本代码是使用sklearn库的线性回归模型和kaggle的“House price prediction”数据集进行模型训练，对房价进行预测。

设计思路如下：

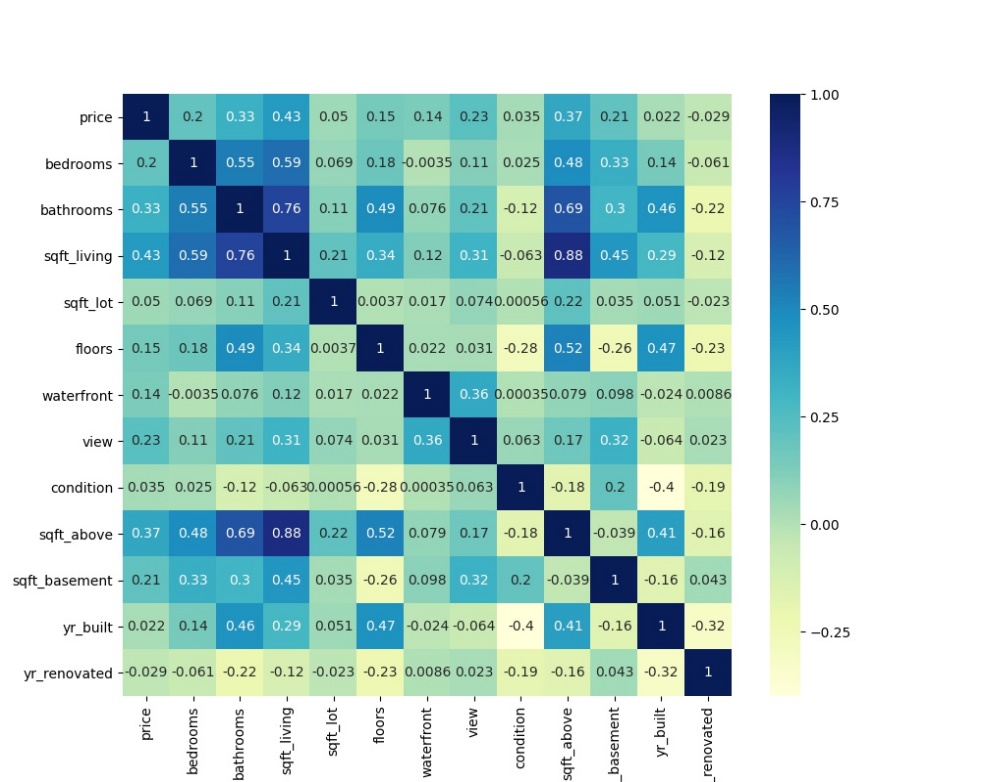
1. **下载数据集**

网址：

<https://www.kaggle.com/shree1992/housedata?select=data.csv>

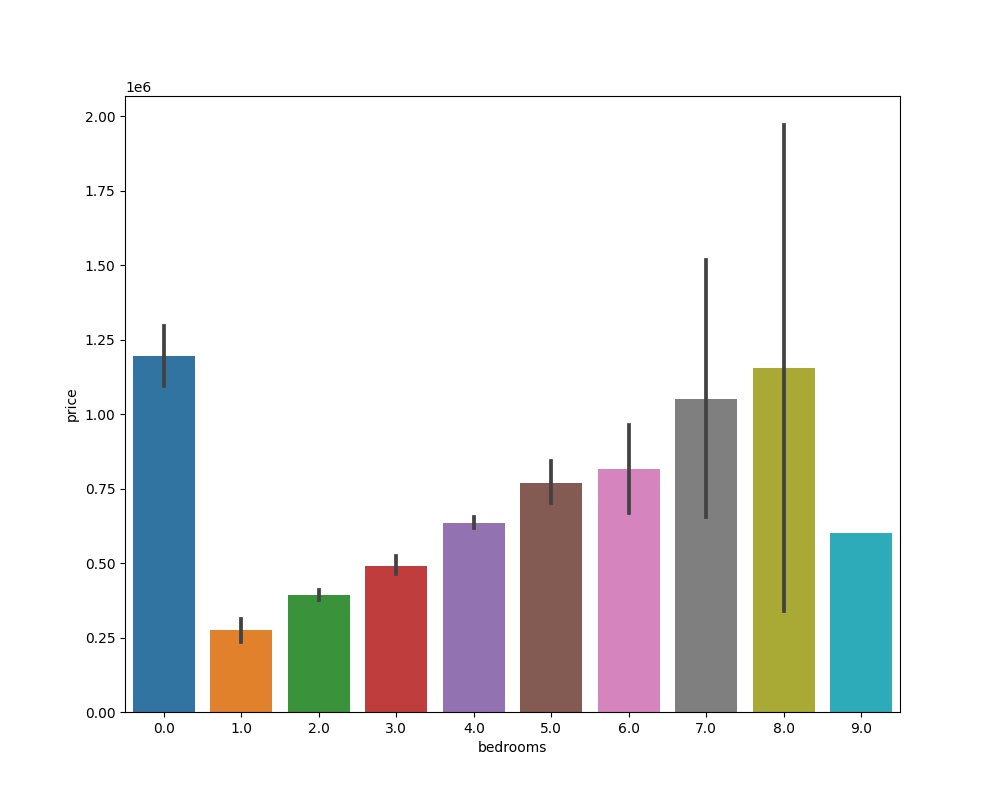
1. **对数据进行分析处理：**

数据中拥有多个属性，18个列，对房价造成或多或少的影响。因为不知道影响程度如何，所有先计算每个属性与房价的关联关系，用图片展示。

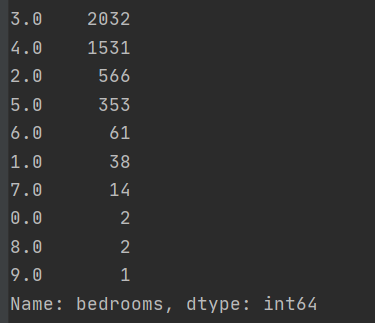


但图片只能展示数值之间的相关性，例如street，city，statezip等文字数据无法分析出。

1. **先对bedrooms进行分析**，得到了bedrooms的数量和房价的关系图



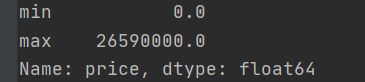
发现bedrooms在1~8时，呈现线性关系，而0和9很突兀，不符合。查看具体bedrooms的分布



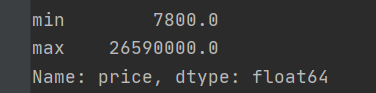
发现是因为bedrooms数量为0和9的数据量过少，所以会产生表格上的突兀，删除这些行。

