商户流失率预测-实验报告

梅佳奕 10165300206

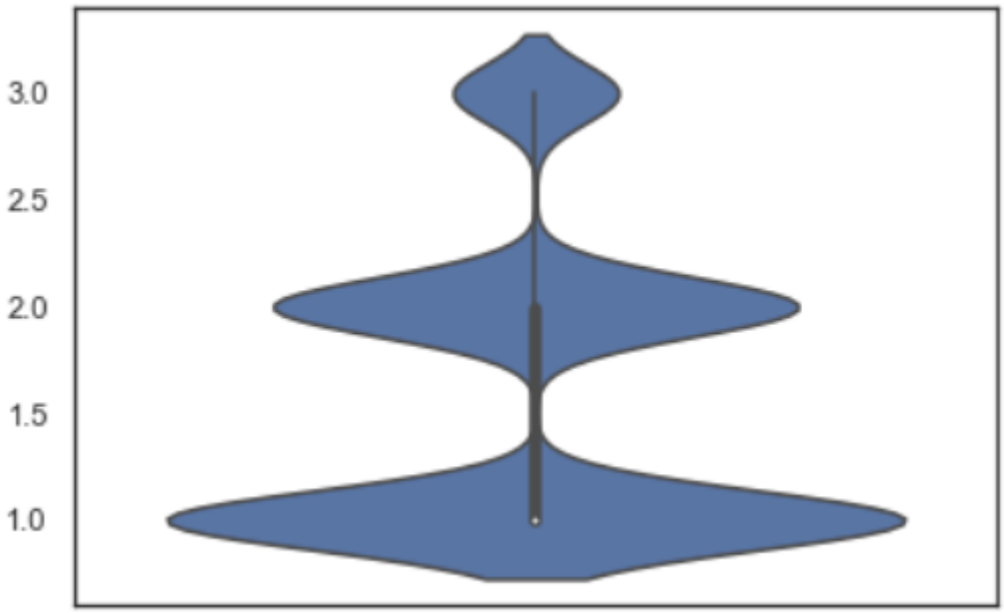
阶段二：特征值制作

经过提取，得到以下特征：

- **table merchant:**

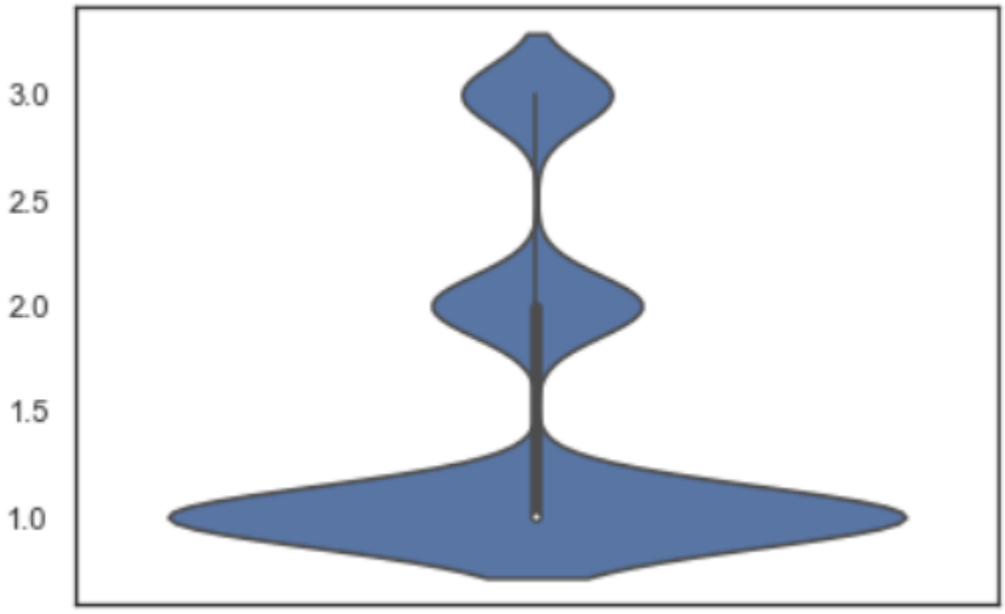
表名：table_datamining_merchant	
列名	comment
merchant_id	商户ID
industry_level1	一级行业 [1,2,3]
province	省份 [1,2,3]

