**因子1：bid\_age**

含义：当前时刻的ap1或bp1，按照这个值已经持续了多少时间。

dataframe.rolling()函数提供滚动窗口计算的函数。一次取一个窗口大小为k，并对其执行一些所需的数学运算。

pd.dataframe.apply(func,axis,raw:bool)，raw是True的话，function收到的是ndarray对象；否则收到的是Series对象。

**因子2：arrive rate**

含义：交易成交的频率，可以看出行情好坏。

res = trade['ts'].diff(n).fillna(0) / n

**因子3：twap\_diff**

含义：历史平均交易价格与当前交易价格的差值，反映市场状况。

**因子4：cofi（尝试用bp1-bp5做一个因子）**

含义：比较当前时刻与前一时刻bp1和ap1的差值，乘上bv1或av1。

**因子5：oir\_diff**

含义：比较当前av1-av5和bv1-bv5的大小。

其他的注意点：

数据集中，ts是时间戳，单位是ms。

**因子6：pmtv（尝试改下diff的值，做一个因子）**

含义：(当前时刻累计交易量-前一时刻累计交易量)\*(当前时刻bp1-前一时刻bp1)

**因子7：price\_impact**

含义：当前ap1与5个ap整体值的比较；bp1与5个bp整体值的比较。

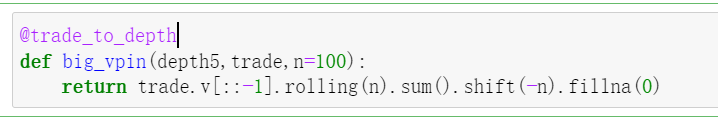
**因子8：weight price to mid**

含义：给5个level的volumn赋予退化权重

,

**因子9：withdraws**

**因子10：big\_vpin**



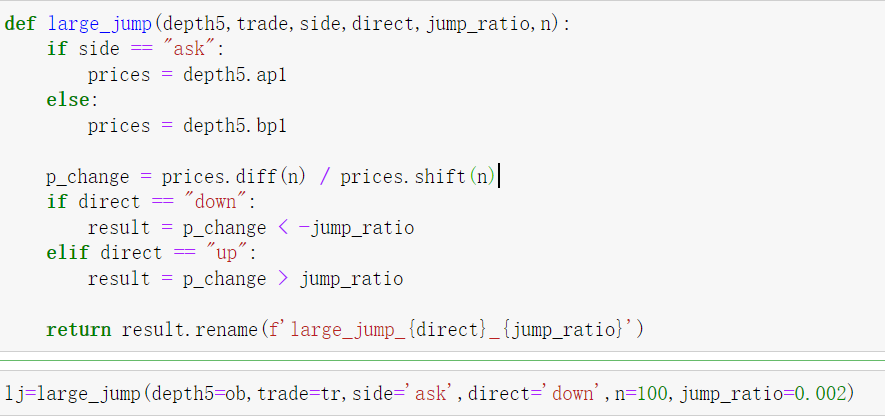
**因子11：fair\_spread**

含义：

相关系数高。

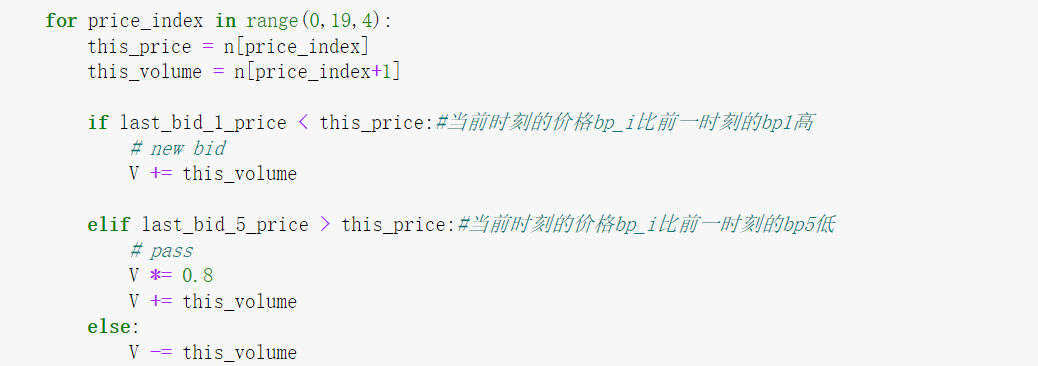
**因子12：large\_jump**

含义：价格(ap1 or bp1)变化是否超过特定的阈值（有买和卖两个方向）。



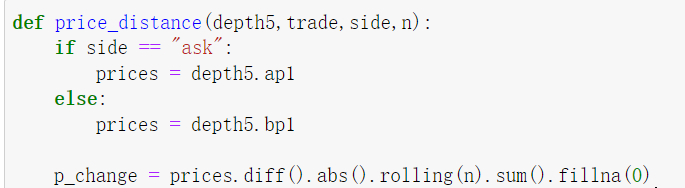
**因子13：order\_flow**

含义：先记录前一时刻的bp1和bp5，以及V=bv1+bv2+bv3+bv4 +bv5；再比较当前时刻价格和前一时刻bp1与bp5的大小，根据比较结果调整V的大小。（ask和bid类似）



