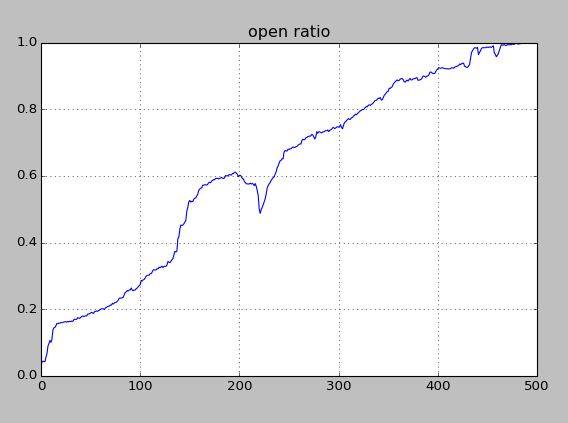
2017-02-01

本文主要基于count\_user\_pay.csv进行分析

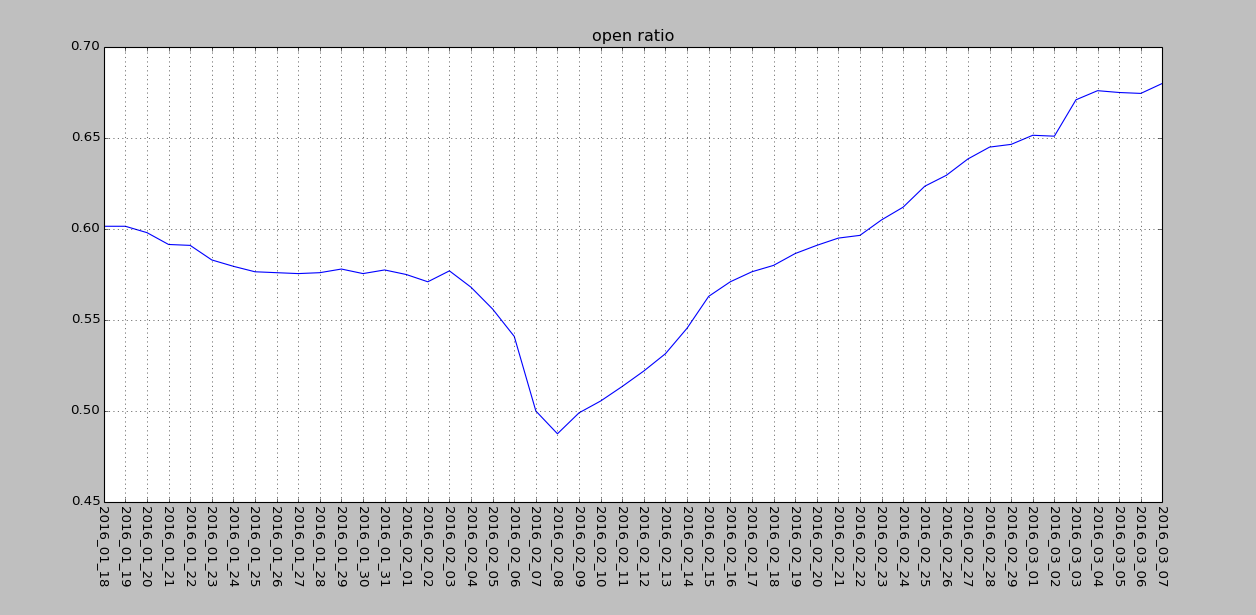
>>> open\_ratio():

用于分析商家在第[0,488]天（对应2015-07-01 ~ 2016-10-31）中的开张率。例如：图（100,0.23）的含义是在第100天，2000家商铺大概有2000\*0.23家开张。（ “开张率”是形象的称呼，实际上应当是 2000\*0.23的商家有客流，其余的在这一天都没有客流。）



