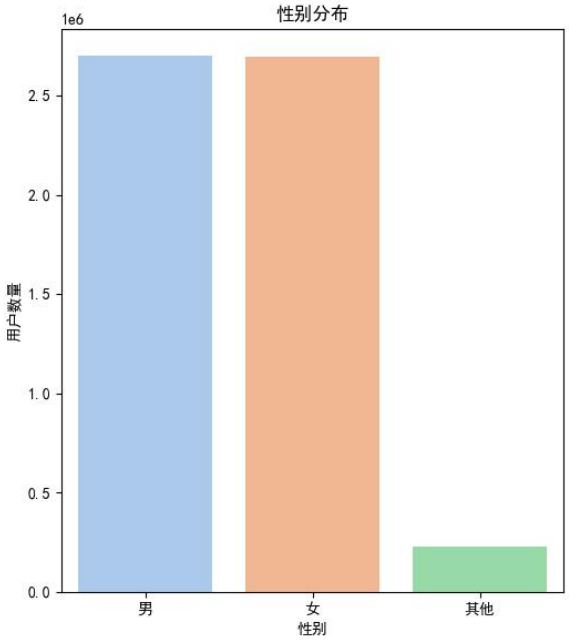

收入分布箱线图
1e6
1.0
0.8
0.6
0.4
0.2
0.0
收入（元）

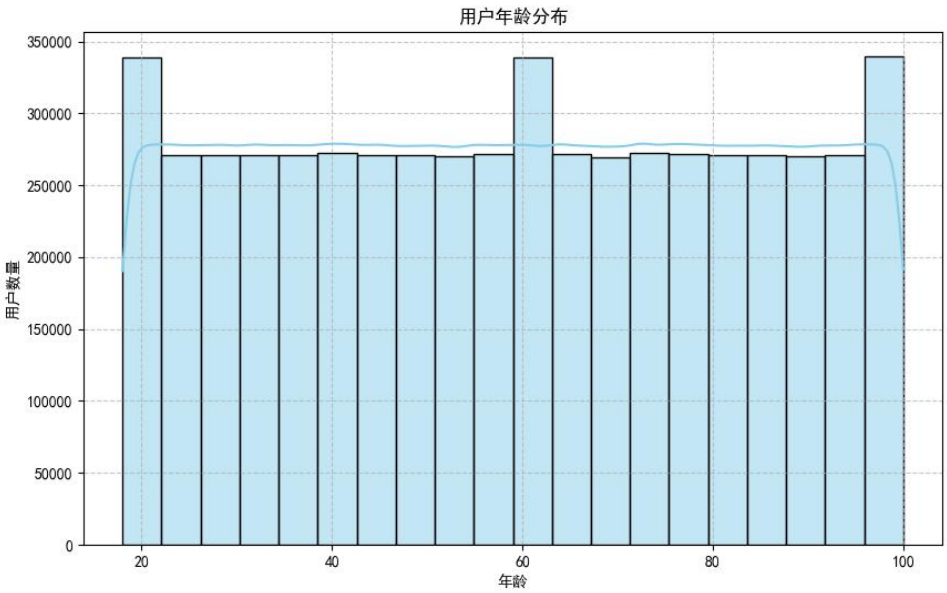

收入等级分布
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
用户数量
低
中低
中高
高
收入等级

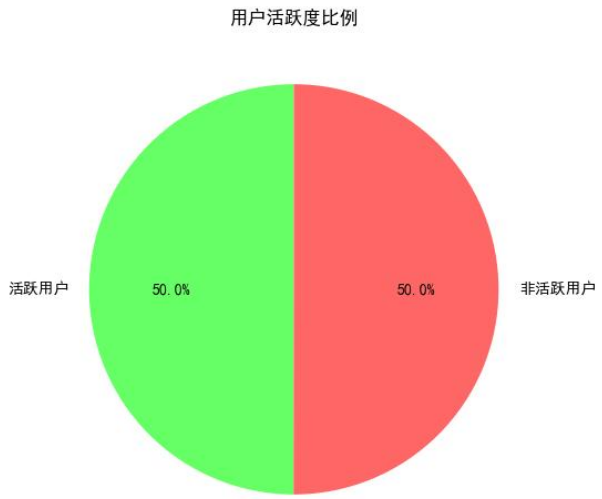

Top 5 国家分布
日本
巴西
美国
法国
印度
0
100000
200000
300000
400000
500000
用户数量

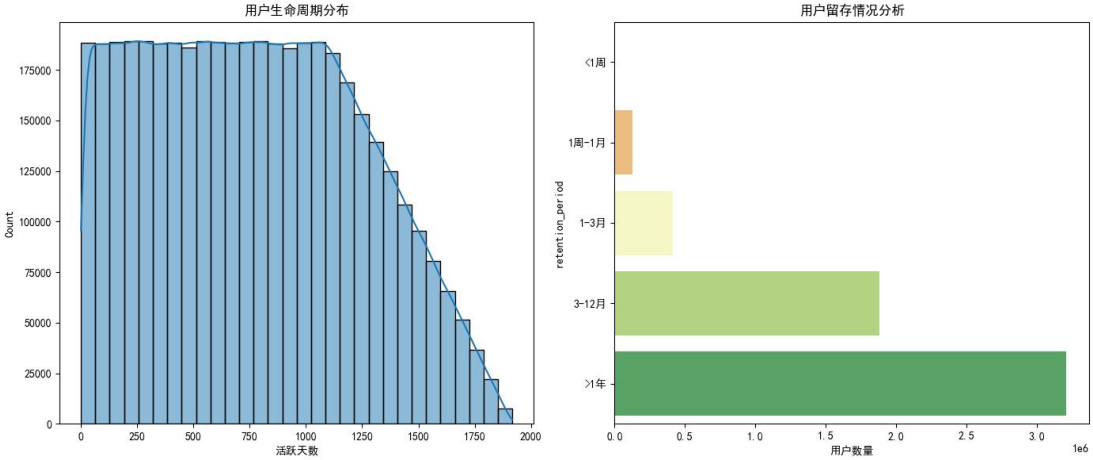

Top 5 省份分布
新疆
香港
广西
台湾
山东
0
20000
40000
60000
80000
100000
120000
用户数量


性别比例
其他
4.00270223618%
男
48.00053238869%