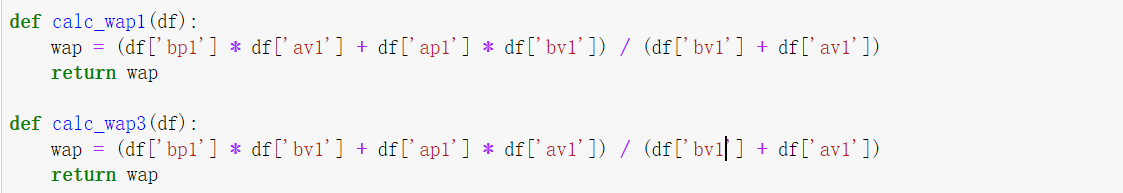
**因子14：price\_distance**

含义：当前时刻的价格与前一个时刻价格的差值，取绝对值。再对当前时刻前n个时刻的差值求和。



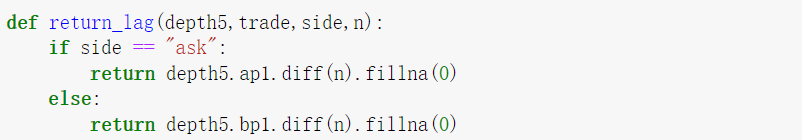
**因子15：price\_swap**

含义：原先是ap1\*av1+bp1\*bv1，改成ap1\*bv1+bp1\*av1；再求这两个值的差。



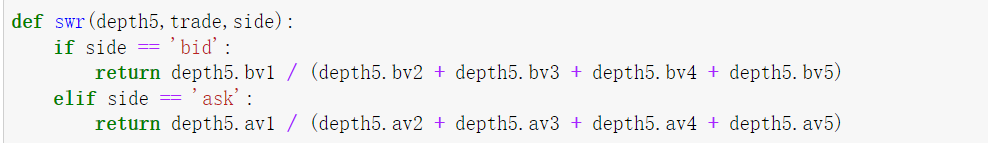
**因子16：return\_lag**

含义：ap1和bp1价格的差值。



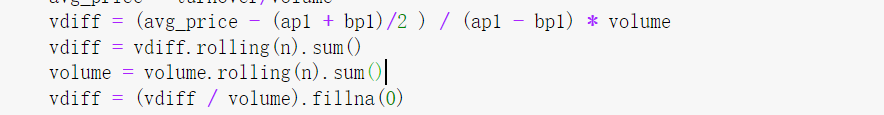
**因子17：swr**

含义：



**因子18：tick\_vpin**

含义：avg\_price = trade.p/trade.v；volume=trade.v；



**因子19：volume**

含义：当前时刻的bv1或av1

**因子20：volume\_at\_same\_price**

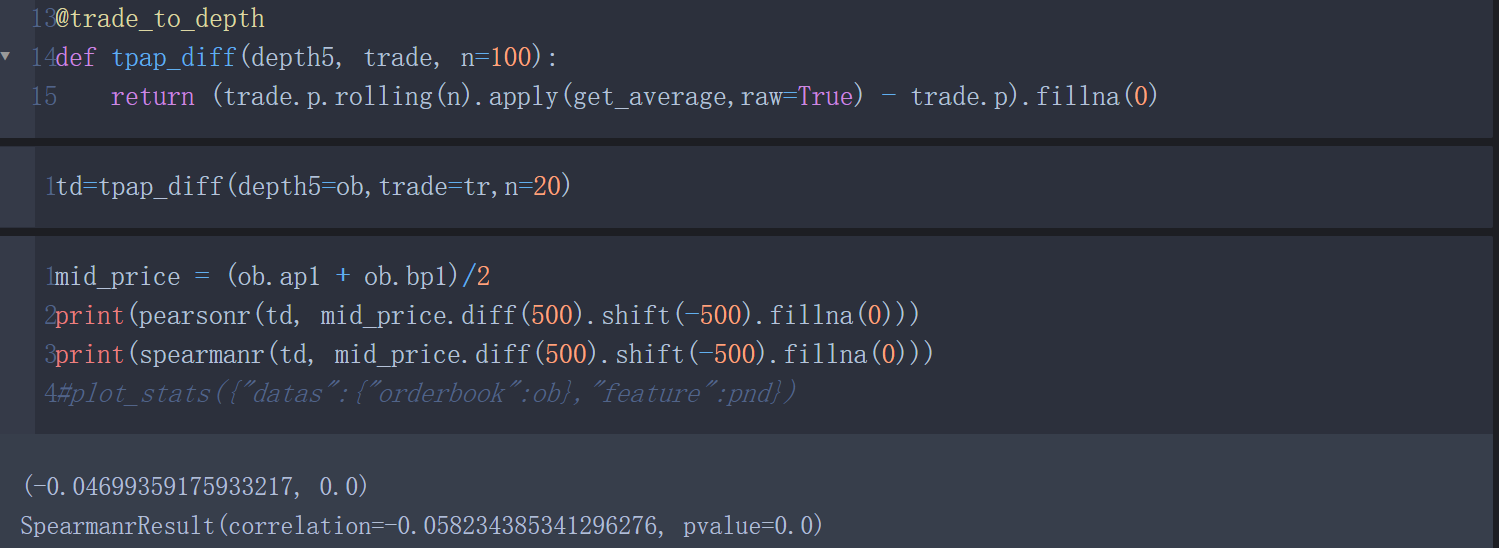
含义：对trade数据，同一个price持续的时间。

**新的测试变量：**

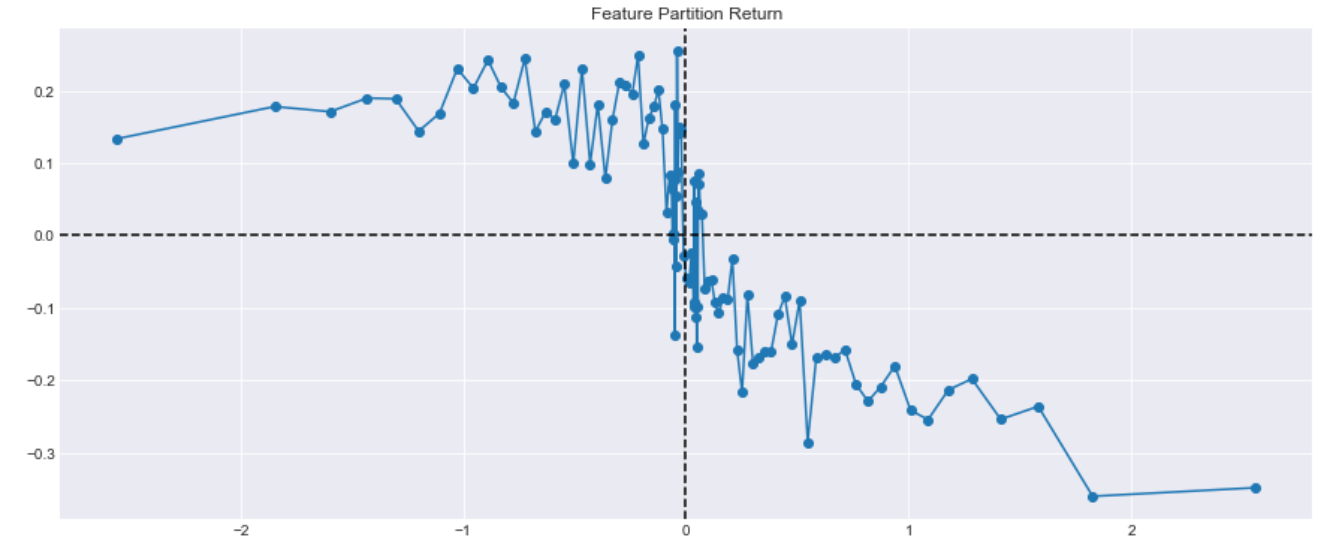
对trade来说，price的值较高，因此采用两个数相减，比两个数相除会更有可解释性些。