观察上图可以发现，在第200~250天之间，开张率出现明显的下滑，我将观察范围缩小。

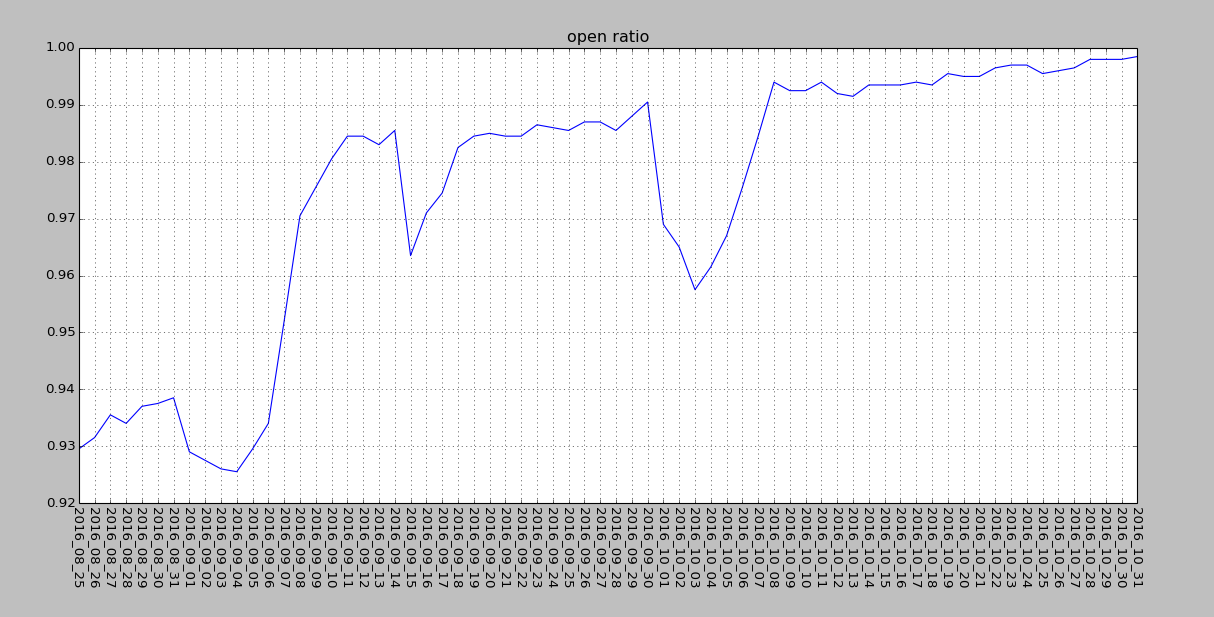
>>> open\_ratio(200,250)



观察范围缩小之后，发现波谷是2016-02-08（春节）。因此，在春节前后，一部分商家选择关门或者客流量为0。（为了显示清晰，只有当查找范围小于100天，既 250-200<100 的时候，横坐标才会显示日期。）

在第420天到最尾，开张率几乎稳定在峰值。放大观察：

>>> open\_ratio(420)



从2016-08-31 ~ 2016-09-11 经历了先下滑，后上涨（我个人觉得可能跟学生的开学有关）。

2016-09-15（中秋节）到达波谷。

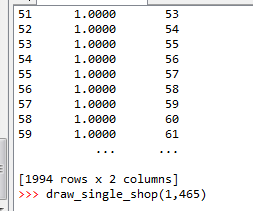
2016-10-01 ~ 2016-10-07（国庆节）也对开张率产生较大影响。

从2016-10-08开始，开张率稳定在99%以上，既每天还有<200家的商家客流量为0.这部分数据适合用来构建训练模型。

进一步分析2016-10-08 ~ 2016-10-31之间各个商家的开张情况。

>>> every\_shop\_open\_ratio(0.5,465)