47.99676537514%
女


性别分布
1e6
2.5
2.0
1.5
1.0
0.5
0.0
用户数量
男
女
其他
性别


用户年龄分布
350000
300000
250000
200000
150000
100000
50000
0
用户数量
20
40
60
80
100
年龄

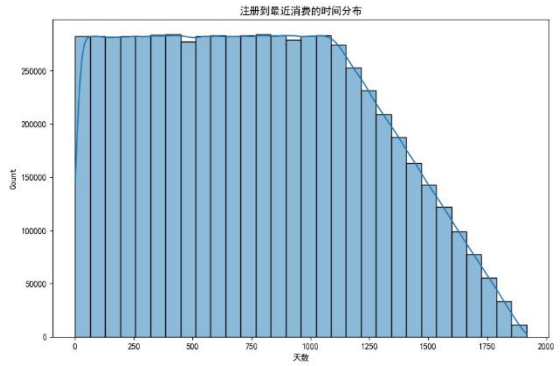

用户活跃度比例
活跃用户
50.0%
50.0%
非活跃用户

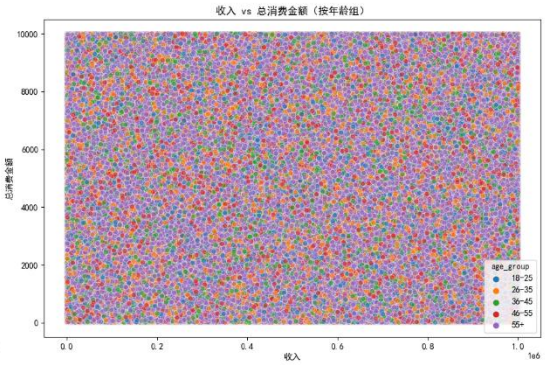

用户生命周期分布
175000
150000
125000
100000
75000
50000
25000
0
Count
0
250
500
750
1000
1250
1500
1750
2000
活跃天数
用户留存情况分析
<1周
1周-1月
1-3月
3-12月
>1年
retention_period
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1e6
用户数量

