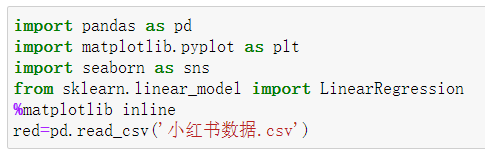
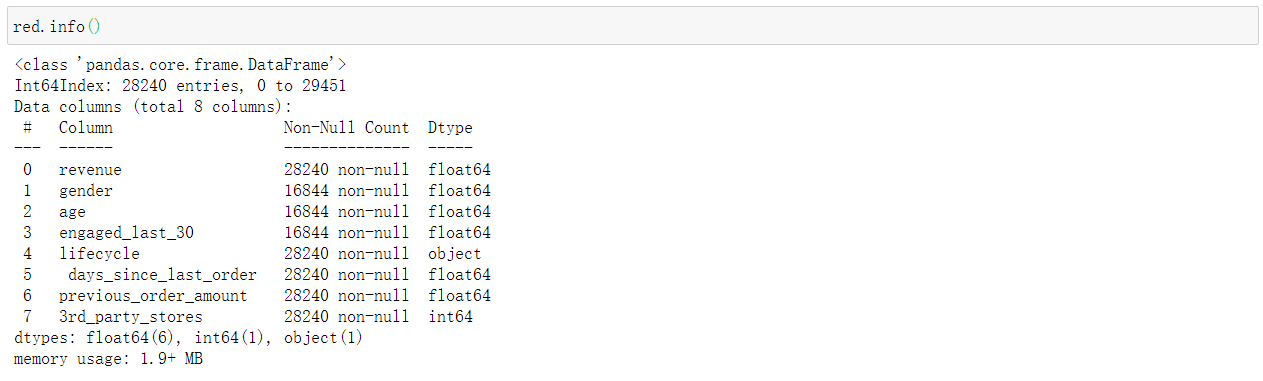
### WEEK6

1. 数据预处理

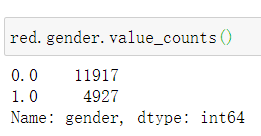
导入依赖库及数据：

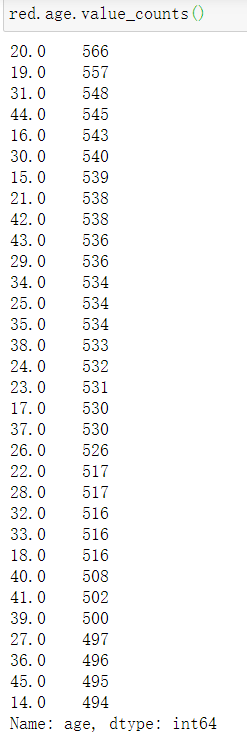


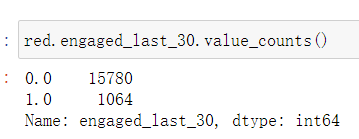
首先用info()方法初步检查是否存在空值：



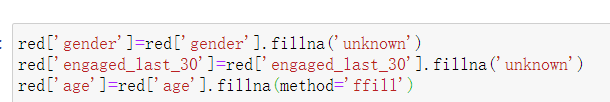
可以看到数据中gender，age和engaged\_last\_30这三个变量存在着空值，且空值数量较大，因此决定选用补充的方式来进行处理，并且要对这三个用value\_counts()方法来观察具体数值来决定补充的数据类型：