1. industry_level 行业等级 $\in \{1, 2, 3\}$



2. province 省份 $\in \{1, 2, 3\}$

1、2来自merchant信息表，是merchant的属性，在merchant_id与transaction表merge得到。由背景信息得：一个merchant_id可能对应多个store_id，所以一个store_id对应唯一确定的 **industry_level** 、 **province** 。



- **table transaction:**

  根据阶段一的分析，在该表中保留可用的字段为(以颜色标出）：

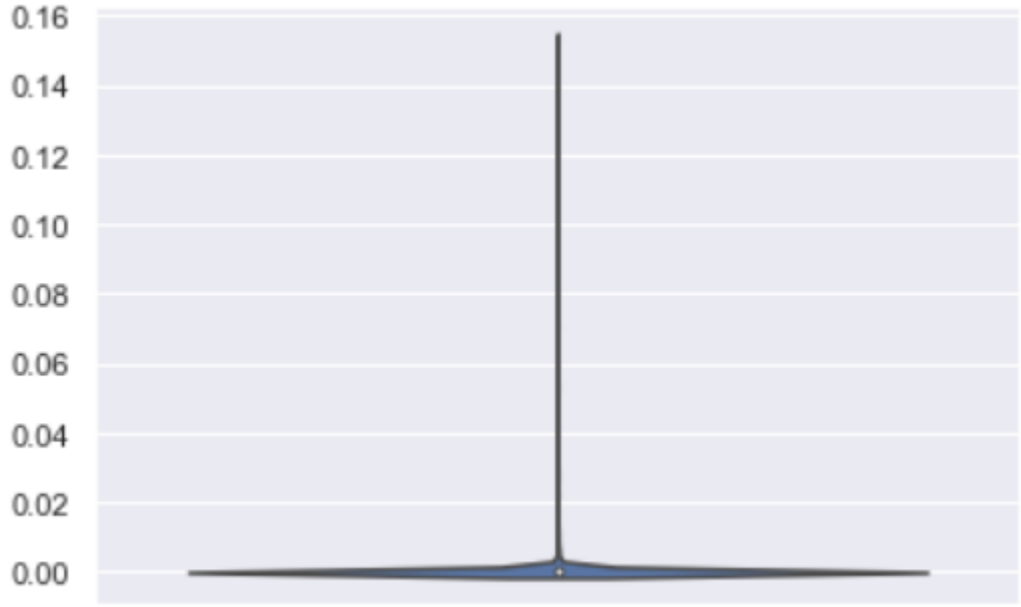
表名：table_transaction	
列名	comment
id	交易id
type	交易类型 30-付款 10-取消 11-退款
status	状态 0 创建 2000 成功 其他参见wiki
store_id	门店id
merchant_id	商户id
terminal_id	终端id
payer_uid	付款人在支付服务商的用户ID
pay_way	一级主支付方式
sub_pay_way	二级支付方式
effective_amount	向支付通道请求的金额
original_amount	交易金额
received_amount	商户实收金额
paid_amount	消费者实际支付金额
ctime	创建时间
bankcard_credit	信用卡支付金额
bankcard_debit	储蓄卡支付金额
wallet_weixin	微信余额支付金额
wallet_alipay	支付宝余额支付金额
wallet_alipay_finance	支付宝余额宝支付金额
alipay_huabei	花呗支付金额
alipay_point	集分宝支付金额
hongbao_wosai	喔嚓红包
hongbao_wosai_mch	喔嚓商户红包 免充值
discount_wosai	喔嚓立减
discount_wosai_mch	喔嚓商户立减 免充值
discount_channel	支付通道 折扣(立减优惠)
discount_channel_mch	折扣(立减优惠) 支付通道商户 免充值
discount_channel_mch_top_up	折扣(立减优惠) 支付通道商户 充值
hongbao_channel	支付通道红包
hongbao_channel_mch	支付通道商户红包 免充值
hongbao_channel_mch_top_up	支付通道商户红包 充值
card_pre	支付通道商户预付卡
card_balance	支付通道商户储值卡
merchant_sn	商户sn
is_liquidation_next_day	是否直清 0-否 1-是
trans_currency	交易币种
pt	日期分区