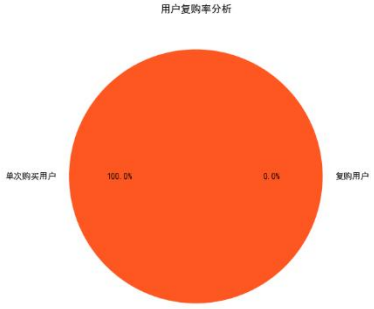
对 30G 数据集进行处理

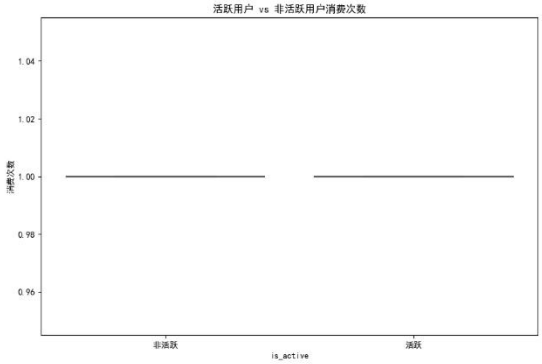
```
==========================================
30G数据集预处理开始

[缺失值统计]
                          缺失数量  缺失比例(%)
id                           0      0.0
last_login                   0      0.0
age                          0      0.0
income                       0      0.0
gender                       0      0.0
country                      0      0.0
address                      0      0.0
purchase_history             0      0.0
is_active                    0      0.0
registration_date            0      0.0
login_history                0      0.0
last_login_year              0      0.0
last_login_month             0      0.0
last_login_day               0      0.0
registration_date_year       0      0.0
registration_date_month      0      0.0
registration_date_day        0      0.0
```

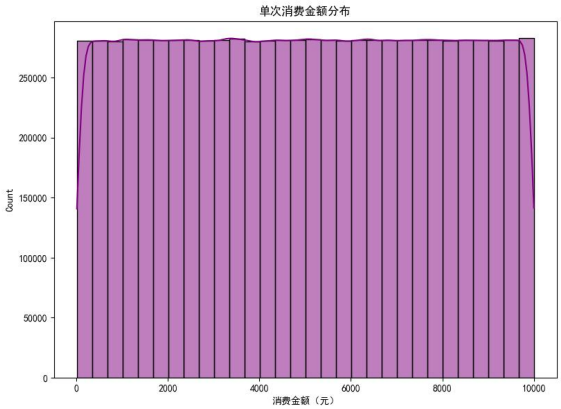
```
[重复值统计] 0
[删除重复值后数据量统计] 8437500
[预处理完成] 耗时: 171.65秒 | 内存使用: 95.7%

可视化分析 30G数据集
- 平均登录频率: 6.0次
- 用户复购率: 0.00%
- 中位生命周期: 553.0天
- 用户留存分布: >1年: 0.5702, 3-12月: 0.3345, 1-3月: 0.0730, 1周-1月: 0.0223, <1周: 0.0000

========================================
构建用户画像
- 年龄中位数: 59.0000岁
- 年龄分布: {55+: 0.5490, 46-55: 0.1220, 36-45: 0.1219, 26-35: 0.1219, 18-25: 0.0852}
- 性别分布: {男: 0.4803, 女: 0.4797, 其他: 0.0400}
- 高频国家: 法国, 俄罗斯, 美国, 日本, 中国
- 高频城市: 湖南, 江西, 天津, 福建, 青海
- 收入平均水平: 499903.4821元
- 收入分布: {高: 0.2500, 中高: 0.2500, 中低: 0.2500, 低: 0.2500}
- 消费平均水平: 5005.9562元
- 消费频率: 1.0000
- 高频消费产品: ['耳机', '饮料', '汽车装饰', '零食', '智能手表']
- 活跃度平均水平: 0.5000
[
画像分析完成] 耗时: 45.82秒 | 内存使用: 94.7%

========================================
总耗时: 12.00分钟
```