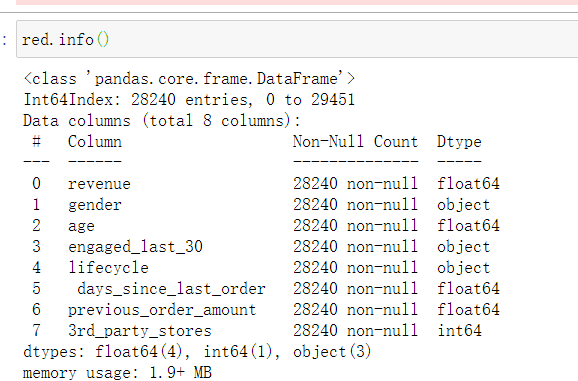




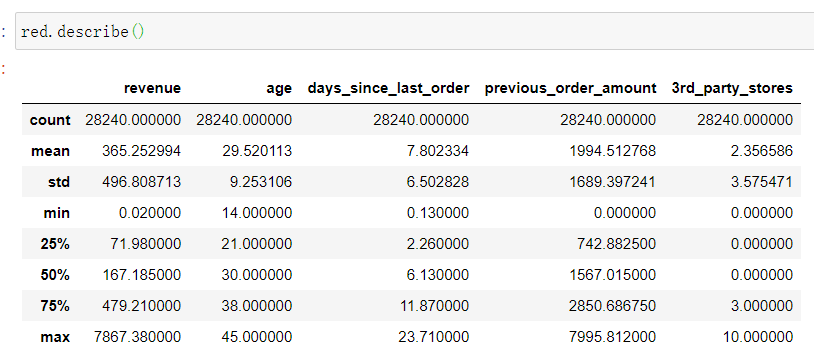
因此可以判定gender和engaged\_last\_30应该填充’unknown’,而age的分布是14到45之间的均匀分布，所以可以用向前填充或者向后填充的方式，在这里使用的是ffill：

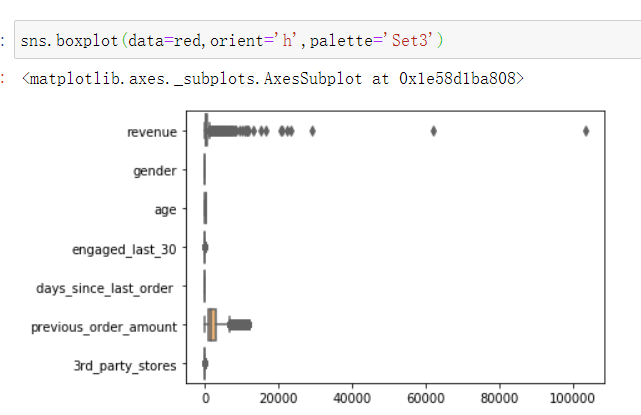


然后再使用info()检查空值的替换情况



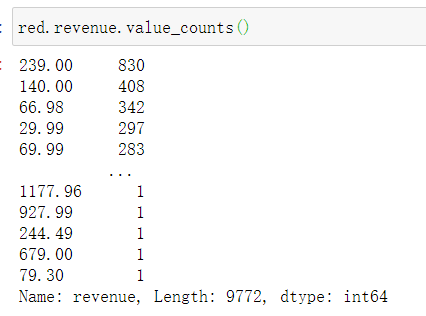
可以看出数据中存在的空值已经填充完成，接下来就是异常值的处理，先使用describe()方法和箱型图来观察数据的整体的分布情况：