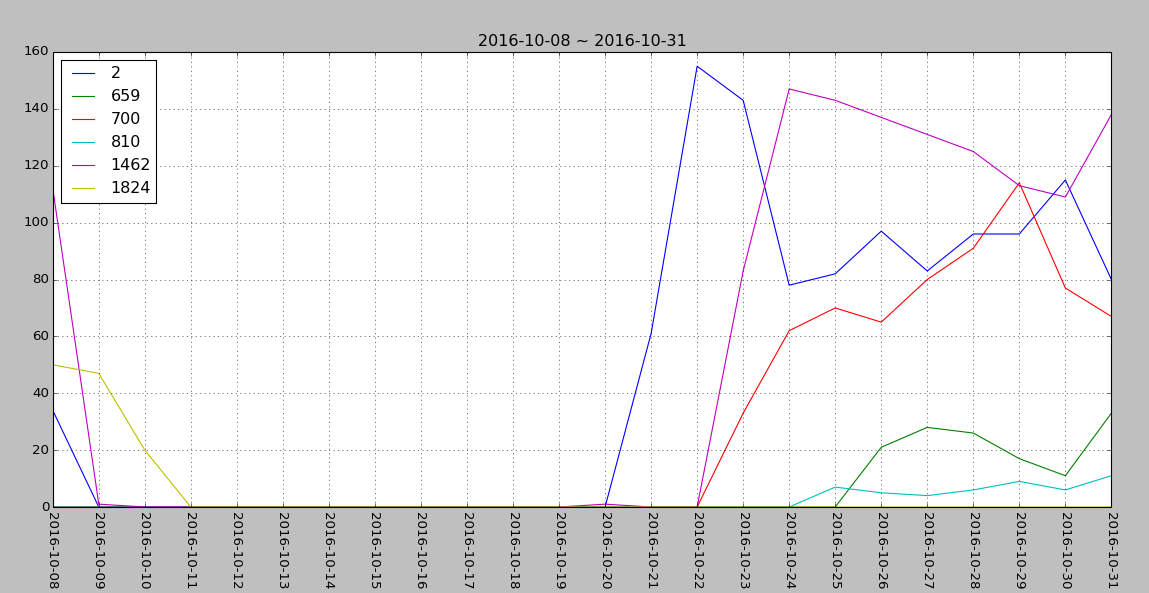
# 筛选出第465天至最后一天（2016-10-08 ~ 2016-10-31）开张率大于50%的商家。



结果表明：1994个商家在这段时间中，开业率都大于50%。我反向筛选出剩余的6个开业率小于50%的商家,并作图（纵坐标是客流量，既当日用支付宝买单的人数）。

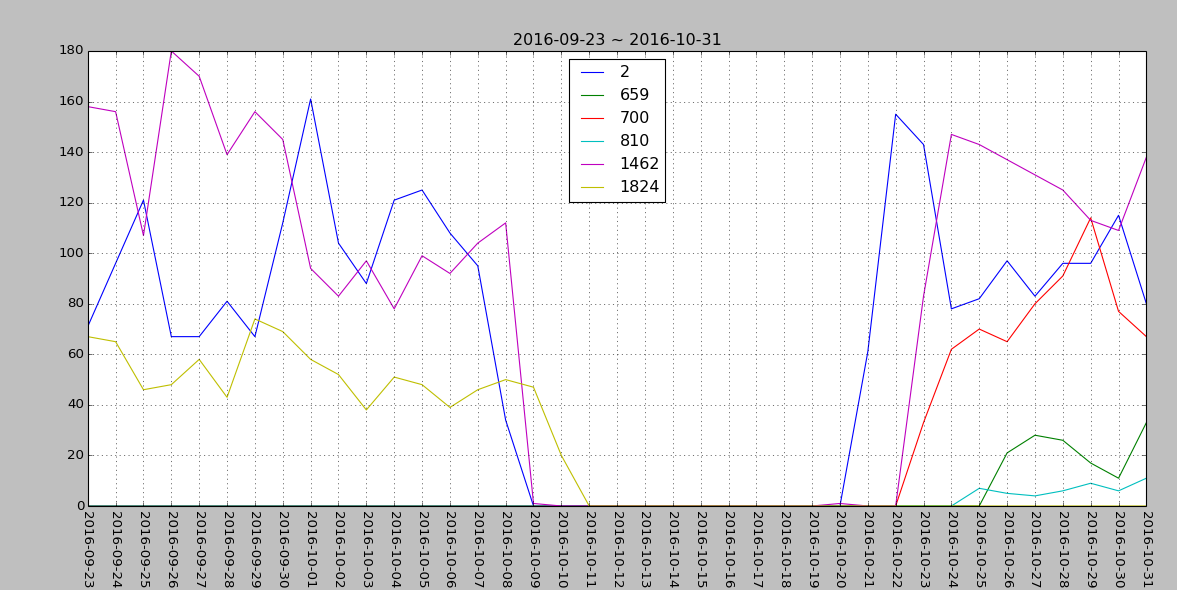
>>> df = every\_shop\_open\_ratio(0.5,465,smaller=1)

