一、实验要求

数据集：Hotel booking demand，酒店预定需求

数据集信息：包含城市酒店和度假酒店的预订信息，包括预订时间、入住时间、成人、儿童和/或婴儿的数量以及可用停车位的数量等信息。

数据量：32列共12W数据。

基于这个数据集，可进行以下问题的探索：

（1）基本情况：城市酒店和假日酒店预订需求和入住率比较；

（2）用户行为：提前预订时间、入住时长、预订间隔、餐食预订情况；

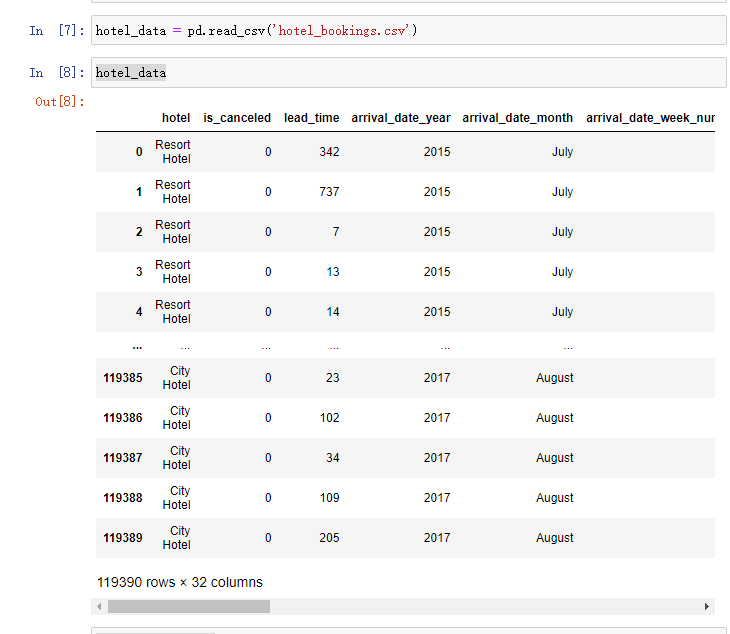
（3）一年中最佳预订酒店时间；

（4）利用Logistic预测酒店预订。

（5）也可以自行发现其他问题，并进行相应的挖掘。

二、数据处理

（1）我们导入了数据集，并先观察了数据集的大体分布情况及信息：



从上面的输出中可以看出，Hotel booking demand数据集的主要包括以下特征：

is\_canceled:标志这间房间是否已经被取消。

lead\_time:预订日期和到达日期之间经过的天数

arrival\_date\_year:到达的年份

arrival\_date\_month：到达的月份

arrival\_date\_week\_number：到达的星期

arrival\_date\_day\_of\_month：到达的日期