3. price_reduced_rate （支付通道中）优惠率 $\in (0,1)$

计算公式：

$$price\_reduced\_rate_{store\ x} = \frac{Count_{store\ x}(original\_amount - effective\_amount > 0)}{Count_{store\ x}(allrecords)}$$

通过**实付金额effective_amount**与**发起交易请求的金额original_amount**之间的差异，可知每一条交易**是否在支付通道中得到优惠**，再进行计数，能得到$store_x$所有交易纪录中发生优惠情况的比率。



从图可以看出，该特征维度的分布真的很稀疏，是典型的长尾分布，有效性待议。

4. reduce_range （支付通道中）折扣率 $\in (0,1)$

以$store_x$所有交易纪录中发生优惠情况的比率来估计优惠信息显然/不够，所以在引入$\#(original_amount - effective_amount)$具体金额，来评估支付通道中发生的优惠的力度。

计算公式：

$$reduce\_range_{store\ x} = \frac{Sum_{store\ x}(original\_amount - effective\_amount)}{Sum_{store\ x}(original\_amount)}$$

相当于是：将在支付通道中发生的优惠，转换成折扣力度。

需要注意的是，支付通道中产生的优惠是不以店铺折扣、活动形式张贴宣传写出的，有可能以”用户的积分抵扣“的形式存在，属于通道方（收钱吧）给出的优惠，店铺最终到账依然是original_amount；在这个过程中，惠及的是用户。考虑用户心理：趋于再次以受惠形式购买，影响作用于用户、当前交易发生后。

同样地，分布稀疏。有效性待议。

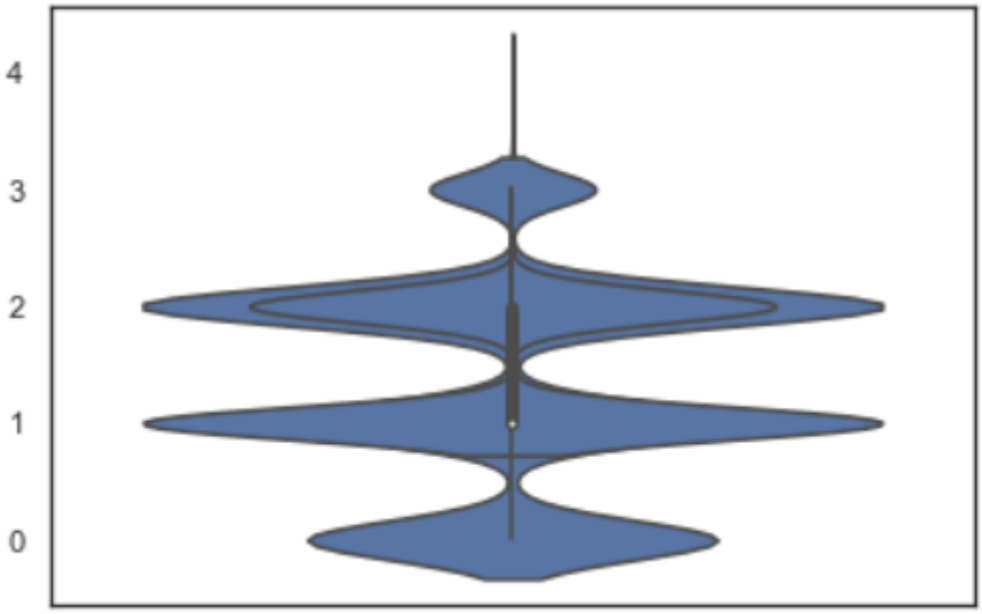
5. discount （店铺）折扣率 $\in N$

除了在支付通道中会产生优惠，店铺提供的满减、折扣、会员等活动对用户刺激更大，且通常直接作用于当前交易。与该特征有关的字段为：**discount_channel、discount_channel_mch、discount_channel_mch_top_up、hongbao_channel。**由于这些信息都极其稀疏（阶段一：$K_{col} > 0.9$），且金额较小，所以将它们弱化成bool类型以表示：店铺是否启用该优惠方式。

