BiBDR | 博雅大数据学院 BOYA BIG DATA COLLEGE | 数据酷客 COOKDATA

北京市大数据人才培训示范基地

数据科学“云实训”项目训练营

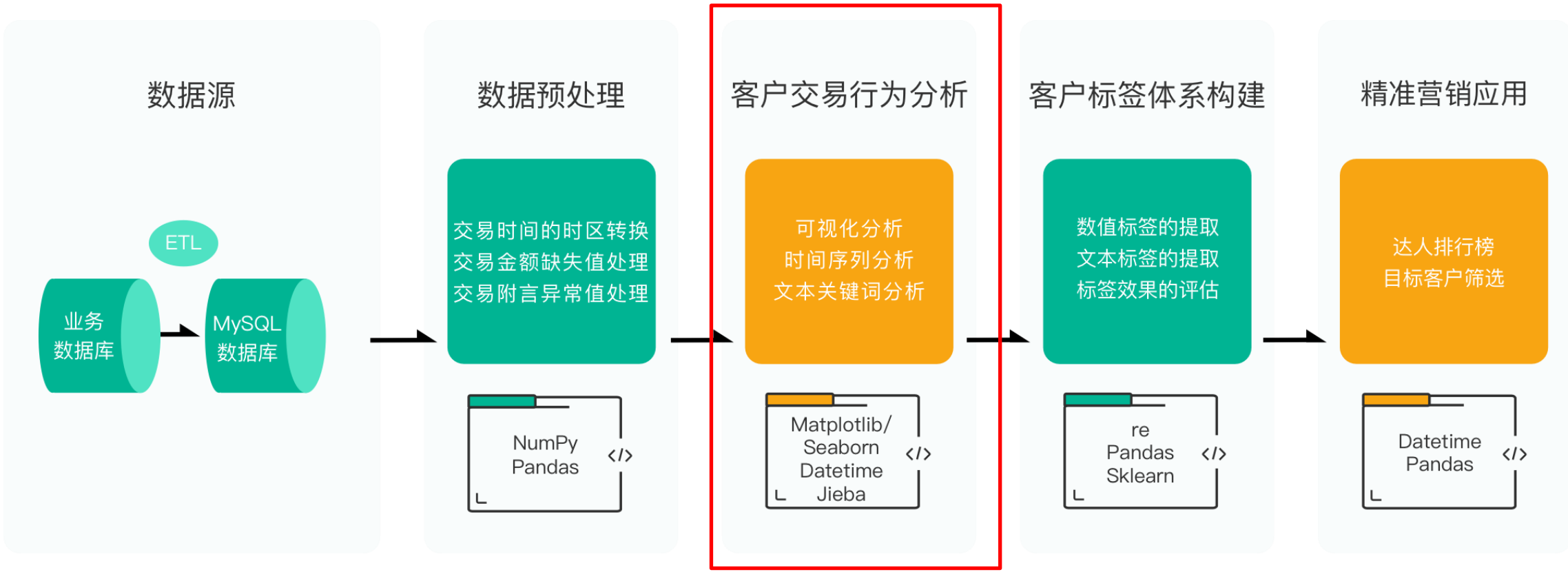
第三课：客户交易行为分析

讲师：吴佳佳

数据酷客

数据科学人工智能

项目流程


数据源
ETL
业务
数据库
MySQL
数据库
数据预处理
交易时间的时区转换
交易金额缺失值处理
交易附言异常值处理
NumPy
Pandas
客户交易行为分析
可视化分析
时间序列分析
文本关键词分析
Matplotlib/
Seaborn
Datetime
Jieba
客户标签体系构建
数值标签的提取
文本标签的提取
标签效果的评估
re
Pandas
Sklearn
精准营销应用
达人排行榜
目标客户筛选
Datetime
Pandas