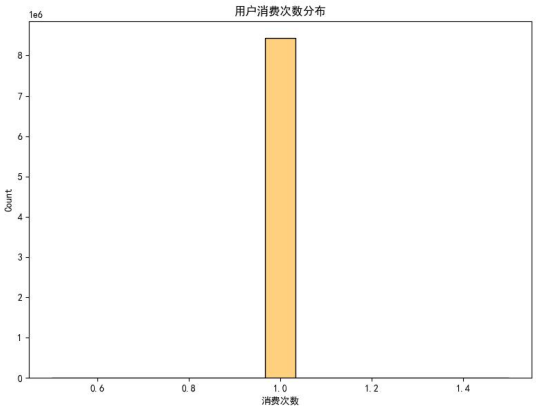


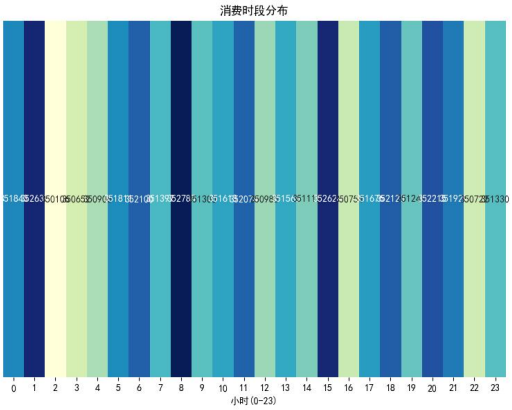



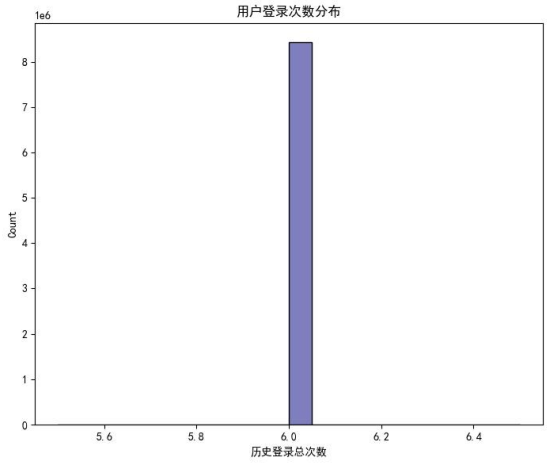



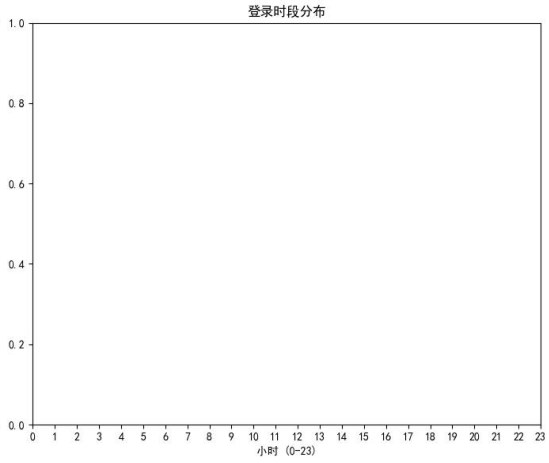

单次消费金额分布
Count
消费金额（元）

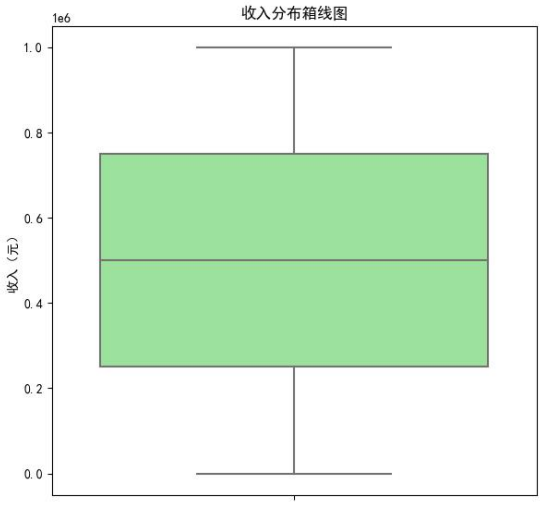

用户消费次数分布
1e6
Count
消费次数

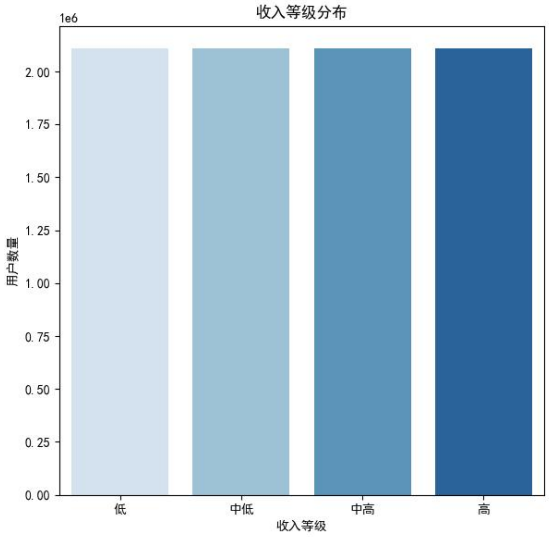

消费时段分布
小时(0-23)