计算公式：

$$discount_{store\ s} = Count_{store\ x}(bool(discount\_channel) + bool(discount\_channel\_mch) + bool(discount\_channel\_mch\_top)\_up) + bool(hongbao\_channel)$$

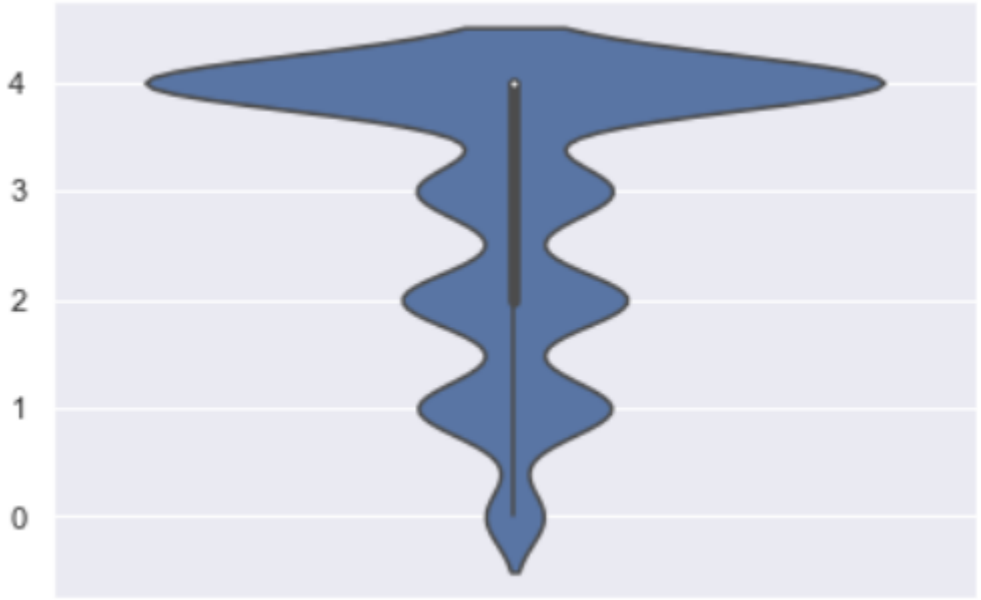
该特征的含义是店铺提供优惠活动的种类，即活动丰富程度。活动越丰富越容易吸引消费。



6. payw_offer 提供支付方式种类 $\in \{1, 2, 3, 4\}$

店铺提供支付方式的多寡将影响消费的便利性，降低/抬高消费门槛，从而影响营业额。经过阶段一的筛选，和次维度有关的字段为：**bankcard_debit、wallet_weixin、wallet_alipay、alipay_huabei。**

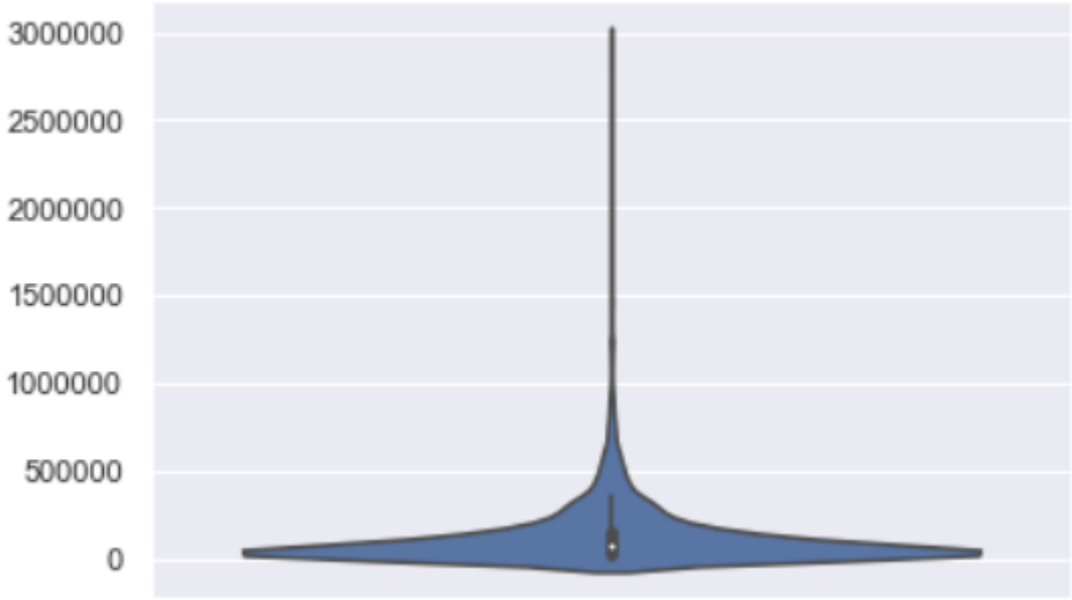
$$pay\_offer_{store\ x} = Count_{store\ x}[bool(bankcard\_debit) + bool(wallet\_weixin) + bool(wallet\_alipay) + bool(alipay_huabei)]$$