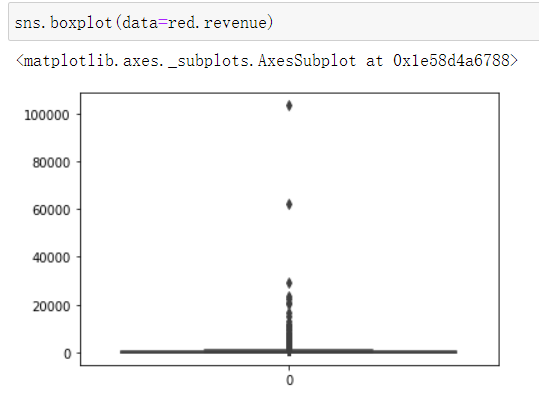


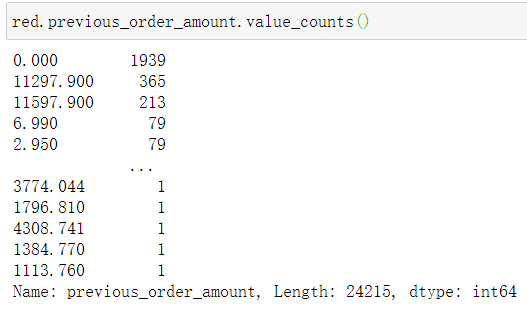


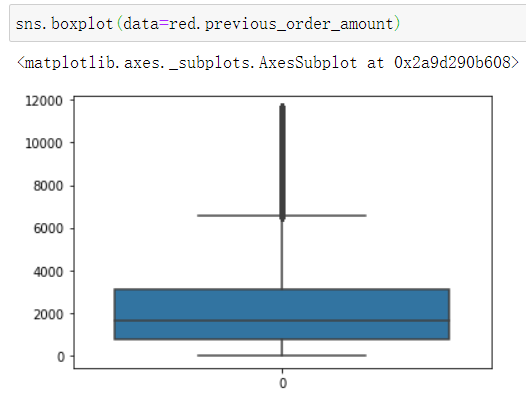
可以从表格中观察到revenue和previous\_order\_amount这两项的方差相比其他变量要大得多，因此这两项很有可能存在离群值，并且这一推断在箱型图中也得到了很好地印证。

因此需要单独针对这两项进行value\_counts()和箱型图分析来对异常值进行更好的处理：



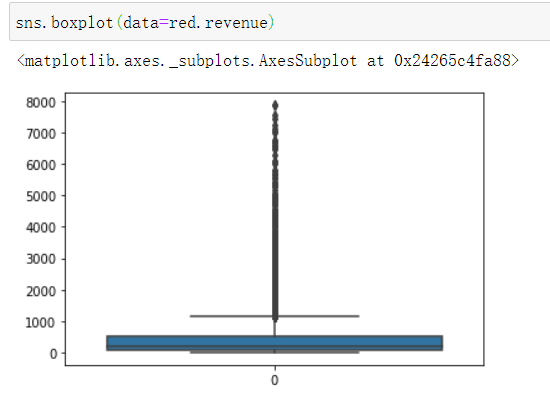




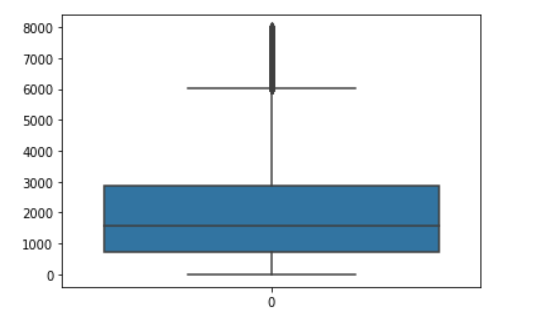


从分布和箱型图的结果可以判断,revenue大部分数据集中于小于8000的范围内，而previous\_order\_amount的也集中于小于8000的范围内，因此处理及结果如下图：









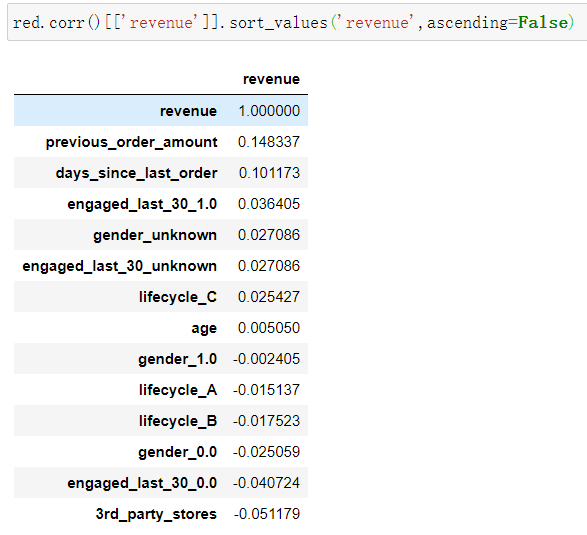
处理后的结果如图,且处理前后的数据量差距小于百分之0.5%，在原则上保留了大部分的数据。

