代码分为4个部分：

table\_merge 将其余表格的有用信息加入到train和test这3张表格中。

preprocess\_1 对merge之后的表格进行基本的处理，比如某些特征由string转换成int，便于之后算法训练能顺利进行，不会因为数据类型等其他原因而报错。

baseline3\_split\_dataset 将train和test这些完整的表格，按照日期划分成训练用，验证用，测试用的分离的数据集。

baseline3 用分离的数据集进行模型的训练，验证，测试集提交等一系列流程。

【table\_merge】

从母酒店维表提取出glon和glat

从物理房型维表和子酒店维表提取出totalrooms

将这些特征分别加入到训练和测试的订单数据维表中。

获得以\_update结尾的3张表格。

【preprocess\_1】

对\_update后缀的表格进行进一步处理。

对object类型的特征（supplierchannel等）进行类别值编码，由string转变成int类型。

从特征orderdate中提取出hour和minute特征。

根据特征arrival\_day和etd\_day，获得特征day\_stay，表示房客的住房总天数。

将处理后的结果以\_pre2结尾保存。

【baseline3\_split\_dataset】

对特征zone进行统一化处理：将-1,0,np.nan三种都代表null的形式，统一化为-1表示。

对特征city进行低频值去除：出现频次少于3次的（即1到2次）city名称，用np.nan代替，作为缺失值处理。

根据orderdate将完整的数据集做如下分裂：

提取日期在8月31日到9月13日的数据作为train，预测9月14日的无房概率（noroom）；

提取日期在9月1日到9月14日的数据作为train，预测testA和B的无房概率。

【baseline3】

使用分离得到了两个训练集，训练模型，并获得test的预测结果。

模型使用单xgboost。