6. daily_amount 日均交易额 $\in [0, +\infty)$

以**original_amount**作为标准的交易额数据，计算店铺自有纪录营业的起始日期（训练集2018-04-01，测试集2018-06-01），至抛出点的日均交易额。

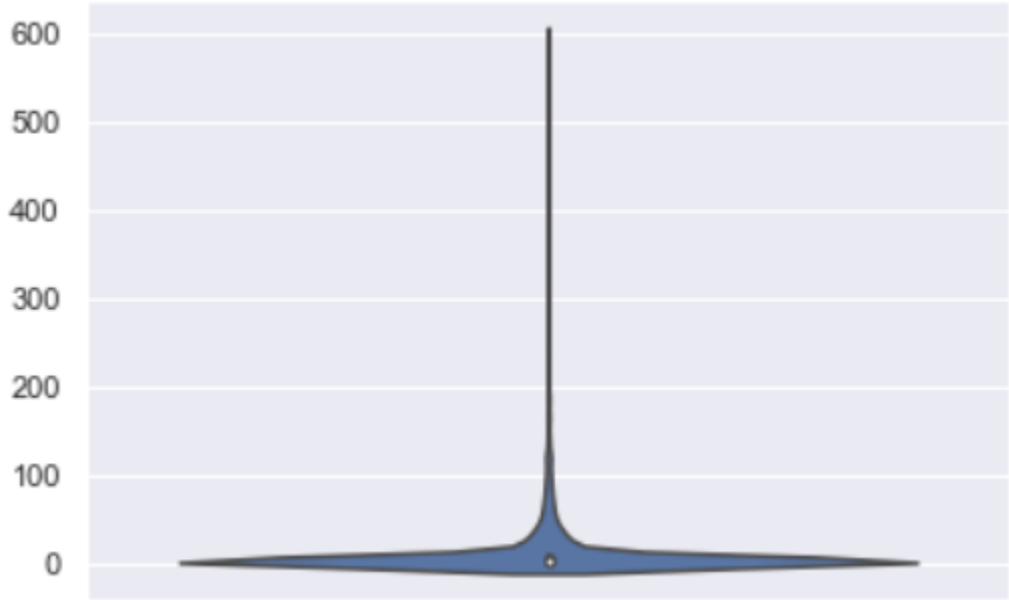
$$daily\_amount_{store\ x} = \frac{Sum_{store\ x}(original\_amount)}{\#days}$$



7. daily_count 日均交易笔数 $\in [0, +\infty)$

以**original_amount**作为标准的交易额数据，计算店铺自有纪录营业的起始日期（训练集2018-04-01，测试集2018-06-01），至抛出点的日均交易笔数。