最后针对一些类别型的变量，我们需要衍生出哑变量来方便后续的回归分析。



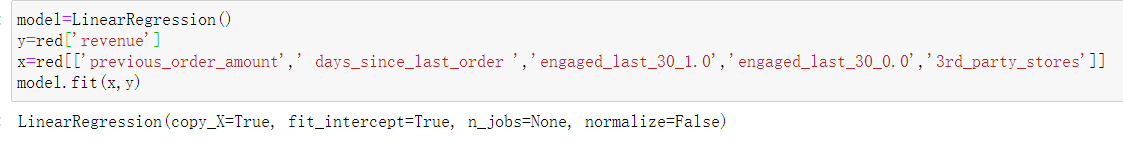
1. 使用Python建立线性回归模型

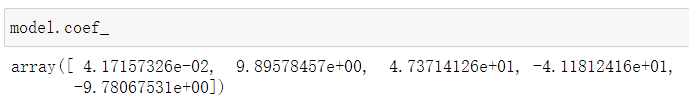
在建立回归模型前，因为考虑到变量较多，因此我们要通过corr()方法来筛选出相关性更高的变量来更有效地建模：

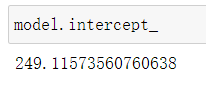


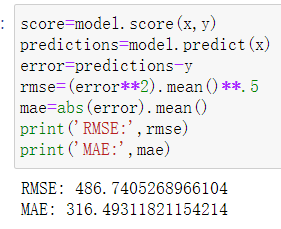
分析：

我们可以观察到previous\_order\_amount，days\_since\_last\_order这两项有着一定的正相关性，3rd\_party\_stores有着一定的负相关性，且engaged\_last\_30\_0.0和engaged\_last\_30\_1.0在其中也起到了一定的相关性作用，因此我们可以选用上述的几个变量来作为预测因子。









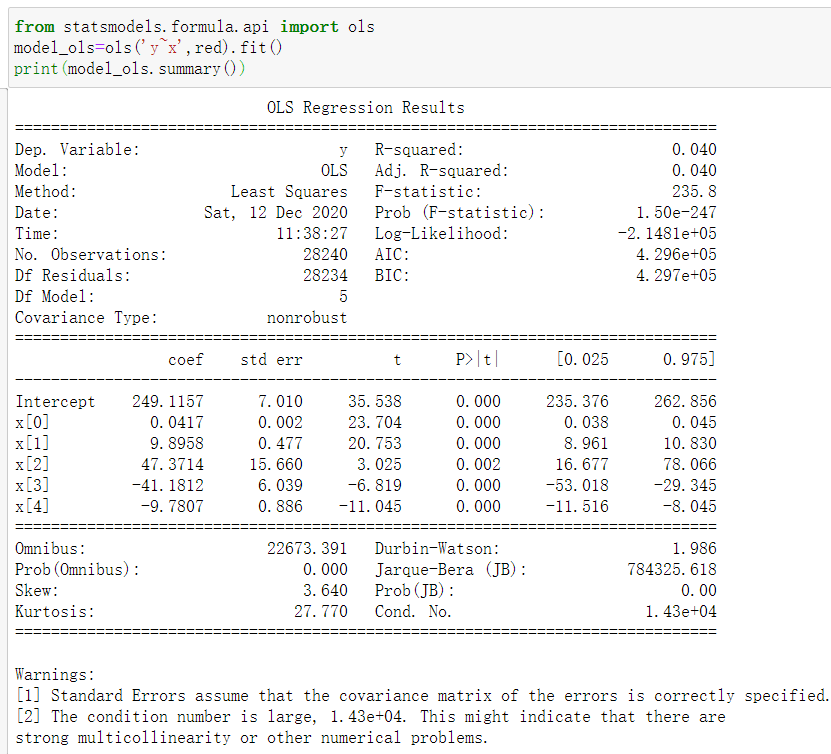
上图分别是模型的建立、相关因数、截距、以及残差评估，

为了选择更好的模型，之前还建立了bfill和ffill但未进行异常值处理的模型，以下是它们的残差：





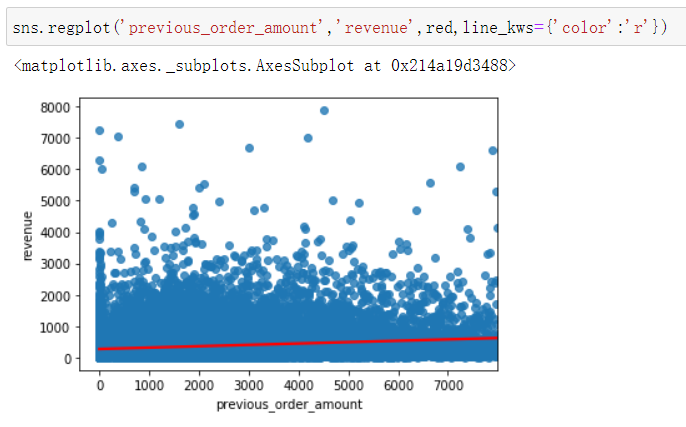
可见填充方式不论是向前还是向后都没有明显的差异，异常值的处理对模型的拟合效果提升有很大的帮助。

