

**研究生课程报告**

**分类、预测与聚类**

|  |  |
| --- | --- |
| 学 院 ： | 计算机学院 |
| 专 业 ： | 电子信息 |
| 指 导 教 师 ： | 汤世平 |
| 姓 名 ： | 李云彤 |
| 学 号 ： | 3220200910 |

2021 年 5 月 31 日

1. **实验要求**

数据集：Hotel booking demand，该数据集包含城市酒店和度假酒店的预订信息，包括预订时间、停留时间，成人/儿童/婴儿人数以及可用停车位数量等信息。

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；

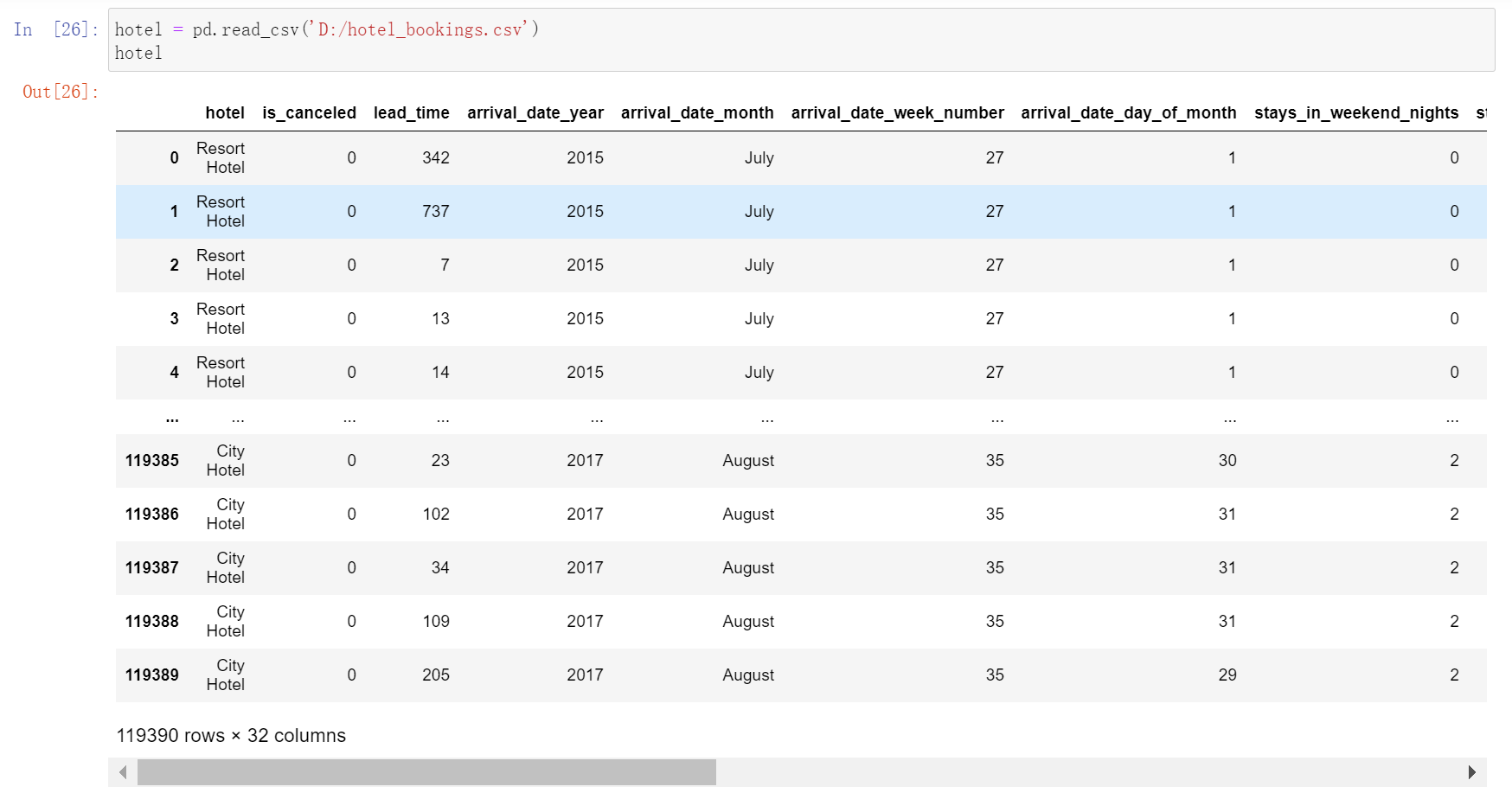
（3）一年中最佳预订酒店时间；

（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。

1. **数据预处理**

首先导入数据集，显示数据集的大体分布情况：



Hotel booking demand数据集的主要属性包括：