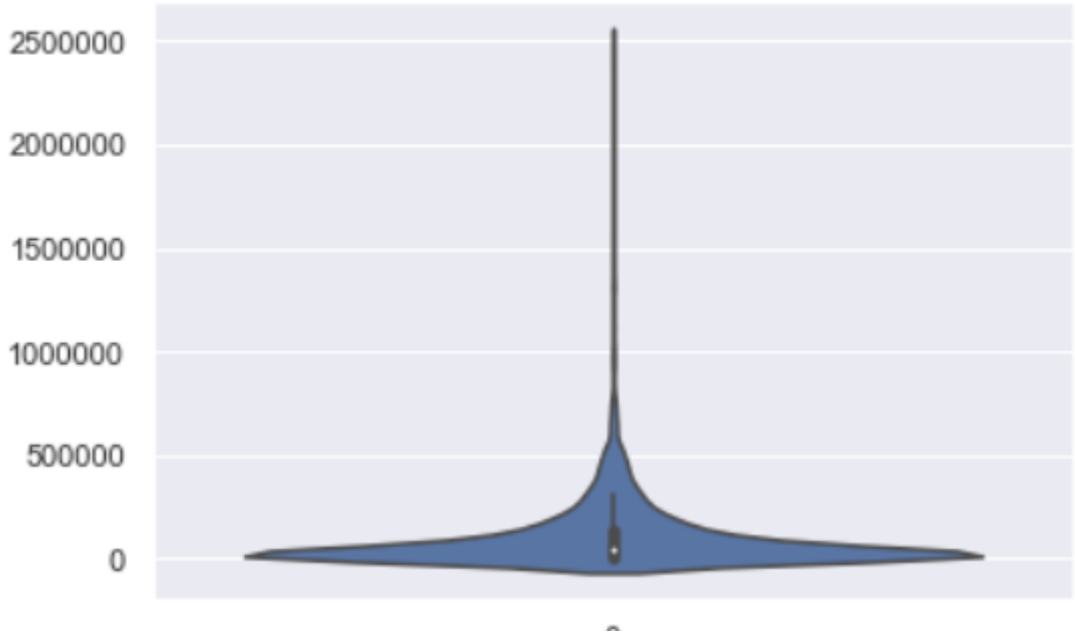
$$daily\_count_{store\ x} = \frac{Count_{store\ x}(original\_amount > 0)}{Count_{store\ x}(all)}$$



8. original_amount_mean 单笔交易均额 $\in [0, +\infty)$

鉴于店铺经营种类不同，用单笔交易均额的方式，可以区分出大宗买卖与小额零售的交易类型；即便对于同一类型的交易，我们常用"人均消费"的方法刻画它的受众群体。在这里，用**单笔交易金额**的信息来刻画这一维度，有助于对店铺类型、层级加以区分。

$$original\_amount\_mean_{store\ x} = Mean_{store\ x}(original\_amount)$$

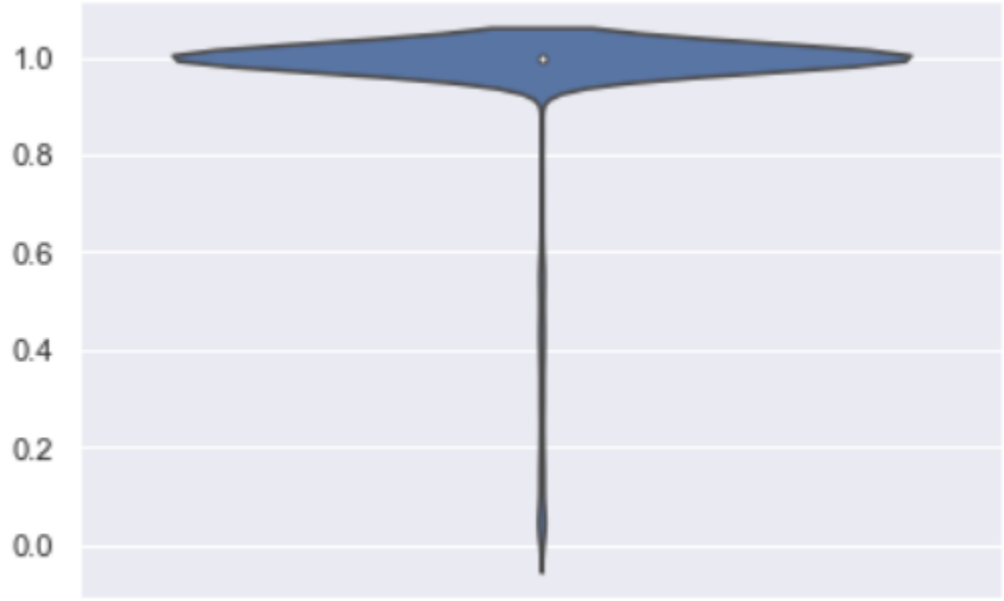


9. operate_state 经营状况 $\in [0,1]$

要考虑行业经营状况，需将店铺营业信息与行业信息联合考虑。店铺发展除了受到行业本身的影响（**industry_level**行业等级），还受到行业内竞争的影响；行业内竞争反应在经营状况（交易额）上，就是同等时间内，该行业市场中每家店铺分得"蛋糕"大小。