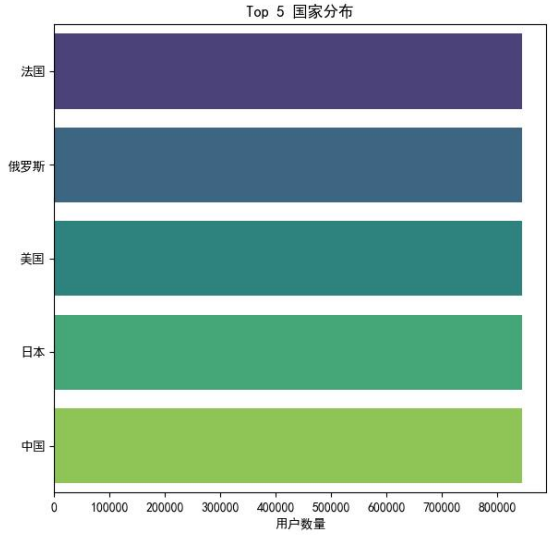

消费类别词云
耳机
汽车装饰
智能手表
零食
饮料

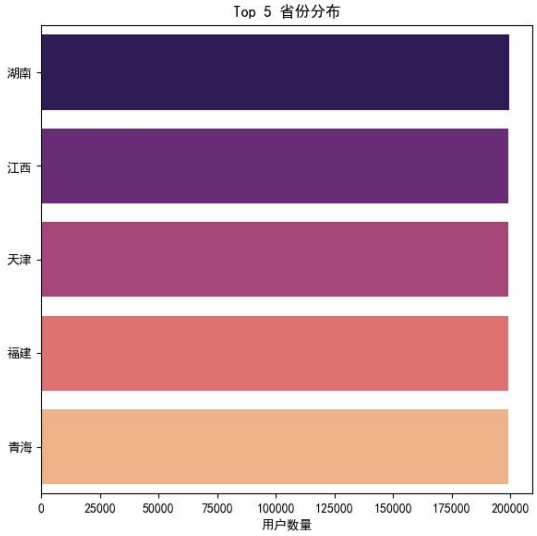

用户登录次数分布
1e6
Count
历史登录总次数

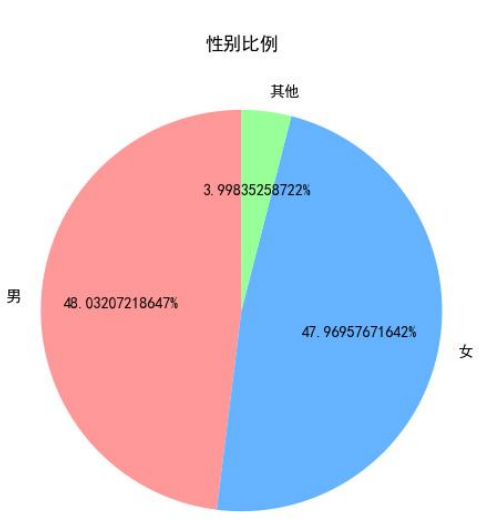

登录时段分布
小时（0-23）

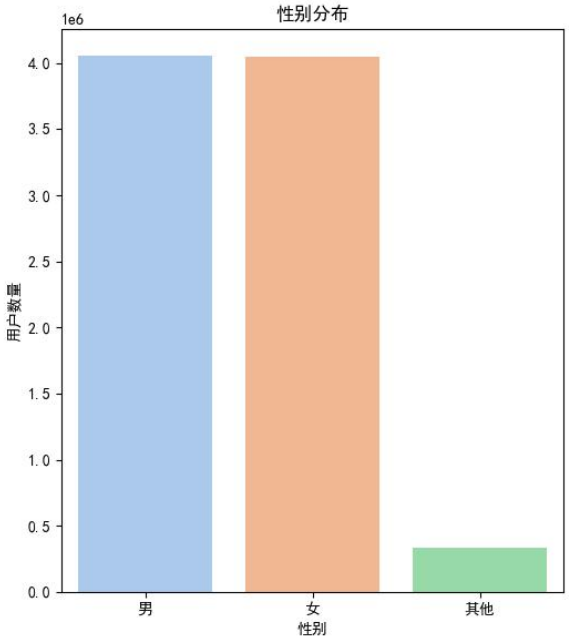

收入分布箱线图
1e6
1.0
0.8
0.6
0.4
0.2
0.0
收入（元）

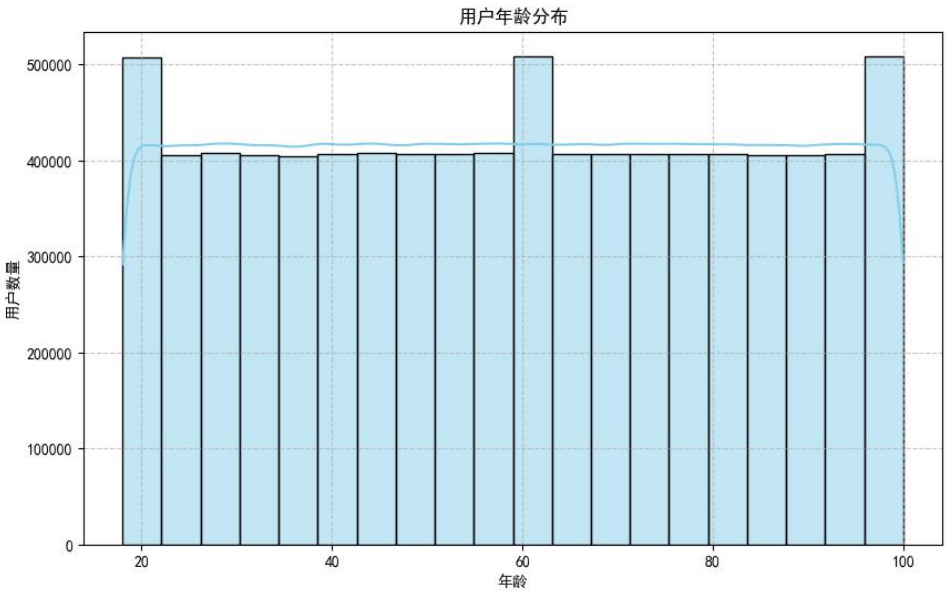

收入等级分布
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