**测试因子1：tpap\_diff**

含义：当前trade的价格与前n个时刻trade平均价格的差值。（相比twap\_diff，没有用trade.v做加权）

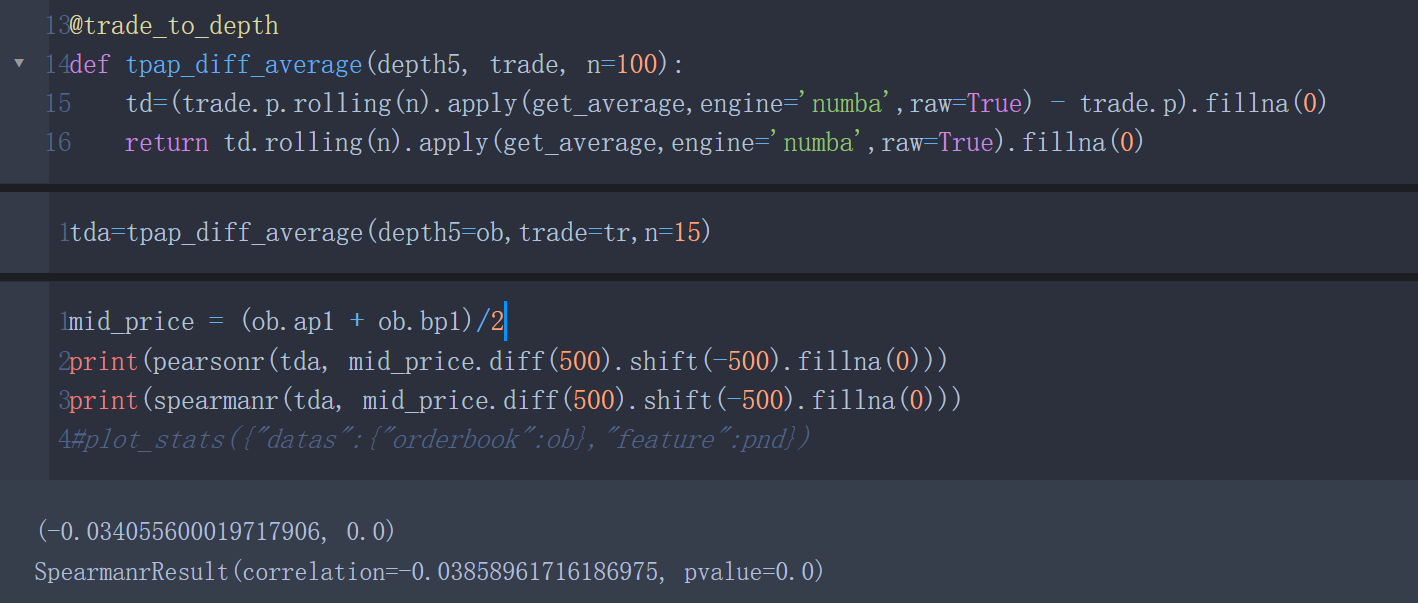
****

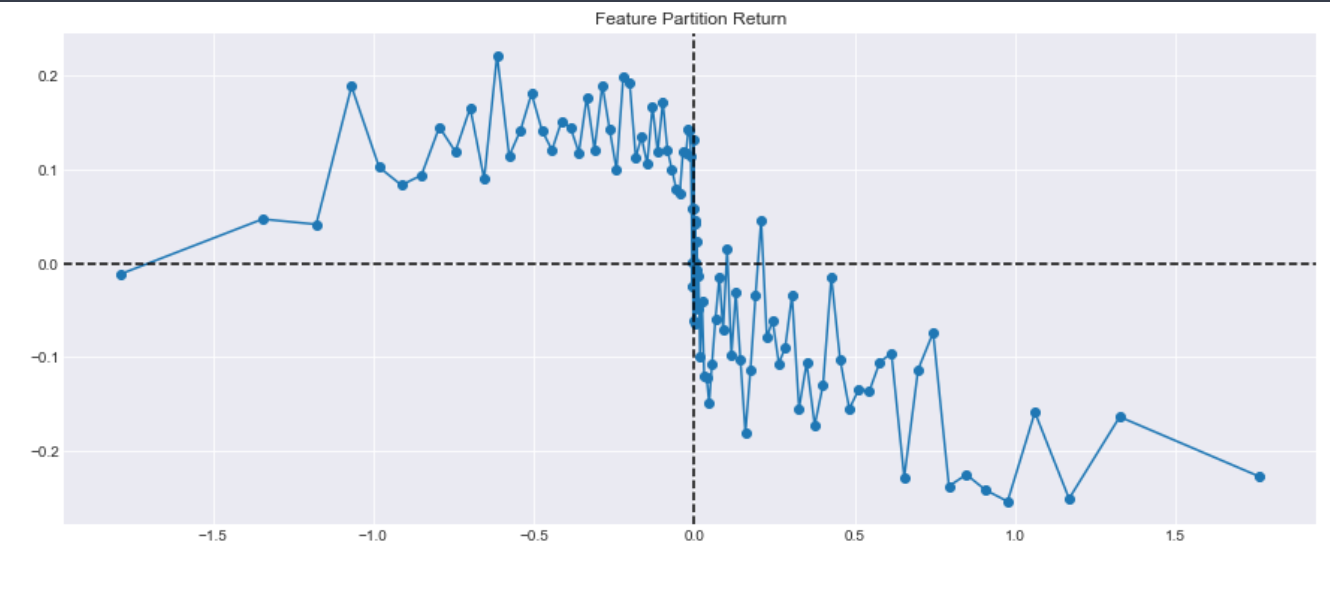
