本程序是采用线性回归模型和kaggle的数据集进行模型训练，对美国金县的房价进行预测。此外还使用Xgboost模型，作为对比。

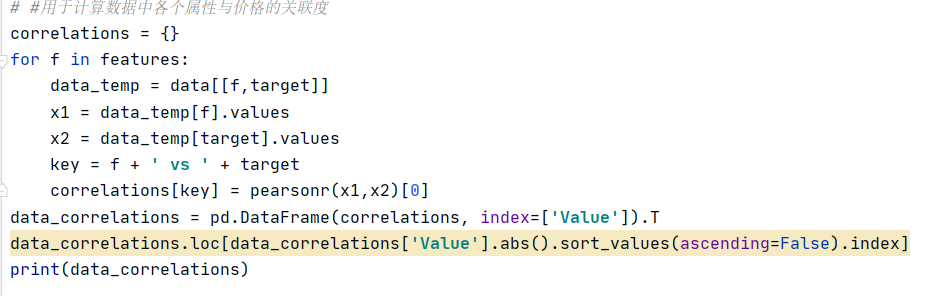
设计思路如下：

1. **下载数据集**

网址：https://www.kaggle.com/harlfoxem/housesalesprediction

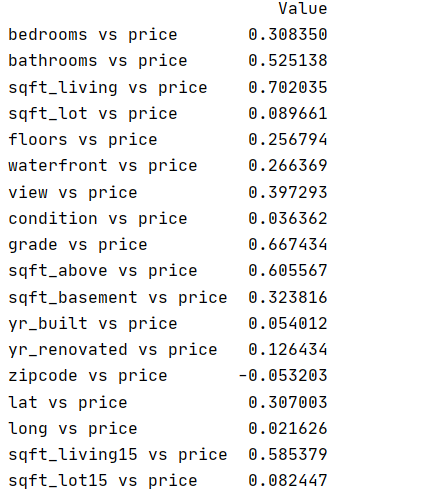
1. **对数据进行处理：**

因为数据中拥有多个属性，都能对房价造成影响。所以计算每个属性与房价的关联关系，并按从大到小排序。从中选择与价格关联度最高的前十个属性，将其作为新的数据集。这样子可以提高训练出来的模型预测的精准度。输出关联度的代码如下：



前十个与价格关联度最高的属性为：'sqft\_living', 'grade', 'sqft\_above', 'sqft\_living15', 'bathrooms', 'view', 'sqft\_basement', 'bedrooms', 'lat', 'waterfront'。

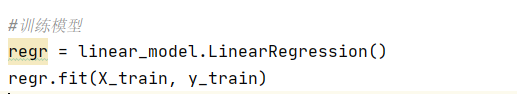
该段代码执行结果如下所示：



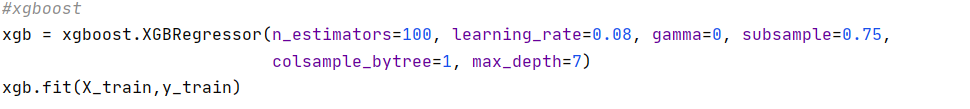
使用sklrean. train\_test\_split（）函数，对数据集进行划分 ，切分成训练集和测试集

1. **训练模型：**

使用线性回归模型：



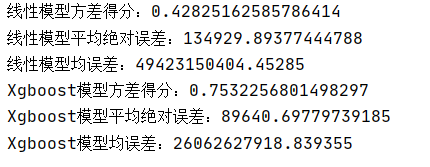
使用Xgboost模型



1. **利用模型进行预测，评估模型并上传预测结果至S3：**

预测模型是采用模型自带的predict()函数。

评估模型线性从三个方面进行评价，分别是回归模型的方差得分（其值取值范围是[0,1]，越接近于1说明自变量越能解释因变量的方差变化，值越小则说明效果越差）、平均绝对误差（用于评估预测结果和真实数据集的接近程度的程度，其值越小说明拟合效果越好）、均方差（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）。并使用Xgboost模型的作为对比：



可见，线性模型的预测结果较为一般，相比较而言，Xgboost表现更好。

使用方法：

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