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

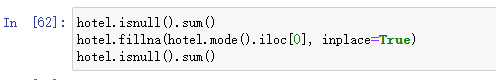
market\_segment：市场名称

distribution\_channel：订购渠道

is\_repeated\_guest：是否为回头客

country：国家

进行缺失值处理，由于出现缺失值的情况并不多，因此利用出现最频繁的数据来填补缺失值：



**三、基本情况与用户行为分析**

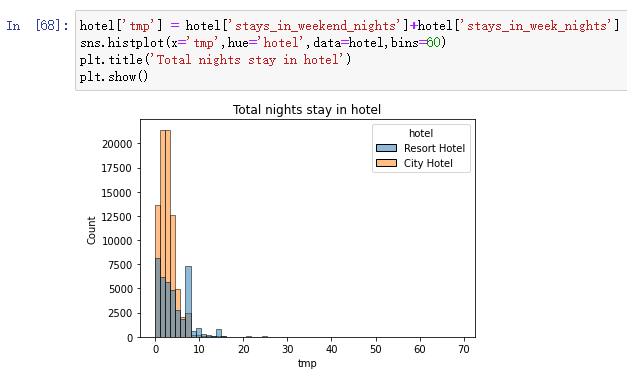
首先观察入住和取消情况，通过图中可以看出城市酒店的取消比例相对比较高，而度假酒店的入住率更高。



新老客户入住和取消占比：可以看出酒店的老用户数量相较于新用户的数量较少，而且新用户取消预订的比例较高，老用户的取消比例很低。



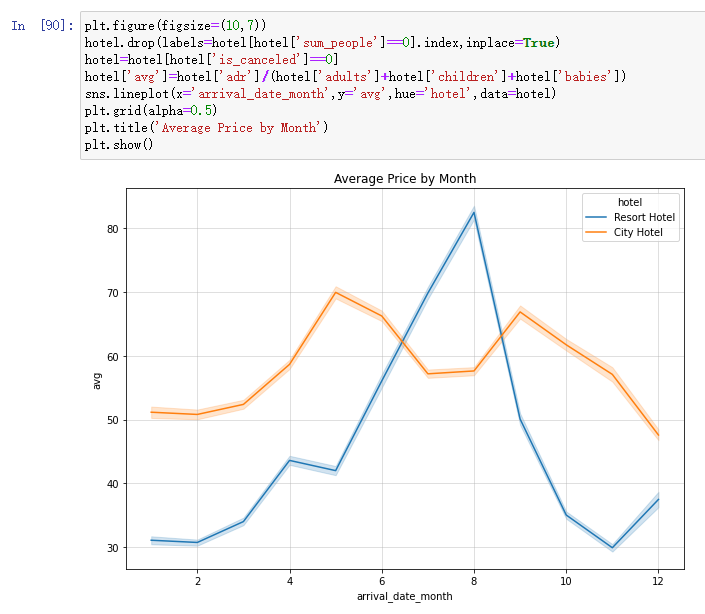
用户入住时长分析：城市酒店和度假酒店入住客人的入住时长一般都不会超过10天，城市酒店大部分客户的居住时间为1-4天，而度假酒店的客户的居住时间大部分为1天或7天。