stays\_in\_weekend\_nights：客人在周末预订或留宿的天数

stays\_in\_week\_nights：客人在周中预订或留宿的天数

adults：大人的数量

children：孩子的数量

Babies：婴儿的数量

meal：订餐的类型

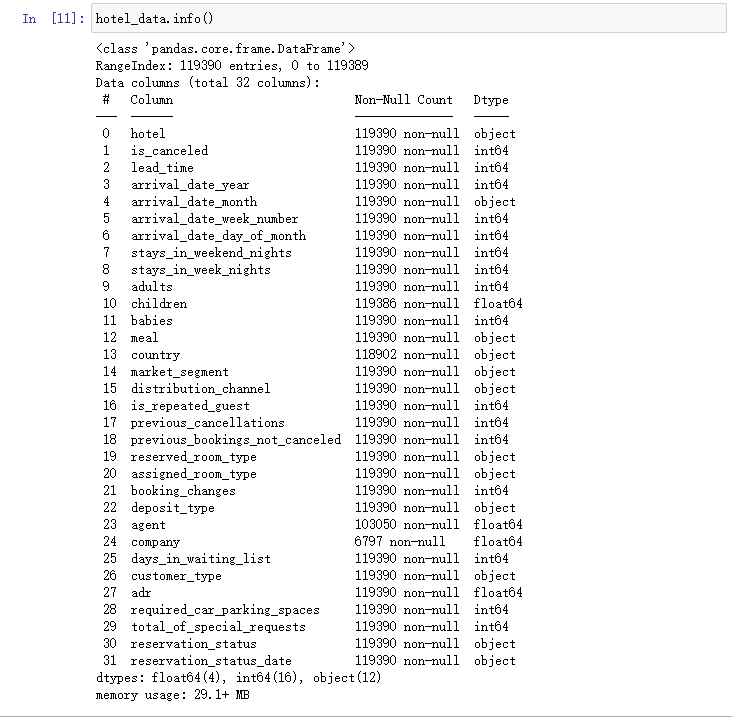
market\_segment：市场名称

distribution\_channel：订购渠道

is\_repeated\_guest：是否为回头客

country：国家

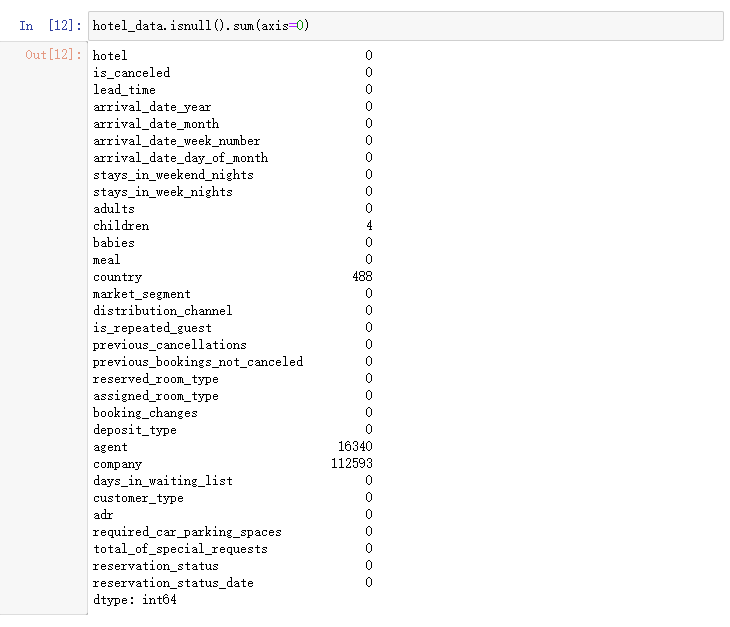
利用info（）函数查看了hotel\_data二维表的详细信息和数据缺失情况如下：



从上面可以看到，children, country, agent, company 这四个variable 需要填补缺失数据。

（2）数据缺失值处理

根据hotel\_data.info()，知道有32个特征，且存在特征为NULL，但并不直观，因此能够用.isnull().sum()来求出每一特征有多少个NULL值。



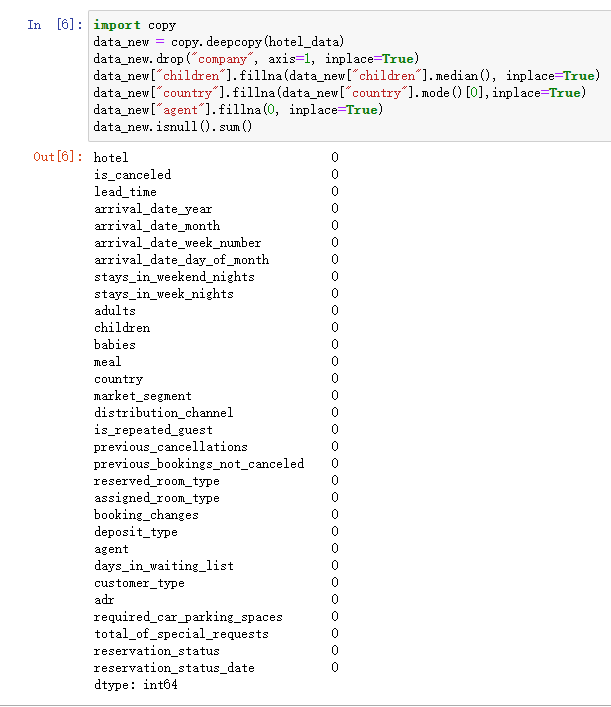
Company特征缺失值112593/119390超过80%，因此能够直接删除；