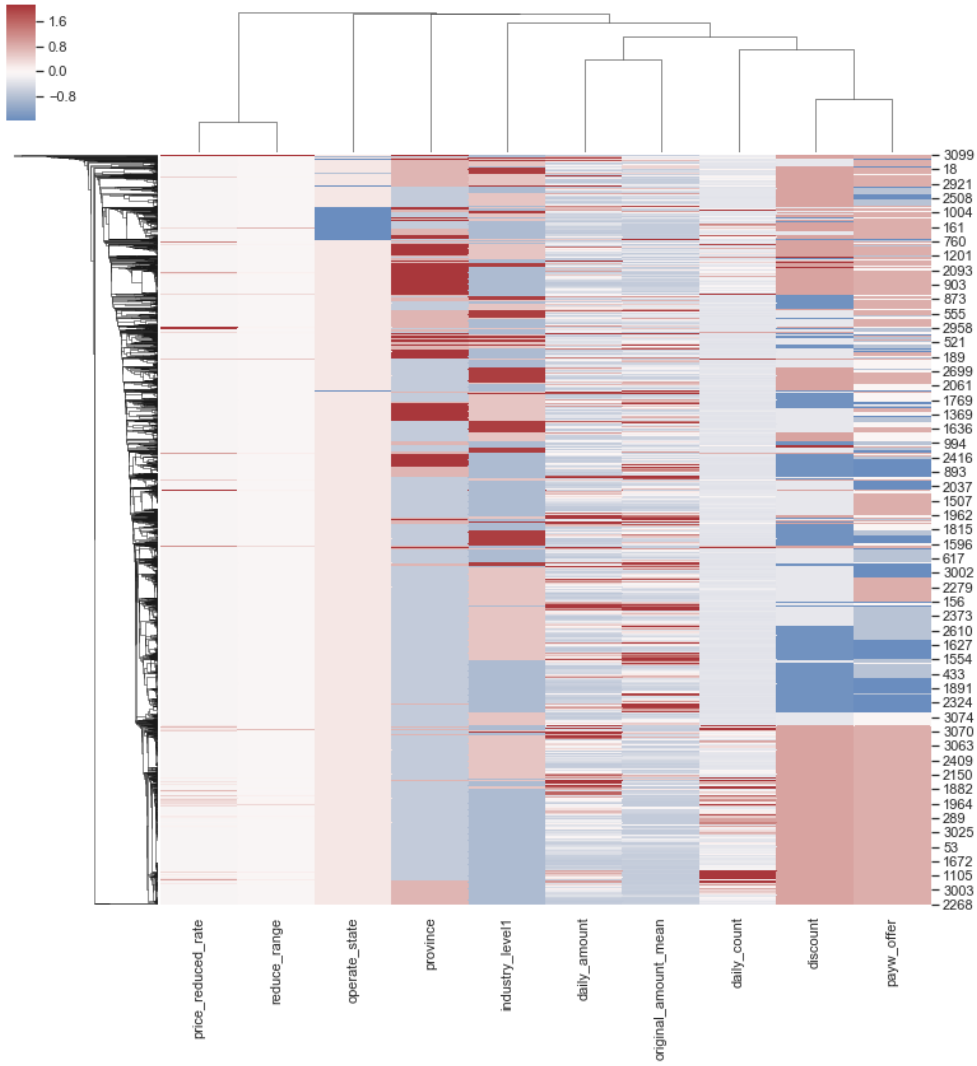
$$operate\_state_{store\ x} = \frac{Sum_{store\ x}(original\_amount)}{Sum_{merchant\ \chi}(original\_amount)}, store\ x \in merchant\ \chi$$