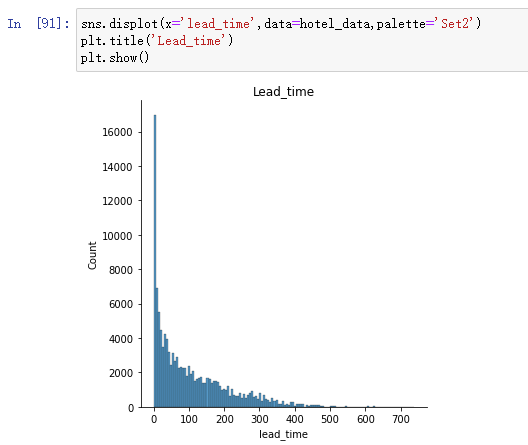
订餐情况：大部分用户都会订餐，其中订早餐所占比例最高。而度假酒店的订餐比例要超过城市酒店，三餐都订的比例很少，可能是因为中午客户一般不会回酒店就餐。



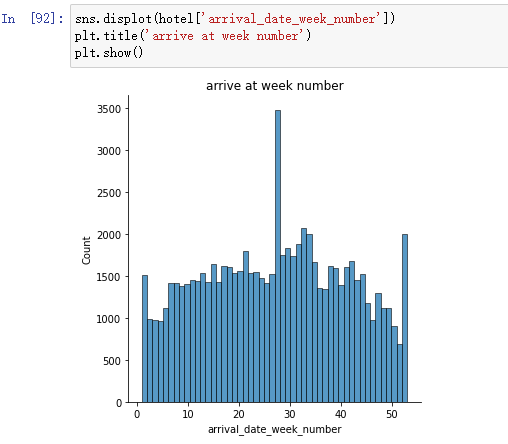
人均花销分布：城市酒店的价格在夏季降低，而度假酒店的价格会在夏季显著升高，甚至接近城市酒店同期价格的两倍，可能由于夏季人们倾向于去度假放松。城市酒店的价格分布比较均匀，而度假酒店价格起伏很大，推荐人少钱少的淡季去旅行。



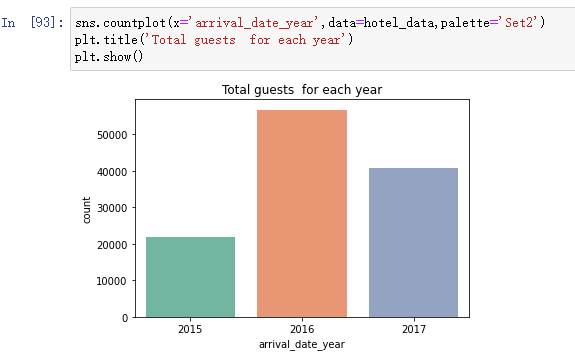
客户提前预定情况分布：大量的用户会选择提前1-2天预定，几乎占据所有客户数量的一半，但仍然有不少的用户会提前超过一年来进行预定。