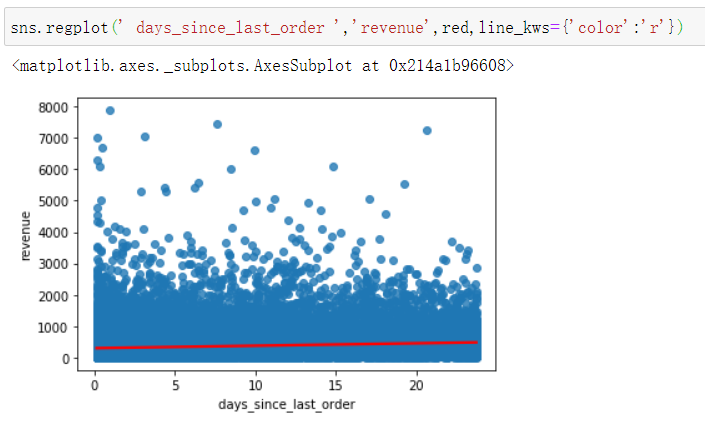


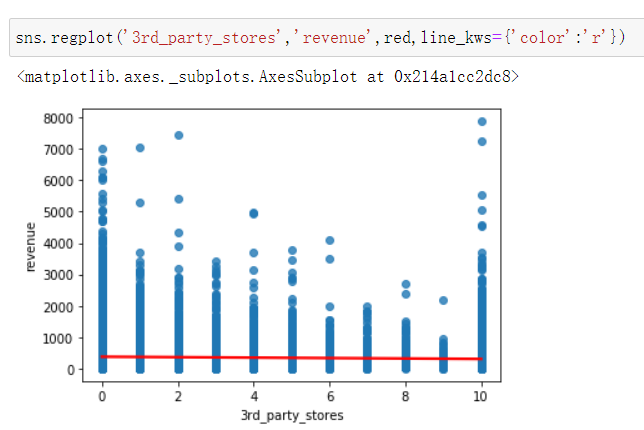
从模型评估中可以看到，R评分较低，证明模型自变量的影响较小，Prob（F-statistic）<0.05，通过了F显著性检验，P（t）也小于0.05,通过了t检验，AIC和BIC较大，模型拟合不佳，Cond. No.较大，存在共线性问题，后续需要进行逐步回归、PCA等。

1. 预测用户的消费金额变化

结果：







分析：用户的消费金额会随着累计订单的金额以及上次下单的时间的增加而有一定的提升，其原因可能是用户会随着在小红书的累计花费的增加而产生一定的用户粘性，从而增加消费投入到小红书的可能性，距离上次订单的时间的增加，用户的购买欲会随之上升，可见小红书用户所比较青睐的产品应该是更新换代较快的，较为潮流的产品。在第三方app中的数量无法进行有效的回归分析。

1. 找到对用户消费影响较大的因素

对用户消费影响较大的因素，应该是用户的订单累计金额以及下单的时间，因为根据小红书的消费主要人群，是一些喜欢新鲜事物的年轻人，所以用户的消费金额决定了这些人对于小红书这一app的认可程度，用户的粘性，而下单的时间长短，也决定了小红书的消费是具有时效性的，消费人群是需要定期消费的，因此小红书需要在留住一些忠实客户的基础上，也要关注一些高新时髦产品的推广和引进。实际上，用户的活跃性也有着一定的决定性作用，在ols表中也可以看出，活跃的用户会更有倾向去消费，而不活跃的用户则不太可能去消费，因此小红书可以给出一些比如留言抽奖、有奖提供建议等活动，调动用户在该产品上的积极性，也会对产品的消费金额的提升有所帮助。