这个因子比较好的地方是负x对应的y一般都为正；正x对应的y一般为负。



**测试因子2：tpap\_diff\_mean**

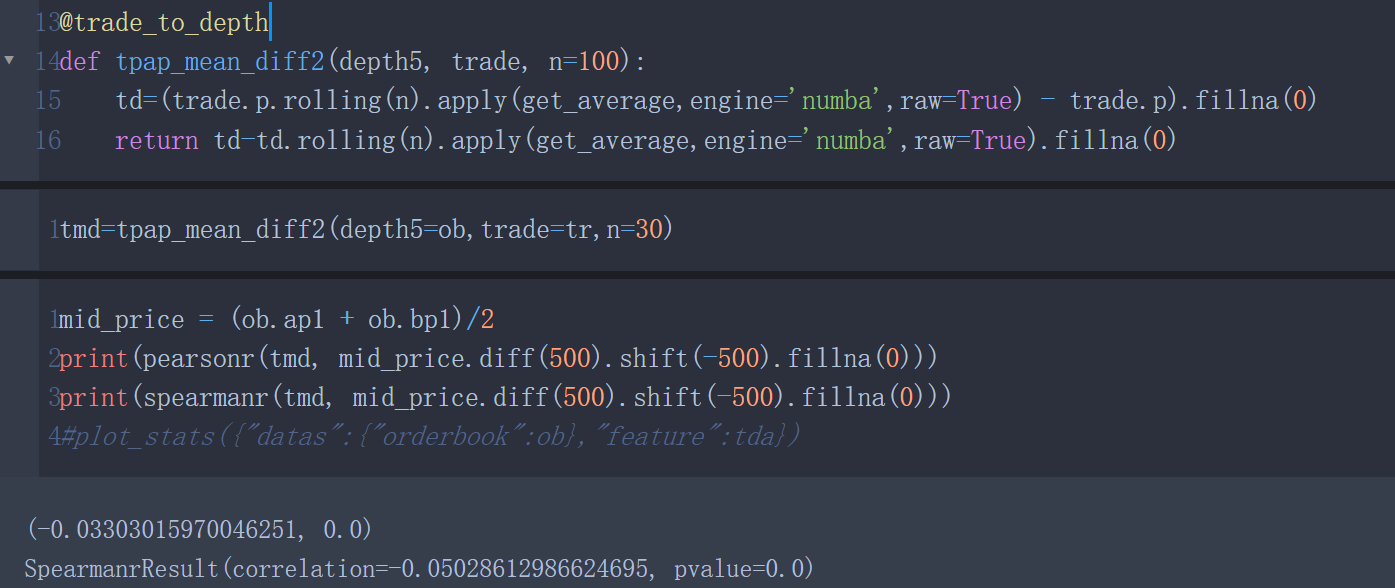
含义：对测试因子1的前n个时刻，取平均值。





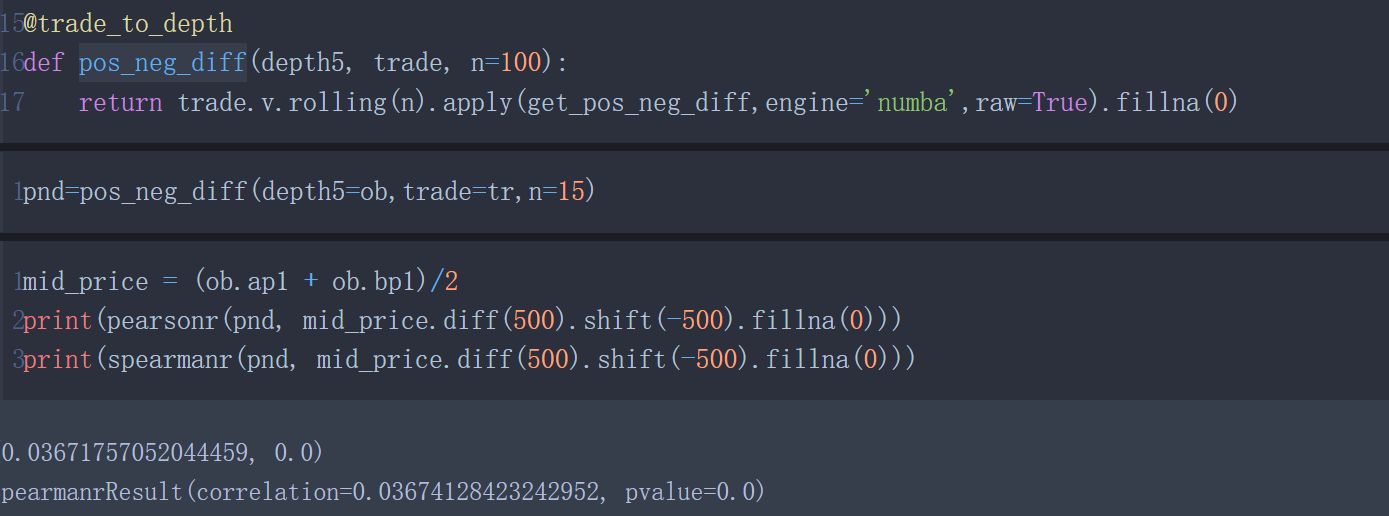