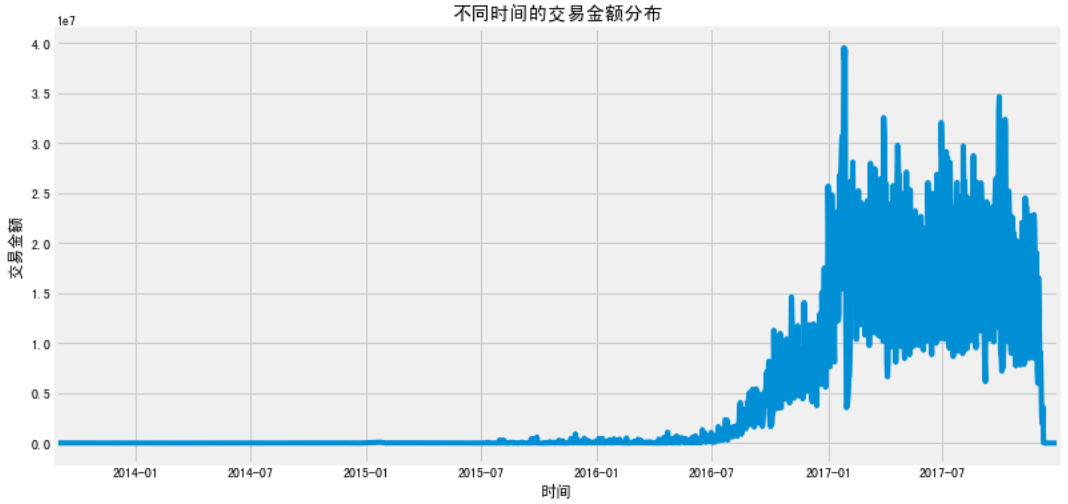
目录

本部分主要包括以下四个步骤：

- 时间维度的分析：对交易时间进行分析，探索交易随时间的分布规律
- 交易属性的分析：对交易金额和次数进行分析，探索不同客户的交易属性
- 文本数据预处理：为了便于后续分析，对交易附言的文本进行预处理
- 文本数据的分析：对预处理后的文本进行分析，例如绘制词云分布图和提取关键词等