在时间对齐的问题上，要注意对于不同的店铺，分母"行业交易总额"总是统计到与分子"该店铺抛出日期"一样的时刻。



- **Overview：**

1. 先作图看看10个维度特征的分布情况：



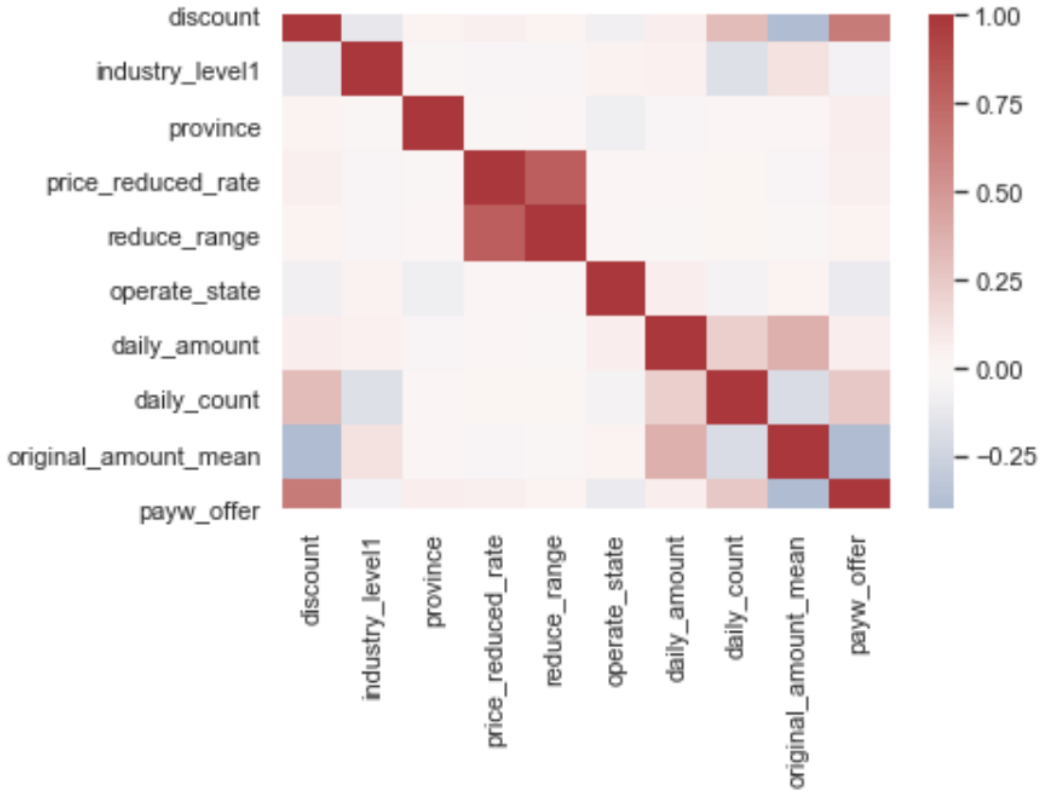
此处参数选择 z_score=1：

z_score : int or None, optional

> Either 0 (rows) or 1 (columns). Whether or not to calculate z-scores for the rows or the columns. Z scores are: z = (x - mean)/std, so values in each row (column) will get the mean of the row (column) subtracted, then divided by the standard deviation of the row (column). This ensures that each row (column) has mean of 0 and variance of 1.

这张分布图可以看出哪些特征分布稀疏(与上文单个分析时结论一致）、哪些特征分布较为对称（红蓝对比）。非稀疏的特征中，除了**operate state经营状况**和**daily_count日均交易笔数**外，其他特征的分布都较对称，可用性较高。由两个非对称特征推测，店铺中存在部分经营状态完全不活跃，存在部分低价走量的经营模式。

2. 再看特征间的相关性：



在相关性程度上，颜色浅接近白色的可以认为不相关；从图看出，在最重要的几个特征中，与营业额有关的特征之间存在相关性，从上文的解释意义来看，它们之间就存在冗余信息。这是在设计特征时故意留下的，因为不知道就模型上的表现来看，不同特征（信息的不同组织形式）在不同分类方式下哪个表现会更好。

所以，此处查看相关信息是只是为了大致确定模型参数max_features的大小，而非筛选特征。