**测试因子3：tpap\_mean\_diff2**

含义：测试因子1减测试因子2。

****

**测试因子4：pos\_neg\_diff**

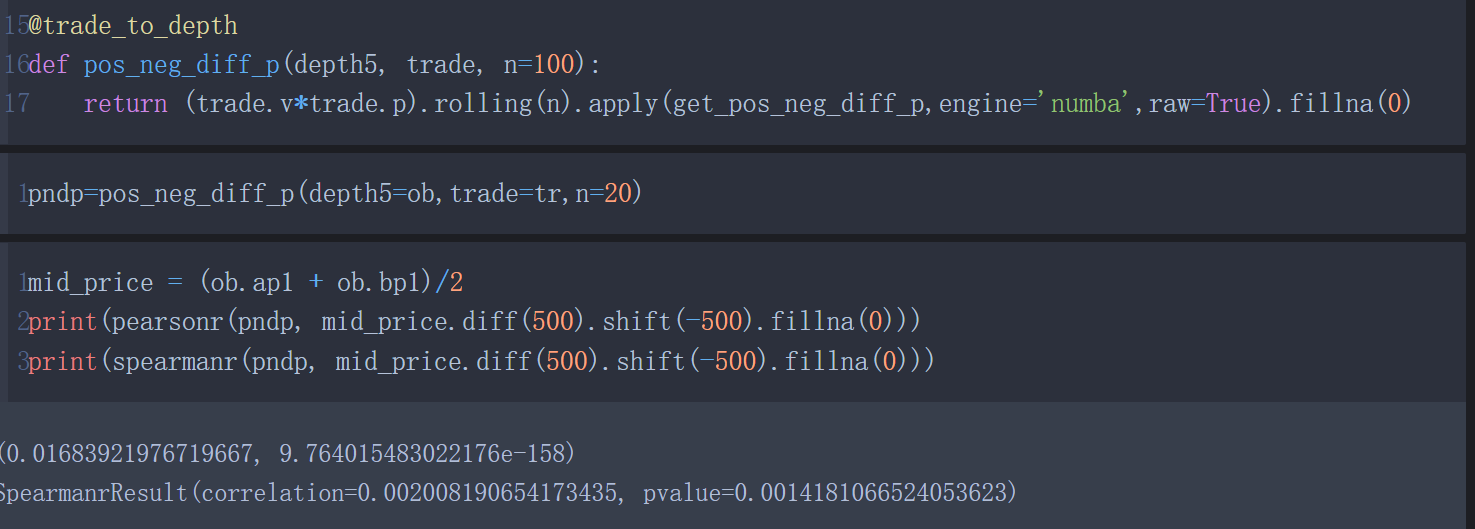
含义：最近的n个时刻，trade.v为正和trade.v为负的**个数**的差值（更一般的，trade.v的正负差值）。