时间维度的分析

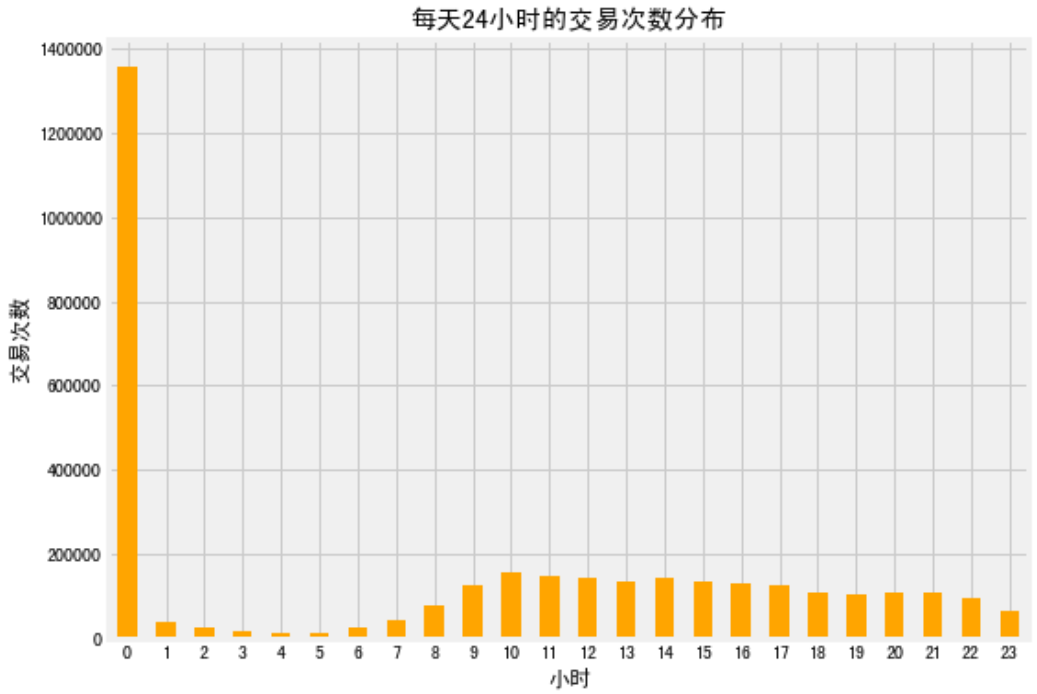
总交易次数和总交易金额随时间变化的折线分布图如下：



交易主要集中在2016年7月至2017年12月在本项目后续环节，我们选取该时间段的交易记录进行重点分析

时间维度的分析

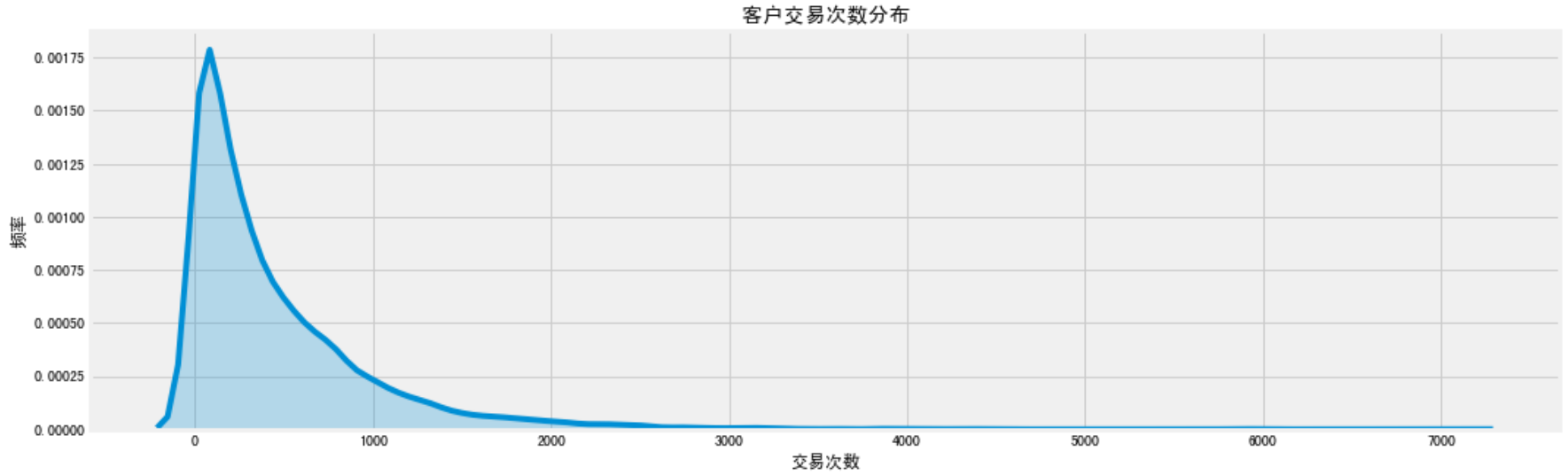
统计每天24小时的各时间段交易次数：



- 大部分交易的时间集中在0点附近，这可能是金融机构进行清算的时间段
- 凌晨1点至7点交易数量较少，其余时间段交易数量分布较为均衡

交易属性的分析

从每个客户出发，可视化分析不同客户的交易次数分布：

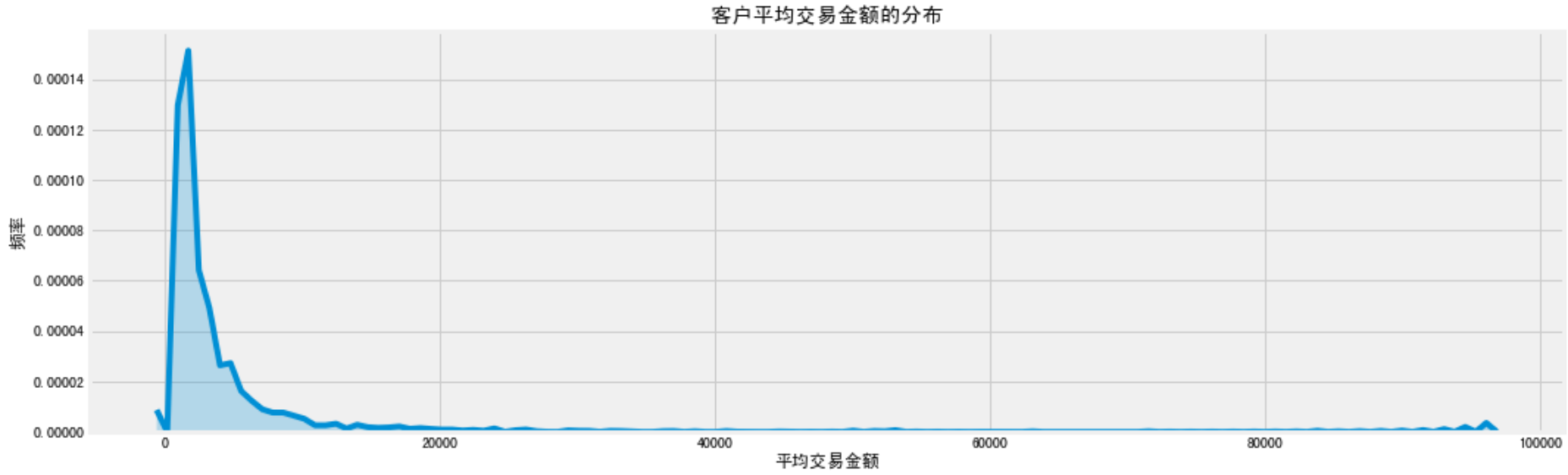


可以看到客户的交易次数从0-7000不等，分布主要集中在0-1000在之后的指标构建中，

根据客户的交易记录构建指标，如果在此平台交易记录太少，我们认为此类为休眠客户

交易属性的分析

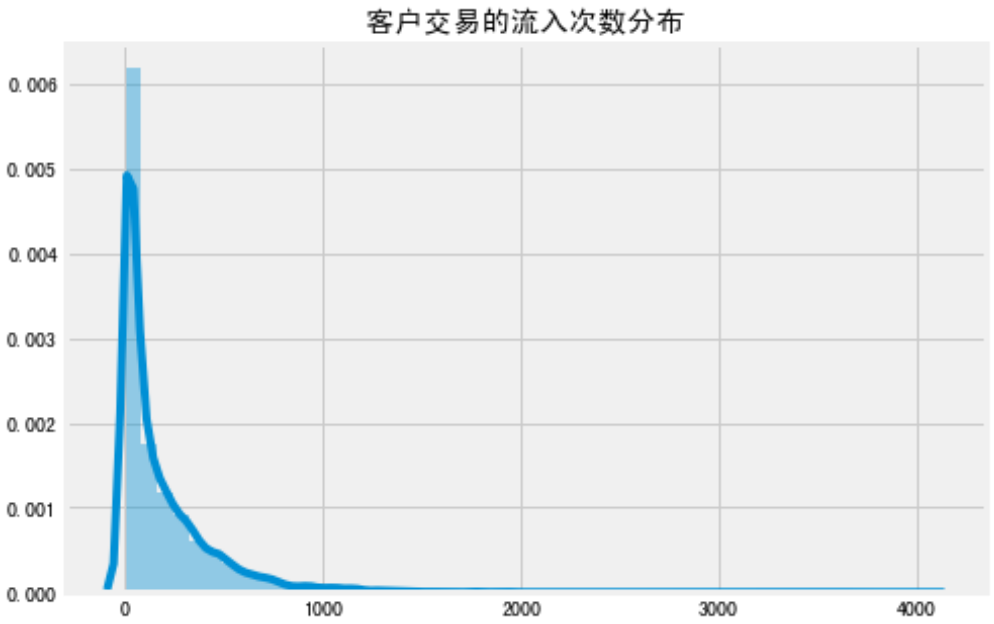
从每个客户出发，可视化分析不同客户的平均交易金额分布：