>>> draw\_multi\_shops(df.shop\_id,465)



我将左边的时间轴延伸：

>>> draw\_multi\_shops(df.shop\_id,450)

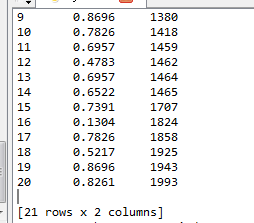


结果表明：这些商铺都选择在国庆期间继续营业，而在国庆之后休业一周。

我继续增大threshold = 0.9

>>> every\_shop\_open\_ratio(0.9,465,smaller=1)

# 筛选出第465天至最后一天（2016-10-08 ~ 2016-10-31）开张率大于90%的商家。

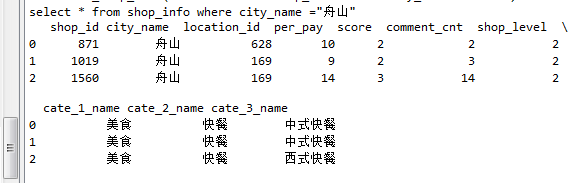


结果表明：只有21家在2016-10-08 ~ 2016-10-31开张率小于90%。

接下来分析“星期”对数据的影响，这里我以舟山为例：

>>> fetch\_shop\_info('select \* from shop\_info where city\_name ="舟山"')

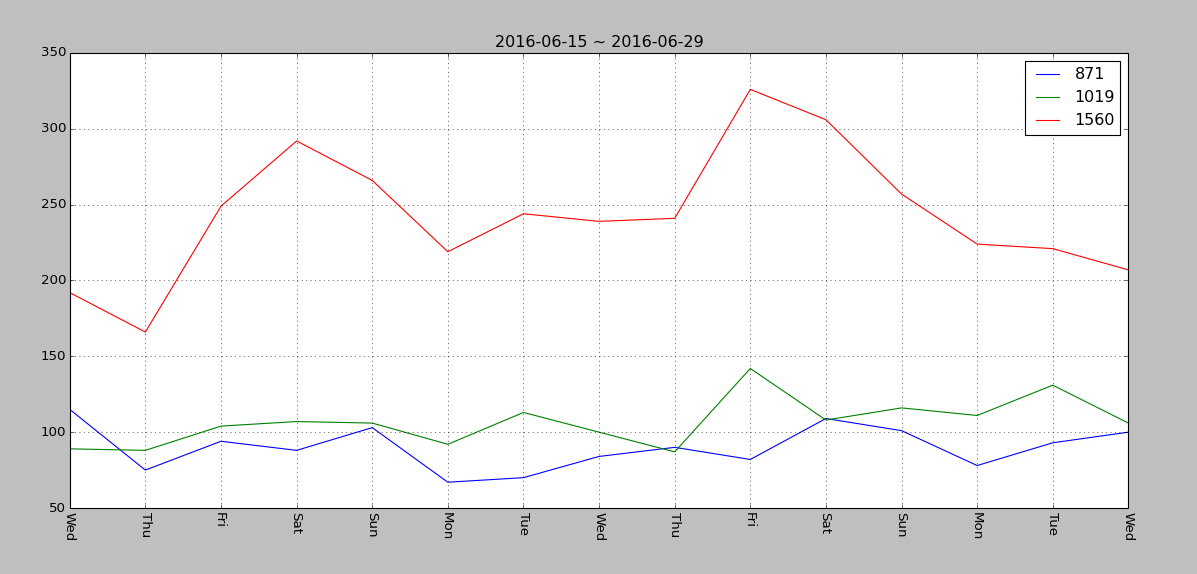
# fetch\_shop\_info:在该函数直接输入SQL语句，结果会以DataFrame的形式返回。