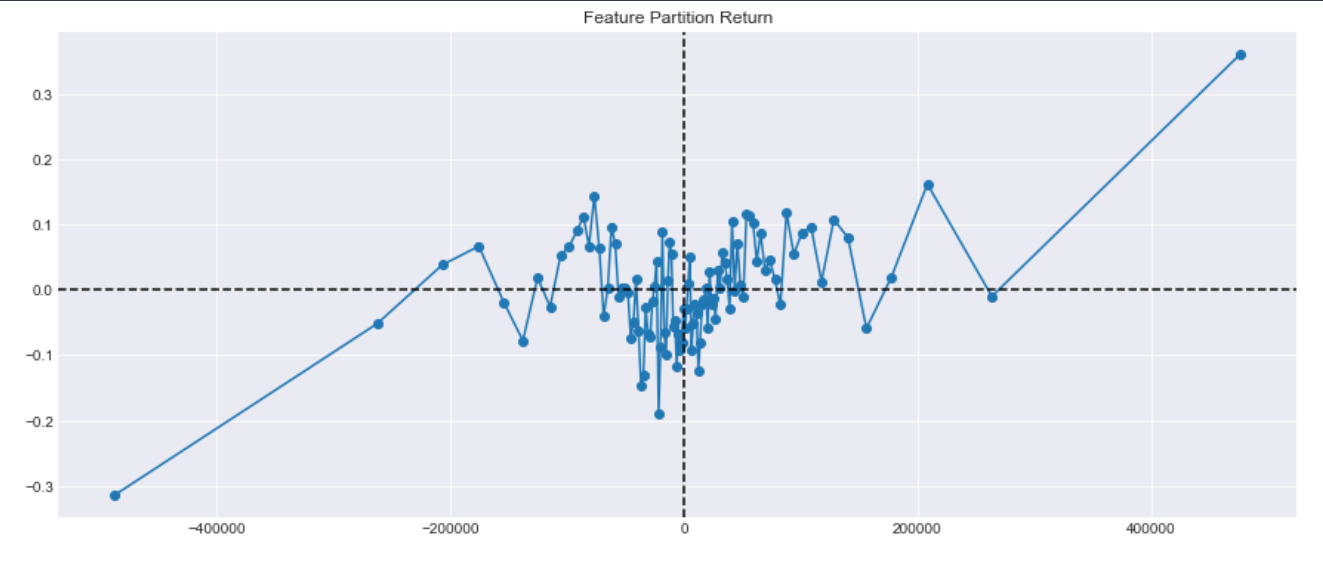
相关性较高，plot\_stats的情况也较好。

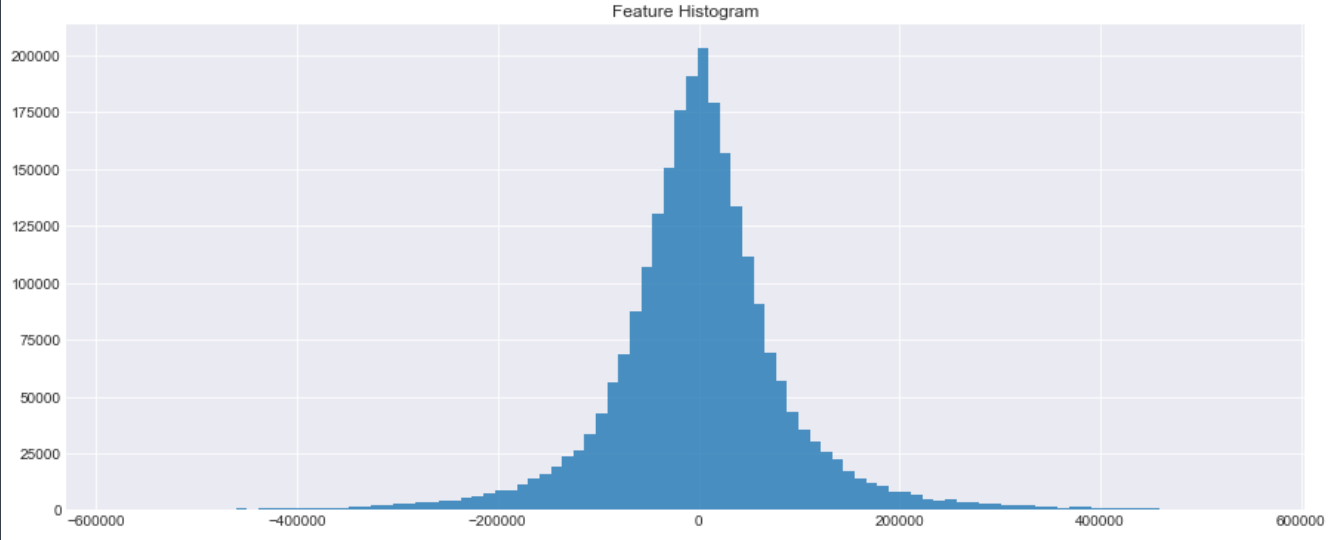
**测试因子5：pos\_neg\_diff\_p**

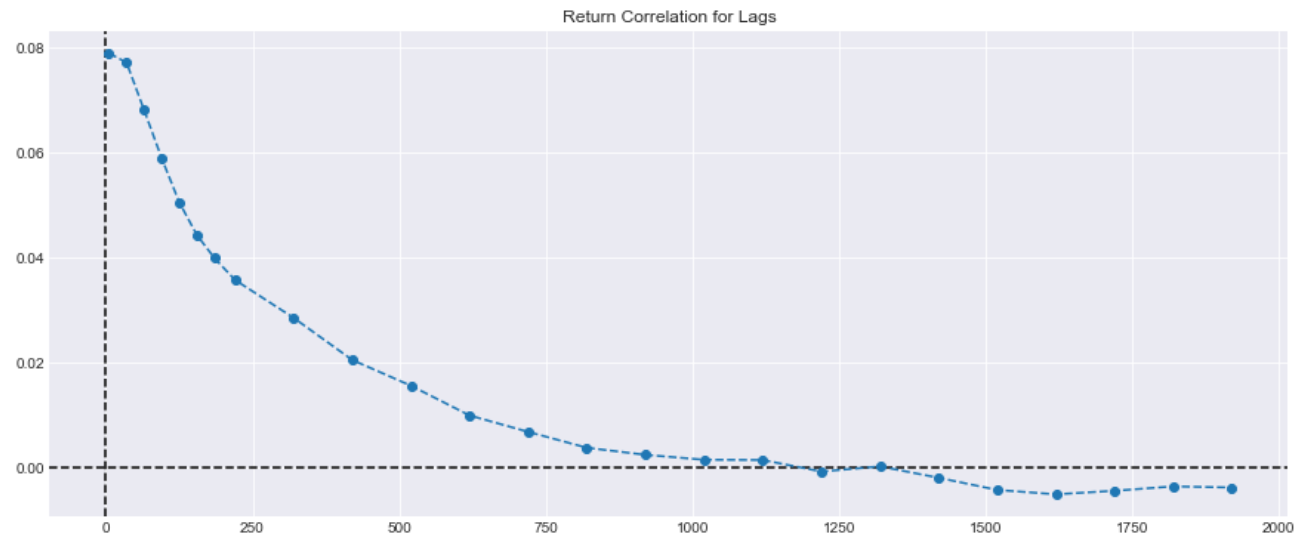
含义：最近的n个时刻，trade.v为正和trade.v为负的时间点交易总价的差值。