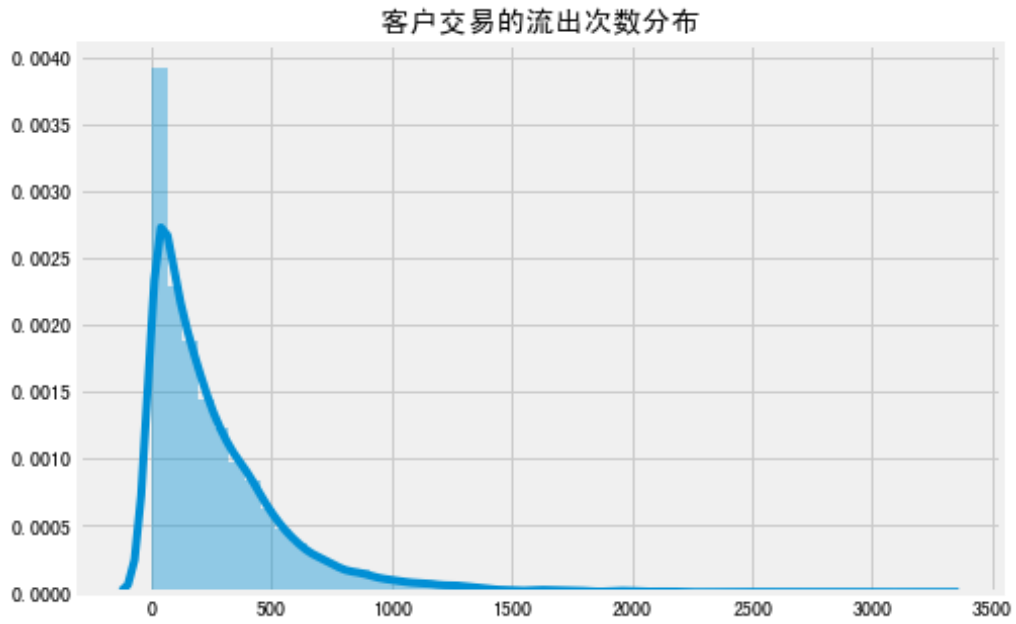


可以看到客户的平均交易金额跨度较大，分布主要集中在0~10000，呈长尾分布

交易属性的分析

从每个客户出发，可视化分析不同客户的流入流出次数分布：

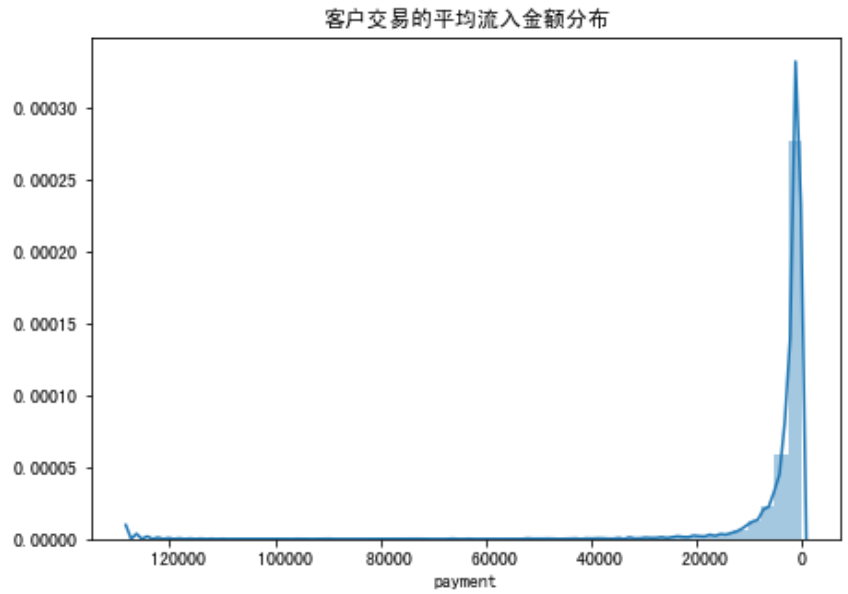


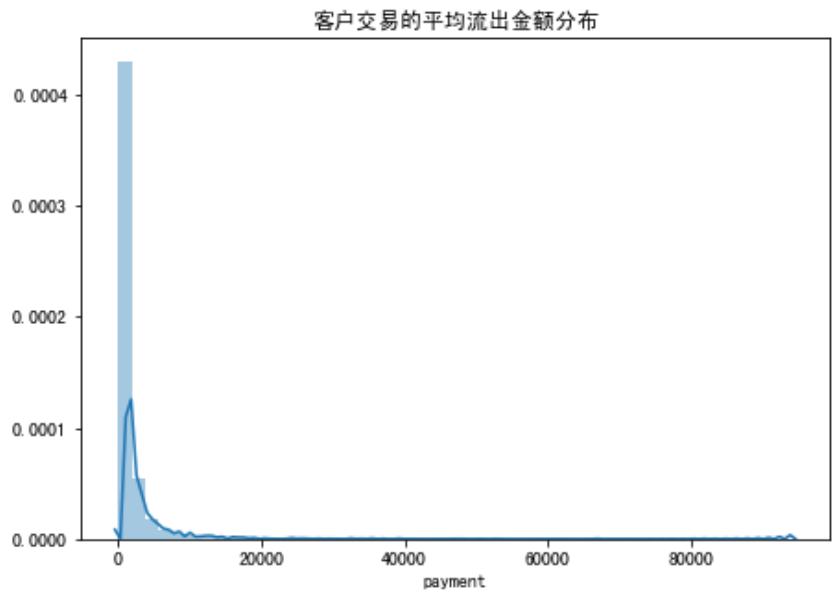


可以看到客户整体的流入流出次数均集中在0-1000次

交易属性的分析

从每个客户出发，可视化分析不同客户的流入流出金额分布：





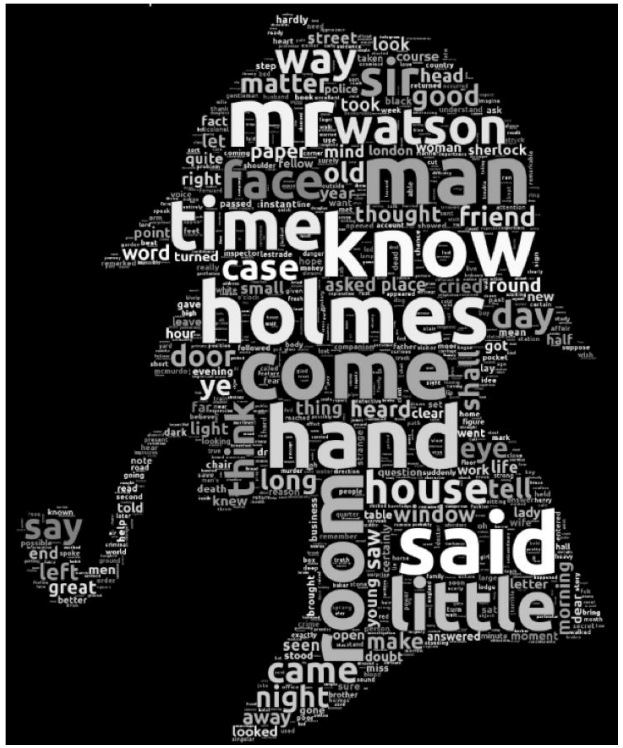
可以看到流入流出平均金额均集中在0-20000之间，金额流入的跨度比金额流出的跨度略大

文本数据预处理

- 中文文本分词：将中文的句子切分成有意义的词语
- 去除停用词：根据事先设置好的停用词，规避掉一些特殊符号或者常用但无意义的词语
- jieba库可以进行很多中文文本的处理，包括分词、去停用词、关键词抽取和词性标注等

词云图

词云是一种可视化描绘词语出现在文本数据中频率的方式，以词频数为每个词分配权重，用字体的大小、粗细来表示权重的高低。福尔摩斯探案集词云图如下：