>>> df = fetch\_shop\_info('select \* from shop\_info where city\_name ="舟山"')

>>> draw\_multi\_shops(df.shop\_id,350,364,week=1)

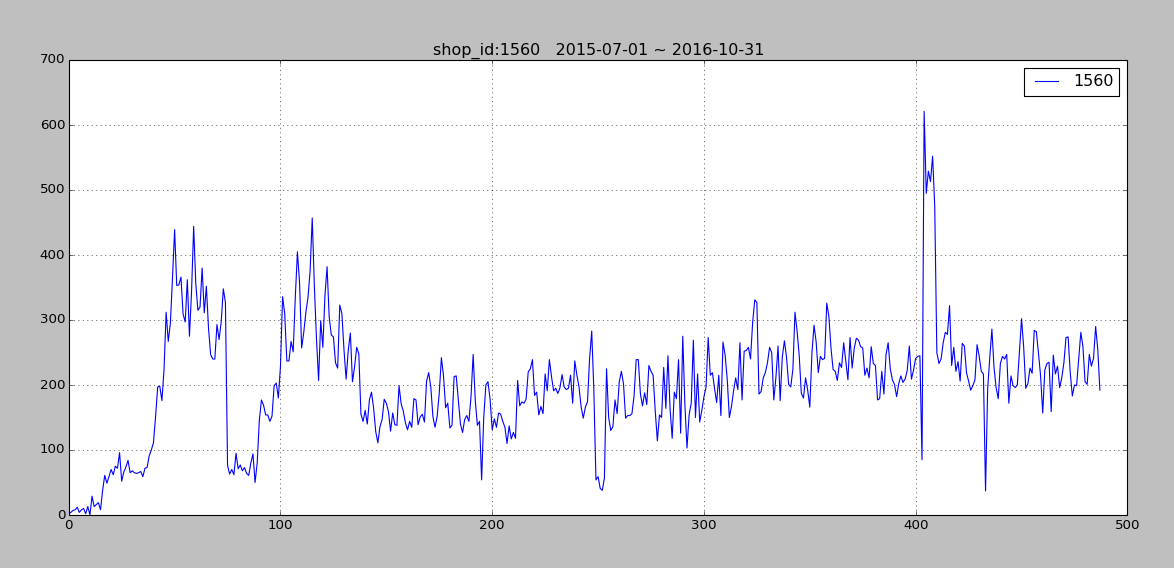
# 当week=True的时候，横坐标会用“星期”来表示；



结果似乎显示红色：1560店铺的周末销量更好。

我单独分析1560.

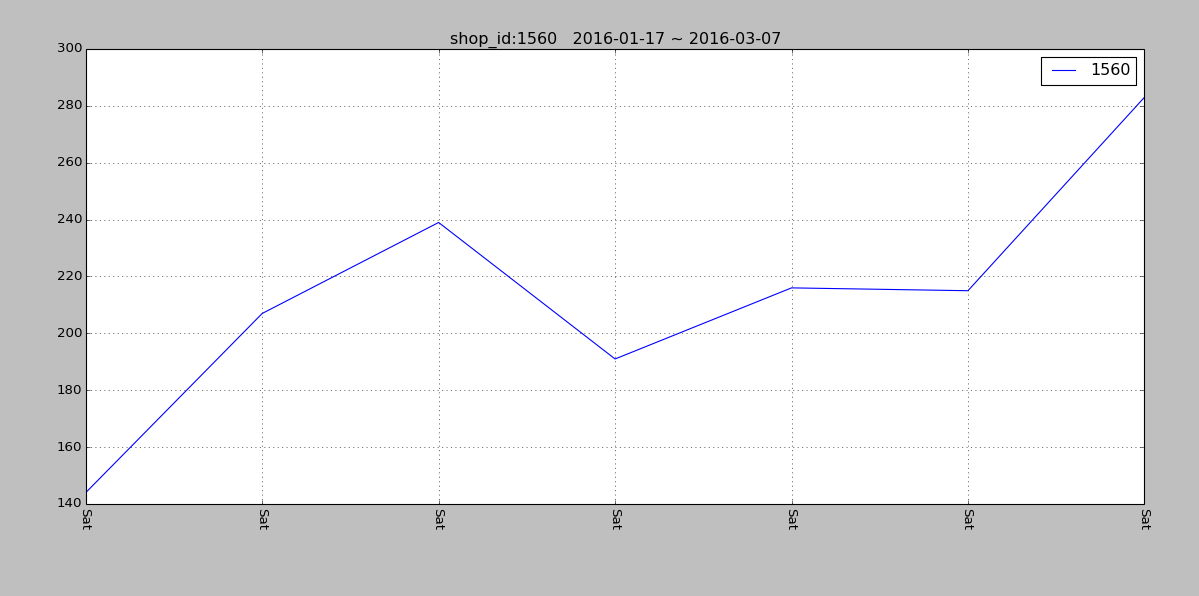
>>> draw\_single\_shop(1560)