1. **分析price数据**

查看下房价的最大最小值



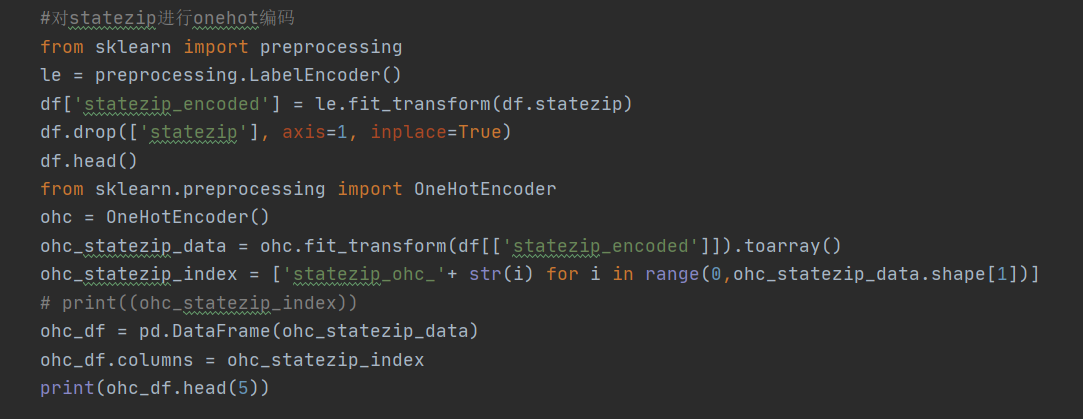
发现有房价为0的错误数据，共49行，删除这些行。



删除后，结果如上图。

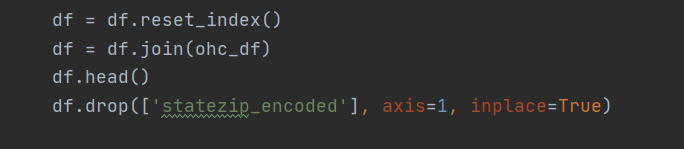
1. **处理statezip数据**

因为它是string类型的，不能用在sklearn的线性回归模型。就尝试使用独热编码，转为数字。

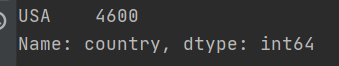


处理的时候，发现得到的只是array（），还是需要自己去弄index的名字，然后创建DataFrame。之后，和原来的df进行join操作，再去除statezip这一列string数据。

这里遇到了坑，因为之前已经删除了一些行，然后join的话，会出现NaN的情况，导致后面调用线性回归报错。需要再reset\_index()



1. **处理city、street、country等数据**

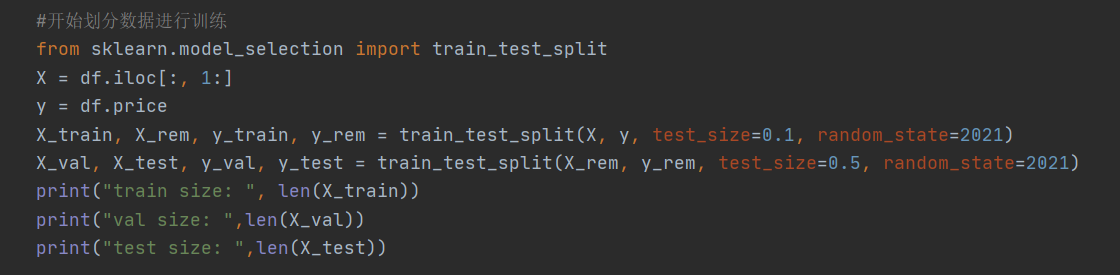


Country全是美国，删掉这一列。

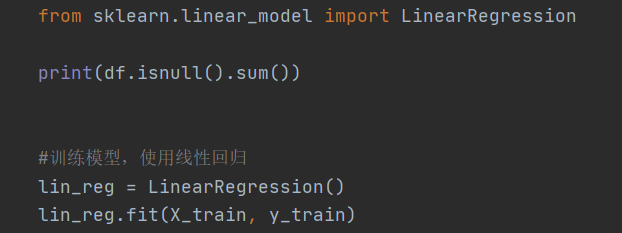
Street、city已经被statezip包含了这些信息，不想再弄独热编码，直接删了。

1. **训练模型：**

先划分数据，分为 train，validation，test三部分

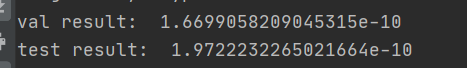


使用线性回归模型：



1. **利用模型进行预测，评估模型并上传预测结果至S3：**

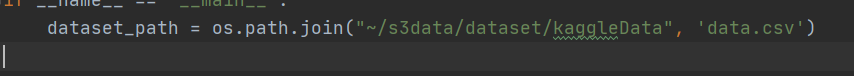
使用均方差作为误差分析（该指标计算的是拟合数据和原始数据对应样本点的误差的平方和的均值，其值越小说明拟合效果越好）。



发现结果很强

使用方法：

1. 使用pycharm，配置好与服务器的连接。
2. 上传dataset至S3，在homework的package的init文件中，设置dataset路径



1. 依次点击【Tools→Deployment→Upload to …】，确保将程序上传到服务器上
2. 点击运行，等待结果输出