词云图的绘制

使用Python词云库wordcloud绘制交易附言的词云图：

可以看到文本内容主要集中在消费、转账、购物等方向在构建客户标签体系和画像时，

我们着重从这些方面对客户进行刻画，分析其中的规律

文本关键词的提取

tf-idf(term frequency–inverse document frequency)是一种用于信息检索与数据挖掘的加权技术，常用于挖掘文章中的关键词

- 词频(tf - term frequency)：
  词在文本中的出现的频率，代表了词在文档中的重要程度

- 逆文档频率(idf - inverse document frequency)：
  语料库中文档总数与包含当前词的文档数的比值，如果一个关键词在很少的文本中出现，通过它更容易锁定目标，那么权重应该较大

tf-idf算法

$$tf(t) = \frac{\text{词}t\text{在文档中出现的次数}}{\text{文档的总词数}}$$

$$idf(t) = \log\left(\frac{\text{语料库中文档的总数}}{\text{包含词}t\text{的文档数}+1}\right)$$

$$tf\text{-}idf(t) = tf(t) \cdot idf(t)$$

- 分母之所以要加1，是为了避免分母为0（即所有文档都不包含该词）
- log表示对得到的值取对数，在编程语言中我们通常取自然对数e

文本关键词的提取

('充值', 0.558863427344304)
('支付宝', 0.3475876939471043)
('转账', 0.3168968658650 1015)
('淘宝', 0.226583908354 8234)
('购物', 0.212031856358 1532)
('支付', 0.187322078363 40334)
('消费', 0.129582113243 0632)
('余额', 0.113736043599 5011)
('还款', 0.110796945671 86835)
('提现', 0.103674606282 06764)
('ATM', 0.0845512512961 2921)
……

可以看到，排名前50的关键词主要集中在客户的消费、转账、购物等方向，和词云的结论基本相同

数据酷客

数据科学人工智能

加入数据酷客交流群