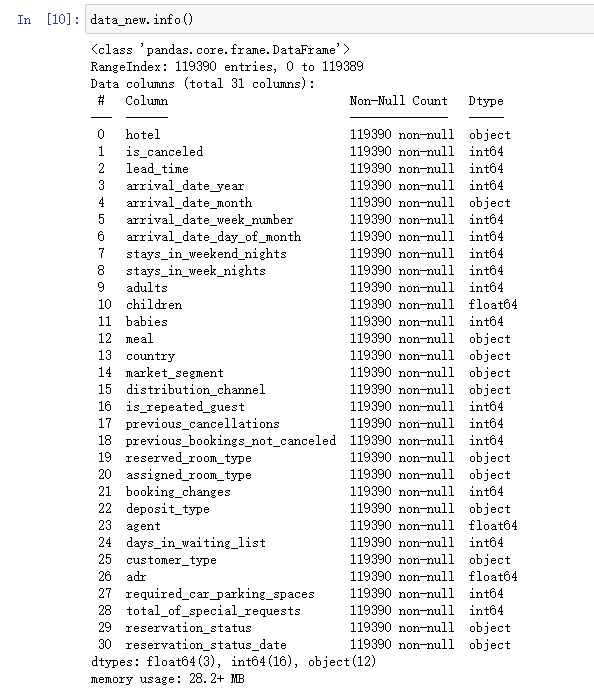
children字段中是数值型变量，且偏态分布，即用中位数替代

country字段为类别型变量，用众数替代

假设agent中缺失值表明未指定任何机构，即null=0

处理后数据如下：





三、基本情况

城市酒店和假日酒店预订需求和入住率比较

根据数据可以得知共有两种酒店，分别是城市酒店和度假酒店，首先查看两者的预约状况：通过下图可以看出，城市酒店的取消比例相对比较高，而度假酒店的入住率更高。



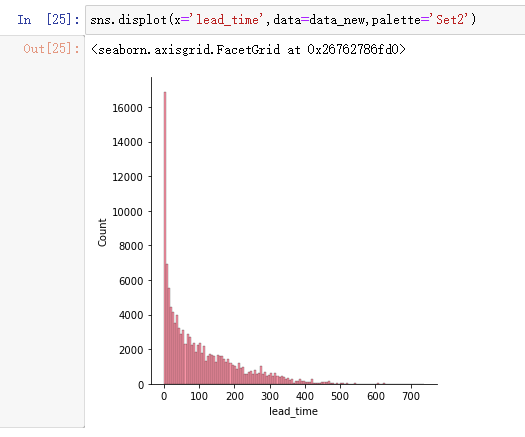
新老顾客取消情况：可以看到无论是哪种类型的酒店，新客取消预订的占比远大于老客，并且老用户相对新用户较少，老用户更少取消预订。



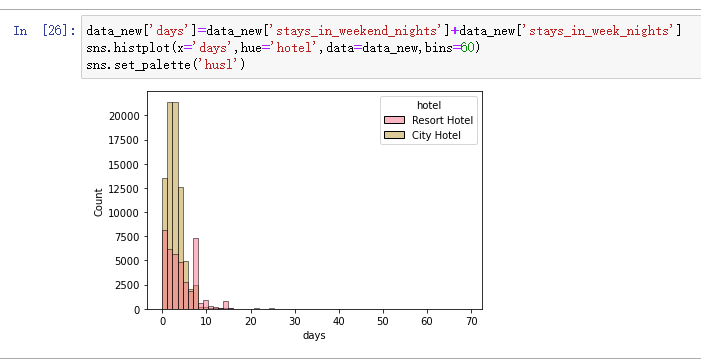
四、用户行为

提前预订时间、入住时长、预订间隔、餐食预订情况

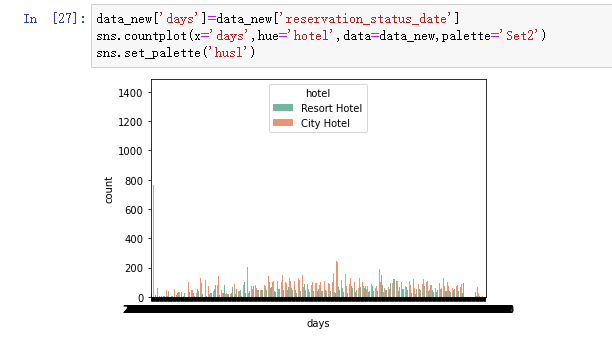
提前预定时间：由下图可以看出，大部分用户都会选择提前1、2天预订，有的用户甚至提前超过一年就预定了。



入住时长：城市酒店和度假酒店入住客人的入住时长一般都不会超过10天，可以发现度假酒店的入住时长普遍都要大于城市酒店，前者的中位数为4天，后者为2天，入住时间大于15天的也是度假酒店的数量更多。