客户到达的周数分布：一年之内每周客户到达的数量总体稳定，7-8月份到达的用户数较多，而1-2月份到达的用户数较少。



每年客户总数分布：可以看出2016年的用户总量最多，甚至达到了前一年用户总量的2倍以上，而2017年的用户总量有所下降。



每月客户数量分布：可以看出每月的用户数量大体分布趋势和每周的分布情况类似，7-9月份的人数明显上升，同时全年城市酒店的预定数量都要超过度假酒店。



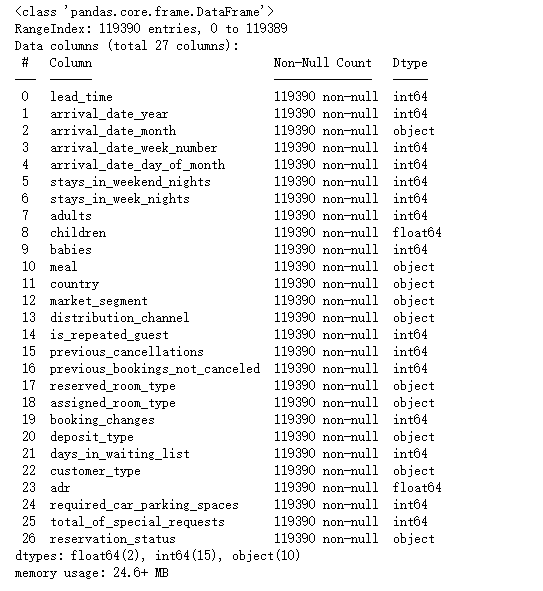
**四、利用Logistic预测酒店预订**

首先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：

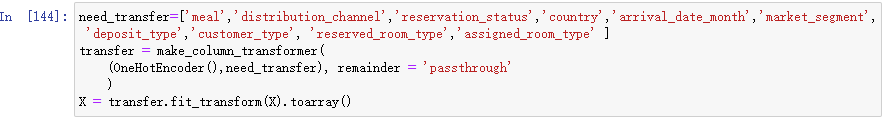


其中X作为特征，Y作为标签。查看X的大体分布：

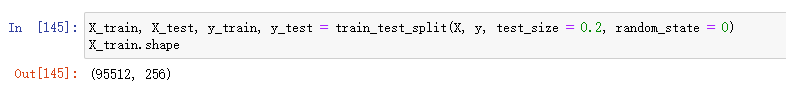




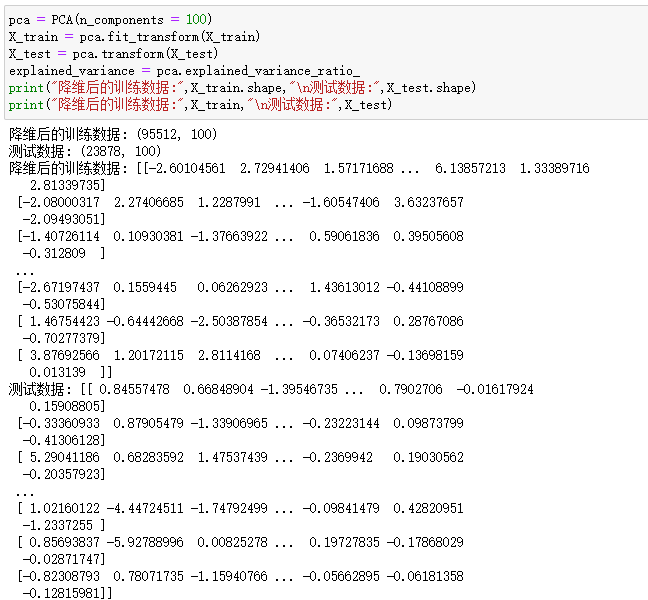
由于逻辑回归算法只能处理数值特征，而数据集中的特征既包含数值特征又包含非数值特征，所以使用python的sklearn库中的OneHotEncoder函数将非数值特征转化为数值特征再进行后续的处理：



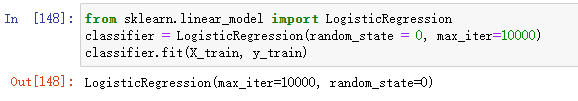
划分数据集，将80%的数据作为训练集，20%的数据作为测试集：



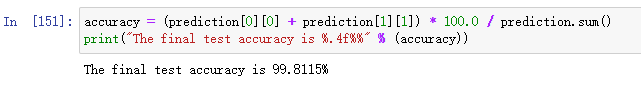
对于非数值特征要进行进一步的处理，将非数值特征转化为数值特征后数据的维数增加至了256，需要进行降维。将数据集进行标准化后利用python的PCA库进行降维，将数据维数降低至100。



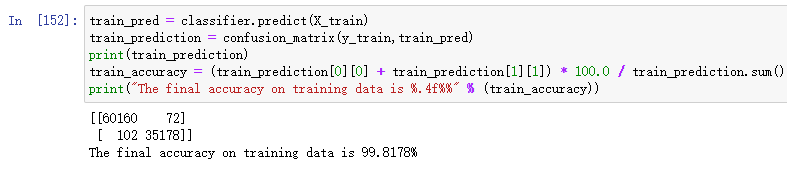
利用python的Sklearn库的LogisticsRegression函数对训练集进行训练：



对训练集进行预测：



对测试集进行预测：



训练后在训练集和预测集上都取得了99.8%以上的准确率，说明所选特征以及训练方法比较准确。