用户数量
低
中低
中高
高
收入等级

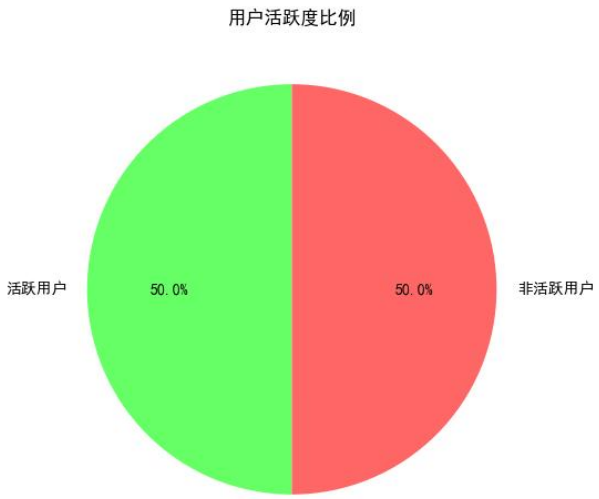

Top 5 国家分布
法国
俄罗斯
美国
日本
中国
0
100000
200000
300000
400000
500000
600000
700000
800000
用户数量

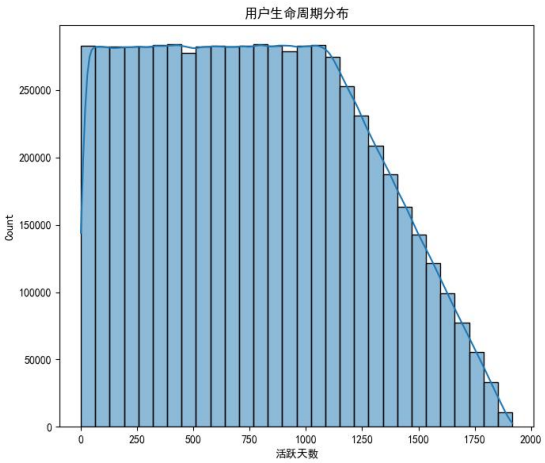

Top 5 省份分布
湖南
江西
天津
福建
青海
0
25000
50000
75000
100000
125000
150000
175000
200000
用户数量

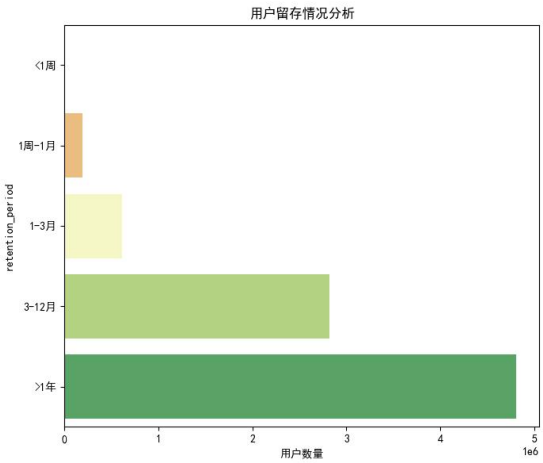

性别比例
其他
3.99835258722%
男
48.03207218647%
47.96957671642%
女


性别分布
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
用户数量
男
女
其他
性别


用户年龄分布
用户数量
年龄
0
100000
200000
300000
400000
500000
20
40
60
80
100


用户活跃度比例
活跃用户
50.0%
50.0%
非活跃用户


用户生命周期分布
Count
活跃天数
0
50000
100000
150000
200000
250000
0
250
500
750
1000
1250
1500
1750
2000


用户留存情况分析
retention_period
<1周
1周-1月
1-3月
3-12月
>1年
用户数量
1e6
0
1
2
3
4
5