预订间隔：可以发现预定间隔都为0天，意味着每天都会有人来预定酒店

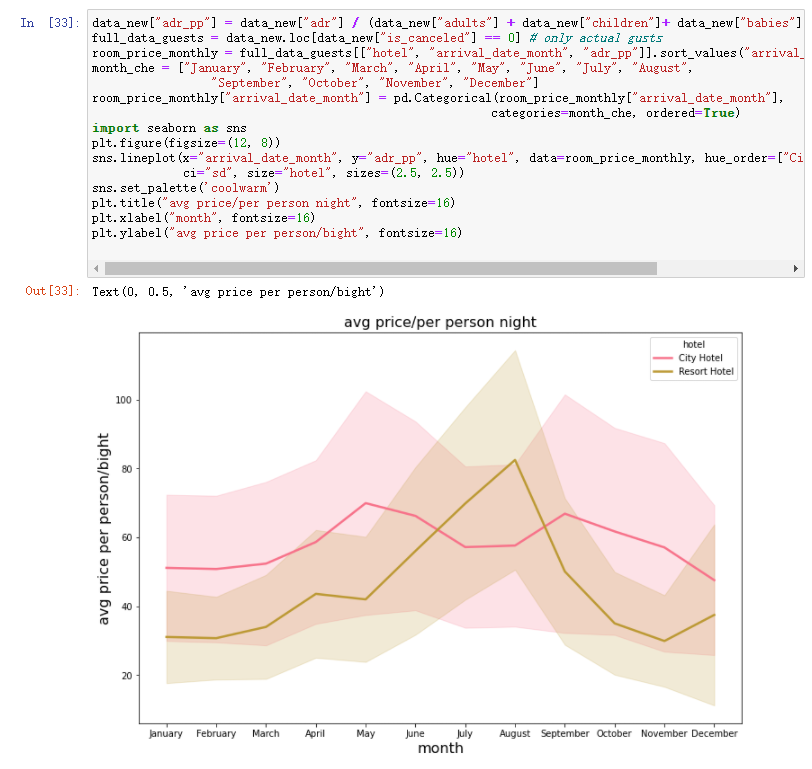


餐食预订情况：两种酒店的餐食预定情况是很类似的，大部分用户都会订餐，其中订早餐所占比例最高。而度假酒店的订餐比例要超过城市酒店，三餐都订的比例很少。



五、一年中最佳预订酒店时间

从下面两图可知：城市酒店在5月份和9月份人均价格有两个小高峰，在7-8月份价格降低；度假酒店在7-8月份人均价格处于高峰阶段测。并且，城市酒店在夏季降价，度假酒店在夏季涨价，并且，度假酒店夏季远超城市酒店价格，城市酒店价格起伏不大；城市酒店4-5月（春季）和9-10月（秋季）为预约旺季，房价也相应提升，度假酒店3-5月份（春季）和10月份（秋季）为预约旺季，房价稍有上浮，对于两家酒店来讲6-8月份均为淡季，但发现度假酒店在7-8月淡季反而房价很高，远高于其余月份，11月-来年1月份（冬季）也是预约淡季。推荐在人少和价格低的淡季去旅行。

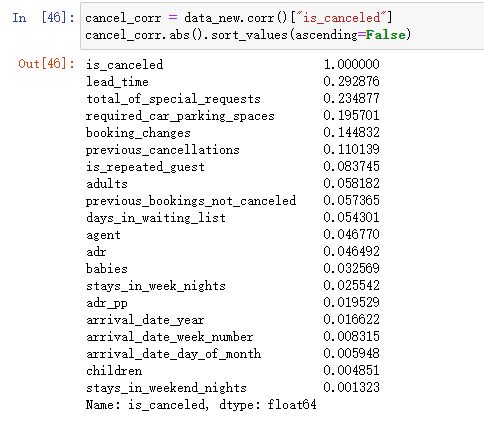




五、利用Logistic预测酒店预订

计算每一个特征与"is\_canceled"的相关性，因为有些是类别变量，因此不能参与计算，去除与预测无关的几个属性:agent、company、hotel以及reservation\_status\_data,。is\_canceled属性反应了客户最终是否取消了预定，因此可以将该属性看作特征属性，划分特征和标签。

我们首先计算每一个特征与"is\_canceled"的相关性，因为有些是类别变量，因此不能参与计算：

可知：除了"is\_canceled"外，前5个（从预约到到店时间，客户提出的特殊要求的数量，要求停车场，对预订进行的更改的数量，客户在当前预订以前取消的先前预订的数量）与"is\_canceled"相关性较大

由上图可知：除了"is\_canceled"外，前5个（从预约到到店时间，客户提出的特殊要求的数量，要求停车场，对预订进行的更改的数量，客户在当前预订以前取消的先前预订的数量）与"is\_canceled"相关性较大。

利用sklearn.linear\_model来创建逻辑回归模型，并进行训练测试：



训练后在训练集和预测集上都取得了80.8%以上的准确率，说明所选特征以及训练方法比较准确。