相关性不高，但plot\_stats较好



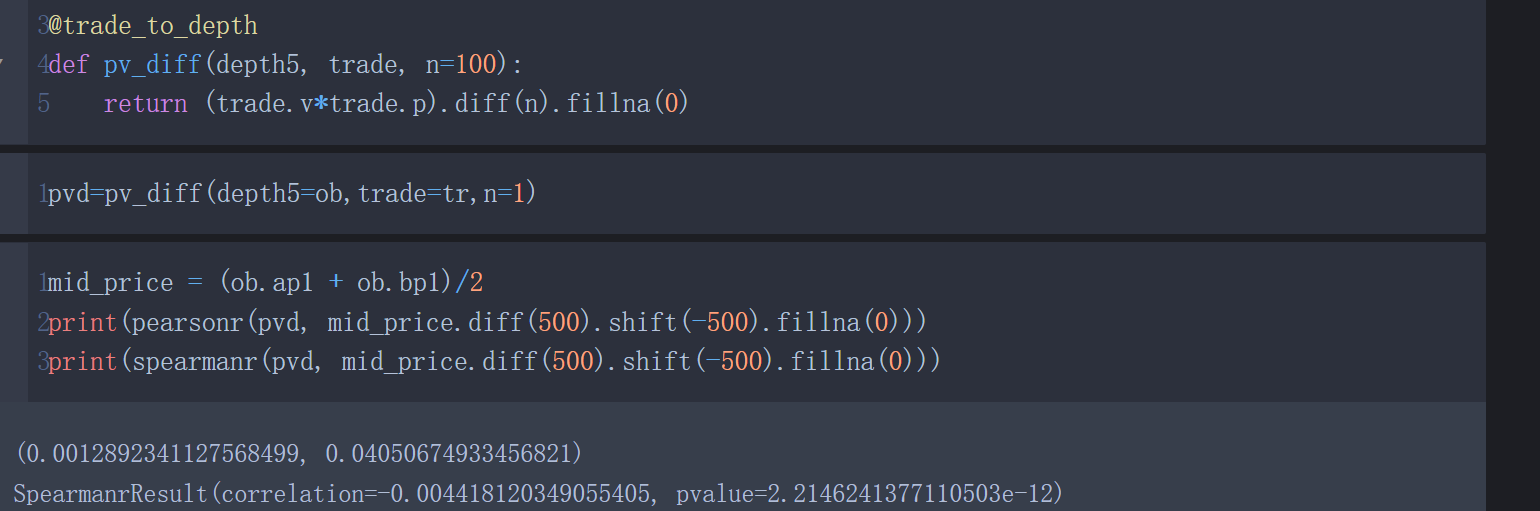




**测试因子6：pv\_diff**

含义：当前时刻v\*p减去前一时刻v\*p

从意义上来讲可行，比如前一时刻v为负值， v为正值，相减后，值较大，体现买卖主动性的变化。

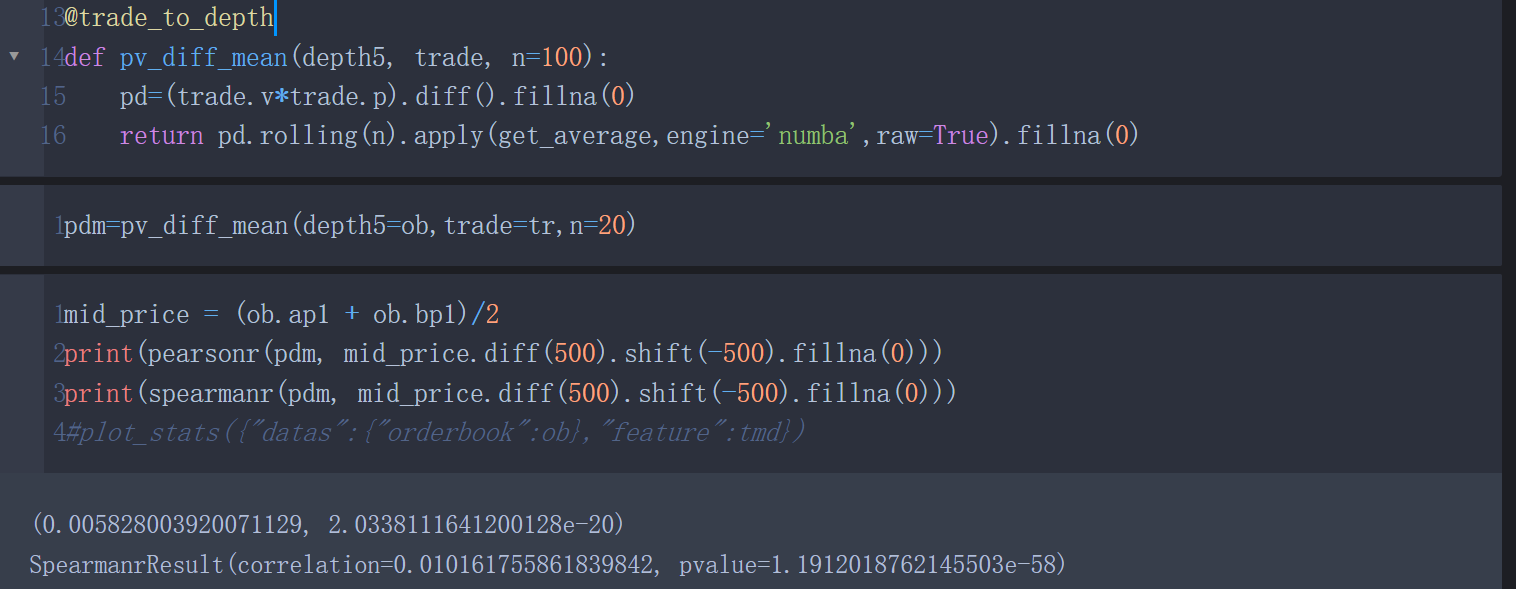


但总的相关性较低。

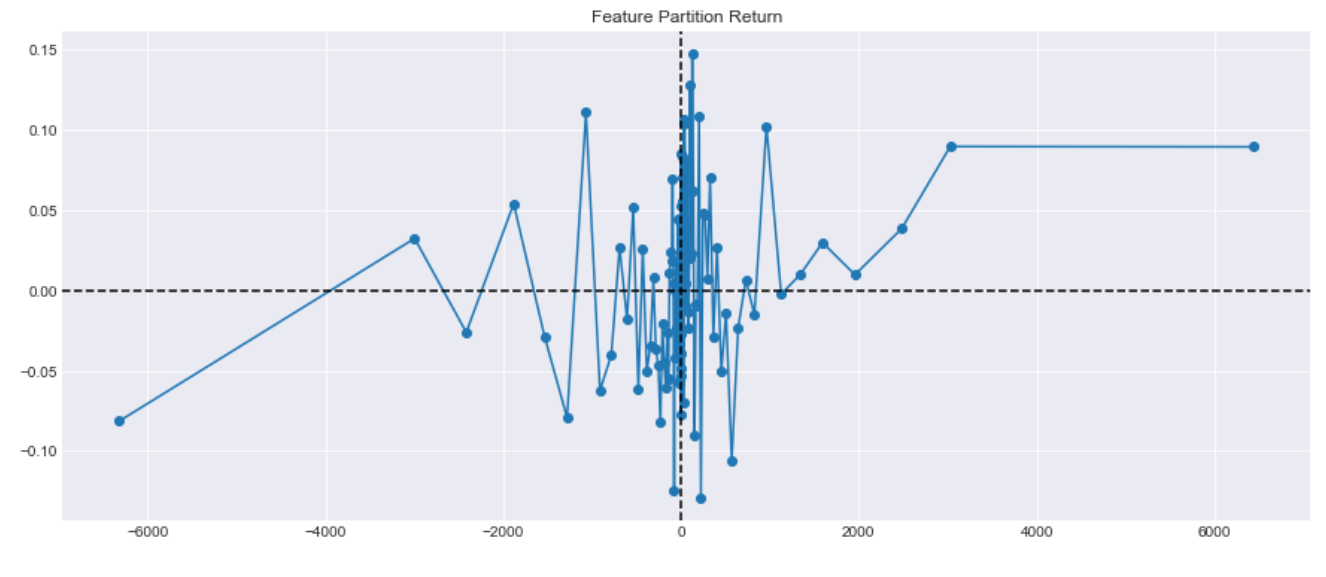


**测试因子7：pv\_diff\_mean**

含义：测试因子6前n个时刻的平均值。



相关性不高。



**测试因子8：tpap\_std**

含义：测试因子1（当前trade的价格与前n个时刻trade平均价格的差值）除以前n个时刻trade价格的标准差。

测试因子8，相比较测试因子1，在n=20时，皮尔逊相关系数从-0.470提高到-0.057；斯皮尔曼相关系数从-0.058提升到-0.065。而且相关系数图更接近于线性。但是变量呈现双峰分布。

|  |
| --- |
|  |
| 测试因子1 |
|  |
| 测试因子8 |