结果表明：全年开业。

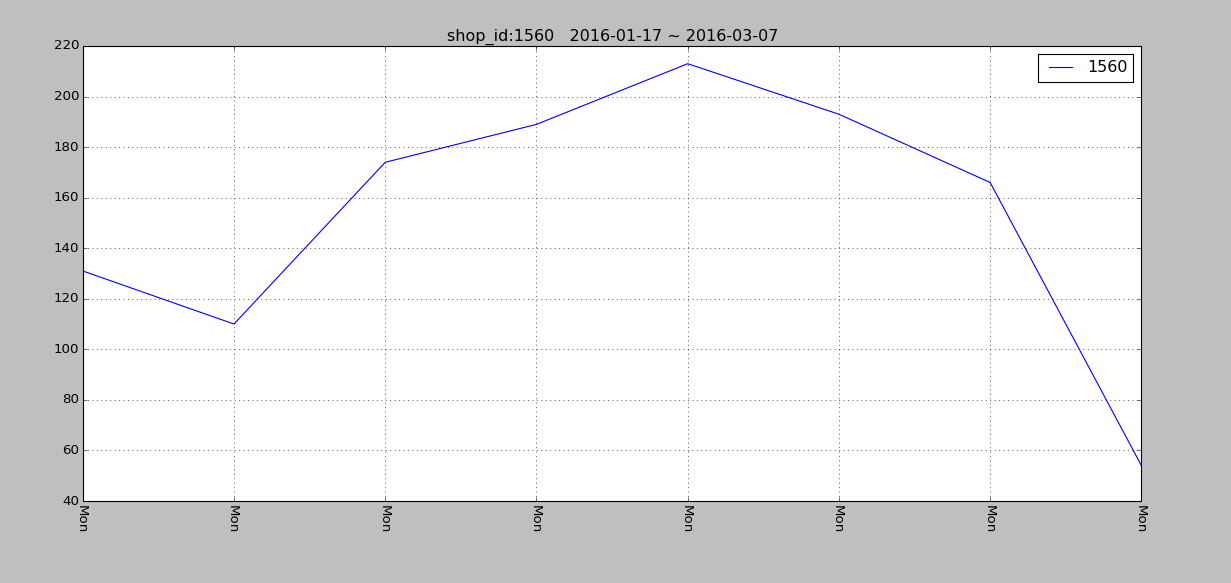
>>> draw\_single\_shop(1560,200,250,week=1,fr='W-SAT')

# 1560号 第200-250天范围内，每个周六的客流量。



>>> draw\_single\_shop(1560,200,250,week=1,fr='W-MON')

#同样时间段中，每个周一的客流量。



综上：

1. 重大的国家法定节假日会对开张率产生较大影响，但是由于待预测范围：2016-11-01 ~ 2016-11-14不存在法定节假日，因此需要注意的是：取训练集的时候应当尽量避开之前日期中的法定假日。
2. 2016-10-08之后的数据能够比较好的体现商家运营状态，少数特殊的商家人工处理即可。
3. “星期”在部分商家的客流量上有较大影响。